Current Issues and Applications

Week 11

Part I: Current Issues

Current issues/questions in Semantic Theory

- Where is the border between semantics and syntax?
- Where is the border between semantics and pragmatics?
- What semantic universals exist, and how can they be captured formally?
- How can we apply and evaluate model theory at scale?

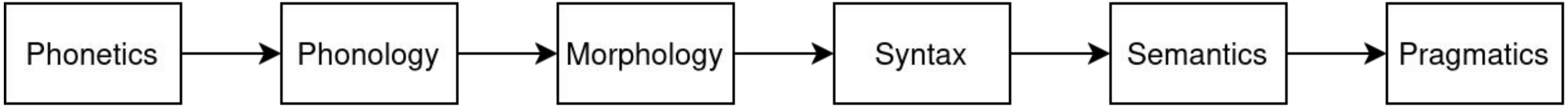
Pragmatics

- **Pragmatics** is the study of the interaction between meaning and context
  - Basic idea: semantics gives us the core/templatic meaning of a sentence, which is then completed by pragmatics

- For example: “*this rock is hot*”
  - On Earth (average temperature: 15°C): >45°?
  - On Venus (average temperature: 464°C): >700°?

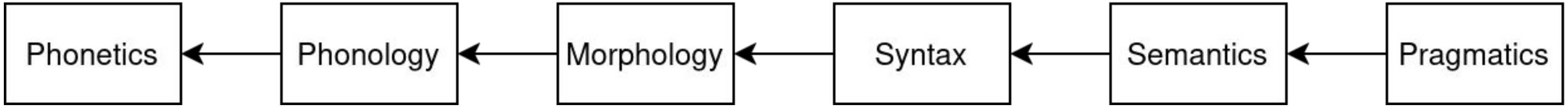
Feed-forward model of linguistic processing/production

Processing:



Phonetics → Phonology → Morphology → Syntax → Semantics → Pragmatics

Production:



Phonetics ← Phonology ← Morphology ← Syntax ← Semantics ← Pragmatics

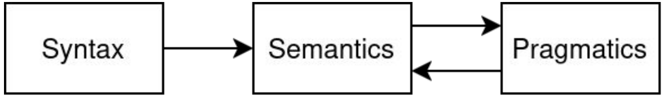
The semantics-pragmatics interface: anaphora

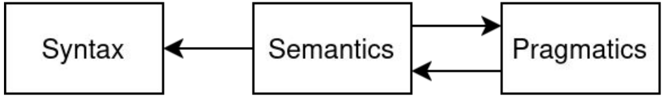
- Consider the following:

  the woman who met the teacher thought she would be happier

- Contextual information (pragmatics) is required to resolve the referent of “she”

- But (to a degree) semantic meaning *determines* contextual information:
  - *the man who met the teacher thought she would be happier*
  - *the woman who met the male teacher thought she would be happier*





The semantics-pragmatics interface: presuppositions

- Flashback: consistency and informativity constraints
  - Informativity: the resolved DRS should not be entailed by our background knowledge
  - Consistency: the resolved DRS must be satisfiable (taking background knowledge into account)
    - **Local consistency**: no sub-DRS can be inconsistent with any superordinate DRS

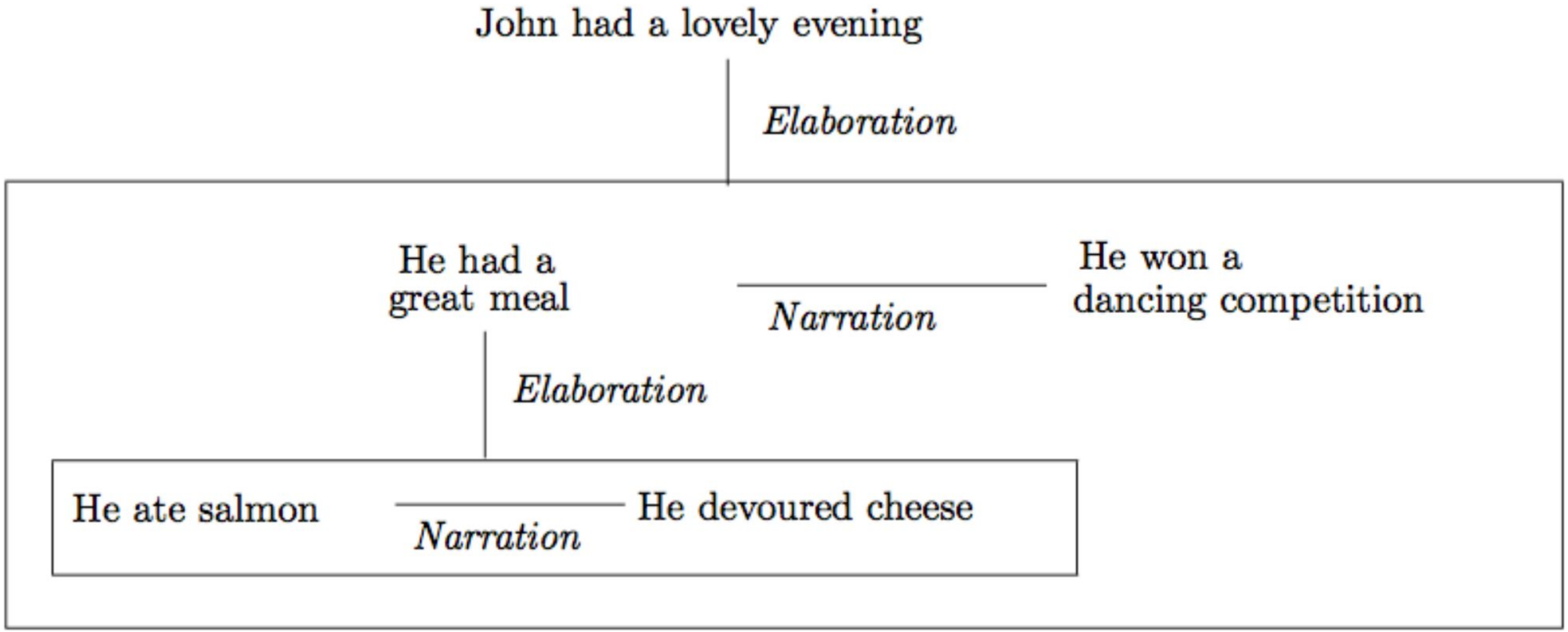
- "if there are no more communists in Paris, every communist here will revolt"
  ≫ "there are communists here"
  - **Unless** *here* = *Paris*

The semantics-pragmatics interface: rhetorical structure

- "*John had a great evening last night. He had a great dinner. He ate salmon. It was delicious.*"
- "*John had a great evening last night. He had a great dinner. He ate salmon. He ate so fast, he barely stopped to breathe. He was in heaven. It was delicious.*"
- "*John had a great evening last night. He had a great dinner. He ate salmon. He devoured lots of cheese. Then, he won a dancing competition.* ??*It was delicious.*"

- **Discourse relations** connect segments of discourse

The semantics-pragmatics interface: rhetorical structure



John had a lovely evening

Elaboration

He had a great meal

Narration

He won a dancing competition

Elaboration

He ate salmon

Narration

He devoured cheese

Semantic universals

- What universal regularities govern the meaning of terms? And how can we formalize them?
    - For example, there is no language with a term for *grue* (Goodman, 1955) $\textit{grue}(x) := \exists e_0 \forall e[(e < e_0 \rightarrow \textit{green}(e, x)) \land ((e_0 < e \lor e_0 \cdot e) \rightarrow \textit{blue}(e, x))]$

- Which subclasses of generalized quantifiers represent meanings of natural language noun phrases?

- What are the basic units of lexical meaning (*semes*)?

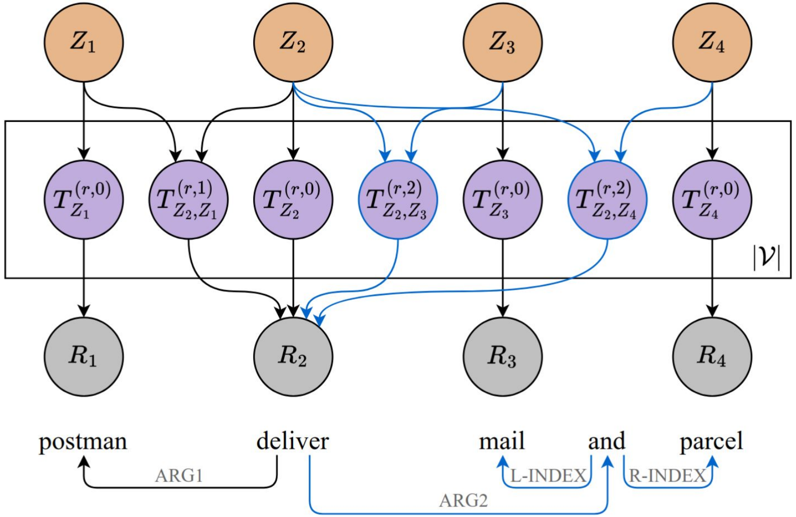
Model theory at scale

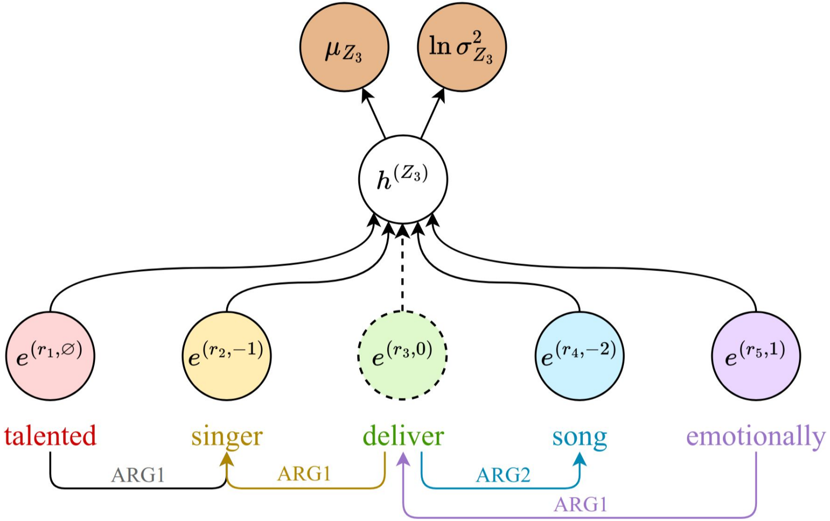
- The biggest problem with model theory: it isn’t feasible to construct large-scale models by hand
  - To do this, we need data-driven (distributional) approaches

- Problem: distributional data is too messy for discrete truth values ({⊤, ⊥})
  - Real-world data can be contradictory
  - Obvious facts are often left unstated
  - Automatic semantic parsing can be unreliable
    - Ambiguity

FDSAS

- Functional Distributional Semantics at Scale (FDSAS; Lo et al., 2023)
  - Probabilistic, truth-conditional distributional semantic model
  - FDSAS learns probabilistic *regions of truth* for predicates





Part II: Applications

LLM Evaluation

> *“... linguistic semantics is relevant to NLP because it helps us to understand how language works and thus positions us to critically evaluate claims that systems such as LLMs understand. Furthermore, if we want to build things that actually do understand, then linguistic semantics (and pragmatics, etc.) are going to be important”* (Emily Bender)

- We need semantic concepts to evaluate LLMs, and to determine what they’re doing wrong:
    - If our goal is human-like LMs, we need to understand human-like language use

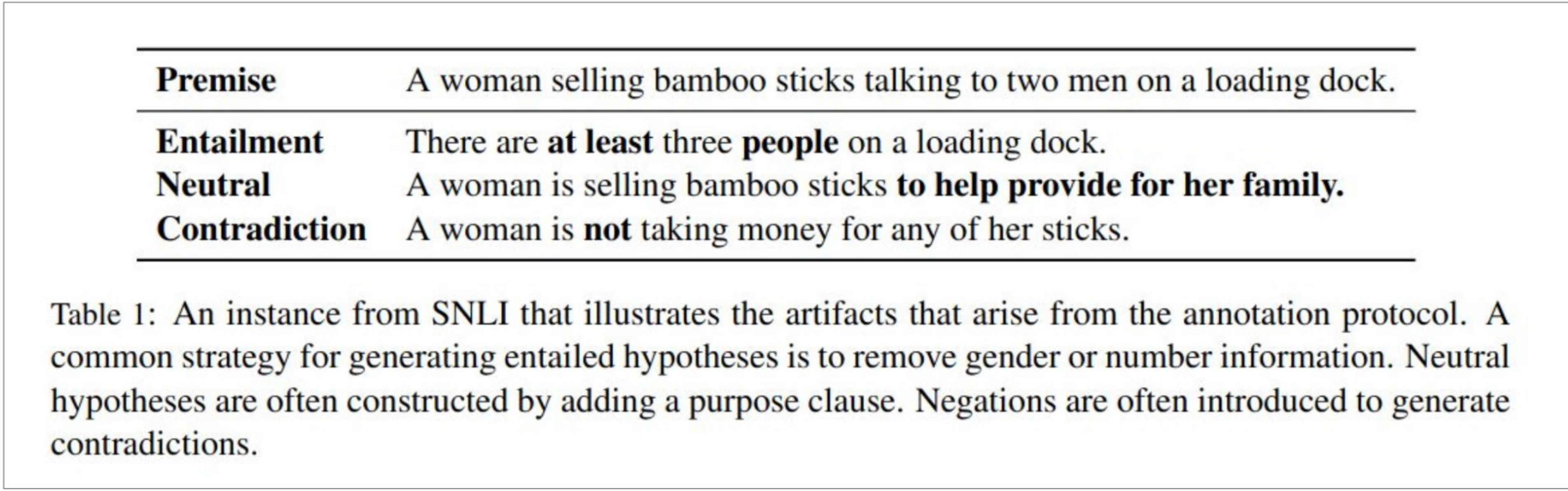
LLM Evaluation: NLI

- Natural language inference (NLI): classifying inferential relations. Given a pair of sentences (P, H), determine the label L:
    - L = *entailment*: $P \vDash H$
    - L = *contradiction*: $P \vDash \neg H$
    - L = *neutral*: $\neg((P \vDash \neg H) \vee (P \vDash H))$

- NLI evaluates LLM's ability to model entailment
    - Entailment = basic notion of (truth-conditional) meaning: understanding meaning ⊨ understanding entailment

NLI: right for the wrong reasons

- Models can reach high performance on this task without even seeing the premise (*P*), because the hypothesis (*H*) contains cues that indicate the label (Gururangan et al., 2018, Poliak et al., 2018)



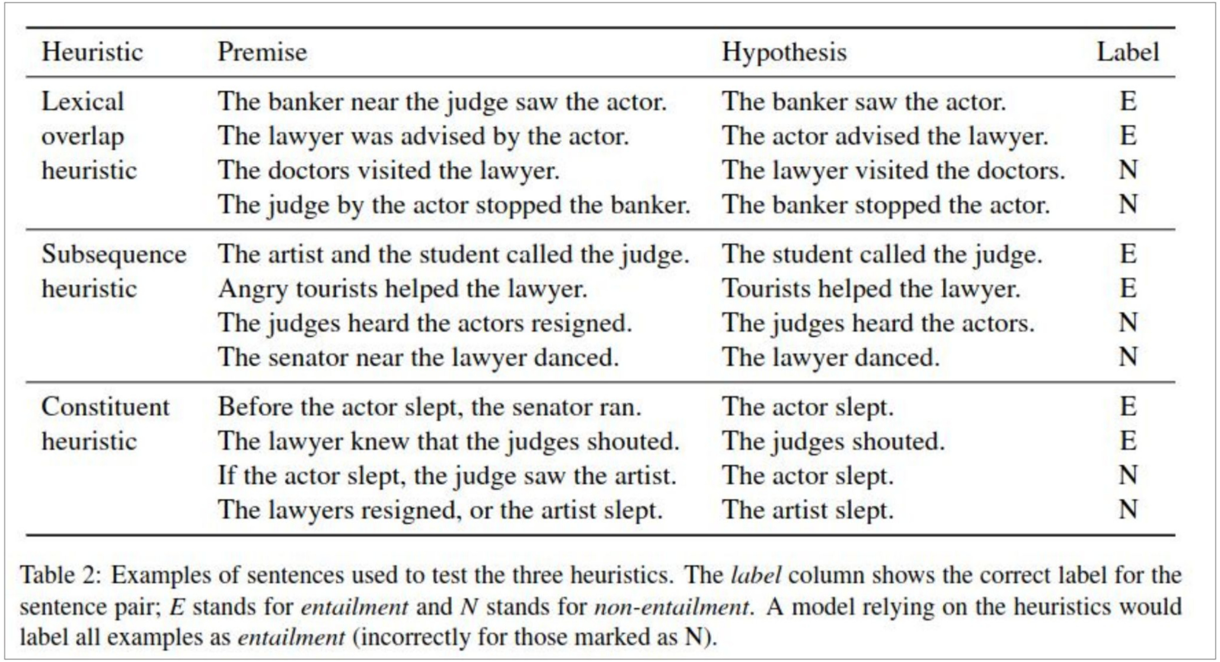
Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Table 1: An instance from SNLI that illustrates the artifacts that arise from the annotation protocol. A common strategy for generating entailed hypotheses is to remove gender or number information. Neutral hypotheses are often constructed by adding a purpose clause. Negations are often introduced to generate contradictions.

(Gururangan et al., 2018)

NLI: right for the wrong reasons

- Models can reach high performance on the task relying on heuristics between premise and hypothesis



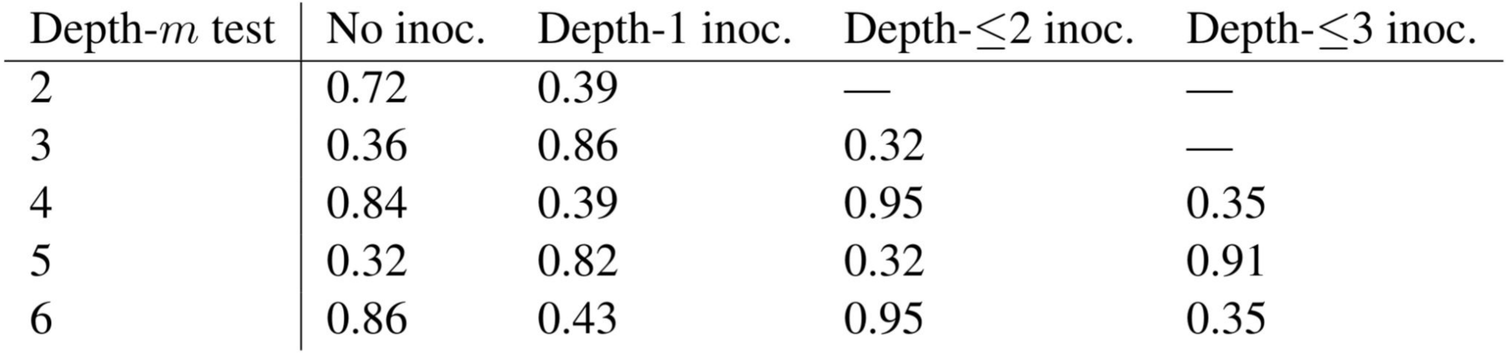
Heuristic	Premise	Hypothesis	Label
Lexical overlap heuristic	The banker near the judge saw the actor.	The banker saw the actor.	E
	The lawyer was advised by the actor.	The actor advised the lawyer.	E
	The doctors visited the lawyer.	The lawyer visited the doctors.	N
	The judge by the actor stopped the banker.	The banker stopped the actor.	N
Subsequence heuristic	The artist and the student called the judge.	The student called the judge.	E
	Angry tourists helped the lawyer.	Tourists helped the lawyer.	E
	The judges heard the actors resigned.	The judges heard the actors.	N
	The senator near the lawyer danced.	The lawyer danced.	N
Constituent heuristic	Before the actor slept, the senator ran.	The actor slept.	E
	The lawyer knew that the judges shouted.	The judges shouted.	E
	If the actor slept, the judge saw the artist.	The actor slept.	N
	The lawyers resigned, or the artist slept.	The artist slept.	N

Table 2: Examples of sentences used to test the three heuristics. The *label* column shows the correct label for the sentence pair; *E* stands for *entailment* and *N* stands for *non-entailment*. A model relying on the heuristics would label all examples as *entailment* (incorrectly for those marked as N).

(McCoy et al., 2019)

NLI and negation

- Models fail to learn the logical role of negation with respect to NLI (Sullivan, 2024):
  - (P, H) = *entailment* ⇔ (P, ¬H) = *contradiction*
  - (P, H) = *contradiction* ⇔ (P, ¬H) = *entailment*



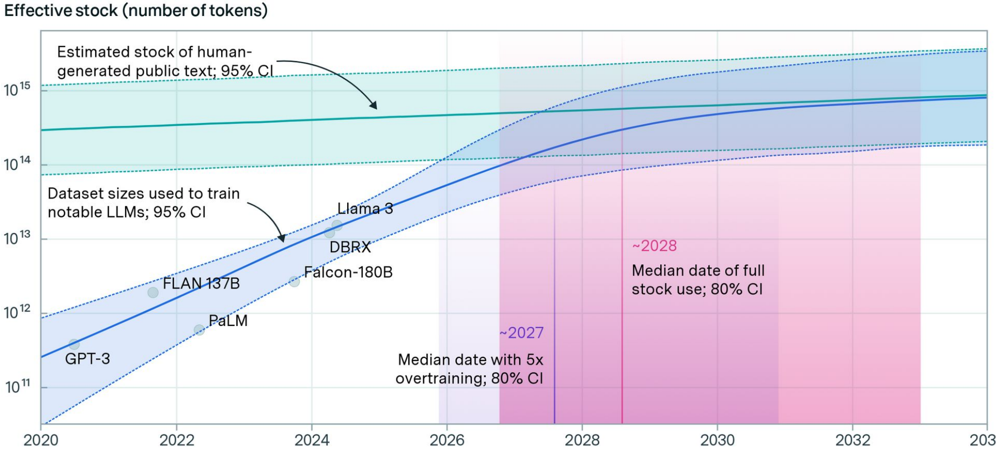
Depth-m test	No inoc.	Depth-1 inoc.	Depth-$\leq$2 inoc.	Depth-$\leq$3 inoc.
2	0.72	0.39	—	—
3	0.36	0.86	0.32	—
4	0.84	0.39	0.95	0.35
5	0.32	0.82	0.32	0.91
6	0.86	0.43	0.95	0.35

(Sullivan, 2024)

We are running out of training data

- Larger and larger LLMs require more and more training data:
  - GPT-2 (Radford et al., 2018): 1.5 billion parameters
  - GPT-3 (Brown et al., 2020): 175 billion parameters
  - GPT-4 (OpenAI, 2023): >1 trillion parameters (estimated)

- Villalobos et al. (2024): high-quality English training data will be exhausted sometime between 2026 and 2032



https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data

Linguistically-informed LMs

- Linguistically-informed LMs can improve over textual models, without using additional training data (e.g. Xu et al. 2021; Zhou et al., 2020; Zhang et al., 2020; etc.)

The Accelerated Learning Hypothesis (ALH): (Sullivan, 2025)

1. *Linguistically-informed LMs immediately begin learning more complex patterns, because the (aspects of) linguistic knowledge incorporated into such models obviates the need to learn elementary linguistic phenomena.*
2. *This accelerated learning of complex patterns allows linguistically-informed LMs to learn from less data than their textual counterparts.*

Semantics is better than syntax

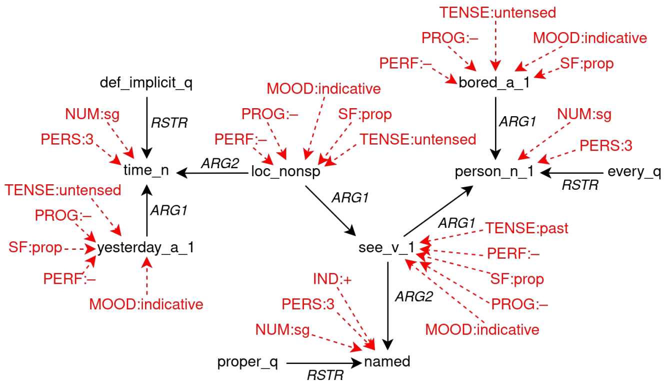
- Wu, Peng, and Smith (2021) and Prange, Schneider, and Kong (2022) show that semantically-informed LMs outperform syntactically-informed LMs

- Sullivan (2025): this is due to a syntactic/morphological de-noising effect:
  - Argument structure is explicitly annotated:
    - “*John saw Mary*” ⇒ *see*(*john*, *mary*)
    - “*Mary was seen by John*” ⇒ *see*(*john*, *mary*)
  - Morphological features are explicitly annotated*:
    - “*goose*” ⇒ $\textit{goose}_{\text{NUM:SG}}$
    - “*geese*” ⇒ $\textit{goose}_{\text{NUM:PL}}$

*in some frameworks

GFoLDS

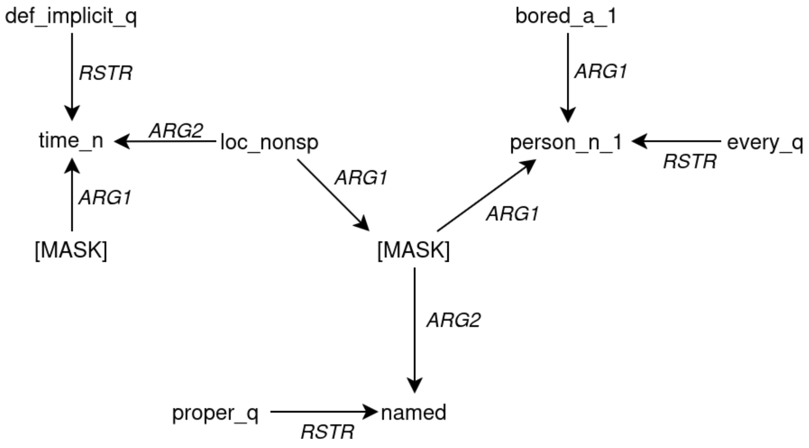
- Graph-based Formal-Logical Distributional Semantics (GFoLDS; Sullivan, 2025): pretrained graph transformer (Wu et al., 2021) over DMRS graph representations



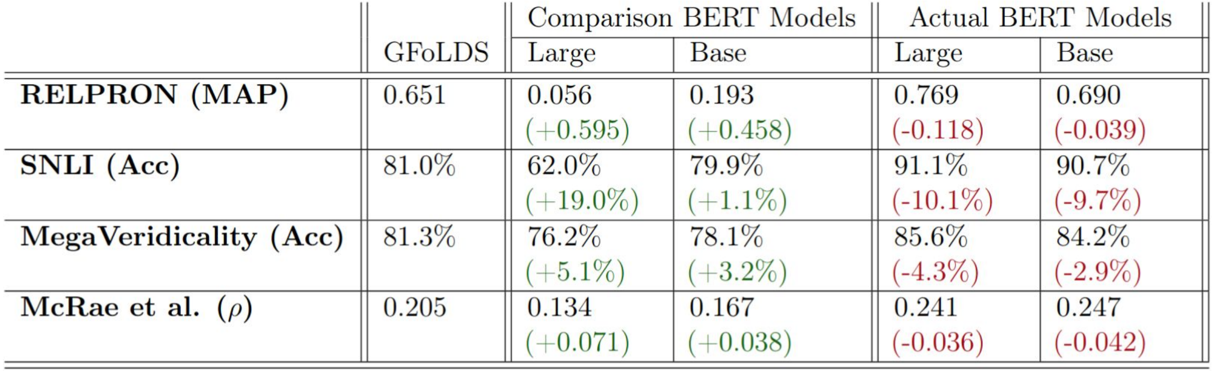
“Every bored person saw Mary yesterday”

GFoLDS: pretraining

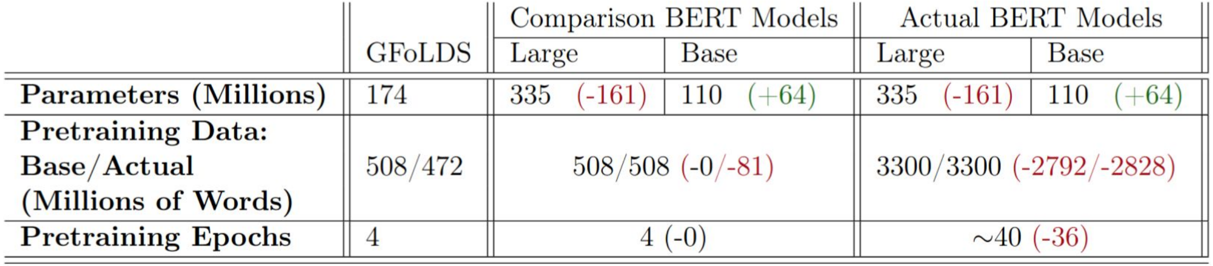
- GFoLDS is pretrained on 17.5 million sentences: ~6.5x smaller than BERT's pretraining dataset
- Pretraining objective: masked node modeling (MNM)



GFoLDS: experimental results



		Comparison BERT Models		Actual BERT Models	
	GFoLDS	Large	Base	Large	Base
RELPRON (MAP)	0.651	0.056 (+0.595)	0.193 (+0.458)	0.769 (-0.118)	0.690 (-0.039)
SNLI (Acc)	81.0%	62.0% (+19.0%)	79.9% (+1.1%)	91.1% (-10.1%)	90.7% (-9.7%)
MegaVeridicality (Acc)	81.3%	76.2% (+5.1%)	78.1% (+3.2%)	85.6% (-4.3%)	84.2% (-2.9%)
McRae et al. (ρ)	0.205	0.134 (+0.071)	0.167 (+0.038)	0.241 (-0.036)	0.247 (-0.042)



		Comparison BERT Models		Actual BERT Models	
	GFoLDS	Large	Base	Large	Base
Parameters (Millions)	174	335 (-161)	110 (+64)	335 (-161)	110 (+64)
Pretraining Data: Base/Actual (Millions of Words)	508/472	508/508 (-0/-81)		3300/3300 (-2792/-2828)	
Pretraining Epochs	4	4 (-0)		~40 (-36)	

Accelerated Learning Hypothesis

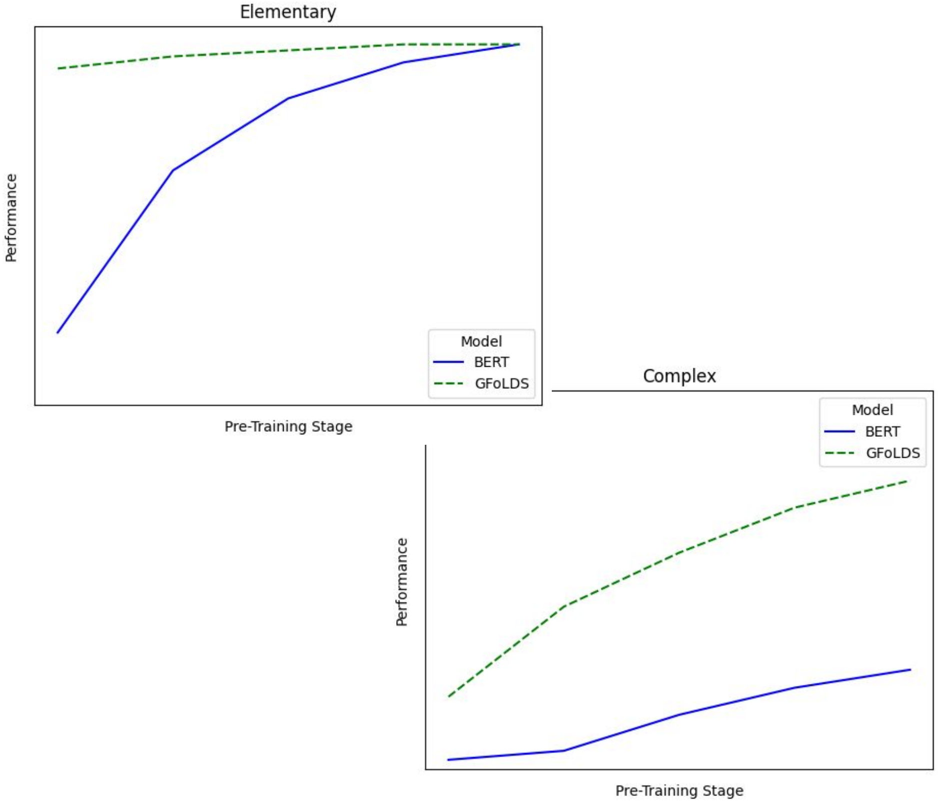
- Evaluated GFoLDS and the BERT comparison models at 80 intervals throughout pretraining

- Measure performance on:
  - Elementary tasks:
    - POS prediction
    - Quantifier agreement prediction
  - Complex task:
    - RELPRON

The Accelerated Learning Hypothesis (ALH):

1. *Linguistically-informed LMs immediately begin learning more complex patterns, because the (aspects of) linguistic knowledge incorporated into such models obviates the need to learn elementary linguistic phenomena.*
2. *This accelerated learning of complex patterns allows linguistically-informed LMs to learn from less data than their textual counterparts.*

ALH: prediction

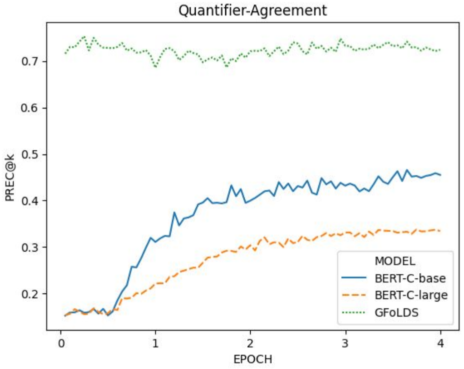


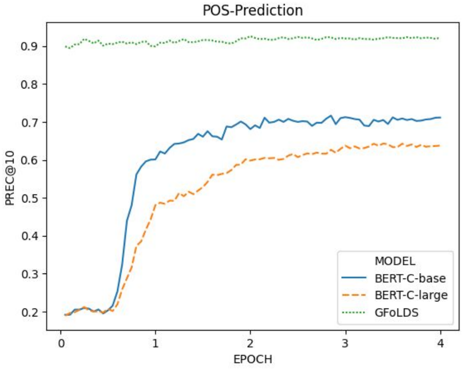
The Accelerated Learning Hypothesis (ALH):

1. *Linguistically-informed LMs immediately begin learning more complex patterns, because the (aspects of) linguistic knowledge incorporated into such models obviates the need to learn elementary linguistic phenomena.*
2. *This accelerated learning of complex patterns allows linguistically-informed LMs to learn from less data than their superficial counterparts.*

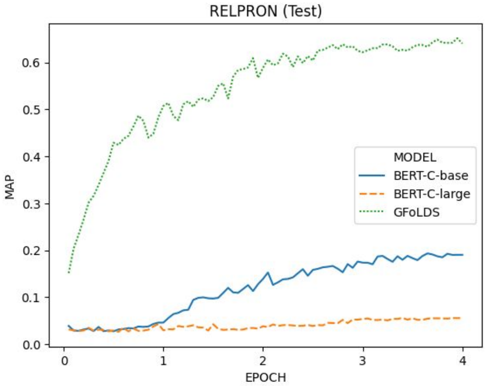
ALH: experimental results

Elementary:


Quantifier-Agreement


POS-Prediction

Complex:


RELPRON (Test)

But why do we need semantic *theory*?

- Semantic-theoretic correctness ensures representational consistency:
  - Semantic theory allows us to verify that our representational framework is accurately modeling meaning