A Dataset Details

A.1 Heuristics Used in Template Generation

In order to make the template generation process more efficient, we apply some heuristics to eliminate templates that would not result in valid sentences in any of our artificial languages. We eliminate templates with the following properties:

1. Shorter than 3 words (the shortest valid sentence in all grammars is 3 words),
2. Starting with a conjunction,
3. Ending with a conjunction,
4. Containing 2 consecutive conjunctions,
5. Containing 2 consecutive prepositions,
6. Starting with subject or object markers,
7. The total number of subject and object markers is greater than the number of NPs,
8. A complementizer appears in the template without a complement verb.

A.2 Restrictions Applied to Parser

In order to parse our templates and assign them to the suitable languages, we adapt the NLTK CCGChartParser (Bird et al., 2009) by disabling type raising, which is included in Combinatory Categorial Grammar (CCG) (Steedman, 1996) and implement and integrate the permutation operation as defined by Briscoe (1997, 2000), which is included in Generalized Categorial Grammar (GCG) (Wood, 2014).

In the NLTK CCGChartParser, restrictions can be applied to prevent composition, crossing, and substitution by adding ",","." or "_", respectively, before the argument when defining the grammar. When we implement permutation, we introduce an additional character "@" that prevents permutation from being applied.

When defining our grammars, we restrict permutation to categories with S functors only, i.e., verbs. Additionally, in order to restrict the subject and object markers to only combine with NP, we restrict composition when defining the NP_SUBJ and NP_OBJ categories in the grammar.

Using GCGs to create our artificial languages can allow for flexible word orders as a result of permutation. This would result in OSV sentences being present in SOV datasets, VSO sentences being present in VOS datasets and vice versa. We inhibit permutation when parsing templates into OSV, SOV, VOS and OVS languages, except in the sentences where a REL category is present. This way, there is a clearer distinction between these languages.

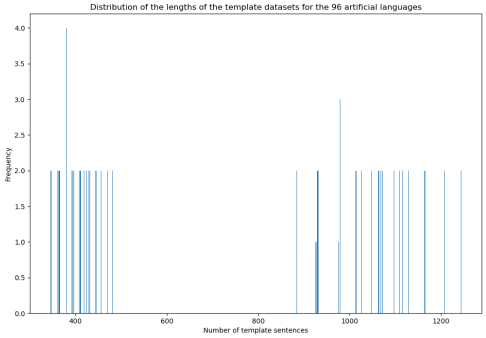

Figure 6: Histogram showing the distribution of the number of templates in the 96 artificial languages

A.3 Dataset Statistics

We calculate statistics for our 96 artificial languages and the templates from which we generate the sentences to provide more insight into the properties of the datasets.

We calculate the average sequence length for the templates and sentences used in evaluation, and they are both approximately 9.35 words long. We count the number of sequences in each template and plot the distribution of them in Figure 6. The smallest and largest template files consist of 354 and 1244 template sequences, respectively. We calculate the average template size as 723.8 sequences.

We show the number of overlapped sentences and overlapped templates, and percentage of overlapped sentences and templates in Figures 7,8,9, and 10. As shown in the heatmaps, there is some overlap in the templates for the different languages (Figures 9 and 10). However, there is negligible overlap between the datasets used for experiments (Figures 7 and 8).

B Limitations

In this work, we use artificial languages to evaluate our LMs' inductive biases. Artificial languages, though controlled, often do not reflect many of the properties and complexities of natural languages, such as subject-verb agreement, lexical ambiguity,

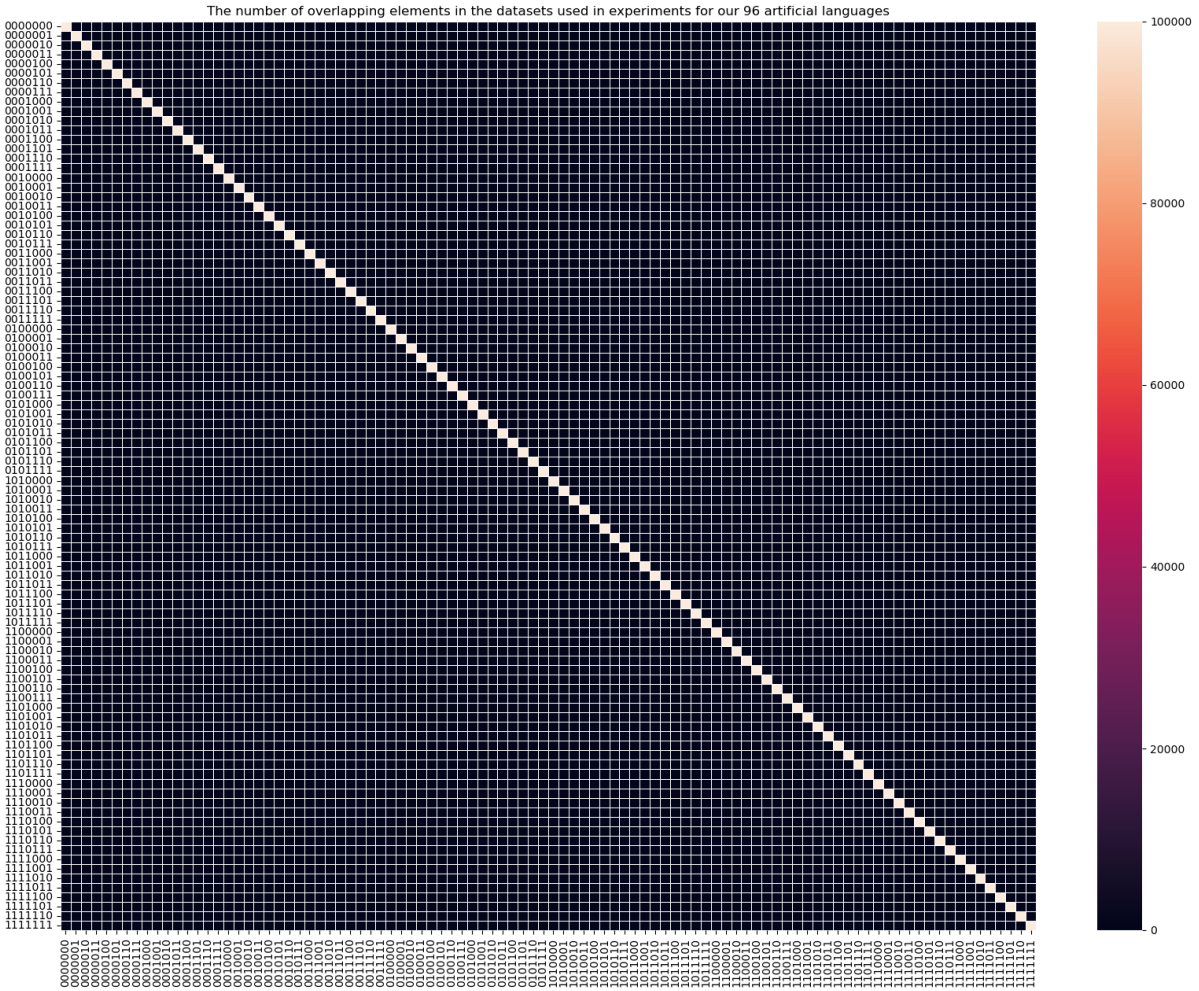

Figure 7: Heatmap showing the overlap between the datasets for the 96 original grammars.

and long-distance dependencies. We do not currently distinguish between nouns of different pluralities or verbs of different tenses in our lexicon. Although our study is in the direction of resolving such limitations with GCG, in the future, we plan to extend our lexicon and grammar to include more detail and more realistic properties of natural language step-by-step. Especially, our artificial languages can go beyond context-free, and allow us to evaluate the different types of longer-distance dependencies, which we have not explored in detail in this work, but plan to address in the future.

Such future work should also include more in-depth ablations on what kind of additional complexity, compared to the existing PCFG data, affected the results. The evaluation framework also has room to be extended; for example, we can evaluate the compositional generalization of LMs using out-of-domain, longer sequences in evaluation. It will also be fruitful to integrate the perspective of interpretability research to answer how and why LMs struggled with specific word order languages internally.

Lastly, while the training paradigms we use in this work are very commonly used, our tested LMs are limited with respect to, e.g., their parameter size, types, and training procedures. In the future, we would like to develop a better understanding of the learning dynamics and explore LM learning of our ALs using different learning paradigms.

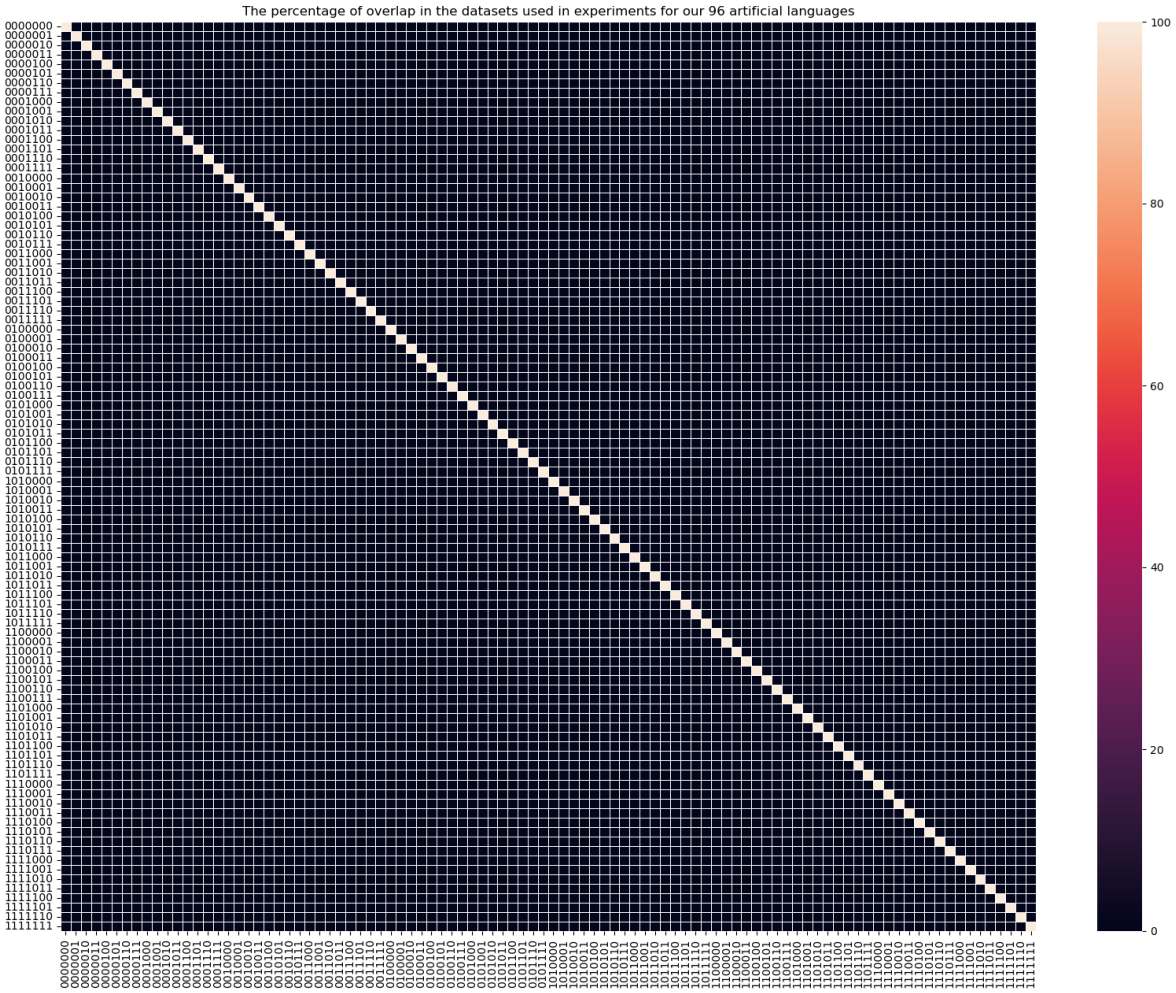

Figure 8: Heatmap showing the percentage overlap between the datasets for the 96 grammars.

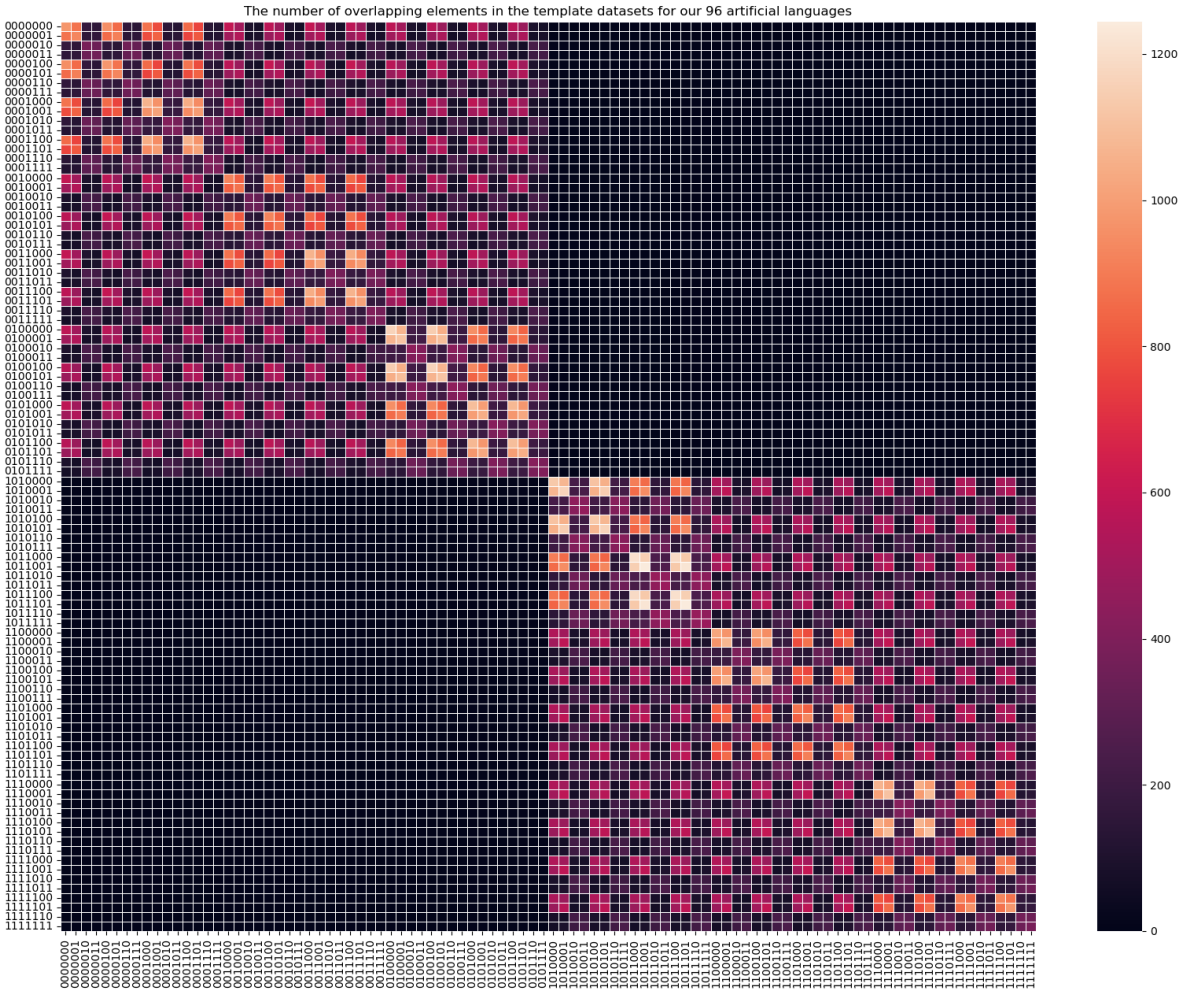

Figure 9: Heatmap showing the overlap between the template datasets of the original 96 grammars.

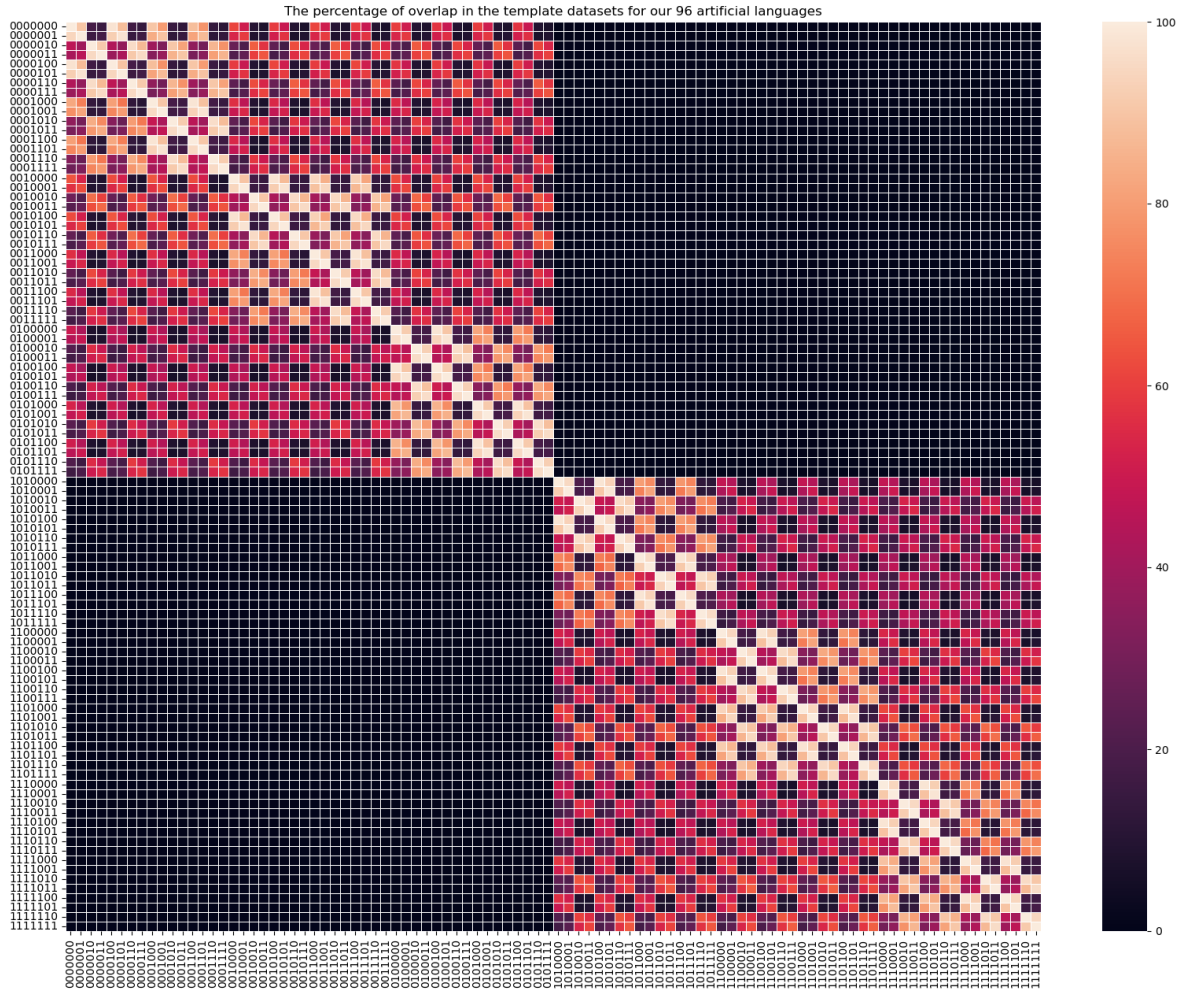

Figure 10: Heatmap showing the percentage overlap between the template datasets for 96 artificial languages.