Syntax and grounding in adjective learning
Supplementary material for Study 2

Adèle Hénot-Mortier (MIT)

February 27, 2025

Queen Mary University of London

Theoretical background

Extra background on positive and negative adjectives

- Intuitively positive vs. negative adjectives pattern differently in several respects...
  - Positive (rather than negative) adjectives are used to **ask unbiased degree-related questions**.
  - Positive (rather than negative) adjectives are used to **form unbiased comparatives/equatives**.
  - Negative (rather than positive) adjectives may **feature overt negative morphology**.

(1) a. How **tall** is Jo? $\rightsquigarrow$ Jo may be tall or short.

b. How **short** is Jo? $\rightsquigarrow$ Jo is short.

(2) a. Jo is as **tall** as Al. $\rightsquigarrow$ Both may be tall or short.

b. Jo is as **short** as Al. $\rightsquigarrow$ Both are short.

(3) a. in-competent; im-modest; un-lucky; dis-honest ...

b. *un-small; *im-messy; *un-poor; *dis-arrogant ...

An account of the ITA: Krifka, 2007

The Inference Towards the Antonym (Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018)

(*not* A) $\Longrightarrow$ A' where A and A' are antonyms. (♡)

ITA Pragmatic Mitigation Condition (Krifka, 2007)

(*not* A) $\not\Longrightarrow$ A', if $\textsc{Cplx}$(*not* A) $\gg$ $\textsc{Cplx}$(A') (♢)

Negative Adjectives Complexity Hypothesis (Büring, 2007a, 2007b)

$\forall \mathbf{A^-}.\ \mathbf{A^-} = \textsc{Not-}\mathbf{A^+}$, therefore:

$\textsc{Cplx}(\mathbf{A^-}) = \textsc{Cplx}(\textsc{Not-}\mathbf{A^+}) \sim \textsc{Cplx}(\textit{not}\ \mathbf{A^+})$ (♠)

$\textsc{Cplx}(\textit{not}\ \mathbf{A^-}) = \textsc{Cplx}(\textit{not}\ \textsc{Not-}\mathbf{A^+}) \gg \textsc{Cplx}(\mathbf{A^+})$ (♣)

(4) a. He is not **tall**. $\overset{\heartsuit+\spadesuit+\diamondsuit}{\Longrightarrow}$ He is **short**.

b. He is not **short**. $\overset{\heartsuit+\clubsuit+\diamondsuit}{\not\Longrightarrow}$ He is **tall**.

Refinement of Krifka's predictions: Ruytenbeek et al., 2017

- Intuition: the decomposition **A**$^{-}$ = NOT-**A**$^{+}$ is made **particularly salient when the adjective is transparent**.
- This means that (♠) and (♣) hold even "more unambiguously" for morphologically transparent pairs, which leads to a stronger interaction between the ITA and adjective polarity.
- In other words, **the ITA contrast is expected to be stronger for transparent antonyms (cf. (5)) as opposed to opaque ones**.

(5) a. John is not **lucky**. Paul is **unlucky** too.

b. ## John is not **unlucky**. Paul is **lucky** too.

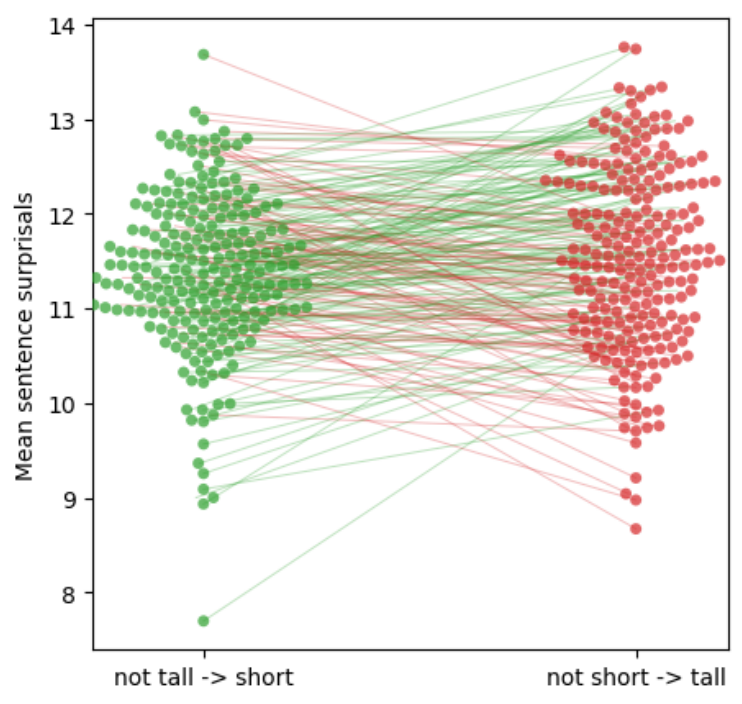
Preposed paradigm, sentence-level, other models

XLNet transparent + opaque

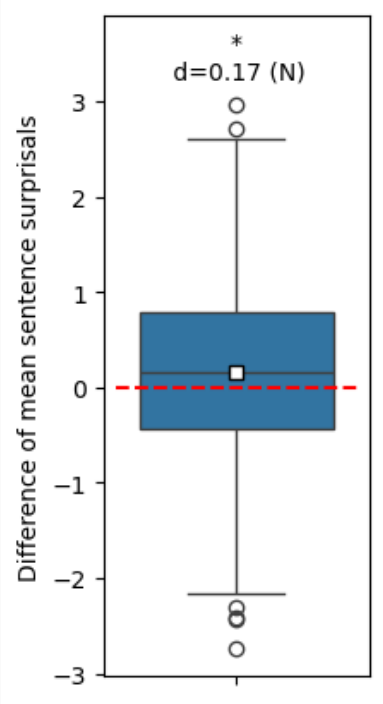
(6) "Preposed *too*"

a. He is not A^+. She too is A^-.

b. # He is not A^-. She too is A^+.

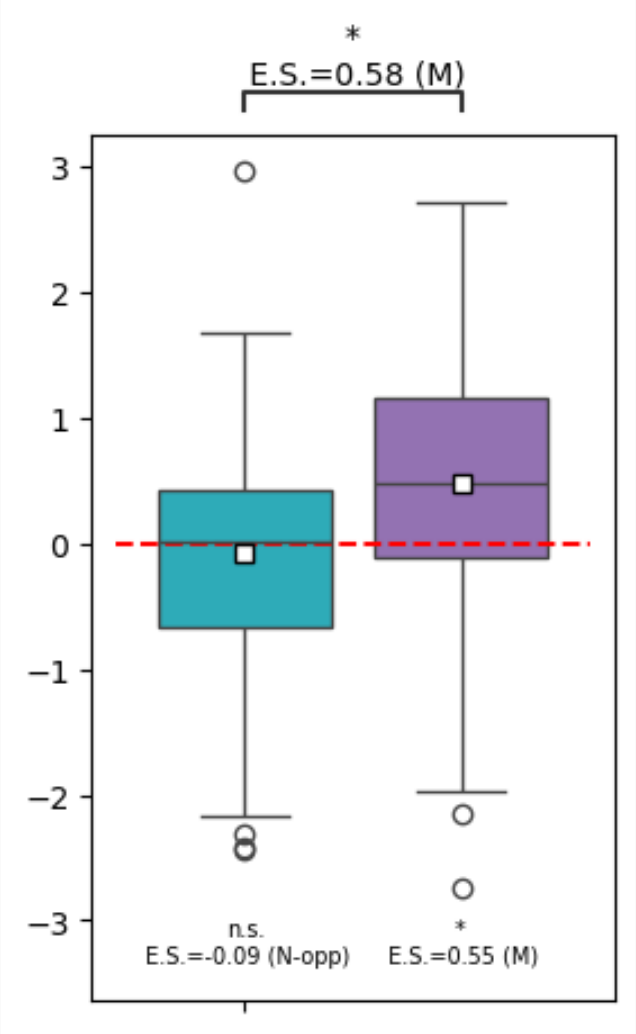


$\mathscr{S}$(6a) and $\mathscr{S}$(6b). Links are sentence pairings. Green links show differences in the expected direction, red links in the opposite direction.



$\mathscr{S}$(6b)−$\mathscr{S}$(6a). White square is the mean. '*' means $p < .05$; effect size is Cohen's d. N=Negligible.

XLNet transparent vs. opaque



$\mathscr{S}$(6b)−$\mathscr{S}$(6a), **transparent** vs. **opaque** pairs. White squares are means. Within-group p-values are BY-corrected. Effect size is Cohen's d. N-opp=Negligible opposite, M=Medium.

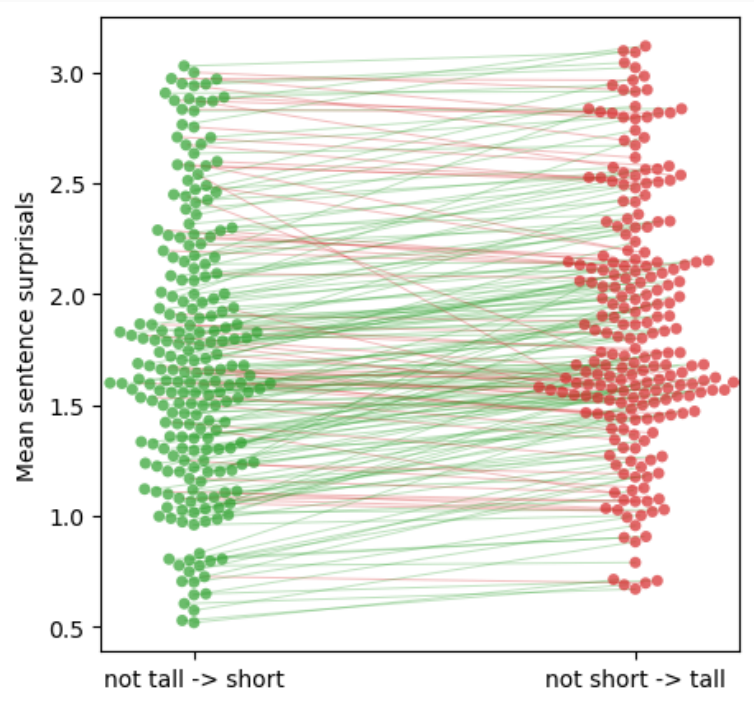
- Morphologically transparent pairs are associated with a stronger contrast than opaque pairs.
- In fact, only the transparent group gives rise to a significant contrast in ITA.

BERT transparent + opaque

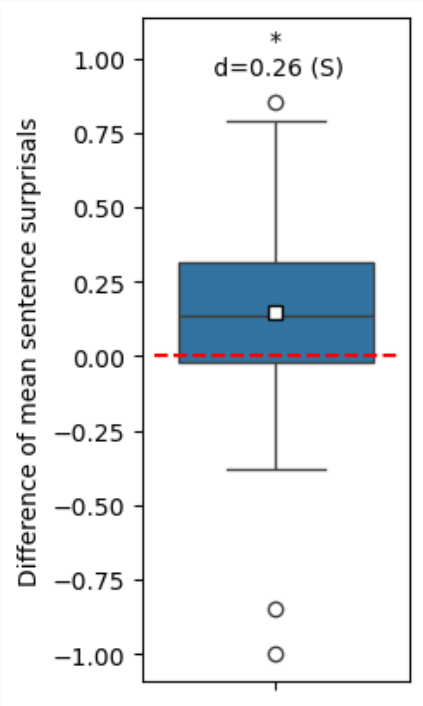
(6) "Preposed *too*"

a. He is not A^+. She too is A^-.

b. # He is not A^-. She too is A^+.

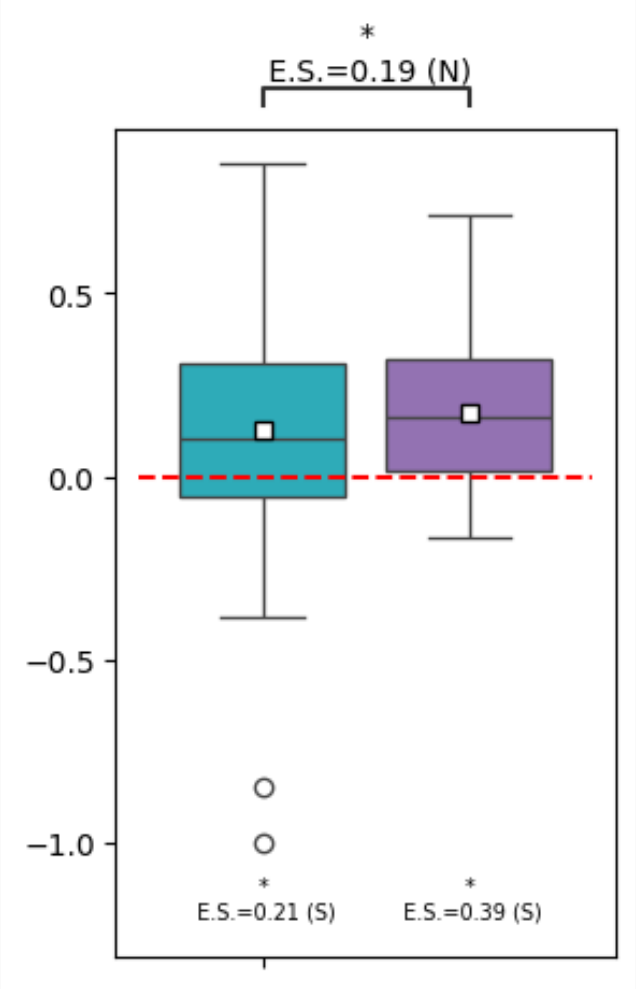


$\mathscr{S}$(6a) and $\mathscr{S}$(6b). Links are sentence pairings. Green links show differences in the expected direction, red links in the opposite direction.



$\mathscr{S}$(6b)−$\mathscr{S}$(6a). White square is the mean. '*' means $p < .05$; effect size is Cohen's d. S=Small.

BERT transparent vs. opaque



$\mathscr{S}$(6b)−$\mathscr{S}$(6a), **transparent** vs. **opaque** pairs. White squares are means. Within-group *p*-values are BY-corrected. Effect size is Cohen's *d*. S=Small, N=Negligible.

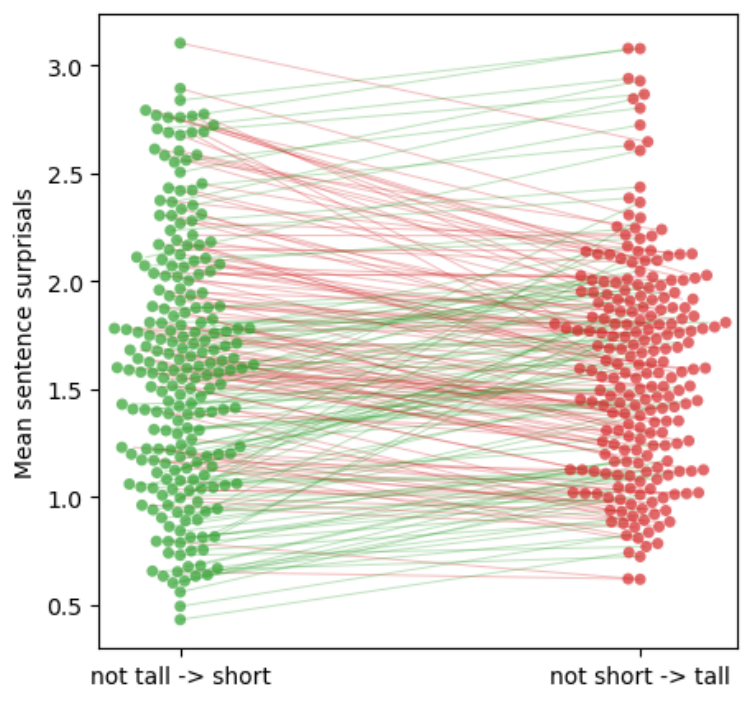
- No significant difference between morphologically transparent and opaque pairs.
- Significant, small contrast in ITA in both groups.

RoBERTa transparent + opaque

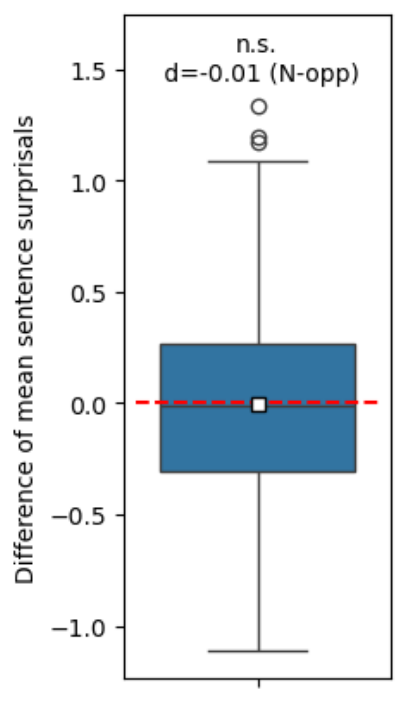
(6) "Preposed *too*"

a. He is not A^+. She too is A^-.

b. # He is not A^-. She too is A^+.

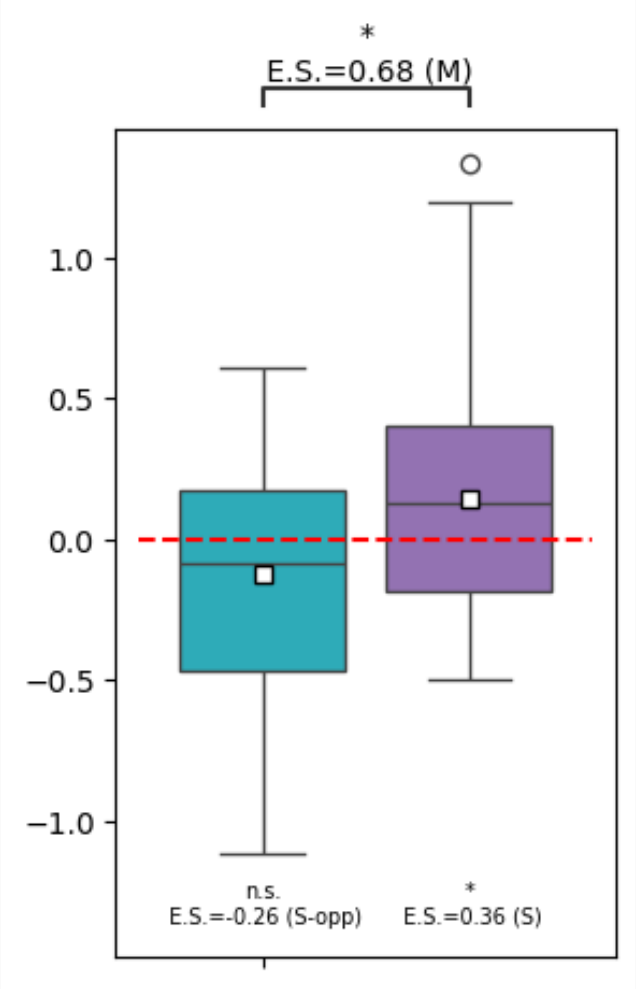


$\mathscr{S}$(6a) and $\mathscr{S}$(6b). Links are sentence pairings. Green links show differences in the expected direction, red links in the opposite direction.



$\mathscr{S}$(6b)−$\mathscr{S}$(6a). White square is the mean. '*' means $p < .05$; effect size is Cohen's d. N-opp=Negligible opposite.

RoBERTa transparent vs. opaque



$\mathscr{S}$(6b)−$\mathscr{S}$(6a), transparent vs. opaque pairs. White squares are means. Within-group p-values are BY-corrected. Effect size is Cohen's d. S-opp=Small opposite, S=Small, M=Medium.

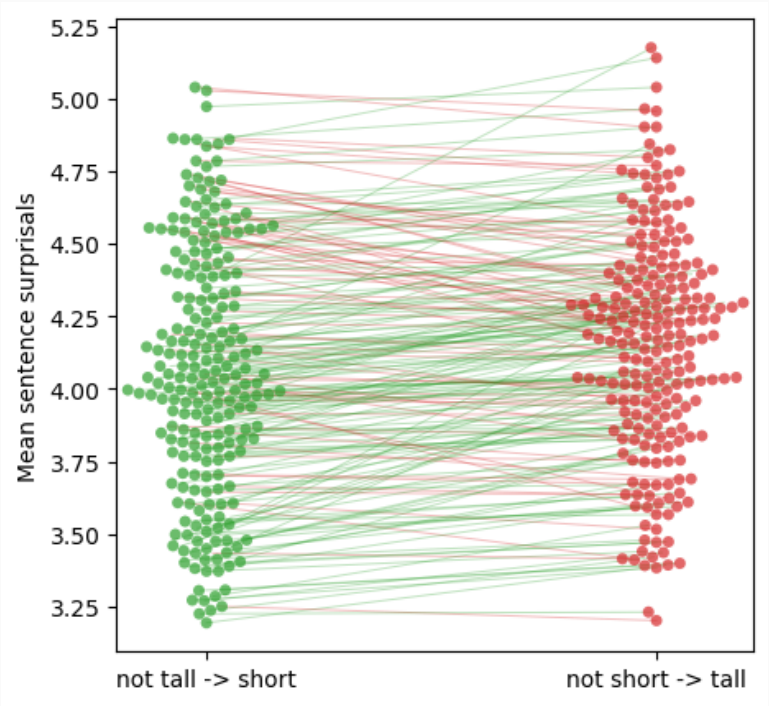
- Morphologically transparent pairs are associated with a stronger contrast than opaque pairs.
- In fact, the transparent group gives rise to a significant contrast in ITA in the right direction, while the opaque group gives rise to a contrast, *in the wrong direction*!

Mistral 7B transparent + opaque

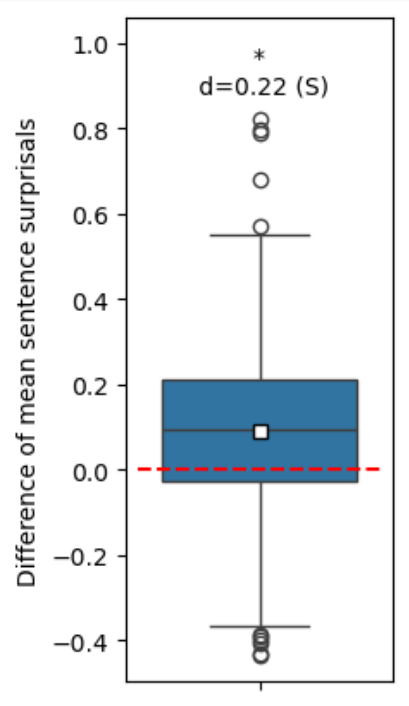
(6) "Preposed *too*"

a. He is not A^+. She too is A^-.

b. # He is not A^-. She too is A^+.

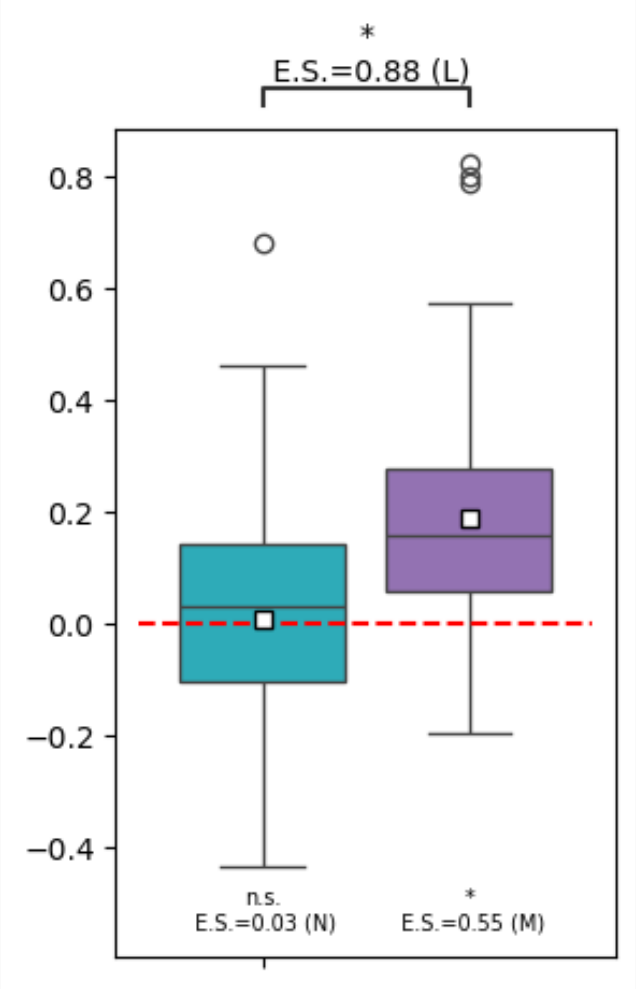


$\mathscr{S}$(6a) and $\mathscr{S}$(6b). Links are sentence pairings. Green links show differences in the expected direction, red links in the opposite direction.



$\mathscr{S}$(6b)−$\mathscr{S}$(6a). White square is the mean. '*' means $p < .05$; effect size is Cohen's d. S=Small.

Mistral 7B transparent vs. opaque



$\mathscr{S}$(6b)−$\mathscr{S}$(6a), **transparent** vs. **opaque** pairs. White squares are means. Within-group *p*-values are BY-corrected. Effect size is Cohen's *d*. N=Negligible, M=Medium, L=Large.

- Morphologically transparent pairs are associated with a stronger contrast than opaque pairs.
- In fact, only the transparent group gives rise to a significant contrast in ITA.

Other paradigms tested

All three "paradigms"

- 3 kinds of minimal pairs were assessed in 3 different sub-experiments. All pairs of sentences were counterbalanced for gender and filled with the 111 possible (**A**$^+$, **A**$^-$) antonymic pairs.

(6) "Preposed *too*" (does more justice to left-to-right LLMs)

a. He is not **A**$^+$. She too is **A**$^-$.

b. # He is not **A**$^-$. She too is **A**$^+$.

(7) "Postposed *too*" (very close to the stimuli in Ruytenbeek et al., 2017)

a. He is not **A**$^+$, and she is **A**$^-$ too.

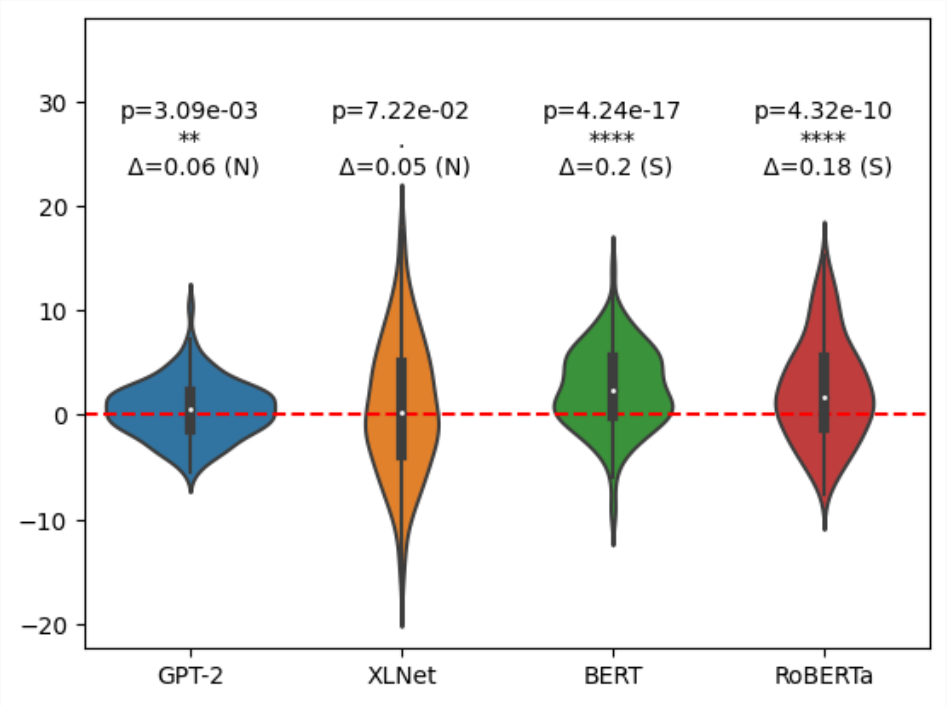
b. # He is not **A**$^-$, and she is **A**$^+$ too.

(8) "Meta"

a. He is not **A**$^+$ means that he is **A**$^-$.

b. # He is not **A**$^-$ means that he is **A**$^+$.

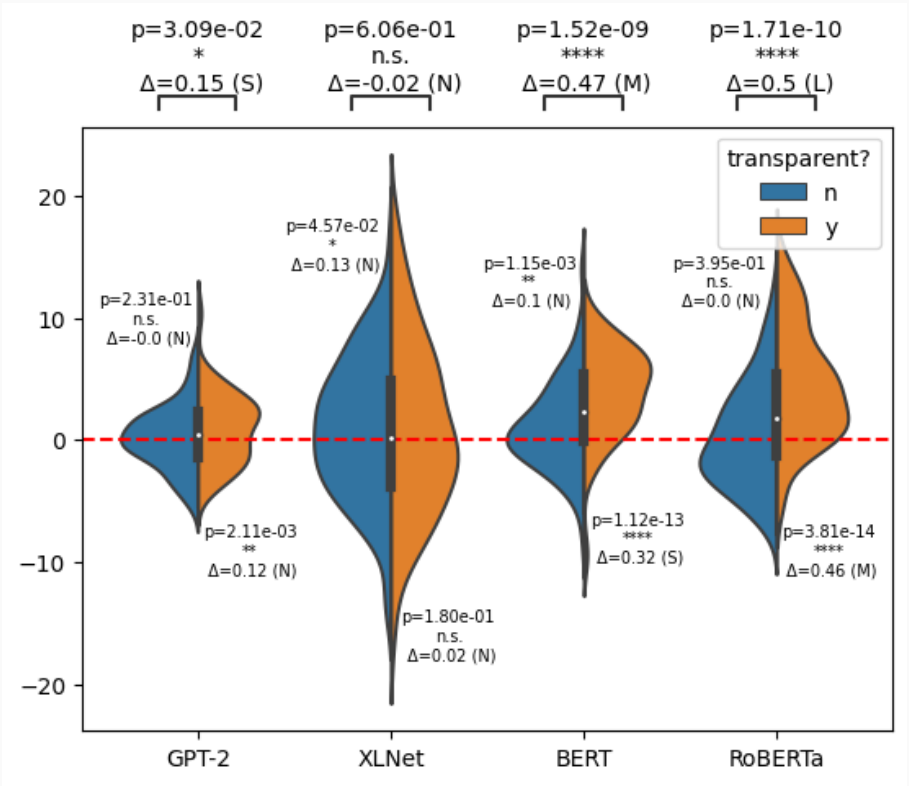
Sentence-level results for all models, postposed too paradigm



Paired differences in sentence surprisal between (5′b) and (5′a), *p*-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- All models but one (XLNet) exhibit a significant contrast in ITA strength, but the effect sizes are negligible (GPT-2) or small (BERT/RoBERTa).
- Because *too* appears after the critical adjectives, this paradigm expectedly favors bidirectional models.

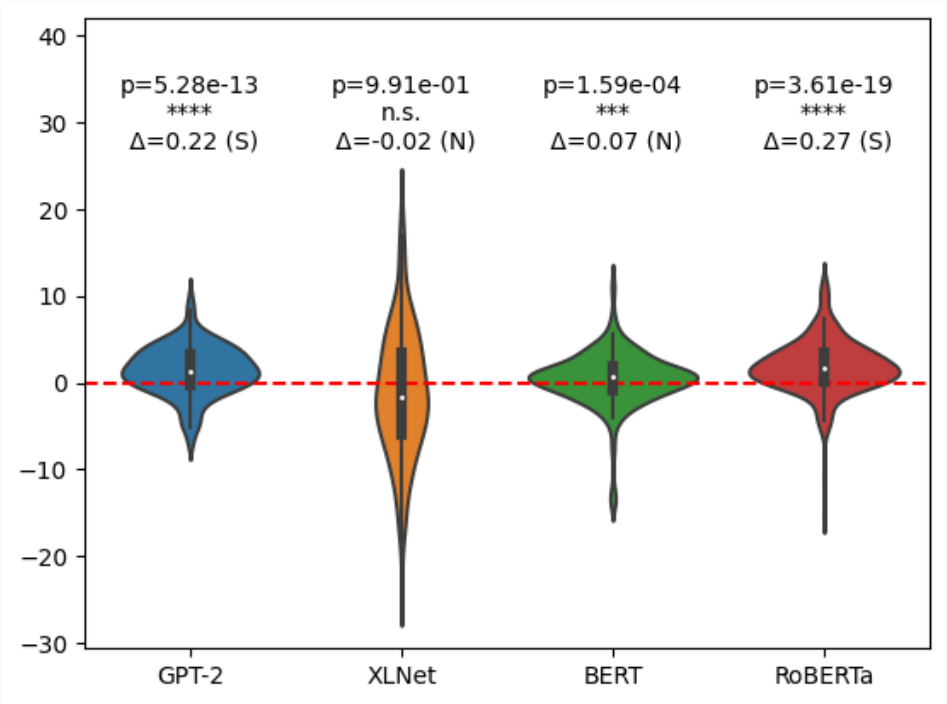
Postposed too paradigm at the sentence-level: group-by-group



Paired differences in sentence surprisal between (5′b) and (5′a), group-by-group (T vs. O), *p*-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- BERT is the only model for which H1 is individually verified by both the T- and O-group.
- BERT also verifies H2, meaning, the T-group is associated to a bigger contrast in ITA strength than the O-group (medium effect size).

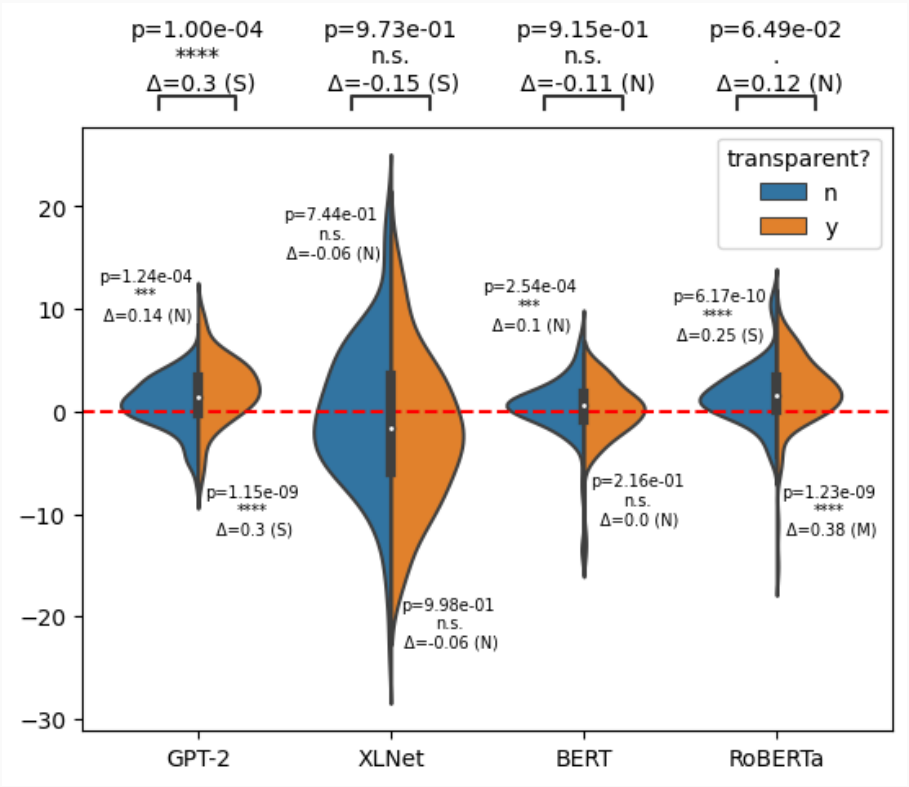
"Meta" paradigm at the sentence-level:both groups



Paired differences in sentence surprisal between (8b) and (8a), *p*-value computed using a Wilcoxon test, effect sizes with Cliff's Δ.

- All models but one (XLNet) exhibit a significant contrast in ITA strength, but the effect sizes are negligible (BERT) or small (GPT-2/RoBERTa).

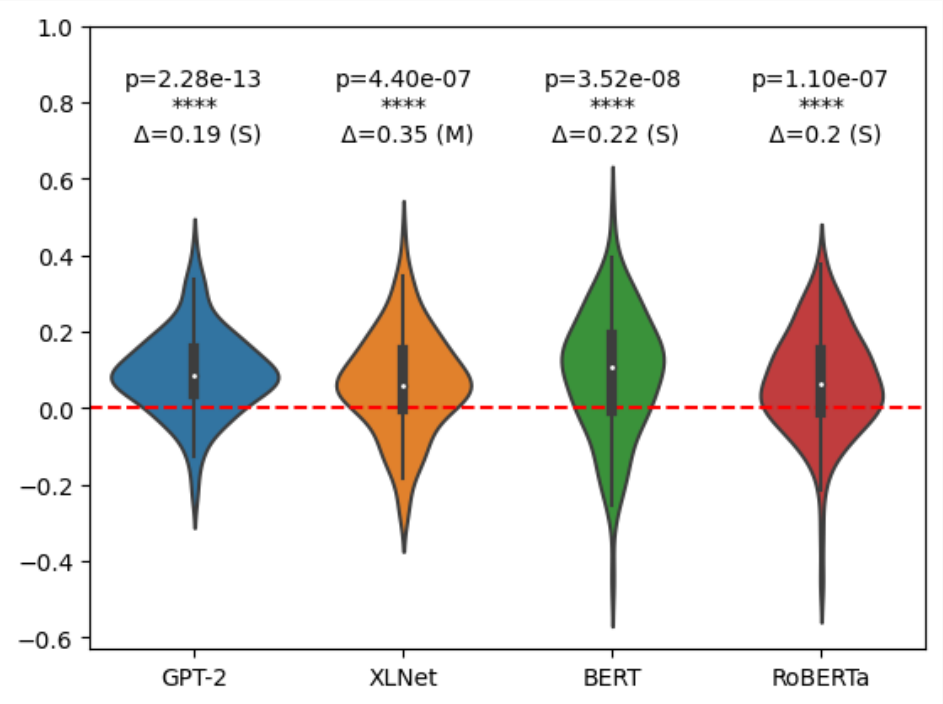
“Meta” paradigm at the sentence-level:group-by-group



Paired differences in sentence surprisal between (8b) and (8a), group-by-group (T vs. O), *p*-value computed using a Wilcoxon test, effect sizes with Cliff’s Δ.

- GPT-2 and RoBERTa are the two models for which H1 is individually verified by both the T- and O-group.
- But only GPT-2 clearly verifies H2 (RoBERTa is characterized by a negligible effect size...).

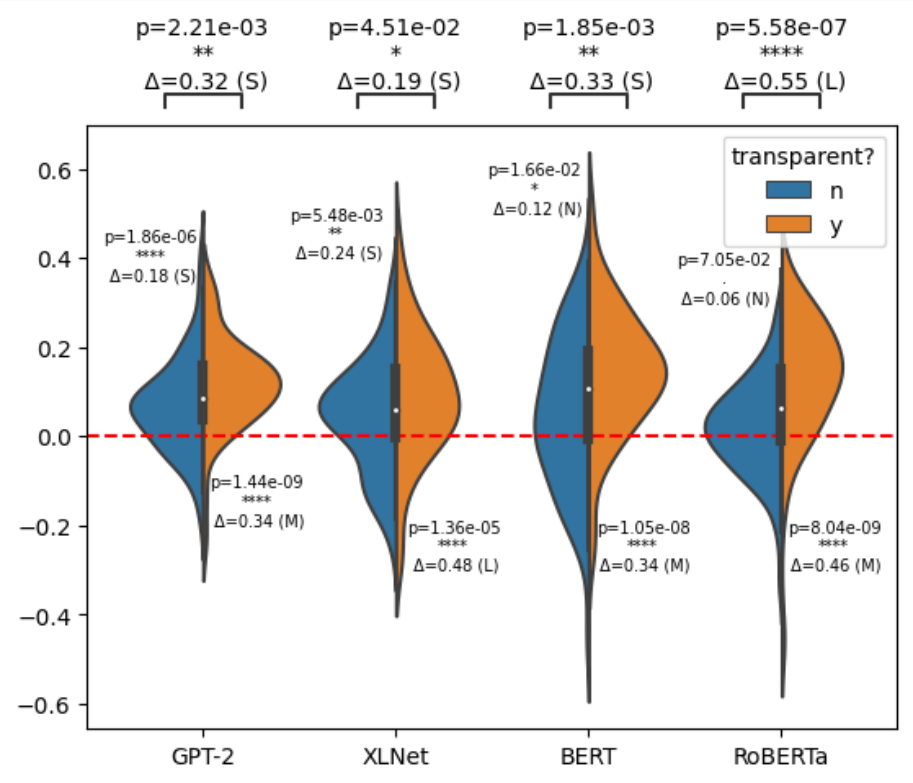
Results for H1, both groups



Paired differences in cosine similarities between $(\overrightarrow{\text{not } \mathbf{A}^+}, \overrightarrow{\mathbf{A}^-})$ and $(\overrightarrow{\text{not } \mathbf{A}^-}, \overrightarrow{\mathbf{A}^+})$, p-value computed using a Wilcoxon test, effect sizes using Cliff's Δ.

- All models exhibit a **significant contrast in cosine similarities (and by proxy ITA strength) as a function of adjective polarity**, with small-to-medium effect sizes.
- This suggests that H1 translates into a topological inequality within the LLMs' vector spaces!

Results for H1, group-by-group, and H2



Paired differences in cosine similarities between $(\overrightarrow{\text{not } \mathbf{A}^+}, \overrightarrow{\mathbf{A}^-})$ and $(\overrightarrow{\text{not } \mathbf{A}^-}, \overrightarrow{\mathbf{A}^+})$, group-by-group p-values computed using a Wilcoxon test, and between-group p-values using a Mann-Whitney U-test. Effect sizes are Cliff's Δ.

- GPT-2 and XLNet are the two models for which H1 is individually verified by both the T- and O-group.
- Both models also verify H2, meaning, the T-group is associated to a bigger contrast in ITA strength than the O-group (small effect sizes).
- **Quite encouraging results overall but...**

Word-level

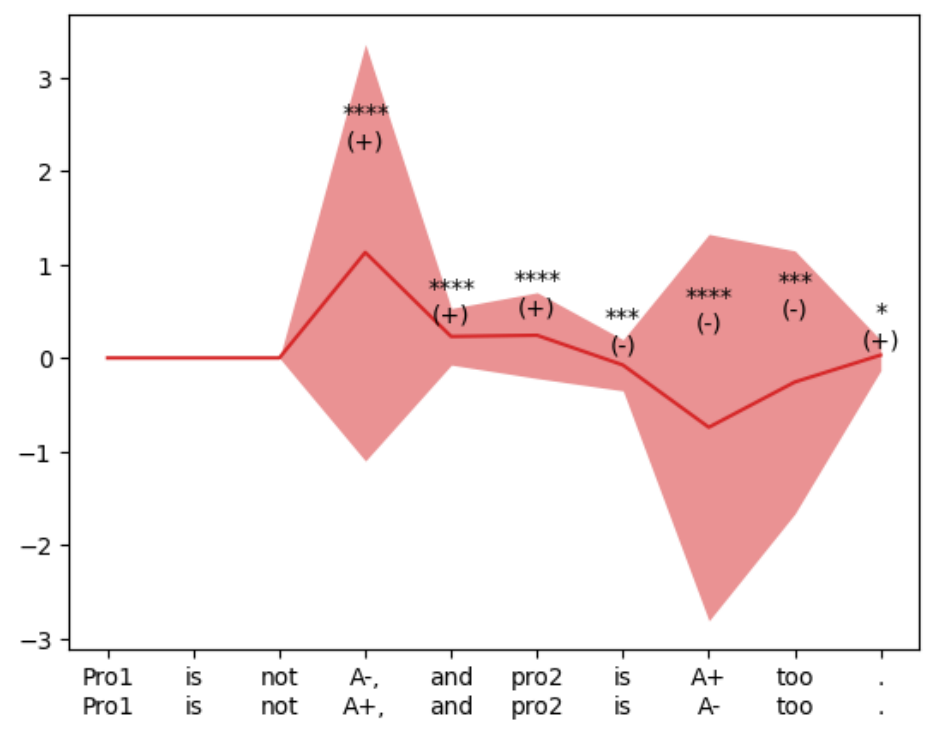
Predictions of left-to-right processing on word-level surprisal

- **What do the best performing models do at the word-level?**
- From a language processing standpoint, we expect the positive contrasts in surprisal witnessed in the sentence-level assessments to be **driven by the occurrence of the second adjective**:
  - given what precedes it, this adjective is expected to be ok (i.e. not surprising) when **negative**;
  - and less ok (i.e. quite surprising) when **positive**.

(6) a. He is not **A**$^{+}$. She too is **A**$^{-}_{\odot}$.

b. # He is not **A**$^{-}$. She too is **A**$^{+}_{\odot}$.

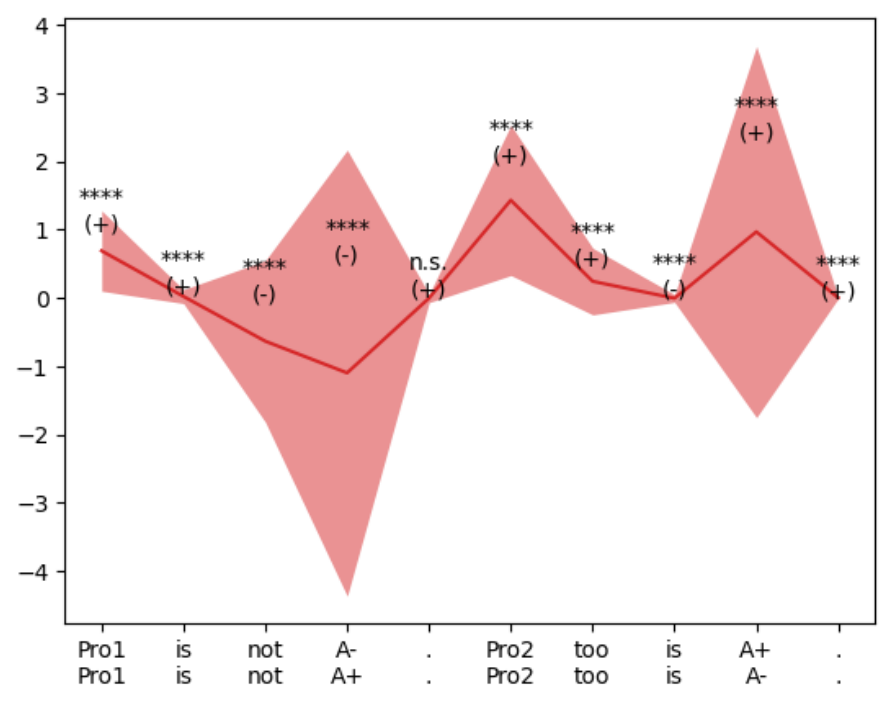
Word-level processing: GPT-2



Paired word-by-word differences in surprisal between (2″b) and (2″a), *p*-values computed using Wilcoxon tests. Red line is the mean, red enveloppe is the standard deviation. Similar plots for the two other paradigms.

- $\mathbf{A}^{-}$ is significantly more surprising than $\mathbf{A}^{+}$ after negation (position 4)...
- but also in position 8 (second occurrence), against the expectations...
- **The effect witnessed at the sentence-level was driven by the wrong element of the sentence!!!**
- BERT and RoBERTa did better but evaluating bidirectional models at the word-level is also trickier.

Word-level processing: BERT



Paired word-by-word differences in surprisal between (8b) and (8a), *p*-values computed using Wilcoxon tests. Red line is the mean, red enveloppe is the standard deviation. Similar plots for the two other paradigms.

- $\mathbf{A}^-$ is significantly less surprising than $\mathbf{A}^+$ after negation (position 4)...
- and also significantly less surprising than $\mathbf{A}^+$ in position 9.
- The effect witnessed at the sentence-level makes sense at the word-level.
- But some amount of negative surprisal may have "transferred" from position 9 to position 4, due to the model's bidirectionality.

Neural assessment

Measuring the ITA in the embedding space

- In this task, we abandon stimuli sentences to focus on the **internal (vector) representations assigned by the original standard LLMs to $\mathbf{A}^+$, $\mathbf{A}^-$, and their respective negations**: $\overrightarrow{\mathbf{A}^+}$, $\overrightarrow{\mathbf{A}^-}$, $\overrightarrow{\text{not } \mathbf{A}^+}$, $\overrightarrow{\text{not } \mathbf{A}^-}$.[2]
- A common measure of semantic proximity in such vector spaces is cosine similarity:

$$CosSim(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 . \vec{v}_2}{||\vec{v}_1|| \times ||\vec{v}_2||} \in [-1; 1]$$

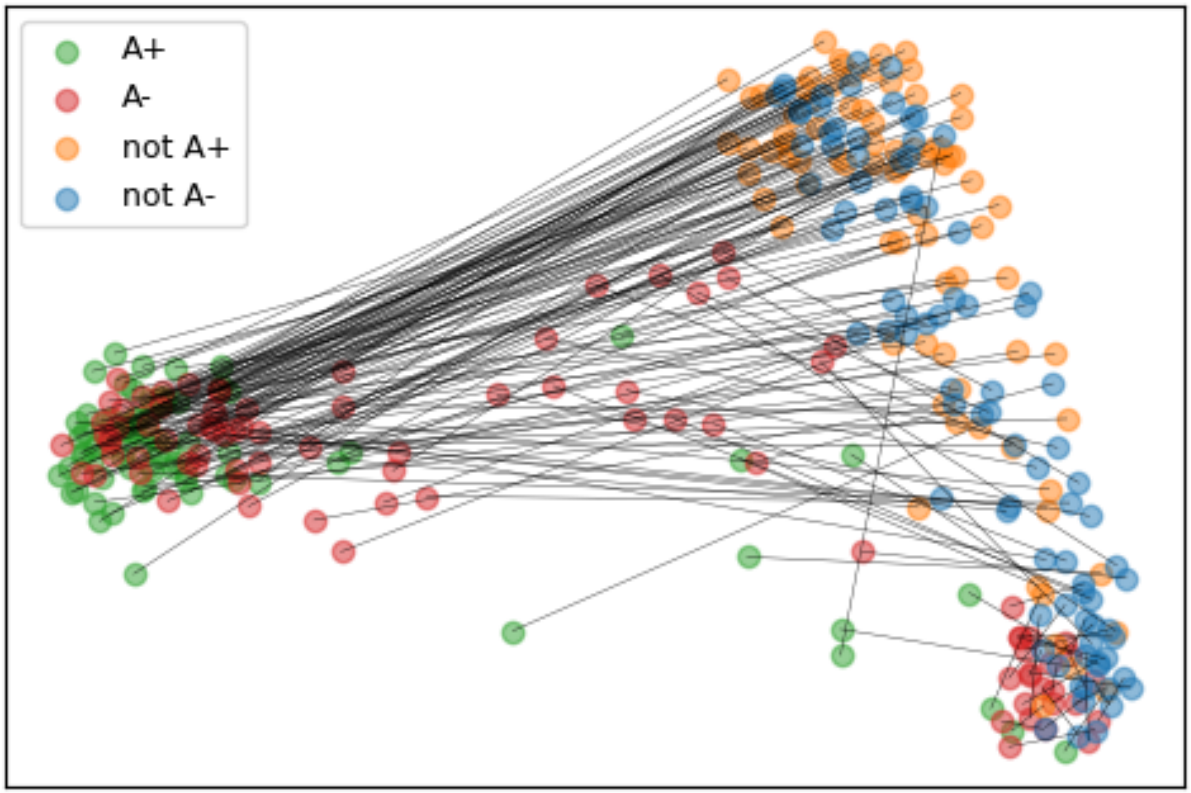
- If H1 translates into the LLMs' vector space, we then expect $\overrightarrow{\text{not } \mathbf{A}^+}$ to be closer to $\overrightarrow{\mathbf{A}^-}$ than $\overrightarrow{\text{not } \mathbf{A}^-}$ is close to $\overrightarrow{\mathbf{A}^+}$, i.e.:

$$CosSim(\overrightarrow{\text{not } \mathbf{A}^+}, \overrightarrow{\mathbf{A}^-}) - CosSim(\overrightarrow{\text{not } \mathbf{A}^-}, \overrightarrow{\mathbf{A}^+}) > 0$$

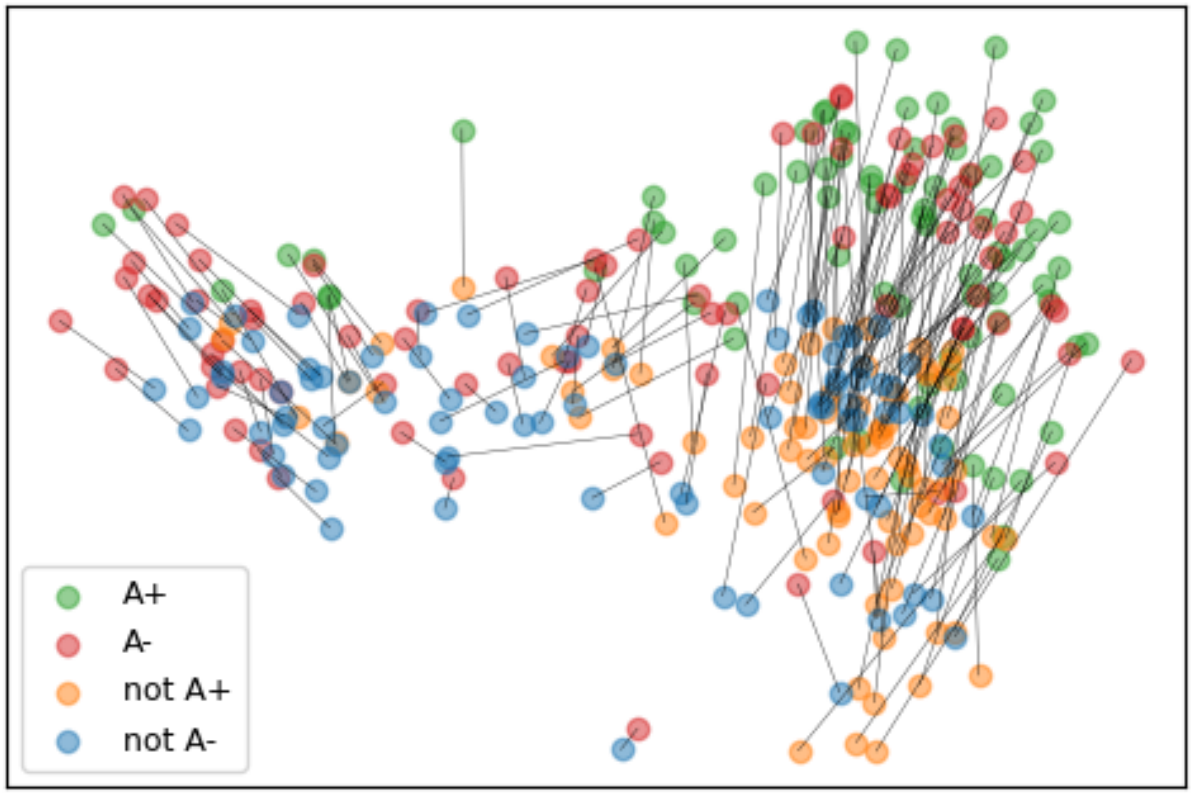
- Moreover, H2 predicts that this difference should be bigger for T-antonyms as opposed to O-antonyms.

[2] In practice, we included the copula *is* as a left context to get those representations.

XLNet Embedding



BERT Embedding



RoBERTa Embedding