Universal linguistic inductive biases via meta-learning

R. Thomas McCoy,[1] Erin Grant,[2] Paul Smolensky,[3,1] Thomas L. Griffiths,[4] and Tal Linzen[1]
tom.mccoy@jhu.edu, eringrant@berkeley.edu, smolensky@jhu.edu, tomg@princeton.edu, tal.linzen@jhu.edu
[1]Department of Cognitive Science, Johns Hopkins University
[2]Department of Electrical Engineering & Computer Sciences, University of California, Berkeley
[3]Microsoft Research AI, Redmond, WA USA
[4]Departments of Psychology and Computer Science, Princeton University

Abstract

How do learners acquire languages from the limited data available to them? This process must involve some inductive biases—factors that affect how a learner generalizes—but it is unclear which inductive biases can explain observed patterns in language acquisition. To facilitate computational modeling aimed at addressing this question, we introduce a framework for giving particular linguistic inductive biases to a neural network model; such a model can then be used to empirically explore the effects of those inductive biases. This framework disentangles universal inductive biases, which are encoded in the initial values of a neural network's parameters, from non-universal factors, which the neural network must learn from data in a given language. The initial state that encodes the inductive biases is found with meta-learning, a technique through which a model discovers how to acquire new languages more easily via exposure to many possible languages. By controlling the properties of the languages that are used during meta-learning, we can control the inductive biases that meta-learning imparts. We demonstrate this framework with a case study based on syllable structure. First, we specify the inductive biases that we intend to give our model, and then we translate those inductive biases into a space of languages from which a model can meta-learn. Finally, using existing analysis techniques, we verify that our approach has imparted the linguistic inductive biases that it was intended to impart.

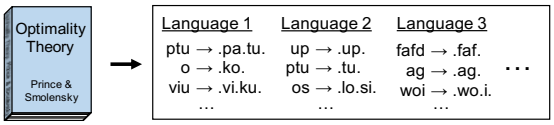
Keywords: meta-learning, inductive bias, language universals, syllable structure typology, neural networks

Introduction

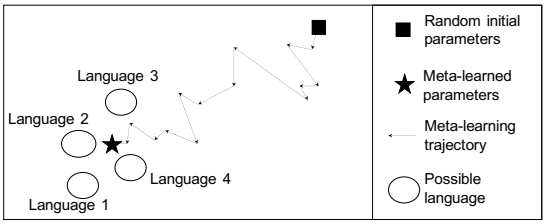
Human learners can acquire any of the world's languages from finite data. The acquisition of a particular language involves two factors: data from that language, and the learner's **inductive biases**, which are the factors that determine how the learner will generalize beyond the particular utterances in the data (Mitchell, 1997). Many inductive biases are shared by all humans (e.g., because of shared brain anatomy or shared communicative goals), so these biases exert universal pressures on language acquisition. A central task of linguistics is to determine which inductive biases affect language acquisition and how those biases interact with learning to yield a learner's linguistic knowledge.[1]

There are two major approaches for modeling the interplay between universal inductive biases and learning. One approach, probabilistic modeling, typically follows a top-down methodology that commits to a representation and an inference algorithm. Such strong commitments allow targeted investigations (e.g., Perfors, Tenenbaum, and Regier, 2011) but can be too restrictive, making it difficult for these models to represent all possible languages. In contrast, neural network modeling takes a bottom-up, data-driven approach that gives greater flexibility in representing the full span of languages. In recent work, neural networks trained on naturally-occurring data have shown success at learning linguistic phenomena such as subject-verb agreement (Gulordava et al., 2018). Since neural networks do not have overt biases specific to language, their successes give insight into which aspects of language are learnable from realistic input paired with domain-general biases. However, these models often generalize in different ways from humans (McCoy et al., 2019), and they require far more training data than humans, indicating that their learning is underconstrained (van Schijndel et al., 2019). To address these problems, it would be necessary to give these models additional inductive biases that would appropriately constrain their learning to be more human-like, but their bottom-up nature makes it difficult to build in additional biases (Griffiths et al., 2010).

Step 1: Translate a desired set of inductive biases into a space of languages.



Step 2: Have a model "meta-learn" from these languages to find a parameter initialization from which the model can acquire any language in the space.



Step 3: Verify that meta-learning has imparted the desired inductive biases by training the model on analysis datasets.

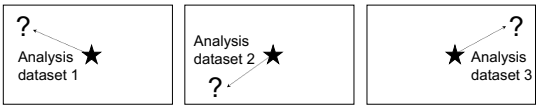


Figure 1: Summary of our approach. The steps shown above give a pre-specified set of inductive biases to a model; alternately, Step 1 could be skipped by having the model meta-learn from existing languages, in which case the approach would discover a set of biases sufficient to acquire those existing languages rather than imparting a pre-specified set of biases (see Figure 8).

[1]Though the term *inductive biases* often refers to cognitive biases, we use it to encompass all pressures that shape the language that a learner learns; see Figure 2 and the Background section.

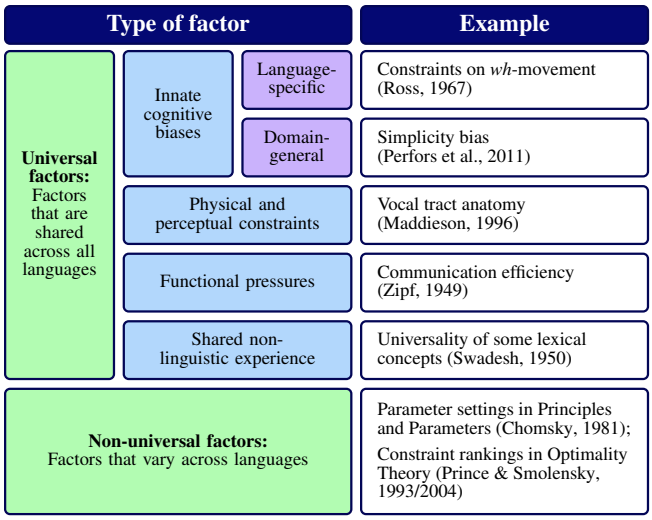


Figure 2: Factors that shape languages and hypothesized examples.

In this work, we propose a computational modeling framework for imparting a hypothesized set of universal linguistic inductive biases in a way that is compatible with the flexibility of neural networks. Our approach is based on **meta-learning**, a technique in which a learner is exposed to a variety of tasks, each of which comes with a limited amount of data (Thrun & Pratt, 1998; Hochreiter et al., 2001). This process instills in the learner a set of inductive biases which allow it to learn tasks similar to those it has seen before from limited data. In our setting, each "task" is a different language, and the inductive biases that result from meta-learning are encoded in a neural network's initial state. This initial state is found in a data-driven manner; by controlling the data, we can influence which inductive biases will be encoded in the initial state, and the initial state can then be analyzed to verify that it encodes the universal inductive biases that it is intended to encode.

As a first case study, we show the effectiveness of this approach on the acquisition of a language's syllable structure, a paradigmatic example of universal linguistic inductive biases. We define a set of inductive biases relating to syllable structure that we intend to give our model, and we then translate this set of inductive biases into a space of possible languages from which we have a model meta-learn. Through analysis of the meta-learned initial state, we verify that meta-learning has successfully imparted the inductive biases that it was intended to impart; for example, the model has meta-learned that the presence of certain input-output mappings in a language implies the presence of other input-output mappings.[2]

Background

Universal linguistic inductive biases Evidence for universal inductive biases that shape language acquisition primarily comes from two areas. First, in typology (the taxonomy of observed language types), certain grammatical structures are much more common than others even across unrelated languages (Greenberg, 1963), and at least some of these patterns appear to arise from learners' inductive biases (Culbertson et al., 2012). Second, in acquisition, the argument from the poverty of the stimulus (Chomsky, 1980) notes that all language learners generalize in similar ways despite being faced with stimuli that are consistent with multiple generalizations. We use the phrase **universal linguistic inductive biases** for any pressures that universally affect acquisition, including the types of innate, language-specific constraints sometimes termed *Universal Grammar*, as well as other influences such as articulatory or information-maximizing considerations; see Figure 2 for a categorization of universal pressures. We group these factors together because our framework could be used to impart any type of inductive bias regardless of what source that bias might have in the real world.

Several linguistic formalisms provide theories of the universal/non-universal distinction. In the Principles and Parameters framework (Chomsky, 1981), universal principles interact with non-universal parameter settings; in Optimality Theory (Prince & Smolensky, 1993/2004), a universal set of constraints interacts with non-universal rankings of these constraints. In contrast, our approach does not require any formal characterization of universal or non-universal factors; instead, due to the data-driven nature of the approach, these factors are characterized purely in terms of the behaviors they would lead to when paired with particular types of training data. If a formal characterization of the model's inductive biases is desired, it must come from an analysis of the trained model, because the meta-learning process itself does not provide a formal characterization of the biases it imparts.

Learning and meta-learning The models we use are artificial neural networks, which are governed by a large number of numerical parameters such as connection weights. At the core of our approach are two processes for determining the values of those parameters: **standard training** and **meta-training**. Standard training iteratively minimizes error within a single training language: the model starts with a particular set of initial values for its parameters and is exposed to a training set of example input-output pairs from the language to be learned. For each training example, the output that the model generates is compared to the target value, and the model's parameters are adjusted to decrease the difference between the predicted output and the correct target. Ideally, after many such updates, the model will perform well not only on its training set but also on a test set, which contains unseen examples drawn from the same language as the training set. Standard training requires the model to begin with some initial parameter values; it is the task of meta-training to set these initial values based on data.

Standard training requires only a single language. Meta-training, by contrast, samples multiple languages from a distribution of possible languages, $p(L)$. The particular form of meta-learning that we use is *model-agnostic meta-learning* (MAML; Finn, Abbeel, & Levine, 2017): The

[2]Our code is at `https://github.com/tommccoy1/meta-learning-linguistic-biases`; there is also a demo at `http://rtmccoy.com/meta-learning-linguistic-biases.html`.

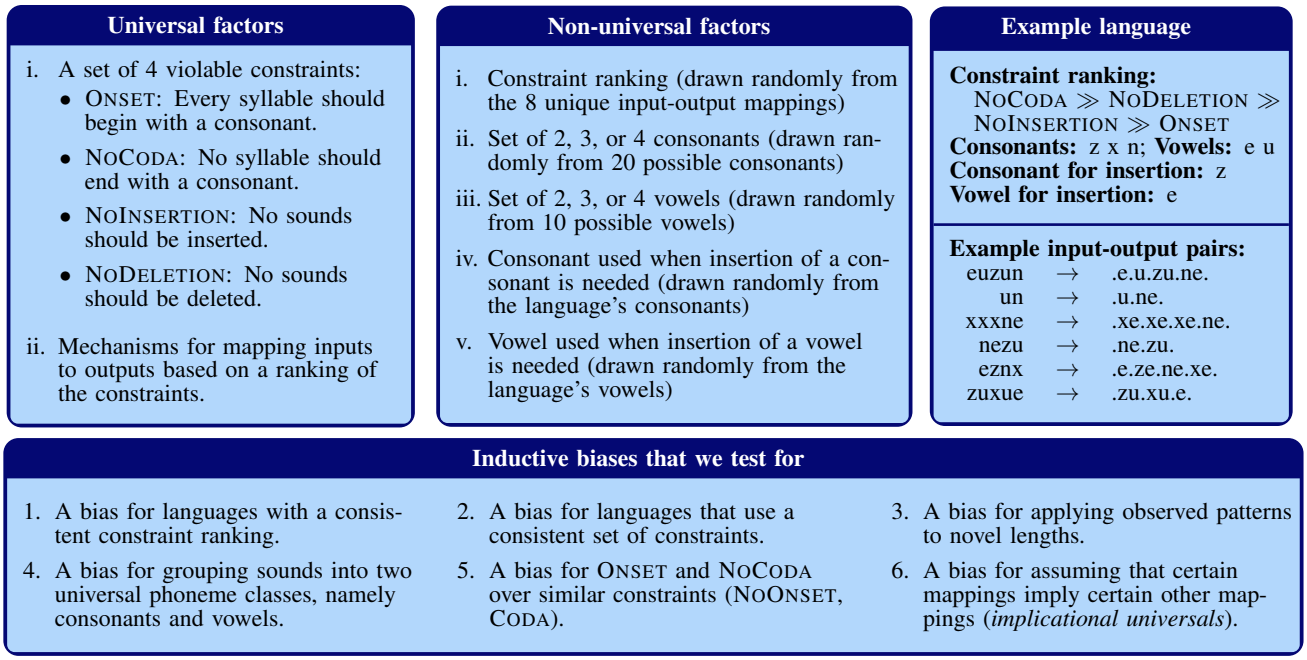


Figure 3: A summary of basic syllable structure theory in Optimality Theory (Prince & Smolensky, 1993/2004). The top middle panel posits 8 unique input-output mappings because many of the 24 orderings of the 4 constraints are equivalent in the outputs they produce. The top right panel gives an example language; the chosen constraint ranking leads to a language where no syllables end in a consonant, and where violations of this restriction are fixed by vowel insertion rather than consonant deletion. Periods (present in the output but not the input) indicate syllable boundaries. The bottom panel lists the inductive biases that we use as behavioral tests of the universal factors.

model's initial state, M_0, is determined by a set of initial parameter values; then, for each sampled language $L_i \sim p(L)$, we train our model on the training set of L_i using standard training, to yield a trained model M_i. Crucially, we then compute an adjustment of the initial state M_0 using M_i's loss on the unseen test examples from L_i; M_i is discarded after the adjustment of M_0. Intuitively, we tweak M_0 so that, if we were to train the model on L_i again, it would learn L_i in fewer iterations. As meta-training proceeds, the initial model state M_0 (the square in Figure 1, Step 2) moves to a point from which it can readily learn any language from the distribution of meta-training languages (the star in Figure 1, Step 2). We hypothesize that, if we construct $p(L)$ to encode the inductive biases that we wish our model to have, then meta-learning that aims to facilitate acquisition of languages in $p(L)$ will give a model these inductive biases.

Overview of the approach

The goal of our approach is to give a model a set of inductive biases hypothesized to be relevant for human cognition; once such a model has been created, it could then be used to empirically investigate the effects that those inductive biases have. The following sections walk through our 3-step approach, using a case study in syllable structure typology.

Step 1: Defining the space of learning problems

To apply our method, we must first define the inductive biases that we wish to impart; that is, we must define what innate knowledge we wish our model to have. For this purpose, we focus on the domain of syllable structure (the study of how words in a language are divided into syllables; Jakobson, 1962), and in particular we adopt the Optimality Theory account of Prince and Smolensky (1993/2004),[3] because this account provides a clear characterization of which factors are universal and which are non-universal.

In this account, each word has an input form, which is the form it takes before any phonological processes have applied, and an output form, which is the result of phonological processes acting on the input. For example, in English, the prefix *in-* combines with the word *possible* to create the input *inpossible*, which is then mapped to the output *impossible* through a place-assimilation process. The input-output mapping is determined by a ranked set of constraints, where the set of constraints is universal, but their ranking is non-universal. For syllable structure, we use four constraints. Two of them evaluate the output alone: ONSET favors output syllables that begin with a consonant, and NOCODA favors output syllables that end with a vowel.[4] If the input does not satisfy these constraints (e.g., *kep*), the output could be made to satisfy them by either inserting or deleting phonemes (e.g., *.ke.pa.* or *.ke.*).[5] However, insertions and deletions are discouraged by the remaining constraints, NOINSERTION and NODELETION. When two constraints conflict with each other, the con-

[3] We use the simplified account from Sec. 6.1 of Prince and Smolensky (1993/2004), leaving out the subsequent refinements.

[4] A syllable's **onset** and **coda** consist of, respectively, syllable-initial and syllable-final consonants (e.g., for *kep*, *k* and *p*, resp.).

[5] We use periods (".") to indicate word and syllable boundaries.

flict is resolved by a priority-ranking of the constraints; this ranking differs across languages. In a language where NOINSERTION and NODELETION outrank NOCODA, the input *kep* would map to the output *.kep.*, because NOCODA cannot be satisfied without violating a higher-ranked constraint. Under other rankings, the input *kep* could map to *.ke.* or *.ke.pa.*.

Based on this framework, we defined a set of inductive biases that we intend to give to our model via meta-learning (Figure 3, bottom panel). These biases were chosen to provide behaviorally-defined versions of the universal factors in the Optimality Theory framework. For example, this framework includes a universal mechanism for mapping inputs to outputs based on a constraint ranking, a mechanism which could not be directly observed in our model's behavior. Thus, we instead defined inductive biases that encode properties of this mechanism, such as a bias for languages with a consistent constraint ranking, to encode the fact that the mechanism employs a consistent ranking within each language.[6] We then translated these inductive biases into a space of possible languages and used that space to sample languages for use in meta-learning (Figure 3, top middle and top right panels).

Step 2: Meta-training

The next step is to train a meta-learner on the set of learning problems defined in Step 1. In our case study, the initial state M_0 and the language-specific state M_i are the parameters of a sequence-to-sequence neural network (Sutskever et al., 2014) that maps a word's input form to a predicted output form. This architecture has two components: the **encoder** is fed the input one phoneme at a time and outputs a vector that encodes the entire input; this vector encoding is fed to the **decoder**, which generates the output symbols one at a time, ending with a special end-of-sequence token.[7] We apply meta-learning to such a model, allowing it to meta-learn from a set of 20,000 unique languages (called the *meta-training set*). For each language, the model was trained on 100 examples from that language and then tested on 100 held-out examples; the model's meta-training objective is thus learning to perform *100-shot learning*, that is, acquiring the ability to learn a new language from only 100 examples. After every set of 100 meta-training languages, we evaluated how well the model could perform 100-shot learning on each of 500 held-out languages; we terminated meta-training when there had been 10 consecutive evaluations without improvement, and then evaluated the meta-trained model on its ability to perform 100-shot learning on a final set of 1,000 held-out languages called the *meta-test set*. Performance on 100-shot learning was measured as the proportion of inputs in a language's test set for which the model generated an exactly correct output sequence of phonemes and syllable boundaries after observing 100 examples from the language's training set. We compared the model whose parameters were initialized using meta-learning to a baseline model whose parameters were initialized randomly; aside from initialization, both models use the same learning procedure to learn each language.

Meta-learning results The model with meta-learned initial parameters had an average 100-shot accuracy (i.e., the accuracy after exposure to 100 examples) of 98.8% on the languages in the meta-test set. By contrast, the 100-shot accuracy for a randomly-initialized model was only 6.5%. In this case study of syllable structure typology, then, meta-learning succeeded at imparting the ability to learn languages in our distribution of languages from a small number of examples. To evaluate whether meta-learning imparted the specific inductive biases that we intended it to impart (Figure 3), we next analyzed the weight initialization found through meta-learning by examining the learning behavior it produces.

Step 3: Verification of the acquired inductive bias

Ease of learning Our first approach for studying our model's inductive biases is to evaluate how easily they learn languages that differ from each other in controlled ways. We quantify ease of learning as the minimum number of training examples that a model needs from a language to reach 95% accuracy on that language's test set.[8]

We first use this technique to test whether the meta-learned inductive bias produces learning behavior that favors the **set of constraints** defining our syllable structure typology. Recall that our space of languages was defined with the constraints ONSET and NOCODA. We now test our model on languages defined by these constraints, as well as languages defined by alternate constraint sets in which ONSET is replaced with NOONSET, or NOCODA with CODA, or both.

Across language types, the model initialized with meta-learning required far fewer examples than the randomly-initialized model (Figure 4, top). Importantly, though, meta-learning did not improve performance equally across languages: The ONSET/NOCODA languages were 5.6 times easier to learn than languages defined by other constraints for the model initialized with meta-learning, compared to 1.2 times easier for the random model (Figure 6a), suggesting that meta-learning has imparted an inductive bias favoring languages that are consistent with the meta-training constraints.

Has the model initialized with meta-learning simply memorized the types of languages it has seen, rather than learning the more abstract constraints of ONSET and CODA? Figure 5 suggests that meta-learning has imparted some de-

[6]Not all of these biases are necessarily present in humans; e.g., languages differ in which phonemes can be syllabic nuclei, whereas one of our target biases is knowledge of a universal class of potential nuclei (i.e., the vowels). We do not intend to propose a theory of syllable structure acquisition but rather to demonstrate how meta-learning could instantiate such a theory in a neural network.

[7]The particular model that we used was an LSTM (Hochreiter & Schmidhuber, 1997) with a single hidden layer of size 256, an embedding layer of dimension 10 (which learned distributed representations of phonemes), and no attention. The inner loop optimization of MAML used stochastic gradient descent with a learning rate of 1.0 and batch size 100, while the outer loop optimization used Adam with a learning rate of 0.001 (Kingma & Ba, 2015).

[8]Specifically, we selected the number of examples to be the smallest multiple of 100 for which the model converged to at least 95% accuracy, without restricting the number of training iterations.

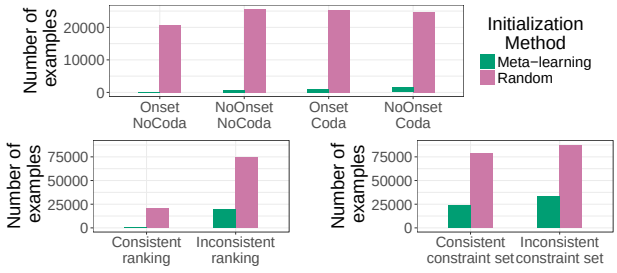


Figure 4: The number of examples needed to learn a language to 95% accuracy (lower is better). Each bar is an average of 80 to 100 languages. Meta-learning improves performance on all conditions.

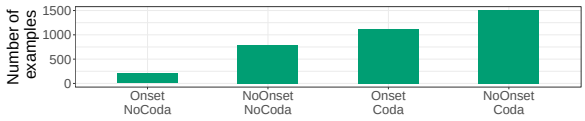


Figure 5: Data from the top panel of Figure 4 re-plotted at a different scale: The number of examples needed by the model initialized with meta-learning to learn languages with different sets of constraints.

gree of more abstract knowledge, because, of the types of languages that were not present during meta-learning, the model has an easier time learning ones that have one of the correct constraints (i.e., NOONSET/NOCODA languages and ONSET/CODA languages) than ones that have neither correct constraint (i.e., NOONSET/CODA languages).

We now test whether the model has meta-learned that **there must be a consistent constraint ranking within a language**. We test our models on languages governed by the constraints used during meta-learning (ONSET, NOCODA, NOINSERTION, and NODELETION), but with no consistent ranking of constraints within the language. This is done by independently choosing a random constraint ranking for each input structure; e.g., we might select a ranking for VV such that any input of the form VV maps to .CV.CV., while VVV might receive a ranking that maps VVV inputs to the empty string.[9]

For the model initialized with meta-learning, languages with a consistent ranking were 97.1 times easier to learn than languages without a consistent ranking, compared to only 3.6 times easier for the randomly initialized model (Figure 6b). This improvement in learning relative to random initialization suggests that meta-learning has strengthened the model's bias for languages generated by a consistent constraint ranking.

We next test whether our models have a bias for the fact that **within a language, a single set of constraints can consistently generate all input-output mappings** (the previous test was about the constraint *ranking*, while this one is about the constraint *set*). We evaluate our models on languages with a consistent set of constraints but no consistent ranking across inputs, as in the previous experiment; but now we allow the set of constraints generating a given language to include any of the output constraint combinations discussed above (ONSET/CODA, ONSET/NOCODA, NOONSET/CODA,

[9]We use C and V as shorthands for *consonant* and *vowel*; CV is shorthand for *any syllable of the form consonant-vowel*.

(a) Ratios comparing languages with an incorrect set of constraints to those with ONSET and NOCODA.

(b) Ratios comparing languages with an inconsistent constraint ranking to those with a consistent ranking.

(c) Ratios comparing languages with an inconsistent set of constraints to those with a consistent set.

Figure 6: Each subplot shows the ratio of the average number of examples need to learn a language with a property inconsistent with typology to the average number needed to learn a language with the analogous consistent property; higher is better. In each case, the model initialized with meta-learning favors the typologically consistent language type more strongly than does the randomly initialized model. The ratios derive from the results shown in Figure 4.

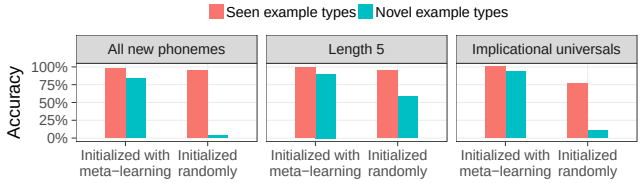


Figure 7: Results on poverty-of-the-stimulus experiments. Both models perform well on the example categories they have seen before, but the performance of the randomly-initialized model plummets when it is tested on novel types of examples; the meta-trained model exhibits less of a performance drop in these cases.

or NOONSET/NOCODA). We compare the learning of such languages to languages with no consistent set of constraints (and also no consistent constraint ranking), such that, for each input template (e.g., CCVC), there is a randomly-selected set of constraints and a random ranking for those constraints.

On average, the languages with a consistent constraint set were 1.4 times easier to learn for the model initialized with meta-learning than languages without a consistent constraint set, compared to 1.1 times easier for the randomly initialized model (Figure 6c). This result suggests that meta-learning has moderately strengthened the model's bias favoring languages that can be generated by a single set of constraints. For such constraint-set consistency to greatly increase the learnability of a language, it appears necessary that the language also be generated by a single constraint ranking.

Poverty of the stimulus As a second way to study biases, we use a poverty-of-the-stimulus approach: for a given language, we train on a dataset lacking a certain *class* of examples, and test generalization to the withheld class.

We performed three such experiments. In the **all new phonemes** setting, the training set for each language only contained 2 to 4 unique consonants and 2 to 4 unique vowels, as before, but now every example in the test set consisted entirely of consonants and vowels that were not present in the language's training set. The randomly initialized model has no hope of succeeding in this case, as it has no way to know whether each novel character is a consonant or vowel, but the model initialized with meta-learning could have learned these distinctions during meta-learning because the division between consonants and vowels is consistent across the meta-training languages. The model initialized with meta-learning

performs strongly here (Figure 7, left), suggesting that meta-learning has imparted universal consonant/vowel classes.

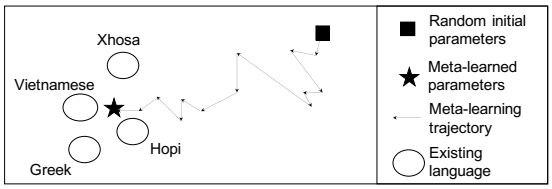
In the **length 5** setting, each language's training set only contained examples with an input length of at most 4, but its test set examples were all of length 5. The model initialized with meta-learning also performed strongly here (Figure 7, middle). Note that, during meta-learning, inputs with lengths up to 5 appeared; thus, the length 5 setting only requires generalization within the bounds seen during meta-learning. We also tested how models generalized from lengths at most 5 to length 6; in this case, the model initialized with meta-learning only achieved 59% accuracy. This suggests that meta-learning imparts a bias favoring languages in which the types of mappings that apply to short strings also apply to longer strings, but that this bias is restricted to the lengths present during meta-learning.

Last is the **implicational universals** setting. The space of languages that we have defined is restricted such that the presence of certain input-output mappings implies the presence of certain other input-output mappings. For example, the presence of the mapping VC $\rightarrow$ .CVC. indicates that the language will insert a consonant at the start of any syllable that does not start with one, which means that the language will also have the mapping V $\rightarrow$ .CV. To test whether a model has an inductive bias for this association, we can train it solely on examples of the form VC $\rightarrow$ .CVC. and then see how it handles V inputs; a naive model is unlikely to know how to handle this input, while a model that knows the implication would know to transform V to .CV. Our space of languages (Figure 3) predicts 24 dependencies of this form; when we test all of these dependencies, we find that the randomly-initialized model performs poorly while the model initialized with meta-learning performs well (Figure 7, right). This suggests that the model has meta-learned these implicational universals.

Conclusion

We have demonstrated how meta-learning can impart universal inductive biases specified by the modeler. This example-based approach to imparting inductive biases does not require an explicit theory of the biases in question; rather, imparting the biases only requires these biases to be translated into a distribution of possible languages. While the meta-learned biases are not as transparent as those encoded in probabilistic symbolic models, analysis of the model's learning behavior can be used to evaluate whether meta-learning has produced the desired biases, as we have shown. In our case study, we found evidence that meta-learning had successfully imparted all of our target inductive biases (or strengthened them, in cases where the biases were already present), including both some abstract biases (e.g., a bias for languages with a consistent constraint ranking) and some more concrete biases (e.g., a bias for treating certain phonemes as vowels). These results show that linguistic inductive biases that have previously been framed in symbolic terms can be reformulated in the context of neural networks, facilitating cognitive modeling

Step 1: Have a model meta-learn from many natural languages.



Step 2: Analyze the inductive biases that the model has meta-learned to gain insight into the biases that are relevant to the world's languages.

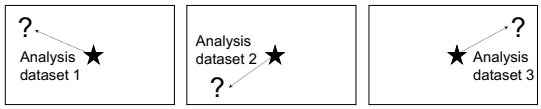


Figure 8: Meta-learning from natural-language data.

that combines the power of neural networks with the controlled inductive biases of symbolic approaches.

One important feature of the proposed approach is that it imparts soft biases rather than hard constraints. For example, after meta-learning, the model could learn attested language types more readily than unattested types—but it still could learn the unattested ones. This capability is at odds with some theories that predict that unattested language types should be unlearnable, but there are reasons to believe that the consistent patterns seen in language typology and language acquisition may be best viewed as biases rather than constraints: almost all linguistic universals have exceptions (Evans and Levinson, 2009; though see Smolensky and Dupoux, 2009); for example, the Arrernte language has been argued to be an exception to the syllable structure typology we have adopted (Breen & Pensalfini, 1999). Further, humans in artificial language learning experiments are capable of learning "unnatural" languages (Moreton & Pater, 2012).

Several other works have discussed meta-learning from a cognitive perspective (Lake et al., 2017; Griffiths et al., 2019; Lake, 2019; Grant et al., 2019), and in applied settings meta-learning has been applied to language to create technology in low-resource languages (Gu et al., 2018; Ponti et al., 2019; Kann et al., 2020). Our novel contribution is the use of meta-learning to analyze the interplay between data and linguistic inductive biases.

Our approach can be used to test the behavioral effects of a particular inductive bias (e.g., to test if the bias has the explanatory power hypothesized in a cognitive theory): All that is required to create a model with a specific inductive bias is a way to translate the bias into a distribution of meta-training languages, as we have demonstrated with Optimality Theory. In our experiments, we knew what factors defined the space of languages, and we showed that the inductive biases found through meta-learning reflected these factors; alternatively, this technique could be applied to naturally-occurring linguistic data for which we do not know the underlying data-generating process, to lend insight into the inductive biases that shaped this data (Figure 8). Finally, this framework is general enough that it can be straightforwardly applied to cognitive domains other than language (e.g., vision).

Acknowledgments

For helpful comments, we are grateful to Colin Wilson, Paul Soulos, the members of the Johns Hopkins Computation and Psycholinguistics Lab, and the members of the Johns Hopkins Neurosymbolic Computation Lab. This research was supported by NSF Graduate Research Fellowship No. 1746891, NSF INSPIRE grant BCS-1344269, NSF grant BCS-1920924, and contract number FA8650-18-2-7832 from the Defence Advanced Research Projects Agency. Our experiments were conducted using the Maryland Advanced Research Computing Center (MARCC).

References

Breen, G., & Pensalfini, R. (1999). Arrernte: A language with no syllable onsets. *Linguistic Inquiry*, *30*(1).

Chomsky, N. (1980). Rules and representations. *BBS*, *3*(1).

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3).

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *BBS*, *32*(5).

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. ICML*.

Grant, E., Peterson, J. C., & Griffiths, T. L. (2019). Learning deep taxonomic priors for concept learning from few positive examples. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Greenberg, J. H. (1963). *Universals of language*. Cambridge, MA: MIT Press.

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences*, *29*, 24–30.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *TiCS*, *14*(8).

Gu, J., Wang, Y., Chen, Y., Li, V. O. K., & Cho, K. (2018). Meta-learning for low-resource neural machine translation. In *Proc. EMNLP*.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proc. NAACL*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hochreiter, S., Younger, A., & Conwell, P. (2001). Learning to learn using gradient descent. *Proc. ICANN*, 87–94.

Jakobson, R. (1962). *Selected writings: Volume 1: Phonological studies*. Mouton.

Kann, K., Bowman, S. R., & Cho, K. (2020). Learning to learn morphological inflection for resource-poor languages. *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proc. ICLR*.

Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. In *Proc. NeurIPS*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *BBS*, *40*.

Maddieson, I. (1996). Phonetic universals. *UCLA Working Papers in Phonetics*.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proc. ACL*.

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, *45*(37).

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning, Part II: Substance. *Language and linguistics compass*, *6*(11).

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3).

Ponti, E. M., Vulić, I., Cotterell, R., Reichart, R., & Korhonen, A. (2019). Towards zero-shot language modeling. In *Proc. EMNLP-IJCNLP*.

Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Wiley.

Ross, J. R. (1967). *Constraints on variables in syntax*. Unpublished doctoral dissertation, MIT.

Smolensky, P., & Dupoux, E. (2009). Universals in cognitive theories of language. *BBS*, *32*(5).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proc. NeurIPS*.

Swadesh, M. (1950). Salish internal relationships. *International Journal of American Linguistics*, *16*(4).

Thrun, S., & Pratt, L. (1998). *Learning to learn*. Kluwer Academic Publishers.

van Schijndel, M., Mueller, A., & Linzen, T. (2019). Quantity doesn't buy quality syntax with neural language models. In *Proc. EMNLP-IJCNLP*.

Zipf, G. K. (1949). Human behavior and the principle of least effort.