Tom McCoy: Research statement (for a linguistics audience)

1 One-page summary of research agenda

What type of computational system is the mind? I approach this question from the perspective of language, spanning the divide between linguistics and natural language processing (NLP). I focus on two core topics: ***reconciling neural and symbolic computation*** and ***characterizing the learning biases that guide language acquisition***. My work on these topics spans across linguistic subfields including phonology, morphology, syntax, and semantics, and it also connects to areas of cognitive science outside of linguistics.

Neural vs. symbolic computation: For millennia, linguists have viewed language as a symbolic system, in which discrete units (e.g., words) are combined in structured ways (e.g., in syntax trees). Recently, however, NLP has seen tremendous progress with a very different type of system: neural networks. These models encode information in vectors of continuous numbers and process those vectors using mathematical operations. Though they seem poorly suited for language, neural networks are the state of the art for a range of linguistic tasks (e.g., machine translation), far outperforming approaches motivated by symbolic linguistic theories.

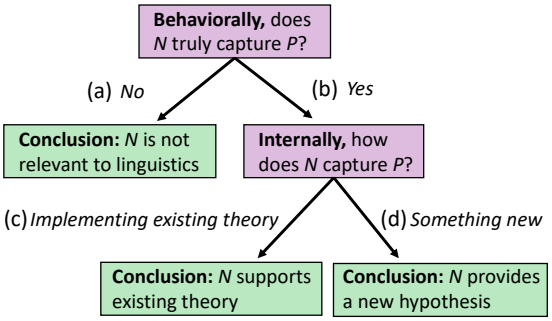

Figure 1: Understanding the relevance of neural network N for linguistic phenomenon P.

Figure 1 illustrates my approach for understanding whether and how neural networks should inform linguistics. We have shown (Section 2.1.ii) that some neural networks, despite their state-of-the-art status, fail to capture even the most basic aspects of compositional semantics (Fig. 1a; e.g., treating *the owl saw the fox* as synonymous with *the fox saw the owl*). However, other neural networks (Section 2.1.iii) display substantial linguistic abilities, generating novel syntactic and morphological combinations (Fig. 1b). Do these successes require us to revise linguistic theory to build in neural computation? We have analyzed neural networks trained to perform symbolic tasks (Section 2.2.i) and have shown that these models do not necessitate revisions to symbolic conceptions of language because their vector representations implicitly implement symbolic analyses of syntax (Fig. 1c).

Linguistic learning biases: Using human experiments and computational simulations, I study what learning biases guide the acquisition of structural properties of syntax and phonology, with plans to expand into morphology (Figure 2). In human experiments focusing on **recursion**, we have shown that people robustly extrapolate the recursive pattern of center embedding beyond the sentence sizes they have seen (Section 3.i). In ongoing work, we are studying which learning biases drive this generalization (e.g., a general simplicity bias, or more specific biases for **headedness** and **context-freeness**). In computational simulations, we have shown that generic neural network architectures fail to generalize in human-like ways for several syntactic phenomena, but neural networks built around **hierarchical structure** generalize correctly, suggesting that human-like generalization requires a hierarchical bias (Section 3.ii). We have also developed a new method that enables the creation of neural networks with targeted linguistic biases (Section 3.iii); we have applied this method to **syllabification** and plan to use it to evaluate hypotheses about the learning biases underlying acquisition of **word order** and **reduplication**. Through such experiments, we aim to investigate which formal properties of language are innate and which are learned. Using these insights about how people acquire language so rapidly, we also plan to reduce the data hunger of NLP models, enabling NLP to better handle languages that have little available training data.

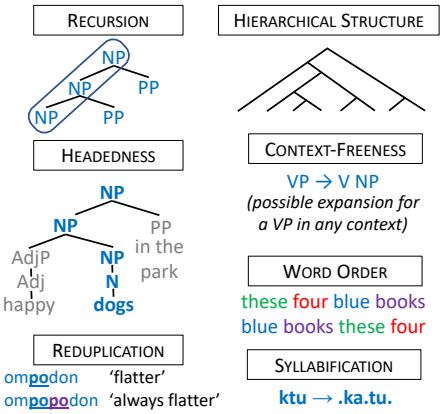

Figure 2: Some formal properties of linguistic structure.

2 Understanding the linguistic relevance of neural networks

In this section, I discuss in more detail the approach outlined in Figure 1.

2.1 Behaviorally, do neural networks capture the phenomena that linguists seek to explain?

i. Methodology: Targeted behavioral evaluation Neural networks perform well on the standard tests used by the engineering community, but these tests are formed from naturally-occurring corpora that might mainly contain "easy" examples that can be solved using shallow heuristics rather than by mastering language. To overcome the ambiguity of standard evaluations, I perform targeted, linguistically-motivated evaluations that reveal which linguistic phenomena a model has captured.

ii. Finding: Brittle heuristics In the task of natural language inference, a model must determine whether one sentence entails another. In principle, inference requires an understanding of syntax. The evaluation set that we created, HANS (McCoy, Pavlick, and Linzen 2019), tests whether inference models capture syntax, or if they instead use three shallow heuristics, such as assuming that sentence S entails any sentence whose words all appear in S (e.g., assuming that *the owl saw the fox* means the same thing as *the fox saw the owl*). Even BERT, a state-of-the-art model which scores close to humans on a standard evaluation, performs poorly on HANS (Figure 3), consistent with the hypothesis that it has adopted this heuristic. Thus, despite appearances to the contrary, these models are likely not relevant to linguistic theory (Figure 1a).

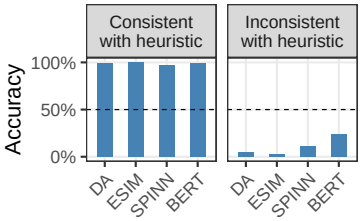

Figure 3: Inference models succeed on examples that can be solved with a shallow heuristic, but fail when attention to syntax is needed.

iii. Finding: Generative competence Current language models can generate grammatical, coherent text. It is unclear how they do so: do they have true generative abilities, or—as critics claim—are they simply copying from their training set? In McCoy, Smolensky, Linzen, Gao, & Celikyilmaz (2023), we analyze whether text generation models are overly reliant on copying. Here the conclusion is more positive than with HANS: on a variety of linguistic levels, models show an impressive degree of novelty. For instance, the model GPT-2 generated the sentence *The **Sarrats** were lucky to have her as part of their lives*, which includes a novel plural word (*Sarrats*) accompanied by the proper syntactic consequences of this word's plurality: a plural verb, *were*, and a plural coreferential pronoun, *their*. Thus, some neural networks have a non-trivial amount of generative competence (Figure 1b), motivating the focus of the next section: understanding *how* such models represent linguistic structure.

2.2 Internally, how do neural networks represent symbolic structure?

Our behavioral evaluations of novelty discussed above give clear evidence that neural networks can process language well; yet their representations are vectors of continuous values, which look very different from the symbolic structures used in linguistic theory. How do neural networks encode linguistic structure in vector space? The answer to this question determines how linguists should view the success of neural networks: Either these models are implementing—and therefore supporting—existing theories, or they have discovered a new approach to language that may require us to revise our theories to incorporate neural mechanisms.

i. Finding: Implicit symbolic structure Drawing on mathematical methods from cognitive science, we have developed a technique for testing the hypothesis that models' vector representations are implicitly symbol structures (McCoy, Linzen, Dunbar, & Smolensky 2018). This approach has yielded symbolic analyses that produce close approximations to models' vector representations. For instance, models that process words in linear order encode sequential positions (*first, second...*), while models that process words in accordance with a tree encode tree positions (*root, left child of root...*). We can use our analyses to make targeted interventions to modify a neural network's output (Soulos, McCoy, Linzen, & Smolensky 2020), verifying that the representational structure we have revealed is causally linked to model behavior.

In Lepori & McCoy (2020), we analyzed a neural network in which each word's representation is a single vector that can in principle encode any feature(s) of the word's context. We found that these representations encode linguistic dependencies. For instance, *himself* in (1a) encodes *politician*, while *him* in (1b) encodes *person*, showing that this model has implicitly learned to respect binding theory principles.

(1) a. The person believes that the **politician** loves **himself**.
b. The **person** believes that the politician loves **him**.

Such results show that, despite their apparent incompatibility with symbolic structure, at least some neural networks implicitly rediscover symbolic linguistic theories such as the basic binding theory principles. Thus, these models in fact *support* existing views of language as a symbolic system (Figure 1c). Though current work is still far from this goal, such insights from artificial neural networks might eventually help us understand how language is encoded in the biological neural network of the brain.

3 Linguistic inductive biases in humans and machines

The research discussed above focuses on characterizing what computational class language belongs to (that is, neural vs. symbolic computation). I also study the computational system deployed in language *acquisition*, focusing on inductive biases—the factors that guide how a learner learns from, and generalizes beyond, experience. A learner's inductive biases encompass both the characterization of the hypothesis space (e.g., a detailed notion of Universal Grammar, or a more general hypothesis space that has Merge as its only language-specific component) and the process used to search that space (e.g., a constraint ranking algorithm in Optimality Theory, or the numerical computations underlying the Tolerance Principle).

i. Finding: Extrapolation of recursion I study people's inductive biases using psychological experiments based on artificial language learning: train people on sentences from a specially-designed language, and then test how they generalize. This work has produced the first robust demonstration that people extrapolate the recursive syntactic pattern of center embedding beyond the sentence sizes they have seen (McCoy, Culbertson, Smolensky, & Legendre 2021), carefully avoiding the confounds that have arisen in prior work.

ii. Finding: Importance of a hierarchical inductive bias In Yedetore, Frank, Linzen, and McCoy (2023), we analyzed neural network models trained on utterances made by parents to their children in the CHILDES corpus. Using this corpus allowed us to bring our models into closer contact with linguistic questions, compared to previous models which were trained on corpora that were not representative of what children acquire language from (e.g., all of Wikipedia). We have used our CHILDES-trained models to study English polar question formation (e.g., turning *The dog can bark* into *Can the dog bark?*) because there are longstanding—but controversial—claims that strong innate biases are necessary to make this phenomenon learnable from the primary linguistic data that children receive. Both neural models that we tested (LSTMs and Transformers, two very different state-of-the-art architectures) do not have such strong biases, and they fail to learn the correct question-formation rule, bringing a new type of empirical evidence to support the poverty-of-the-stimulus argument that human language acquisition involves some strong syntactic biases.

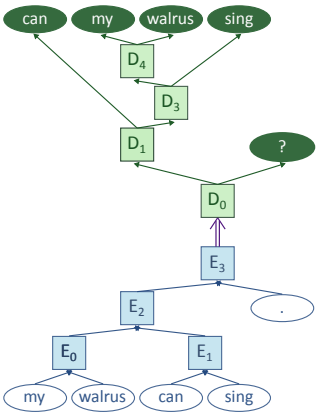

Figure 4: A tree-structured neural network.

How can we create models that better account for human-like generalization? Using synthetic datasets, we found that a hierarchical bias—the bias often hypothesized to underlie acquisition of English polar questions—can be imparted by using a tree-structured network (Figure 4), whose computations are guided by a syntax tree rather than following linear order (McCoy, Frank, & Linzen 2020). In later work, we showed that using tree-structured architectures also improves syntactic knowledge for models trained on natural text (Lepori, Linzen, & McCoy 2020). These results point toward structured architectures as a viable path for aligning models' biases with humans' biases.

iii. New method: Targeted inductive biases via meta-learning Structured architectures can impart some abstract inductive biases (e.g., formal universals), but they may not be flexible enough to impart all the biases that we wish to study (e.g., substantive universals, which are harder to capture in an abstract architectural structure). In McCoy, Grant, Smolensky, Griffiths, & Linzen (2020), we introduced a more flexible approach based on **meta-learning**: We first instantiate our desired biases as a distribution over synthetic languages. A neural network then meta-learns from languages sampled from this distribution to acquire our target biases; in meta-learning, exposure to many languages teaches a model about the commonalities across languages, enabling it to learn new languages more readily. We used this approach to impart biases that encode an account of syllabification from Optimality Theory. A model with inductive biases resulting from meta-learning learns syllabification patterns from only 200 examples, vs. 20,000 examples without such biases. These biases also improved accuracy from 6% to 88% on targeted linguistic evaluations.

More recently, in McCoy & Griffiths (2023), we scaled up this approach to a new setting where we distilled the syntactic priors of a Bayesian model into a neural network. The resulting system was able to learn syntactic patterns from a small number of examples. It was also able to learn aspects of English syntax from a naturalistic corpus, outperforming a standard neural network in several areas (such as extrapolating recursive syntactic phenomena). These results provide strong evidence that our approach can create neural networks with types of inductive biases traditionally described using symbolic linguistic theories.

The types of models that can be created with this method provide a new theory about the computational structures that underlie language learning: strong inductive biases instantiated in a flexible neural network system. Strong inductive biases account for rapid learning of linguistic patterns—one hallmark of language in humans. The flexibility of the neural network substrate accounts for the ability to learn successfully from unstructured naturalistic data—another key property of human language acquisition. Prior modeling approaches have captured one but not both of these capacities: for instance, Bayesian models can learn linguistic patterns from few examples but typically cannot learn tractably from large-scale naturalistic data, while neural network systems can learn effectively from naturalistic data but require many examples. This new meta-learning approach is the first approach that can capture *both* of these abilities, showing its promise as a way to combine the complementary strengths of neural and symbolic theories of learning.

4 Conclusion

I combine linguistics and NLP to study how to computationally characterize the language faculty. First, I analyze neural networks—which perform surprisingly well at NLP tasks, far outperforming linguistically-motivated models—to assess whether and how they capture linguistic structure. Second, I use computational modeling and experiments with human participants to investigate what inductive biases can explain human language acquisition. By bringing linguistics and NLP into closer contact, I aim to elucidate the linguistic relevance of advances in NLP, and to enable linguistics to better inform practical NLP technology.

References

Michael Lepori and R. Thomas McCoy. 2020. Picking BERT's brain: Analyzing contextual embeddings using representational similarity analysis. In *COLING 2020*.

Michael A. Lepori, Tal Linzen, and R. Thomas McCoy. 2020. Representations of syntax [MASK] useful: Effects of constituency and dependency structure in recursive LSTMs. *ACL 2020*.

R. Thomas McCoy, Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. 2021. Infinite use of finite means? Evaluating the generalization of center embedding learned from an artificial grammar. In *CogSci Conference 2021*.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *TACL*.

R. Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L. Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. In *CogSci Conference 2020*.

R. Thomas McCoy and Thomas L. Griffiths. 2023. Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. arXiv preprint.

R. Thomas McCoy, Tal Linzen, Ewan Dunbar, and Paul Smolensky. 2019. RNNs implicitly implement tensor-product representations. *ICLR 2019*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *ACL 2019*.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *TACL*.

Andrew Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*.

David E. Rumelhart, and James L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 2*.

Paul Soulos, R. Thomas McCoy, Tal Linzen, and Paul Smolensky. 2020. Uncovering the compositional structure of vector representations with Role Learning Networks. In *BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Paul Smolensky. 1990 Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. *ACL 2023*.