Syntax and grounding in adjective learning[1]

Adèle Hénot-Mortier (MIT)

February 27, 2025

Queen Mary University of London

[1] Many thanks to Athulya Aravind, Ido Benbaji, Martin Hackl, Nina Haslinger, Roni Katzir, Yoon Kim, Keely New, Zhouyi Sun, and the audiences of my practice talks at MIT, of SuB28, and my classmates in the Fall 2022 24.S96 seminar and the Spring 2023 6.S986 seminar for their support, comments, and feedback.

Introduction

Which factors enable word learning in humans?

"You shall **know** a word by the **company** it keeps".

J. R. Firth, *Studies in Linguistic Analysis*, 1957

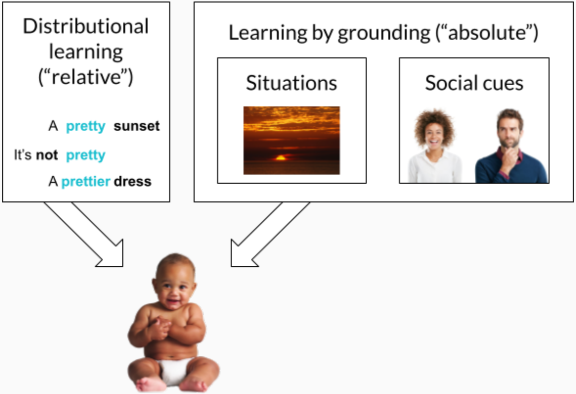
- Distributional Hypothesis (Harris, 1954): words with similar **syntactic environments** have similar **meanings**.

An example of distributional learning

- Distributions encode a lot of information, but comes with challenges!

Sunsets are **so**	**pretty**	
The **red dress** is	**pretti-**	**er** than the **blue** one
Jo finds **Crocs**	**pretty**	
Anglerfish do **not** look	**pretty**	
This is a	**pretty**	**ugly** way to say it

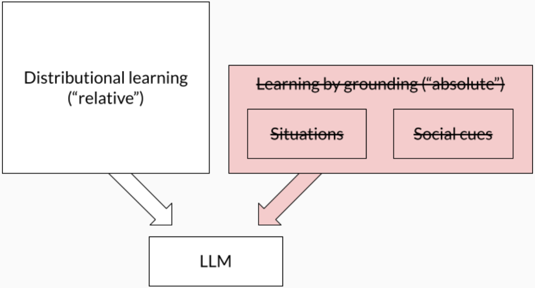
Human word learning results from entangled factors



- Both distributional cues[1] and grounding are used for word learning in humans, but these factors are **hard to disentangle**!

[1]L. R. Gleitman, 1981; L. Gleitman, 1990; Naigles, 1990; Snedeker and Gleitman, 2004; Syrett, 2007; Yuan et al., 2012; Gotowski, 2022.

LLMs can help determine the limits of distributional learning



- **L**arge **L**anguage **M**odels, (LLMs) **typically do not display grounding**.[2] They therefore represent an interesting edge case re:

How far can distributional information alone take us?

[2]Cf. Bender and Koller (2020) for a position paper. Multimodal LLMs exist however (Alayrac et al., 2022 i.a.), and may arguably display more grounding. For this reason they appear less relevant to our research question.

Plan for today

Plan for today

- **Two case studies** focusing on **adjective learning**.
- They vary in **how much distributional information** can be used by LLMs to distinguish adjectives.
- **Successes or failures** inform us re:

How far can distributional information alone take us?

The two studies

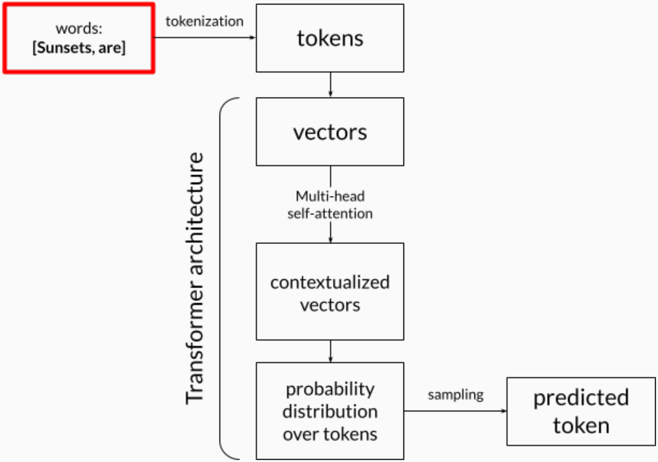
- **Study 1** focuses on the argument structure of adjectives like **tough**, **pretty**, **brave**, and **short**.
  - The observed distinctions are **distributionally clear**...
  - but **intuitively subtle**.
- **Study 2** focuses on antonymic adjectives (e.g. **tall**/**short**) and their behavior under negation.
  - The observed distinction is **intuitively obvious**...
  - but **distributionally subtle**.

Methods

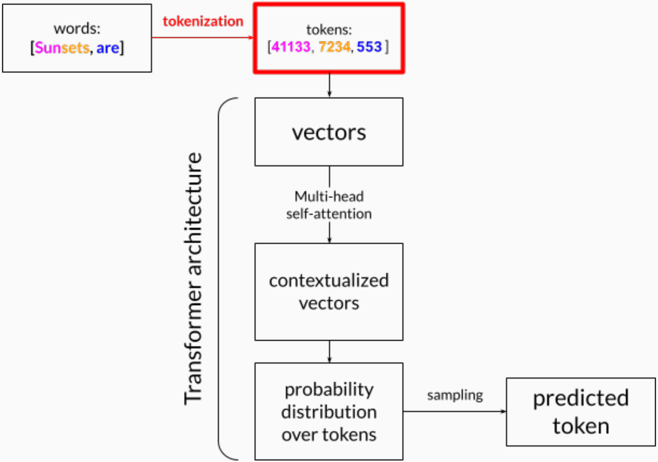
Structure of both studies

- Two kinds of "assessment", inspired by psycholinguistics.
- **"Behavioral":** are LLMs differentially "**surprised**" when processing contrasting sentences that only differ in the adjectives used?
- **"Neural":** are the behavioral contrasts, rooted in the internal **vector** representations assigned by the models to the adjectives ?

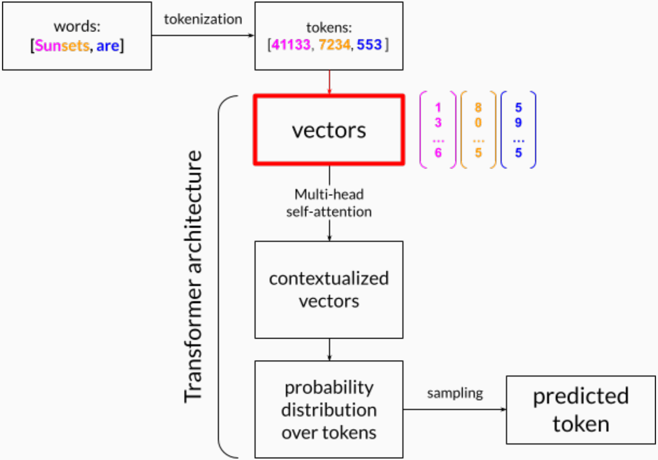
A bird's eye view on LLMs' architecture



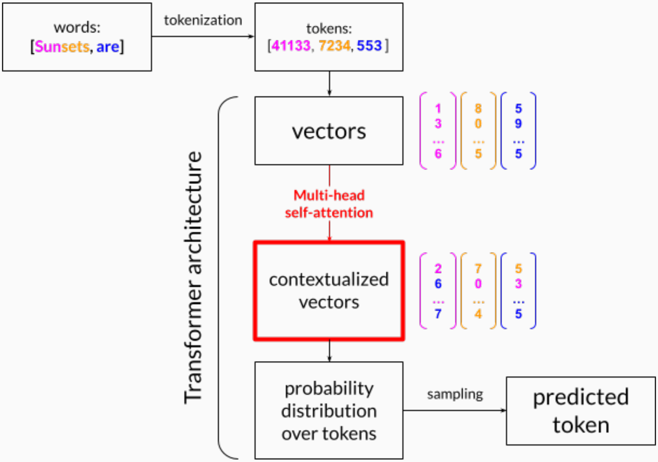
A bird's-eye view on LLMs' architecture



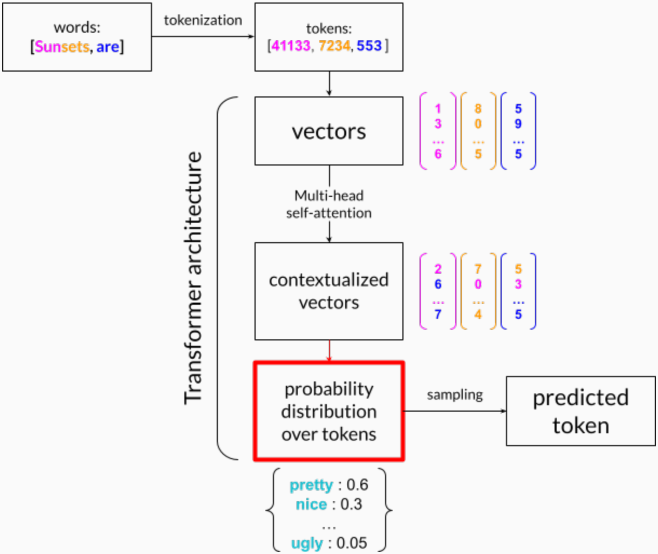
A bird's-eye view on LLMs' architecture



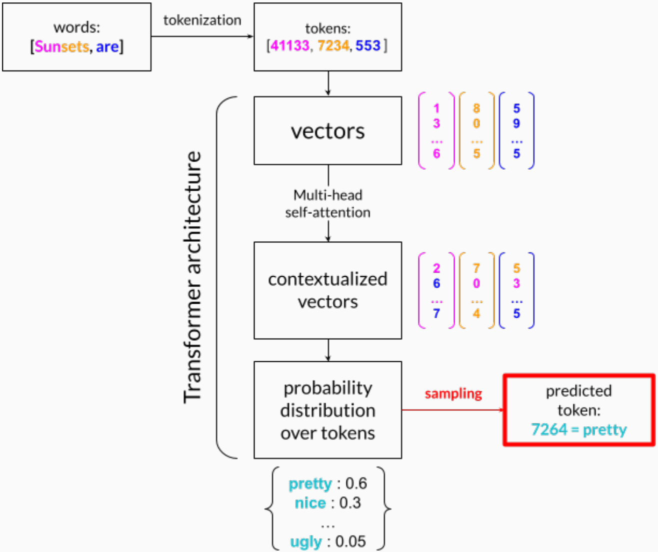
A bird's-eye view on LLMs' architecture



A bird's-eye view on LLMs' architecture

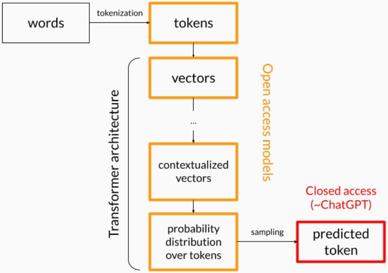


A bird's-eye view on LLMs' architecture



Models tested, and rationale

- Five Transformers: GPT-2, XLNet, BERT, RoBERTa, Mistral7B.[3]
- Though not state-of-the-art, **open-access**.
- Contrast with indirect prompting methods.[4]
- Allow to evaluate the **robustness** of the Transformer architecture.



- **Focus on the best-performing model, GPT-2.**

[4]Vaswani et al., 2017; Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019; Yang et al., 2019; Jiang et al., 2023

[4]Hu and Levy (2023) shows that prompting and probability assessment can yield significantly different outcomes.

Study 1

Distinguishing adjectives through their syntactic distribution.
Code; Surprisal dataset.

Learning to distinguish categories of adjectives

- It seem hard to distinguish adjectives like **short**, **tough**, **pretty**, and **brave** at first blush.

(1) a. This problem is **short**/**tough**/?**pretty**/***brave**.
   b. Jo is **short**/**tough**/**pretty**/**brave**.
   c. This decision is ***short**/**tough**/***pretty**/**brave**.

- **The Distributional Hypothesis can help**: **short**, **tough**, **pretty**, and **brave** can be easily and sharply teased apart in terms of their syntactic distributions.[5]

[5] Supported by syntactic theory: cf. Rosenbaum (1967), Lasnik and Fiengo (1974), Stowell (1991), and Keine and Poole (2017), among many others.

Differences in terms of clausal embedding

- **Short**-like adjectives cannot embed an **infinitival clause**, while the other adjectives can.

(2) **This X is A to VP**

a. * This kid is **short**/**old**/**poor** to ride the rollercoaster.

b. This problem is **tough**/**interesting**/**impossible** to solve.

c. This vase is **pretty**/**harmonious** to look at.

d. This student is **brave**/**rude**/**smart** to point out the issue.

Differences in the availability of an "impersonal" variant

- **Tough**- and **brave**-adjectives can take a **dummy** *it* as subject, while **pretty**- and **short**-like adjectives can't.

(3) **It's Adj to VP**

a. **It**'s **tough** to solve this problem.

b. **It**'s **brave** to point out the issue.

c. * **It**'s **pretty** to look at this vase.

d. * **It**'s **short** to ride the rollercoaster.

Two refinements of the impersonal construction

- The impersonal **tough**-construction allows for an extra experiencer introduced by **for**.

(4) **It's Adj for X to VP**

a. **It**'s **tough** **for** Jo to solve this problem.

b. * **It**'s **brave** **for** Jo to point out the issue.

c. * **It**'s **pretty** **for** Jo to look at this vase.

d. * **It**'s **short** **for** Jo to ride the rollercoaster.

- The impersonal **brave**-construction, allows for an extra theme introduced by **of**.

(5) **It's Adj of X to VP**

a. * **It**'s **tough** **of** Jo to solve this problem.

b. **It**'s **brave** **of** Jo to point out the issue.

c. * **It**'s **pretty** **of** Jo to look at this vase.

d. * **It**'s **short** **of** Jo to ride the rollercoaster.

Four classes of adjectives, three contrasting templates

- Templates (2), (4) and (5) are sufficient to tease apart our adjectives.

Template		short	tough	pretty	brave
(2)	X is Adj to VP	*			
(4)	It's Adj for X to VP	*		*	*
(5)	It's Adj of X to VP	*	*	*	

- These distributional differences correlate with broad semantic differences.

Can LLMs leverage the distributional contrasts between these adjectives, to distinguish between them on psycholinguistics-inspired tasks?

Behavioral assessment

"Templatic" stimuli

- We focus on template (4).[6]

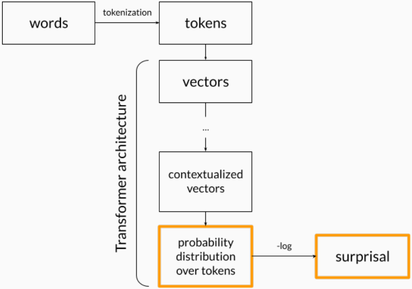
(4) It's $\left\{\begin{array}{l} ^{\checkmark}\textbf{tough} \\ ^{\times}\textbf{short} \\ ^{\times}\textbf{pretty} \\ ^{\times}\textbf{brave} \end{array}\right\}$ for **you** to **rible** this **zud**.

- We filled (4) with 64 adjectives (16 per class), 3 experiencer pronouns, 7 nonce verbs, 7 object nonce nouns.
- (4)+**tough** is **more grammatical** than (4)+{**short**, **pretty**, **brave**}.

[6] Templates (2) and (5) were also tested.

Surprisal as a dependent variable

- The **surprisal** $\mathscr{S}$ of a sentence is its **negative log probability**.
- In humans, word surprisal correlates with processing effort.[7]
- In LLMs, **surprisal differences** may reflect **grammatical contrasts**.[8]



$$\textbf{ungrammatical} \sim \textbf{unlikely} \iff \textbf{surprising}$$

$$* \qquad p \simeq 0 \qquad \mathscr{S} = -\log(p) \simeq \infty$$

[8]Hale, 2001; Levy, 2008

[8]See E. Wilcox et al. (2018), Futrell et al. (2019), and E. G. Wilcox et al. (2023). van Schijndel and Linzen (2021) and Arehalli et al. (2022) however suggest that LLM surprisal underestimates human slowdowns in garden-path effects.

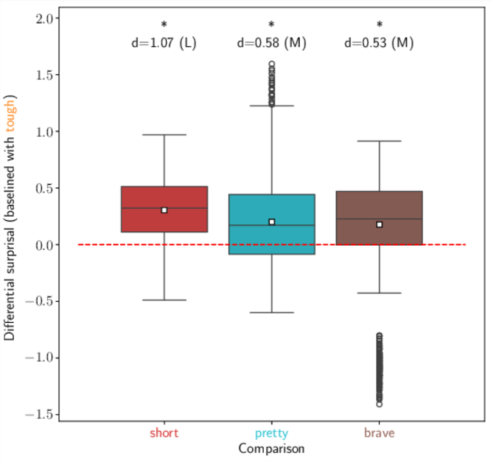
Prediction for template (4)

(4) It's $\left\{ \begin{array}{l} ^{\checkmark}\textbf{tough} \\ ^{\times}\textbf{short} \\ ^{\times}\textbf{pretty} \\ ^{\times}\textbf{brave} \end{array} \right\}$ for **you** to **rible** this **zud**.

- In template (4), **tough**-adjectives should be the **least surprising.**

$$\mathscr{S}(\text{It's } \textbf{short}/\textbf{pretty}/\textbf{brave} \text{ for } \textbf{you} \text{ to } \textbf{rible} \text{ this } \textbf{zud}.)$$
$$-\ \mathscr{S}(\text{It's } \textbf{tough} \text{ for } \textbf{you} \text{ to } \textbf{rible} \text{ this } \textbf{zud}.)$$
$$> 0$$

Focus on GPT-2 and template (4)



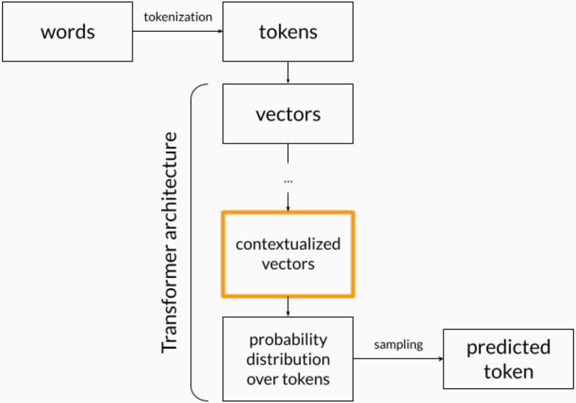
Differential surprisals (**short**, **pretty**, **brave** vs. **tough**) from GPT-2 Large. White squares display the means.

One-sided Wilcoxon test for matched pairs. *p*-values are Benjamini-Yekutieli-corrected. Effect sizes are Cohen's *d*. M=Medium, L=Large.

Neural assessment

What are LLMs' "neural" representations?

Does the "surprise" of LLMs correlate with their internal representation of the relevant adjectives?



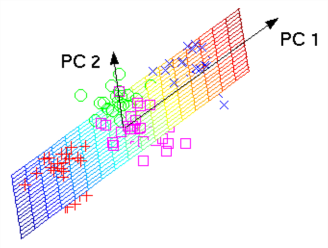
Rationale behind underlying vector spaces

- Pre-LLM underlying vector spaces have been shown to capture semantic relations between words, e.g. analogies of the form "*king* - *man* + *woman* = *queen*".[9]
- Do the vector spaces derived by our LLMs reflect a distinction between **tough**, **pretty**, **brave**, and **short**-like adjectives, in terms of **clustering**?

[9] Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Pennington et al., 2014.

Methodology

- Vectors extracted from the LLMs' penultimate layer and reduced via Principal Component Analysis (**PCA**).
- PCA **kills uninformative dimensions**.



GPT-2's internal representations

Study 1 shows distributional cues take us quite far!

- **Distributional information seems sufficient** to derive meaningful distinctions between **different categories of adjectives**.
- The models' performance contrasts with that of humans:
  - A classification of these adjectives in terms of e.g. polarity or subjectivity may seem more intuitive.
  - Some classes of adjectives appear in constructions that are comparatively late-acquired.[10]
- Did LLMs **learn something brand new**, or were they able to **efficiently encode** a pattern fully present in the input?
- We now investigate **another distinction between adjectives**, that is more primitive, but also more challenging to learn from a purely distributional perspective.

[10] Chomsky, 1969.

Study 2

Distinguishing antonymic adjectives through their behavior under negation.

Code; Surprisal dataset

Antonymic adjectives and the Distributional Hypothesis

- The opposition between positive and negative adjectives is **easy to learn** from an early age.[11]
- But antonyms occur in **very similar distributions**![12]
- Earlier neural networks could not capture anonymity.[13].

[11] Clark, 1972; Jones and Murphy, 2005.
[12] Charles and Miller, 1989; Justeson and Katz, 1991.
[13] Aina et al., 2019.

Positive and negative adjectives lead to distinct inferences

(6) a. Jo is **not tall**. $\rightsquigarrow$ Jo is fairly **short**.
"Inference Towards the Antonym" (ITA)[14]

b. Jo is **not short**. $\not\rightsquigarrow$ Jo is fairly **tall**.

- The ITA requires to distinguish between antonyms...
- but also, to grasp which antonym is **positive** and which one is **negative**, because they differentially interact with negation!
- This has to be done **in the absence of clear distributional cues**.

[14] Horn, 1989; Krifka, 2007; Ruytenbeek et al., 2017; Gotzner et al., 2018.

Behavioral assessment

Operationalizing the ITA contrasts in terms of surprisal

- The template in (7) captures ITA contrasts in terms of felicity.[15]

(7) a. He is not **tall**. She too is **short**.
Presupposes: not **tall** $\sim$ **short**.

b. # He is not **short**. She too is **tall**.
Presupposes: not **short** $\sim$ **tall**.

- These pairs allow us to reuse the "**differential surprisal**" methodology from Study 1.

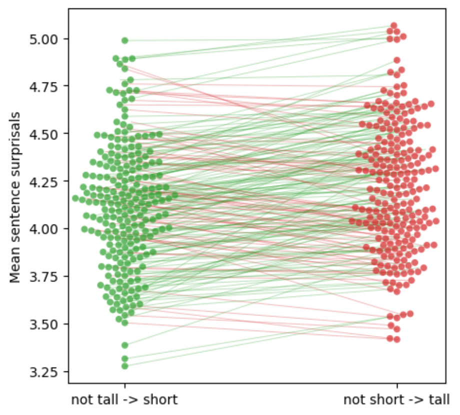
$$\mathscr{S}(7b) - \mathscr{S}(7a) > 0$$

- 111 antonymic pairs were used to measure this difference.

[15] Inspired by Ruytenbeek et al. (2017). Two other templates were tested, one where *too* appears after the second adjectives, and one "meta" template using the predicate *mean* to coordinate the two sentences.

GPT-2: sentence surprisals for $\mathscr{S}$(7b) and $\mathscr{S}$(7a)

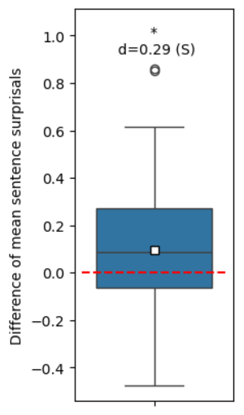
Prediction: $\mathscr{S}(7b) - \mathscr{S}(7a) > 0$



Surprisals of (7a) and (7b). Lines indicate minimal pairs. Green lines are ascending, i.e. are the ones for which the surprisal difference goes in the expected direction. Red lines are descending, and so go in the opposite direction.

GPT-2: differential surprisals for $\mathscr{S}$(7b) and $\mathscr{S}$(7a)

Prediction: $\mathscr{S}(7b) - \mathscr{S}(7a) > 0$



Paired differences in surprisal between (7b) and (7a). White squares display the means.
One-sided Wilcoxon test for matched pairs. Effect size is Cohen's *d*. S=Small.

Refining the ITA

- LLMs display surprisal contrasts that suggest they learned something about adjective polarity. How did they do this?
- **Morphologically transparent** antonyms, like **lucky**/**unlucky** were experimentally shown to display bigger ITA contrasts than **opaque** ones (like **tall**/**short**).[16]

(8) a. He is not **lucky**. She too is **unlucky**.

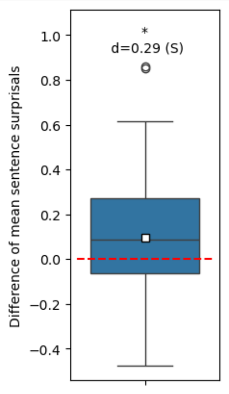
b. # He is not **unlucky**. She too is **lucky**.

- In transparent pairs, **polarity is distributionally encoded**, via a negative morpheme.

Can LLMs pick this up, and why would this matter?

[16] Ruytenbeek et al., 2017.

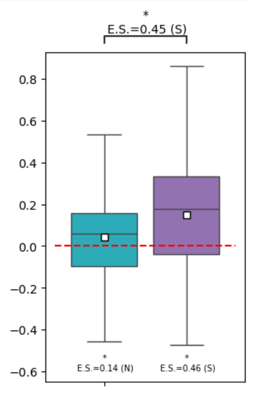
Results for GPT-2 (transparent vs. opaque)



Paired differences in surprisal between (7b) and (7a). All pairs together. White squares display the mean.

Splitting **transparent** (48) and **opaque** (63) pairs

➔



Paired differences in surprisal, **opaque** vs. **transparent** pairs. White squares are means.Within-group p-values are BY-corrected. Effect size is Cohen's d. N=Negligible.

- The ITA contrast is verified only for **transparent** adjectives!

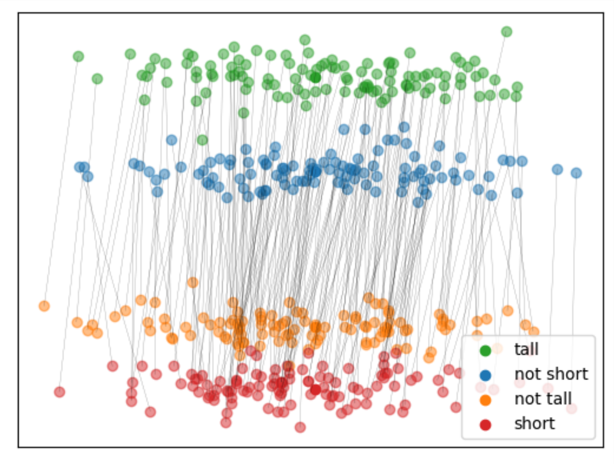
Neural assessment

Methodology

How do LLMs represent positive and negative adjectives, and the influence of negation?

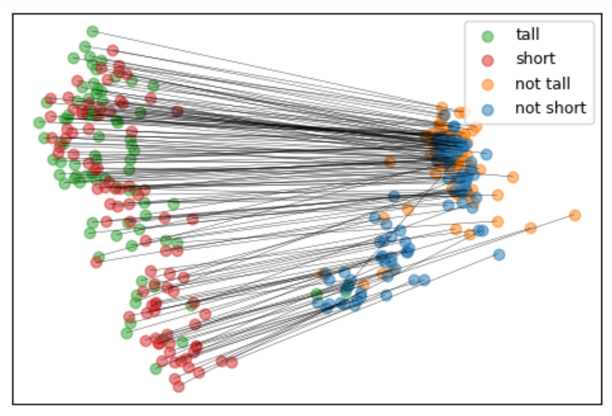
- Vectors for **positive** and **negative** adjectives were extracted, along with vectors for their respective negations (e.g. **not tall**/**not short**).
- Same methodology as in Study 1.

An idealized 2D vector space (Gaussians+Hungarian algorithm)



An idealized 2D representation of the (negated) adjectives' vectors. **tall**/**not short** and **short**/**not tall** cluster together, **short**/**not tall** being closer to each other than **tall**/**not short** due to the ITA. Dark line represent the effect of negation.

GPT-2's underlying vector space



GPT-2's vector space (after PCA). If a word was made of multiple tokens, its vector was computed as the mean of its tokens' vectors. Lines between points track the effect of negation, that is fairly stable, i.e. not contextualized.

- Antonyms and their negations cluster together!

Study 2 outlines the limits of distributional learning

- The ITA was only captured **when adjectives contained distributional information** (negative morphemes) indicating their polarity.
- Additionally, **the underlying vector space was characterized by counterintuitive topological regularities**, suggesting the functional nature of negation was not captured.
- This behavior contrasts with our intuitive understanding of antonyms, that we grasp from early toddlerhood, even for opaque pairs like **tall** and **short**.

Main takeaways from the Studies

- The two studies allow to disentangle the importance of **grounding** from that of **distributional information**, in the context of adjective learning.
- Study 1 showed that **distributional cues can take us far** in distinguishing between adjective meanings, even when the distinctions are intuitively not obvious.
- Study 2 showed that **when distributional information is not clearly present, adjectives cannot be properly distinguished**, even if the criterion is pretty obvious from a human perspective.
- The results of these studies **sharply contrast with children's acquisition of adjectives, and adult intuitions.**
- This suggests that grounding and social cues are crucial for word learning, even when the goal is to simply *distinguish* between meanings.

Zooming out: why these findings matter

- Study 1 and 2 point to important differences in the **environments** in which human and machine are learning, offering insights into:
  - **Human learning**: how much do linguistic biases, and extra-linguistic factors matter for language acquisition?
  - **Machine learning**: where should our efforts lie in improving the models?

Zooming out: fostering a productive interdisciplinary dialogue

- The tools presented today provide a **new type of testbed** for a number of questions that **matter to linguists**.[17]
- The linguistic datapoints we investigated may also **benefit computer scientists** who design and train LLMs, stressing the need for more grounded, reliable, and robust models of natural language.

[17] Study 1 in fact emerged from prior theoretical and experimental work of mine Hénot-Mortier et al., 2022; Hénot-Mortier, submitted

Thank you !

Selected references i

Harris, Z. S. (1954). **Distributional structure.** *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Rosenbaum, P. S. (1967). **The grammar of english predicate complement constructions [Doctoral dissertation, MIT].**

Chomsky, C. (1969). **The acquisition of syntax in children from 5 to 10.** *Research Monograph 57*.

Clark, E. V. (1972). **On the child's acquisition of antonyms in two semantic fields.** *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 750–758. https://doi.org/10.1016/s0022-5371(72)80009-4

Lasnik, H., & Fiengo, R. (1974). **Complement object deletion.** *Linguistic Inquiry*, *5*(4), 535–571. Retrieved April 15, 2022, from http://www.jstor.org/stable/4177842

Gleitman, L. R. (1981). **Maturational determinants of language growth.** *Cognition*, *10*(1–3), 103–114. https://doi.org/10.1016/0010-0277(81)90032-9

Charles, W. G., & Miller, G. A. (1989). **Contexts of antonymous adjectives.** *Applied Psycholinguistics*, *10*(3), 357–375. https://doi.org/10.1017/S0142716400008675

Selected references ii

Horn, L. R. (1989). **A natural history of negation.** University of Chicago Press.

Gleitman, L. (1990).**The structural sources of verb meanings.** *Language Acquisition*, *1*(1), 3–55. https://doi.org/10.1207/s15327817la0101_2

Naigles, L. (1990).**Children use syntax to learn verb meanings.** *Journal of Child Language*, *17*(2), 357–374. https://doi.org/10.1017/s0305000900013817

Justeson, J. S., & Katz, S. M. (1991).**Co-occurrences of antonymous adjectives and their contexts.** *Computational Linguistics*, *17*(1), 1–20. https://aclanthology.org/J91-1001

Stowell, T. (1991, December). **The alignment of arguments in adjective phrases.** In S. Rothstein (Ed.), *Perspectives on phrase structure: Heads and licensing* (pp. 105–135). Academic Press. https://doi.org/10.1163/9789004373198_007

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

Selected references iii

Snedeker, J., & Gleitman, L. R. (2004, January). **Why it is hard to label our concepts.** In *Weaving a lexicon* (pp. 257–294). The MIT Press. https://doi.org/10.7551/mitpress/7185.003.0012

Jones, S., & Murphy, M. L. (2005).**Using corpora to investigate antonym acquisition.** *International Journal of Corpus Linguistics*, *10*(3), 401–422. https://doi.org/10.1075/ijcl.10.3.06jon

Krifka, M. (2007). **Negated antonyms: Creating and filling the gap.** In U. Sauerland & P. Stateva (Eds.), *Presupposition and implicature in compositional semantics* (pp. 163–177). Palgrave Macmillan UK. https://doi.org/10.1057/9780230210752_6

Syrett, K. L. (2007). **Learning about the structure of scales: Adverbial modification and the acquisition of the semantics of gradable adjectives [Doctoral dissertation, Northwestern University].**

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

Yuan, S., Fisher, C., & Snedeker, J. (2012).**Counting the nouns: Simple structural cues to verb meaning.** *Child Development*, *83*(4), 1382–1399. https://doi.org/10.1111/j.1467-8624.2012.01783.x

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). **Efficient estimation of word representations in vector space.** https://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). **Distributed representations of words and phrases and their compositionality.** https://arxiv.org/abs/1310.4546

Pennington, J., Socher, R., & Manning, C. (2014, October). **GloVe: Global vectors for word representation.** In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Keine, S., & Poole, E. (2017).**Intervention in tough-constructions revisited.** *The Linguistic Review*, *34*(2), 295–329. https://doi.org/doi:10.1515/tlr-2017-0003

Ruytenbeek, N., Verheyen, S., & Spector, B. (2017).**Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives.** *Glossa: a journal of general linguistics*, *2*(1). https://doi.org/10.5334/gjgl.151

Selected references v

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017).**Attention is all you need.** *CoRR, abs/1706.03762*.

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

Gotzner, N., Solt, S., & Benz, A. (2018).**Adjectival scales and three types of implicature.** *Semantics and Linguistic Theory*, *28*, 409. https://doi.org/10.3765/salt.v28i0.4445

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018).**What do RNN language models learn about filler–gap dependencies?** *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211–221. https://doi.org/10.18653/v1/W18-5423

Aina, L., Bernardi, R., & Fernández, R. (2019).**Negated adjectives and antonyms in distributional semantics: Not similar?** *Italian Journal of Computational Linguistics*, *5*(1), 57–71. https://doi.org/10.4000/ijcol.457

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019).**Neural language models as psycholinguistic subjects: Representations of syntactic state.** *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 32–42. https://doi.org/10.18653/v1/N19-1004

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.** *arXiv preprint arXiv:1907.11692*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019).**Language models are unsupervised multitask learners.**

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). **Xlnet: Generalized autoregressive pretraining for language understanding.** In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, neurips 2019, december 8-14, 2019, vancouver, bc, canada* (pp. 5754–5764). https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

Bender, E. M., & Koller, A. (2020, July). **Climbing towards NLU: On meaning, form, and understanding in the age of data.** In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

van Schijndel, M., & Linzen, T. (2021).**Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty.** *Cognitive Science, 45*(6). https://doi.org/10.1111/cogs.12988

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). **Flamingo: A visual language model for few-shot learning.** https://arxiv.org/abs/2204.14198

Arehalli, S., Dillon, B., & Linzen, T. (2022, December). **Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities.** In A. Fokkens & V. Srikumar (Eds.), *Proceedings of the 26th conference on computational natural language learning (conll)* (pp. 301–313). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.conll-1.20

Gotowski, M. (2022). **Syntactic bootstrapping in the adjectival domain: Learning subjective adjectives [Doctoral dissertation, Rutgers University].**

Hénot-Mortier, A., Stacey, R., Torma, C., & Aravind, A. (2022). **Two kinds of adjective-infinitive constructions in acquisition [Architectures and Mechanisms of Language Processing 2022 (AMLaP 28)].** https://adelemortier.github.io/files/AMLaP%5C_2022%5C_slides.pdf

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

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). **Mistral 7b.** https://arxiv.org/abs/2310.06825

Nair, S., & Resnik, P. (2023, December). **Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship?** In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: Emnlp 2023* (pp. 11251–11260). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.752

- Wilcox, E. G., Futrell, R., & Levy, R. (2023).**Using Computational Models to Test Syntactic Learnability.** *Linguistic Inquiry*, 1–44. https://doi.org/10.1162/ling_a_00491
- Peng, B., Narayanan, S., & Papadimitriou, C. (2024). **On limitations of the transformer architecture.** https://arxiv.org/abs/2402.08164
- Hénot-Mortier, A. (submitted).**It's tough to be pretty: On the semantic relatedness between tough and pretty predicates.** *Glossa*.

Appendices

Links to Appendix slides

- General supplementary slides
- Supplementary slides Study 1
- Supplementary slides Study 2