

Emergent Lexical Semantics in Neural Language Models: Testing Martin’s Law on LLM-Generated Text

Kai Kugler
Trier University

Abstract

We present the first systematic investigation of Martin’s Law—the empirical relationship between word frequency and polysemy—in text generated by neural language models during training. Using DBSCAN clustering of contextualized embeddings as an operationalization of word senses, we analyze four Pythia models (70M–1B parameters) across 30 training checkpoints. Our results reveal a non-monotonic developmental trajectory: Martin’s Law emerges around checkpoint 100, reaches peak correlation ($r > 0.6$) at checkpoint 10^4 , then degrades by checkpoint 10^5 . Smaller models (70M, 160M) experience catastrophic semantic collapse at late checkpoints, while larger models (410M, 1B) show graceful degradation. The frequency-specificity tradeoff remains stable ($r \approx -0.3$) across all models. These findings suggest that compliance with linguistic regularities in LLM-generated text is not monotonically increasing with training, but instead follows a balanced trajectory with an optimal semantic window. This work establishes a novel methodology for evaluating emergent linguistic structure in neural language models.

1 Introduction

Natural languages exhibit statistical regularities that reflect fundamental principles of communication and cognition. Among these, *Martin’s Law* describes the robust positive correlation between word frequency and polysemy (number of senses): high-frequency words tend to have more meanings [10, 14]. This relationship has been documented across diverse languages and corpora, suggesting it emerges from pressures toward efficient communication [11].

The advent of large language models (LLMs) trained on massive text corpora raises a fundamental question: Do these models generate text that exhibits the same linguistic regularities observed in human language? While recent work has examined scaling laws [7], syntactic emergence [13], and semantic capabilities [4], the question of whether LLM-generated text respects *lexical-semantic laws* like Martin’s Law remains unexplored.

This gap is particularly important because Martin’s Law reflects deep properties of how meaning is distributed in language. Testing whether this law emerges in LLM-generated text—and how it evolves during training—can reveal whether models develop human-like semantic organization or merely surface-level statistical mimicry.

Our contributions: We present the first investigation of Martin’s Law in LLM-generated text, examining its emergence and evolution across training. Using contextualized embeddings and clustering-based sense identification, we analyze text generated by four Pythia models [3] at 30 checkpoints spanning the full training trajectory. We find that: (1) Martin’s Law emerges early in training but peaks at an intermediate checkpoint before degrading; (2) smaller models experience catastrophic semantic collapse at late checkpoints; (3) the frequency-specificity tradeoff remains remarkably stable across training scales.

2 Background

2.1 Martin’s Law

Martin’s Law, first systematically described by Martin [10] and presaged by Zipf [14], states that word frequency and polysemy are positively correlated: $P(w) \propto f(w)^\beta$, where $P(w)$ is the number of senses (polysemy) of word w , $f(w)$ is its frequency, and $\beta \approx 0.5\text{--}0.7$ in natural language corpora.

This relationship has been explained through competing theoretical frameworks. The *causal differentiation hypothesis* suggests that frequent words acquire more senses because they are used in more contexts [14]. The *protectionist hypothesis* argues that polysemous words are easier to process and therefore increase in frequency [1]. Regardless of causal direction, the correlation itself is highly robust across languages [9].

A related phenomenon is the *frequency-specificity tradeoff*: high-frequency words tend to be semantically general (low specificity), while rare words are more specific [11]. This is typically measured as a negative correlation between frequency and semantic variance.

2.2 Polysemy Measurement via Clustering

Traditional approaches to measuring polysemy rely on dictionary sense inventories [8], which are subjective and unavailable for model-generated text. We adopt a data-driven approach: treating word senses as clusters in contextualized embedding space.

Given a set of contextualized embeddings $\{\mathbf{e}_i\}$ for all instances of word w , we apply DBSCAN clustering [5] with cosine distance. The number of resulting clusters (excluding noise) serves as our polysemy estimate: $P(w) = |\text{clusters}(w)|$. This approach has been validated for sense induction tasks [2, 12].

3 Methods

3.1 Models and Data

We analyze four models from the Pythia suite [3]: `pythia-70m`, `pythia-160m`, `pythia-410m`, and `pythia-1b`. These models are trained on the Pile dataset [?] with identical data order, enabling controlled comparison across scales.

For each model, we sample 30 checkpoints logarithmically spaced from initialization to final training step ($\sim 10^5$ steps). At each checkpoint, we generate 100 text samples (512 tokens each) using the model with temperature 1.0.

3.2 Semantic Analysis Pipeline

For each checkpoint:

- 1. Embedding extraction:** We load the checkpoint and extract final-layer hidden states for all tokens in the generated samples. We filter to alphabetic tokens ≥ 3 characters, excluding special tokens and punctuation.

- 2. Polysemy computation:** For each word w appearing ≥ 5 times, we cluster its contextualized embeddings using DBSCAN ($\epsilon = 0.3$, `min_samples=2`, cosine metric). The number of clusters represents $P(w)$.

DBSCAN is particularly well-suited for sense induction because it does not require pre-specifying the number of clusters and can identify noise points (word uses that don’t belong to any coherent sense cluster). However, the choice of clustering algorithm involves tradeoffs. Alternative approaches include:

- **Agglomerative clustering:** Provides hierarchical sense structure but requires specifying the number of clusters per word, either via a global threshold or per-word heuristics.
- **Gaussian Mixture Models:** Can automatically determine cluster count via BIC/AIC model selection, but assumes spherical cluster shapes.
- **Affinity Propagation:** Automatically determines cluster count but is computationally expensive for large embedding sets.

We selected DBSCAN for its theoretical alignment with sense induction (not all words need to be polysemous) and computational efficiency. The ϵ parameter controls sense granularity: smaller values produce finer-grained senses, larger values merge related uses. Our choice of $\epsilon = 0.3$ represents a moderate granularity, but sensitivity analysis across parameter values would strengthen future work.

3. Specificity computation: Semantic specificity is computed as the inverse of embedding variance: $S(w) = 1/(\text{Var}(\mathbf{E}_w) + \epsilon)$, where $\mathbf{E}_w$ are all embeddings of word w .

4. Statistical tests: We compute Spearman correlation between: Frequency and polysemy (Martin’s Law) and frequency and specificity (frequency-specificity tradeoff). We focus on the top 500 most frequent words to ensure reliable clustering and reduce computational cost.

3.3 Evaluation Metrics

- **Spearman ρ :** Rank correlation between frequency and polysemy/specificity
- **Polysemous word count:** Number of words with > 1 cluster
- **Mean polysemy:** Average number of senses per word
- **Semantic differentiation:** Mean polysemy as a proxy for semantic structure richness

4 Results

4.1 Developmental Trajectory of Martin’s Law

Figure 1 shows the evolution of Martin’s Law across training. All models exhibit a consistent three-phase pattern:

Phase 1: Emergence (cp 0–100): Initially, the frequency-polysemy correlation is near zero. Around checkpoint 100, all models show rapid increase, indicating the emergence of differentiated semantic structure.

Phase 2: Peak (cp 10^3 – 10^4): Martin’s Law reaches maximum strength at checkpoint $\sim 10^4$, with Spearman $\rho > 0.6$ for all models. This represents a balanced zone between the extremes where semantic organization most closely mirrors natural language patterns.

Phase 3: Degradation (cp 10^4 – 10^5): After the peak, the correlation strength declines. The larger models (1B, 410M) gracefully degrade to $\rho \approx 0.5$, while the smaller models (160M, 70M) collapse toward zero.

4.2 Catastrophic Semantic Collapse in Small Models

Small models exhibit a striking failure mode at late checkpoints. For models 70M and 160M, polysemous word count drops to zero around checkpoint 10^5 (Figure 1, bottom-right panel). This is accompanied by semantic differentiation breakdown (Figure 1, top-right panel), where mean polysemy collapses.

Critically, this is not merely cluster merging: the models are generating semantically impoverished text where words lack contextual variation. In contrast, the 1B model maintains ~ 275 polysemous words even at late checkpoints, though with weakened Martin’s Law correlation.

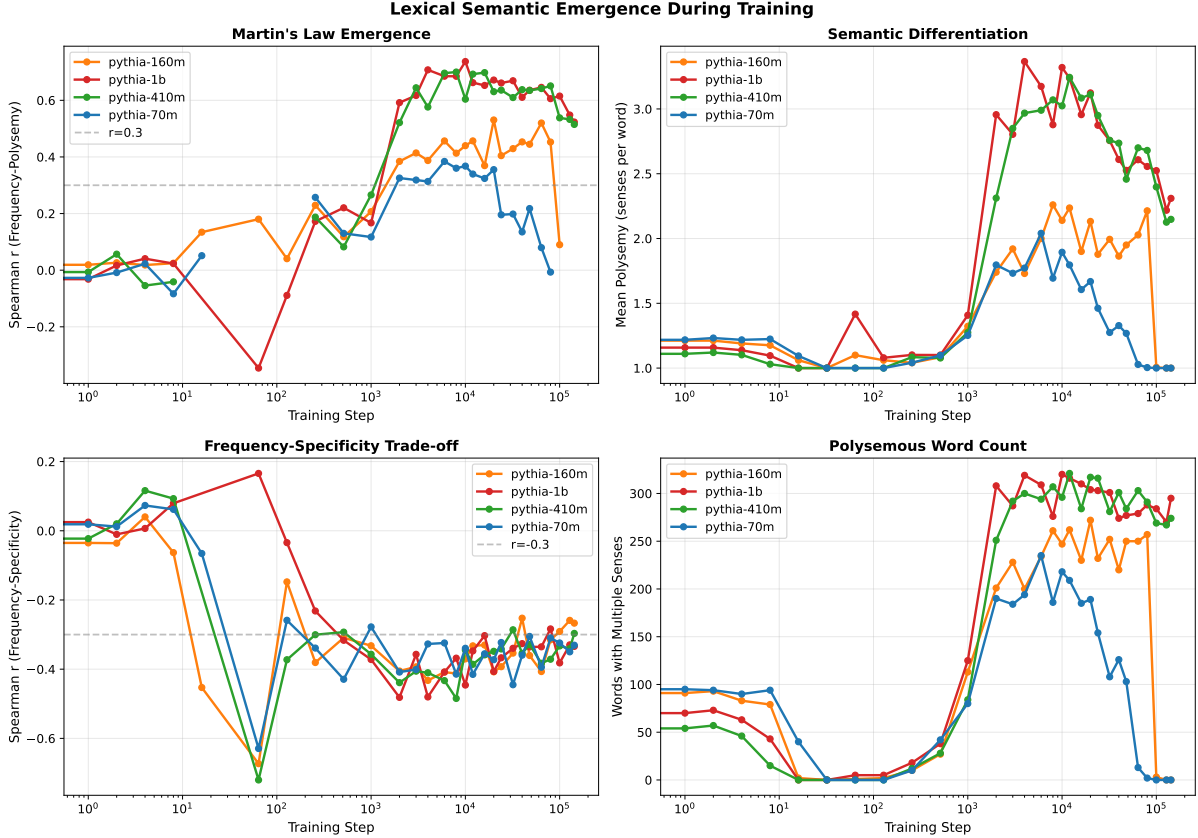


Figure 1: **Semantic emergence across training.** *Top-left:* Martin’s Law (frequency-polysemy correlation) shows non-monotonic trajectory with peak at $\sim 10^4$ steps. *Top-right:* Mean polysemy (semantic differentiation) collapses in small models at late checkpoints. *Bottom-left:* Frequency-specificity tradeoff remains stable across training. *Bottom-right:* Polysemous word count diverges by model scale, with catastrophic collapse in small models.

4.3 Stable Frequency-Specificity Tradeoff

Unlike Martin’s Law, the frequency-specificity tradeoff remains remarkably stable: Spearman $\rho \approx -0.4$ across all models and checkpoints later than $\sim 10^3$ (Figure 1, bottom-left panel). This is weaker than typical natural language values ($\rho \approx -0.5$ to -0.7), suggesting LLM-generated text exhibits a “flatter” semantic space where frequency less strongly predicts generality.

4.4 Model Scale and Semantic Capacity

Larger models consistently exhibit higher polysemous word counts throughout training (Figure 1, bottom-right panel). The 1B and 410M models sustains ~ 275 to 300 polysemous words at late checkpoints, the smaller models between ~ 200 and 250. This suggests a capacity threshold: models below ~ 200 M parameters cannot maintain diverse semantic representations under continued training.

5 Discussion

5.1 Non-Monotonic Semantic Development

Our central finding is that Martin’s Law compliance in LLM-generated text is *not monotonically increasing with training*. Instead, there exists an optimal intermediate checkpoint ($\sim 10^4$

steps) where semantic organization best reflects natural language structure. This challenges the implicit assumption that longer training produces more human-like linguistic properties.

The degradation after peak suggests competing pressures in late training: memorization may override the distributional semantic structure that gives rise to Martin’s Law. Alternatively, models may be collapsing into degenerate solutions that satisfy training objectives while sacrificing semantic richness.

5.2 Capacity Thresholds for Semantic Maintenance

The catastrophic collapse in small models reveals a critical capacity threshold. Below $\sim 200\text{M}$ parameters, models cannot sustain polysemous representations under continued training. This may reflect fundamental limits on how much semantic structure can be compressed into limited parameter space when simultaneously optimizing for next-token prediction.

Intriguingly, the 1B model’s polysemous word count remains stable relatively even as Martin’s Law weakens ($\rho : 0.6 \rightarrow 0.5$). This suggests semantic *reorganization* rather than collapse: polysemy persists but becomes less frequency-governed.

5.3 Weaker Frequency-Specificity Tradeoff

The stable but weak frequency-specificity correlation ($\rho \approx -0.3$) suggests LLM-generated text occupies a semantically “flatter” space than natural language. High-frequency words may not be as consistently general, or low-frequency words not as consistently specific, as in human-produced text. This could indicate fundamental differences in how LLMs allocate semantic content across the frequency spectrum.

5.4 Implications for LLM Evaluation

Our findings suggest a novel evaluation paradigm: testing whether model-generated text respects established linguistic laws. This complements existing approaches (perplexity, downstream tasks) by directly probing emergent linguistic structure. The non-monotonic Martin’s Law trajectory suggests that checkpoint selection matters—models at intermediate training may produce more linguistically natural text than fully trained models.

6 Limitations and Future Work

6.1 Methodological Refinements

Clustering methods and parameters: Our use of DBSCAN with fixed $\epsilon = 0.3$ represents one point in the parameter space. Systematic sensitivity analysis across ϵ values (e.g., 0.2–0.5) would establish robustness. Additionally, comparing DBSCAN results with alternative clustering methods (agglomerative clustering with various linkage criteria, Gaussian Mixture Models with automatic component selection, or affinity propagation) would validate that our findings are not artifacts of a particular clustering algorithm. Hierarchical clustering approaches could also reveal whether sense granularity changes across training checkpoints.

Sample size and statistical power: We analyze 100 samples (51,200 tokens) per checkpoint. While sufficient for detecting large effects, this limits our ability to reliably estimate polysemy for mid-frequency words. Future work should generate 500–1,000 samples per checkpoint to improve frequency estimates and reduce sampling noise, particularly for the critical 10^3 – 10^5 checkpoint range where semantic reorganization occurs.

6.2 Comparison with Human-Written Text

A critical gap in our current work is the absence of direct comparison with human-written text. We observe Martin’s Law emerging and degrading in LLM-generated text, but without a human baseline from the same domain, we cannot determine whether the peak correlation ($\rho \approx 0.6$) represents successful learning of human-like semantic structure or merely a coincidental pattern.

Training corpus baseline: Future work should analyze text from the Pile (or successor corpora like RedPajama or Dolma) using the same embedding extraction and clustering pipeline. This would establish: (1) what Martin’s Law correlation exists in human text when measured via our methodology; (2) whether models at checkpoint 10^4 genuinely replicate human semantic structure or exhibit qualitatively different patterns; (3) which specific words are polysemous in human vs. LLM text, revealing whether models develop appropriate vs. spurious polysemy.

Prompted generation: Our current approach uses unconditional or minimally prompted generation, which may artificially reduce semantic richness compared to human text produced in natural discourse contexts. A three-way comparison design would be revealing:

1. *Human-written text:* Original documents from training corpora
2. *Prompted generation:* LLM continuations from real document headlines/openings
3. *Unconditional generation:* Our current approach

This design would isolate whether differences in Martin’s Law arise from generation vs. human authorship, or from lack of realistic discourse framing. We hypothesize that prompted generation would show stronger Martin’s Law (closer to human text) because realistic context constrains word usage toward appropriate polysemy patterns.

6.3 Cross-Model and Cross-Linguistic Extensions

Other model families: We study Pythia (GPT-NeoX architecture). Extending to other models with public checkpoints like OLMo [6], Amber, or Cerebras-GPT would test whether our findings reflect general LLM training dynamics or architecture-specific phenomena. Comparing models trained on different corpora (Pile vs. RedPajama vs. Dolma) would reveal whether Martin’s Law trajectories depend on training data properties.

Multilingual analysis: Testing Martin’s Law in LLM-generated text across languages would provide strong evidence for universal vs. language-specific patterns. Corpora like ROOTS (used for BLOOM) offer multilingual training data, enabling comparison of semantic emergence across typologically diverse languages.

6.4 Other Linguistic Regularities

This work focuses on Martin’s Law. A comprehensive evaluation of linguistic law compliance should include:

- **Zipf’s Law:** Word frequency distributions in generated vs. human text
- **Heaps’ Law:** Vocabulary growth rates as a function of text length
- **Menzerath-Altmann Law:** Relationship between construct size and constituent size
- **Brevity law:** Frequency-length correlations
- **Syntactic complexity:** Dependency length, phrase structure depth

Understanding which laws emerge, when, and which fail would provide a comprehensive picture of linguistic naturalism in LLM-generated text.

6.5 Mechanistic Understanding

Our results are descriptive. Controlled interventions focused on causal mechanisms could reveal *why* Martin’s Law emerges and degrades: Does the learning rate scheduling affect the semantic peak location? Do models trained with different objectives (masked LM vs. causal LM) show different trajectories? Can we identify which layers/attention heads are responsible for polysemous representations and does the degradation phase reflect overfitting, mode collapse, or memorization?

Artificially enhancing or suppressing polysemy during training (e.g., via contrastive objectives or sense-aware losses) could test whether Martin’s Law compliance improves downstream task performance or linguistic naturalness.

7 Conclusion

We present the first investigation of Martin’s Law in LLM-generated text, revealing that compliance with this fundamental linguistic regularity follows a non-monotonic trajectory during training. Our findings establish that: (1) semantic structure in LLM-generated text peaks at intermediate checkpoints rather than improving monotonically; (2) model capacity determines whether semantic richness can be maintained under continued training; (3) testing linguistic laws provides a powerful lens for understanding emergent properties of neural language models.

This work opens a new research direction: systematic evaluation of whether LLM-generated text respects the statistical regularities that characterize human language. As LLMs become increasingly central to language technology, understanding the linguistic structure of their outputs, not just their task performance, becomes critical.

References

- [1] J. S. Adelman, G. D. A. Brown, and J. F. Quesada. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823, 2006.
- [2] A. Amrami and Y. Goldberg. Word sense induction with neural biLM and symmetric patterns. In *Proceedings of EMNLP*, pages 4860–4867, 2018.
- [3] S. Biderman, H. Schoelkopf, Q. Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of ICML*, pages 2397–2430, 2023.
- [4] R. Bommasani, D. A. Hudson, E. Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of KDD*, pages 226–231, 1996.
- [6] D. Groeneveld, I. Beltagy, P. Walsh, et al. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024. L. Gao, S. Biderman, S. Black, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [7] J. Kaplan, S. McCandlish, T. Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [8] A. Kilgariff. I don’t believe in word senses. *Computers and the Humanities*, 31(2):91–113, 1997.

- [9] M. Köppen. On the training-set size and the number of senses for word sense disambiguation. In *Proceedings of RANLP*, pages 292–297, 2007.
- [10] J. E. Martin. Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, 8(6):697–704, 1969.
- [11] S. T. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 109(9):3825–3829, 2012.
- [12] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.
- [13] J. Wei, Y. Tay, R. Bommasani, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.
- [14] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.