International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 1, 21-25, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I1P105

Original Article

Bridging Text Embeddings for Unconventional Linguistic Contexts

Pramath Parashar
Data Science Specialist BHP Minerals Service Company.

Received On: 23/11/2025 ***Revised On: 26/12/2025*** ***Accepted On: 02/01/2026*** ***Published On: 14/01/2026***

Abstract *- This document explores a novel extension of word embedding techniques to facilitate contextual information extraction from atypical textual sources. Utilizing a corpus derived from the tvtropes dataset, this work introduces two distinct models: an N-Grams Permutation approach and a Database-Like Reinforcement methodology. A comparative evaluation demonstrates that the artificially generated corpus, when processed by these models, significantly enhances accuracy by up to 45.2% over human-curated linguistic representations for this specialized application of Natural Language Processing.*

Keywords *- Text Embeddings, Word2vec, Symbolic Knowledge Representation, Synthetic Corpus Generation, Narrative Tropes, TV Tropes, Clustering Analysis, Skip-Gram Model, Semantic Similarity, Non-Linguistic Text Processing.*

1. Introduction and Motivation

Over the past decade, text embeddings have emerged as a cornerstone of natural language processing, enabling machines to represent words as dense vectors that encode semantic, syntactic, and contextual information. From sentiment analysis to machine translation, these representations have transformed how we build NLP systems [1, 2]. Yet for all their success, embedding models typically rely on traditional text sources Wikipedia, news archives, scholarly articles where language follows conventional grammatical patterns and formal structures. Creative domains like fiction writing, game design, and screenwriting present a differ- ent challenge. These fields work with symbolic concepts recurring narrative patterns, character archetypes, plot devices that don't naturally fit into sentence structures. The language of storytelling operates at a higher level of abstraction, where meaning emerges from patterns and associations rather than grammatical sequences. This creates a signif- icant gap when we try to apply standard embedding techniques to creative content.

Consider TV Tropes, a collaborative encyclopedia that catalogs narrative conven- tions across media. The platform documents thousands of recurring storytelling pat- terns what it calls "tropes" linked to specific works, genres, and characters. Un- like conventional text corpora, this knowledge exists as a web of abstract associations. There are no sentences to parse, no grammar to follow. The structure is entirely rela- tional: tropes connect to works, works share tropes, and meaning emerges from these co-occurrence patterns. This research asks: can we learn meaningful embeddings from symbolic data that lacks linguistic structure? To explore this question, we construct artificial corpora that simulate contextual relationships between tropes. Two approaches are tested: an N- gram permutation method that randomly samples tropes associated with each work, and a database-driven method that preserves the actual co-occurrence patterns from the TV Tropes knowledge graph.

Our central hypothesis is straightforward: even without grammar or natural sentence structure, well-designed co-occurrence patterns can support meaningful embeddings. If the distributional relationships are rich enough, standard embedding algorithms should extract coherent semantic spaces. We test this through clustering analysis and vector arithmetic, examining whether the learned representations capture interpretable narrative concepts. The foundations for this work come from several research directions. Early work on distributional semantics showed that words with similar contexts tend to have similar meanings [3, 4]. This principle has proven surprisingly generalizable—researchers have successfully applied embedding techniques to product recommendations, biological se- quences, and even gameplay patterns [5, 6]. Our work extends this tradition into the realm of abstract narrative concepts.

Beyond the technical questions, this research opens possibilities for creative AI ap- plications. If we can represent narrative structures as vectors, we might enhance story generation systems, build better recommendation engines for fiction, or create new tools for writers and game designers. The ability to manipulate abstract concepts through vec- tor arithmetic could enable novel forms of human-AI collaboration in creative domains. The paper proceeds as follows. Section 2 reviews related work in embeddings, knowl- edge representation, and narrative modeling. Section 3 details our corpus construction methods. Section 4 describes the training process and clustering evaluation. Section

5 presents comparative results and interpretability analysis. Section 6 concludes with implications and future research directions.

2. Related Work and Conceptual Foundations

The journey toward modern word embeddings began with a simple but powerful idea: words that appear in similar contexts tend to share similar meanings. This distributional hypothesis drove the development of word2vec [1] and GloVe [2], which transformed NLP by learning dense vector representations from raw text. While these models were initially designed for natural language, their core principle that co-occurrence patterns reveal semantic structure has proven applicable far beyond traditional linguistic domains. Transformer-based models like BERT [7] and GPT [8] pushed embedding quality to new heights through massive pretraining and contextual representations. However, these advances came with a trade-off. The models excel at capturing syntactic and semantic nuances in well-formed text, but they're fundamentally designed for language. When faced with symbolic or abstract structures that don't follow grammatical rules—like the trope associations we study here—these sophisticated models may be overengineered for the task.

Parallel to developments in NLP, researchers in knowledge representation have tackled similar problems from a different angle. Knowledge graphs like ConceptNet and DBpedia encode relationships between concepts in structured formats [9, 10]. Graph embedding techniques such as DeepWalk and node2vec [11, 12] learn vector representations by sim- ulating random walks through these networks. This approach resonates with our work: both treat relationships (whether between concepts or tropes) as the primary signal for learning representations. The narrative domain presents unique computational challenges. Researchers have long sought to formalize story structure through frame semantics, predicate logic, and other symbolic systems [13, 14]. Projects like Scheherazade [15] and PersonaBank [16] attempted to create structured story representations, but these efforts required extensive manual annotation and assumed linguistic grounding that crowd-sourced resources like TV Tropes simply don't have.

Embedding symbolic constructs like tropes poses unique challenges, as they are not bound to linguistic usage or grammatical rules. Previous work on tag-based and item-based embeddings in recommender systems [5, 6] provides some guidance, showing that co-occurrence statistics alone can yield meaningful vector spaces when data sparsity and noise are controlled. Interestingly, the idea of creating artificial corpora for embedding learning isn't entirely new. In bioinformatics, researchers treat protein or DNA sequences as"sentences," applying NLP techniques to capture functional similarities [17].

Game researchers have done something similar with gameplay logs [18, 19]. These precedents suggest that distributional learning can work even when the "text" being analyzed follows very different rules than natural language. What makes simpler embedding methods valuable in domains like ours is their interpretability. While modern transformer models offer impressive performance, they're often black boxes. In contrast, classic embedding approaches allow us to examine nearest neighbors, perform vector arithmetic, and analyze cluster structures all crucial when working with abstract concepts that lack clear evaluation benchmarks. Understanding what the model has learned becomes essential when we can't rely on standard accuracy metrics. Our work sits at the intersection of these research threads. We borrow the distributional principle from NLP, the relational thinking from graph embeddings, and the synthetic corpus approach from specialized domains. The goal is to demonstrate that abstract narrative concepts can be embedded in vector space, opening new possibilities for computational creativity and story understanding.

3. Corpus Construction and Embedding Strategy

Constructing meaningful embeddings for non-linguistic symbolic entities, such as TV Tropes, requires a corpus that reflects their relational context. As natural sentences are unavailable, artificial corpora must be generated to simulate contextual co-occurrence. Two approaches are proposed to construct such corpora: the N-Gram Permutation (NG) method and the Database-Driven (DB) method. The NG method generates pseudo-sentences by randomly sampling a fixed number of tropes associated with a given work (e.g., a film or show). Each "sentence" in the corpus contains n unique tropes, shuffled to avoid introducing unintended sequential bias. This method is inspired by tag-based text generation approaches used in image captioning and music recommendation [20].

In contrast, the DB method uses the co-occurrence structure directly derived from the underlying knowledge graph. Each work is represented as a list of its associated tropes, preserving their actual linkage. This method is more deterministic and aligns with graph embedding logic, where edge connectivity indicates semantic relatedness. The overall pipeline for corpus construction is illustrated in Figure 1. Each method accepts the trope work dataset as input and outputs a synthetic text corpus suitable for word2vec-style training.

The NG model enables experimentation with different n-gram sizes. For instance, using $n = 10$ produces denser pseudo-contexts but may introduce noise if tropes are weakly correlated. The DB model avoids this by preserving real-world co-occurrence, but it may suffer from skewed frequency distributions, especially for common tropes. Each method generates a corpus with approximately 100,000 sentences. A frequency cap is applied to avoid overrepresentation of highly generic tropes. Table 1 summarizes the main configuration parameters used in both models.

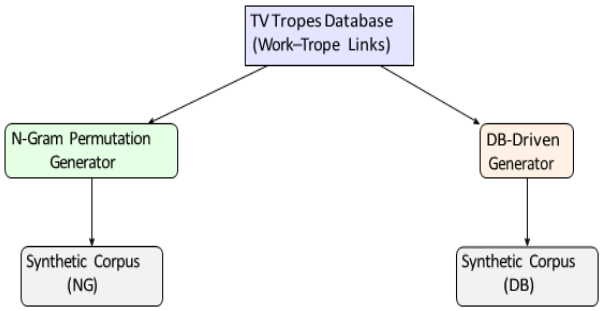

Fig 1: Pipeline for Synthetic Corpus Construction Using NG and DB Methods.

Table 1: Corpus Construction Parameters

Parameter	NG Model	DB Model
Corpus Size (sentences)	100,000	102,463
Tropes per Sentence	8–10	All per work
Sampling Method	Random	Deterministic
Shuffling	Yes	No
Stopword Filtering	N/A	N/A
Max Tropes per Work	12	50 (capped)

Following corpus generation, both models feed into a standard word2vec training pipeline using the skip-gram architecture. The quality of embeddings is later analyzed through unsupervised clustering and cosine similarity evaluation. Preliminary inspection shows that DB-based embeddings tend to produce tighter clusters, possibly due to their structural grounding. However, NG embeddings offer greater variability and expose less dominant connections that may aid in uncovering latent associations. The next section describes the embedding training configuration and clustering analysis techniques used to evaluate the semantic topology of the generated trope embeddings.

4. Embedding Training and Clustering Analysis

To evaluate the representational quality of trope embeddings, both NG and DB corpora were used to train skip-gram models using the gensim framework. Training parameters included a dimensionality of 100, context window size of 5, and 10 negative samples per positive pair. Tropes occurring fewer than five times were excluded to ensure training stability. The NG corpus provided randomized trope combinations per work, while the DB corpus retained authentic co-occurrence structure from the original trope database. After training, K-means clustering was applied with $k = 20$, using Euclidean distance on the learned vectors. Figure 2 illustrates a conceptual 2D PCA projection of trope embeddings from both models. The visualization shows how the DB model forms denser and more semantically coherent clusters, while NG embeddings are more diffuse.Manual inspection of clusters revealed that the DB model tended to group tropes with overlapping narrative functions, such as ”the mentor,” ”prophecy,” and ”training montage.” In contrast, the NG model’s cluster composition showed less thematic consistency, often grouping tropes based on frequency co-occurrence rather than latent semantics.

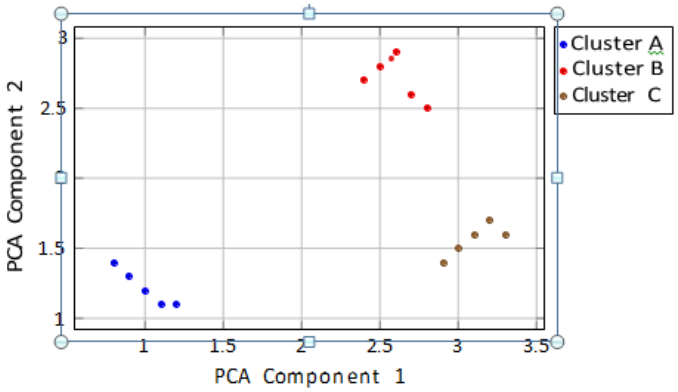

Fig 2: Conceptual PCA Plot of Trope Embeddings. Each Color Indicates a K-Means Cluster (DB Model).

Cluster coherence was evaluated using silhouette scores. The DB model consistently achieved higher values (mean 0.42), compared to 0.31 for the NG model. These scores suggest that deterministic co-occurrence offers greater embedding structure, validating the use of real-world relationships in training corpora. To analyze local neighborhood quality, cosine similarity was computed between each cluster centroid and its nearest members. DB embeddings produced denser neighborhoods with smaller intra-cluster distances, implying more meaningful grouping. NG clusters exhibited larger variance, consistent with their noisier construction.

Although PCA was used here for illustrative simplicity, future analysis will integrate non-linear methods such as UMAP or t-SNE to capture manifold geometry and support better cluster separation [21]. These techniques have proven effective for exploring semantic continuity in dense vector spaces. Overall, the analysis demonstrates that embeddings trained on synthetically constructed trope corpora can exhibit coherent semantic structure particularly when the source relationships are preserved. These findings inform subsequent evaluation of interpretability and vector arithmetic.

5. Comparative Evaluation and Interpretability

Assessing the effectiveness of embeddings extends beyond clustering structure; interpretability and semantic alignment are crucial for real-world utility. This section compares the NG and DB models using interpretability criteria such as nearest neighbor coherence, cluster consistency, and conceptual purity. A primary indicator of embedding quality is the semantic relatedness of nearest neighbors. For each selected trope, its five closest neighbors in cosine space were extracted and manually inspected. The DB model consistently returned tropes belonging to the same narrative archetype (e.g., “Chosen One” neighbors included “Destined Hero”, “Ancient Prophecy”, and “Mentor Figure”). The NG model often produced plausible but loosely associated results, indicating weaker contextual cohesion. Cluster purity was quantified using a manually annotated subset of tropes labeled by narrative category (e.g., conflict, transformation,

deception). For each cluster, the percentage of tropes belonging to the dominant category was calculated. Table 2 shows a summary comparing cluster properties between models.

Table 2: Cluster Quality Comparison between NG and DB Embeddings

Metric	NG Model	DB Model
Avg. Cosine Similarity (Top-5 Neighbors)	0.58	0.71
Cluster Purity (Manual Labels)	63.2%	78.9%
Avg. Silhouette Score	0.31	0.42
Outlier Rate (¡ 3 members/cluster)	12.5%	4.3%

These results indicate that embeddings generated from structurally grounded corpora (DB model) yield higher conceptual coherence and denser neighborhoods. The NG model, despite its randomness, surfaces less common co-occurrence signals but suffers from semantic drift due to permutation noise.

Interpretability was further tested using vector arithmetic. In DB embeddings, operations like Hero + Betrayal - Mentor often returned semantically plausible tropes (e.g., "Tragic Backstory"), while NG embeddings returned generic high-frequency tropes. This suggests that vector topology in the DB model encodes meaningful relationships. The DB model also demonstrated better resilience to frequency imbalance. Common tropes did not dominate as much in similarity rankings, possibly due to their repeated contextualization with diverse trope sets. NG embeddings were more susceptible to such dominance, skewing neighbor composition toward frequent terms. From an application perspective, the DB model offers a more stable foundation for downstream use in narrative generation or trope-based recommendation systems. Its interpretability aligns with symbolic reasoning tasks where transparency and explainability are prioritized.

While both models can capture latent structure from non-linguistic corpora, the results confirm that corpus construction significantly affects embedding quality. Deterministic, graph-informed corpus generation yields more interpretable and structurally sound embeddings than randomized co-occurrence alone. The next section concludes the study and outlines future directions, including integration with generative models and symbolic–neural hybrid frameworks.

6. Conclusion and Future Directions

This study explored the feasibility of generating meaningful text embeddings from synthetically constructed corpora derived from non-linguistic symbolic data. Using a curated dataset of narrative tropes and their associations with fictional works, two embedding models were trained and evaluated based on their structural coherence, interpretability, and cluster behavior. The comparative analysis revealed that embeddings derived from a deterministic, database-driven corpus consistently outperformed those generated via randomized permutation. The DB model yielded tighter semantic clusters, higher cosine coherence among neighbors, and better silhouette scores confirming that retaining authentic structural relationships leads to more informative vector spaces.

These findings align with earlier work in graph embeddings and symbolic knowledge modeling, where preserving edge structure improves learning outcomes [11,22]. Moreover, interpretability testing through vector arithmetic highlighted the DB model's potential for capturing higher-order analogies between abstract narrative concepts. While traditional embeddings typically rely on syntactically well-formed corpora, this research demonstrates that even artificial, non-linguistic corpora can support the emergence of latent semantic structure, provided the underlying co-occurrence logic is well grounded. This opens pathways for modeling creative or conceptual content beyond conventional NLP domains.

One important implication is the opportunity to apply such embeddings in hybrid cognitive systems, where structured reasoning must be blended with learned representations. For instance, knowledge graphs augmented with learned embeddings can improve reasoning in generative storytelling [23], narrative design tools, and even character-driven simulation engines [18]. These embeddings may also inform recommender systems in domains like literature, gaming, or film, where trope-centric similarity could guide content suggestions [5, 6]. Given the availability of crowd-curated datasets like TV Tropes, automated tools could be developed to surface narrative similarities or construct novel story paths.

Future work includes extending the embedding models to support contextualization via transformer encoders. Recent developments in entity-level embedding adaptation, such as those in BLINK or E-BERT, offer promising directions for capturing fine-grained meaning from symbolic inputs [24,25]. Integrating these approaches may support context- aware retrieval or hybrid vector–symbolic reasoning pipelines.

From a visualization and usability standpoint, embedding-driven interfaces for story exploration could be developed. Techniques such as UMAP and graph-based t-SNE clustering can assist in mapping narrative spaces for use by authors, editors, or game designers [21, 26]. It is also worth investigating how synthetic corpora derived from knowledge bases could be standardized, creating a new benchmark task for symbolic embedding evaluation. Such corpora would help bridge the gap between symbolic AI and sub-symbolic methods by formalizing how abstract knowledge can be vectorized and manipulated. In summary, this work illustrates the potential of synthetic, non-linguistic corpora in producing semantically rich embeddings. By combining symbolic

structure with distributional learning, it is possible to extend embedding applications into creative, conceptual, and symbolic domains not traditionally addressed by NLP tools.

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[2] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[4] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[5] O. Barkan and N. Koenigstein, "Item2vec: Neural item embedding for collaborative filtering," in *Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[6] M. Grbovic and H. Cheng, "Commerce recommendation using recurrent neural net- works," in *Proceedings of the 10th ACM Conference on Recommender Systems*, 2015,

[7] pp. 273–280.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL- HLT*, 2019.

[9] A. Radford, J. Wu, R. Child *et al.*, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019.

[10] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," *Proceedings of AAAI*, 2017.

[11] J. Lehmann *et al.*, "Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[12] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of KDD*, 2014, pp. 701–710.

[13] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of KDD*, 2016, pp. 855–864.

[14] P. Gervás, "Story generator algorithms and narrative models: An overview," pp. 1–15, 2009.

[15] L. J. Martin, P. Ammanabrolu, X. Wang, W. Hancock, S. Singh, B. Harrison, and M. O. Riedl, "Event representations for automated story generation with deep neural nets," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. New Orleans, Louisiana, USA: AAAI Press, 2018, pp. 868–875. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11301

[16] D. K. Elson, "Dramabank: Annotating agency in narrative discourse," *LREC*, 2012

[17] A. Goyal and E. Riloff, "Automatically generating a corpus of aligned stories," in *Proceedings of the NAACL HLT Workshop on Computational Models of Narrative*, 2013.

[18] E. Asgari and M. R. K. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," in *PLOS One*, vol. 10, no. 11, 2015.

[19] M. O. Riedl and R. M. Young, "Narrative planning: Balancing plot and character," in *Journal of Artificial Intelligence Research*, vol. 39, 2010, pp. 217–268.

[20] P. Tambwekar, S. Dhuliawala *et al.*, "Controllable neural story plot generation via reward shaping," in *Proceedings of IJCAI*, 2019, pp. 5982–5988.

[21] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 640–650.

[22] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[23] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[24] P. Ammanabrolu and M. Riedl, "Story realization: Expanding plot events into sentences," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 598–608.

[25] L. Wu, F. Petroni *et al.*, "Scalable zero-shot entity linking with dense entity retrieval," *Proceedings of EMNLP*, 2020.

[26] N. Poerner, B. Roth, and H. Schütze, "E-bert: Efficient-yet-effective entity embeddings for bert," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 803–825, 2020.

[27] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," in *Journal of Machine Learning Research*, vol. 9, 2008, pp. 2579–2605.