harapalb at SemEval-2026 Task 4: Multi-Signal Neuro-Symbolic Ensembles for Narrative Similarity

Andrei Tiberiu Carp
Tomorrow University
ING Hubs Romania
tiberiucarp@gmail.com

Abstract

Determining narrative similarity requires moving beyond surface-level semantic overlap toward an understanding of structural, causal, and dynamic alignment. In this paper, we present a hybrid system for the SemEval 2026 Task 4 (Track A) on narrative similarity. Our architecture is a neuro-symbolic ensemble that fuses three complementary signals: (1) action-focused neural embeddings that capture event trajectories while filtering descriptive noise, (2) a symbolic Structural Survival Ratio (SSR) grounded in dependency parsing that measures discrete event preservation, and (3) high-level LLM-based structural comparison across causal, thematic, and sequential dimensions. We also augment this three-signal baseline with a Sequential Trajectory Analysis module applied as a trajectory-based fallback on cases where Signal 3 failed, using the discrete Fréchet distance to compare trajectory shapes and abrupt semantic shift detection to identify plot transition alignment. Unlike static centroid-based approaches, this trajectory analysis preserves narrative *order*, enabling disambiguation in failure cases where the primary vote reduces to a two-signal split. Evaluated on the official test set, our ensemble achieves 68.25% accuracy.

1 Introduction

Recent advancements in Natural Language Processing (NLP) system design reflect a critical transition from purely neural, end-to-end Large Language Models (LLMs) toward hybrid, structured, and modular architectures. While LLMs have achieved the best performances on benchmark tasks, extensive evaluations reveal their tendency to rely on surface-level statistical correlations rather than deep semantic understanding (Han et al., 2024). This results in "black box" systems that struggle with long-context comprehension, multi-hop reasoning, and complex causal dependencies, often leading to hallucinations and a lack of interpretability (Li et al., 2024). Consequently, narrative similarity requires structural alignment beyond embeddings, symbolic reasoning, and multi-channel feature fusion to ensure reliability.

For the task of narrative similarity, where the goal is to identify which of two candidate stories is more structurally similar to an anchor, we propose a modular design. Following this broader methodological shift, our system treats narrative similarity not as a static metric, but as a composite of three distinct semantic/symbolic signals, further augmented by a trajectory-based fallback module grounded in continuous dynamical systems theory (Robinson, 2012) to capture the non-linear "flow" of human narratives.

2 Literature Review

2.1 Re-defining Narrative Similarity

The evaluation of narrative similarity has traditionally been hindered by conflating it with surface-level topical or lexical overlap, relying heavily on topic modeling or exact text matching (Waight et al., 2025). However, true narrative similarity transcends shared vocabulary; it requires an understanding of the underlying sequence of events, their causal links, and thematic resolutions (Gupta et al., 2025). Cognitive and computational studies hypothesize that narrative alignment is fundamentally an abstraction task, where the similarity between two stories equates to the existence of an appropriate "common summary" that captures a core structural schema (Kyprideemou and Michael, 2013). Modern computational narratology thus asserts the necessity of stripping away superficial details to compare these structural and causal foundations.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2238–2244

2.2 Neural Embeddings and Event Representation

To accurately capture the dynamic progression of stories, researchers emphasize event-centric models, often utilizing "scripts" or narrative event chains, which are partially ordered sets of events centered around a common protagonist (Chambers and Jurafsky, 2008). Modern approaches rely on rich lexical resources, such as VerbNet (Kazeminejad et al., 2022), which detail semantic representations and track participant states through subevents. Furthermore, frameworks like GLAMR (Tu et al., 2024) extend Abstract Meaning Representation (AMR) by incorporating structured subeventual interpretations of predicates and tracking the property changes of arguments. When coupled with advanced sentence embeddings, such as SBERT, these action-focused representations provide a highly effective substrate for quantifying the semantic proximity of narrative events beyond exact text matching (Waight et al., 2025).

2.3 Neuro-Symbolic Integration and Dependency-Based Structures

Despite the results obtained by LLMs, these architectures continue to struggle with deep semantic understanding, coreference resolution, and long-range structural reasoning (Papakostas et al., 2025). To overcome these limitations, the state-of-the-art has shifted toward Neuro-Symbolic (NeSy) architectures that combine neural pattern-matching capabilities with the deterministic, rule-based reasoning of symbolic systems (Akter et al., 2025). For instance, separating coreference resolution and dependency extraction into an explicit symbolic pre-processing module before mapping text to an AMR parser enforces strict linguistic constraints and yields structurally sound, semantically richer graphs (Papakostas et al., 2025). Incorporating symbolic structural analysis based on dependency parsing, such as the Structural Survival Ratio (SSR), explicitly anchors neural models, preventing error propagation and ensuring high-fidelity structural alignment between texts (Kádár et al., 2021).

2.4 High-Level LLM Reasoning and Structural Comparison

While symbolic methods provide essential deterministic boundaries, LLMs remain effective at distilling texts down to their core ideas, subjects, and high-level abstractions (Aly et al., 2025). However, relying solely on LLMs for narrative reasoning introduces critical vulnerabilities. Benchmarks like NovelHopQA reveal that even frontier models experience accuracy drops when reasoning across complex multi-hop narrative contexts (Gupta et al., 2025). To effectively harness LLM reasoning, current methodologies provide models with explicit structural guidance (Ma et al., 2025). For example, frameworks like LeStrTP utilize graph neural networks and Markov Chain modeling to identify precise plot boundaries, guiding the LLM to encapsulate complete causal arcs when generating structural comparisons (Ma et al., 2025).

2.5 Multi-Signal Ensembles for Robust Narrative Analysis

The synthesis of these diverse methodologies underscores a broader consensus: the next generation of robust NLP systems must not rely solely on scaling standalone neural models (Barnes and Hutson, 2024). Fusing multi-channel data, such as semantic embeddings, symbolic representations, and temporal structures, captures both linguistic nuances and dynamic structural behaviors (Yu and Xu, 2021; Jiao et al., 2024). Ensembling action-focused neural embeddings, deterministic symbolic logic, and high-level LLM structural evaluation consistently outperforms neural-only baselines, providing a transparent and accurate framework for narrative similarity tasks (Bollikonda, 2025; Hou, 2025). These findings motivate our ensemble design: Signal 1 operationalizes event-centric embeddings, Signal 2 provides the deterministic SSR anchor, Signal 3 harnesses LLM structural reasoning with explicit guidance, and Signal 4 addresses the temporal ordering gap as a trajectory-based fallback when the primary vote is incomplete.

3 System Architecture

We address SemEval 2026 Task 4, Track A (Hatzel et al., 2026). Given a triplet consisting of an anchor story (S_{anchor}), and two candidate stories (S_A and S_B), the objective is to predict which candidate (S_A or S_B) is more structurally and narratively similar to the anchor story.

Our system is a multi-channel ensemble that combines three independent semantic/symbolic

signals via a majority vote. The overall architecture is illustrated in Figure 1. Signals 1-3 form the primary majority vote (Semantic/Symbolic Baseline). Signal 4 is evaluated as a trajectory-based fallback module exclusively on the subset of examples where Signal 3 was unavailable, reducing the primary vote to a two-signal split. Accuracy badges reflect individual test-set performance; $+1.25\%$ denotes Signal 4's marginal contribution on this failure subset.

3.1 Signal 1: Action-Only Embeddings

To capture the raw trajectory of events while filtering out descriptive noise, we isolate the action content. We utilize `spaCy` (`en_core_web_sm`) to extract specific parts of speech: verbs, nouns, proper nouns, pronouns, and auxiliaries from each story. Nouns and proper nouns are retained because they anchor the agent and object roles of each action; removing them empirically reduced development accuracy, suggesting they encode structural participant information rather than purely topical content.

The extracted text is embedded using `all-mpnet-base-v2` (sentence-transformers), yielding normalized embeddings. We compute the L_2 distance between the anchor embedding $\mathbf{e}_{\text{anc}}$ and each candidate embedding $\mathbf{e}_c$:

$$d(S_{\text{anchor}}, S_c) = \|\mathbf{e}_{\text{anc}} - \mathbf{e}_c\|_2 \quad (1)$$

Since `all-mpnet-base-v2` produces unit-normalized embeddings, cosine similarity and L_2 distance are monotonically equivalent; the choice does not affect ranking. The system predicts the candidate with the smaller distance. This lightweight semantic signal serves as a baseline, achieving 63.5% development accuracy.

3.2 Signal 2: Structural Survival Ratio (SSR)

Acting as our symbolic reasoning module, SSR measures how much of the anchor's discrete event structure "survives" in the candidates. We extract concrete events using dependency parsing. Each event $e \in \mathcal{E}$ is represented as a 6-tuple:

$$e = (\text{lemma, class, agent, patient, outcome, neg})$$

We perform graded abstraction matching across four weighted levels to find the best match for each anchor event in the candidate story:

- **Exact** ($w = 1.0$): Same lemma, roles, and outcome.
- **Class** ($w = 0.7$): Same action class, roles, and outcome.
- **Class+Outcome** ($w = 0.5$): Same action class and outcome; roles relaxed.
- **Role-only** ($w = 0.3$): Same agent/patient types; predicate ignored.

The weights were set heuristically to reflect decreasing structural fidelity at each abstraction level and were not tuned on the development set; a systematic grid search is left for future work. The final SSR score is:

$$\text{SSR}(S_{\text{anc}}, S_c) = \frac{\sum_{i=1}^{|\mathcal{E}_{\text{anc}}|} \max(w_{i,c})}{|\mathcal{E}_{\text{anc}}|} \quad (2)$$

The candidate with the higher SSR is predicted as more similar. This strict symbolic mapping achieved 56.0% accuracy.

3.3 Signal 3: LLM Structure Comparison

To capture high-level narrative abstractions (themes, complex causal links), we employ `gpt-5-mini` in a two-stage process. First, an extraction phase prompts the LLM to return a Pydantic `NarrativeStructure` object containing the chronological event chain, agent-action-goal triples, causal links, themes, setting, and genre. Second, a comparison phase presents the three extracted structures to the LLM. The model analyzes structural similarity across five dimensions (sequence, roles, causality, theme, genre) and returns a discrete prediction (A or B) alongside a structured explanation. The abstraction-first design is intentional: feeding full narrative texts directly to the LLM risks length-induced drift and surface-lexical anchoring, which is precisely the failure mode we aim to avoid. Structured extraction constrains the comparison to causal and sequential dimensions. This signal achieved 65.5% accuracy.

3.4 Signal 4: Sequential Trajectory Analysis

While static vector centroids capture aggregate semantic content, they discard narrative *order*. We evaluate a trajectory-based fallback module on cases where Signal 3 failed to return a valid prediction (due to API error or unavailability), reducing the primary vote to a two-signal split between Signals 1 and 2. Although not integrated

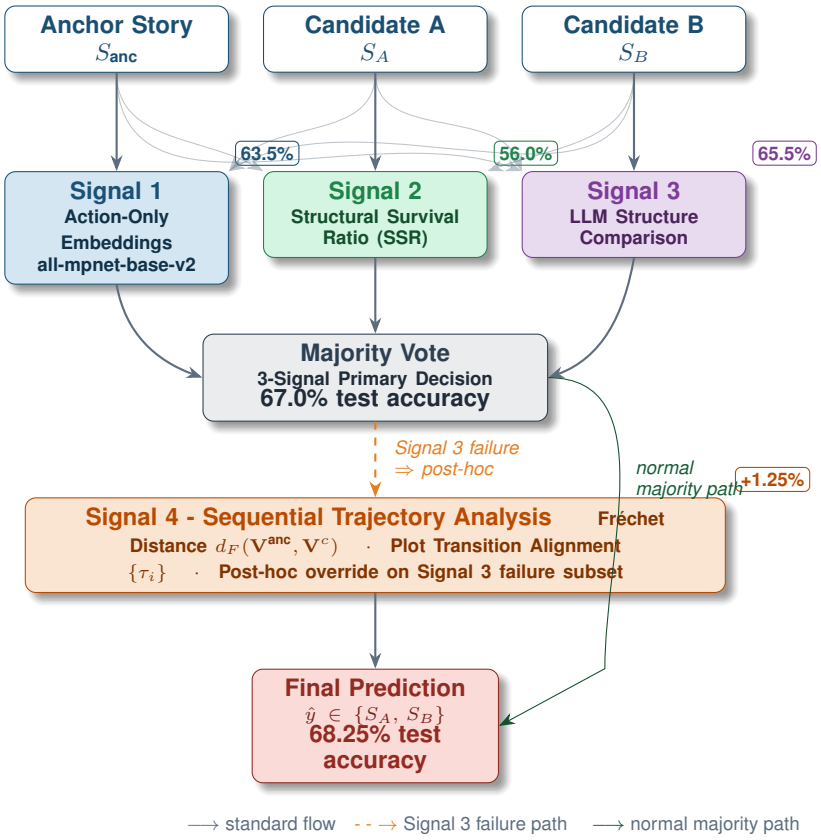

Figure 1: The neuro-symbolic ensemble for narrative similarity.

into the primary voting pipeline, this module tests whether order-aware trajectory comparison can recover structurally valid decisions when the main abstraction signal fails. Each story is encoded as an ordered sequence of verb-event chunks (grouped around syntactic verb phrases) using `all-mpnet-base-v2`. This produces an ordered sequence of vectors $\mathbf{V} = \{v_1, v_2, \ldots, v_n\}$ representing the narrative's progression through semantic space, under the assumption that text order reflects narrative chronology.

Fréchet Distance Comparison. We compare trajectory shapes using the discrete Fréchet distance (Alt and Godau, 1995), a well-established measure of curve similarity that respects temporal ordering. Formally, given two trajectories $\mathbf{P} = \{p_1, \ldots, p_m\}$ and $\mathbf{Q} = \{q_1, \ldots, q_n\}$, the discrete Fréchet distance is defined as:

$$d_F(\mathbf{P}, \mathbf{Q}) = \min_{\sigma} \max_{i} \|p_{\sigma_P(i)} - q_{\sigma_Q(i)}\|_2 \quad (3)$$

where σ ranges over all valid monotone traversals of the two sequences. Intuitively, this captures whether two stories "travel through" semantic space in a structurally congruent way, independent of surface vocabulary overlap. The candidate $S_c \in \{S_A, S_B\}$ with the smaller Fréchet distance to the anchor S_{anc} is predicted as more similar:

$$\hat{y}_4 = \arg\min_{c \in \{A, B\}} d_F(\mathbf{V}^{\text{anc}}, \mathbf{V}^c) \quad (4)$$

Plot Transition Detection. We additionally identify abrupt semantic shifts between consecutive chunks as candidate *plot transition points*. Specifically, a transition is flagged at position i when the distance to the narrative attractor increases by more than 15% relative to the preceding step:

$$\tau_i = \mathbf{1}[\, d_{i+1} > 1.15 \cdot d_i \,], \qquad i = 1, \ldots, n-1 \quad (5)$$

where d_i denotes the distance to the narrative attractor at position i, defined as the centroid of the story's verb-event chunk embeddings. This relative criterion avoids tuning an absolute distance threshold and is robust to differences in absolute embedding scale across stories of varying length. The resulting transition sequence

$T = \{\tau_1, \ldots, \tau_{n-1}\}$ characterizes the structural "rhythm" of the narrative. We use the alignment of transition sequences between the anchor and each candidate as a secondary feature, rewarding candidates whose plot-shift patterns mirror those of the anchor. When the two sub-signals disagree ,i.e. the candidate with the lower Fréchet distance differs from the candidate with higher transition-sequence alignment, Fréchet distance takes priority, as it captures the global trajectory shape; transition alignment acts as a secondary tiebreaker only when Fréchet distances are within $\delta < 0.005$.

Application. Signal 4 was not integrated as an inline pipeline component. It was applied as a trajectory-based fallback on the subset of development examples where Signal 3 failed to return a valid structural comparison, reducing the vote to a two-signal split between Signals 1 and 2. On this subset, the candidate with the lower Fréchet distance and higher transition-sequence alignment was selected as the final prediction, contributing a marginal gain of $+1.25\%$ over the three-signal baseline.

3.5 Decision Logic

The primary decision is determined by a majority vote of Signals 1-3. Since each signal produces a binary prediction (S_A or S_B), the vote always resolves to a 2-1 or 3-0 outcome under normal operation, a strict three-way tie is not possible. Signal 4 was not part of the live voting pipeline; it was applied as a trajectory-based fallback exclusively on the subset of examples where Signal 3 was unavailable due to API failure, as described in Section 3.4. In those cases, the two-signal contest between Signals 1 and 2 may produce a 1-1 split, which Signal 4 resolves by selecting the candidate with the lower Fréchet distance and higher transition-sequence alignment relative to the anchor.

4 Experimental Setup and Results

4.1 Results

Our system (accessible in this GitHub repository) was evaluated on the official test set for Task 4. As shown in Table 1, the ensemble approach outperforms any individual module. The organizer's GPT-4o-mini baseline compares stories as raw text; our Signal 3 uses `gpt-5-mini` on structured narrative abstractions (Section 3.3).

System / Signal	Test Accuracy
Signal 1 (Action-only)	63.5%
Signal 2 (SSR)	56.0%
Signal 3 (LLM Comparison)	65.5%
Signal 4 (Trajectory, standalone)	–
2-Signal Vote (Signals 1 + 3 only)	64.2%
3-Signal Majority Vote	67.0%
+ Trajectory Fallback (Final)	**68.25%**
Organizer baselines (Track A) (Hatzel et al., 2026)	
Random	50.00%
Jaccard Similarity	56.25%
GPT-4o-mini (raw text)†	67.00%
Best participating system (Track A)	∼78.0%

Table 1: System performance on the Task 4 Track A test set, alongside organizer-provided baselines and the top-ranked participating system on the official leaderboard (Hatzel et al., 2026).

Signal 4 is a trajectory-based fallback; its marginal contribution is captured by the $+1.25\%$ gain over the 3-signal vote on the Signal 3 failure subset. The ablation removing SSR (2-Signal Vote, Table 1) drops performance by 2.8 percentage points ($64.2\% \rightarrow 67.0\%$), confirming that despite Signal 2's low standalone accuracy, it provides a meaningful structural anchor the neural signals alone cannot replicate.

Our final system (68.25%) outperforms all three organizer-provided Track A baselines, including GPT-4o-mini (67.00%). Notably, the organizer's GPT-4o-mini baseline operates on raw narrative texts, whereas our Signal 3 uses `gpt-5-mini` on structured abstraction outputs; Signal 3's lower standalone accuracy (65.5%) reflects this constrained input regime rather than a model capability difference. The gap to the best participating system (∼78%) reflects the abstraction-first design (Section 3.3), trading raw accuracy for interpretability and reduced surface-lexical anchoring. The $+1.25\%$ gain from Signal 4 corresponds to approximately 5 additional correct predictions on the 400-example test set and should be interpreted as indicative rather than statistically confirmed.

4.2 Confidence Scoring

Beyond aggregate accuracy, understanding *when* the ensemble is reliable is operationally valuable, for instance, to flag low-confidence predictions for human review. By observing agreement across signals, we derive a lightweight confidence heuristic that strongly correlates with predictive accuracy (Table 2). The "Low" tier corresponds to

examples where Signal 3 failed and the trajectory-based fallback (Signal 4) was applied.

Confidence	Criterion	% Dev	Acc.
High	All 3 signals agree	34%	76.5%
Medium	2-of-3 signals agree	54%	65.0%
Low	Signal 3 unavailable	12%	58.0%

Table 2: Per-confidence-tier accuracy on the development set.

5 Error Analysis and Case Studies

An analysis of the cases where baseline models failed reveals the value of the trajectory-based fallback (Signal 4).

5.1 When Semantic Gravity Fails

Purely semantic embeddings (Signal 1) frequently collapse distinct themes if vocabulary overlaps heavily. We observed a notable failure case in our development set involving a gothic horror anchor story concerning a widower, a dubious doctor, and village murders. Candidate A was a medical-ethics drama about a doctor and euthanasia of a mentally handicapped child. Candidate B was a gothic mystery involving a restoration artist, insane sisters, and village murders.

When mapped via static semantic centroids, the system incorrectly preferred Candidate A ($L_2 = 0.965$) over Candidate B ($L_2 = 1.024$). The vector space over-indexed on the shared vocabulary of "doctor" and "madness/mental illness," failing to recognize the profound genre disconnect. On this example, Signal 3 failed to return a valid prediction due to an API error, reducing the vote to a 1-1 split between Signal 1 (favouring A) and Signal 2 (favouring B via higher SSR), and triggering the trajectory-based fallback.

5.2 Mitigation via Sequential Trajectory Analysis

This failure was successfully resolved by Signal 4 (Trajectory Analysis). By breaking the texts into sequential narrative chunks and mapping their trajectory curves across the embedding space, the system detected that Candidate B's temporal shape aligned closely with the anchor's trajectory (Fréchet distance: 1.075 for B vs. 1.086 for A). The dynamic analysis correctly identified that the sequence *Mystery* $\rightarrow$ *Murder* $\rightarrow$ *Madness* was a tighter structural match than *Medical Clinicality* $\rightarrow$ *Euthanasia*.

5.3 Remaining Vulnerabilities

Notwithstanding these mitigations, the SSR module (Signal 2) exhibits a persistent bias against implicit causality. When an anchor narrative explicitly states an outcome that is only implied within a candidate story, the dependency parser fails to extract the relevant tuple; this results in a survival ratio that does not accurately reflect semantic alignment. Consequently, instances predicated on latent world-knowledge continue to pose significant challenges for both symbolic parsers and the `gpt-5-mini` architecture. Future work should evaluate whether larger dependency parsers improve SSR quality.

6 Conclusion

The work indicates that optimal system design favors modularity. By utilizing LLMs for flexible semantic representation while anchoring their outputs using deterministic symbolic logic (SSR) and dense action embeddings, our ensemble achieves higher consistency and structural accuracy. The post-hoc application of phase-space trajectory analysis, raising performance to 68.25%, confirms that order-aware trajectory comparison is a valuable complement to static ensemble methods when the primary LLM signal is unavailable.

Several directions merit further investigation. Replacing the hand-crafted SSR with a learned event-extraction model could improve implicit causality coverage, and comparing the abstraction-first LLM design against raw-text comparison would quantify the accuracy–interpretability tradeoff directly. Most importantly, integrating Signal 4 as a live inline component with a principled confidence-based trigger, rather than as a fallback for API failures, would test whether trajectory analysis provides consistent gains beyond failure recovery.

References

Mst Shapna Akter, Md Fahim Sultan, and Alfredo Cuzzocrea. 2025. Neuro-symbolic methods in natural language processing: a review.

Helmut Alt and Michael Godau. 1995. Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(1–2):75–91.

Walid Mohamed Aly, Taysir Hassan A Soliman, and Amr Mohamed AbdelAziz. 2025. An evaluation of large language models on text summarization

tasks using prompt engineering techniques. *arXiv preprint arXiv:2507.05123*.

Emily Barnes and James Hutson. 2024. Natural language processing and neurosymbolic ai: the role of neural networks with knowledge-guided symbolic approaches. *DS Journal of Artificial Intelligence and Robotics*, 2(1):1–13.

Manaswini Bollikonda. 2025. Bridging symbolic logic and neural intelligence: hybrid architectures for scalable, explainable ai.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Abhay Gupta, Kevin Zhu, Vasu Sharma, Sean O'Brien, and Michael Lu. 2025. Novelhopqa: Diagnosing multi-hop reasoning failures in long narrative contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26145–26162.

Yujin Han, Lei Xu, Sirui Chen, Difan Zou, and Chaochao Lu. 2024. Beyond surface structure: A causal assessment of llms' comprehension ability. *arXiv preprint arXiv:2411.19456*.

Hans Ole Hatzel, Ekaterina Artemova, Haimo Stiemer, Evelyn Gius, and Chris Biemann. 2026. SemEval-2026 Task 4: Narrative similarity and narrative representation learning. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Zhe Hou. 2025. Neural-symbolic reasoning: Towards the integration of logical reasoning with large language models. *Authorea Preprints*.

Tianzhe Jiao, Chaopeng Guo, Xiaoyue Feng, Yuming Chen, and Jie Song. 2024. A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Computers, Materials & Continua*, 80(1).

Ákos Kádár, Lan Xiao, Mete Kemertas, Federico Fancellu, Allan Jepson, and Afsaneh Fazly. 2021. Dependency parsing with structure preserving embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1684–1697, Online. Association for Computational Linguistics.

G Kazeminejad, M Palmer, S Brown, and J Pustejovsky. 2022. Componential analysis of english verbs. frontiers of artificial intelligence.

Elektra Kypridemou and Loizos Michael. 2013. Narrative similarity as common summary. In *2013 Workshop on Computational Models of Narrative*, pages 129–146. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. 2024. Making long-context language models better multi-hop reasoners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2475.

Yuanchi Ma, Jiamou Liu, Hui He, Libo Zhang, Haoyuan Li, and Zhendong Niu. 2025. Boundary matters: Leveraging structured text plots for long text outline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 49–63.

Christos Papakostas, Christos Troussas, Akrivi Krouska, and Cleo Sgouropoulou. 2025. A hybrid neuro-symbolic pipeline for coreference resolution and amr-based semantic parsing. *Information*, 16(7):529.

Rex Clark Robinson. 2012. *An introduction to dynamical systems: continuous and discrete*, volume 19. American Mathematical Soc.

Jingxuan Tu, Timothy Obiso, Bingyang Ye, Kyeongmin Rim, Keer Xu, Liulu Yue, Susan Windisch Brown, Martha Palmer, and James Pustejovsky. 2024. Glamr: augmenting amr with gl-verbnet event structure. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7746–7759.

Hannah Waight, Solomon Messing, Anton Shirikov, Margaret E Roberts, Jonathan Nagler, Jason Greenfield, Megan A Brown, Kevin Aslett, and Joshua A Tucker. 2025. Quantifying narrative similarity across languages. *Sociological Methods & Research*, 54(3):933–983.

Lei Yu and Yang Xu. 2021. Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 920–931.