

Some Challenges for Developing Computational Linguistic Resources

Sandiway Fong

University of Arizona

Email: sandiway@arizona.edu

Website: sandiway.arizona.edu

Megamind Plus Academy, Iraq via Zoom
9pm (Baghdad), September 19th 2026

Theme of talk

- Develop **Linguistic Resources**
 - for an under-resourced language
 - useful for computation

Linguistic Theory

- What can (modern) **Linguistic Theory** tell us?
- **Archaeology**: explosion of **symbolic works** in the **fossil record**
 - *coincides with the appearance of **modern humans** (200-300 tya)*
- (Chomsky 2021)
 - Language/thought, **I-Language**, an **authentic species property** (recent)
 - *Our closest relatives, otherwise intelligent apes, cannot begin to grasp the most elementary rudiments of language even with intensive training. They have about the same auditory system as humans, but acquire nothing from the sounds that lead a human infant, almost reflexively, to develop complex systems for constructing and expressing thought.*
- Basic structure of I-Language should be simple (**Merge**):
 - *the result of some [...] small rewiring of the brain [...] and has not changed [...] since.*



Knowledge of Language

- **I-Language** (Chomsky)
- We distinguish between **explicit** knowledge and **implicit** knowledge that all speakers possess
- Example: **Knowledge** (*explicit*):
 - you know how to read and write a **Python** program
 - **conscious propositional knowledge**
 - an explicitly learned skill (*could include an implicit component*)
- **Knowledge of Language** (*implicit*):
 - You obviously *wouldn't expect the native speaker of the language to be able to tell you anything very sensible about it, or for that matter, to tell you anything very sensible about any other aspect of his **knowledge**. He might, on occasion, be able to, but he might very well not be able to.*
 - *that knowledge is not consciously extricable.* (Chomsky, 1968 talk)

Knowledge of Language

- (Chomsky, 1968 talk)
 - *I don't see any way to avoid the conclusion that **knowledge of a language** involves some sort of **internal representation**.*
 - *... a **system of rules**, of what is called nowadays a **generative grammar**, meaning a **system of principles** somehow mentally represented.*
 - *Nobody knows anything about [the] **physical representation**, but which we can describe abstractly that expresses what the speaker of the language intuitively knows, ..., though it's **not the kind of knowledge that he can tell you about very often**.*
 - *then the next question you want to ask is **how is that knowledge put to use?***
 - *For example, ..., **can I construct a model**, ..., which **accounts for the ability of an ordinary human to accept some sort of signal written or spoken, let's say, and to produce the sort of structural description**,...*



Computational Linguistics:
Parser

Knowledge of Language

- (Chomsky, 1968 talk)
 - *All operations in all languages [...] are **structure-dependent**.*
 - *It's an **operation** that doesn't depend on the substantive content of the **string of elements** in question, but rather on an **abstract structure** associated with it.*
- Example (*aux-fronting question-formation in English*):
 - John's **having been** here yesterday **was** a surprise
 - the boy who **was** here **was** sick



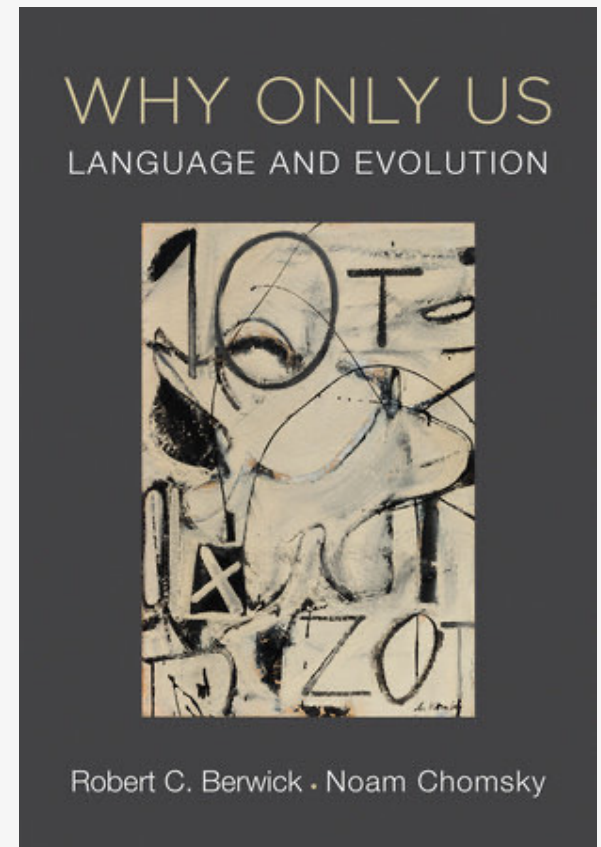
Knowledge of Language

- **Induction** (*from visible data*):
 - *I think you find that many of the principles that are present in the functioning of the grammar are not many... are **not obtainable from the data by any imaginable induction**.*
 - *And furthermore, I think one can say not only that these **principles are common to all normal speakers of English**, but that **very similar principles**, if you move to the right level of abstractness, are also **true of every other language** that we know of them.*

Knowledge of Language

- **Innateness**

- Chomsky (1968), see also Berwick & Chomsky's (2015) *Why Only Us* 
- *On what grounds is the **data in the acquisition of language** held to be so **limited** that **only the theory of innate knowledge** can account for the acquisition of language?*
- *I would argue two things. First of all, that **[induction]** have **absolutely no generality**, that **they would be unique [to] language acquisition**, as far as we know.*

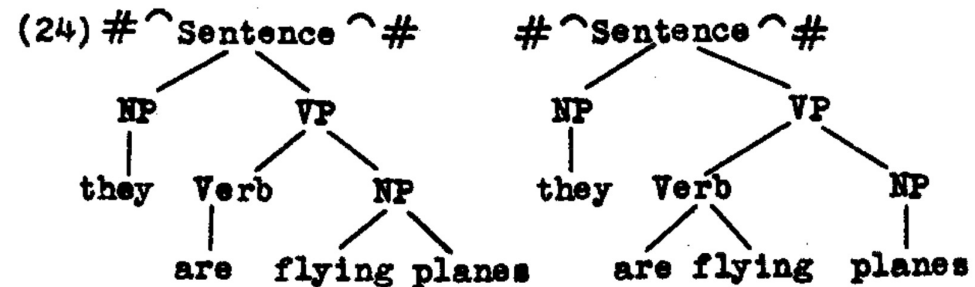


Knowledge of Language

- **Language:** *discrete infinity: infinite employment of finite means*
 - **Galileo** expressed his amazement at what is, in fact, an astonishing fact; with a **finite number of symbols** one can construct in the mind an **infinite number of linguistically formulated thoughts** and can even go on to reveal to others who have no access to our minds their innermost workings.
- 2023 Keio Lecture 2 (00:36) (Chomsky)
- **Innateness**
 - W. von Humboldt (*On Language*, 1836)
 - "Language is an **involuntary emanation of the mind**, no work of nations, but a gift fallen to them by their inner destiny."
 - "Every language is produced by the **original tendency** or '**original talent**' shared by all human beings."

Parsers

But this set of rules provides us with two non-equivalent derivations of the sentence "they are flying planes", reducing to the diagrams:



(Chomsky 1956)

Hence this sentence will have two phrase structures assigned to it; it can be analyzed as "they - are - flying planes" or "they - are flying - planes." And in fact, this sentence is ambiguous in just this way; we can understand it as meaning that "those specks on the horizon - are - flying planes" or "those pilots - are flying - planes." When the simplest grammar automatically provides nonequivalent derivations for some sentence, we say that we have a case of constructional homonymity, and we can suggest

Parsers

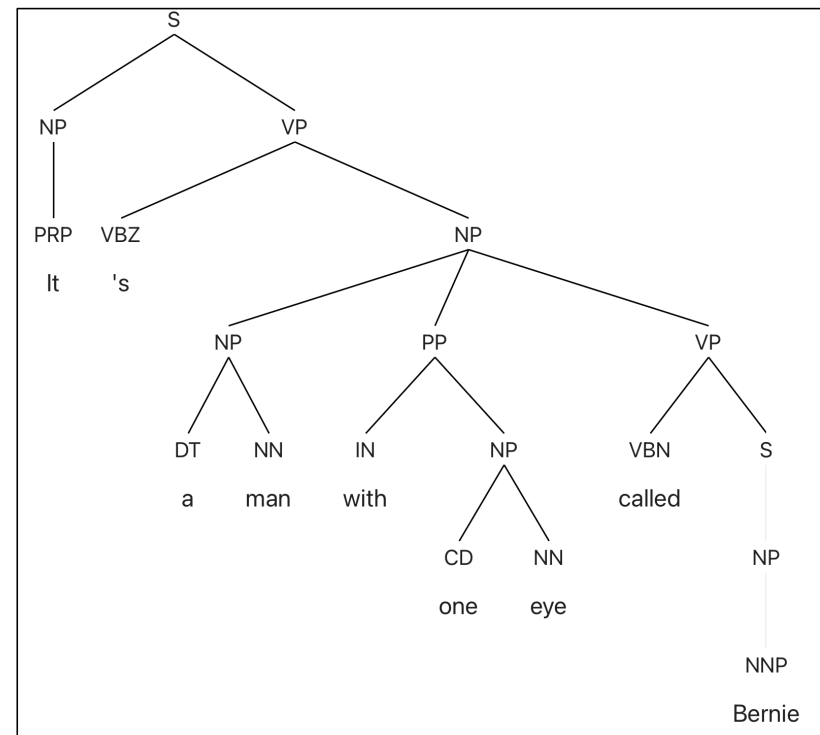
Mark: Who's the cook in this cafeteria?

Gary: It's a man with one eye called Bernie.

Mark: What's the other eye called?

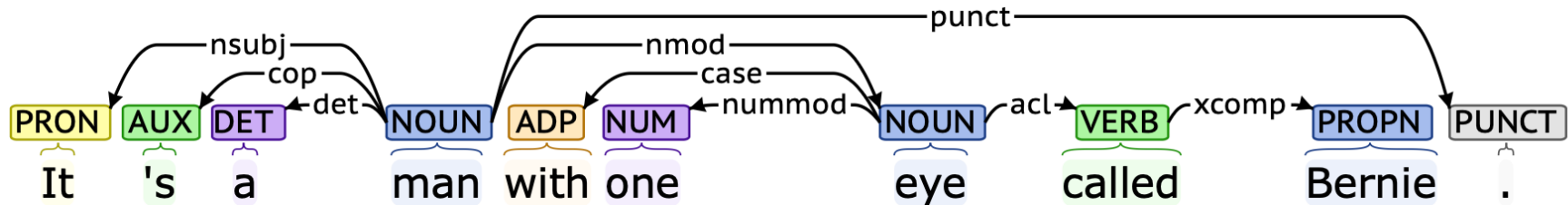
parser.kitaev.io

Berkeley Neural Parser



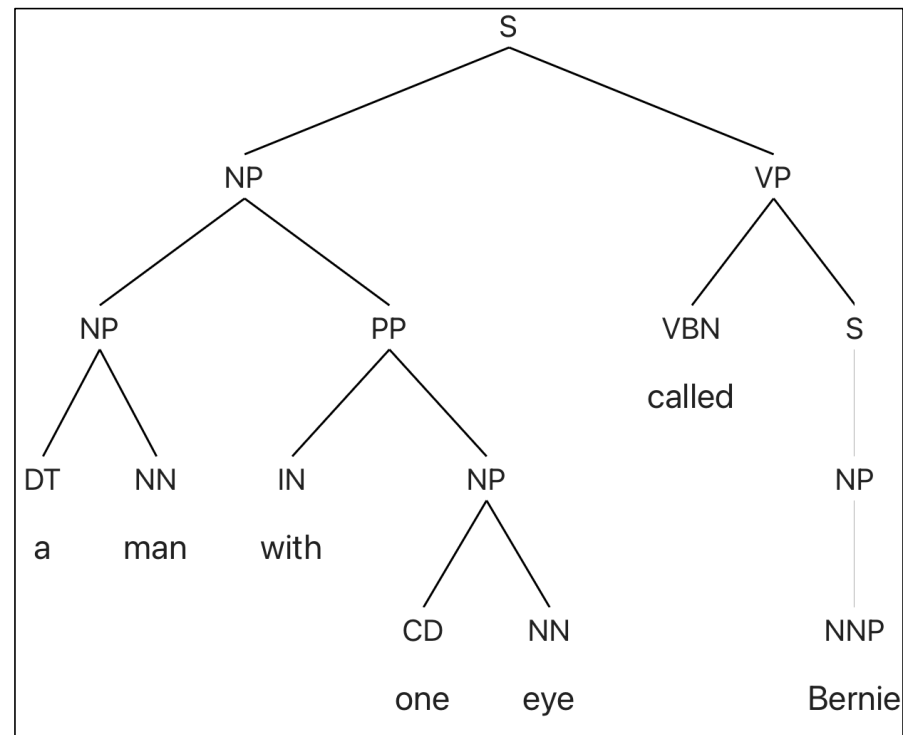
Dependency Grammar

- Stanza (<https://stanza.stanford.edu>)



- An open clausal complement (**xcomp**) of a verb or an adjective is a predicative or clausal complement without its own subject.
- **acl** stands for finite and non-finite clauses that modify a nominal.
- <https://universaldependencies.org/#language-en>

Parsers



Do **LLMs** understand language?

- ChatGPT (**GPT-5.6 Luna**): *what is the joke?*
 - *Cows kill more Americans each year than sharks*

The joke is based on **misdirection** and an **absurd comparison**.

"**Cows kill more Americans each year than sharks do**" sounds like it is setting up a surprising statistical fact. The humor comes from imagining cows as a more dangerous enemy than sharks—even though cows are peaceful farm animals.

- *Cows kill more Americans each year than sharks **do***

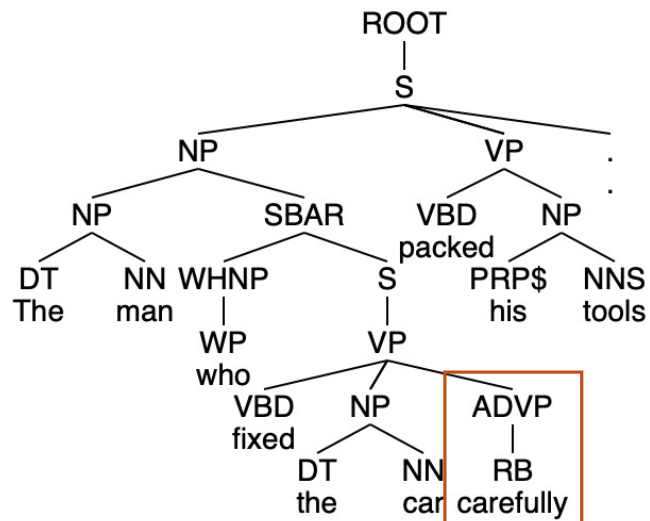
The joke is based on **unexpected danger** and **misplaced fear**.

People are famously afraid of **sharks** because sharks are portrayed as terrifying killers in movies like *Jaws*. 🦈 But cows seem completely harmless and even comical. 🐮

Stanford PCFG Parser *k*-Best Mode

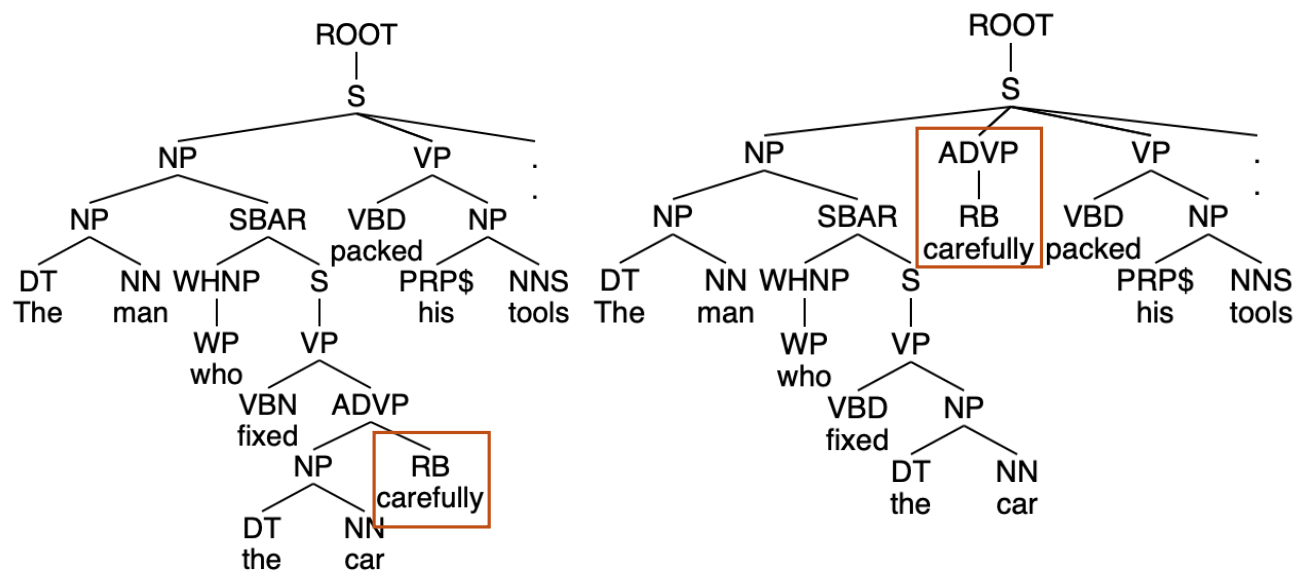
Parse 1 with score -76.55970804393291

Parse 2 with score -77.13352148234844



Parse 3 with score -77.16154663264751

Parse 5 with score -77.32319156825542



Parse 4 with score -77.19331233203411

Problems with Probabilistic Models

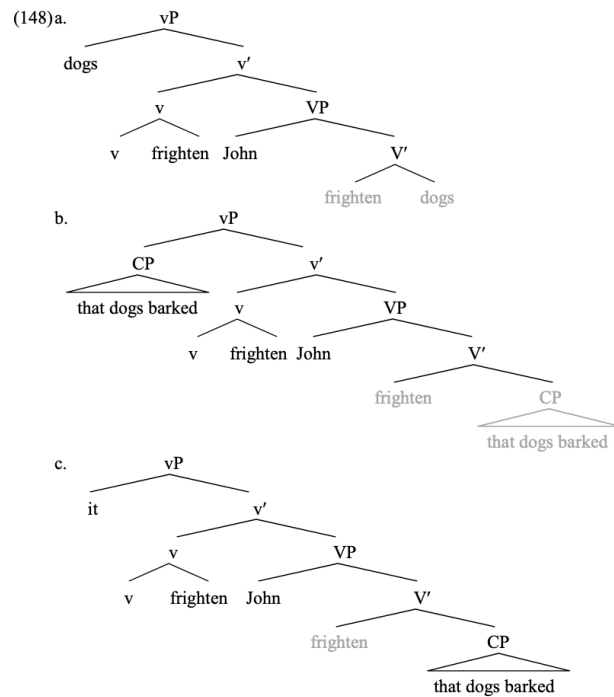
Ranking results (5!):

- #1 *Furiously ideas sleep colorless green.*
- #2 *Furiously green ideas sleep colorless.*
- #3 *Green ideas sleep furiously colorless.*
- ...
- #23 *Colorless green ideas sleep furiously.*
- ...
- #36 *Furiously sleep ideas green colorless.*
- ...
- #120 ...

- Chomsky (1956):

In § 2.4 we argued that there is no significant correlation between order of approximation and grammaticalness. If we order the strings of a given length in terms of order of approximation to English, we shall find both grammatical and ungrammatical strings scattered throughout the list, from top to bottom. Hence the notion of statistical approximation appears to be irrelevant to grammar. In § 2.3 we pointed

Problems with Syntactic Theories



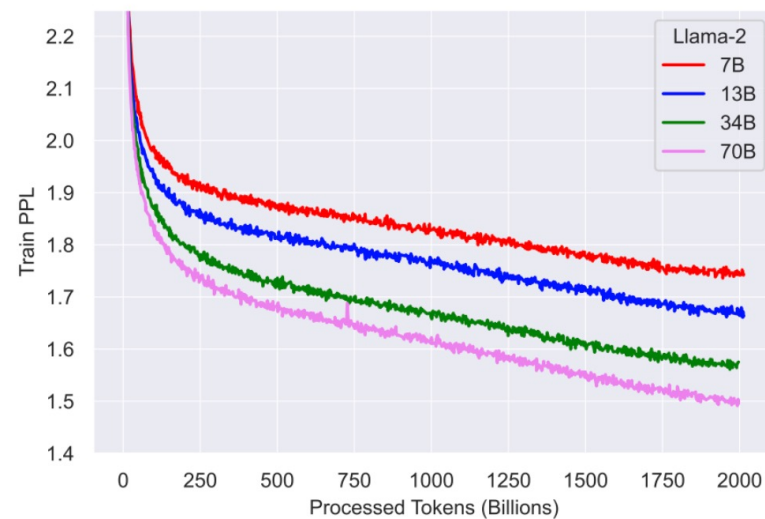
- **Pre-SMT**

- too complex

Large Language Models (LLMs)

- **LLMs** only work well with massive amounts of data to train on
- This creates a **Digital Divide**

Iteration on pre-training is very expensive

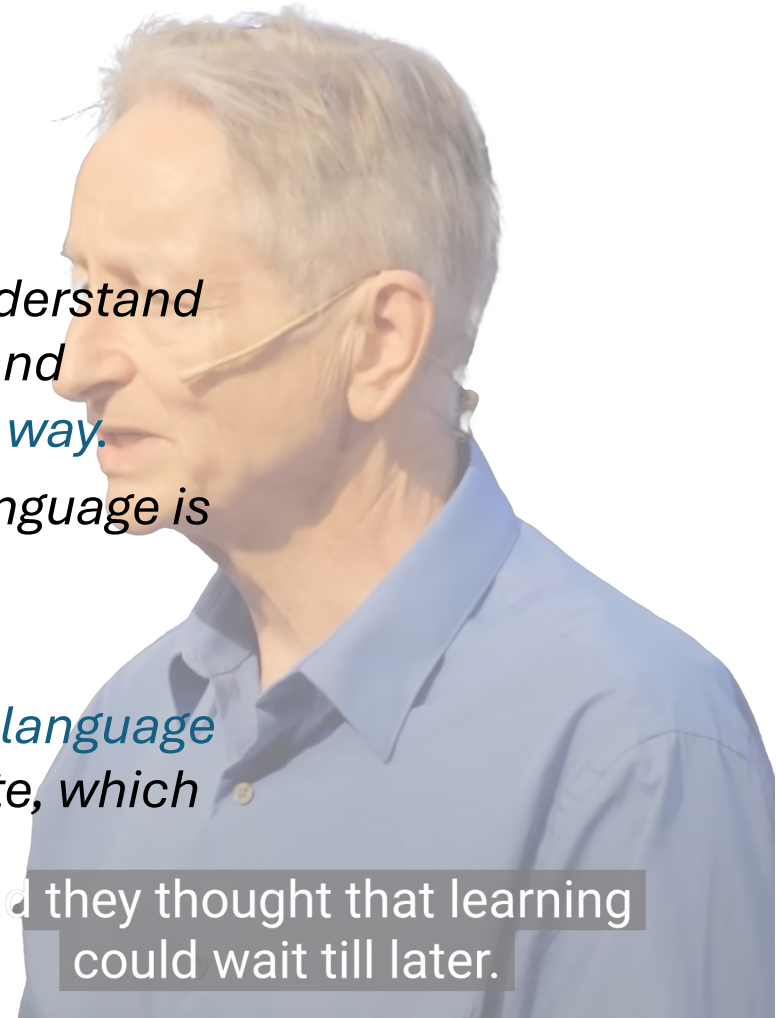


[Llama 2: Open Foundation and Fine-Tuned Chat Models](#), Hugo Touvron, Louis Martin, Kevin Stone et al. (2023)

Hinton vs. Chomsky

- Hinton *Royal Institution* lecture (2025):
 - *That's what understanding is when you understand language and when these [LLMs] understand language. We understand in just the same way.*
 - *And that's a much better model of what language is than anything the linguists ever had.*
 - *The linguists of course hate it.*
 - *[Linguists] also thought that knowledge of language was innate, knowledge of syntax was innate, which is just stupid.*

And they thought that learning
could wait till later.



LLMs vs. Linguists

- **Modern language models refute Chomsky's approach to language**

(Piantadosi, *lingbuzz* 2023; 2024)

- *The rise and success of [LLMs] undermines virtually every strong claim for the innateness of language that has been proposed by **generative linguistics**.*
- ***Modern machine learning** has subverted and bypassed the entire theoretical framework of Chomsky's approach, including its core claims to particular insights, principles, structures, and processes.*

- **How Linguistics Learned to Stop Worrying and Love the Language Models**

(Futrell and Mahowald 2025)

- ***Formal linguistics** has not presented an alternative model with the demonstrated **practical language-learning capabilities of neural models**.*
- *Finding these **internal mechanisms** is not a prerequisite for the reality of linguistic structure — **the mechanisms may be irreducibly complex**.*

Chomsky reply

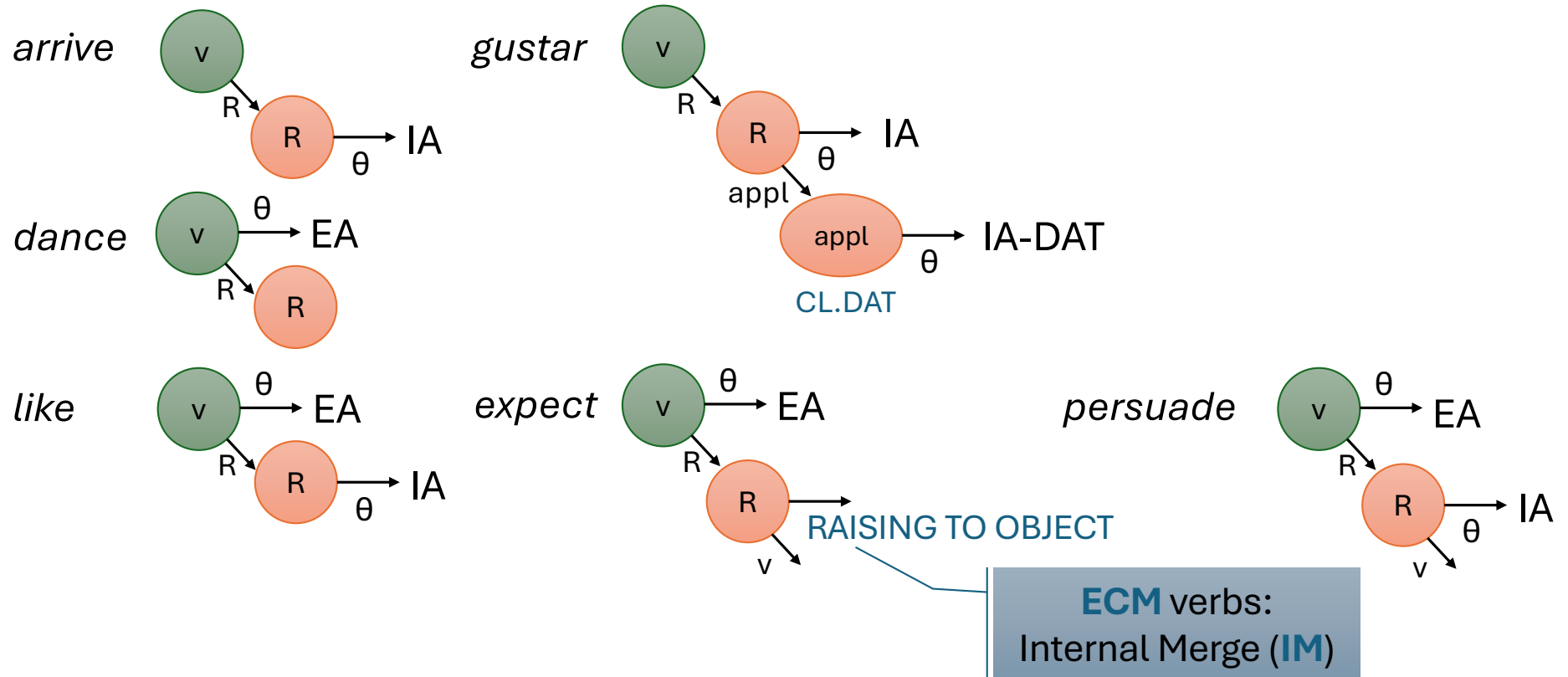
- Chomsky talk (2023):
 - ... so the enthusiasm for deep learning, like the *state of the art GPT-3* system of language processing that makes headlines, because it can produce text that looks like actual language, and we can ask the same question, **how does it fare with impossible languages?**
 - Does it just do just as well or better on systems that violate the basic principles of language.
 - **Andrea Moro:** *there are no impossible languages for [LLMs]!*
 - So, therefore, this multimillion dollar deep learning neural network trained on a 45 TB corpus of text, and running on the world's largest supercomputers **totally knows nothing about language**

Tokenization

- **Language-particular** (linear) process (*explicitly learned*):
 - *important for computational linguistics* (**stemmer**)
 - decoder for **inflectional morphology**:
 - Person/Number/Gender/Case System (*split ergativity*)
 - Clitics/Passive/Causative morphemes
 - Serial Verbs (*monoclausal*), e.g. Japanese, Chinese
 - **derivational morphology** (English):
 - adjective **-ly** (adverb), nominal **-ize** (verb), verb **-tion** (noun), $\emptyset$ (verb)
- Example (Spanish):
 - **Words:** *dos trenes llegan*
 - **Tokens:** $\text{llegar}_{\theta} \text{ }_{VR:\text{present}} \text{INFL}_{-v:*\text{2pl}/\text{3pl}} \text{tren}_{\text{dos}}$
 - **Word/affix order:** *det-noun* vs. *noun-det* (e.g. Malay)

```
vStem(anduve, Root, TNS, AGR).  
Root = andar, (walk)  
TNS = past,  
AGR = 1sg.
```

θ -theory: *small number of molecules*



θ -theory

- Language acquisition:

Agreement (MS):

Me gusta el libro
me.SG.DAT please.**SG**.PRES the.**SG**.M book.**SG**.M

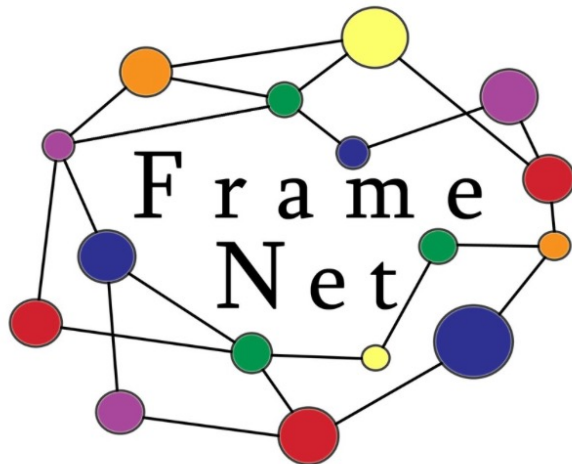
Me gustan los libros
me.SG.DAT please.**PL**.PRES the.**PL**.M book.**PL**.M

"I like the book/books"

$\{v_{\text{appl}}, \{IA, \{R_{\text{appl}:\theta}, \{\text{appl}_{\theta}, IA\text{-DAT}\}}\}\}$

Need **LEX**

Welcome to FrameNet!

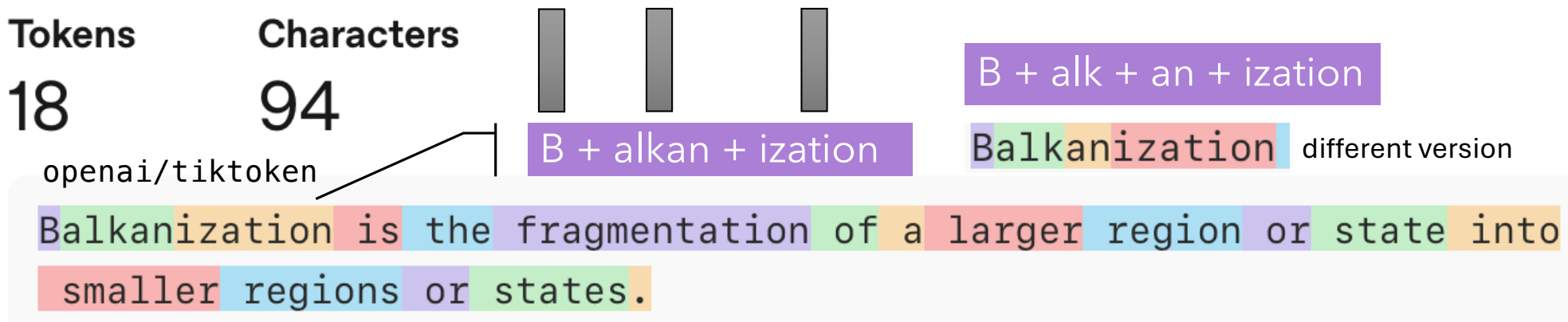


please.v

Frame Element	Type
[1] Experiencer	Core
[2] Stimulus	Core
[3] Degree	Peripheral
[4] Manner	Peripheral
[5] Means	Peripheral
[6] Time	Peripheral
[7] Depictive	Extra-Thematic
[8] Result	Extra-Thematic
[9] Circumstances	Extra-Thematic
[10] Explanation	Extra-Thematic

Word Embeddings: *vectors for **tokens***

- *How does a **LLM** know the meaning of unknown words?*
- **Answer:** it can't
- **Tokenization:** sub-word (~4 characters) (OpenAI's **BPE**)



- **Note:** not a *bag-of-words* scheme: a unique **position vector** is added to each **token vector**; e.g. **GPT-2**: 1024 token context size

Python: openai/tiktoken

- **Example:**

- `pip3 install tiktoken`
- `import tiktoken`
- Example:
- `>>> import tiktoken`
- `>>> enc = tiktoken.get_encoding("o200k_base")`
- `>>> enc.encode("sod-it-why-not")`
- `[82, 368, 46343, 12, 60291, 36155]`

- **Encode/decode:**

- `enc.encode(string)`
- `enc.decode(list)` `enc.decode_single_token_bytes(int)`

- **Encodings:**

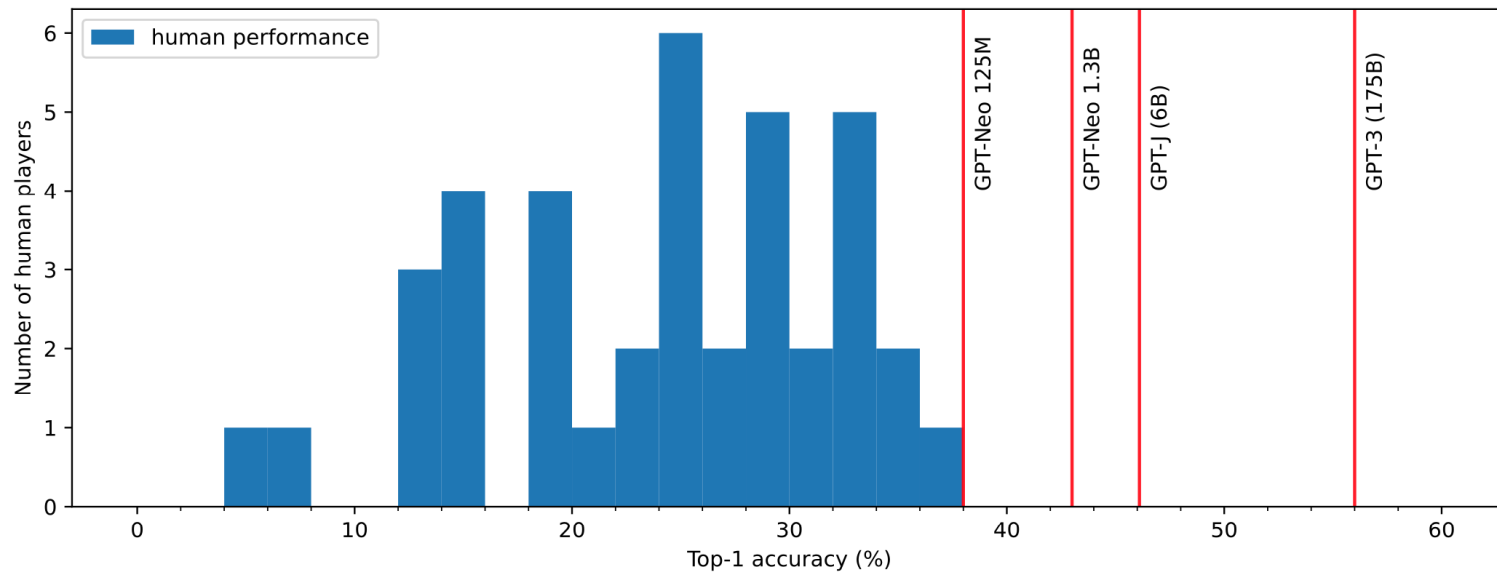
- `o200k_base` `gpt-4o, gpt-4o-mini`
- `cl100k_base` `gpt-4-turbo, gpt-4, gpt-3.5-turbo, ...`
- `r50k_base` `GPT-3 models like davinci`

GPT-2 on ungrammatical/nonsense input

fe fi fo fum fumf . tibii	hat sat pat ive a d ia loo urna sat	HUMUHUMUNUKU C HUMU HUMUN
--	--	--

Top-1 Prediction

- **GPT** (*Generative Pre-trained Transformer*)
 - *unsupervised training*, good at *next token prediction* (Shlegeris et al. 2024)
 - **better than humans!**



Can LLMs determine syntactic structure?

- *The structures present in large language models can be seen in detailed analyses, where as the model generates text after training, **its internal states represent latent aspects of syntactic structure and semantic meaning.***

