A model-theoretic approach to evaluating the strong generative capacity of large language models

Tommi Buder-Gröndahl

Introducing myself

Cognitive scientist working at the intersection of linguistics, philosophy, and AI

Postdoctoral researcher at the Department of Digital Humanities of the University of Helsinki

Education

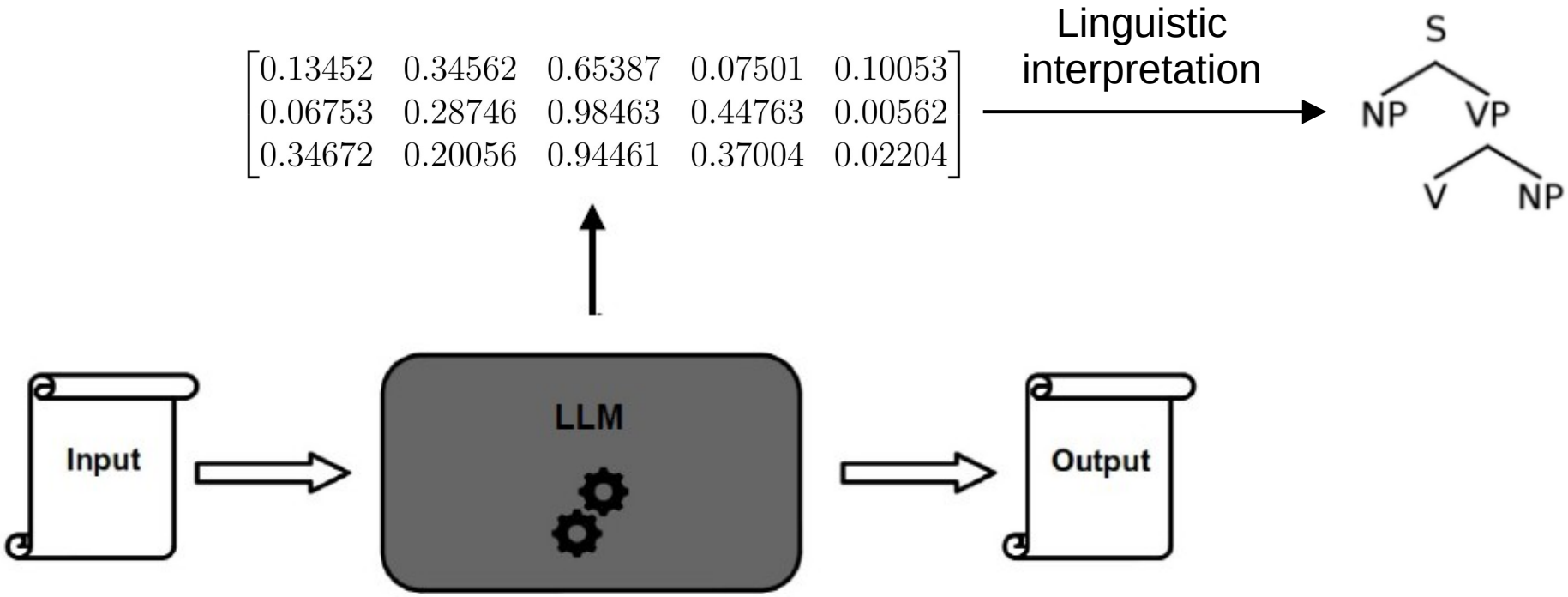
- DSc in computer science (Aalto University, 2021)
- PhD in cognitive science (University of Helsinki, 2023)

tommi.grondahl@helsinki.fi
https://tommi-bg.fi/
https://bsky.app/profile/tommibg.bsky.social

Current/forthcoming work

- PI of the project *Linguistic Interpretation of Leading Techniques in Natural Language Processing*, funded by the Research Council of Finland (2022–2025)
- Next fall (2025), starting as a postdoc on the project *AI for Reinforcing Democracy (AIDemoc)*, led by Anna-Mari Wallenberg (Uni. Helsinki) and funded by the Strategic Research Council of the Research Council of Finland (2025–2030)

Linguistic interpretation of LLMs: motivation



Linguistic interpretation of LLMs: challenges

Competence vs. performance (Dupre 2021)

- LLMs are typically evaluated based on performance: success across input-output tasks
- Linguists often aim to analyze competence: abstract aspects of cognitive architecture

Strong vs. weak generative capacity (Buder-Gröndahl 2023)

- Classification of expressions (based on e.g. acceptability) concerns *weak generative capacity*: which *strings* a grammar produces
- Linguists often aim to analyze *strong generative capacity* instead: which *structural descriptions* a grammar produces (Ott 2017)

Weak vs. strong generative capacity

"Let us say that a grammar *weakly generates* a set of sentences and that it *strongly generates* a set of structural descriptions (…) Suppose that the linguistic theory T provides a class of grammars G_1, G_2, ..., where G_i weakly generates the language L_i and strongly generates the system of structural descriptions Σ_i. Then the class {L_1, L_2, ...) constitutes the *weak generative capacity* of T and the class {Σ_1, Σ_2, ...) constitutes the *strong generative capacity* of T."

(Chomsky 1965: 60)

Strong equivalence: first attempt

$\Sigma(G)$ = set of structural descriptions (SDs) generated by G

G_1 and G_2 are strongly equivalent iff $\Sigma(G_1) = \Sigma(G_2)$ (Chomsky & Miller 1963)

Problem: grammars can be strongly equivalent without having identical SDs (Kuroda 1976)

- LTAGs and HPSGs (Yoshinaga et al. 2002)
- LCFRSs and sPRCGs (Mery et al. 2006)
- TAGs and monadic linear CFTGs (Kepser & Rogers 2011)
- TAGs and CCGs (Schiffer & Maletti 2021)
- MGs and MCFGs (Stabler 2013)
- MGs and single-movement normal form MGs (Graf et al. 2016)
- (...)

Strong equivalence: second attempt

$\Sigma(G)$ = set of structural descriptions (SDs) generated by G

G_1 and G_2 are strongly equivalent iff SDs in $\Sigma(G_1)$ are isomorphic to SDs in $\Sigma(G_2)$
(Kornai & Pullum 1990)

Problem: *which relations* need to be preserved? (Miller 1999: 5)

> "it is not altogether clear how to compare the strong generative capacities of linguistically significant models (…) Although many of these models give sets of labeled phrase structure trees, there generally is information produced from these trees which differs from model to model"
>
> (Rounds et al. 1987: 351)

Plan: build on prior formal definitions of SGC

Miller (1999): generic definition of SGC for linguistic formalisms (phrase-structure, dependency)

Rogers (1998): definition of linguistic formalisms with the SGC of context-free grammar (CFG)

Both use model theory, but in different ways:

- Miller: model-theoretic meta-semantics *for* linguistic theories/grammars
- Rogers: defining linguistic theories/grammars *as* models of a logical meta-language

Aim: utilize these for LLM-interpretation

- Put the analyses together to yield an explicit generic account of SGC for CFG
- Replace relevant parts of the account with LLM-friendly notions

Miller (1999): Model-theoretic definition of SGC

Interpretation function $IF_{T \to ID}$ maps SDs in formalism T to *interpretation domain* ID

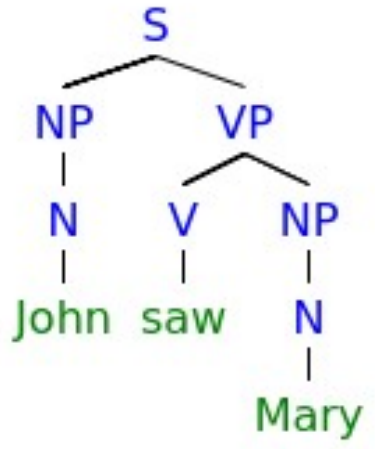
Interpretation domain for constituency (ID_C): set of *constituent structures*

- Occurrences S_O: set of pairs <v, i> where v is a vocabulary item and i is a (unique) index
- Constituent structure Γ on S_O: set of non-empty subsets of S_O

CFG:

$$\begin{aligned}
&\text{S} \rightarrow \text{NP VP}\\
&\text{NP} \rightarrow \text{N}\\
&\text{VP} \rightarrow \text{V NP}\\
&\text{N} \rightarrow \text{John} \mid \text{Mary}\\
&\text{V} \rightarrow \text{saw}
\end{aligned}$$

SD:



Constituent structure:

Γ= { {<John, 1>, <saw, 2>, <Mary, 3>},
{<saw, 2>, <Mary, 3>},
{<John, 1>},
{<saw, 2>},
{<Mary, 3>} }

Miller (1999): Model-theoretic definition of SGC

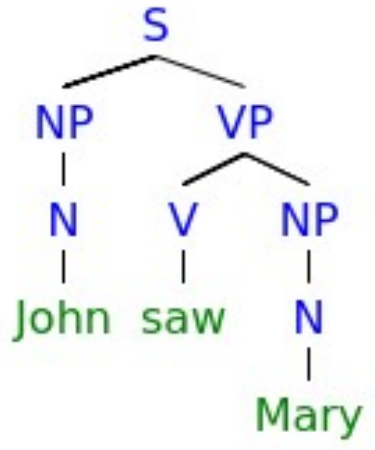
Interpretation function for constituency for CFGs ($IF_{CFG \rightarrow C}$)

- V_N: non-terminal vocabulary
- V_T: terminal vocabulary
- $IF_{CFG \rightarrow C}$: $\Sigma(CFG) \rightarrow ID_C$: $\sigma \rightarrow \Gamma$ such that for all occurrences of A in σ, $A \in V_N$, the set of occurrences of elements of V_T dominated by A belongs to Γ

CFG:

$$\begin{aligned}
&\mathrm{S} \rightarrow \mathrm{NP\ VP} \\
&\mathrm{NP} \rightarrow \mathrm{N} \\
&\mathrm{VP} \rightarrow \mathrm{V\ NP} \\
&\mathrm{N} \rightarrow \mathrm{John} \mid \mathrm{Mary} \\
&\mathrm{V} \rightarrow \mathrm{saw}
\end{aligned}$$

SD:



Constituent structure:

Γ= { {<John, 1>, <saw, 2>, <Mary, 3>},
{<saw, 2>, <Mary, 3>},
{<John, 1>},
{<saw, 2>},
{<Mary, 3>} }

Miller (1999): Model-theoretic definition of SGC

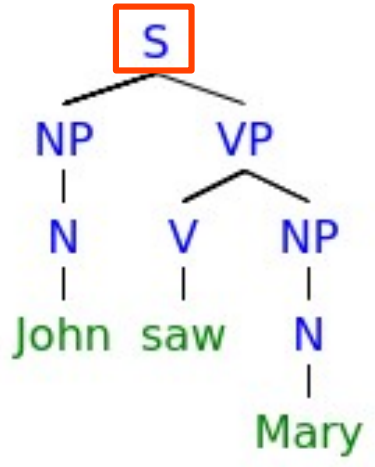
Interpretation function for constituency for CFGs ($IF_{CFG \rightarrow C}$)

- V_N: non-terminal vocabulary
- V_T: terminal vocabulary
- $IF_{CFG \rightarrow C}$: $\Sigma(CFG) \rightarrow ID_C$: $\sigma \rightarrow \Gamma$ such that for all occurrences of A in σ, $A \in V_N$, the set of occurrences of elements of V_T dominated by A belongs to Γ

CFG:

$$S \rightarrow NP\ VP$$
$$NP \rightarrow N$$
$$VP \rightarrow V\ NP$$
$$N \rightarrow John \mid Mary$$
$$V \rightarrow saw$$

SD:



Constituent structure:

Γ= {{<John, 1>, <saw, 2>, <Mary, 3>},
{<saw, 2>, <Mary, 3>},
{<John, 1>},
{<saw, 2>},
{<Mary, 3>} }

Miller (1999): Model-theoretic definition of SGC

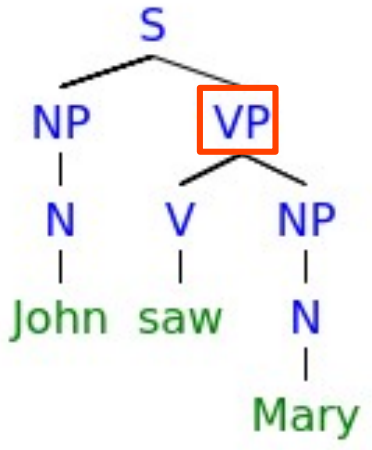
Interpretation function for constituency for CFGs ($IF_{CFG \rightarrow C}$)

- V_N: non-terminal vocabulary
- V_T: terminal vocabulary
- $IF_{CFG \rightarrow C}$: $\Sigma(CFG) \rightarrow ID_C$: $\sigma \rightarrow \Gamma$ such that for all occurrences of A in σ, $A \in V_N$, the set of occurrences of elements of V_T dominated by A belongs to Γ

CFG:

$$\begin{aligned}
&\text{S} \rightarrow \text{NP VP}\\
&\text{NP} \rightarrow \text{N}\\
&\text{VP} \rightarrow \text{V NP}\\
&\text{N} \rightarrow \text{John} \mid \text{Mary}\\
&\text{V} \rightarrow \text{saw}
\end{aligned}$$

SD:



Constituent structure:

Γ= { {<John, 1>, <saw, 2>, <Mary, 3>},
{<saw, 2>, <Mary, 3>},
{<John, 1>},
{<saw, 2>},
{<Mary, 3>} }

Miller (1999): Model-theoretic definition of SGC

SGD of grammar/theory T: subset of an Interpretation Domain that can be represented by T

Subset of ID_C that can be represented by CFGs must satisfy the following:

- S_O belongs to Γ (i.e. there is a constituent dominating all occurrences: the root non-terminal)
- If E_1 and E_2 belong to Γ, and the intersection of E_1 and E_2 is non-empty, then:
    1. E_1 is included in E_2; or
    2. E_2 is included in E_1

Can be expanded further

- Allowing syncategorematic occurrences: For all $\langle v, i \rangle \in S_O$: $\{\langle v, i \rangle\} \in \Gamma$
- Labels: relating each (non-syncategorematic) constituent to a label from V_N
- Linear order: treat indices of occurrences as ordered, disallow discontinuous constituents

Rogers (1998): Logical account of CFGs

$L^2_{K,P}$: monadic second-order metalanguage for expressing typical syntactic properties

Models of sentences of $L^2_{K,P}$ are syntax trees

Proof: ϕ is a sentence of $L^2_{k,p}$ iff

models of ϕ = trees generated by some CFG

Rogers (1998): Logical account of CFGs

$\mathbf{L^2_{K,P}}$

- individual constant symbols: ***K***, predicate symbols: ***P***
- variables: $\mathbf{X = X^0 \cup X^1}$, where: $\boldsymbol{X^0}$ range over individuals and $\boldsymbol{X^1}$ over sets of individuals
- relations: $\lhd$ (parent), $\lhd^*$ (domination), $\lhd^+$ (proper domination), $\prec$ (left-of order), $\approx$ (equality)
- logical connectives, quantifiers, and grouping symbols: $\wedge$, $\vee$, $\neg$, $\forall$, $\exists$, (,), [,]

Rogers (1998): Logical account of CFGs

Tree axioms A_1^T:

A1 $(\exists x)(\forall y)\ [x \lhd^* y]$

A2 $(\forall x,y)\ [(x \lhd^* y \wedge y \lhd^* x) \rightarrow x \approx y]$

A3 $(\forall x,y,z)\ [(x \lhd^* y \wedge y \lhd^* z) \rightarrow (x \lhd^* z)]$

A4 $(\forall x,y)\ [x \lhd y] \rightarrow (x \lhd^+ y \wedge (\forall z)\ [x \lhd^* z \wedge z \lhd^* y) \rightarrow (z \lhd^* x \vee y \lhd^* z)])]$

A5 $(\forall x,z)\ [z \lhd^+ x \rightarrow (\exists y)[y \lhd x]]$

A6 $(\forall x,z)\ [x \lhd^+ z \rightarrow (\exists y)[x \lhd y \wedge y \lhd^* z]]$

A7 $(\forall x,y)\ [x \prec y \leftrightarrow (\neg x \lhd^* y \wedge \neg y \lhd^* x) \wedge y \not\prec x]$

A8 $(\forall w,x,y,z)\ [(x \prec y \wedge x \lhd^* w \wedge y \lhd^* z) \rightarrow w \prec z]$

A9 $(\forall x,y,z)\ [(x \prec y \wedge y \prec z) \rightarrow x \prec z]$

A10 $(\forall x)\ [(\exists y)\ [x \lhd y] \rightarrow (\exists y)\ [x \lhd y \wedge (\forall z)\ [x \lhd z \rightarrow z \not\prec y]]]$

A11 $(\forall x)\ [(\exists y)\ [x \prec y] \rightarrow (\exists y)\ [x \prec y \wedge (\forall z)\ [x \prec z \rightarrow z \not\prec y]]]$

A12 $(\forall x)\ [(\exists y)\ [x \prec y] \rightarrow (\exists y)\ [x \prec y \wedge (\forall z)\ [x \prec z \rightarrow y \not\prec z]]]$

Induction axioms:

$\mathbf{A_{WF-D}}$ $(\forall X)\ [(\exists x)[X(x)] \rightarrow (\exists x)[X(x) \wedge (\forall y)\ [y \lhd^+ x \rightarrow \neg X(y)]]$

$\mathbf{A_{WF-L}}$ $(\forall X)\ [(\exists x)[X(x)] \rightarrow (\exists x)[X(x) \wedge (\forall y)\ [y \prec x \rightarrow \neg X(y)]]$

Rogers (1998): Logical account of CFGs

Tree axioms A_1^T:

A1	Every tree contains a root non-terminal
A2	Domination is anti-symmetric
A3	Domination is transitive
A4	Domination does not allow a node to fall properly between a node and its parent
A5	Every node except the root has a parent
A6	Every non-trivial path from x includes a child of x
A7	All nodes are related by domination or order (not both)
A8	Subtrees rooted at ordered nodes are ordered
A9	Order is transitive
A10	Order has a minimum: a parent has a left-most child
A11	Order is discrete in one direction (left-most element)
A12	Order is discrete in other direction (right-most element)

Induction axioms:

$\mathbf{A_{WF-D}}$	Proper domination is well-founded: there is no infinite sequence of nodes properly dominating each other
$\mathbf{A_{WF-L}}$	Left-of ordering is well-founded: there is no infinite sequence of nodes left of each other

Rogers (1998): Logical account of CFGs

Models of $L^2_{K,P}$: $\langle \mathcal{U}, \mathcal{I}, \mathcal{P}, \mathcal{D}, \mathcal{L}, \mathcal{R}_p \rangle_{p \in P}$, where:

- $\mathcal{U}$ is a non-empty domain
- $\mathcal{I}$ is a function from *K* to $\mathcal{U}$ (constants)
- $\mathcal{P}$ is a relation interpreting $\lhd$ (parent)
- $\mathcal{D}$ is a relation interpreting $\lhd^*$ (domination)
- $\mathcal{L}$ is a relation interpreting $\prec$ (left-of order)
- $\mathcal{R}_p \subseteq \mathcal{U}$ is a set interpreting *p* for each $p \in P$ (predicates)

Models of $L_{\varnothing\varnothing}$ (empty domains for $\mathcal{I}$ and $\mathcal{R}_p$): $\langle \mathcal{U}, \mathcal{P}, \mathcal{D}, \mathcal{L} \rangle$

Intended models: isomorphic to a tree domain in natural interpretation when restricted to $L_{\varnothing\varnothing}$

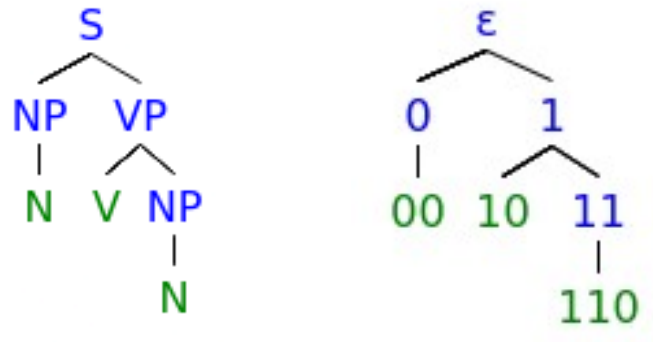
Rogers (1998): Logical account of CFGs

Tree domain (Gorn 1967): non-empty set $T \subseteq \mathbb{N}^*$ where for all $u,v \in \mathbb{N}^*$ and for all $i,j \in \mathbb{N}$:

- if $u \cdot v \in T$, then $u \in T$ ("·" denotes concatenation)
- if $u \cdot i \in T$ and $j < i$, then $u \cdot j \in T$

Natural interpretation of tree domain T: $T_{\nmid} = \langle T, \mathcal{P}_T, \mathcal{D}_T, \mathcal{L}_T \rangle$ where

- $\mathcal{P}_T = \{\langle u, u \cdot i \rangle \in T \times T \mid u \in \mathbb{N}^*, i \in \mathbb{N}\}$
- $\mathcal{D}_T = \{\langle u, u \cdot v \rangle \in T \times T \mid u,v \in \mathbb{N}^*\}$
- $\mathcal{L}_T = \{\langle u \cdot i \cdot v, u \cdot j \cdot w \rangle \in T \times T \mid u,v,w \in \mathbb{N}^*, i < j \in \mathbb{N}\}$



$T = \{\varepsilon, 0, 00, 1, 10, 11, 110\}$

$\mathcal{P}_T = \{\langle \varepsilon, 0\rangle, \langle \varepsilon, 1\rangle, \langle 0, 00\rangle, \langle 1, 10\rangle, \langle 1, 11\rangle, \langle 11, 110\rangle\}$

$\mathcal{D}_T = \{\langle \varepsilon, \varepsilon\rangle, \langle \varepsilon, 0\rangle, \langle \varepsilon, 00\rangle, \langle \varepsilon, 1\rangle, \langle \varepsilon, 10\rangle, \langle \varepsilon, 11\rangle, \langle \varepsilon, 110\rangle,$
$\langle 0, 0\rangle, \langle 0, 00\rangle, \langle 1, 1\rangle, \langle 1, 10\rangle, \langle 1, 11\rangle, \langle 1, 110\rangle,$
$\langle 00, 00\rangle, \langle 10, 10\rangle, \langle 11, 11\rangle, \langle 11, 110\rangle\}$

$\mathcal{L}_T = \{\langle 0, 1\rangle, \langle 0, 10\rangle, \langle 0, 11\rangle, \langle 0, 110\rangle,$
$\langle 10, 11\rangle, \langle 10, 110\rangle\}$

Rogers (1998): Logical account of CFGs

Tree domain (Gorn 1967): non-empty set $T \subseteq \mathbb{N}^*$ where for all $u,v \in \mathbb{N}^*$ and for all $i,j \in \mathbb{N}$:

- if $u \cdot v \in T$, then $u \in T$ ("·" denotes concatenation)
- if $u \cdot i \in T$ and $j < i$, then $u \cdot j \in T$

Natural interpretation of tree domain T: $T_{\natural} = \langle T, \mathcal{P}_T, \mathcal{D}_T, \mathcal{L}_T \rangle$ where

- $\mathcal{P}_T = \{\langle u, u \cdot i \rangle \in T \times T \mid u \in \mathbb{N}^*, i \in \mathbb{N}\}$
- $\mathcal{D}_T = \{\langle u, u \cdot v \rangle \in T \times T \mid u,v \in \mathbb{N}^*\}$
- $\mathcal{L}_T = \{\langle u \cdot i \cdot v, u \cdot j \cdot w \rangle \in T \times T \mid u,v,w \in \mathbb{N}^*, i < j \in \mathbb{N}\}$

Intended models: isomorphic to a tree domain in natural interpretation when restricted to $L_{\varnothing\varnothing}$

- $\mathbf{B_x} = \{y \mid \langle y, x \rangle \in \mathcal{D}\}$ is finite (i.e. finite path from the root to any node)
- $\mathbf{L_x} = \{y \mid (\exists z)[\langle z, y \rangle, \langle z, x \rangle \in \mathcal{P} \text{ and } \langle y, x \rangle \in \mathcal{L}\}$ is finite (i.e. finite number of left siblings)
- allow induction proofs of node depth (A_{WF-D}) and the number of left siblings (A_{WF-L})

Comparing Rogers (1998) + Miller (1999)

Rogers (1998): Definition of trees in CFG

- models of $L^2_{K,P}$ that satisfy tree axioms A1–A12 and induction axioms A_{WF-D} and A_{WF-L}
- isomorphic to a tree domain in natural interpretation when restricted to $L_{\varnothing\varnothing}$

Miller (1999): Definition of interpretation domain for trees in CFG

- $IF_{CFG \to C}$: $\Sigma(CFG) \to ID_C$: $\sigma \to \Gamma$ such that for all occurrences of A in σ, $A \in V_N$, the set of occurrences of elements of V_T dominated by A belongs to Γ

Both use model theory, but for different purposes:

- Rogers defines trees (in CFG) as models of $L^2_{K,P}$
- Miller gives model-theoretic semantics for trees

LLM-states

LLM ($\mathcal{M}$)

- Input: sequence of words(/tokens): $w_1, \ldots, w_n$
- Produces *contextual encoding* + *positional encoding* for each word: $\mathbf{h_1}, \ldots, \mathbf{h_n}$
- Each $\mathbf{h_i}$ consists of activation values for m nodes: $\langle a^i_1, \ldots, a^i_m \rangle$ (concatenated across layers)
- *Complete state* of $\mathcal{M}$ for input I = $\langle w_1, \ldots, w_n \rangle$:

  $S^{\mathcal{M}}_I = \langle a^1_1, \ldots, a^n_m \rangle$ (i.e. concatenation of all contextual encodings)
- Equivalently, $S^{\mathcal{M}}_I$ can be expressed as a set of triplets <i, j, a>, where i indicates the input token position, j indicates the $\mathcal{M}$-node, and a indicates its activation value
- All *states* of $\mathcal{M}$ for input I: the power set of $S^{\mathcal{M}}_I$ (i.e. all subsets of node activations)

LLM-states

Plausible(?) assumptions about modeling phrase-structure in $\mathcal{M}$:

1. Nodes in phrase-structures are interpreted as model states
2. Relations between nodes are interpreted as relations between interpretations of nodes

Purpose of assumptions: preventing trivial mappings (Buder-Gröndahl 2023)

- Any sufficiently complex system can be mapped to finite abstract structures (Searle 1992)
- Anything can be mapped to finite-storage computation (Putnam 1988, Sprevak 2018)

LLMs as models of $L_{\varnothing\varnothing}$

Models of $L_{\varnothing\varnothing}$: $\langle \mathcal{U}, \mathcal{P}, \mathcal{D}, \mathcal{L} \rangle$ (Rogers 1998)

- $\mathcal{U}$ is a non-empty domain
- $\mathcal{P}$ is a relation interpreting $\lhd$ (parent)
- $\mathcal{D}$ is a relation interpreting $\lhd^*$ (domination)
- $\mathcal{L}$ is a relation interpreting $\prec$ (left-of order)

$\mathcal{M}$ as a model of $L_{\varnothing\varnothing}$:

- $\mathcal{U}^{\mathcal{M}}$: $\mathcal{M}$-states that enter into at least one of relations $\{\mathcal{P}^{\mathcal{M}}, \mathcal{D}^{\mathcal{M}}, \mathcal{L}^{\mathcal{M}}\}$
- $\mathcal{P}^{\mathcal{M}}$, $\mathcal{D}^{\mathcal{M}}$, $\mathcal{L}^{\mathcal{M}}$: relations between $\mathcal{M}$-states that satisfy axioms A1–A12 and A_{WF-D} and A_{WF-L}

Interpretation domain for constituency in LLMs

Interpretation domain for constituency (ID_C): set of constituent structures (Miller 1999)

- Occurrences S_O: set of pairs <v, i> where v is a vocabulary item and i is a (unique) index
- Constituent structure Γ on S_O: set of non-empty subsets of S_O

Interpretation function for constituency for CFGs ($IF_{CFG \rightarrow C}$) (Miller 1999)

- $IF_{CFG \rightarrow C}$: $\Sigma(CFG) \rightarrow ID_C$: $\sigma \rightarrow \Gamma$ such that for all occurrences of A in σ, $A \in V_N$, the set of occurrences of elements of V_T dominated by A belongs to Γ

→ Change occurrences of non-terminal or terminal nodes to their $\mathcal{M}$-state interpretations

Interpretation domain for constituency in LLMs

Interpretation domain for constituency in $\mathcal{M}$ ($ID^{\mathcal{M}}_{C}$): set of constituent structures on *terminals*

- $\mathcal{T}^{\mathcal{M}}$ = {X | X $\in$ $\mathcal{U}^{\mathcal{M}}$ and there is no Y such that <X, Y> $\in \mathcal{D}^{\mathcal{M}}$}
- Constituent structure $\Gamma^{\mathcal{M}}$ on $\mathcal{T}^{\mathcal{M}}$: set of non-empty subsets of $\mathcal{T}^{\mathcal{M}}$

Interpretation function for constituency for $\mathcal{M}$-states ($IF_{\mathcal{M}\rightarrow C}$)

- $IF_{\mathcal{M}\rightarrow C}$: $\mathcal{U}^{\mathcal{M}} \rightarrow ID^{\mathcal{M}}_{C}$: X $\rightarrow$ $\Gamma^{\mathcal{M}}$ such that for all Y $\in$ {Z | Z $\notin \mathcal{T}^{\mathcal{M}}$ and <X, Z> $\in \mathcal{D}^{\mathcal{M}}$}, the set {W | W $\in \mathcal{T}^{\mathcal{M}}$ and <Y, W> $\in \mathcal{D}^{\mathcal{M}}$} belongs to $\Gamma^{\mathcal{M}}$

Interpretation domain for constituency in LLMs

Interpretation domain for constituency in $\mathcal{M}$ ($ID^{\mathcal{M}}_{C}$): set of constituent structures on *terminals*

- $\mathcal{T}^{\mathcal{M}}$ = {X | X $\in$ $\mathcal{U}^{\mathcal{M}}$ and there is no Y such that <X, Y> $\in \mathcal{D}^{\mathcal{M}}$}
- Constituent structure $\Gamma^{\mathcal{M}}$ on $\mathcal{T}^{\mathcal{M}}$: set of non-empty subsets of $\mathcal{T}^{\mathcal{M}}$

Interpretation function for constituency for $\mathcal{M}$-states ($IF_{\mathcal{M}\rightarrow C}$)

- $IF_{\mathcal{M}\rightarrow C}$: $\mathcal{U}^{\mathcal{M}} \rightarrow ID^{\mathcal{M}}_{C}$: X $\rightarrow$ $\Gamma^{\mathcal{M}}$ such that for all Y $\in$ {Z | Z $\notin \mathcal{T}^{\mathcal{M}}$ and <X, Z> $\in \mathcal{D}^{\mathcal{M}}$}, the set {W | W $\in \mathcal{T}^{\mathcal{M}}$ and <Y, W> $\in \mathcal{D}^{\mathcal{M}}$} belongs to $\Gamma^{\mathcal{M}}$

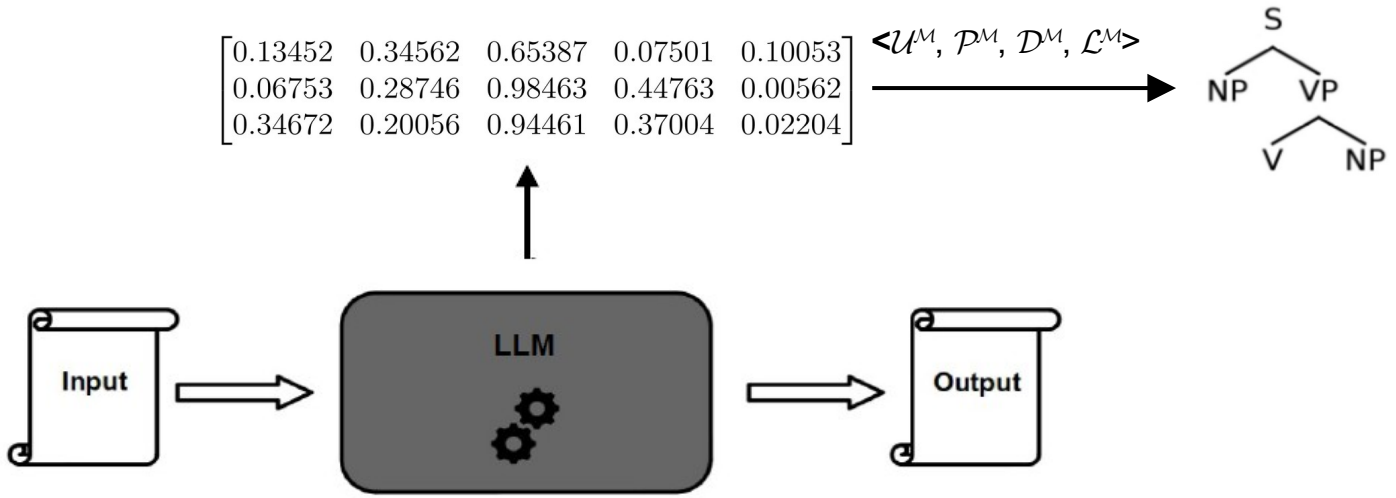
"for all non-terminal nodes Y dominated by X"

"the set of all terminal nodes dominated by Y"

Generic account of LLM-interpretation (for CFG)

Interpretation of $\mathcal{M}$: specification of $<\mathcal{U}^{\mathcal{M}}, \mathcal{P}^{\mathcal{M}}, \mathcal{D}^{\mathcal{M}}, \mathcal{L}^{\mathcal{M}}>$ + resulting constituents via $IF_{\mathcal{M}\to C}$

- $<\mathcal{U}^{\mathcal{M}}, \mathcal{P}^{\mathcal{M}}, \mathcal{D}^{\mathcal{M}}, \mathcal{L}^{\mathcal{M}}>$ must satisfy tree axioms A1–A12 and induction axioms A_{WF-D} and A_{WF-L}
- Equivalently, $<\mathcal{U}^{\mathcal{M}}, \mathcal{P}^{\mathcal{M}}, \mathcal{D}^{\mathcal{M}}, \mathcal{L}^{\mathcal{M}}>$ must be isomorphic to a tree domain in natural interpretation
- SGC of $\mathcal{M}$: subset of $ID^{\mathcal{M}}_{C}$ that function as constituents via $IF_{\mathcal{M}\to C}$



Links to empirical work on LLM-interpretation

Behavioral methods:

- Directly relevance only for WGC; but possible indirect relevance for SGC
- Grouping tokens on some behavioral basis, comparing these to Miller-type constituents

Probing: mapping $\mathcal{M}$-states to linguistic formalisms (Belinkov 2022)

- Theoretical problem: trivially available mappings to all kinds of SDs (Buder-Gröndahl 2023)
- Possible alleviation for constituency: need to satisfy axioms A1–A12 and A_{WF-D} and A_{WF-L}

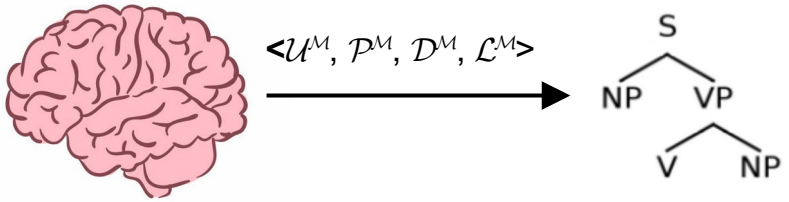
Mechanistic interpretation:

- Identifying $\mathcal{M}$-internal cirquits for syntactic tasks → restricting candidates for $\mathcal{U}^{\mathcal{M}}$
- Initial challenge: methods like activation patching are based on altering input (Meng et al. 2022); how to obtain information on different possible interpretations on the *same* input?

Broader relevance for cognitive science

How could there be “trees in the head”?

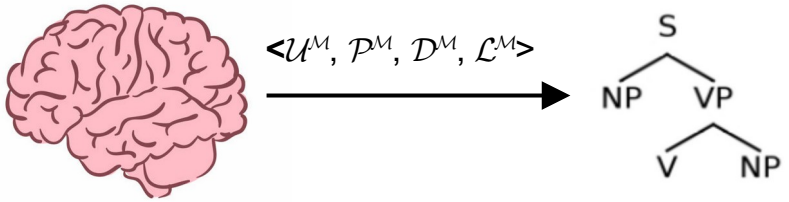
- Extend $\mathcal{M}$ to be a placeholder of any concrete computational system (e.g. a human brain)
- $\mathcal{M}$ functions as an intended model $L_{\varnothing\varnothing}$ iff there is a specification of $\langle \mathcal{U}^{\mathcal{M}}, \mathcal{P}^{\mathcal{M}}, \mathcal{D}^{\mathcal{M}}, \mathcal{L}^{\mathcal{M}} \rangle$ that satisfies A1–A12 and $A_{WF-D} - A_{WF-L}$; i.e. is isomorphic to a tree domain in natural interpretation
- SGC of $\mathcal{M}$: subset of $ID^{\mathcal{M}}_{C}$ that function as constituents via $IF_{\mathcal{M}\rightarrow C}$



Broader relevance for cognitive science

How could there be “trees in the head”?

- Extend $\mathcal{M}$ to be a placeholder of any concrete computational system (e.g. a human brain)
- $\mathcal{M}$ functions as an intended model $L_{\varnothing\varnothing}$ iff there is a specification of $\langle \mathcal{U}^{\mathcal{M}}, \mathcal{P}^{\mathcal{M}}, \mathcal{D}^{\mathcal{M}}, \mathcal{L}^{\mathcal{M}} \rangle$ that satisfies A1–A12 and A_{WF-D} – A_{WF-L}; i.e. is isomorphic to a tree domain in natural interpretation
- SGC of $\mathcal{M}$: subset of $ID^{\mathcal{M}}_{C}$ that function as constituents via $IF_{\mathcal{M}\rightarrow C}$



Thank you!

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations.

Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics* 48: 207–219.

Buder-Gröndahl, T. (2023). The ambiguity of BERTology: what do large language models represent? Synthese 203: 15.

Chang, T.A. & Bergen, B.K. (2024). Language Model Behavior: A Comprehensive Survey. *Computational Linguistics* 50: 293–350.

Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge: MIT Press.

Chomsky, N. and Miller, G.A. (1963). Introduction to the formal analysis of natural languages. In Handbook of Mathematical Psychology, vol. II, edited by R.D. Luce, R.R. Bush, and E. Galanter: 269-322. New York: Wiley.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. *Advances in Neural Information Processing Systems* 36.

Dupre, G. (2021). (What) Can Deep Learning Contribute to Theoretical Linguistics? Minds and Machines 31 (4):617-635.

Gorn, S. (1967). Explicit Definitions and Linguistic Dominoes. In Hart, J.F. & Takasu, S. (eds.). Systems and Computer Science, Proceedings of the Conference held at Univ. of Western Ontario, 1965. University of Toronto Press.

Graf, T., Aksënova, A., De Santo, A. (2016). A Single Movement Normal Form for Minimalist Grammars. In: Foret, A., Morrill, G., Muskens, R., Osswald, R., Pogodalla, S. (eds) Formal Grammar. FG FG 2015 2016. Lecture Notes in Computer Science vol 9804. Berlin, Heidelberg: Springer.

Kepser, S. & Rogers, J. (2011). The equivalence of tree adjoining grammars and monadic linear context-free tree grammars. Journal of Logic, Language and Information, 20 (3) :361–384.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Proceedings of the 36th International Conference on Neural Information Processing Systems*: 22199 –22213.

Kornai, A. and Pullum, G. K. (1990). The X-bar Theory of Phrase Structure. Language, 66: 24–50.

Linzen, T. & Baroni, M. (2021). Syntactic Structure from Deep Learning. *Annual Review of Linuistics* 7: 195–212.

Mahowald, K. (2023). A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*: 265–273.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. ArXiv preprint ArXiv:2202.05262.

References

Mery, B., Amblard, M., Durand, I., Retoré, C. (2006). A Case Study of the Convergence of Mildly Context-sensitive Formalisms for Natural Language Syntax: from Minimalist Grammars to Multiple Context-Free Grammars. [Research Report] RR-6042, INRIA: 67.

Miller, P.H. (1999). Strong Generative Capacity. The Semantics of Linguistic Formalism. Stanford: CSLI publications.

Miller, P.H. & Pullum, G. (2001). Review of "A descriptive approach to language-theoretic complexity" by James Rogers. Computational Linguistics 27(2): 304–308.

Ott, D. (2017). Strong generative capacity and the empirical base of linguistic theory. Frontiers in Psychology, 8.

Putnam, H. (1988). *Representation and reality*. MIT Press.

Rogers, J. (1998). A Descriptive Approach to Language-Theoretic Complexity. Stanford: CSLI Publications.

Rounds, W.C., Manaster-Ramer, A. & Friedman, J. (1987). Finding natural languages a home in formal language theory. In A. Manaster-Ramer (ed.). Mathematics of Language: 349- 359. Amsterdam: John Benjamins.

Saparov, A. & He, H. (2023). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *International Conference on Learning Representations*.

Searle, J. (1992). The Rediscovery of the Mind. Cambridge: MIT Press.

Sprevak, M. (2018). Triviality arguments about computational implementation. M. Sprevak & M. Colombo (Eds.), Routledge handbook of the computational mind (pp. 175–191). London: Routledge.

Schiffer, L.K., Maletti, A. (2021). Strong Equivalence of TAG and CCG. Transactions of the Association for Computational Linguistics 9: 707–720.

Stabler, E. (2013). Two Models of Minimalist, Incremental Syntactic Analysis. Topics in Cognitive Science 5(3): 611–633.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., . . . Polosukhins, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing*: 6000–6010.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D.. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*: 24824–24837.

Wu, Z., Geiger, A., Icard, T., Potts, C., & Goodman, N. (2023). Interpretability at scale: identifying causal mechanisms in Alpaca. *Proceedings of the 37th International Conference on Neural Information Processing Systems*: 78205–78226.

Yoshinaga, N., Miyao, Y., & Tsujii, J. (2002). A Formal Proof of Strong Equivalence for a Grammar Conversion from LTAG to HPSG-style. Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks: 187–192.