brains and languages in the fantasy world of NLP

Yosef Grodzinsky

www.grodzinskylab.com
yosef.grodzinsky@mail.huji.ac.il

a neural explanation of language - the holy grail of our business

- But what could it look like?
- A neural explanation of what exactly?
- What are the neural and behavioral variables?

NLP's answer:

- language is a huge bag of sequentially layed out words and sentences.
- To understand it, we just need to calculate probabilisties of occurrence of words and sentences in context. Nothing else is needed.
- DNNs offer the finest, most advanced and suitable tools.
- Finding the neural substrate of these tools is a breakthrough.

what I hope to do here

- *A bit on what language models (GPT-2, BERT) do (but not on how they do it)*
  - ➢ what they use as their primary learning input
  - ➢ what tests they used to evaluate their models
- *Connection to the brain (these DNNs are the brain's processors)*
- *Critique*
  - ➢ representativeness of stimuli
  - ➢ DNN performance on certain language tasks
  - ➢ psycho-biological relevance of the models' architecture
  - ➢ Where there is good fit between models' output and human signals, and where there isn't
- *Hints regarding a modular, structure-based, alternative*

DNNs for language: a. “unidirectional” models (GPT-2) (here: moving window of length *l*=4)

#	1	2	3	4	5	6	7	8	9	10	#
.	The	man	asked	the	woman	to	sit	on	the	chair	.
.	The	*p(1/#)*									
.	The	**man**	*p(2/1)*								
.	The	man	**asked**	*p(3/1+2)*							
	The	man	asked	**the**	*p(4/1+2+3)*						
	The	man	asked	the	**woman**	*p(5/1+2+3+4)*					
		man	asked	the	woman	**to**	*p(n/n-4…n-1)*				
			asked	the	woman	to	**sit**	*p(n/n-4…n-1)*			
				the	woman	to	sit	**on**	*p(n/n-4…n-1)*		
					woman	to	sit	on	**the**	*p(n/n-4…n-1)*	
						to	sit	on	the	**chair**	*p(n/n-4…n-1)*

Next word prediction

$$p(w_n)=p(w_n/(w_{n-1}\&w_{n-2}..\&\#))$$

DNNs for language: b. “Bidirectional”, collocations-based models (Bidirectional Encoder Representations from Transformers – BERT)

#	1	2	3	4	5	6	7	8	9	10	##
.	The	man	asked	the	woman	to	sit	on	the	chair	.
.	*$p(w_1)$*	man	asked	the	woman	to	sit	on	the	chair	.
.	The	*$p(w_2)$*	asked	the	woman	to	sit	on	the	chair	.
.	The	man	*$p(w_3)$*	the	woman	to	sit	on	the	chair	.
.	The	man	asked	*$p(w_4)$*	woman	to	sit	on	the	chair	.
.	The	man	asked	the	*$p(w_5)$*	to	sit	on	the	chair	.
.	The	man	asked	the	woman	*$p(w_6)$*	sit	on	the	chair	.
.	The	man	asked	the	woman	to	*$p(w_7)$*	on	the	chair	.
.	The	man	asked	the	woman	to	sit	*$p(w_8)$*	the	chair	.

$$p(w_n)=p(w_n/(w_{n-1}\&w_{n-2}..\&\#))*p(w_n/(w_{n+1}\&w_{n+2}..\&\#))$$

some i/o characteristics

- **Input**: sentences (labeled or unlabeled); linear but no hierarchical information
- **Sources**: strings found on the web; experimental data pertaining to ‘naturally occurring’ strings.
- **Output**: probability of occurrence in a specific serial position; choices among alternative sequences.

- ***Absent**: grammatical objects or rules, distinction between grammatical and ungrammatical strings*

  ➢ “Every time we fire a linguist the performance of our system goes up”

(Jelinek, IBM, 1985)

model assessment tasks

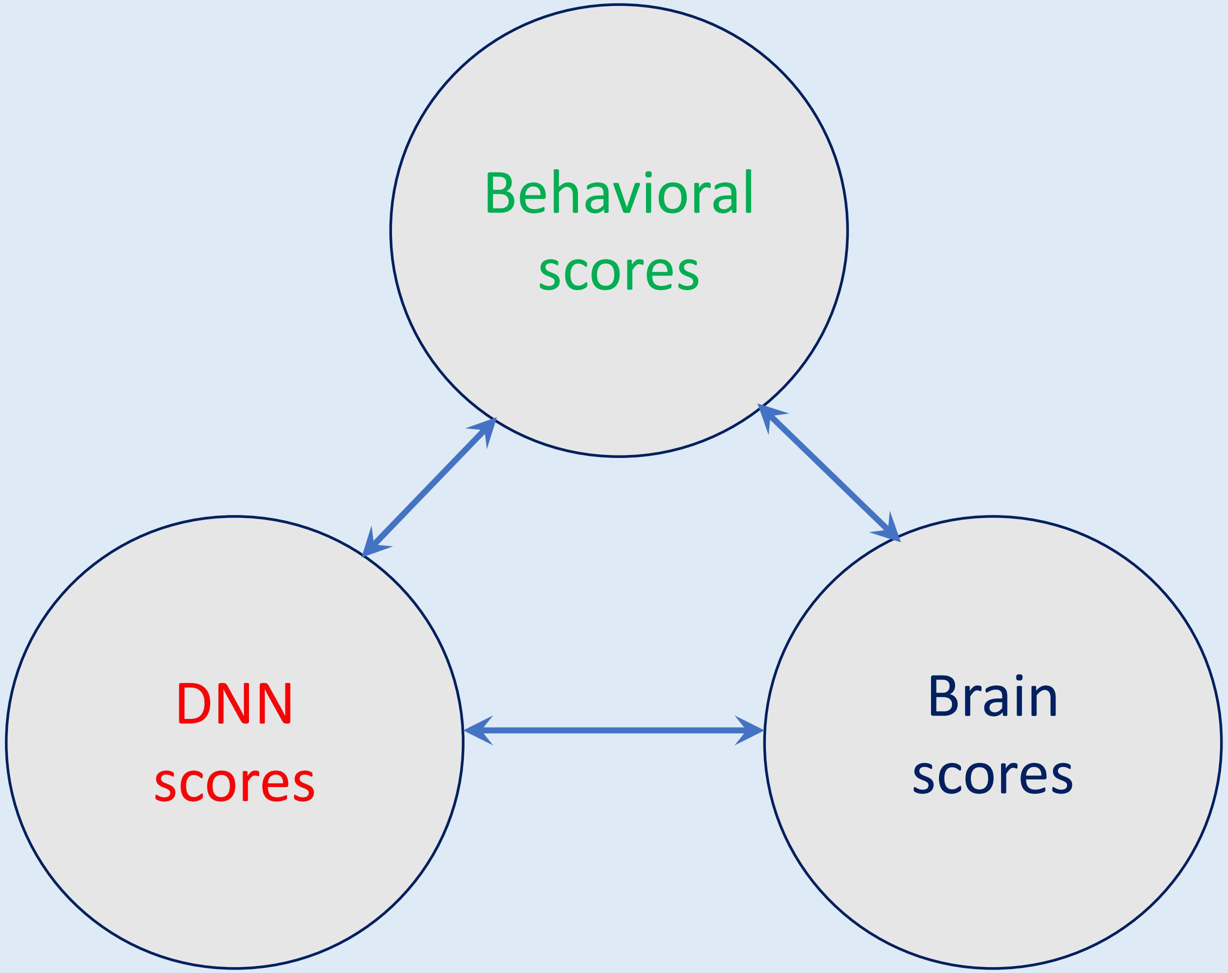
Tasks

- Next word prediction
- Next sentence prediction (out of choices provided)
- Very limited, corpus-based, question answering
- Related tasks on the corpus

GPT-2

- Improves on pre-training
- Ourperforms other systems

test: next word prediction in a “real-world” text –
humans, machines, and brains’ performance on the same inputs



Two very recent, high visibility works

Shared computational principles for language processing in humans and deep language models

Ariel Goldstein[1,2], Zaid Zada[1,8], Eliav Buchnik[2,8], Mariano Schain[2,8], Amy Price[1,8], Bobbi Aubrey[1,3,8], Samuel A. Nastase[1,8], Amir Feder[2,8], Dotan Emanuel[2,8], Alon Cohen[2,8], Aren Jansen[2,8], Harshvardhan Gazula[1], Gina Choe[1,3], Aditi Rao[1,3], Catherine Kim[1,3], Colton Casto[1], Lora Fanda[3], Werner Doyle[3], Daniel Friedman[3], Patricia Dugan[3], Lucia Melloni[4], Roi Reichart[5], Sasha Devore[3], Adeen Flinker[3], Liat Hasenfratz[1], Omer Levy[6], Avinatan Hassidim[2], Michael Brenner[2,7], Yossi Matias[2], Kenneth A. Norman[1], Orrin Devinsky[3] and Uri Hasson[1,2]

The neural architecture of language: Integrative modeling converges on predictive processing

Martin Schrimpf[a,b,c,1], Idan Asher Blank[a,d,2], Greta Tuckute[a,b,2], Carina Kauf[a,b,2], Eghbal A. Hosseini[a,b], Nancy Kanwisher[a,b,c,1], Joshua B. Tenenbaum[a,c,3], and Evelina Fedorenko[a,b,1,3]

Goldstein et al., *Nat. Neuro.*, 2021; Schrimpf et al., *PNAS*, 2021

Goldstein *et al.*, 2021

“we provide empirical evidence that **the human brain processes incoming speech similarly to an autoregressive DLM**”

(Goldstein et al., NN, 2021:369)

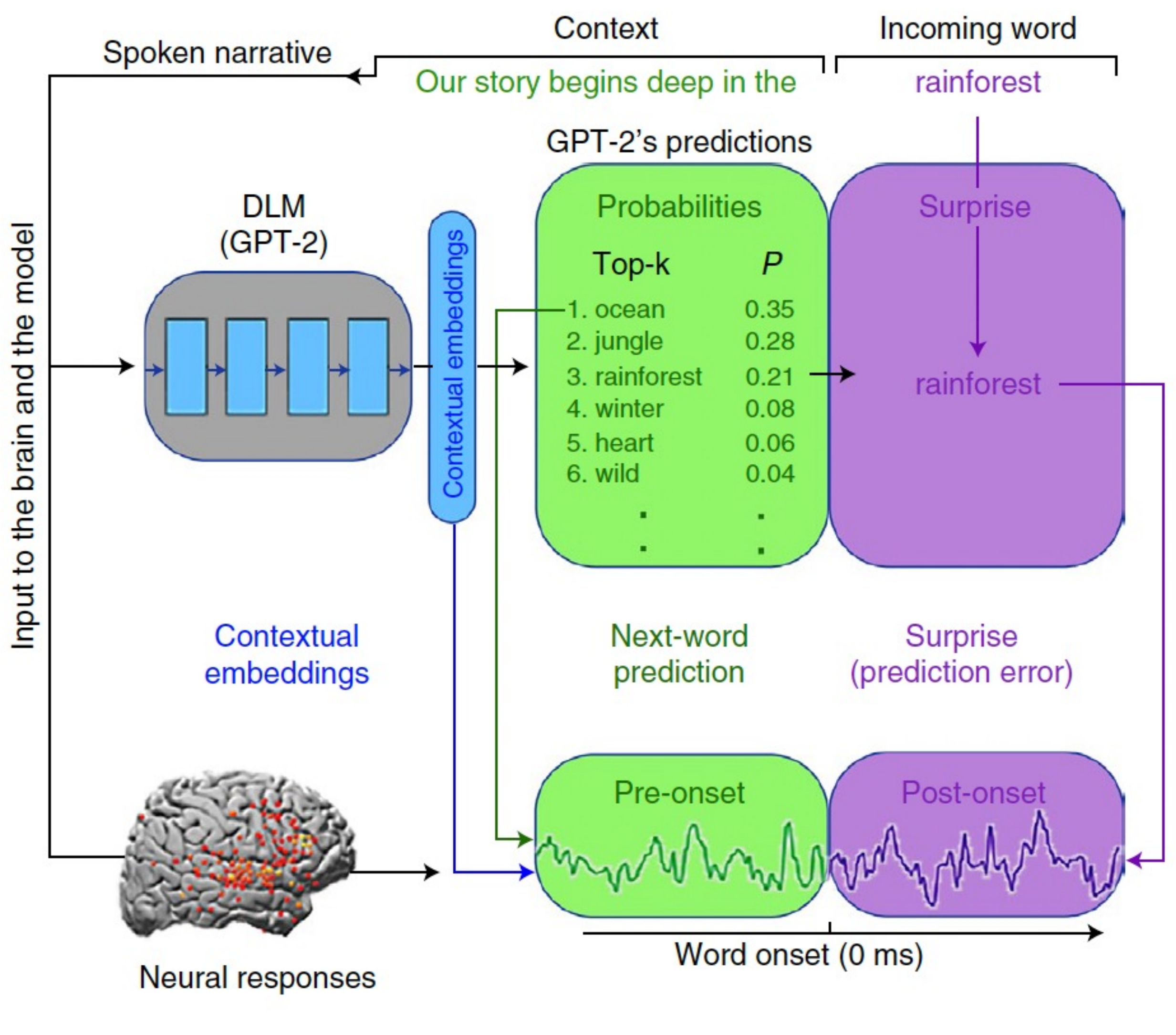
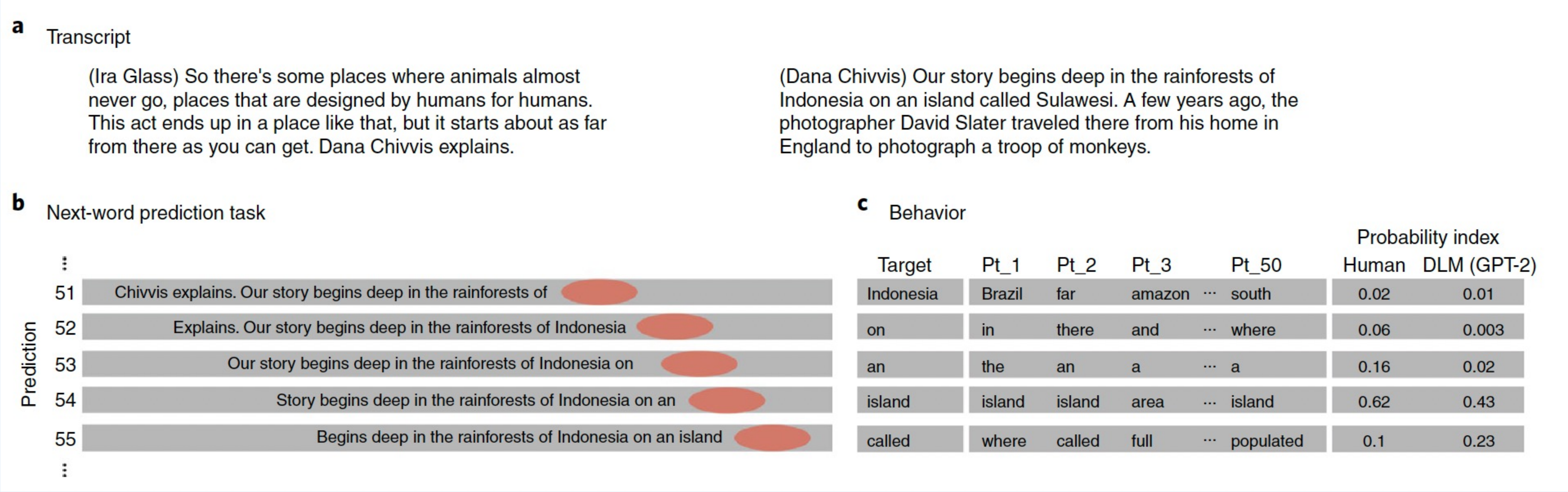
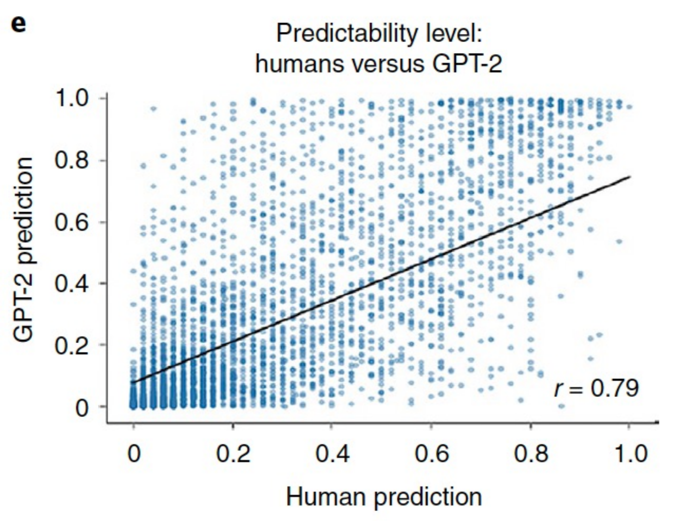


Fig. 1 | Shared computational principles between the brain and autoregressive deep language models in processing natural language.

next-word prediction on random stories (*“This American Life”*)

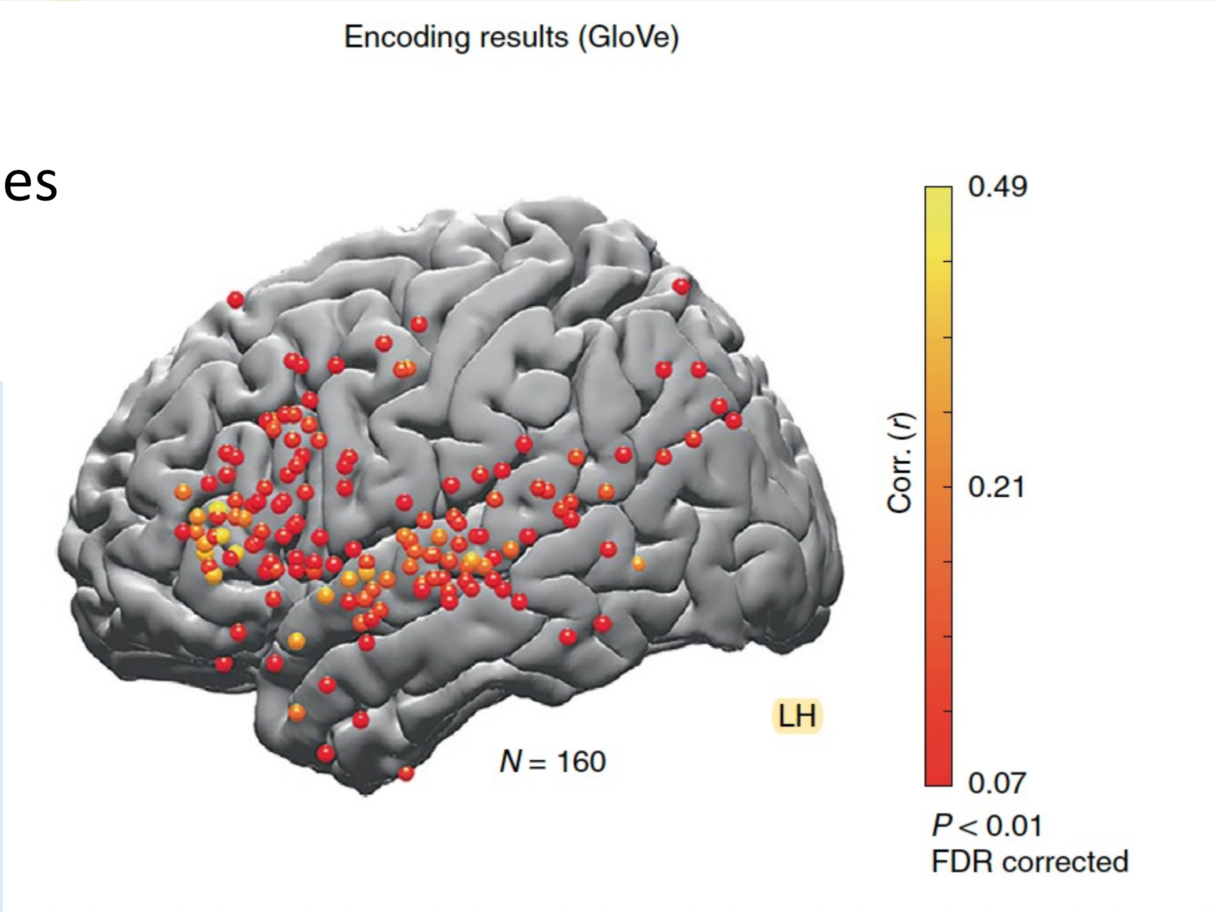


Human predictability scores versus GPT-2’s predictability scores for each upcoming word in the podcast

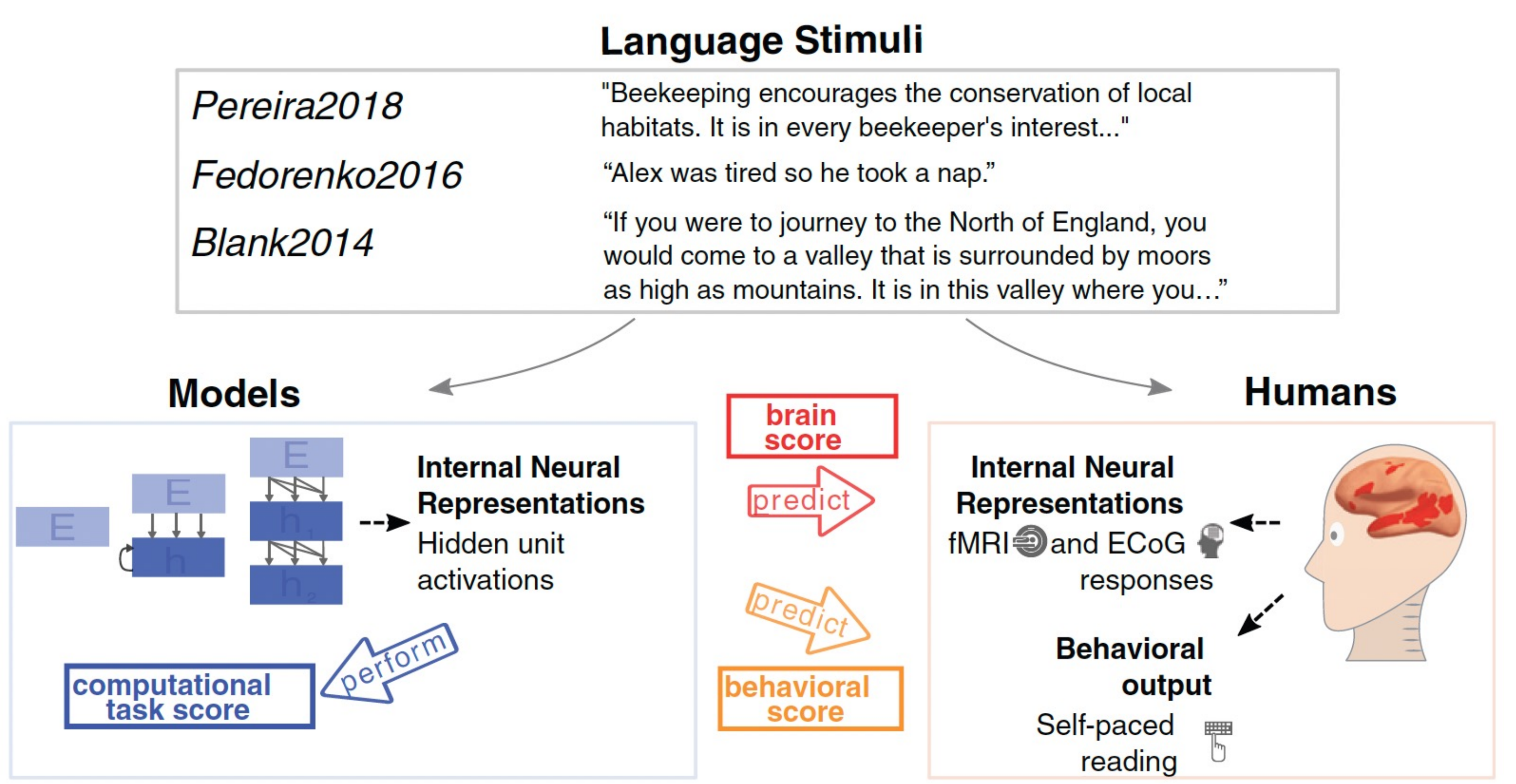


ECoG data

"Electrodes (160/1,339, in 9 patients) with significant correlation at the peaked lag between predicted and actual word responses for semantic embeddings"
(Goldstein et al., 2021: 373)

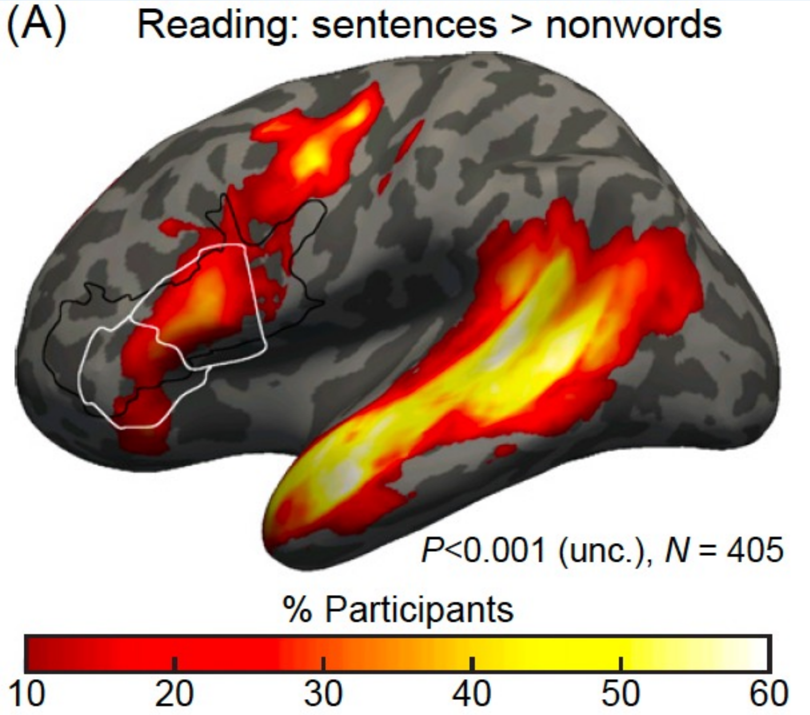


Schrimpf et al., 2021
fMRI & ECoG data



Artificial Neural Network (ANN) models "that perform better at predicting the next word in a sequence, also better predict brain measurements… GPT-2 consistently outperforms all other models and **explains almost all variance in both fMRI and ECoG data from sentence-processing tasks**." (Schrimpf et al., *PNAS*, 2021)

selected brain areas and stimuli



A large cluster is identified by a test

Sample materials

Condition	Example
Sentences	STEVE WAS LATE TO SCHOOL BECAUSE HE OVERSLEPT [probe: SCHOOL]
	THE RED BALLOON ROSE UP INTO THE CLOUDS [probe: WENT]
Word-lists	RAIN THE WORK BEHIND REACHED GREW KIDS OPENED [probe: GREW]
	STOOD THE TIED CANDLE INTO SHED THE QUICKLY [probe: WALLET]
Jabberwocky	THE GAR WAS SWARBING THE MUME FROM ATAR [probe: ATAR]
	TOMAL HOTHED THE BLESPY NULO DURING THE VAYLANT [probe: FLORKY]
Nonword-lists	PHREZ CRE EKED PICUSE EMTO PECH CRE ZEIGELY [probe: PHREZ]
	PIV WUBA WOS PAFFING DEBON TRIENED LE KIF [probe: LOME]

Skiing (passage 1)

I hesitantly skied down the steep trail that my buddies convinced me to try. I made a bad turn, and I found myself tumbling down. I finally came to a stop at a flat part of the slope. My skis were nowhere to be found, and my poles were lodged in a snow drift up the hill.

Exemplars from test materials

Schrimpf's Stimuli

Participants read semantically and syntactically diverse sentences, presented one word at a time…the datasets varied in the imaging modality (fMRI/ECoG), the nature of the materials (unconnected sentences/ passages/stories), the grain of linguistic units to which responses were recorded (sentences/words/2-s-long story fragments), and presentation modality (reading/listening).

(Fedorenko et al., 2016)

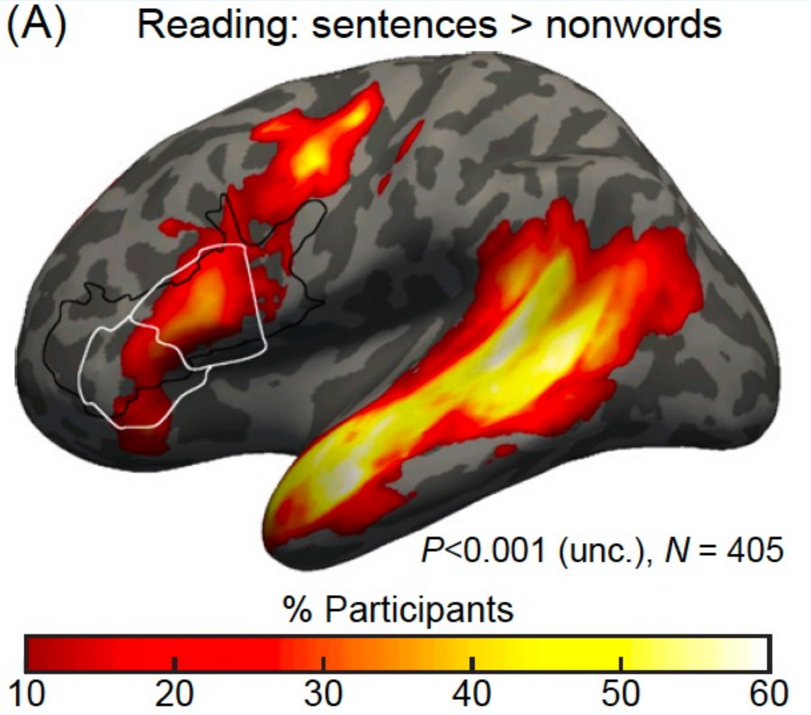
Sample materials

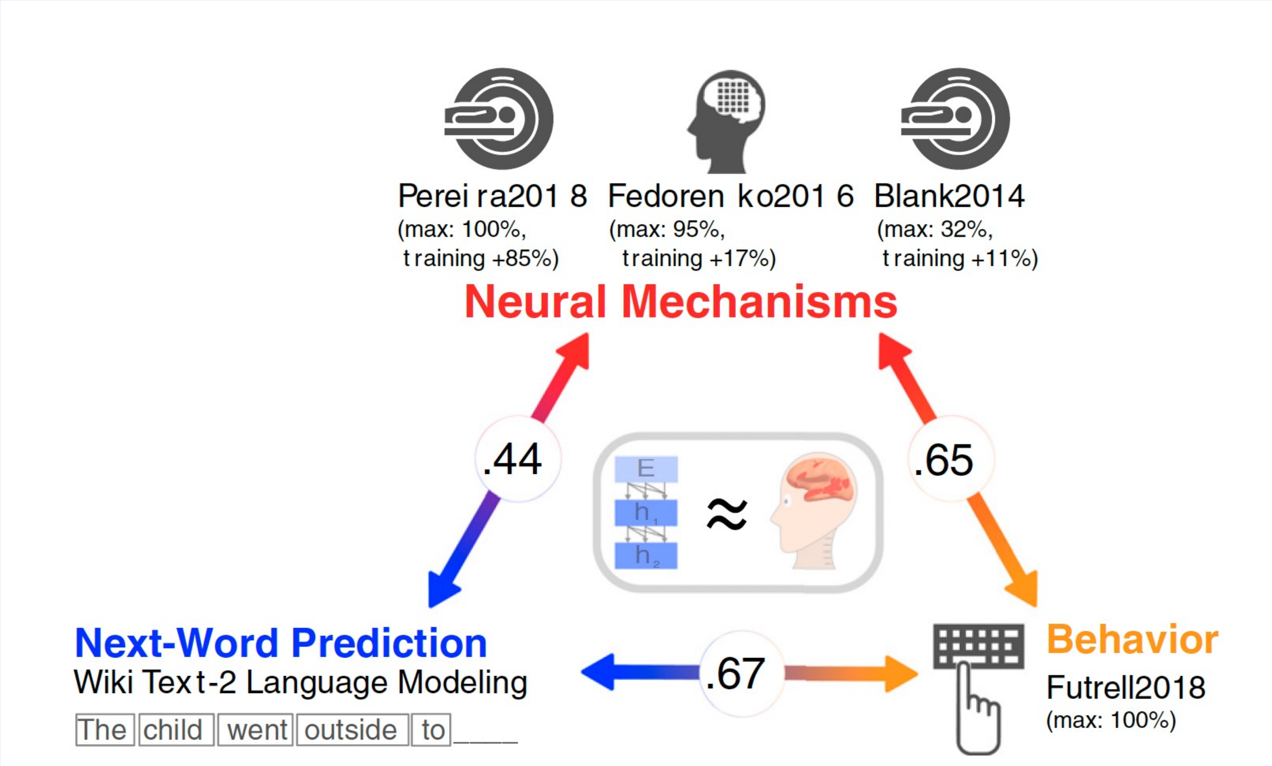
Condition	Example
Sentences	STEVE WAS LATE TO SCHOOL BECAUSE HE OVERSLEPT [probe: SCHOOL]
	THE RED BALLOON ROSE UP INTO THE CLOUDS [probe: WENT]
Word-lists	RAIN THE WORK BEHIND REACHED GREW KIDS OPENED [probe: GREW]
	STOOD THE TIED CANDLE INTO SHED THE QUICKLY [probe: WALLET]
Jabberwocky	THE GAR WAS SWARBING THE MUME FROM ATAR [probe: ATAR]
	TOMAL HOTHED THE BLESPY NULO DURING THE VAYLANT [probe: FLORKY]
Nonword-lists	PHREZ CRE EKED PICUSE EMTO PECH CRE ZEIGELY [probe: PHREZ]
	PIV WUBA WOS PAFFING DEBON TRIENED LE KIF [probe: LOME]

Skiing (passage 1)

I hesitantly skied down the steep trail that my buddies convinced me to try. I made a bad turn, and I found myself tumbling down. I finally came to a stop at a flat part of the slope. My skis were nowhere to be found, and my poles were lodged in a snow drift up the hill.

Schrimpf's results





"**Specific ANN language models are beginning to approximate the brain's mechanisms for processing language** (Middle, gray box)".

a happy conclusion

“we provide empirical evidence that **the human brain processes incoming speech similarly to an autoregressive DLM.**”

(Goldstein et al.)

“**Specific ANN language models are beginning to approximate the brain’s mechanisms for processing language.**”

(Schrimpf et al.)

What's the problem?

- *Are the stimuli used representative of **the range of human language**?*
  - Tests were on grammatically similar contexts
  - Tests presented very similar perceptual complexity
- *Are DNNs **similar to human speakers**?*
  - DNNs fail on tasks that require complex linguistic ability (Google Translate, AI21, GPT-2) - the correlations drop sharply with such tasks
  - DNNs rely on gigantic training sets, but human language acquisition doesn't
- *Is DNN architecture biologically relevant and **localizable in the brain**?*

$\Rightarrow$ The punchline of this talk is simple: these works did not measure "language," but some generic prediction done *in the context of linguistic activity*

Adversarial examples in language: Grammaticality judgment and perceptual complexity

- *Questions can be long-distance, but are restricted*

I heard that bill thinks that John loves Mary — **OK**

Who did I hear that bill thinks that John loves ~~Mary~~? — **OK**

I heard the rumor that John loves Mary — **OK**

Who** did I hear the rumor that John loves ~~Mary~~? — ***ODD

Who is the person such that I heard the rumor that John loves her?

- *Next word prediction is highly structure-dependent*

Which girl do you think ~~**that girl**~~ likes John

Which girl do you think John likes to play with ~~**that girl**~~ these days

Have the students who failed the exam take the supplementary!

slowing

but AI modeling doesn't recognize that linguistic rules define the boundaries of language (cf. Jelinek)

(Ira Glass) So there's some places where animals almost never go, places that are designed by humans for humans. This act ends up in a place like that, but it starts about as far from there as you can get. Dana Chivvis explains.

(Dana Chivvis) Our story begins deep in the rainforests of Indonesia on an island called Sulawesi. A few years ago, the photographer David Slater traveled there from his home in England to photograph a troop of monkeys.

Language Stimuli

Pereira2018	"Beekeeping encourages the conservation of local habitats. It is in every beekeeper's interest..."
Fedorenko2016	"Alex was tired so he took a nap."
Blank2014	"If you were to journey to the North of England, you would come to a valley that is surrounded by moors as high as mountains. It is in this valley where you..."

Stimuli: “Localizer sentences”

1	'NEVER'	'AGAIN'	'DID'	'HE'	'ENTER'	'INTO'	'THE'	'RITUAL'	'OF'	'SHOWING'	'THE'	'APARTMEN
2	'THEN'	'ANGELINA'	'TURNED'	'AND'	'WITH'	'AN'	'EASY'	'GRACE'	'WALKED'	'TOWARD'	'THE'	'KITCHEN.'
3	'HE'	'SEEMED'	'TO'	'BE'	'LOOKING'	'AT'	'A'	'POINT'	'ABOVE'	'THE'	'LITTLE'	'WINDOW.'
4	'JUST'	'THE'	'BAREST'	'SUGGESTIO	'OF'	'A'	'HEEL'	'IS'	'FOUND'	'ON'	'TEENAGE'	'PUMPS.'
5	'HIS'	'WIFE'	'WAS'	'IN'	'DELICATE'	'HEALTH'	'AND'	'NURSING'	'AN'	'INFANT'	'WITH'	'MEASLES.'
6	'THE'	'TARGET'	'CHART'	'QUICKLY'	'AND'	'BRIEFLY'	'TELLS'	'YOU'	'WHICH'	'ADDITIVES'	'DO'	'WHAT.'
7	'KYOTO'	'IS'	'THE'	'ANCIENT'	'CAPITAL'	'OF'	'JAPAN'	'AND'	'STILL'	'ITS'	'CULTURAL'	'CENTER.'
8	'THIS'	'HAPPENED'	'IN'	'THE'	'MIDDLE'	'OF'	'A'	'DRINKING'	'BOUT'	'WITH'	'ANOTHER'	'BUM.'
9	'HE'	'SAT'	'UP'	'AND'	'WATCHED'	'AS'	'THEY'	'PULLED'	'THEMSELVE	'OVER'	'THE'	'STERN.'
10	'MIKE'	'PASSED'	'THROUGH'	'IT'	'AND'	'MOVED'	'TOWARD'	'THE'	'DARK'	'MASS'	'OF'	'HORSES.'
11	'AT'	'ONCE'	'A'	'BEVY'	'OF'	'DOGS'	'WAS'	'SNAPPING'	'AND'	'SNARLING'	'AROUND'	'HIM.'
12	'IT'	'WAS'	'A'	'ROUGH'	'LONG'	'RIDE'	'THROUGH'	'THE'	'MUD'	'AND'	'POT'	'HOLES.'
13	'I'	'WENT'	'TO'	'VISIT'	'ALFRED'	'IN'	'THE'	'KINGSTON'	'HOSPITAL'	'A'	'FEW'	'TIMES.'
14	'IF'	'WE'	'LOOK'	'AT'	'RECENT'	'ART'	'WE'	'FIND'	'IT'	'PREOCCUPIE	'WITH'	'FORM.'
15	'THE'	'REPORTER'	'NODDED'	'AS'	'HE'	'MOVED'	'UP'	'BESIDE'	'HIM'	'AT'	'THE'	'BAR.'
16	'IN'	'THE'	'STARLIGHT'	'HE'	'COULD'	'SEE'	'THE'	'TREES'	'STRIPPED'	'OF'	'THEIR'	'LEAVES.'
17	'I'	'WAS'	'HELD'	'UP'	'A'	'BIT'	'TRYING'	'TO'	'MAKE'	'A'	'LEFT'	'TURN.'
18	'THE'	'OTHER'	'PATRONS'	'WERE'	'TAXI'	'DRIVERS'	'AND'	'ART'	'STUDENTS'	'AND'	'SMALL'	'SHOPKEEPE
19	'HE'	'SAT'	'DOWN'	'ON'	'AN'	'OLD'	'BOX'	'AND'	'FOCUSED'	'ON'	'THE'	'PROBLEM.'
20	'THIS'	'IS'	'AN'	'ASSUMPTIO	'WITH'	'WHICH'	'FEW'	'WOULD'	'BE'	'DISPOSED'	'TO'	'QUARREL.'
21	'THERE'	'ARE'	'THOUSAND	'OF'	'SQUARE'	'MILES'	'OF'	'SALT'	'PAN'	'WHICH'	'ARE'	'HIDEOUS.'
22	'HE'	'STOPPED'	'PACING'	'TO'	'STARE'	'AT'	'HAL'	'WITH'	'HIS'	'PALE'	'BLUE'	'EYES.'
23	'A'	'NUMBER'	'OF'	'CONSIDERA	'SUGGEST'	'THAT'	'THIS'	'OCCURS'	'EARLY'	'IN'	'THE'	'PROCESS.'
24	'IT'	'IS'	'VERY'	'MUCH'	'A'	'MATTER'	'OF'	'BUILDING'	'THE'	'FOUNDATIC	'OF'	'COMMUNIT

12-words long; changing number of syllables; arbitrarily selected meanings; a narrow variety of syntactic types

Stimuli

Participants read semantically and syntactically diverse sentences, presented one word at a time…the datasets varied in the imaging modality (fMRI/ECoG), the nature of the materials (unconnected sentences/ passages/stories), the grain of linguistic units to which responses were recorded (sentences/words/2-s-long story fragments), and presentation modality (reading/listening).

(Fedorenko et al., 2016)

- ***A single contrast*** (Ss vs. scrambled words)

⇒ all sentences led to similar brain activity (against an uneven baseline)

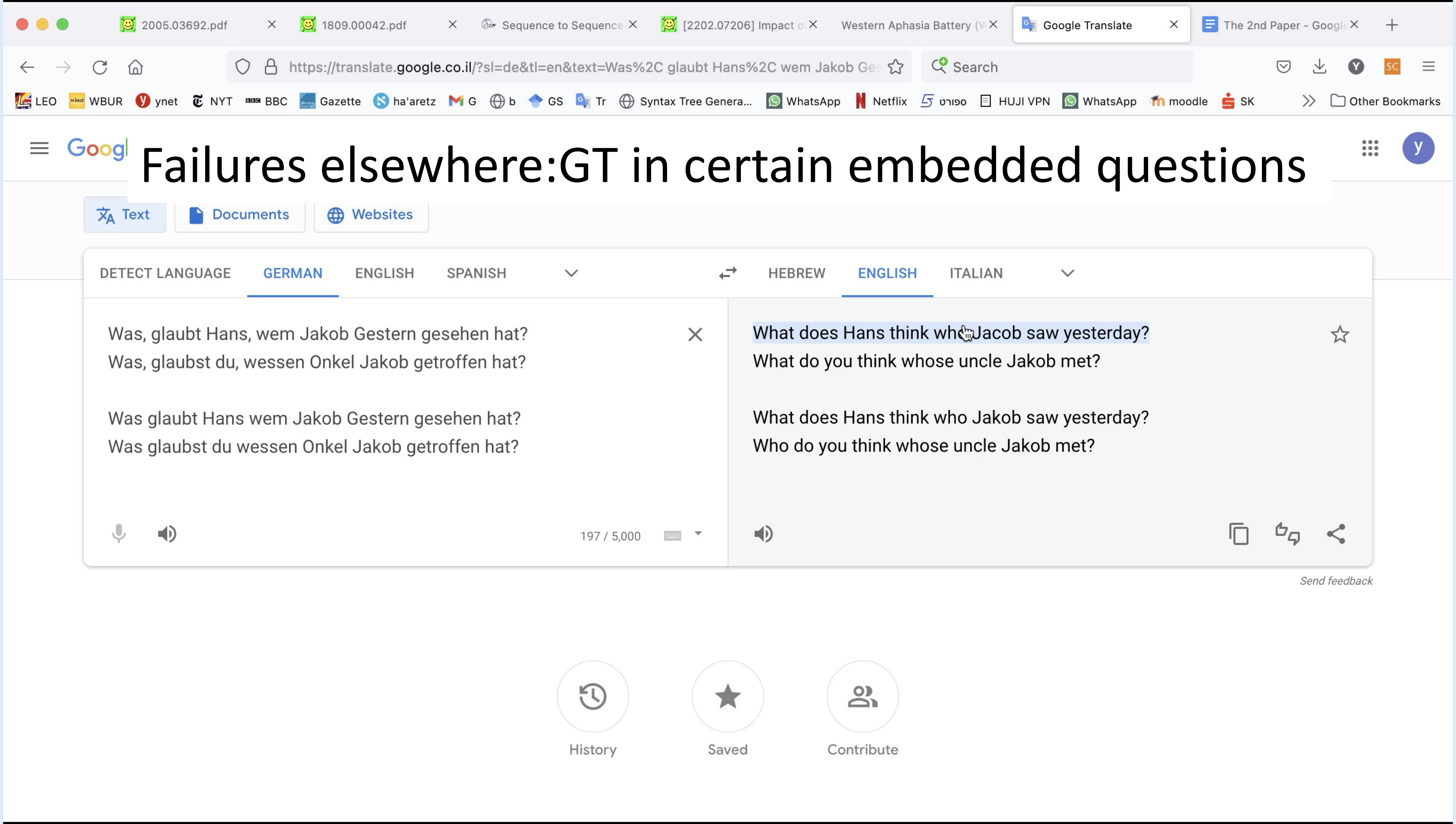
- ***Hardly any***
  - complex sentences - embeddings
  - logical operators (negations,
  - disjunctions, modals, quantifiers)
  - ellipsis
  - ambiguities – syntactic and semantic

- ***No questions***

⇒ Complexity was not measured

- **No weighting** of the stimulus materials

Failures elsewhere:GT in certain embedded questions



failure to detect ungrammaticality that requires complex rules

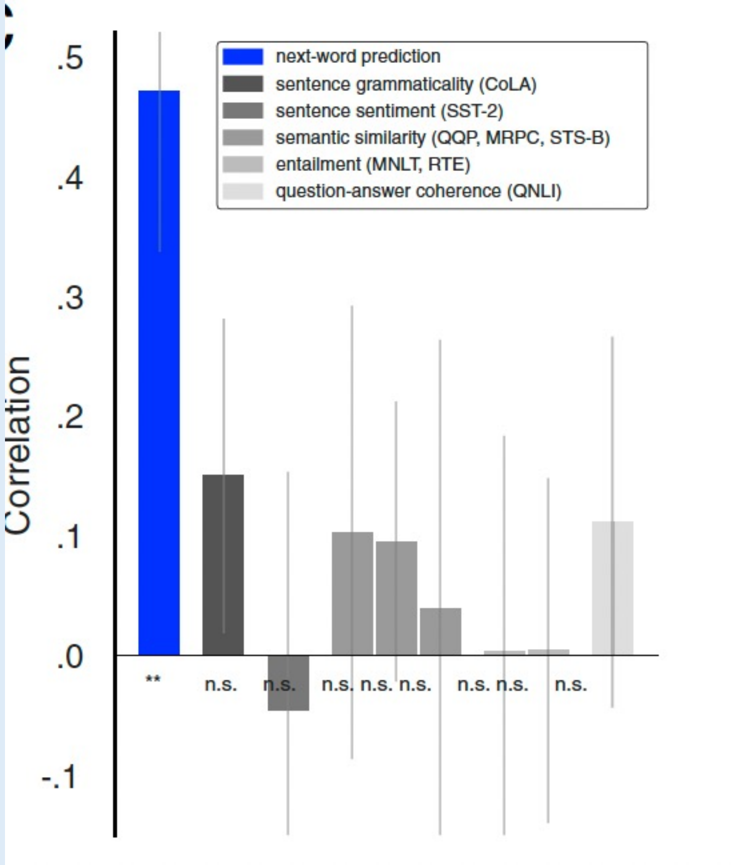
	RNN	Multitask	*n*-gram	Humans	# sents
SUBJECT-VERB AGREEMENT:					
Simple	0.94	1.00	0.79	0.96	280
In a sentential complement	0.99	0.93	0.79	0.93	3360
Short VP coordination	0.90	0.90	0.51	0.94	1680
Long VP coordination	0.61	0.81	0.50	0.82	800
Across a prepositional phrase	0.57	0.69	0.50	0.85	44800
Across a subject relative clause	0.56	0.74	0.50	0.88	22400
Across an object relative clause	0.50	0.57	0.50	0.85	44800
Across an object relative (no *that*)	0.52	0.52	0.50	0.82	44800
In an object relative clause	0.84	0.89	0.50	0.78	44800
In an object relative (no *that*)	0.71	0.81	0.50	0.79	44800
REFLEXIVE ANAPHORA:					
Simple	0.83	0.86	0.50	0.96	560
In a sentential complement	0.86	0.83	0.50	0.91	6720
Across a relative clause	0.55	0.56	0.50	0.87	44800
NEGATIVE POLARITY ITEMS:					
Simple	0.40	0.48	0.06	0.98	792
Across a relative clause	0.41	0.73	0.60	0.81	31680

Marvin & Linzen, *Proc. ACL*, 2018

A General Language Understanding Evaluation (GLUE) benchmark uncovers failure (Schrimpf)

Wang et al., *IRCL*, 2019

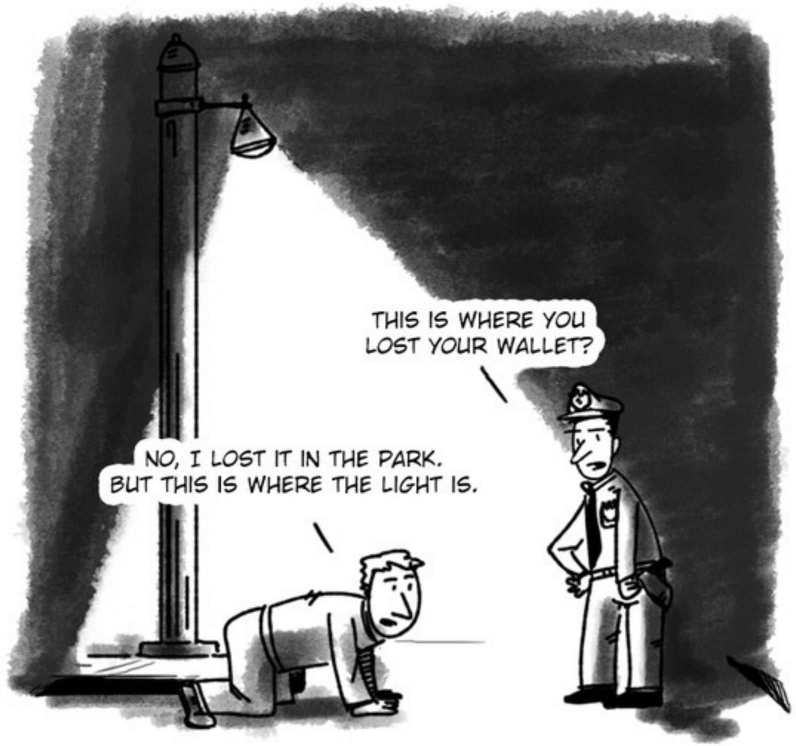
Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

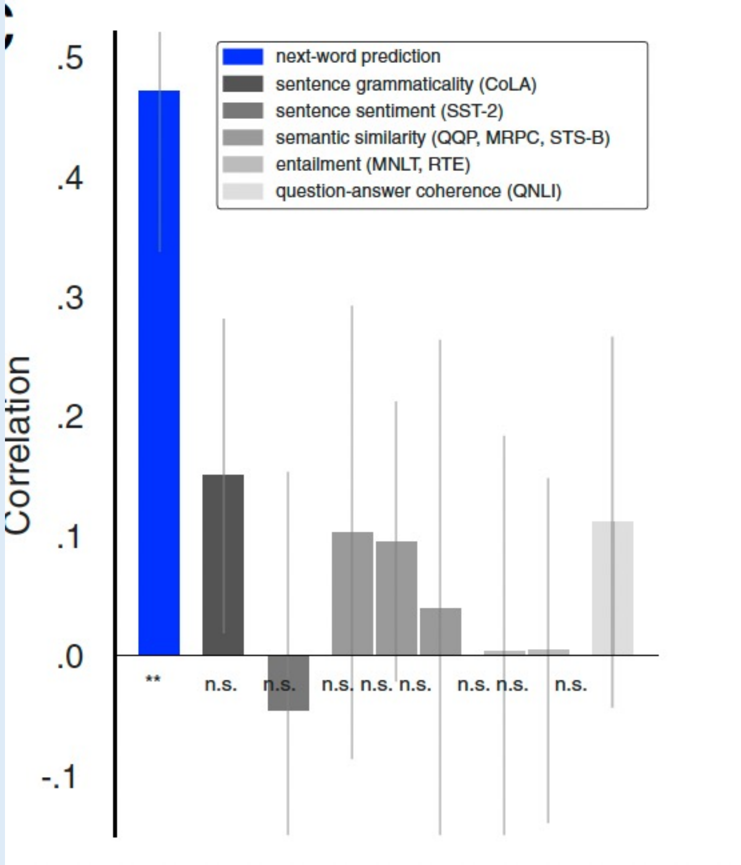


"Model performance on diverse benchmarks from the GLUE suite of benchmarks, including judgments about syntactic and semantic properties of sentences, was not predictive of brain or behavioral scores" (Schrimpf, 2021)

"Model performance on a next-word-prediction task selectively predicts brain scores" Schhrimpf et al., Fig. 3

A General Language Understanding Evaluation (GLUE) benchmark uncovers failure (Schrimpf)





“Model performance on diverse benchmarks from the GLUE suite of benchmarks, including judgments about syntactic and semantic properties of sentences, was not predictive of brain or behavioral scores” (Schrimpf, 2021)

“Model performance on a next-word-prediction task selectively predicts brain scores” Schhrimpf et al., Fig. 3

puzzling psycho-biological properties of the models: size of dataset in AI training and in human development

- *The size of training sets in AI and in human language acquisition*

AI: The GPT-2 model has 1.5 billion parameters; it was trained on 8m web pages

#words in a web page (10-25K), i.e., **~80 billion sentences**

A 4-year old child: (18000 x 30 x 48) / 8= **3.24 million sentences**

W/Day /Mo. #mo. W/S

At age 4, children master almost all of complex syntax

what children can do at age 3: VP-ellipsis

He drove a bus and Dennis drove a bus

He drove a bus and Dennis ***did, too*** ***~~[drive a bus]~~***

*CHI (Age 3): what kind of bus does he have?

*MOT: he has a Volkswager bus.

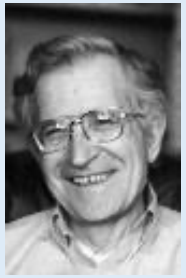
*CHI: Dennis does **too**.

*MOT: that's right Gary and Dennis have the same car.

*CHI: uhhuh [=yes]

DNNs has no biological feasibility with respect to training sets/acquisition

Grodzinsky et al., *OUP, 2020*



Selective deficits in aphasia: The shape of the syntax comprehension disorder

Same words, same meaning, different syntax

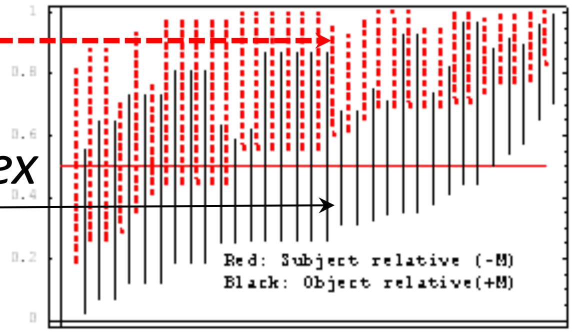
a. The woman who is covering the girl is old

b. The woman who the girl is covering is old

simple

complex

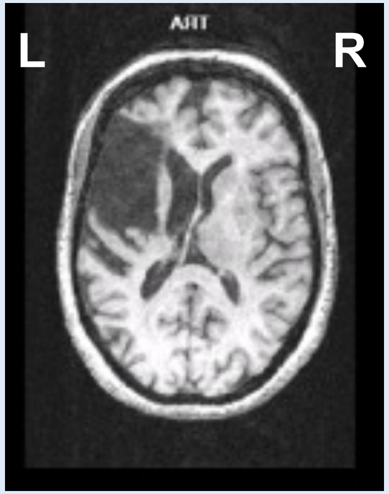
% correct



high

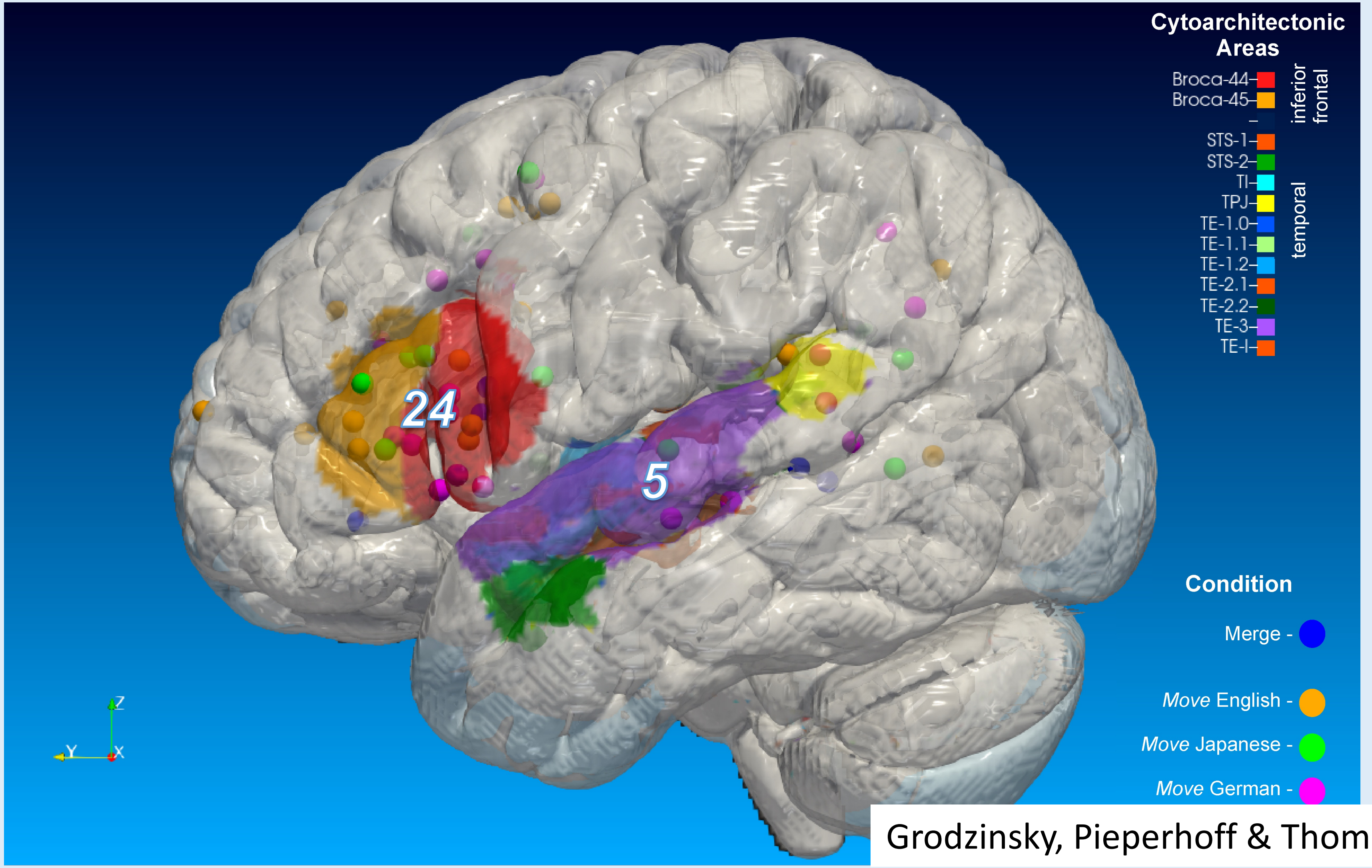
low

האישה שמכסה את הבחורה היא זקנה



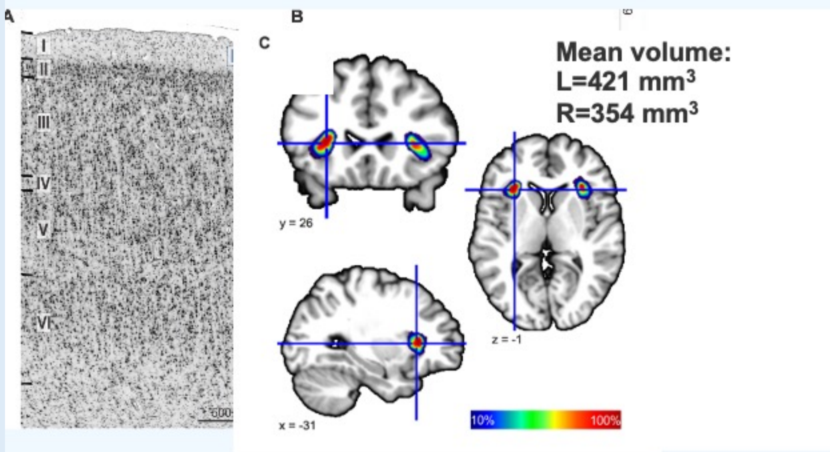
Drai & Grodzinsky, *Br. Lang*. 2006

psycho-biological relevance of the models' architecture: neural specificity of complex syntax (once much noise is removed)



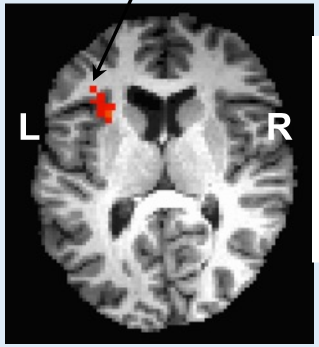
Grodzinsky, Pieperhoff & Thomspon, *Cortex*, 2021

semantics: the neural cost of (implicit) negation



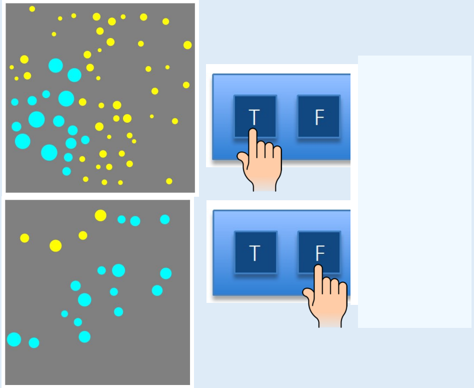
Id7 (Insular dysgranular area 7)

NetNeg cluster

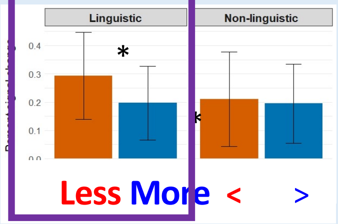


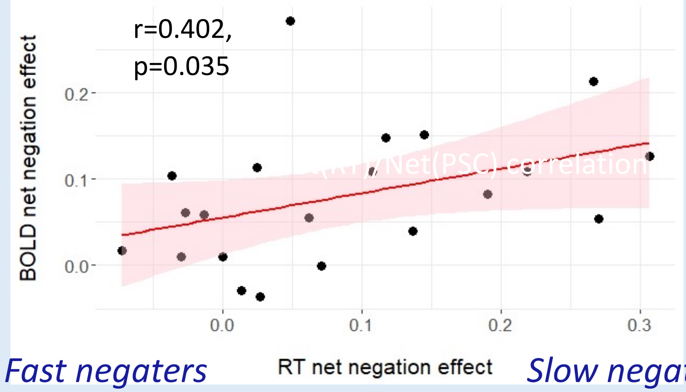
More than half of the circles are blue

Less than half of the circles are yellow



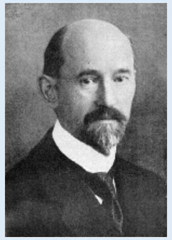
fMRI

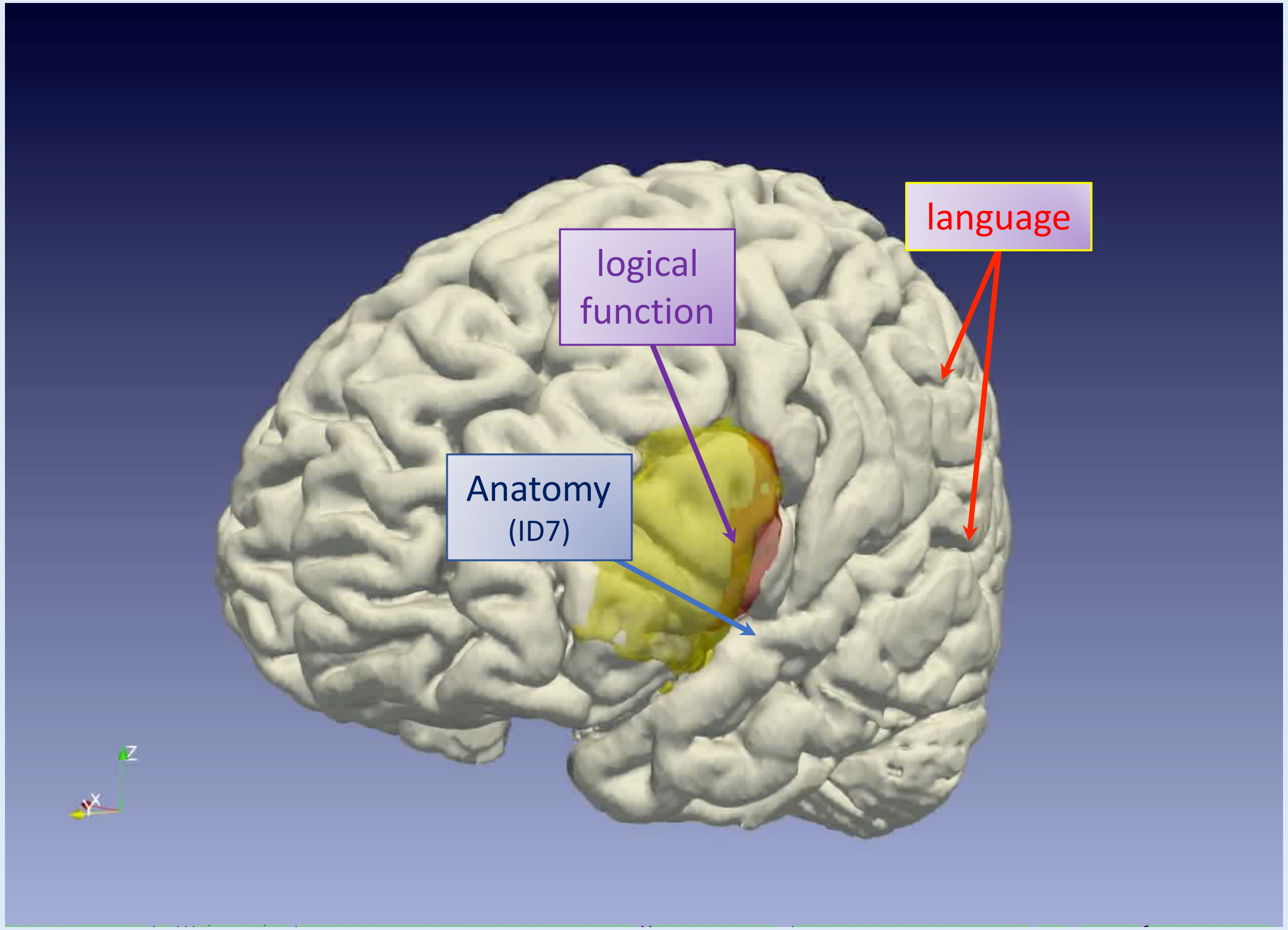


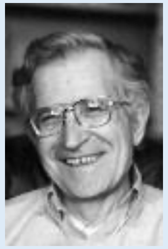


RT

comparing Id7's locus to the language regions

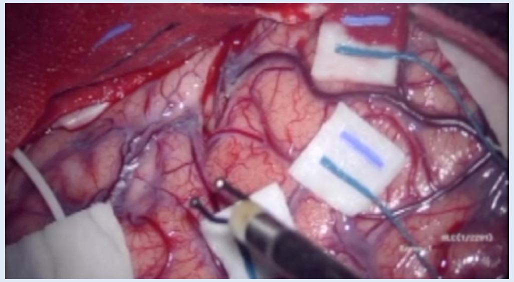


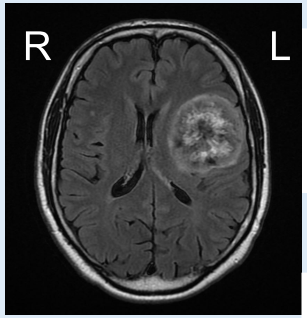




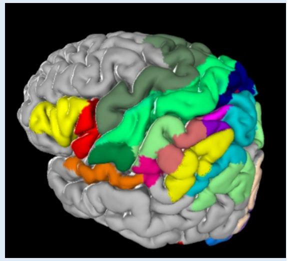
No overlap with the Broca's region

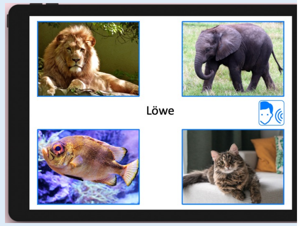
Application: intra-operative navigation with these materials





Onconeurosurgical Unit, Uniklinik Düsseldorf





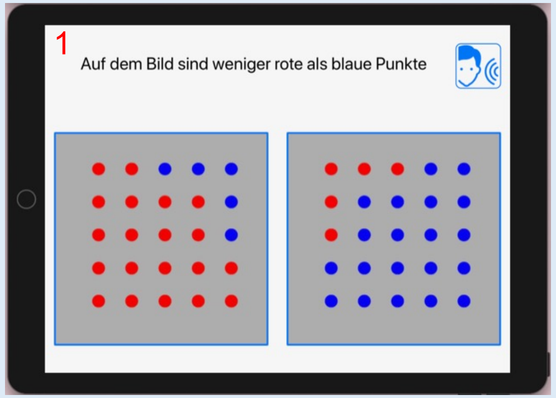
Test goal: maximize resection; minimize functional loss

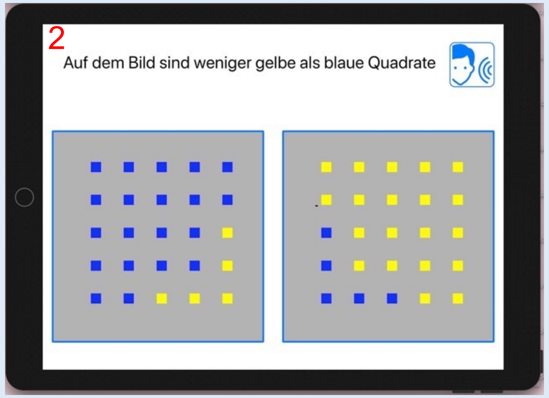
Test procedure:

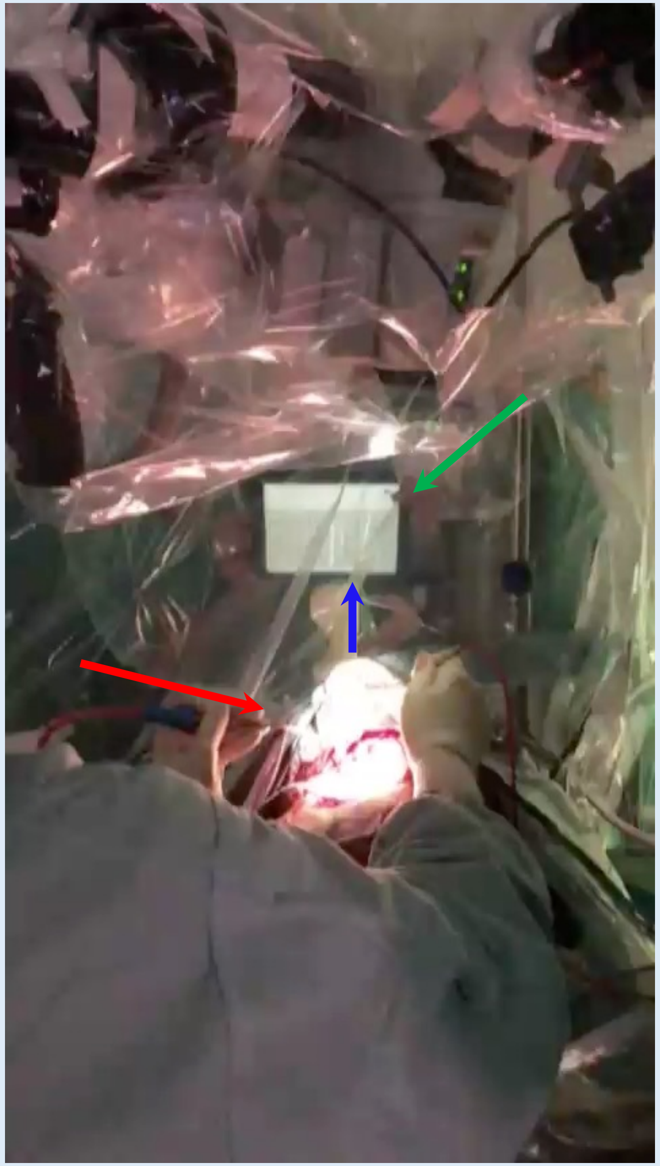
1. Pre-OP determination of stimulation points.
2. Patient is awakened.
3. Direct Electrical Stimulation with a concomitant linguistic stimulus.
4. Results are displayed, supporting surgical decisions.
5. Tests: words, syntax, semantics.

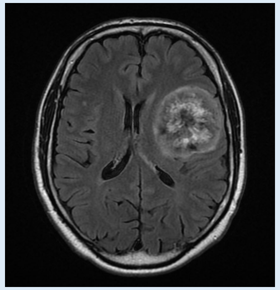
the semantic Polarity intra-op test, aimed to improve functional resolution – *more/less*

Onconeurosurgical Unit, Uniklinik Düsseldorf

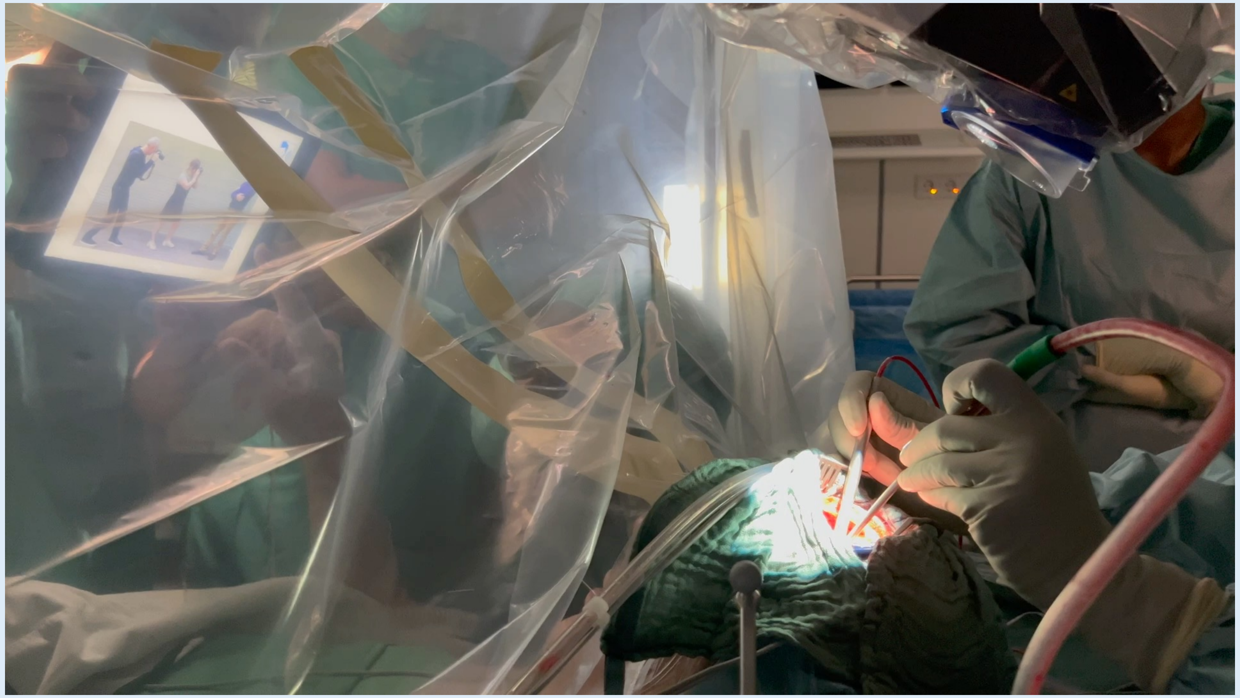








a syntactic recognition intra-op test, aimed to improve functional resolution – questions



conclusions

- Next word prediction is a task that hardly probes language
- DNNs have thus far been successful on everyday tasks, but have failed to exhibit serious human-like behavior

- The human brain exhibits a modular functional anatomy
- The studies noted above blur the both function and anatomy
- With a bit of luck, this modular approach can even be clinically useful

- “I've never fired anyone, and a linguist least of all”

(Jelinek, 2005)

the end

yosef.grodzinsky@mail.huji.ac.il

Funding:

NIH (NIDCD)

Canada Research Chairs (CRC)

Canadian Social Science and Humanities

Research Council (SSHRC)

Humboldt Foundation Research Award

Forschungszentrum Jülich

ELSC, HUJI

Israel Science Foundation

Isabelle Deschamps — *Québec*

Galit Agmon, Yonatan Loewenstein

Andreas Haida, I-An Tan, Assaf Brown

Kedem Snir — *HUJI*

Katrin Amunts, Stefan Heim,

Francesca Ianilli, Peter Pieperhoff,

Christiane Jockwitz , Svenja Caspers,

Nora Bittner — *FZ Jülich*

Marcel Kamp, Marion Rapp, Michael Sabel

Uniklink Düsseldorf

Virginia Jaichenco, María Elina Sánchez

U. Buenos Aires