ELM 3 Program - June 12-14, 2022

Main Session Program

DAY 1 - June 12

Online Symposium on Language and Thought. June 12 9:00-10:45 (virtual)

Panel Talk: *Sandra Waxman* p. 1
Principled and precise links between object naming and object representation in prelinguistic infants
Panel Talk: *Alexis Wellwood* p. 2
Semantics at the language-mind interface
Panel Talk: *Paul Pietroski* p. 3
Logically Negative Thoughts without Negaters

Main-1.1. June 12 11:00-12:30 (virtual)

Mathias Barthel, Rosario Tomasello, Mingya Liu p. 4
Prediction and integration of discourse-level meaning are functionally related: EEG and reading time evidence
Stavroula Alexandropoulou, Nicole Gotzner p. 6
The effect of standards on scalar implicature processing of gradable adjectives: A web-based eye-tracking study
Fabienne Martin, Florian Schäfer, Despina Oikonomou, Felix Gölcher, Artemis Alexiadou p. 8
The 'no-agent' scalar implicature triggered by anticausatives is stronger when the causative alternative is structurally-defined

---

Main-1.2. June 12 16:00-17:30 (virtual)

Michelle Denise Olvera Hernández, Asela Reig Alamillo p. 10
Context and connective effects on the processing of concessive discourse relations: a VWP experiment
Natalia Talmina, Barbara Landau, Kyle Rawlins p. 12
Pragmatics of spatial language comprehension
Noa Attali, Lisa Pearl, Gregory Scontras p. 14
Navigating ambiguity: The usefulness of context and prosody for naturalistic scope interpretations

---

DAY 2 - June 13

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

Main-2.1. June 13 9:00-10:30 (in person + stream on Zoom)

Josh Knobe p. 16
Dual Character Concepts
Julian Grove, Aaron Steven White p. 17
Modeling the prompt in inference judgment tasks

Main-2.2. June 13 11:00-12:30 (in person + stream on Zoom)

Monica L. Do, James R. Kesan p. 19
Language Production for Source-Goal Motion Events: Factors Affecting Goal Mention
Ankana Saha, Yağmur Sağ, Jian Cui, Kathryn Davidson p. 21
Mandarin demonstratives as strong definites: An experimental investigation

Ugurcan Vurgun, Yue Ji, Anna Papafragou p. 23
Aspectual Coercion: A New Method to Probe Aspectual Commitments

---

Main-2.3. June 13 15:00-16:00 (in person + stream on Zoom)

Anouk Dieuleveut, Ira Noveck p. 25
Devoir, or pouvoir, that is the question
Kristen Syrett, Misha Becker p. 27
Syntactic structure supports the acquisition of emotion and mental state adjectives

---

Main-2.4. June 13 16:30-17:30 (in person + stream on Zoom)

Breanna Pratley, Jed Sam Guevara, Adina Camelia Bleotu, Kyle Johnson, Brian Dillon p. 29
Both Principle B and Competition Are Necessary to Explain Disjoint Reference Effects
Antoine Cochard, Angeliek van Hout, Hamida Demirdache p. 31
"Liz can buy a croissant or a donut… Both together, right?" Distinguishing target Free Choice from non-target Modal AND in Child French

---

DAY 3 - June 14

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

Main-3.1. June 14 9:00-10:30 (in person + stream on Zoom)

Invited Talk: *Elsi Kaiser* p. 33
Experiments in (non-truth-conditional) linguistic meaning: Exploring subjective predicates and perspective-taking
Alexandros Kalomoiros, jacopo romoli, Matthew Mandelkern, Florian Schwarz p. 34
Presuppositions project asymmetrically, unless they don't

Main-3.2. June 14 11:00-12:30 (in person + stream on Zoom)

Thomas Sostarics, Eszter Ronai, Jennifer Cole p. 36
Relating Scalar Inference and Alternative Activation: A view from the Rise-Fall-Rise Tune in American English
Paul Marty, jacopo romoli, Yasutada Sudo, Richard Breheny p. 38
On the salience of linguistic alternatives in the inference task for scalar implicatures
Morwenna Hoeks, Maziar Toosarvandani, Amanda Rysling p. 40
Focus slowdowns arise due to the computation of alternative sets, not unpredictability

---

Main-3.3. June 14 15:00-16:00 (in person + stream on Zoom)

Hayley Ross, Najoung Kim, Kathryn Davidson p. 42
Fake reefs are sometimes reefs and sometimes not, but are always compositional
Yifan Wu, Helena Aparicio p. 44
Disagreements do not automatically raise the standard of precision

---

Main-3.4. June 14 16:30-17:30 (in person + stream on Zoom)

Invited Talk: *Kate Davidson* p. 46
Semantic/pragmatic universals and variation via crosslinguistic experimentation

---

Online Parallel Session Program

DAY 1 - June 12

Parallel-1.Ia. June 12 13:30-14:30 (virtual)

Sebastian Walter, Stefan Hinterwimmer p. 47
An experimental investigation of perspective alignment in gesture and speech
Nitzan Trainin, Einat Shetreet p. 49
'Exhausting' Theory of Mind resources impairs speaker-specific lexical alignment
Stephanie Solt, Roland Mühlenbernd, Mariya Burbelko p. 51
Social meaning and pragmatic reasoning: The case of (im)precision
Chengjie Jiang, Ruth Filik p. 53
Expecting the unexpected: Examining the interplay between world knowledge and context in relatively unconstraining scenarios
Sol Lago, Petra Schulz, Esther Rinke, Elise Oltrogge, Carolin Dudschig, Barbara Kaup p. 55
Insensitivity to truth-value in negated sentences: does linear distance matter?
Muffy Siegel, Florian Schwarz p. 57
Local Accommodation Continues to be Backgrounded

Parallel-1.Ib. June 12 13:30-14:30 (virtual)

Zhuang Qiu, Casey D. Felton, Zachary Nicholas Houghton, Masoud Jasbi p. 59
The Effect of Experimental Paradigms on Scalar Implicature Estimation
Anna Teresa Porrini, Luca Surian, Nausicaa Pouscoulous p. 61
The importance of speaker knowledge and cooperation in priming scalar implicatures
Radim Lacina, Nicole Gotzner p. 63
Only the (informationally) stronger survive: A probe recognition study with scale-mates and antonyms
Benjamin Weissman p. 65
How does a speaker's intent to deceive affect scalar inference and lie judgments?
Casey D. Felton, Masoud Jasbi p. 67
Quantifying Non-Implicature Sources of Disjunction Exclusivity
Yasutada Sudo, Lisa Bylinina, Stavroula Alexandropoulou p. 69
Priming acceptability judgments of NPI any

Parallel-1.IIa. June 12 15:00-15:50 (virtual)

Kurt Erbach, Cornelia Ebert, Magnus Poppe p. 71
Experimental findings for a cross-modal account of dynamic binding in gesture-speech interaction
Hyewon Jang p. 73
A type of sarcasm that current theories fail to explain – evidence from sarchasm
Shirly Orr p. 75
The lying/misleading distinction from the viewpoint of truth evaluators
Oliver Bott, Torgrim Solstad p. 77
Abductive inferences in causal discourse: Evidence from eyetracking during reading
Zarina Levy-Forsythe, Aviya Hacohen p. 79
On a grammaticized lexical count-mass distinction in classifier languages: Experimental evidence from Tashkent Uzbek

---

Parallel-1.IIb. June 12 13:30-14:30 (virtual)

Sebastian Walter p. 81
Indirect discourse as mixed quotation: Evidence from self pointing gestures
Paola Pinzón-Henao, Jennifer Barbosa, Angelina Pasquella, Paul Muentener, Laura Lakusta p. 83
Development of Mechanistic Support Language in Spanish Speakers in Colombia
Anna Pryslopska, Titus von der Malsburg p. 85
Towards a psycholinguistic model of bracketing paradoxes
Emily Sadlier-Brown, Carla Hudson Kam p. 87
Evaluating context-independent meaning in two English discourse particles
Inbal Kuperwasser, Einat Shetreet p. 89
Group membership impact on referential communication

Posters 1 - DAY 2 - June 13 13:30-14:30 (in person)

The in-person Poster sessions are held in the SAIL Room on the ground floor of **Levin**.

Natalia Jardon, Elena Marx, Eva Wittenberg p. 91
Perfect ever after: An empirical investigation of tense-based event construals in English and Spanish
Sven Smeman, Maaike Smit, James A Hampton, Yoad Winter p. 93
Counting uncountables and measuring countables – unpreferred, not ungrammatical
Daiki Asami, Chao Han, Jacob Burger, Deanna Dunlop, Yue Lu, Effah Yahya M Morad, Chenyue Zhao, Arild Hestvik p. 95
'Negation-blind' N400 disappears when priming is controlled
Adina Camelia Bleotu, Mara Panaitescu, Anton Benz, Andreea Nicolae, Gabriela Bilbiie, Lyn Tieu p. 97
Coloring disjunction in child Romanian
Yuli Feng p. 99
On the Interpretational Flexibility of Mandarin Chinese Dabufen
Premvanti Patel, Kristen Syrett, Athulya Aravind p. 101
Using bounds set by modals to investigate the status of partial objects and count nouns
Angela Cao, Aaron White, Dan Lassiter p. 103
Graded Causatives
Chiara Saponaro, Desiré Carioti, Maria Teresa Guasti p. 105
Talking about Distributivity: How Cognitive Factors Influence Children's Language
Cassandra Kim, Ariel Starr p. 107
From words to memory: Evidence of language guiding motion event reconstruction
Andrea Beltrama, Joyce He, Florian Schwarz p. 109
It's not just Imprecision: Stereotypes guide Vagueness Resolution in Implicit Comparisons
Ebru Evcen, David Barner p. 111
Already Perfect: Conditional Statements
Christian Muxica, Jesse Harris p. 113
Context rather than semantic priming drives the early availability of focus alternatives
Leah Doroski, Raquel Montero, Maribel Romero p. 115
Spanish Neg-raising: Always in the mood for Neg-raising, sometimes in the mood for NPIs
Giuseppe Ricciardi, Kevin Reuters p. 117
Exploring the Agent-Relativity of Truth
Shenshen Wang, Chao Sun, Richard Breheny p. 119
Getting to the Truth is More Cognitively Demanding – Another Look at the Role of Working Memory in Negation Processing
Silvia Curti, Desiré Carioti, Maria Teresa Guasti p. 121
Do all Telic-Perfective Sentences (Always) Culminate? An Exploratory Study on Event Culmination in Italian Monolingual Adults.
Eleanor Muir, Simge Topaloglu, Jesse Snedeker p. 123
The Role of Working Memory in Scalar Implicature Computation in ADHD and Non-ADHD Individuals
Mieke Slim, David Barner, Roman Feiman p. 125
Learning the logic in language: Acquiring the meanings of all, every and each
Mingya Liu, Stephanie Rotter p. 127
Semantic and Social Meaning Match: experiments on modal concord in US English
André Eliatamby, Lyn Tieu p. 129
The role of definiteness in ad hoc implicatures
Elizabeth Coppock, David Beaver, Emily Richardson p. 131
Ordering is not ranking: A study of ordinals vs. degree modifiers in nested definites
Sarah Hye-yeon Lee, Anna Papafragou p. 133
Conceptual Signatures of Atomicity Across Languages
Chao Sun, jacopo romoli, Yasutada Sudo, Richard Breheny p. 135
Putting donkeys into context
Fabian Schlotterbeck, Polina Berezovskaya p. 137
In German, 'less'-comparatives must be less ambiguous than 'exactly'-differentials, experimental data shows

Yoolim Kim, Carolyn Jane Anderson p. 139
Parenthesized Modifiers in English and Korean: What They (May) Mean

Posters 2 - DAY 3 - June 14 13:30-14:30 (in person)

The in-person Posters sessions are held in the SAIL Room on the ground floor of **Levin**.

Adina Camelia Bleotu, Andreea Nicolae, Mara Panaitescu, Gabriela Bilbiie, Anton Benz, Lyn Tieu p. 141
A nonce investigation of a possible conjunctive default for disjunction
Andrea Beltrama, Joyce He, Florian Schwarz p. 143
Integrating social information into pragmatic reasoning in real time
Lyn Tieu, Yawovi Godo, Lydia Mei, Andreea Nicolae p. 145
Experimentally investigating the strengthening properties of disjunction in French: When exclusivity meets free choice and ad hoc implicatures
Shaokang Jin, Richard Breheny p. 147
Priming relevant and non-relevant features in metaphorical and literal contexts
Mieke Slim, Roman Feiman, Mora Maldonado p. 149
Priming between universal quantifiers in negated scopally ambiguous sentences
Sarah Hye-yeon Lee, Anna Papafragou p. 151
Cross-domain event primitives are reflected in motion verb learning across languages
Emil Eva Rosina, Kristina Liefke p. 153
Experientiality markers in memory reports: A semantics-pragmatics puzzle
Letizia Raminelli, Desirée Carioti, Jakob Wünsch, Maria Teresa Guasti p. 155
Assessing scalar meaning: a first exploratory study on some Italian focus particles
David Strohmaier, Simon Wimmer p. 157
Contrafactives, learnability, and production
Jonathan Palucci p. 159
Pseudo-scoping out of tensed clauses: cumulation vs. buildups
Irene Mognon, Amber L. Marree, Petra Hendriks p. 161
Reduced sensitivity to underinformativeness? Using a ternary judgment task to assess scalar implicature generation in L2 and L1
Anton Benz, Torgrim Solstad, Oliver Bott, Martin Kahnberg, Andrea C. Schalley p. 163
A conceptual analysis of verbs of pushing and pulling
Ghyslain Cantin-Savoie, Grégoire Winterstein, Denis Foucambert p. 165
The effect of context on the online processing of adversatives: an eyetracking study
Daniel Asherov, Gabor Brody, Vincent Rouillard, Athulya Aravind p. 167
Pragmatics of human-AI communication
Karl Mulligan, Kyle Rawlins p. 169
Identifying QUDs in Naturalistic Discourse
Katherine Howitt, Colin Phillips, Jeffrey Lidz p. 171
4 year old children really do know the strong crossover constraint
Tiana V. Simovic, Craig Chambers p. 173
Pronoun Interpretation Reveals the Robustness and Flexibility of Perspective Reasoning
Laila Johnston, Daniel A. Smits, Ellie Pavlick, Roman Feiman p. 175
The Structure of Ad-Hoc Alternatives
Christiana Moser, Bahar Tarakcı, Ercenur Ünal, Myrto Grigoroglou p. 177
Conceptual and language-specific effects on multimodal recipient event descriptions
EryingQin, Richard Breheny, Chao Sun p. 179
Does 'a couple' pattern with scalars or numbers - Insights from the inference and 'so' tasks
Vic Tianlan Wen, Kirby Conrod, Dan Grodner p. 181
Online Processing of, and Adaptation to, Nonbinary Pronouns
Jennifer Arnold p. 183
Learning discourse patterns through exposure: Mixed input helps identify informative categories
Robin Lemke p. 185
Investigating fragment usage with a gamified utterance selection task

Lilia Rissman, Sebastian Sauppe, Arrate Isasi-Isasmendi, Anna Merin Mathew, Kamal Kumar Choudhary, Susan Goldin-Meadow, Balthasar Bickel p. 187
Do speakers of nominative vs. ergative languages think about Agency in different ways?

Qiawen Liu, Gary Lupyan p. 189
Why is "tree skin" better than "human bark": Semantic centrality predicts asymmetries in metaphorical extensions

Principled and precise links between object naming and object representation in prelinguistic infants

Sandra Waxman
Northwestern University

Abstract:
The power of human language derives not only the from its own precision and complexity, but also from of its intricate links to conceptual representations. But how, and how early, is a link between language and cognition forged in the first place? Long before they produce their first words, infants have already begun to forge this link. They represent objects flexibly, informed not only by whether but how the objects are named. In today's talk, I focus specifically on our flexibility in representing the very same object (e.g., the family dog) as either a unique individual (Rover) or a member of an object category (eg., a dog). This flexibility is supported by language: how an object is named - either as a unique individual or a member of a category - is instrumental to how we represent it. This representational flexibility is available to infants as early as 7 months of age. Moreover, infants' representations of objects – created in the context of naming – are sufficiently robust to support their reasoning about objects in dynamic events, and sufficiently precise to support linguistic analysis.

Semantics at the language-mind interface

Alexis Wellwood
University of Southern California

Abstract:
Standard models in linguistics and philosophy of language suppose that "the meaning relation" in natural language is both compositional and functional, pairing linguistic expressions with their contributions to the truth conditions of the sentences in which they occur. Such models hew quite closely to the primary data on which semantic theories are based—truth value judgments in context—and are compatible with approaches that pair linguistic expressions with their contributions to thought only insofar as sentences determine thoughts. In this talk, I discuss experimental findings on people's understanding of plural comparatives like The red dots are bigger than the blue dots. Minimally, these findings challenge the assumption that the relationship between sentence and thought is functional. More substantively, I take the relevant phenomenology to suggest that (i) even non-specialists expect sentences to (determinatively) express thoughts, but (ii) in fact, sentences merely provide instructions for thought assembly. This discussion highlights the need for a new foundational formal model that can predictively relate morphosyntax to nonlinguistic cognition.

Logically Negative Thoughts without Negaters

Paul Pietroski
Rutgers University

Abstract:
For the purposes of this talk, I'll assume that sentences like 'Aristotle was not dumb' and 'None/Some/All/Most of the dots are purple' correspond to mental sentences of some Language of Thought (LoT) that is available to humans without special training. I'll review some motivations for suspecting that the relevant LoT doesn't have a negation operator that can (i) combine with a complete thought T to form the logically negated thought ~T, or (ii) combine with a mental predicate P to form the complement predicate ~P, which applies to whatever P does not apply to. I'll suggest a different way of thinking about logically negative thoughts. The suggestion invites experimental investigation.

Prediction and integration of discourse-level meaning are functionally related

The relation of prediction and language processing have recently received increasing attention in psycholinguistics [1, 2], with prediction being investigated in semantics and discourse level pragmatics [3]. To date, predictive processing has mainly been investigated indirectly, with critical measures being taken *after* the critical language input had been presented. Especially in EEG studies, ERPs observed after the critical input have been compared between more vs. less predictable conditions [4]. If these post-word differences between conditions are effects of prediction, i.e., of processes executed *before* the presentation of the critical linguistic material, then (i) the effects of these processes should already be observable while predictions are generated, and (ii) the effects before and after the critical words should be found to be related.

In the present study we investigate the processes of discourse level prediction and their relation to language input processing. We visually presented short discourses in German including conditional sentences containing either the conditional connective *if* or *only if.* Within the presented discourses, the conditional sentences with these different connectives allowed for more or less predictable discourse continuations. Consider the following example:

Sentence 1: *Leon besuchte seine Eltern und dachte sich:*
(Leon visited his parents and thought:)
Sentence 2: ***Wenn / Nur wenn*** *die Blumenstrause hubsch 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 Blumenstrause nicht hubsch.*
(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.)

S1 set the scenario context. The conditional sentence S2 contained either *if* or *only if*. After S3, which, in critical trials, negated the antecedent of the conditional in S2, *only if* discourses allowed for a strong prediction of a negated conditional consequent in S4, while bare *if* discourses did not allow for a strongly constrained prediction [5, 6]. S4 finally either negated the consequent of the conditional in S2, containing the critical quantifier *none*, or confirmed it, containing the quantifier *one*. We thus tested a 2 × 2 design, with two levels of conditional and two levels of discourse continuation, disclosed at and by the critical quantifier.

In Exp. 1 we gained first indirect evidence for the differences in predictability of the discourse conclusion presented in S4 in a self-paced reading study presenting 108 discourses like the exemplified one to 29 participants. We found negated quantifiers to be read significantly faster in discourses with *only if* conditionals than in discourses with bare *if* conditionals (**Fig. 1**).

In order to gain more direct evidence for the effects being due to predictive processing, the target processes need to be observed in situ, i.e. before the critical discourse continuation is presented [7, 8]. Measuring participants' EEG signal, and changing the presentation procedure to even-paced visual presentation, we tested 144 items in 38 subjects in Exp. 2. Analyzing subjects' brain responses across trials before the critical quantifier, we observed a significantly increased Prediction Potential (PP) [9], a slowly building negative brain wave before the critical input, in *only if* scenarios as compared to bare *if* scenarios, indicating that subjects built stronger expectations about the upcoming discourse continuation in *only if* scenarios as compared to *if* scenarios (**Fig. 2A**). This finding supports previous linguistic analyses on the semantics of the two conditional connectives. Additionally, in response to the presentation of the critical quantifier, negative quantifiers (*none*) led to significantly decreased P300 responses in *only if* scenarios as compared to *if* scenarios (**Fig. 2B**). These results match the previous effect observed in reading times in Exp. 1, giving reason to assume that discourse continuations containing negative quantifiers were easier to be integrated into the discourse representation after they were made predictable in *only if* scenarios as compared to bare *if* scenarios.

Notably, in the constraining discourse contexts containing *only if*, where strong PPs were observed, the size of the word-induced P300 component in response to both expected and unexpected discourse continuations was found to be predictable by the size of the PP before the critical word (**Fig. 3**). The greater the PP before the onset of the critical word, the greater the word-induced P300 component in response to unexpected, positive quantifiers, but the smaller the P300 in response to expected, negative quantifiers. In other words, the stronger the expectations generated by participants in the constraining context condition (*only if*), the greater the word-induced processing effort for the integration of the new information in cases where the input was unexpected (*one*), and the smaller the processing effort for word-induced discourse updating when the input matched the expectations (*none*).

This is the first work observing the Prediction Potential for predictions on the discourse level, i.e., triggered by predictions across sentences. We find that the observed Prediction Potential and the word-induced P300 are functionally related. The correlations of prediction effort or commitment before the discourse continuation, as indicated by the Prediction Potential, and the processing effort for integration of the presented discourse continuation, as indicated by the P300, are taken as evidence for a direct link between pre-activation of expected discourse continuations and reduced (or increased) costs of input processing. Our results thus demonstrate that the mental processes of discourse understanding are functionally interconnected with processes of discourse prediction.

Figure 1. Reading times in Experiment 1.

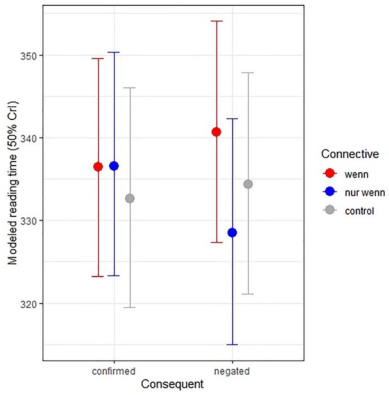


Figure 2. Prediction Potentials (panel A) and word-induced ERPs (panel B) in Experiment 2.

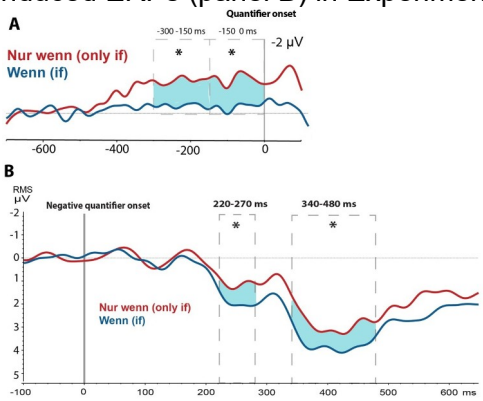
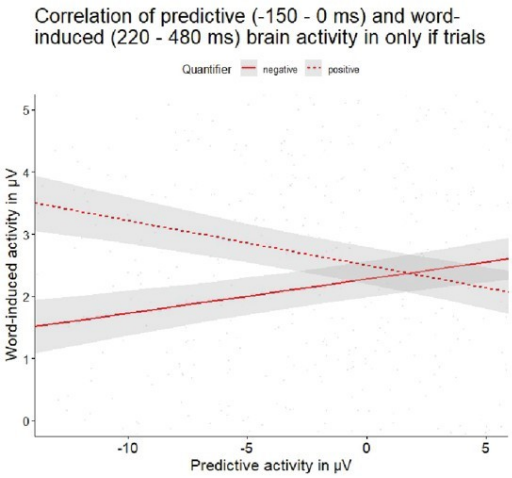


Figure 3. Correlations of Prediction Potential and P300 in *only if* trials in Experiment 2.



References

[1] Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Beh Brain Sci.*

[2] Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *PNAS*.

[3] Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When Peanuts Fall in Love. *J Cogn Neurosci*.

[4] Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting. Ann Rev Psych.

[5] Herburger, E. (2015). Only if: If only we understood it. *Sinn und Bedeutung.*

[6] Herburger, E. (2019). Bare conditionals in the red. *Ling Phil.*

[7] Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs. *Cognition*,

[8] Pulvermüller, F., & Grisoni, L. (2020). Semantic Prediction in Brain and Mind. *TICS.*

[9] Grisoni, L., Miller, T. M., & Pulvermüller, F. (2017). Neural Correlates of Semantic Prediction and Resolution in Sentence Processing. *J Neurosci.*

The effect of standards on scalar implicature processing of gradable adjectives: A web-based eye-tracking study

Properties of measurement scales underlying the meaning of gradable adjectives (e.g., Kennedy, 2007) have been found to affect the availability of pragmatic inferences for these terms (Gotzner et al., 2018). The type of standard value on the measurement scale invoked by gradable adjectives is such a property, which divides gradable adjectives into relative and absolute adjectives: while for relative adjectives the value on the underlying measurement scale that serves as a standard of comparison is contextually determined, for absolute adjectives this is typically a fixed, context-invariant value (Rotstein & Winter, 2004; Kennedy & McNally, 2005). Crucially, it has been argued that scalar implicatures (SIs) of relative adjectives (*warm*$\rightsquigarrow$'warm but not hot') are not derived in all contexts presumably because one needs to be able to resolve the standard for each of the two scale-mates (*warm* vs. *hot*). Absolute adjectives, on the other hand, are more robust SI triggers, especially if the stronger scale-mate is endpoint-denoting (van Tiel et al., 2016), heightening its salience as an alternative for SI computation (Gotzner et al., 2018; Alexandropoulou et al., 2022).

Present study—The present study investigates how the type of standard affects the incremental computation of SIs triggered by gradable adjectives. This will allow us to assess whether lower and upper bounds of gradable adjectives are computed incrementally during compositional interpretation. We build on the visual world (VW) eye-tracking studies by Aparicio et al. (2015, 2018), demonstrating that the processing of relative adjectives hinges on the visual presence of an object (Contrast object) that helps fixing the standard invoked by the relevant adjective (so-called *referential contrast effect* (RCE); cf. Sedivy et al., 1999), whereas the processing of minimum (min) standard absolute adjectives relies solely on linguistic information. Hypothesizing that these semantic differences also factor into the online computation of SIs, we expect to find differential RCEs for relative and min-standard absolute adjectives during incremental interpretation.

Methods—We conducted a web-based eye-tracking experiment using a similar referential communication task to Aparicio et al.'s and the VW paradigm. English native speakers (N=241, recruited from Prolific) were first presented with a visual display of 4 images (see examples in Fig. 1) and 3s later they heard a referring instruction (e.g., *Click on the picture of the warm water with the purple spoon*, Fig. 1(a)/(b)). The instruction is temporarily ambiguous (up to *water* in Fig. 1(a)/(b)) between two referents in the visual scene, i.e., the Target and the Competitor. Importantly, the SI triggered by the adjective in the instruction (*warm*$\rightsquigarrow$'warm but not hot', Fig. 1(a)/(b)) is false of the Competitor, which presents a higher degree of the property encoded by the critical adjective (cf. *warm*). If one were to disambiguate between Target and Competitor by generating the SI associated with the critical adjective of the instruction, this should be reflected in a high(er) proportion of looks to the Target over the Competitor. Participants' task was to click on the correct image after the end of the auditory instruction. Note that the final *with*-PP of the instruction (see Fig. 1) disambiguates the sentence. Participants' eye-movements were collected

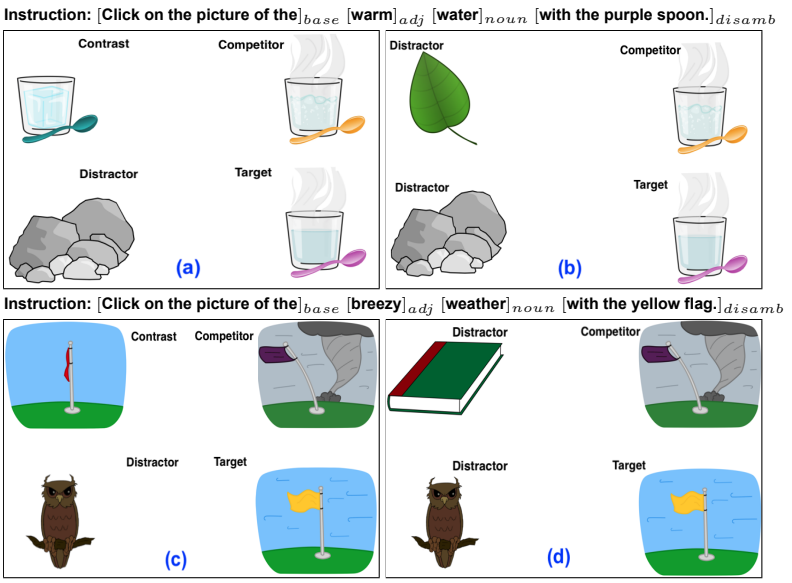


Fig. 1: Example item of relative Horn scale <*warm, hot*> and of min-standard Horn scale <*breezy, windy*> in contrast ((a), (c)) and no-contrast conditions ((b), (d).

from instruction onset until after a selection was made, and were recorded using PCIbex (Zehr & Schwarz, 2018) and the WebGazer.js algorithm (Papoutsaki et al., 2016).

We manipulated the Adjective Type used in the instruction (relative/min-standard) and the presence/absence of a Contrast object in the visual scene (ContrastCond: contrast/no contrast). The Contrast object can be described by the noun (*water*/*weather* in Fig. 1) but not by the adjective of the instruction (*warm*/*breezy* in Fig. 1). We tested 3 relative and 3 min-standard adjective Horn scales (from van Tiel & Pankratz, 2021), the weak scalemate of which has been found by van Tiel & Pankratz to trigger SIs in a picture verification task with pictures like the Competitor images (Fig. 1).

We hypothesize that disambiguation by deriving the SI of the critical adjective of the instruction will be facilitated by processing the comparison standard information of the adjective, and specifically that it will happen differentially for the two adjective types. We predict that disambiguation will be supported by the presence of the Contrast object for relative adjectives, while for min-standard adjectives this should be less likely the case (differential RCE). Therefore, it is expected that participants will fixate on the Target image faster in the contrast condition of relative adjective items than in the respective no-contrast condition, where their looks will be divided between Target and Competitor for longer, whereas such a difference is less likely to be observed between the contrast and no-contrast conditions of min-standard adjectives (Time*AdjectiveType*ContrastCond interaction).

Results—We fit logistic mixed-effects models for three time windows (*adj*(ective), *noun*, *disamb*(iguation)) predicting Target over Competitor looks in terms of time (centered), Adjective Type (sum-coded) and ContrastCond (sum-coded), including the maximal converging random-effect structure justified by our design. Our results revealed a significant 3-way interaction in the *disamb* window (Time*AdjectiveType*ContrastCond: $\beta = 7.62, SE = 2.75, z = 2.78, p < 0.01$), reflecting ongoing processing of ambiguous information. More precisely, this effect reveals that participants converge on the Target faster in the contrast than the no-contrast condition of relative adjectives, while this difference is smaller for min-standard adjectives (see Fig 2).

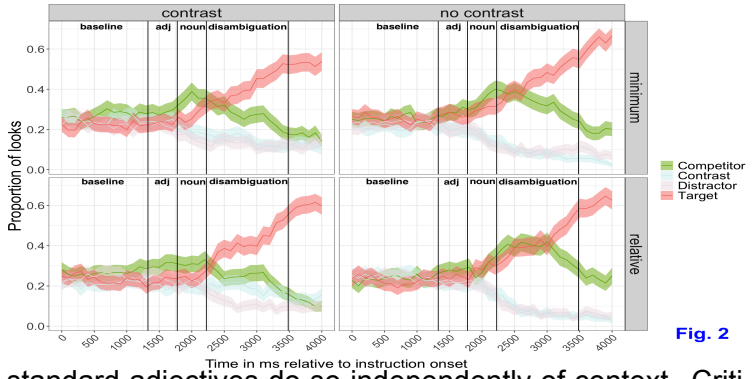


Fig. 2

Discussion—Our finding is in line with our hypothesis: Relative adjectives rely on contextual information to resolve their meaning, while minimum-standard adjectives do so independently of context. Critically, in the contrast condition, the Contrast object lowers the standard for the critical adjective in the relative adjective condition (e.g., *warm*) compared to the no-contrast condition. This happens because in the contrast condition the relevant comparison class includes lower degrees, e.g., of temperature, as compared to the no-contrast condition (see also Barner & Snedeker, 2008; Solt & Gotzner, 2012). Consequently, the degree instantiated by the Competitor is further away from the standard degree for *warm* in the contrast vs. no-contrast condition. In the scalar diversity literature (van Tiel et al., 2016; Gotzner et al., 2018), it is argued that semantic distance is crucial for SI calculation, and a semantically distant alternative to *warm* is highly unlikely to be communicated when uttering *warm*. Hence, when the speaker utters a weak scalar like *warm*, she is more likely to convey that the Competitor degree is excluded in the contrast than in the no-contrast condition.

Overall conclusions—The present study demonstrates that lexical-semantic properties of gradable adjectives are essential to SI processing, and more generally that semantics and pragmatics are highly intertwined during incremental adjective interpretation. We also conclude that web-based eye-tracking may yield fine-grained enough data, advocating for its application in the experimental semantics and pragmatics research.

The 'no-agent' scalar implicature triggered by anticausatives is stronger when the causative alternative is structurally-defined

1. Schäfer & Vivanco (2016) propose that **anticausative** (**AC**) expressions such as in (1a) form scales with their (lexical) causative counterparts as in (2) ($\langle break(y), break(x,y)\rangle$). Under this view, AC statements should exhibit a similar behavior as other items triggering scalar implicatures (Grice 1967, Horn 1972, Gazdar 1979, Noveck 2001 a.m.o): they should be felt less natural in a context fulfilling the stronger alternative (e.g., to describe a broken window and a smiling boy with a slingshot in the hand in front of it), because the AC alternative is too weak in such contexts as it triggers a **'no-agent' scalar implicature** (SI), i.e. an inference that there is no agent involved in the event denoted by the AC (see (1a/b)).

(1) a. The window broke.
b. $\rightsquigarrow \neg$(Someone broke the window)

(2) Someone broke the window.

2. In English, ACs and causatives, however, are not of equal formal complexity: While ACs involve a vP denoting a set of events endured by the theme argument, causatives have on top of this vP a Voice-projection introducing an external argument variable (Kratzer 1996 a.o.). English ACs therefore do not have causative counterparts as *structurally-defined* alternatives (Katzir 2007, Katzir and Fox 2011). Structurally-defined alternatives for a structure ϕ are at most as complex as ϕ. This holds if they obtain via deletion or substitution. More complex structures do not count as alternatives, unless they are salient in the discourse (i.e. are *contextual* alternatives). In English, causatives are therefore at best contextual alternatives of ACs. On this view, (1a) is not expected to trigger the SI (1b), unless a causative statement such as (2) is salient in the context.

3. Languages like French differ from English in that a subset of their ACs receive morphological marking (*se* in French), either optional or compulsory, depending on their morphological class ($-$*se*, $+$*se*, or $\pm$*se* verbs, cf. Doron and Labelle 2011 a.o.). For instance, (1a) is translated in French either as in (3) or (4), as *casser* is a $\pm$*se* verb.

(3) La fenêtre casse/ the window breaks.

(4) La fenêtre se casse/ the window se breaks

Under Alexiadou et al.'s (2015) and others view, the anticausative morphology has no semantic impact; e.g., (3/4) are truth-conditionally equivalent. ACs such as (4) are not semantically reflexive, and both marked and unmarked ACs are logically entailed by their causative counterparts. But they differ in syntactic complexity: while unmarked ACs have no Voice projection just as in English, marked ones involve a Voice projection (cp. (6b) and (6c) on p.2), and are therefore *syntactically* transitive although they have exactly the same inchoative semantics as their unmarked counterparts (Alexiadou et al. 2015, Schäfer 2017). This is because in marked ACs, Voice is semantically expletive: it denotes the identity function and hosts an expletive argument (*se*) in its specifier. On this view, marked ACs and causatives *do* have the same structural complexity; e.g., *Ana casse la fenêtre* 'Ana breaks the window' and (4) both involve a Voice projection on top of vP, see (6a/b). This means that *marked ACs have causatives as structurally-defined alternatives* (unlike unmarked ones). Adopting Katzir's 2007 and Fox & Katzir's 2011 characterization of alternatives, we thus put forward the hypothesis in (5).

(5) The no-agent SI triggered by AC statements is stronger when the corresponding causative statement is a structurally-defined alternative.

4. We tested the hypothesis in (5) through an online acceptability judgement study with native speakers of English and French. Participants (N=80 per language, 70 for French and 63 for English after exclusion) were asked to answer the question *Is the sentence below a natural description of what you see in the pictures?* through a [1-5] scale (1=not at all natural; 5=perfectly natural). Three factors were involved. **Agentivity:** whether the picture representing the change of an object depicts an agent or not ($+$AG vs. $-$AG pictures; see Figure 1b). **Syntactic frame:** whether the

test sentence is a (short) passive (e.g., *The ladder has been tipped over*) or an AC (e.g., *The ladder has tipped over*). We tested short passives of causatives rather than transitive causatives in order to keep the number of overt arguments constant across conditions. **Morphology:** whether the anticausative is morphologically marked or unmarked (relevant for French only). 25 verbs were tested across 50 items in English; 18 +*se* verbs and 9 −*se* verbs were tested across 54 items in French (±*se* verbs were not used to avoid the problem of competition between forms). The visual stimuli were the same across languages. Participants were divided in two groups; all of them saw all pictures, but the pairing between sentence types and pictures was different between groups.

5. Our predictions were as follows. As descriptions of +AG pictures, passives should be fully acceptable. In the same +AG condition, *marked* ACs should be felt infelicitous by participants sensitive to the non-literal meaning of our test sentences: the corresponding causative expressions being structurally-defined alternatives, marked ACs should trigger a clear no-agent SI (clashing with the presence of the agent in the +AG condition). *Unmarked* ACs should be rated better than marked ACs, since the corresponding causative is not a structurally-defined alternative. Furthermore, marked and unmarked ACs should be rated well as descriptions of −AG pictures. Passives were expected to be slightly penalized in the same −AG condition, since the intervention of an invisible agent, although always plausible, needs to be accommodated for the passives to describe −AG pictures felicitously.

6. Our predictions were confirmed by the results (see Fig. 1a). Both in English (left panel) and French (right panel), passives were at ceiling in the +AG condition. AC statements were rated less well than passives as descriptions of the same +AG pictures in both languages (p<.01). But in French, marked ACs receive much lower ratings than unmarked ones, confirming hypothesis (5). In the same +AG condition, the means for unmarked ACs is high but the responses somewhat scattered, which we take to indicate that unmarked ACs do trigger a SI, but a rather weak one. Assuming that passive test sentences play the role of contextual alternatives in our experiment, this confirms that the SI is weaker when the alternative is contextual only. Turning to −AG pictures, ACs received higher ratings than passives in both languages, as expected. Responses for passives were somewhat spread out, suggesting that participants differed in their readiness to accommodate the intervention of an invisible agent.

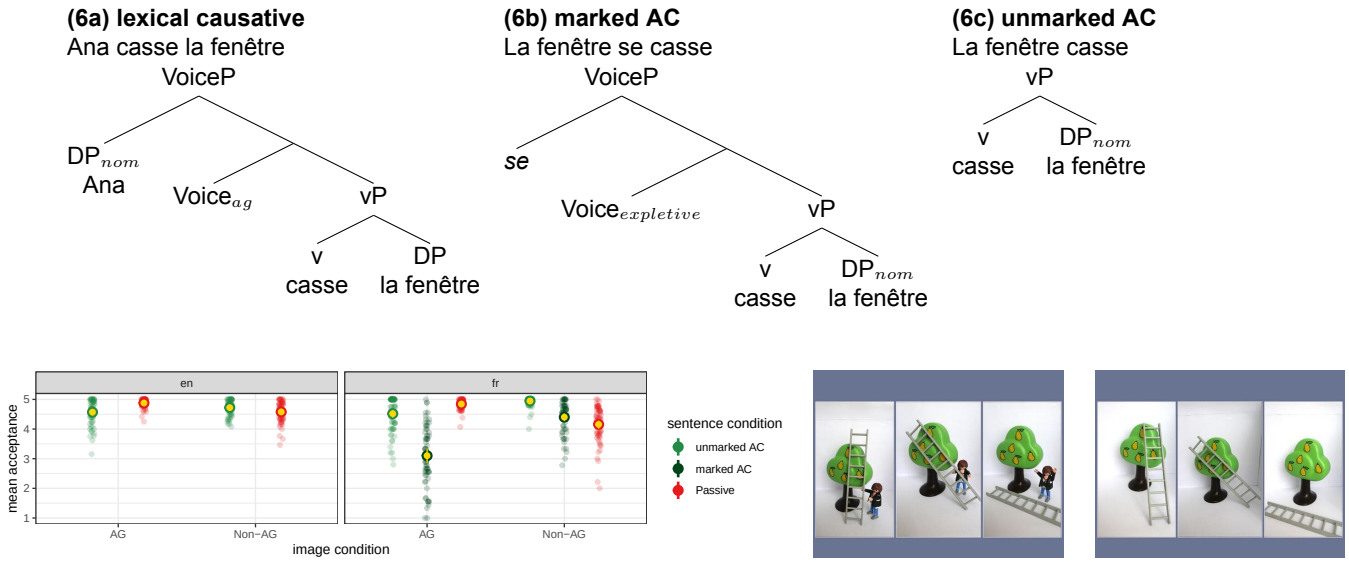


Figure 1: (a) Results: Mean acceptance per person **(b)** Two pairs of visual stimuli (+AG and −AG pictures);

Context and connective effects on the processing of concessive discourse relations: a VWP experiment

Understanding a discourse involves interpreting discursive relations. Concessive relations (a.k.a. negative causal relations) have been proved to be more costly to process than other relations (Xu et al., 2017). Connectives can explicitly mark discourse relations and, by constraining expectations about the upcoming discourse, guide the interpretation and reduce their processing cost (Köhne et al., 2013). In turn, it is known that discourse relations are interpreted in contexts, but how context and connectives interact in the processing of discourse relations is a question that has not been sufficiently addressed. This study investigates how previous context and connectives affect the processing of concessive discourse relations. We address the following research questions: a) Does the presence of a biasing context reduce the cognitive cost of processing a concessive discourse relation? and b) Does the connective have the same facilitating effect independently of the biased or neutral context?

The study consists of a *Visual World Paradigm* experiment, in which 39 Mexican participants listened to 20 stimuli in Spanish with the following form: context sentence; cause sentence; negative consequence sentence (see Fig. 1). Half of the stimuli had a biasing context: it favored the anticipation of the negative consequence (congruent with the Target image); in the other half, the context was neutral (congruent with the Target or Competitor image). Half of the stimuli contained the connector *pero* (but) preceding the negative consequence sentence, and the other half had no connector. Participants listened to the auditory stimuli while looking at four pictures on the screen (Fig. 2): two Distractors, Target (congruent with the heard negative consequence) and Competitor (congruent with the cause sentence). Participants' task was to choose the image that best matched the content at the end of the auditory stimuli. The stimuli were divided into windows for the analysis (Fig. 1). We measured both response times and looking times at the objects in each of the windows, using a Tobii Pro X2-30. The data were analyzed using linear mixed effects regression models (lmerTest package in R (Bates et al. 2015). The models included Object (Target, Competitor, Distractors), Context (biasing / neutral) and Connector (present/absent) as fixed effects and Item and Participant as random effects.

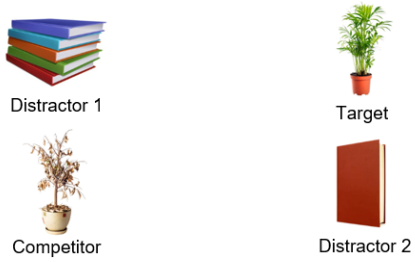
The results are as follows. Response times are significantly affected by the interaction of Context and Connective: items in the condition of neutral context without connective require significantly longer response times than the rest. The eye-tracking data shed light on the integration of both signals: in the Context window, participants discard distractors and, in the Context-extended window, looking times are significantly longer for T vs. C only in the biasing context window, as expected. In the Cause window, and even more clearly in the Cause-extended window, whose linguistic content is compatible with the Competitor, looking times are significantly influenced by Object, Context and their interaction: in the neutral context condition, C receives significantly longer looking times than T, which is practically discarded; in the Biasing context condition, on the contrary, the activation of T -due to the effect of the biasing context- is maintained, therefore, T receives significantly longer looking times than C. Finally, the effect of the connective and its interaction with the biasing context is observed in the Consequence window: T receives significantly longer looking times than C in the condition with connective (with biasing and neutral context), as well as in the condition with biasing context and without connectives. In items without connective and with neutral

contexts, looking times to T and C in the Consequence condition are not significantly different. In these items, the looking preference for T is captured later, after the end of the auditory stimuli. The results indicate that the biasing context reduces the cost of processing a concessive relation, and its facilitating effect is comparable to the effect of an explicit connective. The experiment also sheds light on the online processing of these utterances: the negative consequence in our stimuli, which is preactivated as a result of the integration of the biasing context, remains activated throughout the stimulus, despite the presence of a sentence congruent with the positive consequence. Finally, the facilitating effect of the connective is notorious when there is a neutral context, but it is not perceived in our results when the context has already created the expectation of a negative consequence. The rigid meaning of connectives (Blackmore, 1997) is often assumed to have a constant, pervasive facilitating effect on utterance processing. This study shows that the rigid meaning does not always translate into facilitating effects, as these effects seem to be stronger or present only when the relation itself is difficult to process, but disappears when the relation is intrinsically easy to process (Aragón, 2021) and when other discourse elements already reduce the cost of the relation.

Figure 1. *Structure of auditory stimuli, by Context and Connective*

CONTEXT	**Biasing Context**	**Context-extended**	**Cause**	**Cause-extended**	**Connective**	**Negative consequence**
	1700 ms	700 ms	1700 ms	700 ms	700 ms	1700 ms
neutral	*Esta planta es de Ana* This is Ana´s plant		*Estuvo muchos días al sol* It was in the sun for many days		*(pero)* (but)	*No se secó ni un poco* It did not dry out a bit
biasing	*Esta planta es muy resistente.* This plant is very resistant		*Estuvo muchos días al sol* It was in the sun for many days		*(pero)*	*No se secó ni un poco* It did not dry out a bit

Figure 2. *Example of visual stimuli (Labels not displayed in the experiment)*



References

Aragón, S. G. (2021). *Procesamiento de las relaciones de coherencia causal* [Tesis de maestría, Universidad Autónoma del Estado de Morelos]. Repositorio UAEM.

Blakemore, D. (1997). "Non-truth conditional meaning", *Linguistische Berichte*. 8, 92-102.

Köhne, J., & Demberg, V. (2013). The time-course of processing discourse connectives. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35(35). https://escholarship.org/uc/item/3ng7w640

Xu, X, Chen, Q., Panther, K. U. & Wu, Y. (2017). Influence of Concessive and Causal Conjunctions on Pragmatic Processing: Online Measures from Eye Movements and Self-Paced Reading. *Discourse processes*, 55(4), 387–409. https://doi.org/10.1080/0163853X.2016.1272088

Pragmatics of spatial language comprehension

Spatial prepositions like *in* and *on* are primary tools for describing spatial relations in English. However, their inventory is limited, and speakers must make generalizations about what kinds of relations can be described with the same spatial term [1,2]. Geometric approaches (GA) to the meaning of spatial prepositions propose that the relations between objects are characterized mostly by geometric features, such as direction and distance, with the information about objects and their functions having limited effect [3,4]. In contrast, functional approaches (FA) have argued that functional prepositions *in* and *on* encode rich information about the functions of objects, such as their mechanisms of containment and support [5,6,7]. These views generate distinct predictions: the FA predicts that speakers' intuitions about the acceptability of *in* and *on* depend primarily on the existence of functional relationships between objects, while the GA predicts these intuitions depend on whether a given spatial configuration is geometrically canonical. For example, GA predicts that a description of the type *X is on Y* should be equally acceptable for any objects X and Y as long as X is in a canonical location described by *on* (e.g. supported from below by Y and in contact with Y). If X is a non-canonical location (e.g. it is not in contact with Y), the description *X is on Y* should not be acceptable for any X and Y. On the other hand, FA predicts that the acceptability of *X is on Y* depends not just on the locations of X and Y, but on the properties of X and Y: if Y functionally provides support for X, then the exact configuration of X and Y plays a less important role.

Design. The experiment had a 2 (Scene Type) x 3 (Position) mixed (between/within subjects) design. Participants in the **Real Object** (N=100) condition viewed images of everyday objects, while the **Abstract Shape** (N=100) condition included images of two-dimensional geometric shapes, whose contours and locations matched the images in the Real Object condition. There was a within-subject manipulation of Position: the objects/shapes (for example, a Spanish dictionary on a lamp or a pink square on a green rectangle) appeared in either **Ideal, Competitor, or Distractor** configurations (Figure 1)**.** Ideal configurations depicted objects in the canonical locations described by *in* and *on* (e.g. the Spanish dictionary was supported from below by and in contact with the lamp). Competitor configurations depicted the same objects in non-canonical configurations, with a distractor object now occupying the canonical position (e.g. the Spanish dictionary was not in direct contact with the lamp, but a Calculus textbook was). In Distractor configurations, there was no relationship of containment and support between objects (the Spanish dictionary and the lamp were shown side by side). Each trial showed the same pair of objects across three configurations alongside a description (e.g. *The Spanish dictionary is on the lamp*), and participants had to select the images that fit that description.

Predictions. Both FA and GA predict that Ideal configurations will be acceptable examples of spatial relationships such as *The Spanish dictionary is on the lamp* or *The pink square is on the green rectangle* for both Real Object and Abstract Shape items The predictions of GA vs. FA differed in whether participants would consider Competitor configurations to be acceptable. According to the GA, only the Ideal configurations would be considered acceptable for both Real Object and Abstract Shape trials, as Competitor configurations are geometrically non-canonical. The FA, however, predicts that Competitor configurations will be acceptable in the Real Object condition: since participants have more information about the functional and force-dynamic relationships between the depicted objects, their exact geometric configuration should carry less weight. Therefore, according to the FA, there will be an interaction between Scene Type and Position, such that Competitor scenes will be selected more often in the Real Object condition.

Analysis. We fit a mixed effects logistic regression model with the image choice as the dependent variable, fixed effects of Scene Type and Position and a random intercept for trial number. We found a significant interaction between Scene Type and Position ($\beta = 1.3238$, $SE = 0.3788$, $t = 3.495$, $p = 0.0005$) such that participants selected more Competitor configurations in the Real Object condition.

Discussion. Participants' choices were significantly affected by the Abstract Shape vs. Real Object manipulation: participants were more likely to accept Competitor configurations in the Real Object condition, as predicted by the FA. This suggests that speakers rely on functional and force-dynamic relationships between objects – rather than the geometry of the scene alone – when interpreting *in* and *on*.

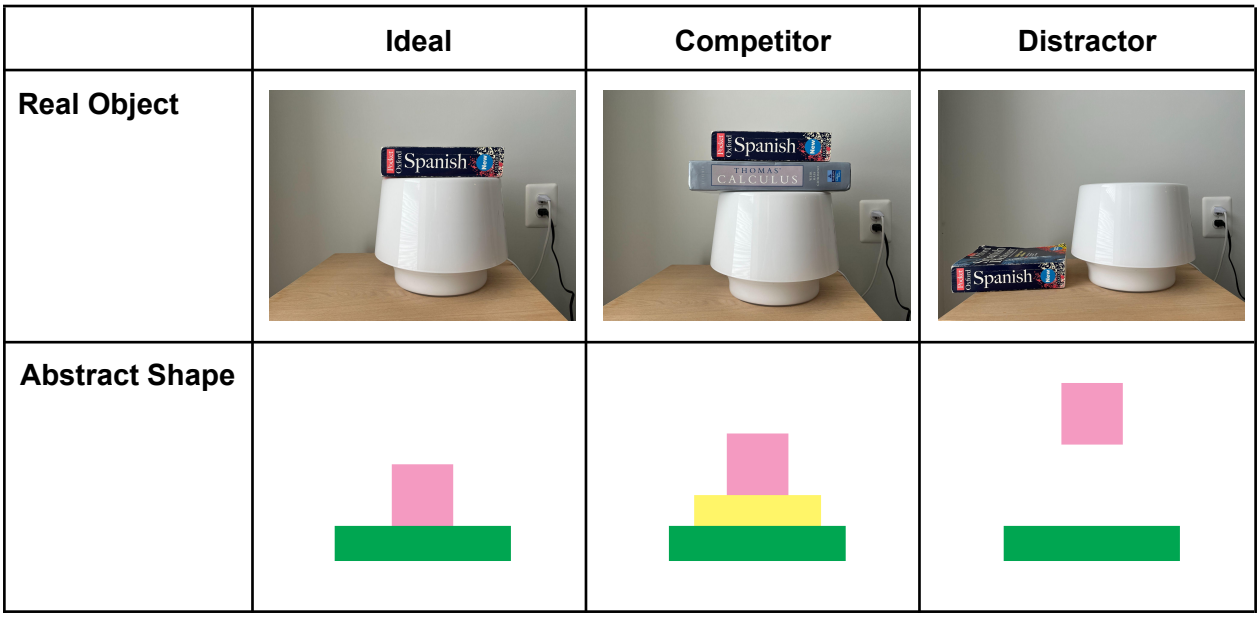


Figure 1. On each trial, participants saw three images in either the Real Object or Abstract Shape condition alongside a description (e.g. *The Spanish dictionary is on the lamp* or *The pink square is on the green rectangle*). Participants were asked to select all images that fit the description.

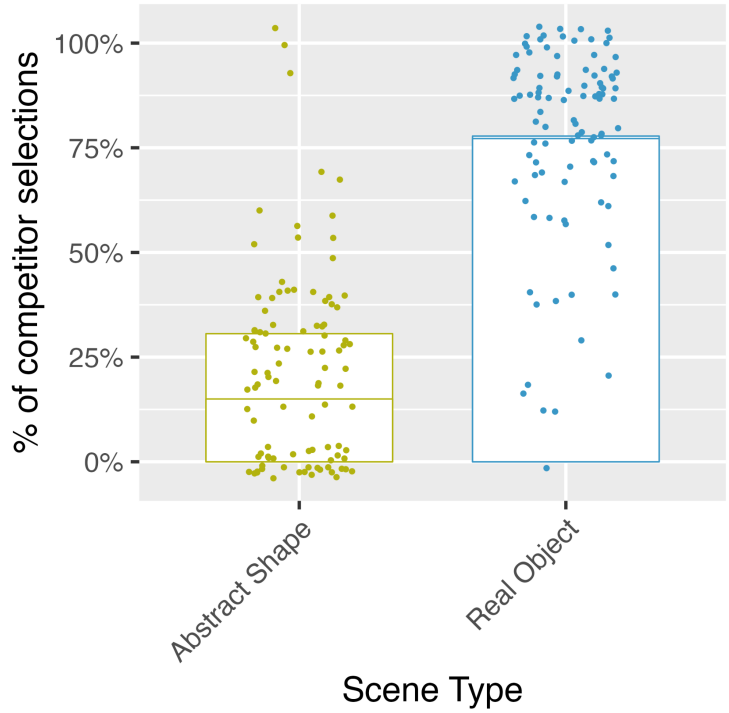


Figure 2. Competitor scenes were selected more often in the Real Object condition (for scenes such as *The Spanish dictionary is on the lamp*) than in the Abstract Shape condition (for scenes such as *The pink square is on the green rectangle*).

References. [1] Talmy, 1985. *Language Typology and Syntactic Description*. [2] Landau & Jackendoff, 1993. *Behavioral and Brain Sciences*. [3] Zwarts, 1997. *Journal of Semantics*. [4] Zwarts & Winter, 2000. *Journal of Logic, Language and Information*. [5] Landau, 2017. *Cognitive Science*. [6] Herskovits, 1986. [7] Coventry & Garrod, 2004.

Navigating ambiguity: The usefulness of context and prosody for naturalistic scope interpretations

Natural languages are full of potentially ambiguous expressions—at least, when these expressions are considered as text out of context—but we seem to be very good at navigating ambiguity and understanding each other. Two key sources of potentially disambiguating information are context and prosody. Our question is, to what extent can listeners use context and prosody to interpret a potentially ambiguous utterance in everyday conversation? We focus on *every*-negation scope ambiguity (e.g., *Every vote doesn't count*) as a case study of ambiguity. In prior work, we gathered naturalistic uses of this ambiguity from conversation recordings. Here, we compare interpretations of these naturalistic uses as text-only, audio-only, text-in-context, and audio-in-context. We find that both context and prosody contribute significant and partially-redundant information.

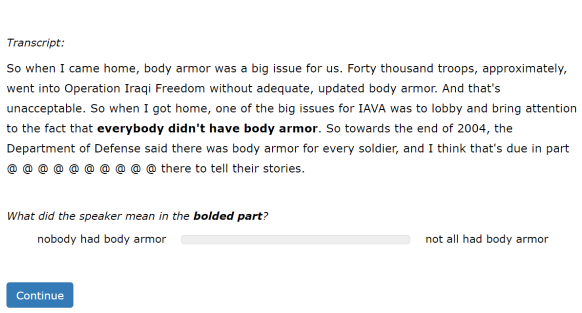
Background. Utterances like *Every vote doesn't count*, with a quantified subject and verb negation, are potentially ambiguous between a surface scope interpretation *every*>*not* (*No vote counts*) and an inverse scope interpretation *not*>*every* (*Not all votes count*). A striking facet of prior research on scope interpretation is both a strong expectation that prosody matters and a lack of clear evidence that it does (e.g., Halliday, 1967; Jackendoff, 1972; Liberman and Sag, 1974; Ladd, 1980; Ward and Hirschberg, 1985; Büring, 1997). A larger question emerges from this body of work about how redundant prosody is with context, since many describe the information provided by prosody as information that might also be provided by context. In one of the only experimental studies investigating prosody, Syrett et al. (2012) found a speaker-specific but no cross-speaker mapping between interpretation and prosody; conversely, listeners show a success rate between 53% and 77% at matching between what they hear and what the speaker intended (Syrett et al., 2014). This weak and variable mapping between prosody and interpretation may be due to many reasons, highlighting the value of understanding the extent of the disambiguating information in both context and prosody of naturalistic data.

Methods. We ran an experiment on Prolific (N=94 monolingual English speakers) to annotate the 63 conversational *every*-negation items collected in past work from radio and TV interviews. Participants judged the speaker's intended meaning on a sliding scale between paraphrases of the item's surface and inverse scope interpretations, in a 2x2 design with factors context (with or without context) and modality (text or audio): each item appeared in each of four conditions (text, audio, text-in-context, audio-in-context). Figure 1a shows an example trial. Each participant judged twenty items (5 randomly-selected items in each of the 4 conditions) in a random order. Between 2 and 15 judgments were collected per item in each condition.

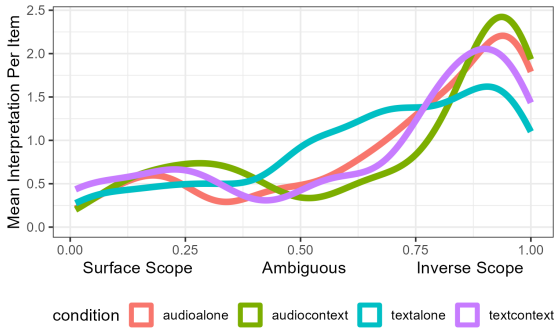
Results. To test the amount of additional information provided by context and prosody, we coded a variable (int-diff) for each item that encodes the absolute value difference in interpretations between the text-only condition and the three other conditions (e.g., for a hypothetical item that received an average interpretation of 0.6—60% inverse—in text-only, 0.8 in text-in-context, 0.9 in audio-only, and 0.9 in audio-in-context, the corresponding int-diff values would be 0, 0.2, 0.3, and 0.3). We then used a mixed effects model predicting int-diff by an interaction of context and modality, with random intercepts for item and participant, using the lme4 package in R (Bates et al., 2015). The main effects of context (β=-0.1455, SE=0.00524, p<2e-16) and modality (β=0.01765, SE=0.005232, p=0.000757) were significant, as was their interaction (β=0.1416, SE=0.007424, p<2e-16). As a measure of the confidence of interpretations in the different conditions, we compared the entropies of the mean interpretations in the four different conditions. We estimated the Shannon entropy, using the entropy package in R (Hausser et al., 2012), of each mean interpretation distribution, where mean interpretations were calculated using the non-parametric bootstrap method from the

Hmisc package in R (Harrell Jr and Harrell Jr, 2019). We found that entropy decreased between conditions in the following order: text-only (5.89) > audio-only (5.86) > text-context (5.83) > audio-context (5.69). Figure 1b shows the four distributions of mean responses per item.

Discussion. In spite of variation, we found that both context and prosody contribute significant information to the interpretations of naturalistic ambiguity, with context providing more confidence than audio, and with the audio information partially redundant with the contextual information. In future work, we investigate more specifically where the disambiguating aspects of context and prosody are redundant with each other. In a previous study that only considered text-in-context interpretations, we identified a specific contextual cue that predicts interpretations; in another study, we identified potential acoustic cues. Future work will test how these contextual and acoustic cues, alone and in interaction, predict interpretations of naturalistic items, using the experimental paradigm we introduce in this study and on the basis of a larger corpus of naturalistic items.



(a) Sample trial of a text-in-context condition.



(b) Distribution of responses.

Figure 1: Sample trial from the experimental task, and the distributions of mean interpretations per item in each of the four conditions (text-only, audio-only, text-in-context, and audio-in-context).

References

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

D. Büring. *The meaning of topic and focus: The 59th Street Bridge accent*, volume 3. Psychology Press, 1997.

M. A. K. Halliday. Intonation and grammar in british english. In *Intonation and grammar in British English*. De Gruyter Mouton, 1967.

F. E. Harrell Jr and M. F. E. Harrell Jr. Package 'hmisc'. *CRAN2018*, 2019:235–236, 2019.

J. Hausser, K. Strimmer, and M. K. Strimmer. Package 'entropy'. *R Foundation for Statistical Computing: Vienna, Austria*, 2012.

R. S. Jackendoff. Semantic interpretation in generative grammar. 1972.

D. R. Ladd. The structure of intonational meaning (bloomington), 1980.

M. Liberman and I. Sag. Prosodic form and discourse function. In *Chicago Linguistics Society*, volume 10, pages 416–427, 1974.

K. Syrett, G. Simon, and K. Nisula. Prosodic disambiguation of scopally ambiguous sentences. In *Proceedings of the Meeting of the North East Linguistic Society*, volume 43, pages 141—152. GLSA (University of Massachusetts), 2012.

K. Syrett, G. Simon, and K. Nisula. Prosodic disambiguation of scopally ambiguous quantificational sentences in a discourse context. *Journal of Linguistics*, pages 453–493, 2014.

G. Ward and J. Hirschberg. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, pages 747–776, 1985.

Dual Character Concepts

Josh Knobe
Yale University

Abstract:
Imagine a person who has a job as a physics professor, but who never changes her beliefs in light of empirical data and is always trying to prop up a preconceived dogma. Is it right to say that this person is a *scientist*? In cases like this one, people often feel torn. They say that (a) there is clearly a sense in which this person is a scientist, but also (b) in a deeper sense, this person is not a scientist at all. Examples like this one suggest that people have two different criteria for determining whether a person counts as a scientist, and concepts of this type are therefore known as "dual character concepts." Over the past ten years or so, experimental research has led to many important findings about dual character concepts. In this talk, I will be reviewing this research and exploring some of the unresolved theoretical questions that arise out of it.

Modeling the prompt in inference judgment tasks

Introduction. A major question in the literature on presupposition projection is whether factive inferences (e.g., *Jo {loves, doesn't love} that Mo left* $\rightsquigarrow$ *Mo left*) are necessary, as classically assumed (Kiparsky and Kiparsky 1970; Karttunen 1971), or not (Tonhauser, Beaver, and Degen 2018). Recent work by Grove and White (2023) addresses this question by fitting statistical models encoding these two assumptions about factive inferences to inference judgment data aimed at capturing factive inferences' strength (Degen and Tonhauser 2021). Grove and White find that models characterizing factive inferences as necessary (henceforth, *discrete models*) fit the inference judgment data better than models that assume they are not (*gradient models*).

Contribution 1. We address a potential flaw in Grove and White's use of Degen and Tonhauser's data for comparing their models: the way participants were asked to respond may artificially improve the discrete models' performance. With the aim of putting the discrete and gradient models on more equal footing, we present two new datasets that keep all other aspects of Degen and Tonhauser's materials constant but which manipulate the natural language prompt participants are given. Consistent with Grove and White 2023, we find that discrete models fit the data better than gradient models for both datasets, supporting Grove and White's claim.

Contribution 2. We show that jointly modeling both the compositional semantics of the target sentence—i.e., the sentence containing the presupposition trigger—and the compositional semantics of the natural language prompt within Grove and White's framework substantially improves fit to response distributions. This finding suggests that it is important to model the interaction between the meaning of a target sentence and the meaning of a prompt when analyzing experimental data.

Degen and Tonhauser's data. Degen and Tonhauser provide experimental participants with a background fact, paired with a predicate taking a complement clause related to that fact.

(1) a. **Fact (which Elizabeth knows):** Zoe is a math major.
Elizabeth asks: "Did Tim discover that Zoe calculated the tip?"
b. Is Elizabeth certain that Zoe calculated the tip?

Participants are asked to provide an answer to the prompt in (1b) on a sliding scale with 'yes' on the left and 'no' on the right. Degen and Tonhauser collect responses for twenty clause-embedding predicates taking one of twenty possible embedded clauses, each paired with either a "high prior" fact or a "low prior" fact. ((1a) illustrates the high prior fact for the given clause.)

Grove and White's models. The aggregate measures of different predicates' factivity derived from inference judgment data show substantial gradience (White and Rawlins 2018; Degen and Tonhauser 2022), and hence constitute potential evidence for variation among predicates in the strength of such inferences. Grove and White ask if this gradience arises due to *metalinguistic uncertainty*—uncertainty about whether a predicate is factive or not—or *contextual uncertainty*—uncertainty inherently associated with predicate meanings. If the uncertainty is metalinguistic, factive inferences may nevertheless be discrete; different predicates would in turn differ in the frequencies with which they trigger such inferences. If it is contextual, predicates would license inferences with varying degrees of certainty, similar to the manner in which a vague predicate, such as *tall*, can license uncertain inferences about the heights of individuals of which it is predicated.

Grove and White fit four models to Degen and Tonhauser's data, varying whether uncertainty about either background world knowledge or factivity is encoded as metalinguistic or contextual. Their models are the *discrete-factivity* model (DF), which regards uncertainty about factivity as metalinguistic and uncertainty about world knowledge as contextual; the *wholly-gradient* model (WG), which regards both kinds of uncertainty as contextual; the *discrete-world* model (DW), which regards uncertainty about factivity as contextual and uncertainty about world knowledge as on a

par with metalinguistic uncertainty; and the *wholly-discrete* model (WD), which regards both kinds of uncertainty as (on a par with) metalinguistic uncertainty. They find that DF performs the best, as assessed by expected log pointwise predictive densities (ELPDs), lending support to the classical view of factivity as a fundamentally discrete phenomenon.

While Grove and White's results are promising, they are consistent with the possibility that the nature of the question prompt exemplified in (1b) biases experimental participants toward making discrete 'yes' or 'no' judgments, even while the contribution to inference judgments made by factive predicates may be gradient. Because the prompt in (1b) is a polar question, and 'yes' and 'no' label the slider response, participants may effectively treat their response as a binary forced choice by providing an answer near 'yes' if they are sufficiently certain about the relevant inference, and an answer near 'no' if they are not. If so, an *a priori* advantage is conferred on models regarding the contribution to inference of factive predicates as discrete and, thus, models which regard uncertainty about factive inferences as metalinguistic. Our manipulations of the prompt address this concern, while our new models explicitly target the semantics of the question prompt.

Varying the prompt. We conduct two experiments identical to Degen and Tonhauser's, but which vary the prompt. In both, participants are provided with a *degree* question, which is either about the speaker's degree of certainty (2a) or degree of *likelihood* that the speaker is certain (2b).

(2) a. How certain is Elizabeth that Zoe calculated the tip?
b. How likely is it that Elizabeth is certain that Zoe calculated the tip?

The prompt in (2a) was paired with a slider labeled 'not at all certain' on the left and 'completely certain' on the right, while the prompt in (2b) was paired with 'impossible' and 'definitely'.

Modeling. We obtained the Stan code used to fit each of the four models of factivity from Grove and White, and we constructed two additional models which extend DF, in order to implement a semantics for *certain* and *likely* which allows them to attend to distinct lexical scales. Specifically, to model the prompt in (2a), we assume that the degree introduced by *certain* ranges over degrees of *confidence* rather than degrees of probability (following, e.g., Klecha 2012), and thus that its scale is truncated relative to that of *likely* (yielding the *discrete-factivity-certain* model (DF+C)). To model the prompt in (2b), we assign a semantics to *likely* on which it introduces a degree corresponding to a *probability*, and where this degree is computed based on the corresponding semantics for *certain* (yielding the *discrete-factivity-likely-certain* model (DF+LC)).

Results. We compare the (rounded) ELPDs (s.e. in parentheses) of the four original models of Grove and White with our models of the prompts in (2), each fit to the two new datasets.

Experiment	n	DF+C	DF+LC	DF	WG	DW	WD
(2a)	285	2466 (67)	2360 (64)	2183 (65)	1653 (66)	1837 (63)	2000 (56)
(2b)	292	2064 (56)	2052 (56)	1966 (57)	1821 (60)	1524 (48)	1540 (44)

Among the original models, DF continues to perform the best on both datasets. Meanwhile, we find that DF+C performs the best on the dataset containing the prompt in (2a), as expected, while DF+C and DF+LC perform about equally on the dataset containing the prompt in (2b).

Conclusions. Our results (i) confirm that the model comparisons obtained by Grove and White do not reflect an *a priori* bias conferred on the discrete models by the experimental task, but rather these models' abilities to capture the distributions of degrees of certainty associated with the inferences generated for the predicates and complement clauses tested; and (ii) suggest that it is important to develop explicit, semantically-motivated linking hypotheses when modeling inference data, not only about the nature of the natural language expression under investigation, but about the question prompt used to elicit an inference. Future research in this line will aim to understand why the model of the prompt in (2a) performs equally well on the dataset containing (2b).

Language Production for Source-Goal Motion Events: Factors Affecting Goal Mention

Anonymous Otter 1 (Otter School) & Anonymous Otter 2 (Otter School)
{Otter 1 email}

When describing an event in the world, how do people decide what to mention and what to omit? One factor is audience design: speakers tend to omit what's already known or highly inferable to listeners and mention what's unknown. However, recent work investigating descriptions of source-goal motion events (e.g., an octopus$_{FIGURE}$ swimming from a treasure chest$_{SOURCE}$ to a coral reef$_{GOAL}$), found that while factors related to audience design could dramatically affect the mention/omission of sources; goals – surprisingly – were mentioned whether they were or were not already known to addressees.[1,2] These studies suggest that pragmatic factors related to audience design do not affect message generation for conceptually core event components (i.e., goals, [3-5]) versus conceptually peripheral event components (i.e., sources) in the same way.

Exp1. (n=61) aims to replicate the surprising goal results from [1] using a design that more clearly eliminates ambiguity about the knowledge state of the addressee: we explicitly told speakers in Goal Common Ground (GCG) conditions that addressees would be shown only the last frame of the event on a separate display (Fig1). Speakers in No Common Ground (NCG) conditions where told addressees could not see any part of the event. Prior work has shown that the conceptual status of goals in events with animate (e.g., octopus) versus inanimate (e.g., pirate flag) figures does differ.[6-8] So, animacy of the figure in motion as also varied between-subjects.

Results showed that speakers mentioned goals upwards of 95% of the time – surprisingly, even (i) in GCG conditions, where they were already known to interlocutors and (ii) in Inanimate conditions, where goals are not considered conceptually core ([6-8], Fig2). This pattern was not driven by insensitivity to the knowledge state of the addressee: speakers in GCG conditions used significantly more definite determiners than those in NCG conditions (Fig3; β = 6.20, SE = .85, |z| = 7.29). Thus, in line with [1], audience design did **not** affect speakers' decisions about whether to mention/omit goals (e.g., during message generation); but did determine how they talked about them (e.g., during linguistic encoding). As such, **Exp1b** asked whether goal mention was driven in part by the need to convey the telicity of the event (e.g., "The octopus swam from the treasure chest" describes a different, atelic event). We re-analyzed GCG utterances from Exp1 and found that in roughly 70% of utterances telicity was only inferrable via goal mention. This suggests that communicating telicity is one reason speakers in both Animate and Inanimate GCG conditions still mentioned even pragmatically uninformative goals.

Exp2. asked why speakers didn't produce telic descriptions like "the octopus {came, swam over} from the lamppost". **Exp2a** tested the possibility that doing so requires speakers to not only be aware of addressees' knowledge states, but also to put themselves in the 'cognitive shoes' of the addressee. We made addressee perspective more salient using the GCG-Shared condition: speakers (n=16) watched the event and with the last frame still visible, turned their computer screen towards the addressee, then described the event from the same physical perspective as the addressee. Contra a perspective-taking account, goal mention rates were no different in GCG-Shared versus Exp1 GCG conditions ($p > .4$) for Animate and Inanimate events. **Exp2b** is ongoing and tests the possibility that goal mention may also depend on whether the manner of motion (e.g., swim vs float vs go) is also pragmatically important to mention to addressees.

Conclusions: Goals are resilient to pragmatic factors because they communicate multiple, core aspects of an event that are otherwise uninferrable to addressees – including (but not limited to) the intentionality of the figure in motion [3-5], and the telic nature of the event. These results shed light on why some event components are less sensitive to pragmatic factors than others. They also bear on the relationship between non-linguistic versus linguistic representations of animate and inanimate source-goal events. Finally, we discuss implications of other exploratory analyses (e.g., order of goal vs source mention) that point to other differences in the way that people talked about animate versus inanimate motion events.

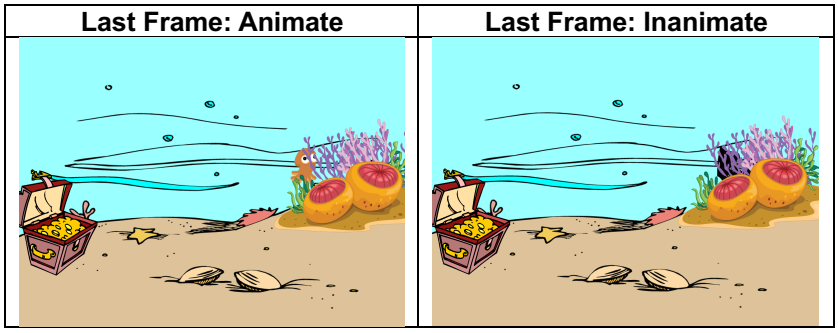


Fig1 Sample stills showing the last frame of the animate (octopus) item and corresponding inanimate (flag) item. Source and Goal arrows shown here were not visible to participants.

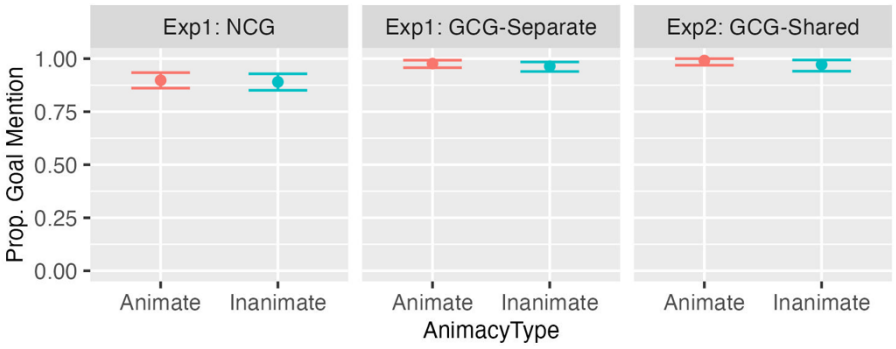


Fig1 Proportion of Goal Mentions in Exps. 1 & 2. Error bars show +/- 1 SE. In Exp1: NCG, addressees saw no part of the event. In Exp1: GCG-Separate, they saw the last frame of the event on their own separate computer screen. In Exp2: GCG-Shared, they saw the last frame on the speaker's computer screen.

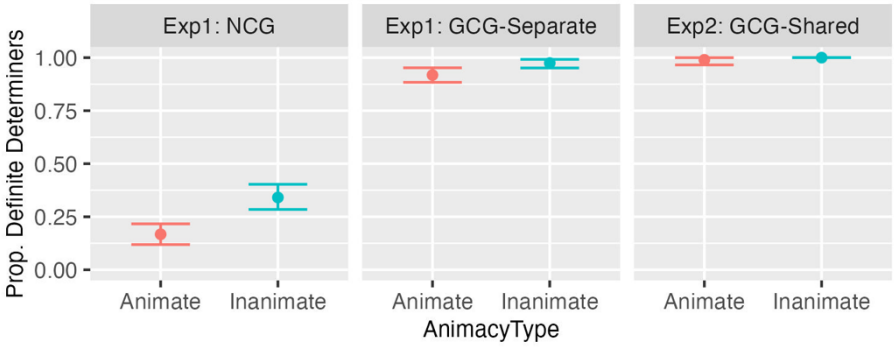


Fig2 Proportion of definite determiners used when referencing goal landmarks in Exps. 1 & 2 with error bars showing +/- 1 standard error.

References

[1] [Redacted] (2020). Cognitive and pragmatic factors in language production: Evidence from source-goal motion events. Cognition, 205, 104447.
[2] [Redacted] (2022). Encoding Motion Events During Language Production: Effects of Audience Design and Conceptual Salience. Cognitive Science, 46(1)
[3] Regier, T., & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. Cognitive Science, 31, 705–719.
[4] Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. Cognition, 96(1), 1–33.
[5] Lakusta, L., Wagner, L., O'Hearn, K., & Landau, B. (2007). Conceptual Foundations of Spatial Language: Evidence for a Goal Bias in Infants. Language Learning and Development, 3(3), 179–197.
[6] Lakusta, L., & Landau, B. (2012). Language and Memory for Motion Events: Origins of the Asymmetry Between Source and Goal Paths. Cognitive Science, 36(3), 517–544.
[7] Lakusta, L., & Carey, S. (2015). Twelve-Month-Old Infants' Encoding of Goal and Source Paths in Agentive and Non-Agentive Motion Events. Language Learning and Development, 11(2), 152–175.
[8] [Redacted] (2023). Conceptual and pragmatic factors influencing the representations of core event components. Poster presented at AMLAP 29, San Sebastian, Spain.

Mandarin demonstratives as strong definites: An experimental investigation

This study argues based on new experimental data that Mandarin demonstratives exhibit strong definiteness in a manner not observed with standard demonstratives (e.g. in English) (Jenks 2018).

Definiteness in Mandarin. Building on Schwarz (2009, 2013), Jenks (2018) proposes that Mandarin, a determinerless language, lexically distinguishes uniqueness-based, i.e., *weak* (Frege 1892, Russell 1905), and anaphoric, i.e., *strong* (Heim 1982, Roberts 2003) definites—bare nouns are used for a unique referent in a situation and demonstratives establish anaphoric links to an existing discourse referent, as in (1a), with the exception of subject positions, where bare nouns are felicitous as anaphors since they are continuing topics (not due to being strong definites). Dayal & Jiang (2022), with a different follow-up to (1) as in (1b), claim that Mandarin bare nouns are felicitous in both uniqueness and anaphoric contexts and demonstratives are standard demonstratives.

(1) Jiaoshi li zuo zhe yi ge nansheng yi ge nüsheng
classroom inside sit prog one cl boy one cl girl
'There is a boy and a girl sitting in the classroom.'

a. Wu zuotian yudao #(**na ge**) **nansheng**
I yesterday meet that cl boy
'I met the boy yesterday.'

b. **Nüsheng** zuo zai **nansheng** pangbian.
girl sit dur boy side
'The girl was sitting next to the boy.'

Dayal & Jiang (D&J) link the contrast between (1a) and (1b) to the situations invoked by the follow-up sentences. When the initial situation in (1) remains unchanged, speakers opt for the simpler of two felicitous options, the bare noun (1b). If the situation expands (1a) (e.g., including a new participant), the demonstrative is preferred, as bare nouns might become infelicitous if the extended situation fails the uniqueness requirement of the definite. Demonstratives, though, would remain felicitous, as they have an anti-uniqueness requirement (*the sun* vs. *#that sun*, e.g., Robinson 2005), which can be satisfied in a wider situation.

Anaphoric demonstratives. Experimental work has shown that the acceptability of anaphoric demonstratives (vs. definites) depends on both the situation extension in the follow-up sentence and the number of discourse referents (NPs) introduced initially. Saha (2023) and Saha *et al*. (2023) obtained acceptability judgments from one language with determiners (English) and two determinerless languages (Turkish, Bangla) encoding definiteness distinctly: Turkish via bare nouns, Bangla by preposing the NP before the classifier. Context manipulated situation (**same** (2a) vs. **new** (2b)) and number of NPs (**one vs. two**):

(2) {[$_{OneNP}$ A boy]/ [$_{TwoNP}$ A boy and a girl]} entered the classroom.
a. The/That boy sat down in the front row.
b. I had noticed the/ that boy at a coffee shop yesterday.

Across all these languages (English and Turkish in Saha *et al.* 2023, and Bangla in Saha 2023), definites were near ceiling in these contexts and rated significantly higher than demonstratives, while the acceptability of demonstratives varied significantly and were highest in One NP contexts and in New Situations (Fig 1). Saha *et al.* (2023) accounts for this pattern by adopting a focus-driven information structural approach to demonstratives. Following insights from Schwarz (2009) and D&J (2022), they assume that that anaphoric definites and demonstrative descriptions are similar in including an anaphoric index argument, and they argue that demonstratives essentially differ from definites in evoking focus alternatives on the index argument, ((3b) vs (3c)).

(3) a. the boy (no focus with DP): $[\![[\text{DEF } 1\ ]\ \text{boy}]\!]^o = \iota x[boy(x) \wedge x = g(1)]$ e.g. 1 NP cases
b. the BOY (as opposed to the girl) e.g. 2 NP cases
$[\![[\text{DEF } 1\ ]\ \text{boy}_F]\!]^f = \{\iota x[boy(x) \wedge x = g(1)], \iota x[girl(x) \wedge x = g(2)]\}$
c. THAT boy (as opposed to another boy) e.g. 1 NP, New Situation cases
$[\![[\text{DEM } 1_F\ ]\ \text{boy}]\!]^f = \{\iota x[boy(x) \wedge x = g(1)], \iota x[boy(x) \wedge x = g(3)]\}$

Our Study: Design & Methods. We adapted the experimental paradigm in Saha (2023) and Saha *et al.* (2023) to Mandarin to test contrasting claims in Jenks (2018) and D&J (2022). The acceptability of definites vs. demonstratives were tested across 12 scenarios varying both subject/object position and animacy. Participants (N=64) read short scenarios and were presented with two possible continuations after each, one with a demonstrative and one with a bare noun (order counterbalanced), and rated the acceptability of each continuation using a slider bar. Scenarios varied between participants in a 2x2x2 Latin Square design by number of discourse referents (one vs. two) and situation (same vs. new) [See (4)]. New situations introduced a new participant (e.g. speaker or someone else) and a temporal change from the initial situation.

(4) {[$_{1NP}$ yi ge nanhai]/ [$_{2NP}$ yi ge nanhai he yi ge nvhai]} zoujin le jiaoshi.
one cl boy one cl boy and one cl girl walk.into perf classroom
'A boy/A boy and a girl walked into the classroom.'

a. {∅/na ge} nanhai zuozai qianpai.
∅/that cl boy sit.at front.seat
'The/That boy sat at the front.'

b. wo zuotian zai shudian jian guo {∅/na ge} nanhai.
I yesterday at bookstore see perf ∅/that cl boy
'I saw the/that boy at the bookstore yesterday.'

Results & Discussion. The data was fit with a mixed effects linear model in R, which found a main effect of demonstratives rated significantly higher than definites across the board (micro-variations in ratings for subject vs. object positions were not checked for) with no significant effect of either Situation or number of NPs. Within definite responses, we found a main effect of situation: Definites were significantly more acceptable in Same Situation follow-ups (Fig. 1). The strong preference for demonstratives as anaphors supports Jenks' claim of strong definiteness (contra D&J 2022). However, in line with D&J, definite bare nouns are also felicitous (though less preferred) in anaphoric contexts. *Demonstratives:* The contrast of the Mandarin data against the consistent patterns found in English, Bangla, and Turkish establish that anaphoric demonstratives in Mandarin do not behave like demonstratives but pattern more closely with anaphoric definites in these languages. We suggest that Mandarin demonstratives allow for the absence of focus on the index, akin to (3a) and (3b), unlike standard anaphoric demonstratives, *e.g.,* (3c). *Definites:* We see an effect of situation in the relative acceptability of anaphoric definites; they are less preferred in New Situations, as claimed by D&J, although definites do not surpass demonstratives in acceptability within Same Situations. We argue that this stems from the ability of Mandarin sentences with bare nouns to also have generic readings due to lack of tense and aspectual marking, as well as indefinite readings for postverbal bare nouns (e.g., Cheng & Sybesma 1999). Demonstratives would be unambiguously anaphoric, driving their preference across the board. In Same Situations, there is a bias towards referring to the entities introduced previously, so definites fare better as anaphors in Same (vs. New) Situation.

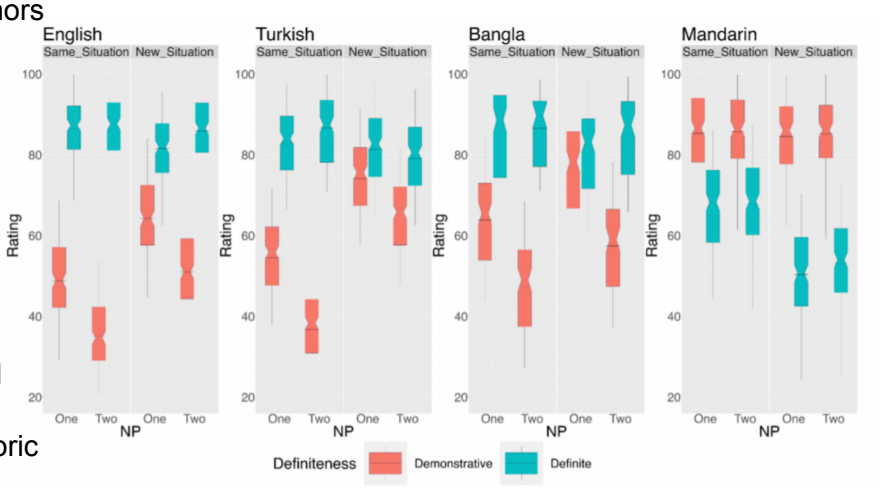


Figure 1: Anaphoric Definites vs Demonstratives: English, Turkish (Saha *et al.* 2023), and Bangla (Saha 2023) vs Mandarin (our present study)

Aspectual Coercion: A New Method to Probe Aspectual Commitments

1. Introduction. Aspectual theories in semantics distinguish telic verb phrases denoting bounded events with an inherent endpoint (e.g., *draw a balloon*) from atelic verb phrases denoting unbounded events that lack an inherent endpoint (e.g., *do some drawing*; Krifka 1998; van Hout 2016). Aspectual coercion occurs when a sentence combines elements that mismatch in aspectual terms (e.g., the otherwise telic VP *draw a balloon* with a durative adverbial such as *for 10 seconds* or the atelic VP *do some drawing* with a delimited adverbial such as *in 10 seconds*). Aspectual mismatches force a reinterpretation to align temporal expectations (Jackendoff 1991; Moens & Steedman 1988). Previous research on aspectual coercion has yielded divergent results in terms of whether coercion incurs processing costs (Bott 2010, Dölling 2014, Pickering et al. 2006, Piñango et al. 1999); however, these studies primarily relied on reaction times during reading or lexical decision tasks, and have not fully addressed whether aspectual coercion leads to a true shift in event understanding (i.e., a genuine commitment to a coerced interpretation). Using a novel paradigm, here we measure event perception while viewers have to verify coerced and non-coerced aspectual sentences against dynamic visual events. This method, informed by findings on real-time event apprehension (Ji & Papafragou, 2022), reveals how people interpret sentences with mismatched linguistic aspectual cues and use them as a zoom lens to process visual events.

2. Stimuli. We created 21 videos, each featuring a woman performing an action (e.g., drawing a balloon, mean: 10.4 sec, range: 6.5-14.8 sec). Preliminary studies showed that these videos were perceived as bounded – i.e., having an inherent endpoint. Each video was edited to include a 30 ms visual interruption (one frame removed from the timeline) at either the midpoint (50%) or a late point (80%) of the action. One-third of critical events had a midpoint interruption, another third had an endpoint interruption, and the rest, as control items, had no interruptions. The logic of interruption placement is explained below.

3. Experiment 1 - telic to atelic coercion. 192 monolingual English speakers on Prolific saw a scenario where a woman, post-surgery, performed various exercises for motor skill recovery. Each trial began with a sentence describing the exercise. Three between-subjects conditions were based on the type of sentence: 'Telic' (e.g., 'Ebony should draw a balloon'), 'Telic+IN'' ('…draw a balloon in 10 seconds'), and 'Telic+FOR', or coerced atelic ('…draw a balloon for 10 seconds'). Following the video, participants were asked whether the actor did the exercise (where answers for critical items should always be *Yes*). We found that participants in all conditions indeed gave *Yes* answers to that question (Telic: 98.6%, Telic+IN: 97.8%, Telic+FOR: 94.9%). Participants were also asked whether there was a glitch in the video. Answers served as a key metric: they indicated whether participants' perception of event boundaries was influenced by the aspectual framing of the sentences, thus providing a direct link between the linguistic aspect and cognitive event processing. The placement of interruptions was crucial, as previous research has shown that, for events perceived as bounded, interruptions at late points are more likely to be missed compared to midpoints, while for unbounded events, interruption detection remains consistent across the timeline (Ji & Papafragou 2022). This is because, for bounded event construals, endpoints are important and attract attention, thereby causing failures to detect external distractors such as interruptions. Unbounded event construals, however, by definition have no canonical endpoints so there is no difference in attention allocation between midpoints and endpoints. Here, if a sentence with a coercive adverbial successfully elicited a coerced reading, the interruption detection

performance was expected to pattern with the new reading. Thus, participants in 'Telic' and 'Telic+IN' conditions were expected to perceive events as bounded, and thus be more likely to miss late-point interruptions. In contrast, participants in the 'Telic+FOR' (i.e., coerced Atelic) condition should interpret events as unbounded, and not incur a similar cost for late-point interruption detection. Indeed, we found a significant interaction between Condition and Interruption type (χ^2 = 8.39, p = 0.0151) (see Fig. 1). In both Telic and Telic+IN conditions, a marked difference in detecting midpoint and late interruptions was found (Telic: odds r. = 0.69, p = 0.04; Telic+IN: odds r. = 0.55, p = 0.0027), suggesting a bounded event interpretation. By contrast, in the Telic+FOR condition, there was no significant difference in detecting midpoint vs. late point interruptions, indicating an unbounded (coerced) event construal (odds r. = 1.29, p = 0.246).

4. Experiment 2 - atelic to telic coercion: The procedure was as in Exp. 1 with a new set of sentences. A separate group of 192 participants was assigned to one of 3 conditions: 'Atelic' (e.g., 'Ebony should do some drawing'), 'Atelic+FOR' ('...do some drawing for 10 seconds'), and 'Atelic+IN', i.e., Coerced Telic ('…do some drawing in 10 seconds'). As in Exp.1, participants always considered the woman to have done the exercise in critical trials, i.e., all sentences matched the videos (Atelic: 98.2%, Atelic+FOR: 94.1%, Atelic+IN: 97.1%). Turning to glitch detection, the hypothesis was that the IN Adverbial would result in participants perceiving events as bounded, with a lower late-point glitch detection. Conversely, the 'Atelic' and 'Atelic+FOR' conditions were expected to lead to an unbounded event interpretation, with no midpoint-late point glitch difference (see Fig. 2). Again, there was a significant interaction between Condition and Interruption type (χ^2 = 11.55, p = 0.003). In the presence of a coercive adverbial (Atelic+IN), accuracy changed between midpoint (64%) and late point (55%) interruptions, per a coerced, bounded construal (odds r. = 0.56, p = 0.002), whereas the other conditions showed more balanced detection rates (Atelic: odds ratio = 1.13, p = 0.52; Atelic+FOR: odds ratio = 1.35, p = 0.12), aligning with an unbounded event interpretation.

Figure 1 - Experiment 1

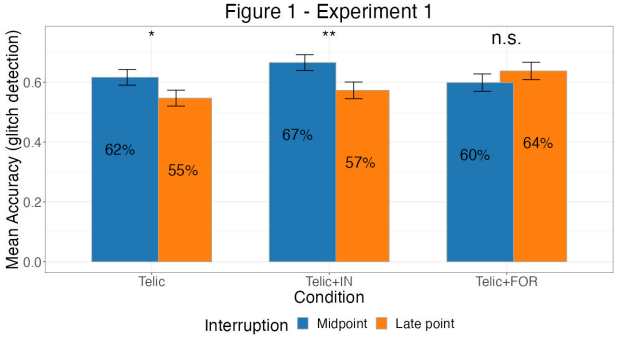
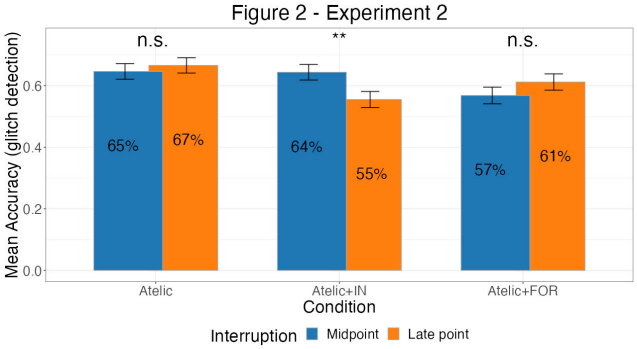


Figure 2 - Experiment 2



6. Conclusion. This study goes beyond traditional measures of coercion processes such as processing costs and reading times to reveal participants' commitments to aspectual meanings via a novel event perception paradigm. We find that, for both directions of aspectual coercion, people's event construals align with coerced sentence readings**.** By linking aspectual coercion in language to distinct patterns in visual event perception, we capture the nuanced ways people cognitively engage with and interpret linguistic cues to aspect, offering clear evidence of commitments to coerced meanings (as opposed to more open-ended, or underspecified aspectual-semantic content).

References: Bott, O. (2010). The processing of events; Dölling, J. (2014). Topics in the semantics of verbs; Jackendoff, R. (1991). Cognition; Ji, Y. & Papafragou, A. (2022). JML; Krifka, M. (1998). Events and grammar; Moens, M. & Steedman, M. (1988). Computational Linguistics; Pickering, M. et al. (2006). Discourse Processes; Piñango, M. et al. (1999). Journal of Psycholinguistic Research

Devoir, ou *pouvoir,* that is the question

Modals can express different forces: possibility (e.g., "you *can"*) or necessity (e.g., "you *must"*). Modal force raises a *Subset problem* for learners [1,2,3]: given that necessity entails possibility, how can children realize that *must* is stronger than *can*? Existing acquisition studies suggest that children struggle with necessity modals particularly, contrasting with their early mastery of possibility modals [3,4,5]. Yet, most of these studies focus on English, where necessity modals are much rarer than possibility modals in talk to children (children's "input") [3], suggesting that the delay could just be due to a lack of exposure. In this study, we show by looking at French, that this isn't sufficient: despite more exposure, French children also struggle with necessity modals.

Background. [3] ran a corpus study of children's modal productions and input from the Manchester Corpus [6] (CHILDES database [7]). They show that English children use possibility modals like *can* early, frequently, but struggle with necessity modals like *must*/*have to*, using them later on, less frequently, and crucially not in an adult-like way. To make results directly comparable, we stayed as close to [3]'s methods as possible, applying them to a French corpus.

Corpus. We used the Lyon Corpus [8] (5 children; age range: 1;00-3;00), and the Paris corpus [9] (6 children; age range: 0;7-6;03), on CHILDES [7]. We extracted and coded modal utterances for force (Possibility: *pouvoir*; Necessity: *devoir*/*falloir*/*avoir-à*), type of modality ('epistemic' vs 'root'), and negation. **Results**. Modal utterances represent 3.8% of all adult utterances (vs 5.8% in English), and 1.9% of child utterances between age 2 and 3. **Table 1** summarizes counts of adult and child productions comparing French and English. We find that in French adult talk, necessity modals are more frequent (62% of all their modal utterances, vs 28% in English). Yet, French children produce more possibility (62%). As in English, they also produce possibility modals earlier (mean age of 1st production: *pouvoir*=1:11; *falloir*=2:03; *devoir*=2:11; *avoir-à*=5:06).

Experiment 1. To test child usage, we use a paradigm introduced by [3]. Its goal is to determine whether children use necessity and possibility modals in an adult-like way. (Adult) participants are presented with mother-child dialogues extracted from the corpus and asked to guess the force of a blanked out modal, by picking between two options: either a possibility (*pouvoir*) or a necessity modal (*devoir*/*falloir*). The modal is uttered either by a child (**Fig1-i**) or by her mother (**Fig1-ii**). **Procedure**. All experiments were coded with PCIbex and run online. Overall, participants had to judge 40 dialogues, presented in a randomized order (20 controls, 20 trials: 10 possibility, 10 necessity, randomly selected out of a list of 20 dialogues randomly extracted from the corpus). **Conditions**. We had three groups based on the speaker's age: 2-3-year-olds, 4-5-year-olds, Adults (used as baseline). We ran two versions varying the necessity modal (**Exp1_*d***: *devoir*; **Exp1_*f***: *falloir;* we don't test *avoir à* because it is too rare). We test only 'root' modals because epistemic uses are too rare in children's production ([10]), and we excluded negated utterances to avoid issues from the scopal interaction of modals and negation ([11]). Force was tested within subject, Age and Lemma between subjects. **Participants.** 358 French participants were recruited on Prolific (60 per condition, 2 failed to record data) (166 F, 186 M, 6 NB; mean age: 32.8yrs). We removed 11 participants whose accuracy scores on controls was <75% (3.1%) (**Exp1_*d***: ADU: 59; CHI2-3: 56; CHI4-5: 59; **Exp1_*f***: ADU: 59; CHI2-3: 54; CHI4-5: 58). **Results. Fig2** summarizes the mean accuracy for each condition. We use generalized linear mixed effects models, built with a maximal random effect structure, testing Accuracy (dependent variable, binomial), with Force as fixed effect and Subject and Item as random factors, and compare them with reduced models without Force as a fixed effect [12,13]. ***Effect of Force.*** For adult production, participants are accurate at guessing force with no difference between possibility and necessity contexts (general mean accuracy: P: 78%; N: 77%). For child production, we find higher performance on possibility than necessity in both age groups (2-3yo: P: 75%; vs N: 60%; **Exp1_*d***: $\chi^2(1)=4$, $p=.04$*, **1_*f***: $\chi^2(1)=4.1$, $p=.04$; 4-5yo: P: 82% vs N: 64%; **1_*d***: $\chi^2(1)=5.3$, $p=.02$*; **1_*f***: $\chi^2(1)=6.5$, $p=.01$*). ***Effect of Age***. Comparing Child groups to Adult, we find significantly lower accuracy for necessity contexts in all age groups. For possibility contexts, we find a difference in **Exp1_*d***, but not in **1_*f***.

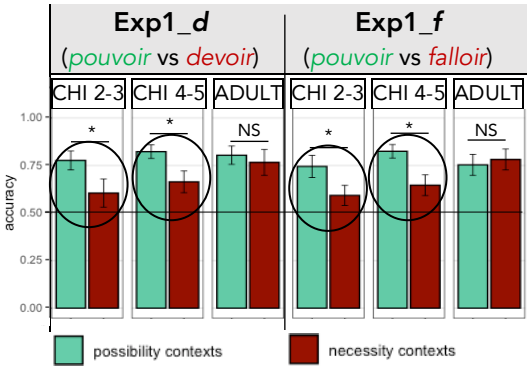
Figure 1. Experimental stimuli: example trials (*pouvoir* vs *devoir*)

(i) **Exp1**: Children's production	(ii) **Exp1**: Mothers' production	(ii) **Exp2**: With role reversal
ENFANT : ... t'en laisses un petit coup. MAMAN : merci. ENFANT : voilà. MAMAN : merci. ENFANT : arrête d'aller là avec le ptit chevaux ENFANT : vous arrêtez d'aller là. ENFANT : parce que c'est après. ENFANT : qu'on _____ aller après. peut doit	AUTRE ADULTE : oui. MAMAN : oui AUTRE ADULTE : une soucoupe. ENFANT : sont un peu vieilles. MAMAN : oui sont un peu abîmées tordues. MAMAN : ah celle-là elle marche bien. AUTRE ADULTE : merci beaucoup. MAMAN : tu _____ souffler dessus. peux dois	MAMAN : ... t'en laisses un petit coup. ENFANT : merci. MAMAN : voilà. ENFANT : merci. MAMAN : arrête d'aller là avec le ptit chevaux MAMAN : vous arrêtez d'aller là. MAMAN : parce que c'est après. MAMAN : qu'on _____ aller après. peut doit
CHILD: you leave a small bit/ MOTHER: thank you/ CHILD: Here you go/ MOTHER: Thank you/ CHILD: Stop going there with the little horse/ CHILD: You stop going there/CHILD: cause it's after/CHILD: that we ___ go after.	OTHER ADULT: Yes/ MOTHER: Yes/ ADULT: a plate/ CHILD: They're kind of old/ MOTHER: Yes they're a little TRANSLATE/ MOTHER: oh, this one works well/ ADULT: Thanks a lot/ MOTHER: you ____ blow on them.	MOTHER: you leave a small bit/ CHILD: thank you/ MOT: Here you go/ CHILD: Thank you/ MOT: Stop going there with the little horse/ MOT: You stop going there/ MOT: cause it's after/ MOTHER: that we ____ go after.

Table 1. Counts and percentage of modal uses in French and English, by force and speaker

		2-3-year-olds count (%mod utt)	3-5-year-olds count (%mod utt)	Adults count (%mod utt)
French	POSSIBILITY	850 (62%)	516 (58%)	2008 (38%)
	NECESSITY	529 (38%)	370 (42%)	3108 (62%)
	falloir	492 (36%)	298 (34%)	2659 (53%)
	devoir	21 (2%)	66 (7.4%)	403 (8%)
	avoir-à	16 (1%)	6 (1%)	46 (1%)
	ALL	**1379 (100%)**	**886 (100%)**	**5114 (100%)**
Eng	POSSIBILITY	3798 (79%)	Not assessed.	13500 (72%)
	NECESSITY	1002 (21%)		5353 (28%)
	ALL	**4800 (100%)**		**18853 (100%)**

Figure 2. Mean accuracy, Exp1 (n=347)



Experiment 2. We ran a follow-up study using the same dialogues, but switching the roles of child and mother, to see whether performance could come from some participants' expectations for children to use more possibility modals, rather than children effective misuses. **Fig1-iii** illustrates the manipulation. Half of the trials had the reversed speakers; half kept the original speaker, allowing us to replicate results from **Exp1**. We excluded contexts when role reversal was too odd, based on a naturalness rating with French naive participants (prop. excluded: 52%). We had four groups (**2_*d***: *pouvoir* vs *devoir*; **2_*f***: *pouvoir* vs *falloir*; judging either adult or child). From a participant's perspective, Exp2 was identical to Exp1. **Participants**. 120 French participants who hadn't taken part in Exp1 were recruited on Prolific (30 per condition) (66 M, 49 F, 2 NB, 3 unknown; mean age: 32.5yrs). We excluded 2 participants due to low accuracy on controls. **Results.** We find that adults' judgements remain stable: we replicate results from Exp1, both on unchanged dialogues (**Table 2**, row (**ii**) vs (**iii**)) and on role reversed contexts (row (**ii**) vs (**iv**)).

Table 2. Results (mean accuracy) of Exp2 (n=118)

		Exp_d (*pouvoir* vs *devoir*)				**Exp_f** (*pouvoir* vs *falloir*)			
		CHI (2-3yo)		ADULT		CHI (2-3yo)		ADULT	
		POSS	NECE	POSS	NECE	POSS	NECE	POSS	NECE
i	**Exp1** (all contexts)	78%	61%	80%	77%	75%	59%	75%	78%
ii	**Exp1** (kept in Exp2)	80%	66%	80%	72%	78%	57%	73%	79%
iii	**Exp2** (unchanged)	82%	65%	79%	72%	76%	59%	71%	85%
iv	**Exp2** (role reversed)	82%	66%	74%	64%	78%	54%	62%	83%

Discussion. We replicate [3]'s findings in French: children master possibility modals early, but struggle with necessity modals. They use them later on, less frequently, and crucially, don't use them in an adult-like way: they use them when adults expect possibility modals. While we still don't know the source of their difficulties, our study shows that they are not limited to English, and that "lack of exposure" can't explain them: French children actually hear more necessity than possibility modals in their input. Are children are confused about the meaning of necessity modals? Or, is it simply that they don't know yet in which contexts they are appropriate? These are questions to discuss, and call for extension to other logical scales where similar *Subset problems* arise, like *some/all* or *sometimes/always*.

References. [1] Berwick, 1985. [2] Wexler and Manzini, 1987. [3] Dieuleveut et al., 2022. [4] Noveck, 2001. [5] Ozturk and Papafragou, 2013. [6] Theakston, 2001. [7] MacWhinney, 2000. [8] Demuth and Tremblay, 2008. [9] Morgenstern and Parisse, 2007. [10] Cournane and Tailleur, 2020. [11]. Iatridou and Zeljstra, 2013. [12] Barr, 2013. [13] Team, R. 2013.

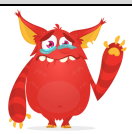
Syntactic structure supports the acquisition of emotion and mental state adjectives

Introduction: Learning the meaning of adjectives presents a challenge to young children, even for adjectives that label salient perceptible properties (Booth & Waxman, 2003; Mintz & Gleitman, 2002; Waxman & Markow, 1998, a.o.). How, then, can children acquire adjective meaning for abstract states, like 'happy' or 'confident'? A well-established finding is that syntactic bootstrapping supports the acquisition of abstract verb meaning (Landau & Gleitman, 1985; Gleitman, 1990), because a verb's argument structure (i.e., the number and position of NPs and the complements it takes) correlates with its meaning. As a result, learners use the presence and type of a sentential complement to deduce that some (but not all) verbs like *think*, *know*, *want*, or *believe* denote mental states (Gleitman et al., 2005; Hacquard & Lidz, 2019). To date, little work has systematically extended this hypothesis to adjectives. Doing so is promising for understanding more about the word learning process and the range and power of syntactic bootstrapping, since some (but not all) adjectives also take complements, even sentential complements. Here, we investigate how syntactic cues from adjectival syntactic complements support the acquisition of one particular type of abstract adjective meaning: adjectives denoting *emotions and mental states*. We demonstrate that while such adjectives may be infrequent in the input, a significant percentage of the time they appear with syntactic complements. We then show across three word learning experiments that both young children and adults actively recruit these syntactic cues to narrow the hypothesis space to an emotion/mental state adjective meaning.

Corpus Search: We analyzed speech of caregivers to English-learning children ages 2-5 years in 44 corpora (CHILDES; MacWhinney, 2000). Here we focus on one aspect of speech: presence and type of syntactic complements within utterances containing an adjective. Out of over 36,000 adjectives, nearly 12,000 denoted size, color, or physical sensation/perception, while only 1,800 labeled emotions or mental states. However, while only 4.2% of *all* adjectives, and 1.5% of color/size adjectives took a syntactic complement, approximately 27% of emotion and mental state adjectives did. Moreover, emotion/mental state adjectives were significantly more likely to appear in predicative (v. prenominal) position, and with animate subjects. Complements were comprised of five main types: ADJ+PP (*about NP, at NP, of NP*) and ADJ+ finite/non-finite sentential clause. We leverage these complements across our word learning studies.

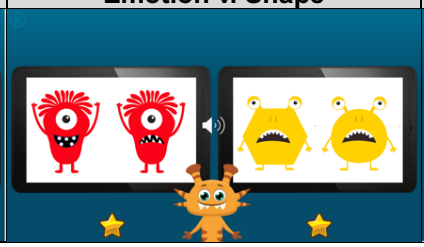
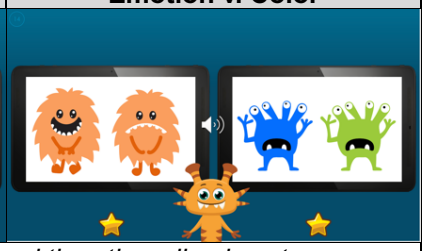
Experiment 1: 51 children (3;0-6;6) and 75 adults participated in a binary forced-choice task manipulating presence/absence of adjectival complement as our between-subject factor. See Figure 1 for sample trial structure. There were five target adjective(+complement) trials. Our dependent measure was choice of emotion match at test. Both children and adults were significantly more likely to choose the emotion match in the Complement condition than in the Baseline condition ($p < .0001$), the latter of which was no different from chance ($p = .51$) (Baseline: children: 45.8%, adults: 40.0%; Complement: children: 75.6%, adults: 93.2%).

Table 1: Sample trial structure for Experiment 1

Familiarization Phase	**Contrast**	**Re-Exposure**	**Test Phase**	
				
Color(yellow)+Emotion (happy)	×Color+×Emotion	✓Color +✓Emotion	✓Emotion ×Color	×Emotion ✓Color
These aliens are both… **Baseline condition** *…daxy.* **Complement condition** *…daxy about something.*	*Uh oh! This alien is NOT daxy…*	*Yay! This alien IS daxy.*	*Here are some new aliens!* *Which one is daxy?*	

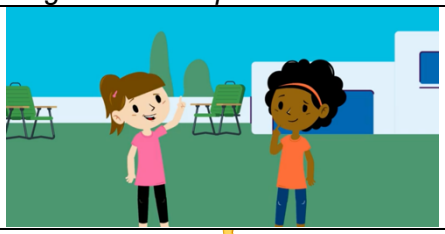
Experiment 2: 54 children (3;8-6;6) and 45 adults participated in a binary forced-choice task again manipulating presence/absence of adjectival complement as our between-subject factor. There were 9 target emotion+other property trials (counterbalancing side, and property of shape, size, and color) and four non-target trials. See Figure 2 for sample target trials. Our dependent measure was the assignment of a star to the emotion contrast. While both children and adults were above chance in the Baseline condition ($p < .0001$), perhaps because emotion was always an available contrast choice, they were significantly more likely to choose the emotion contrast in the Complement condition than in the Baseline condition (Baseline: children: 73.5%, adults: 71.0%; Complement: children: 88.3%, adults: 97.6%) ($p < .05$).

Table 2: Sample target trials for Experiment 2

Emotion v. Size	Emotion v. Shape	Emotion v. Color
		
Find the screen where… one alien is ADJ (+ complement) and the other alien is not.		

Experiment 3: 58 children (17: 4;11-5;10; 20: 6;1-6;11; 21: 7;0-8;4) and 38 adults participated in an asynchronous word learning study administered on Qualtrics in which participants watched animated Powtoon videos of two characters engaged in dialogue using a novel noun and adjective, then provided their best guess as to the novel adj's meaning. See Table 3 for a sample trial. There were 9 trials: 5 targets (ADJ+target complement), 2 baseline (no complement; see Table 1), 2 controls (ADJ+complement and an expletive or gerundive subject, consistent with subjective adjectives, e.g. *It is troby to do something*). No animacy cues were provided for the novel nouns. Both children and adults were likely to guess adjectives for the novel word, *and* were most likely to guess an emotion/mental state adjective for the target trials. Moreover, all conditions differed significantly from each other (Baseline: children: 20.7%, adults: 8.1%; Control: children: 9.9%, adults: 5.4%; Complement: children: 38.6%, adults: 75.0%). Children were increasingly more adultlike with age. Thus, given the presence of *syntactic* cues with no *visual* cues, participants were able to converge upon an abstract emotion/mental state adjective meaning.

Table 3: Sample dialogue for one target trial of Experiment 3

A: Hey, look, there's a derkum! Do you see the derkum?		*B: Oh yes, I see the derkum.*
A: You know, that derkum is troby.		*B: You're right. That derkum is troby.*
A: Mm hmm, the derkum is troby about doing something.		*B: I agree. The derkum is troby about doing something.*
What do you think the word "**troby**" means? 👇Type your guess in the space below. 👇		

Conclusions: Acquiring abstract meaning presents an inherent challenge in word learning. Syntactic complements are a reliable distributional cue in child-directed speech known to support the acquisition of mental state verbs. We show that both children and adults are able to recruit these cues to deduce abstract adjective meaning, arriving at an emotion/mental state interpretation. This research thus extends the syntactic bootstrapping mechanism beyond verbs to adjectives, highlighting the potency of syntax for supporting the acquisition of word meaning.

Both Principle B and Competition Are Necessary to Explain Disjoint Reference Effects

Introduction. Many languages exhibit a restriction against pronouns expressing local coreference [1]. It remains debated whether this is due to an explicit grammatical constraint against local pronominal coreference (classical *Principle B*) [1,6], or if it instead reflects *Competition*, a pragmatic reasoning process selecting between competing alternative forms [3-5]. To evaluate these approaches, we conducted two experiments using *Evans Sentences* as in (1) [2]. These apparent violations of Principle B have been critical to the development of *Competition* theories emphasizing distinctness of meaning in context [3-5] and taken to indicate that Principle B governs bound variable anaphora rather than coreference [3].

(1) Sarah said that everyone voted for Michael, but she lied. Only Michael$_i$ voted for him$_i$.

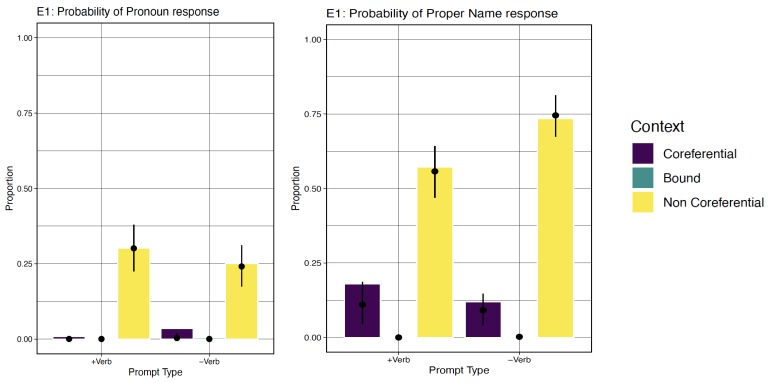
Competition claims coreference is possible when the context makes bound variable anaphora unavailable. For example, coreference in (1) should be available because the context distinguishes a bound variable interpretation (*Only Michael*(*x voted for x*)) from coreference (*Only Michael*(*x voted for Michael*)). Competition then expresses a requirement to use a reflexive form when the meaning is indistinguishable from a bound variable interpretation (e.g. *Rule I* or similar [3, 5]). *Principle B* [1,6] and *Competition* [3,4,5] make differing predictions about the production and comprehension of these sentences. In production, *Principle B* prohibits pronouns for local coreference, and so producers should always find some other way to express local coreference in any context. *Competition* allows producers to use pronouns in contexts that prohibit bound variable anaphora. Since pronouns are better than alternative possibilities like repeated names, we expect to see pronouns selectively in these contexts. In comprehension, *Principle B* predicts rejection of pronouns with a local antecedent, but if participants do accept it, there should be no correlation with context. *Competition* predicts comprehenders to allow coreference when they can associate it with contexts that prohibit bound variable anaphora. Our experiments support both predictions, revealing the need for both an explicit constraint against local coreference *and* Competition in deriving Principle B effects. In short, we found an overall preference for reflexives for both coreferent and bound meanings, as predicted by *Principle B* [1,6], but when participants did accept the pronoun form in the comprehension experiment, they preferred a coreferential interpretation, as predicted by *Competition* [3].

Experiment 1 (N_{subj} = 36). *Which forms do participants produce, given a meaning?* Participants completed natural SMS exchanges [7]. We manipulated the context so that participants had to choose a form to express a locally bound, locally coreferential, or locally non-coreferential dependency (Table 1); we further manipulated whether the prompt contained the repeated verb or not (+/- verb). We created 48 critical items, distributed via Latin square and randomized with 48 filler items.

Table 1. Sample item, manipulating CONTEXT

A: By the way, Zachary said that everyone listened to…	
(*Bound: x listened to x*)	**…themselves.**
(*Coreferential: x listened to Ashley*)	**… Ashley.**
(*NonCoreferential: x listened to Jacob*)	**…Jacob.**
B: He lied! I overheard that only Ashley listened to _____.	

Fig. 1. Probability of Pronoun and Name responses in E1



Participants (i) overwhelmingly preferred the reflexive form in both the Coreferential (88.4%) and Bound (100%) contexts, and (ii) produced almost no pronouns coreferential with the local subject (see Fig. 1). In production, we see a strong preference for abiding by Principle B, without influence from

the greater discourse context. However, Competition may reflect an interpretive strategy [4]. We address this in Exp. 2.

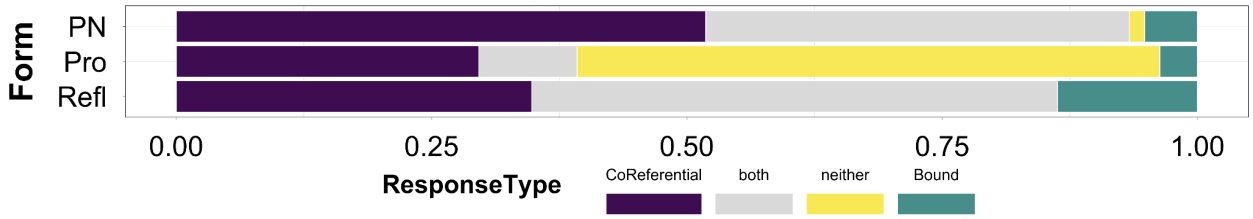
Experiment 2 (N_{subj} = 54). *Do comprehenders accept pronouns with local antecedents, and if so, in what contexts?* Participants were shown an SMS exchange, using Experiment 1's stimuli. The final sentence was complete, but the critical context sentence was blanked out. We manipulated the form that participants saw (Table 2). They were asked to choose the best sentence to fit in the context blank, and could choose the Bound sentence, Coreferential sentence (Table 1), *Both,* or *Neither*. Principle B predicts that the pronoun form should be unacceptable with a local antecedent, and thus we should expect only *Neither* responses in the *Pronoun* condition, since both contexts force an interpretation with a local antecedent. Competition predicts a strong preference for coreference in the Pronoun condition.

Table 2. Sample item, manipulating FORM

A: By the way… _____.	
B: He lied. I overheard that only Ashley listened to…	
(*Reflexive*)	**…herself.**
(*Pronoun*)	**…her.**
(*ProperName*)	**…Ashley.**

Our results support both predictions (Fig. 2). Consistent with Principle B, *Pronoun* was the only condition where participants rejected both contexts at a high rate (57%; PN=1.5%, Refl=0%), supporting the dispreference found in Exp 1. In an analysis that excluded 'neither' responses, participants endorsed the coreferential context at a higher rate for Pronouns (91.4%) than for Reflexives (86.3%); for the purposes of this analysis, we treated 'Both' and 'Coreferential' responses as endorsement of the latter.

Fig. 2. Exp. 2: Distribution of participants' responses



Discussion. In both of our experiments, we find strong avoidance of local antecedents for pronouns no matter the context, suggesting a grammatical constraint against local pronominal coreference (*Principle B*). At the same time, when comprehenders do assign an interpretation to pronouns with local antecedents, they systematically associate them with coreferential, not bound readings, as predicted by *Competition*. Together, our results suggest that syntactic context has more influence on availability of local coreference than discourse context, but both are required for a complete theory. Our results also suggest an asymmetry between production and comprehension of these sentences, as in Experiment 1 participants almost categorically avoided the pronoun form, but in Experiment 2, they were able to systematically interpret it. This may mean that Competition reflects comprehender-side Gricean [8] or Bayesian [9] reasoning processes that complement, but do not fully explain, the constraint against locally coreferent pronominals.

References. [1] Chomsky, N. (1981). *Lectures on government and binding*. [2] Evans, G. (1980). Pronouns. [3] Grodzinsky, Y., & Reinhart, T. (1993). *The innateness of binding and coreference*. [4] Reinhart, T. (2006). Interface strategies: Optimal and costly computations. [5] Roelofsen, F. (2010). *Condition B effects in two simple steps*. [6] Heim, I. (2007). *Forks in the Road to Rule I*. [7] Kroll, M.I. (2020). *Comprehending ellipsis*. [8] Marty, P. P. (2017). Implicatures in the DP domain. [9] Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference.

"Liz can buy a croissant or a donut… Both together, right?"
Distinguishing target Free Choice from non-target Modal AND in Child French

Background: Acquisition studies have shown that (i) while children give non-target interpretations of plain disjunction (1a), where OR is not interpreted exclusively as in the adult grammar (1b) but instead conjunctively as AND (1c) ([5-6]), (ii) they have the target interpretation of modal disjunction (2a), correctly drawing the conjunctive Free Choice (FC) inferences in (2b-c) ([3], [7-8]). A widely accepted account proposed by Singh et al. (2017), relying on Fox' (2007) double exhaustification analysis of adult conjunctive FC inferences (2b-c), is that children have adult-like semantics but cannot retrieve the stronger alternative AND – allowing for the non-adult conjunctive strengthening of OR to the meaning of AND (1c).

(1) **Plain disjunction**

a. Liz bought the donut or the croissant.		$(P \vee Q)$
b. ⇝ Liz bought the donut or the croissant but not both.	[Adult]	$(P \vee Q) \wedge \neg(P \wedge Q)$
c. ⇝ Liz bought both the donut and the croissant.	[Non-adult]	$(P \wedge Q)$

(2) **Modal disjunction**

a. Liz can buy the donut or the croissant.	$\Diamond(P \vee Q)$
b. ⇝ Liz is allowed to buy the donut.	$\Diamond P$
c. ⇝ Liz is allowed to buy the croissant.	$\Diamond Q$
d. ⇝ Liz is not allowed to buy both.	$\neg\Diamond(P \wedge Q)$

Proposal: Putting to test the conclusion that children derive genuinely adult-like FC inferences, we empirically tested an alternative interpretation of children's responses to the FC inferences of modal disjunction: children interpret $\Diamond(P \vee Q)$ as $\Diamond(P \wedge Q)$, just like they interpret $(P \vee Q)$ as $(P \wedge Q)$. Crucially, $\Diamond(P \wedge Q)$ (modal AND) is not equivalent to $(\Diamond P \wedge \Diamond Q)$ (FC inference), since the former entails the latter, but not conversely. Suppose that P and Q can both hold $(\Diamond P \wedge \Diamond Q)$ but *not* simultaneously, then $\Diamond(P \wedge Q)$ comes out as false (3), e.g. John's being in Paris and John's being in London are mutually exclusive situations: they can hold at different times (3b)/(3c), but not simultaneously (3d).

(3)

a. John might be in Paris or in London.	[1] [9]
b. ✓ John is in Paris.	$\Diamond P$
c. ✓ John is in London.	$\Diamond Q$
d. ✗ John is in Paris and he is also in London.	$\neg\Diamond(P \wedge Q)$

Previous designs do not distinguish $(\Diamond P \wedge \Diamond Q)$ from $\Diamond(P \wedge Q)$ which is critical for the interpretation of the results. To this effect, we develop an experimental paradigm with mutually exclusive scenarios to tease apart the two interpretations and thus settle whether children have genuine FC construals of modal disjunction. Our findings extend the empirical observation that children derive a conjunctive interpretation of OR to modal contexts $\Diamond(P \wedge Q)$. This novel observation follows on the proposal that children have the adult semantics for (2), but exhaustify below the modal.

Method: 57 French children (M=5;5 | 3;11-6;9) and 37 adults (M=34 | 22-68) participated in a truth-value judgment task adapted from [4]. In the setup, a shop employee describes what a customer can buy given the number of coins in her purse (4). The task is to judge whether the employee said it right. The conditions in Table 1 vary the price of the objects. Condition 1 replicates prior studies by ensuring the falsity of FC inferences. Condition 2 tests mutually exclusive situations, while condition 3 renders true the $\Diamond(P \wedge Q)$ interpretation.

Hypotheses: If children have the target FC interpretation of modal disjunction (row 1 in Table 1), they should accept test sentence (4) for Condition 2 and reject it for Condition 1. In contrast, if children have a non-target conjunctive interpretation of modal disjunction (row 2 in Table 1) they should reject (4) for both Conditions 1 and 2, while accepting it for Condition 3.

(4) **Test sentence:** "With 1 coin, Liz can buy a croissant or a donut."

Interpretation of ◇(P ∨ Q)	Condition 1	Condition 2	Condition 3
			
◇P ∧ ◇Q (FC)	No	**Yes**	No
◇(P ∧ Q)	No	**No**	Yes

Table 1: Conditions and expected answer patterns

Results: Fig. 1 shows that **Condition 1** was rejected by adults, but not always by the children. Post-hoc simple pairwise comparisons showed that this difference is statistically significant ($p < 0.0001$). Most children were consistent: 38 always rejected Condition 1, 11 always accepted it, and 8 were at chance. **Condition 2** was always accepted by adults and also mostly by children; the difference with adults was not significant ($p = 0.21$). The individual pattern analysis in Table 2) revealed that children who rejected Condition 1 were split in two groups: 22 accepted Condition 2 (giving a FC interpretation) and 11 rejected it (giving a modal AND interpretation). **Condition 3** was mostly accepted by children in contrast to the adults - in particular by those children who rejected both Conditions 1 and 2. Overall, mixed model analyses with *condition* and *age* as fixed factors showed that **Age** was not a significant predictor for children ($p = 0.55$).

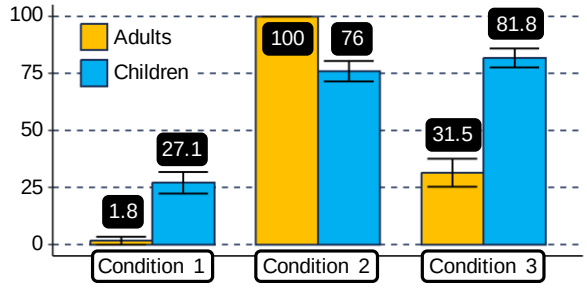


Figure 1: Mean percentage of acceptance for each condition

Table 2: Individual patterns

		Condition 2	
		Accept	Reject
Condition 1	Accept	11 (◇P ∨ ◇Q)	
	Reject	22 (◇P ∧ ◇Q)	11 (◇(P ∧ Q))

Discussion: Using a novel and more detailed design, our study reveals that FC is not so early acquired contrary to previous claims. Indeed, 22 out of the 57 children (38%) were in fact consistent non-adult interpreters of modal disjunction: 11 had Modal AND (◇(P ∧ Q)), and 11 did not derive FC inferences (◇P ∨ ◇Q). Moreover, 1 in 3 children who seemed adult-like on Condition 1 turned out to be modal AND ◇(P ∧ Q) interpreters once Condition 2 was taken into account. Having shown that the non-adult conjunctive interpretation of OR reported in the literature extends to modal OR, we straightforwardly extend Singh et al's account of conjunctive OR (lack of access to the stronger alternative) to modal contexts on a simple assumption: namely, that double exhaustification takes place below the modal:

(5) a. Exh(Exh(◇(P ∨ Q))) ⇔ ◇P ∧ ◇Q Double Exhaustification above ◇ ⇝ Free Choice
b. ◇(Exh(Exh(P ∨ Q))) ⇔ ◇(P ∧ Q) Double Exhaustification below ◇ ⇝ Modal AND
c. Exh(Exh(P ∨ Q)) ⇔ (P ∧ Q) Double exhaustification of OR ⇝ AND

The exhaustification procedure that leads to the modal AND interpretation is thus exactly on a par with the one that strengthens the meaning of plain disjunction to conjunction (5c).

Selected references: **[1] Ciardelli**, Groenendijk and Roelofsen. (2014). *Approaches to Meaning: Composition, Values, and Interpretation*. **[2] Fox**. (2007). *Presupposition and implicature in compositional semantics*. **[3] Huang** and Crain. (2020). *Language acquisition* 27(1). **[4] Liu**. (2017). *Interpreting Disjunction under Deontic Modals: An Experimental Investigation*. **[5] Singh**, Wexler, Astle-Rahim, Kamawar, and Fox. (2016). *Natural Language Semantics* 24(4). **[6] Tieu**, Yatsushiro, Cremers, Romoli, Sauerland and Chemla. (2017). *Journal of Semantics* 34. **[7] Tieu**, Romoli, Zhou and Crain. (2016). *Journal of Semantics* 33(2). **[8] Zhou**, Romoli and Crain. (2013). *Proceedings of SALT 23*. **[9] Zimmermann**. (2000). *Natural Language Semantics* 8.

Experiments in (non-truth-conditional) linguistic meaning: Exploring subjective predicates and perspective-taking

Elsi Kaiser
University of Southern California

Abstract:
The information that we encounter conveys both objective facts about the world and people's subjective opinions. This distinction is also reflected in language: Words that express opinions (e.g. fascinating, frightening) differ from words conveying more objective facts (e.g. wooden, Philadelphian): Subjective adjectives are perspective-sensitive and reflect someone's opinion/attitude, whereas objective adjectives express factual information. Indeed, when two people disagree about matters of taste, neither is in the wrong: It is widely observed that there is nothing contradictory when one person says "That cheesesteak was tasty!" and the other responds "No, it was not tasty" (faultless disagreement) -- in contrast to disagreements about objective facts. The question of how (and whether) to capture such phenomena using truth-conditional semantics is a foundational question that has attracted extensive attention formal semantics and philosophy, but has traditionally not been explored from an experimental perspective. In this talk, I will present a series of psycholinguistic studies from my lab that use a variety of experimental methods to explore three inter-related questions concerning subjectivity: First, how good are we at noticing subjective information, at recognizing something as a subjective opinion? Second, how accurately and how automatically do we keep track of whose opinion is being expressed? Third, when faced with opposing opinions, do we really regard the disagreement as faultless, with neither person being in the wrong? As I will show in my talk, the processing of subjective adjectives is constrained in semantically and syntactically principled ways, but also guided by contextual and social considerations that go far beyond the predicate itself. These results call for an approach to subjective adjectives that integrates not only lexical factors, but also sentence-level, interlocutor-level and social factors.

Presuppositions project asymmetrically, *unless* they don't

Overview. The theory of presuppositions aims to predict and explain how presuppositions project or are filtered in different environments. Early theories derived this behavior by stipulating projection properties on a connective-by-connective basis (Karttunen 1974, Heim 1983, a.o.). But this is explanatorily unsatisfying (Soames 1989, Schlenker 2008 a.o.). More recent work tries to derive projection properties from the truth-conditions of connectives together with global facts about language processing (Schlenker 2008 a.o.). In particular, *asymmetries* in projection are explained on this approach by the sequential nature of linguistic processing. A striking prediction of this kind of approach is that since asymmetries are due to a global feature of the linguistic system, asymmetry will be a uniform feature of projection across different connectives. Existing experimental literature, however, has found differential (a)symmetries across connectives. Only left-to-right filtering appears possible across conjunction (e.g., (1a) vs (1b); see Mandelkern et al. 2020). By contrast, disjunction exhibits right-to-left filtering as well (e.g., (1c) vs (1d); see Kalomoiros 2023).

(1) a. Mary studied in Tokyo, and John studied in Japan too.
b. #John studied in Japan too, and Mary studied in Tokyo.
c. Mary didn't study in Japan, or John studied in Japan too.
d. John studied in Japan too, or Mary didn't study in Japan.

We contribute to this debate by testing order effects for presuppositions triggered under 'unless'. We present an experiment showing that *unless*-sentences exhibit costless symmetry: that is, both left-to-right and right-to-left filtering are equally possible for 'unless'. These results extend the empirical picture for theories of presupposition, and, given existing findings about conjunction, extend the challenge for processing-based accounts of presupposition.

Experiment. We adapted the acceptability paradigm from Mandelkern et al., 2020. Critical items consisted of two conditions differing in Order: PsFirst, with initial Unless-clauses containing a presupposition (based on either *too*, *again*, or the prefix *re-*), and a consequent whose negation strictly entailed that presupposition; and PsSecond, identical but with reversed clause-order, (2a-b). (This contrasts with the only prior relevant experimental study on *unless* by Chemla & Schlenker 2012, who compared presuppositions in the antecedent of *unless*-clauses with those in the consequent.) Both were presented in *Explicit Ignorance* (EI) contexts (Simons 2001), that explicitly leave open whether the presupposition holds. If the presupposition projects, it should conflict with this ignorance, leading to decreased acceptability (which should not arise if filtering is available).

Importantly, our design employed consequents whose negation strictly entails the presupposition of the antecedent, to rule out potential symmetry effects due to *cancellation/local accommodation* (Gazdar 1979, Heim 1983); e.g., (1d) could be seen as triggering local accommodation to avoid a presupposition settling the truth of the other disjunct (Hirsch & Hackl 2014). In our stimuli, the presupposition of the antecedent is compatible with the consequent to rule out a parallel possibility in *unless*-sentences. Minimally varied non-presuppositional controls provided a baseline for potential Order effects independent of presuppositions: NoPsFirst and NoPsSecond, identical to corresponding critical items but with presupposition triggers removed, also shown in EI contexts for maximal comparability (2c-d). There were also additional controls, namely *unless* sentences with a presupposition in the antecedent, and an unrelated consequent that didn't allow for filtering, (3) (SimplePs). These appeared in EI and Support (S) contexts. The former requires local accommodation to prevent the presupposition from clashing with the context. The latter directly supported the presupposition in the context, with no recourse to local accommodation needed. The difference in acceptability between EI-SimplePs vs S-SimplePs thus acts as a baseline for the cost of local accommodation.

Methods. 200 participants were recruited. They saw relevant contexts paired with a sentence, and were asked to evaluate the sentence's naturalness on a 7-point scale.

Predictions. Processing accounts predict **PsFirst to be less acceptable than PsSecond**, going beyond potential Order-effects in NoPs controls and **resulting in an interaction between Order and Ps**. They may allow for symmetric filtering at a cost (reflected in decreased acceptability)

relative to left-to-right filtering; but this cost should be less than the cost for local accommodation, predicting the following: we can categorize the EIPsFirst and EISimplePsFirst conditions as exhibiting NoPriorS(upport) (they do not involve preceding material supporting their presupposition); conversely, EIPsSecond and SSimplePsFirst exhibit PriorS(upport). Then, the effect on acceptability of switching from PriorS to NoPriorS should be greater in the SimplePs cases, than in the PsFirst/Second cases; thus, **a (No)PriorS $\times$ Simplicity interaction is predicted**.

Results. We find no difference between PsFirst vs. PsSecond ($p = 0.56$), and no Order$\times$Ps interaction ($p = 0.8$), (Fig 2); we do find a (No)PriorS$\times$SIMPL interaction ($p < 0.01$) (Fig 3). We used a Bayesian analysis to assess credence in the null Order$\times$Ps interaction (using the Mandelkern et al. conjunction interaction (see Fig 1) as our priors). We find extreme evidence in favour of the null interaction model ($\text{BF}_{10} < 0.01$) (Jeffreys 1939).

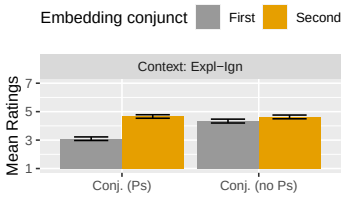


Figure 1: Mandelkern et al. Order$\times$Ps interaction

Discussion. To account for our data, one could stipulate that the filtering profile of *unless* is symmetric (essentially making filtering part of the lexical entry). But this is clearly not explanatory (Soames 1989, Schlenker 2009 a.o.). More satisfactorily, if we can treat *unless* as more parallel to disjunction (i.e. *Unless* A, B $\approx$ A *or* B), as a first approximation, we could account for the results via theories that predict symmetric disjunction (George 2008, Kalomoiros 2023), but other theoretical moves should be explored as well.

(2) ***EI Context:*** John and Mary are siblings and want to study abroad. Options include Tokyo and Kyoto in Japan, or Beijing and Shanghai in China. Mary is interested in studying in Japan: she would go to Kyoto on her own, but she doesn't want to go to an enormous city like Tokyo, unless John also comes with her to Japan (if not to Tokyo, then at least to Kyoto). I don't know what they ended up deciding so ***I have no idea whether Mary is currently studying in Tokyo or whether she even decided to go to Japan.*** However, given the above, I know that:

a. Unless John is studying in Japan too, Mary is not studying in Tokyo. PsFirst
b. Mary is not studying in Tokyo, unless John is studying in Japan too. PsSecond
c. Unless John is studying in Japan, Mary is not studying in Tokyo. NoPsFirst
d. Mary is not studying in Tokyo, unless John is studying in Japan. NoPsSecond

(3) ***EI/S Context:*** John and Mary are siblings and are trying to figure out whether to study abroad or not. Options for studying abroad are China or Japan. Mary has a preference for Japan over China, but at the same time she will be unhappy if she's studying abroad and John isn't with her. I don't know what either of them have decided so ***I have no idea whether Mary is studying in Japan.***/ ***I know that in the end Mary went to Japan, but I have no idea what John did.*** However, given the above, I know that:

a. Unless John is studying in Japan too, Mary is unhappy. SimplePs

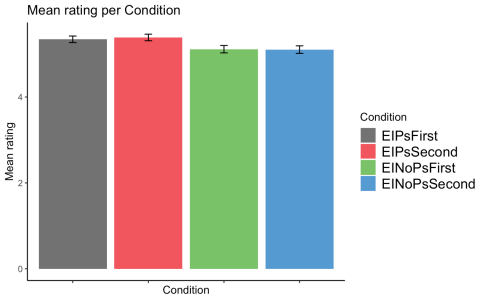


Figure 2: Order $\times$ Ps

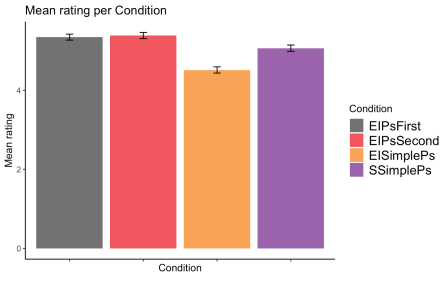


Figure 3: (No)PriorS $\times$ Simplicity

Relating Scalar Inference and Alternative Activation: A view from the Rise-Fall-Rise Tune in American English

The rise-fall-rise (RFR) tune in American English is notorious for its varied and often conflicting description in both its pragmatic function and its phonological description. Phonological theory (Pierrehumbert, 1980) predicts 3 RFR-shaped tunes that differ in pitch accent (monotonal H* or bitonal L+H* or L*+H), yet it is unclear whether reported variation in the semantic/pragmatic function of RFR is due to differences among studies in the intonational materials used, or whether some (or perhaps all) accounts might be unified under a broad class of RFR intonational patterns. A common thread among accounts relates RFR to higher alternatives; i.a. conveying uncertainty (Ward and Hirschberg, 1985), unclaimability (Constant, 2012), or salience (Göbel, 2019) of some higher alternative. These accounts make different predictions for RFR's effect on scalar inference calculation (SI, e.g., *some→ some but not all*). Experimental work has used SI as a probe to adjudicate among accounts of RFR, finding higher rates of SI calculation when RFR is used (de Marneffe and Tonhauser, 2019; Göbel and Ronai, 2023, though cf. Buccola and Goodhue, 2023); however, these studies tested only a single RFR tune with non-specific phonetic description. Thus, there lacks a study comparing the interpretation of the various RFR-shaped tunes. Finally, there is a processing question: results from cross-modal priming show that some but not all rising pitch accents modulate the activation of focus alternatives in processing (i.a., Husband and Ferreira 2016). Since alternatives in SI and focus have been claimed to be related both in semantic/pragmatic theory (i.a., Fox & Katzir, 2011) and psycholinguistics (i.a., Gotzner & Romoli, 2022), differences in the interpretations of RFR-shaped tunes may be reflected in processing, as indexed by alternative activation. We present a systematic investigation of 3 RFR-shaped tunes. Using both SI judgment and cross-modal lexical decision tasks, we test 1) differences among RFR-shaped tunes in their interpretation, and 2) whether any such effects are mirrored in processing. **Materials:** We wrote polar question+indirect answer dialogues for 72 different adjective pairs that form a scale, e.g., <*tough*, *impossible*> Q: *I haven't gone running since before the pandemic, do you think I could do a half marathon?* A: *That distance would be* ***tough***. In a text-only norming task, undergraduate participants (n=48) read the dialogues, provided acceptability ratings, and answered questions such as "Would you conclude that that distance would **not** be **impossible**?", where a "Yes" (as compared to "No") response means that SI was calculated. We found that the dialogues were overall acceptable compared to incongruent fillers and we replicated previous findings that SI calculation rates vary across scales (i.a., van Tiel et al., 2016). We chose 64 items to record in 6 intonation conditions with one of 3 pitch accents ("neutral" H* and focus-marking L+H* and L*+H) and one of 2 edge-tones (fall, L-L%, or fall-rise L-H%) and standardized the pitch contours using pitch resynthesis in Praat.

Exp 1: These auditory materials were used in a follow-up SI task. Online participants (Prolific, n=83) listened to a dialogue in one of 6 intonation conditions then answered questions like "Would you conclude […] not be impossible?" If RFR conveys uncertainty about a higher alternative, and not belief in its negation, then we predict lower SI rates for RFR compared to falls. But if RFR instead functions more broadly to mark the salience of higher alternatives, we predict higher SI rates for RFR, as salient alternatives are available for SI calculation. Such effects could potentially be seen with just **one** RFR or with **any** RFR-shaped tune. **Results:** A Bayesian logistic mixed effects model shows a main effect of edge-tone: all RFR-shaped tunes yielded higher SI rates compared to falls (posterior probability of direction=100%). We find slight gradience between the pitch accents within the RFR tunes, with an average SI rate ranking of L*+H > L+H* > H*, which is reversed for falls (Fig. 1).

Exp 2: In a cross-modal lexical decision task, participants listened to the recorded dialogues (*…tough)* and, after a 750ms delay, wereshown the higher alternative (*impossible*). Participants then judged whether the displayed string was a word or non-word. We measure the reaction time (RT) of participants' judgments and predict that if RFR invokes higher alternatives (as shown in Exp 1), then *impossible* should be facilitated (=faster RTs). To control for activation arising from semantic similarity, adjectival scales are also tested in the opposite prime-target order, where participants listen to *…impossible* and judge *tough.* The predicted RFR effect on the activation of higher alternatives (e.g. *impossible*) should be greater than (1) the activation of *tough* when *impossible* is uttered with RFR and (2) the activation of *impossible* when *tough* is uttered with other tunes. This task is administered in the laboratory using low-latency hardware in a sound-attenuated booth, as well as online, where the hardware and environment of the participant cannot be as easily controlled. Preliminary results from in-person data collection (Undergraduates, n=46/target 60) (Fig.2) suggest a main effect of displayed alternative such that RT is faster for higher alternatives after accounting for word frequency and length (p.d.=98%). Moreover, H*L-H% shows evidence of an interaction, yielding additional facilitation for higher alternatives (p.d.=91.9%). We do not find evidence of such an effect with the other RFR-shaped tunes (p.d.<85%) nor the falls (p.d.<65%). Online participants (Prolific, n=60) show longer and more varied RTs with no notable pattern of facilitation across intonation conditions, suggesting that this in-person effect is subtle and not robust to noise arising from unconstrained environmental factors.

Discussion: The findings of Exp. 1 replicate prior work and are most compatible with accounts of RFR that do not invoke uncertainty. But the combined findings from the two tasks present a puzzle: in Exp. 1, while all RFR-shaped tunes increase SI rates compared to falls, L*+HL-H% increases the likelihood of SI most strongly. Since SI arises via retrieval and negation of a higher alternative, we expect the processing signature of this tune to show stronger facilitation in lexical decision. Yet in the lexical decision task, only H*L-H% provides evidence of facilitation. We discuss a possible account of this pattern in terms of pitch range, following Ward & Hirschberg (1992): expanded pitch range in the bitonal RFRs may invite competing inferences, e.g., related to speaker arousal. In the SI task, participants might be more willing to accommodate the SI-enriched interpretation because it is explicitly probed via the question in the trial. But this task effect is not present in priming, and the competing inferences may mask potential facilitation.

Implications: By controlling the intonational variation in the auditory materials and the experimental setting, we provide novel psycholinguistic evidence of the relationship between RFR-shaped tunes and scalar alternatives. We find a distinction between lower-scaled/monotonal RFR (H*) and higher-scaled/bitonal RFR (L+H*/L*+H), but overall, all RFR tunes behave differently from falls. The variation among RFR-shaped tunes in our experiments is emblematic of the variation seen in prior work, suggesting within-category variation may reflect more particularized inferences beyond RFR's conventional connection with higher alternatives.

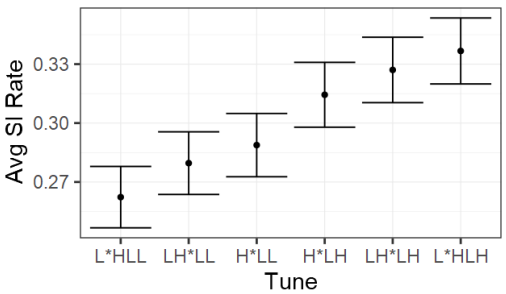


Fig. 2: Average SI rates for each tune in Exp. 1, _LL=Fall & _LH=RFR

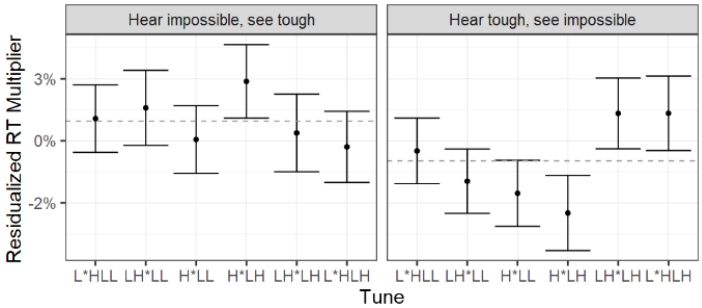


Fig. 1: Residual speedup (-x%) or slowdown (+x%) in RT controlled for log word frequency, log length, and block.

On the salience of linguistic alternatives in the inference task for scalar implicatures

Background. Variability in rates of Scalar Implicatures (SIs) has been observed across many studies: between contexts, individuals, participant groups, and scalar expressions. Here we focus on another kind of variability – the fact that inference tasks tend to result in higher rates of with-SI response than comparable verification tasks. [1] explicitly demonstrates this fact using the two tasks with the same sentences. It can also be observed in comparing outcomes for scalar expressions appearing both in inference tasks like [2] and verification tasks like [3]. A by now standard inference task stimulus is shown in Fig.1a (based on [2]). To account for the raised rates in inference tasks, [1] conjectures that, by asking the participant if the speaker excludes the alternative, the probe question strongly suggests that it is relevant. Another factor that may be at play is that the probe question references the linguistic alternative. According to some views, salience of the alternative expression itself can impact positively on SI availability [4, 5]. This view has recently been challenged in [6, 7]. [7] argues that mere salience of the scalar expression is relatively inert in promoting SI. We report on a study that tested these competing ideas on the efficacy of alternative salience by manipulating whether the alternative was explicitly mentioned, implicitly present, or entirely absent in the probe, while holding constant the meaning of the question asked and thereby the relevance of alternatives.

Experiment. Our test trials are illustrated in Fig.1. In each condition, the lexical content of the target statement was manipulated to test whether the presence of the alternative has any effect on SI rates above that of making the proposition expressed by the alternative contextually relevant. For these purposes, we introduced two novel ANTONYM probes in addition to the standard NOT-ALT probe. Unlike in the NOT-ALT probe, the query in the ANTONYM probes expressed the falsity of the alternative of interest by other linguistic means than referencing the stronger alternative and embedding it under negation. We further distinguished between ANTONYM and ANTONYM* probes in order to detect if implicit activation of the alternative promotes SI. This is possible in the former, as opposed to the latter, since the scalar expression is employed in the probe and this itself may trigger a SI, involving a representation of the alternative in its derivation. ANTONYM* probes were variants of the ANTONYM probes in which neither the weak scalar expression, nor its stronger scalemate appeared. These probes were created mainly by using a blank paraphrase of the weak scalar expression (as in Fig.1c), or else by replacing that expression with a lexical antonym (e.g., replacing *tried to* with *failed to*). We tested 12 lexical scales (see Table 1). Probe was a between-group factor. Participants ($n = 164$) were assigned to one of three lists containing 36 target items (3 instances of 12 scales) plus 10 control items. We hypothesized that if raising the salience of an alternative has a boosting effect on SI rates above that of suggesting its relevance, the proportions of Yes-responses in the test trials should be lower in either of the ANTONYM conditions than in the NOT-ALT conditions.

Main results. The distribution of by-participant mean rates was very similar in all three probe conditions, as shown in Fig.2. We fitted a Bayesian mixed effects logistic regression model to the data. The hypothesis that ANTONYM(*) should yield lower rates of acceptance than NOT-ALT was tested using the hypothesis function of `brms`. The posterior probability of ANTONYM yielding lower rates of acceptance than NOT-ALT was 49% with an evidence ratio of 0.96, and the difference was estimated to be 0.01 with 90% quantiles being [-0.54,0.55]. For ANTONYM*, the posterior probability was 40% with an evidence ratio of 0.67, and the difference was estimated to be 0.08 with 90% quantiles being [-0.47,0.64]. Fig.3 shows the mean rates by Scale and Probe type. For each scale, we fitted a GLMER model with a logit link function, predicting participants' responses from the fixed effect of Probe (treatment coded). The results of the model comparison tests showed that including Probe as a predictor led to a significantly improved fit over the null model for only two scales, ⟨permit, require⟩ and ⟨few, lot⟩. For both these scales, the estimated marginal means were significantly higher in the ANTONYM conditions than in the NOT-ALT conditions. We conclude that the by-scale rates of SIs were largely unaffected by the Probe manipulation, consistent with the results of the global analysis.

Discussion. Our results show that SI rates are much the same across all three probe conditions and they provide evidence against the hypothesis that making the alternative contextually salient has a boosting effect on SI rates above that of merely raising the relevance of that alternative. These findings, on the other hand, are in line with the idea that the probe question generally biases participants to think that the alternative is relevant, enhancing the likelihood that the SI reading be endorsed and accounting in turn for the inflated rates of SIs yielded by the inferential paradigm.

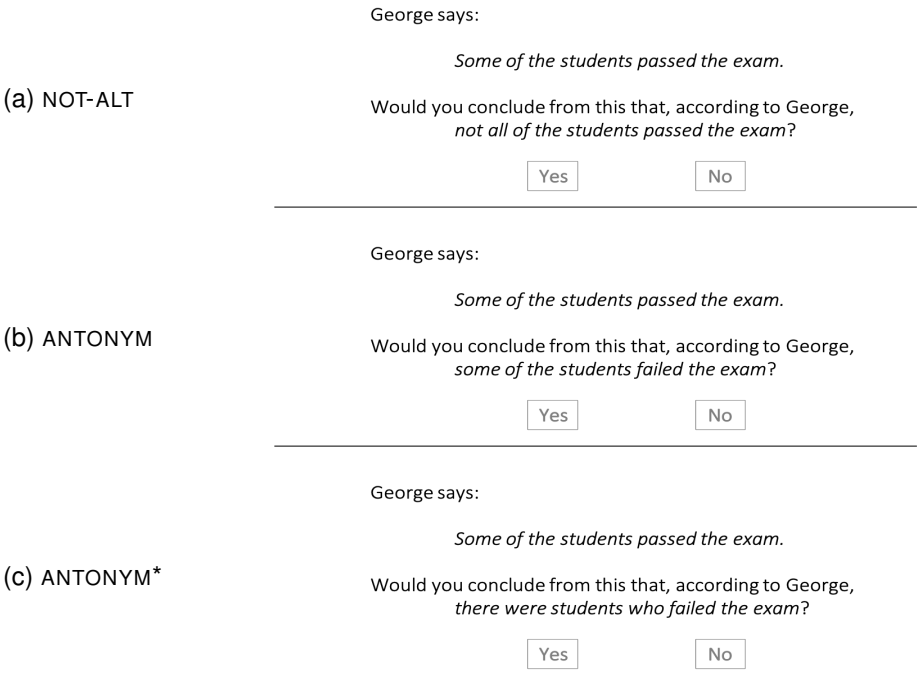


Figure 1: Example test trials in the (a) NOT-ALT, (b) ANTONYM and (c) ANTONYM* conditions, here for the scale ⟨some, all⟩. A Yes-response in these trials indicates that an SI is drawn.

Table 1: Scales tested in the experiment by category.

Category	Scales
Adjective	⟨possible, certain⟩, ⟨good, excellent⟩, ⟨difficult, impossible⟩
Adverb	⟨sometimes, always⟩
Connective	⟨or, and⟩
Determiner	⟨some, all⟩, ⟨a few, a lot⟩
Verb	⟨allow, require⟩, ⟨may, have to⟩, ⟨permit, require⟩, ⟨try, succeed⟩, ⟨participate, win⟩

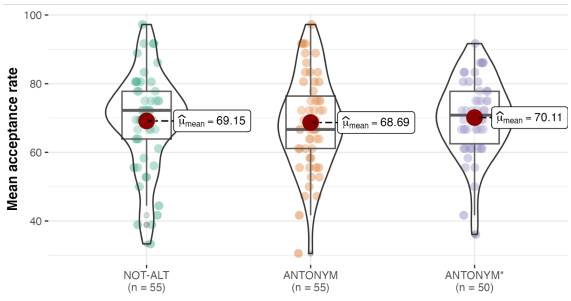


Figure 2: Percentage of Yes-responses to the test trials by Probe condition.

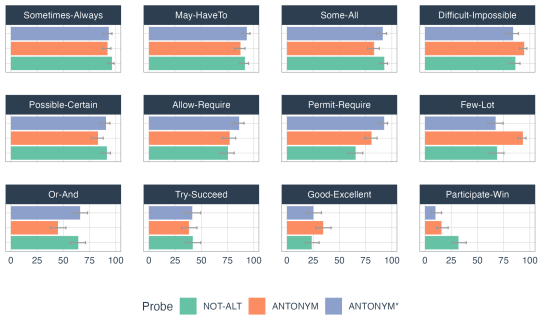


Figure 3: Percentage of Yes-responses to the test trials by Scale and Probe condition.

References [1] Geurts, B. & Pouscoulous, N. (2009). *Embedded implicatures?!?* [2] van Tiel, B., van Miltenburg, E. Zevakhina, N. & Geurts, B. (2016). *Scalar diversity* [3] van Tiel, B., Pankratz, E. & Sun, C. (2019). *Scales and scalarity: Processing scalar inferences.* [4] Barner, D., Brooks, N. & Bale, A. (2011). *Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference* [5] Rees, A. & Bott, L. (2018). *The role of alternative salience in the derivation of scalar implicatures.* [6] Skordos, D. & Papafragou, A. (2016). *Children's derivation of scalar implicatures: Alternatives and relevance.* [7] Marty, P., Romoli, J., Sudo, Y., & Breheny, R. (2024). *Implicature priming, salience, and context adaptation.*

Focus slowdowns arise due to the computation of alternative sets, not unpredictability

Comprehenders have been argued to expend more resources processing foci than non-foci, as evinced by longer reading times [1-4] and more accurate responses in both memory [5-6] and error/change detection tasks [7-8]. In three reading studies, we disentangle four potential causes for these focus effects, listed in (1). Under (1a), slowdowns on foci have been explained by appealing to newness [2], which should always require more processing effort than material that has recently been processed, but foci need not be new [9]. Experiments (E)1-3 here found slowdowns on given foci [4] and fully predictable foci, contra what would be expected if material were more costly to process for the reasons given by (1b-c). We argue that such slowdowns are instead driven by (1d), the computation of *contrastive alternatives,* i.e., expressions that can substitute for and contrast with the focus [10]. This suggests that the allocation of resources is guided, not just by prioritization of importance or (un)predictability, but also by representations of the relevant contrasts in discourse that are not reducible to non-linguistic concepts.

E1. (n=56) used context questions in different conditions to manipulate the Size of focus in a subsequent target sentence (held constant within each item), to obtain reading time measures on wide and given foci. Of particular interest was whether readers would slow down on the left edge / beginning of a wide focus. 60 target sentences as in (2) were presented using the Maze task [11-12]. Bayesian mixed effects models in `brms` [13] were fit to log and raw RTs on all |target| regions. Only effects reliable in both measures are reported here (Table 1). **Results.** Models revealed reliable slowdowns on the verb in the VP focus condition, on the first noun in the NP1 focus, and on the second noun in the NP2 focus condition, thus replicating the given focus slowdown and indicating slowdowns throughout foci larger than one word. Focus slowdowns thus cannot be explained by newness. But, since the goal of conversation is often taken to be expansion of the common ground [14-15], perhaps focus slowdowns arise because comprehenders spend more time reading information not already established in the common ground (1b). Or, (1c) since conversation may primarily be involved with the resolution of a series of (implicit) questions [16], foci may slow down reading because they answer such questions.

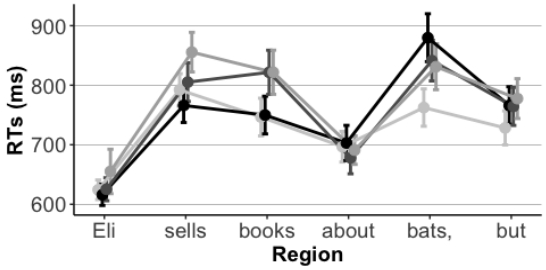
E2. (n=48) crossed focus Size (WIDE vs NARROW) with focus Type (NEW focus vs second-occurrence focus/SOF) to test this. Target foci in SOF conditions were always entailed by their contexts ((3a) already entails that someone read a book about bats), and answered neither an explicit question nor the current (implicit) QUD, e.g., in the context of (3a) this would be *Who only read a book about bats?* The |target| region in these stimuli was always the first object NP as this word was focused in the WIDE but not the NARROW conditions, and WIDE-NARROW RT differences there thus index focus marking. Maze RTs for 48 items like (3) were analyzed as in E1. **Results.** Models revealed a main effect of focus Size (faster RTs in WIDE than NARROW conditions), a main effect of focus Type (faster RTs on SOF than NEW foci), and an interaction between focus Size x Type, such that the focus Type effect was only reliable in the WIDE focus conditions. E2 thus found wide and given focus slowdowns even for SOF foci. This suggests that comprehenders generally encode what contrastive alternatives are relevant in a discourse context, and that contrast among such alternatives guides the allocation of resources during sentence comprehension, not newness, entailment or answerhood. **E3.** aimed to show that contrast plays a role in discourse comprehension even when the need to consider alternatives is not explicitly signaled by a particle. The particle was removed from E2's SOF materials, thus creating conditions in which the |target| was either the second occurrence of a BOUND focus as in (4b) or that of a FREE focus as in (4d). **Results** again revealed both a main effect of focus Size and focus Type, as well as an interaction indicating focus slowdowns in both BOUND and FREE conditions. **In sum,** these findings go against a general understanding in which linguistic material expressing less crucial information is somehow more shallowly parsed. Future work should determine whether the obtained effects carry over to other measures in which effects of focus have been found.

(1) **Potential reasons for prioritization:** Material is prioritized if it is…
(a) discourse new (i.e., unmentioned, see [6])
(b) not already entailed by the discourse context
(c) an answer to the current question-under-discussion (QUD)
(d) standing in some (implicit) contrast with other material

(2)
a. Does Eli sell books about **bats**, or **gorillas**? — NARROW NP
b. Does Eli sell **books about bats**, or **articles about gorillas**? — WIDE NP
c. Does Eli **sell books about bats**, or **edit articles about gorillas**? — VP
d. Does Eli sell books about bats? — BROAD
Target: I think Eli said he |sells|$_1$ |books|$_2$ about |bats|$_3$, but I'm not entirely sure.

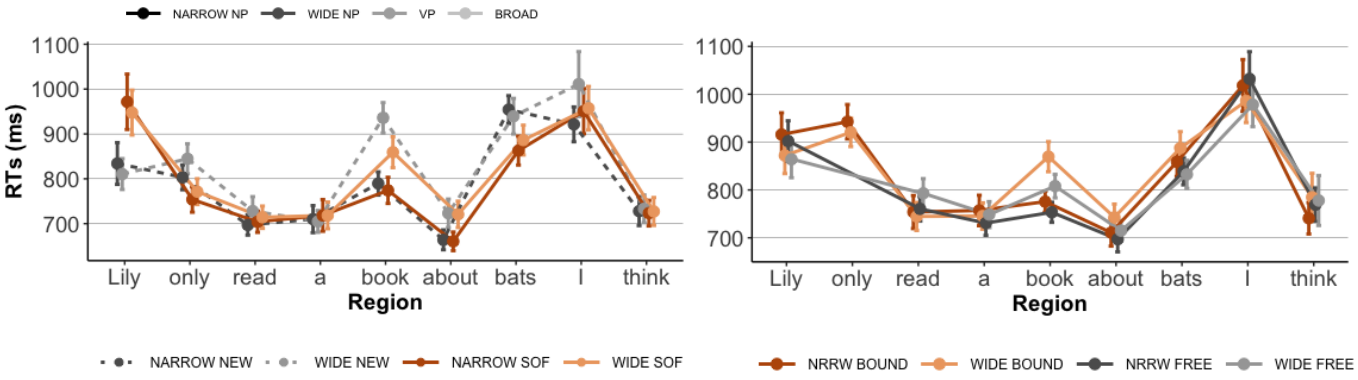
(3)
a. Abby read a book about **penguins$_F$**, Bob read a book about **gorillas$_F$** — NARROW NEW
b. Abby read **a report about penguins**, Bob read **an article about gorillas$_F$** — WIDE NEW
c. Abby read a book about **penguins$_F$**, but Bob only read a book about **bats$_F$** — NARROW SOF
d. Abby read **a report about penguins**, but Bob only read **a book about bats$_F$** — WIDE SOF
Target: {And (NEW) | No, (SOF)} Lily$_{F1}$ only read a |book| about bats

(4)
a. Abby read a book about **penguins$_F$**, but Bob only read a book about **bats$_F$** — NARROW BOUND
b. Abby read **a report about penguins**, Bob only read **a book about bats$_F$** — WIDE BOUND
c. Abby read a book about **penguins$_F$**, but Bob read a book about **bats$_F$** — NARROW FREE
d. Abby read **a report about penguins**, but Bob read **a book about bats$_F$** — WIDE FREE
Target: No, Lily$_{F1}$ { only (ASSOC) | __ (FREE) } read a |book| about bats



	Verb (*sells*)			NP1 (*books*)			NP2 (*bats*)		
	β	**(err.)**	**95% Cr.I.**	**β**	**(err.)**	**95% Cr.I.**	**β**	**(err.)**	**95% Cr.I.**
Intcpt	2.88	(.01)	[2.86,2.9]	2.85	(.01)	[2.83,2.87]	2.86	(.01)	[2.84,2.89]
NRRW	-.01	(.01)	[-.03,.00]	0.00	(.01)	[-.01,.02]	0.06	(.01)	[.04,.07]
WIDE	0.00	(.01)	[-.01,.02]	0.04	(.01)	[.02,.05]	0.04	(.01)	[.02,.05]
VP	0.03	(.01)	[.01,.05]	0.04	(.01)	[.03,.06]	0.03	(.01)	[.01,.05]

Table 1: Posterior estimates per region of E1 (logRTs)



	β	**(error)**	**95% Cr.I.**
Intcpt	2.90	(0.01)	[2.87, 2.92]
Type	0.05	(0.01)	[0.02, 0.04]
Size	0.03	(0.01)	[0.03, 0.07]
Type x Size	0.04	(0.01)	[0.01, 0.06]

Table 2: Posterior estimates of E2 (logRTs)

	β	**(error)**	**95% Cr.I.**
Intcpt	2.88	(0.01)	[2.85, 2.91]
Type	0.04	(0.01)	[0.03, 0.05]
Size	0.02	(0.01)	[0.01, 0.03]
Type x Size	0.01	(0.01)	[–.01, 0.03]

Table 3: Posterior estimates of E3 (logRTs)

References [1] Birch & Rayner (1997) *Mem. & Cogn.* **[2]** Benatar & Clifton (2013) *JML.* **[3]** Lowder & Gordon (2015) *Psych. Bulletin & Review.* **[4]** Hoeks et al., (2023) *JML.* **[5]** Birch & Garnsey (1995) *JML.* **[7]** McKoon et al. (1993) *JML.* **[8]** Bredart & Modolo (1988) *Acta Psychologica.* **[9]** Sanford & Sturt (2002) *Trends in Cog. Sci.* **[10]** Rooth (1992). *Nat. Lang. Semantics.* **[11]** Forster et al. (2009) *Behav. Res. Methods.* **[12]** Boyce et al. (2020) *JML.* **[13]** Bürkner (2017) *J. Stat. Soft.* **[14]** Stalnaker (1978) *Pragmatics.* **[15]** Lewis (1979) *Semantics from Different Points of View.* **[16]** Roberts (2012) *Sem. & Prag.*

Fake reefs are sometimes *reefs* and sometimes not, but are always compositional

Summary. In semantics, adjective modification is typically handled with set intersection, such that ⟦yellow flower⟧ = ⟦yellow⟧ ∩ ⟦flower⟧. Thus a *yellow flower* is a *flower*. Such an account, however, runs into problems for adjectives like *fake* or *counterfeit*, which typically have a privative entailment: a *fake fire* is not a *fire* and a *counterfeit dollar* is not a *dollar*. Moreover, privativity cannot easily be encoded as a property of adjectives like *counterfeit*, since e.g. a *counterfeit watch* is judged to be a *watch*, a subsective entailment (Martin, 2022). We gather judgments on over 300 English adjective-noun bigrams (57 novel; i.e., zero corpus frequency), and show that privativity depends on the adjective, noun and context, and can be manipulated for the very same adjective-noun bigram by presenting it in different contexts. This is difficult to explain if privativity is seen as a property of the adjective (del Pinal, 2015; Partee, 2010). Moreover, we find no difference between novel AN bigrams and high frequency ones, suggesting that this is still a case of productive composition and not the result of convention or memorized idiosyncrasy. Our results support compositional accounts like Martin (2022) and Guerrini (2022) which treat privativity as context-dependent.

Data. We test 305 adjective-noun bigrams obtained by crossing 38 nouns with 12 adjectives, filtering out bigrams rated to be impossible to assign a meaning in a separate study. 6 typically privative adjectives are matched with 6 typically subsective adjectives of similar corpus frequency: *artificial, counterfeit, fake, false, former, knockoff; homemade, illegal, multicolored, tiny, unimportant, useful*. The nouns are selected to yield a high quantity of zero-frequency bigrams (19% after filtering), as counted in a ∼200B word corpus (Raffel et al., 2020). Representative high-frequency bigrams include *fake fire* and *counterfeit watch*; zero-frequency bigrams include *fake reef* and *false concert*.

Experiment 1. We recruited 510 native English speakers on Prolific (15 excluded). Each participant saw 12 questions (of which 4 fillers) of the form *Is an A N still an N?* (Fig. 1), yielding 10+ ratings/item. Mean bigram ratings are shown in Fig. 3. We find that each "privative" adjective yields graded variation from privative to subsective depending on the noun, and that "subsective" adjectives are less clearly subsective with certain nouns (e.g. *homemade cat*). Further, we find no effect of frequency on rating variance (typ. subsective: $R^2 = 0.009$, typ. privative: $R^2 = 0.014$), showing that participants behave similarly for high-frequency and novel adjective-noun bigrams, rather than e.g. having a conventionalized/memorized meaning or entailment only for high-frequency bigrams. Moreover, some zero frequency bigrams like *knockoff image* have quite low variance ($\mu = 4.90, \sigma^2 = 0.10$), showing that participants compose even novel bigrams systematically.

Experiment 2. We select 6 pairs of AN bigrams from Experiment 1 with similar middling ratings and high variance, such that one bigram is zero/low frequency and the other is high frequency: *counterfeit diamond/dollar, fake reef/fire, fake scarf/drug, fake glance/plan, false concert/war* and *former accusation/house*. For each, we construct two contexts designed to bias the reader towards a subsective or privative entailment respectively (Fig. 2). We recruited 40 native English speakers

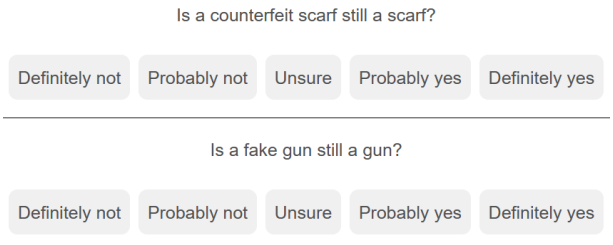


Fig. 1: Sample questions in Exp. 1.

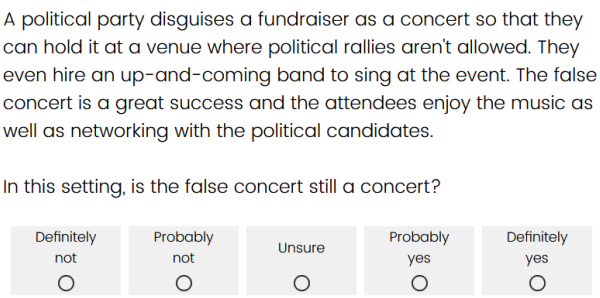


Fig. 2: Subsective-biased context in Exp. 2.

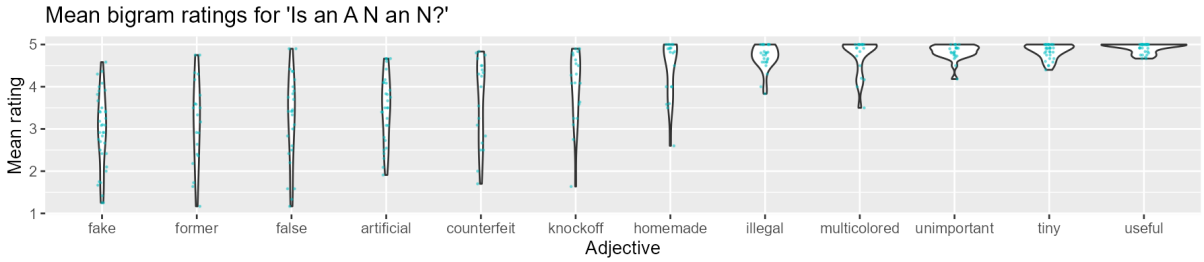


Fig. 3: Mean bigram ratings for Exp. 1, where 1 is most privative and 5 is most subsective.

on Prolific (1 excluded); each participant saw 12 items (of which 6 fillers), yielding 10 ratings/item. We find that for some bigrams (*fake fire, fake plan, false concert*), the contexts bias participants' entailments very effectively, though other bigrams have more mixed results (Fig. 4) due to item-specific effects (*counterfeit dollar*) or unintended effects of the specific context wording (*fake reef*). We conclude that these entailments are indeed context-dependent and that variation in imagined context may explain some of the variance in Exp. 1. Further, we see no frequency-related patterns in this experiment (e.g. high-frequency bigrams like *fake fire* having less manipulable entailments), showing that deriving entailments from AN bigrams is not conventionalized/memorized and is instead derived from productive use of world knowledge and context. Finally, the ability to manipulate the entailments of novel bigrams such as *false concert* again supports a compositional account.

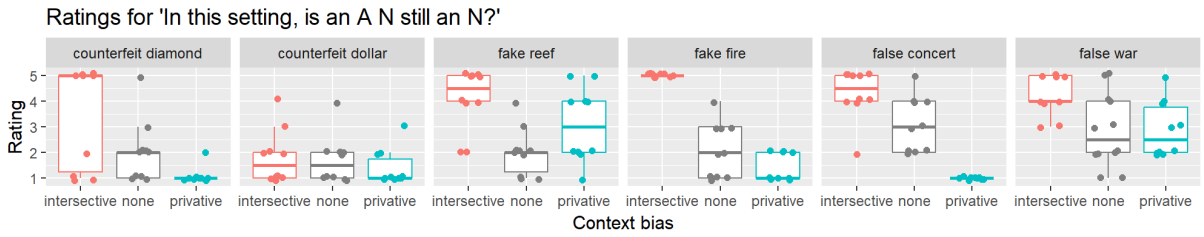


Fig. 4: Selected results for Exp. 2, where 1 is most privative and 5 is most subsective. The ratings from Exp. 1 are shown in gray.

Discussion. Our experiments reveal significant variation in privative entailments among so-called privative adjectives and pose problems for any theory (del Pinal, 2015; Partee, 2010) which treats privativity as a property of the adjective. We find that the entailment drawn depends on the adjective and noun (Exp. 1) as well as the context (Exp. 2). This noun and context-dependent variation is equally possible with novel adjective-noun bigrams, and we do not find any effects of frequency/convention, supporting a compositional account of adjective-noun modification nonetheless. One way to capture within-adjective variation without resorting to polysemy is by adapting del Pinal's qualia-based proposal (Martin, 2022): first, all adjectives compose with nouns as functions over noun qualia. For example, *fake* overwrites the telic and agentive qualia of *gun*. A second step evaluates this new bundle of qualia for noun membership to derive subsective/privative entailments. We can adapt this second step to account for context, which influences which qualia matter for determining noun membership. More broadly, the data from these experiments open the door for more detailed accounts which explain how exactly each case of variation is derived.

References. • Del Pinal, G. (2015). "Dual Content Semantics, privative adjectives, and dynamic compositionality". *Semantics and Pragmatics* 8 • Guerrini, J. (2022). "Keeping fake simple". *LingBuzz* • Martin, J. (2022). "Compositional Routes to (Non)Intersectivity". PhD thesis • Partee, B. H. (2010). "Privative adjectives: Subsective plus coercion" • Raffel, C. et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *JMLR* 21.140

Disagreements do not automatically raise the standard of precision

Speakers often choose to utter imprecise sentences that strictly speaking are false (1a). The *standard of precision* [1-3] governing a discourse can be negotiated through metalinguistic disagreements: in (1b), Andy's challenge signals that a stricter standard of precision (SoP) should be adopted. Here we investigate whether metalinguistic disagreements like (1b) result in an automatic update of the SoP. We consider two hypotheses: **Hypothesis 1 (H1)** states that challenging the SoP automatically updates this discourse parameter, superseding previous parametrizations. **Hypothesis 2 (H2)** states that metalinguistic challenges act as a request to shift the SoP, but do not directly update it. Unlike H2, H1 predicts that disagreements should decrease the acceptability of a previous imprecise utterance. Contra H1, we find that imprecise utterances continue to be perceived as felicitous even after the SoP has been challenged, suggesting that any potential updates to the SoP ought to take place in subsequent conversational moves.

Experiment 1 (Exp1): We created twenty-four five-point scales instantiating different Maximum Standard adjectival properties (e.g., *empty*) to varying degrees (Fig.1a). Each scale was normed (n=30) to ensure that the lower scalepoints (1-4) tolerated some amount of imprecision. The goal of Exp1 (n=30) was to gather interpretational preferences for individual scale points in isolation to be used as a baseline in the analysis of Experiment 2. Participants saw individual images accompanied by a description of the form '*This* [object] *is* [adjective]' (Fig. 2a), and were instructed to choose one of three answers: '*Yes*', '*Unsure*' or '*No*'. Exp1 results are shown in Fig. 2b.

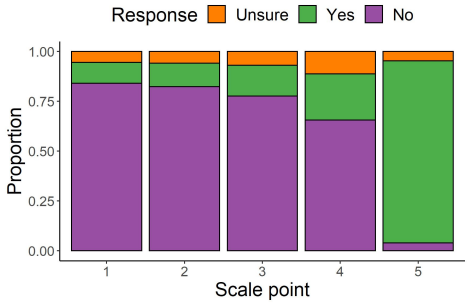
Experiment 2 (Exp2): (n=60) The goal of Exp2 was to assess whether metalinguistic disagreements modulate the acceptability of imprecise utterances. Participants saw the same stimuli used in Exp1 with the only difference that the initial assertion '*This* [object] *is* [adjective]' was followed by an utterance of the form '*No, this* [object] *is not* [adjective]' (Fig. 3a). Participants' task was to choose one of three options: *'Both of them can be right,' 'Only the* {*first, second*} *speaker is right*. The scale points in the 24 scales tested were distributed in 5 lists following a Latin-square design. Twenty-four disagreements about properties not subject to imprecision (e.g., *checkered*) were included as fillers.

Results: Responses (Fig. 3b) were binarized such that selections of '*both of them can be right*' (henceforth *Both*) were coded as 1, while the remaining two levels were coded as 0. A logistic mixed effects regression model was fitted to this new binomial variable using SCALE POINT as a fixed effect. Scale point 5 (S5) was coded as the reference level. Random intercepts and slopes by items and participants were also included. All comparisons were significant, with lower scale points (S1-S4) receiving higher proportions of *Both* responses compared to S5 (all p's < 0.05). Next, we constructed two new binary variables coding whether participants selected '*Only the* {*first, second*} *speaker is right*' (henceforth *First* and *Second*) respectively. The same procedure was followed for Exp1 '*Yes*' and '*No*' responses. The four binomial variables were appended and coded based on 1) whether the observation belonged to Exp1-2 (EXPERIMENT); and 2) whether the imprecise utterance was *accepted* (i.e., '*Yes*' in Exp1, and '*First*' in Exp2) or *rejected* (i.e., '*No*' in Exp1, and '*Second*' in Exp2, see Fig. 4). We refer to this factor as ACCEPTABILITY. A series of mixed effects models were fitted to the data pertaining to each scalepoint, with EXPERIMENT, ACCEPTABILITY and their interaction as fixed effects. Random intercepts and slopes by item and participant were also included. The interactions were significant in S1-4 (all p's < 0.05; S5: p's > 0.05). Simple effect analyses revealed the interactions were driven by higher rates of '*No*' responses compared to *Second* responses (S1-4: all p's < 0.05). No significant differences were detected between '*Yes*' and *First* responses (S1-5: all p's > 0.05).

Discussion & Conclusion. Our results suggest that imprecise utterances are not deemed unacceptable when embedded in a disagreement dialogue. This is shown by the fact that *First* responses were comparable to '*Yes*' responses in S1-4. Conversely, proportions of *Second*—a choice compatible *only* with a higher SoP—were lower than '*No*' responses in S1-4. These lower rates were due to participants displaying a higher preference for *Both*—an option compatible with a lower SoP—in S1-4 compared to S5. The current findings are therefore incompatible with H1, but can be better accommodated by H2. In further research, we address how the discourse commitments [4] incurred by subsequent conversational moves (e.g., concessions, vs. retractions) update the SoP.

(1) a. *Shelly:* This bottle is empty.
b. *Andy:* No, this bottle is not empty, there's a bit of water in it.

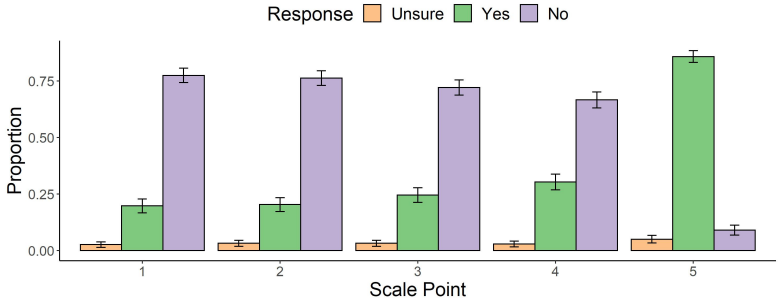
(a) Norming Study Item Example.



(b) Norming Results.

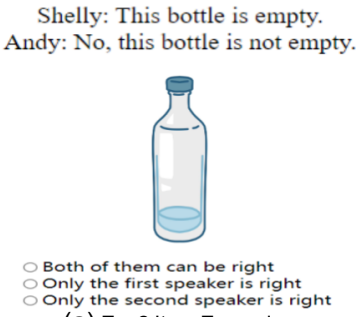
Figure 1: Norming Study.

(a) Exp1 Item Example.

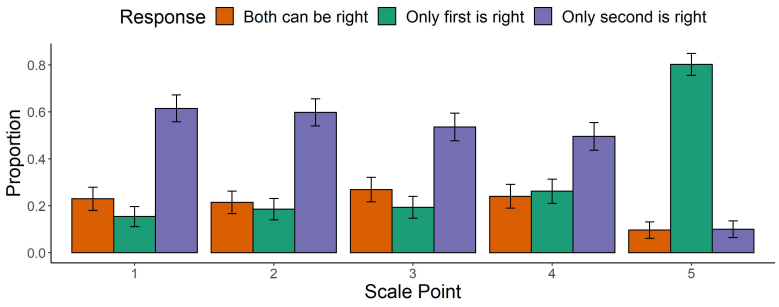


(b) Exp1 Results.

Figure 2: Experiment 1.

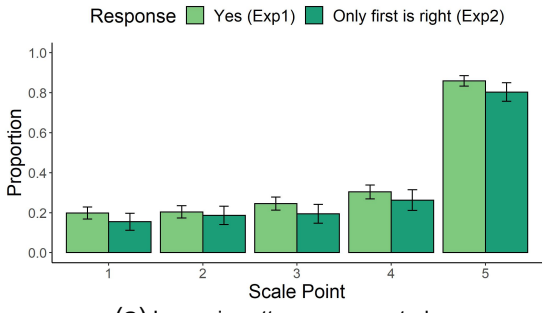


(a) Exp2 Item Example.

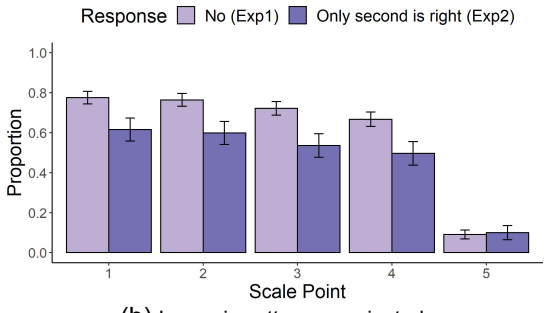


(b) Exp2 Results.

Figure 3: Experiment 2.



(a) Imprecise utterance accepted.



(b) Imprecise utterance rejected.

Figure 4: Exp1-2 comparison.

References: [1] Lewis, D. (1979). *Scorekeeping in a language game.* J. Philos. Log. | [2] Lasersohn, P. (1999). *Pragmatic halos* Lang. | [3] Klecha, P. (2018). *On unidirectionality in precisification*. L&P. | [4] Lauer, S. (2012). *On the pragmatics of pragmatic slack*. Proc. SuB.

Semantic/pragmatic universals and variation via crosslinguistic experimentation

Kate Davidson
(Harvard)

Formal theoretical approaches to semantics and pragmatics have for most of their history tended to focus more on universals than variation, with several notable exceptions in areas such as definiteness, tense, modals, quantificational structures, and expressions of gradability. Similarly, experimental approaches to meaning -both in psycholinguistics and experimental pragmatics- have tended to rely on data from a small number of well studied languages. In this talk I will present three recent studies in our lab that use experimental methodologies to directly probe crosslinguistic semantic variation and the related crosslinguistic variation in available pragmatic alternatives (comparing and contrasting among both signed and spoken languages), highlighting the value to linguistic theory of directly applying insights from semantic fieldwork to experimentation and bringing experimental methodology to semantic fieldwork, and how we have dealt with both linguistic and logistical challenges in collecting this data.

An experimental investigation of perspective alignment in gesture and speech

Summary. Hinterwimmer et al. (2021) experimentally investigated the hypothesis that perspective in gesture and speech are by default aligned, i.e., when a character's or protagonist's perspective is conveyed in the speech signal, this utterance is preferably aligned with a *character viewpoint gesture* (CVG). If an utterance expresses an observer's perspective, by contrast, it is more likely accompanied by an observer viewpoint gesture (OVG). Their results, however, showed an overall preference for CVGs. They argued that there were pragmatic factors (e.g., informativity) at play blocking the hypothesized perspective alignment. The study reported here further investigates Hinterwimmer et al.'s (2021) hypothesis by comparing two different CVGs paired with a verbal utterance in a rating study. The results suggest that, contrary to Hinterwimmer et al.'s (2021) hypothesis, multiple perspectives can be simultaneously expressed in gesture and speech.

Background. Normally, an utterance expresses the speaker's perspective or viewpoint. Therefore, all perspective-dependent expressions (e.g., relational expressions such as *left* and *right*) are by default interpreted from the speaker's perspective (e.g., Harris and Potts, 2009). It is possible, however, to shift the perspective from the speaker to some other individual which is salient in the current discourse. Examples are instances of *reported speech*. Perspective can also be expressed in gesture (McNeill, 1992). A common distinction in this line of research is the one between CVGs and OVGs. CVGs depict an event from a first-person perspective, OVGs depict an event from a third-person perspective. There is very little research on how perspective taking in the two modalities interacts. Hinterwimmer et al. (2021) posited the hypothesis that the perspectives expressed in gesture and speech should be aligned. They ground their hypothesis on previous research which has found that i) gesture and speech convey a joint multimodal message which is planned by one central cognitive process and later passed on to different communication channels (e.g., de Ruiter, 1998) and ii) perspective in gesture and speech have the same conceptual source (Parrill, 2010). In order to test their hypothesis, they designed a forced-choice study where they paired videotaped utterances in *free indirect discourse* (FID), which clearly expressed a salient protagonist's perspective, or a more general statement describing an event from an observer's perspective with a CVG and an OVG. Participants then had to select the version of the utterance which they considered more natural. They predicted that CVGs were preferred in the FID condition, while OVGs were predicted to be preferred in the condition where the event was described from an observer's perspective. However, contrary to their hypothesis, they found an overall preference for CVGs regardless of the perspective expressed in speech. Hinterwimmer et al. (2021) hypothesized that this might be due to pragmatic factors which block the default perspective alignment, e.g., that CVGs are more salient than OVGs due to their differences in size. This was experimentally validated by Walter et al. (2023). Therefore, the hypothesis that perspective in gesture and speech are preferably aligned is difficult to investigate when comparing CVGs and OVGs. It thus seems more promising to investigate cases where one can compare two occurrences of the same type of viewpoint gesture. In sentences where two protagonist's perspectives are introduced, it is therefore hypothesized that a CVG conveying the more prominent protagonist's perspective is preferred over a CVG conveying the less prominent protagonist's perspective.

Experimental study. An experimental rating study was conducted (2x2 design) in order to investigate this hypothesis. In each item, two protagonist's perspectives were introduced: a prominent one and one which was less prominent. The sentences were either aligned with a CVG from the more prominent protagonist's perspective or a CVG conveying the less prominent protagonist's perspective (factor Gesture). Moreover, the more prominent perspective was either introduced from a first-person perspective via a first-person pronoun or from a third-person perspective via a proper name (factor Referential Expression). The second protagonist was always introduced by an indefinite. 24 experimental items were construed along the lines of the example in (1). The exper-

imental items were split up according to a Latin square design and interspersed with 25 fillers. 40 native speakers of German were recruited via Prolific for participation. Participants had to rate the items on a 7-point Likert scale for naturalness (1 = completely unnatural; 7 = completely natural).

(1) a. Gestern Abend ist mir etwas Krasses passiert. Ich war im Park spazieren und auf einmal kam ein Typ auf mich zu und hat mich ohne Vorwarnung so heftig geschubst, dass ich fast hingefallen wäre, weil ich das Gleichgewicht verloren habe.

b. Gestern Abend ist Paula etwas Krasses passiert. Sie war im Park spazieren und auf einmal kam ein Typ auf sie zu und hat sie ohne Vorwarnung so heftig geschubst, dass sie fast hingefallen wäre, weil sie das Gleichgewicht verloren hat.

'Yesterday evening something crazy happened to me/Paula. I/she was taking a walk in the park when suddenly some guy walked to me/her and nudged me/her so strongly that I/she nearly fell because I/she lost my/her balance.'

Prominent CVG: Speaker is staggering backwards and flailing about.

Not prominent CVG: Speaker performs a nudging gesture.

In (1) the CVG conveying the backwards staggering aligns with the perspective which is more prominent on the level of the speech signal, since it conveys the speaker's (1a) or Paula's perspective (1b) on the described event. It should therefore be preferred over the nudging CVG, which expresses the perspective which is less prominent on the level of the speech signal. Based on the hypothesis that there is perspective alignment in gesture and speech, a main effect for Gesture is predicted. Moreover, since introducing a perspective by means of a first-person pronoun makes that perspective even more prominent, an interaction between the two factors is predicted. The results show that the conditions were all rated equally well (first-person + prominent CVG: M = 5.43, SD = 1.53; first-person + not prominent CVG: M = 5.39, SD = 1.47; proper name + prominent CVG: M = 5.47, SD = 1.43; proper name + not prominent CVG: M = 5.33, SD = 1.53). An ordinal mixed-effects model was fitted onto the data and yielded neither a main effect for the factor Gesture nor significant interactions.

Discussion and conclusion. The results show that both CVGs were equally acceptable regardless of the prominent perspective in the speech signal. Moreover, the factor Referential Expression did not have any influence on the ratings either. The results thus do not confirm the hypothesis that perspective in gesture and speech are preferably aligned. In contrast to Hinterwimmer et al.'s (2021) study there were no intervening pragmatic factors which might have blocked perspective alignment in this study. The most plausible conclusion is therefore to reject the hypothesis that there is a preference for perspective alignment in the two modalities. Rather, multiple perspectives can be simultaneously expressed. Future research should investigate whether there are any constraints for expressing multiple perspectives in the two modalities.

References

de Ruiter, J. P. (1998). *Gesture and Speech Production*. University of Nijmegen PhD Thesis.

Harris, J. A. and C. Potts (2009). Perspective-shifting with appositives and expressives. *L&P, 32(6)*.

Hinterwimmer, S., U. Patil, and C. Ebert (2021). On the interaction of gestural and linguistic perspective taking. *FiC, 6*.

McNeill, D. (1992). *Hand and Mind: What Gestures reveal about Thought*. UChicago Press.

Parrill, F. (2010). Viewpoint in speech-gesture integration: Linguistic structure, discourse structure, and event structure. *LaCP, 25, 5*.

Walter, S., C. Ebert, and S. Hinterwimmer (2023). Are there salience differences between character and observer viewpoint gestures? Poster at *XPrag X*.

‘Exhausting’ Theory of Mind resources impairs speaker-specific lexical alignment

Speakers can recognize inter-speaker variability in various pragmatic phenomena and adapt to the speakers’ different preferences of language (e.g.,[1],[2]). Furthermore, it has been repeatedly shown that interlocutors align with regard to their referential choices in what is commonly known as *lexical alignment* [3-6]. Moreover, we recently showed that in addition to alignment, individuals actively store speaker-specific lexical ‘stylistic’ choices, and that they use this knowledge to generalize speaker-specific information both in the linguistic and the social domains [7]. In this study, we aimed to examine the cognitive processes involved in the different stages of detecting, aligning with-, and generalizing speaker-specific language use. Specifically, we were interested in examining how these phenomena relate to (a) Theory of Mind (ToM), a social function and (b) Executive Functions (EF). It has been shown that performing cognitively demanding tasks can interfere with performance in subsequent language-related tasks [8]. Following this, we examined in this study whether performing a task that requires either using ToM (Reading the Mind in the Eyes Test (RMET)[9]), or inhibition-control (Flanker [10]) interferes with the ability to store, generalize, and align with speaker-specific language use.

Methods. Native Hebrew speakers (N=70, so far) took part in an online interactive picture selection task. Participants were led to believe they were engaging in an interactive task with other naïve participants. In fact, the ‘other participants’ were simulated by a computer program. Each participant was exposed to two different speakers, differing in their naming preferences for real-world objects, such that one speaker consistently produced disfavored words and the other one – their favored alternatives. Participants were assigned to one of three conditions. In one condition, participants performed the Reading the Mind in the Eyes Test before the experimental task; in the second condition, participants performed the Flanker task before the experimental task; the third condition was a control condition in which participants did not perform any task before the experimental task. In the experimental task, there were two different roles: *Directors* and *Matchers*. *Directors* instruct *Matchers* which image to choose, within an array of real-world objects. There were 5 steps in the task, always presented in the same order. (1) In the exposure phase, the participants acted as *Matchers* and were instructed by two other simulated participants (each in their turn) which image to choose. (2) In the alignment-test phase, participants acted as *Directors* and were required to instruct the simulated participants who were their directors (who supposedly now act as *Matchers*, each in their turn/block) which image to choose. (3) In the detection-test phase, participants were presented with an image on each trial and were asked if one of the simulated participants had used a certain word to describe the image. (4) The linguistic generalization phase included a task similar to the detection-test in which we asked participants if it is possible – hypothetically – that a given speaker would produce a certain utterance (of three different types - (a) common/uncommon adjective orders; (b) Sentences with non-canonical constituent order; (c) favored/disfavored words). (5) The social generalization phase included a rating task with a visual analog scale – for each speaker - asking about social and personality traits of each speaker (cooperation, book reading, number of friends, non-native language, and autistic traits).

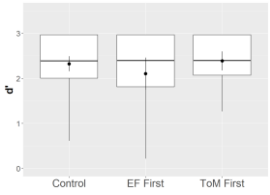


***Figure* 1.** d' by condition.

Results. *Detection.* The ability to correctly map inter-speaker variability was analyzed using the d’ measure of the Signal Detection Theory [11], calculated per participant. In all conditions, the signal (speaker-word association) was reliably detected (control: $t(31) = 14.43$, $p < 0.001$; EF First: $t(19) = 8.03$, $p < 0.001$; ToM First: $t(17) = 18.3$; $p < 0.001$; Fig. 1) The d’ distributions did not significantly differ between the conditions.

Alignment. We fitted a mixed-effects logistic regression model predicting the odds of producing the less common alternative for each image by condition and speaker status. This model revealed a significant interaction ($p < 0.001$; Fig. 2), such that the odds of producing the disfavored word were higher when interacting with the uncommon speaker than with the common speaker, but only in the control ($Z = 8.21$; $p < 0.001$) and in the EF First ($Z = 6.80$; $p < 0.001$) conditions, and not in the ToM First condition ($Z = 1.33$; p = 0.18).

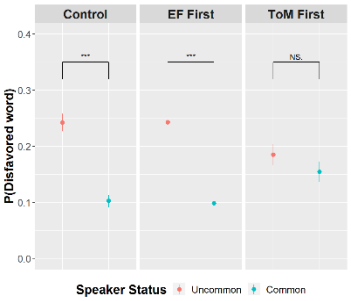


Figure 2. Probability of producing the disfavored word in the alignment-test phase by condition and speaker status.

Linguistic generalization. we conducted a separate analysis for each linguistic phenomenon, fitting three separate logistic regression models considering condition and speaker's status. We included only the uncommon forms of each phenomenon and analyzed the odds of accepting the association of each utterance to a given speaker (Fig. 3). For the lexical items, this model revealed that the odds of a positive response were higher for the uncommon speaker than for the common one, under all conditions. The other two phenomena did not reveal any significant effects.

Social generalization. We analyzed the ratings for each of the 5 questions separately. For each question, we fitted a mixed-effects ordered beta regression model predicting the numeric rating by condition and speaker status (Fig. 4). To sum up the results, in the control condition, we saw effects of speaker status and interactions for the cooperation, number of books, number of friends, and autism questions. These effects were absent in both the EF First and the ToM first conditions. To conclude, using ToM impairs speaker-specific lexical alignment, suggesting ToM is involved in this process. Furthermore, because social generalization was not observed in both the EF-First and in the ToM-First conditions, it seems that generalizing social information based on language-use requires available resources of both abilities.

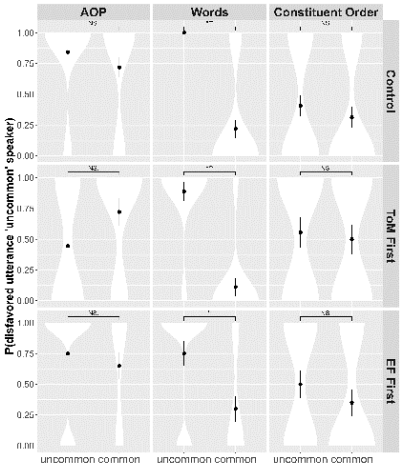


Figure 3. Probability of assigning the disfavored utterance to a speaker by condition and speaker status.

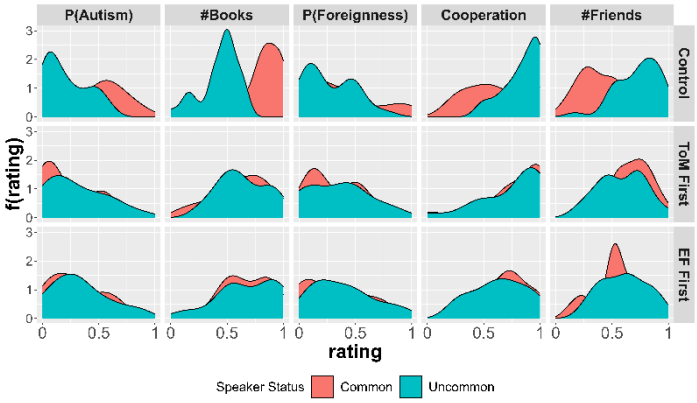


Figure 4. Ratings for each of the social questions, by condition and speaker status.

References. [1] Schuster, S., & Degen, J. (2020). *Cognition*. **[2]** Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). *Frontiers in psychology*. **[3]** Brennan, S. E., & Clark, H. H. (1996). *Journal of Experimental Psychology: Learning, Memory, and Cognition*. **[4]** Brown-Schmidt S. (2009). *Journal of memory and language*. **[5]** Clark, H. H., & Wilkes-Gibbs, D. (1986). *Cognition*. **[6]** Garrod, S., & Anderson, A. (1987). *Cognition*. **[7]** The authors. Under Review. **[8]** Saratsli, D., Trice, K. M., Papafragou, A., & Qi, Z. (2023). *Psyarxiv.* **[9]** Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. and Plumb, I. (2001). Journal of Child Psychology and Psychiatry **[10]** Eriksen, B.A., Eriksen, C.W. (1974). *Perception & Psychophysics*. **[11]** Swets, J. A., Tanner Jr, W. P., & Birdsall, T. G. (1961). *Psychological review*.

Social meaning and pragmatic reasoning: The case of (im)precision

Introduction: A speaker's choice between linguistic alternatives can prompt their hearer to draw pragmatic inferences about facts of the world (e.g. the inference from the utterance of *some* that *all* does not obtain). But such choices can also invite inferences about the properties, ideologies and/or stances of the speaker herself; that is, they can convey **social meaning**. Recently, there has been growing interest in exploring the connections between social meaning (traditionally studied within sociolinguistics; e.g. Eckert 2012) and pragmatic reasoning and processes (Burnett 2019 on sociophonetic variation, Acton 2019 on the definite article, Beltrama & Papafragou 2023 on relevance and informativity). In the present work, we investigate this topic from the perspective of the phenomenon of **numerical imprecision**, i.e. the choice of the level of granularity at which numerical information is reported (e.g. describing a time as *8:03* vs. *around 8 o'clock*). Previous work has shown that the choice of precision level can convey social meaning (Beltrama 2019, Beltrama, Solt & Burnett 2022; the latter henceforth BSB). Speakers who use precise forms are perceived as more intelligent/articulate/confident (status- or competence-related) than those who use approximate forms, but also more pedantic/uptight, while those who use approximate forms are seen as more likeable/friendly/laidback (likeability-related). We extend this research here.

Present research: The goal of the present study is to test the following broad **hypothesis: The social meaning of (im)precision is derived via pragmatic reasoning about the needs of the situation, the epistemic state of the speaker, and the reasons for their choice of form.** Specifically, we hypothesize that the competence-related associations of precise forms derive from the inference that the speaker knows the exact value (i.e. has a high knowledge level), something that often is not the case for a speaker who uses an approximation. Conversely, the likeability-related associations of approximate forms are hypothesized to derive from the inference that the speaker, in a situation where high precision is not required, is rounding off to make the information easier to understand (van der Henst et al. 2002). Finally, the association of precise forms with pedantry derives from the inference that the speaker is being more precise than required in the utterance situation, highlighting their knowledge and not engaging in hearer-oriented simplification. This pragmatic view of social meaning leads to the following **predictions**:

Prediction 1: context dependence: The measured social meaning of (im)precision will be modulated by the utterance context, in particular the degree of precision required: the competence-related associations of precise forms will be most pronounced in a situation where high precision is required (e.g. making a police report), whereas the likeability-related associations of approximate forms and pedantry-related disadvantages of precise forms will be most pronounced in contexts where high precision is not required (e.g. a casual chat with friends). BSB found certain contextual effects of this nature, but these were not entirely robust; this may relate to the complexity of the study design (12 conditions), but also to the fact that the tested scenarios could not be directly linked to contextual precision needs. We address this here.

Prediction 2: correlation with motivations: The social meaning of alternative numerical forms will be correlated with the motivation attributed to the speaker for their choice of form. In particular, if the perceived motivation for the use of an approximate form is lack of precise knowledge, this is expected to correlate with lower competence ratings, whereas if it is desire to make the information easier to understand, this is expected to correlate with higher likeability ratings.

Pretest: As a first step, 16 scenarios were created in which a speaker asks a question requiring a numerical answer; each had 2 versions, one expected to require a precise answer (HighPr), the second expected to prefer an approximate answer (LowPr). These were tested in an online experiment (Prolific; n=174) in which participants saw the scenario/question and 2 possible answers (precise, approx) and indicated which of the two was more appropriate, or if both were equally appropriate. Based on the results, 6 scenarios were selected that showed the greatest difference between HighPr and LowPr, the precise answer preferred in the former and the approximate answer in the latter. These were used as the basis for the main experiment.

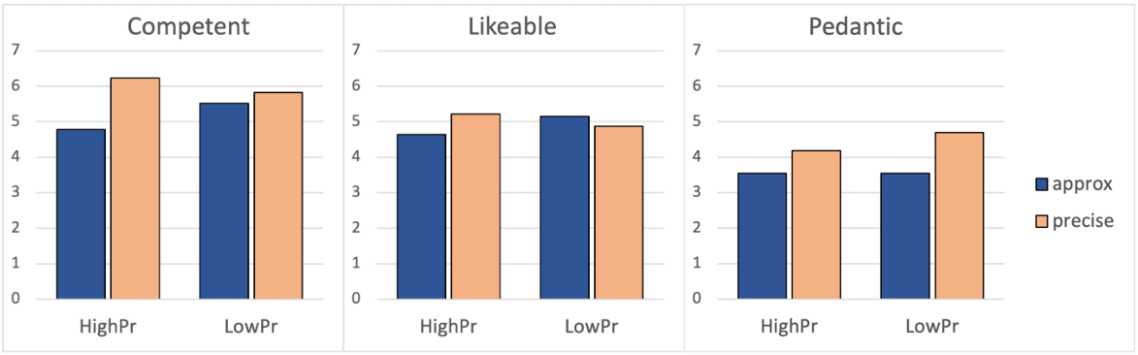


Figure 1. Mean ratings by context and form – selected attributes

Experiment: A pre-registered matched guise study (Campbell-Kibler 2007) was conducted, using as stimuli scenarios (selected via the pretest) in which one speaker asks a question and a second speaker answers it with a numerical expression. Two factors were manipulated: *context*, i.e. required precision level (HighPr, LowPr) and *numerical form* (precise, approx). For example:

HighPrecision: Jamie's new bicycle was stolen. Fortunately it was insured. *Insurance agent: "How much did the bicycle cost? I'll start the paperwork right away."*	**LowPrecision:** Jamie has a new bicycle and is telling a friend about it. The friend is interested and wants to know more**.** *Friend: "How much did the bicycle cost? I'd love to get one like it."*
Jamie: "The bicycle cost $509.55 [precise] / about $500 [approx]"	

The study was executed online via Prolific in a 1-item, fully between-subjects design (n=362 total, ~90/condition, randomly assigned to 1 of 6 scenarios). Participants rated the second speaker on 6 attributes using a 7-point Likert scale: *competent*, *knowledgeable*, *well-prepared* (competence-related), *likeable*, *helpful* (likeability-related) and *pedantic*. They then indicated what motivation they attributed to the second speaker for their choice of form, via free text and multiple choice.

Results are shown in Fig. 1. A linear mixed-effects model was fit to the ratings for each attribute (lmer package in R; Bates et al. 2015), with context, form and their interaction as fixed factors and random intercept for scenario; significance testing was via likelihood ratio. As predicted, for each of the 3 competence-related attributes, a significant main effect of form was found (precise higher; $p < 0.001$ for all), as well as a significant interaction of context and form (greater effect in HighPr; *competent/well-prepared* $p<0.001$, *knowl.* $p<0.05$). No main effect of form was found for *likeable* (a departure from BSB), but as predicted there was a significant interaction of context and form ($p<0.001$), with the relative strength of approximate relative to precise greater in LowPr than HighPr. For *pedantic*, a main effect of form was found (precise higher, $p<0.001$), with no significant interaction though a numerical difference in the predicted direction. Finally, *helpful* patterned (unexpectedly) with the competence-related attributes. Regarding inferred speaker motivations, "to make the information easier to understand" as a reason for using approximate was correlated significantly with higher ratings on *likeable*/*helpful*, whereas "speaker didn't have the exact information" was a near-significant predictor of lower ratings on *competent*/*well-prepared*.

Conclusions and Future Work: The observed effects of context and the correlations between inferred speaker motivations and social meaning are largely in line with the stated predictions, thereby supporting the hypothesized pragmatic source of the social meaning of (im)precision. A follow-up experiment (currently in progress) investigates the further role of speaker knowledge, contrasting the above conditions with ones in which the speaker is known to have the exact information at hand (e.g. a receipt for the bicycle purchase), which we predict will reduce the competence associations of precision and increase the likeability associations of approximation. We furthermore pursue modeling these findings in a probabilistic game-theoretic framework in which social meaning derives from inferences about the speaker's decision strategy.

Expecting the unexpected: Examining the interplay between world knowledge and context in relatively unconstraining scenarios

Real-world implausible information induces processing difficulties unless licensed by the context [1]. However, since most studies used explicit contextual cues to indicate a strong bias towards plausibility violations, it remains unclear how context and world knowledge interact in relatively unconstraining scenarios (e.g., a dream) where both plausible and implausible information seem acceptable. On the one hand, since comprehenders lack enough cues to form a specific prediction that shares a sufficient overlap with the irreal setting of the context, they may expect something real-world plausible (plausibility-driven approach) [2]. On the other hand, since "dreams" are usually associated with unusual events in real life, comprehenders may expect something implausible in a general way even without specific cues (context-driven approach) [3].

Exp 1 (sentence completion task, N = 52) had two conditions: factual versus dream contexts (Table 1, 24 targets, 26 fillers, in English). Each scenario described either a real-life experience or dream, ending with a "preposition + noun phrase" structure. The noun phrase was truncated for participants to complete. If comprehension is plausibility-driven, there should be no difference between the contents of the completions in the two contexts; if comprehension is context-driven, completions should have lower plausibility and higher variability in dream than in factual contexts. **Results*:*** (1) Two raters not involved in the study rated the **plausibility** of completions (Cohen's Kappa = 0.96), and the plausibility was higher in factual than in dream contexts ($p < .001$, using LMM). (2) The **variability** (indexed by entropy) of completions was higher in dream than in factual contexts ($p < .001$, using permutation-based ANOVAs).

Exp 2 (self-paced reading, N = 104) crossed context (factual vs. dream) and plausibility (plausible vs. implausible) in a 2 × 2 within-subjects design (Table 2, 24 targets, 60 fillers, in English). The materials were identical to Exp 1, except that they ended with a critical noun that was either plausible or implausible, followed by spillover regions. If comprehension is plausibility-driven, the same plausibility effect should be found in RTs for both factual and dream scenarios; if comprehension is context-driven, the plausibility effect should be attenuated or even reversed towards the end of the dream scenario but not the factual scenario. Log-transformed RTs were analyzed with LMM. **Results:** (1) **Critical & spill1 regions**: no significant effects. (2) **Spill2 & spill3 regions**: a plausibility effect (spill2: $p = .005$; spill3: $p < .001$). (3) **Spill4 region**: a context × plausibility interaction ($p = .008$), due to a plausibility effect in the factual ($p = .005$) but not the dream condition ($p = .252$). (4) **Spill5 region**: a context × plausibility interaction ($p = .010$). There was a plausibility effect in the factual condition ($p = .046$), but this effect was reversed in the dream condition ($p = .010$) due to longer RTs for plausible words in the dream than factual condition.

Conclusions: The current results provide novel evidence that context is powerful enough to bias comprehension towards world knowledge violations even when there are no explicit constraints indicating this bias (although this effect only emerged at the final region). This indicates necessary extensions for language comprehension models (e.g., the RI-Val Model [3]), by highlighting that information with extremely low cloze probability (i.e., information unrelated to both context and world knowledge in any direct way) can still be preferred in certain scenarios.

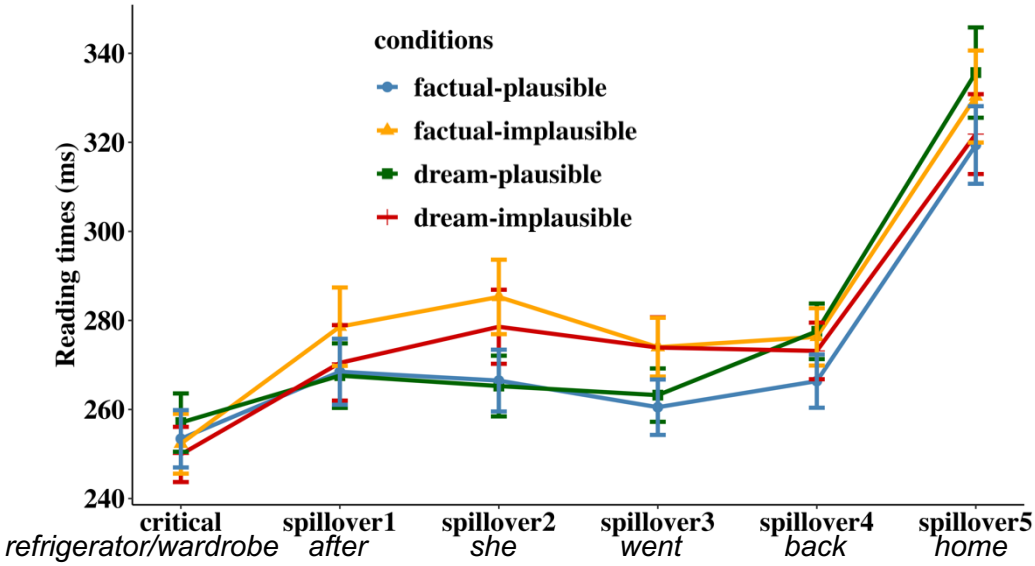
Table 1. Exp 1 example stimuli (a sentence completion task)

Factual	Dream
Mary is telling her friend what she **did** on Sunday. That day, she drove to the nearest grocery store with her husband, bought some fresh meat and vegetables, and then put them in _____.	Mary is telling her friend what she **dreamt** on Sunday. In her dream, she drove to the nearest grocery store with her husband, bought some fresh meat and vegetables, and then put them in _____.

Table 2. Exp 2 example stimuli (self-paced reading, stimuli presented word-by-word)

Factual	Dream
Mary is telling her friend what she **did** on Sunday. That day, she drove to the nearest grocery store with her husband, bought some fresh meat and vegetables, and then put them in the refrigerator$_{\text{plausible}}$ vs. wardrobe$_{\text{implausible}}$ after she went back home.	Mary is telling her friend what she **dreamt** on Sunday. In her dream, she drove to the nearest grocery store with her husband, bought some fresh meat and vegetables, and then put them in the refrigerator$_{\text{plausible}}$ vs. wardrobe$_{\text{implausible}}$ after she went back home.

Figure 1. Exp 2 mean RTs per region (the error bars represent the 95% confidence intervals)



References

[1] Nieuwland, M. S., & van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111.

[2] Cook, A. E., & O'Brien, E. J. (2013). Knowledge activation, integration, and validation during narrative text comprehension. *Discourse Processes*, *51*(1–2), 26–49.

[3] Kuperberg, G. R. (2016). Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, *31*(5), 602–616.

Insensitivity to truth-value in negated sentences: does linear distance matter?

Sentences are usually easier to understand when they are true vs. false, but this generalization is challenged by negated sentences. For example, picture recognition studies have shown faster comprehension of true vs. false affirmative sentences, but a reduced or absent effect of truth-value for negated sentences [1,2]. This motivated the claim that negative sentences like "The package is not wrapped" are understood in two steps: comprehenders first represent the counterfactual/alternate state-of-affairs expressed by the affirmative proposition—'The package is wrapped'—and later, in a second step, represent the actual state. However, recent findings suggest that the representation of a counterfactual state can be diminished or even avoided altogether when negative sentences are pragmatically licensed by context and/or the question-under-discussion is prominent [3,4]. We investigate whether the linear position of the negator in a sentence can similarly modulate the processing of negation. We hypothesized that an earlier negator position may facilitate comprehension by stopping the activation of a counterfactual interpretation, or by facilitating its inhibition. To date, only one study addressed this prediction but it did not find an effect of the negator position [5]. But this study differed from previous studies in that it used brain responses to a single word rather than post-sentence response times. To fill this gap and to examine whether the negator position affects the activation of counterfactual states, we conducted a conceptual replication of [5] with a picture recognition task.

Design. German-speaking adults read 40 sentences word-by-word and decided whether a subsequent picture depicted an object in the sentence. The target answer was always 'yes' for the experimental items (Table 1). Experiment 1 (n = 69) used a Polarity (affirmative/negative) × State-of-affairs (actual/alternate) design. Picture type and the adjectival predicate were used to manipulate the state of affairs, resulting in 8 Latin-square lists (collapsed to four in the analyses). Experiment 2 (n = 72) focused on negative sentences. Following [5], an earlier position of the negator was implemented as a greater linear distance between the negator and the predicate (3–4), compared to a shorter distance (1–2). A Distance (close/far) × State-of-affairs (actual/alternate) design assessed whether more distance enhanced participants' sensitivity to truth value.

Results and discussion. Experiment 1 showed longer response times in negative than affirmative sentences, consistent with more processing difficulty. Further, response times were faster for actual vs. alternate state-of-affairs in affirmative, but not in negative sentences, resulting in a significant Polarity × State-of-affairs interaction ($t = -2.4$, $p = .02$; Figure 1). Thus, sensitivity to truth value was reduced in negated sentences. Experiment 2 did not find evidence of a difference due to the negator's position (non-significant Distance×State-of-affairs interaction: $t = -0.68$, $p = .49$). This was because far distance sentences did not show a facilitation for actual (true) states—descriptively, response times were even longer than in the alternate condition. Thus, there was no evidence that the earlier negator fostered sensitivity to the truth value of negated sentences. Ongoing work is examining whether the type of negation may have influenced the results: While the negation in (1–2) simply negates a specific state of affairs, the negation in (3–4) is a metalinguistic negation, rejecting a previous assertion (e.g., 'The package is wrapped').

Table 1. Sample item in Experiment 2. Experiment 2 featured only negative sentences (all fillers were affirmative). Experiment 1 comprised close distance negative sentences together with their affirmative counterparts (e.g., 'The package is wrapped'). The target picture for the actual conditions is surrounded by a dotted line. The target picture for the alternate conditions is not framed. In another 4 lists, target pictures and predicates were reversed.

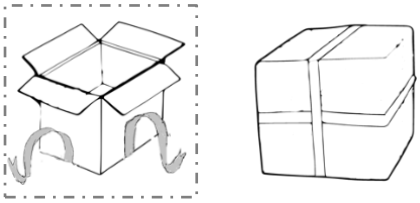
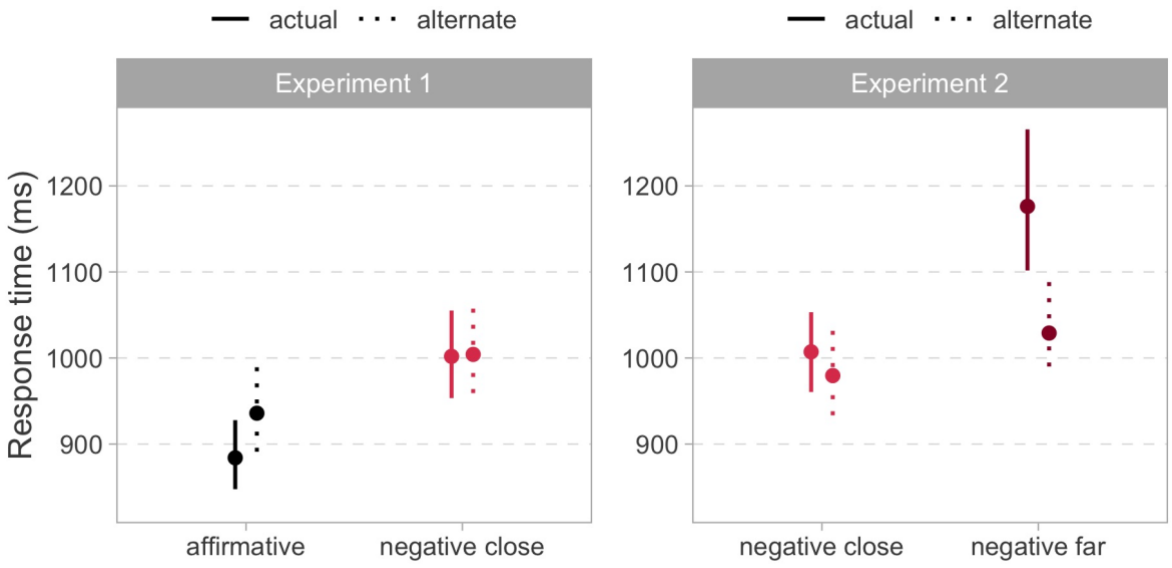
Experimental conditions	Pictures (one picture shown per trial)
1. **Close distance, actual** Das Paket ist nicht eingepackt. ‚The package is not wrapped'	
2. **Close distance, alternate** Das Paket ist nicht ausgepackt. ‚The package is not unwrapped'	
3. **Far distance, actual** Es stimmt nicht, dass das Paket eingepackt ist. ‚It is not true that the package is wrapped'	
4. **Far distance, alternate** Es stimmt nicht, dass das Paket ausgepackt ist. ‚It is not true that the packet is unwrapped'	

Figure 1. By-condition response time averages for correct responses, with error bars showing 95% confidence intervals. Response times are displayed in milliseconds for interpretability, but the statistical analyses were performed on reciprocally transformed response times using linear mixed-effects models with maximal random effects structures by participants and items.



References

[1] Clark & Chase (1982) *Cogn. Psychol*. [2] Kaup et al. (2005) *Cogsci.* [3] Tian et al. (2010) *Q. J. Exp. Psychol.* [4] Darley et al. (2020) *Cognition* [5] Dudschig et al. (2019) *Lang. Cogn. Neurosci.*

Local Accommodation Continues to be Backgrounded

Presuppositions may fail to project, as in (1) below. To derive such local interpretations, standard semantic local accommodation accounts posit an operation in embedding environments that turns content lexically marked as presupposed into non-backgrounded content and conjoins it with the clause's entailed content (Heim, 1983). Such accounts predict that locally accommodated presuppositions (LocAcc) differ from globally projecting ones (GlobAcc) in lacking the presuppositional property of backgroundedness. (A prominent class of recent pragmatic accounts arrives at a parallel prediction via their central claim that all and only backgrounded material projects (Simons et al, 2010; Tonhauser et al, 2018)). However, an experimental study by Siegel and Schwarz (2023) finds LocAcc to be backgrounded: using a picture-matching task in which reduced cognitive salience serves as proxy for presuppositional backgroundedness (Schwarz, 2016), they find evidence for backgrounding of the presupposition of the additive particle *also* in the scope of *if*. The present study extends this approach both empirically and methodologically by testing, in questions, a trigger of a different type, the change-of-state verb *continue*. In order to meet the challenges of testing embedded material not easily pictured, we introduce a novel methodology. Participants are given a task where they must reveal concealed information in order to answer questions or verify statements. The reduced cognitive salience associated with backgrounded material is reflected in what aspects of the interpretation participants attend to in choosing what information to reveal. We compared the hypothesized backgroundedness of LocAcc *continue* to non-presuppositional controls (see details below). Standard LocAcc accounts predict equivalence among these, given their view of LocAcc as non-backgrounded information. But our results indicate that locally interpreted content contributed by *continue* reflects greater backgroundedness than the controls, parallel to Siegel and Schwarz's findings for *also*. A similar pattern holds for global accommodation conditions, supporting parallel backgroundedness across accommodation and trigger types.

Design. We measure the relative attention paid to identical information in 3 conditions, presented by *continue* via LocAcc (1a), by the explicit, non-backgrounded conjunction paraphrases representing their meaning posited by semantic accounts (Heim, 1983) (b), and by a non-presuppositional elision as a further control more closely matching LocAcc surface forms (c). 6 item variants in both LocAcc and parallel GlobAcc examples (2) were shown (both factors between subjects).

(1) (a) I'm looking into Rob's health habits, and I have no idea whether he used to smoke. Is it the case that he continues to smoke now? [CONT condition]
(b) I'm looking into Rob's health habits, and I have no idea whether he used to smoke. Is it the case that he used to smoke, and he smokes now? [CONJ condition]
(c) I'm looking into Rob's health habits, and I have no idea whether he used to smoke. Is it the case that he did, and he smokes now? [DOES condition]

(2) I'm looking into Rob's health habits. I called to find out whether he used to smoke, and it turns out that he continues to smoke now [CONT] **/** that he used to smoke, and he smokes now [CONJ] / that he did, and he smokes now [DOES].

In the critical LocAcc CONT condition (1a), the trigger *continue* conveys presuppositionally that Rob used to smoke, but projection is blocked by the explicit ignorance context in the first clause (Simons, 2001; Abusch, 2010). Control conditions (1b) and (1c) introduce 'Rob used to smoke' as non-presuppositional content: (1b) is the semantic account's conjunctive paraphrase of the local interpretation, differing from (1a) in explicitly mentioning Rob's having previously smoked. (1c) conveys 'Rob used to smoke' implicitly but non-presuppositionally, using ellipsis.

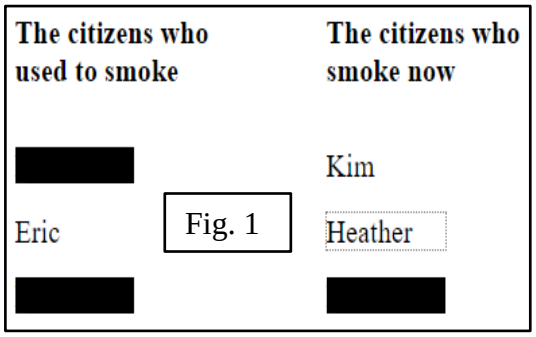


Fig. 1

Task. Participants are told that they will be helping town officials check on outside investigators by trying to answer the investigators' questions highlighted in (1) or verify their claims (2). To do this, participants must seek information about Rob (and other citizens) by clicking to uncover up to three of the black boxes in either of two lists of names we provide. Lists are labelled with the presupposed content of *continue* on the left and its entailed content on the right, as in Fig. 1, in which 3 boxes have been clicked to reveal names. If the information that Rob used to smoke is less salient in (1a), where it is introduced presuppositionally through LocAcc, than in (1b), where it is introduced as an explicit conjunct, we expect more frequent clicks on the righthand column when participants attempt to answer the question in (1a) than when they answer (1b). (1c), in which 'Rob used to smoke' is neither presuppositional nor explicit, controls for potential impact of explicitness independent of backgroundedness. Higher right-click rates for (1a) than for (1c) are thus attributable to *continue*'s presuppositional nature, beyond the implicitness also at play in (1c). GlobAcc (2), where backgroundedness is expected across theories, provides a baseline.

Procedure. 155 participants from our university's subject pool participated online via the PCIbex platform for course credit. Each participant saw 6 critical items representing the 6 item variants, all in a single condition (CONT, CONJ, or DOES) and 21 fillers, in a randomized order.

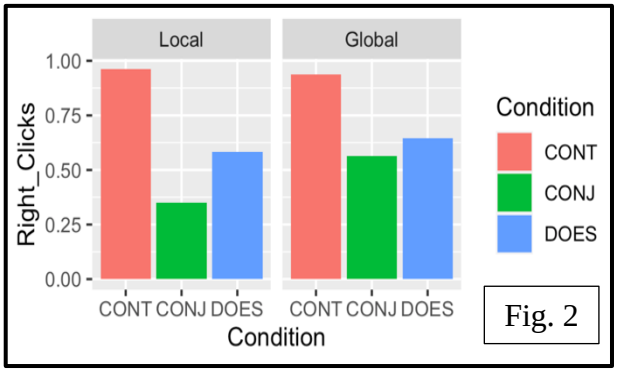


Fig. 2

Results. Participants failing to give the expected answer for 5 of 6 selected fillers were excluded from data analysis, leaving 134 participants. The number of right clicks exhibited the pattern in Fig. 2, with the presuppositional CONT yielding the highest, the explicit conjunctive paraphrase CONJ the lowest, and the elliptical DOES in between. In a mixed effect model analysis, the CONT condition differed significantly from the two non-presuppositional ones, and patterns were similar for LocAcc and GlobAcc in this respect.

Discussion. Using a novel methodology measuring salience during an information-seeking task, we find that the presupposition of *continue* is less salient than its non-presuppositional, lexically equivalent counterparts. In the context of previous findings for *also*, this indicates that the relevant content introduced by presupposition triggers, whether change of state or additive, is lexically encoded as backgrounded, even when interpreted locally, a finding inconsistent with the strongest versions of pragmatic theories. This is of substantial theoretical importance, severing backgrounding from (non-)projection in a way not captured by any existing accounts. Semantic LocAcc accounts a la Heim might be amended accordingly, e.g., by modeling all accommodation as adding information to the relevant context, global or local, in some way that retains its backgrounded discourse status. Other theoretical perspectives will need to explore alternatives to incorporate the implications of this data as well.

References. Abusch, D. 2010. Presupposition triggering from alternatives. *JoS* 27(1)37-80.
Heim, I.1983. On the projection problem for presuppositions. *Proc. WCCFL* 2, 114–125. Stanford.
Schwarz, F. 2016. False but slow. *JoS* 33(1). 177-214.
Siegel, M.and F. Schwarz. 2023. Local Accommodation is Also Backgrounded. *Proc. SuB* 27, 609-624.
Simons, M. 2001. On the Conversational Basis of Some Presuppositions. *Proc. SALT* 11, 431– 448.
Simons, M.,et al. 2010. What projects and why? *Proc. SALT* 20, 309-327.
Tonhauser, J.,et al. 2018. How Projective is Projective Content? *JoS* 35(3),495-542.

The Effect of Experimental Paradigms on Scalar Implicature Estimation

Background: An intriguing feature of human language is the ability to enrich the literal meanings of utterances with pragmatic implicatures (Grice, 1975; Gazdar 1980; Horn 1972; Levinson 2000; Chierchia 2004). Experimental research on the processing and acquisition of Scalar Implicatures (SIs) relies on behavioral tasks that measure the rate at which SIs are computed within an experimental paradigm. Two paradigms have dominated the experimental pragmatics literature: the *Truth Value Judgment Task* (TVJT) (Gordon, 1998) and the *Picture Selection Task* (PST) (Gerken & Shady, 1998). Yet, the effects of task choice on implicature rate has remained underexplored. Here we report the results of three studies testing participants in the TVJT, PST and a variant of the PST called the Hidden Card Task (HCT) using three different linguistic scales in English: “ad hoc”, “or-and”, and “some-all”.

Methods: In Exp.1, participants responded to both TVJT and PST trials in a single Qualtrics survey. In TVJT trials, participants saw a sentence and a card with animal pictures. They were asked to judge the sentence as true or false. In PST trials, participants saw a sentence and two cards. They were instructed to choose the card that best matched the sentence. In TVJT critical trials (Fig.1a), the description was logically true but pragmatically infelicitous. A “false” judgments counted as evidence for SI computation. In the critical PST trials (Fig.1b), the sentence was logically compatible with both cards, but the implicature of the sentence only matched one card, and thus, choosing that card counted as evidence for implicature computation. To make sure that the within-subjects design did not affect the findings, Exp.2 replicated Exp.1 with a between-subjects (TVJT vs. PST) design. Exp.3 examined a variant of the Picture Selection Task called the Hidden Card Task (HCT) which is being increasingly used in the context of priming research (e.g. Bott & Chemla, 2016). The stimuli used in Exp.3 were adopted from the same inventory of stimuli for the PST in Exp.1 and 2 with an important modification: one card in the stimuli was replaced by a “Better Picture” card. For the critical trials, the “Better Picture” card always replaced the card that matched the implicature of the sentence, while for the control conditions, the “Better Picture” card randomly replaced one of the two cards in the trial (Fig.1c). Each experiment had 18 critical trials and approximately 30 to 40 control trials per task. We recruited 50 participants for each experiment.

Results: For all three experiments, the probability of computing SIs was modelled as a function of task type, scale (“some-all”, “or-and”, and “ad hoc”) along with their interactions using logistic mixed-effects models (Bürkner, 2017). We found main effects of task type, scale and their interactions on the estimated rate of SI computation in both Exp.1 (see Fig.2) and Experiments 2-3 (see Fig.3). Compared with the baseline “or-and” trials, participants in PST computed more SIs in “some-all” trials as well as “ad hoc” trials. For the “or-and” trials, the rate of computing SIs in PST (baseline) was the same as that in TVJT (β = 2.50, CI = [-4.17, 9.69]) and HCT (β = 0.05, CI = [-6.12, 6.37]); however, for the “some-all” trials and “ad hoc” trials, the rates of computing SIs were significantly decreased in the TVJT and HCT as compared with PST.

Conclusions: We found that the estimated rate of SIs is significantly affected by the choice of experimental task and lexical scale. For “ad hoc” and “some-all” scales, TVJT and HCT reported a lower implicature rate than PST. There was no difference in implicature rates for the “or-and” scale across the three tasks. These findings suggest that TVJT and HCT can potentially underestimate participants’ pragmatic abilities, which is central to debates in children’s pragmatic development. They also highlight the special status of exclusivity implications and the possibility that they are fundamentally different from (other) SIs. Finally, our studies stress the need for a more careful attention to the pragmatics of experimental tasks themselves and how they affect participants’ linguistic behavior.

References: Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, 91, 117–140. **Bürkner, P. C. (2017).** brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80(1), 1–28. **Chierchia, G. (2004).** Scalar Implicatures, Polarity Phenomena, and the Syntax/Pragmatics Interface, in A. Belletti (ed.), *Structures and Beyond: The Cartography of Syntactic Structures*, Volume 3. OUP, 39-103. **Gazdar, G. (1980)**. Pragmatics and logical form. *Journal of Pragmatics*, 4(1), 1-13. **Gerken, L., & Shady, M. E. (1998)**. The picture selection task. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for Assessing Children's Syntax.* The MIT Press. 125–145. **Gordon, P. (1998)**. The truth-value judgment task. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for Assessing Children's Syntax*, The MIT Press. 211-231. **Grice, H. P. (1975)**. Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts*. Academic Press. 41–58. **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.

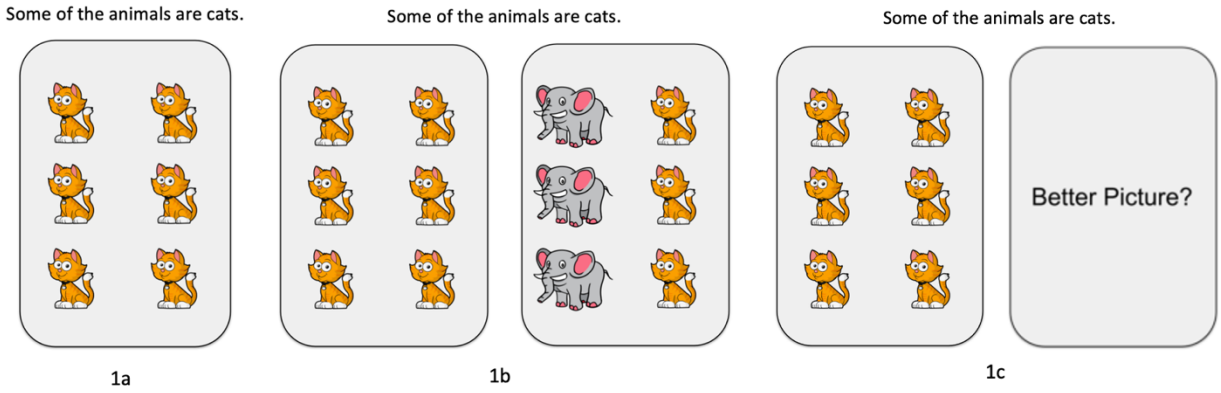


Fig.1 An example of a critical item in TVJT (1a), PST (1b), and HCT (1c). This example concerns the "some-all" scale, while other experimental items may use the "or-and" scale or the "ad hoc" scale. In addition to the images of cats and elephants, images of dogs were also used in the design of the cards. The position of the two cards in PST and HCT was randomized in the experiment.

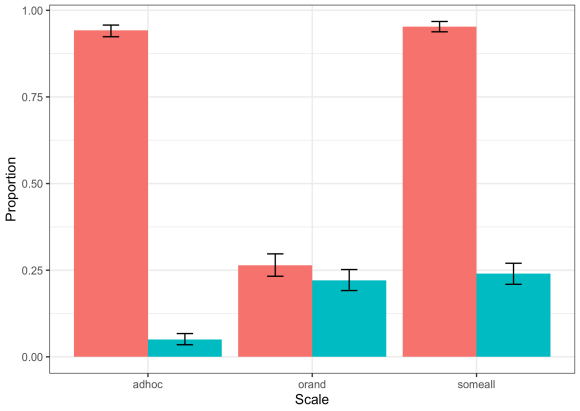


Fig.2: Rate of SI computation estimated by TVJT and PST in Experiment 1. The y-axis shows the percentage of deriving SI for a given scale ("ad hoc" vs "or-and" vs "some-all") in each task (TVJT vs PST), with zero meaning zero percent and one meaning 100 percent. Confidence intervals were computed using bootstrapping methods.

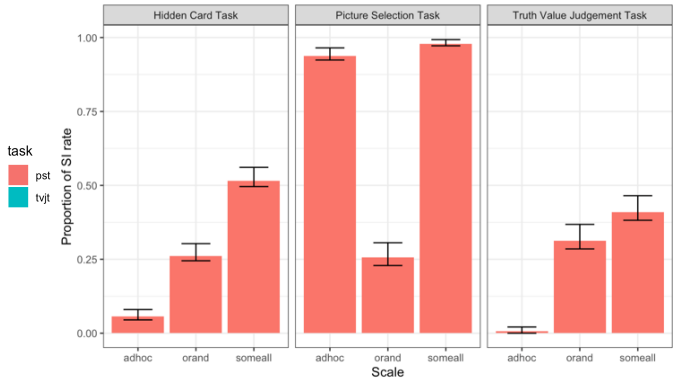


Fig.3: Rate of SI computation estimated by HCT, PST and TVJT in Experiment 2 and 3. The y-axis shows the percentage of deriving SI for a given scale ("ad hoc" vs "or-and" vs "some-all") in each task (HCT vs PST vs TVJT), with zero meaning zero percent and one meaning 100 percent. Confidence intervals were computed using bootstrapping methods.

The importance of speaker knowledge and cooperation in priming scalar implicatures

According to Post-Gricean approaches to implicatures, the speaker's cooperative intention and knowledgeability, as well as the contextual relevance of the implicature, all contribute to whether or not an implicature will be derived in a given context of utterance. Previous studies on implicature priming have investigated the derivation mechanism for scalar implicatures (e.g. Bott & Chemla, 2016) but did not take into account the role of speaker cooperation and speaker knowledge in their experimental design. In two priming experiments, we investigated the effect of the presence of a cooperative and knowledgeable interlocutor on the derivation of both scalar and ad-hoc implicatures.

Experiment 1 was conducted online on 195 English-speaking adults and involved the presence or absence of knowledgeable and cooperative interlocutors as a between-subjects variable in a structural priming task modelled after Bott & Chemla (2016). Participants played a game, in which they were shown two cards and had to pick the winning one based on a description. The game included two types of trials: primes and targets. In target trials, only one of the two cards was visible and the other was covered, and the description of the winning card included either a lexical (<some/all>) or an ad-hoc scalar expression. Examples of target trials for Experiments 1 and 2 are given in Figure 1. Crucially, the description was adequate for the visible card only if the participant did not derive the implicature.

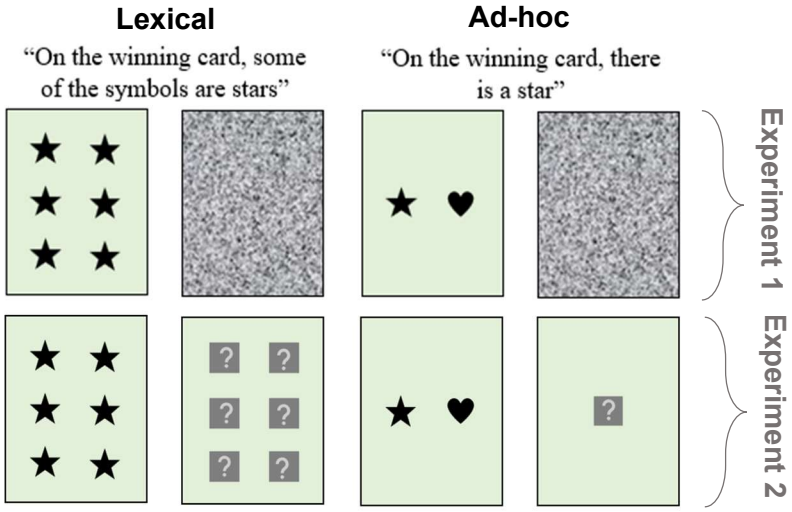
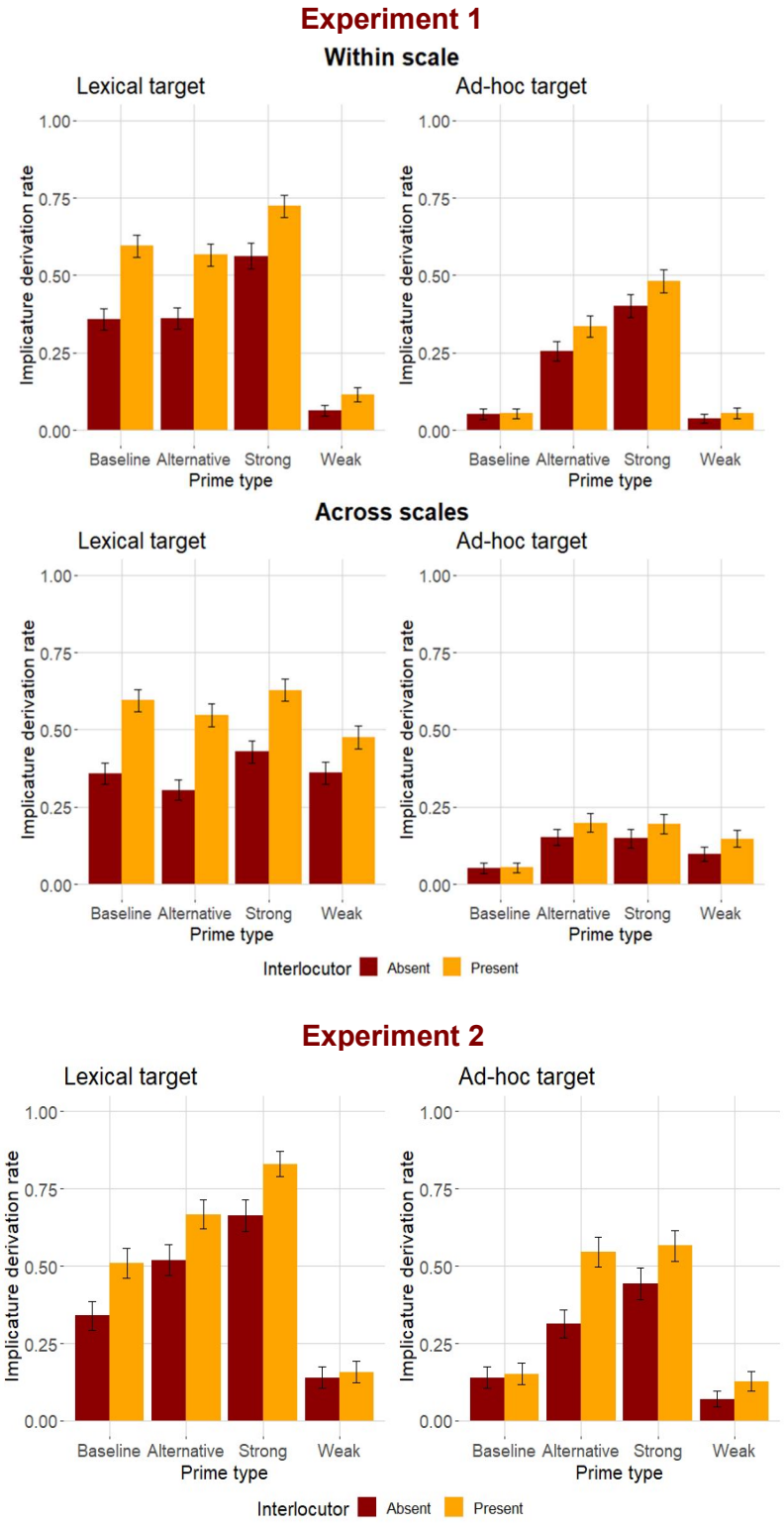


Figure 1: Examples of target items in Experiments 1 and 2.

The choice of the covered card in target trials was taken as a measure of implicature derivation. Each target trial was preceded by two prime trials, which could be of four types: Strong, Weak, Alternative, and Baseline. Strong primes induced the strong reading of the sentence, eliciting an implicature (e.g., some and not all), while weak primes elicited a weak reading (e.g., some and possibly all). Alternative primes provided the more informative alternative to the scalar item used (e.g., all). Finally, Baseline primes aimed to establish how participants understood the target trials in the absence of direct priming; these items were shown to participants separately in the first block of the experiment. Implicature priming was tested both within (e.g. lexical primes and lexical target) and across scales priming (e.g. lexical primes and ad-hoc target). The main modification made to the task compared to previous experiments was the addition of the presence or absence of a knowledgeable and cooperative interlocutor in the instructions. It was predicted that the presence of an interlocutor would increase the rates of implicature derivation overall and allow for across-scale priming.

The data were analysed with Generalised Linear Mixed Models and the results confirmed priming is possible both within and across the two scales, but more importantly that the presence of an interlocutor has a positive effect on implicature derivation and allows for priming effects across different scales. Unexpectedly, we also found that the presence of an interlocutor interacted positively with the lexical scale.

A summary of the main results of both experiments is given in Figure 2.



In Experiment 2, we tried to address a potential confound: in this paradigm, the context is only partially available in target trials. This creates an asymmetry between lexical and ad-hoc scales since alternatives are dependent on the context only in the latter case. A modification of experimental items was implemented to limit potential contextual alternatives by covering only the symbols in target trials instead of the whole card. This second experiment did not include across-scale priming, and 110 English-speaking adults took part in it.

The manipulation worked, as the interaction between interlocutor presence and lexical scale was no longer detected in Experiment 2, while other effects were replicated.

The results of the two experiments are consistent with previous findings. More importantly, however, they highlight the role of communicative context and interlocutors in the process of implicature derivation and provide some evidence for a shared derivation mechanism for lexical and ad hoc scalar implicatures, which depends on perspective-taking and intention-reading. The results also yield important methodological consequences for testing pragmatic phenomena, as they show the importance of providing participants with an adequate conversational context.

Figure 2: Summary of the results of Experiments 1 and 2

Reference: Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. Journal of Memory and Language, 91, 117–140.

Only the (informationally) stronger survive: A probe recognition study with scale-mates and antonyms

Background: Most theoretical accounts assume that scalar implicatures involve alternatives, but it is an open question which kinds of alternatives listeners reason about (see Chemla & Singh, 2014, and Gotzner & Romoli, 2022 for an overview). The standard view following Horn (1972) is that only stronger scale-mates should play a role in the inferential process. For example, when hearing a sentence like *Zack's carpet was dirty,* listeners should activate and negate the scale-mate *filthy*. Recent psycholinguistic studies have used lexical priming paradigms to show that listeners indeed activate alternatives within pragmatic processing (de Carvahlo et al., 2016; Ronai & Xiang, 2023). Interestingly, some studies indicated that non-entailed alternatives such as antonyms (*clean*) also play a role in the inferential process, contrary to the standard view (e.g., Peloquin & Frank, 2016; Lacina et al., 2023). This is consistent with the Alternative Activation Account (Gotzner, 2017), which proposes a two-stage process: First, all semantically associated meanings are activated (e.g., *filthy* and *clean*) and second, grammatical and contextual restrictions select those alternatives that are relevant for implicature computation (*filthy*). This view predicts that only the strong scalars should survive in the representation of the final product of pragmatic processing. Here, we test this prediction in experiments using the probe recognition task, which taps into eventual discourse representations (Gernsbacher & Jescheniak, 1995) and has been used to test relevant alternatives in focus processing (Gotzner et al., 2016). We hypothesised that only stronger scalars (*filthy*) should be included in the final discourse representation while antonyms (*clean*) should no longer be represented since they are not part of the relevant set of alternatives for scalar implicature computation.
Method: Our native English speakers were exposed to sentences in the RSVP mode such as *Zack's carpet was…* The prime words ending the sentence were either related (*dirty* or *clean*) or unrelated (*patterned*) to the same target words (*filthy*). In Exp 1, the related probe word was the weak scalar (*dirty*) and in Exp 2, the antonym (*clean*). The unrelated primes were the same in both experiments and they served as a control that is irrelevant for pragmatic processing and associative priming. Participants were asked to read the sentences and then indicate whether a probe word, which appeared 2000ms after the stimulus, was present in the previous sentence. We used the same sentence frames as Ronai & Xiang (2023) and Lacina et al. (2023).
Results: For both experiments, we ran linear mixed effects models on the log-RT data of correct probe rejections with the fixed effect of relatedness. In Exp 1 (N = 74, Items = 60), target words were rejected slower when they followed weak scalars compared to unrelated words (β = 0.0337, SE = 0.0099, df = 52.11, t = 3.416, p = 0.00124**). This was not the case in Exp 2 (N = 78, Items = 60), where antonymic primes did not significantly differ from unrelated primes: β = 0.0087, SE = 0.0076, df = 75.78, t = 1.144, p = 0.256. A combined analysis of both experiments showed that the interaction of prime type (weak scalar or antonym experiment) and relatedness was significant: β = 0.0245, SE = 0.0109, df = 146.3, t = 2.260, p = 0.0253*.
Discussion: Our data from Exp 1 showed an interference effect in the probe recognition task with strong scalars—weak scalar primes made the recognition of the strong scale-mate slower. This result is reminiscent of the findings regarding unmentioned focus alternatives (e.g., Gotzner et al., 2016) and show that strong scalar terms are being represented by comprehenders in the

mental model of the discourse during comprehension. What this suggests is that at a point in processing where both the sentence and any of its pragmatic inferences have presumably been dealt with, the stronger term is present in the minds of comprehenders, arguably as a part of the finished enriched meaning of the sentence with its associated scalar implicature.

In contrast, Exp 2 showed that this was not the case when antonyms were presented as primes and a cross-experiment analysis revealed an interaction effect. Thus, strong scalars but not antonyms seem to be retained in the final discourse representation. Lacina et al. (2023) reported that in the earlier stages of processing, both weak scalars (*dirty*) and antonyms (*clean*) activated the targets (*filthy*). Taken together with the current results, we find support for the Alternative Activation Account: while all semantic associates might be activated in the process of implicature derivation, only the strong scalars are retained in the eventual representation, being the only relevant alternatives. The initial broad activation of all associates is a result of how the brain organises information in semantic networks across domains (e.g., Onifer & Swinney, 1981) while there has to be additional specialised mechanisms for the computation of scalar implicatures that identify relevant alternatives that are being negated.

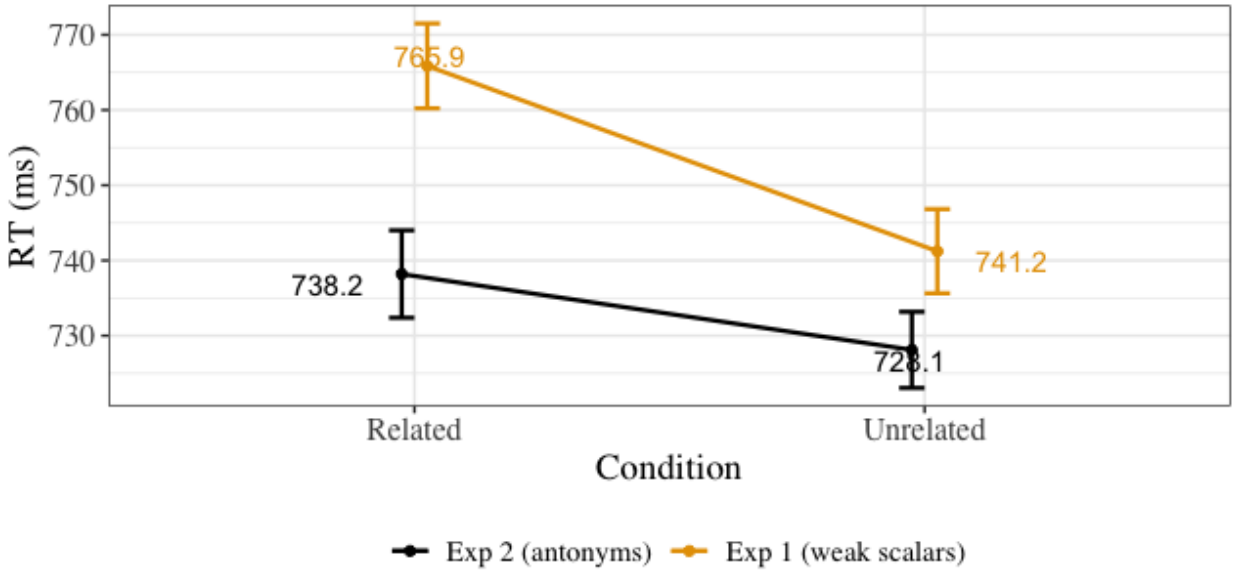


Figure 1: Mean response times of correct rejections by condition in Experiments 1 and 2 with associated standard errors.

Selected References: De Carvalho, A., Reboul, A. C., Van der Henst, J. B., Cheylus, A., & Nazir, T. (2016). Scalar implicatures: The psychological reality of scales. *Frontiers in psychology*, 7, 1500.; Gernsbacher, M. A., & Jescheniak, J. D. (1995). Cataphoric devices in spoken discourse. *Cognitive psychology*, 29(1), 24-58.; Gotzner, N. (2017). *Alternative sets in language processing: How focus alternatives are represented in the mind*. Springer.; Horn, L. (1972). *On the semantic properties of logical operators in English*. University of California, Los Angeles.; Lacina, R., Alexandropoulou, S., Ronai, E., & Gotzner, N. (2023). The Priming of Informationally Weaker Alternatives: Antonyms and Negation. Poster presented at the 10th Experimental Pragmatics conference, September 20 - 22 Paris, France.; Ronai, E., & Xiang, M. (2023). Tracking the activation of scalar alternatives with semantic priming. *Experiments in Linguistic Meaning*, 2, 229-240.

How does a speaker's intent to deceive affect scalar inference and lie judgments?

The question of whether false implicatures are lies has interested theoreticians at the semantics/pragmatics interface for several years,[1,2] with more recent work turning to experimental evidence to help clarify the picture.[3,4] Despite differing results, this experimental work has begun to highlight elements of context that explain variation in such judgments, with a particular eye toward intention to deceive. If context establishes that a speaker has a clear intention to deceive a hearer, participants will reliably rate a false implicature from that speaker as more of a lie than in context without such an intention established.[5]

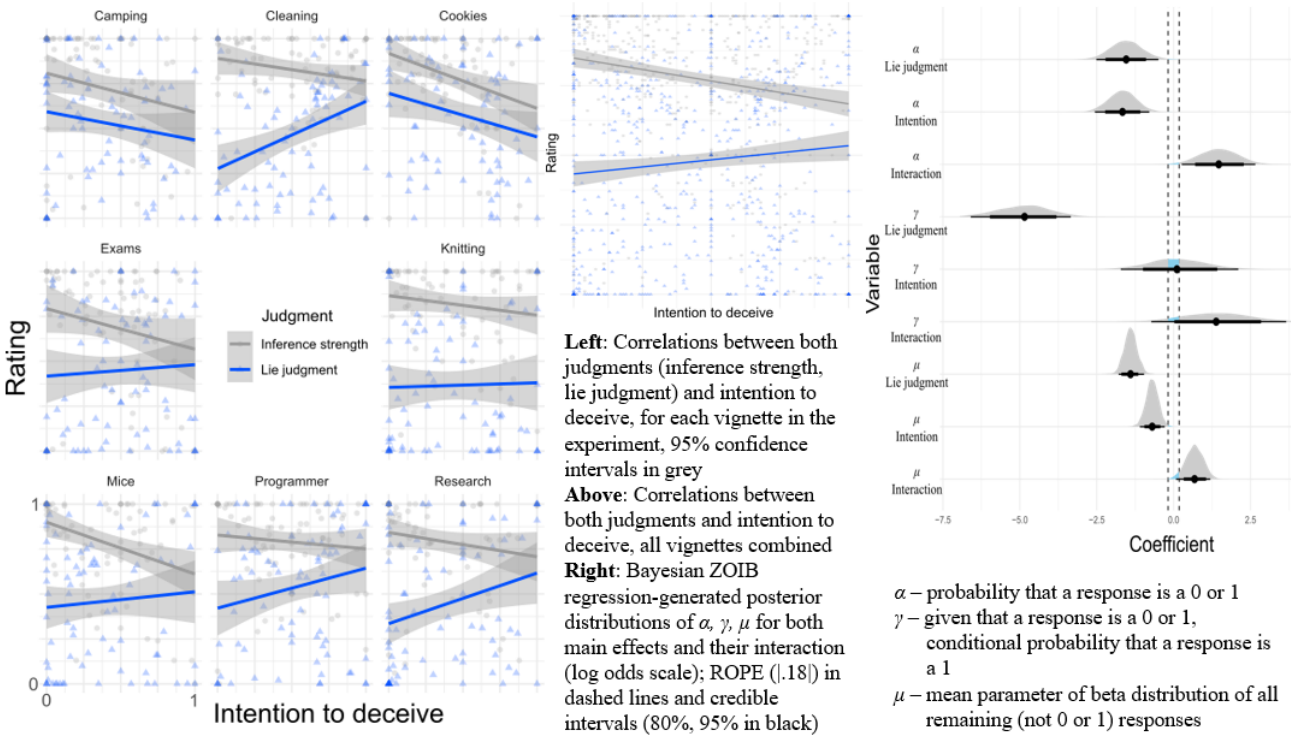
The present research adds to this strand of false implicature research by connecting to recent work investigating the effects of elements of context on the strength of scalar inference.[6] That study found that speaker competence (i.e., whether the speaker knows whether *all* is true when they use *some*) and prior probability (i.e., the a priori likelihood of *all* being true) significantly affect the strength of the *some but not all* inference made by participants. Inference strength is closely tied to lie judgments as well, as recent proposals argue that the level of commitment attributed to the speaker with respect to false implicated content modulates the degree to which that utterance will be considered a lie.[3,7]

The present work aims to investigate whether perceived intention to deceive significantly affects (a) strength of scalar inference drawn and (b) lie judgment of the utterance; in addition, this research will compare the magnitude with which intention to deceive affects (a) and (b). Eight vignettes were crafted, each of which led to a speaker delivering a line licensing a *some but not all* inference. For each vignette, two versions were created: one with a clear motivation for the speaker to try to deceive the hearer, and one without such a motivation explicitly provided; this intention to deceive is counterbalanced within-subjects across the vignettes.

Participants in the main experiment first saw a vignette without the critical utterance and made a sliding-scale judgment about the likelihood of the speaker intending to deceive the hearer in the situation. Following a comprehension question, the second judgment probed either inference strength or lie judgments, varying between subjects. For the former, the critical utterance was added to the vignette, and participants provided a sliding-scale scalar inference judgment. For the latter, the truth of the situation was revealed (i.e., that *all* is, in fact, true) and the critical utterance is added to the vignette, prompting a sliding-scale lie judgment.

Thus far, 240 native English-speaking participants were recruited via Prolific (avg. age = 37.4, sd = 13.7, 117F/120M/3 other); data collection is continuing to 320. Each participant sees 16 items total – 8 *some but not all* items and 8 fillers without scalar implicatures. For participants in the lie judgment condition, the revealed truth of the critical *some* items is balanced between critical cases where *all* is actually true and distractor cases where *some* is actually true or *none* is actually true. Target and filler items are counterbalanced across the two intention conditions, and trial order is randomized for all participants.

A Bayesian mixed effects model regresses sliding scale ratings against fixed effects of speaker's intention to deceive, the judgment being made (scalar inference strength vs. lie judgment), and their interaction, with random intercepts and slopes by item and random intercepts by participant. Zero-one-inflated-beta regression is used due to inflation at 0 and 1 (the sliding scale extremes).



Left: Correlations between both judgments (inference strength, lie judgment) and intention to deceive, for each vignette in the experiment, 95% confidence intervals in grey
Above: Correlations between both judgments and intention to deceive, all vignettes combined
Right: Bayesian ZOIB regression-generated posterior distributions of α, γ, μ for both main effects and their interaction (log odds scale); ROPE (|.18|) in dashed lines and credible intervals (80%, 95% in black)

α – probability that a response is a 0 or 1
γ – given that a response is a 0 or 1, conditional probability that a response is a 1
μ – mean parameter of beta distribution of all remaining (not 0 or 1) responses

Results (thus far) highlight the complexities in assessing the relationship between these judgments. There is a consistent negative effect of intention to deceive on inference strength whereby scalar inference gets weaker as intention to deceive increases. There is an inconsistent positive effect of intention to deceive on lie judgments whereby lie ratings get higher as intention to deceive increases.

These findings appear to complicate the commitment-based account, though they do not necessarily refute it. This analysis adds to the growing body of research investigating effects of context on the strength of scalar inferences. It also begins to quantify the preliminarily-documented finding that intention to deceive affects lie judgments of false implicatures. Lastly, it helps to clarify the relationship between context, commitment, and message interpretation, or at least helps to highlight the complexity in such a relationship.

[1] Meibauer, J. (2014). Lying at the Semantics-Pragmatics Interface. Berlin: De Gruyter Mouton.
[2] Saul, J. M. (2012). Lying, Misleading, and What is Said: An Exploration in Philosophy of Language and in Ethics. Oxford: Oxford University Press.
[3] Reins, L. M., & Wiegmann, A. (2021). Is lying bound to commitment? Empirically investigating deceptive presuppositions, implicatures, and actions. Cognitive Science, 45(2).
[4] Weissman, B., & Terkourafi, M. (2019). Are false implicatures lies? An empirical investigation. Mind & Language, 34(2), 221–246.
[5] Wiegmann, Alex. (2022). Lying with deceptive implicatures? Solving a puzzle about conflicting results. Analysis, 1–11.
[6] Tsvilodub, P., Van Tiel, B., & Franke, M. (2023). The role of relevance, competence, and priors for scalar inferences. Experiments in Linguistic Meaning, 2, 288.
[7] Meibauer, J. (2023). On commitment to untruthful implicatures. Intercultural Pragmatics, 20(1), 75–98.

Quantifying Non-Implicature Sources of Disjunction Exclusivity

A positive disjunction (A or B) in natural language typically receives one of two logical interpretations: inclusive (A or B or both) or exclusive (A or B but not both). Within the Gricean paradigm, the inclusive interpretation is often considered as the primary meaning of disjunction words such as “or” while the exclusive interpretation is attributed to other factors and mechanisms. It is recognized that both scalar implicatures (Grice, 1978; Horn, 1972; Gazdar, 1980) and prior expectations on the exclusivity/compatibility of the disjuncts (Geurts, 2006) can contribute to exclusivity implications, yet no study so far has measured their respective contributions. We present two experimental studies that first pulls apart the sources of exclusivity in examples used in the disjunction literature (Exp. 1), and second tests the role of the syntactic category of disjuncts and their lengths as a potential source of exclusivity (Exp. 2).
Experiment 1: Prior Compatibility vs. Scalar Implicatures: Motivating Experiment 1 is Geurts (2006)’s observation that certain disjuncts cannot co-occur (e.g. The car is in the garage or on the street) and thus must be interpreted exclusively sans any scalar reasoning. Others are merely unlikely to co-occur, e.g. John is singing or screaming. We hypothesized that some amount of exclusivity may stem from this partial incompatibility. First, we collected norming data on the compatibility of the disjuncts in 47 items pulled from published work on exclusive disjunction (Example 1). N = 50 subjects rated how likely the two separate disjuncts were to be true together from 0% to 100%, aiming to assess compatibility. Next, we exposed a different N = 50 participants to the full disjunctions from the literature and asked them to rate how possible it is that both disjuncts were true based on the sentence they just read, aiming to assess exclusivity. Our results show that prior beliefs about disjunct compatibility were not only highly predictive of exclusivity (R2 adjusted = 0.488; Figure 1), but also that disjunctions judged by the literature to be exclusive due to scalar implicatures tended to have less compatible disjuncts (Figure 2). This suggests that not only does prior compatibility contribute to the overall exclusivity of a disjunction, it may also play a confounding role in previous theoretical work that has historically assumed scalar implicatures as the primary source of exclusivity. However, we also show that the use of “or” and a disjunction introduces exclusivity above and beyond the prior expectations of the exclusivity of disjuncts, which is most likely due to scalar implicatures.
Experiment 2: Syntactic Structure and Disjunct Length: Experiment 2 was motivated by the observation that coordinating clauses (e.g. John likes tea or John likes coffee) tends to imply exclusivity more than coordinating NPs (e.g. John likes tea or coffee) (Jasbi, 2018). Because varying the syntactic category of disjuncts necessarily varies the length of said disjuncts, it was crucial to control for disjunct length as well as the syntactic category of the phrases. 32 disjunction frames were created in the style of Example 2 that varied in syntactic category within items and NP length across items, split into 4 latin square groups so each participant only saw each sentence frame once. N= 60 participants rated the exclusivity of these sentences in the same manner as participants did in Experiment 1. Their data was analyzed using mixed effects linear regression with main effects of syntactic category, disjunct length, and their interaction, and random effects of item, participant, and item by participant. Because the analysis was within items, statistical control for prior compatibility was superfluous. We found no significant effects of either length or syntactic category, suggesting that the syntactic category and length of disjuncts do not introduce a strong and robust exclusivity implication independent of prior compatibility and scalar implicatures.
Conclusion: Our results demonstrate that most exclusivity implications are composite implications that feed from minimally 2 sources: prior beliefs about compatibility and scalar implicatures, but that neither the syntactic category of disjuncts nor their length has a consistent effect. These studies also suggest that future work in semantics and pragmatics should be careful not to overestimate the role of scalar implicatures in generating exclusivity implications and pay closer attention to non-implicature sources of exclusivity. Finally, follow-up work should test other potential non-implicature sources of exclusivity, such as prosody (Roelofsen & Pruitt, 2013) and the presence of “*either*”, incorporating them into a comprehensive model that uses

multiple factors in predicting the interpretation of disjunction across contexts. Our results serve as a demonstration of a many-to-one model of pragmatic meaning that complements existing theories of implicature while accounting for non-implicature sources of disjunction exclusivity.

Example 1: Prior Compatibility Stimuli:
(a) “John is singing.” (Norming Disjunct A)
(b) “John is screaming.” (Norming Disjunct B)
(c) “How likely is it that someone is both singing and screaming?” (Compatibility Probe)
(d) “John is singing or screaming.” (Original Disjunction)

Example 2: Syntactic Category and Disjunct Length Stimuli
“John likes tea or John likes coffee” (Coordinated Clauses+Proper Name, 3 word disjuncts)
“John likes tea or he likes coffee” (Coordinated Clauses+Pronoun, 3 word disjuncts)
“John likes tea or likes coffee” (Coordinated VPs, 2 word disjuncts)
“John likes tea or coffee” (Coordinated NPs, 1 word disjuncts)
(Stimuli had NPs ranging from 1-8 words)

References: Gazdar, G. (1980). Pragmatics and logical form. *Journal of Pragmatics*, 4(1), 1-13. Geurts, B. (2006). Exclusive disjunction without implicature. [Ms., University of Nijmegen]. Grice, H. P. (1978). Further notes on logic and conversation. In *Pragmatics* (pp. 113-127). Brill. Horn, L. R. (1972). On the semantic properties of logical operators in English. [Doctoral Dissertation, UCLA]. ProQuest Dissertations and Theses. Jasbi, M. (2018) Learning Disjunction. [Doctoral Dissertation, Stanford University]. Stanford Digital Repository. Pruitt, K., & Roelofsen, F. (2013). The interpretation of prosody in disjunctive questions. *Linguistic inquiry*, 44(4), 632-650.

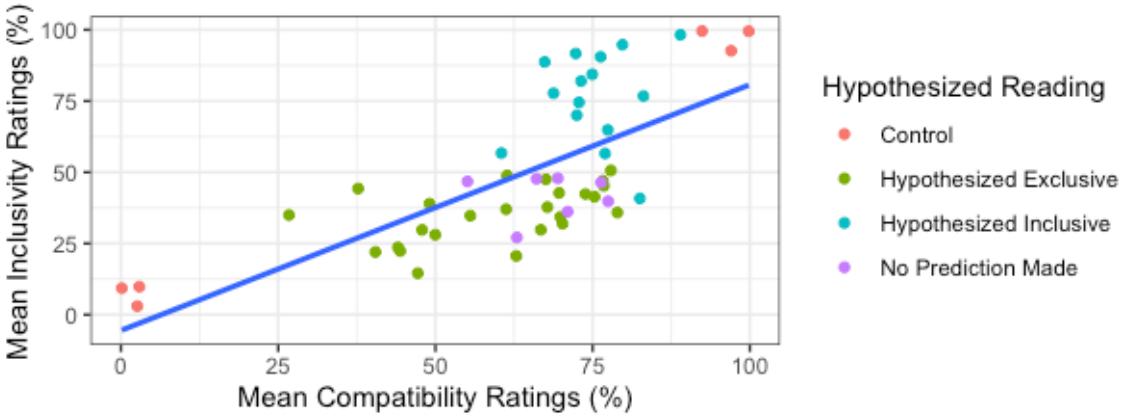


Figure 1: Correlation Between Compatibility and Inclusivity

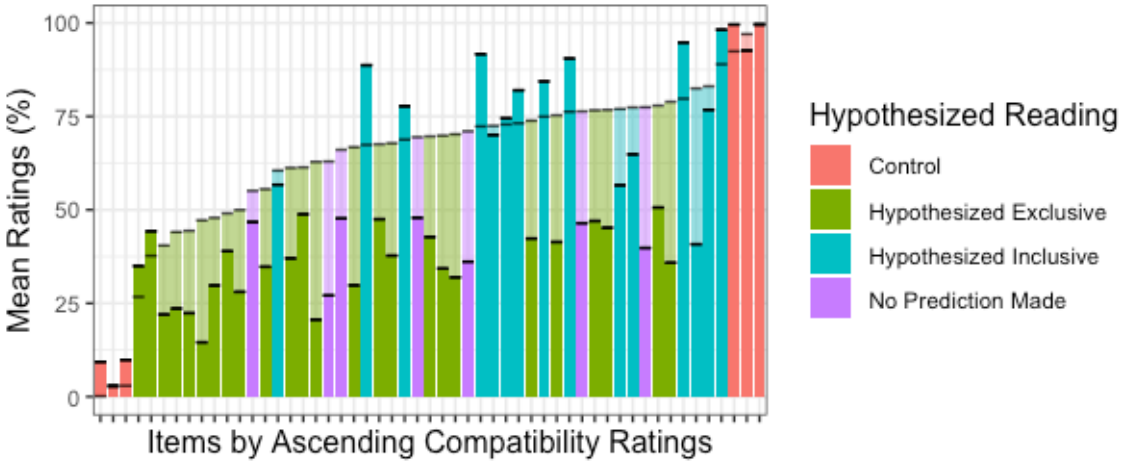


Figure 2: Change in Item Means Between Tasks (prior compatibility is translucent)

Priming acceptability judgments of NPI *any*

Summary We report on a priming experiment whose results indicate that (i) acceptability judgments of the Negative Polarity Item (NPI) *any* can be primed, but (ii) only unacceptable sentences of the same type, i.e., those that contain unlincensed *any*, trigger priming effects. While these findings from a single experiment on their own admittedly have only indirect implications on theories of NPI licensing, we argue that our paradigm has far-reaching methodological importance for theoretical linguistics, offering a novel way of directly testing theoretical predictions. We will illustrate this with the so-called bagel problem for certain Russian NPIs and the source of island effects.

NPI *any* in non-monotonic environments Weak NPIs like *any* are canonically licensed in Downward Entailing (DE) environments (Fauconnier 1975, 1979, Ladusaw 1979, 1980), but it is also known that they are licensed in certain non-monotonic (NM) environments (Linebarger 1980, 1987). For NM environments with DE at-issue meaning and non-DE presupposition, von Fintel (1999) proposes that weak NPIs are insensitive to presuppositions. However, there are instances of weak NPIs in certain NM environments that this account does not explain. Among those, we focus on NPI *any* under *exactly* n (Heim 1984, Rothschild 2006, Crnič 2011).

(1) Exactly two restaurants served any vegan dishes.

Previous experimental research found that the acceptability judgments of such sentences are not as crisp as those of NPI *any* in plainly DE environments (Alexandropoulou, Bylinina & Nouwen 2020). Using the experimental method of *priming*, our experiment investigates how these acceptability judgements are affected by preceding sentences. To the best of our knowledge such priming effects on acceptability judgments have not been systematically investigated before.

Priming Priming has been extensively used to investigate mental representations in various domains of psycholinguistics, most relevant of which in the context of our research is the so-called *structural priming* (Bock 1986; see Pickering and Ferreira 2008 for an overview). To illustrate, participants in Bock's (1986) study repeated prime sentences, appearing either in active or in passive form, and then described a picture. When doing so, they were more inclined to utter a sentence in passive when they had repeated a passive sentence (a 'prime'), than when they had repeated an active prime. This is taken as evidence for the psychological reality of some mental representation that encodes the voice information, but is abstract enough to not include the specific lexical items of the primes. This experimental technique has more recently been used to argue for mental representations of quantifier scope (Raffray & Pickering 2010, Chemla & Bott 2015, a.o.) and scalar implicatures (Bott & Chemla 2016, Meyer & Feiman 2021, a.o.). In the present study, we employed the structural priming paradigm to address our investigation into how the acceptability of *any* in NM environments is affected by the (un)acceptability of different types of primes.

Material, method, and procedure We collected acceptability judgments of 16 sentences that contain *exactly* n as subject and NPI *any* as object, as in (1). As weak NPIs are considered to be judged as more acceptable for smaller n's (Heim 1984, Rothschild 2006, Crnič 2011), we used numerals between *two* and *eight* (each in two target items). Each target item was preceded by two primes (as in most previous structural priming experiments). There were six types of primes altogether. They contained *no* or *some* as the subject quantifier and one of the following as the object quantifier: (a) NPI *any*, (b) a bare plural, or (c) *many* + singular NP. Regardless of the subject quantifier, (b) is expected to be grammatical, and (c) is expected to be ungrammatical, while (a) should be sensitive to the subject quantifier. Therefore, there were six types of primes, as exemplified in (2) and (3). The experiment also contained 72 filler items with varying acceptability.

(2) a. No artists sold any paintings.
b. No artists sold paintings.
c. No artists sold many painting.

(3) a. Some artists sold any paintings.
b. Some artists sold paintings.
c. Some artists sold many painting.

90 participants were recruited on Prolific. They were randomly assigned to one of the six priming conditions. Each of them provided acceptability ratings of 120 sentences (16 target items, each preceded by 2 primes, plus 72 filler items) on a 7-point Likert scale (labelled 'Completely ungrammatical' on the left and 'Completely grammatical' on the right), after reading instructions that were

modelled after those used by Sprouse, Schütze & Almeida 2013 and completing one practice item. Two participants were excluded from the analysis for low accuracy on filler items ($< 75\%$).

Results The ratings of primes and target items are summarized in Figure 1. The acceptability judgments of primes (left column, Figure 1) are overall as expected. We fitted an ordinal mixed effect regression model to the target data using the `ordinal` package (Christensen 2022) for R. Two fixed effect variables, Subject and Object, were each treatment-coded with *Some* and *Any* as reference levels. The model also had by-item variance on the intercept as the sole random effect (including any other random effect resulted in estimation error). The model reveals that target items following Some+Any primes were judged as more acceptable than those following No+Any primes ($\beta = -1.14, p < 0.001$). We also observe that target items following Some+Any primes were judged as more acceptable than those following the other two kinds of primes containing *some* as subject (BarePlural: $\beta = -0.83, p < 0.001$; Many: $\beta = -1.04, p < 0.001$). Moreover, the significant positive interaction effects (BarePlural: $\beta = 0.99, p < 0.001$; Many: $\beta = 0.88, p < 0.001$), which counteract the negative effect of Subject, suggest that there is not much difference among target items following the three types of primes containing *no* as subject.

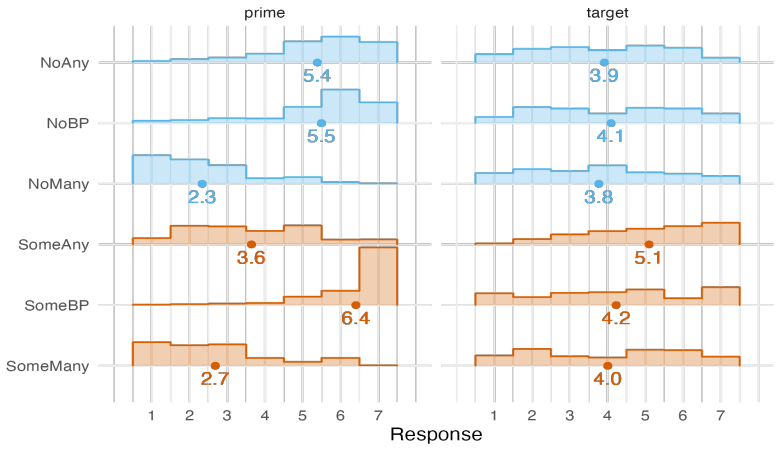


Figure 1: Ratings of primes and target items by condition. The numbers are mean ratings and the histograms represent distributions of by-subject mean ratings.

Discussion The experimental results indicate that acceptability judgments of NPI *any* under *exactly* n can be primed, but only by unacceptable primes containing unlicensed NPI *any* (i.e. Some+Any). It is especially notable that the kind of unacceptability triggered by the number mismatch between *many* and a singular NP exhibited no comparable priming effects. This selective nature of NPI priming gives credence to the existence of a mental representation dedicated to NPI licensing. We illustrate here two potential ways of making use of this finding to directly investigate theoretical issues in future research. The first one is the so-called 'bagel problem' for Russian NPIs. Russian has two series of NPIs, wh+*libo* and wh+*nibud'*, which are licensed in all environments where NPI *any* is licensed, except under negation (Haspelmath 1997, Pereltsvaig 2004). One way to understand this pattern is by assuming that these Russian NPIs are weak NPIs on a par with NPI *any*, but have further licensing conditions. In that case, we expect unlicensed instances of uncontroversially weak NPIs (in Russian or English) to trigger priming effects on wh+*libo*/wh+*nibud'*. The second theoretical issue we discuss here is how island effects are to be explained. It has long been suggested that at least some island effects—especially the so-called weak islands (see, e.g., Szabolcsi 2006)—are to be explained non-syntactically (see Newmeyer 2016 for an overview). Testing what has priming effects on the acceptability of which islands may provide direct evidence for some of these theoretical explanations.

Lastly, we also note that unlike unlicensed NPI *any* (i.e. Some+Any), licensed NPI *any* (i.e. No+Any) had no noticeable priming effects. We claim that this is part of a general property of priming that only 'unexpected events'—in our case unlicensed *any*—trigger priming effects. This is explained by the hypothesis that the mechanism behind priming is an adaptation mechanism (Fine, Jaeger, Farmer & Ting 2013, Jaeger & Snider 2013, Waldon & Degen 2020, Marty, Romoli, Sudo & Breheny to appear). Applying this hypothesis to our case, we claim that the adaptation mechanism lowered the standard for the overall acceptability/grammaticality of NPI *any*, upon exposure to unlicensed instances (cf. 'syntactic satiation effect'; Snyder 2000).

Experimental findings for a cross-modal account of dynamic binding in gesture-speech interaction

We report experimental results of two experiments on pronoun and presupposition binding across modalities. We show that (1) ordinary pronouns (in the spoken/written domain) can be dynamically bound to gesturally introduced discourse referents and (2) that presuppositions induced by presupposition triggers in the spoken/written domain (as e.g. *again* or *too*) can be bound and satisfied by propositions that have been introduced in the gestural domain.

Background. Ebert, Ebert & Hörnig (2020) (based on Ebert & Ebert 2014) suggest a formal framework for gesture semantics where certain iconic and pointing co-speech gestures introduce discourse referents that can serve as antecedents in anaphoric reference. Crucially, this necessitates a unidimensional dynamic system that allows for binding effects across dimensions and, in this case, modalities. Based on the dynamic system of Anderbois et al. (2015) that can handle binding effects across dimensions (with appositives introducing non-at-issue material), Ebert, Ebert & Hörnig (2020) suggest that gestures behave and can be handled on a par with appositives since both contribute propositional non-at-issue information by default. Furthermore and crucially, pointing gestures and iconic gestures introduce discourse referents for rigid designators as their core 'lexical' meaning, i.e. when a pointing gesture is performed this triggers the introduction of a discourse referent that is identified with the rigid concept of the gesture referent. This discourse referent (DR) can then be anaphorically picked up by a pronoun in later discourse. Importantly, in this dynamic semantic framework it is predicted that gesturally introduced DRs allow for anaphoric binding across dimensions, i.e. gesturally introduced DRs can be referents of speech pronouns.

While the introduction of DRs by gesture has been claimed and implemented in the formal system of Ebert, Ebert & Hörnig (2020), this has not been experimentally demonstrated. Here we show that dynamic binding across dimensions can be made with respect to both pronouns and presupposition triggers. It can be shown that gesture can introduce discourse referents which can be picked up by a speech pronoun later-on (as illustrated in (1)). Furthermore, gestures can introduce propositional content that can serve as presupposition binders for presupposition triggers in speech (see ex. (2)).

In the constructed example (1a), the pointing co-speech gesture in the form of extending an index-finger towards a piece of cake as opposed to other baked goods is assumed to introduce a DR for the gesture concept for the referent of said piece of cake and allows it to be bound to the pronoun "it" in the hypothetical follow-up (1b). If (1a) had included a hand-over-stomach gesture to indicate having eaten (1c) and crucially not introducing a DR, then presumably "it" in (1b) cannot be bound. In our experiment, we add as a control (1d) as a possible follow-up. While it seems unlikely that (1a) would be followed by (1d) where a confirmatory response is given that ignores the pointing gesture, (1d) could presumably follow (1c) where no specific referent is indicated. Similarly with presupposition triggers like *again*, the jogging gesture in (2a) - adding the propositional content that Paul was jogging (when the speaker met him) - is assumed to be an additional propositional information given in the visual modality via gesture that can serve to satisfy the presupposition that is triggered by *again* in (2b), namely that Paul went jogging before. In the absence of such a gesture the presupposition triggered by *again* would not be satisfied, at least under the assumption that people don't commonly meet while jogging and hence such a proposition cannot be accommodated. Conversely, a follow up like (2d) is presumably odd following a jogging gesture (2a) under the assumption that people do not jog in cafes, but following (2c) ought to be fine assuming people often meet in cafés.

(1) a. Have you eaten[pointing to a piece of cake]?
b. It was too sweet for me.
c. Have you eaten[placing hand over stomach]?
d. Yeah, a few too many cookies.

(2) a. Yesterday I met Paul[jogging gesture]
b. He went jogging again today.
c. Yesterday I met Paul[pointing backwards]
d. Was it in the café again?

Experiments. Two experiments were designed in German to test the contrasts demonstrated in (1) and (2). Given the similarity in contrasts, albeit distinct form of gesture and anaphora, the designs were complementary and allowed each to be used as filler for the other. Both experiments had two factors each with two levels, yielding two treatment factor levels (felicitous

or infelicitous). Experiment 1 had the levels GESTURE (pointing (1a) or iconic (1c)) and to-be-bound-PRONOUN (present (1b) or absent (1d)), and Experiment 2 the levels: GESTURE (pointing (2c) or iconic (2a)) and to-be-bound-PRESUPPOSITION (present (2b) or absent (2d)). Each participant participated in each of the within subject conditions in (3)-(4). The minimal pairs resembling (1) and (2) were distributed across four groups of participants. We recruited 60 native German speaking participants via Prolific, following the 2x2 repeated-measures design in Brysbaert (2019). In a variation of the covered-box task (cf. Fanslow et al. 2019), the sentence pairs were presented with the context, e.g. (1a) presented in video form, and the follow-up, e.g. (1b), being presented in written form as one choice in a pair of alternatives, the other being 'covered' (lit. "[geschwärzt]" ('redacted')). Participants were instructed that one of the alternatives was a reasonable follow-up to the context and the other wasn't, and they should select whichever they believe to be more reasonable.

(3) a. GESTURE—pointing + PRONOUN—present (felicitous, (1a)+(1b))
b. GESTURE—pointing + PRONOUN—absent (infelicitous (1a)+(1d))
c. GESTURE—iconic + PRONOUN—present (infelicitous (1c)+(1b))
d. GESTURE—iconic + PRONOUN—absent (felicitous (1c)+(1d))

(4) a. GESTURE—iconic + PRESUPPOSITION—present (felicitous, (1a)+(1b))
b. GESTURE—iconic + PRESUPPOSITION—absent (infelicitous (1a)+(1d))
c. GESTURE—pointing + PRESUPPOSITION—present (infelicitous (1c)+(1b))
d. GESTURE—pointing + PRESUPPOSITION—absent (felicitous (1c)+(1d))

Results. Starting with the pronoun experiment, for items with pointing gestures, follow-ups with pronouns meant to be bound to the gesture DR were largely accepted (3a, n=115), and, surprisingly, those without such a pronoun were accepted nearly as much (3b, n=105). As expected, with iconic gestures, pronouns that could not be bound to a DR were not accepted (3c, n=63) unlike those with other continuations (3d, n=133). In the presupposition experiment, items with iconic gestures plus follow-ups with presuppositions meant to be bound to iconic gestures were largely accepted (4a, n=129), and those with such presuppositions absent less so (4b, n=86). As expected, the same items albeit with pointing gestures plus follow-ups with to-be-bound-presuppositions were generally not accepted (4c, n=72) and those without were accepted (4d, n=138). Responses for each experiment were analyzed with a 2x2 ANOVA with the within-subject factors. A significant interaction of GESTURE+PRONOUN was found ($F(1,716)$ 39.54, $p < 0.001$, $\eta 2$ 0.055) as well as GESTURE+PRESUPPOSITION ($F(1,716)$ 75.32, $p < 0.001$, $\eta 2$ 0.105)—i.e. the null hypothesis of no interaction between gesture and anaphora is unlikely.

Discussion. There are two key contrasts targeted in this study: (i) when pronouns have gesture DR vs. when they have no obvious referent (cf. (1a+1b) vs. (1c+1b) and (ii) when presupposition triggers can be bound to a gesture-introduced proposition vs. when they have no obvious referent (cf. (2a+2b) vs. (2c+2b). In both contrasts the former has been assumed to be felicitous, and the latter not, and the interaction between gesture and binding found in the experiments support these assumptions. In other words, we have provided experimental findings that substantiate the introspectively supported claims of the need for a cross-modal account of dynamic binding in gesture-speech interaction.

References. Brysbaert, M. 2019 How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, 2(1): 16, pp. 1–38. DOI: https://doi.org/10.5334/joc.72 ● Anderbois, Scott & Brasoveanu, Adrian & Henderson, Robert. 2013. At-issue proposals and appositive impositions in discourse. *Journal of Semantics* 32(1). 93–138. ● Ebert, Cornelia & Ebert, Christian. 2014. Gestures, demonstratives, and the attributive/referential distinction. *Talk at Semantics and Philosophy in Europe 7, Berlin: ZAS*. ● Ebert, Christian & Ebert, Cornelia & Hörnig, Robin. 2020. Demonstratives as dimension shifters. In Franke, Michael & Kompa, Nikola & Liu, Mingya & Mueller, Jutta L. & Schwab, Juliane (eds.), *Proceedings of Sinn und Bedeutung* 24. 161–178. ● Fanselow, G. & Zimmermann, M. & Philipp, M., (2022) “Accessing the availability of inverse scope in German in the covered box paradigm”, *Glossa: a journal of general linguistics* 7(1).

A type of sarcasm that current theories fail to explain – evidence from *sarchasm*

Overview. In this work in progress, I examine multiple instances from my experiment data that fall under the category of *sarchasm*, an utterance that is intended to be sarcastic but missed by the listener or overhearer (Fox Tree et al., 2020). Missed instances of sarcasm provide a unique window for thinking about the use and interpretation of sarcasm. I show that there is a particular subtype of sarcasm found in real data, which current theories of sarcasm (or verbal irony[1]) fail to explain. I propose a new framework that can address this phenomenon.

(1) *Context*: Your friend was sure it would not rain today but you realize it is raining.
Response: What a great day.

Theories of sarcasm. In the **Gricean theory**, sarcasm is identified when there is a blatant violation of maxim of quality. The response in (1) is therefore is sarcastic because the speaker is being untruthful. In **Echoic theories**, a speaker "echoes" (as opposed to "uses") an utterance to convey a negative attitude. An echoic utterance alludes to the thoughts or utterances of others, which reminds the listener of norms or failed expectations and allows for the interpretation of sarcasm. (1) is sarcastic since the listener would know that the speaker is merely 'echoing' the previous thought that it was not going to be rainy, in order to express her negative attitude towards it. In the **Pretense theory**, a speaker (S) 'pretends' to be an alternative speaker (S') speaking to an alternative listener (H'). S poses a negative attitude towards the utterance of S', and H' is ignorant and takes the utterance literally, while H understands it all. In (1), the speaker thinks that the weather is bad but pretends to be a person who thinks that the weather is good, and has a pretend-listener who would believe it and intends for the actual listener to understand all of it. In the **Implicit display theory**, sarcasm occurs if the speaker has an unmet expectation and conveys a negative attitude toward the failed expectation through the utterance. The speaker in (1) had an expectation that her friend's belief would be true but expresses her negative attitude when the belief turned out to be wrong. It is not the focus of this work to discuss the limitations of individual theories. Instead, I show data that suggest that there is another type of sarcasm that current theory as a whole cannot explain.

Data. I use data collected from four (two production and two comprehension) online experiments. In each production experiment, participants (N=60 and N=128) were provided with contexts (N=32 and N=40), responded freely, and rated how sarcastic their responses were from *1: not at all* to *6: completely*. In each comprehension experiment, new participants (N=360 and N=512) rated how sarcastic they found the same responses as external evaluators. Neither speakers nor evaluators were given sarcasm definitions in order to obtain natural data. I selected the instances to which the speakers gave the highest sarcasm rating (6), which I consider as having sarcastic intent. Of 584 such instances, I selected the ones that external observers gave lower than 4 on average (*sarchasm*). I have identified 251 such instances and show examples below.

Limitations of previous theories. In (2), the speaker points out how blind Steve is to his own flaw by bluntly pointing it out to him.

(2) *Context:* Steve has a brother Bill. Bill often feels annoyed by his friend. The reasons that Bill finds his friend annoying are the same as the reasons why you find Steve annoying (for example, both Steve and Bill's friend always ask for money and never pay it back). Steve says, "why is my brother even friends with that guy? I don't get it."
Response: Well you should know, shouldn't you?

[1] I treat *sarcasm* and *verbal irony* synonymously following recent work. The default terminology is *sarcasm*. See Fox Tree et al. (2020) and references therein.

The maxim of quality has not been violated (Gricean theory), nor does the speaker provide any echoic utterance (Echoic theory). If the speaker was engaged in a pretense (Pretense theory), the real listener (H) would have to figure out that the utterance is sarcastic, but given how true and direct the response is, the listener would be faced with a garden-path situation at best. We do spot a failed expectation (Implicit display theory), which is that the speaker expects Steve to be aware of his own flaws at the presence of a similar example. But it is not obvious whether a direct remark would embed a negative attitude at the failure of expectation, which is required for an utterance to be sarcastic.

(3) *Context:* Steve gives you a watering can on your birthday while smiling at you with a strange expression. But you don't even have a single plant.
Response: Umm?? What's this for?

(3) provides a similar type of sarcasm: no violation of the maxim of quality, no echoing, no pretense. The failed expectation is also not clear in this case because even if the listener knows that the speaker does not expect a watering can, it is still possible that the speaker is just being unassuming and asking a genuine question. But the speaker still meant for the response to be sarcastic even though it is unlikely to provide the listener the cues necessary to interpret sarcasm, violating the cooperative principle (Grice, 1975). So how do we explain that such utterances are sarcastic?

Proposal of a new framework. I argue that sarcasm has a variant in which the speaker makes a reasonable remark in a direct manner but actually suppresses her desires to be more emotive, which often leads to *sarchasm*. The reason for muting emotion could be, among others, to save face (Jorgensen, 1996), avoid being rude (dews et al, 1995), or keep the amicable relationship to the listener (gibbs, 2000). The intentional suppression of attitude is deemed sarcastic by the speaker because she knows the underlying emotion behind the utterance, but the listener often misses it unless obvious or external cues are available. This type of sarcasm could be considered as 'reverse sarcasm', in which the speaker wishes to convey an attitude but (ironically) does so by being direct instead of choosing the literal/straightforward (emotionally strong) reaction.

(4) *Context:* You are having a small party at your house. Steve, a little tipsy, starts mixing ketchup, mustard, potato chips, and orange juice and says "hey, look, I made something delicious!"
Response: As long as you eat it buddy, you do you, and don't make a mess!

Then we can interpret the response in (4) as sarcastic. The speaker wishes to point out the silliness of Steve's behavior and does it by making reasonable requests, therefore muting her emotional reaction to him. If Steve also understands the silliness of his own behavior, he might get the sarcasm in the speaker's remark. Otherwise, it will likely become an instance of *sarchasm*.

Implications. The new proposal aligns with prior work that discusses the communicative functions of sarcasm (muting of criticism & face-saving). Sarcasm is used to subdue the criticism embedded in a message (Dews et al., 1995) or to save face by appearing less rude and fairer (Jorgensen, 1996). A new finding that emerged from the data I showed is that the muting of the negative message can go as far as turning an utterance into a direct remark that is reasonable given the context, and thus create a garden-path-like utterance for the listener. But as long as there is intentionally suppressed emotion behind the utterance, it will still count as intended sarcasm, but it will be missed by some listeners. The proposal I made in this work suggests that theory of sarcasm may need to separate intended and perceived sarcasm to thoroughly grasp the complexity of the phenomenon.

The lying/misleading distinction from the viewpoint of truth evaluators

Background. The two dominant definitions of Lying face a challenge when attempting to distinguish between lying and misleading claims. According to the traditional view, lying involves false *explicit* content, whereas misleading claims involve false *implicated* content (1,7, but see 2). Recent empirical studies, however, indicate that speakers can be perceived as lying even when the believed-false content is implicated (e.g., 3). According to a more recent view, speakers are lying if they are perceived as committed to the false communicated content (5, 8). While this view effectively accounts for the empirical evidence, it seems to obscure the boundary between lies and misleading claims. That is to say, in a commitment-based framework for lying and given that commitments are measured on a continuum, it is unclear what constitutes the boundary between lies and misleading claims (as even a merely misleading claim involves some degree of commitment).

Objectives. This study aims to draw a clearer distinction between lies and misleading claims. To achieve this, I address a methodological issue that pertains to the inconsistency within the empirical literature regarding the means by which the difference between lying and misleading claims is gauged. Furthermore, I address the data through the oft-overlooked role of the truth evaluators. By doing so, I aim to highlight a conceptual difference between lying and misleading that pertains to the mindset of those tasked with their evaluation.

The current study. This study examines participants' disposition when asked to truth evaluate a literally true statement that indirectly conveys a believed-false claim. In this study, indirectness was manipulated using explicit, rather than implicated, expressions through the epistemic modals, *possible* and *possibly*, which express uncertainty. Participants were introduced to a context story about a party. Following the story, targets appeared in one of four conditions, see Table 1.

Table 1 – Experimental procedure

Both Aron and Dan like Jill, and they both know this about each other. One day, Aron invited Jill to a party, and she agreed. They danced all night and had a lot of fun. The day after the party, Aron met Dan and told him about the party. Dan immediately asked him if Jill came to the party too.	**Context story**, similar for all participants
Aron said: "It is possible that Jill came to the party."	**Target – Condition 1**: literally true *p* + possible
Aron said: "Jill possibly came to the party."	**Target – Condition 2**: literally true *p* + possibly
Aron said: "Jill did not come to the party."	**Target – Condition 3**: a literally false *p*
Aron said: "Jill came to the party."	**Target – Condition 4**: a literally true *p*
Did Aron lie?	**Task**, similar for all participants
|----------| No, he did not Yes, he did	**Measurement tool**, similar for all participants

Four screening/comprehension questions followed the task. Failure in any one of these questions led to removal from the analysis.

Method. A web-based paradigm with a between-subjects design was employed in Prolific (4). 320 participants (age range: 19–45, M = 31.85, SD = 7.15, 160 females) were recruited until a sample size of ~80 participants per condition, after exclusion, was met. All participants were native US English speakers and were paid regardless of their success in comprehension questions.

Results. A summary of the means' lie ratings per condition shows that the literally false claim received a high lie rating (Cond.3; *M* = 97.3, *SD* = 11.4) and that the literally truthful claim received a low lie rating (Cond.4; *M* = 6, *SD* = 12.2)—as expected. The hedged literally true claims received an intermediate lie rating when accompanied by the objective epistemic modal *possible* (Cond.1; *M* = 50.9, *SD* = 32.6) and by the subjective epistemic modal *possibly* (Cond.2; *M* = 56.6, *SD* = 12.2).

Data was analyzed using a Bayesian Zero-One-Inflated-Beta (ZOIB). The *emeans* package was used for later pairwise comparisons (6). Because this analysis uses a Bayesian framework, it is important to note that there are no clear thresholds to determine significance. Traditionally, if the coefficient intervals do not include 0, it can be deduced with adequate confidence that a significant effect was observed. The model revealed a significant effect of condition for the question, "Did [the protagonist] lie?" It specifically showed that the Highest Posterior Density (HPD) interval of Condition 1 with Condition 3 and of Condition 1 with Condition 4 did not include 0, indicating that an objective epistemic modal with a literally true claim is considered neither a full-fledged lie nor a truthful claim—and similarly for condition 2. The HPD interval of the comparison between Condition 1 and Condition 2 included 0, indicating that the truth evaluations in the two conditions were not significantly different. Lastly, the HPD interval of Condition 3 with Condition 4 did not include 0, indicating that the evaluation of full-fledged lies differs significantly from that of truthful claims.

***Discussion*.** These findings indicate that hedging a literally true claim using epistemic modals is a misleading act. It, thus, also indicates that misleading is not restricted to implicated content. These findings, however, here and in other studies, do not directly explain the lying/misleading distinction. To do this, it is essential to adopt the truth evaluators' perspective (rather than the content's explicitness/speaker's commitment).

A closer look at the truth evaluators' behavioral patterns suggests that two distinct mindsets underlie the evaluation of different forms of deception. In misleading claims, participants are conflicted, probably by the presence of two opposing truth values. They resolve this conflict by leaning towards one of the truth-values (as evident through the bimodal distribution and its wide range). In full-fledged lies and truthful claims, truth evaluators experience no such conflict (as evident in the skewed distribution with its narrow range).

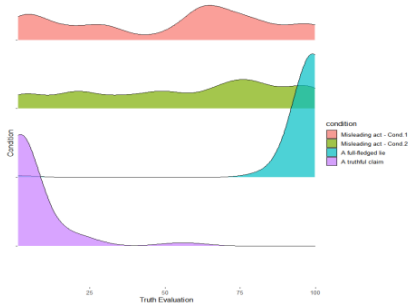


Figure 1 – The patterns of deception

***Future Directions*.** To the extent these patterns generalize, they provide insights into the mindsets of truth evaluators when evaluating different forms of deception. An ongoing experiment explores this using other stories and other modes of deception (e.g., politeness).

References: **(1)** Adler, J. E. (1997). Lying, Deceiving, or Falsely Implicating. The Journal of Philosophy, 94(9), 435–452.; **(2)** Meibauer, J. (2005). Lying and falsely implicating. Journal of Pragmatics, 37(9), 1373–1399.; **(3**) Orr, S., Ariel, M., & Peleg, O. (2017). The case of literally true propositions with false implicatures. In I. Chiluwa (Ed.), Deception and deceptive communication: Motivations, recognition techniques and behavioral control (pp. 67–107). Nova Science Publishers, Inc.; **(4)** Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance, 17, 22–27; **(5)** Reins, L. M., & Wiegmann, A. (2021). Is Lying Bound to Commitment? Empirically Investigating Deceptive Presuppositions, Implicatures, and Actions. Cognitive Science, 45(2) e12936.; **(6)** Russell V., L. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means (R package version 1.8.3).; **(7)** Saul, J. M. (2012). Lying, Misleading, and What is Said: An Exploration in Philosophy of Language and in Ethics. Oxford University Press.; **(8)** Viebahn, E. (2021). The Lying-Misleading Distinction: A Commitment-Based Approach. The Journal of Philosophy, 118(6), 289–319.

Abductive inferences in causal discourse: Evidence from eyetracking during reading

When we interpret causal statements in discourse, we not only integrate causes and effects incrementally, but also immediately take relevant world knowledge into consideration in doing so (Köhne-Fuetterer et al. 2021, Kuperberg et al. 2011, Xiang/Kuperberg 2015, Xu et al. 2017). Accordingly, Kuperberg et al. (2011) showed that two-sentence discourses violating domain knowledge immediately give rise to an N400 effect even in the absence of explicit discourse marking. The present study contributes to this line of research by providing first online evidence that even more fine-grained subtypes of inferential processes occur in online processing. Consider (1):

(1) *Weil [Alex sich an die Aufbauanleitung hielt]$_{cause2}$, [ging die Spülmaschine kaputt]$_{effect}$.*
Because [Alex followed the assembly instructions]$_{cause2}$, [the dishwasher broke down]$_{effect}$.

In isolation, (1) seems *anomalous*. World knowledge predicts an effect to the contrary (cause2 ⇒ ¬effect; cf. Aliseda 2006): Without other evidence, following the instructions prevents a machine from breaking down. (1) should therefore either be rejected, or taken to constitute a partial explanation, leading to the introduction of an additional cause via abductive inferencing (Aliseda 2006).

The anomaly may disappear once (1) is embedded in a larger context. One can think of a number of situations in which (1) could make sense. Consider, for instance, the complex cause in (2), in which a *cause1* has been added to *cause2* in (1):

(2) *Weil [die Aufbauanleitung einen Fehler enthielt]$_{cause1}$ und [Alex sich an die Anleitung hielt]$_{cause2}$, [ging die Spülmaschine kaputt]$_{effect}$.*
'Because [the assembly instructions contained an error]$_{cause1}$ and [Alex followed the instructions]$_{cause2}$, [the dishwasher broke down]$_{effect}$.'

Taken together, the erroneous assembly instructions (cause 1) and Alex following these (cause 2) may be taken to *fully explain* the *effect*.

The present study investigated how partial explanations like (1) are processed in discourse. Moreover, we compared two types of causal relations differing in their involvement of domain knowledge. In addition to anomalous sequences as in (1), where the opposite effect is expected, we included situations where world knowledge doesn't make a particular prediction, introducing what we characterize as *novel* causal relations (cause2⇏effect and cause2⇏¬effect):

(3) *Weil [Maria sich auf die Bank setzte]$_{cause2}$, [bekam sie einen schlimmen Ausschlag]$_{effect}$.*
Because [Mary sat on the bench]$_{cause2}$, [she got a bad skin rash]$_{effect}$.

(3) doesn't contradict world knowledge: Rather, cause and effect seem unrelated: Whatever the possible effects of sitting on a bench are, getting a skin rash is usually not among them.

Materials: 15 anomalous and 15 novel discourses were constructed in three discourse order variants according to a 3x2 design (discourse order x causal relation). Discourses with the two causes conjoined within a *because* clause (*because cause1 and cause2, effect*, cf. (2)) served as controls. *Left dislocation* conditions, where *cause1* preceded the *because* clause (*cause1. because cause2, effect*), tested how easily causes can be integrated when not embedded under a causal connective. In the *right dislocation* condition of most interest here, *cause1* followed the *because* clause (*because cause2, effect. cause1, you see*). Importantly, all three orderings contained exactly the same 'nucleus' (*because cause2, effect*). All discourses were preceded by a two-sentence sequence introducing all referents and ended with a sentence concluding the story.

Pretests: Materials were pretested with respect to several aspects. Most importantly, the causal connectedness of anomalous and novel causal relations was rated (N=24) on a scale from -3 (*highly contradictory*) to +3 (*highly natural*) with 0 explicitly requested to indicate *no causal*

connection. We tested three conditions: (i) similar to (1)/(3): *because cause2, effect*; (ii) negating *cause2*: *because* ***negated(****cause2****)****, effect* ("because Alex did **not** follow the instructions, the dishwasher broke down"), and (iii) *because cause1 and cause2, effect*, as in (2). As expected, anomalous relations were rated oppositely in the positive (i) and negative (ii) cases (mean ratings: -2.0 vs. 2.1), whereas novel and negated novel cases both had no causal connection (0.0 vs. 0.1). Crucially, both types were rated as natural when they were part of a complex cause (*anomaly* 2.0; *novelty* 2.1). Another pretest (N=30) established that all three conditions for both *anomaly* and *novelty* items were rated equally plausible as a whole.

Predictions concerning **right dislocation** (*because cause2, effect*) as in (1) and (3) were captured in the framework of Halpern/Pearl (2005). Both anomalous and novel causal relations invoke a causal network consisting of a cause and an effect variable. However, the networks differ in one important respect: Anomalous relations violate established world knowledge, predicting a contrary distribution of cause and effect. Consequently, integrating *because cause2, effect* leads to a contradictory causal model calling for revision. Readers are therefore predicted to regress from the effect region to earlier parts of the discourse to check whether they had parsed *cause2* incorrectly. In novelty cases, on the other hand, the simple model invoked by *because cause2, effect* isn't contradictory, but insufficient. This is predicted to lead to abductive reasoning as to how the model could be plausibly *extended*. We thus expected integration difficulty right at the effect clause, that is, enhanced first-pass times on the effect clause, but less regressive eye-movements than for *anomaly*. In the **left dislocation conditions** (*cause1. because cause2, effect*), we assumed incremental discourse interpretation with immediate access to the global discourse representation (Hagoort/van Berkum, 2007). Thus, integrating the effect clause shouldn't be more difficult than in the control condition. Similarly, in Halpern/Pearl's theory, left dislocation (and control) provide full explanations, for which no abductive modelling effort is required.

Eyetracking experiment: Participants (N=27) read the discourses plus 30 filler texts while their eye-movements were monitored using an EyeLink 1000 system. In line with our predictions, *left dislocation* didn't differ from control at any segment. By contrast, *right dislocation* led to longer first-pass times and more regressions from the *effect* ROI. Furthermore, the effects differed for *anomaly* and *novelty*. **Inferential statistics** analyzing residual first-pass times of the *effect* ROI revealed a reliable interaction: Whereas *novelty* led to significantly longer first-pass times than control (mean difference: 169.7ms; p<.01), *anomalous right dislocations* didn't differ reliably from control (mean difference: -6.7ms). A logit mixed effects model analysis of first-pass regression ratios revealed an opposite pattern with significantly more regressions out of anomalous effect clauses (16.3%, control: 7.4%; p<.05) than for novel ones (12.6%, control: 8.9%; p=.34). Analyses of the second-pass times of *cause2* revealed the same interaction. *Right dislocation* led to longer second-pass times (SPT) than control, but this effect was more pronounced for *anomaly* (mean SPT: 679.1ms, control: 250.0ms) than *novelty* (mean SPT: 467.1ms, control: 292.6ms), as shown by a significant interaction (estimate = -262.05, t=2.42, p<.05).
In **conclusion**, the eye-tracking record of *anomalous* vs. *novel right dislocation* shows that subtle world knowledge distinctions and their associated inferential profiles are reflected in different temporal profiles when inferring from partial to full explanations during text comprehension.

REFERENCES

Aliseda (2006): *Abductive Reasoning*. Springer. **Hagoort/van Berkum** (2007). Beyond the sentence given. *PTRS*, B: 362. 801–811. **Halpern/Pearl** (2005). Causes and Explanations: A Structural-model Approach. *BJPS*, 56. 843–911. **Hobbs et al.** (1993). Interpretation as Abduction, *AI*, 63 (1-2): 69–142. **Köhne-Fuetterer et al.** (2021). The online processing of causal and concessive discourse connectives. *Linguistics* 59. 417–448. **Kuperberg et al.** (2011). Establishing Causal Coherence across Sentences. *JCN*, 23(5): 1230–1246. **Xiang/Kuperberg** (2015). Reversing expectations during discourse comprehension. *LaCoN* 30. 648–672. **Xu et al.** (2017). Influence of Concessive and Causal Conjunctions on Pragmatic Processing. *Disc. Proc.* 55. 387–409.

On a grammaticized lexical count-mass distinction in classifier languages: Experimental evidence from Tashkent Uzbek

Background Traditionally, nouns in classifier languages (CLs) were claimed to have uniform unindividuated (i.e., mass) semantics (e.g., Sharvy 1978). More recent literature argues that nouns crosslinguistically may be either underspecified (e.g., Borer 2005) or flexible w.r.t. count-mass (e.g., Pelletier 2012). Within these frameworks, then, the count reading is obtained only at the syntactic level, e.g., via classifiers.

Alternatively, others have argued that the count-mass distinction is, in fact, encoded in the semantics of nouns, even in CLs (e.g., Cheng & Sybesma 1998, Chierchia 2010, Rothstein 2010), a position supported by experimental data showing that despite the absence of count syntax, speakers of CLs have access to the core non-uniform semantics of nouns (e.g., Barner et al. 2009, Li et al. 2009). Importantly, though, even scholars recognizing the non-uniform nature of nouns in CLs assume that the linguistic count-mass distinction in these languages merely aligns with the cognitive object-substance distinction (e.g., Chierchia 2021).

One very recent exception is Erbach et al. (2021), who present preliminary empirical evidence suggesting that while there is considerable overlap between the linguistic categories *count-mass* and the cognitive categories *object-substance* in Japanese, a CL, the two are not fully aligned.

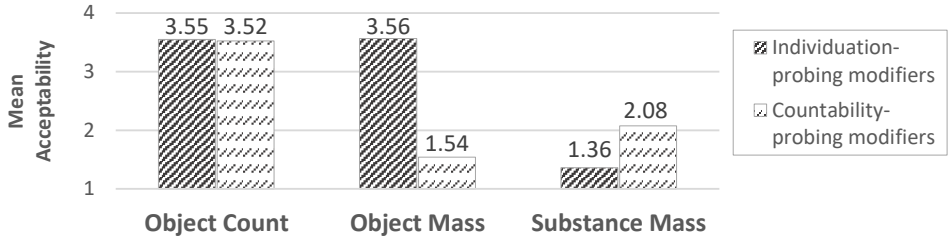
Current study Taking Erbach et al.'s exploratory findings as a starting point, the goal of the current study is to establish the existence of a lexicalized count-mass distinction in Tashkent Uzbek (TU), an obligatory classifier dialect of Uzbek. Specifically, we want to systematically demonstrate that nouns in TU are *not* uniformly unindividuated, and more importantly, that the count-mass distinction in TU – just like in English – *transcends* the cognitive object-substance distinction.

Methods We developed an experimental paradigm to elicit acceptability ratings of sentences with a range of modifier+noun combinations. Three nominal categories were tested: object count (e.g., *xat* 'letter'), substance mass (e.g., *qor* 'snow'), and so-called *object mass* (e.g., *mebel* 'furniture'). The modifiers were of two types: a) those sensitive to notional (un)individuation, and b) modifiers sensitive to morphosyntactic countability. Individuation-probing modifiers included an adjective of size *katta* 'big' and a reciprocal *bir-biriga o'xshash* 'similar to each other'. Countability-probing modifiers included a cardinal numeral followed by either a general classifier *-ta* (i.e., *uchta* 'three.CL') or by a collective suffix *-ala* (i.e., *ikkala* 'both'). The experimental design, along with some example items are presented in the table below.

		Modifier Type	
		Individuation-Probing	**Countability-Probing**
Noun Type	**Object Count**	Xonada **katta televizor** o'rnatildi. Room.LOC big TV installed.PSV 'A big TV was installed in the room.'	Vazirlikda **ikkala xat** imzolandi. Ministry.LOC two.COLL letter signed.PSV 'Both letters were signed at the ministry.'
	Object Mass	Zavodda **katta mebel** ishlab chiqarildi Factory.LOC big furniture produced.PSV 'Big furniture was produced in the factory.'	Yo'lda **ikkala pochta** yo'qoldi. Road.LOC two.COLL mail lost.PSV 'Both mails were lost on the road'
	Substance Mass	Rasmda **katta qor** chizildi Picture.LOC big snow drew.PSV 'Big snow was drawn in the picture.'	Laboratoriyada **ikkala gaz** suyultirildi Lab.LOC two.COLL gas liquefied.PSV 'Both gases were liquefied at the lab.'

There were 6 items in each condition, for a total of 36 experimental items. Examples from each sentence type are provided below. The task was conducted online via Qualtrics^{xm}. Verbal stimuli were presented as fully randomized audio files. Adult TU speakers (n=40) were asked to determine the likelihood that the test sentences could be produced by a native speaker of TU. Judgments were noted on a 4-point scale, with only the extreme ends explicitly labeled 1= *past* ('low'); 4= *baland* ('high').

Results and analysis A summary of the results is plotted in the graph below, presenting the mean scores for each modifier type across conditions.



The graph reveals that acceptability ratings in the object count condition are at near ceiling for both types of modifiers. This is in stark contrast with the results observed in the substance mass condition, where both modifiers receive low ratings. Particularly striking are the results of the object mass condition, in which speakers' judgments are sharply polarized as a function of modifier type. While modification by an individuation-probing modifier essentially mirrors the response pattern in the object count condition, countability-probing modifiers yield judgments that closely pattern with those in the substance mass condition.

A final, minor note concerns the slightly elevated ratings of sentences with countability-probing modifiers in the substance mass condition. We attribute this to the contextual mass-to-count shift enabled by the availability of the 'standard packaging' and the '(sub)kinds' reading, typical for substance mass nouns.

To analyze the significance of the findings, we performed a Paired-Samples T Test. We found a main effect of Noun Type (p <0.001). Additionally, a significant interaction of Noun Type and Modifier Type was found in the Object Mass and the Substance Mass conditions (p <0.001), but not in the Object Count condition (p = 0.5567).

Discussion Our data affirm the existence of two canonical noun classes in TU (object count and substance mass), which is clearly at odds with claims that nouns in CLs have uniform semantics. Most notably, our study also provides robust evidence for the existence of an additional, non-canonical nominal class, namely, *object mass nouns*. Morphosyntactically, object mass nouns pattern with mass nouns, i.e., they are incompatible with number coding; unlike canonical substance mass nouns, however, object mass nouns refer to individuals (cf. Barner & Snedeker 2005). As such, object mass nouns represent a dissociation between the linguistic count-mass distinction and the cognitive object-substance distinction (cf. Carey & Spelke 1996). Accordingly, under the view that in CLs, the linguistic and the cognitive distinctions fully align, such non-canonical nouns are predicted to be entirely absent in CLs such as TU. This prediction is not borne out by the results of the current study.

In sum, to the best of our knowledge, no existing research to date has been able to offer such clear evidence for three distinct nominal classes (object count, substance mass, and object mass) in a CL. These previously unavailable, systematically controlled, experimental data strongly indicate that a grammaticized lexical count-mass distinction is, in fact, encoded in the semantics of nouns in (at least some) CLs. Hence, our findings pose a serious challenge for the prevailing typology of noun semantics, which assumes a fundamental distinction between number-marking languages such as English and CLs like TU.

Selected References: Barner et al. (2009). Language, thought, and real nouns. *Cognition* 111. 329–344. | **Borer, H. (2005).** *In name only*. Oxford University Press. | **Cheng, L. & Sybesma, R. (1998).** Yi-wan tang, yi-ge tang: Classifiers and massifiers. *The Tsing Hua Journal of Chinese Studies*, *28*(3), 385–412. | **Chierchia, G. (2010).** Mass Nouns, Vagueness, and Semantic Variation, *Synthese, 174*: 99-149. **| Erbach et al. (2021).** Object Mass Nouns as an Arbiter for the Count–Mass Category. In *Things and Stuff: The Semantics of the Count-Mass Distinction* (pp. 167-192). Cambridge: Cambridge University Press.

Indirect discourse as mixed quotation: Evidence from self pointing gestures

Summary. The results of an experimental rating study are reported suggesting that self pointing gestures aligned with a third-person pronoun are available in German *indirect discourse* (ID) utterances. Following a proposal by Ebert and Hinterwimmer (2022) for self pointing in *free indirect discourse* (FID), self pointing in ID is interpreted as a *character viewpoint gesture* (CVG) from the matrix subject's perspective. Crucially, it is argued that in ID a perspective shift to the matrix subject can take place. It is proposed that ID is an instance of mixed quotation involving a demonstration (cf. Clark and Gerrig, 1990; Davidson, 2015) where self pointing is quoted from the matrix subject's original utterance.

Background. Davidson (2015) proposes a formal account of quotation under which it is demonstrational (cf. Clark and Gerrig, 1990). Ebert and Hinterwimmer (2022) report that self pointing aligned with a third-person pronoun is acceptable in German FID utterances, providing evidence in favor of an analysis of FID as mixed quotation (Maier, 2015), since self pointing aligned with a third-person pronoun can be treated as a demonstration with the pointing gesture being a quoted CVG (McNeill, 1992) from the protagonist's perspective. As a control condition, they also tested for self pointing in ID utterances and hypothesized them to be unacceptable because they have not been claimed to involve (mixed) quotation. However, self pointing was surprisingly acceptable in ID. The study reported here further explores this by means of a rating study that pairs self pointing with German ID utterances where either the speaker's or the matrix subject's perspective was made prominent. Since a perspective shift is more readily available when the matrix subject's perspective is prominent in ID (cf. Anderson, 2019), it is hypothesized that self pointing is more acceptable in those cases.

Experimental study. A rating study with 16 experimental items was conducted to test this hypothesis. German ID utterances were paired either with a self pointing or a beat gesture (factor Gesture). The gestures were always aligned with a third-person pronoun marked with a focal accent in order to increase the overall naturalness of the gesture. Assuming that ID can involve demonstrations if a perspective-shift toward the matrix subject takes place and that self pointing on a third-person pronoun is a demonstration, it is predicted that self pointing is more acceptable in the matrix subject condition than in the speaker condition. The beat gesture condition was included as a control condition as they should always be acceptable. Moreover, the ID utterances were manipulated in such a way that in one version the matrix subject's perspective was more prominent and in the other the speaker's perspective was more prominent (factor Perspective). The study was thus of a 2x2 design. The verbal stimuli (i.e., without a gesture) were tested in a pre-study in which participants had to select whose perspective was more prominent in the ID utterance: the speaker's or the matrix subject's. The items were then videotaped for the pilot study. An example of an experimental item is given in (1) (square brackets indicate gesture-speech alignment).

(1) a. **Matrix subject perspective:** Pia ging es erbärmlich. Sie fragte sich, warum ihre beste Freundin Anna, diese gottverdammte Saufziege, gestern Abend mal wieder [IHR] zu viel Wein nachgeschüttet hat, obwohl sie doch so wenig verträgt.

'Pia was feeling miserable. She wondered why her best friend Anna, that damned lush, had poured [HER] too much wine again last night, even though she couldn't handle it.'
+ self pointing or beat gesture

b. **Speaker perspective:** Pia ging es erbärmlich. Sie fragte sich, warum ihre beste Freundin Anna, die aber eigentlich nur die letzte Pfütze aus der Weinflasche loswerden wollte, gestern Abend mal wieder [IHR] zu viel Wein nachgeschüttet hat, obwohl sie doch so wenig verträgt.

'Pia was feeling miserable. She wondered why her best friend Anna, who was just trying to get rid of the last drops in the wine bottle, had poured [HER] too much wine again last night, even though she couldn't handle it.'

+ self pointing or beat gesture

The items were divided onto four lists according to a Latin square design and interspersed with 30 filler items. 60 native speakers of German participated in the study. Their task was to rate the utterances for acceptability on a 7-point Likert scale (1 = completely unacceptable; 7 = completely acceptable). An interaction of the two factors was predicted since self pointing was hypothesized to be more acceptable when the matrix subject's perspective is more prominent than when the speaker's perspective is more prominent in the ID utterance. The results show that self pointing was equally acceptable in both conditions of the factor Perspective (matrix subject: M = 4.47, SD = 1.95; speaker: M = 4.54, SD = 1.96). Beat gestures were also acceptable in both conditions of Perspective (matrix subject: M = 5.04, SD = 1.68; speaker: M = 5.24, SD = 1.67). An ordinal mixed-effects model with Gesture and Perspective as fixed effects and participants and items as random intercepts was fitted onto the data. It yielded a main effect for the factor Gesture (p < .001, z = -6.212). The rating differences for the two gesture types were significant in the matrix subject as well as the speaker condition (matrix subject: p < .001, z = -3.928; speaker: p < .001, z = -4.948).

Discussion and conclusion. The model output suggests that, contrary to the prediction, self pointing aligned with a third-person pronoun is acceptable irrespective of the prominent perspective in the ID utterance. However, it confirms the hypothesis that self pointing is acceptable in ID when the matrix subject's perspective is made prominent. The results thus go beyond the initial hypothesis. Following Ebert and Hinterwimmer (2022), this suggests that in ID a perspective shift toward the matrix subject takes place thus allowing for the interpretation of self pointing gestures as demonstrations of CVGs from the matrix subject's perspective. From a theoretical perspective, the results of the present study indicate that demonstrations in Davidson's (2015) sense and thus (mixed) quotation can also be present in instances of ID. This is in line with previous findings that some indexicals (e.g., temporal expressions in German and *tomorrow* in English) can shift in ID utterances (Plank, 1986; Anderson, 2019). Davidson's (2015) demonstration needs to be extended, however, so that it is able to also capture gestures (Ebert and Hinterwimmer, 2022). This proposal can be formally implemented by modifying Davidson's (2015) account, which can then also straightforwardly model the aforementioned shifting behavior of some indexicals in ID.

References

Anderson, C. J. (2019). Tomorrow isn't always a day away. Proceedings of *SuB 23*.

Clark, H. H. and R. J. Gerrig (1990). Quotations as demonstrations. *Language, 66(4)*.

Davidson, K. (2015). Quotation, demonstration, and iconicity. L & P, 38(6).

Ebert, C. and S. Hinterwimmer (2022). Free indirect discourse meets character viewpoint gestures. Proceedings of *LE 2020*.

Maier, E. (2015). Quotation and unquotation in free indirect discourse. *M & L, 30(3)*.

McNeill, D. (1992). *Hand and Mind: What Gestures reveal about Thought*. UChicago Press.

Plank, F. (1986). Über den Personenwechsel und den anderer deiktischer Kategorien in indirekter Rede. *Zeitschrift für germanistische Linguistik, 14(3)*.

Development of Mechanistic Support Language in Spanish Speakers in Colombia

Background Beyond basic spatial relations (e.g., teddy on table), we know little about how children learn to talk about mechanical support events (e.g., objects attached/hung from a surface via glue, magnet, etc.), and map them onto linguistic structures. Moreso, the majority of the research that has been done focuses on children learning English - a language that has several verbs that lexicalize support via a specific mechanism (e.g., glue, tape, clip, etc.)[1]. The broad goal of the current study is to deepen our understanding of spatial language acquisition by diversifying the populations that have been studied. Specifically, we explore how 4- to 6-year-old monolingual Spanish-speaking children and adults in Colombia, encode mechanical support events.

Figure 1. Examples of stimuli

Note. Images depict the static result state; stimuli are *dynamic* events of a hand attaching the figure object to the ground object. There are also examples of visible mechanisms; in hidden mechanism events the fastener is obscured by the figure and never shown to the participant.

Consider the mechanical support event depicted in Figure 1; *different* verbs can be used to encode the *same* spatial configuration (e.g., 'la niña pusó/ colgó/ pegó el papel al arbol' '*the girl put/ hung/ stuck/ taped the paper to the tree).* Typically in Spanish, the Basic Locative Construction (estar = *be on*), encodes a static state (e.g., 'la foto esta en la pared' = '*the picture is on the wall*'), whereas Put verbs (poner = *put*, colocar = *place*), act similarly semantically for dynamic events[1]. Moreover, rooted in Levin's English classification of verbs (1993), Verbs of Putting in a Spatial Configuration (e.g., colgar = *hang*, lean = *inclinar)* encode the spatial orientation of the figure object to the ground object without indicating the causal mechanism used in the support relation (i.e., 'Ella cuelga la foto de la puerta' = '*She hangs the picture from the door*' specifies that the picture is oriented in a downward orientation from the door, but the mechanism of support remains unclear).

Verbs of Attaching however, can either encode the specific mechanism in the lexical verb (often as a denominal) (e.g., enganchar = *hook)* or they can refer to a specific descriptor of the mechanism (e.g., sticky), without specifying the mechanism (e.g., pegar = *stick).* Since Verbs of Attachment encode the mechanism or provide a descriptor of the mechanism in the verb, we refer to these as Mechanism Verbs. General Verbs of Putting (poner = *put)* or Verbs of Putting in a Spatial Configuration (colgar = *hang)* are considered Non-Mechanism Verbs.

Recent findings show Spanish, contrary to English, has relatively fewer lexical verbs that encode the specific mechanism used (e.g., pegar = *stick)*. Several Specific Verbs of Attaching, commonly denominals in English, may be less available in Spanish than in English (e.g., 'tape' and 'pin.'). Thus, at least in terms of describing dynamic support relations, Spanish descriptions may compensate for the lack of lexical verbs that encode the mechanism (denominals) by using a separate adverbial clause to encode the mechanism (e.g., 'pegar con cinta' = '*stick with tape').* In terms of development, we consider how the limited availability of Mechanism Verbs may make learning easier (thus predicting little, if any, significant developmental change) or harder, because the mechanism is encoded outside of the main verb as an adverbial clause (thus predicting developmental change). We ask, 1) How do monolingual Spanish speakers encode dynamic mechanical support events? And 2) How may these descriptions change over development in monolingual Spanish speakers?

Procedure Spanish monolinguals, four to six-year-olds (*N* = 28), and adults (*N* = 25) were tested in Manizales, Colombia. Participants viewed videos of dynamic events where an agent attached a figure (paper) to a ground (tree or door) with a mechanism (clip, tape, pin), and were asked to describe the event (Fig. 1). Participant utterances (N = 304) were transcribed and coded for the type of verb; Mechanism Verb (e.g., pegar = *stick*), Non-Mechanism Verb (e.g.,

Simple verb: poner = *put* or Orientation verb: colgar = *hang*). We also coded whether the mechanism was encoded as the main verb, an adverbial phrase ('lo colgó con un gancho' = '*she hung it with a clip'*), or an individual clause "le pusó un clip y lo pusó ahí' = '*she put a clip on it and put it there'*).

Results Spanish-speaking adults were equally likely to use Non-Mechanism and Mechanism verbs (Figure 2). Further, when they did encode the mechanism (which was less than 60% of the time; Figure 3), they encoded it in a variety of linguistic structures, not only the main verb, thus motivating future cross-linguistic research on the encoding of spatial relations across languages and over development. Spanish-speaking children showed a similar pattern to their adult counterparts; binary logistic regressions showed no difference between children and adults for use of Mechanism or Non-Mechanism verbs (*p*s > .10). However, within the class of Non-Mechanism verbs (e.g. colgar, poner), children use more simple verbs (poner) compared to adults (*p* =.026).

Our findings suggest that both child and adult monolingual Spanish speakers encode the mechanism in a clause outside the main verb. In addition, children use more simple verbs (e.g. poner) than adults, whereas adults use more orientation verbs (e.g., colgar), suggesting developmental change in the acquisition of orientation verbs from childhood to adulthood. Implications for linguistic theory and spatial language acquisition will be discussed, including consideration of whether and how this pattern observed for Spanish in the domain of mechanical support compares to the encoding of path and manner in the domain of manner of motion[2].

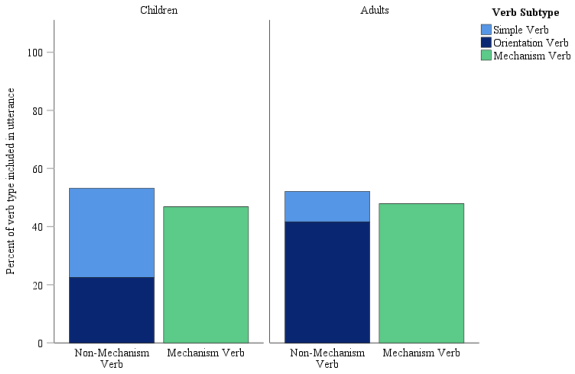


Figure 2. Percentage of verb types; Mechanism and Non-Mechanism verbs (i.e., Simple and orientation verbs) used in monolingual Spanish-speaking children and adults' mechanistic support descriptions

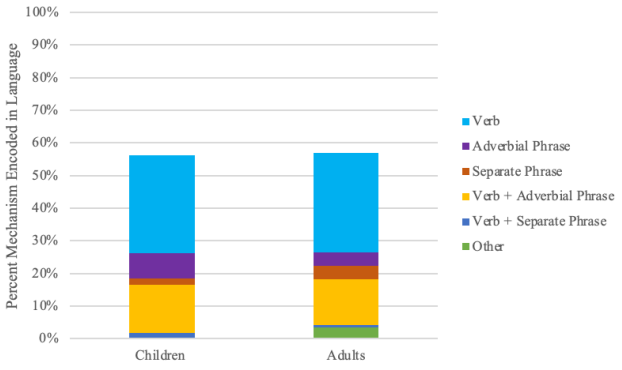


Figure 3. Percentage of trials that encoded the mechanism (and how it was encoded) in Spanish-speaking children's and adults' dynamic event descriptions

References

[1]Levin, B. (1993). *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press.

[2]Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description* (pp. 36-149). Cambridge: Cambridge University Press.

Towards a psycholinguistic model of bracketing paradoxes

German nominal compounds modified by an adjective typically have a canonical reading (1) in which the adjectives modifies the second noun of the compound. However, in some constructions, the adjective can equally or even preferentially modify the first noun (2). The latter construction is referred to as a *bracketing paradox* (Winkler 2015). These constructions appear to have different syntactic and semantic bracketing, seemingly violating compositionality principles (Bergmann 1980; Frege 1892). From a grammatical standpoint, the adjective should apply to the second noun or to the compound as a whole (3), but—crucially—not to the first noun (Bergmann 1980). How, then, are bracketing paradoxes licensed, whether odd (4) or unremarkable (5a)? Context, world knowledge, and pragmatic factors are potential contributors to interpretation preferences, along with morphosyntactic agreement, and the semantic compatibility between the adjective and nouns. Language economy and how lexicalized the compound is likely also play a role (Maienborn 2020). This multitude of possible factors calls for a broad empirical basis to enable further progress; empirical data on this phenomenon is, however, virtually non-existent. Our study begins to close this gap and lays the foundations for a comprehensive model of bracketing paradoxes.

(1) [Französischer[Sprachlehrer]] *French* language.*teacher* — canonical reading

(2) [[Französischer Sprach]lehrer] *French language*.teacher — bracketing paradox

(3) Verrückter Chemieprofessor (Crazy chemistry.professor) — ✖ Chemie ✔ Professor

(4) ?Vierstöckiger Hausbesitzer (Four.story house.owner) — ✔ Haus ✖ Besitzer

(5) a. Psychologische Beratungsstelle (Psychological counseling.center) — AN_1N_2

b. Psychologische Beratung (Psychological counseling) — AN_1

c. Psychologische Stelle (Psychological center) — AN_2

Experiment 1 investigated the role of semantic compatibility between the adjective and the individual nouns in the adjective-nominal-compound construction. 36 participants were asked to evaluate 204 AN_1N_2 in one of 3 conditions, as in (5). They assigned 1–5 scores on the dimensions of naturalness, comprehensibility, and stylistic form. The ratings across scales were highly correlated ($r \geq 0.95$). We, therefore, used the mean of these ratings which was scaled to the interval $[0, 1]$ for analysis. All but three items received good ratings for either AN_1 or AN_2 or for both (Fig. 1A). This is due to our attempt to exclude constructions where the adjective was a poor match for both nouns, as these are unlikely to be produced. As a result, AN_1 and AN_2 ratings were negatively correlated ($r = -0.5$). A Bayesian Beta regression modeled the averaged and scaled ratings of the AN_1N_2 constructions as a function of the corresponding AN_1 and AN_2 ratings along with their interaction (Fig. 1A–C). As expected, high AN_2 ratings were predictive of high AN_1N_2 ratings ($\beta = 6.3$, 95%-CrI $[4.7, 8.2]$, Fig. 1B). However, AN_1 ratings, too, had a positive, albeit smaller effect on AN_1N_2 ratings ($\beta = 3.3$, 95%-CrI $[1.7, 5.2]$). Crucially, there was an interaction of the AN_1 and AN_2 ratings ($\beta = -4$, CrI $[-6.2, -2.0]$, Fig. 1C): When AN_2 ratings were low, AN_1 ratings had a substantial positive effect. When AN_2 ratings were high, higher AN_1 ratings slightly *reduced* the AN_1N_2 ratings, suggesting a perceived conflict.

Experiment 2 investigated which noun in a compound is modified by the adjective, as this is not necessarily determined by the ratings obtained in Exp. 1. 20 participants indicated for 235 AN_1N_2-phrases (5a) whether the adjective modifies N_1, or N_2, or whether they were unsure. Participants overwhelmingly selected one of the nouns, with only $< 3\%$ "unsure" answers (Fig. 1D). Therefore, we excluded "unsure" answers from the analysis. 30% of compounds exhibited a flexible attachment preference, with 6 to 14 votes for either N_1 or N_2. A Bayesian logistic regression modeled the choice of attachment site (N_2 or not) as a function of the corresponding AN_1 and AN_2 ratings from Exp. 1 and their interaction. There were two main effects ($\beta_{N1} = 1.3$, 95%-CrI $[-2.5, -0.2]$; $\beta_{N2} = 8.0$, 95%-CrI $[6.5, 9.6]$) as well as an interaction ($\beta = -5.6$, 95%-CrI $[-7.6, -3.7]$, Fig. 1E–F).

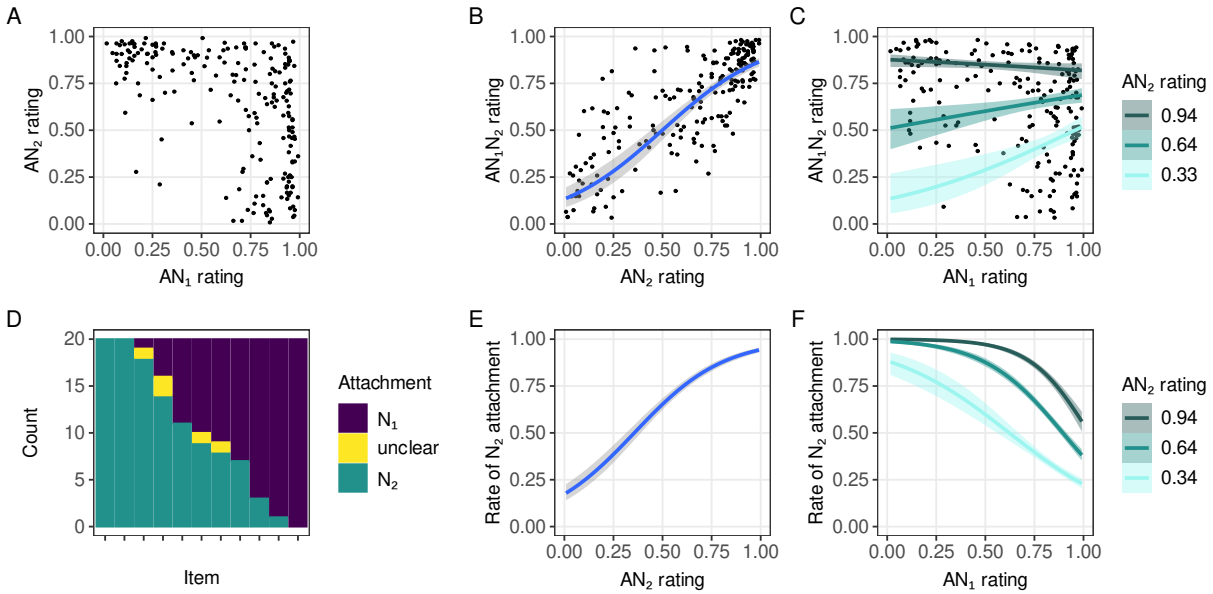


Figure 1: **A:** Exp. 1. Relationship between AN_2 and AN_1. **B:** Exp. 1. Relationship between AN_1N_2 and AN_2. **C:** Exp. 1. Relationship between AN_1N_2 and AN_1. Lines correspond to AN_2 groups. **D:** Exp. 2. Representative selection of items in Exp. 2 ordered by number of N_1/N_2/unsure answers. **E:** Exp. 2. Relationship between AN_1 rating and N_2 adjective attachment. **F:** Exp. 2. Relationship between attachment preference and AN_1 rating. Lines correspond to AN_2 groups.

High AN_2 ratings and low AN_1 led to more N_2 attachment. When AN_2 ratings were low, high AN_1 had a stronger effect on N_2 attachment. When AN_2 ratings were high, AN_1 had a lesser influence on N_2 attachment.

Conclusions: Contrary to grammatical and strictly compositional constraints on their relationship, the first noun plays an important role in the acceptability of a nominal compound modified by an adjective. This is in spite of the second noun's dominance over the adjective and compound. This result aligns with the role of semantic and pragmatic factors on such constructions, which may favor an otherwise grammatically unavailable attachment site. When both nouns are good matches for the adjective, acceptability is slightly reduced suggesting a perceived conflict or competition between possible attachment sites. Thus, even though both nouns have a positive effect on a compound's acceptability, their effects are not additive. In the absence of a suitable head noun candidate, the first noun becomes an attractive modification target for the adjective. This work suggests that the interpretation of bracketing paradoxes is not a clear-cut choice between the nouns, and there is much uncertainty and disagreement on interpretation between readers. Open questions include how do a speaker and listener agree on an interpretation and what distinguishes natural (5a) and unnatural (4) sounding bracketing paradoxes.

Bergmann, Rolf (1980). "Verregnete Feriengefahr und Deutsche Sprachwissenschaft. Zum Verhältnis von Substantivkompositum und Adjektivattribut." In: *Sprachwissenschaft* 5.3, pp. 234–265.

Frege, Gottlob (1892). "Über Sinn und Bedeutung." In: *Zeitschrift für Philosophie und philosophische Kritik* NF 100, pp. 25–50.

Maienborn, Claudia (2020). "Wider die Klammerparadoxie: Kombinatorische Illusionen beim Adjektivbezug auf NN-Komposita." In: *Zeitschrift für Sprachwissenschaft* 39.2, pp. 149–200.

Winkler, Julia (2015). "Kleine Geschichte der 'schiefen Attribute'." In: *ZAS Papers in Linguistics* 58, pp. 124–139.

Evaluating context-independent meaning in two English discourse particles

Background Linguistic meaning is divisible into two categories: context-independent and context-dependent (e.g. Gutzmann, 2014). Whereas the context-independent meaning of a lexical item is stable across contexts (it is considered lexically encoded), a context-dependent meaning is the product of a lexical item's use in a particular context. For many categories (e.g. discourse markers and connectives) there is often disagreement over what a word's meaning(s) is/are, as well as whether a given meaning is context-independent or context-dependent (e.g. see Ariel and Mauri [2019] for 'or'). We present an experiment designed to help determine whether language users understand the proposed meanings of two English discourse markers as being context-independent. We specifically ask whether, holding all contextual information steady, the audibility of a discourse marker's segmental information (i.e. its lexical information being interpretable) affects listeners' judgments on the extent to which speakers are demonstrating the meanings in question.

We focus on two discourse markers, *apparently* and *actually.* There is disagreement in the literature as to what these words mean and what they are used for (see e,g. Glougie [2016] for discussion). Our experiment is restricted to testing for two proposed dimensions of meaning: certainty and surprise. Considering both context-independent and context-dependent analyses, *actually* has been associated with speaker certainty and related notions such as being in the possession of reliable evidence for a claim (Biber & Finegan, 1988, Glougie, 2016, Sarfo-Kantankah & Ben Kudus Yussif, 2019). *Apparently* has been associated with speaker uncertainty (Mittwoch, Huddleston and Collins, 2002, Glougie, 2016, Carretaro and Zamorano-Mansilla, 2019). The uncertainty meaning of *apparently* is often argued to be a pragmatic function stemming from a core evidential meaning (e.g. Glougie, 2016). X and Y (2021) note that, like certain indirect evidentials in other languages, *apparently* can be used in contexts of speaker surprise (DeLancey, 2001). We therefore test three hypotheses: 1) Actually encodes speaker certainty; 2) Apparently encodes speaker uncertainty; 3) Apparently encodes speaker surprise.

Methods All utterances containing *apparently* (n=24) were extracted from PhonBank's video-taped Providence corpus (Rose & MacWhinney, 2014, Demuth, Culbertson & Alter, 2006). Utterances were all naturally produced by adults in speech around children (this study is part of a larger study on acquisition). For each *apparently* token, the utterance containing *actually* that occurred closest in time was also extracted. The resulting 48 short video clips formed the regular condition stimuli set. For a second condition, the target word was low-pass filtered to remove segmental information; only prosodic information was audible. The rest of the utterance was unaltered, meaning the only difference between the conditions was whether the target word was identifiable. 294 participants were recruited from linguistics classes at a North American university. They received a course credit for participating. After exclusions (technical issues, n=48; non-native English speakers, n=74; diagnosed hearing disability or hearing loss, n=8), 164 participants were included in the analysis. The design was between subjects. Participants were asked to watch each video clip and answer the question "How surprised does the speaker seem?" or "How certain does the speaker seem?" (with 7 being "extremely surprised/certain" and 1 being "extremely unsurprised/uncertain"). Because there were two questions asked of each clip, participants answered a total of 96 questions each.

Predictions We predicted that, for *apparently*, participants in the regular condition would rate speakers as seeming more surprised and less certain than in the low-pass filter (LPF) condition. For *actually*, we predicted that participants in the regular condition would rate speakers as seeming more certain than in the low pass filter condition. If the expected differences are found,

then encoded lexical information (context-independent meaning) must be at least partly responsible for listeners' beliefs about a speaker's level of certainty, uncertainty or surprisal. If no differences are found between conditions, this would suggest that either these words do not have these meanings at all, or that these meanings are not encoded in the words themselves, but are merely aspects of the larger contextual conditions in which these words tend to be used—which were the same in both conditions.

Results 2 tailed, paired t-tests on mean token ratings in the two conditions indicate that for *apparently*, participants in the regular condition rated speakers as seeming more surprised than participants in the LPF condition (REG m=4.33(1.32); LPF m=3.93(1.28); $p<0.001$). (Interestingly, this difference was also true of *actually* test items, where a difference in surprise ratings was not expected. In fact, the surprise use finds some support in the literature, e.g. Greenbaum [1969].) Participants also rated speakers as seeming less certain in the regular condition than in the LPF (REG m=3.76(1.41); LPF m=4.23(1.29); $p<.0001$). For *actually*, participants in the regular condition rated speakers as seeming more certain than in the LPF condition (REG m=4.89(1.37); LPF m=4.62(1.42); $p<.01$).

Conclusion Participants' ratings on how surprised or certain speakers seemed were affected by whether or not the target word was identifiable. All results were in the directions predicted. Although there are many proposed meanings for these words, the results suggest the words do have the hypothesized meanings (perhaps among others): native English speakers may consider surprise and uncertainty part of the context-independent meaning of *apparently* and may consider certainty part of the context-independent meaning of *actually*. At minimum, it would seem that lexically-encoded meaning interacts significantly enough with the surrounding context to alter participants' understanding of a speaker's level of certainty or surprise.

References

Ariel, M., & Mauri, C. (2019). An 'alternative' core for or. *Journal of Pragmatics*, *149*, 40-59.
Biber, D. & E. Finegan. 1988. Adverbial Stance types in English. *Discourse processes* 11:1, 1-34.
Carretaro, M., & Zamorano-Mansilla, J. R. (2019). Disentangling epistemic modality, neighbouring categories and pragmatic uses: the case of English epistemic modal adverbs.
DeLancey, S. (2001). The mirative and evidentiality. *Journal of pragmatics*, *33*(3), 369-382.
Demuth, K., J. Culbertson, & J. Alter. 2006. Word-minimality, Epenthesis, and Coda Licensing in the Acquisition of English. Language & Speech, 49, 137-174.
Glougie, J. R. S. (2016). *The semantics and pragmatics of English evidential expressions: the expression of evidentiality in police interviews* (Doctoral dissertation, University of British Columbia).
Greenbaum, S. (1969). Studies in English adverbial usage. Longmans, Green & Co. Ltd.
Gutzmann, D. (2014). Semantics vs. pragmatics. *The companion to semantics.* Oxford: Wiley.
Mittwoch, A., Huddleston, R. D., & Collins, P. (2002). The clause: adjuncts. In R. Huddleston & G. Pullum (Eds.), *The Cambridge Grammar of the English Language* (pp. 663-784). Cambridge University Press.
Rose, Y., & MacWhinney, B. (2014). The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development. In Durand J., Gut U., & Kristoffersen, G. (Eds.), *The Oxford handbook of corpus phonology* (pp. 380-401). Oxford, UK: Oxford University Press.
Sarfo-Kantankah, K. S., & Yussif, B. K. (2019). The use of actually in a non-native English parliamentary context: a corpus study. *Word*, *65*(4), 234-251.
X & Y. 2021. Polysemy and ambiguity in indirect evidential use around children. Paper presented at Dubrovnik Conference on Cognitive Science, Dubrovnik, Croatia.

Group membership impact on referential communication

Arriving at the speaker's intended meaning involves linguistic, cognitive and social processes, which include incorporating knowledge concerning the speaker's identity. Previous research focused on social characteristics of the speaker or listener, but often overlooked effects of group membership and specifically intergroup interactions (i.e., when speaker and listener are not part of the same social group). Intergroup interactions have been shown to deplete executive functions resources (Richeson & Trawalter, 2005) and interfere with theory of mind abilities (Hackel et al., 2014).

Additionally, previous research showed that intergroup settings affect interpretation when the content is group-relevant (Beltrama & Schawrz, 2021). Importantly, this is also the case in group-neutral interpretations for jokes (Morisseau et al., 2017), as well as regularized scalar implicatures (the authors, in prep).

In the current study, we expand our finding to another well-tested case, which is directly related to ToM abilities, the communication of referring expressions. To effectively use referents, interlocutors have to consider which objects are shared and which are privileged. This requires representing the knowledge of the others (Heller et al., 2012). If the ability to represent the knowledge of outgroup members declines, then a more egocentric perspective is expected in intergroup settings.

To test this hypothesis, we employed the Director's Task (Keysar et al., 2000). In this task, participants are presented with an array of objects in grid display (Fig 1.). A confederate director instructs them on which object to choose. Critically, some of the cells in the grid are only privileged to the participants. In critical trials, privileged objects are competitors for the object mentioned by the director. If participants are able to represent the director's perspective, they should ignore those cells completely. Yet, previous studies have shown that participants do consider the competitor to some extent (e.g., Barr, 2008). We assume both more errors and longer processing times when interacting with an outgroup member than when interacting with a neutral speaker.

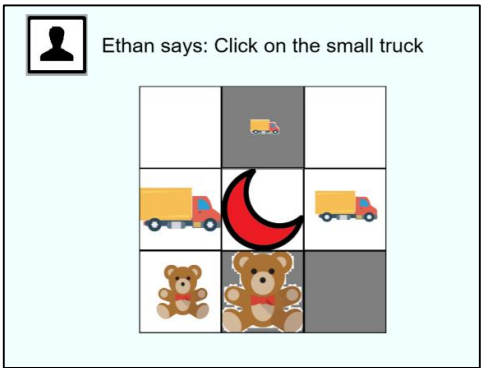


Fig 1. An example of a critical trial – the smallest truck is privileged (as indicated by the grey background) so an accurate response would be to choose the medium sized truck.

We conducted an online experiment (N=72, preliminary results). Participants were American native English speakers who identified themselves as Democrats. To avoid intergroup task effects, we divided the participants into three groups: (i) an ingroup condition where the director was a member of their own group (democrat), (ii) an outgroup condition where the director was a member of the other group (republican), (iii) a control group, to serve as a baseline (no party affiliation mentioned).

In the experimental groups, participants first had to indicate their political affiliation by clicking on the appropriate party logo and to answer a group identification questionnaire (adapted from Leach et al., 2008). All the participants were then told they will play a "game" with another player (who was actually a virtual-decoy) who played as the director in the game. In the experimental groups, the party affiliation of the speaker was constantly highlighted.

We modelled the rates of correct (non-privileged) responses with a fixed effect group (control/ingroup/outgroup; Fig 2a.). The model did not reveal an effect of group ($p = 0.11$). We then modelled the RTs for correct responses in both control (no privileged option) and critical trials with fixed effects of group and trial-type, as well as the interaction between the two (Fig 2b.). The model revealed an interaction ($p < 0.05$) where RTs for critical trials were significantly longer than for control trials in the outgroup condition ($p < 0.05$), but not in the ingroup and

control conditions. There were no main effects of group or trial type (p =0.64 and p = 0.51). We did not find correlations for level of identification.

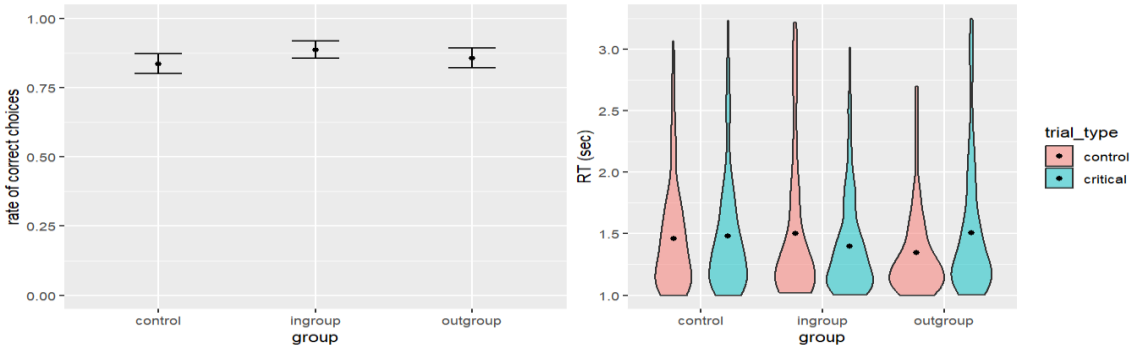


Fig 2. a. rate of correct responses in critical trials by group; b. RT for correct responses in the control and critical trials.

Our preliminary results show that a high-threat intergroup setting impacted the processing time of referring expressions, though it did not affect accuracy. This suggests an egocentric perspective is considered more often in cases where the speaker is an outgroup member, perhaps due to difficulty in representing the knowledge of the. This processing cost can, in turn, result in more inefficient communication.

Notably, Savitsky et al. (2010) suggested that increased familiarity between interlocutors (friends rather than strangers) causes listeners to adopt a more egocentric perspective. They argued that this is because listeners erroneously attributed a similar perspective to their familiar interlocutors. Thus, these results are interesting in that they show: a. that an egocentric perspective may also be reached by a lesser identification with the speaker; b. that increased *similarity* between the interlocutors in terms of group membership (i.e., ingroup interactions) do not lead to the adoption of egocentric perspectives. This may suggest a difference between two types of 'familiarity' - frequency of interaction or similarity between interlocutors (as dissociated by Brown & Levinson, 1987).

References:

Richeson, J. A., & Trawalter, S. (2005). *Journal of Personality and Social Psychology*, *88*(6), 934–947; **Hackel, L. M., Looser, C. E., & Van Bavel, J. J.** (2014). *Journal of Experimental Social Psychology*; **Leach, C. W., Van Zomeren, M., Zebel, S., Vliek, M. L., Pennekamp, S. F., Doosje, B., ... & Spears, R.** (2008). *Journal of personality and social psychology*; **Beltrama, A., & Schwarz, F.** (2021). *Semantics and Linguistic Theory*; **Morisseau, T., Mermillod, M., Eymond, C., Henst, J.-B. V. D., & Noveck, I. A.** (2017). *Interaction Studies*; **Heller, D., Gorman, K. S., & Tanenhaus, M. K.** (2012). *Topics in cognitive science*; **Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S.** (2000). *Psychological Science*; **Barr, D. J.** (2008). *Cognition*, *109*(1); **Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A.** (2011). *Journal of experimental social psychology*, *47*(1), 269-273.

Perfect ever after:
An empirical investigation of tense-based event construals in English and Spanish

In English and Spanish, just as in many other languages, speakers can use simple past tense (1) or perfect tense (2) to describe something that happened in the past:

(1) Julius travelled to London
(2) Julius has travelled to London

But what is the difference in meaning between (1) and (2)? A longstanding descriptive observation is that while (1) and (2) both refer to the past event, (2) establishes a link to the present time in a way that (1) does not, therefore creating a more complex meaning. Formal accounts of the perfect differ in their assumptions about the nature of such a link, and the semantic contribution of the perfect overall (e.g., Iatridou et al. 2001; Klein 1994); but one prominent analysis concludes that the perfect creates a state from a previous event (Moens 1987, Parsons 1990). Here, we test this "perfect-as-state" analysis empirically. We present the results of two behavioral studies (total N=960) in English, a Germanic language, and Spanish, a Romance language. Our results show that the perfect tense in both languages leads to event construals that have more in common with states than events in the simple past: (1) refers to a past travelling event, whereas (2) refers to Julius' acquired property of having been to London.

In our studies, we used *boundedness* as a tool to tease apart the construals of past vs perfect events. Objects with boundaries can be counted (e.g. 'three apples'), whereas unbounded substances cannot (*'three applesauces'; only sortal reading possible); this property of individuability has been shown for bounded objects as well as events (Barner et al. 2008; Wittenberg & Levy 2017). We applied it to the domain of tense: If the perfect denotes a state, it should be unbounded, like mass nouns and durative verbs, triggering lower rates of individuation for events in perfect tense compared to events in past tense.

Experiment 1 replicated Barner et al. (2008) and extended it to Spanish: We manipulated nominal syntax (count vs. mass) and event type (durative vs. punctual); in addition to these conditions, we also manipulated tense (past vs. perfect; all between subjects). In the critical trials (n=12), participants read a set of vignettes describing two characters performing actions, normed to be either unbounded and durative, such as *dancing*, or bounded and punctual, such as *jumping*. A question followed, using a light verb in past or perfect form, followed by a noun either in mass syntax (e.g., *Who did/has done more dancing/jumping*; *bailar/saltar ¿quién hizo/ha hecho más?*) or count syntax (e.g., *Who did/has done more dances/jumps; bailes/saltos ¿quién hizo/ha hecho más?*). Participants had two response choices: One character did more of the action in number of times, and the other character did more of the action in a different, pre-tested dimension (e.g., jumping higher, dancing longer). The number-based choice therefore served as measure of individuation. **Results (Fig.1):** We successfully replicated Barner et al.'s (2008) results in two languages (N=240ea.), finding that speakers quantified events in count syntax more than in mass syntax. This was primarily driven by event type: punctual events resulted in high individuation rates regardless of nominal syntax, whereas durative events in mass syntax yielded lower

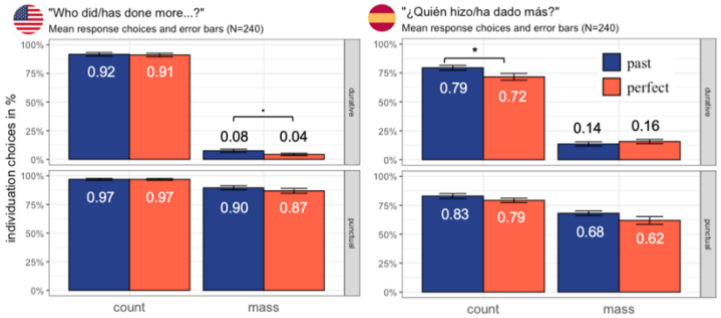


Figure 1. Mean individuation responses for English (left) and Spanish (right). In both English and Spanish, we replicated Barner et al's (2008) findings. For visual clarity, we only illustrate the significant effects of pairwise comparisons (=significant; •=marginally significant); for other results, please refer to the text. Error bars represent Standard Errors.*

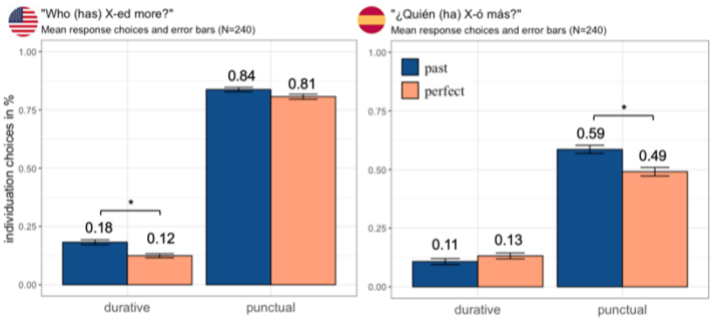


Figure 2. Mean individuation responses for English (left) and Spanish (right). For visual clarity, we only illustrate the significant effects of pairwise comparisons (=significant;•=marginally significant); for other results, please refer to the text. Error bars represent Standard Errors.*

individuation rates compared to count syntax (all effects and interactions *Df*s=1, χ^2s>118.36, *p*<0.001). There was a marginally significant effect of tense in Spanish (*Df*=1, χ^2= 3.33, *p*=0.07), whereas in English the trend was only numerical, but both pointed into the predicted direction: less individuation in the perfect tense, compared to the past tense. **Experiment 2** (N=240ea.) tested the effect of tense on individuation without nominal syntax as intermediating factor. Instead of using light verb constructions, such as asking *Who did more jumps?*, we used full verb forms in past tense (e.g., *Who jumped/danced more? ¿Quién saltó/bailó más?*) and perfect tense (e.g., *Who has jumped/danced more? ¿Quién ha saltado/ bailado más?*). Other than that, the procedure followed that of Experiment 1. **Results (Fig.2):** A main effect of event type confirmed that durative events give rise to significantly less individuation than punctual events (*Df*s=1, χ^2s>689.12, *p*<0.001). Crucially, we also found the predicted effect of tense: Perfect tense led to less individuation, both in English and in Spanish, as predicted by the perfect-as-state hypothesis (*Df*s=1, χ^2s >10.97, *p*<0.001).

Discussion: The pattern of results from four experiments clearly indicates: In the absence of strong cues of individuation such as nominal syntax, the perfect tense leads to more stative event construals compared to past tense, constituting the first empirical evidence supporting the 'perfect-as-state' hypothesis, so far advocated only on theoretical grounds (Bybee et al. 1994; Sánchez-Marco 2012; Dowty 1979; Katz 2003). Our findings lay the groundwork for further investigations across languages and tense systems.

References

Barner D., L. Wagner & J. Snedeker. 2008. Events and the ontology of individuals: Verbs as a source of individuating mass and count nouns. *Cognition* 106, 805-832.

Bybee, J.L., R. Perkins & W. Pagliuca. 1994. *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: University of Chicago Press.

Dowty, D. 1979. *Word meaning and Montague Grammar*. Dordrecht: Reidel.

Iatridou, S., E. Anagnostopoulou & R. Izvorsky. 2001. Observations about the form and meaning of the Perfect. In Michael Kenstowicz (ed.) *Ken Hale: A Life in Language*, 189-238. Cambridge MA: MIT Press.

Katz, G. 2003. On the stativity of the English Perfect. In A. Alexiadou, M. Rathert & A. von Stechow (eds.) *Perfect Explorations*. 205-234. Berlin: Mouton de Gruyter.

Klein, W. 1994. *Time in Language*. London: Routledge.

Mittwoch, A. 2008. The English resultative perfect and its relationship to the experiental perfect and the simple past tense. *Linguistics and Philosophy 31*, 323-351.

Moens, M. 1987. *Tense, aspect and temporal reference*. PhD Thesis. University of Edinburgh.

Parsons, T. 1990. *Events in the Semantics of English*. MIT Press.

Portner, P. 2003. The (temporal) semantics and (modal) pragmatics of the perfect. *Linguistics and Philosophy 26* (4), 459-510.

Sánchez-Marco, C. 2012. *Tracing the development of Spanish participial constructions: an empirical study of semantic change*. PhD Thesis. UPF.

Wittenberg, E., & R. Levy 2017. If you want a quick kiss, make it count: How choice of syntactic construction affects event construal. *Journal of Memory and Language*, *94*, 254-271.

Counting uncountables and measuring countables – unpreferred, not ungrammatical

English nouns are categorized as 'count' or 'mass' according to grammaticality with numerals (*one tube*/**toothpaste*). This suggests that meanings of count/mass nouns (CNs/MNs) are discrete /continuous, respectively [2,5]. One piece of evidence for this hypothesis is that in comparatives (*more tubes*/*toothpaste*), CNs (MNs) usually trigger counting (measurement, resp.). However, this pattern is complicated by many findings. Comparatives with *object mass nouns* (OMNs) like *mail* or *baggage* often support counting like the CNs *packages* or *bags*. Notably, with stimuli like fig.1 and the question *who has more mail*/*packages*, answers equally showed strong preference for counting [1]. And the puzzle goes beyond OMNs. Some CNs support non-cardinal measurement: e.g. *Anna put more oranges in the punch than Ben* may compare quantities of orange juice rather than numbers of oranges [6]. Conversely, substance MNs (SMNs) like *sand* may trigger counting, e.g. in *there are more stars in the universe than sand on earth*, which compares the number of grains of sand to the number of stars [7]. This leads us to two inter-related problems:

CN problem: To what extent can CNs compromise their count-based interpretation?
MN problem: To what extent can MNs compromise their measure-based interpretation? Specifically, do OMNs support counting as strongly as CNs?

We hypothesize that all nominal comparatives allow both measurement and counting. The choice between strategies is affected both by the discreteness (continuity) of CN (resp. MN) denotations, and by the perception of real-world objects as discrete or continuous [5], which may trigger a shift in denotations. To test this hypothesis, we study four types of contrastive pairs of comparatives:

(1) a. CN vs. OMN: *A has more packages*/*mail than B*
b. CN vs. SMN: *A has more <u>rocks</u> than clay* / *more <u>rock</u> than clay*
(2) a. CN vs. *number of*: *A needs more bananas* / *a greater number of bananas than nuts*
b. SMN vs. *volume of*: *A has more rock than clay* / *a greater volume of rock than clay*

(1a,b) test the effect of the mass/count distinction on quantity judgements, where (1a/b) favors counting/measurement, respectively. (2a,b) test the appearance of measurement with bare CNs (vs. the baseline of 'number of' CNs) and of counting with bare SMNs (vs. 'volume of' SMNs). A central methodological point in this study of "exceptional" strategies is the following assumption:

When choosing a count/*measure interpretation, a speaker needs to consider the most probable way of comparing the salient perceptions of quantities using the given noun.*

For example, let us consider the following sentences in relation to fig.1 (BrE=British English):

(3) a. Anna has more mail (BrE: post) than Ben b. Ben has more mail (post) than Anna
(4) a. Anna has more packages than Ben b. Ben has more packages than Anna

When asked on truth-value of one sentence in (3-4) in isolation, most speakers use counting as a default. Thus, judgements on fig.1 are predominantly positive/negative on (3a,4a)/(3b,4b) resp., as [1] observe. To test the presence of an (unpreferred) measurement strategy, we show a speaker <u>*both*</u> (3a) and (3b) (or (4a) and (4b)), asking her to choose between two statements:

(I) "I imagine either one of the two sentences might be used to describe the situation"
(II) "Only one of the sentences can be used felicitously"

When (II) is selected, we ask on the identity of the unique felicitous sentence. If, despite bias towards counting, OMNs allow measurement more readily than CNs, we should expect speakers to accept (3b) with fig.1 – either by choosing the ambiguous strategy (I) or by unambiguously choosing (3b) in (II) – more frequently than (4b). This guides our testing of all the cases in (1-2).

<u>*Materials & procedure*</u>. For the four sentence types (1a-b,2a-b) we selected nouns as follows:

CN vs. OMN: *packages-post, bags-baggage, instruments-equipment, sofas-furniture, weapons-weaponry, stationery items-stationery*
CN vs. SMN: *rocks-rock* (+*clay*), *chocolates-chocolate* (+*flour*), *stones-stone* (+*soil*), *ropes-rope* (+*sand*)
CN vs. *number of*: *bananas+hazelnuts*, *apples+almonds*, *cod fillets+peas, potatoes+olives*
SMN vs. *volume of*: *rock+clay, chocolate+flour, stone+soil, rope+sand*

This led to 6/4/4/4 ×2 sentence pairs as in (3) and (4). After training with choices like (I) and (II) in contrasts unrelated to mass/count, each participant was presented with one pair of sentences as in (3) or (4) together with a description of a situation where counting and measurement should lead to different answers. She was asked to choose between (I) and (II), and specify her selected sentence in case she chose (II). The four types of stimuli led to five experiments, where situations were described graphically, or, in cases where stimuli proved hard to depict, textually:

Exp1 – OMN vs. CN: graphical, e.g. fig.1 for each of the sentences in (1a)
Exp2(a) – CN vs. SMN: graphical, e.g. fig.2 for each of the sentences in (1b)
Exp2(b) – CN vs. SMN: textual, e.g. the description below for each of the sentences in (1b)
Anna bought: rock(s)-10 pieces of 3kg each (total 30kg); clay- 4 lumps of 25kg each (total 100kg)
Exp3 – CN vs. *number of*: textual, e.g. the description below for each of the sentences in (2a)
Anna needs: 300gr bananas (about 3 medium ones); 100gr hazelnuts (about 60 average nuts)
Exp4 – SMN vs. *volume of*:graphical, e.g. fig.2 for each of the sentences in (2b)

Using *Prolific*, 479/320/321/320/320 different speakers of British English (309/205/178/222/202 female, mean age 42.1/41.9/42.5/39.5/42.2, resp.) were recruited for these five experiments.

Results. In Exp1/2 the CN was expected to show less measurement than the other noun (OMN/SMN, resp.). In Exp3/4, the '*number*/*volume of* phrase was expected to show less measurement/counting than the other noun (CN/SMN, resp.). These expectations were supported by the total acceptance rates of the exceptional strategy, in terms of selecting ambiguity (I) or by selecting the exceptional strategy unambiguously (II). Rates are reported below with respective Odds Ratios and 95% Confidence Intervals and p-values according to Fisher Exact Test:

Exp1- *measure* CN (OMN) 37 (110) of 239 (240): OR=0.22 (95% CI [0.14,0.34], *p<0.00001*)
Exp2(a) - *count* SMN (CN) 58 (104) of 161 (159): OR=0.30 (95% CI [0.19,0.47], *p<0.00001*)
Exp2(b) - *count* SMN (CN) 40 (64) of 160 (161): OR=0.51 (95% CI [0.32,0.82], *p<0.01*)
Exp3 - *measure* CN (*num. of*) 105 (30) of 159 (161): OR=0.12 (95% CI [0.07,0.20], *p<0.00001*)
Exp4 - *count* SMN (*volume of*)58 (18) of 161 (159):OR=0.23 (95% CI [0.13,0.41], *p<0.00001*)

Per items significant differences appeared with 4/3/2/4/3 out of the 6/4/4/4/4 items respectively.

Conclusions. In five experiments, participants were shown a situation where both measurement and counting are pragmatically possible. The variable between participants in each experiment was the linguistic stimulus. The questions tested whether participants experienced ambiguity, or unambiguously preferred one of the strategies. The reactions show decisively that while there is considerable variability in individual responses (likely due to the ambiguity in the task), there is also a clear linguistic hierarchy in terms of tolerance towards measurement/counting. *Number of* phrases are less tolerant towards measurement than bare CNs (Exp3), which are in turn less tolerant towards measurement than MNs (Exp1). Conversely, *volume of* phrases are less tolerant towards counting than MNs (Exp4), which are less tolerant towards counting than CNs (Exp2). Obtaining these results was made possible by acknowledging that pragmatics may overrule linguistic preferences, hence testing for perceived ambiguity is the key for discovering non-salient strategies. Comparative strategies are shown to be inherently ambiguous with both MNs and CNs, though with a substantial role for the mass/count distinction in disambiguating them.

Figure 1: *more post*/*packages*

Figure 2: *more rock*/s *than clay*

[1] Barner & Snedeker 2005. Quantity judgments. *Cognition*. **[2]** Chierchia 1998. Plurality. In *Events & Grammar*. **[3]** Grimm &. Levin 2012. Who has more furniture? In *Mass/Count in Linguistics*. **[4]** Rothstein 2017. *Semantics for counting*. CUP. **[5]** Scontras et al. 2017. Who has more? *LSA*. **[6]** Snyder 2021. Counting. *L&P*. **[7]** Winter 2021. Mixed comparatives, *CSSP*, Paris.

'Negation-blind' N400 effect disappears when lexical priming is controlled

Daiki Asami, Chao Han, Jacob Burger, Deanna Dunlop, Yue Lu,
Effah Yahya M Morad, Chenyue Zhao, and Arild Hestvik
University of Delaware

Introduction. Prior ERP studies of truth-value and negation computation (Fischler et al. 1983; Palaz et al. 2020; among others) have argued for a classic two-step account of negation processing (Clark & Chase 1972). Their evidence comes from an interaction between sentence form (presence or absence of "not") and truth value in an N400 effect. Specifically, false affirmative (FA, 1b) sentences yielded a larger N400 compared to true affirmatives (TA, 1a)—the truth-sensitive N400 effect (Hagoort et al. 2004)—whereas the inverse was observed for negative sentences: true negative (TN, 2a) sentences caused a larger N400 than the false negative (FN, 2b) as if the N400 was 'blind' to negation and reflected only the truth value of the internal positive proposition (i.e., 'a robin is a tree').

(1) a. A robin is a bird. (TA) b. A robin is a tree. (FA)
(2) a. A robin is not a tree. (TN) b. A robin is not a bird. (FN)

In the two-step model of negation processing, when comprehending a negative sentence such as (2a), the meaning of the to-be-negated proposition, which is false, is computed in the first step, and then negation is applied to flip its truth value in the second step (Clark & Chase 1972). Under the assumption that the N400 is elicited during the first step, the negation-blind N400 effect follows.

However, the prior studies arguably contained a confound: the stimuli that generated the larger N400 contain no lexical priming relation between subject and object, whereas the control stimuli contained the lexical priming relation between subject and object. The N400 effect is known as an inverse index of priming: when a semantically related word pair (e.g., 'doctor' and 'nurse') is compared to an unrelated pair (e.g., 'car and 'nurse'), the primed word generates a reduced N400 compared to the unprimed word (Holcomb 1988). The subject and object in (1a/2b) are semantically related, but the subject and object in (2a/1b) where the truth-sensitive N400 effect was observed, are semantically unrelated, thereby providing an independent source of the observed N400 effect. The goal of the current study was to examine if the negation-blind N400 pattern persists even when this priming confound was removed. To this end, we conducted an ERP experiment with comparative constructions where the subject and object were unrelated in terms of animacy as well as semantic category (Table. 1). We predicted that if the previously observed interaction was unrelated to priming, the negation-blind N400 patten would replicate; if not, it would disappear. Our result was consistent with the second prediction.

Methods. 30 people participated in our ERP experiment with the 2x2 within-subject design, manipulating truth value (true vs. false) and sentence form (affirmative vs. negative). Each condition had 40 sentences. We also used 40 fillers (160 + 40 = 200 sentences).

	Sentence form	
Truth value	Affirmative	Negative
True	A tiger is bigger than a guitar.	A mouse is not bigger than a guitar.
False	A tiger is smaller than a guitar.	A mouse is not smaller than a guitar.

Table 1: Sample stimuli in the four conditions (two truth values × two sentence forms).

Each stimulus was visually presented in four chunks (e.g., A tiger / is / bigger than / a book) with 175ms duration and 800ms ISI. Participants made a speeded truth value judgment via button press at the object chunk. EEG was timelocked to the object with -200-to-1000ms epochs.

Behavioral Results. For accuracy, we observed main effects of truth value and sentence form: the false condition had higher accuracy than the true condition (87% vs. 83%, $F(1,29)=11$, $p<0.01$)

and the affirmative condition had higher accuracy than the negative condition (91.1 vs. 79.3%, $F(1,29)=93.5$, $p<.0001$). Truth value interacted with sentence form such that FN accuracy was higher than TN accuracy (82.8 vs. 75.7%, $F(1,29)=8.14$, $p<.01$). RT analysis revealed main effects of truth value and sentence form: the true condition was judged faster than the false condition (1342 vs 1390 ms, $F(1,29)=4.79$, $p<0.05$) and the negative condition took shorter to judge than the affirmative condition (1190 vs. 1542 ms, $F(1,29)=132.4$, $p<0.0001$). The results mirrored prior findings (Clark & Chase 1972; Carpenter & Just., 1976; Fischler et al. 1983; Palaz et al. 2020).

ERP Results. In looking for the N400, we used a data-driven sequential PCA technique (Dien, 2010, 2012) to identify the temporal and spatial components of the brain response to truth value. We used the two difference waves, false-minus-true for affirmatives and negatives, as inputs. Inspection of the resulting temporal factors corresponding to the difference between true and false sentences revealed no N400, but instead a late left-anterior negativity. We next used the factor loadings to constrain selection of a time window of 504-680ms and a left-anterior electrode cluster and calculated the mean voltage per cell and subject as dependent measures for a 2x2 repeated measure ANOVA. This revealed a main effect of negation ($F(1, 29) = 5.52$, $p = .026$), main effect of truth value ($F(1, 29) = 6.12$, $p = .020$), and interaction between the truth value and negation ($F(1, 29) = 5.53$, $p = .026$), driven by a greater difference for negatives. Figure 1 shows the mean waveforms for the regionalized channels:

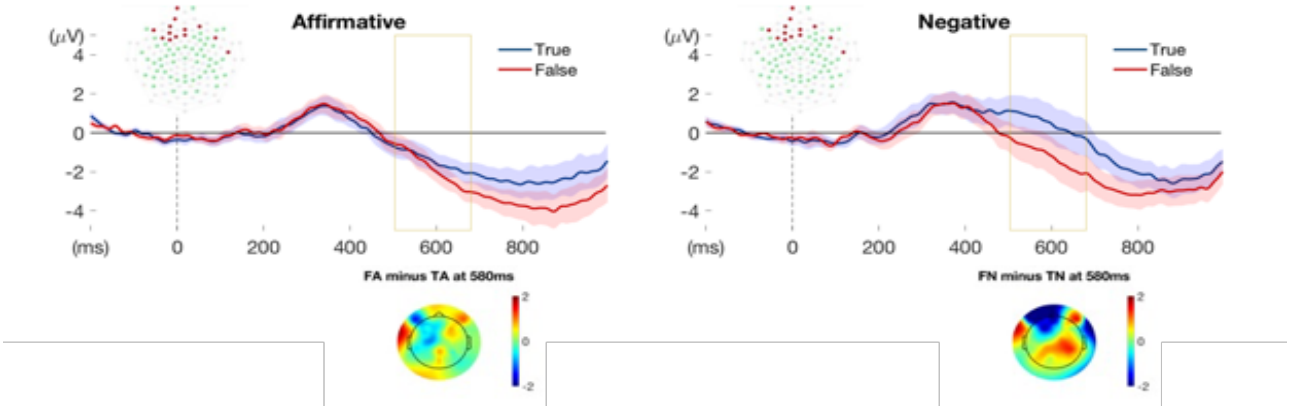


Fig. 1: Waveforms with 84% CI for main effect of truth; difference wave topoplot at peak latency

Discussion. The main finding is that when controlling for a lexical priming relation between subject and object, no N400 index to truth value was observed, and consequently no negation-blind N400. This suggests that previous N400 evidence for the two-step negation processing was wholly due to the lexical priming confound. Despite the lack of the N400 effect, we did observe a statistically significant brain response modulation by truth value, which suggests that the true ERP index of truth value computation is not the N400 but the LAN. This matches the findings of Hagoort et al., (2004) who identified the left inferior prefrontal cortex as being related to truth-value computation based on world knowledge. We attribute the relative lateness of this LAN to the relatively more difficult judgment task, which is seen by the longer RTs than those in previous studies (e.g., Fischler et al. 1983).

Selected References. Carpenter, P. A. & Just M.A (1975) in *Psychological Review*; Clark, H., & Chase, W. G. (1972) in *Cognitive Psychology*; Dien, J. (2010) in *Journal of Neuroscience Methods*; Dien, J. (2012) in *Developmental Neuropsychology*; Fischler, I. et al. (1983) in *Psychophysiology*; Hagoort, P. et al. (2004) in *Science*; Palaz, B. et al. (2020) in *Psychophysiology*

Coloring disjunction in child Romanian

Introduction Children have been argued to be more logical than adults in their interpretation of quantifiers, modals,[1] and disjunction.[2] However, recent studies suggest that children's performance may vary with the task: while children may find truth value judgments challenging, they appear more adult-like in act-out tasks,[3] ternary reward tasks,[4] felicity judgment tasks,[5,6] and coloring and erasing tasks,[7-9] the latter corresponding to more engaging tasks that allow children more freedom of action. We investigated disjunction in child Romanian using the Coloring Book Task (CBT),[10,11] used previously to investigate the acquisition of passives and binding,[10,11] adjunct control,[12-14] PP-modification,[15] and implicatures of quantifiers.[7-9] Importantly, the method has generally elicited more adult-like behavior from children. In our version of the task, children colored images based on their understanding of the disjunctive test sentences.

Current experiment Romanian children rarely interpret disjunction exclusively in TVJTs,[16,17] preferring inclusive or conjunctive interpretations (treating '(either) or' as meaning 'and'). We here use the CBT to determine whether children interpret *sau…sau* 'either…or' more exclusively in this task. We tested 34 5-year-old Romanian monolinguals and 40 adult controls. Participants were introduced to a puppet Bibi whose wishes they had to fulfill by coloring objects, erasing the color of objects, or taking no action. They saw displays of vehicles/fruits/shapes/ vegetables in which none, some, or all objects were colored (Figs.1-3).

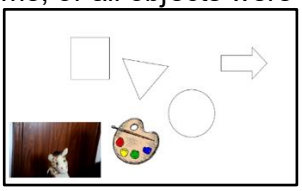


Fig.1 0-Object Scenario

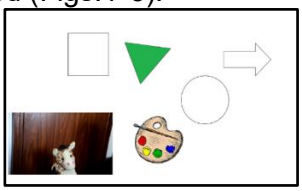


Fig.2 1-Object Scenario

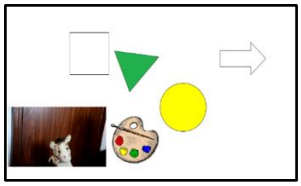


Fig.3 2-Object Scenario

They then heard a recorded statement left by Bibi on WhatsApp as in (1), and they had to fulfill her wish. The materials consisted of 6 warm-up statements balanced for action (coloring/erasing/ doing nothing), 36 critical sentences and 15 balanced fillers. The experiment tested disjunctive sentences (1a) in three scenarios: the 0-Object Scenario (containing no colored objects-see Fig.1), the 1-Object Scenario (containing 1 colored object-see Fig. 2), and the 2-Object Scenario (containing 2 colored objects-see Fig. 3), similarly to [3]. We also tested conjunctive and negative sentences as controls (1b,c) in these three scenarios.

(1) a. Bibi: Aş vrea sa aibă culoare sau triunghiul sau cercul.
'I would like either the triangle or the circle to have color.'
b. Bibi: Aş vrea sa aibă culoare triunghiul și cercul.
'I would like the triangle and the circle to have color.'
c. Bibi: Aş vrea să nu aibă culoare nici triunghiul nici cercul.
'I would like neither the triangle nor the circle to have color.'

For disjunctive *sau…sau* 'either…or' statements, we expected adults to color one object in the 0-Object Scenario, do nothing in the 1-Object Scenario, and erase the color of an object in the 2-Object Scenario, while we expected more variability in children's answers given previously reported inclusive/conjunctive behavior (Table 1). Nevertheless, given the CBT's success in eliciting adult-like performance, we expected some proportion of exclusive responses.

Table 1. Predicted responses for disjunctive statements per participant type in the three scenarios

Scenario *Initial Situation*	Inclusive Participants *A or B, possibly (A and B)*	Exclusive Participants *(A or B) but not (A and B)*	Conjunctive Participants *A and B*
0-Obj	Color 1 or 2 objects	Color 1 object	Color 2 objects
1-Obj	Do nothing or color 2nd object	Do nothing	Color 1 object
2-Obj	Do nothing	Erase 1 object	Do nothing

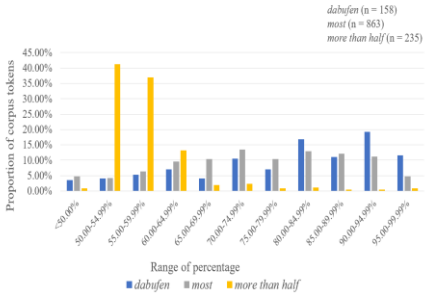
Results: Adults generally behaved as predicted, i.e., they consistently preferred exclusive interpretations. Turning to children, we observe that they were close to adult-like on the conjunctive controls (89%) and the negative controls (83.3%). For the disjunctive statements, however, more non-adult-like responses were observed overall. Importantly, there was variation depending on scenario. In the 0-Object Scenario, 86% of children's responses were adult-like (coloring one object); the remaining responses involved coloring two objects instead of one. In the 1-Object Scenario, 52.2% of responses were adult-like (doing nothing); the remaining responses involved coloring a second object. In the 2-Object Scenario, 44.11% of responses were adult-like (erasing the color of one object); the remaining responses involved leaving both objects colored. An individual analysis revealed that 10/34 children were consistently exclusive, 3/34 were consistently conjunctive, and the rest showed mixed (inclusive/conjunctive/exclusive) behavior.
Discussion: Importantly, we see that children seem to be more adult-like with disjunction in this task compared to previous studies which used TVJTs.[2,16,17] For example, just like adults, almost all children colored an object in the 0-Object Scenario when hearing a disjunctive statement. However, there were many non-adult-like responses in the 1-Object and 2-Object Scenarios: in the 1-Object Scenario, around half the children chose to color in a second object, and in the 2-Object Scenario, around half of the children chose to do nothing. Based on the relatively high accuracy on the controls, we assume that children's coloring responses essentially reflect their linguistic understanding of disjunction in line with we call a *Meaning in Action Principle* (Make the sentence true according to the semantic/pragmatic meaning of disjunction). Interestingly, while most children colored one object in the 0-Object Scenario, they varied in their behavior in the other scenarios: they would sometimes color nothing or color one more object in the 1-Object Scenario, and they would erase one object or simply leave the two objects colored in the 2-Object Scenario. We take this behavior to suggest that some children may be at a developmental stage where they oscillate between inclusive and exclusive interpretations for the complex disjunction *sau...sau*, in contrast with adults, who consistently favor the exclusive interpretation. Additionally, we argue that our results cannot be accounted for on non-linguistic grounds. We consider two possible non-linguistic cognitive constraints, which we term (i) *Maximal Preference*, whereby more colored objects are to be preferred (as a strategy for maximizing Bibi's happiness), and (ii) *Minimal Effort Preference*, whereby the least effort is employed as a means of satisfying the request. Teasing apart the role of non-linguistic preferences is difficult when they go in a similar direction with *Meaning in Action*: in the 0-Object Scenario and in the 1-Object Scenario, having only one colored object is not only in line with inclusive/exclusive meanings, but it is also in line with *Minimal Effort.* However, in the 2-Object Scenario, the adult-like answer (to erase the color of one object) involves both more effort and fewer colored objects than the non-adult-like answer (to do nothing), i.e. it clashes both with *Minimal Effort* and *Maximal Preference*, yet, even in this condition, a non-trivial proportion of children provided exclusive answers.
Conclusion The present findings support the use of the CBT as a method of eliciting adult-like interpretations in children. Unlike the TVJT, which may simply show that children are more pragmatically tolerant than adults,[4] the CBT is a preference-based task, combining linguistic comprehension with non-linguistic production. In line with previous studies,[7-13] preference-based tasks like the CBT elicit more adult-like responses from children. Our findings also suggest that at least some children in this age range can interpret disjunction exclusively – contra many findings from TVJT-based studies,[16,17] which show that Romanian children tend to be inclusive in their comprehension of *sau...sau*.
References [1] Noveck 2001. [2] Tieu et al. 2017. [3] Pouscoulous et al. 2007. [4] Katsos & Bishop 2011. [5] Chierchia et al. 2001. [6] Foppolo, Guasti & Chierchia 2012. [7] Bleotu 2018. [8] Bleotu 2019. [9] Nuninga et al. 2023. [10] Zuckerman et al. 2016. [11] Zuckerman & Pinto 2018. [12] Gerard et al. 2017. [13] Gerard et al. 2018. [14] Gerard & Lidz 2018. [15] Hall & Pérez-Leroux 2022. [16] Bleotu et al. 2023a. [17] Bleotu et al. 2023b.

On the Interpretational Flexibility of Mandarin Chinese *Dabufen*

This paper probes into the interpretational mechanism of Mandarin Chinese proportional quantifier *dabufen*. The existing studies on *dabufen* (e.g., Lin 1998) treat the expression as an equivalent of the English *most* and assign it a conventional GQT definition which ensures its proportional interpretation of 'above 50%'. However, *dabufen* differs from *most* in terms of its syntactic distribution, semantic interpretation and internal morphological makeup, which prompts us to propose that it encodes a weaker adjectival semantics, meaning sufficiently large parts as compared to a contextually determined neutral range. After pinning down the semantics of *dabufen*, we also conduct a truth value judgement experiment and perform a clustering analysis to uncover the manifestation of the weaker interpretational mechanism among native speakers.

Syntactic difference between *dabufen* and *most*: Syntactically, *dabufen* can be preceded by demonstrative determiners and free choice *renhe*, both of which are typical positions hosting predicates. Also, it can occur after adjectival modifiers, after the predicative copula *shi*, or in the scope of the existential *you*. To a large extent, the distribution of *dabufen*-N patterns like typical weak quantifiers, and forms a sharp contrast with that of strong quantifiers like English *most* Chinese universal quantifiers headed by *mei* and *suoyou* .

Semantic difference between *dabufen* and *most*: Our corpus search results in Fig 1. shows the parallel between *dabufen* and *most*, both of which mainly represent percentages between 50% and nearly 100%. However, a close look at the cases where *dabufen* and *most* express proportions below 50% uncovers the subtle difference: Both *most* and *dabufen* allow the NP-external relative superlative reading, which explicitly requires that the portion they associate with be the largest as compared to other alternatives (Hackl 2009), as in (2); however, in still other cases, the use of *dabufen* does not exert such a strict requirement and can simply mean a sufficiently high proportion, as in (2).



(1) As to the size of the institutions, most developmental students in the sample were enrolled at small-size (28.7%), followed by medium-size (26.6%), large size (25.0%), and extra-large colleges (19.6%). (COCA)

(2) Context: 42% of the interviewee choose to improve their health through diet and exercise.
Zhe shuoming dabufen-ren-de-guannian shi zhengque de. (BCC)
this indicate DABUFEN-person-DE-idea BE correct DE
'This indicates that DABUFEN people's ideas of health are right.'

Fig 1 Proportional ranges of *dabufen* and *most*

Furthermore, *dabufen*, but not *most*, can be modified by indefinite *yi* 'one' which marks indefiniteness and variability, and in such cases, it is more common for *dabufen* to express proportions below 50%; *dabufen* is also more susceptible to the influence of contextual regulators like *xiangduieryan* 'relatively speaking' and can diverge from its default interpretation of 'above 50%' to refer to lower percentages.

Internal semantic composition of *dabufen*: In terms of morphological makeup, *Dabufen* can be dissected into *da* 'large', a gradable adjective, and *bufen* 'part', and it can be further modified by degree modifiers like *geng* 'more', *zui* 'most' and *ji* 'extremely'. In this light, we opt for an

adjectival semantic analysis in the spirit of Solt (2009, 2016) to characterize the meaning of *dabufen*. Essentially, *da* is a gradable adjective encoding proportions as its dimension of measurement, as in (3a), and the degree argument can be bound by operators like POS and -est. When bound by POS, the degree interval expressed picked out by *da* is compared with a contextually-determined neutral range, which is normally set to the mid-point of the proportional scale and derives the meaning of 50%. Yet, with appropriate contextual support, the neutral range can be scaled down to lower points which derives the reading of 'a high yet below 50% proportion'. For instance, when modified by indefinite 'one' which indicates variability, or the possible existence of more than one large portion, the neutral range of comparison tends to be set lower. When bound by an implicit -est operator, *dabufen* can express a relative superlative reading, meaning 'the largest part'. (3b) shows the derivation of 'dabufen people came', where *bufen* is formalized as an abstract partition of a collection of entities.

(3) a. $[\![da]\!] = \lambda N \lambda d \lambda x [N(x) \wedge \mu(x)/\ \mu \oplus N \geq d]$

b. $[\![da\text{-}bufen\text{-}person\text{-}came]\!] = \lambda d \lambda x [\ x \leq \oplus person \wedge \mu(x)/\ \mu \oplus person \geq d \wedge *came(x)]$

Existential closure: $\lambda d \exists x [\ x \leq \oplus person \wedge \mu(x)/\ \mu \oplus person \geq d \wedge *came(x)]$

$[\![POS]\!] = \lambda I\ \forall\ d \in Ns\ [I(d)]$

$[\![POS\ da\text{-}bufen\text{-}person\text{-}came\ ]\!] = \forall\ d \in Ns\ \exists x [\ x \leq \oplus person \wedge \mu(x)/\ \mu \oplus person \geq d \wedge came(x)]$

Truth value judgement experiment: An experiment was conducted to investigate the availability of the 'relative superlative', 'sufficiently large' and 'more then half' readings of *dabufen* by carefully controlling the proportional information in the context, and we conclude that there are two populations of native speakers of systematic patterns of interpretation by performing a clustering analysis: One (n=62) hardly accepts the superlative interpretation or the sufficiently large interpretation, and one (n=71) endorses these weaker readings. We conjecture that this population split might reflect individual differences in terms of mental calculation strategies.

References

Hackl, M. 2009. On the grammar and processing of proportional quantifiers: *most* versus *more than half*. *Natural Language Semantics*, 17, 63-98.

Lin, J.-W. 1998. Distributivity in Chinese and its implication. *Natural Language Semantics* 6. 201-243.

Solt, S. 2009. The Semantics of Adjectives of Quantity. City University of New York dissertation.

Solt, S. 2016. On measurement and quantification: The case of *most* and *more than half*. *Language*, 92, 65-100.

Using bounds set by modals to investigate the status of partial objects and count nouns

Previous work has revealed a surprising pattern: faced with a display such as Figure 1 and asked to 'count the forks', children, unlike adults, treat discrete fork-parts and whole forks on par, counting 6 (Shipley & Shepperson 1990, Brooks et al. 2011, Srinivasan et al. 2013, a.o). In recent work, Syrett & Aravind (2022) argue that children's treatment of partial objects is consistent with the underlying semantics for count nouns, which are **vague** and **context-sensitive.** Where children and adults diverge is in their ability to restrict a count noun's application in a given context. Supporting this hypothesis, they showed that preschoolers are less likely to allow a count noun like 'fork' to pick out a partial form if the speaker specifies a goal of using the fork for eating. However, Syrett & Aravind employed tasks that probed categorization – i.e., whether or not a count noun like 'fork' can apply to an object – and did not highlight counting or quantification. Thus, it remains an open question whether contextual factors can influence how children and adults resolve the ambiguous status of partial objects in a numerically-oriented task. The current research seeks to fill this gap.

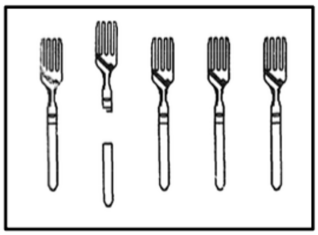


Figure 1: Display of forks from Shipley & Shepperson (1990).

Background and hypotheses: We manipulated contextual requirements with a goal-oriented introduction phase, followed by a modal statement, as in (1)-(2).

(1) To get a star, you have to have three balls. *(universal modal; lower bound, 'at least')*

(2) To get a star, you're allowed to have three balls. *(existential modal; upper bound, 'at most')*

The difference in these modals lies in how they trigger varying bounding conditions for numerals in their scope. Universal modals induce lower bound interpretation of numerals: the *minimum* number of balls required to meet the requirement is 3, and surpassing the lower bound is acceptable. In contrast, existential modals induce an upper bound: the *maximum* number of balls allowed is 3, though deviation below this upper limit is permissible. We manipulated whether the set of objects on display counting towards these limits included a partial object. See Figure 2. The key question is whether the partial object is treated as meeting or exceeding the limit. If so, given (1), the lower limit is met; otherwise, it is not.

Figure 2: Sample of *have to* stimuli with 2.5 objects

Experiment: Adults (N=73) were randomly assigned to two between-subject modal groups (*have to* or *allowed to*). Children (N=21/30 run, mean age 4;10) participated in the *have to* variant of the study. (Data collection with *allowed to* is ongoing.) Both groups were shown characters possessing a combination of whole and partial objects alongside sentences such as those in (1) or (2), and asked, "Is what they have okay?" The child task was set up as a counting game among aliens (see Figure 2), in which they were asked to assign a calculator ("no") or star ("yes") for each trial. Otherwise, the design was identical to adults. Trial types (see Table 1) featured controls probing the availability of bounded readings, critical items with whole and partial objects, and a strictly whole object comparison set.

Trial Type	Number of Objects
Control	2 whole
	4 whole
Critical	2 whole, 1 partial
	3 whole, 1 partial
	3 whole

Table 1: Experiment trial types

Results: Figure 3. Adults patterned as expected, accepting sets of 3 whole objects and greater in the *have to* condition (reflecting an 'at least' reading), and accepting sets of 3 whole objects and fewer in the *allowed to* condition. A partial object *did not* help meet the lower limit for *have to* (2.5 vs 3: β=-6.804, p<.001), yet incurred a penalty for exceeding the upper limit for *allowed to* (3 v. 3.5: β=-4.94, p<.001). Taken together, adults' behavior on partial-object trials suggests that they considered these objects as affecting the numerosity of the counted set, but in a more granular way: *for adults, a partial ball increases the size of the set by a fraction less than 1***.** Children in the *have to* condition differed from adults in two ways. They largely did not accept 2-whole or 4-whole trials, reflecting an exact numerical preference. Consistent with this preference, they were also significantly less likely to accept 3-whole-1-partial scenarios than 3-whole ones (β=-4.37, p<0.01). Crucially, they did not distinguish between 2-whole-1-partial and 3-whole scenarios (β=-1.74, p=.11). Thus, a partial object and a whole object have comparable status in helping to meet the required lower bound of the modal: *for children, a partial ball increases the size of the set by 1***.**

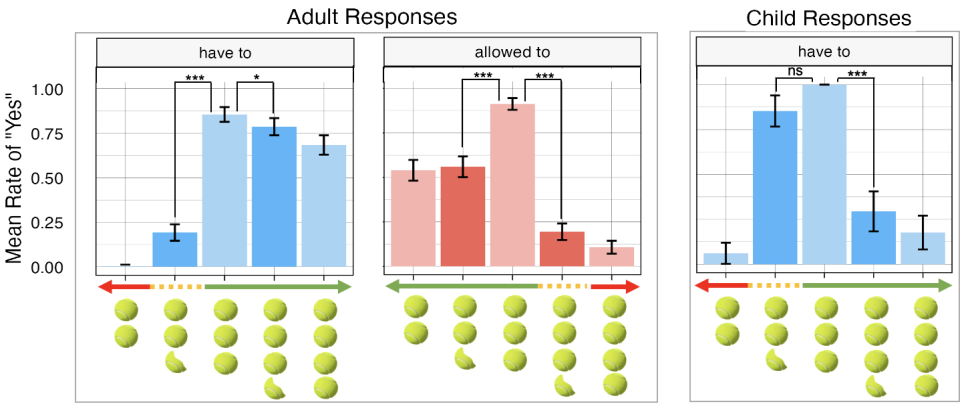


Figure 3: Mean "Yes" responses (+/-1 SEM) by modal and trial type for each population

Discussion: Consistent with previous work, children treated partial objects on par with wholes when counting instances of a count noun. This behavior is reinforced by their strong preference for 'exact' interpretation of numerals, well-attested in earlier work (e.g., Papafragou & Musolino 2003; Musolino 2004). Crucially, for children, partial objects help satisfy this exact interpretation. Adults opt for an 'at least' reading with 'have to' and an 'at most' reading with 'allowed to', but in neither case did they flexibly shift their criteria for a noun's application to let partial objects meet limits set by modals. Instead, they employed a more fine-grained counting system, quantifying a partial object as a fractional portion. Thus, in a numerically-oriented task, the child-adult difference is again reinforced. We consider two possibilities consistent with these results. One ties the child-adult distinction to differences in the measurement scales accessible to the two populations: unlike adults, children are unable to count and measure in fractional quantities. Another possibility is that differences in recruiting contextual information underlies the child-adult difference in numerical tasks as well, more in line with Syrett and Aravind's hypothesis. Contexts where object *quantity* matters, rather than object *kind*, lead adults to opt for a more granular measurement scale; children, despite in principle having access to such scales, fail to do so.

Selected References: Brooks, N. et al. (2011). Piecing together numerical language. *Dev. Science;* Shipley, E. & Shepperson, B. (1990). Countable entities: Developmental changes. *Cognition*; Srinivasan, M. et al. (2013). Sortal concepts and pragmatic inference in children's early quantification of objects; *Cognitive Psychology;* Syrett, K. & Aravind, A. (2022). Context sensitivity and the semantics of count nouns. *Journal of Child Language.*

GRADED CAUSATIVES

Introduction. Semanticists have long been interested in how concepts present in causal relationships are lexicalized (C&H'15; B&S'21; N&S'22; L'00; S'11; S'76). The predominant approach to analyzing verbs of causing has been to argue that they convey some version of *sufficiency*, which is measured given parameters of a causal situation. Here, we provide experimental evidence for a differentiating and multi-faceted semantics of three causing verbs using explicitly-defined causal models, which represent how participants reason about the stimuli. This approach enables us to quantify concepts including sufficiency and use them as predictors.

Contribution. We focus on the constructions *C caused/made/forced E* and argue that **H1.** *cause*, *make*, and *force* are in an asymmetric entailment relation, and that **H2**. this entailment relation is structured not by sufficiency, intentionality, or alternatives alone, but by an interaction of these three. Our experiment uses tic-tac-toe (ttt) sequences defined using structural causal models (SCMs; P'09). The use of SCMs enable us to make predictions about verb selection by defining probability distributions across counterfactual scenarios.

Possible scales. We postulate that graded causatives have a semantics built around threshold values on a continuous scale, similar to gradable adjectives. We consider three measures that are relevant features of causal relationships: ALT, INT, and SUF. Firstly, previous work (F'69; P'00) argues that *the number of alternative actions available to the causee* can distinguish between causal relationships in which the causer is (or is not) culpable for the action taken by the causee. This feature is also of interest for differentiating the semantics of causal verbs, since it provides the contrast in (1) *The child was* {*made/?forced*} *to get into the car, although she could've chosen to do otherwise*. So, our first measure ALT quantifies the number of alternative actions available to the causee. This postulates that w.r.t. (1), the threshold ALT for *force* is less than the threshold ALT for *make*. In three-state ttt sequences as in Fig. (A), ALT is measured as the number of empty squares in the third state. So, $\text{ALT}(Y_1) = 5$. Secondly, the notion of *intention* is also strongly related to alternatives (W&M'06) and relevant for distinguishing causal situations (C'18). For example, consider that the pirate's intention is what distinguishes (2) *The pirate* {*intentionally/?accidentally*} *forced the prisoner down the plank.* Building on this intuition, our second model (INT) is based on the 'degree of intention' proposed by H&K-W'18 (see their paper for details), which is roughly the probability of reaching the goal state given the current action versus given alternative actions. This is why (3) *Player O placing at location 2* is more intentional in Z_1 than Z_2. Specifically, any alternative to *Player O placing at location 2* in (A), e.g. *Player O placing at location 5*, would make it highly probable that *Player X* wins at the next time-step, thereby largely decreasing the probability of reaching the goal-state of *Player O*. The same is not true for (B). Thirdly, the notion of *causal sufficiency* has been well-represented in previous literature on causal verb selection – G'23 argues that *cause* entails local sufficiency, while L&N'18 and N&L'20 argue that *make* conveys (non-probabilistic) causal sufficiency. Intuitively, this distinguishes between causing and enabling verbs – in (4a/b) *The pirate* {*made/let*} *the prisoner walk down the plank*, we can say that likely *the prisoner walks down the plank* in (4a) while it is less clear whether this result comes about in (4b). Thus, our third model (SUF) is P'18's 'probability of sufficiency', which is defined as the probability that the event $X = 1$ would be sufficient to produce outcome $Y = 1$. Descriptively, SUF denotes the capacity of C to produce the outcome E in situations where the agent of C did some action other than the one encoded in C. Intuitively, *Player X placing at location 1* in Y_1 is more sufficient in bringing about *Player O placing at location 2*, than *Player X placing at location 7* in Y_2 is for bringing about the same. This is because in sequences where settings Y_1 and Y_2 don't result in *Player O placing at location 2* at the next time-step, it is more likely that Y_1 will *eventually* lead to *Player O placing at location 2* to block X's clear three-in-a-row than Y_2, which does not present that danger to *Player O*. To conclude our measurements, observe that our definitions have been applied to ttt sequences, which can be defined as partial setting of a SCM. This means that given some setting of variables in a SCM, we

can apply functions ALT, INT, and SUF that output a numerical value. Minimally, a probabilistic SCM has a set of exogenous variables with an associated probability distribution, a set of endogenous variables, and a set of deterministic functions that assigns a value to each exogenous variable given values of some subset of exogenous and endogenous variables (see C-et-al'18 for technical detail). This framework can encode any causal process as a directed acyclic graph (DAG). Thus, we choose the game of ttt as experimental stimuli, since an entire game-tree can be efficiently stored as a DAG (and consequently defined as a SCM). In this way, an endogenous board-state variable is stored as a conjunction of location-demarcation assignments. So, given a statement such as *Player X placing at location 3 made Player O place at location 5*, we can measure the number of possible alternative actions that Player O could have taken besides placing at location 5, the degree of intention of that Player X had for bringing about the event of Player O placing at location 5, and the probability that Player X placing at location 3 would bring about Player O placing at location 5.

Experiment. Our stimuli consist of 30 two-frame ttt sequences, filtered from 21 full games to represent the range of possible ALT, INT, and SUF values. Participants were asked to rate whether sentences such as *Player X forced Player O to place at location 3* are accurate in describing the stimulus (see example in Fig. 2). We recruited 109 L1 English participants, of which 19 were excluded for failing attention check(s).

Results/Analysis. We find that holding the set of stimuli constant, participants were less likely to determine *made* than *caused* as accurate in describing a scenario, and less likely to determine *forced* than *made* as accurate (Fig. 3). This supports **H1**, since the semantic interpretations of weaker predicates are entailed by the use of stronger ones (M-et-al'10). Regarding **H2**, we fit (I) an initial Bayesian linear regression using participant judgements as the outcome variable and model such using a Bernoulli distribution. The predictors include the `verb` used in the sentence presented to participants, the `ALT`, `INT`, and `SUF` value of the associated stimuli as fixed effects, as well as their interactions. The results (full model results in Tab. 1) provide evidence that besides the different levels of `verb`, `SUF` and the three-way interaction of `ALT:INT:SUF` has a non-zero effect on the response variable. We then fit a second regression (II) that predicts judgements using only `verb` and `SUF`. We find that $WAIC(I) = 1941.93$ ($SE = 32.24$), $WAIC(II) = 2003.09$ ($SE = 28.18$), and $WAIC(I) - WAIC(II) = -61.15$ ($SE = 17.24$), indicating that (I) is the better fit, and that our results are better explained by including all three predictors and their interactions, than by SUF alone. Next, we fit follow-up regressions similar to (I), except without `INT` and all of its interactions (III), and without `ALT` and all of its interactions (IV). Comparing (I) to (III), we get $WAIC(III) = 1961.96$ (SE : 30.69) and $WAIC(I) - WAIC(III) = -20.02$ ($SE = 10.42$), indicating that since (III) does reliably worse, the predictor `INT` does matter despite including 0 in its CrI in regression (I). Comparing (I) to (IV), we get $WAIC(IV) = 2001.33$ ($SE = 28.57$) and $WAIC(I) - WAIC(IV) = -59.40$ ($SE = 16.79$), indicating that since (IV) does reliably worse, the predictor `ALT` also matters (despite also including 0 in its CrI in regression (I)). To conclude, our Bayesian analysis demonstrate that all three features – ALT, INT, and SUF – have reliable effects on participant judgements of *cause*, *make*, and *force*. This work demonstrates that these causatives not only encode information about sufficiency, but also intention and possible alternative actions.

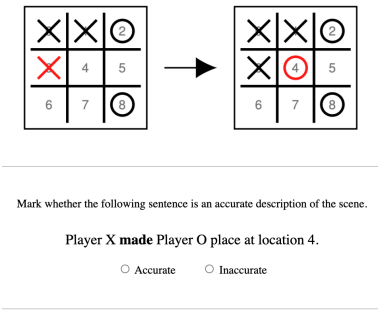


FIGURE 2. Example of experiment question.

	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	-3.96	0.75	-5.42	-2.49
verbmade	-0.35	0.13	-0.61	-0.09
verbforced	-0.62	0.14	-0.90	-0.36
SUF	5.97	1.48	3.10	8.88
INT	0.19	1.92	-3.60	3.85
ALT	0.32	0.19	-0.07	0.68
SUF:INT	-4.97	3.49	-11.74	1.89
SUF:ALT	0.08	0.50	-0.90	1.07
INT:ALT	-0.25	0.49	-1.21	0.71
SUF:INT:ALT	2.72	1.17	0.34	5.02

TABLE 1. Full model results for (I).

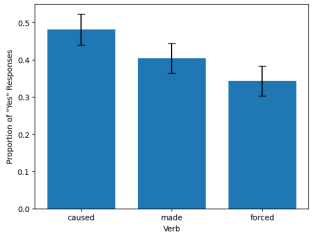


FIGURE 3. Proportion of "Yes" w/ 95% CIs.

Talking about Distributivity: How Cognitive Factors Influence Children's Language

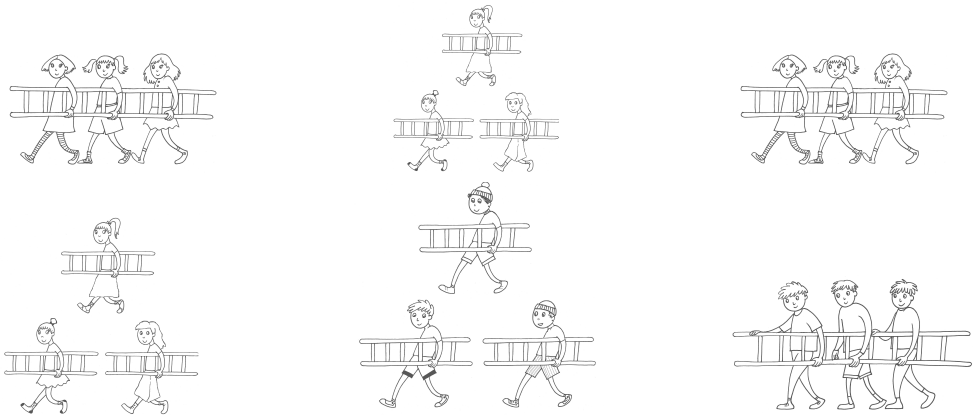
Plural sets of entities are represented as groups or collections of individuals: a sentence out of context (e.g., "The girls are carrying a ladder") receives a distributive reading if the predicate refers to the atomic members or a collective reading if it refers to the whole plurality. Previous accounts suggest that the distributive representation includes an additional semantic operator (e.g., 1). Comprehension experiments show that adults interpret an ambiguous sentence as collective, hinting at easier processing costs (2). However, children accept the distributive reading more often than adults (e.g., 3), casting doubts on its presumed greater difficulty. The current study investigates these interpretations in a novel way, by comparing the same group of preschoolers in both comprehension and production. In the idea that language could be a mirror of the mind (4), we study how children describe distributive and collective scenes to explore whether the two structures differ in complexity. Furthermore, we investigate whether cognitive factors, such as the ability to take the other's perspective, may influence children's performance in the linguistic tasks.

We tested 23 Italian monolingual children (10 females; age in months *M*=68.81, range=64–76). In the first session, they participated in a production task: they saw 24 (18 experimental) trials displaying two images of transitive actions and described them. Based on the within-participants Contrast Type factor, the conditions were *mixed* (distributive vs. collective image), *distributive* (two distributive) and *collective* (two collective; Figure 1). Participants provided two descriptions, one per image, and we coded each trial as marked if at least one contained a collective or distributive marker (e.g., *insieme*, "together", or *ciascuno*, "each"). In the second session, children performed the Dimensional Change Card Sort (5), testing the executive function of shifting, and the Perspective Taking task (adapted from 6), testing the ability to switch quickly from their perspective to another one. Children saw a character in a room and judged a sentence describing how many dots they or the character saw on the walls; the two perspectives might differ (inconsistent trials, seeFigure 2). Lastly, children performed the Raven matrices as a measure of nonverbal reasoning and a linguistic comprehension task: they had to choose between a collective and a distributive image while listening to sentences ambiguous or marked for distributivity or collectivity.

In the production task, a mixed effects logistic regression on Marking, with Contrast Type as the fixed effect and the participant as the random intercept, revealed that the mixed condition had more marked descriptions than both the collective ($p<.001$) and the distributive one ($p<.01$). Children showed a very low tendency to produce linguistic marking (*M*=8.3%): they expressed more markers in the mixed condition (*M*=17%), followed by the collective (*M*=6%) and the distributive one (*M*=2%) (Figure 3). In the comprehension task, they were capable of correctly understanding the collective (accuracy *M*=93%) and distributive (*M*=86%) sentences; in the ambiguous condition, they showed a higher preference for the collective images (*M*=93%). From a cognitive point of view, the percentage of linguistic marking in the production task did not significantly correlate with the shifting or the perspective-taking score; still, it correlated positively with the Raven matrices (r=0.4, $p<.05$). Instead, by looking at the comprehension task, a correlation (r=-0.4, $p<.05$) between the egocentric bias in the perspective taking and the interpretation of the ambiguity emerged.

Children were generally not fully sensitive to the necessity for expressing markers disambiguating the two readings. Nevertheless, as expected, they produced more markers when the contrast was explicit. Children at this age are likely too young to produce these markers, even though they clearly understand them. In the comprehension task, they preferred the collective interpretation of an ambiguous sentence; this is in contrast with previous studies, but ours presented both ambiguous sentences and distributive or collective sentences in trials randomly ordered: children might have benefited from the contrast and reached a tendency similar to the adults', who consider the collective reading as the default one. Finally, we found that some cognitive factors may play a role in comprehending these linguistic structures: children who were more ahead in cognitive devel-

opment produced more linguistic markers overall. Furthermore, the more the participants were anchored to an egocentric bias, the more they chose the collective image; the more they took the other's perspective, the more they chose the distributive image. Hence, the capacity to shift quickly from different perspectives may influence linguistic processing, and good perspective-taking abilities may reverse the preferred interpretation. However, this ability should be fully developed in adults, but still, they prefer the collective reading. We will have more reliable conclusions once we finish the current data collection on older children (7 years of age) and adults.



(a) (b) (c)

Figure 1: Example conditions (in vertical): a) mixed, b) distributive, c) collective.

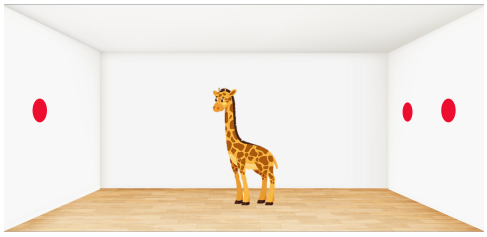


Figure 2: Example trial in the PT task.

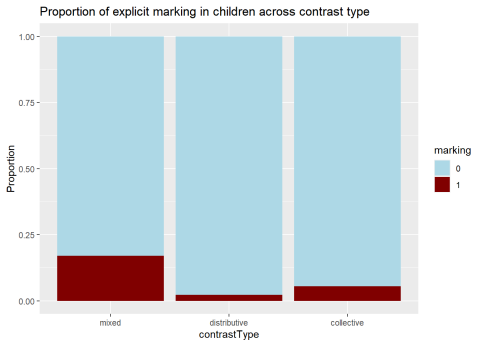


Figure 3: Proportion of linguistic marking.

References

[1] L. Champollion, "Covert distributivity in algebraic event semantics," *Semantics and Pragmatics*, vol. 9, pp. 15–1, 2016.

[2] L. Frazier, J. M. Pacht, and K. Rayner, "Taking on semantic commitments, ii: collective versus distributive readings," *Cognition*, vol. 70, pp. 87–104, 1999.

[3] K. Syrett and J. Musolino, "Collectivity, distributivity, and the interpretation of plural numerical expressions in child and adult language," *Language acquisition*, vol. 20, 10 2013.

[4] M. T. Guasti, A. Alexiadou, and U. Sauerland, "Undercompression errors as evidence for conceptual primitives," *Frontiers in Psychology*, vol. 14, 2023.

[5] P. Zelazo, "The dimensional change card sort (dccs): A method of assessing executive function in children," *Nature protocols*, vol. 1, pp. 297–301, 02 2006.

[6] L. M. Sacheli, E. Arcangeli, D. Carioti, S. Butterfill, and M. Berlingeri, "Taking apart what brings us together: The role of action prediction, perspective-taking, and theory of mind in joint action," *Quarterly Journal of Experimental Psychology*, vol. 75, no. 7, pp. 1228–1243, 2022. PMID: 34609238.

From words to memory: Evidence of language guiding motion event reconstruction

Two primary verb categories exist in human languages: manner and path. Manner verbs describe how a subject moves (e.g., *shoot* and *swim*), whereas path verbs indicate the direction of movement (e.g., *enter* and *rise*) [1]. Languages are categorized as manner or path based on the predominant verb class, resulting in manner languages (e.g., English and German) and path languages (e.g., Turkish and Spanish). These typological variations in language influence non-linguistic perception and memory in different ways [2, 3]. Prior research shows that language usage influences manner and path information prioritization. Linguistic production data supports this idea: individuals demonstrate a preference for verbs in the major verb category of one's language [e.g., 4, 5, 6, 7]. Evidence from gaze data, which shows that individuals first attend to the aspect of motion encoded most frequently in their language when preparing to speak [e.g., 5, 6] also advocates for language effects on aspect saliency. The current study had two main goals. First, it aimed to test whether language continues to affect manner versus path saliency when encoding from language to an internal representation. Past studies investigating motion saliency generally used a paradigm in which participants viewed depictions of events and then linguistically encoded them [e.g., 4, 5, 6, 7]. We designed a new paradigm in which participants read a linguistic event and then recalled the event from memory during an image-selection task, therefore reversing the classic paradigm. If one's language experience affects the saliency of different aspects of motion, the aspect of motion encoded by a language's majority verb class should be recalled by the speakers of that language (e.g., manner language speakers will find manner of motion to be more salient than the path of motion). Because our paradigm requires the target aspect of motion to be held in memory, observing differences in saliency between the language groups would indicate that language influences recall in addition to online processing. The second goal was to explore how the type of sentential element and the order of presenting manner and path information influenced the saliency of the motion aspects. Past studies focused on aspects of motion encoded in verbs, but other sentential elements known as modifiers also encode manner and path (e.g., adjectives, adverbs, prepositions) [1]. Little work has been done on how the location of motion information and sentential element affect cognition. The new linguistic-to-visual paradigm allowed us to test this question.

Procedure & materials. This experiment was conducted as an online survey. English monolinguals (N = 63) and Spanish-English bilinguals (N = 21, data collection is ongoing) participated in a linguistic encoding task followed by a forced-choice memory task. Participants completed four blocks. In each block, participants read six English paragraph vignettes with an embedded target event phrase and then completed six memory questions in which they selected the image that best corresponded to the target phrase events (Table 1; Figure 1).

Results & discussion. Memory task responses were analyzed using logistic mixed effects models and revealed significant interaction effects between language group and aspect of motion recalled ($\beta_{interaction}$ = -.94, p < .01). Further significant effects of the target phrase condition on image selection were revealed by the model when comparing the path with path+manner modifier conditions ($\beta_{verb\text{-}type}$ = -1.36, p <.001) and the two modifier conditions ($\beta_{verb\text{-}type}$ = -1.3, p < .001). English monolinguals selected more manner images after reading both the manner-framed and path-framed phrases, whereas Spanish-English bilinguals selected more manner images after reading manner-framed events but more path images after path-framed events (Figure 2). When presented with extra path or manner information via the modifiers, both monolinguals and bilinguals selected manner images when manner information was present in the phrase. This pattern of results suggests that the manner of motion was overall more salient than the path of motion. Taken together, our results support the idea that typological variance gives rise to differences in memory for motion events [2, 3]. These findings align with prior studies in demonstrating that language type affects aspects of motion saliency

and verb selection [e.g., 4, 5, 6, 7] and further highlight the interplay between language and perception. In addition, greater path information saliency in manner- and path-language bilinguals compared to manner language monolinguals—even when engaged with a manner language—patterned consistently with the hypothesis that L1 and L2 language systems are intertwined rather than independent [7]. Furthermore, regardless of the presentation order, sentential element type, and participant's language experience, manner was more salient than the path of motion. One potential explanation for this could be because the manner of motion is closer to the agent than the path of motion in that the agent performs the manner whereas the path is external to the agent. Finally, the current findings demonstrate the validity of our new paradigm by corroborating the results of past studies. Through this paradigm, researchers can reach more linguistically diverse populations to further the understanding of the interplay between language and perception.

Condition	Target phrase	Paragraph Vignette
Manner	A bunny **leapt** for the carrots.	The parents put the harvested carrots in a neat stack. While the parents had their backs to the pile of carrots, **[ ]**. The neighbor's car backfiered. The startled bunny decided to look for food elsewhere and the carrots were saved.
Path	A bunny **headed** for the carrots.	
Path + manner modifier	A bunny headed for the carrots **energetically.**	
Manner + path modifier	A bunny leapt **directly** for the carrots.	

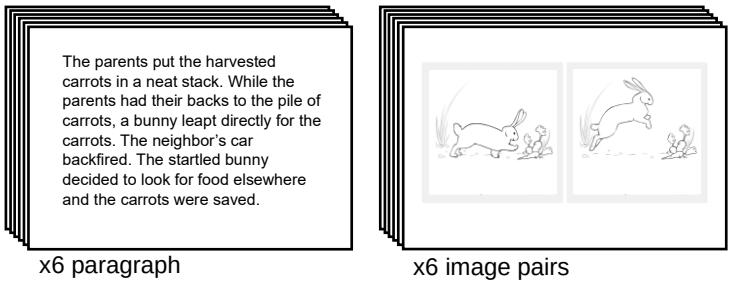
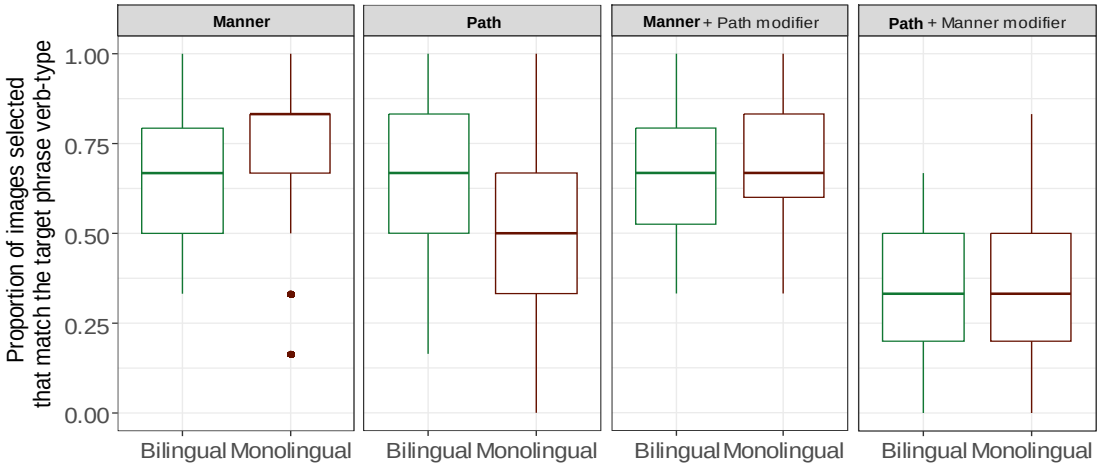


Table 1 (above)**.** Example the four conditions by each target and the paragraph they were embedded in.
Figure 1 (left)**.** Trial structure of one block.
Figure 2 (below)**.** Average proportion of images selected that match the motion in the target phrase verb (bolded) by the two participant groups.



References. [1] Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), Language typology and syntactic description (pp. 36-149). Cambridge: Cambridge University Press. **[2]** Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". **[3]**Slobin, D. I. (2003). Language and thought online: Cognitive consequences of linguistic relativity. Language in mind: Advances in the study of language and thought, 157192. **[4]** Naigles, L. R., Eisenberg, A. R., Kako, E. T., Highter, M., & McGraw, N. (1998). Speaking of Motion: Verb Use in English and Spanish. Language and Cognitive Processes, 13(5), 521–549. **[5]** Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. Cognition, 108(1), 155–184. **[6]** Bunger, A., Skordos, D., Trueswell, J. C., & Papafragou, A. (2021). How children attend to events before speaking: Crosslinguistic evidence from the motion domain. Glossa: a journal of general linguistics, 6(1). **[7]** Filipović, L. (2011). Speaking and remembering in one or two languages: Bilingual vs. monolingual lexicalization and memory for motion events. International Journal of Bilingualism, 15(4), 466-485.

It's not just Imprecision: Stereotypes guide Vagueness Resolution in Implicit Comparisons

Recent work highlighted a bi-directional relation between the social and descriptive dimensions of meaning [1, 5, 2, 7, 4]. For instance, speakers are associated with different social stereotypes based on the *precision* level they choose (High → ***Nerdy***; Low → ***Chill***); and numerals are interpreted *more precisely* when uttered by Nerdy (vs. Chill) speakers [3]. These results show the key interplay of social information and pragmatic interpretation, raising two questions: (**i**) Do social information effects hold for processes of indeterminacy resolution besides numerical imprecision? (**ii**) What specific dimension of pragmatic reasoning leads comprehenders to adjust their interpretation based on a specific stereotype? One possibility is that Nerdy speakers are perceived as especially *attentive to literal meaning* (**Hyp.A**), thus committed to using expressions fully in line with their truth-conditions; alternatively, Nerds might be perceived as *attentive to details* more broadly (**Hyp.B**), and committed to incorporating such details in their descriptions. Prior work on imprecision does not differentiate these options, as both predict more precise interpretations for Nerdy speakers. We thus turn to a new domain: *vague* predicates' interpretation in implicit comparisons.

ICs, Vagueness, & Similarity. *Implicit Comparatives* (**ICs**) with vague predicates, in (1), are subject to a *Similarity Constraint* (**SC**; [8]): the two objects described must *significantly* diverge along the relevant dimension (Cxt 2), resulting in infelicity if they don't (Cxt 1).

(1) Route A is long, but Route B is not. (Road A = 600 miles)
#Cxt 1: Road B = 595 miles ✓**Cxt 2**: Road B = 295 miles ??**Cxt 3**: Road B = 495 miles

The SC is ultimately rooted in the semantics of vague predicates: *long* is true of an object iff it exceeds a contextually relevant threshold *by a significant amount* [6, 8]; but this can't hold for A but not B if A,B only minimally differ. What remains underexplored is how and when comprehenders adjust the threshold of what counts as "different enough" to satisfy the SC to possibly accept the use of ICs in intermediate cases (Cxt.3). We address this by exploring how this process is shaped by social information about the speaker, guided by Hyp.A-B above. Per Hyp.A, Nerdy speakers should be perceived as more committed to strictly adhering to ICs' truth-conditions than Chill speakers, and thus as more *hesitant* to use ICs with similar objects, to avoid risking violating the SC. Per Hyp.B, Nerds should be perceived as more *detail-oriented* than Chill speakers, and thus more *prone* to using ICs with similar objects, since this allows them to express subtle distinctions.

ICs and Imprecision The SC crucially doesn't hold for ICs with Maximum Standard adjectives, which are imprecision-prone, but not vague [8, 9]): using an IC to represent 100% vs. 95% full tanks (e.g.,"A is *full* but B is not") generates a statement that is both highly granular, and perfectly truth-conditionally compliant. Regardless of Hyp. A-B, Nerds should thus be expected to use ICs in this way especially frequently, making it possible to test findings on social effects on numerical imprecision in the adjectival domain, and to assess how the resolution of semantic (vagueness) vs. pragmatic (imprecision) indeterminacy is shaped by social information.

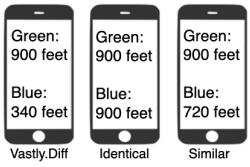
Methods We implemented a variant of [3]'s covered screen task (n=360, from Prolific). The stimuli introduced textual scenarios where one speaker, after looking at their phone, makes a statement containing an IC with a vague adjective. In our first factor, we manipulated the **Identity** of the speaker with three levels:Nerdy, Chill; No.Social (no social information provided, ∅ below).

(2) Rachel and Arthur, {who have been described as [Chill/Nerdy] /∅}, want to go for a swim. Arthur checks his phone and says: "*Green Lake is wide, but Blue Lake is not*".

Participants would then see one phone image with a VISIBLE and one with a Covered screen, selecting the former if they thought the speaker's statement was based on its content, and the latter otherwise. In our second factor, we manipulated the **Similarity** between the two objects being described (e.g., Green vs. Blue Lake), measured as the Object2/Object1 ratio, with three levels: *Vastly.Different* (ratio=0.35; SC clearly satisfied); *Identical* (ratio=1.00; SC clearly violated); and the critical *Similar* condition (ratio=[0.50-0.80]), with SC's status uncertain and con-

tingent on comprehenders' reasoning – range selected based on prior norming). 12 items were distributed in 4 lists (3 each for Vastly.Different/Identical, 6 for Similar). Of the 16 fillers, 8 had ICs with Maximum Standard (e.g., *full* adjectives used to describe near-identical objects (ratio=0.95).

Predictions. VISIBLE-rates, indicating participants accepting the IC, should be at floor/ceiling for Identical/Vastly.Different, with no Identity effect. For the critical Similar condition, we expect intermediate VISIBLE-rates. **Hyp.A** predicts an Identity effect with VISIBLE-rates: Nerdy $<$ No.Social $<$ Chill; **Hyp.B** predicts Nerdy $>$ No.Social $>$Chill. For ICs with absolute adjectives we expect Nerdy $>$ No.Social $>$Chill regardless of Hyp. A-B.



F1: VISIBLE screens in Similarity

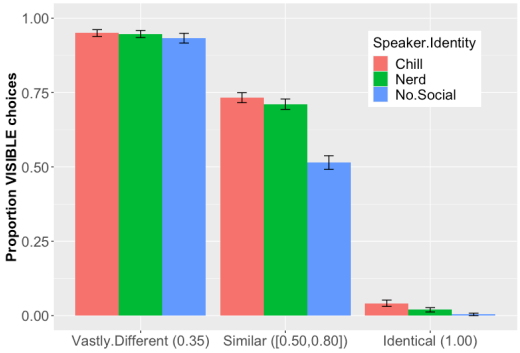
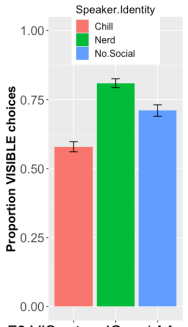


Fig.2: VIS-rates for ICs w/ Vague Adjectives

Results.We fit a ME logistic regression with Similarity (ref=Similar) and Identity (ref=No.Social) as predictors, random intercepts+slopes for Identity for Items, random intercepts for Subjects. VISIBLE-rates (**F2**) are at (near) floor/ceiling in Identical and Vastly.Different, with higher (β=2.48; $p<0.001$) and lower (β=6.47;$p<0.001$) rates than in the Similar condition and no Identity difference. In the critical Similar condition we found an Identity effect, with VISIBLE-rates for *both* Nerdy (β=0.93; $p<0.001$) and Chill (β=1.09; $p<0.001$) higher than No.Social. A ME regression on absolute adjective fillers (**F3**) showed higher VISIBLE-rates for Nerdy (β=0.64; $p<0.05$) and lower for Chill (β=-0.75; $p<0.05$) vs. No.Social.

Discussion Social information affects comprehenders' resolution of the Similarity Constrain – hence, their assessment of whether an Implicit Comparative with a vague predicate is appropriate in the context. This is shown by the higher VISIBLE-rates for both Nerdy and Chill speakers relative to the No.Social condition. The specific pattern, however, does not neatly align with either Hyp.A or B. The observed higher VISIBLE-rate for Nerdy than No.Social aligns with Hyp.B, supporting the idea that these speakers are perceived as especially committed to representing detail, leading comprehenders to accept a relative small difference between the two objects as justifying the use of an IC; yet, the higher VISIBLE-rates for Chill than No.Social is unexpected under this hypothesis. We consider two explanations. One is that the Identity manipulation simply didn't work. But the absolute adjective data speak against this: *consistent* with [3], Nerdy speakers' descriptions *are* indeed interpreted more precisely than Chill ones', suggesting that the social manipulation affected interpretation as expected, and that imprecision resolution is similarly affected by social information across numerals and adjectives. The second option is that comprehenders' similar behavior across the two social identities for vague adjectives is based on a bias towards adopting a charitable interpretation, seeking whatever justification can be found to see the facts on the visible screen as in line with the IC. On this view, comprehenders would then recruit social information to accept the statement-to-scenario pairing in the context in whatever way is consistent with the specific stereotype – by perceiving Nerdy speakers as especially detail-oriented, and of Chill speakers as inclined to be looser with the truth-conditions of ICs. In sum, our findings shed novel light on the interface between social and pragmatic reasoning by: (i) suggesting that the interplay between stereotypes and interpretation, besides imprecision, is also observed in vagueness resolution; (ii) replicating prior results on the effects of social information on imprecision in a different grammatical domain.



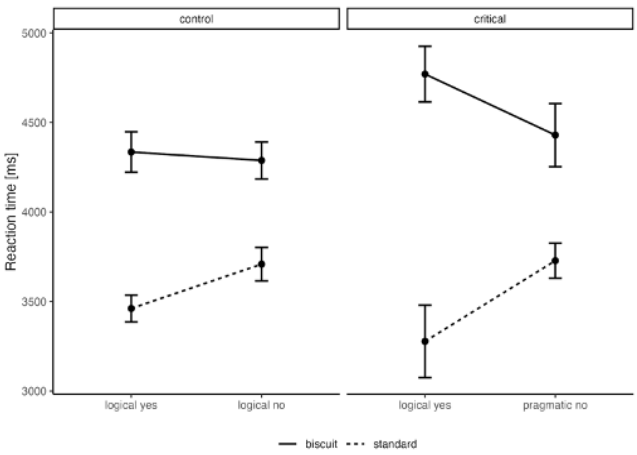
F3:VIS-rates, ICs w/ AAs

[**1**] Acton & Potts 2014. That straight…*J.Slx*• [**2**] Beltrama 2020. Social meaning …. *LgLxComp* • [**3**] Beltrama & Schwarz 2021. Imprecision, Identity …*SALT 31* • [**4**] Beltrama, Solt & Burnett 2022. Context …. *LinS* • [**5**] D'Onofrio 2018. Personae…*LinS* • [**6**] Glass 2015. Strong necessity …. *PWPL* • [**7**] Graff Fara 2000. Shifting Sands. *Phil. Topics* •[**8**] Henderson & McCready. 2020. Dogwhistles, Trust…*AC* •[**9**] Kennedy 2007. Vagueness …*L&P* • [**10**] Solt 2015. Vagueness and Imprecision. *A.R of Ling*

Already Perfect: Conditional Statements

Conditional statements often convey implied meanings beyond their literal content: The standard conditional 'If you mow the lawn, you'll receive $5' is logically true even when the lawn is not mowed and the person receives $5 anyway (e.g., for a different chore). However, listeners often judge this sentence as false in those situations, treating it exhaustively with an '*if and only if*' meaning, known as Conditional Perfection (CP).[1] However, in other cases, sometimes called 'biscuit conditionals', the pragmatic interpretation is infelicitous.[2] For example, in 'If you are hungry, there are biscuits in the cupboard', perfection is not possible, since the outcome (biscuits being in the cupboard) does not depend on the condition (being hungry), making a logical interpretation more fitting. Here we exploit this well-attested difference to investigate how people arrive at the pragmatic interpretation as opposed to the literal, logical one. In two sets of studies, we investigate if computing CP is linked to processing cost and whether the listener starts with the logical (not-perfected) meaning of the conditional and then enriches it via implicature (CP-later hypothesis)[10,12] or instead *begins* with the perfected meaning and retreats to the weaker meaning if supported by context (CP-first hypothesis).[13,14] These hypotheses are associated with different processing costs: an enrichment cost for the CP-later and a weakening cost for the CP-first.

Exp 1: This experiment included 3 reaction time (RT) studies where participants read sentences in the form of (p → q) and then saw pictures in one of the three conditions [control: (p & q), (p & ¬q); critical: (¬p & q)] and evaluated whether the fictional character told the truth (Table 1). In **Exp. 1a** (N=151), both the experimental group reading standard "if" sentences and the control group reading "only if" sentences, where CP is obligatory, showed a clear preference for pragmatic responses. No significant differences in RTs were observed, indicating no additional processing cost for CP. In **Exp. 1b** (N=75), we tested biscuit conditionals and found that they generated longer overall, but there was no difference between the control and critical trials, suggesting no weakening cost either. Note that the RT measures were collected *after* the conditional statement had been read and interpreted, which could pose an issue if participants formed interpretations while reading sentences, leading to RT differences during reading but not in the response phase. Thus, in **Exp. 1c** (N=72), we recorded both the reading and reaction time for each trial, manipulating standard and biscuit conditionals within subjects. The results showed that it took longer to interpret biscuit conditionals, which required a logical interpretation, compared to standard conditionals, which were perfected (β=0.22, SE=0.04, t=5.45, p<0.001). Notably, the logical interpretation of biscuit conditionals was also slower than that of control trials (Conditional*Condition: β=-0.08, SE=0.03, t=-2.61, p<0.01), indicating that computing logical, non-perfected, meanings are costly whereas deriving CP comes without a processing cost (see the Figure on the right).



Exp 2: While data regarding processing costs are informative, they may not conclusively reveal the machinery behind CP. To provide converging evidence, in **Exp. 2a** (N=91), we asked participants to verify sentence-picture pairings (similar to Exp 1) while simultaneously memorizing visual dot patterns, varying in memory load from low to high. Drawing on existing research on scalar implicatures, [11, 14] we hypothesized that an increase in memory load would reduce their capacity to compute pragmatic inferences. Thus, if CP is an inference on top of the logical meaning, then it is less likely to arise under a high cognitive load. Manipulating conditional type (standard, biscuit) and cognitive load (low, high) within-subjects, we found that participants perfected standard conditionals (92%) while the logical responses for biscuits were below chance (41%), irrespective of the degree (high vs low) of the cognitive load. The degree of cognitive load did not influence interpretations of either type of conditionals. The complexity of conditional utterances, paired with our use of a picture-sentence verification task, might have been sufficient to exhaust participants' cognitive resources in *both* load conditions, unique to this study. Supporting this possibility, in Exp 1, we found that participants predominantly (60-80%) provided logical responses for biscuit conditionals when there was no load manipulation. This difference between our prior work and the subsequent study that added load suggests a potential effect, albeit not between the low and high load conditions. Considering these, we ran a No-Load version of the same experiment in **Exp 2b** (N=46) and compared these data to the Load (high & low load combined) conditions. Results revealed an effect of both Load (β=-0.19, SE=0.07, t=-2.58, p<0.01) and Conditional (β=0.47, SE=0.07, t=6.44, p<0.001), such that both types of conditionals were interpreted less *logically* when there was load, and standard conditionals were less logical than biscuit conditionals overall.

Discussion: Results indicated that standard conditionals are understood with a pragmatic meaning without extra effort. In fact, the pragmatic meaning remains even under cognitive load, leading to converging evidence for the CP-first hypothesis. In contrast, a richer pragmatic inference might be necessary to establish the logical interpretation for biscuit conditionals, requiring more resources. Each of these results contrasts with findings regarding other forms of implicature, suggesting that conditional statements - and conditional perfection - may require a unique analysis.

Table 1: Sample stimuli used in **Exp 1** & **Exp 2**

conditional	sample stimuli	[p&q]	[p & ¬q]	[¬p & q]	
standard	Ms. Blicket: If the weather is sunny, I will wear purple.				Did she tell the truth? **Yes/No**
biscuit	Ms. Blicket: If your phone is dead, there is a charger in the drawer.				

References: [1] Geis & Zwicky, 1971; [2] Austin, 1961; [3] Cornulier, 1983; [4]Horn, 2000; [5] von Fintel, 2001; [6] van der Auwera 1997; [7]Marcus & Rips, 1979; [8] van Tiel & Schaeken, 2016; [9] Barrouillet et al., 2000; [10] Bott & Noveck 2004; [11] De Neys & Schaneken, 2007; [12] Noveck et al., 2011; [13] Huang & Snedeker, 2009; [14] Chemla & Bott, 2011; [15] Marty & Chemla, 2013

Context rather than semantic priming drives the early availability of focus alternatives

I. Summary

Successful interpretation of any utterance containing focus requires a comprehender to infer the set of alternatives intended by the speaker [1]. Prior cross-modal forced-choice task studies have endorsed a two-stage model of this process [2, 3, 4]. Under this view, an initial context-insensitive stage of semantic priming provides a second context-sensitive stage with the lexical activation necessary to represent focus alternatives as such.

We present results from a cross-modal probe recognition task experiment challenging this view. We found that alternative status, as modulated by discourse context, influenced the speed of recognition for probe words that were not semantically primed by their focus. We observed this effect immediately after focus was encountered, contrary to the predictions of the two-stage model.

II. Background

Under the two-stage model, identifying alternatives is a *destructive* process. In the first stage, immediately after encountering focus, semantic priming takes place activating a large set of *associates* (i.e., words semantically primed by the focus). In the second stage, a context-sensitive mechanism selects relevant alternatives from among these *associates* and maintains their activation, eventually yielding the appropriate alternative set. In line with this, prior studies found that, after encountering focus, relevant alternatives are only represented following a delay [2, 3, 4].

However, [5] pointed out that none of these studies tested contextually relevant *non-associate* alternatives (i.e., those not semantically primed by their focus). The authors argued that this confound might have obscured the early availability of focus alternatives. They performed a cross-modal probe recognition task experiment with discourses containing a focus (e.g., *violin*) and two relevant alternatives used as probes: one *associate* alternative (e.g., *guitar*) and one *non-associate* alternative (e.g., *pizza*). Contrary to the predictions of the two-stage model, they found that both alternatives were correctly recognized faster than a non-alternative control (e.g., *house*) immediately after the focus was encountered (i.e., 0ms SOA).

[5] took their results to support a *constructive* model in which discourse context alone is utilized to build a representation of the alternative set. Under this view, the early representation of an item as a focus alternative crucially depends upon the surrounding discourse. The present study more directly investigates the potentially context-sensitive nature of this early processing.

III. Method

We modified [5]'s materials (see Table 1) and ran an in-person cross-modal probe recognition task experiment (N=57) in a 2x2 (context x probe word) within-subjects design. In the *two-alt* context, subjects listened to a discourse in which both an *associate* (e.g., *guitar*) and *non-associate* (e.g, *pizza*) were alternatives to a focus (e.g., *violin*). In the *one-alt* context, subjects listened to a discourse in which the *associate* was an alternative, but the *non-associate* was simply mentioned. Immediately after encountering the focus (i.e., 0ms SOA), subjects performed speeded recognition of either the *associate* or the *non-associate* as a written probe.

IV. Results

Given the *one-alt* context, subjects were on average faster to correctly recognize the *associate* probe ($M = 995, SE = 23$) than the *non-associate* probe ($M = 1138, SE = 27$). Given the *two-alt* context, subjects were also on average faster to correctly recognize the *associate* probe ($M = 1038, SE = 23$) than the *non-associate* probe ($M = 1098, SE = 22$). We fit a linear mixed model to the **log-transformed** response times. We observed the predicted interaction ($\beta = 0.03, t = 3.97$). An interaction interpretation is supported by pairwise comparison of the estimated marginal means which indicated that the *non-associate* probe only elicited longer response times in the *one-alt* context, when it was not a relevant focus alternative ($\beta = -0.16, t = -7.12, p < 0.01$).

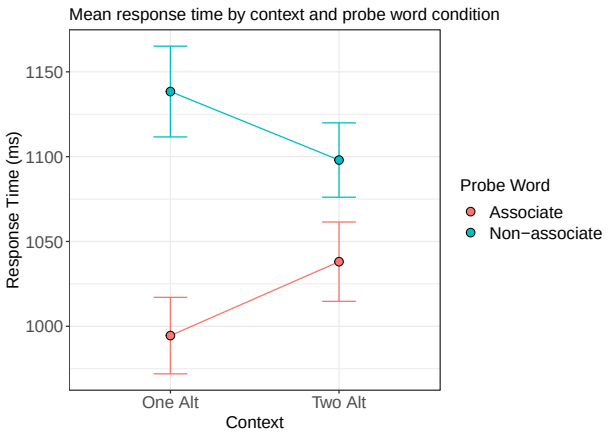


Figure 1: Error bars indicate standard error. Incorrect responses, long responses (>2500ms), and short responses (<200ms) not analyzed.

CONTEXT (AUDIO)	
TWO-ALT:	**ONE-ALT:**
A. Jonah brought the guitar and the pizza to band practice at the new house	**A.** After eating leftover pizza, Jonah brought the guitar to band practice at the new house
B. No, he only brought the [violin]$_F$	
PROBE WORD (VISUAL)	
ASSOCIATE:	**NON-ASSOCIATE:**
GUITAR	PIZZA

Table 1: Example item depiciting context and probe word conditions. The focus alternatives for each context condition occur in a red font.

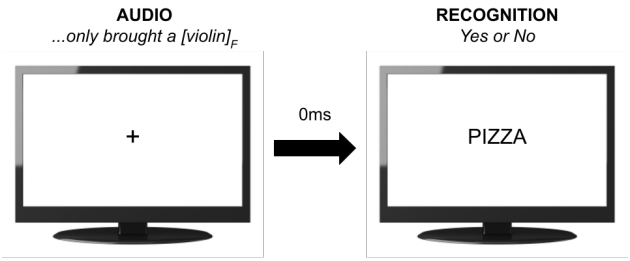


Figure 2: Schema of the cross-modal probe recognition task

V. Discussion

We take the *two-alt* context condition to partially replicate [5]'s findings. As in their study, we found no significant difference in response times between *associate* and *non-associate* probes, as both are contextually relevant alternatives. We take the significant response time penalty observed for the *non-associate* probe in the *one-alt* context to support a *constructive* model of selecting alternatives. Our results suggest that the early availability of alternatives is primarily driven by the discourse context. It is unclear how a *destructive* model dependent upon semantic priming, such as the two-stage model, could capture this early context-sensitive behavior.

[1] Mats Rooth. A theory of focus interpretation. *Natural Language Semantics*, 1992.
[2] Matthew Husband and Fernanda Ferreira. The role of selection in the comprehension of focus alternatives. *Language, Cognition and Neuroscience*, 2016.
[3] Nicole Gotzner, Isabell Wartenburger, and Katharina Spalek. The impact of focus particles on the recognition and rejection of contrastive alternatives. *Language and Cognition*, 2016.
[4] Nicole Gotzner and Katharina Spalek. The life and times of focus alternatives: Tracing the activation of alternatives to a focused constituent in language comprehension. *Language and Linguistics Compass*, 2019.
[5] Author 1 and Author 2. Constructing alternatives: Evidence for the early availability of contextually relevant focus alternatives. In *Alternatives in Grammar*. Palgrave, To Appear.

Spanish Neg-raising: Always in the mood for Neg-raising, sometimes in the mood for NPIs

BACKGROUND. So-called Neg-Raising (NR) predicates like *creer* 'believe', when negated, give rise to two interesting effects: **(a)** they can be interpreted as if negation were in the embedded clause (**NR inference**) and **(b)** they license strict NPIs like *en meses* 'in months' and punctual *hasta las siete* 'until seven' in the embedded clause (**NPI-licensing**), as in (1) (Lakoff, 1969; Horn, 1978; Gajewski, 2007). Non-NR predicates like *asegurar* 'assure' do not give rise to these effects.

(1) María no cree [que el tren llegue hasta las siete]
Mary not believe that the train arrives$_{SUBJ}$ until the seven
⇝ 'Mary believes the train won't arrive until seven' (NR)

However, there is a debate in the literature on how the mood of the embedded complement impacts these two effects in Spanish. On the one hand, it has been long observed that indicative (IND) blocks the licensing of strict NPIs, and this has been used as evidence for the claim that the NR inference is blocked too (Rivero, 1971; Harrington & Pérez-Leroux, 2016; a.o.), see (2). On the other hand, a few have claimed that the NR inference is still available with IND mood (Bolinger, 1968; Fignoni, 1982; Siegel, 2009); but, to the best of our knowledge, they make no mention of whether, in those cases, strict NPIs are also licensed. In fact, given that certain interveners disrupt the licensing of NPIs in general (Homer, 2008; Gajewski, 2011), it might be that IND mood in Spanish disrupts strict NPI-licensing even when the NR inference obtains. This leads to the three alternative hypotheses in (3):

(2) *?María no cree [que el tren llega hasta las siete] (?NR)
Mary not believe that the train arrives$_{IND}$ until the seven

(3) a. **Hyp A**: IND blocks both the NR inference and the licensing of strict NPIs.
b. **Hyp B**: IND allows both for the NR inference and for the licensing of strict NPIs.
c. **Hyp C**: IND allows for the NR inference but blocks the licensing of strict NPIs.

In this paper, we experimentally test these hypotheses, leading to evidence for **Hyp C**.

EXPERIMENTAL DESIGN. We ran a 2x3 study with two simultaneous experiments comparing indicative to subjunctive mood (IND vs. SUBJ) in three sentence types: with a non-NR predicate, with a NR predicate, and with a NR predicate and a strict NPI (NNR vs. NR vs. NR+NPI), see (4). We tested their acceptability on a 1-7 Likert scale (**exp1**) and their ability to convey a NR interpretation (**exp2**). Participants were first asked how acceptable they found the sentence, and, if they rated the sentence as 4 or higher, they were asked whether or not the sentence communicated the NR interpretation ("yes"/"no" response). The materials included 36 critical items using two strict NPIs, *until* and *in years/months*, and six NNR and six NR predicates, all split equally among the predicates and counterbalanced across participants following a Latin Square Design. There were 12 filler items as well as four attention check trials spaced evenly throughout the experimental items. Native Spanish speakers of Peninsular Spanish (n=48) were recruited in Prolific to participate in the experiment, which was implemented using PCIbex (Zehr and Schwarz, 2018).

(4) (Translated version of an example item set)
a. John didn't know that Valeria had(IND/SUBJ) visited the museum that year. (NNR)
b. John didn't believe that Valeria had(IND/SUBJ) visited the museum that year. (NR)
c. John didn't believe that Valeria had(IND/SUBJ) visited the museum in years. (NR+NPI)

Q: *On a scale of 1 to 7, how acceptable does this sentence sound to you?*
Q: *To the extent that the sentence is acceptable, can it have the following interpretation?*
Interpretation: John (knew/believed) that Valeria didn't visit the museum (that year/in years).

RESULTS. For experiment 1, a linear-mixed effects regression model with Acceptability Rating (1-7) as dependent variable and Mood and Sentence Type as independent variables was run in R using the packages lme4 and lmerTest. Participants and items were added as crossed random effects. The model indicated a main effect of both Mood ($p<.0001$) and Sentence Type ($p<.0001$) and, importantly, an interaction between the two (χ^2=31.48, $p<.0001$). Additional post-hoc analysis was conducted using the emmeans()-function to investigate the nature of the interaction. The overall results showed that **(i)** although constructions with NPIs were generally less grammatical than those without NPIs in both IND ($p<.0001$) and SUBJ ($p<.0001$), the effect was larger within IND, and crucially that **(ii) strict NPIs with IND were less grammatical than with SUBJ** ("4"vs."6", $p<.0001$). The raw data are plotted in the box-plot in Fig 1.

For experiment 2, a mixed-effects logistic regression model was run with "yes"/"no" response as dependent variable with the same independent variables Mood and Sentence Type (reference level: NR). The model indicated a main effect of Sentence Type ($p<0.0001$) but no effect of Mood ($p=0.52$) and no interaction ($p=0.29$), thus indicating that **(iii) IND does not block the NR inference**. We then removed Mood as a main effect and reran the model with only Sentence Type. The results indicated that **(iv)**, though the NR constructions were indeed usually interpreted with NR interpretations, the constructions with NPIs produced slightly fewer NR interpretations than those without NPIs. These data are shown in Fig 2 with corresponding confidence intervals.

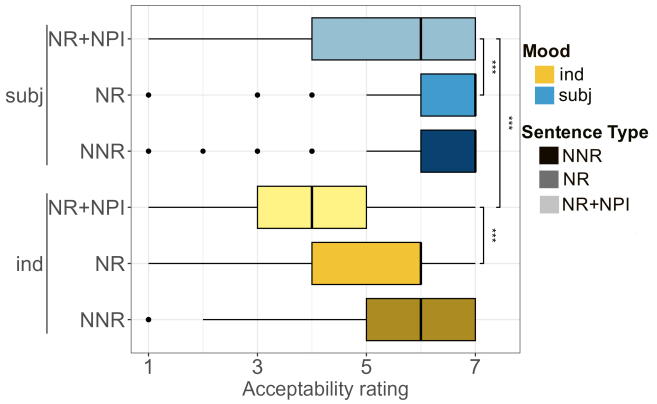


Figure 1: Median acceptability ratings.

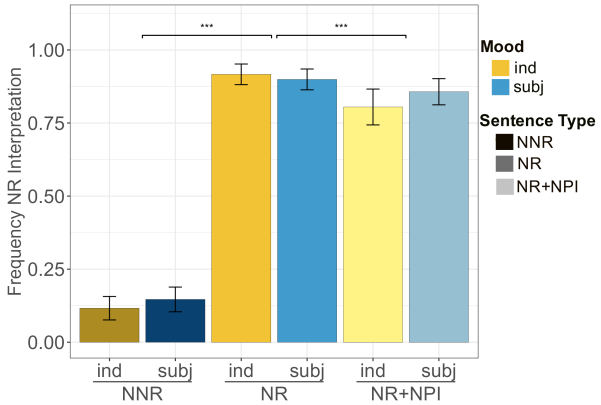


Figure 2: Mean frequency of NR interpretations.

DISCUSSION. Result **(iii)** that IND does not block the NR inference excludes **Hyp A**. Further, result **(ii)** that strict NPIs are less grammatical with IND than with SUBJ argues against **Hyp B**. Result **(i)** also militates against **Hyp B**: while sentences containing a strict NPI seem to involve an extra "tax" compared to their non-NPI counterparts, this "tax" is more substantial with IND (2-pt median difference) than with SUBJ (1-pt median difference). Finally, **Hyp C** correctly predicts the combined results from experiments 1 and 2. Two other results are of interest. Result **(i)** on the additional "tax" of strict NPIs might indicate a potential processing cost from the licensing of NPIs which could be further explored. Result **(iv)** that NR constructions with NPIs produced fewer NR interpretations, even if only slightly, is surprising for all current analyses of NR and strict NPIs and calls for additional investigation.

CONCLUSION. Our results controlling for mood in Spanish show that, contra common practice, the (un)grammaticality of strict NPIs should not be used as an indication of the NR inference.

SELECTED REFS. Bolinger, D. 1968. Postposed main phrases: an English rule for the Romance subjunctive. *CJL*14. • Gajewski, J. R. 2011. Licensing strong NPIs. *NLS*19. • Homer, V. 2008. Disruption of NPI licensing: The case of presupposition. *SALT*18. • Horn, L. R. 1978. Remarks on neg-raising. In *Pragmatics*. • Lakoff, R. 1969. A syntactic argument for negative transportation. *CLS*5. • Rivero, M.-L. 1971. Mood and presupposition in spanish. *FoL*. • Zehr, J. and Schwarz, F. 2018. Penncontroller for Internet Based Experiments (IBEX).

Exploring the Agent-Relativity of Truth

Imagine a context where a speaker is justified in making a claim but its content does not correspond to the facts. Recent work in experimental philosophy and semantics has uncovered that English speakers tend to split on the truth value of sentences like "Joe might be in Boston" in such contexts (Knobe & Yalcin 2014, Phillips & Khoo 2019, Phillips & Mandelkern 2020). Subsequent empirical studies even suggest that truth-value judgments of simple declarative sentences like "Joe is in Boston" show surprisingly high variability (Reuter & Brun 2021, Ricciardi & Martin 2022). These findings pose a challenge to the commonly assumed view in formal semantics that what makes a sentence true or false is just its correspondence with the facts (**correspondence sense of "true"**). So, what underlies the variability in truth-value judgments? Reuter & Brun 2021 hypothesized that such variability is due to an inherent ambiguity of the term "true" between the correspondence sense and a **coherence sense**, according to which a sentence is true or false depending on whether its content coheres with the speaker's set of beliefs at the time when she utters the sentence. The study we present investigates the following question:

Question: What is the key determinant in activating a particular sense of "true" across contexts?

In this paper, we hypothesize that the key determinant is **agent-relativity**. More specifically, we predict that if the focus is on the sentence uttered by a speaker as in "Is it true that [sentence]?", people will tend to apply a correspondence sense of "true". In contrast, if the focus is on the agent making a statement as in "Is it true what A said?", people will be more inclined to apply a coherence sense of "true".

Methodology. We designed a two-response options questionnaire with a 2 x 5 design where participants first read one of two stories adapted from previous works (**see 1**) and then answered one of five questions (**see 2**). We recruited 400 participants from Prolific Academic who were randomly assigned in batches of 40 to one of the ten conditions.

(1) The two stories read by participants

Story A: Party
Maria and Peter are students and meet up for a late dinner. Peter asks Maria whether Tom is at the party that they intend to go to after dinner. Maria answers that **Tom is at the party**. After all, Tom had told her that he would be at the party. When they arrive at the party, it turns out that Tom has changed his plans, and is not at the party.

Story B: Boston
Sally and George are meeting up in a cafe in the afternoon, talking about whether Joe is currently in Boston. Yesterday, Joe told Sally that he would have a job interview in Boston at 5 pm today and he would fly there early in the morning. So, Sally states: "**Joe is in Boston**". Just then, George gets an email from Joe. The email says that the job interview was canceled and that he is still in Berkeley. So George says: "No, he isn't in Boston. He is in Berkeley."

(2) The five critical questions (A = speaker, S = sentence)

Question type	Label	Response Options

Has A said the truth?	(A Said Truth)	Yes/No
Was A's answer true or false?	(A's Statement)	True/False
Is it true what A said?	(What A Said)	Yes/No
Is the underlined statement true?	(Pure Statement)	Yes/No
Is it true that S?	(Fact)	Yes/No

Results. We ran χ2-square tests to assess the impact of the two independent variables, *scenario* and *question type*, on participants' responses. For *scenario*, the analysis revealed a significant effect with $\chi2 = 13.33$, $p < 0.01$, Cramer's $V = 0.183$. Regarding *question type*, the test showed a significant effect with $\chi2 = 100.55$, $p < 0.001$, Cramer's $V = 0.502$. For Story A Party, the proportions of 'Yes/True' responses by question type were as follows: Truth (77.5%), A's statement (55%), What A Said (52.6%), Pure Statement (12.5%), and Fact (0%). For Story B Boston, the proportions of 'Yes/True' responses by question type were as follows: Truth (60%), A's Statement (32.5%), What A Said (25%), Pure Statement (17.5%), and Fact (2.5%). **See also Figure 1**.

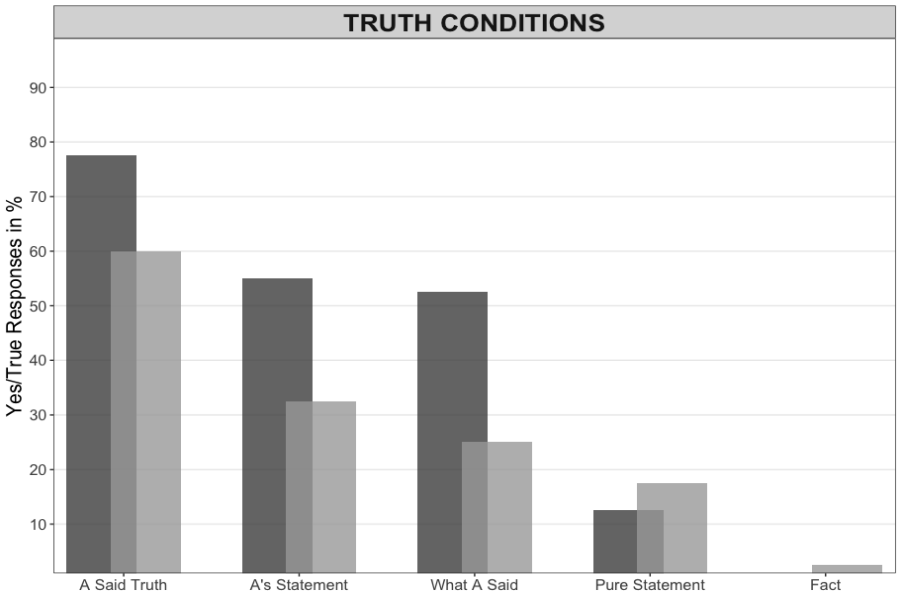


Figure 1. Yes/true responses in % of participants in the five question types tested. Dark grey columns show the results for the Party story, light grey columns for the Boston story.

Conclusion. In this work, we investigated what drives variance in truth-value judgments across different contexts. Our results show that when the task instructions are phrased in agent-relative terms like in "Has A said the truth" or "Was A's answer true or false?", a large proportion of people judge the sentence to be true; instead, when the task instructions focus on the sentence itself like "Is the underlined statement true?" or "Is it true that S?", most participants converge in judging the sentence to be false. We take these findings as suggesting that with task formulations focusing on the agent, many people seem to access the coherence sense of "true" whereas with task formulations focusing on the sentence people converge on the correspondence sense. Overall, our findings indicate the need for further investigations into how naive participants interpret the terminology employed in experimental semantic tasks. These investigations are crucial to test further hypotheses on the exact meaning of the two senses of "true", as well as possible pragmatic factors that drive people's responses.

Getting to the Truth is More Cognitively Demanding – Another Look at the Role of Working Memory in Negation Processing

For negative sentences, the results of visual probe recognition task [1] show that participants may take longer to respond to images that match the true states of affairs (soas) than mismatch images, which depict the positive argument of negation. [2-4] argues that attention to the positive soa soon after reading a negative sentence is the outcome of normal parallel language processes which compute both the content and the relevance of the utterance (QUD) given the same linguistic source and discourse information. [2-4] maintain that sentential negation alone can trigger a strong cue to a type of context where the positive soa is entertained as a live possibility. Therefore, preference for an image consistent with the positive soa after reading simple negative sentences suggest inferences about context may be stimulated first. Our idea is simply that, participants' expectations about visual probes are influenced by inferences based on the interpretation of the linguistic stimulus in context. In 'the banana is not peeled', parsing the subject and predicate ('peeled') directly promotes inferences about the denied state of the banana, while inferences about the asserted state would draw on associated world knowledge not directly encapsulated in linguistic expressions. In this new work, we consider the effect of working memory on negation processing. Given the idea that inferring the actual scenario of negative sentences is more resource intensive, especially in comparison to the affirmative sentences, we contend that for simple negative sentences in [1], individuals with more working memory (WM) resources are more likely to integrate background inferences about the positive context and activate the true soa at an earlier stage. We present the results of two fully-normed, probe task experiments based on [1-2], where the participants' WM capacity is manipulated in a dual-task (Exp.1) and measured in a WSPAN task (Exp.2). The results bring convergent evidence that inferring aspects of the content for simple negative sentences requires more cognitive resources than computing the expected context.

The Norming Task: Participants (N=46) completed an object-name probe task which used the same nouns (N=28) and images as in Experiment 1 and 2. Their task was to decide if the object had been mentioned in the preceding screen. Filler nouns (N=28) counterbalanced for response.
Results: A LME model predicting the Log (RT) from match showed no significant ME of match (p=.284).

Experiment 1: Participants (N=40) in the no-memory load group only did the probe recognition task, which asked to first read a sentence and then to decide whether the item in the image had been mentioned in the sentence. The other group (N=41) additionally completed a memory load task, which consisted of remembering a simple grid pattern at the beginning of each trial and recreating it after the probe task response. The probe task has a 2 (polarity) * 2 (match) within-group design. See Table 1.

Results: A LME model was constructed to predict the Log(RT) from polarity, match and WM load. Results showed highly significant MEs of polarity and match (ps<.001), interactions between WM load and match (p=.007), and between polarity and match (p=.005). Crucially, the three-way interaction was significant (p=.05). We further broke down the interaction by the load group which revealed that no-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). See Figure 1 (left).

Experiment 2: Participants (N=72) undertook two tasks in the following order: (a) Word span task (WSPAN) ([5-6]); (b) probe recognition task. The design of (b) is the same as the probe task of Exp. 1.

Results: Analysis of just the probe task showed significant main effects of polarity (p=.03), match (p=.01) and an interaction between polarity and match (p=.002). Then we constructed a LME model predicting Log(RT) from polarity, match and WSPAN score. There was a significant interaction between polarity and match (p=.001), and an interaction between match and WM score (p=.04). Additionally, there was a marginal three-way interaction (p=.08). To follow up, we separately looked into the data of High (top 25%) and Low (bottom 25%) WSPAN score participants. The post hoc analyses revealed that High WM group showed a main effect of match (p<.001) and also an interaction between polarity and match (p=.03) whereas the Low WM group showed no main effect of match only a significant interaction between polarity and match (p=.02). See Figure 1 (right).

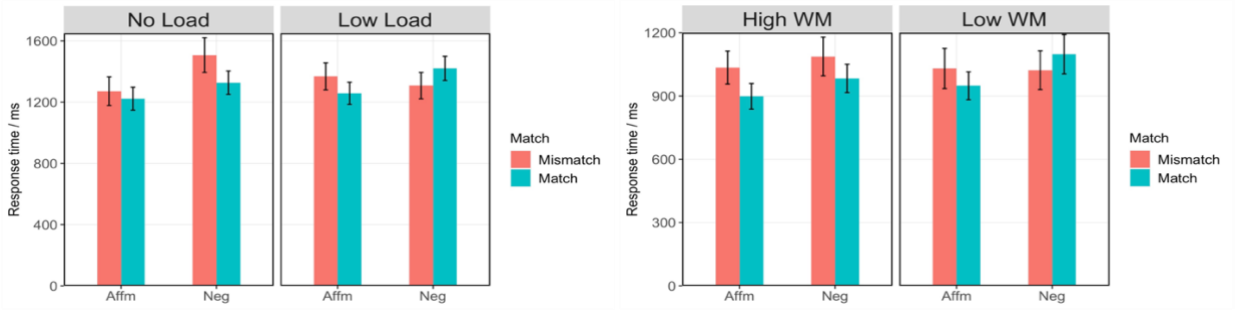
Discussion: The results of norming task show that given only the nouns there was no preference for one state over the other. In low load/negative trials of Exp.1, the response delay for negative compared to positive soa indicates that WM load has a greater impact on processes that arrive at the expectations for the actual content. For Exp.2, regardless of polarity, HWM individuals' responses were most influenced by inferences about the true soas while LWM individuals do not consistently show this. Two experiments jointly attest the costs involving in getting to the truth of simple negative sentences.

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 the probe task. 2 (Polarity) * 2 (Match) design.

Figure 1. Exp. 1 (Left). Mean RT for each condition of No/Low Load groups. Error bars represent 95% confidence intervals.

Exp. 2 (Right). Mean RT for each condition of High and Low WM groups. Error bars represent 95% confidence intervals.



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. 50, 1511-1534. [5] Engle, Tuholski, Laughlin, & Conway (1999). J. of Experimental Psychology: General. 128(3), 309-331. [6] La Pointe, & Engle (1990). J. of Experimental Psychology: Learning, Memory, and Cognition. 16, 1118-1133.

Do all Telic-Perfective Sentences (Always) Culminate? An Exploratory Study on Event Culmination in Italian Monolingual Adults.

In the traditional analysis, telic-perfective sentences entail event culmination [1, 2]. Therefore, these sentences can be judged true only when used to describe a culminated event. According to Krifka's mereological theory, telicity is compositionally derived at the VP level from the combination of the verb's semantics and its direct object. Moreover, telicity is defined by the quantization properties of predicates and the notions of 'homogeneity' and 'cumulativity'. Nevertheless, there is agreement in the literature that telic predicates are not a homogeneous class: some predicates, indeed, exhibit what has been called 'variable telicity', allowing for both telic and atelic interpretations [3]. Moreover, according to van Hout [4], several studies investigating children's acquisition of event culmination have found a relatively high acceptance rate of non-culmination in adults – although adult data are not the primary focus of the research, they are consistently collected in these studies. Finally, a new line of research has recently focused on non-culmination in some adult languages, highlighting that culmination entailment does not always hold. Indeed, in some cases, the culmination reading is only an implicature that can be canceled within the same sentence (i.e., Mandarin, Hindi, etc.) [5]. Despite this suggestive evidence, experimental work on adult native speakers' interpretation of telic-perfective sentences is rather limited and far from being conclusive.

This study aims to establish whether telic-perfective sentences always entail event culmination or not. Specifically, we are interested in (i) whether native Italian adult speakers accept the 'non-culmination' reading and (ii) under what conditions this reading is accepted (is it equally acceptable across different verb classes?).

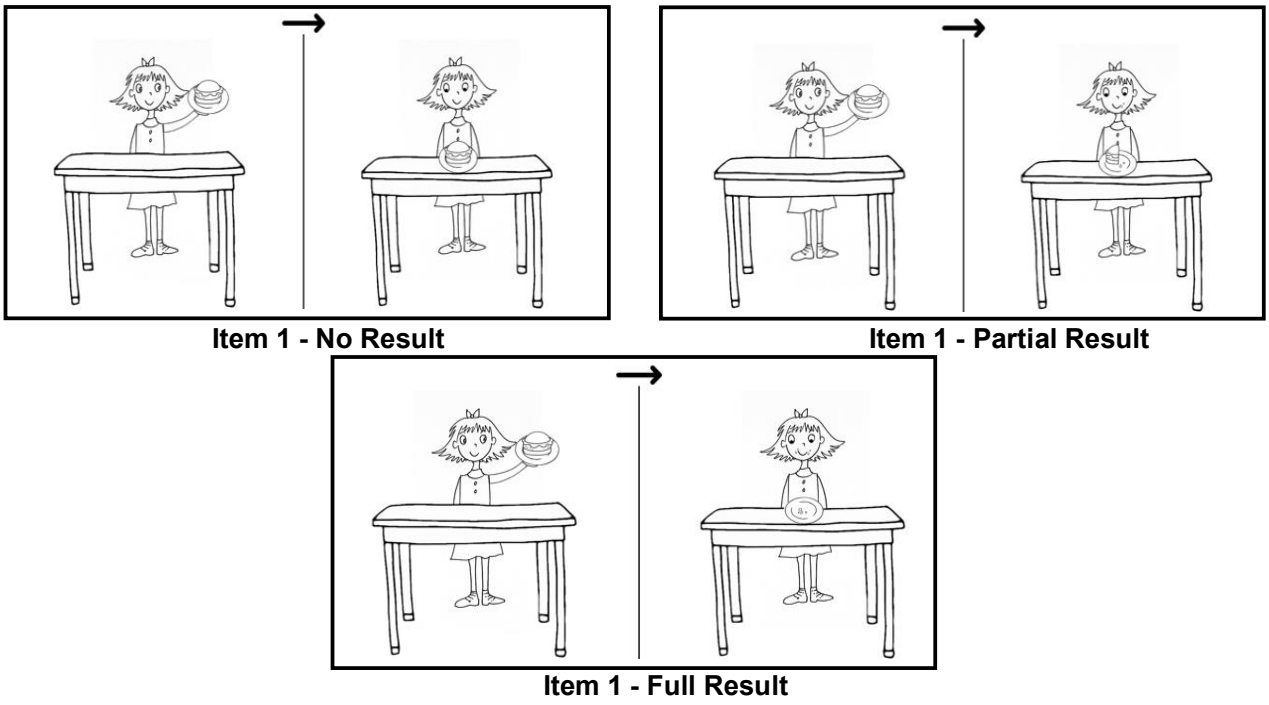
We decided to compare different verb classes since, according to van Hout [4], different verbs may trigger different acceptance rates of 'non-culmination'. Variation across verb types, with different acceptance rates of 'non-culmination', is not expected in a mereological theory of telicity based on quantization, but it seems in line with a Scalar Approach, as the one proposed by Rappaport Hovav, Kennedy, and Levin, among others [6, 7, 8]. In summary, in predicates whose meaning encodes a 'two-point scale' (i.e., to open), culmination is an entailment and cannot be canceled. On the other hand, in predicates whose meaning lexicalize a 'multi-point scale' (i.e., to wipe), culmination is not entailed. The 'culmination reading' is indeed a conversational implicature, hence cancelable. Therefore, based on the Scalar Approach, we expect predicates to behave differently according to the type of scale they lexicalize. CoS_P verbs, lexicalizing a 'two-point scale' should be rejected in a visual context where the event has not culminated since culmination is an entailment and cannot be canceled. On the other hand, CoS_D verbs, lexicalizing a 'multi-point scale', should also be accepted as a description of a 'partial result' event since the 'culmination reading' is a conversational implicature, hence cancellable. As for Incr_T verbs, according to Rappaport Hovav [7], the scale associated with this class of verbs has a different status. Indeed, the 'volume/extent scale' is not directly encoded in the verb but is provided by the verb's object (i.e., VP level). Therefore, we may expect a difference in the acceptance rate of 'non-culmination' for this class compared to the previous ones, in which the scale is lexically encoded in the verb.

To achieve our goals, a group of 60 native Italian speakers (F = 38, Male = 21, Not Binary = 1; age on average = 27.68, SD = 9.43) was recruited through the SONA System and Prolific platforms and administered a truth-value-judgment task. The experimental stimuli consisted of 99 pairs of images: one-third depicted a 'no result' situation, the second-third a 'partial result' situation, and the last a 'full result' situation ('degree_of_event') (see Figure 1). The events, in total, were 33, divided into 3 different verb classes ('verb_type'): punctual change of state verbs (i.e., open the window), durative change of state verbs (i.e., melt the ice cube), and incremental theme verbs (i.e., eat the sandwich). Participants' task was to determine whether the sentence described the rightward picture correctly by pushing a 'yes' or 'no' button on the screen. The crucial condition is the 'non-culminating' situation, namely, the 'partial result' image.

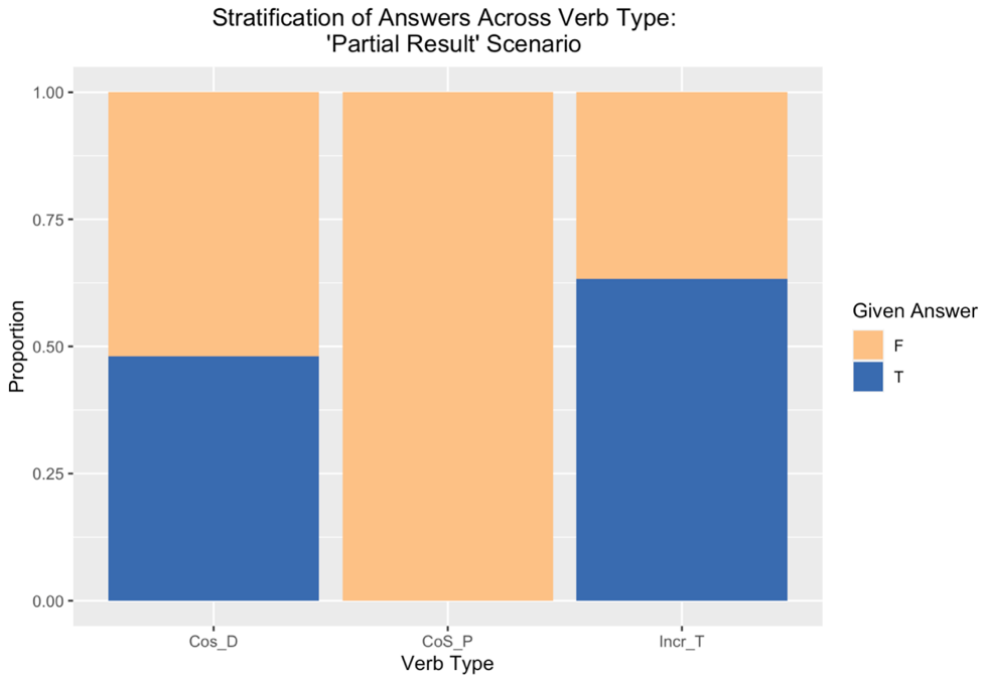
Data were analyzed using a logistic mixed model, with 'given_answer' as dependent variable, 'verb_type' and 'degree_of_event' as factors, and 'age' as a nuisance covariate. Participants' ID codes were included as a random effect. Statistical Analysis revealed main effects of the factors 'verb_type' ($X^2_{(1)}$ = 39.45, p-value = < .001) and 'degree_of_event' ($X^2_{(1)}$ = 735.9, p-value = < .001), but no significant interaction. Participants' age significantly influenced the type of response ($X^2_{(1)}$ = 3.85, p-value = < .04). Based on posthoc analyses on the main effect of 'degree-of-event", different answers were obtained for 'no' vs. 'full result' (p <.0001) but no difference emerged for 'partial' vs. 'full result' (p = 0.4) and 'no' vs. 'partial' (p = 0.99). As expected, participants, on average, did not accept the 'partial result' scenario for CoS_P verbs. On the other hand, a higher degree of acceptance of the 'partial result' scenario was recorded for CoS_D and Incr_T verbs. Data seems compatible with the Scalar Approach since CoS_P verbs were never accepted in the 'partial result' scenario, as they lexicalize a "closed scale". Nevertheless, contrary to the predictions of the Scalar Approach, CoS_D, and Incr_T verbs did not behave differently. One reason for that may be the fact that visual context and sentences were presented in an "out-of-the-blue" fashion. As suggested by previous studies [9], contextual information may influence the acceptance of the 'non-culmination' reading, a possibility that we intend to investigate in a follow-up study by adding contextual background and/or the agent's goal as additional variables.

Figures and Graphs.

- *Figure 1. Example of the experimental item 'eat the sandwich' in the 3 scenarios.*



- *Graph 1. Proportion of 'True' VS 'False' answers across verb type: partial result scenario*



References.

[1] Krifka, M. (1989). 'Nominal Reference, Temporal Constitution and Quantification in Event Semantics,' in R. Bartsch, J. van Benthem and P. van Emde Boas (eds.), Semantics and Contextual Expressions 75-115. Dordrecht: Foris. [2] Krifka, M. (1998). 'The Origins of Telicty,' in Events and Grammar, S. Rothstein, ed., Dordrecht: Kluwer. [3] Martin, F. (2019). *"Non-culminating accomplishments"*. Language and Linguistics Compass, 13(8), e12346. [4] van Hout, A. (2018). "*On the acquisition of event culmination*". Semantics in language acquisition, 96-121. [5] Martin, F. et al. (2020). "*Children's non-adultlike interpretations of telic predicates across languages*". Linguistics, 58(5), 1447-1500. [6] Hovav, M. R. (2008). Lexicalized meaning and the internal temporal structure of events. *Theoretical and crosslinguistic approaches to the semantics of aspect*, *13*. [7] Hovav, M. R., Alexiadou, A., Borer, H., & Schäfer, F. (2014). Building scalar changes. *The syntax of roots and the roots of syntax*, 259-281. [8] Kennedy, C., & Levin, B. (2008). Measure of change: The adjectival core of degree achievements. [9] Mathis, A., & Papafragou, A. (2022). Agents' goals affect construal of event endpoints. *Journal of Memory and Language*, *127*, 104373.

The Role of Working Memory in Scalar Implicature Computation in ADHD and Non-ADHD Individuals

This study investigates the real-time processing of scalar implicatures in people with or without ADHD. A scalar implicature arises when the logical meaning of a sentence departs from its pragmatically enriched reading. The most well-known example of scalar implicatures are observed in sentences with under-informative quantifiers. For example, the scalar term 'some', can mean "some, and possibly all", but speakers typically compute an implicature and interpret it to mean "some, but not all". Studies have demonstrated that accessing the latter, pragmatically enriched interpretation, requires more cognitive effort as it relies on greater use of working memory resources (De Neys & Schaeken, 2007; Dieussaert, Verkerk and Gillard, Schaeken, 2011; Marty, Chemla, Spector, 2013; Antoniou, Cummins and Katsos, 2016; Cho, 2020). We also know that working memory deficits are a clinical characteristic of ADHD, and individuals with ADHD struggle more under cognitive load than neurotypical individuals (Kofler, Rapport, Bolden, Sarver and Raiker 2010; Kim, Liu, Glizer, Tannock and Woltering 2014).

Taking these findings into account, in this present study, we wanted to investigate how working memory load impacts scalar implicature computation in a sentence verification task, for both non-ADHD individuals and individuals with ADHD. We hypothesised that if working memory plays a role in scalar implicature computation, and if adults with ADHD have more a limited working memory capacity compared to neurotypical adults, then the working memory load should affect their performance more than neurotypical adults' performance. Our aims were to: 1. Replicate the finding that working memory limitations impair scalar implicature derivation, and 2. Find out whether adults with ADHD differ in scalar implicature computation compared to neurotypical adults. We collected data from 81 participants (41 ADHD, 40 non-ADHD) from the Prolific platform to complete our study. Participants completed an ADHD trait scale, in addition to a dual Truth Value Judgement and Memory Load Task to measure scalar implicature computation. This study was a direct replication of the original De Neys & Schaeken (2007) study, but with the addition of an ADHD group. For examples of sentences and to see the structure of a single trial *refer to Figures 1 and 2*, respectively.

We observed no effects of memory load ($\beta = 0.621$, $SE = 0.322$, $z = 1.93$, $p > 0.05$) or diagnostic status ($\beta = -0.872$, $SE = 0.915$, $z = -0.953$, $p > 0.05$) on the acceptance of under-informative statements. However, we did observe a significant interaction between ADHD status and memory load ($\beta = 1.27$, $SE = 0.431$, $z = 2.94$, $p < 0.01$), such that the non-ADHD participants were more likely to accept these sentences as true under high memory load, compared to the ADHD participants who had a baseline tendency to accept these sentences as true irrespective of memory load condition *(see Figure 3).* These findings suggest that individuals with and without ADHD might differ in their computation of scalar implicatures. This aligns with what we predicted based on the previous findings that people with ADHD have a lower working memory capacity and therefore might be less likely to generate scalar implicatures due to insufficient working memory resources.

To our knowledge, this study was the first to test scalar implicature computation in this population. This not only enhances our understanding of the role of working memory in scalar implicature computation and how diverse cognitive abilities affect scalar implicature computation, it also allows us to understand how individuals with ADHD process language in real-time and how executive dysfunction, specifically working memory deficits, might impact pragmatic language comprehension more generally.

Target Utterance	*Utterance Status*	*Expected Response*
Some trout are fish	***True but under-informative***	***True or False***
Some birds are magpies	*True and informative*	*True*
Some pigeons are insects	*False*	*False*

Figure 1. Example Sentences from Truth Value Judgement Task: True but under-informative, True and False

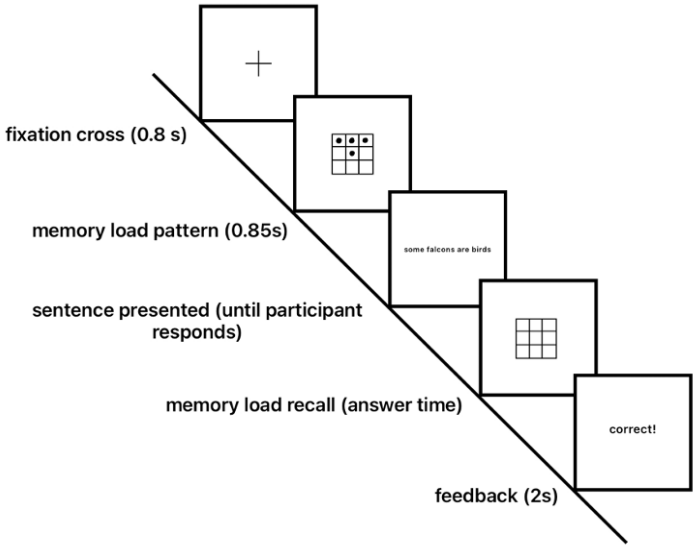


Figure 2. Structure of a Single Trial (High Load)

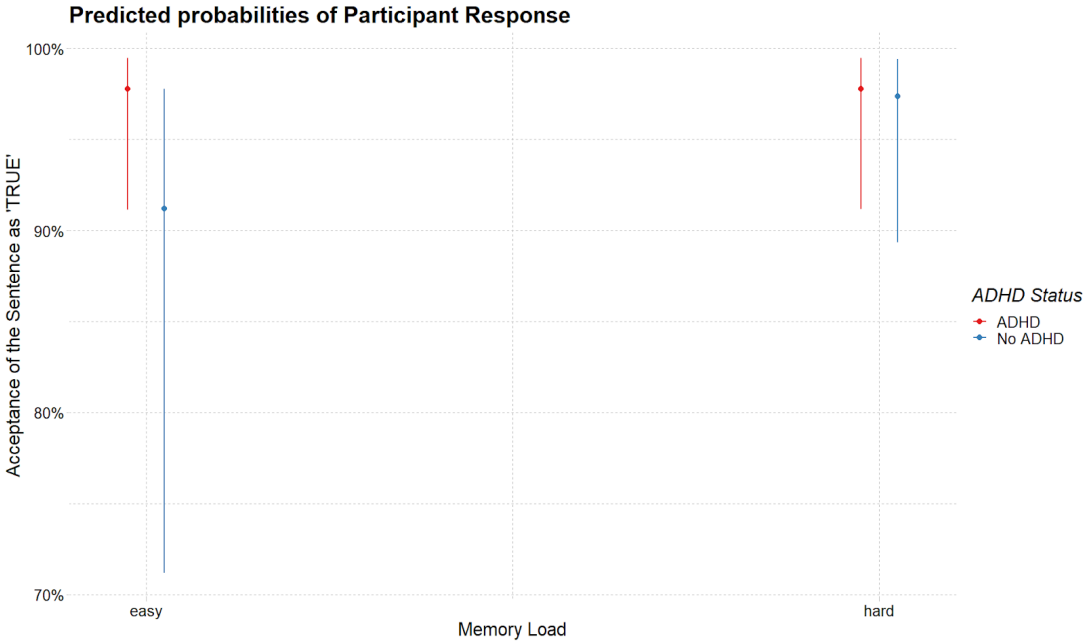


Figure 3. Plot showing the interaction between Memory Load Condition and Diagnostic Status on Participants Acceptance of Under-Informative Statements.

Learning the logic in language: Acquiring the meanings of *all, every* and *each*

Natural languages contain a vocabulary of words that specify semantic relations between the elements in a sentence, like the universal quantifiers *all, each* and *every.* Although the relations specified by these words are all universal (i.e., they specify the 'for all' relation) they differ on other dimensions, such as distributivity. *Each* necessarily specifies a distributive relation: the predicate must separately apply to each individual member of the quantified set. The distributivity of *every* is weaker, while *all* can be used when the predicate applies to the quantified set collectively (e.g., Roberts, 1987; Tunstall, 1998). Previous studies on the acquisition of universal quantifiers often assumed that children treat them as universal from the outset, and only become sensitive to differences in distributivity later in development (see Syrett, 2019, for overview). However, it is also possible that the universality of *each* and *all* have different sources. In particular, the universal force of *each* might be a byproduct of its distributivity – of applying the predicate to each individual until none are left (see also, Knowlton, et al., 2022). In that case, children might not understand *each* as universal until whenever they also understand it as distributive. We tested these alternatives by directly comparing children's understanding of the universality of different quantifiers. Do children acquire the universality of different quantifiers at different points in development or all at once?

In Experiment 1, children (3-7 years old, *n* = 110) were shown five toy fruits and an Elmo puppet. They were asked *Can you give Elmo {each/every/all/some/a/dax} (of the) fruit?*, with *dax* serving as a baseline for how children respond when they don't know the quantifier's meaning. Results from a mixed-effect model revealed that older children were more likely to give a universal response (i.e., the maximal number of items) when prompted with any of the universal quantifiers (*all*, *each*, *every*) than younger children (which was not the case for *dax*). This suggests that the universality of these quantifiers is acquired gradually in development. However, the analysis also revealed differences between quantifiers: Averaged across ages, children were more likely to interpret *all* and *every* as universal than *each*, and even among 7-year-olds, *each* was only interpreted universally in about 75% of trials (Fig. 1).

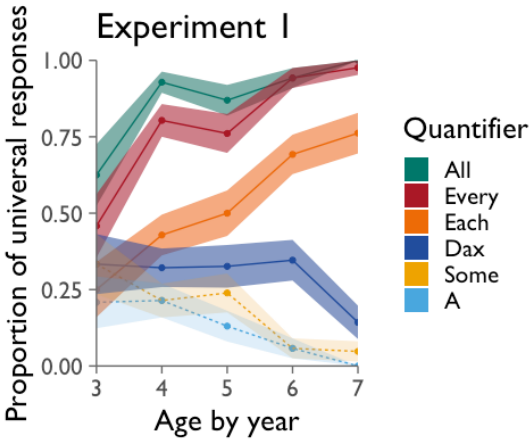


Fig. 1 Proportion of trials in which the maximal number of items were given in Experiment 1, split up per age (plotted in years) and quantifier. The shaded area represents the standard error. The quantifiers *some* and *a* are plotted for completeness, but not included in our analyses.

In Experiment 2, we focused on *each* specifically. Children (4-7 years old, *n* = 78) watched an animation of Cookie Monster taking a bite out of zero, two, or three out of three cookies. They were then asked *Did Cookie Monster bite each/the/two/dax (of the) cookies?* In our main analyses, again conducted with mixed-effect models, we tested whether children differentiated *each* from *dax*. When Cookie Monster bit two of the three cookies, the correct response would be to say

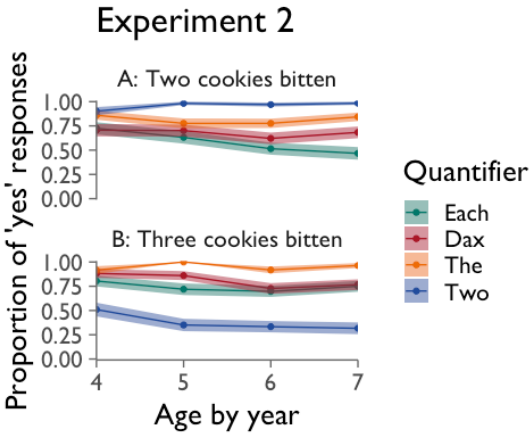


Fig. 2 Proportion of 'yes' responses, split up by event outcome, quantifier, and age.

'no' to the question of whether he bit *each of the cookies*. However, our analyses revealed that children were just as likely to say 'no' to the questions with *each* as to those with *dax*, and even the oldest children only provided correct 'no' responses for *each* on 50% of trials when two cookies were bitten (Fig 2a). When Cookie Monster bit three of the three cookies, the correct response would be to say 'yes' to the questions with *each*. Again, our analyses revealed that children were just as likely to say 'yes' to the questions with *each* as to those with *dax*, and even the oldest children responded with 'yes' to the questions with *each* in only about 75% of trials (Fig 2b). These findings reinforce the conclusion that children do not interpret *each* as universal until late in development.

In our ongoing Experiment 3, we are testing whether the late acquisition of *each* as a universal persists across sentences that might encourage a more distributive interpretation. Children (3-6 years old, *n* = 66 so far) are presented with three toy fish and a pile of toy fruits, and asked *Can you give {each/every/all/some/a/dax} (of the) fish fruit?*. In this experiment, a distributive interpretation may be more accessible than in the previous experiments because the questions can be answered by pairing fish and fruit one-to-one. We have not conducted inferential statistics due to ongoing data collection, but preliminary results (Fig. 3) show 4- and 5-year-olds already predominantly giving universal responses when prompted with *each*, and 6-year-olds nearing ceiling. We're currently investigating whether this pattern holds in a truth-value judgement task (Experiment 4). These observations suggest that constructions which encourage a distributive interpretation of *each* may thereby create a universal interpretation, via a one-to-one mapping between quantified individuals and predicates (e.g., fish and fruit).

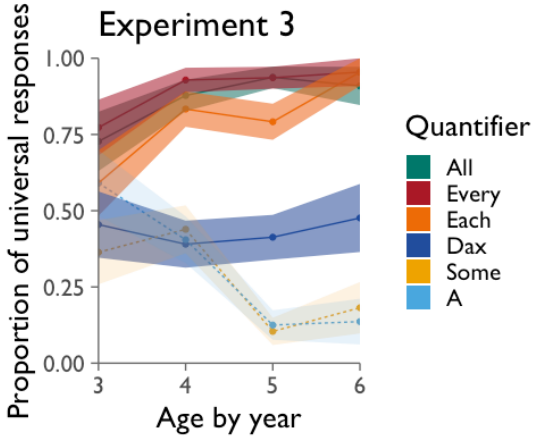


Fig. 3 Proportion of universal responses in Experiment 3, split up per age (in years) and quantifier. The shaded area represents the standard error. The quantifiers *some* and *a* are plotted for completeness, but not included in our analyses.

Our findings reveal that children learn that *all* and *every* are universal quantifiers before they learn that *each* is, at least in contexts in which each is not also clearly distributive. This suggests that different universal quantifiers are learned in a dissociable manner, possibly due to differences in the underlying cause of their universal force. In particular, children may understand that *each* has universal force only once they understand it as distributive.

References

Brooks, P. J., & Braine, M. D. (1996). What do children know about the universal quantifiers all and each?. *Cognition, 60*(3), 235-268.

Knowlton, T., Trueswell, J., & Papafragou, A. (2022). A Mentalistic Semantics Explains "Each" and "Every" Quantifier Use. *Proceedings of the Annual Meeting of the Cognitive Science Society,* 44, 395-401. Retrieved from https://escholarship.org/uc/item/5qt582m2

Roberts, C. (1987). *Modal subordination, anaphora, and distributivity* [Doctoral dissertation, University of Massachusetts Amherst].

Syrett, K. (2019). Distributivity. In C. Cummins & N. Katsos (Eds.), *The Oxford Handbook of Experimental Semantics and Pragmatics* (pp. 143-155). Oxford University Press

Tunstall, S. L. (1998). *The interpretation of quantifiers: Semantics and processing.* University of Massachusetts Amherst [Doctoral dissertation, University of Massachusetts Amherst].

Semantic and Social Meaning Match: experiments on modal concord in US English

Introduction: Recently, formal and experimental linguistics show a growing interest in studying social meaning of language users' choice among functionally similar variants, integrating formal grammar with methods of sociolinguistics, language comprehension and perception [1, 2]. Here, we report a case study on modal concord (MC) in US English: MC (e.g., *may possibly*) refers to the phenomenon where two co-occurring modal elements of epistemic modality and the same force (possibility/$\Diamond$ or necessity/$\Box$) give rise to the interpretation of one single modality (SM) [3]. In comparison to SM, MC has a more restricted use — given their (arguably) equivalent semantics, MC and SM can function as alternative choices in different contexts of use, so what is the mechanism behind the choice for SM vs. MC, and how is the choice processed and perceived?
Experiments – method. We conducted two experiments in US English, Exp1 without context and Exp2 with context — Both used 24 items (and 17 fillers): each item consisted of an introduction sentence (S1), which was fixed for Exp1 and included the CONTEXT manipulation for Exp2, and the critical sentence (S2), see (1). CONTEXT was manipulated via social relations (distant vs. close), which have been shown to affect linguistic choice, a.o. also choice among modal expressions [4, 5]. In both experiments, participants rated (S2) w.r.t. its **(i) interpretation** using the speaker commitment ratings, see (1)-(Q1), **(ii) grammaticality** (additionally its contextual **appropriateness** in Exp2), and **(iii) social meaning** relating to speaker properties in nine dimensions (low/high socioeconomic status —SES, low/high education, in/formal, im/polite, obedient/rebellious, un/cool, cold/warm, un/friendly, un/confident), all on a 7-point Likert scale (1-7 for low-high). **Exp1 – without context (subjects: N=101)** used a 2x2 design with the factors NUMBER (MC vs. SM) and FORCE (P vs. N). **Exp2 – with context (subjects: N=160)** used a 2x2x2 design with a third factor CONTEXT (distant vs. close). We computed ordinal models for the ratings of each question separately (see Figure 1); p-values were obtained using log-likelihood ratio tests.

(1) (S1-Exp1) Somebody says: ..
(S1-Exp2) A man talks to his {**boss**$_{distant}$ / **mother**$_{close}$}: ..
(S2) "I {**may possibly**$_{MC}$/**may**$_{SM}$}$_{\Diamond}$ / {**must certainly**$_{MC}$/**must**$_{SM}$}$_{\Box}$ have lost my keys."
(Q1) Does the person believe that they have lost their keys?

Experiments – Main results. (i) interpretation: In both Exp1/2, significantly higher speaker commitment ratings were received for $\Box$ vs. $\Diamond$ and for MC vs. SM, i.e. **$\Box$>$_{Int}\Diamond$, MC>$_{Int}$SM**. Furthermore, there was a significant NUMBER*FORCE interaction with a cross-over effect: **MC$_{\Box}$>$_{Int}$SM$_{\Box}$; MC$_{\Diamond}$<$_{Int}$SM$_{\Diamond}$**. — This finding challenges the semantic equivalence assumption for MC and SM: We will leave it under-specified for now as to whether the weakening effect of *may possibly* vs. *may* and the strenghthening effect of *must certainly* vs. *must* is a semantic or pragmatic (i.e., via enriched meanings) effect. • **(ii) grammaticality/appropriateness:** In both Exp1/2, MC was rated as less grammatical (above point 4 though) than SM, and in Exp. 2 MC was rated as less appropriate: **MC<$_{G/A}$SM**. — This finding is in line with the more restricted distribution of MC vs. SM. • **(iii) social meaning:** In Exp1, MC was rated as less friendly/warm/cool/rebellious than SM. Certain measures showed a significant NUMBER*FORCE interaction: MC$_{\Diamond}$ was rated as significantly lower than SM$_{\Diamond}$ in SES/education/confidence levels; MC$_{\Box}$ was rated as more formal/confident than SM$_{\Box}$. Furthermore, MC$_{\Diamond}$ was rated as more rebellious than MC$_{\Box}$. Exp2 largely replicated the results of Exp1 — MC$_{\Diamond}$ was rated as significantly lower in SES/education/confidence levels, but as more rebellious than SM$_{\Diamond}$. MC$_{\Box}$ was rated as more formal/confident than SM$_{\Box}$. Furthermore, CONTEXT showed a significant main effect in the formality measure: distant conditions received higher ratings than close conditions. No interactions with CONTEXT were significant.

Conclusion: Our findings (i)/(iii) show that (simplifying here) weaker statements give rise to more negative perceptions and stronger ones to more positive perceptions, providing convergent evidence for the correlation between semantic (or narrow-pragmatic) meaning and social meaning. In our study, context via interlocutor relation manipulations did not have a strong influence on the perception of MC; it remains to be explored as to the effect of other situational parameters.

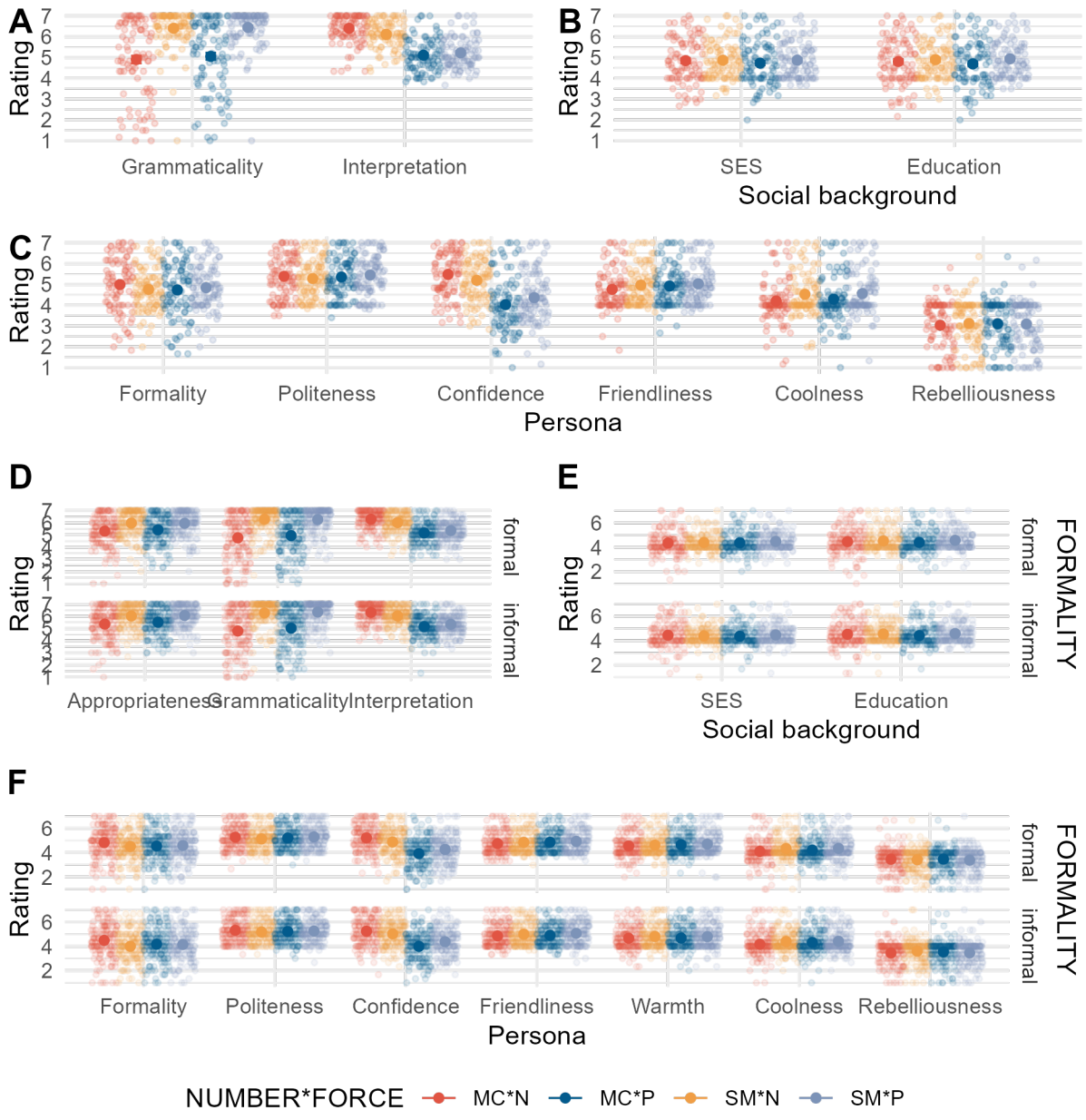


Figure 1: Means and subject means (opaque vs. transparent dots) of Exp1/2 (A/B/C vs. D/E/F).

Selected references: [1] Beltrama, A. (2020). Social meaning in semantics and pragmatics. [2] Burnett, H. (2019). Signalling games, sociolinguistic variation and the construction of style. [3] Geurts, B. and J. Huitink. (2006). Modal concord. [4] Glass, L. (2015). Strong necessity modals: Four socio-pragmatic corpus studies. [5] Pescuma et al. (2023). Situating language register across the ages, languages, modalities, and cultural aspects.

The role of definiteness in ad hoc implicatures

Summary: This study investigates how ad-hoc implicatures and the definiteness presupposition of *'the'* interact. Using a truth value judgment task (Crain & Thornton 2000), we examine whether English-speaking adults interpret "*Mary bought **a** striped sweater"* differently from "*Mary bought **the** striped sweater"* in a context where there are two possible referents, one which is best described with one adjective (e.g. *"striped"*) and the other which is best described with two adjectives (e.g. *"striped and spotted"*). Contrary to what standard models of implicature generation predict, we find that uses of *'the'* are rejected more frequently than uses of 'a' when the item bought is best described with two adjectives. This shows that the use of the indefinite blocks the generation of potential ad-hoc implicatures, which suggests that the processing of presuppositional content takes precedence over the processing of (ad-hoc) implicatures.

Ad-hoc implicatures and reference disambiguation: Under standard accounts of meaning enrichment, ad-hoc implicatures (Hirschberg 1991) are generated by treating a contextually provided alternative as false. When p is used in a context where $p \wedge q$ is a relevant alternative, an implicature that $\neg(p \wedge q)$ is generated. For example, in a context where there is a person with glasses and a person with both glasses and a hat, adults and even preschool-aged children interpret *"My friend has glasses"* as referring to the person with only glasses (Stiller et al. 2015). The ad-hoc implicature appears to disambiguate the two possible referents who both match the literal interpretation of *"My friend has glasses."* Note, however, that the denotation of *"my friend"* independently implies there is a unique relevant friend being described.

Manipulating uniqueness: In this study we test whether ad-hoc implicatures provide reference disambiguation when the description of the possible referent doesn't imply uniqueness. We compare how the definite article, *'the'*, and the indefinite article, *'a',* are interpreted when two contextually provided referents match the literal denotation of the NP. In a scenario where there is a sweater with stripes and a sweater with both stripes and spots (Figure 1), we assess how adult English speakers interpret (1) and (2):

Figure 1. Critical target image, paired with (1) in the definite condition and (2) in the indefinite condition.

(1) Mary bought **the** striped sweater.
(2) Mary bought **a** striped sweater.

Experiment: *Participants:* 60 English native speakers were recruited through Prolific and randomly assigned to either the *'a'* or *'the'* condition. Participants were paid at an average pay rate of £11.75/hour for the task, which took on average 6m9s to complete. *Procedure:* The task was a truth value judgment task, implemented and hosted on Qualtrics. Participants were given a back story about characters who were shopping at the store. On each trial, they saw a picture containing three items, and a shopping basket under one of the items. A puppet named Raffie described which item the character purchased (using either a definite or an indefinite description), and participants had to indicate whether Raffie was right or wrong by clicking on 'Yes' or 'No'. *Materials:* Noun phrase type (definite *'the'* vs. indefinite *'a'*) was a between-subject variable. Critical target trials involved weak/under-informative descriptions containing one adjective, such as *"Mary bought a/the striped sweater"*, to describe a context in which there was both a sweater with stripes and a sweater with stripes and spots, and Mary had bought the one with stripes *and* spots (see Figure 1, paired with (1)/(2)). If participants computed the ad-hoc implicature that the sweater Mary bought didn't contain spots, they were expected to reject the test sentence; if not, they would accept the test sentence on its literal meaning. The experiment also included

unambiguously true and unambiguously false 1- and 2-adjective controls, in which the test sentences were clearly true or clearly false descriptions of the purchased item (see Figures 2 and 3 for examples). We also included clearly true and clearly false filler items which involved descriptions that did not contain any adjectives. In all, each participant saw 2 training items, followed by 30 test items: 12 ambiguous target trials containing either *'a'* or *'the'*, 6 clearly true/clearly false 1-adjective controls, 6 clearly true/clearly false 2-adjective controls, and 6 adjective-less fillers. Order of presentation was randomized across participants.

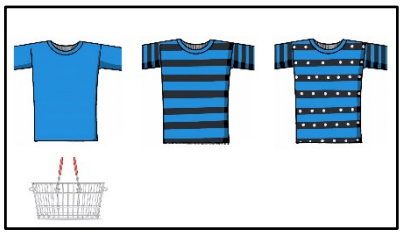


Figure 2. Clearly true control image paired with the sentence: *'Max bought a/the plain shirt'.*

Figure 3. Clearly false control image paired with the sentence: 'Ellie bought a/the rainbow-coloured and polka-dotted dress.'

Results: One participant was excluded for failing to score at least 12/18 (two thirds) accuracy on the unambiguous control and filler trials, leaving a total of 59 participants for analysis (29 in the *'a'* condition and 50 in the *'the'* condition). For these participants, accuracy was above 95% for all unambiguous filler and control conditions. Figure 4 displays the average proportion of *yes*-responses in the target *'a'* and *'the'* conditions (dots represent individual participant means). Mean acceptance in the indefinite *'a'* condition was 93%, compared with 55% in the definite *'the'* condition. We fitted a mixed effect logistic regression model on responses to the target conditions, with definiteness as a fixed effect, and random effects for subject and item. Model comparisons revealed a significant effect of definiteness ($\chi2(1)=15$, $p<.0001$), with participants more likely to reject the underinformative target statements when it contained the definite article *'the'*.

Discussion: The indefinite article was more referentially ambiguous than the definite article. Participants accepted *'a'* more often than *'the'* in scenarios where two referents matched the literal (at-issue) description of the purchased object. Standard accounts of implicatures predict no difference in how (1) and (2) disambiguate the object NP. Both have a contextually provided alternative, *"the/a striped and spotted sweater"*, and negating these alternatives should pick out the sweater with stripes and no spots. Our findings suggest that computing presuppositional content blocks the generation of the ad-hoc implicature that would disambiguate the referent. This could be because *'a'* carries a non-uniqueness presupposition (Hawkins 1978), or because listeners are sensitive to why terms with stronger presuppositions are avoided (Heim 1991). Either way, presuppositional processing takes precedence over implicature generation.

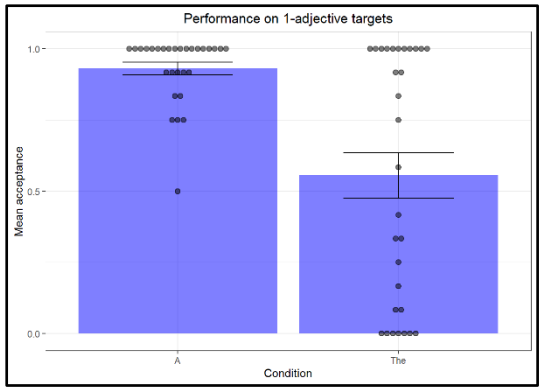


Figure 4. Performance on critical 'a' and 'the' targets.

References: Crain, S., & Thornton, R. (1998). *Investigations in universal grammar: a guide to experiments on the acquisition of syntax and semantics*. (Language, speech, and communication). • **Hawkins J. (1978)**. Definiteness and indefiniteness: a study in reference and grammaticality prediction • **Heim, I. (1991).** Artikel und Definitheit [Articles and definiteness]. In A. von Stechow & D. Wunderlich (eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*. • **Hirschberg, J.L. (1991).** *A theory of scalar implicature*. • **Stiller, A.J., N.D. Goodman & M.C. Frank. (2015).** Ad-hoc implicature in preschool children. *Language Learning and Development* 11.

Ordering is not ranking: A study of ordinals vs. degree modifiers in nested definites

This study probes how the semantics of ordinals relates to the semantics of comparatives and superlatives. We examine this question with the help of a picture task in which participants are asked to locate objects described by nested descriptions like *the candle on the first/closer/closest table*, with an ordinal, comparative or superlative modifier in the inner noun phrase. We show that ordinals systematically lack the 'relative readings' (as we call them) first observed by Haddock (1987) for unmodified nested descriptions like *the rabbit in the hat*, in which the inner definite is understood with enriched content, as in *the rabbit in the hat with a rabbit in it*. As Bumford (2017) observes and explains via scope movement, nested descriptions with superlatives like *the rabbit in the biggest hat* have relative readings too, in this case paraphrasable as *the rabbit in the biggest hat with a rabbit in it*. This present study shows that superlatives and ordinals differ in their propensity to give rise to such readings (and comparatives easily allow them).

The differences we observe are in line with prior work showing differences between ordinals and superlatives (Bylinina et al., 2014). However, the results present difficulties for accounts of the semantics of ordinals on which they are entirely parallel to (Bhatt, 2006) or contain superlatives (Alstott, 2023). Such accounts would predict relative readings with both ordinals and superlatives in nested descriptions, *contra* what we found in the experiments we will report. We discuss two strategies for explaining the contrast, one building on Bylinina et al.'s idea that ordinals do not undergo scope movement, and another building on the idea that ordinals depend on a contextually salient linear ordering with a basis that is preferably iconic to the natural numbers.

In both of our experiments, participants were presented with displays involving objects placed on a sequence of locations. In Experiment 1 this was a series of tables (see Figure 1); in Experiment 2 it was a series of stairs (Figure 2). Relative to the same display throughout the experiment, participants were asked a series of questions like *What's next to the cat on the closest table?* Participants were instructed to write "doesn't make sense" if the question does not make sense. All target trials were set up so that a relative reading would be the only one available, given the display. Rejection ("doesn't make sense") thus signalled the absence of a relative reading.

Prompts varied in the **number of objects** described by the relevant noun (e.g *cat*): 2 or 3. The **type of modifier** could be either ORDINAL (e.g. *first*) or DEGREE (comparative like *closer* or superlative like *closest*). In the DEGREE condition, the modifier was comparative in the case of two objects, and superlative in the case of three objects, as comparatives are more felicitous than superlatives with comparison classes of size 2. Our main focus is on nested descriptions containing modifiers in the EMBEDDED noun phrase, as in *What's next to the cat on the closest table?* but as a control, we included **construction**s where the modifier appears in the MATRIX position within the noun phrase, as in *What's on the closest table with a cat on it*, where the complement of the adjective is explicitly restricted by information from the noun. Two items were constructed for each of the 8 conditions, and participants saw all 16 items. Order w.r.t. both modifier type and sentence type was counterbalanced across lists, and fillers were evenly interspersed with target trials. For both experiments, 40 native speakers of English were recruited via Prolific (different groups of 40).

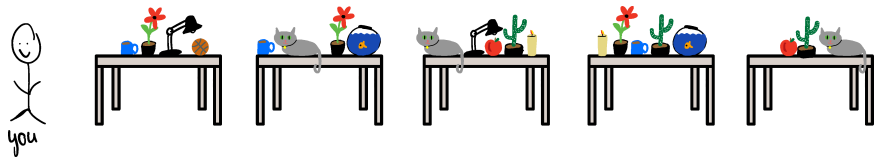


Figure 1: Exp. 1 display

The results of Experiments 1 and 2 are shown in Figure 3. We found the same pattern in both experiments. With the modifier in MATRIX position (*first table with a cat*), there was almost no rejection. A strong majority of respondents rejected relative readings for nested descriptions with ORDINAL modifiers in the EMBEDDED position (*cat on the first table*), as shown in Figure 3. Relative readings for nested descriptions containing DEGREE modifiers were sometimes rejected, but significantly less often than with ORDINALS. Interestingly, rejection was significantly more common with superlatives than with comparatives; we suspect that this is due to the absence of a competing absolute reading with comparatives.

Figure 2: Exp. 2 display

We conclude that ordinals are substantially less susceptible to relative readings than degree modifiers, in nested descriptions. One strategy for explaining this result is to adopt Bylinina et al.'s stipulation that ordinals cannot undergo scope movement, made in order to explain the absence of 'upstairs *de dicto*' readings with ordinals. This assumption alone does not suffice to block relative readings, though, because in order to generate focus-related relative readings of ordinals as in Bhatt's (2006) *John$_F$ gave the first telescope to Mary*, Bylinina et al. assume that ordinals expect an implicit comparison class. So one would need a theory of why the comparison class argument of *first* in *the cat on the first table* cannot be set to 'with a cat on it'.

Our explanation relies on the familiar idea that an ordinal expects an ordering that can be provided by context. The ordering is a function f from a 'basis' to satisfiers of the modified predicate. The basis is a linearly ordered set like a sequence of times (as in *second train*) or locations (*second stair*). The *nth table* is the nth object in a sequence $\langle f(i_1), f(i_2), f(i_3), \ldots \rangle$. We posit further that the more iconic a sequence is to the natural numbers, the more accessible it is as a basis for the ordering. The more evenly spread out a sequence is, as measured by a perceptually salient distance metric, the more iconic it is to the natural numbers. In our experiments, the sequence of locations corresponding to the full set of tables is more iconic to the natural numbers than the sequence over the subset containing cats. The highly iconic basis fixes the reading of an embedded ordinal to be absolute (low scope), even on pain of global reference failure. Superlatives do not rely on a linear ordering and therefore have a more flexible range of scope options.

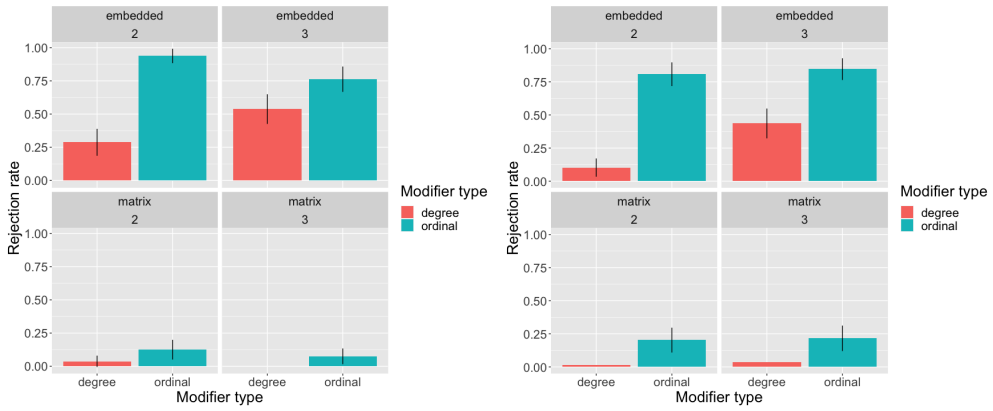


Figure 3: Results of Experiments 1 (left) and 2 (right). Error bars show 95% CI.

References. Alstott, J. 2023. Ordinal numbers: Not superlatives, but modifiers of superlatives. *SALT 33*. • Bhatt, R. 2006. *Covert modality in non-finite contexts*. • Bumford, D. 2017. Split-scope definites: Relative superlatives and Haddock descriptions. *L&P*. • Bylinina, L. et al. 2014. A non-superlative semantics for ordinals and the syntax and semantics of comparison classes. • Haddock, N. 1987. Incremental interpretation and CCG. In *Proc. IJCAI 10*.

Conceptual Signatures of Atomicity Across Languages

Logico-semantic theories suggest that *atomicity* underlies the representation of both telicity in the semantics of verbal predicates and the mass/count distinction in the semantics of nominals ([1]-[3]; cf. [4]-[7]). It is plausible that atomicity has a counterpart in non-linguistic cognition: atomicity for temporal entities would underlie the distinction between bounded events whose representation includes inherent endpoints and unbounded events whose representation lacks such boundaries [8]. Similarly, atomicity for spatial entities would underlie the distinction between objects that possess inherent spatial boundaries and substances that lack such boundaries [9]. Here we aim to **(a)** uncover the non-linguistic features that could provide the basis for conceptual atomicity (bounded events and objects) across the domains of temporal and spatial entities and **(b)** test whether these conceptual features precede the linguistic encoding of boundedness and objecthood.

We propose that a well-defined internal structure is a distinguishing feature of atomicity (see also [7], [9]). Two predictions follow from this hypothesis: (1) No Restructuring: viewers should be more sensitive to structural changes to atomic entities (bounded events and objects) than to non-atomic entities (unbounded events and substances); (2) Distinct Parts: subparts of atomic entities should be more likely to be perceived as distinct from one another than subparts of non-atomic entities. We test these predictions in Experiments 1 (1a: events, 1b: objects) and 2 (2a: events, 2b: objects) respectively, across English- and Mandarin-speaking adult participants.

We also test how conceptual representations of atomicity arise in the mind. We hypothesize that conceptual atomicity precedes and structures the linguistic encoding of atomicity. Alternatively, the conceptual signature of atomicity might arise because of familiarity with the way atomicity is encoded in the viewer's language. Only the first hypothesis predicts that non-linguistic atomicity would be conceptualized in similar ways cross-linguistically.

We compare speakers of Mandarin Chinese and English because the two languages differ in the linguistic encoding of boundedness and objecthood. While English speakers can use different predicates (e.g. *fix/drive a car*) to denote boundedness contrasts, in Mandarin, mono-morphemic verbs (e.g. *kai* "drive") are generally inherently unbounded ([10], [11], [12]). In the nominal domain, while English speakers can specify objecthood in language via count/mass syntax (*a vase/clay*). Mandarin lacks count-mass syntax, thus all nouns can appear in their bare form ([13]).

No Restructuring *Experiment 1a (Events)*: We created 16 videos of bounded events (e.g. cutting the paper in half) and 16 videos of closely related unbounded events (e.g. cutting pieces from the paper). We confirmed that naïve viewers construe the videos along these lines in a prior norming study in which people were asked if the event "had a beginning, midpoint and endpoint" (M=90% vs. 17.5% for bounded vs. unbounded events). Each video was edited so that the temporal order of the second and third quarters of the video was flipped. Participants (English N=24; Mandarin N=24) watched the original video followed by the restructured video, and were asked to decide whether the two videos were identical. English-speaking participants were more likely to accurately judge the original video and the structurally disrupted video as different for Bounded Events (M=77.7%) than for Unbounded Events (M=60.9 %) (*glmer*, $p<0.001$), as were Mandarin-speaking participants (Bounded M=74.8%, Unbounded M=68.6%, $p<0.05$). As expected, both groups of participants were better at detecting structural disruptions to bounded events than to unbounded events.

Experiment 1b (Objects): We used 16 pairs of object (e.g. vase) and substance (e.g. clay) images, which were confirmed to be construed along these lines (1-7 scale; 1=object, 7=substance;

M=2.91 vs. 4.81 for objects and substances, respectively). We created structurally disrupted versions of each entity by flipping the order of the second and third vertical quadrants of the image (Table 1). Participants (English N=24; Mandarin N=24) were briefly (100ms) shown the original entity, followed by the structurally disrupted entity (100ms). They were asked to identify whether the two entities were identical. English speakers were more likely to accurately judge the original entity and the structurally disrupted entity as different for Objects (M=87.8%) than for Substances (M=58.6%) ($p<0.001$), as were Mandarin-speaking participants (Objects M=87%, Substances M=67.6%, $p<0.001$). Taken together, Experiments 1a and 1b show that regardless of one's native language, the cognitive system is better at detecting structural disruptions to atomic entities than to non-atomic entities.

Distinct Parts *Experiment 2a (Events)*: We segmented each original video from Experiment 1a into nine temporal segments, and used the fifth and the eighth segments (roughly, a middle and close-to-the-boundary segment). Participants (English N=24; Mandarin N=24) watched each segment and were asked to decide whether the two videos were identical or not. As expected, English-speaking participants were more likely to accurately identify the two segments as distinct for Bounded (M=71.4%) than for Unbounded events (M=66.5%) ($p<0.05$), as were Mandarin-speaking participants (Bounded M=57.5%, Unbounded M=53.1%, $p<0.05$).

Experiment 2b (Objects): Using the 16 original images used in Experiment 1b, we took two different segments from each image. One segment was cropped at the center, and another segment was cropped at the top right corner (again, a middle and close-to-the-boundary segment). Participants (English N=24; Mandarin N=24) saw each subpart and were asked to identify whether the two segments they saw were identical. As expected, English-speaking participants were more likely to accurately identify the two subparts as distinct for Objects (M=73.6%) than for Substances (M=54.4%) ($p<0.001$)), as were Mandarin-speaking participants (Objects M=78.6%, Substances M=57.1%, $p<0.001$). Again, Experiments 2a and 2b show that regardless of one's native language, the cognitive system is more likely to perceive two subparts of atomic entities as distinct from one another than subparts of non-atomic entities.

Discussion Together, these results throw light onto the nature of entity categories in the human mind: both English-speaking and Mandarin-speaking viewers process atomic and non-atomic entities differently, with only the former having a well-defined (temporal/spatial) structure with integrally-ordered, distinct parts. We propose that these key conceptual characteristics organize atomicity and can be used to individuate entities. These features of non-linguistic atomicity are potentially universal and are conceptualized in similar ways cross-linguistically. Furthermore, these conceptual features can be used to map entity concepts onto foundational semantics in natural language.

Table 1. Sample entity images (Exp.1b)

condition	original	structurally disrupted
object		
substance		

References [1] Bach 1986. *Ling&Phil.* [2] Jackendoff 1991. *Cognition.* [3] Taylor 1977. *Ling&Phil.* [4] Champollion 2015. *Theoretical Linguistics*. [5] Champollion 2017. OUP. [6] Filip 2012. OUP. [7] Wellwood et al. 2018. In *Oxford studies in experimental philosophy.* [8] Ji & Papafragou 2020. *Cognition*. [9] Prasada et al. 2002. *Cognition.* [10] Lin, 2004. *MIT.* [11] Sybesma, 1997. *Journal of East Asian Linguistics*. [12] Tai, 1984. *CLS.* [13] Chierchia, 1998. *Events and Grammar.*

Putting donkeys into context

Overview. Recent competing theoretical accounts of donkey sentences rest on claims about what readings are available in different contexts, when different types of determiner are involved. That donkey sentences may be open to more than one reading is widely acknowledged. Specifically, a sentence like (1) may be understood with a so-called Universal (U-) reading, requiring each girl who baked a cake to have iced all of the cakes they baked. Alternatively, it may have an Existential (E-) reading, requiring each girl who baked a cake to have iced some of those cakes.

(1) Every girl who baked a cake iced it.

A long tradition sees this U-/E- ambiguity as related to a similar ambiguity with plural definite descriptions (see [1]). Recent proposals regarding donkey sentences, [2,3], follow distinct proposals for plural definites, [4,5]. [2] follows [4] in setting out a trivalent approach, while [3] follows [5] in explaining U-readings as resulting from a form of (obligatory) scalar strengthening on the quantifier's scope. According to each account, the canonical reading for (1) accounts for U-reading intuitions; but each allows for a mechanism which accommodates E-readings in contexts where, for example, the Question partition locates states as depicted in Fig.1 in the same cell as states that support the U-reading - as per our Forbidden contexts in Exp.1a,b.

The key difference between accounts lies in proposals about quantifiers that have different monotonicity properties. Specifically, based on the fact that SI strengthening tends to not occur in DE contexts, [3] predicts an asymmetry in the availability of U-readings for (1) vs. (2):

(2) No girl who baked a cake iced it.

[2] assumes that U-readings for (2) are in principle available when the QuD locates states as in Fig. 2 in the cell that supports their assumed default E-reading, as in Obligatory contexts below. In addition, according to both [2] and [3], donkey sentences with positive existential determiners, as in (3), should prefer U-readings. However, it has been argued, based on introspection, that such readings are hardly available. Accordingly, [2-3] have proposed separate mechanisms which may explain the apparent lacuna. Finally, these proposals treat singular donkey sentences (as in (1-3)) on a par with plural versions (where the indefinite and pronoun are plural), whilst previously it was argued that singular donkey sentences may not be assimilated to plural, [9]. Experiments 1a,b present sentences like (1-3) in contexts which test these proposals.

(3) More than two girls who baked a cake iced it.

Experiment 1a,b: N=200. Our innovation on recent donkey studies, e.g. [6-8], was to manipulate context – obligatory vs. forbidden. These contexts are illustrated below. For (1) and (3), [2] predicts more False responses to Fig.1 in Obligatory than Forbidden contexts. Regarding (2) and Fig. 2, the target scenario would be assimilated to the same cell as the biased E-reading in the Obligatory context, meaning more False responses for Forbidden than Obligatory. In Exp.1, we manipulated Context and Form (singular, plural) between groups, with Determiner (every, no – Exp.1a; every, more than two – Exp. 1b) within group. We used 4 scenarios (baking/icing cakes; building/painting trains, etc.). Presentation was blocked by scenario with each block introducing a context rule, on which participants were tested on during the course of the block. 3 donkey sentences (T/F/target) + 6 non-donkey filler per block.

Obligatory	**Forbidden**
The teacher told the girls that they can make cupcakes or cookies. ***But cupcakes must be iced for the presentation the next day***	*The teacher told the girls they can make cupcakes or cookies.* ***But cupcakes should not be iced because they need plain cakes for activities the next day.***

Results:For determiner 'every', analyses demonstrate a clear U/E ambiguity, also clear effects of Context, and Number on Target outcomes. For 'no' we see no such effects. In a follow up replication of Exp.1a allowing participants to express judgements with a Likert scale, rather than a binary

judgement, we found a very small effect of context in the 'no' plural condition. Exp.1b established evidence for a U/E ambiguity for 'more than two' as well as an effect of context. Post-hoc, we observed a difference in rates of U-readings for 'every' when blocked with different determiners ('no' vs. 'more than two') s.t. rate of E-readings is greater in Exp.1b. This 'priming effect' was confirmed in a follow-up study looking at singular donkey sentences only, without context.

Discussion: U- and E-readings for 'every' donkey sentences have previously been shown to be available, [6-8], and this is replicated here. We also show a predicted effect of context. Previous verification tasks have not provided any evidence for U-readings for positive existential quantifiers, but here we find evidence for these with context. While effects of context are demonstrated for both these determiners we find none for 'no'; also, replicating [6,7] we fail to detect any robust U-readings for 'no'. The asymmetry between positive and negative quantifiers in context effects is challenging for [2] and more in line with the pattern assumed in [3]. As for the clear effect of number, with singular versions eliciting more U-readings, and the 'priming' effect of determiners on U-/E-readings for 'every', these are not readily predicted by either account. These outcomes will be taken up in discussion.

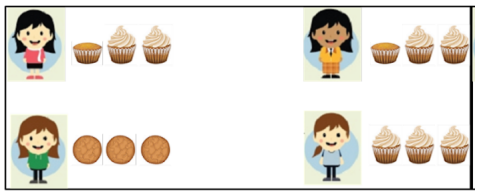


Figure 1: Target condition for *Every*

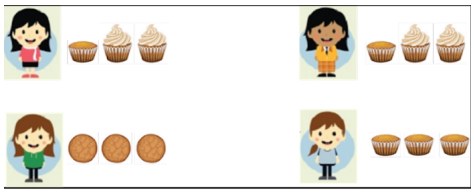


Figure 2: Target condition for *No*

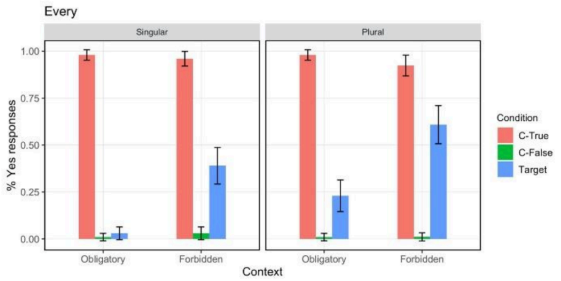


Figure 3: Exeriment 1a rates for *Every*

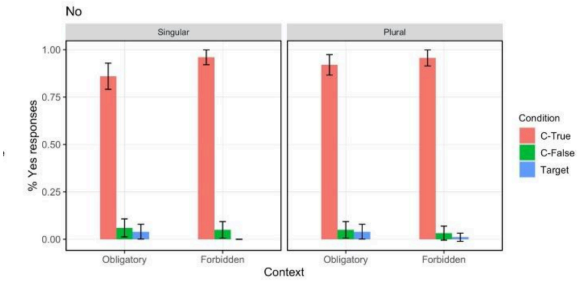


Figure 4: Expt. 1a rates for *No*

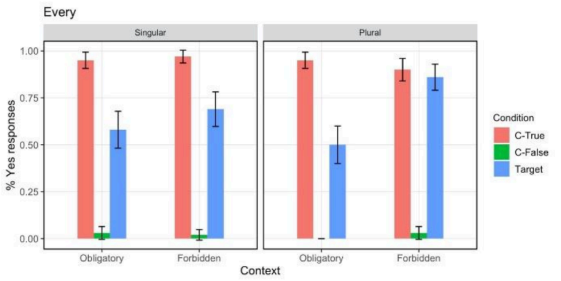


Figure 5: Expt. 1b rates for *Every*

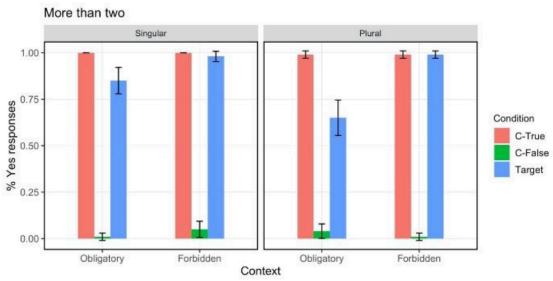


Figure 6: Expt. 1b rates for *More than 2*

References: [1] Krifka (1996) SALT; [2] Champollion et al. (2019) S&P; [3] Chierchia (2022) J.Sem; [4] Kriz (2016) J.Sem; [5] Bar- Lev (2020) L&P; [6] Geurts (2002) L&P. [7] Sun et al (2020) Sinn&B; [8] Denic & Sudo, (2022) JSem; [9] Kanazawa (1994) L&P.

Less-comparatives must be less ambiguous than *exactly*-differentials, experimental data shows

Introduction: According to [1] and contrary to previous conclusions, e.g. in [2], scope mobility of comparative operators does after all surface in a narrow class of cases where intensional verbs are combined with *less*-comparatives or *exactly*-differentials, as in (1). According to this view, (1) has the two readings in (1-a/b) and its less prominent, inverse scope reading in (1-b) imposes no upper limit on the paper's length but only a minimal requirement of 15 pp. Though not uncontroversial and dependent on notoriously subtle judgments, this type of ambiguity influenced subsequent compositional analyses of comparatives (e.g. [3–5], among many others) and was also used as a diagnostics for scope mobility of the comparative operator in cross-linguistic studies, e.g. [6]. We use judgment data from forced-choice experiments to empirically test the availablity of this ambiguity in German and we discuss theoretical implications of our findings.

(1) (This draft is 10 pages.) The paper must be exactly 5 pages longer than that.
   a. linear scope: $\forall w \in Acc : max\{d : long_{\text{w}}(p, d)\} = 15pp$
   b. inverse scope: $max\{d : \forall w \in Acc : long_{\text{w}}(p, d)\} = 15pp$

Methods: We conducted two web-based questionnaire studies, Exps.1 & 2, recruiting participants via `prolific.co`. 12 German items were constructed as exemplified in (2) and (3). All items start with a sentence in which gradable adjectives (e.g, 'long') are degree-modified by exactly-differentials (e.g. 'exactly 10 pages longer than ...') or *less*-comparatives (e.g. 'less long than ...'). In half of the conditions (e.g. (2-a/c)), comparatives are combined with the modal verb 'must'. By hypothesis, the presence of 'must' should cause the purported ambiguity to emerge. Sentences without modals (e.g. (2-b/d) in Exp.1) were used as unambiguous controls against which responses to the modal conditions can be compared. All of these sentences are followed by the same short post-context sentence (also illustrated in (2)). Each sentence doublet is, furthermore, paired with yes-no comprehension questions as shown in (3). There are two types of questions: *Matching* questions probe for the preferred or (in case of the controls) only possible reading, whereas the *mismatching* questions ask about propositions that are incompatible with the preferred readings and would thus receive a "no"-response if this was the only available reading (pairing indicated by the labels in (3); e.g. (2-a) is paired with the matching question in (3-a-m) and mismatching question in (3-a-mm)). Altogether, we thus manipulated the factors MODIFIER TYPE (*exactly* vs. *less*), MODAL (*absent* vs. *present*) and QUESTION (*match* vs. *mismatch*), yielding eight conditions in a $2 \times 2 \times 2$ design. The complete set of experimental items comprised 96 pairs of assertions and questions distributed (together with 48 fillers) over eight lists using a Latin square. Exp.2 was a follow-up, in which *exactly*-controls of Exp.1, e.g. (1-b), were also put into comparative form, e.g. (2-e), because the positive form in Exp.1 led to almost flawless performance, complicating the interpretation of the results. Exp.2 was thus a quasi-replication testing whether the results of Exp.1 reflect differences between the two types of comparatives or were due to characteristics of the controls.

Results: After applying predefined exclusion criteria, data from 62 and 65 (out of 87 and 87) participants were passed on to the statistical analysis of Exps.1 & 2, resp. Although the comparative *exactly*-controls in Exp.2 did in fact lead to a few more errors as compared to Exp.1, as we expected, the general pattern of results was the same in both experiments. Descriptive results are shown in Figure 1. In the modal conditions, questions matching the preferred linear scope interpretation were overwhelmingly answered with "yes" (Exp.1: 94.1%; Exp.2: 88.6%) and mismatching questions with "no" (Exp.1: 83.3%; Exp.2: 91.3%). A logit mixed effects model analysis revealed a significant three-way interaction in both experiments (Exp.1: $z = -2.99, p = .003$; Exp.2: $z = -2.05, p = .041$) which we resolved by conducting separate analyses for the two modifier types. In the *exactly*-conditions, we found the predicted MODAL $\times$ QUESTION interaction (Exp.1: $z = 3.21, p = .001$; marginal in Exp.2: $z = 1.58, p = .064$), whereas no such interaction was found in the *less*-comparatives (Exp.1: $z = -0.30, p = .767$; Exp.2: $z = -0.9, p = .37$).

Discussion: Across both experiments, indication of the purported ambiguity was limited to *exactly*-differentials. We suggest that our results might best be explained by accounts that derive ambiguity from properties of the measure phrases in *exactly*-differentials. For example, the proposal of [7] accounts for the ambiguity in (1) in terms of scope mobility of the measure phrase rather than the comparative operator (contrary to,e.g. [1, 4, 5]). To rule out unattested scope interaction, e.g. with nominal quantifiers (cf. [1, 2, 5, 8]), it can be complemented along the lines of [5] (following [9]) by restrictions that derive from the underlying algebraic structure of degrees. Our data on German invites deliberation about cross-linguistic variation in this sphere. We thus also collected data on the ambiguity in English and addressed potential criticisms of our German data (replaced definite descriptions in the *than*-clause (e.g. *the draft*) with a demonstrative (e.g. *that*; cf. (1)), removed *also* from the comprehension questions in (3-a-mm,c-mm) and embedded items in contexts supportive of the inverse reading). The results for German are replicated in English.

(2) Target sentences and post contexts

a. Das Papier muss genau 10 Seiten länger sein als der Entwurf. So lautet die Vorgabe der Zeitschrift.
The paper must exactly 10 pages longer be than the draft. So sounds the guideline of_the journal
'The paper is required to be 10 pages longer than the draft. That's what the journal's guideline says.'

b. Das Papier ist genau 10 Seiten lang. Das haben die Autoren gesagt.
The paper is exactly 10 pages long. That have the authors said
'The paper is exactly 10 pages long. That's what the authors said.'

c. Der Entwurf muss weniger lang sein als das Papier. So lautet die Vorgabe der Zeitschrift.
The draft must less long be than the paper. So sounds the guideline of_the journal
'The draft is required to be less long than the paper. That's what the journal's guideline says.'

d. Der Entwurf ist weniger lang als das Papier. Das haben die Autoren gesagt.
The draft is less long than the draft. That have the authors said
'The draft is less long than the paper. That's what the authors said.'

e. Das Papier ist genau 10 Seiten länger als der Entwurf. Das haben die Autoren gesagt.
The paper is exactly 10 pages longer than the draft. That have the authors said
'The paper is exactly 10 pages longer than the draft. That's what the authors said.'

(3) Comprehension questions (no translation provided if identical to gloss)

a-m Soll das Papier 10 Seiten länger sein als der Entwurf?
Should the paper 10 pages longer be than the draft
'Should the paper be 10 pages longer than the draft?'

a-mm Darf das Papier auch 15 Seiten länger sein als der Entwurf?
May the paper also 15 pages longer be than the draft
'Is the paper also allowed to be 15 pages longer than the draft?'

b-m Ist das Papier 10 Seiten lang?
Is the paper 10 pages long

b-mm Ist das Papier 14 Seiten lang?
Is the paper 14 pages long

c-m Soll der Entwurf kürzer sein als das Papier?
Should the draft shorter be than the paper
'Should the draft be shorter than the paper?'

c-mm Darf der Entwurf auch länger sein als das Papier?
May the draft also longer be than the paper
'Is the draft also allowed to be longer than the paper?'

d-m Ist der Entwurf kürzer als das Papier?
Is the draft shorter than the paper

d-mm Ist der Entwurf länger als das Papier?
Is the draft longer than the paper

e-m Ist das Papier 10 Seiten länger als der Entwurf?
Is the paper 10 pages longer than the draft

e-mm Ist das Papier 14 Seiten länger als der Entwurf?
Is the paper 14 pages longer than the draft

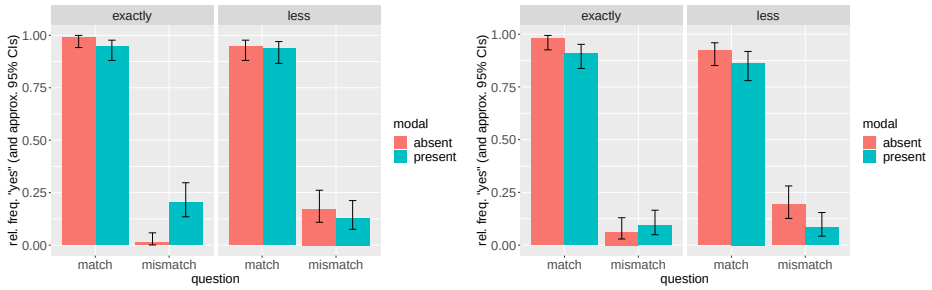


Figure 1: Relative frequency of "yes" responses across conditions in Exps.1 (left) and 2 (right).

Selected References: 1. Heim, I. *Degree operators and scope* in *Proceedings of SALT 10* (2000), 40–64. 2. Kennedy, C. *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison* (University of Chicago, 1997). 3. Bhatt, R. et al. Late Merger of Degree Clauses. *Linguist Inq* **35,** 1–45 (2004). 4. Breakstone, M. Y. et al. *On the Analysis of Scope Ambiguities in Comparative Constructions: Converging Evidence from Real-Time Sentence Processing and Offline Data* in *Proceedings of SALT 21* (2011), 712–731. 5. Lassiter, D. *Quantificational and modal interveners in degree constructions* in *Proceedings of SALT 22* (ed Chereches, A.) (2012), 565–583. 6. Beck, S. et al. Crosslinguistic Variation in Comparison Constructions. *Linguistic Variation Yearbook* **9,** 1–66 (2009). 7. Oda, T. *Degree constructions in Japanese* (University of Connecticut, 2008). 8. Beck, S. DegP scope revisited. *Nat Lang Seman* **20** (2012). 9. Szabolcsi, A. et al. Weak islands and an algebraic semantics for scope taking. *Nat Lang Seman* **1,** 235–284 (1993).

Parenthesized Modifiers in English and Korean: What They (May) Mean

Existing work on parentheticals has focused on their use as appositives, speaker-oriented adverbials, and expressives (McCawley 1982; Ziv 1985; Potts 2002; Dehé & Kavalova 2007). However, there is little work on parentheticals that are marked with parentheses: in Kaltenböck (2007)'s survey of English parentheticals, none contain parentheses. Moreover, work on parentheticals outside of Indo-European is relatively scarce. Our work contributes new cross-linguistic evidence about the meaning contribution of parentheses in one key construction.

We present results from an experiment comparing the interpretation of parenthesized modifiers in Korean[1] and English, manipulating syntactic position and modifier properties (scalar/non-scalar, categorical/continuous). We focus on a parenthesized construction, shown in (1), which in English gives rise to an implication that its non-parenthesized counterpart does not.

(1) Sam studies linguistics for (intellectual) profit. #And actual profit.

(2) Sam studies linguistics for intellectual profit. And actual profit. (Lewen & Anderson 2022)

Lewen & Anderson (2022) refer to the above construction as a *restricted parenthesized parenthetical* and show that it behaves differently than better-studied parentheticals like appositives: not only is its parenthesized content integrated into the host, but it also gives rise to the implication shown in (1) and (2). They posit that the parentheses function as a focus-sensitive operator, invoking and negating a set of alternatives to the parenthesized content. In this paper, we test their hypothesis experimentally. We explore how the semantic properties of the parenthesized modifier affect how alternatives are negated in English, and compare to a language with different conventions on use of parentheticals: Korean.

A key difference between Korean and English is that in Korean, the parenthesized parenthetical can come on either side of the modified noun, as in (3) and (4). Do these syntactic structures correspond to different meanings? Preliminary native speaker judgements suggest that Korean readers may parse the parenthesized parenthetical in (3) as a non-exhaustive example of the kind of gain, while in (4), the parenthetical adds emphasis: i.e., *Sarah hopes to acquire some gain, especially intellectual gain.*

(3) Sam-neun (cicek) iik-ul wuyhay enehak-ul kongpu-ha-p-ni-ta
sam-TOP (intellectual) gain-ACC for linguistics-ACC study-do-AH-IND-DECL

(4) Sam-neun iik-ul (cicek) wuyhay enehak-ul kongpu-ha-p-ni-ta
sam-TOP gain-ACC (intellectual) for linguistics-ACC study-do-AH-IND-DECL

English and Korean thus provide an interesting cross-linguistic comparison. We present experimental results from a study exploring 1) whether alternatives are invoked in each position and 2) what alternatives are negated. We include four categories of modifiers: (1) non-scalar and categorical (e.g. *wool* v. *cotton*), (2) scalar and categorical (*weekly* v. *monthly*), (3) non-scalar and continuous (*morning* v. *afternoon*), and (4) scalar and continuous (*warm* v. *hot*). For each condition, we present ten dialogue sets between A and B in which A presents a question, and B's response contains a parenthetical, as in (5), a scalar categorical example.

(5) A: Are you still doing a lot of volunteer work for the pet shelter?
B: I don't do as much as I used to, but I still help write their (weekly) newsletter.
Question: Which kinds of newsletters do you think B doesn't help to write?
( ) daily

[1] We use the Yale system of romanization for Korean, and the standard abbreviations for grammatical morphemes given by the Leipzig Glossing Rules, with the following addition: AH = addressee honorific.

( ) monthly
( ) Other: _____

Based on the parenthesized information, participants selected one or more options; they could also fill in an Other option. In Korean, we tested an additional manipulation of position: the parenthetical appeared either to the right or left of the modified noun. Data from 32 native Korean and 32 English speakers was collected.

In English, our results confirm Lewen & Anderson's proposal that some alternative is negated; however, we find a contrast between the Non-Scalar and Scalar conditions. In Non-Scalar conditions, participants tend to exclude both alternatives, while in Scalar conditions, they tend to exclude only one. Although we expect the strongest alternative to be excluded, we find equal selections of the weaker and stronger alternatives for Scalar Categorical items. A by-item analysis reveals that this is an effect of averaging across items: most items show a strong preference for one or the other alternative to be excluded. We posit that scale flip occurs in cases where the weaker alternative is excluded.

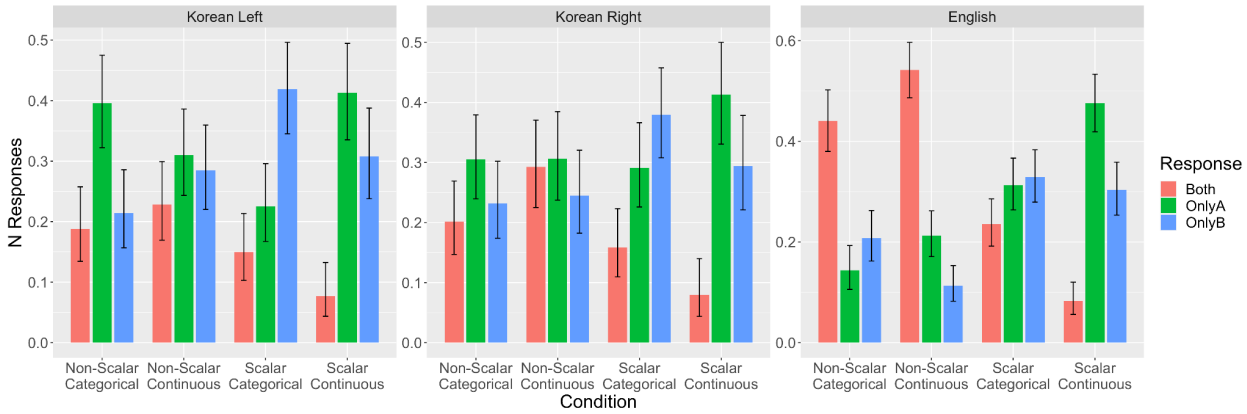


Fig. 1: Responses by Type in Main Conditions. Error bars indicate 95% confidence intervals.

In Korean, we find similar results across both position conditions, suggesting that syntactic position does not correlate with a difference in meaning. In general, Korean participants exclude only one alternative: we find no evidence of the extra implication that arises in English and leads to the exclusion of all alternatives in Non-Scalar conditions.

Our findings corroborate the richness of the (often neglected) semantico-pragmatic space of parenthesized content, and that key differences emerge across languages varying in writing systems and with differential uses in parentheses.

References ● Dehé, N. & Kavalova, Y. (2007). Parentheticals: an introduction. In N. Dehé and Y. Kavalova (Eds.), *Parentheticals*. Amsterdam: Benjamins. ● Kaltenböck, G. (2007). Spoken parenthetical clauses in English. In N. Dehé and Y. Kavalova (Eds.), *Parentheticals*. Amsterdam: Benjamins. ● Lewen, C. B. & Anderson, C. J. (2022). (Some) parentheses are focus-sensitive operators. In D. Gutzmann & S. Repp (Eds.), *Proceedings of Sinn und Bedeutung 22*. ● McCawley, J. D. (1982). Parentheticals and Discontinuous Constituent Structure. *Linguistic Inquiry* 13(1). ● Ziv, Y. (1985). Parentheticals and Function Grammar. In M. Bolkestein, C. de Groot, and J. L. Mackenzie (Eds.), *Syntax and Pragmatics in Functional Grammar*. De Gruyter Mouton.

A nonce investigation of a possible conjunctive default for disjunction

The current paper addresses the question of whether there is a conjunctive default in the interpretation of disjunction by probing into Romanian children's and adults' understanding of nonce functional words. Previewing the results, we find that, when exposed for the first time to sequences of words containing nonce connectives such as *A mo B* or *mo A mo B*, potentially corresponding to '(both) A and B' / '(either) A or B' / 'A not B' / 'neither A nor B', participants tend to associate them with a conjunctive interpretation rather than a disjunctive or negative one. Our findings suggest that a possible reason why children have been reported to interpret disjunction as conjunction in some previous studies may be the existence of a conjunctive default in the interpretation of operators linking A and B. Our findings also raise deeper questions about why speakers default to one interpretation over another, what the set of logical primitives is, and the possible role of frequency in shaping children's hypothesized meanings for logical connectives.

Background on the interpretation of disjunction Adults interpret simplex disjunction (e.g., *or) inclusively* (*The hen pushed one, possibly both*) or *exclusively* (*The hen pushed one but not both*), while they tend to associate complex disjunction (e.g., *either…or*) with exclusive interpretations [1,2]. In contrast, children, treat simplex and complex disjunctions alike, showing *inclusive*, *conjunctive* (*The hen pushed both*) or *exclusive* behavior: French and Japanese children are reportedly *inclusive* and *conjunctive* [3], while German children are *inclusive* or *exclusive* [4].

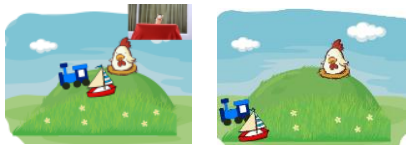
Disjunction in child Romanian Recently, this line of investigation has been extended to Romanian, which makes common use of multiple forms of disjunctions: the complex disjunction (i) *sau…sau* which is built off the simplex *sau*, and (ii) *fie…fie*, which lacks a simplex disjunctive counterpart, as well as two distinct prosodic patterns for *sau*: (iii) a neutral *sau* with no prosodic boundary after the first disjunct, and (iv) a marked *sau*, where both disjuncts are stressed. In two studies by [5], based on the design in [3], Romanian 5-year-olds were shown to be inclusive with all *sau*-based disjunctions, but conjunctive or inclusive with *fie…fie.*

The source of conjunctive interpretations in child language While children's inclusivity is typically explained as a logical interpretation of disjunction, the conjunctive interpretation of disjunction has been a matter of debate. [6,7] argue that it is merely an experimental artifact, which arises when the visual display (discourse context) only contains the objects in the disjunctive utterances. In this context, disjunction would not be informative, which is why children default to conjunction. However, [5] have shown that conjunctive behavior persists even when the background contains additional objects, casting doubt on this explanation. Alternatively, children's conjunctive interpretation is a genuine semantic-pragmatic interpretation, which may originate as a default [8], as an implicature [9], or as an additional meaning of disjunction alongside inclusivity [4]. We here focus on the conjunctive default hypothesis, probing into whether, when exposed to a connective operator unknown to them, participants default to conjunction.

Nonce words paradigms We employ a nonce paradigm. Nonce words have been employed in linguistics studies from as early as the 1950s, to probe into children's ability to learn the meanings of words by drawing on syntactic cues, also known as *syntactic bootstrapping* [10]. Brown (1957) showed experimentally that preschool-aged children could use their knowledge of different parts of speech to distinguish the meanings of nonsense words in English (*Do you see any/ a **sib**?, What is **sibbing**?*). Gleason's (1958) Wug Test used nonce words to explore children's acquisition of plural morphology (*one wug-two wugs*), possessives (*wug's, wugs'*) and verbal morphology (*He zibs*). Interesting experimental work has since ensued ([11-19], a.o.), introducing further paradigms such as the *Human Simulation Paradigm* [20], testing whether adults can infer meaning from context, and *Artificial Language Learning Paradigms* [21,22,23], testing whether adults and children can learn artificial words and what their biases are. These paradigms have been recently employed to probe into logical words such as modals [24] and negation [23].

Current experiments In our investigation, we look at what kinds of meanings adults and children ascribe to a nonce word linking A and B by using the materials in [3], originally designed to test children's interpretation of simple and complex disjunctions. We tested 21 adult native speakers

of Romanian and 17 monolingual children (3;06—5;11) on their interpretation of the nonce words *mo* and *mo…mo*. The same participants took part in the *Mo* Experiment first and the *Mo…mo* Experiment after 1 week. Following [3], we used a modified TVJT presented in Prediction rather than Description Model [9] to license *ignorance inferences*, which characterize disjunctive statements. Participants were introduced to a puppet, Bibi, who made guesses about various situations. They were told that Bibi would sometimes make use of an unknown word, and they had to decide what it meant for Bibi. Importantly, they were told that the unknown word did not refer to something that one could point to, so as not to give it a lexical meaning. Bibi would be familiarized with an animal and two objects (see Fig. 1a) and would then make a guess about what would happen (*The hen pushed the train mo the boat/ The hen pushed mo the train mo the boat*). Participants then saw the outcome (Fig. 1b) and had to say whether Bibi had guessed well. At the end of the experiment, participants were asked what they thought *mo/ mo…mo* meant. Each participant saw 15 sentences: 2 practice trials and 13 experimental items (8 targets, 2 controls, 3 fillers). *Mo/ Mo...mo* test sentences (*Găina a împins (mo) trenul mo barca* 'The hen pushed the train mo the boat') were presented in 1-disjunct-true (1DT) contexts (x4) where only one disjunct was true (*The hen pushed only the train*), and 2-disjunct-true (2DT) contexts (x4) where both disjuncts were true (*The hen pushed both objects)*. We also included false controls in which neither disjunct was true.



a. Start b. Outcome

Fig. 1. Example of a test item in 2DT

Table 1. Results

Group	Interpretation	*Mo*	*Mo...mo*
Children (N= 17)	Conjunctive	12	16
	Negative	1	0
	Mixed	4	1
Adults (N= 20)	Conjunctive	13	16
	Negative	2	2
	Mixed	5	2

Results One adult participant was excluded for failing the fillers. Like adults, children were overwhelmingly conjunctive in their interpretation of utterances containing *mo* and *mo…mo…* (i.e. accepting 2DT targets and rejecting 1DT targets). In the *Mo* Experiment, 13/20 adults and 12/17 children were conjunctive, while in the *Mo….mo…* Experiment, 16/20 adults and 16/17 children were conjunctive. The remainder either opted for a negative interpretation ('A not B' or 'neither A nor B') or oscillated between a conjunctive and a negative interpretation (Table 1).

Discussion Our results suggest that when participants are exposed to nonce words connecting A and B, they default to a conjunctive meaning. Even more strikingly, they seem to default to conjunction even in an experimental set-up where Bibi does not always make correct guesses. These findings can be interpreted in multiple ways. Under *a frequency approach*, it could be argued that participants simply associate the unknown connectors with the interpretation corresponding to the most frequent logical operator linking two elements, namely, conjunction (see [25] for a discussion of corpus evidence that conjunction is more frequent than disjunction). Under *a logical universal primitives approach*, it could be argued that conjunction is more basic than disjunction, since disjunctive interpretations can be reduced to the conjunction of two modalized elements [26]: *possible A* and *possible B*. Conjunction would also have the advantage of conceptual simplicity: (A and B) is simpler than (*possible A* and *possible B*). It is difficult to distinguish between these two approaches, given that frequency may also be a consequence of this bias. Concerning children's interpretation of disjunction, our findings suggest that a conjunctive default could be a possible source for children's interpretation of *fie…fie* as conjunctive, especially if *fie…fie* is less frequent [5], and consequently less familiar for children.

References [1] Spector 2014, [2] Nicolae & Sauerland 2016, [3] Tieu et al. 2017, [4] Sauerland & Yatsushiro 2018, [5] Bleotu et al. 2023a, b, [6] Skordos et al. 2020, [7] Huang & Crain 2020, [8] Roeper 2011, [9] Singh et al. 2016, [10] Gleitman 1990, [11] Naigles 1990, [12] Soja 1992, [13] Höhle et al. 2004, [14] Cristophe et al. 2008, [15] Syrrett et. al. 2010, [16] Yuan & Fisher 2012, [17] Jin & Fisher 2014, [18] Cao & Lewis 2021, [19] Huang et al. 2021, [20] Gillette et al. 1999, [21] Culbertson & Schuler 2019, [22] Maldonado & Culbertson 2021a, [23] Maldonado & Culbertson 2021b, [24] Dieueleveut et al. 2022, [25] Jasbi et al. 2018, [26] Zimmerman 2000

Integrating social information into pragmatic reasoning in real time

Pragmatic reasoning has been found to be shaped by different sources of social information (e.g. Bonnefon et al., 2009; Yoon et al., 2020; Mazzarella et al., 2018; Fairchild and Papafragou, 2018; Lorenzoni et al., 2022; Mahler, 2022) – including the stereotypical *persona* embodied by a speaker. In particular, Beltrama and Schwarz (2021) show that comprehenders adopt less precise interpretations of numerals (e.g., "$200" as "$190-210") for a **Chill** speaker, socially expected to speak loosely, than a **Nerdy** one, socially expected to speak precisely. These findings raise the question of how social information is integrated in meaning interpretation in real time – and specifically whether (**Hyp.A**) social considerations come into play at later stages of the interpretation process; or (**Hyp.B**) they are integrated from the start. Shedding light on this question would allow a novel perspective on how social information fits in the semantics/pragmatics interface. Support for Hyp.A would suggest that social stereotypes effects on meaning interpretation should be seen as the result or high-level, costly pragmatic reasoning, much like it has been suggested for pragmatic maxims in scalar inferences (Bott and Noveck (2004); Pouscoulous et al. (2007)); support for Hyp.B, by contrast, would suggest that it should be seen as information that is quickly integrated, similar to what is the case for linguistic/semantic information encoded as part of the truth-conditional content.

Methods. Adapting Beltrama and Schwarz (2021)'s task, we presented dialogues – visually represented as a cartoon – with one character asking a question ('How much is the flight?') and the other responding with a numeral utterance ('It's $200.') after checking their phone. The characters either embodied a **Nerdy** or **Chill** persona (between-subjects; see **Fig.1A-B**).

"Rachel and Arthur are looking for a one-way plane ticket"

Fig. 1A. Nerdy characters

"Alex and Eva are looking for a one-way plane ticket"

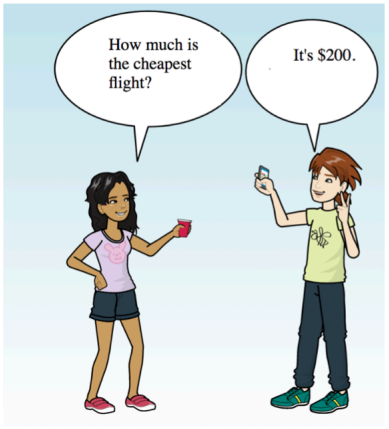


Fig 1B. Chill Characters

Participants had to indicate which of two phones the answer was based on: one displayed a number (visible screen); and one was shown face-down (covered screen). Participants were instructed to select thevisible screen if they thought the speaker was getting their information from this one; and the covered screen otherwise. Two further factors were manipulated. **Match** manipulated how closely the number on the visible phone matched the utterance, with 3 levels: 2 control levels, *Match* (identical) and *Mismatch* (far-off values); and the critical *Imprecise* level, displaying numbers slightly diverging from the uttered one (5-19%). Visible screen selections in the Imprecise condition indicate an imprecise interpretation; Covered screen selections a more precise one.

Screen Fit manipulation

Finally, **Time-Window** varied how long participants had to respond before the trial was aborted, with 4 levels (between-subjects): *SuperShort* (1250 ms); *Short* (2000 ms); *Medium* (2750 ms); *Long* (3500 ms).

24 items were presented in a Latin Square Design – 6 in *Match* and *Mismatch*, and 12 in *Imprecise*, +24 fillers. 768 participants were recruited on Prolific (96 per Persona/Time-Window combination), paid $2.

Results. As shown in Fig.3, covered choices for *Match/Mismatch* are at floor/ceiling as early as the Short window, but show degraded effects in the SuperShort one, suggesting that time-pressure in the latter made picture selection challenging even at the most basic level. We fit a ME logistic regression with random effects for Items/Participants on covered choice rates (excluding the SuperShort window due to lower accuracy in controls) and Persona, Match, Window and their interaction as predictors. We found a main effect of Match (β=0.62, p$<$.0001), reflecting a stepwise decrease from Mismatch to Imprecise and Match; and Persona (β=0.26, p$<$.0001), with higher rates for Nerdy speakers. But the Persona effect was dominated by interactions with Match and Window. Planned comparisons revealed a Persona effect in Imprecise ($p<0.0001$) but not in Match/Mismatch ($ps>0.4$); and – crucially for present purposes – the effect was significant in the Long window ($p<0.05$), but not in the shorter ones ($ps>0.7$).

Discussion. Our findings support **Hyp.A**: information about speaker identities does not affect interpretation in shorter response time windows. This indicates that comprehenders attend to and integrate descriptive linguistic meaning and social meaning in distinct stages, suggesting a stage of combining these two streams of information in processing. Thus, while social information is crucial for resolving meaning, it is dealt with separately from other interpretive cues. These results open a novel perspective on how the sociolinguistic and descriptive dimensions of meaning interact, a growing topic in pragmatics (see also Burnett (2019); Acton (2019)).

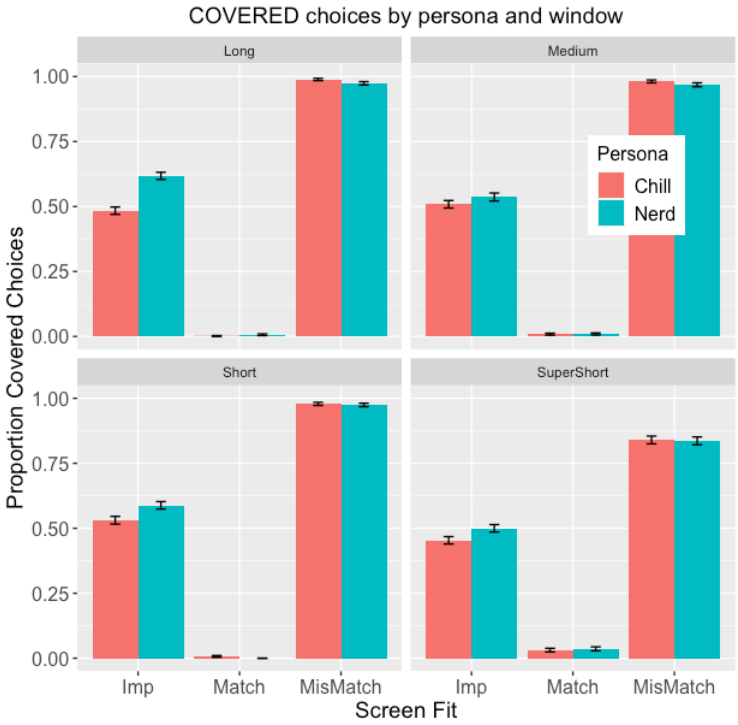


Fig.3. Covered choices by Match, Persona and Window

Experimentally investigating the strengthening properties of disjunction in French: When exclusivity meets free choice and ad hoc implicatures

French has at least two forms of disjunction, the simple 'ou' (1), and the complex 'soit…soit' (2).

(1) Anne a acheté la glace **ou** la tarte.
(2) Anne a acheté **soit** la glace **soit** la tarte. 'Anne bought the ice cream or the pie.'

'Soit…soit' is argued to trigger obligatory exhaustivity effects [1], which in unembedded contexts amounts to an obligatorily *exclusive* reading of the disjunction, namely that not both disjuncts are true. In this study, we investigate the obligatory exhaustivity requirement of 'soit…soit' by looking at the interaction of exclusivity with two other kinds of inferences: free choice [2] and ad hoc implicatures [3]. Exp.1 shows as a baseline that 'soit…soit' is indeed more *exclusive* compared to 'ou' in unembedded contexts. Exps. 2-3 show that when we introduce the possibility of strengthening to free choice or ad hoc implicatures, this difference between 'soit…soit' and 'ou' disappears. The findings support the proposal that 'soit…soit' is associated with obligatory exhaustification, which can be satisfied via exclusivity, free choice (FC), or ad hoc implicatures.

Experiments: All three experiments used the same paradigm. Participants were given a context story about characters shopping at a store. On each trial, a puppet named Rafie would make guesses about what the character would buy (Exp.1/3), or would describe what the character was allowed to buy (Exp.2). Participants had to judge whether the puppet was right or wrong, against the pictured outcome/rules. In each experiment, disjunction type ('ou' vs. 'soit…soit') was a between-subject variable. Participants saw 2 training items, followed by 30 test items (the relevant targets, true and false controls, and true and false fillers, all presented in randomized order).

Exp.1 (Baseline): *Participants*: 60 French native speakers were recruited through Prolific (30 'ou', 30 'soit…soit'). *Procedure:* On each trial, Rafie made a guess about what the character would buy (e.g., *Anne will buy the ice cream or the pie*). On the next screen, participants saw a picture of two items; purchased items were circled in green, while unpurchased items had a red circle and line through them (Fig.1). Participants had to judge whether Rafie's guess matched the pictured outcome. *Materials:* Critical targets (x10) involved both items circled in green, falsifying exclusivity. True controls satisfied exclusivity, while on false controls neither pictured item was purchased. *Results:* Accuracy was >97% on fillers/controls. Fig.2 displays the mean proportion of yes-responses to Excl-False and Excl-True trials. We fit a mixed effect logistic regression model with target type (Excl-False vs. Excl-True), disjunction type ('ou' vs. 'soit…soit'), and their interaction as fixed effects, and random by-participant slopes for target type. Model comparisons revealed effects of target type ($\chi2(1)=15$, $p<.001$) and disjunction type ($\chi2(1)=7.4$, $p<.01$), and a marginal interaction ($\chi2(1)=3.4$, $p=.065$). Importantly, people treated 'ou' differently from 'soit…soit', with more rejections of 'soit…soit' when exclusivity was not satisfied.

Figure 1. Image from Exp.1

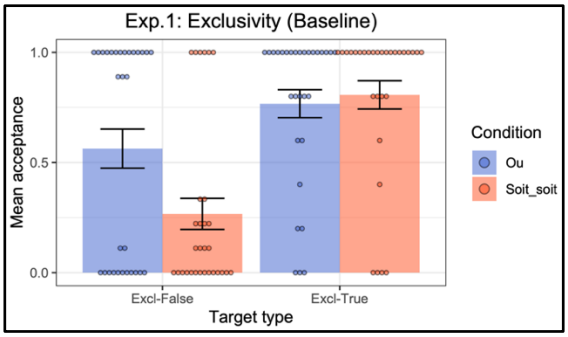


Figure 2. Results from Exp.1

Exp.2 (Adding FC): *Participants*: Another 61 French native speakers were recruited through Prolific (31 'ou', 30 'soit…soit'). *Procedure:* This time, what Rafie had to describe were the rules that Mum had set out for what each character was allowed to buy (e.g., *Anne is allowed to buy the ice cream or the pie*, which generates the FC inference that Anne is allowed to buy the ice cream and Anne is allowed to buy the pie). On each trial, there were three pictures side by side: the first object, the second object, and the third possibility was the combination of the two objects.

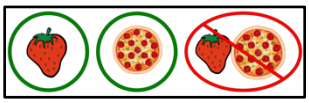


Figure 3. Image from Exp.2.

As explained in the instructions to participants, a green circle around an individual item meant that the character was allowed to purchase that item; a red circle with a line through an item meant that the character was not allowed to buy it; a green circle around the combination of the two items meant that the character could buy both; a red circle with a line through the third possibility indicated the character could not buy both items at the same time (Fig.3). Participants had to judge whether the puppet correctly described the rules. *Materials:* FC-True/Excl-False targets (x5) satisfied FC but falsified exclusivity, and FC-False/Excl-True targets (x5) falsified FC but satisfied exclusivity. *Results:* One participant was excluded for failing controls/fillers. For the remaining 60, mean accuracy was >95% on controls/fillers. Fig.4 displays the mean proportion of yes-responses to the FC-True/Excl-False and FC-False/Excl-True targets. We fit a mixed effect logistic regression model with target type (FC-True/Excl-False vs. FC-False/Excl-True), disjunction type ('ou' vs. 'soit…soit'), and their interaction as fixed effects, and random by-participant slopes for target type. Model comparisons revealed an effect of target type ($\chi2(1)=59$, $p<.001$), no effect of disjunction type ($\chi2(1)=.20$, $p=.66$), and no interaction ($\chi2(1)=.05$, $p=.82$). Importantly, people did not treat 'ou' and 'soit…soit' differently, responding primarily based on the truth/falsity of the FC inference. When the context falsified FC, participants always rejected the targets, suggesting the FC inference is quite strong, if not obligatory; meanwhile the bimodal distribution of participants in the FC-True/Excl-False condition shows that only some participants computed exclusivity *in addition to the FC inference* – even for 'soit...soit'.

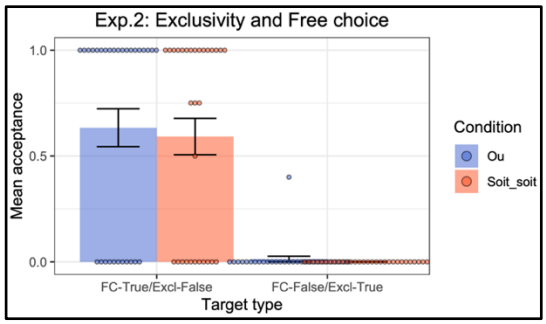


Figure 4. Results from Exp.2.

Exp.3 (Adding ad hoc implicatures): *Participants*: Another 60 French native speakers were recruited through Prolific (30 'ou', 30 'soit…soit'). *Procedure:* The set-up was as in Exp.1, but each picture contained three objects instead of two (allowing for ad hoc implicatures arising from the use of disjunction). *Materials:* Adhoc-True/Excl-False targets verified the ad hoc implicature but falsified exclusivity, while Adhoc-False/Excl-True targets falsified the ad hoc inference but satisfied exclusivity. *Results:* Fig.5 displays the proportion of yes-responses to the targets. Mixed effect logistic regression models revealed no effect of target type, disjunction type, or interaction. Unlike the FC data in Exp.2, the data in Exp.3 suggest that neither ad hoc nor exclusivity inferences are obligatory, with more than half of participants accepting when one of the inferences was false. Importantly, people did not treat 'ou' and 'soit…soit' differently.

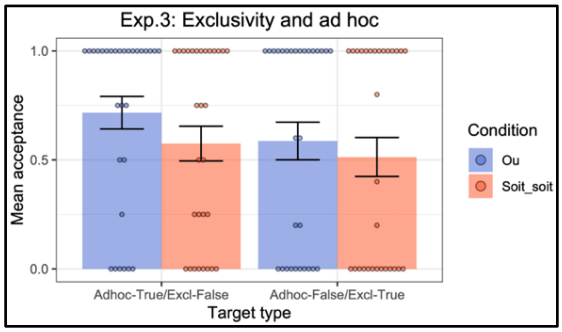


Figure 5. Results from Exp.3.

Discussion: Exp.1 confirms that when making judgments based on exclusivity alone, participants treat 'soit…soit' as more exclusive than 'ou'. However, once another inference is at play, be it FC (Exp.2) or ad hoc implicatures (Exp.3), the difference in the strength of exclusivity of the two disjunctions disappears. These findings are consistent with the idea that it is not exclusivity that is obligatory for 'soit…soit', but rather *strengthening* of some kind [1]. When strengthening via another implicature is an option, the difference between 'ou' and 'soit…soit' disappears, with participants becoming considerably less exclusive with 'soit…soit'.

References: [1] Spector, B. (2014). Global positive polarity items and obligatory exhaustivity. *Semantics & Pragmatics 7.* [2] Fox, D. 2007. Free choice and the theory of scalar implicatures. *Presupposition and Implicature in Compositional Semantics*, 71–120. [3] Hirschberg, J.L. (1991). *A theory of scalar implicature*.

Priming relevant and non-relevant features in metaphorical and literal contexts

This paper presents evidence for a continuity approach to predicate interpretation, based on cross-modal priming evidence. According to [1,2,3], hearers compute the speaker's meaning for an utterance like (1b,d) by selecting those features of CACTUS, which the speaker meant to convey. Here we assume that the same approach holds for (1a,c) and that the primary aim of both metaphorical and literal comprehension processes is to compute speaker's goals that may select potential implications from the predicate's semantic representation.

1. He/It is a cactus.
   a. John fell into a large plant. It was a cactus.
   b. Al's boyfriend is an awkward character and hard to come close to. He is a cactus.
   c. Max forgot to water his friend's house plant while she was away. But it's ok. It is a cactus.
   d. Al's boyfriend likes nothing more than to spend his summers in the desert. He is a cactus.

[4] conducted a cross-modal priming study for metaphors in context where target words were relevant Distinctive Features (DFs) and non-relevant Superordinates (SUPs) (John is a *cactus* – SPIKE/PLANT). Priming effects were found at all ISIs (0ms, 400ms & 1000ms) for DFs and at ISIs 400ms for SUPs (though marginal at 0ms). [5] reported a similar study with literal sentences in which targets were strong and weak associates (cactus – SPIKE/DRY), tested in neutral and weak-associate biasing contexts. Similar to [4], priming was found for both kinds of features at earlier ISIs, but only for relevant features at later ISIs. Taken together, these studies indicate similar patterns for both Lit. and Met. contexts, but neither the prime sentences/context nor target types were the same. Our first aim was to conduct a better controlled comparison between Lit and Met contexts, by using sentences placed in contexts which result in either a literal or metaphorical interpretation (e.g., 2 & 3), and by controlling different types of non-relevant features in addition to relevant features (See **Table 1**).

Items: Following [4, 5], we did a distinctive feature listing task, a brief definition task and a simple association task to select distinctive features (DFs), superordinates (SUPs) and strong associates (SAs). Selected DFs had a lower frequency rank than selected SUP and SA targets. The latter were ranked highest among elicited responses. LSA analysis showed no difference in association between Prime words and any of the three target types. We then constructed 24 strongly constraining literal and metaphorical context sentences so that DFs are related to clear coherence relations and SUPs & SAs are non-relevant (See **Table 1**).

2. Maria's friends looked after her when she was in a difficult situation. They are gems.
3. The objects he dug out of the ground in Brazil impressed every collector. They are gems.

Table 1.

Prime	Distinctive features	Superordinates	Strong associates
Gems	Precious	Stone	Diamond

Cross-modal priming task: Participants (N=360, native English) first listened to context sentences and then made lexical decisions to visual target words offset at either 0ms, 400ms or 1000ms from the Prime. They were employed in a 3 (ISI) * 2 (context) * 3 (target type) * 2 relatedness (related, control) design. Only ISI was a between-group factor. A different set of 12 metaphoric contexts & 12 literal contexts paired with English-like non-words were included as fillers.

Results: A generalized linear mixed-effects model for each ISI showed: **(1). At 0ms**, there was a *context*target type*relatedness* interaction (p<.001). Follow-up analysis showed in literal contexts, no priming was found for any target type; in metaphorical contexts, there was a *target type*relatedness* interaction (p<.001). Priming was found only for DFs (p=.01). **(2)**. **At 400ms**, overall, there was a two-way interaction between *target type* & *relatedness* (p<.001). Follow up analysis on each target type showed priming for DFs (p=.002) and SUPs (p=.03), not for SAs (p=.3). **(3)**. **At 1000ms**, there was a *context*target-type*relatedness* interaction (p=.01). In both literal and metaphorical contexts, there was a *target type & relatedness* interaction (both *p's*<.001); priming was found for only DFs (Lit, p=.005; Met p=.03). (see **Figure 1**).

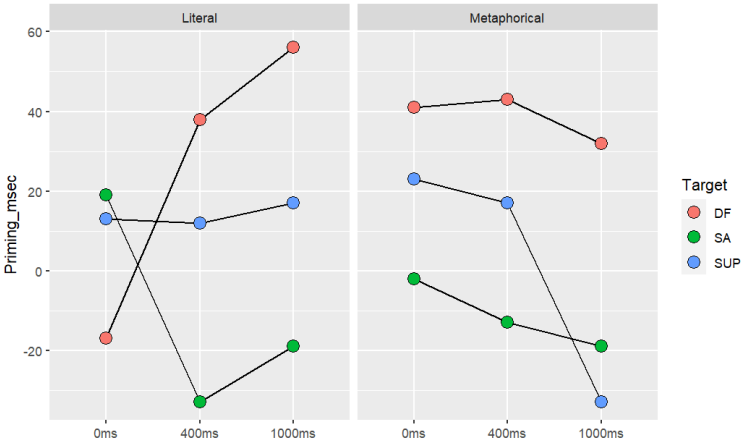


Figure 1. Priming of three types of target words in literal and metaphorical contexts

Discussion: Overall, we find comparable patterns in Lit. and Met. contexts, with clear priming advantages for relevant DFs compared to non-relevant core (SUP) and associate (SA) features. Unlike [4,5], we account for any limited priming for non-relevant features in terms of probabilistic models of the hearer's problem of deciding which set of features the speaker intends, similar to [2,3]. Priming effects of non-relevant features result from strength of priors on feature sets and goal uncertainty. In particular, SUP features such as PLANT for *cactus* are more related to frequently relevant category prototype features, so that even though contexts make a subset of features relevant, the high prior on those defining features makes the posterior for these implications compete with the intended relevant ones. In discussion we will reflect on model details in [2,3] and consider whether their 'literalness prior' (P(c)) needs in fact to be conditioned on a 'wonkiness' variable as per [6]. Also the role of any 'salience' term in speaker's model (see [3]), in light of relative prominence of 'low salient' SUP features. We attribute the lack of priming at 0ms in the literal context for even relevant distinctive features to the fact that our literal contexts overall may not have been as constraining as metaphorical contexts (e.g., "*There were water stations every two miles at that event. It was a marathon").*

References: [1]. Sperber, D. & Wilson, D. (1986). *Proceedings of the Aristotelian Society 86,* 153-172. [2]. Kao *et al*. (2014). *Proc. of CogSci* 36 (36), 719-724. [3] Mayn & Demberg (2022). *Proc. of CogSci* (44), 3154-3160. [4]. Rubio Fernandez, P. (2007). *Journal of semantics* 24(4), 345-371. [5]. Rubio-Fernández, P. (2008). *Journal of semantics* 25(4), 381-409. [6] Degen *et al*. (2015). Proc. of Cogsci.

Priming between universal quantifiers in negated scopally ambiguous sentences

Sentences involving universal quantification and negation give rise to systematic scope ambiguities. For example, the English sentence *Every shark doesn't attack the surfer* can either mean that there is no shark that attacked the surfer (the *universal-wide* interpretation) or that not every shark attacked the surfer (but possibly some did; the *negation-wide* interpretation). Interestingly, quantifiers seem to differ in their scope preferences, even when they carry the same quantificational force. *Each*, for example, has a stronger tendency to take wide scope than *every* or *all.*[1,2] This observation, among other differences between quantifiers, has led to theoretical descriptions that posit distinct mental mechanisms and representations of scope-taking for different quantifiers.[e.g., 3-4]

The representation of scope can be experimentally tested using *structural priming*, a phenomenon in which the use of a linguistic representation is facilitated if the same representation was recently used. When structural priming between sentences occurs, these sentences therefore share some representational resources. Scope configurations are susceptible to such priming.[2,5,6] However, it is not clear whether this priming is dependent on the repetition of the same quantifier.[2,6] In the current project, we investigate this question by examining the representation of quantificational scope in the interpretation of French sentences with a universal quantifier vis-à-vis negation.

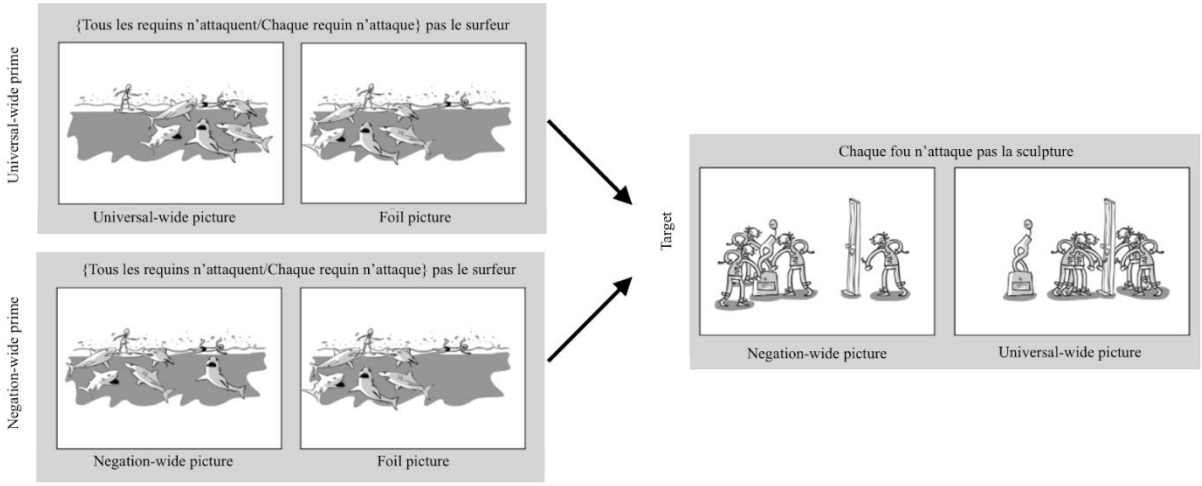


Fig. 1 Procedure and conditions of our sentence-picture matching task.

We used a sentence-picture matching task to test priming of relative scope in French.[2,5] On each trial, participants matched a sentence with one of two pictures. In primes and targets, this sentence contained universal quantification vis-à-vis negation, e.g. *Chaque requin n'attaque pas le surfeur* ("Every shark doesn't attack the surfer"). In the primes, we forced participants to assign a particular reading, because they could choose between a picture depicting that reading and a picture that mismatched any possible reading. In the subsequent targets, participants could freely choose between two pictures matching the two different readings (Fig. 1). Priming occurs if participants' choice of reading on the target trial is affected by the reading they were forced to choose on the preceding prime trial.

In the prime trials, we varied the Prime Scope (between *universal-wide* and *negation-wide*) and the Prime Quantifier (between *chaque* 'every' and *tous les* 'all the') within participants. The target sentences always involved the quantifier *chaque*.[6] If the representation of the scope taken by *tous les* and *chaque* abstracts away from the differences between these words and their meanings, then there should be priming not only between sentences that share the same

quantifier (from *chaque* to *chaque*), but also across different quantifiers (from *tous les* to chaque). Native speakers of French took part in the experiment (n = 144).

Fig.2 shows the proportion of target responses compatible with the universal-wide interpretation for both primes. A Bayesian logistic regression model revealed that in both prime quantifier conditions, participants were less likely to select the universal-wide picture in the targets following a negation-wide prime than in those following a universal-wide prime (β = -0.240, 90%CI = [-0.35, -0.13], SE = 0.06, P(β<0)=1)). The model also revealed an interaction: priming was larger in the within-quantifier *chaque* condition than in the between-quantifier *tous les* condition (β = -0.11 , 90%CI = [-0.21, -0.02], SE = 0.06, P(β<0.97)=1; Fig. 2).

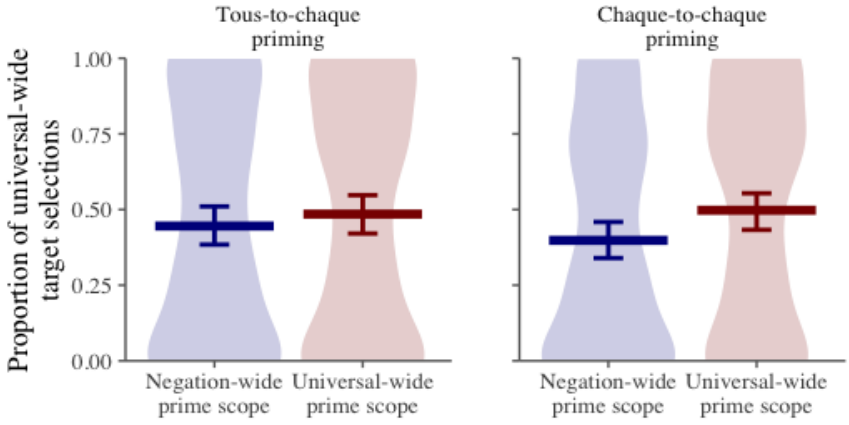


Fig. 2 Results of our sentence picture-matching task. The horizontal bars denote the mean, and the outline of the shaded

Altogether, our results show that scopal configurations can be primed between different universal quantifiers (although we also find that priming within the same quantifier is larger than between quantifiers). This suggests that there are commonalities in the representation of scope between different quantifiers[8], which contradicts theories that posit quantifier-specific mechanisms for scope taking.[e.g., 3,4] Instead, our results suggest that people rely on more general mechanisms in the assignment and representation of relative scope.[6, 7, 8]

References

1. Ioup, G. (1975). Some universals for quantifier scope. In *Syntax and Semantics* volume 4 (pp. 37-58). Brill.
2. Feiman, R., & Snedeker, J. (2016). The logic in language: How all quantifiers are alike, but each quantifier is different. *Cognitive psychology, 87*, 29-52.
3. Beghelli, F., & Stowell, T. (1997). Distributivity and negation: The syntax of each and every. *Ways of scope taking*, 71-107.
4. Steedman, M. (2012). *Taking scope: The natural semantics of quantifiers*. Mit Press.
5. Raffray, C. N., & Pickering, M. J. (2010). How do people construct logical form during language comprehension?. *Psychological science, 21*(8), 1090-1097.
6. Slim, M. S., Lauwers, P., & Hartsuiker, R. (2023). *Revisiting the logic in language: The scope of 'each' and 'every' universal quantifier is alike after 'all'.* PsyArXiv. https://psyarxiv.com/jgcxy/
7. Gil, D. (1995). Universal quantifiers and distributivity. In *Quantification in natural languages* (pp. 321-362). Dordrecht: Springer Netherlands.
8. Fodor, J. D. (1982). The mental representation of quantifiers. In *Processes, beliefs, and questions: Essays on formal semantics of natural language and natural language processing* (pp. 129-164). Dordrecht: Springer Netherlands.

Cross-domain event primitives are reflected in motion verb learning across languages

Languages vary in the components of a motion event they prefer to lexicalize in verbs. English often packages the manner in the verb ("He **ran** into the store"). Spanish typically encodes path in the verb ("Él **entró** en la tienda corriendo") ([1]). These verb lexicalization biases affect the way novel motion verbs are acquired cross-linguistically ([2]-[6]) but are malleable ([7], [8]). The **manner-path** distinction in the spontaneous-motion domain is semantically similar to the **means-result** distinction in the caused-motion domain ([6]), for example a girl kicking a ball into a bucket, where kicking is the means and the sending-into-a-bucket is the result. We test whether taught lexicalization biases in spontaneous-motion shape means-result verb learning in the caused-motion domain, and do so consistently across languages. If so, this would give evidence for both the structure of the lexicon ([9]), and the flexibility of verb lexicalization biases.

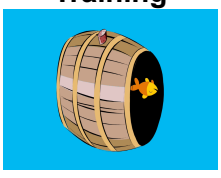
Participants Adult native speakers of English (N=78) and Spanish (N=76) were assigned one of the three training types: No-training, Manner-verb training, and Path-verb training.

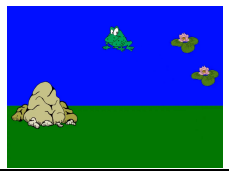
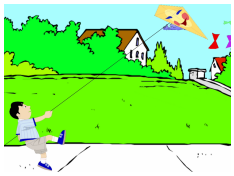
Training The training groups were told that they will be learning an alien language. They were trained on eight novel verbs each associated with a videoclip depicting spontaneous-motion. On each trial, participants first saw a clip with a novel verb. Afterwards, they saw a manner-match clip and a path-match clip. Participants in the Manner-verb training condition were told that the manner-match clip, but not the path-match clip, is an instance of the verb, and vice versa for the Path-verb training condition. (Table1)

Testing After training, participants were tested on four novel spontaneous-motion events and eight caused-motion events. On each trial, participants watched a clip along with a novel verb. Afterwards, they saw a manner-match (or means-match, for caused-motion) clip, and a path-match (or result-match, for caused-motion) clip, and were asked whether they accept them as instances of the verb (Table1).

Results/Discussion On spontaneous-motion trials, both English- and Spanish-speaking participants trained with manner or path-verbs generalized these lexicalization patterns to new spontaneous motion events (*glmer*, p's<0.05). This suggests that learned lexicalization patterns affect novel spontaneous-motion verb conjectures. On caused-motion trials, both English- and Spanish-speaking participants who learned manner or path-biases in the spontaneous motion domain (≥ 75% accuracy) transferred these lexicalization patterns to new caused motion events (p's<0.05). After learning novel motion verbs that encoded manner or path, both English- and Spanish-speaking adults formed corresponding lexicalization biases that influenced the acquisition of subsequently encountered motion verbs across domains. The overall data pattern indicates that there are underlying commonalities between Manner/Path and Means/Result, suggesting that higher-order generalizations operate over conceptual or lexical dimensions that are not specific to a particular kind of event (spontaneous motion or caused motion).

Table 1. Example of training and test trials. (The order of the means/manner and the path/result testing trials was counterbalanced across verbs.)

Phase	Video type	Scene	Language (English, Spanish)	
Training 	Initial video	Fish flips through barrel	This is gorping. Esto es dojar.	
	Manner match	Fish **flips** under barrel	Manner-verb training	This is gorping. Esto es dojar.
			Path-verb training	This not gorping. Esto no es dojar.
	Path match	Fish bobs **through** barrel	Manner-verb training	This is not gorping.

				Esto no es dojar.
			Path-verb training	This is gorping. Esto es dojar.
Testing – spontaneous motion 	Initial video	Frog jumps to the front of a rock	This is bligging. Esto es sarar.	
	Manner match	Frog **jumps** to the top of a rock	Was that bligging? (Y/N) ¿Eso fue sarar? (Y/N)	
	Path match	Frog hops **to the front of a rock**	Was that bligging? (Y/N) ¿Eso fue sarar? (Y/N)	
Testing – caused motion 	Initial video	A boy pulls on a kite string; the kite comes down from the sky	This is nolding. Esto es chellar.	
	Means match	A boy **pulls** on a kite string; the kite moves slightly in the air	Was that nolding? (Y/N) ¿Eso fue chellar? (Y/N)	
	Result match	A boy clasps a kite string; the kite **comes down from the sky**	Was that nolding? (Y/N) ¿Eso fue chellar? (Y/N)	

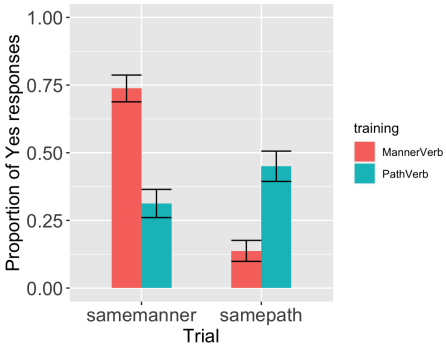


Figure 1. English-speaking participants' responses on **spontaneous motion** test trials (error bars are ±SE)

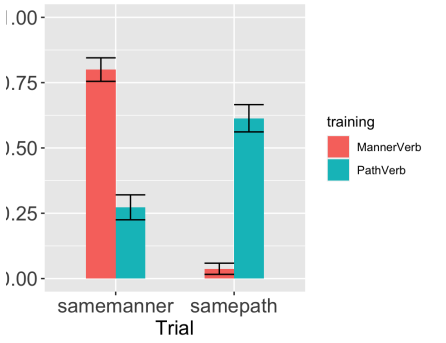


Figure 2. Spanish-speaking participants' responses on **spontaneous motion** test trials (error bars are ±SE)

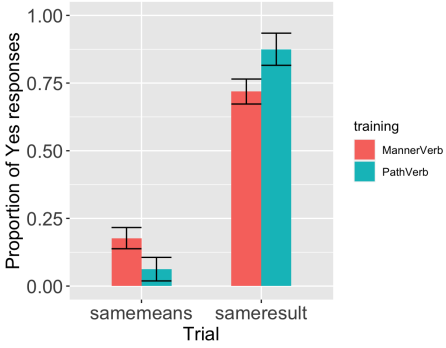


Figure 3. English-speaking participants' responses on **caused motion** test trials (error bars are ±SE) (Participants who successfully learned intended biases for spontaneous motion; ≥ 75% accuracy)

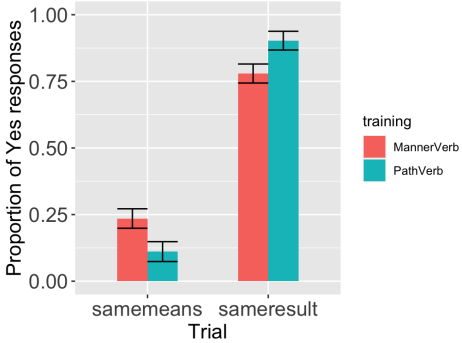


Figure 4. Spanish-speaking participants' responses on **caused motion** test trials (error bars are ±SE) (Participants who successfully learned intended biases for spontaneous motion; ≥ 75% accuracy)

References

[1] Talmy (1985). In *Language typology and syntactic description.* **[2]** Hohenstein (2005). *Journal of Cognition and Development*. **[3]** Naigles & Terrazas (1998). *Psychological Science*. **[4]** Maguire et al. (2010). *Cognition*. **[5]** Papafragou, Massey & Gleitman. (2002). *Cognition.* **[6]** Papafragou & Selimis (2010). *Language Learning and Development*. **[7]** Shafto, Havasi, & Snedeker (2013). *Developmental Psychology*. **[8]** Geojo (2015). Harvard Dissertation. **[9]** Rappaport Hovav & Levin. 1988. Building verb meanings.

Experientiality markers in memory reports: A semantics-pragmatics puzzle

In a nutshell. We give experimental support that German free relative *wie-* ['how'] complements embedded under the memory predicate *noch wissen* ['still know'] mark the remembering of a personally experienced event. Our main experiment, based on scale judgements, raises questions about the pragmatics-semantics interface of this phenomenon, and about the robustness of experiential memory markers in general. Two complementary studies address these questions.

Background. The complex German memory predicate *noch wissen* (lit. 'still know'), can combine with a declarative *dass* ['that'] clause (1b) and with an eventive-*wie* ['how'] free relative (1a):

(1) a. *Ich weiß noch, wie Oma im Meer geschwommen ist.*
I know still how Grandma in-the sea swim is
'I remember Grandma swimming in the sea.'

b. *Ich weiß noch, dass Oma im Meer geschwommen ist.*
I know still that Grandma in-the sea swim is
'I remember that Grandma was swimming in the sea.'

Dass-clauses and *wie*-free-relatives can be coordinated under *noch wissen*. Therefore, we assume a uniform semantics for *noch wissen* in (1a) and (1b) (cf. Sadock and Zwicky, 1975), such that these sentences form a minimal pair.

Most theories of memory distinguish experiential remembering (i.e. recall of a personally experienced event) from 'fact-only' remembering, i.e. recall of general facts based on indirect evidence (Tulving, 1972). In our experiments, we introduce the siblings Red and Blue (wearing name-matching clothes; alongside their control cousin Pinkie) to personify these kinds of experience:

(2) a. Red spent the summer two years ago with Grandma and saw her swimming in the sea.

b. Blue spent that summer abroad and was told about Grandma's swimming much later.

Based on our intuitions and in line with literature on non-manner 'how' (Umbach et al., 2022), we expect (1a) to unambiguously report experiential memory (Red, (2a), 1st picture) while (1b) is expected to report both fact-only (Blue, (2b), 2nd picture) and experiential memory. By confirming this, we provide the first empirical evidence for experientiality markers in memory reports.

Our **Main Experiment** is a Qualtrics online study asking for judgements on a scale from 1 (*gar nicht richtig*, 'not correctly at all') to 7 (*völlig richtig*, 'absolutely correctly') for sentences describing a given scenario. We recruited 40 German native speakers via Prolific, excluded three based on low control performance, and tested within-subjects. Independent variables were the complementizer (values: WIE, DASS; see (1)) and the character uttering the sentence (values: RED, BLUE; see (2)), resulting in four items per scenario and 16 target items – 4 per condition – in sum, augmented with 16 controls. Based on our background assumptions and literature-informed expectations sketched above, we formulated two hypotheses. **Hypothesis A:** Higher ratings for the RED+WIE than for the BLUE+WIE condition; and **Hypothesis B:** Higher ratings for BLUE+DASS than for BLUE+WIE. Both together would show that *wie* in *noch wissen* reports is an experientiality marker in the sense that it disambiguates for experiential memory in contrast to *noch wissen, dass*.

Descriptive Statistics: Hypothesis A was clearly confirmed with an extremely strong contrast (see table for means of all datapoints and standard deviation; see Figure 1 for quartiles and outliers). Hypothesis B was also confirmed, but with a weaker contrast due to the lower-than-expected rating of BLUE+DASS. We are confident that these contrasts will be shown to be highly significant in our inferential statistics performed in January, applying an ordinal cumulative link mixed effect model (for motivation of the choice, see Liddell and Kruschke, 2018).

	RED	BLUE
WIE	$\mu = 6.80$ ($\sigma = 0.54$)	$\mu = 3.78$ ($\sigma = 1.98$)
DASS	$\mu = 6,69$ ($\sigma = 0.61$)	$\mu = 4.80$ ($\sigma = 1.91$)

Speaker-ID Experiment: The set-up and the phrasing of the scale labels of the Main Experiment were aimed at truth-conditional semantics (see Zhu and Ahn, 2023, for the influence of instructive formulations on results). To control for pragmatic competition, we ran a smaller experiment (n=30, 4 target + 4 control items) in a speaker-identification format (inspired by Davis and Landau, 2021) with the same background story and sentences as the Main Experiment. In the speaker-ID format, participants had to select exactly one character who uttered the sentence. The independent variable here was the complementizer. The nature of the evidence (RED vs. BLUE) was turned into the dependent variable, leading to a 2x2 set-up. We confirmed **Hypothesis I** – RED was selected more often (84%) in the WIE condition – and **Hypothesis II** – BLUE (64%) in the DASS condition.

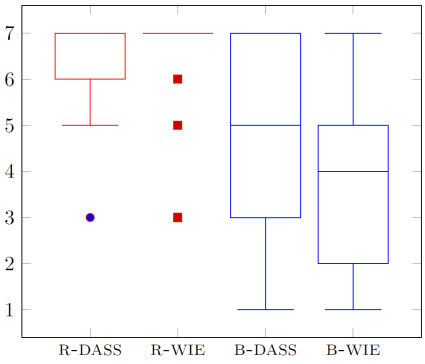


Figure 1: Ratings Main Exp.

Discussion: Experiment 1 on its own suggests that *wie* is an episodicity marker in the sense of decreased acceptability of *noch wissen, wie* in a fact-only scenario. It leaves room, in principle, to reason that this could be due to pragmatic competition with *noch wissen, dass* which is preferred for independent reasons in the fact-only case. The fact that the preference for BLUE in the DASS condition was much weaker than the preference for RED in the WIE condition suggests the opposite: *wie* is limited to experiential remembering semantically, *dass* can be used in both cases. Maximizing precision in the fact-only case leads to a pragmatic preference of *noch wissen, dass*.

A Puzzle: That BLUE+DASS scored much lower than RED+WIE in our Main Experiment is a surprise: Since BLUE+WIE has even lower ratings, there are participants who do not grant Blue *any* kind of remembering even though they have reliable indirect evidence. The high σ for BLUE+DASS and a look at the individual participants' answers suggest a divide: One group of participants is in line with our semantic-pragmatic explanation above while others have stricter conditions on memory – in the scale format of our Main Experiment, that is! If experiential remembering was just always 'the real' memory, we wouldn't expect a preference for BLUE in the DASS condition in the Speaker-ID Experiment. The solution we suggest is that the forced choice design gives rise to the pragmatic competition we intended while our judgement scale design is sensitive to the accomodation of different Questions Under Discussion (QUD): That the grandchildren are said to exchange stories of the old time might lead some people to accommodate a QUD like 'Who was there when that happened?'. We plan to test this explanation by contrasting the Main Experiment with a version of it introducing a QUD like 'Who knows the most facts about Grandma?'

The question remains how broad the phenomenon of experientiality markers is. Our **English Scale Experiment** (n=29, 8 target items) is a first hint that it might be quite robust. It is mostly equivalent to the Main Experiment, but with the memory predicate *remember* and the hypothesized marker gerundive *-ing* small clauses in contrast to *that*-clauses (i.e. the translations in (1); inspired by Bernecker, 2010). The results, including the puzzle on BLUE+DASS, closely resemble the German results. This suggests a phenomenon that ranges over languages, memory predicates, and structures marking experientiality. A preregistered study (n=100) with our scale design contrasting German present and past tense in the complement of *noch wissen* follows in February 2024.

Bernecker, S. (2010). *Memory.* Oxford University Press. · Davis, E. E. and B. Landau (2021). *Seeing vs. seeing that.* In Proceedings of ELM 1, pp. 125-135. · Liddell, T. and J. Kruschke (2018). *Analyzing ordinal data with metric models: What could possibly go wrong?* Journal of Experimental Social Psychology, 328-348. · Sadock, J. M. and A. M. Zwicky (1975). *Ambiguity tests and how to fail them.* In J. Kimball (Ed.), Syntax and Semantics, Vol. 4, pp. 1-36. New York. · Tulving, E. (1972). *Episodic and semantic memory.* In E. Tulving and W. Donaldson (Eds.), Organization of Memory, pp. 381-402. New York: Academic Press. · Umbach, C., S. Hinterwimmer, and H. Gust (2022). *German wie-complements: Manners, methods and events in progress.* Natural Language and Linguistic Theory 40, 307-343. · Zhu, Z. and D. Ahn (2023). *Effects of instruction on semantic and pragmatic judgment tasks.* In Proceedings of ELM 2, pp. 322-33.

Assessing scalar meaning: a first exploratory study on some Italian focus particles

Well-established analyses of *even* claim that the contribution of this scalar-additive operator to the sentences in which it appears is of both an additive and a scalar inference [1], [2], [3]. Sentences with scalar-additive particles not only entail the corresponding sentences without the particle. They also imply that at least one among the associate's [4] focus alternatives satisfies the predicate under consideration, and that the alternative(s) at issue are somewhat ordered [5]. When *even* appears under negation, this ordering gets reversed. As far as Italian is concerned, the perfect correspondent of *even* is held to be *persino* [6]. However, other particles can trigger the additive and scalar inferences in a similar way. Among these, *pure* and *anche,* and their negative counterparts *neppure* and *neanche* [7], [8], [9], [10]. At present, little experimental work has been carried out on Italian focus operators. To our knowledge, only one study has been conducted on the processing of *persino* [11]. Hence, it is yet to be determined whether the strength of the scalar inference may remain constant across different particles. Based on the existing literature, there is reason to believe that *persino* might carry a stronger scalar inference compared to other operators – *anche* and *pure* being compatible with, but not committed to, a scalar interpretation.

We thus designed a multiple-choice cloze test targeting *persino-persino...non, pure-neppure* and *anche-neanche*. We first presented participants with a drawing and a short story and then asked them to fill in a concluding sentence choosing one of these six particles provided as alternatives (Figure 1). The pictures and stories were ideated to make participants build some expectations about which of the three characters presented was the most or least likely to carry out the action described – an action which could eventually be either successfully accomplished or failed by all. In this way, we aimed at observing whether the rate of particle selection may vary depending on which character the concluding sentence focusses on and, possibly, on the outcome of the stories. Via within-subject manipulation of Particle Polarity (positive-negative) and Focussed Character (likely-middle), the experiment indeed consisted of 80 (48 experimental) trials which could appear under four possible conditions.

The responses of 89 monolingual Italian adults (43 females; age M=32) were analysed via multinomial logistic regressions on Particle Type, with Polarity and Character as fixed effects and Participant as random intercept. The model returned a significant interaction between Polarity and Character (β=1.23, p<.001). The most influencing effect was that of Character (β=-3.32, p<.001), even though the effect of Polarity was significant, too (β=-0.38, p<.001). As shown in Figure 2, *persino-persino…non* was the most selected alternative in both positive and negative settings with focus on the likely character (78.3% and 48.3%, respectively), which seems to indicate that this is the alternative that participants perceived as carrying the strongest scalar inference among the ones provided. Interestingly, though, the choice rate of *persino…non* in negative settings was lower than that of *persino* in positive ones. *Pure-neppure* was rather chosen more in negative (27.3%) than in positive frameworks (3.1%) in association with the likely character. In negative situations, it was chosen at an almost equal rate when coupled with the likely (27.3%) and the middle character (26.5%), and its selection rate on the likely was almost on a par with that of *neanche* (24.4%). All this seems to indicate that *pure-neppure* was understood to bring a somewhat weaker scalar inference than *persino-persino…non.* Last, *anche-neanche* was selected more with the middle than with the likely character, be it in positive (82.3%) or negative contexts (68.2%), which suggests that this might have been felt as the "less scalar" among the alternatives provided.

All in all, these data not only seem to point out that different Italian scalar-additive operators are associated with different scalar force. They also seem to imply that, despite appearances, the positive-negative pairs selected are not the mirror image of each other – a fact certainly worth scrutinizing in further detail.

Figures

Figure 1 - An example of experimental trial

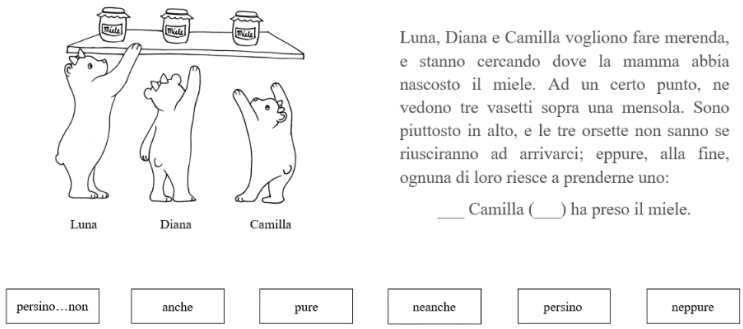
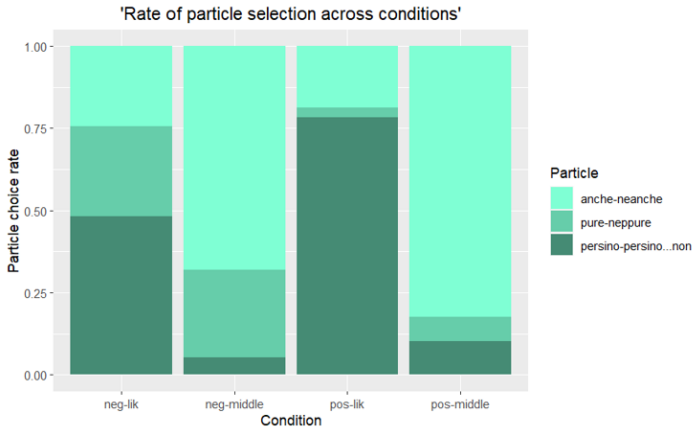


Figure 2 - Proportion of particle selection across conditions



References

[1] Horn, L. (1969). A presuppositional analysis of only and even. *Chicago Linguistic Society (CLS) 5*, 98-107.
[2] Karttunen, L. & Peters, S. (1979). Conventional implicature. In C-K. Oh & D.A. Dinneen (Eds.), *Syntax and Semantics: Presupposition* (Vol. 11, pp. 1-56). Academic Press.
[3] Rooth, M. (1985). Association with focus. PhD thesis. University of Massachusetts at Amherst.
[4] Krifka, M.: 1998, Additive particles under stress. In D. Strolovitch & A. Lawson (Eds.), *Proceedings of Semantics and Linguistic Theory VIII* (pp. 111-129). Cornell University.
[5] König, E. (1991). *The Meaning of Focus Particles*. Routledge.
[6] Gast, V. & van der Auwera, J. (2011). Scalar additive operators in the languages of Europe. *Language 87*(1), 2-54.
[7] Tovena, L.M. (2006). Dealing with alternatives. *Proceedings of Sinn und Bedeutung* (Vol. 10, No. 2, pp. 373-388).
[8] Tovena, L. M. (2005). Discourse and addition. In K. von Heusinger & C. Umbach (Eds.), *Proceedings of the Workshop Discourse Domains and Information Structure ESSLLI 2005*, (pp. 47-56).
[9] Mari, A. & Tovena, L.M. (2006). A unified account for the additive and scalar uses of Italian neppure. In C. Nishida & J-P.Y. Montreuil (Eds.), *New Perspectives on Romance Linguistics: Morphology, Syntax, Semantics, and Pragmatics*. (Vol. 1, pp. 187-200). John Benjamins Publishing Company.
[10] Panizza, D. & Sudo, Y. (2021, February 23-26). *Not you, too! A two-stage exhaustification account for Italian additives/miratives.* [Conference presentation]. IGG 46, Siena, Italy.
[11] Bello Viruega, I. & Nadal, L. (2023). *Processing scalarity: an experimental study on the Italian focus operator perfino*. [Manuscript submitted for publication].

Contrafactives, learnability, and production

Abstract for ELM3

Natural languages appear to universally feature factive verbs like *know* (Goddard, 2010), whereas no clear example of a so-called contrafactive has been found yet (see, e.g., Holton, 2017; Glass, 2023; Roberts and Özyildiz, 2023). A contrafactive is the mirror image of a factive attitude verb like *know*. Although both *x factives that p* and *x contrafactives that p* entail that x believes that p, the former presupposes that p is true, whilst the latter presupposes that p is false.

Strohmaier and Wimmer (2022; 2023) have proposed that the stark difference in how common factives and contrafactives are arises partly because the meaning of a contrafactive is harder to learn than that of a factive. They tested this hypothesis by conducting two computational experiments using artificial neural networks—more specifically, Transformers, which are the foundation of current state-of-the-art results in natural language processing and show greater convergence with human processing than other approaches (Vaswani et al., 2017; Caucheteux and King, 2022). Their networks were trained to predict the truth value of factive, non-factive and contrafactive ascriptions, given a representation of the state of the world and a representation of the world as the attitude holder takes it to be (which may or may not be accurate). The networks' predictions were then expressed in a probability that the target ascription is true. Importantly, Strohmaier and Wimmer's experiments provide initial support for their hypothesis: in both cases the networks' loss drops faster for factives than for contrafactives.

However, their experiments are subject to at least two limitations. First, they understand an assignment of probability 0 to an ascription as claiming that the ascription is definitely not true, which leaves open whether the ascription is false or undefined due to presupposition failure. Thus, their experiments are not sensitive to a key feature of factives and contrafactives: their presuppositions. Second, their experiments effectively consider the comprehension (or evaluation) of attitude ascriptions fed into the networks. Their results therefore do not speak to the relative difficulty of learning how to produce factive and contrafactive ascriptions. This leaves open the possibility that it is easier to learn how to produce contrafactive ascriptions than factive ones. But if this possibility was to be realized, would the meaning of a contrafactive be harder to learn than that of a factive overall? This would depend on the difficult question of how to weigh the learnability of production and comprehension in assessing the overall learnability of the target expressions.

To address the two limitations facing Strohmaier and Wimmer, we conducted a computational experiment, using another Transformer, in which our network produces factive, non-factive or contrafactive ascriptions, given a representation of the world as the attitude holder takes it to be, information or a lack thereof about whether this representation is correct, plus a demand that the network produces an ascription with a certain truth-value (true, false, or presupposition failure) and, in doing so, not only uses an embedded clause with a certain truth-value (true, false, unknown), but also produces the most informative ascription possible (thereby satisfying Grice's maxim of quantity and Heim's maximize presupposition). Because similar, learnability-focused work on other semantic universals (e.g., Steinert-Threlkeld, 2020) has also been limited to testing comprehension, our approach serves as proof of concept for a new experimental paradigm that can be used in learnability-based explanations of semantic universals.

To illustrate the input and output of our network, let's say we provide it with the attitude holder representation 'buy lorelai tomato chili stew dinner now', information that this representation is incorrect, and demand that it produces a true ascription with a false embedded clause. Given these inputs, the network is trained to produce a contrafactive ascription. Again, say we provide the network with the attitude holder representation 'cook lorelai mushroom pepper rice lunch now',

information that this representation is correct, and demand that it produces a true ascription with a true embedded clause. Now, the network is trained to produce a factive ascription.

Our Transformer models closely follow the paper by Vaswani et al. (2017) using the pyTorch implementation. The main difference is that output for each position in the sentence is constrained to the words allowed in that position using position-specific linear layers, i.e. our model capitalises on the fixed word order of the artificial language. This minimizes the role of syntax, which is of benefit since we are primarily interest in lexical semantics (and a limited number of pragmatic principles). We used a custom loss that encoded the semantic-pragmatic success conditions.

We explored 41 different hyperparameter settings using a randomised search and 5-fold cross-validation. In all but two of those settings, the model failed to learn the semantics of the target expressions. We evaluated the two successful settings on the held out test-data and in addition varied the original random seed for each of them four times, leading to 10 evaluations overall. The varying of the random seed allows us to test whether the results are robust to a random change in the initial conditions of the neural network.

While we do find small differences in the speed in which attitude verbs are learned by the model, these are not robust to changes in the random seeds. Furthermore, insofar as any trends are discernible, contrafactives appear to be acquired faster than factives.

Our empirical contributions thus are:

1. Transformer models can produce factive, non-factive, and contrafactive ascriptions, learning both semantic conditions and pragmatic principles.
2. Variation of random initialisation can affect the learning differences between attitude verbs, contrary to the hypothesis that contrafactives are consistently harder to learn than factives.

These results stand in clear contrast to those previously found by Strohmaier and Wimmer (2022; 2023) and underline the importance of considering production in modelling lexical acquisition.

References

Caucheteux, C. and J.-R. King (2022). “Brains and algorithms partially converge in natural language processing”. In: *Communications Biology* 5.1, p. 134. DOI: 10.1038/s42003-022-03036-1.

Glass, L. (2023). “THE NEGATIVELY BIASED MANDARIN BELIEF VERB yǐwéi*”. In: *Studia Linguistica* 77.1, pp. 1–46. DOI: 10.1111/stul.12202.

Goddard, C. (2010). “Universals and Variation in the Lexicon of Mental State Concepts”. In: *Words and the Mind: How words capture human experience*. Oxford: Oxford University Press.

Holton, R. (2017). “I—Facts, Factives, and Contrafactives”. In: *Aristotelian Society Supplementary Volume* 91.1, pp. 245–266. DOI: 10.1093/arisup/akx003.

Roberts, T. and D. Özyildiz (2023). “Bad attitudes: Impossible meanings and the false belief gap”.

Steinert-Threlkeld, S. (2020). “An Explanation of the Veridical Uniformity Universal”. In: *Journal of Semantics* 37.1, pp. 129–144. DOI: 10.1093/jos/ffz019.

Strohmaier, D. and S. Wimmer (2022). “Contrafactives and Learnability”. In: *Proceedings of the 23rd Amsterdam Colloquium*. Ed. by M. Degano et al. Amsterdam, pp. 298–305.

— (2023). “Contrafactives and Learnability: An Experiment with Propositional Constants”. In: *Logic and Engineering of Natural Language Semantics*. Ed. by D. Bekki, K. Mineshima, and E. McCready. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, pp. 67–82. DOI: 10.1007/978-3-031-43977-3_5.

Vaswani, A. et al. (2017). “Attention is All you Need”. In: *31st Conference on Neural Information Processing Systems*, pp. 1–11.

Pseudo-scoping out of tensed clauses: cumulation vs. buildups

0. Introduction. Recent papers have argued that tensed clauses are not scope islands for universal quantifiers (Barker, 2022; Hoeks et al., 2022). One reason is that (1-a) allows a reading where, instead of a particular student responsible for every speaker's ride, the student can vary by speaker (henceforth, a *variation reading*). In contrast, (1-b) does not allow a varying reading.

(1) a. A student made sure that [every invited speaker had a ride].
b. A student claimed that [every speaker had a ride]. ✗ $\forall > \exists$

Assuming tensed clauses are not scope islands for universals requires imposing some predicate sensitive restriction on scope taking, to rule out the variation reading in (1-b). Hoeks et al. (2022) propose that QR is only possible if the eventuality described by the quantification is in some intuitive sense ''build up to' over time by the individual cases of the quantification ("buildup approach"). The lexical semantics of *make sure* inherently requires such a buildup, but *claim* does not, resulting in the impossibility of QR in (1-b). This abstract offers a different explanation for (1): (1-a) doesn't in fact involve scope taking, but receives its variation reading through a cumulative inference (CI) ("cumulating approach"). (1-b) is impossible because tensed clauses are scope islands after all.

1. Evidence for cumulating approach. We propose that the cumulativity responsible for variation readings is not the prototypical kind involving a relation between two pluralities. Rather, CIs involve a cumulative contribution between the members of a subject plurality resulting in the truth of the embedded proposition (Harada, 2022). In (2), the predicate *make sure* licenses an inference combining the contributions of *Ann* and *Bea*, resulting in the truth of the embedded proposition: *that every problem was error-free*. Licensing this inference depends on the semantics of *make sure*. Crucially, CIs are not available with every embedding predicate when there's a conjoined subject: *claim* can't cumulate contributions together like *make sure*, as illustrated in (3).

(2) CONJOINED SUBJECT/VARYING INDEFINITE CONTEXT: *[Ann and Bea are teaching assistants. The professor asked the teaching assistants to review four homework problems. Ann made sure the first and second problems were error-free, but didn't look at the third and fourth problems. Bea made sure the third and fourth problems were error-free, but didn't look at the first and second problems.]*
{Ann and Bea/A teaching assistant} made sure that every problem was error-free.

(3) CONJOINED SUBJECT/VARYING INDEFINITE CONTEXT: *[Ann and Bea are teaching assistants. The professor asked the teaching assistants to review four homework problems. Ann claimed that the first and second problems contained errors, but had no issues with the other problems. Bea claimed that the third and fourth problems contained errors, but had no issues with the other problems.]*
{#Ann and Bea/#A teaching assistant} claimed that every problem contained errors.

This predicate-sensitivity of CIs is not limited to *make sure* and *claim*; it also correlates with apparent inverse scope. We ran a series of acceptability rating tasks to show that the same predicates which license CIs give rise to apparent inverse scope. The task involved 10 predicates: 5 which license CIs (*make sure, confirm, establish, prove, verify*—henceforth, *cumulating predicates*) and 5 which don't license CIs (*claim, notice, confess, heard, believe—non-cumulating predicates*). Sample contexts for conjoined subject and varying indefinite conditions are illustrated in (2)–(3) with the bolded target sentences. Controls involved non-conjoined/non-varying indefinites that simply referred to a single individual. Figure 1 illustrates a higher acceptability of CIs with

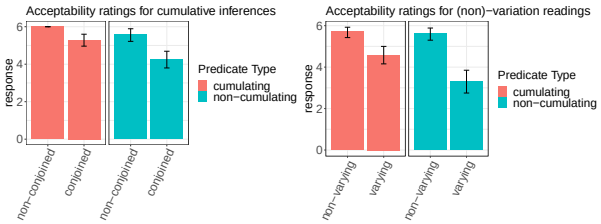


Figure 1: Left: CIs with plural subjects. Right: Variation readings with singular indefinites.

plural subjects (left-hand plot) and variation readings with indefinites (right-hand plot) for cumulating predicates (red bars) compared to non-cumulating predicates (blue bars). The significance of this interaction in mixed effects models supports the empirical generalization in (4). These results can be made sense of if tensed clauses are scope islands after all and apparent wide scope is illusional, derived indirectly via CIs, which cumulate the contribution of each witness of the indefinite.

(4) THE CUMULATING CORRESPONDENCE: A clause embedding predicate will license variation readings (i.e. apparent wide scope of a universal) whenever the predicate licenses CIs.

2. Evidence against buildup approach. Apart from (1), the crucial empirical argument for the buildup approach in Hoeks et al. (2022) is that the variation reading should become available for embedding predicates like *claim* (and others, like *heard, found, become aware* and *believe/come to believe*—hencerforth *buildupicle predicates*), when additional cues force a buildup reading. The two manipulations given are (i) adding a buildup adverbial like *by 8pm*, and (ii) using perfect aspect. This is illustrated by Hoeks et al. (2022) in (5) which they report licenses a variation reading.

(5) By 8pm, a student had claimed that every professor had a ride. ✓$\forall > \exists$

Hoeks et al. (2022) furthermore report that some other embedding predicates do not allow scope taking even with these cues to buildup, for example *is confident, is sure, is aware, is convinced, realize* and *remember* (henceforth *non-buildupicle predicates*). We ran an acceptability rating experiment to test these predictions. The task compared buildupicle and non-buildupicle predicates in buildup and non-buildup contexts. Sample contexts and target sentences are provided in (6)–(7). Controls involved non-varying indefinites that simply referred to a single individual. Results are in Figure 2.

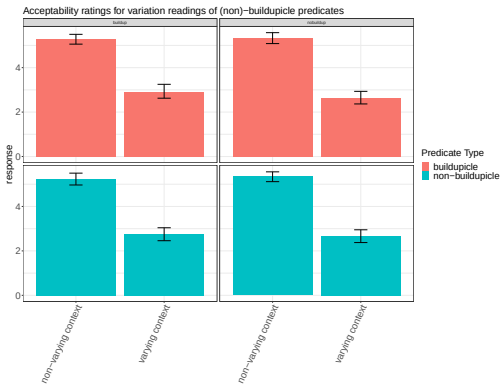


Figure 2: Left: Non-varying and varying indefinite contexts involving buildups. Right: Non-varying and varying indefinite contexts involving no buildup.

(6) BUILDUP, VARYING INDEFINITE CONTEXT: *[Ann, Bea and Carol are students. During yesterday's talk, the speaker presented three theories in total. When the speaker presented the first theory, Ann claimed it was wrong. When the speaker presented the second theory, Bea claimed it was wrong. Finally, when the speaker presented the third theory, Carol claimed it was wrong.]* **By the end of the talk, a student had claimed that every theory was wrong.**

(7) NO BUILDUP, VARYING INDEFINITE CONTEXT: *[Ann, Bea and Carol are students. At yesterday's talk, the speaker presented three theories. During the final discussion, Ann claimed the first theory was wrong, Bea claimed the second theory was wrong and Carol claimed the third theory was wrong.]* **A student claimed that every theory was wrong.**

Figure 2 illustrates no difference in acceptability for buildupicle predicates (red bars) compared to non-buildupicle predicates (blue bars) or between buildup (left-hand plots) and non-buildup contexts (right-hand plots). Buildipcle and non-buildupicle predicates are rated worse than non-varying controls and just as bad as the non-cumulating predicates from experiment 1. Thus, the empirical claim made by Hoeks et al. (2022) concerning (5) is not borne out. Variation readings are unavailable for these predicates in contrast to predicates that license CIs, which allow variation readings even without buildup cues, as shown in experiment 1. The cumulating approach dispenses with the need for QR to derive the variation reading, and in the process dispenses with imposing a buildup constraint on QR. The CI in some sense captures the intuition, however, that the truth conditions of (1-a) involve adding up the individual cases toward the overall reading.

References. Barker, C. (2022). Rethinking scope islands. In *LI* 53(4), 633-661. | Hoeks, M., Özyıldız, D., Pesetsky, J., Roberts, T. (2022). Event plurality & quantifier scope across clause boundaries. In *SALT* (Vol. 1, pp. 443-462). | Harada, M. (2022). *Locality effects in Composition with Plurals and Conjunctions* (PhD Thesis, McGill University). |

Reduced sensitivity to underinformativeness? Using a ternary judgment task to assess scalar implicature generation in L2 and L1

Natural language utterances can often receive more than one interpretation. For instance, the literal meaning of (1) corresponds to (2). However, (1) can also be interpreted as in (3):

(1) Some of my friends studied linguistics
(2) Literal interpretation: At least one of my friends studied linguistics
(3) Pragmatic interpretation/Scalar Implicature: Not all of my friends studied linguistics

According to the Standard Pragmatic Model (Grice, 1975 and subsequent work), the interpretation in (3) is pragmatically derived via an inferential process (Scalar Implicature/SI generation) whereby comprehenders take the usage of the weaker term to imply the negation of the stronger alternative on the same scale (*some ~> some but not all*)

Despite the fact that both (2) and (3) are easily accessible to typical adult language users, children strongly prefer literal interpretations and do not generate SIs at adult-like rates until relatively late in language development (e.g., Noveck, 2001). Interestingly, a similar pattern is found in (adult) L2 speakers: in the classic binary choice tasks, L2 speakers tend to accept underinformative sentences like “Some elephants are mammals” more frequently than L1 speakers, and the SI rate appears modulated by L2 proficiency (Khorsheed et al., 2022).

Do children and L2 speakers perform similarly for similar reasons? According to the Pragmatic Tolerance Account (Katsos & Bishop, 2011), children generate fewer SIs than adult (L1) speakers not because they lack the necessary pragmatic competence, but rather because - unlike adults - they are generally tolerant towards pragmatic violations. Indeed, in Ternary Judgment Tasks (TernJT), a task in which instead of binary response options (“False”, “True”), participants are given a ternary scale with an intermediate option (“A bit true”), children and adults perform alike: they judge underinformative *some*-sentences choosing the intermediate option. According to Katsos and Bishop (2011), this finding suggests that children, albeit more tolerant towards violations, are as sensitive as adults to underinformativeness and, when given the chance, can demonstrate an adult-like pragmatic competence.

With this study, we aimed to investigate whether pragmatic tolerance plays a role also in L2 pragmatic processing. Specifically, we hypothesize that the processing difficulties connected to comprehending a foreign language might make L2 speakers pragmatically more tolerant than adult L1 speakers: if this is the case, pragmatic tolerance, not a difficulty with SI generation, may be responsible for the reduced rate of SIs attested in L2.

Method

Ninety-one participants (43 L1 Dutch speakers and 48 Dutch L2 speakers of English) took part in our experiment. L2 proficiency was assessed by means of the LexTALE task (Lemhöfer & Broersma, 2012) and used to divide (by median split) the L2 participants in two groups (Low vs. High Proficiency). Following Bott and Noveck (2002), the experiment included *some*-Underinformative sentences (“Some pets are dogs”) and 5 types of control sentences with the quantifiers *all* and *some* (*some*-True, *some*-False, *all*-True, *all*-False, *all*-FalseAbsurd). Participants performed a TernJT: they were asked to judge the sentences by choosing between “False”, “A bit true”, or “True”.

Results

Performance on control conditions was as expected in L1 and L2 groups: false sentences were overwhelmingly rejected and true sentences accepted; the middle option was hardly ever selected. Participants’ responses in the critical condition *some*-Underinformative are shown in Figure 1. Regression analysis confirmed that the intermediate option was less likely to be selected compared to the other responses ($\beta = -3.7$, $p < .001$) and the L1 and L2 groups did not

differ in their tendency to choose the intermediate option as opposed to the other choices.

Furthermore, to assess participants' tendency to accept the underinformative sentences, we created a factorial outcome variable with two levels: "acceptance" ("True") vs. "other response" ("False" and "A bit true") and found that the L2_Low Proficiency group was more likely (β = 1.24, $p < .05$) than the other two groups (L1 and L2_High Proficiency) to fully accept *some*-Underinformative ("True").

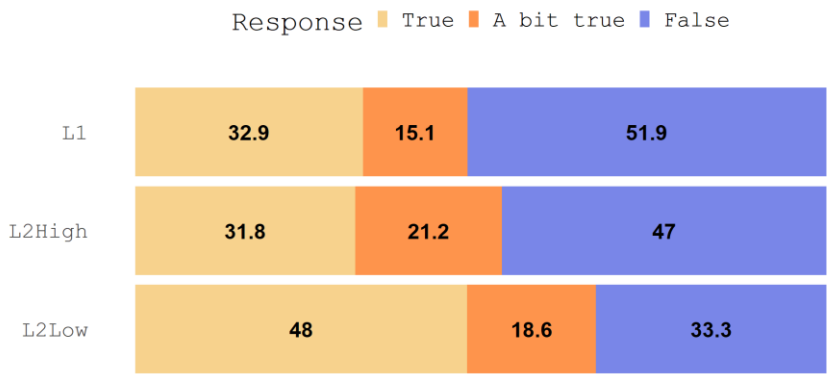


Figure 1: Percentage of responses of the L1 and L2 (High Proficiency vs Low Proficiency) groups on the Ternary Judgment Task

Discussion and Conclusions

In line with previous literature, our study brings additional support to the observation that L2 proficiency modulates the rate of acceptance of underinformative sentences: our L2_Low Proficiency group accepted *some*-Underinformative utterances 48% of the time (vs. 32% in L2_High Proficiency). At the same time, our study does not suggest that L2 speakers differ from L1 speakers in terms of pragmatic tolerance: despite the availability of an intermediate option, L2 speakers (irrespective of proficiency) were not likely to judge *some*-Underinformative sentences more often as "A bit true". Taken together, these findings suggest that, despite the use of a TernJT, L2 speakers show a reduced sensitivity to underinformativeness (modulated by proficiency) that is not attributable to a tolerant attitude towards pragmatic violations.

Finally, our study suggests that the reliability of TernJTs for gauging inferential skills should not be taken for granted. In fact, neither our L2 groups nor, importantly, our L1 group, behaved as expected in the TernJT: even these latter participants failed to preferentially select the intermediate response. An unexpected behavior in the control group has emerged before in previous studies with TernJTs (e.g., Wampers et al., 2018); this high variability in the performance of the control group casts doubt on the idea that the TernJT can be used as a fine-grained, more sensitive measure to assess and uncover differences in the pragmatic skills of different populations.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (Vol. 3, pp. 41–58). Academic Press.

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

Khorsheed, A., Rashid, S. M., Nimehchisalem, V., Imm, L. G., Price, J., & Ronderos, C. R. (2022). What second-language speakers can tell us about pragmatic processing. *PLoS ONE*

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343.

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

Wampers, M., Schrauwen, S., de Hert, M., Gielen, L., & Schaeken, W. (2018). Patients with psychosis struggle with scalar implicatures. *Schizophrenia Research*.

A conceptual analysis of verbs of *pushing* and *pulling*

Within the theory of *Conceptual Spaces* (Gärdenfors, 2000), concepts are analysed as regions in multi-dimensional spaces which are derived from (fine-grained) semantic dimensions. Such dimensions are assumed to be derivable to a large extent from perception (Gärdenfors 2000, 2007; Gärdenfors & Warglien 2012). Previous research has provided profound evidence for a geometrical organization of concepts in the (direct) sensory domain, such as colour (Berlin et al., 1969), olfaction (Majid et al., 2018), static spatial relations (Levinson et al., 2006), and even prototypical instances of motion events (Giese et al. 2008, Malt et al. 2014). However, less progress has been made in the conceptual space of actions and events involving both an agent and a patient such as events of *pushing* and *pulling*.

Recent studies have used simple 2D videos to elicit naming of basic pushing and pulling events focusing on the difference between verb- and satellite-framed languages (e.g. Hickmann et al., 2018; Montero-Melis, 2021). However, these studies do not allow a fine-grained identification of the relevant semantic properties needed to develop a semantic analysis of such verbs, which must be considered an important desidaratum in cognitive linguistics.

Based on the assumption that "the fundamental cognitive representation of an action consists of the pattern of forces that generates it" (Gärdenfors and Warglien 2012: 498; cf. also Talmy 1988), the present study presents the results of a free production experiment that aimed at assessing in more detail which semantic dimensions make out the domain of *pushing* and *pulling* as a fundamental domain of physical interaction between agents and patients. Pinpointing conceptual boundaries requires investigating *peripheral event instances*, which leads to large number of combinatorial possibilities to be tested in a systematic explaration of conceptual spaces. We approached this problem by presenting participants with short 3D video clips in which a computer-animated agent moved a barrel a short distance, allowing for fine-grained adjustments of potentially impactful properties. Among the numerous dimensions possibly involved, we manipulated four: i) the angle of contact between agent and object, ii) the strength of force used by the agent, iii) the duration of contact, and iv) the agent's orientation (facing the object or the direction of movement). In our study, the main research goal was to determine the predictors that trigger the production of different verbs and to classify them in semantic verb clusters. The role of modifiers of various types is not discussed in this presentation.

Methods. The 3D videos involved a human-like agent causing the movement of a barrel (see Fig. 1). The 3 second videos were created using a state-of-the-art physics engine according to a $7\times2\times2\times2$ fully within-design with the factors **Angle between human and barrel** (0, 45, 90, 105, 120, 135, 180), **Barrel movement** (*continuous* vs. *instantaneous*), **Facing direction** (*towards barrel* vs. *forward in direction of movement*) and **Force** (*heavy* vs. *light*). This resulted in a total of 52 trials (at 0 degrees, facing direction cannot be differentiated). We recruited 81 native speakers of German (45 female; mean age: 24.5) via Prolific, who were told that they should provide descriptions rich enough to categorize the videos for a second group of participants. After each video, participants were prompted to answer the question *What does the person do with the barrel?* (in German), for which the following prompt was provided: *The person ….*

Data. We gathered a corpus of 4212 descriptions (word length range: 3–70, mean 8.7). We annotated the main matrix verbs that expressed movement of the barrel (in addition to a number of other properties not yet finalized). We found 95 different matrix verb constructions with 9 matrix verbs that have a frequency $> 0.5\%$: *ziehen* 'pull' (1635), *schieben* 'push' (1156), *drücken* 'push' (195), *schubsen* 'shove' (195), *stoßen* 'poke' (176), *gehen* 'walk' (173), *bewegen (reflexive)* 'move oneself' (102), *bewegen* 'move' (71), *laufen* 'walk' (29). **Results.** K–means clustering ($k = 3$) for

Movement (*continuous* vs. *instantaneous*) identified 3 clusters definable by binary features: *bewegen (refl.)*, *gehen*, *laufen* with feature +cont(inuous); *schubsen*, *stoßen* with +inst(antaneous); and the other verbs are unmarked in relation to the cont/inst distinction. For all −cont-verbs the barrel is realized as the direct object, for all +cont-verbs it is embedded in a PP (e.g., *'move oneself with the barrel'*). For verbs produced by at least 15 participants, we fitted linear mixed effect models with random intercepts for participants and Cos_Angle (Cos), Movement,[1] Force, and Facing as fixed effects. Predictors vary for individual verbs. We found the following stable patterns with respect to Cos: Cos was no significant predictor for the remaining +cont-verbs; all other verbs are either positively or negatively correlated with Cos, see Tab. 1, except *bewegen* (*move*) which also did not correlate with Cos. Other predictors (Force, Facing) may correlate with individual verbs, but we found no general pattern correlated to verb clusters.

Discussion. For the factors manipulated in the videos, conceptually clearly distinguishable verb clusters can only be defined by the Movement feature, which tells us whether the agent moves together with the barrel (+cont), or is unmoved (+inst), and the Cos of the angle. Interestingly, the results provide little evidence that verbs are categorized according to the Force applied to the barrel (as predicted by Gärdenfors/Warglien's theory). It is rather the movement and position of the agent in relation to the barrel that determine production of verbal descriptions.

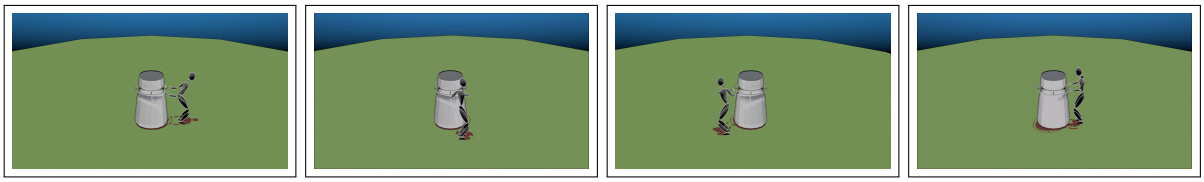
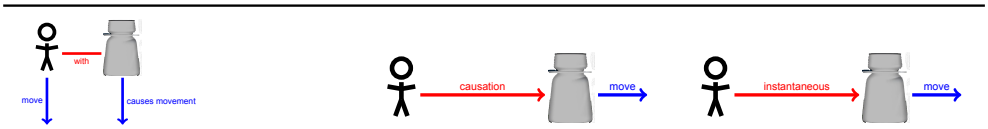


Figure 1: Stills for $180°, 105°, 135°, 0°$ with left/right movement and agent facing forward or to object.



+ continuous	unmarked		+ instantaneous
+ Sin_Angle	+ Cos_Angle	− Cos_Angle	+ Cos_Angle
bewegen (REFL)-PP (*move self with*)	schieben (*push*)	ziehen (*pull*)	stoßen (*push*)
gehen-PP (*walk with*)	drücken (*press*)		schubsen (*push*)

Table 1: Verb clusters: semantic feature (red), use correlated (blue).

Bibliography: • Berlin, B. et al. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California Press. • Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: The MIT Press. • Gärdenfors, P. and M. Warglien (2012). Using conceptual spaces to model actions and events. In: *Journal of semantics* 29.4, 487–519. • Hickmann, M. et al. (2018). Caused motion across child languages: a comparison of English, German, and French. In: *Journal of Child Language* 45.6, 1247–1274. • Levinson, S. C. et al., eds. (2006). *Grammars of Space: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press. • Majid, A. et al. (2018). Differential coding of perception in the world's languages. In: *Proceedings of the National Academy of Sciences U.S.A.* 115.45, 11369–11376. • Montero-Melis, G. (2021). Consistency in Motion Event Encoding Across Languages. In: *Frontiers in Psychology* 12.625153. • Talmy, L. (1988). Force dynamics in language and cognition. In: *Cognitive science* 12.1, 49–100.

[1] We dropped Movement for +cont- and +inst-verbs. Angle was re-scaled to cosin to resolve convergence issues.

The effect of context on the online processing of adversatives: an eyetracking study

The meaning of adversative connectives such as English *but* relies on an opposition between the properties and entities of the propositions it conjoins. The exact **nature** of that opposition is discussed in most works on the semantics of *but*. This research focuses on identifying **what** in the conjuncts is used as information to be subject to this opposition, and **when/how** this information is taken into account in the processing of adversative conjunctions, especially in relationship with contextual information. We rely on sentences such as (1) which involve an interaction between adversative conjunctions and comparative structures involving gradable predicates.

1. Alex is tall, but (less/ * more) tall than Riley.

Author et al. (2014) report that speakers judge that superiority comparatives (*more X than*) in the conjunct introduced by *but* are degraded compared to inferiority ones; a difference which disappears when the target sentence is placed in a context that facilitates the contrast as in (2).

2. We are looking for a stunt double to replace Riley, an actor, in a movie. The stunt double must be of the same height as Riley for the scene to be believable. The hunt proves to be difficult because Riley is tall. Alex is considered as a potential double.

Using a self-paced reading paradigm, the same study shows that despite the effect of context on offline acceptability judgments, the superiority cases still show significantly longer reading times in the post-*but* regions, suggesting that online processing remains affected by these constructions, and that contextual information is integrated at a later stage of the interpretation of the conjunction. Building on these results, we use an eye-tracking paradigm to investigate in more detail the processes of interpretation of sentences like (1). Specifically, one issue at stake is how early contextual information is integrated in the processing of adversatives. On one hand, relevance theoretic accounts of adversative consider that the interpretation of adversatives relies on the identification of a *pivot inference* made *accessible* by the first conjunct and that gets contradicted by the second conjunct (Blakemore, 2002): the more accessible that pivot, the easier the interpretation of the conjunction. On such accounts, we thus expect that contexts as in (2) should facilitate every aspect of the interpretation of (1). In contrast, within theories like argumentation within language (AwL: Anscombre and Ducrot, 1983, Author, 2019), the search for the pivot inference is first driven by lexical information, and then complemented by contextual information. Given that a predicate *P* and a form like *more P* are lexically not in opposition (Author, 2019), we expect to observe an effect of the choice of construction ("*more"* / "less") on measures that reflect the processing of information, even within contexts that facilitate the interpretation.

We considered two binary variables in the experiment: one for the nature of the `Context` (*Neutral/Helping)* and a `Valence` for the choice of construction *(Positive: "more than"/Negative: "less than")*. Materials for the experiment were produced in Quebec French, using "*mais"* as an adversative with target items comparable to example (1) and the context in (2) (as a *Helping* context, *Neutral* contexts involved material unrelated to the target predicate). We used 20 target items, meaning that participants saw each combination of conditions 5 times. 40 filler items were interspersed with target items, for a total of 60 items, presented using a pseudo-random design. 55 native speakers of Quebec French were recruited, sat in front of a computer screen equipped with a Tobii Pro Fusion 250 Hz eye tracker and were asked to read the sentences on the screen and answer a comprehension question for each item. Participants were compensated 15$.

Measures were considered at two levels of analysis: at the sentence level, we relied on the total duration ratio (total fixation time in milliseconds divided by the number of characters in a sentence), taken as an indicator of overall sentence reading difficulty (Clifton et al., 2007). At the word level we used go-past duration, one of numerous indicators for higher-level processing

during online reading (Cook & Wei, 2019). Each measure was taken as the dependent variable in a linear mixed effect model (lme4 R package, Bates et al., 2014), using random intercepts for items and participants and assessing the significance of factors via the R package "moments" (Komsta & Novomestky, 2015). At the sentence level (Fig. 1), we found a marginally significant interaction between `Valence` and `Context` (t = 1.682, p= 0.093), and at the word level (Fig. 2), an effect of `Context` (t=1.825, p=0.0681), as well as an interaction between `Valence` and `Context` (t=-1.740, p=0.0819).

Figure 1 : Total fixation duration by experimental conditions

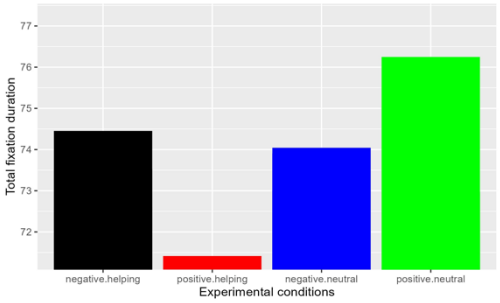
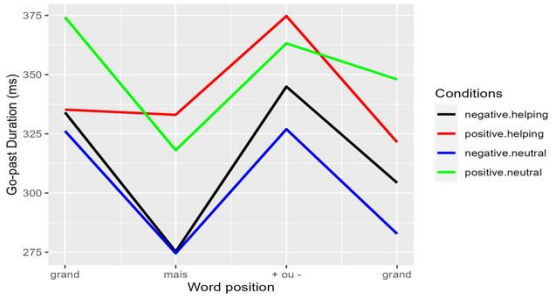


Figure 1 shows the effect of `Context:` in the *positive* condition *helping* contexts lower the total time spent reading the sentence, unlike in the *negative* condition in which `Context` has no effect. This is consistent with the AwL hypothesis that inferiority comparatives are lexically opposed in a way that make them compatible with the semantics of *but* without recourse to context, contrary to superiority comparatives, which prompt readers to access the necessary pivot via abduction of the contextual information. When that inference is contextually accessible, the total reading time is overall reduced, though not in the *Negative.helping* cases. Nevertheless, as shown on Fig. 2, participants still take more time to go past *"mais"* and the second gradable adjective on a first reading in the *Positive* cases, irrespective of the nature of context. This is again congruent with the AwL-based hypothesis that the processing of information is initially lexically based: the contrasted predicates in the *Negative* are in lexical opposition, unlike the ones in the *Positive*, which accounts for their higher processing times.

Figure 2 : Go-past duration per word position



Our results thus seem to generally support the hypothesis that the interpretation of adversative conjunctions relies on the lexical properties of its conjunct, before integrating potential contextual information. In that way, our results seem to contradict the predictions of Relevance Theory. Note however that those predictions seem borne out at the sentence level (in a marginally significant way), suggesting that RT might be on the right track as far as secondary inferential processes are concerned. Overall, we thus take our results to be consistent with a two-time interpretation process: extracting the opposition pivot from lexical properties first, then searching for it in context (or the memory of context) if that failed. This is in line with the general claims of AwL about "integrated pragmatic" effects in the semantics of certain linguistic expressions. Further work will analyze other eye-tracking measures, in particular backward regressions to investigate which elements are perceived as problematic in the processing of *Positive* cases.

References: Anscombre, J-C. and Ducrot, O. (1983) *L'argumentation dans la langue* Pierre Mardaga ♥ Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*. ♥ Blakemore D. (2002) *Relevance and Linguistic Meaning. The semantics and pragmatics of discourse markers.* CUP ♥ Clifton Jr, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. *Eye movements*, 341-371. ♥ Cook, A. E., & Wei, W. (2019). What can eye movements tell us about higher level comprehension? *Vision*, *3*, 45. ♥ Komsta, L., & Novomestky, F. (2015). Moments, cumulants, skewness, kurtosis and related tests. *R package version*, *14*(1).

Pragmatics of human-AI communication

Introduction. Within linguistics and the philosophy of language, discourse is formalized in terms of mental content – beliefs, goals and motivations of conversation participants – with the joint goal of mutual understanding (Grice 1975 et seq.). The development of large language models (LLMs) introduces new kinds of communicative settings where the standard mentalistic approach to discourse may not be appropriate. Because LLMs don't have a human mind, and likely don't have the same kind of motivation structure as humans do, people might employ distinct strategies when talking to machines. In this work, we explore whether such an alternative strategy for communication is actually employed for machine-generated linguistic content. Focusing on a distinction between asserted and presupposed content in human communication and the different use conditions governing each type, we ask: (1) whether humans uptake information differently when generated by an AI which they are told is unreliable, and (2) whether information processing is affected by whether the content is packaged as asserted vs. presupposed.

Background. We take as our starting point a model of discourse based on proposals by Stalnaker (1974, 1978). On this model, sentences used in communication contribute to the conversational **common ground**, the set of shared beliefs among discourse participants. The model distinguishes two kinds of linguistic content in the way they affect the common ground. **Asserted** content is put forth with the explicit intent to change the listener's beliefs and expand the common ground. In contrast, **presupposed** content must already be part of the common ground, or be accommodated, before that common ground can be updated with the assertion. Crucially, novel information packaged in these two forms have distinct effects on belief change: asserted content is presented to the listener as up for debate, giving them the option of accepting or rejecting. Novel presuppositions, on the other hand, are things the speaker expects a cooperative listener to tacitly add to their own beliefs, and in turn to the common ground. Listeners infer based on the utterance what the speaker wishes to take for granted, and trusting them not to mislead, shifts to the intended common ground.

Hypotheses. The common ground model takes exchange of information as grounded in the beliefs and intentions of interlocutors, and it is possible that the way humans intake information from machines, which lack such a mental apparatus, is different. The model, furthermore, distinguishes the type of belief revisions a listener is expected to do on the basis of whether a piece of new information is asserted vs. presupposed. Presuppositions can lead listeners to adjust their beliefs without much deliberation or discussion. This type of tactic belief change – which relies on reasoning about what the speaker wants to be common ground – may not happen when communicating with an AI. In that case, new information should be treated as new and up for debate, irrespective of how it is packaged. Another possibility is that presupposition accommodation is automatic, and people are prone to accept and go along with AI presuppositions, even when they may challenge AI assertions.

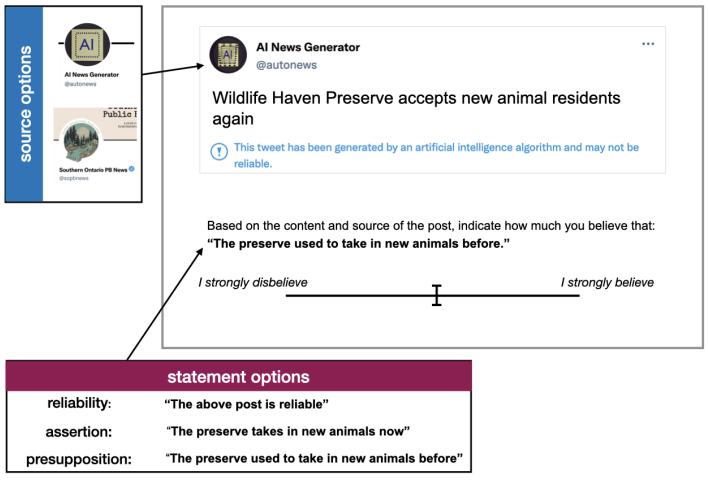


Figure 1: structure of the experimental task

Experiment. Participants (N=205) were asked to read constructed social media posts, and a potentially related follow-up statement, after which they evaluated the extent to which they believe that statement on a slider from “strongly disbelieve” to “strongly believe” (Figure 1). Crucially, each post contained a presupposition trigger (e.g,. *again*). In a 2x4 within-subjects design, we manipulated two factors: (1) **source of information**: human (a news outlet, Southern Ontario Public Broadcasting) vs. AI (AI algorithm tasked with constructing news-like posts; AI-posts indicated that the post’s content may not be reliable); (2) **follow-up statement type:** all participants saw 4 types of follow-up statements: (i) explicit statements about the reliability of the post (*Reliability* condition), (ii) asserted content from the post (*Assertion* condition), presupposed content from the post (*Presupposition* condition), and (iv) unrelated content from the post (*Unrelated* condition). Unrelated trials were used for exclusion and do not figure in analyses.

Results. See Figure 2. There are three findings of note. We found a main effect of source (β= -25.38, p<.001): participants indicated lower belief in AI content overall compared to human-generated content, perhaps unsurprisingly given that they were told that the AI content was unreliable. This finding shows that humans can modulate their trust in information based on source, at least when reliability issues are highlighted. Second, and strikingly, we found a significant difference (β= -4.05, p<.001) between participants' ratings of AI reliability and their endorsement of AI content (assertions and presuppositions): participants endorsed AI-generated content significantly more than they endorsed its reliability. In other words, perceived low reliability of AI did not fully prevent participants from updating their beliefs with the content it produced. Finally, we found a small but significant difference between AI-assertions and AI-presuppositions, with participants indicating greater belief in presupposed content (β= 1.62, p=.01). This suggests that people are ready to accommodate, rather than challenge, AI-presuppositions, despite the conversational setting not obviously licensing such behavior.

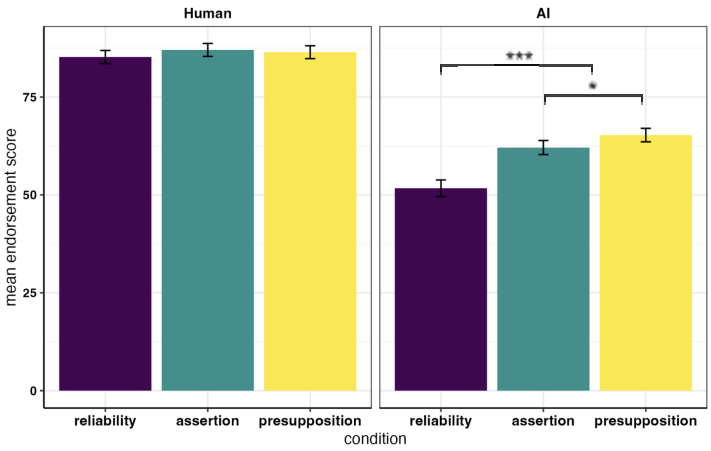


Figure 2: Mean endorsement scores, indicating level of belief in each follow-up; error bars represent 95% CI

Conclusions. AI-generated language presents new questions, both theoretical and practical, about how our beliefs evolve over the course of a conversation. In this study we found that, despite the fact that machines might lack human-like mental states, people treated AI-generated language as constrained by the same principles as those found in human language. On a theoretical front, this finding implies that humans tend to perceive any natural language as human-like. On a practical front, it raises questions about how humans can be aided to encode AI language appropriately, rather than imbuing it with human motivations.

References. Grice, H. P. (1975). Logic and Conversation. In P. Cole and J. L. Morgan (eds.), *Syntax and Semantics, Vol. 3, Speech Acts* (pp. 41-58). New York: Academic Press. Stalnaker, R. (1974). Pragmatic presuppositions. In M. Munitz and P. Unger (eds.), *Semantics and Philosophy* (pp. 197-214). New York: NYU Press. Stalnaker, R. (1978). Assertion. Syntax and Semantics 9: pp. 315-332.

Identifying QUDs in Naturalistic Discourse Anonymous ELM 3 Submission

The Question Under Discussion (QUD) model has been an influential theoretical device in pragmatics, but efforts to derive QUDs from naturalistic data are few. In this work, we crowdsource QUD annotations of radio interviews. We address a fundamental issue at the center of QUD theory: can discourse agents reliably infer an implicit question or questions being addressed in naturalistic discourse? The secondary question we address is whether, as most QUD theories presuppose, there are multiple salient QUDs. We compare several similarity metrics for questions and answers, demonstrating that our user interface encourages annotators to obey useful theoretical constraints like Q-A Congruence. Overall, we find moderate annotator agreement forming qualitatively identifiable clusters, consistent with the existence of multiple contextually-restricted immediate QUDs. We further find, unexpectedly, that annotators are unreliable at reconstructing masked overt questions, suggesting that explicit questions may correspond to QUD/topic shifts.

Background Roberts 2012/1996 characterizes discourse as a game in which possible moves (utterances) are guided by whether they help answer the immediate QUD, usually a single implicit question. QUDs help formalize Gricean Relevance, important also in theories of focus, exhaustivity, coherence, etc. Existing resources fall broadly into two camps: rigorous, theory-grounded approaches, such as the hierarchical annotations in De Kuthy et al. 2018 and Hesse et al. 2020, albeit limited in scope by ontological complexity; or large, crowdsourcing approaches working with various kinds of implicit question, such as evoked questions (Westera et al., 2020) or elaborations (Wu et al., 2023), albeit not necessarily targeting theoretical properties of immediate QUDs.

Procedure We selected 10 complete two-party dialogue transcripts from INTERVIEW (Majumder et al., 2020), a corpus of NPR interviews in American English, split by sentence and annotated with turn information. Episodes were chosen to have between 29 and 32 sentences, of which at least 5 were overt questions (μ=5.5). 10 native English speakers per episode were recruited on Prolific, resulting in 100 unique sets of annotations. For each episode, annotators read the dialogue one sentence at a time, in a moving two-sentence window to simulate linear processing, as inspired by Westera et al. 2020. For each new sentence, annotators were prompted to (i) write a question that can be answered by that sentence, and (ii) select a contiguous span from that sentence best representing the answer to their question (Fig. 1). Annotators could opt to mark "no clear question" (e.g., for moves like *Good morning.*) While participants were free to write any question that the sentence addresses, we assume that discourse context makes certain potential QUDs more likely.

Evaluation We consider several similarity metrics for measuring QUD agreement. The first is *token edit distance* (ED), which counts the minimum number of words that must be inserted, deleted, or substituted to transform one array of tokens into another. This metric is useful for measuring answer similarity (μ=6.6), since all answers are forced by our interface to be subsets of the target sentence. Measuring similarity among questions is more challenging. Assuming Q-A Congruence, we hypothesize that annotators who select similar (low ED) answer spans are more likely to be writing similar QUDs, since they place focus on the same information. To test this hypothesis, we look at how answer ED correlates with three question similarity metrics: *question edit distance* (μ=8.0); rescaled *BERTScore* (Zhang et al., 2020) (μ=0.37), which encodes two sentences using a large transformer language model and measures the cosine similarity of their embeddings (between –1 and 1, where higher values are more similar); and *Wh-word agreement* (μ=0.39). Examples of these metrics applied to the collected data below in (1) are given in Table 1.

(1) a. Who else had been watching the radar? [**One of my graduate students**]
b. Who saw the occurrence and effects on the radar? [**my graduate student**]
c. Where are the clouds coming from? [**southwest about five miles**]

Results We find a moderate correlation for answer ED and question ED (Spearman's ρ=0.41), as well as for answer ED and BERTScore using DeBERTa (ρ=−0.37), the model recommended

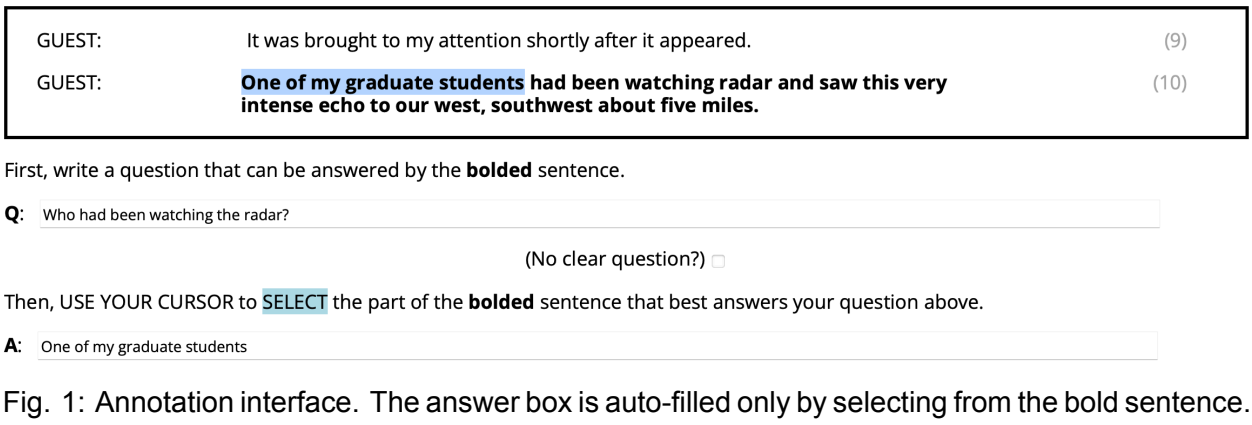


Fig. 1: Annotation interface. The answer box is auto-filled only by selecting from the bold sentence.

Metric	(a,b)	(b,c)	(a,c)
Token ED (A)	3	4	5
Token ED (Q)	6	8	7
BERTScore (Q)	0.41	0.14	0.12
Wh-word (Q)	1	0	0

Table 1: Similarity metrics on (1).

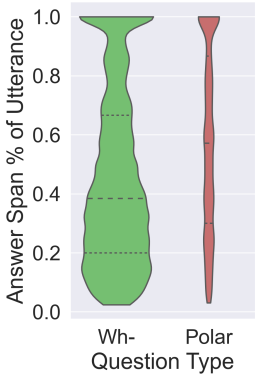


Fig. 2: Ans. spans.

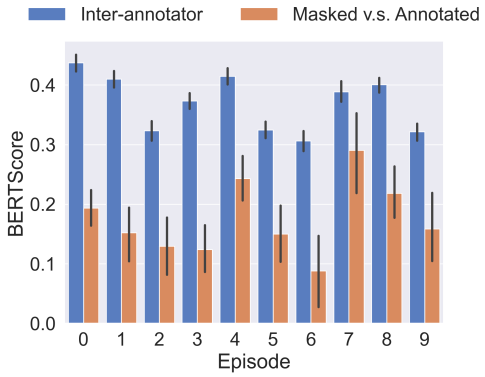


Fig. 3: Mean question similarity.

by the BERTScore authors. We also find correlations for Wh-word agreement with answer ED (ρ=−0.32) and BERTScore (ρ=0.49). The vast majority of QUDs written are Wh-questions, though polar questions exhibit an interesting pattern. With no explicit instruction to do so, for polar QUDs, annotators often select the entire sentence as their answer span (a response consistent with theoretical predictions about focus), while Wh-QUDs have short, constituent-sized spans (Fig. 2).

Masking questions Under most theories of QUDs, in normal circumstances, explicitly asked questions become the new QUD. To see whether annotator-written QUDs match actual questions, we masked all explicitly asked questions, keeping the sentence preceding it intact for context. We found that across episodes, annotators write QUDs consistently less similar to the masked question than to one another (Fig. 3), yet the mean inter-annotator BERTScore for QUDs on post-masked trials is not significantly different from inter-annotator agreement on normal trials. One possibility is that discourse participants may opt to ask explicit questions precisely in contexts with unpredictable topic shifts, making recovery difficult. Another is that our question similarity metrics fail to account for more general superquestions, a limitation of our linear, non-hierarchical method.

Conclusion Our results suggest that naturalistic discourse involves multiple compatible QUDs, but annotators are able to robustly extract these QUDs. The next step is extracting annotations about QUD hierarchy and relations among questions — a challenge we leave to future work.

References **De Kuthy et al. (2018)**. “QUD-based annotation of discourse structure and information structure”. *LREC 2018*. **Hesse et al. (2020)**. “Annotating QUDs for generating pragmatically rich texts”. *Workshop on discourse theories for text planning 2020*. **Majumder et al. (2020)**. “INTERVIEW”. *EMNLP 2020*. **Roberts (2012)**. “Information structure in discourse”. *Semantics and Pragmatics 5*. **Westera et al. (2020)**. “TED-Q”. *LREC 2020*. **Wu et al. (2023)**. “Elaborative simplification as implicit questions under discussion”. arXiv:2305.10387.

Probing 4 year old children's knowledge of the strong crossover constraint

This study uses a new task to probe four-year-olds' knowledge of a syntactic constraint on bound pronouns: strong crossover (CO). Previous studies on children's knowledge of CO use truth value/acceptability judgment tasks (Crain and Thornton 1998; McDaniel and McKee, 1986). Their findings have been widely taken to demonstrate that children as young as four years old respect CO. However, the materials in prior studies may have been pragmatically biased towards a crossover-respecting interpretation, raising the possibility that these tasks overestimated children's knowledge. In this study, we introduce a new methodology to probe children's knowledge of CO by removing the biases toward a crossover respecting interpretation. Our results show that while the design effectively reveals adult knowledge of crossover, children do not behave in an adult manner, suggesting that previous evidence that children do respect CO in embedded Wh questions should be treated with caution.

Tests of grammatical constraints that rely on interpretation must be sensitive to children's tendency to commit to interpretations and their reluctance to revise these commitments (Trueswell et al 1999, Omaki et al 2014). Prior tasks had pragmatic factors that allowed children to arrive at an interpretation of the pronoun before being exposed to the relevant syntactic configuration. Crossover respecting interpretations, then, could arise without ever making reference to the syntax. Our task addresses this concern by making both a bound and unbound interpretations available. Additionally, the task requires the child to make an inference about what happened in the world on the basis of their interpretation of a statement, rather than asking them to match a statement to what they already know happened in the world.

In the task, participants are first introduced to two dragons, Stella and Peter. Peter is trapped in a castle behind many locked doors. Peter explains that keys to each door are hidden in boxes and he has friends who can whisper a clue about the keys to Stella. The participant is then asked if they will help Stella if she gives them the clue (12 trials, 4 critical). On critical trials, the clue contains an embedded Wh-phrase and a pronoun which can either enter into a binding relation or not as a function of the presence of a crossover configuration. At the point of the clue, all conditions consist of the same actions and dialogue.

At the test sentence "I know who said she has the key" (non crossover) or "I know who she said has the key" (crossover), children must evaluate the relationship between who and she, while the salient nature of the whisperer should bias them toward interpreting the pronoun as the whisperer in the critical sentences in both conditions. Crucially, unlike in earlier studies, the sayer is never in doubt: it is the whisperer. The dependent variable, whose box the child chooses, corresponds to who the child believes has the key. That belief can only be determined by the form of the clue. Given the opportunity, 4 year olds readily bind pronouns (Koster & Koster, 1986; Thornton & Wexler, 1999), so any preference away from a bound interpretation (e.g. whisperer = key-haver) in the crossover condition should be taken as evidence that the clue has pulled them away from their prior bias to bind the pronoun.

32 adults (16 per condition) completed the study. Results from the adult study (shown in Figure 1) demonstrate that the task effectively reveals adults' sensitivity to crossover and that the control conditions adequately control for all other factors. 41 4-year-old children (range=4;0-4;11, mean =4;5; 21 in crossover condition; data collection ongoing, target N=48) were recruited from local preschools. In the test trials, the children do not differ by condition: children behave at chance. Their behavior is identical in the irrelevant trials (Figure 2).

Our results suggest caution when interpreting other findings of children's early knowledge of crossover. This design is effective at revealing the crossover constraint in adults, so children's failure may reveal a lack of knowledge of the crossover constraint. However, the results could simply be masking children's knowledge. They may have decided the clues are not particularly helpful, and thus not used them to make their choice. They may have been thrown off by the pragmatic oddness of the clue in the noncrossover condition. They may not prefer disjoint reference in cases where binding is not available; while adults prefer disjoint reference in these instances, it is not required by the grammar.

In this study, we corrected for one potential bias toward a crossover respecting interpretation in past studies of CO. Our results suggest that children do not reveal knowledge of the crossover constraint as readily as adults do. To further probe their knowledge, we must continue to modify experimental procedures and avoid extra grammatical biases.

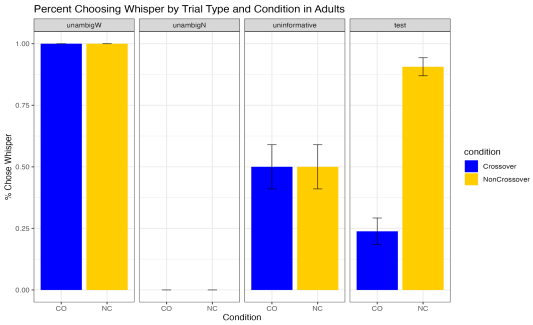


Figure 1: Adult's Baseline Response. Control items UnambigW and UnambigN were designed to probe willingness to choose both characters given unambiguous clues– Adults performed at ceiling. Uninformative condition tested baseline preference for each character without a disambiguating clue– Adult were at chance in both conditions. Test condition compares response to "I know who she said has the key" and "I know who said she has the key"

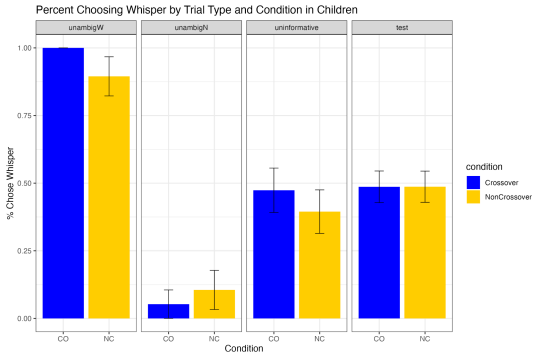


Figure 2: Results from Children. Children perform near ceiling in both unambiguous controls suggesting an ability to complete the task. In test trials, there is no effect by condition mirroring the irrelevant trials.

References

Crain, S., & Thornton, R. (1998). *Investigations in Universal Grammar: A guide to experiments on the acquisition of syntax and semantics.* The MIT Press.

Koster , J., & Koster , C. (1986). The acquisition of bound and free anaphora. Paper presented at the 11th annual Boston University conference on language development, Boston.

McDaniel, D., McKee, C. (1992). Which Children Did They Show Obey Strong Crossover?. In: Goodluck, H., Rochemont, M. (eds) Island Constraints. Studies in Theoretical Psycholinguistics, vol 15. Springer, Dordrecht.

Omaki, A., White, I. D., Goro, T., Lidz, J., & Phillips, C. (2014). No fear of commitment: Children's incremental interpretation in English and Japanese wh-questions. Language Learning and Development, 10, 206–233.

Thornton, R., & Wexler, K. (1999). *Principle B, VP ellipsis, and interpretation in child grammar*. (Current studies in linguistics; Vol. 31). The MIT Press.

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73*(2), 89–134.

Pronoun interpretation reveals the robustness and flexibility of perspective reasoning

The understanding of perspective is recognized as an essential component of semantic and pragmatic processing, influencing a wide range of processes including the interpretation of nominal expressions [e.g., 1,2], adjectives [3,4,5], and appositives and epithets [6,7], among others. In studies of language processing, it has often been claimed that the computation of perspective is challenging, entailing that perspective cues might not be used effectively during natural listening or reading [2,8]. In the present study, we explore both the robustness and complexity of perspective taking from a new angle, focusing on dramatic changes in interpretation that result from comprehenders' grasp of the felicity conditions of speech acts and epistemic authority [9,10]. This work builds on the assumption that pronoun interpretation is a by-product of understanding the overall discourse [11,12], and in turn can provide important insights into perspective-taking processes.

Consider the assertion "Jane told Annie that she likes spaghetti". Intuitively, the grammatically-ambiguous pronoun refers to Jane, as a report of "liking spaghetti" is best understood as reflecting Jane's intention to tell Annie something that Annie did not already know (cf. the pragmatic convention to "be informative" [13,14]). This reasoning explains why intuitions reverse with a *question* (cf. [15], "Jane asked Annie if she likes spaghetti", where "she" is now preferentially interpreted as referring to Annie). The latter case contrasts with the notion that pronoun interpretation is heavily guided by a bias toward subjects/first-mentioned entities [16,17]. We tested materials of this kind in antecedent judgment and self-paced reading tasks to validate and further understand how perspective reasoning influences pronoun resolution.

Experiments 1a-b (each: n_{subs}=54, n_{trials}=24) were offline antecedent judgment tasks. **Experiment 1a** assessed judgments of ambiguous *subject* pronouns in sentences like "Madeline [asked/told] Anna [if/that] she remembers when the lecture starts." Intuitively, a character *asking* an interlocutor about the information expressed in the subordinate clause should lead readers to interpret the pronoun as coreferring with the main-clause object, whereas *telling* should entail main-clause subject selections. The results overwhelmingly supported this intuition: Participants picked the antecedent we expected to be "perspectivally-congruent" 99.8% of the time, and there was no order-of-mention bias. In **Experiment 1b**, we tested *object* pronouns like "Nina [asked/told] Mary [if/that] modern art interests her more than classics." The results followed the same pattern, with the "congruent" antecedent selected 99.4% of the time.

Experiments 2a-b (each: n_{subs}=60, n_{trials}=24) used self-paced reading to clarify the scope of information used in the judgments. The critical sentences used in **Experiment 2a** were identical to Experiment 1a, but now contained unambiguous pronouns, where gender marking compelled coreference with either the "perspectivally-congruent" (1a-b) or "incongruent" (1c-d) antecedent:

(1a) Madeline asked Oscar if he remembers when the lecture starts.
(1b) Madeline told Oscar that she remembers when the lecture starts.
(1c) Madeline asked Oscar if she remembers when the lecture starts.
(1d) Madeline told Oscar that he remembers when the lecture starts.

Cases (1c-d) should entail processing costs relative to (1a-b) because of the forced link with the perspectivally-incongruent character. The critical question was whether the interpretive patterns arise from (i) shallow lexical cues (e.g., the verbs *ask/tell* signal which character possesses at-issue knowledge, making the effects emerge at the pronoun) or (ii) deeper/more rational forms of linguistic reasoning drawing on global sentence information. On the latter account, referential decisions would reflect a consideration of the complete or nearly-complete subordinate clause (i.e., downstream of the pronoun). Reading time was measured at the pronoun, subordinate verb, and sentence-final regions. Mean reading times are shown in Fig. 1. Consistent with a deep reasoning account, the effect of congruency (slower reading times in the incongruent condition) was not apparent until the sentence-final region, confirmed with linear mixed-effects

modelling (β=7.70, *SE*=1.78, *t*=4.32, *p*<.001). **Experiment 2b** used the object pronoun sentences from Experiment 1b, where, e.g., the perspectivally-incongruent sentences were:

(2a) Nina asked Isaac if modern art interests her more than classics.
(2b) Nina told Isaac that modern art interests him more than classics.

The results corroborated Experiment 2a (Fig. 2), where the location of the incongruency effect suggests readers use global sentence information (β=4.60, *SE*=1.79, *t*=2.57, *p*<.05).

To further assess the richness and flexibility of perspective reasoning, **Experiment 3** (n_{subs}=60, n_{trials}=20) assessed the potential for a preceding context sentence to "switch" the default patterns in the *ask* vs. *tell* sentences seen in Expt. 1, with materials like the following:

(3a) Molly, who is unfamiliar with Japanese currency, was talking to her tour guide, Hana. Molly asked Hana if she had enough cash to buy a sandwich.
(3b) Molly, a tour guide, was talking to Hana, who is unfamiliar with Japanese currency. Molly told Hana that she had enough cash to buy a sandwich.

Readers' judgements reflected a preference for subject antecedents 68% of the time for *ask* and 23% for *tell*, overriding **Experiment 1a-b's** near-categorical object selections for *ask* and subject selections for *tell*. Readers significantly changed their antecedent selection preference when presented with context sentences (relative to neutral baseline sentences, where the context sentence was not presented: β=-2.48, *SE*=0.25, *z*=-10.1, *p*<.001, via generalized linear mixed-effects modelling). Thus, the context sentences readily shift the understood subject of the embedded clause despite the "cues" stemming from the main verb. This outcome provides even more compelling evidence that the interpretive patterns reflect full-blown perspective reasoning.

In summary, Experiments 1a-b show extremely robust effects of perspective on pronoun resolution. Experiments 2a-b confirm that interpretation is not driven by lexical cues but instead involves a consideration of global sentence content, which we argue is a rational processing strategy considering the different ways that subsequent sentence information can influence interpretation. Experiment 3 further demonstrates that shallow lexical cues are insufficient as an explanation and highlights the flexibility of linguistic perspective taking. Together, the findings underscore the robustness of perspective reasoning in language understanding.

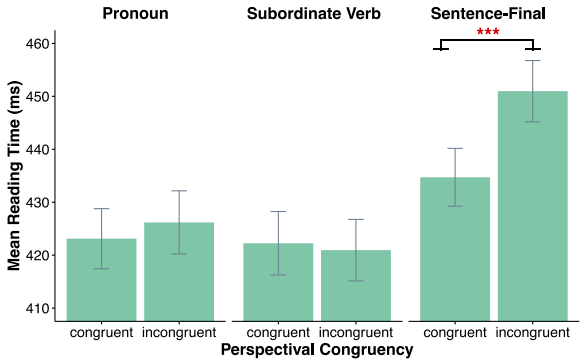


***Figure 1**: Mean RTs per condition by region, subject pronouns.*

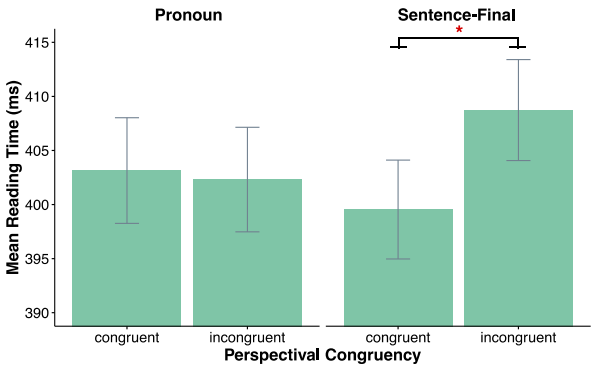


***Figure 2**: Mean RTs per condition by region, object pronouns.*

References: **[1]** Clark & Marshall (1981). In *Elements of discourse understanding*. **[2]** Keysar et al. (2000). *Psych. Sci*. **[3]** Lasersohn (2005). *Ling. & Phil.* **[4]** Nadig & Sedivy (2002). *Psych. Sci.* **[5]** Heller et al. (2008). *Cognition*. **[6]** Harris & Potts (2009). *Ling. & Phil.* **[7]** Kaiser (2015). *Sem. & Ling. Theory*. **[8]** Weingartner & Klin (2005). *Mem. & Cognition*. **[9]** Searle (1969). *Speech Acts*. **[10]** Westra & Nagel (2021). *Cognition*. **[11]** Hobbs (1979). *Cog. Sci.* **[12]** Kehler (2002). *Coherence, reference, and the theory of grammar.* **[13]** Grice (1975). In *Syntax & Semantics Vol. 3*. **[14]** Smyth (1995). *J. Child Lang.* **[15]** Brown-Schmidt et al. (2008). *Cognition*. **[16]** Gordon et al. (1993). *Cog. Sci.* **[17]** Arnold et al. (2000). *Cognition.*

The Structure of Ad-Hoc Alternatives

Understanding what a speaker means requires not only understanding what they said, but also considering what they could have said instead [1]. Of course, there are many things a speaker *could* have said, too many to consider them all. In every ad hoc context, listeners must consider only those alternatives they take to be relevant for informing the meaning of what the speaker actually said [2, 3]. In this study, we investigate which alternatives those are.

Imagine you are at your friend's house and you say, “I’m thirsty.” Your friend opens their fridge and says, “I only have milk.” As the listener, you might infer that your friend does not have water (negating this alternative), while remaining agnostic about whether they have, say, meat. How did you make these judgments? Did you generate a set of drinkable alternatives that did not include meat, such as {*water*, *juice*, *milk*}, and negate everything but milk? Or did you have some implicit ranking of alternatives, such as *water* > *juice* > *milk* > … > *meat*, and negate only those that were higher ranked than milk? Existing proposals entangle these data structures – sets and orderings (Horn scales [2]). In this study, we distinguish and probe each of them and their boolean combinations.

Design: We investigate whether a listener's beliefs about different alternatives change in response to an utterance, given some context. In our task, we call “trigger” items the words that evoke alternatives by being in the scope of a focus particle (“*only milk”*, above). We call “query” items the potential alternatives (*water, meat, etc.*). We test whether people change their beliefs about different queries, given contexts with each of the same words as triggers. We investigated four different alternative structures: Set, Ordering, Set-Ordering Conjunction, and Set-Ordering Disjunction. These structures make some overlapping predictions about which alternative query item gets negated [regions I.A. and O.B. in Figure 1], but diverging predictions in other cases [regions I.B. and O.A. in Figure 1]. Predictions diverge when a query item is inside the set but ranked below the trigger (does saying “I only have water” indicate not having milk?), and when a query item is outside the set of alternatives a listener would normally consider but ranked above the trigger (does saying “I only have meat” indicate not having soup?).

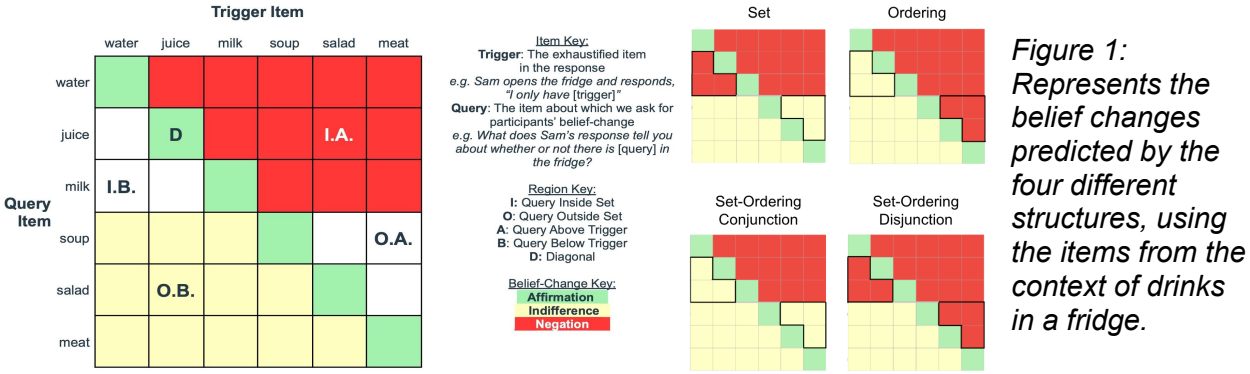


Figure 1: Represents the belief changes predicted by the four different structures, using the items from the context of drinks in a fridge.

Materials: A series of stimulus creation experiments were first done to generate 6 alternatives for each of 16 contexts. The alternatives ‘inside’ the set were generated by replacing the trigger phrase [the ‘Response’ in Figure 2] with “Sure, I have __” and asking participants to fill in the blank with 6 items that would satisfy the request. The alternatives ‘outside’ the set were generated by replacing the trigger phrase with “Looks like I don't have anything for that” and asking participants for 6 items that the friend could still have, which would *not* satisfy the request. Items generated in each context were aggregated across participants, and experimenters selected 6 of the most common distinct items for each context. The Orderings were then generated by asking a different group of participants to rank-order the 3 ‘inside’ the set items or the 3 ‘outside’ the set items, in each context, for their fit in the phrase “I have __.” Finally, another group of participants were asked to rank-order all 6 words for each

context, confirming that the items meant to be ‘inside’ were all ranked higher than those meant to be ‘outside’.

Methods and Results: We pre-registered our methods and analysis plan. 213 participants were tested through Amazon Mechanical Turk. Each participant read each context (16 trials) [Figure 2]. The combination of triggers and queries were pseudo-randomly assigned to each context to ensure the same number of responses per region per participant, and across query-trigger pairs within a region [Figure 1 shows the different regions], producing a total of 3,408 responses [Figure 3].

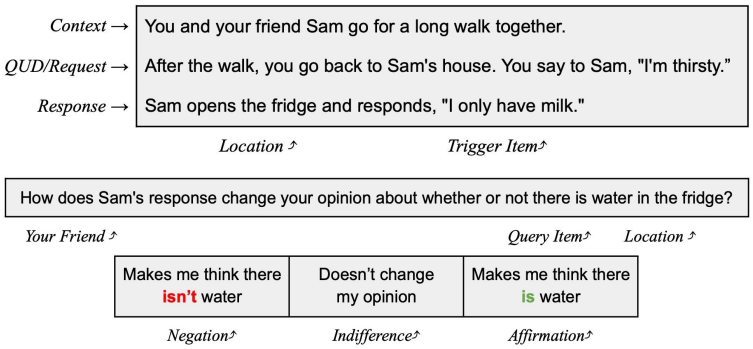


Figure 2: All trials had the same template: context, a request that defines the question under discussion (QUD), and a response to the QUD. Following each story, participants answered a multiple choice question designed to measure the direction of change in belief caused by the trigger item.

Our dependent variable is whether participants responded with a `Negation`. In a mixed-effects logistic model, we found significant effects of whether the query was inside vs. outside the `Set` ($\chi^2 = 61.19$, $p < 0.001$) and whether the query was higher or lower than the trigger in the `Ordering` ($\chi^2 = 53.15$, $p < 0.001$) but no interaction. We subsequently tested the pairwise differences between theoretically critical regions [Figure 1]. Items in regions `O.A.` and `I.B.` were each significantly more negated than in region `O.B.` (`O.A. > O.B.`; $\beta = 0.19$, $p < 0.001$), and `I.B. > O.B.` ($\beta = 0.40$, $p < 0.001$). This combination uniquely diagnoses a Set-Ordering Disjunction structure [Figure 3].

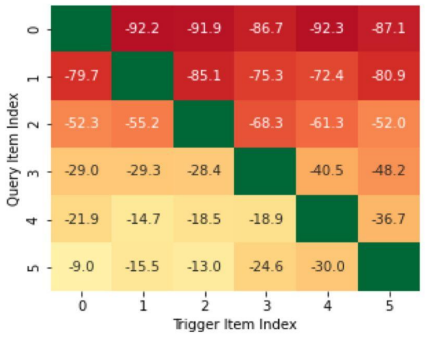


Figure 3: Aggregated Results. Represents the average negation from participants across all contexts. Larger magnitudes indicate more negation. Indexes 0-2 are inside the Set, 3-5 are outside. 0 is highest in the Ordering, 5 is lowest.

Conclusion: In aggregate, we find evidence for Set-Ordering Disjunction. This suggests that, at least in some contexts, people negate all plausible alternatives, even if there would not have been better responses than the focused trigger word (e.g. *only having water* implies not having milk, even though milk would have been a worse option). At the same time, they also (at least sometimes) negate higher ranked alternatives even when these are outside the set of what would normally be considered relevant responses (e.g. *only having meat* negates having even normally-irrelevant soup). Aggregate results might reflect combining more set-like structures and more ordering-like structures. This raises a further possibility: perhaps the very structure (not just the content) of how alternatives are relevant varies across contexts.

References: [1] Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics 3: Speech Acts*. [2] Horn, L. R. (1972). On the semantic properties of logical operators in English. *University of California, Los Angeles*. [3] Gotzner, N., & Romoli, J. (2022). Meaning and alternatives. *Annual Review of Linguistics, 8*.

Conceptual and language-specific effects on multimodal recipient event descriptions

When describing events, speakers often do not include all event participants involved.[1] One reason for such omissions is the conceptual prominence of each participant role. Prior research shows that across languages, conceptually peripheral roles (e.g., RECIPIENTS, INSTRUMENTS) are mentioned less than conceptually prominent ones (e.g., AGENTS, PATIENTS).[2] However, not all conceptually peripheral roles are born equal. For instance, certain verbs conceptually "require" a recipient (e.g. *The person sends a message to their friend*), while others "allow" a recipient (e.g. *The woman bounced the ball (to her friend)*).[3] Although this theoretical distinction is confirmed by English speakers' judgments,[4] it is unclear how it affects speakers' syntactic choices in free event descriptions across languages. Further, speech is not the only modality used to describe events, and it is possible that omission of a participant role in speech is compensated by its inclusion in gesture.[5] Here, we investigate how underlying conceptual requirements (i.e. the require-allow distinction) influence the content of multimodal possession-transfer event descriptions across languages. We use two typologically distinct languages (English, Turkish) that differ in the grammaticality of event participant omissions (Turkish allows argument drop, while English does not) and the use of gesture (Turkish culture is high-gesture).[6]

Sixty participants (30 L1 Turkish, 30 L1 English) described short videos of everyday events (n=36) to a naïve interlocutor with maximal informational needs (friend of the speaker who could not see the events). Test events involved 12 possession-transfer events (6 require-recipient, 6 allow-recipient; Fig.1). We coded for recipient mentions in speech and gesture within the same clause as the main verb that described the event (e.g., *The woman bounced the ball to her friend*). We hypothesized that speakers should mention recipients more frequently when conceptually required than allowed, across both languages and modalities. Given that language-specific event encodings in speech also persist in gesture,[5-7] we anticipated that recipients would be dropped more frequently in Turkish than in English in both modalities.

Beginning with recipient mentions in speech, a mixed-effects logistic regression showed no effect of Verb Type (p = .807, n.s.). Contrary to our predictions, speakers of both languages mentioned required and allowed recipients equally frequently ($M_{Require}$=0.84, M_{Allow}=0.77). Crucially, the model yielded a significant effect of Language (β=-0.696, SE=0.242, z=-2.874, p=.0041) in the expected direction: English speakers mentioned recipients more frequently than Turkish speakers (M_{ENG}=0.84, M_{TUR}=0.77). Next, we analyzed recipient mentions in gesture. Observation of the data indicated that these were all gestures that co-occurred with mentions in speech (Fig.2). Similar to the analysis of speech, there was no effect of Verb Type (p = .599, n.s.), but a significant effect of Language (β=1.893, SE=0.548, z=3.452, p<.001). Interestingly, however, this effect was in the opposite direction than in speech: recipient gestures were used more frequently in Turkish than in English (M_{ENG}=0.20, M_{TUR}=0.33). Finally, we analyzed recipient mentions in both modalities. This analysis revealed an effect of Language, with recipients being mentioned more in English than in Turkish (β = -0.674, SE = 0.258, z = -2.612, p = .009, M_{ENG} = 0.71, M_{TUR} = 0.70).

In sum, our findings show that language-specific encoding patterns heavily affect mention of recipients in free event descriptions across modalities. When both speech and gesture were considered, speakers of Turkish used recipients less frequently than speakers of English. Similar to prior research,[6] we found that recipient gestures were used more frequently in Turkish than in English. However, these were co-speech gestures that did not add additional information beyond what was encoded in speech. Taken together, these findings suggest that argument drop in Turkish persists across modalities. Contrary to our predictions, the require-allow distinction did not affect speakers' mentions of recipients in any modality. We conclude that linguistic planning for recipient event roles is more heavily affected by language-specific encoding options than the gradient conceptual prominence of the roles.

References

1. Papafragou, A., & Grigoroglou, M. (2019). The role of conceptualization during language production: Evidence from event encoding. *Language, Cognition and Neuroscience, 34* (9), 1117–28.
2. Ünal, E., Richards, C., Trueswell, J., Papafragou, A. (2021). Representing agents, patients, goals, and instruments in causative events: A cross-linguistic investigation of early language and cognition. *Developmental Science*, *24*(6), 1-13.
3. Levin, B. (1993). English verb classes and alternations: A preliminary investigation. Chicago: University of Chicago Press.
4. Rissman, L., Rawlins, K., & Landau, B. (2015). Using instruments to understand argument structure: Evidence for gradient representation. *Cognition, 142,* 266-290.
5. Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and language, 48*, 16-32.
6. Azar, Z., Özyürek, A., & Backus, A. (2020). Turkish-Dutch bilinguals maintain language-specific reference tracking strategies in elicited narratives. *International Journal of Bilingualism*, *24*(2), 376–409.
7. Ünal, E., Manhard, F., & Özyürek, A. (2022). Speaking and gesturing guide event perception during message conceptualization: Evidence from eye movements. *Cognition, 225, 105127.*

Figures

(*Kadın) arkadaşına bır sır anlattı*

a. 'The woman told a secret to her friend'

(*Kadın*) *topu arkadaşına sektirdi*

b. 'The woman bounced the ball to her friend'

Figure 1. Example stimuli and potential descriptions in Turkish and English for possession transfer events where the recipient is (a) conceptually required or (b) conceptually allowed.

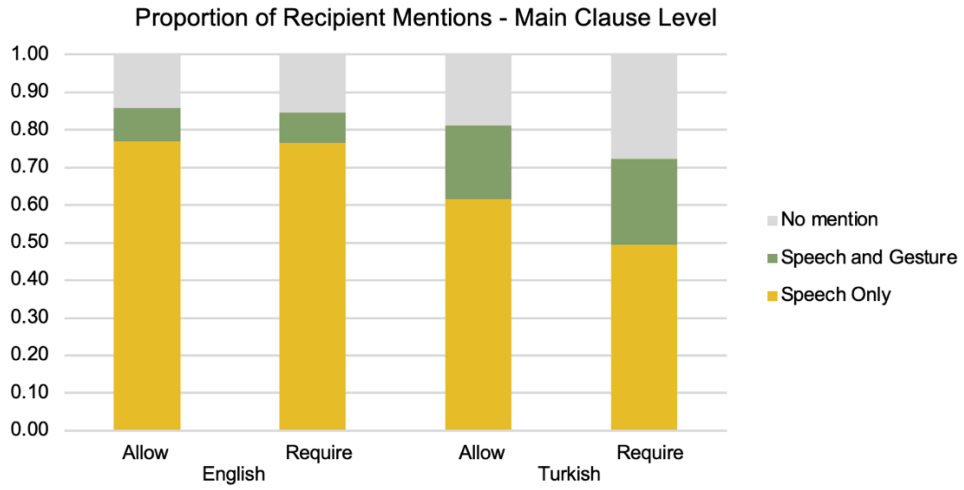


Figure 2. Mean proportion of Recipient mentions in speech and gesture within the same clause as the main verb, across verb types and language groups.

Does ‘a couple’ pattern with scalars or numbers - Insights from the inference and ‘so’ tasks

Background. Paucal quantifiers typically denote a range of small numbers. For example, ‘a couple’ can mean more or less the same thing as ‘two’, but it is also often used for a broader range of cardinal values than just two, depending on certain properties the objects in question are perceived to have. Thus, on the one hand, ‘a couple’ has properties akin to other indefinite expressions like ‘some’ and ‘a few’. On the other hand, it has a semantic core that is somewhat more determinate, like the numeral ‘two’. A growing body of experimental research, such as Sun & Breheny’s (2022) inference tasks, points to differences in outcomes for tasks when ‘some’ and numerals like ‘two’ are compared. The first experiment of this research adopts their inference tasks, and shows that ‘a couple’ in some ways behaves like numbers and in some ways like other scalars. To continue this theme of finding whether ‘a couple’ patterns with ‘some’ or numerals, the other experiment replicates Sun et al.’s (2018) ‘so’ task which explores the correlation between the naturalness of an ‘X so not Y’ construction and the rates of scalar inferences (SIs) measured in the inference task.

Experiments. Sun & Breheny (2022) have tested numbers and scalar expressions in inference tasks and established that scalars (e.g. some) are sensitive to a manipulation that can change the contextual relevance of alternatives (all), whilst ‘exactly’ readings of numbers are not. Our Exp. 1 mirrors their study so as to see if manipulating contexts has an effect on interpreting ‘a couple’.

Exp.1 (n = 60) was a partial replication of Sun & Breheny’s study investigating ‘a couple’, ‘possible’ and ‘some’ in inference tasks with two types of probe questions (see Fig.1, left). One type, referred to as ‘not Alt’ probe, was intrinsically a standard inference task where the probe question asked participants whether they could infer the negation of a scalar alternative (e.g. not many), according to a speaker character’s statement containing a scalar expression (e.g. a couple), and Target Response corresponding to inferring the SI was a ‘Yes’ response. The other type of probe question, called ‘could Alt’ probe, asked participants whether ‘many’ might not be excluded for the same statement, and Target Response was a ‘No’ response. Note that participants could also give a ‘No’ response when they were uncertain about the speaker’s intended meaning, irrespective of the probe type. In light of Sun & Breheny, the interpretations of ‘some’ and ‘possible’ were affected by the manipulation of probes, because there were more Target Responses for ‘not Alt’ than ‘could Alt’ probes. This suggested that probe questions had an effect on making the SI contextually relevant, so participants were more certain about inferring the SI as part of the intended meaning, which led to more Target Responses for the ‘not Alt’ probe. Responses to numbers in their study showed a reversed pattern, indicating that the different probes had no effect on which reading would become available for participants, so they gave a ‘No’ response due to uncertainty, which led to more Target Responses for the ‘could Alt’ probe and fewer Target Responses for the ‘not Alt’ probe. As illustrated in Fig. 1 (right), our results replicated the pattern in Sun & Breheny between ‘not Alt’ and ‘could Alt’ probes for ‘possible’ ($p = .07$). Crucially, the probability of Target Responses was greater for the ‘not Alt’ probe compared to the ‘could Alt’ probe for ‘a couple’ ($p < .01$), suggesting that paucal quantifiers, such as ‘a couple’, behave like a genuine scalar expression, not like numbers. However, we note that, for ‘a couple’, the rate of Target Responses to ‘not Alt’ probe was significantly lower than that for ‘possible’ ($p = .04$), which was similar to that found by Sun & Breheny when numerals were compared to scalars in ‘not Alt’ trials. This may be due to the fact that like numbers, inferences for ‘a couple’ are more independent from the context. Given that Sun et al. ran a ‘so’ task that is a follow up of the inference task in their Exp.1, we also conduct Exp.2 that mirrors Sun et al.’s ‘so’ task to continue this aim of seeing if ‘a couple’ behaves more like ‘some’ or numbers.

Exp.2 (n = 103) adapted Sun et al.’s ‘so’ task. Fig. 2 (left) is an example item. We used 48 scalars including 43 of them investigated in Sun et al.’s study along with ‘a couple/high number’, ‘a couple/many’ and some other scalars to construct experimental sentences for Exp.2. The

experimental sentences were of the form ‘X so not Y,’ where X is informationally stronger than Y; for example, ‘The student sharpened many of the pencils, so not a couple of the pencils.’ As can be seen in Fig. 2 (left), we employed a between-subject design. Two groups, the partitive group and the non-partitive group, were created. All the other scalars in the two groups were the same, except for ‘a couple/high number’ in the non-partitive group, whilst ‘a couple/many’ in the partitive group. Each participant was randomly assigned to one of the two groups and judged 47 experimental sentences. Participants were asked to indicate how natural these constructions are on a 1 (very unnatural) - 7 (very natural) Likert scale. As illustrated in Fig. 2 (right), we found a significant difference between ‘a couple’ and ‘some’ in both groups (‘a couple/high number’: $p < .001$; ‘a couple/many’: $p = .003$). However, we did not find a significant difference, when comparing number to ‘a couple/high number’ or to ‘a couple/many’. Similarly, there was no significant difference between ‘a couple/high number’ and ‘a couple/many’.

Discussion. Although, in Exp.1, paucal quantifiers such as ‘a couple’ behave like genuine scalar items such as ‘some’, Exp.2 shows that ‘a couple’ can be similar to ‘number’, when compared to scalars like ‘some’. Overall, our findings indicate two natures of ‘a couple’, and then the follow up ‘so’ task would further show the numbers’ nature of ‘a couple’, particularly in the context of numbers, compared to ‘many’.

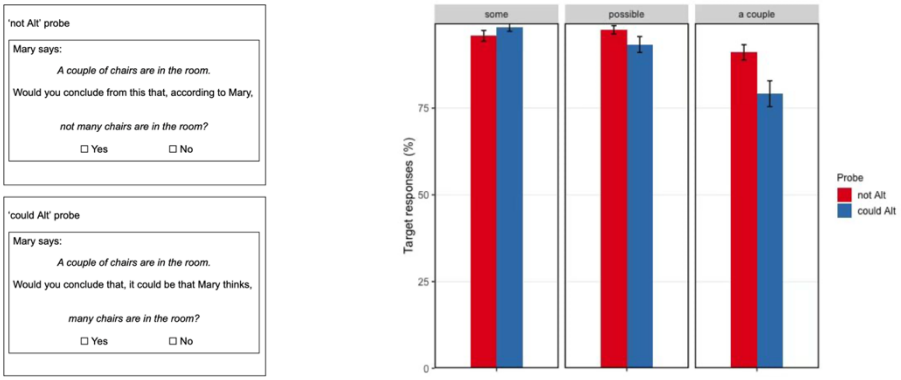


Fig. 1. Example trials (left) and results (right) for Exp.1.

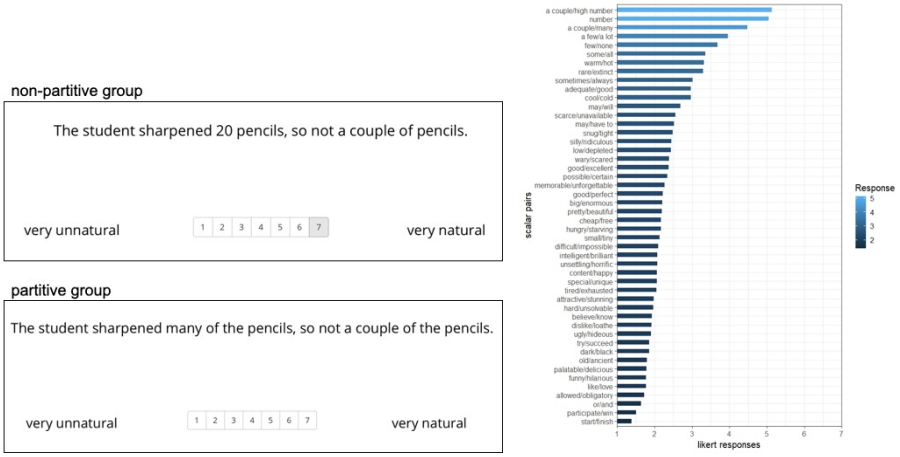


Fig. 2. Example trials (left) and results (right) for Exp.2.

Selected references: Sun, Chao, Ye Tian & Richard Breheny, 2018, *A Link Between Local Enrichment and Scalar Diversity* • Sun, Chao & Richard Breheny, 2022, *The role of Alternatives in the interpretation of scalars and numbers: Insights from the inference task*

Online Processing of, and Adaptation to, Nonbinary Pronouns

Recent years have seen a surge in usage of English nonbinary pronouns associated with increased salience of trans identities (Minkin 2021). These include definite specific singular *they* with referents of known gender, as well as neopronouns such as *xe, ze, fae,* and *thon*. Acceptability judgment studies have shown their grammaticality to be in transition (Rose et al 2023). For *they*, English speakers fall under one of three categories based on their acceptance of *they* (Camilliere et al. 2021): non-innovators, who only license indefinite antecedents (1); innovators, who also allow non-gendered specific antecedents (1-2); and super-innovators, who accept any animate antecedent (1-3).

(1) Someone$_i$ slept because they$_i$ were tired
(2) The student$_i$ slept because they$_i$ were tired
(3) Sarah$_i$ slept because they$_i$ were tired.

The present study used a web-based Maze task (Boyce et al. 2022) to investigate processing costs for *they*, *ze*, and *s/he* with definite singular referents, as well as whether difficulty changes throughout an experiment as participants are exposed to these usages. One possibility is that the novel *ze* will be more difficult than the more common *they* throughout the experiment. Alternatively, *ze* may be more difficult initially than *they* but may actually exhibit more rapid adaptation over the course of the study. Note that *they* is referentially and pragmatically more ambiguous than *ze*. *They* can be used to refer to many different types of antecedents (e.g., plurals, indefinites, generics, institutions). Nonbinary individuals are likely the least common antecedent for *they*. *Ze* is solely and explicitly a nonbinary pronoun. This may facilitate adaptation.

Experiment. 112 participants were trained on the use of either *they* or *ze*, then asked to read sentences about named individuals "who would be referred to with their pronouns." The names were highly associated with one binary gender or equibiased between binary genders, (established via a web-based survey on a separate group of participants). Sentences contained a critical pronoun (binary/nonbinary within participants, *they/ze* between participants) that matched its antecedent's gender features to varying degrees (intermediate/weak). 100 stimuli were developed and divided among four presentation lists using a Latin square design and pseudorandomly interspersed with 25 strongly matched controls.

Table 1. Example stimuli. Instructions: "This is a story about [name], who uses [pronouns] pronouns."

	Strong match	**Intermediate match**	**Weak match**
Binary pronoun	Amanda was studying for the bar because she wanted to be a lawyer.	Alex bought a new phone because he broke the old one.	Alice bought a new phone because he broke the old one.
Non-binary pronoun	-	Alex bought a new phone because they/ze broke the old one.	Alice bought a new phone because they/ze broke the old one.

At each point of a sentence, participants were presented with two words: the grammatically correct word, and a length- and frequency-matched foil word that was incompatible with the unfolding sentence. Participants had to select the correct word. RTs and error rates at the pronoun were recorded to assess processing difficulty. Participants also completed an acceptability survey of *they* with various antecedents in order to be classified as noninnovators, innovators, or superinnovators.

Results. Accuracy was at ceiling in all conditions (>98% in each condition) demonstrating that participants recognized all pronoun types as more grammatical than the foils. RTs were analyzed with maximal mixed effect models. We found a main effect of nonbinary pronoun type where *ze* elicited significantly greater difficulty than *they* ($\beta = -37.2$, $t = -2.76$, $p < 0.01$), likely due to its status as a neologism in a closed class (pronouns). There was also a main effect of presentation order ($\beta = -41.9$, $t = -8.40$, $p < .001$), and an interaction where reaction times

decreased over the course of the experiment at a greater rate for *ze* than *they* (β = -36.7, t = -2.69, p < 0.01). Non-innovators experienced greater difficulty with nonbinary pronouns than innovators and superinnovators (β = -95.7, t = -2.03, p < 0.05). For *ze*, non-innovators also showed more adaptation than innovators and superinnovators (β = 2.0, t = 3.75, p < 0.001), and innovators more than superinnovators (β = 1.4, t = 2.87, p < 0.01). No effect of match was found for nonbinary pronouns. Thus gender equibiased names did not significantly ameliorate difficulty with nonbinary pronouns.

Discussion. *Ze* was more difficult than *they*, but participants also adapted more quickly to *ze* than *they*. This supports the hypothesis that *ze* is easier to learn because it is less ambiguous than *they*. Another possibility is that learning is error based: The larger the error, the larger the adaptation. However such a mechanism should have led to fast adaptation in the binary weak match conditions, which was not observed. Superinnovators experienced less difficulty with nonbinary pronouns, but also less adaptation than the other clusters. They were previously shown to be younger, more familiar with, and more accepting of trans identities (Camilliere et al. 2021). Their processing fluency may have reached a ceiling early in the study due to prior exposure to, and acceptance of, nonbinary pronouns.

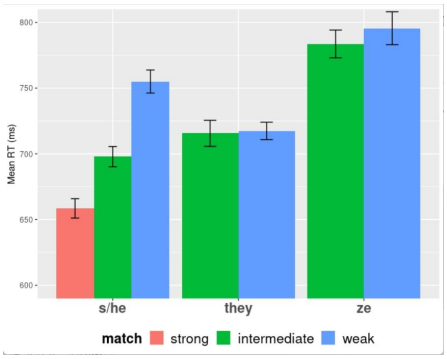


Figure 1: Mean RT by pronoun type and match.

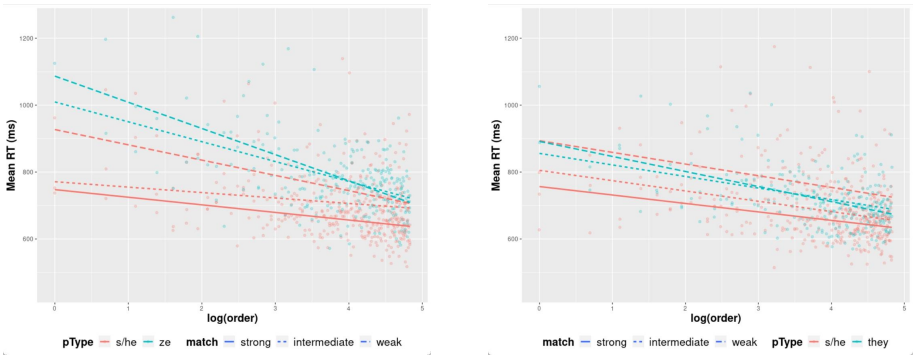


Figure 2: Mean RT by order for *ze* vs *s/he* (left) and *they* vs *s/he* (right).

References.

Boyce, V., Futrell, R., & Levy, R. P. (2019). Maze Made Easy: Better and Easier Measures of Incremental Processing Difficulty. https://doi.org/10.31234/osf.io/b7nqd

Camilliere, S., Izes, A., Leventhal, O., & Grodner, D. J. (2021). They is Changing: Pragmatic and Grammatical Factors that License Singular they. Proceedings of the Annual Meeting of the Cognitive Science Society.

Minkin, R. (2021, July 27). Rising shares of U.S. adults know someone who is transgender or goes by gender-neutral pronouns. Pew Research Center. https://www.pewresearch.org/short-reads/2021/07/27/rising-shares-of-u-s-adults-know-someone-who-is-transgender-or-goes-by-gender-neutral-pronouns/.

Rose, E., Winig, M., Nash, J., Roepke, K., & Conrod, K. (2023). Variation in acceptability of neologistic English pronouns. Proceedings of the Linguistic Society of America, 8(1), 5526. https://doi.org/10.3765/plsa.v8i1.5526

Learning discourse patterns through exposure: Mixed input helps identify informative categories

While much of language is learned during childhood, adults continue to adapt to the most frequent patterns in local contexts. Speakers tend to imitate the structure of syntactic primes (e.g., Bock, 1986) and comprehenders are biased toward recently-heard syntactic structures (e.g., Thothathiri & Snedeker, 2008). Adaptation also occurs at the discourse level, where the interpretation of ambiguous pronouns is biased toward recently encountered patterns (e.g., Kaiser, 2009; Contemori, 2019). Johnson & Arnold (2023, exp. 2) tested the interpretation of ambiguous pronouns like "{Ana sent a text to Liz / Ana got a text from Liz} and then she took a screenshot." Here people favor the subject (Ana) as the referent of "she", following the well-known subject-bias (e.g., Stevenson et al., 1994), but they also favor the goal (Liz for "send", Ana for "got"), so the preference for Ana is stronger for "got" than "send" (Langlois & Arnold, 2020). Johnson & Arnold showed that these biases are malleable. If people have recently read numerous examples of unambiguous pronouns referring to the nonsubject (half goal, half source), they are somewhat more likely to pick nonsubject antecedents (Exp. 2a). But if people have recently read many pronouns referring to the source (half subject, half nonsubject), they shift their interpretation in favor of the source (Exp. 2b). This shows that people adapt to the property of antecedents that is most informative. In Exp. 2a, the subject/nonsubject distinction was informative, and goal/source was not; the reverse pattern held for Exp. 2b.

This raises questions about how people respond to linguistic input that could be informative about multiple patterns. Given "Matt got a book to Ana and he…", do people learn that pronouns refer to goals? Or to subjects? We hypothesize that over a lifetime of input, people may abstract across exemplars to learn biases related to both syntactically-driven categories (e.g., subject antecedents) and semantically-driven categories (e.g., goal antecedents). When either one varies in the local context, people shift their biases to adapt.

In two experiments we tested how people respond to input that is either uninformative about the relevant category to learn (Exp. 1, 116 subjects), or informative (Exp. 2, 80 subjects). Experiment 1 used Johnson and Arnold's methods and stimuli, but all the exposure stories used goal-source verbs ("sent" type; see Table 1). In the subject-exposure condition, all 32 exposure stories used pronouns referring to the subject/goal; in the nonsubject-exposure condition, all exposure pronouns referred to the nonsubject/source. Interspersed were 12 stories with ambiguous pronouns using either goal-source or source-goal verbs, and we probed interpretation with questions (Table 2). The key question was whether exposure to goal-source stories would influence pronoun interpretation for both verb types or not.

Results showed it did not (Figure 1). For stories with matching verbs, there was a strong exposure effect: more subject/goal interpretations for subject/goal-exposure than for nonsubject/source-exposure. There was no exposure effect for the mismatching verbs. We know that exposure effects are not specific to thematic role, because Ye & Arnold (2023) found that exposure generalizes across verbtype. Thus, participants may have learned both syntactically- and semantically-conditioned patterns that canceled out for the source-goal verbs, or failed to learn either, or a mix.

Experiment 2 tested whether mixed input can direct participants' attention to the syntactic dimension of pronoun antecedents. Using the same goal-source exposure stimuli as Exp. 3, we replaced 8 exposure trials with "joint action" verbs (see Table 1), where the pronoun refers to either a subject/agent or a nonsubject/comitative role. This thematic role variability may signal that the informative dimension is syntactic role, and not thematic role. If so, exposure should generalize to test trials with the source-goal verb.

Results showed that indeed exposure generalized (Figure 1). Both experiments contrasted exposure stories with 100% subject vs. 100% objPP antecedents. But in Exp. 1, with only one verbtype, people didn't learn anything special about syntactic position per. In Exp. 2 we

varied the thematic roles (subject antecedents were 75% goal/25% agent while nonsubject antecedents were 75% source/25% comitative), and exposure modulated the subject bias for both verbtypes.

This study shows that discourse patterns are inferred from the input by abstracting over multiple exemplars, and not just through immediate priming from the previous trial. It also shows that people can extract generalizations like “pronouns tend to refer to subjects” from exposure to complex inputs.

==========

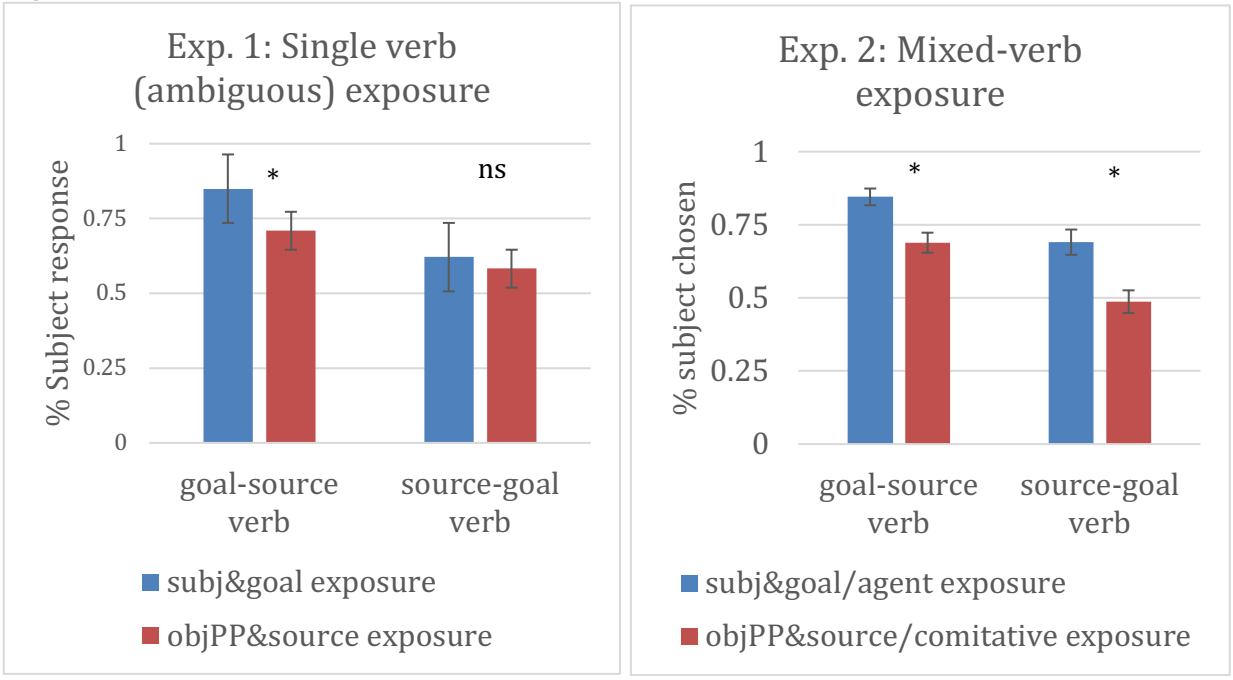
Table 1. Example exposure stimuli:

Goal-source verb; Subject pronoun: Ana and Matt were taking an English lit class. Ana borrowed the book from Matt and then she looked up a reference.
Goal-source verb; Nonsubject pronoun: … and then he looked up a reference.
(Exp. 2 only) Joint-action verb; Subject pronoun: Liz and Will were spending the weekend together. Liz set up a picnic in the park with Will and then she ate some sandwiches.
(Exp. 2 only) Joint-action verb; Nonubject pronoun: … and then he ate some sandwiches.

Table 2. Example critical (ambiguous) stimuli:

Goal-source verb: Will and Matt were taking an exam in class. Will borrowed a pencil from Matt and then he began his exam. Did Matt begin his exam? (no = subject interpretation)
Source-goal verb: Will and Matt were taking an exam in class. Will loaned a pencil to Matt and then he began his exam. Did Matt begin his exam? (no = subject interpretation).

Figure 1. Results from Exp. 1 and Exp. 2



References:

Contemori, C. (2019). Changing comprehenders’ pronoun interpretations… Second Language Research. ✠ Johnson, E., & Arnold, J. E. (2023). The Frequency of referential patterns…. JEP:LMC. ✠ Kaiser, E. (2009). Effects of anaphoric dependencies … Discourse anaphora…(pp. 121–129). Springer. ✠ Langlois, V. J., & Arnold, J. E. (2020). Print exposure .... Cognition. ✠ Stevenson, R. J., et al.. (1994). Thematic roles… LCP. ✠ Thothathiri, M., & Snedeker, J. (2008). Give and take… Cognition. ✠ Ye, Y. & Arnold, J. E. (2023). Learning the statistics … *Cognition.*

Investigating fragment usage with a gamified utterance selection task

Why do we use fragments? Fragments like (1a) (Morgan, 1973) can often be used to perform the same speech act as the corresponding sentence (1b).

(1) [Passenger to conductor before entering the train:]
   a. To Paris?
   b. Does this train go to Paris?

The syntax of fragments is relatively well researched, but the question of why and when speakers use fragments is not. Some syntactic accounts propose licensing conditions on fragments (e.g. Merchant, 2004; Barton and Progovac, 2005) based on information structure or recoverability, but fragments are not always used when they are licensed, as the acceptability of (1b) in this context shows. Intuitively, the advantage of fragments is that they allow the speaker to get a message across with less production effort. However, fragments can be enriched in different ways (see e.g. (2) for (1a)) and thus increase the risk of being misunderstood.

(2) a. How long does it take to travel to Paris?
   b. Have you ever been to Paris?

The choice between a fragment and a sentence probably consists in a trade-off between a gain in efficiency and the risk of communication failure. In what follows, I present a game-theoretic formalization of this reasoning and an pseudo-interactive experiment testing its predictions.

A game-theoretic account of fragment usage The model I propose is based on Franke's (2009) account of implicature: There is (i) a set of messages $m \in M$ that a speaker can to communicate and (ii) a set of utterances $u \in U$ which can used for this purpose. The speaker selects the utterance which is most optimal; the hearer receives it and figures out which message the speaker had in mind. The hearer computes $p(m|u)$ based on the prior likelihood of m and a denotation function $[[\cdot]]$, which returns 1 if u can be derived by grammatically licensed omission from m and 0 otherwise (see equation 1). The speaker in turn tries to maximize $L_0(u, m_i)$ for their intended m_i while keeping the production cost for u as low as possible.

$$L_0(u, m) = \frac{Pr(m) \times [[u]]_m}{\sum_{m'} Pr(m') \times [[u]]_{m'}} \tag{1}$$

Empirically founded model parameters In order to compute L_0 posterior probabilities with equation 1, I estimated M, $Pr(M)$, U and $[[u]]_m$ for all $m \in M$, $u \in U$ from a data set collected by Lemke (2021) with a production study. The data set contains about 100 utterances for each of 24 context stories (4) based on the DeScript corpus of script knowledge (Wanzare et al., 2016). The utterances were transformed into simplified representations like (3a) (pooling synonyms and to excluding ungrammatical omissions of function words, see Lemke (2021) for details), each of these representing a message like (3a). Its relative frequency is used as $Pr(m)$ in the model. Since all of the "words" in representations like (3a) can be freely omitted, this yields the set of utterances in (3b), for which $[[u]]_{(3a)} = 1$.

(3) "Pour the pasta into the pot"
   a. `pour pasta pot.GOAL¡`

Experiment design Since the game-theoretic account is inherently interactive, I test its predictions with an interactive utterance selection design (similar to Rohde et al. (2012) for referring expressions). The production cost for utterances is implemented by an explicit cost term. Currently, the participant plays the speaker role and the listener role is simulated by the computer, who – in a initial step – behaves maximally rationally, i.e. as predicted by the

model. In each trial ($n = 15$), the participant is presented a context story and an message to communicate with one out of six utterances (see fig. 1, showing the German implementation). Their task consists in selecting one of the utterances to communicate the message. In order to model utterance cost, subjects are assigned an account of virtual coins they can spend for sending utterances (starting with 500 coins): Sentences (cost: 100) are more expensive than fragments (cost: 30) and successful communication is rewarded with 120 coins. In the experiment, there are three conditions (i) the "target utterance" (most likely given the fragment) is highlighted, (ii) the competitor (less likely, but possible), (iii) the distractor is highlighted. According to model predictions, subjects should use fragments more often in the target than in the competitor condition, and most often in the unambiguous distractor condition,

Preliminary results and dicussion Data collection is ongoing, but the results of the first list of a pilot study indicate that – as expected – the rate of fragment choice is highest in the unambiguous distractor condition (46%). Furthermore, fragments are used more often (20%) when they refer to a predictable message than when they refer to an unpredictable one (14%). The analysis of the further data currently being collected will show whether this pattern is consistent. If it were, it would provide empirical support for a rational and game-theoretic account of fragment usage. Interestingly, the data collected so far also indicate a strong overall bias for using sentences, even in the unambiguous distractor condition, which will be also subject to further research.

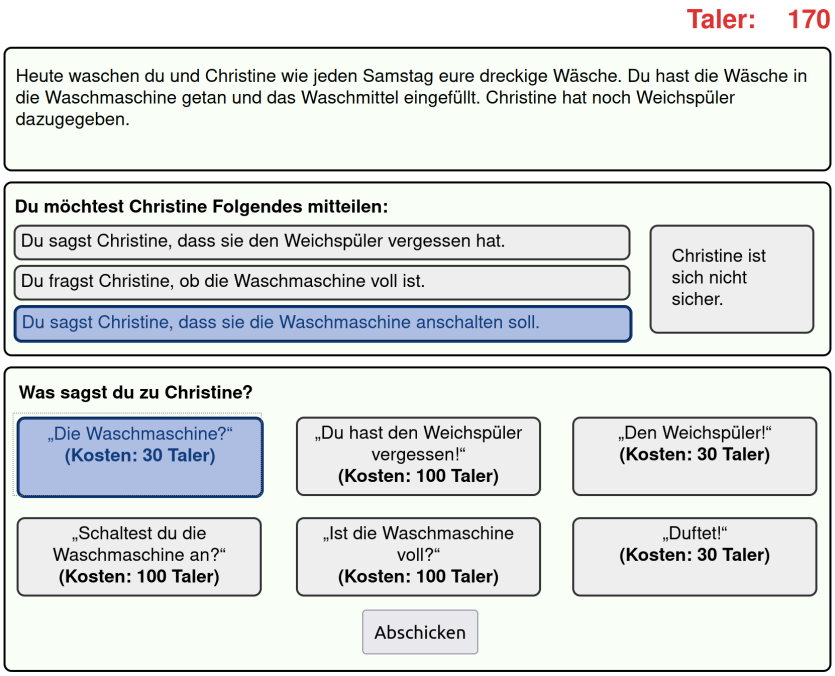


Figure 1 Sample utterance selection display (German) showing the context story, three messages and six utterances.

References •Barton, E. and Progovac, L. (2005). Nonsententials in Minimalism. In Elugardo, R. and Stainton, R. J., editors, Ellipsis and Nonsentential Speech, 71–93. Springer, Dordrecht. •Franke, M. (2009). Signal to Act: Game Theory in Pragmatics. PhD thesis, Universiteit van Amsterdam. •Lemke, R. (2021). Experimental Investigations on the Syntax and Usage of Fragments. Number 1 in Open Germanic Linguistics. Language Science Press, Berlin. • Merchant, J. (2004). Fragments and ellipsis. Linguistics and Philosophy, 27(6):661–738. •Morgan, J. (1973). Sentence fragments and the notion 'sentence'. In Kachru, B. B., Lees, R., Malkiel, Y., Pietrangeli, A., and Saporta, S., editors, Issues in Linguistics. Papers in Honor of Henry and Renée Kahane, 719–751. University of Illionois Press, Urbana. • Rohde, H., Seyfarth, S., Clark, B., Jaeger, G., and Kaufmann, S. (2012). Communicating with cost-based implicature: A game-theoretic approach to ambiguity. In Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue, 107–116. •Wanzare, L. D. A., Zarcone, A., Thater, S., and Pinkal, M. (2016). DeScript: A crowdsourced corpus for the acqui¸sition of high-quality script knowledge. In Proceedings of LREC 2016, 3494–3501, Portoroz, Slovenia.¸

Do speakers of nominative vs. ergative languages think about Agency in different ways?

Introduction Event roles such as Agent and Patient have been argued to be cross-linguistically universal and crucial for language evolution [1-3]. One challenge to this universal view is that Agent-marking syntactic structures in different languages express different semantic categories [4]. For example, intransitive (one-participant) verbs (e.g., *jump*, *arrive, die*) range on a semantic continuum from more activity-oriented (e.g., *jump*) to more state-oriented (e.g., *die*). In English, the arguments of activity-oriented verbs and state-oriented verbs are expressed in the same way (all are marked by nominative case). In Basque, by contrast, more activity-oriented verbs mark their arguments with what is known as ergative case, while the arguments of more state-oriented verbs are nominative-marked. Hindi is an ergative/absolute language like Basque, but in Hindi arguments of intransitive verbs do not receive ergative case. We investigate whether these different syntactic systems correspond to English, Basque, and Hindi speakers conceptualizing Agency in different ways. Specifically, we test two ways in which Agent roles might differ. First, English, Basque, and Hindi speakers might represent Agent in terms of different prototypes. In linguistic theory, event roles are often analyzed in terms of proto-Properties: for example, being intentional and playing a causative role are properties of proto-Agents whereas being affected is a property of proto-Patients [5]. The proto-Properties that constitute Agency may differ for English, Basque, and Hindi speakers. Second, these speakers might diverge in how they conceptualize the single participant in an intransitive event (e.g., one who jumps, one who arrives) with respect to the Agent category. Consistent with how arguments of intransitive verbs are marked in these languages, English speakers might represent an individual who *arrives* as more Agentive than Basque or Hindi speakers do. We tested these hypotheses using an event categorization task in which participants learned to sort pictures of transitive (two-participant) events into Agent and Patient piles, building on Rissman and Lupyan [6]. At test, we asked participants to generalize these categories to transitive events with more or less prototypical Agents and Patients, testing our first question, and to generalize these categories to intransitive events, testing our second question.

Method We recruited 108 English, 109 Basque, and 72 Hindi speakers who completed the study online. In the training phase of the experiment, participants saw 28 images of one figure acting on another. Either the Agent or the Patient was shaded red (see Figure 1). Participants learned to group the pictures into “Agent” and “Patient” categories (labelled Category “A” or “B”), receiving accuracy feedback on every trial. Participants then completed a test phase where they viewed new images and decided whether the scenes belonged to Category “A” or “B”. This test phase included both transitive and intransitive scenes. The *transitive scenes* featured more or less prototypical Agents and Patients (e.g., the roles in Figure 1 being more prototypical; the roles in a scene of one person whispering to another being less prototypical). We used the prototypicality norms in Rissman and Lupyan [6], who normed the transitive scenes for six of Dowty’s proto-Properties: intentionality, causation, movement, change of state, affectedness, and being stationary. The *intransitive scenes* featured both activity-oriented events (e.g., jumping, running) and state-oriented events (e.g., someone grabbing their stomach as if sick); see examples in Figure 2. Across all participants, we tested 48 transitive scenes and 48 intransitive scenes. Each participant viewed 48 transitive trials (half with a red Agent and half with a red Patient) randomly interspersed with 24 intransitive trials (showing a single, red-shaded individual). No feedback was provided on the test trials.

Figure 1. A sample training picture

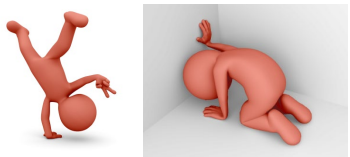


Figure 2. Sample intransitive scenes

Results & Discussion Test accuracy for transitive scenes was high: English, Basque, and Hindi speakers correctly categorized the pictures into Agent and Patient categories on 90% of trials (CI_{95} = [88%, 92%]). The same proto-Properties predicted generalization accuracy in the three languages. Participants were more accurate when the Agent was more intentional (English: b = .49, CI_{95} = [.22, .73]; Basque: b = .52, CI_{95} = [.24, .81]; Hindi: b = .55, CI_{95} = [.14, .96]) and when the Agent caused the event (English: b = .26, CI_{95} = [.004, .52]; Basque: b = .46, CI_{95} = [.18, .74]; Hindi: b = .50, CI_{95} = [.1, .9]). These results suggest that English, Basque, and Hindi speakers represent transitive event roles in highly similar ways.

Does this similarity extend to intransitive scenes, for which the three languages use diverging grammatical systems? Rates of classifying the intransitive pictures into the Agent category are shown in Figure 3. Basque and English speakers tended overall to sort intransitive pictures as Agents, and rates of classifying individual scenes in the Agent category were strongly aligned across these two languages: $r(46) = .83$, $p < .001$. For Hindi speakers, by contrast, Intransitive scenes were equally likely to be categorized as Agents or Patients. In addition, Agent sorting rates for individual scenes were not significantly correlated between Hindi and English ($r(46) = .27$, $p > .1$) or between Hindi and Basque ($r(46) = .15$, $p > .1$). These results suggest that the syntactic difference between Hindi, English, and Basque (where intransitive arguments in Hindi do not receive ergative case) may have influenced participants' conceptualization of these roles.

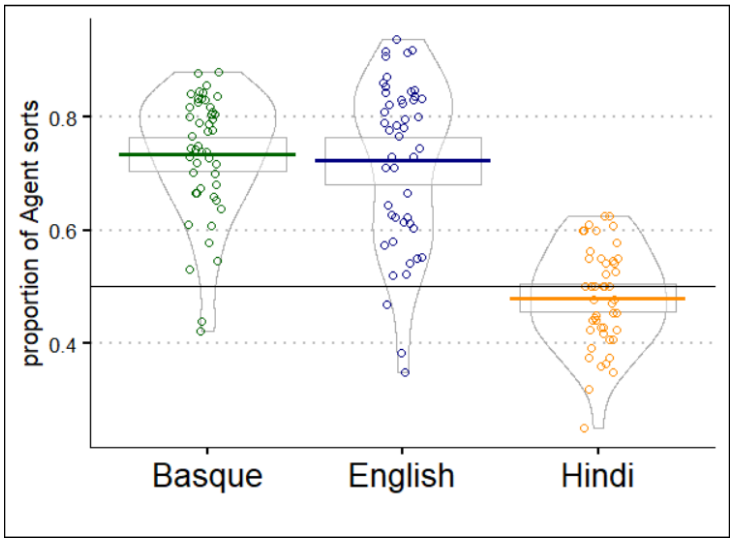


Figure 3. Rates of classifying individual intransitive scenes into the Agent category for Basque vs. English vs. Hindi speakers. Horizontal lines show mean proportion of Agent sorts.

In summary, English, Basque, and Hindi speakers represent transitive Agents in terms of the same prototype, despite the syntactic differences between these languages. Nonetheless, participants sorted the intransitive pictures in divergent ways. This suggests a partial role for syntax in the task: participants were sensitive to the semantics of the intransitive events (a jumping person was more likely to be categorized as an Agent than a sick person) but participants may also have been influenced by the syntactic groupings in their language. This raises the question of whether Hindi speakers conceptualize Agency in different ways than Basque and English speakers do.

References

1. Strickland, B., *Language reflects "core" cognition: A New theory about the origin of cross-linguistic regularities.* Cognitive Science, 2017. **41**: p. 70-101.
2. Zuberbühler, K. and B. Bickel, *Transition to language: From agent perception to event representation.* WIREs Cognitive Science, 2022. **13**(6): p. e1594.
3. Rissman, L. and A. Majid, *Thematic roles: Core knowledge or linguistic construct?* Psychonomic Bulletin & Review, 2019. **26**(6): p. 1850-1869.
4. Comrie, B., *Language universals and linguistic typology: Syntax and morphology*. 1989, Chicago, IL: University of Chicago Press.
5. Dowty, D., *Thematic proto-roles and argument selection.* Language, 1991. **67**(3): p. 547-619.
6. Rissman, L. and G. Lupyan, *A dissociation between conceptual prominence and explicit category learning: Evidence from agent and patient event roles.* Journal of Experimental Psychology: General, 2022. **151**(7): p. 1707.

Why is “tree skin” better than “human bark”: Semantic centrality predicts asymmetries in metaphorical extensions

Background. People have a remarkable ability to draw analogies between different domains, an ability often showcased by metaphors. For instance, we frequently interpret the concept of life through the lens of a journey, where life is viewed as a path we travel on, starting at birth and encountering various challenges along the way. However, such mappings often exhibit an asymmetry – for example, we rarely if ever use life to understand journeys. A common explanation for this asymmetry is that metaphors typically map from more concrete to more abstract domains, rather than the other way around. Conceptual metaphor theorists (Kovecses, 2010; Lakoff & Johnson, 1980) proposed that our physical experience provides a natural basis for understanding more abstract ideas, and concreteness explains why in most everyday metaphors the (more concrete) source and the (more abstract) target are not reversible. Concreteness is also proposed to explain the metaphorical extension of meaning. For example, terms denoting sensory experiences are regularly used to communicate more abstract concepts like rationality, as in the phrase “You are blinded by love” to mean that one is not acting rationally. The opposite mapping is rather more difficult to conceive.

However, concreteness falls short of explaining asymmetries when mappings between two relatively concrete domains. For example, English and Russian use “balls” and “eggs” respectively to refer to testicles. Opposite mappings are rarely if ever attested. English speakers find it relatively easy to understand a novel mapping such as the use of “skin” to refer to bark (as in “tree skin”) as done in Mandarin Chinese among other languages. The reverse (equally unfamiliar) mapping—using “human bark” to refer to “skin” seems rather more strained. Researchers have attempted to explain these asymmetries in several ways. For example, Bottini and Casasanto (2013) argued that the source domain may be relatively more familiar, perceptually available, imageable, memorable. Dancygier and Sweetser (2014) further suggest that the source domain's higher intersubjective accessibility – its ease of being accessed and shared among multiple speakers – makes metaphors a valuable tool in communication for aligning understanding of less accessible domains. Aligning with the accessibility account, Winter and Srinivasan (2022) proposed that word frequency is a good explanation for asymmetry in cross-domain mapping, as more frequent words are easier to access, more familiar, and more memorable, making them ideal sources of metaphorical meaning extension. Consistently, they found frequency as a robust predictor of asymmetry in the metaphorical extension of meaning across languages.

However, a reliance on word frequency as an explanation begs the question of why words from the source domain are more frequent in the first place. In a series of studies, Liu et al. (2023) found that—controlling for multiple confounds—word frequency was predicted by measures of semantic centrality: the number of connections the word and its surrounding words have (as measured by, e.g., Laplacian centrality), and the ability of the word to connect less interconnected words (as measured by, e.g., Burt’s constraint). These network properties not only predicted synchronic word frequency, but centrality measures taken at one point predicted which words decreased and which words increased in frequency later, suggesting a potential causality link between network centrality and word frequency. Here, we extend this approach to examine whether network centralities can help explain the asymmetry in metaphorical extensions.

Method. We used data from Urban (2011) that contains 71 concept pairs that have cross-linguistic asymmetries in their semantic extensions (e.g., skin ~ bark, ball ~ testicle). We matched the translation equivalents of concept pairs in English from Urban (2011) with concreteness data (Brysbaert et al., 2014) and word frequency (Google Ngram). We also use two network centralities: Burt’s constraint and Laplacian centrality as computed from English word associations (De Deyne et al., 2019) as proxies for semantic centrality. For concepts with multiple English equivalents (e.g., 'road/street/way'), we calculated the average frequency, concreteness, and centrality values across these terms. We then applied a mixed logistic regression model, predicting whether a concept is the source domain of that concept pair from the fixed effects: log frequency, concreteness, Burt’s Constraint, and Laplacian Centrality (all standardized as

z-scores) of that concept. The model also included random intercepts and random effects for frequency, concreteness, and centrality measures by concept pair. This regression model was estimated using the *brms* package in R (Bürkner, 2017), with a weakly informative prior (normal distribution with mean 0 and standard deviation 1).

Results and discussion. Winter and Srinivasan (2022) found word frequency was a robust predictor of asymmetry in the semantic extension of meaning across multiple languages, and the concreteness of a word doesn't predict whether it's more likely to be the source of extension. Our re-analysis shows that semantic centralities are better predictors ($Burt's\ constraint: \beta = -1.4, SE = 0.68, 95\%\ credible\ interval\ [-0.12, 2.78]$, odds: 4 to 1; Laplacian Centrality: $\beta = 1.04, SE = 0.53, 95\%\ credible\ interval\ [0.14, 2.18]$, odds: 3 to 1]. The results suggest that the less a word is "constrained" by its neighbors (by bridging neighbors that are not interconnected among themselves), and the more connections a word and its neighbors have, the more likely the word will become the source of metaphorical semantic change. Importantly, when we include semantic centralities as predictors, word frequency ceases to be a significant predictor ($\beta = 0.51, SE = 0.46, 95\%\ CI\ [-0.35, 1.45]$. Concreteness, hypothesized by conceptual metaphor theories to explain the asymmetry is also not in fact predictive of it ($\beta = .07, SE = 0.34, 95\%\ CI\ [-0.64, 0.72]$.

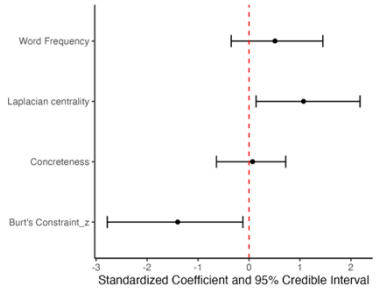


Figure 1 Standard Coefficients in predicting asymmetry of semantic changes. Error bars indicate 95% Credible Interval.

This finding highlights the significant role of a word's semantic centrality in understanding the dynamics of semantic extension and metaphorical asymmetry. Traditional metrics like concreteness fall short of explaining why concepts of similar concreteness are used metaphorically to represent each other. The accessibility hypothesis instead posits that more accessible words are likelier to become sources in metaphorical extensions. We posit a linking hypothesis that connects centrality, accessibility, and frequency by arguing that words with a central position in the network are more frequently activated during speech comprehension and production. This higher activation level is due to the increased input these words receive from their neighboring connections, enhancing their accessibility and, consequently, the likelihood of their use and extension to new meanings. Furthermore, words that serve as bridges in less connected network segments tend to have higher contextual diversity and wider semantic ranges, making them better candidates for metaphorical extension in contrast to words situated in densely interconnected clusters, which often have narrower and more redundant semantic contexts. The current analysis does not definitively establish a causal link between network centrality and asymmetry in semantic change. However, future longitudinal studies could provide deeper insights by e.g., analyzing how words with similar levels of metaphorical usage but differing network positions influence the likelihood of metaphorical extension at a later time. Additionally, it's also possible to experimentally manipulate a word's position within a participant's semantic network to see if it causes changes in the propensity of that word to be metaphorically extended.

References. Bottini, R., & Casasanto, D. (2013). Space and time in the child's mind: Metaphoric or ATOMic? |Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas.|Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan.|Dancygier, B., & Sweetser, E. (2014). *Figurative Language*.|De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. | Kovecses, Z. (2010). *Metaphor: A Practical Introduction*. |Lakoff, G., & Johnson, M. (1980). Conceptual Metaphor in Everyday Language. |Liu, Q., De Deyne, S., Jiang, X., & Lupyan, G. (2023). Understanding the Frequency of a Word by its Associates: A Network Perspective. |Urban, M. (2011). Asymmetries in overt marking and directionality in semantic change.|Winter, B., & Srinivasan, M. (2022). Why is Semantic Change Asymmetric? The Role of Concreteness and Word Frequency and Metaphor and Metonymy.