



Learning the (un)attested: Modeling acquisition of island constraints using symbolic and neural language models

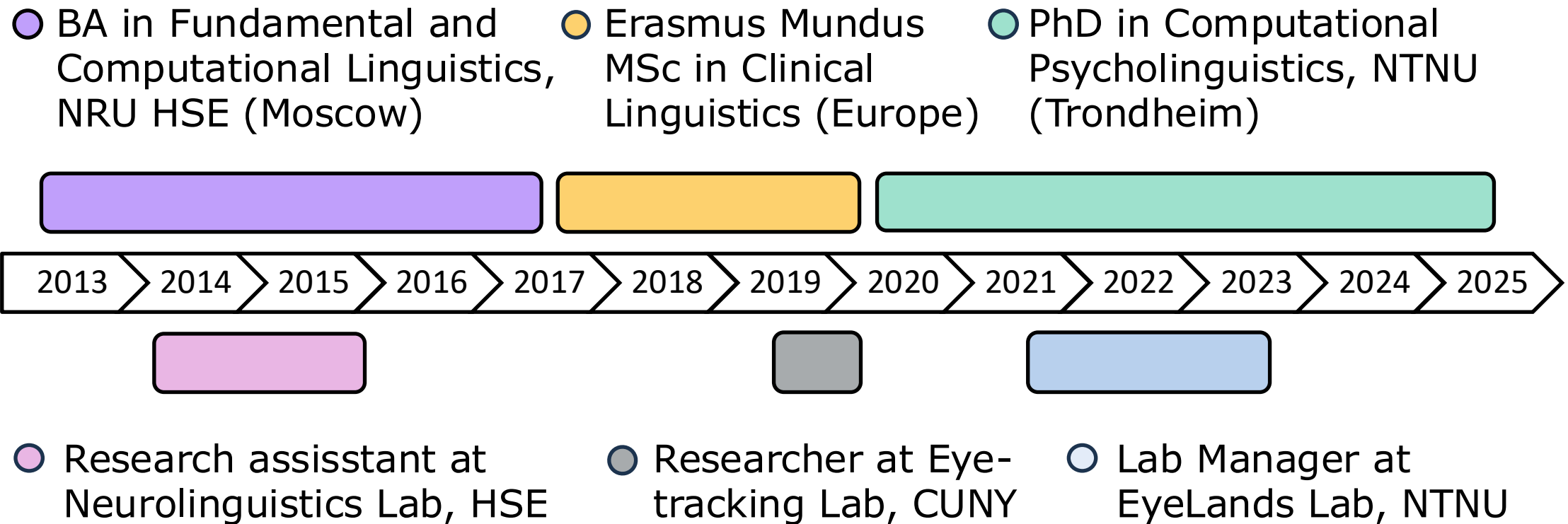
LTG Seminar presentation by Anastasia Philipps (born Kobzeva)

Oslo, 20th of January 2026



About me

- You can call me Nastia 😊



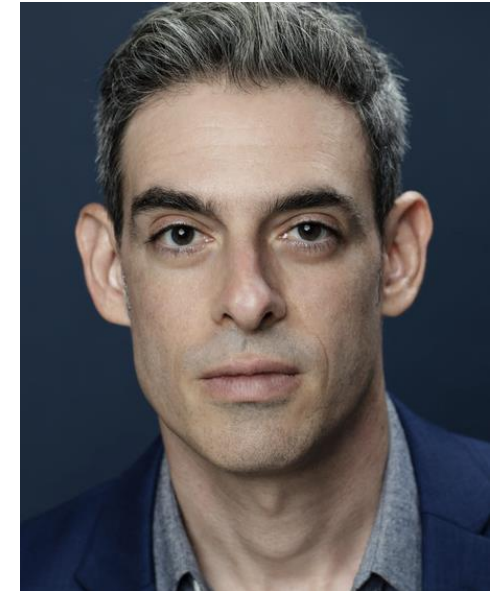
PhD project

Supervisors:

- Computational psycholinguistics
- Testing learnability of constraints on filler-gap dependencies (FGDs)
- Modeling and behavioral experiments in Norwegian + input corpora analyses



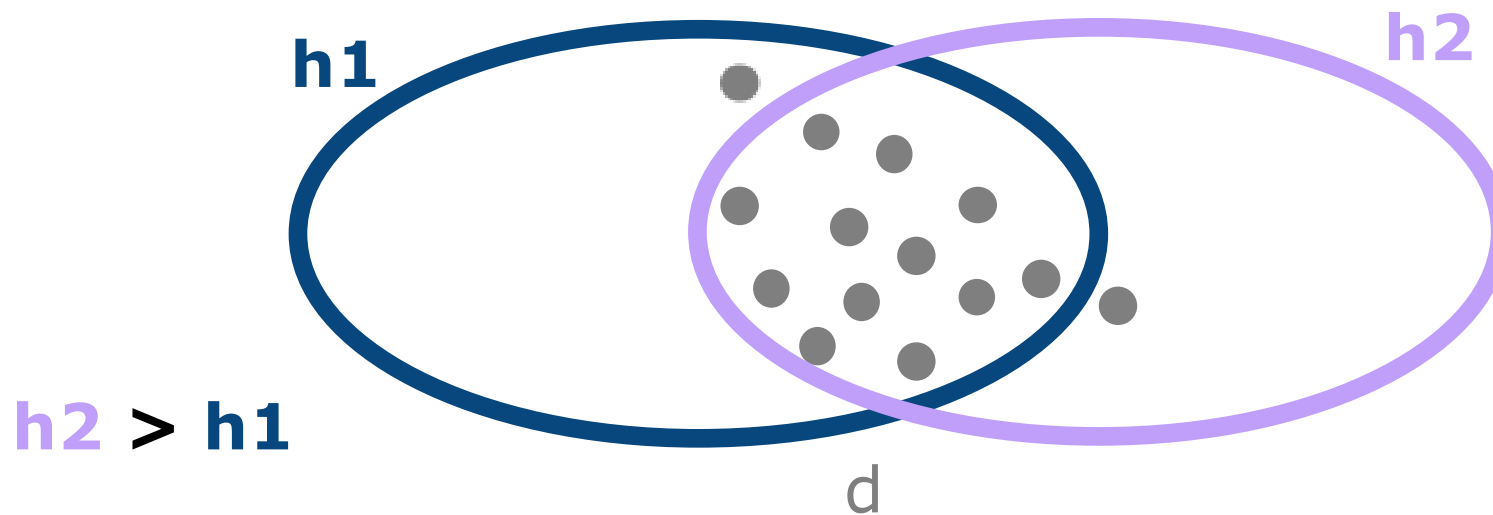
Dave Kush
University of Toronto



Tal Linzen
New York University

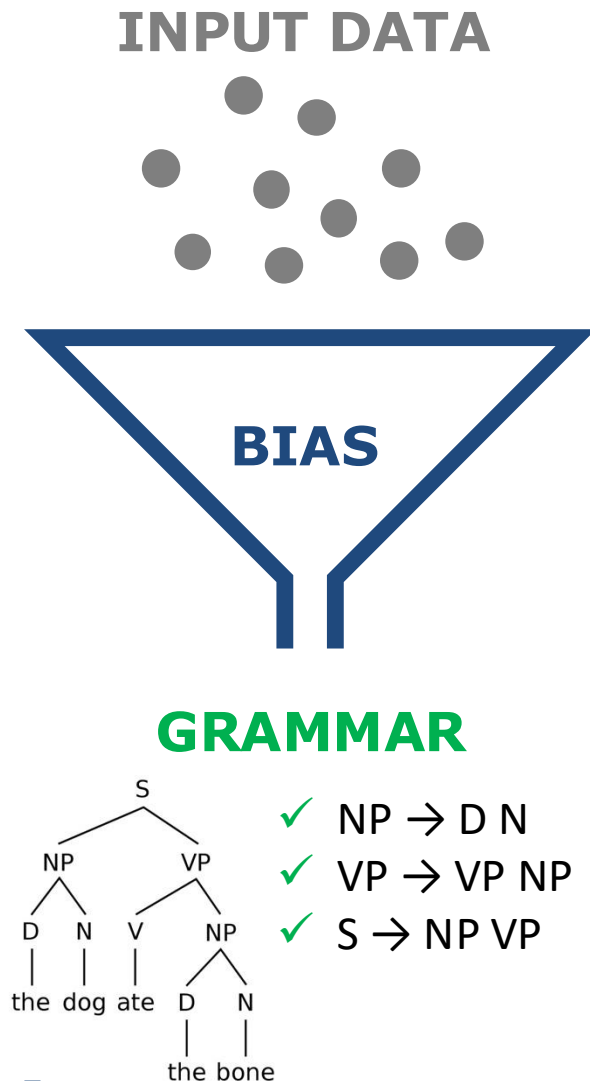
Background

- Learners make complex generalizations from limited input data
 - **Poverty of the Stimulus (POS):**
Input data that children encounter is compatible with multiple hypotheses about their target grammar



→ There must be some **learning biases** that guide acquisition

Learning biases



- The nature of learning biases is a topic of debate
 - Innate vs. acquired
 - Domain-general vs. domain-specific
- **Linguistic nativist:** innate & domain-specific biases
- **Empiricist:** innate domain-general biases
- Computational modeling allows us to test what kind of learning biases are sufficient for an acquisition task

FGDs and island constraints

- FGDs are contingencies between fillers and gaps:
 - He said that she bought wine for her friend.
 - *What_i* did he say that she bought _____i? wh-FGD
 - He knows about *the wine_i* that she bought _____i. RC FGD
- FGDs seem to be unconstrained in length:
 - *What_i* did he say they thought he knew that she bought _____i?
- However, there are island constraints on FGD formation:
 - **What_i* did he wonder whether she bought _____i? whether-island
 - **What_i* has the letter about _____i created problems? subject island

Island learnability challenge

- The input lacks evidence that islands exist
 - Absence $\neq$ unacceptability: highly-embedded FGDs are absent too
 - Learning about islands requires abstract generalization beyond surface frequency patterns
 - Islands are acquired relatively early despite their structural complexity (De Villers et al., 2008)
 - Relative cross-linguistic consistency of islands
- Support nativist accounts (Philipps, 2013) → innate domain-specific biases for island acquisition



Modeling island acquisition in English

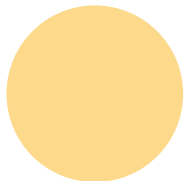
Can islands be learned with fewer domain-specific biases?

Pearl & Sprouse: Symbolic model

- Strong representational bias: syntactic structure as input
- Simple statistical inference
- Reduced complexity of domain-specific biases relative to traditional proposals

Wilcox et al.: Neural lang. models

- Weak representational bias: raw text as model input
- Powerful statistical inference
- Proof-of-concept simulations that islands can be learned by domain-general learners



Islands in Norwegian

- Overall, Norwegian is very similar to English when it comes to FGDs
 - Allows long-distance dependencies with gaps in several syntactic positions
 - Sensitive to some of the same island constraints

*Hva_i har brevet om _____i skapt problemer?

subject island

`*What_i has the letter about _____i created problems?'

*Hva_i spiste du kake da han spiste _____i?

adjunct island

`*What_i did you eat cake when he ate _____i?'

Islands in Norwegian

- Allows FGDs into embedded questions (EQs) – environments that are islands in English (wh-/whether-islands)
 - FGDs into EQs are accepted by native speakers and found in written corpora (Kush et al., 2018; 2019; 2021)

Vi var redde for noe vi ikke visste [hva ____ var].

`*We were afraid of something we did not know [what ____ was].'

- Excellent testing ground for exploring whether variable judgment patterns can be induced from the input



Main research questions



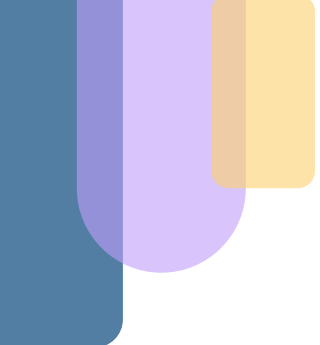
1. Can the models induce the (non-)island status of different domains in Norwegian?



2. What kind of evidence does the input provide, and how does it influence model generalizations?



3. Can the models generalize beyond the input – for example, across constructions?



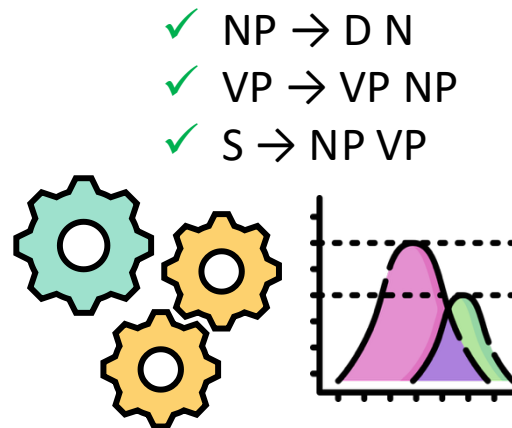
Research dissemination

- C1** Kobzeva et al. (2022). LSTMs Can learn basic wh- and relative clause dependencies in Norwegian. *CogSci Proceedings*
- C2** Kobzeva et al. (2023). Neural networks can learn patterns of island-insensitivity in Norwegian. *SCiL Proceedings*
- J1** Kobzeva & Kush (2024). Grammar and expectation in active dependency resolution: Experimental and modeling evidence from Norwegian. *J Cog Science*
- J2** Kobzeva & Kush (2025). Acquiring constraints on filler-gap dependencies from structural collocations: Assessing a computational learning model of island-insensitivity in Norwegian. *Language Acquisition*
- J3** Kobzeva et al. (2025). Learning filler-gap dependencies with neural language models: Testing island sensitivity in Norwegian and English. *J Memory & Language*

Modeling approaches

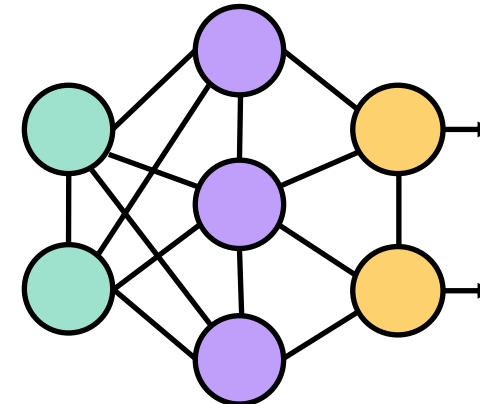
Symbolic cognitive models

- Traditional approach
- Top-down & rule-based
- Strong inductive biases



Neural language models

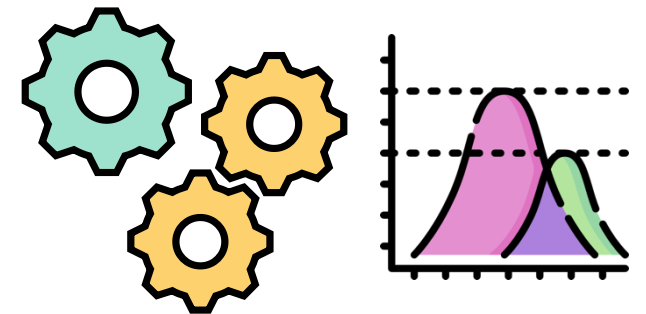
- Approach rooted in the connectionist tradition (Rumelhart & McClelland, 1986; Elman, 1990)
- Bottom-up & data-driven
- Weakly-biased



Symbolic cognitive models

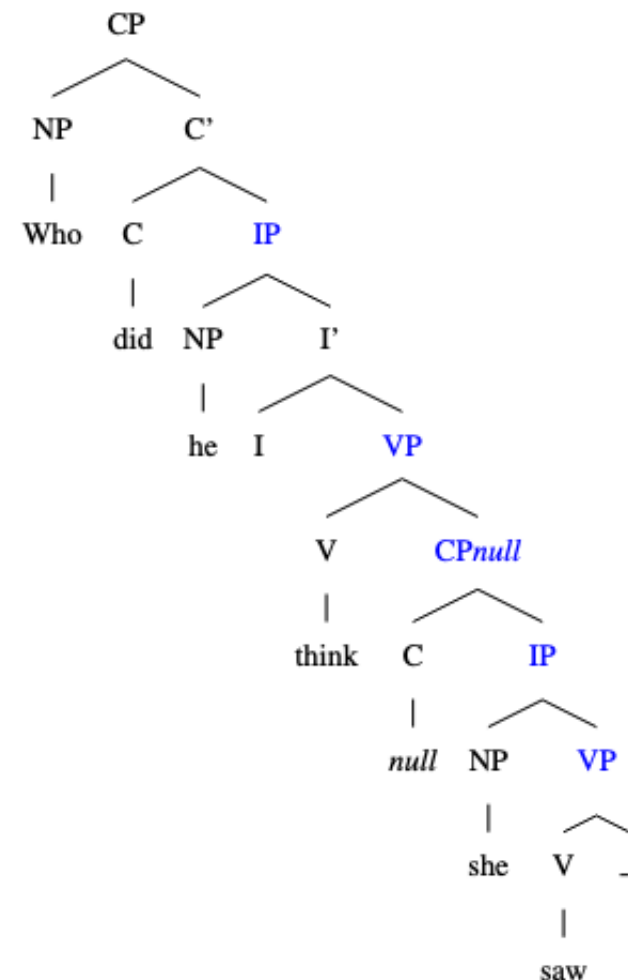
- Symbolic representations
 - E.g., syntactic categories
- Cognitively plausible components
 - Based on previous research findings
- Explicit linking hypotheses about learning mechanisms & outcomes in models and humans
 - E.g., probability linking to acceptability
- Inference mechanism
 - Counts, reinforcement learning, Tolerance principle, Bayesian inference, etc. (Pearl, 2023)

✓ $NP \rightarrow D N$
✓ $VP \rightarrow VP NP$
✓ $S \rightarrow NP VP$



Symbolic model of Pearl & Sprouse

- Annotated CHILDES data as model input
- Behavioural AJT data as model target
 - Probability linking to acceptability
- Tracking probabilities of trigrams of **container nodes**: phrasal tree nodes that constitute the path between the filler and the gap (IP, VP, CP...)
- START and END nodes are added to track FGD boundaries
- $P(\text{FGD}) = P(\text{START-IP-VP}) * P(\text{IP-VP-CP}_{\text{null}}) * P(\text{VP-CP}_{\text{null}}\text{-IP}) * P(\text{CP}_{\text{null}}\text{-IP-VP}) * P(\text{IP-VP-END})$



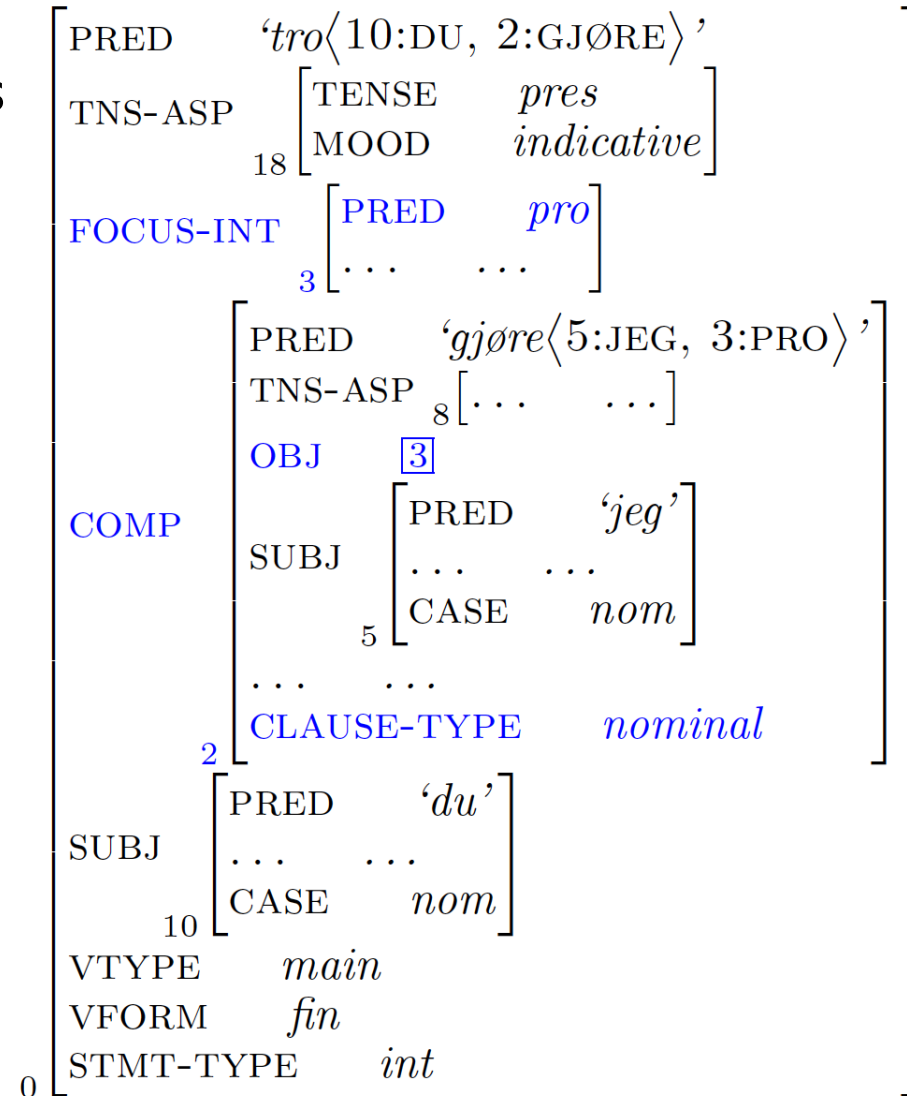
Applying the model to Norwegian

- As model input we used data from children's fiction corpus (4 mil tokens) annotated within Lexical-Functional Grammar

Hva tror du jeg gjorde ____?

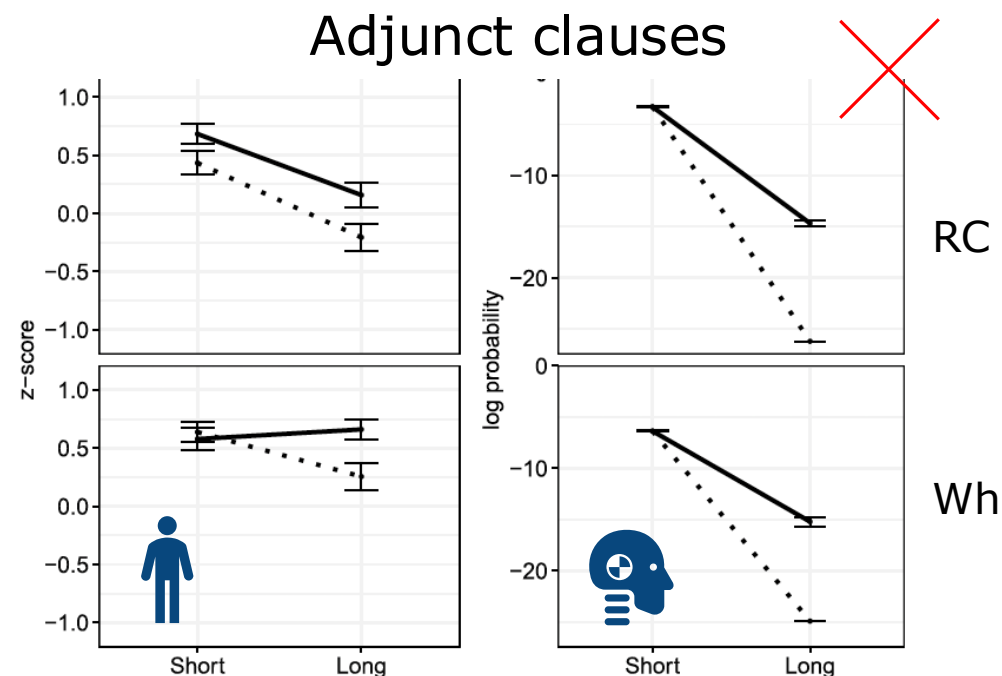
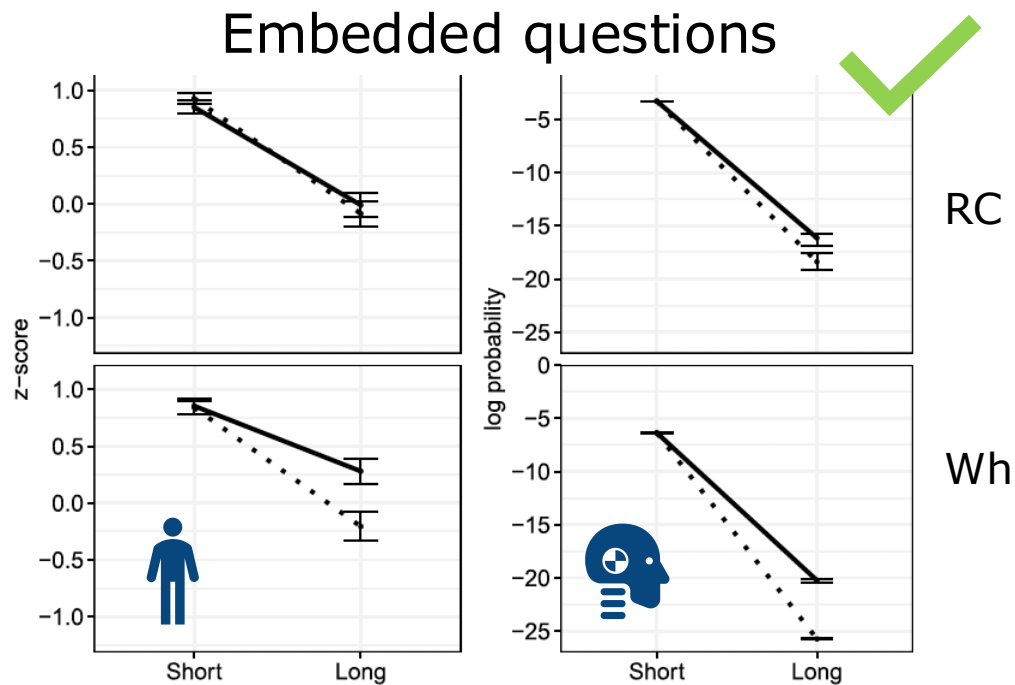
'What do you think I did ____?'

- F-structure labes used as container nodes instead of phrasal tree nodes
 - START COMP_{nominal} OBJ END vs.
 - START IP VP CP_{null} IP VP END
- Two model implementations:
 - Trigram and bigram-based probabilities



Results of the symbolic model

- The model learned that EQs are not islands in Norwegian but failed on other cases (e.g., adjunct clauses for RC-dependencies)
- Input data lacked examples like "Who do you think __ came?"
- The richer tagset of LFG allowed for bigram-based probability calculation

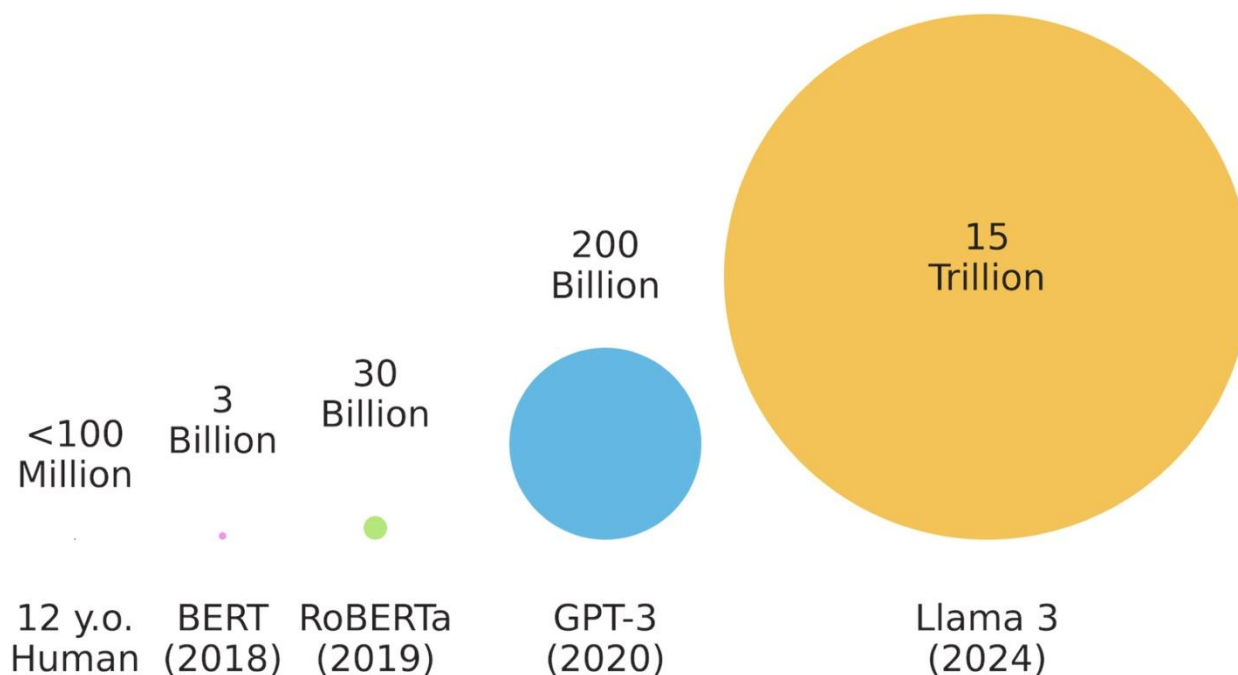


Neural language models

- Rooted in the connectionist tradition (Rumelhart & McClelland, 1986; Elman, 1990)
 - Early models were seen as plausible cognitive models of how computation proceeds in human brain
- NLMs compute a probability distribution over words in their vocabulary to predict the next word in a sequence of words
 - Next word prediction → Similar to human sentence processing (Hale, 2001; Levy, 2008)
- Taken to be domain-general learners without hierarchical bias towards the structure of natural languages
- Develop distributed representations over the course of training

Neural language models

- The exact nature of NLM biases is unknown (black boxes)
- Input usually differs quantitatively and qualitatively from humans
 - For early models, this wasn't always the case, and the input was often abstracted into hierarchical labels (Lewis & Elman, 2001)
 - Modern NLMs are often trained on huge amounts of Internet text



Utility of NLMs in LA research

- Can we use NLMs to study human linguistic cognition?
 - Hotly debated topic
- Spectrum of various positions, according to which NLMs can:

Teach us nothing
about language

Be useful in principle,
but are limited at
present

Provide proof-of-
concept for usage-
based theories

Serve as actual
theories of LA

Bolhuis et al., 2024;
Chomsky et al., 2023

Ziv et al., 2025; Lan et
al., 2024; Vázquez
Martínez et al., 2024;

Futrell & Mahowald,
2025; Warstadt &
Bowman, 2022

Piantadosi, 2023

Neural language models

- It is challenging to assess what internal components of NLMs have
 - **Targeted evaluation**: testing how NLMs fare as incremental processors in psycholinguistic experiments (Futrell et al., 2019)
 - **Dependent variable**: surprisal (inverse log probability)

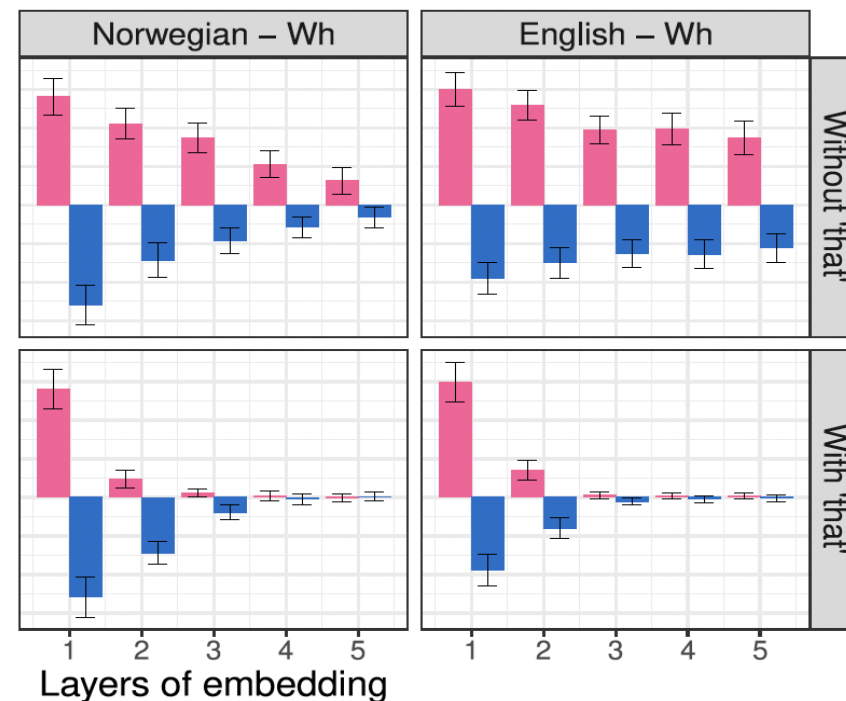
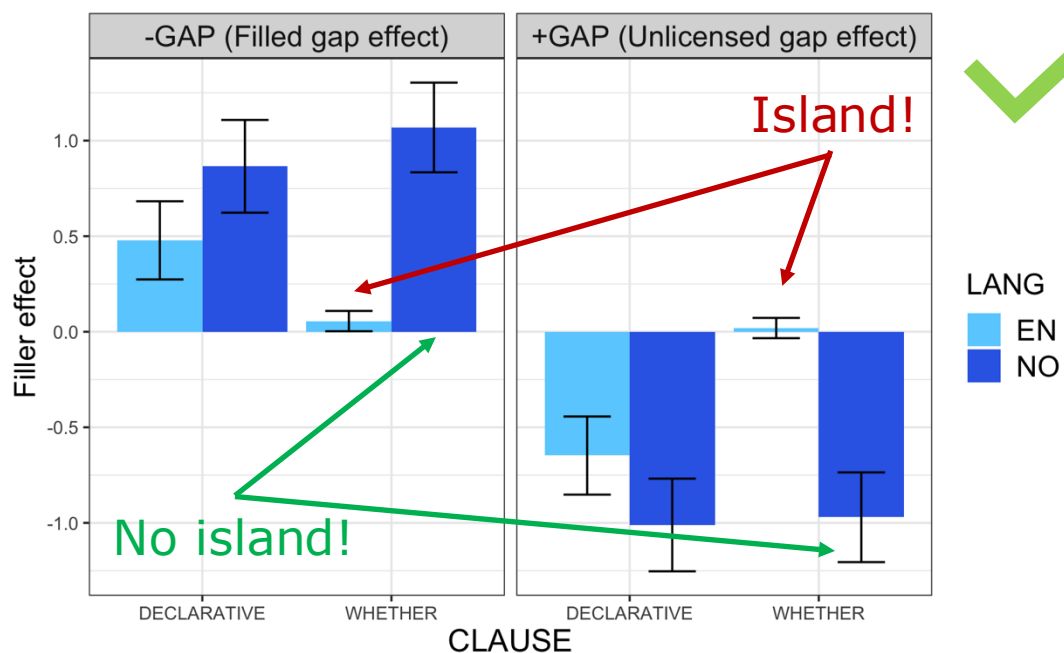
She knows **that** the priest revealed *the secret* at the party. -FILLER

*She knows **what** the priest revealed *the secret* at the party. +FILLER

- Trained LSTM RNNs and GPT-2 models on Norwegian Bokmål Wikipedia (113 million tokens)
- Controlled cross-linguistic comparison between English and Norwegian

NLM results

- The models could learn that EQs are islands in English but not in Norwegian and correctly predicted the island status of other domains (subject and adjunct phrases) but failed on some other cases
- ... like with overt complementizers (that) in multiply embedded clauses



Main findings I



1. Island status of different domains in Norwegian

- ✓ Both models learned that FGDs can span multiple clauses
- ✓ Correctly induced the island status of subject phrases
- ✓ Picked up some patterns of island-insensitivity pertaining to EQs
- ✗ The symbolic model could not induce the non-island status of adjuncts clauses and of EQs for wh-FGDs
- ✗ NLMs didn't learn the island status of English wh-clauses & showed complementizer-dependent generalization on unboundedness

Main findings II



2. Evidence found in the input to the models

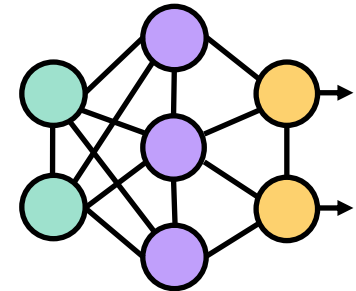
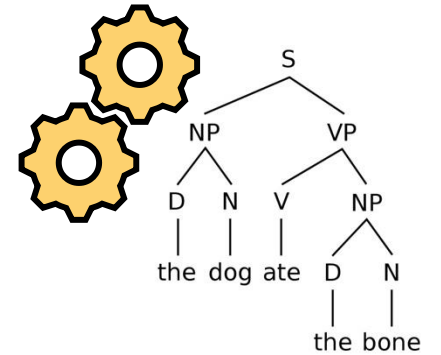
- ✓ Both training corpora contained examples of topicalization and relativization out of EQs – typologically marked structure
- ✓ Child directed-texts featured more such examples → conversational nature of the structure
- **Poverty of the stimulus is real:**
 - Island-crossing examples were limited and lexically restricted
 - Neither corpora contained examples of wh-FGDs into EQs
 - → Learners must generalize beyond the input

Main findings III



3. Model generalization capacity

- Symbolic model: generalization capacity depends on the granularity of the structured representations
 - Too granular distinctions → data sparsity, limited generalization
 - The decision to treat wh- and RC-FGDs separately
- NLMs could generalize across some constructions
 - Similar results across wh- and RC-dependencies
 - Occasional overgeneralization and undergeneralization



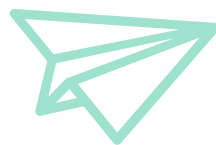
Results summary

- Some patterns of island-insensitivity are learnable from the input
 - Both model types succeeded on learning the (non-)island status of EQs
 - Trade-off between representational and statistical inference biases
- Both types of models showed limitations
 - To fully approximate human linguistic behavior, additional biases or model modifications might be necessary
- Modeling studies allow us to explore how learning is shaped by an interaction of input and biases & help refine our learning theories



Main take-aways

- Computational modeling is a powerful tool that can be used to understand linguistic development
 - If we have enough empirical data to base the models on
 - The results can help refine our theories and propose new ones
- The choice of a particular modeling approach depends on:
 - One's modeling goal
 - One's linguistic tradition
 - Prior research in the field



References

- Kush, D., Lohndal, T., & Sprouse, J. (2018). Investigating variation in island effects: A case study of Norwegian wh-extraction. *Natural Language & Linguistic Theory*, 36 (3), 743–779.
- Kush, D., Lohndal, T., & Sprouse, J. (2019). On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language*, 95 (3), 393–420.
- Kush, D., Sant, C., & Strætken, S. B. (2021). Learning Island-insensitivity from the Input: A Corpus Analysis of Child- and Youth-Directed Text in Norwegian. *Glossa: A journal of general linguistics*, 6 (1), 1–50.
- De Villiers, J., Roeper, T., Bland-Stewart, L., & Pearson, B. (2008). Answering hard questions: Wh-movement across dialects and disorder. *Applied Psycholinguistics*, 29 (1), 67–103.
- Phillips, C. (2013a). On the nature of island constraints II: Language learning and innateness. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 132–158). Cambridge University Press.
- Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN Language Models Learn about Filler-Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP* (pp. 211–221).
- Wilcox, E., Futrell, R., & Levy, R. (2023). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 1–44.
- Pearl, L., & Sprouse, J. (2013a). Computational models of acquisition for islands. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 109–131). Cambridge University Press.
- Pearl, L., & Sprouse, J. (2013b). Syntactic Islands and Learning Biases: Combining Experimental Syntax and Computational Modeling to Investigate the Language Acquisition Problem. *Language Acquisition*, 20 (1), 23–68.