Book of Abstracts

The conference is supported by the DFG Priority Program "XPrag.de: New pragmatic theories based on experimental evidence".

Local organizers:

Petra B. Schumacher

Lena Straßburger

Barbara Tomaszewicz

Hanna Weiland-Breckle

Program

Tuesday, June 20, 2017

19:00	Informal get-together	Maybach

Wednesday, June 21, 2017

12:30	Registration	
13:00	Welcome	
13:15	**Jennifer Rodd**	How do listeners understand the meanings of ambiguous words?
14:15	*Coffee break*	
14:45	Catherine Davies, Kremena Koleva and Ekaterini Klepousniotou	Do speaker-specific cues influence ambiguous word interpretation?
15:15	Paolo Canal, Luca Bischetti, Simona Di Paola and Valentina Bambini	Social abilities help us detecting jokes: An EEG study on the temporal dynamics of humor comprehension
15:45	Andrea Beltrama	Subjective assertions are weak: an experimental study on perspective-dependent meaning
16:15	*Coffee break*	
16:45	Elspeth Wilson and Napoleon Katsos	Speaker epistemic state and ad hoc quantity implicatures in children
17:15	Kyriakos Antoniou, Alma Veenstra, Mikhail Kissine and Napoleon Katsos	How does childhood bilingualism and bilectalism affect the interpretation and processing of implicature?
17:45	Irene Symeonidou and Wing Yee Chow	Through the eyes of a teenager: complexity of real-time Theory of Mind inferences in language comprehension
18:15 – 19:45	Poster session I	

Thursday, June 22, 2017

9:30	**Bruno Galantucci**	Experimental Semiotics: What is it? What is it good for?
10:30	*Coffee break*	
11:00	Nausicaa Pouscoulous and Giulio Dulcinati	Quantity implicatures in a competitive game
11:30	Diana Dimitrova, Brian McElree and Petra Schumacher	Speed and accuracy tradeoff of meaning composition
12:00	Felix Frühauf, Berry Claus, Sophie Repp, Manfred Krifka and Anna Marlijn Meijer	Two response systems for German 'ja' and 'nein'? Evidence from usage preference data and interpretation data
12:30	*Lunch break*	*on your own (see restaurant map)*
14:00	Ming Xiang, Chris Kennedy and Allison Kramer	Threshold adaptation and its time course
14:30	Christina Kim and Louisa Salhi	Visual contrast, discourse contrast and conceptual convention
15:00	Judith Holler, Kobin Kendrick and Stephen Levinson	Turn-timing and the body: Gestures play a core role in coordinating conversation
15:30	*Coffee break*	
16:00 – 17:30	Poster session II	
18:30	*Conference dinner*	UndSohn (preregistration only)

Friday, June 23, 2017

9:30	**Kathryn Davidson**	Combining continuous and discrete representations in speech, sign, and gesture
10:30	*Coffee break*	
11:00	Stavroula Alexandropoulou, Jakub Dotlaèil and Rick Nouwen	Pragmatic effects attested in online interpretation of *more than* and *at least*
11:30	Teodora Mihoc and Kathryn Davidson	Testing a PPI analysis of superlative modified numerals
12:00	Alice Rees and Lewis Bott	Investigating shared representations in implying and inferring
12:30	*Lunch*	*on your own (see restaurant map)*
14:00	**Mikhail Kissine**	What Autism Spectrum Disorder can teach us about pragmatics
15:00	*Coffee break*	
15:30	Martien Wampers and Walter Schaeken	Scalar Implicatures And The Literal-First Hypothesis: Theory Of Mind And Working Memory Effects In Pragmatic Inferences By Patients With Psychosis
16:00	Bob van Tiel and Mikhail Kissine	Pragmatic impairment is selective in autism: evidence from quantity implicatures
16:30	Oliver Bott	Immediate use of discourse context in aspectual coercion - An eyetracking during reading study
17:00	Farewell	
17:30	Cultural event	Street art tour or historic city walk

Poster session I (Wednesday, June 21, 2017, 18:15 – 19:45)

1. *Alix Kowalski and Yi Ting Huang. Listeners encode multiple meanings when generating scalar inferences → canceled, but see poster on OSF*
2. Jeffrey Geiger and Ming Xiang. Ellipsis in context: The interaction of identity and discourse salience
3. Claudia Poschmann. At-issue: Non-restrictive relative clauses
4. Sophie Egger, Bettina Braun and Nicole Dehé. The realization of bouletic bias: Evidence from German questions
5. Guilio Dulcinati and Nausicaa Pouscoulousinati. Scalar implicatures in non-cooperative contexts
6. Richard Breheny, Chao Sun and Ye Tian. Rates of scalar inferences beyond 'some' – A corpus study
7. Cecília Molnár, Beáta Gyuris and Katalin Mády. Evidential bias and polar questions – the division of labour in Hungarian
8. Daniele Panizza and John M. Jr. Tomlinson. Pragmatic inferences towards prototypical meanings. A visual world study
9. Laia Mayol. Asymmetries between interpretation and production in Catalan pronouns
10. Tamás Káldi, Anna-Christina Boell and Anna Babarczy. Contextual effects on the processing of Hungarian pre-verbal focus sentences: an eye-tracking study
11. Simona Di Paola, Nausicaa Pouscoulous and Filippo Domaneschi. Metaphorical Developing Minds: The role of multiple Factors in the Development of Metaphor Comprehension
12. Erlinde Meertens, Andrea Beltrama and Maribel Romero. Polar Questions, "or not" Alternative Questions and Complement Alternative Questions: an experimental study
13. Nadine Bade. Processing antipresuppositions
14. Cécile Barbet and Guillaume Thierry. When 'some' triggers a scalar inference out of the blue. An electrophysiological study of a Stroop-like conflict elicited by single words
15. Eva Link, Holger Schneider, Kristina Schopf, Marcel Schwille, Franziska Rück and Barbara Kaup. Does it matter who is producing an utterance? – Effect of speaker identity in utterances without self-reference
16. Elisa Kreiss, Judith Degen, Robert Hawkins and Noah Goodman. Mentioning atypical properties of objects is communicatively efficient
17. Francesca Foppolo, Francesca Panzeri, Greta Mazzaggio and Luca Surian. Find a friend or a scale mate: comparing ad hoc and scalar implicatures
18. Charlotte Out, Martijn Goudbeek and Emiel Krahmer. Alignment in naturalistic dialogue: Language production in interactive reference production
19. Jarang Kwak, Haejin Kim, Soyoung Kwon and Donghoon Lee. Influence of interpersonal variables during utterance comprehension: A neurophysiological investigation with the Korean honorific system
20. Anton Benz, Nicole Gotzner and Lisa Raithel. Embedded implicature: What can be left unsaid?
21. Alma Veenstra and Napoleon Katsos. When children accept under-informative utterances: Lack of competence or pragmatic tolerance?
22. *Binh Ngo and Elsi Kaiser. Referential form production in Vietnamese: Effects of modality and topicality → canceled, but see poster on OSF*

Poster session II (Thursday, June 22, 2017, 16:00 – 17:30)

1. Stefan Hinterwimmer, Umesh Patil and Andreas Brocher. Do German demonstrative pronouns avoid prominent perspectival centers?
2. *Judith Holler, Kobin Kendrick and Stephen Levinson. Turn-timing and the body: Gestures play a core role in coordinating conversation → Talk on Thursday 15:00*
3. Paula Rubio-Fernandez and Julian Jara-Ettinger. A new Director task: Modelling common ground through referential specificity
4. Diana Mazzarella, Emmanuel Trouche, Hugo Mercier and Ira Noveck. Believing what you're told: Politeness and scalar inferences
5. Debora Rossi, Simona Di Paola and Filippo Domaneschi. The aging factor in presuppositions processing
6. Filippo Domaneschi and Simona Di Paola. The processing costs of presupposition Accommodation
7. Verena Keite, Ralf Klabunde and Eva Belke. Alternatives in processing ad-hoc implicatures
8. Corien Bary, Daniel Altshuler, Kristen Syrett and Peter de Swart. Factors licensing embedded present tense in speech reports
9. Saskia Brockmann and Nadine Bade. Evidence for global pronoun resolution
10. Myrto Pantazi, Mikhail Kissine and Olivier Klein. Automatic content accommodation: Direct perception and meta-cognitive vigilance
11. Margaret Kroll and Matthew Wagers. Interaction of parentheticals, (not-)at-issue content, and working memory
12. Heather Burnett and Barbara Hemforth. A Bayesian game-theoretic approach to cross-linguistic variation
13. Elli Tourtouri, Francesca Delogu and Matthew Crocker. Over-specification and uniform reduction of visual entropy facilitate referential processing
14. Maria Spychalska, Ludmila Reimer, Petra Schumacher and Markus Werning. Scalar implicatures in the context of full and partial information. Evidence from ERPs
15. Stephanie Solt, Jon Stevens and Brandon Waldon. "Some" approximations: an experimental investigation
16. Jérémy Zehr and Florian Schwarz. Returning to non-entailed presuppositions again
17. Ye Tian and Chris Cummins. Top-down and bottom-up cues to speech acts
18. Chao Sun and Richard Breheny. On the compositional interpretation of scalar quantifiers: The role of the residue set
19. Amanda Pogue and Michael Tanenhaus. Exploring how speakers mark, and listeners assess, certainty
20. Stanley Donahoo and Vicky Tzuyin Lai. What the hell? What swearing can tell us about conventional implicatures
21. Sarah Dolscheid, Franziska Schleussinger and Martina Penke. Different pragmatic interpretations of German 'eine' (a/one) in children and adults
22. Andreas Trotzke. Approaching the pragmatics of exclamations experimentally
23. Viola Schmitt and Daniele Panizza. What *and* means: a study on the intersective vs. non-intersective construal of VP-*and*

Canceled talk: Galit W. Sassoon, Natalia Meir, Julie Fadlon, David Anaki and Petra B. Schumacher. The acceptability, processing and neural signature of nominal gradability. → see slides on OSF

Wifi @ KOMED

Network: XPrag2017
Password: Cologne2017

Conference dinner @ UndSohn (ground floor of KOMED) on Thursday at 18:30. (Few more tickets left. Check at registration desk!)

Cultural event (Friday at 18:00)

- *Historic city walk (Old Town)*

Enjoy a guided walking tour around Cologne's Old Town and learn about the city's history. We will visit the magnificent Cologne Cathedral and the excavations at the Roncalli square including the Dionysos mosaic to learn about the 1st-century AD Roman town founded by Agrippina. We will visit the historic City Hall, the Old Market and the Heumarkt. The tour ends in Martin's quarter, a perfect spot to have a refreshing Kölsch on your own.

Duration: 1,5 hours, about 2km

Meeting point: "Kreuzblume" in front of Cathedral (tram station: Central station/Dom).
Go there with Barbara. Meet in front of KOMED at 17:30.

- *Street art tour (Belgian quarter)*

This guided walk leads you through the street art scene of the Belgian Quarter. This is a bustling meeting spot for Cologne's youth with a distinct identity. You can learn about the local and international artists like Tika, Hendrik ECB Beikirch, Sepe & Chazme and Mark Jenkins. We will visit the exhibitions of Die Kunstagentin, 30works, Ruttkowski; 68 or Kunst&So.

Duration: 1,5 hours, about 2km

Meeting point: Hahnentorburg at Rudolfplatz (tram station: Rudolfplatz).
Go there with Lena and Petra. Meet in front of KOMED at 17:30.

Pragmatic effects attested in online interpretation of *more than* and *at least*

Stavroula Alexandropoulou[1], Jakub Dotlačil[2] & Rick Nouwen[1]
[1]Utrecht University, [2]University of Amsterdam

We present results of an eye-tracking reading study on the interpretation of two types of numeral modifiers (NMs), viz., *at least* and *more than*, in three kinds of context, thereby probing the inferences triggered by such modifiers, and their status.

Motivation. Since Geurts & Nouwen (2007), it's been an uncontroversial and well-established fact that superlative NMs, unlike their comparative counterparts, trigger ignorance effects. Only very recently has this fact been called into question: Westera & Brasoveanu (2014) and Mayr & Meyer (2014) argue that comparatives too give rise to ignorance, if there is a *how many* Question Under Discussion (QUD). Coppock et al. (2016) too observe that, while an answer to a polar question, see (1), could imply that B knows the exact number of apples in the case of *more than* but not of *at least*, the use of either NM in B's answer in (2) conveys speaker ignorance, as B is explicitly asked to name the precise number of apples Joe ate.

(1) A: Did Joe eat any apples?
B: Yes, he ate *at least*/*more than* 3 apples.

(2) A: How many apples did Joe eat?
B: He ate *at least*/*more than* 3.

The present study sets out to directly probe ignorance effects with *more than* and *at least* with a *how many* QUD by means of an online experiment. We are, moreover, concerned with yet another type of inference of NMs, which has been neglected by the existing literature, namely, *speaker indifference*. In B's answer with *at least*, in either (1) or (2), if 3 is a relevant number in the context, there is an additional reading whereby B knows the exact quantity of apples but s/he regards it as relevant to only mention a lower bound, not caring about the exact number. This inference together with ignorance as well as *acknowledgment of disagreement* appears to form a family of inferences, also displayed by free relatives (Condoravdi, 2015), epistemic indefinites (Chierchia, 2013), disjunction (Lauer, 2013), and has often been treated on a par with ignorance in the sense that it is derived via a (similar) pragmatic mechanism (see, e.g., Lauer, 2013). In this study, we investigate speaker indifference effects with both *at least* and *more than*, and further evaluate their status relative to ignorance.

Present study. In order to directly examine speaker ignorance and indifference effects of *at least* and *more than*, we ran an eye-tracking reading experiment measuring what happens in real time when interpreting those NMs in a context with an ignorant, an indifferent or a plain knowledgeable/authoritative speaker, and an implicit *how many* QUD. So we manipulated the factors **Context** and **NM** in a 3×2 design. Dutch native speakers read texts in Dutch like the following (translated into English; *target* is in glosses):

> *Intro*: Sophie is a figure skater and very dedicated. Normally, she trains for four hours in the weekend, but last weekend she trained as intensively as possible.
> IGNORANCE: I'm not sure how much exactly, but this is what I think:
> INDIFFERENCE: I could tell you exactly how much, but it's not that important.
> AUTHORITY: I can tell you how much because I talked to her yesterday.
> *Target*: Sophie has last weekend AT LEAST/MORE THAN eight hours on the ice practiced.

The context setup is inspired by Breheny et al. (2006), who tested the online interpretation of scalar terms in a self-paced reading task. They found a slowdown at the region of the scalar expression when the preceding context supported a scalar implicature vs. when being compatible with the lower-bound-only reading, and attributed this finding to online implicature calculation being costly. Our starting point is an analysis in which ignorance and indifference inferences behave in a way fully parallel to the finding of Breheny et al. (2006) on scalar implicatures: ignorance and indifference inferences are computed online

and come about via a costly pragmatic mechanism. If such an analysis is on the right track, we expect to find a slowdown in IGNORANCE and in INDIFFERENCE contexts at the modified numeral. Lastly, the semantic meaning of *at least* and *more than* is fully compatible with AUTHORITY contexts. Such contexts are incompatible with ignorance effects, but not with indifference effects, so such contexts are expected to yield no or optional indifference inferences. For this reason, AUTHORITY was the baseline for the **Context** factor in our study. MORE THAN was the reference level for the **NM** factor. We tested 36 items, with 72 fillers and a Latin square design. 37 native speakers of Dutch (33 female, mean age: 23.7, age range: 18–42) participated in the experiment. The **NM** type did not affect text coherence ($z = .840$, $p = .401$) in a pretest where people had to judge how compatible the *Target* is given the preceding context, on a Likert scale from 1 (*not compatible*) to 7 (*compatible*).

Results & discussion. Linear mixed-effects regression analyses revealed:

(i) a processing penalty for IGNORANCE contexts with AT LEAST at the region of "eight hours" in re-reading probability (positive AT LEAST*IGNORANCE interaction in overall analysis & positive IGNORANCE effect in AT LEAST subset analysis, both $z > 2$, $p < .05$) as well as for INDIFFERENCE contexts with AT LEAST. In a previous experiment testing *at least* (Alexandropoulou et al., 2016), we found the same effect in IGNORANCE contexts (vs. AUTHORITY) at the region "eight hours", where the interpretation of the whole modified numeral phrase is completed. We interpreted this effect as being due to ignorance implicature calculation, in support of pragmatic accounts of ignorance like that in Büring (2008) or in Schwarz (2016) (a.o.), which derive ignorance as a Quantity implicature. As the present study gets rid of previous possible confounds (e.g., using a round number in *Target* or introducing a contrast with another number, see *Intro*, makes the ignorant speaker's *Target* utterance more natural), the replication of our previous finding strengthens our conclusion that ignorance with *at least* is a pragmatic inference computed online. The INDIFFERENCE effects we found are likewise to be attributed to a costly pragmatic mechanism responsible for the derivation of indifference effects, exhibiting a status similar to that of ignorance.

(ii) a slowdown in IGNORANCE contexts with MORE THAN at the spillover region "on the ice" (negative AT LEAST*IGNORANCE interaction, positive IGNORANCE effect in overall & in MORE THAN subset analyses, in early & late measures, all $t/z > 2$, $p < .05$) and likewise in INDIFFERENCE contexts with MORE THAN in "last weekend" (where subjects already see *more than*) up to "on the ice". One possibility is that these effects suggest that ignorance and indifference inferences are available with *more than* too, and are in fact derived by a pragmatic process. This would go against the claim that *more than* has no ignorance implication (see Coppock and Brochhagen, 2013). Another possibility, which could potentially explain the different processing profiles of the two NMs in our experiment, is that the attested processing cost is due to a Manner implicature: subjects find *at least* a better cue to ignorance and indifference and, hence, wonder why the speaker did not use *at least* instead, with this reasoning inducing a slowdown (cf. Degen & Tanenhaus, 2011 for similar results due to competition between *some* and number terms).

Conclusion. We provide evidence of the unexplored speaker indifference effects with numeral modifiers and of their pragmatic status, similar to ignorance. Furthermore, we replicated our previous finding suggesting that ignorance effects of *at least* are pragmatic inferences that are computed online. Crucially, we found a processing penalty for *more than* in contexts with an ignorant or an indifferent speaker, showing that pragmatic reasoning is involved in real-time comprehension of *more than*, too, in such contexts, as a Manner and/or an ignorance/indifference implicature. More generally, our findings contribute extra evidence that pragmatic reasoning occurs online and is costly.

Selected references: 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*. **Geurts, B. & Nouwen, R. (2007).** At least et al.: The semantics of scalar modifiers. *Language*. **Lauer, S. (2013).** *Towards a dynamic pragmatics*. **Schwarz, B. (2016).** Consistency preservation in Quantity implicature: The case of *at least*. *S&P*.

References

Alexandropoulou, S., Dotlacil, J., and Nouwen, R. (2016). *At least* ignorance inferences come at a processing cost: Support from eye movements. In *Proceedings of SALT 26*. Austin, Texas.

Breheny, R., Katsos, N., and 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.

Büring, D. (2008). The least *at least* can do. In Chang, C. B. and Haynie, H. J., editors, *West Coast Conference on Formal Linguistics (WCCFL)*, volume 26, pages 114–120, Somerville, Massachusetts. Cascadilla Press.

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

Condoravdi, C. (2015). Ignorance, indifference, and individuation in *wh*-ever. In Alonso-Ovalle, L. and Menéndez-Benito, P., editors, *Epistemic Indefinites: Exploring Modality Beyond the Verbal Domain*, pages 213–243. Oxford University Press.

Coppock, E. and Brochhagen, T. (2013). Raising and resolving issues with scalar modifiers. *Semantics & Pragmatics*, 6(3):1–57.

Coppock, L., Ciardelli, I., and Roelofsen, F. (2016). Implicatures of modified numerals: quality or quantity? Talk at Sinn und Bedeutung 21, Edinburgh, September 5.

Degen, J. and Tanenhaus, M. (2011). Making Inferences: The Case of Scalar Implicature Processing. In Carlson, L., Hölscher, C., and Shipley, T., editors, *Annual Conference of the Cognitive Science Society*, volume 33, pages 3299–3304.

Geurts, B. and Nouwen, R. (2007). At least et al.: The semantics of scalar modifiers. *Language*, 83(3):533–559.

Lauer, S. (2013). *Towards a dynamic pragmatics*. PhD thesis, Stanford University.

Mayr, C. and Meyer, M.-C. (2014). More than "at least". Talk at "Two days at least" workshop, Utrecht, September 10.

Schwarz, B. (2016). Consistency preservation in Quantity implicature: The case of *at least*. *Semantics & Pragmatics*, 9:1–47.

Westera, M. and Brasoveanu, A. (2014). Ignorance in context: The interaction of modified numerals and QUDs. In Snider, T., D'Antonio, S., and Weigand, M., editors, *Semantics and Linguistic Theory (SALT)*, volume 24, pages 414–431.

How does childhood bilingualism and bi-dialectalism affect the interpretation and processing of implicature?

Kyriakos Antoniou[a, b, c], Alma Veenstra[a, b], Mikhail Kissine[b], & Napoleon Katsos[a]
[a]University of Cambridge
[b]Université Libre de Bruxelles
[c]Hellenic Open University

Past research has revealed a negative effect of bilingualism on vocabulary proficiency and a positive effect on pragmatics, Theory of Mind, and executive control (see in [1]). Focusing on pragmatics, studies with pre-schoolers reported superior bilingual performance in detecting violations of Gricean maxims and in understanding scalar implicatures (SIs) (see in [1]). A recent study, however, found no differences between older multilingual, bi-dialectal, and monolingual children (6-9 years) in various implicatures (e.g. novel metaphors and SIs) [1].

In this study, we aimed to investigate whether a bilingual advantage might be found for implicatures that have not been previously tested (irony, contrastive implicatures) and whether such an advantage might be evident at the processing level. Based on [1], we hypothesised that there would be no group differences for most implicatures. However, we expected that a bilingual advantage might be found in irony for two reasons. First, irony is the most difficult and late-developing implicature for children [2]. It has been suggested that a bilingual cognitive advantage for older children and young adults might be found only when using very demanding and more sensitive tasks [3]. Second, a previous study reported that bilingual children (like adults) relied more than monolinguals on tone of voice to judge a speaker's emotion, but only when the paralinguistic cue was inconsistent with semantic content (possibly because bilinguals used their superior inhibition to focus on intonation) [4]. This situation resembles irony where intonation indicates a different interpretation than the utterance's literal meaning. Bi-dialectals were tested because (1) it has been suggested that they show smaller language delays than bilinguals due to the close similarity of their dialects (and language affects implicature) [1]; (2) bi-dialectals can be recruited from the same country and schools as monolinguals and, hence, cultural differences between the two groups are minimal (and cannot confound results).

Forty-four bilingual (in Dutch and French; 121-144 months old) and 46 bi-dialectal children from Belgium (in Dutch and West Flemish; 121-155 months old), and 48 Dutch-speaking monolinguals from the Netherlands (ages 121-145 months) were given: (1) a picture-selection task (in Dutch) on implicatures (testing irony, scalar, relevance, manner, contrastive implicatures, and novel metaphors). There were 12 critical and 32 filler items. Accuracy and reaction times (RTs) were recorded. For irony, for instance, children heard conversations ending with an ironic reply (e.g. *Yes, you know how much I like fruits* with an ironic intonation), and had to give the speaker one of three items (one compatible with an ironic, one with a literal interpretation, and one irrelevant). (2) The Word Definitions Test [5] and the Peabody Picture Vocabulary Test (PPVT) [6] for vocabulary. (3) The Family Affluence Scale (FAS) [7] and parental education levels for socioeconomic status (SES). We measured Vocabulary and SES because research suggests that they affect children's cognitive skills (see in [1]).

Percentage accuracy and mean RTs for accurate responses in critical items by implicature and group are presented in table 1. There was sufficient variation in all sub-tests (accuracies from 44% for irony and metaphors-82% for manner) besides relevance (93%). A Principal Component Analysis (PCA) on accuracies in each sub-test (excluding relevance because of ceiling performance) returned three components, with Scalars and Contrastive scores loading on the first factor, Manner and Metaphor scores loading on the second and Irony loading on a third. These results are largely in line with theory and developmental evidence on implicature

in that: (1) implicatures based on the maxim of quantity (scalar and contrastive implicatures) are related; (2) relevance implicatures are the easiest to understand (ceiling performance) with quantity implicatures following; (3) irony is one of the most difficult implicatures for children and is a distinct pragmatic phenomenon [2]. We also formed composite scores by averaging variables that were conceptually and statistically related to increase reliability: Vocabulary (from Word Definitions and PPVT), SES (from FAS, and parental education levels) and two Pragmatics composite scores (based on PCA results). Finally, analyses on background factors indicated differences in age ($F(2, 135)=3.625$, $p<.05$), SES ($F(2, 135)=80.56$, $p<.05$), and Vocabulary ($F(2, 135)=9.316$, $p<.05$), in that bi-dialectals tended to be older than bilinguals ($p=.08$) and monolinguals ($p=.06$); monolinguals had a higher Vocabulary than the other groups ($ps<.05$); and bilinguals had a higher SES than the others, while monolinguals had a higher SES than bi-dialectals ($ps<.05$). Age, SES, and Vocabulary were covaried in subsequent between-group analyses to control for these differences (see [1] that this is a valid use of ANCOVA).

A between-group analysis was conducted on Pragmatics (Pragmatics-1 vs Pragmatics-2, vs Irony) with age, Vocabulary, and SES covaried. Results indicated that neither the Group effect ($F(2, 124)=1.30$, $p>.05$) nor the Pragmatics x Group interaction ($F(4, 184.776)=.729$, $p>.05$) were significant. Similar results were obtained when Vocabulary was not covaried. Moreover, we obtained largely null results when performing similar analyses for each sub-test on RTs for correct responses in critical items and on difference scores calculated by subtracting RTs in fillers from RTs for correct responses in critical items (to control for baseline processing speed). Bi-dialectals, however, showed a trend for faster RTs ($ps=.07$) and smaller difference score ($ps=.08$) than bilinguals in Irony (with Vocabulary covaried or not) and significantly faster RTs than monolinguals in scalars (but only when Vocabulary was covaried) ($p<.05$).

Results show no consistent differences between bilingual, bi-dialectal, and monolingual children in implicature. This is true (1) despite bilinguals'/bi-dialectals' lower vocabulary, (2) for both implicature comprehension and processing, and (3) for late-developing implicatures, such as irony. These results suggest that (1) bilinguals'/bi-dialectals' lower language proficiency does not impede their implicature understanding and (2) that implicature comprehension possibly depends on other cognitive skills besides language proficiency. We discuss what these cognitive skills might be and, finally, suggest the possibility that a bilingual advantage is found only in the preschool years, when pragmatic skills are still at a very early stage of development.

Table 1: Percentage accuracy (A) and mean reaction times (RTs) by implicature and group.

	Relevance		Scalars		Contrastive		Manner		Metaphor		Irony	
	A	RTs	A	RTs	A	RTs	A	RTs	A	RTs	A	RTs
Monolinguals	90	2710	80	1949	72	2932	79	3138	50	5166	45	5006
Bilinguals	90	2439	75	1713	60	3483	82	3179	40	5291	42	8228
Bi-dialectals	95	2070	67	1651	57	2904	85	2228	40	4917	45	3457

[1] Antoniou, K. & Katsos, N. (2017). The effects of childhood bilectalism and multilingualism on implicature understanding. *Applied Psycholinguistics*, 1–47.
[2] Happé, F. G. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, *48*(2), 101-119.
[3] Kroll, J. F., & Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology*, *25*(5), 497-514.

[4] Yow, W. Q., & Markman, E. M. (2011). Bilingualism and children's use of paralinguistic cues to interpret emotion in speech. *Bilingualism: Language and Cognition, 14*(04), 562-569.
[5] Kort, W., Schittekatte, M., & Compaan, E. (2008). CELF-4-NL: clinical evaluation of language fundamentals. Pearson.
[6] Schlichting, L. (2005). Peabody picture vocabulary test III-NL. Amsterdam: Hartcourt Assessment.
[7] Currie, C. E., Elton, R. A., Todd, J., & Platt, S. (1997). Indicators of socioeconomic status for adolescents: the WHO Health Behaviour in School-aged Children Survey. *Health education research, 12*(3), 385-397.

Processing Antipresuppositions

Nadine Bade

Summary: The aim of this paper is to show that presuppositions and antipresuppositions (Percus 2006) are processed differently. Data from a reading time study comparing the German definite and indefinite determiner are presented. The results suggest that both the definite and indefinite are processed immediately. In addition, there are earlier effects of the indefinite only when it introduces a new referent but not when its antipresupposition is not satisfied, which suggests that different cognitive processes are involved when calculating presuppositions and antipresuppositions.

Theory: It has been observed that presupposition triggers have to be used if their presupposition (PSP) is fulfilled in the context. This is standardly assumed to be a result of the principle *Maximize Presupposition* (Heim 1991). According to theories working with *MaxPres* PSP triggers are ordered on a scale of a presuppositional strength with their non- presuppositional counterparts (Sauerland 2008, Percus 2006, Chemla 2008). One of these scales orders the definite and indefinite determiner. The indefinite yields the inference that the PSP of the definite is false ("antiuniqueness") due to this competition, which is why it is infelicitous in (1). The definite has to be used as a result of *MaxPres*.

(1) The / # A father of the victim came.

The inference arising from not using the presuppositionally stronger version has been called an antipresupposition (Percus 2006). It has been argued to have an epistemically weak status and project out of negation (Sauerland 2008). It is thus distinguished theoretically from PSPs and implicatures. There have so far only been few experimental investigation of antipresuppositions. Previous data suggest that they show late effects in processing (Kirsten et al. 2014) and that they do project under negation (Bade 2016). This is predicted under theories where antipresuppositions and presuppositions are considered to have a different status. Another view on the competition between the definite and indefinite is that they both come with their own context restrictions. For example, the oddness of (1) might have to do with the fact that the indefinite comes with a novelty condition whereas the definite comes with a familiarity condition (Heim 1982); or that the indefinite comes with an antiuniqueness presupposition as well (Kratzer 2004). Under these alternative views the definite and indefinite should behave paralelly in processing.

The Study: For the study 24 items were created in four conditions. A 2x2 design was used

crossing the conditions DEFINITE (with levels def/indef) and MATCH (with levels match/mismatch). The definite was considered matching when its PSP was verified in the context, it was mismatching when the PSP was not. For the indefinite the opposite was the case, if the indefinite has the antipresupposition that the PSP of the definite is false it should be mismatching when the definite was matching. Contexts were created were a referent was introduced with a simple description. The definite appeared with a relative clause matching the description of the context in the match condition and introducing a new description in the mismatching condition. Thereby the matching condition for the definite moreover fulfilled familiarity and the matching condition for the indefinite fulfilled novelty, see sample item below.

(2) A man entered the bar. {The/A} man who {entered the bar/ was sitting at the bar} smiled and ordered another beer.

Statistical analysis was done using linear mixed effect models and the lmer function in R. It revealed significant interactions between DEF and MATCH on words 5, 6 ,7, 10 and the final word (matrix clause) in the target (p<.05). The results suggest that, as has been observed before, unfulfilled presuppositions lead to immediate processing difficulties compared to fulfilled ones. For the indefinite the matching condition was significantly slower than the mismatching one early on, which shows that checking the contexts for whether the indefinite is appropriate also happens immediately. The fact that the mismatch condition was significantly faster than the match condition suggests that disobeying *MaxPres* is detected early on and the sentence is discarded as infelicitous. However, in the match condition a new discourse referent has to be introduced which leads to increased processing efforts. This confirms the speculation expressed in Kirsten et al. 2014 that introducing new discourse referents with the indefinite is costly. Unlike found in Kirsten et al., however, these processing costs are not more enduring compared to unfulfilled presuppositions. Rather, they are decreasing at the end of the sentence compared to the mismatching definite. This suggests that the accommodation process for the definite involves additional cognitive processes than just introducing a referent.

Conclusion: The data overall confirm theories where definites and indefinites are consid- ered non-equal in complexity and restrictions they impose. Just violating *MaxPres* seems to be a rapid process compared to violated presuppositions and novel indefinites.

Selected References: Chemla, E. (2009) "An epistemic step for antipresuppositions" In *Journal of Seman- tics*. Heim, I. (1991). "Artikel und Definitheit". In *Semantics: an international handbook of contempo- rary research*. Percus, O. "Antipresuppositions" In *Theoretical and Empirical Studies of Reference and Anaphora*. Sauerland, U. "Implicated presuppositions" In *Sentence and Context*..

When some triggers a scalar inference out of the blue
An electrophysiological study of a Stroop-like conflict elicited by single words

Cécile Barbet & Guillaume Thierry
School of Psychology, Bangor University

Several studies in experimental pragmatics have concluded that scalar inferences (hereafter SIs, e.g. 'some X are Y' implicates 'not all X are Y') are cognitively costly context-dependent pragmatic computations delayed relative to semantic computations (see e.g., Bott and Noveck 2004; De Neys and Schaeken 2007; Huang and Snedeker 2009; but see e.g., Grodner et al. 2010; Degen and Tanenhaus 2015; Politzer-Ahles and Gwilliams 2015).

However, it still remains unclear whether strong contextual support is necessary to trigger such inferences. Here we tested if the SI 'not all' triggered by some can be evoked in the absence of any linguistic context. We investigated event-related potential (ERP) amplitude modulations elicited by Stroop-like conflicts in participants (27 native speakers of English) instructed to indicate whether strings of letters were printed with all their letters in upper case or otherwise. In a randomized stream of nonwords and distractor words, the words *all*, *some* and *case* were presented either in capitals or featured at least one lower case letter.

As expected, we found a significant conflict-related N450 modulation (see e.g., West 2003; Szucs and Soltész 2010; Tillman and Wiens 2011) when comparing e.g. `aLl` with `ALL`. Surprisingly, and despite the fact that most responses from the same participants in an off-line sentence-picture verification task were "logical" (the participants largely accepted as good descriptions sentences such as 'Some circles are red' when all of the circles depicted were red), we also found a similar modulation when comparing `SOME` with e.g. `SoMe`, even though `SOME` could only elicit such a Stroop-like conflict when construed pragmatically. No such modulation was found for e.g. `CasE` vs. `CASE` (the neutral contrast), see Fig. 1.

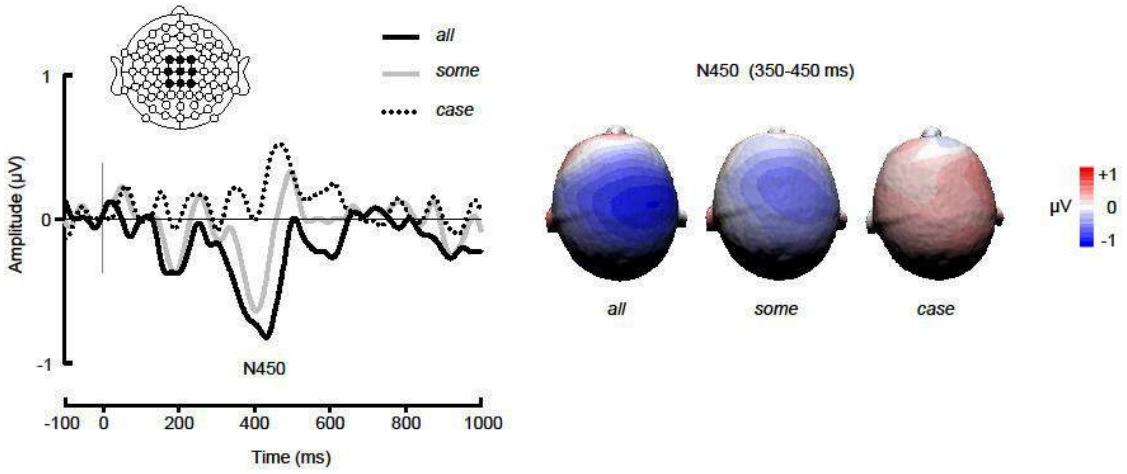


Figure 1. **Stroop-like conflict effect on ERPs corrected for physical differences**. *Left*, Grand-average difference (incongruent minus congruent) corrected ERP waveforms elicited over the central region (linear derivation of FC1, FCz, FC2, C1, Cz, C2, CP1, CPz, CP2) in the semantic test (*all*, solid black line), pragmatic test (*some*, solid grey line), and neutral control (*case*, dotted black line) conditions. *Right*, Topographies of the N450 effect for *all*, *some*, and *case*.

These results suggest that *some* can appear incongruent with the concept of 'all' in the absence of strong contextual support. Furthermore, there was no correlation between N450 effect magnitude (`SOME` minus e.g. `sOmE`) and pragmatic response rates recorded in the sentence-picture verification task. Interestingly, most of the participants of this study could be considered "logical" since almost 80% of the under-informative *some*-statements were considered good descriptions in the off-line task. Yet, the same participants exhibited a Stroop-like conflict when presented with the pragmatically

incongruent stimulus SOME in the ERP experiment. This seems to indicate that "logical" behaviour may stem from cognitive strategising rather than mere linguistic processing.

The N450 conflict effect observed for SOME is overall incompatible with a strong context-dependency view of the SI 'not all', given that in a situation of minimal linguistic context, SOME is not construed logically. This study shows for the first time that the pragmatic meaning of *some* can be accessed in the absence of linguistic support, and thus, that the SI 'not all' triggered by *some* should be construed as context-*sensitive* rather than context-*dependent*, that is, more or less salient depending on the context rather than contingent upon it.

References

Bott, Lewis and Ira Noveck (2004). 'Some utterances are underinformative: the onset and time course of scalar inferences'. In: *Journal of Memory and Language* 51, pp. 437–457.

De Neys, Wim and Walter Schaeken (2007). 'When People Are More Logical Under Cognitive Load. Dual Task Impact on Scalar Implicature'. In: *Experimental Psychology* 54(2), pp. 128–133.

Degen, Judith and Michael K Tanenhaus (2015). 'Processing Scalar Implicature: A Constraint-Based Approach'. In: *Cognitive science* 39.4, pp. 667–710.

Grodner, D. J. et al. (2010). '"Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment'. In: *Cognition* 116.1, pp. 42–55.

Huang, Yi Ting and Jesse Snedeker (2009). 'Online interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface'. In: *Cognitive Psychology* 58.3, pp. 376–415.

Politzer-Ahles, Stephen and Laura Gwilliams (2015). 'Involvement of prefrontal cortex in scalar implicatures: evidence from magnetoencephalography'. In: *Language, Cognition and Neuroscience* 30.7, pp. 853–866.

Szucs, Dénes and Fruzsina Soltész (2010). 'Stimulus and response conflict in the color–word Stroop task: a combined electro-myography and event-related potential study'. In: *Brain research* 1325, pp. 63–76.

Tillman, Carin M and Stefan Wiens (2011). 'Behavioral and ERP indices of response conflict in Stroop and flanker tasks'. In: *Psychophysiology* 48.10, pp. 1405–1411.

West, Robert (2003). 'Neural correlates of cognitive control and conflict detection in the Stroop and digit-location tasks'. In: *Neuropsychologia* 41.8, pp. 1122–1135.

Factors licensing embedded present tense in speech reports

Corien Bary, Daniel Altshuler, Kristen Syrett and Peter de Swart

Introduction According to Ogihara (1995), the truth of the complement in (1) at the actual utterance time n (i.e. when (1) is uttered) is not a prerequisite for the use of an embedded present tense. What matters is the cause of the belief (the state that made John think that Mary is in the room): the present tense can be used only if this cause still holds at n.

(1) John said that Mary is in the room.

Empirical evidence suggests that this hypothesis (KEY OBSERVATION) is only one of the factors involved in licensing a felicitous usage of the embedded present. We present two experiments which show that the cause of belief still holding is a sufficient, but not a necessary factor. We identify two additional factors (predicate type and who is aware of the falsity of the belief). These factors collectively suggest that the old idea of 'current relevance' (Costa 1972, McGilvray 1974) is right and cannot be translated into one single concrete factor (contra prevailing theories) but rather corresponds to a cluster of factors. Thus, this paper presents a case study of how the applications of experimental methods (see also Altshuler et al. 2015 for a related corpus study) may lead to different kinds of theories than introspection-based ones.

Background Ogihara (1995) considers various contexts for (1), a present under past sentence. He motivates the KEY OBSERVATION by comparing (2a) with (2b).

(2) John and Bill are looking into a room. Sue is in the room.
John (near-sighted): 'Look! Mary is in the room.'
Bill: 'What are you talking about? That's Sue, not Mary.'

a. John: 'I'm sure that's Mary.'
One minute later, Kent joins them. Sue is still in the room.
Bill (to Kent): '**John said that Mary is in the room**. But that's not true. The one that is in the room is Sue.'

b. John: 'I'm sure that's Mary.'
Sue leaves the room. One minute later, Kent joins them.
Bill (to Kent): # '**John said that Mary is in the room**.'

Klecha (2015) questions the KEY OBSERVATION with the counterexample in (3):

(3) Mary puts a balloon under her shirt. John then observes her in this state, and then says to everyone: 'Mary is pregnant!' Later that day, Mary takes the balloon out from under her shirt and pops it. Bill, aware of everything that happened, says to Mary: '(Earlier today,) John told everyone that you're pregnant.'

In this scenario, the cause of John's belief that Mary is pregnant, i.e. the balloon under her shirt, is absent by the time of Bill's report. Nevertheless, the present tense is acceptable. Why should this be? A direct comparison of (2) and (3) reveals a key set of factors that might play a role in the acceptability: the use of the verb of speech (*say* versus *tell*), whether or not the audience of the reported utterance still believes the complement, and the duration of the state in the complement (being in a room vs. being pregnant). To arrive at a better understanding, we conducted two experiments to investigate the effect of each of these factors.

Exp1: rating task The experiment followed a fully-crossed 2 x 2 x 2 x 3 design, with Latin square presentation of stimuli lists. There were two between-subject factors (MATRIX VERB

(*say* vs. *tell*) and TENSE OF EMBEDDED VERB (past vs. present)) and two within-subject factors (PREDICATE TYPE (individual-level vs. stage-level)) and WHO WAS AWARE of the fact that it was a false belief (A: the reporter alone; B: the reporter and the reported speaker; C: the reporter, the reported speaker, and the audience)). For each of the between-subject conditions there were 12 experimental items, divided equally among the within subject factors.

Each item began with a brief scenario introducing two key individuals (Ind-1, who becomes the Reporter, and Ind-2, who becomes the Reported Speaker) and some friends, the Audience. Ind-2 remarks aloud to Ind-1 that an Ind-3 has an I-level or S-level property P, an exclamation acknowledged by the entire group. The scenario then diverges based on who is aware of the falsity of the belief in 3 conditions: (A) only the Reporter (Ind-1), (B) both the Reporter and the Reported Speaker (Ind1+2), (C) the Reporter, Reported Speaker and the entire group. All items ended with the target sentence indicating that a few minutes later, another person (Ind-4), who is not part of the group, arrives. Ind-1 reports to Ind-4: "You won't believe this, but ≪Ind-2 [told us/said] that Ind-3 [was/is] P≫." Participants (n=88) were asked to rate the acceptability of the target sentence in ≪ ≫ on a 5-point scale.

Exp2: forced choice task To complement the acceptability ratings from Exp1, we conducted a forced-choice task, in which participants (n=41) explicitly had to choose between present and past tense for the embedded clause. The items had the same structure as in Exp1, but the matrix verb was always *tell* since the matrix verb was not a significant factor in Exp1.

Results The main results indicated:

(i) In both experiments: no interaction between item type (test vs. control) and tense ($p>.05$). In particular, present tense is not significantly better in the control items (with the cause of belief present) than in the test items (with the cause of belief absent). Thus, – contra the KEY OBSERVATION – the cause need not still be present for the use of a present tense;

(ii) In both experiments: the past is significantly preferred over the present with stage level predicates ($p<.01$). In Exp2 there was also a strong preference for the present with individual level predicates ($p<.001$). This is consistent with the contrast between (2b) and (3);

(iii) In Exp1: if everyone is aware that the complement is false (condition C) the past tense is significantly better than the present tense ($p=.04$). Thus, when the falsity of the belief is common knowledge, the present tense is less acceptable.

Discussion The factors above motivate the following three sufficient (but not necessary) conditions for a felicitous use of the present tense: (i) the cause of belief holding at the actual utterance time n (KEY OBSERVATION); (ii) if, had the complement been true at the time of the report, it would still hold at n (predicate type effects); (iii) the audience of the original utterance still having this belief at n (knowledge condition effects). Note that (iii) indicates that tracking other people's beliefs affects our choice of grammatical morphemes, even in the case of beliefs of people who are not participating in the actual conversation. We discuss the implications of these findings for the (use of the) notions of acquaintance relations (Abusch 1997, Ogihara 1995) or time concepts (Heim 1994) adhered to in the prevailing theories to explain the KEY OBSERVATION. We stress that (i)–(iii) form a natural class: they all indicate what must hold at n, suggesting that the long-standing intuition of 'current relevance' corresponds to a cluster of factors rather than one single concrete one.

Andrea Beltrama (University of Konstanz) andrea.beltrama@uni-konstanz.de

Subjective assertions are weak: an experimental study on perspective-dependent meaning.

The issue – Sentences containing subjective predicates – e.g., *beautiful* in (1) – intuitively feature a perspective-dependent flavor, contrary to sentences describing objective facts (as in (2)).

(1) **Subjective**: Paris is beautiful. (2) **Objective**: Paris is in France

While authors have long debated on whether this intuition tracks a lexical distinction between subjective and factual predicates, much remains to be explored on whether, and how, the difference between (1) and (2) is reflected at the illocutionary level. We show that assertions with subjective predicates (henceforth **SAs**) display a different discourse behavior from objective assertions (henceforth, **OAs**), unveiling a genuine empirical difference between subjective and factual speech.

Background – A wide open issue in the study of subjectivity revolves around whether assertions like (1) should be treated on a par with (2), that is, as a regular proposal to update the Common Ground with *p* or whether they merely presentational moves, which update the speaker's individual commitments but don't aim at increasing the CG (Dechaine at al. 2014). An intermediate position is that SAs do target the CG, but rely on a *weaker* norm of assertion than OAs, where *p* can be asserted as long as the speaker judges it to be true, but is only added to the CG if all participants in the conversation judge it as true (Stephenson 2007; see Coppock 2014 for a variant). We test these proposals experimentally, comparing the behavior of SAs and OAs with respect to two distinctive parameters of assertions.

Exp1: Silent Replies and CG Update – Adding *p* to the CG represents the unmarked outcome of an assertion (Stalnaker 1978 a.o.). As such, while rejection needs to be overtly signaled with a denial, silence typically leads accepting the proposal, on a par with an explicit "Yes" reply (Farkas&Bruce 2010). Exp1 compares SAs and OAs on this ground. If SAs work like regular assertions, silent responses should lead to updating the CG with *p*. This should not be observed, by contrast, if SAs are merely presentational, in which case no proposal is made at all; or if they rely on a weak norm of assertion, in which case an explicit response would be required from all participants before an update. 2 factors were crossed in a 3x3 design. Each trial consisted of a written dialogue in which Greg makes one of three possible moves – OA, SA or Polar Question (PQ) – and Mary provides one of three possible responses – Confirmation, Denial or Silence. Following each dialogue, participants were asked to assess on a 1-7 scale (7="totally agree") the statement "It is now part of Greg and Mary's mutual knowledge that *p*", which operationalized the idea that the CG has been updated with *p*. The higher the score, the higher the likelihood that the update went through.

Greg: {**OA**: "Paris is in France"/**SA**: "Paris is beautiful"/**PQ**: "Is Paris in France?"}
Mary: {**Conf.**: "Yes, indeed!"/**Den.**: "No, not really!"/ **Silence**: [Keeps listening, says nothing.]}
Statement to assess: "It is now part of G and M's mutual knowledge that {Paris is beautiful/is in France.}"

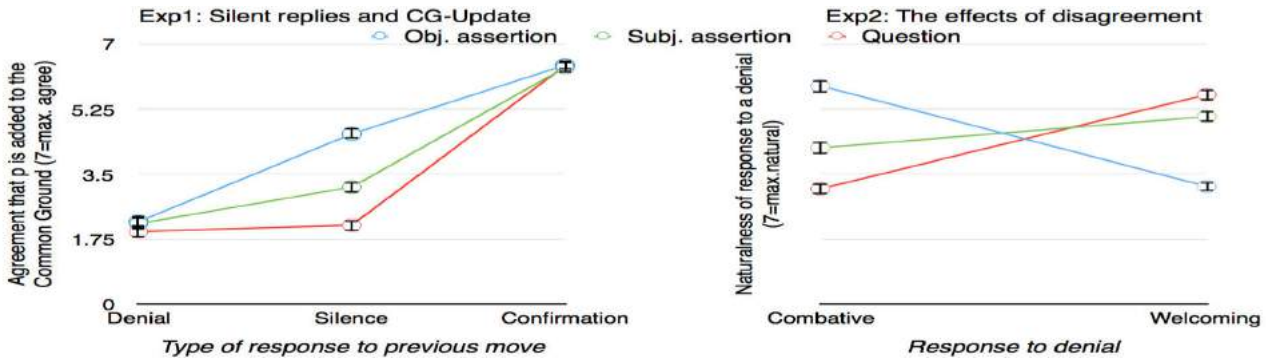
27 items, each with a different set of predicates, were distributed in 9 lists with a Latin Square Design. 54 native speakers of English were recruited on MTurk. The results are plotted on page 2. A mixed effects model with random intercepts for Subject/Item revealed main effects of Move and Response and an interaction Move:Response (*ps* $<.001$). Confirmation and denials led to respectively high and low CG-acceptance scores across moves. Silent responses lead to high and low scores respectively following OAs and PQs; following SAs, however, they record a higher score than PQs, but a much lower one than OAs (*ps*$<.001$). This indicates that silent replies were taken as a cue to update the CG with OAs, but not with SAs.

Exp2: The Effect of Disagreement – A converse property of assertions is that denials are highly

marked and lead the conversation into a state of *crisis* (F&B), which needs to be acted upon before the participants can move on. Exp2 compares OAs and SAs by looking at the naturalness of two types of reactions to a denial: "Aha, interesting to hear!", which signals a welcoming disposition towards disagreement; and "No way! That can't be true", which signals willingness react to the denial. 2 factors were crossed in a 3x2 design. Each trial consisted of a written dialogue in which Greg makes one of three moves (OA, SA or a PQ); Mary responds with a denial; and Greg follows up with one of the two reactions above. Subjects provided a 1-7 naturalness judgment (7=perfectly natural) on the final reaction. An example is below.

Greg: {**OA**: "John is 18."/**SA**: "J. is a great teacher!"/**PQ**: "Is J. 18?"}.
Mary: "No, he's not."
Greg: {**Welcoming**: "Aha, interesting to hear!"/**Combative**:"No way! That can't be true"}

If SAs do not differ from OAs, in both cases denials should engender a crisis, making a welcoming response odd. However, if SAs have no or weaker assertoric force, disagreement should be less disruptive, making it more natural for the interlocutors to welcome it. 18 items were distributed in 6 lists with a LSD (20 fillers). 54 subjects were recruited on MTurk. To ensure that welcoming and combative replies were perceived as such, subjects were explicitly instructed to assume that Greg was *not* being sarcastic. A mixed effects model with random intercepts for Subject/Item showed an interaction Move:Response ($p < .001$). As predicted, combative responses are rated higher than welcoming ones with OAs ($p < .0001$). Concerning SAs, welcoming replies are rated higher than combative ones ($p < .001$), similar to PQs; however, the two types of response are respectively rated considerably lower/higher than with PQs (all $ps < .001$).



Discussion – While OAs feature the canonical behavior of canonical assertions, SAs turn out to be different on two counts: **(i)** when followed by a silent response, they do not systematically lead to a CG update; (**ii**) they allow the listener to welcome the ensuing disagreement, rather than inducing a crisis. While this argues against the idea that SAs are just canonical assertions, note that SAs also behave differently from questions. In particular, the fact that combative responses to denials, though dispreferred to welcoming ones, are still more natural for SAs than for PQs provides evidence against the view that SAs have merely presentational force. Quite the contrary, these speech acts *do* make a positive proposal for increasing the CG, which can justify the speaker's effort to stand by the assertion after it has been rejected. Concerning the specific illocutionary profile of SAs, our findings suggest that these moves are *not* categorically biased towards the addition of p to the CG, contrary to what has been argued for regular assertions (see F&B); rather, their discourse profile, at the very least, must project disagreement as an equally unmarked outcome, explaining the failure of silent responses to default to CG Update, and the non-disruptive nature of denials.

Embedded implicature: What can be left unsaid?

Anton Benz & Nicole Gotzner, ZAS, Berlin.[1]

There has been a sharp debate about implicature of complex sentences, a variety of theoretical approaches have been developed [e.g. 1, 2, 3, 4, 5], and conflicting experimental evidence has been produced [e.g. 6, 7]. The relevant complex sentences are sentences in which an implicature trigger like '*some*' is embedded under a quantifier, which may itself be an implicature trigger. For example, the sentence (A-E) '*Each girl found some of her marbles*' potentially gives rise to the inference that each girl found some but not all of her marbles. In the course of this debate, a view took hold according to which sentence meaning is highly ambiguous, and different implicatures are just different readings that language speakers may entertain [in particular 1, 5]. In this talk, we are guided by the standard neo–Gricean view [8] that considers implicature a part of communicated meaning. Therefore, our main research question is: What can be reliably communicated by sentences containing embedded or unembedded '*some*'? In the following, we operationalise this research question and develop a new interactive experimental paradigm that tests both the production and interpretation of embedded '*some*'. We started out with the following basic idea: A speaker who wants to communicate a certain proposition can express all he wants to express literally, or he may take advantage of implicature, and leave certain aspects unsaid. This will lead to a shortening of utterances. Hence, our main research question can be reformulated as follows: To what extend can a description be shortened without jeopardizing communicative success? The shortest descriptions will then reveal all the implicatures that can be communicated reliably. To turn this idea into a testable theory, we formulated two cognitive principles that guide the elimination of linguistic material related to embedded '*some*': (ENA-Elim) the simplification of '*some but not all*' to '*some*', and (N-X-Elim) the elimination of '*none found X*'. For example, together they allow the simplification of literal '*Some found all, some some but not all, and none none*' (E–A : E–ENA : N–N) to '*some all and some some*' (E–A : E–E).[2] Our assumption was that utterance simplifications based on (ENA-Elim) and (N-X-Elim) communicate the intended message as reliably as the corresponding literal description, and all further simplification leads to unreliable communication.

With utterances composed of sentences of the form (X–Y) '*X of the girls found Y of the marbles*' with X and Y chosen from quantifier phrases '*none*', '*some*', '*any*', '*some but not all*', '*some and possibly all*', and '*all*', seven different worlds can be semantically distinguished depending on whether there are some who found none (E–N), some who found some but not all (E–ENA), or some who found all (E–A). As a next step towards a testable hypothesis, we defined a critical production strategy for the seven possible worlds, shown in (1) below, by application of the two elimination rules to a literal production strategy also shown in (1).

The main test hypotheses were: (I) The critical strategy is as successful at communicating the state of the world as the corresponding literal strategy; (II) any further reduction of utterance length makes the utterance significantly less reliable than the corresponding literal description. Further, the model predicts utterances of differential length for different possible worlds. In the following, we present an experimental study that tests the efficiency of this strategy for all seven worlds. Specifically, we tested whether the strategy is successful, and how it compares to strategies pursued by naive participants, in particular whether they produce shorter utterances, and if so, whether these utterances are still successful. The experiments indicate that the critical strategy is among the shortest strategies with almost maximal communicative success.

[1]This work was supported by the Bundesministerium für Bildung und Forschung (BMBF) (Grant Nr. 01UG1411), and the Deutsche Forschungsgemeinschaft (DFG) (Grant Nr. BE 4348/4-1).

[2]E–A : E–ENA : N–N –(N-X-Elim)→ E–A : E–ENA –(ENA-Elim)→ E–A : E–E.

Interactive Best Response Paradigm. Previous experiments on embedded implicature using picture verification tasks and acceptability judgements have obtained a substantial proportion of literally interpreting subjects [e.g. 6, 7]. This renders their experimental designs inappropriate for our task. Since our goal is to test the communicative success of utterances involving embedded *some*, we developed an interactive task involving both the production and interpretation of sentences in a collaborative scenario.

Methods: Participants in our experiment were presented with a scenario involving six girls who each own a set of four special edition marbles (based on 9). While the girls are playing the marbles get lost and they have to find them again. During the experiment, participants took two different roles. (1) The speaker had to describe a picture representing how many marbles each girl found. (2) The hearer received a message from the speaker and had to buy sweets to reward the girls. The speaker was allowed to produce up to five sentences by typing in one the following words into a sentence frame: *all, some, none, some but not all, some and possibly all* and *any* (in German). The speaker could produce a description consisting of a conjunction of up to five sentences of the form X–Y. Subsequently, the hearer received the sentences the speaker produced and had to choose the appropriate sweets as rewards. The reward system was defined such that a girls gets...

- chocolate if she finds all 4 of her marbles
- candy if she finds fewer than 4 of her marbles
- a gummy bear when she finds none of her 4 marbles (as a consolation prize).

Seven possible worlds were represented by seven items in total. The system randomly paired two participants for a given production-interpretation trial and each participant took a certain role three times. In total, 53 native German participants took part in the experiment. Participants took the experiment in groups of varying sizes: there were groups with 4 players, with 2 players, and groups with 3 players in addition to the experimenter, who played the critical strategy.

Results: We analyzed participants' success rate (expected utility) as a function of whether the hearer selected the appropriate sweets depending on the picture the speaker saw. Overall, the average participant had a high success rate of 89% (average length 2.09 compared to 1.71 (critical) and 2.5 (literal)), showing that participants understood the task. A t-test showed that the critical strategy was significantly better than the average participant strategy and it was also significantly shorter in terms of sentence length (p-values $<.001$). Interestingly, when participants produced exact/literal descriptions such as *Each girl found some but not all of her marbles* the communicative success was not better compared to utterances where the short form was used (1).

(1)

world	critical	% success	literal	% success
□	N–Any	97%	N–Any	97%
■	A–E	93%	A–ENA	92%
■	A–A	98%	A–A	98%
◨	E–E : E–N	95%	E–E : E–N : N–A	88%
◨	E–A : E–N	98%	E–A : E–N : N–ENA	93%
◨	E–A : E–E	93%	E–A : E–ENA : N–N	82%
◨	E–A : E–E : E–N	100%	E–A : E–ENA : E–N	93%

Reducing utterance length further can result in three utterances: E–E (39%■, 22%◨, 34%◨, 4%■), E–A (12%■, 69%◨, 19%■), and E–N (68%◨, 5%◨, 13%◨, 9%□, 5%■). For all of them the success rate was significantly lower than for the utterances of the critical strategy. The data, therefore, confirmed both main hypotheses: The critical strategy is as successful as the corresponding literal strategy, and shortening it further significantly reduces communicative success. The results, thereby, support the thesis that the two proposed elimination principles (ENA-Elim and N-X-Elim) characterise what can be left unsaid.

[1] Chierchia, G., Fox, D., and Spector, B. (2012) Scalar Implicature as a Grammatical Phenomenon. In Maienborn, C., von Heusinger, K., and Portner, P., (eds.), *Semantics: An International Handbook of Natural Language Meaning*, Vol. 3, pp. 2297–2331 De Gruyter Mouton Berlin.

[2] Sauerland, U. (2004) Scalar implicatures in complex sentences. *Linguistics and Philosophy*, **27**, 367–391.

[3] Franke, M. Signal to Act: Game Theory in Pragmatics PhD thesis Universiteit van Amsterdam (2009) ILLC Dissertation Series DS-2009-11.

[4] Benz, A. (2012) Implicatures of Complex Sentences in Error Models. In Schalley, A., (ed.), *Practical theories and empirical practice*, pp. 273–306 John Benjamins Amsterdam.

[5] Potts, C., Lassiter, D., Levy, R., and Frank, M. C. (2016) Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics,* **33**, 755–802.
[6] Geurts, B. and Pouscoulous, N. (July, 2009) Embedded Implicatures?!?. *Semantics and Pragmatics,* **2**(4), 1–34.
[7] Chemla, E. and Spector, B. (2011) Experimental evidence for embedded scalar implicatures. *Journal of Semantics,* **28**(3), 359–400.
[8] Levinson, S. C. (1983) Pragmatics, Cambridge University Press, Cambridge.
[9] Degen, J. and Goodman, N. D. (2014) Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In Bello, P., Guarini, M., McShane, M., and Scassellati, B., (eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pp. 397–402.

IMMEDIATE USE OF DISCOURSE CONTEXT IN ASPECTUAL COERCION – AN EYETRACKING DURING READING STUDY

OLIVER BOTT, UNIVERSITY OF TÜBINGEN & TU DORTMUND

According to the principle of compositionality the meaning of a complex expression is entirely determined by the meaning of its parts and their syntactic combination. However, linguistic expressions are also highly context dependent, and the interpretation system is therefore not only dependent on the parts of complex expressions in a bottom-up fashion, but also has to be open to top-down influences of the context of utterance. A number of studies on the online composition of meaning have used the phenomenon of coercion to address the time-course of compositional interpretation within the sentence. However, to date only a single study (Traxler et al. 2005) has explicitly investigated the effects of contextual information on the resolution of sentences involving complement coercion (e.g., *begin the book*).The present experiment studied the interplay between sentential and contextual information during the online resolution of yet another coercion type, viz. aspectual coercion (see Piñango et al 1999, Paczynski et al. 2015, a.o.). In particular, we tested whether contextual information is immediately used to resolve compositional conflicts during online interpretation. Consider (1).

(1) Als es ihm heute gelang, in einer Stunde zu joggen, freute er sich sehr.
When he managed today in one hour to jog, he was very happy.

When uttered out of the blue, sentence (1) is hardly interpretable. The *in-adverbial* requires a telic event predicate of the accomplishment type (Vendler 1957, Dowty 1979), but *Peter jogged* expresses an atelic activity. However, if the sentence makes reference to a spatially bounded path such as *five miles* it becomes perfectly interpretable (cf. Krifka 1992). German sentences of type (1) were embedded in a discourse context which introduced a bounded path in the preceding discourse such as (2) translated from German.

(2) Half a year ago Peter started to jog a distance of eight kilometers every day. When he started he was quite slow but he has become faster and faster.

Based on the pragmatic literature (e.g. Recanati 2010 vs. Cappelen & Lepore 2005) two alternative hypotheses were formulated. The *Composition in Context Hypothesis* predicts immediate availability of the bounded path from the context and hence no temporary mismatch effect when composing the verb of the target sentence with the adverbial. Alternatively, the initial compositional interpretation of the target sentence could operate strictly locally encapsulated from contextual information. The latter hypothesis predicts an initial semantic mismatch at the underlined critical region *to jog*, followed by contextually driven repair by enriching the activity into an accomplishment. This should be reflected by an immediate slow-down relative to an aspectual control condition; cf. (3).

(3) Half a year ago Peter started to jog every day. When he started he could barely jog for ten minutes but he is becoming better and better.
Als es ihm heute gelang, eine ganze Stunde zu joggen, ...
When he managed today for one hour to jog, ...

In order to compare the timing of aspectual enrichment with the breakdown of aspectual interpretation, a mismatch condition was included. The atelic contexts of the control condition were combined with the *in*-modified target sentences.

While the two hypotheses clearly differ with respect to their predictions concerning the time course of interpretation, both hypotheses predict that the coercion condition should eventually be repaired making use of contextual information. This was confirmed in two offline acceptability rating experiments.

Pretests: 24 items were constructed in the aspectual enrichment, control and mismatch conditions. The first pretest (N=30) elicited discourse sensicality judgments on a seven-point scale. The statistical analysis of the ratings revealed that the aspectual enrichment condition was fully acceptable. Mismatch, by contrast, was perceived as nonsensical at the same level as clearly nonsensical fillers. The interpretation data thus confirm that the contextual support in the aspectual enrichment condition made the target sentences fully acceptable and that the telic target sentences do not fit an atelic context.

The second pretest (N=20) confirmed that without supporting context the telic target sentences were not fully well formed but required further contextual support. Decontextualized target sentences with *for-adverbials* were rated much better than sentences with *in-adverbials*. However, the target sentences with *in-adverbials* were still rated better than clearly nonsensical fillers suggesting that participants were well aware of the fact that the sentences with *in-adverbials* might turn out to be well-formed given appropriate contextual support.

Eyetracking Experiment: 48 participants read the pretested discourses together with 66 fillers while their eye gaze was monitored. After each trial they provided a sensicality judgment.

The analysis of acceptance rates did not reveal any reliable differences between coercion and control. Aspectual mismatch, by contrast, was rejected as uninterpretable equally often as were nonsensible fillers. Thus, participants perceived a clear aspectual mismatch in the mismatch condition and computed aspectually enriched interpretations of the target sentences in the coercion condition.

The analyses of first fixation durations, first-pass times and proportions of regressions out of the critical verb region consistently revealed processing costs of the aspectual mismatch condition relative to the control and the coercion condition. Crucially, none of the analyzed eyetracking measures related to first-pass reading indicated any reliable differences between the coercion and the control condition. This is fully consistent with the *Composition in Context Hypothesis*. However, coercion led to significantly longer second-pass times of the critical verb region than the control condition suggesting that the integration of contextual information from the preceding context is in fact not cost free but requires building up a more complex discourse model than in the control condition.

References

- Cappelen, H. & Lepore, E. (2001). *Insensitive semantics*. Oxford: Blackwell.
- Dowty, D. (1979). *Word meaning and Montague Grammar*. Dordrecht: Reidel
- Krifka, M. (1992). Thematic relations as links between nominal reference and temporal constitution. In Sag, I. & Szabolcsi, A. (eds.): *Lexical matters*. Stanford: SUP. 29–53.
- Paczynski, M. et al. (2015). When events change their nature. *JCN*, 26(9), 1905–1917.
- Piñango, M.M. et al. (1999). Real-time processing implications of aspectual coercion at the syntax-semantics interface. *JPR*, 28(4), 395–414.
- Recanati, F. (2010). *Truth-conditional pragmatics*. Oxford: OUP.
- Traxler, M. et al. (2005). Context effects in coercion. *JML*, 53, 1–25.
- Vendler, Z. (1957). Verbs and times. *Philosophical Review*, LXVI, 143–160.

Rates of scalar inferences beyond 'some' – A corpus study

Richard Breheny (University College London), Chao Sun (University College London), Ye Tian (Universite Paris Diderot)
chao.sun.13@ucl.ac.uk

In a large-scale corpus-based web study, Degen (2015) extracted 1363 utterances containing *some*-NPs from the Switchboard corpus. For each utterance, they measured the rate of the scalar inference (SI) from *some* to *some but not all* using a paraphrase task. Their findings showed that around half the time *some* is used, an SI reading is not judged to be available. Little is known about how frequently scalar expressions of different lexical categories give rise to SIs in *real use*. In an inference task, van Tiel et al. (2016) showed that different scalar terms give rise to SIs at different rates. Here, we adopt Degen's paraphrase task using a Twitter corpus we constructed to investigate whether the rates of SI derivation vary to the extent found in van Tiel et al.'s inference task. We do find variability in the rates of SIs across different scalar expressions, but not the same degree of variability found when items are presented out of context in an inference task. A modest amount of this variance could be explained by factors which van Tiel et al. found contributed to the variances in the experimental setting. Our study yields several interesting results, mentioned below.

Collecting a Twitter corpus: We selected 28 out of 43 scalar expressions found in van Tiel et al. (2016). There were 2 quantifiers (e.g. <some, all>), 1 adverb (<sometimes, always>) and 25 adjectives (e.g. <intelligent, brilliant>). For each scale, we extracted tweets containing the weak scalar term - with a minimal length of 30 characters. Then we conducted part-of-speech (POS) tagging on each tweet and used regular expressions to filter out tweets where scalar expressions appear in environments which the inferences are unavailable or less likely to arise (see Table 1).

environment	example
in the scope of negation	I'm not really hungry.
in the scope of conditional antecedents	If the weather was warm, we would have some people over for a small party in our backyard.
in the scope of wh-questions or polar questions	Do you get adequate vitamin D?

Table 1: Environments prohibit the scalar inference

To perform the final exclusion, we conducted a word sense disambiguation task on Amazon Mechanical Turk to obtain human annotation on tweets containing polysemous scalar expressions. Considering <old, ancient> for example, in (1a) the sense of *old* meaning "existing a long time" is on the same scale as the core meaning of *ancient*. However, in (1b) the sense of *old* meaning "previous" was not on the same scale as the strong term. Cases like (b) need to be excluded because in these cases the strong term is not contextually available which make it infelicitous to investigate the rate of SIs. We consulted the Merriam-Webster dictionary and found 20 out of 28 our scalar expressions have at least two meanings.

(1a) I'm in an ***old*** abandoned train station w/ a translator working on the script.

(1b) That means my ***old*** boss has been approaching a breakdown for the last 2 years.

80 M-Turk workers were recruited and each annotated 50 tweets of a particular scalar expression. In total, 4000 tweets were annotated, 200 tweets per scale. We presented workers with a tweet containing the scalar expression, e.g. *warm* (I guess he wants his home to feel ***warm*** and inviting.) and ask them to choose the meaning of *warm* from the following three sense labels: (if none are appropriate, workers can click 'none of the above' option) (a)

having a fairly high temperature; (b) friendly and affectionate; (c) light and bright colors. (a) is the sense that could be understood on the same dimension as the strong term, whereas (b-c) are the relatively common senses listed in the dictionary. Based on our results, we excluded tweets in which weak terms evoke senses that are not on the same scale as strong terms.

Corpus-based paraphrase task: We ran a paraphrase task based on Degen (2015) to measure the frequencies of SIs triggered by the 28 scalar expressions. After the final exclusion, we ended up with 3075 tweets in total. We randomly selected 50 tweets for each scale as the target sentences. On each trial, participants read an utterance containing a scalar expression *X* (the weak term, in red) and a nearly identical utterance, expect that the negation of the stronger term *not Y* (in green) was inserted (Figure 1). Participants were asked to rate on a seven point scale to indicate how similar is the statement with *X but not Y* to the statement with *X*. 550 participants each judged 28 items – one item per scale.

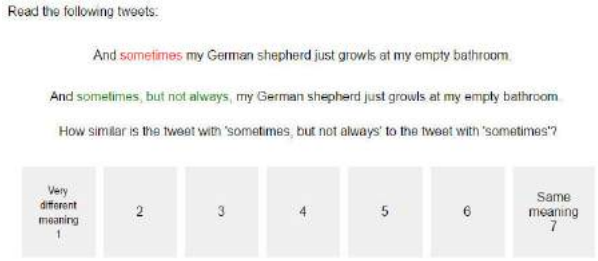


Figure 1 paraphrase task example item

Results: The responses were coded into three categories: low (ratings were 3 or lower), median (ratings were 4), and high (ratings were 5 or above). We considered high ratings as an indicator of SIs being drawn. Inspecting Figure 2, the frequency of SI varies across scalar expressions, from 27% for <adequate, good> to 86% for <sometimes, always>. These results correlated with the results of van Tiel et al. (2016) ($r=0.81$, $p<.001$), suggesting that, to some extent, the results yield from the inference task based on artificial examples could reflect frequencies of SIs triggered in *real use*. However, Levene's test for equality of variances showed that variances of two studies are not equal ($F(1,54)=14.69$, $p<.001$). Visual inspection of Figure 2 suggests that there is less variation on the paraphrase task. In particular, adjective scalar expressions give rise to SIs more frequently in real use. We replicate the result in Degen (2015) for 'some' and note that actual rates of SIs for this item and other terms like, 'possible' and 'allowed' are far lower than rates found on the inference task.

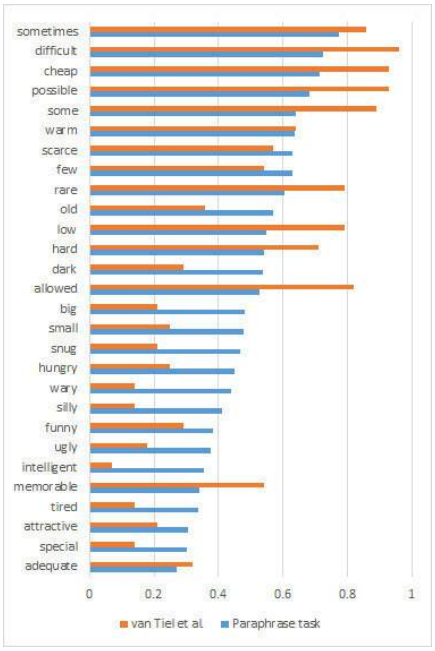


Figure 2 shows the percentage of 'High' ratings for 28 scalar expressions. Percentage of SI responses from van Tiel et al. (2016, Experiment 2) are shown in orange.

The variability displayed in the frequencies of SIs call for an explanation. Multiple linear regression analyses were conducted to predict the frequencies of SIs from possible factors explored in van Tiel et al. (2016), including association strength, grammatical class, word frequencies, semantic relatedness, semantic distance, and boundedness. As van Tiel et al., found with their inference task results, only semantic distance and boundedness are substantial factors. In this case, these factors together accounted for 43% of the variance. Future studies need to explain where the remaining variance comes from.

Reference: [1] Degen, Judith. 2015. *Semantics and Pragmatics* 8(11). 1–55. [2] van Tiel, B. van Miltenburg, E. Zevakhina N. & Geurts, B. (2016), *Journal of Semantics*, 33: 137-175.

Evidence for global pronoun resolution
Saskia Brockmann & Nadine Bade

Summary: By presenting data from a study conducted on cataphoric pronouns in German we show that the semantics of different temporal connectives determines whether people are looking for a referent immediately or delay the pronoun resolution.

Theory: In previous experimental work, it has been shown that the process of looking for a referent happens immediately upon encountering a pronoun (Chow 2014). Moreover, under a standard semantic view a sentence is considered inappropriate if it contains pronouns without a referent (Heim & Kratzer 1998). We will adopt a dynamic semantic framework where the meaning of a sentence is a function from contexts to contexts. Contexts are modelled as world-assignment pairs. We model a context update where a proposition p is added to the current context with the help of the operator "Assert". The operator introduces definedness conditions which make sure that all free occurrences of indices on pronouns are part of the assignment function g and all presuppositions of p are satisfied in the context (see the definition in (1)).

(1) $[\![\text{Assert}_c]\!] = \lambda p \in D_{\langle gst,gst\rangle}$: $\forall w,g[c(w)(g) \rightarrow i \in dom(g)\ \&p(w)(g) = 1 \text{ or } p(w)(g) = 0]$. p(c)

In the case of "and" we assume a stepwise update to prevent asymmetries as in (2).

(2) a. #He also cooked dinner and Peter cleaned the kitchen.
b. Peter cleaned the kitchen and he also cooked dinner.

The lexical entry of "and" in (3) captures that the first conjunct is updated before the second. The ASSERT operator must thus be in the first conjunct, see (4).

(3) a. $[\![\text{and}]\!] = \lambda q \in D_{\langle gst,gst\rangle}.\ \lambda p \in D_{\langle g,st\rangle}.\ \lambda c.\ p(q)$
b. $[\![[\ [\ \text{ASSERT q }]\ [\text{and p }]]\!] = [\![\text{and}]\!]([\![p]\!])([\![\text{ASSERT}]\!]([\![q]\!]))$

For sentences with subordinate temporal clauses headed by "before" and "after" the same asymmetry does not arise (see (4)).

(4) a. Before he also cooked dinner, Peter cleaned the kitchen.
b. Peter cleaned the kitchen before he also cooked dinner.

Following an adapted version of "before" and "after" in Penka (2008), "after" (and parallel "before") relates a temporal phrase and a point in time at which another temporal phrase took place (see (5) and the analysis of (6-a.) in (6-b.)).

(5) a. $[\![\text{after}]\!] = \lambda t.\ \lambda t'.\ \lambda P \in D_{\langle i,t\rangle}.\ P(t)\ \&\ t' > t$.
b. Mary arrived after John left.
c. ($\exists t''' < t_{now}$) Mary arrives at t'''& t'''>the earliest t such that ($\exists t'' < t_{now}$) t'' = t & John left at t)

In this analysis, the tense of the matrix clause is dependent on the interpretation of the temporal phrase. Consequently, "Assert" can only scope above the overall clause. On the basis of this analysis, our hypothesis for the study is that participants delay the update process when they hear "before" or "after". To test this, we used a 2x2 design crossing the conditions CLAUSE TYPE and GENDER MATCH. We created items where a cataphoric pronoun appeared in the matrix or temporal clause and either ambiguously or unambiguously referred to a given referent. The ambiguity was created by a match of the gender of the previously introduced referent and the pronoun. These cases were counterbalanced with mismatching pronoun and referent pairs. See one sample item in the four conditions below:

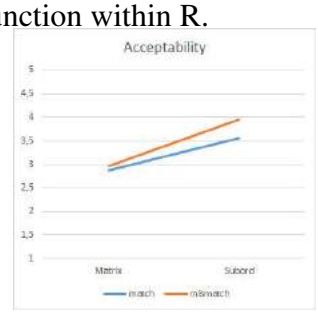
(6) Eine Unternehmerin kommt ins Büro.
An entrepeneur.FEM enters the office.
a. Er bringt einen Kaffee, bevor der Sekretär die Unternehmerin spricht.
He brings a coffee, before the secretary.MASK the entrepeneur.FEM talks-to.
(MATRIX CLAUSE, MISMATCH)

b. Bevor er einen Kaffe bringt, spricht der Sekretär die Unternehmerin.
Before he a coffee brings, talks-to the secretary.MASK the entrepeneur.FEM.
(SUBORD CLAUSE, MISMATCH)

c. Sie bringt einen Kaffee, bevor die Sekretärin die Unternehmerin spricht.
She brings a coffee, before the secretary.FEM the entrepeneur.FEM talks-to.
(MATRIX CLAUSE, MATCH)

d. Bevor sie einen Kaffe bringt, spricht die Sekretärin die Unternehmerin.
Before she a coffee brings, talks the secretary.FEM the entrepeneur.FEM.
(SUBORD CLAUSE, MATCH)

The Study: 24 German native speakers participated in the experiment. They were asked to judge the acceptability of the sentences within the given context on a scale from 1 - 5 (5 meaning fully acceptable). The analysis was done using linear mixed effect models using the lmer function within R.



Gender	Matrix	Subordinate
match	2,88	3,55
mismatch	2,96	3,96

There was a significant interaction between clause type and gender match ($p<.05$). There were moreover significant main effects of both gender match ($p<.03$) and clause type ($p<.001$): Mismatching pronouns were overall judged better since they disambiguate the reference. In addition, subordinate clauses containing the pronoun were judged better than main clauses. The interaction by a simple effect showing that subordinate clauses containing a mismatching pronoun were judged significantly better than matrix clauses with a mismatching pronoun ($p<.001$). This suggests that in matrix clauses people immediately look for a referent. In the mismatch condition, the unavailability of a referent was clear from the start and even when the sentence later on provided one this was not found appropriate. In subordinate clauses, however, the mismatching pronoun was found more acceptable, suggesting that participants knew that more information was to come and waited for the possibility of resolution rather than rejecting the sentence at this point.

Conclusion: Our results suggest that, indeed, participants are aware of the fact that a context update is delayed in the case of subordinate clauses. However, when encountering a matrix clause, people expect to update the context and want to assign a referent right away.

References

Chow W.Y. & Lewis, S. & C. Phillips (2014). "Immediate sensitivity to structural constraints in pronoun resolution." In: *Frontiers in Psychology*.

Penka, D. & A. von Stechow (2008). "Phrasal complements of *before* and *after*." In: *Empirical Issues in Syntax and Semantics 7*. Ed. by O. Bonami & P. Cabredo Hofherr, pp. 1–17.

A Bayesian Game-Theoretic Approach to Cross-Linguistic Variation

Heather Burnett & Barbara Hemforth
(LLF, CNRS-Université Paris Diderot)

1. Introduction. An important current research question in psycholinguistics concerns the mechanisms through which different interpretations of superficially similar constructions can arise across languages (see Frazier & Clifton 1995, Grillo & Costa 2015 for cross-linguistic variation in the interpretation of relative clauses; Carminati 2002, Hemborth et al. 2010, de la Fuente et al. 2016 for variation in pronoun resolution). Many of these researchers suggest that Gricean-type principles may be underlying the observed differences; however, they do not provide formal accounts of the relationship between interactive reasoning, syntactic variation and pragmatic interpretation. This paper argues that Bayesian signaling game models, particularly *Iterated Best Response (IBR)* or *Rational Speech Act (RSA)* models (Franke 2009, Frank & Goodman 2012 et seq.), can significantly contribute to making these proposals maximally explicit. Although these models have been shown to be useful in the analysis of many pragmatic phenomena in English, the potential of this framework for analyzing cross-linguistic pragmatic variation has yet to be explored. In IBR/RSA models, speakers' actions (and consequently listeners' interpretations) are optimized according to the inventory of syntactic forms available in the language. Therefore, this architecture is ideal for capturing the link between syntactic variation and variation in interpretation.

We build a simple RSA model of differences in pronominal resolution preferences between German, English and French that takes into account differences in the inventory of syntactic constructions between them, and we show how our model straightforwardly predicts the patterns of pronominal reference observed in psycholinguistic experiments. We therefore conclude that game-theoretic models constitute valuable tools for investigating the link between the syntactic properties of a language and the pragmatic reasoning processes of its speakers.

2. Cross-linguistic data. Consider an utterance with two possible referents (subject and object) (1). Using both visual world and questionnaire methodology, Author et al. 2010 and Baumann et al. 2014 have show that, when such an utterance is followed by a separate sentence containing a pronoun (2), German, French and English listeners highly prefer to interpret the pronoun as referring to the subject rather than the object.

(1) a. **Der Briefträger** hat **den Straßenfeger** getroffen... **German**
b. **The postman** met **the street sweeper**... **English**
c. **Le facteur** a rencontré **le balayeur**... **French**

(2) a. . Dann ging **er** nach Hause. **German**
b. . Then **he** went home. **English**
c. . Puis **il** est rentré à la maison. **French**

However, these authors also observe cross-linguistic variation when the following utterance is an adjunct on the main clause (3). In German and English, listeners still prefer to resolve the pronoun to the subject (see also Bouma & Hopp 2007, Kehler & Rohde 2016). However, in French, they are most likely to interpret the pronoun as referring to the object (see also Colonna et al. 2012).

(3) a. ... bevor **er** nach Hause ging. **German**
b. ... before **he** went home. **English**
c. ... avant qu'**il** rentre à la maison. **French**

For example, in Hemforth et al. 2010's questionnaire, the percentage of subject interpre-

tations across and within sentences is shown in the table below:

Language	Between sentence	Within sentence
German	95%	85%
English	95%	70%
French	95%	**20%**

3. An RSA model for German/English/French. RSA models formalize aspects of Gricean reasoning in terms of Bayesian signaling games. In a signaling game, there are two players: speaker (S) and listener (L). S has a fact about the world that they want to communicate to L. To model the pronoun resolution data, we will assume that the space of propositions under consideration consists of *The individual denoted by the subject of (1) went home* (p_S) and *The individual denoted by the object of (1) went home.* (p_O). S has a set of interpreted syntactic forms that they can choose from to send to L. As Hemforth et al. 2010 observe, German, English and French crucially differ both on the inventory of forms in the language and on the patterns of use of syntactic variants. Unlike German, both English and French possess variants of (3) containing a null PRO (4), which is obligatorily interpreted as referring *de se* to the subject (Chierchia 1989).

(4) ... **before going** home. **(Eng.)** ... **avant de** rentrer à la maison. **(Fren.)**

Furthermore, English and French differ in the relative frequency of the PRO form: in corpus studies of English, the overt pronoun form was found to be 4.32 times more frequent than the PRO form; whereas, the PRO form was found to be 1.58 times more frequent in French studies (Baumann et al. 2014). As is common in RSA, we encode such grammatical (dis)preferences through assigning a higher **cost** to the dispreferred syntactic structure than to the preferred one. Thus, we propose that the (relevant) inventories of syntactic forms across the three languages are as follows:

	German		English		French	
Form (m)	$[\![m]\!]$	Cost(m)	$[\![m]\!]$	Cost(m)	$[\![m]\!]$	Cost(m)
Overt (*er/he/il*)	$\{p_S, p_O\}$	0	$\{p_S, p_O\}$	0	$\{p_S, p_O\}$	1.5
PRO			$\{p_S\}$	1	$\{p_S\}$	0

Following Arnold 2001, we assume that hearing a DP in subject position increases L's expectation that this DP will serve as a referent in the subsequent discourse, which explains the cross-linguistic subject preference between sentences. We therefore take L's beliefs after hearing (1), but prior to hearing (3), to be represented by the prior probability distribution $Pr(p_S) = 0.95; Pr(p_O) = 0.05$. We then apply the RSA iterated solution concept to this architecture (with soft-max temperature parameter = 1), and generate the predicted probabilities of subject interpretations, which mirror the experimental results.

Language	Between sentence	Within sentence
German	95%	95%
English	95%	72%
French	95%	15%

Thus, using these models, we show explicitly how cross-linguistic variation in pronoun resolution can be reduced to variation in the syntax of different languages.

Social abilities help us detecting jokes: An EEG study on the temporal dynamics of humor comprehension

Paolo Canal[1], Luca Bischetti[2], Simona Di Paola[1,3] & Valentina Bambini[2]
[1]Laboratorio di Linguistica G. Nencioni, Scuola Normale Superiore, Pisa
[2]NETS, IUSS Institute for Advanced Study, Pavia
[3]DISFOR – Psychology Unit, University of Genoa

Introduction: Humor refers to anything that tends to make others laugh and is a universal aspect of human experience and communication [1]. Traditional theories of humor processing [2] posit that humor comprehension is a two-stage process in which the perception of an incongruity in a playful context is followed by its resolution. When the resolution is successful, one "gets" the joke and the typical feeling of mirth. Recent research in the Relevance Theory framework underlines how resolution is achieved through different types of inferential – pragmatic – processes filling the gap between what is coded and what is eventually interpreted [3]. As a useful technique to investigate the temporal sequence of cognitive mechanisms, Event-Related brain Potentials (ERP) generally found support for two-stage accounts of humor comprehension. However, the results are far from being consistent. Late positive effects (P600/LPC) have been discussed in many studies [4,5,6,7,8,9,10], sometimes linked to inferential processes [7,8,10]. These positivities were often accompanied by negative effects, interpreted as N400 effects [4,7,8,9,10], even though their scalp distribution was not canonical [7,8,9]. In addition, several studies reported sustained negative effects over frontal left electrodes, suggesting the involvement of the Left Anterior Negativity (LAN) [4,5,6]: because of these temporal and topographic differences, no agreement exists on the functional meaning of these effects. These discrepancies may be due to the inter-individual variability in the ERP response to humor. Researchers often dig into such variability by splitting participants into groups based on performance [4,9], sex [5] or verbal abilities [5]. However, these studies might have failed to characterize differences that could be due to more general cognitive or socio-cognitive abilities, which likely play a role in incongruity detection and resolution. Here we investigated the effect of verbal working memory and social skills on the ERP, using a relatively large sample of participants. We selected these abilities as predictors of the ERP response based on behavioral and neuroimaging studies [1,12].

Method: 70 jokes taken from Italian repertoires were included as materials. Each joke consisted of a three-sentence context (presented sentence by sentence) followed by a final punchline (presented word by word), including a humorous trigger word in a non-final position. For each joke, a non-humorous, straightforward-ending counterpart was created by replacing the humorous trigger word with a different word matched for frequency, length, and grammatical class, as in the following example: *A man goes to the grocery store to buy apples. / The grocer asks: "Would you like the red ones or the green ones?"/ And the man says: / "It doesn't really matter, 'cause I peel them anyway"* (humor) - *"It doesn't really matter, 'cause I pay them the same price"* (non-humor). Materials were rated for funniness on 7-point Likert scale [humor=3.97; non-humor=2.07] and cloze probability [humor=30.63%; non-humor=9.64%]. 52 right-handed participants (31F; 24y on average) took part in the study. We collected measures of verbal working memory through a sentence-span task and of social skills through the Autism-Spectrum Quotient (AQ). Single trial ERPs were analyzed with linear mixed models, focusing on 3 regions of interest (Frontal Left, Centro-Parietal and Parietal) and 3 time-windows (early – 300-500ms; middle – 500-700ms; late – 700-1000ms).

Results: Reliable effects of Humor emerged during the LAN [Frontal Left; early; -0.75μV, $t=-2.40$, $p<0.05$], sustained LAN [Frontal Left; middle; -1.33μV, $t=-4.05$, $p<0.001$], P600 [Parietal; middle; +0.75μV, $t=2.18$, $p<0.05$] and LPC [Parietal; late; +1.46μV, $t=4.27$,

$p<0.001$], but not during the N400 [Centro-Parietal; early; +0.16μV, $t<1$] (See Figure 1). AQ predicted the ERP response during the LAN time window [$\Delta\beta=-0.53$, $t=2.52$, $p<0.05$], revealing that the size of the negativity increased as AQ scores increased. No other effects were observed.

Discussion: Results further support the two-stage model of humor comprehension. Assuming that the detection of an incongruity precedes its resolution, the effect on the early time window suggests that it can take place as early as after 300ms, consistently with all previous investigations. The scalp distribution of the effect also suggests that the component involved in this processing stage is a LAN rather than an N400, in contrast with [4,7,8,9,10]. Indeed, the role of the N400 component was already questioned by studies finding no reliable effects [11] or frontally distributed negativities during the N400 time-window [7,8,9]. We thus argue that incongruity detection is associated with left-anterior negativities, whose functional characterization as sensitive to unexpected information might extend beyond the morpho-syntactic domain. Those studies reporting genuine N400 effects may have used kinds of jokes that not only presented an incongruity but were also strongly unpredictable. The later stage of resolution was mirrored by an enhanced positivity affecting the ERPs for a long time interval ranging from 500 to 1100ms, in line with previous research on humor and, more generally, with the view of the P600/LPC as reflecting interpretative efforts [13]. As for individual differences, verbal working memory does not seem to play a role in humor processing in the young healthy population, differing from evidence reported in patients and elderly people. Yet, the effect of AQ as predictor of the size of the LAN suggests that participants with less developed socio-cognitive skills pay more effort in the detection of the incongruence between setup and punchline. This finding matches with previous evidence on the interplay of AQ and irony – another inference-based process that might be associated with humorous effects [3] – showing that socially disinclined participants are less able to discern ironic utterances [14]. We speculate that socio-cognitive skills increase the ability to understand the playful context in which humor incongruity occurs. In sum, this study shed new light on the temporal dynamics of humor processing, linking its sequential stages to a LAN followed by a long-lasting positivity, and on the speaker-based variation of the early phase of such pattern, depending on the individual's socio-cognitive skills.

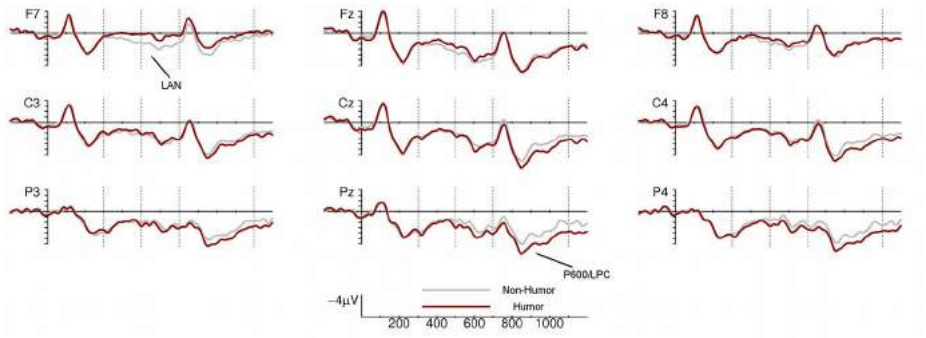


Figure 1. ERPs from nine representative electrodes: Grand Averages of Humor (dark red) and Non-Humor (grey) conditions.

References: [1] Vrticka, Black & Reiss. *Nat Rev Neurosci*, 14 (12), 860–868 (2013). [2] Suls. In *The psychology of humor: Theoretical perspectives and empirical issues,* Goldstein, Ed (Academic Press, 1972). pp. 81–100. [3] Yus. *Humor and Relevance*. (John Benjamins, 2016). [4] Coulson & Kutas. *Neurosci Lett*, 316(2), 71–74 (2001). [5] Coulson & Lovett. *Cognitive Brain Res,* 19, 275-288 (2004). [6] Coulson & Williams. *Neuropsychologia,* 43(1), 128–141 (2005). [7] Feng, Chan & Chen. *J Neurolinguist,* 32, 59-70 (2014). [8] Du, Qin, Tu, Yin, Wang, Yu & Qiu. *Int J Psychol,* 48(2), 149-157 (2013). [9] Ku, Feng, Chan, Wu & Chen. *J Neurolinguist,* 42, 49-62 (2017). [10] Shibata, Terasawa, Osumi, Masui, Ito, Sato & Umeda.

Brain Res, 1657, 215-222 (2017). [11] Mayerhofer & Schacht. *Front. Psychol.* 6:550 (2015). [12] Uekermann, Daum & Channon. *Soc Cognition,* 25(4), 553-572 (2007). [13] Bambini, Bertini, Schaeken, Stella & Di Russo. *Front. Psychol.* 7:559 (2016). [14] Spotorno & Noveck. *J Exp Psychol Gen.* 143(4), 1649-1665 (2014).

Do speaker-specific cues influence ambiguous word interpretation?

Cat Davies[1], Vincent Porretta[2], Kremena Koleva[1], Ekaterini Klepousniotou[1]

[1] University of Leeds, UK; [2] University of Alberta, Canada

Speaker identity has been shown to be an influential factor in language processing across multiple linguistic domains, e.g. phonetics, syntax, reference, and pragmatics. Addressees use information from speakers' previous discourse to make predictions about incoming linguistic material and to restrict the choice of potential interpretations. For example, addressees disambiguate words during the earliest moments of processing based on whether a particular speaker had previously produced the same word (Creel et al., 2008). Addressees also disambiguate syntactic structures based on previously modelled attachment preferences by particular speakers (Kamide, 2012). In pragmatic processing, partner-specific stored information may be especially helpful in disambiguating intended meaning. For example, contrastive inferences are suspended if a particular speaker has been habitually over-informative (Grodner & Sedivy, 2011), and addressees adapt to speaker-specific biases in the intended meaning of scalar quantifiers (Yildirim et al., 2016).

Our study used polysemous words with metaphorical extensions, e.g. *head; chair; fork,* which can be interpreted to refer to a dominant, literal meaning, as well as to a lower-frequency, metaphorical meaning, to investigate the extent to which speaker-specific cues influence semantic interpretation. Using an exposure-test design, speaker identity was manipulated by training participants to associate a specific speaker with a highly literal or a highly metaphorical style. At test, participants responded to video instructions from each speaker to *'click on the X'* while their eye movements were tracked using the visual world paradigm. We hypothesised that participants would ultimately resolve reference to the literal target (LT, e.g., dinner fork) rather than the metaphorical target (MT, e.g., fork in the road) in both speaker-style conditions due to its meaning dominance. However, if addressees use speaker-specific information to disambiguate referring expressions, we predicted that participants would experience interference from the MT in the metaphorical speaker condition, indexed in that condition by i) longer reaction times for resolution to the LT in the metaphorical style condition; and ii) a lower proportion of looks to the LT while processing the ambiguous noun.

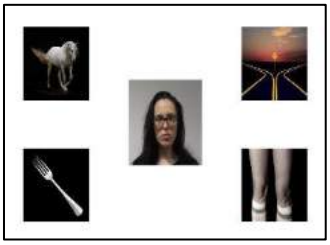


Figure 1. Example test item. Instruction: *click on the fork.*

As expected, across speaker conditions, 89% of referring expressions were resolved to the LT and 10% to the MT (the remaining 1% were unresolved before timing out). Contrary to our prediction, there was no effect of speaker style on reaction times. Given the dominance of LT responses, we examined gaze data from noun onset to trial end on trials resolving to the LT. GLMER was used to analyse LT preference (i.e., looks to the LT vs. looks to the MT) as a function of speaker style. As Figure 2 shows, a significant effect was found in two critical time windows. In the early window (400-850ms), participants' preference for the LT was significantly reduced in response to the metaphorical speaker (estimate = -1.69, $SE = 0.64$, $p < .01$), as hypothesised. Conversely, in the late window (850-1300ms), participants' preference for the LT was significantly greater in response to the metaphorical speaker (estimate = 1.62, $SE = 0.57$, $p < .01$). This suggests early anticipation and interference of the MT in response to the metaphorical speaker. The later preference for the LT in this condition is likely due to participants double-checking the initial

interpretation. These patterns reflect listeners' assumptions that the metaphorical speaker may have intended the expressions to have a non-literal meaning.

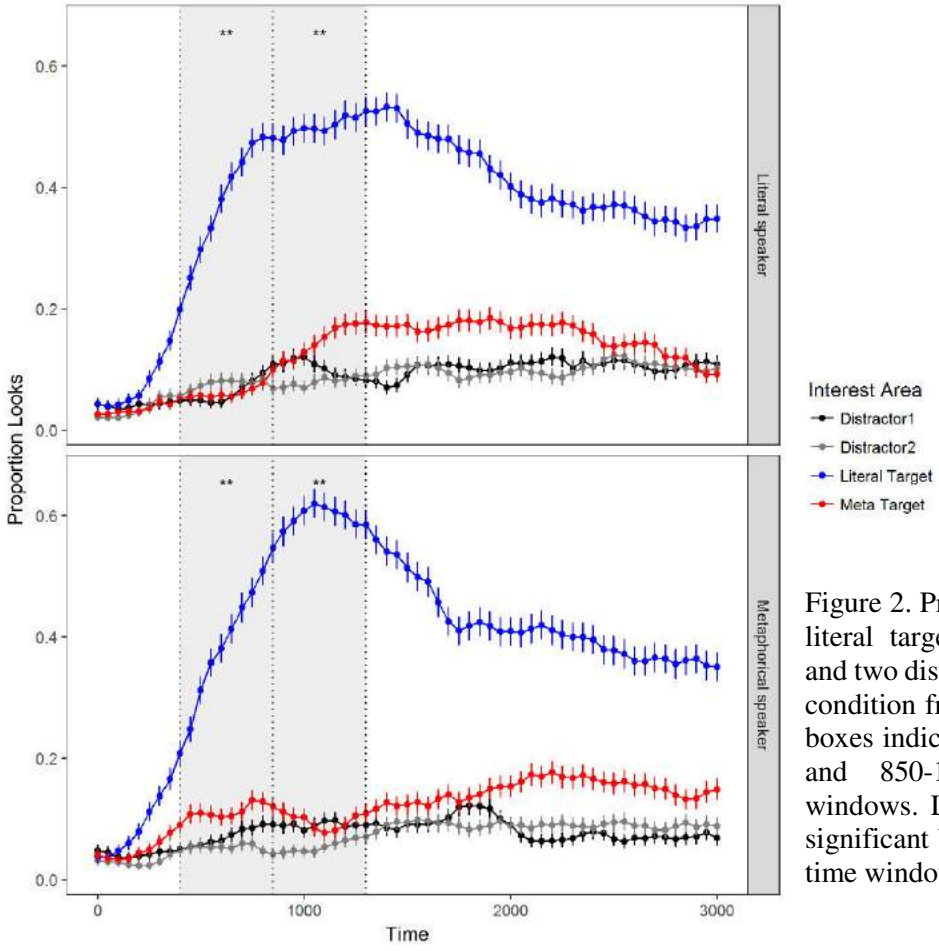


Figure 2. Proportion of looks to the literal target, metaphorical target, and two distractors by speaker-style condition from noun onset. Shaded boxes indicate 400-850ms ('early') and 850-1300ms ('late') time windows. Literal target preference significant by speaker style in both time windows ($p < .01$).

Our results support accounts proposing that semantic comprehension involves rapid integration of multiple cues including those of a social nature (Rodd, 2017). We provide evidence that speaker style is a contextual determinant in semantic disambiguation using polysemous words. Our findings extend the literature on partner-specific effects to the domain of semantic processing.

References

Creel. S.C., Aslin, R.N. & Tanenhaus, M.K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, 106, 633-664.

Grodner, D. & Sedivy, J. (2011). The effects of speaker-specific information on pragmatic inferences. In N. Pearlmutter & E. Gibson (eds). *The Processing and Acquisition of Reference.* MIT Press: Cambridge, MA.

Kamide, Y. (2012) Learning individual talkers' structural preferences. *Cognition,* 124, 66-71.

Rodd, J. (2017). Resolving ambiguities in spoken language: Evidence from skilled adult comprehenders. Presentation at *Multi-disciplinary approaches to understanding social communication development and disorder* (University College London, 13/02/2017).

Yildirim, I., Degen, J., Tanenhaus, M.K., and Jaeger, T.F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*. 87, 128-143.

Speed and accuracy trade-off and their link to neural processes of meaning composition

Diana V. Dimitrova[1], Brian McElree[2] & Petra Schumacher[1]
[1]University of Cologne, Germany; [2]New York University, USA

When listeners interpret a message, they activate the meaning of words from memory and integrate them into a discourse representation. Numerous studies have tested the predictive role of context in creating meaning, however they have neglected the contribution of adjectives, which strongly affect the computation of meaning in noun phrases (NPs). While some adjectives like "white" specify the denotation set of the noun and are less context-dependent ("white diamond"), other adjectives like "real" must be enriched since their meaning is context-dependent ("real diamond"). So-called "real"-adjectives might be pragmatically over-informative and therefore readers might need to compute a contrast set to arrive at their interpretation. In contrast, "fake"-adjectives negate the meaning of the noun and lead to a contradiction (a fake diamond is a diamond in some respect and not a diamond in another respect) (Kamp & Partee 1995). Previous ERP studies suggest that "fake"-adjectives initiate processes of reanalysis since listeners need to repair the contradiction, which gives rise to a Late Positive Component (Schumacher et al. 2018). In contrast, "real"-adjectives do not cause extra processing costs (Schumacher et al. 2018). The process of composition is also modulated by the adjective's polarity: negative adjectives like "fake" cause higher processing costs, which is reflected in a higher N400 amplitude relative to positive adjectives (Herbert et al., 2008; Schumacher et al., 2018). How do enrichment and polarity differences in adjectives affect compositional processing? We designed a behavioral study to test how enrichment and polarity modulate the speed and accuracy of composition and what neural mechanisms underlie these processes.

We applied the innovative multi-response Speed-Accuracy-Tradeoff (SAT) task (Foraker & McElree 2011) where the speed and accuracy of a behavioral response are measured as a dynamically developing response function at pre-determined response lags, ranging from incomplete (stimulus onset) to complete processing (5s post stimulus). 22 German participants read sentences on a computer screen presented in segments like "The tradesman | buys | a real diamond". Two factors were varied to build the four experimental conditions: *composition* (neutral: "white", "flawed" vs. enriched adjectives: "real", "fake") and *polarity* (positive: "white", "real" vs. negative adjectives: "flawed", "fake"). Adjective type was determined by pretests on polarity. Upon display of the target NP "a real diamond", a series of 15 tones (1 kHz, 50ms duration, 350ms lag latency) was played. Participants indicated by key press if the sentence was meaningful; they could change their response by switching to a different key. The SAT function (Figure 1) was computed based on three parameters: (i) *asymptote*, the response accuracy (d') at each time lag, (ii) *rate*, the response speed at each lag, and (iii) *intercept*, the time point at which accuracy departs from chance. D' was calculated by scaling the four experimental conditions against an implausible condition "The tourist buys a flying diamond".

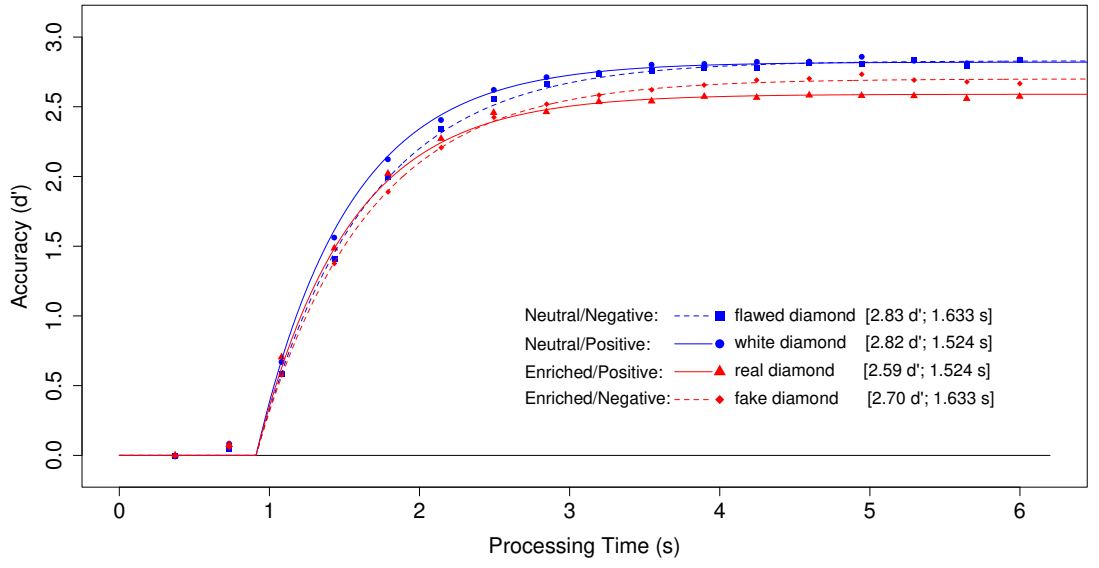
The results show first that accuracy judgments are significantly lower for enriched NPs ("real/fake diamond") vs. neutral NPs ("white/flawed diamond"), with "real"-type adjectives having the lowest asymptote. This suggests that listeners may not always arrive at an enriched representation of "real diamond"-NPs, since their interpretation strongly depends on subjective judgment and sentience, which vary across individuals. In contrast, "fake diamond"-NPs seem to be more easily interpretable, since the contradiction (a fake diamond is a diamond in some respect and not a diamond in another respect) must be resolved during composition. The finding that "fake"-type combinations are more accurate than "real"-type combinations also narrows down the possible explanations of the Late Positivity observed in previous ERP research: it precludes well-formedness as a potential explanation and substantiates the claim that processing costs are associated with reconceptualization. Second, concerning polarity, the SAT data show that it

modulates processing rate: negative adjective-noun combinations ("flawed/fake diamond") had a lower rate and thus required more processing time than positive adjective-noun combinations ("white/real diamond"). This result is in line with prior ERP studies that found enhanced processing demands for negative information (Herbert et al. 2008; Schumacher et al. 2018). The longer processing time for negative adjectives further supports the claim for a negative bias in information processing (e.g., Alves et al. 2017). These demands are observable independent of the type of composition (neutral vs. enriched). We conclude that processes of enrichment are modulated by the polarity of adjectives and the type of composition. Contradictions arising during compositionality must be resolved and engender processing costs (reflected by Late Positive ERP effects) while combinations with more vague, over-informative adjectives may not be fully interpreted (indicated by lower accuracy).

Example stimuli

(1a) Enriched/Positive: The tradesman buys a real diamond.
(1b) Neutral/Positive: The tradesman buys a white diamond.
(2a) Enriched/Negative: The tradesman buys a fake diamond.
(2b) Neutral/Negative: The tradesman buys a flawed diamond.

Figure 1: SAT function to the four experimental conditions.



References

Alves, H., Koch, A. S., & Unkelbach, C. (2017). Why good is more alike than bad: Processing implications. *Trends in Cognitive Sciences*, *21*, 72–82.

Foraker, S., & McElree, B. (2011). Comprehension of Linguistic Dependencies: Speed-Accuracy Tradeoff Evidence for Direct Access Retrieval From Memory. *Language and Linguistics compass*, *5*(11), 764-783.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*(2), 129-191.

Schumacher, P. B. (2013). When combinatorial processing results in reconceptualization: toward a new approach of compositionality. *Frontiers in Psychology*, *4*, 677.

Schumacher, P. B., Brandt, P., & Weiland-Breckle, H. (2018). Online processing of "real" and "fake": The cost of being strong. In: Castroviejo, E., McNally, L., & Weidman Sassoon, G. (Eds.) *The Semantics of Gradability, Vagueness, and Scale Structure: Experimental Perspectives*. Heidelberg: Springer.

Metaphorical Developing Minds: The role of multiple Factors in the Development of Metaphor Comprehension

Simona Di Paola[a], Filippo Domaneschi[a], Nausicaa Pouscoulous[b]

[a] *Department of Educational Sciences, Psychology Unit - University of Genoa, Italy*
[b] *Division of Psychology and Language Sciences - University College London, UK*

Metaphor understanding is traditionally thought to emerge late in childhood or even adolescence (Winner, 1988/1997), although recent findings suggest that even pre-schoolers can understand metaphor in perhaps more age appropriate paradigms (e.g. Özçaliskan, 2005). In any case, one wonders which skills scaffold the development of metaphorical abilities. Metaphor comprehension is a complex process relying on multiple higher-order cognitive abilities with different developmental paths, such as alternative naming and analogical reasoning (Rubio-Fernández & Grassman, 2016). This study was aimed at (i) teasing apart the contribution of these two cognitive abilities necessary for metaphor understanding: *Alternative Naming* (i.e. accepting two labels for the same referent) and *Analogical Reasoning* (i.e. detecting similarities across objects); (ii) assessing their developmental trajectories within a single experimental paradigm which included a metaphor task; (iii) further characterizing metaphor developmental trajectory by identifying possible enhancing/impeding factors.

We tested 3- (N: 20; age range: 3;1–3;9; mean age: 3;4) and 4-year-olds (N: 20; age range: 4;1-4;11; mean age: 4;5) in three tasks: Metaphor Comprehension, Alternative Naming and Analogical Reasoning. The general procedure consisted of a picture-matching paradigm adapted from Morriseau et al. (2013): the children were presented with several pictures arranged on a grid and were asked to move them according to the experimenter's instructions in order to match a given configuration. In the metaphor and alternative naming tasks, the experimenter asked the children to remove the pictures from the grid by referring to them either metaphorically or using an alternative label. In the analogy task, children were asked to choose the picture which best completed a given pattern and place it next to the other pictures of the sequence.

<u>*Metaphor Task*</u>: Eight triplets of pictures were shown to participants – each triplet corresponding to one trial. For each triplet/trial, the experimenter referred to the target picture either using a metaphor of the form [The X with the Y] or literally (e.g. *'Give me the glass with the antennae/glass with the straws'* for a picture depicting a glass with two straws, see Fig 1a). There were four trials in each condition – metaphorical and literal. Children could choose one of three pictures: (i) *Target* (e.g. a glass with 2 straws); (ii) *Control I*, a literal competitor where both target and vehicle were literally shown (e.g. a glass and a girl wearing an antennae-headband); (iii) *Control II*, showing the metaphor target deprived of the relevant property (e.g. a glass with no straws).

<u>*Alternative Naming Task*</u>: Children saw 13 pictures on the grid that they had to reconfigure according to the experimenters' instructions. Eight were target pictures: four of them were referred to with the same term as they had been previously (ST condition) and four with a new term, (NT condition; e.g. *'Give me the Lollipop/Candy',* see Fig. 1b).

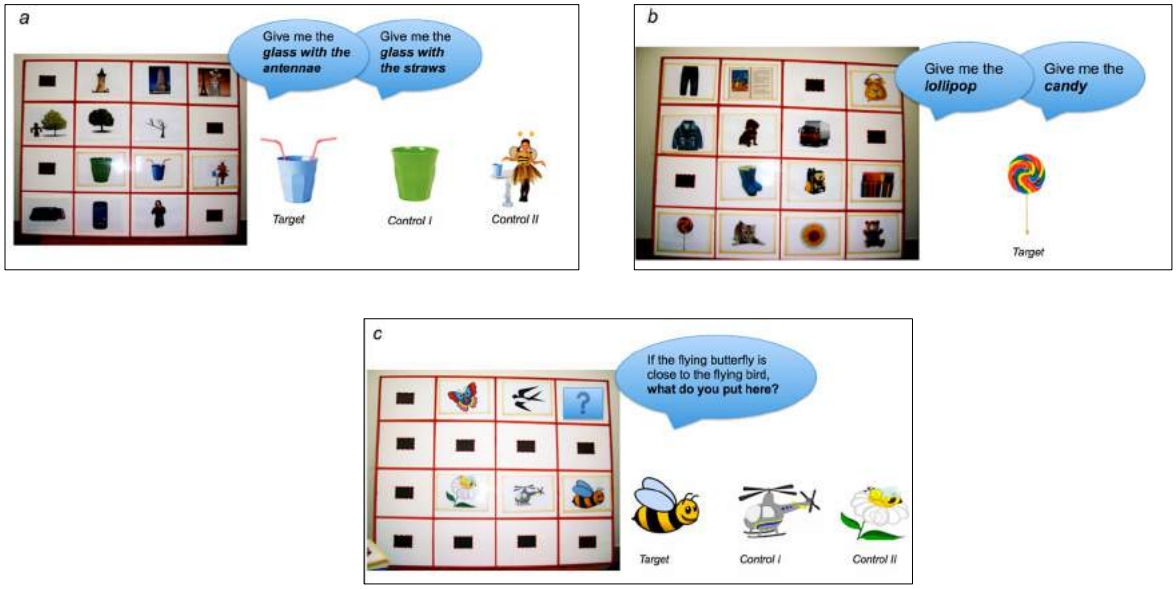
<u>*Analogy Task*</u>: In each of four trials, children had to choose the picture (out of three) which best completed a sequence of two pictures. The first two pictures of the sequence always shared a relational feature on which an analogy could be based (e.g. flying for the *animals that fly* analogy; see Fig. 1c). The three alternatives were: (i) *Target picture*, showing the relevant relational feature applied to the relevant object (e.g. a flying bee); (ii) *Control I*, showing the relevant property but on the irrelevant object (e.g. a helicopter); (iii) *Control II*, showing an irrelevant property but on the relevant object (e.g. a sleeping bee).

Vocabulary assessment: A picture naming-and-pointing game was administered to assess children's comprehension and production of the vocabulary used in the metaphor task.

Accuracy was coded for all tasks. Additionally, reaction times were collected for the metaphor and the alternative naming tasks. Data analysis was carried out separately for each task using Linear Mixed Models (LMM) statistics and Kendall's Tau correlations. Moreover, a Generalized Linear Mixed Models (GLMM) statistics was conducted to test if Alternative Naming and Analogy significantly predicted pre-schoolers' metaphor understanding. In the metaphor task, while all children exhibited more difficulties interpreting metaphorical than literal expressions ($p<0.0001$), their accuracy improved with age ($p <0.05$), with 4-year-olds only performing above chance. Both at ages 3 and 4 children performed well above chance in the Alternative Naming Task, with residual difficulties in 3-year-olds (NT vs. ST: $p<0.001$) significantly lowered by four (NT vs. ST: p=n.s.). Both 3-and-4-year-olds performed at chance in the Analogy Task, with no significant difference between groups (p = n.s.). Importantly, the GLMM statistics indicated that both Alternative Naming and Analogy significantly predict pre-schoolers' understanding of metaphor. Specifically, the faster children were in the alternative naming task the more accurate they were in the metaphor task ($p<0.0001$) and children with more developed analogical abilities showed a better performance in the metaphor task ($p<0.0001$).

Our findings suggest that Alternative Naming and Analogy play a role in the development of metaphoric competence. By age 4, children's difficulties with alternative naming are fully solved and are likely not to increase the cognitive demands imposed by metaphor interpretation. It might nonetheless still be a minor source of difficulty for 3-year-olds. Analogical perception, on the other hand, may hinder the ability to understand metaphors in 3-year-olds and even, to a lesser extent, in 4-year-olds. Overall, the development of metaphoric competence is likely to depend from a cluster of cognitive abilities including alternative naming and analogical-reasoning skills. Each cognitive ability within this cluster might enhance or impede preschoolers' interpretation of a metaphor depending on its developmental trajectory.

Figure 1: Example of the material used in one trial for each of the experimental task: (a) Metaphor Task; (b) Alternative Naming Task; (c) Analogy Task.



References

Morriseau, T.; Davies, C. & Matthew, D. (2013). How do 3- and 5-year-olds respond to under- and over-informative utterances? *Journal of Pragmatics*, 59: 26-39.

Özçaliskan, S. (2005). On learning to draw the distinction between physical and metaphorical motion: Is metaphor an early emerging cognitive and linguistic capacity? *Journal of Child Language,* 32: 291-318.

Rubio-Fernández, P. & Grassman, S. (2016). Metaphors as Second Labels: Difficult for Preschool Children? *Journal of Psycholinguist Research,* 45: 931-944.

Winner, E.(1988/1997). *The Point of Words: Children's Understanding of Metaphor and Irony.* Harvard University Press.

Sarah Dolscheid, Franziska Schleussinger, Martina Penke

Department for Rehabilitation and Special Education

University of Cologne

Different pragmatic interpretations of German 'eine' (a/one) in children and adults

Children seem to treat cardinal numbers (like 1, 2, 3) and quantifiers (like *some*) differently with respect to pragmatic principles (e.g. Hurewitz et al., 2006). For instance, 3-year-olds reject the claim that an alligator has *two* cookies when in fact he has *four*, whereas they accept that the alligator has *some* of the cookies when in fact he has *all* of them (Hurewitz et al., 2006). Children thus assign upper bounded interpretations to numbers but not to quantifiers. A similar observation holds for the numeral *one* vs. the indefinite determiner *a*. That is, English-speaking children do not accept two strawberries as a correct response to the question "Is there one strawberry in the red circle?", but they do if the question includes *a* instead of *one* strawberry (Barner et al., 2009). Unlike English, however, many languages do not draw a distinction between the indefinite determiner *a* and the numeral *one* (e.g. Sarnecka et al., 2007). In German, for instance, '*eine'* serves both functions. This raises the question of how German-speaking children and adults interpret the ambiguous term *eine*.

To find out, we tested 37 German-speaking children (3- to 6-year-olds) and 31 adults in a Truth-Value Judgment task (based on Barner et al., 2009). Participants were asked to answer the following question "Ist da eine Erdbeere in der Schüssel" 'Is there a/one strawberry in the bowl?', while they were presented with different numbers of strawberries (i.e., zero, one, or two; see Figure 1a). German-speaking adults predominantly showed an upper bounded interpretation of *eine*. Only the minority of adult speakers (32%) accepted two strawberries as a correct response to the ambiguous term. In contrast, the majority of the children (89%) considered two strawberries a correct response for *eine* (see Figure 1b). Unlike English-speaking children who draw a distinction between *one* and *a*, German-speaking children seem to interpret *eine* as the determiner *a* and not in an upper bounded way (i.e., exactly one). This is also in contrast to adult speakers of German who prefer an exact interpretation of *eine*.

Is it possible that context has an impact on adults' interpretation of *eine*? To test this assumption, we administered a modified version of the Truth-Value Judgment task. While the exact same question was asked, participants were presented not only with different numbers of strawberries but also with other types of fruit (e.g. two bananas, three oranges; see Figure 2a). Again, the majority of children accepted two strawberries as a correct response to *eine* (92%). However, also 61% of the adults accepted two strawberries as a correct response in the 'multi-fruit' context, resulting in a less exact interpretation of the term *eine* (see Figure 2b). Our results thus show that – depending on contextual information – the same question may or may not elicit an upper bounded interpretation in adults. In sum, our findings shed light on developmental aspects of quantifier acquisition as well as on factors that can influence the pragmatic interpretation of the ambiguous term *eine* in German.

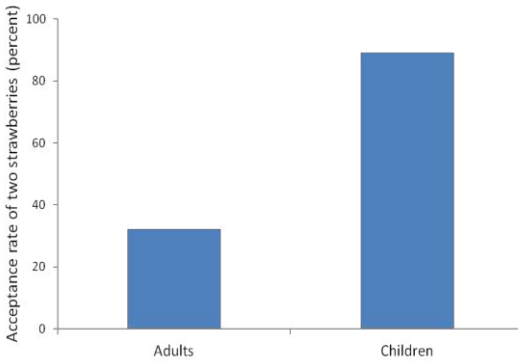
Figure 1

Truth-Value Judgment task ('strawberries only' condition)

a)

Ist da eine Erdbeere in der Schüssel? Is there a/one strawberry in the bowl?

b)



Children accepted two strawberries significantly more often as a correct response than adults (Fishers exact test; p<.0001).

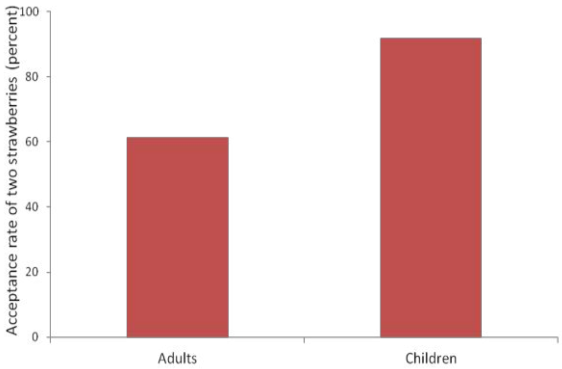
Figure 2

Truth-Value Judgment task ('multi-fruit' condition)

a)

Ist da eine Erdbeere in der Schüssel? Is there a/one strawberry in the bowl?

b)



Children still accepted two strawberries significantly more often than adults (Fishers exact test; p<.001). However, adult participants accepted two strawberries significantly more often in the 'multi-fruit' condition compared to the 'strawberries-only' condition (McNemar's test: p<. 01).

References

Barner, D., Chow, K., & Yang, S. J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology*, *58*(2), 195-219.

Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language learning and development*, *2*(2), 77-96.

Sarnecka, B. W., Kamenskaya, V. G., Yamana, Y., Ogura, T., & Yudovina, Y. B. (2007). From grammatical number to exact numbers: Early meanings of 'one', 'two', and 'three' in English, Russian, and Japanese. *Cognitive Psychology*, *55*(2), 136-168.

The Processing Costs of Presupposition Accommodation

Filippo Domaneschi and Simona Di Paola

Department of Educational Sciences, Psychology Unit - University of Genoa, Italy

Introduction: The study of the timing of availability of presuppositions (PSPs) in on-line language comprehension is crucial for characterizing PSPs as either a semantic or a pragmatic phenomenon (cf. Schwarz 2015:13): on the one hand, if PSPs are conceived as information conventionally encoded in the lexical meaning constituting a condition for the context update (Heim 1990, Heim & Kratzer 1998), then in processing a presupposing utterance, we should expect delays during the sentence processing, before the asserted content is computed. On the other hand, if PSPs are the result of pragmatic inferences based on the truth-conditional content (Simons 2002), delays should be expected after the asserted content is computed, as with conversational implicatures.

Some preliminary behavioural studies have suggested that the processing times of PSPs vary according to three processing conditions: satisfaction (SAT), accommodation (ACC) and falsification (FAL). Schwarz (2007) has shown that the overall reading times (RTs) for a sentence containing the focus particle *auch* are longer in ACC than in SAT. With a word-by-word paradigm, Tiemann et al. have found that, with different PSP triggers, ACC takes longer than FAL on the trigger region (Tiemann et al. 2011) and that ACC elicits longer processing times than SAT on the critical word *wieder* (Tiemann et al. 2015), suggesting that, at least with this trigger type, ACC as compared to SAT, starts immediately during the sentence processing.

Research Questions: two aspects about the on-line processing times of PSP accommodation are still on the way to be clarified: (i) does ACC compared to SAT elicit longer processing times independently of the PSP trigger in use? Or is this a difference related to specific trigger types? (ii) What is the time-course of presupposition accommodation? Or, in other words, are presuppositions accommodated online during the sentence processing or off-line after the asserted content is computed?

Method & Procedure: Within a self-paced reading times paradigm followed by a true/false task, participants (N: 42; mean age = 25.06) were asked to read 40 stories and answer 3 verification questions after each story. The stories (Table 1) were composed of 2 context sentences followed by 1 target sentence presented word-by-word. Four types of PSP triggers were used: *definite descriptions* (DD, N: 10), *change of state verbs* (CSV, N:10), *iterative expressions* (IT, N:10) and *focus-sensitive particles* (FC, N:10). Items were presented in 2 conditions: satisfaction (SAT), where the presupposed information activated by the trigger in the target sentence was made explicit in context sentence 1, and a neutral condition (NEU) where it was not and prompted accommodation. The verification questions were 2 distractors and 1 target question verifying the content of the presupposition. We collected participants' RTs on the word-by-word target sentence. We identified the following main regions of interest (Table 2): (i) for all the trigger types, the triggering point (T1); (ii) for CSV, IT and FC, the computational point (T2), where the content of the PSP becomes fully available.

Results: The high percentage of correct answer to the verification questions in the NEU condition (i.e. 74.89%) suggests that participants have mostly accommodated the presuppositions. RTs data revealed (i) significantly longer reading times for NEU than SAT on

T1 and T1+1 ($p<0.05$ in both regions) for all the PSP triggers; and (ii) a significant interaction ConditionXTrigger Type in T1+1 ($p< 0.005$) and in T2 ($p< 0.05$). Post-hoc comparisons revealed that the longest reading times were elicited in NEU with DDs in T1+1 (DD vs. CSV: $p< 0.05$; DD vs. FC: $p<0.005$; DD vs IT: $p<0.05$) and with ITs in T2 (IT vs. CSV: $p<0.05$; IT vs. FC: $p<0.05$).

Discussion: Data collected suggest that, independently of the PSP trigger in use: (i) ACC takes longer than SAT, reflecting the cognitive cost associated with a process of context repair; (ii) presuppositions seem to be processed online given that accommodation takes place immediately and proceeds incrementally while the sentence unfolds (i.e. effects on T1, T1+1 and T2); and (iii) different triggers differently affect the cognitive load of processing presuppositions: DDs and ITs are more cognitively demanding than other triggers at different phases of sentence processing. Overall, by extending the preliminary existing results, this study provides evidence for the on-line processing of presuppositions and supports the predictions of the semantic accounts of PSPs according to which PSPs are accommodated before the asserted content is computed.

Condition	**Context sentence 1**	**Context sentence 2**	**Target sentence**	**Verification questions**	
SAT	Before her pregnancy Gaia smoked ten cigarettes per day	The possible fetal diseases scare her a lot	From the very beginning she has *given up smoking* but her worries remained the same	*Target*	Was Gaia used to smoke?
				Distractor	Does Gaia have three kids?
NEU	Gaia is at the third month of her first pregnancy			*Distractor*	Is Gaia peaceful about her pregnancy?

Table 1. Example of an item with CSV in condition SAT and ACC. Literal translation from Italian

Trigger	*Word number*														
	1	2	3	4	5 (T1)	6	7 (T2)	8	9	10	11	12	13	14	15
DD	Un	mese	fa	il	grafico	ha	presentato	le	dimissioni	per	problemi	con	il	suo	capo
	One month ago, the designer has submitted his resignation due to problems with his boss														
IT	Marco	ha	dimenticato	di	nuovo	le	chiavi	e	purtroppo	è	rimasto	chiuso	fuori	dall'	ufficio
	Mark has forgotten again the keys and unfortunately he is remained closed out the office														
FC	Da	giovane	è	stato	anche	in	Australia	dove	ha	incontrato	la	sua	compagna	di	vita
	When he was young he also visited Australia where he met his current partner														
CSV	Fin	da	subito	ha	smesso	di	fumare	ma	le	sue	paure	sono	rimaste	sempre	uguali
	Since the beginning she has given up smoking but her worries remained the same														

Table 2. Example of target sentence for each trigger type presented word-by-word.

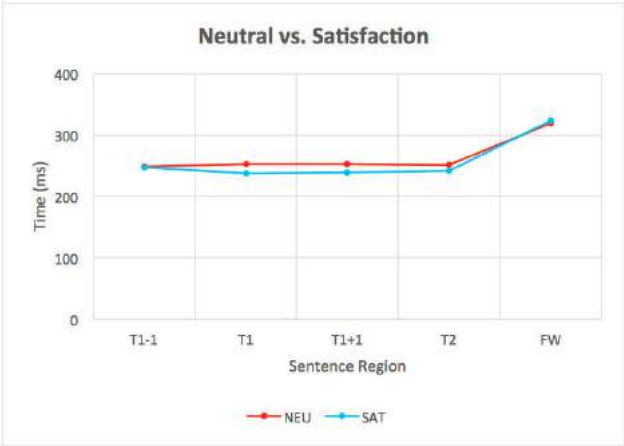


Figure 1. Mean reading times in conditions NEU vs SAT.

References

Domaneschi, F., Carrea, E., Greco, A, Penco, C. (2014). "The cognitive load of presupposition triggers. Mandatory and optional repairs in presupposition failure", in *Language and Cognitive Processes*, 29 (1), pp. 136-146.

Tiemann, S., Schmid, M., et al. (2011). "Psycholinguistic Evidence for Presuppositions: Online and Off-line Data", in Reich, Ingo et al. (eds.), *Proceedings of Sinn & Bedeutung*, 15, 581–595.

Tiemann, S., Kirsten, M., Beck, S., Hertrich, I., Rolke, B. (2015). Presupposition Processing and Accommodation: An Experiment on *wieder* ('again') and Consequences for Other Triggers, in Schwarz (2015) (Ed.), *Experimental Perspectives on Presupposition, Studies in Theoretical Psycholinguistics*, Dordrecht: Springer, 39-65.

Schwarz F. (2007). "Processing presupposed content". *Journal of Semantics* 24:373– 416.

What the hell? What swearing can tell us about conventional implicatures

Stanley Donahoo and Vicky Tzuyin Lai, University of Arizona

Expressives are not-at-issue, speaker-oriented, conventionally implicated content (Potts 2005). How is the expressive dimension (Potts, 2007; McCready, 2010) of language processed and represented? The present study focusses on the most clearly expressive items, swear words.

The study of swear words is important to linguistics and cognition in general, but has been a neglected area experimentally. In clinical populations, those with aphasia or who have had stroke can often recite lengthy chunks of memorized material, such as prayers, song lyrics, or greetings; in many cases, these automatic chunks include swearing (Van Lancker & Cummings, 1999). Infamously, pathological use of swear words is a defining characteristic of Gilles de la Tourette syndrome (Shapiro & Shapiro, 1982). Finally, some work on swearing's effect on memory has been done. Using a Stroop task, MacKay et al. (2004) show that swear words cause a slowdown in naming relative to 'neutral' words.

We have begun to explore the mental underpinnings of swearing, from a behavioural and electrophysiological perspective, influenced by an account which is rooted squarely in pragmatic theory. Potts (2005, 2007) provides a detailed account of conventional implicatures, arguing that they have six properties: independence, nondisplaceability, perspective dependence, descriptive ineffability, *immediacy*, and repeatability. Ever since Grice (1975) characterized them, implicatures as a whole have remained problematic to our current models of communication. Swear words thus provide an ideal and yet unexplored testing ground for exploring implicatures and the expressive dimension.

The present study examined swear words using a lexical decision task and will use EEG next. Stimuli are 30 swear words (e.g. *shit*, *damn*), 30 negatively valenced but non-swear words (*kill*, *sick*), 30 open class neutral words (e.g. *wood*, *lend*), 30 closed class neutral words, as swear words are a closed class as well (e.g. *while*, *whom*), and 120 pseudowords, for a total of 240 items. Norms for valence and other dimensions were obtained from recent corpus work (Warriner, Kuperman, & Brysbaert, 2013; Balota et al., 2007). A summary of the stimuli can be seen in Table 1.

Table 1. Mean values for psycholinguistic variables for each stimulus type.

	Psycholinguistic Measure					
	Letter Length	Subtlex Freq.	Log Subtlex	Ortho N	Num Phonemes	Valence
Swear Words (*shit*)	4.96	119.93	3.39	8.24	3.84	3.38
Negative Valence Words (*kill*)	4.89	119.47	3.59	6.18	4.14	2.45
Open Class Neutral Words (*wood*)	5.00	116.69	3.25	6.39	4.04	6.15
Closed Class Words (*while*)	4.97	113.45	3.36	6.37	3.83	X[1]

[1] Values for only 6 of the 28 words were listed in the Warriner corpus, so we chose not to list an average for this category.

p value for t-tests comparing Swear Words with all others	.95	.56 (.99 for swear and neg only)	.55	.91	.51	NA

Participants had to decide whether a letter string presented on a computer screen was a word of English or not. We hypothesise that contrary to other negatively valenced words, swearing will cause an increase in reaction time, as this content may possibly be initially processed by a less efficient, non-linguistic channel. Results of 27 participants were analysed using linear-mixed effects modelling in R (Baayen, 2008). Any *t* value greater than 2.0 was deemed significant. Effects of type, excluding RTs more than 2.5 standard deviations from the mean, were robust (*t*=3.49 for Negative Valence; *t*=2.88 for Neutral; *t*=2.81 for Nonwords) with respect to swear words. Closed Class words were not different with respect to swear words (*t*=.23). Behavioural results are displayed in Table 2.

Table 2. $\log^{-1}$RT in ms for each stimulus type. Significance is with respect to swear words as the baseline (0~*, .001~**, .01~*, .05~.).**

Swear	Negative Valence	Open Class Neutral	Closed Class	Nonwords
547.1	511.7***	517.1**	557.8	598.5**

These behavioural results show that swear words are more effortful for subjects than other words that are similar in their negative affect, meaning that there is more to the expressive dimension than merely a heightened emotional state. Our results are situated within the Potts framework, and ramifications for theories of implicature, both conventional and conversational, will be discussed. In particular, Potts's principle of immediacy will be examined.

We are currently setting up the EEG version of the experiment. We predict an increased N400 response to swear words relative to the negative valenced words. King and Kutas (1995) showed that, if closed class words are contextually unexpected in a sentence context, they will induce an N400. Additionally, the N280 will also allow us to better explore the result that the swear words behave similarly to the other closed class words. EEG data collection to corroborate these behavioural findings is underway, and will help to isolate where in the brain these words are being processed.

With these results, we contribute a new data set for probing our understanding of the central properties of implicatures, further demarcating the semantic-pragmatic boundary.

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior research methods*, *39*(3), 445-459.

Grice, H. P. (1975). 'Logic and Conversation' in P. Cole and J. Morgan (eds.) Syntax and Semantics Volume 3: Speech Acts.

Mackay, D. G., Shafto, M., Taylor, J. K., Marian, D. E., Abrams, L., & Dyer, J. R. (2004). Relations between emotion, memory, and attention: Evidence from taboo Stroop, lexical decision, and immediate memory tasks. *Memory & Cognition*, *32*(3), 474-488.

McCready, E. S. (2010). Varieties of conventional implicature. *Semantics and Pragmatics*, *3*, 8-1.

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

Potts, C. (2007). The expressive dimension. *Theoretical linguistics*, *33*(2), 165-198.

Shapiro, A. K., & Shapiro, E. (1982). Tourette syndrome: history and present status. *Advances in neurology*, *35*, 17.

Van Lancker, D., & Cummings, J. L. (1999). Expletives: Neurolinguistic and neurobehavioral perspectives on swearing. *Brain research reviews*, *31*(1), 83-104.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, *45*(4), 1191-1207.

Quantity implicatures in a competitive game

Giulio Dulcinati (UCL) *Nausicaa Pouscoulous (UCL)*

In Grice's account (1989) quantity implicatures are afforded by the assumption that the speaker is cooperative and therefore informative. Grice does not give an account of how the expectations of hearers are affected when they cannot assume that the speaker is cooperative. However, from Grice's account we can derive the prediction that rational hearers should not expect a non-cooperative speaker to abide by the maxim of quantity and therefore they should not derive quantity implicatures. By contrast, Sperber et al. (2010) would predict that hearers derive relevant implicatures regardless, but might then decide not to believe them if the speaker is not trustworthy. We set out to test these predictions using a signalling game.

In our experiment 140 Native English speakers played an online game with another (virtual) player. In each round of the game they saw two cards (a winning card and a losing card) and they read a short description of the winning card written by the other player. Their goal in each round was to click on the winning card. They were assigned to one of three conditions:

- a *cooperative condition*, where the other player's goal was to help them click on the winning card
- a *competitive condition*, where the other player's goal was for them to click on the losing card
- a *competitive-truthful condition*, where the other player was also playing *against* them but was not allowed to lie.

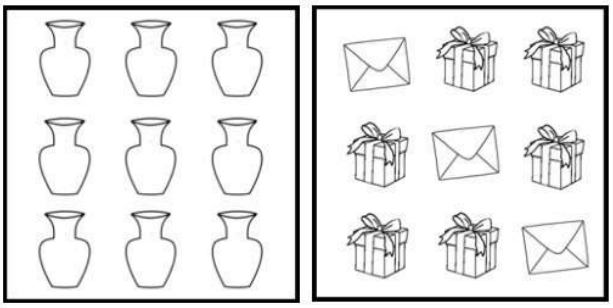


Fig. 1 *Cards associated with the description "On the winning card none of the objects are vases"*

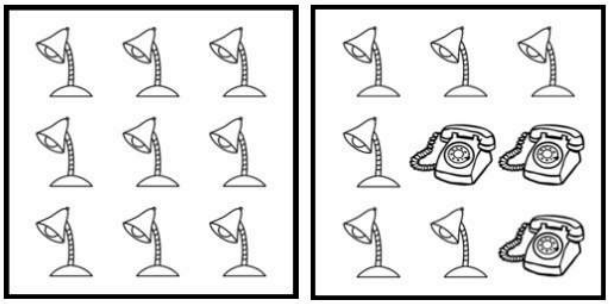


Fig. 2 *Cards associated with the description "On the winning card most of the objects are lamps"*

Each participant saw 16 control items, where the description was true of one card but false of the other (e.g., **Fig.1**), and 16 experimental items, where the description was true of both cards but could give rise to a quantity implicature (either a scalar implicature with *most* or *some* or an *ad hoc* quantity implicature) which was only true of one of the two cards (e.g., **Fig.2**).

We analysed the frequency of 'true' responses for control items (i.e. clicking on the card that fits the description) and 'pragmatic response' for experimental items (i.e. clicking on the card that fits the quantity implicature of the description) in each condition (**Fig.3**). We found that the rate of pragmatic responses in experimental items was significantly lower in the competitive (W_{subj}=448, p<0.001; V_{items}=136, p<0.001) and competitive-truthful (W_{subj}=601, p<0.001; V_{items}=136, p<0.001) conditions compared to the cooperative condition. This indicates that our manipulation did have an effect. If participants in the competitive conditions did not infer implicatures at all, they should have no preference between the pragmatic response and the other response. Therefore, we compared the rate of pragmatic

choices with chance level (p=0.5) and found that it was significantly higher than chance in the competitive-truthful condition (V_{subj}=720.5, p<0.001; V_{items}=136, p<0.001) but not in the competitive condition (V_{subj}=648, p=0.541; V_{items}=101, p=0.090). This preference for the pragmatic response in the competitive-truthful condition indicates that participants are inferring *and* accepting implicatures to some extent, which is not in line with the prediction we derived from the Gricean account. A plausible explanation for this result and for the difference between the two competitive conditions is that participants consider false implicature to be lies (Meibauer, 2014) and therefore tend to accept the content of the implicature in the competitive-truth condition but not in the competitive condition where lying is allowed. However, if participants considered false implicatures to be lies to the same extent as false assertions, they should not choose the pragmatic response in experimental items less frequently than they choose the 'true' response for control items in the competitive-truthful condition. We compared the responses to control and experimental items in the competitive-truthful condition and we found that the rate of pragmatic responses to experimental items was significantly lower than the rate of 'true' responses to control items (V_{subj}=486, p<0.001; W_{items}=0, p<0.001). This suggests that false implicatures are not considered lies to the same extent as false assertions. We also investigated whether there was a difference between scalar and ad hoc implicatures in this respect and we found that the rate of pragmatic choices was significantly higher for scalar implicatures than for ad hoc implicatures in the competitive-truthful condition (V_{subj}=285.5, p=0.004; W_{items}=62, p<0.001) but not in the cooperative condition (V_{subj}=87, p=0.205; W_{items}=31, p=0.922). Since this difference is present in the competitive-truthful condition but not in the cooperative condition, its cause is probably not a simple difference in the availability of the inferences but the fact that false scalar implicatures are more likely to be considered lies than false ad hoc implicatures.

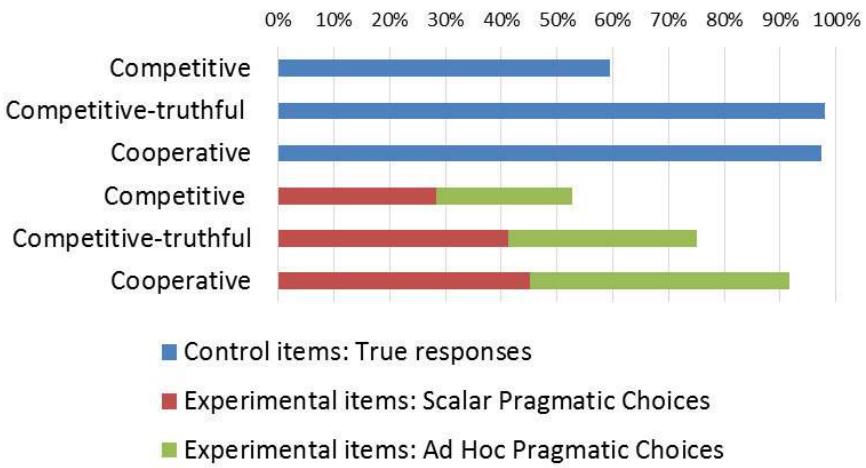


Fig. 3 *Frequency of true/pragmatic responses in each condition*

In conclusion, our results suggest that hearers faced with a non-cooperative speaker infer quantity implicatures but they are less likely to believe them than if the speaker were a cooperative speaker. This is more in line with the view of Sperber et al. (2010) than Grice (1989). These results are also informative for the status of false implicatures with respect to lies (Meibauer, 2014).

References:

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

Meibauer, J. (2014). *Lying at the semantics-pragmatics interface*. Berlin: Mouton de Gruyter.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language 25(4)*, 359-393.

The realization of bouletic bias: Evidence from German questions

Sophie Egger, Bettina Braun, Nicole Dehé
University of Konstanz

In everyday life we use subtle ways to communicate desires, often without explicitly saying so. Asking questions is one way to indirectly utter desires. Questions with an additional non-truth-conditional aspect are referred to as *biased* [1]: they are not plainly information seeking but additionally express an attitude towards one of the possible answers, e.g., a wish or desire (*bouletic bias*) [1-5]. We hypothesize that speakers successfully convey their desires when expressing them in a biased question, given that interlocutors seem aware of it. In order to better understand what leads to this success, we investigate the prosodic and morphosyntactic realization of bouletic bias in German polar (PolQs) and alternative questions (AltQs). Since PolQs highlight one particular alternative from a set of propositions [6, 7], we expect them to be more appropriate to mark bias. In AltQs, however, both alternatives are equally epistemically available which makes them suitable to offer an unbiased (i.e., neutral) choice [6]. We hence predict that speakers produce more positive PolQs in biased and more AltQs in unbiased contexts.

In a production experiment, we used 32 situational contexts, evoking either a biased (16 contexts) or a neutral question (16 contexts, within-items design; see Table 1). They were presented together with either a PolQ or an AltQ (manipulated between-items, 32 trials per participant). Each trial started with a context displayed on screen. By pressing a button, participants saw the target question which they were asked to produce (part 1). After another button press, they were given the possibility to rephrase the question in a way that seemed most natural (part 2). Part 1 thus enables us to perform a fine-grained acoustic analysis in a segmentally stable environment, whereas part 2 directly reveals the morphosyntactic structure preferred for biased and neutral questions, respectively. Sixteen speakers (Ø = 23.3 years, 12 male) produced 512 target questions (256 biased, 256 neutral).

Our prosodic analysis follows those in previous work about the realization of epistemic bias in PolQs by [8]. So far, we manually annotated a subset of 128 productions from part 1 (4 different contexts: 32 neutral/ 32 biased AltQs, 32 neutral/ 32 biased PolQs) according to GToBI [9]. Results showed that AltQs were generally produced with a low plateau ((H+)L* L-%) in both conditions (neutral: 89%, biased: 85%). Neutral PolQs were mostly produced with a final high rise (L* H-^H%, 68%), while biased PolQs showed either a final high rise (L* H-^H%, 44%) or a low rise (L* L-H%, 34%). We also found differences in pitch accent placement: biased PolQs are more often produced with a pitch accent on the pronoun (*Willst DU das Schoko-Eis?,* 'Do YOU want the chocolate ice cream?') than neutral PolQs (neutral: 10%, biased: 28%). The pitch range in the final rise in neutral PolQs is higher than in biased PolQs (neutral: 10.3st ; biased : 9.8st). In AltQs we find the reverse picture: in biased AltQs the pitch range in the final fall is higher than in neutral AltQs (neutral: 7.3st; biased: 8.3st).

The 375 target questions (201 biased, 174 neutral) produced in part 2 were coded for syntactic type (AltQ, PolQ, tag-question, *wh*-question, other). Participants predominantly produced PolQs in the biased condition (74%) and AltQs in the neutral condition (69%); see Figure 1. The question types presented in part 1 were changed in 70% of the biased contexts from AltQ to PolQ, and in 67% of the neutral contexts from PolQ to AltQ, showing strong preferences for particular question types according to context.

Our findings corroborate the assumption that positive PolQs tend to convey bias [6, 7], while AltQs function as neutral questions more readily [6]. There appear to be some differences in the preferred intonational realization across conditions. Also, speakers use an increased pitch range to compensate for the non-prototypical morphosyntactic structure when producing AltQs in biased contexts and PolQs in neutral contexts. However, we leave key phonetic differences in voice quality, segmental durations or the exact realization of the intonation contours (e.g., slope, peak alignment) for future analyses.

Neutral condition	Biased condition
You and one of your friends are going on vacation and driving with an intercity-bus. You are able to get two seats next to each other. It doesn't matter to you, where you sit, but you **don't know** which seat your friend prefers. Therefore you ask him…	You and one of your friends are going on vacation and driving with an intercity-bus. You are able to get two seats next to each other. You would like to have the window seat and hope that your friend wants to sit at the aisle. You ask him…
Speaker intention:	
I want to know whether you want the window seat or the aisle seat.	*I want you to take the aisle seat.*
Target questions: (either PolQ or AltQ presented on screen in part 1)	
PolQ: *Do you want to sit by the aisle?*	
AltQ: *Do you want to sit by the window or by the aisle?*	

Table 1: *Example of a neutral and a biased context with speaker intention and example of a PolQ and AltQ used in both conditions (question type was manipulated between-items).*

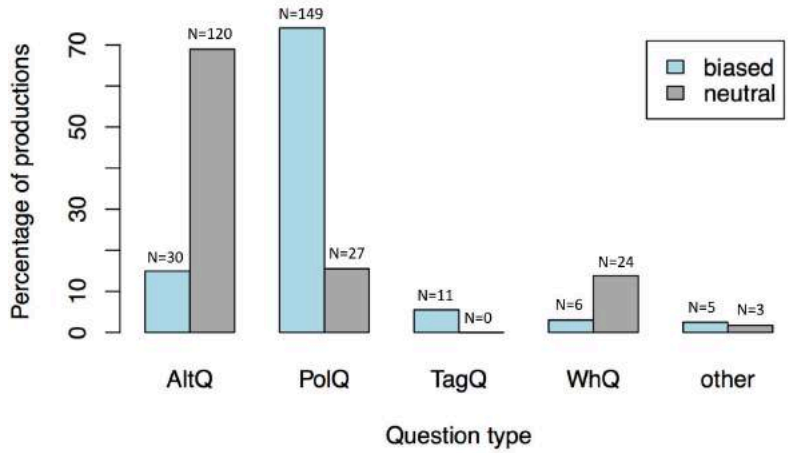


Figure 1: *Percentage of productions of each question type per condition.*

[1] Sudo, Y. 2013. Biased polar questions in English and Japanese. *Beyond expressives: Explorations in use-conditional meaning*, 275-296.

[2] Reese, B. 2007. *Bias in Questions* (PhD Dissertation). University of Texas, Austin, TX.

[3] Huddleston, R., & Pullum, G.K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

[4] Reese, B. 2006. The meaning and use of negative polar interrogatives. *Empirical Issues in Syntax and Semantics*, 331-354.

[5] van Rooij, R., & Šafárová, M. 2003. On Polar Questions. Talk at the *Semantics and Linguistic Theory, 13* (Seattle, WA).

[6] Biezma, M., & Rawlins, K. 2012. Responding to alternative and polar questions. *Linguistics and Philosophy*, *35(5)*, 361-406.

[7] Huddleston, R. 1994. The contrast between interrogatives and questions. *Journal of Linguistics, 30(2),* 411-439.

[8] Domaneschi, F., Romero, M., & Braun, B. (2017). Bias in polar questions: Evidence from English and German production experiments. *Glossa: a journal of general linguistics, 2(1)*: 26. 1–28,

[9] Baumann, S., & Grice, M. 2006. The intonation of accessibility. *Journal of Pragmatics, 38(10),* 1636-1657.

Find a friend or a scale mate: comparing ad hoc and scalar implicatures

Francesca Foppolo[a], Francesca Panzeri[a], Greta Mazzaggio[b] & Luca Surian[b]
[a]University of Milano-Bicocca & [b]University of Trento

Children's pragmatic abilities have been the matter of a vivid debate since at least Chierchia et al. (2001) and Noveck (2001). Several studies in the past years investigated children's derivation of pragmatic inferences by testing different items in different languages and populations and by means of different tasks. Overall, differences have been found across ages, types of items and tasks (cf. Skordos & Papafragou, 2016 for a review). In general, pre-schoolers have difficulties in the computation of the scalar implicature (SI) related to *some*, while a better performance has been documented in the case of non scalar or ad hoc scales, even in younger kids (Katsos & Bishop, 2011; Stiller, Goodman & Frank, 2015). Children's difficulty have been explained by different hypotheses: children are more tolerant of pragmatic violations than adults (Katsos & Bishop, 2011); children have difficulties in lexicalizing the scale and/or retrieving the lexical alternatives (Barner et al., 2011; Foppolo et al., 2012; Tieu et al., 2015); children do not (always) recognize what is conversationally relevant (Skordos & Papafragou, 2016).

Our study. In our experimental study, we tested 58 pre-school children (age range (in months): 45-72, MA = 60,58) split in two age groups (29 5-6 year olds, labelled "old"; 29 3-4 year olds, labelled "young"). Participants were administered three tasks:

- a classic Truth Value Judgment task in which children had to judge sentences like *"He put some of the cookies in the box"* in a situation in which all the cookies are in the box (Figure 1). We also assessed children's competence with scalar quantifiers *some* and *all* in true and false situations.
- a novel task for scalar implicature computation in which participants had to find the correct target (among 4 pictures) by exploiting a sentential cue (Figure 2). The task is a classical picture selection (PST), with the novelty that the relevant *all*-alternative was provided as a visual contrast (Figure 2).
- a PST for Ad Hoc scales modelled after Surian & Job (1987) and Stiller, Goodman & Frank (2015) in which participants had to find the correct target (among 4 pictures) by exploiting a sentential cue (Figure 3). The paradigm was analogous to that employed for *some.*

Results. Main results are plotted in Figure 4. Data were analysed by means of mixed models in which children's performance was modelled after Age and Task. We found: (i) a significant effect of age in the derivation of the scalar implicature connected to *some*, both in the TVJT (accuracy (old vs. young): 71% vs. 36%, p = .0135), and in the PST (accuracy (old vs. young): 66% vs. 38%, p = .0059), but not in the case of Ad Hoc implicatures, in which the two age groups did not differ (old = 81%, young = 77%, p = .3566); (ii) a significant difference between scalar and ad hoc implicatures when using the same task (PST): while overall accuracy was 52% in the case of *some*, it was 79% in the Ad Hoc scales (p = .0051). Interestingly, children's performance with the scalar item *some* did not improved in the PST compared to the classical TVJT (53% vs. 52%, p = .542).

Discussion. Our findings add an additional piece to the understanding of children's failure and success with scalar inferencing. In particular, we show that, in a task that is designed to enhance contextual relevance of the alternatives, children succeed with ad hoc implicatures but fail with scalar implicatures. We interpret these results in light of a lexical hypothesis to SI: children are able to derive the *some but not all* implicature only at a developmental stage in which they have lexicalized the scale <some, all>, i.e. they know that *some* and *all* are scale mates ordered on an informativeness scale. Before that stage, their performance is equivocal. In the case of ad hoc implicatures, in which a lexicalization is not required, their performance is good even at a young age. This, in turn, demonstrates that children are not, in general, more logical or more tolerant than adults; indeed, they are capable of generating alternatives, are sensitive to informativeness and are capable of deriving pragmatic inferences, provided that they can match scale mates in a scale, an operation that takes more time for scalar quantifiers.

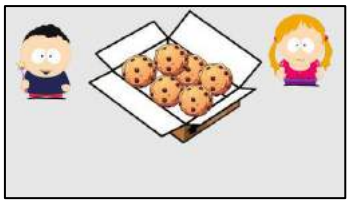


Figure 1. Scalar Implicatures – Truth Value Judgment Task

Target sentence
He put some of the cookies in the box.

Figure 2. Scalar Implicatures – Picture Selection Task
Lead-in sentence
Guess which one is my birthday cake, I give you a cue.
Target sentence
On my birthday cake, some of the candles are burning.

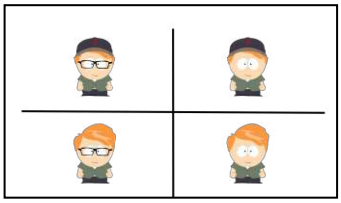


Figure 3. Ad Hoc Implicatures – Picture Selection Task
Lead-in sentence
Guess who is my friend, I give you a cue.
Target sentence
My friend wears glasses.

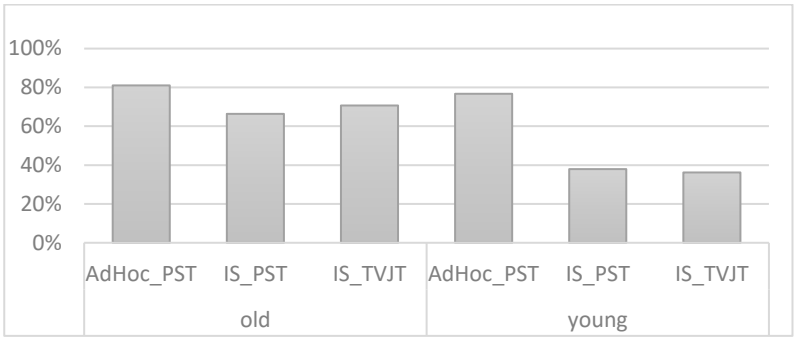


Figure 4. Children's accuracy (by age group and type of task)

References

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, 118(1), 84-93.

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

Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language learning and development*, 8(4), 365-394.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1), 67-81.

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

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

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

Surian, L., & Job, R. (1987). Children's use of conversational rules in a referential communication task. *Journal of Psycholinguistic Research*, 16(4), 369-382.

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

Two response systems for German *ja* and *nein*?
Evidence from usage preference data and interpretation data

Felix Frühauf[1], Berry Claus[1], Sophie Repp[2], Manfred Krifka[1] and A. Marlijn Meijer[1]
[1] Humboldt-Universität zu Berlin | [2] Universität zu Köln

The German response particle system is a three particle system: *ja*, *nein*, and *doch*. In responses to positive assertions and questions, *ja* and *nein* are translation equivalents of *yes* and *no*; *ja* affirms and *nein* rejects a positive antecedent. *Doch* is a dedicated particle for rejecting responses to negative antecedents (see 1.B.i). In affirming responses, both, *ja* and *nein*, can be used (see 1.B.ii), with *ja* signaling the truth of the antecedent (*truth-based strategy*; cf. Jones 1999) and *nein* signaling that the response has negative polarity (*polarity-based strategy*; cf. Jones 1999).

(1) A: Anna raucht nicht. ('Anna doesn't smoke.')
B: i. Doch. (= She does.) ii. Ja/Nein. (= She doesn't.)

Two recent approaches to response particles (Roelofsen & Farkas 2015, Krifka 2013) can account for the pattern for *ja* and *nein* with competing theoretical analyses. In a nutshell, Roelofsen & Farkas propose that *ja* and *nein* do double duty: *ja* can indicate that the antecedent is true or that the response is positive, and *nein* can signal that the antecedent is false or that the response is negative. Krifka assumes that negative utterances introduce two propositional discourse referents, a negative one and its positive counterpart, which both can be picked up with *ja* and *nein*, with *ja* asserting the targeted discourse referent and *nein* asserting its negation. Both approaches predict that by default *nein* is preferred over *ja* in affirming responses to negative antecedents. In two acceptability-judgment experiments, however, Claus et al. found the opposite pattern: overall, higher ratings for *ja* compared with *nein*. A closer inspection of the data revealed differences among participants. The majority of participants showed a preference for *ja* over *nein*. Yet, a notable minority displayed a preference for *nein* over *ja*. One possible account for the opposite preference patterns is that they reflect two different response systems (e.g. truth-based vs. polarity-based). Yet, what casts some doubt on this account is that the overall difference between the ratings for *ja* and *nein* was rather small and that both particles received mostly rather high ratings. However, the method may have been crucial for this finding. When judging acceptability, people might allow for variation that they are familiar with, e.g. different usage preferences. Thus, participants might have judged a particle response that does not correspond to their response system as quite acceptable because they are acquainted with the usage pattern resulting from the alternative response system.

The goal of the **present study** was to gain further insight into preference patterns for *ja* and *nein* in affirming responses to negative utterances by using different methods. Experiment 1 tapped into usage preferences, with the following rationale: if there are indeed two different response systems, then a production-like task should reveal unequivocal usage preferences. Experiment 2 tapped into interpretation preferences. If speakers of German are indeed familiar with two different usage patterns (possibly reflecting different response systems), then they should take into account both patterns when interpreting bare particles.

In **Exp 1**, participants (*n*=43) were presented with 48 short dialogues (see Table 1). The dialogues consisted in a (negative or positive) assertion and a response, comprising a response particle and a (negative or positive) full response clause. In all target items (*n*=12), the polarity of both the assertion and the response clause was negative. Participants were presented

Der Gärtner hat den Rasen noch nicht gesät. 'The gardener hasn't sown the lawn yet.'
□ Ja, er hat den Rasen noch nicht gesät.
□ Nein, er hat den Rasen noch nicht gesät.
□ Doch, er hat den Rasen noch nicht gesät.
'Ja/Nein/Doch, he hasn't sown the lawn yet.'

Table 1: Sample of the dialogues of Exp 1

with three versions of the response, differing only in the response particle. Their task was to indicate which of the responses they themselves would use given they had the knowledge conveyed by the response clause. They were explicitly allowed to choose more than one of the response options. For the target items, i.e. affirming responses to negative assertions, there was a clear overall pattern of response choices: 56% *ja*-response, 27% *nein*-response, 17% choice of both *ja*- and *nein*-response ($p < .001$). An inspection of the individual choice patterns revealed that the majority of participants, approx. 50%, had a clear preference for *ja*. However, approx. 20% showed a clear preference for *nein* and approx. 15% chose the *ja*-response and the *nein*-response about equally often. The choice patterns of the remaining participants are inconclusive. Taken together, the findings from Exp 1 are consistent with the findings by Claus et al. and extend them by revealing more clear-cut preference patterns.

Exp 2 employed an interpretation choice task. Participants (n=45) were presented with modified versions of the 48 dialogues from Exp 1, i.e. the three response versions were replaced by one single response consisting of a bare particle (*ja*, *nein*, or *doch*). In all target items (n=16), the assertion was negative and the response particle was either *ja* or *nein*. The participants were asked to indicate how they interpreted the bare response particle by choosing one of three options (see Table 2). They were instructed to choose the third option if they were not sure how to interpret the response particle.

□ The gardener hasn't sown the lawn yet. (affirming response)
□ The gardener has sown the lawn already. (rejecting response)
□ Without additional information it is not clear to me, what the response means.

Table 2: Sample of the choice options in Exp 2, translated from German. The information in parentheses was not presented to the participants.

The choice pattern for *ja* differed from the choice pattern for *nein* ($p < .001$). However, both response particles were interpreted as affirming responses in the vast majority of cases (*ja*: 93.1%, *nein*: 84.7%). A closer data inspection revealed that most participants (≈ 80%) interpreted both *ja* and *nein* consistently as affirming responses. Only a small number of participants showed a clear difference between the two particles: Six participants interpreted *ja* as affirming and *nein* as unclear or rejecting, and three interpreted *nein* as affirming and *ja* as unclear. The finding that most participants did not show a clear distinction between *ja* and *nein* is consistent with the assumption that speakers of German are acquainted with two different usage patterns of response particles and take this variation into account when interpreting bare particles.

Conclusion: The present findings, stemming from usage preference data and interpretation data, together with the previous findings (Claus et al.), stemming from acceptability data, indicate that there are two different usage preferences for *ja* vs. *nein* in affirming responses to negative antecedents and that speakers of German are familiar with them. The two usage preferences possibly reflect two different response systems. In the framework proposed by Roelofsen & Farkas, the two systems could be characterized in terms of truth-based vs. polarity-based strategies. In Krifka's account, the two different usage preferences would rather be attributed to differences in negation processing.

References

Claus, Meijer, Repp & Krifka (accepted pending revisions). Puzzling response particles: An experimental study on the German answering system. *Semantics & Pragmatics*. | Jones (1999). *The Welsh answering system*. Berlin: de Gruyter. | Krifka (2013). Response particles as propositional anaphors. In *Proceedings of SALT 23*. | Roelofsen & Farkas (2015). Polarity particle responses as a window onto the interpretation of questions and assertions. *Language* 91. 359-414.

Ellipsis in context: The interaction of identity and discourse salience

Jeffrey Geiger & Ming Xiang (University of Chicago)

The possibility of exophoric (antecedentless) verb phrase ellipsis (VPE) [1-2, i.a.] presents a challenge to traditional accounts of VPE based solely on linguistic identity [3-8, i.a.]. Whether salient nonlinguistic and linguistic information have the same status in VPE interpretation [1] or nonlinguistic information is only incorporated via accommodation [2] has not been resolved. In two experiments, we examine the interaction between discourse salience and linguistic identity. In Experiment 1, we show that salient nonlinguistic information can be recruited to (re)construct an antecedent for a VPE site, even in the presence of an overt antecedent. Experiment 2, however, shows that linguistic identity supersedes discourse salience as a locus for antecedent construction. Our results support a model in which linguistic information contributes more strongly to VPE interpretation than nonlinguistic information, which can affect interpretation via accommodation.

Experiment 1 (subj n=146 AMT workers) Each trial featured a nonlinguistic context presented as a comic strip and a simultaneously presented text dialogue between two characters. In the example in Table 1, the father always uttered the VPE Reply as a response to the son's request. The *Linguistic Antecedent* (son's utterance) could be absent (Exophoric), present with no numeral (Unmodified), or present with a numeral (Modified). The *Nonlinguistic Context* (comic strip) made no reference at all to the numerosity of the referent (Unavailable), made the numerosity recoverable but not salient (Available), or made the numerosity highly salient (Salient). These manipulations created 9 (3x3) conditions. Finally, for each of the 9 conditions, the *VPE Interpretation* question solicited ratings (on a 1-7 scale, also shown simultaneously) for the VPE to be interpreted as not containing a numeral modification ("buy candy bars"; Unmodified Interpretation) or containing a numeral modification ("buy five candy bars"; Modified Interpretation) (9x2=18 conditions total). There were 6 critical trials and 10 fillers. The goal was to examine whether and how the salience of the numeral information supplied by the nonlinguistic context can modulate VPE interpretations.

Figure 1 shows the results. For the Exophoric conditions, paired comparisons showed that the numeral-modified VPE interpretation increased its rating as a function of the increased salience of the numeral information in the Nonlinguistic Context (p's $<.05$), confirming that a VPE antecedent can be reconstructed from salient nonlinguistic context. When there is an overt linguistic antecedent (non-exophoric conditions), the VPE interpretation that is identical to the antecedent is always rated higher than the non-identical one (p's $<.001$). However, Salient numeral information from the nonlinguistic context boosted the rating for the Modified Interpretation when the linguistic antecedent was Unmodified (p's $<.01$), suggesting that salient nonlinguistic information can be used to enrich the linguistic antecedent, albeit in a restricted manner.

Experiment 1 showed that salient discourse information dominates VPE interpretation when there is no explicit linguistic antecedent, but otherwise linguistic identity is preferred for VPE interpretation. One explanation of these findings is to assume VPE is resolved around a salient question under discussion (QUD), which is supplied either by salient discourse information in the nonlinguistic environment or by an explicit linguistic antecedent, assuming that an uttered antecedent is automatically more salient than the implicit discourse information in the environment. An alternative account is to acknowledge VPE resolution under identity and accommodation of a structure reflecting salient non-linguistic information as two separate but interacting mechanisms. Experiment 2 aims to tease these two accounts apart.

Experiment 2 (n=164) shared the same design as Experiment 1 except: the VPE Reply utterance was replaced with the complete unmodified or modified VPE Interpretation (e.g., "We can't buy any candy bars."), and the subjects provided a 1 to 7 rating of how *coherent* they thought the Reply was given the prior information. Assuming that the exchange is coherent only when the reply properly addresses the QUD raised by the previous context and/or utterance, the coherence rating task tracks what QUDs can be raised by salient linguistic and nonlinguistic contexts. If the unified QUD account can completely explain the results from Experiment 1, we should expect the results from Experiment 2 to closely track those from Experiment 1. This prediction is largely borne out (Figure 2). However, an important finding is that the coherence ratings of the two types of Replies are not significantly different with an Unmodified Antecedent and Salient Context ($p>.4$). This shows that the linguistic antecedent does not contribute more strongly to the QUD than the non-

linguistic context does, so the QUD account cannot explain why the antecedent-identical reading is preferred in Experiment 1. Linguistic identity plays a larger role in VPE interpretation than it is implicitly granted in the QUD account.

Conclusion: In two experiments, we showed that VPE interpretation considers both linguistic and nonlinguistic information, but that discourse salience is subordinate to linguistic identity as a locus of ellipsis resolution. Ellipsis resolution based solely on a discourse-salient QUD is not supported. The results support a model of ellipsis interpretation in which resolution under identity is dominant, but a new structure reflecting salient nonlinguistic information can be accommodated.

Table 1: Factors & levels for Experiment 1

Nonlinguistic Context	Antecedent	Reply	VPE Interpretation
Unavailable: Father and son stand in grocery store aisle near candy bars.	*Exophoric:* [no antecedent]	Father: We can't.	*Unmodified:* On a scale from 1 to 7, where 1 is the least likely and 7 is the most likely, how likely do you think it is that the father meant: We can't buy any candy bars.
Available: Son places five candy bars in cart at one time.	*Unmodified:* Son: I want to buy candy bars!		
Salient: Son conspicuously places five candy bars in cart one at a time.	*Modified:* Son: I want to buy five candy bars!		*Modified:* ...We can't buy five candy bars, but maybe we could buy fewer.

Figure 1: Experiment 1 results. Horizontal split: Antecedent type. Horizontal axis: Nonlinguistic Context type. Vertical axis: Mean likelihood rating. Error bars: Standard error.

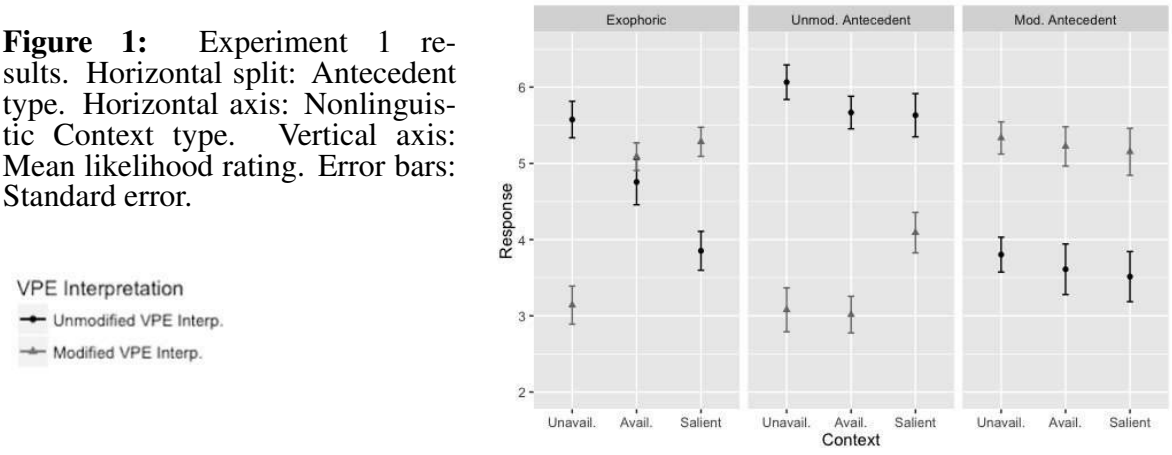
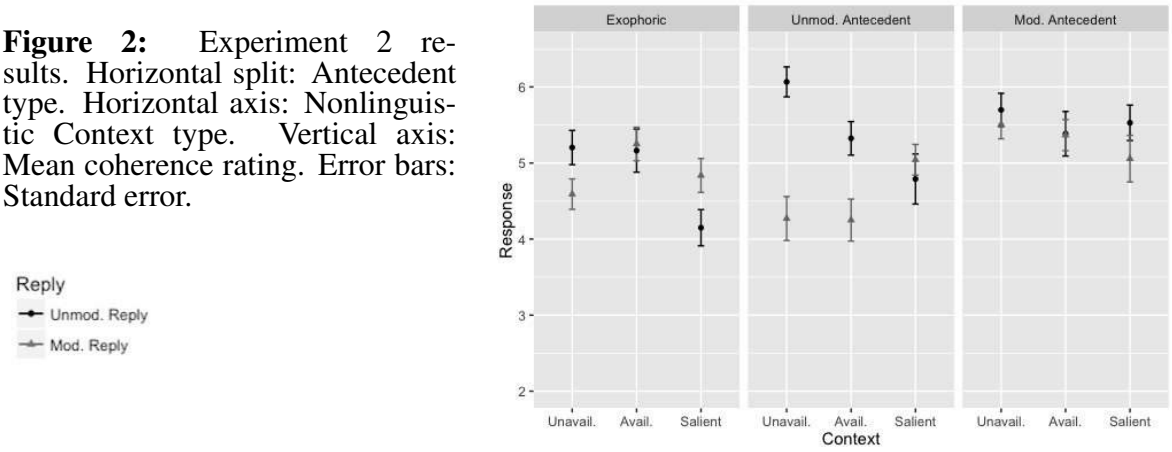


Figure 2: Experiment 2 results. Horizontal split: Antecedent type. Horizontal axis: Nonlinguistic Context type. Vertical axis: Mean coherence rating. Error bars: Standard error.



References: [1] Miller & Pullum (2013); [2] Merchant (2004); [3] Hankamer & Sag (1976); [4] Fiengo & May (1994); [5] Chung, et al. (1995); [6] Dalrymple, et al. (1991); [7] Hardt (1993); [8] Merchant (2001)

Do German demonstrative pronouns avoid prominent perspectival centers?

Stefan Hinterwimmer (*University of Cologne*)*, Umesh Patil* (*University of Osnabrueck), Andreas Brocher* (*University of Cologne)*

German demonstrative pronouns of the *der*/*die*/*das*-paradigm (DPros) are known to avoid prominent discourse referents as antecedents or binders. Prominence is usually defined in terms of the notions subjecthood, topicality, and agentivity (see Bosch et al., 2007; Hinterwimmer, 2015; Schumacher et al., 2016). Recently, Hinterwimmer and Bosch (2016) (HB) argued that individuals functioning as *perspectival centers* are also prominent and thus avoided by DPros, where they define perspectival centers as follows: An individual *x* is the perspectival center with respect to a proposition *p* if *p* expresses the content of an utterance or thought of *x*. This assumption automatically accounts for the observation that DPros can typically not be bound by subjects of propositional attitude verbs (Wiltschko, 1997). HB claim, however, that subjects of propositional attitude verbs can bind DPros that are contained in the respective complement clause, provided that there is a more prominent perspectival center available.

HB consider two scenarios where this is the case: First, a sentence with a propositional attitude verb is itself the complement clause of a higher propositional attitude verb. Since the subject of the higher propositional attitude verb is a more prominent perspectival center than the subject of the lower propositional attitude verb, binding by the latter but not the former should be possible. Second, a speaker uses an epithet to refer to the subject of a propositional attitude verb and thereby makes her own perspective maximally prominent. Since the subject of the propositional attitude verb is consequently the less prominent perspectival center, it should be able to bind a DPro contained in the respective complement clause.

(1) a. Klaus behauptet, dass Lisa denkt, dass dessen Ferrari eine sinnvolle Investition war.
b. Lisa behauptet, dass Klaus denkt, dass dessen Ferrari eine sinnvolle Investition war.
Klaus/Lisa claims that L./K. thinks that his (DPro) Ferrari was a sensible investment.

We conducted three experiments to test the predictions of HB. The first two studies were sentence reading, eye-tracking experiments in which we tested the first prediction. In Experiment 1, 24 participants read 16 single sentences like the ones in (1a-b). Eight sentences were of the kind (1a, *dispreferred*) and eight of the kind (1b, *preferred*). Thus, in the *dispreferred* condition (1a), the matrix subject was male, agreeing in gender with the DPro, and the embedded subject was female. In the *preferred* condition, the matrix subject was female and the embedded subject male (which then agreed in gender with the DPro). Regions of interest were the noun phrase headed by the DPro *dessen* and including a noun (e.g., *dessen Ferrari*) as well as the immediately following word (spillover). These regions were identical between conditions. Eighty filler sentences were interspersed with the experimental sentences. Yes/no comprehension questions appeared after 25% of items (overall accuracy was 96%).

Analysis of log-transformed reading times showed that, compared to the *preferred* condition, readers slowed down in the noun phrase region of the *dispreferred* condition. This was evident for gaze durations, $t(22) = 2.17$, $p = .041$, and total reading times, $t(22) = 3.14$, $p = .005$. Interestingly, the observed reading slow-down persisted for the spillover word (gaze durations:

$t(22) = 2.03$, $p = .055$; total times: $t(22) = 1.91$, $p = .069$). Taken together, these data suggest that DPros can more easily be interpreted as bound by the subject of embedded propositional attitude verbs than by the subject of matrix propositional attitude verbs. Now, in order to ensure that the observed contrasts were not due to recency effects, we conducted a second experiment, which was very similar in materials and design to Experiment 1. In Experiment 2, we varied the pronoun that headed the critical noun phrase, while holding binder-bindee distance constant. In both the *preferred* and *dispreferred* conditions, the matrix subject was male and the embedded subject female. However, in the *preferred* condition (2b), the pronoun of interest was a personal pronoun (PPro) *sein*, while in the *dispreferred* condition (2a) it was the DPro *dessen*. Other than pronoun type, sentences were again identical between conditions. After having run 13 out of the 24 participants, we already see a reliable reading slow-down for noun phrases in the *dispreferred* compared to the *preferred* condition, and this slow-down shows up in gaze durations, $t(12) = 2.49$, $p = .028$, regression path times, $t(12) = 3.93$, $p = .002$, and total reading times, $t(12) = 2.41$, $p = .033$. These effects spilled over to the subsequent word for gaze durations, $t(12) = 1.91$, $p = .079$. These data confirm that the contrasts in Experiment 1 were not due to recency.

(2) a. Klaus behauptet, dass Lisa denkt, dass dessen Ferrari eine sinnvolle Investition war.
b. Klaus behauptet, dass Lisa denkt, dass sein Ferrari eine sinnvolle Investition war.
Klaus claims that Lisa thinks that his (DPro/PPro) Ferrari was a sensible investment.

Addressing the second hypothesis, we conducted an eye-tracking study using a visual-world paradigm. We used short discourses, as the one in (3), and manipulated the "perspectival center-hood" of the topic of the current discourse topic (R1 = *der Polizist*): It was either mentioned again by a PPro (*er* in (3a)) or by an epithet (*der nette Wachtmeister* in (3b)). The discourse also introduced another human masculine referent (R2 = *der Fotografen*) as well as two non-human referents as distractors. The DPro, *der*, occurred in the complement clause of the third sentence. The display showed these four referents together with an unmentioned distractor object. Results show that R1 was less preferred than R2 in terms of focussing frequencies. But, in the epithet condition, R1 was reliably more preferred than in the PPro condition. In sum, then, R2 seems to be generally preferred as binder of a DPro, since it is less prominent than R1 in terms of discourse topicality, subjecthood, and agentivity in both conditions and because it is not the perspectival center with respect to the proposition denoted by the entire sentence or the one denoted by the embedded clause. At the same time, the fact that being referred to by an epithet boosts availability of R2 as binder of the DPro despite being maximally prominent with respect to discourse topicality, subjecthood, and agentivity, supports the claims made in HB.

(3) Sentences 1 and 2 (same in both conditions): Eine gute Nachricht. Der Polizist hat gerade das Motorrad abgestellt und redet mit dem Fotografen.
Good news. The policeman has just parked the motorcycle and talks to the photographer.
Sentence 3 prelude:
(a) Er erzählt soeben dem Fotografen, der eigentlich wegen der Kängurus hier ist,…
(b) Der nette Wachtmeister erzählt soeben dem Fotografen, der eigentlich…
Sentence 3 postlude (same in both conditions): *dass der im Lotto gewonnen hat.*
He/the_nice_sergeant has just told the photographer who is here because of the kangaroos that DPro has won the lottery.

References: Bosch, P. et al. (2007). The non-subject bias of German demonstrative pronouns. Hinterwimmer, S. & Bosch, P. (2016). Demonstrative pronouns and perspective. Hinterwimmer, S. & Bosch, P. (to appear). Demonstrative pronouns and propositional attitudes. Schumacher, P. et al (2016). Thematic role as prominence cue during pronoun resolution in German.

Turn-timing and the body:
Gestures play a core role in coordinating conversation

Judith Holler, Kobin H. Kendrick, & Stephen C. Levinson

Conversation is the core niche of human language use and it is based on a turn-taking system. How we coordinate who says what and when is a significant pragmatic and psycholinguistic challenge. This becomes particularly evident when we consider that conversational turn-taking is remarkably fast, with gaps between speaking turns averaging around just 200 ms (Stivers et al., 2009). Considering that the production of single word utterances takes a minimum of 600 ms alone (Indefrey & Levelt, 2004), language production and comprehension must largely run in parallel; that is, while listening to an on-going turn, a next speaker has to predict the upcoming content, understand the speech act, and start preparing their own turn to be able to launch it on time (Levinson, 2013, 2016).

Considering that the primordial site of conversation is face-to-face social interaction where participants do not just speak but make use of a host of visual signals to communicate, a fundamental question arises: what is the role of the body in the coordination of speaking turns in conversation? In order to investigate this question, we carried out two studies, one quantitative analysis of multimodal conversational corpus data, and one based on spontaneous conversation combined with experimentally manipulating the availability of bodily signals.

For study 1, we analyzed a corpus of 7 casual face-to-face conversations between English speakers by identifying all question-response sequences (N=281), as well as the gestures that accompanied the identified set of questions, and the timing of these gestures with respect to the speaking turns they accompanied. Moreover, we measured the length of all inter-turn gaps in our set. To gain a first insight into whether gestures contribute to conversational coordination we asked whether the length of the gap between turns varied systematically as a consequence of questions being accompanied by gesture. Our results revealed that this is indeed the case: Questions with a gestural component were responded to significantly faster than questions without a gestural component. This finding holds when we consider head and hand gestures separately, when we control for points of possible turn completion in the verbal utterance prior to turn end, and when we control for the complexity associated with question type. Furthermore, our findings revealed a second, independent effect; namely, even within the group of questions accompanied by gestures, those questions whose gestures retracted prior to turn end were responded to faster than questions whose gestures retracted following turn end.

Study 2 is based on conversations that involved a within-participants manipulation: 20 dyads talked while they were able to see one another as well as while they were not. As for study 1, we measured the gaps between turns and compared the face-to-face with the no vision condition. The findings are in line with those from study 1 in that gaps between turns are shorter when interlocutors have bodily signals at their disposal, thus suggesting that bodily signals play important coordinative functions. A further experimental study is currently underway testing which types of gestures and other bodily signals facilitate early responding to speaking turns and what

mechanisms lie beneath this effect. Results are expected in time for the conference and will further elucidate the issue at hand.

In sum, the two studies suggest that the body plays an important role in the coordination of face-to-face conversation. Rather than burdening our cognitive system, gestures i) facilitate language processing, even in the rich and cognitively challenging context of conversational interaction, and ii) they seem to play a role also in the prediction of upcoming turn ends. Both of these contributions appear to contribute to interlocutors being able to respond fast in face-to-face conversation. The findings suggest an urgent need for adapting existing turn-taking models that focus primarily on the verbal modality (Sacks et al., 1974).

References

Indefrey, P., and Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition, 92,* 101–144.

Levinson, S. C. (2013). Action formation and ascription. Ins J. Sidnell and T. Stivers (Eds.), *The Handbook of Conversation Analysis*, pp. 101–130. Malden, MA:Wiley-Blackwell.

Levinson, S. C. (2016). Turn-taking in human communication, origins, and implications for language processing. *Trends in Cognitive Sciences, 20,* 6-14.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50,* 696–735.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. U.S.A., 106,* 10587.

Contextual effects on the processing of Hungarian pre-verbal focus sentences: an eye-tracking study

Tamás, Káldi and Anna, Babarczy – Research Institute for Linguistics (HAS)

The interpretational characteristics of the Hungarian pre-verbal focus (preVf) has been subject to heated debate for a long time. While the view that preVf tends to have an exhaustive interpretation (exhaustivity henceforth) is not generally questioned, the status of exhaustivity is. The current study uses eye-tracking to investigate the effects of linguistic context on the processing of preVf. Specifically, we examine the hypothesis that exhaustivity emerges as a result of implicature generation.

An array of empirical studies revealed that exhaustivity is variable, emerges late in processing and may be context dependent, and thus concluded that exhaustivity has implicature status (Onea & Beaver 2011, Kas & Lukács 2013, Gerőcs et al, 2014). Furthermore, based on eye-tracking data Káldi et al (2016) claim that exhaustivity has the status of *scalar* implicature, where the exhaustive reading corresponds to the upper-bounded, whereas non-exhaustive reading to the lower bounded interpretation. Káldi et al (2016) compared the interpretational characteristics of preVf and lexically marked focus sentences (*only*-f, henceforth) in two visual-world experiments: one experiment used a forced-choice sentence-picture matching task while the other one posed no restriction on the number of images that could be chosen. The results revealed that eye-gaze converged on the exhaustive images at the same rate in both sentence conditions in the forced choice task, whereas in the multiple choice task fixation patterns showed hesitation between the exhaustive and non-exhaustive images in the preVf sentence condition relative to the *only*-f sentence condition in the post-verb period (starting approx. 1000ms after the focused element). One important limitation of this study, however, is that the context-dependence of exhaustivity could only be inferred indirectly, as it emerged as a result of the manipulation of the experimental task, and not because of the linguistic context in which preVf (and *only*-f) sentences had to be interpreted.

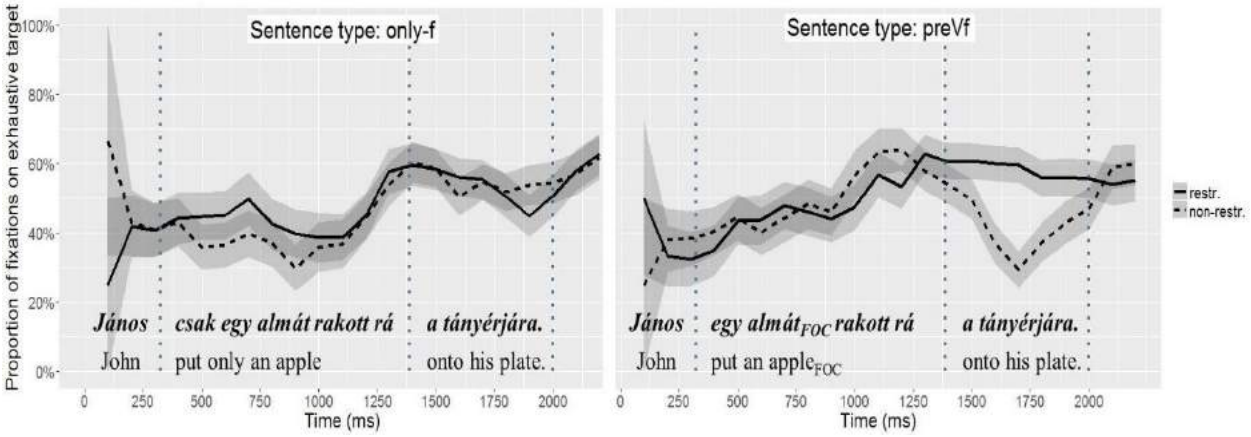
To overcome this limitation, we have conducted a visual-world experiment in which we introduced a direct manipulation of the linguistic context. Each critical trial included three sentences (see table): the Introductory sentence listed the possible referents in the universe of discourse, the second sentence either restricted the number of choices (restrictive condition) or not (non-restrictive condition), and the third, Test sentence contained either an *only*-focus (*only*-f condition) or a preVf sentence (preVf condition).

Intro.	*Az asztalon volt egy tál tele gyümölcsökkel. Volt benne egy csomó alma, körte, barack.* There was a bowl full of fruit on the table. There were a lot of apples, pears, peaches.
Cont.	*Minden vendég rakhatott a tányérjára ezek közül EGYET/NÉHÁNYAT.* Every guest could put ONE/SOME fruit(s) onto their plates.
Test	*János (csak) egy almát rakott rá a tányérjára.* John put (only) an apple onto his plate.

In each trial participants listened to the three sentences consecutively, and were shown a set of four images simultaneously with the last sentence. The set of four images contained an exhaustive target (e.g. an apple), a non-exhaustive target (an apple & a pear), and an exhaustive and a non-exhaustive distractor. The experimental task was to choose the image or images that best corresponded to the linguistic stimuli. 21 adult native Hungarians participated in the experiment. We

measured looks to the four images while participants heard the Test sentences. As the emphasis of the current study is the processing related differences between the two types of focus construction, we predicted that in the case of *only*-f sentences we will not see a difference in the proportion of looks to the exhaustive target in the two context conditions, as lexically marked focus should be insensitive to such variation. In the case of preVf sentences, however, we expected eye-gaze patterns to differ between the restrictive and the non-restrictive context conditions in the post-verb Interest Period (IP) (approx. 1000ms post focused NP onset). Based on Huang & Snedeker (2009) and Káldi et al (2016) divergence in this IP can be regarded as a correlate of implicature generation.

As predicted, the results of the experiment revealed no divergence in the proportion of fixations on the exhaustive target image in the two context conditions for the *only*-f sentences, whereas the proportion of fixations indicates a high degree of hesitation in the case of preVf sentences in the post-verb IP in the non-restrictive context condition relative to the restrictive condition (Context x Sentence Type x IP interaction ($F(2, 40) = 4.19$, $p = .02$) in GLM with all three variables as repeated measures factors). Based on the results we conclude that processes related to the exhaustive interpretation of preVf sentences are context dependent. Additionally, we consider these results as further evidence supporting the pragmatic status of the exhaustive interpretation of preVf sentences.



References

Gerőcs Mátyás, Babarczy Anna & Surányi Balázs 2014. Exhaustivity in Focus: Experimental Evidence from Hungarian. In: Joseph Emonds & Markéta Janebová (eds.): Language Use and Linguistic Structure. Olomouc: Palacky University. 181–94.

Kas Bence & Lukács Ágnes 2013. Focus sensitivity in Hungarian adults and children. Acta Linguistica Hungarica 60(2): 217–45.

Káldi Tamás & Babarczy Anna 2016. A magyar fókusz és a skaláris implikatúrák: Egy szemmozgáskövetéses kutatás eredményei. In: Kas Bence (ed.): *„Szavad ne feledd!"* 190 – *Tanulmányok Bánréti Zoltán tiszteletére.* Budapest: MTA Nyelvtudományi Intézet. 333–46.

Huang, Y., & Snedeker, J. 2009. On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. Cognitive Psychology, 58, 376-415.

Onea, Edgar & Beaver, David 2011. Hungarian focus is not exhausted. In: Ed Cormany, Satoshi Ito & David Lutz (eds.): Proceedings of the 19th Semantics and Linguistic Theory Conference. eLanguage. 342–59.

Alternatives in processing ad-hoc implicatures

Verena Keite, Ralf Klabunde and Eva Belke
Department of Linguistics, Ruhr University Bochum

Considering alternatives is one of the fundamental tasks in pragmatic reasoning. The generation of implicatures requires hearers to reason about alternative expressions the speaker could have uttered but did not. Recent work by van Tiel & Schaeken (2016) suggests that differences between the alternatives of different types of implicatures critically influence their processing: Scalar implicatures were associated with a processing cost, while free choice inferences, conditional perfection, and exhaustivity in "it"-clefts caused no delay in reaction times (see also Chemla & Bott (2014) who find no processing costs for free choice inferences).

Based on Katzir's (2007) formal, structurally oriented approach to scalar implicatures that explains scalar behavior by structural complexity, van Tiel & Schaeken (2016) and Chemla & Bott (2014) argue that scalar implicatures differ from other kinds of inferences in that in constructing the alternatives, the scalar term needs to be substituted with a more informative one. According to this view, these lexical processes are what makes processing scalar implicatures effortful, not constructing and reasoning about alternatives as such.

However, this account is unlikely to apply to ad-hoc scalar implicatures as these are based on alternatives that arise from the context and therefore do not require lexical access. In the present study, we studied effects of the alternatives' complexity on ad-hoc implicatures. Since Katzir's (2007) notion of complexity is confined to structural complexity and remains rather vague, we based our study on Hirschberg's (1985) more detailed account on alternatives as sets with a partial order defined on them (posets). We hypothesized that retrieving alternatives from the context and constructing posets of alternatives in processing ad-hoc implicatures is effortful. Accordingly, easily accessible posets of alternatives will facilitate the generation of ad-hoc implicatures compared to more difficult ones. The present study was designed to test this prediction.

We devised an experimental paradigm that required participants to initially learn three pseudowords and their meanings (training phase). This way, we made participants establish different contexts and different sets of alternatives, respectively; i.e., we used the experimental task to make the alternatives available that need to be considered in the processing of scalar implicatures (see Degen & Tanenhaus 2016). In the experimental phase, participants' generation of ad-hoc implicatures was tested in a picture selection task (see Degen, Franke, & Jäger (2013) and Stiller, Goodman, & Frank (2015) for similar paradigms).

Two different posets of alternatives were examined in a between participants design: In the partial poset condition, participants learnt the pseudowords OSIM, EGAT, and ULOS with the following meanings: 'a monster that has arms', 'a monster that has legs', and 'a monster that has horns'. Here, we expected them to learn words for three out of the eight sets included in the power set P_1 ({arms, legs, horns}). In the full poset condition, participants learnt the same pseudowords, but they now referred to arms, legs and the conjunction of arms and legs,

covering the whole power set P_2 ({arms, legs}) except for the empty set. In both conditions, participants were trained on the meanings of the pseudowords in 12 unambiguous trials.

For the experimental phase nine trials were constructed per condition. In each trial three pictures of monsters were presented, featuring a set of features from the power set P_1 (partial poset condition) or from the power set P_2 (full poset condition), respectively. The three pictures corresponded to three referents: One of the monsters only had the feature in question and no other feature from the relevant poset. If the participant generated the ad-hoc implicature, she selected this monster (implicature target). Another monster had more than one feature, i.e. it was compatible with the literal meaning of the pseudoword (logic target). The third monster did not have the feature in question (distractor).

With this design, we assessed how the accessibility of sets of alternatives affects the generation of ad-hoc implicatures of 180 native speakers of English recruited from Amazons Mechanical Turk. In target choice trials, participants performed above chance level in both conditions, selecting the implicature target in 90.74% of the experimental trials in the partial poset condition and in 95.63% of the experimental trials in the full poset condition.

In a logistic mixed-effects model predicting implicature target choice as a function of condition (partial poset/full poset) with random effects of participants and items, the fixed effect of condition was significant. We compared the full model to a model without the effect in question. The AIC was lower for the full model indicating that the model with condition as a predictor fits the data better. However, the BIC was lower for the null model, suggesting that the difference between conditions was subtle.

We take our findings to suggest that the generation of ad-hoc implicatures depends on the accessability of the relevant poset of alternatives. In the full poset condition, participants learnt pseudowords for all sets of the relevant poset (except for the empty set) and saw stimuli from this poset, facilitating the retrieval and construction of the relevant poset of alternatives. As a result, they generated more ad-hoc implicatures than participants in the partial poset condition. In future research, we will investigate further how the properties of alternatives (e.g. the set size, the salience of alternatives) affect the generation of scalar implicatures.

References:

Chemla, E., & Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition*, *130*(3), 380-396.

Degen, J., Franke, M., & Jäger, G. (2013). Cost-Based Pragmatic Inference about Referential Expressions. *Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci'13)*, 376–281.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, *40*(1), 172-201.

Hirschberg, J. (1985) *A Theory of Scalar Implicature.* Univ. of Pennsylvania, Tech. Rep. MS-CIS-85-86.

Katzir, R. (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, *30*(6), 669-690.

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

van Tiel, B., & Schaeken, W. (2016). Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive Science*.

Visual contrast, discourse contrast and conceptual convention

Christina S. Kim and Louisa Salhi (University of Kent)

How standards of comparison for gradable adjectives like *large* or *small* are established based on contextual information has been an active area of research [1-5], but much of this work sets aside the question of what features of the context comprehenders rely on to determine an appropriate comparison class, and how different cues are weighed against each other. The present study uses the Visual World Paradigm [6] to investigate how comprehenders integrate visual and discourse context to referentially disambiguate an expression like *small square*. It builds on [7], which showed that a contrast set in the visual context (tall glass, short glass) increased listeners' expectation that a contrast set member would be described using a modifier (*tall*) due to the need to disambiguate from the other contrast set member, even when e.g. a taller object in terms of absolute height (pitcher) was also present in the display (visual contrast effect). This study extends this paradigm to include the prior discourse as an additional source of contrast.

Experiment1 asks to what extent contrast across discourse functions like visual contrast to aid referential disambiguation, and how discourse and visual contrast are integrated when both are present and provide conflicting cues to contrast. Participants listened to pairs of sentences like (1) accompanied by pairs of displays like (2). Target type (whether the target word, *square* in (1), was a part of a discourse or visual contrast set) was crossed with the presence of an additional contrast set (discourse contrast if the target a visual contrast set member, and vice versa). The visual contrast effect [7] was replicated: when a contrast set was in the visual context, fixations converged on the target referent in the 200-100ms preceding the onset of the target word (t=3.02, p <.0001).

There was also evidence that discourse contrast has a similar facilitative effect on resolving reference: when both visual and discourse contrast were present, the discourse contrast set member competed with the visual contrast set member, as indicated by later convergence on the target referent when both sources of contrast were present (100-200ms post target onset for discourse contrast targets, t=2.15, p <.05; 300-400ms post target onset for visual contrast targets, t=2.23, p <.05), than when only one contrast was present (300-400ms post target onset for discourse contrast targets, t=2.09, p <.05; see above for visual contrast targets). Target and competitor fixations were fit with mixed-effects logistic regression models in analysis windows aligned to linguistically-determined events (pre-adjective, adjective-to-target, post-target), with Target type (discourse, visual contrast), Number of contrast sets (one, two), Time, and their interactions as predictors. There were more competitor fixations for two-contrast than one-contrast conditions in the adjective-to-target (β=.042, SE=.0013, p <.0001) and post-target windows (β=.071, SE=.0027, p <.0001). However, discourse contrast was a less salient cue than visual contrast. Discourse contrast competitors gave rise to a weaker competitor effect than visual contrast competitors: there was a larger competitor advantage for visual contrast competitors in both adjective-to-target (β=.034, SE=7.25e^{-6}, p <.0001) and post-target windows (β=.047, SE=.0014, p <.0001). In addition, comprehenders recovered faster from discourse contrast competitors (100-200ms post target, t=2.15, p <.05) than from visual contrast competitors (300-400ms post target, t=2.23, p <.05).

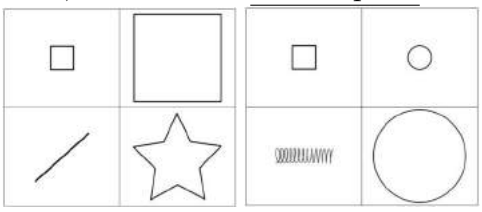
While discourse contrast appeared to be a weaker cue than visual contrast in Exp1, in conversation, discourse contrast is often far more salient than visual contrast, simply because many conversations are not about the visual environment. In addition, richer discourses have

additional internal dependencies that the pairs of sentences in Exp1 do not. For instance, coherence relations [8] and Question Under Discussion structure [9] have been shown to influence online discourse interpretation [10-11], as has prosodically marked contrast [12]. The conceptual pact literature [13-15] also suggests that how something has been referred to in prior discourse influences how a listener expects the same item to be referred to subsequently. **Experiment2** asks whether prior experience describing classes of items in a particular way modulates the strength of discourse or visual cues to contrast.

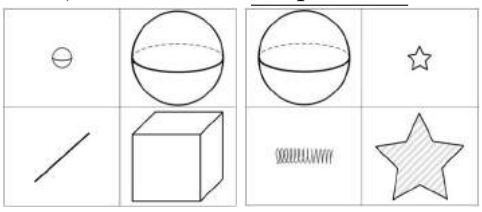
Exp2 differed from Exp1 in two respects. First, test trials were preceded by a training block in which participants categorized one class of objects (2D shapes, e.g. square, circle) in terms of size (*large*, *small*), and another (3D shapes, e.g. cube, sphere) by whether they were *striped* or *solid*. Second, in addition to the four conditions from Exp1, the test block included two-context conditions where the competitor contrast item was from a different training category than the target, as in (3-4). If prior experience associating different category members with particular modifiers leads to expectations that the same conventions will continue to be followed, different category competitors (whether discourse or visual contrast) should be weaker competitors to the target referent than same category competitors.

To assess same v. different category competitor effects, target and competitor fixations from the two-contrast conditions were fit with mixed-effects regression models using the same analysis windows as for Exp1, with Target type (discourse, visual contrast), Competitor type (same, different category contrast), Time, and their interactions as predictors. There were more competitor fixations for same-category than different-category competitors in the adjective-to-target (β =.053, SE=.0034, p <.0001) and post-target windows (β =.053, SE=.0034, p <.0001), suggesting that unexpected modifier-category pairings were weaker competitors with target referents than expected ones. However, within different category conditions, comprehenders recovered more slowly from discourse contrast competitors (600-700ms post target onset, t =2.72,p <.01) than from visual contrast competitors (convergence on target 200-300ms post target onset, t =2.43,p <.05). The strong discourse competitor effect may be because this is the only condition that requires comprehenders to shift from one dimension of modification (e.g. *small/large*) to another (e.g. *striped/solid*) within a discourse (3-4); this suggests comprehenders may expect that, regardless of category-specific modification history, speakers will modify discourse referents in consistent ways.

(1) Click on the large square.
Now, click on the small square.

(2) 

(3) Click on the small sphere.
Now, click on the striped star.

(4) 

References [1] Klein 1980. *LP*. [2] von Stechow 1984. *JoS*. [3] Graff 2000. *Phil Topics*. [4] Kennedy & McNally 2005. *Language*. [5] Sassoon & Zevakhina 2012. *SALT22*. [6] Tanenhaus, et al. 1995. *Science*. [7] Sedivy, et al. 1999. *Cognition*. [8] Kehler 2002. CSLI. [9] Roberts 1996. OSUWPL. [10] Rohde et al. 2011. *Cognition*. [11] Clifton & Frazier 2012. *CogPsych*. [12] Fraundorf et al. 2010. *JML*. [13] Brennan & Clark 1996. *JEP:LMC*. [14] Metzing & Brennan 2003. *JML*. [15] Yoon & Brown-Schmidt 2013. *JML*.

Listeners encode multiple meanings when generating scalar inferences

Alix Kowalski and Yi Ting Huang
University of Maryland, College Park

During spoken-language comprehension, each sentence's interpretation is built on a moment-to-moment basis. However, little is known about when the interpretation itself is encoded in memory. For example, in scalar inferences, listeners overwhelmingly prefer the pragmatic meaning of "*some*", but initially consider its semantic meaning.[i] Is the semantic meaning included in the final interpretation of the sentence? Or is it replaced by the pragmatic inference? One possibility is that the system waits until after pragmatic analysis to encode an interpretation into memory. Consistent with traditional models of sentence processing, this would result in a single interpretation of each sentence.[ii] Alternatively, the processing system may interpret and encode all interpretations under consideration, before pragmatic analysis and regardless of whether they fit with the context.

The current study uses two tasks to investigate whether the semantic meaning of "some" is encoded in memory prior to a scalar inference. First, during the word-learning task,[iii] participants (n = 40) heard instructions like "*Click on the girl that has some of the blickets*" while their eye-movements were recorded to a display (Fig. 1) featuring a subset of objects (girl with 2-out-of-4 items), and a total-set of objects (girl with 3-out-of-3 items). Thus, both sets are consistent with the semantics of "*some*" but only the subset is consistent with the implicature. Filler trials featured the quantifiers "*two*", "*three*", and "*all"*. The target is the subset character in "*some/two*" trials and the total-set character in "*all/three*" trials. Second, during the recall task, participants saw objects that were previously associated with the subset and total-set (Fig. 2), and were instructed to *"Click on the blicket*." Importantly, to examine participants' memory for the semantic meaning of "some" during the recall task, we calculated the proportion of matches between responses on the word-learning and recall tasks. A response was coded as a match if the same object is selected in both word learning and recall. We only analyzed the matches for accurate word-learning trials because we want to probe memory for interpretations made via pragmatic inference. If the semantic meaning is overridden by pragmatic inference, recall for the subset object should be as high for "some" as "two/three/all". Alternatively, if the semantic meaning is encoded in memory prior to the inference, it may interfere with recall and lead to fewer matches for "some" trials than "two/three/all" trials.

During the word-learning task, we analyzed proportion of looks to the target character (target over competitor looks) following quantifier onset. Figure 3 illustrates that Target looks were generally lower for quantifiers compared to number words. This is likely because the exact semantics of number words isolates the domain of quantification to the basic level, and generates a clear expectation that the up-coming novel word will distinguish the objects. The quantifier terms refer to relationships between individuals within a set, so listeners might entertain the possibility that the novel word is a superordinate category that refers to both object kinds. Critically, comparisons to chance indicate that looks converged on the target following "all" (200ms), "two" (100ms), and "three" (100ms). In contrast, a preference for the target in the "some" condition emerged at 700ms. This confirms that semantic analysis precedes pragmatic inference.[v] Moreover, participants selected the subset on 85% of trials, indicating that the pragmatic inference was made (Fig. 4). Figure 5 shows that recall of the novel object labels was significantly greater for trials that did not feature a pragmatic inference ("*two"*: 75%, "*all*": 74%, "*three*": 77%) than trials that did ("*some*": 59%). There was a significant main effect of scale

type ($p < 0.05$) and strength (lesser or greater) ($p < 0.05$), as well as a significant interaction between scale type and strength ($p < 0.05$). Although traditional models of sentence processing assume that comprehension results in a single and accurate representation of what was said, these findings suggest that listeners encode the semantic meaning of "some" in memory prior to making the scalar inference. This results in an interpretation that features multiple meanings.

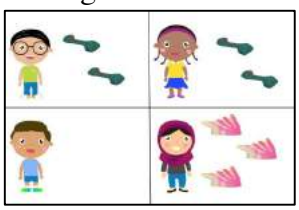


Figure 1. *Sample display from the word-learning task.*

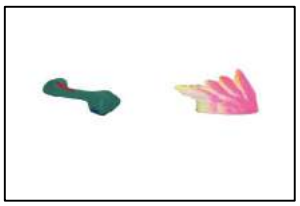


Figure 2. *Sample display from the recall task.*

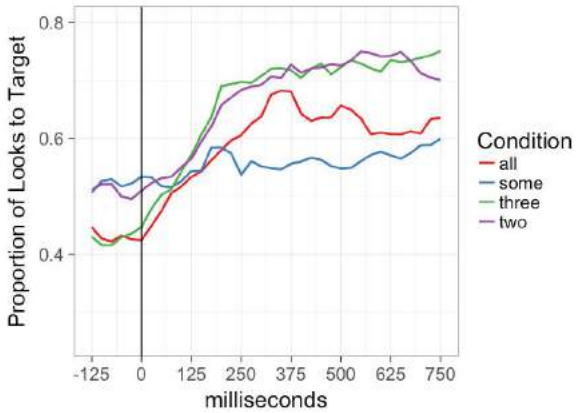


Figure 3. Proportion of looks to the target referent following the onset of the quantifier.

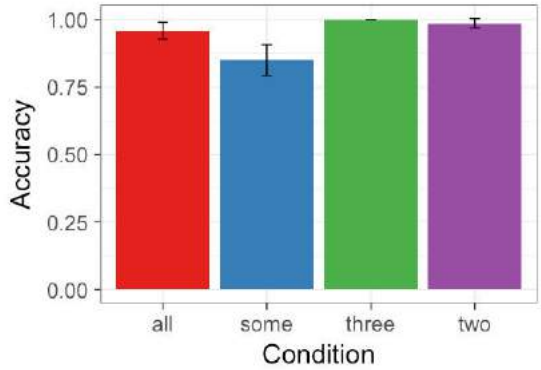


Figure 4. Word learning accuracy by condition.

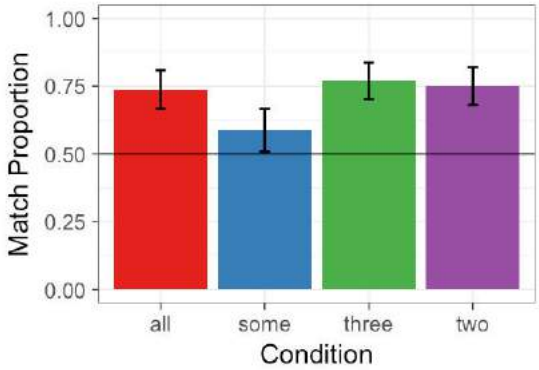


Figure 5. Percent word-learning and recall matches by condition.

---

[i] 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.

[ii] MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological review, 101(4), 676.

[iii] Huang, Y.T., & Arnold, A. (2016). Word learning in linguistic context: Processing and memory effects. Cognition.

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

Mentioning atypical properties of objects is communicatively efficient

Elisa Kreiss[1], Judith Degen[2], Robert X.D. Hawkins[2], and Noah Goodman[2]
[1]University of Osnabrück, [2]Stanford University

What governs how much information speakers include in referring expressions? One pressure is for speakers to include just enough information for the interlocutor to correctly select an intended referent from among a set of potential referents [3]. This amounts to calling the target object a "banana" in 1a), where there is no competing banana; but a "yellow banana" in 1c), where there is a competing (brown) banana. However, speakers also have a well-documented preference to mention properties of objects "overinformatively", especially color [4]. For example, speakers are likely to call the banana in 1a) a "brown banana" some of the time. More precisely, speakers tend to mention atypical rather than typical properties of objects overinformatively [5] [6] [7]. An account of why more typical properties are less likely to be mentioned is lacking. Some have proposed that it is due to a speaker-internal pressure to mention salient properties; others have proposed that speakers aim to facilitate the listener's visual search. We ask: when should a rational speaker with the goal of communicating an intended referent mention an object's color?

Model. We model reference production within the Rational Speech Act framework [1]. Taking inspiration from [2], utterances (simple nouns like "banana", simple color adjectives like "blue", and modified noun phrases like "blue banana") are taken to have a graded semantics: rather than assuming that the bananas shown in Fig. 1a)-1d) are equally good instances of "banana" or that all shades are equally "blue", we empirically elicited object-utterance typicality values on MTurk for all possible utterances. The pragmatic speaker selects utterances proportionally to the probability that a literal listener using a graded semantics will select the correct object. The listener is more likely to select a typical yellow banana upon hearing "banana," thus it is more informative for the speaker to mention "COLOR banana" when the intended referent is atypical.

Production experiment. In order to evaluate the RSA model quantitatively, we collected freely produced referring expressions in a multi-player online reference game experiment using contexts such as those depicted in Fig. 1. 60 pairs of participants were recruited through MTurk and randomly assigned to speaker and listener role. Speakers used a chat window to produce a referring expression that would allow the listener to click on the target object. Once listeners made a choice by clicking on an object, feedback was provided to both participants. Stimuli were photo-realistic depictions of food items that occurred in three different colors, which differed in typicality. Conditions differed in whether mentioning color was "informative" (necessary for uniquely establishing reference, 1c-d) or "overinformative" (1a-b); and whether there was a competitor from another food category of the same color (1b/1d) or not (1a/1c).

Results are visualized in Fig. 2**.** For ease of exposition, we focus on whether or not color was mentioned at all (though the RSA model predicts the entire utterance distribution for each of the unique 1085 contexts). Color was mentioned more often in informative than in overinformative contexts (β=5.27,p<.0001) and more often when there was no color competitor than when there was (β=.67,p<.0001). Crucially, there was a main effect of typicality in the expected direction – the more typical an object was for the simple nominal expression, the less likely color was mentioned (β=-4.11,p<.0001), replicating previous studies. This was the case even when color was informative – in these cases, participants preferred to sometimes say "banana" for the very typical banana even though there was another banana present. BDA suggests the graded semantics model captures these data much better than a deterministic semantics model (r=.8). We conclude that the systematicity with which speakers redundantly mention color implicates a system geared towards communicative efficiency rather than towards wasteful

overinformativeness. We discuss potential extensions of this approach to other production phenomena, such as optional instrument mention.

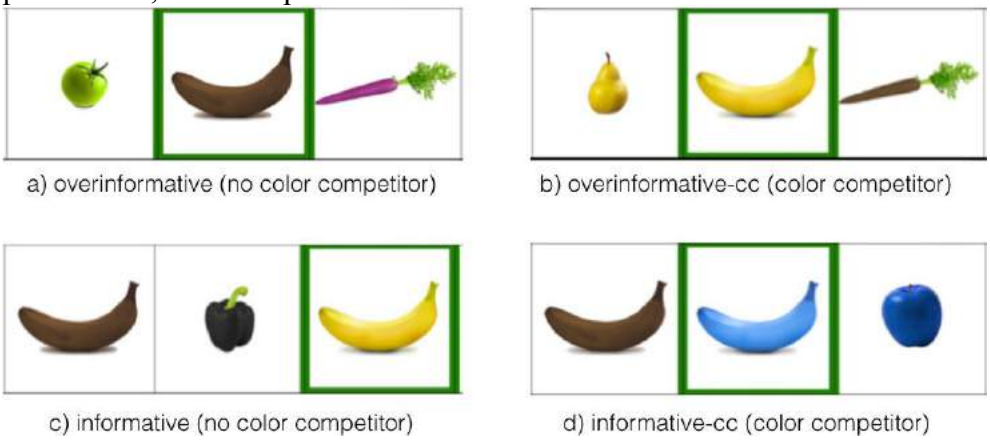


Fig. 1. Examples of relevant informativeness and color competitor presence conditions.

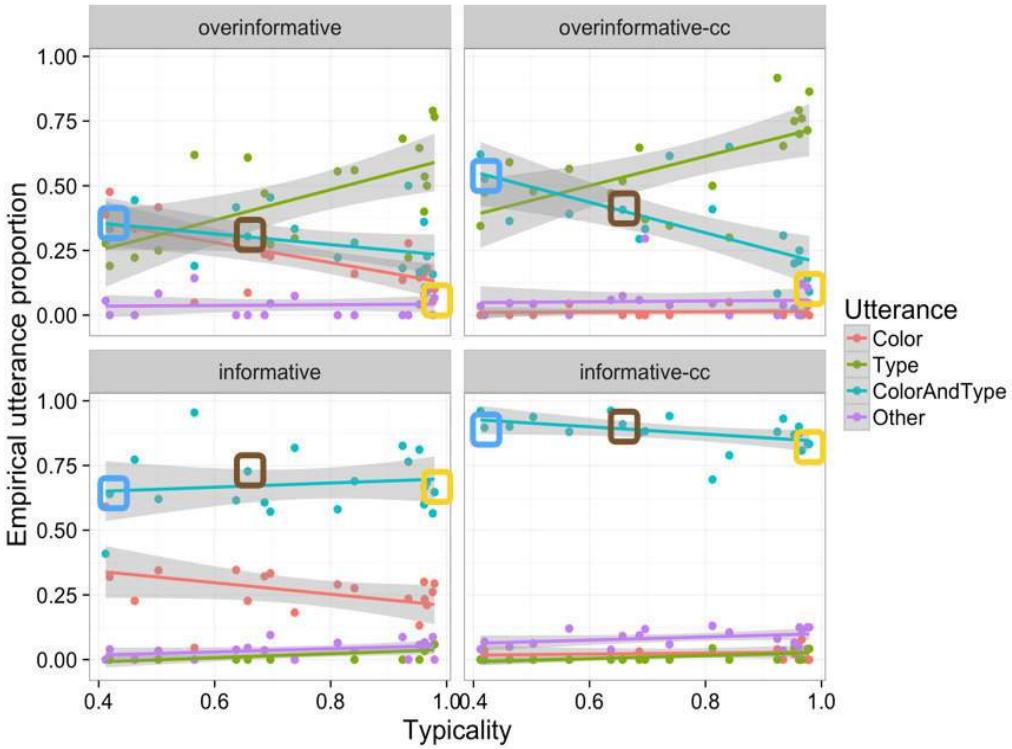


Fig. 2. Proportion of Color ("blue"), Type ("banana"), and ColorAndType ("blue banana") utterances as a function of mean object typicality for the Type utterance, across conditions. "COLOR banana" cases are circled in their respective color.

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

[2] Graf, C., Degen, J., Hawkins, R. X. D., & Goodman, N. D. (2016). Animal, dog, or dalmatian ? Level of abstraction in nominal referring expressions. In *Proceedings of CogSci 38*.

[3] Grice, H. P. (1975). Logic and Conversation. *Syntax and Semantics, 3* , 41–58.

[4] Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics, 27*(1), 89–110.

[5] Rubio-Fernandez, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology, 7* (153).

[6] Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of psycholinguistic research, 32*(1), 3–23.

[7] Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: the case of color typicality. *Frontiers in Psychology, 6*(July), 1– 12.

Is working memory sensitive to at-issueness? Experimental evidence from at-issue appositives

Margaret Kroll & Matt Wagers
University of California, Santa Cruz

Introduction This project presents new experimental evidence that challenges a view of discourse processing in which at-issue and not-at-issue content rely on different sets of working memory resources. Under the assumption that restrictive relatives and appositives differ in contributing at-issue and not-at-issue content, respectively, a recent finding by Dillon et al. that acceptability ratings are much more sensitive to the length of restrictive relatives than to the length of appositives supports a model of discourse processing in which the parsing operations that construct at-issue and not-at-issue content proceed independently [1, 2]. However, the assumption that appositives always contribute not-at-issue content is challenged in corpus and experimental work showing that appositives can sometimes be interpreted as at-issue [3, 4]. We present a new experimental design to control the at-issue status of appositive content, allowing us to directly test whether it is the at-issue status of appositives and restrictive relatives that is driving the observed acceptability differences. We find that, counter to the predictions of Dillon et al., acceptability ratings show the same sensitivity to appositive length whether the appositive contributes not-at-issue *or* at-issue content. We argue that the observed acceptability differences between restrictive relatives and appositives cannot be attributed to the not-at-issue status of appositive content, and offer an alternative explanation in which the differences are attributed to the burdening of particular prosodic domains.

Background The at-issue/not-at-issue distinction splits utterance content into primary and secondary information, respectively [2, 5, 6]. Not-at-issue content consists of projective meaning that does not contribute to resolving the current Question Under Discussion (QUD) [7, 8, 9]. It traditionally includes presuppositions, appositives, and parentheticals. Potts [2] influentially proposed that at-issue and appositive content are logically and compositionally independent. Subsequent work has complicated this view by demonstrating that appositive content can often behave as at-issue, such as appositives' ability to be targeted by polarity response particles [3, 4] and their failure to project in certain environments [10]. These observations suggest that appositives can sometimes be interpreted as contributing at-issue content [3, 4, 6].

Dillon et al. [1] show that the contribution of appositives to the perceived complexity of their embedding clause is less than the contribution of a comparable restrictive relative. To illustrate, adding the bolded material in (1) inside a restrictive relative decreases the acceptability of the entire sentence more so than adding the bolded material inside an appositive in (2).

(1) The fox that is reading a poem **the host highly recommended** is sitting on the ottoman.

(2) The fox (who is reading a poem **the host highly recommended**) is sitting on the ottoman.

Dillon et al. conclude that the parsing operations that construct not-at-issue structures consume resources independently from those that construct at-issue main clauses. The authors show in a series of follow-up experiments that this interaction effect is not due to attentional differences nor due to retrieval interference at the main verb.

Current Study While Dillon et al. rely on the canonical status of appositives as contributors of not-at-issue content, recent work suggests that appositive constructions **can** contribute at-issue content. We use this observation to probe the claim that the acceptability differences observed in sentences like (1)-(2) are due to the (not-)at-issue status of the appositive clause: **If the length effect in appositives is attenuated because of the clauses' not-at-issue status, then we expect the effect to strengthen when the appositive is interpreted as at-issue.**

Experiment 1 Exp. 1 extends the Dillon et al. findings. It used materials adjusted to more closely match syntactic structures across content types, a different subject pool, and different fillers. We found that when sentences are presented in out-of-the-blue contexts, adding additional length to a restrictive relative clause led to a greater decrease in the acceptability of the containing sentence than adding length to an appositive clause in a corresponding sentence (an interaction effect of sentence length and clause structure type, $p < .01$; $N_{subj} = 24$, $N_{items} = 24$). This finding is consistent with the Dillon et al. results.

Experiment 2 Exp. 2 directly tests whether the observed acceptability effects are affected by the at-issue status of appositives. To do this, we embedded target sentences in a multi-exchange discourse organized as a simulated text message exchange between two interlocutors. For each item, the target sentence was presented as one interlocutor's answer to an explicit question from the other interlocutor. The design utilizes two observations: 1) only at-issue content can address the current QUD; and 2) appositive relative clauses are able to address part of a coordinated-question QUD [11], and in doing so contribute at-issue content. Exp. 2 uses a 2x2x2 design that crosses the conditions LENGTH (*long* and *short* sentences), STRUCTURE type (*parenthetical* and *restrictive*), and AT-ISSUENESS of the appositive clause (*not-at-issue* and *at-issue*). The at-issue status of appositives was controlled for by varying whether the target sentence appeared as the answer to a single QUD (not-at-issue condition) or to a coordinated QUD (at-issue condition). Two conditions are exemplified in Figure 1.

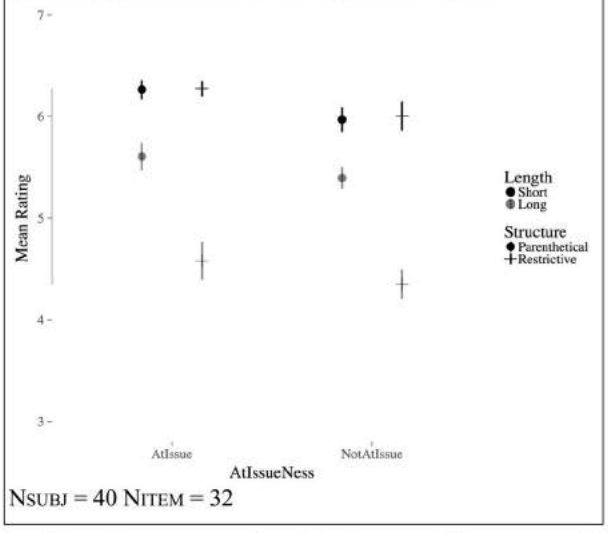
FIGURE 1. EXPERIMENT 2 PARTIAL ITEM EXAMPLE

Condition	QUD	Short Parenthetical
At-issue	Where is the bear standing and what is it wearing?	The bear (who is standing on the ball) is wearing a hat.
Not-at-issue	What is the bear wearing?	

We found main effects of LENGTH, STRUCTURE, and AT-ISSUENESS, and an interaction of LENGTH and STRUCTURE (all *p*s < .001). No other effects reached significance. Crucially, we found no interaction of LENGTH, STRUCTURE, and AT-ISSUENESS. As Figure 2 shows, **no length effects were affected by AT-ISSUENESS.**

Experiment 3 Exp. 3 is identical to Exp. 2 but uses items containing final appositives–appositives whose matrix anchor is an object–instead of medial appositives. Final appositives have been argued to more easily contribute at-issue content [11]; therefore, these items constitute a stronger test than Exp. 2. Even so, the results for Exp. 3 show the same pattern of length effects that we found in Exp. 2.

FIGURE 2. EXPERIMENT 2 MEAN RATINGS DATA



NSUBJ = 40 NITEM = 32

Conclusion The contributions of this project are twofold: we develop a new methodology for studying at-issueness experimentally, and present evidence complicating the picture of discourse processing in which at-issue and not-at-issue content draw from separate resources in working memory. We show that, counter to predictions, at-issue appositives burden working memory to the same extent as not-at-issue appositives. We account for the acceptability differences between restrictive relatives and appositives by proposing that appositive prosody facilitates how input is chunked in short term memory *independently* of the at-issue status of the input material [12, 13, 14].

References [1] Dillon, B. et al. 2014. Pushed aside. [2] Potts, C. 2005. *The logic of conventional implicatures*. [3] Koev, T. and K. Syrett. 2014. Experimental evidence. [4] AnderBois, S. et al. 2011. Crossing the appositive/at-issue meaning boundary. [5] Chierchia, G. & McConnell-Ginet, S. 2000. *Meaning and grammar.* [6] Simons, M. et al. 2010. What projects and why. [7]

Roberts, C. 1996/2012. Information structure in discourse. [8] Ginzburg, J. 1996. Interrogatives. [9] Roberts, C. et al. 2009. Presupposition, conventional implicature, and beyond. [10] Potts, C. & J. Harris. 2009. Perspective-shifting with appositives and expressives. [11] Koev, T. 2013. *Apposition and the structure of discourse*. [12] Redeker, G. 2006. Discourse markers as attentional cues. [13] Fodor, J.D. 2002. Prosodic disambiguation in silent reading. [14] Drury, J.E. et al. 2016. Punctuation and implicit prosody in silent reading.

Influence of Interpersonal Variables during Utterance Comprehension: A Neurophysiological Investigation with the Korean Honorific System

Jarang Kwak, Haejin Kim, Soyoung Kwon, & Donghoon Lee
Department of Psychology, Pusan National University

This study investigated the influence of social information about interlocutors, which can be extracted from terms of addressee, on the comprehension of the following utterance with a neurophysiological method. In verbal communication, the use of an address term functions to indicate the relative status of the addressee as well as social distance between the conversation partners. Based on the social information the expectation of the level of politeness and corresponding linguistic forms for following utterances can be established. In our ERP experiment Korean participants performed an acceptability task with single utterances that began with a term of address and ended with a sentence-final verb. In Korean, which has advanced linguistic devices for politeness, the use of an honorific form of the verb is expected for politeness purpose when the addressee possesses a relatively higher status than the speaker. The pragmatic agreement between the honorific form of the verb and the social status of the addressee yields critical conditions in this study, status-match vs. status mismatch. More importantly, we manipulated the social distance information (i.e., close vs. distant) between the conversation partners by different types of address terms used in private relationship or official relationship. Behavioral data indicated that the misuse of verb honorific form is more or less acceptable when a close social distance between the speaker and the addressee has been expected from the address term of an interlocutor. From the ERP data, we expected the N400 effect when the social status information implied by an address term mismatched the use of an honorific form of a critical verb, which was observed as previous ERP studies with the Chinese honorific system (Jiang et al., 2013). Critically, we hypothesized that the N400 effect from the mismatch of honorific forms would be variable in terms of the social distance between interlocutors. The ERP results supported our hypotheses (Figure 1). The N400 was significantly observed for the status-mismatch (e.g., "Boss, I've finished$_{\text{less respectful}}$ the work.") as compared to the status-match condition. However, the N400 effect disappeared when the addressee is a figure in a close relationship (e.g., "Father, I've finished$_{\text{less respectful}}$ the work."). The current ERP evidence suggests that the listener build expectation about the politeness of the upcoming utterance based on the social status as well as distance information reflected on the term of addressee.

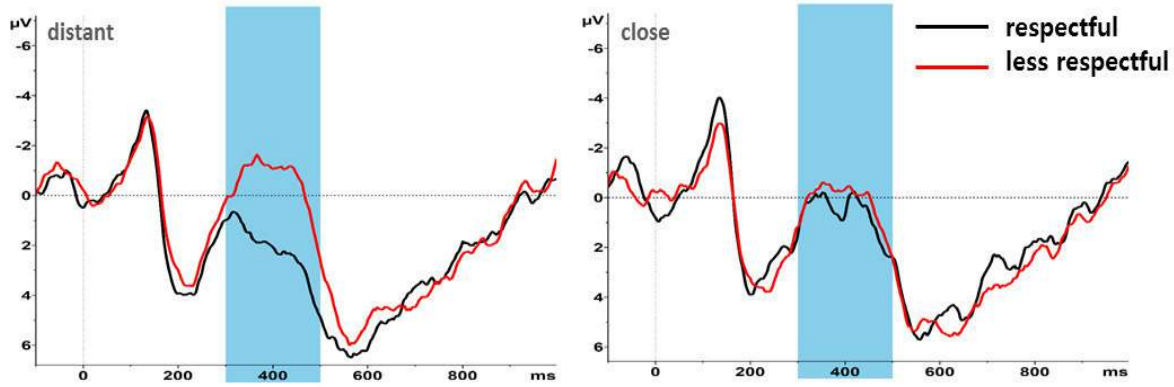


Figure 1. Grand average waveforms at Cz electrode showing more negative-going waves over 300-500 ms window for the status-mismatch compared with the status-match condition. Note that the N400 effect size varied as a function of social distance information.

< Reference>

Jiang, X. M., Li, Y., & Zhou, X. L. (2013). Is it over-respectful or disrespectful? Differential patterns of brain activity in perceiving pragmatic violation of social status information during utterance comprehension. Neuropsychologia, 51(11), 2210–2223.

Does it matter who is producing an utterance? – Effect of speaker identity in utterances without self-reference

Eva Link, Holger Schneider, Kristina Schopf, Marcel Schwille, Franziska Rück, & Barbara Kaup
University of Tübingen

There is a debate, whether extra-linguistic information about a speaker is integrated as fast as semantic information (one-step model of comprehension) or delayed until a second step (two-step). It seems that the information of a speaker gets rapidly integrated (van Berkum, van den Brink, Tesink, Kos & Hagoort, 2008). To assess the question, whether self-reference – the use of the first person pronoun – might have caused the early integration, we conducted a self-paced reading study, where we employed utterances, which did not directly refer to the speaker. The prolonged reading times for target words mismatching the speaker indicate that readers do integrate speaker identity early on during comprehension. Future ERP-studies with this material might provide more insight into the time course of the integration of speaker identity.

In the debate about the time course of language understanding there is the question whether comprehenders integrate extra-linguistic information as fast as they integrate semantic information (one-step model of comprehension) or whether the integration of extra-linguistic information is delayed until a second step during comprehension (two-step models of comprehension). A study by van Berkum, van den Brink, Tesink, Kos, and Hagoort (2008) demonstrated that comprehenders rapidly integrate the characteristics of a speaker during language comprehension, and thus provides evidence for a one-step model of comprehension. However, the sentences employed in this study all contained a first person singular pronoun referring directly to the speaker of the utterance (e.g., "Every evening I drink some wine before I go to sleep"). This overt reference to the speaker of the utterance might have encouraged listeners to integrate the corresponding extra-linguistic information at an early stage during comprehension. In the current study, we were interested in whether speaker identity is also taken into account early on during the comprehension process when the utterances do not directly refer to the speaker of the utterance.

To address this question, we conducted a self-paced reading study. Participants read 40 utterances which were plausible for a certain group of people but implausible for another without containing any direct reference to the speaker. One word was manipulated and determined the plausibility of the utterance, for example:

> Der schöne Teil des Tages beginnt erst nach der *Schule/Arbeit* so richtig.
> The nice part of the day does not really start until after *school/work*.

In this example, the sentence version with the word *school* is typical for a child, but atypical for an adult. The same applies vice versa to the version with the word *work*. The match vs mismatch in the sentences did not exclusively relate to age, as in the example sentence above, but also to gender (male/female) or political orientation (liberal/conservative) among others. A picture of the speaker, whose stereotypical appearance either matched or mismatched the content of the utterance, was shown above the sentence (see Figure 1). We presented all of the items with a matching and a mismatching speaker, though the participants always saw only one of the four possible combinations.

Sixty-seven native speakers of German took part in the study. One participant did not score above 80% in the control questions randomly asked about filler sentences and was therefore excluded. The remaining 66 participants (3 male) were 18 to 33 years old (M = 20.84, SD = 2.79).

___ _____ ____ ___ _____ _____ ____ ____ ___ Schule __ _____.

Figure 1. Exemplary trial with a picture of a matching speaker (Reines, 2011) and utterance. Here, the participant has reached the target word (*school*).

We performed paired t-tests on reading times for the target word (e.g., “Schule” vs. “Arbeit”, as well as for the pre-target word (e.g., “der”), the post-target (e.g., “so”), and the final word of the sentence (e.g., “richtig”). Reading times did not differ significantly for the pre-target word (7ms, $t_1(65) = 1.84$, p=.07, t_2 (39) = 1.11, n.s.). However, reading times were significantly longer in the mismatching than in the matching condition for the target word (20ms, $t_1(65) = 2.92$, $p = .005$, $t_2(39) = 2.03$, $p = .05$). This effect was still visible on the post-target word (16ms, $t_1(65) = 2.82$, $p = .007$, $t_2(35) = 2.32$, $p = .03$). There was no mismatch effect on the final word of the sentence (6ms, $t_1 < 1$, $t_2<1$).

As expected, there was no significant difference in reading times before participants reached the manipulated target word. As soon as a mismatch occurred between speaker and sentence, reading times were significantly prolonged. The effect persisted on the post-target word and was no longer visible at the end of the sentence. These results indicate that readers do indeed integrate the characteristics of the speaker of an utterance early on during the comprehension process even when the utterance does not directly refer to the speaker. As such, these results provide further evidence for one-step models of comprehension according to which comprehenders rapidly integrate information from all available sources early on during comprehension. As self-paced reading is not of high temporal resolution, future ERP-studies with these materials might provide more insight concerning the time course with which extra-linguistic information concerning speaker identity gets integrated during the comprehension of utterances without direct reference to the speaker.

Reines, S. (2011). Junge [Online image]. Retrieved February 14, 2017 from https:// fotoreines.wordpress.com/willkommen-auf-meiner-homepage/

van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M. & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, *20*, 580-591.

Asymmetries between interpretation and production in Catalan pronouns

Laia Mayol (Universitat Pompeu Fabra)

Introduction. The literature on Romance null-subject languages has often postulated a division of labor between Null and Overt pronouns: Nulls prefer to retrieve an antecedent in subject position, whereas Overts prefer an antecedent in a lower syntactic position (Carminati, 2002). However, recent research on English pronouns (Rohde and Kehler, 2014) has shown that pronoun interpretation and production are sensitive to different set of factors and, instead of being mirror images of each other, are related probabilistically in a Bayesian fashion. It is an open question whether null-subject language pronouns can be accounted in the same way. We show that both Null and Overts exhibit the same asymmetry in Catalan, by means of two discourse completion studies: production is only sensitive to grammatical factors, while interpretation is also sensitive to pragmatic factors (i.e. context type and rhetorical relation).

Experiment 1. A discourse-completion study was carried out in which the context sentence used a transfer of possession verb (TPV), see (1). TPV contexts are interesting because a pronoun following it may not have the usual subject preference (Stevenson et al., 1994): i.e. a very natural follow-up of (1) would be to explain what Pere did with the book. Experiment 1 had 3 conditions: Free (participants could continue as they wished), Overt and Null (participants had to continue with an Overt and a Null, respectively). Three lists were generated with 18 items and 18 fillers. 90 Catalan native speakers participated in the experiment, yielding 1620 completions (see (2) for an example with a Null). We will only discuss the 1152 completions in which the subject unambiguously referred either to the subject or object of the context sentence.

(1) El Robert li va passar un llibre al Pere.
'Robert passed a book to Peter.'

(2) Va donar -li les gràcies.
Gave him the thanks.

The data in the Null and Overt conditions, in Table 1, are interpretation probabilities: i.e. of how participants understood the pronouns presented to them. In the Null condition, the pronoun displayed a mild object preference, and in the Overt condition, the pronoun exhibited a very strong object preference. In the Free condition, we coded what type of expression was chosen to refer to either the subject or the object. These data, in Table 2, are, thus, production probabilities.

Table 1 (*)[1]	Null	Overt	Free
Subject	38	10	24
Object	62	90	76

Table 2 (*)	Subject	Object
Null	72	29
Overt	6	15
Proper Name	21	56

The results have uncovered an asymmetry between interpretation and production. Although Nulls displayed a mild interpretation bias towards the object, they have strong production bias towards the subject. When participants could choose a form they overwhelmingly chose a Null to refer to the subject. However, when they had to interpret a pronoun, the pragmatic biases of VTPs came into play and the Null pronoun lost its subject bias. The data has also uncovered another asymmetry concerning Overts. Although interpreters have a very strong bias to interpret an Overt as referring to the object, this is not how they use it when they can choose which form to use. The results also show that, while Nulls are more subject biased than Overts (as expected), it is possible for Nulls to have an overall Object preference, depending on pragmatic factors.

We further coded the rhetorical relation established between the context and the completion sentence to examine the role of pragmatic factors. Here we discuss the results of two rhetorical relations which have been shown to display opposed biases: Occasion and Elaboration (Kehler et. al. 2008). In an Occasion, the events described by the two sentences are temporally ordered, while in an Elaboration, the two sentences provide descriptions of the same eventuality. In the

[1] * indicates statistically significant differences at the .05 level in a mixed-effect logistic regression with item and participant as random effects.

former the end state for the first eventuality is important for the coherence relation, while in the latter it is not. This affects pronoun interpretation: pronouns in an Occasion , (3-a), are biased towards the object, since it is the most salient referent at the end state of the VTP sentence. In contrast, pronouns in an Elaboration, (3-b), are biased towards the subject, since Elaborations do not focus on the end state of the first event.

(3) a. John handed a book to Bob. He began reading it.
b. John handed a book to Bob. He did so slowly and carefully.

Table 3 shows the subject bias of Nulls and Overts in the two rhetorical relations. As before, Nulls are more subject biased than Overts, but their biases are greatly affected by the rhetorical relation: Elaboration is much more subject-biased than Occasion. Table 4 shows the production data: the subject bias of Null pronouns by rhetorical relation. The difference is not statistically significant showing again that production, unlike interpretation, is insensitive to pragmatic factors.

Table 3 (*)	Null	Overt
Occasion	11	0
Elaboration	75	41

Table 4	%
Occasion	80
Elaboration	71

Experiment 2 had the same design as the previous experiments, but included implicit causality verbs, which impute the cause of the event they denote either to the subject (ICV1; *surprise*) or the the object (ICV2; *congratulate*). The experiment contained 30 critical items (15 ICV1s and 15 ICV2s) and 20 fillers. 78 native speakers of Catalan participated yielding 2340 completions, of which we analyze the 1963 that unambiguously referred to the previous subject or object.

Table 5 shows the percentage of subject reference of both pronouns depending on verb type: Nulls are more subject-biased than Overts, and ICV1s trigger more subject references than ICV2s. Thus, pronoun interpretation is sensitive to both grammatical and pragmatic factors. Table 6 shows the choice of referring expression in the Free condition for subject reference: Nulls are the favorite form regardless of whether the context is ICV1 or ICV2. Production is not affected by the pragmatic factors (i.e. verb type) that did affect interpretation.

Table 5 (*)	Null	Overt
All	65	20
VIC1	76	51
VIC2	52	15

Table 6	ICV1	ICV2
Null	87	93
Overt	4	2
Proper name	9	5

A Bayesian account. The data supports a model in which interpretation and production are related in a Bayesian fashion, as in (4). $P(subj \mid pronoun)$ is the interpretation probability: the probability that a pronoun refers to the subject. $P(pronoun \mid subj)$ is the production probability: the probability that the speaker uses a pronoun given that she wants to refer to the subject. These two probabilities are not mirror images of each other; a third probability plays a role, $P(subj)$, which is the probability that the subject will be mentioned regardless of the form (Null, Overt or Proper name).[2]

(4)

$$P(subj \mid pronoun) = \frac{P(pronoun \mid subj)\,P(subj)}{P(pronoun)}$$

In experiment 1, the observed probability $P(subject \mid null)$ was 38% (see table 1) and the expected probability applying the formula in (4) is 44%. In experiment 2, the observed probability $P(subject \mid null)$ was 65% (see Table 5) and the expected probability is 64%. For both experiments a linear regression test over item means was conducted, which yielded statistically significant correlations between observed and predicted means.

References. Carminati, M. N. (2002). *The processing of Italian subject pronouns*. PhD thesis, UMasss. • Rohde, H. and Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8):912–927. • Stevenson, R. J., Crawley, R. A., and Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.

[2] $P(pronoun)$ is the probability that a pronoun is used and is computed summing the denominator over all possible referents. It contributes a constant factor and normalizes the probabilities over all possible referents to 1.

Believing what you are told: Politeness and scalar inferences

Diana Mazzarella, Emmanuel Trouche, Hugo Mercier, Ira Noveck
Institut des sciences cognitives Marc Jeannerod

Recent studies in experimental pragmatics have investigated the effect of politeness on the derivation of scalar inferences. Most notably, Bonnefon and colleagues (2009, 2011) claim that when the scalar utterance is face-threatening ('Some people hated your speech') the scalar inference is blocked. Furthermore, contrary to evidence showing that scalar inferences come with extra cost (since Bott & Noveck, 2004), they suggest that, in face-threatening contexts, the *semantic* interpretation - *at least some people hated your speech* - is arrived at slowly and effortfully (as compared to face-boosting contexts). Their claims rest on tasks such as Table 1's, where participants are presented with a scenario that ends with a scalar utterance, which is either face-threatening (*Some people hated your speech*) or face-boosting (*Some people loved your speech*). At that point, participants are asked to evaluate the utterance through what we label as the *semantic compatibility question*. Bonnefon and colleagues report a significantly higher percentage of 'Yes' answers to the semantic compatibility question in the face-threat condition than in the face-boost one. Furthermore, 'Yes' answers take significantly longer, but only in the face-threat condition.

We argue that, while intriguing, their analysis conflates the interpretation of *some* with a different mechanism, one in which a participant decides whether or not to *accept* a speaker's intended meaning. To be clearer, we distinguish between the *derivation* of the scalar inference, which can arise as part of the comprehension process, and its *epistemic assessment*, which can result in the possibility that the addressee will subsequently reject such pragmatically generated output. The gap between comprehension and acceptance is typically bridged by *epistemic trust* (Sperber et al., 2010). Crucially, in face-threatening contexts, the addressee may have reasons to doubt the truth of the pragmatically refined meaning (what the speaker communicates) because he thinks that the speaker is trying to be kind and polite (rather than strictly honest). We tested our hypothesis through a series of MTurk studies inspired by Bonnefon et al.'s task (see Table 1). Our main experimental innovation is that we separated the presentation of the scalar utterance from participants' evaluation of it. In this way, reaction times to each part could be measured separately. These will be referred to as $RT_{UTTERANCE}$, and $RT_{QUESTION}$; their combination will be referred to as RT_{TOTAL}.

Data preparation: In order to retain the cleanest data possible, we removed from our analysis participants who (i) clicked on the relevant screen more than necessary; (ii) exceeded the following reaction times: $RT_{UTTERANCE}$ > 20s, $RT_{QUESTION}$ > 30s; and, (iii) were identified as outliers (SD's exceeding ± 2.5).

Study 1 (N =292): The behavioural results of Study 1 replicate Bonnefon et al.'s finding showing that participants are more likely to answer 'Yes' to the semantic compatibility question in the face-threat condition (45%) than in the face-boost one (32%) (Fisher exact test, $p = .02$). With regard to RTs, Study 1 does not replicate Bonnefon et al. (2011). Overall RTs (RT_{TOTAL}) show that participants *tend* to be slower to answer 'Yes' when asked the semantic compatibility question in the face-threat condition ($F(1,288) = 2.75$, $p = .10$). However, as anticipated, by separating the scalar utterance from the listener's response to the question, it appears that slowdowns occur at the epistemic assessment stage ($RT_{QUESTION}$), with a main effect of condition ($F(1,288) = 4.21$, $p = .04$) and a tendency towards an interaction ($F(1,288) = 3.51$, $p = .06$). Similar analyses for the scalar utterance ($RT_{UTTERANCE}$) do not yield significant effects, all p's > .29.

Study 2 (N=294): Study 2 encourages participants to derive the scalar inference by (i) increasing its relevance and by (ii) characterizing the speaker as epistemically more knowledgeable with regard to the question at issue (see Table 1). As predicted, the face-threat/face-boost distinction is much clearer here ('Yes' answers: 45% vs. 12.5%) than in Study 1 (Fisher exact test, $p < .001$). Again, there are no new effects to report with respect to RT_{TOTAL} (p's > .10) and $RT_{UTTERANCE}$ (p's > .28). However, it is noteworthy that those who ultimately respond 'Yes' to the semantic compatibility question tend to read the scalar utterance faster than those who ultimately respond 'No', which goes counter to Bonnefon et al's main claim. As far as the $RT_{QUESTION}$ is concerned, we observed a significant effect of condition ($F(1,290) = 4.91$, p = .03), response type ($F(1,290) = 5.06$, p = .03), as well as an interaction that tends towards significance ($F(1,290) = 3.18$, p =.08). This trend indicates that those who respond 'Yes' in the Face-threat condition take an exceptionally long time to answer the semantic compatibility question.

Overall, we find no evidence that the scalar utterance is interpreted at different speeds across the two (face-threat vs. face-boost) conditions. Our data suggest, instead, that the process of epistemic evaluation, which operates when answering the semantic compatibility question, is the source of the reaction time differences. This undermines any claim that suggests that participants slow down while *drawing* a semantic reading of the utterance on line. These data open up an interesting direction of research within the field of experimental pragmatics as they highlight the importance of taking into consideration the cognitive distinction between comprehension and acceptance, which has been neglected in this literature so far.

References [1]Bonnefon, J-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition, 112*, 249-258. [2] Bonnefon, J-F., De Neyes, W., & Feeney, A. (2011). Processing scalar inferences in face-threatening contexts. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX. [3] Bott, L., & Noveck, I.A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Cognition*, *51*, 437–457. [4] Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 24*(4), 359-393.

Table 1. The table below displays the Speech story in the original version from Bonnefon et al. (2011) (translated from Dutch), as well as in the modified versions of Study 1 and Study 2. Relevant changes to the original story are in bold.

Bonnefon et. al. (2011)	**Study 1**	**Study 2**
Imagine you gave a speech at a small political rally. You are discussing your speech with Denise, who was in the audience. There were 6 other people in the audience. You are considering whether to give this same speech to another audience.	Imagine you gave a speech at a small political meeting. You are discussing your speech with Denise, who was also there. There were 6 other people in the audience that day. You tell Denise that you are thinking about giving the same speech to another group.	Imagine you gave a speech at a small political meeting. You are discussing your speech with Denise, who was also there. There were 6 other people in the audience that day and **you know that Denise spoke with all of them about it later**. **You tell Denise that you would like to know the audience's reaction.**
Hearing this, Denise tells you that 'Some people loved/hated your speech.'	Hearing this, Denise tells you that 'Some people loved/hated your speech.'	Hearing this, Denise tells you that 'Some people loved/hated your speech.'
Given what Denise told you, do you think that it is possible that everybody loved/hated your speech?	Given what Denise told you, do you think that it is possible that everybody loved/hated your speech?	Given what Denise told you, do you think that it is possible that everybody loved/hated your speech?

Erlinde Meertens, Andrea Beltrama, and Maribel Romero (University of Konstanz)
{erlinde.meertens,andrea.beltrama,maribel.romero}@uni-konstanz.de

Polar Questions, "or not" Alternative Questions and Complement Alternative Questions: an Experimental Study.

The puzzle – It has been argued that questions with seemingly identical semantic content have different pragmatic properties. In particular, Bolinger (1978) observed that "or not" Alternative Questions (henceforth, NAQ), contrary to their polar counterparts (PQ), are infelicitous in non-canonical uses – e.g., to make invites, draw inferences, or pose rhetorical questions (in (1-3)); and are instead especially appropriate to force the addressee to respond to an information-seeking question that previously went unanswered (in (4), see also Biezma 2009).

(1) **Invite**: Do you want a drink (# or not)?
(2) **Inference**: You are wet. Is it raining outside (# or not)?
(3) **Rhet**: Are you crazy (# or not)?
(4) **Info-seeking, asking 2nd time**: ✓Did you do your homework or not?

Such contrasts raises an issue: is the restricted illocutionary range of NAQs driven by he specific semantic-pragmatic properties that differentiate alternative from polar questions, or by general pragmatic principles? We address this issue by comparing the distribution of PQs and NAQS with Complement Alternative Questions (CAQ), a type of AQ that spells out the disjuncts differently.

(5) **PQ**: Is it a boy?
(6) **NAQ**: Is it a boy or not?
(7) **CAQ**: Is it a boy or a girl?

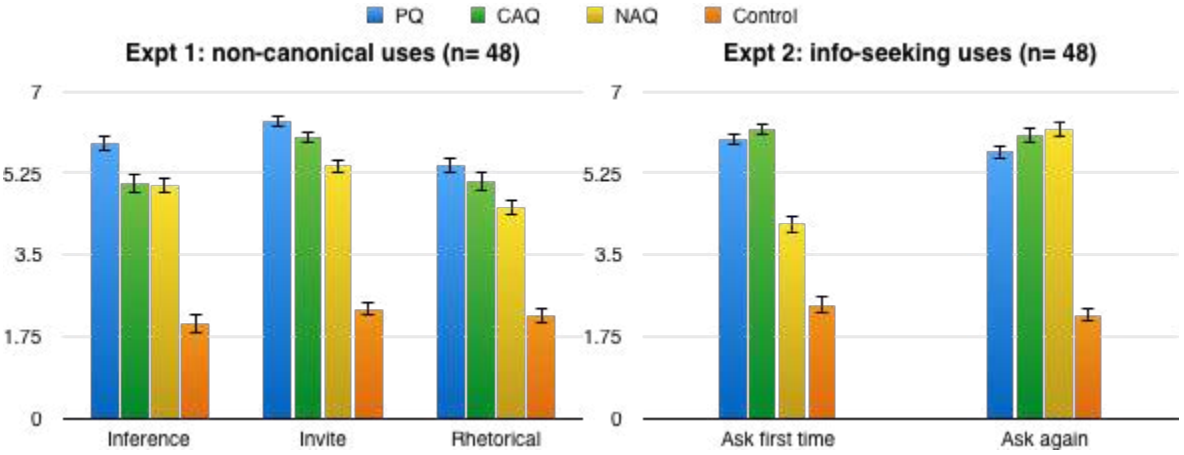
Background – Two competing analyses have been outlined to explain the facts in (1-4). Pragmatic accounts argue that PQs and NAQs have equal denotations {p;¬p}. What makes them different is that PQs, by virtue of pronouncing only one alternative, assign *p* a special pragmatic status – e.g., a higher Utility Value (van Roij and Safarova 2003); NAQs, by pronouncing both, signal *indifference* between the alternatives, driving infelicity in contexts where a preference of the speaker is expected (as in (1-3)). Semantic accounts hold that PQs denote a singleton set {p}, while NAQs denote two exhaustive, mutually exclusive alternatives {p, ¬p} (Biezma 2011, Biezma and Rawlins 2012). This property imbues NAQs with a flavor of *insistence*, which makes them a good strategy to force the addressee to respond to a previously unanswered question (in (4)), but infelicitous when the illocutionary goal is inconsistent with hard-pressing the listener (as in (1-3)). While both sound, such accounts share a limitation: by exclusively contrasting PQs and NAQs, they cannot shed light on whether the illocutionary restrictions of the latter reflect a general property of alternative questions; or if they are instead due to the effects of spelling out the second disjunct as "not *p*", rather than as a full proposition. CAQs emerge as an ideal testbed to address this issue: similarly to NAQs, they (i) pose logically opposite alternatives and (ii) pronounce both of them; yet, they spell out the second disjunct in full. We predict that if what causes the illocutionary restrictions in (1-3) is a general property of alternative questions – be that pragmatic indifference o semantics-driven insistence – CAQs should feature the same restrictions as NAQs. If the illocutionary range of NAQs, instead, is driven by the "or not" formulation of the second disjunct, such restrictions shouldn't hold for CAQs. We test this hypothesis by exploring the distribution of these questions in two rating experiments: Exp 1 for non-canonical uses; and Exp 2 for info-seeking uses.

Methods – Each trial consisted of a dialogue, at the end of which one participant asks a question. Subjects rated the naturalness of the question on a 1(min)-7(max) scale. Two factors were manipulated: (**i**) the question, with 4 different conditions: Polar, Negative Alternative, Complement Alternative, plus a control; (**ii**) the illocutionary goal of the speaker uttering the question, with Exp 1 comparing Inference, Invite and Rhetorical questions and Exp 2 comparing info-seeking questions used discourse-initially and to re-ask a question. The illocutionary goal of the speaker was

specified by the preceding context. 24 items were distributed in 4 lists with a LSD (24 fillers). 48 native speaker of English were recruited on MTurk for each study. Below is an example of an item.

Invite: Joe and Fred are at a party. Joe receives a call from Mary, who invites him over to her own party. John wants to invite Fred to join:
John: "Hey, do you want to ... {**PQ**:come to Mary's?/**NAQ**:come to Mary's or not?/**CAQ**:come to Mary's or do you want to stay here?/**Control**: Do you want a beer?}"

Results – The average ratings of Expt 1 and Expt 2 are plotted below. For both studies, mixed effect models revealed an interaction between Question Type and Context ($p < .001$). Paired comparisons within each type of context revealed the following contrasts. Expt 1: for Inferences, PQs were rated higher than NAQs and CAQs ($p<.001$), for Invites, PQs and CAQs were rated higher than NAQs ($p<.001$); for Rhetorical questions PQs and CAQs were rated higher than NAQs ($p<.001$). Expt 2: for discourse-initial uses, PQs and CAQs were rated higher than NAQs ($p<.0001$); for ask-again, NAQs were rated higher than PQs ($p<.01$) but did not differ from CAQs, while CAQs show a trend towards being rated higher than PQs ($p<.1$). The control condition was rated low across contexts.



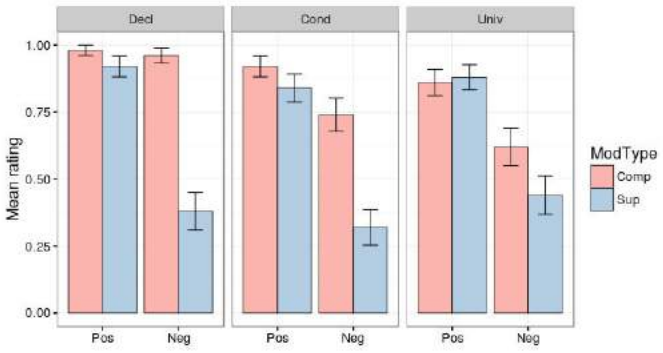
Discussion – On one hand, NAQs display a highly specialized illocutionary profile, confirming the claims from the literature. On the other hand, CAQs feature a significantly larger range: while they are as bad as NAQs to draw inferences and as good to re-ask a question, CAQs, can also be successfully used to make invites and ask rhetorical questions, as well as to pose discourse-initial inquiries. These findings support the idea that the illocutionary restrictions on NAQs are crucially underlied by the way in which the second disjunct is spelled out. We suggest that, by expressing the second disjunct through the negation of the first, NAQs bring about a complex effect: not only they lead the listener to pick between two exhaustive/exclusive alternatives; they also signal that one of them is to be preferred/is more important than the other. As such, NAQs emerge as a marked strategy to pose an alternative question in comparison to CAQs, which frame them as perfectly equivalent. Building on Horn's (1984) division of pragmatic labor, we suggest that NAQs' markedness restricts them to contexts where the combination of insistence and and emphasis on *p* is maximally functional for the speaker to achieve their illocutionary goal – i.e., those in which the speaker aims to re-ask a question to wrestle an answer from the listener; CAQs' unmarked status, by contrast, makes them less anchored to such "ask again" contexts, granting them the flexibility to operate in a broader range of situations. As such, exhaustivity/exclusivity and indifference alone, while still crucial, are not fine-grained enough to derive the restricted illocutionary range of NAQs.

Teodora Mihoc (Harvard U) and Kathryn Davidson (Harvard U)

Testing a PPI analysis of superlative modified numerals

Introduction. Comparative modified numerals (CMs) and superlative modified numerals (SMs) have equivalent truth conditions: *John saw **less than 4** stars* = *John saw **at most 3** stars* = John saw 0/1/2/3 stars (Cohen & Krifka 2011). However, they're known to differ in other ways: SMs require ignorance (Geurts & Nouwen 2007; Nouwen 2010; Coppock & Brochhagen 2013; Mayr 2013; Kennedy 2015; Mendia 2015), and SMs appear to be worse than CMs under negation: *John didn't see less than 4/??at most 3 stars* (Nilsen 2007, Cohen & Krifka 2011, 2014 and Spector 2014, 2015). To account for this difference, Spector (2015) treats SMs as underlyingly disjunctive positive polarity items (PPIs). We present 3 experiments in which we test **whether SMs pattern with other PPI items**: (a) PPIs are anti-licensed/not acceptable under a negation operator; (b) PPIs are acceptable in restrictors even when those define a DE-environment; and (c) PPIs can again become acceptable if the anti-licenser is itself in the scope of a DE-operator / in a DE-environment (Szabolcsi 2004, Nicolae 2012, Spector 2014). For each prediction we test whether the ways that SMs diverge from CMs would have been expected on a PPI account, and show that to the extent that a PPI account remains tenable, there still remains much data to be accounted for.

Fig. 1: Example trial: superlative modifier in negative declarative (Answers: Yes/No)



(Decl) I have/don't have [modifier] 3 [card suit].
(Cond) If you have/don't have [modifier] 3 [card suit], then we have something in common.
(Univ) Everyone who has/doesn't have [modifier] 3 [card suit] has something in common with me.

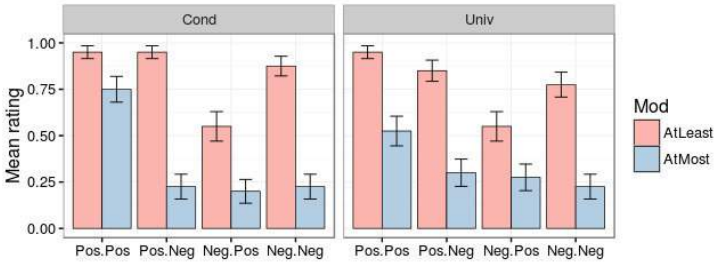
Fig. 2: Exp. 1 results grouped by sentence type, polarity, and modifier type

Experiment 1 (n = 25). Because SMs *require* ignorance and CMs *permit* it, our stimuli present (partial) ignorance via a card game, inspired by Cremers & Chemla (2016) (Fig. 1). Twenty-four acceptability judgments were presented which crossed modifiers (CMs: *less than*, *more than*, SMs: *at least*, *at most*), sentence type (declarative, conditional, universal) and polarity (positive, negative). Data were analyzed using mixed effects logistic regression.

Our results confirmed prediction (a): we found a significant difference between CMs and SMs under negation in simple declaratives ($\beta = 4.28, z = 5.24, p < .0001$). Second, supporting prediction (b), the effect of negation on SMs does not generalize to other downward-entailing environments: there was no difference between CMs and SMs in positive conditionals and universals. However, we did not find the same support for prediction (c): the contrast between CMs and SMs under negation remained when the combination occured under conditionals ($\beta = 2.25, z = 4.48, p < .0001$), and universals ($\beta = 0.92, z = 2.01, p = 0.044$) (Fig. 2). However, it has also been suggested that judgments of SMs under negation that are further

embedded in a conditional/universal are subject to a pragmatic polarity match between the antecedent/restrictor and the consequent/scope (Cohen & Krifka 2014), which could account for the failure of Exp. 1 to support prediction (c). We therefore designed a follow-up study (Exp. 2) to vary the negative or positive valence of the continuation.

Experiment 2 (n = 40). Design was similar to Exp. 1 except with positive/negative continuations of conditionals and universals (see Fig. 3). Partially supporting the alternative hypothesis, *at least* in a negative antecedent/restriction improved with polarity match in the consequent/scope (in conditionals: $\beta = 1.92, z = 3.21, p = 0.003$; in universals: $\beta = 1.15, z = 0.026, p = 0.027$), but *at most* did not. Further support for the heterogeny of SMs comes from another Exp. 2 finding that *at most* is judged unacceptable in conditionals and universals even when the only negative meaning is in the verb *lose* in the second argument (conditionals: $\beta = 2.62, z = 4.71, p < .0001$; universals: $\beta = 1.07, z = 2.15, p = 0.031$), a surprise under any account that treats the interaction of SMs with negation as only due to PPI anti-licensing. Given the improvement of *at least* under negation in Exp. 2 when further embedded in conditionals/universals, Exp. 3 tests whether this improvement generalizes to another combination of two DE operators: matrix and embedded negation.



(Cond) If you have/don't have [modifier] 3 [card suit], you win/lose.
(Univ) Everyone who has/doesn't have [modifier] 3 [card suit] wins/loses.

Fig. 3: Exp. 2 results grouped by sentence type, polarity, and modifier

Experiment 3 (n = 45). Design was similar to Exp. 1-2 except for the introduction of clausal embedding (and its supporting experimental context) in order to directly compare the behavior of local negation in two DE contexts: one under matrix negation versus the other in the antecedent of a conditional. Sentences varied by crossing four modifiers (as above) with sentence type (negative declarative (*Scyther doesn't know that he ...*), conditional (*If Scyther knew that he ...*)) and polarity (positive, negative for embedded clause/scopes). Results did not uniformly support prediction (c) for *at least*: *at least* under two negations is significantly worse than in a negated antecedent ($\beta = -1.9, z = -2.5, p = 0.025$) and differences between SMs persisted such that *at least* was judged worse than *at most* in every condition.

Conclusions. Overall, our data found some support for the predictions of the PPI account of SMs, but also raised multiple unresolved issues. First, why are SMs not always "rescued" when the anti-licensor is in the scope of another DE operator (*contra* prediction (c))? Second, why does the positive/negative valence of the predicate in the consequent play a role in the acceptability of SMs? Finally, what accounts for the difference between *at least* and *at most*? The empirical pattern is predicted neither under a strict PPI account nor under a simple account of processing complexity (given that SMs are acceptable in conditionals and universals but not under negation). We conclude that a full account of how SMs differ from CMs is still an open question and needs to be sensitive to semantic, pragmatic, and processing factors.

Selected references: Cohen, A., & Krifka, M. (2014). Superlative quantifiers and meta-speech acts. Linguistics and philosophy. Cremers, A., & Chemla, E. (2016). Experiments on the acceptability and possible readings of questions embedded under emotive-factives. Spector, B. (2014). Global positive polarity items and obligatory exhaustivity. Semantics and Pragmatics. Spector, B. (2015). Why are Class B modifiers global PPIs? Hurford disjunctions as a model for Class B modifiers. Handout at Workshop on Negation and Polarity in Jerusalem.

Evidential bias and polar questions – the division of labour in Hungarian

Cecília Sarolta Molnár, Beáta Gyuris and Katalin Mády
Research Institute for Linguistics, HAS, Budapest

Introduction. The paper discusses the results of two experiments that investigated the evidential bias properties of positive and negative polar question forms in Hungarian.

Previous work. The division of labour between forms expressing positive vs. negative polar questions (PPQ vs. NPQ) have been discussed by Ladd (1981), Büring and Gunlogson (2000), Farkas and Bruce (2010), Krifka (2017), Romero and Han (2004), Reese (2007), Sudo (2013), and van Rooij and Safárová (2003), among others. There is general agreement that the choice between a PPQ and an NPQ in a particular situation is based (at least) on the availability of evidence, the speaker's beliefs, expectations stemming from the norm/rules or what the speaker desires, and the goals of the interaction. Büring and Gunlogson (2000) propose, based on English and German data, that PPQs and NPQs are licensed in the absence of "compelling contextual evidence" (CCE) for the proposition corresponding to the negative and the positive answer, respectively.

Aims and hypotheses. In the current study, which is the first of its kind on Hungarian, we followed the simplest possible design. We concentrated on the influence of CCE on the choice between PPQs and NPQs in contexts that did not to make reference to any type of previous speaker belief but were compatible with all (in view of Arnhold et al. 2016 and Roelofsen et al. 2013 who observed significant interaction between evidential and epistemic biases).

We investigated the choice between positive and negative PQs that can each be realized in terms of two string-identical but prosodically different form types, which are also string-identical to the corresponding declaratives (cf. Gyuris in print for further discussion): i) positive and negative polar *interrogative* form types marked by a final rise-fall tone, with a peak on the penultimate syllable, and ii) positive and negative *declarative* forms that are pronounced with a rise-fall tone on each stressed word, licensed in contexts where English 'rising interrogatives' (Gunlogson 2003) are used. Relevant examples are shown in (1):

(1) a. *Esik az eső?* b. *Nem esik az eső?*
falls the rain not falls the rain
'Is it raining?' 'Isn't it raining?'

The following hypotheses were made:
Hypothesis 1: In a neutral context (i.e. one lacking CCE for any of the answers) the PPQ is preferred to the NPQ.
Hypothesis 2: In the presence of CCE for the positive answer, only the PPQ form is felicitous.
Hypothesis 3: NPQs are only felicitous if there is no CCE for the positive answer.

Materials and methods. The hypotheses were tested in two experiments using 2-alternative forced choice tests. The critical items were presented in writing, which masked the prosodic distinction between PQs expressed by interrogatives and declaratives. In both experiments there was one experimental factor with two levels, and two response types:

	factors	*responses*
Exp. 1	CCE for the positive answer vs. neutral context	PPQ vs. NPQ
Exp. 2	CCE for the negative answer vs. neutral context	PPQ vs. NPQ

Each item consisted of a context description, followed by a PPQ and an NPQ alternative that participants had to chose from, depending on which they would ask in the context. Two lists

were created according to a latin square design, including 16 experimental trials and 32 fillers. Data were collected via an online query form. Each experimental list was filled in by 21 to 45 participants (mean age 38.5 y.), totalling in 752 responses in Exp. 1 and 1168 in Exp. 2. Generalised mixed-effect models with random slopes were applied to the data, evidence as fixed effect and participant and item as random effects.

Results: PPQs were clearly preferred over NPQs (81% of all occurrences in the two experiments). Statistical analysis revealed that positive evidence given by the preceding context did not have an impact on the choice of question type ($p > 0.1$ in both lists of Exp. 1). However, negative evidence in the context increased the preference for NPQs substantially as opposed to the neutral context condition ($p < 0.001$ in both lists of Exp. 2), as shown below.

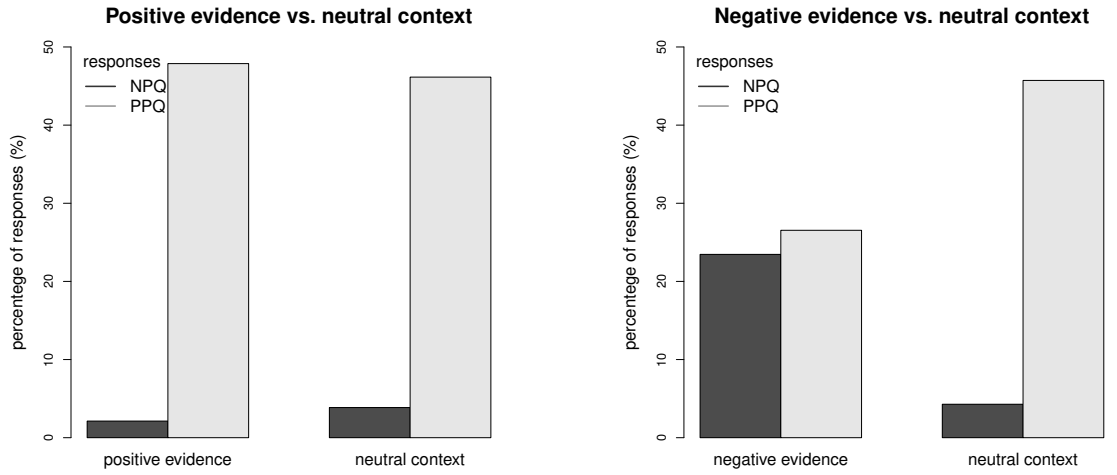


Figure 1: *Percentages of preference for NPQs and PPQs with positive evidence vs. neutral context (left) and negative evidence vs. neutral context (right).*

Discussion. All three hypotheses were confirmed by the data. Due to the fact that participants could interpret the forms in question as 'rising declaratives', the issue of the infelicity of positive forms in the face of positive evidence observed by Roelofsen et al. (2013) did not arise. The fact that in the face of negative evidence, more PPQs (26,54%) than NPQs (23,46%) were chosen comes as a surprise. The paper will discuss possible ways of accounting for it.

References. Arnhold, A. et al. (2016). Syntax and prosody of negative polar questions. Ws. "Questions, Answers and Negation". ZAS Berlin. • Büring, D. and Gunlogson, C. (2000). Aren't positive and negative polar questions the same? Ms. • Farkas, D. and Bruce, K. (2010). On reacting to assertions and polar questions. *JoS* . Gunlogson, C. (2003). *True to Form.* Routledge. • Gyuris, B. (in press) New perspectives on bias in polar questions. A study of Hungarian *-e*. *IRP*. • Krifka, M. (2017). Negated polarity questions as denegations of assertions. In Lee, C. et al. (eds.) *Contrastiveness in Information Structure* . . . Springer. • Ladd, D. R. (1981). A first look at the semantics and pragmatics of negative questions and tag questions. *CLS* 17. • Reese, B. (2007). *Bias in Questions.* Diss., U. Texas. • Roelofsen, F. et al. (2013). Positive and negative questions in discourse. *SuB* 17. • Romero, M. and Han, C.-h. (2004). On negative yes/no questions. *L& Ph* 27. • Sudo, Y. (2013). Biased polar questions in English and Japanese. In Gutzmann, D. and Gärtner, H.-M. (eds.) *Beyond Expressives,*. Brill. • van Rooij, R. and Safárová, M. (2003). On polar questions. *SALT XIII.*

Alignment in Naturalistic Dialogue: Language Production in Interactive Reference Production

Charlotte Out (c.out@uvt.nl)
Martijn Goudbeek (m.b.goudbeek@uvt.nl)
Emiel Krahmer (e.j.krahmer@uvt.nl)
Tilburg center for Cognition and Communication, Tilburg University, PO Box 90153, 5000 LE, The Netherlands

Speakers trying to distinguish one object from others often use referring expressions such as 'the red chair' and 'the large couch'. According to Dale and Reiter (1995)'s Incremental Algorithm, there is a fixed preference order of attributes, in which definite attributes (e.g., color) are preferred over less definite attributes (e.g., size). According to the Incremental Algorithm, speakers will never use a dispreferred attribute when a preferred attribute is sufficient for identification. In contrast, Goudbeek and Krahmer (2012) found that speakers describing pictures of furniture might use a dispreferred attribute (orientation) instead of a preferred attribute (color) when they are primed to do so. Inspired by the Interactive Alignment Model (Garrod & Pickering, 2004), they conjuncture this is due to speakers aligning with their conversational partner by using the same linguistic representations, to make sure conversation goes smoothly.

However, Goudbeek and Krahmer (2012) used a relatively artificial paradigm: speakers interacted with a computer and were primed by a pre-recorded computerized female voice. We aimed to replicate this study creating a more naturalistic setting involving two human participants in naturalistic dialogue.

Following their study, we used pictures depicting furniture items (a fan, a chair, a couch, and a desk) in four different colors (blue, green, red, and grey) and two different sizes (large or small). There were three types of trails: color trials, size trials and filler trials. Both participants view the same pictures, but in a different layout. Participants engaged in a computer task together, taking turns identifying the target picture (accompanied by two distractors) to their conversational partner.

Our experiment went as follows. Participant A describes the target picture (framed by a red border on the screen) to participant B. Depending on the trial, participant A used (was *forced* to use) either a preferred or dispreferred attribute to describe the target picture to participant B. In the color prime, the target picture had a different color (e.g. red) than the distractors (e.g. both green), but the same size (all large). In the size prime, the target picture had a different size (e.g. large) and the distractors (e.g. both small), but the same color (e.g. all green, see Figure 1, square 1).

Second, participant B indicated the matching picture by pressing a key of the corresponding number on their keyboard, e.g. '1' (see Figure 1, square 2). Third, the participants switched roles: now participant B was the director and participant A the

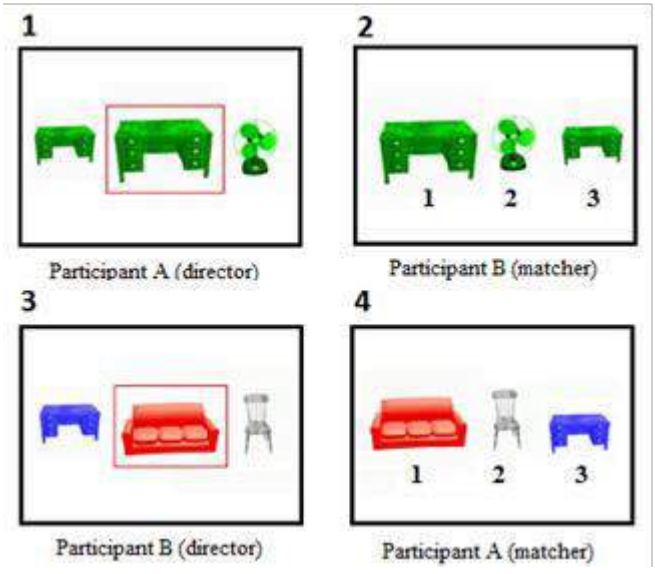


Figure 1. Example of a size trial in the director-matcher task

matcher. In contrast to the previous turn, now the target picture differed in both color and size (e.g., large and red) from the distractors (e.g., one small and blue, one small and grey). This gave participant B the *choice* to use either the preferred or dispreferred attribute to distinguish the target picture from the distractors (see Figure 1, square 3). In case participant B aligned with participant A, they used color when participant A (i.e., in color trials) used the preferred attribute, and size when participant A used the dispreferred attribute (i.e., in size trials). Finally, participant A selects the matching picture (see Figure 1, square 4). In this way, we induce priming or preferred or dispreferred properties (and potentially alignment) in a naturalistic setting.

For our statistical analyses, we used the proportion of attribute use as our dependent measure to create a measurement for alignment, including overspecification (the speaker using both the preferred and dispreferred attributes).

Our results indicate that participants generally preferred to use color (M = .80, SE = .03) over size (M = .52, SE = .03), $F(1, 68) = 33.67$, $p < .0001$, $\eta^2 = .33$. Type of prime had a significant main effect on attribute choice, $F(1, 68) = 47.36$, $p < .0001$, $\eta^2 = .41$. Participants primed with color used the preferred attribute color (M = .85, SD = .21) significantly more than the dispreferred attribute size (M = .31, SD = .28). In contrast to the (statistically non-significant) difference found by Goudbeek and Krahmer (2012), participants primed with size did not show a preference for using size (M = .72, SD = .29) over color (M = .75, SD = .31, see Figure 2) in the size priming condition, but, importantly, they did use size substantially more than in the color priming condition.

In conclusion, we were able to replicate the findings by Goudbeek and Krahmer (2012), showing that regarding referential expressions, speakers do not only align the choice of attributes in their referential expressions when interacting with a computer, but also in a naturalistic interaction with another human.

This experiment is part of a larger project studying the effect of emotion on language production. In future studies we aim to study the underlying mechanism of the language production (of referring expressions) of emotional speakers.

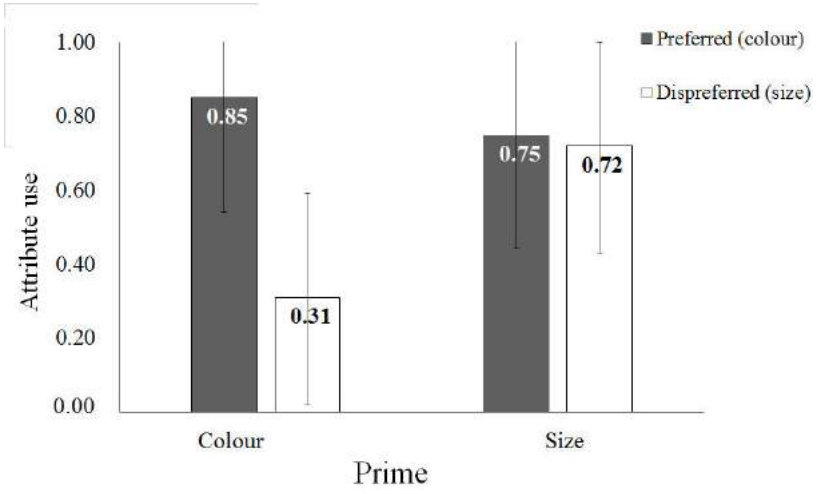


Figure 2. Proportion of preferred and dispreferred attributes per Prime (Color or Size)

References

Dale, R., & Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 19, 233–263.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8, 8–11.

Goudbeek, M., & E. Krahmer (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. *Topics in Cognitive Science*, 4, 269-289.

Pragmatic inferences towards prototypical meanings. A visual world study.

Daniele Panizza, University of Goettingen & John M. Tomlinson Jr., Z.A.S.

When understanding sentences such as "A bee is looking for food on the *field*", a listener must select a specific instance - a hyponym - of *field* from conceptual knowledge (e.g. flower field, grass field, soccer field etc.). One question is how listeners restrict the conveyed interpretation of a simple or broad concept ('a field') towards its most prototypical denotation ('a grass field'). Neo-Gricean frameworks, e.g., view this type of inference as a default inference (M-implicature [1], but see also [2], [3]). Various studies have investigated the how polysemy is processed differently from lexical ambiguity (cf., e.g., [4]), however no study to date has investigated the processing mechanisms underlying pragmatic narrowing, i.e. going from a broad concept to more narrow, specific, in many cases the prototypical one. In eye-tracking study, [5] showed that listeners incrementally enrich the interpretation of scalar adjectives (e.g. 'a tall glass') and exploit the information in the visual context, such as the presence of a smaller glass, to anticipate the identification of the target. The primary goal of the present study is to investigate whether there is a processing bias (i.e. eye gaze preference) for the prototypical interpretation over the non-prototypical one. If there is a bias, a further question is whether this bias is found in both ambiguous and unambiguous scenarios or whether it only emerges in latter scenarios that include two contrasting competitors. If a bias towards the prototype is due to pragmatic narrowing, we expect to find faster prototypical target identification in the scenarios including a referential competitor. In contrast, if both ambiguous and unambiguous conditions display such an effect, the bias should be attributed to greater conceptual/lexical association between the event described in the sentence and the prototypical picture (e.g. a bee is more likely to look for food on a flower field)., We also included sentences involving lexically ambiguous homonyms (2) to explore whether they display difference in processing with respect to hyponyms. In a visual world experiment based on [5], forty-five participants identified referents for sentences such as (1) and (2) in two kinds of visual scenarios: an ambiguous scenario including two possible referents for the critical word (e.g. a soccer field and a flower field, as in fig.1) vs. an unambiguous scenario (fig.2) with only one possible referent (e.g. a flower field). Each sentence the speech stream was divided into 500 ms time windows, time-locked at the noun (*field*) and disambiguation information (*flower/soccer*).

(1) A bee is looking for food on the *field*... (a) it's a flower field / (b) it's a soccer field.

(2) A boy saw the *boxers* in front of the hall... playing with each other/as they finished training.

Prototypes vs. non-prototypes. As predicted, the targets in unambiguous scenarios were identified much earlier than in ambiguous ones (fig.3, 4). Main effects of *ambiguity* were found in five consecutive time windows starting one and a half seconds prior to disambiguation (2.5s: $p<.001$). *Prototypicality* did not affect the identification of the target. In contrast, targets in ambiguous scenarios were identified 500 ms before the disambiguation point in prototypical hyponyms (fig. 4) but not in the non-prototypical ones (3.5s: $p=.03$). This resulted in a main effect of *prototypicality* in the disambiguation time window (4s: $p=.01$) as well as an interaction between *ambiguity* and *prototypicality* in the time regions immediately before (3.5s: $p=.02$) and after (4s: $p<.01$) the disambiguation. **Hyponyms vs. Homonyms.** In the overall analysis (fig 4), main effects of *ambiguity* were found starting one second after the onset of the sentence (1s: $p<.004$, 1.5s: $p=.01$, 2.s: $p<.01$, etc.). While with unambiguous scenarios targets were identified more quickly for homonyms than in hyponyms (main effects of *kind* at 1.5s ($p=.04$) and 2s ($p=.07$)), hyponyms in ambiguous scenarios where disambiguated more quickly than homonyms, as shown by main effects of *sentence type (hyponym vs. homonym)* after the disambiguation (5s: $p<.01$; 5.5s: $p=.04$) but with opposite directionality. **Discussion.** Overall, prototipicality resulted in anticipated target disambiguation in ambiguous visual scenarios (i.e. flower field vs. soccer field). Participants showed a strong bias towards for the prototypical hyponym (flower field*)* vs. the less prototypical one (soccer field). Critically, this effect was selective for ambiguous scenarios, similarly to what reported by studies where referential ambiguity was affected by pragmatic inferencing ([5]). Thus, listeners were able to incrementally assess conceptual knowledge of an event to resolve referential ambiguity and committed to the prototypical interpretation before to actually hearing the disambiguation. The finding that homonyms were disambiguated more quickly with unambiguous scenarios,

whereas, hyponyms were disambiguated more quickly with ambiguous scenarios, also supports this explanation, as well as ruling out alternative explanations such as higher probabilistic association or visual saliency of prototypes.

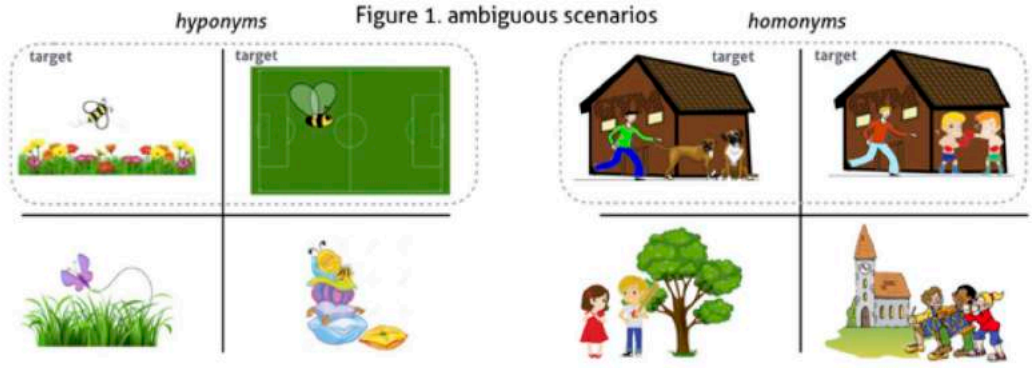


Figure 1. ambiguous scenarios

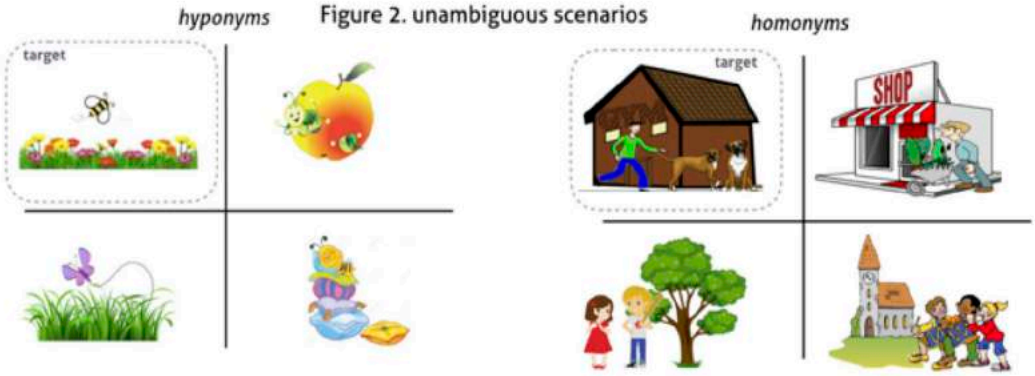


Figure 2. unambiguous scenarios

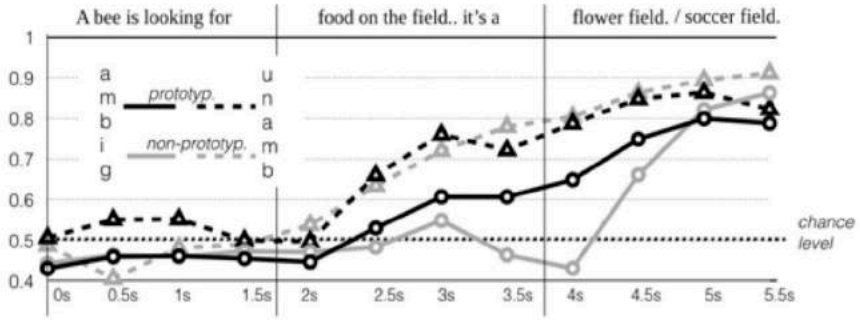


Figure 3. Target preference (T/T+D) in hyponym conditions

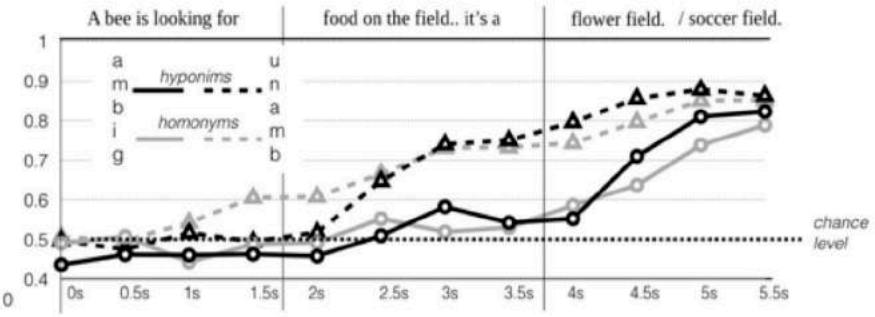


Figure 4. Target preference (T/T+D) for hyponyms vs. homonyms

References. [1] Horn, L. (1999). Toward a new taxonomy for pragmatic inference. In D. Schiffrin (Ed.), Form and use in context: Linguistic applications. [2] Levinson, S. (2000). Presumptive meanings. [3] Wilson, D. (2003) Relevance and Lexical Pragmatics. Italian J. of Ling. [4] Klepousniotou E. (2002). The processing of lexical ambiguity: homonymy and polysemy in the mental lexicon. Brain Lang. [5] Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G. & Carlson, G. N. (1998). Achieving incremental semantic interpretation through contextual representation. Cognition.

Myrto Pantazi[ab], Mikhail Kissine[a], Olivier Klein[b]

[a]Center of Research in Linguistics LaDisco, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, 1050, Brussels, Belgium

[b]Center for Social and Cultural Psychology, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, 1050, Brussels, Belgium

Automatic Content Accommodation: Direct Perception and Meta-Cognitive Vigilance

An important debate in linguistics, philosophy of language and experimental psychology is how we validate the content of statements we understand. Are we capable of assessing it and filtering it out, in case it is erroneous? Or do we rather tend to automatically believe it?

We present 4 experiments testing whether participants can disbelieve the content of statements that are explicitly presented as false. In Experiment 1 we asked participants to listen to statements about two ostensible judicial cases in the form of short crime reports. The participants were informed that the two reports contained both true (e.g. female speaker) and false (e.g. male speaker) statements, as indicated by the voice of the speaker. Crucially, the content of the false statements in the one report was aggravating the crime described while the content of the false statements in the other report was attenuating the crime described. Participants had a strong tendency to judge the "aggravated" perpetrator in a more severe manner compared to the "attenuated" perpetrator. Additionally, in a memory test, participants misremembered more false statements as true than true statements as false, showing that we have a pervasive tendency to believe statements' content.

In Experiment 1, the true statements in the reports outnumbered the false ones, just like in real life most of the utterances we hear are expected to be truthful (see Grice's, 1975, Maxim of Quality). One could conjecture, thus, that participants' tendency to believe the false statements presented in Experiment 1 was largely due to the fact that participants were in a context where most of the statements were true. In Experiment 2 we rendered the number of the true and false statements in the reports equal. Still, Experiment 2 replicated the above-mentioned results, even in a context where the tendency to believe statement content is not ecologically valid.

In the last two experiments we tested two factors that might potentially increase participants' vigilance, and reduce their pervasive tendency to believe statements' content: accountability (Experiment 3) and financial incentives (Experiment 4). Participants who were informed that they would have to account for their judgments in our experimental setting, still, tended to be influenced in their judgments by the false statements, as well as to misremember the false statements as true. Nevertheless, offering participants financial incentives for accurately judging the perpetrators, eliminated these tendencies.

The present findings go against widespread post-Gricean pragmatic theories (Sperber & Wilson, 1995) assuming that linguistic communication passes by complex meta-representations of the speakers' communicative intentions. We posit a *Direct Perception Mechanism* (DPM) entrenched in the statement comprehension process (e.g. Millikan, 2005; Recanati, 2002). Specifically, we argue that addressees tend to automatically accommodate the contents of statements they hear and read in a way that resembles visual perception. Such a DPM is particularly evolutionarily plausible if language is viewed as a mechanism evolved to facilitate

information exchange (Jackendoff & Pinker, 2005; Pinker & Jackendoff, 2005) among cooperative agents (Kissine & Klein, 2013).

Nevertheless, in those contexts where statements' content is inaccurate, the DPM is inadequate. Experiments 3 and 4 suggest that while it is hard to disbelieve statements we hear and read, there are cases where people can be sufficiently vigilant. However, we argue that the operation of vigilance is costly and consists in effortful meta-cognitive processes, which is the reason why the tendency to believe is relatively hard to override. We will present a fully-fledged cognitive model of statement validation, defining the conditions under which the DPM operates, and those where people will be vigilant and manage to filter out statements' content.

References

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

Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition*, *97*(2), 211–225. http://doi.org/10.1016/j.cognition.2005.04.006

Kissine, M., & Klein, O. (2013). Models of communicatoin, epistemic trust and espistemic vigilance. In J. Laszlo, J. Forgas, & O. Vincze (Eds.), *Social Cognition and Communication*. New York: Psychology Press.

Millikan, R. G. (2005). Language: A Biological Model. *Language: A Biological Model*, 1–240. http://doi.org/10.1093/0199284768.001.0001

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition*, *95*(2), 201–236. http://doi.org/10.1016/j.cognition.2004.08.004

Recanati, O. I. S. (2002). Does Linguistic Communication Rest on Inference ? *Mind & Language*, *17*(1–2), 105–126.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Exploring how speakers mark, and listeners assess, certainty

Amanda Pogue (apogue@ur.rochester.edu) & Michael K. Tanenhaus
Brain and Cognitive Sciences, University of Rochester

Successful communication would seem to require speakers to signal their degree of certainty about an utterance through lexical choices and prosodic markings (e.g., It's a dog, I think/THINK it's a dog) and listeners to successfully use these signals and adapt to variations in how different speakers signal uncertainty and whether a speaker is likely to overestimate or underestimate how certain she should be. We are beginning address a range of questions about how uncertainty is marked by speakers and inferred by listeners, and how listeners make use of this information when calibrating specifically or generally from a speaker. First, we asked whether speakers have conscious access to features of productions that can mark certainty (Pre-task). Second, we ask whether listeners have a stable preference for different lexical structures that mark uncertainty, and whether degree of uncertainty is indicated by different production cues (Experiment 1). Third, we introduce a task that asks speakers to produce labels for objects they are uncertain about, in order to see how uncertainty influences the utterances speakers choose for communicating their labels to an interlocutor (Experiment 2). Finally, we discuss a new project investigating how uncertainty expressions affect an interlocutors' choice of a name versus a description in a referential communication task.

We consider the possibility that the speaker might have two motivations for marking uncertainty in their productions: first, to appear to be producing accurate information, and second, to signal to the listener that they are a reliable speaker. These two motivations are motivated by Grice's (1975) principles of cooperation, such that if speakers want to be seen as cooperative, they should want to provide only the information they know to be true, or in this case, to do so with certainty, only when certain of the truth. Secondly, we consider the possibility that some speakers have better access to their own production choices for marking uncertainty than others. In the current tasks, and in our future research we aim to explore how speakers mark uncertainty, whether by lexical choice, prosodic contour, disfluencies, hesitations, etc, and how interlocutors use these uncertainty markings. Our goal is to extend beyond the current work in the field suggesting that listeners are sensitive to cues to speaker knowledge (see: Smith & Clark, 1993; Brennan & Williams, 1995; Swerts & Krahmer, 2005), to show that this ability allows listeners to adjust their expectations for an interlocutors' likely referential knowledge.

Pre-task: We predicted that speakers would modulate their speech when they were asked to mark uncertainty compared to just reading aloud the same sentences. Eight adult speakers of American English were recruited from the University of Rochester community, and were told that they would be making recordings for a future study. They were randomly shown individual phrases about birds that used a variety of possible hedges for uncertainty, and were asked to read each one out loud. Then they were told to imagine that they were in a task where they briefly saw the birds, but that they might be unsure that they had correctly seen the bird, either because it was displayed too briefly, or because it was partially occluded. They were shown the same phrases in random order again, and were asked to speak them out loud. We predicted that if speakers are aware of how to explicitly mark uncertainty that listeners would give lower certainty ratings when the speaker was told to produce the utterance as if she were unsure.

Experiment 1: Participants on Amazon's Mechanical Turk were randomly assigned to a read-text (n=16) or listen (n=160) task. They were told that previous workers were asked to describe pictures of birds, but that sometimes the pictures flashed quickly or were not fully visible. Their task was to rate how certain they thought the speaker was on a slider scale of 1 to 100 (1= not at

all certain, 100=completely certain). Participants either read each of the 8 target sentences plus 2 control sentences (to ensure that they were paying attention), or heard the sentences. Each participant in the listen condition heard all 8 of the target sentences; sentences were randomly ordered, and the listener heard one of 16 possible productions (8 speakers x 2 certainty levels) for that sentence. After completing these ratings, participants were asked to rank order the 8 sentences in order of certainty. **[Results]** Regardless of condition we find a stable order of the rated certainty of the utterances, and their rankings. Rank orderings, and overall ratings are provided in Table 1. We also looked at the relative differences between the ratings for each of the speakers between the read, and uncertainty emphasis productions and found that for 7/8 of the productions listeners on average rated the speakers as more uncertain when they had been instructed to mark uncertainty. We also found individual differences in the amount of certainty conveyed. Overall some speakers averaged more uncertainty between the types of productions (mean differences of 7-10 points on the certainty scale, max: 30-40 points), whereas others overall showed less difference, or even sounded more confident when they should have marked uncertainty (mean differences of 0.5-1, or an increase in certainty by 5, max: 3-11).

Phrase	Read-text Confidence	Listen (Read)	Listen (Uncertainty)	Read-text Rank	Listen Rank	Experiment 2 (mean confidence)
It could be a goose	36.994	37.706	36.283	7.125	7.063	25.163
It might be a robin	39.294	41.094	37.494	6.375	6.375	28.798
I think it's a falcon	49.918	48.918	50.919	5.688	5.644	46.458
It looks like a hummingbird	57.080	61.362	52.797	5.25	5.381	45.828
I'm pretty sure it's a woodpecker	65.476	68.110	62.842	4.063	4.319	68.577
I'm sure that it's a sparrow	84.220	87.510	80.930	2.688	2.919	80.300
It's a blackbird	86.777	88.864	84.689	2.625	2.525	91.765
It's definitely a canary	90.935	90.246	91.624	2.188	1.775	93.192

Table 1: Results from Experiment 1 and Experiment 2

Experiment 2 directly manipulated the likely degree of uncertainty. Participants (n=32) on Mechanical Turk were asked to label objects, rate their confidence in their label, and then selected one of 8 possible phrases they would use to describe that item to another person. We used line drawings from a classic perceptual recognition study (Biederman, 1987) in which parts of the images were occluded in a way that either did or did not preserve the underlying components (geons) of the image. We manipulated exposure duration, 120 or 220 ms, presented with a random dot mask (to avoid afterimage completion). Participants saw 17 images, and 3 control images (complete, simple pictures) randomly presented. **[Results]** Participants were more confident in their labels when items were presented for longer durations (p <.02), and when the deleted information preserved geons (p <.001). We also found a relationship between the confidence ratings and the kind of phrase participants used to communicate what they saw to another person (see: Table 1). Speakers' certainty in their own label, resulted in similar ordering of the phrases (7/8), as those determined by the listeners in Experiment 1.

We are now using utterances modeled on the ones we have tested in a confederate study in which a naïve participant and the confederate learn rare and common names for dogs and kitchen utensils together based on Ibarra, Runner & Tanenhaus (2017), who found that judgments of relative expertise modulated a directors use of a name versus a description following a shared learning task. The confederate will use expressions that indicate greater or lesser uncertainty. We predict that the confederate's degree of uncertainty will affect the (naïve) director's item-specific and category-based use of names in a subsequent referential communication task.

References: http://amandapogue.github.io/docs/PogueTanenhausXPragReferences.pdf

At-issue: Non-Restrictive Relative Clauses

Claudia Poschmann / Goethe-Universität Frankfurt a.M.

Outline: Recent studies (e.g. Schlenker 2009, Koev 2013, Jasinskaja 2016) have challenged the view that non-restrictive relative clauses (NRRCs) are inherently projective and non-at-issue (e.g.McCawley 1982, Potts 2005, Simons et al. 2010). This talk presents the results of two experiments in German on NRRCs embedded in *if*-clauses. The results confirm a claim of Schlenker (2009) that NRRCs can contribute conjunctively to the at-issue meaning of their matrix clause giving rise to embedded readings. This embeddability is dependent on position, discourse structure and other pragmatic factors.

Experiments: In a first questionnaire, with 62 German native speakers and 18 items, we tested the availability of embedded readings depending on the CLAUSE-TYPE of the embedded construction (NRRCs, "and"-conjunctions, V2-parenthesis) and the PREDICATE TYPE (event vs. state). Each item consisted of a little context-story and a target sentence. The participants had to judge whether the target sentence was appropriate as part of a summary of the information given by the story. The stories were constructed such that both the wide-scope reading and a potential modal subordination reading of the target sentences were explicitly ruled out. For example in (1) it is unclear, whether Gerd can be saved even if he reaches Dr. Meier, since we don't know whether Dr. Meier has got the right anti-dot available. Thus, if the participants only got a wide-scope or modaly subordinated reading (similar to (1-d)), according to which Gerd is saved as soon as he reaches Dr. Meier (because in this case Dr. Meier will for sure inject him the right anti-dot), they were expected to reject the target as part of a summary of the context. Only if the participants interpreted the NRRC as contributing conjunctively to the antecedent of the *if*-clause (such as the conjunction in (1-b)), were they expected to accept the target sentence as an appropriate summary of the context-story. (1-a) to (1-c) give an example for a test item with event-predicate type in the three clause-type conditions of the first experiment. In a second follow-up experiment with 22 participants and 12 items, we directly compared the interpretation of sentence-internal NRRCs (1-a) and the corresponding matrix clauses in sentence-final position ((1-d)), again each with event and state predicate.

(1) **Story:**
Gerd wurde von einer Schlange gebissen und hat nur wenig Chancen zu überleben. Denn das Gift wirkt schnell tödlich. Wenn überhaupt, kann er nur noch Dr. Meier erreichen, der ganz in der Nähe wohnt. Ob dieser jedoch über das äußerst seltene Gegengift verfügt, ist mehr als ungewiss. Nur falls Dr. Meier ihm noch rechtzeitig das richtige Gegengift verabreicht, kann er gerettet werden. *(Gerd got bitten by a snake. There is only little chance that he will survive. The dot is quite deadly. His only chance is to reach Dr. Meier in time, who lives close by. But its quite unlikely that Dr. Meier has got the anti dot, Gerd needs. Only if Dr Meier gives him the anti dot in time, can Gerd be saved.)*
Target-Sentence:

a. Wenn Gerd rechtzeitig Dr. Meier erreicht, der ihm das passende Gegengift verabreicht, kann er gerettet werden. *(If Gerd reaches Dr. Meier in time, who gives him the right anti-dot, can he be saved)*
b. Wenn Gerd rechtzeitig Dr. Meier erreicht und der ihm das passende Gegengift verabreicht, kann er gerettet werden. *(If Gerd reaches Dr. Meier in time and*

he gives him the right anti-dot, can he be saved)

c. Wenn Gerd rechtzeitig Dr. Meier erreicht (der verabreicht ihm das passende Gegengift), kann er gerettet werden. *(If Gerd reaches Dr. Meier in time (he gives him the right anti-dot), can he be saved)*

d. Wenn Gerd rechtzeitig Dr. Meier erreicht, kann er gerettet werden. Er gibt ihm das passende Gegengift. *(If Gerd reaches Dr. Meier in time, can he be saved. He will give him the right anti-dot.)*

Results: The results of both experiments indicate that NRRCs with event predicate can indeed be interpreted as truly embedded. In the first experiment, we found a highly significant effect of CLAUSE TYPE ($p < 0.001$) as well as a significant effect of PREDICATE TYPE ($p < 0.001$). NRRCs with event predicates got overall acceptance rates about 49 percent, lower than the corresponding and-conjunctions (0.92), but significantly higher than the corresponding matrix-clause-parenthesis (0.21). NRRCs with state predicate, by contrast, rated nearly as low (0.25) as the corresponding matrix-clause parenthesis. A highly significant contrast ($p < 0.001$) between NRRCs with event predicate and the corresponding matrix clause parenthesis indicates that the observed embeddability is not only a discourse effect or a last resort repair strategy but the result of a structural embedding of the NRRCs. The results of the follow-up experiment confirmed these effects. The NRRC with event predicates rated significantly ($p < 0.001$) higher (0.51) than those with state predicates (0.29) and significantly higher than the postponed matrix clauses (0.09), on which a variation of the predicate type had no effect.

Analysis: The findings challenge the assumption that NRRCs are inherently projective and non-at-issue. We will briefly sketch an analysis according to which NRRCs are always attached low to their head-DP by a tentative relation that is locally abstracted from. If the NRRC is in situ, this relation is projected to the matrix-level, where it is instantiated by a suitable discourse relation. If the NRRC is extraposed, the NRRC is moved from its DP-modifying position, where it leaves a trace, to the right edge of a clause, where the trace is bound and at the same time the missing connective is instantiated by conjunction. In the in situ case, the NRRC is interpreted as an independent speech act and various factors such as the position of the NRRC (Koev 2013) and discourse structure (Jasinskaja 2016) will decide whether this speech act is currently at-issue or not. In the latter case the NRRC is interpreted as contributing locally to the at-issue-content of the matrix clause. We will discuss two options why the predicate-type of an NRRC might affect the availability of embedded readings: (i) An NRRC can be interpreted semantically with low scope only if its proposition is anaphorically dependent on the proposition expressed by the matrix clause. Coordinating discourse relations make anaphoric use of an event described in a preceding proposition. (ii) Event predicates allow the NRRC to stand in a coordinating discourse relation (Asher/Lascarides 2003) to the proposition expressed by the matrix clause and are thus more easily conjoinable to the matrix proposition than NRRCs with subordinating discourse relations. We are evaluating a third experiment, designed to disambiguate between the two options.The results will be presented, too.

References: Asher, N. & Lascarides, A. (2003). Logics of Conversation. Cambridge University Press./ Jasinskaja, K. (2016). Salience and (not-)at-issue status of subordinate clauses. In: Proceedings of SuB 2016, 111-112./ McCawley, J.D.(1982). Parentheticals and discontinuous constituent structure. In: Linguistic Inquiry, 13(1): 91–106. / Koev, T. K. (2013). Apposition and the Structure of Discourse. PhD thesis, Rutgers./ Potts, C. (2005). The logic of conventional implicatures. Oxford University Press./ Schlenker, P. (2009). Supplements without bidimensionalism. In: Proceedings of the Amsterdam Colloquium. / Simons, M. et al. (2010). What projects and why. In: Proceedings of SALT, 20, 309–321.

Scalar implicatures in non-cooperative contexts

Nausicaa Pouscoulous (UCL) *Giulio Dulcinati (UCL)*

Grice's account (1989) does not make direct predictions about non-cooperative situations. However, since hearers cannot expect non-cooperative speakers to be *as informative as is required*, we can assume that this account would predict that they should not infer quantity implicatures. In our study we tested how lack of cooperation affects scalar implicature derivation with a paradigm similar to the one used by Bonnefon, Feeney and Villejoubert (2009). We designed our experiment to discriminate between three hypotheses of how the lack of cooperation in the speaker might affect the implicatures drawn by the hearer:

1. For hearers, there is no difference between a cooperative and a non-cooperative speaker with regards to drawing and accepting implicatures (i.e. our null hypothesis).
2. Hearers draw less implicatures from a non-cooperative interlocutor compared to a cooperative one (i.e., the prediction we derived from the Gricean account).
3. Hearers infer implicatures from cooperative and non-cooperative speakers to the same extent but they are more likely to reject them in the case of non-cooperative speakers (cfr. Sperber et al. 2010).

We constructed five short stories (70-100 words) describing situations in which the reader is given the perspective of a character. The reader's character needs to know a piece of information and they ask another character in the story. For example, in one situation they are about to take an exam in the company where they work and they *don't remember whether you need to need to answer all of the open questions in order to pass*, so they ask a colleague who has just taken the exam. We constructed a cooperative version and a non-cooperative version of each story. In the cooperative version the character is motivated to help the reader (e.g. everyone in the company gets a bonus if enough employees pass the exam), whereas in the non-cooperative version the character benefits from the reader being ignorant or misinformed (e.g. it's a competitive selection exam and only one person can pass). In both versions the character answers with a statement containing the expression *some* which may give rise to a scalar implicature (e.g., *Some of the open answer questions must be answered*). Each story was followed by three yes-no questions: an epistemic question (e.g., *Given what she told you, do you think it's possible that all of the open questions must be answered?*); a meaning question (e.g., *Do you think she meant that you don't need to answer all the open questions?*) and a deception question (e.g., *Do you think she was trying to mislead you?*).

In our internet-based experiment 425 native English speakers were randomly assigned to read only one version of one of the five stories. We analysed their yes-no responses to the three follow-up questions (**Fig. 1**). In the epistemic question, participants in the non-cooperative condition answered yes (i.e., thought that 'all' could be the case) significantly more than participants in the cooperative condition ($X^2(1, N=425)=36.42$, $p<0.001$). Hypothesis 1 would have predicted no difference between the two conditions in the responses to the epistemic question, so this result indicates that there is some difference in interpretation or acceptance of scalar implicatures between the cooperative and non-cooperative condition. In the meaning question, there was no significant difference in the rate of yes answers (i.e., thinking that the character intended to communicate that 'not all' was the case) between participants

in the cooperative and non-cooperative condition ($X^2(1, N=425)=0.55$, $p=0.46$). Hypothesis 2 would have predicted participants in the non-cooperative condition to infer less implicatures than in the cooperative condition and therefore to have lower rates of yes responses in both the epistemic and the meaning question. Hypothesis 3 instead is consistent with this pattern of results as it did not predict participants to daw less implicatures in the non-cooperative condition but it predicted them to reject the content of the implicatures more often than in the non-cooperative condition. Finally, in the deception question, participants in the non-cooperative condition answered yes (i.e., though that the character was trying to mislead them) significantly more than participants in the cooperative condition ($X^2(1, N=425)=67.61$, $p<0.001$). This last result fits how participants responded to the epistemic and meaning questions and suggests that participants in the non-cooperative condition recognized the intended scalar implicatures as false implicatures (Meibauer, 2014). In conclusion, the results of our experiment indicate that, other things being equal, hearers infer implicatures to the same extent from cooperative and non-cooperative speakers, but they are less likely to accept the content of the implicature from non-cooperative speakers. This conclusion is not consistent with the prediction we derived from Grice's (1989) account but it is consistent with the model proposed by Sperber et al. (2010) of how the interpretation process interacts with epistemic vigilance towards the source of the information. Our results echo findings by Mazzarella, Trouche, Mercier and Noveck (2016) on the effect of politeness on the derivation of scalar implicature.

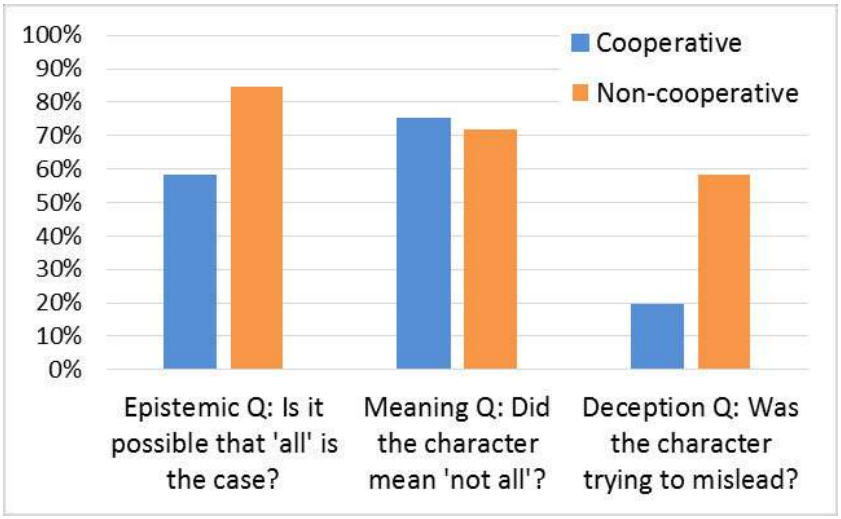


Fig.1 *Frequency of yes responses to each of the three follow up questions*

References:

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

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

Mazzarella, D., Trouche, E., Mercier, H., & Noveck, I.A. (2016). Believing what you are told: Politeness and scalar inferences. *Poster presented at AMLaP, Bilbao, Spain*

Meibauer, J. (2014). *Lying at the semantics-pragmatics interface*. Berlin: Mouton de Gruyter.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language 25(4)*, 359-393.

Investigating shared representations in implying and inferring

Alice Rees & Lewis Bott, Cardiff University

Utterances communicate much more than the literal meaning of the words. Consider the following exchange: A: "I hear Helen's husband is rich and intelligent." B: "Well, he's rich." Speaker B *says* that Helen's husband is rich, but *implies* that he is not intelligent. The linguistic and psychological mechanisms underpinning implicatures like these have been intensely studied since Grice (1975) (e.g. Bott & Chemla, 2016; Breheny, Ferguson, & Katsos, 2013; Chierchia, 2004; Noveck, 2001). However, this research has focused predominantly on the listener, and not the speaker. Here we present two experiments that investigate implicature *production* and test whether the psychological procedures involved in implying overlap with those involved in inferring.

Inferring and implying are different processes. Inferring is carried out by the listener and involves going from speech to meaning, whereas implying is carried out by the speaker and involves going from meaning to speech. This can also be seen by noting that the standard Gricean account of how implicatures are derived (by the listener) must be substantially adapted to before it makes sense from the perspective of the speaker (e.g. how can the speaker generate alternatives to their own utterances?). However, while there must necessarily be some differences between implying and inferring, there may be overlap. Some procedures may be used in both directions even if the system as a whole is different.

If implicature processes overlap, a listener who infers might subsequently be likely to imply. In other words priming of implicatures might exist between interlocutors (as with other linguistic structures, e.g. Pickering & Ferreira, 2008). To test this we adapted a dialogue-based communication game (see Branigan et al, 2000) for implicatures.

Overview.

A participant and a confederate took turns describing and identifying a referent card from a set of four, each of which included one or two images. The structure of the images is shown in Figure 1. Experimental trials referred to either the A or AB card. Crucially, one item was duplicated across these cards (a pencil in Figure 2). Thus, if the speaker used an unmodified expression to describe the A card, "The card with the pencil," they relied on the listener to derive an inference to disambiguate the referent (since the speaker did not say pencil and book, they must mean only the pencil). Alternatively, they could use an explicit modifier, "The card with just the pencil". The **dependent measure** was whether the participant (as speaker) chose an unmodified form (an *implicit* construction) or a modified form (an *explicit* construction).

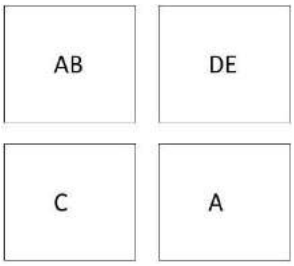


Figure 1. Image structure

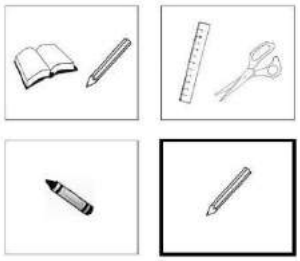


Figure 2. Example trial

We implemented a priming manipulation using the confederate. There were two forms of priming: global and local. Global priming was between subjects. For one group, the confederate used predominantly implicit constructions, and for the other, she used predominantly explicit constructions. Local priming was within participants. Here, the confederate used an implicature on some trials but not others, and the DV was the rate of implicatures on subsequent trials. If there are overlapping inference and implication processes, the rate of implicatures used by the participant should depend on that used by the confederate, both globally and locally.

Experiment 1. N =35

Global priming. Participants adopted the conversational style of their partner (Figure 3). When their partner was using

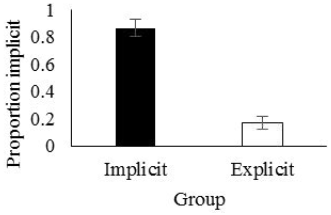


Figure 3. Proportion of implicit responses in each group

implicatures, participants were more likely to also use implicatures ($F(1,31)=125.11$, $p < .001$).

Local priming. We also manipulated trial-by-trial use of implicatures by the confederate. Participants were more likely to use an implicature immediately after the confederate had used one than when she did not, $F(1,31)=8.08$, $p=.008$. These findings support the hypothesis that there are overlapping processes between implying and inferring.

Experiment 2. N=35. Experiment 1 used a confederate as the interlocutor. However, we have no way of knowing whether participants believed our deception. Our results could therefore be a consequence of participants believing that the conversational partner was an experimenter. In Experiment 2 we tested this by manipulating whether the partner was presented as an experimenter or another participant. The basic design was exactly the same. The only difference was that one group of participants were told that their partner was an experimenter and the other group was not.

Partner role. We found no significant differences between when the partner was presented as an experimenter or as a confederate ($F(1, 36) = 1.13$, $p = 30$).

Global priming. We replicated the findings from Experiment 1. Participants in the implicit condition produced more implicit utterances than those in the explicit condition ($F(1, 36) = 45.72$, $p < .001$, 95% CI = 1.97 – 3.65). The global priming effect was significant in the experimenter ($F(1, 16) = 19.25$, $p < .001$, 95% CI = 1.53 – 4.39) and the confederate condition ($F(1, 16) = 30.06$, $p < .001$, 95% CI = 1.65 – 3.68).

Local priming. As in Experiment 1, the local priming effect was significant, $F(1, 32) = 6.64$, $p = .015$). However, the results were only marginal when considering experimenter and confederate condition separately ($F(1, 6) = 4.18$, $p = .058$; $F(1, 16) = 3.01$, $p = .100$).

Overall, we replicated the global and local priming effects from Experiment 1 and we found no evidence that the findings were dependent on whether the participants believed that the partner was a confederate.

Conclusion. We found that participants can be primed to produce implicatures in dialogue. This suggests that there are overlapping processes in implying and inferring. When a listener infers, they activate implicature processes, which in turn makes it more likely that those processes will be used in subsequent production.

Previous research has suggested that implications are produced as a consequence of socio-pragmatic factors such as politeness and efficiency (Holtgraves & Yang, 1990; 1992; Levinson, 2000). These factors cannot explain our results because in Experiment 1, we did not manipulate any social factors and participants systematically varied their choice of construction across conditions, and in Experiment 2 we manipulated the social status of the conversational partner but found no difference in rates of implicature production. Our findings therefore suggest sociopragmatic factors are only one source among many that influence the decision about whether a speaker is implicit or explicit.

More generally our study takes some initial steps into understanding implicature production. The nature of these processes and the extent to which comprehension and production overlap in pragmatics are interesting topics for the future.

Selected References

Bott, L. & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language.*

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

Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and Beyond.* Oxford UK: Oxford University Press.

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

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*, 427–459.

A new Director task: Modelling common ground through referential specificity

Paula Rubio-Fernández (Massachusetts Institute of Technology; *prubio@mit.edu*) &
Julian Jara-Ettinger (Yale University; *julian.jara-ettinger@yale.edu*)

Over two decades, the Director task (DT) has increasingly been employed as a test of the use of Theory of Mind in communication, first in psycholinguistics and more recently in social cognition research. In this task, a participant follows the instructions of a confederate Director to move around various objects in a vertical grid of squares. The confederate sitting on the other side of the grid is ignorant of the contents of some of the cells because they are occluded on her side. A long series of studies has revealed that participants suffer ***interference*** from their privileged perspective when interpreting the Director's instructions (e.g., Keysar et al., 2000, 2003; Barr, 2008; Lin et al., 2010).

Keysar and colleagues and more recently social cognition researchers (e.g., Apperly et al., 2010; Dumontheil et al., 2010a, 2010b) have interpreted the poor performance normally observed in the DT as evidence of ***'limited use of Theory of Mind in communication'***. In this paper we challenge that conclusion and argue instead that the design of the DT itself is what limits participants' perspective-taking abilities, by imposing artificial demands on their selective attention. While seemingly uncomplicated, the design of the DT forces participants to supress a universal assumption in human communication: namely, that people know more than what they can see and are therefore able to refer to entities outside their visual field – unlike the Director.

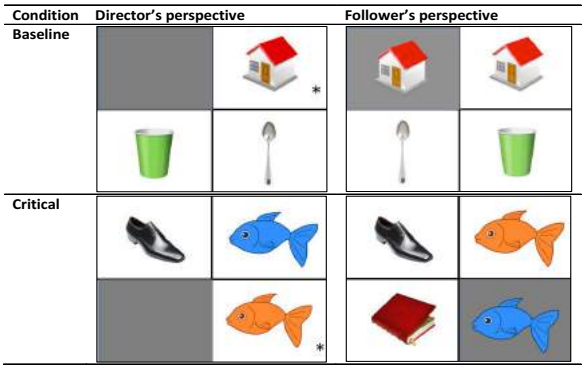
However, exclusively focusing on the objects that the Director can see in the grid need not be evidence of Theory of Mind use. In fact, the standard metrics of interference used in the DT cannot determine the extent to which participants' performance is dependent on Theory of Mind or selective attention. In other words, participants may suffer interference from their own perspective because they are failing to adopt the Director's perspective, or because they do not have enough executive control to inhibit their own. Thus, the aim of this study was to challenge the key assumption in the DT: that ***when participants consider the hidden objects as possible referents for the Director's instructions, they need not be failing to use their Theory of Mind***.

Pairs of naive participants played a new DT on two computers showing 2×2 grids of objects. One of the four cells had a grey background and contained an object in the Follower's grid, but was empty in the Director's (see Table 1). The Critical trials included a subtle manipulation: unbeknownst to the participants, the position of the grey cell was shifted in the Director's grid, so that a figure that appeared on a grey background in the Follower's grid now appeared on a white background in the Director's. The Director would see two fish of different colours, for example, and ask the Follower for 'the orange fish'. However, in the Follower's grid the blue fish appeared on a grey background, thus inviting the question: ***if the Director cannot see the blue fish, why does she call the target 'the orange fish', and not just 'the fish'?***

A direct measure of suspicion was collected in a post-test questionnaire and participants' fixations on the grey cell were taken as an indirect measure of suspicion. Only 4 participants (19%) reported not to have noticed anything peculiar in the Director's instructions. As predicted, those participants also said they had tried to focus their attention on the three white cells and block the grey cell from their view. According to the standard metrics of interference, those 4 participants would be 'model perspective takers'; however, those participants were actually underusing their Theory of Mind. By contrast, 16 participants (76%, sig. above chance) were suspicious that the Director sometimes knew about the contents of the grey cell because she used colour to distinguish the target. Also as predicted, those participants paid increasing attention to the grey cell in the second half of the task, when critical trials were administered and they grew suspicious of the Director's perspective. LMM comparisons confirmed that the suspicious participants' fixations on the grey cell

had a U-shaped distribution, first decreasing because of practice and then increasing because of suspicion (see Figure 1).

The results of this study confirm that ***the DT is not a reliable test of Theory of Mind use in communication***. According to the standard metrics of interference, 'optimal performance' is possible by using selective attention alone, while in this study the most sophisticated Theory of Mind use was revealed by those who kept track of the hidden cell.



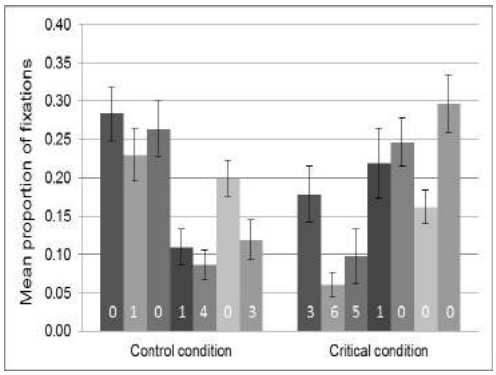


Table 1 (left): Sample trials from the Baseline condition and the Critical condition. The asterisk indicates the target object. The position of the objects was scrambled in the Follower's grid to avoid that the Director would use coordinates. **Figure 1 (right)**: Mean proportions of fixations on the grey cell in the seven trials of the Baseline condition and the seven trials of the Critical condition (by order of presentation from left to right) by those Followers who were suspicious of the Director's perspective (N=16). The number at the base of each bar indicates how many Followers did *not* fixate on the grey cell in that trial.

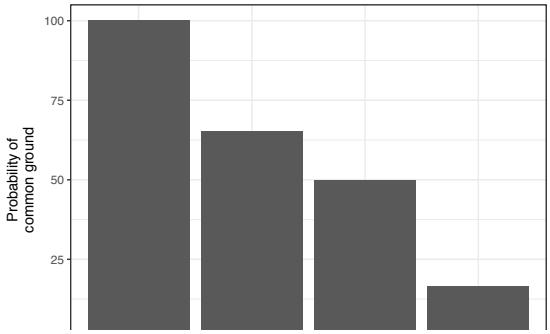
Following up on this study, we have tried to ***model common ground through referential specificity***. We presented MTurkers (N=40) with 4-figure displays, similar to those in the new DT. Participants had to estimate the likelihood that the speaker knew about the object in the grey cell given her instructions. To manipulate ***referential specificity*** we used **colour adjectives** ('the orange fish' vs. 'the fish'), **size adjectives** ('the small suitcase' vs. 'the suitcase') and **category level** ('the Porsche' vs 'the car'). To manipulate ***speaker's knowledge*** we used a **direct-reference condition** ('Select the blue fish'), a **hidden contrast condition** ('Select the orange fish', as in the new DT), a **hidden and shared contrasts condition** (adding a red fish on a white cell to the previous condition) and an **ambiguous condition** ('Select the fish' when there was a hidden blue fish).

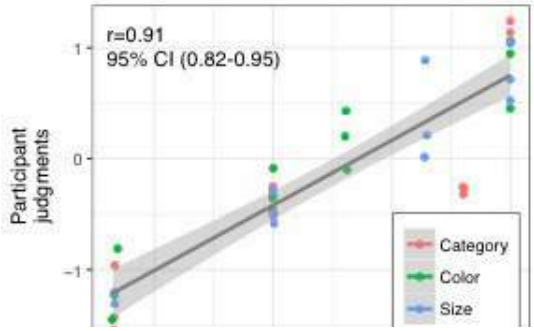
We modelled the task through a speaker model that produces utterances (*Utt*) given a referent (*Ref*) by doing Bayesian inference over the listener's probability of recovering the referent given the utterance:

$$p_s(Utt|Ref) \propto p_l(Ref|Utt)p(Utt)$$

Our listener model jointly infers the referent (*Ref*) and the speaker's knowledge (*Cg*) by reasoning about the probability that the speaker would produce the heard utterance given different hypotheses about the speaker's knowledge, and the referent:

$$p_l(Ref, Cg|Utt) \propto p_s(Utt|Ref, Cg)p(Ref)p(Cg)$$





The acceptability, processing and neural signature of nominal gradability

Galit W. Sassoon[a], Natalia Meir[a], Julie Fadlon[b], David Anaki[a] and Petra Schumacher[c]

Bar Ilan University[a], University of California San Diego[b], Cologne University[c]

Introduction: There is a dearth of studies investigating semantic Event-Related Potentials (ERP) effects in adjectives. To meet this gap, our study compares nouns to adjectives. It combines formal semantic theorizing with psycho- and neurolinguistic experimentation to test the hypothesis that acceptability and processing of nominal predicates in comparison constructions ("more P") is predicted by the ease of shift toward an interpretation based on **dimension counting**.

Categorization research reveals prevalence of dimension counting in classification under adjectives like *conservative/liberal* or *optimistic/pessimistic*, which can relate to politics, religion, sex, family, dress code, music, theoretical views, or some or all of those together (cf. Bartch 1986). According to Sassoon (2017), classification under such multidimensional adjectives is based on **a counting scale** reflecting the number of dimensions whose norms each entity exceeds. With a maximal/relative standard of membership, x *is optimistic* is true iff x is within the normal range of all/many optimism dimensions, and *x is more optimistic than {y, pessimistic}* is true iff x is *optimistic* in more respects than the number of respects in which {y is optimistic, x is pessimistic}.

As for nouns, dimension counting, although not as common as in multidimensional adjectives, is possible, especially in social nouns – labels of human traits or human-made objects (e.g., *linguist, scarf*). It is rare in natural-kind nouns (e.g., *duck, oak*; Hampton et al. 2009). Classification under natural-kind nouns is based on a causal model of the world (e.g., Gelman 2003). The absence of one dimension can cast doubt on the presence of the underlying cause and thus nullify all the other dimensional contributions. Thus, a donkey with some zebra property is classified neither as a zebra nor as a donkey. This can be modeled using multiplicative averaging, where x *is a donkey* is true iff x's averaged degree of similarity to the prototype in the dimensions of *donkey*, **the weighted product** of its dimensional degrees, $\mathbf{f_1(x)^{W1} \times \ldots \times f_n(x)^{Wn}}$, is above a norm. Any low degree (e.g., 0 or ½) reduces the product significantly (e.g., Medin & Schaffer 1978; Estes 2004).

By contrast, in social nouns the causal connections between dimensions (e.g., *intended* vs. *actual function*) are much looser. Each dimension has a constant additive effect on classification; e.g., typically, a linguist works in linguistics departments, investigates languages, and reads Chomsky's work, but a person violating some of these features may still count as a linguist. Thus, x *is a linguist* is typically modeled as true iff x's **weighted sum** of degrees in the dimensions, $\mathbf{w_1f_1(x) + \ldots + w_nf_n(x)}$, is above norm. One 0 degree hardly affects the sum (Rosch & Mervis 1975; Tversky 1977; Hampton et al. 2009). Significantly, upon a shift to equally important dimensions and binary dimensional scales consisting of 1 and 0, **additive (but not multiplicative) dimension-integration reduces to dimension-counting:** classification in social nouns may depend on whether entities were within the norm in *sufficiently many* dimensions (Wattenmaker 1995).

The dimension-counting hypothesis **contrasts** with the view that grammar and conceptual structure (dimension-integration) are dissociated, where mainly the mere lexical category or semantic-type of adjectives rather than nouns matters (e.g., **birder*; #*more bird than that one;* Kennedy 1999; Baker 2003; Neeleman et al. 2004), especially views that deny the role of dimension counting in gradability. Intuitively, rather than counting dimensions of multidimensional adjectives, speakers weigh the dimensions by importance and sum up the entities' weighted degrees in those dimensions thus creating one complex dimension (Bylinina 2013; Kennedy 2013; Lassiter 2015, a.o.), i.e. adjectives resemble nouns conceptually differing mainly in syntax/type.

Predictions: The dimension-counting hypothesis predicts that in comparatives (**A**) a higher acceptability of multidimensional adjectives stems from their readily available dimension-counting scales, (**B**) additive nouns whose interpretation may shift to dimension-counting would be judged more acceptable than multiplicative nouns, but the shift will exert processing time. Moreover, since conceptual violations typically engender an N400 ERP effect, while syntactic violations (among others) evoke a P600 effect (Bornkessel & Schlesewsky 2006; Friederici 2004), we expect the behavioral effect predicted in (A) to engender a P600 effect for comparatives with nouns vs. adjectives, and the effect in (B) to engender a more pronounced N400 effect for comparatives with multiplicative-natural-kind nouns relative to additive-social nouns.

Stimuli: Eight conditions (0-7) with 40 sentences each and 80 fillers were prepared. Subjects and predicates were balanced for length, frequency, morphological complexity and specificity.

		Comparison	**Baseline**
Adjectival	**Nat**	*(0) Leumat| ha-toref| ha-hu|, ha-toref| ha-ze| yoter amic*; Compared to that predator, this predator is more **brave**.	*(4) Ha-toref| ha-ze| amic*; 'This predator is **brave**'.
	Social	*(1) Leumat| ha-hoker| ha-hu|, ha-hoker| ha-ze| yoter| macliax;* 'Compared to that researcher, this researcher is more **successful'**.	(5) Ha-hoker| ha-ze| macliax; 'This researcher is **successful**'.
Nominal	**Nat**	*(2) Leumat| ha-toref| ha-hu|, ha-toref| ha-ze| hu| yoter arie*; 'Compared to that predator, this predator is more **a lion**'.	*(6) Ha-toref| ha-ze| hu| arie;* 'This predator is **a lion**'.
	Social	*(3) Leumat| ha-hoker| ha-hu|, ha-hoker| ha-ze| hu| yoter biolog';* Compared to that researcher, this researcher is more **a biologist**.	(7) Ha-hoker| ha-ze| hu| biolog; 'This researcher is **a biologist**'.

Participants were 29 healthy right-handed native Hebrew speakers (age range 18- 33).

Procedure: The stimuli were presented in a word by word reading paradigm followed by a naturalness judgment task (1-5-point scale). ERPs (mean amplitude and peak amplitude) were analyzed using 5-way ANOVA with Location (anterior/posterior), Laterality (left/central/right), Structure (baseline/comparison), Predicate (nominal/adjectival) and domain of subject (natural-multiplicative/social-additive) in time windows 325-500ms (N400) and 600-800ms (P600).

Results and discussion: Predictions A-B were born out by both the behavioral and ERP data. First, **naturalness ratings** for adjectives were high in both comparative and basic forms, while nouns were rated less acceptable in comparatives. In comparatives only, additive-social nouns manifested higher acceptability over multiplicative-natural-kind nouns. Second, **the RTs** indicated the predicted processing cost for social relative to natural nouns in comparatives, providing evidence for a conceptual dimension counting approach. Speakers easily tap the acceptability of multidimensional adjectives (0-1) and unacceptability of multiplicative natural nouns (2) in comparatives, but spend time on reanalysis to render additive social nouns (3) more acceptable, consistently with the predicted semantic shift to 'adjectival' dimension-counting scales. Third, **the ERP analysis** of the P600 window manifested a higher cost for nouns (2-3) than adjectives (0-1) in comparatives, but not in baseline, this is consistent with both a conceptual and syntactic/type view. The N400 window revealed a central posterior cost for processing of multiplicative-natural-kind nouns as opposed to additive-social ones in comparatives, but not in baseline. Such findings are consistent with a conceptual incongruency. Thus, our ERP data further help unravel the nature of violation in "x *is more a duck"*. The evidence is most consistent with a conceptual dimension-counting view (alternatives will be discussed in detail).

Selected references: **Baker, M.C.** (2003). *Lexical Categories*. CUP; **Bartsch, R.** (1986). Context-dependent interpretations. *GRASS* 7; **Bornkessel, I. et al.** (2006). *Psych Review 113*(4), 787–821; **Bylinina, L.** (2013). PhD diss.,

Utrecht; **Estes, Z.** (2004). *Psychonomic Bulletin & Review* 11, 1041–7; **Friederici, A. D.** (2004). *Current Neurology and Neuroscience Reports 4*(6), 466–70; **Gelman, S. A.** (2003). *The essential child.* OUP; **Hampton, A. J. et al.** (2009). *Memory & cognition 37*(8), 1150–63; **Kennedy, C.** (2013). *Inquiry 56*(2-3), 258–77; **Lassiter, D.** (2015). *Handbook of contemp. semantic theory*, 143–67; **Sassoon, W. G**. (2017). *Language* 93(1); **Wattenmaker, W. D.** (1995). *Cog Psy* 28.

What *and* means: a study on the intersective vs. non-intersective construal of VP-*and*

Viola Schmitt, University of Vienna & Daniele Panizza, University of Goettingen

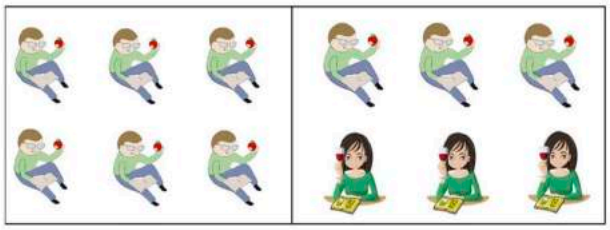
English sentences with VP-conjunctions such as "the six people are eating$_P$ *and* reading$_Q$" have two construals: an *intersective* construal (IC), where it is true iff each of the six boys is both eating and reading and a *non-intersective* construal (NIC), where it is true iff some of the people are eating, some are reading and each of them is either reading or eating. The linguistic literature disagrees w.r.t. what the core semantics of *and* is: some accounts claim that it is IC ([1], [7]) whereas other accounts claim it is NIC ([3],[2]). Furthermore, it has been claimed that the IC is the basic interpretation of VP-conjunction [1], that NIC are only found if the predicates are disjoint [8] (as in "the boys are sitting$_P$ and standing$_Q$") or more frequently interpreted as disjoint [5] and that NIC are more easily accessible with in contexts where "P and Q" is followed by "but not R" [6]. If so, this represents a problem for NIC analyses. Our experiment was designed to investigate the availability of IC and NIC in adults and 6- to 10-year-old children in scenarios where P and Q are disjoint or conjoint, whether there is a preference for IC or NIC scenarios and whether a continuation "and not R" affects the interpretation. We employed the Semantic Choice Task [4], where two scenarios including six characters performing an action are presented simultaneously on the screen and participants must choose one or reject both, while listening to a sentence. (1) exemplifies an item with non-disjoint predicates, (2) one with disjoint predicates. The material in brackets was included in half of the conditions.

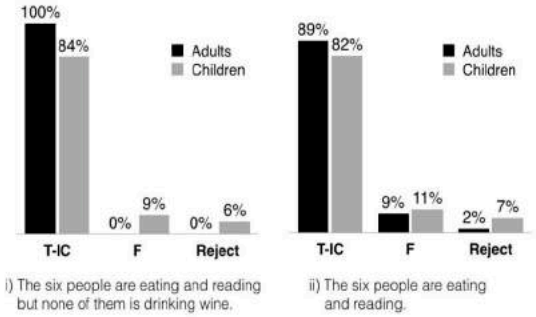
(1) The six people are eating$_P$ and reading$_Q$ (but none of them is [drinking wine]$_R$)

(2) The six children are sitting$_P$ and standing$_Q$ (but none of them is [lying down]$_R$)

Conditions involving a true scenario (T-NIC/T-IC) vs. a false scenario (F) allow us to test for the access to one construal; Conditions involving two true scenarios allow us to test for preferred construals. **Condition 1** tests the availability of IC with non-disjoint predicates (i.e. (1)): It contrasts a T-IC-scenario with an F-scenario (fig.1). Participants consistently selected the T-IC-scenario over the F-scenario and the Rejection option (R) in the items including the continuation and those excluding it (fig.1). **Condition 2** tests the availability of NIC with disjoint predicates (i.e.(2)) by contrasting a T-NIC-scenario with an F-scenario (fig.2). Both groups consistently chose the T-NIC-scenario in the items with and without the continuation (fig.2). **Condition 3** tests the accessibility of NIC with non-disjoint predicates (i.e. (1)) and the preference for either scenarios where P and Q overlap in some individuals or scenarios where they don't (fig.3). For items including the continuation, the non-overlapping scenarios are T-NIC-scenarios, the overlapping ones F-scenarios (continuation: none of them is R). Here, both children and adults selected the T-NIC-scenario (fig.3), the difference between the two groups was not significant ($p=.41$). Both groups selected the rejection option more often than in Conditions 1 ($p=.02$) and 2 ($p=.026$), but the acceptance rate was much higher than what is reported by [5] for analogous cases. For items without the continuation, both scenarios are T-NIC-scenarios. Here both adults and children preferred the non-overlapping scenario (fig.3); the difference between the groups was again not significant ($p=.17$). **Condition 4** tests the preference between the T-IC-scenario from Condition 1 with the non-overlapping T-NIC-scenario from Condition 3 (fig.4). In the items with the continuation adults displayed a strong preference for the T-IC scenario, but children only showed a mild preference for it (fig.4). In the items without the continuation, the preference for the TIC was more attenuated in both groups (fig.4). The difference between the groups preference was significant ($p<.01$) but the presence of the continuation did not have any effect ($p=.35$). **Conclusion**: The experiment shows that NIC of VP-conjunction are not exceptional/tied to particular semantic configurations: children and adults generally access NIC of VP-conjunctions in configurations where P, Q are disjoint and where they not disjoint. The semantic configuration plays only a marginal role: with non-disjoint predicates both adults and children have more rejections (27% and 18%) but still accept the NIC in the great majority of cases. The continuation plays no significant role, either. Yet the experiment also reveals two interesting facts about preference: adults strongly prefer T-IC- over T-NIC-scenarios, whereas children have a much smaller preference, suggesting again that the NIC is clearly available for children.

Figure 1 - *Condition 1*

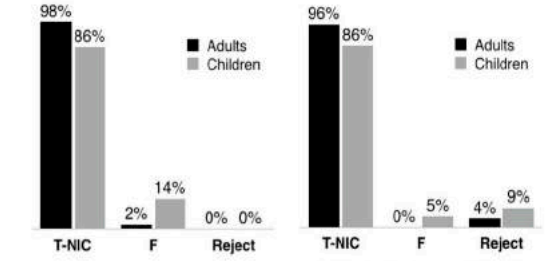




i) The six people are eating and reading but none of them is drinking wine.

ii) The six people are eating and reading.

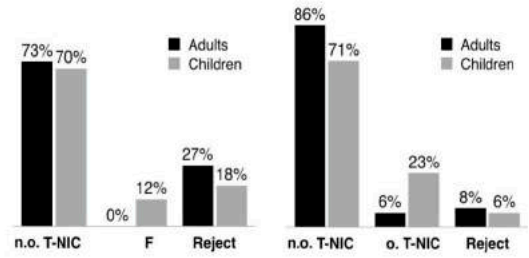
Figure 2 - *Condition 2*



i) The six children are sitting and standing but none of them is lying down

ii) The six children are sitting and standing.

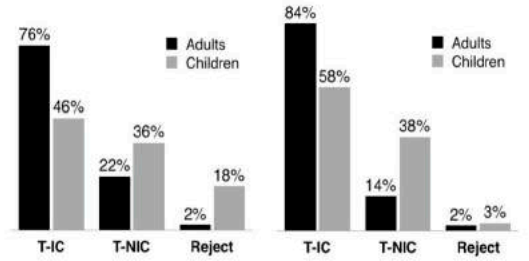
Figure 3 - *Condition 3*



i) The six people are eating and reading but none of them is drinking wine.

ii) The six people are eating and reading.

Figure 4 - *Condition 4*



i) The six people are eating and reading but none of them is drinking wine.

ii) The six people are eating and reading.

[1] Lucas Champollion. Ten Men and Women Got Married Today: Noun Coordination and the Intersective Theory of Conjunction. *Journal of Semantics*, 10.1093/jos/ffv008(0):1–62, 2015.

[2] Caroline Heycock and Roberto Zamparelli. Friends and colleagues: Plurality, coordination, and the structure of DP. *Natura Language Semantics*, 13(3):201–270, 2005.

[3] Manfred Krifka. Boolean and non-boolean 'and'. In László Kálmán and László Pólos, editors, *Papers from the Second Symposium on Logic and Language*, pages 161–188, Budapest, 1990. Akadémiai Kiadó.

[4] Karoliina Lohiniva and Daniele Panizza. When Pragmatics helps Syntax: An Eye Tracking Study on Scope Ambiguity Resolution in 4- to 5-Year Old Children. In *Proceedings of 40th annual Boston University Conference on Language Development*, pages 216–228, 2016.

[5] Eva B. Poortman. Between Intersective and 'Split Interpretations of Predicte Conjunction. In *Proceedings of the Forma and Experimental Pragmatics Workshop*, pages 36–42, Tübingen, 2014.

[6] Viola Schmitt. *More pluralities*. PhD thesis, University of Vienna, 2013.

[7] Yoad Winter. *Flexibility Principles in Boolean Semantics*. MIT Press, Cambridge, Massachusetts, 2001.

[8] Yoad Winter. Plural Predication and the Strongest Meaning Hypothesis. *Journal of Semantics*, (18):333–365, 2001.

Some approximations: an experimental investigation

Stephanie Solt (ZAS), Jon Stevens (OSU) and Brandon Waldon (ZAS)

The problem. The sentence in (1) illustrates a curious use of *some* in which it combines with a numerical expression (the '*some* + numeral' construction, or SN for short):

(1) Some 20 cars were involved in the accident.

Authors including Sauerland & Stateva (2007) and Anderson (2014) align *some* on this use to approximators such as *about* and *approximately* (i.e. *some 20* ≈ *about 20*), an idea that may be formalized via the mechanism of scale granularity (per (2)) or pragmatic halos (per (3)):

(2) $[\![\textit{some twenty}]\!]^{gran} = \text{coursest(gran)}[\![\textit{twenty}]\!]$ (Sauerland & Stateva 2007)

(3) $[\![\textit{some twenty}]\!]^{C} = f([\![\textit{twenty}]\!] \cup \textit{halo}_C([\![\textit{twenty}]\!]))$ (Anderson 2014)

However, such analyses fail to capture distributional restrictions that distinguish *some* from approximators such as *about*.

i) Sum-based interpretation: In contrast to true approximators, SNs are restricted to occurring with numerical expressions that can be interpreted as the sum or aggregation of some type of entity or unit of measure (e.g. pluralities, temporal/spatial extents, but not clock times):

(4) a. The meeting lasted some 3 hours. / We drove some 30 miles. / *It's some 3 o'clock.
b. It's about/roughly/approximately 3 o'clock.

In support of this, a corpus analysis (COCA; Davies 2008-) found no examples of non-sum-based measures in SNs; for approximators these accounted for 3-5% of non-cardinality tokens.

ii) Lack of degree interpretation: SNs differ further from approximators in being infelicitous in mathematical statements and as answers to *how many?* questions (with no overt pronoun).

(5) Seven times fourteen is about 100 / approximately 100 / roughly 100 / ??some 100.

(6) Q: How many students passed the test? A: 50 / about 50 / ?some 50 / some 50 of them.

From this we conclude that SNs do not have denotations in the degree domain (contra (2),(3)), but rather interpretations that are more closely integrated with that of the following NP.

An even more central issue is that while some speakers attribute an approximative interpretation to SNs, others do not. This is reinforced by corpus examples in which SNs are used to convey a precise known value; (7b), for example, could not be paraphrased as 'she bore him about 6 children'. This is unexpected on an analysis that treats *some* as an approximator.

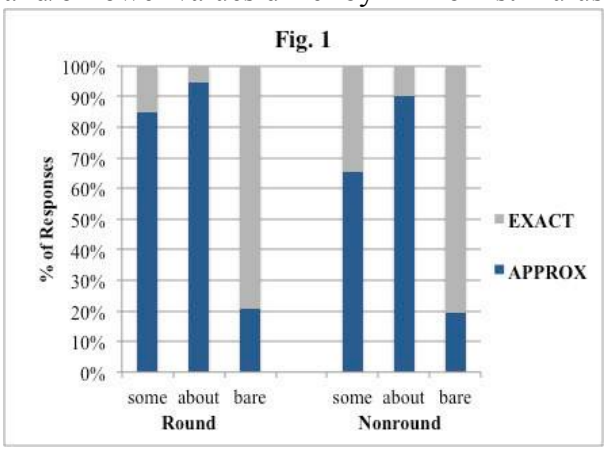
(7) a. Of some 206 students who responded to the survey, 52% were female.
b. She bore him some 6 children, 3 of them boys.

Objectives. In order to develop a revised analysis of SNs that captures the patterns in (i) and (ii), it is necessary to understand the source of the variation in the presence of the approximating effect. We consider two hypotheses: H1: Speaker Variation: some speakers interpret *some* as an approximator, while others assign it a different meaning. H2: Derived Approximation: The approximative effect in examples such as (1) is not contributed by *some* itself, but rather derives from a possibility already available (if latent) for bare numerals. Krifka (2007) observes that round numbers allow approximate interpretations, while non-round numerals are interpreted precisely. If this is the source of the approximate interpretation of SNs, we predict *some*+round number to be interpreted approximately, but *some*+non-round to be interpreted precisely.

Experiment. To test the above hypotheses, native English speakers (n=75) were recruited via Amazon MTurk for an online interpretation study, in which they saw sentences such as (8) and were asked to indicate their interpretation of the numerical expression by providing a range:

(8) The company added {about/some/∅} {50/47} new jobs in the first half of the year.
How many jobs did the company add in the first half of the year? Between ___ and ____.

A 3x2 design was employed, including 3 modifier conditions (*some*, *about*, bare) in 2 numerical conditions (round, non-round). Items were distributed across 5 lists: each respondent saw 1 item with a round number in all 3 modifier conditions, and a second item with a non-round number in all 3 modifier conditions. There were an additional 20 fillers, for a total of 26 items/participant. Presentation order was randomized for each participant. Data from 3 participants were excluded due to incomplete/inaccurate responses on filler items. The remaining responses were coded as EXACT (upper and lower values differ by at most 1 from stimulus value) or APPROXIMATE (upper and/or lower values differ by >1 from stimulus value).



As seen in Fig. 1, bare numerals elicited mostly EXACT responses and *about*+*n* primarily APPROXIMATE responses. *Some* constructions patterned distinctly from both, in the round condition eliciting mainly APPROXIMATE responses, but in the non-round condition exhibiting mixed behavior. A generalized logistic regression model found a significant difference between *some* and both bare ($z=7.8$, $p<0.001$) and *about* ($z=-4.2$, $p<0.001$). At the respondent level, the most common pattern (22/72) was to give APPROXIMATE responses to 'about' and 'some' trials in both round and non-round conditions, and EXACT responses to 'bare' trials in both conditions. But the second most common (16/72) was to treat *about* and *some*+round as APPROXIMATE but bare and *some*+non-round as EXACT. In total, roughly a third of participants distinguished *some* from *about* by giving *some* an EXACT interpretation in at least one condition where *about* received an APPROXIMATE interpretation.

Conclusions and preliminary analysis. The experimental results provide support for both of the above hypotheses. Some respondents do not differentiate *some* from the true approximator *about*. But others do (Speaker Variation), giving responses consistent with an analysis on which the approximative effect derives from a possibility inherent in the interpretation of bare round numbers themselves (Derived Approximation). We propose a two-part analysis: i) for some speakers, SNs are truth-conditionally equivalent to bare numerals, differing only in that they introduce a plural discourse referent formed via aggregation of atomic entities; in combination with round numbers they inherit the potentially approximate interpretation of the latter, an effect facilitated by the pragmatic principle of associating marked forms with marked meanings (Horn 1984); ii) for other speakers, this approximative meaning has been conventionalized into the semantics of *some* itself. In support of this, diachronic corpus data show that in earlier stages of English, SNs frequently (half of tokens) occurred with other markers of approximation (e.g. *some 15 or 20 men*), which we hypothesize to have encouraged semantic reanalysis.

REFERENCES. Anderson, C. 2014. Approximation of complex cardinals using *some*. *WECOL 2013*. **Davies, M.** 2008-. *The Corpus of Contemporary American English: 520 million words, 1990-present.* **Horn, L.** 1984. Towards a new taxonomy of pragmatic inference. In D. Schiffrin (ed.), *Meaning, Form, and Use in Context.* **Krifka, M.** 2007. Approximate interpretations of number words. In G. Bouma et al. (eds) Cognitive foundations of interpretation. **Sauerland, U. & P. Stateva.** 2007. Scalar vs. epistemic vagueness. *SALT 17*.

Scalar implicatures in the context of full and partial information. Evidence from ERPs.

Maria Spychalska*, Ludmila Reimer, Petra schumacher*, Markus Werning****
(***University of Cologne; ** Ruhr-University of Bochum**; m.spychalska@gmail.com)

It is considered underinformative to say *Some cards contain cats* if all cards contain cats, even though semantically it is true. This phenomenon is described in Gricean pragmatics [4, 5] in terms of scalar implicature: if the speaker uses a semantically weak quantifier *some*, the listener may infer that the she is not in a position to use the stronger alternative *all*. Assuming that the speaker is informed (*competence assumption*), the listener may infer that the stronger alternative is believed by the speaker to be false. From the psycholinguistic perspective the main question has been whether this implicature is processed incrementally – as a fast, automatic inference upon encountering the quantifier *some*, or whether it is only inferred at the later stage during sentence processing [1, 7, 6, 8]. Most experiments investigating this issue have involved paradigms where full information relevant for the sentence evaluation is available to all parties involved. In such contexts, underinformative sentences tend to trigger divergent truth-value judgments. Using ERPs, [9] showed that this intuitive truth-value evaluation determines the way the implicature is processed: underinformative sentences were associated with larger N400 ERPs relative to informative sentences only for subjects who evaluated them as false (*pragmatic responders*), whereas no such effect was observed for those participants who evaluated underinformative sentences as true.

Up to date, there is relatively little evidence regarding the role of the speaker's competence assumption for the implicature processing (related work [2, 3]). In our ERP experiment we investigated the processing of the scalar implicature in the context of partial information, i.e. when the assumption of the speaker's competence is violated. The experiment uses a paradigm where participants evaluate appropriateness of the speaker's utterances about a card game situation. The target scenarios consist of (i) the speaker's avatar; (ii) four open cards placed on the table; and (iii) two cards face down (whose content cannot be seen) placed on the side of the speaker (Tab 1, Fig 2). The subject is informed that the speaker doesn't know what is on the face-down cards. The speaker's utterances are presented auditorily and either refer all cards including the face-down cards (*Some cards **in the game** contain As*), or to the cards on the table only (*Some cards **on the table** contain As*). By manipulating whether the critical noun A refers to (i) the object category contained by every visible card; (ii) the object category contained by a subset of visible cards; (iii) another object category not presented at the screen, we compare cases where the sentence's truth-value and pragmatic felicity is either known or unknown to the speaker.

The results indicate an N400 effect for false relative to informative and underinformative sentences, both for the **table** and **game** context. Unlike in [9], in the context of full information, no effect is found for the implicature violation (**Table-Underinformative** vs. **-True**) for pragmatic responders, i.e. those who reject **Table-Underinformative** sentences as not appropriate utterances. We argue that the context of our experiment does not endorse the scalar implicature due to the presence of additional partial information scenarios. Among the available alternatives, *some* can be considered the most optimal quantifier to express uncertainty, which endorses its logical (*some and possibly all*) interpretation. Consequently, even for the pragmatic responders (31%) the implicature is not incrementally processed. For the partial information context, we observe that sentences that are *known* to be informative (**Game-True**) form a significant negativity relative to *potentially* underinformative sentences (**Game-Underinformative**) ($p < .014$), as well as relative to **Table-Underinformative** sentences ($p < .004$). This effect supports the hypothesis that *some* is interpreted as means of expressing uncertainty: it indicates that *some cards in the game* is more optimally used to describe the quantity of those objects that occur in all visible cards (and thus may also be present on the face-down cards), whereas for objects that occur only in a subset of visible cards, more appropriate quantifying expressions are available (e.g. *some cards on the table*).

	Some cards (in the game/on the table) contain...		
place	Noun		
	cats	*balls*	*dogs*
in the game	***Game-Underinformative*** Unknown infelicitous Yes/No	***Game-Informative*** Known felicitous Yes	***Game-False*** Unknown false No
on the table	***Table-Underinformative*** Known infelicitous Yes/No	***Table-informative*** Known felicitous Yes	***Table-False*** Known false No

Table 1: For each critical word the table provides: the condition's label (first line), semantic/pragmatic value of the sentence in that condition (second line), expected resp. possible response (third line). **Note:** A "no" response in **Table-Underinformative** condition indicates a pragmatic interpretation. Participants were highly **consistent** in their choice of the logical (ca. 70%) or pragmatic (ca. 30%) interpretation. A "no" response in **Game-Underinformative** condition is considered a *strong pragmatic interpretation* and was chosen by only one participant in the whole tested group.

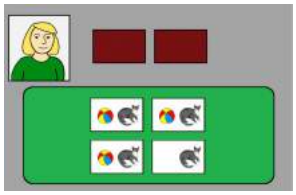


Figure 1: Schematic illustration of a target visual scenario: In the experiment sentences are presented as auditory stimuli during the presentation of the visual scenario. **Note:** Filler trials are used to balance the materials (i) with a different number of object categories presented, (ii) with other quantifiers (*all*, *no*, *more than three/two*, *fewer than four/three*, *three/four*), (iii) with additional cards outside the table being face-up (both the speaker and the listener can see what these cards present), or with no additional cards outside the table dealt (in this way we highlight the relevance of the face-down cards in the target trials). Filler trails allowed also to control that the subject understood the task and was able to make a distinction between **table** and **game** sentences.

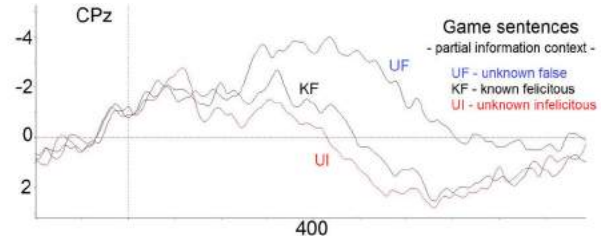


Figure 2: Grand averages (N=23) for the **Game**-sentences (partial information). Cluster-based permutation statistics: Significant negativity for *unknown false* (**Game-False**) relative to *known felicitous* (**Game-Informative**) as well as *unknown infelicitous* (**Game-Underinformative**) conditions (effects with $p < .0001$). Significant negativity for the *known felicitous* relative to the *unknown felicitous* condition ($p < 0.014$).

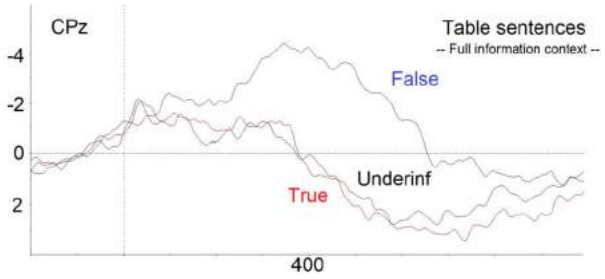


Figure 3: Grand averages (N=23) for the **Table**-sentences (full information context) at the critical sentence-final noun. Cluster-based permutation statistics: Significant negativity for the *false* relative to *true* ($p < .0001$) and *underinformative* ($p < .0001$) conditions. No significant differences between *true* and *underinformative* conditions; no effect due to divergent evaluation of *underinformative* sentences.

[1] Lewis Bott and Ira A. Noveck. Some utterances are underinformative: The onset and time course of scalar implicatures. *Journal of Memory and Language*, 51(3):437–457, 2004. Doi:10.1016/j.jml.2004.05.006.

[2] Richard Breheny, Heather J. Ferguson, and Napoleon Katsos. Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition*, 126(3):423–40, 2013. Doi:10.1016/j.cognition.2012.11.012.

[3] Noah D. Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5:173–184, 2013. Doi:10.1111/tops.12007.

[4] Herbert Paul Grice. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics*, volume 3 of *Speech Acts*, pages 41–58. Academic Press, New York, 1975. Reprinted in *Studies in the Way of Words*.

[5] Laurence R. Horn. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin, editor, *Meaning, Form, and Use in Context: Linguistic Applications*, pages 11–42. Georgetown University Press, 1984.

[6] Mante S. Nieuwland, Tali Ditman, and Gina R. Kuperberg. On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, 63(3):324–346, 2010. Doi:10.1016/j.jml.2010.06.005.

[7] Ira A. Noveck and Andres Posada. Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85(2):203–210, 2003. Doi:10.1016/S0093-934X(03)00053-1.

[8] Stephen Politzer-Ahles and Robert Fiorentino. The realization of scalar inferences: Context sensitivity without processing cost. *PloS one*, 8(5), 2013. Doi:10.1371/journal.pone.0063943.

[9] Maria Spychalska, Jarmo Kontinen, and Markus Werning. Investigating scalar implicatures in a truth-value judgment task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, 31(6):817–840, 2016. Doi:10.1080/23273798.2016.1161806.

Chao Sun and Richard Breheny (University College London)

On the compositional interpretation of scalar quantifiers: The role of the residue set

Recent visual world studies examined whether comprehenders interpret *some* as *some and not all* in the same timecourse as they compute the semantic interpretation of *all*. [1] reported that referential disambiguation based on pragmatic *some* was delayed compared to *all*, whereas [2-3] found no evidence for such delay. [4] manipulated target set sizes and found stronger target bias after hearing *all* than *some* only when the set size was big. So far, the timecourse question remains unsolved and how set size interact with scalar processing is unclear. Here we demonstrate that people have prior expectations about the target set size in a display given the quantifier use and these expectations influence target bias formation. Unlike previous studies, we also examine the time course question by comparing looks to the 'residue set' after hearing quantifiers and numerical determiners. Looks to the residue set reflect incremental integration of compositional interpretation of quantifiers and are not affected by other expectations. We find the timecourse of gaze bias based on pragmatic *some* is not different to that for *all*.

Exp.1 Given [4], our hypothesis is that people have prior expectations that an agent with a total set of something will possess a relatively large set of objects. We asked participants to indicate on a slider scale which image fits better with a statement containing a quantifier, (fig.1). The statement could equally be true of both. Participants (N=39) judged 2 critical items and 4 fillers, of which two were clearly unambiguous and two ambiguous (e.g. the girl has red and green apples). **Results:** for both quantifiers, participants prefer the agent with the larger set as the referent (both ps<.001). And the preference for *all* to be used with a big set was stronger than that for *some* (p =.024). The big set preference in *all* is consistent with our hypothesis. The result for *some* could reflect a preference for using *some* with a set containing three over two. Note that in [1] *all* targets have three items, *some* targets have two.

Exp.2 We re-examine the timecourse question and test how prior expectations influence scalar processing. Participants viewed a visual display (fig.2) while listening to an instruction of the form "Click on the [girl/boy] that has [Det] of [name's] [noun]", [Det] is one of *some*, *all*, *two*, *three* (fully counterbalanced). The residues of partitioned sets were in the centre. **Prediction 1.** Less looks to the residue set for *numbers* than *all* because anticipating a referent in *number* conditions does not require checking the residue set. **Prediction 2.** Only if *some* is interpreted as 'some and not all' should there be also less looks to the residue set in *numbers* than *some*. **Prediction 3.** Given exp.1's results, in *all* and *some* condition, looks to the target should increase faster when the target set size was big compared to when it is small. **Prediction 4.** Anticipatory looks to the target should increase faster in *all* than in *some* when the target set size was big. **Results**. Residue set results show rapid integration of pragmatic *some*. As shown in fig.3, during disambiguation regions, looks to the residue set decrease faster in numbers than both *all* and *some* ([Det]: *p*s<.001; [name's]: *p*s<.001). Critically, growth curve analysis reveals that during [Det] window for *all* and *some*, there is a quadratic increase in looks to the residue set (rise/fall), but such trend is not found in numbers. With regard to prediction 3 and 4, shown in fig.4, the target bias in big *all* is stronger than in small *all* (both windows, *p*s<.001) and a marginal diff. between big and small *some* in ([Det]: p=0.09). We also find looks in big *all* condition increase more rapidly than looks in *big* some ([Det]: p=.02). These results show that prior expectation has bigger influence on how target bias develop over time in *all*. Independent of size, we find target bias emerged earlier and stronger in *numbers* than in *all* and *some* (for both windows, *p*s <.001), whereas *all* and *some*, did not differ. **Conclusion** prior expectations affect target identification when set size is not

controlled. Our results render the interpretation of previous visual world data, incl. in [1], problematic. When prior expectations are controlled, overall target results and residue set results indicate that enriched *some* was as fast as *all*.

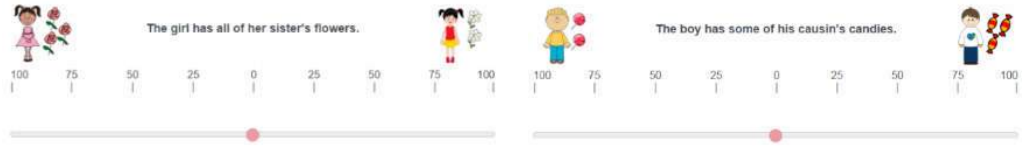


Figure 1 experimental items used in exp.1

(a) (b)

Figure 2: 2(a) can be paired with instructions 'Click on the boy that has *all/three* of Susan's apples' or ''Click on the girl that has *some/two* of Susan's pears'. 2(b) can be paired with instructions ''Click on the girl that has *all/two* of Susan's pears' or 'Click on the boy that has *some/three* of Susan's apples'.

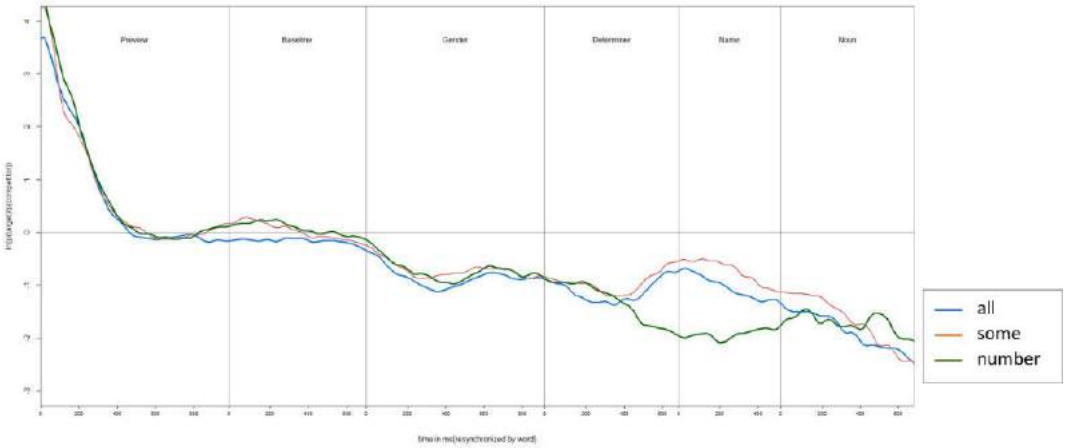


Figure3 Time course plot of looks to the residue set (empirical logit) by *Determiner* from the instruction onset to the instruction offset

7

6.8

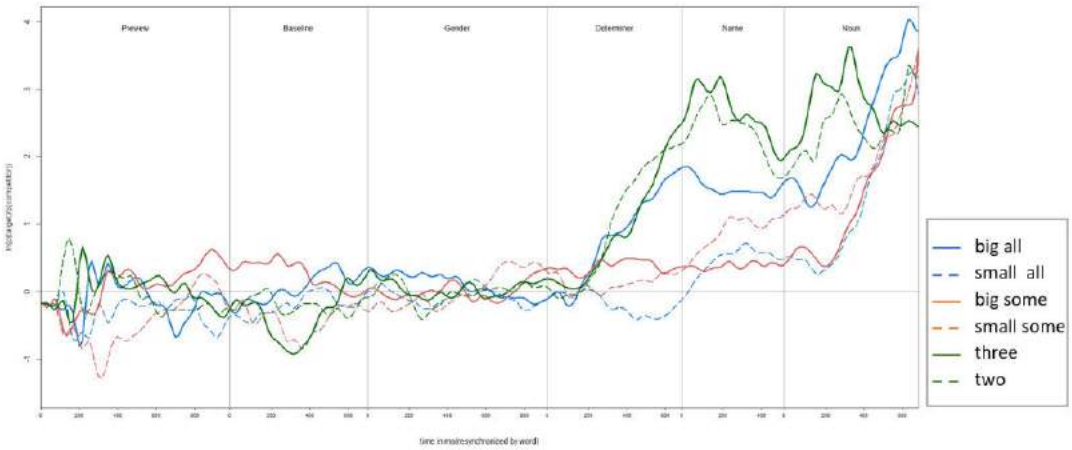


Figure 4. Log ratio of looks to target over competitor by Determiner and Target size from the display onset to the instruction offset

References: [1] Huang, Y. T., & Snedeker, J. (2009). *Cognitive Psychology*, 58(3), 376–415. [2] Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). *Cognition*, 116(1), 42–55. [3] Breheny, R., Ferguson, H. J., & Katsos, N. (2012). *Language and Cognitive Processes*, 28(4), 443–467. [4] Degen, J. & Tanenhaus, M. K. (2016). *Cognitive Science* 40 (1):172-201.

Rates of scalar inferences beyond 'some' – A corpus study

Chao Sun (University College London), Ye Tian (Universite Paris Diderot), Richard Breheny(University College London)
chao.sun.13@ucl.ac.uk

In a large-scale corpus-based web study, Degen (2015) extracted 1363 utterances containing *some*-NPs from the Switchboard corpus. For each utterance, they measured the rate of the scalar inference (SI) from *some* to *some but not all* using a paraphrase task. Their findings showed that around half the time *some* is used, an SI reading is not judged to be available. Little is known about how frequently scalar expressions of different lexical categories give rise to SIs in *real use*. In an inference task, van Tiel et al. (2016) showed that different scalar terms give rise to SIs at different rates. Here, we adopt Degen's paraphrase task using a Twitter corpus we constructed to investigate whether the rates of SI derivation vary to the extent found in van Tiel et al.'s inference task. We do find variability in the rates of SIs across different scalar expressions, but not the same degree of variability found when items are presented out of context in an inference task. A modest amount of this variance could be explained by factors which van Tiel et al. found contributed to the variances in the experimental setting. Our study yields several interesting results, mentioned below.

Collecting a Twitter corpus: We selected 28 out of 43 scalar expressions found in van Tiel et al. (2016). There were 2 quantifiers (e.g. <some, all>), 1 adverb (<sometimes, always>) and 25 adjectives (e.g. <intelligent, brilliant>). For each scale, we extracted tweets containing the weak scalar term - with a minimal length of 30 characters. Then we conducted part-of-speech (POS) tagging on each tweet and used regular expressions to filter out tweets where scalar expressions appear in environments which the inferences are unavailable or less likely to arise (see Table 1).

environment	example
in the scope of negation	I'm not really hungry.
in the scope of conditional antecedents	If the weather was warm, we would have some people over for a small party in our backyard.
in the scope of wh-questions or polar questions	Do you get adequate vitamin D?

Table 1: Environments prohibit the scalar inference

To perform the final exclusion, we conducted a word sense disambiguation task on Amazon Mechanical Turk to obtain human annotation on tweets containing polysemous scalar expressions. Considering <old, ancient> for example, in (1a) the sense of *old* meaning "existing a long time" is on the same scale as the core meaning of *ancient*. However, in (1b) the sense of *old* meaning "previous" was not on the same scale as the strong term. Cases like (b) need to be excluded because in these cases the strong term is not contextually available which make it infelicitous to investigate the rate of SIs. We consulted the Merriam-Webster dictionary and found 20 out of 28 our scalar expressions have at least two meanings.

(1a) I'm in an ***old*** abandoned train station w/ a translator working on the script.

(1b) That means my ***old*** boss has been approaching a breakdown for the last 2 years.

80 M-Turk workers were recruited and each annotated 50 tweets of a particular scalar expression. In total, 4000 tweets were annotated, 200 tweets per scale. We presented workers with a tweet containing the scalar expression, e.g. *warm* (I guess he wants his home to feel ***warm*** and inviting.) and ask them to choose the meaning of *warm* from the following three sense labels: (if none are appropriate, workers can click 'none of the above' option) (a)

having a fairly high temperature; (b) friendly and affectionate; (c) light and bright colors. (a) is the sense that could be understood on the same dimension as the strong term, whereas (b-c) are the relatively common senses listed in the dictionary. Based on our results, we excluded tweets in which weak terms evoke senses that are not on the same scale as strong terms.

Corpus-based paraphrase task: We ran a paraphrase task based on Degen (2015) to measure the frequencies of SIs triggered by the 28 scalar expressions. After the final exclusion, we ended up with 3075 tweets in total. We randomly selected 50 tweets for each scale as the target sentences. On each trial, participants read an utterance containing a scalar expression *X* (the weak term, in red) and a nearly identical utterance, expect that the negation of the stronger term *not Y* (in green) was inserted (Figure 1). Participants were asked to rate on a seven point scale to indicate how similar is the statement with *X but not Y* to the statement with *X*. 550 participants each judged 28 items – one item per scale.

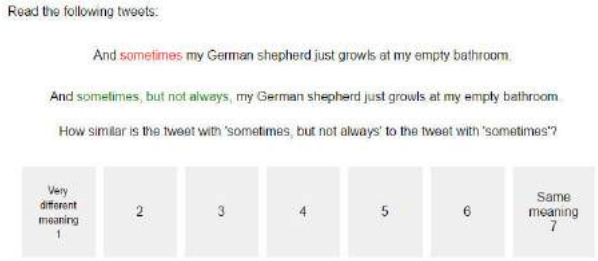


Figure 1 paraphrase task example item

Results: The responses were coded into three categories: low (ratings were 3 or lower), median (ratings were 4), and high (ratings were 5 or above). We considered high ratings as an indicator of SIs being drawn. Inspecting Figure 2, the frequency of SI varies across scalar expressions, from 27% for <adequate, good> to 86% for <sometimes, always>. These results correlated with the results of van Tiel et al. (2016) ($r=0.81$, $p<.001$), suggesting that, to some extent, the results yield from the inference task based on artificial examples could reflect frequencies of SIs triggered in *real use*. However, Levene's test for equality of variances showed that variances of two studies are not equal ($F(1,54)=14.69$, $p<.001$). Visual inspection of Figure 2 suggests that there is less variation on the paraphrase task. In particular, adjective scalar expressions give rise to SIs more frequently in real use. We replicate the result in Degen (2015) for 'some' and note that actual rates of SIs for this item and other terms like, 'possible' and 'allowed' are far lower than rates found on the inference task.

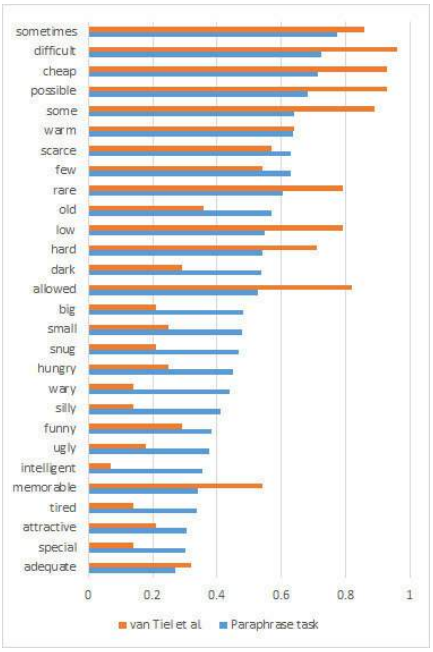


Figure 2 shows the percentage of 'High' ratings for 28 scalar expressions. Percentage of SI responses from van Tiel et al. (2016, Experiment 2) are shown in orange.

The variability displayed in the frequencies of SIs call for an explanation. Multiple linear regression analyses were conducted to predict the frequencies of SIs from possible factors explored in van Tiel et al. (2016), including association strength, grammatical class, word frequencies, semantic relatedness, semantic distance, and boundedness. As van Tiel et al., found with their inference task results, only semantic distance and boundedness are substantial factors. In this case, these factors together accounted for 43% of the variance. Future studies need to explain where the remaining variance comes from.

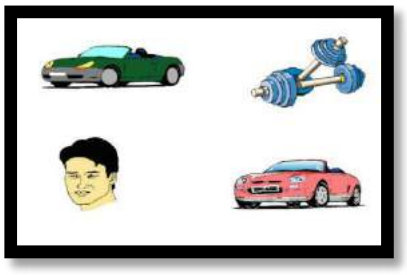
Reference: [1] Degen, Judith. 2015. *Semantics and Pragmatics* 8(11). 1–55. [2] van Tiel, B. van Miltenburg, E. Zevakhina N. & Geurts, B. (2016), *Journal of Semantics*, 33: 137-175.

Through the eyes of a teenager: complexity of real-time Theory of Mind inferences in language comprehension.

Irene Symeonidou[1], Wing Yee Chow[1], Heather Ferguson[2], Richard Breheny[1]

[1] Division of Psychology and Language Sciences, UCL; [2] School of Psychology, University of Kent

A number of studies have provided evidence for structural and functional changes in the brain areas involved in theory of mind (ToM) (the "social brain") not only during childhood, but also during adolescence, [1]. Recent findings suggest that the online use of ToM shows a prolonged development through late childhood and adolescence, [2]. In order to investigate the role of Theory of Mind (ToM) in language comprehension and its developmental trajectory, we adopted a visual world paradigm from [3] to examine how quickly younger (9-13) and older adolescents (14-18.5) can use knowledge about a character's preferences and desires to make complex ToM inferences and predict that character's subsequent behaviour during discourse, in comparison to adults (25-36). Participants were presented with two sentences in each trial. Sentence (i) introduced a property of a character and (ii) described the character performing an action that is consistent with context given in sentence (i) (**Figure 1**).



Open Context (basic preference and intentions are consistent)
i) *Tom is always telling people that his favourite colour is pink.*
ii) *Last week Tom bought a new car and he deliberately chose a pink car.*

Secret Context (basic preference and intentions are in conflict)
i) *Tom does not want anyone to know that his favourite colour is pink.*
ii) *Last week Tom bought a new car and he deliberately chose a green car.*

Figure 1. Example of visual stimulus that participants viewed while the heard the target sentence (sentence ii).

In the open condition, the character's basic preferences and high-order desires match, whereas in the secret condition the character's basic preferences and high-order intentions are in conflict. The visual display (see **Figure 1**) was presented during the second sentence and was the same for both conditions. Participants also under took standard tests of inhibitory control (IC), working memory (WM) and an Empathy Quotient test. In addition, adolescents and adults also undertook an on-line false belief task, adopted from [4].

Results. Fifty-two participants took part in this study (Adults n=17, M = 27.32 years, SD = 3.57; Adolescents I n=18, M = 16.70, SD = 1.39; Adolescents II, n=17 M = 11.81, SD = 1.43). Statistical analyses were carried out with mixed effect regression models for each time window separately: i) ambiguous noun (e.g. "car") ii) post-ambiguous noun, iii) adverb (e.g. "deliberately"), iv) transitive verb, and v) disambiguating noun. For each time-window a referent preference score was calculated as in [3]: log(Open/Secret) = *ln* ($P_{(Open)}$ / $P_{(Secret)}$). $P_{(Open)}$ refers to the sum of looks to the open referent divided by the total number looks to all ROI's within that trial and $P_{(Secret)}$ is the sum of looks to the secret referent divided by the total number of looks to all ROI's within that trial. Results showed that adults start making anticipatory eye movements towards the target in the *open condition* (pink car) and the *secret condition* (green car) early on from the ambiguous noun ('car') Est.=0.92, SE=.31, t=2.97, $p < .01$), (**Figure 2**). In contrast, both adolescent groups only begin to anticipate the target in both conditions during the transitive verb region ('choose') Est.=0.64, SE=.30, t=2.12, $p < 0.5$. On participants' WM, IC and EQ, younger adolescents

performed significantly worse in all individual measures from older adolescents and adults but older adolescents did not differ from adults. On the on-line False-Belief task, adolescents did not differ from adults.

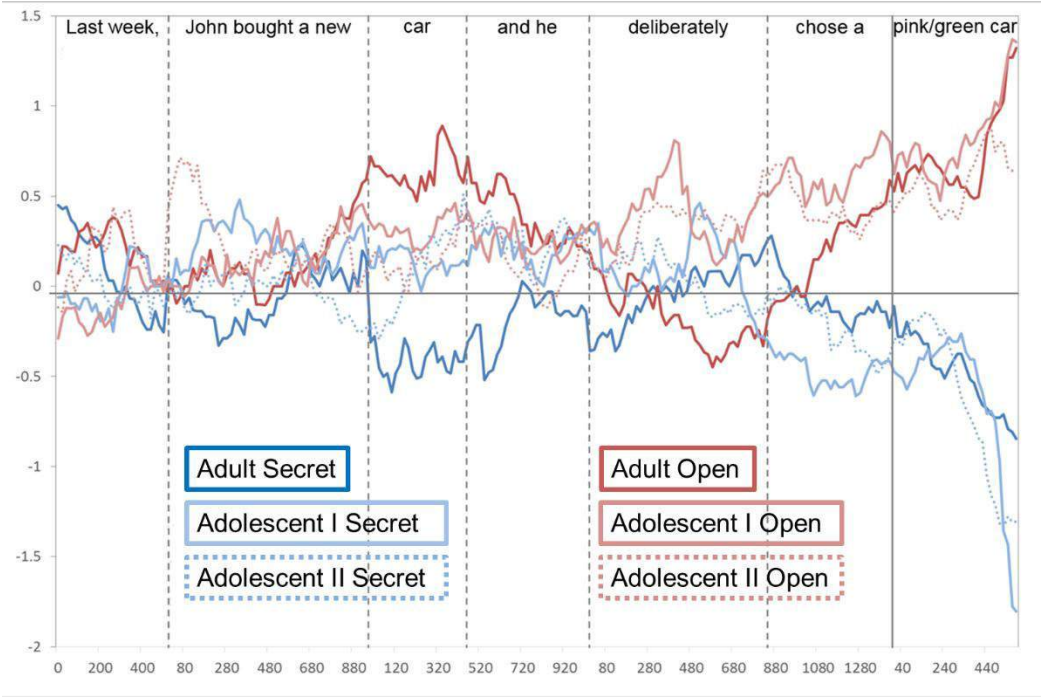


Figure 2. The average log(Reality/Belief) score for each condition and age group. Note that the dashed and vertical lines indicate absolute onsets and average offsets of words in the target sentence.

Discussion. These results suggest age-related differences between adolescents and adults in their online use of ToM. Critically, when reasoning about the character's basic preference and high-order intentions, adults anticipated the target early on in the 'car' region. Adolescents only begin to anticipate the target in both conditions during the 'choose' region, suggesting that adolescents may not be able to use information about others' mental states for language comprehension as quickly as adults. Although WM and IC have been shown to be factors in the application of ToM in previous studies [2,5], our results suggest these are not the only factors in the on-line application of ToM. Our results also point to the conclusion that not all on-line comprehension tasks that involve ToM are equal in terms of ToM inferential complexity, and that such complexity does not necessarily correlate with other task demands, like IC.

References:

[1] Blakemore, S-J. (2008). *Nature Reviews Neuroscience*, 9 (4), 267-277. **[2]** Symeonidou, I., Dumontheil, I., Chow, W., Breheny, R. (2016). *JECP*, 149, 81-97. **[3]** Ferguson, H. & Breheny, R. (2011). *Cognition*, *119*, 179-196. **[4]** Ferguson, H. J., Apperly, I., Ahmad, J., Bindemann, M., & Cane, J. (2015). *Cognition*, 139, 50-70. **[5]** Brown-Schmidt, S. (2009b). *Psychonomic Bulletin & Review*, *16*(5).

The Aging Factor in Presuppositions Processing

Debora Rossi, Simona Di Paola and Filippo Domaneschi

Department of Educational Sciences, Psychology Unit - University of Genoa, Italy

Introduction: In the psycholinguistic literature, the decline in pragmatic processing with aging in healthy subjects has been studied mainly in relation with two typical pragmatic processes: the turn-taking system (Murphy et al. 2006) and figurative language (Byrd et al. 1991). Apparently, what is still missing is a research line on the effect of age on the processing of presuppositions (PSPs), namely, the information implicitly communicated as taken for granted, which are another core level of pragmatic processing.

Research questions: The present work takes a first step in the direction of a psycholinguistic investigation on the potential impact of the aging factor on the processing of PSPs. In particular, we address three research questions: (i) does the processing of PSPs in online language comprehension involve higher processing costs with healthy older adults as compared to younger speakers? (ii) Does the aging factor affect the ability to recover from the discourse mental model information introduced as presupposed? (iii) Does the aging factor affect the ability of updating the discourse mental model with presupposed information?

Materials: 60 three-sentence stories have been created, with two context sentences and a target sentence. Target sentences contained a definite description (DD) or a change-of-state verb (CSV) presented in condition of satisfaction (SAT) or of accommodation (ACC) – Figure 1. Each story was followed by 3 questions: a *target question* verifying the content of the presupposition activated by the target sentence and two *distractor questions.*

Methods and procedure: In a self-paced reading times paradigm (cf. Tiemann et al. 2011), 21 young adults (mean age: 22.47) and 20 elderly adults (mean age: 63.6) read the stories and answered the 3 true/false questions. *Context sentences* were presented as a whole on the screen. The *target sentences* were presented word-by-word. A *Verbal Working Memory Ability* test was administered too. We collected: (i) participants' word-by-word reading times on the target sentences; (ii) response times to the target questions, and (iii) accuracy (i.e. correct responses to target questions).

Age-related Results: *Word-by-word Reading times* (e.g. sentence region: T1= *give up*, T2+1= *pictures*; see Figure 1) – Figure 2. *T1*: all participants were slower in ACC than in SAT (Condition: $p<0.05$) and even more with DDs than CSVs ($p<0.001$). *T2+1*: processing a PSP was costlier for elderly than for younger adults ($p<0.05$), with elderly participants' higher processing costs for CSVs than for DDs ($p<0.05$). *Response times* (Figure 3) revealed a significant GroupXConditionXTrigger Type interaction ($p<0.0001$): recovering an accommodated PSP triggered by DDs elicited longer response times for elderly subjects'. *Accuracy*: no Group effect was observed. *Working Memory*: WM differently accounts for the recovering of information triggered by CSVs in condition of accommodation depending on age group (CondXGroupXTriggerXWM= $p<0.05$): even for elderly participants with higher WM scores response times were longer ($\beta=1237.840$; $t=1.755$) than the younger participants', for whom the higher the WM scores the faster the response times ($\beta= -115.254$; $t= -0.355$).

Discussion: data collected show that aging affects PSPs processing. First, in online language comprehension older adults exhibit higher processing costs with CVSs (*Reading times*), presumably because they involve a demanding temporal mental representation (Domaneschi et al. 2014). Second, since PSPs constitute a condition for the understanding and appropriateness of

an utterance, updating the mental discourse model with presupposed information does not seem to decline across the life span (*Accuracy*). Rather, what seems to decline is the ability to recover from the discourse mental model information introduced in the context as presupposed (*Response times*). Finally, beyond other more explored levels of pragmatic processing, the decline with age of working memory ability seems to affect PSPs processing.

Trigger Type	Condition	Context sentence 1	Context sentence 2	Target sentence
DD	*SAT*	Enrico and Marta will have dinner in *a pub with a pianist* tonight.	They have chosen this pub to celebrate their first wedding anniversary.	After the dinner, because of the anniversary, *the pianist of the pub* will sing a serenade.
	ACC	Enrico and Marta will have dinner in a very suggestive pub tonight.		
CSV	*SAT*	Paolo really loved *taking pictures* on the mountains during the weekend.	He has always used his free time to go along with his passions.	Now he works on Sundays, so he *has given up taking pictures* because it was time-consuming.
	ACC	Paolo has always devoted his weekends to his hobbies.		

Figure 1 Example of an item with DD and CSV in condition SAT and ACC (literal translation from Italian).

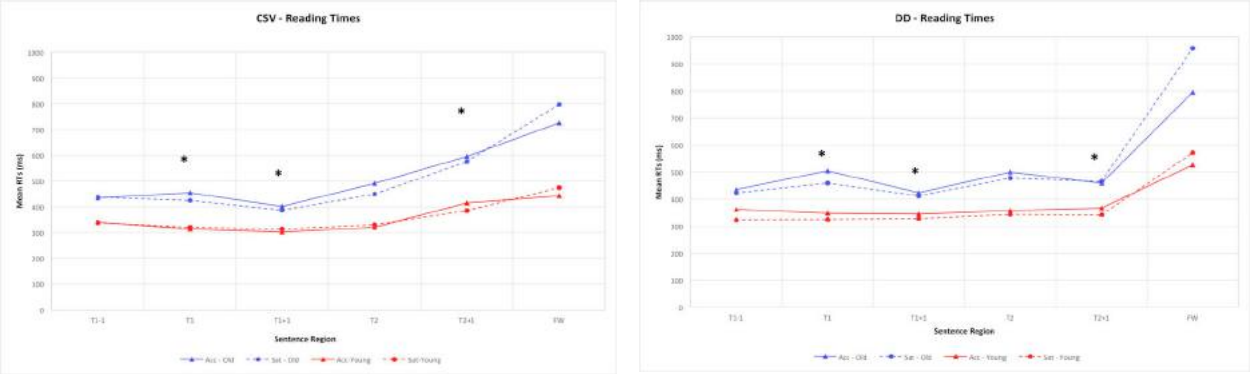


Figure 2. Reading times at different sentence regions for CSVs (2a) and DDs (2b) across conditions and groups.

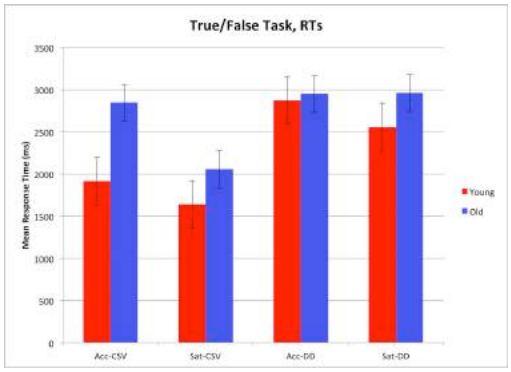


Figure 3 Response times on the true/false for CSVs and DDs across groups in the two experimental conditions

References

Byrd, M. (1991). Adult age differences in the ability to read and remember metaphor, *Educ Gerontol*, 17(4), 297–313.

Domaneschi, F., Carrea, E., Penco, C., & Greco, A. (2014). The cognitive load of presupposition triggers: mandatory and optional repairs in presupposition failure, *Language, Cognition and Neuroscience*, 29(1), 136-146.

Murphy, D. R., Daneman, M., & Schneider, B.A. (2006). Why do older adults have difficulty following conversations?, *Psychol Aging*, 21(1), 49–61.

Tiemann, S., Schmid, M., Bade, N., Rolke, B., Hertrich, I., Ackermann, H., Knapp, J., & Beck, S. (2011). Psycholinguistic evidence for presuppositions: on-line and off-line data. In I. Reich et al. (Eds.), *Proceedings of Sinn & Bedeutung*, 15 (pp. 581-595). Saarbrücken: Saarland University Press.

Top-down and bottom-up cues to speech acts

Ye Tian[1] & Chris Cummins[2]
[1]Laboratoire de linguistique formelle, Université Paris Diderot ; [2]Department of Linguistics and English Language, University of Edinburgh

When we are presented with an utterance, we not only interpret its semantic meaning, but also its discourse purpose, and, on the other side of the coin, its relevance to the broader context. These two notions are often construed in the frameworks of speech act and Question Under Discussion (QUD) (Ginzburg, 2012; Roberts, 2012; to appear). This reasoning can take two approaches: a top-down approach, where we reason about the speaker's discourse goals and consider how the utterance contributes to their realisation; or a bottom-up approach, where we use the content of the utterance itself as a basis for inferring what the speaker is attempting to achieve. Computationally, these two approaches can be synthesised within a cue-based approach to speech act recognition/ disambiguation, in which high-level and low-level considerations are both used as probabilistic cues to the successful classification of a speech act. However, little is known about how these two sources of information are integrated by hearers in offline interpretation and in online processing. In this paper, we argue that a better understanding of this process is necessary, not only because of its theoretical implications for our analysis of discourse in general, but also because of its methodological implications for experimental semantics and pragmatics.

We all know that our understanding of the overarching discourse goals influences the speech act interpretation of an utterance. What we don't know is what hearers do when extended context is unavailable. They may use a bottom-up approach and reason about contextual relevance using local cues, or they may use a top-down approach and imagine a (stack of) discourse goals which allow rich inferences to be drawn on the current utterance. For example, if someone hears (1) out-of-the-blue, they might accommodate the QUD (2) and interpret (1) literally; or they might imagine a richer discourse goal, accommodate the QUD (3), and interpret (1) to imply that the date wasn't good.

(1) The coffee was not bad.
(2) Was the coffee bad?
(3) How did your date go?

Experiment: inferences about prior and subsequent context for speech acts

The goal of our research is to shed light on the interplay between high-level and low-level factors that bear upon the recognition of speech acts. A first step is to determine the degree of variability in the inferences participants are able or willing to draw based on decontextualized utterances, about the nature of the current discourse context. We uses a Cloze task in which participants are presented with isolated utterances and asked to suggest the context in which these utterances took place – either providing the preceding turn, the following turn, or both, according to their preference. If the bottom-up approach is primary, we would expect more consistent and literal interpretations; if the top-down approach is primary, we would expect more varied and enriched interpretations.

We constructed 34 potentially ambiguous utterances as experimental items. They were constructed such that each could be interpreted as instantiating at least two distinct speech acts, as shown for examples (5) to (7) below. In addition, 26 unambiguous fillers were

constructed. 63 participants were recruited via Amazon Mechanical Turk. They read each utterance and could fill in the preceding utterance, the following utterance, or both.

(5) I am on my way. (ambiguous between answer to a "where" question and acceptance of a request)
(6) I'm not working on Saturday. (ambiguous between answer to a question and acceptance of a suggestion/invitation)
(7) Are you wearing that shirt? (ambiguous between question and complaint)

Results: for each response, we coded the fine-grained speech act (e.g. "information-seeking question", "offer" etc.), as well as whether the approach is "bottom-up" (literal interpretations) or "top-down" (enriched interpretations). The participants filled the preceding context in 77% of responses and the following context in 82% of responses. On average, each utterance received 4 different interpretations. Although declaratives received more distinct interpretations than interrogatives did, this may be a reflection of the content of the utterances rather than their sentence type. Among the interrogatives, we included some items that have been argued strongly to cue particular "non-literal" interpretations, and these potentially conventionalised items did indeed appear to admit more homogeneous interpretations. Items such as "Can you find out the name of this song?" and "Can you fix my bike?" were widely interpreted as requests, although the formally similar "Can you email your boss?" was variously interpreted as either a request (5) or as a suggestion (9).

(8) B: Can you email your boss? A: When I get the chance, sure.
(9) A: What should I do about this situation?
B: Can you email your boss? A: Yeah that's a good idea.

69% of interpretations were top-down, with interrogative sentences attracting the highest rate of top-down interpretations (including requests, offers, and implied answers to previous questions). An interpretation is more likely to be top-down when the preceding context is filled (71% top-down) than when it is left empty (60%). Whether the following context is filled doesn't make a difference in the proportions of top-down readings.

Table 1 summary of results

Type of item	number of items	% preceding cntxt filled	% following cntxt filled	number of speed acts interpreted	% top-down readings
positive declarative	14	84%	75%	4.21	62%
negative declarative	6	82%	79%	3.83	53%
interrogative	14	67%	90%	3.07	83%

Overall, our results show that when presented with utterances context out of the blue, both approaches can be used. More often, participants take the 'top-down' approach: attribute rich overarching conversational goals and thereby derive additional pragmatic interpretations of utterances. Inferring from our data to speech-act recognition in natural settings (when at least some context is available), it is likely that top-down considerations play a more primary role interpreting the speech act of an utterance and its discourse relevance to the broader context. Our results also demonstrates that certain utterances are indeed ambiguous when presented out of context, to an extent that makes them suitable for further experimental investigation of the online interplay between top-down and bottom-up interpretative processes.

References: Ginzburg, J. (2012). The interactive stance. CSLI: Center. Oxford University Press. Roberts, C. (2012). Information structure: toward an integrated theory of formal pragmatics. *Semantics and Pragmatics*, 5(7):1-19. Roberts, C. (to appear). Speech acts in discourse context. To appear in D. Fogal, D. Harris and Matt Moss (eds.), *New Work on Speech Acts*. Oxford University Press.

Over-specification and uniform reduction of visual entropy facilitate referential processing

Elli N. Tourtouri, F. Delogu, Matthew W. Crocker
Saarland University

Over-specifications (OS) are expressions that provide more information than minimally required for the identification of a referent, thereby violating Grice's 2nd Quantity Maxim [1]. In Figure 1, for example, the expression "Find the blue ball" identifies exactly one object in all panels, but only in the top displays is the adjective required to disambiguate the target. In recent years, psycholinguistic research has tried to test the empirical validity of Grice's Maxim, resulting in conflicting findings. That is, there is evidence both that OS hinders [2,3] and that it facilitates [4,5] referential processing. The current study investigates the influence of OS on visually-situated processing, when the context allows both a minimally-specified (MS) and an OS interpretation of pre-nominal adjectives (cf. Fig.1). Additionally, as the utterance unfolds over time, incoming words incrementally restrict the search space. In this sense, information on "blue" and "ball" is determined not only by their probability to occur in this context, but also by the amount of uncertainty about the target they reduce — in information theoretic terms [6]. A greater reduction of the referential set size on the adjective (A&C) results in a *more uniform* reduction profile (Uniform Reduction, UR), as the adjective reduces entropy by 1.58 bits and the noun by 1 bit. On the other hand, a moderate reduction of the set size on the adjective (B&D) results in a *less uniform* reduction profile (Non-uniform Reduction, NR): the adjective reduces entropy by .58 bits and the noun by 2 bits. This study examines whether, above and beyond any effects of specificity, the rate at which incoming words reduce visual entropy also affects referential processing.

Methods. We conducted an eye-tracking experiment crossing **Specificity** (MS vs. OS) and **Entropy Reduction** (UR vs. NR). Participants (N=24, mean age=25) were presented with displays such as the ones in Figure 1, and after 2sec heard an instruction in German to *Find the* ADJ TARGET, mentioning either the colour or pattern of the target object. Research on the production of OS has demonstrated that they are commonly used by adult speakers, both pre- and post-nominally [7,8,9], and with various types of adjectives [10]. As rational speakers would unlikely encode redundant information so consistently, if it hindered listeners' processing, we expected OS to be more or, at least, as beneficial as MS for referential processing, and, as found in [5], we expected to observe this effect on the noun. Regarding Entropy Reduction, either of two outcomes were expected: a) a preference for UR, indicating that expressions reducing *visual entropy* more uniformly across the utterance are more efficient for referential processing — consistent to the predictions of UID [11] — b) a facilitation for NR, suggesting that the *gradual restriction of referents*,

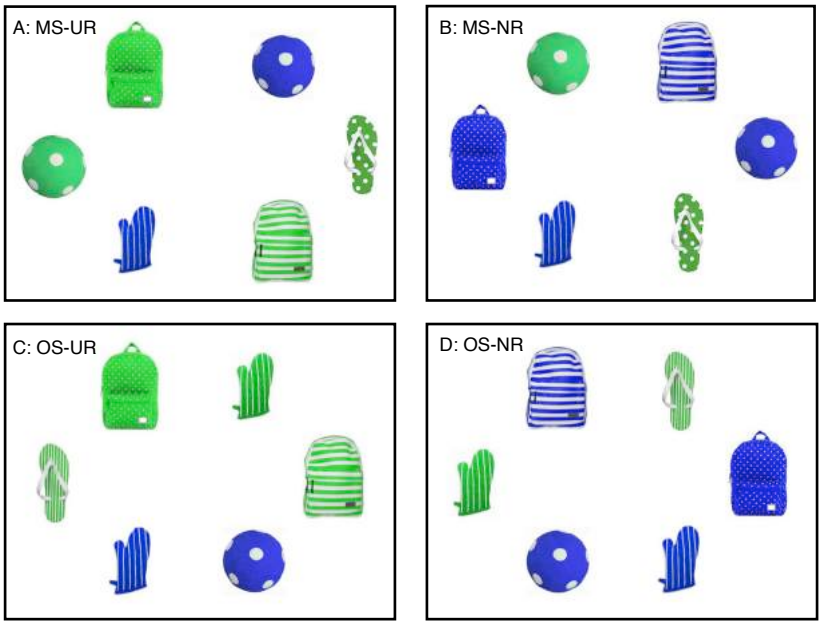


Figure 1. Sample visual stimuli, combined with the instruction "Find the blue ball".

rather than the rate of entropy reduction, facilitates processing. To examine these questions we compared inspection probabilities to objects of interest in the adjective and noun regions across conditions. As information about the target became incrementally available, different comparisons were interesting per region. On the adjective, since the target object was not yet known, we were interested in inspections to *single* (cf. the mitten in A&B, and the ball in C&D) and *contrast* (cf. the blue ball in A&B, and the blue mitten in C&D) objects in UR vs. NR. On the noun, we compared inspections to the *target* (the blue ball: MS in A&B, and OS in C&D) across conditions. In addition, we present results from the Index of Cognitive Activity (ICA), a novel measure of cognitive effort that is based on rapid pupil dilations that are due to load reflex, separating them from those due to light reflex or noise [12,13]. Higher ICA values are associated with greater cognitive workload. Finally, we also report Reaction Times across conditions.

Results. On the adjective, only one comparison yielded significant results, i.e. *contrast* objects were inspected more frequently in UR vs. NR.[1] On the noun, analyses of inspection probabilities to the *target* produced effects for colour items. Specifically, there was a marginal effect of Specificity, with more inspections in OS vs. MS (p=.06), suggesting a preference for OS. Furthermore, we observed a main effect of Entropy Reduction, with more inspections to the target in UR vs. NR (p=.048). Analyses of the ICA produced main effects of Entropy Reduction and Specificity for both colour and pattern items (cf. Fig.2), such that ICA values were lower for UR vs. NR (p=.003) and for OS vs. MS (p<.001). RTs also resulted in two main effects, with faster responses in UR vs. NR (p=.002) and OS vs. MS (p=.023).

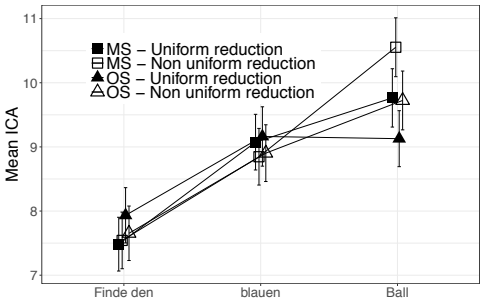


Figure 2. Mean ICA per condition and region. Error bars represent 95% CI.

Discussion. We present evidence confirming previous findings that redundant adjectives facilitate processing of the upcoming noun in situated comprehension [5]. Even though only colour items yielded higher inspection probabilities for OS vs. MS, ICA values and RTs were reduced also for pattern items, suggesting that, while pattern is less salient than colour, its mention is similarly beneficial. These results indicate a general advantage for OS in referential processing. In addition, we showed that uniform reduction of visual entropy, resulting from a more drastic decrease of referents on the adjective — while not accompanied by greater load in that region — is associated with a reduced cognitive effort when processing the noun. We entertain two explanations regarding the absence of an Entropy Reduction effect on the adjective. First, it may be that our manipulation of entropy reduction on the adjective was not that distinct between UR and NR. Secondly, it is possible that entropy reduction elicits end-state effects, showing up after the full entropy reduction profile of an utterance has evolved, and such effects cannot be observed "on the fly". We conclude that efficient processing is determined by both the degree of specificity of the reference, and its contribution to the uniform reduction of visual entropy across the utterance.

References. [1] Grice (1975) In Cole & Morgan. [2] Engelhardt et al (2011) *Brain Cogn* [3] Davies & Katsos (2013) *J pragmat* [4] Arts et al (2011) *J pragmat* [5] Tourtouri et al (2015) *CogSci* [6] Hale (2006) *Cognitive Sci* [7] Pechmann (1989) *Linguistics* [8] Engelhardt et al (2006) *JML* [9] Rubio-Fernández (2016) *Front Psychol* [10] Tarenskeen et al (2015) *Front Psychol* [11] Jeager (2010) *Cognitive Psychol* [12] Marshall (2000) US Patent 6,090,05 [13] Demberg & Sayeed (2016) *PLoS ONE*

[1] Since in NR more entities bear the mentioned feature (cf. the blue rucksacks in B&D), attention is spread across more objects. Therefore, we do not take this result to reflect any preference for a contrastive (MS) interpretation of the adjective.

Approaching the pragmatics of exclamations experimentally

Andreas Trotzke (Stanford University & University of Konstanz)

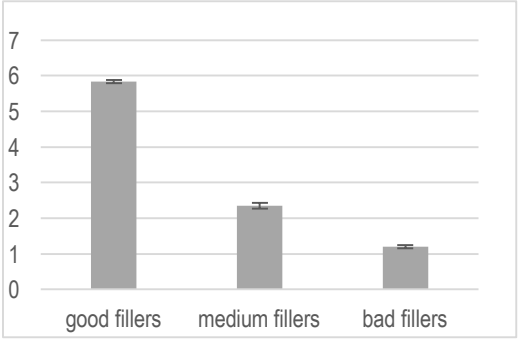
A standard assumption is that sentence exclamations like (1) count as an assertion and can thus be denied, whereas exclamative cases such as (2) do not make a contribution to the discourse that could be denied (or affirmed) directly. It is controversial whether exclamatives can nevertheless be 'weakly denied' by phrases like *not really* etc. (examples and judgments by Rett 2008, 2011):

(1) A: (Wow,) John bakes delicious desserts!
B: No (he doesn't), these are store-bought. John's actually a terrible cook.

(2) A: (My,) What delicious desserts John bakes!
B: ?No (he doesn't), these are store-bought. John's actually a terrible cook.
B': Not really; these are store-bought. John's actually a terrible cook.

These judgments have so far not been assessed empirically. Our paper follows the methodology of an increasing number of studies in pragmatics that incorporate experiments in order to obtain reliable and robust judgments (Sauerland & Schumacher 2016). Specifically, we reexamine a prominent theory of exclamations (Rett 2011): It is argued that the difference we see in (1) and (2) falls out of the fact that only exclamative clauses and not sentence exclamations denote degree properties and not propositions. That is, while (1) can be associated with a non-scalar expectation (i.e., that the desserts John bakes would not be delicious), (2) can only be associated with a scalar expectation (that the desserts John bakes would not be as delicious as they are). In our study, in addition to cases like (1) and (2), we also included a potentially interesting construction from Germanic languages other than English: German *dass*-exclamatives (see also Dutch) display a dedicated exclamative syntax (lacking V-to-C movement) and, at the same time, do not allow the scalar-expectation reading of *wh*-exclamatives (d'Avis 2002; Truckenbrodt 2013).

Materials. The experimental items were manipulated at two levels: EXCLAMATION FORM, that is, whether the relevant case is a sentence exclamation (6), a *wh*-exclamative (7), or a *dass*-exclamative (8), and DENIAL, that is, whether the utterance by Speaker B is a strong (SD) or a weak denial (WD), see below. For each combination, there were four examples. Sentence exclamations can also be associated with scalar expectations (e.g., accomplished by using focus on the adjective). To ensure that sentence exclamations receive a non-scalar interpretation, cases included explicit degree statements featuring deictic *so* ('so'), which blocks a scalar reading of the whole exclamation (Truckenbrodt 2013). In addition, we constructed four fillers we expected to get good judgments ('good' fillers, [3]), four fillers we expected to get bad judgments ('bad' fillers, [4]), and four fillers we expected to receive mixed judgments ('medium' fillers, [5]). Taken together, there were thus 36 stimuli in total; stimuli were divided into 2 lists, each consisting of 24 items. ***Participants.*** We collected judgments from 112 native German speakers. The experimental items were presented through an online questionnaire, and participants had to rate the acceptability of Speaker B's reactions on a scale ranging from 1 (= very bad) to 6 (= very good).

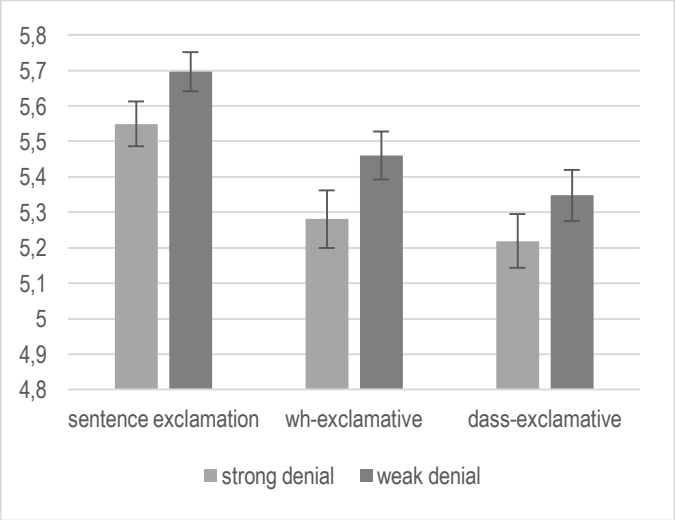


(3) A: Linda hat einen schlauen Sohn.
'Linda has a smart son.'
B: Nein, das stimmt nicht.
'No, that's not right.'

(4) A: Wie ist sein Name?
'What's his name?'
B: Nein, das stimmt nicht.
'No, that's not right.'

(5) A: Hat Boris gestern eingekauft?
'Has Boris done the shopping yesterday?'
B: Nein, das finde ich nicht.
'No, I disagree.'

(6) A: Wow! Peter kann so lecker kochen!
wow Peter can so delicious cook
'Wow! Peter is such a great cook!'
B: {SD: Nein, / WD: Nicht wirklich,}
er wärmt immer nur Fertiggerichte auf.
'{SD: No, / WD: Not really,} he always warms up convenience food.'



(7) A: Wahnsinn! Was für schwierige
madness what for difficult
Matheaufgaben Katrin lösen kann!
math.problems Katrin solve can
'Man! What difficult math problems Katrin can solve!'
B: {SD: Nein, / WD: Das stimmt nicht ganz,} sie schlägt immer im Lösungsbuch nach.
'{SD: No, / WD: That's not quite right,} she always looks the solution up in the textbook.'

(8) A: Wow! Dass die Maria so
wow that the Maria so
schön aussieht!
beautiful looks
'Wow! I'm amazed that Maria is so beautiful!'
B: {SD: Nein, / WD: Nicht wirklich,} sie benutzt lediglich sehr viel Make-up.
'{SD: No, / WD: Not really,} she just uses a lot of makeup.'

Results. Figure 1 shows that fillers were judged as we had expected. Acceptability of bad fillers was lowest (1.2), acceptability of medium fillers was about in the middle of the provided scale (2.4), and acceptability of good fillers was at ceiling (5.8). The results of a one-way ANOVA of FILLER TYPE on acceptability judgments show that the main effect of FILLER TYPE on acceptability judgments was highly significant ($F(1304, 101) = 652.13$, $p < .001$). These data on fillers show that participants not only understood the task well, but that they also used the full range of options for their judgments. Figure 2 shows that weak denial is always preferred over strong denial, also in the case of sentence exclamations. A two-way ANOVA (3 x 2) revealed significant main effects of both EXCLAMATION FORM ($F(14, 48) = 6.96$, $p < .001$) and DENIAL ($F(4, 32) = 3.87$, $p < .001$), but there was no significant interaction ($F(.07, 36) = .04$, $p > .05$). Overall, it is striking that all judgments of exclamation items were in accordance with our category of 'good fillers' and thus at ceiling (ranging from 5.2 to 5.7), suggesting that the often-cited infelicity of certain reactions to particular exclamation forms (e.g., strong denial in the context of *wh*-exclamatives) is actually a very subtle matter. However, paired t-tests show that the difference between strong and weak denial is significant within both the sentence-exclamation ($p < .01$) and the *wh*-exclamative condition ($p < .01$), but not significant in the *dass*-exclamative condition ($p > .05$), supporting the theoretical claims in the literature that semantic content featuring non-scalar expectations (as in *dass*-exclamatives) increases the acceptability of strong denial in the context of exclamations.

Pragmatic impairment is selective in autism: evidence from quantity implicatures

Bob van Tiel & Mikhail Kissine — Université Libre de Bruxelles

Pragmatic deficits have long been recognised as one of the main nosological markers of Autism Spectrum Disorder (ASD). The latest edition of the DSM thus states that people with ASD tend to exhibit "[d]ifficulties understanding what is not explicitly stated (e.g., making inferences) and nonliteral or ambiguous meanings of language" (p. 48). In spite of this diagnostic criterion, a number of studies have found that people with ASD derive *scalar inferences* (e.g., the inference from "some" to "some not all") at the same rate as neurotypicals (e.g., Pijnacker et al., 2009).

One might conclude from this finding that people with and without ASD are equally adept when it comes to reasoning about the speaker's intentions for being underinformative. Such a generalisation would be in line with the social motivation theory of ASD, which holds that people with ASD are in fact capable pragmatic reasoners but often lack the motivation to engage in pragmatic reasoning (Chevallier et al., 2010). There are, however, compelling reasons to doubt that findings for scalar inferences can be generalised across the entire family of quantity implicatures.

In particular, scalar inferences have two features that are not shared by all varieties of quantity implicature. First, scalar infererences are closely connected to certain lexical expressions, to the extent that a number of theorists have argued that they are an aspect of lexical meaning rather than involving pragmatic inferencing (e.g., Levinson, 2000). Second, assuming that scalar inferences are bona fide inferences, their derivation is simple in that it can be reduced to constructing and negating alternatives, without considering the speaker's beliefs and intentions. In other words, scalar inferences are *lexicalisable* and their derivation is potentially *non-mentalistic* (LEX+ / MENT–).

In order to determine whether these two features shaped the observation that people with ASD derive scalar inferences at the same rate as neurotypicals, and more generally to what extent people with ASD are able to reason about the speaker's intentions for being underinformative, we extended the scope of investigation to four types of inferences that are often explained as quantity implicatures: scalar inferences, distributivity inferences, conditional inferences, and exhaustivity inferences.

Scalar inferences (LEX+ / MENT–)
Some of the shapes are red.
⇝ Not all of the shapes are red.

Distributivity inferences (LEX– / MENT+)
Each of the shapes is red or green.
⇝ There are both red and green shapes.

Conditional inferences (LEX– / MENT–)
Each of the shapes is red if it is a circle.
⇝ Not all of the shapes are red.

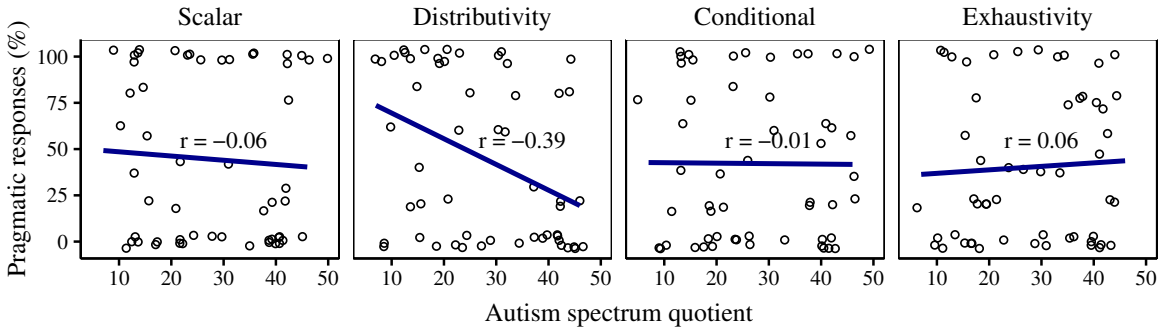
Exhaustivity inferences (LEX± / MENT–)
It is the circle that is red.
⇝ Only the circle is red.

Scalar inferences but not distributivity and conditional inferences are lexicalisable. Whether or not exhaustivity inferences can be encoded in the semantics of clefts is a matter of current debate. The derivation of scalar, conditional, and exhaustivity inferences is potentially non-mentalistic in that it can be reduced to constructing and negating alternatives. By contrast, the derivation of distributivity inferences also involves reasoning about what the speaker would have implicated had

she uttered one of the alternatives (or, in grammaticalist parlance, their derivation involves double exhaustification, cf. Crnič, Chemla, & Fox, 2016 for experimental evidence).

Following van Tiel and Schaeken (2017), we conducted a sentence-picture verification task. 62 participants with (28) and without (34) an official diagnosis of ASD (mean age: 34, range: 18–64, 32 females) read sentences that were followed by a picture and had to indicate if the sentence was true or false in that picture. Pictures showed either two (exhaustivity inferences) or five (the other three varieties of quantity implicature) coloured geometrical shapes. In target situations, sentences were true on their literal construal but false if the corresponding quantity implicature was derived. In control situations, sentences were unambiguously true or false. Example sentences, corresponding quantity implicatures, and target situations are provided in the figure above.

Participants also filled out the autism spectrum quotient (AQ) test (Baron-Cohen et al., 2001). The AQ test is a self-test consisting of 50 multiple choice questions and provides a measure of the extent to which one exhibits traits that are symptomatic of ASD. The figure below plots for each participant their AQ and the proportion of pragmatic (i.e., 'false') responses.



The results confirm previous observations that the rate of pragmatic responses for scalar inferences is independent of one's AQ. The same pattern was found for conditional and exhaustivity inferences (all Z's < 1). However, the proportion of pragmatic responses for distributivity inferences significantly decreased with one's AQ (β = -0.17, SE = 0.08, Z = -2.18, p = .03). Indeed, the effect of AQ on the rate of pragmatic responses for distributivity inferences significantly differed from its effect on the other varieties of quantity implicature (all p's $< .01$). These results were confirmed when participants were categorised based on whether they had been diagnosed with ASD.

These results indicate that the observation that people with ASD derive scalar inferences at the same rate as neurotypicals cannot be ascribed to these inferences being lexicalisable, since the same result was found for conditional and exhaustive inferences. However, people with ASD experience difficulties when the derivation of inferences involves more complex reasoning about the speaker's mental states, as was the case for distributivity inferences. Hence, structural differences in the derivation procedure affect the ease of computing pragmatic inferences for people with ASD. Interestingly, these difficulties were not reflected in the response times, which were equally high as for scalar inferences. The pragmatic deficits of people with ASD are thus selective, which speaks against the social motivation theory of ASD and a monolithic conception of pragmatics in general.

References: **[1]** Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). In: *J Autism Dev Disord*, *31* (1), 5–17. **[2]** Chevallier, C., Wilson, D., Happé, F., & Noveck, I. (2010). In: *J Autism Dev Disord*, *40* (9), 1104–1117. **[3]** Crnič, L., Chemla, E., & Fox, D. (2015). In: *Nat Lang Semant*, *23* (4), 271–305. **[4]** Levinson, S. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. **[5]** Pijnacker, J., Hagoort, P., van Buitelaar, J., Teunisse, J.-P., & Geurts, B. (2009). In: *J Autism Dev Disord*, *39* (4), 607–618. **[6]** van Tiel, B. & Schaeken, W. (2016). Forthcoming in: *Cognitive Sci*.

When children accept under-informative utterances: Lack of competence or pragmatic tolerance?

Alma Veenstra and Napoleon Katsos
University of Cambridge

Binary judgment on under-informative utterances (e.g., judging the truth of a sentence such as "*some horses jumped over the fence*" in a situation where all horses did) is the most widely used methodology to test children's ability to generate implicatures. Accepting under-informative utterances is considered a failure to generate scalar implicatures. Children who do not realize that there is a more informative alternative that the speaker could have used will accept the under-informative utterance, as it is logically—but not pragmatically—true. Studies following this reasoning have concluded that children as old as 8 and 10 years have not yet acquired scalar implicature [1], and that with training, explicit instruction and helpful context, the performance of young children is still rather unstable [2][3][4].

Another, more recent, line of research argues that it is the binary judgment paradigm that possibly obscures children's true performance. The Pragmatic Tolerance Hypothesis posits that although under-informative utterances are pragmatically infelicitous, some children may not find this violation grave enough to warrant a downright rejection when asked whether the utterance is right or wrong. When given multiple options, instead of two as in the binary judgment paradigm, children as young as 5 years do seem sensitive to under-informativeness and no longer opt for the most positive option [7][8][9].

We present off-line and response time evidence for the Pragmatic Tolerance Hypothesis. Seventy-five Dutch-speaking 4- to 9-year-old children completed both a binary judgment task (Experiment 1) and a ternary judgment task (Experiment 2). In Experiment 1, the participants were asked to judge the statements of a fictional character about a visual display. Critically, some of the utterances were under-informative, for instance when the character said "*in the basket there is a shoe*" when there were both a shoe and a ball in the basket. Judgments and response times were collected. In Experiment 2, the participants were asked to reward the character with a small, medium, or large strawberry, corresponding to how accurate they judged her description. Here, only participants' judgments were collected.

Comparing the results from Experiments 1 and 2 revealed that there were three main types of participants: children who accepted under-informative utterances in the binary judgment task and opted for the large strawberry in the ternary judgment task; children who accepted under-informative utterances in the binary judgment task and opted for the small or medium strawberry in the ternary judgment task; and children who penalized under-informative utterances in both tasks. We argue that this demonstrates a developmental pattern, where children evolve from pragmatically oblivious speakers to pragmatically tolerant speakers to fully competent pragmatic speakers.

Half of the participants who accepted under-informative utterances in Experiment 1, penalized them in Experiment 2. The response times in Experiment 1 showed that these children experienced a significant slow-down in the under-informative utterances compared to simple true utterances, suggesting that they detected the pragmatic violation even though they did not reject it. In contrast, the participants who accepted under-informative utterances in both tasks, did not show this slow-down in the binary judgment task and were equally fast in accepting under-informative utterances as they were in accepting simple true utterances. Had we only used the judgments from the binary task, these pragmatically tolerant children would have been incorrectly categorized as not having yet acquired implicature.

Taken together, these results suggest that data from binary judgment tasks should be interpreted with caution as they seem to systematically underestimate children's competence

with pragmatics. In addition, other measures, such as response times, are necessary to distinguish between children who accept under-informative utterances due to a lack of pragmatic competence and children who accept such utterances due to pragmatic tolerance.

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

[2] 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.

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

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

[5] Davies, C., & Katsos, N. (2010). Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations? *Lingua, 120*(8), 1956-1972.

[6] Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67-81.

[7] Katsos, N., & Smith, N. (2010). Pragmatic tolerance and speaker-comprehender asymmetries. In Franich, K., Iserman, K. M., and Keil, L. L. (eds.), *Proceedings of the 34th Boston University Conference in Language Development*. Cascadilla Press, MA, USA, pp 221-232.

Scalar Implicatures And The Literal-First Hypothesis: Theory Of Mind And Working Memory Effects In Pragmatic Inferences By Patients With Psychosis

Martien Wampers[1] & Walter Schaeken[2]
[1]Research Group Psychiatry, University of Leuven, Belgium
[2] Brain and Cognition, University of Leuven, Belgium

Introduction: Successful social interactions rely heavily on one's ability to go beyond the explicit, literal content of conversational statements and grasp the actual, intended meaning for in daily communication. The message that one wants to express is often not explicitly mentioned. For decades, researchers have illustrated the difficulties patients diagnosed with psychosis experience when they have to decode the non-literal content of conversational statements. These difficulties include trouble grasping the figurative meaning of proverbs and metaphors and problems with understanding humor and irony (e.g., Bambini et al, 2016; Brüne & Bodenstein, 2005;Sponheim et al., 2003). In the present study, we aim to gain more insight in the ability of people with psychosis to derive scalar implicatures (SIs). SIs are among the most studied types of pragmatic inferences but, to the best of our knowledge, have not yet been studied in people with psychosis.

SIs are based on linguistic expressions like *some*, *or*, *must* etc. Such expressions are part of a scale of informativeness. Examples of such scales are: <*All/many/some* >, <*Must/may*>, <*Always/often/sometimes*>. The statement (1) "*Some patients were attentive"* will be generally interpreted as (2) "*Some but not all patients were attentive*" and not as (3) "*All patients were attentive*". However, on a strictly semantic level "*some*" means "*some and possibly all*". The (implicit) addition of "*but not all*" does not follow logically but is the result of a SI. A popular explanation for an SI starts with the observation that the speaker did not use the alternative "*All patients were attentive*" A likely explanation for not uttering the "*all*"-sentence is that this sentence is not the case, otherwise she, being a cooperative speaker, would have said so. Combining the previous premises leads therefore to the interpretation that the speaker intended to say that she has a good relationship with some but not all of her colleagues. One of the consequences of this view is that listeners access the literal interpretation of an utterance before computing conversational implicatures such as scalar inferences. This viewpoint is sometimes referred to as the literal-first hypothesis, and argues that the enriched interpretation is associated with a processing cost (e.g., De Neys & Schaeken, 2007; Degen & Tanenhaus, 2015; Noveck, 2001). However, not all theorists agree. Some argue that it is possible for an utterance to get an enriched interpretation right from the start without any processing costs. The precondition for this immediate and automatic enrichment to happen is that it is supported by the context (e.g., Chierchia, Fox, & Spector, 2012; Récanati, 1995). Both sides of the debate have emphasized the importance of psycholinguistic evidence to decide whether the literal-first hypothesis is correct. We believe that insights might come also from patients' studies. In the present study, we test in three experiments how people with psychosis respond to underinformative statements containing scalar expressions. We expect patients with psychosis to have problems deriving SIs. They will respond less pragmatically when confronted with the scalar expression "some" than controls.

Experiment 1: We focused on the scalars *some/all* and tested the hypothesis that, in a binary sentence verification task (true/false), patients with psychosis would choose the pragmatic interpretation of *some* (i.e., the "not all" interpretation) less often than controls, in favor of the logical interpretation (i.e., "all"). The patient group consisted of 25 adults diagnosed with schizophrenia according to DSM-IV by an experienced psychiatrist. All patients were outpatients. The second group, the control group, was matched to the patient group with respect to age and educational level On average, patients derived less SIs than controls, which is in line with our hypothesis and this difference was marginally significant. Moreover, the number of participants that consistently derives SIs, is significantly higher in the control group than in the patient group. These differences are not due to

differences in the ability to perform the task since both patients and controls attain high levels of accuracy on the filler items. Rather, these findings suggest that patients with schizophrenia are less likely to derive SIs.

Experiment 2: We tested the same hypothesis as in Experiment 1, but now in young hospitalized individuals with psychosis. The patient group consisted of 17 young psychotic patients, who were diagnosed with psychosis according to DSM-IV by an experienced psychiatrist. All but two patients were hospitalized. The control group was matched to the patient group based on age and educational level. Apart from the patient group, there were two changes compared to Experiment 1. Instead of a binary judgment task, we used a ternary judgment task. Additionally, we investigated whether the amount of pragmatic interpretations is associated with theory of mind (ToM) ability. At group level, patients preferred the logical interpretation, yet a clear association between ToM and the amount of logical answers was found. Only patients with an impaired ToM preferred the logical interpretation.

Experiment 3: We examined whether working memory (WM) influenced the amount of pragmatic responses. Moreover, different scalar implicatures (*might/must*, *warm/hot*, *or/and*, *good/excellent*, *big/enormous*) were studied. The patient group consisted of 21 adult psychotic patients, who were diagnosed with psychosis according to DSM-IV by an experienced psychiatrist. The control group was matched to the patient group based on age and educational level. Like healthy controls, individuals with psychosis showed scalar diversity: not all scalars were treated alike. In the clinical group, an effect of WM was observed for some of the scalars, but not all.

Discussion: Taken together, these results can only be interpreted in a nuanced manner. The general picture is in line with the predictions of the literal-first hypothesis, as indicated by the working memory and ToM-effect. However, the limitations of these effects make it clear that a strict literal-first view is not consistent with the data. Moreover, we will briefly address the claim that a decreased pragmatic ability is a core feature of psychosis.

References:

Bambini V., G. Arcara, M. Bechi, M. Buonocore, R. Cavallaro, M. Bosia (2016) The communicative impairment as a core feature of schizophrenia: Frequency of pragmatic deficit, cognitive substrates, and relation with quality of life. *Comprehensive Psychiatry 71*, 106-120.

Brüne, M. (2005). ‘‘Theory of Mind’’ in Schizophrenia: A Review of the Literature. *Schizophrenia Bulletin, 31*(1), 21-42. doi:10.1093/schbul/sbi002

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

De Neys W., & Schaeken W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. Experimental Psychology, 54(2), 128–133. doi:10.1027/1618-3169.54.2.128

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

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. Cognition, 78(2), 165–188. doi:10.1016/s0010-0277(00)00114-1

Récanati, F. (1995). The alleged priority of literal interpretation. Cognitive Science, 19(2), 207–232. doi:10.1207/s15516709cog1902_2

Speaker epistemic state and ad hoc quantity implicatures in children

Elspeth Wilson and Napoleon Katsos
Department of Theoretical and Applied Linguistics, University of Cambridge

Background Gricean and neo-gricean accounts of implicature assume that the listener takes into account the speaker's epistemic state. For example, the listener infers from the utterance 'John ate an apple for lunch' that *John ate only an apple* if he knows or can assume that the speaker is fully informed about John's lunch – this is known as the epistemic step (Sauerland, 2004) and has been demonstrated in studies with adults (e.g., Breheny, Ferguson & Katsos, 2013; Politzer-Ahles & Fiorentino, 2013). Children succeed with ad hoc implicatures from 3 years, where relevant information is in common ground (e.g., Stiller, Goodman, & Frank, 2015), and are able to match an under-informative utterance to a partially-knowledgeable speaker from 5 years (Hochstein, Bale, Fox & Barner, 2014; Papafragou, Friedman & Cohen, 2016). They also start to reason about others' epistemic states relatively early, for example predicting another's actions based on their false belief from age 3 or 4 (Wellman, Cross & Watson, 2001). Here, we present the first study to our knowledge that investigates children's ability to take into account speaker epistemic state in ad hoc implicature derivation. Our findings support a two-step developmental trajectory: first, children learn pragmatic inferences (assuming full relevant common ground) and reasoning about epistemic states, and then, second, learn to integrate the two processes.

Experimental study We tested English-speaking children aged 5;3-6;4 (N=34) and adults (N=36) in a novel experimental design that combined an ad hoc implicature picture-matching task (Horowitz & Frank, 2015), with the director task testing reference and perspective-taking (e.g., Nilsen & Graham, 2009) – Figure 1. Participants collected double-sided picture cards and put them in a 'card box', following the puppet's instructions 'pick the card with Xs'. There were four conditions (6 trials per condition; 6 lists across participants): unambiguous; common ground ad hoc implicature; privileged ground ad hoc implicature; and privileged ground semantic (checking perspective-taking with no pragmatic inference).

Fig.	Condition + Utterance	Correct
1A	Unambiguous "Pick the card with apples"	Apples
1A	Common ground ad hoc "Pick the card with bananas"	*Only* bananas
1A	Privileged ground ad hoc "Pick the card with pears"	Pears and bananas
1B	Privileged ground semantic "Pick the card with oranges"	Oranges – in common ground (L)

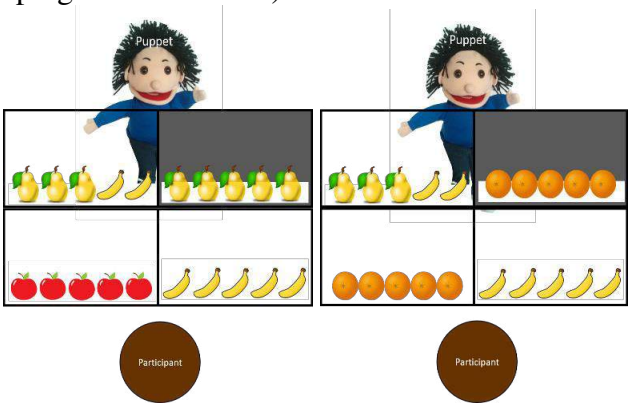


Figure 1A (left) and 1B (right): example experiment displays

In the critical privileged ground ad hoc condition, the card with only Xs was in privileged ground, while the card with Xs and Ys was in common ground. If participants take into account the puppet's epistemic state (not knowing about the card in privileged ground), they would not derive an ad hoc implicature, and instead choose the card with Xs and Ys – for the puppet, 'the card with Xs' is an optimally informative description for the card with Xs and Ys. Participants were asked every 4 trials which cards the puppet could / could not see, and whether he knew what was on them. Adults completed an online version of the task (via Prolific Academic). Children also did a Sally-Anne False Belief task (Wimmer and Perner, 1983).

Results All children passed the Sally-Anne task, except for one who is excluded from the analysis. They also invariably answered correctly the questions about which cards the puppet could see and know about. Adults were at ceiling in all conditions except privileged ad hoc; children were at ceiling only in the unambiguous and common ground ad hoc conditions. As the data was largely bimodally distributed, we coded participants as passers (scoring 5/6 or 6/6) or failers (otherwise)[1]. There were significantly more child passers in the privileged ground semantic than privileged ground ad hoc condition (McNemar's $\chi^2 = 8.5$, $p = .003$; Table 2A), and a significant association of age and performance with more adult than child passers in both privileged ground conditions (Fisher's exact test $p < .001$; Tables 2B, 2C).

A. Child:	Ad hoc Passer	Ad hoc Failer
Semantic Passer	4	10
Semantic Failer	0	19

B. Privileged:	Ad hoc Passer	Ad hoc Failer
Adult	27	9
Child	4	29

C. Privileged:	Semantic Passer	Semantic Failer
Adult	36	0
Child	14	19

Table 2: Chi-squared contingency tables

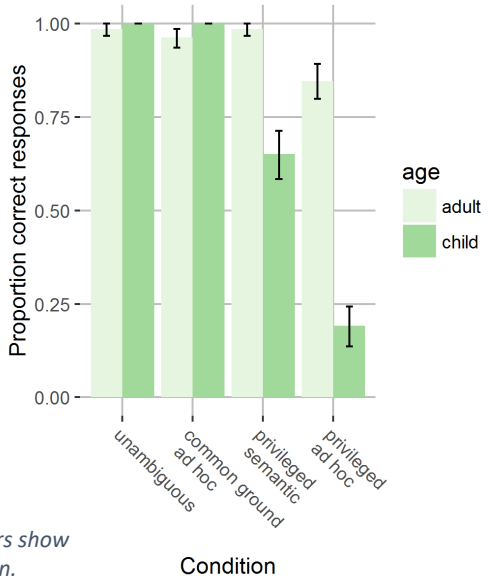


Figure 2: Percentage of correct choice for adults and children. Error bars show bootstrapped 95% confidence intervals for between-subject comparison.

Discussion and conclusion Our results support a two-step development hypothesis. In contrast to adults, children mostly persisted in deriving ad hoc implicatures when the speaker was ignorant of the relevant picture (choosing the one the puppet could not see), despite reasoning correctly about someone's false beliefs or ignorance and exceling in ad hoc implicatures when relevant information is in common ground. Some children also failed to take into account the speaker's perspective when the utterance was semantically ambiguous. Integrating knowledge of the speaker's epistemic state into utterance interpretation therefore seems to be a challenge for children. Our findings support the proposal that children develop the ability to integrate contextual and linguistic information gradually (Papafragou and Skordos, 2016), and raise the question: is integration of full theory of mind always required for some pragmatic competence?

Select References **Horowitz, A., & Frank, M**. (2015). Sources of developmental change in pragmatic inferences about scalar terms. *37th Cog Sci.* **Nilsen, E., & Graham, S.** (2009). The relations between children's communicative perspective-taking and executive functioning. *Cogn Psychol*, *58*(2), 220-249. **Papafragou, A., & Skordos, D.** (2016). Scalar Implicature. In J. Lidz, W. Snyder, & J. Pater (Eds.), *Oxford Handbook of Developmental Linguistics*. Oxford: OUP. **Politzer-Ahles, S., & Fiorentino, R.** (2013). The Realization of Scalar Inferences: Context Sensitivity without Processing Cost. *PLOS ONE*, *8*(5), e63943. **Stiller, A, Goodman, N., & Frank, M**. (2015). Ad-hoc Implicature in Preschool Children. *Lang Learn Dev*, *11*, 176-190.

[1] The maximal mixed effects logistic regression model (Barr, et al., 2013) failed to converge due to ceiling / floor performances and small random effect sizes (*lme4* in R: R Core Team, 2016; Bates, et al., 2015). A model with condition and age as fixed effects (sum coding), by-item (list) random slope, and by-subject random intercept, indicated a main effect of age ($\beta = 1.99$, $p < .01$) – children performed worse than adults – and condition (common ground ad hoc $\beta = 1.93$, $p < .001$; privileged ad hoc $\beta = -3.88$, $p < .001$; privileged semantic $\beta = -1.08$, $p < .001$).

Threshold adaptation and its time course: An investigation of gradable adjectives

Ming Xiang, Chris Kennedy, Allison Kramer (University of Chicago)

It is well established that the interpretation of gradable adjectives is heavily context dependent, since the threshold of determining whether an object qualifies as having a particular adjectival property "X" varies across contexts. One source of variability is the speaker: different speakers may have different thresholds for the same adjective. For successful communication to happen, speakers need to dynamically align thresholds. This study is a first step towards experimentally investigating whether and how hearers adjust their adjective thresholds based on speaker input.

The procedure was adopted and modified from the speech adaptation studies in Vroomen et al. (2007) and Kleinschmidt and Jager (2015). A one-to-five scalar continuum was created for three adjective properties: a relative adjective "tall", a minimum adjective "bent", and a maximum adjective "plain" (see Figure 1A), which represent three different kinds of lexical scales that lead to distinct thresholds (Kennedy 2007). The experiment consisted of a pre-calibration, an exposure/testing and a post-calibration session. In the **pre-calibration session**, each participant was presented with all the images from each adjective scale multiple times, and made a binary judgment to the question "Is this tall/bent/plain?". The presentation of images followed a Gaussian distribution, with the image from scale position 3 presented most frequently for each adjective. For each participant, the most ambiguous scale point image for each adjective was individually determined, and was used for the next session. In the **exposure/testing session**, for each adjective, participants were exposed to a sequence of 24 repetitions of an auditory statement that describes a simultaneously visually presented image. At six different intermediate exposure trial positions (the 2^{th}, 4^{th}, 8^{th}, 13^{th}, 20^{th} and 24^{th}), participants received three test trials. The test trials always consist of the "most ambiguous" image individually chosen for each participant from his/her pre-calibration, and also an additional image from the scale position right above the most ambiguous image or right below it. Participants made a binary judgment as to whether each test image was "tall/bent/plain". Since the goal of the testing trials was to examine whether and how participants adapt and revise their own thresholds under the influence of the exposure trials, in a between-subject design, we manipulated four different kinds of exposure sequences, defined by the combination of the auditory statements and the images they describe (Figure 1B). In the Ambiguous.Positive and the Ambiguous.Negative conditions, the exposure image was the most ambiguous image, and the auditory description was "This is X" or "This is not X". In the Prototypical.Positive condition, the exposure image was from the highest #5 scale position, and the auditory description was "This is X". In the Prototypical.Negative condition, the exposure image was from the lowest #1 scale position, and the auditory description was "This is Not X". For each of the four exposure sequences, before the first auditory statement, the speaker set up a discourse scenario that specified a relevant conversational goal (e.g. "For a party, I need a tall candle" or "For a party, I need a candle that is not tall"). After completing all the exposure/testing trials, participants carried out a **post-calibration session** for each adjective, following the same procedure as the pre-calibration.

Results and discussion Thirty native speakers were recruited for each exposure condition (120 total MTurkers). Figure 2 shows that participants' thresholds were significantly influenced by the exposure sequences, measured by the change from pre-calibration to the post-calibration acceptance judgments. There are two main findings. First, the Ambiguous.Negative exposure made participants less likely to accept an image as having the property X, whereas the Ambiguous.Positive exposure made participants more likely to do so. However, the effect of the Negative and Positive statements with the Prototypical images is the opposite from the Ambiguous image exposure conditions. To explain this, we propose that participants initially maintain probabilistic distributions over the possible threshold values for each gradable adjective, and upon hearing "This is X" with an exposure image, they recalibrate the mean of the initial threshold distribution by shifting the mean towards the direction of the scale position indicated by the exposure trial image. Importantly, in addition to maintaining and adjusting the threshold distribution for a category X, they also simultaneously maintain and adjust the threshold distribution for a category "Not X". Upon observing a negative exposure trial "This is not X", they shift the mean of the "Not X" threshold distribution towards the exposure image as well. Since the two categories "X" and "Not X" are not completely independent from each other, shifting the mean of one distribution simultaneously leads

to the shift of the other mean as well. Figure 3 demonstrates the process of threshold adaptation with three examples. Adaptation that involves shifting the mean of an original category boundary distribution has been found for categorical perception in speech (Kleinschmidt and Jager 2015) and quantifier interpretation (e.g. *some* and *many*, Yildirim et al., 2016). The current study extends this general adaptation strategy to adjective processing. In addition, our results also showed that even exposures to statements that the hearer would have no dispute about can still trigger readjustment behavior for the hearer (e.g. the two Prototypical exposure conditions).

The second finding in Figure 2 is that for absolute adjectives *plain* and *bent*, although the direction of the adaption is similar to the relative adjective *tall*, the adaptation effect is only present under some of the exposure conditions. We suggest that absolute and relative adjectives are sensitive to the same adaption strategy, but the end-point oriented lexical semantics of absolute adjectives leads to a very different initial threshold distribution for the participants, which in turn shapes the output of the adaptation strategy.

Finally, the testing trials during the exposure sequence were designed to examine the time course of the adaptation behavior. Analysis on these trials revealed that the adaptation effect emerged at the earliest point we tested (i.e. after the 2nd trial during the exposure sequence), and the effect size did not change at the later testing trials, suggesting that adaptation happens very early, and is insensitive to the frequency of the exposure trial (c.f. Kleinschmidt and Jager 2015 for categorical speech perception).

Conclusion Listeners actively and quickly revise their thresholds of gradable adjectives by re-estimating the mean of the original threshold distribution based on speaker input. The lexical semantics of different adjectives also modulates the output of the adaptation strategy.

Figure 1: **A**. (**Left)**: Image Stimuli on a 1-5 scale. B. (**Right)**: The four exposure sequences.

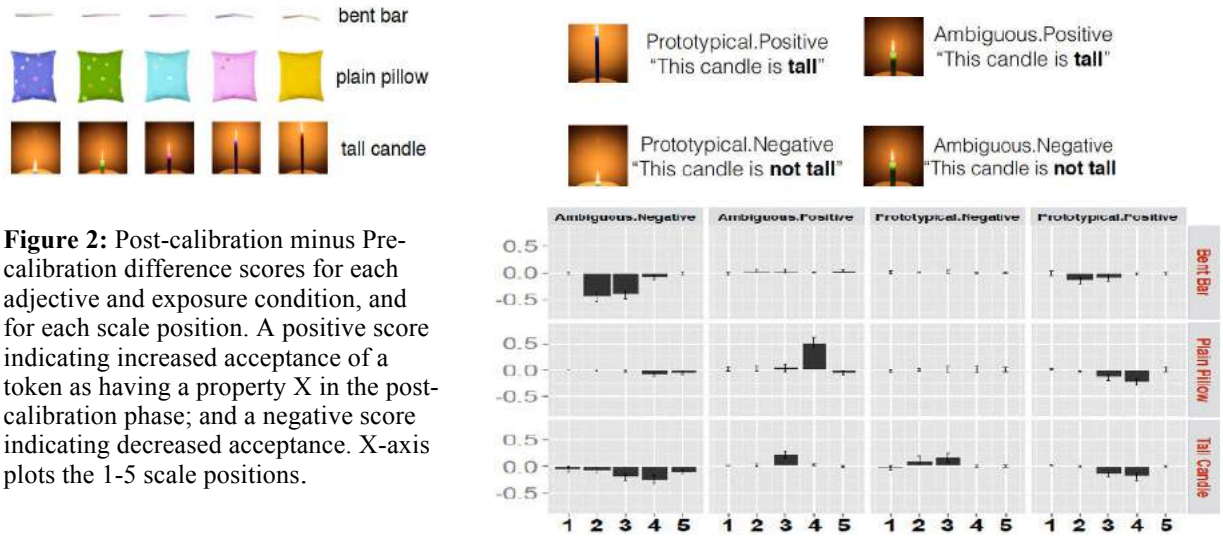
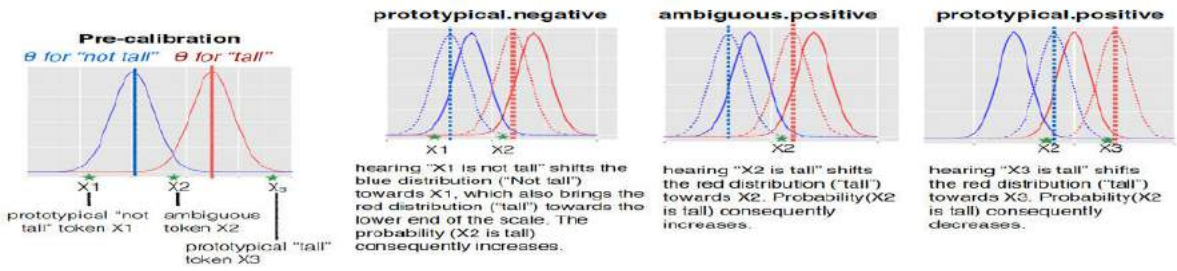


Figure 2: Post-calibration minus Pre-calibration difference scores for each adjective and exposure condition, and for each scale position. A positive score indicating increased acceptance of a token as having a property X in the post-calibration phase; and a negative score indicating decreased acceptance. X-axis plots the 1-5 scale positions.

Figure 3: Schematic adaption behavior for "tall" under three different exposure conditions. The x-axis of each plot represents degrees on a height scale. Solid line distributions represent hearers' original threshold distributions, and dashed line distributions represent the new distributions after adaptation.



Returning to Non-entailed Presuppositions Again

Jérémy Zehr & Florian Schwarz *University of Pennsylvania*

Introduction. A key question in presupposition theory concerns the relationship between presupposed and entailed content. Recent work by Klinedinst (2012) and Sudo (2012) proposes a new perspective on differences between classes of presupposition triggers, with an empirical split roughly mirroring Abusch's hard vs. soft distinction and related notions. They propose that triggers differ in whether or not their presuppositional content simultaneously affect the calculation of the presuppositions and of the entailments of the sentences in which they appear. Sudo shows that conventionally entailed contents can be singled out from exclusively presupposed contents in that only the former scope under non-monotone quantifiers such as *exactly one* (also see Tonhauser et al. (2013)'s notion of obligatory local effects). Recent experimental results by Zehr & Schwarz (2015) are in line with this distinction: the content presupposed by *stop* scopes under *exactly one*, suggesting that it also affects entailment. Crucially, this is not the case for *also*: the presupposed content doesn't scope under the quantifier. Various authors (Zeevat, 1992, Glanzberg, 2005, Domaneschi et al. 2013) advance different proposals sharing a line of thinking, which suggest an *explanation* for this split in triggers' categories. While the verb *stop* cannot be left aside when composing the meaning of the sentence – as it would result in something uninterpretable – leaving aside the additive particle *also* would not affect the good running of the interpretative process. As a result, since *also* exclusively contributes presupposed information, ignoring it results in a viable interpretation where the presupposed content plays no role under the scope of *exactly one*. This strategy is not viable for *stop*, and its content winds up contributing under the scope of the quantifier.

We test this removability/independence (RI) hypothesis experimentally by comparing *return* to *(go) again* (as well as *go back*). These are intuitively equivalent in terms of their overall meaning, but differ precisely with regards to the crucial hypothesized property: ignoring *return* is fatal to the interpretation process, but ignoring *again* or *back* leaves us with perfectly interpretable sentences. In addition, we include *stop* and *also* for additional points of reference and to ensure comparability with Zehr & Schwarz's results. Our results provide clear evidence against the RI hypothesis, as *return* patterns with *again* (as well as *also*), and not with *stop*. At the same time, our data provide further support for Sudo's entailment-contrast proposal.

Design. We employed a picture selection task with a covered box (Huang et al. 2013) using sentences with non-monotonic quantifiers, parallel to Zehr & Schwarz (2015). The design utilized recordings of minimally varied sentences with identical pictures to test whether or not the presupposed information is considered for evaluation of the quantifier *exactly one*.

Exactly one kid { returned / went back / went / also went / stopped going } to the

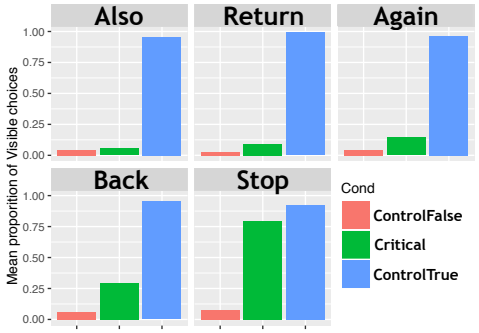
Jackson aquarium {again} on Wednesday. also/again/back/return

stop

Zehr & Schwarz's (2015) results for *also* (with stress on *Wednesday*) and *stop* provide baseline and ceiling for the role of presupposed content respectively: their subjects generally did not consider the presupposed information of *also* (of going to the aquarium before Wednesday – highlighted in green here) in assessing the *exactly one* claim for *also*, instead using only the entailed information (highlighted in red here), and thus rejected the visible picture. In contrast, the presupposed information of *stop* (highlighted in green here) generally WAS considered for counting-purposes and the picture accepted. Based on the RI hypothesis, we expect *again* and *back* to pattern with *also*, and *return* with *stop*. Though *return* conveys the same overall information as the other two, it cannot be felicitously removed (nor can its prefix *re-*) while it contributes, independently from the presupposition, to the entailed content, unlike in the case of *again* and *back*.

Procedure. 150 participants were recruited via Prolific.ac to participate in a 15 minute study for £1.30. Stimuli were presented via Ibex. Participants saw target pictures and 'covered box' variants where relevant details were occluded, and had to decide which of the two matched the sentence they heard. Trigger-type was a between-participant factor, so that 30 participants saw 12 items per condition for each trigger.

Results and Discussion. The predictions of the RI hypothesis were not borne out. Target acceptance rates for *return*, *back* and *again* were overall much closer to *also* than to *stop*. This suggests that the presuppositions of these triggers generally do not figure in the evaluation of the *exactly one* claim. Interestingly, mixed-effect regression models show that *back*, rather than *return*, stands out as factoring the presupposition into the evaluation of the quantifier more often (though still far less so than *stop*). A tentative explanation relates this result to effects of prosody on at-issueness, along the lines of an influential proposal by Simons, Tonhauser & colleagues, where material becomes presupposed precisely if it is not at-issue in terms of addressing the QUD. Conversely, stressing a presuppositional expression increases the chances of it becoming at-issue (also see Tonhauser et al. (2016)). Post-hoc analyses of mean F0-values in our recordings indeed reveal a higher mean for *back* compared to *again*, but crucially the prosody-based explanation doesn't extend to *return*: it exhibited *higher* mean F0-measurements, but *lower* target-acceptance rates, than *back* and even *stop*.



Ultimately, our results are in line with an entailment contrast *à la Sudo/Klinedinst*, but they leave us without an explanation for why triggers belong to the class they belong. We see two potential lines of explanation worth of investigation. The first (suggested by a reviewer) appeals to non-presuppositional alternatives: while *go* naturally appears as a non-presuppositional alternative to *return*, *go again* and *go back*, it is harder to find one for *stop*. The second draws on Abrusán (2016)'s proposal. It hypothesizes that *stop* introduces a timespan of *going* events as inherently connected to a timepoint of reaching a state of not going, making it impossible for our participants to restrict their attention to that timepoint; the other triggers by contrast introduce multiple *going* events mapped to *isolable* timepoints.

Selected References. Abrusán, M. 2016. Presupposition cancellation. NLS **Domaneschi et al.** 2013. The cognitive load of presupposition triggers. L&CP. **Sudo, Y.** 2012. On the semantics of phi features on pronouns. PhD thesis. **Zehr, J. & Schwarz, F.** 2015. Entailed vs. non-entailed presuppositions. NELS 46.