ELM 2 Program - May 18-20, 2022

Main Session Program

Zoom link for ALL Main sessions (virtual and hybrid)

DAY 1 - MAY 18

Panel - Computational Perspectives on Meaning. May 18 9:00-10:45 (virtual)
Chair: Shane Steinert-Threlkeld. **Tech host:** June Choe

Main-1.1. May 18 11:15-12:45 (virtual)
Chair: Vicky Lai. **Tech host:** June Choe

Main-1.2. May 18 16:00-17:30 (virtual)
Chair: Judith Tonhauser. **Tech host:** June Choe

DAY 2 - MAY 19

All in-person Main sessions are held in the Tedori Auditorium in the **Levin Building**.

Main-2.1. May 19 9:00-10:30 (in person + stream on Zoom)
Chair: Kathryn Davidson. **Tech hosts:** Yiran Chen & June Choe

---

Main-2.2. May 19 11:00-12:30 (in person + stream on Zoom)
Chair:. Gillian Ramchand **Tech hosts:** Yiran Chen & June Choe

DAY 3 - MAY 20

All in-person Main sessions are held in the Tedori Auditorium in the **Levin Building**.

Parallel Session Short Talk Program

DAY 1 - MAY 18

Parallel-1.A.1. May 18 13:30-14:30 (virtual) **Zoom Link**
Chair: Rosario Tomasello. **Tech host:** Sarah Lee

Parallel-1.A.2. May 18 13:30-14:30 (virtual) **Zoom Link**
Chair: Oliver Bott. **Tech host:** Ariel Mathis

Parallel-1.A.3. May 18 13:30-14:30 (virtual) **Zoom Link**
Chair: Elsi Kaiser. **Tech host:** Ugurcan Vurgun

Parallel-1.A.4. May 18 13:30-14:30 (virtual) **Zoom Link**
Chair: Andrea Beltrama. **Tech host:** Karen Li

Parallel-1.B.1. May 18 14:45-15:45 (virtual) **Zoom Link**
Chair: Lisa Levinson. **Tech host:** Ariel Mathis

Parallel-1.B.2. May 18 14:45-15:45 (virtual) **Zoom Link**
Chair: Andrea Beltrama. **Tech host:** Ugurcan Vurgun

Parallel-1.B.3. May 18 14:45-15:45 (virtual) **Zoom Link**
Chair: Lyn Tieu. **Tech host:** Sarah Lee

DAY 2 - MAY 19

All in-person Parallel sessions are held in the **Perelman Center for Political Science and Economics**.

Parallel-2.1. May 19 13:30-14:30 (in person: PCPE 100 + stream on Zoom) **Zoom Link**
Chair: Yiran Chen. **Tech host:** Ugurcan Vurgun

Parallel-2.2. May 19 13:30-14:30 (in person: PCPE 101 + stream on Zoom) **Zoom Link**
Chair: Alex Göbel. **Tech host:** Alex Kalomoiros

Parallel-2.3. May 19 13:30-14:30 (in person: PCPE 200 + stream on Zoom) **Zoom Link**
Chair: Alexandre Cremers. **Tech host:** Lefteris Paparounas

Parallel-2.4. May 19 13:30-14:30 (in person: PCPE 202 + stream on Zoom) **Zoom Link**
Chair: Remus Gergel. **Tech host:** Karen Li

DAY 3 - MAY 20

All in-person Parallel sessions are held in the **Perelman Center for Political Science and Economics**.

Can neural networks identify speaker commitment?

Marie-Catherine de Marneffe

When we communicate, we infer a lot beyond the literal meaning of the words we hear or read. In particular, our understanding of an utterance depends on assessing the extent to which speakers are committed to the events they describe. An unadorned declarative like "The cancer has spread" conveys firm speaker commitment of the cancer having spread, whereas "There are some indicators that the cancer has spread" imbues the claim with uncertainty. When I say, "I don't think you should go", you take me to believe that you should not go. In this talk, I will investigate how well BERT, a current neural language model, performs on predicting speaker commitment of embedded events in English. I will show that, although BERT achieves very good results, it does so by exploiting surface patterns that correlate with certain speaker commitment labels in the training data, but it fails on items that necessitate pragmatic knowledge. These results highlight directions for improvement to build robust natural language understanding systems.

Learning Grounded Word Representations

Ellie Pavlick

This talk will discuss the potential of neural network models to learn grounded and structured lexical concepts by modeling the physical world. I will discuss results from two recent sets of experiments. In the first, we train large neural network models on a sequence prediction task--i.e., modeling the future trajectories of objects in motion--and find that many verb concepts (e.g., roll vs. slide, push vs. hit) emerge organically from such training. In the second, we train a neural network on a simple object-naming task and investigate the extent to which the learned conceptual representations exhibit desirable internal compositional structure. Taken together, these projects provide a preview of the possible role of neural networks in both theoretical and empirical lexical semantic research.

Montague Grammar Induction

Aaron Steven White (joint work with Gene Louis Kim)

We propose a computational modeling framework for inducing combinatory categorial grammars from arbitrary behavioral data. This framework provides the analyst fine-grained control over the assumptions that the induced grammar should conform to: (i) what the primitive types are; (ii) how complex types are constructed; (iii) what set of combinators can be used to combine types; and (iv) whether (and to what) the types of some lexical items should be fixed. In a proof-of-concept experiment, we deploy our framework for use in distributional analysis. We focus on the relationship between s(emantic)-selection and c(ategory)-selection, using as input a lexicon-scale acceptability judgment dataset focused on English verbs' syntactic distribution (the MegaAcceptability dataset) and enforcing standard assumptions from the semantics literature on the induced grammar.

The role of relevance, competence and priors for scalar implicatures

Polina Tsvilodub (Osnabrück University, ptsvilodub@uos.de), Bob van Tiel (Radboud University), Michael Franke (University of Tübingen)

If someone says "Anna ate some cookies", the hearer might infer the upper-bounded reading that Anna ate *some, but not all* cookies. Similarly, given "Donald ate a donut or a pretzel.", one might infer that Donald ate either *the donut or the pretzel, but not both* (i.e., an exclusive interpretation). Both inferences are usually explained as a variety of *scalar implicature* (SI). SIs rely on lexical scales consisting of words ordered in terms of informativeness, like ⟨some, all⟩ and ⟨or, and⟩. If a speaker uses an informationally weaker term (e.g., "some"), they may imply that the corresponding stronger alternative (e.g., "all") is false [3,5]. Crucially, prior research suggests that the robustness of SIs is influenced by different contextual factors [1]. We investigate the effects of three factors: (1) the *competence* of the speaker about the truth of the stronger alternative, (2) its *relevance* to the listener, and (3) the *prior probability* that it is true [2,4,6]. We explore how these three factors *interactively* influence the robustness of SIs associated with the triggers "some" and "or".

In our web-based rating study, participants read background stories which were designed to vary in terms of the strength of the three factors (`high` or `low` with respect to: prior probability$\times$competence$\times$ relevance, for each trigger), manipulated within-subjects. On critical trials, participants were asked to rate three sentences on a scale ranging from "certainly true" to "certainly false" ($\widehat{=}$0–100), one per factor. The story ended with one of the characters in the story making an utterance containing "some" or "or". Participants then had to indicate the probability of an SI-enriched paraphrase of that utterance. We thus obtained judgements on the contextual factors, and on the robustness of the SI (see `https://tinyurl.com/3ru9sdja` for experiment details). Based on the literature, we expected higher likelihood ratings for the SI-enriched paraphrase if the alternative was perceived as highly relevant, the speaker was judged as highly competent, and the stronger alternative was viewed as a priori unlikely [2,4,6]. Each participant saw four stories per trigger, sampled from 32 stories/triggers, randomly shuffled with eight structurally similar attention checks and comprehension questions.

We analysed data from 206 participants recruited on Prolific. Their ratings were z-scored within each factor by-participant. We regressed the implicature likelihood ratings against the fixed effects of trigger, all factor ratings within-story, all interactions and maximal random effects, using a Bayesian linear mixed effects model. Participants' factor ratings by-story agreed well with the designed classification of the stories (Fig. 1, red vs. blue color on x-axis). As predicted, participants were more likely to derive the SIs of "some" and "or" when judging speaker competence as high ($P = 0.999$ for "some", $P = 0.993$ for "or" for effect sizes being > 0.05, Fig. 2 for all results). They were also more likely to derive the SI of "some" when judging the prior probability of "all" as low ($P = 1$ for effect < 0.05), which was not the case for "or". Finally, we did not observe credible effects of relevance for either trigger. Given the unexpected absence of prior effects for "or", we computed exploratory pairwise correlations of all predictors. While no correlations were found for "some", we found a significant correlation between the explanatory factors prior and relevance for "or" ($R^2 = -0.106, p < 0.01$). An exploratory model comparison of two models, one containing the relevance effect over one containing the prior effect, each combined with competence, revealed mild evidence in favor of the prior as a better explanatory factor than relevance (Bayes Factor $= 3.70$). While supporting the view that SIs for both triggers rely on epistemic reasoning affected by speaker competence, our results indicate that prior and relevance might be closely connected for "or". Ultimately, our results suggest that the interdependence of the three factors is more complex than just the sum of the effects anticipated in the literature, and provide further insights into how they might enter into people's decision as to whether or not to derive an SI.

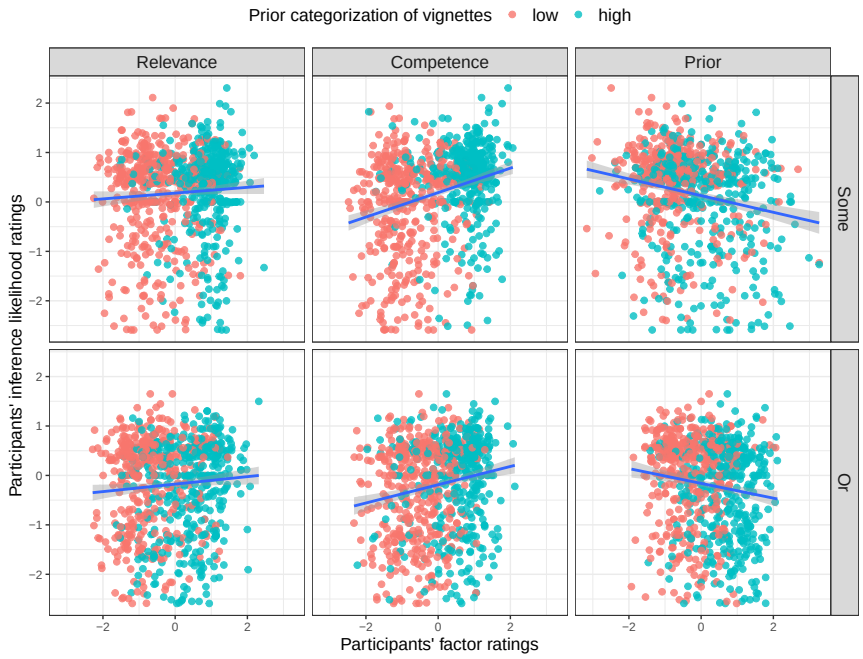


Figure 1: Relating ratings for relevance, competence and prior statements (x-axis) to ratings for the strength of pragmatic enrichments (y-axis). The top row shows ratings for "some" (enriched to "some, but not all"). The bottom row shows ratings for "or" (enriched to "A or B, but not both"). Ratings for stories categorized as low (red) w.r.t. a given factor are on average lower (x-axis) than for those categorized as high (blue). The apparent effect of prior for "or" is mitigated by collinearity with relevance.

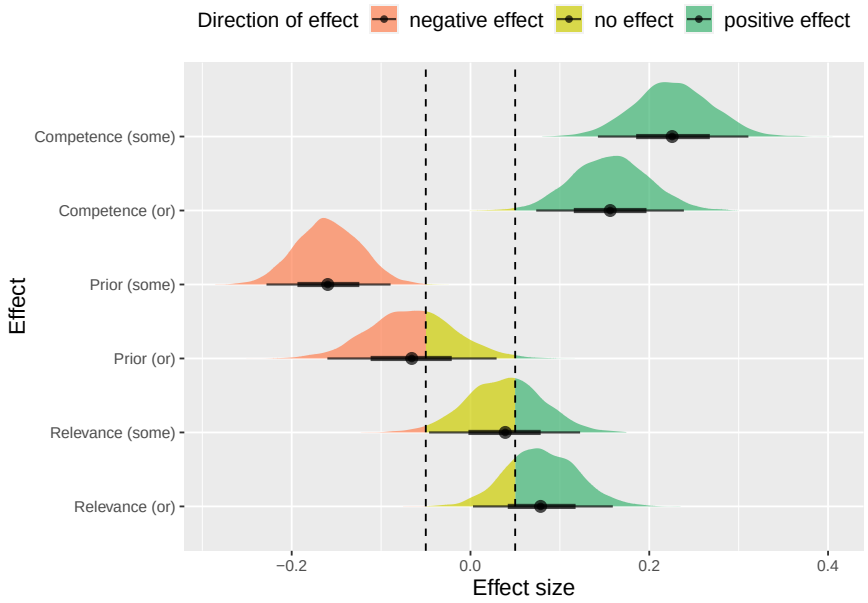


Figure 2: Distributions of posterior samples for the effects (x-axis) of each predictor for each trigger (y-axis). The colors indicate the fraction of the distribution density corresponding to a positive, no or a negative effect. The dashed lines indicate the Region Of Practical Equivalence that we defined as (-0.05, 0.05) in which posterior samples are considered equivalent to 0 (i.e., no credible effect).

References: [1] Degen, J., In *Semantics & Pragmatics*, 2015 [2] Degen, J., Tessler, M. H., & Goodman, N. D., In *Proceedings of CogSci*, 2015 [3] Geurts, B., *Quantity Implicatures*, 2010 [4] Goodman, N. D., & Stuhlmüller, A., In *Topics in Cognitive Science*, 5(1), 2013 [5] Horn, L. R., *On the semantic properties of logical operators in English*, 1972 [6] Sperber, D., & Wilson, D., *Relevance: communication and cognition*, 1995

Accounting for free choice: Revisiting the challenge for the implicature approach

Background: A sentence containing disjunction in the scope of a possibility modal, such as (1-a), gives rise to the FREE CHOICE inference in (1-b). This inference presents a well known puzzle in light of standard treatments of modals and disjunction (Kamp 1974 and much subsequent work). To complicate things further, FREE CHOICE tends to disappear under negation: (2-a) doesn't merely convey the negation of (1-a), but rather the stronger DOUBLE PROHIBITION reading in (2-b). A prominent approach to the FREE CHOICE-DOUBLE PROHIBITION pattern is based on a standard meaning of modals and disjunction and generates FREE CHOICE as an implicature (Fox 2007, Klinedinst 2007, Romoli & Santorio 2018, Bar-Lev 2018, a.o.). This approach successfully captures the basic pattern and a variety of more complex data points related to free choice, but has recently been challenged by experimental data presented in Tieu, Bill, and Romoli (2019) (hereafter 'TBR').

The challenge: To illustrate, consider a context like Fig.1, in which Sue is only allowed to buy the hamburger. In this context, the implicature approach predicts a difference in status across the two polarities: the positive (1-a) is literally true, but with a false implicature, while the negative (2-a) is plainly false. TBR investigated this prediction using a ternary judgment task (Katsos Bishop 2011). Participants were presented with sentences like (1-a) and (2-a) as uttered by a puppet and their task was to reward the puppet with a small, medium, or large strawberry, depending on whether the sentence was completely right, completely wrong, or neither. TBR reported that participants primarily selected the intermediate reward for both the positive (1-a) and negative (2-a) in the given context. In contrast, when presented with simple disjunctive sentences like the positive (3-a) and the negative (3-b) in a context where both disjuncts were true, participants exhibited the asymmetric pattern of responses expected on the implicature approach: a preference for the intermediate reward when the (positive) sentence was logically true but with a false implicature, and the minimal reward when the (negative) sentence was plainly false. TBR took the parallel responses to (1-a) and (2-a), combined with the divergent responses in the equivalent disjunction sentences, to pose a challenge for the implicature approach.

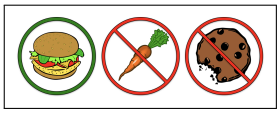


Figure 1: TBR's FC target image, paired with positive and negative FC ((1)-(2)) and CONJUNCTION ((4)) targets in our Exp.1.

Potential confound: TBR's results, however, can also be explained as participants having chosen the intermediate reward in an attempt to be charitable to the puppet. The puppet mentioned two things (the hamburger and the carrot) and she turned out to be right about one of them. So while the sentence on its FC meaning is not compatible with the pictured context, there is a sense in which the puppet's guess was partially right, and this could underlie the reported intermediate responses.

Current study: We report on two experiments that build on TBR's study but which control for the potential confound mentioned above. In Exp.1, we tested TBR's free choice conditions against a corresponding baseline using simple conjunctions. The goal was to test the following prediction of the charitable strategy hypothesis: if the participants selected the intermediate reward in the FC conditions because the puppet was partially right (namely about one of the mentioned items), then given the context in Fig.1, we should observe the same kind of behaviour for simple conjunctions like (4-a) and (4-b). Here everyone agrees, at least for the positive case in (4-a), that the sentence is simply false in the given context. A choice of the intermediate reward, then, would be evidence for the charitable strategy, while a difference between the two conditions would corroborate the challenge for the implicature approach. We furthermore hypothesised that because the disjunctive statements explicitly mentioned two items, they might especially invite the charitable strategy to reward the puppet for getting one of the two things right. So, in Exp.2, we moved to a variant of free choice involving FC 'any', in order to avoid overt disjunctions that would explicitly mention specific items. It was not possible to compare FC 'any' against a 'some' baseline, given the latter's positive polarity properties would make it impossible to test the negative counterpart; hence we decided to compare FC 'any' to the indirect scalar implicature of negated 'every' (*not every but some*). This had two advantages: (i) for both conditions we could compare literally true sentences with false implicatures to literally

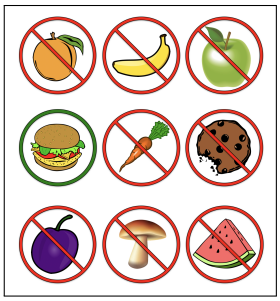


Figure 2: Exp.2 test image paired with positive and negative ANY targets in (5).

false sentences with true implicatures; (ii) the experimental context and visual stimuli could be made comparable (see Fig.2 for an 'any' target; the 'every' targets also featured an array of 9 items, all (LiteralFalse/ImplicatureTrue) or none (LiteralTrue/ImplicatureFalse) of which were circled in green).

Methods: Participants were English native speakers recruited via Amazon Mechanical Turk. We report here on data from 38 participants in Exp.1 (20 FC, 18 CONJUNCTION) and 27 participants in Exp.2 (13 ANY, 14 EVERY), all of whom displayed at least 75% accuracy on unambiguous controls. Participants' task was to decide, given a pictured scenario, whether a puppet's guess had been right ('big strawberry'), wrong ('small strawberry'), or neither ('medium strawberry'). In all, participants saw 8 targets, 8 controls, and 6 fillers.

Results: The results of Exp.1 are presented in Fig.3 (left). Participants' responses to targets qualitatively resembled those in TBR's experiment, with mostly intermediate responses to positive and negative FC targets. Strikingly, the same pattern was observed for the conjunction targets, where no implicature is involved. Cumulative link mixed models with Condition (FC vs. CONJUNCTION), Polarity, and their interaction as fixed effects did not reveal an effect of Condition or interaction ($p > .05$). The results support the suggestion that participants were adopting the aforementioned charitable strategy. Moving to Exp.2 (Fig.3, right), mixed models revealed a significant interaction between Condition and the truth value of the implicature ($\chi^2(1) = 6.3, p < .05$): the status of the implicature had a significant effect on 'every', as expected for a scalar implicature, but not on 'any'.

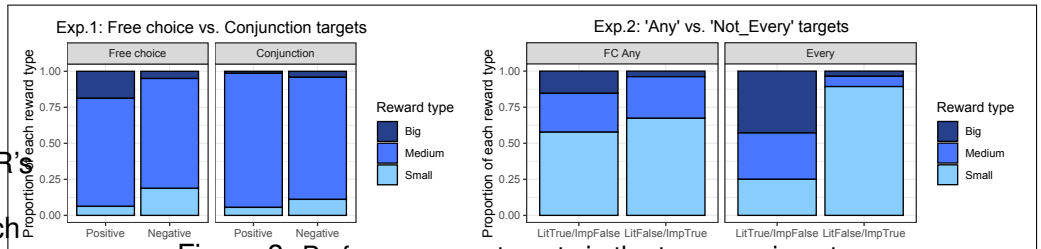


Figure 3: Performance on targets in the two experiments.

Discussion: Our study makes two contributions. First, we investigated a possible confound associated with TBR's challenge to the implicature approach to free choice. We observed a similar pattern of responses for items where no implicature is involved (i.e. (4)), suggesting participants may simply have been responding charitably to the target items (the puppet was right about at least one of the mentioned items). This strategy could have been especially encouraged by the fact that overt disjunction explicitly mentions two items, so we investigated the phenomenon of free choice using 'any', as compared to a(n indirect) scalar implicature baseline. Here we obtain results that converge with TBR's original findings: parallel responses to positive and negative free choice, in contrast to the implicature baseline. The results of Exp.2 clarify the empirical landscape by controlling for the potential confound, and further point us to semantic accounts of free choice like Goldstein (2018), which predicts the observed parallel (undefined) status for positive and negative FC, in contrast to standard cases of implicature.

(1) a. Sue is allowed to buy the hamburger or the carrot.
b. ⇝ *Sue is allowed to buy the hamburger and she is allowed to buy the carrot*

(2) a. Sue is not allowed to buy the hamburger or the carrot.
b. ⇝ *Sue is not allowed to buy the hamburger and she is not allowed to buy the carrot*

(3) a. Sue will buy the hamburger or the carrot.
b. Sue will not buy the hamburger or the carrot.

(4) a. Sue will buy the hamburger and the carrot.
b. Sue will not buy the hamburger and the carrot.

(5) a. Sue is allowed to buy any item. b. Sue is not allowed to buy any item.

(6) a. Sue didn't buy every item. *(in 0/9 vs. 9/9 context)*

Selected References

Bar-Lev, M. 2018. Free choice, homogeneity and innocent inclusion. • Fox, D. 2007. Free choice and the theory of scalar implicatures. • Goldstein, S. 2018. Free choice and homogeneity. • Tieu, L., C. Bill & J. Romoli. 2019. Homogeneity or implicature: An experimental investigation of free choice.

The investigation of quantity implicatures during typical development: a systematic review

Anna Teresa Porrini, Luca Surian
University of Trento

During the last two decades, many experimental studies have concentrated on the investigation of quantity implicatures, and in particular of scalar implicatures, which are generated through the use of lexical items that belong on a scale of informativeness. From a developmental point of view, experimental data suggest that children find these implicatures difficult to process (Guasti et al., 2005; Katsos & Bishop, 2011; Noveck, 2001; Papafragou & Musolino, 2003; Pouscoulous et al., 2007). There is however no straightforward indication of the exact age at which children start deriving this type of implicature, as the available data on the acquisition of scalar implicatures during typical development is varied and sometimes appears contradictory (Eiteljoerge et al. 2018; Sullivan et al., 2019). Furthermore, there are different hypotheses regarding the underlying mechanism of the derivation of these implicatures, and the reasons for the difficulties that children seem to have in deriving them (Foppolo et al. 2012; Katsos & Bishop, 2011; Pouscoulous et al., 2007).

The present work is part of a bigger review project on the acquisition of conversational implicatures in typically developing children, and it is meant to concentrate on quantity implicature in an attempt not only to shed light on what theory amongst the most accredited within experimental pragmatics is most supported by the data, but also on methodological issues regarding the investigation of this type of implicature. The references for this review were selected through the PRISMA method. The criteria for eligibility were that the articles should be peer-reviewed, published articles written in English, they should contain empirical data on the comprehension of quantity implicatures in first language acquisition during typical development, and there needed to be a classification of what type of implicature was being tested and in what way, with examples. Furthermore, the authors needed to have performed a replicable statistical analysis on the data and there needed to be indication of the age range and mean age of the participants. In order to make the data more easily comparable, the last criterion was that the articles should all present their results in term of percentage of success in implicature derivation (or a measure that could be converted to this).

In the end, 39 papers were deemed eligible for this study, all published between years 2001 and 2021. Within these, a total of 141 different findings in terms of percentage of success was obtained, summing up the different experiments, implicature types, tasks and groups tested within these 39 references. The minimum age tested was 2 years old and the maximum age tested was 13 years and 4 months old. Information on how many findings were found for each age group can be found in Table 1.

Mean age in years	Findings per age group
2	2
3	11
4	36
5	43
6	8
7	20
8	3
9	6
10	8
11	4

Table 1

Quantity implicatures taken into consideration could be those derived via the use of a scalar lexical terms or those derived via a contextually given ad-hoc scale. There is therefore a distinction made between scalar implicatures, which count 111 findings, and ad-hoc implicatures, which count 30.

The experiments were run in eight different languages, namely Dutch, English, French, Greek, Italian, Japanese, Mandarin Chinese and Spanish. The results are generalizable beyond the scope of just one language, as there is no detectable difference in percentage of success among the eight languages; in fact, while a MANOVA shows significant effect of language on performance ($F = 2.772$, $p < 0.05$), a subsequent Tukey test reveals that there is no statistically significative difference between any two languages.

Among the 39 papers, six different types of task were used to test scalar implicatures, with four possible types of output variables (binary, ternary, quaternary or performative). A summary of the tasks used and how many findings were collected with each can be seen in Table 2.

Task type	Findings per task type
Action based	10
Communicative context assessment	2
Felicity judgment	30
Referent selection	50
Speaker selection	9
Truth value judgment	40

Table 2

An exploratory analysis of the data done through simple linear regressions suggests that, as expected, performance improves overall with age, which is demonstrated by a positive correlation between the mean age in months and the percentage of success in implicature derivation ($R^2 = 0.078$, $p < 0.001$). This improvement is however more evident for certain tasks than it is for others: in particular, a comparison between the Referent selection task and the Truth value judgment task, which are the two methodologies that count more than 30 findings each, shows that while age does not seem to predict a better performance in the latter case ($R^2 = 0.049$, $p = 0.168$), it does in the former ($R^2 = 0.283$, $p < 0.001$).

The data also suggest that ad-hoc implicatures are easier to derive for children as compared to scalar implicatures, as Fig.1 shows. A t-test confirmed that the difference in percentage of success between the two implicature types is in fact significant ($t = 5.376$, $p < 0.001$).

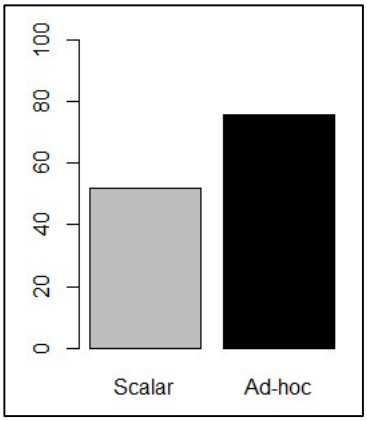


Figure 1: mean percentage of success per implicature type

The data will be analyzed further, through statistical methods, in order to account for interactions among factors. However, it will also be analyzed qualitatively, by grouping the main conclusions drawn by the authors of each paper and the modifications made to the methodologies. Aside from age, task, implicature type and output variable type, other potential predictors of better performance will be taken into consideration in this review, such as presence of pre-training, number of participants and trials, age span of the participants and other linguistic, cognitive and socio-economic factors that were studied within the 39 references.

References

Eiteljoerge, S. F. V., et al. (2018). "Some pieces are missing: Implicature production in children." *Frontiers in Psychology*, 9:1928.

Foppolo, F., et al. (2012). "Scalar Implicatures in Child Language: Give Children a Chance." *Language Learning and Development*, 8(4): 365-394.

Guasti, M. T., et al. (2005). "Why children and adults sometimes (but not always) compute implicatures." *Language and Cognitive Processes*, 20(5): 667-696.

Katsos, N. and D. V. M. Bishop (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.

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

Pouscoulous, N., et al. (2007). "A developmental investigation of processing costs in implicature production." *Language Acquisition*, 14(4): 347-375.

Sullivan, J., et al. (2019). "Differentiating scalar implicature from exclusion inferences in language acquisition." *Journal of Child Language*, 46(4): 733-759.

Explanations over Consequences: Explaining Implicit Causality and Consequentiality Biases
Torgrim Solstad & Oliver Bott (Bielefeld University)

In psycholinguistics, the phenomena of *Implicit Causality* (I-CAUS) and *Consequentiality* (I-CONS) have received much attention for their *coreference* properties, that is, whether the sequence *Peter VERB-ED Mary because/and so...* is biased towards subject- or object-coreferent continuations ([1-11]). For instance, Stimulus-Experiencer (*fascinate*) and Experiencer-Stimulus (*admire*) verbs display strong I-CAUS biases (*because...*) to Stimulus arguments and I-CONS biases (*and so...*) to Experiencers. Consequently, lexical/verb-based accounts ([10-11]) have provided unified accounts of I-CAUS and I-CONS based on shared argument structure. On these *One-Mechanism Accounts*, explanations and consequences specify entities introduced by the verb. More precisely, explanations typically specify the sub-lexical causing eventuality (a property of or event associated with the Stimulus): *Peter admired Mary because she ...*. Conversely, consequences target the caused eventuality (a property of the Experiencer): *Peter admired Mary and so he ...*.

Besides coreference, however, [12] showed that I-CAUS/I-CONS verbs are also *coherence*-biased: Prompts without a connective (*Peter VERB-ED Mary. ...*) lead to the production of explanations over consequences ([12]). Thus, there is a discrepancy between strong coreference biases for both *because* and *and so* versus an overall coherence bias towards explanations. Based on this, we propose the *Two-Mechanism-Account*: Whereas I-CAUS coreference and the coherence bias are both driven by underspecified, explanation-triggering slots in these verbs, I-CONS is governed by the Contiguity Principle ([13]). This principle involves general discourse mechanisms from which we infer a subsequent eventuality, separate from the lexically specififed state of the I-CAUS/I-CONS verb ([13-14]). It comes into play whenever an explicit *and so* overrides the preference to fill underspecified explanatory slots.

Four written production experiments in German each employed 20 Stimulus-Experiencer and 20 Experiencer-Stimulus verbs in different *name*$_1$ *verb-ed name*$_2$ sentence frames (see Materials). These verbs display strong I-CAUS and I-CONS bias and trigger explanations without connectives ([3,10,12,15]). First, **Exp. 1** confirmed the mirror I-CAUS and I-CONS biases from previous research. **Exp. 2 and 3** investigated continuations after full stops (with more uniform verb classes than [12]). **Exp. 2** found a clear preference for explanations over consequences after *Name*$_1$ *verb-ed Name*$_2$*.* prompts. **Exp. 3** expanded this design by enforcing continuations about either Stimulus or Experiencer ([16]). *One Mechanism Accounts* predict continuations about the Experiencer to trigger consequences, as I-CONS is inherently tied to Experiencer arguments on that account. Still, continuations focusing on the Experiencer were mostly explanations. Exp. 2 and 3 thus confirmed *Two Mechanisms*: There is a strong preference for explanations over consequences – even in conditions consistent only with I-CONS. Finally, **Exp. 4** provided more direct evidence for *Two Mechanisms*. Prompts enforced bias-congruent or -incongruent continuations for I-CAUS and I-CONS, respectively: *Peter annoyed Mary because/and so he/she....* We annotated whether continuations specified semantic properties of the Stimulus or Experiencer ([15]). On *One-Mechanism* accounts, bias-congruent I-CAUS prompts should lead to specifications of Stimulus properties and I-CONS prompts should specify the psychological state of the Experiencer. However, participants only provided such specifications in *because* continuations. For I-CONS, end-state specifications were almost never provided ($<$1%). Instead, consequences disjoint from and subsequent to the experiencer end-state were provided, in line with the Contiguity Principle. What is more, only for *because* prompts a difference in continuation strategy between bias-congruent and -incongruent conditions could be observed.

Conclusions: I-CAUS is grounded in verb semantics triggered by semantic underspecification ([11,15]). I-CONS, however, relies on a general Contiguity Principle. I-CONS bias is found

because experiencers are holders of end-point states from which discourse continues in the case of consequences. However, it is only found for explicit marking, which overrides explanatory preferences in the verb. The results have intriguing implications for real-time comprehension ([17]).

Materials and descriptive statistics (GLMER analyses not reported here)

Experiment 1 (N=52)

1. *Stim-Exp, IC, NP1 bias*: 87% NP1
   Peter störte Maria, weil ...
   'Peter annoyed Mary because ...'
2. *Stim-Exp, ICons, NP2 bias*: 95% NP2
   Peter störte Maria, sodass ...
   'Peter annoyed Mary so ...'
3. *Exp-Stim, IC, NP2 bias*: 96% NP2
   Peter bewunderte Maria, weil ...
   'Peter admired Mary because ...'
4. *Exp-Stim, ICons, NP1 bias*: 78% NP1
   Peter bewunderte Maria, sodass ...
   'Peter admired Mary so ...'

Experiment 2 (N=52)

1. *Stim-Exp*: 58% Expl.; 21% Cons.; 6% Contrast
   Peter störte Maria. ...
2. *Exp-Stim*: 60% Expl.; 15% Contr.; 10% Cons.
   Peter bewunderte Maria. ...

Experiment 3 (N=52)

1. *Stim-Exp, Subject focus, IC congruent*:
   84% Expl(anations), 4% Cons(equences)
   [Peter] störte Maria. ...
2. *Stim-Exp, Object focus, ICons congruent*:
   43% Expl., 44% Cons.
   Peter störte [Maria]. ...
3. *Exp-Stim, Subject focus, ICons congruent*:
   49% Expl., 32% Cons.
   [Peter] bewunderte Maria. ...
4. *Exp-Stim, Object focus, IC congruent*:
   77% Expl., 3% Cons.
   Peter bewunderte [Maria]. ...

Experiment 4 (N=56); Proportions of verb-semantically triggered specifications

1. *Stim-Exp, IC bias congruent*: 97%
   Peter störte Maria, weil er ('he') ...
2. *Stim-Exp, IC bias incongruent*: 4%
   Peter störte Maria, weil sie ('she')...
3. *Stim-Exp, ICons bias incongruent*: 1%
   Peter störte Maria, sodass er ('he') ...
4. *Stim-Exp, ICons bias congruent*: 1%
   Peter störte Maria, sodass sie ('she') ...
5. *Exp-Stim, IC bias congruent*: 98%
   Peter bewunderte Maria, weil sie ('she') ...
6. *Exp-Stim, IC bias incongruent*: 9%
   Peter bewunderte Maria, weil er ('he') ...
7. *Exp-Stim, ICons bias congruent*: 0.4%
   Peter bewunderte Maria, sodass er ('he') ...
8. *Exp-Stim, ICons bias incongruent*: 0%
   Peter bewunderte Maria, sodass sie ('she') ...

References

[1] Garvey/Caramazza (1974): Implicit causality in verbs, *LI* • [2] Brown/Fish (1983): The psychological causality implicit in language, *Cognition* • [3] Au (1986): A verb is worth a thousand words, *JML*. • [4] Stewart et al. (1998): Implicit Consequentiality, Ms. • [5] Stewart et al. (2000): The time course of the influence of implicit causality information, *JML* • [6] Stevenson et al. (2000): Interpreting pronouns and connectives, *LaCP* • [7] Koornneef/Van Berkum (2006): On the use of verb-based implicit causality in sent. comprehension, *JML* • [8] Pykkönen/Järvikivi (2010): Activation and persistence of IC information, *EP* • [9] Cozijn et al. (2011): The time course of the use of IC information in the processing of pronouns, *JML* • [10] Crinean/Garnham (2006): Implicit causality, implicit consequentiality and sem. roles, *LaCP* • [11] Hartshorne et al. (2015): The causes and consequences explicit in verbs, *LCaN* • [12] Kehler et al. (2008): Coherence and coreference revisited, *JoS* • [13] Kehler (2004): Discourse coherence, in Horn/Ward (eds.): *Handbook of Pragmatics*. • [14] Pickering/Majid (2006): What are implicit causality and consequentiality?, *LaCP* • [15] Bott/Solstad (2021): Discourse expectations, *Linguistics* • [16] Fukumura/van Gompel (2010): Choosing anaphoric expressions, *JML* • [17] Garnham et al. (2020): Anticipating causes and consequences, *JML*

Learnability and constraints on the semantics of clause-embedding predicates

Mora Maldonado, Jennifer Culbertson and Wataru Uegaki

Summary. Responsive predicates (RPs) are clause-embedding predicates like English *know* and *guess* that can take both declarative and interrogative clausal complements. The meanings of RPs when they take a declarative complement and when they take an interrogative complement are hypothesized to be constrained in systematic ways. Here we investigate whether one such constraint—C(lausal)-distributivity—is reflected in learning. To preview, we find that adults learning a novel clause-embedding predicate in the lab infer the constraint without explicit evidence.

Constraints on RP meanings. Since Karttunen (1977), a major question for the semantics of question-embedding is the relationship between the interpretation of a given RP when it embeds a **declarative** complement (e.g., *Jo knows **that** it is raining*) and when it embeds an **interrogative** complement (e.g., *Jo knows **whether** it is raining*). A number of proposals have been made in the form of constraints on the meanings of RPs. Two examples of such constraints are given below.

(1) **Veridicality constraint:** An RP is veridical w.r.t. declarative complements iff it is veridical w.r.t. interrogative complements (Spector & Égré 2015, i.a.), where V is veridical w.r.t. interrogative complements iff $\ulcorner x\ Vs\ Q \urcorner$ together with $\ulcorner p \urcorner$ entails $\ulcorner x\ Vs\ that\ p \urcorner$.

(2) **C-distributivity:** $\ulcorner x\ Vs\ Q \urcorner \Leftrightarrow$ there is an answer p to Q s.t. $\ulcorner x\ Vs\ p \urcorner$ (Theiler et al. '18).

Compared to the rich theoretical literature on these constraints (e.g., Spector & Égré; Theiler et al. 2018), relatively few attempts have been made to assess the validity of these constraints from empirical grounds. Notably, Sterinert-Threlkeld (2019) tested (1) in learnability experiments using neural nets, and Roelofsen & Uegaki (2021) surveyed the cross-linguistic validity of several constraints including (1) and (2). Nevertheless, it remains unclear whether human learners are sensitive to these kinds of constraints. In this study, we tested the hypothesis that RPs satisfy (2). From this hypothesis, we derive a novel learning-based prediction: when learning a new RP, learners will infer that it is C-distributive. We tested this hypothesis for two different RPs: 'falsely believe' (FALSEBEL) and 'has a correct belief that p is false' (KNOWFALSE). The former would be C-distributive if *Jo* FALSEBEL *whether it's raining* is true only in situations where Jo believes a false answer to the question of whether it's raining. The latter, if *Jo* KNOWFALSE *whether it's raining* is true only in situations where Jo believes a true answer.

Experimental design. Participants are randomly assigned to one of two possible conditions. In both conditions, they learn a new verb *lem*, which can be combined with declarative and interrogative complements. Conditions differ on whether *lem* means KNOWFALSE or FALSEBEL. Participants are first trained on how to use the predicate *lem* with declarative complements, in sentences of the form *Jo lems that p*, where *p* is one of [*it's raining outside*, *it's sunny outside*, *it's snowing outside*]. The training consists of: (a) Exposure phase: participants are shown the situations where they can use a sentence of the form *Jo lems that p* (positive evidence only; Fig.1A); and (b) Acceptability phase: Participants are shown different situations and asked to decide whether a sentence of the form *Jo lems that p* could be used to describe them (Fig.1B). The situations illustrate where *lem* can be used and where it cannot be used. Participants are given feedback on their answers, so they get both positive and negative evidence. For example, in the FALSEBEL condition, participants are shown that they cannot use *lem* in a situation where Jo has a true belief about the weather. Participants are then tested on their interpretation of sentences of the form *Jo lems Q*, where *Q* is *what the weather is like* (Fig.1C). Participants are asked whether the sentence *Jo lems Q* can be used in the following three situations: (i) When Jo believes a true answer to *Q* (True answer); (ii) When Jo believes a false answer to *Q* (False answer); (iii) When Jo has no belief

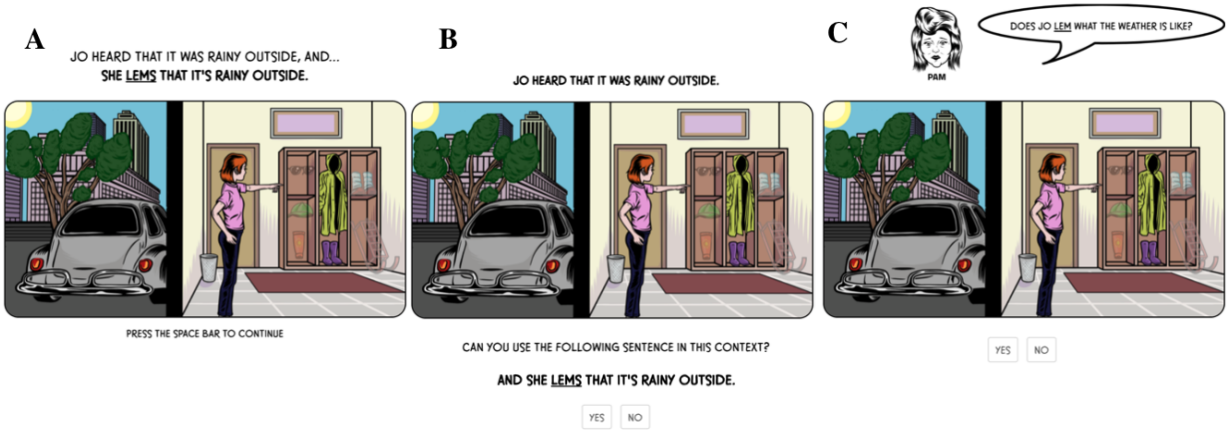


Figure 1: Example trials for Exposure (**A**), Acceptability (**B**) and Testing (**C**) in the FALSEBEL condition.

(No answer). No feedback was given in this part. Learners who infer that *lem* is C-distributive in the FALSEBEL/KNOWFALSE condition are expected to accept the sentence *Jo lems Q* in False/True answer situations, and reject it otherwise (in No answer and True/False answer situations).[1]

Results. 61 English-speaking participants were recruited on Prolific and successfully trained on the use of *lem* with declarative complements (FALSEBEL=40; KNOWFALSE=21). Fig.2 shows the proportion of responses compatible with C-distributivity during testing for each condition, across situations (True, False and No answers). A logit mixed-effects model, including random intercepts per subject (nested by condition) and situation, revealed that the proportion of trials in which *lem* is treated as satisfying C-distributivity is significantly above chance (β =3.73; p =.0024).

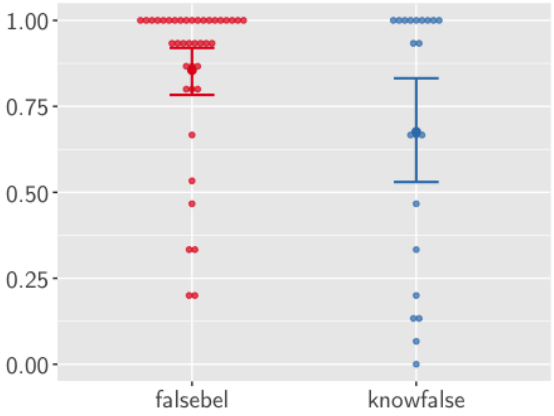


Figure 2: Responses compatible with C-distributivity.

Discussion. Our results suggest that the learning-based prediction derived from the hypothesis that RPs must satisfy (2) is borne out for the novel RPs in our experiment. Note, however, that this finding is mainly driven by the FALSEBEL condition, as the sample in the KNOWFALSE group is too small to confidently infer a pattern.[2] Notably, our results cannot be explained by (1) because (1) doesn't make any prediction about the participants' choices in Testing (both predicates are non-veridical w.r.t. interrogative complements regardless of participants' choices). While our results concerning KNOWFALSE are still tentative, they align Roelofsen & Uegaki ('21) who observe that RPs tend to obey (a version of) (2) cross-linguistically. Importantly, the results also suggest that this constraint might drive inferences during natural language acquisition, thus providing a mechanism for explaining this cross-linguistic tendency. // **References.** Karttunen. 77. Syntax and semantics of questions • Roelofsen & Uegaki. 21. Searching for a universal constraint on ... • Spector & Égré. 15. A uniform semantics for embedded interrogatives • Steinert-Threlkeld. 19. An Explanation of the Veridical Uniformity Universal • Theiler, Roelofsen, & Aloni. 18. A uniform semantics for declarative and interrogative complements.

[1] This experiment, including predictions, design, and analysis was preregistered here.

[2] KNOWFALSE turned out to be very difficult to learn w.r.t. declarative complements to begin with, and for this reason we have not been able to collect our target sample size.

Beyond Surprising: English Event Structure in the Maze

Lisa Levinson, University of Michigan

Are there event structure properties of the lexical representations of verbs that influence reading times above and beyond the probabilistic distribution of those verbs and their arguments?
Background: Previous behavioral studies have found "costs" for lexical semantic verb representations due to the number of sub-events [1]–[4] and event types[4], even in lexical decision where contextual prediction does not play a role. It remains unclear, however, the extent to which this semantic complexity affects sentence processing. Structural verb biases vary both within and across languages *independent of* the event structure of the verbs themselves[2], [5]. Event structural properties thus might not travel through the same "causal bottleneck" [6] of surprisal, but rather make an independent contribution to processing. Prior findings cannot tease apart these factors; while based on stimuli that are controlled for a variety of probabilistic factors, they have not been recently re-evaluated in the context of (a) probabilities from less sparse language models, (b) measures more closely correlated with reading times[7], [8], (c) statistical modeling of multiple stimulus properties, and (d) more focal behavioral tasks such as grammatical maze[9], [10].
Experiment 1 sought to replicate effects of crossing event complexity and transitivity in English (exp 2 of [11]), with added analyses to evaluate the relative contribution of event complexity vs. surprisal. Stimuli (1)-(4) cross verb type (change-of-state (COS) vs. activity) with number of arguments. Transitives in the COS alternation (causatives) are assumed to have more complex events than intransitives (inchoatives)[12]–[14], as well as both activity variants, predicting a pairwise effect in COS verbs and an interaction independent of transitivity itself. Wh-question frames ensured that direct objects would be apparent prior to verb presentation.
Methods: 90 American English speakers completed a self-paced moving window task (with acceptability judgments) presented online via IbexFarm[15].
Results: LME models [16], [17] comparable to those in [11] supported replication of the predicted interaction (β = .04, se = .015, p < .05) and pairwise effect in COS verbs (β = .03, se = .01, p < .01) at verb+1 (Fig 1). LMEs were then fit with additional fixed effects of syntactic surprisals based on verb transitivity probabilities in VALEX[18] and full context lexical surprisals from pre-trained GPT-2[19]. While model comparison showed GPT2 significantly improved model fit (LRT p < .01), neither syntactic surprisal nor event structure did. This may be due to the dispersed and small effects observed via SPRT.
Exp 2: 60 participants completed a grammatical maze task on IbexFarm with the same stimuli, implemented with A-maze [20] using GRNN[21].
Results: As predicted, maze exhibited more focal effects, with no apparent spillover (Fig 2). Even with the full model, the predicted pairwise effect (β = .13 (255ms), se = .03, p < .0001) and interaction (β = .16 (280ms), se = .03, p < .001) were significant at the verb. GPT2 (but not syntactic surprisal) significantly improved model fit (LRT p < .0001), but the event structure interaction also significantly improved fit over GPT2 alone (LRT p < .0001).
In conclusion, these results support an independent contribution of event structure complexity to incremental processing above and beyond surprisal in the slower but more incremental maze task. Comparison of methods suggests that such effects may only be separable with more focal and larger effects that allow for teasing apart multiple fine-grained contributions to sentence processing.

Stimuli (matched across conditions for acceptability and verb frame entropy)

	Sentence		Verb Type	Arguments	Subevents
(1)	What did the explosion	sink near the harbor?	COS	2	2
(2)	When did the boat	sink near the harbor?	COS	1	1
(3)	What did the professor	read for the seminar?	Activity	2	1
(4)	When did the professor	read for the seminar?	Activity	1	1

Exp 1 & 2 full models: log RTs with fixed effects of verb type:num args, scaled and centered [verb frequency, length, syntactic surprisal, GPT2 surprisal], random effects of subjects and items.

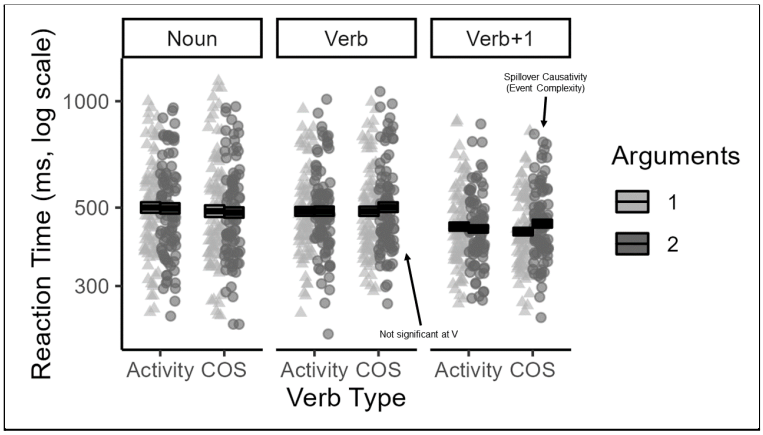


Figure 1: Experiment 1, SPRT. Interaction and pairwise effect emerge at Verb+1 spillover (preposition).

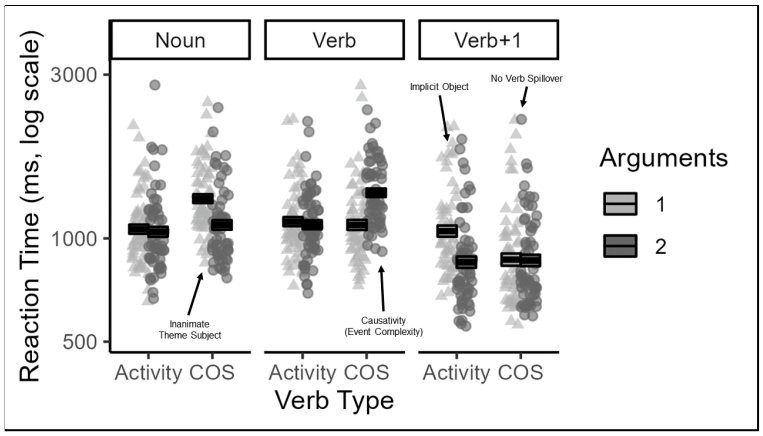


Figure 2: Experiment 2, Maze. Interaction and pairwise effect of event structure at verb. Effect at prior noun likely due to inanimate theme subjects of inchoative verbs, does not spillover to verb. Effect at verb+1 (preposition in all conditions) likely due to implicit object with intransitive activity verbs, also found in [11] and experiment 1 verb+2. No spillover of verb effect to verb+1.

[1] G. McKoon and J. Love, Language and Cognition, 2011. [2] G. McKoon and T. MacFarland, Language, 2000. [3] G. McKoon and T. Macfarland, Cognitive Psychology, 2002. [4] S. Gennari and D. Poeppel, Cognition, 2003. [5] M. Rappaport Hovav, in Perspectives on Causation, 2020. [6] R. Levy, Cognition, 2008. [7] J. Hale, presented at the NAACL, 2001. [8] J. Hale, Language and Linguistics Compass, 2016. [9] K. I. Forster, C. Guerrera, and L. Elliot, Behavior Research Methods, 2009. [10] N. Witzel, J. Witzel, and K. Forster, J Psycholinguist Res, 2012. [11] L. Levinson and J. Brennan, in Morphological Metatheory, 2016. [12] L. Pylkkänen, Introducing Arguments. 2008. [13] A. Alexiadou, E. Anagnostopoulou, and F. Schäfer, in Phases of Interpretation, 2006. [14] M. Rappaport Hovav and B. Levin, in The Theta System, 2012. [15] A. Drummond, Ibex farm. Online server, 2013. [16] D. Bates and M. Maechler, 2009. [17] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, Journal of Statistical Software, 2017. [18] A. Korhonen, Y. Krymolowski, and T. Briscoe, presented at the LREC, 2006. [19] A. Radford et al., OpenAI blog, 2019. [20] V. Boyce, R. Futrell, and R. P. Levy, Journal of Memory and Language, 2020. [21] K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni, arXiv:1803.11138 [cs], 2018.

Geometry and function in spatial terms: Core and more

Barbara Landau

Theories of the meanings of spatial terms often focus on geometric properties as the key to understanding meaning. For example, “The cat is on the mat” might engage geometric properties characterizing the figure (cat, a ‘point') and the ground (mat, a ’surface') as well as the geometric relationship between the two objects (‘on’, coincidence). However, other theorists suggest that geometric properties are far from sufficient to capture the meanings of many spatial terms, and that instead, functional, force-dynamic properties of objects (e.g. support, containment) are crucial to spatial term meanings. In this talk, I argue that both approaches are necessary to understanding the variety of spatial terms that appear in language. To do this, I introduce two new divisions of labor within English spatial prepositions. The first is a division between ‘geometric’ spatial terms in English (including above/below, left/right, north/south/east/west), and ‘functional’ or ‘force-dynamic’ terms (including in, on , and others), with each set of terms drawing on quite different kinds of properties. The second division of labor is within the set of functional/force-dynamic terms; here, the ‘core’ exemplars of a category are encoded with the simplest expressions (e.g. is in/ is on), while ’non-core’ exemplars are encoded through use of a rich set of lexical verbs that help specify the particular kind of force-dynamic properties engaged. The division between geometric and functional / force-dynamic terms has many consequences, including the ease of acquisition of each type in first or second language acquisition, the extent and kind of cross-linguistic variation for each type, and possibly the neural substrate underlying the two types.

Does causality matter?
Impressions of agency influence judgments of both causal and non-casual sentences

Sehrang Joo [a], Sami R. Yousif [a], Fabienne Martin [c], Frank C. Keil [a], and Joshua Knobe [b]

[a] *Department of Psychology, Yale University*
[b] *Program in Cognitive Science and Department of Philosophy, Yale University*
[c] *ERC LeibnizDream Project, Humboldt-Universität zu Berlin*

Imagine a train platform with a line that people aren't supposed to cross—if they do, incoming trains will automatically stop. Suppose that Tom deliberately steps over the line to stand in front of it, and this ends up causing a train delay. In this case, it seems natural to say:

(1) Tom caused the train delay.

Existing research shows that people's willingness to apply this sentence depends in part on the degree to which Tom is exercising agency. Thus, suppose that, instead of acting intentionally, Tom blacks out and falls over the line. Just as in the first scenario, Tom is now too near the edge of the platform, and this leads to a delay. In this case, however, (1) seems like much less natural way to describe what has happened. Indeed, existing research shows that people's endorsement of sentences like (1) are often affected by whether an agent acted intentionally (see e.g., Kirfel & Lagnado, 2021; Lombrozo, 2010; Rose, 2017; Schwenkler & Sytsma, 2020).

This work typically understands these effects as demonstrating something about *causal* cognition in particular. In other words, existing research has focused especially on judgments about causation and on how impressions of agency might impact those judgments.

Consider, however, the following sentence:

(2) Tom crossed over the line.

In (2), there is no longer any information about causation; the path verb *cross* is typically analyzed as devoid of causative semantics. Yet, strikingly, we find it in the experiments described below that people's evaluations of (2) are affected by intentionality in precisely the same way that their evaluations of (1) are. This result suggests that these effects of intentionality are not about how people reason about causation in particular, but instead show that perceptions of agency impact the way people think about a far broader class of sentences.

This raises a question about what gives rise to the effect of intentionality found in sentences like (1) and (2). One possibility is that these effects are not located in how people reason about the verb in the sentence (i.e., *cause* or *cross*), but instead in how they reason about the subject (i.e., *Tom*). To explore this hypothesis, we can look at cases in which the subject is inanimate:

(3) a. The water caused the train delay.
b. The water crossed over the line.

If these sentences require intentionality in order to be acceptable, then people should also be hesitant to accept (3a-b), since the water is not acting (and cannot act) intentionally. In contrast, if the effect of intentionality has something to do with animate agents in particular, then (3) may be acceptable, since the water is not an animate in the first place.

In our experiments, we find that people endorse (3), to the same extent that they endorse (1) and (2) when Tom acts intentionally. These results suggest that intentionality affects the evaluation only of sentences that are about animate agents (and does so whether or not those sentences involve explicit causation).

Experiment 1

Four hundred adult participants were shown one of four short vignettes about a person, Tom, acting with full agency or with a very low degree of agency. For example, in one vignette, participants were told that Tom is waiting for a train and that there is a yellow line on the platform that people aren't supposed to cross. In the full agency condition, Tom deliberately crosses over the line, causing an adverse outcome. In the reduced agency condition, Tom passes out and falls over the line, causing the same outcome. Participants were then asked to evaluate *either* a causal statement (e.g., "Tom caused the train delay.") *or* a statement with one of the four non-causative verbs *hit, touch, enter* and *cross* (e.g., "Tom crossed the line.") on the basis of whether this sentence was a "natural/valid way of describing the event."

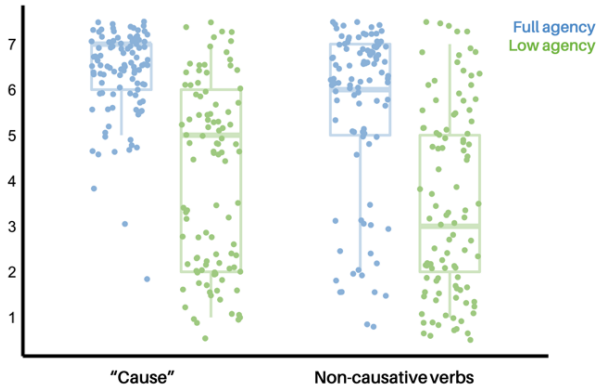


Figure 1. Results from Experiment 1.

Results are displayed in Figure 1. We found no significant interaction between degree of agency and statement type. There was, however, a significant effect of degree of agency within each statement type (*p*s<.001). This means that whether or not Tom acted with full agency affected participants' evaluations of both causal and non-causal statements.

Experiment 2

Six hundred adult participants were again shown one of four short vignettes. Now, however, participants were split into three agency conditions: (1) Tom acting with a very high degree of agency (e.g., Tom, in full control of his actions, deliberately stepping over the line); (2) Tom acting with very reduced agency (e.g., Tom blacking out and falling over the line); and (3) an inanimate object (e.g., a heavy rainstorm floods the train platform, and the weight of the water over the line triggers the same outcome). Participants were again asked to evaluate *either* a causal statement (e.g., "Tom caused the train delay" or "The water caused the train delay") *or* a statement with a non-causative verb (e.g., "Tom crossed over the line" or "The water crossed over the line") on the basis of whether this sentence was a "natural/valid way of describing the event."

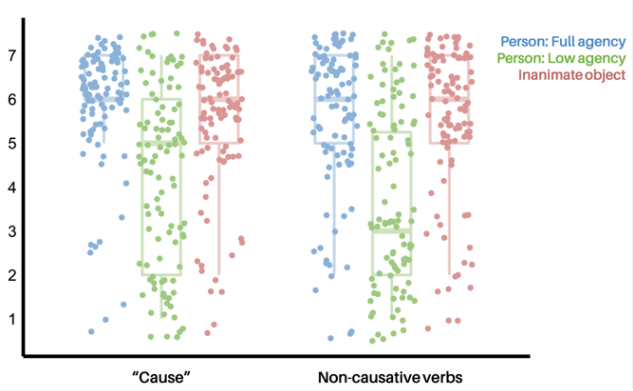


Figure 2. Results from Experiment 2.

Results are displayed in Figure 2. We again found no significant interaction between *degree of agency* and *statement type*—replicating the effect of degree of agency across sentences with both causative and non-causative verbs. Furthermore, degree of agency affected participants evaluations of sentences about Tom, such that sentences describing Tom's actions were rated as more natural/valid when Tom acted intentionally than when he did not (*p*<.001)—but did *not* affect their evaluation of sentences about inanimate objects; participants thought a sentence like "The water crossed over the line" was an acceptable description of the scenario (even though the water obviously had a very low or null degree of agency; *p*=.30).

Conclusion

The effect of intentionality on people's evaluations of sentences like (1) are well-documented. We find, however, that these effects do not arise from something about causal cognition in particular. Instead, they may result from some more general role that agency plays in language. Thus to best understand how people are reasoning about intentional action in these cases, future research should focus not on developing theories that are specific to causal cognition in particular—but instead on developing theories designed to capture more general effects involving the role of agency in language.

A theoretically motivated quantitative model for the interaction between vagueness and implicatures

Alexandre Cremers – Vilnius University

Leffel et al. [4] observed a puzzling contrast between the implicatures of relative and minimum standard adjectives, which they attribute to the fact that the former, but not the latter, are vague:

(1) John is not very tall. ↛ John is tall

(2) The antenna is not very bent ⇝ The antenna is (somewhat) bent

(2) gives rise to the expected structural implicature, by competition with the simpler and more informative alternative *not bent*, but this implicature is absent in (1), unless *very* is stressed. [4] remark that no height can both clearly satisfy *tall* and clearly falsify *very tall*, making the candidate strengthened meaning of (1) akin to *borderline contradictions* such as "tall and not tall" ([6]). By contrast, since *bent* can be interpreted strictly, one can choose a degree arbitrarily close to 0 in order to fully satisfy both *bent* and *not very bent*. [4] propose to generalize [2]'s notion of *innocent exclusion* so that the EXH operator block such borderline contradictions. Doing so captures the initial observation, but we argue that implicatures' sensitivity to vagueness is unlikely to be encoded semantically. Instead, we propose a pragmatic model which makes explicit the intuition of [4], but derives the contrast using the standard definition of EXH: (1) does not give rise to an implicature because *tall but not very tall* is only compatible with a very narrow range of heights, and if the speaker and listener assign slightly different thresholds to *tall*, the heights they consider "tall but not very tall" may not overlap and communication may fail.

Model description: We factor the speaker's uncertainty about the interpretation by implementing higher-order vagueness in the model: not only is there uncertainty about θ, but the distribution of θ is itself uncertain. We adopt [7]'s implementation of supervaluationism in RSA: the utility of a message is its average utility across all possible threshold distributions. Since utility diverges to $-\infty$ as the probability of the message being true approaches 0, a message must be true under all possible interpretations to be usable. In line with the grammatical view of implicatures ([1]) and recent work in the RSA framework ([3]), implicature derivation is treated as a disambiguation problem between parses with and without EXH. We adapt [3]'s Global Intentions model, which differs from the supervaluationist treatment of underspecification: the speaker chooses the pair (message, parse) which best conveys their intention. In particular, this decision rule does not prevent the speaker from using a message u when one of its interpretation is false or likely false (e.g., *not very tall*). Piecing everything together, the model captures the observation in (1) as follows: upon hearing *not very tall*, the pragmatic listener knows that—in principle—the speaker could mean either the exhaustive or literal interpretation. However, no matter which height the speaker had in mind, the exhaustive interpretation has a very low expected utility (across all possible vague denotations for *tall* and *very tall*): in supervaluationist terms, no height makes EXH*[not very tall]* supertrue. By contrast, the literal interpretation is compatible with low heights under any reasonable threshold for *very tall*. The listener therefore concludes that the speaker likely meant the literal interpretation, and that John is somewhat short. Concretely, we assign the following truth-conditions to vague messages, where θ and $\theta{+}\delta$ are the thresholds for POS *adj* and *very adj* respectively, h the degree to convey, and Θ a set of parameters describing the distribution of θ and δ:

$$[\![\text{POS adj}]\!]^{h,\Theta} = P(\theta < h|\Theta); \qquad [\![\text{not POS adj}]\!]^{h,\Theta} = P(\theta \geq h|\Theta); \qquad [\![\text{very adj}]\!]^{h,\Theta} = P(\theta + \delta < h|\Theta)$$

$$[\![\text{not very adj}]\!]^{h,\Theta} = P(\theta + \delta \geq h|\Theta); \qquad [\![\text{EXH not very adj}]\!]^{h,\Theta} = P(\theta < h \leq \theta + \delta|\Theta)$$

We follow [5] in assuming an additional ambiguity between POS and MIN for *late*, but for reasons of space we skip details regarding this aspect of the model (it doesn't play a crucial role in predictions). Our L_0 listener is parametrized by Θ and a parse i. The speaker S_1 selects the pair (u, i) such that u under parse i maximizes expected utility (across all parameter sets Θ). L_1 jointly infers h and i using Bayes' rule, with uniform prior on all parses compatible with u.

$$L_0(h|u,i,\Theta) \propto P(h)[\![u]\!]^{h,i,\Theta} \qquad U_1(u,i|h) = \int \log L_0(h|u,i,\Theta)P(\Theta)d\Theta - c(u)$$

$$S1(u,i|h) \propto \exp(\lambda U_1(u,i|h)) \qquad L_1(h,i|u) \propto P(h)S_1(u,i|h)$$

Implementation and Evaluation: We tested our model on [4]'s Exp 1, which compared relative *tall* and minimum standard *late*. Because we are not interested in explaining vagueness *per se*, only its interaction with implicatures, we fitted a hierarchical Stan model on data from the affirmative constructions *adj* and *very adj* to obtain the distribution of Θ_{tall} and Θ_{late} empirically. As a first approximation, we treat within-participant

fuzziness as indicative of first-order vagueness, and between-participants variance as second-order vagueness: we assume that each participant instantiates a single Θ, and the population variance reflects the distribution of Θ. From the fitted hyperparameters of the distribution of Θ, we computed L_1's posterior on EXH as a function of $(\lambda, c_{\text{adj}}, c_{\text{not}}, c_{\text{very}})$, and fitted participants' responses to *not adj* and *not very adj*, assuming that the acceptability of a message u in this experiment is its expected truth given Θ and a pragmatically inferred probability $P(\text{EXH})$. The Θ fitted for each participant from their responses to *adj* and *very adj* was fed to a new hierarchical model with parameters $(\lambda, c_{\text{adj}}, c_{\text{not}}, c_{\text{very}})$, predicting behavior on *not very adj*. Fig. 1 shows that the model correctly predicts participants' behavior with median by-participant parameters (λ=1.5, c_{adj}=2.0, c_{not}=2.6, c_{very}=2.1). The posterior probability of the exhaustive interpretation is lower with *tall* than with *late* (CIs [.14,.19] vs. [.36,.39]). Crucially, Fig. 2 shows that $P(\text{EXH}|\textit{not very late})$ usually increases with rationality, while $P(\text{EXH}|\textit{not very tall})$ always falls to 0.

Discussion: By combining results and intuitions from the theoretical literature with recent advances in RSA models, we were able to capture the whole range of behaviors in the experimental data. Qualitatively, the model correctly predicts that *not very tall* does not convey "tall but not very tall", while this interpretation can be very salient for *not very late*. We can show that the decision to use supervaluationism for vagueness and Global Intentions for implicatures is crucial: treating vagueness and implicatures uniformly under a single disambiguation rule fails to capture the contrast between *tall* and *late*.

Acknowledgment: This research was funded by the European Social Fund under the measure 09.3.3-LMT-K-712 "Development of Competences of Scientists, other Researchers and Students through Practical Research Activities".

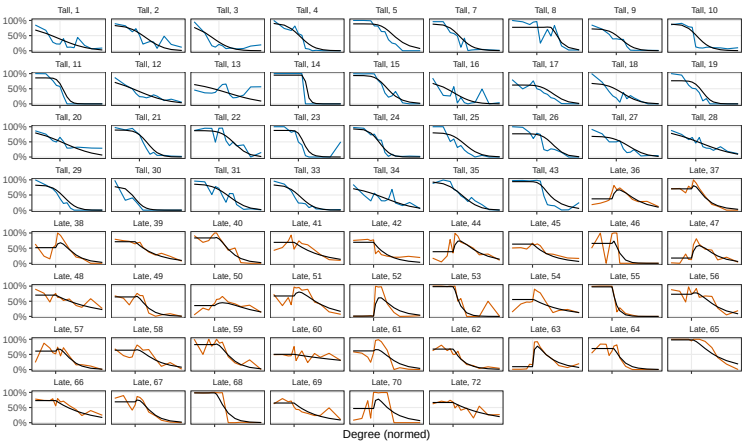


Fig. 1: Individual participants' acceptability of *not very adj* (colored line) and model fit (black line). The implicature translates as reduced acceptability for low degrees.

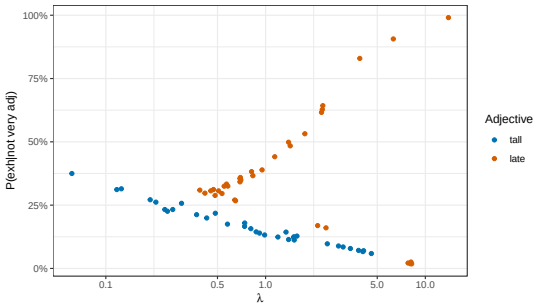


Fig. 2: $P(\text{EXH}|\textit{not very adj})$ as a function of participants' fitted rationality (log-scale).

References

[1] G. Chierchia, D. Fox, and B. Spector. Scalar implicature as a grammatical phenomenon. In *Semantics: An International Handbook of Natural Language Meaning*, volume 3, pages 2297–2331. Mouton de Gruyter, Berlin, 2012.

[2] D. Fox. Free choice disjunction and the theory of scalar implicature. In U. Sauerland and P. Stateva, editors, *Presupposition and implicature in compositional semantics*, pages 71–120. Palgrave Macmillan, New York, NY, 2007.

[3] M. Franke and L. Bergen. Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings. *Language*, 96(2):77–96, 2020.

[4] T. Leffel, A. Cremers, N. Gotzner, and J. Romoli. Vagueness in Implicature: The Case of Modified Adjectives. *Journal of Semantics*, 36(2):317–348, 2019.

[5] C. Qing. Zero or minimum degree? Rethinking minimum gradable adjectives. *Proceedings of Sinn und Bedeutung 25*, 2021.

[6] D. Ripley. Contradictions at the borders. In R. Nouwen, R. van Rooij, U. Sauerland, and H.-C. Schmitz, editors, *Vagueness in communication*, pages 169–188. Springer, 2011.

[7] B. Spector. The pragmatics of plural predication: Homogeneity and non-maximality within the rational speech act model. In A. Cremers, T. van Gessel, and F. Roelofsen, editors, *Proceedings of the 21st Amsterdam Colloquium*, page 435, 2017.

On a concessive reading of the rise-fall-rise contour: contextual and semantic factors

Alexander Göbel & Michael Wagner - *McGill University* (alexander.gobel@mcgill.ca)

Intro. There are various ways in which intonation can affect the meaning of an utterance. However, the meaning contribution of an intonational contour is often subtle and difficult to capture precisely. A useful formal perspective for certain contours has been to draw a connection to the semantics of Focus and Focus-particles (Constant 2012, Bianchi et al. 2016). Here we investigate two readings of the rise-fall-rise contour (RFR; Ward & Hirschberg 1985) and the contextual factors underlying them by drawing a parallel with an ambiguity of *at least* motivated by three auditory ratings studies.

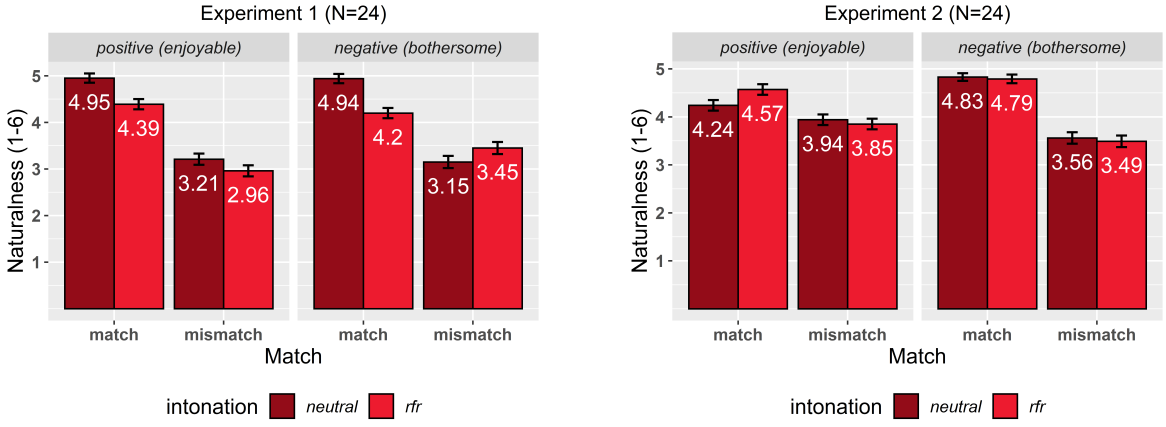
Background. Prior research on the RFR has mostly focused on its effect in replies to questions, where the contour conveys a sense of uncertainty or incompleteness regarding some alternative (Constant 2012, Wagner 2012, Goodhue et al. 2016). However, there are counterexamples to this characterization where all alternatives are resolved without rendering the contour infelicitous, as in (1) (Wagner 2012). Moreover, the RFR has been argued to show an asymmetry in argumentative dialogues like (2) (Göbel 2019). This asymmetry is surprising given that both replies seem to argue against A's statement such that any relevant alternative should be equally (un-)certain.

(1) A: It sucks that Cam didn't feed all of the cats. - B: She fed SOME of them. [AUDIO]

(2) a. A: That was a really bothersome hike today. - B: It was sunny. [AUDIO]
b. A: That was a really enjoyable hike today. - B: ??It was pouring. [AUDIO]

Notably, in (1)-(2) the RFR occurs in reply to a prior assertion. Experiments 1 and 2 were meant to experimentally test the intuition behind (2) and to what extent prior context plays a role.

Exp1&2. We used an auditory rating experiment where participants listened to recordings of 8 dialogues like (2), with A's prompt being either a statement (Exp1) or a question (e.g. *Do you think today's hike was bothersome/enjoyable?*, Exp2), and rated the naturalness of the dialogue on a scale from 1 (completely unnatural) to 6 (completely natural). Both experiments crossed three factors in a within-subject design: the VALENCE of A's utterance (negative $\widehat{=}$ *bothersome*, positive $\widehat{=}$ *enjoyable*); whether the response MATCHed in valence (e.g. *bothersome+pouring*, *enjoyable+sunny*) or not (as in (2)); and the INTONATION of B's response (neutral vs. rfr). We used ordinal mixed effects models to analyze the data, coding neutral and match as reference levels for their respective factors while VALENCE was Helmert coded. For Experiment 1, there was a large preference for match over mismatch and neutral over rfr, but the decrease for mismatch was ameliorated with the RFR. Crucially, this amelioration was larger when the context was of negative valence. Experiment 2 also showed higher ratings for match than mismatch, but a marginal preference for rfr over neutral and notably only a marginal three-way interaction. Additionally, the mismatch penalty was smaller in positive contexts.

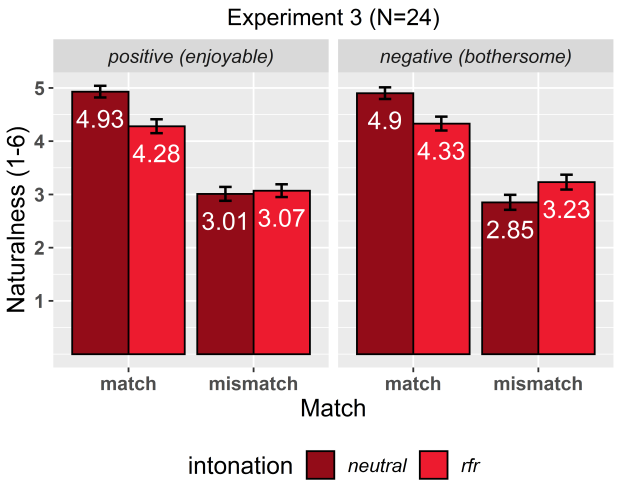


Interim Discussion. The results provide evidence for the existence of a valency asymmetry where the RFR is more acceptable when used in a positive reply than a negative reply. However, this effect seems to be restricted to replies to assertions and is weakened or disappears in replies to questions. We suggest that this context dependency can be understood in terms of an ambiguity similar to that of *at least*, which allows an epistemic interpretation conveying uncertainty (3) and a concessive interpretation conveying an evaluation (4a) (Nakanishi & Rullmann 2009, Biezma 2013, Chen 2018). Interestingly, concessive *at least* also exhibits a valence asymmetry, being prohibited from negative replies to positive statements (4b). We can thus think of the characterization of the RFR from prior research in terms of uncertainty as analogous to epistemic *at least*, while the valence asymmetry in replies to assertions is analogous to concessive *at least*.

(3) A: Did Cam feed the cats? - B: She fed **at least** SOME of them.

(4) a. A: That was a really bothersome hike today. - B: **At least** it was sunny.
b. A: That was a really enjoyable hike today. - B: #**At least** it was pouring.

Exp3 investigated what factors aside from whether the RFR is used in reply to a question or an assertion affect the valence asymmetry, namely by following up on the intuition that adding *also* to the reply weakened the contrast between positive and negative replies. We conducted another auditory rating experiment that only differed from Experiment 1 in that the replies contained *also*. The results patterned largely like those of Experiment 1, with higher ratings for match than mismatch and neutral than rfr, and the mismatch penalty being smaller with the RFR. However, the three-way interaction was no longer significant, failing to provide evidence for a valence asymmetry.



Final Discussion. The explanation we want to suggest for this effect of *also* is that *also* blocks the concessive reading of the RFR by virtue of restricting the alternatives considered in the computation of the contour to those concerned with truth. As a consequence, the RFR no longer conveys valence but merely that a higher alternative is possibly true, which is compatible with both prior assertions. On this account, the meaning of an intonational contour like the RFR thus makes use of the same types of alternatives as those required by a Focus-particle like *also* rather than being calculated separately, for instance as a conventional implicature à la Constant (2012). Our investigation thus sheds light on issues of semantic composition that go beyond the refinement of the contribution of the RFR while highlighting the usefulness of drawing connections to insights on Focus-particles. A remaining open issue concerns the cause of the smaller decrease for a mismatch in positive contexts in Experiment 2. One possible explanation might be that items differed in the extent to which they provided a clear answer to the question. A follow-up study supported this account, showing that replies in positive contexts were taken by participants to provide weaker confirmation in the match condition and weaker denial in the mismatch condition.

References: Bianchi et al. (2016). *Semantics and Pragmatics* • Biezma (2013). *PLC 36* • Chen (2018). PhD Thesis, Rutgers University • Constant (2012). *Linguistics and Philosophy* • Göbel (2019). *SALT 29* • Goodhue et al. (2016). *NELS 46* • Nakanishi & Rullmann (2009). *CLA 2009* • Wagner (2012). *Semantics and Pragmatics* • Ward & Hirschberg (1985). *Language*

Investigating a shared mechanism in the priming of *manner* and *quantity* implicature.

Introduction: Debate exists surrounding the nature of the mechanism that derives *quantity* implicatures (e.g., the inference '*not all of the cars are green*' from '*some of the cars are green*'). According to Gricean-inspired approaches, *quantity* implicature derivation is accounted for as a pragmatic, socially oriented phenomena. In contrast, competing accounts consider *quantity* implicature to be grammatically rooted (Chierchia et al., 2012, 2019; Fox, 2007).

Recently, structural priming paradigms have been adapted by Bott & Chemla (2016) and Rees & Bott (2018) to investigate different types of *quantity* implicature (*scalar quantifiers*, *ad hoc quantity* implicatures and *numerals*). Bott & Chemla conclude 1) *quantity* implicature subtypes can prime their own subsequent derivation (e.g., *scalar* implicature can prime the derivation of succeeding *scalar* implicatures) and 2) between certain subtypes of *quantity* implicature, a cross-priming effect can be observed (e.g., *ad hoc* implicature can prime *scalar* implicature). A cross-priming effect suggests that there are shared mechanisms involved in the derivation of certain subtypes of *quantity* implicature. However, the evidence of a shared derivational mechanism is compatible with both Gricean-inspired accounts and grammatically oriented approaches. As per a Gricean account, certain subtypes of *quantity* implicature require the same considerations of a more informative, unsaid, alternative, and assumptions of speaker cooperation and informativity to be derived – it may be that the relevance of these considerations and assumptions is primed between the experimental trials. In contrast, the grammatical account posits the existence of a covert operator ***O*** (semantically expressed as '*only*'), which is inserted within the syntax of an utterance and triggers the derivation and negation of a more informative alternative (e.g., ' ***O***[*some of the cars*] *are green*' = '*only some of the cars are green not all of them'*)

To utilise a structural priming paradigm as a tool to reach theory-critical conclusions, we investigated whether a structural priming can be used to prime *manner* implicature (e.g., the inference '*we danced in an unusual/uncharacteristic way'* derived from the utterance *'We moved our limbs to the music'.*). Like *quantity* implicatures, *manner* implicatures are triggered by the derivation and negation of an unsaid alternative expression. With *manner* implicature, the 'alternative' is the non-marked, typical expression (e.g., '*we danced [typically]*'), which is negated (e.g., '*we did not dance typically'*) and is triggered by the use of obscure or unduly lengthy utterances (see Horn, 1991, Levinson 2000). Importantly, what is negated is not the semantic content of the alternative, unmarked, expression, but the typical connotations of the expression.

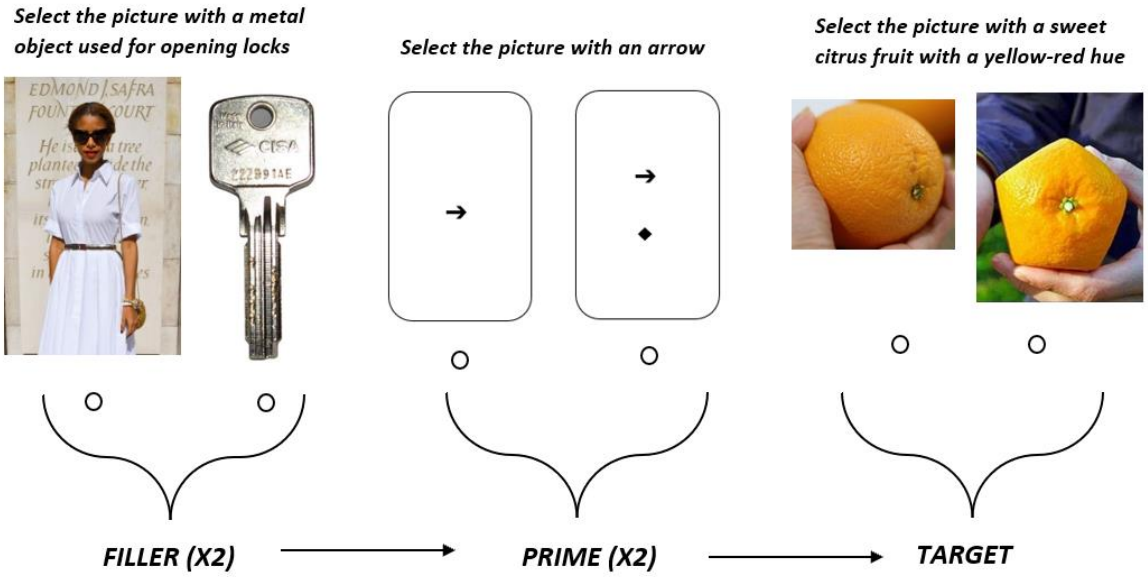
Research Question: the current study investigated two novel questions: 1) Can manner implicatures be primed? and 2) If so, is there cross-priming between *manner* and *quantity* implicatures? **Predictions:** The differences between *manner* and *quantity* implicatures mean that the grammatical approach does not predict any cross-priming between the two types. Unlike in the case of different subtypes of *quantity* implicature, the insertion of a grammatical operator ***O*** will fail to derive a *manner* implicature, as it will derive informationally stronger rather than similar alternatives (e.g '*We* ***O*** *[moved our limbs to the music]*' = we only literally moved our limbs to the music, i.e., we didn't dance). As per a Gricean account, both *manner* and *quantity* implicatures only require consideration of the speaker's cooperative intentions. Therefore, any type of implicature may lead to the priming of another type of implicature.

Experiment 1: aimed to investigate *manner → manner* priming effects. We recruited 180 adult monolingual English speakers. Exp.1's trials comprised of 30 trials: 6 target trials, 12 priming trials and 12 filler trials, presented in a *filler→filler→prime 1→prime 2→target order*. The trials were configured as per *trial 5)* in Fig.1, and both primes and trials involved *manner* implicature. **Results:** The mean rate of manner implicature in the target trials stood as 16.23% (SD = 12.34%); an increase of 4.37% from our preestablished baseline of 11.86% (p= .0221). The baseline rate of implicature, while low, is expected of one-off, context dependent phenomenon. A 4.37% increase from the baseline suggests that the *manner* primes primed implicature derivation in the subsequent target trials.

Experiment 2: after supplementary experiments reconfirming Bott & Chemla's assertion of *quantity → quantity* priming, we conducted a series of cross-priming experiments. The participant selection and paradigm structure were functionally identical to that of Exp.1, except *prime 1* and *prime 2* consisted of *quantity* primes (both *scalar* and *ad hoc*) and the target trials of *manner* implicatures (see Fig.1, *trials 3)* and *4)* for *ad hoc* primes). **Results:** after *ad hoc* primes, we observed a mean rate of manner implicature of 18.04% in the target trials, an increase of 6.18% from the manner baseline (p =0.0022). After *scalar* primes, we saw a mean implicature rate of 15.67% an increase of 3.78% from the manner baseline (p=0.0420). Overall, the priming effect of *manner*, *scalar,* and *ad hoc* primes on *manner* targets is comparable – no single prime type outperforms the others.

Conclusions: Firstly, *manner* implicature is indeed primeable. While the formation of the experimental items was difficult due to the inherently ad-hoc, context-dependant nature of *manner* implicature, the data shows that priming paradigms can be used to investigate the nature of *manner* implicature. Secondly, the increase in *manner* implicature after *quantity* implicature primes has important ramifications for accounts that posit *quantity* implicature as a purely grammatical phenomenon as the observed cross-priming effect suggests that a shared derivational mechanism between *manner* and *quantity* implicature exists. While the presented data cannot rule out a grammatical component of *quantity* implicature, it certainly suggests that *quantity* implicature has similarities with *manner* implicature, and that these similarities are wholly pragmatic in nature.

Figure 1 - a trial set for Experiment 2's ad hoc primes



The development of irony comprehension and epistemic vigilance

Introduction. Irony (e.g., uttering "You're so kind" to criticise someone who has not provided the expected support) has been found to be a relatively late acquisition: several studies suggest that children only start grasping ironical utterances from around the age of 6. However, some studies have suggested that a sensitivity to some aspects of irony (e.g., speaker's using a characteristic tone of voice/facial expression) may arise earlier than this (AuthorX and AuthorY, 2021). The current experiment takes as its starting point the relevance-theoretic *echoic* account of verbal irony (Wilson & Sperber, 2012) and addresses the development of the recognition of irony based on the sensitivity to the inappropriateness of what is said. The aim was to investigate whether children's epistemic vigilance towards utterance content (Sperber et al. 2010) might help them detect the ironical speaker's dissociative attitude towards the proposition literally expressed by her utterance. An utterance echoing a thought that is very inappropriate is likely to be recognized as ironic because (1) it is more likely to be the target of a dissociative attitude, and (2) because of epistemic vigilance mechanisms, the utterance is less likely to be interpreted literally since one is unlikely to assume that the speaker intends her audience to accept such an inappropriate thought. Furthermore, since children are found to often provide literal interpretations of ironical utterances in experimental settings (AuthorX & Author Y, 2020), we hypothesised that the absence of a literal option - and using videos which revealed the ironical speaker's facial expression - would improve children's irony comprehension.

Design. We designed a novel irony task which does not require a verbal response. Participants saw short movies involving a young woman and a hand puppet (see Figure 1). The hand puppet has to guess what is on a picture (e.g., "a completely full glass") and the woman either praises his guess or mocks it ironically. The sentences in both the literal and the ironical condition are identical with respect to wording (e.g., "'Yes, this is really a completely full glass"), but they differ in the speaker's attitude: sincerely happy versus ironical intonation and facial expressions. Based on these audiovisual cues, the participants had to choose which of two images (e.g., a full or an empty glass) is depicted on the card in the speaker's hand. The two pictures represent different points of a scale: a literal option (e.g., a full glass), a medium option (a half full glass) or an extreme option (an empty glass). All participants watched 12 videos varying in utterance type (literal, ironic), picture combination (literal-extreme, literal-medium, medium-extreme) and scale (e.g., full-empty, happy-sad, hard-soft). We measured picture choice and eye gaze while the sentence unfolds.

Figure 1: Screenshot from a video stimulus

Participants. Eighty-nine Norwegian-speaking children between 3 and 7 years of age (range: 3.08 to 7.33 years; mean age: 5.12 years; 41 female) and a control group of 20 adults (range: 20 to 53 years; mean age: 28.5; 16 female) participated in the study.

Results

Picture-selection. The accuracy of picture-selection in children was 89 percent for literal utterances and 12 percent for ironical utterances; for adults it was 97 for literal utterances and 85 percent for ironical utterances. Put differently, children selected the picture most closely aligned with a literal interpretation, regardless of whether a literal or ironical utterance was presented. To give an example, when hearing the sentence "Yes, this is really a completely full glass", children overwhelmingly chose the picture depicting a full glass – and in case this literal picture option was not available, they picked the half full glass over the empty glass.
We analyzed children's picture-selection data with mixed effects logistic regression in R (version 4.1.2.), using the binary response accuracy of picture selection as a dependent variable. The model includes random intercepts for subjects and items, as well as age, condition (irony, literal) and picture combination (literal-extreme, literal-medium, medium-extreme) as fixed effects factors. Children selected the correct picture more often in the literal condition than the irony condition (β = 4.482, z = 18.136, p < .001). Age was weakly significant (β = 0.211, z = 1.972, p = 0.049), mostly driven by the fact that older children tended to be more accurate than younger children in the literal condition, albeit not the irony condition. The type of picture combination did not affect children's accuracy of picture selection.

Gaze. We analysed the proportion of looks while the utterance unfolds (plus 300 ms afterwards) to four areas of interest: the two pictures as well as the faces of speaker and addressee. Both children and adults preferred to look at the picture closest to a literal interpretation, regardless of utterance type (irony, literal). However, when comparing the looks to the two pictures in both conditions, calculating a difference score, children's preference for the picture that is most in line with a literal interpretation turned out to be stronger in the literal condition than in the irony condition (p =.006), similar to adults. Furthermore, children looked more at the speaker in the irony condition compared to the literal condition (0.54 vs. 0.45, p= .020).

Discussion and conclusion

The offline data from the picture-selection task show no evidence of irony understanding in children aged 3 to 7 and, surprisingly, no improvement of irony understanding with age. With just 12 percent correct picture choice in the irony condition, children were substantially below the chance level of 50 percent, showing a strong bias to interpret ironical utterances literally. This was the case even when the ironical utterance was echoing a thought that was very inappropriate in the context, suggesting little help from epistemic vigilance mechanisms. Removing the picture representing the literal interpretation did not improve children's irony comprehension, as they tended to pick the picture closest to the literal option on the scale. By contrast, the gaze data reveal that children looked less at the "literal" picture in the ironical condition compared to the literal condition. Furthermore, children pay special attention to the facial expressions of an ironical speaker. Both findings could indicate a sensitivity for irony, not captured by the offline results. A possible explanation for the observed divergence between offline and online measures could be the high demands of the picture-selection task, requiring children to infer the state of the world solely based on the speaker's tone of voice and facial expressions. The fact that in standard narrative paradigms the majority of children is able to understand irony by the age of 6 (e.g., AuthorX and AuthorY, 2021), our goal to create a simple and sensitive irony task was not successful. However, with our new methodology we were among the first to study the role of facial expressions in children's interpretation of irony, something further studies should explore in more detail.

References. AuthorX and AuthorY, 2020, 2021; Sperber, D. et al. Epistemic vigilance. *Mind & Language, 25*(4), 359–393; Wilson, D. & Sperber, D. 2012. Explaining relevance. *Meaning and Relevance,* Cambridge UP 2012.

Proportions vs. cardinalities: Comparative ambiguities and the COVID pandemic

Elsi Kaiser <emkaiser@usc.edu> University of Southern California

We report a study on quantity comparatives that are ambiguous between **cardinal** and **proportional** readings. In degree-based semantics, comparatives express relations between degrees on a scale [1,2,9]. (i) is true if we compare *cardinalities of people*. But the 'reverse' in (ii) is true if we compare *proportions*: a larger proportion of Ithacans know their neighbors [8]. Here the *scale* tracks 'proportions of a totality', not cardinalities [8].

(i) ***Cardinal*** More residents of New York City than Ithaca know their neighbors
$|NYC_{know\_neighbor}| > |ITH_{know\_neighbor}|$
(ii) ***Proportional*** More residents of Ithaca than New York City know their neighbors
$|\underline{ITH_{know\_neighbor}}| / |ITH_{population}| > |\underline{NYC_{know\_neighbor}}| / |NYC_{population}|$

Though both readings are available, questions remain about whether – in constructions of the type in (i-ii) – one reading is preferred and if so, what modulates this (see also [3,7]), and what this means for how to capture the existence of scales ranging of degrees of proportions [8]. Prior work largely assumes cardinal readings are preferred [8]. We test this experimentally, and suggest that a dispreference for proportional readings, if it exists, could be due to their greater complexity (depend on numerator, denominator). In addition, cardinal vs. proportional readings (at least with certain quantifiers) have been argued to be constrained by predicate type [6,8], in particular stage-level predicates (describing transient properties, e.g. *is feverish*) vs. individual-level predicates (describing stable/permanent properties, e.g. *has a college degree*).

The cardinal-proportional ambiguity has been highlighted by the COVID pandemic, as shown by confusion about public health information (iii). This situation also provides a meaningful, naturalistic context for experiments. We test 2 hypotheses: **(a)** *Simplicity hypothesis:* Cardinality readings are preferred over proportional readings in quantity comparatives (Exp1) and superlatives (Exp2). **(b)** *Predicate hypothesis:* Availability of cardinality vs. proportional interpretations is modulated by predicates (in ways that seem related to individual-/stage-level).

(iii) *Naturalistic example of confusion between cardinal and proportional readings (www)*
A: Alaska has more COVID than California…riiiight. ***B***: No, the percentage goes by their population individually (…) Yeah California is bigger buuuut the percentages are only going off each states numbers. (…) They aren't counting people, only percentages of those people

We test both statements about COVID cases/infections (*stage-level*) vs. vaccinations /vaccinated people (*individual-level*) to assess the predicate hypothesis in a realistic context.

Exp1 Comparatives. 139 native English speakers saw pairs of COVID county dashboards (Fig.1, 8 different pairs) and typed words into blanks (Table 1) to indicate their interpretations. In a pair, one county had higher absolute numbers of COVID cases (or vaccinated people); one had a higher proportional COVID rate (or higher % of vaccinated people): The cardinal/ proportional readings are truth-conditionally distinct. We tested 4 wording types (Table 1). More+*NP* conditions (2a,b) should allow cardinal *and* proportional readings (depending on predicates) and will shed light on our hypotheses. Two control conditions verify availability of cardinal (3a,b) and proportional readings (4a,b). The *'more COVID/vaccinated'* conditions (1a,b) are exploratory, testing if *eliminating direct reference to vaccinated people*/*countable cases* weakens the cardinality bias (which involves 'counting people'). To do this, (1a,b) use place-name meaning transfer [5], e.g. *Lakehorne County has more COVID than Blue Oak County*.

Exp2 (Superlatives, n=129) used 3-dashboard displays and superlative wording (*the most*, Table 1), to see if the results extend to proportional superlatives [4]. On a trial, one county had the highest absolute number; one the highest proportional number; one was in-between.

Results. In both studies (Figs.2,3), comparisons involving *COVID cases (stage-level) receive more cardinal readings than comparisons involving vaccination (individual-level)*. All COVID/vax differences are significant (p's<.05, glmer), except Exp2 'number' conditions. These effects coexist with an overall ***cardinality bias*** in both Exp1,2: All conditions yield above-

chance rates of cardinality readings (p's<.05), except for proportion controls (4a,b, as expected) and the 'more/most vaccinated' construction (1b). This fits our hunch that meaning transfer disprefers cardinal scales – suggesting the cardinality bias is malleable and not hard-wired. The exp1-2 parallelism is compatible with decompositional analyses of *most* [3,4].

Table1 *(people type county names in blanks)*	**Exp 1: Comparatives**	**Exp 2: Superlatives**
1a. People/cases not directly ment'd COVID	_ has more COVID than _	_ has the most COVID.
1b. People/cases not directly mentioned vax	_ is more vaccinated than _	_ is the most vaccinated.
2a. more/most COVID cases	There are more COVID cases in _ than _	_ has the most COVID cases.
2b. more/most vax'd people	There are more fully vaccinated people in _ than _	_ has the most fully vaccinated people.
3a. number of COVID cases *(card)*	The number of COVID cases is higher in _ than in _	_ has the highest number of COVID cases.
3b. number of vax'd people *(card)*	The number of fully vaccinated people is higher in _ than in _	_ has the highest number of fully vaccinated people.
4a. COVID case rate *(prop)*	The rate of COVID cases is higher in _ than _	_ has the highest rate of COVID cases.
4b. vaccination rate *(prop)*	The vaccination rate is higher in _ than in _	_ has the highest vaccination rate.

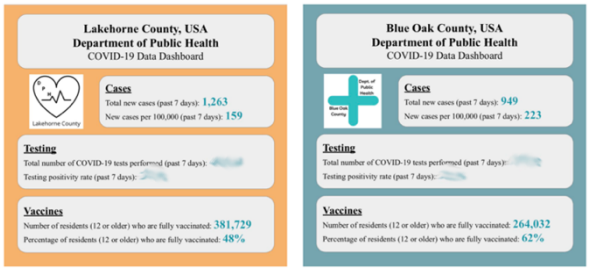


Fig 1. Exp1 example with two county dashboards, e.g. *Lakehorne County, Blue Oak County*. (Proportional COVID cases reported *out of 100,000*, vaccination rates as *%*, following common U.S. practice. Testing info was blurred out, it is irrelevant here.)

Fig2. *Exp1 Comparatives*: Cardinal vs. proportional interpretations

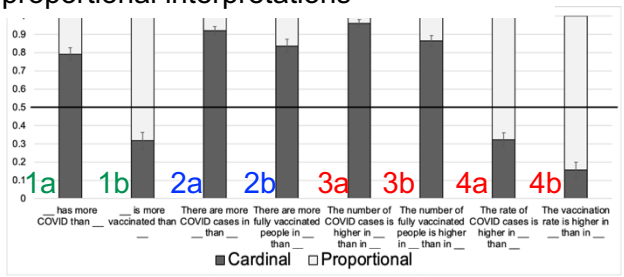
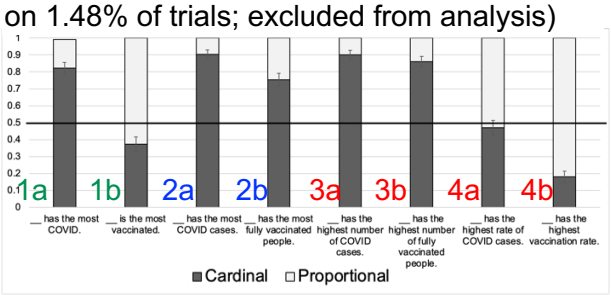


Fig3. *Exp2 Superlatives* (3rd county chosen on 1.48% of trials; excluded from analysis)



We provide new evidence that comparatives and superlatives refer to scales where the degrees *d* range over *degrees of proportion*. We identify a cardinal bias, but it is not rigid and can be weakened in favor of proportional readings by factors seemingly related to stage-/individual-level differences, and by certain linguistic forms (1b), suggesting specific syntactic and semantic factors impact scale interpretation (degrees of cardinality vs. proportion) in ambiguous contexts.

References: [1] Beck'11 *Comparison constructions* [2] Cresswell'77 *Semantics of degree* [3] Hackl'09 *On the grammar and processing of proportional quantifiers* [4] Kotek et al'12 *Many readings of most* [5] Nunberg'95 *Transfers of meaning* [6] Partee'89 *Many quantifiers* [7] Pietroski et al'09 *The meaning of 'most'* [8] Solt'18 *Proportional comparatives and relative scales* [9] von Stechow'84 *Comparing theories of comparison*

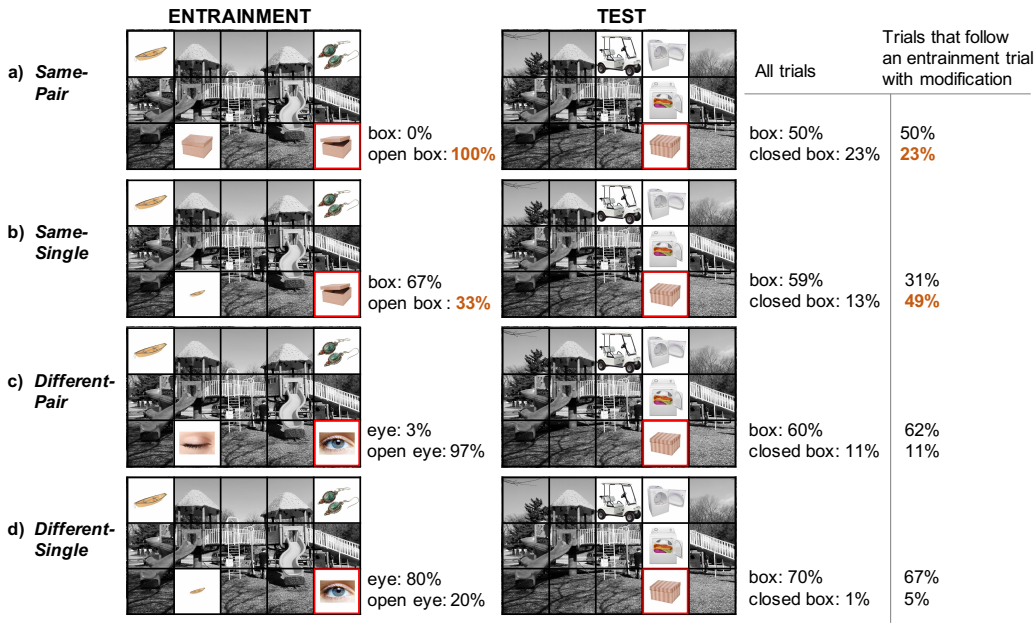
Referential domains, priming and the effect of invisible objects

Si On Yoon (U. of Iowa), Breanna Pratley (U. of Mass Amherst), Daphna Heller (U. of Toronto)

Referring Expressions (REs) are determined not just by properties of the referent, but also by the properties of *other* objects; the set of relevant objects is known as the *referential domain* [1]. The referential domain contains objects in the local visual context from which the referent needs to be distinguished: when there are two boxes in the visual context, referring to one of these requires a modifier (e.g., "*the open box*"). Interestingly, speakers sometimes include a modifier even when the contrasting entity is no longer visible, saying *"the closed box"* when the local visual context contains only a single box, but after they referred to a different (open) box earlier [2,3]. We note that this pattern suggests that previously-mentioned objects are also part of the referential domain, and ask (i) whether *unmentioned* earlier objects can also be part of the referential domain (Exp. 1), and (ii) whether earlier and current objects are all part of a *single* referential domain (Exp. 2).

General Method. Participants (n=24) performed a referential communication task over Zoom. Participants viewed grids of 15 "cards" each, completing 8 trials per grid: 1 ENTRAINMENT trial, 1 TEST trial, and 6 interspersed fillers. On each trial, 4 of the 15 cards were "flipped" to reveal their images, and the participant described a target card for the experimenter to click.

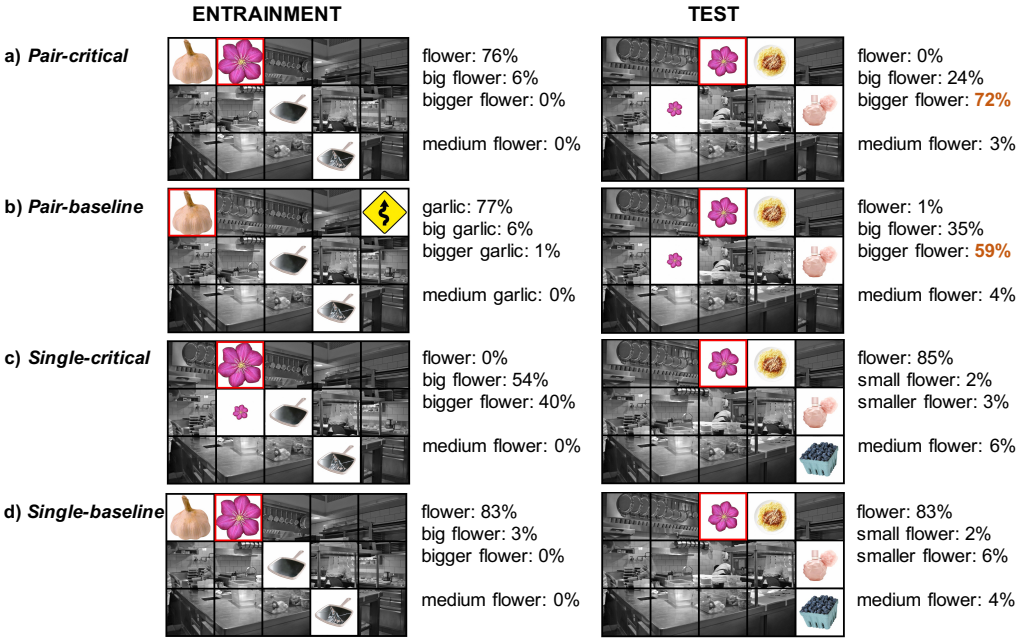
Exp. 1. To examine whether an earlier, unmentioned object is part of the referential domain, we manipulated whether the earlier ENTRAINMENT trial contained a pair of objects (e.g., an open and a closed box) or just a single object (an open box). The TEST trial was held constant: it always included one object (e.g., a striped closed box). If the referential domain only includes the earlier mentioned object (e.g., open box), the later RE should only encode contrast with this object (e.g., *"the closed box"*), regardless of the presence or absence of an earlier closed box. Alternatively, if the earlier contrasting object is part of the referential domain despite being unmentioned, speakers should avoid saying *"the closed box"* because this RE would not distinguish the current target from the earlier closed box. This pressure stands in contrast to a priming effect, whereby saying *"the open box"* earlier should prime *"the closed box"*. For control, we also manipulated whether the ENTRAINMENT trial included the same or a different noun (e.g., box vs. eye).



On ENTRAINMENT trials, speakers produced the modifiers at ceiling for pairs (same: 100%, diff: 97%), and much less for a single object (same: 33%; diff: 20%). This (expected) difference means

that speakers were more likely to be primed by their own modified REs in the pair conditions than in the single condition. To control for priming, we focused on those trials which had a modified RE in ENTRAINMENT (e.g., *"the open box"*). As expected from prior priming studies [e.g., 4], speakers produced more primed modifiers when the noun was repeated. More importantly, the primed RE was much *less* likely when the entrainment trial contained a second, unmentioned box (Same-Pair 23% vs. Same-Single 49%). This indicates that when the primed form (e.g., *"the closed box"*) did not distinguish the current target from the earlier, unmentioned box, speakers avoided using a RE that was sensitive to the historical context. This effect reveals that the earlier, unmentioned object is part of the referential domain.

Exp. 2. To examine whether all three objects are part of one referential domain, and to control for priming, we exploited the fact that the intermediate object in a set of three is called *"medium"* (pilot: 94%), but the same object is called *"big(ger)"* when paired with just one object (pilot: 97%). Participants described the object of intermediate size: (i) the TEST contained either a *Pair* of objects or a *Single* object, and (ii) the ENTRAINMENT trial either completed the set of 3 (Critical), or had one less object (Baseline). Most importantly, the effect of the historical context was again observed: comparatives (e.g., *"bigger"*) were more likely when a third object of the same category was seen earlier (72%) than when it was not (59%): speakers were less likely to call the medium object *"big"* when the historical context contained an even bigger flower. Nevertheless, speakers rarely produced *"medium"* in the critical conditions, revealing that the three objects do not in fact form a *single* referential domain. Importantly, these effects are independent of any priming effects (modifiers are not repeated across ENTRAINMENT and TEST).



Conclusions. We observe a novel effect where an entity is part of the referential domain – thereby affecting referential forms – despite not being physically present in the local context (and thus a potential referent) and not being referred to earlier. This effect reveals that speakers do not just represent the language previously uttered, but also aspects of the non-linguistic context that has given rise to their utterance. More specifically, these patterns could be explained by positing two simultaneous referential domains [cf. 5], one for the historical context, and a second one for the local visual context, with the local context taking precedence over the historical context.

References [1] Roberts (2003). *Linguistics and Philosophy* • [2] van der Wege (2009). *J of Memory and Language* • [3] Yoon & Brown-Schmidt (2013). *J of Memory and Language* • [4] Cleland & Pickering (2003). *J of Memory and Language* • [5] Heller et al., (2016). *Cognition*.

Source-Goal asymmetry in motion events: Sources are robustly encoded in memory but overlooked at test

There is a widely-attested asymmetry between Sources and Goals in people's description and memory of motion events. When describing an event such as a squirrel going from a mailbox to a trash can, people mention the Goal ("to the trash can") more often than the Source ("from a mailbox") and are also better at detecting Goal changes in same-different tests.[1,2,3,4] These findings are often taken as evidence for a homology between linguistic and conceptual representations: the unmentioned event component is less likely to be conceptually salient. However, we show that the Goal/Source memory asymmetry disappears when memory is probed with a forced-choice task. Thus Sources are present in event memory, but may not be attended to in same-different tasks.

In Experiment 1, 80 native English speakers first described video clips depicting motion events (each 5sec) in the same pseudo-random order. In critical events, an agent moved from a Source to a Goal (Fig1A), while fillers didn't include Goal/Source paths (e.g., a ghost moves around the moon). Sources and Goals were left-right counterbalanced and counterbalanced across lists such that Sources in one list were Goals in another. A memory test immediately followed the description task. On each critical test trial, participants saw a variant of the video with either a Source or a Goal change (Fig1B). They indicated whether each video was "exactly the same" as what they saw earlier. As expected, participants were more likely to mention the Goal in their linguistic description (β=1.091, SE=0.085, $p<0.001$) and were more likely to detect Goal than Source changes at the memory test (β=0.412, SE=0.118, $p<0.001$) (Fig2 Top).

Experiment 2 was the same as Experiment 1 except that, in the memory test, participants chose which video they had seen from 4 options: the target, a foil that only differed in the Source, a foil that only differed in the Goal and a foil that differed in both (Fig1C). Foil images were the ones used in Experiment 1. For each trial, whether the event chosen by the participant contained the correct Goal and/or the correct Source was respectively coded. Linguistic description results were identical to Experiment 1 (Fig2). However, the memory results differed. First, as expected, participants were much more likely to be correct in Experiment 2(M=67%) than in Experiment 1(M=33%). Importantly, participants no longer showed the Goal bias, but were in fact more likely to choose the correct Source than Goal (β=-0.267, SE=0.106, $p=0.012$).

The implications of the study are twofold. First, the memory Goal bias in the same-different task cannot be due to lack of Source encoding but to an on-line attentional bias during the test: in the forced-choice task, where contrasts in both Source and Goal are presented at the same time, the bias disappears. Second, the presence of the linguistic asymmetry in the absence of memory asymmetry in Experiment 2 suggests that what is encoded linguistically does not exhaust what is represented at the conceptual level and calls for a finer-grained homology between language production and event encoding in memory.

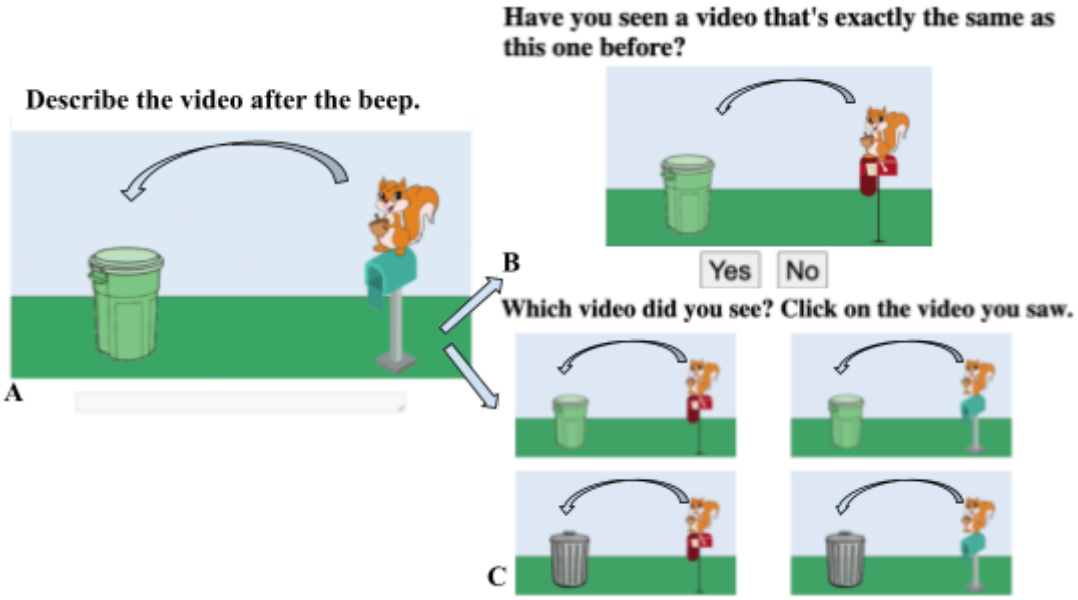


Fig1 A) A Sample Description Trial. B) A Sample Memory Test Trial in Exp1 (Source Change). C) A Sample Memory Test Trial in Exp2. Arrows represent the path of motion in the original video clips. Subjects did not see arrows and instead saw the videos.

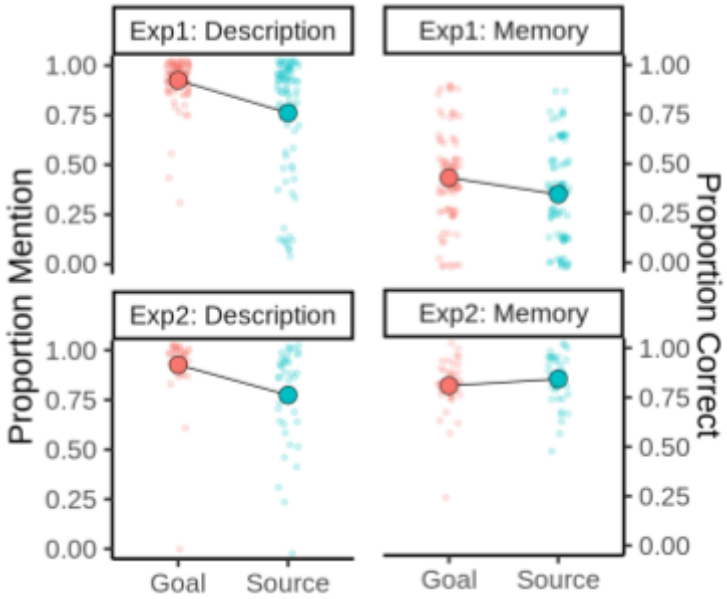


Fig2 Participants' mean proportion of mentioning Goal/Source in their linguistic description and mean proportion of correct Goal/Source response at memory test in Exp1 and Exp2. In the memory test in Exp1, correct response refers to successfully detecting the change of Goal/Source. In the memory test in Exp2, correct response refers to selecting the event that contains the correct Goal/Source.

References:

1. Regier & Zheng, 2007. *Cognitive Science.*
2. Papafragou, 2010. *Cognitive Science*.
3. Lakusta & Landau, 2012. *Cognitive Science*.
4. Do, Papafragou & Trueswell, 2020. *Cognition*.

Modelling the Role of Polysemy in Verb Categorization

A great deal of work has been devoted in recent years to developing computational models of meaning based on the distribution of words in text. Traditional static embeddings [Mikolov et al. 2013] represent each word type as a unique vector, while more recent contextual models [Devlin et al. 2019] generate a unique representation for every instance of a word in context. In this paper, we focus on the role of polysemy in verb categorization. Because verbs generally have multiple possible senses, categorization decisions depend on which sense of a word is being considered. Representing the distinct senses of polysemous words is thus important to modelling how humans categorize sets of verbs' denotations. This paper shows that the contextual information implicit in recent distributional semantic models makes them a good approximation of polysemy and of verb categorization.

To understand the unique role contextual embeddings can play in modeling effects of polysemy on verb categorization, it is important to note that there are at least two different approaches to the relationship between polysemy and context. One account holds that words have a static set of possible senses, and while context helps disambiguate between possible senses of a word, those senses exist independently of context. A stronger claim, though, has been made (for example, by Elman 2009) that different senses of a word are not simply reflected in, but actually *created by* context. Proponents of this claim believe that word meaning is fundamentally context-dependent. Contextual language models like BERT implicitly take this view of word meaning, as they represent each instance of a word in a particular context as a unique embedding. We can then treat classes of contexts as equivalent to senses in these models. This is why contextual embeddings seem well-suited to test Elman-style conception of polysemy and its role in verb categorization. If this view of polysemy is correct, contextual word embeddings should model verb categorization better than static embeddings. This is the hypothesis we test in this paper.

Interestingly, recent work evaluating different word embedding models on verb categorization suggests just the opposite. Majewska et al. [2021] found that contextual models perform poorly compared to older static models when approximating the verb categorization done by participants in their experiments. We argue that this result is due not to the irrelevance of context to categorization, but rather to the way the contextual embeddings were extracted from the model in Majewska et al. [2021]. Although many of the words in Majewska et al. [2021]'s ground truth data are polysemous and are assigned to multiple 'gold' classes by participants, they evaluate models in a one-representation-per-word-form manner. Even when evaluating BERT, which has been shown to encode sense-specific information, this information was thrown away by averaging over all contexts. Because they use polysemous data to test representations which do not encode sense information, Majewska et al. [2021]'s results may not reflect the full potential of contextual architectures to model categorization.

Our paper shows that by accounting for polysemy in the model representations, we can significantly improve the correlation between word embedding clusters and human categories. In particular, retaining sense-level information from contextual BERT embeddings more than doubles its performance, outperforming static embeddings by a large margin. These results demonstrate that sense-specific information is crucial even for categorization of words in isolation, and suggests that contextual embedding models are a good approximation of both polysemy and verb categorization, supporting a contextual account of word meaning. We evaluated two ways of handling polysemy in word embeddings:

1. **Static word2vec embeddings trained on POS-tagged data.** Part-of-speech tagging allows the model to distinguish between, for example, *duck_NOUN* and *duck_VERB*. This strategy

Model	F1-optimal	F1-gold
Majewska et al. [2021] word2vec	0.355	0.326
Majewska et al. [2021] BERT	0.340	0.322
Our POS-tagged word2vec	0.442	0.433
Our multi-prototype BERT	**0.755**	**0.731**

Table 1: Comparison of our methods with results reported in Majewska et al. [2021]. F1 scores reported. 'Gold:' k=17, as in the ground truth. 'Optimal:' best result for k in the range (5,30).

is simplistic as different senses which have the same part of speech are still conflated into one vector (like *get#ACQUIRE* and *get#UNDERSTAND*), but it at least factors out noise from non-verb senses.

2. **Multi-prototype BERT embeddings.** Following the methods of Chronis and Erk [2020], we distill BERT embeddings representing individual tokens into multiple prototype embeddings, which represent each sense of a word, without collapsing every token into a single representation, as in Majewska et al. [2021]. This allows for different senses of a word to be assigned to different clusters, while still generalizing beyond individual instances of a word. Multi-prototype BERT embeddings have been used to improve word pair similarity estimates, but to our knowledge this is the first time that such sense-level representations have been used to study the role of polysemy in other domains.

To evaluate the performance of each method, we use the verb categorization data from SpAVerb [Majewska et al., 2021] as a ground truth, which comprises 825 verbs in 17 semantic classes. This data was derived from a sorting task performed by 10 participants. We use k-means clustering to group verb embeddings into predicted classes. To compare the induced clusters with the ground truth, we use the same F1 metric used by Majewska et al. [2021], which balances precision and recall. Table 1 shows the results of the two methods compared to the results reported in Majewska et al. [2021]. The F1 value for the POS-sensitive word2vec model is slightly higher than reported for a similar model architecture without POS information. Multi-prototype BERT, by contrast, performs dramatically better than any of the results previously reported.

While Majewska et al. [2021] found that contextual models performed poorly compared to static models, our results indicate that properly exploiting its contextual information allows BERT to predict verb categories very well. Retaining sense-level information from BERT by generating multi-prototype representations, rather than generating one representation per word form, more than doubles its F1 score. This boost in performance shows that contextual, sense-specific information is important to human verb categorization, and supports a strongly contextual view of polysemy and word meaning. On a more general level, these results suggest that linguistic input encodes a great deal of information about semantic categories, independently of other perceptual input that humans receive, and that this category information can be extracted from embedding models which are trained on linguistic data alone. Future work is needed, though, to further explore the role of language in forming semantic categories, and whether models like those discussed here can model the flexible, goal-dependent nature of human categorization.

References

G. Chronis and K. Erk. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *CoNLL*, Online, Nov. 2020.

O. Majewska, D. McCarthy, J. J. van den Bosch, N. Kriegeskorte, I. Vulić, and A. Korhonen. Semantic data set construction from human clustering and spatial arrangement. *Computational Linguistics*, 47(1):69–116, 2021.

Context, Convention and Coordination: Insights from Gradable Adjectives

Chris Kennedy

Gradable adjectives denote properties that are relativized to contextual thresholds of application: how long an object must be in order to count as `long' in a context of utterance depends on what the threshold is in that context. But thresholds are variable across contexts and adjectives, and are in general uncertain. This leads to several questions about the meaning and interpretation of gradable adjectives in particular contexts of utterance, including: what truth conditions are they understood to introduce, what information are they taken to communicate, and how (if at all) do language users adapt their understanding of these expressions over time? In this talk, I will report on a series of studies that my colleague Ming Xiang and I have carried out that are prompted by these questions, which provide insights on the role of context, convention and coordination in our use and understanding of this particular class of context-dependent expressions.

2-year olds derive mutual exclusivity inferences from contrastive focus

Gabor Brody[1], Roman Feiman[1,*], Athulya Aravind[2,*]
[1]Brown University; [2] MIT; * contributed equally

Overview: When children hear a novel term in the context of two potential referents – one familiar or already named, and one novel – they tend to assume that the novel word picks out the unfamiliar referent, an effect dubbed "Mutual Exclusivity" (ME). In a typical study (Markman and Wachtel 1988 et seq.), children are presented with a novel object (e.g. a vacuum tube) and a familiar object (e.g. a car) – and asked which one is the "dax"; children as young as 17 months of age (Halberda 2003) reliably look to the novel object. While there are several competing explanations for why children (and adults) in these tasks treat *dax* and *car* as being mutually exclusive in reference, all of them invoke a bias to avoid applying two labels to the same object. This study tests an alternative hypothesis that the exclusivity inference is a consequence of a well-attested grammatical phenomena present in adult language: **focus structure**.

Theoretical background: A standard assumption within linguistic semantics is that representations of sentences contain markers of *givenness* and *focus*, which trigger distinct discourse requirements (Rooth 1992, Buring 2016, Kratzer and Selkirk 2020). G(ivenness)-marking on an expression indicates that its meaning is salient in and recoverable from the preceding discourse. F(ocus)-marking on an expression indicates that its meaning contrasts with a salient alternative in the preceding discourse. In languages like English, these markers affect the prosodic realization of a sentence, such that differences in prosody correspond to systematic differences in interpretation. G-marked expressions are de-accented (1a); F-marked expressions are accented (1b).

(1) A: How did you like the conference?
  a. I liked the talks_{G}. → speaker liked the conference (talks ≈ conference)
  b. I liked the TALKS_{F}. → speaker did not like other salient aspects of the conference

We propose a novel hypothesis that such information-structural cues play a critical role in mutual exclusivity inferences. F-marking on the critical NP (indicated by accenting) should prompt listeners to exclude contrastive alternatives in the context (e.g. the already labeled or familiar object), resulting in an ME inference even if the noun is not novel. But, if the NP is marked as given (indicated by de-accenting), listeners should look for a coreferential salient discourse antecedent, resulting in no ME effect. To test these predictions, we manipulate F- and G-markings on the noun-phrase prompt and test whether children make an ME inference.

Study. Logic and design: Because our study tests whether cues in information structure predict ME effects in a context where the target label could *in principle* apply to both objects, instead of using a novel noun in the carrier phrase, we used "the toy". After a short warm up game, participants were presented with 6 test trials where they saw two novel objects and an on-screen communicator, Foxy. First, Foxy pointed to and labeled one of these objects with a novel label (e.g. "blicket", which was always F-marked to introduce a new referent). Then in the test phrase, Foxy asked the participant to "point to the toy". Crucially, we manipulated between subjects

Figure 1: *Two frames of a test trial. Foxy first introduces one of the objects, then asks the participant to "point to the toy", where NP is either F-marked(accented) or G-marked (de-accented).*

whether “the toy” was F-marked (accented) or G-marked (deaccented).
Hypotheses: We predicted that children who hear F-marking on “the toy” should assume that the expression contrasts with a salient alternative. As the only such alternative is “blicket”, the child should reason that the blicket is distinct from the toy, so the referent of “the toy” has to be the other object. On the other hand, if children hear G-marking on “the toy”, they should assume that its meaning is recoverable from the available discourse. As “blicket” is the only salient antecedent, they should assume that “the toy” refers back to the same object as “blicket”.
Participants: We report findings from a 10 participant pilot study (mean age: 2y;9mo) and an in-progress study with 14 participants (mean age: 2y;6mo; pre-registered full sample size = 50).
Results: In both the pilot sample (Figure 1a) and the in-progress sample (Figure 1b), children in the Focus condition are much more likely to choose the new object – i.e. derive an ME inference – than children in the Given condition. For both of these samples we conducted a mixed effects logistic regression (model syntax: *ChoiceOfNew ~ Condition + (1|Participant)*). Both revealed a significant effect of condition (pilot sample (β = -22.122, SE = 9.59, z = -2.59, p = .01); in-progress sample (β = -5.1743, SE = 1.41, z = -3.67, p < .001))
Discussion. Our findings support the hypothesis that children can use information structure to decide whether the referent of an NP should be recoverable from prior discourse (Given condition) or contrast with previously mentioned referents (Focus condition). This result opens the door to a possible reinterpretation of the ME inference as a result of contrastive focus. While past studies did not systematically manipulate information-structure to our knowledge, we suggest they may have tended to present their linguistic stimuli with prosodic prominence on the novel label, since this is the natural way to introduce new referents, thus marking the expression as focused. The upshot is that the grammar-based model of early ME does not require positing either conceptual or pragmatic biases to derive the inference and can provide principled answers to long-standing questions of when and how ME inferences should be suspended (see Bloom, 2001).

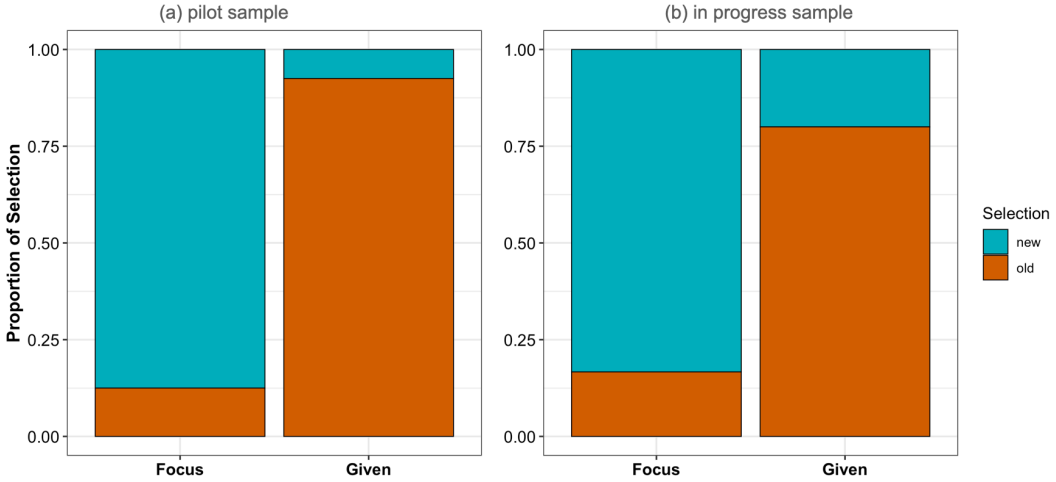


***Figure 2**: Rate of target selection (new versus old object) in pilot sample (a) and in in-progress sample (b) across Focus and Given conditions*

References: Bloom, P. (2002). *How children learn the meanings of words* - MIT Press · Büring, D. (2016). I*ntonation and meaning* - Oxford University Press · Halberda, J. (2003). The development of a word-learning strategy - *Cognition* · Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words - *Cognitive psychology* · Kratzer, A., & Selkirk, E. (2020). Deconstructing information structure - *Glossa* · Rooth, M. (1992). A theory of focus interpretation - *Natural language semantics*

Lexical Aspect Maps Onto Event Apprehension

1. Introduction. Aspectual theories in semantics draw a distinction between telic verb phrases denoting events with an inherent endpoint (e.g., *peel a banana*) and atelic verb phrases denoting events that lack an inherent endpoint (e.g., *peel*; Krifka, 1998; van Hout, 2016). Telicity is assumed to correspond to alternative perspectives on what could be the same underlying eventuality; furthermore, telicity is frequently taken to map onto nonlinguistic event structure – specifically, whether an event is taken to be bounded or not (Filip, 1993; Ji & Papafragou, 2020a, b). Here we present one of the first direct tests of these assumptions. We ask whether prior exposure to telic vs. atelic descriptions of an event influences how the event is mentally processed (i.e., whether it is construed as bounded vs. unbounded).

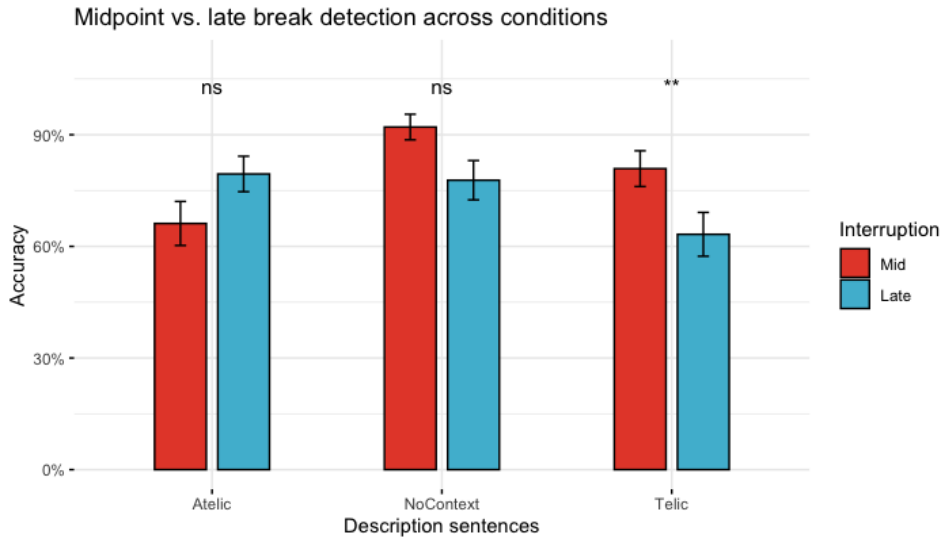
2. Stimuli. We created 15 videos in which the same woman performed an action on an object (drew a picture, peeled a banana, blew a balloon, etc.). Each lasted 10.6 sec on average (range: 7.31-14.81). In a norming study, these videos were overwhelmingly judged as depicting "something with a beginning, midpoint and specific endpoint" (i.e., had bounded construals). Nine of these videos (test items) were then edited to place a visual interruption of .13s at the temporal point corresponding to either 50% of the video (mid-interruption) or 80% of the video (late-interruption). Filler items were left intact. We then rotated mid- and late interruptions across lists of test videos such that each participant only saw one type of interruption per test video.

3. Procedure. Ninety monolingual English speakers were presented with a scenario in which a woman was recovering from surgery and had to perform a range of exercises to regain her fine motor skills. Participants were randomly assigned to one of three (between-subject) Context conditions: Telic, Atelic, and No Context. In the Telic and Atelic conditions, prior to each trial, viewers were presented with a sentence describing the exercise that the woman had to do in either telic ('Draw a picture') or atelic ('Do some drawing') framing. The sentence was displayed for 6.5s, then the video clip started. After watching the video, participants were reminded of the exercise and had to indicate via a key press whether the woman had done the exercise or not. As a secondary task, they then had to identify whether there was a glitch in the video by pressing a key. In the No Context condition, participants were given the same cover story but without descriptions, and only had to determine whether there was a glitch in the video or not.

4. Hypotheses. We predicted that atelic descriptions would lead viewers to construe the event as unbounded (i.e., homogeneous and lacking a boundary at the end), and telic descriptions would encourage a bounded event construal (consisting of discrete steps culminating in a specific endpoint; Ji & Papafragou, 2020a). If so, these construals should have distinct signatures on the mental processing of interruptions during event apprehension. Because what happens at event endpoints is critical for cognitive processing (e.g., Shipley & Zacks, 2008), we expect that viewers should attend more to the content of the videos at endpoints compared to midpoints in the Telic condition, and therefore *miss* irrelevant, content-external Late compared to Mid- interruptions (see Shipley & Zacks, 2008 for evidence that attention leads to ignoring external distractors at event endpoints). However, since unbounded events do not have canonical endpoints - they stop, but do not culminate, endpoints should be treated largely similarly to other time points; hence in the Atelic condition the placement of the interruption should not make a difference. The No Context condition served as a control. Because there was

only a single task here, it was possible that both Mid and Late interruptions would be equally easy to detect.

5. Results. We found a significant interaction between Context and Interruption type ($p<0.001$). A mixed-effects model with interruption type as a fixed effect with random intercepts for participants and items showed that interruptions had a significant effect on accuracy ($\chi^2(2)$: 12.976, $p < 0.01$) in the Telic condition only. Post-hoc testing on midpoint and late point accuracy differences showed that they were only significant in the Telic condition (z: -2.736, Tukey adjusted $p = 0.01$). As expected, participants in the Telic condition had lower accuracy rates for late point interruptions compared to midpoint interruptions (Mid: 81% vs. Late: 63%) but this difference disappeared in the other conditions (Atelic: Mid: 66%, Late: 79%; NoContext: Mid: 92% Late: 78%).



6. Discussion and Conclusion. Our results show that identical events presented with telic compared to atelic descriptions were more likely to be processed as bounded events (as evidenced by how viewers processed temporal points within each event). These data confirm that telicity is a perspective on the temporal profile of otherwise multiply interpretable streams of experience. Our data further support a mapping between linguistic telicity and boundedness in non-linguistic cognition, and a broader alignment between linguistic and cognitive representations (Pinker, 1989).

References: Filip, Hana. 1993. Aspect, situation types and nominal reference: University of California, Berkeley dissertation. Ji, Yue & Anna Papafragou. 2020a. Is there an end in sight? Viewers' sensitivity to abstract event structure. Cognition 197. 104197. Ji, Yue & Anna Papafragou. 2020b. Midpoints, endpoints and the cognitive structure of events. Language, Cognition and Neuroscience 35. 1465 – 1479. Pinker, Steven. 1989. Resolving a learnability paradox in the acquisition of the verb lexicon. Paul H. Brookes Publishing. Shipley, Thomas F & Jeffrey M Zacks. 2008. Understanding events: From perception to action. Oxford University Press. Van Hout, Angeliek. 2016. Lexical and grammatical aspect. In The Oxford Handbook of Developmental Linguistics, Oxford University Press.

Perfective accomplishments don't always denote (maximal) event culmination, even in Russian: Evidence from psycholinguistics

Natasha Kasher & Aviya Hacohen, Ben-Gurion University of the Negev

Overview It is a widely established view in the event-semantics literature that perfective (PFV) telic accomplishments, comprised of a dynamic verb and a quantized incremental theme argument (e.g., Krifka 1989), denote culmination (Parsons 1990). It has also been increasingly recognized over the past two decades that such constructions demonstrate varying degrees of culmination requirements crosslinguistically (see Martin 2019 for a detailed list). However, while PFV non-culminating accomplishments have been found in a variety of languages and language families, the Slavic PFV has been consistently argued throughout the theoretical and psycholinguistic literature to enforce strict culmination requirements on accomplishments within its scope, such that non-culminating interpretations are entirely disallowed for such forms (e.g., Filip 2017), and PFV accomplishments followed by a cancellation phrase (PFV+CNCL) result in a contradiction. This is illustrated by the contrast between Hindi (1) and Russian (2):

(1) maya-ne biskuT-ko khaa-yaa par us-e puuraa nahiin khaa-yaa
Maya.ERG cookie.ACC eat.PERF but it.ACC full not eat.PERF
'Maya ate a cookie (but not completely).' (from Arunachalam & Kothari 2011)

(2) Masha s'ela prjanik (#no ne ves').
Masha. PFV.ate.SG.F gingerbread.cookie.ACC (#but not all).
'Masha ate a/the gingerbread cookie (#but not all of it).'

We report results from a gradable acceptability judgment task, which challenges this generally assumed typology. We show that while Russian PFV accomplishments do carry culmination requirements, they are not stricter than what has been reported for other languages. Moreover, our data reveal high acceptability ratings for (PFV+CNCL), indicating that the culmination inference of the PFV accomplishment is defeasible, even in Russian. We discuss how these results are in line with Kearns' (2007) distinction between the *standard telos* and the *maximal telos*, and what they suggest with respect to the semantics and pragmatics of telic accomplishments.

Methods Experimental items included 8 accomplishment predicates, comprised of an incremental transitive verb + a singular count direct object. Each base accomplishment appeared in three aspectual frames: (1) perfective (PFV); (2) perfective followed by a cancelation phrase (PFV+CNCL); (3) imperfective (IMP), as illustrated in the table for 'draw a/the star':

Condition	Example
1. PFV	*Malčik* ***na****risoval zvezdu.* Boy **PFV**.drew star.ACC ('The boy drew a/the star.')
2. PFV+CNCL	*Malčik* ***na****risoval zvezdu, no odnovo lučika ne xvataet.* Boy **PFV**.drew star.ACC but one ray not sufficient ('The boy drew a/the star, but one point is missing.')
3. IMP	*Malčik risoval zvezdu.* Boy **IMP**.drew star.ACC ('The boy was drawing a/the star.')

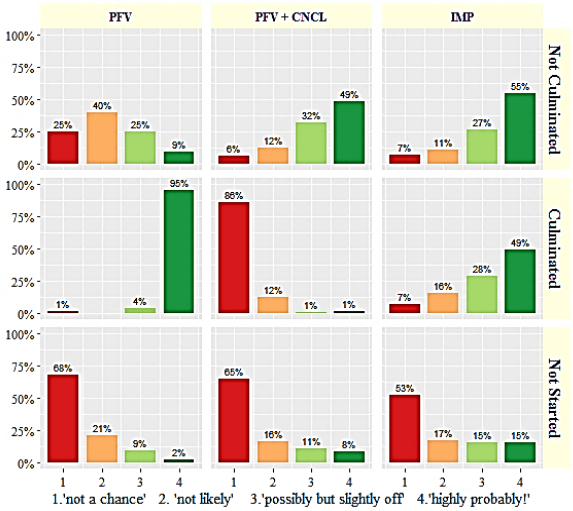
The visual stimuli were short animated video clips, depicting a human character performing the action denoted by the 8 accomplishments. In the 8 test items, the event was shown as ceasing short before reaching culmination, as illustrated for 'draw a/the star'. In the control items, 5 of the videos depicted culminated events and 3 portrayed scenarios where the event denoted by the predicate doesn't even begin. The visual stimuli were presented in one pseudo-randomized order across participants, while the verbal

stimuli were fully randomized for each clip and for each participant. The experiment was conducted using Qualtrics.

33 native Russian adults were instructed to determine how likely it is for a Russian speaker to use each of the five accompanying sentences upon watching the clip. Participants noted their judgments on a 4-point forced-choice scale, with the following labels: 1=*ni maleišego šansa* ('not a chance'); 2=*vrjad li* ('not likely'), 3=*vozmožno, xotja čto-to ne tak* ('possible though slightly off'); 4= *vpolne verojatno!*'('highly probably').

Results & analysis: As can be seen in Figure 1, non-culminating PFV items were scored as 3-4 34% of the time, and items in the PFV+CNCL frame received ratings of 3-4 81% of the time, with 4-scores as high as 49%. This latter finding is particularly surprising given the assumed degradation introduced by the supposed mismatch within the verbal stimuli, as illustrated by (2). Finally, IMP items received rating of 3-4 82%, as expected. An analysis of the non-culminated items using a Friedman's Chi-Square revealed a main effect of aspectual frame (p <0.001). This effect, though, was entirely due to the distribution of the PFV, as confirmed by a Wilcoxon Signed-Rank test showing no significant difference between the PFV+CNCL and the IMP (p=0.470).

Fig.1: Acceptability scores by aspectual frame and culmination



Discussion Our study demonstrates that while Russian PFV telic accomplishments do carry culmination requirements, these inferences are not stricter in Russian than what has been reported for other languages (e.g., Arunachalam & Kothari 2011). Hence, the Russian PFV is not exceptional in terms of the culmination requirements it imposes on telic accomplishments. Moreover, our data reveal that even in Russian, PFV telic accomplishments may in fact be followed by a cancellation phrase without creating a contradiction. We argue that what's being cancelled here is not the culmination inference *per se*, but rather, the **maximal** interpretation of Culmination (cf. Martin 2019, Martin & Demirdache 2020). Our data are in line with Kearns' (2007) proposal that PFV accomplishments only entail the **standard telos**: the onset of a specified endstate; and further, that while the standard telos is part of the **semantics** of PFV telic accomplishments, the maximal telos is only **implicated** by such predicates, and may therefore be cancelled. Crucially, the events depicted in the visual stimuli did not end at some early, arbitrary point; they were all completed up to approximately 80%. This suggests that this range (between approx. 80% completion and 100%) may reflect the margin between Kearns' standard telos and her maximal telos. And further, that any point within this range qualifies as Culmination. Such an approach recognizes the critical role of pragmatics in licensing the maximal interpretation of PFV telic accomplishments, while not abandoning Vendler's original claim that culmination is an integral part of the semantic denotation of accomplishments.

References: Arunachalam, S. & Kothari, A. (2011). An experimental study of Hindi and English perfective interpretation. *Journal of South Asian Linguistics*, *4*(1), 27-42. **Filip, H. (2017)**. The semantics of perfectivity. *Italian journal of linguistics, 29(1)*, 167-200. **Kearns, K. (2007).** Telic senses of deadjectival verbs. *Lingua*, *117*(1), 26-66. **Krifka, M. (1989).** Nominal reference, temporal constitution and quantification in event semantics. *Semantics and contextual expression*, *75*, 115. **Martin, F. (2019).** Non-culminating accomplishments. *Language and Linguistics Compass*, 13(8). **Martin, F. & Demirdache, H. (2020).** Partitive accomplishments across languages. *Linguistics*, *58*(5), 1195-1232. **Parsons, T. (1990)**. *Events in the semantics of English: A study in subatomic semantics*. Cambridge: MIT Press.

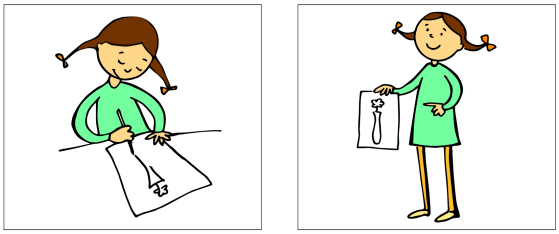
Aspect Processing Across Languages
Visual World Eye Tracking Evidence for Semantic Distinctions

Sergey Minor, Natalia Mitrofanova, Gustavo Guajardo, Myrte Vos and Gillian Ramchand
UiT The Arctic University of Norway

1. Introduction. The Visual World paradigm (VWP) has been a richly productive methodology in the area of linguistic processing, ever since the seminal study of Tanenhaus et al. (1995). The usefulness of the paradigm stems from the general fact that human eye movements or saccades track the focus of linguistic attention, if that attention is given a visual manifestation (Huettig 2015). In this study, we apply the VWP to experimentally probe into the semantic representation of aspectual categories (perfective vs imperfective) across three languages: Russian, Spanish and English. We show that this methodology can reveal subtle differences in processing, reflecting the different meaning and morphosyntactic encoding of aspectual categories in these languages.

2. Processing of Aspect. Previous offline studies have provided evidence that imperfective aspect focuses on the in-progress, activity stage of an event, while perfective aspect triggers a representation of the event as a completed whole, highlighting the final stage and/or the result (goal) state of the event (Madden & Zwann 2003; Ferretti et al. 2007 *a.o.*). The VWP is well-suited to investigate this contrast by employing a visual set-up that counterposes two pictorial event representations which focus on different temporal portions of the depicted event— a snapshot of the ongoing event (OE), and a snapshot of the completed event (CE), i.e. the immediate aftermath of the event. Linguistic cues have be shown to drive anticipatory visual attention (Altmann and Kamide 2007), and aspectual information coded by grammatical morphemes have been shown recently in a number of eye tracking VWP studies to facilitate event recognition (Zhou et al. 2014 for Mandarin and Foppolo et al 2021 for Italian). In each of these two latter cases, the perfective morpheme or functional element triggers preferential looks towards the completed picture, thus corroborating the general semantic judgements that in the context of telic verbs, perfectivity generates a culmination entailment by default. *The present study is the first attempt to explicitly compare typologically different aspectual systems using the same task while eye-tracking.*

3. The Experiment. Each experimental trial included an audio preamble which located the narrative in the past (e.g. *It was a rainy day*), followed by a sentence-picture matching task where the participants were presented with a pair of pictures: one representing an action in progress (OE, Fig. 1a), and one representing the result that obtained after the action was completed (CE, Fig. 1b). While looking at the pictures, the participants heard a sentence in the past tense (e.g. *A girl was drawing/drew a slender vase*). In all the investigated languages we manipulated the aspect of verb in the target sentence (Imperfective vs Perfective verbs in Russian; Imperfect vs Preterite verb forms in Spanish; Past Progressive vs Simple Past verb forms in English). The participants were asked to choose the picture that best corresponded to the sentence. Each experiment included 24 test trials. In all cases, we used accomplishment predicates and reused the picture stimuli between the languages as far as possible. The participants' eye-movements and offline responses were recorded. Overall, we tested 124 Russian speakers, 66 English speakers, and 32 speakers of Argentinian Spanish.



(a) Ongoing event (b) Completed event

Figure 1: 'A girl drawing a vase'

4. Cross-linguistic Differences and Predictions. In a narrative context, all three languages use the imperfective forms for event overlap, conveying the notion of an event that is in progress at a given time interval or reference point (Klein 1994). We thus expected imperfective forms in all three languages to draw attention to the activity stage of telic events (OE pictures). Conversely, all

the perfective verb forms we tested are used to convey sequencing of events in a narrative. There are however subtle differences in the meaning of the perfective forms among the three languages, which we predicted could lead to contrasting results. In the Russian experiment, the aspectual manipulation involved perfective/imperfective aspectual pairs (*risovat'* 'draw^{IMP}' - *na-risovat'* 'draw^{PFV}'; cf. Forsyth 1970, Zalizniak & Šmelev 2000). The Perfective verb in such pairs entails that the event reached an *inherent boundary*, i.e. a lexically specified result state or the maximal possible extension of the event (Klein 1995, Filip 2008, Tatevosov 2013). In Spanish, verbs in the preterite also entail the existence of an event boundary (Fábregas 2015). However, in contrast to the Russian perfective verbs, they do not require an *inherent* boundary (e.g. the attainment of a result state), as evidenced from the fact that adjuncts introducing temporal boundaries license the use of the preterite in Spanish but not perfective in Russian (Janda & Fábregas 2019). Finally, the case of English is especially interesting. Given its role in event sequencing, the Simple Past form of non-stative predicates has been analyzed as a kind of perfective which presents the event as a completed whole (Smith 1991, Wurmbrand 2014, Martin 2019, cf. de Swart 1998 for a dissenting view).Given the existence of non-culminating contexts in English, however, it is unclear how strongly the English past generates completive entailments in practice (cf. Martin 2020).

5. Results. In all three languages, the results revealed at-ceiling preference for the OE picture in the imperfective condition both in the offline task (picture selection; 98% of the trials in Russian, 97% in Spanish, 95% in English) and the online gaze patterns. In the perfective condition, we found robust differences: In Russian, the choice of the result state (CE) picture in the offline task was once again at ceiling (95%); for Spanish it was high, but not quite at ceiling (83%); in English there was no statistical preference for the OE picture in the Simple Past condition (54%, not significantly different from chance, $p = 0.39$). The analysis of the participants' online gaze patterns yielded parallel results (Fig 2). These results confirm our prediction that the imperfective forms in all the three languages draw attention to the in-progress representation of the event. With respect to the perfective forms, our results suggest that perfective accomplishment verbs in Russian strongly highlight the result state of the event. In Spanish, the preterite also highlights event completion, but to a lesser extent than in Russian, in line with its less restrictive semantics in not requiring an inherent boundary.

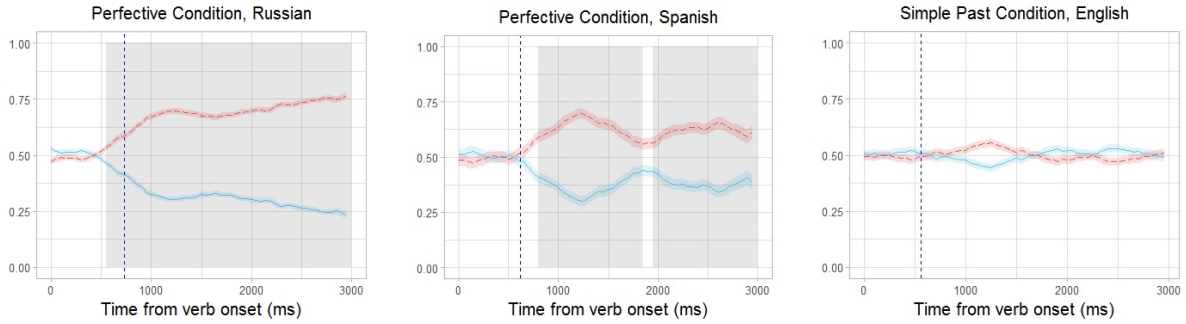


Figure 2: Proportion of looks to the OE (solid blue line) and CE (dashed red line) pictures in the perfective condition. Shading represents the time windows where the probability of looks to the CE picture was significantly above chance. The dashed vertical blue lines mark the average verb offsets.

Our results for the English Simple Past condition are striking. They suggest that even on telic predicates, the Simple Past form does not encode a preferential cognitive salience for either the activity portion of an event *or* its result state. Our result points to a dissociation between the role of verbal categories in encoding narrative sequencing, vs. highlighting particular portions of a complex event structure. These facts do not emerge cleanly if one relies solely on offline judgements of entailment in context, thus highlighting the role and value of online experiments of this type.

References

Altmann, G. T. and Y. Kamide (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language 57*, 502–518.

de Swart, H. (1998). Aspect shift and coercion. *Natural Language and Linguistic Theory 16*, 347–385.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research 1626*, 118–135.

Martin, F. (2019). Non culminating accomplishments. *Language and Linguistics Compass*.

Tanenhaus, M. K., M. J. Soivey-Knowlton, K. M. Eberhard, and J. C. Sedivy (1995). Integration of visual and linguistic information in spoken language comprehension. *Science 268(5217)*, 1632–1634.

Wurmbrand, S. (2014). Tense and aspect in english infinitives. *Linguistic Inquiry 45(3)*, 403–447.

Zhou, P., S. Crain, and L. Zhan (2014). Grammatical aspect and event recognition in children's online sentence comprehension. *Cognition 133(1)*, 262–276.

Negative islands do not block active gap filling

Zirui Huang, Matthew Husband
University of Oxford

While constraints on long distance dependencies are often syntactic in nature, they may also arise from semantic considerations. Negative islands, a type of weak island, selectively constrain certain wh-dependencies that violate Dayal's (1996) maximal informativity presupposition on questions, i.e., that the answer set contains a true answer entailing all the other true ones (Fox & Hackl, 2007; Abrusán, 2011). Negative degree questions like **How tall isn't John?* are judged to be unacceptable because they ask for the minimal height interval that does not contain John's height, even though such an interval does not exist because the true answer set contains two mutually exclusive subsets that do not entail one another, i.e. all intervals below John's height, (0, $\text{height}_{\text{John}}$), and all intervals above John's height, ($\text{height}_{\text{John}}$, ∞).

In general, island constraints have been found to constrain long distance dependency formation in real time. Stowe (1986) showed that comprehenders actively posits gaps for wh-phrases in grammatical positions, demonstrating that filled-gap effects emerged when gaps are grammatically licensed but not when they are grammatically inaccessible, e.g., inside subject islands. Further research has found that comprehenders respects strong wh-island constraints (Traxler & Pickering, 1996; Wagers & Phillips, 2009), reflecting the parser's rapid use of syntactic constraints to avoid positing illicit dependencies in real-time.

Whether comprehenders can use semantic constraints, such as negative islands, in real-time is unclear. Compared to syntactic constraints, it may take comprehenders more time to use presupposition violations to block dependency formation, as their calculation may be more complex. We examined whether negative islands are as effective as wh-islands at blocking illicit gaps in real-time. If comprehenders respect presuppositional dependency constraints, then we expect negative islands (2b) to be as effective as wh-islands (2c) in blocking a filled-gap effect (at *famous*). However, if comprehenders are unable to rapidly use presuppositional constraints to prevent illicit gaps, then we expect to see a filled-gap effect for negative islands, but not for wh-islands. Experiment 1 examined offline acceptability of negative islands with (un)reduced relative clauses, setting up Experiment 2 to use online filled-gap effects to investigate whether comprehenders posit illicit gaps inside negative islands compared to wh-islands.

Experiment 1 acceptability judgements. (N=51, Items=24) We manipulated POLARITY (Positive, Negative) and STRUCTURE (No, Reduced, Unreduced RCs), shown in (1). Results are shown in Figure 1/Table 1. While the presence of negation reduced acceptability overall (*Est.*=0.37, *t*=5.16), there was a significant interaction with structure (*Est.*=0.45, *t*=4.23). NoRC sentences (corresponding incrementally to a potential temporary gap in RRCs) were rated much lower when negation was present, compared to difference in R/URCs (Table 2), suggesting that participants use negative island constraints offline.

Experiment 2 self-paced reading. (N=63, Items=24) We manipulated ISLAND type (No-, Neg-, Wh-Island) in (3) to examine whether comprehenders actively posit a (temporary) gap inside islands. A filled-gap effect emerged in the first spillover region between No-Island and Wh-Island conditions (*Est.*=46.0, *t*=3.12, *p*=.007), showing that Wh-Islands blocked dependency formation relative to No-Islands, but no significant difference was found between No-Islands and Neg-Islands (*Est.*=28.8, *t*=1.87, *p*=.157).

Discussion. Although comprehenders are aware of negative islands offline, online results showed that they were unable to use them to block active dependency formation. This asymmetry suggests that the effects of weak (semantic) islands take time to emerge, unlike strong (syntactic) islands which are more immediate.

(1) Example item in Experiment 1 acceptability judgments

a. How tall did Mary think the girl hoped to be? (NoRC, Positive)
b. How tall did Mary think the girl hoped **not** to be? (NoRC, Negative)
c. How tall did Mary think the girl hoped to be to be famous by her parents was? (RRC, Positive)
d. How tall did Mary think the girl hoped **not** to be to be famous by her parents was? (RRC, Negative)
e. How tall did Mary think the girl who was hoped to be to be famous by her parents was? (URC, Positive)
f. How tall did Mary think the girl who was hoped **not** to be to be famous by her parents was? (URC, Negative)

Table 1: Model summary for Experiment 1

	Est.	***t***	***p***
Polarity	0.37	5.16	<.001
NoRC v RCs	0.86	8.43	<.001
RRC v URC	0.24	2.70	.007
Polarity:NoRC v RCs	0.45	4.23	<.001
Polarity:RRC v URC	-0.05	-0.59	.556

Table 2: Effect of polarity within sentence structures for Experiment 1

Positive – Negative	***Est.***	***t***	***p***
No RC	1.48	10.56	<.001
Reduced RC	0.27	1.94	.053
Unreduced RC	0.38	2.46	.014

(2) Example item in Experiment 2 self-paced reading

a. How tall did Mary think the girl hoped to be **famous** by her parents was before she went to college? (No-Island)
b. How tall did Mary think the girl hoped not to be **famous** by her parents was before she went to college? (Neg-Island)
c. How tall did Mary think the girl who was hoped to be **famous** by her parents was before she went to college? (Wh-Island)

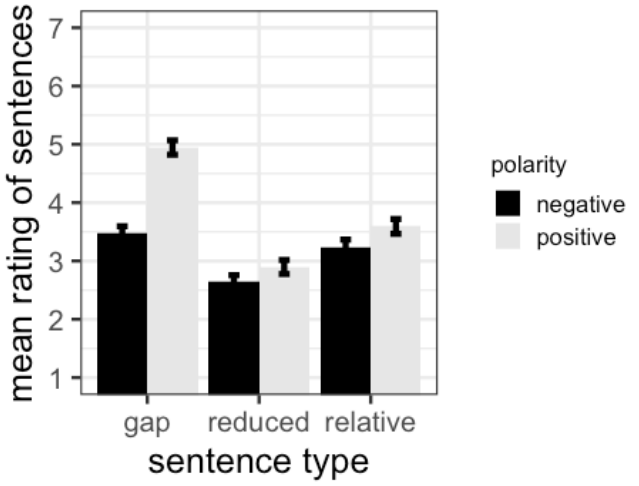


Figure 1. Acceptability task

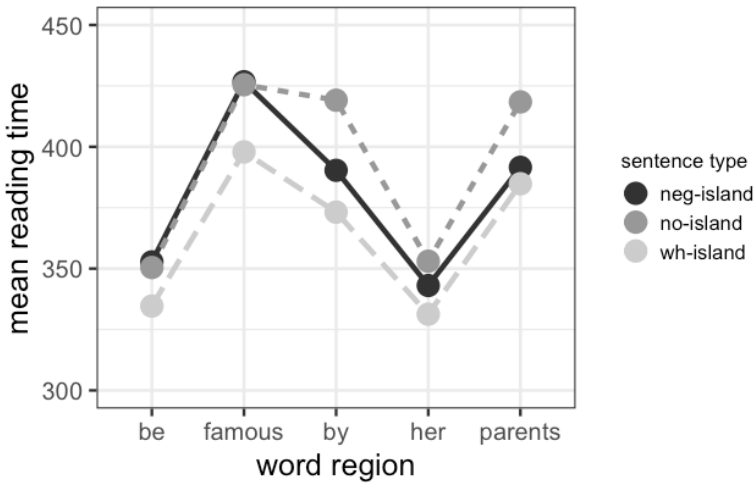


Figure 2.Self-paced reading task

Selected Reference:

Abrusán, M. (2011). Presuppositional and negative islands: A semantic account. *Natural Language Semantics, 19*(3), 257-321.

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, *1*(3), 227-245.

Less than a Sentence is not Enough – An Eyetracking Study on the Incremental Interpretation of Negative Expressions

Fabian Schlotterbeck (University of Tübingen) & Oliver Bott (Bielefeld University)

Online studies of quantifier and negation processing suggest that not all aspects of semantic operators are interpreted immediately. A number of previous studies concluded that downward entailing (DE) quantifiers such as *less than half* lead to severe processing delays as compared to upward entailing (UE) ones, such as *more than half* (e.g. [1]), as does the interpretation of negation (e.g. [2]). Perhaps unsurprisingly, then, the non-incremental interpretation of scopal operators seems to extend to multiply quantified sentences, as suggested by an eyetracking during reading study on the relative scope of quantifiers by [3] with scope interpretation delayed until the end of the sentence.

The purported violation of incrementality has not gone unchallenged, though. In particular, it has been shown that pragmatic factors such as world knowledge and discourse context bear important influences on the time course of scope interpretation (e.g. [4,5,6]). For negation, [7,8] have proposed the Dynamic Pragmatic Account based on *Questions under Discussion* (QUDs, [9]) essentially claiming that the delay is caused by the need to accommodate an appropriate QUD. The present eyetracking during reading study thus investigated the time course of comprehending sentences with quantifiers and negation (see ex. (1)-(5)) with and without discourse context.

Eyetracking Study: Exp. 1 ($N = 48$) established clear complexity differences with overadditive effects of operators, though was not intended to address incrementality yet. Participants read sentences containing UE vs. DE quantifiers in initial position and negated vs. positive predicates (e.g. *not blue* vs. *blue*; see ex. (1)) in the sentence-final region of interest (ROI). Linear scope was fixed because negation appeared in a scope island. The final ROI contained the negation, provided the second semantic argument of the quantifier and completed the sentence. It was this ROI where our manipulation of semantic complexity showed the expected interaction between operators: Regression path durations (RPDs) were longer for DE than UE quantifiers, with a bigger difference in negated conditions than in the positive control condition (Fig. 1 a; all reported effects were significant in (G)LMER analyses). Exp. 2 ($N = 40$) employed these complexity differences as indication of compositional interpretation during reading more natural quantifier-negation sentences out of discourse context. To test for the influence of event information encoded in lexical verbs [cf. 3], the position of the main verb was another factor manipulated (cf. ex. (2) vs. (3)). A pretest established an overwhelming surface-scope preference for the experimental items. Delayed semantic interpretation may be expected in (3) because here the event information of the main verb was presented several words after the negation. In (2), we considered incremental effects likely, hoewever, since the negation was presented simultaneously with the main verb and, as in Exp. 1, completed a minimal sentence [10]. Contrary to this Verb-Dependent Incrementality assumption, complexity effects were delayed to the final ROI, irrespective of verb position, as in [3] (Fig. 1 c). Exp 3. ($N = 48$) embedded clefted versions of the same sentences (ex. (5)) in discourse contexts that introduced positive and negative properties (e.g. *to play* or *not to play*) establishing the QUD "how many" individuals have or lack the property in question. Based on the literature [7,8], we expected incremental effects in such sentences with contextually licensed negation. Contextual embedding led to earlier and sustained effects of negation (Fig. 1 b), but monotonicity of the quantifier still only affected the final ROI of the relative clause and none of the earlier ROIs.

Conclusions: The different time course observed for Exps. 2 and 3 resulting from the contextual establishment of the QUD shows that discourse pragmatics is an important prerequisite for the realtime interpretation of scope. However, finding an effect of monotonicity still only at the end of the clause indicates that multi-operator interpretation proceeds in an essentially non-incremental

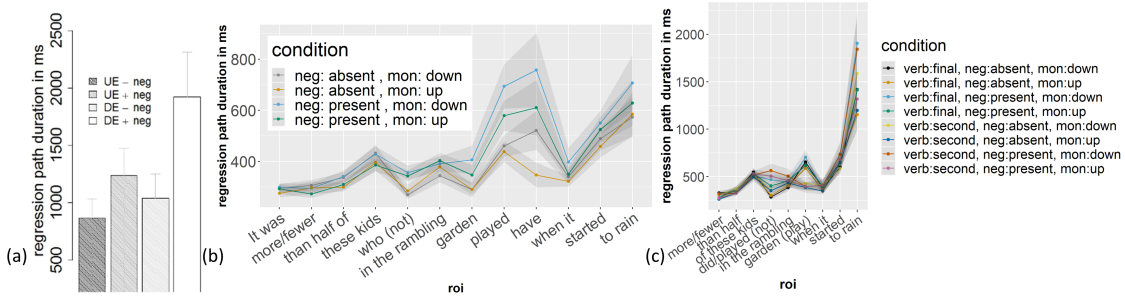


Figure 1: Regression path durations (+95% confidence intervals), i.e. the time of all fixations summed up from first entering a ROI until it is left to the right (a: sentence final ROI in Exp. 1; b: all ROIs in Exp. 3; c: all ROIs in Exp. 2).

way. We consider this a highly interesting finding because the verbal information was information already given in the discourse context. The processing of negative operators thus depends on a larger domain than just the operators themselves such as a complete minimal sentence.

Sample Item Experiment 1 (picture verification task not shown here, results fully consistent with complexity results of the reading stage):

(1) Auf/For {mehr/weniger/more/fewer} als/than die Hälfte/the half der Quadrate/of squares |trifft zu/it's true, |dass sie/that they |(nicht)/(not) blau sind./blue are

Sample Item Experiment 2:

(2) {Mehr/Weniger/More/Fewer} als/than |die Hälfte/half |dieser Kinder/of these kids |spielten (nicht)/played (not) |im weitläufigen/in the rambling |Garten,/garden, |als es/when it |anfing/started |zu regnen./to rain

(3) {Mehr/Weniger/More/Fewer} als/than die Hälfte/half |dieser Kinder/of these kids |haben (nicht)/have (not) |im weitläufigen/in the rambling |Garten gespielt,/garden played, |als .../when ...

Sample Item Experiment 3:

(4) Preceding Context: Ida's parents invited the kids from the neighborhood to her birthday party. After lunch they all played in the garden. When it started to rain, Ida's parents decided to open up the living room for the kids. Some of the kids didn't want to play in the garden anymore whereas others stayed outside and played in the rain.

(5) Es waren/It was {mehr/weniger/more/fewer} |als/than die Hälfte/half |dieser Kinder/of these kids, |die (nicht)/who (not) |im weitläufigen/in the rambling |Garten/garden |gespielt/played |haben/have, |als es/when it |anfing/started |zu regnen./to rain

References: [1] Urbach & Kutas (2010), *JML* 2 (63). • [2] Kaup et al. (2007), in Schmalhofer & Perfetti (eds.): *How is negated information represented?* • [3] Bott & Schlotterbeck (2015), JoS 32. • [4] Nieuwland (2016), *J. Exp. Psychol. Learn.Mem. Cogn.* • [5] Freunberger & Nieuwland (2016), *Brain Research* 1646. • [6] Nieuwland & Kuperberg (2008), *Psychol. Sci.* 19. • [7] Tian & Breheny (2010), *Psychol. Sci.* 19. • [8] Tian & Breheny (2015), in Larivée & Lee (eds.): *Negation & Polarity* • [9] Roberts (1996/2012), *Sem. & Prag.* 5.• [10] Radó & Bott (2012), *Proc. of Amst. Colloq. 18*

Beyond the sentence: Discourse structural effects on reference resolution

Petra Schumacher

Much of the psycho- and neurolinguistic research on reference to date has used short texts consisting of a context sentence and a target sentence as the object of study. This allows for careful control of contextual factors (word order, givenness, agentivity, number of referents, etc.), which is common practice in laboratory experiments. However, the use of minimal context comes at the expense of the naturalness of information transfer and it has various negative implications for theory building: (i) contexts with only two potential referents restrict hypothesis testing; (ii) higher-level discourse structural factors such as discourse topicality or perspective cannot be taken into account; (iii) reference as a phenomenon of common ground management between speakers and addressees is reduced to an artificial communication situation (the lab experiment).

This talk will reflect upon the limitations of many previous studies on reference resolution (my own included) and present research that (gradually) moves away from mini-texts towards more naturalistic contexts involving more elaborate referential spaces. In particular, it will discuss the role of discourse topicality and perspective taking (via evaluation) on the resolution of personal and demonstrative pronouns in German.

Ignorance and Exclusivity in Semi-Cooperative Contexts

Background In ordinary conversations, disjunctive sentences like (1) give rise to EXCLUSIVITY and IGNORANCE inferences. Disjunctive sentences embedded under a negative factive, like (2), have been argued to give rise to parallel inferences at the presupposition level (Marty & Romoli 2021, Spector & Sudo 2017, a.o.). In semi-cooperative contexts, however, IGNORANCE inferences normally drawn from such sentences are cancelled. Thus, in the context of a game show, (1) can be felicitously uttered by a host who is known to know in which boxes there is money. These contexts have recently been discussed as a challenge for the pragmatic view (Fox 2014, Agyemang 2020). On this view, the cancellation of IGNORANCE is explained if the Maxim of QUANTITY is deactivated. This predicts in turn that EXCLUSIVITY, which also depends on QUANTITY, should be also cancelled. Fox (2014), however, argues that EXCLUSIVITY is still derived: intuitively, (1) would be a misleading hint if it turns out there is money in both boxes, unlike a variant of (1), where EXCLUSIVITY is blocked by the addition of 'or both'. Agyemang (2020) offers experimental data in support of this intuition. Testing Fox's game scenarios in a forced-choice task, Agyemang found that, compared to the 'or both'-variant, people were significantly less likely to pick one of the two boxes after hearing (1) when the contestant before them picked one of them and won money (78% vs. 61%). This suggests that (1) can still give rise to EXCLUSIVITY in contexts where QUANTITY is deactivated. We argue, however, that these findings can receive another explanation: EXCLUSIVITY may follow from general assumptions as to how games work. Specifically, hearers may assume that, in order to increase the interest of the game, the game actions most favored by a hint (e.g., choosing box 20/box 25) must not all lead to a winning outcome. This explanation would account for the contrast between (1), where this assumption leads to EXCLUSIVITY, and its variant where this assumption is blocked by 'or both'. Next, we note that the questions of the existence and source of EXCLUSIVITY in semi-cooperative contexts carry over to presuppositional cases, where pragmatic approaches predicts PRESUPPOSED IGNORANCE and PRESUPPOSED EXCLUSIVITY to go together. Thus, in contexts where PRESUPPOSED IGNORANCE is cancelled, do sentences like (2) still give rise to PRESUPPOSED EXCLUSIVITY? And if so, does this inference arise through scalar reasoning or follow from game-related assumptions?

Experiments. We report on two experiments, building on Agyemang's study, inquiring into the source of EXCLUSIVITY in game scenarios and extending this research to presuppositional cases. Exp.1 adds to Agyemang's MONEY conditions (Table 1, A) novel control conditions testing whether the contrasts between OR and OR-BOTH reproduce in set-ups where choosing the alternative-box (e.g., box 25) is *strongly discouraged* by the game rules. In these conditions (Table 1, B), contestants received hints about which boxes are associated with slime: if they picked a wrong box, they were slimed and left the game. If EXCLUSIVITY remains available in these cases, participants should nonetheless prefer the 'alternative-box' option (e.g., box 25) after getting OR than OR-BOTH hints when the contestant before them picked one of the two boxes and got slimed. Exp.2 tested the presuppositional variants of the OR and OR-BOTH hints from Exp.1 in both the MONEY (Table 1, C) and the SLIME conditions (Table 1, D). Hint and Game type were manipulated between subjects, and Previous outcome (whether the previous contestant WON vs. LOST) was manipulated within subjects.

Main results. Results from Exp.1 ($n = 200$) replicate Agyemang's results (MONEY conditions) and show that the target contrasts reproduce in SLIME conditions: people were far more likely to choose the alternative-box after hearing OR than OR-BOTH when the previous contestant picked the other box and lost (74% vs. 27%), despite the strong incentive to choose any other box in these cases. Results from Exp.2 ($n = 200$) are entirely parallel to those from Exp.1: in the MONEY conditions, people strongly preferred to choose any other box upon hearing OR when the previous contestant picked one of the two boxes and found money (27% vs. 70%) whereas, in the SLIME conditions, they strongly preferred to choose the alternative-box in the same critical conditions (60% vs. 20%).

Discussion Our studies make two contributions. First, our results replicate Agyemang's (2020) findings and, consequently, confirm Fox's original judgments while ruling out an independent account explaining EXCLUSIVITY in terms of game-related assumptions. Second, our results extend these findings to presuppositional cases like (2) where similar inference types have been identified, raising a challenge similar to Fox's (2014) original challenge for recent proposals extending the pragmatic approach from the assertion to the presupposition level. We will discuss potential responses from the perspective of the pragmatic approach for both the assertive and presuppositional cases.

(1) There is money in box 20 or 25.
   a. EXCLUSIVITY: *There isn't money in both box 20 and 25*
   b. IGNORANCE: *The speaker doesn't know whether there is money in box 20 and doesn't know whether there is money in box 25*

(2) Previous contestants were unaware that there is money in box 20 or 25.
   a. PRESUPPOSED EXCLUSIVITY: *There isn't money in both box 20 and 25*
   b. PRESUPPOSED IGNORANCE: *The speaker doesn't know whether there is money in box 20 and doesn't know whether there is money in box 25*

A. Example ASSERTIVE items in MONEY conditions (Exp.1, replication of Agyemang's)
There are 100 numbered boxes in total, and 5 of them contain a million dollar prize. The host tells the first contestant that there is money in {box 20 or 25 (OR) / box 20 or 25, or both (OR-BOTH)}. The contestant picks box 20 and {finds a million dollars (WON) / does not win any money (LOST)}.

B. Example ASSERTIVE items in SLIME conditions (Exp.1, novel)
There are 100 numbered boxes in total, and 5 of them are associated with slime. The host warns the first celebrity that slime is associated with {box 20 or 25 (OR) / box 20 or 25, or both (OR-BOTH)}. The celebrity picks box 20 and {nothing happens (WON) / is slimed (LOST)}.

C. Example PRESUPPOSITIONAL items in MONEY conditions (Exp.2, novel)
There are 100 numbered boxes in total, and 5 of them contain a million dollar prize. The host tells the remaining players that previous contestants were unaware that there is money in {box 20 or 25 (OR) / box 20 or 25, or both (OR-BOTH)}. The contestant picks box 20 and {finds a million dollars (WON) / does not win any money (LOST)}.

D. Example PRESUPPOSITIONAL items in SLIME conditions (Exp.2, novel)
There are 100 numbered boxes in total, and 5 of them are associated with slime. The host warns the remaining celebrities that previous contestants were unaware that slime is associated with {box 20 or 25 (OR) / box 20 or 25, or both (OR-BOTH)}. The celebrity picks box 20 and {nothing happens (WON) / is slimed (LOST)}.

Imagine you are the next player in this game. The host does not give you any more information. Which action are you most likely to take? | *Choose box 25* | *Choose any other box*

Table 1: Example items illustrating the experimental conditions in Exp.1 (A,B) and Exp.2 (C,D). Participants chose one of two options: the alternative-box (e.g., box 25) or any other box.

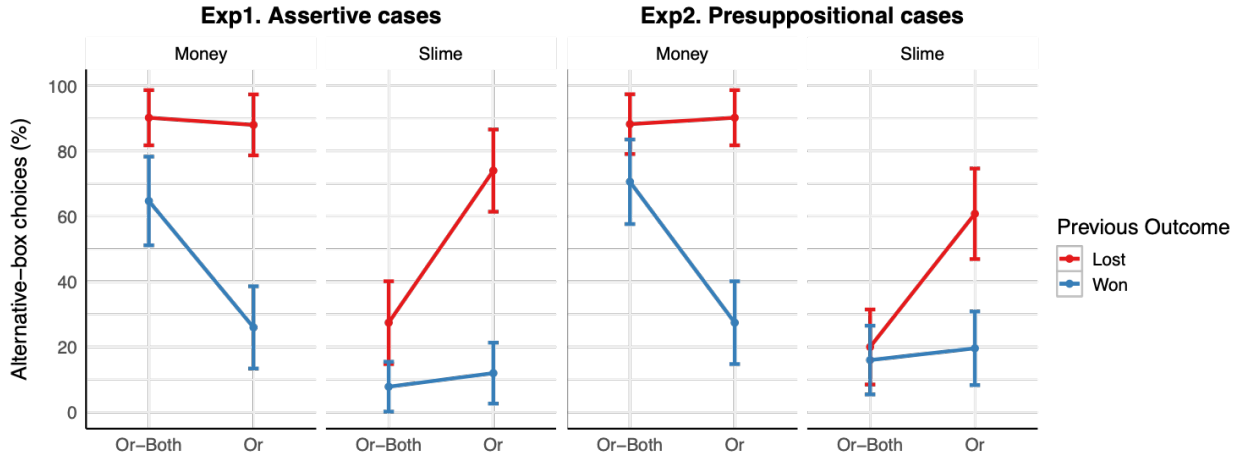


Figure 1: Proportion of 'alternative-box' choices (e.g., 'choose box 25') by Hint type, Game type and Previous outcome in Exp.1 and Exp.2. Error bars denote 95% confidence intervals.

References Agyemang, C. 2020. Scalar implicatures under uncertainty • Chemla, E. 2008. An Epistemic Step for Anti-Presuppositions • Fox, D. 2014. Cancelling the Maxim of Quantity: Another challenge for a Gricean theory of Scalar Implicatures • Marty, P. and Romoli, J. 2021. Presupposed free choice and the theory of scalar implicatures • Spector, B. and Sudo, Y. 2017. Presupposed ignorance and exhaustification: how scalar implicatures and presuppositions interact

You must worry! The interpretation of *mustn't* varies with context and verb complement.

Adina Camelia Bleotu[1,2], Anton Benz[1] & Roxana Mihaela Pătrunjel[2]
[1]The Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), [2]University of Bucharest

The current paper investigates experimentally whether the interpretation of deontic *mustn't* in American English varies with pragmatic context (*lack of necessity/ necessity not to*) and the semantic properties of the modal complement (negative mental activity/physical event).

Background English modals display an irregular behaviour in interaction with negation. While negation has a fixed position in English (i.e., always after a modal), its interpretation is variable (i.e., the negation may scope below/under modality). This is true for modals of a different quantificational force (universal/necessity vs. existential/possibility- see 1a, b), as well as different flavors of the same modal (deontic & epistemic-1b, c) [2].

(1) a. The boy must not/mustn't go to the party. (NECESSARY> NOT)
b. The girl cannot/can't play in the park this evening. (NOT> POSSIBILE)
c. The girl may not be doing her homework. (POSSIBILE> NOT)

Various attempts at generalizations have been put forth, either in terms of the possibility/necessity distinction (e.g., [5]), or in terms of the deontic/ epistemic modality distinction (e.g., [1]), but, as pointed by these authors themselves, there are always exceptions to these generalizations. Moreover, it is unclear what *n't* and *not* represent from a syntactic point of view (sentence or adverbial negation). In the ideal situation, sentence negation translates as external negation (NEG> MODALITY), and adverbial negation translates as internal negation (MODALITY> NEG). However, we find cases where what looks like sentence negation (*n't*) expresses internal negation (see (1a)). While deontic *must not* and *mustn't* are generally argued to express interdiction ([1], [2], [3], [4], [5]), deontic necessity scopes below negation in special polarity-sensitive contexts like contrastive negation-see (2) ([6], [7], [8], [9]).

(2) No student MUST read 5 articles on the topic but one student is encouraged to do so.

We draw attention to some not (obviously) polarity-dependent situations, involving *lack of necessity* contexts and negative mental activities (3), which are also interpreted as *not necessary.*

(3) You mustn't worry. It's just your usual jokester holiday!/ You mustn't feel bad if what you try to do doesn't work. //You mustn't panic.

Current experiment (N = 34 native AE speakers) We tested the intuition that *lack of necessity* contexts and negative mental activities bias the interpretation of *mustn't* towards *not necessary*.

Procedure Our experiment combined a forced choice task with a gradient acceptability task. Participants read sentences in context and had to choose the most suitable interpretation of *mustn't* (either necessity not to or lack of necessity contexts-Table 1). They then had to rate the acceptability of the sentence in context on a Likert scale from 1 to 7.

Materials Participants were presented with 8 critical sentences and 16 fillers (with *needn't* and *shouldn't*). The critical sentences used different verb types (mental/physical) in different pragmatic contexts (*lack of necessity*/*necessity not to*)-Table 1. Participants saw verbs in only one context. For each verb type, 4 verbs were tested: *worry, panic, be sad, be upset* (mental), and *eat, drink, do, speak* (physical). Half of the sentences with *mustn't* had 2nd person subjects and half 3rd person subjects.

Results While 15 participants (Interdiction group) gave mostly *necessary not to* readings, the rest produced more *necessary not* readings in *lack of necessity* contexts and with mental verbs (Figures 1, 2). Importantly, *mustn't* was rated as very acceptable in both *necessity not to* (5.84) and *lack of necessity* contexts (5.53). We computed a linear regression with Verb Type (Mental/Physical) and Context (*Necessity not to/Lack of necessity*) and their interaction as fixed effects and random slopes per Item and Participant. The results show significance for Verb Type ($p < .05$), Context ($p < .01$), and the Verb Type-Context interaction ($p < .05$) The parallel analysis

of expected response times for the forced choice shows significantly longer times for *not necessary* contexts. The person of sentential subjects did not affect the interpretation.

Table 1. Sample experimental items

Forced Choice Task
***Lack of necessity* context**: In *You mustn't worry. The woman will give you money*, the sentence *You mustn't worry* means: ***Necessity not to* context**: In *You mustn't worry. You will get sick otherwise*, the sentence *You mustn't worry* means:
a. It is necessary that you do not worry. b. It is not necessary that you worry.
Acceptability Judgment Task (on a Likert scale from 1 to 7)
***Lack of necessity* context**: How acceptable do you think the sentence *You mustn't worry* is in the context *You mustn't worry. The woman will give you money*? ***Necessity not to* context**: How acceptable do you think the sentence *You mustn't worry* is in the context *You mustn't worry. You will get sick otherwise*?

(Fully unacceptable)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	(Fully acceptable)

Figure 1. Proportion of *necessary not to* readings for the Interdiction Group

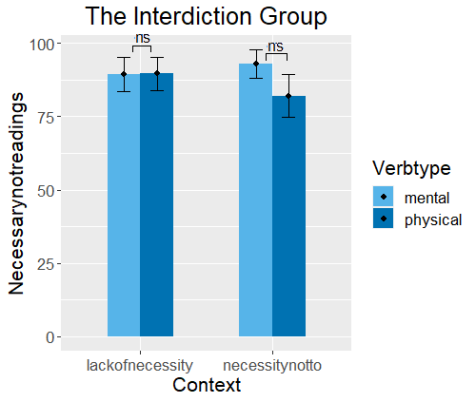
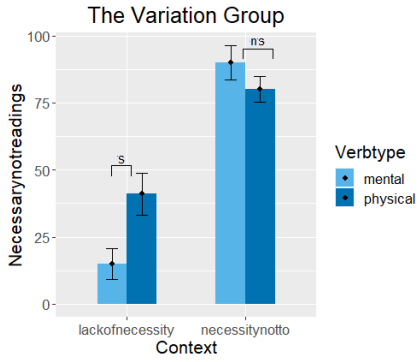


Figure 2. Proportion of *necessary not to* readings for the Variation Group



Discussion For some AE speakers, the scope between modality and negation in *mustn't* varies with pragmatic context and lexical verb. The importance of context has also been noticed in Romance, e.g., where *Negation + Obligation Verb* can contextually express either *not necessary* or *necessary not to*. Several proposals may capture this behavior: i) Neg Raising ([7], [8], [10], [11], deriving the strong reading from the basic order Neg> Modal via negative strengthening, ii) pragmatic weakening [12], [13], arguing the *not necessary* reading obtains as a suggestion from the basic strong *necessary not to*, (iii) ambiguity, arguing *mustn't* is ambiguous between two basic readings (strong/weak). While all accounts could be accommodated to capture context-sensitivity, a pragmatic weakening account starting from an *interdiction* (*necessary not to)* basic reading of *mustn't* is more in line with the high accuracy rates for *necessity not to* contexts compared to *lack of necessity* contexts, as well as *lack of necessity* readings being associated with longer RTs than *not necessary* readings. In addition to context, the type of verb the modal combines with also matters. Mental activities give rise to more *lack of necessity* readings than physical activities in *lack of necessity* contexts. Moreover, physical verbs give rise to more *necessary not* readings in *necessity not to* contexts than in *lack of necessity* contexts. We propose a cognitive account in terms of the difficulty of imposing one's will over another's (private) mental activities.

References: [1] Coates, J. (1983). *The Semantics of Modal Auxiliaries.* [2] Palmer, Frank R. 1990. *Modality and the English Modals.* [3] Papafragou, A. (2000). *Modality: Issues in the Semantics-Pragmatics Interface.* [4] Huddleston, R & Pullum G. (2002). *The Cambridge Grammar of the English Language.* [5] Cormack, A., N. Smith. (2002). Modals and Negation in English. In *Modality and its interaction with the verbal system.* [6] Israel, M. (1996). Polarity sensitivity as lexical semantics. *Linguistics and Philosophy* 19. [7] Homer, V. (2015). Neg-raising and positive polarity: The view from modals. *Semantics and Pragmatics* 8. [8] Iatridou, S. & Zeijlstra, H. (2013). Negation, polarity and deontic modals. *Linguistic Inquiry* 44.[9] Zeijlstra, H. (2017). Does NEG-Raising involve NEG-Raising? *Topoi.* [10] Hacquard, V. (2010). On the event relativity of modal auxiliaries. *Natural Language Semantics* 18. [11] Jeretič, P. (2021). Neg-raising Modals and Scaleless Implicatures. [12] Condoravdi, C. & Lauer, S. (2012). Imperatives: Meaning and Illocutionary Force. *Empirical Issues in Syntax and Semantics* 9. [13] von Fintel, K, & Iatridou, S. (2019). A modest proposal for the meaning of imperatives. *Modality across Syntactic Categories.*

What is the processing cost of (im)precision?

Camilo R. Ronderos, Ira Noveck, Ingrid Lossius Falkum

Semantically, a line is only *straight* when it has the maximal degree of 'straightness' (Kennedy, 2007; Syrett et al 2010; Aparicio, 2015, i.a). Thus, when such a *Maximum Standard Absolute Adjective* (MSAA*)* is used to express imprecision (e.g., 'almost straight'), it is assumed to require a threshold-oriented contextual adjustment (see Lasersohn, 1999; Leffel, 2016).

Two proposals regarding the relationship between (im)precise MSAAs and sentence processing have been put forth. Syrett et al., (2010) argue that imprecision, as a pragmatic adjustment, necessarily adds processing cost to sentence processing compared to precision. Aparicio et al. (2016) speculate that *precise* MSAAs are costlier to process than imprecise expressions because, in general, more contexts support imprecise interpretations. In the current study we test a third hypothesis, namely that processing cost will be mediated by contextual expectations of precision (see Van der Henst et al., 2002, Gibbs & Bryant, 2008, for a related account on number processing). Further, we investigate how distance of a visual referent from the maximum standard can act as a further influencing factor of processing cost.

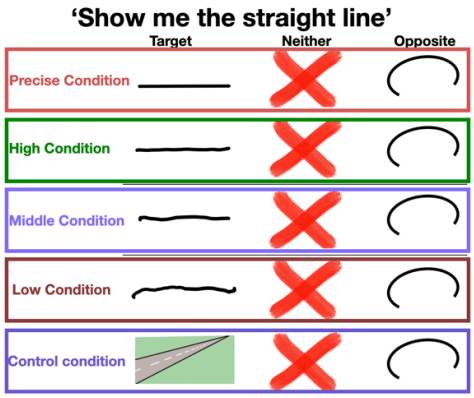
DESIGN We adopted Syrett et al's (2010) task (Figure 1) in two web-based experiments in order to investigate participants' judgements and reaction times when understanding (im)precision. Experiment 1 (200 participants) included 12 critical trials with 6 different MSAAs (*straight, closed, empty, full, round, clean)*, plus 18 filler trials. In each trial, participants read a sentence and saw three images (See Figure 1): a target image that corresponds to an MSAA, an 'opposite' image (that's always incorrect), and an image indicating that neither of the previous two was satisfactory. Their task was to select the image that best matched the sentence. Importantly, the target image had 5 levels of (im)precision: 'precise', 'high' (i.e., slightly imprecise), 'middle', 'low' (i.e., very imprecise), and control (factor: PICTURE TYPE). Levels of imprecision were normed in a pre-test. Experiment 2 (360 participants) was identical except that each trial was preceded by one of two 1-sentence contexts meant to elicit different expectations of precision: loose vs. strict (factor: CONTEXT). Contexts were also normed.

PREDICTIONS In Experiment 1, we expected participants to accept imprecise pictures in the 'high' condition, but at a cost relative to accepting pictures in the 'precise' condition (measured in acceptance-time differences), in line with Syrett et al. (2010). 'Middle' and 'low' conditions should be accepted at rates below chance while the 'control' condition should be rejected. However, for Experiment 2, we predicted context to have a key mediating role. Only precise interpretations would be accepted following the 'strict' contexts, whereas 'precise' and 'high' interpretations would be equally accepted following the 'loose' contexts. In terms of processing time, accepting a precise picture after the 'strict' context should be fastest, but, following 'loose' contexts, there should not be a difference between acceptance-times in the 'precise' and 'high' conditions. These predictions were pre-registered on the project's OSF page.

ANALYSIS We fitted mixed-effects logistic (for picture selection, 1=Target and 0='neither') and linear (for BoxCox-transformed picture acceptance-times) regression models. In Experiment 1, the 'high' condition was indeed accepted significantly less often than the 'precise' condition (~90% vs. ~100%, respectively). The 'middle' (~50%) and the 'low' (~19%) followed. The 'precise' condition showed the significantly shortest acceptance-times (see Figure 2). In Experiment 2, context significantly mediated both acceptance rate and time. Critically, there was a significant interaction in both picture acceptance rate and time between CONTEXT and the 'precise' and 'high' conditions (see Figure 2). Interestingly, the reverse pattern appeared in the rejection times for Experiment 1 and for the 'loose' conditions of Experiment 2: The smaller the degree of imprecision, the longer it took participants to reject it as an appropriate referent of an MSAA.

CONCLUSION Our study shows that without context, processing precision is less effortful relative to imprecision, similar to Syrett et al. (2010). However, once context is taken into account (Experiment 2), this cost can disappear, but only when a visual referent is close to the precise standard ('high' Picture condition). Overall, our findings highlight the pivotal role played by contextual expectations during language processing, as well as how different factors interact during processing to mediate processing effort. We see these results as being broadly in line with constraint-based accounts of pragmatic processing (Degen & Tanenhaus, 2019).

Figure 1: Example image grids (Experiment 1 & 2) and example context (Experiment 2 only). Image grids are color-coded representing the same conditions depicted in Figure 2. Each rectangle shows the three images that participants saw in a given trial for each condition. The target utterance was identical across conditions.

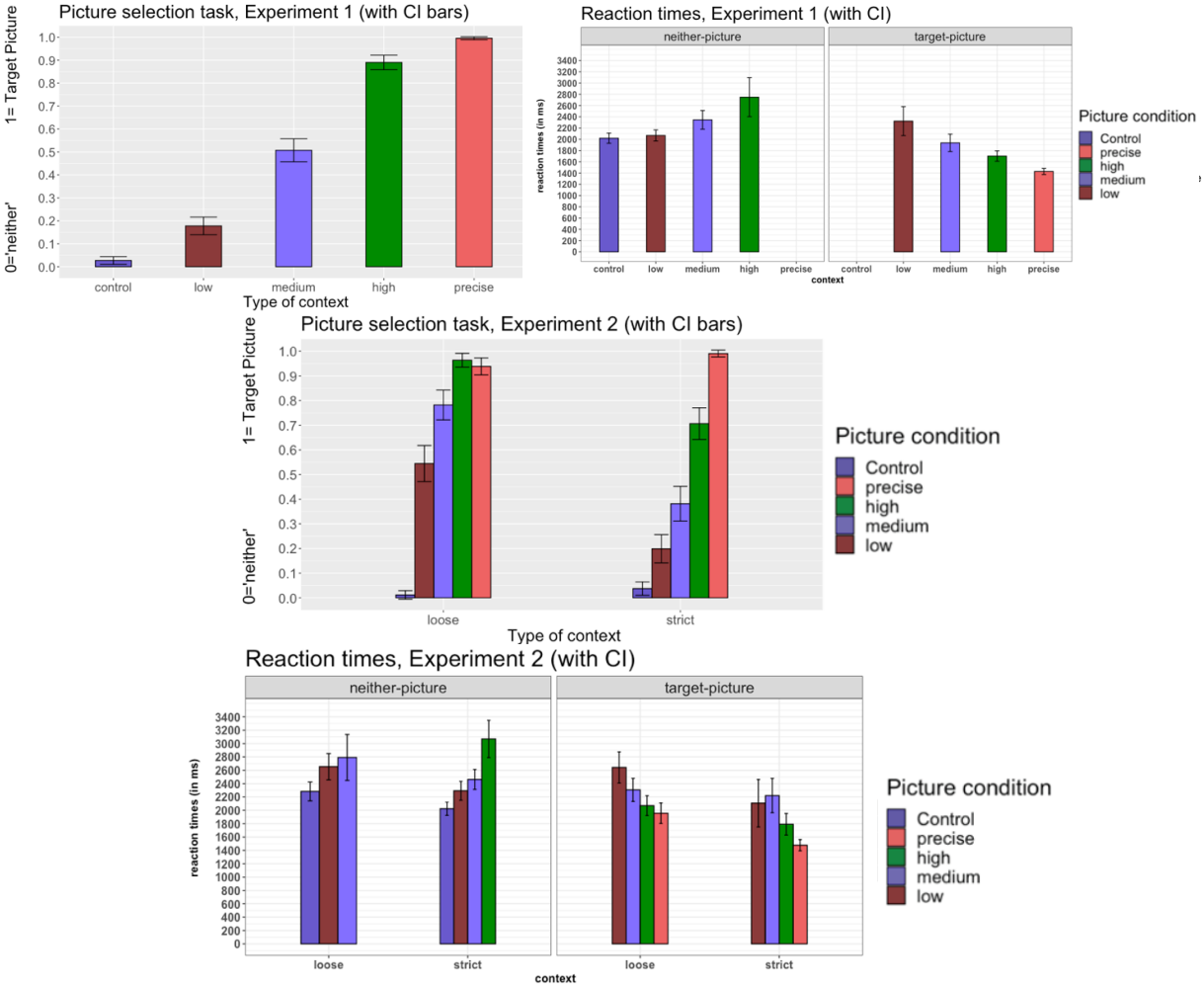


CONTEXT SENTENCES (Exp. 2)

Strict condition: *Jasmine carefully drew a line with a ruler on a piece of paper.*

Loose condition: *Jasmine rashly drew a line with her eyes closed on a piece of paper.*

Figure 2: Results of Experiments. 1 (top panel) & 2 (middle and bottom panels). RTs were transformed for analysis, shown here as raw-RTs for clarity.



References

Aparicio, H., Xiang, M., & Kennedy, C. (2016). Processing gradable adjectives in context: A visual world study. In *Semantics and Linguistic Theory* (Vol. 25, pp. 413-432).

Gibbs Jr, R. W., & Bryant, G. A. (2008). Striving for optimal relevance when answering questions. *Cognition*, *106*(1), 345-369.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, *30*(1), 1-45.

Lasersohn, P. (1999). Pragmatic halos. *Language*, 522-551.

Leffel, T., Xiang, M., & Kennedy, C. (2016). Imprecision is pragmatic: Evidence from referential processing. In *Semantics and linguistic theory* (Vol. 26, pp. 836-854).

Syrett, K., Kennedy, C., & Lidz, J. (2010). Meaning and context in children's understanding of gradable adjectives. *Journal of semantics*, *27*(1), 1-35.

Van Der Henst, J. B., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind & Language*, 17(5), 457-466

A path to ignorance: The default computation of Scalar Implicatures

Alan Bale (Concordia), Maho Takahashi (UCSD), Hisako Naguchi (Concordia), Marguerite Rolland (Concordia), and David Barner (UCSD)

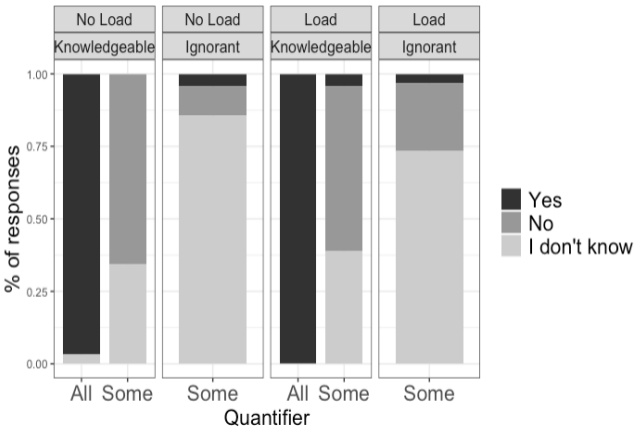
We provide experimental evidence that listeners compute scalar implicatures (SIs) by default, even in contexts where speakers are ignorant about stronger alternatives (i.e., contexts that should yield ignorance implicatures, IIs). Furthermore, these default computations are grammatically encapsulated, in the sense that the computation of SIs seems to be separate from general reasoning processes or the representation of contextual information. We show that people under cognitive load (e.g., engaged in a task that taxes their working memory) over-compute strong SIs. For example, they compute that *some* implies *not all* in contexts where it is obvious that the speaker is using *some* to imply their ignorance about *all*. They over-compute these types of inferences despite overtly acknowledging that the speaker is ignorant about the status of *all*.

1. Background & Controversy: One of the defining properties of natural language is that weak statements often provide information about the status of stronger propositions. This information comes in one of two forms: either such statements imply that stronger propositions are false (SIs, e.g., “some of the rabbits jumped” implies that not all of them did) or they imply that speakers do not know whether stronger propositions are true or false (IIs, e.g., “Franny ate a banana or an apple” implies that the speaker doesn’t know whether or not Franny ate a banana). Although much has been said about the differences between these two types of implication (see [1-9] among others), very little is known about the interaction between them. Some linguists and philosophers have suggested that contextual cues signal whether or not a speaker is knowledgeable about certain stronger propositions (see the discussions in [2-5]). If context indicates that the speaker is most likely knowledgeable about stronger propositions, hearers compute an SI, whereas if context indicates the opposite, hearers compute an II. Others have suggested that SIs are computed by default (see [6-9]). According to them, hearers automatically assume that speakers are knowledgeable/opinionated, only abandoning such an assumption when contextual information makes it impossible to maintain. Currently, however, there is little experimental evidence to differentiate these claims. A notable exception is a recent study that tested the pragmatic abilities of teens on the autism spectrum ([10]). This study demonstrated that autistic teens over-compute SIs in contexts that should only license IIs. Critically, autistic teens answered questions showing that they understood that the speaker was ignorant about the status of stronger alternatives, yet they still interpreted weak statements from the speaker as implying that the stronger alternatives were false. In other words, autistic teens were unable to consistently integrate their knowledge of the context (and the speaker’s state of mind) in order to block the computation of SIs. It remains an open question whether this separation between contextual information and (conscious) knowledge of the context is unique to autistic teens, or whether similar evidence can be found in neuro-typical adults. In the current study we attempt to answer this question by inhibiting the ability of neuro-typical adults to integrate contextual information. We tested their ability to compute implicatures under cognitive load by getting them to perform a memorization task while simultaneously computing SIs & IIs.

2. The Experiment: We tested 60 English speaking university students, ranging in age from 18 to 45 years old (mean 23.5). Participants were divided into a control group, who participated in a scalar implicature task, and an experimental group who performed the same task but under cognitive load. Specifically, experimental participants were asked to memorise a dot pattern at the beginning of each trial, and then recall the pattern at the end. In the scalar implicature task, participants were introduced to a “helpful” speaker who provided them with as much information as he could about the content of three boxes. For each trial, the speaker and the participant

were able to see what was inside the first two boxes, but the third was covered so that the participant could not see what was inside. The first two boxes always contained the same thing (e.g., two tiny orange cubes in each). Trials differed only in terms of what happened with the third, covered box. In some trials, the speaker looked into the third box (knowledgeable-speaker trials) and then made a statement either using the quantifier "some" or "all" (e.g., "Some of the boxes have orange cubes" vs. "All of the boxes have orange cubes"). In other trials, the speaker didn't look inside the third box (ignorant speaker trials) and made a statement with the quantifier "some" (e.g., "Some of the boxes have orange cubes"). At the end of each trial, participants were asked whether the speaker knew what was in the third box. Participants were required to answer this question correctly before moving on. If they did not answer correctly, the scene was replayed. Once they could answer this question correctly, participants were then asked whether the third box contained the same items as the first two (e.g., "Does the third box have orange cubes?"). They were told explicitly that they could respond, "yes", "no" or "I don't know." Given the experimental paradigm, participants without cognitive load were expected to take the speaker's knowledge-state into consideration when determining whether the third box had the same objects in it or not. If the speaker looked inside the third box, then participants were expected to answer "yes" when the speaker used the quantifier "all" (via entailment), but "no" when the speaker used the quantifier "some" (via SI). On the other hand, if the speaker didn't look into the third box and uttered a statement with "some", then participants were expected to answer "I don't know" (via contextual cues and II).

Results: The control group performed as expected, computing SIs in the knowledgeable-speaker condition and IIs in the ignorant-speaker condition (65.6% and 85.6% respectively). In contrast, the test group exhibited a significant increase in their "no" response in the condition where the speaker was ignorant (from 10% to 23.3%), despite the fact that they acknowledged that the speaker did not know what was in the third box. We constructed a generalized linear mixed-effects model that predicted "no" responses on critical *some* trials from cognitive load, knowledge state, and their interaction. The model revealed a main effect of knowledge state (β =-5.06, SE = 0.8, $p < .001$), and importantly, an interaction effect between knowledge state and cognitive load (β =2.62, SE = 0.86, $p < .01$). This reflected the fact that the "no" responses (and thus SIs) were more frequent under cognitive load in conditions of speaker ignorance, where no implicature was supported.



3. Discussion: The results reported here are compatible with the conclusion that people compute SIs by default, even in contexts that do not support such an inference. Such results complicate the traditional analysis of implicatures through a neo-Gricean perspective. The participants in this study explicitly acknowledged that speakers were ignorant about the status of scalar alternatives, yet still frequently computed SIs. Under the traditional neo-Gricean approach, hearers must assume that speakers are knowledgeable/opinionated about alternatives in order to derive an SI. The results here suggest that the computation of SIs must be separated from general reasoning about the speaker's epistemic state.

Refs:
[1] Grice (1975). Logic and conversation. In Speech acts, pp. 41–58.
[2] Horn (1989). A natural history of negation. Chicago University Press.
[3] Leech (1983). Principles of Pragmatics. Longman.
[4] Soames (1982). How presuppositions are inherited: a solution to the projection problem. LI 13: 483–545.
[5] Matsumoto (1995). The conversational condition on Horn scales. L&P 18: 21–60.
[6] Gazdar (1979). Pragmatics: implicature, presupposition, and logical form. Academic Press.
[7] Levinson (1983). Pragmatics. Cambridge University Press.
[8] Sauerland (2004). Scalar implicatures in complex sentences. L&P 27: 367–391.
[9] Guerts (2010). Quantity Implicatures. Cambridge University Press.
[10] Hochstein et al. (2018). Scalar implicature in absence of epistemic reasoning? The case of Autism Spectrum Disorders. LL&D 14(3), 224-240.

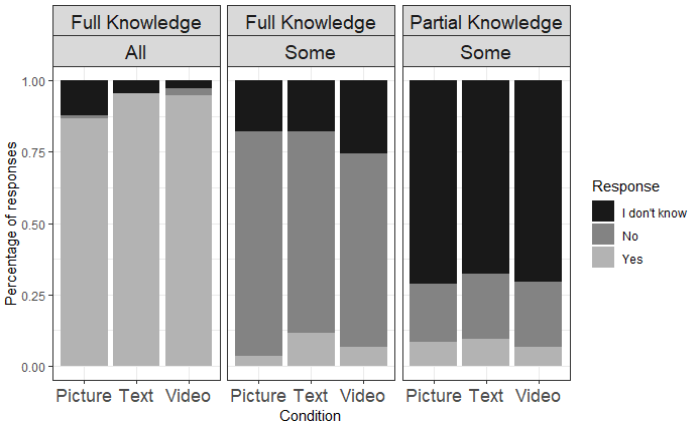
Sensitivity to speaker knowledge in online tests of scalar implicature

How is language comprehension impacted by how we experience contextual information? In two experiments, we asked whether online methods differed from in-person assessments of scalar implicature that relied on mental state reasoning - a task we reasoned might be especially sensitive to testing modality. We tested participants in one of four conditions: (1) in-person with a live-experimenter, (2) online with video stimuli, (3) online with pictures and text, or (4) online with text only stimuli. Across the experiments, no consistent differences emerged between modalities, suggesting that online methods provide valid measures of implicature under a variety of circumstances, even when relatively sophisticated mental state reasoning is involved. In particular, written stimuli were just as valid as video stimuli, if not more so.

Background: Research in semantics and pragmatics has recently witnessed rapid growth in the use of experimental methods that test large groups of participants to support robust statistical inference.[1-3] To facilitate this, many researchers have turned to online testing platforms such as Mechanical Turk and Prolific,[4-9] which include large groups of participants who speak diverse languages. However, while the validity of these online methods has been investigated in certain restricted test cases (e.g., acceptability judgments)[7-9] little is known about the impacts of online methods when testing pragmatic inferences, which often rely on subtle contextual parameters such as the knowledge states of particular speakers. For example, the computation of scalar implicatures (e.g., *some* implies *some but not all*) requires the hearer to assume that the speaker is knowledgeable about potential scalar alternatives. If contextual cues indicate that the speaker is not knowledgeable, hearers will derive an ignorance implicature instead (e.g., *some* implies *some and perhaps all*). What's unclear is whether such inferences about speaker states differ when an actual speaker, with actual mental states, is physically present vs. when a speaker is merely described via text, or otherwise represented via images or video. While implicatures have been documented across a variety of modalities,[1] it's unclear to what extent differences across these studies might be attributable to experimental modality. To investigate this question and probe the validity of remote testing methods, we tested participants' sensitivity to speaker knowledge when computing implicatures by presenting them with speakers across four modalities: in-person, remote video, remote photos, and text-only remote testing.

The Experiments: In Exp. 1, 90 English-speaking participants were recruited via Prolific, 30 per condition. These data were compared to existing data from 30 participants tested in-person by an experimenter who presented videos of the speaker on a laptop computer. Participants saw/read vignettes about a speaker (Mary) who had three boxes in front of her. Conditions differed in the modality: vignettes were presented either as video clips, still images, or short paragraphs of text. In each trial, the contents of the first two boxes were revealed to the participant and both always contained the same object types (e.g., apples). Mary then either looked inside the third box without revealing the contents to the participant, or did not look inside, and made a statement about the contents of the boxes using either 'some' or 'all'. There were therefore three types of trials: those where Mary looked in all three boxes and said 'all' (e.g., "All of the boxes have apples."; full knowledge/all), those where Mary looked in all three boxes and said 'some' (e.g., "Some of the boxes have apples."; full knowledge/some), and those where Mary looked in two out of three boxes and said 'some' (partial knowledge/some). Participants then answered a question about the contents of the third box (e.g., 'Do you think that there are apples inside the third box?'), by choosing "Yes", "No", or "I don't know".

Expected responses for each condition were as follows: full knowledge/all should lead to “Yes”, full knowledge/some should lead to “No” (as a result of computing a scalar implicature), and partial knowledge/some should lead to “I don’t know”. Participants completed a total of 9 trials (3 of each type). In Exp. 2, we conducted an exact replication of Exp. 1, but doubled the number of participants to 60 per condition, 180 total. The goal of Exp. 1 was to verify the reliability of effects observed in Exp. 1.



Results: Data from Exp. 1 were analyzed with the existing in-person data. We constructed a generalized linear model (GLM) predicting the proportion of participants’ “No” response to the trials with ‘some’ based on modality, knowledge state, and their interaction. The in-person condition was treated as the baseline. The model revealed a significant main effect of knowledge state (β=-2.84, SE=0.42, $p < 0.001$), as well as an interaction effect between modality and knowledge state; in particular, the proportion of “No” responses in partial knowledge trials increased with the online/video modality (β=1.25, SE=0.53, p=0.02; see Figure). As the expected response on partial knowledge trials was “I don’t know,” this effect suggests that participants in the online video condition were slightly more likely to compute scalar implicatures even though the speaker’s knowledge state (i.e., not knowing what is inside the third box) did not support doing so. In order to test whether participants were simply less attentive in an online setting, we reran the GLM model predicting the proportion of “I don’t know” responses. Shifting modalities from in-person to online did not result in an increase in these responses, suggesting that online participants were not overall less certain than in-person participants. For Exp. 2 we again created a GLM predicting the proportion of “No” responses to ‘some’ based on modality (picture vs. text vs. video). Contrary to Exp. 1, a chi-square test found no significant effect of modality (Deviance=1.49, df=2, p=0.47) with a larger sample size (n=60 per modality). A model predicting “I don’t know” responses found no significant effect of modality, replicating Exp. 1 (Deviance=2.12, df=2, p=0.35), again suggesting that modality did not affect participants’ attentiveness. **Conclusion**: We find no reliable impact of testing modality on how participants compute scalar implicature. Online text-only stimuli were just as likely to generate implicatures as richer modalities that featured images and video, despite the role of mental state reasoning in the tasks. **Refs**: **[1]** Chemla & Singh (2014). Remarks on the experimental turn in the study of scalar implicature, Pt I. L&LC. **[2]** Cummins & Katsos (2019). The Oxford Handbook of Experimental Sem. & Prag. **[3]** Devitt, M. (2011). Experimental semantics. P&PR. **[4]** Erlewine & Kotek (2016). A streamlined approach to online linguistic surveys. NLLT. **[5]** Munro et al. (2010). Crowdsourcing and language studies. **[6]** Fort et al. (2011). Amazon mechanical turk: Gold mine or coal mine? CL. **[7]** Sprouse, (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. BRM. **[8]** Schnoebelen & Kuperman (2010). Using Amazon mechanical turk for linguistic research. Psihologija. **[9]** Gibson et al. (2011). Using Mechanical Turk to obtain and analyze English acceptability judgments. L&LC.

Investigating discourse referent salience patterns of negative quantifying expressions

Eva Klingvall, Lund University & Fredrik Heinat, Linnæus University

In this talk, we report the results from three studies investigating discourse salience patterns of negative quantifying expressions (e.g. 'not all', 'few') in Swedish, from both a hearer (comprehender) and a speaker (producer) perspective. Salience from these two perspectives has been argued to rely on different information structural properties. For hearers, sentence TOPICS are often more salient than non-topics (COMMENTS), while for speakers, FOCUSSED material is often more salient than BACKGROUNDED material (e.g. Chiarcos, 2010; Molnár and Vinckel-Roisin, 2019). The hearer perspective has been extensively studied in the context of pronoun resolution (e.g. Ariel, 1990; Gundel et al., 1993). Previous research on quantifying expressions in English has shown that for negative quantifying expressions (monotone decreasing), such as *not all, not many, few*, both the set of entities for which some property is true, the REFERENCE SET, and the set of entities for which the property is *not* true, the COMPLEMENT SET, are available for anaphoric reference. Although both sets are possible, speakers generally prefer to refer back to the COMPLEMENT SET (e.g. Moxey and Sanford, 1987, and subsequent work):

(1) Not many kids were outside in the morning.
  a. They were building a snow castle. REFSET
  b. They stayed inside instead. COMPSET

In three sentence continuation studies, we investigated which of these sets speakers referred back to, and what linguistic form they used to refer to this set. The aim was to find out what discourse topic speakers selected and how this selection reflected both hearer and speaker salience of discourse entities. In Experiment 1, 244 participants read the sentence in (2) but with one of the eight QEs in (3) instead of 'QE' (six negative ones, plus two positive ones included as a control condition), and wrote a continuation of it. As indicated in the translation in (2), the word *de* can be either a personal pronoun, which can appear with or without modifiers, a demonstrative pronoun, or a definite article.

(2) QE föräldrar var på klassmötet igår och de . . .
'*QE parents were in the school meeting yesterday and they/the/those . . .*'

(3) a. **Negative Quantifying Expressions**
inte exakt alla 'not exactly all', *inte precis alla* 'not precisely all' *inte riktigt alla* 'not quite all', *få* 'few', *inte många* 'not many', *nästan inga* 'almost no
  b. **Positive Quantifying Expressions**
några 'some', *nästan alla* 'almost all'

For all negative quantifiers except *få* ('few'), the linguistic form used as an anaphor indicated that the COMPLEMENT SET was most salient from a hearer perspective while reference to the REFERENCE SET required a more marked structure. However, for all quantifiers it was the REFERENCE SET that was most salient from the speaker perspective, most often selected as the discourse topic. In Experiment 2, we had a closer look at the quantifier *få* ('few'), investigating whether relative and cardinal readings of this quantifier (see e.g. Partee, 1989) resulted in different patterns and could shed some light on the exceptional behaviour of *få* in Experiment 1. Sixty-one participants read the sentence in (2), with one of the quantifying expressions *färre än tio* ('fewer than ten') (cardinal) and *färre än hälften* ('fewer than half') (relative) in place of 'QE', and wrote a continuation of it. The results were similar to those for *få* in Exp 1, with no clear difference between the cardinal

and the relative quantifying expressions. Thus, the participants referred back to the REFERENCE SET, using an unmodified pronoun. The REFERENCE SET was thus most salient from both the hearer and speaker perspective.

In Experiment 3, we investigated whether the discourse salience patterns of negative QEs are affected by the status of the clause in which the anaphoric NP is found. The sentence fragment read by the participants (192) was therefore modified to include a complementizer, *att* ('that'), before the final word, *de* ('they/the/those'). In this way, the participants were prompted to write a continuation where *de* would be (part of) the subject of the *that*-clause that would itself function as the subject of a co-ordinated structure. Instead of 'QE' the six negative quantifying expressions from Experiment 1, in (3a), were used.

(4) QE föräldrar var på klassmötet igår **och att de . . .**
'QE parents were in the class meeting yesterday ***and that they/the/those . . . '***

With this form of the prompt, the participants selected the COMPLEMENT SET as discourse topic to a much larger extent than in the other two experiments. For all quantifiers except *få* ('few'), the COMPLEMENT SET was the most salient set from both the hearer and the speaker perspective in this experiment. The quantifier *få* again showed a different behaviour but notably to a lesser degree than in Experiment 1. Experiment 3 thus showed that the speaker salience pattern is also dependent on whether the subject of the continuation is an entity or a proposition. The three experiments showed that the discourse referent that is re-mentioned in production is not necessarily the one that is most salient in comprehension, supporting views that hearer- and speaker-salience should be distinguished (e.g. Chiarcos et al., 2011). This distinction is important not least in the study of reference patterns of quantifying expressions.

References

Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London and New York: Routledge.

Chiarcos, Christian. 2010. Mental salience and grammatical form. Toward a framework for salience in natural language generation. Ph.d. thesis, University of Potsdam, Potsdam.

Chiarcos, Christian, Berry Claus, and Michael Grabski. 2011. Introduction: Salience in linguistics and beyond. In *Salience: Multidisciplinary perspectives on its function in discourse*, ed. Christian Chiarcos, Berry Claus, and Michael Grabski, 1–26. Berlin: De Gruyter.

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69:274–307.

Molnár, Valéria, and Hélène Vinckel-Roisin. 2019. Discourse topic vs. sentence topic. Exploiting the right periphery of German verb-second sentences. In *Architecture of topic*, ed. Valéria Molnár, Verner Egerland, and Susanne Winkler, 293–333. De Gruyter Mouton.

Moxey, Linda M., and Anthony J. Sanford. 1987. Quantifiers and focus. *Journal of Semantics* 5:189–206.

Partee, Barbara. 1989. Many quantifiers. In *Proceedings of the fifth Eastern states conference on Linguistics*, ed. Joyce Powers and Kenneth de Jong, 383–402. Colombus: The Ohio State University.

Introduction What semantic representations underlie the meanings of numerals? Let us assume that numerals denote either (i) numbers, when they non-ambiguously pick out a number — for instance, the denotation of English *two* is in (1) (we leave aside the question of whether this results from truth-conditional meaning or from pragmatic enrichment, cf. Spector 2013); or (ii) sets of numbers, when their meaning is 'more than n', for some number n (cf. Table 1).

(1) $[\![\text{two}]\!] = 2$ (2) $[\![\text{two}]\!] = 1 + 1$ (3) $[\![\text{two}]\!] = s(1)$

(2) and (3) are denotationally equivalent to (1), with s the successor function. However, does our semantic representation of *two* involve 2 as a semantic primitive, or $1 + 1$, or $s(1)$? As semantic representations cannot be observed, they need to be inferred. Multiple approaches to this challenge have been developed, inferring semantic representations from behavioural data (Hackl 2009, Pietroski et al. 2009, Lidz et al. 2011, Piantadosi et al. 2012, 2016, Knowlton et al. 2021) or from typological generalizations (Züfle and Katzir 2021). We put forward a novel approach to inferring semantic representations using data on the optimality of the languages' simplicity/informativeness trade-off. We use numerals as a case study, building on the simplicity/informativeness trade-off analysis by Xu et al. (2020). Importantly, the approach can be applied to any domain for which cross-linguistic semantic data is available.

Hypotheses In this project, we focus on numerals denoting numbers or sets of numbers 1–10. We follow a tradition in semantics and philosophy of language to think about semantic representations in terms of combinations of primitive concepts (Fodor 1975, Pietroski 2018), and assume that the semantic representations of numerals are composed from a certain set of primitive number concepts PRIM, functions $+$, $-$ and successor s ($s(n) = n + 1$), and relation 'greater than' $>$ ($> n = \{x \in \mathbb{N} | x > n\}$) (cf. also Xu et al. 2020). Semantic primitives and operations may have different complexities (e.g., $+$ may be semantically more complex than $>$). This can be modeled as *weight* w_x for a primitive or operation x. Our research question is what PRIM and weights underlie semantic representations of numerals. We explore ten hypotheses, according to which PRIM contains $[1, \ldots, n]$, with $n \in \{1, \ldots, 10\}$. For each of these hypotheses, we consider all possible assignments of two values (1 and 2) to w_{PRIM} (the weight assigned to elements of PRIM), w_+, w_-, w_s, $w_>$. This amounts to $2^5 \times 10 = 320$ hypotheses.

Method Natural languages differ in terms of how complex they are to represent and in terms of how informative they are (i.e. how precise a communication they allow for). For instance, focusing on numbers 1-10, some languages have numerals for only a few of them, while others have numerals for each of them (cf. Table 1) – the former are simpler, but the latter are more informative. Simplicity and informativeness are in a tension: languages cannot both be maximally simple and maximally informative. This tension is known as the *simplicity/informativeness trade-off problem*. There can be many *optimal solutions* to this problem: the set of optimal solutions is called the *Pareto frontier*. More specifically, a language is (Pareto) optimal if there is no other language that has both lower complexity and higher informativeness. Computational modeling of cross-linguistic semantic data has demonstrated that natural languages optimize the simplicity/informativeness trade-off (Kemp and Regier 2012, Steinert-Threlkeld 2019, Denić et al. 2021, Uegaki 2020, Xu et al. 2020). Importantly, simplicity of a a language is assumed to be a function of the semantic representations underlying its expressions: these studies thus stipulate underlying semantic representations of languages' expressions, and analyze the optimality of the simplicity/informativeness trade-off under those stipulations. In the present work, we reverse the direction of the analysis: we assume that languages trade optimally simplicity and informativeness, which allows us to infer the semantic representations underlying their expressions. Concretely, the aforementioned 320 hypotheses will be evaluated as follows: as we assume that natural languages should be optimal solutions to the problem, we will have evidence against a hypothesis if under that hypothesis natural languages are not at the Pareto frontier. This approach connects to recent work by Zaslavsky et al. (2021), who show that two different hypotheses about cognitive biases involved in personal pronoun systems lead to different trade-off results.

Complexity of numeral systems We take the complexity of a combination of primitives to be the sum of weights of primitives and operations involved in the combination. For instance, if $w_1 = 1$ and $w_+ = 2$, the complexity of '1+1' is 4. We assume that the semantic representation underlying a numeral is the lowest complexity combination of primitives compatible with its denotation. The complexity of a language is defined as the sum of complexities of semantic representations underlying its numerals.

Informativeness of numeral systems The informativeness of a language $I(L)$ is defined in (6) (cf. Skyrms 2010, Steinert-Threlkeld 2019, Denić et al. 2021). It corresponds to the probability that the

Numeral systems	Languages
1, 2, 3	Bare, !Xóõ
1, 2, 3, more than 3	Achagua, Araona, Hixkaryana, Krenak, Mangarrayi, Martuthunira, Pitjantjatjara
1, 2, 3, 4, more than 4	Awa Pit, Kayardild
1, 2, 3, 4, 5, more than 5	Barasano, Imonda, Rama, Yidini
1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Hup, Waskia, Wichi

Table 1: 18 exact restricted languages per their numeral systems inventory (Xu et al. 2020, Comrie 2013). For instance, Krenak has terms for 1,2,3, and a term that can be used for any number greater than 3.

communication will be successful given the speaker's probability to use an expression e to communicate a number n, $P_S(e|n)$, as in (4), the listener's probability to guess n upon hearing e, $P_L(n|e)$ as in (5), and the need probability to communicate about different numbers $P(n)$, which we assume to be approximated by a power law distribution as in (7), following Dehaene and Mehler (1992), Piantadosi (2016).

Natural languages Our natural language sample consists of the 18 exact restricted languages from Xu et al. (2020). They can be divided into 5 classes (Table 1) in terms of their numeral systems inventory denoting numbers 1–10. Informally, these are languages for which it's not the case that for every natural number they have a numeral non-ambiguously denoting it.

Hypothetical languages We generate all numeral systems ($N = 1534$) with (i) numerals denoting one of the numbers 1-10, and/or (ii) an expression meaning 'more than n', for some number n. We assume that expressions within a single language don't overlap in their denotations.

Results For each of the the 320 hypotheses: (i) we compute the complexity and informativeness of natural and hypothetical languages; (ii) we find the set of optimal languages (the Pareto frontier); (iii) we compute the average distance $\overline{D}$ of natural languages from the Pareto frontier. If languages are optimal solutions to the trade-off problem, we can discard all hypotheses for which $\overline{D} \neq 0$. We find that there are 10 out of 320 hypotheses for which $\overline{D} = 0$. These 10 hypotheses can be compressed into 2 families of hypotheses: (i) PRIM = $\{1\}$ and $w_s > w_>$; or (ii) PRIM = $\{1,2\}$ and $w_{\text{PRIM}}, w_+, w_s > w_>$. Interestingly, no hypothesis where PRIM contains $[1, \ldots, n]$, with $n \in \{3, \ldots, 10\}$, results in natural languages being Pareto optimal.

Discussion (i) The 310 hypotheses not resulting in the optimal simplicity/informativeness trade-off can only be discarded under the assumption that natural languages are optimal solutions to the trade-off problem. This assumption may be too strong: natural languages may be very good solutions, but not necessarily optimal. If this is the case, our approach cannot provide categorical evidence against certain hypotheses, but can nonetheless be used to evaluate their plausibility: if natural languages are far from the Pareto frontier under a specific hypothesis, that makes the hypothesis unlikely to be true. (ii) It would be interesting to explore more fine-grained weight assignments, which may reveal that under specific assumptions, PRIM other than $\{1\}$ or $\{1,2\}$ can result in the optimal simplicity/informativeness trade-off. (iii) The approach we develop adds to existing approaches to inferring semantic representations (cf. *Introduction*), creating novel opportunities to compare and integrate findings from multiple approaches.

Conclusion We have developed a new methodology for studying semantic representations underlying a semantic domain. We applied the method to numerals; importantly, the method can be applied to other semantic domains for which cross-linguistic semantic data is available. We thus hope that it will be a valuable tool for studying semantic representations underlying truth-conditional meanings.

$$(4) \quad P_S(e|n) = \frac{[\![e]\!](n)}{\sum_{e' \in L} [\![e']\!](n)} \qquad (6) \quad I(L) = \sum_{n \in N} \sum_{e \in L} P(n) P_L(n|e) P_S(e|n)$$

$$(5) \quad P_L(n|e) = \frac{[\![e]\!](n)}{\sum_{n' \in N} [\![e]\!](n')} \qquad (7) \quad P(n) \propto n^{-2}$$

Selected references: Denić et al. (2021). Complexity/informativeness trade-off in the domain of indefinite pronouns. *SALT 2020* | Fodor (1975). The language of thought. *HUP* | Kemp & Regier. Kinship categories across languages reflect general communicative principles. *Science* | Xu et al. (2020). Numeral systems across languages support efficient communication. *Open Mind* | Zaslavsky et al. (2021). Lets talk (efficiently) about us: Person systems achieve near-optimal compression. *CogSci 2021*

A corpus-based study of (non-)exhaustivity in *wh*-questions

A key issue in *wh*-question interpretation regards the distribution of exhaustive (Mention-All, MA) vs. non-exhaustive (Mention-Some, MS) question readings (see (1) and (2)):

(1) Who came to the party?
   a. Who is every person that...? MA
   b. Who is a person that...? ?MS

(2) Where can I find coffee?
   a. What is every place that...? MA
   b. What is a place that...? MS

Linguists' intuitions have typically concluded that MA is generally appropriate, while MS is marked [1-10]. Linguistic factors have been noted to generate variation in readings, including the specific *wh*-word—e.g., *who*-questions are biased for MA, while *where*/*how*-questions are biased for MS [11-12]—and existential (priority) modality—e.g., *can* purportedly licenses MS, as in (2) [3-5,7-8]. Recent work [13] tested these judgements in lab-controlled experiments with artificial stimuli and found evidence for some biases; however, [13] showed these biases can be overridden by features of the context like speaker/discourse goals [2,7,10-12]. To-date there is no systematic investigation of *naturally occurring questions* that tests the intuitions reported in the literature. We ask: (Q1) How much does question interpretation vary in natural discourse contexts? Is there indeed a bias for MA? (Q2) Is the distribution of interpretations modulated by linguistic form? We addressed these questions in a two-part study.

Methods. **Step 1: Naturalistic Stimuli from a Corpus Database.** Using TGrep2 and the Tgrep2 Database Tools [14-16], we extracted all occurrences of *wh*-questions (10,009) from the Switchboard corpus [17] and coded the questions for syntactic structure (e.g., embedded, root), *wh*-word, and presence of modality. To curate stimuli for step 2, we focus on root and embedded questions, leaving 2070 unique *wh*-questions. The distribution of *wh*-word and modality is reported in Table 2. **Step 2: Paraphrase Rating Task.** The remaining cases were divided into 31 lists with occurrence of critical factors roughly proportional to the overall database. Participants (n=1740) on Prolific were presented with each question and the 10 preceding lines of dialogue, and asked to rate the likely intended meanings (paraphrases), using a slider task (Fig. 1). Question paraphrases were selected to reflect MS/MA readings: *a* indicates MS ((1b)/(2b)), *every* MA ((1a)/(2a)), while the two readings converge in *the*-paraphrase (*what/who is the place/person*). There was a fourth option (*something else*) in case no other was appropriate.

Results. Questions with highest ratings for *something else* (17%) were excluded because they were rhetorical (see Tab. 1). *The*-paraphrases, where MS=MA, had the highest mean rating (.55), suggesting that only one reading was possible for most cases. Data were analysed using linear mixed effects regression. To investigate the posited MA bias, we compared *every* vs. *a* ratings, as these represent MA and MS (Fig. 2): although there was no bias for *every* contrary to literature (Q1), means for MA were higher than MS agregating over root and embedded questions. However, significant 3-way interactions between paraphrase and linguistic form factors partially support reports from the literature (Q2). First, the presence of a modal resulted in higher ratings of *a* [5-9,10] but not *every* for all except for *where*-questions. Second, *how*,*why*, and *when*-questions all showed a bias for MS, confirming [3-4, 10]; *who* and *where* show no bias (except for 'the') in contrast, and finally *what* questions revealed a bias shifting from MA to MS with a modal present.

Conclusion. In contrast to theoretical predictions, we find no bias for MA question readings in naturalistic dialogue (Q1). With respect to (Q2), we find support for some, but not all, observations about the effect of linguistic form on question interpretation reported in the literature. We suggest that MS/MA readings result from reasoning about the speaker's goal in the context, consistent with a constraint-based account [18] on which hearers integrate multiple sources of information to determine meaning. These results also highlight the importance of large-scale experiments for insight into more realistic meaning distributions [19].

Paraphrase	Example	Mean
every (MA)	*Where have you skied?*	.66
	Where's it all going?	.59
a (MS)	*Where do you like to eat?*	.57
	How would you achieve that?	.51
the (MS=MA)	*Where you going to school?*	.99
	Where do you work?	.99
something else	*Who knows?*	.61
	How can you watch that?	.53

Table 1: For each paraphrase, examples of questions that resulted in high ratings on that paraphrase.

Figure 1: Paraphrase Rating Task: Participants evaluate intended question meanings by moving the slider next to paraphrases, assigning a numerical value between 0-1 to generate a proper probability distribution. Combined ratings must sum to 1.

References. [1] Karttunen (1977), [2] Groenendijk & Stokhof (1984), [3] George (2011), [4] Nicolae (2013), [5] Fox (2014, 2018), [6] Chierchia & Caponigro (2013) [7] Dayal (2016), [8] Xiang (2016, 2020), [10] van Rooij (2003), [11] Ginzburg (1995), [12] Asher & Lascarides (1998), [13] Moyer & Syrett (2019), [14] Rohde (2005), [15] Jaeger (2006), [16] Degen & Jaeger (2011), [17] Godfrey et al. (1992), [18] Degen & Tanenhaus (2019), [19] Degen (2015)

Wh-word	+Modal	-Modal
What	5.3%	41.28%
How	3.93%	23.38%
Where	1.72%	9.38%
Why	1.07%	4.69%
Who	0.35%	5.00%
When	0.24%	1.45%

Table 2: Distribution of *wh*-words and modality in Switchboard root and embedded questions. % of total (2070).

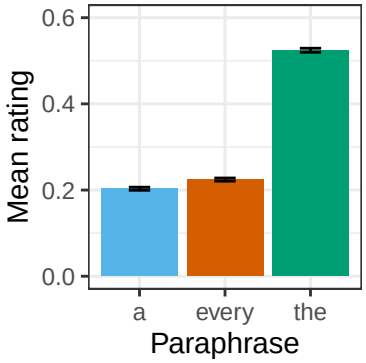


Figure 2: *Every*-paraphrases were not preferred over *a*-paraphrases.

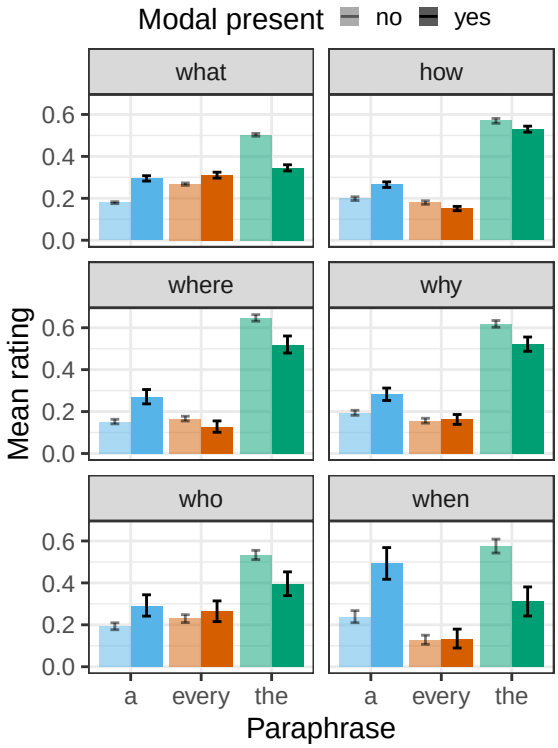


Figure 3: Significant 3-way interactions confirm some but not all intuitions from literature about linguistic form factors.

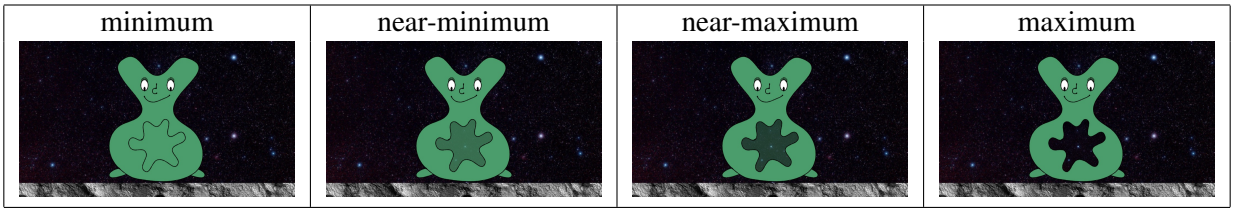
Generalizating NPIs to positive uses in an Artificial Language Jeremy Kuhn and Mora Maldonado

Overview Negative Polarity Items (NPIs) are characterized by a polarity-sensitive use, restricted to downward entailing (DE) environments. However, in many languages, the very same lexical items also have positive uses that appear in upward entailing (UE) environments. For example, under negation, the NPI *any* has an existential meaning (*I didn't talk to anybody* = *I didn't talk to a single person*). But certain UE environments allow *any* to appear with a universal (or free choice) meaning (*I talked to anyone was interested* = *I talked to everyone who was interested*). Similar positive uses can be found for other NPIs in English (*ever*, *yet*, *anymore*) and other languages (e.g. French *encore*).

Ladusaw (1979) observes that the positive and negative uses of NPIs are often systematically related: they are *logical duals*. If an NPI is licensed by (and scoping under) negation, its positive counterpart carries the meaning the word would need to receive if it were interpreted as scoping above the negation, in order to derive the same sentential meaning. For *any*, existential force under negation becomes universal force (since $\neg\exists = \forall\neg$). This observation offers a potential diachronic explanation of the systematic ambiguity. When the syntactic distribution of a logical item is restricted so that it always appears in the presence of negation, the meaning of the item is ambiguous between two denotations depending on its scope relative to negation: A¬ or ¬B. The typological data above can then be explained by the hypothesis that when the use of an NPI is extended to new, UE environments, an attractive interpretation is the wide-scope dual meaning (A).

Using an artificial language learning paradigm, we test how learners generalize the meaning of NPIs when they appear in positive environments. We teach English speaking participants an artificial language which includes a negative marker *em* ('not') and a degree modifier *tup*, roughly equivalent to English 'at all'. During training, participants are exposed to sentences in which *tup* is restricted to negative sentences (i.e. *tup* never occurs without *em*). At test, participants are asked to interpret sentences where the degree modifier appears on its own, without negation. We evaluate whether learners are more likely to assign a universal meaning (as attested in the typological data) or an existential meaning to this sentence.

Methods Participants were taught a miniature language consisting of four predicates, four proper nouns, one negative marker and one degree modifier. All predicates denote gradable properties with closed scales (e.g., transparent/opaque). For each property, we define four possible scale points: 'minimum', 'near-minimum', 'near-maximum', 'maximum,' as shown below for the noun *Greenie* and the predicate *pleet*:



Participants were first trained on the following non-target sentences: (i) *simple positive (SP)* (e.g., 'Greenie pleet'), used when the predicate applies to a maximum or near-maximum degree; (ii) *simple negative (SN)* (e.g., 'Greenie em pleet'), used for minimum or near-minimum degrees; and (iii) *negative NPI (Neg-NPI)* (e.g., 'Greenie em pleet tup'), used only when the predicate applies to a minimum degree. Crucially, participants had no evidence of the use of the degree modifier in absence of negation. At test, participants were asked to interpret these *positive NPI (Pos-NPI)* held-out sentences (e.g., 'Greenie pleet tup'). Participants had to decide whether the Pos-NPI sentence can be used when the noun applies to the predicate to a near-minimum degree or to a maximal degree. These two choices correspond to the two dual meanings that could be posited for the NPI: existential and the universal, respectively. After test, we asked subjects for translations of all four sentence types. (This experiment was preregistered here.)

Results 49 English speaking participants were recruited on Prolific and successfully trained on non-target sentences (i.e. accuracy rates above 75%). Fig. 1 (left side) shows the proportion of trials on which participants chose the 'maximum-degree' meaning for Pos-NPI sentences during the test phase. A logit mixed-effects model showed that the proportion of responses compatible with these maximum meanings is significantly below chance ($\beta = -1.6; p = .0137$), revealing an overall preference for 'near-minimum' meanings. However, a visual inspection of Fig. 1 reveals two clusters of participants. While approximately 50% of participants consistently derive 'near-minimum' meanings (driving the statistical effect reported above), a second group, which corresponds to 25% of our sample, systematically select the 'maximum-degree' meaning (binomial test: $p < .05$). This suggests the existence of two populations who generalize in different directions.

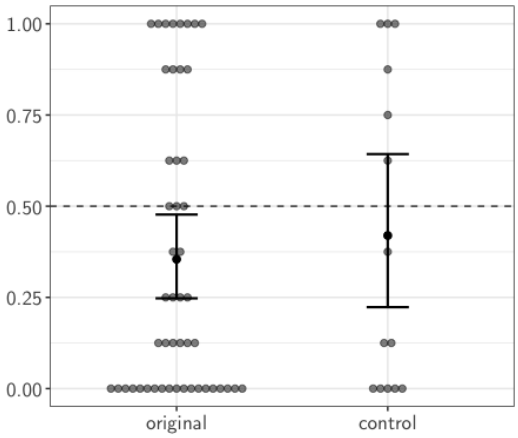


Figure 1: Proportion of 'maximum' responses

Translations provided at the end of the experiment give further insight into the make-up of these two groups. 'Maximum degree' responders systematically translate Pos-NPI sentences as using universal degree modifiers like 'completely' or 'very,' and Neg-NPI sentences as involving either the same words (e.g. 'Greenie is very transparent') or 'at all' (e.g. 'Greenie is not transparent at all'). Among the 'near-minimum degree' responders, translations of Pos-NPI sentences are less consistent. Notably, though, no subjects translated the meaning using existential degree modifiers like 'a bit' or 'somewhat'. On the other hand, a number of participants gave the sentence a *negative* meaning: 'X tup' is translated as 'not X.' For such subjects, the 'near-minimum' meaning is presumably chosen as the one that is comparatively closer to the minimum.

Control To investigate whether 'near-minimum' responses arise from a negative interpretation of the Pos-NPI sentences, we modified the original experiment, replacing the 'near-minimum' choice with a 'minimum' choice in test trials. Fig.1 (right side) shows pilot results for 14 participants. While these results are preliminary, the existence of a group ($\sim$50% of participants) that consistently derive minimum interpretations supports the hypothesis that in both experiments, non-'maximum' responders interpret the NPI as negation.

Discussion The results reported here show that several different strategies are adopted when extending the meaning of NPI items to contexts without a licensor. These strategies correspond to two meaning shifts attested in diachronic typology. First, one group of participants assign a 'maximum degree' interpretation to Pos-NPI sentences, thus displaying a pattern of generalizing the NPI meaning to its wide-scope dual. In the appropriate sociolinguistic contexts, such a population could explain the emergence of positive *any*, *ever*, *yet*, and *anymore*. A second group of participants assigns a 'near-minimum' degree interpretation to Pos-NPI sentences. Translations and a control experiment suggest that this is not due to an existential interpretation, but rather due to a negative interpretation of the NPI, possibly due to a repair strategy with reconstructed negation. Interestingly, this generalization corresponds to Jespersen's cycle (1917), in which a minimizing NPI is reinterpreted as contributing negation itself.

Further syntactic and semantic factors may influence the generalization strategy adopted by participants. Syntactically, a word order that privileges a specific scopal configuration may make a wide-scope dual interpretation more or less accessible. Semantically, properties of the predicates may also affect generalization preferences. In future work, we intend to use the present paradigm to test the strength of these factors, which may make specific predictions about the kind of diachronic change a given language is likely to undergo.

Refs. Jespersen 1917. *Negation in English and other languages*. • Ladusaw 1979. *Polarity sensitivity as inherent scope relations*.

Exhaustivity in preschoolers' clefted focus interpretation: Identification in context

The issue One aspect of sentence interpretation that seems to become adult-like relatively late in the course of language development involves inferences triggered by focus (Höhle et al. 2016). A key inference of this type (at least when focus is used to answer an explicit or implicit question, called Question Under Discussion, QUD) is exhaustivity, namely, that replacing the focused element with any of its possible (non-weaker) alternatives would yield false alternative answers to the same QUD. Previous research has uncovered that children do not compute this inference at adult-like levels before seven years of age, even in cleft(-like) syntactic constructions (Heizmann 2012, Tieu & Križ 2017, Pintér 2018). The nature of this limitation, however, is still unclear. Specifically, it is not known whether preschoolers' non-exhaustive interpretations are merely due to their difficulties in accurately identifying the focus and the relevant alternatives to it in the context (=Hypothesis1), or they also reflect some deeper-running limitation hindering the computation of the exhaustivity inference itself in clefts (=Hypothesis2). We report on a comprehension study of five-to-six-year-old children whose aim is to adjudicate between these two alternative hypotheses, as applied to pre-verbal focus in Hungarian.

Motivation According to one possible approach to the exhaustivity of focus, this inference is essentially similar in its logical structure to scalar implicatures associated with scalar items like *some*, whose acquisition is better researched. While these latter inferences have also been found to be acquired late in early studies, more recent results show that when adequate contextual support is provided as part of the experimental task to highlight the relevance of scalar alternatives, scalar inferences appear to be already present at much earlier ages (Chierchia et al. 2011, Foppolo et al. 2012, Papafragou & Tantalou 2014, Guasti et al. 2015). By analogy, Hypothesis1 holds that children's non-exhaustive interpretations of focus are caused by their difficulties in exploiting the context to identify the focus and its set of alternatives. By contrast, Hypothesis2 takes the delay compared to (other) scalar inferencing to be real in that it assumes that, while children's difficulties in utilizing the context to properly identify the focus and its relevant alternatives might contribute to protracted acquisition, yet this is not the key factor. If so, then this latter factor must be sought in the meaning of clefted focus.

The experiment The study consists of two sub-experiments (= TASK) based on sentences containing a fronted focus. In Subexp1 children had to correct false assertions on the basis of picture stimuli (a task adapted from Szendrői et al. 2018). Congruent corrections of the element in focus reflect successful identification of the focus and its relevant alternatives. Subexp2 employed a TVJ task, using sentence–picture pairs to test the acceptance or (partial or full) rejection of non-exhaustive interpretations of focus. Both sub-experiments were conducted with the same thirty-two 5-6-year-old children (mean age: 5;10) in two sessions one week apart, which differed (in addition to the lexicalizations used) in the presence of a congruent *wh*-question before each test sentence in the second session (= CONTEXT).

Predictions Adding an explicit *wh*-question was expected to enhance the accurate identification of the focus and its contextual alternatives (by boosting the latter's relevance). According to Hypothesis1, this should yield an increase in congruent corrections in Subexp1, and a concomitant rise of (at least) the same extent in the rate of exhaustive responses in Subexp2. While Hypothesis2 is also compatible with an increase of congruent/exhaustive responses in Subexp1/Subexp2, it crucially predicts that in Subexp2 any such contextual effect of the presence of an explicit question should be limited: the proportion of exhaustive responses in Subexp2 is expected to rise by a smaller rate (if at all) than the increase of congruent responses in Subexp1.

Results and discussion The presence of a *wh*-question enhanced children's exhaustive interpretations in Subexp2 less than it helped their focus-corrections in Subexp1 (while adult controls (N=12) were at ceiling in the *wh*-question condition of both sub-experiments), yielding a significant interaction between TASK and CONTEXT. This outcome confirms the predictions of Hypothesis2 over those of Hypothesis1: the key factor hindering children's focus-exhaustification cannot simply be poor identification of focus and its relevant alternatives. We argue that of

competing approaches to exhaustivity in cleft(-like) focus constructions, DeVeaugh-Geiss et al.'s (2018) suggests an illuminating answer to what the relevant factor may be instead, and one that also accounts for a difference between our 5- and 6-year-olds. In terms of their approach, children's non-exhaustive clefted focus interpretations may be due to their inability to identify a maximal discourse referent associated with the background, or, in terms of QUDs, a maximal QUD. Indeed, Roeper et al. (2007) found that young children interpret questions as non-maximal, and start interpreting them as maximal only at 6-7 years. This ties in with a marked difference between 5-year-olds (n=16) and 6-year-olds (n=16) in Subexp2: the presence of the question significantly raised exhaustive responses in the latter, but not in the former age group.

Sample item of Subexp1

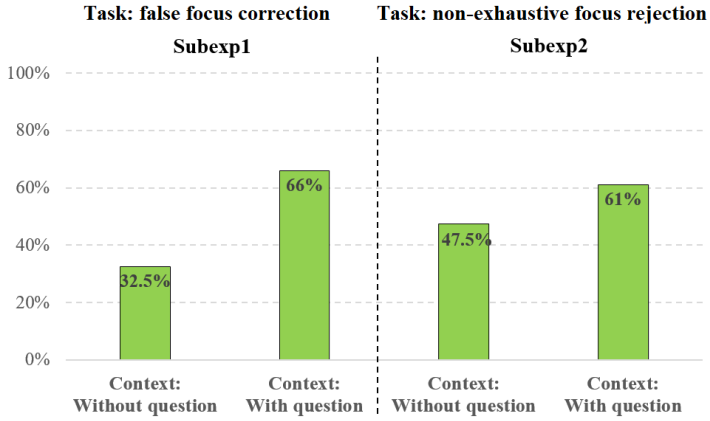
[KI]$_{FOC}$ emelte fel a teknős-t?
who lifted PRT the turtle-acc
'WHO lifted the turtle?'

[A KROKODIL]$_{FOC}$ emelte fel a teknős-t.
the crocodile lifted PRT the turtle-acc
'It is the crocodile who lifted the turtle.'

Sample item of Subexp2

[KI]$_{FOC}$ fogott ki egy halacská-t?
who caught PRT a fish-acc
'WHO caught a fish?'

[A KISMACKÓ]$_{FOC}$ fogott ki egy halacská-t.
the bear caught PRT a fish-acc
'It is the bear who caught a fish.'



Results (children)

Significant effects (GLMM):

- CONTEXT: $\chi^2(1) = 40.99$, $p < 0.001$
- **CONTEXT * TASK interaction**: $\chi^2(1) = 9.23$, $p = 0.002$

References De Veaugh-Geiss, J.P., S. Tönnis, E. Onea, & M. Zimmermann (2018) That's not quite it: An experimental investigation of (non-)exhaustivity in clefts. *Semantics & Pragmatics,* 11, Art. 3. // Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Lang Learning and Development* 8 // Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Lang & Cogn Proc 20.* // Heizmann, T. (2012). *Exhaustivity in questions & clefts; and the quantifier connection: A study in German and English.* PhD diss., Amherst. // Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Lang Acq* 12. // Szendrői, K., Bernard, C., Berger, F., Gervain, J., & Höhle, B. (2018). Acquisition of prosodic focus marking by English, French, and German 3-, 4-, 5- and 6-year-olds. *Journal of Child Lang* 45. // Tieu, L., & Križ, M. (2017). Connecting the exhaustivity of clefts and the homogeneity of plural definite descriptions in acquisition. In M. LaMendola, & J. Scott (eds.), BUCLD 41.

Conceptual Foundations of Telicity: Viewers' Spontaneous Representation of Boundedness in Event Perception

Yue Ji[1], Anna Papafragou[2]
[1] Beijing Institute of Technology, [2] University of Pennsylvania

Foundational semantics literature distinguishes between *telic* verb phrases denoting *bounded* events with an inherent endpoint (e.g., *fix a car*) and *atelic* verb phrases denoting *unbounded* events that lack an inherent endpoint (e.g., *drive a car*; Bach, 1986, Krifka, 1998). Telicity is frequently assumed to build on conceptual notions (Filip, 1993; Ji & Papafragou, 2020), but little research has explored sensitivity to a cognitive bounded-unbounded distinction. Here we fill this gap. Building on the finding that endpoints are critical components in both memory and language (e.g., Lakusta & Landau, 2012; Gold et al., 2017; Papafragou, 2010), we hypothesize that the salience of endpoints should only characterize bounded events; in unbounded events, endpoints should be treated largely similarly to other time points. To test this hypothesis, we inserted a brief interruption into videos that were biased towards a bounded vs. unbounded event construal. Viewers of bounded events would be more likely to neglect an interruption close to the endpoint since the developments near the endpoint would draw their attention and the external interruption would be missed. For viewers of unbounded events, the placement of the interruption should not make a difference: these events do not have canonical endpoints - they stop, but do not culminate.

We created 20 pairs of videos containing events that encouraged either a bounded or an unbounded construal (see Figure 1). These construals were confirmed in a norming study where "bounded" videos were more likely to depict "something with a beginning, midpoint and specific endpoint" than unbounded ones. Each video was then edited to place a visual interruption of .03s at the temporal point corresponding to either 50% of the video (mid-interruption) or 80% of the video (late-interruption). In Exp.1, 64 adults watched 10 test videos drawn from either the Bounded or the Unbounded construal group, half with a mid-interruption and half with a late-interruption (along with 10 filler videos without any interruption) and indicated whether they detected an interruption after watching each video. A significant interaction between Interruption Placement (Mid vs. Late) and Event Construal (Bounded vs. Unbounded) was found (z=2.70, p =.007; Figure 2a). As expected, participants processing bounded event representations had more difficulty detecting late-interruptions (M=79.7%) compared to mid-interruptions (M=95.3%; z=-3.53, p<.001), but this difference disappeared among viewers representing unbounded events (for late-interruptions, M=95.8%; for mid-interruptions, M=93.8%; p > .581). Exp.2 was identical but participants had to press a key as soon as they detected an interruption during a video. An analysis of response times revealed an interaction between Interruption Placement and Event Construal (t=-1.97, p=.049; Figure 2b). Participants watching videos construed as bounded events had longer response times for late-interruptions (M=882 ms) compared to mid-interruptions (M=760 ms; t=5.27, p<.001) but the difference was smaller for unbounded events (for late-interruptions, M=710 ms; for mid-interruptions, M=669 ms; t =3.10, p =.002).

Together, our data show that viewers spontaneously compute boundedness, or the temporal texture of dynamic events, during event perception. This finding supports the homology between aspect and event cognition and speaks to the language-cognition interface.

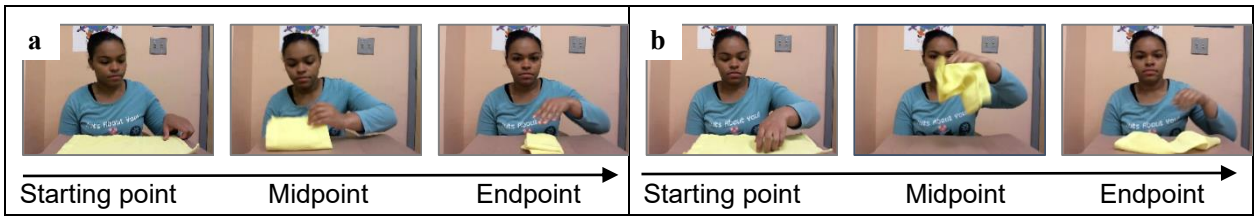


Figure 1. Examples of (a) a bounded construal (fold up a handkerchief), (b) an unbounded construal (wave a handkerchief).

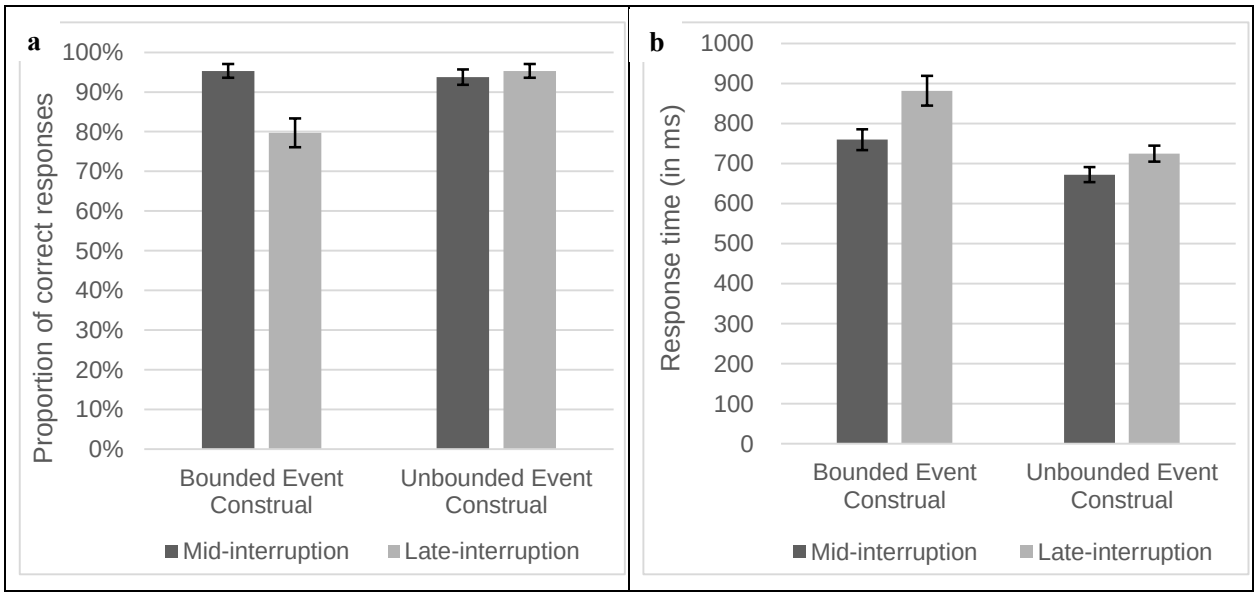


Figure 2. (a) Proportion of correct responses in Experiment 1. Error bars represent ±SEM. (b) Response time (in ms) for correctly identifying an interruption in Experiment 2. Error bars represent ±SEM.

References

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy, 9,* 5-16.

Filip, H. (1993). *Aspect, situation types and nominal reference* (Unpublished doctoral dissertation). University of California at Berkeley, Berkeley, CA.

Gold, D., Zacks, J., & Flores, S. (2017). Effects of cues to event segmentation on subsequent memory. *Cognitive Research: Principles and Implications*, *2*, 1–15.

Ji, Y., & Papafragou, A. (2020). Is there an end in sight? Viewers' sensitivity to abstract event structure. *Cognition, 197,* 104197.

Krifka, M. (1998). The origins of telicity. In S. Rothstein (ed.), *Events and Grammar*. Dordrecht: Kluwer.

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal. *Cognitive Science*, *36*, 517-544.

Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science, 34,* 1064-1092

Far from independent: Matrix-driven temporal shift interpretations of English and German past-under-past relative clauses

Elena Marx & Eva Wittenberg, *Central European University, Vienna, Austria*

Complex sentences allow speakers to describe multiple events, and express relations between them: In “the girl kissed the boy who was next to the traffic light”, the kiss and the boy’s situation relate to each other. One dimension of this relationship is temporal. In two pre-registered studies using English and German, we investigate which interpretations are available for past-under-past relative clauses: is the sentence true if the girl kisses the boy after (Fig.1A, back-shifted) or before (Fig. 1B, forward-shifted) he is next to the traffic light?

Formal accounts of embedded tense conceive of the interpretation of tense in relative clause as only dependent on utterance time, not on the matrix clause’s tense [1-3]. Therefore, these approaches predict that back-shifted (1A) and forward-shifted (1B) interpretations for the situation described in a relative clause (*standing next to a traffic light*), relative to a matrix event (*being kissed*), should be acceptable: Both are past relative to utterance time.

Semantically however, relative clauses can be conceptualized as anchored to a main event [4-5]: the relative clause tense (“who was next to the traffic light”) is interpreted relative to that anchoring event’s tense (“kissed”). This account predicts forward-shift interpretations (1B) in past-under-past relative clauses to be inacceptable because here, the embedded past tense describes a situation (*standing next to a traffic light*) that happens after the anchor (*kissing*).

Conceptually building on a French acquisition study [6], in **Exp.1**, 50 native English speakers watched 8 clips (+20 fillers) of events like 1A and 1B, then read descriptions (like in Fig.1), and rated whether they matched the movie. *While*-clauses (Fig.1), which should always be rated unacceptable, served as controls. **Results** (statistical analysis: mixed-effects models, significance assessed using pairwise model comparisons): As predicted, *while*-descriptions were rated worse than relative clause descriptions (Fig. 2, left panel, main effect of clause-type: Df=1, χ^2=129.5, p<0.001). There was also a main effect of shift: forward shifts were less acceptable than backward shifts (Df=1, χ^2=29.4, p<0.001). Crucially, the interaction between clause-type and shift-type was significant (Df=1, χ^2=17.65, p<0.001), and pairwise comparisons revealed that it was driven by the relative clauses: While *while*-clauses were unacceptable across shift-types, (β=-.08, t=-0.9, p>0.37), participants rated sentences containing relative clauses significantly higher when they described back-shifted clips compared to forward-shifted clips (β=-.06, t=-5.9, p<0.001). **Exp.2**: replication in German (N=18), resulting in a similar pattern of results (Fig.2, right panel: main effect of clause-type: Df=1, χ^2=28.1, p<0.001, main effect of shift-type: Df=1, χ^2=10.9, p<0.001, interaction: Df=1, χ^2=10.4, p<0.01), with the interaction driven by the relative clause (*while*-clauses: β=-.001, t=-0.17, p>0.87, relative clauses: β=-.07, t=-3.7, p<0.001).

Our results indicate that tense interpretation in relative clauses is dependent on the matrix clause – at least when the matrix sentence describes a salient anchoring event, and the relative clause a backgrounded situation. Further experiments will assess whether the same pattern holds when this mapping between syntax and semantics is switched (“The boy who the girl kissed stood next to the traffic light”), and whether forward shifts are ameliorated in contexts with discourse focus on relative clauses (“This story is about a boy”). Whereas none of these manipulations are predicted to change interpretations under a syntactic account, a semantic account [5] predicts an amelioration of forward shifts for discourse-salient relative clauses as they can serve as conceptual anchors for past-under-past sentences.

Overall, our findings provide insight into the representation of events and how temporal semantic features are linked to main and dependent clauses. Using language to describe mental representations is a selective process in which speakers must decide which information they want to communicate, and choose their expressive means accordingly. In this regard, complex sentences convey a temporal perspective on event structure which is not solely determined by grammatical principles. By contrast, our results suggest that speakers take factors of event structure into consideration when they map temporal relations onto linguistic structure.

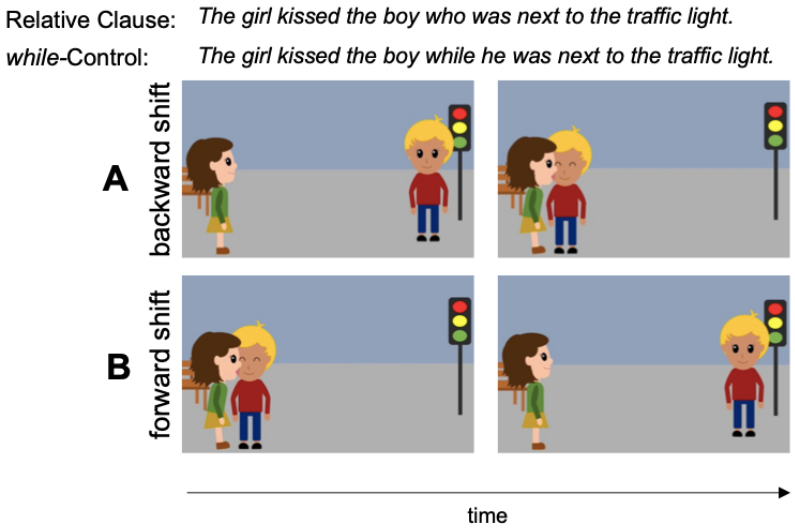


Figure 1. Temporal arrangements of the embedded situation (*standing next to a traffic light*) relative to the main event (*kissing*) in the video clips. Videos under https://osf.io/6ae5m/?view_only=fa7a501f340d4a538ee604c3faa3be7c .

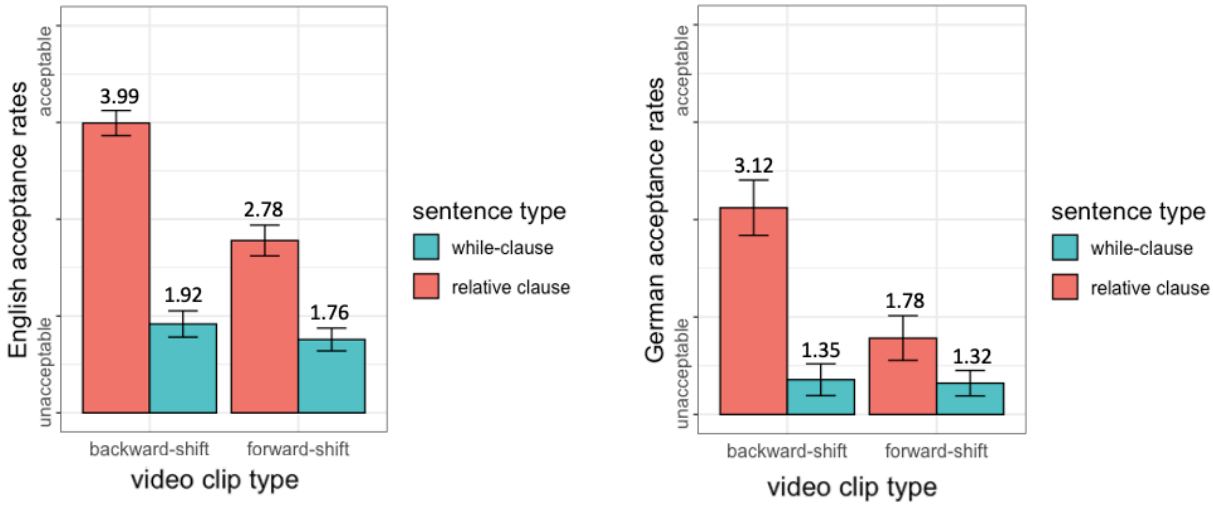


Figure 2. Mean ratings in Experiment 1 (left, English) and Experiment 2 (right, German). Error bars denote Standard Errors.

References

[1] Abusch, D. (1997). Sequence of Tense and Temporal De Re. *Linguistics and Philosophy*, *20*, 1–50.
[2] Ogihara, T. (1995). The Semantics of Tense in Embedded Clauses. *Linguistic Inquiry*, *26*(4), 663–679.
[3] Stowell, T. (2007). The syntactic expression of tense. *Lingua*, *117*(2), 437–463.
[4] Carroll, M., Stutterheim, C. von, & Klein, W. (2003). Two ways of construing complex temporal structures. In F. Lenz (Ed.), *Deictic conceptualisation of time, space and person* (pp. 97–133).
[5] Klein, W. (2000). An analysis of the German Perfekt. *Language*, *76*(2), 358–382.
[6] Demirdache, H., & Lungu, O. (2008). Sequence of tense in (French) child language. *Linguistic Variation Yearbook*, *8*(Ea 3827), 101–130.

When Transformer models are more compositional than humans: The case of the depth charge illusion

Dario Paape (University of Potsdam)

English native speakers often interpret the sentence *No head injury is too trivial to be ignored* to mean that head injuries, even seemingly trivial ones, should never be ignored. However, this interpretation is not compositionally licensed: The embedded degree phrase is internally incongruous (sensible: *too serious to be ignored*), and the verb *ignored* should not be negated (cf. *No missile is too small to be banned*) [1]. Nevertheless, participants complete the sentence *No head injury is too trivial to be* ___ with the verb *ignored* or a semantically similar continuation about 80% of the time [2]. This effect is known as the "depth charge" illusion. Among the proposed explanations for the illusion are processing errors and superficial interpretation [1,2], pragmatic inference about the intended meaning [3], and the existence of a stored, non-compositional grammatical template [4].

An interesting question to ask is whether the illusion also appears in giant Transformer-based language models like `GPT-3` and `BERT` [5,6]. Transformer models show an impressive ability to generate coherent text, but struggle with complex grammatical structures [7] and semantic mechanisms such as negation and entailment [8]. For instance, `BERT` will produce the word *Apple* with equal probability in the sentence *iOS is developed by* ___ compared to the sentence *iOS is not developed by* ___ [9]. In order to provide compositional completions for depth charge sentences, a Transformer model would need to identify the scope of the negation, as well as its interaction with the degree phrase *too trivial to X*. Given their limitations and partial reliance on heuristics [10], Transformers could show a stronger depth charge illusion than humans. On the other hand, Transformers do not process sentences incrementally; they can use all the information in the sentence in parallel [11]. This may give them a compositional advantage over humans: It has been suggested that human compositional processing is foiled in depth charge sentences due to the incremental combination of *no* and the second negative element *too*, which masks the incongruity [2,4]. In sum, Transformers may behave differently from humans with regard to the illusion, but the direction is not clear.

We conducted an experiment with four giant Transformer models: Two versions of `GPT-3` (`ada` with 2.7 billion parameters and `davinci` with 175 billion parameters), `Jurassic-1-Jumbo` (175 billion parameters) [12], and `RoBERTa`, which is `BERT` with additional training (125 million parameters) [12]. The input items were 32 depth charge sentences that have previously been used with human participants [2,3]. We included a control condition with *some* instead of *no*, which reduces the illusion to about 10% in humans [2], and a condition with *enough* instead of *too*, which allows for a sensible compositional interpretation. A variety of additional controls were tested to check whether the models are sensitive to negation and the meaning of degree constructions, as shown in **Table 1**. The dependent variable is the log probability of the verb (e.g., *ignored*) in each sentence.

As shown in **Figure 1**, the Transformer models show a higher log probability for *ignored* in sentences with *no* than in sentences with *some*, similar to humans. This is despite the fact that the models have apparently encoded the necessary knowledge to handle the construction: The control conditions all show behavior that is consistent with compositionality. However, when looking at actual sentence completions generated by Transformers, patterns emerge that set them apart from humans. First, they often fail in the control conditions, producing transparently incongruous sentences (see examples in **Table 2**). Second, they tend to produce many compositional continuations for depth charge sentences. For instance, in the negated *head injury* item (1b) in **Table 1**, `RoBERTa` ranks the verbs *addressed* (14%), *treated* (9%) and *considered* (7%) higher than the verb *ignored* (5%).

Taken together, the results show that Transformer models exhibit human-like behavior in that they fall for the depth charge illusion, but also suggest that Transformers may be more compositional than humans in cases where incremental processing creates a bottleneck of complexity.

(1) (a) No head injury is trivial enough to be ignored. ☑ (compositionally sensible)
(b) No head injury is too trivial to be ignored. ☒ (depth charge)
(c) Some head injuries are too trivial to be ignored. ☒ (not compositionally sensible)
(d) No head injury is so trivial as to be ignored. ☑
(e) No head injury is so trivial as to not be ignored. ☒
(f) Head injuries that are too trivial will be ignored. ☑
(g) Head injuries that are not too trivial will be ignored. ☒
(h) Head injuries that are trivial are more likely to be ignored. ☑
(i) Head injuries that are trivial are less likely to be ignored. ☒

Table 1. Example item showing the constructions tested in the experiment. 32 different sentences were used.

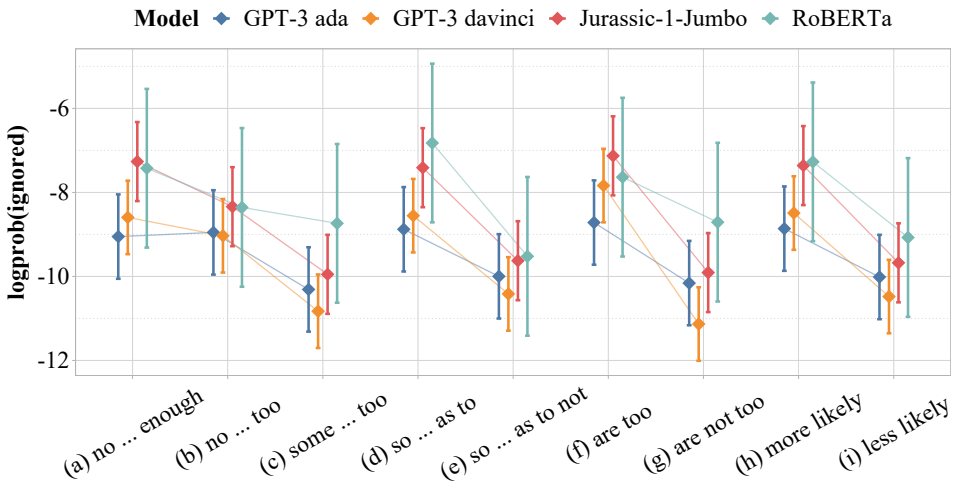


Figure 1. Log probability of the critical verb (e.g., *ignored*) by construction and model.

GPT-3 ada

No head injury is too trivial to be counted as a crime. ☑ (compositional)
Some head injuries are too trivial to be taken lightly. ☒ (non-compositional)
Head injuries that are trivial are more likely to be fatal. ☒
Head injuries that are trivial are less likely to be fatal. ☑

GPT-3 davinci

No head injury is too trivial to be ignored. Any recent head injury, no matter how minor, should be included in the patient's history. ☒
Some head injuries are too trivial to be treated, Dr. Benson acknowledged. ☑

Jurassic-1-Jumbo

No head injury is too trivial to be noticed by a parent. ☑
No head injury is too trivial to be ignored. All head injuries need to be taken seriously. ☒

RoBERTa

Head injuries that are too trivial will be punished. ??
Some head injuries are too trivial to be ignored. ☒

Table 2. Example completions by model.

References. [1] Wason & Reich (1979, Q J Exp Psychol). [2] Paape et al. (2020, J Semant). [3] Zhang et al. (2021, AMLaP presentation). [4] Fortuin (2014, Cogn Linguist). [5] Brown et al. (2020, arXiv:2005.14165). [6] Devlin et al. (2018, arXiv:1810.04805v2). [7] van Schijndel et al. (2018, arXiv:1909.00111). [8] Hossain et al. (2020, Proc EMNLP). [9] Hosseini et al. (2021, arXiv:2105.03519). [10] McCoy et al. (2019, arXiv:1902.01007). [11] Kahardipraja et al. (2021, arXiv:2109.07364). [12] Lieber et al. (2021, white paper, AI21 labs). [13] Liu et al. (2019, arXiv:1907.11692).

Transparency in the Processing of Temporal Ambiguity: The Case of Embedded Tense

Giuliano Armenante[1,2] & Vera Hohaus[1,3] & Britta Stolterfoht[1]
1 Eberhard Karls Universität Tübingen, 2 Universität Potsdam, 3 Leibniz-Zentrum für allgemeine Sprachwissenschaft, Berlin

Summary. We report the results of one acceptability rating study and two self-paced reading studies on the form-meaning mismatch in the interpretation of past-under-past in complement clauses in English. Across the three experiments, we find an offline and online preference for the backward-shifted interpretation, in line with predictions of the structural approach to the ambiguity when assuming a processing preference for morphological transparent interpretation.

Background. In English, embedded tenses in certain configurations give rise to ambiguities (but see Altshuler & Schwarzschild 2013, Altshuler 2016), as in the case of past tense in a stative complement clause (CP) embedded under a past-marked verb of reported speech, in (1). Structural approaches to the ambiguity (prominently, Ogihara 1996) derive SIM from the Logical Form for BACK by additional morpho-syntactic technology, such as a deletion operation like (2) that is licensed in this kind of configuration.

(1) *Oliver said*$_{\text{past}}$ [$_{\text{CP}}$ *that Amber* **[$_{\text{AUX}}$ *was*$_{\text{past}}$] [$_{\text{ADJ}}$ *sick*]**]
a. Oliver said: “Amber is sick.” (simultaneous reading, SIM)
b. Oliver said: “Amber was sick.” (backward-shifted reading, BACK)

(2) a. SIM-LF: [PAST$_{t^*}$ [λt Oliver say$_{w@,t}$ [λw ~~$\lambda t'$ PAST$_{t'}$~~ [$\lambda t''$ John sick$_{w,t''}$]]]]
b. BACK-LF: [PAST$_{t^*}$ [λt Oliver say$_{w@,t}$ [λw $\lambda t'$ PAST$_{t'}$ [$\lambda t''$ John sick$_{w,t''}$]]]]
with t^* the utterance time and $w_@$ the actual world

Previous experimental findings are overall inconclusive (in particular, Dickey 2001). The data from one of the adult-control groups in Hollebrandse (2000) suggests a slight acceptability preference for SIM; Gennari (2004) observes an advantage for overlapping temporal intervals in reading times, but employs a design that relies on additional manipulations.

Experimental Hypotheses. The three experiments Exp1-3 reported below investigate two competing processing hypotheses H1 and H2 derived from structural approaches, WYSIWYG and Structural Simplicity. Under H1, comprehension is driven by morphological transparency, and an embedded past tense in a configuration like (1) should initially always be interpreted as such, favouring BACK. Under H2, comprehension is driven by structural simplicity at Logical Form (also keeping the number of times in the semantic representation low), favouring SIM.

Experiments 1 and 2 both adopted the same 2x2 design, with factors EvalT (past vs future) and interpretation (BACK vs SIM). Participants (N_{Exp1}=43, N_{Exp2}=40) saw 88 trials, consisting of 40 experimental items embedded within 40 fillers. Each trial had a context picture of the type in Fig.1 establishing the EvalT and the intended reading, followed by a sentence like (1), which was presented word-by-word in Exp1, and rated on a scale from 1-6 in Exp2. H1 predicts an acceptability preference for BACK over SIM, which should also be reflected in longer reading times for SIM over BACK on the embedded auxiliary AUX or adjective ADJ in (1). H2 predicts a preference and reading time advantage for SIM over BACK.

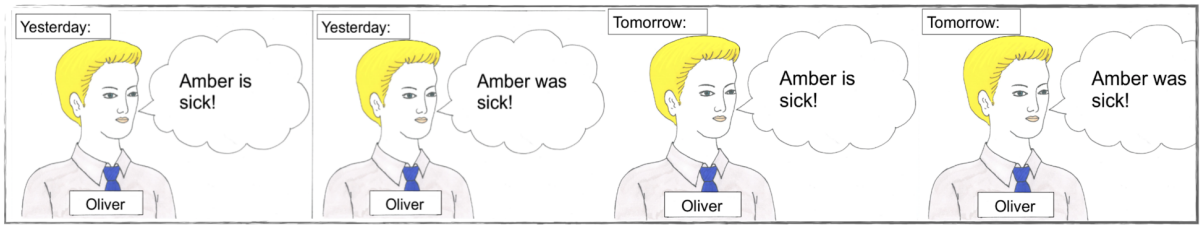


Figure 1: Sample context pictures used in the Self-Paced Reading Experiment (Exp1) and the Rating Study (Exp2).

The results from Exp1 show no significant reading-time difference between BACK and SIM for past context at AUX, but a marginal effect at ADJ ($t(42) = 1.939$, $p = .059$), with longer reading times for SIM; see also Fig.2 below. In Exp2 we observed higher ratings for BACK compared to SIM ($M_{BACK,fut} = 5.54$, $M_{SIM,fut} = 4.16$, $M_{BACK,past} = 5.98$, $M_{SIM,past} = 5.38$). For the control conditions (BACK,fut vs SIM,fut) this is expected, since SIM readings are unavailable when embedded under a matrix predicate that is not in the past tense. For the experimental conditions, this statistically significant difference ($t(39) = 4.921$, $p < .001$) supports the WYSIWYG hypothesis.

Experiment 3 relied on disambiguation by continuation rather than context. A 2x2 design was adopted, with factors SoT (BACK vs SIM) and ambiguity (+amb vs –amb). Participants (N=68) saw 64 trials, consisting of 16 experimental items with 48 fillers. Experimental trials such as [BACK,+amb] in (3) involved a context sentence establishing two time intervals, followed by the target, where the continuation disambiguated the locally ambiguous embedded past. Assuming incremental processing, H1 predicts longer reading times for SIM over BACK for the critical region, bolded in (3), and potentially the spillover region. H2 predicts that BACK forces a revision of a previously assigned SIM, resulting in longer reading times in those regions.

(3) Context: After last week's final rehearsal,
last night, John's band finally gave a concert, where I spoke to him about Mary…

Target: *John* | *said* | *that* | *Mary* | *was sick*, |
so | *that's why* | *she* | *missed* | ***the concert***$_{SIM}$ / ***the rehearsal***$_{BACK}$ | *with* | *great* | *regret*.

A repeated-measures ANOVA in SPSS reveals a marginally significant interaction between SoT and amb for the critical region ($F_{1,67} = 3.127$, $p = .082$, $\eta^2 = .045$), resulting from longer reading times for SIM as opposed to BACK against the [-amb] baseline condition (see Fig.2).

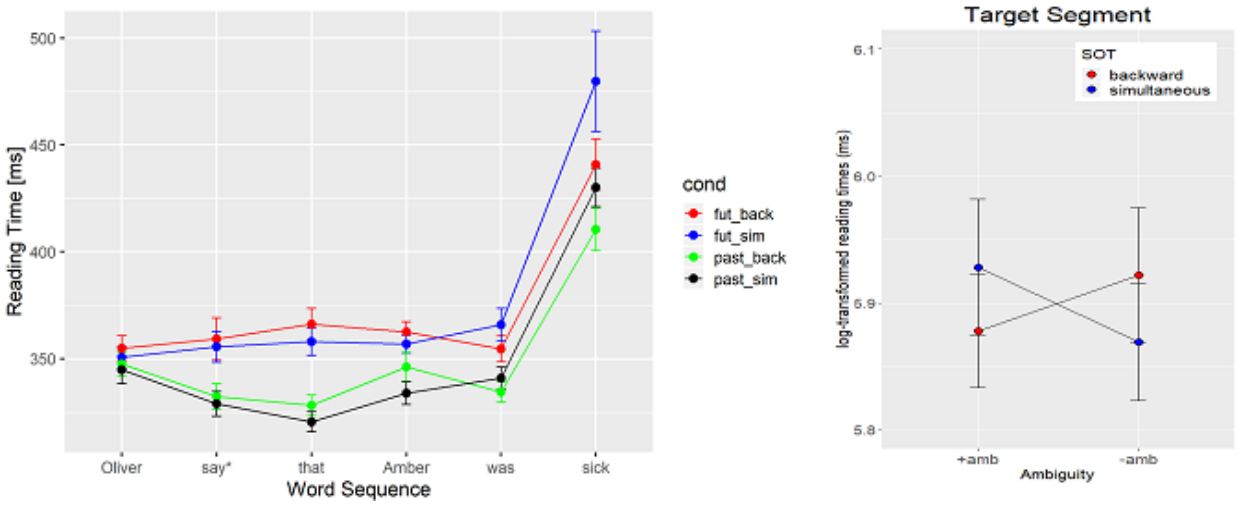


Figure 2: Reading-Time Results from Experiments 1 and 3.

Discussion. We find an acceptability preference for BACK over SIM, which is also reflected in the reading times in Exp3, contra some of the findings in the previous research literature. Taken together, Exp1-3 provide preliminary but converging evidence in favour of the WYSIWYG hypothesis and a processing strategy that is guided by morphological transparency, rather than Structural Simplicity. While further research into the processing of the sequence of tenses is needed, these findings are also relevant for other cases of ambiguity processing that involve form-meaning mismatches (in the interpretation of phi-features of pronouns, for instance).

References. D. ALTSHULER & R. SCHWARZSCHILD (2013), "Moment of Change, "Cessation Implicatures and Simultaneous Readings", *SuB Proceedings* 17, 45-62. ::: D. ALTSHULER (2016), *Events, States and Times* (Berlin: De Gryuter). ::: M. W. DICKEY (2001), *The Processing of Tense* (Dordrecht: Kluwer). ::: S. P. GENNARI (2004), "Temporal References and Temporal Relations in Sentence Comprehension", *Journal of Experimental Psychology 30*(4), 877-890. ::: B. HOLLEBRANDSE (2000), "The Acquisition of Sequence of Tense", UMass dissertation. ::: T. OGIHARA (1996), *Tense, Attitudes, and Scope* (Kluwer: Dordrecht).

Effects of referent lifetime knowledge on processing of verb morphology

Daniela Palleschi[1,2,3], Camilo Rodríguez Ronderos[1,4], Pia Knoeferle[1,2,3]
[1]Humboldt-Universität zu Berlin, [2]Einstein Center for Neurosciences Berlin, [3]Berlin School of Mind and Brain, [4]University of Oslo
daniela.palleschi@hu-berlin.de

Background In the 'Perfect' Lifetime Effect, experiential readings of the English Present Perfect are felicitous with a living referent, but not dead referent (ex. 1; Klein, 1992; Meyer-Viol, 2011; Mittwoch, 2008). Such predicates in the Past Simple would be felicitous with the dead, but "odd" with the living if no completed past reference time is defined (ex. 1b; Partee, 1984). Meanwhile, processing of the English Present Perfect has been shown to be influenced by lexically defined time reference (Roberts & Liszka, 2013) as well as a visually depicted scene (Altmann & Kamide, 2007). In two experiments, we explored the processing of lifetime-tense congruence in the Present Perfect and Past Simple as well as the influence of the source of lifetime information by manipulating the presence of long-term knowledge of a referent.

Present Study We presented participants with lifetime context sentences defining the lifetime of referents who are well-known (Experiment 1; ex. 2) or unknown (Experiment 2; ex. 3), thereby manipulating the presence of long-term knowledge. This was followed by critical sentences describing an accomplishment of this person in the Present Perfect (ex. 3a) or the Past Simple (ex. 4b). Our stimuli contained two two-level factors (*tense*: Present Perfect (PP), Past Simple (PS); *lifetime congruence*: congruent, incongruent). The *congruent* conditions were *living-PP* and *dead-PS*, and *incongruent* conditions were *dead-PP* and *living-PS*.

Procedure Lifetime context sentences and critical sentences were presented to native British speakers (n = 160/experiment) in two cumulative self-paced reading experiments. Each trial was followed by a binary naturalness judgement task. Within each experiment, longer reading times and lower proportions of acceptances were expected for the *incongruent* conditions (dead-PP, living-PS), reflecting processing costs and awareness of the violations, with larger nested *lifetime congruence* effects expected for the Present Perfect than the Past Simple. If the presence of long-term knowledge in addition to contextually defined lifetime strengthens activation of the temporal (lifetime) constraints, then earlier and/or stronger effects of lifetime congruence would emerge in Experiment 1 compared to Experiment 2. Linear mixed models were fitted to reading-time data, and genearlised linear mixed models to binary response data. Self-paced reading time results were corrected for multiple comparisons (p-values multiplied by 5, the number of regions analysed per experiment).

Results In Experiment 1, a main effect of lifetime congruence emerged in naturalness responses (Fig. 1; z = -12.6, p < .001), total sentence reading times (Fig. 2; t = *7.2*, p < .001;), and self-paced reading times from the *verb+1* region (Fig. 3; **verb+1**: t = 2.9, p < .05; **verb+2**: t = 3.6, p < .01; **verb+3**: t = 4.6, p < .001; **verb+4**: t = 7.8, p <.001). An interaction effect of lifetime congruence and tense was found, with larger nested lifetime effects for the Present Perfect than Past Simple in total sentence reading times (**PP**: t = 7.5, p < .001; **PS**: t = 3.99, p < .001) and self-paced reading times from the verb+3 region, with significant nested effects in the Present Perfect only (**verb+3**: t = 5.0, p < .001; **verb+4**: t = 8.5, p < .001). In Experiment 2, a main effect of lifetime congruence was likewise found in naturalness responses (Fig. 1; z = -10.1, p < .001), total sentence reading times (Fig. 2; t = 6.5, p < .001), and self-paced reading times from the *verb*+3 onward (Fig. 3; **verb+3**: t = 3.6, p < .01; **verb+4**: t = 5.3, p < .001), with no interaction effects.

Conclusion The earlier emergence of main lifetime congruence effects in Experiment 1 compared to Experiment 2 suggests that the dual presence of both long-term and contextually defined lifetime information strengthened the activation of the temporal lifetime constraints. The additional finding of an interaction effect in Experiment 1 reading times, with a larger effect of lifetime congruence in the Present Perfect compared to Past Simple, implies a larger cost for integrating the Present Perfect in a completed past time frame than for integrating the Past Simple in an on-going time frame, similar to findings in Roberts and Liszka (2013). Taken together, these results suggest that temporal constraints on the English Present Perfect and Past Simple extend to referent lifetime during incremental processing, and that the source of lifetime information influences the temporal emergence of effects.

Example sentences

1a	Einstein visited/*has visited Princeton.	*dead*
1b	Chomsky ?visited/has visited Princeton.	*living*
2a	Beyoncé is an American performer. She lives in California.	*famous - living*
2b	Whitney Houston was an American performer. She died in California.	*famous - dead*
3a	Sophie Laverty is an American performer. She lives in California.	*unknown - living*
3b	Sophie Laverty was an American performer. She died in California.	*unknown - dead*
4a	She *has performed* in many arenas, according to Wikipedia.	*Present Perfect*
4b	She *performed* in many arenas, according to Wikipedia.	*Past Simple*

Figures

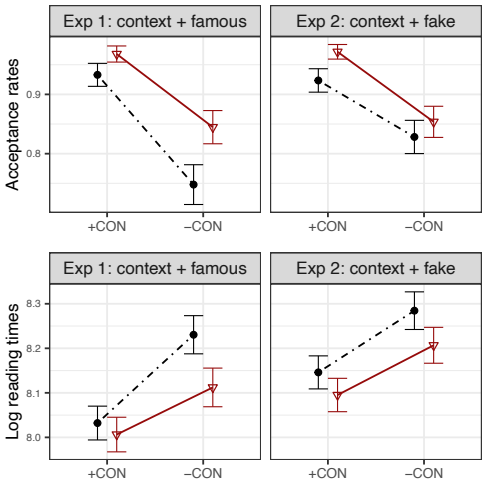


Figure 1 *(top row): naturalness acceptance rates for Experiments 1 and 2* (+CON = *congruent,* -CON = *inconguent*)
Figure 2 (bottom row): *mean total reading times for critical sentences; Figure 3 legend applies*

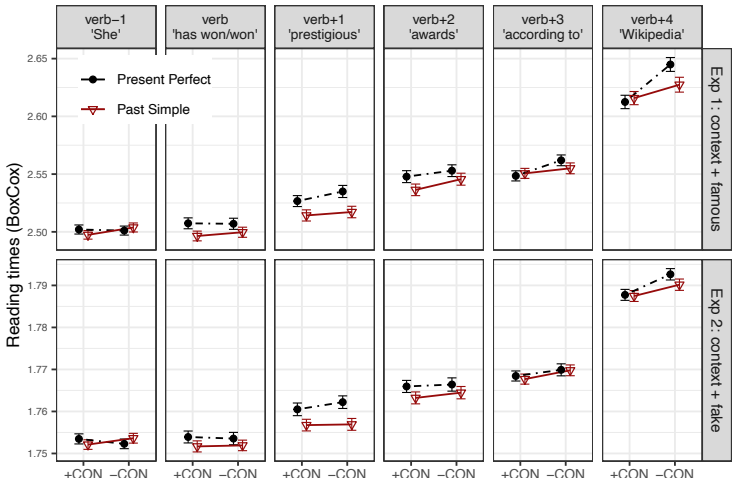


Figure 3*: BoxCox transformed reading times across sentence regions for Experiemnt 1 (top row) and Experiment 2 (bottom row); lifetime congruence is plotted on the x-axis* (+CON = *congruent,* -CON = *inconguent)*

References

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, 57, 502–518. https://doi.org/10.1016/j.jml.2006.12.004

Klein, W. (1992). The present perfect puzzle. *Language*, 68(3), 525–552.

Meyer-Viol, W. P. M. (2011). Reference time and the English past tenses Author. *Linguistics and Philosophy*, 34(3), 223–256.

Mittwoch, A. (2008). The English Resultative perfect and its relationship to the Experiential perfect and the simple past tense. *Linguistics and Philosophy*, 31(3), 323–351.

Partee, B. H. (1984). Nominal and temporal anaphora. *Linguistics and Philosophy*, 7(3), 243–286. https://doi.org/10.1007/BF00627707

Roberts, L., & Liszka, S. A. (2013). Processing tense/aspect- agreement violations on- line in the second language: A self-paced reading study with French and German L2 learners of *English. Second Language Research*, 29(4), 413–439. https://doi.org/10.1177/0267658313503171

Amusing or aggressive? A cross-cultural study in sarcasm interpretation and use

Ning Zhu & Ruth Filik, School of Psychology, University of Nottingham

There is some debate regarding whether sarcasm mutes the negativity of criticism [1] or enhances condemnation [2]. Previous research suggests that the emotional impact of sarcasm may depend on the perspective taken by the rater [3]. However, the findings are mixed so far. There is also some evidence that, besides linguistic factors, individual differences factors influence sarcasm interpretation and use (e.g., [4]). For example, research in children suggests that theory of mind ability (ToM) might be associated with sarcasm interpretation [5]. However, studies examining neurotypical adults' ToM and sarcasm comprehension are relatively rare. While some studies suggest that there might also be cultural differences in sarcasm interpretation and use [6] [7] [8], relatively little is known about how these differ across Western and Eastern cultures. To address these gaps in the literature, the present study investigated individual differences in sarcasm interpretation and use in participants in the UK and China.

Experiment 1 (with UK participants) had a 2 (comment type: literal, sarcastic) * 3 (perspective: speaker, recipient, reader) within-subjects design. We created 48 experimental scenarios in six conditions (see Table 1 for an example) and combined them with 16 filler scenarios. We collected ratings on sarcasm, aggression, amusement, and politeness of the target comment. We also examined effects of ToM (assessed by the Faux Pas test [9]), empathy (assessed by the interpersonal reactivity index [10]), and sarcasm use tendency (indicated by scores on the sarcasm self-report scale [4]) in sarcasm interpretation, and effects of ToM and empathy in sarcasm use. Experiment 2 was a replication of Experiment 1, but with Chinese participants.

We used linear mixed models in R to analyse the rating data, with *comment type* and *perspective* as fixed factors, and intercepts and slopes for all the fixed effects (including interactions) across participants and scenarios as random effect structure [11]. We conducted two-tailed Pearson correlations to assess the relationship between individual differences factors (e.g., ToM) and the rating measures. We used independent samples *t*-tests to examine cultural differences across the UK and China.

Key results from Experiment 1 showed that UK participants rated sarcasm as being more amusing and polite than literal criticism, supporting the Tinge hypothesis [1], which suggests that sarcasm mutes the negativity of criticism. Theory of mind ability positively predicted sarcasm use and interpretation (in ratings of sarcasm and amusement). Sarcasm use tendency had positive correlations with ratings of amusement and politeness, and negative correlation with ratings of aggression. Key results from Experiment 2 showed that Chinese participants rated sarcasm as being more amusing, but also more aggressive than literal language. Theory of mind ability positively predicted sarcasm interpretation (in ratings of sarcasm). Sarcasm use tendency had positive correlations with ratings of amusement and politeness, and also ratings of sarcasm. Compared with UK participants, participants from China rated sarcasm as being more aggressive and less amusing and they were less likely to use sarcasm in daily life.

We found that sarcasm interpretation and use tendency varied across cultures. Whereas Western participants tended to consider sarcasm as amusing, participants from Eastern cultures tended to view sarcasm as also being aggressive, which in turn affects their use of sarcasm. In relation to theoretical accounts, that is, whether sarcasm mutes the negativity of criticism [1] or enhances condemnation [2], we suggest that it may depend on the cultural background of the perceiver. Thus, the Tinge Hypothesis [1] may need to be modified to take culture into account. Practical implications of the findings include the need for speakers to consider the recipients' cultural background when using sarcasm, in order to avoid confusion over speaker intent.

Table 1

Example Experimental Scenario in All Conditions

Condition		Scenario
Perspective-speaker	Literal	You were building a very complicated structure out of Lego. Person B came over to help. Unfortunately, Person B unintentionally knocked some of it down. You said to Person B: 'You are a bad helper.'
	Sarcastic	You were building a very complicated structure out of Lego. Person B came over to help. Unfortunately, Person B unintentionally knocked some of it down. You said to Person B: 'You are a good helper.'
Perspective-recipient	Literal	Person A was building a very complicated structure out of Lego. You came over to help. Unfortunately, you unintentionally knocked some of it down. Person A said to you: 'You are a bad helper.'
	Sarcastic	Person A was building a very complicated structure out of Lego. You came over to help. Unfortunately, you unintentionally knocked some of it down. Person A said to you: 'You are a good helper.'
Perspective-reader	Literal	Person A was building a very complicated structure out of Lego. Person B came over to help. Unfortunately, Person B unintentionally knocked some of it down. Person A said to Person B: 'You are a bad helper.'
	Sarcastic	Person A was building a very complicated structure out of Lego. Person B came over to help. Unfortunately, Person B unintentionally knocked some of it down. Person A said to Person B: 'You are a good helper.'

References. [1] Colston, 1997. *Discourse Processes.* [2] Dews & Winner, 1995. *Metaphor and Symbol.* [3] Pexman & Olineck, 2002. *Discourse Processes*. [4] Ivanko et al., 2004. *Journal of Language and Social Psychology*. [5] Happé, 1993. *Cognition*. [6] Blasko et al., 2021. *Canadian Journal of Experimental Psychology*. [7] Oprea & Magdy, 2020. *Proceedings of the ACM on Human-Computer Interaction*. [8] Rockwell & Theriot, 2001. *Communication Research Reports*. [9] Stone et al., 1998. *Journal of Cognitive Neuroscience.* [10] Davis, 1980. *JSAS Catalog of Selected Documents in Psychology.* [11] Barr et al., 2013. *Journal of Memory and Language.*

The role of context and working memory in the MIE — A window on metaphor processes

Shaokang Jin[1], Richard Breheny[1]

[1] University College London

The Metaphor Interference Effect (MIE) emerges when participants take more time to judge metaphors (e.g.(1)) as *literally false* than their scrambled counterparts (e.g.(2)).

1. **Some cats are princesses.**
2. **Some flutes are princesses**

[1,2] propose that the MIE is a kind of Stroop effect, wherein an automatically generated metaphoric interpretation conflicts with the task of finding and evaluating a literal interpretation. In previous work, replicated below, we place metaphors in a strongly constraining context and find that the MIE is eliminated. This outcome was contrary to expectations if the MIE was a stroop-like effect, since context should further promote the competing metaphoric meaning. We attribute previous MIE results to uncertainty surrounding de-contextualised metaphor items: language processes require background knowledge to derive figurative meanings and, without specific indications of relevance, an item like (1) can have many meanings (spoilt, bossy, lazy, haughty), depending on which implications are deemed relevant. We contend that a lack of discourse context keeps all such meanings 'live', draining resources and leading to longer latencies on the explicit task. In this new work, (i) we test our hypothesis about meaning uncertainty leading to longer latencies; (ii) we reconsider research on working memory and metaphor. Regarding (ii), [3] shows the MIE is lower for a High WM group than LWM and they attribute this to WM abilities overcoming Stroop interference. We contend instead that HWM individuals have more resources to deal with meaning uncertainty while completing the secondary task. In the current study we follow the individual difference analysis procedure of [3] but with our context/no-context design. Regarding (i), we ran a separate norming study on our metaphor sentences, eliciting participant interpretations and used an LSA-based analysis to measure similarity. Overall, our results replicate our previous effect of context (no MIE in context) and also the effect of WM in [3], in the no-context condition. The novel comparison supports our contention about the MIE and effect of WM. In addition, our LSA analysis reveals a correlation between perceived ambiguity of context-less metaphors and MIE.

Experiment 1. Participants (N=96 native English) completed two tasks in the following order: (a) Word span task (WSPAN) [4,5]; (b) Literal truth judgement task. In (b), participants were employed in a *2 (Within-group: Sentence form) * 2 (Between-group: Context)* design. Following [1], they made literal truth decisions to 24 metaphors (highly apt & novel) & 24 scrambled items, as well as 12 literally false & 60 literally true fillers, in either a no-context or a context condition (see Table 1). The context sentence was formulated so that target sentence was an elaboration and thus context strongly constrained figurative meaning. Literal fillers counterbalance response biases.

Results. Overall MIE Effect: We found a Sentence form * Context interaction (β=-35, se=6.06, p<.001): there was a large MIE in the no-context condition (p<.001), but no MIE in the context condition (p=.15) – see Fig. 1. Following [3], we analysed data for High (+1SD) and Low (-1SD) WSPAN participants and find a three-way interaction between WM, form & context (β=5.84, se=1.88, p=.002). With *no context*, we replicate the finding in [3] -- the MIE for High-WM (6ms, p=.59) < Low-WM (133ms, p=.001). With *context*, the MIE for High-WM was reduced to the negative value (-67ms, p=.53); the MIE for Low-WM was also eliminated (4ms, p=.36) – see Fig. 2.

Experiment 2 – Metaphor Interpretation Task. Participants (N=48 native English) were presented with the same list of metaphors (N=24) used in Experiment 1 and instructed to decide on the number of different interpretations that they can think of for each metaphor and

write down their interpretations.

Results. Figurative meaning uncertainty was measured by calculating semantic similarity between different interpretations of each metaphor using functions in the R package *LSAfun* [6]. A generalized linear mixed-effects model quantifies the semantic similarity of figurative meanings on *literally false* response of metaphors shows that the lower meaning similarity predicts the longer latency – with the meaning similarity decrease by 0.1 value leading to the latency increase by 71.2 msec (β=-712, se=47, p<.001).

Discussion. We attribute the negative MIE in context for HWM to the fact that automatic language processes attempt sense-making of even scrambled sentences. This suggests that a single constrained figurative meaning in context hardly interferes with the secondary task. Moreover, Exp. 2 confirms our hypothesis that delay on the literal truth judgement task results not from interference of a figurative meaning, but from figurative meaning uncertainty.

Table 1. Sample of critical items used in the literal truth-value judgement task

Conditions	CONTEXT	TARGET
NO-CONTEXT	/	Some friendships are wines. (metaphor)
	/	Some tickets are wines. (scrambled)
CONTEXT	Their friendship gets better with age.	Some friendships are wines. (metaphor)
	Their friendship gets better with age.	Some tickets are wines. (scrambled)

Note - the metaphors used in the study were highly apt and novel ones selected from a sample of 200 metaphors which were subjected to two previous pre-tests of familiarity and aptness norming

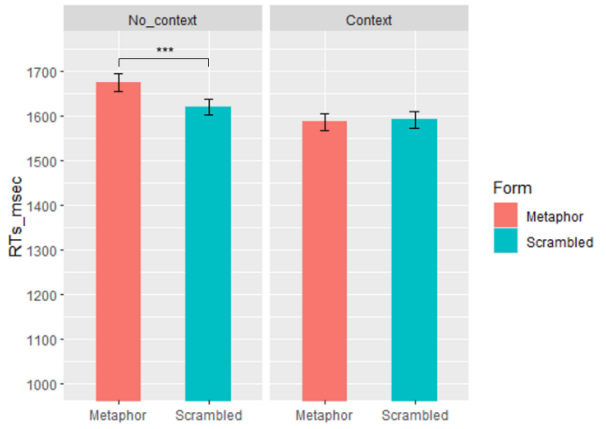


Figure 1. Overall Mean RT (and standard errors of the mean) of *literally false* responses to metaphors and scrambled sentences in two context conditions

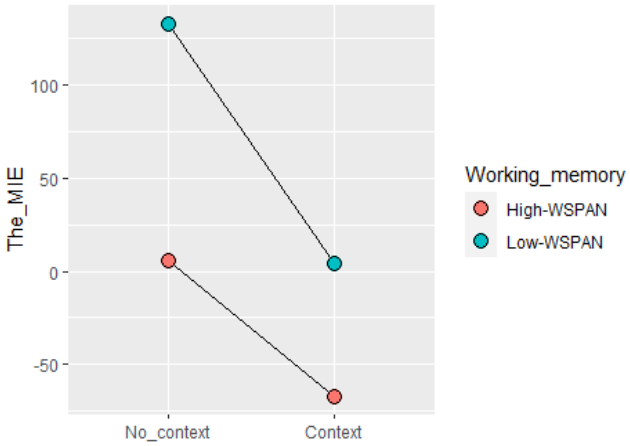


Figure 2. The MIE (metaphor RT – scrambled RT) for High-WSPAN participants and Low-WSPAN participants in two context conditions

References. [1] Glucksberg, Gildea & Bookin, 1982. *J. of Verbal Learning and Verbal Behavior* 21, 85-98. [2] Gildea & Glucksberg, 1983. *J. of Verbal Learning and Verbal Behavior* 22, 577-590. [3] Pierce, MacLaren & Chiappe, 2010. *Psychon Bull & Rev* 17, 400-404. [4] Engle *et al*., 1999. *JEP: General* 128, 309-331. [5] La Pointe & Engle, 1990. *JEP: Learning, Memory, and Cognition* 16, 1118–1133. [6] Günther, Dudschig & Kaup, 2015. *Behavior Research Methods* 47, 930–944.

Title: Pragmatic and knowledge lenience towards foreigners

Anna Lorenzoni[1], Elena Pagliarini[2], Francesco Vespignani[1] & Eduardo Navarrete[1]

[1]Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università degli studi di Padova, Italy

[2] Dipartimento di Studi Linguistici e Letterari, Università degli studi di Padova, Italy

Background: The identity of the interlocutor is an essential cue for successful communication. For example, a sentence like 'I have a large tattoo on my back' could be considered a credible statement if made by an adult, but an ironic one if made by a child. Here, we focused on the linguistic identity of the interlocutor. Recently, some researchers have highlighted the idea that the evaluation of an utterance is affected by accented speech. In a paper by Lev-Ari & Keysar (2010), speakers uttered typically unknown world-knowledge facts statements (e.g., 'Ants don't sleep'), with either a native or a foreign accent. Participants judged foreign-accented trials to be less true than native-accented statements. The authors interpreted their findings according to a 'fluency-intelligibility' account, where foreign-accented speech leads to a decrease in fluency and ease of understanding. Critically foreign-accented speech may not only affect message intelligibility but may also lead to an implicit categorization of the speaker as an outgroup individual (foreign) in terms of cultural and social heritage. Our main aim here was to explore whether the identification of an individual as a native or foreign speaker has an impact per se on **unknown** statement judgments. Critically, to avoid any influence of the auditory signal, we used a written modality presentation of the statements.

In a recent study, Fairchild and Papafragou (2018) also used written materials to isolate the influence of speaker identity on the acceptability of the scalar implicature. In their study, participants tended to accept more a series of under-informative written sentences ('Some dogs are mammals') when attributed to a foreign speaker compared to native speakers. In the two studies we present here, we first aim to replicate the Fairchild and Papafragou study on **scalar implicature** (Study 1); then we used the similar procedure to test **unknown** statements (Study 2). Two different experiments were conducted within each study. In experiments 1a and 2a we used the same methodology developed by Fairchild and Papafragou (2018). In experiments 1b and 2b, the same procedure was used with the difference that we added face photographs to each of the two speakers to increase the association between speaker and sentence.
Study 1: 244 native Italian speakers participated in the study (99 and 145 for experiment 1a and 1b, respectively). The experimental set was composed of 20 under-informative sentences with the quantifier 'some'. Furthermore, three filler conditions (20 sentences each) were added: true filler sentences containing 'some' ('Some hair is brown'); true filler sentences containing 'all' ('All snow is cold'); and false filler sentences containing 'all' ('All women are doctors').
Following Fairchild and Papafragou (2018), four bio-descriptions were created. Each short-bio either gave a description of a native Italian speaker with a strong Roman accent (Native speaker condition) or a native speaker of Moldovan with a strong Moldovan accent (Foreign speaker condition). In addition, for experiment 1b, two colour photographs of real women's faces were selected. The experiment consisted of two blocks: a native and a foreign language block (counterbalanced between participants). The sentences within each block were evenly distributed among the four types of sentences (10 of each),and presented in a random order. At the start of each block, one of the four speaker bio-descriptions was presented, and participants were instructed to read it carefully. The participants were then instructed that they would be reading 40 sentences that were originally uttered by the speaker they had just read. The sentences were presented in a random order. For each trial, a sentence appeared in the centre of the screen together with the ratings scale below. The speaker bio-description was presented at the top of the screen. Participants had to rate how each sentence made sense on a five-point scale (1-

"Completely no sense" and 5-"Completely sensible"). For experiment 1b the same procedure was used with the following differences: the two bio-descriptions were presented at the beginning of the experimental session together with one face image; sentences were presented together with the face at the top of the screen instead of the bio-description; the 80 sentences were presented in a random order with a short break after 40 sentences.

Descriptive statistics are reported in Table 1. Analyses were performed on the rating responses of the critical sentence condition. Ordinal logistic regression was used in the form of a mixed cumulative link model (*clmm* in R). In the mixed models, the factor Speaker (Native vs. Foreigner) and Experiment (1a vs. 1b) was introduced as fixed effect. The participant and item were included in the model as random factors. Two models were constructed, with and without interaction of the two fixed effects. The fits of the two models were compared using Akaike's information criterion (*AIC*). The model with the lowest AIC would have the best fit. The comparison between the two models revealed that the best model was the one without interaction. The model shows a main effect of the Speaker (SE=0.06, z= -2.01, p=.04) due to the fact that ratings for Under-Informative sentences were higher in the Foreign speaker condition (M=2.55, SD=1.48) than in the Native speaker condition (M=2.49, SD=1.47). The main effect of Experiment was not significant (p=.14).

Study 2: 239 native Italian speakers participated in Study 2 (114 for experiment 2a and 125 for experiment 2b). The experimental set was composed of 20 unknown sentences ('The capital of Botswana is Gaborone'). Furthermore, two filler conditions, 20 sentences each, were added: true filler sentences ('To play tennis, you need to have a racket') and false filler sentences ('Arachnophobia is the fear of having fun'). The same task, presentation modality, and analyses as for Study 1 were used. The comparison between the two models revealed that the best model was the one without interaction. Results from *clmm* also revealed a main effect of the Speaker (SE=0.06, z= -2.13, p=.03), with ratings for Unknown sentences were higher in the Foreign Speaker (M=2.99, SD=0.86) condition than in the Native Speaker (M=2.95, SD=0.88) condition. The main effect of Experiment was not significant (p=.25). See Table1.

Speaker	Study 1 (Under-informative)		Study 2 (Unknown)	
	Experiment 1a	Experiment 1b	Experiment 2a	Experiment 2b
Native	2.34 (1.40)	2.58 (1.52)	2.98 (0.87)	2.92 (0.89)
Foreign	2.45 (1.41)	2.62 (1.52)	3.02 (0.83)	2.96 (0.88)

Table 1. Average of the ratings in Study 1 and Study 2 divided by manipulation and type of experiment. Standard deviations are reported in parentheses.

Discussion: Our results showed that the categorization of speakers as foreign or native speakers per se modulates the acceptability of statements independently from differences of processing linked to fluency. In Study 1, we replicated in Italian previous findings reported in English. We interpret 'pragmatic lenience' toward foreign speakers on the basis of beliefs of comprehenders about the lower linguistic competence of foreign speakers. In Study 2, our results were in the opposite direction with respect to the findings of Lev-Ari and Keysar (2010). A possible explanation for the advantage for foreigners may rely on the different attribution of general knowledge to foreign and native speakers when an unknown sentence is presented. Something we will call 'knowledge lenience' toward foreign speakers. Together, our results suggest that native speakers do not only tend to forgive lack of linguistic competence of foreign speakers, by accepting as more sensible under-informative statements, but they also tend to trust more foreign speakers in situations of lack of knowledge.

References:

- Fairchild, S., & Papafragou, A. (2018). Sins of omission are more likely to be forgiven in non-native speakers. Cognition, 181(December 2016), 80–92. https://doi.org/10.1016/j.cognition.2018.08.010

- Lev-Ari, S. & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. Journal of Experimental Social Psychology, 46(6), 1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025

Can slurs be used without being mentioned? Evidence from an inference judgement task

Maria Esipova (University of Oslo)

Background Different types of content behave differently under ellipsis (see, e.g., Esipova 2019 for a brief overview and references therein). Thus, not-at-issue inferences that are an inextricable part of an item's lexical meaning can't be ignored in elliptical environments that require some form of identity with an antecedent instance of that item. This is true, e.g., for presuppositions of verbs encoding a stage of an event (*start*, *stop*, *continue*...), factive predicates (*know*, *regret*...), etc.:

(1) Pam stopped smoking, {but Kim didn't / and Kim did, too / and so did Kim}.
(i) $\rightarrow$ Pam used to smoke. (ii) $\rightarrow$ Kim used to smoke.

On the other end of the spectrum are pure expressives, which are always ignored under ellipsis:

(2) A: Did you bring a fucking gun to my house?
B: No, I didn't. / Yes, I did. / Yes, I did so. / Yes, I brought one.
(i) $\rightarrow$ A is experiencing strong emotions. (ii) $\not\rightarrow$ B is experiencing strong emotions.

Now, there are a few potentially relevant differences between these two cases: (i) the target inference in (1) is part of the lexical meaning of the head of the recovered VP, but the target inference in (2) is contributed by an adjunct inside an NP that is in turn inside the recovered VP or (contestably) targeted by *one*; (ii) the presupposition of *stop* is a precondition for the antecedent in (1) to make sense, but that's not the case for the contribution of *fucking* in (2); and perhaps most importantly, (iii) acts of producing expressives like *fucking* are purely performative, i.e., the speaker achieves their goal (here, expressing their emotions) by virtue of producing a certain form (use via mention), and there is no way to achieve this goal without performing this act (no use without mention).

Question Slurs, however, are a more complex case: (i) the "prejudice inference" is part of the lexical meaning of a slur, which can be the head of the antecedent constituent targeted by different types of ellipsis (like *stop*, unlike *fucking*); (ii) despite that, this inference is not crucial for the at-issue content of the sentence containing the slur to make sense (unlike *stop*, like *fucking*); (iii) slurs can be used performatively (use via mention) and can even have a performative effect of offense by virtue of being uttered in the absence of such intent on the speaker's part (mention without use), but it is unclear if the prejudice inference can be preserved if a slur is recovered but not uttered (use without mention). In this paper, I look at paradigms like (3) to assess the effect of different factors on the presence/strength of the prejudice inference and thus shed further light on the nature of this inference (the exchanges are set in a fictional universe where humans co-exist with centaurs, dwarves, elves, orcs, etc. and happen in the context of a criminal investigation):

(3) a. *Context: 'Tusky' is a slur for orcs.*
Detective: Did you see a tusky?
Witness: Yes. ('Bare') / Yes, I did. ('VPE') / Yes, I saw one. ('One') / Yes, I saw a tusky. ('Slur') / Yes, I saw an orc. ('Nonslur')
b. *Context: 'Tusky' is a slur for orcs. This slur can also be used as a verb meaning 'to crawl' (for any race), because orcs are stereotyped as living in caves and, thus, having to crawl through narrow spaces all the time. The detective is asking a question about a human.*
Detective: What happened next? Did he tusky under the table?
Witness: Yes. ('Bare') / Yes, he did. ('VPE') / Yes, he did so. ('So') / Yes, he tuskied under the table. ('Slur') / Yes, he crawled under the table. ('Nonslur')
Question: How likely do you think that this witness is prejudiced against orcs?

Hypothesis I hypothesized that the "prejudice likelihood" inferred from responses like those in (3) is gradient and is affected by several syntactic, semantic, and pragmatic factors: 1. Maintaining that slurs do have performative effects, I expected the likelihood to be highest when the witness utters

the slur themselves ('Slur'), ostensibly both using and mentioning it. 2. I also expected the likelihood to be lowest when the witness tacitly corrects the detective by using the neutral term instead ('Nonslur'), thus, neither mentioning nor using the slur and, furthermore, indirectly challenging the detective on their use of the slur in an attempt to minimize complicity (see, e.g., Cepollaro 2020 and references therein on unchallenged slurs). 'Nonslur' responses are, thus, expected to have lower prejudice likelihood ratings than any of the elliptical responses. 3. Finally, I hypothesized that the prejudice component of slurs is not exclusively performative, i.e., it does allow for use without mention. Thus, when the slur is obligatorily recovered, e.g., when it is the head of the constituent targeted by a proform requiring lexical identity with said head (more obviously in 'One' for nouns; less obviously in 'VPE' and 'So' for verbs), the prejudice likelihood is expected to be higher than in elliptical responses where the slur is not necessarily recovered ('Bare' and 'VPE' for nouns; 'Bare' for verbs). So, if all parts of the hypothesis are correct, we expect the following picture:

(4) Predicted prejudice likelihood ratings (from lowest to highest)
   a. Nouns: 'Nonslur' < 'Bare'/'VPE' < 'One' < 'Slur'
   b. Verbs: 'Nonslur' < 'Bare' < 'VPE'/'So' < 'Slur'

Methods The experiment involved 10 conditions (2 parts of speech, with 5 response types for each). Each participant saw 2 trials per condition and 2 attention checks (22 trials total); the trials looked similarly to (3). Participants assessed the prejudice likelihood by dragging a slider on a pseudo-continuous scale (mapped to 0–100) from 'Not at all likely' to 'Very likely'. Participants were recruited on Prolific (final N = 128) and paid £1.25 for completing the task.

Results The results are visualized in Fig. 1. The statistically significant contrasts fully matched the prediction in (4a) for nouns, but only partially matched the prediction in (4b) for verbs:

(5) Statistically significant contrasts in prejudice likelihood ratings (from lowest to highest)
   a. Nouns: 'Nonslur' < 'Bare'/'VPE' < 'One' < 'Slur'
   b. Verbs: 'Nonslur' < 'Bare'/'VPE'/'So' < 'Slur'

Discussion The results for noun slurs corroborate all parts of the original hypothesis, suggesting that slurs do make performative contributions (which is why actually saying a slur gives rise to a stronger effect than using it without mentioning), but are not exclusively performative (which is why the prejudice inference still persists to some extent if the slur is recovered, but not uttered). This calls for a hybrid analysis for slurs that doesn't reduce their prejudice component to just a presupposition (as in Schlenker 2007) or just a performative effect on the context (as in Potts 2007). The results for verb slurs corroborated parts 1 and 2, but not 3 of the hypothesis, possibly because: (i) in the absence of perfect English counterparts, verb slurs were harder to intuit about, so the data were more noisy, (ii) due to (i) and the less direct link between the meaning of a verb slur and the targeted group, the contrasts were overall less pronounced, and (iii) the identity requirements for verbs in VPE and *do so*-replacement are less clear than for nouns in *one*-replacement.

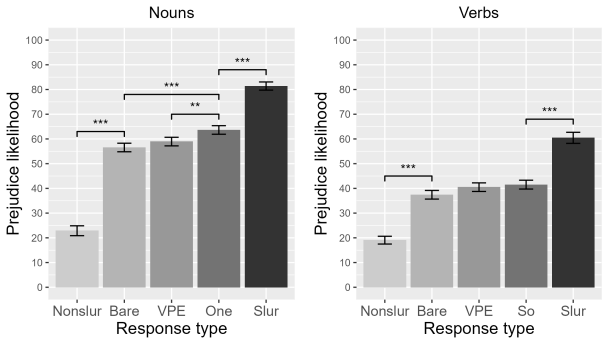


Fig. 1: Bar charts showing mean prejudice likelihood ratings of different types of responses to antecedent utterances with noun and verb slurs, with SE and key significant contrasts indicated.

References Cepollaro. 2020. isbn/9781793610522 Esipova. 2019. lingbuzz/004676. Potts. 2007. doi:10.1515/TL.2007.011 Schlenker. 2007. doi:10.1515/TL.2007.017

Irony Regulates Negative Emotion – in Speakers and Listeners

Valeria A. Pfeifer & Vicky Tzuyin Lai
Department of Psychology & Cognitive Science Program, University of Arizona

Verbal irony is when a speaker uses words whose literal meaning is the opposite of the speakers intended meaning. For example, when someone looks at the giant buffet at a potluck and exclaims: “That’s hardly enough food!”. Verbal irony is commonly used to express negative emotions (Roberts & Kreuz, 1994), yet it is unclear what irony does to negativity and why irony is useful for expressing negative emotions. Some argue that irony can be used to milden negativity, known as the *tinge hypothesis* (Dews & Winner, 1995). This is supported by empirical evidence from ratings, eyetracking, and Event-related potentials (ERPs) (e.g. Filik et al. 2017, Pfeifer & Lai, 2021). However, past studies mainly considered the speaker, or the statement itself. Here, we propose that irony can effectively reduce negative emotion not just in speakers, but also in listeners, making irony a vital communicative tool to regulate negative emotions in social situations, for example conversations.

Our hypothesis was that irony would reduce negative feelings when compared to literal language. We used a block-design where participants (N = 54) saw images of negative events (N = 128, mean negativity = 3.01 on a 4-point scale (1 = weak, 4 = strong negativity), e.g. flies on a pie, flat tire) and were instructed to imagine the situation was happening to them. In the verbal block, they then read either literal (N = 32) or ironic (N = 32) statements about the situation, presented word-by-word, before viewing the same picture for a second time. In the non-verbal block, they either saw “attend” (N = 32) or “reinterpret” (N = 32) to indicate if they should regulate their emotions or attend to them, before viewing the same picture for a second time. In both cases, participants rated how negative they felt (1 = weak, 4 = strong) after the second image presentation was completed. Electroencephalography (EEG) was recorded throughout the experiment. Participants also reported their language background, success of the *reinterpret* strategy and frequency of irony use. All statements were normed for ironicity.

Averages of behavioral responses are displayed in Figure 1. Paired-t-tests showed that ironic statements led to lower ratings of negativity compared to literal statements (p = .025), and that *reinterpret* led to lower ratings of negativity than *attend* (p < .001). There also was a positive relationship between how frequently participants used irony in daily conversations and how negative a literal statement made them feel (p = .049, r = .26), such that using more irony in daily life led to feeling more negative after reading literal statements during the experiment, but no such (reverse) relationship was present for ironic statements.

ERPs (N = 43, 11 excluded due to excessive noise) are displayed in Figure 1. ERPs were time-locked to the onset of the literal/ironic word in the verbal block, and to the onset of the attend/reinterpret instructions in the non-verbal block, respectively. Irony elicited a larger prolonged negativity compared to literal statements from 300-900ms, visible on the whole scalp. *Reinterpret* elicited a larger positivity compared to *attend* in 300-500ms, and in frontal channels from 800-1600ms.

We interpret the findings as follows. Behaviorally, irony significantly lowers negative emotion elicited by a negative image compared to literal. While irony is more effective than literal language, it is not as effective as actively regulating one’s emotion via cognitive reappraisal. Neurally, similar evidence is found. Irony creates a contrast between the image and the statement, as evident by the enhanced negativity in the traditional N400 timewindow (300-500 ms), and such contrast lingered and continued to be processed (500-900 ms). Cognitive Reappraisal, however, elicits a larger positivity compared to attending to emotions, likely indexing the cognitive effort used in actively regulating emotion. Together with behavioral results, this suggests that irony is successful in decreasing negative emotion, but it accomplishes this in different ways from cognitive reappraisal: rather than actively focusing on

regulating one's response, readers of ironic statements experience a contrast to the situation, which results in less negative emotion, possibly by creating distance, or via Theory of Mind involvement. In other words, irony can be a successful tool that regulates negative emotion, without requiring active participation from the listener. This is important, as it suggests that pragmatically, irony not only mildens negativity in speakers (Pfeifer & Lai, 2021, Filik et al. 2017) but also in recipients, thus, demonstrating that pragmatic functions of irony can be both self- and other serving. Based on the current and previous data, we propose a model of the pragmatic functions of irony (Figure 2) that uses self- and other-serving functions to explain how irony can be simultaneously more hurtful and more amusing (Boylan & Katz, 2013).

Figure 1: *Left*: Average ERP waveforms for non-verbal and verbal blocks, timelocked to the critical word (verbal) or the onset of the instructions (non-verbal). Non-verbal block shows frontal and parietal channels, verbal block shows central channels. *Right*: Average ratings of negativity on a 1-4 scale (1 = weak to 4 = strong).

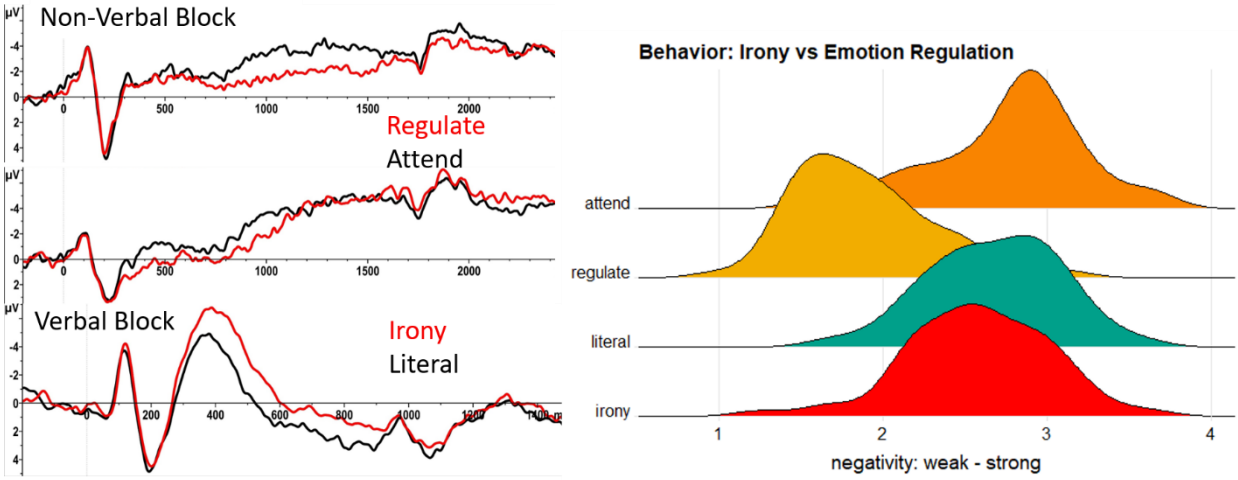
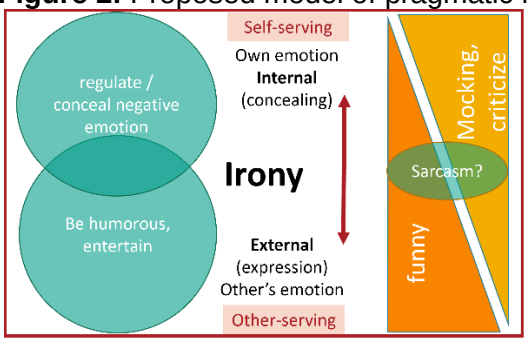


Figure 2: Proposed model of pragmatic functions of irony



References: Roberts, R. M., & Kreuz, R. J. (1994). Why do people use figurative language?. *Psychological science*, *5*(3), 159-163.; Dews, S., & Winner, E. (1995). Muting the meaning a social function of irony. *Metaphor and Symbol*, *10*(1), 3-19.; Pfeifer, V. A., & Lai, V. T. (2021). The comprehension of irony in high and low emotional contexts. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*.; Filik, R., Brightman, E., Gathercole, C., & Leuthold, H. (2017). The emotional impact of verbal irony: Eye-tracking evidence for a two-stage process. *Journal of Memory and Language*, *93*, 193-202.; Boylan, J., & Katz, A. N. (2013). Ironic expression can simultaneously enhance and dilute perception of criticism. *Discourse Processes*, *50*(3), 187-209.

Five degrees of (non)sense: Investigating the connection between bullshit receptivity and susceptibility to semantic illusions

Dario Paape (University of Potsdam)

The sentence *The invisible is beyond new timelessness* is bullshit [1]. Bullshit is characterized by *unclarifiable unclarity*: It has no clear meaning that could be explained without significant deviation from the original form of the statement [2]. Yet, bullshit statements are often judged to be true or even profound [1,3]. Individual differences in bullshit receptivity have been partly attributed to differences in acquiescence bias and/or interpretive charity [1,3]. Participants who see patterns in visual noise also tend to endorse bullshit sentences, suggesting that they tend to "creat[e] meaning where no meaning exists" [4]. Such a tendency may extend to other forms of nonsense. Semantic illusions arise in sentences such as *No head injury is too trivial to be ignored* ("depth charge" illusion) [5] or *More people have been to Russia than I have* (comparative illusion) [6], which are often perceived as sensible despite being compositionally incongruous. We have two aims: To investigate the possible correlation between individual bullshit receptivity and susceptibility to semantic illusions, and to investigate the possible shared role of interpretive charity and illusory pattern perception.

As a cover story for our experiment, we told participants that we had created an artificial intelligence (AI) system that can create natural-sounding utterances but also occasionally produces nonsense. 100 participants were asked to rate the stimulus sentences' meaningfulness and naturalness on a 7-point Likert scale. In addition to bullshit statements and semantic illusion sentences, we included sensible sentences and transparently nonsensical sentences as controls, as shown in **Table 1**. As a measure of illusory pattern perception, participants were shown randomly generated two-dimensional dot patterns (see **Figure 1**) and told that these represented the AI's neuronal activations. They were told to indicate for each pattern whether they saw any meaningful structure in the activations or not. Reaction times were collected for both tasks.

Ratings were analyzed with a hierarchical cumulative logit model. To control for differences in the use of the Likert scale, the model contained subject-specific adjustments (random effects) to the sizes of the rating "bins". Crucially, we estimated the correlations between subject-specific adjustments to the mean ratings — relative to the average — across sentence types, as well as between ratings and the pattern perception measure. For example, if perceived meaningfulness is due to interpretive charity, the subject-wise adjustments across sentence types should be positively correlated, as charitable subjects should give higher ratings across the board.

Our results indicate that interpretive charity plays a role in bullshit receptivity: Participants who gave higher ratings to nonsense and sensible sentences also gave higher ratings to bullshit sentences (**Figure 2**). Furthermore, there is a negative correlation between ratings for nonsense and sensible sentences, due to subjects being more or less extreme in their negative perception of nonsense and their positive perception of sensible sentences. Participants who made stronger distinctions showed a stronger effect of sentence length on reading time, suggesting that they read more attentively [7]. There is no indication of a correlation between bullshit receptivity and susceptibility to semantic illusions, nor of a general correlation with illusory pattern perception. Only the comparative illusion shows some indication of a positive correlation with pattern perception, as well as of a negative correlation with attentive reading. By contrast, the depth charge illusion shows some indication of a positive correlation with attentive reading, and of a negative correlation with nonsense acceptability. Overall, our results suggest that there may be no general individual trait that explains bullshit receptivity and susceptibility to semantic illusions. However, the results raise interesting possibilities for future research: The depth charge and comparative illusions may involve different cognitive mechanisms, and may be differentially related to attention and depth of processing.

(1) (a) **Sensible**
Your teacher can open the door, but you must enter by yourself.

(b) **Nonsense**
One can say that flowers with a lot of old nettles do not limp without great experience value.

(c) **Bullshit**
The invisible is beyond new timelessness.

(d) **Dept charge illusion**
No head injury is too trivial to be ignored. (correct: . . . to be treated)

(e) **Comparative illusion**
More people have been to Russia than I have.

Table 1. Example sentences used in the experiment. Half of the sensible sentences were “profound” like the example, while the other half were more mundane (*Newborns need constant attention*). The bullshit condition contained an equal number of “pseudo-profound” bullshit [1], scientific bullshit [3], and “International Art English” [8] sentences.

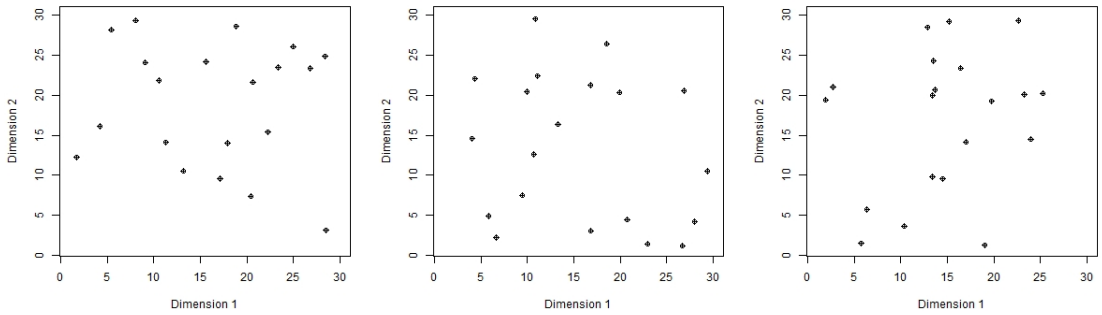


Figure 1. Example dot patterns (“neuronal activations”) used in the pattern recognition task.

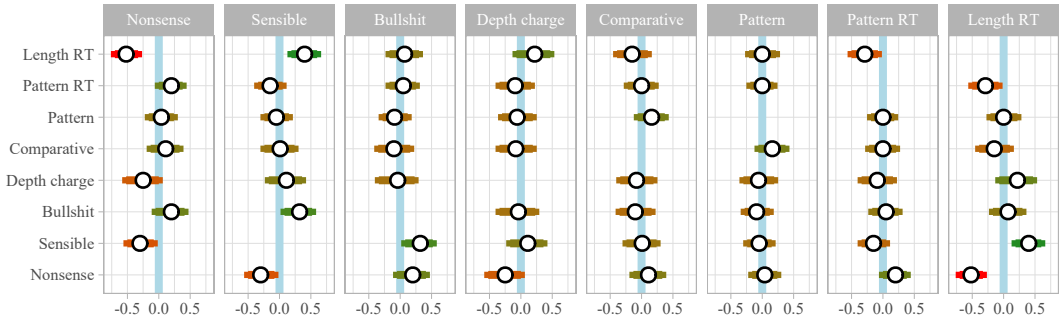


Figure 2. Estimates and 95% credible intervals of subject-level random-effects correlations.

References. [1] Pennycook et al. (2015, Judg Dec Mak). [2] Cohen (2002, *Deeper into bullshit*). [3] Evans et al. (2020, Judg Dec Mak). [4] Walker et al. (2019, Judg Dec Mak). [5] Wason & Reich (1979, Q J Exp Psychol). [6] Wellwood et al. (2018, J Semant). [7] Schad et al. (2012, Cognition). [8] Turpin et al. (2019, Judg Dec Mak).

Context matters: Changes in the affective representation of a word in younger and older adults

Li-Chuan Ku[1, 2] & Vicky Tzuyin Lai[1, 2]
[1] Department of Psychology & [2] Cognitive Science program, University of Arizona

Do younger and older adults differ in their processing of positive or negative meanings in language [1]? Based on the automatic vigilance hypothesis (AVH) [2], humans of all ages attend to negative information, a.k.a. 'negativity bias', as it threatens perceivers' well-being. However, based on the socioemotional selectivity theory (SST) [3], such preference changes into 'positivity bias', as people age and re-prioritize positive information for their emotional well-being. Gaps in knowledge are that most studies focused on single words [4], their absolute valence values, and younger adults [5, 6]. Here we investigated whether younger (YAs) and older adults (OAs) update the affective representations of a (same) word in negatively and positively valenced context fluently, using EEG. We hypothesized that if the AVH holds, negative contexts should lead to more negative evaluations of all target words. In contrast, if the SST holds, positive contexts should lead to more positive evaluations of all target words. If neither holds, the very same word before and after the emotional contexts should show the same neural representations.

We conducted an online (Exp 1: N_{YA}=60, N_{OA}=43) and an ERP study (Exp 2: N_{YA}=41, N_{OA}=23 and ongoing). Stimuli consisted of 320 three-sentence vignettes with positive/negative target words and positive/negative contexts (=adjectives in 2nd sentence; Table 1). Target words were all low-arousing, as our prior data on single words indicated that positivity bias in OAs was revealed in low-arousing words. Word valence ratings were obtained from both YAs and OAs based on affective norms. Word properties (length, frequency, concreteness) were matched between conditions for target words and for contexts/adjectives. Exp 1 participants read the first two sentences in each vignette and rated the valence of the target word from 1 (very negative) to 9 (very positive). Exp 2 participants read each vignette word-by-word and did a valence judgment task (Figure 1).

Participants with high depression scores, cognitive impairment, program error, or excessive alpha were excluded. For Exp 1 (N_{YA}=36, M_{age}=19.7; N_{OA}=36, M_{age}=65.4), participants' age, cognitive ability (Wisconsin Card Sorting Task; Digit-Symbol Substitution Task), and affect scores (PANAS-trait) were entered in a regression model as predictors of the valence ratings. For Exp 2 (N_{YA}=24, M_{age}=18.9; N_{OA}=14, M_{age}=68.9), changes in the affective representations are reflected by the ERP differences between the 1st and 2nd occurrences of the target words (ERP effects hereafter).

In Exp 1, increasing age (β_{Age}=.342, p=.026) and positive affect (β_{PA}=.314, p=.014) separately predicted more positive evaluation for positive target words in positive contexts (Figure 2), consistent with 'positivity bias'. In Exp 2, in YAs, P2 effects (180-300 ms) were reduced for positive targets (p=.048), suggesting automatic attention to negative targets (Figure 3). Also in YAs, LPP effects (600-800 ms) were enhanced for target words in negative contexts (p=.008), suggesting sustained attention to negative contexts. In OAs, there was no interaction, but simple comparison supported an enhanced P2/LPP effect for positive words in positive contexts.

Altogether, YAs support the AVH, as first reflected by a reduced P2 effect to positive targets, and then an enhanced LPP effect to negative contexts. While there is no robust support of the SST, OAs show steady reactions to positive words in positive contexts, in P2/LPP effects first, and then Exp 1 valence ratings.

Table 1. Stimulus examples

	Positive target (in green)	Negative target (in red)
Positive context (in bold)	The pianist had a new performance. Her skills were **remarkable**. The pianist practiced every day.	The dentist often worked with children. They found him **trustworthy**. The dentist cared about them.
Negative context (in bold)	The pianist had a new performance. Her skills were **rusty**. The pianist practiced every day.	The dentist often worked with children. They found him **formidable**. The dentist cared about them.

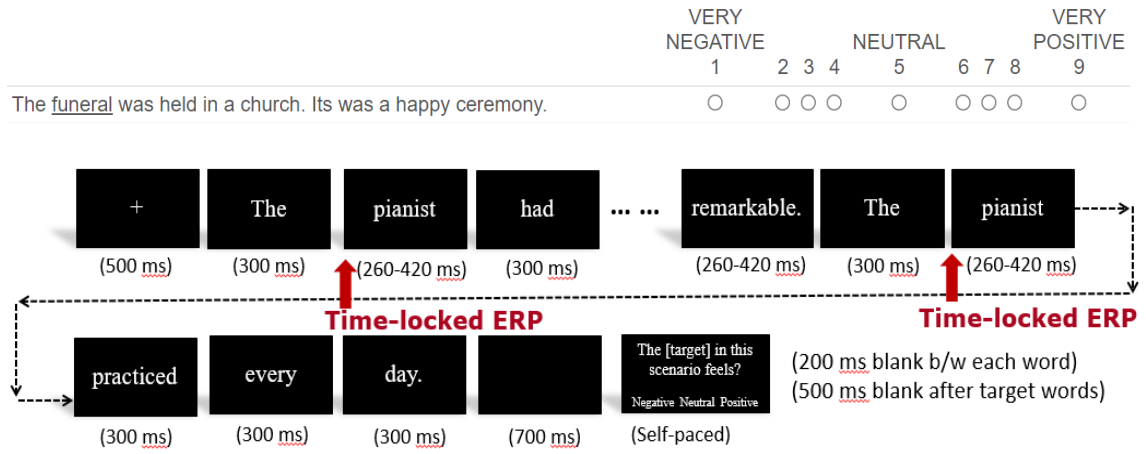


Figure 1. An example of a trial in Experiment 1 (top) and Experiment 2 (bottom)

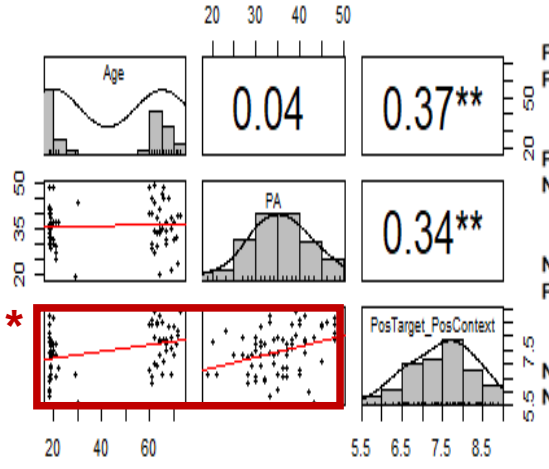


Figure 2. Exp 1: Correlation plots between age, positive affect (PA), and the valence ratings on positive targets in positive contexts (PosTarget_PosContext)

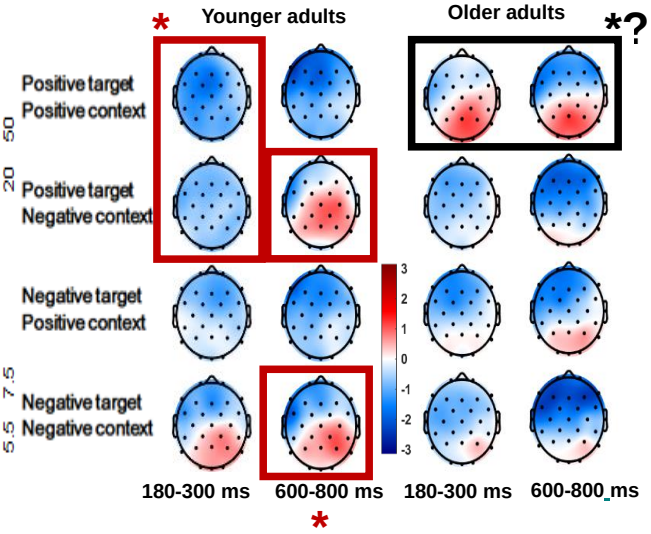


Figure 3. Exp 2: Scalp topography of difference amplitudes between the 2nd and 1st occurrences of the target words[1]

References. [1] Kauschke et al., 2019. *Frontiers in Psychology*. [2] Estes & Adelman, 2006. *Emotion*. [3] Carstensen, 2006. *Science*. [4] Ku et al., 2020. *Cognitive, Affective, & Behavioral Neuroscience*. [5] Delaney-Busch & Kuperberg, 2013. *Cognitive, Affective, & Behavioral Neuroscience*. [6] Lüdtke & Jacobs, 2015. *Frontiers in Psychology*.

[1] Data collection for older adults is still ongoing due to delay caused by Covid-19 pandemic.

Accessing children's pragmatic competence through intonational production

Line Sjøtun Helganger[1] & Ingrid Lossius Falkum[2]
[1]University of South-Eastern Norway; [2]University of Oslo

Introduction: In this study, we investigate the question of what type of pragmatic competence children have, and how early it arises in development. We use Norwegian children's intonational productions as a way of accessing their pragmatic competence. Although not yet adult-like, it is assumed that the ability to use intonation functionally (i.e., to signal an utterance's information structure) is largely established by the age of five. However, children's acquisition of intonation in the period prior to five years of age is still a quite unexplored field of research. Furthermore, there have been few attempts to combine suprasegmental phonology with cognitive pragmatic theory in the study of language acquisition (Wharton, 2020). Thus, the question of how children's ability to master intonation as a communicative device develops, is a largely unresolved question.

We investigate the pragmatic function of intonation by focusing on utterances realized with so-called 'polarity focus' (PF) in Norwegian, where the polarity of a proposition is highlighted through intonational means: By accentuation of a 'polarity carrier' (most commonly the finite verb) followed by an additional accentuation later in the utterance, the speaker signals whether she believes that a metarepresented proposition is a true or false description of some state of affairs (Fretheim, 2002). Consider the conversation in (1)[1]:

(1) A: *Jeg kommer meg ikke til butikken!*
I come me not to grocery store-DEF ('I cannot get to the grocery store!')

B (who knows A has an electric car): *Bilen ER LADET.*
car-DEF IS CHARGED ('The car is charged (despite what you seem to think)')

The proposition expressed by A in (1) is that A cannot get to the grocery store. By responding with a PF utterance, B communicates, by prosodically highlighting the finite verb *er* ('is'), that B dissociates herself from a (false) belief that she attributes to A, that he cannot use his (electric) car to drive to the grocery store because it is discharged. B's use of PF allows her to communicate that there is an opposition between what she thinks A thinks, and her own belief.

We hypothesize that the ability to produce PF utterances in contexts where the speaker dissociates from an inferred (false) belief (e.g., (1) above), is acquired around four years of age, together with the emergence of explicit theory of mind abilities (Wellman et al., 2001). However, the minimal requirement for PF utterances is the ability to produce multiword utterances realized with two accentuations (Fretheim, 2002). We therefore expect that the earliest productions of PF utterances occur after two years of age in less complex contexts. To test this, we designed an experiment to elicit PF utterances in increasingly more complex contexts, based on the assumption that negation increases utterance complexity (Just & Carpenter, 1971), and with dissociation from an inferred belief as the most complex condition.

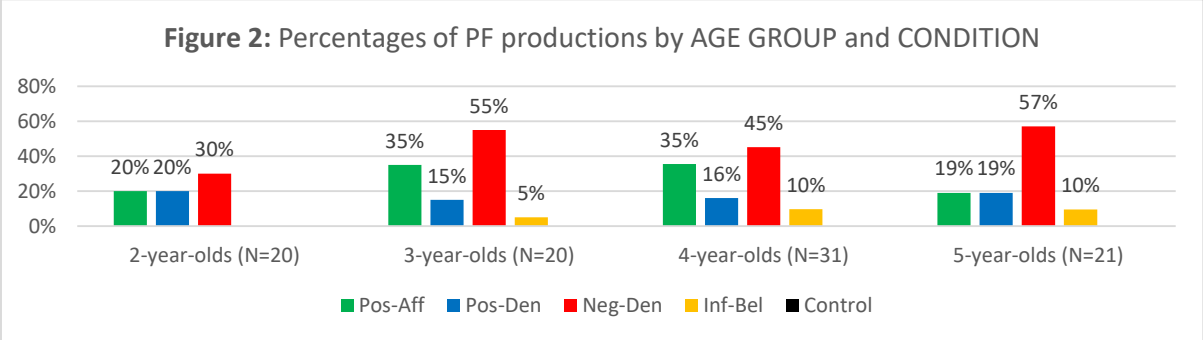
Method: Participants include 92 Norwegian-speaking children aged 2;2-5;9 who take part in a semi-structured elicitation task. An experimenter and a handpuppet initially show the participant some toys (e.g., rubber ducks) in an unstructured conversation, in which the handpuppet demonstrates that he is a bit forgetful. The structured elicitation task has four PF conditions with increasing complexity and one control. Notice that the participants are not given any instructions for how to respond in this task. In the PF conditions, the puppet initially states his (positive or negative) prior belief about what he thinks is depicted in a set of upcoming still life pictures (see Fig. 1 for an example item). Depending on the condition, the prior belief is either a match or a mismatch as a description of the picture's motive. The crucial task for the participant is to produce a target utterance in response to the puppet's declared belief (e.g., *gutten LESER bok*, i.e., 'the boy DOES read a book'). In the fourth condition, the use of PF is relevant as a response only if the participant has inferred a (false) belief of the handpuppet (e.g., the PF utterance *du HAR BADEENDENE dine*, i.e., 'you DO have your rubber ducks' in response to the handpuppet's utterance *I wish I had something to play with while taking a bath*,

[1] Upper case letters indicate a focal accentuation (i.e., a tonal rise to an extra high tone).

suggesting that he has forgotten about the rubber ducks that they played with in the initial unstructured section). In the control condition, the handpuppet presents a neutral belief about what is depicted (e.g., *I don't know what the girl does*), where use of PF is not relevant due to there being no proposition to attribute or to highlight the polarity of.

Figure 1 Example item: Negative-denial condition (Neg-Den)
Experimenter: *Here is a picture of a boy*.
Prior (negative) belief: *I believe that the boy is not reading a book*
Visual stimuli: A boy reading a book
Elicitation question (if needed): *Does not the boy read a book?*
Possible PF response: *gutten LESER bok*
boy-DEF READS book ('The boy DOES read (a) book')

Preliminary results: PF is produced in all four PF conditions and there are no PF productions in the control condition (see Fig. 2). Furthermore, there are PF productions in all age groups, and even participants from the youngest age group produced PF in as much as three out of four PF conditions, leaving only the Inferred (false) belief condition without PF productions in this age group.



We fitted a Generalized Linear Mixed Model of the PF productions as a count response with an upper bound and age as a covariate using a binomial error distribution and the glmer function of the lme4 package in R (version 4.1.2) to investigate the development of production of PF with age. The results show no effect of age (p = 0.191).

Discussion: As expected, we find the ability to use PF to express the dissociation from an inferred false belief to arise around four years of age. Already from two years of age, children have the ability to use PF in contexts with increasing complexity. Strikingly, the Negative-Denial (Neg-Den) condition has by far the highest percentage of PF production across the age groups, which suggests that this is the most natural (or more familiar) context for PF.

An objective of this study is to contribute to a deeper understanding of the cognitive abilities necessary for the production of PF, and what this can tell us about intonational competence as part of a broader pragmatic competence. Our data suggest that already from two years of age children are able to convey their affirmation or denial of an attributed proposition or thought; and, by doing so, they demonstrate an early intention reading ability. The mastery of PF production can be seen as an early linguistic manifestation of children's abilities to (i) consider the knowledge states of others, and (ii) to convey an attitude to an attributed proposition. In its most complex form, the use of PF also indicate an ability to infer the false beliefs of others, arising around four years of age.

References:

Fretheim, T. (2002). Intonation as a constraint on inferential processing. *Speech Prosody 2002*. http://sprosig.org/sp2002/pdf/fretheim.pdf. **Just, M.A., & Carpenter, P.A. (1971).** Comprehension of negation with quantification. *J.of Verbal Learning and Verbal Behavior, 10*(3). doi:https://doi.org/10.1016/S0022-5371(71)80051-8. **Wellman, H. M., Cross, D., & Watson, J. (2001)**. Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child Development, 72*(3). doi:10.1111/1467-8624.00304. **Wharton, T. (2020).** Acquiring prosody. In K.P. Schneider & E. Ifantidou (Eds.), *Developmental and Clinical Pragmatics* (Vol. 13). Berlin: Mouton de Gruyter.

Affect encoding in word embeddings

Yuhan Zhang[1], Wenqi Chen[1], Ruihan Zhang[2], Xiajie Zhang[2]
[1] *Harvard University,* [2] *Massachusetts Institute of Technology*

An increasing trend in natural language processing has been investigating what syntactic and semantic knowledge can be learned by large neural networks (NN) in word embeddings (e.g., Ettinger, 2020; Linzen et al., 2016; Manning et al., 2020). **Here we ask whether word embeddings that are fed into NNs encode intricate lexical semantic meaning.** In particular, we focus on the affect meaning of words, which, according to Osgood et al. (1957), involves three dimensions -- "valence" represents the pleasantness of a word (*nightmare* vs. *love*); "arousal" represents the intensity of emotion invoked by the word (*napping* vs. *abduction*); "dominance" represents the level of control exerted by the word (*weak* vs. *powerful*). We adopted three different analytical methods–principal component analysis, cosine similarity analysis, and a supervised classifier probe–to investigate whether word embeddings encodes information along the three affect dimensions that resemble human judgments. A positive correlation will indicate that the affective meaning is well captured by word embeddings. The human judgments of words' affective values on three dimensions came from the VAD dataset–where 20k English words were annotated by raters based on their perceived values on the aforementioned affective dimensions (Mohammad, 2018). The tested word embeddings were GloVe (Pennington et al., 2014), vanilla BERT embeddings (Devlin et al., 2019), embeddings from BERT-based model fine-tuned on a GoEmotion dataset with 27 emotion categories (Demszky et al., 2020), and our BERT-based contextualized word embeddings derived from aggregating the context-specific word embeddings from a vanilla BERT with the IMDB movie review dataset (Maas et al., 2011).

Figure 1 shows the PCA results with the correlation coefficients between word vectors from human judgments and the two principal components of each word's embeddings. The significant correlation coefficients indicate that the affective meaning is captured by word embeddings. Interestingly, each type of word embedding encodes the affect meaning differently and with diverse strengths. This pattern also parallels with the correlation coefficients of pairwise cosine similarities for 80 strong affect words in Table 1. For our supervised probe, Figure 2 displays the pipeline: we added a linear classifier layer after word embeddings for binary classification tasks on each of the three affect dimensions. The validation and affect word sample test results indicate that BERT-based contextualized embedding performed the best and that the valence dimension was the easiest to predict. Besides, the attention-based model such as BERT is better at capturing the affect meaning of the words compared with unsupervised learning-based models such as GloVe.

Then we ask whether affect-enriched word embeddings improve the performance in downstream affect-related tasks. We compared the performance of (i) the vanilla BERT model and (ii) the BERT-based model fine-tuned on the human labeled VAD dataset, on the task of predicting positive/negative IMDB movie reviews. Figure 3 shows that the affect-enriched model performed better than the vanilla BERT in the 10 epochs under investigation. Noticeably, the performance of the affect-enriched model had been superior to the vanilla model from the very first epoch. As shown in Figure 4, the affect-enriched model improved rapidly as the training progressed. Thus, fine-tuning BERT on affect datasets enhanced the model's performance on downstream sentiment analysis tasks, especially in the small-data regime.

Above all, we provide positive evidence that word embeddings from statistical learning and large neural network models do capture the affect meaning of words, but in different ways which might result from their individual training algorithm. The classifier result indicates that the easiness to predict intricate dimensions of affect meaning differs by the word embedding type, which invites future investigation into neural networks' deep knowledge about meaning. We further show that affect-enriched word embeddings enhance the downstream sentiment-related tasks, which is informative and translatable to other NLP tasks.

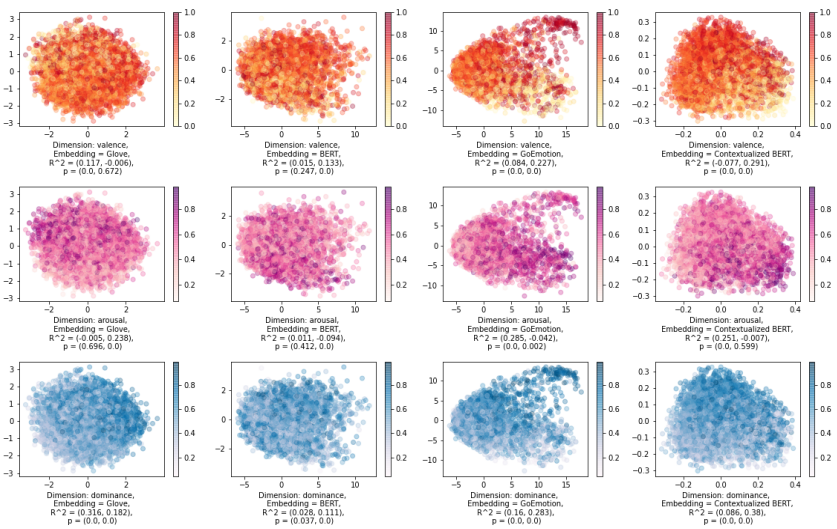


Fig.1 Two components PCA representations of Glove, vanilla BERT, BERT with GoEmotion, Contextualized BERT (Human ratings were color coded in a spectrum. Each dot represents a word. 5586 words are represented. The darker the dot, the more prominent the human rating in the respective dimension.)

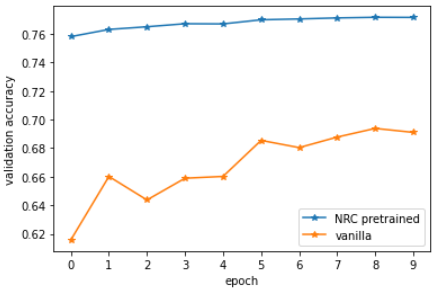


Fig.3 Validation accuracy of two models across 10 training epochs. Blue: the BERT model pretrained on human judgment dataset. Yellow: the vanilla BERT model.

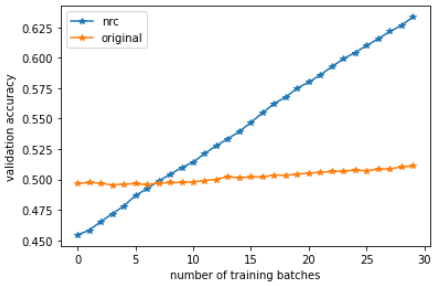


Fig.4 Performance of two models in the small-data regime. Each batch contains 32 training samples. Blue: the BERT model pretrained on human judgment dataset. Yellow: the vanilla BERT model.

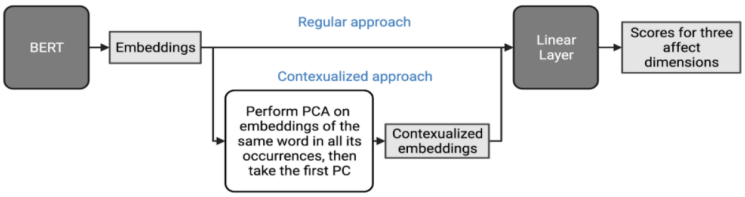


Fig 2. Pipelines of two different approaches for extracting word embeddings. The regular approach is to extract the embedding from the last hidden layer in NLP models such as BERT. The contextualized approach is to perform PCA on embeddings of the same word in the IMDB dataset and take the first principal component of the multiple occurrences.

CORR (p)	VAD	Glove	BERT	GoEmotion BERT	Contextualized BERT
VAD	1.000 (.00)				
Glove	**0.272** (.00)	1.000 (.00)			
BERT	0.116 (.00)	0.148 (.00)	1.000 (.00)		
GoEmotion BERT	**0.252** (.00)	0.172 (.00)	0.013 (.47)	1.000 (.00)	
Contextualized BERT	**0.314** (.00)	**0.710** (.00)	**0.240** (.00)	**0.204** (.00)	1.000 (.00)

Table1. Spearman correlation coefficients (p value) of pairwise cosine similarities between each and the rest of word embedding types from human ratings(0-1) and four types of word embeddings. Correlation coefficients are in bold when larger than 0.2.

Pretrained Embeddings	Validation Accuracy			Affect Word Sample Accuracy		
	Valence	Arousal	Dominance	Valence	Arousal	Dominance
Glove	0.75	0.7	0.73	0.84	0.74	0.75
Context-free BERT	0.76	0.73	0.74	0.93	0.85	0.82
Contexualized BERT	**0.85**	**0.77**	**0.85**	**0.95**	**0.88**	**0.9**
BERT trained on GoEmotion	0.68	0.68	0.7	0.92	0.76	0.76

Table2. Performance of different word embeddings in predicting VAD Lexicon classification labels. This result comes from the linear classification probe model where the labels are the VAD binary classes. Validation accuracy is the prediction accuracy on validation set(2000 words) and affect word sample accuracy is the prediction accuracy on 130 affect words.

Real-time processing of indexical and generic expressions: Insights from, and implications for, COVID-related public health messages

Elsi Kaiser and Jesse Storbeck, University of Southern California <emkaiser@usc.edu>

Much work on real-time referential processing has centered on 3rd person *anaphoric pronouns* (e.g. *she/he*) and has generated foundational insights about antecedent retrieval, salience, and discourse representations. However, the real-time processing of *indexical pronouns*, whose reference changes from one context to another (e.g. *I, you, we*) has received less attention. Indeed, most existing psycholinguistic accounts of pronoun processing are effectively accounts of *anaphoric* pronoun processing, even though indexicals are some of the most frequent and communicatively central parts of language, and semantically fundamentally different from anaphoric pronouns (e.g. Braun 2001). Pronouns with generic reference (e.g. 'you' = people in general, akin to 'one') have also received little attention in real-time processing work.

We use COVID-related health messages to investigate the processing of indexical and generic expressions, with the dual aims of (i) contributing to our understanding of how these experimentally under-researched expressions are processed in real time and (ii) exploring whether the ease of comprehending public health messages related to the COVID pandemic (as measured by reading time) is influenced by type of referring expression.

Prior work on 1st and 2nd person pronouns in COVID messages (and health messages generally) is limited, with mixed results. E.g. Tu et al. (2021) found that COVID stay-at-home messages with *you* are more effective than ones with *we* when it comes to shaping people's self-reported likelihood of staying at home vs. going to a friend's party in a hypothetical scenario. Kaiser (2021)'s work on *you*, *we* and *people/everyone* in subject position found that COVID messages about masks and social distancing with *people/everyone* were rated more convincing by democrats, while non-democrats showed no clear pronoun-type effects.

However, these studies did not measure processing ease. Given the value of easily-understood health messages (e.g. CDC's *Health Communication Playbook*), identifying differences in processing ease of different expressions has theoretical and applied relevance.

Experiment. Using self-paced reading, we tested COVID-related behavior recommendations (ex.1). The study had 39 targets and 48 native English-speaking, U.S.-based participants recruited via Prolific who participated via PCIbex (Zehr & Schwarz 2018). We manipulated referring expression type (*we*, *you*, *people;* within-subjects Latin Square design). The referring expression was preceded by a short preamble phrase (eg. 'on account of the pandemic') and followed by an auxiliary verb, the main verb and the rest of the sentence.

(1) On account of the pandemic, **we/you/people** should get the vaccine to prevent further spread of COVID-19…..

COVID messages with deontic modality like (1) have the advantage of allowing for minimal triplets where the communicative function of *we/you/people* is as similar as possible: in messages like (1), the communicative goal is constant regardless of which of the forms is used. (This is not the case in examples like 'We go to Italy in the summer' vs. 'People go to Italy in the summer.' Thus, health messages are well-suited for comparing the different expressions.)

In (1), presented out of context, *you* is (in principle) ambiguous: On an indexical interpretation, it refers to the addressee (e.g. Brunyé et al. 2009 on *you* triggering a participant perspective). On a generic interpretation, *you* refers to people in general (like 'one'). *We* is also potentially ambiguous between indexical and generic readings (Holmberg 2017), but its generic use is less frequent than generic *you*. In contrast, *people* is not indexical and only receives a generic-type interpretation. We test 3 hypotheses about the processing ease of these forms:

Indexicality hypothesis*:* Indexically-interpreted pronouns refer to highly salient referents, and do not require evoking/constructing a new discourse referent or even a generic

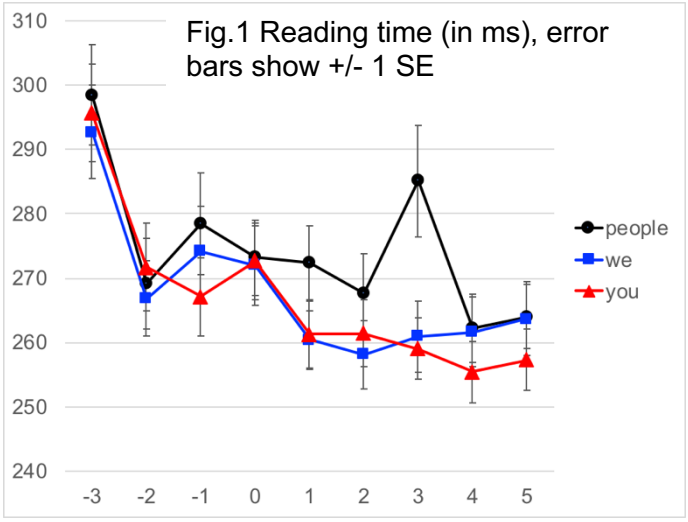
operator/referent. If this special property of indexicals is hard-wired into the representation of these forms, then we might expect any expressions that *can* in principle have an indexical interpretation to be easier to process than a form that *can never* be indexical: you, we < people. The special status of indexicals receives initial indirect support from the results of Warren & Gibson (2002, 2005) on the processing of indexicals in a different context.

***Perspective-taking hypothesis*:** Although indexicals refer to salient referents, they are also perspective-sensitive: The referent of indexical *we* or *you* depends on who the speaker and the addressee are. If this perspective-sensitivity is hard-wired into our processing of certain pronouns (regardless of context), then – in light of prior work suggesting that perspectival processing is cognitively costly (e.g. Keysar et al. 2000, Ferguson et al. 2017) – *you* and *we* may be harder to process than *people* which has no indexical component (people < you, we).

Genericity hypothesis: Given that all 3 forms can receive generic interpretations in contexts like (1), this could render the indexical readings of *we* and *you* irrelevant/unavailable in this context. If so, we may see no differences between the three forms (you = we = people).

Results: Reading time (RT) data is in Fig.1. Mixed-effect regression models were used to analyze log-transformed RTs. Overall, messages with *people* elicit RT slowdowns relative to messages with *you* and *we*, which do not differ – supporting the **Indexicality Hypothesis**.

Specifically, there are no effects of referential form before the critical region, and no effects at the critical region itself (*you/we/people*, "0" in Fig.1). At spillover region 1, *people* conditions are marginally slower than the *you* (p=0.061) and *we* (p=0.0768) conditions. At spillover region 2, there are no significant differences. At spillover region 3, *people* conditions are significantly slower than *you* (p<.005) and *we* (p<.005) conditions. At spillover region 4, there is still a marginal slowdown in *people* conditions relative to *you* conditions (p=0.09) but no other differences. Spillover region 5 shows no significant differences.



(As the word *people* is longer and less frequent than *we* or *you*, the marginal slowdowns in spillover region 1 may be due to these surface factors. Crucially, these effects are not significant in region 2. Thus, it seems reasonable to view the slowdown in region 3 as a meaningful indication that *people* sentences trigger slowdowns relative to *you* and *we* for reasons independent of word length/frequency.)

We have also conducted a **between-subjects version** of this study (n=48 new people) with the same items and method but with referring expression type manipulated between-subjects, which **replicates** the finding that *people* conditions are read more slowly than *you* or *we* conditions.

Conclusions. This study takes initial steps to explore the real-time processing of non-anaphoric pronouns by focusing on indexical and generic forms in COVID health messages. To the best of our knowledge, this is the first real-time study to test how -- in COVID health messages -- different forms (*you, we, people*) impact reading time, which we take to reflect ease of processing. Our results point to an increased processing load in messages with the non-indexical form *people* (relative to pronouns *we* and *you*) which (i) we interpret as providing initial support for the Indexicality Hypothesis, and which (ii) also has practical implications for the construction of easily-understood public health messages.

4-year-olds' interpretation of additive *too* in question comprehension

Hisao Kurokami, Daniel Goodhue, Valentine Hacquard, and Jeffrey Lidz
University of Maryland, College Park

Introduction: Additive particles like English *too* contribute an additive presupposition to sentence meaning: e.g., in uttering (1), a speaker not only asserts that Mickey ate a banana but also presupposes that Mickey ate something else in addition.

(1) Mickey ate a BANANA too

The previous literature is divided over when children understand this additive presupposition. [1, 3, 6, and 7] report children's difficulty with the additive presupposition in various languages, well into their school years. However, these initial studies tested children's comprehension in contexts where the presupposition was not supported. Later studies address this problem, but again the findings are split: [2] find that German-acquiring children as young as 3 understand the additive presupposition of *auch* 'also', though the study may overestimate children's comprehension, given their high (*auch*-less) baseline; [5] find that English-acquiring 4-year-olds understand the additive presupposition of *also*, but not at the level of German-acquiring children in [2]. Here we adapt [5]'s task with some methodological improvements to test whether 4-year-olds perform better with the English particle *too*, which appears much more frequently in children's input than *also*. We find that they do: 4-year-olds successfully consider *too*'s presupposition and use that information to restrict the range of possible answers to a *wh*-question.

Experiment: Alongside a puppet, participants listen to short stories about Mickey, Minnie, and Donald, who each complete some tasks (e.g., eating fruit as in (2)). After each story, the experimenter asks the puppet a question like *Who ate the most fruit?* in (3). The puppet first responds by recounting the story, as in (4). This plays a crucial role in setting up a natural context in which the additive presupposition of *too* is supported ([4]). Having forgotten some details, the puppet proceeds to ask a target question like *Who ate a BANANA (too)?* in (5), with or without *too* (between participants design). In [5]'s original design, the puppet's recount of the story contained a VP-ellipsis (e.g., *Mickey and Minnie did* <~~*eat an apple*~~>). We eliminated this potential confound as resolving an ellipsis and assessing a presupposition simultaneously could place extra demand on children's processing, hindering their performance.

(2) **Sample story:** Mickey, Minnie, and Donald are going to eat fruit for breakfast. There are apples and bananas to eat. Mickey says, "I just woke up so I'm not that hungry. I'll just eat one fruit." Look, Mickey eats an apple! Mickey then says, "that was delicious. I'll eat another fruit!" Look, Mickey eats a banana! Minnie says, "eating too much fruit is not good for me. I'll just eat one fruit." Look, Minnie eats an apple! Donald says, "I ate a lot for dinner yesterday, so I'm not hungry. I'll just eat one fruit." Look, Donald eats a banana!

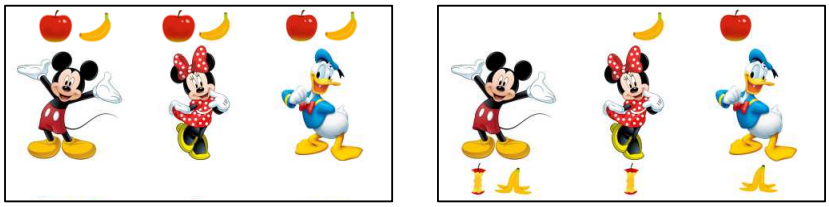


Figure 1. First and last scenes from an animated PowerPoint slide accompanying (2)

(3) **Experimenter's question:** Alright, Charlie. Who ate the most fruit in this story?
(4) **Puppet's recount of the story:** Well, let's see. There were apples and bananas to eat. Donald didn't eat an apple, but Mickey and Minnie did eat an apple.
(5) **Test question:** And who ate a BANANA (too)?

The dependent variable is whether or not participants answer the target question with the "two-action character" (e.g., Mickey, who ate both an apple and a banana in (2)). If sensitive to *too*'s presupposition, participants in the TOO-condition should choose the two-action character since it is the only character that satisfies both the truth-conditional content of the question (i.e., *x ate a banana*) and the presupposition of *too* (i.e., *x ate something else in addition to a banana*). No such preference is predicted for the NO-TOO-condition, as both banana eaters are truth-conditionally valid answers.

Results: Figure 2 summarizes the results from 32 English-acquiring children (age 4;0-5;0, mean 4;5), displaying the mean % of two-action character responses across two conditions (TOO vs NO-TOO). Since there was no variance in the amount of two-action character responses in the NO-TOO condition, a logistic regression would be inappropriate. Instead, we calculated the 95% confidence interval for the TOO-condition to see if it excludes the results from the NO-TOO condition, and it does: the 95% confidence interval for the TOO-condition is 52.77% and 87.23%.

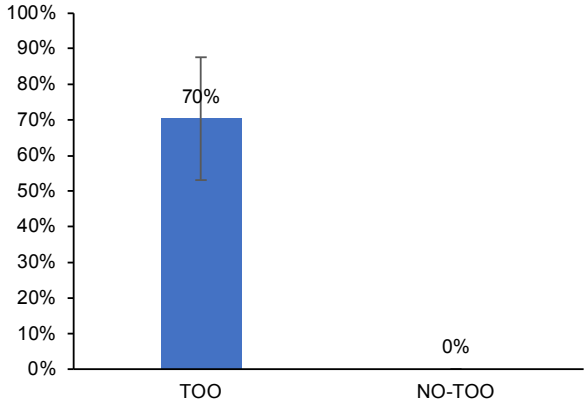


Figure 2. mean % of two-action character responses across two conditions with error bars indicating 95% confidence interval

Discussion: We find that children in the TOO-condition show a strong preference for the two-action character response. In contrast, children in the NO-TOO-condition never gave this type of response, despite being truth-conditionally valid. Since the only difference between conditions is the presence/lack of *too* in the test questions, it's safe to assume that the change in children's behavior is driven by *too* and its presupposition, and that children at this age know *too*'s presupposition. And because our (*too*-less) baseline is zero, we can be confident that our experiment doesn't overestimate children's comprehension. Furthermore, we see an increase in children's performance compared to [5] (20% more two-action character responses in the TOO-condition). Further research will determine whether children's improved performance relative to [5] is due to the difference in the additive particle tested (*too* vs. *also*), or to methodological improvements (no VP-ellipsis in (4)). We also plan to test younger, as well as adult controls on *too* and *also*.

References: [1] Bergsma, W. 2006. (Un)stressed *ook* in Dutch. / [2] Berger, F., & Höhle, B. 2011. Restrictions on addition: Children's interpretation of the focus particles *auch* 'also' and *nur* 'only' in German. / [3] Hüttner, T., Drenhaus, H., van de Vijver, R., & Weissenborn, J. 2004. The acquisition of the German focus particle *auch* 'too': Comprehension does not always precede production. / [4] Kripke, S. 2009. Presupposition and anaphora: Remarks on the formulation of the projection problem. / [5] Kurokami, H, D. Goodhue, V. Hacquard & J. Lidz. Children's interpretation of additive particles *mo* 'also' and *also* in Japanese and English. / [6] Matsuoka, K. 2004. Addressing the syntax/semantics/pragmatics interface: The acquisition of the Japanese additive particle *mo*. / [7] Matsuoka, K., Miyoshi, N., Hoshi, K., Ueda, M., Yabu, I., & Hirata, M. 2006. The acquisition of Japanese focus particles: *Dake* (only) and *mo* (also).

To honor or not to honor: Korean honorifics with mixed status conjoined subjects

Christopher Davis (University of the Ryukyus) & Sunwoo Jeong (Seoul National University)

Background Korean is a language with a rich honorific system, including both addressee-oriented and argument-oriented honorific elements (Lee 1973, 1985; Yun 1993; Kim & Sells 2007; Portner et al. 2018; Choi & Harley 2019, i.a.). This talk focuses on *subject-oriented honorifics*, which are signaled by the verbal suffix *-si* (and optionally the case marker *-kkeyse*; see below). The presence of this suffix signals honorification of the grammatical subject, and its felicitous use is conditioned by the relative social status of the referent of the subject NP and the speaker. In (1), with a high status (elder) subject referent, the honorific form is felicitous, while the non-honorific form is not. With a low status (younger) subject referent, these judgments are reversed.

(1) halapenim-i pata-ey {#ka-ess-ta | **ka-*si*-ess-ta**}.
grandfather-NOM sea-DAT {#go-PST-DECL | **go-HON-PST-DECL**}.
"The grandfather went to the ocean."

(2) ai-ka pata-ey {**ka-ess-ta** | #ka-*si*-ess-ta}.
child-NOM sea-DAT {**go-PST-DECL** | #go-HON-PST-DECL}.
"The child went to the ocean."

These usage patterns, which are also found in Japanese and Yaeyaman, are modeled by Davis (2021) using complementary pragmatic constraints, *UnderHonor and *OverHonor, which militate respectively against underhonoring high status referents (accounting for (1)) and overhonoring low status referents (accounting for (2)). Davis points out that these constraints come into conflict in the case of conjoined subjects with mixed status referents, like the sentence in (3):

(3) ai-wa halapenim-i pata-ey {ka-ss-ta | ka-*si*-ess-ta}.
child-CONJ grandfather-NOM sea-DAT {go-PST-DECL | went-HON-PST-DECL}.
"The child and grandfather went to the ocean."

Experiment 1 We aim to find out how speakers of Korean resolve the conflict in (3), as well as get a firmer empirical understanding of the core contrast in (1)/(2). We also test a suggestion of Kim & Sells (2007) that the order of conjuncts modulates the resolution of cases like (3).
Materials. Stimuli sentences were created by crossing 4 types of subjects with 3 types of honorific marking, as exemplified in (4), where the professor is contextually established as the speaker's advisor, and Yura as the speaker's younger friend. The 4 types of subjects varied in number, status, and conjuct order for conjoined subjects: ***high***, ***low***, ***high-low***, and ***low-high***. The 3 types of honorific marking were: **0** (no subject honorific marking), **HON1** (verbal honorific suffix *si-*), and **HON1+2** (a combination of the verbal honorific suffix *si-* and the honorific nominative case marker *-kkeyse*). As noted by Kim & Sells (2007), *-si* can be used in the absence of the *-kkeyse*, but not vice versa; we included honorific sentences with and without *-kkeyse* to check for any differences between these two honorification strategies.

(4) { kyoswunim | yura | kyoswunim-kwa yura | yura-wa kyoswunim }
{ professor$_{\textbf{\textit{high}}}$ | Yura$_{\textbf{\textit{low}}}$ | professor-and Yura$_{\textbf{\textit{high-low}}}$ | Yura-and professor$_{\textbf{\textit{low-high}}}$ }
-{ i/ka | kkeyse } nonmwun-ul ssu-{ ∅ | *si* }-ess-supnita
-{ NOM | NOM.**HON2** } paper-ACC write-{ **0** | **HON1** }-PAST-DEC

Procedure. 47 Native Korean speakers were recruited as participants. Each participant saw 64 sentences: 8 items crossed with 8 of the 12 possible conditions. Presence vs. absence of honorifics (**no honorifics** vs. honorifics) and subject type were tested within subjects, whereas hon-

orific subtype (choice between **HON1** and **HON1+2**) was tested between subjects. The stimuli were presented in random order. In a given trial, participants were asked to rate the naturalness of a given sentence along a 7-point Likert scale.

Results. We established 3 main empirical findings, each supported by significant interactions between, and robust effects of, subject type and honorific marking (details omitted for space).

First, the results confirm the general pattern noted in (1)-(2). For singular high status subjects (***high***), **HON1** and **HON1+2** were the preferred options, whereas for singular low status subjects (***low***), **0** was the preferred option (two leftmost panels). At the same time however, the results also reveal an asymmetry: Honorifics paired with low status subjects are judged to be more categorically unacceptable than non-honorifics with high status subjects. We interpret this asymmetry as follows: Honorific forms make a positive requirement on the status of the subject referent. Thus, using honorific forms with a low-status subject results in semantic infelicity (or false entailments). By contrast, non-honorific forms are semantically unmarked. The combination with high status subjects is only bad by a kind of pragmatic implicature.

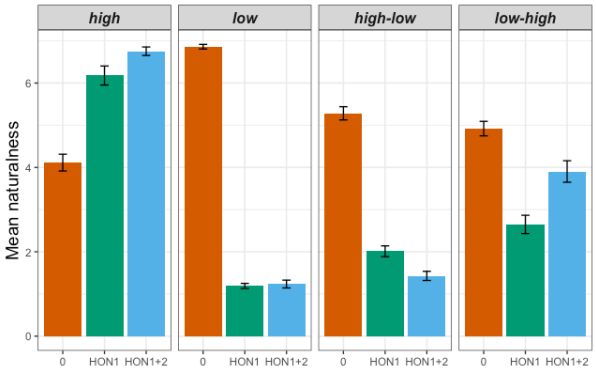


Figure 1: Means & 95% CIs

The second main finding is that there is an overall preference for choosing the non-honorific form in mixed subject cases. Assuming the constraints proposed in Davis (2021), this suggests that in Korean, *OverHonor outranks *UnderHonor. While there is thus an overall preference for using non-honorific forms with mixed subjects, we also found an effect of word order in the acceptability of mixed subjects with honorifics, akin to an 'agree with closet conjunct' effect; in essence, the acceptability of sentences with honorification is boosted in the case of ***low-high***, where the honorific marker(s) appear closer to the high status conjunct. This effect was strongest for **HON1+2**.

These findings are based on a comparison of the mean acceptability scores using mixed effects regression models. Examining the results for individual participants, however, we observe that the overall patterns conflate several distinct patterns/strategies that vary systematically across participants, similar to what Davis (2021) found for Japanese speakers (cf. Han et al. (2016)).

Experiment 2 While the results above suggest that Korean speakers generally prefer non-honorific forms with mixed subjects, they also indicate that conjunct order, which is *not* semantic/pragmatic in nature, modulate this overall preference. In experiment 2 (which we are currently running), we probe for semantic factors that modulate this pattern. In particular, we hypothesize that the naturalness of mixed subject sentences with honorific marking will be boosted when co-occurring with the adverb *hamkkey* 'together', and more degraded in combination with *kakca* 'each'. This hypothesis is based on the intuition that *kakca* forces interpreters to consider each conjunct individually, including the mismatched one, whereas *hamkkey* may enable interpreters to apply the predicate to the plurality denoted by the conjoined NP without considering the status of each conjunct.

Additional Discussion In the full talk, we discuss the ramifications of the experimental evidence for theories of subject honorification, focusing on (i) whether the phenomenon should be modeled via syntactic agreement, and (ii) the semantics and pragmatics of honorification. We also consider how the inter-speaker variation noted in experiment 1 should be modeled, and its consequences for theories of semantics and pragmatics; in the spirit of Han et al. (2016), we argue for the existence of three distinct strategies for resolving honorific conflicts in Korean, each of which is relatively

categorical at the individual level, but giving rise to variation at the population level.

Processing conditionals in context: reading time and electrophysiological responses

Mathias Barthel
Humboldt University Berlin, Institute for the German Language, Mannheim
barthel@ids-mannheim.de

Rosario Tomasello
Freie Universität Berlin

Mingya Liu
Humboldt University Berlin

Background: The concept of conditionality is central to human thought and action. In the formal semantic literature, it has been long debated how to compositionally derive the different meanings that conditionals in natural language convey (e.g., Kratzer 1986, von Fintel 2011). In this paper, we focus on the role of conditional connectives from a semantic and processing perspective. We compare conditionals of the form 'If P, Q' to conditionals of the form 'Only if P, Q' based on the literature of conditional perfection (Geis & Zwicky, 1971; Van der Auwera, 1997; Horn, 2000) and that of *'only if'* (Herburger 2015, 2019). Conditional perfection describes the observation that an *if*-conditional (e.g., 'If you mow the lawn, I will give you 5 dollars.') can receive a stronger – biconditional (i.e., *if and only if*) – interpretation. Following this conception, 'If P, Q' triggers a pragmatic inference of 'If not-P, not-Q' or 'Only if P, Q'.

For the case of 'Only if P, Q', Herburger (2015, 2019) questions whether comprehenders generally draw the inference "If P, Q". For sentences such as "Only if you work hard do you succeed.", she argues that they do not presuppose that "all (normal) instances of hard work will be rewarded by success" in contrast to their *if*-counterparts. We tested the comprehension of contextually embedded conditionals with 'If' versus 'Only if' in a self-paced reading experiment (Experiment 1) and a follow-up EEG experiment (Experiment 2) in German.

Experiment 1: In the self-paced reading experiment, 29 participants (mean age (sd) = 28.5 (8.1) years) read 108 critical short scenarios of four sentences such as (1).

(1) Sentence 1: *DE: Leon besuchte seine Eltern und dachte sich:*
(EN: Leon visited his parents and thought:)
Sentence 2: ***Wenn / Nur wenn*** *die Blumensträuße hübsch sind, bringe ich einen mit.*
(If / Only if the bouquets are pretty, I will take some with me.)
Sentence 3: *Wie sich zeigte, waren die Blumensträuße nicht hübsch.*
(As became apparent, the bouquets were not pretty.)
Sentence 4: *Von denen brachte er* ***einen / keinen*** *mit und ging weiter.*
(Of those he took one / none and went on.)

After an initial context sentence (S1), participants read a conditional sentence with the conditional connective *If* or *Only if* (S2), followed by a sentence negating the antecedent P (S3), followed by a sentence either confirming or negating the consequent Q (S4). The materials thus yield a 2 x 2 design, with Conditional Connective ('If' vs. 'Only if') and Consequent (true or false) as factors. S1 to S3 were presented sentence-by-sentence, while S4 was presented word-by-word. Participants could move on to the next sentence or word by pressing the space bar as soon as they were finished with reading the current sentence or word. Reading times on the positive or negated quantifier (*ein / kein* 'one / no') in S4 served as the critical measure. Following the logic presented in the background above, we predicted reading times of the negated quantifier (i.e., the negated consequent) to be shorter in the case of *'Only if'* compared to *'if',* as 'If not-P, not-Q' is a semantic inference in *'Only if'* and only potentially a pragmatic one in *'If'*. Reading times of the positive quantifier are predicted to be either identical between *'if'* and *'only if'*, since readers should not expect the positive quantifier in either of them, or to be longer in *'only if'* compared to *'if'* (Herburger 2015, 2019).

Reading times for the critical positive quantifier were statistically equivalent between conditional connectives (β=0.13, CI=[-8.6, 8.86], BF_{10}=0.99), reading times for the negative quantifier were shorter for *'Only if'* conditionals than for simple *'if'* conditionals (β=-12.06, CI=[-20.41, -3.81], BF_{10}=121.45) (Figure 1). These findings indicate that the negative quantifier is processed faster after *'Only if'* than after *'If'* conditionals, in line with their semantics.

These results show that comprehenders form distinct predictions about discourse continuations based on differences in the lexical semantics of the tested conditional connectives, shedding light on the role of conditional connectives in the online interpretation of conditionals in general.

Experiment 2: The study aimed to investigate whether the differences in reading times described above may be reflected at the level of brain responses by employing electroencephalography (EEG). To this end, we used an extended set of experimental materials (144 critical items) in an adapted procedure, where both S1 to S3 as well as words in S4 were presented for a fixed duration for participants to silently read for comprehension (1600 ms for S1 to S3; 150 ms for the words in S4, with 500 ms blank in between each sentence/word). In line with the semantics of *'Only if'*-conditionals, the negated quantifier should be pre-activated to a higher degree as compared to simple *'If'*-conditionals, and processing of the negated quantifier should thus be easier in *'Only if'*-conditionals. Hence, we expect greater amplitudes in the N400 component for the negative quantifier in *'If'* conditionals than *'Only if'* conditionals, reflecting the varying degrees of discourse expectations (Kutas & Federmeier, 2011). Informed by the results of the self-paced reading experiment, we predict no difference in the amplitude of the N400 component for the positive quantifier.

Testing of 38 subjects (mean age (sd) = 25.5 (4.9) years) had been delayed due to lab-closures and has only recently been finished, so that final analyses were not ready by the time of submission but will be presented by the time of the conference.

References:

von Fintel, K. (2011). Conditionals. In C. Maienborn, K. von Heusinger, & P. Portner (Eds.), *Semantics: An international handbook of natural language and meaning.* 1515–1538. Berlin: de Gruyter.

Geis, M. L., & Zwicky, A. M. (1971). On Invited Inferences. Linguistic Inquiry, 2(4), 561– 566.

Herburger, E. (2015). Only if: If only we understood it. Proceedings of Sinn Und Bedeutung, 19, 304–321.

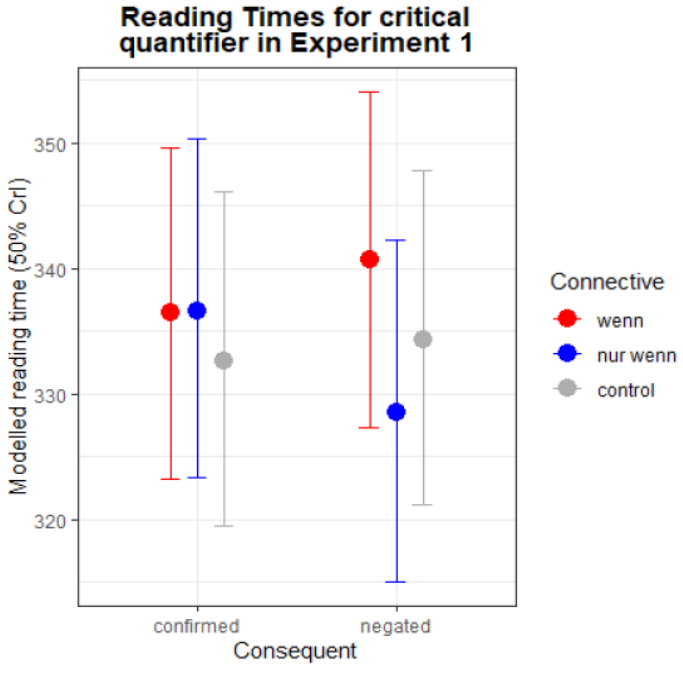
Herburger, E. (2019). Bare conditionals in the red. Linguistics and Philosophy, 42(2), 131–175.

Horn, L. R. (2000). From if to iff: Conditional perfection as pragmatic strengthening. Journal of Pragmatics 32(3), 289–326.

Kratzer, A. (1986). Conditionals. Chicago Linguistics Society, 22(2), 1–15.

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). Annual Review of Psychology, 62(1), 621–647.

Van der Auwera, J. (1997). Pragmatics in the last quarter century: The case of conditional perfection. Journal of Pragmatics, 27(3), 261–274.



Social identity modulates inferences about speaker commitment to projective content

In sentences like *Ken didn't hear that the minimum wage is too low*, the content of the complement (CC) (*the minimum wage is too low*) can "project" out of the entailment-cancelling environment, such that the speaker is taken to be committed to the truth of the CC (e.g., Kiparsky & Kiparsky 1970, Karttunen 1971). Recent experimental work has shown that listener beliefs about the truth of the CC – listener CC beliefs – modulate projection (Degen & Tonhauser 2021). Another type of listener beliefs that might influence projection judgments are *perceived speaker beliefs*: listener beliefs about what the speaker believes with respect to the CC. Indirect evidence for this hypothesis comes from Mahler (2020), who found that social information about the speaker influenced projection judgments for utterances similar to the one in Fig. 1: 'liberal' CCs such as *the minimum wage is too low* were more projective when the speaker was affiliated with a Democrat vs. Republican group (the reverse pattern was found for 'conservative' CCs). Building on Mahler (2020), we directly investigate the role of perceived speaker beliefs in projection inferences. In addition to replicating Mahler's (2020) finding, we find evidence that the effect of social information on projection inferences can be partially attributed to perceived speaker beliefs. As in Degen & Tonhauser (2021), listener CC beliefs also influenced projection inferences. The role of listener CC beliefs in the presence of social information suggests that listeners consider their own beliefs about the CC even when those beliefs are potentially misaligned with perceived speaker beliefs – a finding that has implications for the design and interpretation of comprehension experiments. Overall, the findings are in line with contentions that social and semantic-pragmatic domains of meaning are interconnected (e.g., Burnett 2019; Acton 2021; Beltrama & Schwarz 2021).

Experiment The experiment was conducted online via Prolific. Three experimental blocks were presented in the following order: the listener beliefs block, the speaker evaluation block, and the projection block. The blocks are discussed in a different order than the experiment for expositional clarity. In the projection block (block 3), each target sentence consisted of a 3rd person subject, a clause-embedding predicate, and a complement clause, embedded under negation. 24 of the target sentences involved "political" CCs conveying positions on 12 political topics: half of the CCs conveyed liberal positions on the topics (e.g., *the minimum wage is too low*), and the other half conveyed conservative positions on the same topics (e.g., *the minimum wage is too high*). There were also 12 "neutral" CCs about apolitical topics. Each participant saw 18 target sentences, 6 each with conservative, liberal and neutral CCs, with 6 (of 12 total) predicates. A sample trial is illustrated in Fig. 1. Projection was measured by asking participants about the speaker's certainty with respect to the CC, as in Mahler (2020) and Degen & Tonhauser (2021). Participants responded by adjusting a slider labeled from "no" (0) to "yes" (1). On each trial of the speaker evaluation block (block 2), participants saw one of the speaker profiles associated with the political target sentences from the projection block, but the target sentences themselves were not presented. Participants adjusted sliders in response to questions about their impressions of the speaker, including a question about

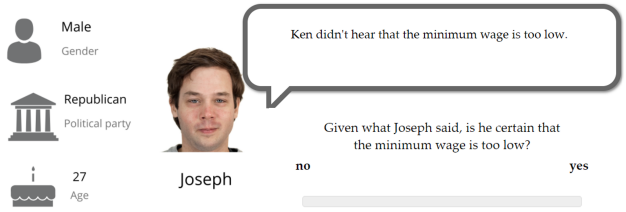
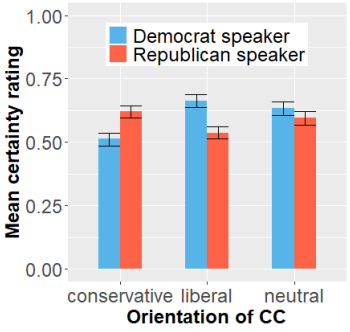


Fig. 1: Example trial in projection block

the speaker's likelihood of believing the CC from the projection block (e.g., *how likely is Joseph to believe that the minimum wage is too low?*). On each trial of the listener beliefs block (block 1), participants were presented with a question about their beliefs with respect to one of the political CCs from the projection block (e.g., *how much do you believe that the minimum wage is too low?*).

Results Data from 212 participants was analyzed using linear mixed-effects models. As illustrated in Fig. 2, certainty ratings for neutral CCs did not differ according to the speaker's political affiliation ($\chi^2(1) = 3.53; p = 0.06$). However, for political items (analyzed in a separate model), certainty ratings were predicted by a significant interaction between the speaker's political affiliation and the CC orientation ($\chi^2(1) = 118.37; p < 0.002$), such that conservative CCs received higher ratings with Republican speakers ($\beta = 0.12, SE = .015$) and liberal CCs received higher ratings with Democrat speakers ($\beta = 0.11, SE = 0.02$). As illustrated in Fig. 3, listener CC belief ratings (from block 1) were a weak but significant predictor of certainty ratings ($\chi^2(1) = 9.958; p < 0.01; \beta = 0.04, SE = 0.01$), while perceived speaker belief ratings were a stronger predictor ($\chi^2(1) = 165.09; p < 0.002; \beta = 0.19, SE = 0.01$). The embedding predicate was also a significant predictor of certainty ratings across all analyses.



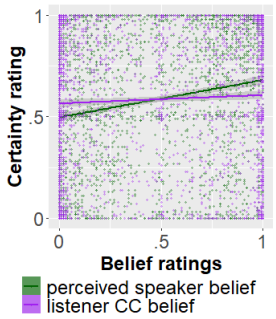


Fig. 2 (left): mean certainty ratings as a function of speaker political affiliation and the orientation of the CC, with 95% confidence intervals. **Fig. 3** (right): individual certainty ratings as a function of perceived speaker belief ratings (green) and listener CC belief ratings (purple) with lines-of-best-fit.

Discussion Our findings replicate the effect of social information on projection found in Mahler (2020), and further suggest that the effect can be partially attributed to listener beliefs. These include perceived speaker beliefs, informed by social information about the speaker, as well as listener beliefs about the CC itself. The role of listener CC beliefs is on one hand consistent with Degen & Tonhauser's (2021) finding. However, it is also somewhat surprising given that listeners' political beliefs (potentially) diverge from the the political beliefs attributed to the speaker. In practice, it seems either that listeners use their own beliefs to "fill in the gaps" when they are not very confident about the speaker's beliefs, or they simply cannot ignore their own beliefs when interpreting someone else's utterance. This has an important implication for the assumptions that researchers make in experimental work on meaning: even when experimental tasks are setup to investigate participants' (a.k.a. listeners') judgments about a speaker's meaning, participants may also consider their own beliefs about what the speaker has said in making that judgment. Moreover, the importance of participants' own beliefs illustrates the multifaceted way in which projection inferences are socially-mediated. These inferences depend not only on social information about the speaker, but also participants' subjective beliefs that shape and are shaped by their social identities.

Selected References Beltrama, A. & F. Schwarz. 2021. Imprecision, personae, and pragmatic reasoning. SALT 31. • Degen, J. & J. Tonhauser. 2021. Prior beliefs modulate projection. Open Mind 5. • Mahler, T. 2020. The social component of the projection behavior of clausal complement contents. LSA Proceedings 5(1).

Can 'hard words' become easy? Mapping evidential meanings onto different forms

Why are some word meanings harder for children to acquire than others? According to a prominent hypothesis, this difficulty stems from the complexity of the underlying concepts.[1] On an alternative proposal, the difficulty often lies in the mapping between linguistic expressions and concepts, even if the concepts themselves are available.[2,3] Here, we offer a novel argument for the role of mapping factors in acquiring a well-known 'hard' case: evidentiality (i.e., the linguistic encoding of the speaker's information source). [4-6]

Using an artificial language learning paradigm, we compare adult learners' acquisition of a single evidential meaning expressed by different linguistic forms (a novel verb/morpheme/adverb). Our goal is to see whether mapping the *same concept* onto *different forms* yields different learning outcomes. We expect the learnability of evidential meanings to differ depending on the linguistic and extra-linguistic (pragmatic) properties of the forms that encode these meanings and, correspondingly, the tools that learners use to extract the commonalities within a particular set of events during form-to-meaning mappings. In a control condition, the same meanings are encoded by a non-linguistic stimulus.

In our experiment, 280 English speakers were shown 5 videos in which a girl gained access to an event through observation (Visual Access), and 5 that involved a third character's report (Reportative Access; mixed order; Fig.1). At the end of each video, the speaker described what happened and marked her own evidential access through an alien *verb* ('I *gorp* she lit the lamp'), *morpheme* ('She *litgorp* the lamp') or *adverb* ('*Gorpingly*, she lit the lamp'). In a fourth, control condition, the speaker uttered a regular English sentence ('I lit the lamp') but her access was marked by a non-linguistic form - a red frame placed around the video. Participants had to figure out what the novel form meant. We crossed two between-subject factors: Form (verb/morpheme/adverb/frame) and Evidential Access Meaning (visual or reportative). Participants later completed a Production task: they watched 8 new videos (4 per access type) and had to use the target form if appropriate. They also completed a Comprehension task: they watched 24 videos (12 per access type) and detect any errors in the use of the form.

We hypothesized that evidential meanings should be more easily discoverable for verbs compared to the control condition because of verb syntax (overt finite sentence complementation); for morphemes and adverbs, no such advantage over the control condition should exist (for adverbs, their placement suggested but did not require sentential scope). Additionally, across linguistic and non-linguistic forms, we hypothesized that pragmatic factors should prioritize marking indirect, potentially unreliable access (e.g., reported information) over direct, more reliable access (e.g., visual perception).[7] Our results confirmed both of these predictions. Evidential verbs were learned better compared to the non-linguistic control (β=-3.12, z=-6.4, p<0.001) but evidential morphemes were harder (β=3.28, z=7.07, p<0.001) and evidential adverbs showed no difference from the control condition. Throughout, reportative evidentials were acquired more easily than visual evidentials (β=-1.29, z=-4.7, p<0.001).

Our findings provide novel evidence in support of the claim that what makes lexical meanings easy or hard to learn, regardless of their conceptual presuppositions, often lies in the transparency of the correspondence between those meanings and the linguistic forms that express them.

Figure 1. Sample screenshots from one video for each Access type: (A) Reportative, (B) Visual. Videos always had the same ending (Panel 5). In that panel, the girl in white either uttered an evidential sentence with an alien verb: "I gorp she lit the lamp", morpheme: "She litgorp the lamp", or adverb: "Gorpingly, she lit the lamp", or offered an unmarked sentence ("She lit the lamp") while a red frame marked the video of the target access throughout the event.

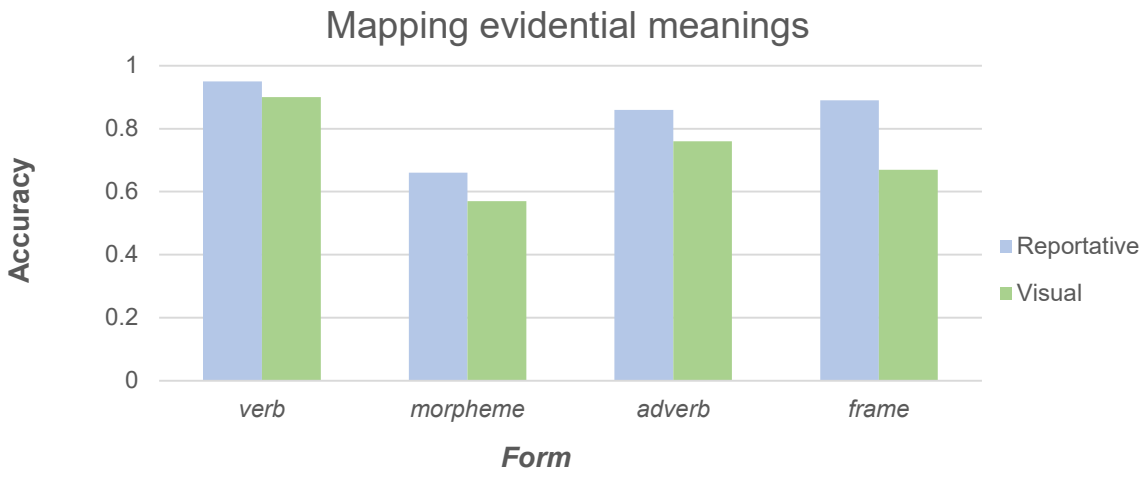


Figure 2. Accuracy in acquiring evidential forms.

References

1. Smiley, P., & Huttenlocher, J. E. (1995). Conceptual development and the child's early words for events, objects, and persons. In M. Tomasello & W. Merriman, *Beyond names for things*. Hillsdale, NJ: Erlbaum.
2. Gleitman, L. R. (1990). The structural sources of verb meanings. *Language Acquisition, 1,* 3–55.
3. Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. (2005). Hard words. *Language Learning and Development*, *1*(1), 23–64.
4. Fitneva, S. (2018). The acquisition of evidentiality. In A. Aikhenvald (Ed.), *The Oxford Handbook of Evidentiality*. Oxford, UK: Oxford University Press
5. Aksu-Koç, A. (1988). *The acquisition of aspect and modality: The case of past reference in Turkish.* Cambridge, England: Cambridge University Press.
6. Ozturk, O., & Papafragou, A. (2016). The Acquisition of Evidentiality and Source Monitoring. *Language Learning and Development*, 12(2), 199–230.
7. Saratsli, D. Bartell, S., & Papafragou, A. (2020). Cross-linguistic frequency and the learnability of semantics: artificial language learning studies of evidentiality. *Cognition*, 197.

Getting to the truth is not easy as putting it in context – A dual task study of negation processing

Shenshen Wang[1], Chao Sun[2], Richard Breheny[1]

[1]University College London, [2]Leibniz Centre for General Linguistics

In negation research, a widely discussed procedure involves the presentation of a visual probe soon (250ms) after reading a sentence. [1] finds that response latencies for images that match states of affairs consistent with the truth of negative sentences are *longer than* mismatch images, which match the positive argument of negation. While this kind of evidence has been argued to support a two-step model of negation processing, [2-4] argue that the prominence of the positive after reading the negative sentence is a result of normal parallel language processes which compute information about the Relevance of a sentence as well as its content, from the same linguistic and contextual cues. They argue that, in the absence of further cues, negation itself provides evidence about a type of context in which the positive state of affairs is at issue. [2-4] manipulated expected QUD using information structural (clefting) or contextual (explicit questions) cues and find reversed effects when the QUD has a negative predicate, rather than positive. We assume that visual probe tasks require participants to generate expectations of visual features in a display, given an object named in the sentence. We conclude then that for simple negative sentences in [1], expectations about context can be generated prior to content. This is consistent with an idea that inferring things about a negative's state of affairs is difficult, especially in comparison to the positive. But this idea has never been directly tested before. We present a dual-task study to determine the relative costs of inferring content and relevance for negative sentences. Our results point to an additional cost to infer the true state of affairs for negative sentences.

Experiment: Two groups of participants undertake a probe task based on [1]. Participants (N=40) in the no-memory load task only did the probe recognition task. In the memory load group, participants (N=41) completed an additional task, which consisted of remembering a simple grid pattern at the beginning of each trial and recreating it after the probe task (Fig. 1). The probe task has a 2(polarity)*2(match) design. Participants read a sentence and then a visual probe is presented at 250ms. The task is to decide if the object in the image is mentioned in the sentence. In test items, images of the mentioned object are either presented in a state which matches the state implied by the sentence (Match) or Mismatches. See Table 1. Fillers counterbalance for response and polarity. Comprehension questions for 25% of trials.

Results: See Figure 2. We constructed a linear mixed-effects model predicting the logarithmised reaction time (RT) from polarity (affirmative or negation), match (match or mismatch) and WM load (no- or memory load). The results showed highly significant main effects of polarity and match (ps<.001). There were significant interactions between WM load and match (p=.007), and between polarity and match (p=.005). Crucially, the three-way interaction was significant (p=.05). To further examine the interaction, we broke down the analyses by load condition using a fitted mixed-effects model predicting RT from polarity and match for each load group. The post hoc analyses revealed that the no-memory load group showed only main effects of match and polarity (ps<.001), whereas the memory load group showed an interaction between polarity and match (p=.001).

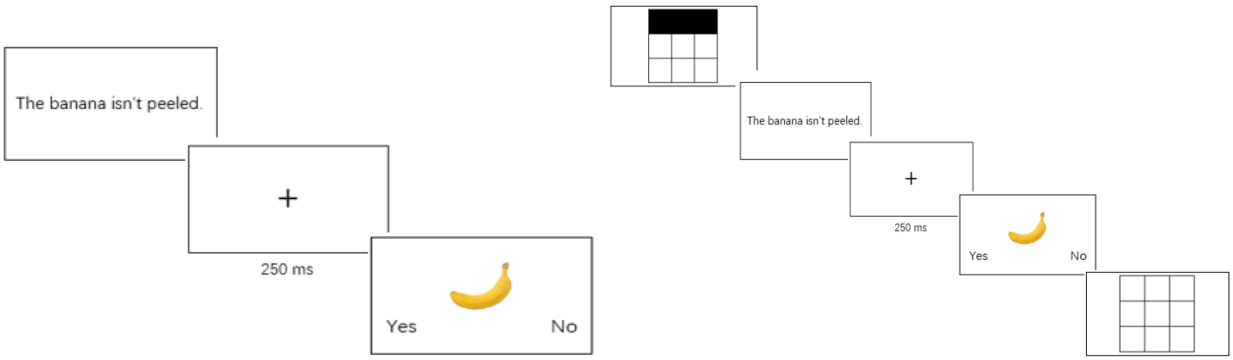
Discussion: Our no load results do not replicate those found in [1]. Here, Match latencies are faster than MisMatch for negative sentences. We attribute this to our items having a negative state of affairs (soa) which is simpler to infer (*not peeled banana*, *not open door*), while [1] use predicates with less obvious antonymic states (*bird not in the air*). Overall increased RT for

negatives in no-load group are nevertheless consistent with idea of competition between positive context and negative content soas. In WM/Negative trials, we see evidence for increased advantage for positive over negative soa (RT(M)>RT(MM)) and this suggests that WM load has a greater impact on processes that arrive at expectations for the true soas (content), rather than what soa is under discussion (context). Thus we have rather direct evidence that inferring content for negative sentences comes at a cost which is more susceptible to resource limitations than inferring relevant context. This is consistent with our contention that linguistic stimuli are processed to compute how an utterance is meant to be relevant in parallel with computations to derive inferences from semantic interpretation of sentence.

Polarity	Match	Example Sentence	Display
Affirmative	Match	The banana is peeled.	
	Mismatch	The banana is peeled.	
Negative	Match	The banana isn't peeled.	
	Mismatch	The banana isn't peeled.	

Table 1. Example items for probe task. 2(Polarity) * 2(Match) design.

Figure 1. Procedures of no-memory load and memory load tasks (e.g. is of a Negative-Match trial).



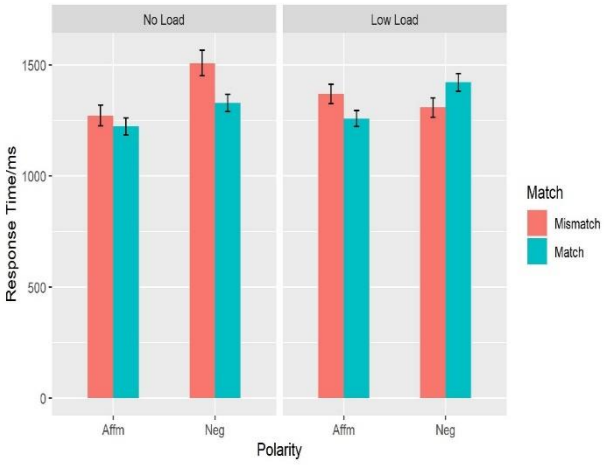


Figure 2. Mean RT for each polarity, match, and WM group. Error bars represent standard errors of the mean.

References: **[1]** Kaup, Yaxley, Madden, Zwaan, & Lüdtke (2007). QJEP, 60, 976–990. **[2]** Tian, Breheny, & Ferguson. (2010). QJEP, 63(12), 2305–2312. **[3]** Tian, Ferguson, & Breheny, (2016). LCN. 31, 683-698. **[4]** Wang, Sun, Tian, & Breheny. (2021). J. of Psycholinguistic Research.

Multiple pressures to explain the 'not all' gap

Jeremy Kuhn and Lena Pasalskaya
Institut Jean Nicod (CNRS), Ecole Normale Supérieure, PSL University

Overview Horn (1973) famously observes that languages frequently lexicalize three corners of Aristotle's square of opposition (*all*, *none*, *some*), but rarely lexicalize the concept 'not-all'. This generalization is robust across languages and across domains of quantification: times (*always*, *never*, *sometimes* vs. **not-always*) and worlds (*required*, *forbidden*, *allowed* vs. **not-required*). Horn (1973) explains part of this observation using pragmatic mechanisms: specifically, *some* implicates *not all* (by competition with *all*), and *not all* implicates *some* (by competition with *none*). The two statements are thus contextually equivalent, so natural language does not need to lexicalize all four meanings—three corners of the square suffice.

Why then do languages lexicalize *some* and not *not-all*? Two hypotheses have been proposed. On the MARKEDNESS HYPOTHESIS, monotone decreasing operators are inherently more difficult to process than monotone increasing operators, possibly due to a simpler cognitive representation (Katzir & Singh 2013). On the INFORMATIVITY HYPOTHESIS, the properties denoted by nouns, verbs, and adjectives generally hold of a minority of objects (e.g. more things are not purple than are purple). As a consequence, '*Something is P*' is usually more informative—and thus more useful—than '*Not all things are P*' from a probabilistic perspective (Enguehard & Spector 2021). Note that these hypothesized pressures are not mutually exclusive.

Here, we describe new predictions of these theories, which we test in an experimental setting. First, we present crosslinguistic data that suggests that the pressure to not lexicalize *not-all* is weaker for modal quantification than for individual quantification. We show that this can in principle be explained by the informativity hypothesis (but not the markedness hypothesis) since the relevant probabilistic properties depend on contingent facts about the lexicon and the world.

We then measure these probabilistic properties in an online experiment in which subjects evaluate the surprisingness of quantificational statements. The results provide evidence for a *combination of both pressures*. Overall, the pressure from markedness is stronger than the pressure from informativity, but informativity still plays a role to explain differences between different domains.

Differences between domains? Typologically, there may be evidence that differences exist between the three domains of quantification. While the lexicalization biases can be found in some form for each, the biases seem to be less strong for modal quantification than they do for individual quantification. In English, for example, the paradigm *possible*, *necessary*, *impossible*, *unnecessary* fills all four corners of modal quantification. In French Sign Language, deontic '*not-all*' modals include the morphologically complex PAS-BESOIN (derived from universal affirmative BESOIN) as well as the morphologically simplex PAS-LA-PEINE. But neither English nor LSF has a single word to express 'not-all' for individual quantification.

The informativity hypothesis has the ability to explain such differences between quantificational domains. For example, there are many activities that people ought to do, but don't. Consequently, while (1b) is probably more surprising than (1a), the judgment for (2) is less clear. The modal *not required* will thus be more informative than the individual quantifier *not everybody*.

(1) a. John is required to help.
b. John is not allowed to help.

(2) a. Everybody helped.
b. Nobody helped.

If such facts hold generally across the verbal lexicon, they will affect lexicalization biases.

In contrast, on the markedness hypothesis, there is no difference between quantification over individuals, times, or worlds. In each case, the representation of a monotone decreasing operator is equally complex; there should thus be no differential effect between domains.

Experiment The informativity hypothesis is grounded on intuitions about the lexicon (specifically, the supposition that lexicalized properties generally hold of a minority of objects), but Enguehard & Spector (2021) do not test this assumption experimentally. We did so here. Subjects were asked to judge the degree to which the situations described by quantified statements were surprising, on a continuous scale from 'Not at all surprising' to 'Very surprising.' We tested 'All' (*everybody*/*always*/*required*) and 'None' (*nobody*/*never*/*not allowed*) for 75 of the most frequent English verbs and adjectives; on each screen, subjects judged two quantified sentences with the same predicate, as in (3). The experiment had one block for each domain: subjects saw the same predicates for quantification over individuals, times, and worlds.

(3) a. Everybody said something. Not at all surprising ..... Very surprising
b. Nobody said anything. Not at all surprising ..... Very surprising

The informativity hypothesis makes two predictions, shown in (4). First, if the general tendency to not lexicalize *not-all* arises from informativity, then *All* statements should usually be more surprising than *None* statements for each domain. Second, if the weakening of this tendency for the modal domain arises from informativity, then the difference in surprisingness of *Required* minus *Not allowed* should be less than that of *Everybody* minus *Nobody*.

(4) Prediction #1: $All_D > None_D$ for each domain D
Prediction #2: $All_{\text{world}} - None_{\text{world}} < All_{\text{indiv}} - None_{\text{indiv}}$

Results As shown in Figure 1, the experimental results manifestly did not confirm Prediction #1. For each domain, *None* statements were judged to be more surprising than *All* statements. But, as shown in Figure 2, Prediction #2 was borne out: the *All* − *None* measure was significantly lower for modal quantification than for individual (or temporal) quantification (on a Wilcoxon Signed-Rank test: $z = 2.9307, p = .00338$).

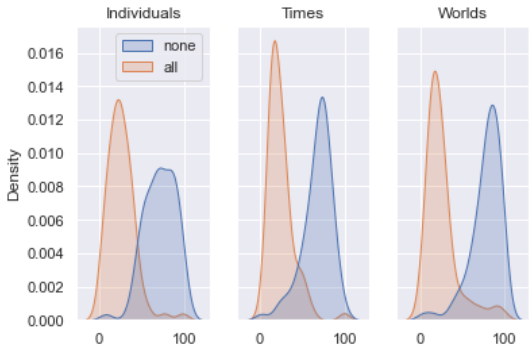


Figure 1: Distribution of surprisingness by item.

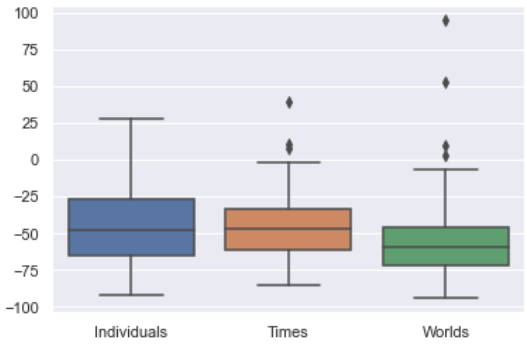


Figure 2: Box plots of *All* − *None* by item.

Discussion These results can be explained as arising from a combination of the two pressures. The experimental results show that *Not-all* statements are in fact usually more informative than *Some* statements. This suggests that any informativity bias (in favor of *not-all*) is overridden by a markedness bias (against *not-all*). On the other hand, evidence of an informativity bias emerges in the differential effects: *not-required* is even more informative than *not-everybody*, leading to exceptional lexicalization of *not-all* in the modal domain. Inspecting the data by item supports this interpretation: the trend appears to be driven by predicates like *help*, *understand*, and *be sure*, which carry a strong moral imperative that may not be satisfied in practice.

One notable finding is that the underlying supposition of the informativity hypothesis (above: 'more things are not purple than purple') doesn't actually hold for how people use language in practice. Certainly '*Everybody did the homework assignment*' is is very surprising if one quantifies over a random sample of

the 5000 students at a small college, but RELEVANCE plays a enormous role restricting the domain to just those individuals who are expected to do the homework.

Finally, more typological work is needed to establish the differential lexicalization tendencies. Useful examples may come from sign languages, which frequently show suppletive negative forms.

References Enguehard & Spector (2021). Explaining gaps in the logical lexicon of natural languages. *S&P*. • Horn (1973). *On the semantic properties of logical operators in English*. UCLA thesis. • Katzir & Singh (2013). Constraints on the lexicalization of logical operators. *L&P*.

Addressing unexpected questions in discourse

Swantje Tönnis and Judith Tonhauser
Stuttgart University

Previous research has assumed a broad range of linguistic phenomena to be sensitive to questions in discourse (e.g., Roberts 1996/2012, Beaver & Clark 2008, Rojas-Esponda 2014, Onea 2016). There have, however, only been few experimental investigations of the question-based structure of discourse (e.g., Kehler & Rohde 2017, Westera & Rohde 2019); in particular, there are no investigations on when unexpected questions can be addressed. In this paper, we contribute to filling this gap by providing evidence for two hypotheses: Exp 1 uses a novel experimental design to show that, in German narrative discourses, questions that are expected to be addressed become more unexpected as the discourse proceeds. Exp 2, a case study on German clefts, shows that relatively unexpected questions can in fact be addressed (in line with Tönnis 2021), not only the most recently-introduced question (as in Roberts 1996/2021).

Data and previous research: Most previous research focused on how to address expected questions. Roberts (2012), for instance, claimed that a discourse move must address the top-most question on the QUD stack, or sub-questions thereof. Extending this, Rojas-Esponda (2014) proposed that it is also possible to address super-questions of the top-most question. Kehler & Rohde (2017) assumed that addressees form a probability distribution over possible questions that the ensuing utterance is going to address. Tönnis (2021) pointed out that this distribution changes when discourse proceeds. For example, the question Q1 is more expected in (1) than in (2).

(1) *When Lilly joined breakfast the rolls were already gone.*
Q1: Who ate the last roll? [relatively expected question]

(2) *When Lilly joined breakfast the rolls were already gone. There weren't any croissants or toast either. So she went to the bakery nearby.*
Q1: Who ate the last roll? [relatively unexpected question]

Roberts (1996/2012) and Rojas-Esponda (2014) predict that Q1 cannot be addressed in the next sentence of (2), given that it is neither the top-most question nor a super-question. Following Onea (2016) and Kehler & Rohde (2017), Tönnis (2021) argued that in German Q1 in (2) can in fact be addressed, namely by a cleft (*It was Benni who ate the last roll*), which she assumed to mark that a relatively unexpected question is addressed. Tönnis' (2021) discourse analysis assumed that an expectedness value is assigned to each possible question at each stage of a discourse. This value represents how strongly the addressee expects the respective question to be addressed in its context. She assumed that the expectedness of a question is higher the smaller the distance of the question to the question-raising sentence is. The question Q1 is raised by the sentence in (1), and is predicted to be more expected in context (1) than in context (2). Exp 1 tests this prediction while the Exp 2 tests whether the expectedness of the addressed question affects the acceptability of German clefts.

Previous experiments mainly focused on eliciting questions which are evoked in discourse. Kehler & Rohde (2017) used continuation tasks, which showed that linguistic cues affect the identification of the QUD. Westera & Rohde (2019) used an elicitation task to investigate which questions arise to readers in text snippets taken from corpora. However, those methods only covered expected questions. In our paradigm, it is possible to also target unexpected questions, which is necessary to test Tönnis' (2021) hypotheses.

Experiment 1 (n=80): Expectedness was measured for 16 German questions in 2 conditions: after the first sentence of a discourse, as in (1), and after the third sentence of a discourse, as in (2). For each discourse, an array of 5 different questions was presented consisting of a question raised by the first (Q1:*Who ate the last roll*), second (Q2:*What could Lilly have for breakfast instead?*) and third (Q3:*What did Lilly buy at the bakery?*) sentence, a very unexpected control (Q–:*What*

was the weather in Colombia?), and a relatively expected control (Q+:*What did Lilly do next?*). Participants were asked to rate the expectedness of each question to be addressed in the next sentence on a sliding scale from 'absolutely unexpected' (coded as 0) to 'very expected' (coded as 100). The expectedness of Q1 was evaluated while the other questions served as baselines.

Results Exp 1: The mean ratings of Q1 were significantly higher after the first sentence than after the third sentence, see Fig. 1 for the expectedness means of all five questions in both conditions. The result was confirmed by a linear mixed effects model (R, *lme4*) that predicted the expectedness rating of Q1 from a fixed effect of number of context sentences (reference level: one context sentence) with participant and item as random effects and a by-participant slope (β = -29.3, *SE* = 2, *t* = -15, $p < .001$).

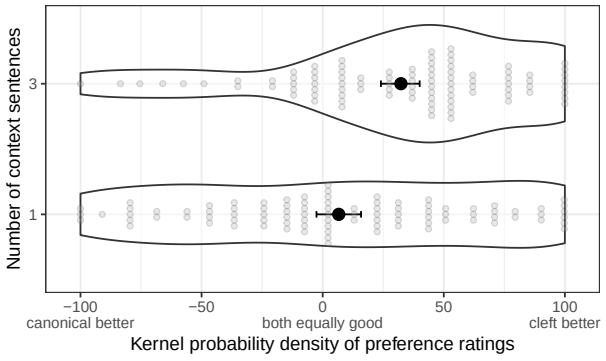
Figure 1: Mean expectedness by question and number of context sentences.

Experiment 2 (n=120): Relative preference ratings for German clefts (e.g., *It was Benni who ate the last role*) compared to their canonical variants (*Benni ate the last role*) were measured for 16 target items both after the first and after the third sentence. The cleft and the canonical sentence both addressed the question raised by the first sentence. Participants were told that the next sentence of the text was illegible, and they were asked to indicate their relative preference between the two alternatives (A and B) on a slider ranging from 'A (canonical) much better' (coded as [-100,0]) to 'B (cleft) much better' (coded as [0,100]), and 'equally good' in the middle.

Results Exp 2: There was a significantly stronger preference for the cleft after the third sentence than after the first sentence, see Fig. 2. This result is supported by a linear mixed effects model that predicted the relative preference rating from a fixed effect of number of context sentences (reference level: one context sentence) with participant and item as random effects and a by-participant slope (β = 26, *SE* = 5.8, *t* = 4.5, $p < .001$). Given the results of Exp 1, this means that clefts are more acceptable when they address a relatively unexpected question.

Discussion: The results support Tönnis' (2021) hypotheses. Exp 1, furthermore, revealed that our method is suitable to attest different levels of expectedness of questions. Crucially, it can also investigate relatively unexpected questions, which cannot be elicited using the designs described above. Exp 2 showed that the QUD may very well be a relatively unexpected question as long as it is addressed with a cleft. This result speaks in favor of more flexible discourse models with respect to which questions can be addressed. The method we introduce could be used as a general paradigm for investigating further phenomena affected by discourse expectations.

Figure 2: Preference ratings by number of context sentences. Black dots represent means with 95% CIs. Light dots represent participants' means.



Selected references • Kehler & Rohde (2017). Evaluating an expectation-driven question-under-discussion model of discourse interpretation. *Discourse Processes*. • Roberts (2012). Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. *Semantics and Pragmatics*. • Tönnis (2021). *German es-Clefts in Discourse. A Question-Based Analysis Involving Expectedness*. • Westera & Rohde (2019). Asking between the lines: Elicitation of evoked questions in text. *Amsterdam Colloquium*.

Predicting the f*ing word: an eye-tracking study on negative expressive adjectives**

Camilo R. Ronderos, Filippo Domaneschi

Theoretical work on negative expressive adjectives such as “fucking” (Table 1) has argued that these convey a speaker-oriented attitude, which constitutes a separate, expressive, dimension of meaning (Potts, 2005, Harris & Potts, 2009, Gutzmann, 2019, i.a.). As such, their meaning is computed outside of compositional meaning construction. Their relative syntactic flexibility supports this perspective: they can be attached to different constituents keeping an identical expressive meaning (see the final two sentences in Table 1) (Frazier et al., 2014).

Recent experimental work has shown that expressives convey emotional and social content (Donahoo & Lai, 2020). However, it’s unclear what it means for an expressive to be *speaker-oriented* from a language processing perspective: Can comprehenders automatically and rapidly retrieve a speaker’s perspective via the expressive, or is this a delayed and effortful inferential process? Further, what purpose does their syntactic flexibility serve?

In the current study, we address these questions by claiming that negative expressives serve the specific comprehender-oriented purpose of reducing processing effort. To investigate this, we tested two novel hypotheses using an eye-tracking, Visual World Paradigm (VWP). We hypothesized that (i) comprehenders can automatically and locally use expressives as indices of a speaker’s perspective in order to anticipate an upcoming referent - but only if they have knowledge of the speaker’s perspective. We also hypothesized that (ii) an expressive’s syntactic flexibility allows for even earlier anticipation of a referent, representing an added cognitive benefit.

DESIGN We created a VWP where 60 native Italian speakers[30f;MeanAge=24.72;SD=7.04] read a discourse context. They then heard a spoken utterance completing the discourse (see Table 1) while visualizing four images (Figure 1). Participants had to then select the correct visual referent and answer a subsequent comprehension question. In 10 critical items, the context introduced two potential referents (Target and Competitor images, Figure 1). The discourse either introduced a speaker’s negative attitude towards the Target referent, or had a neutral statement (Factor ATTITUDE, levels ‘neutral’ vs. ‘supportive’). The spoken utterance contained a negative expressive that either modified the target referent (In-Situ) or the subject of the sentence (Ex-Situ) (Factor: EXPRESSIVE POSITION, levels: In-Situ vs. Ex-Situ), resulting in a 2X2 design. Participants also saw 18 filler trials, which had different combinations of number of referents and speaker’s intentions in order to prevent participants from developing a strategy.

ANALYSIS We calculated proportions of looks to the images, time-locked to the beginning of the disambiguating word (*Cappello,* in target sentence of Table 1). The purpose was to investigate any anticipation effects by analyzing four 500 ms. time-windows: three prior to the onset of the disambiguating word, and one after (Figure 2). A preference for the target image in regions 1-3 would suggest that anticipation occurred. We fitted maximal LMEMs to each region using the log-ratio of looks to target image divided by looks to competitor image as dependent variable. Positive log-ratios represent a preference for the target image. We included (treatment contrast-coded) ATTITUDE and EXPRESSIVE POSITION and their interaction as predictors.

RESULTS *InSitu*- and *ExSitu*-supportive conditions both showed an anticipation effect in region 3 (both positive and significantly different from 0), supporting hypothesis (i). The ExSitu-Supportive condition was positive and significantly different from zero in all four regions. This suggests an earlier anticipatory effect brought on by the early appearance of the expressive, in line with hypothesis (ii). Neutral conditions were only positive in region 4, suggesting no anticipation. They were also significantly different from both *InSitu*- and *ExSitu*-supportive conditions in region 3. This suggests that knowledge of the speaker’s perspective is crucial to understand the meaning of a negative expressive, in line with hypothesis (i).

CONCLUSION Our study proposes a pivotal role for negative expressives during language processing: They aid in anticipatory sentence comprehension by automatically indexing a speaker’s perspective. Their syntactic flexibility is a tool that can be used to ease processing load by allowing for anticipation to take place even earlier in comprehension. This amounts to a unique processing benefit for comprehenders.

CONTEXT	Elena e Martino vogliono acquistare un regalo di compleanno per una loro amica. Martino vuole acquistare un cappello. Elena and Martino want to buy a birthday present for a friend of theirs. Martino wants to buy a hat	
	NEUTRAL	SUPPORTIVE
	A Elena piace tanto fare shopping e vorrebbe acquistare uno zaino Elena loves shopping and would like to buy a backpack	Elena ripete più volte che il cappello è bruttissimo e vorrebbe invece acquistare uno zaino Elena repeats several times that the hat is very ugly and would like to buy a backpack
	Quando suona alla porta il corriere, Elena dice: When the delivery man rings the door, Elena says:	
	IN SITU	EX-SITU
	Il corriere ci ha portato **il cazzo** di cappello The delivery man brought us the **fucking** hat	***Il cazzo*** di corriere ci ha portato il cappello The **fucking** delivery man brought us the hat

Table 1 *Example critical item in all four conditions. Original in Italian with English translation. The context was presented in written form, while the target sentence was played through speakers.*

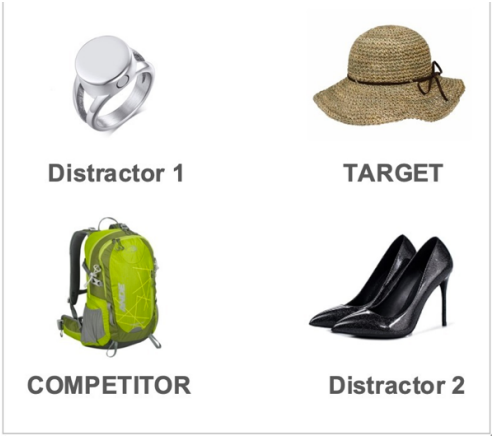


Figure 1 *Example images. Top right and bottom left represent Target and Competitor images respectively.*

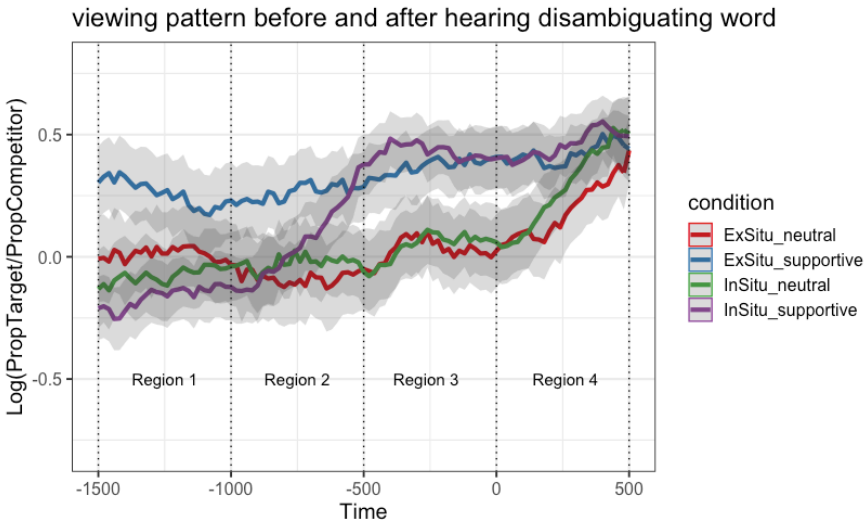


Figure 2 *Log-gaze probability ratios of looks to target to looks to competitor, time-locked to the disambiguating word ('hat', in the example shown in Table 1). Values above zero signify a preference for the target picture. Gray ribbons are confidence intervals.*

References

Donahoo, S. A., & Lai, V. T. (2020). The mental representation and social aspect of expressives. *Cognition and Emotion*, *34*(7), 1423-1438.

Frazier, L., Dillon, B., & Clifton Jr., C. (2014) A note on interpreting *damn* expressives: transferring the blame. *Language and Cognition,* 1-14.

Gutzmann Daniel (2019). The Grammar of Expressivity, Oxford Studies in Theoretical Linguistics, Oxford.

Harris J.A., Potts C. (2009), Perspective-shifting with appositives and expressives, *Linguistics and Philosophy*, 32, pp. 523–552.

Potts C. (2005), *The Logic of Conventional Implicatures*. Oxford Studies in Theoretical Linguistics. Oxford: Oxford University Press.

The enduring effects of default focus in *let alone* ellipsis: Evidence from pupillometry

Jesse Harris, UCLA (jharris@humnet.ucla.edu)

Introduction. In contrastive clausal ellipsis, the remnant is placed in focal contrast with its correlate (Winkler 2018 for review). A particularly intriguing case is focus-sensitive coordination (FSC), like *John can't run a mile, let alone a marathon* (Fillmore et al., 1988). Harris (2016) analyzed the material following the coordinator (*let alone*) as a focus-marked remnant to clausal ellipsis (e.g., *let alone* $[_{FOC}$ *a marathon*$]_1$ ~~*John run* t_1~~*)*; see also Toosarvandani (2010). To interpret the remnant (*a marathon*), the processor locates the contrasting correlate phrase (*a mile*) in the prior clause from among other same-category competitors using multiple, possibly competing, preferences. Experimental and corpus research finds that the nearest possible correlate is vastly preferred (*Locality Bias*; Harris & Carlson, 2015). Similar biases have been observed for other clausal ellipsis structures, like sluicing (Frazier & Clifton, 1998), and replacives (Carlson, 2013). However, semantic and prosodic parallelism have also been shown to interact with Locality (Harris & Carlson 2015), suggesting a general, but violable, preference for pairing a remnant with a correlate that is maximally similar along multiple dimensions.

The tradeoff between Locality and prosodic marking in *let alone ellipsis* was explored by Harris & Carlson (2018). In an auditory corpus of radio interviews, every correlate and remnant in an FSC bore pitch accent, usually an L+H* contrastive accent (79% on correlates and 73% on remnants). The corpus revealed a strong Locality bias: 88% of remnants contrasted with the most local correlate. In auditory naturalness ratings studies, they observed a penalty for non-local (subject) correlates over local (object) correlates. Although pitch accent on subject correlates reduced the penalty for violating Locality, it did not eliminate it.

To explain why the preferred correlate did not simply match the location of pitch accent, they proposed that correlate selection was subject to *Enduring Focus*: "Locations that typically bear default focus continue to provide potential locations for focus, regardless of overt markers of focus", a constraint that might be particularly strong in ellipsis processing. An unaccented Local object noun would therefore continue to provide a tempting correlate, despite lacking overt pitch accent. However, it is unclear whether the impact of default focus is limited to post-sentence interpretation or is active in real time, as well. This study employs pupillometry, an implicit measure of cognitive load or effort, to assess whether default focus locations tempt the processor during online auditory sentence processing.

Method and design. Pupillometry measures minute changes in pupil diameter associated with a stimulus, typically peaking between 700 and 1200ms after stimulus offset (Laeng, et al., 2012). Increased pupil dilation is associated with greater cognitive load, and crucially, does not appear to be under strategic control. Pupil size has recently been explored as a dynamic measure of language comprehension (Schmidtke, 2017 for review), finding increased pupil size for syntactically complex sentences (Engelhardt et al., 2010), metrical violations (Scheepers et al., 2013), and inadequate or misleading pitch accent (Zellin et al., 2011; Breiss et al., 2021).

20 quartets crossed Pitch Accent location (Object/Subject PA) and Remnant contrast (Subject/Object Remnant), operationalized as animate and inanimate nouns, respectively; Table 1. Sentence stimuli were produced with contrastive L+H* accent on the correlate and the remnant, an L-H% boundary tone before *let alone* and after the remnant, as is typically found in corpora. Two seconds of acoustically identical material was spliced into the recording after the remnant following 100ms of computer-generated silence that served as the baseline for measuring pupil change. In half of the items, the Subject Remnant was locally plausible as an object to the verb (*Jonah sent Daniel*); the other half were not (*#The patient ate her family*). Although Harris & Carlson (2018) found no effects of local plausibility in ratings, implausible nouns have been shown to produce N400 online penalties in gapping constructions (Kaan et al., 2004).

Results. 48 native English speakers with self-reported normal hearing listened to sentences over high-quality headphones. Pupil size was recorded with a high-speed eye-tracker for 2 seconds after the offset of the remnant on acoustically identical material within a quartet. Data cleaning followed the recommendations of Mathôt et al. (2018). After removing blinks and other artefacts, and interpolating missing points with spline-smoothing, the data was down-sampled to 10Hz to reduce autocorrelation. The data were then normalized by trial to reflect *change in pupil size over time* by subtracting the mean pupil size obtained from the 100ms baseline, rather than absolute pupil size. Time-series analyses were conducted to capture changes in pupillary excursion. The best-fitting model was a generalized additive mixed effects model (van Rij et al., 2019) with subject-as-object plausibility as a 3-way interactive factor.

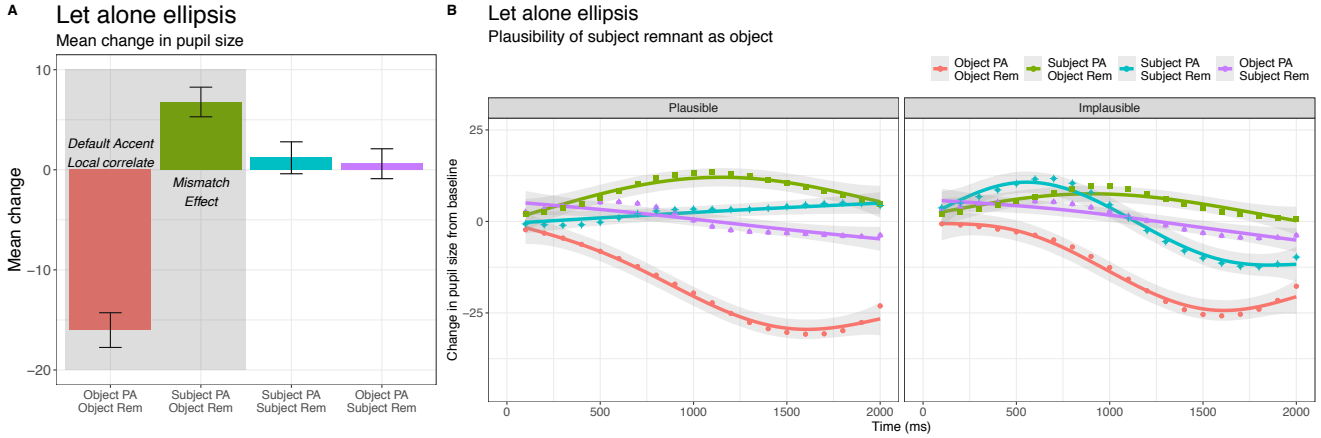
As expected, the baseline condition with both default object accent and a local correlate (Object PA-Object Rem) elicited the lowest pupil response overall; see the leftmost condition in Fig 1A for illustration. Pupil response was greater for subject (vs. object) accent, t = 7.74, p < .001, as well as for subject (vs. object) remnants, t = 7.78, p < .001. The predicted interaction was observed for which pitch accent had little to no effect on pupil size for subject remnants in comparison to the large effect of pitch accent on object remnants, t = -7.21, p < .001. This interaction was further moderated by local plausibility, t = 1.97, p < .05, shown in Fig 1B. In emmeans, a penalty for subject remnants was observed when the subject was plausible as an object, t = 2.52, p < .05, but not when it was implausible as an object, t = 1.09. In both cases, subject remnants appeared to be more taxing than the baseline.

Conclusion. The findings largely support Harris & Carlson's interaction between Locality and Enduring Focus in online auditory comprehension. In *let alone* ellipsis, subject remnants elicited a processing cost and failed to show a mismatch penalty when the object correlate bore contrastive accent. The study also presents a novel use of pupillometry to explore the real-time influence of prosodic information to resolve ellipsis structures in sentence comprehension.

Table 1. Sample materials. Accent (Object/Subject) x Remnant contrast (Object/Subject).

	Subject Plausible as Object		**Subject Implausible as Object**	
Pitch Accent	*Object Accent*	*Subject Accent*	*Subject Accent*	*Subject Accent*
Host clause	Jonah wouldn't send a POSTCARD, let alone	JONAH wouldn't send a postcard, let alone	The patient didn't eat DINNER, let alone	The PATIENT didn't eat dinner, let alone
Object Remnant	a LETTER	a LETTER	DESSERT	DESSERT,
Subject Remnant	DANIEL	DANIEL	her FAMILY	her FAMILY
Critical region	[100ms silent baseline] during visiting hours at the local hospital.		[100ms silent baseline] and the parents started to get a little worried.	

Figure 1. Let alone ellipsis. (A) Mean pupil response. (B) Pupillary response for 2000ms.



Social identity & charity: when less precise speakers are held to stricter standards

Expanding work at the socio-semantics interface ([1-2-3] i.a.), we explore the impact of social information on imprecision resolution in a T(ruth)-V(alue) J(udgment) task. We find that imprecise statements from speakers socially expected to be *less* precise are strikingly held to *more* stringent evaluation standards, suggesting a more nuanced interplay between social and semantic meaning than previously thought, while shedding new light on how social factors impact TVJ responses.

RECENT WORK unveiled a bi-directional relationship between social and pragmatic properties of (im-)precision with numerals: comprehenders infer social properties from speakers' levels of (im)precision ([2]), and conversely adjust their precision thresholds based on speaker identity – as recently shown in in a picture selection task ([4]). In this study, participants saw numeral utterances (*It's 3 o' clock*) along with a phone displaying a slightly divergent number ("2:51"), as well as a face down phone; they had to select which phone they thought the speaker was basing their utterance on. Screens showing divergent numbers were selected more often with speakers embodying a **Chill** (vs. **Nerdy**) persona, indicating higher propensity to accept imprecise numerals from speakers socially expected to speak less precisely – especially for comprehenders who did not themselves identify with the speaker's stereotypical traits. These findings raise the question as to whether speaker identity similarly affects the acceptance of an imprecise description when comprehenders are conversely asked to determine whether a given description fits a state of affairs – the type of inference typically involved in TVJ tasks, a standard experimental paradigm for interpretation judgements ([5-6-7]). Beyond offering a potential cross-paradigm validation of [4]'s findings, this extension is also of general methodological value, as it constitutes a first step towards investigating the role of social information in TVJ tasks – a widely used measure in experimental studies of meaning, whose sensitivity to speaker identity considerations is uncharted.

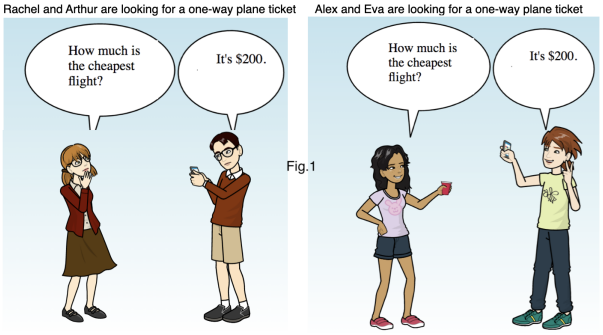


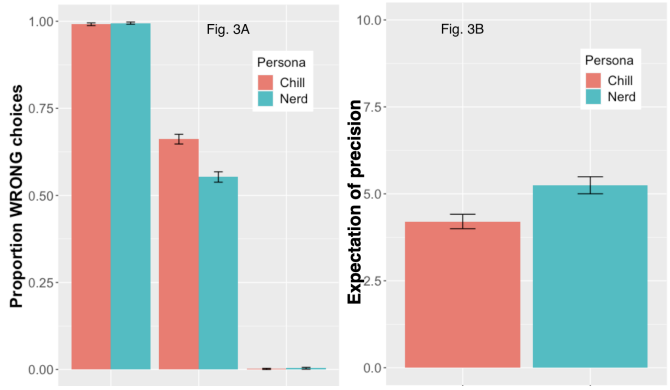
Fig.1

DESIGN. Following [4], we presented dialogues with one character asking a question and the other providing a numeral utterance response after checking their phone. Crossing two factors in a 2x3 design, **Speaker Persona** and **Match** were manipulated. The former was between subjects, with levels *Nerdy*, expected to speak precisely, and *Chill*, expected to speak imprecisely (Fig.1), normed for precision expectations. The latter manipulated how closely the uttered numeral and the number on the phone matched, with 3 levels (Fig.3): *Match*; *Mismatch*; or *Imprecise* (with a 5-19% range of divergence).

Fig.2
Match MisMatch Imprecise

Participants (n=196; via Prolific) assessed whether, given the number on the screen, the utterance was **Right** or **Wrong**. 24 items were counterbalanced across 4 lists, each with 6 items in *Match* and *Mismatch*, and 12 in *Imprecise* (+ 24 fillers). At the end, participants indicated on a 1(min)-10(max) scale how precisely they expected the character to speak, and to what extent they saw themselves in the character's stereotype (=**Similarity**). If social information affects TVJs for imprecise numerals in the same way as picture selection choices ([4]), imprecise descriptions by Chill speakers should be accepted more often, leading to lower rates of WRONG responses (**H1**). Persona effects should also be more prominent for participants who do not identify with the speaker (**H2**). **RESULTS**. Having confirmed ceiling/floor WRONG response rates for Match/Mismatch and intermediate ones for Imprecise, we fit a ME logistic regression on the Imprecise condition data with Persona as a predictor. The rate of WRONG responses is higher for Chill than for Nerdy speakers (β=2.17, p$<$.05; Fig.3A), suggesting

more stringent precision thresholds for the former – *contra* the findings in [4] – even though Chill speakers were still rated to be *less* precise (p <0.001 in the post-questionnaire; Fig.3B).



To test possible modulation by participants' own identity, we fit a second ME model on the Imprecise condition data looking at the interaction of Persona (ref=Chill) and Similarity (ref=1) (Fig.4A): we find a simple effect of Persona at low levels of similarity, with a higher rate of WRONG responses for Chill (β=3.17, p<.05); and a near-significant interaction Persona/ Similarity (β=0.42, p=.08), with the persona effect decreasing as participant-speaker similarity increases (as in [4]). Again precision expectations show the opposite pattern of choices: Nerdy speakers license higher precision expectations than Chill at low Similarity (Fig.4B).

DISCUSSION. These findings provide further evidence that the social identity of the speaker affects comprehenders' behavior in a task that requires computing an imprecision threshold. Contrary to what happens in picture selection [4], in a TVJ task, comprehenders are *less inclined* to accept imprecise statements from Chill speakers; in both paradigms, however, the persona effect is maximally prominent for participants who don't identify with the speaker (similar to phonetic processing [8-9]).

We propose that the different patterns are grounded in the distinct epistemic implications of rejecting an imprecise numeral in the two paradigms. While in picture selection ([4]) rejecting the imprecise number is compatible with taking the speaker to be truthful, a WRONG choice in a TVJ is crucially prejudicial – it commits the respondent to implying that the speaker is violating Quality. Accordingly, Chill speakers' stereotype as imprecise makes it easier to see them as violating Quality than Nerdy speakers, socially perceived as more accurate, leading respondents to be more charitable towards Nerdy than Chill speakers – even though numerals uttered by the former are actually expected to be more precise, and thus (in principle) more likely to prompt a WRONG response. We conclude that social information can impact comprehenders' assessment of utterances in two different ways: it can yield adjustments in precision thresholds with response behavior aligned with precision expectations (as in [4]); or it can yield higher levels of charity towards one persona as opposed to the other, in contrast with precision expectations. This shows that social information affects TVJs', and that these effects might go in the opposite direction of those observed in other tasks tapping into meaning intuitions, complementing methodological work investigating how different experimental tasks inform our understanding of interpretation ([10-11]).

[**1**] Acton & Potts 2014. That straight talk. *J. of Sociolx*•[**2**] Beltrama, Solt & Burnett. Context, (im)precision, and social perception. *Language in Society* • [**3**] Burnett 2019. Signalling Games, Sociolinguistic Variation and the Construction of Style. • [**4**] Beltrama & Schwarz 2021. Imprecision, personae & pragmatic reasoning. *SALT 31*• [**5**] Crain & Thornton 1998. Investigations in Universal Grammar • [**6**] Bott & Noveck 2004. Some utterances are underinformative. *J.MemLan* • [**7**] Papafragou & Musolino 2003. Scalar implicatures... *Cognition* •[**8**] Niedzielski 1999. The Effect of Social Information on the Perception of Sociolinguistic Variables. *J of Lg & Soc. Psych.* •[**9**] Wade 2021. Experimental evidence for expectation-driven linguistic convergence. *Language* • [**10**]Scontras & Pearl 2021. When pragmatics matters . . . *Glossa*• [**11**]Waldon & Degen 2020. Modeling Behavior in TVJT. *Society for Computation in Linguistics*.

Tracking the activation of scalar alternatives with semantic priming

Introduction. We investigate the psycholinguistic mechanisms underlying scalar implicature. Using semantic priming with lexical decision, we find facilitated reaction times to scalar alternatives, which provides evidence that they are retrieved and activated in the computation of scalar implicature.

Background. In calculating scalar implicature (SI), hearers infer messages beyond what is literally, explicitly said by the speaker. In (1), the SI *not all* is inferred, while in (2), *not excellent* is inferred.

(1) Mary ate some of the cookies. → SI: Mary ate some, but not all, of the cookies.

(2) The movie is good. → SI: The movie is good, but not excellent.

The standard assumption is that the inferential process that gives rise to SI involves hearers reasoning about what the speaker could have said, but did not. But it is an open question precisely what psycholinguistic mechanisms underlie this inferential process. Additionally, theoretical accounts disagree on what is involved in the inferential process. Neo-Gricean accounts typically assume that hearers infer the negation of informationally stronger alternatives that the speaker could have said, and that these alternatives are determined via the lexicon or grammar (i.a. Horn, 1972; Katzir, 2007) —e.g., because <*some, all*> form a lexical scale, and *all* is stronger than *some*, hearers derive *not all* upon encountering *some*. But there exist other, Post-Gricean accounts, which take scalar inference to be a contextually driven, conceptual process, whereby utterances undergo strengthening as an instance of ad hoc concept construal, with lexical scales playing no special role (i.a. Sperber and Wilson, 1995).

In this study, we investigate the psycholinguistic reflexes of SI. Specifically, we use semantic priming with lexical decision to test whether lexical alternatives are retrieved and activated in the processing of SI-triggering sentences. The general logic of our experiments is to probe whether alternatives like *all* and *excellent* are recognized with facilitated reaction times in a lexical decision task when they follow a relevant SI-triggering sentence. Similar methods have been successfully used to investigate the activation of alternatives in sentential focus. For instance, Husband and Ferreira (2015) (see also Braun and Tagliapietra, 2009; Yan and Calhoun, 2019) auditorily presented participants with sentences such as *The murderer killed the NURSE last Tuesday night*, and found that visually presented focus alternatives, e.g., *doctor*, were recognized faster as a word of English. The activation of alternatives in SI has, however, not been tested in this way. (For work on priming and scalar inference more generally, see Schwarz et al. (2016), who subliminally primed participants with the stronger alternative before they read the weaker one in an SI-triggering sentence, as well as de Carvalho et al. (2016), who investigated whether scalar terms prime each other in the absence of a sentential context.)

Experiment 1: Sentential priming. Capitalizing on the scalar diversity phenomenon (i.a. van Tiel et al., 2016), our testing ground for the activation of alternatives is 60 lexical scales (adjectival, verbal, adverbial and quantifier). In Exp. 1, participants (N=46) saw an SI-triggering sentence such as *The movie is good*, which was presented word-by-word. Participants then saw the scalar alternative *excellent*, and had to indicate by button press whether this word was a word of English or not. We refer to this experimental condition as the "related" condition. In the "unrelated" condition, participants were first presented with a sentence that was unrelated to the lexical scale, e.g., they saw *The movie is foreign* before making a lexical decision on *excellent*. In addition to the 60 lexical scales, there were 60 fillers items with non-words (e.g., *kleens*, *spraize*), which were preceded by unrelated sentences.

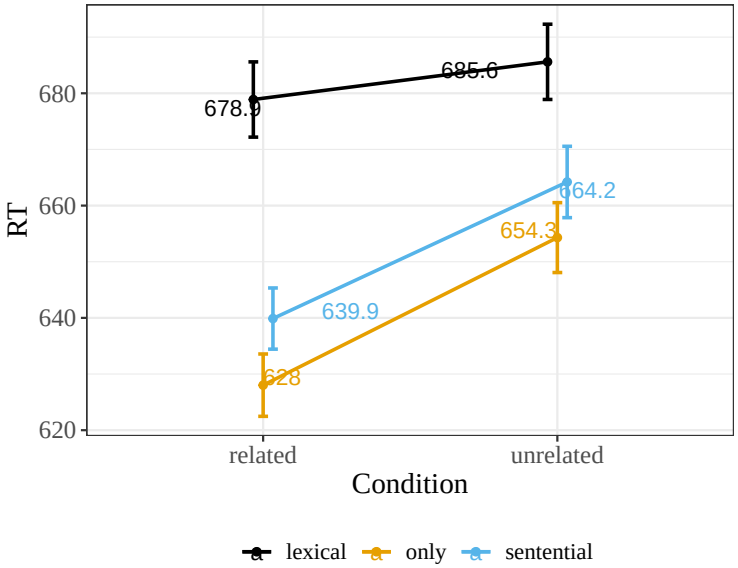
Predictions. If lexical scalar alternatives like *all* and *excellent* are reasoned about, and retrieved in the process of SI-calculation, then we should see facilitated reaction times in the related condition, as compared to the unrelated condition. That is, *excellent* should be recognized faster when it follows an SI-triggering sentence where it serves as a stronger alternative, than when it follows an unrelated sentence. On the contrary, if lexical alternatives do not play a role in the processing of SI, then there should be no difference in reaction times between the related and unrelated conditions.

Experiment 2: Priming with "only". For comparison, we conducted a version of the experiment

where the prime sentences also included the focus particle *only*. That is, participants (N=43) saw sentences like *The movie is only good* before they had to make a lexical decision on *excellent*. Because the exclusion of alternatives in sentential focus is encoded in the semantics (Rooth 1992, 1985), and previous work has found that focus alternatives are indeed primed, we have a strong prediction that we should find facilitated reaction times in Exp. 2, which can then provide a baseline for Exp. 1.

Experiment 3: Lexical priming. In Exp. 3, we investigated the lexical priming of stronger alternatives, given the weaker alternative, but without any sentential context. This is to rule out the possibility that semantic priming might occur unrelated to SI processing, simply because pairs of scalar terms are semantically related. Here, participants (N=44) were presented with single words (*good* vs. *foreign*) as the prime, and responded to *excellent* afterwards —otherwise, the design of the experiment was the same as Exp. 1. If semantic priming arises due to similarities in meaning between scalar terms such as *good-excellent*, then we should see facilitated reaction times in the related condition in Exp. 3, serving as a control for SI-related priming in the sentential experiments.

Results and discussion. The figure on the right shows the results of our experiments. In the control, lexical experiment (Exp. 3), we found no effect of Condition ($p = 0.26$): targets in the related condition were not recognized significantly faster than in the unrelated condition. This shows that pairs of scalar terms are not sufficiently semantically related to result in semantic priming, and therefore any priming effect we find in sentential experiments is due to alternative retrieval, not just mere meaning similarity. In the sentential experiment (Exp. 1), we indeed found a significant effect of Condition ($p < 0.5$): targets were recognized faster following an SI-triggering sentence. This provides evidence for the retrieval and activation of alternatives in SI processing, and supports Neo-Gricean accounts of SI, in which hearers reason about particular lexical alternatives. On the other hand, such results are not predicted by theoretical accounts of SI that dispense with lexical scales, such as Post-Gricean accounts.



The experiment with *only* (Exp. 2) also revealed a significant effect of Condition ($p < 0.01$). Exp. 1 and 2 pattern alike: analyzing the two data sets together, we find no significant difference between them ($p = 0.59$). This suggests that alternatives like *excellent* are similarly activated no matter whether the sentence that is processed is *The movie is good* or *The movie is only good*. This presents a puzzle: in a separate set of experiments, we investigated the rate of inference calculation for SI-triggering sentences and sentences with *only*, and found that the latter lead to higher rates of inferences. The lack of a difference between the current Exp. 1 and 2 suggests that activation of alternatives, as measured via priming, does not track the rate of inference from the corresponding sentences.

Conclusion. In a series of semantic priming experiments, we have addressed an open question regarding the psycholinguistic processing of scalar implicature. We have found evidence that lexical alternatives (*all*, *excellent*) are retrieved and activated in the real-time processing of SI-triggering sentences. In addition to informing our understanding of the mental representations behind pragmatic reasoning, our findings also help adjudicate between theoretical accounts of SI, and are most in line with Neo-Gricean theories.

Group membership impact on pragmatic inferences

Inbal Kuperwasser[a], Yoav Bar-Anan[b], Einat Shetreet[a,c]
[a] Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel
[b] Department of Psychology, Tel Aviv University, Tel Aviv, Israel
[c] Department of Linguistics, Tel Aviv University, Tel Aviv, Israel

Correspondence to: inbalk@mail.tau.ac.il

Understanding the meaning of non-literal language involves linguistic, cognitive and social processes, including using our knowledge about the speaker's identity. Previous research highlighted the importance of the social characteristic of either the speaker or the listener (e.g., occupation, social rank, accent, etc.), but group membership effects on pragmatic processing have been largely overlooked. This is despite their theoretical likelihood, because both group membership (e.g., Hackel et al., 2014) and pragmatic processing (e.g., Fairchild & Papafragou, 2021) have been tied to Theory of Mind and executive functions. In this study, we asked whether high threat intergroup settings impacted the interpretation of a well-studied pragmatic phenomenon, scalar implicatures (SI). SIs concern sentences with weak scalar terms (e.g., *some*), for which listeners typically assume the speaker cannot use the more informative expression *all*, leading them to reject the logical meaning (*some and possibly all*), and adopt the pragmatic interpretation (*some but not all*).

We conducted an online experiment (N=180) using a simple judgement task. Participants were American native English speakers who identified themselves as either Democrats or Republicans. To avoid intergroup task effects, we divided the participants into three groups: (i) an ingroup condition where the participants were exposed to members of their own group, (ii) an outgroup condition where the participants were exposed to members of the other group, (iii) a control group, to serve as a baseline. The number of Democrats and Republicans was balanced across the groups.

In the experimental groups, participants first had to indicate their political affiliation by clicking on the appropriate party logo and answered a group identification questionnaire (adapted from Leach et al., 2008). In the control group, participants were asked general personality questions not concerned with group (adapted from Chang et al., 2016). All the participants were then told they will play a “game” with other (virtual-decoy) players in the game (4 in total, gender balanced). In the experimental groups, the party affiliation of the speakers was constantly highlighted (Fig 1). The “game” was a simple judgment task where the participants were asked to decide whether a statement given by one of the other players matched a picture shown on screen. The 8 critical trials (of a total of 24) included statements with the scalar term *some* with a picture where all the entities in the picture shared the relevant trait (Fig 2). A 'matching' response was categorized as logical, and 'not matching' response as pragmatic.

In a mixed-effects model, we modelled the rates of pragmatic responses with a fixed effects of group (control/ingroup/outgroup) and a random effect of party affiliation (without a random slope). The model revealed an effect of group, and follow-up pairwise comparisons showed significant difference between all groups ($p < 0.001$ for the control vs. ingroup and the control vs. outgroup comparisons, and $p=0.019$ for the ingroup vs. outgroup comparison; Fig 3), so that the control condition had the highest percentage of pragmatic response, followed by the outgroup condition and then the ingroup condition. We further examined the relation between

pragmatic responses in the ingroup and outgroup conditions with the scores of the group identification questionnaire, but found no significant correlation.

We showed that a high-threat intergroup setting impacted the interpretation of SIs and generally increased logical interpretations. We assume that this effect originates from different reasons in the two groups. Because both “yes” and “no” responses are acceptable with underinformative statements, participants in the ingroup condition were likely to present ingroup favoritism, and tended to agree with the speaker from their own group (meaning to say the picture 'matched' the statement), leading to many logical responses. This could not be true for the outgroup condition. We therefore hypothesis that in this group the effect may be the result of resource depletion due to the need to inhibit attitudes in intergroup settings, or of difficulty in mentalizing. To elucidate these results, we are currently conducting a version of this study without direct judgment to control for the ingroup favoritism.

Fig 1 – Example of a speaker's introduction before each statement

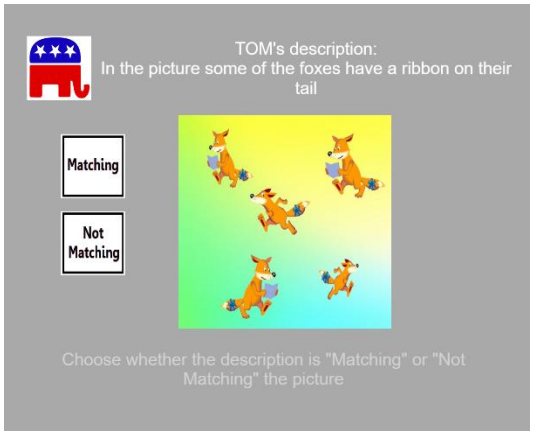


Fig 2 – Example of a critical trial in the experiment. The speaker uses *'some'* to describe an *'all'* situation

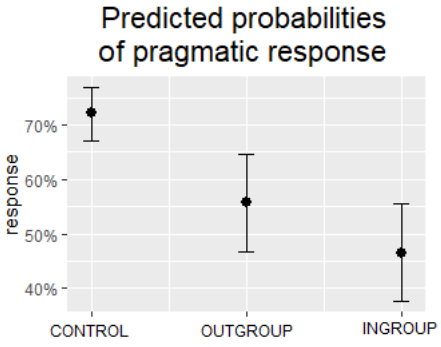


Fig 3 – The predicted probabilities plot of a pragmatic response (*some but not all*) to SI by group type (control/outgroup/ingroup)

References:

Chang, L. W., Krosch, A. R., & Cikara, M. (2016). Effects of intergroup threat on mind, brain, and behavior. *Current Opinion in Psychology*, *11*, 69-73.

Fairchild, S., & Papafragou, A. (2021). The Role of Executive Function and Theory of Mind in Pragmatic Computations. *Cognitive Science*, *45*(2), e12938.

Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology*, *52*, 15-23.

Leach, C. W., Van Zomeren, M., Zebel, S., Vliek, M. L., Pennekamp, S. F., Doosje, B., ... & Spears, R. (2008). Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *Journal of personality and social psychology*, *95*(1), 144.

On the interpretation of German *einige*: The effect of tense and cardinality

Maya Cortez-Espinoza and Lea Fricke, Karl-Franzens-Universität Graz

Introduction. Scalar Implicatures (SIs) have been a research interest since Horn in 1972 (see Breheny, 2019 for an overview). However, to our knowledge, the effect of tense on SIs has only once been experimentally investigated in a study on the exclusivity inference of *or* (Chierchia et al., 2000). In line with these previous results, we hypothesize that the SI *not all* is less likely drawn in future than in past tense sentences involving *einige* 'some'. Differences in the question under discussion (QUD) (Roberts, 2012) are argued to strongly influence the computation of SIs (Kuppevelt, 1996 and Zondervan, 2011). We assume that probabilities of questions to be the immediate QUD differ between past and future tense. For the past, a QUD inquiring about details is more likely as receiving precise information is likely as well. Therefore, the SI is drawn for past tense, see (1-b). With the future being inherently uncertain, a more general QUD asking for rough estimations can be argued to be more likely. Therefore, (1-a) can be interpreted without the SI.

(1) a. John will eat some apples.
interpreted as: 'John will eat some or all the apples.'
b. John ate some apples.
interpreted as: 'John ate some and not all apples.'

More concretely, we hypothesize that in contexts that violate the SI, sentences like (1-a) receive higher acceptance rates than sentences like (1-b). Besides investigating this hypothesis, we tested sets of varying size as representatives of German *einige* 'some' in our experiment. Previous experiments on English *some* (van Tiel/Geurts, 2013, Degen/Tanenhaus, 2015) suggests that higher cardinalities are regarded as more typical representatives of a phrase like *some N* than smaller cardinalities (≤ 3). Based on these data and introspection, we expect a prototypicality effect, such that larger numbers are more typical representatives of *einige* than smaller numbers. That is, we expect sentences like in (1), interpretated with the SI, to be more acceptable in a context in which John ate 6 apples than in a context in which he ate only 2. We also expect singleton sets to be particularly bad due to the plurality inference of the plural NP that serves as the first argument of *einige* (Tieu et al., 2014). We employed an experimental paradigm, which aims to foster rational behavior in participants by financially rewarding them for choosing the optimal response to each stimulus.

Method. We tested 32 participants (mean age = 23.8 years, SD = 5.5 years, 15 female and 17 male participants), who saw 20 test items and 30 fillers, of which some served as controls. The target sentences were conditional statements containing the scalar term *einige*. They were presented in the context of a story about 9 candidates in a reality show who did activities together. The stimuli had the form of bets about activities to happen on the show and the participants' task was to decide on whether if were won, thereby judging the truth of the target sentences. Each judgement had a direct impact on the budget they received in the beginning of the experiment. Figure 1 is a example of a stimulus for past tense. We had an additional hypothesis about upward and downward entailing environments that we will not report on due to a procedure error that happened on the level of participant instruction which turned part of the data invalid. For this reason, only the unaffected data with the scalar term in the conditional is analysed and shown. The stimuli were shown along with a table which indicated for each candidate whether she was involved in the activity in question with the candidate number of positives ranging from 0 to 9 ('cardinality' in the following). The 0-context yields a false target sentence, the 9-context yields an SI-violation and the numbers 1 – 8 constitute different manifestations of *einige*. Additional context resolved the antecedent of the conditional as true. Besides the item manipulations, participants

Bet – Front Side

If the group was instructed to prove its patience, some participants finished a puzzle with 1000 parts.

Context – Back Side

Person	Finished a puzzle with 1000 parts
Alice	Yes
Bella	Yes
Diana	Yes
Flora	Yes
Gabi	Yes
Helena	Yes
Ida	Yes
Clara	No
Elsa	No

The group was instructed to prove its patience.

Figure 1: example bet - English translation

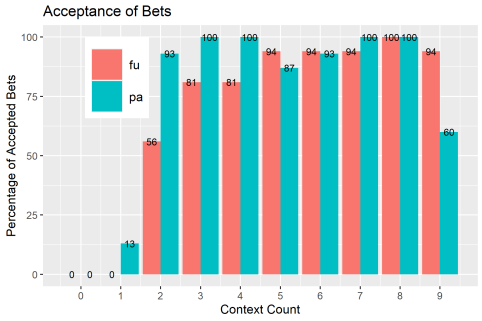


Figure 2: results

were assigned one of two tense levels for which all bets appeared in the according tense.

Results and discussion. Figure 2 shows the acceptance rates of bets by *cardinality* and *tense*. It can be seen that SI-violations are more strongly penalized in past tense than in future tense. Furthermore, the graph shows cardinality 1 in particluar and low cardinalities in general to be bad representatives of *einige*. This seems to be the case especially in future tense. For the inferential statistical analysis, we used the software R (R Core Team, 2017). We constructed three binomial regression models to test our three hypotheses. Model 1 tested whether the interpretation of scalar terms depended on factor tense. For this, we assumed a 2-level factor *SI support* with the levels [+support] (the union of cardinalities 1 – 8) and [-support] (cardinality 9). We found a significant main effect of *tense* ($p < 0.05$) along with an interaction of *tense* and *SI support* ($p < 0.01$), which confirms that in the past tense, more SIs are computed. Model 2 tested the hypothesis that higher cardinalities are better representatives of *einige*. In this model, we included the factor *cardinality* with levels corresponding to the cardinalities 1 – 8 (cardinalitities 0 and 9 were excluded as they do not represent *einige*) and the factor *tense*. There was a significant main effect of *cardinality* ($p < 0.01$) as well as a main effect of *tense* ($p < 0.05$). The same effects occur when we change the levels of *cardinality* from 1 – 8 to 2 – 8. Other than the descriptive results suggest, there was no significant interaction between the two factors. Model 3 tested whether cardinality 1 is a significantly worse representative of *einige* than the other numbers. For this, we assumed a 2-level factor *plurality* with [+plurality] (the union of cardinalities 2 – 8) and [-plurality] (cardinality 1). A main effect of *cardinality* was found ($p < 0.001$) along with a main effect of *tense* ($p < 0.01$), no interaction was found. To sum up, we found that tense influences whether an SI is drawn or not and that higher cardinalities are better representatives of *einige*. These findings add a new dimension to the discussion on scalar implicatures. Future work should replicate the findings concerning the effect of tense. Moreover, it would be interesting to investigate this effect for English comparing *will*- and *going to*-future forms which differ in the certainty with which an event is said to occur.

References: Breheny. 2019. Scalar Implicatures. In Cummins/Katsos (eds.) *The Oxford Handbook of Experimental Semantics and Pragmatics*. Oxford: Oxford University Press. Chierchia, Crain, Guasti, Thornton. 1998. Some'and 'or': A study on the emergence of logical form. *Proceedings of BUCLD 22* 97-108. Somerville: Cascadilla. Geurts, van Tiel. 2013. Embedded scalars. *Semantics and Pragmatics* 6(9) 1-37. R Core Team. 2017. R: A language environment for statistical computing. Technical report, R Foundation for Statistical Computing, Vienna. Roberts. 2012. Information structure in Discourse: Towards an integrated Formal Theory of Pragmatics. *Semantics and Pragmatics* 5(6). Degen, Tanenhaus. 2015. Processing scalar implicature: a constraint-based approach. *Cogn Sci* 39(4) 667-710. Van Kuppelvelt. 1996. Inferring from Topics: Scalar Implicatures as Topic-Dependent Inferences. *Linguistics and Philosophy* 19(4). Zondervan. 2011. The role of QUD and focus on the scalar implicature of most. In Meibauer/Steinbach (eds.) Experimental Pragmatics/Semantics. Amsterdam: Benjamins.

The role of grammatical cues in tracking object location in transfer-of-possession events: A visual-world eye-tracking study

Sarah Hye-yeon Lee[1,2] & Elsi Kaiser[1]
[1]University of Southern California, [2]University of Pennsylvania

Source-goal events involve an *object* (Figure) moving from the Source to the Goal (e.g. [3], [5], [6]). Tracking the changes that objects undergo is fundamental to event comprehension ([2]). We investigate how grammatical properties of transfer-of-possession sentences (grammatical aspect, verb semantics) influence comprehenders' mental representations of object *location* changes during real-time processing. We test two main questions:

(a) How do grammatical factors influence sentence-final object location representations? We consider two non-mutually-exclusive hypotheses: The *Grammatical Aspect Hypothesis* predicts that whether the event is described as ongoing (imperfective aspect: *Liam was throwing …*) or completed (perfective aspect: *Liam threw …*) influences people's representation of object location.

The *Verb Semantics Hypothesis* predicts that whether the transfer-of-possession verb entails (semantically guarantees) successful transfer influences representation of object location. *Give*-type verbs (e.g. *give*, *hand*, *bring*) entail successful transfer, but *throw*-type verbs (e.g. *throw, kick*, *toss*) do not (e.g. [4]).

(b) Are mental representation of object locations dynamically updated in real-time? Building on prior work (e.g. [1]), we hypothesize that listeners use grammatical cues to update object location representations as the sentence unfolds.

Experiment We used visual-world eye-tracking to investigate effects of grammatical aspect and verb semantics (Table 1) on representations of object locations. Eyegaze data was collected using webcam-based *Webgazer.js* ([7]) and PCIbex ([9]). Participants heard transfer-of-possession sentences (e.g. Table 1) about "the ball" and saw scenes with Source and Goal characters (Fig. 1). The ball was never depicted. Instead, participants were asked to imagine that 'we freeze the world' at the moment described by the sentence they heard, and then to mouse-click on where they think the ball is.

Click data (final interpretations) (N=65) There are more SOURCE region clicks in imperfective than perfective aspect ($z=6.33$, $p<0.0001$; *glmer*); more GOAL region clicks in perfective than imperfective aspect ($z=-6.94$, $p<0.0001$), supporting the *Grammatical Aspect Hypothesis*. To test the *Verb Semantics Hypothesis*, we compare clicks on the Center area (and the MIDDLE region) with *give*- vs. *throw*-type verbs, and we indeed find more MIDDLE clicks with non-guaranteed-transfer *throw*-type verbs (Fig. 3, $z=-8.22$, $p<0.00001$).

Eyegaze data (N=56) Looks to the SOURCE/GOAL regions were analyzed from the beginning of the main verb to the end of the sentence, offset by 400ms (to account for a systematic delay in Webgazer recordings, e.g. [8]). Proportions of SOURCE looks were higher in imperfective than in perfective aspect (Fig. 4; $t=2.29$, $p=0.026$, *lmer*). Proportion of GOAL looks were higher in perfective than in imperfective aspect (Fig. 4; $t=-15.85$, $p < 0.0001$). Goal-advantage scores (=GOAL minus SOURCE looks) revealed the same pattern (main effect of grammatical aspect; $t=-2.21$, $p=0.032$). These results suggest that grammatical aspect drives the real-time updating of object location representations.

Discussion Our results suggest that both *grammatical aspect* (as shown by gaze data and click data) and *verb semantics* (as shown by the click data) guide the process of constructing event representations. Eye-gaze patterns show that participants use grammatical aspect to dynamically update the object location representations as the sentence unfolds. That is, the process of mapping language onto mental event representations is a dynamic, real-time process. This finding is in line with prior work by [1]. Our study further shows that a temporal-semantic grammatical cue such as grammatical aspect is a relevant cue during this dynamic process. The study also suggests that the novel webcam-based eye-tracking method can provide informative data for psycholinguistic research.

Table 1: Design (For completeness, both dative argument realizations were tested (V *the ball to GOAL* & V *GOAL the ball*). Because they did not pattern differently, we collapse them.

Grammatical aspect	*give*-type verb	*throw*-type verb
imperfective	Liam was giving Paige the ball.	Carly was throwing Oliver the ball.
perfective	Liam gave Paige the ball.	Carly threw Oliver the ball.

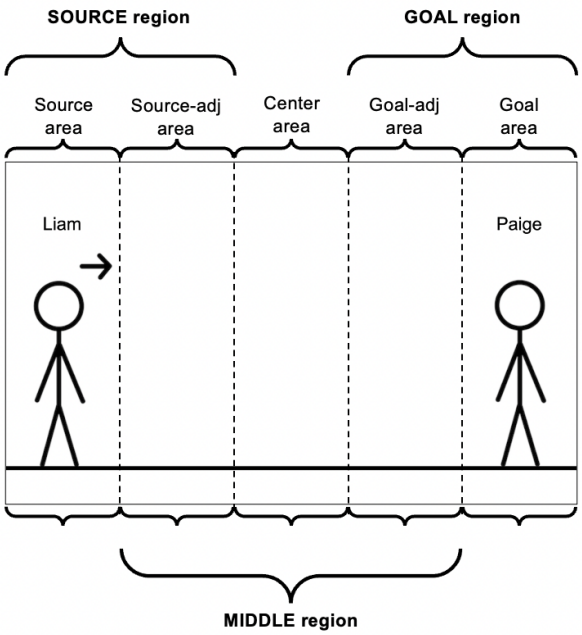


Figure 1. Visual scene

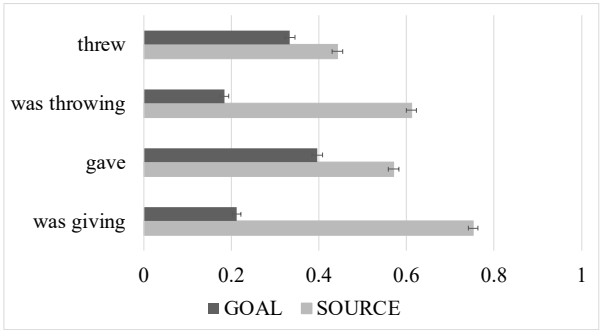


Figure 2. SOURCE vs. GOAL region clicks

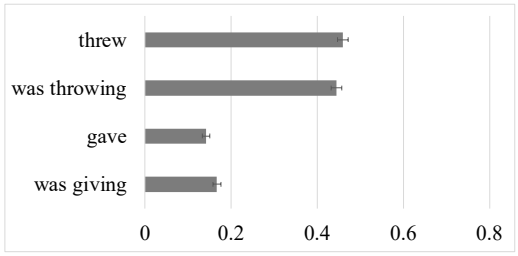


Figure 3. MIDDLE region clicks

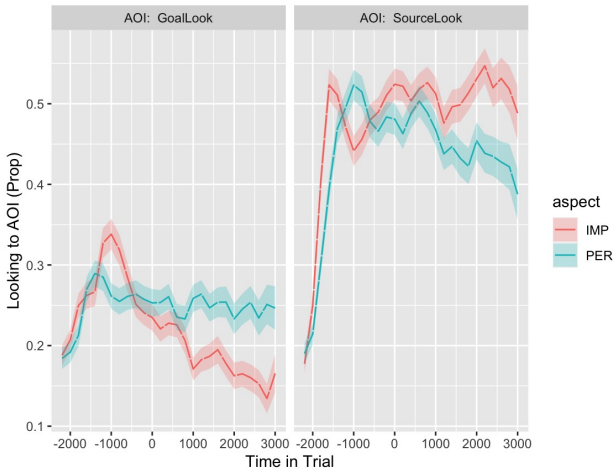


Figure 4. Proportions of GOAL region looks by grammatical aspect (left), Proportions of SOURCE region looks by grammatical aspect (right); Center area looks not plotted; 0 on the x-axis indicates the onset of the verb; Data is collapsed by participant for plotting

References

[1] Altmann & Kamide. 2009. *Cognition.*
[2] Altmann & Ekves. 2019. *Psychological Review*.
[3] Jackendoff. 1983. *Semantics and cognition.*
[4] Rappaport Hovav & Levin. 2008. *Journal of Linguistics.*
[5] Talmy. 1983. How language structures space. In *Spatial orientations: Theory, research, and application.*
[6] Talmy. 1985. Lexicalization patterns: Semantic structure in lexical forms. In *Language typology and syntactic description.*
[7] Papoutsaki et al., 2016. *Proceedings of the International Joint Conference on Artificial Intelligence.*
[8] Slim & Hartsuiker. 2021. Visual world eyetracking using WebGazer.js.
[9] Zehr & Schwarz. 2018. PennController for Internet Based Experiments (IBEX).

Acquisition of subordinate nouns as pragmatic inference: Semantic alternatives modulate subordinate meanings

June Choe[1], Anna Papafragou[1]
[1]University of Pennsylvania

A major aspect of word learning involves identifying the level of specificity encoded by word meanings (Quine, 1960). Evidence suggests that learners show a bias for mapping words to basic-level (*dog*), as opposed to subordinate-level meanings (e.g., *poodle*; Markman, 1990; Waxman & Markow, 1995; Waxman et al., 1991, 1997), but the circumstances that allow learners to generalize word meanings beyond the basic level are still under debate (Xu & Tanenbaum, 2007; Spencer et al., 2011; Lewis & Frank, 2018; Wang & Trueswell, 2019).

Here, we begin with the assumption that learners make pragmatically-driven inferences about the hypothesis space over which possible word meanings are proposed and evaluated. Unlike past accounts that framed the acquisition of subordinate-level nouns as a question of how various sources of perceptual information in the referential world interact and converge on a specific concept (e.g., Xu & Tanenbaum, 2007), we ask under what discourse contexts learners expect to hear a word with a narrower meaning. In the case of basic vs. subordinate meanings, identifying the intended meaning involves selecting the appropriate level of informativeness for a novel word, with subordinate meanings being more informative. In two online experiments, we probed the nature of these pragmatic inferences by testing the role of contrast in adult learners' basic- vs. subordinate-level generalization of novel words from single trials. We hypothesized that the rate of basic-level generalizations for an ostensive target label (e.g., 'mipen' paired with a red apple) would decrease if the target is followed by a semantic alternative at the subordinate level (e.g., a green apple), under the assumption that the presence of the alternative makes it clear that the more informative (subordinate-level) categories are relevant to the task (Barner, Brooks & Bale, 2011; Skordos & Papafragou, 2016). Additionally, we hypothesized that this effect of contrast would be linguistic, as opposed to conceptual, and should thus be stronger when the alternative was labelled rather than simply present and unlabelled. In two experiments, we tested these hypotheses respectively.

In Experiment 1 (n=50), a foreign-language speaker named Sally told participants that they would be learning words from her native language. There were 10 trials (4 with natural kinds, 4 with artifacts, and 2 catch trials), each divided into two parts. In the learning phase (**Fig. 1**), Sally labelled the target with a novel word. In the *contrast* condition, she then introduced an alternative at the same subordinate level with a different label. In the *no-contrast* condition, no such contrast was introduced. In the test phase, participants were asked to select all matches for the target label from an array of images (two subordinate-level matches to each of the target and the contrast, three unseen basic-level exemplars, three superordinate exemplars, and eight unrelated exemplars; **Fig. 2**). Basic-level responses contained all matches to the target and the contrast and any number of other basic-level exemplars. Consistent with our predictions, we found a significant negative effect of contrast (p <.0001) from a logistic regression model fitted to basic-level responses to the target label at test (**Fig. 3A**).

Experiment 2 (n=90) was similar but sought to disentangle the effect of labelling from the mere presence of the alternative referent. We manipulated the labelling of the semantic alternative (*labelled* vs. *un-labelled*) and, to guard against possible presentation effects, we counterbalanced the order in which the target appeared in the learning phase relative to the alternative (first vs. second). A logistic mixed-effects model fitted to basic-level responses revealed a significant main effect of labelling (p <.0001), indicating a strong shift to subordinate-level generalizations for the target label when the alternative was labelled **(Fig. 3B).**

In sum, semantic alternatives facilitate mappings to subordinate-level meanings, and especially so when the alternative is labelled. This suggests that learners can use linguistically marked contrast to reason about the level of specificity for a word's meaning expected from an ostensive labelling event.

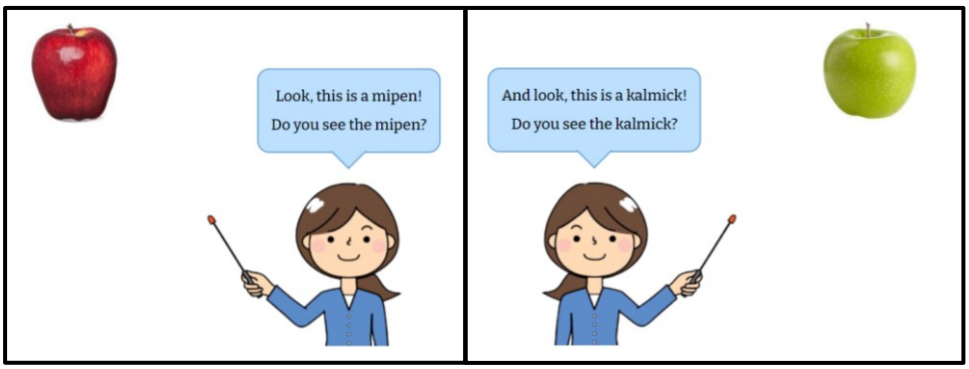


Figure 1. The learning phase. In Experiment 1, the target (A) or the target and the contrast (A and B) appeared to the left then right of Sally, for seven seconds each with one second in between. In Experiment 2, the speech bubble for the *un-labelled* condition read: "(And) look here! Do you see this?"

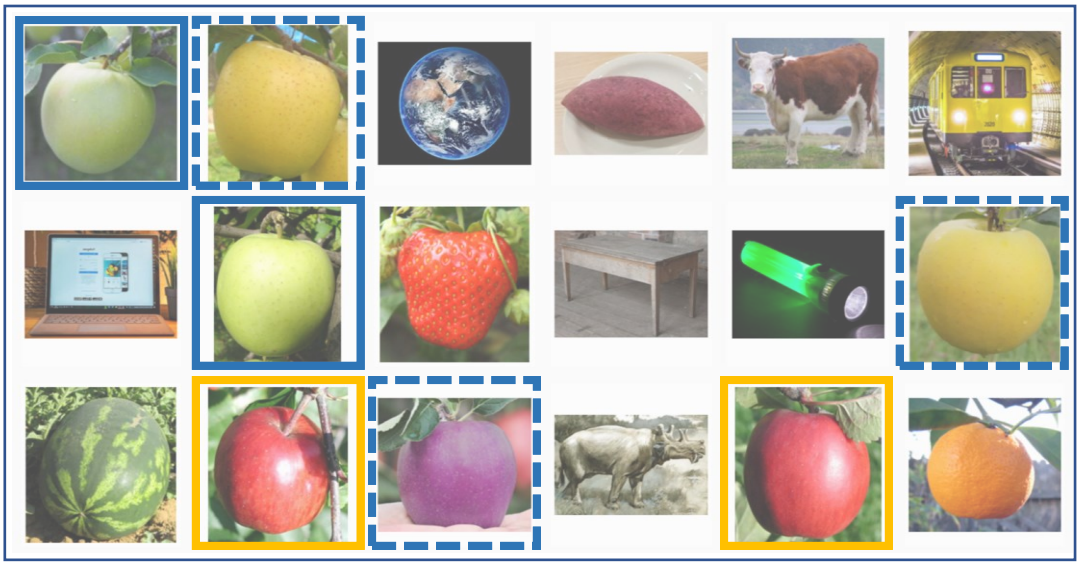


Figure 2. The testing phase. Sally reappeared to give instructions "Do you see any other mipens below? Click on all the mipens you see!" Choices were coded as 'subordinate (orange), 'basic' (orange + blue), and 'other' for all other responses. When the learning phase introduced a contrast label, responses including the alternative subordinate-level images (solid blue) were coded as 'other'.

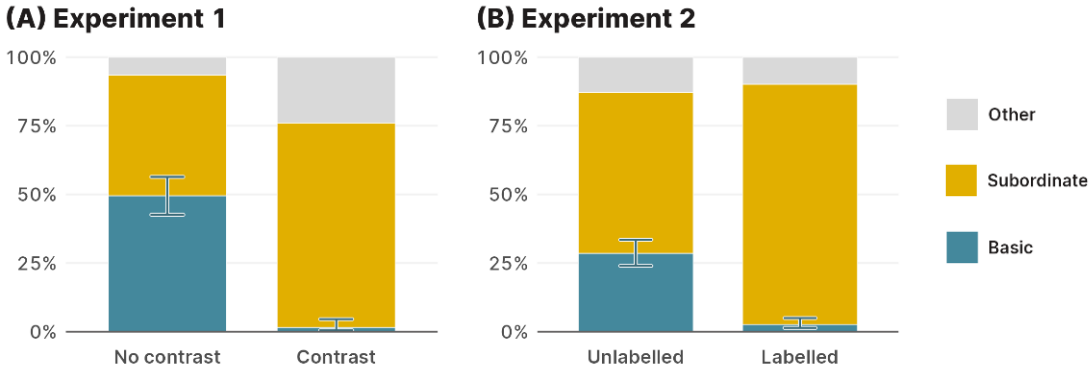


Figure 3. Coded responses to the target label at test in Experiment 1 (A) and Experiment 2 (B)

References: [1] Quine, 1960. MIT Press. [2] Markman, 1990. *Cog. Sci*. [3] Waxman & Markow, 1995. *Cog. Sci.* [4] Waxman, 2003. Oxford UP. [5] Waxman et al., 1991. *Child Dev.* [6] Waxman et al., 1997. *Dev. Psy*. [7] Xu & Tanenbaum, 2007. *Psy. Rev.* [8] Spencer et al., 2011. *Psy. Sci*. [9] Jenkins et al., 2015. *Cog. Sci.* [10] Lewis & Frank, 2018. *Psy. Sci*. [11] Wang & Trueswell, 2019. *Cog. Sci.* [11] Barner, Brooks & Bale, 2011. *Cognition*. [12] Skordos & Papafragou, 2016. *Cognition*.

Are they touching? Contact and pronoun choice in English prepositional phrases

Shannon Bryant, Harvard University

Introduction. In English, both reflexive pronouns (*herself*) and personal pronouns (*her*) can be used in locative prepositional phrases (LPPs) to refer back to the sentence subject [1-3]. Prior theoretical work has proposed that the choice between forms depends on the nature of the spatial relation expressed by the preposition, in particular whether the relation is one of direct physical contact. According to [4-5], reflexives are more acceptable when physical contact holds, while personal pronouns are more acceptable in the absence of physical contact:

(1) a. Corporal Crump pinned the medal beside him/*himself (on the wall).
b. Corporal Crump pinned the medal onto *him/himself. ([5]:15)

Similarly, [6-8] report that use of the reflexive gives rise to an inference of physical contact, whereas the personal pronoun is neutral in this respect:

(2) a. When he woke up, John found a rope around himself.
He had been tied up. / *It described a neat circle 4 meters in diameter.
b. When he woke up, John found a rope around him.
He had been tied up. / It described a neat circle 4 meters in diameter. ([8]:54)

In this study, we experimentally tested the relationship between physical contact and pronoun choice in English LPPs, looking at both the impact of contact on pronoun acceptability (Exp. 1) and the impact of pronoun choice on the likelihood of inferring contact (Exp. 2). Our results support the proposal that reflexives are favored in contexts in which contact holds, though they point to a flexible association between reflexives and contact rather than a fixed semantic requirement.

Exp. 1: Acceptability rating survey. To test the effect of physical contact on reflexive and personal pronoun acceptability, we created 18 sets of target sentences by varying pronoun type and relation type (∓ CONTACT) across three types of embedding verb (HAVE, PERCEPTION, MOTION):

	+CONTACT	-CONTACT
HAVE	Chloe had some glitter on her(self).	Chloe had some glitter next to her(self).
PERC.	Chloe noticed some glitter on her(self).	Chloe noticed some glitter next to her(self).
MOTION	Chloe poured some glitter on her(self).	Chloe poured some glitter next to her(self).

Sentences were paired with short supporting contexts, each naming two people, the second of whom served as the subject of the target sentence; stereotypically gendered names were used to constrain pronoun interpretation. Following [9], we presented minimal sentence pairs side-by-side with slider bars to help draw out relative preferences between the reflexive and personal pronoun in a given construction (**Fig 1a**). Ratings were collected online from 122 participants, each of whom saw 18 target questions (3 per condition) as well as 4 control and 22 filler questions.

Responses from 31 participants were excluded from analysis due to failure on catch trials, leaving an average of 270 observations per condition. Consistent with the observations in [4-5], reflexives received higher ratings on average in +CONTACT sentences than in -CONTACT sentences across all three verb types, while personal pronouns showed the opposite pattern (**Fig. 1b**). Results from a linear mixed effects analysis revealed relation type to be a significant predictor of both reflexive acceptability (β=0.507, $p<0.001$) and personal pronoun acceptability (β=-0.389, $p<0.001$).

Exp. 2: Contact inference survey. Exp. 1 stimuli were designed to bias participants towards either a +CONTACT or a -CONTACT reading. A subsequent norming study confirmed the overall efficacy of this manipulation: +CONTACT stimuli were overwhelmingly interpreted as involving contact, -CONTACT stimuli as not involving contact. However, norming results also revealed some variation in interpretation, particularly for MOTION/-CONTACT stimuli. This opened up the possibility that pronoun choice could influence whether or not contact is inferred for these sentences, in line with [6-8]. To test this, we included 10 of the MOTION/-CONTACT sentence pairs in a binary choice

inference survey. Sentences were presented one at a time followed by a Yes/No question of the form *Was the X touching Y?* (**Fig. 2a**). Inferences were collected online from 30 participants, each of whom saw 10 target questions (5 reflexive, 5 personal pronoun) and 5 filler questions.

Though 'No' responses were considerably more frequent than 'Yes' responses regardless of pronoun type, sentences containing reflexives gave rise to contact inferences more often than did sentences containing personal pronouns (**Fig. 2b**). A logistic mixed effects analysis showed the effect of pronoun type on the likelihood of inferring contact to be significant (β=-0.678, p=0.02).

Discussion. Experiments 1 and 2 lend empirical weight to the purported relevance of contact to pronoun choice in English LPPs. However, contrary to the strongest claims in the literature, contact was not found to impose strict complementarity between reflexives and pronouns, nor did reflexives uniformly prompt inference of contact, ruling against an analysis that builds contact into the denotation of the reflexive. Instead, we suggest that these findings reflect association of the reflexive with two features of *event structure*—spatial contiguity and affectedness— following from its canonical use in transitive constructions.

Figure 1a: Question from Experiment 1

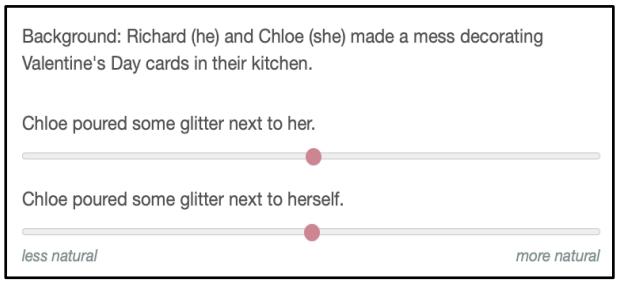


Figure 2a: Question from Experiment 2

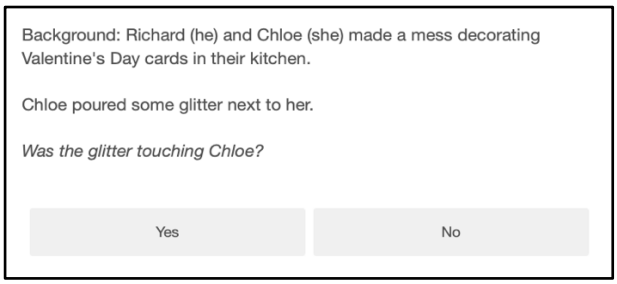


Figure 1b: Z-scored ratings from Exp. 1

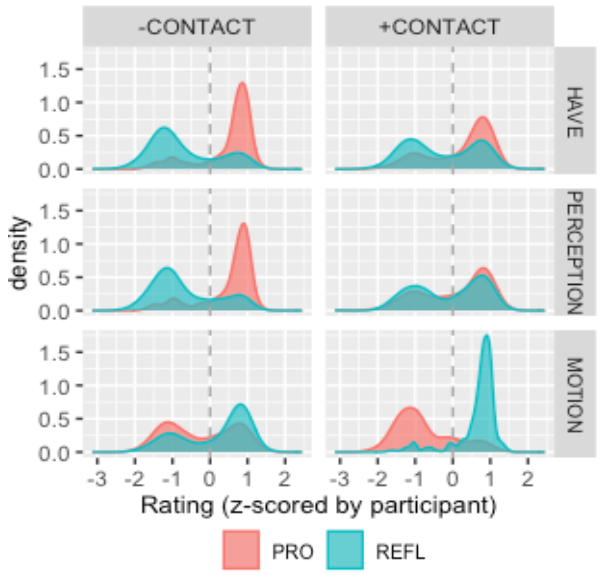
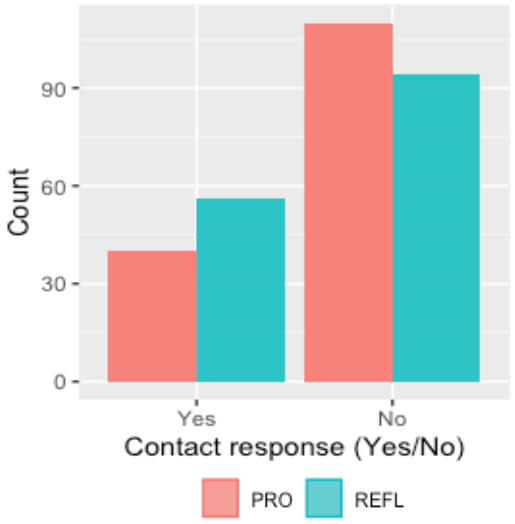


Figure 2b: Contact inferences from Exp. 2



References. [1] Chomsky 1981. *Lectures on Government and Binding*. [2] Reinhart & Reuland 1993. Reflexivity. [3] Büring 2005. *Binding theory.* [4] Faltz 1985. *Reflexivization: A Study in Universal Syntax*. [5] Wechsler 1997. Prepositional phrases from the twilight zone. [6] Kuno 1987. *Functional Syntax: Anaphora, Discourse, and Empathy.* [7] Van Hoek 1997. *Anaphora and Conceptual Structure*. [8] Rooryck & Vanden Wyngaerd 2007. The syntax of spatial anaphora. [9] Marty, Chemla & Sprouse 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks.

Trouble finding the words: Lexical differences affect how English and Chinese speakers communicate categories

Lilia Rissman, Qiawen Liu and Gary Lupyan

Background. Languages vary substantially in how they lexicalize the same concepts — for example, some languages have distinct lexical items for “niece” and “nephew” but others do not (Wallace & Atkins, 1960). We investigate the impact of such cross-linguistic differences on communication — whether having a conventional term for a concept facilitates communication of that concept. We focus on superordinates such as *beverages* and *vehicles* — nouns that convey broad categories of individuals. Languages have different inventories of superordinates (Kemmerer, 2019; Mihatsch, 2007). For example, ||Gana divides living things not into categories of *plant* and *animal* but into categories such as *kx’ooxo* (‘living things which are edible’) (Harrison, 2007). We ask whether the availability of a superordinate term leads speakers to communicate more effectively about that category than if a superordinate term were absent.

As humans, the range of ideas we want to convey is far wider than the discrete set of morphemes present in any one language. The expressive capacity of language comes largely through combinatoriality – composing morphemes into larger units and phrases that convey complex thoughts. Given this expressive capacity, it might be that communication is not hampered by the absence of a superordinate term. For example, although English has no superordinate that is translationally equivalent to ||Gana *kx’ooxo,* we can convey its meaning through the complex, ad hoc category description “living things which are edible” (see Barsalou, 1983 on ad hoc categories). From this perspective, languages are fully intertranslatable with each other — as articulated by Harnad (1996), “whether it does so analytically, synthetically, or even entirely holophrastically, a language must provide the resources for marking distinctly all the categories we distinguish.” An alternative to this perspective is that an ad hoc description is more limited than a superordinate in its ability to convey a category, because the ad hoc description only approximates the meaning of the superordinate, or because speakers differ in their ability to construct ad hoc descriptions on the fly, or because speakers and receivers interpret ad hoc descriptions in different ways. We tested these two alternatives by comparing how English and Chinese speakers communicate about categories for which there is a superordinate in one language but not the other. These languages have been shown to lexicalize semantic space in strikingly different ways (e.g., Saji et al., 2011). At the same time, as members of industrialized societies in increasing contact, speakers of these languages share a range of category knowledge about artifacts, foods, and the natural world.

Method. Participants completed a referential matching game (a ‘Director-Matcher’ task). 77 American English speakers and 80 Chinese speakers played the Director role. Directors viewed a 3 x 3 grid with a noun in each cell. Three of these nouns (e.g., *beer, soda, juice*) were highlighted. Directors were instructed to write a clue that would enable another person to choose the highlighted and only the highlighted words from the same grid. 86 American English speakers and 124 Chinese speakers played the Matcher role. Matchers viewed the Director’s clue along with the 3 x 3 grid (without highlighting) and selected the nouns corresponding to the clue. Matcher’s and Director’s grids had the same words, but in a different order.

We selected the words in the grid based on 10 English and 10 Chinese superordinate terms shown in Table 1. For each term, there is no direct translational equivalent in the other language. Each trial corresponded to a superordinate term and the nouns in the 3 x 3 grid for each trial were of the following types: 1) three Targets, which were typical members of the category denoted by the superordinate (e.g., *beer, soda, juice* for the term *beverages*), 2) two Lure Distractors, which were semantically similar to the Targets but were not members of the superordinate category (e.g., *vinegar, oil*), and 3) four Non-Lure Distractors, which were

semantically dissimilar to the Targets and Lures (e.g., *motorcycle, star, tree, sleet*). We constructed six grids for each superordinate term. The grids based on the English and the Chinese superordinates were translated into Chinese and English, respectively. We selected nouns that would be familiar to both English and Chinese speakers (e.g., *beer*; *píjiǔ* 'beer'). Directors and Matchers saw two grids per term, resulting in 40 trials per participant. Each Matcher was yoked to a single Director and saw all clues produced by that Director. All studies were conducted online.

Table 1. English and Chinese superordinates

English terms	Chinese terms (English gloss)
appetizers	nóngchǎnpǐn (agricultural products and livestock)
beverages	huàzhuāngpǐn (cosmetics and facial products)
crafts	dìxíng (terrain and water features)
crimes	jiājù (furniture and home décor)
drugs	fúshì (apparel, shoes, and jewelry)
pests	shuǐyù (bodies of water)
precipitation	shēngwù (living things)
skills	diànqì (electrical appliances and devices)
snacks	tiáowèi pǐn (food seasonings)
vehicles	fēngjǐng (scenic places to visit)

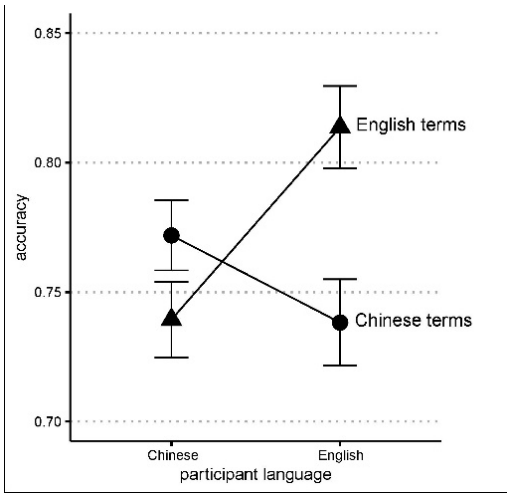


Figure 1. Mean Matcher accuracy across Language and Condition. Error bars show 95% confidence intervals.

Results and Discussion. We calculated Matcher accuracy, defined as the Hit rate per trial (correctly choosing a Target) minus the False Alarm rate per trial (incorrectly choosing a Distractor). Mean accuracy for each condition is shown in Figure 1. We modeled accuracy using linear mixed-effects regression and the *lme4* package for R (Bates, Maechler, Bolker, & Walker, 2014). Our model included Subject and Term random intercepts and Subject-by-Condition random slopes. With reaction time as a covariate, the main effects of Language and Condition were non-significant ($b = -.05$, $SE = .06$, $p < .1$; $b = -.11$, $SE = .11$, $p > .1$). That is, English speakers were not more accurate than Chinese speakers (or vice versa) and categories derived from English terms were not more difficult than categories derived from Chinese terms (or vice versa). We did, however, observe a significant interaction between Language and Condition: English speakers were more accurate for categories derived from English terms ($b = .36$, $SE = .041$, $p < .001$). This demonstrates that English speakers were more effective at conveying categories when there was an English superordinate term available, *mutatis mutandis* for Chinese speakers. These results dispute the view that languages are all mutually intertranslatable. Instead, when language users have to create ad hoc, non-conventional category descriptions, the descriptions they create appear to be less effective than conventional superordinates for conveying the same categories.

Barsalou (1983). Ad hoc categories. *Memory & cognition* | Bates et al. (2014). lme4: Linear mixed-effects models using S4 classes. | Harnad (1996). The origin of words: A psychophysical hypothesis. *Communicating meaning: evolution and development of language* | Harrison (2007). *When languages die: the extinction of the world's languages and the erosion of human knowledge*. | Kemmerer (2019). *Concepts in the brain: The View from cross-linguistic diversity*. | Mihatsch (2007). Taxonomic and meronomic superordinates with nominal coding. *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts.* | Saji et al.

(2011). Word learning does not end at fast-mapping… *Cognition*. | Wallace & Atkins (1960). The meaning of kinship terms. *American Anthropologist*.

The color/size asymmetry in redundant modification replicates cross-linguistically

Speakers produce redundant color adjectives more frequently than redundant size adjectives [1-3]. For example, in contexts like Fig. 1, where size is sufficient for establishing the target referent, speakers frequently produce color redundantly ("the small green apple" instead of "the small apple"). Whether this asymmetry is the result of an asymmetry in the referential utility of mentioning lexicalized concepts -- e.g., because color is more perceptually salient than size and thus likely to increase the probability of communicative success [3-5] -- or the result of incremental language processing pressures [6,7] is an open question. Cross-linguistic studies of redundant modification are important to this debate: similar cross-linguistic rates of redundant modification across languages that differ in relevant syntactic properties would implicate lexicalized concepts as the source of redundant modification. In contrast, lower prevalence of redundant modification in languages with post-nominal modification implicates a strong role for incrementality. Thus far, studies addressing redundancy in referring expressions have mostly been conducted on a handful of pre-nominal modification languages (e.g., English [1-4], German [8] and Dutch [9]). Notable exceptions include [10, 11], who observed less redundant color modification in Spanish, a post-nominal modification language, than in English, providing initial evidence for the role of incrementality. However, these studies were conducted on a set of contexts in which only redundant color but not redundant size modification was investigated.

Aim. We ask whether the propensity for redundant modification and in particular the color/size asymmetry replicate in two particularly interesting languages, which we compare to English: Spanish and Central Taurus Sign Language (CTSL, see Tab. 1 for details). As a language in its infancy, CTSL provides us with the unique opportunity to test the extent to which, with no established conventions, redundant modification patterns mirror those previously documented.

Methods. Participants (see Tab. 1) played an interactive reference game (see Fig. 1). On each trial, participants saw a display of objects. The director was asked to communicate the target object marked by a green border in their display to the addressee, who selected an object. On half of trials, color was sufficient for unique reference, and on the other half, size was sufficient. Participants were recorded during the task and their responses were transcribed and translated to English for analysis. Productions of both color and size were coded as redundant.

Results. Modification in CTSL and Spanish was overwhelmingly postnominal (~90%). Both CTSL signers and Spanish speakers were more likely to redundantly mention color than size **(**β=4.95**,** 95% CI = [3.73, 6.32], see Fig. 2**)** at rates similar to those previously reported [3]. Compared to English, there was no evidence that rates of redundant modification differed in Spanish (β=-0.02, CI = [-1.75, 1.61) or in CTSL (β=1.21, CI = [-0.13, 2.56]). These null results may be due to low power for Spanish and CTSL; data collection for Spanish is still ongoing.

Discussion. In neither predominantly post-nominal adjective language was redundancy lower than in the pre-nominal English baseline. These results are at odds with those of [10, 11] and with explanations of redundancy that ascribe a large explanatory role to incremental pressures. Instead, the results preliminarily suggest that the underlying systematicity in redundant adjectival modification is due to lexicalized concepts that differ fundamentally in referential utility. With the aim of rigorously comparing such explanations, in ongoing work we evaluate quantitative computational theories of referring expression production that differ in whether speakers plan utterances in anticipation of incremental pressures, and/or whether adjectival modifiers receive a Boolean intersective semantic analysis or a noisy, continuous one [3,7].

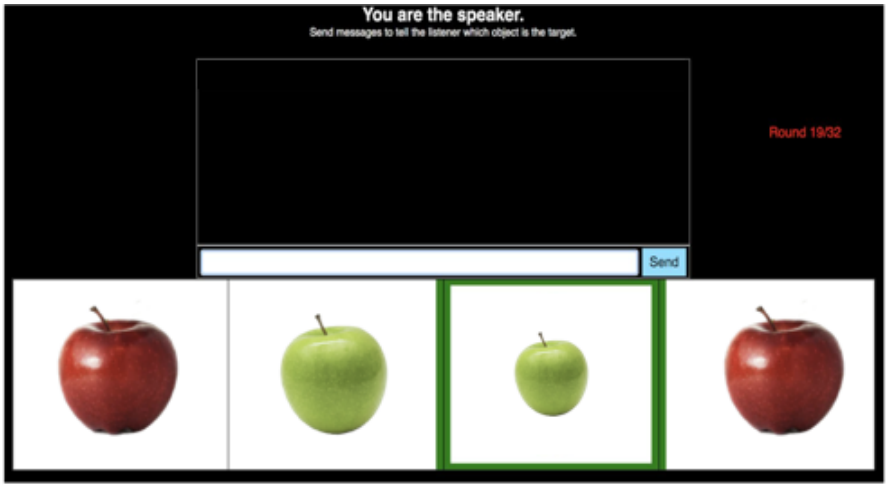


Fig. 1: Example display from the director's perspective on a size sufficient trial.

Language	English	Spanish	CTSL
Syntactic features	Pre-nominal adjectives	Post-nominal adjectives	No established ordering conventions
Participants	60 dyads (re-analysis of data reported by [3])	9 dyads (data collection still in progress)	11 dyads
Modality	written	written	signed
Examples	*small green apple*	*manzana verde pequeña la pequeña, verde*	APPLE GREEN SMALL SMALL APPLE GREEN

Tab. 1: Tested languages and their features. CTSL is an emerging village sign language that arose naturally within the last half century in a small isolated community in Southern Turkey as a result of high incidence of hereditary deafness, and in the absence of a conventionalized language model. Both Spanish and CTSL allow for post-nominal modification and split pre- and post-nominal modification, shown in table. CTSL also allows purely pre-nominal modification.

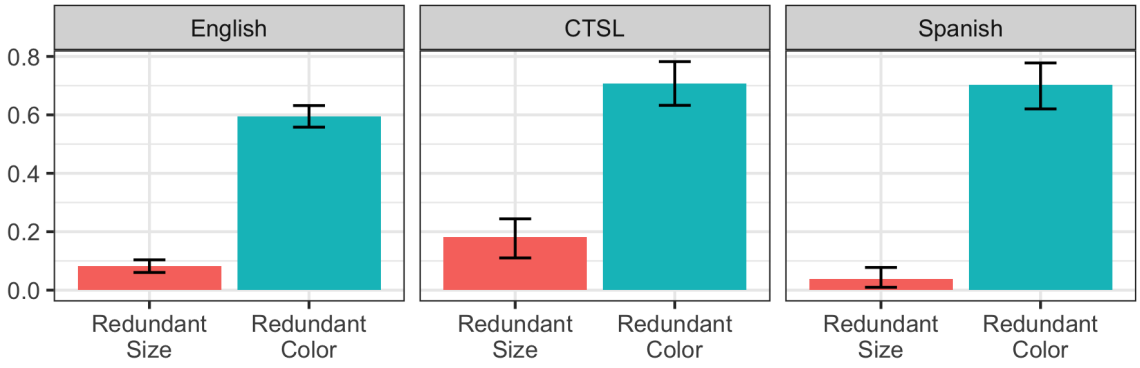


Fig. 2: Proportion of redundant "color and size" mentions by condition and language.

References [1] Pechman (1989) [2] Sedivy (2003) [3] Degen et al. (2020) [4] Kursat & Degen (2020) [5] Rubio-Fernandez et al. (2020) [6] Rubio-Fernandez & Ettinger (2020) [7] Waldon & Degen (2021) [8] Belke (2001) [9] Koolen et al. (2013) [10] Rubio-Fernandez (2016) [11] Wu & Gibson (2020)

Modals in natural language optimize the simplicity/informativeness trade-off

Nathaniel Imel and Shane Steinert-Threlkeld

Introduction A language can be simple and uninformative (e.g. containing a single expression). A language can be complex and informative (e.g. containing unique expressions for each possible meaning). A language cannot be both simple and informative: these two pressures trade-off against each other. A recent line of work develops the idea that meanings cross-linguistically are optimized for efficient communication, i.e. they optimally balance these two competing pressures [1]. This approach successfully explains the semantic variation observed in domains both of content words (e.g. kinship [2], color [3]) and function words (e.g. quantifiers [4], indefinites [5], boolean connectives [6]). This paper shows that modals cross-linguistically [7, 8] can be seen as optimizing this trade-off.

Measures In modeling (efficient) communication with modals, we take the object of communication to be the correct transmission of a pair of a *force* and a *flavor*. At this level of modeling, the meaning of a modal is a set of such pairs, allowing us to capture variability in flavor (e.g. for English *may*) as well as variability in force, as recently argued to be present in Lilloet Salish [9], Nez Perce [10], Washo [11], and Old English [12].

We measure the complexity of a modal in terms of the shortest formula in a language of thought [13]. In particular, we use a basic propositional language with atoms for each possible force and each possible flavor. For a modal, we write a disjunctive normal form capturing all of the force-flavor pairs it can express, and then apply a minimization algorithm based on [14]. The complexity is the number of atoms in this shortest formula; the complexity of a language is the sum of the complexity of the modals therein.

We measure the informativeness of a modal system (following [4, 5]) in terms of the probability of successful communication between a speaker who wants to convey an intended force-flavor pair to a listener, who must guess which one is intended solely on the basis of hearing a modal expression from the speaker. More formally:

$$I(L) := \sum_{\mathbb{M}} P(\mathbb{M}) \sum_{m \in L} P(m|\mathbb{M}) \sum_{\mathbb{M}' \in m} P(\mathbb{M}'|m) \cdot u(\mathbb{M}', \mathbb{M})$$

$$\text{where } u(\mathbb{M}', \mathbb{M}) = 0.5 \cdot \mathbb{1}_{\text{force}(\mathbb{M})=\text{force}(\mathbb{M}')} + 0.5 \cdot \mathbb{1}_{\text{flavor}(\mathbb{M})=\text{flavor}(\mathbb{M}')}$$

Here, $P(\mathbb{M})$ is a prior probability (assumed to be uniform) over the pairs; $P(m|\mathbb{M})$ represents the speaker, where m is a modal, and $P(\mathbb{M}'|m)$ the listener. The utility function u gives partial credit: the listener gets half credit for correctly guessing each of the force and the flavor, and so full credit for correctly guessing the intended pair. Finally, *communicative cost* is inversely related to informativeness: $C(L) := 1 - I(L)$.

In the absence of a robust dataset of the modal systems of many languages, we proceed by generating a large number of artificial languages and using proposed semantic universals to measure how natural such languages are. In particular, Nauze [15] proposed what we may call the *Single Ambiguity Universal (SAU)*: a modal may be ambiguous in either force or flavor, but not both. For a given language, we measure its Nauze degree as the proportion of modals that satisfy the SAU. As a refinement, Vander Klok [16] suggested that within both the epistemic / root domains, the system as a whole may only exhibit one kind of ambiguity. See [8] for discussion. We record for each langauge whether or not it satisfies Vander Klok's refinement.

Results Figure 1 presents the main results. We experiment with a meaning space containing 2 forces and 3 flavors. Each point is a language; the x-axis is communicative cost,

and the y-axis is complexity. The black line is the Pareto frontier: the set of languages for which no other language is both simpler and more informative. Triangles are Vander Klok languages. The color of a language is its Nauze degree.

We catalog several particular results. All optimal languages (those on the frontier) satisfy Vander Klok's generalization, with the exception of a single language on the bottom-right, which corresponds to a language with a single, highly-ambiguous modal (à la the Washo verb -*e*ʔ [11]). In particular, the Vander Klok languages ($N = 2255$) have mean optimality of 0.957 compared to a mean optimality of 0.797 for the remaining languages ($N = 65023$). More generally: Nauze degree is highly correlated with optimality (Pearson $r = 0.55$). This shows that languages which have more modals satisfying Nauze's SAU tend to do better at optimizing the simplicity/informativeness trade-off.

Discussion To summarize: our experiments show (i) that modal systems optimized for efficient communication satisfy Vander Klok's generalization and (ii) that languages with more Nauze modals tend to be more efficient for communication. These results show that trading off very general pressures for simplicity and informativeness may shape the semantic variation in the modal systems of the world's languages.

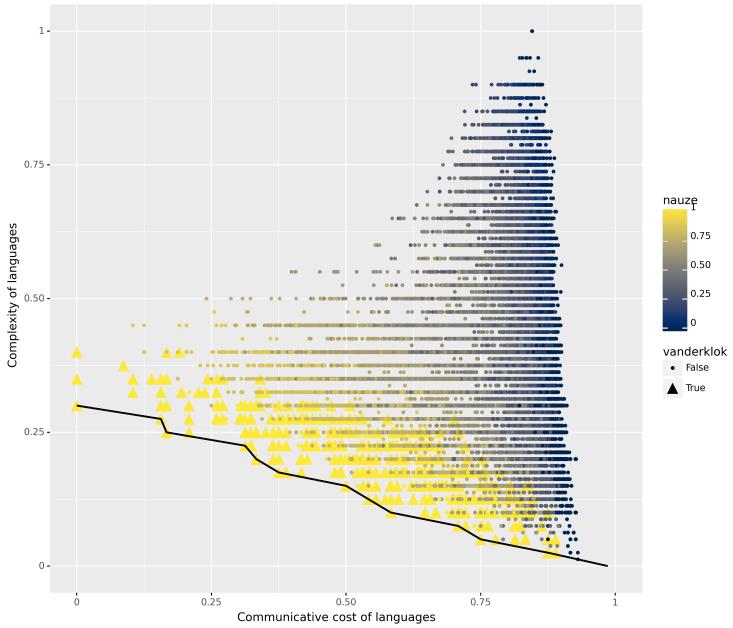


Figure 1: The modal systems sampled, plotted with communicative cost on the x-axis and complexity on the y-axis. Black: the Pareto frontier of optimal languages. Triangles satisfy Vander Klok's generalization. Color corresponds to Nauze degree.

1. Kemp, C. *et al.* Semantic Typology and Efficient Communication. *Annual Review of Linguistics,* 1–23 (2018).
2. Kemp, C. *et al.* Kinship Categories across Languages Reflect General Communicative Principles. *Science* **336,** 1049–1054 (2012).
3. Zaslavsky, N. *et al.* Efficient Compression in Color Naming and Its Evolution. *Proceedings of the National Academy of Sciences* **115,** 7937–7942 (2018).
4. Steinert-Threlkeld, S. Quantifiers in Natural Language: Efficient Communication and Degrees of Semantic Universals. *Entropy* **23,** 1335 (2021).
5. Denić, M. *et al. Complexity/Informativeness Trade-off in the Domain of Indefinite Pronouns* in *Proceedings of Semantics and Linguistic Theory (SALT 30)* **30** (2020), 166–184.
6. Uegaki, W. The Informativeness / Complexity Trade-off in the Domain of Boolean Connectives. *Linguistic Inquiry* (2021).
7. Kratzer, A. *The Notional Category of Modality* in *Words, Worlds, and Context* (eds Eikmeyer, H.-J. *et al.*) 38–74 (Walter de Gruyter, 1981).
8. Matthewson, L. *Modality* in *The Cambridge Handbook of Formal Semantics* (eds Aloni, M. *et al.*) 525–559 (Cambridge University Press, Cambridge, 2019).
9. Rullmann, H. *et al.* Modals as Distributive Indefinites. *Natural Language Semantics* **16,** 317–357 (2008).
10. Deal, A. R. Modals Without Scales. *Language* **87,** 559–585 (2011).
11. Bochnak, M. R. *Variable Force Modality in Washo* in *Proceedings of North-East Linguistic Society (NELS) 45* (eds Bui, T. *et al.*) (2015), 105–114.
12. Yanovich, I. Old English *motan, Variable-Force Modality, and the Presupposition of Inevitable Actualization. *Language* **92,** 489–521 (2016).
13. Piantadosi, S. T. *et al.* The Logical Primitives of Thought: Empirical Foundations for Compositional Cognitive Models. *Psychological Review* **123,** 392–424 (2016).
14. Feldman, J. Minimization of Boolean Complexity in Human Concept Learning. *Nature* **407,** 630–633 (2000).
15. Nauze, F. D. *Modality in Typological Perspective* (Universiteit van Amsterdam, 2008).
16. Vander Klok, J. *Restrictions on Semantic Variation: A Case Study on Modal System Types* in *Workshop on Semantic Variation* (2013).

Crosslinguistic differences on the Present Perfect Puzzle: an experimental approach

Martín Fuchs & Martijn van der Klis - Utrecht University

Introduction. The *present perfect puzzle* states that "the present perfect does not go with an adverbial referring to the past" (Klein 1992: 526), so the *Simple Past* has to be used instead:

(1) Chris **has left* / *left* York **today at six**. (adapted from Klein 1992: 546, (ex.45))

Yet other languages (French, Italian, German) do allow their corresponding PERFECT markers to combine with past referring adverbials (e.g., Squartini & Bertinetto 2000), showing that this constraint does not hold crosslinguistically. The Dutch PERFECT *Voltooid Tegenwoordige Tijd* is also not affected by it, *as* (2a) shows. Rather, as (2b) exemplifies, the PERFECT is said to be preferred over the PAST form in such contexts (van der Klis et al. 2021).

(2a) Chris *heeft* York **vandaag om zes uur** *verlaten*. 'Chris *has* York **today at six** *left*'
(2b) *Chris *verliet* York **vandaag om zes uur**. 'Chris *left* York **today at six**'

Peninsular Spanish appears to reflect an intermediate point in its availability to combine with past-referring temporal adverbials (e.g. Harris 1982):

(3) Chris se *ha ido* / #*fue* de York **hoy a las seis**. 'Chris *has left* / *left* York **today at six'**.
(4) Chris se **ha ido* / *fue* de York **ayer**. 'Chris *has left* / *left* York **yesterday'.**

As (3) indicates, Spanish is not subject to the *present perfect puzzle* as long as the temporal adverbial (*hoy a las seis* 'today at six') creates the relation E=R $\subseteq$ day(S). That is, when the event E is temporally located within the day of utterance S, the Spanish PERFECT form –the *Pretérito Perfecto Compuesto*– can be used. Conversely, when the event E is anchored to a past reference time R before the day of utterance S, as in (4), with the adverb *ayer* 'yesterday', only the (Perfective) PAST –the *Pretérito Indefinido*– is allowed. This has led some authors to define the Spanish PERFECT as a hodiernal marker (e.g., Schwenter 1994).

Other work in English has provided indications that *deictic* temporal adverbials (i.e., adverbials whose reference is calculated with respect to the speaker's time/space center of reference) behave differently with respect to their (in)compatibility with the PERFECT (e.g., Hitzeman 1995). Different from (1), the *Present Perfect* seems to be able to combine with deictic past-time referring adverbials that include the speech time S, like *this afternoon*, as (5) shows:

(5) Chris *has left* / *left* York **this afternoon**.

To our knowledge, the role of deixis in the acceptability of the Spanish and Dutch PERFECT forms has not been studied. Here we experimentally test the acceptability of different past time adverbials with the PERFECT and PAST markers of English, Spanish, and Dutch. We consider a twofold distinction of temporal adverbials. First, (3) and (4) indicate variation between adverbials related to the day of utterance and those that are not. Second, (1) and (5) drive a distinction between deictic and non-deictic adverbials. Finally, (2a) and (2b) suggest that Dutch prefers its PERFECT over the PAST across the board.

Method. We investigate English, Spanish, and Dutch use of PERFECT and PAST markers in combination with different temporal adverbials distinguished by two variables: (i) **+/-T**: In +T cases, adverbials relate to day (S) by being included in it (e.g., *this morning*) or including it (e.g., *this month*). This is a broader notion of strict hodiernality that intends to incorporate the 'extended now' (e.g, Portner 2003). Conversely, -T adverbs, such as *last month*, do not include or are included in day (S); (ii) **+/-D**: In +D adverbs, the temporal reference of the adverbial is deictic in nature. For example, to place *yesterday* on the timeline, we need information about the speaker's current temporal location. Conversely, -D adverbials, such as *in November*, can be placed on the timeline independently from the speaker's center of reference.

We ran an online acceptability judgment task using a 2x2x2 design with three independent variables (+/-T, +/-D, and marker). We created 64 stimuli (+96 fillers) in a Latin Square design. 160 subjects per language rated sentences on a 5-point Likert scale. Each stimulus was displayed separately and was accompanied by an introductory context. All sentences presented an

achievement to control for lexical aspect. An example item in English is shown in (6):

(6) Peter and Theresa are planning to go to a concert next weekend. Peter offers to go get the tickets later today, but Theresa tells him: I *purchased / have purchased* mine **this morning / at midnight / last month / in November.** It was cheaper that way.

Results. Mean acceptability scores are reported in Table 1. Linear mixed-effect analysis (random intercepts: subject and item) show in English a significant effect of T*D*Marker ($\chi^2(2) = 6.373$; $p < .05$). In all T/D adverbial combinations, there is a significant effect of marker favoring the PAST over the PERFECT. There seems to be a less categorical difference in the +T,+D condition, but a post-hoc test still shows the effect of grammatical marker ($\beta = 0.394$; $p = .035$). Interestingly, if we subdivide +T,+D adverbials by considering whether the adverb includes the day (S) or is included in it, we find that in the first case the difference across markers is still significant ($\chi^2(1) = 6.7711$; $p < .01$) and favors the PAST ($\beta = 0.5931$; $p < .001$), but when it comes to adverbs included in the day (S), the difference disappears ($\chi^2(1) = 0.5942$; $p = .4408$; PERFECT = 4.25; PAST = 4.38). Spanish presents a significant interaction of T*Marker ($\chi^2(1) = 47.12$; $p < .001$), with no effect of deixis. In -T adverbials, there is a main effect of grammatical marker ($\chi^2(1) = 57.07$; $p < .001$), favoring the PAST ($\beta = 1.353$; $p < .001$), but crucially, in the +T condition, there is no significant effect of marker ($\chi^2(1) = 0.016$; $p = .90$). Finally, Dutch only presents a main effect of marker ($\chi^2(2) = 32.117$; $p < .001$), favoring the PERFECT over the PAST in all conditions ($\beta = 0.8031$; $p < .001$).

Type of adverbial	Marker	English	Spanish	Dutch
+T, +D (*this morning*)	PERFECT	4.03	4.05	4.28
	PAST	4.42	4.31	3.37
+T, -D (*at midnight*)	PERFECT	3.34	4.33	3.78
	PAST	4.33	4.03	3.14
-T, +D (*last month*)	PERFECT	3.42	3.14	4.37
	PAST	4.51	4.53	3.58
-T, -D (*in November*)	PERFECT	3.44	3.21	4.07
	PAST	4.53	4.53	3.19

Table 1. Mean acceptability ratings per type of adverbial and tense-aspect marker in each language.

Discussion. Spanish speakers accept the PERFECT when the adverb is linked to the present. However, there is no preference for the *Pretérito Perfecto* in +T conditions: the *Pretérito Indefinido* receives similar ratings in these cases. English speakers prefer the *Simple Past* in all conditions but they accept the *Present Perfect* with deictic hodiernal adverbials, especially when the adverb is included in the day (S) (e.g., *this morning*). As expected, Dutch speakers prefer the PERFECT over the PAST across the board. In sum, our work provides evidence that both deixis and hodiernality play a role in PERFECT-PAST crosslinguistic variation. While Dutch allows the PERFECT to refer to past events unconstrainedly, Spanish restricts its use to events that are connected to the day of utterance, and English only allows it as far as these events are properly included in day (S) and are computed from the speaker's center of reference.

References. Harris, M. 1982. The 'past simple' and 'present perfect' in Romance. In M. Harris & N. Vincent (eds.), *Studies in the Romance Verb*. 42-70. // Hitzeman, J. 1995. A Reichenbachian Account of the Interaction of the Present Perfect with Temporal Adverbials. *NELS 25*, 17. // Klein, W. 1992. The Present Perfect Puzzle. *Language* 68, 525-552. // van der Klis, M., Le Bruyn, B. & de Swart, H. 2021. A multilingual corpus study of the competition between past and perfect in narrative discourse. *Journal of Linguistics*. 1-35 // Portner, P. 2003. The (Temporal) Semantics and (Modal) Pragmatics of the Perfect. *Linguistics and Philosophy* 26, 459-510. // Schwenter, S. 1994. The Grammaticalization of an Anterior in Progress. *Studies in Language* 18, 71–111. // Squartini, M., & Bertinetto, P. 2000. The Simple and Compound Past in Romance languages. In Ö. Dahl, *Tense and Aspect in the Languages of Europe*. 403-440.

Machine classification of modal meanings: An empirical study and some consequences

Aynat Rubinstein, Valentina Pyatkin, Shoval Sadde, Reut Tsarfaty, and Paul Portner

Introduction. We discuss the linguistic relevance of a computational study on modality (Authors 2021) which sets out to detect modals in texts without assuming they come from a closed class of lexical items, to classify their meaning in terms of modality type (or "modal flavor"), and to identify the eventuality they modalize. Building on a linguistically motivated annotation of modal meaning in news text (Rubinstein et al. 2013), we show that, while the detection of modal auxiliaries is trivial, detection and classification of a more open, semantically defined class is difficult. We also show that jointly performing the tasks of classifying the modality type and identifying the modalized eventualities produces superior results to doing either separately. We also discuss the distribution of modality types across our typology and the learnability of the subtypes. Our results suggest that the standard typology due to Kratzer (1981, 1991) should be restructured by grouping together epistemics and certain circumstantials as a "facts and knowledge" class.

The study. We use the **taxonomy** in Table 1 for classifying senses. It is based on a *coarse-grained* split between Priority modality (Portner 2009) and Plausibility modality and six *finer-grained* sub-types adapted by Rubinstein et al. (2013) in their annotated corpus. The taxonomy unifies and harmonizes the different modal senses offered by previous computational studies (Ruppenhofer & Rehbein 2012; Marasović & Frank 2016; Baker et al. 2012; Mendes et al. 2016). Examples with modals of a variety of parts of speech (POSs) are shown.

Priority	
Norms and Rules	*the ballot which* **must** *be held by the end of March*
Desires and Wishes	*extend our full* **support** *to the George W. Bush administration*
Plans and Goals	*a* **necessity** *emerged to enter the Pilgrim's House*
Plausibility	
State of Knowledge	*The ship is* **believed** *to carry illegal immigrants*
State of the World	*The disease* **can** *be contracted if a person is bitten*
State of the Agent	*They are* **able** *to do whatever they want*

Table 1: Proposed Taxonomy with Examples from GME.

	Baseline		RoBERTa	
	Aux V	All	Aux V	All
Modal/Not	99.04	68.24	99.9	73.2
Coarse-Grained	93.29	63.94	93.3	68.9
Fine-Grained	73.48	55.23	78.5	58.14

Table 2: F1 on Auxiliary Verbs (*can, could, may, must, should, shall*) vs. All triggers, Majority Vote Baseline vs. RoBERTa.

Rules	Intentions	Knowledge	World	Agent
60.42 (50.94)	46.1 (39.11)	59.27 (50.95)	54.64 (52.58)	72.72 (67.39)

Table 3: F1 results RoBERTa (vs. Baseline) for fine-grained senses. *Wishes/Goals* unified due to data sparsity.

For training and testing our models, we use the **Georgetown Gradable Modal Expressions Corpus** (GME; Rubinstein et al. 2013), a corpus obtained by expert annotations of the MPQA Opinion Corpus (Wiebe et al. 2005). We processed the corpus by extracting modal triggers and their prejacents into a CoNLL-formatted file. We added lemmas, POS tags, and dependencies using spaCy (Honnibal et al. 2020). As opposed to previous work, which trained and evaluated only on sentences known to contain modals, we use the entire dataset. We also accommodate sentences that contain multiple modals with different senses. We experiment with three **tasks**: (i) classifiying the sense of words specified by fiat as modal, (ii) detecting modal words and classifying their sense, and (iii) identifying also the modalized event. The results for the second task are shown in Table 2 (all results will be discussed in the talk), comparing a majority vote baseline to a fine-tuned RoBERTa-based classifier (Liu et al. 2019). The results show that detecting modality at the fine-grained level beyond the small set of modal auxiliary verbs is not trivial, with RoBERTa performing significantly better and far better than chance. The breakdown of RoBERTa's F1 scores is given in

Table 3. The largest label-wise gain in absolute points in comparison to the baseline is for *Rules* (∼10) and *Knowledge* (∼8), and the smallest is for *World* (∼2).

Consequencs for semantics. Rubinstein et al.'s (2013) annotation effort had already noted that the distinction between *Knowledge* (epistemic) and *World* (circumstantial) modality is often very unclear. An example from the corpus is given in (1):

(1) That will facilitate their **possible** convergence later with the international system.

On the *World* reading, (1) is based on some event taken as evidence, and the accessible worlds are the ones where that event is the same in relevant respects. The assertion of (1), on this reading, is that the circumstances of the US following certain standards of the Kyoto Protocol make it possible that its policies will converge at a later date. On the *Knowledge* reading, the same kind of evidence is relevant, but in addition, the mental state of the author plays a crucial role. In other words, the evidence isn't enough, and we need to include private knowledge of the author (i.e. that US officials intend to work towards covergence once the political situation changes) to understand why (1) is a justified assertion. Thus, the two readings differ in whether the speaker's mental state is involved *in addition to* an evidential event, not *instead of* it.

We see evidence for this view in the experiment in cases where the GME Corpus and the model differ in that one assigns *Knowledge* and the other *World*. There are 36 such cases, of which we judge both annotation and model to be correct in 17 of them (i.e., true examples of ambiguity). We judge only the model to be correct in 7 examples, only the annotators to be correct in 8, and 4 where neither was correct. Overall, the model errs on the side of annotation as *World* over *Knowledge*; this may be partially due to the fact that the model did not use extrasentential context. We also note that most of the confusion occurs with particular high frequency lexemes, in particular *would* (n=13), *could* (n=5), and *possible* (n=3), with idiosyncratic confusion around *clear(ly)* (n=3).

When a modal is embedded under an epistemic (or doxastic) operator, it is typically forced to take into account some knowledge (Hacquard 2006; Yalcin 2007). In (2), the modal background from *would* involves both relevant circumstances and the judgment of Mr. Carmona or other individuals at the company he represents. Annotators were correct in this case, whereas the model did not detect the *Knowledge* signal given by the embedding verb.

(2) Mr. Carmona said that operations **would** return to normal at the oil company.

The idea of collapsing epistemic modality with some cases of circumstantial modality is not new (Hacquard 2010, Kratzer 2012, p. 24), but our computational study sheds new light on the issues. We find that the model makes the smallest gains over baseline for the class of non-ability circumstantial modals (even setting aside cases which humans annotated as ambiguous between epistemic and circumstantial). We believe that collapsing circumstantial modality (perhaps not including ability modals) with epistemics would lead to more reliable classification, and we suggest that this change would reflect the linguistic reality that "epistemic modal" is not the class we thought it was. Examples like (1) and (2), and the comparison of the annotation and computational model, suggest that epistemic modality should be understood of as a sub-type of circumstantial modality.

Summary. We have shown that state-of-the-art NLP models can extract a significant amount of detailed information on the meanings of modal elements from annotated text. Perhaps more interestingly, they reveal patterns that fail to align with our standard theoretical assumptions, but which ultimately may be vindicated by a reassessment of the relevant categories. We have made this point with regard to the categories of epistemic/circumstantial modality, and in the presentation we will expand upon it regarding the split between modals and attitude verbs and the distinction between bouletic and teleological modality.

Title: Non-Doxastic Attitude Ascriptions and Semantic Meaning

Authors: Wojciech Rostworowski, Katarzyna Kuś, Bartosz Maćkiewicz (University of Warsaw)

Abstract: The aim of this talk is to provide new experimental evidence on (non-doxastic) attitude ascriptions and their entailment properties. We report two experiments using truth-value and acceptability judgement tasks, whose results suggest that the attitude verbs like 'want', 'fear' or 'glad' require a hyperintensional notion of meaning, including not only a truth conditional aspect but also the informational structure.

The problem under investigation emerges from the discussion on definite descriptions in attitude-verbs contexts. It has been observed that the statements ascribing a non-doxastic attitude to a subject (roughly, an attitude which does not involve *believing* in a proposition, e.g., 'hopes', or contains an extra component in addition to the belief, e.g., 'is glad') do not preserve their truth conditions once we substitute an embedded definite description with a corresponding 'there'-clause. For example, compare (1a) and (1b):

1. a. Hans wants the ghost in his attic to be quiet tonight.
   b. Hans wants *there to be* a (unique) ghost in his attic *and* for it to be quiet. (Elbourne 2010: 2)

Ascription (1a) seems to have different truth conditions than (1b) and it is possible for the latter to be intuitively false when the former is true (e.g., when Hans does not want to have any ghosts in his attic, but he actually believes that there is one and wants for that one ghost to be quiet). It is a matter of dispute whether there is a truth-conditional difference between the complement clauses in (1a) and (1b) (Russell 1905 vs Strawson 1950; for experimental findings see: Abrusán & Szendrői 2013, Schwarz 2016). A number of theorists (e.g., Heim 1991, Elbourne 2010, Schoubye 2013) has taken the contrast between (1a) and (1b) to be evidence against the Russellian interpretation of descriptions, thus explaining the contrast by positing a genuine semantic difference between the subclauses. One explanation appeals to the *presuppositional* status of definite descriptions – the subclause in (1a) presupposes the existence of a ghost while the subclause in (1b) does not, as the existence claim is a part of its assertoric content. However, further evidence suggests that contrastive ascriptions like (1) do not have to feature definites, but also various types of indefinite expressions (Schoubye 2013), and do not necessarily involve presuppositional differences (Blumberg 2017, Rostworowski 2018). For instance, (2a) is different from (2b):

2. a. Anne wonders whether the dictator has been assassinated.
   b. Anne wonders whether the dictator is dead and has been assassinated.

The subclauses in (2a) and (2b) do not, however, differ in terms of their presuppositions. In particular, 'being dead' is not presupposed by 'being assassinated', as it does exhibit typical projection behavior; instead, it seems to be an ordinary *entailment* (Rostworowski 2018: 1317-1323). Altogether, the theoretical literature indicates that the problem of substitutions in the scope of non-doxastic attitude verbs is more general and concerns the nature of these verbs rather than the issue of proper treatment of definite descriptions/presuppositions.

The aim of our first experiment (Study I) was to investigate to what extent the predictions of theoreticians about the contrast between a-type and b-type ascriptions are confirmed by evaluations of ordinary language speakers. Study I employed 2 (type of ascription: a-type vs b-type) x 2 (task: acceptability vs true value judgment) x 4 (non-doxastic attitude: fear, want, feel sorry, glad) mixed design. The first two factors were between-subject manipulation, the last one was within-subject. The study participants were presented with a set of simple contexts where a protagonist could be naturally ascribed a (non-doxastic) attitude of a-type, but not b-type. After each context, the participants were asked to evaluate an ascription of a given attitude (a-type or b-type ascription, depending on the experimental condition), that is, to say whether the ascription is true/acceptable in the context. Our informants were also requested to indicate how confident they are in their judgments. The main finding is a statistically significant effect of the type of ascription, with a-type rated much higher than b-type ($p < 0.001$).

The results of Study I indicate that there is a difference in both acceptability and truth conditions between a-type and b-type ascriptions, i.e., the former are more acceptable/regarded as 'true' in the contexts investigated in the study. This is in line with theoretical predictions. However, it is interesting that b-ascriptions are not fully rejected/regarded as 'false' in those contexts.

The aim of the second experiment (Study II) was to further explore the problem by investigating the nature of the discrepancy between a-type and b-type ascriptions. Roughly, there are two possible routes for the explanation to go: (i) we have a genuine semantic non-equivalence between the subclauses in a-type and b-type ascriptions, which goes beyond the mere presuppositional differences, and consequently generates non-equivalent readings of the ascriptions; (ii) a-type and b-type ascriptions are different for pragmatic reasons, in particular, in the contexts under investigation – where the a-type formulation is perfectly acceptable – b-type is 'infelicitous' as it violates the principle of 'contextual redundancy' (for details, see Blumberg 2017; cf. Fox 2008). The two approaches (i) and (ii) have different predictions about the status of Conjunction Elimination "under" attitude verbs (i.e., an inference to e.g. '*S* wants *p*' based on that *S* wants *p* and *q*). According to (i), it may be a true semantic entailment (as the two ascriptions have different sets of entailments); for (ii) it must be derivable on pragmatic basis, i.e., it is akin to a conversational implicature. In Study (II), we test this particular prediction by appealing to 'cancelability' (Grice 1989), that is, we check whether our informants find it coherent to ascribe an attitude towards a conjunctive proposition to a person and to deny that the person has the attitude towards the conjuncts in isolation.

Study II employed a within-subject design (non-doxastic attitude vs semantic entailment vs implicature). In this study, the participants were presented with two-sentence discourses. The first sentence attributed a non-doxastic attitude to a subject (e.g., 'Anne feels sorry that she went to the forest and found no mushrooms'). The second sentence denied that the subject had the attitude towards a single conjunct alone (e.g., 'In fact, she doesn't feel sorry about being in the forest'). The participants were asked whether such discourses were coherent on the 7-point pseudo-Likert scale. The discourses with attitude ascriptions were contrasted with discourses with canceled implicatures on the one hand, and with canceled semantic entailments, on the other hand. The main finding of Study II is that the discourses with attitude ascriptions were judged as generally incoherent – similarly to the discourses with cancelled semantic entailments (the ratings significantly below the midpoint, $p < 0.001$) – and much different from those with canceled implicatures, which were deemed to be coherent (although weakly; the ratings significantly above midpoint, $p < 0.01$).

The results of Study II confirmed the prediction that Conjunction Elimination is supported by the considered attitude verbs and that the inference is semantic rather than pragmatic. This is a significant empirical result in light of the observation that non-doxastic attitude verbs do not generally support entailments (e.g., Asher 1987, Kaplan 2005: 985). More importantly, the results suggest that there is a genuine semantic difference between ascriptions like (2a)/(2b) and hence attitude verbs operate on the semantic content of the complement clauses taken to include not only truth conditions but also the information structure that goes beyond presuppositions.

References:

Abrusán, M. & Szendrői, K. 2013. Experimenting with the king of France: topics, verifiability, and definite descriptions. *Semantics & Pragmatics* 6, 1-43.

Asher, N. 1987. A typology for attitude verbs and their anaphoric properties. *Linguistics and Philosophy* 10, 125-198.

Blumberg, K. 2017. Ignorance implicatures and aon-doxastic attitude verbs. In A. Cremers, T. van Gessel, F. Roelofsen (eds.) *Proceedings of the 21st Amsterdam Colloquium*,135-144.

Elbourne, P. 2010. The existence entailments of definite descriptions. *Linguistics and Philosophy* 33(1), 1-10.

Fox, D. 2008. Two short notes on Schlenker's theory of presupposition projection. *Theoretical Linguistics*, 34(3), 237-252.

Grice, P. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Heim, I. 1991. Artikel und Definitheit. In A. von Stechow, D. Wunderlich (eds.) *Semantik: ein internationales Handbuch der zeitgeňossischen Forschung*, 487–535. Berlin: Walter de Gruyter.

Kaplan, D. 2005. Reading 'On Denoting' on its centenary. *Mind* 114, 933–1003.

Rostworowski W. 2018. Descriptions and non-doxastic attitude ascriptions. *Philosophical Studies* 175(6), 1311-1331.

Russell, B. 1905. On denoting. *Mind*, 14, 479–493.

Schwarz, F. 2016. False but slow: evaluating statements with non-referring definites. *Journal of Semantics*, 33, 177-214.

Strawson, P. 1950. On referring. *Mind* 59(235) 320-344.

Schoubye, A. 2013. Ghosts, murderers, and the semantics of descriptions. *Noûs*, 47(3), 496-533.

Not that "fake" - Adults interpret the present counterfactual's "fake" past te

Maxime Tulling, Johanna Bunn & Ailís Cournane (New York University)

Counterfactual constructions such as the present counterfactual conditional (1a) express alternatives that are contrary to the actual state of affairs (cats do not have wings). The past morphology ("***had***") in such constructions is sometimes called "fake"[1], since the construction refers to an alternative state in the present. To refer to an alternative state in the past, one should use the past counterfactual construction (1b), where the past perfect expresses one layer of past temporal orientation and one layer of "fake" counterfactual past.

(1) a. If cats **had** wings, the human race would have become extinct due to flying tigers

b. If tyrannosauruses **had had** telescopes, they wouldn't have gone extinct.

In practice however, this is not always how adult native speakers of English express counterfactuals about the past. Crutchly[2,3] reported that adults spontaneously produce a wider range of tense combinations in their counterfactual constructions, showing that utterances with simple present in the antecedent (3) can also be used to encode a past counterfactual meaning. Prescriptively, the past perfect (*had taken*; *had lived*) is expected here.

(2) a. if they **took** my wages into consideration they would have let us buy next door even

b. if I **lived** with him first, I would never of married him (Crutchley, 2013, 15&16)

A small group of adults rating this type of utterance (which accounted for ~15% of spontaneous past CF conditionals) could not agree on the grammaticality of this construction[2]. In the current study, we investigated adult's interpretations of present and past counterfactual utterances. Since spontaneous production can be influenced by speech errors or context, we aimed to test people's intuitions about the meaning of counterfactual utterances in a controlled paradigm, asking the following questions: **1**) Can the present counterfactual convey past counterfactual meaning? **2**) Does the prescriptive rule reflect an older stage of a change-in-progress, and do younger people allow for this interpretation to a greater extent than older people?)

Hypotheses: We hypothesize that the present counterfactual construction can be understood as having past counterfactual meaning by reinterpreting the "fake" past tense marker to indicate past temporal orientation. We hypothesize this to be language change in progress and expect to find a generational effect, where older participants are more conservative than younger ones.

Methods: 50 adults were recruited online via Prolific and divided into 5 equal age groups: 18-28 (M=22.1-years, SD=2.53), 28-38 (M=31.3-years, SD=3.18), 38-48 (M=41.8-years, SD=3.30), 48-58 (M=50.9-years, SD=3.13) and 58-68 (M=63.0-years, SD=3.42). All participants completed an animated referent selection task that was designed for children and hosted on PCIbex Farm[1], to test the interpretation of past and present counterfactual constructions. Three identical characters ("kippies") choose milkshakes from the Milkshake Man. After the kippies pay with a coin of the same flavor as the milkshake they picked, the Milkshake Man produces a target utterance (e.g. "If that kippie had drunk a banana milkshake, he would have given me a banana coin") referring to one of the kippies. Utterances were divided into four main conditions: CONTROL, PAST, PRESENT COUNTERFACTUAL, and PAST COUNTERFACTUAL (Figure 1A). Participants were asked to select the kippie the Milkshake Man is talking about. The three possible referents are compatible with a **Past** (having drunk the mentioned milkshake), **Present Counterfactual** (holding a different milkshake) or **Past Counterfactual** (having drunk a different milkshake) interpretation of the utterance. Participants completed 8 trials (2 per condition) total. Order of presentation was pseudo-randomized and location of possible referents and milkshake flavors was balanced across the experiment. Counterfactuals always mentioned a banana milkshake (to facilitate a counterfactual reading, since the Milkshake Man particularly loves those coins).

Results: We excluded 6 participants for failing control trials. For those remaining (n=44), we

[1] Demonstration of experiment available here: https://farm.pcibex.net/r/rRfFjE/

calculated the percentage of responses per condition and age group (Figure 2). As expected, adults picked the **Past** referent on PAST trials, and the **Past CF** referent on PAST COUNTERFACTUAL trials, almost at ceiling. For the PRESENT COUNTERFACTUAL trials however, participants are split between choosing the **Present** or **Past CF** referent. This split was observed across all age groups. While adults selected more **Past CF** referents for PRESENT COUNTERFACTUAL WISHES than for CONDITIONALS, this difference was not significant at the group level, $\chi 2=2.6$, $p=.27$.

Discussion: The results of this study show that the present counterfactual can be interpreted as having a past temporal orientation, corroborating observations from (spontaneous) production[2,3]. Surprisingly, this was the case 50% of the time, and one participant commented sometimes two referents were possible. We found evidence against our hypothesis that this availability of a past tense interpretation is due to language change in progress, reporting the same pattern of results across all 5 age groups. What thus seems to be the case, is that participants can interpret the past tense morpheme in present counterfactuals either as a "fake" past tense, or as a real past tense, which raises questions about semantic accounts that rely on the past tense morpheme to derive counterfactuality[1,4,5] and the necessity of double tense marking in past counterfactuals.

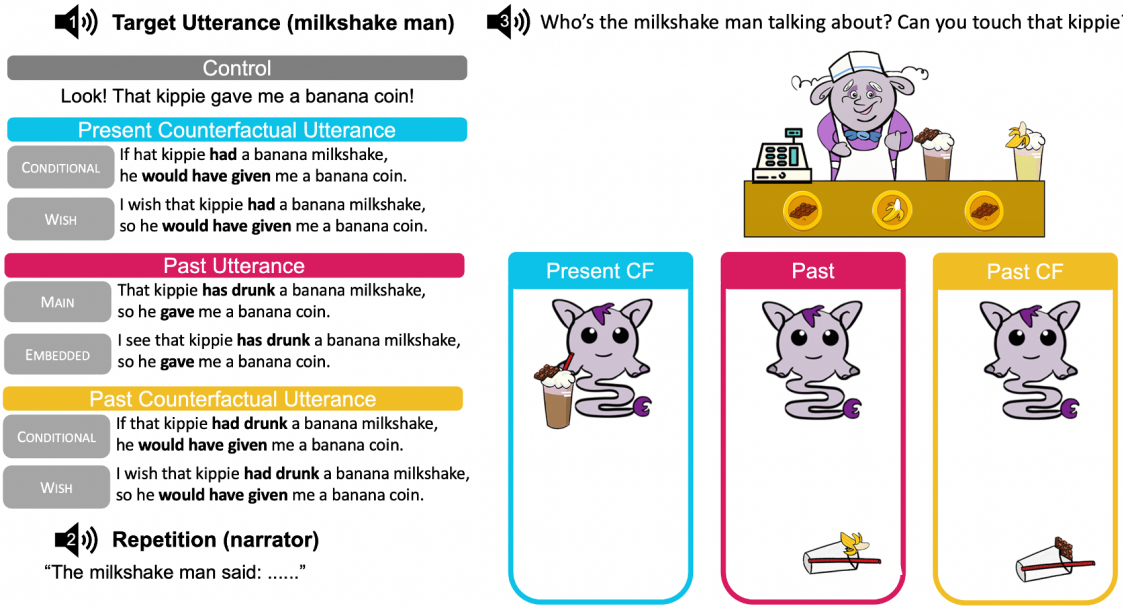


Figure 1. **A.** Target utterances divided per condition. **B.** Task Design showing three possible referents, each corresponding to be the target referent of one of the main utterance conditions.

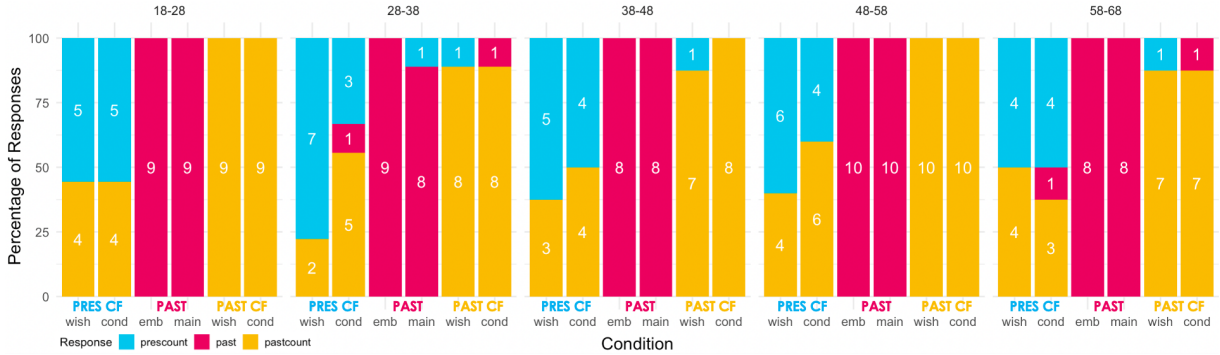


Figure 2. Count and percentage (y-axis) of responses per utterance condition split by age group.

References: [1] Iatridou, S. (2000). [2] Crutchley, A. (2004). [3] Crutchley, A. (2013). [4] Ippolito, M. (2006). [5] Karawani & Zeijlstra (2013).

Effects of entity relatedness and definiteness on bridging inferences

Mandy Simons, Carnegie Mellon University

Hannah Rohde, University of Edinburgh

An interpreter encountering an NP in discourse must decide whether its referent is part of a situation already constructed in their mental model (bridged interpretation), or is a new, unrelated entity. In (1), an entity *the living room* is introduced in a context sentence; subsequent NPs can be related to this situation or interpreted to refer to an unrelated entity.

(1) a. Jane was standing in the living room. The window...
b. Jane was standing in the living room. The congresswoman...
c. Jane was standing in the living room. A congresswoman...

A bridged interpretation for (1a) is easily inferable, with the window understood as a window in the living room (e.g., *The window was open and Jane could feel a breeze*) but a non-bridged reading may also be coherent (e.g., *The window of a car that drove by was rolled down and she could hear music blaring*). Work in formal pragmatics identifies two factors that are claimed to contribute to bridging inferences: entity relatedness (Asher & Lascarides 1998; Prince 1992) and definiteness (Clark 1975; Roberts 2003). For (1b), a bridged interpretation may be disfavored given the atypicality of a congresswoman in the living room, and for (1c) the indefinite determiner may undermine the referent's givenness (no bridge: *The/A congresswoman announced that the state was going into lockdown*). Experimental work on NP processing has found early robust effects of entity relatedness and also some influence of definiteness but only in later measures (N400 vs P600 in ERP; Schumacher 2009). This work, however, doesn't establish how to determine whether a comprehender has indeed established a bridged interpretation. Such interpretations are often assumed to arise (as in Clark's examples or Schumacher's materials), but to test what factors support bridging inferences, we need a clear metric of whether comprehenders indeed infer a bridged interpretation. Here we present three experiments that manipulate entity relatedness and definiteness while testing for the presence of bridging. The results show that entity relatedness affects the interpretation of an NP (offline & online); definiteness alone does not influence the interpretation but high-related definites favor bridged interpretations (online).

Experiment 1. Our goal is to test which properties of an NP encourage a bridged interpretation by assessing how participants treat that NP in a story continuation task. Participants (N=54, mturk) wrote story continuations for 40 targets, 40 fillers. Target items described a context followed by the potentially bridgeable NP (high vs low related; def vs indef), as in (2).

(2) [high, def] Ian likes to work at a large desk. The chair _____
[high, indef] Ian likes to work at a large desk. A chair _____
[low, def] Hilda created a nice arrangement of fruit. The chair
[low, indef] Hilda created a nice arrangement of fruit. A chair _____

The annotation process showed many cases of unambiguous bridging (*Ian likes to work at a large desk. The chair ... fits nicely underneath*) and cases of likely non-bridging (*Hilda created a nice arrangement of fruit. A chair... was on the porch*), but many were equally coherent assuming a bridged or a non-bridged interpretation. Relying only on the annotators' intuitions risked a circular treatment in which the coding would reflect the annotators' own sensitivity to the manipulated factors rather than their effect on participants. For example, similar continuations might be treated as bridges for high-related entities (*Ian likes to work at a large desk. The chair… is very comfy*) and non-bridges for low-related entities (*Hilda created a nice arrangement of fruit. The chair… is very nice*). Experiment 1 was thus inconclusive, but we used the resulting continuations to create a different paradigm in Experiment 2.

Experiment 2. To assess the potential bridges in participants' continuations in Expt 1, we asked new participants (N=55, mturk) to complete dialogues that probe whether the continuations generated in Expt 1 are interpreted by the new participants as bridged or not, as in (3,4). We used the Expt 1 data with 4 randomly selected continuations per NP per condition.

(3) Speaker A: Ian likes to work at a large desk. The chair leans back and was quite expensive.
Speaker B: Wait, sorry, I wasn't listening. Which chair are you talking about?
Speaker A: _____

(4) Speaker A: Hilda created a nice arrangement of fruit. The chair had dust on it.
Speaker B: Wait, sorry, I wasn't listening. Which chair are you talking about?
Speaker A: _____

Our assumption was that a bridged interpretation would yield Speaker A replies that repeated content from the context sentence. To illustrate with two sample replies, participants wrote *The one at the big desk Ian likes to work at* for (3) and *The chair that had dust on it* for (4), where the former indicates a bridged reading and the latter does not. We used a string similarity word-overlap metric whose scores we modelled with linear mixed-effect regressions. In using a continuous measure to identify bridging, we acknowledge that bridging may be a matter of degree and not a binary feature. For context~reply similarity, we found higher scores in the high-related than low-related condition ($p<.001$) and no effect of definiteness or interaction, suggesting that bridging primarily reflects entity relatedness. For continuation~reply similarity, we found higher scores for the low-related than high-related condition ($p<.001$); we also found a relatedness X definiteness interaction ($p<.001$), suggesting a pattern whereby more cases of non-bridging emerged in the low-related condition, particularly for indefinites in that condition.

Experiment 3. In a self-paced reading paradigm, we assessed RTs at the point in a sentence where a potential bridge is cancelled, with the prediction that factors that support a bridging inference will increase the processing difficulty if that bridge must be cancelled. Participants (N=100, prolific) read passages (40 target, 24 filler), with target items that consisted of a context sentence followed by a continuation with multi-word chunks for the determiner-noun, the start of a relative clause (RC), bridge-incompatible content in the RC, and two or more spillover regions.

(5) Context sentence: Jane_was_in_the_living_room.
[high, def] The_window that_was_in **her_dream** suddenly_came to_mind.
[high, indef] A_window that_was_in **her_dream** suddenly_came to_mind.
[low, def] The_knife that_was_in **her_dream** suddenly_came to_mind.
[low, indef] A_knife that_was_in **her_dream** suddenly_came to_mind.

For RTs at the critical bridge-incompatible region (bold in (5)), a linear mixed effects model showed a main effect of relatedness ($p<.05$) with slower RTs for high-related nouns. There was also a relatedness X definiteness interaction ($p<.05$), whereby the high-related definite condition had the slowest RTs. We take these results to show that entity relatedness is a core component for bridged interpretations and that definiteness, while it alone does not trigger bridging (i.e., low-related definites didn't give rise to the slowdown that would be evidence of bridging), acts together with high relatedness to make bridged interpretations more likely.

Overall, these results contribute to theoretical models of bridging and more broadly to a growing literature challenging analyses of definiteness as an unambiguous signal of givenness or contextual uniqueness.

Asher & Lascarides 1998. *Jrnl of Semantics*. **Clark 1975.** In Schank & Nash-Webber's *Theoretical issues in natural language processing*. **Prince 1992**. In Mann & Thompson's *Discourse description: Diverse linguistic analyses of a fund-raising text*. **Roberts 2003.** *Linguistics & Philosophy*. **Schumacher 2009.** In Lalitha Devi et al.'s *Anaphora processing*.

Commitment vs. discourse orientation : experimental and computational perspectives

Grégoire Winterstein, Ghyslain Cantin-Savoie, Samuel Laperle, Josiane Van Dorpe and Nora Villeneuve
Département de Linguistique – Université du Québec à Montréal

In this work, we argue in favor of distinguishing between (i) the *commitment* associated with an utterance, i.e. the set of language-external situations compatible with the meaning of an utterance, and (ii) the *discourse orientation* of the utterance, *i.e.* the discourse possibilities made available by the utterance. This distinction is rooted in early observations by Anscombre and Ducrot (1983) about the differences between the informational and argumentative content of natural language expressions. We illustrate these differences with the adverbs *almost* and *barely*, as seen in (1).

1. a. John is done with his beer.
   b. John is *almost* done with his beer.
   c. John is *barely* done with his beer.

Intuitively, an utterance of (1b) entails that (1a) is false (Ducrot 1972, Jayez & Tovena 2008). Nevertheless, in many cases, substituting (1b) for (1a) will not affect the overall felicity of the discourse. For example, if someone asks who needs another beer from the bar, both (1a) and (1b) would intuitively convey that John might need one. This contrasts with (1c), which seems to commit its speaker to the truth of (1a), but would be understood as conveying that John does **not** need a beer. This is unexpected: if being *almost* done with one's beer is grounds for ordering another, then being *completely* done (as is entailed by both (1b) and (1c)) should be even better grounds. But in the case of (1c), the opposite seems to be true. This conflict underlines the above distinction between *commitment* and *discourse orientation*. Jayez & Tovena (2008) account for these observations using a multilayered semantics in which *almost* conveys the negation of its prejacent via a conventional implicature (CI), and has the at-issue content in (2a) (adapted from J&T, where std(P) gives the standard degree of the property P). Conversely, *barely* conveys the truth of its prejacent via a CI, and has the at-issue content in (2b).

2. a. [[**almost**]] = $\lambda P.\lambda x.\ deg(P)(x) = d\ \&\ d > std(P) - \varepsilon$
   b. [[**barely**]] = $\lambda P.\lambda x.\ deg(P)(x) = d\ \&\ d < std(P) + \varepsilon$

J&T argue that the discourse orientation of an utterance is determined solely by its at-issue content, ignoring CIs. Given the meanings in (2), *almost* picks out higher degrees than *barely*, which grounds their different discourse orientations.

In this work, we report the results of two experiments that verify the empirical validity of the distinctions reported above for analogous contrasts in French. Having established that naïve speakers do indeed distinguish between commitments and discourse orientation, we then discuss a third experiment showing the implications of that distinction for computational language models, from the perspective of textual entailment and natural language inference.

The first experiment was designed to test commitment. Participants (n=43) were presented with out-of-context sentences. The target items instantiated one of the 3 conditions exemplified in (1): (i) no modification (labeled ø), as in (1a), (ii) the use of `presque` (≈'almost') as in (1b), or (iii) `à peine` (≈'barely') as in (1c). Participants used a slider to situate the meaning of the sentence relative to two extrema. Extrema were chosen so that the ø-condition would *a priori* denote the middle of the range, or some portion to its right. For example, in (1), the extrema were *30 minutes before John drinks the last drop of his beer* and *30 minutes after John drank the last drop of his beer*. Each participant saw 9 target items (3 in each condition), interspersed with 18 distractor items. Results are summarized in Fig. 1. We fitted linear mixed effect models with random intercepts for items and

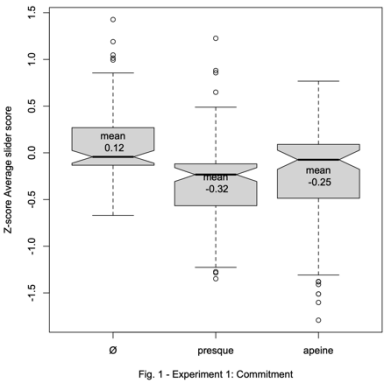


Fig. 1 - Experiment 1: Commitment

participants, and assessed the significance of our main factor via model comparison using likelihood ratio tests. We found a significant effect ($\chi(1)=34.741$, $p<0.001$), with the `presque` (≈'almost') condition being scored significantly below the other two. These results support the hypothesis that, with respect to commitment, the ø and `à peine` (≈'barely') sentences correspond to comparable situations, distinct from those denoted by the `presque`-sentences, which are situated "below" the other two conditions.

The second experiment was designed to test discourse orientation. It involved the same conditions and material as in experiment 1, but sentences were presented after a context that *a priori* licenses the ø-condition (as in the context described for (1) above). Participants (n=30) judged the naturalness of the target sentence in the given context, using a 7-point Likert scale. Results are summarized in Fig. 2. Model comparison between ordinal mixed models with random intercept and slopes for items and participants shows a significant effect of the factor under study ($\chi(1)=66.03$, $p<0.001$), supporting the hypothesis that as far as discourse orientation is concerned, the ø and `presque` sentences behave similarly, and are scored significantly higher than the `à peine` ones (in the full talk, we discuss contexts in which the relative acceptability of `à peine` and `presque` sentences is reversed). In summary, experiment 1 shows that with respect to commitment, ø and `à peine` group together to the exclusion of `presque`, while experiment 2 shows that, with respect to discourse orientation, ø and `presque` group together to the exclusion of `à peine`.

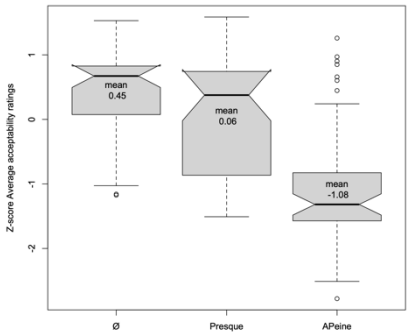


Fig. 2 - Experiment 2: Discourse orientation

These two experiments establish that speakers of French are indeed sensitive to the difference between commitment and discourse orientation. Our third experiment was designed to test whether state of the art language models such as BERT (Devlin et al., 2019) are similarly sensitive to this distinction. Distributional information plays a central role in shaping these models. For example, one of the training objectives of the BERT model is to predict whether, in a discourse of the form A B, the B sentence is a natural continuation of A (though this objective is not part of the training of all models, distributional information remains fundamental in their design Gastaldi, 2020). We therefore hypothesize that these models would be sensitive to the similarities exhibited in experiment 2 (*i.e.* discourse orientation), rather than to truth-conditional entailments like the ones of experiment 1 (*i.e.* commitment). To check this prediction, we tested inference patterns based on 200 semi-randomly selected examples from the French Wikipedia that involve the adverbs under study: *presque* and *à peine*. For each sentence, we used the pre-trained French CamemBERT model for natural language inference (Martin et al. 2020) to test whether the model predicts the truth of these sentences' prejacent. If the model is sensitive to commitment, it should predict the truth of the prejacent in the *à peine* cases and its negation for *presque*, and vice-versa if the model is sensitive to discourse orientation. The table on the right summarizes the average predicted probability of the prejacent and its negation for the *presque* and *à peine* cases. These results support the hypothesis that these models ground inference in discourse orientation rather than commitment.

	Infer prejac.	Infer ¬prejac.
presque	99.60	20.52
à peine	55.99	98.68

In the full paper, we discuss error patterns found in using BERT-like models for natural language inference tasks (*e.g.* Jiang & de Marneffe 2019), and show how these errors can be analyzed using the distinction between commitment and discourse orientation. We also argue that commitment and discourse orientation are confounded in the general case, which accounts for why this distinction has not yet been recognized in the computational literature.

References: Anscombre, JC & O. Ducrot *L'argumentation dans la langue* ♠ Devlin J. et al. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ♠ Ducrot O. (1972) Dire et ne pas dire ♠ Gastaldi, JL (2020) Why Can Computers Understand Natural Language? The Structuralist Image of Language Behind Word Embeddings. ♠ Jayez J. & L. Tovena (2008) Presque and almost: how argumentation derives from comparative meaning ♠ Jiang N. & MC de Marneffe (2019) Evaluating BERT for natural language inference: a case study on the CommitmentBank. ♠ Martin L. et al. (2020) CamemBERT: a Tasty French Language Model.

Testing the Influence of QUDs on Conditional Perfection

Britta Grusdt, Mingya Liu, Michael Franke

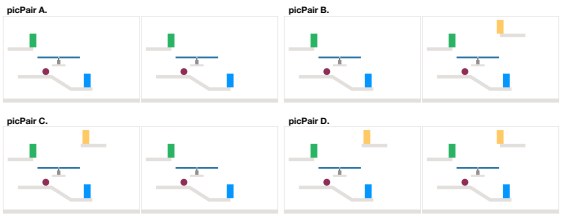
Introduction A longstanding subject of research in the context of natural language conditionals, e.g., expressions of the form "If p, q" ($p \rightarrow q$), is their interpretation as biconditionals, a phenomenon that became known as *Conditional Perfection* (CP) (Geis & Zwicky, 1971). The perfected interpretation of $p \rightarrow q$ involves an additional pragmatic inference (besides $p \rightarrow q$): "*q* only if *p*" or directly "If not p, not q" ($\neg p \rightarrow \neg q$). The degree to which a conditional is perfected seems to vary strongly between conditionals — leading to the question about the factors that influence whether and to what extent a conditional is perfected. This is what we aim to investigate here. More conceretely, we aim to test a theory proposed by von Fintel (2001) which predicts the occurrence of CP to be influenced by a question-under-discussion (QUD): when the QUD puts the focus on the consequent (what if *p*?), the conditional is interpreted as an exhaustive list of consequences of the antecedent *p*, hence CP is not expected, whereas when the QUD shifts the focus to the antecedent (will *q*?), an exhaustive list of conditions for *q* is expected thereby triggering perfection.[1] This theory has been tested empirically before (Cariani & Rips, 2016; Farr, 2011) yielding conflicting results. We will present a novel experiment using visual stimuli (scenes of block arrangements) that explicitly show a very constrained context and should thereby not elicit latent, uncontrolled beliefs, which likely happens in experiments that use text-based stimuli (see Cariani & Rips, 2016).

Experiment 300 native English speaker were recruited via the online Platform Prolific. The cleaned data comprises data from 282 participants (103 male, 175 female, 1 other) with a mean age of 32.8 (range 18 – 84).[2] **Design & Material**. We use a 3×4 within-subject design, manipulating the QUD, as encoded in an question (*neutral*, *if-p*, *will-q*) of an interlocutor, and the shown stimulus (picPair A-D). Each stimulus is a pair of what we call an exhaustive (left picture) and a non-exhaustive situation (right picture). In exhaustive situations the consequent-block (blue block in Fig. 1(a)), only falls when the antecedent-block (green block) falls and in non-exhaustive situations, there is a second reason for the consequent-block to fall, either because of its position on the edge or because of another falling block (yellow block). **Hypothesis**. According to the theory from von Fintel (2001), we should see an effect of the QUD on the selection rate of the exhaustive situation: participants are expected to choose the exhaustive situation more often with QUD=*will-q* than with QUD=*if-p* since contrary to the non-exhaustive situation, the exhaustive situation represents a biconditional interpretation of the conditional. **Procedure**. First, participants saw 8 training trials with animations of block arrangements to get familiar with the physical behavior of the blocks. In the subsequent test phase (12 critical + 6 control trials) participants first read a dialogue between two persons, Ann and Bob. After participants finished reading Ann's question and Bob's response[3], they were shown two situations and were asked to select the one that they rated as more likely described by Bob. **Results**. Figure 1(b) shows the proportion of participants who selected the exhaustive situation as the situation that Bob is more likely to describe. We run a Bayesian logistic regression model (using brms, Bürkner, 2017) that predicts participants' choice (exhaustive vs. non-exhaustive situation) based on the QUD and the picture pair, using default priors, varying intercepts and slopes per participant for both predictors and an interaction term.

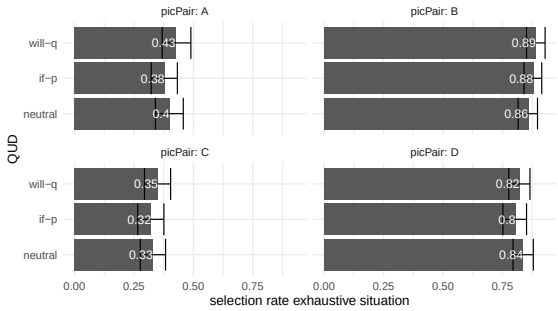
[1] Levels QUD:*neutral*: "Which blocks do you think will fall?", *if-p*: "What happens if the antecedent-block falls?", *will-q*: "Will the consequent-block fall?"

[2] Anonymized link to preregistration: `https://osf.io/47w85?view_only=dd070669fad44969b698698f7e413dc3`.

[3] In all critical trials, Bob's response is "If the antecedent-block falls, the consequent-block will fall" where 'antecedent-' and 'consequent-' were replaced by the appropriate randomly assigned color, 'BLUE' or 'GREEN'.



(a) 4 critical stimuli where, for each pair, the exhaustive situation is on the left and the non-exhaustive situation on the right.



(b) Bars are average selection rates for the exhaustive situation separately for each QUD and stimulus (picPair), errorbars are 95% bootstrap confidence intervals.

Figure 1: Stimuli and results of critical trials.

We find strong evidence for the hypothesis formulated above for stimuli A (posterior probability 0.95). For the remaining stimuli the posterior probabilities are 0.70 (B), 0.83 (C) and 0.71 (D). The overall effect of the QUD in the predicted direction has an estimated posterior probability of 0.95.

Discussion & Conclusion Our overall results show a tendency in line with our hypothesis based on the QUD-account on CP even though the data is not conclusive. Two aspects are particularly interesting thereof: on the one hand, the effect of the QUD on the selection rate of the exhaustive situation (larger for *will-q* than for *if-p*) seems to be larger for stimuli A+C than for B+D. On the other hand, in the former two stimuli, the conditional does not tend to be perfected to the same extent as in the latter two: the selection rate for the exhaustive situation is constantly below 0.5 in A+C but close to ceiling in B+D. A possible explanation for both observations may lie in the set of *salient* alternative utterances available to the speaker. In B+D, the second cause for the consequent-block to fall in the non-exhaustive situation can clearly be communicated with a salient alternative conditional, 'green or yellow $\rightarrow$ blue' which would be more informative than the uttered conditional 'green $\rightarrow$ blue'. This may explain the large values of the selection rates of the exhaustive situation which also makes a potential effect of the QUD harder to detect. Contrary to that, in A+C, the second cause is visible in the non-exhaustive situation, but there is no salient alternative conditional.[4] Quite the opposite, there is an alternative conditional for the exhaustive situation that discriminates both: 'only blue $\rightarrow$ green' which might explain the large difference in the selection rates for the exhaustive situation across QUDs in A+C as compared to B+D. In a follow-up experiment, we plan to investigate the interaction between QUDs, context and alternative utterances.

References

Bürkner, P.-C. (2017). **brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1).

Cariani, F., & Rips, L. J. (2016). Experimenting with (Conditional) Perfection. *Unpublished manuscript*.

Farr, M. (2011). Focus influences the presence of conditional perfection: Experimental evidence. *Proceedings of Sinn & Bedeutung*, *15*, 225–239.

Geis, M. L., & Zwicky, A. M. (1971). On invited inferences. *Linguistic inquiry*, *2*(4), 561–566.

von Fintel, K. (2001). Conditional strengthening. *Unpublished manuscript*.

[4] Alternatives in the non-exhaustive situation for A+C are rather "blue falls' or "blue might fall'.

The information structure of word order alternations

Giuseppe Ricciardi (Harvard University), Edward Gibson (MIT)

Some researchers have argued that the noncanonical variant of a word-order alternation with two possible orders to present two NPs is systematically dispreferred when the first NP (NP1) is discourse-new and the second NP (NP2) is discourse-old (e.g., Birner & Ward 1998); we refer to this as the "**co-dependence hypothesis**". In support of co-dependence, Clifton & Frazier (2004) and Brown et al. (2012) showed that "old-new" was strongly preferred over "new-old" in the noncanonical NPNP variant (*John gave [the teacher]$_{NP1}$ [a book]$_{NP2}$*) of the dative alternation, but not in the canonical NPPP variant (*John gave [a book]$_{NP1}$ to [the teacher]$_{NP2}$*). Although the current data from the dative alternation are consistent with this view, the proposal is descriptive, without independent motivation. An alternative view holds that these findings are determined by ***discourse preferences independently affecting each NP***: for the canonical NPPP both NPs are subject to an "old-over-new" preference - in line with the assumption that old information is easier to process than new information (e.g., Arnold et al. 2000, 2013) - whereas for NPNP, NP1 is subject to the default "old-over-new" preference but NP2 is subject to the opposite "new-over-old" preference - in line with the hypothesis that the production of NPNP structures is driven by a "new-final" requirement ("non-co-dependence hypothesis"). To discriminate between these two hypotheses, an acceptability rating study (E1) was conducted investigating all four combinations of the two critical NPs' discourse-status (not just those that differed in discourse status, as in previous work). We also conducted a second study (E2) to test whether the finding from Clifton & Frazier (2004) and Brown et al. (2012) extends to a different case of English word order alternation, i.e. 'locative inversion', as in (2).
Methodology: In both studies, participants rated the second of two sentences within the context of the first. We manipulated the second sentence by crossing the word orders (E1: NPPP/PPNP/NPNP; E2: NPvPP/PPvNP) with the two NPs' status (old/new) (see 1 and 2). **Predictions**: The co-dependence hypothesis predicts only for the noncanonical variants (E1: PPNP/NPNP; E2: PPvNP) an interaction effect between the two NPs' status such that new-old is rated worst of the four, with no differences among the other three. The non-co-dependence hypothesis instead predicts for all variants except for PPNP/NPNP two main effects of the NPs' status, such that sentences with an old NP1/2 will be preferred over those with a new NP1/2; for PPNP/NPNP it predicts a main effect of NP1 in the default old-over-new direction and a main effect of NP2 in the opposite direction (new-over-old). **Results: E1:** Focusing on the two conditions where the two NPs differ in discourse status, we replicated previous findings: for PPNP/NPNP but not for NPPP "old-new" sentences were rated better than "new-old" sentences. However, for no word order did we find a significant interaction between the two NPs' status (ps>.1). Instead, for NPPP we found main effects of the two NPs such that "old" is better than "new"; for PPNP/NPNP we found a main effect of NP1 in the default old-over-new direction and a main effect of NP2 in the opposite direction (new-over-old). **E2:** For NPvPP, we found main effects of the two NPs in the old-over-new direction. For PPvNP, we found a significant old-over-new main effect of NP1 and only a numerical one for NP2. Again, for no word order did we find a significant interaction between the two NPs' status (ps>.4). **Conclusion**: Although we replicated findings from previous works concerning the dative alternation, we showed that these results are determined by the combination of independent discourse preferences for each NP: two preferences in the same-direction ("old-over-new") in the canonical order and two preferences in opposite directions ("old-over-new" for NP1 and "new-over-old" for NP2) in the noncanonical orders. Furthermore, we showed that the findings about the dative alternation don't extend to the locative alternation case, where we found main effects of the two NPs in the same direction "old-over-new" across word orders. Overall, our findings support a view where the information structure of word order alternations is affected by a general preference for old over new NPs which can be overwritten when a NP occurs in a non-canonical position.

(1) E1 (Dative Alternation) example item (N = 64; N items = 24)

old-old
Context: A professor was exhausted because he had been working together with an administrator on the first draft of a grant all day long.
The professor sent the grant to the administrator [NPPP] / **The professor sent (to) the administrator the grant** [PPNP/ NPNP]

old-new for NPPP and new-old for PPNP / NPNP
Context: A professor was exhausted because he had been working on the first draft of a grant all day long.
The professor sent the grant to an administrator [NPPP] / **The professor sent (to) an administrator the grant** [PPNP/NPNP]

new-old for NPPP and old-new for PPNP / NPNP
Context: A professor was exhausted because he was writing long emails to an administrator all day long about personality conflicts among the faculty.
The professor sent a grant to the administrator [NPPP] / **The professor sent (to) the administrator a grant** [PPNP/NPNP]

new-new
Context: A professor was exhausted because he was writing long emails all day long about personality conflicts among the faculty.
The professor sent a grant to an administrator [NPPP] / **The professor sent (to) an administrator a grant**

(2) E2 (Locative Alternation) example item (N = 57; N items = 24)

old-old
Context: The police officer entered the room and saw a hunting weapon, a broken chair, a box, and a scary painting.
The weapon lay behind the box. [NPvPP] / **Behind the box lay the weapon**. [PPvNP]

old-new for NPvPP and new-old for PPVNP
Context: The police officer entered the room and saw a hunting weapon, a broken chair, an open cupboard, and a scary painting.
The weapon lay behind a box. [NPvPP] / **Behind a box lay the weapon**. [PPvNP]

new-old for NPvPP and old-new for PPvNP
Context: The police officer entered the room and saw an empty bottle, a broken chair, a box, and a scary painting.
A weapon lay behind the box.

new-new
Context: The police officer entered the room and saw an empty bottle, a broken chair, an open cupboard, and a scary painting.
A weapon lay behind a box. [NPvPP] / **Behind a box lay a weapon**. [PPvNP]

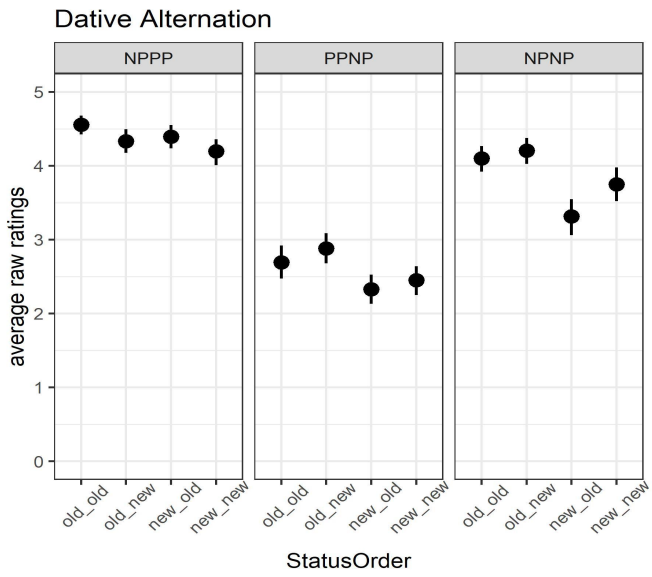


Fig 1 Mean ratings for E1 by discourse status order condition.

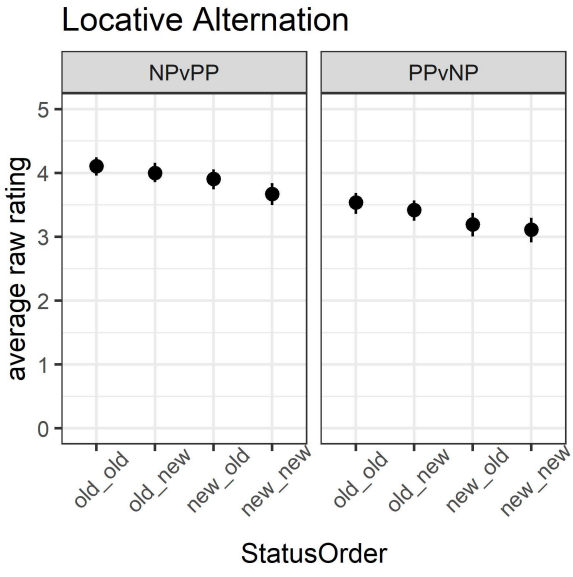


Fig 2 Mean ratings for E2 by discourse status order condition.

Generating Discourse Connectives with Pre-trained Language Models: Do Discourse Relations Help? *

Symon Jory Stevens-Guille† Aleksandre Maskharashvili† Xintong Li‡ and Michael White†
†The Ohio State University ‡Baidu Research

1 Motivation and Setup

Traditional approaches to discourse have shown the essential importance of discourse (rhetorical) relations in providing coherence to a text [1, 2, 3]. Current approaches to natural language generation (NLG) employing pre-trained models have been shown to excel in generating well-formed text [4], but their ability to produce coherent texts structured with the help of discourse connectives is understudied [5]. Therefore, the study of how well pre-trained models realize discourse relations is of significant interest in the NLG community.

We report results of our experiments using BART [6] and the Penn Discourse Tree Bank [7] (PDTB) to generate texts with correctly realized discourse relations. We address a question left open by previous research [8, 9] concerning whether conditioning the model on the intended discourse relation—which corresponds to adding explicit discourse relation information into the input to the model—improves its performance.

BART, being a transformer [10] based language model, is trained on purposefully corrupted data so that the model learns to 'denoise' the corrupted input in the process of reconstructing the original data. Fine-tuning BART on different versions of input and output lets us probe whether the underlying language model needs or benefits from explicit cues to consistently reconstruct an adequate discourse connective. The PDTB is one of the few corpora developed to identify discourse dependencies between texts. It provides a well-developed ontology of discourse relations; these discourse relations are used to annotate the Wall Street Journal corpus. We consider versions of the corpus differing in (i) whether the order of the arguments in the output is explicitly encoded in the input, (ii) whether the output is the connective or the connective embedded in the corresponding WSJ text, (iii) whether a discourse relation is included in the input and how specific it is. The third is the most important difference since it corresponds to whether the model is conditioned on discourse relation information. We refer to models conditioned on discourse relations by BART_{D+} and models not conditioned on discourse relations by BART_{D-}.

In order to determine how well the models perform in realizing discourse relations, we employ standard metrics, e.g. precision, recall, F-1, and devise some new metrics inspired by psycholinguistic and corpus studies to determine the degree to which the models' preferences for realizing different discourse relations correspond to reported human preferences for realizing those relations [11, 12, 8]. While space precludes reporting of the results on these new metrics in the abstract, we intend to report them subsequently.

2 Results and Discussion

Our results show that fine-tuning BART on the different versions of PDTB inputs and outputs mentioned in the foregoing consistently produces discourse connectives which match those used in the text. The BART_{D+} models nonetheless outperform the BART_{D-} models. It's noteworthy that the best BART_{D+} model matched (recall = 79%) on hundregs of additional data points compared to the best BART_{D-} (recall = 71.3%) model (McNemar's Test Statistic 313; $p < .000$). With respect to matching on explicit connectives, the BART_{D+} (69.8%) model matched significantly more than the BART_{D-} (54.3%) model (McNemar's Test Statistic 157; $p < .000$). With respect to matching on implicits the BART_{D+} (89.2%) model is slightly but significantly worse than the BART_{D-} (90.6%) model (McNemar's Test Statistic 118; $p < .025$), though this seems to reflect the overprediction of implicits by the BART_{D-} model.

*This research was supported by a collaborative open science research agreement between Facebook and The Ohio State University. E-mail: stevensguille.1@buckeyemail.osu.edu

The results reported above are in line with the view that information concerning discourse relations should be present in the inputs of neural approaches to NLG [13, 14, 5], which has not typically been the case. When the metrics are extended to include whether non-matching connectives chosen by the model fit the intended discourse relation, the best BART_{D+} model continues to outperform the best BART_{D-} model. When producing non-matching connectives, we find that the chosen connectives of the BART_{D+} models correspond to the intended discourse relations more frequently than those produced by the BART_{D-} models.

The main conclusion one can draw from our results is that discourse relation information is essential for consistently generating matching discourse connectives beyond the sentence level. While large-scale human judgement experiments on our model's predictions are the most obvious next step, the improvement of the BART_{D+} models over the BART_{D-} models with respect to exact matching is encouraging, especially in light of recent results showing that humans don't uniformly accept substitution of discourse connectives which express the same discourse relation [8]. With respect to whether mere arguments suffice to predict the discourse connective holding between them, our results indicate that the purely distributional meaning of texts induced by the models under-determines the meaning of explicit discourse connectives. Directly conditioning on explicit discourse relations significantly improves the match between discourse connective produced and discourse relation intended to be expressed.

To sum up, our results suggest that the intended discourse relation cannot always be inferred from the arguments using pre-trained models. Inclusion of the discourse relation in the input provides an immediate boost to control over output coherence.

References

[1] W. C. Mann and S. A. Thompson, *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles, 1987.

[2] A. Lascarides and N. Asher, "Segmented discourse representation theory: Dynamic semantics with discourse structure," in *Computing meaning*, pp. 87–124, Springer, 2008.

[3] A. Kehler and A. Kehler, *Coherence, reference, and the theory of grammar*. CSLI publications Stanford, CA, 2002.

[4] M. Kale and A. Rastogi, "Text-to-text pre-training for data-to-text tasks," in *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 97–102, 2020.

[5] A. Maskharashvili, S. Stevens-Guille, X. Li, and M. White, "Neural methodius revisited: Do discourse relations help with pre-trained models too?," in *Proceedings of the 14th International Conference on Natural Language Generation*, (Aberdeen, Scotland, UK), pp. 12–23, Association for Computational Linguistics, Aug. 2021.

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[7] B. Webber, R. Prasad, A. Lee, and A. Joshi, "The penn discourse treebank 3.0 annotation manual," *Philadelphia, University of Pennsylvania*, 2019.

[8] F. Yung, M. Scholman, and V. Demberg, "A practical perspective on connective generation," in *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pp. 72–83, 2021.

[9] W.-J. Ko and J. J. Li, "Assessing discourse relations in language generation from gpt-2," in *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 52–59, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[11] J. D. Murray, "Connectives and narrative text: The role of continuity," *Memory & Cognition*, vol. 25, no. 2, pp. 227–236, 1997.

[12] T. Sanders, "Coherence, causality and cognitive complexity in discourse," in *Proceedings/Actes SEM-05, First International Symposium on the exploration and modelling of meaning*, pp. 105–114, University of Toulouse-le-Mirail Toulouse, 2005.

[13] A. Balakrishnan, V. Demberg, C. Khatri, A. Rastogi, D. Scott, M. Walker, and M. White, "Proceedings of the 1st workshop on discourse structure in neural nlg," in *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, 2019.

[14] S. Stevens-Guille, A. Maskharashvili, A. Isard, X. Li, and M. White, "Neural NLG for methodius: From RST meaning representations to texts," in *Proceedings of the 13th International Conference on Natural Language Generation*, (Dublin, Ireland), pp. 306–315, Association for Computational Linguistics, Dec. 2020.

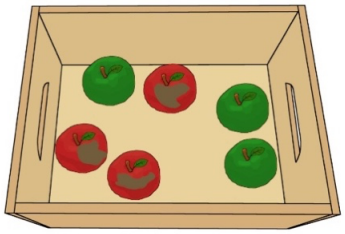
The rise and particularly fall of presuppositions: Evidence from duality in universals

Remus Gergel*, Maike Puhl*, Simon Dampfhofer°, Edgar Onea°
* Saarland University, ° University of Graz

SYNOPSIS: Our paper contributes to the larger endeavor to use experimental linguistics to elucidate diachronic issues (cf. Zhang, Piñango & Deo 2018, Fedzechkina & Roberts 2020). We provide first experimental evidence regarding the direction in which language change proceeds regarding the historical loss or acquisition of presuppositions (PSP) in lexical material, a topic debated in the theoretical literature (e.g. Eckardt 2006, 2009 vs. Gergel 2020).
BACKGROUND: It has long been noticed diachronically that implicatures tend to conventionalize (and thereby disappear as implicatures), but tendencies of PSPs remain under-investigated. Eckardt (2006, 2009), treating implicatures and PSPs together, claims them (non-experimentally) to be subject to *Avoid Pragmatic Overload.* Simplifying: when there are too many side messages, dispense with them/some. One possible diachronic consequence that we *could* derive: (some) PSPs are prone to be lost over time. *Prima facie* contrarily, Gergel (2020) argues for a diachronic version of *Maximize Presuppositions*. A possible consequence we *could* derive: the marking of PSP triggers may be increasing over time. Since both approaches are theoretically well motivated in the diachronic context, novel experimental evidence would be welcome to help elucidate the debate.
GOAL: In this paper, we approach the issue precisely from such an experimental perspective, by focusing on reinterpretive learning of actual and potential PSP triggers in change following the assumption of Eckardt (2006) that semantic change is typically caused by adults and thus diachronic processes resemble of second, rather than first, language acquisition. Since this topic has never been experimentally discussed before, we start with an exploratory experiment on one single lexical item that shifts between the meanings of BOTH / ALL. Thereby we assume (Heim & Kratzer 1998) that words like *both* are universal quantifiers but additionally presuppose their restrictor cardinality to be two. Thus the theoretical issue is operationalized at the level of our study as follows: Will participants find it easier to re-acquire an item they learned as meaning BOTH used as ALL (both→all) or vice versa?

METHOD: We conducted an exploratory study with 25 native speakers of German (11m/14f) with mean age 23.1 (SD 3.2) from a South-Eastern-Region (in a German speaking country), split into two groups, which determined whether they would learn a nonce word *gure* in the meaning BOTH or ALL during training. Subsequently they were exposed to contexts leading to a reinterpretation towards the respective other meaning. Whence the denomination of the groups: *both→all* vs. *all→both*. Participants were asked to imagine visiting a fictitious community of German speakers in the US guided by a native speaker who studies with them in Vienna. To make this as plausible as possible, we used spoken stimuli produced in a remote and little prominent variant of a Mosel-Franconian (West-German) dialect. Training: Participants were first taught a non-word (*gure),* which would represent either BOTH or ALL, spoken by an older member of the community. This was done via example: they were presented images on a computer screen and then heard sentences containing the target non-word describing the situation. The old person would then tell the participants whether the sentence was true in the situation presented or not. If the sentence was not true, he in addition provided a reason why it was false. After three training items each, participants were asked to rate the truth of five sentences themselves (on a binary scale). After each judgement, they received written feedback from the older speaker whether their choice was correct. During training, all items were shown in fixed order. We also included six filler items containing two different non-words with no presuppositional meaning intended. Training was successful in all cases, we do not report results on the training phase here for lack of space. Main experiment: Participants were asked to imagine visiting a reunion of younger members of the community. There, two characters are of importance: their friend F, who having been abroad for some time is not up to date with current language developments (within the younger members of the community), and a high prestige competent local speaker S. In this context participants are faced with examples showing that *gure* is used by S precisely in the respective opposite meaning of what they learned from the old person (i.e. both→ all or vice versa). F, by contrast represents an older low-prestige stage of the language. Participants were again shown

pictures and heard sentences containing the target word. The "BOTH and ALL conditions were met in half of the items each. Participants were then asked to rate their agreement for the sentence in the presented situation on a scale from 1 to 10. After rating, they read text from both younger speakers commenting on the situation (but not explicitly the truth of the sentence), with either both agreeing or one of them remarking they found something odd. The translation of an example item is shown below.



Someone utters: *Gure red apples are rotten.*
Task: Rate acceptability on a scale from 1 to 10.

all→both:

S: That's not right. "Gure" is something my grandma would say in this case!

both → all:

F: Didn't she see the third apple?

S: Why? She said *gure*. She was right

RESULTS: We analyzed the data with a linear model with group and order of test items as predictors and judgment values and times as dependent variables in R (plotting with Lüdecke 2021). These jointly provide a good overall insight into the speed of learning the new usage of *gure* in the younger community.


```
lm(formula = judgment ~ group * order, data = experiment)
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.206019   0.749599   2.943   0.0036 **
groupboth-->all          0.001246   1.039506   0.001   0.9990
order                    0.334722   0.133207   2.513   0.0127 *
groupboth-->all:order    0.240919   0.184725   1.304   0.1935
```

```
lm(formula = t_judgment ~ group * order, data = experiment)
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2971.5     613.0   4.848 2.35e-06 ***
groupboth-->all          3418.9     850.0   4.022 7.92e-05 ***
order                    -147.3     108.9  -1.352 0.177704
groupboth-->all:order    -541.3     151.1  -3.584 0.000417 ***
```

The graphs show model predictions of judgment and judgment time respectively depending on the order in which target items were presented in the main experiment. Left-hand graph: Low values of judgment represent accepting the nonce word as learned during training, while higher values mean participants accepted the new meaning introduced by the competent young speaker. There is a tendency of learning the *both→all* direction of reinterpretation more quickly and reliably, however this interaction did not come out significant in the model. Right-hand: Higher judgment times in the early phase suggest that progress through repetitions facilitates a speeding-up in the experiment which amounts to learning effects. The linear model shows the group:order interaction is highly significant. The two findings converge, in that the direction of reinterpretation PSP→non-PSP usage takes place at a higher speed than the opposite one.

DISCUSSION: This is a novel finding given the theoretical literature discussed above. It suggests this direction of language change (PSP→no-PSP) is the more likely one. Of course, given a number of limitations of our experiment, this suggestion is very preliminary: we only studied one item, our results are only significant for the judgment times and for the particular item used, etc. However, our experiment paves the way towards a novel paradigm of language change - semantic language processing interface much in line with studies of change (from different perspectives) such as Zhang, Piñango & Deo 2018, Fedzechkina & Roberts 2020, and others.

SELECTED REFERENCES: Eckardt, R. 2006. *Meaning Change in Grammaticalization*. OUP. || Fedzechkina, M., & Roberts, G. 2020. Learners sacrifice robust communication as a result of a social bias. https://doi.org/10.31219/osf.io/usfhz. || Gergel, R. 2020. Sich ausgehen: Actuality entailments and further notes from the perspective of an Austrian German motion verb construction. *Linguistic Society of America*, *5*(2), 5-15. || Lüdecke, D. 2021. sjPlot: Data Visualization for Statistics in Social Science. https://CRAN.R-project.org/package=sjPlot ||

Zhang, M., M. Piñango & A. Deo. 2018. Real-time roots of meaning change: Electrophysiology reveals the contextual-modulation processing basis of synchronic variation in the location possession domain. *40th Annual Conference of the Cognitive Science Society*, 2783–2788.

Comparing Global and Local Accommodation: Rating and Response Time Data

Alexander Göbel (McGill) & Florian Schwarz (UPenn)

Preview. Presuppositions (PSPs) are commonly characterized as backgrounded content that is taken for granted by the speaker (Stalnaker 1974). However, not all presuppositions need to be explicitly satisfied in the prior discourse in order to be felicitous, i.e. they can be accommodated, either at the utterance level (Global Accommodation = GA; Lewis 1979) or in embedded contexts (Local Accommodation = LA; Heim 1983). While GA and LA share a common label, it is an open question whether they also share an underlying mechanism. We present a speeded acceptability rating study that directly compares Global and Local Accommodation for five PSP triggers (*again*, *still*, *even*, *regret*, *discover)* in order to bring behavioral evidence to bear on this issue.

Background. One unified formal treatment of GA and LA is Beaver & Krahmer (2001)'s *A* operator, which turns presupposed content into asserted content, such that GA and LA can be reduced to a difference in the syntactic position of this operator at LF. In contrast, von Fintel (2008) makes a conceptual argument that GA is a pragmatic operation, where the hearer adjusts the context to match the meaning of an utterance, whereas LA is a semantic operation adjusting the meaning of a sentence to avoid a clash with the context (typically seen as a last resort). An intermediate position comes from Klinedinst (2016), who argues that only triggers that entail their PSP allow for a unified treatment of GA and LA, while other triggers require distinct mechanisms.

Design. We crossed ACCOMMODATIONTYPE (*global* vs *local*) and CONTEXT (*PSP met* vs *PSP unmet*) in a 2x2 Latin-square design, using short dialogues as in (1). These consisted of four clauses, with the second context clause either supporting the relevant PSP or expressing Explicit Ignorance with regard to it. The third - target - clause contained the PSP trigger. ACCTYPE was manipulated by making the target clause a root clause (followed by *so*; global) or an *if*-clause (local). Additionally, the context clause was either uttered by speaker A in the global condition or by speaker B in the local condition, in order to ensure accommodation at the appropriate level.

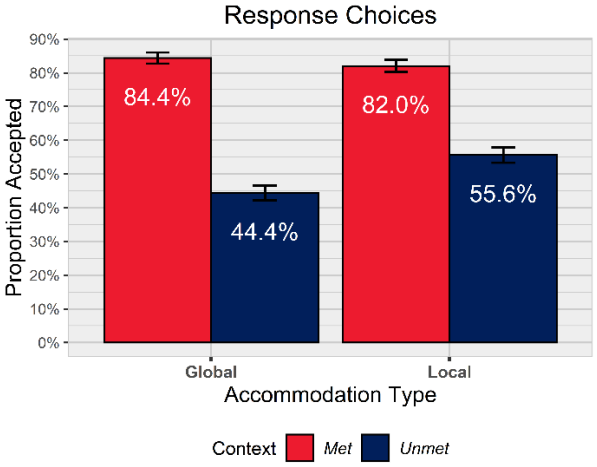
(1a) Global: A: Linda loves traveling,
{and last year she went to Vietnam. (*PSP met*)
but I don't know whether she's been to Vietnam before. (*PSP unmet*)}
B: She went to Vietnam **again** this year,
so she probably picked up some Vietnamese already.

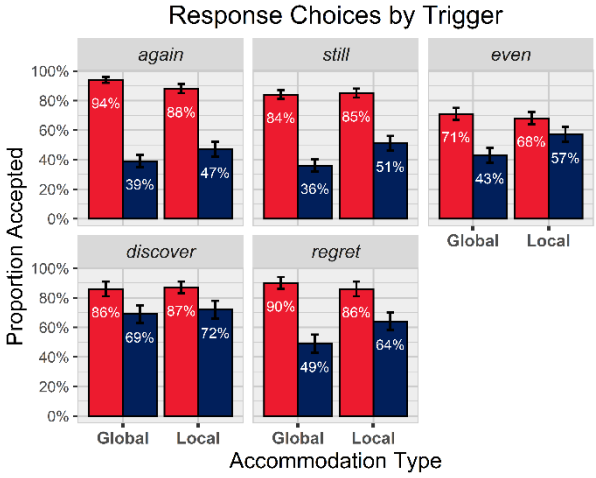
(1b) Local: A: Linda loves traveling.
B: {Yeah - last year she went to Vietnam. (*PSP met*)
Yeah - though I don't know whether she's been to Vietnam before. (*PSP unmet*)}
If she went to Vietnam **again** this year,
then she probably picked up some Vietnamese already.

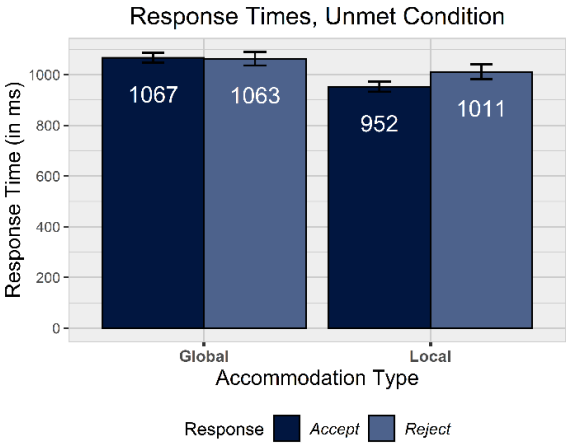
Method. Each trial began with a button displayed at the center bottom of the screen and large thumbs-up and -down icons at the top left and right respectively. Button click started a character-by-character unfolding of the text (at 60ms/char). 500ms before the end of the target clause, participants were prompted to quickly indicate acceptability of the discourse so far by moving their cursor to one of the icons. The initial choice had to happen within 2 seconds. (Error messages were displayed if the cursor was moved too early or did not reach an icon within the time limit; the setup aimed to also provide insights from mouse tracking data, but these are inconclusive so far.) Upon selection, the final clause unfolded, and participants could adjust their up/down choice.

Results. *Ratings*. Final acceptance rates by condition are shown below. A mixed effects logistic regression model showed a significant decrease in acceptability for *unmet* conditions (β=-1.86, $p<.001$), as expected. This effect was more pronounced in the global condition, as reflected in a significant interaction (β=.75, $p<.01$) (and corresponding simple effect in the unmet condition

(β=.58, p<.01),). Similar patterns were found across individual triggers, with some variation in the extent of the unacceptability rates, e.g., for the cognitive factive *discover*, as shown on the right.







<u>*Response Times*</u>. RTs were calculated from the prompt to respond during the unfolding of the target sentence to initial mouse selection of up- or down-icon. A 2x2 linear mixed effects model across all conditions found significantly faster RTs for *local* (β=-119, p<.001) but no other effects. Focusing on the *unmet* conditions, where accommodation is at play, we ran a second model predicting RT from the interaction of Global/Local and RESPONSE CHOICE and found a significant interaction (β=.89, p<.05), with faster acceptance choices for *local* than *global* (β=-135, p<.001) (but no simple effects of CONTEXT within responses).

Discussion. Counter to claims that LA is a last resort mechanism that's only marginally available (if at all, for certain triggers), we find it to be readily available, just like GA - in fact, it is more acceptable than GA overall. Whether or not this difference speaks against a unified mechanism remains somewhat open. To the extent that LA and GA generally rise and fall together across triggers (with the exception of *even-lex*), this can be seen as supporting a shared mechanism, as long as the LA advantage can be independently accounted for (e.g., due to particular properties of our task and stimuli). Some of the trigger differences align with Klinedinst (2016)'s hypothesis, showing comparable LA and GA costs for *discover* but larger cost for GA than LA for *regret* (cf. Djärv et al. (2017)'s account of cognitive factives as entailing their PSP). Trigger variation clearly requires further scrutiny for a fuller picture of the accommodation mechanism(s) at play.

Our RT findings are surprising as well, in that there was no processing cost for either type of accommodation. LA Accept responses being faster than GA ones provides a novel comparison across accommodation types, that aligns with the acceptability pattern. Moreover, the fast LA RTs contrast with prior studies reporting slowdowns in RTs for LA (Chemla & Bott 2013; Romoli & Schwarz 2015), though these involved slightly different comparisons. But most importantly, our paradigm provided explicit contextual support for LA, whereas prior work offered the choice of an LA interpretation of a sentence in isolation. Prior claims that LA is hard to access may thus have to be reevaluated to take into account the role of contextual motivation, leaving more direct comparisons of relevant manipulations of contextual support as an important direction for future work.

Effects of instruction on semantic and pragmatic judgment tasks

Ziling Zhu & Dorothy Ahn, Rutgers University

Background. Experimental linguistic work is defined by its design, procedures, and statistical analysis (Kirk, 2012; Myers, 2017). There have recently been more discussions on how to optimize procedures for sentence judgment tasks, featuring two considerations: instruction (Schutze, 2005; a.o.) and response scale (Schutze & Sprouse, 2013; a.o.). Instruction variation was claimed to be trivial for morphological (Aronoff & Schvaneveldt, 1978), syntactic (Schutze & Sprouse, 2013), and pragmatic (Veenstra & Katsos, 2018) judgment tasks. This study fills the research gap for experimental semantics and pragmatics, revealing that instruction is a significant factor in identifying and distinguishing between semantic and pragmatic violations in sentence judgment tasks. Furthermore, we show that English and Mandarin speakers respond differently to different keywords in the instructions, highlighting the need for language and study-specific norming procedures.

Methods. To investigate the effects of instruction in sentence judgment tasks, we compared participants' responses to four commonly used instructions shown in (1) against the same set of sentence stimuli. A total of 24 syntactically well-formed sentences were tested in the stimuli, and we grouped them into three categories based on their semantic and pragmatic felicitousness: (i) 8 `semantically odd` (logical contradiction and thematic mismatch), (ii) 8 `pragmatically odd` (redundant information), and (iii) 8 `neutral`. An example of each sentence type is shown in (2).

(1) a. Does this sound natural to you?
b. Does this sound acceptable to you?
c. Does this sound grammatical to you?
d. How likely is it for a native speaker to say this?

(2) a. Jake is a married bachelor. (`semantically odd`)
b. Yuki arrived. Yuki sat down. Yuki turned on her laptop. (`pragmatically odd`)
c. Mason thinks it's raining outside. (`neutral`)

In order to test for language-specific effects, we also created a Mandarin version of the English study with the instructions as in (3).

(3) a. yixia neirong ting-qilai ziran ma?
following contents hear-impression natural Q-PART?
'Do the following contents sound natural?'
b. yixia neirong ting-qilai fuhe yufa ma?
following contents hear-impression fit grammar Q-PART?
'Do the following contents sound grammatical?'
c. yixia neirong ting-qilai ke jieshou ma?
following contents hear-impression can accept Q-PART?
'Do the following contents sound acceptable?'
d. nin renwei muyu wei hanyu de ren, you duo-da keneng
you think native.language be Mandarin GEN person, have how-big possibility
shuo-chu yixia neirong?
say-out following contents?
'How likely do you think is it for a native speaker of Mandarin to say the following contents?'

We used a between-subject study so that each participant would only see one question type for all 24 test items. Participants were asked to respond on a 7-point Likert scale.

Eighty-one native English speakers and 81 native Mandarin speakers (18-64; gender-balanced) were recruited via Prolific. They were asked to provide some demographic and language background information, and then were presented with the 24 sentence stimuli (randomized in order).

Predictions. If instruction variation is trivial for semantic and pragmatic judgment tasks, we would

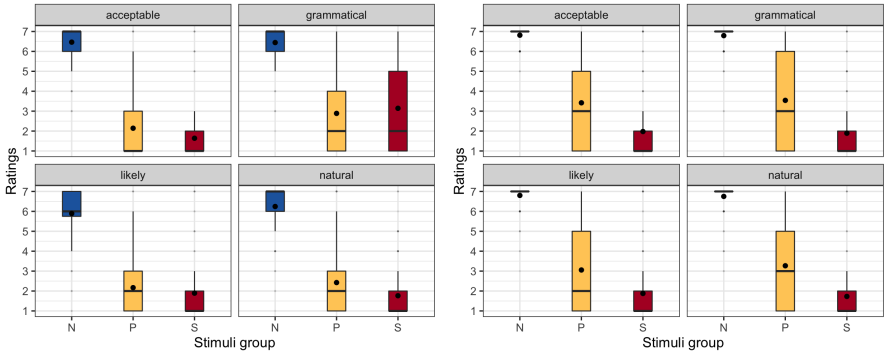


Figure 1:
Ratings as function of stimuli group, grouped by instruction type
N: `neutral`
P: `pragmatically odd`
S: `semantically odd`
English(L), Mandarin(R)

predict that instruction type would not change the rating results for each test sentence. Otherwise, different instructions would lead to different ratings of the same stimuli.

Results. We fit a Cumulative Link Mixed Model in R to compare ratings in different conditions (Fig. 1). For English, the results showed a main effect of stimuli group ($p < 0.001$), instruction type ($p < 0.001$), and significant interaction ($p < 0.001$). For Mandarin, we only found a main effect of stimuli group ($p < 0.001$), and not instruction type ($p > 0.1$), and no significant interaction ($p > 0.1$).

Across the stimuli groups, all instruction types reliably distinguished between `odd` and `neutral` stimuli ($p < 0.001$) for both English and Mandarin. Between `semantically` and `pragmatically odd` sentences, for English, all instruction types led to significantly different responses except for `grammatical` ($p > 0.1$); for Mandarin, all instruction types led to significantly different responses ($p < 0.001$). Moreover, the instruction type `natural` was the most effective in teasing apart the stimuli groups for both languages.

Discussion: Our experiment reveals the significance of instruction type in semantic and pragmatic sentence judgment tasks. First, we confirm the intuitive choice, made by previous researchers, of using 'natural' in the instruction design (Cremers & Chemla, 2017; Zlogar & Davidson, 2018; Hara et al., 2014; a.o.). Second, we highlight the need to include control sentences with standard ratings to evaluate semantic and pragmatic violations more accurately. For instance, Sprouse et al. (2020) use a set of previously-tested sentences as fillers to calibrate newly collected grammaticality judgments in their syntax study. Our preliminary data can serve a similar role in semantic and pragmatic judgment tasks.

The current study also draws attention to cross-linguistic differences in sentence judgment tasks. While `natural` is the best keyword to distinguish between semantic and pragmatic oddness for both languages, the range of responses spreads wider in Mandarin than in English. Hence, language-specific norming studies with control sentences are crucial in order to effectively compare cross-linguistic judgments.

More generally, our study speaks to the general concern on the validity of sentence judgment tasks used for semantic and pragmatic research. The grouping of the stimuli into pragmatically odd, semantically odd, and neutral sentences is not independently motivated and thus potentially theory-internal. However, our results suggest that the paradigm of sentence judgment tasks can identify at least some distinction between logically illicit sentences (`semantically odd`) and sentences that are logical but not discourse-natural (`pragmatically odd`).

Aronoff, M., & Schvaneveldt, R. 1978. Testing morphological productivity. *Annals of the New York Academy of Sciences*, 318(1). **Cremers, A., & Chemla, E. 2017.** Experiments on the acceptability and possible readings of questions embedded under emotive-factives. *Natural Language Semantics*, 25(3). **Hara, Y., Kawahara, S., & Feng, Y. 2014.** The prosody of enhanced bias in Mandarin and Japanese negative questions. *Lingua*, 150. **Kirk, R. 2012.** *Experimental design: Procedures for the behavioral sciences.* **Myers, J. 2017.** Acceptability judgments. *Oxford Research Encyclopedia of Linguistics.* **Schütze, C. T. 2008.** Thinking about what we are asking speakers to do. *Linguistic Evidence.* **Schütze, C. T., & Sprouse, J. 2013.** Judgment data. *Research methods in linguistics.* **Sprouse, J., Messick, T., & Bobaljik, J. 2020.** Gender asymmetries in ellipsis: An experimental comparison of markedness and frequency accounts in English. *Journal of Linguistics.* **Veenstra, A., & Katsos, N. 2018.** Assessing the comprehension of pragmatic language: Sentence judgment tasks. *Methods in Pragmatics.* **Zlogar, C., & Davidson, K. 2018.** Effects of linguistic context on the acceptability of co-speech gestures. *Glossa.*

To parse or not to parse: symmetric filtering in negated conjunctions

Alexandros Kalomoiros & Florian Schwarz
University of Pennsylvania

Intro: We present experimental evidence for symmetric filtering of presuppositions in conjunctions inside conditionals (typically predicted to be asymmetric) in case the presuppositional conjunct is negated. Such a pattern is predicted by a parsing-based approach like *Limited Symmetry* (Kalomoiros 2021), but not by traditional accounts that are constituent-based.

Constituent Approaches: A key characteristic of standard approaches to presupposition projection is that projection is calculated recursively on the constituent structure of a sentence. For instance, on dynamic accounts (Heim 1983 a.o.), the rule for a conjunction $(p\ and\ q)$ is to 'update the context C with p, $C + p$, and then update the result with q, $(C + p) + q$'. This rule updates constituent-by-constituent, requiring each context to be updated to entail any presuppositions of the constituent that is under update. Such constituent-based mechanism can be combined with an order constraint (requiring update to proceed from left to right), resulting in asymmetric filtering across connectives; alternatively, update can be unordered, allowing symmetric filtering across the board (Schlenker 2009, Rothschild 2012). Thus, either all filtering is asymmetric, with symmetry perhaps being available at a cost, or all filtering is symmetric by default. However, recent experimental work points to the conclusion that conjunctions are categorically asymmetric (Mandelkern et al 2020), but disjunctions are symmetric with regard to projection (Kalomoiros & Schwarz 2021).

Limited Symmetry: A system that derives symmetry for disjunction but asymmetry for conjunction through a single mechanism is *Limited Symmetry* (Kalomoiros 2021). This is a parsing-based system which makes distinct predictions from constituent-based systems. Consider a language $\mathcal{L}$:

(1) $\phi := p_i \mid p'_j p_k \mid (not\ \phi) \mid (\phi\ and\ \phi) \mid (\phi\ or\ \phi) \mid (if\ \phi.\ \phi)$ ($i, j, k \in \mathbb{N}$; indices omitted below)

$p'p$ represents a statement that presupposes p' and asserts p; it is interpreted as conjunction: $w \models p'p$ iff $w \models p'$ and $w \models p$. There are two core ideas: **i)** sentences are parsed from left to right, symbol by symbol, against a context C. Hence $(p'p\ and\ q)$ is associated with a parsing list `[(,(p'p,(p'p and, (p'p and q, (p'p and q)]`. Note how this gives us access to non-constituent elements like `(p'p and`. At every parsing point t_i on this list, the parser attempts to compute the sets of worlds where the sentence is True or False for every possible continuation d ($\mathbb{T}$/$\mathbb{F}$). **ii)** We assume that for every $\mathcal{L}$-sentence S we have access to a [-presup] version of S, where all the primed bits have been removed; e.g. [-presup]($p'p$) = p. If at a parsing point t, $\mathbb{T}$/$\mathbb{F}$ can be computed, then the following presupposition constraint must be respected: Given a sentence S and parsing point t_i, all the worlds in $\mathbb{T}$/$\mathbb{F}$ at t_i must be worlds in the $\mathbb{T}$/$\mathbb{F}$ computed at the corresponding parsing point t'_i for [-presup](S). If this fails, it leads to infelicity. **This constraint is a subsethood condition amounting to the standard condition requiring presuppositions not to introduce new info**.

Negated Conjunction: Consider now a sentence of the form $(if\ ((not\ p'p)\ and\ q).\ r)$, where $q \models p'$. Assume a material implication semantics for conditionals. On a constituent-based approach that proceeds from left to right, presuppositions project from the negation, a first conjunct and the antecedent of a conditional, so such approaches predict projection, requiring the global context to entail the presupposition p'. Applying *Limited Symmetry*, we reason as follows: No $\mathbb{F}$ set of worlds can be computed before we have parsed the whole conditional. But at parsing point `(if (not p'p) and`, we already know that the entire conditional is True in all worlds $\mathbb{T} = \{w | p'(w) = 1\ and\ p(w) = 1\}$. The corresponding parsing point for the [-presup] version of this sentence is `(if(not p) and`. At this parsing point, $\mathbb{T} = \{w | p(w) = 1\}$. Thus, $\mathbb{T}_{[+presup]} \subseteq \mathbb{T}_{[-presup]}$, so the presupposition constraint is respected. The parse moves on. At parsing point `(if((not p'p) and q`, $\mathbb{T}_{[+presup]} = \{w | (p'(w) = 1\ and\ p(w) = 1)\ or\ q(w) = 0\} \subseteq \mathbb{T}_{[-presup]} = \{w | p(w) = 1\ or\ q(w) = 0\}$.

Again the presupposition constraint is respected. We omit the computation for the parsing step where the entire sentence is parsed (it's lengthy), but no violations of our constraint turn up. So, this is a case where *Limited Symmetry* predicts filtering of a presupposition, whereas mainstream approaches predict projection. Crucially, once the negation is gone (i.e. $(if\ (p'p\ and\ q.\ r))$, *Limited Symmetry* predicts projection (same as constituent approaches). To tease the two approaches apart, we designed an experiment contrasting $(if\ ((not\ p'p)\ and\ q).\ r)$ and $(if\ (p'p\ and\ q.\ r)$.

Design: We selected 6 triggers (*again, stop, continue, find out, happy, aware*), which we presented in the following two conditions: **i)** A negated conjunction inside the scope of a conditional (NegConj); **ii)** A non-negated conjunction inside the scope of a conditional (SimpleConj). Both of these conditions were presented in Support (S) and Explicit Ignorance (EI) contexts. Overall then, there were four conditions: {EI/S}NegConj, {EI/S}SimpleConj:

(2) **Contexts:** Sue likes to keep close tabs on her husband, Donald. One day I saw a ticket from the Berlin opera in Donald's office …
…I don't know whether Donald ever visited Germany, so I thought: (EI)
…I know that he visited Germany recently, so I thought: (S)

(3) If Sue didn't find out that Donald visited Germany and he visited Berlin, then that would be very strange. (NegConj)
If Sue found out that Donald visited Germany and he visited Berlin, then she must know about the opera ticket. (SimpleConj)

Predictions: *Limited Symmetry* predicts that negated conjunction conditionals should be equally felicitous in S and EI contexts, since the presupposition is supported by the context in the former case, and filtered in the latter case. Simple conjunctions should be less felicitous in an EI context than in a S context, since the projecting presupposition clashes with the EI context. Overall, an interaction is predicted: the difference in acceptability between EI vs S contexts (Context type) should be greater for SimpleConj compared to NegConj (Conjunction type). No such interaction is predicted by mainstream approaches: EISimpleConj and EINegConj should be equally degraded.

Participants & Procedure: 163 participants (all native English speakers) were recruited from our university's subject pool. Each participant saw three items (from three distinct triggers) in each condition in a Latin square design. There were also 12 fillers (24 items in total, randomised). Participants had to indicate on a 9-point scale how felicitous a sentence was in the given context.

Results: Our results are strikingly in line with the *Limited Symmetry* predictions (Fig 1). We tested for the relevant differences by fitting linear mixed-effects regressions. First, there is a significant difference between EISimpleConj and SSimpleConj ($p < 0.05$). At the same time, there is no significant difference between EINegConj and SNegConj. This leads to a significant interaction between Conjunction type and Context type: the difference in acceptability between EI and S contexts is significantly larger ($p < 0.05$) for SimpleConj.

Discussion: These results run counter to predictions of standard theories of projection. But a potential worry is that the felicity of EINegConj is due to special availability of a mechanism like local accommodation under negation. To control for this, we re-run the experiment, adding two local accommodation conditions (LocAcc): a conditional containing a negated presupposition in the antecedent, in an EI context vs an *S* context. Preliminary results (N = 172) show that the felicity difference is larger for LocAcc than for NegConj. This suggests that the felicity increase in EINegConj is not due to a local accommodation-like mechanism. Nevertheless, the re-run of the experiment replicates the interaction for Conjunction

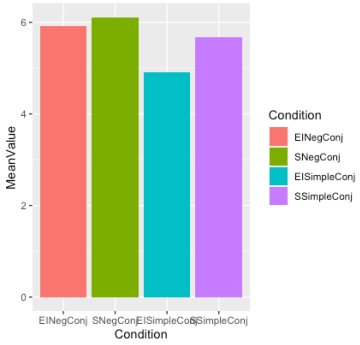


Figure 1: Mean acceptability

type vs Context type only for a subclass of triggers (more details in the presentation). This creates further questions, but reinforces the idea that at least some triggers behave as *Limited Symmetry* predicts.

Selected Ref: Deriving the (a)-symmetries of presupposition projection. Forthcoming in NELS 52.

Corpus evidence for the role of world knowledge in ambiguity reduction: Using high positive expectations to inform quantifier scope

Noa Attali, Lisa S. Pearl, and Gregory Scontras
Department of Language Science, UC Irvine

Investigations into interpretations of quantifier-negation utterances (e.g., *Every vote doesn't count*, which is ambiguous between *No vote counts* and *Not all votes count*) have found variation: child and adult interpretations of *every*-negation diverge (e.g., Musolino, 1999), adult interpretations of utterances with different quantifiers vary (e.g., *every*- vs. *some*- vs. *no*-negation; Attali et al., 2021), and even adult interpretations of different *all*-negation constructions alone (Carden, 1973) and in context (Heringer, 1970) show considerable disagreement. Can we concretely identify factors to explain some of this variation and predict tendencies in individual interpretations? Here, we show that a type of expectation about the world, which can surface in the linguistic contexts of *every*-negation utterances that we find in a speech corpus, predicts experimental preferences for the *not all* interpretation of different *every*-negation utterances. These findings suggest that world knowledge, as set up in a linguistic context, helps to effectively reduce the ambiguity of potentially-ambiguous utterances for listeners.

High positive expectations. In their computational cognitive model of this ambiguity, Scontras and Pearl (2021) demonstrate that a kind of world knowledge we term a "high positive expectation" (**hpe**) can explain some variation in behavior with *every*-negation utterances. For example, in *Every vote doesn't count*, an hpe is the prior belief that it's highly likely that every vote *does* count – that is, that the worlds consistent with the non-negated utterance (*Every vote does count*) are likely. This world knowledge quantitatively specifies a pragmatic factor in previous proposals meant to capture truth value judgment results (e.g., Musolino and Lidz, 2006; Gualmini, 2004).

In particular, an hpe could contribute to the felicity of using *every*-negation with a *not all* interpretation, thereby reducing the ambiguity of the utterance for listeners. For speakers, Scontras and Pearl's model predicts that they tend to endorse *every*-negation as a true description of a scenario consistent with the *not all* interpretation when *every*-negation conveys that an hpe is false (e.g., that some votes are, in fact, not counted). For listeners, Attali et al. (2021) find that *not all* interpretations are preferred on average for *every*-negation, and that when Scontras and Pearl's model is applied to predict these listener interpretations, it does so successfully if given an hpe: it accounts well for the qualitative and quantitative pattern of average cross-speaker interpretation preferences for the experimental, out-of-context sentences *Every/Some/No marble isn't red*. The modeled pragmatic listener prefers the *not all* over the *none* interpretation, when given an hpe, because the listener assumes a cooperative, efficient speaker. A cooperative speaker wants to say something true, and there are more ways for the *not all* interpretation to be true compared with the *none* interpretation; an efficient speaker wants to be informative, and it's highly informative to update a strongly biased, salient belief (e.g., that every vote does count; see Attali et al., 2021).

So here, we ask to what extent an hpe accounts for interpretations of different *every*-negation utterances in naturalistic contexts. As a case study, when a local linguistic context seems to express an hpe, is *not all* a more likely interpretation than *none*?

Corpus data and behavioral experiment. We identified 390 uses of *every*-negation in the radio and TV transcripts (1990-2012; $\approx$9 million clauses) in the Corpus of Contemporary American English (Davies, 2015). Following Degen (2015), we crowd-sourced interpretation preferences of these uses in their immediate contexts (three preceding sentences and one following sentence). For each item, participants (N = 208) completed a paraphrase-endorsement task (Scontras and Goodman, 2017), choosing on a sliding scale between *none* and *not all* paraphrases of the

potentially-ambiguous clause. In line with previous findings, we found a preference for *not all* interpretations and a high degree of variation (see Fig. 1).

Identifying high positive expectations in linguistic contexts. As a preliminary measure, the first author hand-coded categorically for the presence/absence of an overt hpe expression in each preceding context (finding that 59/390 (15%) had an hpe). For an automatic and principled measure of the expression of an hpe in the linguistic context, we calculated the degree of lexical overlap between the preceding linguistic context and a string representing the positive expectation (hpe). That is, for each item (e.g., *Every vote doesn't count*), we first coded hpe as the potentially-ambiguous utterance without negation (e.g., *Every vote does count*). We then coded for the extent to which the hpe appeared in the preceding context as the longest common substring (**LCS**) similarity (Needleman and Wunsch, 1970), calculated using the R stringdist package (van der Loo, 2014). Each LCS was equal to the longest sequence formed by pairing words from the preceding context string and hpe, while keeping their order intact; the dissimilarity d_{lcs} was then the number of unpaired words left over in both strings, and LCS similarity was $-d_{lcs}$. Thus LCS similarity ranges from 0 to $-\infty$, with higher values indicating a greater probability that the context contained a high positive expectation. For example, if the preceding context was *Every vote does count* for an utterance with the hpe *Every vote does count*, LCS similarity would equal 0. On the other hand, if the preceding context was *What is going on?*, LCS similarity would equal -8 (since all eight words in the two strings would be unpaired). The disadvantage of LCS similarity is its noisy potential to underestimate the presence of an hpe (e.g., it would discount the context *All votes should matter*); but it provides an automatic continuous measure. (Other lexical overlap implementations yield similar results.)

Results. Using the preliminary categorical hand-coding, we found that 50/59 (85%) of the utterances with hpes were on average better paraphrased by *not all* than *none*. Using the continuous and automatic LCS measure to assess if an hpe predicts a *not all* preference per item, we ran a linear mixed effects model predicting logit-transformed mean item responses by LCS similarity, with random intercepts for participants (see Fig. 2). To determine whether an hpe captures individual judgment variation, above and beyond mean item-level variation, we predicted logit-transformed item responses by LCS similarity, with random intercepts for participants and items. Both models found that LCS similarity was a significant predictor of a *not all* preference ($p < .001$ in both). Interestingly, a version of both models which calculated LCS similarity using overlap with the following – rather than preceding – context, found LCS similarity of the following context not to be a significant predictor of either item-level or judgment-level interpretations.

Conclusion. Our corpus analysis supports the plausibility of an hpe, expressed in the preceding linguistic context, playing a role in *not all* interpretation preferences for *every*-negation utterances. These results align with the previous modeling results and pragmatically-oriented proposals from truth value judgment studies. We note that we might underestimate the role of an hpe, because our automated method of identifying it is only one of many, and may be noisy; moreover, such an aspect of world knowledge could affect interpretations without necessarily receiving expression in the immediate discourse. In general, our findings support the theory that negation use is more felicitous in affirmative contexts (e.g., Wason, 1961), such as contexts containing an hpe.

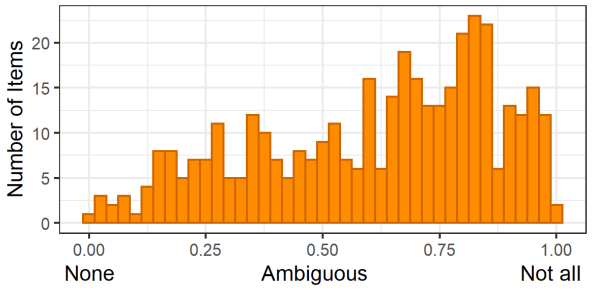


Figure 1: Histogram of average item interpretation from the corpus analysis.

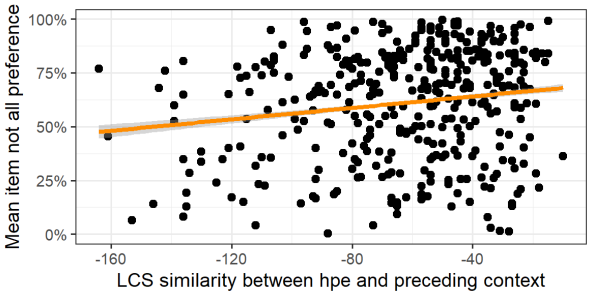


Figure 2: Preceding hpe and average *not all* item preference.

References

N. Attali, G. Scontras, and L. S. Pearl. Pragmatic factors can explain variation in interpretation preferences for quantifier-negation utterances: A computational approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

G. Carden. Disambiguation, favored readings, and variable rules. *New ways of analyzing variation in English*, pages 171–82, 1973.

M. Davies. Corpus of Contemporary American English (COCA). 2015. URL `https://doi.org/10.7910/DVN/AMUDUW`.

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

A. Gualmini. Some knowledge children dont lack. *Linguistics*, 42(5):957–982, 2004.

J. T. Heringer. Research on quantifier-negative idiolects. In *Chicago Linguistic Society*, volume 6, page 95, 1970.

J. Musolino. Universal grammar and the acquisition of semantic knowledge: An experimental investigation into the acquisition of quantifier-negation interaction in english. 1999.

J. Musolino and J. Lidz. Why children aren't universally successful with quantification. *Linguistics*, 44(4):817–852, 2006.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

G. Scontras and N. D. Goodman. Resolving uncertainty in plural predication. *Cognition*, 168:294–311, 2017.

G. Scontras and L. S. Pearl. When pragmatics matters more for truth-value judgments: An investigation of quantifier scope ambiguity. *Glossa: a journal of general linguistics*, 6(1), 2021.

M. P. van der Loo. The stringdist Package for Approximate String Matching. *The R Journal*, 6(1):111–122, 2014. doi: 10.32614/RJ-2014-011. URL `https://doi.org/10.32614/RJ-2014-011`.

P. C. Wason. Response to affirmative and negative binary statements. *British Journal of Psychology*, 52(2):133–142, 1961.

Incremental theme verbs do not encode measures of change: experimental evidence from German-speaking adults

Merle Weicker, Lea Heßler-Reusch, Petra Schulz (Goethe University Frankfurt)

This study investigates German-speaking adults' interpretation of incremental theme verbs such as *drink*, probing the role of the incremental theme argument and of adjectives with an upper closed scale (e.g., *empty*) regarding telicity. Our findings provide experimental evidence for the claim that incremental theme verbs do not lexicalize measure of change functions (see Kennedy, 2012). Moreover, our results indicate that the telic interpretation of incremental theme verbs is construed semantically when combined with an upper closed scale adjective, whereas it is construed pragmatically when it is combined with a quantized NP (Filip, 2008).

In some events an entity changes as a result of participating in this event; these events can be described *i.a.* by degree achievement verbs morphologically related to gradable adjectives (1) (taken from Kennedy, 2012: 107) and by incremental theme verbs (2).

(1a) The sink emptied. (1b) The sink emptied completely. (1c) … but not entirely.

(2a) Maria drank the beer. (2b) Maria drank up.

In principle, (1a) and (2a) allow both telic and atelic readings, that is they can denote culminating or non-culminating events (Filip, 2008; Hay et al., 1999; Kennedy, 2012). For instance, continuation of (1a) and (2a) with (1c) is felicitous. Accordingly, it has been argued that the telic readings in (1a) and (2a) arise pragmatically via a generalized conversational implicature, which is cancellable (Filip, 2008). In contrast to (1a) and (2a), (1b) and (2b) only allow a telic reading. Continuation of (1b) and (2b) with (1c) is infelicitous (Hay et al., 1999; Schulz et al., 2001). These latter examples are instances of semantic telicity, i.e., the culmination point is entailed and this entailment is not cancellable. Following this line of reasoning, the degree modifier *completely* and the resultative particle *up* have been referred to as strong telicity markers (e.g., Schulz, 2018).

Kennedy (2012) analyzes both (1) and (2) as involving a measure of change function, i.e., a function that measures the degree to which an object changes as a result of its participation in an event. In degree achievements (ex. 1), this function is encoded by the verb. In the case of degree achievement verbs that are morphologically related to gradable adjectives, telicity properties are argued to be attributed to the scalar structure of the adjectival core (Deo et al., 2013; Hay et al., 2013; Kennedy/Levin, 2008; Winter, 2006). This adjectival base indicates the standard against which the change is measured, and if the scale has a maximal endpoint as with *empty*, the change should reach this endpoint. Accordingly, the default interpretation of the corresponding event is telic (e.g., Kennedy/Levin, 2008). In the case of incremental theme verbs (ex. 2), the measure of change function is not encoded by the verb but inherent to its argument (Kennedy, 2012). Telicity is related to the amount of incremental change required; if it is specified by a quantized NP (e.g., *the beer*, *ten cookies*), and its default interpretation is telic (Krifka, 1989).

Previous comprehension studies on incremental theme verbs found that adults treated resultative particles like *up* and quantized NPs differently, e.g., when added to verbs of consumption (see overview in van Hout 2018). While *eat/drink up* were restricted to telic interpretations, *eat the cheese/drink the tea* were allowed to denote events without event culmination in about half of the cases.

Building on these lines of theoretical and experimental research, our study investigates the scalar properties of adjectives in resultative constructions (ex. 3a) and contrasts them with ordinary incremental theme verb structures (ex. 3b).

(3a) Er trank den Saft leer. (3b) Er trank den Saft.

'He drank the juice empty.' 'He drank the juice.'

Adjectives with an upper closed scale such as *empty* have been argued to cause telicity via a homomorphism between the adjective's scale and the event (Wechsler, 2005), i.e., the endstate sub-event corresponds to the culmination point of the adjectival scale. Accordingly, adjectives with an upper closed scale should be strong telicity markers, as evidenced by the infelicitous continuation of (3a) with (1c). In contrast to a range of empirical studies of verb particles like *up*

(see van Hout, 2018) empirical research on resultative adjectives structures is scarce. The existing data suggests that adults are sensitive to the lexical restrictions of the participating verbs (Richter/van Hout, 2013), leaving open how adults interpret resultative adjective constructions.

Our study asked whether German-speaking adults (N = 21, mean age = 25 years) assign a semantically telic interpretation to sentences such as (3a) and a pragmatically telic interpretation to sentences such as (3b). A novel Truth Value Judgement task was developed with 4 conditions, varying event type (CULMINATING/NON-CULMINATING) and structure (ADJECTIVE/NO ADJECTIVE), with 4 items per condition. The incremental theme verbs *drink*, *wipe, blow-dry*, *iron* were combined with the adjectives *empty*, *clean*, *dry*, *flat*, respectively. All events were presented as animated paintings (see Fig. 1 for the stills) accompanied by prototypical sounds of the respective actions, e.g., slurping for (3a/b). In the CULMINATING condition, the boy drank the juice completely, in the NON-CULMINATING condition some juice was still in the glass after the drinking had stopped. After watching the video clips, participants answered yes/no questions (e.g., *Hat er den Saft (leer) getrunken?*, has he the juice (empty) drank). If the adjective causes the shift from pragmatic to semantic telicity, non-culminating events should be consistently rejected in the ADJECTIVE but not in the NO ADJECTIVE condition.

The mean number of yes-answers per condition is given in Table 1. We fitted a generalized mixed effects model (*lme4*, Bates et al., 2021) to participants' answers with event type and structure and their interaction as fixed effects. Participants and item were entered into the model as random effects. There was a main effect of event type (β = -6.3428, SE = 1.2260, z = -5.174, p < .001), indicating that non-culminating events were more often rejected than culminating events, and a main effect of structure (β = -2.2104, SE = 1.0714, z = -2.063, p < .05), showing that adjective structures were less often accepted than structures without the adjective. The significant interaction between event type and structure (β = -4.4208, SE = 2.1428, z = -2.063, p < .05) was further inspected via pairwise comparisons (*emmeans*, Lenth et al., 2021), revealing a significant difference between structures with and without adjective for non-culminating events (p = .002), but not for culminating events.

As expected, adults interpreted ordinary incremental theme verb structures as pragmatically telic, allowing the telicity implicature to be cancelled in 70% of the cases for non-culminating events. This result is in line with the analysis of quantized NPs as weak telicity markers. In contrast, resultative structures were interpreted as semantically telic, as evidenced by over 90% rejections of non-culminating events. This latter result provides first experimental evidence that upper closed scale adjectives like *empty* are strong telicity markers when combined with incremental theme verbs. Moreover, the interpretative contrast between structures with and without adjective supports Kennedy's (2012) theoretical analysis that incremental theme verbs do not introduce measure of change functions as part of their lexical meaning. With regard to resultatives, this finding suggests that two measure of change functions are available, provided by the adjective and by the quantized NP, but the adjectival one dominates the nominal one.

Table 1. Mean number of yes-answers per condition (max = 4).

Condition	Mean	SD	%
Culminating event, adjective	3.95	.21	98.8
Non-culminating event, adjective	0.38	.57	9.5
Culminating event, no adjective	3.95	.21	98.8
Non-culminating event, no adjective	2.81	.13	70.2

Fig.1. Example culminating event.

Selected references

Filip, H. (2008). Events and maximalization. **van Hout, A. (2018).** On the acquisition of event culmination. **Kennedy, C. (2012).** The composition of incremental change. **Kennedy, C./Levin, B. (2008).** Measure of Change: The Adjectival Core of Degree Achievements. **Wechsler, S. (2005).** Resultatives under the 'Event-Argument Homomorphism' Model of telicity.

Speaker reliability: calibrating confidence with evidence

Mélinda Pozzi & Diana Mazzarella

Cognitive Science Center, University of Neuchâtel, Switzerland

Overconfidence is typically damaging for one's reputation as a reliable source of information. Research in psychology shows that, when deciding whether to trust a speaker, addressees do not exclusively rely on the speaker's confidence ("confidence heuristics"), but consider, whenever possible, whether the speaker's degree of confidence matches with the accuracy of their claim. As a result, a confident speaker whose messages turn out to be false will typically lose their credibility (Tenney et al., 2007; 2008; 2011; Vullioud et al., 2017).

Crucially, though, preliminary findings from Tenney et al. (2008) indicate that an overconfident speaker does not suffer any reputational costs if their mistake is taken to be *justified*. This suggests that the speaker's perceived reliability as a source of information depends on whether their confidence matches with the quality of the evidence at their disposal ("confidence-evidence calibration"). If this is the case, then, even an accurate informant should lose their credibility if the evidence available to them does not warrant the degree of confidence expressed (bad confidence-evidence calibration).

The present study has two aims. First, replicating Tenney et al. (2008) results showing that overconfidence does not backfire if inaccuracy is justified by strong evidence: an inaccurate confident speaker who communicates false information that is justified by strong evidence does not lose their credibility (hypothesis 1 – experiment 1). Second, investigating whether confidence can backfire if accuracy is not justified by enough evidence (the speaker is accurate "by chance"): an accurate confident speaker who makes a claim that is not supported by enough evidence will lose their credibility (hypothesis 2 – experiment 2). Our study is pre-registered here: https://osf.io/fbv8g/?view_only=8d90bab9d82a43e1a7928e4de4aca7ef

We conducted two online experiments in which participants were presented with two testimonies concerning a car accident, judged the credibility of the two witnesses (one confident and one unconfident male witness), and were asked to choose which of the two depositions they believe. The material was adapted from Tenney et al. (2008, Experiment 2). In experiment 1, both witnesses were inaccurate but were justified by strong evidence. In experiment 2, both witnesses were accurate but had weak evidence. In both experiments, we measured participants' credibility judgments (on a scale from 1 to 6) and believability choices (*Who do you believe?*) at three distinct times: (1) participants have no information about accuracy and strength of evidence, (2) participants get feedback about accuracy, (3) participants get feedback about evidence.

The first experiment (N = 108) replicated Tenney et al. (2008) results. There was a significant interaction effect of confidence and time on credibility ($F(2, 321) = 93.018$, $p < .001$), and believability changed over time as predicted ($\chi^2(2) = 9.663$, $p = 0.008$). In the absence of any information about accuracy and strength of evidence (Time 1), the confident witness was rated as more credible and was more likely to be believed than the unconfident witness. At Time 2, when both witnesses turned out to be inaccurate, the confident witness lost his credibility to the benefit of the unconfident witness. At Time 3, when the inaccuracy was found to be justified by strong evidence, the confident witness' credibility was restored. The second experiment (N = 109) supported our second hypothesis. There was a significant interaction effect of confidence and time on credibility ($F(2, 324) = 35.115$, $p < .001$), and believability changed over time as predicted ($\chi^2(2) = 45.942$, $p < 0.001$). In the absence of any information about accuracy and strength of evidence (Time 1), the confident witness was rated as more credible and was more likely to be believed than the unconfident witness. At Time 2, when both witnesses turned out to be accurate, the confident witness kept his credibility. At Time 3, when the testimony of the witnesses was found to be warranted by weak evidence, the confident witness suffered a reputational loss.

This study shows that reputation management in communication depends on how well the speaker's confidence is calibrated to her evidential basis.

References

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science, 18*(1), 46–50. http://dx.doi.org/10.1111/j.1467-9280.2007.01847.x

Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology, 44*(5), 1368–1375. http://dx.doi.org/10.1016/j.jesp.2008.04.006

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental psychology, 47*(4), 1065–1077. http://dx.doi.org/10.1037/a0023273

Vullioud, C., Clément, F., Scott-Phillips, T., & Mercier, H. (2017). Confidence as an expression of commitment: Why misplaced expressions of confidence backfire. *Evolution and Human Behavior, 38*(1), 9–17. http://dx.doi.org/10.1016/j.evolhumbehav.2016.06.002

Figures

Experiment 1 - hypothesis 1

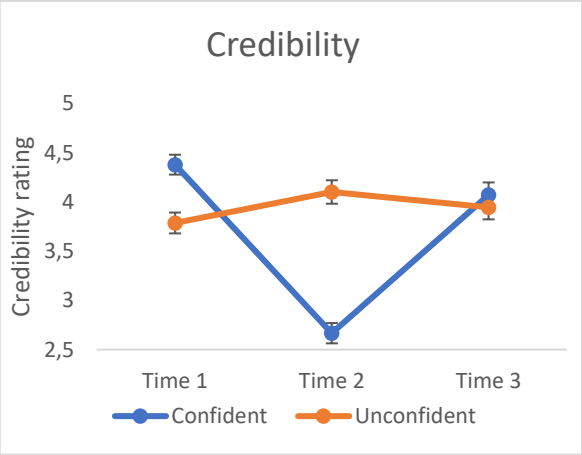


Figure 1. Credibility scores of the confident (blue) and unconfident (orange) witness, on a scale from 1 "not credible" to 6 "credible".

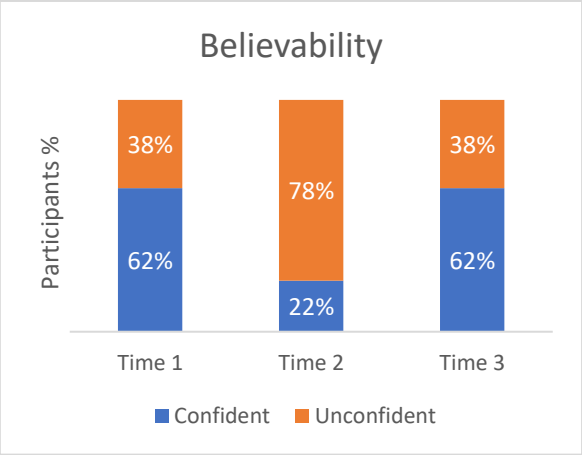


Figure 2. Percentage of participants who believed the confident (blue) or unconfident (orange) witness.

Experiment 2 – hypothesis 2

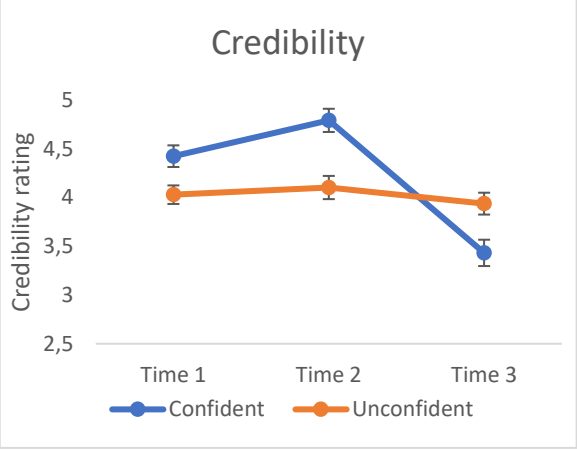


Figure 3. Credibility scores of the confident (blue) and unconfident (orange) witness, on a scale from 1 "not credible" to 6 "credible".

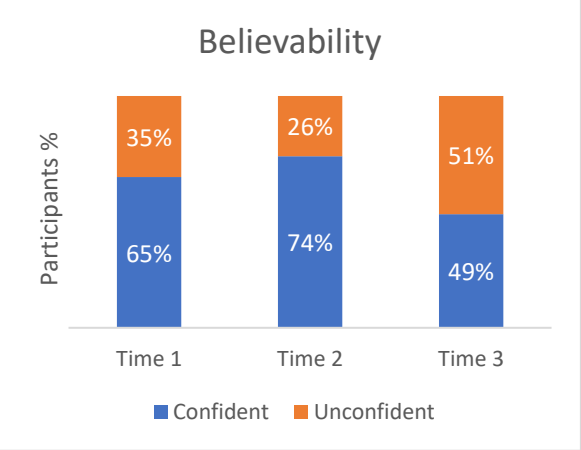


Figure 4. Percentage of participants who believed the confident (blue) or unconfident (orange) witness.

Task effects on the processing of predicate ambiguity: Distributivity in the Maze

John Duff (jduff@ucsc.edu), Adrian Brasoveanu & Amanda Rysling (UC Santa Cruz)

Some behavior in sentence processing experiments can be modulated by features of the task [1, e.g.]. But these interactions have limitations: for instance, question difficulty modulates re-reading, but not first-pass reading [2]. Here we examine the limitations of a recently-documented task effect in comprehension, **early semantic commitments in the Maze task** [3].

In (A-)Maze tasks [4, 5], participants proceed through a sentence by choosing between correct continuations and high-surprisal distractors. Incorrect choices result in early termination of the sentence, and optimal performance requires incremental comprehension. [3] used the Maze to investigate the comprehension of polysemes, lexical items with multiple related senses—e.g., *newspaper* as printed object or organization. While polysemes usually remain underspecified until the end of a sentence, [6, *i.m.a.*], [3] found that in the Maze, participants commit to a sense early, and face reanalysis costs for later revision. They concluded that, faced with a task where specification is strategic, participants commit to meanings earlier.

How early do Maze participants commit? We use the task here to examine an ambiguity that can be anticipated in sentences with plural subjects. In (1), *taught two classes* can admit either a **collective** reading true for the subject as a whole, or a **distributive** reading true of each member of the subject [7]. Adverbs *together* and *each* can disambiguate, respectively. Dominant theoretical accounts [e.g. 8] suggest distributivity involves additional implicit structure. [9] and [10] observe that in reading, late (post-predicate) *each* is associated with a slowdown on following words. They conclude that predicates with plural subjects receive a default collective reading at the predicate, such that late *each* triggers costly reanalysis.

Given that verbal meaning is predicted online from features of preceding arguments [11-12], and plural subjects regularly introduce collective/distributive ambiguities, collective readings could in principle be decided upon before the predicate itself, at the subject. But neither eye-tracking [9] nor SPR [10] find evidence of reanalysis when *each* appears before the predicate.

We might predict a different pattern in the Maze. A **powerful task effects** hypothesis, where strategic demands can motivate even anticipatory commitments, could predict reanalysis costs for *each* and following words even when it occurs early. On the other hand, a **restricted task effects** hypothesis, where anticipatory commitments are impossible despite strategic value, would predict reanalysis only for post-predicate *each*, as observed in other tasks.

Methods Prolific and (ongoing) student participants read 32 critical items (1) based on those used in [9], crossing Position (EARLY/LATE) and Meaning (TOGETHER/EACH) of a critical adverb. We analyze residualized log response latencies summed over two regions of interest, the predicate and a three-word spillover. LME models will be fit over complete samples in `brms`, taking a positive Meaning x Position interaction as critical evidence for **restricted task effects**.

Our **SPR** (n = 40 of 48) results are so far visually consistent with previous findings, suggesting we will find an interaction in at least the spillover region such that LATE *each* prompts reanalysis in particular. LATE conditions are also numerically read faster than EARLY across the board, as noted by [9]. This provides a baseline for evaluation of our Maze results.

Our **Maze** (n = 23 of 48) results are less clear at present: LATE conditions again appear faster in general. We see no evidence of an interaction, but surprisingly, *each* seems to be associated with faster reading in the spillover, contrary to predicted reanalysis costs. 81% of trials were completed without errors, and at the moment we see no notable relationships between particular conditions and the error rate in any region of interest.

Discussion This study contributes towards a broader understanding of task effects on the resolution of different types of semantic ambiguity. In particular, we hope to resolve whether task pressures of the Maze can induce anticipatory commitments to the structure of verbal meaning. If the patterns in the current sample are borne out in full, they could suggest that task-based

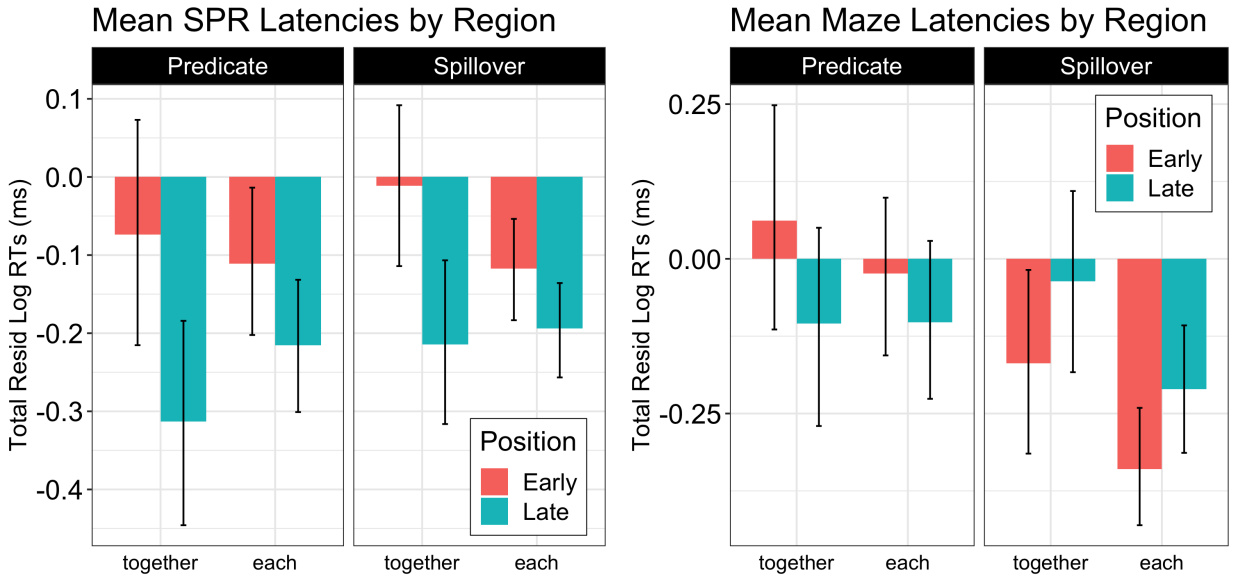
early commitment is indeed strikingly **powerful**: comprehenders in the Maze are not only more likely to commit to lexical meanings upon encountering a word, but they will even resolve predictable ambiguities in advance. On the other hand, if further data reveals the predicted interaction holds in both SPR and the Maze, this task effect, like others, could be shown to be **restricted** in its ability to impact linguistic processing.

What remains puzzling is the direction of comprehenders' apparent online bias in the Maze. In the current sample, Maze participants seem to default to a distributive interpretation. While certain predicates have been shown to bear a distributive bias that can surface online [10], norming of our items (n = 36) reveals a stable offline preference for collective readings. Should the online distributive bias persist, it would seem to be somehow related to performance in the Maze task, an unexpected possibility that would merit additional investigation.

(1) Reportedly, Burt and Sierra…

{ EARLY: together / each } taught two classes { LATE: together / each } | when the school | opened.

— — — — — *predicate* — — — — — — *spillover* —



FIGURES 1 & 2: Partial results from the self-paced reading and Maze tasks. Error bars represent bootstrapped 95% CIs. Log RTs were residualized by length and position of words with a random intercept for subjects.

REFERENCES **[1]** Hammerly et al. (2019) *Cog Psych* 110. **[2]** Weiss et al. (2018) *QJEP* 71(1). **[3]** Duff et al. (2021) *CUNY* Short Talk. **[4]** Forster et al. (2009) *Behav Res Meth* 41(1). **[5]** Boyce et al. (2020) *JML* 111. **[6]** Frazier & Rayner (1990) *JML* 29. **[7]** Landman (1995) "Plurality" in *The Handbook of Contemporary Semantic Theory*, Blackwell. **[8]** Lasersohn (1995) *Plurality, Conjunction and Events*, Kluwer. **[9]** Frazier et al. (1999) *Cognition* 70. **[10]** Dotlačil & Brasoveanu (2021) *Glossa* 6(1). **[11]** Konieczny & Döring (2003) *Proc. of ICCS 4*. **[12]** Levy & Keller (2013) *JML* 68.

Great variability exists in the cardinal values of linguistic quantifiers like *many*, *few*, *several*, etc. For example, the sentence "There are *many* students in Erin's class" maps to a very different set of cardinal values than "Andy had *many* cups of coffee last week." This is a problem for semanticists, who seek to formally describe meaning, because there is no clear way to resolve this context-dependent variability. Following the work of [1], we develop and experimentally validate a Bayesian model of quantifier semantics which represents these quantifiers as cumulative density thresholds along a probability distribution of expected values. Considering the previous examples, Figures 1a and 1b illustrate these expected value distributions and the cardinal threshold values for the lower-bound of *many* used in these contexts. The semantics of *many* here are defined with respect to θ_{many}, a stable cumulative probability density shown as the area underneath these curves (Equation 1a). Our hypothesis is that this threshold remains stable across contexts, and differences between probability distributions over expected values introduce the contextual variation (Hypothesis 1). As the figures reveal, the shapes of the expected value curves shift the cardinal threshold value (x_{min}). For example, with the threshold $\theta_{many} = 0.4$, then the expected values imply that "many cups of coffee" is at least 8 cups, while "many students" is at least 17 students.

In addition to testing this preliminary hypothesis, we also examine whether or not the upper- and lower- bounds of *several* ($\theta_{several-max}$ and $\theta_{several-min}$) align respectively with θ_{many} (Hypothesis 2) and the upper-bound of *few* (θ_{few}) (Hypothesis 3). Experiment 1 empirically elicits values for the imputation of the expected value curve for sixteen different contexts, two of which are described above. Experiment 2 then probes participants' truth-conditional semantics regarding the quantified utterances in these contexts.

In order to compare the viability of these hypotheses, we perform three independent analyses, all using Bayesian hierarchical models to estimate these threshold values from the experimental data. Five models were fit to test three hypotheses. One model implements the null hypothesis, which rejects the claims of all three hypotheses and fits individual thresholds for each quantifier for each context (Model A). The remaining four models accept Hypothesis 1 and implement all possible combinations of acceptance or rejection of Hypotheses 2 and 3. Models were fit using rstan and technical specifications are given on page 3.

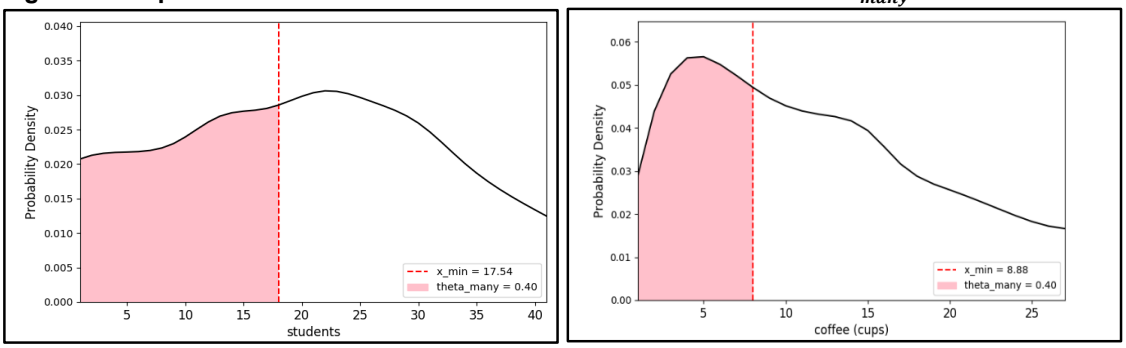
In our first analysis, we investigate what overlap exists between the context-independent thresholds implemented in Model A for a given threshold value. We find that some thresholds are more consistent than others and might represent more context-stability, namely θ_{many} (Figure 3) and $\theta_{several-min}.\theta_{few}$ also demonstrates an overlapping interval for most contexts, although not quite as many. This provides significant evidence for the existence of a context-stable threshold for these quantifiers. However, $\theta_{several-max}$ (Figure 4) is much less consistent than the other three. This leads us to believe that probabilistic threshold values for bounds on quantifiers might exist on a spectrum of context-dependence.

Our second analysis computed the WAIC Information Criterion [2] to perform model comparison. While Model A achieved the best performance, Model B did not perform significantly worse. This result, in combination with the findings of [1], leaves open the question of whether a context-stable threshold model is more appropriate than the null-hypothesis.

Our final analysis compares our context-stable thresholds against the combined threshold values $\theta_{several-few}$ and $\theta_{several-many}$. Figure 2 shows that these thresholds are all significantly different from one another, strong evidence against Hypotheses 2 and 3.

The results of our modeling and analysis suggest some degree of context-stability in cumulative density thresholds, albeit with varying degrees of stability by quantifier and by upper- versus lower-bound. We plan in future to investigate a larger, more diverse set of quantifiers to explore this spectrum in further detail. Our study does not however provide support for Hypotheses 2 and 3 as semantic phenomena. Further research will provide a pragmatic approach to the coincidence or lack thereof between thresholds of this nature.

Figure 1 – Expected Value Distributions and Context-Stable Threshold θ_{many}



(a) "There are many students in Erin's class." **(b) "Andy had many cups of coffee last week."**

Figure 2 – 95% Credible Intervals for Stable Thresholds in Models B-E

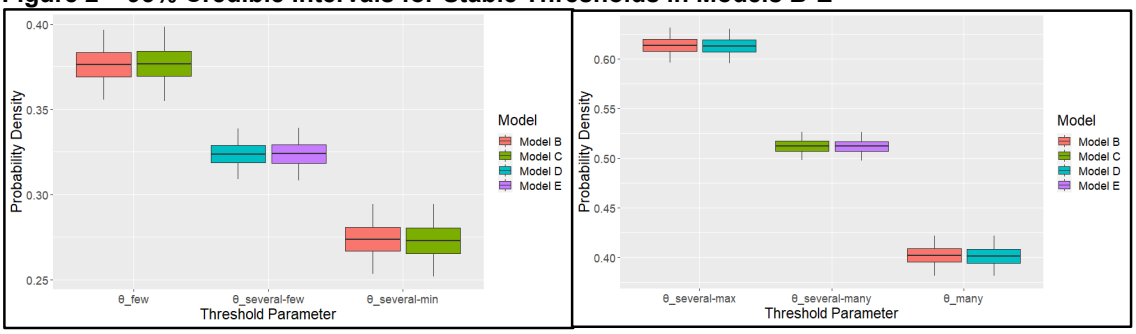


Figure 3 – θ_{many} Distributions for Model A – ITM

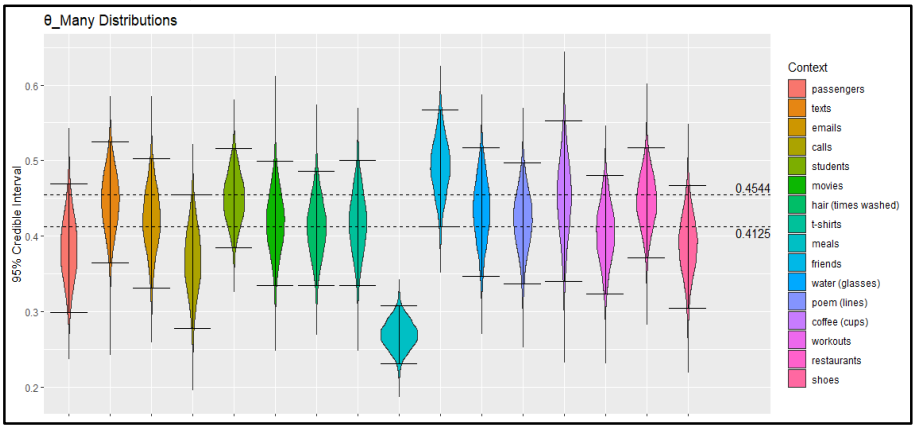
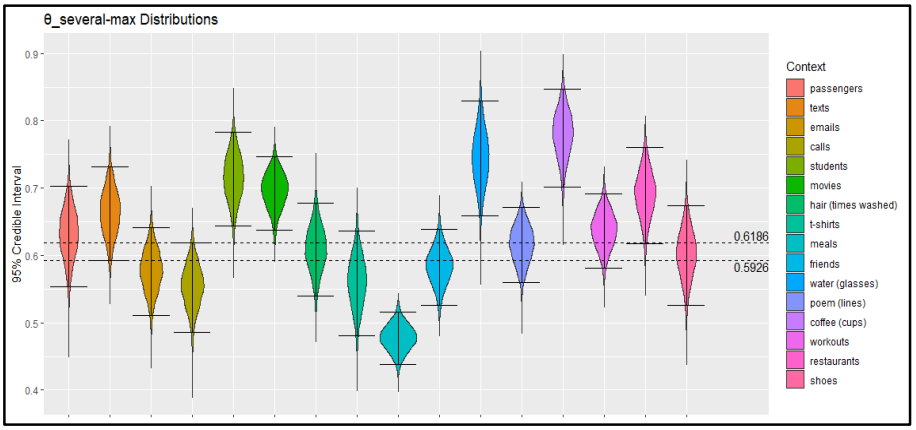


Figure 4 – $\theta_{several-max}$ Distributions for Model A – ITM



References
[1]Schöller & Franke (2017)
[2]Gelman, Hwang & Vehtari (2013)

Universal quantification without language? Ten-month-old infants represent the universality of visually presented properties.

Universal quantification–the logical operations lexicalized with "all", "every", and "each"–supports the universal (i.e., without exception) application of a predicate to the things that fall under a concept. By means of logical quantification, the human mind can represent universality over an infinite number of entities (e.g., "EVERY natural number is divisible by 1"). However, inferring such a representation requires sophisticated deductive abilities. In contrast, in everyday life, we can also easily notice universality in our immediate visual experiences (e.g., "look, ALL the apples in front of us are green!"). Our ability to detect universality based on observable data paves the way for studying forms of universal quantification beyond language (e.g., at the interface with vision) and their cognitive development. With six experiments, we provide initial evidence that preverbal infants represent the universality of visually-presented actions in much the same way as adults.

The distinction between individual-implicating (first-order) and group-implicating (second-order) forms of universal quantification is fundamental to logic and language (Knowlton et al., 2021). When prompted with quantified sentences (e.g., is each dot blue? Are all the dots blue?), speakers verify the universality of visually-presented properties by recruiting a-linguistic cognitive systems dedicated either to the tracking of individuals (i.e., the object-file system, Scholl 2001) or of groups (i.e., the ensemble systems; Alvarez, 2011) in the environment. These core systems are already in place in the first year of life. Can human infants deploy them to detect universality without using linguistic quantifiers?

Here we ask whether mastery of words like "all" and "each" is required to think about universality and to detect it in visual scenes, just as it has been argued that children need the words “one” "two", "three", "four", and "five" to represent cardinalities of exactly 5 items (Carey, 2009; Frank et al. 2012). In contrast to this strong Whorfian view, we offer initial evidence that 10-month-old infants have access to preverbal forms of universal quantification long before acquiring quantifier words, in line with the proposal that precursors of logical capacities may be in place in infancy (Cesana-Arlotti et al. 2018).

In an initial series of experiments, adult participants watched simple animations, with no linguistic descriptions, involving agents performing goal-directed actions (e.g., scenes of three/five/eleven chevrons, EACH chasing a ball alone, or of three/five/eleven chevrons, ALL chasing one ball together; see Fig.1 for one example of our procedure). In the EACH situations, adults were less likely to detect universality when the number of agents exceeded working memory limits (i.e., > 4 agents), indicating that universality was represented across multiple discrete events (e.g., $chevron_1$_$chasing_ball_1$; $chevron_2$_$chasing_ball_2$; …). In ALL situations, adults were equally likely to notice the universality no matter how many agents were present, in line with the computing of universality based on a single collective representation (e.g., an ensemble). The interaction between the number of agents and the distribution of goals indicates two representations of universality (group- and individual-universality).

Next, we asked whether 10-month-olds notice the universality of goal-directed actions similarly to adults in five visual-habituation experiments (See Fig.2 for a description of our procedure). In Experiments 2 and 3 (n = 24 each), infants who were habituated to ALL videos with three chasers successfully dishabituated to EACH videos with three chasers ($p = 0.008$), and vice versa ($p = 0.01$; Fig.1). This result shows that infants encoded different representations of our ALL and EACH movies. However, it remains unclear how such difference was encoded: along some low-level perceptual dimension (e.g., variability in the orientation of the chevrons' tips), or else in terms of the contrast between group- and individual-universality?

In three ongoing studies, we habituate infants to 5-agents "all" videos (Experiment 4, n = 27/30), 5-agents "each" videos (Experiment 5, n = 15/30), or 3-agents "each" videos (Experiment 6, n = 0/30). In all three cases, we test for dishabituation to "broken-chasing"

movies in which the chevrons are not pointing toward the target they chase but toward empty locations. Thus, the change in the variability of the orientations of their tips is equated across the three experiments (see Fig.2), and, thus, equally detectable. In contrast, preverbal representations of group- and individual-universality predict that infants, like adults, will fail to form a robust representation of 5-agent EACH chasing (as 5 is above their working memory limit for individual items; Feigenson, 2004), but will succeed in the 5-agent ALL and 3-agent EACH chasing conditions. Preliminary analyses initially confirm our predictions: in Experiment 4, infants dishabituate to broken chasing ($p = 0.02$), while in Experiment 5, they do not. These results point to a preverbal precursor of linguistic quantification in infants' representations of the universality of visually-presented actions. This supports the idea that language acquisition is not a prerequisite for basic forms of universal quantification.

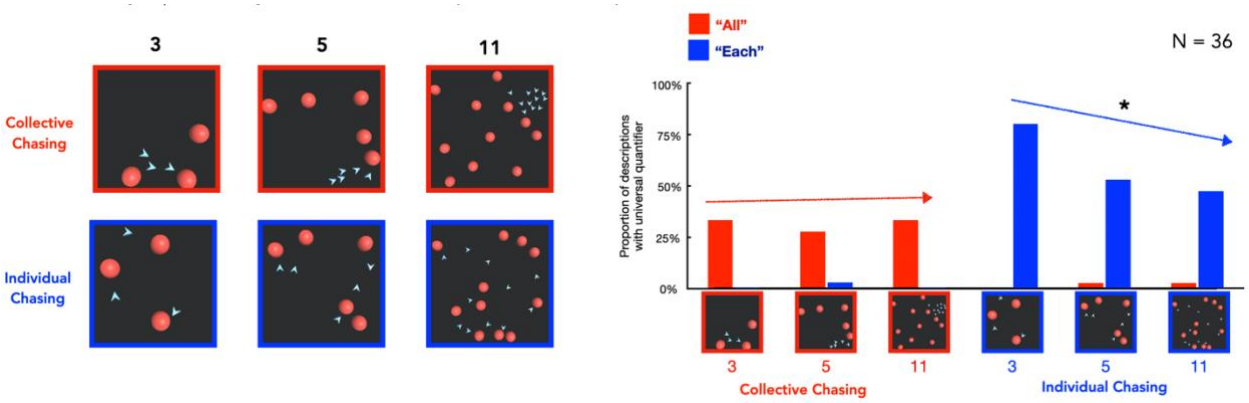


Fig. 1. Design and results of Experiments 1 (adults). In a MOT design, participants were asked to describe our movies. Across six conditions, we varied the distribution of the target and the number of agents. Each of the participants was presented with each of the six movies, exactly one time. After each movie, the participants were asked to describe it. At no time during the experiment were participants told to use quantifiers to describe the movies. We measured the proportion of trials where quantifiers were used to apply CHASING.

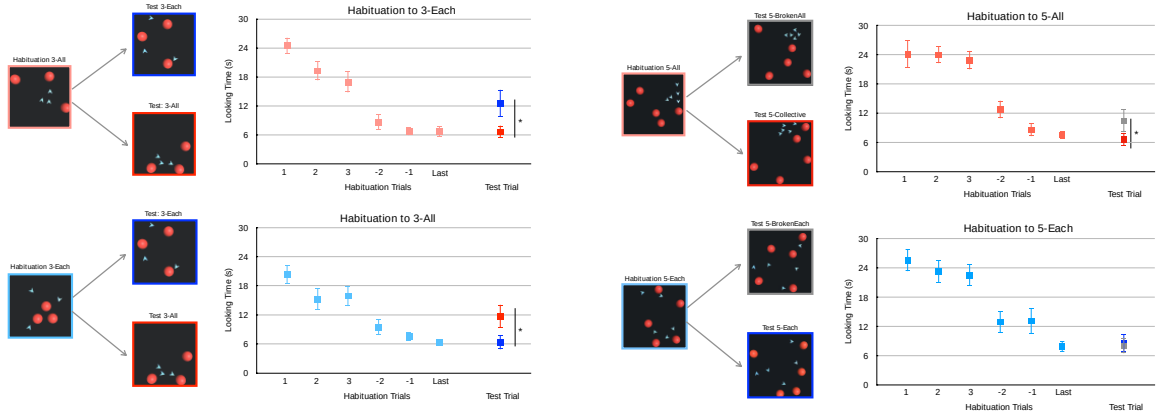


Fig. 2. Design and results of Experiments 2-5 (infants). In our procedure, an infant saw a sequence of videos of the same type (e.g., 3 agents, each chasing a ball alone), and, each time, we recorded how long the baby watched the video before getting bored and looking away. When an infant's looking times dropped under a critical threshold, she was tested either with a movie of a new type or with a new instance of the familiar movie type. If infants encode representations of the movies that support the detection of the change between movie types, they will retrieve interest in and look longer at the novel one. Infants looking patterns confirmed our predictions.

A psycho-semantic explanation of “each” and “every” quantifier use

“Each” and “every” can be used to express the same truth-conditions but differ in their contexts of use. A long-standing observation is that “each” is somehow more individualistic than “every” [A]. On standard (truth-conditional) approaches, capturing this difference requires additional machinery [B]. We adopt an alternative, mentalistic semantics for the two quantifiers and show that it correctly predicts a host of known and newly-observed constraints on how “each” and “every” are pragmatically used.

On this alternative [C,D], quantifier meanings are mental representations with different properties. In particular, “each” treats its first argument as independent individuals, whereas “every” groups its first argument. So despite shared truth-conditions, processing sentences with “each” or “every” leads to the assembly of distinct mental representations: “each” implicates the cognitive system for parallel-individuation [E]; “every” implicates the system for ensemble representation [F]. We propose that this meaning difference predicts a host of usage differences (all consistent with the long-held intuition that “each” is more individualistic [A]).

First, since parallel-individuation is subject to more stringent working memory constraints than ensemble representation [G], “every” should be preferred when the domain of quantification is larger as opposed to smaller. Second, since parallel-individuation treats individuals independently whereas ensemble representations describe many individuals with summary statistics (e.g., their average size) [H], “every” should be preferred when the speaker intends to license a global generalization as opposed to a statement about the locally-established domain. Third, though both quantifiers are ‘distributive universals’ [B,I], “every” groups the domain, and thus should be better suited to collective predicates, which apply to groups. In a series of experiments, we show that people’s preferences for “each” vs. “every” confirm these predictions.

In three forced-choice judgment experiments conducted on Prolific, participants chose between “each” and “every” for 12 sentences in minimally-different pragmatic contexts, manipulated within-subjects. They were asked to “pick which sentence best continues the story”. In Exp1 (n=100), the context either established a small or large domain (“three” vs. “three thousand martinis”; see example in (1)). Participants were more likely to pick “every” for the large compared to the small domain ($p<.001$; Fig1). Exp2 (n=100) established a small domain (see example (2)) and the quantificational phrase either referred back to that domain or explicitly went beyond it. Participants were more likely to pick “every” when quantification projected beyond the locally-established domain ($p<.001$; Fig2). Exp3 (n=100) sentences either contained collective predicates, which apply to groups as a whole (“gathered in the hall”) or distributive predicates, which apply to individuals (“went to their locker”; see example (3)). Participants were more likely to pick “every” given a collective predicate ($p<.001$; Fig3).

Finally, Exp4 (n=198) confirmed the domain size differences more directly: participants were asked how many martinis someone had in mind after they said “each/every martini needs an olive”. Participants were more likely to provide an answer ≤3 in the “each” than the “every” condition ($\chi^2=11.97$, $p<.001$; Table2).

The current results demonstrate that fine-grained differences in semantic representations affect canonical patterns of use in predictable ways, thereby offering natural links between the psycho-semantics and pragmatics of quantifiers. By treating the output of semantics as mental representations that are more finely articulated than propositions (/truth-conditions), we can explain these otherwise puzzling patterns.

(1) a. The bartender at the local tavern has made three martinis. (SMALL DOMAIN)
He said that {**each**/every} martini he made had an olive.
b. The bartender at the local tavern has made three thousand martinis. (LARGE DOMAIN)
He said that {each/**every**} martini he made had an olive.

(2) a. The bartender at the local tavern made a few martinis.
He said that {**each**/every} martini that he made has an olive. (LOCAL DOMAIN)
b. The bartender at the local tavern made a few martinis.
He said that {each/**every**} martini that's worth drinking has an olive. (GLOBAL DOMAIN)

(3) a. Math class at the local middle school lasts a full hour.
After class, {**each**/every} student went to their locker. (DISTRIBUTIVE PREDICATE)
b. Math class at the local middle school lasts a full hour.
After class, {each/**every**} student gathered in the hall. (COLLECTIVE PREDICATE)

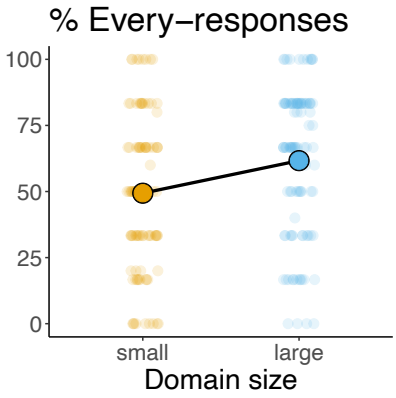


Figure 1 - Exp1

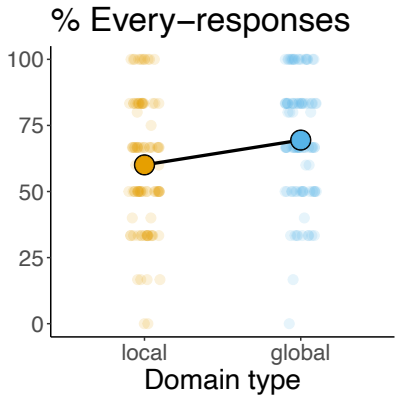


Figure 2 - Exp2

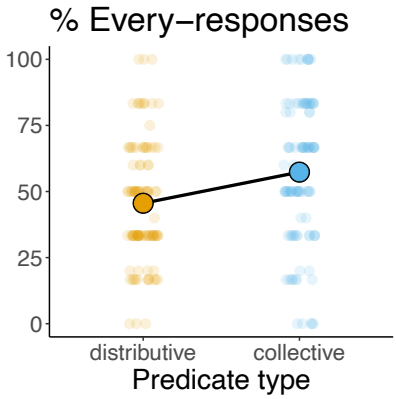


Figure 3 - Exp3

Table 1: Results of mixed-effects binomial regression with effects coding

Experiment	Estimate	SE	z value	P(z)
1: Domain size	0.6995	.132	5.30	<.001 ***
2: Domain type	0.5707	.132	4.31	<.001 ***
3: Predicate type	0.58906	.129	4.56	<.001 ***

Table 2: Responses to the Exp4 question:
If someone said: {*each*/*every*} *martini needs an olive,*
how many martinis would you guess they have in mind?

Quantifier	**≤3**	4-5	≥6	Infinitely many	Exhaustive (e.g., "all of them")
Each	**62**	10	12	0	9
Every	**29**	13	21	5	30

References

[A] Vendler (1962) *Each* and *every*, *any* and *all* **[B] Beghelli & Stowell (1997)** Distributivity and negation: the syntax of *each* and *every*. **[C] Knowlton, Pietroski, Halberda & Lidz (2021)** The mental representation of universal quantifiers. **[D] Knowlton (2021)** The psycho-logic of universal quantifiers. **[E] Kahneman, Treisman & Gibbs (1992)** The reviewing of object files: Object-specific integration of information. **[F] Whitney & Yamanashi Leib (2018)** Ensemble perception. **[G] Feigenson & Carey (2005)** On the limits of infants' quantification of small object arrays. **[H] Haberman & Whitney (2012)** Ensemble perception: Summarizing the scene and broadening the limits of visual processing. **[I] Dowty (1987)** Collective predicates, distributive predicates, and *all*.

Reading times show effects of contextual complexity and uncertainty in comprehension of German universal quantifiers

Introduction: Current semantic and pragmatic theory offers detailed models of meaning-related processes in a wide range of linguistic phenomena. In addition to classical approaches, these models do not only intend to explain the compositional derivation of sentence meaning in general, but also focus on phenomena like incremental meaning composition [e.g. 1], the complexity of meaning representations ([e.g. 2, 3]) and contextual effects on the behavior of speakers and listeners [e.g. 4, 5]. Despite these recent advances, relating predictions derived from semantic and pragmatic theory to processes during online comprehension remains elusive up until today, while theory-driven syntactic considerations have been implemented into models of on-line sentence comprehension for decades. This is especially surprising as highly comparable linking hypotheses could be developed on the basis of recent semantic and pragmatic models. For example, one could assume that complex meaning representations are generally avoided, or that highly expected sentence continuations lead to facilitation during incremental processing. We attempt to bridge this gap by studying how complexity and uncertainty in sentence meaning affect on-line sentence comprehension. To this end, we examined how restrictive processes incrementally interact with other aspects of quantifier meaning, comparably to previous studies [6, 7]. In the current experiment, we combined self-paced reading with picture-sentence verification to test how reading times are affected by meaning-related processes.

Methods: In each trial of the current self-paced reading experiment (N = 41), participants first inspected a picture context showing geometrical objects inside and outside of a container shape (one of the Fig. in 1a-1d) and then read a German universally quantified sentence as in example (1-a-c). Half of the sentences contained a restrictive relative clause, which could lead to a possible meaning change by introducing a particular subset reading. Sentences were presented word-by-word (with punctuation displayed separately) using the moving-window technique, and participants then performed a truth-value judgment task. For sentences following the simple contexts 1b and 1c, a truth-value judgment is possible already on the color adjective whereas, with contexts 1a and 1d, the judgment has to be delayed until the preposition is encountered. If we additionally assume that, by default, participants expect true utterances [6], their expectations would diverge between contexts 1b vs. 1c on the color adjective and between contexts 1a vs. 1d on the preposition. Based on these considerations as well as conclusions from previous studies, we predicted an effect of truth value on the color adjective in the former (context 1b vs. 1c) and on the preposition in the latter conditions (context 1a vs. 1d). To test for effects of pictorial complexity and truth values, we statistically analyzed reading times on the color adjective and the preposition of the restricted sentences using a linear mixed effects model. As fixed effects the model included the factors PICTURE COMPLEXITY (levels simple (1b, 1c) vs. complex (1a, 1d)) and GLOBAL TRUTH VALUE (levels true vs. false) in a 2x2 factorial design. Note that in this analysis contexts such as 1a and 1d were aggregated on the color adjective according to the GLOBAL TRUTH VALUE of the specific sentence-picture combination and the two prepositions were also aggregated.

Results: Across conditions, truth-value judgments were correct in the majority of cases ($87.1\%-91.9.8\%$). Word reading times are shown in Fig 2. PICTURE COMPLEXITY led to significantly longer reading times. These effects were sustained over several regions and turned out to be reliable on the color adjective ($t = 2.331, p = .022$) as well as on the preposition ($t = 5.982, p < .001$). In contrast, GLOBAL TRUTH VALUE affected reading times only after a truth-value judgment was possible (PICTURE COMPLEXITY $\times$ GLOBAL TRUTH VALUE interaction on the colour adjective: $t = 1.928$ (pairwise comparisons: $t = 2.504$ vs. $t = 0.169$ in the simple and complex conditions, resp.); main effect of GLOBAL TRUTH VALUE on the preposition: $t = 1.741, p = 0.0819$).

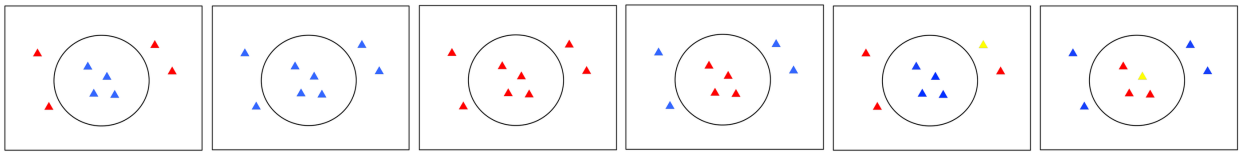
Discussion: The current results showed that picture complexity and truth values affected reading times in the expected direction. First, with regard to the complexity effect, we assume that a theory that describes how representations of context in memory affect the online construction of meaning representations could offer a plausible explanation for this finding. While we are not aware of such a theory, we think memory-

based approaches to syntactic processing could be instructive [e.g. 8, 9]. Second, in line with previous ERP results [e.g. 6], the truth-value effect on the colour adjective could either reflect a local truth evaluation, or, alternatively, a facilitation of sentence continuations that are expected because they still allow for true descriptions of the context. To distinguish between these two types of explanations, we are currently conducting a follow-up experiment using pictures like shown in Figures 1e and 1f, in which one triangle is colored differently, e.g. yellow instead of red. In such pictures, no truth-value judgment is possible on the adjective but, in contrast to the complex pictures in Fig. 1a and 1d, the actually presented colour adjective (e.g. *blue*) is the only one that still allows for true sentence completions in our design. Results form this follow-up will also be informative with respect to the potential role of salience, e.g. of colour terms primed by the picture contexts. In sum, by investigating quantifier restriction in a self-paced reading task, the current study showed that meaning-related processes may incrementally affect on-line sentence comprehension. Together with our planned follow-up studies, the current study is intended to be the basis for implementing formal-pragmatic and semantic considerations on theories of on-line sentence comprehension.

(1) a. Alle Dreiecke sind blau, die innerhalb des Kreises sind.
All triangles are blue that inside of_the circle are

b. Alle Dreiecke sind blau, die außerhalb des Kreises sind.
All triangles are blue that outside of_the circle are

c. Alle Dreiecke sind blau



(a) (b) (c) (d) (e) (f)

Figure 1: Picture contexts used in the current (1a-1d) and follow-up experiment (1a-1f)

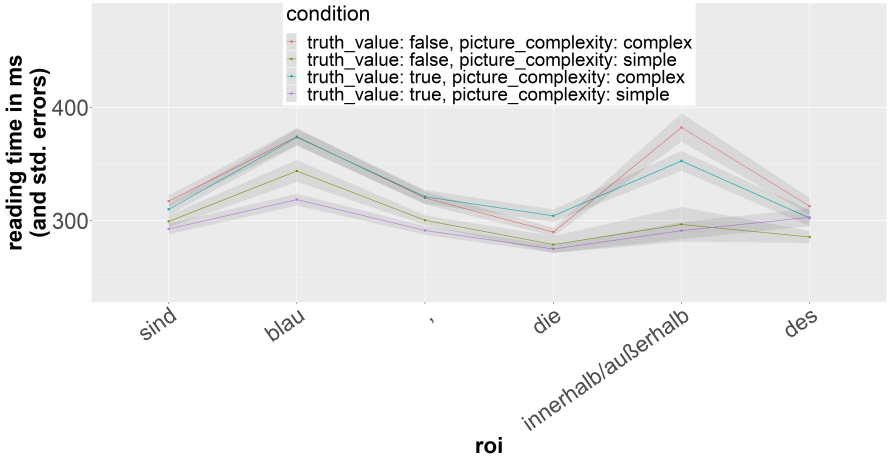


Figure 2: Reading times in regions 3 to 8 (conditions aggregated)

References:

1. Bott, O. et al. *J Semant* **34,** 201–236 (2017). 2. Pietroski, P. et al. *Mind and Language* **24,** 554–585 (2009). 3. Szymanik, J. *Quantifiers and Cognition. Logical and Computational Perspectives* (Springer, 2016). 4. Frank, M. C. et al. *Science* **336,** 998–998 (2012). 5. Van Tiel, B. et al. *PNAS* **118** (2021). 6. Augurzky, P. et al. *Lang Cog* **9,** 603–636 (2017). 7. Augurzky, P. et al. *Cognitive Sci* **43** (2019). 8. Lewis, R. et al. *Cognitive Sci* **43,** 375–419 (2005). 9. Futrell, R. et al. *Cognitive Science* **44** (2020).

Visual boundaries in sign motion: processing with and without lip reading cues

Julia Krebs[1], Evie Malaia[2], Ronnie B. Wilbur[3] & Dietmar Roehm[1]
[1] University of Salzburg, Austria; [2] University of Alabama, AL, USA; [3] Purdue University, IN, USA

Sign languages allow investigation of the hypothesis that language processing builds on neural circuitry underlying general, non-linguistic abilities – such as the ability to identify, parse, and interpret actions. Sign languages utilize articulator motion profiles similar to motion profiles of observed events, conveying event-based semantics and constructing grammatical features such as aspect. Studies of unrelated sign languages indicate that event structure, expressed by verbs and their arguments, is overtly expressed in verb sign dynamics, manifesting Event Visibility (cf. review in Malaia & Milković, 2021). For instance, signs denoting an event with an endpoint (telic verbs, e.g. English 'fall') have a sharper final movement with rapid deceleration to a stop. In contrast, verbs denoting an ongoing event, or one without an inherent endpoint (atelic verbs, e.g. English 'sleep'), might be conveyed by a steady movement without rapid acceleration profile (Wilbur 2008). Remarkably, visual event structures of sign language verbs are recognized by hearing non-signers without any knowledge of sign language. In an alternative-forced-choice task, hearing non-signers were found to associate unfamiliar (pseudo-)signs involving a dynamic visual boundary with telic events (Strickland et al. 2015). Hearing non-signers also were found to neurally process the perceptual-kinematic difference between atelic and telic verbs in American Sign Language (Malaia et al. 2012).

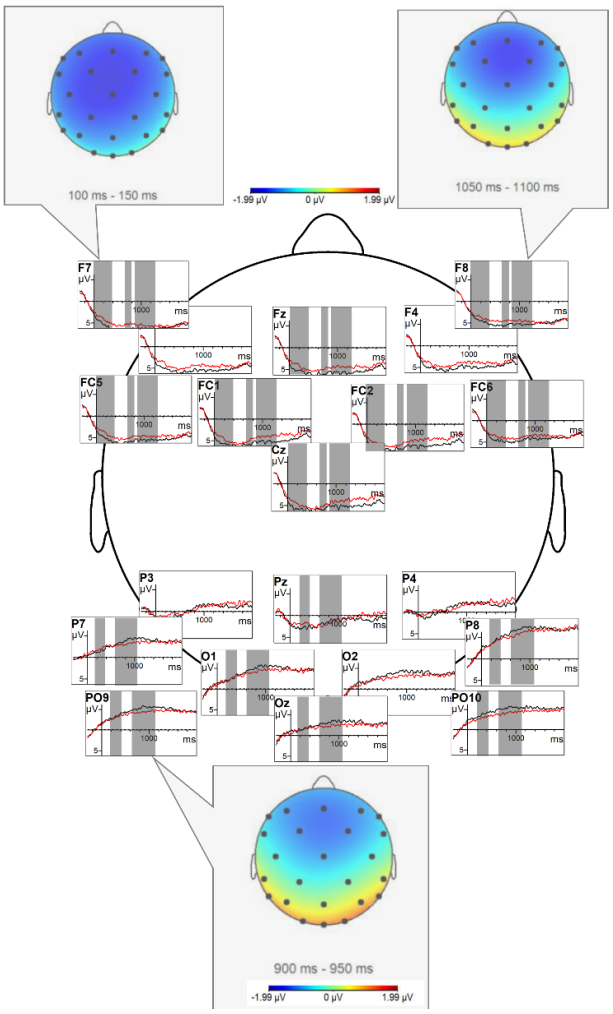


Figure 1. Telic/atelic sign processing without non-manual cues

In this study, we first assessed the timeline of neural processing mechanisms in non-signers processing telic/atelic signs to understand the pathways for incorporation of physical-perceptual motion features into the linguistic system. Experiment 2 further probed the possible impact of visual information provided by lip-reading (speech decoding based on visual information from the face of the speaker, most importantly, the lips) on the processing of telic/atelic signs in non-signers.

Hearing German speaking non-signers (N=27) were presented with telic and atelic verb signs unfamiliar to them, which they had to classify in a two-choice decision task (cf. Strickland et al. 2015). The stimuli consisted of signs from unrelated sign languages (Turkish, Italian, Croatian and Dutch). Behavioral data analysis confirmed that non-signers could classify telic/atelic verbs, whereby telic events were easier to classify than atelic events. Processing differences for atelic compared to telic sign stimuli were revealed at the neurophysiological level (Figure 1). Beginning from sign onset (i.e. target handshape positioned in target location), statistically significant neural differences in processing appeared anteriorly (0-200ms, 650-800ms, 850-1300ms), posteriorly (600-1050ms), and in a broadly distributed manner (200-400ms). The timing and distribution of ERP effects appear to reflect both the differences in perceptual processing of verb types and the integration of perceptual and linguistic processing required by the task. These findings suggest that non-signers use visual-perceptual features of signs while engaging higher cognitive processing for classifying the percepts linguistically. Non-signers appear to segment visual sign language input into discrete

events as they try to map the observed sign language form to a linguistic concept that might represent the sign. The mechanism might be indicative of the potential pathway for co-optation of perceptual features into the linguistic structure of sign languages.

In Experiment 2, the participants were presented with telic and atelic signs of Austrian Sign Language (ÖGS), which both evidence a distinct telic/atelic motion profile (Krebs et al. 2021), and are accompanied by mouthing information (movement of the mouth forming (part of) the German corresponding word). Behavioral data revealed that participants responded more accurately, faster, and with more certainty to the classification task. ERP findings differ from those of Experiment 1: ERP effects for telic compared to atelic signs started in later time windows, extended into later time windows, and showed a primarily posterior distribution (Figure 2).

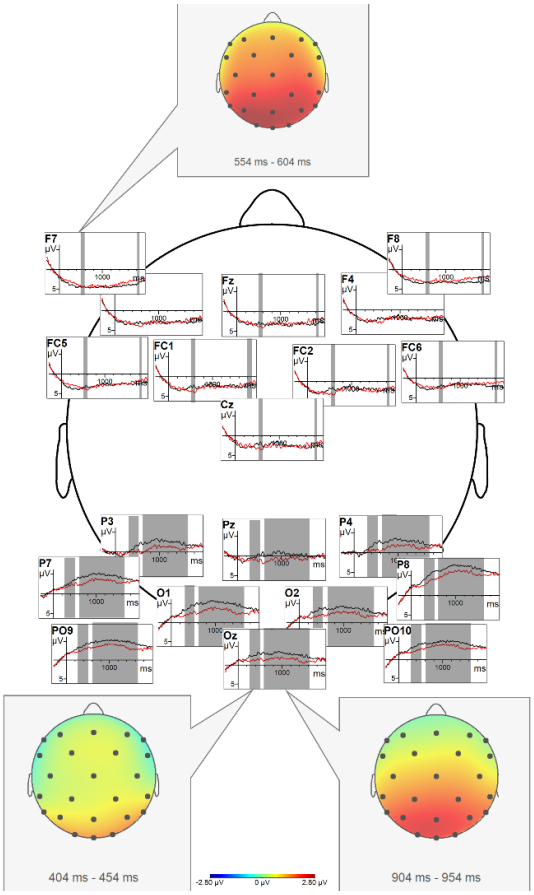


Figure 2. Telic/atelic sign processing with non-manual cues

The findings suggest that non-signers rely on information provided by mouthing, if available. In this case non-signers pay more attention to lip-reading (as self-reported after the experiment), as opposed to tracking visual motion profiles in the stimuli. Because linguistic information provided by lip movement is part of audio-visual spoken language processing, it was easier for non-signers to classify the signs in Experiment 2 compared to Experiment 1. The ERP effects for telics vs. atelics observed in Experiment 2 also reflected the qualitatively different mapping/integration processes for telic compared to atelic verbs. However, a different strategy was used by the participants in the two experiments, leading to different ERP patterns in both experiments. In line with previous work (e.g. Malaia et al. 2009; Ji & Papafragou 2020), the differences in ERP effects during processing of telic vs. atelic stimuli observed in both experiments appear to indicate easier event segmentation in response to telic stimuli.

References

Ji, Y., & Papafragou, A. (2020). Is there an end in sight? Viewers' sensitivity to abstract event structure. *Cognition*, *197*, 104197.

Malaia, E., Wilbur, R.B., & Weber-Fox, C. (2009). ERP evidence for telicity effects on syntactic processing in garden-path sentences. *Brain and Language, 108*, 145-158.

Malaia, E., Ranaweera, R., Wilbur, R.B. & Talavage, T.M. (2012). Event segmentation in a visual language: Neural bases of processing ASL predicates. *Neuroimage, 59*, 4094-4101.

Malaia, E.A., & Milković, M. (2021). Aspect: Theoretical and experimental perspectives. In: J. Quer, R. Pfau & A. Herrmann (eds.), *The Routledge Handbook of Theoretical and Experimental Sign Language Research*. London: Routledge, 194-212.

Krebs, J., Strutzenberger, G., Schwameder, H., Wilbur, R.B., Malaia, E. & Roehm, D. (2021). Event visibility in sign language motion: Evidence from Austrian Sign Language (ÖGS). *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 362-368.

Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M. & Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences, 112*, 5968-5973.

Wilbur, R.B. (2008). Complex Predicates involving Events, Time and Aspect: Is this why sign languages look so similar? In: J. Quer (ed.), *Signs of the time.* Signum Verlag, 217-250.

Exploring the pragmatic import of non-truth-conditional discourse connectives

Cecile Larralde[1,2], Nausicaa Pouscoulous[3], Ira Noveck[2]

Grice famously made a distinction between conversational and conventional implicatures [1]. One important feature of conventional implicatures is that they, unlike their conversational counterparts, are assumed to contain a pragmatic component that is bound to a lexical item. To appreciate conventional implicatures, consider (1a-c.):

(1) Mary ate two apples {a. and/b. but/c. so} Luke ate one.

As one can see, the meaning of *and* in (1a) is compatible with a logical conjunction; in contrast, the discourse connectives in (1b) and (1c) additionally convey contrast and causality, respectively. These non-truth-conditional features are the focus of the current investigation.

Note that we adopt Grice's nomenclature of *conventional implicature* for historical reasons and convenience but not necessarily to indicate full endorsement of his account. In fact, several other accounts of DCs have been proposed since Grice first introduced the concept of conventional implicature. Sanders et al. have carefully examined the different types of discourse relations marked by DCs and proposed what they called a taxonomy of coherence relations [2]. Blakemore [3], Wilson [4] and Hall [5] have argued that DCs encode procedural meaning which makes the hearer shape expectations about the upcoming discourse. From a semantic perspective, DCs are generally seen as interacting with the entailments of a sentence. For example, the *but* in (1b), could be understood as denying the entailment that *Luke ate as many apples as Mary* [6,7]. Our overall goal is to determine whether those pragmatic features that are assumed to be intrinsic to the meaning of individual DC's add processing costs to them. All accounts would be edified by such an investigation.

Experimental studies using eye-tracking or ERP paradigms have reported fast integration of DCs to discourse representation in context-rich paradigms [8–12]. To our knowledge however, no studies have examined how the very presence of DC's, such as *but* or *so*, themselves in context poor scenarios force a reader to infer the corresponding discourse relation. Past studies on scalar implicatures, which have demonstrated that the pragmatic interpretation of an expression incurs higher processing costs relative to a straight-forward semantic reading, were the inspiration for the current study e.g., see [13], [for a review, see 14]. That said, scalar implicatures require contextual licensing [15] so it is unclear whether the discovered additional cognitive costs pertain to computing the contextual information, to the scalar inference itself or to both. Here, we set up a paradigm in which the DC's *but* and *so* arise as part of a sentence whose context is minimal as we aim to determine whether they are responsible for slowdowns with respect to *and*. In addition, we determine whether the slowdowns are arguably linked to creating discourse expectations. To anticipate, we indeed report that the DC's *but* and *so* lead to slowdowns while their pragmatic features appear to lead to precise discourse expectations.

[1] Corresponding author: cecile.larralde.15@ucl.ac.uk
[2] Université Paris Cité, LLF, CNRS
[3] University College London

EXPERIMENT: In this pre-registered study (OSF link not disclosed to preserve anonymity), we tested 79 native English speakers on an online reaction time and truth evaluation task. Each participant completed 108 trials (36 fillers). As shown in Figure 1, trials displayed a fixation cross, a three-letter word, and then a two-part statement about the letters in the target word. Three dependent variables were recorded: 1) participants' reading time of Part 1, which includes the connective; 2) participants' accuracy in evaluating the full sentences and; 3) their (Part 2) answer reaction time.

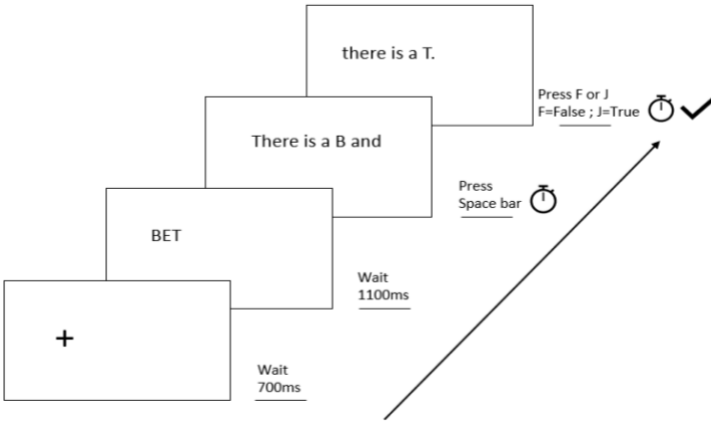


Figure 1 The unfolding of a single trial

The 72 test sentences were carefully designed to remove all possible sources of inference outside of the *but* and *so* implicatures. All told, our set up amounts to a 3 (and/but/so) X 2 (affirmative/negative expression of Part 2) X 2 (true/false statement) design. The example in Figure 1 is an *and-affirmative-true* trial. Keeping BET as the target word for expository purposes, other possible trials could be described as *so-affirmative-false* ("There is a B so there is a K."), or as *but-negative-true* ("There is a B but there is no F.") and so on. Note too that the task was designed to keep participants vigilant to each trial so filler items would include cases such as *There is no X but there is a B*.

PREDICTIONS: 1) As we indicated above, we expect the pragmatic import bound to *but* and *so* to lead to further inferencing when compared to *and*. We thus predict that Part 1's which end with *but* and *so* to be read more slowly than those which end with *and*. 2) For the answer reaction time (the truth-value-judgement of the trial), we predict that the processing of a negative Part 2 to be facilitated by the presence of *but* since this connective should prepare participants to process the contrasting negation. 3) For Part 2 experimental items that have true affirmatives, we also predict reaction times to be slower when they arise in the wake of *but* and *so* rather than for *and* in Part 1, due to the absence of any contextual contrast or causal link.

RESULTS: The reaction times in Part 1 and Part 2 were analysed using a Bayesian linear mixed effects model in R brms() [16]. Results revealed that Part 1's ending with *but* (1274.53*ms*) and *so* (1275.72*ms*) were indeed read on average more slowly than those ending with *and* (1239.09*ms*). The statistical analysis of the data (see the posterior distributions of the log-transformed reading times of the connectives in Figure 2) confirmed this difference. Furthermore, participants required more time to evaluate sentences containing *but* and *so* when the DC-specific inference was not realized in Part 2. However, when the *but*-contrast arose in the presence of a negation in Part 2, reaction times were not affected relative to *and* trials (Figure 3).

CONCLUSION: Our results revealed that the *but* and *so* sentences were costly to process relative to the logical *and*-reading. This suggests that even when pragmatic information is lexicalised in a DC, it is not as fast as those that arguably do not include such pragmatic information. Furthermore, the answer reaction time data suggests that participants created DC-specific expectations for the post-connective part of the sentence. Ongoing work aims to replicate these findings while avoiding cases whose Part 2's render the statement infelicitous.

REFERENCES: [1]Grice (1975), [2]Sanders et al. (1992), [3]Wilson (1994), [4]Blakemore (2000), [5]Halll (2007), [6]Winter et al. (1994), [7]Umbach (2005), [8]Xiang & Kuperberg (2015), [9]Koehne-Fuetterer et al. (2021), [10]Canestrelli et al. (2013), [11]Koehne & Demberg (2013), [12]Schwab & Liu (2020), [13]Chevallier et al.(2008), [14]Noveck (2018), [15]Breheny et al. (2006), [16]Buerkner (2018)

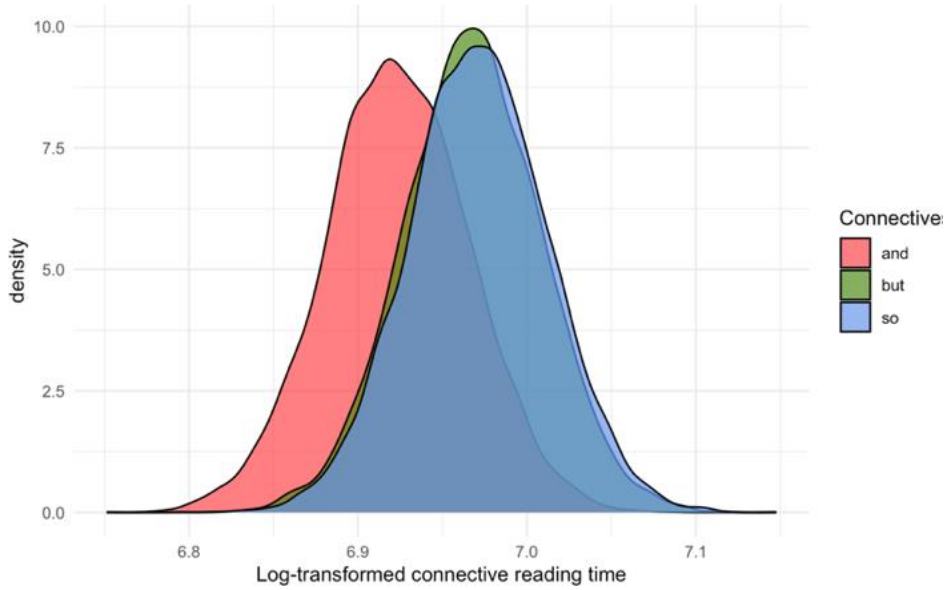


Figure 2 Posterior distribution of the log-transformed reading times for Part 1

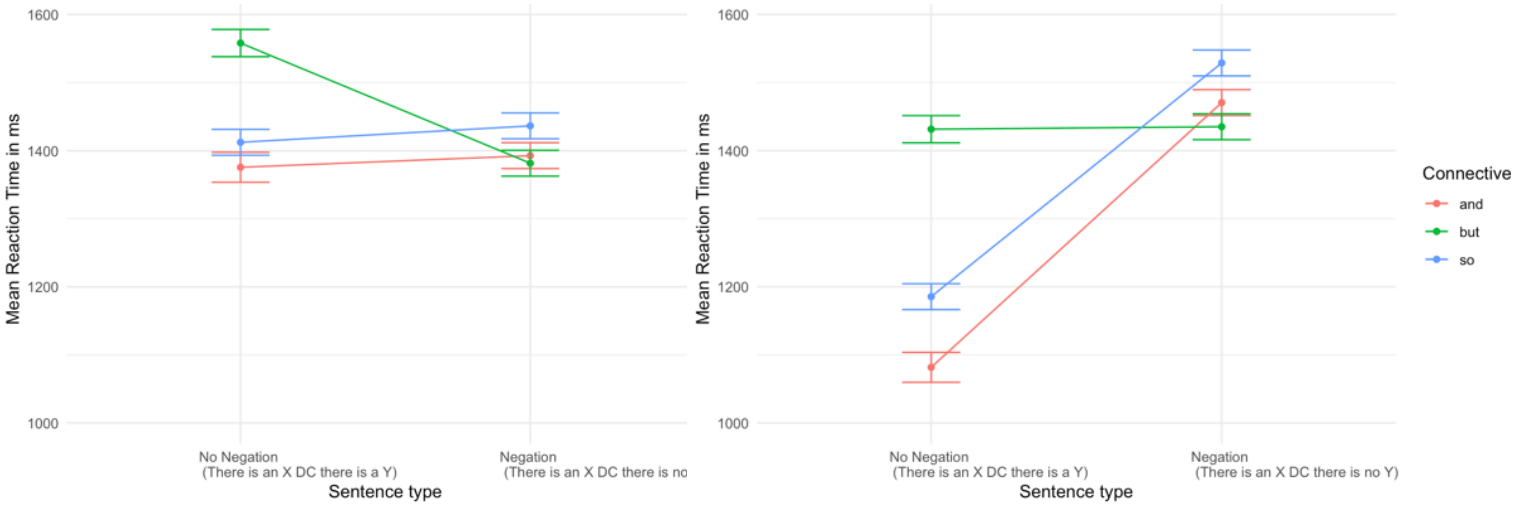


Figure 3 Mean answer reaction time (Part 2) by connective and sentence type.
Truth-value of sentence: top =TRUE, bottom =FALSE

Logical connectives: An extendable experimental paradigm

Overview The meaning of words like *not*, *and*, *or* and their relation to logical connectives such as negation, conjunction and disjunction has been an integral part of semantic and pragmatic theory (Grice 1978, Horn 1972, Gazdar 1980, etc.). Two central topics in this area are the status of exclusive interpretation of disjunction via implicature and the interpretation of negated conjunction and disjunction. Exclusivity implicatures can be viewed as default inferences (e.g. Levinson 2000) that are suspended in certain linguistic or grammatical environments such as questions and antecedent of conditionals (Chierchia 2004). Negated conjunction and disjunction are argued to follow de Morgan laws but vary in scope parameters cross-linguistically (Szabolcsi 2002, Szabolcsi & Haddican 2004, Crain 2012). Experimental studies have investigated these topics and the interpretation of connectives words separately (e.g. Chevallier et al. 2008; Lungu et al. 2021), but few have looked at them together within the same experimental paradigm and crosslinguistically.

We present a study that tests participant judgments, first here in written English, for different combinations of connectives *not*, *and*, *or*, and *either-or* across three different linguistic environments: questions, statements, and antecedent of conditionals, with the goal of extending the paradigm to other languages as well as spoken English and acquisition studies. First, there was not a large effect of these linguistic environments in shaping exclusivity inferences or connective interpretations generally, and across environments, disjunction was typically interpreted as inclusive. Second, while the negation of a disjunction ($\neg[p \vee q]$) was interpreted as the conjunction of negatives ($\neg p \wedge \neg q$), the negation of a conjunction ($\neg[p \wedge q]$) received two interpretations across environments: conjunction of negatives ($\neg p \wedge \neg q$) and disjunction of negatives ($\neg p \vee \neg q$). These results are compatible with theories that allow variable scope relations between English negation and conjunction (e.g. Winter 2000), and raise challenges for theories assuming uniform scope or default suspension of implicatures based on the linguistic environment.

Methods The study was designed as a card selection task to minimize metalinguistic task demands, with an eye toward future crosslinguistic/acquisition work. In each trial, participants (N=150) viewed six cards with the following cartoon images: 1.a cat 2.a dog 3.an elephant 4.a cat and a dog 5.a dog and an elephant and 6.a cat and an elephant. They also saw a written English sentence, and were asked to select "the cards that matched". Participants could select any single or combination of these cards by clicking on each card. We varied the sentences to test three types of linguistic environments (between-subjects, N=50 per environment): questions (e.g. *which has a cat or a dog?*), statements (e.g. *Bob selected the cards that had a cat or a dog*), and antecedent of conditionals (e.g. *Select a card if it has a cat or a dog*). The cards stayed the same throughout the study and the verb combining with nominal (e.g. *cat*) was always *have*.

Given a linguistic environment, the study had 7 experimental trial-types, among which we expected more potential variation in answers: 1.simple positive (e.g. *has a cat*), 2.simple negative (e.g. *doesn't have a cat*), 3.positive disjunctive (e.g. *has a cat or a dog*), 4.negative disjunctive (e.g. *doesn't have a cat or a dog*), 5.complex positive disjunctive (e.g. *has either a cat or a dog*), 6.complex negative disjunctive (e.g. *doesn't have either a cat or a dog*), 7.and negative conjunctive (e.g. *doesn't have a cat and a dog*). There were also 7 control trial types (Figure 1, Left), among which we expected less variation, and which were presented in a randomized block after the randomized experimental block.

Results Since the results for all three linguistic environments were similar we only discuss the question environment here shown in Figure 1 (Right panel). In positive disjunctive trials (e.g. *has a cat or a dog?*), the majority of responses were compatible with an inclusive interpretation, as measured by inclusive disjunctive control trials (e.g. has a cat or a dog or both?). This was also the case in complex positive disjunctive trials with *either-or* (e.g. *has either a cat or a dog?*). We tested

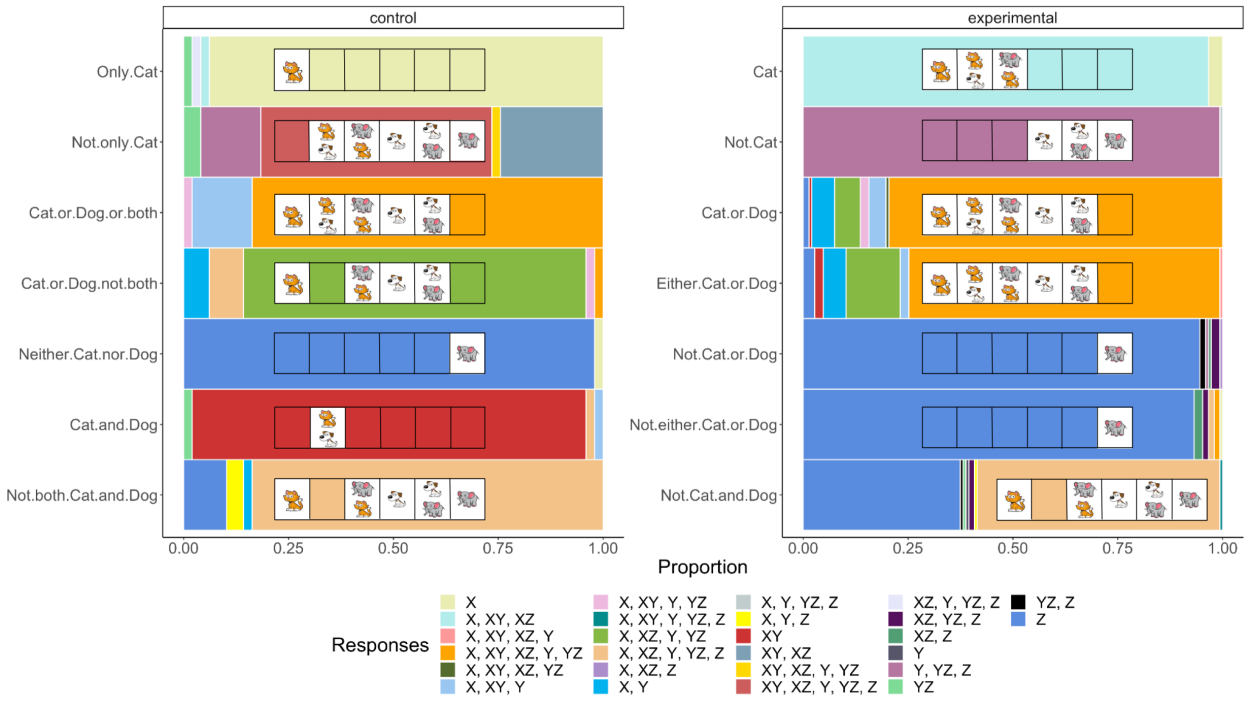


Figure 1: Results for the question environment for control (left) and experimental (right) trials. The x-axis shows proportion responses and the y-axis the trial-types with example sentences. Majority responses are annotated with the card selection. Given example sentences here, the legend's choices of cards (Responses) should be interpreted as X=cat, Y=dog, Z=elephant.

whether participants included the card with both animals in the disjunctive trials using a Bayesian mixed-effects logistic regression with random intercepts and slopes for subjects and fixed effect of trial-type (or vs. either-or) and linguistic environment (question, statement, conditional) and did not find evidence for the effect of *either* or linguistic environment in exclusivity inferences (95% CIs for all coefficients contained zero, 4 chains, 2000 iterations, 1000 warm-up, $\hat{R}$=1).

In negative disjunctive trials, whether simple (e.g. *doesn't have a cat or a dog?*) or complex (e.g. *doesn't have **either** a cat or a dog?*), responses were similar to complex negative trials with *neither-nor* (e.g. *has neither a cat nor a dog?*). In other words, the negation of a disjunction ($\neg[p \vee q]$) was interpreted as the conjunction of negatives ($\neg p \wedge \neg q$). However, in negative conjunctive trials (e.g. *doesn't have a cat and a dog?*) responses were consistently split between a *neither-nor* interpretation and a *not-both* interpretation in all linguistic environments (e.g. *doesn't have both a cat and a dog?*). In other words, the negation of a conjunction ($\neg[p \wedge q]$) was either a disjunction of negatives ($\neg p \vee \neg q$) or conjunction of negatives ($\neg p \wedge \neg q$).

Although limited in (imageable/existential) scenarios, this task provides a unified way to test the interaction of logical operators that allows for comparisons across languages and development.

References Chevallier, Noveck, Nazir, Bott, Lanzetti, & Sperber. 2008. Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology*. Chierchia, G. 2004. Scalar implicatures, polarity phenomena, and the syntax-pragmatics interface. In: Belletti (ed), *Structures and Beyond*. Crain, S. 2012. *The emergence of meaning*. CUP. Gazdar, G. 1980. Pragmatics and logical form. *Journal of Pragmatics*. Grice, H P. 1978. Further notes on logic and conversation. In Cole P. (eds): *Pragmatics*. Horn, L. R. 1972. On the semantic properties of logical operators in English. *UCLA Dissertation*. Levinson, S. C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press. Lungu, Falus, Panzeri. 2021. Disjunction in negative contexts: across-linguistic experimental study. *Journal of Semantics*. Szabolcsi & Haddican 2004. Conjunction meets negation: A Study in Cross-linguistic Variation. *Journal of Semantics*. Winter, Y. 2000. On some scopal asymmetries of coordination. In Bennis, Everaert, & Reuland (eds), *Interface Strategies*.

Non-Boolean Conditionals

Paolo Santorio
University of Maryland, College Park

Alexis Wellwood
University of Southern California

1. Overview. Standard theories predict that indicative conditionals (ICs) behave in a Boolean fashion when interacting with *and* and *or*. We test this prediction by investigating probability judgments about sentences of the form $\ulcorner a \to b \ \{ \text{ AND/OR } \} \ c \to d \urcorner$. Our findings are incompatible with a Boolean picture. This is challenging for truth-conditional theories of ICs, as well as for several other theories. Some trivalent theories hold promise for providing an account of our data.

2. Background. Boolean interpretations of *and* and *or* entail constraints about probabilities of compounds (see e.g. Adams 1998). The following two are relevant here:

***and*-drop.** If $\mathrm{A} \not\models \mathrm{B}$, $Pr(\mathrm{A}) > Pr(\mathrm{A} \wedge \mathrm{B})$ ***or*-drop.** If $\mathrm{A} \not\models \mathrm{B}$, $Pr(\mathrm{A} \vee \mathrm{B}) > Pr(\mathrm{A})$

These constraints apply to all sentences of natural language that express propositions. Thus, if truth-conditional theories of ICs are correct (see a.o. Stalnaker 1968, Kratzer 2012), the sentences in (1) are predicted to conform to the constraints on the right below.

(1) a. If Lea danced, Mia danced, or, If Lea didn't dance, Nina danced.
b. If Lea danced, Mia danced.
c. If Lea danced, Mia danced, and, If Lea didn't dance, Nina danced.

$$Pr(1a) > Pr(1b)$$
$$Pr(1b) > Pr(1c)$$

3. Experiment 1. Our experiment tests ***and*-drop** and ***or*-drop** for natural language ICs. Rather than relying on assumptions about probabilities of conditionals (like Stalnaker's Thesis; see Stalnaker 1970), subjects were asked to assign probabilities on the basis of observed frequencies.

After an exposition period (Experience phase), subjects were presented with several sentences and asked to perform a likelihood estimation task (Test Phase). Three main variables were manipulated: presence and type of connective (And vs Or vs None; within); compatibility of the two antecedents, when sentences involved two ICs (Compatible vs Incompatible; between); and frequency of the event described in the consequent, given the antecedent (50/50 vs 75/25; between).

In Experience, participants viewed 24 animations of 1 shape (Incompatible conditions) or 1-2 shapes (Compatible) traveling by "car" into a "tunnel", whereupon they changed into 1 of 2 colors (**Fig.1**). Then, participants answered (2), and were included only if they answered "yes" to both ($N = 153$). In Test, participants viewed two sets of 4 "mystery car" animations, and gave likelihood estimates for (i) the simple ICs in (3) and (ii) the compounds schematized in (4).

(2) a. If the SQUARE enters the tunnel, it always turns RED or YELLOW.
b. If the CIRCLE enters the tunnel, it always turns GREEN or BLUE.

(3) a. If the car was carrying the SQUARE, the SQUARE turned { RED / YELLOW }. $s \to r$, $s \to y$
b. If the car was carrying the CIRCLE, the CIRCLE turned { GREEN / BLUE }. $c \to g$, $c \to b$

(4) a. $s \to r$ { AND / OR } $c \to g$
b. $s \to y$ { AND / OR } $c \to b$

Findings. Our participants overestimated input frequencies in the 50/50 condition ('balanced' inputs occurred 50% of the time, mean estimate 68%) and in the lower frequency events of the 75/25 condition ('lower' input 25%, estimate 46%; cp. 'higher' input 75%, estimate 75%). Importantly for us, the ordering between estimates was accurate, and they were significantly different,

$F(1,148) = 8.15, p < .005$.[1] Also, and crucially, likelihood estimates were *not* impacted by the factors Compatibility or Connective, ps $> .53$. See **Fig.2** (L).

Discussion. ***and*-drop** or ***or*-drop** predict lower probability estimates for $\ulcorner s \to r$ OR $c \to g \urcorner$ over $\ulcorner s \to r \urcorner$, and for $\ulcorner s \to r \urcorner$ over $\ulcorner s \to r$ AND $c \to g \urcorner$. This asymmetry was not observed, revealing non-Boolean behavior. The contrast between estimated probabilities in the 50/50 and 75/25 conditions shows that subjects did make discriminating probabilistic judgments.

4. Analysis Our findings are challenging for all theories that vindicate ***and*-drop** and ***or*-drop**. Conversely, they can be explained by some trivalent theories (in particular Bradley 2002; see Rothschild 2014, Lassiter 2019 a.o. for similar views). Every clause A has definedness conditions $\mathrm{D}(\mathrm{A})$ and truth conditions $\mathrm{T}(\mathrm{A})$. $\mathrm{A} \to \mathrm{B}$ is defined iff A is true and B is defined, and true iff A and B are true. $\mathrm{A} \wedge \mathrm{B}$ ($\mathrm{A} \vee \mathrm{B}$) is defined iff at least one of A and B is defined, and true iff all (at least one of) the defined conjuncts (disjuncts) are true.

$$[\![\mathrm{A} \to \mathrm{B}]\!] = \begin{cases} \text{def. at } w \text{ iff } w \in \mathrm{T}(\mathrm{A}) \text{ and } w \in \mathrm{D}(\mathrm{B}) \\ \text{true at } w \text{ iff } w \in \mathrm{T}(\mathrm{A}) \cap \mathrm{T}(\mathrm{B}) \end{cases} \qquad [\![\mathrm{A} \wedge (\vee) \mathrm{B}]\!] = \begin{cases} \text{def. at } w \text{ iff } w \in \mathrm{D}(\mathrm{A}) \text{ or } w \in \mathrm{D}(\mathrm{B}) \\ \text{true at } w \text{ iff: if } w \in \mathrm{D}(\mathrm{A}),\ w \in \mathrm{T}(\mathrm{A}) \\ \quad \text{and (or) if } w \in \mathrm{D}(\mathrm{B}),\ w \in \mathrm{T}(\mathrm{B}) \end{cases}$$

Combined with a notion of trivalent probability (see Cantwell 2006), this semantics predicts failures of ***and*-drop** and ***or*-drop**.

5. Experiment 2 (control). One could worry that our findings reflect a flawed novel experimental paradigm. In response, we tested non-conditional sentences. We replaced the sentences in (3) with those in (5), modified the Test animations so that the mystery car initially shows the two shapes, and replaced the sentences in (4) with conjunctions/disjunctions of (5a) and (5b).

(5) a. The SQUARE turned { RED / YELLOW }. r, y

b. The CIRCLE turned { GREEN / BLUE }. g, b

Findings (n=83). We found a main effect of Connective in Experiment 2, $p < 0.0001$, due to estimates for *and* differing significantly from *or* and none (*and* 58.5%, *or* 68.6%, none 66.1%) , both ps < 0.007. This shows expected Boolean behavior at least in the 50/50 condition, alleviating concerns that our paradigm wouldn't be sensitive enough to detect such behavior. See **Fig.2** (R).

Fig.1: (a) 1-, 2-traveler scenes, (b) mystery scene, (c) Incompatible & (d) Compatible event schemas.

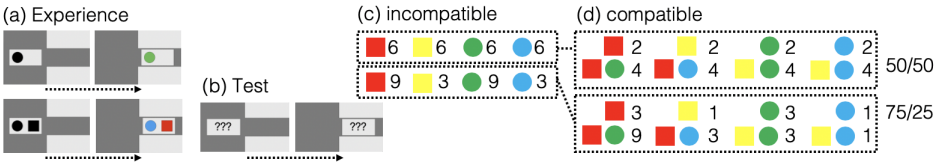
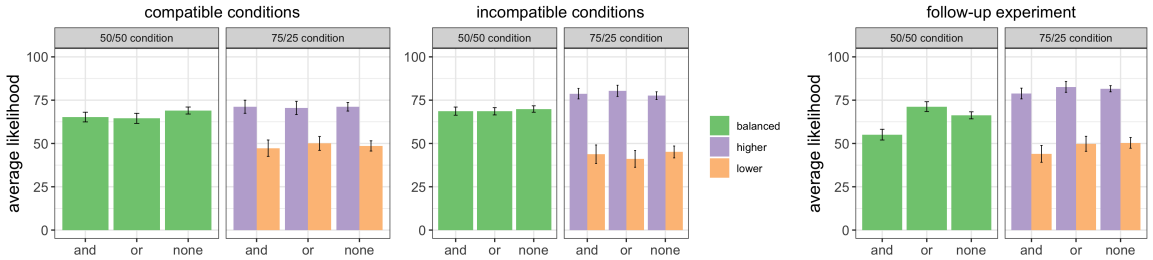


Fig.2: Results for Experiment 1 (L) and Experiment 2 (R).



[1] We report the results of a 3x2x2 ANOVA with a within-subject error term for connective type.