Methods to Evaluate Verb Argument Structure Knowledge in Language Models and Embeddings

James v. Bruno, David Yi, Jiayu Han, Peter Zukerman

Probing Linguistic Knowledge in Neural Networks



Motivations:

- Does neural networks encode enough grammatical information?
- If so, what extent do the features learned by neural networks resemble the linguistic competence of humans?

Why do we want to probe?

- Would be helpful in downstream tasks
- Analyses of results can contribute to the scientific questions in linguistics: the role of prior grammatical bias in human language acquisition.

Roadmap

- Verb Argument Structure Alternations
- Probing Linguistics Knowledge of verbs in Embeddings
- Probing Linguistics Knowledge in Pretrained Language Models
- Our work
- Q&A

Verb Argument Structure Alternations

Levin (1993) comprehensively describes many classes of Verb Argument Structure Alternations.

1 example (out of many): The Spray/Load Alternation

- a. Lucy sprayed the wall with paint.
 b. Lucy sprayed paint on the wall.
- 2) a. Lucy covered the towel with sand.b. *Lucy covered sand on the towel

The Spray-Load Alternation (Arad, 2006)

- (1) a. Lucy sprayed the paint on the wall.
 - b. Lucy sprayed the wall with paint.
- (2) a. Ben loaded hay on the truck.
 - b. Ben loaded his truck with hay.
- (3) Jan plant bomen in de tuin. a. John plants the Garden. trees in tuin met bomen. b. Jan beplant de John be-plants the garden with trees.

- (4) a. kabe ni penki o nuru wall on paint ACC Paint (VERB) 'smear paint on the wall'
 - b. kabe o penki de nuru wall ACC paint with Paint (VERB) 'smear the wall with paint'

- (5) a. János rámázolta a festéket a falra.
 John onto-smeared-he-it ACC Paint-ACC the wall-onto 'John smeared paint on the wall.'
 - b. János bemázolta a falat festékkel. John in-smeared-he-it ACC The wall Paint-with 'John smeared the wall with paint.'

Argument-structure based account (Levin and Rappaport, 1988)

- (8) LOAD: <Agent, Locatum, Goal>
- (9) a. LOAD: x < y, $P_{loc}z >$ b. LOAD: x < y, $P_{with}z >$

Linking Rules

Lexical Entry

(10) LOAD_a: <Agent, Locatum, Goal> (locative variant) LOAD_b: <Agent, Theme, Locatum> (*with*-variant)

Or just two different entries altogether?

How about a richer

semantic representation?

- (14) a. LOAD: [x cause [y to come to be at z] / LOAD]
 - b. LOAD: [[*x* cause [*z* to come to be in a STATE]] BY MEANS OF [*x* cause [*y* to come to be at *z*]] / LOAD]

The Aspectual Interface Hypothesis (Tenny, 1987)

- The direct object "measures out the event".
 - E.g. for *eat an apple*, the eating event is over when the apple is consumed.
- "Load verbs denote an event that can be measured out in two different ways both by the Theme and by the Goal. Since measuring out is associated with direct objects, either the Theme or the Goal may be realized as direct objects"
- Lucy sprayed the wall with paint.
 - The spraying event is measured out according to the status of the wall.
- Lucy sprayed paint on the wall.
 - The spraying event is measured out according to the status of the paint.

The Aspectual Interface Hypothesis (Tenny, 1987)

- Lucy covered the towel with sand.
 - The covering event is measured out according to the status of the towel.
- *Lucy covered sand on the towel
 - The covering event cannot be measured out according to the status of the sand. The towel is either covered, or it's not.

What are the Lexical Properties of Spray-Load Verbs? (Pinker, 1989)

Ingredient 1: In general, the location has be construeable as undergoing a change of state.

- a. Lucy sprayed the wall with paint.b. Lucy sprayed paint on the wall.
- 4) a. Lucy covered the towel with sand.b. *Lucy covered sand on the towel

What are the Lexical Properties of Spray-Load Verbs? (Pinker, 1989)

Ingredient 2: Content-Oriented vs. Container-Oriented

- Content-Oriented verbs obligatorily take a locatum, with an optional location.
 - (20) a. Lucy piled the books (on the shelf).
 - b. Lucy piled the shelf *(with books).
- Container-Oriented verbs obligatorily take a location, with an optional locatum.
 - (21) a. Lucy stuffed the turkey (with breadcrumbs).
 - b. Lucy stuffed the breadcrumbs *(into the turkey).

What are the Lexical Properties of Spray-Load Verbs? (Pinker, 1989)

Those ingredients allow us to say:

- Container-oriented verbs that alternate must specify not only the change of state in the container, but also the manner in which the substance is moved into the location
 - *Lucy covered sand on the towel
 - Lucy stuffed breadcrumbs into the turkey.
- Content-oriented verbs that alternate must specify not only the manner in which the substance is moved, but also the change of state in the location
 - Lucy piled the shelf with books.
 - *Lucy poured the glass with water.

So what's really important here, in the context of 575?

- Verb argument structure alternations are a lexical property of the verb.
- Verb argument structure alternations are identifiable by the kinds of tokens in the neighborhood of the verb. *"You shall know a word by the company it keeps"* (Firth, 1957).

• That suggests that the alternation class may be encoded in embeddings.

Verb Argument Structure Alternations in Word and Sentence Embeddings

Verb Alternation Classes

Verb Frame	Example Sentences		
Caus.	Jessica dropped the vase.	Jessica blew the bubble.	
Inch.	The vase dropped .	*The bubble blew.	
Dative-Prep.	Liz gave a gift to the boy.	Liz administered a test to the kid.	*Liz charged \$50 to Jon.
Dative-2-Obj.	Liz gave the boy a gift.	*Liz administered the kid a test.	Liz charged Jon \$50.
SprLowith	Sue loaded the truck with wood.	Sue coated the deck with paint.	*Sue swept the bin with sand.
SprLoLoc.	Sue loaded wood onto the truck.	*Sue coated paint on the deck.	Sue swept sand into the bin.
no-there	Fear remained in my mind.	A girl focused on the quiz.	
there	There remained fear in my mind.	*There focused on the quiz a girl.	
UObjRefl.	Ada clapped her hands.	Ada permed her hair.	*Ada exercised herself.
UObjNo-Refl.	Ada clapped.	*Ada permed.	Ada exercised.

*Important Note: The sentences are formed in such a way that only the **main verb** alternation information determines grammaticality judgements.

LaVA Dataset

The LaVA (Lexical Verb-Frame Alternations) dataset includes 515 verbs annotated for membership in 10 verb frame classes.

Human annotations note 1 for membership, 0 for non-membership, and 'x' where membership is unknown (or non-existent).

verb	sl	sl_noloc	sl_nowith	inch	non_inch	there	non_there	dat_both	dative_to	dat_do	refl_op	refi_only
fed	0	0	0	0	1	0	1	1	0	0	0	0
served	0	0	0	0	x	0	x	1	0	0	0	0
gave	0	0	0	0	x	0	x	1	0	0	0	0
left	0	0	0	0	x	0	x	1	0	0	0	0

LaVA Dataset

The LaVA corpus presents 5 of the largest syntactic verb frame alternations provided by Levin (1993):

- Causative-Inchoative
- Dative
- Spray-Load (as seen earlier)
- there-Insertion
- Understood-Object

On Sparsity:

Due to how verb argument structure alternations function, negative samples can not always be obtained. For example, no English verbs can appear in the inchoative but not the causative. There are also no verbs that can only appear in the *there* frame but not the no-*there*. This leads to sparsity in annotations, which causes trivial word-level classifications.

Experiment 1: From Word Embeddings to Argument Structures

Objective

- For each alternation class, build a multi-label classifier that predicts whether a verb participates in a particular syntactic frame

 $p(s) = \sigma(W_2(f(W_1x)))$

Modeling Details

- Input (x): Word embedding representation of verb v
- Alternation Class: causative-inchoative
- Syntactic frame (s): Inchoative
- Output p(s): Probability that verb v participates in frame s
- **Training**: Single-layer MLP with 4-fold Cross Validation

LaVA Dataset

Levin class	CAUS.	-INCH.	DATIVE		SPRAY	-LOAD	there-INSERTION		UNDERSTOOD-OBJEC	
	Inch.	Caus.	Prep.	2-Obj.	with	Loc.	no-there	there	Refl.	No-Refl.
Positive	70	120	63	72	90	81	50	145	11	81
Negative	140	(0)	356	405	220	229	185	(0)	466	396
Total	210	120	419	477	310	310	235	145	477	477

LaVA Dataset (Kann et al. 2019)

Why does Causative (NEG) have 0 examples?

The vase **dropped** (inchoative) /Jessica **dropped** the vase (causative) * The bubble **blew** (inchoative) / Jessica **blew** the bubble (causative)

Where do the word embeddings come from?

Word Embeddings

- GloVe Embeddings: 300d embeddings trained on 6B Tokens
- **Custom** Embeddings: Trained on **100M** tokens from the British National Corpus (BNC) using a single-directional LSTM w/ LM Objective

Why these Embeddings?

- Trained on similar amount of data that humans are exposed to during language acquisition
- Large pretrained models (i.e. BERT) trained on "several orders of magnitude more data than humans see in a lifetime" than custom embeddings
 - 3.3B tokens v.s. 100M

Evaluation: Matthew's Correlation Coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Why MCC?

- Special case of Pearson's Correlation Coefficient for Binary Classification
- Generalizes better to imbalanced distributions than accuracy/F1-score
 - -1: Complete disagreement between predictions and observations
 0: Average score of two unrelated distributions
 - 1: Perfect correlation between predictions and observations

Results

LaVA Results (Kann et al. 2019)

		CAUSATIVE Inch.	–INCHOATIVE (Caus.)	1.1.1	tive 2-Obj.			o there-IN: no-there		100 B	RSTOOD-OBJECT Non-Refl.
CoLA: Majority BL	Acc.	66.7	(100.0)	85.0	84.9	71.0	73.9	78.7	(100.0)	97.7	83.0
CoLA: MLP		0.555 81.0	0.0 (100.0)	0.32 86.6			0.253 72.9	0.459 84.3	0.0 (100.0)	0.0 97.7	0.219 79.0
GloVe: Majority BL	Acc.	66.8	(100.0)	85.0	85.3	71.0	74.6	79.1	(100.0)	97.6	81.5
GloVe: MLP	MCC Acc.	0.672 85.5	0.0 (100.0)	0.0 85.0		0.585 83.9	0.145 73.4	0.536 85.8	0.0 (100.0)	0.0 97.6	0.3 73.2



The FAVA (Frame and Alternations of Verbs Acceptability) dataset consists of ~10,000 sentences containing the verbs in LaVA in different verb frames and labeled for grammaticality.

Annotations are 1 for accepted and 0 for unaccepted sentences.

dat	0	christopher tipped a week 's salary to james .
dat	1	christopher tipped james a week 's salary .
dat	0	jason tipped 20 pounds to rebecca .
dat	1	jason tipped rebecca 20 pounds .
inch	1	rebecca steered the car .
inch	1	the car steered .
inch	1	rebecca steered the bicycle .
inch	1	the bicycle steered .
inch	1	james steered the truck .

Detour: CoLA Dataset(Corpus of Linguistic Acceptability)

Label	Sentence	Source
*	The more books I ask to whom he will give, the more he reads.	Culicover and Jackendoff (1999)
1	I said that my father, he was tight as a hoot-owl.	Ross (1967)
1	The jeweller inscribed the ring with the name.	Levin (1993)
*	many evidence was provided.	Kim and Sells (2008)

Data: 10,657 sentences labeled for grammatical acceptability that analyze different types of linguistic phenomena

- 17 in-domain, 6 out-of-domain

A few examples:

- Comparatives (Culicover and Jackendoff, 1999)
- Islands (Ross, 1967)
- Verb Alternations (Levin, 1993)
- General syntax (Kim and Sells, 2008)

Experiment 2: Sentence Embedding Probing

- Linguists would classify a word by interrogating whether sentences with a given verb and frame are acceptable.
- Analogously, a MLP model is used to calculate the probability that a sentence S is acceptable.

 $p(S) = \sigma(W_2(tanh(W_1x)))$

Where do these Sentence Embeddings come from?

Sentence Encoder trained by Warstadt et al. (2018) on "Real/Fake" discrimination task for downstream CoLA task

Input: ELMo-style Word Embeddings

Training (12M sentences)

- Real: 6M sentences from BNC(British National Corpus)
- Fake: 3M generated by LSTM + 3M permuted from BNC

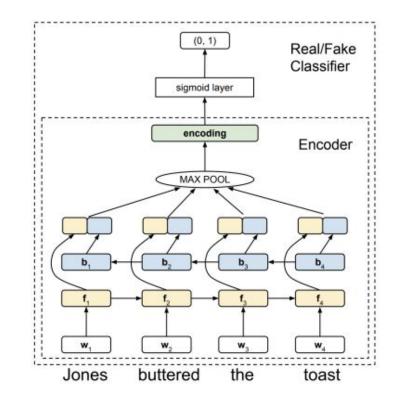


Figure 1: Real/fake model. w_i = word embeddings, f_i = forward LSTM hidden state, b_i = backward LSTM hidden state. Figure from Warstadt et al. (2018).

Results

		Comb.	CAUSATIVE-INCHOATIVE	DATIVE	Spray–Load	there-INSERTION	UNDERSTOOD-OBJECT
w/o CoLA	MCC	0.290	0.603	0.413	0.323	0.528	0.753
	Acc.	64.6	85.4	76.0	66.2	72.9	87.4
w/ CoLA	MCC	0.361	0.464	0.329	0.261	0.523	0.638
	Acc.	68.7	81.2	59.0	63.4	72.5	81.8
Majority BL	MCC	0.0	0.0	0.0	0.0	0.0	0.0
	Acc.	66.6	77.6	82.1	60.3	77.5	53.7

1. Easiest alternation was UNDERSTOOD-OBJECT

alternation

Blink -> her eyes, Clap -> his hands, etc..

2.Alterations involving **only transitive** verb frames (i.e. SPRAY-LOAD) were generally **more difficult** than those with **at least one intransitive frames** (i.e.CAUSATIVE-INCHOATIVE)

3.**No relationship** between **# training examples** and performance

Verb Frame	Example Sentences		
Caus.	Jessica dropped the vase.	Jessica blew the bubble.	
Inch.	The vase dropped .	*The bubble blew.	
Dative-Prep.	Liz gave a gift to the boy.	Liz administered a test to the kid.	*Liz charged \$50 to Jon.
Dative-2-Obj.	Liz gave the boy a gift.	*Liz administered the kid a test.	Liz charged Jon \$50.
SprLo <i>with</i>	Sue loaded the truck with wood.	Sue coated the deck with paint.	*Sue swept the bin with sand
SprLoLoc.	Sue loaded wood onto the truck.	*Sue coated paint on the deck.	Sue swept sand into the bin.
no-there	Fear remained in my mind.	A girl focused on the quiz.	
there	There remained fear in my mind.	*There focused on the quiz a girl.	
UObjRefl.	Ada clapped her hands.	Ada permed her hair.	*Ada exercised herself.
UObjNo-Refl.	Ada clapped.	*Ada permed.	Ada exercised.

Takeaways

 Models achieved moderate correlation (0.5-0.7) in 5/12 acceptability experiments, all except one achieved > 0.3

- Easiest alternation was UNDERSTOOD-OBJECT alternation
 - Blink -> her eyes, Clap -> his hands, etc..
- Alterations involving only *transitive* verb frames (i.e. SPRAY-LOAD) were generally more difficult than those with at least one *intransitive* frames (i.e. CAUSATIVE-INCHOATIVE)

- No relationship between **# training examples** and performance

Takeaways

- Pros:
 - word-level and sentence-level datasets: LaVA, FAVA
 - Probing word embeddings
 - Probing sentence embeddings
- Cons:
 - FAVA is not from natural sentences
 - Not easy to tell linguistic knowledge is in neural networks or in the probing models

BliMP: The Benchmark of Linguistic Minimal Pairs for English

BliMP: The Benchmark of Linguistic Minimal Pairs

Motivations:

- Existed evaluating datasets only focus on a small set of linguistic phenomena
- Probing by additional models cannot tell whether the linguistic knowledge is in the Neural Networks

Minimal Pairs + New probing paradigm

Minimal Pairs: Pairs of minimally different sentences that contrast in grammatical acceptability and isolate specific phenomenon in syntax, morphology, or semantics.

a. The cats annoy Tim. (grammatical)

b. * The cats annoys Tim. (ungrammatical)

New probing paradigm: probing LMs without an additional supervised model

 Observe whether LMs assign a higher probability to the acceptable sentence in each minimal pairs

BliMP Dataset Overview

• 12 linguistic phenomenon categories, 67 individual datasets(different linguistic paradigms), each containing 1000 minimal pairs.

Phenomenon	Ν	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	Many girls insulted themselves.	Many girls insulted herself.
ARG. STRUCTURE	9	Rose wasn't disturbing Mark.	Rose wasn't boasting Mark.
BINDING	7	Carlos said that Lori helped <u>him</u> .	Carlos said that Lori helped himself.
CONTROL/RAISING	5	There was bound to be a fish escaping.	There was <u>unable</u> to be a fish escaping.
DETNOUN AGR.	8	Rachelle had bought that chair.	Rachelle had bought that chairs.
Ellipsis	2	Anne's doctor cleans one important	Anne's doctor cleans one book and
		book and Stacey cleans a few.	Stacey cleans a few important.
FILLER-GAP	7	Brett knew what many waiters find.	Brett knew that many waiters find.
IRREGULAR FORMS	2	Aaron broke the unicycle.	Aaron broken the unicycle.
ISLAND EFFECTS	8	Which <u>bikes</u> is John fixing?	Which is John fixing <u>bikes</u> ?
NPI LICENSING	7	The truck has clearly tipped over.	The truck has <u>ever</u> tipped over.
QUANTIFIERS	4	No boy knew fewer than six guys.	No boy knew <u>at most</u> six guys.
SUBJECT-VERB AGR.	6	These casseroles disgust Kayla.	These casseroles disgusts Kayla.

Table 1: Minimal pairs from each of the twelve linguistic phenomenon categories covered by BLiMP. Differences are underlined. N is the number of 1,000-example minimal pair paradigms within each broad category.

Warstadt et al. (2020)

BliMP Dataset Overview

	UID	5-812	IST LST	151	GP1	2 Hum	Acceptable Example	Unacceptable Example	1pfx
ANAPHOR AGREEMENT	anaphor_gender_agreement anaphor_number_agreement	44 52	88 95	91 97	99 100	96 99	Katherine can't help herself . Many teenagers were helping themselves .	Katherine can't help himself. Many teenagers were helping herself .	
	animate_subject_passive	54	68	58	77	98	Amanda was respected by some waitresses.	Amanda was respected by some picture.	1
	animate_subject_trans	72	79	70	80	87	Danielle visited Irene.	The eye visited Irene.	
	causative	51	65	54	68	82		Aaron appeared the glass.	
RGUMENT	drop_argument	68	79	67	84	90	The Lutherans couldn't skate around.	Aaron appeared the glass. The Lutherans couldn't disagree with.	
	inchoative	89	72	81	90	95	A screen was fading.	A screen was cleaning.	
TRUCTURE	intransitive	82	73	81	90	86	Reasonable design and the second states of the seco		
	passive_1	71	65	76	89	99	Some glaciers are vaporizing. Jeffrey's sons are insulted by Tina's supervisor.	Some graciers are scaring. Jeffrey's sons are smiled by Tina's supervisor.	
	passive_1 passive_2	70	72	74	79	86	Most cashiers are disliked.	March and have and Alated	
	passive_2 transitive	91	87	89	49	87	A lot of actresses' nieces have toured that art gallery.	Most cashiers are flirted. A lot of actresses' nieces have coped that art gallery.	
	transitive	91	-		49		A lot of actresses' meces have toured that art gallery.		
	principle_A_c_command	58	59	61	100	86	A lot of actresses that thought about Alice healed themselves.	A lot of actresses that thought about Alice healed herself.	1
	principle_A_case_1	100	100	100	96	98	Tara thinks that she sounded like Wayne.	Tara thinks that herself sounded like Wayne.	
	principle_A_case_2	49	87	95	73	96	Stacy imagines herself praising this actress.		
INDING	principle_A_domain_1	95	98	99	99	95	Carlos said that Lori helped him.	Carlos said that Lori helped himself.	1
	principle_A_domain_2	56	68	70	99 73	75	Mark imagines Erin might admire herself.	Carlos said that Lori helped himself. Mark imagines Erin might admire himself.	~
	principle_A_domain_3	52	55	60	82	83	Nancy could say every guy hides himself.	Mark imagines Erin might admire himself . Every guy could say Nancy hides himself .	
	principle_A_reconstruction	40	46	38	37	78	It's herself who Karen criticized.	It's herself who criticized Karen.	
	existential there object raising	84			92	0.0	William has declared there to be no guests getting fired.	William has obliged there to be no guests getting fired.	
	existential_there_object_raising existential_there_subject_raising	84	66 80	76 79	92 89	90 88	William has declared there to be no guests getting fired. There was bound to be a fish escaping.	William has obliged there to be no guests getting fired. There was unable to be a fish escaping.	
ONTROL/	expletive_it_object_raising	72	63	72	58	86	Regina wanted it to be obvious that Maria thought about Anna.	Regina forced it to be obvious that Maria thought about Anna.	
ISING	exprense_n_ooject_raising	33	34	45	72	75	Lake mean it for to talk to		
	tough_vs_raising_1 tough_vs_raising_2	33	34 93	45	72 92	75 81	Julia wasn't fun to talk to. Rachel was apt to talk to Alicia.	Julia wasn't unlikely to talk to. Rachel was exciting to talk to Alicia.	
	tougn_vs_raising_2	-n	93	80	- 92	81	Racnei was apt to taik to Alicia.	Rachel was exciting to talk to Aficia.	
	determiner_noun_agreement_1	88	92	92	100	96	Craig explored that grocery store.	Craig explored that grocery stores.	1
	determiner noun agreement 2	86	92	81	93	95	Carl cures those horses.	Craf curs that horses. Card curs that horses. Phillip was lifting this mice . Those ladies walk through that oases . Tracy prejuses those lucky urys .	
ETER-	determiner_noun_agreement_irregular_1	85	82	88	94	92	Phillip was lifting this mouse.	Phillin was lifting this mice.	·····
INER-	determiner_noun_agreement_irregular_2	90	86	82	94 93	85	Those ladies walk through those oases.	Those ladies walk through that passes	
DUN	determiner_noun_agreement_with_adj_1	50	86	78	90	96	Tracy praises those lucky guys.	Tracy praises those lucky guys.	
JR.	determiner_noun_agreement_with_adj_2	53	76	81	96	94	A		
JK.	determiner_noun_agreement_with_adj_2 determiner_noun_agreement_with_adj_irregular_1	55	70	01		85	Some actors buy these gray books.	Some actors buy uns gray books.	
		52	83 87	77 86	88 93		Some actors buy these gray books. This person shouldn't criticize this upset child.	Some actors buy this gray books. This person shouldn't criticize this upset children.	· · · · · · · · · · · · · · · · · · ·
	determiner_noun_agreement_with_adj_irregular_2	52	87	86	93	95	That adult has brought that purple octopus.	That adult has brought those purple octopus.	
LLIPSIS	ellipsis_n_bar_1 ellipsis_n_bar_2	23 50	68 67	65 89	88 86	92 78	Brad passed one big museum and Eva passed several. Curtis's boss discussed four sons and Andrew discussed five sick sons.	Brad passed one museum and Eva passed several big. Curtis's boss discussed four happy sons and Andrew discussed five sick.	
	wh questions object gap	53	79	61	84	85	Joel discovered the vase that Patricia might take.	Joel discovered what Patricia might take the vase.	
		82	92	83	95	98	Chervl thought about some dog that upset Sandra.	Chand down by share when some days much be	
	wn_questions_subject_gap	86	94	86	88	85		Bruce knows who that person that Dawn likes argued about a lot of guys.	
ILLER	wh_questions_subject_gap wh_questions_subject_gap_long_distance	86	96	86	88		Bruce knows that person that Dawn likes that argued about a lot of guys.	Bruce knows who that person that Dawn likes argued about a lot of guys.	
AP	wh_vs_that_no_gap	83	96 97 97	86	97	97	Danielle finds out that many organizations have alarmed Chad.	Danielle finds out who many organizations have alarmed Chad.	
	wh_vs_that_no_gap_long_distance	81	97	91	94	92	Christina forgot that all plays that win worry Dana.	Christina forgot who all plays that win worry Dana.	
	wh_vs_that_with_gap	18	43	42	56	77	Nina has learned who most men sound like.	Nina has learned that most men sound like.	
	wh_vs_that_with_gap_long_distance	20	14	17	56	75	Martin did find out what every cashier that shouldn't drink wore.	Martin did find out that every cashier that shouldn't drink wore.	
RREGULAR	irregular_past_participle_adjectives irregular_past_participle_verbs	79 80	93 85	91 66	78 90	99 95	The forgotten newspaper article was bad. Edward hid the cats.	The forgot newspaper article was bad. Edward hidden the cats.	
		48	_	_					-
	adjunct_island complex NP island	48	67 47	65 58	91 72	94 80	Who has Colleen aggravated before kissing Judy? Who hadn't some driver who would fire Jennifer's colleague embarrassed?	Who has Colleen aggravated Judy before kissing? Who hadn't Jennifer's colleague embarrassed some driver who would fire?	
	coordinate_structure_constraint_complex_left_branch	32	30	36	42	90	Who hadn't some driver who would fire Jennifer's colleague embarrassed? What lights could Spain sell and Andrea discover?	Who hadn't Jennifer's colleague embarrassed some driver who would hre? What could Spain sell lights and Andrea discover?	
	coorumate_structure_constraint_complex_left_branch	32	30	36 74	42		what rights could Spain sell and Andrea discover?	what could spain sell lights and Andrea discover?	
					88	91	Who will Elizabeth and Gregory cure?	What could span son igns and Andrea discover: Who will Elizabeth cure and Gregory?	
	coordinate_structure_constraint_object_extraction	59	71					What would David cure snake?	
	coordinate_structure_constraint_object_extraction left_branch_island_echo_question	96	32	63	77	91	David would cure what snake?	what would David cure shake?	
	coordinate_structure_constraint_object_extraction left_branch_island_echo_question left_branch_island_simple_question	96 57	32 36	63 36	77 82	99	Whose hat should Tonya wear?	ND 1177 1	
	coordinate_structure_constraint_object_extraction left_branch_island_echo_question left_branch_island_simple_question sentential_subject_island	96 57 61	32 36 43	63 36 37	35	99 61	Whose hat should Tonya wear? Who have many women's touring Spain embarrassed.	Whose should Tonya wear hat? Who have many women's touring embarrassed Spain.	
	coordinate_structure_constraint_object_extraction left_branch_island_echo_question left_branch_island_simple_question	96 57	32 36	63 36	77 82 35 77	99	Whose hat should Tonya wear?	Whose should forware that we have a share? Whose should forware wear hat? Who have many women's touring embarrassed Spain. What could Alan discover who has run around?	~
	coordinate_structure_constraint_object_extraction left_branch_island_ecbo_question left_branch_island_simple_question sentential_subject_island wh_island	96 57 61	32 36 43	63 36 37	35 77	99 61	Whose hat should 'Tonya wear? Who have many women's touring Spain embarrassed. What could Alan discover he has run around?	Whose should Tonya wear hat? Who have many women's touring embarrassed Spain. What could Alan discover who has run around?	~
	coordinate_structure_constraint_object_extraction left_branch_island_sche_quession left_branch_island_simple_quession sentential_subject_island wh_island matrix_question_npi_licensor_present	96 57 61	32 36 43 47 2	63 36 37 20	35 77 67	99 61 73 98	Whose hat should Tonya wear? Who have many women's touring Spain embarrassed. What could Alan discover he has run around? Should Monica ever grin?	Whose should Tonya wear hat? Who have many women's touring embarrassed Spain. What could Alan discover who has run around? Monica should ever grin.	
FECTS	coordinate_structure_constraint_object_extraction left_branch_island_echc_question left_branch_island_simple_question semtential_subject_island wh_island matrix_question_npi_licensor_present mpi_present_1	96 57 61 56	32 36 43	63 36 37 20 1 61	35 77 67 55	99 61 73	Whose hat should Torya wear? Who have many women's touring Spain embarrassed. What could Alan discover he has run around? Should Monica ever grin? Even these trucks have often alowed.	Whose should Tonya wear hat? Who have many women's touring embarrassed Spain. What could Alan discover who has run around? Monica should ever grin. Even these trucks have ever slowed.	
PI	coordinate_structure_constraint_object_extraction left_branch_island_echc_question left_branch_island_simple_question semtential_subject_island wh_island matrix_question_npi_licensor_present mpi_present_1	96 57 61 56 1 47 47	32 36 43 47 2 54 54 54	63 36 37 20 1 61 48	35 77 67 55 62	99 61 73 98 83 98	Whose has should Tonya wear? Who have many sources is couring Spain embarrassed. What could Alan discover he has run around? Should Monica ever grin? Even these trucks have Ofen alowed. Many Auchematic also roll.	Whose should Tonya wen haf? Who have many women's touring emburassed Spain. What could Alan discover who has run around? Monies should ever grin. Even these trucks have ever slowed. Many skateboards ever roll.	
FECTS	coordinate_structure_constrain_object_extraction left_branch_island_eches_question sentential_subject_island wh_island matrix_question, npi_licensor_present matrix_question, npi_licensor_present matrix_clessor_present	96 57 61 56 1 47 47 47 57	32 36 43 47 2 54 54 54 93	63 36 37 20 1 61 48 80	35 77 67 55 62 100	99 61 73 98 83 98 98 92	Whose has should Tonya wear? Who have many sources is couring Spain embarrassed. What could Alan discover he has run around? Should Monica ever grin? Even these trucks have Ofen alowed. Many Auchematic also roll.	Whose should Tonya wen haf? Who have many women's touring emburassed Spain. What could Alan discover who has run around? Monies should ever grin. Even these trucks have ever slowed. Many skateboards ever roll.	
PECTS	coordinate_structure_constrain_object_extraction left_branch_island_eches_question sentential_subject_island wh_island matrix_question, npi_licensor_present matrix_question, npi_licensor_present matrix_clessor_present	96 57 61 56 1 47 47 47 57 30	32 36 43 47 2 54 54 54 93 36	63 36 37 20 1 61 48 80 45	35 77 67 55 62 100 85	99 61 73 98 83 98 92 72	Whose has should Tonya wear? What could Alan discover he has run around? Should Monie or engris of the structure of the str	Whose theolad Tonya wear hard? Who have many sources 1 source genebarassed Spain. What could Alan discover who has run around? Manica should are grain. Even those turks have even slowed. Many skateboards ever roll. Even fall would ever compain. Those doctors who only Karla respects ever conceal many anakes.	ý
PECTS	coordinate.structure_constraint_object_extraction left_branch_island_sche_question estentorial_ubject_island wh_stand natrix_question_npl_icensor_present npl_present_1 npl_present_2 nubj_proj.coge estential_ubjection_proj.censor_present	96 57 61 56 1 47 47 47 57 30 93	32 36 43 47 2 54 54 93 36 100	63 36 37 20 1 61 48 80 45 99	35 77 67 55 62 100 85 89	99 61 73 98 83 98 92 72 93	Whose has should Tonya wear? What could Alan discover he has run around? What could Alan discover he has run around? Should Monica were grin? Even these trucks have often showed. Many skateboards also roll. Only Bill would ever complain, agents ever conceal many indust. Those banks had not ever liced.	Whose should Tonya wear hat? Who have many women's storing embarased Spain. What could Alin discover who has run around? Monica should ever grift. Even these tracks have ever alowed. Many skateboardwe ever oil. Even fill would ever companie. To bose basks hat mostly wear first.	
PI	coordinate.structure_constrain_object_extraction left_branch_island_exbo_question elemental_subject_island wh_sland matrix_question_npi_licensor_present npi_present_1 npi_present_2 only_npi_licensor_present elemental_negution_npi_licensor_present elemental_negution_npi_scope	96 57 61 56 1 47 47 47 57 30	32 36 43 47 2 54 54 54 93 36	63 36 37 20 1 61 48 80 45 99 53	35 77 67 55 62 100 85 89 95	99 61 73 98 83 98 92 72 93 81	Whose has should Tonya wear? What could Alan discover he has run around? Should Monie verg rui? Even these runch have often skowed. Many skateboards also rull. Only Bill would ever complain. Only those doctors who Karlar respects ever conceal many snakes. Those banks had not ever incl.	Whose theolad Tonya wear hat? Who have many women's touring embarased Spain. What could Alan discover who has na around? Monica should ever grin. Even have trucks have ever slowed. Many skarboardwei ever could. Even fall would ever compain. Those doction who only Marin aropects ever conceal many studes. Those busis had really ever field.	ý
PI	coordinate.structure_constraint_object_extraction left_branch_luturd_extb_question left_branch_luturd_extb_question matrix_question_npi_licensor present mpl_present_i	96 57 61 56 47 47 57 57 30 93 45 91	32 36 43 47 2 54 54 93 36 100 23 96	63 36 37 20 1 61 48 80 45 99 53 94	35 77 67 55 62 100 85 89 95 99	99 61 73 98 83 98 92 72 93 81 94	Whose has should Tonya wear? What could Alan discover he has run around? Should Monica ever grin? Even these tracks law of the aboved. Many stateboards also roll. Only the even could have the state of the state of the state of the state discover who Kafa respects ever conceal many makes. Those tartles that are boring April could not ever break those couches. These aren't may light darkening.	Whose theolad Tonya wear hat? Who have many sources it source embarassed Spain. What could Alan discover who has run around? Monica should ever grain the second se	ý
SLAND FFECTS IPI ICENSING	coordinate.structure_constraint_object_extraction left_branch_island_eches_question esternatia_ubject_island wh_siand matrix_question_npti_lecensor_present npi_present_1 npi_present_2 npi_present_3 npi_present_4 npi_present_4 npi_present_5 npi_present_6 npi_presen	96 57 61 56 47 47 47 57 30 93 45 91 62	32 36 43 47 2 54 54 93 36 100 23 96 16	63 36 37 20 1 61 48 80 45 99 53 94	35 77 67 55 62 100 85 89 95 99 24	99 61 73 98 83 98 92 72 93 81 94 76	Whose has should Tonya weer? What could Alan discover he has run around? Should Monie veer gin? Even these trucks have often aloved. Many stachoout also roll. Only Bill would ever omplain. Only Bill would ever complain. Only Bill would ever complain. Those banks had not ever lied. Those trucks that should a boring April could not ever break those couches. There aren't many lights darkning. Each book is there dischring means.	Whose theolad Tonya wear hat? Who have many sources it source genbarrased Spain. What could Alan discover who has run around? Manica should ever grain. Even inhore tracks have ever allowed. Many akarboardword ever coil. Even Bill would ever complain. Those doctors who only Karla respects ever conceal many snakes. Those hacks had raily ever lick. Those tracks had aro lob oring April could ever heak those couches. Three rave's all lights darkening. Three is cards book disubting Margaret.	ý
PI	coordinate.structure_constraint_object_extraction left_branch_island_sche_question left_branch_island_sche_question enternotat_abusge_island wh_stand matrix_question_npi_licensor_present npi_present_1 npi_present_2 npi_present_2 npi_present_3 npi_present_3 npi_present_4 npi_pres	96 57 61 56 47 47 47 57 30 93 45 91 62 45	32 36 43 47 2 54 54 93 36 100 23 96 16 63	63 36 37 20 1 61 48 80 45 99 53 94	35 77 67 55 62 100 85 89 95 99 24 84	99 61 73 98 83 98 92 72 93 81 94 76 91	Whose has should Tonya wear? What could Alan discover he has run around? What could Alan discover he has run around? Should Monica were grin? Even these trucks have often showed. Many skateboards also roll. Only Bill would ever complain, agent ever conceal many instacs. Those truths that are boring April could not ever break those couches. Those truths that are boring April could not ever break those couches. Those truths that are boring April could not ever break those couches. Those truths that are boring April could not ever break those couches. Those truths that are boring April could not ever break those couches.	Whose should Tonya wear hat? Who have many women's storing embarased Spain. What could Alm discover who has run around? Monica should ever grint. Even these tracks have ever alowed. Many skateboard ever could be also also also also also also also also	\$
PI CENSING	coordinate.structure_constrain_object_extraction left_branch_island_exbq.gestion estentiat_abject_island wh_stand npi_present_1 npi_present_2 npi_present_2 npi_present_2 npi_present_3 npi_present_4 npi_present_4 npi_present_4 npi_present_4 npi_present_4 npi_present_4 npi_present_4 estimatia_hreg_auntifern_1 existentia_hreg_auntifern_1 existentia_hreg_auntifern_1 existentia_hreg_auntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estimatia_hregaauntifern_1 estima	96 57 61 56 47 47 47 57 30 93 45 91 62 45 17	32 36 43 47 2 54 54 93 36 100 23 96 16	63 36 37 20 1 61 48 80 45 99 53 94	35 77 67 55 62 100 85 89 95 95 99 24 84 78	99 61 73 98 83 98 92 72 93 81 94 76 91 85	Whose has should Tonya weer? What could Alan discover he has run around? Should Monie veer gin? Even these trucks have often aloved. Many stachoout also roll. Only Bill would ever omplain. Only Bill would ever complain. Only Bill would ever complain. Those banks had not ever lied. Those trucks that should a boring April could not ever break those couches. There aren't many lights darkning. Each book is there dischring means.	Whose theolad Tonya wear hat? Who have many sources it source genbarrased Spain. What could Alan discover who has run around? Manica should ever grain. Even inhore tracks have ever allowed. Many akarboardword ever coil. Even Bill would ever complain. Those doctors who only Karla respects ever conceal many snakes. Those hacks had raily ever lick. Those tracks had aro lob oring April could ever heak those couches. Three rave's all lights darkening. Three is cards book disubting Margaret.	ý
PI	coordinate.sincure_constrain_object_extraction left_branch_island_exbs_question left_branch_island_exbs_question matrix_question_npi_licensor present npi_present_1 matrix_question_npi_licensor present nult_proj_licensor_present nult_proj_licensor_present nult_proj_licensor_present estimatia_negation_npi_licensor_present semential_negation_npi_scope lettimetia_here_quantifien_1 existentia_here_quantifien_2 superlaive_quantifien_2 superlaive_quantifien_2 superlaive_quantifien_2 superlaive_quantifien_1 superlaive_quantifien_1 superlaive_quantifien_2	96 57 61 56 47 47 47 57 30 93 45 91 62 45 17 24	32 36 43 47 2 54 54 93 36 100 23 96 16 63 83 76	63 36 37 20 1 61 48 80 45 99 53 94 14 84 85 77	35 77 67 55 62 100 85 89 95 95 99 24 84 78 83	99 61 73 98 83 98 92 72 93 81 94 76 91 85 81	Whose has should Tonya wear? What could Alan discover he has run around? What could Alan discover he has run around? Should Monica were grin? Even these trucks have offen aloved. Many skateboards also roll. Many skateboards also roll. Only those decions who Keint respects ever conceal many snakes. Those burks had not ever lied. Those turits that are boring April could not ever break those couches. Those burks had not ever lied. Those turits that are boring April could not ever break those couches. Those barks had not ever lied. An solar arrived at at most its lakes. An acture arrived at at most its lakes.	Whose should Tonya wear hat? Who have many women's touring embarased Spain. What could Alm discover who has run around? Monica should ever grin. Even these tracks have ever dowed. Many skateboard ever coll. These hills would ever colling and the state of the state of the These hills would ever colling and the state of the state of the Those banks has and colly ever incl. Those banks has and colly ever incl. Those banks has and colly ever incl. Those tracks has and colly ever incl. Those tracks has light dowering. There is each book disturbing Margaret. No man has revealed at least five forefs. No action arrived at at most state. No state of the state of the state.	
PI ICENSING	coordinate.structure_constrain_object_extraction left_branch_island_exbs_question elemental_subject_island wh_sland matrix_question_npi_licensor_present npi_present_1 npi_present_2 only_npi_licensor_present elemental_negation_npi_licensor_present elemental_negation_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope	96 57 61 56 47 47 47 57 30 93 45 91 62 45 17 24	32 36 43 47 54 54 93 36 100 23 96 16 63 83 76 63	63 36 37 20 1 61 48 80 45 99 53 94 14 84 85 77 60	35 77 67 55 62 100 85 89 95 99 24 84 78 83 68	99 61 73 98 83 98 92 72 93 81 94 76 91 85 81 86	Whose has about Tonya weer? Who have many wome's tooring Spain embarransed. What could Alan discover he has run around? Should Monice were grin? Even these trucks have often slowed. Many skateboards also roll. Only Bill would ever grin? Those banks had not ever lied. Those trucks that are boring April Could not ever break those couches. There aren't many lights darkening. Each books is there dawthing Magnet. No man has revealed more than five (refs. An atter arrived at most is hales. An actor arrived at most sin hales. A sketch of lights desen't appear. Boys that aren't dawthing Nanlas suffer.	Whose theolad Tonya wear hat? Who have many sources it storing embarased Spain. What could Alan discover who has na around? Menica should ever grin. Even show trucks have ever alowed. Many skateboards ever coil. Even shill would ever compain. Those docts who only Mart arogester ever conceal many stakes. Those tarks had a real be ever field. Those tarks that aro ho bring April could ever heak those couches. There is each bood disturbing Margaret. No actor arrived at it most the fact. No actor arrived at it most the fact. A sketch of lights don't appear. Boys that are(i disturbing Margaret.	
PI CENSING UANTIFIERS	coordinate.structure_constrain_object_extraction left_branch_island_exbs_question elemental_subject_island wh_sland matrix_question_npi_licensor_present npi_present_1 npi_present_2 only_npi_licensor_present elemental_negation_npi_licensor_present elemental_negation_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope	96 57 61 56 47 47 47 57 30 93 45 91 62 45 17 24 22 73	32 36 43 47 2 54 54 93 36 100 23 96 16 63 83 76 63 81	63 36 37 20 1 61 48 80 45 99 53 94 14 84 85 77 60 78	35 77 67 55 62 100 85 89 95 99 24 84 78 83 68 95	99 61 73 98 83 98 92 72 93 81 94 76 91 85 81	Whose has about Tonya weer? Who have many wome's tooring Spain embarransed. What could Alan discover he has run around? Should Monice were grin? Even these trucks have often slowed. Many skateboards also roll. Only Bill would ever grin? Those banks had not ever lied. Those trucks that are boring April Could not ever break those couches. There aren't many lights darkening. Each books is there dawthing Magnet. No man has revealed more than five (refs. An atter arrived at most is hales. An actor arrived at most sin hales. A sketch of lights desen't appear. Boys that aren't dawthing Nanlas suffer.	Whose should Tonya wear hat? Who have many women's storing embarased Spain. What could Alm discover who has run around? Monica should ever grin. Even have tracks have ever alowed. Many skateboards ever conjunit. Even hill would ever conjunit. Those bedre who on ever hild. Those turtles that are not boring April could ever break those couches. There aren't all light darkening No runn has revealed at listed for breks. No rate arrived at at most is klacks. A sketch of light dor't appear. A sketch of light dor't appear. A sketch of light dor't appear. A sketch of light dor't appear.	
PI CENSING UANTIFIERS UBJECT- ERB	coordinate.structure_constrain_object_extraction left_branch_island_senge_extra enternoil_abject_island wh_stand matrix_question_npi_licensor_present npi_present_1 n	96 57 61 56 47 47 47 57 30 93 45 91 62 45 17	32 36 43 47 54 54 54 54 54 54 54 54 54 54 54 54 54	63 36 37 20 1 61 48 80 45 99 53 94 14 84 85 77 60	35 77 67 55 62 100 85 89 95 99 24 84 78 83 68 95 96	99 61 73 98 83 98 92 72 93 81 94 76 91 85 81 86	Whose has should Tonya weer? What could Alan discover he has run around? Should Monie veer gain? Even these much have often clored. Many stachowski oks roll. Only Bill would ever complain. Only Bill would ever complain. Only Bill would ever complain. Only Bill would ever complain. Only Bill would ever complain. Those turks that are being April could not ever break those couches. These area's that are being April could not ever break those couches. These area's that are being April could not ever break those couches. These area's that around a star being April could not ever break those couches. These area's that around a star around a star bill and a star around a star bill and a star around a star bill and around a star around a star bill and a star around a star bill and a star around a	Whose should Tonya wear hat? Who have many women's storing embarased Spain. What could Alm discover who has run around? Monica should ever grin. Even have tracks have ever alowed. Many skateboards ever conjunit. Even hill would ever conjunit. Those bedre who on ever hild. Those turtles that are not boring April could ever break those couches. There aren't all light darkening No runn has revealed at listed for breks. No rate arrived at at most is klacks. A sketch of light dor't appear. A sketch of light dor't appear. A sketch of light dor't appear. A sketch of light dor't appear.	
PI CENSING UANTIFIERS	coordinate.structure_constrain_object_extraction left_branch_island_exbs_question elemental_subject_island wh_sland matrix_question_npi_licensor_present npi_present_1 npi_present_2 only_npi_licensor_present elemental_negation_npi_licensor_present elemental_negation_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope existentian_npi_scope	96 57 61 56 47 47 47 57 30 93 45 91 62 45 17 24 22 73	32 36 43 47 2 54 54 93 36 100 23 96 16 63 83 76 63 81	63 36 37 20 1 61 48 80 45 99 53 94 14 84 85 77 60 78	35 77 67 55 62 100 85 89 95 99 24 84 78 83 68 95	99 61 73 98 83 98 92 72 93 81 94 76 91 85 81 86 95	Whose has should Tonya were? What could Alan discover he has run around? What could Alan discover he has run around? Should Monica were grin? Even these trucks have often showed. Many skateboards also roll. Only Bill would ever complain, genes ever conceal many induces. Those back has but over list. Those trucks that are boring April could not ever break those couches. These sees in many lights darkening. These sees in many lights darkening. A sector of signs down't appear. A sector arrived at at most is lakes. How the down't appear. How the area's dimensional sectors. A sector arrived at at most is lakes.	Whose theolad Tonya wear hat? Who have many sources it storing embarased Spain. What could Alan discover who has na around? Menica should ever grin. Even show trucks have ever alowed. Many skateboards ever coil. Even shill would ever compain. Those docts who only Mart arogester ever conceal many stakes. Those tarks had a real be ever field. Those tarks that aro ho bring April could ever heak those couches. There is each bood disturbing Margaret. No actor arrived at it most the fact. No actor arrived at it most the fact. A sketch of lights don't appear. Boys that are(i disturbing Margaret.	

Warstadt et al. (2020)

Data Generation

Datasets

- All minimal pairs are *artificially generated* from a vocabulary of 3,000 words, each lexical item annotated with morphological, syntactic, and semantic features

Example: Causative Frame

sentence_good: "Aaron breaks the glass." sentence_bad: "Aaron appeared the glass.", Linguistics_term (major): "argument_structure", UID (minor): "causative"



Comparing FAVA/CoLA and BliMP

CoLA/FAVA

- Supervised Binary Acceptability Judgments
- No "generally accepted method" to obtain acceptability predictions from unsupervised model
 -> need to use something like Logistic Reg. / MLP
- Sentences are pulled directly from wide variety of Linguistic corpora for CoLA (not the case for FAVA)

BliMP

- Unsupervised Acceptance Probabilities using LM objective
- Can use unsupervised LMs like GPT-2, Transformer-XL, LSTMs, etc.. directly to model probability
- Sentences are artificially generated, acceptability judgments from authors and validated through Amazon Mechanical Turk

Results

Phenomenon	UID	5-gra	m LST	N TXL	GPT.	Human
Anaphor agreement	anaphor_gender_agreement anaphor_number_agreement	44 52	88 95	91 97	99 100	96 99
Argument structure	animate_subject_passive animate_subject_trans causative drop_argument inchoative intransitive passive_1 passive_2	54 72 51 68 89 82 71 70	68 79 65 79 72 73 65 72	58 70 54 67 81 81 76 74	77 80 68 84 90 90 89 79	98 87 82 90 95 86 99 86
	transitive	91	87	89	49	87

Performance on Verb Argument Structure Classes

Warstadt and Bowman (2019): "Performance is also high on sentences with marked argument structures, indicating that argument structure is relatively easy to learn"

- Analyzing the performance of BERT, GPT, etc.. on CoLA

Warstadt et al. (2020): "We note that the reported difficulty of these phenomena contradicts Warstadt and Bowman's (2019) conclusion that argument structure is one of the strongest domains for neural models."

Hypotheses:

- Supervised v.s. Unsupervised datasets
- Disproportionate amount of "Argument Structure" related sentences in CoLA

Other Interesting Takeaways

5-gram	0.34	0.39	0.58	0.59	1
LSTM	0.49	0.63	0.9	1	0.59
TXL	0.48	0.68	1	0.9	0.58
GPT-2	0.54	1	0.68	0.63	0.39
human	1	0.54	0.48	0.49	0.34
	human	GPT_2	TYI	I STM	5_gram

numan GPI-2 TAL LSTM 5-gram

Figure 1: Heatmap showing the correlation between models' accuracies in each of the 67 paradigms.

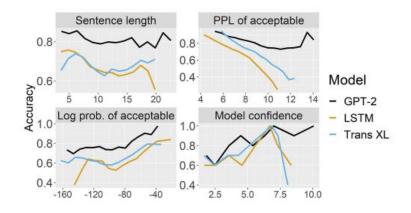


Figure 2: Models' performance on BLiMP as a function of sentence length, perplexity, log probability of the acceptable sentence, and model confidence (calculated as $|\log P(S_1) - \log P(S_2)|$).

Our Work

We are reproducing the experiments of Kann et al. (2019) by analyzing BERT (and other LLM) embeddings

- Without the constraints of studying "to what extent do the features learned by ANNs resemble the linguistic competence of humans"
- Probe linguistic knowledge of frozen BERT representations **without** additional finetuning for both word/sentence-level embeddings

Essential idea: Use "better" embeddings (static -> contextual) and dumb down the classifier to tackle "Probe Confounder Problem" (Hewitt and Liang, 2019)

- Classifier: Simple Logistic Regression (LR) Classifier
- Control Task: Compare between LR, MLP-1, MLP-2

Q&A