Tom McCoy: Research statement (for a computer science audience)

How can we create computational systems that learn language as rapidly and robustly as humans do? By studying this topic, I hope to both improve models used in AI and to advance linguistics by showing which computational mechanisms underlie humans' linguistic abilities. My research has three main components. First is **rethinking how we measure progress in natural language processing (NLP)**. The prevalent approach is to score a model on a test set that is similar to the model's training set, but improvements on such benchmarks do not always indicate meaningful progress because it is unclear what drives the improvements: enhanced linguistic abilities or brittle, dataset-specific heuristics. I instead develop hypothesis-driven techniques that allow us to *understand* NLP models, illuminating how they perform so well and diagnosing areas where they still fall short, such as requiring far more data than humans do. My second focus is **understanding the linguistic inductive biases of humans**, the factors that guide learning and enable rapid generalization. That is, what strategies do people use to learn so much more effectively than current models in AI? My final focus—which is informed by insights from the first two—is **improving the inductive biases of NLP models** so that they can learn more quickly and generalize better. In particular, I study how linguistically-motivated architectures and training techniques can improve models' generalization abilities.

1 Rethinking how we measure progress in NLP: Hypothesis-driven evaluation and analysis

i. Hypothesis-driven evaluation: Standard NLP test sets are sampled from the same distribution as the training set. This approach has a major flaw: a model can score well on the test set by learning heuristics that succeed for frequent types of examples but fail on rarer cases, meaning that the model has not actually solved the intended training task.

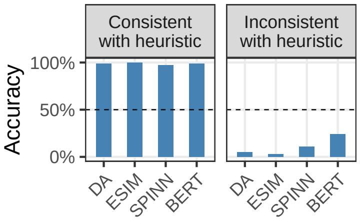

Figure 1: Inference models succeed on examples that can be solved with a shallow heuristic, but fail when attention to syntax is needed.

One focus of my research is addressing this possibility through a different evaluation paradigm: hypothesis-driven evaluations that illuminate what strategies models are using. We applied this approach to the task of natural language inference (NLI) to create the HANS dataset (McCoy, Pavlick, & Linzen 2019), which evaluates whether NLI models have adopted three syntactic heuristics that we hypothesized they were likely to adopt, 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 a state-of-the-art model (BERT) performs poorly on HANS (Figure 1), consistent with the hypothesis that BERT has adopted this heuristic.

A major benefit of this approach is that, when a model fails on HANS, we have a clear hypothesis about what it is doing wrong, which can guide further work aimed at improving the model. For instance, we built from this hypothesis to create additional training examples that emphasize syntax. Adding these examples to the training set substantially decreased the model's reliance on heuristics that ignore word order (Min, McCoy, Das, Pitler, & Linzen 2020). Many other groups have also used HANS to motivate approaches for counteracting spurious heuristics, illustrating the impact that a carefully-designed evaluation can have.

Another area for which we have performed hypothesis-driven evaluation is natural language generation (NLG). Current language models (e.g., GPT-4) can generate coherent, grammatical passages of text. However, it is unclear how they achieve this success: do they have true generative abilities, or—as critics claim—are they simply copying text from their training set? In McCoy, Smolensky, Linzen, Gao, & Celikyilmaz (2023), we analyze whether NLG 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*. Such examples show that some neural networks have a non-trivial amount of generative competence.

ii. Analyzing vector representations of symbolic structure: It has long been assumed that processing language requires representations of symbolic structure. Behavioral evaluations—such as our evaluations of novelty discussed in Section 1.i—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 within vector space?

Drawing on mathematical methods from cognitive science, we have developed a technique for analyzing models' vector representations as implicit encodings of symbol structures (McCoy, Linzen, Dunbar, & Smolensky 2018). The resulting analyses are much more interpretable than the original vectors. For instance, we showed that models trained to copy sequences encode position counting from left to right (*first, second...*); but models trained to reverse sequences encode position counting from right to left (*last, second-to-last...*). Analysis of more complex models has revealed more convoluted representational schemes, such as a scheme that includes the position *second-to-last word in the first of two clauses joined by a conjunction*.

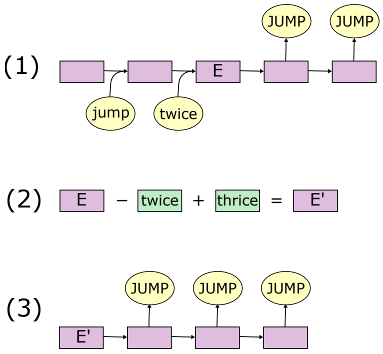

Figure 2: (1): Model being analyzed, which encodes the input (*jump twice*) as vector E then outputs *JUMP JUMP*. (2) Subtracting our analysis's predicted vector for *twice* and adding our analysis's predicted vector for *thrice* turns E into a new vector, E'. (3) Behavioral result of replacing E with E'.

Our analysis provides a closed-form equation for approximating a model's internal vector representations. This equation gives such a close approximation that we can use it to make targeted representational interventions to modify a model's behavior (Figure 2; Soulos, McCoy, Linzen, & Smolensky 2020), verifying that the representational structure we have revealed is causally linked to model behavior.

2 Understanding linguistic inductive biases in humans

Children can learn language from a surprisingly small amount of data, and they readily generalize what they learn to novel situations and novel linguistic structures. In contrast, current neural models require far more training data and also generalize much less robustly. How do humans learn so rapidly and robustly, and how can we replicate those strengths in machines? These questions come down to the topic of *inductive biases*, which are the factors that guide how learners learn and generalize.

I study people's inductive biases using the paradigm of artificial language learning: teach people a specially-designed language and then test how they generalize it. This work has produced the first demonstration that people robustly extrapolate the recursive syntactic pattern of center embedding beyond the sentence sizes they have seen (McCoy, Culbertson, Smolensky, & Legendre 2021). For an ongoing second experiment, we have enriched a Bayesian model of language acquisition (Perfors, Tenenbaum, & Regier 2011) to test whether the bias we have observed specifically favors grammars generating unboundedly large syntactic structures, or more generally just favors simpler grammars. Our aim in such experiments is to characterize people's inductive biases precisely enough to build those biases into computational models.

I also use neural networks as cognitive models to study what types of networks show the most human-like learning behavior. For instance, in Yedetore, Linzen, Frank, & McCoy (2023), we analyzed neural network models trained on the CHILDES corpus, which contains utterances made by parents to their children. Using this corpus allows us to bring our models into closer contact with cognitive questions, compared to existing models which are trained on corpora that are not representative of what children acquire language from (e.g., all of Wikipedia).

3 Improving models' inductive biases

The investigations of humans' biases discussed above give insight into biases that could be useful for models to have as well, but how can we give these biases to a model?

i. Meta-learning from synthetic languages: In McCoy, Grant, Smolensky, Griffiths, & Linzen (2020), we developed an approach for using *meta-learning* to give targeted inductive biases to a model. We first instantiate our desired biases as a distribution over synthetic languages. The model 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. This approach enabled our model to learn linguistic mappings from only 200 examples, vs. 20,000 examples without meta-learning. Meta-learning also enabled robust generalization, yielding 88% accuracy on out-of-distribution tests, vs. 6% without meta-learning.

More recently, in McCoy & Griffiths (2023), we have extended this method to a more general setting, where we used it to distill the syntactic priors of a Bayesian model into a neural network. Like a Bayesian model, this system can learn many syntactic patterns from a small number of examples; like a neural network, it can also learn aspect of English syntax from naturalistic data. This approach thereby bridges the divide between two successful yet distinct research traditions (Bayesian approaches and neural networks), yielding a single system that displays the complementary strengths of both approaches.

ii. Architectures with built-in linguistic structure: I have studied how we can give models a hierarchical inductive bias—a bias for generalizing based on hierarchical syntactic structure rather than linear order—which is a bias long argued to play a role in human language acquisition. Using synthetic datasets, we found that we could robustly impart this bias through use of a structured model whose computations are guided by syntax trees, but other approaches (multi-task learning and syntactic annotation of input data) had only minor effects (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 improving models' inductive biases.

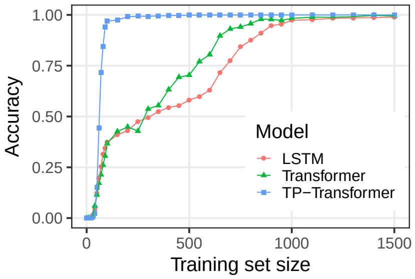

Figure 3: The TP-Transformer (argued for in our position piece) learns a symbolic task much more rapidly than standard NLP architectures.

In a position piece (Smolensky, McCoy, Fernandez, Goldrick, & Gao 2022), we argue for incorporating another type of structure into models—namely, a structure that disentangles contentful information (e.g., the identity of a word) from positional information (e.g., where the word appears). Collaborators at Microsoft have implemented the type of architecture that we argue for and have shown its strength on a variety of tasks. My own experiments have shown the rapid symbolic generalization that this architecture enables (Figure 3).

4 Conclusion

I combine cognitive science and machine learning to analyze models of language and then use insights from those analyses to improve how models learn and generalize. Throughout these different threads of research, I focus on models' inductive biases and internal representations of linguistic structure. My long-term goal is to deepen our understanding of these two factors, allowing us to replicate the fast, robust learning capabilities that still set humans apart from even our most impressive computational models.