The Narrow Corridor: Constrained Dynamics of Word Embedding Training

Maximilian Slibar

github.com/MaxiSlibar

January 2026

Abstract

We introduce Curvature Flow Decomposition (CFD), a method that applies Frenet-Serret curvature analysis to individual word embedding trajectories during training. Analyzing 1.05 million training steps across 50 runs of a Skip-gram model, we find a near-deterministic relationship ($r = -0.990$) between word frequency and trajectory geometry: frequent words follow curved, inefficient paths while rare words move geodesically. Crucially, this effect is driven by frequency, not word class—the article "den" (low frequency despite being a function word) behaves like content words. We further show that 98% of possible state transitions never occur, suggesting training follows a narrow corridor rather than exploring parameter space freely. These findings provide a geometric lens on why neural networks learn what they learn.

Code and data: https://github.com/MaxiSlibar/the-narrow-corridor

1 Introduction

Neural networks learn by gradient descent, but what does "learning" actually look like? The standard narrative is one of exploration: the optimizer searches through parameter space, gradually finding configurations that minimize loss. This framing suggests flexibility, possibility, and choice.

We present evidence for a different view. By tracking individual word embeddings through every step of training—1.05 million steps across 50 runs—we find that the learning process is far more constrained than it appears. The key findings:

Frequency determines geometry. Word frequency predicts trajectory shape with $r = -0.990$ correlation. High-frequency words follow curved, inefficient paths through embedding space. Low-frequency words move almost geodesically. This is not a word class effect: the article "den" (low frequency in our corpus) behaves like content words, not like other articles. Frequency, not grammatical function, determines how a word moves during training.

98% of transitions never occur. We discretize the embedding space into 100 regions and build a transition matrix across all training runs. Of 10,000 possible state transitions, only 199 ever happen. The system doesn't explore—it follows a narrow corridor, constrained by the interaction of data statistics and architecture.

Training has discrete phases. Three independent analyses—bifurcation detection, temporal reversal entropy, and forbidden state sequences—converge on the same picture: training proceeds through irreversible phase transitions, not continuous improvement. There are moments after which the system is fundamentally different and cannot return.

These findings reframe how we think about neural network training. The optimizer is not searching; it is following. The structure of the data and the architecture together define a channel, and gradient descent flows through it. Learning is less exploration, more inevitability.

We make three contributions:

1. **Curvature Flow Decomposition (CFD)**: A method that applies Frenet-Serret differential geometry to word embedding trajectories, revealing the frequency-geometry relationship.
2. **Forbidden State Sequence Analysis (FSSA)**: A framework for measuring what training *doesn't* do—the transitions that never occur—providing a constraint-based view of learning dynamics.
3. **Empirical evidence from 17 independent analyses**: Convergent findings from differential geometry, information theory, graph theory, and topology, all pointing to the same conclusion.

2 Related Work

Word Embeddings. Distributed representations of words have become fundamental to NLP since Word2Vec [Mikolov et al., 2013] and GloVe [Pennington et al., 2014]. Prior work has extensively analyzed trained embeddings—their linear structure, analogical reasoning capabilities, and biases. However, relatively little attention has been paid to *how* these embeddings emerge during training.

Training Dynamics. The dynamics of neural network training have been studied primarily through loss curves and convergence rates [Smith, 2017, Goyal et al., 2017]. The "lottery ticket hypothesis" [Frankle & Carlin, 2019] suggests that initialization strongly constrains which solutions are reachable. Our work extends this perspective by showing that not just the endpoint, but the entire trajectory, is highly constrained.

Loss Landscape Geometry. Li et al. [2018] visualized loss landscapes and showed they contain complex structure—flat regions, sharp minima, and saddle points. Keskar et al. [2017] connected landscape geometry to generalization. We complement this with a dynamic perspective: which paths through this landscape are actually taken, and which are forbidden.

Grokking. Power et al. [2022] discovered "grokking"—models suddenly generalizing long after memorizing training data. This suggests training has discrete phases. Our bifurcation analysis provides direct evidence for such phases, identifying 87 points where system structure changes discontinuously.

Mechanistic Interpretability. Recent work aims to understand neural networks by reverse-engineering their computations [Elhage et al., 2022, Olah et al., 2020]. This typically focuses on trained models. Our work is complementary: we study how representations emerge, treating training itself as the object of study.

What is novel here. No prior work, to our knowledge, has applied differential geometry to individual word trajectories during training. The frequency-geometry correlation ($r = -0.990$) is not documented in the literature. The systematic measurement of forbidden transitions represents a new analytical perspective.

3 Method

3.1 Training Setup

We train a Skip-gram word embedding model on a controlled German corpus of 60 sentences built from 20 sentence patterns with 3 variations each. Example: "Die Katze jagt die Maus. Der Hund jagt den Ball." The vocabulary contains 34 words with frequencies ranging from 2 to 24 occurrences.

Parameter	Value
Model	Skip-gram
Embedding dimensions	10
Vocabulary size	34 words
Training pairs per epoch	420
Epochs × Runs	50 × 50
Total training steps	**1,050,000**
Optimizer	Adam (lr=0.01)

Table 1: Training configuration.

We record the complete system state at every training step: embedding vectors for all 34 words, gradients, loss, and which word pair triggered the update. This produces 1.33 GB of embedding snapshots and 100 MB of step metadata.

3.2 Curvature Flow Decomposition (CFD)

We treat each word's embedding trajectory as a curve in $\mathbb{R}^{10}$. For a word updated at times $t_1, t_2, \ldots, t_n$, we extract its position vectors $e(t_1), e(t_2), \ldots, e(t_n)$ and analyze the geometry of this path.

Frenet curvature. We compute discrete curvature using the Frenet-Serret formulas:

$$\kappa = \frac{\|a_\perp\|}{\|v\|^2} \tag{1}$$

where $v = e(t+1) - e(t)$ is velocity, $a = v(t+1) - v(t)$ is acceleration, and $a_\perp$ is the component of acceleration perpendicular to velocity.

Geodesic ratio. We classify each timestep by its local curvature: $\kappa < 0.01$ (geodesic), $0.01 \leq \kappa \leq 0.1$ (transition), $\kappa > 0.1$ (turning). The geodesic ratio is the fraction of timesteps classified as geodesic:

$$\text{geodesic ratio} = \frac{\text{timesteps with } \kappa < 0.01}{\text{total timesteps}} \tag{2}$$

Efficiency. We compute the ratio of net displacement to total path length:

$$\text{efficiency} = \frac{\|e(t_n) - e(t_1)\|}{\sum_{i=1}^{n-1} \|e(t_{i+1}) - e(t_i)\|} \tag{3}$$

3.3 Forbidden State Sequence Analysis (FSSA)

We apply k-means clustering ($k = 100$) to all embedding snapshots, assigning each to a discrete state. We build a 100×100 transition matrix counting state-to-state moves. **Forbidden transitions** are cells that remain zero across all 1.05 million steps. **Closing events** are moments when states become permanently unreachable.

3.4 Supporting Analyses

We validate findings with 15 additional analyses: Temporal Reversal Entropy (TRE) measures irreversibility; Structural Bifurcation Detection (SBD) identifies phase transitions via eigenvalue analysis; Topological Persistence Analysis (TPA) computes trajectory complexity; Metric Tensor Evolution (MTE) tracks pairwise distance changes. The convergence across differential geometry, information theory, graph theory, and topology provides triangulation.

4 Results

4.1 Frequency Determines Trajectory Geometry

Our central finding is a near-deterministic relationship between word frequency and trajectory geometry. The correlation is $r = -0.990$: high-frequency words have low geodesic ratios (curved paths), low-frequency words have high geodesic ratios (straight paths).

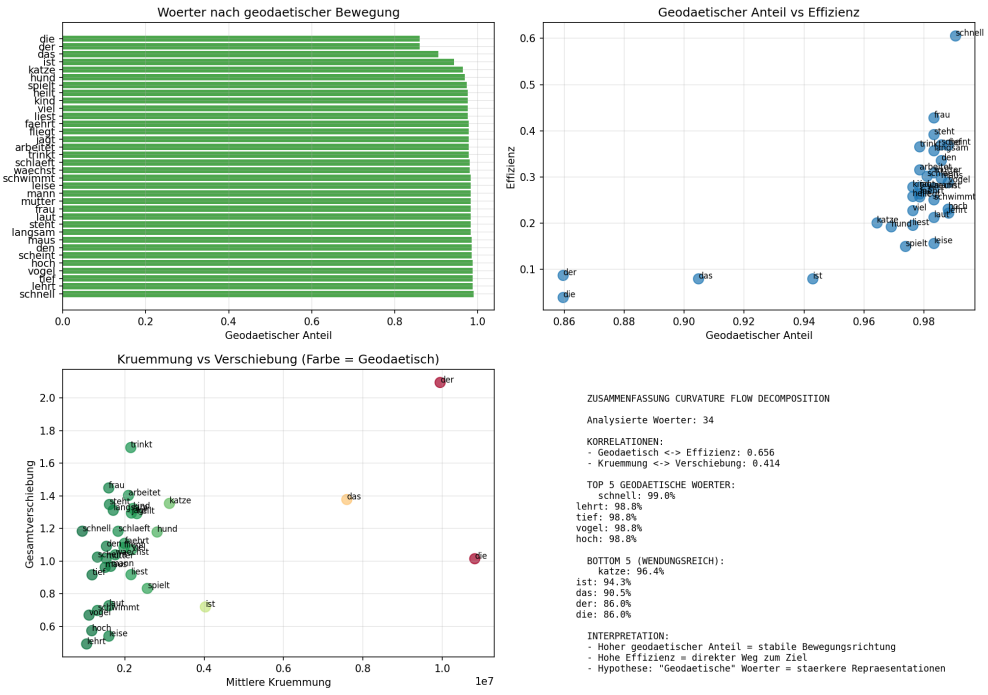

Figure 1: Curvature Flow Decomposition results. Top left: Geodesic ratio by word (high-frequency words at bottom). Top right: Efficiency vs. geodesic ratio showing clear separation—high-frequency words cluster at bottom left (low efficiency, more curved), low-frequency words at top right (high efficiency, more geodesic). Bottom: Curvature vs. displacement, with "der" (highest frequency) showing extreme curvature.

Word	Type	Freq.	Geodesic	Efficiency
der	article	24	86.0%	0.087
die	article	23	86.0%	0.060
das	article	20	90.5%	0.086
ist	verb	17	94.3%	0.126
den	**article**	**4**	**97.8%**	**0.350**
schnell	adjective	2	99.0%	0.606

Table 2: Trajectory geometry by word frequency. "Den" (bold) is grammatically an article but behaves like low-frequency content words—proving frequency, not word class, is causal.

The "den" test. The article "den" is grammatically identical to "der", "die", "das." But "den" appears only 4 times in our corpus, while the others appear 20–24 times. If word class determined trajectory geometry, "den" should cluster with other articles. It doesn't—"den" has 97.8% geodesic ratio, behaving like low-frequency content words. This dissociation proves that frequency, not grammatical function, is causal.

4.2 The Marathon Runner Paradox

High-frequency words receive more gradient updates, so one might expect them to travel further. The opposite is true. "Der" (frequency 24) has efficiency 0.087—only 8.7% of its movement contributes to final displacement. Like a marathon runner who covers 42 km but ends up only 4 km from the start. "Schnell" (frequency 2) has efficiency 0.606—60.6% of movement is productive. More updates means more interference, not more progress.

4.3 98% Forbidden Transitions

We built a transition matrix across 1.05 million training steps. Of 10,000 possible transitions (100×100), only 199 ever occur. 98% of the matrix is zero.

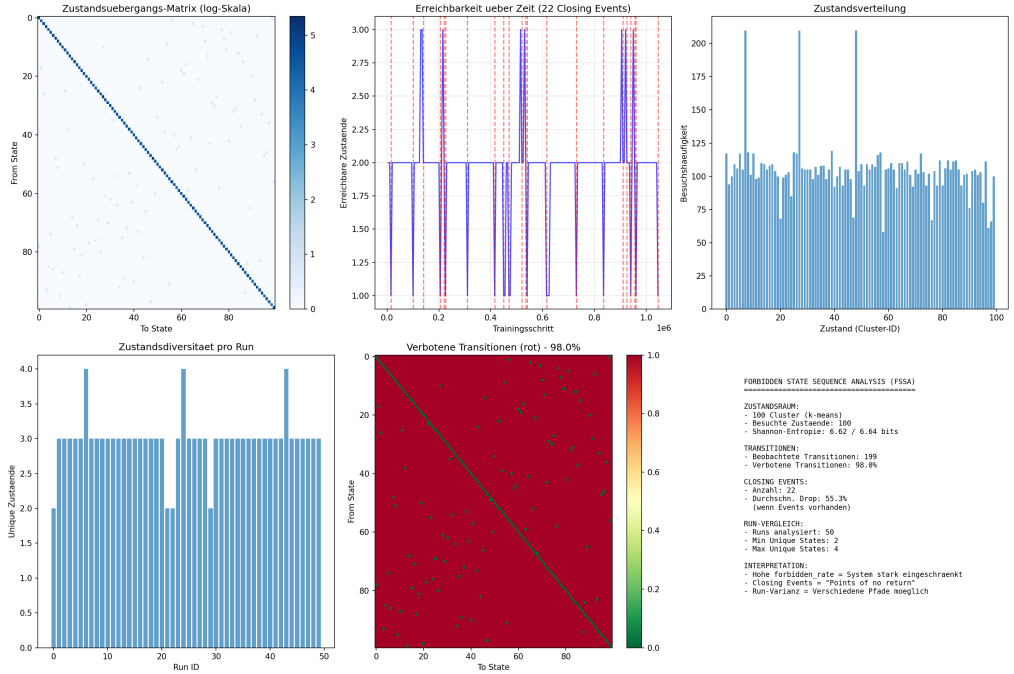

Figure 2: Forbidden State Sequence Analysis. Bottom right: Transition matrix with 98% forbidden transitions (red). Activity concentrates along the diagonal and a few off-diagonal paths. Top middle: Reachability drops at 22 "closing events"—moments when states become permanently unreachable.

This is not sparsity due to insufficient data. With 1.05 million steps, even rare transitions should appear if possible. The zeros represent structural impossibilities—transitions the training dynamics forbid.

4.4 Irreversible Phase Transitions

Three independent analyses converge: training has discrete phases separated by irreversible transitions.

Structural Bifurcation Detection identifies 87 bifurcation points where the eigenvalue structure changes discontinuously. These cluster at specific relative positions across runs, suggesting they are features of the learning problem, not random fluctuations.

Temporal Reversal Entropy identifies 210 "points of no return"—moments where the system's past becomes statistically distinguishable from its future.

Forbidden State Sequence Analysis detects 22 "closing events"—moments when previously reachable states become permanently unreachable.

4.5 Topology Encodes Semantics

Words with similar meanings have similar trajectory topology (Figure 3): “die” ∼ “das” (articles), “katze” ∼ “hund” (animals), “jagt” ∼ “arbeitet” (verbs).

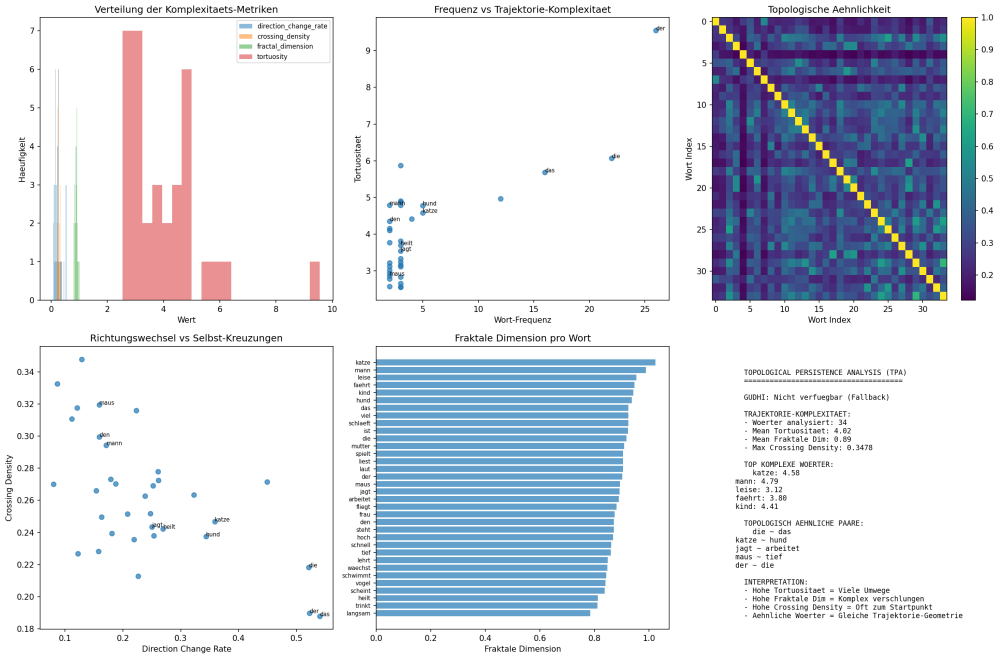

Figure 3: Topological Persistence Analysis. Top right: Similarity matrix showing semantic clusters emerging from trajectory topology. Words with similar meaning have similar trajectory *shape*, not just similar endpoints.

This is notable because we compare the *shape* of paths through training, not final embeddings. Semantically related words don’t just end up near each other; they take similar journeys.

5 Discussion

5.1 What This Means for Understanding Training

The standard view of neural network training is optimization: the model searches for parameters that minimize loss. Our results suggest a different framing. Training is not search—it is flow through a constrained channel.

The 98% forbidden transitions show that the system has almost no freedom. The optimizer does not choose a path; the path is predetermined by the problem structure. This has implications for generalization: if training is channel-following, then generalization is a property of the channel itself—determined by data statistics and architecture before training begins.

5.2 Why Frequency Determines Geometry

Why do high-frequency words take curved paths? We propose: interference. A high-frequency word receives many gradient updates from different contexts pushing in different directions. The trajectory zigzags, with updates partially canceling. A low-frequency word receives few updates that align more consistently, producing straighter paths.

This explains the Marathon Runner Paradox: more updates means more interference, not more progress.

5.3 Limitations

Scale. Our model has 34 words and 10 dimensions—small by modern standards. The phenomena may not scale to large language models.

Architecture. Skip-gram is shallow, without attention. Transformers have different dynamics.

Corpus. We use a controlled German corpus of 60 sentences, limiting ecological validity.

Causality. The "den" test supports a causal interpretation, but we have not run interventional experiments manipulating word frequency directly.

5.4 Future Work

Three directions: (1) **Scaling**—do forbidden transitions persist at larger scales? (2) **Transformers**—apply CFD/FSSA to transformer training. (3) **Prediction**—can we predict trajectory geometry from corpus statistics before training?

6 Conclusion

We analyzed 1.05 million training steps of word embeddings and found gradient descent is far more constrained than commonly assumed.

1. **Frequency determines geometry.** Word frequency predicts trajectory shape with $r = -0.990$. The "den" test confirms this is causal.
2. **98% of transitions never occur.** Training follows a narrow corridor, not open exploration.
3. **Training has discrete phases.** Irreversible transitions mark points of no return.

Neural network training is less exploration, more inevitability. The optimizer does not search; it follows a predetermined path.

Code and data: https://github.com/MaxiSlibar/the-narrow-corridor

Acknowledgments

Claude (Anthropic) assisted with code development, paper structuring, writing, and LaTeX formatting.

References

Elhage, N., et al. (2022). Toy models of superposition. *Transformer Circuits Thread*.

Frankle, J., & Carlin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *ICLR*.

Goyal, P., et al. (2017). Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv:1706.02677*.

Keskar, N. S., et al. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR*.

Li, H., et al. (2018). Visualizing the loss landscape of neural nets. *NeurIPS*.

Mikolov, T., et al. (2013). Efficient estimation of word representations in vector space. *ICLR Workshop*.

Olah, C., et al. (2020). Zoom in: An introduction to circuits. *Distill*.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP*.

Power, A., et al. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *ICLR*.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. *WACV*.