How do Large Language Models process scalar Hurford Disjunctions?

Adèle Hénot-Mortier (MIT)

November 19, 2024

Harvard Language and Cognition Talk Series

Intro to the Intro

- I can't seem to manage my priorities so this whole thing appeared over the past week.
- As a result it's very much work in progress, and the conclusions still do not appear super clear to me.
- I'm happy to get any kind of input, on the theoretical side as well as on the technical side, especially stats – I suck at stats.
- I thank Athulya Aravind, Amir Anvari, Danny Fox, Martin Hackl, Nina Haslinger and Viola Schmitt for advising me on the more theoretical aspects of this project. I thank Forrest Davis who encouraged me to pursue that kind of research 2 years ago, and at that time told me about the ImpPres paper! I also thank Roger Levy who allowed me to follow and then TA for the Computational Psycholinguistics class at MIT, which turned out to be super helpful for all my computational projects.

Introduction

Hurford Disjunctions

- Disjunctions featuring **entailing disjuncts** tend to be odd (Hurford, 1974). Such disjunctions are called Hurford Disjunctions (**HD**).

(1) a. # Jo owns a dog or a poodle.

b. # Jo owns a poodle or a dog.

- In (1) above, oddness does not seem to depend on the order of the disjuncts (logically weaker vs. logically stronger).
- Oddness extends to disjunctions featuring merely **compatible disjuncts** (Singh, 2008b). In (2), the sentences seem to imply dogs are never hypoallergenic (poodles are)...

(2) a. # Jo owns a hypoallergenic pet or a dog.

b. # Jo owns a dog or a hypoallergenic pet.

Hurford's Constraint

- Descriptively, **Hurford's Constraint** amounts to the observation that disjunctions should not feature compatible disjuncts. Here we do not discuss how to model this constraint in an explanatory way.[1]
- We take it as a premise and see **how specific disjunctions can escape it**.

[1] See Katzir and Singh, 2014; Meyer, 2015; Mayr and Romoli, 2016 (i.a.) for different views on how Hurford's Constraint can be explained.

Escaping Hurford's Constraint: non-scalar case

- It is possible to "repair" the HDs in (1) and (2) by making the disjuncts explicitely incompatible.
- The sentences may seem a bit convoluted and may need to answer specific kinds of questions, by are not as odd as the ones without repair.

(1) Does John own a Siamese, a poodle, or another kind of dog?

   a. Jo owns a dog **that is not a poodle**, or a poodle.

   b. Jo owns a poodle or a dog **that is not a poodle**.

(2) Context: John loves all dogs but poodles, and tends to have pet allergies.

   a. Jo owns a hypoallergenic pet or a dog **that's not hypoallergenic**.

   b. Jo owns a dog **that's not hypoallergenic** or a hypoallergenic pet.

Escaping Hurford's Constraint: scalar case

- Interestingly, whenever the compatibility between disjuncts can be broken by (covert) pragmatic reasoning, HDs can improve (Gazdar, 1979).
- For instance in the HD (3), the entailment between *some* and *all* can be broken if *some* is taken to implicate *not all*. Under this assumption, the 2 disjuncts of (3) are no longer compatible!
- Likewise in (4), the entailment between *not all* and *none* can be broken if *not all* is taken to implicate *some* (might be harder to get).

(3) Al ate some or all of the biscuits.
$\rightsquigarrow$ Al ate some but not all or all of the biscuits.

(4) Al ate not all or none of the biscuits.
$\rightsquigarrow$ Al ate some but not all or none of the biscuits.

- We call HDs featuring scalar items like (*some*, *all*) and (*not all*, *none*), **scalar HDs**. They will be the focus of this talk.

Oddness asymmetries in scalar HDs

- The felicity of scalar HDs is subject to an asymmetry: scalar HDs in which the stronger disjunct precedes the weaker one, still feel odd.

(5) ?? Al ate all or some of the biscuits.

(6) ? Al ate none or not all of the biscuits.

- The judgments are subtle,[2] especially in the case of (6) vs. (4). But if the pragmatic enrichment story is on the right track to explain scalar HDs, this means that **it's harder to pragmatically strengthen the weaker disjunct in scalar HDs if it appears after the stronger one** (Singh, 2008a).
- In other words, whatever covert pragmatic mechanism allows to obviate Hurford's Constraint in scalar HDs, it is order-sensitive.[3]

[2] I think the subtleness of the contrasts in such sentences in fact tells us something about the nature of the incremental constraint affecting pragmatic enrichments – if such a constraint is real. Some data in the Appendix advocate for this view, and against the idea that the preference for weak-to-strong disjuncts orderings are frozen.

[3] Which notion of "order" is relevant – linear? syntactic? – remains to be investigated.

Empirical investigation of the ordering preferences in scalar HDs (Fox & Spector, 2018)

- Fox and Spector (2018) had a cursory look at the Corpus of Contemporary American English and showed that the preference for weak-to-strong orderings in scalar HDs is a **clear statistical tendency, not a sharp constraint**.

	Canonical order	Reverse order
some or all	396	53
some or many	7	0
some or most	8	1
most or all	164	152
many or all	14	2
can or must	1	0
may or must	0	0
sometimes or always	3	2
sometimes or often	19	7
often or always	16	14
possible or certain	1	0

Empirical investigation of the asymmetry in the (some, all) case

- Google search: about 1,480,000,000 results for *some or all* pair (predicted ok); about 24,000,000 results for the infelicitous *all or some* pair (ratio=62). Caveat: high degree of ellipsis may cause those pairs to behave like frozen expressions.
- Some examples below, from Facebook.

(7) Some or all (predicted ok)

a. Some or all of the snow on the ground will most likely be here next spring.

b. See some or all movies, just one low price. Come and go as you please.

c. They might account for some or all of dark matter.

(8) All or some (predicted bad)

a. Hey I was wondering if all or some of the snakeskin Demartini guitars are custom shop.

b. Huge Lot of Canning Jars, Equipment & Supplies, Buy All or Some.

c. Hope you can join us for all or some of these events leading up to Monday's eclipse.

Empirical investigation of the asymmetry in the (not all, none) case

- Google search: about 340,000+4,820,000=5,160,000 results for the *not all or none/no* pair (predicted ok); about 293,000+2,150,000=2,443,000 results for the infelicitous *none/no or not all* pair (ratio=2).
- Caveat: *(not all) or (none)* and *not (all or none)* are string-ambiguous.
- Less clear contrast than with *some or all*; consistent with intuitions.
- **This anyway shows that scalar HDs of different kinds can be found on the Internet, and are associated with specific statistical tendencies.**

LLM's processing of scalar HDs

- LLMs constitute an interesting testing ground for scalar HDs, because:
  - They can be evaluated w.r.t. **felicity**, in the form of surprisal measures.
  - Modulo fine-tuning, they can be evaluated w.r.t. (pragmatic) **inferences** – understood as a classification task.
- Do LLMs draw the kind of pragmatic inferences that the theory predicts to be crucial for rescuing scalar HDs?
- Do LLMs judge scalar HDs as more degraded than disjunction following the same structure but featuring incompatible items (e.g. *all*/*no*)? Furthermore, are they sensitive to the asymmetry between e.g. (3) and (5)?
- Do the putative contrasts in felicity connect to how LLMs evaluate compatibility/entailment between disjuncts?
- Answering these questions may shed light on whether oddness in scalar HDs may be driven by frequency-driven preferences, or by deeper logical considerations.

Background on scalar implicatures and their investigation with LLMs

Scalar implicatures (Neo-Gricean view)

- Let's have a word on the kind of pragmatic mechanism that allows to derive *not all* from *some*. For clarity we sketch the Neo-Gricean approach here (Gazdar, 1979; Sauerland, 2004 i.a.).
- Such inferences are usually called **scalar implicatures**, and are based on reasoning about why what could have been reasonably said, was *not* said.
- For instance, if Jo tells me that *Al ate some of the biscuits*, I might think that Jo could have told me something more informative, namely that *Al ate all of the biscuits*, if John knew it were true, and if it was relevant to say it.
- Given that Jo did not use the more informative sentence, I may infer that **Jo does not believe that Al ate all of the biscuits**.
- Moreover, if I can be sure that Jo had access to enough evidence regarding Al and the biscuits (i.e. Jo is **opinionated** on whether or not *Al ate all of the biscuits*), I may strengthen this inference by concluding that **John believes that Al did not eat all of the biscuits**.

"Reverse" scalar implicatures (Neo-Gricean view)

- One can apply a similar reasoning to an utterance like *Al did not eat all of the biscuits*, where the stronger item is under negation.
- Indeed, a more informative alternative to *Al did not eat all of the biscuits*, is *Al ate none of the biscuits*.
- Assuming the speaker is truthful, maximally informative and opinionated on the matter, one can then draw the implicature that it's not the case *Al ate none of the biscuits*, i.e. *Al ate some of the biscuits*.
- Implicatures triggered by items under negation are sometimes called "reverse" of "indirect" scalar implicatures.[4]
- The two kinds of implicatures are schematized in (9) below.

(9) a. some $\rightsquigarrow$ not all — scalar implicature

b. not all $\rightsquigarrow$ some — reverse scalar implicature

[4] Although reverse scalar implicature might be harder to get, Cremers and Chemla (2014) show experimental evidence supporting the claim that both kinds of implicature (direct, reverse) share the same processing signature.

Local scalar implicatures and how they get constrained

- One way to explain how pragmatics can help in scalar HDs, is to assume that **pragmatic reasoning can occur locally** at the level of the individual disjuncts (Spector et al., 2008; Chierchia et al., 2012).
- In (3) for instance, if *some* can be locally understood as *some but not all*, the 2 disjuncts can be correctly predicted to be incompatible.

(3) Al ate some or all of the biscuits.
$\rightsquigarrow$ Al ate some but not all or all of the biscuits.

- The infelicity of the reverse order (5) can then be captured assuming local pragmatic reasoning is incrementally constrained; one idea is that pragmatic enrichments should only be done if they happen to be *incrementally* non-weakening (Fox & Spector, 2018).
- In (5), strengthening *some* to mean *some but not all* after processing *all or ...* does not bring any new information: with or without this strenthening, the disjunction would mean the same thing!

(10) all or (some but not all) $\equiv$ some $\equiv$ all or some

LLMs and scalar implicatures

- Jeretic et al. (2020) showed that LLMs like BERT were somewhat capable of drawing scalar implicatures for the (*some*, *all*) pair.
- However, other scalar pairs were often treated as synonymous (i.e. the items were judged to entail each other).
- To reach this conclusion, LLMs were fine-tuned to perform Natural Language Inference (**NLI**), i.e. to classify pairs of sentences as entailments, contradictions, or as logically compatible (=neutral).
  - The dataset used for fine-tuning was **MultiNLI** (Williams et al., 2018), and was shown to contain very few instances of pure scalar implicatures.
  - After fine-tuning, the models performed NLI on a new dataset called ImpPres, containing pairs of sentences involving various kinds of scalar items, or presupposition triggers.
- **Here, we focus on the subset of ImpPres pertaining to quantifier-based scalar implicatures** (involving scalar items *some*, *all*, *no*, *not all*), because that is where Jeretic et al.'s NLI models were the most successful.

NLI, logical entailment, pragmatic entailment

- The Natural Language Inference task at stake in Jeretic et al. (2020) is s.t.:
  - Two sentences S_1 and S_2 are concatenated to form S_1+S_2.
  - S_1+S_2 is classified according to 3 possible labels: contradiction, neutral, entailment. Labels are intended to be s.t.:
    - If S_1 and S_2 are contradictory, S_1+S_2 is a **contradiction**;
    - If S_1 entails S_2 S_1+S_2 is an **entailment**;
    - Otherwise, S_1+S_2 is **neutral**.
- **Natural Language Inference might be expected to conflate logical and pragmatic entailment...**
  - A NLI model behaves "**logically**" if for instance it categorizes the (*some*, *all*) pair as neutral (because *some* does not entail nor contradict *all* from a purely logical point of view);
  - It behaves "**pragmatically**" if for instance it categorizes the (*some*, *all*) pair as contradictory (given the implicature *some* $\rightsquigarrow$ *not all*).
- The question Jeretic et al., 2020 tried to answer was then: do LLMs robustly behave logically, or pragmatically, across various pairs of sentences involving scalar items?

Pairs of items (quantifier subset of ImpPres)

- To create the quantifier subset of ImpPres, the pairs of scalar quantifiers below were inserted in 100 semi-automatically generated "frames".

Item 1	Item 2	Logically	Pragmatically
no	not all ($\rightsquigarrow$ some)	Entailment	Contradiction
no	some ($\rightsquigarrow$ not all)	Contradiction	
no	all	Contradiction	
not all	no	Neutral	Contradiction
not all	some ($\rightsquigarrow$ not all)	Neutral	Entailment
not all	all	Contradiction	
some	no	Contradiction	
some	not all ($\rightsquigarrow$ some)	Neutral	Entailment
some	all	Neutral	Contradiction
all	no	Contradiction	
all	not all ($\rightsquigarrow$ some)	Contradiction	
all	some ($\rightsquigarrow$ not all)	Entailment	Contradiction

Unpacking the non-trivial predictions

- (*no*, *not all*):
  - *no* logically entails *not all*; *not all* is compatible with *no* (=neutral).
  - pragmatically, *not all* $\rightsquigarrow$ *some* (reverse scalar implicature), and *some* contradicts *no*, so, pragmatically, *no* and *not all* are contradictory.
- (*not all*, *some*):
  - *not all* and *some* are logically compatible (=neutral).
  - pragmatically, *not all* $\rightsquigarrow$ *some* (reverse scalar implicature), and *some* $\rightsquigarrow$ *not all* (scalar implicature), so we expect entailment both ways.
- (*all*, *some*):
  - *all* logically entails *some*; *some* is compatible with *all* (=neutral).
  - pragmatically, *some* $\rightsquigarrow$ *not all* (scalar implicature), and *not all* contradicts *all*, so, pragmatically, *some* and *not all* are contradictory.

Some (but not all) sentences from ImpPres

(11) { No, Not all, Some, All } guys should practice.

(12) { No, Not all, Some, All } actresses were falling asleep.

(13) a. The Borgias boycott { no, some, all } college campuses.

b. The Borgias don't boycott all college campuses.

(14) a. Some rabbit might irritate { no, some, all } governments.

b. Some rabbit might not irritate all governments.

Next steps

- Test newer NLI models on IMPPRES, to see if they can behave pragmatically in at least a subset of the cases;
- Collect surprisals returned by the non-NLI variants of those models, on disjunctive sentences derived from the IMPPRES sentences (some being scalar HDs).
  - Test if scalar HDs are more surpising than their counterparts featuring incompatible items (*e.g.* *no* and *all*).
  - Test if surprisal is sensitive to linear order, especially in scalar HDs.
- Test if surprisal correlates with entailment or non-contradiction scores between disjuncts (as computed *via* NLI).

Performing Natural Language Inference on scalar pairs

Models tested

- We test newer (though not super recent) NLI models[5] on the ImpPres dataset.
- These models constitute improvements of the early Transformer model BERT (Devlin et al., 2018). BERT is a bidirectional encoder trained (mostly) on a Masked Language Modeling (MLM)[6] task.
  - **RoBERTa-Large** (Liu et al., 2019): like BERT, but purely MLM-trained, and with optimized hyperparameters.
  - **DeBERTa-Large** (He et al., 2020): builds on RoBERTa with disentangled attention[7] and enhanced mask decoder training.[8]
  - **BART-Large** (Lewis et al., 2019): an sequence-to-sequence model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. Supposedly overcomes the shortcomings of BERT and GPT.

[5] obtained by fine-tuning on MNLI; models available on HuggingFace. See why I did not "ask Chat-GPT" in te Appendix.

[6] i.e. trained at predicting a masked token, given the other tokens of the sentence and their positions.

[7] separate representations for the tokens, and their positions.

[8] token positions are re-incorporated at the final token prediction stage.

Methodology

- For each pair of sentences from the "quantifier" subset of the ImpPres dataset, we perform Natural Language Inference, i.e. we make our models classify the relation between the sentences of the pair as contradiction, neutral, or entailment.
- This classification step returns three confidence scores (one per label); **for each pair of sentences we take the predicted label to be the one with the highest score**.
- For each *kind* of pair (e.g. (*some*, *all*), (*no*, *not all*) etc), we count the number of predicted contradiction, neutral, and entailment labels across the sentences instantiating the pair. The total sums up to 100, because each condition involves 100 pairs of sentences.
- **A model will succeed on a certain kind of pair, if the proportion of labels consistent with either the "logical" or the "pragmatic" hypothesis is significantly above 50%.**[9]

[9] For a sample size of 100 and a confidence level of 95%, 61% is the lowest proportion for which the lower bound of the confidence interval is above 50%. 61% is thus our threshold for model success.

NLI results for RoBERTa across pairs

	item 1	item 2	contradiction	neutral	entailment
	no	not all ($\rightsquigarrow$ some)	32	41	27
	no	some ($\rightsquigarrow$ not all)	56	20	24
	no	all	52	27	21
☞	not all ($\rightsquigarrow$ some)	no	26	42	32
☞	not all ($\rightsquigarrow$ some)	some ($\rightsquigarrow$ not all)	37	25	38
	not all ($\rightsquigarrow$ some)	all	49	22	29
☞	some ($\rightsquigarrow$ not all)	no	80	19	1
☹	some ($\rightsquigarrow$ not all)	not all ($\rightsquigarrow$ some)	84	6	10
	some ($\rightsquigarrow$ not all)	all	36	6	58
	all	no	47	37	16
☞	all	not all ($\rightsquigarrow$ some)	85	9	6
☞	all	some ($\rightsquigarrow$ not all)	32	2	66

Table 1: RoBERTa-Large-MNLI. ☞ means model overall succeeds on the pair of items; ☹ means it fails. Orange cells are consistent with the model being "logical" but not "pragmatic"; blue cells are consistent with the model being "pragmatic" but not "logical"; green cells are consistent with the model being either logical or pragmatic.

Take away from RoBERTa

- Does not capture all cases of unambiguous contradictions; and even if it does, **no reciprocity**; e.g. (*all*, *not all*) is correctly categorized as a contradiction, but (*not all*, *all*) is not. Same for (*some*, *no*) vs. (*no*, *some*).
- In the other cases, where logical and pragmatic behavior are disentangled, the model succeeds via a **mix of both strategies**.
  - When *not all* is involved, the model tends to be a bit more logical; in the (*all*, *some*) pair, it appears more pragmatic. Might suggest **reverse scalar implicatures are harder** to get based on the available data.
  - Again, lack of reciprocity in the judgments.

NLI results for DeBERTa across pairs

	item 1	item 2	contradiction	neutral	entailment
	no	not all (⇝ some)	43	52	5
☹	no	some (⇝ not all)	5	83	12
	no	all	43	29	28
☞	not all (⇝ some)	no	63	36	1
☞	not all (⇝ some)	some (⇝ not all)	4	42	54
	not all (⇝ some)	all	40	16	44
☞	some (⇝ not all)	no	65	11	24
☞	some (⇝ not all)	not all (⇝ some)	18	8	74
☹	some (⇝ not all)	all	15	1	84
	all	no	59	6	35
☞	all	not all (⇝ some)	63	1	36
☞	all	some (⇝ not all)	3	0	97

Table 2: DeBERTa-Large-MNLI. ☞ means model overall succeeds on the pair of items; ☹ means it fails. Orange cells are consistent with the model being "logical" but not "pragmatic"; blue cells are consistent with the model being "pragmatic" but not "logical"; green cells are consistent with the model being either logical or pragmatic.

Take away from DeBERTa

- More success than RoBERTa: succeeds on the same pairs, plus (*some*, *not all*).
- Still quite bad with unambiguous contradictions: e.g. (*some*, *no*) is correctly classified, but (*no*, *some*) is overall classified as neutral (!)
- In the other cases, where logical and pragmatic behavior are disentangled, the model **sometimes succeeds via a mix of both strategies, sometimes appears purely logical, sometimes purely pragmatic**:
  - Mixed (though more logical) for the cases involving *not all* as first item;
  - Logical for (*some*, *not all*);
  - Pragmatic for (*all*, *some*).
- Again, the underrepresentation of pragmatic behavior in cases involving *not all* suggests **reverse scalar implicatures may be harder**.
- Again, **lack of reciprocity**, except for the pair (*some*, *not all*). But reciprocity is not that surprising in this case, because the expected patterns are the same!

NLI results for BART across pairs

	item 1	item 2	contradiction	neutral	entailment
☞	no	not all	66	34	0
☞	no	some (⇝ not all)	80	19	1
☹	no	all	37	62	1
☞	not all	no	89	11	0
☹	not all	some (⇝ not all)	80	20	0
	not all	all	56	44	0
☞	some (⇝ not all)	no	97	3	0
☹	some (⇝ not all)	not all	98	1	1
☞	some (⇝ not all)	all	73	17	10
☞	all	no	77	23	0
☞	all	not all	79	21	0
☞	all	some (⇝ not all)	56	25	19

Table 3: BART-Large-MNLI. ☞ means model overall succeeds on the pair of items; ☹ means it fails. Orange cells are consistent with the model being "logical" but not "pragmatic"; blue cells are consistent with the model being "pragmatic" but not "logical"; green cells are consistent with the model being either logical or pragmatic.

Take away from BART

- More success than DeBERTA, but also **more failures**:
  - Succeeds on 4 pairs where DeBERTa failed or was inconclusive: (*no*, *not all*/*some*), (*some*, *all*), (*all*, *no*).
  - Fails on 2 pairs where DeBERTa succeeded: (*not all*, *some*) and (*some*, *not all*). Fails on 1 pair where DeBERTa was inconclusive: (*no*, *all*).
- Does better overall with unambiguous contradictions and captures symmetry with (*no*, *some*). But (*no*, *all*) is only captured one-way.
- In the other cases, where logical and pragmatic behavior are disentangled, the model succeeds via a **mix of both strategies, being overall more logical**:
  - Fully logical with (*no*, *not all*);
  - More logical for (*not all*, *no*), (*some*, *all*), (*all*, *some*).
- Again, items involving *not all* lead to less pragmatic behavior, which suggests **reverse scalar implicatures may be harder**.
- **More reciprocity** than with previous models: (*no*, *not all*); (*no*, *some*), (*some*, *all*) as successfully categorized both ways

Summary across the 3 models

item 1	item 2	RoBERTa	DeBERTa	BART
no	not all			✓ (L>P)
no	some		✗	✓
no	all			✗
not all	no	✓ (P>L)	✓ (L>P)	✓ (L>P)
not all	some	✓ (L>P)	✓ (L>P)	✗
not all	all			
some	no	✓	✓	✓
some	not all	✗	✓ (L>P)	✗
some	all		✗	✓ (L>P)
all	no			
all	not all	✓	✓	✓
all	some	✓ (P>L)	✓ (P>L)	✓ (L>P)

- Surprising that the (*all*, *no*) cases were never labeled as contradictory.
- **When they do sensible things, models appear overall more "logical"**. Some evidence of pragmatic inferences in the (*all*, *some*) case and more marginally in the (*not all*, *no*) case.

Next steps

- We have just seen that the models at stake do not consistently draw scalar implicatures.
- The question is then: what do they do with scalar HDs?
- **Given that the models are not consistently pragmatic, we expect scalar HDs to give rise to "high" measures of surprisal – where surprisal is seen as a proxy for infelicity.**
- What "high" means should be fleshed out.
- If the models assign scalar HDs low surprisals anyway, this might be the sign that their judgment for such sentences rely on superficial, frequency-based cues, rather than on some "deep" reasoning about the logical relation between disjuncts.
- This hypothesis will be fleshed out by checking if the models exhibit an asymmetry between weak-to-strong (felicitous) scalar HDs vs. strong-to-weak (degraded) scalar HDs.

Assessing the (in)felicity of scalar HDs and other scalar disjunctions

Evaluating (in)felicity with LLMs

- **Surprisal** (negative log probability of a word/token/sentence) was shown to correlate with **processing effort** (Hale, 2001; Levy, 2008).
- "Processing effort" may collapse grammaticality and felicity; still, **we may expect surprisal in grammatical sentences to reflect infelicity**.

$$\begin{aligned}\textsc{Infelicity}(w_t) &\simeq \textsc{Surprisal}(w_t) \\ &= -\log P(w_t|w_1 \dots w_{t-1})^3\end{aligned}$$

$$\textsc{Infelicity}(w_1 \dots w_t) \simeq \sum_{i=1}^{t} \textsc{Surprisal}(w_i)$$

- Surprisal is not super informative as an absolute value: different sentences exhibiting the same degree of grammaticality and felicity may significantly differ in terms of surprisal, just because they use different lexical items.
- **Surprisal is informative as a relative measure**: surprisal *contrasts* between sentences forming a minimal pair tend to be informative.

[3] In the case of BERT-like bidirectional models, this formula is adapted to MLM: the probability of a word is computed given its left *and* right context.

Modeling the prediction

- Forgetting for now about the ordering of the disjuncts, scalar HDs are expected to lead to overall "high" levels of surprisal. But what should be the baseline?
- An natural idea is to compare scalar HDs to minimal variants where one item is changed and makes the resulting two disjuncts incompatible.
  - For instance, if the HD features the pair (*some*, *all*), one could compare it to a disjunction where *some* is changed into *no*, because (*no*, *all*) is contradictory.

    (15) a. Al ate some or all of the biscuits.

    b. Al ate no or all of the biscuits.

- Problem: **a disjunction can be infelicitous in various ways!** For instance *some or no* features incompatible disjuncts but is tautological, and so sounds quite odd too. Therefore comparing e.g. *some or all* to the minimal variant *some or no* does not yield any clear prediction.

Scalar HDs and baselines

- When choosing the baseline(s) we consider two dimensions: the degree of compatibility between disjuncts, and the sentence's informativity.

Disjuncts (either order)	"Logical" relation between disjuncts	"Pragmatic" relation between disjuncts	Logical informativity	Pragmatic informativity
no/some	C	C	T	NT
no/all	C	C	NT	NT
not all/all	C	C	T	NT
no/not all	E	C/E[10]	NT	NT
not all/some	N	EE	T	NT
some/all	E	C/E	NT	NT

Table 4: Summary of the logical vs. pragmatic relations between disjuncts featuring scalar items *no*, *not all*, *some*, *all*; and of the overall informativity of the resulting disjunctions. C=contradictory disjuncts; E=entailing; EE=equivalent; N=compatible. T=tautological; NT=non-tautological.

[10] Depending on linear order.

Scalar HDs and baselines

Disjuncts (either order)	"Logical" relation between disjuncts	"Pragmatic" relation between disjuncts	Logical informativity	Pragmatic informativity
no/some	C	C	T	NT
no/all	C	C	NT	NT
not all/all	C	C	T	NT
no/not all	E	C/E	NT	NT
not all/some	N	EE	T	NT
some/all	E	C/E	NT	NT

Table 4: Summary of the logical vs. pragmatic relations between disjuncts featuring scalar items *no*, *not all*, *some*, *all*; and of the overall informativity of the resulting disjunctions. C=contradictory disjuncts; E=entailing; EE=equivalent; N=compatible. T=tautological; NT=non-tautological.

- The first 3 rows of the above table describe disjunctions that are not HDs, regardless of whether we reason logically or pragmatically. Among them, only the *no or all* disjunction is informative at both the logical *and* the pragmatic level: that would be the best baseline.

Scalar HDs and baselines

Disjuncts (either order)	"Logical" relation between disjuncts	"Pragmatic" relation between disjuncts	Logical informativity	Pragmatic informativity
no/some	C	C	T	NT
no/all	C	C	NT	NT
not all/all	C	C	T	NT
no/not all	E	C/E	NT	NT
not all/some	N	EE	T	NT
some/all	E	C/E	NT	NT

Table 4: Summary of the logical vs. pragmatic relations between disjuncts featuring scalar items *no*, *not all*, *some*, *all*; and of the overall informativity of the resulting disjunctions. C=contradictory disjuncts; E=entailing; EE=equivalent; N=compatible. T=tautological; NT=non-tautological.

- The last 3 rows of the above table describe HDs. *No or not all* and *some or all* can be made non-HD *via* pragmatics. Both disjunctions are informative, at both the logical and pragmatic level.
- *Not all or some* cannot be rescued *via* pragmatics (the disjuncts remain compatible – in fact, they become equivalent). But pragmatics makes the disjunction informative by excluding *all* and *none*.

Prediction

- We use *no or all* as a baseline.[11] **It should be less surprising than the other non-HDs, which are tautological**.
- **It should also be less surprising than HDs**, which are infelicitous at the logical level, and sometimes also at the pragmatic level. Moreover:
  - If the model is "logical", we expect all HDs to pattern similarly.
  - If the model is "pragmatic", we expect the "rescuable" *no or not all* and *some or all* to be better than *not all or some*.

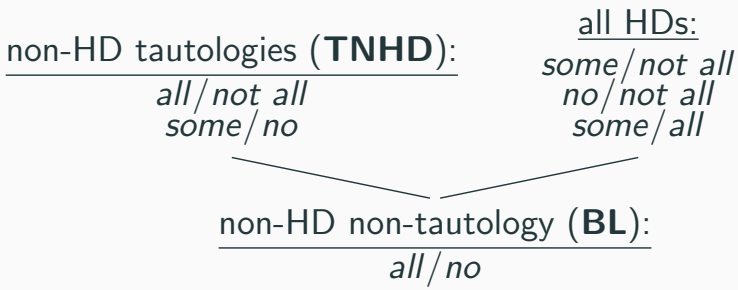


Figure 1: Logical view

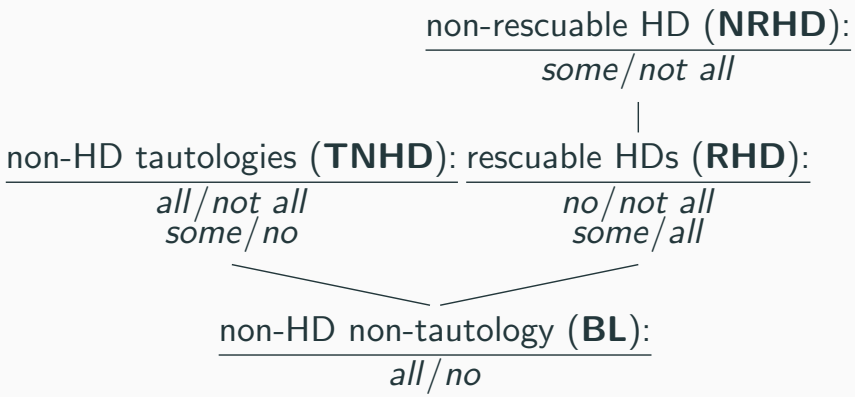


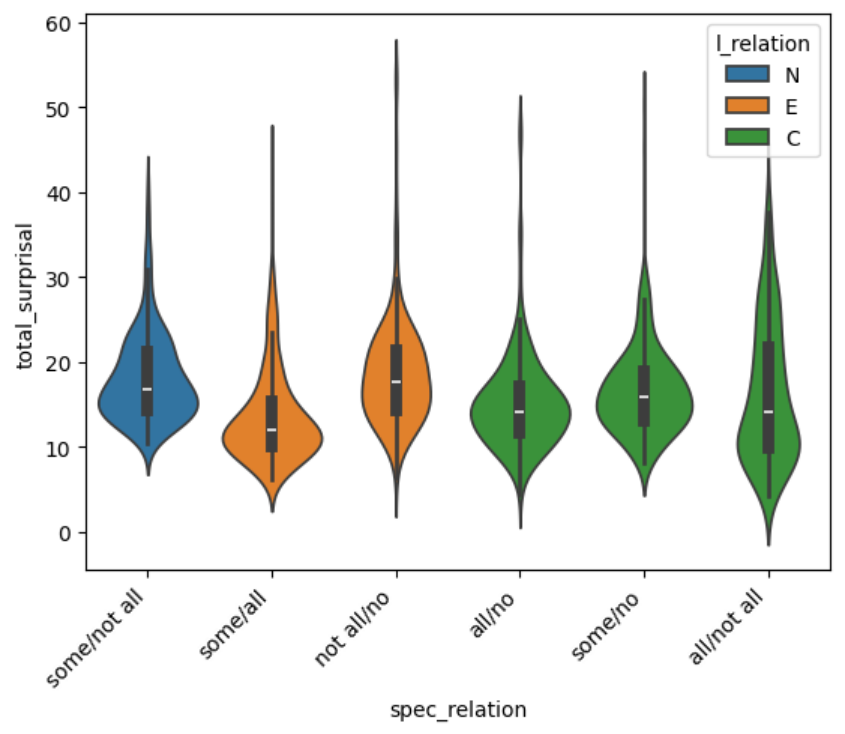
Figure 2: Pragmatic view

[11] This is not perfect w.r.t. the *not all or some* HD, because *no or all* does not minimally differ from it. But it's the best option given that all minimally differing non-HDs (*some or no*, *all or not all*) are logically tautological.

Methodology

- General prediction: **HDs and non-HD tautologies should be more surprising than the baseline**.
- Pragmatics-specific prediction: **the non-rescuable HD some or not all should be more surprising than the other, rescuable HDs**.
- For each type of "degraded" disjunction (non-HD tautology, rescuable HD, non-rescuable HD), we go over each sentence S of that type:
  - We compute its surprisal $\textsc{Surprisal}(S)$, using the Python `minicons` library (Misra, 2022): ;
  - We find its baseline counterparts S' and S'' of the form *all or no*, or *no or all*.
  - We compute $\textsc{Surprisal}(S)$-$\textsc{Surprisal}(S')$ and $\textsc{Surprisal}(S)$-$\textsc{Surprisal}(S'')$ and add those scores to our set of differential scores.
  - Once all sentences have been scanned, we test if the differential scores are significantly above 0 (one-tailed Wilcoxon test for matched pairs).

RoBERTa-Large: testing the general prediction

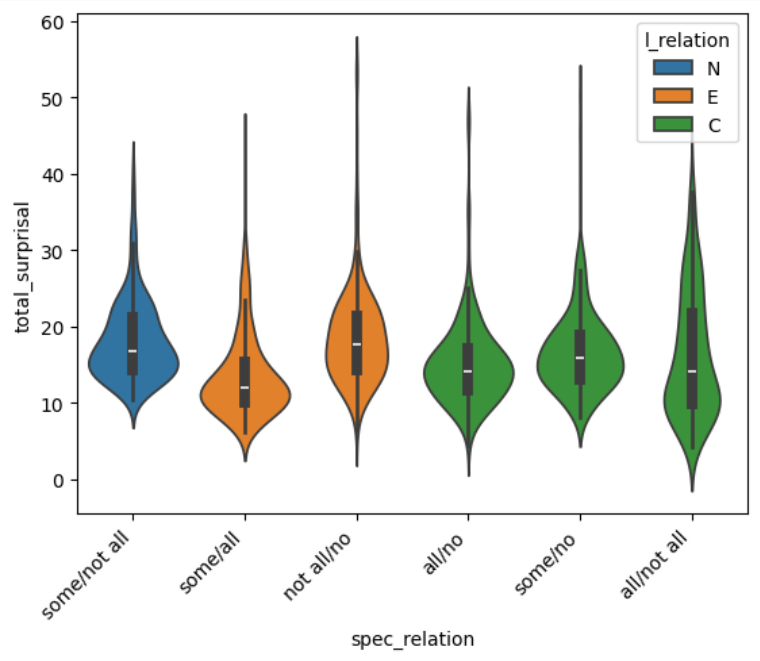


Degraded pair	Baseline pair	Comparison type	Holm-Sidak p-value	Greater surprisal?
not all, some	all, no	NRHD-BL	1.24e-24	True
all, some	all, no	RHD-BL	1.00e+00	False
not all, no	all, no	RHD-BL	3.65e-32	True
no, some	all, no	TNHD-BL	1.30e-13	True
all, not all	all, no	TNHD-BL	1.57e-01	False

RoBERTa-Large: comments on the general prediction

- The non-rescuable HDs based on the pair (*not all*, *some*) are significantly more surprising than the baseline.
- The rescuable HDs exhibit a mixed pattern:
  - Those based on the (*not all*, *no*) pair appear more surprising than the baseline, but those based on the (*some*, *all*) pair do not. Might suggest the former kind is less "rescuable" than the later kind – or maybe, it's just less frequent.
  - Somewhat consistent with the observation that RoBERTa was better with direct scalar implicatures (*some* $\rightsquigarrow$ *not all*) as opposed to reverse scalar implicatures (*not all* $\rightsquigarrow$ *some*) when performing NLI.
- Tautological non-HDs also exhibit a mixed pattern: those based on the pair (*no*, *some*) appear more surprising than the baseline, while those based on the (*all*, *not all*) pairs are not.

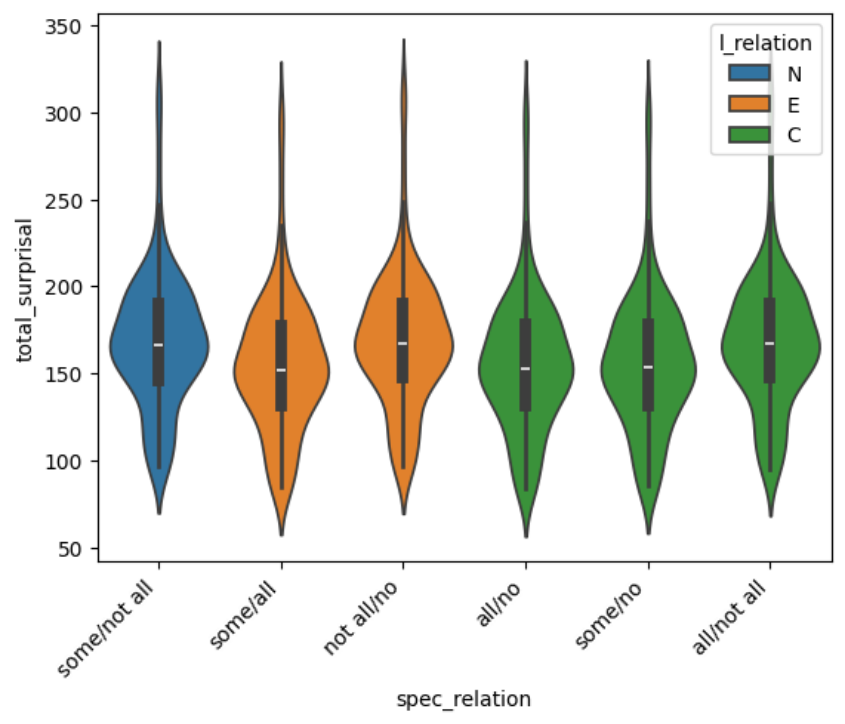
RoBERTa-Large: testing the pragmatic prediction



- The non-rescuable HD *not all or some* is expectedly more surprising than *some or all*, which was already less surprising than the baseline.
- But it is not more surprising than the other rescuable HD *not all or no*. Consistent with the idea RoBERTa does not rescue *not all or no*.

Degraded pair	Baseline pair	Comparison type	Holm-Sidak p-value	Greater surprisal?
not all, some	all, some	NRHD-RHD	1.55e-27	True
not all, some	not all, no	NRHD-RHD	8.77e-01	False

DeBERTa-Large: testing the general prediction

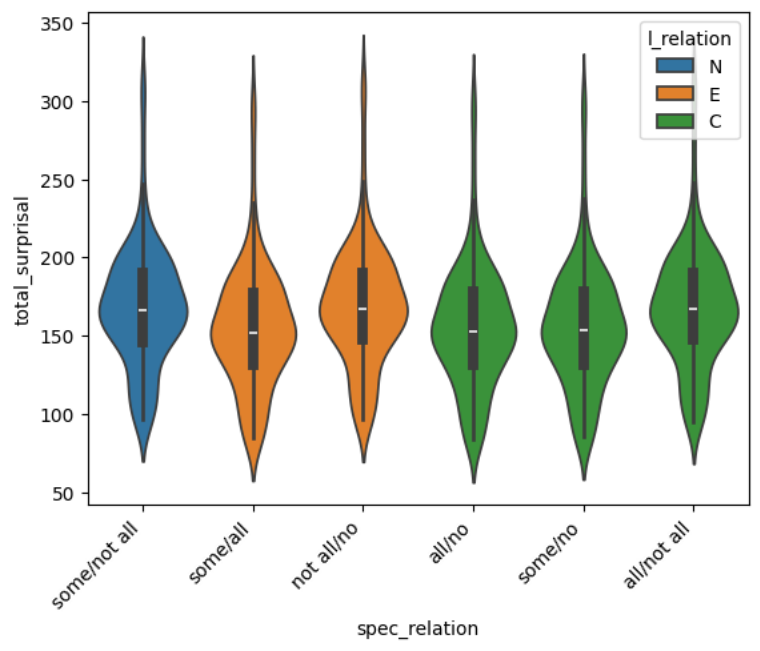


Degraded pair	Baseline pair	Comparison type	Holm-Sidak p-value	greater surprisal?
not all, some	all, no	NRHD-BL	6.83e-67	True
all, some	all, no	RHD-BL	1.00e+00	False
not all, no	all, no	RHD-BL	6.83e-67	True
no, some	all, no	TNHD-BL	2.25e-10	True
all, not all	all, no	TNHD-BL	6.83e-67	True

DeBERTa-Large: comments on the general prediction

- The non-rescuable HDs based on the pair (*not all*, *some*) are significantly more surprising than the baseline – same as with RoBERTa.
- The rescuable HDs exhibit the same mixed pattern as with RoBERTa.
- Tautological non-HDs appear consistently more surprising than the baseline.

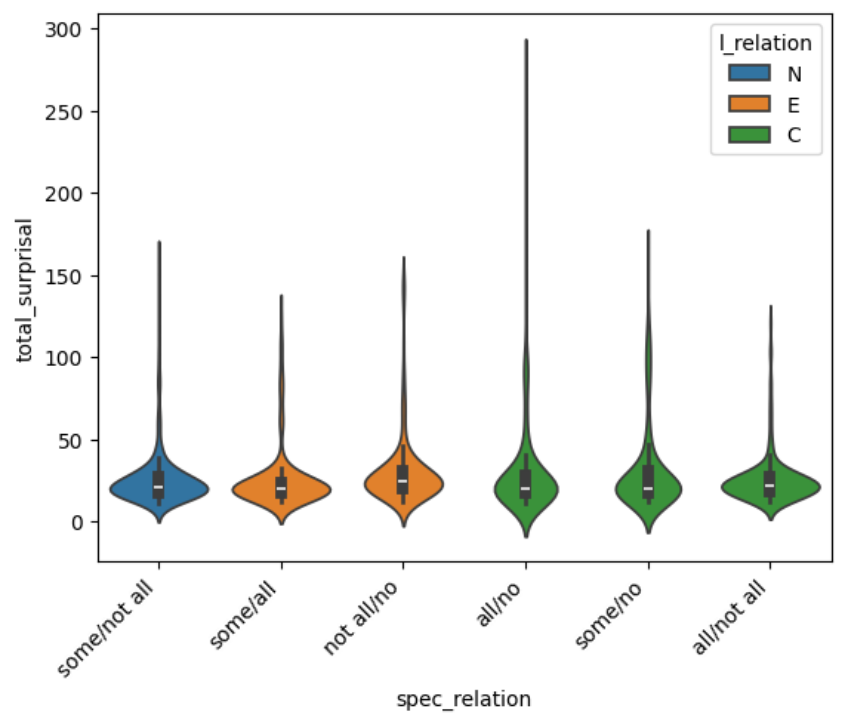
DeBERTa-Large: testing the pragmatic prediction



- Same pattern and comments as with RoBERTa: results seem to suggest a failure at pragmatics for HDs involving *not all*.

Degraded pair	Baseline pair	Comparison type	Holm-Sidak p-value	Greater surprisal?
not all, some	all, some	NRHD-RHD	1.44e-34	True
not all, some	not all, no	NRHD-RHD	1.00e+00	False

BART-Large: testing the general prediction



Degraded pair	Baseline pair	Comparison type	Holm-Sidak p-value	Greater surprisal?
not all, some	all, no	NRHD-BL	9.03e-01	False
all, some	all, no	RHD-BL	9.30e-01	False
not all, no	all, no	RHD-BL	3.58e-11	True
no, some	all, no	TNHD-BL	1.80e-02	True
all, not all	all, no	TNHD-BL	7.32e-01	False

BART-Large: comments on the general prediction

- The non-rescuable HDs based on the pair (*not all*, *some*) is not more surprising than the baseline. Unexpected given that this disjunction is degraded from both a logical point of view and a pragmatic point of view!
- The rescuable HDs exhibit the same mixed pattern as with RoBERTa and DeBERTa.
- Tautological non-HDs appear do not consistently more surprising than the baseline...
- **Overall poor results which clash with BART's ability to perform NLI**... (I'm wondering if I messed up something with sentence scoring for this model).
- We don't test the pragmatic prediction given that the non-rescuable HD *not all or some* did not appear more surprising than the baseline to start with.

Next steps

- So far we have not considered the order between disjuncts as relevant; i.e. we grouped together the surprisals associated with disjunctions featuring the same kinds of disjuncts, in either order. We assumed the infelicitous orders, if they exist, would drive the surprisal contrasts.
- So maybe the weakness of the surprisal contrasts measured for all 3 models is due to the fact that we did not care about order; **in the next section we investigate ordering asymmetries in scalar disjunctions, particularly scalar HDs**.

Testing for asymmetries in scalar HDs

Expected Asymmetries

- We now turn to the effect or order. As shown in the table below, **no or not all and some or all are the two kinds of HDs whose felicity varies under the pragmatic view**:
  - *some or all* and *not all or no* should end up having contradictory (C) disjuncts, because the weaker item appears first and thus can be appropriately strengthened.
  - *all or some* and *no or not all* should end up having entailing (E) disjuncts, because the weaker item appears second and thus cannot be strengthened.

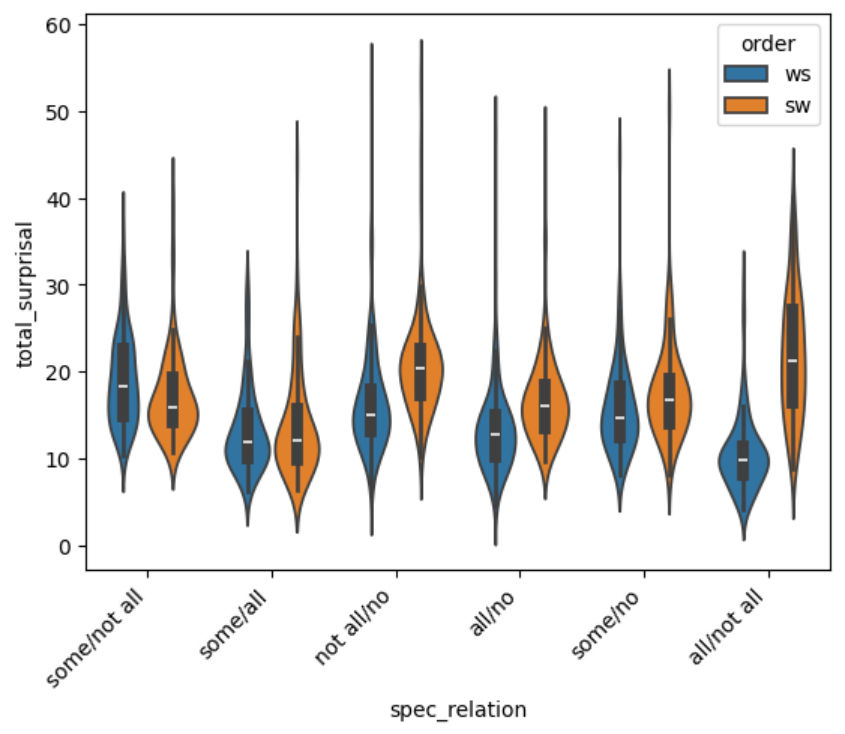
Disjuncts (either order)	"Logical" relation between disjuncts	"Pragmatic" relation between disjuncts	Logical informativity	Pragmatic informativity
no/some	C	C	T	NT
no/all	C	C	NT	NT
not all/all	C	C	T	NT
no/not all	E	C/E	NT	NT
not all/some	N	EE	T	NT
some/all	E	C/E	NT	NT

Table 4: Summary of the logical vs. pragmatic relations between disjuncts with scalar items; and of the overall informativity of the resulting disjunctions.

Prediction

- The ordering of items in conjunctions has been shown to be influenced by many factors (semantic, metrical, frequency-related) (Benor & Levy, 2006). We may expect such factors to be at play in disjunctions, too – in particular non-HDs.
- For that reason, we do not make any prediction regarding non-HDs and the non-rescuable HD *some or not all*. It's likely they'll exhibit order-based asymmetries, but we do not make any assumption regarding their directionality.
- **However, we expect a pragmatics-driven effect of linear order in the case of the rescuable HDs some or all and not all or no.**
- For each possible pair of scalar items, we compared the paired differences of surprisal between both orderings (weak-to-strong vs. strong-to-weak) of each disjunction featuring the items.
- We use two-sided Wilcoxon tests, given that some of the predictions are non-directional.

RoBERTa-Large: weak-to-strong vs. strong-to-weak

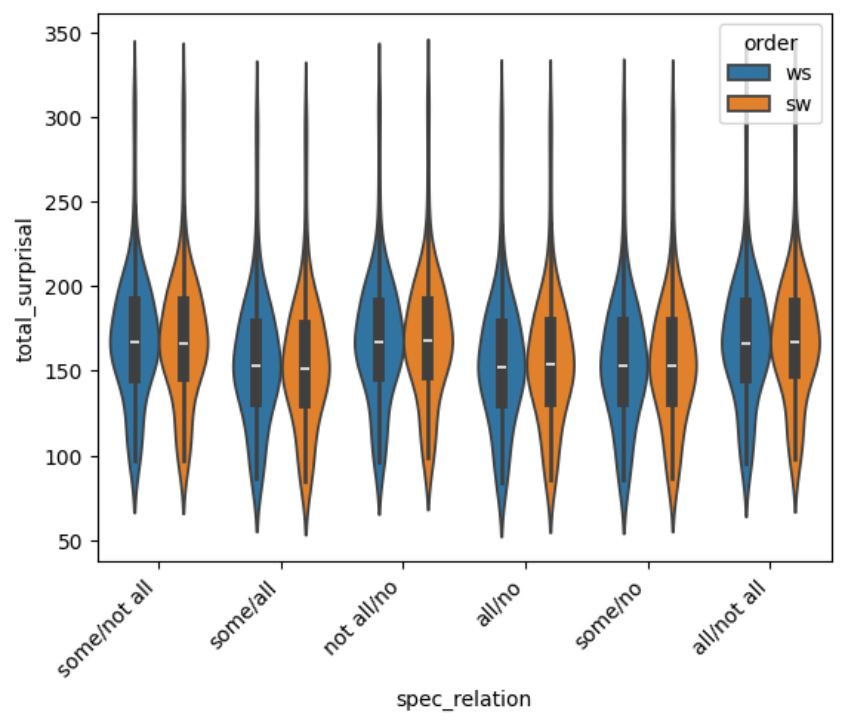


Item 1 ("weak")	Item 2 ("strong")	Pair type	Holm-Sidak p-value	Preferred order
some	all	RHD	1.54e-01	None
not all	no	RHD	3.36e-16	1-2
some	not all	NRHD	6.45e-04	2-1
all	no	BL	7.62e-16	1-2
some	no	TNHD	8.60e-10	1-2
all	not all	TNHD	5.76e-17	1-2

Take away from RoBERTa-Large

- Surprisingly perhaps, **no ordering effect on the some or all pair**...maybe consistent with the fact that RoBERTa did not fully capture the *some* $\rightsquigarrow$ *not all* implicature in the NLI task?
- **An ordering effect is present in the right direction for the not all or no pair**...
  - It's unlikely that this contrast is explained by other independent factors, since usually short, unmarked elements tend to go first, so under that view *no* should probably go first. Plus, it does not seem that there is a systematic bias for having *not all* first; in the non-target disjunctions (NRHD, TNHD), *not all* is not consistently preferred as a first element
  - Still, the result is a bit surprising given that the ordering effect in *some or all* (intuitively easier) was not captured. Also surprising given that RoBERTa did not fully capture the relevant *not all* $\rightsquigarrow$ *some* implicature in the NLI task.

DeBERTa-Large: weak-to-strong vs. strong-to-weak

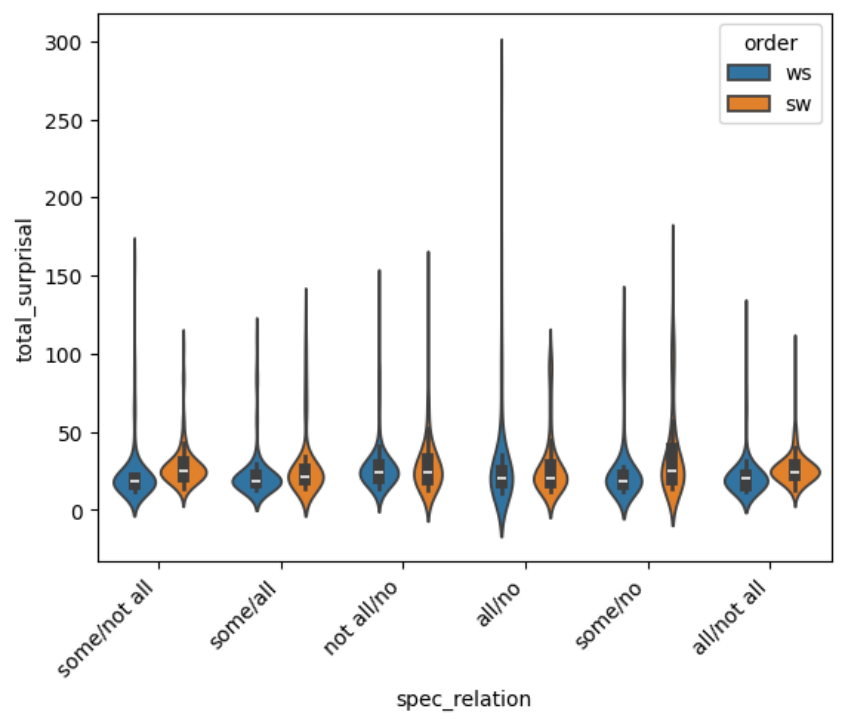


Item 1 ("weak")	Item 2 ("strong")	Pair type	Holm-Sidak p-value	Preferred order
some	all	RHD	6.39e-12	2-1
not all	no	RHD	3.80e-14	1-2
some	not all	NRHD	6.79e-02	None
all	no	BL	3.07e-17	1-2
some	no	TNHD	4.15e-11	1-2
all	not all	TNHD	4.15e-11	1-2

Take away from DeBERTa-Large

- **Unexpected effect of order on the some or all pair**...strange given that DeBERTA-NLI behaved pragmatically for the (*all*, *some*) pair – although not for the (*some*, *all*) pair (categorized as entailment).
- **An ordering effect is present in the right direction for the not all or no pair**...Again quite surprising given that DeBERTa-NLI was behaving rather logically with the (*all*, *not all*) and (*not all*, *all*) pairs – i.e. did not draw the *not all* $\rightsquigarrow$ *some* reverse scalar implicature.
- Preferred order for the non-target items are overall consistent with the previous model, RoBERTa.

BART-Large: weak-to-strong vs. strong-to-weak



Item 1 ("weak")	Item 2 ("strong")	Pair type	Holm-Sidak p-value	Preferred order
some	all	RHD	5.97e-10	1-2
not all	no	RHD	8.82e-01	None
some	not all	NRHD	5.97e-10	1-2
all	no	BL	5.27e-01	None
some	no	TNHD	4.87e-10	1-2
all	not all	TNHD	1.77e-07	1-2

Take away from BART-Large

- **An ordering effect is present in the right direction for the some or all pair**...strange given that BART-NLI behaved more logically for both the (*all*, *some*) and the (*some*, *all*) pair – i.e. did not consistently draw the *some* $\rightsquigarrow$ *not all* implicature.
- **No ordering effect for the not all or no pair**... consistent with BART-NLI behaving logically on the relevant pairs of items.
- Preferred orders for the non-target items are overall consistent with the 2 previous models, RoBERTa and DeBERTa (except for *some or not all*).

Interim conclusion and next step

- The effect of order in pragmatically "rescuable" scalar HDs of the form *some or all*/??*all or some*, and *not all or no*/?*no or not all*, was not clearly reflected by surprisal contrasts.
- If anything, it was clearer for the *not all or no*/?*no or not all* sentences, which is surprising given that this contrast is anything but crisp intuitively, and that the NLI tasks revealed models were doing quite bad overall with reverse scalar implicatures.
- This may suggest different possible things:
  - The theory is wrong?
  - The models mainly rely on non-pragmatic cues to judge scalar HDs.
  - The models are consistent, but on a sentence-by-sentence basis; maybe looking at aggregated NLI performance and aggregated surprisal measures, is not the way to go?
- In the next and last section, **we investigate, on a sentence-by-sentence basis, if NLI performance correlates with surprisal measures**.

Assessing the correlation between NLI and surprisal scores at the sentence level

Taking stock

- We have seen our 3 LLMs are **far from being consistently pragmatic**.
- We have also seen that scalar HDs are not consistently judged as more surprising than an informative, non-HD baseline.
- But can we check if the models are at least internally consistent regarding Hurford violations?
- Meaning, does the strength of the entailment[12] relation between disjuncts correlate with surprisal, on a sentence-by-sentence basis?

[12] scores measuring mere logical compatiblity (i.e. non-contradiction) gave worse results when plotted against surprisal for DeBERTa, and better results for BART. Plots can be found in the Appendix.

Methodology and prediction

- For each disjunction in our dataset, **we compute a "bleached" surprisal measure neutralizing the effect of its frame** (i.e. the lexical items present in the sentence that are different from *or* and the 2 scalar items appearing in it).
  - This is done to avoid any extra noise when correlating surprisal measures for the disjunctions with entailment scores.
  - To do this, we compute the differential surprisal between the surprisal of the whole disjunction, and those of each disjunct, and use the average of these 2 differences as our bleached surprisal measure.

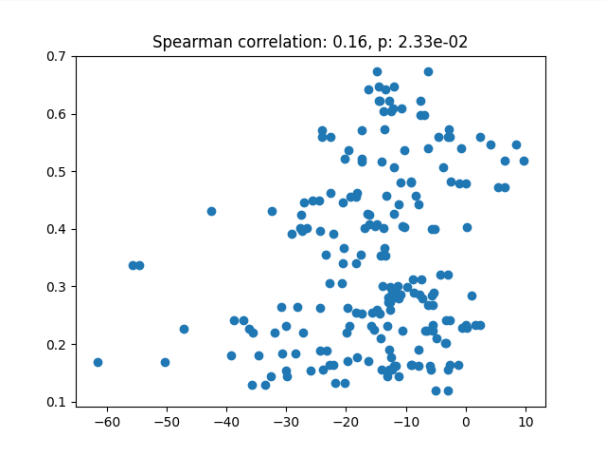
$$\mathrm{S}_{\textsc{Corrected}}(D_1 \vee D_2) = \mathrm{S}(D_1 \vee D_2) - \frac{\mathrm{S}(D_1) + \mathrm{S}(D_2)}{2}$$

- For each disjunction, **we compute the entailment scores between the 2 disjuncts, in either direction**. We take the maximum of these 2 scores to be the "entailment score" of the disjunction.

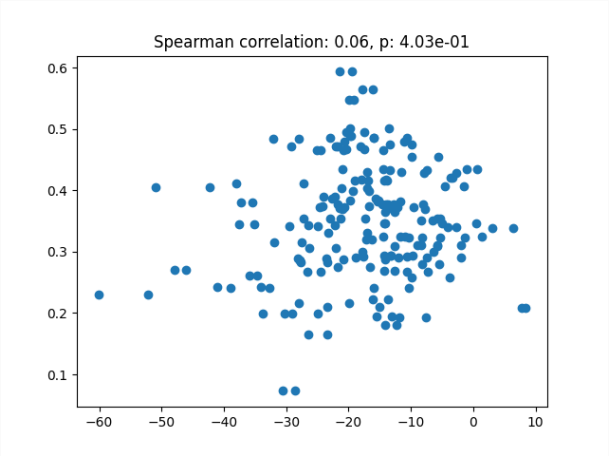
$$\mathrm{E}(D_1 \vee D_2) = \textsc{Max}(\mathbb{P}_{NLI}(D_1 \to D_2), \mathbb{P}_{NLI}(D_2 \to D_1))$$

- E should correlate with a Hurford violation, i.e., with $\mathrm{S}_{\textsc{Corrected}}$

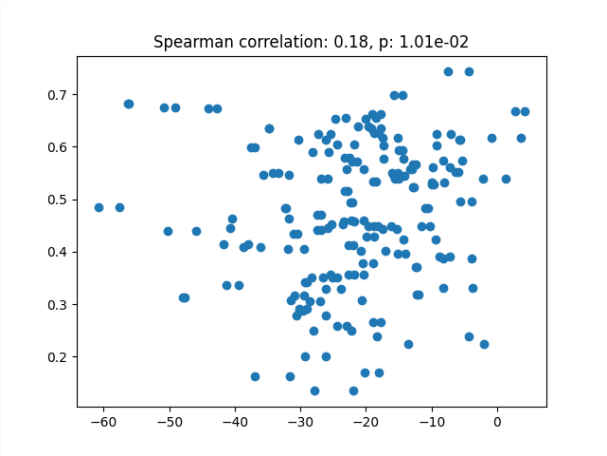
RoBERTa-Large, max entailment score



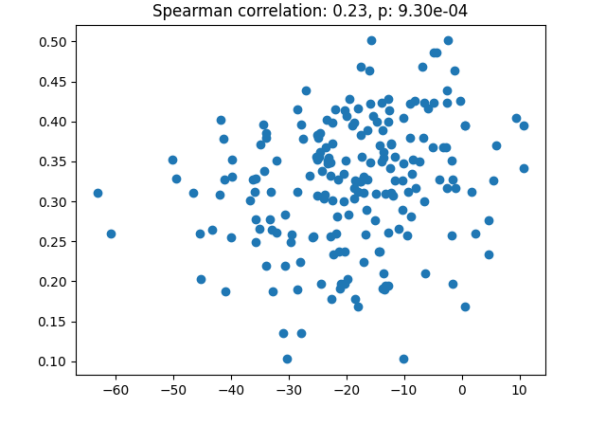
(a) NRHD – (*some*, *not all*)*



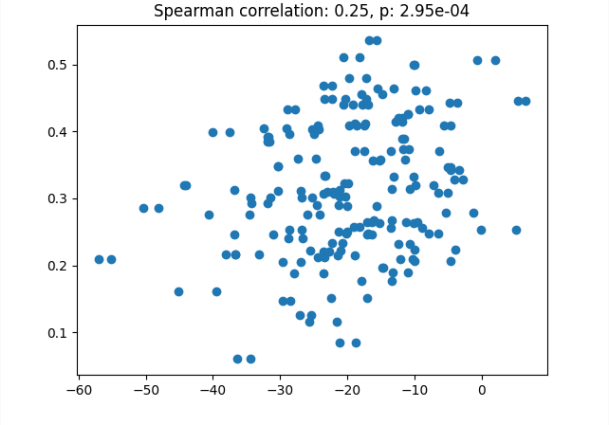
(b) RHD – (*not all*, *no*)



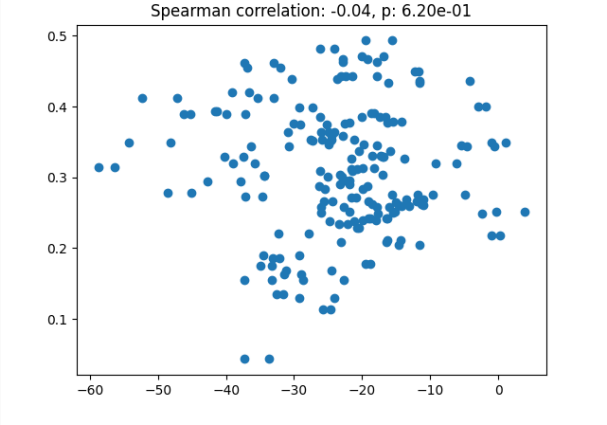
(c) RHD – (*some*, *all*)*



(d) TNHD – (*all*, *not all*)*



(e) TNHD – (*some*, *no*)*

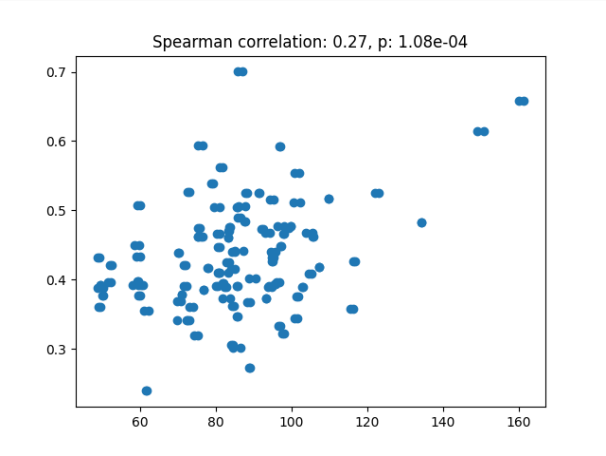


(f) BL – (*all*, *no*)

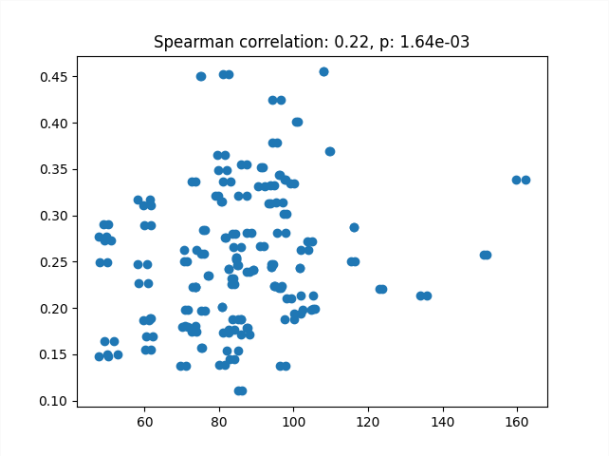
Take away from RoBERTa

- Correlation significant for 4/6 pairs, overall quite weak (~ 0.20)
- No correlation found for our non-tautological, non-HD "baseline", *all or no*...
- No correlation found for the rescuable HD *not all or no*.

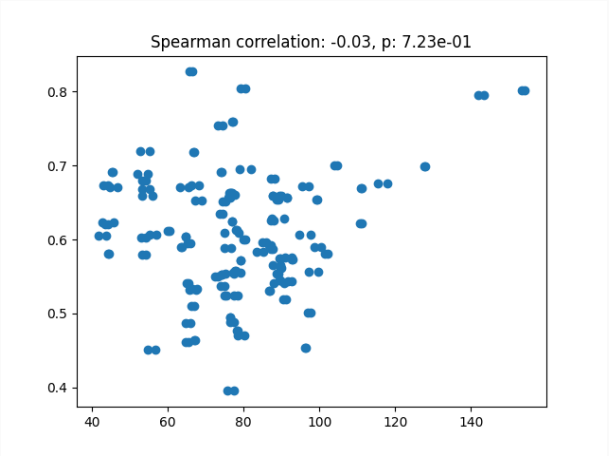
DeBERTa-Large, max entailment score



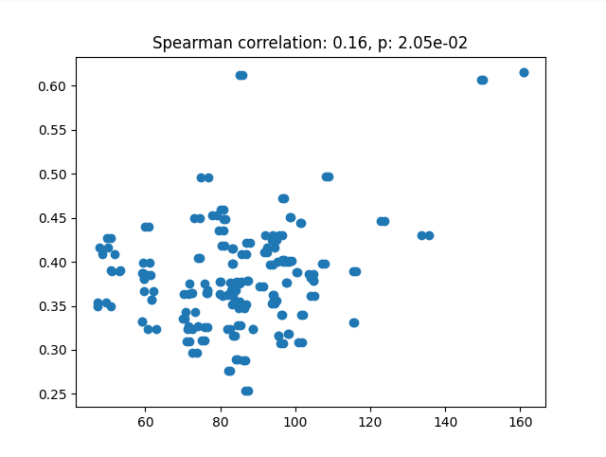
(a) NRHD – (*some*, *not all*)*



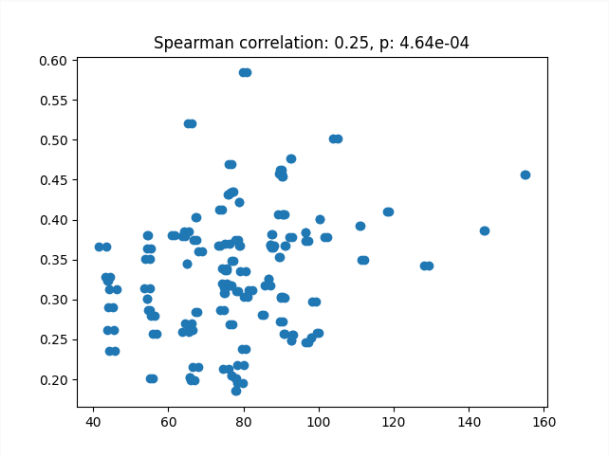
(b) RHD – (*not all*, *no*)*



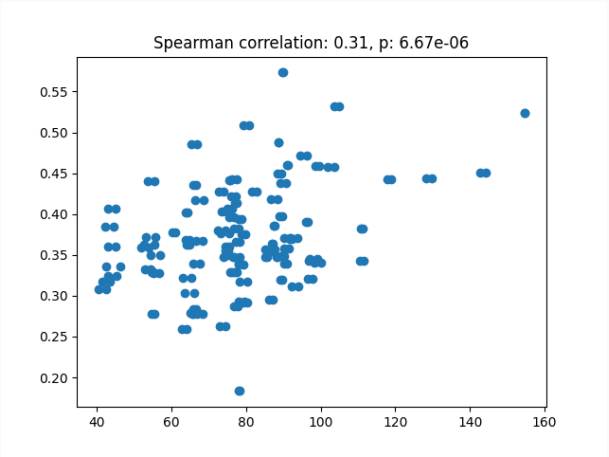
(c) RHD – (*some*, *all*)



(d) TNHD – (*all*, *not all*)*



(e) TNHD – (*some*, *no*)*

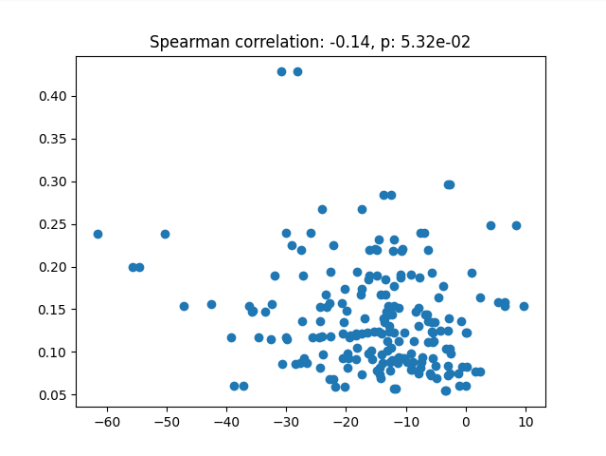


(f) BL – (*all*, *no*)*

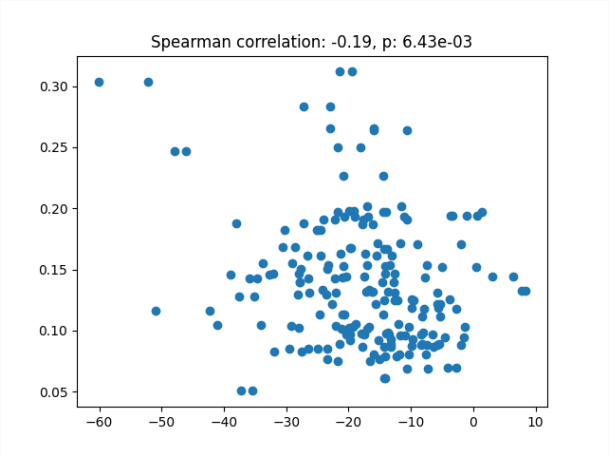
Take away from RoBERTa

- Correlation significant for 5/6 pairs, overall quite weak ($\sim 0.16 - 0.30$)
- No correlation found for the rescuable HD *some or all*. Looking at the plot, it seems bimodal: some datapoint seem to follow a positive correlation trend, while some others seem to follow a negative correlation trend. Separating weak-to-strong vs. strong-to-weak disjunctions might help clarify what is going on here.

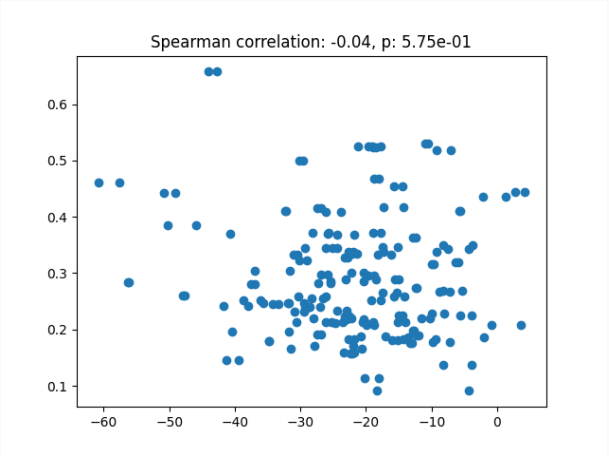
BART-Large, max entailment score



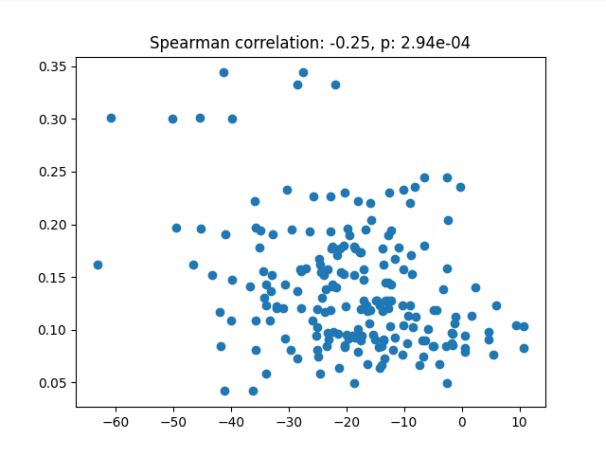
(a) NRHD – (*some*, *not all*)



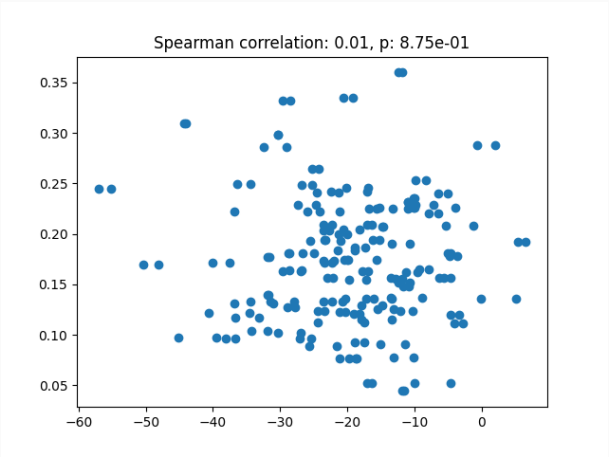
(b) RHD – (*not all*, *no*)*



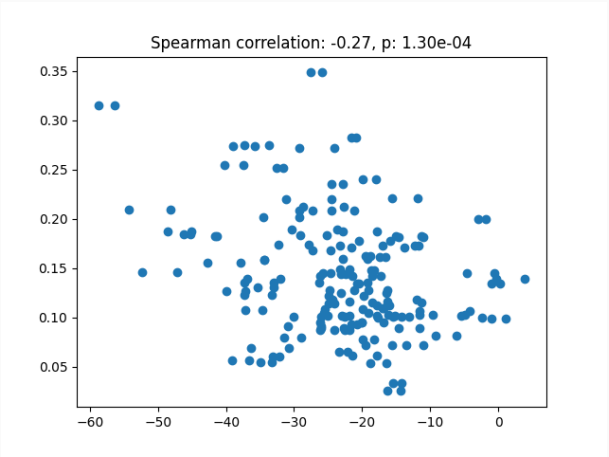
(c) RHD – (*some*, *all*)



(d) TNHD – (*all*, *not all*)*



(e) TNHD – (*some*, *no*)



(f) BL – (*all*, *no*)*

Take away from BART

- That's a mess. It makes me wonder if I messed up the surprisal scoring part with this model (?) However, looking at maximum *compatibility* scores (instead of entailment score), yields slightly better results for BART.
- Overall surprising, because BART seemed to do sensible stuff during the NLI task.

Conclusion: let's start with the bad stuff

- We investigated 3 LLMs on a subset of the ImpPres dataset focusing on scalar quantifiers *no*, *not all*, *some*, and *all*, and on disjunctions formed out of this dataset.
- We showed that LLMs tested on Natural Language Inference were not very consistent and typically used a **mix of logical and pragmatic strategies** to classify pairs of sentences – logical behaviors being more prominent. In particular, the models were **not doing so well on clear-cut cases of contradiction, and on cases involving reverse scalar implicatures** of the form *not all* $\rightsquigarrow$ *some*.
- We showed that the same LLMs assessed on disjunctions *via* surprisal measurements **were not consistently treating scalar HDs as degraded**, despite the fact that NLI showed they were on average not behaving very pragmatically. We also showed that **ordering asymmetries predicted by pragmatic theory were not robustly captured**, especially when it came to the less-subtle one *some or all* vs. *all or some*.

Conclusion: the not so bad stuff

- Still, we showed in the last section that when focusing on disjunct-entailment/disjunction-surprisal correlations on a sentence-by-sentence basis, **some LLMs showed some degree of internal consistency**: the higher the entailment score between (unordered) disjuncts, the higher the corrected surprisal.

To explore

- Regarding the quantifier dataset:
  - Singularity of the **not all-sentences**: they are the only ones which sometimes exhibit negation on the main verb, which makes them structurally distinct from the other sentences. Is that a problem?
  - Effect of **ellipsis** in HDs: here we only tested disjunctions of full sentences – what about disjunctions with a higher degree of ellipsis?
  - **Context**: what happens in terms of surprisal if we introduce a specific context before the target disjunctions (e.g., questions making alternatives salient?)
- Regarding other datasets, earlier models were shown to hardly draw any implicature on **other scalar pairs**. Do newer models do better? How do they judge HDs featuring such items?
- Regarding the analyses:
  - **Word-level analysis** of surprisal measures: where exactly in the sentence do the models get confused?
  - Effect of **disjunct ordering w.r.t. the correlation** between disjunct-entailment and disjunction-surprisal .

Thank you very much for your attention !

Selected references i

Hurford, J. R. (1974).**Exclusive or inclusive disjunction.** *Foundations of language*, *11*(3), 409–411.

Gazdar, G. (1979).**Pragmatics, implicature, presuposition and lógical form.** *Critica*, *12*(35), 113–122.

Hale, J. (2001).**A probabilistic earley parser as a psycholinguistic model.** *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*. https://doi.org/10.3115/1073336.1073357

Sauerland, U. (2004).**Scalar implicatures in complex sentences.** *Linguistics and Philosophy*, *27*(3), 367–391. https://doi.org/10.1023/b:ling.0000023378.71748.db

Benor, S., & Levy, R. (2006).**The chicken or the egg? a probabilistic analysis of english binomials.** *Language*, *82*(2), 233–278. https://doi.org/10.1353/lan.2006.0077

Levy, R. (2008).**Expectation-based syntactic comprehension.** *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Singh, R. (2008a). **Modularity and locality in interpretation [Doctoral dissertation, MIT].**

Selected references ii

Singh, R. (2008b).**On the interpretation of disjunction: Asymmetric, incremental, and eager for inconsistency.** *Linguistics and Philosophy*, *31*(2), 245–260. https://doi.org/10.1007/s10988-008-9038-x

Spector, B., Fox, D., & Chierchia, G. (2008).**Hurford's Constraint and the Theory of Scalar Implicatures.** *Manuscript, MIT and Harvard*.

Chierchia, G., Fox, D., & Spector, B. (2012). **Scalar implicature as a grammatical phenomenon.** In K. von Heusinger, C. Maienborn, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning*. de Gruyter.

Cremers, A., & Chemla, E. (2014). **Direct and indirect scalar implicatures share the same processing signature.** In *Pragmatics, semantics and the case of scalar implicatures* (pp. 201–227). Palgrave Macmillan UK. https://doi.org/10.1057/9781137333285_8

Katzir, R., & Singh, R. (2014).**Hurford disjunctions: Embedded exhaustification and structural economy.** *Proceedings of Sinn und Bedeutung*, *18*, 201–216. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/313

Meyer, M.-C. (2015).**Deriving hurford's constraint.** *Semantics and Linguistic Theory*, *24*, 577. https://doi.org/10.3765/salt.v24i0.2518

Selected references iii


Mayr, C., & Romoli, J. (2016).**A puzzle for theories of redundancy: Exhaustification, incrementality, and the notion of local context.** *Semantics and Pragmatics*, *9*(7). https://doi.org/10.3765/sp.9.7


Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018).**BERT: pre-training of deep bidirectional transformers for language understanding.** *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805


Fox, D., & Spector, B. (2018).**Economy and embedded exhaustification.** *Natural Language Semantics*.


Williams, A., Nangia, N., & Bowman, S. (2018, June). **A broad-coverage challenge corpus for sentence understanding through inference.** In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1112–1122). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101


Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019).**BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.** *CoRR*, *abs/1910.13461*. http://arxiv.org/abs/1910.13461

Selected references iv


Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019).**Roberta: A robustly optimized BERT pretraining approach.** *CoRR, abs/1907.11692*. http://arxiv.org/abs/1907.11692


He, P., Liu, X., Gao, J., & Chen, W. (2020).**Deberta: Decoding-enhanced BERT with disentangled attention.** *CoRR, abs/2006.03654*. https://arxiv.org/abs/2006.03654


Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020, July). **Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition.** In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8690–8705). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.768


Misra, K. (2022).**Minicons: Enabling flexible behavioral and representational analyses of transformer language models.** *arXiv preprint arXiv:2203.13112*.


Hu, J., & Levy, R. (2023). **Prompting is not a substitute for probability measurements in large language models.** https://arxiv.org/abs/2305.13264

Appendix

Repairing scalar HDs via overt operations

- Overt operations (*only*, FOCUS marking), whose overall effect if close to that of covert pragmatic reasoning, can help improve (3). They however don't work for (4)

(16) a. Al ate all or **only** some of the biscuits.

b. Al ate all or SOME of the biscuits.

- It's an independent question why these overt operations can rescue scalar HDs, while covert pragmatic reasoning cannot.
- To account for the difference between overt and covert operations, one can build on the idea that *only* and FOCUS, even though they seem to mimic pragmatic reasoning overtly, bring about inferences that are not introduced at the same level as pragmatic reasoning (assertion vs. presupposition).

Some arguments adapted from Fox and Spector (2018) showing the felicity of scalar HDs is not dictated by surface orderings (I)

- One could think our grammar has a rule that just imposes a weak-to-strong ordering of Hurford Disjuncts over a strong-to-weak ordering.
- This falls short once we look at slight variants of the scalar HDs studied here.
- Example 1: Scalar HDs embedded under a universal. In (17), the contrast between the two orders gets weaker, if it does not totally disappear. Hard to think how memorized order can interact with operators like *must* outscoping the disjunction.

(17) a. Jo **must** finish some or all of the HW exercises by tomorrow.

b. Jo **must** finish all or some of the HW exercises by tomorrow.

Some arguments adapted from Fox and Spector (2018) showing the felicity of scalar HDs is not dictated by surface orderings (II)

- One could think our grammar has a rule that just imposes a weak-to-strong ordering of Hurford Disjuncts over a strong-to-weak ordering.
- This falls short once we look at slight variants of the scalar HDs studied here:
- Example 2: Scalar HDs with universally quantified scalar disjuncts. In (18), the contrast between the two orders gets weaker, if it does not totally disappear. Hard to think how memorized order can interact with operators like *must*, which do not change the strength ordering when they apply to both disjuncts.

(18) a. Jo **must** finish some of the HW exercises by tomorrow, or they **must** finish all.

b. Jo **must** finish all of the HW exercises by tomorrow, or they **must** finish some.

Some arguments adapted from Fox and Spector (2018) showing the felicity of scalar HDs is not dictated by surface orderings (III)

- One could think our grammar has a rule that just imposes a weak-to-strong ordering of Hurford Disjuncts over a strong-to-weak ordering.
- This falls short once we look at slight variants of the scalar HDs studied here:
- Example 3: HDs with scalar disjuncts that are "separated" on their scale by a salient alternative (e.g., *some* and *all* separated by *most*). In (19), the contrast between the two orders gets weaker, if it does not totally disappear. Hard to think how memorized order can interact with the context.

(19) Context: if Jo finished some but not most exercises, they get a B; if they finished **most but not all**, they get an A+, if they finished all, they get an A.

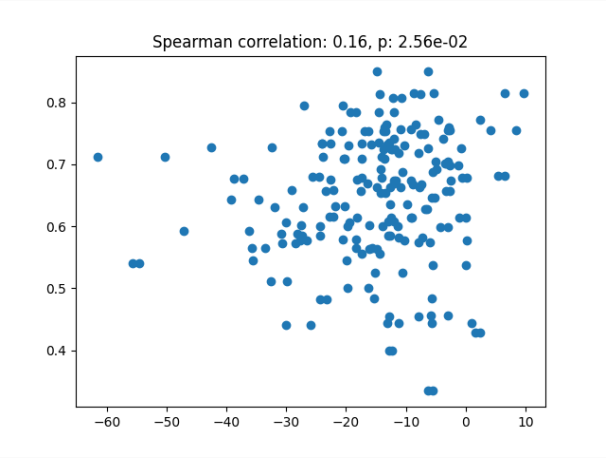
a. Jo finished some or all of the exercises.

b. Jo finished all or some of the exercises.

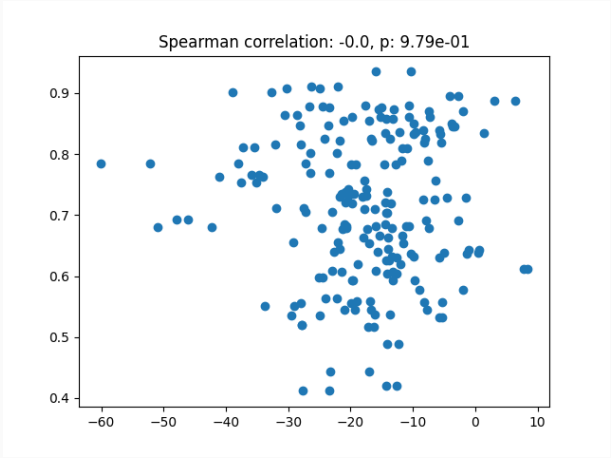
Why not ask Chat-GPT

- First, "asking Chat-GPT" vs. accessing probabilities of possible outputs (either sentence probabilities/surprisals, or label probabilities in the case of NLI), correspond to two distinct tasks:
  - "Asking Chat-GPT" amounts to asking the model to perform introspection on its own "preferences". Absolutely not trivial a model can do this reliably and meaningfully. Asking Chat-GPT is more of a meta-linguistic task: the model will spit out what people *say* about the linguistic phenomenon, not necessarily what is *done* about it.
  - In fact, there is some quantitative evidence that prompt engineering and direct investigations of the output probabilities do not yield robustly similar results (Hu & Levy, 2023).
- Second, from a practical perspective, earlier models are freely available and easy to use in streamlined tasks, and also lighter so that it's faster to iterate if something goes wrong.

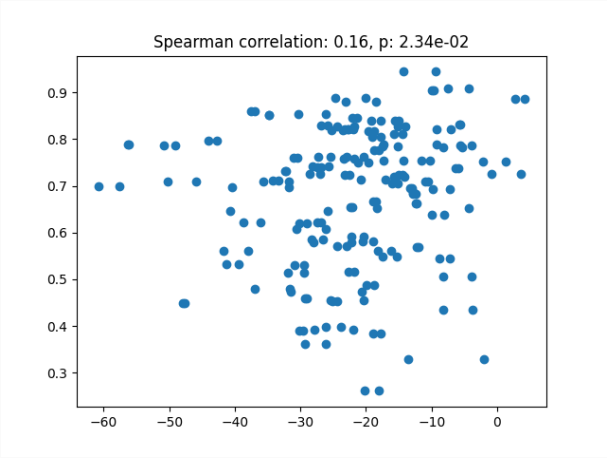
RoBERTa-Large, max compatibility score



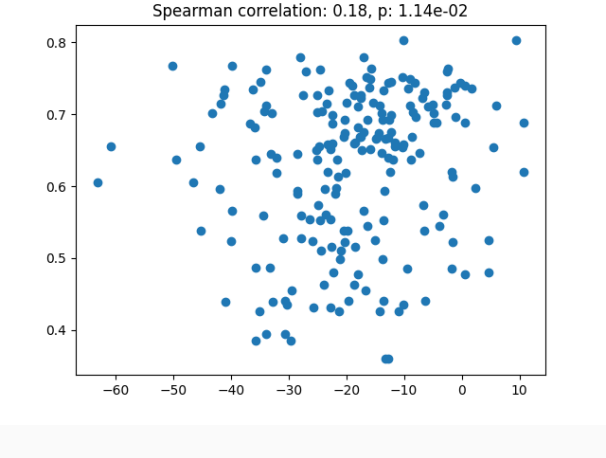
(a) NRHD – (*some*, *not all*)*



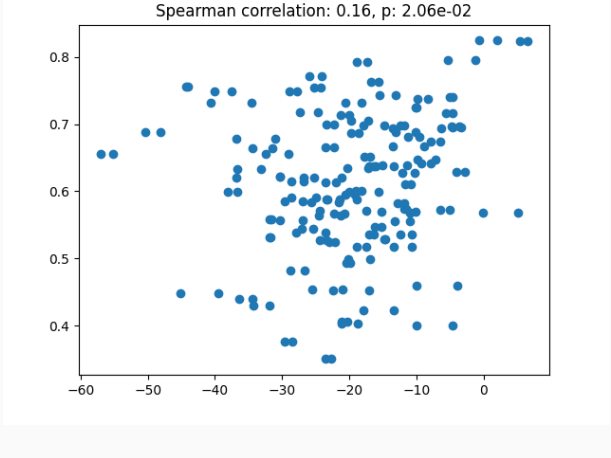
(b) RHD – (*not all*, *no*)



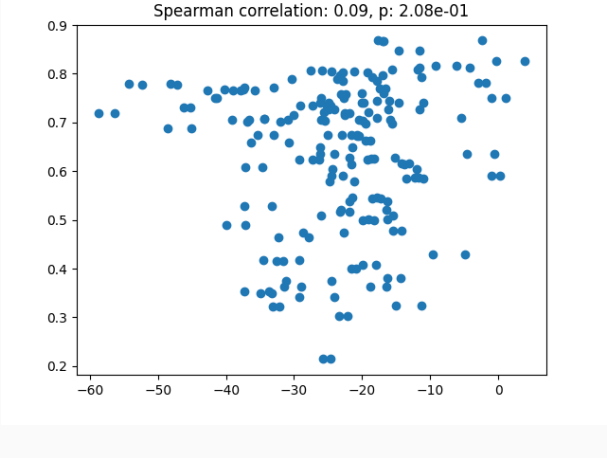
(c) RHD – (*some*, *all*)*



(d) TNHD – (*all*, *not all*)*

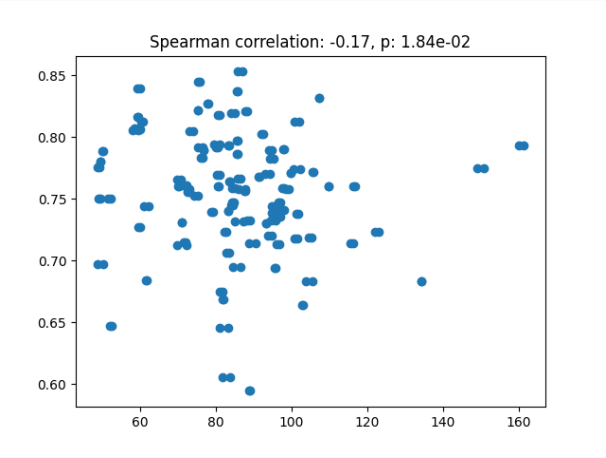


(e) TNHD – (*some*, *no*)*

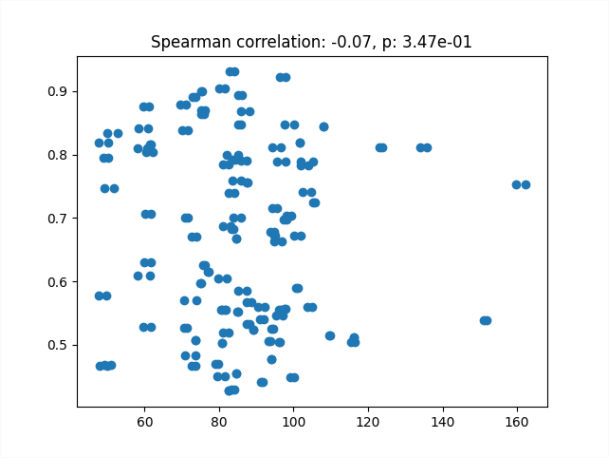


(f) BL – (*all*, *no*)

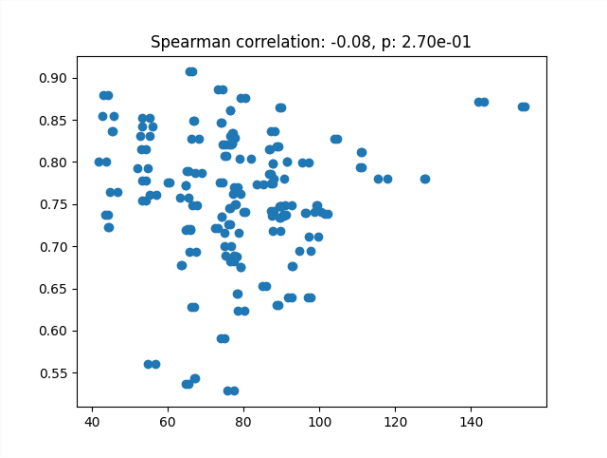
DeBERTa-Large, max compatibility score



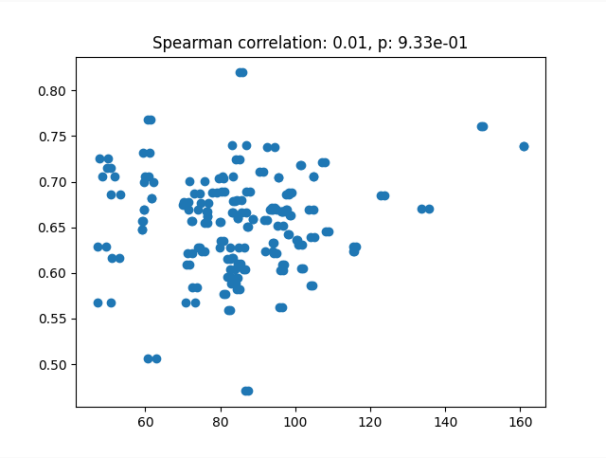
(a) NRHD – (*some*, *not all*)*



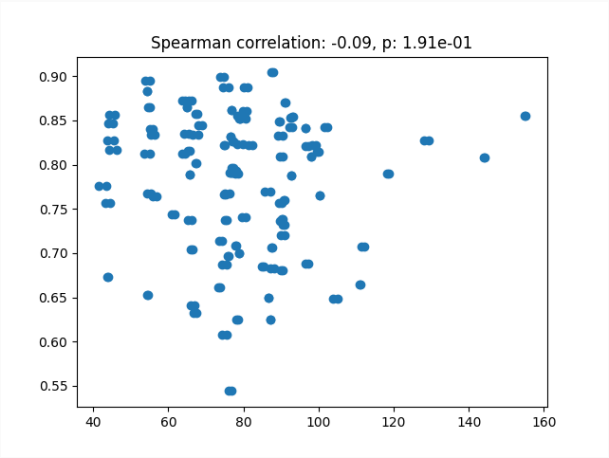
(b) RHD – (*not all*, *no*)



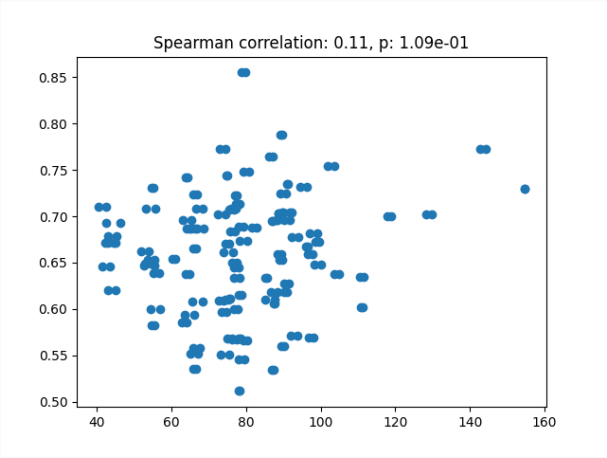
(c) RHD – (*some*, *all*)



(d) TNHD – (*all*, *not all*)

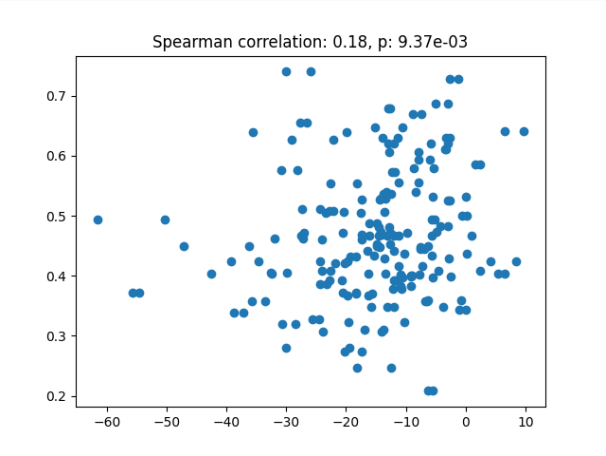


(e) TNHD – (*some*, *no*)

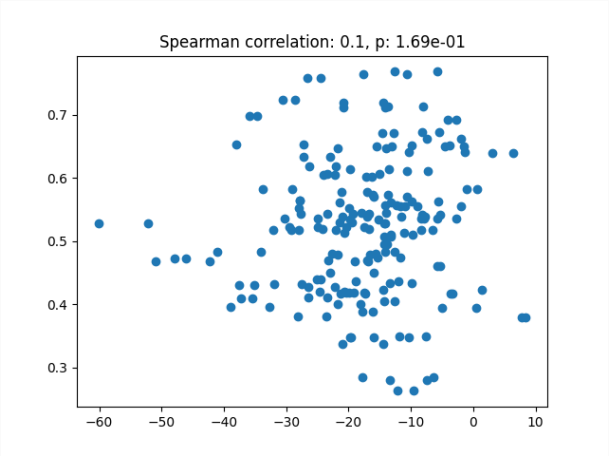


(f) BL – (*all*, *no*)

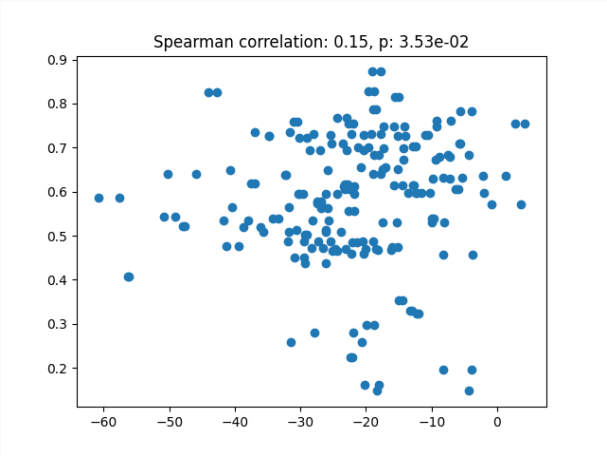
BART-Large, max compatibility score



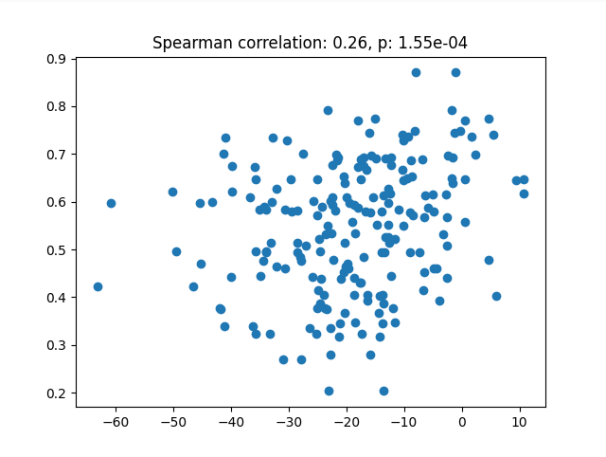
(a) NRHD – (*some*, *not all*)*



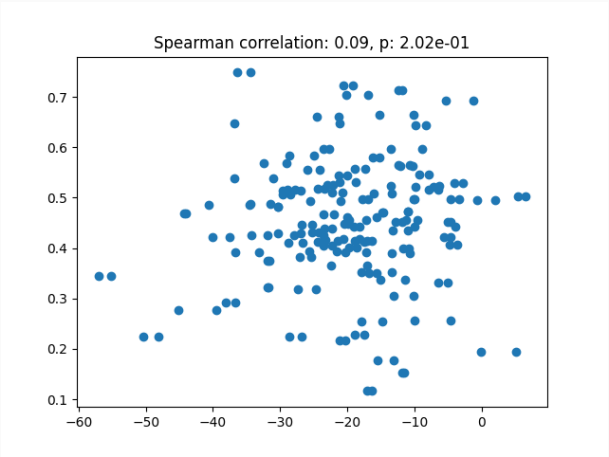
(b) RHD – (*not all*, *no*)



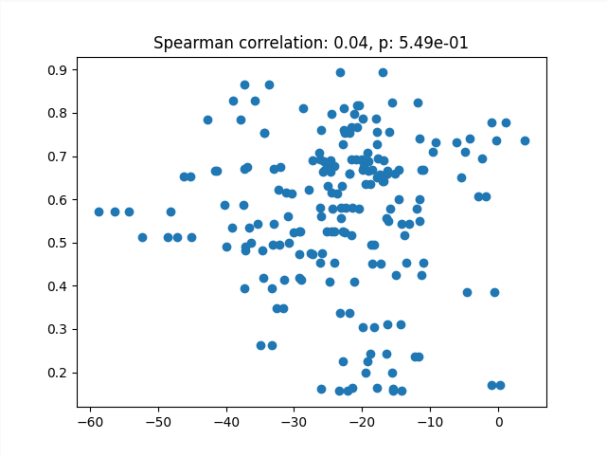
(c) RHD – (*some*, *all*)*



(d) TNHD – (*all*, *not all*)*



(e) TNHD – (*some*, *no*)



(f) BL – (*all*, *no*)