Analyzing the Linguistic Priors of Language Models with Synthetic Languages

Alessio Tosolini[1,2] **Terra Blevins**[3,4]
[1]Yale University
[2]Paul G. Allen School of Computer Science, University of Washington, Seattle
[3]Faculty of Computer Science, University of Vienna
[4]Khoury College of Computer Sciences, Northeastern University
Correspondence:, alessio.tosolini@yale.edu, t.blevins@northeastern.edu

Abstract

While modern language model architectures are often assumed to be language-agnostic, there is limited evidence as to whether these models actually process the vast diversity of natural languages equally well. We investigate this question by analyzing how well LMs learn carefully constructed artificial languages containing a variety of verbal complexity, ranging from simple paradigms to covering far more verb classes than occur in natural languages. Rather than learning all languages equally efficiently, models trained on these languages show strict preferences for processing simpler languages. Furthermore, while *some* observed model preferences mimic human linguistic priors, we find that they correspond to model memorization of its training data rather than generalization from it. This finding suggests that while model behavior often mimics human language understanding, the underlying causes of their proficiencies are likely very different.

1 Introduction

Transformer-based language models (LMs) are often assumed to be language-agnostic, or to learn all natural languages equally well. This has lead to their widespread use for different languages (Scheible et al., 2024; Ahmed et al., 2024, i.a.) and multilingual modeling (e.g., Üstün et al., 2024).

However, there is immense linguistic diversity in the world's languages, and human learners acquire aspects of these languages at different rates. For example, children take longer to learn the opaque Dutch gender system, mastering it by age six (Tsimpli, 2014), while children master the transparent Spanish gender system by three and a half, if not sooner (Lew-Williams and Fernald, 2007). It remains an open question as to whether this complexity similarly affects model acquisition of different languages: previous work exploring the differences in language modeling capabilities presents mixed results on the effect of morphological complexity on language modeling (Cotterell et al., 2018; Mielke et al., 2019; Park et al., 2021; Arnett and Bergen, 2024), and typological differences can impact the performance of models intended to be language-agnostic (Gerz et al., 2018). Furthermore, there is limited evidence whether LMs are even constrained to learning natural linguistic phenomena as humans are (Kallini et al., 2024).

We address this question by testing if LMs demonstrate human-like learning patterns when acquiring new, artificial languages. Specifically, we ask: **Do LMs exhibit linguistic priors favoring certain conjugation paradigms over others?**. We center our behavioral analysis on a single grammatical feature—verb conjugation—in a wide variety of linguistically plausible and implausible settings as a controlled case study into the effect of linguistic grammatical complexity on transformer-based modeling of language.

To evaluate LMs for these linguistic priors, we first construct artificial languages using a probabilistic context-free grammar (PCFG). These languages cover a wide range of (plausible and implausible) conjugation complexity while controlling for other confounding variables found in natural languages. We then test how proficiently and efficiently language models learn these languages by measuring their mastery of both subject-verb agreement (a commonly used linguistic test for LMs, see Gulordava et al., 2018), as well as a novel behavioral experiment for *verb class identification* in these languages throughout the training process.

Our experiments find that language models acquire more complex languages (i.e., those with more verb classes) more slowly. However, they achieve close to 100% accuracy on seen verbs given enough data, even in cases where the number of verb classes is far larger than naturally occurs in human languages. The models also perform significantly worse on novel verbs than those seen during

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 7–15

training, with the performance degradation increasing with the number of verb classes; this indicates that these models do not learn to generalize from the standard conjugation patterns shown to them during pretraining.

These findings suggest both that (1) these models are *not* language-agnostic, but are instead sensitive to the complexity of the target language, and that (2) behavior that resembles human-like language learning in models may actually be *memorization* of the training data, rather than *generalization* to the underlying linguistic rules. Put another way, correlations between model and human behavior do not necessarily indicate that their underlying mechanisms are the same. In light of these findings, we recommend future work analyzing model language learning to incorporate evaluations that disentangle these factors when probing language model behavior.

2 Methodology

This section presents our method for generating artificial languages with the desired characteristics (Section 2.1) and our behavioral experimental setting to test model proficiency on subject-verb agreement in these languages (Section 2.2).

2.1 Artificial Language Generation

To evaluate how well LMs can learn languages across different verb settings, we generate artificial languages with the desired features using a Probabilistic Context-Free Grammar (PCFG), an extension of context-free grammars that assigns probabilities to transitions between states, allowing for the stochastic generation of sentences. We focus our analysis on these artificial languages to control for various confounding factors found in natural languages, including but not limited to semantics, irregularities, ambiguity, and dialectal variation, that make direct comparisons difficult.

We define our PCFG with a set of parameters describing the language's word formation, syntax, and inflectional rules. For verb paradigms, this parameterization allows us to perform controlled ablations across various experimental settings. Specifically, for our experiments, we generate ten languages for each of the {1, 2, 3, 5, 8, 16, 32, 64} verb class settings and report the average performance and standard error in a given setting. There is no overlap in the suffixes between any two verb classes, and verb paradigms are fully regular.

Other parameterization of our PCFG is informed by common natural distributions of language features to ensure our artificial languages are as similar to natural ones as possible. In each language, the number of roots generated per part-of-speech approximates 1% of English senses in Kaikki (Ylönen, 2022), with nouns approximating 0.5% of senses since jargon is often overrepresented in nouns (Table 1). As Zipf's Law is ubiquitous in human language at many scales (Williams et al., 2015), the distribution from which words are selected is drawn from a Zipfian distribution. A skew of 1.2 is used for our word distribution, based on the empirical distribution found in the American National Corpus (Piantadosi, 2014). The verb class assigned to a verb is similarly drawn from a Zipfian distribution with a skew of 1.

We also allow for features (such as nominative for subjects) to be passed between states in the PCFG during generation (Figure 1); this enforces subject-verb agreement on person and number features within each sentence. A more detailed explanation of creating the artificial languages and generating sentences is given in Appendix A.

Part of Speech	Items	Kaikki Senses
Adjective	2000	199759
Determiner	1	387
Noun	4000	856855
Preposition	15	1337
Pronoun	6	1053
Verb	2000	220457

Table 1: Word counts per part of speech in our artificial languages versus Kaikki sense counts for English.

2.2 Model Training and Evaluation

When training language models on our artificial languages, we consider three factors: the verb

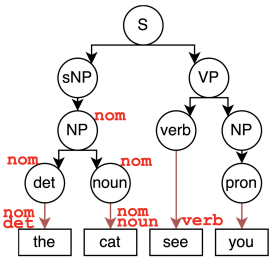
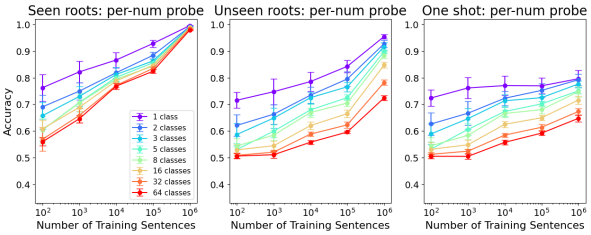
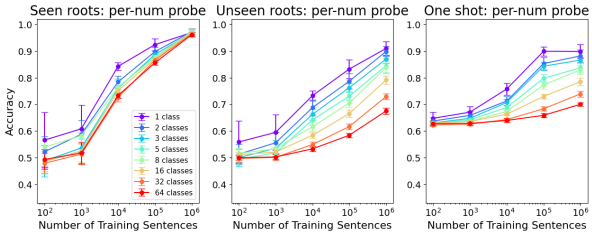


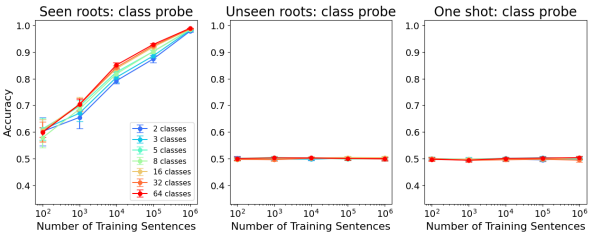
Figure 1: Sample English PCFG. The nominative feature nom passes from the child of the subject noun phrase *sNP* to its descendants, allowing for subject-verb agreement to be enforced later in the generation pipeline.



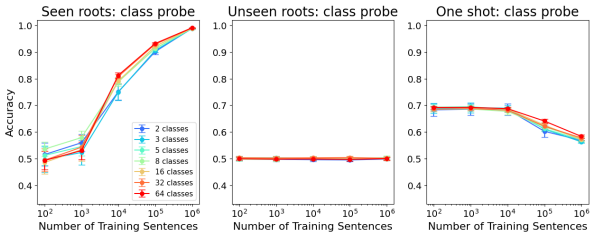
(a) Per-num behavioral test on GPT-2 trained from scratch



(b) Per-num behavioral test on finetuned GPT-2



(c) Class behavioral test on GPT-2 trained from scratch



(d) Class behavioral test on finetuned GPT-2

Figure 2: Behavioral test results for models trained on the generated languages at different training data sizes.

paradigm (or number of verb classes in a language), the size of the training dataset, and the model training scheme. We construct training datasets of $10^{2,3,4,5,6}$ grammatical training sentences from each of the generated languages; the two training schemes considered are training a randomly initialized GPT-2 model from scratch versus finetuning the pretrained GPT-2 model for English (Radford et al., 2019), both of which have approximately 124 million parameters. We train LMs on each combination of these settings (and across all ten languages per verb paradigm).

We then evaluate these models on how well they learn the artificial languages they are trained on by testing whether they can distinguish grammatical examples of the language from ungrammatical ones.[1] Specifically, we construct evaluation sets consisting of minimal pairs of grammatical and ungrammatical sentences and measure the perplexity of the model on each sentence; models that are well-fit to the training language should prefer (achieve a lower perplexity on) the grammatical sentences.

We consider two types of behavioral tests to probe how well the models learn subject-verb agreement and the verb classes. These tests determine the error type shown during inference in a minimally different pair of (grammatical, ungrammatical) sentences: the *per-num* test (Case 1), where the verb in the ungrammatical sentence takes a suffix marking a different person and/or number feature for that class:

(1) a. El perro *escucha* el gato.
"The dog *hears* the cat."
b. *El perro *escucho* el gato.
"The dog *hear* the cat."

and the *class* test (Case 2), where the ungrammatical verb takes a suffix from a different verb class agreeing with the subject's person and number:

(2) a. El perro *escucha* el gato.
"The dog *hears* the cat."
b. *El perro *escuche* el gato.
"The dog *hears*[2] the cat."

The evaluation sets contain 5,000 sentence pairs; we define accuracy as the percentage of test sentence pairs where the grammatically correct sentence's perplexity is less than that of the ungrammatical sentence. Thus, we measure the cases where the model assigns a higher likelihood to the grammatical case as a proxy for how well it models conjugation in the generated languages.

Testing covers three settings, varying whether test verb roots are seen during training: *seen roots*, where the model is evaluated on verb roots from the training; *unseen roots*, which evaluates the model on held-out verb roots to test model generalization; and *one-shot*, where the model is given one demonstration using a hitherto unseen verb before being

[1] This is a common approach for surfacing linguistic knowledge in LMs (e.g., Liu et al., 2019), particularly in the case of subject-verb agreement (Gulordava et al., 2018).

[2] Note that there is no equivalent, ungrammatical English translation for the *class* test, as English does not have verb classes that correspond to multiple regular conjugation paradigms like in Spanish.

tested on that same verb root. Given the set s of LMs trained across each (data scale, verb paradigm, and training scheme) combination, we evaluate s on all described (i) test types and (ii) evaluation settings. We report the mean performance and standard error across the ten runs for each combination.

3 Results

This section presents our language learning experiments. We find that while LMs model simpler inflectional paradigms more easily (indicating that they are not agnostic to language complexity), they struggle to generalize to new verbs across experimental settings, including on linguistically plausible verbal paradigms found in natural languages.

3.1 Per-Num Agreement Evaluation

Figures 2a and 2b show *per-num* behavioral test outcomes (where negative samples contain incorrect subject-verb agreement) on models trained from scratch and finetuned GPT-2, respectively. Across settings, adding verb classes to the generated languages generally corresponds to worse performance on (and slower acquisition of) subject-verb agreement by LMs.

For seen roots, all models achieve high accuracies of 97.5% or greater at the largest data size (1M training sentences). However, acquisition time varies across model and verb class settings: languages with more verb classes consistently need more data to achieve comparable accuracies to those with fewer classes[3]. Pretrained GPT-2 also learns to prefer correctly conjugated seen verbs *slower* than models trained from scratch.

Unsurprisingly, agreement accuracy on unseen roots is lower than on seen roots across comparable experiments.[4] However, we see the same relative trends here as on seen roots: more data improves conjugation accuracy (though now with larger gaps between the best- and worst-performing LMs), and finetuned GPT-2 continues to underperform in limited data settings. For unseen verbs, though, the performance gap between the randomly initialized and finetuned LMs is smaller, particularly on languages with eight or more verb classes.

Finally, we find that providing the model with one correctly conjugated demonstration does not consistently improve accuracy over the unseen verb ("zero-shot") setting. In many cases, the models perform similarly in both settings, and high-data regimes often perform *worse* when given a correct example. This, in addition to the unseen verb results, suggests the models do not learn abstract conjugation patterns when trained on these languages.

3.2 Verb Class Evaluation

Figure 2c presents the *class* behavioral test (where negative samples contain a verb that is correctly conjugated, but with the wrong class pattern) results on models trained from scratch; Figure 2d shows the corresponding results for finetuned GPT-2. Unsurprisingly, we observe random chance performance (50%) on unseen verbs for both the randomly initialized and finetuned models–as the models cannot predict the correct class for verbs not seen during training.

More surprisingly, randomly initialized models are also unable to outperform random chance in the one-shot setting, suggesting that these models can not generalize knowledge about underlying verb classes during inference. While one-shot evaluations of the finetuned model outperform this in low-data settings (achieving $\sim 68\%$ accuracy), this is roughly what would occur if the model always chooses sentences where the prompt and test verb are identical (occurring $\sim \frac{1}{6}$ of the time across conjugations), and chooses randomly otherwise; this performance also occurs on the *per-num* test (Figure 2b). Thus, this behavior is likely caused by the pretrained GPT-2 exhibiting a strong copying preference (Olsson et al., 2022), but not generalizing beyond that.

On seen verbs, model performance again generally improves with more data, but we see smaller performance gaps across languages with different verb class counts, particularly at smaller data scales. Furthermore, model accuracy with more verb classes tends to be *higher* than those with fewer classes, though with more variation than observed with *per-num* probing. The discussion offers a possible hypothesis for this phenomenon.

4 Discussion

This paper investigates whether LMs exhibit linguistic priors for natural and unnatural conjugation paradigms. Our probing experiments find that LMs are much more efficient at modeling person-number agreement for languages with simpler verb

[3]E.g., Training from scratch on 100 sentences with one verb class achieves a mean accuracy of 76.2%, while it requires 10k sentences to get a similar accuracy over 64 classes.

[4]Limited generalization has been observed for other linguistic tasks in transformers (Liu and Hulden, 2022).

paradigms, mirroring human learning of languages. They also corroborate prior work indicating that neural LMs prefer human languages to unnatural ones (Alamia et al., 2020; Kallini et al., 2024). However, this primarily holds for verbs seen during training; models perform much worse at judging subject-verb agreement on novel verb roots, in contrast with the strong generalization shown by human speakers on this task (e.g. Berko, 1958).

Furthermore, we find that LMs adopt unnatural verb paradigms[5] almost as well, given enough training data. This result, in conjunction with degraded performance exhibited on unseen verbs, indicates that model learning of the generated languages is likely heavily dependent on **memorization** rather than **generalization** of the training data, particularly in the *class* behavioral test setting. While this trade-off has been documented in LMs for downstream NLP tasks (Tänzer et al., 2022; Zheng and Jiang, 2022), we find that it also affects the model when learning lower-level linguistic knowledge.

Even more unnaturally, models trained on languages with more complex paradigms are slightly *better* at identifying correct verb classes, with the best performance occurring on 32 and 64 classes. –far beyond what appears in most natural languages. We hypothesize that this behavior is due to how models and their inputs are parameterized: as the number of classes increases, the set of verb roots a suffix can follow (according to the training data) becomes smaller, allowing the model to be more confident about the bigram's conditional probability. However, this finding contrasts sharply with human language learning, where many unrelated paradigms are typologically improbable due to the unreasonable amount of memorization required for humans to model them correctly.

Based on these results, we argue that while language model learning of verbal paradigms may resemble human learning, the underlying mechanisms driving these behaviors are likely very different. Future work comparing model behavior with humans should control for these similarities by also looking at the underlying mechanisms driving model performance.

5 Limitations

Using carefully constructed artificial languages allows us to isolate syntactic complexity's effects on language learnability and to consider a broad, systematic complexity distribution. However, this means that these languages are not natural (particularly regarding the absence of semantics), which limits the findings presented here. Future work should replicate these experiments in a more natural setting to verify that our findings remain valid in such conditions.

Another limitation of this work is the size of the language models: computational limitations and the number of models considered in our experiments (800 trained LMs across experimental settings) limited the model size considered to one setting, GPT-2 Small (124M parameters). Finally, there are many aspects of complexity in natural language, with the number of verb classes being just one aspect. Whether our findings hold for other linguistic phenomena, such as noun classes (i.e. gender), freedom in word order, degree of syncretism, morphophonological alternations, etc. remains an open area for future research.

[5]I.e., more verb classes than in most natural languages.

Acknowledgments

We would like to thank Luke Zettlemoyer for feedback on early stages of this work. We would additionally like to thank Claire Bowern for access to her lab compute resources for model training in the later stages of this work.

References

Murtadha Ahmed, Saghir Alfasly, Bo Wen, Jamal Addeen, Mohammed Ahmed, and Yunfeng Liu. 2024. AlclaM: Arabic dialect language model. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 153–159, Bangkok, Thailand. Association for Computational Linguistics.

Andrea Alamia, Victor Gauducheau, Dimitri Paisios, and Rufin VanRullen. 2020. Comparing feedforward and recurrent neural network architectures with human behavior in artificial grammar learning. *Scientific Reports*, 10(1):22172. Publisher: Nature Publishing Group.

Catherine Arnett and Benjamin K Bergen. 2024. Why do language models perform worse for morphologically complex languages? *arXiv preprint arXiv:2411.14198*.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 2 (Short Papers), pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Casey Lew-Williams and Anne Fernald. 2007. Young children learning spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, 18(3):193–198. PMID: 17444909.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21:1112 – 1130.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Raphael Scheible, Johann Frei, Fabian Thomczyk, Henry He, Patric Tippmann, Jochen Knaus, Victor Jaravine, Frank Kramer, and Martin Boeker. 2024. GottBERT: a pure German language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21237–21250, Miami, Florida, USA. Association for Computational Linguistics.

Michael Tänzer, Sebastian Ruder, and Marek Rei. 2022. Memorisation versus generalisation in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7578.

Ianthi Tsimpli. 2014. Early, late or very late: Timing acquisition and bilingualism: A reply to peer commentaries. *Linguistic Approaches to Bilingualism*, 4.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Preprint*, arXiv:1406.5181.

Tatu Ylönen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 1317–1325. European Language Resources Association.

Xiaosen Zheng and Jing Jiang. 2022. An empirical study of memorization in nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6265–6278.

A Methodological Details

A.1 Artificial Language Generation

An overview of the pipeline used to generate sentences is described in Figure 3.

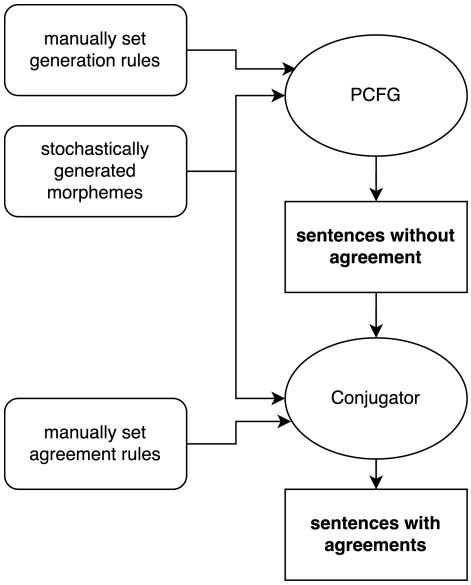


Figure 3: Overview of Sentence Generation with Artificial Languages.

A.1.1 Morpheme Generation

The parts of speech and number of morphemes that belong in each category are set manually. For this experiment, about 8000 morphemes were used in total. The number of morphemes for each part of speech was chosen to approximate 1% of the number of Kaikki's senses for each of English's part of speech (Ylönen, 2022). Approximately 0.5% of nouns were used instead of 1% since nouns disproportionately have technical vocabulary or jargon in Kaikki's database which we do not care to replicate. Additionally, only one determiner, which inflects for number, and six pronouns for each combination of number (singular, plural) and person (1st, 2nd, 3rd) were chosen to simplify conjugation paradigms.

A.1.2 Probabilistic Context Free Grammar

Probabilistic context-free grammars (PCFG) are an extension of a context-free grammar where each input state's production rules take a probability. Since the final result of a PCFG resembles a syntactic tree, it allows us to create sentences with customizable linguistic structures. Pseudocode for a simplified English grammar is provided in Figure 4. All generations start with an "S" state. Generation rules may apply categorically, such as rule (1), which determines that sentences produce a subject noun phrase followed by a verb phrase with probability 1. Other states may have two or more possible outcomes, as demonstrated by rule (3), which determines that noun phrases may produce a determiner followed by a noun or a pronoun with equal probability. All preterminal states are lowercase, while non-preterminal and non-terminal states are required to have at least one capital letter. This simplified grammar does not include adjectives or prepositions and thus is incapable of handling recursion, but the full grammar for the artificial languages does.

```
generation_rules = [
    S → [sNP, VP], 1
    sNP → [NP.nom], 1
    NP → [det, noun], 0.5, [pron], 0.5
    VP → [verb, NP], 0.7, [verb], 0.3
]
```

Figure 4: Generation Rules for a simplified grammar.

In order to handle conjugation, states are assigned features, as demonstrated by rule (2). Not demonstrated is the deletion of a feature and the addition of a tag feature that allows for long-distance agreement. Unless explicitly stated by a generation rule, a state passes all of its features to its children. This can be seen in Figure 1, where the subject NP with feature `nom` (short for nominative, i.e., subject of a sentence) passes on the feature to all of its children states.

Preterminal nodes are represented by the part of speech that will beget a morpheme from the vocabulary. A simplified vocabulary can be seen in Figure 5. In order to mimic the naturalistic distribution of words in human languages, generation rules for preterminal states function differently from the rest of the PCFG. The morpheme that is chosen by the preterminal state is chosen according to a Zipfian distribution with skew = 1.2. Additionally, the part of speech of a terminal node is added to its set of features.

Rules dictating universal features of certain parts of speech may be added at this step as well. For example, in our simplified toy vocabulary, all nouns are assumed to take the feature `3rd`. This is not included in the toy grammar's rules, but will be essential in determining which terminal states agree with which other terminal states.

```
vocabulary = {
    det: [“the”],
    noun: [“cat”, “dog”],
    verb: [“see”, “miss”],
    pron: [“you”]
}
```

Figure 5: Vocabulary for a simplified grammar. The morphemes in this example are not stochastically generated for a clearer example.

A.1.3 Conjugator

The conjugator works by (i) determining which morphemes agree with other morphemes and (ii) applying each inflection rule to each word in the sentence that has a given feature set.

After the PCFG generates a sentence without agreement, the first step is determining which lemmas agree with which other lemmas. We define agreement as a terminal state *copying* a state from another terminal state according to rules set by the user. For example, as demonstrated by Figure 6, we can see our toy grammar's verbs are required to copy one feature relating to number and one feature relating to person from the nominative noun constituent, as seen by the definition of `agreement_rules`. In the toy example in Figure 1, the verb *see* copies the feature `3rd` and `sg` (applied through rules not shown in previous figures) from *cat* to have the feature set `verb, 3rd, sg`. Note that if any word seeking agreement finds more than one word to agree with, or is unable to find exactly one of each of the features it aims to copy, generation fails.

The second step is applying inflections to the sentence, now that all words requiring conjugation have copied features from the word that they are agreeing with. Inflections apply to any word that changes form based on some features. For example, as demonstrated by Figure 6, we can see our toy grammar's verbs take a suffix *-s* iff it has features `3rd` and `sg` or otherwise it does not take any inflections, as seen by the definition of `conjugations`. Conjugations may also apply to words which did not gain features from the agreement rules, such as nouns pluralizing with the suffix *-s* if it gained the `pl` feature from a generation rule.

We generate datasets on 8 different numbers of verb classes: {1, 2, 3, 5, 8, 16, 32, 64}. In all datasets, any verb agrees with the subject of the sentence in person and number, for a total of 6

```
agreement_rules = [
    # Verbs agree with the nom nouny word
    # Verbs must then copy the word's number
    # Similarly, they then copy the person
    {"verb": [["nom", "nouny"],
              [["sg", "pl"],
               ["1st", "2nd", "3rd"]]]}
]
conjugations = [
    ["verb", {
       “-s”: 3rd.sg,
       “-”: otherwise
    }],
    ["noun", {
       “-s”: pl,
       “-”: otherwise
    }]
]
```

Figure 6: Pseudocode for agreement rules and conjugations for a simplified grammar. Nouny is defined as being either the feature `pron` or `noun`.

possible suffixes given a verb root (3 person x 2 number). In datasets with n verb classes where $n > 1$, verbs are assigned to one of n classes. Each of these classes has a unique set of suffixes for each combination of person and number for subject-verb agreement, for a total of $6n$ verbal suffixes per language.

B Replicability Details and Miscellanea

This section provides additional details on our experimental setting for documentation and replicability purposes.

The language models trained in these experiments have the GPT-2 Small architecture with 124.4M trainable parameters. We consider both a randomly initialized version of this architecture and the pretrained GPT-2 hosted through Huggingface,[6] which is released under the MIT license; intended use of this artifact beyond the license is not clearly stated.

In our experiments, we trained 800 models (10 runs, 8 languages per run, 5 models with varying training data amounts per language, 2 base models - English pretrained vs. randomly initialized), and the training dataset sizes ranged from 100 sentences to one million sentences; we evaluated each model

[6] https://huggingface.co/openai-community/gpt2, as accessed before and on 02/15/2025.

across six settings (two behavioral tests, three root settings) with 5,000 sentences each; though, models trained on languages with one verb class were not evaluated on the *class* behavioral test. There is no overlap in sentences from the training and test set. We use 10 GPUs for both training and evaluation, one for each run of 80 models. Training and evaluating 80 models took approximately one day per GPU. Given our training batch size of 1 and single training epoch, this corresponds to between 100 to 1,000,000 training passes per model.