

Vagueness and pragmatic reasoning in quantified sentences

Alexandre Cremers

BKKI, Filologijos Fakultetas, Vilniaus Universitetas

alexandre.cremers@gmail.com

Abstract

Vagueness has been extensively studied as a standalone topic, but little is known of its interactions with other semantic and pragmatic phenomena. We present an experiment testing a Rational Speech-Act model from Cremers (2022) which aims to offer a general account of the pragmatic use of vague sentences and the implicatures they give rise to on a new class of sentences where a vague term restricts a universal quantifier. We found that an implicature predicted by the model is present, but a separate prediction on borderline cases isn't verified. We discuss consequences for the modeling of pragmatic uses of vague sentences.

1 Vagueness and implicatures

Standard theories of implicatures (be it Neo-Gricean or grammatical) predict that a sentence like (1) should compete with an alternative without the modifier, and this would lead to an implicature that Mary is (somewhat) tall (Simons, 2001). Yet, Leffel et al. (2019) show that as long as *very* is not stressed, this implicature is absent (if anything, participants conclude that Mary is rather short). They explain this observation as the result of an interaction between the vagueness of the adjective and the mechanism deriving implicatures.

- (1) Mary is not very tall \nrightarrow Mary is tall

Informally, Leffel et al. (2019) argue that no height can at the same time count as clearly 'not very tall' and clearly 'tall', and this would prevent the implicature from being derived. To capture this, they extend Fox's (2007) mechanism of innocent exclusion (which blocks some contradictory implicatures in a grammatical theory) to also rule out "borderline contradictions" (vague sentences such as 'Peter is tall and not tall'; Ripley, 2011). Cremers (2022) later recast this proposal in slightly more pragmatic terms within the Rational Speech-Act framework (Frank and Goodman, 2012), using a probabilistic version of supervaluationism from Spector (2017). We refer to Cremers (2022) for a general presentation of the model, but in short, the utility of a candidate vague utterance takes into account all possible (first-order) vague denotations, in proportion to their likelihood (second-order vagueness). This, combined with explicit reasoning on possible strengthened meanings (Franke and Bergen, 2020), assigns a probability close to zero to the exhaustified reading in (1).

Nevertheless, few other cases of potential interactions between vagueness and implicatures have been tested, and given the complexity and degrees of freedom in these RSA models, this raises concerns of possible overfit. We test quantified vague sentences which also have the potential to raise implicatures in order to test the model, both qualitatively and quantitatively.

2 Quantified sentences and predictions

- (2) Every tall student laughed $\overset{?}{\rightarrow}$ not every student laughed

Take (2) for instance. Superficially, this example shares a lot with (1): in both cases, a vague term occurs in a downward entailing environments and a more informative alternative

can be derived by deletion (*very* in (1), *tall* in (2)). Unlike (1), the proposals discussed above predict that (2) should have an implicature that some students didn’t laugh (the strengthened meaning is **not** a “borderline contradiction” in this case since a short student can validate the implicature). Cremers (2022) makes an additional prediction: for (2) to be felicitous, borderline tall students should laugh too (a student who counts as tall under at least *some* denotations but didn’t laugh would invalidate the sentence). In probabilistic terms, this theory predicts a slight decrease in laughing probability for short students (due to the implicature) but a marked increase for borderline-tall students (due to the supervaluationist mechanism). As usual with vagueness, it is difficult to delineate between clearly-not-tall and borderline-tall (higher-order vagueness problem), and specific predictions from the model depend on the value of all model parameters. In §3.3 we present details specific to our experiment (worlds, messages, costs...)

3 Experiment

3.1 Design

We designed an experiment to test how participants interpret sentences like (2), with both relative (vague) and absolute (non-vague) gradable adjectives. Participants first saw a table with 8 items and their dimensions on a scale relevant to the gradable adjective, and were asked about three of them “Would you say that the following [items] are [adj]?” (Q1) on a slider from ‘no’ to ‘yes’. The three items were chosen to represent a clear true, clear false, and borderline case (or a case very close to the boundary for absolute adjectives). We then probed participants’ out-of-the-blue prior beliefs about an unrelated predicate (e.g., playing video games) for each of the three items tested on the previous question, using a slider from ‘certainly the case’ to ‘certainly not the case’ (Q2). Participants had no reason to make different a priori judgments for this predicate on the different items, unless they assumed a correlation with the vague predicate, so measuring them all was safer. Finally, the story introduced someone familiar with the matter, who uttered a sentence like (2). We then asked participants to convey their posterior belief about the predicate of Q2 given what they heard, again using continuous sliders (Q3).

We tested 12 gradable adjectives (6 relative, 3 min standard, 3 max standard) in the restrictor of either ‘every’ or ‘no’. Each adjective came with a different scenario (different items, different predicate), all listed in Table 1. For each adjective, we tested two potential “borderline” cases (of which each participant saw only one, randomly selected). After rejecting participants who took the survey multiple times, 417 participants were recruited on Amazon’s MTurk and paid \$0.47 for their participation. We excluded 51 participants whose responses on Q1 suggested they were not focused on the task,¹ plus one participant whose data wasn’t properly recorded.

3.2 Results

Data and scripts available at <https://github.com/Alex-Cremers/vague-restrictors-AC2024>.

Figure 1 shows participants’ response to Q1 (i.e., how much they agree with an attribution of the adjective for each item that was tested). We see that even for absolute adjectives, it is

¹The initial criterion was to reject all participants whose response at one or both scale ends fell on the wrong side of the slider, e.g., participants who assigned the shortest student a “tallness” of over 50%, or the tallest one, less than 50%. This led to the rejection of 89 participant (21%), which an anonymous reviewer found surprisingly high for this task. On further inspection, it turned out that many participant seemed to genuinely find the lowest value to fall to some extent in the positive extension of some predicates. This was particularly true of ‘hot’ (possibly because the experiment was run in winter?). I therefore updated the criteria to include all participants who saw a relative adjective as long as their responses increased with the underlying measure, even if it started too high or ended too low. I didn’t apply such an inclusion criterion to absolute adjectives, because their scale ends had to be clear cases, even under a loose interpretation. I thank the reviewer for drawing my attention to this.


	Adjective	Measure	Item	Scope Predicate
relative	tall	ft, in	students	plays video games
	powerful	hp	car	has a touchscreen
	hot	°F	day	was a workday
	expensive	\$	car	has the gas tank door on the left side
	large	sq ft	apartment	has West-facing windows
	young	yr	employee	is left-handed
min	late	time	employee	works in HR
	spicy		(Thai) dish	contains lemongrass
	profitable	\$	company	is a construction company
max	complete	%	investigation	was about credit card fraud
	safe	crime rate	neighborhood	is situated North of city center
	full	free seats	flight	departed in the morning

Table 1: Adjectives by type, and associated items and predicates. ‘late’ was relative to 8:30am. For ‘young’, ‘full’ and ‘safe’, the measure is inversely correlated to measurement units.

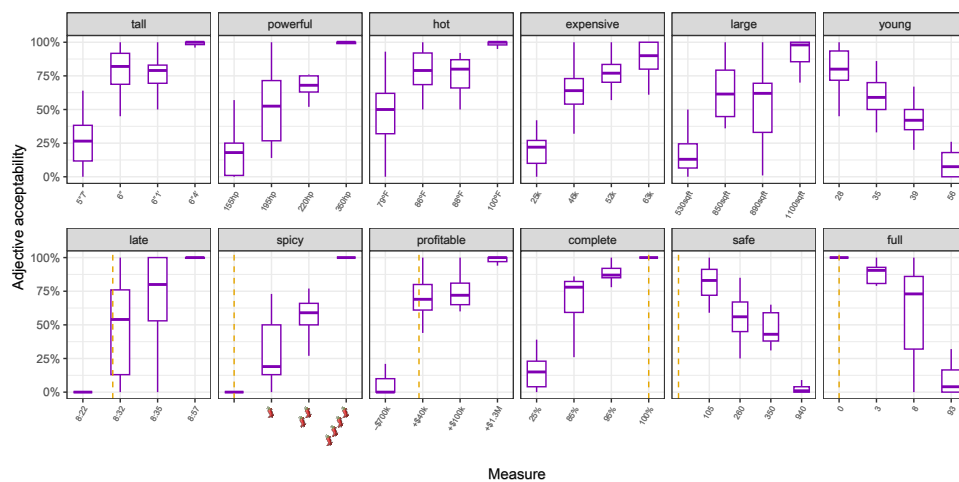


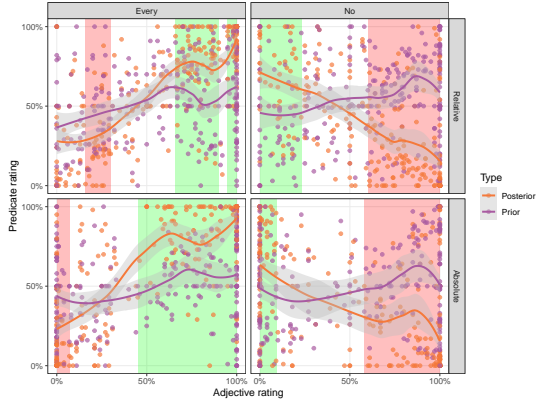
Figure 1: Responses to Q1 as a function of the measure given to participants. For closed scales, we indicate the approximate position of the relevant boundary (dashed orange line).

possible to find cases where ratings are mixed (be it genuine vagueness or a different phenomenon). Figure 2a shows responses to Q2 and Q3 as a function of Q1. A linear mixed-effects model on responses to Q2 with the interaction of scaled adjective rating (answers to Q1) and adjective as a predictor showed a significant relation between Q2 and Q1 ($\chi^2(12) = 122.7, p < .001$), with coefficients ranging from -0.31 for ‘late’ (late employees are less likely to work in HR) to 0.81 for ‘powerful’ (powerful cars are more likely to have a touchscreen), confirming that we cannot assume the target predicates to be independent from the vague restrictors, a fact which will complicate the Bayesian model. We then run a theory-neutral analysis aimed at identifying areas where the posterior differs significantly from the prior. The results are presented in Table 2b. In all 4 cases we found a large cluster corresponding to the literal contribution of the sentence, as well as a smaller cluster of opposite sign at the bottom of the scale, corresponding to the existential implicature. We therefore confirm the presence of this implicature with ‘no’ and ‘every’, for both relative and absolute adjectives, in line with the prediction of Leffel et al. (2019).

3.3 Model evaluation

Implementation: For simplicity, we restrict our attention to relative adjectives only and leave the analysis of absolute adjectives for future research.² We also simplify the model compared to

²The case of minimum standard adjectives is complicated by two facts: they can be ambiguous between a strict reading and a relative-like reading (Qing, 2021), and the threshold of their relative-like interpretation tends



(a) Prior (Q2) and posterior (Q3) ratings on the scope predicate as a function of applicability of the restrictor adjective (Q1). Significant positive/negative clusters are highlighted in green/red. Smoothing lines added for illustrative purposes.

Quant.	Adj. class	cluster range	C	p
every	relative	[16%, 30%]	-42	.002
		[66%, 90%]	85	< .001
	absolute	[95%, 100%]	35	.004
		[0%, 7%]	-29	.008
no	relative	[45%, 100%]	213	< .001
		[0%, 23%]	66	< .001
	absolute	[60%, 100%]	-220	< .001
		[0%, 9%]	28	.008
		[58%, 100%]	-170	< .001

(b) Significant clusters of differences between prior and posterior judgments. We adapted the non-parametric permutation test proposed in Maris and Oostenveld, 2007. The analysis was run independently for each Quantifier and Adjective class (relative vs. absolute).

Figure 2: Responses to Q2 and Q3 as as function of Q1, and cluster analysis results.

Cremers (2022) by fitting a unique set of costs and rationality parameters for all participants (instead of using by-subject random effects on these parameters).

In the context of our experiment, the question of the vague restrictor (e.g., tall) is settled, and we take the set of worlds to represent the $2^8 = 256$ possible ways to realize the binary scope predicate (e.g., laughing), which correspond to the cells of the unique QUD we will consider (“which students laughed?”). Table 2 presents the set of messages we considered. The quantifier ‘no’ is formally identical to ‘every’ if we flip the target predicate (“no tall student laughed” is equivalent to “every tall student didn’t laugh”).

We assume that the threshold θ for the adjective follows a normal distribution parametrized by $\Theta = (\mu, \sigma)$ (first-order vagueness). Θ itself is assumed to follow a hybrid multivariate normal/lognormal distribution parametrized by $\Omega = (m_\mu, m_\sigma, s_\mu, s_\sigma, \rho)$ (Fletcher and Zupanski, 2006) (second-order vagueness plays a crucial role in the RSA-SvI). Following Cremers (2022), we obtained these parameters by fitting an independent Bayesian mixed-effects model for each adjective on responses to Q1, where each participant is assigned a Θ .

For $u_{\forall \text{adj}}$, we propose the literal interpretation in (3), where $d_{\text{last}}(w)$ is the height of the tallest student who didn’t laugh in w (every tall student laughed if and only if the tallest student who didn’t laugh, if any, doesn’t exceed θ_{tall}), and d_{max} is the height of the tallest student (which unlike d_{last} is settled in this case).³

$$(3) \quad L_0(w|u_{\forall \text{adj}}, \Theta) \propto P(w)P(d_{\text{last}}(w) \leq \theta < d_{\text{max}})$$

To compute (3), we need the prior on all 256 worlds, but Q2 only probed participants’ priors on 3 items, so we first used their response to fit a prior on the whole 8 item scale, and then

not to follow a normal distribution (in Cremers (2022), they are modeled with exponential distributions).

³ d_{max} is there to account for the presupposition that the restrictor isn’t empty (at least one student must count as tall, otherwise the target sentence becomes a null message), which we assimilate with falsity.

	message	cost
u_0	“...”	c_0
u_{WE}	“A, B, and C”	$n \times c_1$
u_{SE}	“only A, B, and C”	$n \times c_1 + c_{\text{SE}}$
$u_{\forall \text{adj}}$	“every tall student”	$c_{\text{every}} + c_{\text{adj}}$
u_{\forall}	“every student”	c_{every}

Table 2: Messages considered in the model, and associated cost. List answers receive a cost dependent on the length of the list. We used a different cost for ‘no’, and each adjective had its own cost.

derived priors for worlds.⁴ This was again done independently for each adjective via mixed-effects models with random by-participant intercepts and slope, using the logit of priors squeezed onto $[0.05, 0.95]$ as the dependent variable and normalized degrees as the regressor.

Having obtained the parameters Ω and the prior $P(w)$, we were able to precompute the informativity component of U_1 for all the messages, in each world, and for each participant. We used it to fit a single Bayesian model on data from Q3 with the remaining parameters of the model: the costs variables ($c_0, c_1, c_{SE}, c_{every}, c_{no}, (c_{adj})$) and the rationality parameter α . We obtain the L_1 posterior on proposition φ by summing posterior on world-parse pairs on both parses of the target sentence across φ -worlds.

We compared this model to two simpler “literal” models where instead of $L_1(\varphi(X)|u)$ we assumed that the participants’ responses follow $\int P(\theta)L_0(\varphi(X)|u^i, \Theta)d\Theta$ for $i = \text{LIT}$ or EXH . The model adopts the grammatical theory of implicatures, so we can have a “literal” listener with an exhaustified parse (but it involves no reasoning on parses, second-order vagueness, or alternative messages).

parameter	prior	mean	SD	90% CrI
c_0	$\mathcal{N}_+(0, 5)$	0.17	0.17	[0.01, 0.51]
c_1	$\mathcal{N}_+(0, 0.5)$	2.28	0.35	[1.74, 2.88]
c_{SE}	$\mathcal{N}_+(0, 1)$	3.59	0.81	[2.22, 4.89]
c_{every}	$\mathcal{N}_+(0, 1)$	1.10	0.58	[0.22, 2.11]
c_{no}	$\mathcal{N}_+(0, 1.5)$	3.26	1.00	[1.68, 4.97]
$c_{expensive}$		1.48	0.72	[0.32, 2.71]
c_{hot}		0.69	0.49	[0.07, 1.62]
c_{large}	$\mathcal{N}_+(0, 1)$	0.58	0.51	[0.03, 1.60]
$c_{powerful}$		0.80	0.65	[0.05, 2.09]
c_{tall}		1.58	0.76	[0.39, 2.91]
c_{young}		1.67	0.76	[0.43, 2.95]
α	$\Gamma(4, 1)$	1.69	0.20	[1.38, 2.03]

(a) Prior and posterior distribution for all parameters of the RSA-SvI L_1 model. \mathcal{N}_+ indicates a truncated Gaussian. c_0 is the cost of the null message, c_1 the linear coefficient for list answers.

Model	elpd _{loo}	Δ_{elpd}	SE_{Δ}	p_{loo}
L_0^{EXH}	-96.4	0	0	1.9
L_0^{LIT}	-117.6	-21.2	4.4	1.8
L_1	-118.0	-21.5	5.0	9.2

(b) PSIS LOO-CV model comparison. The expected log predictive density (elpd) is a measure of how well the model predicts unseen data (Gelman, Hwang, and Vehtari, 2014), and naturally penalizes models with too many degrees of freedom. Δ_{elpd} is the difference with the best model, and SE_{Δ} the standard error on this difference. p_{loo} is the estimated effective number of parameters. Pareto influence values indicated that the approximation was valid (all $\hat{k} < 0.34$ across all three models, values under 0.7 are considered good).

Table 3: Posterior distributions of the parameters in the L_1 model (a) and comparison with the L_0 models (b).

Results: All models converged perfectly. Figure 3 shows the aggregated predictions of each of the three models and the measured posteriors. Table 3a gives the priors for each parameter in the L_1 model and the mean, SD, and quantiles of the posterior distribution. Looking at parses, we found that the L_1 model assigned a higher posterior probability to the exhaustified parse for all participants but one outlier (who had a posterior probability of 27%). The posterior probability for other participants ranged from 50% to 92%, with an average of 55%.

The models were compared using the PSIS Leave-one-out CV approximation (Vehtari, Gelman, and Gabry, 2017; Vehtari, Simpson, et al., 2024). The details are presented in Table 3b. In short, the L_0^{EXH} model outperforms the two other models, which are indistinguishable.

⁴We assumed that the predicate on each individual item is independent. For instance, a world in which only students A, B, and E laughed would receive prior probability $p_{APB}(1 - p_C)(1 - p_D)p_E(1 - p_F)(1 - p_G)(1 - p_H)$, where $p_A \dots p_H$ are the value fitted from the participant’s response to 3 of these items.

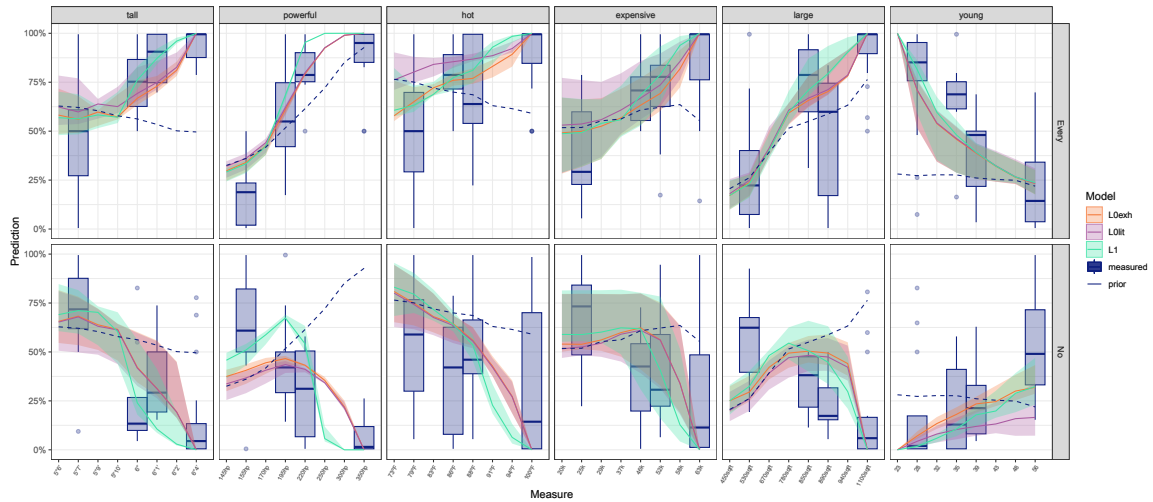


Figure 3: Predictions of the pragmatic L_1 model and the two L_0 models, compared to priors and measured posteriors, by degree. The ribbons indicate the median and quartiles of the predictions at each degree. The “supervaluationist” prediction of the L_1 model manifests itself as the green curve standing above the orange one in the upper-mid range with ‘every’, under for ‘no’ (most visible with ‘powerful’).

4 Discussion

Our results expand our empirical knowledge of the complex interactions between vagueness and pragmatic reasoning and confirm Leffel et al.’s (2019) prediction that an implicature should this time be present. Yet, they seem to invalidate an additional prediction about borderline cases from Cremers (2022), a concrete implementation of this theory.

As usual, falsifying a probabilistic model isn’t straightforward. We showed that a simple L_0 model performed better on relative adjectives in our data, but we already know that it would not capture Leffel et al.’s (2019) data. Since the supervaluationist component that allowed Cremers (2022) to capture the contrast between absolute and relative adjectives (probabilistic supervaluationism on second-order vagueness) makes an incorrect prediction, we could imagine recasting Leffel et al.’s explanation without it, assuming the explanation for the puzzle (1) can be saved. Yet, it remains unclear whether supervaluationism is the main culprit for the poor fit here. In particular, the fact that both L_0 and L_1 models underestimate the impact of the “not every” implicature points to the well-known exacerbated effect of priors in the RSA framework (Degen, Tessler, and Goodman, 2015; Schreiber and Onea, 2021; Cremers, Wilcox, and Spector, 2023). Alternatively, one of our simplifying assumptions might be the issue (independence assumption to derive priors on worlds, shared adjective denotations across participants). Finally, we postponed an evaluation of the model on absolute adjectives, which may favor the L_1 model (after all, it was initially designed to account for a contrast between absolute and relative adjectives).

To conclude, our data suggests that giving a precise, explicit account of the interaction between vagueness and pragmatics remains elusive (and we haven’t touched on speakers with partial knowledge Egré et al., 2023). Understanding the pragmatics of vague sentences is challenging on empirical grounds (requiring precise quantitative data) and theoretical grounds (models tend to be computationally more complex than for other phenomena), but it could eventually shed new light on long-standing questions in the literature on vagueness, in particular, why does human communication rely so heavily on vague terms?

Acknowledgements. This research was made possible by funding from the Research Council of Lithuania and the European Social Fund under Measure 09.3.3-LMT-K-712.

References

- Cremers, Alexandre (2022). “A Rational Speech-Act model for the pragmatic use of vague terms in natural language”. In: *Proceedings of CogSci 44*. Ed. by J. Culbertson et al., pp. 149–155.
- Cremers, Alexandre, Ethan Wilcox, and Benjamin Spector (2023). “Exhaustivity and anti-exhaustivity in the RSA framework: Testing the effect of prior beliefs”. In: *Cognitive Science* 47.5, e13286. DOI: 10.1111/cogs.13286. URL: <https://arxiv.org/abs/2202.07023>.
- Degen, Judith, Michael Henry Tessler, and Noah D Goodman (2015). “Wonky worlds: Listeners revise world knowledge when utterances are odd.” In: *CogSci*.
- Egré, Paul et al. (2023). “On the Optimality of Vagueness: “Around”, “Between”, and the Gricean Maxims”. In: *Linguistics and Philosophy* 46.5. DOI: 10.1007/s10988-022-09379-6.
- Fletcher, Steven J and Milija Zupanski (2006). “A hybrid multivariate normal and lognormal distribution for data assimilation”. In: *Atmospheric Science Letters* 7.2, pp. 43–46.
- Fox, Danny (2007). “Free Choice Disjunction and the theory of Scalar Implicature”. In: *Presupposition and implicature in compositional semantics*. Ed. by Uli Sauerland and Penka Stateva. New York, NY: Palgrave Macmillan, pp. 71–120.
- Frank, Michael C and Noah D Goodman (2012). “Predicting pragmatic reasoning in language games”. In: *Science* 336.6084, pp. 998–998.
- Franke, Michael and Leon Bergen (2020). “Theory-driven statistical modeling for semantics and pragmatics: A case study on grammatically generated implicature readings”. In: *Language* 96.2, pp. 77–96.
- Gelman, Andrew, Jessica Hwang, and Aki Vehtari (2014). “Understanding predictive information criteria for Bayesian models”. In: *Statistics and computing* 24, pp. 997–1016.
- Leffel, Timothy et al. (2019). “Vagueness in Implicature: The Case of Modified Adjectives”. In: *Journal of Semantics* 36.2, pp. 317–348. ISSN: 0167-5133. DOI: 10.1093/jos/ffy020.
- Maris, Eric and Robert Oostenveld (2007). “Nonparametric statistical testing of EEG-and MEG-data”. In: *Journal of neuroscience methods* 164.1, pp. 177–190.
- Qing, Ciyang (2021). “Zero or minimum degree? Rethinking minimum gradable adjectives”. In: *Proceedings of Sinn und Bedeutung* 25, pp. 733–750. DOI: 10.18148/sub/2021.v25i0.964.
- Ripley, David (2011). “Contradictions at the borders”. In: *Vagueness in communication*. Ed. by Rick Nouwen et al. Springer, pp. 169–188.
- Schreiber, Alexander and Edgar Onea (2021). “Are Narrow Focus Exhaustivity Inferences Bayesian Inferences?” In: *Frontiers in Psychology* 12.
- Simons, Mandy (2001). “On the conversational basis of some presuppositions”. In: *Proceedings of SALT*. Ed. by R. Hastings, B. Jackson, and Z. Zvolensky. Vol. 11, pp. 431–448.
- Spector, Benjamin (2017). “The pragmatics of plural predication: Homogeneity and Non-Maximality within the Rational Speech Act Model”. In: *Proceedings of the 21st Amsterdam Colloquium*. Ed. by Alexandre Cremers, Thom van Gessel, and Floris Roelofsen, p. 435.
- Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Statistics and Computing* 27.5, pp. 1413–1432. DOI: 10.1007/s11222-016-9696-4.
- Vehtari, Aki, Daniel Simpson, et al. (2024). “Pareto Smoothed Importance Sampling”. In: *Journal of Machine Learning Research* 25, pp. 1–57.