

言語モデルは言語学に 何のエビデンスも与えないのでしょうか？

MBZUAI 栗林樹生

Evidence-based Linguistics Workshop 2025（招待講演）@国語研

自己紹介

- 栗林 樹生（くりばやし たつき）
 - Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
 - 2年強ポスドクをしていました
 - 人工知能領域で強い、若い大学
 - 情報科学（博士）東北大学大学院情報科学研究科



- 今年8月よりAssistant professor (PI)
 - MsC/PhD（来年夏入学）募集中
 - 学部プログラムもできました (<https://mbzuai.ac.ae/study/undergraduate-program/>)
 - AI特化型の大学で、他大学では類を見ないほど詳細かつ最先端の講義も満載（ML securityで1講義など、私もCogsci×NLPの講義予定）なので、修士で基礎固めに学びに来るとするのもアリだと思います
 - 講義演習→論文化などのケースも見ます
 - ビジター（1ヶ月程度から）も募集中
 - ポスドク（1名）も探し中
 - 今日の話のような言語科学との学際的領域をすすめていきます

Tatsuki Kuribayashi

Assistant Professor of Natural Language Processing

Research Interests

Professor Kuribayashi's research focuses on unique interdisciplinary topics bridging NLP to the science of human language. This involves the exploration of the cognitive plausibility of NLP models as well as the role of modern NLP in understanding language acquisition, processing, and communication. This spans multiple areas beyond NLP, such as computational psycholinguistics, linguistic typology, and information theory.



#	Institution	Count	Faculty
1	▶ Peking University 🇨🇳	157.9	124
2	▶ Carnegie Mellon University 🇺🇸	157.8	92
3	▶ Tsinghua University 🇨🇳	129.3	116
4	▶ Stanford University 🇺🇸	116.6	50
5	▶ KAIST 🇰🇷	106.1	68
6	▶ Univ. of California - Berkeley 🇺🇸	104.5	63
7	▶ Shanghai Jiao Tong University 🇨🇳	103.0	111
8	▶ Chinese Academy of Sciences 🇨🇳	100.1	63
9	▶ Univ. of California - San Diego 🇺🇸	95.7	69
10	▶ Nanyang Technological University 🇸🇬	94.9	50
11	▶ MBZUAI 🇦🇪	91.6	55
12	▶ Massachusetts Institute of Technology 🇺🇸	91.5	73
13	▶ University of Maryland - College Park 🇺🇸	86.9	54
14	▶ ETH Zurich 🇨🇭	86.7	31
15	▶ Univ. of Illinois at Urbana-Champaign 🇺🇸	86.0	62
16	▶ Zhejiang University 🇨🇳	81.1	83
17	▶ HKUST 🇮🇪	75.5	53
18	▶ Johns Hopkins University 🇺🇸	73.9	48
19	▶ Seoul National University 🇰🇷	72.7	66
20	▶ National University of Singapore 🇸🇬	72.1	48
21	▶ University of Washington 🇺🇸	69.8	45
22	▶ Harbin Institute of Technology 🇨🇳	68.4	83

CSRanking (AI/CV/NLP トップ会議論文数)

- 出身は工学（情報科学）
 - 今思い返せば、人工「知能」の方に関心があって、認知科学みたいな分野を期待していたかもしれない
 - 今思い返せば、大学入試では人間工学分野も視野に入れていた（人間にとって負荷の少ない文ではなく、腰に負担の少ない椅子とか開発してたかもしれない）
- 研究はしばしば計算（心理）言語学領域
 - 言語系であれば東大の大関先生としばしば協働
 - 基本的にNLPの学会に論文を出しており、NLP研究者として言語に接近を試みる方向（NLP-輸出->言語学）
 - NLP国際会議は、共著含め、かれこれ40本ぐらい出してきたので、土地勘だけはある
 - Cognitive Modeling and Computational Linguistics (CMCL) 国際ワークショップ運営（2024-）
 - 国語研にスポンサーとして支えていただいております、改めて感謝申し上げます

- 人類の歴史の中で、人間の言語を流暢に話す人間以外のもの（言語モデル）を初めて目にしている
- 人間の言語に何を示唆していると思えば良い？

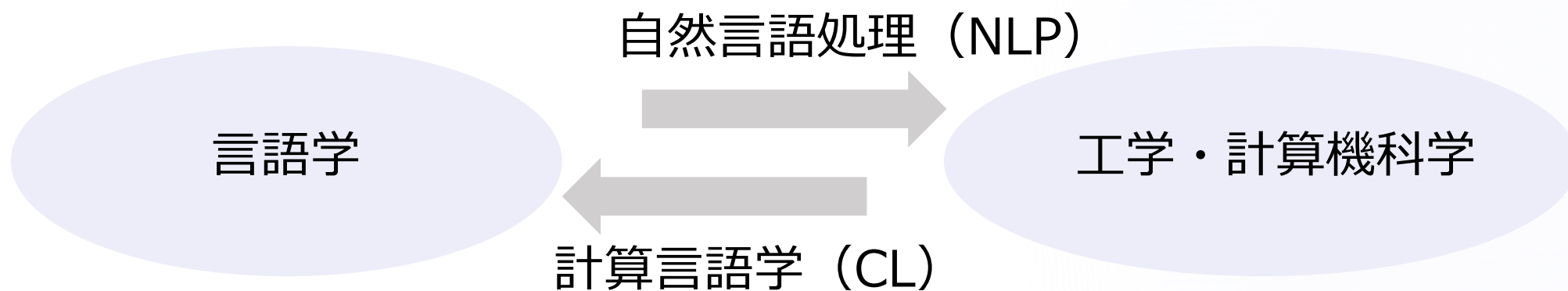
スコープ外（目的感はある程度自明）

- NLPモデルに言語的な特徴量・性質を与え、特定タスクの性能向上を狙う
- 古典的な基礎タスク（構文・意味解析など）を解くことを目的にする
- 言語研究用に便利自動化ツールを作る
- 言語学的な視点から言語モデル評価データを作成する
- 何か新しい言語学的な概念を持ってきて、それを言語モデルが知っているか単に分析する（プロービング）

自然言語処理分野について

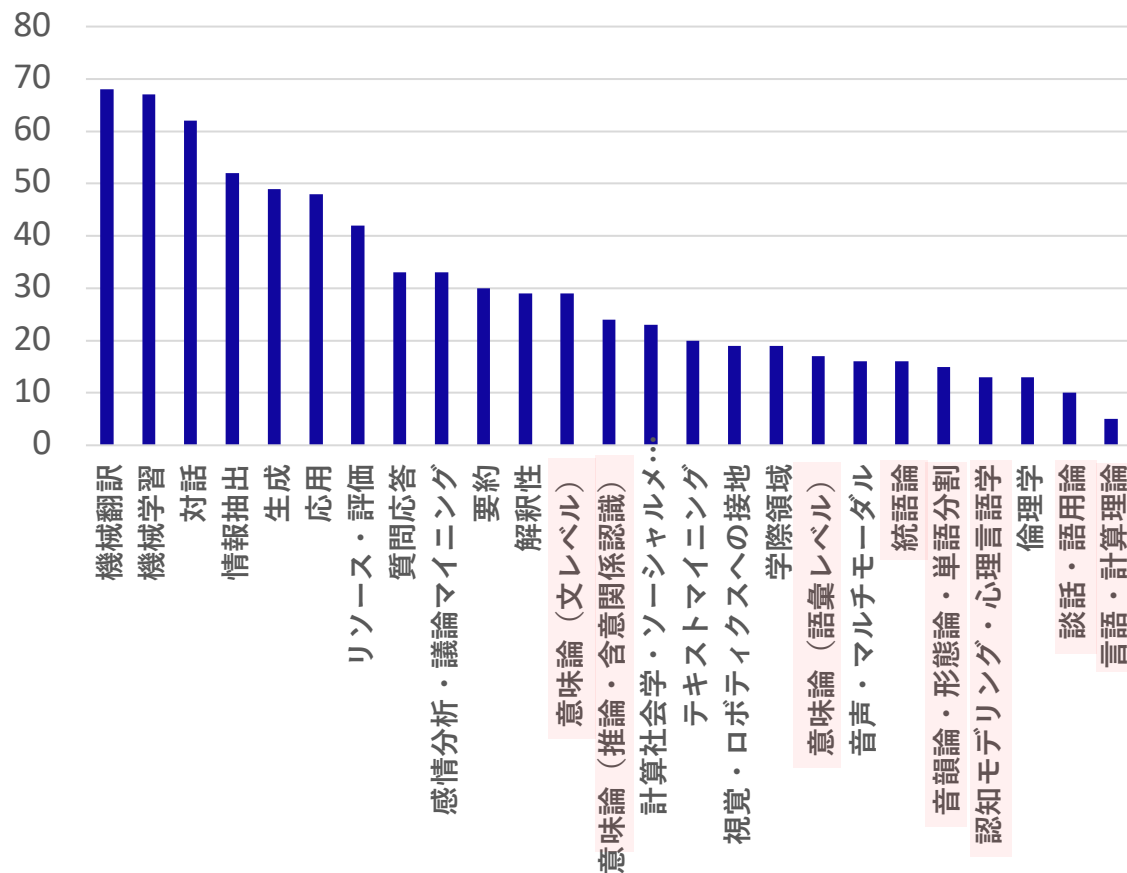
自然言語処理分野

- 自然言語処理（NLP）：
計算機に人間の言葉进行处理させる分野
 - 機械翻訳、対話、要約…
- 計算言語学（CL）：「自然言語処理」のセクシーな言い方
言語的な問いに対して計算・数理的モデリングで迫る分野



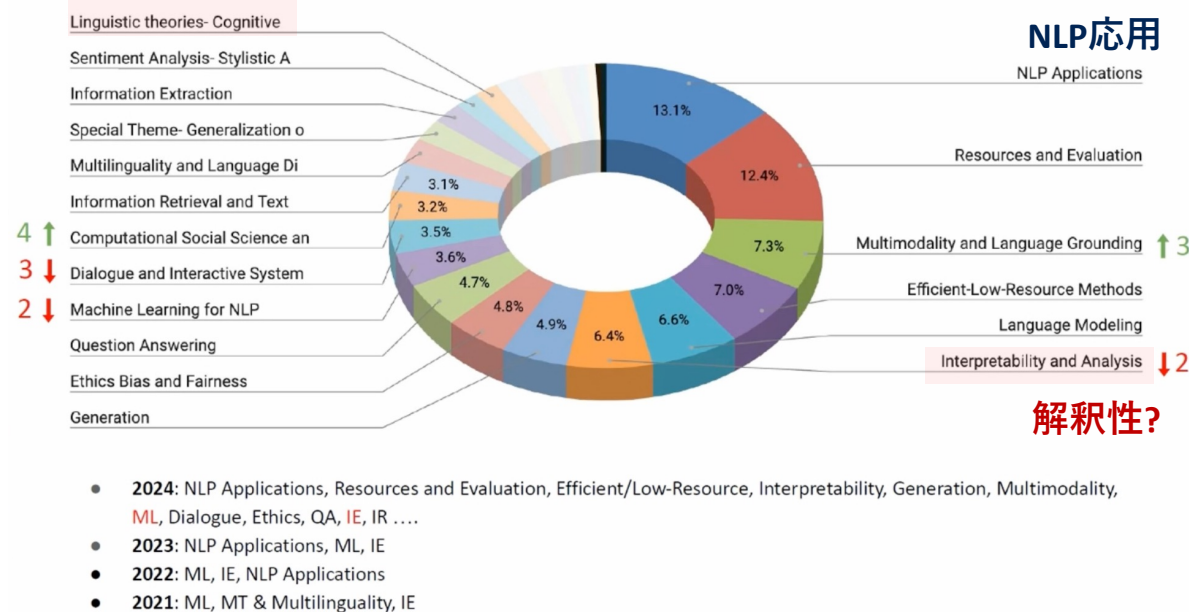
自然言語処理分野の動向

サブ領域別論文数@ACL 2020



サブ領域別論文数@ACL 2025

言語理論・認知モデリング



ここ5年で、少なくとも、
形態論・統語論・意味論のような
言語学に根ざした分野の括りがほぼ消滅

一方、分野全体としては言語科学を楽しんでいる

- 受賞論文 in NLPトップ国際会議

- **Mission: Impossible Language Models** (ACL 2024 best paper)
Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, Christopher Potts
- **Semisupervised Neural Proto-Language Reconstruction** (ACL 2024 best paper)
Liang Lu, Peirong Xie, David R Mortensen
- **Visual Grounding Helps Learn Word Meanings in Low-Data Regimes** (NAACL 2024 best paper)
Chengxu Zhuang, Evelina Fedorenko, Jacob Andreas
- **Between Circuits and Chomsky: Pre-pretraining on Formal Languages Imparts Linguistic Biases** (ACL 2025 outstanding paper)
Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, Tal Linzen
- **A New Formulation of Zipf's Meaning-Frequency Law through Contextual Diversity** (ACL 2025 outstanding paper)
Ryo Nagata, Kumiko Tanaka-Ishii
- **Using Information Theory to Characterize Prosodic Typology: The Case of Tone, Pitch-Accent and Stress-Accent** (ACL 2025 SAC highlights)
Ethan Wilcox, Cui Ding, Giovanni Acampa, Tiago Pimentel, Alex Warstadt, Tamar I Regev

言語の話題が相対的に減っているのは、この大規模言語モデル時代にそういう論文を書く人が相対的に少ないためだと思われる（聴衆が減ったわけではない）

自然言語処理分野の課題？

Cf. 人工知能分野全体のNLP (LLM) 化 → NLP分野のアイデンティティは？

- **言語学・人文学系の人も、更に自然言語処理分野をかき回してくれると嬉しいなあ（願望）**
 - 自然言語処理 - 言語 = アイデンティティ??
 - おそらく研究のネタはたくさん眠っている
 - 分野自体は学際的な話題に寛容
- **なぜか日本のNLP研究者（特に一部の若手）は異様に言語愛がある**
 - 言語学系の研究者（特に若手）がNLP学会で寂しくならないはず…！
 - 日本：NLPから計算言語学へシフトする人がいる
 - 日本外：言語学・認知科学から計算言語学へシフトする人が多い？

準備：大規模言語モデル

- コーパス上で次の単語をうまく当てられるようにニューラルネットを学習
 - そういう方法で言語データで訓練されたニューラルネットワークを、ニューラル言語モデル（言語モデル）と呼ぶ
 - ニューラルネットワーク自体は画像でも音声でも汎用に使われる道具立て
- 簡単なイメージ
 - コーパスをぶつ切りにして最後の単語を消し、次の単語何でしょう問題を作りまくる
 - 言語モデルは消された単語を当てるように訓練される（先読みのモデル）
 - コーパス全体を通して次の単語の分布（1単語全賭けではなく）を考えることになる

おめでとう



言語モデル



あけまして

ございます



言語モデル



あけまして おめでとう

！



言語モデル



あけまして おめでとう

自然言語処理分野における言語学との接点

自分のきっかけ

- 英語言語において、言語モデルの文法能力の分析が盛んに行われる（2018--）
 - 文法的にありえない単語に低い確率を付与するか？
 - *A **sketch** of lights **don't**...
- 日本語特有の現象ってなんかないかな… 比較的語順が自由なので、語順選好とか面白そう
- 大規模コーパスに基づく日本語二重目的語構文の基本語順の分析
(Ryohei Sasano and Manabu Okumura. A Corpus-Based Analysis of Canonical Word Order of Japanese Double Object Constructions. ACL2016)
 - 構文解析器で100億文を解析し、様々な仮説を定量的に検証（動詞によって基本語順は異なる、省略されにくい格は動詞の近くに、Pass/showタイプ、二格名詞が着点を表す場合…）
- コーパス頻度を言語モデルが計算する確率に置き換えたらどうだろう？
(Tatsuki Kuribayashi, Takumi Ito, Jun Suzuki, Kentaro Inui. "Language Models as an Alternative Evaluator of Word Order Hypotheses: A Case Study in Japanese." ACL 2020)

言語モデルで計算してみる

- 言語モデルは文の確率を計算できる

彼を鳥に例えた



確率("彼を鳥に例えた")=0.0001

※値は適当

?鳥に彼を例えた

確率("鳥に彼を例えた")=0.00000001

- コーパス頻度を言語モデルの出力確率で置き換えて語順分析をしても、既存の知見を再現
 - 言語モデルは語順の統計を極めてうまく捉えている
 - 一旦言語モデルを訓練してしまえば、構文解析器などを使うよりも、ある意味楽に、かつやわらかくコーパス統計情報（頻度）にアクセスできる
- ※当時はChatGPTのように聞いて答えているわけではなく、文の確率を直接計算している
 - どのようにメタ言語知識を問い合わせるとよいかは様々分析あり
 - Prompting is not a substitute for probability measurements in large language models (Hu & Levy, EMNLP 2023)
 - How to Make the Most of LLMs' Grammatical Knowledge for Acceptability Judgments (Ide et al., NAACL 2025)

言語モデルで計算してみる

- 言語モデルが人間の語順選好を捉えている -> それすなわち？
 - 語順選好はコーパスをたくさん読めば獲得できるということ？
 - （ただし先ほどの検証では、コーパスの頻度と人間にとっての文の自然さを同一視して議論している）
- 実際はもっと非自明な事が起きている
 - A **sketch** of lights {**doesn't**/*don't} appear
 - “NP PP NP VP”のような文をコーパスから全て消去して、言語モデルを訓練しても、正しい数の一致に高い確率を割り当てる
(Patil+,24. Filtered Corpus Training (FiCT) Shows that Language Models can Generalize from Indirect Evidence)
- 言語モデル（群）が何らかの言語現象Yを再現できる
 - これは一体、言語に対して何を示唆しているのか？（今日のテーマ）

- **言語学的な学術的問いに対して、言語モデルを通して、どんなエビデンスを提供できるか**
 - 特に「言語モデルが言語のYYYを知っている」「言語モデル集団が何らかの現象（文法化とか）を再現」と言ったときに、言語に対して何を言っていることになるのか？
 - （どうやったら言語モデルがYを知っている・獲得していると主張できるかはスコープ外）
- 人間の言語を流暢に扱える人間以外の何か（言語モデル）が、人類史上初めて存在するので、それが言語科学に対して何を示唆するか考えたい
 - 何も示唆しないかもしれないことが、研究を始めない理由にはならない
- この問い自体がホットトピック
 - Futrell, Richard, and Kyle Mahowald, 2025. "How linguistics learned to stop worrying and love the language models."
 - Alex Warstadt, and Samuel R Bowman, 2022. "What artificial neural networks can tell us about human language acquisition"
 - Marco Baroni, 2021. "On the proper role of linguistically-oriented deep net analysis in linguistic theorizing"

言語モデルは人間と異なるので、人間の言語処理と関係ない？

- 🙄 言語モデルと人間は違う！
 - そのとおり。言語モデルがこういう現象を示した->つまり人間では？
- 複雑な対象のうち、関心のない部分を捨象して、目的に合わせた単純化をするのが数理モデル
 - 気象予報モデルが実際にコンピュータ内でその気温を再現していないことは、そのモデルが役に立たないことを意味しない
 - 高校物理は空気抵抗を無視しているので、なんの意味もない！？
- 言語モデルを人間・言語のモデルだとしたときに、なにを削ぎ落としている？
 - 削ぎ落ちていない部分については、適切なモデルとして活用できるかもしれない
- 鈴木陽登, 菅原朔. 言語研究における科学的理解と言語モデル (NLP2025)
- 坪井祥吾, 菅原朔. 言語モデルのふるまいと多重実現 (NLP2025)

ちなみに：楽観的な見方の一例

- 解こうとしている問題が難しいほど、その解法のバリエーションは少ないかもしれない

- 同時に満たすべき要件が多いほど、成功の仕方は限定され、失敗のバリエーションは多くなる

- $A \wedge B \wedge C \wedge D \wedge E \dots$

- どれかひとつでもFalseであれば、全体がFalseになる

- アンナ・カレー二ナの法則「すべての幸せな家庭は似ている。不幸な家庭は、それぞれ異なる理由で不幸である。」

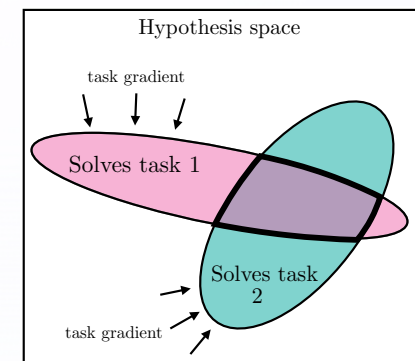


Figure 6. **The Multitask Scaling Hypothesis:** Models trained with an increasing number of tasks are subjected to pressure to learn a representation that can solve all the tasks.

Huh, M., Cheung, B., Wang, T., Isola, P. (2024). Position: The Platonic Representation Hypothesis. ICML2024

- 人間の言語活動の再現が難しい課題であるとするならば、それを唯一再現できている人間と言語モデル同士はそう遠くないかもしれない（！？？）

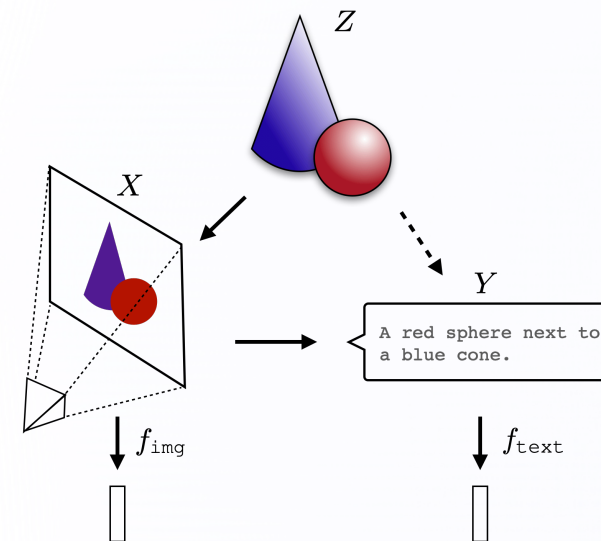
- Rosa Cao and Daniel Yamins, 2021. "Explanatory Models in Neuroscience: Part 2 -- Constraint-Based Intelligibility."

ちなみに：楽観的な見方の一例

- Platonic Representation hypothesis:
モデルの種類やモダリティが異なっても、最近の優秀な大規模言語モデルの内部表現同士は似てきている
 - 同じもの（真の世界の事象の同時分布）を違うスクリーンを通して見ているに過ぎず、被写体が同じなのだから、うまく近似できれば、いかなるスクリーンでもやがて真のモデルに収束するだろう
- 人間と言語モデルの場合はどうなのか？？

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.



言語獲得・創発編

基本方針：構成論的なアプローチ

- ある現象が再現するための**十分条件**を考える
- どういう条件で、その現象が生じるかシミュレーションを通して理解する
 - 地球上でどうやって有機物が発生したか？フラスコ内で原始地球の環境（水, CO₂, N₂…）を再現し、色々と条件を変えて、有機物が発生する条件を探る
 - （実現方法は複数ある可能性があるので、必要条件については言えない）
 - （参考：http://masa.o.oo7.jp/constructive_approach.html）
- 言語の場合は：
 - 学習モデルが母語を獲得できる条件は何か？（生成文法・認知科学的な問い）
 - なんで猫に話し続けても話し出さないのに、言語モデルの場合話し出すのか？
 - その現象はなんらかの必然性をもって生じているのか？
 - 言語横断的に見られる普遍性はどのように再現されるか？

基本方針：構成論的なアプローチ

- （何らかの調査を経て）「言語モデルはYを獲得できていた」とする

言語モデルの学習設定

すなわち

- アーキテクチャAを用いた、B層・パラメータ数Cの言語モデル（学習者）を、D法で初期化し、学習データEと目的関数Fのもと、最適化アルゴリズムGを用いて、Hステップ学習し…のもとで、学習モデルはYを獲得できる。
 - これは真
- 問題点：条件が冗長・Yについてなにが分かったのかよくわからない
 - ただ一例を見たに過ぎない
- 問い：いかにして条件を簡潔に一般化できるか、人間と対応付けられるか、何らかの方法で解釈するか。

方向性①：色々試す（成功例からのトップダウン探索）

- アーキテクチャAを用いた、B層・パラメータ数Cの言語モデル（学習者）を、D法で初期化し、学習データEと目的関数Fのもと、最適化アルゴリズムGを用いて、Hステップ学習し…その結果、言語知識Yは獲得できる。

アーキテクチャA'でもA''でもA'''…でもYの獲得は可能。

- B層・パラメータ数Cの言語モデル（学習者）を、D法で初期化し、学習データEと目的関数Fのもと、最適化アルゴリズムGを用いて、Hステップ学習し…その結果、言語知識Yは獲得できる。

層の数はどうだろう？

もっとゆるい条件

方向性①：色々試す（成功例からのトップダウン探索）

- アーキテクチャAを用いた、B層・パラメータ数Cの言語モデル（学習者）を、D法で初期化し、学習データEと目的関数Fのもと、最適化アルゴリズムGを用いて、Hステップ学習し…その結果、言語知識Yは獲得できる。

アーキテクチャA'でもA''でもA'''…でもYの獲得は可能。

- B層・パラメータ数Cの言語モデル（学習者）を、D法で初期化し、学習データEと目的関数Fのもと、最適化アルゴリズムGを用いて、Hステップ学習し…その結果、言語知識Yは獲得できる。

層の数はどうだろう？

もっとゆるい条件

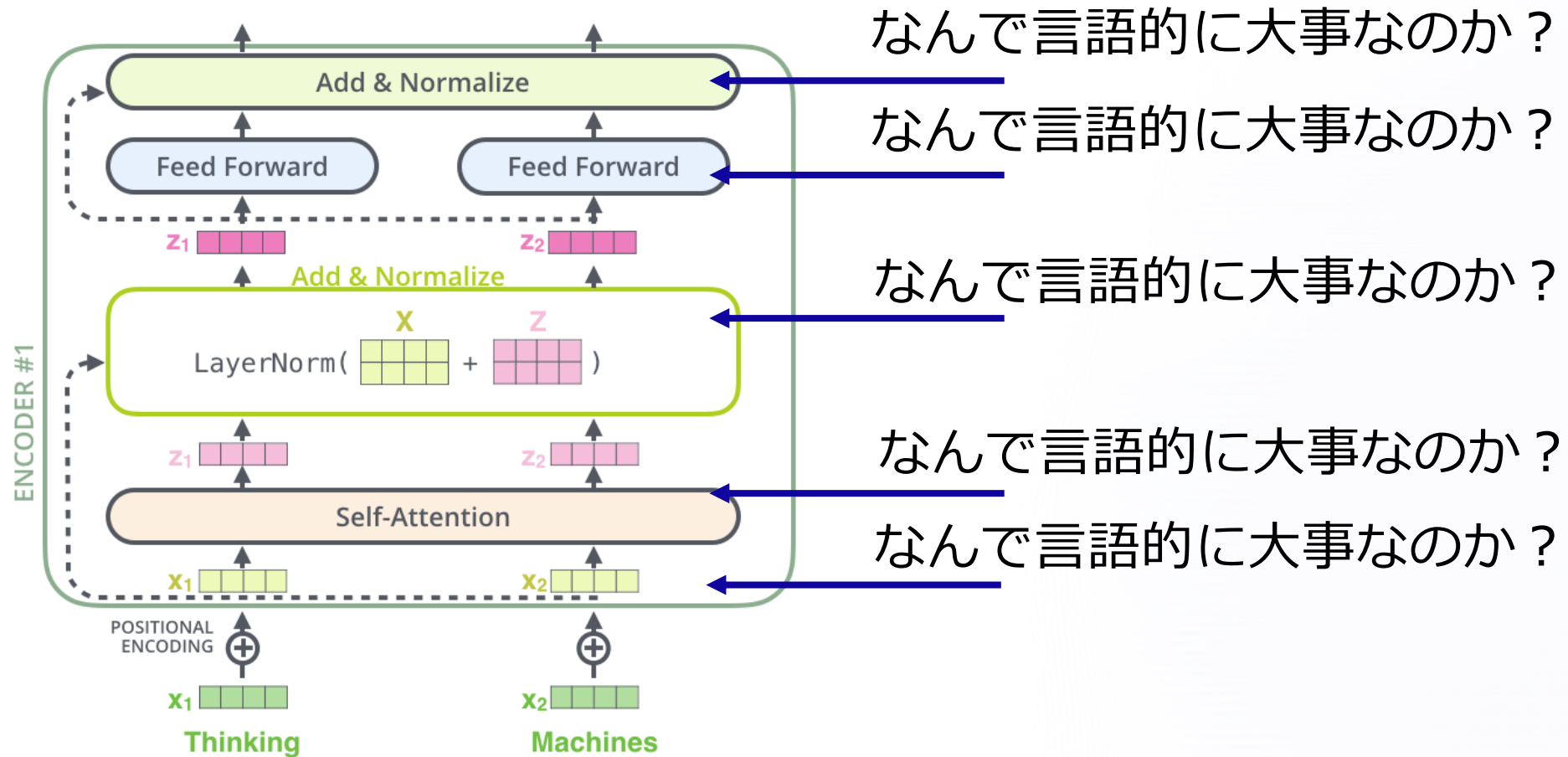
大規模言語モデルの訓練が線形回帰モデルの学習ぐらい軽くなっている未来ならできるかも（現状、網羅的探索は厳しい）

方向性①：色々試す（成功例からのトップダウン探索）

- 厳密に統制したA/Bテスト（アブレーション）ができること自体、人間の実験と比較したときの強み
 - 特にAとBの差分に直観が働く場合
 - **その検証は人間を用いた実験では本質的に困難であるため、仕方なく近似として概念実証に言語モデルを使う**
- 例：特定のモダリティは言語獲得にどう寄与するか？
 - そのモダリティの有無以外の条件が揃った2群の子どもで実験するのはほぼ不可能
- 例：モデル内部の特定の箇所を壊したら、言語能力はどうなるか？
 - 人間を壊すのは倫理的にだめ
- 例：ある規則を獲得するのに、どの程度の証拠が必要か？
- 例：どんな社会的要因（言語モデル集団）で、言語は変わっていくか？
- ...

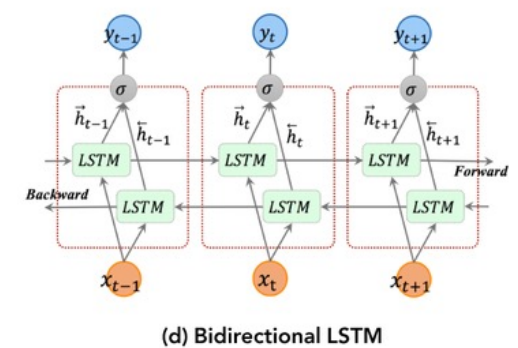
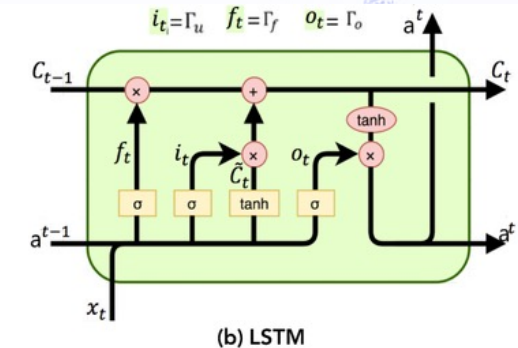
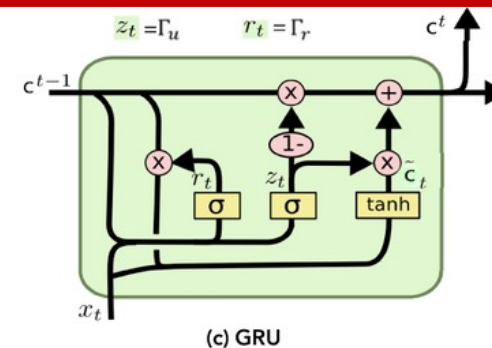
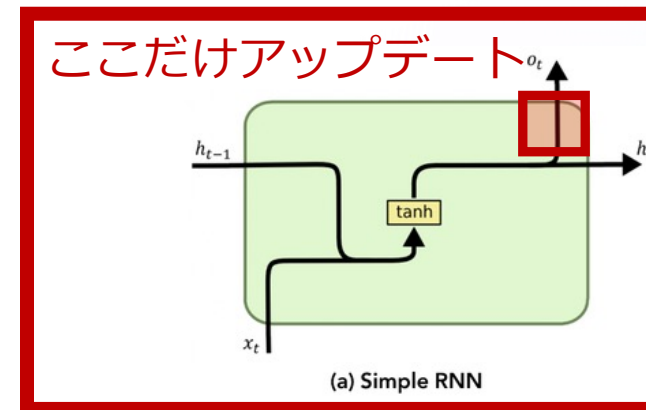
方向性①：色々試す（成功例からのトップダウン探索）

- 限界：特定の要因は直感的な解釈が困難（というかこのレベルの解釈が必要か？後述）



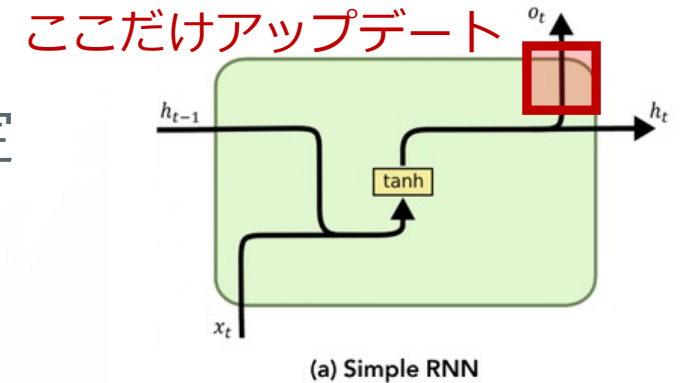
方向性①：色々試す（単純な設定からのボトムアップ探索）

- 逆に可能な限りシンプルな条件から考えてみる
- Syntactic Learnability of Echo State Neural Language Models at Scale.
(Ryo Ueda, Tatsuki Kuribayashi, Shunsuke Kando, Kentaro Inui. 2025)
 - 1層の単純なRNNを**特定の手法**で初期化し、出力層のみの訓練で文法知識を獲得できるか？
 - 学習誤差が時間方向に伝播しない
(cf. 人間の脳)



方向性①：色々試す（単純な設定からのボトムアップ探索）

- 回帰行列の初期値：漏れ率とスペクトル半径という2つのハイパーパラメータでRNNの初期値を制御・固定
 - 小さすぎると、前の文脈をすぐ忘却
 - 大きすぎると、前の文脈が増幅（ハウリング）してしまい混沌
- 様々なハイパーパラメータのもと文法性獲得能力を評価
- 過去の情報を「ちょうどよく」保持できる記憶力（カオスの縁と呼ばれ、複雑系科学・人工生命分野などで議論される）のもとで、言語モデルが文法をある程度獲得できる
 - 少なくとも、同パラメータ数・学習データで訓練されたTransformerと同程度の性能達成
 - つまり、重要なのはちょうどよい（短期）記憶の保持能力



問わねばいけない：成功・失敗条件がどれだけ言語特有か

- 言語モデルのあるパラメータAをいじると、言語の獲得ができなくなった
 - Aは本当に言語を語る際に避けられないのか？
- 例：学習率を1000倍にしたら言語獲得ができなくなった！学習率は言語獲得の理論に関係するに違いない
 - その設定で他モダリティ（画像・プログラミング言語・タンパク質系列…）の学習もできないのであれば、単にそれはニューラルモデルの最適化の話では？
ここからさらに神経科学や脳科学に繋がられる方がいるかもしれないが…
- 言語ドメインだけで成功・失敗する知見が得られたら、言語の特有性について何か言えるかもしれない
 - 言語に閉じて言語の研究をする限界かもしれない

方向性②：言語モデル as ドメイン一般的な学習者

- 言語モデル（ニューラルネット）は基本的に、画像・プログラミング言語・タンパク質系列・音声も含めて何でも学習できる**ドメイン一般的な学習モデル**
- 「言語モデル」が「言語のモデル」なのかは非自明
 - Raphaël Millière. 2024. Language Models as Models of Language (The Oxford Handbook of the Philosophy of Linguistics)
- 自然言語処理分野がたまたま最初に、**言語データ**に対してニューラル**モデル**を適用するために活用していたやり方
 - 系列中の次の要素を予測するようにニューラルモデルを訓練

方向性②：言語モデル as ドメイン一般的な学習者

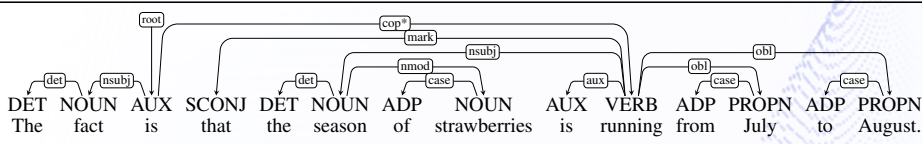
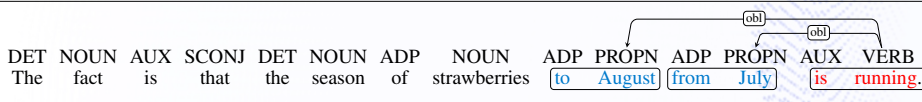
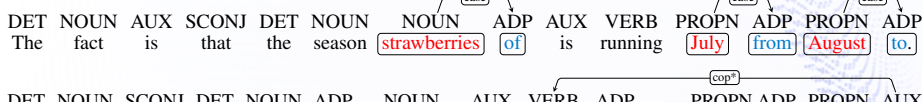
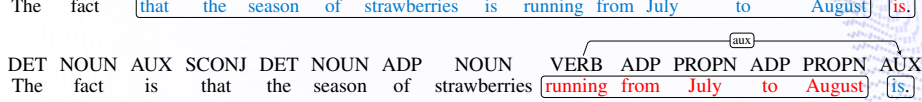
- アーキテクチャAを用いた、B層・パラメータ数Cの言語モデル（学習者）を、D法で初期化し、学習データEと目的関数Fのもと、最適化アルゴリズムGを用いて、Hステップ学習し…その結果、言語知識Yは獲得できる。

解釈

- 言語に特化しない手法でも、**ドメイン一般的（domain-general）な学習バイアス**のもとで、言語知識Yを獲得できる
 - 言語の獲得等が要請する事前知識・バイアスは、他のドメインの学習に必要なものと変わらない（Cf. 生成文法；例えば、そもそもマージ操作などは言語特有な能力として語るべきなのか）
 - Steven T Piantadosi. 2024. Modern language models refute Chomsky's approach to language.
 - に対する反論として：Jordan Kodner, Sarah Payne, Jeffrey Heinz. 2023. Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023)

- Can Language Models Learn Typologically Implausible Languages?
(Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, Alex Warstadt. TACL2025)
 - 個別言語の語順はドメイン一般的な学習者（言語モデル）にとっての学習しやすさで説明がつくか？
 - 英語・日本語の語順を架空のものに変更し、それで言語モデルの学習が滞るかを調査

- 結果：架空の語順では学習が（わずかに）滞る
- ドメイン一般的な学習者のバイアスのみでも語順の説明はある程度可能か
 - 入力で離れた位置にある要素どうしの依存（すなわち長距離依存）を捉えるのは苦手とか

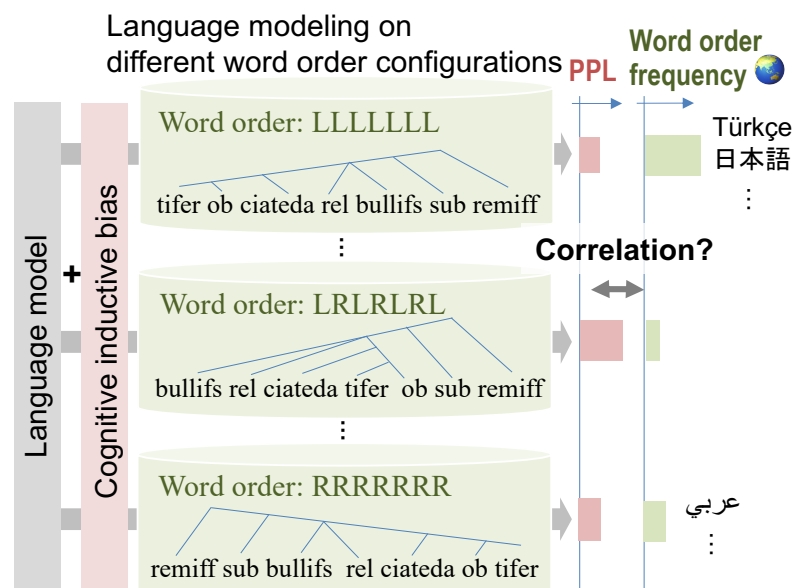
Correlation Pair	Example
Original	
<V, O>	
<Adp, NP>	
<Cop, Pred>	
<Aux, V>	
<Noun, Genitive>	

関連する批判

- 言語モデルは画像・プログラミング言語・タンパク質系列・ありえない言語も含めて何でも学習できるので、言語について何も語らない
(Noam Chomsky, Ian Roberts, and Jeffrey Watumull. "Noam chomsky: The false promise of chatgpt." *The New York Times* 2023) (ただの記事)
- 言語獲得には特別な学習バイアスが必要であるということが前提の批判
 - 伝統的に、言語学（特に統語論）の主要な関心は、可能な言語（文）と不可能な言語（非文）を区別する方法
- そもそも言語モデルは自然言語から逸脱した架空言語（例えば、動詞の活用が、必ず2単語後ろの単語に反映されるとか）も、実在する言語と同程度に学習できるか？
 - ありえない言語では、言語モデルの学習が確かに遅くなる（言語モデルは峻別可能）
 - Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, Christopher Potts. "Mission: Impossible Language Models." ACL2024 best paper.

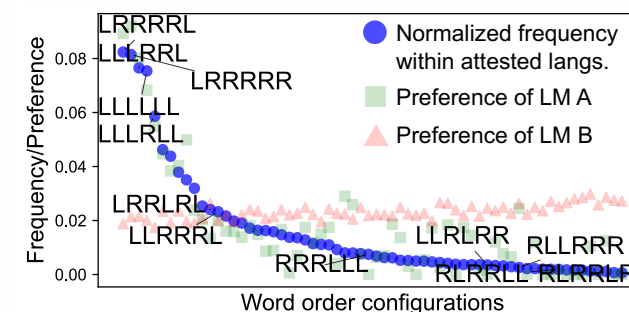
せっかくなので最近の研究紹介

- どんな言語モデルで、類型論的な語順頻度を、よりうまく再現できるか？
 - 文脈自由文法上で、語順をパラメタライズ
 - パラメタの組み合わせで異なる語順のコーパスを作成。どれが学習しやすいか実験
 - どのような学習条件（言語モデルの構造など）で、世界の語順分布を再現できるか



Param.	L	R
S	Cat eats.	Eats cat.
VP	Cat mouse eats.	Cat eats mouse.
PP	Cat table on eats.	Cat on table eats.
NP	Small cat eats.	Cat small eats.
Rel	Likes milk that cat eats.	Cat that likes milk eats.
Case	Cat-sub eats.	Sub-cat eats.

Table 1: Word-order parameters and example constructions with different assignments, L or R (See Apps. A and B and White and Cotterell (2021) for details).



言語モデルのワーキングメモリの弱さと、適切なアルゴリズム（left-corner）を用いた統語言語モデルが良い

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, Timothy Baldwin.
“Emergent Word Order Universals from Cognitively-Motivated Language Models.” ACL2024

せっかくなので最近の研究紹介

- そもそもどうやったら「言語モデルがどのような言語を好むか」という問いによりうまく答えられるか？
 - 例えば、どのような語順を好むか
 - Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, Brian Roark. Are All Languages Equally Hard to Language-Model? ACL2018
- データの問題

方法	自然さ	要因のアブレーション
自然言語を用いて比較	OK	(言語同士が複数の側面で異なってしまう)
人工言語を用いて比較	(形式言語など過度にシンプルになりがち)	OK
リアルで制御可能な人工言語がほしい	OK	OK

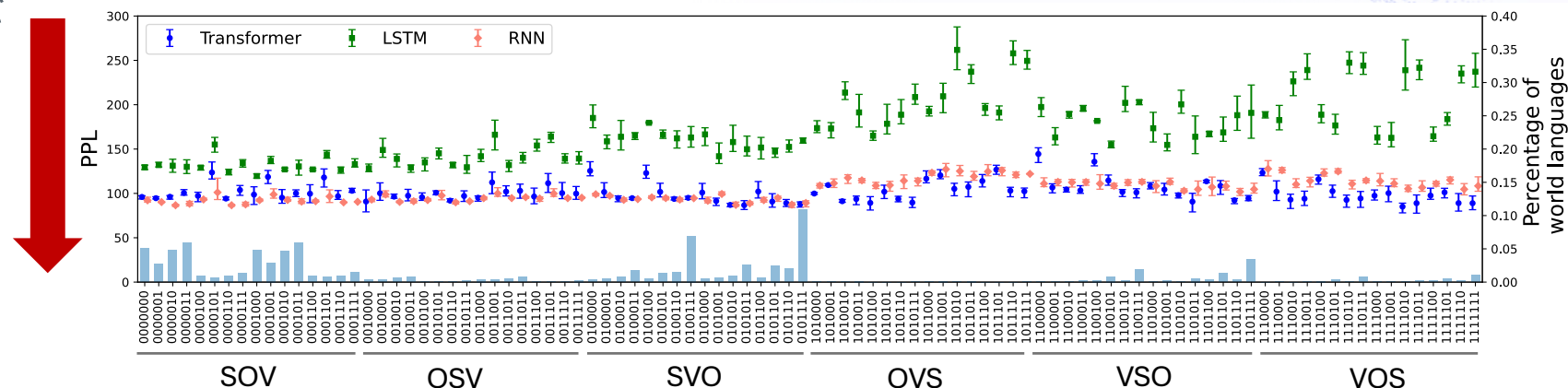
せっかくなので最近の研究紹介

- Which Word Orders Facilitate Length Generalization in LMs? An Investigation with GCG-Based Artificial Languages. (*Nadine El-Naggar, *Tatsuki Kuribayashi, Ted Briscoe. EMNLP 2025)

- CCGをパラメタライズして語順だけ違うコーパスを生成
 - 日本語みたいな語順、英語みたいな語順… (96パターン)

Param.	Description	0 (head-final)	1 (head-initial)
S	Order of subject and verb	VI $\rightarrow$ S\NP _{SUBJ}	VI $\rightarrow$ S/NP _{SUBJ}
		VT $\rightarrow$ (S\NP _{SUBJ})\NP _{OBJ}	VT $\rightarrow$ (S/NP _{SUBJ})\NP _{OBJ}
		VCOMP $\rightarrow$ (S\NP _{SUBJ})\SCOMP	VCOMP $\rightarrow$ (S/NP _{SUBJ})\SCOMP
VP	Order of object and verb	VT $\rightarrow$ (S\NP _{SUBJ})\NP _{OBJ}	VT $\rightarrow$ (S\NP _{SUBJ})\NP _{OBJ}
		VCOMP $\rightarrow$ (S\NP _{SUBJ})\SCOMP	VCOMP $\rightarrow$ (S\NP _{SUBJ})\SCOMP
		REL $\rightarrow$ (NP _{SUBJ} \NP _{SUBJ})\NP _{OBJ}	REL $\rightarrow$ (NP _{SUBJ} \NP _{SUBJ})\NP _{OBJ}
O	Order of subject and object	Subject occurs before the object	Object occurs before the subject
COMP	Position of complementizer	COMP $\rightarrow$ SCOMP\S	COMP $\rightarrow$ SCOMP/S
PP	Postposition or preposition	PREP $\rightarrow$ (NP\NP)\NP	PREP $\rightarrow$ (NP\NP)\NP
ADJ	Order of adjective and noun	ADJ $\rightarrow$ NP\NP	ADJ $\rightarrow$ NP\NP
REL	Position of relativizer	REL $\rightarrow$ (NP _{SUBJ} \NP _{SUBJ})\NP _{OBJ}	REL $\rightarrow$ (NP _{SUBJ} \NP _{SUBJ})\NP _{OBJ}

- 複雑な並列構造や弱文脈依存な構文が人工言語コーパスに自然に入る (Cf. PCFG)
 - そのうえでどの語順が学習しやすいか調査



方向性③：経験論？の支持

- アーキテクチャAを用いた、B層・パラメータ数Cの言語モデル（学習者）を、D法で初期化し、学習データEと目的関数Fのもと、最適化アルゴリズムGを用いて、Hステップ学習し…その結果、言語知識Yは獲得できる。

解釈

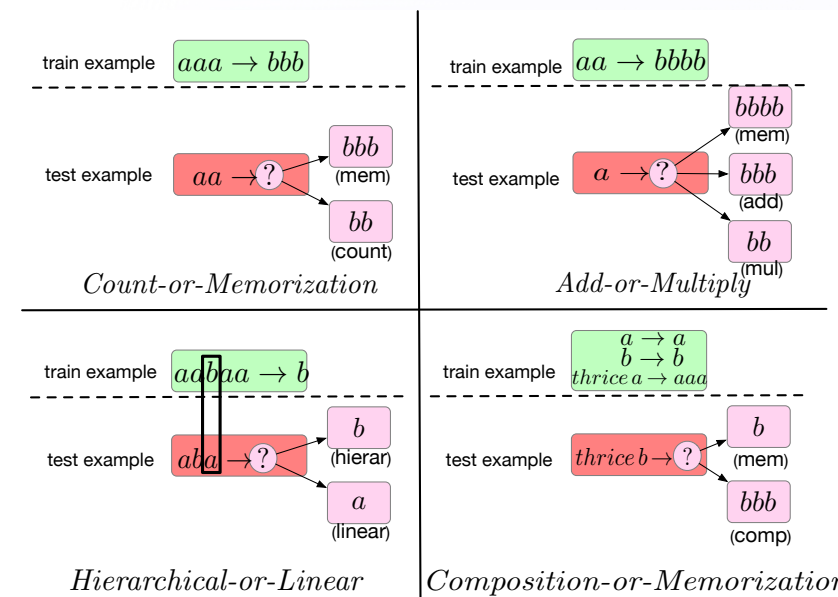
- **事前知識を初期化時に持たないニューラルネットワーク**でも、たくさんテキストデータを学習すれば、言語知識Yを獲得できる

本当に経験だけか？

- 事前知識を初期化時に持たないニューラルネットワークでも、たくさんテキストデータを学習すれば、言語知識Yを獲得できる
- 論点1：アーキテクチャを定義した時点で（何かハイパラを設定した時点で）帰納バイアス（学習者の癖）は入る
 - 帰納バイアスがなければ学習は不可能（Tom M. Mitchell. 1980. The Need for Biases in Learning Generalizations.）
 - モデル初期化後、一事例からの汎化傾向がアーキテクチャごとに異なる
 - その帰納バイアスと言語獲得の関係性について論じることになる

	length l	FPA $\uparrow$		L , nats $\downarrow$	
		count	mem	count	mem
LSTM-s2s no att.	40	1.00	0.00	0.01*	97.51
	30	0.97	0.00	0.01*	72.67
	20	0.07	0.00	2.49*	55.67
	10	0.00	0.00	88.27	48.67*
LSTM-s2s att.	40	0.99	0.00	7.84*	121.48
	30	0.96	0.02	1.14*	83.48
	20	0.70	0.16	5.73*	49.33
	10	0.00	0.20	98.12	8.46*
CNN-s2s	{10, 20, 30, 40}	0.00	> 0.90	> 592.92	<1.31*
Transformer	{10, 20, 30, 40}	0.00	> 0.97	> 113.30	<11.14*

(a) Count-or-Memorization



本当に経験だけか？

- 論点2：言語モデルを定義した時点でいろいろなことを仮定している
 - 単語区切り・トークナイズできる（トークナイザの学習はどこで？）
 - 文字を認識・出力できる
 - 大量のテキストデータを読める
 - 次の単語の予測精度を上げようとしている
- データのみから言語を獲得できるかに関する議論は、音声モデルなり、ロボットなりで検証してもいいと思う
 - 書き言葉の習得は言語獲得の後半戦

本当に経験だけか？

- 生音声データで音声モデルを訓練した場合と、音素や単語に転写したあと言語モデルを訓練した場合で、後者のほうが圧倒的に言語を学習しやすい
 - トークナイザで仮定されている能力が結構すごい？

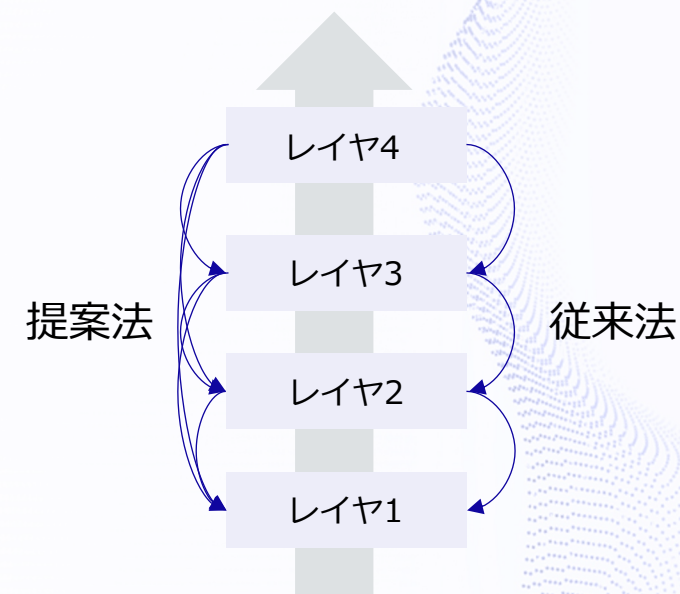
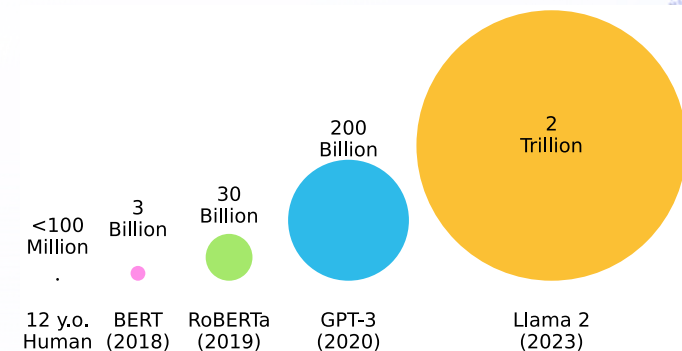
Table 3: **The BabySLM benchmark.** Lexical and syntactic accuracies obtained by different language models trained on developmentally plausible corpora of speech, phonemes, or words. Numbers are computed on the test set, and performances on the development set are reported using small font size. The starred cumulated duration and number of words are estimates based on the 1.2 M of words present in the 128 hours of speech from Providence. Data plausibility indicates the extent to which the training set is close to the real sensory signal available to infants.

System	Input	Training set	Cumulated duration (h)	Number of words (M)	Data plausibility	Lexical acc. (%)	Syntactic acc. (%)
Random baseline	—	—	0	0	—	49.2 52.5	49.3 50.0
STELA [27]	speech	SEEDLingS	1024	9.6*	+++	49.5 45.4	50.3 50.5
STELA [27]	speech	Providence	128	1.2	++	56.8 47.1	50.3 51.1
LSTM	phonemes	Providence	128	1.2	+	75.4 75.2	55.1 55.9
LSTM	words (BPE)	Providence	128	1.2	+	—	65.1 65.3
BabyBERTa [9]	words (BPE)	AO-CHILDES	533*	5	+	—	70.4 70.4

Lavechin, Marvin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. "BabySLM: language-acquisition-friendly benchmark of self-supervised spoken language models." In INTERSPEECH 2023.

凄まじい経験ではないか？

- 言語モデルの学習環境を人間のそれになるべく近づけて議論しよう
 - BabyLM Challenge (<https://babylm.github.io/>)
 - データの量的制限 (100M tokens)
 - データの質的変更 (親子の対話データ; CHILDESなど)
 - マルチモーダルな入力 (Round 2)
 - 他の大規模モデルとのコミュニケーションを許容 (Round 3)
- 初回優勝手法：レイヤ間の残差結合について、直前レイヤだけでなく、それより前のすべてのレイヤと直接繋ぐ！
 - Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every layer counts BERT.
 - それは一体、言語的にどういった示唆があるのだろうか (もしかしたら本当に意味があるかもしれない)
 - リーダーボード的な枠組み vs. 科学的探求



(少し脱線) 帰納バイアスの研究にする

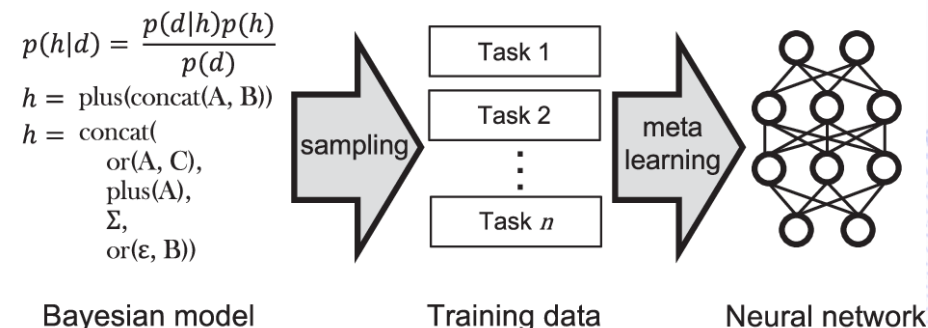
- この事例からこう汎化してほしいという帰納バイアス (prior) を明示的に言語モデルに取り込み、帰納バイアスと言語獲得の関係の研究を行う
 - 単純な帰納バイアス ($aabb \rightarrow a^n b^n$) が、複雑な言語の獲得にどう影響を及ぼすか？
 - このとき言語モデルは単なる容器のような使われ方

b

- メタ学習で実装可能
 - 学習の仕方の学習 (2階微分を伴う)

例 :

- Modeling rapid language learning by distilling Bayesian priors into artificial neural networks (McCoy&Griffiths, 25)
- Human-like systematic generalization through a meta-learning neural network (Lake&Baroni, 23)



言語・人間を測る編

言語モデルはなんのモデルか？

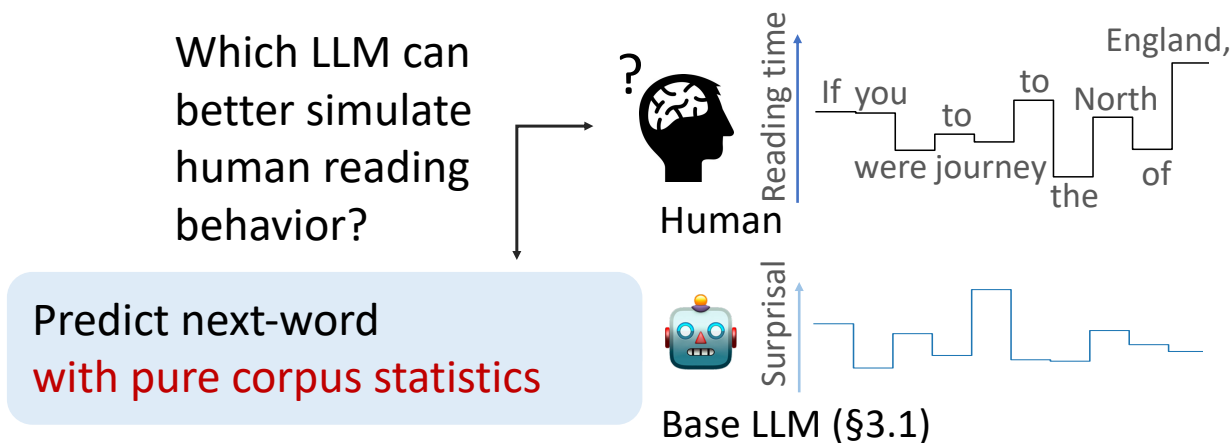
レベル	関心	認知科学	言語モデル
計算	何の問題を解いているか？ 理想的な解法は何か？	認知のベイズモデルと いった原則から理解され る振る舞い	目的関数 ← 人間が設定 しているので 自明
アルゴリズム	その解法をどのような表現 やプロセスで近似している か？	反応時間やエラーパター ンから推測される心的表 層やプロセス	内部表現や振る 舞いの分析
実装	表現やアルゴリズムがどの ように物理的に実現されて いるか？	fMRIなどで観測される神 経回路	回路分析・ ニューロンの解 釈など

Alexander Ku, Declan Campbell, Xuechunzi Bai, Jiayi Geng, Ryan Liu, Raja Marjeh, R. Thomas McCoy, et al. 2025. “Levels of Analysis for Large Language Models.”

言語モデルは紛れもなく次単語予測のモデル
人間の言語処理のうち予測に基づく側面については掘り下げられそう？

言語モデルを使って初めて測れる値で分析する

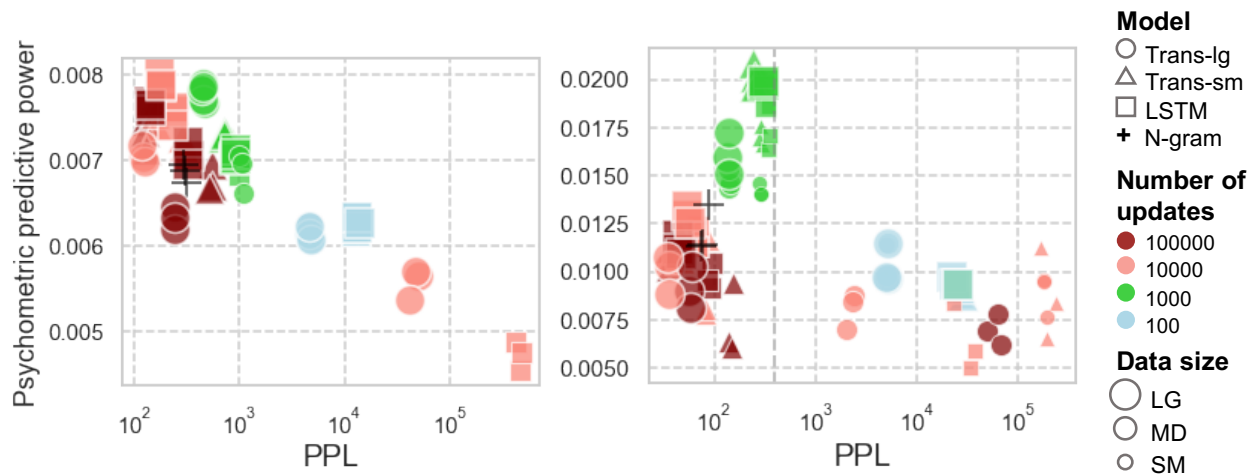
- ある現象Yが言語モデルから得られる値Xで説明できるか？
 - 言語モデルをある種のアノテーションツールとして使用
- 例：人間の単語ごとの読み時間（処理負荷）の違いが、単語の予測しやすさで説明できるか？
 - 人間は無意識に先読みをしているのか？
 - 読み時間がかかる単語（高負荷） $\propto$ 予測を裏切る単語
 - 言語モデルは次の単語をなるべく正確に予測するように訓練されているので、（コーパス統計として）比較的正確な単語の予測しやすさを近似可能



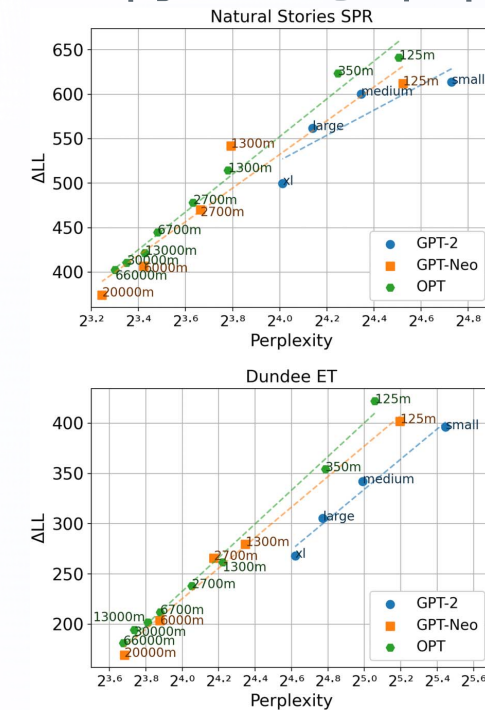
Tatsuki Kuribayashi, Yohei Oseki, Timothy Baldwin.
"Psychometric Predictive Power of Large Language Models." Findings of NAACL2024

言語モデルを使って初めて測れる値で分析する

- 人間の予測に基づく処理負荷（人間の言語モデル）は、どれほどコーパス上の次単語予測確率と相関するか？
 - コーパスを正確にモデリングできる言語モデルが現れて初めて検証可能
- 予測の正確な（パープレキシティの低い）言語モデルから得られる確率値ほど、人間の読み時間をうまく説明できるか？
 - そうとも限らない
 - コーパス上の言語モデル ≠ 人間の頭の中の言語モデル



Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, Kentaro Inui. "Lower Perplexity is Not Always Human-Like." ACL2021



Byung-Doh Oh and William Schuler. 2023. "Why Does Surprisal from Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?" TACL

言語モデルを使って初めて測れる値で分析する

- シンプルな指標としてしばしばサプライザル $-\log_2 p(\text{word}|\text{context})$ が使われるが、もちろんそれに限らない
 - エントロピー、レニーエントロピー、エントロピーリダクションなど、しばしば認知モデリングで活用される

- 情報密度の分布

- 一様に分布する、周期的に波打っている等

- テキストコーパスのもつ情報と他モダリティ（音韻など）の関係

- どれほど相補的か？

- 情報理論的アプローチはとりわけCogSciで盛んか

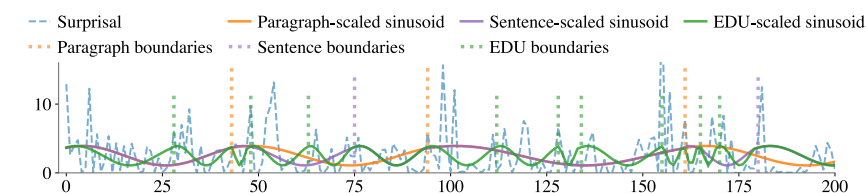
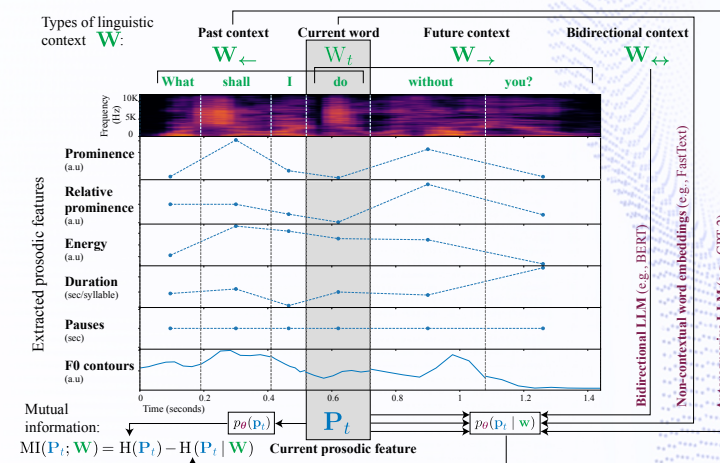


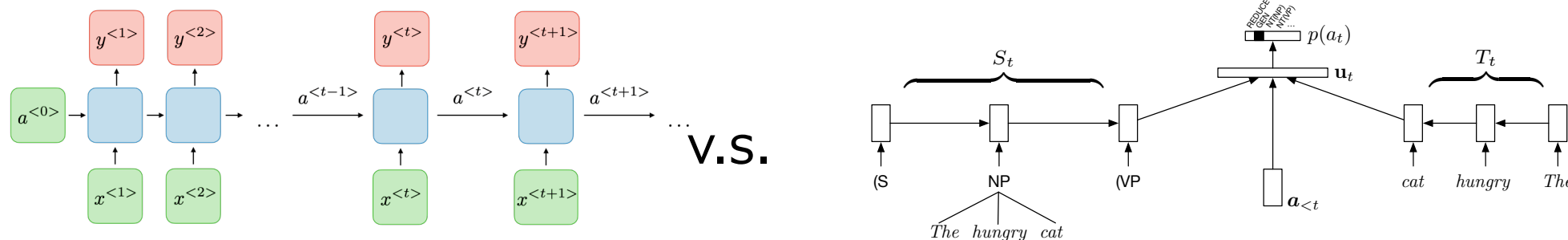
Figure 1: **Illustration of Harmonic Regression on Surprisal Contours.** Surprisal contours, unit boundaries, and first-order sinusoids for the first 200 tokens from a Wall Street Journal article (document wsj_1111 in the English RST Discourse Bank). Time scaling (§3.2) is applied according to the lengths of elementary discourse units (EDUs), sentences, and paragraphs. Here, we set the coefficients of the sinusoids to 1 for illustrative purposes. See Fig. 3 and App. H.1 for realistic decompositions.

Tsipidi+, 25. The Harmonic Structure of Information Contours. ACL2025.



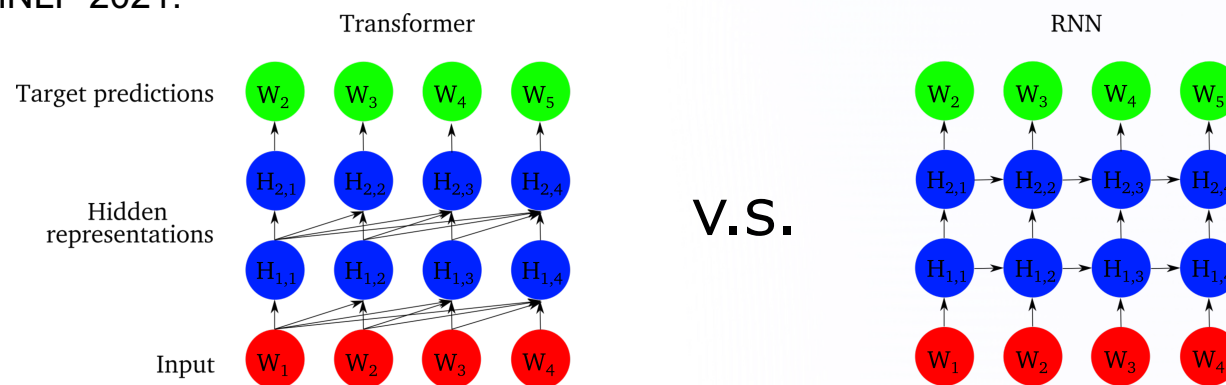
Wolf+, 23. Quantifying the redundancy between prosody and text. EMNLP 2023.

- どのような言語モデルから得られる値で対象をうまく説明できるか？
- ただの言語モデルから得られる単語確率 vs. 統語構造を考慮したモデル (パーザ) から得られる単語確率
 - 人間の予測処理はどれほど文法にバイアスを受けているか？



Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. 2021. “Modeling Human Sentence Processing with Left-Corner Recurrent Neural Network Grammars.” EMNLP 2021.

- Self-attention機構は認知的に妥当か？

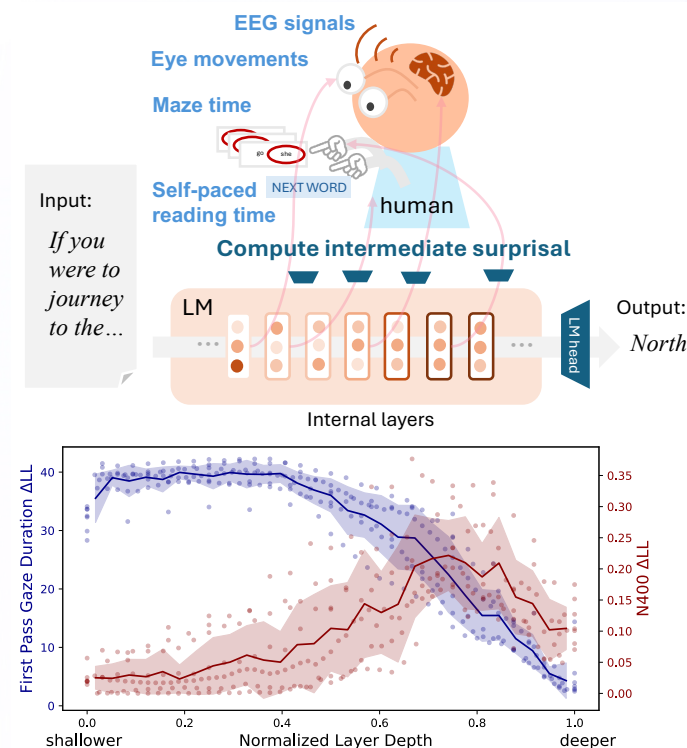


Danny Merkx, and Stefan L. Frank. 2021. “Human Sentence Processing: Recurrence or Attention?” CMCL2021.

次なるモデルは？

- ガーデンパス文のような統語的曖昧性の解消にかかる（再解析）コストを言語モデルの予測だけでは説明できない
 - Huang, Kuan-Jung, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. “Large-Scale Benchmark Yields No Evidence That Language Model Surprisal Explains Syntactic Disambiguation Difficulty.” *Journal of Memory and Language*.
- 自ら再解析する言語モデルはどうやったら作れるか？
- サプライザルという指標が良くないだけで、言語モデルの内部的には再解析が起きていないか/起こせないか？

- Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, Timothy Baldwin. 2025. "Large Language Models Are Human-Like Internally." TACL.
- 言語モデルのどの部分から得られる次単語予測で人間の読み活動をうまく説明できるか？
 - 言語モデルは層が積み重なっているモデル
- 言語モデルの前半・後半層で人間の異なる振る舞い・生理データと対応がとれやすい
 - とりわけ、First pass gaze durationのような速い反応は前半層、N400のような遅い反応は後半層
 - 読み活動データの実時間スケールと層方向の時間スケールが関係ありそう



まとめ

- 言語モデルが言語研究にどのような示唆を与え得るか
- 人間の実験では実現できない概念実証や、ただコーパスを眺めるだけでは測定困難な値を用いた分析が可能
 - 獲得のアブレーション
 - 情報理論的量のアノテーション
- 言語にコミットしない一般的な方法で、言語現象をどれほど説明できるか
 - 言語の特殊性を明らかにしようとする（典型的な）言語学の試みとは、ある種逆側から境界に迫っている
- 基本方針として、ある現象を再現できる十分条件を明らかにしていく
 - そしてその条件について、人間との対応や言語特有さの観点から解釈していく