LIAAD INESCTEC at SemEval-2026 Task 4: Unsupervised Narrative Similarity via Discourse Representation Structures and Sentence Embeddings

Evelin Amorim
INESC TEC
Porto, Portugal
evelin.f.amorim@inesctec.pt

Alípio Jorge
INESC TEC, FCUP
University of Porto
Porto, Portugal
amjorge@fc.up.pt

Purificação Silvano
INESC TEC, CLUP, FLUP
University of Porto
Porto, Portugal
msilvano@letras.up.pt

Abstract

In this paper, we describe an unsupervised approach using Discourse Representation Structures (DRS) for the SemEval-2026 Task 4. This task was Narrative Similarity and was formulated in two different tracks. Our team only developed a solution for track A, where the input is composed of a triplet: an anchor story, a story A, and a story B. The output in this formulation is to predict which story, A or B, is more similar to the anchor story. Our approach parsed each story and transformed it into a DRS format, then we leveraged its structure and extracted features, performing ablation experiments in the development dataset. While our strategy matches the 0.5975 accuracy of a strong neural baseline on the official blind test set, we demonstrate that the symbolic layer provides critical explainability and structural insights that surface-level embeddings lack.

1 Introduction

Narrative similarity is a task whose objective is to measure how similar two given narratives are. Beyond semantic surface similarity, two stories can share the same abstract themes, course of action, and outcome (Hatzel et al., 2026). Identifying these nuanced correspondences is helpful for a number of tasks, for instance, the dominant narrative in news (Guimarães et al., 2025) or the interpretation of historical movements (Lai et al., 2021). However, capturing these subtle similarities remains a significant challenge for Large Language Models (LLMs), as standard semantic embeddings and general reasoning mechanisms often fail to isolate structural plot patterns from concrete surface-level details.

The SemEval-2026 Task 4 proposed two tracks to tackle the narrative similarity task. In track A, each instance of the dataset is a triplet composed of an anchor story, a story A, and a story B. The output should be True if the story A is the most similar to the anchor story, or False otherwise. In track B, each instance is only one story, and the output should be an embedding representing the story. Our system paper describes only a solution for track A. In our solution, we employed a symbolic representation combined with a neural Pretrained language model (PLM). The small size of the test set (400 triplets) limits the generalization of purely neural approaches and highlights the black-box nature of sentence embeddings, which offer no transparency into why two stories are deemed similar. A symbolic representation of narratives is powerful because it captures underlying structural features. This paper serves as a baseline study to explore whether these structural features can complement or even explain the decisions made by neural models.

A symbolic representation of narratives can be powerful because it captures their underlying features. Chambers and Jurafsky (2009) proposed a technique based on coherent event chains to train an unsupervised model to be applied in a Cloze test. The results indicated that the correct verb was recovered within the returned candidate list in 714 of the 740 Cloze tests, showing the model's strong recall of narrative event knowledge. A more common symbolic representation is the use of the Abstract Meaning Representation (AMR). Shou et al. (2022) employed AMR to measure the similarity between sentences in seven STS datasets (Agirre et al., 2012, 2013, 2014, 2015; Bethard et al., 2016; Cer et al., 2017). The authors compared with contrastive learning approaches, and the AMR-based strategy presented the best Spearman correlation results. Finally, Amorim et al. (2024) proposed a toolkit that represents a narrative into a Discourse Representation Structure (DRS) proposed by Kamp and Reyle (1993), but its application for narrative similarity remains unexplored.

The Discourse Representation Theory (DRT) is a linguistic theory whose goal is the modeling

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2193–2199

of the underlying semantic structure of the discourse (Kamp and Reyle, 1993; Geurts et al., 2024). To measure narrative similarity, we propose to employ DRS. In our proposal, this structure comprises four layers for the representation: entities (event and participant), semantic relations, temporal relations, and the logical layer. The combination of these layers allows the extraction of rich features that will represent each story. Although the accuracy of our proposal (0.5975 in the blind test set and 0.6000 in development) remains at parity with text-only baselines, this work provides a "cautionary tale´´ regarding the difficulty of capturing structural plot similarity. Crucially, our approach offers a level of explainability—dissecting specific event densities and grammar profiles—that purely semantic models cannot provide.

The rest of the paper is organized as follows. Section 2 describes the extracted features and the computation of the similarity. Section 3 introduces the dataset, and provides some statistics on the extracted features. Section 4 presents the ablation experiments and their results. Section 5 offers a brief error analysis. Finally, Section 6 concludes the paper.

2 Method

Our proposal consists of two stages: the first parses the story into a DRS file, and the second extracts features and computes similarities between stories. We describe these two phases in detail below.

2.1 DRS construction

To build the DRS structure from a text, we employed the text2story Python toolkit (Amorim et al., 2024). In this toolkit, PLMs can be used in a pipeline to identify the main narrative components (events, participants, and temporal expressions) and their relations (temporal, semantic, and coreferential). Table 1 details the models for each component. For components where multiple models are used (participants and events), the toolkit employs a hybrid merging strategy. Regarding event identification, the system performs a union of attributes across multiple annotators. While for participant identification, candidates are validated based on two primary criteria: (1) an argument extracted via Semantic Role Labeling (SRL) is confirmed as a participant if it is cross-validated as a Named Entity by either SpaCy or NLTK, and (2) in the absence of a Named Entity match, an SRL argument is retained if its lexical head is identified as a noun or as a pronoun.

Entity	Model / Source
Participant	SpaCy (`en_core_web_lg`) NLTK (`maxent_ne_chunker`) SRL (Oliveira et al., 2021)
Event	Lusa BERT Event Type (Amorim, 2025) SRL (Oliveira et al., 2021)
Time Expression	HeidelTime (Strötgen and Gertz, 2010)
Semantic Relations	SRL (Oliveira et al., 2021)
Temporal Relations	Rule-Based Approach[a]
Coreferential Rel.	Maverick (Martinelli et al., 2024)

Table 1: Summary of models employed for narrative component extraction.

[a] The temporal inference engine enriches the Discourse Representation Structures (DRS) by applying rules grounded in Reichenbach's temporal framework and Vendler's aspectual classes (Altshuler, 2016), which categorize events (situations) into states (non-dynamic situations), processes (dynamic, durative, atelic situations), or transitions (dynamic, telic situations). These linguistic principles are operationalized through a combination of the text2story toolkit for initial component extraction and a specialized NLTK-based inference engine that establishes logical ordering, such as narrative progression and causal-temporal anchoring.

After this processing, we obtain a file whose structure is similar to the mockup example in Table 2[1]. An event is represented by a variable (in this example `d`, the verb *stolen*), which `occursBefore` an event `f` (the verb *goes*), and `completedBefore` the utterance time. Other attributes of the event are also incorporated in the variable definition (`Type, PoS, Tense, Aspect, Vform`). The participant is represented by a unique ID (T1 is "The old grandmother Tina"), which has a semantic relation of type agent with the event `a`. Another two participants, T47 and T48, represent the same entity, and have a coreferential relation linking them (`objIdentity`).

2.2 Features

We developed two types of features: DRS-based features and the text-based features. The text

[1] This is only an excerpt of the story from the dev dataset that is: "The old grandmother Tina arrives in town to attend the wedding of his nephew Alberto with his girlfriend Ileana. Upon arrival she discovers that she has been stolen of a medallion that her late husband had given her. He goes to the police station to file a complaint and get the dear object back, but given the length of the investigation, he decides to carry out the search for the thief himself, combining a great deal of mess. Eventually, by chance, he finds the thief, who lives in the same hotel, also managing to have an entire gang of criminals arrested. The grandson Alberto can marry the beautiful Ileana and the grandmother Tina will be appointed, by merit, an honorary colonel of the female police".

Type	DRS Formal Representation
Event	`([d], [event(d), Pos(Verb), Tense(Past), Aspect(Perfective), occursBefore(d, f), completedBefore(d, now), Vform(Participle)])`
Actor	`T1: The old grandmother Tina`
Relation	`agent(a, T1) ∧ objIdentity(T47, T48)`

Table 2: Mockup of the DRS structure including temporal constraints and coreferential relations.

feature is the cosine distance between the anchor story and a given story using the model `all-MiniLM-L6-v2` as the representation. The DRS features consist of 24 structural metrics, including counts and ratios for narrative components (events, participants), grammar profiles (aspectual, tense, aspect), and relational patterns (temporal n-grams, semantic role links, and logic-enriched distributions). The 24 features were derived from the formal components of DRS representations using counts and overlap ratios. While the choice of DRS components is theoretically motivated by the formalism (Kamp and Reyle, 1993), the specific operationalization of each feature was empirically driven. The code for our implementation and additional details can be found in our repository[2].

3 Data

Regarding the data set, for the development of the strategy, we used the development set published by Hatzel et al. (2026). The submission of the system was made using test data also published by Hatzel et al. (2026). Additional statistics using the PLMs employed by our pipeline are described in Table 3.

Statistic	Dev Set		Test Set	
Narrative Components				
# Events	3,125		18,590	
Avg. Events per Triplet	15.62 7.51	±	15.49 7.35	±
# Participants	7,371		44,189	
Avg. Participants per Triplet	36.85 14.98	±	36.82 15.76	±
Narrative Relations				
# Temporal Relations	4,408		25,588	
Avg. Temp. Rel. per Triplet	22.04 21.56	±	21.32 23.38	±
# Semantic Relations	2,249		14,007	
Avg. Sem. Rel. per Triplet	12.24 5.61	±	11.67 5.12	±
# Coreferential Relations	2,860		18,538	
Avg. Coref. Rel. per Triplet	14.30 7.60	±	15.45 8.46	±

Table 3: Detailed statistics of the dataset, partitioned by development and test sets. Values are presented as total counts and averages per triplet with standard deviation (±).

4 Experiments

estrtur In Table 4, the DRS-only experiments focus on features extracted from the DRS file. We selected the most complex one to experiment with and analyze the results. The hybrid experiments were tests that weigh the cosine text similarity and the DRS feature similarities. Several weights were tested to observe the relevance for each type of aspect, the semantic similarity of the text, or the structure similarity of the developed features.

We establish as a baseline a simple cosine similarity with a SentenceBERT model (`all-MiniLM-L6-v2`). Comparing this with a DRS-only approach in the development set, it is possible to notice that three of the features were able to achieve the same result: temporal_relation_semantic, cosine, and event_type. The temporal_relation_semantic, unlike simple overlap counts, performs a 'fuzzy' match by: (1) collapsing specific temporal labels into functional categories (*Sequential* vs. *Simultaneous*), (2) using BERT embeddings to find the most similar event-pairs across stories (e.g., matching 'arrives-attends' in Story A to 'appears-joins' in Story B), and (3) calculating a symmetric similarity score based on the average max-alignment of these temporal triples. This aligns with the aspect of the course of action within a story. The cosine features get all the DRS features as feature vectors and perform a cosine similarity calculation between two stories. Each feature vector includes the normalized ratios of all numerical discourse properties, such as event aspectual types (Transition, Process, State), temporal densities, and semantic role distributions (e.g., Agent-to-Event ratios). The event_type feature is based on the three core event ratios from the DRS: State (static situations), Process (durative, atelic actions, like dance), and Transition (changes of state, telic actions, like "arrives´´ or "arrested´´). These ratios are treated as a probability distribution that defines the pace of the narrative. For example, a story with a high Transition ratio is seen as fast-paced, with more narration segments, while one with a high State

[2] https://github.com/evelinamorim/SemEval-2026-Task-4/

Method	Accuracy (Dev)	Accuracy (Test)
Baselines		
Text-only (SBERT)	0.5500	**0.5975**
DRS-only Experiments		
DRS-only (cosine)	0.5500	**0.5850**
DRS-only (event_type)	0.5500	0.4850
DRS-only (tense)	0.5000	0.4900
DRS-only (temporal_relation)	0.4950	0.4800
DRS-only (temporal_relation_semantic)	**0.5550**	0.5300
DRS-only (temporal_density)	0.5050	0.5200
DRS-only (event_count_ratio)	0.5100	0.5250
DRS-only (event_sequence)	0.4800	0.4875
DRS-only (event_trigram)	0.4950	0.4800
DRS-only (logic_overlap)	0.5100	0.4525
DRS-only (logic_sim)	0.4950	0.4800
DRS-only (aggregate)	0.5200	0.4875
Hybrid Experiments		
Hybrid simple (DRS + SBERT)	0.5900	0.5700
Hybrid sweep (text=0.20, struct=0.80)	0.5300	0.5175
Hybrid sweep (text=0.25, struct=0.75)	0.5500	0.5225
Hybrid sweep (text=0.30, struct=0.70)	0.5600	0.5250
Hybrid sweep (text=0.35, struct=0.65)	0.5700	0.5275
Hybrid sweep (text=0.40, struct=0.60)	0.5850	0.5425
Hybrid sweep (text=0.45, struct=0.55)	0.5650	0.5425
Hybrid sweep (text=0.50, struct=0.50)	0.5850	0.5600
Hybrid sweep (text=0.55, struct=0.45)	0.5900	0.5700
Hybrid sweep (text=0.60, struct=0.40)	0.5850	0.5775
Hybrid sweep (text=0.65, struct=0.35)	**0.6000**	0.5900
Hybrid sweep (text=0.70, struct=0.30)	0.5950	0.5900
Hybrid sweep (text=0.75, struct=0.25)	0.5900	**0.5925**
Hybrid sweep (text=0.80, struct=0.20)	0.5650	0.5900
Hybrid sweep (text=0.85, struct=0.15)	0.5600	0.5875
Hybrid sweep (text=0.90, struct=0.10)	0.5550	**0.5925**

Table 4: Consolidated ablation results on Development and Test sets (Track A). The best scores for each experiment category (DRS-only and Hybrid) are highlighted in bold.

ratio is seen as more descriptive. Event ratios are compared using Jensen-Shannon (JS) Divergence, a symmetric measure of distance between probability profiles, which is then transformed into a similarity score.

Regarding the test set, there is some variation in the importance of the features. Still, the representation of the overall nature of the narrative, conveyed by the DRS cosine similarity feature, is in the top features. However, in the test set, the text seems to play a more relevant role. The best performance for the development set has text with a weight of 0.65 and for the DRS features a weight of 0.35. Differently, in the test set, the best results are with the text weights as 0.75 and 0.5925. Additionally, the result in the test set is different from the CodaBench, which was 0.5975. This is probably due to environmental differences between the CodaBench, and ours. Nonetheless, the trend is still clear from these experiments, hybrid methods presents superior performance compared to the DRS-only strategies.

5 Error Analysis

To better understand the interaction between semantic embeddings (SBERT) and symbolic structures (DRS), we analyzed specific instances where the Hybrid Simple model (Text=0.55, Struct=0.45) failed to match the Ground Truth.

The Table 5 shows a summary of an instance from an instance of the development set. The text similarity of Story A is higher, 0.27, than that of the Story B, 0.21. However, the structure similarity is much higher for Story A, which pushes the final result for the Story A. Table 6 dissects even more this example. The Story B has significantly less events and participants than Story A. Another observation is regarding the grammar profile of both stories. The PLM used in the event tense classification, Spacy, classified all the events in the Story B as present tense, which is correct. However, regarding the tense profile, Story A is more similar to the anchor. For the classification of the aspectual types of events, the PLM (Lusa BERT Event type) used

Narrative Summary	EC	PC	Txt	Str	Tot
Anchor: May, a single mom, wants to become an actress. Her next-door neighbor August, a bodybuilder, wants to become a worthy successor to his hero, Arnold Schwarzenegger. They live in a district of Los Angeles known as Echo Park, not far from Dodger Stadium, and dream of a better life. Jonathan, a pizza delivery boy, arrives at May's door and is immediately smitten with her. As he entertains her young son, Henry, she goes out to pursue an acting opportunity that has come along, only to discover that it involves disrobing in private residences, delivering "strip-o-grams." May gives the strip-o-grams a try and August tries to meet his idol at a reception at the Austrian embassy, while Jonathan worries that the two are more than mere neighbors.	14	37	–	–	–
Story A (Wrong): Jane Falbury (Judy Garland) is a farm owner whose actress sister, Abigail (Gloria DeHaven), arrives at the family farm with her theater troupe. They need a place to rehearse, and Jane and her housekeeper, Esme (Marjorie Main), reluctantly agree to let them use their barn. The actors and actresses, including the director, Joe Ross (Gene Kelly), repay her hospitality by doing chores around the farm. Although Joe is engaged to Abigail, he begins to fall in love with Jane after Abigail leaves him in an angry fit. Similarly, although Jane is engaged to Orville (Eddie Bracken), she falls in love with Joe.	13	45	**0.27**	**0.91**	**0.59**
Story B (Truth): Harold Hall, a young man with little or no acting ability, desperately wants to be in the movies. After a mix-up with his application photograph, he gets an offer to have a screen-test, and goes off to Hollywood. At the studio, he does everything wrong and causes all sorts of trouble. But he catches the fancy of a beautiful actress, and eventually the studio owner recognizes him as a comic genius.	6	17	0.21	0.53	0.54
Result: Predicted **A** | The ground truth is **B** (**Model Failure due to Structural Mimicry**)					

Table 5: An instance of the development set for error analysis. EC (Event Count) and PC (Participant Count) illustrate how Story A's structural density (i.e., features values) mirrors the Anchor, leading to an inflated Structural Similarity (Str) score compared to the shorter, thematically closer Story B.

shows a bottleneck since the events in Story B had no type recognized by the model. In the page of the model, the reported overall macro-$f1$ for the three types is 0.39, which is a considerably low performance.

Feature Distribution	Anchor	Story A	Story B
Structural Scale			
Event Count	14	13	6
Participant Count	37	45	17
Grammar Profile			
Present Tense %	71.4%	69.2%	**100%**
Past Tense %	14.3%	15.4%	0.0%
Transition Type %	7.1%	7.7%	0.0%
Event Logic Sim	–	**0.9880**	0.9535

Table 6: Detailed feature comparison for example in Table 5. Story A mimics the Anchor's multi-modal distribution (tenses/types), whereas Story B's shorter length results in a singular, sparse profile.

In a more quantitative error analysis in the test set, we observe that the systems (DRS vs SBERT) agree in 48.5% of the cases, showing that a complementary architecture is justified. Analyzing short stories, like Story B from Table 5, SBERT has a strong advantage. For stories with 6 to 15 events, the accuracy of baseline SBERT reaches 60.7%, while DRS has 51.3%. The gap decreases for stories with 16 to 30 events, with SBERT achieving 57.7% and DRS reaching 52.6%.

6 Conclusion

We presented an unsupervised system for narrative similarity (Track A of SemEval-2026 Task 4), combining Discourse Representation Structure features with sentence embeddings. Stories are first parsed into DRS format using the text2story toolkit, extracting events, participants, temporal relations, semantic relations, and coreference chains. Structural features derived from these layers are combined with cosine text similarity via a weighted hybrid score, requiring no training. Our system achieved 0.5975 accuracy on the official blind test set. Ablation experiments on the development set show that hybrid methods consistently outperform both text-only and DRS-only baselines, and that the contribution of DRS is modest but stable across evaluation conditions. The results confirm that structural narrative representations can complement surface-level text similarity, even in an entirely unsupervised setting. There are, however, some limitations in the proposal. Despite being based on DRS, the extracted features capture more stylistic or distributional patterns than narrative structure itself, except for the temporal relation semantic feature, which

partially captures the narrative course of action. Additionally, the experiments rely on a lightweight sentence encoder; with a stronger language model as the text baseline, the marginal contribution of DRS features may diminish further. Finally, the presented experiments lack statistical significance, and the small dev set (200 examples) limits reliable hyperparameter tuning.

The current system represents each story as a fixed-dimensional feature vector derived from aggregate statistics of its DRS components, which discards structural information encoded in the graph topology. A natural extension is to replace these hand-crafted features with learned graph representations. Concretely, each DRS layer could be encoded by a dedicated neural module: a Graph Neural Network for the temporal event graph, a participant graph encoder for coreference chains, and a multilayer perceptron over logic predicate distributions. The resulting component embeddings would then be fused into a single story embedding trained end-to-end with triplet loss. This architecture would directly address Track B of the task, which requires a standalone story embedding rather than a pairwise comparison score. Early experiments in this direction during the development phase suggested that the temporal and logic components carry the most discriminative signal, consistent with the ablation results reported here. A second direction concerns the weight combination strategy. The current system uses a fixed weighted average tuned manually on the development set. Learning these weights from supervision — for instance, via a small logistic regression head over the similarity scores — could improve robustness, particularly as the dataset grows. A third direction can be refining the model that classifies the event type. Preliminary results indicate that LLMs can be very effective in few-shot learning of this feature, and a more sophisticated architecture can beat the current simple BERT model of the Lusa BERT Event Type. Finally, the DRS pipeline employed here is English-only. The text2story toolkit supports Portuguese, and extending the approach to other languages would be a straightforward adaptation, making the method applicable to multilingual narrative analysis tasks such as cross-lingual dominant narrative detection in news corpora.

Acknowledgments

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2025 (https://doi.org/10.54499/UID/50014/2025). The authors would also like to acknowledge the project StorySense, with reference 2022.09312.PTDC (https://doi.org/10.54499/2022.09312.PTDC).

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Daniel Altshuler. 2016. *Events, States and Times: An Essay on Narrative Discourse in English*, 1 edition. De Gruyter, Berlin/Boston.

Evelin Amorim. 2025. Bert-lusa event type classifier. `https://huggingface.co/evelinamorim/bert-lusa-eventype-classifier`. Accessed: 2026-02-27.

Evelin Amorim, Ricardo Campos, Alipio Jorge, Pedro Mota, and Rúben Almeida. 2024. text2story: A

python toolkit to extract and visualize story components of narrative text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15761–15772, Torino, Italia. ELRA and ICCL.

Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors. 2016. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Bart Geurts, David I. Beaver, and Emar Maier. 2024. Discourse representation theory. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, winter 2024 edition. Metaphysics Research Lab, Stanford University.

Nuno Guimarães, Purificação Silvano, Ricardo Campos, Alipio Jorge, Ana Filipa Pacheco, Dimitar Iliyanov Dimitrov, Nikolaos Nikolaidis, Roman Yangarber, Elisa Sartori, Nicolas Stefanovitch, Preslav Nakov, Jakub Piskorski, and Giovanni Da San Martino. 2025. NarratEX dataset: Explaining the dominant narratives in news texts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20408–20434, Suzhou, China. Association for Computational Linguistics.

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.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht.

Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400, Online. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Sofia Oliveira, Daniel Loureiro, and Alípio Jorge. 2021. Improving portuguese semantic role labeling with transformers and transfer learning. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE.

Ziyi Shou, Yuxin Jiang, and Fangzhen Lin. 2022. AMR-DA: Data augmentation by Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3082–3098, Dublin, Ireland. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.