Compositional generalisation with linguistically motivated meta-learning

Averie Ho Zoen So
Term Paper for Summer Semester Seminar 2022: Compositional Generalisation
Language Science and Technology, Saarland University
{averieso@coli.uni-saarland.de}

Abstract

Compositional generalisation is a hallmark of natural language and human cognitive capacity, which allows humans to understand and produce novel utterances. Various compositional generalisation benchmarks have shown that neural networks struggle with out-of-distribution generalisation. On the other hand, research shows that a meta-learning framework can improve performance in out-of-distribution linguistic tasks. The current paper applies the meta-learning framework to COGS, a generalisation benchmark that targets lexical and structural out-of-distribution generalisation abilities. The process of adapting the COGS dataset to the meta-learning procedure is linguistically motivated in several aspects, including task division, the targeted inductive biases and the order tasks presented during meta-learning. It is found that the meta-learned model does not outperform the standard supervised training procedure, but it replicates certain aspects the challenges of the COGS dataset, namely a clear distinction in difficulty between structural and lexical generalisation tasks. Additionally, a few technical aspects are revealed to be important in order to properly evaluate the effectiveness of meta-learning on compositional generalisation ability.

1 Introduction

Compositionality is a key feature of natural language, referring to the property that "the meaning of a whole is the meanings of its parts and of the way they are syntactically combined" (Partee et al., 1995). It can be seen as a hallmark of human communication, which enables the comprehension and production of novel linguistic utterance systematically but with infinite possibilities.

Despite major advances in modelling language with neural networks, their ability to compositionally generalise is unclear. This question came into light with recent research which aims to investigate this fundamental property of natural language in successful language models. More specifically, various compositional generalisation benchmarks have been developed to probing this ability in neural models (Lake and Baroni, 2018; Kim and Linzen, 2020). One of these benchmarks, SCAN (Lake and Baroni, 2018), found that RNNs fail to generalise the meaning of "dax" to "dax twice" - a task that a human child can easily accomplish without being taught to do so. This demonstrates that while existing neural models of language may be successful in various tasks such as translation, parsing and question answering, they do not process linguistic input as humans do. The lack of generalising abilities could be an explanation for major shortcomings of modern neural networks, namely the fact that they require a huge amount of linguistic data to acquire a language, whereas only minimal linguistic input is required for human language acquisition (Baroni, 2020).

One possible explanation for this disparity is that humans are largely constrained in language acquisition while neural networks are unconstrained (McCoy et al., 2020). Following the theory of Principles and Parameters (Chomsky, 1981), language acquisition can be understood as a process in which universal principles interact with language-specific parameters. In other words, there exists universal, innate inductive biases that guide the learner in language acquisition, allowing them to learn efficiently despite the huge variation in the data they may be faced with. As such, by introducing constraints relating to syllable structure to neural networks, McCoy et al. (2020) has demonstrated that neural networks can also acquire aspects of a language from few-shot learning.

While McCoy et al. (2020) shows promising results in introducing linguistic inductive biases via meta-learning, it has only been applied to a phonological task for syllable structure transformation. It remains a question whether this framework can be applied to more complex levels of language such as semantic and syntactic patterns, which are strongly tied with generalisation abilities.

The connection between syntax, semantics and generalization originates from the Principle of Compositionality and the related properties of natural language: productivity and systematicity (Szabó, 2012). Productivity refers to our ability to produce and understand new utterances that we have never heard of before and systematicity refers to the property that if one is able to understand a complex expression, they would also be able to understand other complex expressions that are made up of those components. Hence, there exists constraints in natural language that allows us to reuse them and produce new sentences with our intended meaning consistently. A simple example to demonstrate this is that if we know the meaning of *"John likes Mary"*, we would also know the meaning of *"Mary likes John"*, and that is dependent on our knowledge of the individual meanings of *"Mary"*, *"John"* and *"likes"* (semantics), as well as the grammatical rule that the noun before the verb is the subject and that after the verb is the object (syntax).

As that these constraints are considered essential components to language acquisition, one can also consult child language acquisition literature to understand how linguistic inductive biases may be formed. One of these constraints is the noun bias: the observation that among young children, nouns are acquired earlier than other linguistic categories such as verbs, suggesting a universal cognitive bias which preferentially maps words to objects rather than other concepts (Gentner, 1982). However, this bias is disputed since more recent evidence suggests that it is not universal (Tardif et al., 1999).

Therefore, the goal of this paper is to extend the application of the meta-learning framework proposed by McCoy et al. (2020) to COGS (Kim and Linzen, 2020), which is a synthetic semantic parsing dataset that maps natural language input to its logical form. In particular, this would be conducted in a way that is linguistically motivated, which includes targeting linguistic inductive biases and experimenting with a noun-biased meta-learning procedure. The adapted datasets and meta-learning implementation is available on the Github repository[1].

The rest of this paper is organised as follows: the next two sections give more in depth introductions to the main components of this paper, the COGS dataset (Section 2) and the meta-learning approach (Section 3). Section 4 is a literature review of related work in compositional generalisation and meta-learning in NLP. Then, the methodology (Section 5) as well as the various experiments and results (Section 6) are described. Lastly, the discussion and conclusion of the paper are presented in Sections 7 and 8 respectively.

2 COGS

COGS (Kim and Linzen, 2020) aims to evaluate the ability of neural networks to compositionally generalize. In this dataset, the input is an English string and the goal is to map the sentence to a logical form. The dataset is specifically split in ways to probe abilities that can only be achieved if the neural network can compositionally generalize, including: new combinations of familiar syntactic structures, new combinations of familiar words and familiar structures, and out-of-distribution generalization. Experiments on various model architectures show that neural networks can generalize well in in-distribution splits, but not out-of-distribution splits, which suggests their limited ability in compositional generalization.

Compared to other similar compositional generalisation benchmarks such as SCAN (Lake and Baroni, 2018) and CFQ (Keysers et al., 2019), COGS is advantageous as it captures syntactic constructions such as verb argument structure alternation and verbs with multiple classes while the natural language fragments in its predecessors focus on a small proportion of the English grammar and vocabulary (SCAN: imperatives, directions and repetitions, CFQ: questions and imperatives).

Other than the variety in linguistic construction, COGS is also novel in proposing a 'generalization split', where previous studies mostly concentrated on out-of-distribution (SCAN), maximum compound divergence (CFQ) and input length (both). The generalization splits aim to replicate generalizations that children are able to do, such as *a hedgehog* being a subject only in training in *a hedgehog ate the cake*, but has to be generalized to an object in *the baby liked the hedgehog* during testing. Instead of input length, the closest equiv-

[1] https://github.com/averieso/meta-for-cogs

alent split in COGS is depth of recursion, which correlates with input length, but more closely to how natural language behaves in reality.

3 Meta-learning with MAML

McCoy et al. (2020) proposes a meta-learning framework which introduces pre-defined inductive biases in sequential neural networks by a careful design of meta-learning datasets. Their experiments show success on a phonological task which involves syllable structure transformation. By framing each synthetic 'language' as individual tasks during meta-learning, the learner is exposed to a variety of tasks that share universal, high level inductive biases (eg. four constraints of syllable structure rules) but differ in language-specific ways (eg. priority of the constraints). The form of meta-learning is done via *model-agnostic meta-learning* (MAML, Finn et al. 2017), which focuses on learning initial parameters (Lee et al., 2022).

After meta-learning from these tasks, the model is able to succeed in a learning syllable generation task for an unseen language with much fewer training examples than a randomly initialised model. Post-hoc analyses further show that the intended inductive biases have been successfully acquired by the model. Furthermore, some generalisation abilities are demonstrated without being encoded at all during meta-learning, such as applying the same patterns to longer sequence. This demonstrates that inductive biases introduced via meta-learning can be used to constrain language acquisition in neural models, in a way similar to the innate principles in humans.

3.1 From COGS to meta-learning

In the meta-learning setup, the data must be separated into different "tasks" where all the tasks are assumed to be variants that follow broadly similar principles.

Since COGS is originally trained with the standard supervised setting, the training data does not differentiate between different types of generalisations. It is therefore an open question how to best separate COGS into "tasks" that would be suitable for meta-learning. To do this, one may first consult the categories and sub-tasks in Sections 3.1 to 3.5 of Kim and Linzen 2020:

- *3.1. Novel Combination of Familiar Primitives and Grammatical Roles* (9 sub-tasks)
- *3.2. Novel Combination Modified Phrases and Grammatical Roles* (1 sub-task)
- *3.3. Deeper Recursion* (2 sub-tasks)
- *3.4. Verb Argument Structure Alternation* (6 sub-tasks)
- *3.5. Verb Class* (3 sub-tasks)

There is a focus in COGS on the generalisation ability surrounding only nouns and verbs. Specifically, 3.1 targets nouns specifically while 3.4 and 3.5 target verbs. These three categories correspond to lexical generalisations, while the others correspond to structural generalisations. Considering that 18 out of 21 generalisation types are lexical, it would be reasonable to split the dataset by lexical items.

The advantage of this method is obtaining a reasonable number of tasks which mostly align with the generalisation types of COGS. Additionally, dividing the dataset by lexical item would also mean that the tasks are created based on the part-of-speech properties, where syntactic categories such as nouns and verbs are considered universals in linguistic theories such as Greenberg (1963) and Chomsky (1957). Thus, this approach also echos the division of task by universal linguistic inductive biases in McCoy et al. (2020).

However, the limitation is that some parts of the training set would be excluded (eg. if number of instances of a lexical item is too low to form a task), some instances in the training set would be repeated and no tasks would explicitly align with structural generalisations.

Therefore, following the sub-classes in the generalisation set of COGS, the current paper aims to impart these inductive biases. Further details of how the "tasks" are created is described in Section 5.1.1.:

1. distinction between nouns and verbs and their sub-classes (eg. proper noun, common noun)
2. nouns can take on different grammatical roles in a sentence (eg. subject, object)
3. verb argument structure can be alternated (eg. active, passive)

4 Related Work

4.1 Meta-learning in NLP

Meta-learning, also referred to as Learning to Learn, is a machine learning approach aimed at improving model generalizability and data efficiency (Lee et al., 2022). Originally used in image processing (Hospedales et al., 2021), it is gaining popularity in NLP research as well. For example Nooralahzadeh et al. (2020) meta-learned on 15 languages for QA and NLI tasks and was able to achieve zero-shot and few-shot performance for low-resource languages. Other than cross-lingual settings, meta-learning has also been used in domain adaptation in MT for low resource domains, such as adapting from news and legal data (high resource) to medical records (low resource) (Li et al., 2020).

4.2 Meta-learning for compositional generalisation

There has been a few approaches using meta-learning to improve performance on neural network generalisation abilities. Lake (2019) aims to enable the model to acquire higher level and more abstract patterns via meta-learning, such as using mutual exclusivity to resolve ambiguity of unknown tokens. Instead of learning model initialisation parameters, they store prior knowledge in external memory, then use an attention mechanism to add information from the external memory as context. The experiments show that external memory can improve SCAN performance in terms of generalising to new primitives (jump) and new combinations (jump around right), but the model still fails to generalise to longer lengths.

Conklin et al. (2021) is the most similar to Mccoy et al's and the current approach in terms of utilising model initialisation with MAML in the meta-learning phase, although a different variant is used (Domain Generalisation variant, DG-MAML). Compared to McCoy et al. (2020), their bias is more general: memorisation inhibition. This is achieved by sampling similar training data as test data, measured by tree structure or string similarity. While their model succeeds in improving model performance in SCAN and COGS tasks, the inductive bias is not linguistically motivated as it specifically targets a property of traditional supervised learning rather than one of human cognitive abilities. Additionally, they found that each model only performs relatively well on some but not all COGS tasks, meaning that the biases that they have induced are not completely explainable in how the models achieve generalisation abilities. It is unclear that the generalisation abilities will remain in broader NLP tasks such as low-resource MT.

4.3 Pre-training for compositional generalisation

Other than meta-learning, various forms of ”pre-training” to improve compositional generalisation have been explored. For example, Oren et al. (2021) efficiently induced compositional generalisation abilities in the Schema2QA dataset (transforming natural language questions to query templates) by pre-training models with synthetic training sets created from a synchronous grammar and by strategically selecting a small but structurally-diverse samples from the synthetic dataset. Shi et al. (2020) is another approach which mimics human ability to learn new concepts by connecting them to prior knowledge and shows success in various types of architectures on SCAN tasks as well as general NLP tasks including MT and semantic parsing.

5 Methods

5.1 Dataset

Since the COGS dataset is not originally intended for meta-learning, some re-distribution is required to generate datasets that are applicable to the meta-learning procedure. In general, this involves splitting the COGS dataset into individual “tasks” as well as separating the exposure examples from the in distribution training instances, which would makeup the meta-learning training set, to the standard training set. Essentially, this re-formulates the standard training component as a *zero-* or *one-shot learning task.*

5.1.1 Meta-learning dataset

The meta-learning phase frames each unique token within the COGS in-distribution datasets as a task, which includes all the datasets in COGS (train, development, test) except the generalization set. For each unique word, all sentences that contain the word are selected, and each collection forms a task. As the lexical generalization types are directly related to only nouns and verbs, only words that belong to these two classes are selected, while other words such as function words are removed. Note that this also means that instances can be repeated across different tasks if they contain multiple noun and verbs.

Each meta-learning task consists of its own training, development and test set. We sample 100 instances for each task, in which 40 instances are in the training set, while 30 are in the development and test set respectively. If a task has less than 100 instances, it would be removed. For tasks that have more than 100 sentences, only 100 sentences are randomly selected to ensure the training is not over-fitting to some words. The remaining number of task is 201. Given that COGS was not designed with the meta-learning objective, and thus was not created with a wide variety of vocabulary, this is a significantly smaller number of tasks than the original implementation, which had 20,000 tasks.

5.1.2 Standard training dataset

COGS In COGS, there are 21 types of generalizations to be assessed, which is the number of tasks for the standard training set. The standard training set is a combination of the exposure examples from the original training set and the generalization set.

To replicate the setup as closely as possible to the original COGS dataset, the exposure example (n=1) is the training set, the development set is empty, and the test set would have 1000 instances. Each generalization type is assessed with one lexical item. For example, all and only cases of *Paula* in the generalization set assess the generalization type from a proper noun primitive to a proper noun object. As such, each exposure example from the original training set and the corresponding task in the generalization set are matched to create a standard training task.

However, this setup is not applicable to all tasks, since the 3 structural generalization tasks (*cp recursion, pp recursion, and obj pp to subj pp*) do not have exposure examples. In that case, a dummy training instance (*cake:cake*) would be placed in the training set for the task.

Inconsistent primitives In order to investigate whether the specific inductive biases in Section 3.1 are successfully imparted, additional synthetic tasks are created to form an "inconsistent primitives" dataset. This targets only the tasks where the exposure example is a primitive, specifically: *primitive noun to subject (common noun), primitive noun to subject (proper noun), primitive noun to object (common noun), primitive noun to object (proper noun)* and *primitive verb to infinitival argument*.

For each of these exposure examples, the logic form is changed to a different type of part-of-speech. For example, *Paula*, a proper noun, is introduced as a common noun and a verb in the training phase of the inconsistent tasks:

Proper noun (consistent)
Paula → *Paula*

Common noun (inconsistent)
Paula → *LAMBDA a . Paula (a)*

Verb (inconsistent)
Paula → *LAMBDA a .LAMBDA e . Paula . agent (e , a)*

With only the exposure example changed and the test set identical, inconsistencies should be created, where, for example, a verb appears in positions where a noun should be. This results in 10 additional tasks, where each of the 5 tasks using a primitive exposure example would have 2 additional inconsistent tasks.

5.2 Training Details

The model is a sequence-to-sequence unidirectional LSTM, which is used in McCoy et al. (2020) and in some experiments of COGS (Kim and Linzen, 2020). We use the implementation from McCoy et al. (2020), which provides more in-depth details of the model architecture[2]. The main architectural difference is that the LSTM in Kim and Linzen (2020) has 2 layers while the implementation of McCoy et al. (2020) has only 1 layer. While this difference can be critical to the experimental results, it is not modified in the current experiments.

Additionally, the loss function was modified from the original implementation, such that the loss is calculated on the proportion of correct tokens in the whole string rather than a binary classification (eg. a string is either correct or incorrect). This modification was necessary since the model was unable to train with 0 development accuracy.

In terms of hyper-parameters, since the COGS dataset has considerably longer sequence than the original implementation, the maximum length of output is changed from 20 to 60. The embedding size and hidden size are both set at 512, following the setup in COGS. The learning rate is 0.1 within training instances in a task and 0.001 across tasks.

5.2.1 Meta-training

MAML for meta-learning is used in both McCoy et al. (2020) and Conklin et al. (2021). From the

[2] available online: https://github.com/tommccoy1/meta-learning-linguistic-biases

201 in-distribution tasks, we split 100 tasks to be the meta-training set, 50 in the meta-development set and 51 in the meta-test set. For each task in the meta-training set, the model is trained on 40 examples and then evaluated on 30 held-out examples (see Section 5.1.1). After every set of 5 tasks, the model is evaluated on how well it can learn on 40 examples for the 50 tasks in the meta-development set. Meta-training is terminated after 10 consecutive evaluations without improvement.

5.2.2 Standard training

After meta-training, the model is evaluated based on its performance on learning held-out tasks in the meta-test set as well as the generalisation set, where the former is still in-distribution compared to the meta-train set, while the latter is out-of-distribution. The standard training uses the same hyper-parameters as in meta-training.

6 Experiments and results

Model performance with various settings on the meta-dev, meta-test and generalisation set is presented in Table 1. Overall, the *40-shot* meta-learning setting resulted in the best performance but only by a small margin.

model	**Dev.**	**Test**	**Gen.**
40-shot	0.895	0.882	0.781
70-shot	0.893	0.799	0.706
Noun-biased	0.866	0.832	0.759

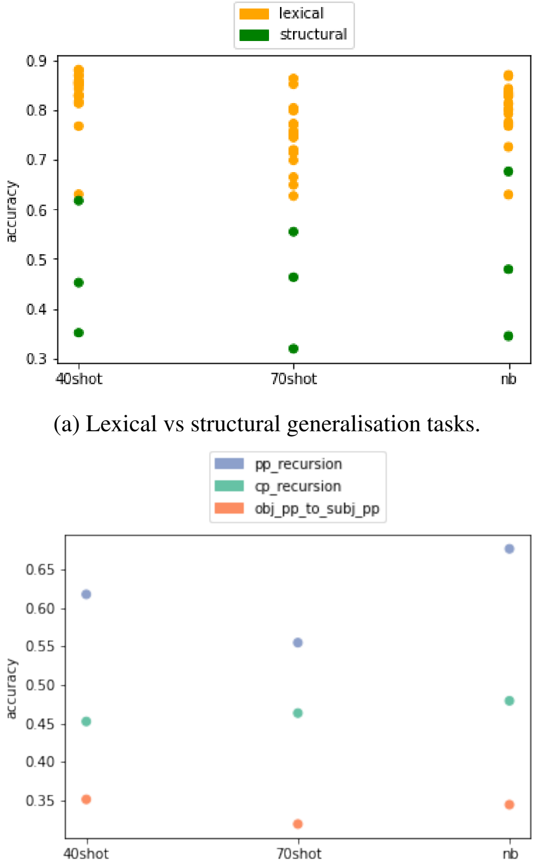
Table 1: Average accuracy by proportion of tokens that matches the target output. Note that this is not directly comparable to the accuracy in Kim and Linzen (2020), which is based on exact matches with the target sequence.

6.1 Meta-learning results

The model with meta-learned initial parameters had 88.2% accuracy on the 40-shot meta-test set. This is significantly lower than the meta-learning results of McCoy et al. (2020), which had 98.8% on the syllable transformation task during meta-learning, as well as the model in Kim and Linzen (2020) with standard training procedure, which achieved 99% accuracy.

6.2 *k-shot* meta-learning

A *70-shot* meta-learning experiment is conducted in order to more closely replicate the original *100-shot* learning setup in McCoy et al. (2020). The distribution of meta- train, dev and test sets were changed from 40-30-30 (*40-shot*) to 70-10-20 (*70-shot*). Although the 70-shot setting had more training examples to learn from, performance on the generalisation set did not improve.



(a) Lexical vs structural generalisation tasks.

(b) Individual structural generalisation tasks.

Figure 1: Average accuracy of proportion of correct tokens by generalisation tasks.

6.3 Effect of noun bias in meta-learning

This experiment aims to investigate whether changing the order in meta-learning would have an effect in generalisation ability by ordering the meta-learning dataset, where all tasks that are nouns are presented first, followed by the verb tasks. In this experiment, only the order of tasks is changed, and it remains a 40-shot learning task. The performance of the noun-biased model did not differ much from the random order models.

6.4 Lexical vs structural generalisation

Accuracy for lexical and structural generalisation tasks are shown in Figure 1a. As in Kim and Linzen (2020), lexical generalisation tasks achieve better performance than structural generalisation tasks across all models.

A closer look into the structural task performance in Figure 1b reveals that the relative performance on these three tasks remain consistent across models: *pp recursion* has the highest accuracy, followed by *cp recursion* and *obj pp to subj pp* has the lowest accuracy. This is also similarly found in Kim and Linzen (2020) where the pp recursion task is the only structural task that achieved non-zero accuracy[3].

6.5 Verification of inductive biases

model	Consistent	Inconsistent
40-shot	0.860	0.860
70-shot	0.794	0.794
Noun-biased	0.821	0.821

Table 2: Average accuracy by task consistency.

Following McCoy et al. (2020), whether the inductive biases have been successfully imparted can be evaluated on how well the model learns tasks that do not conform to the intended biases. Table 2 shows that the model does not differentiate between consistent and inconsistent tasks, as the accuracy remains identical across tasks in all models.

7 Discussion

The current paper has investigated whether a linguistically motivated meta-learning approach can lead to compositional generalisation ability in a neural model, assessed by performance on the COGS dataset. Specifically, the linguistically motivated aspects include 1) splitting meta-learning tasks by lexical items, thereby targeting linguistic inductive biases and 2) conducting an experiment with a noun-biased meta-learning dataset.

Overall, while current approach to meta-learning does not outperform the standard supervised training procedure, it replicates some aspects of generalisation ability in the original COGS experiments by Kim and Linzen (2020), such as the finding that structural generalisation tasks are more difficult than lexical generalisation tasks for a neural model, and that some structural tasks are easier than others, such as *pp recursion*. However, the results do not support that linguistically motivated features helps to improve generalisation ability in the neural model. Firstly, the noun-biased model does not outperform models that are trained with a random order. Secondly, providing the model with exposure examples that violate grammatical rules does not degrade its performance. This suggests that the current experiments have not successfully imparted the targeted inductive biases.

While these results could suggest that linguistically motivated aspects in the current approach are not beneficial to improving compositional generalisation ability, unlike alternative approaches which are not linguistically motivated (Conklin et al., 2021), there are several significant shortcomings in the above experiments which could lead to a lack of improvement in generalisation ability.

Mainly, the inability to reach close to perfect meta-learning results as in previous studies suggests that the current model is too simple for the COGS dataset. The setup in McCoy et al. (2020) was simpler than COGS, in terms of token variability and sequence length. While the model is also an LSTM as in Kim and Linzen (2020), the distinction is that the original COGS experiments were conducted on a model with two hidden layers, where the current model has only one.

Secondly, the current experiments are expected to benefit from a version of COGS which is adapted for the meta-learning training procedure. While dividing tasks by lexical item has several advantages, the resulting number of tasks are only about 1% of the number in McCoy et al. (2020), and does not fully utilise the COGS training dataset. The finding that exposure examples has no effect on the model's performance could suggest that a one-shot task for the COGS dataset may not be suitable. Hence, a version of COGS which allows for a larger number of tasks during meta-learning can better demonstrate the effectiveness of meta-learning.

Other than addressing these limitations, future research can potentially investigate a multilingual version of COGS which targets universal linguistic inductive biases, rather than solely English. A multilingual dataset would perhaps be more suitable for the meta-learning framework proposed by McCoy et al. (2020), since it is intended to investigate universal biases, while the current experiments with COGS assumes some universality across the individual tasks (eg. that nouns and verbs share some commonality).

8 Conclusion

In this paper, the COGS dataset is used to assess compositional generalisation ability of meta-

[3] accuracy based on an exact match with the target sequence

initialised models. While the meta-initialised model does not outperform the model trained with a standard supervised setting, it replicates some aspects from previous findings, in particular the difficulty of structural generalisation tasks in comparison to lexical generalisation tasks. The current experiments also show that several linguistically motivated features do not bring improvements to compositional generalisation ability. However, some technical limitations such as model architecture and dataset design would have to be addressed in order to confirm the validity of this conclusion.

References

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton and Co., The Hague.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris.

Henry Conklin, Bailin Wang, Kenny Smith, and Ivan Titov. 2021. Meta-learning to compositionally generalize. In *ACL/IJCNLP (1)*, pages 3322–3335.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Dedre Gentner. 1982. *Why Nouns are Learned Before Verbs: Linguistic Relativity Versus Natural Partitioning*. BBN report. University of Illinois at Urbana-Champaign.

Joseph Harold Greenberg. 1963. Universals of language.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2021. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Najoung Kim and Tal Linzen. 2020. Cogs: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Brenden M Lake. 2019. Compositional generalization through meta sequence-to-sequence learning. *Advances in neural information processing systems*, 32.

Hung-yi Lee, Shang-Wen Li, and Ngoc Thang Vu. 2022. Meta learning for natural language processing: A survey. *arXiv preprint arXiv:2205.01500*.

Rumeng Li, Xun Wang, and Hong Yu. 2020. Metamt, a meta learning method leveraging multiple domain data for low resource machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8245–8252.

R. Thomas McCoy, Erin Grant, Paul Smolensky, Thomas L. Griffiths, and Tal Linzen. 2020. Universal linguistic inductive biases via meta-learning. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562.

Inbar Oren, Jonathan Herzig, and Jonathan Berant. 2021. Finding needles in a haystack: Sampling structurally-diverse training sets from synthetic data for compositional generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10793–10809.

Barbara Partee et al. 1995. Lexical semantics and compositionality. *An invitation to cognitive science*, 1:311–360.

Ning Shi, Boxin Wang, Wei Wang, Xiangyu Liu, Rong Zhang, Hui Xue, Xinbing Wang, and Zhouhan Lin. 2020. From scan to real data: Systematic generalization via meaningful learning. *arXiv preprint arXiv:2003.06658*.

Zoltán Gendler Szabó. 2012. The case for compositionality. In *The Oxford Handbook of Compositionality*. Oxford University Press.

Twila Tardif, Susan A Gelman, and Fan Xu. 1999. Putting the “noun bias” in context: A comparison of english and mandarin. *Child development*, 70(3):620–635.