The Syntactic Productivity of Large Language Models

Anonymous ACL submission

Abstract

Do Large Language Models (LLMs) produce output that exhibits syntactic productivity similar to human language? The problem is formally equivalent to a major issue in child language research where conclusions must be drawn about the underlying grammar solely on the basis of a child's production data. We apply a mathematically rigorous and independently validated benchmark to quantify syntactic productivity with specific focus on Determiner-Noun (D×N) combinations. Human language corpora show the statistical profile of syntactic productivity but LLM-generated texts do not.

1 Introduction

The success of LLMs has spurred significant research to understand their capacities for representing linguistic structures in comparison to human language learners and users.

A prominent approach has focused on the development and use of benchmarks to probe for specific linguistic properties in LLMs. These range from extracting structures from internal representations (e.g., Hewitt and Manning, 2019; Tenney et al., 2019; McCoy et al., 2020; Tucker et al., 2021; Papadimitriou et al., 2021), to building tasks inspired by psycholinguistic processing studies (e.g., Chowdhury and Zamparelli, 2018; Wilcox et al., 2018; Hu et al., 2020), to classic acceptability rating tasks that theoretical linguists use to infer grammatical knowledge (e.g., Linzen et al., 2016; Warstadt et al., 2020; Huebner et al., 2021; Sinclair et al., 2022). While the benchmarking approach has provided valuable insights into LLMs' linguistic capacity, they are by design limited to the specific structural properties identified by the researcher and may provide an insufficiently representative coverage of linguistic phenomena (McCoy et al., 2019; Vázquez Martínez, 2021; Wang et al., 2022; Guest and Martin, 2023; Vázquez Martínez et al., 2023). With the rise of generative AI models, it is increasingly important to develop evaluation methods for open-ended LLM output (Chang et al., 2024).

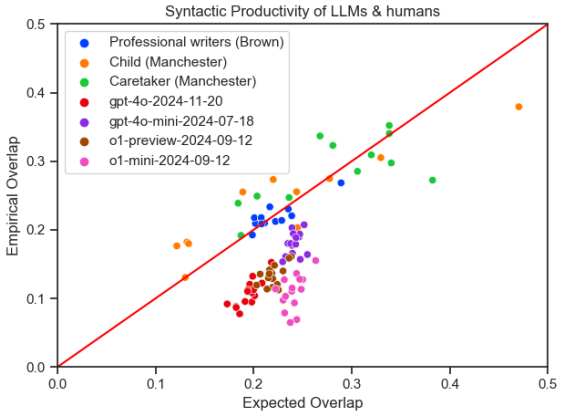

Figure 1: Syntactic productivity measure (overlap; Section 2) of human language corpora (children, their caretakers, and professional writers) and 4 LLM-generated corpora from the OpenAI API. Each point indicates a corpus. Human corpora show measures comparable to the expectations under a fully productive grammar (Section 3) but LLM corpora show significantly lower measures of productivity (Section 4). The red reference line indicates a perfect match between the two.

In this paper, we introduce a novel approach to LLM evaluation with specific focus on syntactic productivity.[1] Our approach draws inspiration from the study of child language, where researchers frequently need to assess a learner's underlying grammar based solely on a corpus of their language production. Section 2 reviews a well-established statistical test (Yang, 2013) with specific reference to syntactic productivity. Section 3 applies the test to child and caretaker speech in the CHILDES database (MacWhinney, 2000) as well as the Brown Corpus (Kučera and Fran-

[1] We will update this footnote with a link to the GitHub repository in the deanonymized version.

cis, 1967). Results show that, as expected, these human language samples exhibit the statistical hallmarks of productivity. In Section 4, the test reveals that LLM-generated narrative text fails to show the statistical properties of productivity. Section 5 discusses the implications of our findings and directions for future research.

2 A Statistical Test for Productivity

The defining feature of language is its infinite productivity, as new words and sentences can always be generated. Understanding the nature and development of productivity has been a central problem in linguistics, cognitive science and now AI.

A revealing method for uncovering productivity, as shown in the celebrated Wug test (Berko, 1958), is to provide the language learner with novel input and assess whether appropriate output forms can be generated. However, such experimental approaches have certain task-related complications that limit their applications. For example, while children learn the English past tense suffix (-ed) before age 3 as shown by occasional over-regularization errors (e.g., *goed*; Kuczaj 1977), not even first graders consistently produced -ed on the Wug test (Berko, 1958) as children often struggle learning and using a novel word in an artificially induced setting. Comprehension studies also carry extra cognitive demands. Even 4-year-olds fail to completely accurately distinguish the temporal reference of "was" and "is" in an experimental setting (Valian, 2006).

Hence, the investigation of early child language has often focused on children's naturalistic production, which is least subject to performance constraints while also providing the most accessible type of acquisition data. In particular, the combination of determiners (D) and nouns (N), or D×N for short, has been a major focus in child language research (Pine and Martindale, 1996; Valian et al., 2009; Pine et al., 2013). This is because determiners, especially singular determiners *the* and *a*,[2] are highly frequent and thus well represented in child language. Despite its simplicity, D×N fully exhibits the hallmark of syntactic productivity: Any singular noun used with *the* can also be used with *a*. A simple metric, dubbed *overlap* (Pine and Lieven, 1997), has been widely used to quantify productivity: the proportion of singular nouns used with both *the* and *a* out of those used with either. The overlap value is bounded between 0 and 1: A higher value would be stronger evidence for productivity, but as we will see shortly, this intuition needs to be qualified.

Many previous studies of D×N focus on the comparison of overlap values in children and their caretaker's language. However, any corpus of caretaker language is only a small sample of a learner's input data. Moreover, adults talk more and have larger vocabularies than children, so it has been difficult to develop "fair" comparisons across samples. A statistical test for syntactic productivity (Yang, 2013; Goldin-Meadow and Yang, 2017) sidesteps these issues. This test calculates the expected value of D×N overlap in a corpus under the assumption that D×N is fully productive i.e., statistically independent.

The statistical test for overlap builds on two key statistical properties of language, one universal and the other specific to D×N in English. First, the test assumes that the frequencies of words, especially open class words such as nouns, follow Zipf or inverse power law distribution (Zipf, 1949; Baroni, 2009). As such, if a corpus contains n unique nouns in D×N combinations, the noun with rank r has the expected probability $1/(r^a H_{n,a})$ where $H_{n,a}$ is the generalized harmonic number $\sum_{i=1\ldots n} 1/i^a$ with a as the exponent of inverse power law. In most cases a is approximately 1 following Zipf's original formulation but deviation from 1 can be accommodated in the calculation.

Second, it is observed that in D×N combinations, nouns tend to have a "favorite" determiner that combines far more frequently than the other. For example, *bathroom* greatly favors *the* over *a* but for *bath*, the reverse is true. This imbalance, referred to as *bias* (b), is defined as follows:

$$b = \frac{\sum_{i=1}^{n} \max(C_{\text{the}\times i},\ C_{\text{a}\times i})}{\sum_{i=1}^{n} (C_{\text{the}\times i} + C_{\text{a}\times i})} \qquad (1)$$

where $C_{\text{the/a}\times i}$ is the frequency of *the*/*a* combined with noun i. The bias value is not part of the grammar *per se* nor does it require learning: It is unlikely that children track the frequency of bodily functions ("the bathroom") or hygienic practices ("a bath"). Rather, the bias value is the vagaries of life reflected in language use. As *bath* and *bathroom* illustrate, not all nouns have the same favorite determiners. Situational factors may also skew the

[2]The phonological variant *an* is treated as *a* as it is an independent developmental process.

bias: a pediatrician will have more balanced use for *the* and *a* for the noun *baby* than the parent of a newborn. Nevertheless, as we show in Section 3, the bias value in aggregate is remarkably stable across samples of English at $b = 0.82$.

Taken together, these two statistical properties greatly enhance the applicability of the test; the full formula is given in Appendix A. For a corpus, one only needs S, the total number of D×N combinations, and n, the number of unique singular nouns. The exponent of Zipf's Law can be obtained from frequencies of the n nouns, and one can easily compute the expected overlap value and compare it to the empirical value. If there are no significant differences, one can conclude that the D×N combinations are in fact consistent with a fully productive grammar.

3 Syntactic Productivity in Humans

The first set of human language analysis is based on the Manchester corpus (Theakston et al., 2001). There are 12 dyads of typically developing children and their caretakers, and the transcripts are based on regular recording sessions between age 2 and 3. The Manchester Corpus is the largest longitudinal records of English language development for this age group and has been frequently used in child language acquisition.

Following previous work (Pine et al., 2013), a D×N combination is extracted if D is *the* or *a* and N is a singular noun that immediately follows D or with one non-noun intervening word. Data extraction used the spaCy dependency parser (Honnibal and Johnson, 2015) which also provides POS tagging of the transcripts. The statistical conclusions of our study remain unchanged if we use the POS annotation provided in CHILDES.

We found that in the Manchester Corpus, the nouns in both child and caretaker language show excellent fit for the original Zipf's Law with an average exponent of $a = 1.03$. Furthermore, as noted earlier, D×N combinations in English are heavily biased toward one of the two determiners. The bias value estimated from COCA based on Eq 1 is 0.82. Remarkably, the bias value across the 12 dyads of children and caretakers is almost identical (mean = 0.814, sd=0.03), and there is no significant difference between the bias value in child language samples and caretaker language samples (paired t-test p=0.612). Thus in all studies we have used the universal bias value $b = 0.82$ for expected overlap calculation. These values of a and b were used to calculate the expected overlap value. The results are shown in Figure 1 with additional details in Figure C1 (Appendix C). There are no statistically significant difference between expected and empirical values in the Manchester Corpus (paired t-test: $p = 0.334$ for children and $p = 0.733$ for caretakers).

The second set of human language analysis is based on the Brown Corpus (Kučera and Francis, 1967), a collection of professional print materials across a wide range of genres. To make suitable comparisons with the Manchester Corpus, we grouped successive files in the Brown Corpus into 12 samples. The D×N combinations were extracted with spaCy following the method used for Manchester Corpus. The nouns in each sample do not follow the canonical Zipf's Law with exponent of 1. Rather, the average exponent of the Brown Corpus samples is 0.771. We believe that this is due to the nature of the Brown Corpus, where each file is a relatively short document about a particular topic. Collectively, the most frequent nouns in each sample are much closer in frequency. By contrast, the speakers in the dialog samples in the Manchester Corpus had more focused and extensive conversations about fewer topic nouns. For the Brown corpus analysis, we used the exponent $a = 0.771$ along with the universal bias value $b = 0.82$ to calculate the expected overlap value in comparison to empirical values. Figure 1 summarizes the results with additional details in Figure C2 (Appendix C). Once again, the expected and the empirical overlap values are not statistically significantly different (paired t-test $p = 0.586$).

Note the overlap test is not limited to D×N but is applicable to any rule that combines a two-member closed class category with an open class category. To further establish the robustness of the test, we extracted the Manchester Corpus verb lemmas inflected with either -ed or -ing from the Manchester corpus: the overlap measures the proportion inflected with both. Note that -ed and -ing are not fully interchangeable due to irregular verbs. Thus, the empirical overlap for verb lemmas over -ed and -ing must be *lower* than the expected value, the latter of which is computed on the assumption of full interchangeability. Indeed, across the 24 dyad samples, the empirical values are significantly lower than the expected values (paired t-test $p < 0.001$). However, once the irregular verbs are removed, the empirical overlap value of verb lemmas for -ed

and -ing are not significantly different from the expected value across the 24 dyad samples (paired t-test $p = 0.852$) because the two suffixes are indeed fully interchangeable for regular verbs.

Taken together, the analyses of human language illustrate the robustness of the test for detecting both true positives of productivity such as adult usage in Manchester and Brown as well as true negatives, such as the counterfactual application to verbal inflection. We can examine whether LLMs constitute a true positive or a true negative of syntactic productivity.

4 LLMs Fail Productivity Test

To evaluate the syntactic productivity of LLMs, we obtained text generated by the four most advanced OpenAI models available to us at the time of writing: `gpt-4o-mini-2024-07-18`, `gpt-4o-2024-11-20`, `o1-mini-2024-09-12`, and `o1-preview-2024-09-12`.

For each model, we composed a set of 15 `NARRATIVE_TOPICS` spanning different genres (e.g., a science fiction story, an academic job talk, an economics survey, among others), each with three more follow up topics that keep the discourse coherent. To prompt the models, we constructed a list of `NARRATIVE_TEMPLATES` that can be filled in with each of the 15 topics and follow ups. We additionally included a `SYSTEM_PROMPT` that instructs the model to write as coherently and in as much detail as possible in order to pass the Turing test. This yielded 15 long-form narratives for each of the four OpenAI models.

As in Manchester Corpus, `D×N` combinations are extracted from LLM texts using spaCy. Empirical analysis shows that on average, the inverse power law exponent of the nouns across 60 texts is $a = 0.745$ – analogous to that in the Brown corpus — which is used in the expected overlap calculation. We first used the human universal bias value $b = 0.82$ to calculate the expected value of `D×N` overlap for the LLMs in comparison to the empirical values. The results are summarized in Figure 1 with additional details in Figure C3 (Appendix C). The expected values are significantly higher than the empirical values (paired t-test $p < 0.001$ for all four models). The LLM text showed a higher average bias value (0.92) than human texts but $b = 0.92$ still resulted in expected values significantly higher than the empirical values ($p < 0.05$ for all four models). We thus conclude that unlike human language learners and users, LLMs do not generate `D×N` combinations in a fully productive way.

5 Related and Future Work

Our work is most relevant to efforts on AI-generated text detection, as current commercial solutions tend to operate at unadvisable False Positive rates, perform poorly on out-of-sample data (i.e. from a different generator) and are susceptible to adversarial attacks (Dugan et al., 2024, 2025). While high proportion of human participants are at chance in discriminating between human and AI text (Jannai et al., 2023; Jones and Bergen, 2024; Clark et al., 2021), there is high variance in participant performance, as there are outliers in the Real or Fake Text (RoFT; (Dugan et al., 2023)) dataset who perform well above chance, as well as participants who improve significantly on the task. A recent study also suggests that one factor to consider is the level of exposure an individual has had to AI-generated text, as annotators who frequently used LLMs for writing-related tasks were able to reliably identify AI-generated text despite adversarial modifications to make them seem more human-like (Russell et al., 2025). Therefore, AI-generated text does have certain distinct profiles. Along with methods that make use of syntactic templates (Shaib et al., 2024), hierarchical parse trees and discourse motifs (Kim et al., 2024a), etc., our work can be seen as a formal and quantitative metric grounded in combinatorial productivity, the fundamental property of human language. Approaches that incorporate linguistic features may lead to more robust and accurate AI-text detection.

While our results point to a significant difference in syntactic productivity between humans and LLMs, it is difficult to ascertain the nature of such discrepancies. A possibility may be LLMs' over reliance on the memorization of lexically specific combinations (Juzek and Ward, 2025). It is a mathematical fact that memorization and retrieval of `D×N` combinations in the input will necessarily reduce the overlap value in the output text. Even if a noun is combined with both *the* and *a* in the input, retrieving these combinations as holist collocations will always incur a positive probability that only one determiners is included in the output. We plan to focus on this issue in future research: Testing the statistical productivity of additional combinatorial processes that meet the criteria of statistical independence and full interchangeability.

6 Limitations

While the productivity test can be applied to many combinatorial processes, it has two inherent limitations. First, the closed class category can only have two members (e.g., *the* and *a* in D). Adding more members (e.g., *this* and *that*) makes the mathematical formulation intractable. Second, the test assumes that the categories combine in fully interchangeable and thus statistically independent ways. While processes such as those studied in the present paper can be characterized as such, this is not the case for all rules in language, at least not in a way than lends readily to the test. For example, not all transitive verbs can passivize ("John resembles Bill" cannot be passivized as "*Bill was resembled by John"), not all dative verbs can appear in both the double object construction (*tell* but not *say*) and the *to*-dative construction (*tell* but not *ask*). In addition, the choice of syntactic process to test must still be decided by the researcher as one needs to know the "ground truth": Which process is genuinely productive and can be subjected to the stringent definition of productivity pursued here.

More practically, the volume of the text needed to achieve statistically significant results is modest but not trivial. The test requires at least 1,000 D×N combinations in each sample, which in turn may require tens of thousand of words in the source text. For each of 60 samples generated by the OpenAI models, we needed a minium of roughly 1,000 lines of text after significant efforts to supply coherent prompts and keep the models both on topic and stop them from repeating text they had already generated. In a setting where one may want to find out whether the source of a particular text was AI or human, 1000+ lines of text are rare to come by, unless the text in question were a whole novel. Therefore, the utility of our test as a tool for text detector is currently quite limited.

Finally, we must acknowledge the limitations of our prompting and text generation methods. We wrote all prompt topics by hand in order to ensure diversity of theme and genre. More diversity, more prompt topics, or perhaps more followups to the topics could have been collected, with or without the assistance of AI, to ensure more generalizable conclusions. Yet the cost incurred to produce the final dataset exceeded $500. We make our data publicly available in the hopes that it be useful to other researchers who study linguistic phenomena in long-form AI-generated text.

References

Marco Baroni. 2009. Chapter 37: Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*, pages 803–822. Mouton de Gruyter.

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

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or Fake Text?: Investigating Human Ability to Detect Boundaries between Human-Written and Machine-Generated Text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12763–12771. Number: 11.

Liam Dugan, Andrew Zhu, Firoj Alam, Preslav Nakov, Marianna Apidianaki, and Chris Callison-Burch. 2025. GenAI content detection task 3: Cross-domain machine generated text detection challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 377–388, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Susan Goldin-Meadow and Charles Yang. 2017. Statistical evidence that a child can create a combinatorial linguistic system without external linguistic input: Implications for language evolution. *Neuroscience and Biobehavioral Reviews*, 81(Part B):150 – 157.

Olivia Guest and Andrea E. Martin. 2023. On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*, 6(2):213–227.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. 2023. Human or Not? A Gamified Approach to the Turing Test. *arXiv preprint*. ArXiv:2305.20010 [cs].

Cameron R. Jones and Benjamin K. Bergen. 2024. People cannot distinguish GPT-4 from a human in a Turing test. *arXiv preprint*. ArXiv:2405.08007 [cs].

Tom S Juzek and Zina B. Ward. 2025. Why does ChatGPT "delve" so much? exploring the sources of lexical overrepresentation in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411, Abu Dhabi, UAE. Association for Computational Linguistics.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024a. Threads of Subtlety: Detecting Machine-Generated Texts Through Discourse Motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024b. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

Stan A. Kuczaj. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*, 3rd edition. Lawrence Erlbaum, Mahwah, NJ.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Julian M Pine, Daniel Freudenthal, Grzegorz Krajewski, and Fernand Gobet. 2013. Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3):345–360.

Julian M Pine and Elena VM Lieven. 1997. Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, 18(2):123–138.

Julian M Pine and Helen Martindale. 1996. Syntactic categories in the speech of young children: The case of the determiner. *Journal of child language*, 23(2):369–395.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. *arXiv preprint*. ArXiv:2501.15654 [cs].

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement

of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Herbert S Terrace, Laura-Ann Petitto, Richard J Sanders, and Thomas G Bever. 1979. Can an ape create a sentence? *Science*, 206(4421):891–902.

Anna Theakston, Elena Lieven, Julian Pine, and Caroline Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, 28:127–52.

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if this modified that? syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online. Association for Computational Linguistics.

Virginia Valian. 2006. Young children's understanding of present and past tense. *Language Learning and Development*, 2(4):251–276.

Virginia Valian, Stephanie Solt, and John Stewart. 2009. Abstract categories or limited-scope formulae? The case of children's determiners. *Journal of Child Language*, 36(4):743–778.

Héctor Javier Vázquez Martínez. 2021. The acceptability delta criterion: Testing knowledge of language using the gradience of sentence acceptability. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 479–495, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Héctor Javier Vázquez Martínez, Annika Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729, Seattle, United States. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Charles Yang. 2013. Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, 110(16):6324–6327. Publisher: Proceedings of the National Academy of Sciences.

George Kingsley Zipf. 1949. *Human Behavior And The Principle Of Least Effort*.

A The Productivity Test

The following describes the statistical test for productivity (Yang, 2013).

The basic idea behind the productivity test is straightforward. Consider the combination of the determiner category D with two members (*the* and *a*) and the noun category N with n members. As discussed in Section 2, we assume that the n members follow Zipf's Law with exponent a: the expected probability of the member with rank r is:

$$p_r = \frac{1}{r^a H_{n,a}} \text{ where } H_{n,a} = \sum_{i=1}^{n} \frac{1}{i^a}$$

Suppose there are S combinations of D×N. The expected overlap value of the r-th ranked noun, or E_r, is

$$\begin{aligned} E_r &= 1 - (1 - p_r)^S \\ &\quad - [(b * p_r + 1 - p_r)^S - (1 - p_r)^S] \\ &\quad - [((1 - b) * p_r + 1 - p_r)^S - (1 - p_r)^S] \end{aligned}$$

where b is the bias value described in the text and repeated here:

$$b = \frac{\sum_{i=1}^{n} \max(C_{\text{the}\times i},\ C_{\text{a}\times i})}{\sum_{i=1}^{n} (C_{\text{the}\times i} + C_{\text{a}\times i})} \qquad (2)$$

And the expected overlap value for all n members in the D×N combinations is:

$$E = \frac{1}{n}\sum_{i=1}^{n} E_r$$

The syntactic productivity test is not limited to determiners and nouns but can be applied to any two combinatorial categories, as long as the closed class category has only two members and the open class category frequency can be approximated by Zipf's Law. Moreover, it can be applied to detect both the presence and absence of productivity. For example, Goldin-Meadow and Yang (2017) adapted the test to the combinatorial structure of homesign, the gestural system created by deaf children in the absence of sign language input. The test finds that homesign combinations are fully productive, providing independent evidence for traditional behavioral analysis. On the other hand, the test has been applied to the ASL sign combinations produced by Nim Chimpsky. Results show that Nim's sign combinations show considerably less diversity than would be expected under a fully productive system, again supporting conclusions based on frame-by-frame sign analyses (Terrace et al., 1979).

B Long-form AI Text Generation

In order to accurately evaluate the syntactic productivity of the LLMs, we needed a sample of text from each model whose raw count of D×N pairs (S) and unique nouns (N) is comparable to that of the human data we use as a baseline. While we would most easily obtain AI-generated text from previously generated detection tasks, these generally consist of short documents between 200 and 500 tokens (Kim et al., 2024b). We therefore need to generate multiple long-form texts of at least 1,000 lines or more for each LLM under evaluation. The method was described in the text.

C Detailed Results of Syntactic Productivity Analyses

The syntactic productivity scores for children and their mothers in the CHILDES Manchester corpus are plotted together in Figure C1. The syntactic productivity measures of the 12 Brown Corpus samples are plotted in Figure C2. The syntactic productivity measures of the narrative texts generated by four OpenAI models are shown in Figure C3.

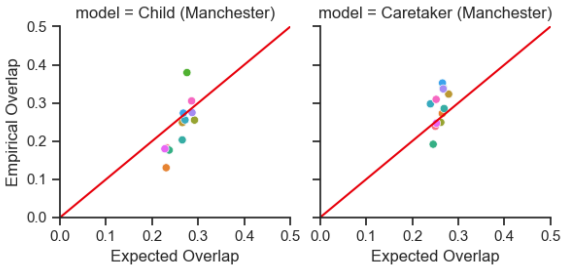

Figure C1: Scatter plot of expected and empirical productivity measure (D×N overlap) for the 12 children and their corresponding caretakers from the Manchester Corpus (Theakston et al., 2001). No statistically significant difference is found (paired t-test $p = 0.334$ children and $p = 0.771$ for caretakers).

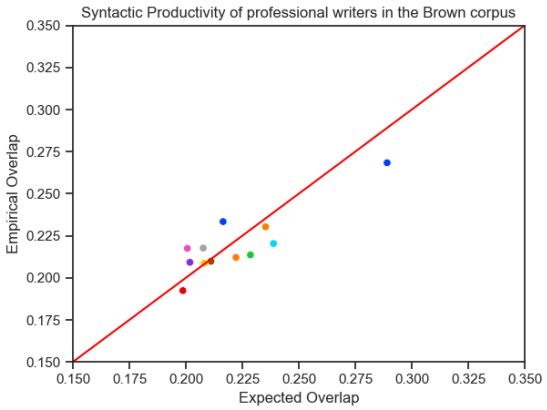

Figure C2: Scatter plot of expected and empirical productivity measured (D×N overlap) for 12 sections of the Brown corpus (Kučera and Francis, 1967). No statistically significant difference is found (paired t-test $p = 0.562$.

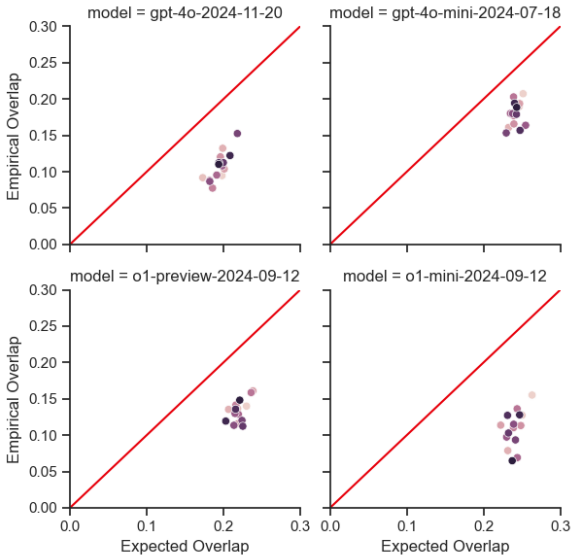

Figure C3: Scatter plot of expected and empirical productivity measures (D×N overlap) for 15 samples of narrative texts generated by OpenAI models. The empirical values of overlap are considerably lower than the expected values under full productivity (paired t-test, $p < 0.001$).