**MILLER — Automated Comparison**

<1 figure, 4 tables, and equations>

**Automated Comparison of Narrative and Character Function Similarity Using Graph Theory**

Miller, B., Shrestha, A., Olive, J. and Gopavaram, S.

Automated comparison of narrative structure facilitates narrative stylometry, cross-document co-reference, and the alignment of events in the retellings of stories from different perspectives. With attention to the second and third outcomes, we introduce a method for the unsupervised population and comparison of graph representations of narrative texts. Populated with extracted story elements drawn from documents, graph substructure similarity analysis techniques are used to measure the similarity of narrative order across stories and for the similarity of character function across those stories. Designed to scale, our first application of this method is against a test corpus of 10 variations of the Aarne-Thompson type 333 story, ‘Little Red Riding Hood’. Preliminary experiments correctly coreferenced differently named entities from story variations and indicated the relative similarity of events in different iterations of the tale despite their appearing in differing order. Next steps include heterogeneous corpora, non-fiction corpora, and more precise comparison.

# \* \* \*

Navigating graphs of sentence-level grammars has been an achieved goal of natural language processing (NLP), at least in English, since the development of the Penn Treebank and the success of statistical parsers in the mid-1990s (Magerman, 1995). Adapting sentence-level approaches to the problem of higher-level analysis such as cross-document correlations of narrative remains an open challenge. The goal of this research is to advance our prior work in narrative information extraction (Miller et al., 2013) and visualization (Shrestha et al., 2013) for narrative similarity assessment, event alignment, and cross-document coreference.

Analytical problems of core interest to humanities scholars are not bound to sentential units. They focus on semantics, representation, ideology, influence, pragmatics, narrative—problematics that frequently bridge, fracture, and co-referentially scatter throughout documents and corpora. Discourse analysis (Jørgensen and Phillips, 2002) and textTiling (Hearst, 1997) are two methods used to circumvent sentential boundaries by segmenting documents into logical blocks. As with latent semantic analysis (Dumais, 2004), these blocks can then be characterized in ways that facilitate comparison, extending sentential navigation to macro-level structures. Our position builds on tree-based textual analysis to argue that comparison of the graph representations of narrative documents using techniques from graph theory advances both NLP and humanistic inquiry. This work scales as extraction, lookup, and comparison operate unsupervised. Identifying events described in a document, producing a graph of actors and locations implicated in those events, and comparing the underlying matrices facilitate various types of cross-document narrative analysis. Graph similarity measures provide our primary comparison method to highlight similarities, as exemplified in Figure1.

Additionally, condensing narrative texts in a manner consistent with narratological theory enables testing of narratological schema. Formal analysis of narrative has been a relatively successful aim of many.

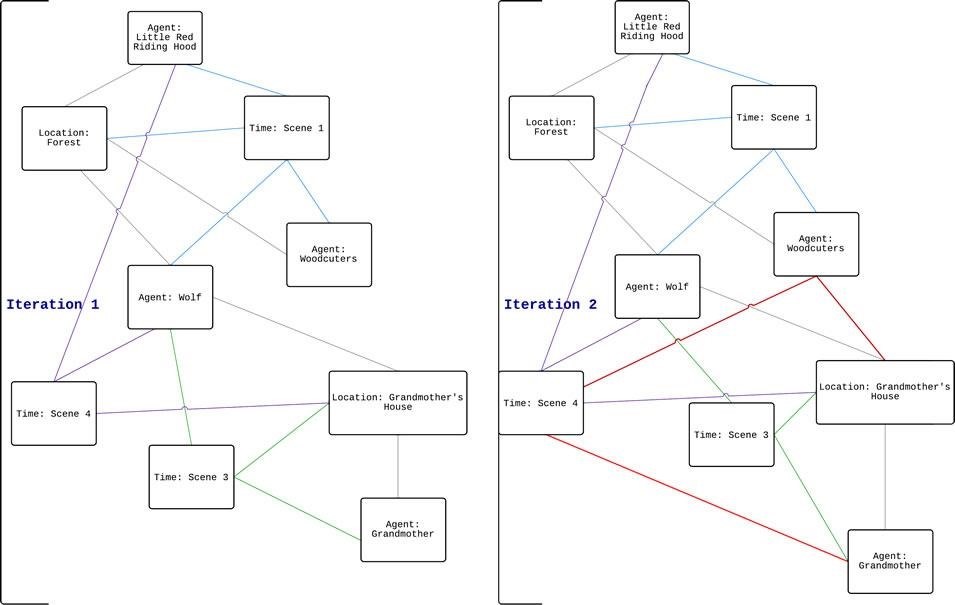


Figure 1. Side-by-side comparison of two simplified iterations of the ‘Little Red Riding Hood’ fairy tale. Linked location, time, and agent units are shown, with red edges indicating change between tellings.

Structuralist topologies developed by the Russian formalists such as Propp (1968), and later work by Genette (Genette and Lewin, 1983), Bal (1997), and others, yielded many complementary top-down models for organizing narrative. These schema distinguish between fabula and discourse, or events to be retold and the manner in which retelling takes place. Narrative, broadly defined, is the ordering of fabula by discourse. Discourse order is the relationship between the temporality of events and their representation as part of a narrative (Genette and Lewin, 1983). This structural perspective serves humanists well when analyzing single narratives or small corpora.

Computational models developed from formalist approaches have been the subject of compelling experiments. Like work by Finlayson (2012) on analogical story merging and Fisseni on story comparison (Fisseni and Löwe, 2014), our work presents a bottom-up method to help test formalist top-down narratological schema. Unlike theirs, our work implements unsupervised comparison of narrative graphs, a proxy of narrative structure, across a corpus.

Narrative graphs facilitate (1) stylometric comparison, (2) clustering of documents according to their event-relationship characteristics, (3) cross-document co-reference, and (4) alignment of event descriptions across narratives. To test three and four, we implemented a method as described below.

**Method**

Computation of similarity between the substructure of the graphs required extraction of entities, creation of the graph, and finally comparison.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | lrrh | wolf | grandmother | woodcutters | forest | gm house |
| lrrh | 1 | 0 | 0 | 1 | 1 | 0 |
| wolf | 0 | 1 | 0 | 0 | 1 | 0 |
| grandmother | 0 | 0 | 1 | 0 | 0 | 0 |
| woodcutter | 1 | 0 | 0 | 0 | 1 | 0 |
| forest | 1 | 1 | 0 | 1 | 0 | 0 |
| gm house | 0 | 0 | 0 | 0 | 0 | 1 |

Table 1. Adjacency matrix created from one version of ‘Little Red Riding Hood’. An edge in the graph or 1 (in the adjacency matrix) between two entities signifies that these entities interacted within the given set of events.

1. *Extraction.* Events were automatically marked in the narratives using the NLP tool, EVITA (Saurí et al., 2005). EVITA increments event tag ID numbers; that sequence proxies discourse order. Sequential narrative discourse order generated from EVITA was used since fiction frequently lacks dates and timestamps—necessary features for temporal taggers like SUTime and GUTime. Entity extraction and classification were accomplished using the Stanford Named Entity Recognizer (NER). Each graph includes elements corresponding to event name, event type, actor, location, and discourse order. Extracted events from a given story contain the following information: document *ID*, event *verb*, event *verblemma*, event *verbhypergram*, event *type*, entity 1, entity *n*, entity 1 type, entity *n* type, event *number*.

2. *Graph creation.* Based on changes in topic, the set of events *E* was divided into *k* parts. A sub-graph *G* was created for each event set, *e*0 *< ex < ek* and was represented as an adjacency matrix as shown in Table 1. An adjacency matrix is a common way of internally representing graphs for computation such that an element *aij* in the matrix equals 1 if there is an edge between *ith* and *jth* nodes in the graph and 0 otherwise. In our case, the adjacency matrix represents the entity::entity relation as mentioned in the narrative for the set of events *ex*.

3. *Similarity analysis.* Many domain-specific algorithms to compute similarity have been developed. Most are based on neighborhood analysis. In this paper, we propose our own similarity analysis method inspired by the work of Blondel et al. (2004).

We begin by modeling each narrative as a 3*D* matrix where the first two dimensions represent the adjacency matrix and the third dimension represents the discourse order of the sets of events. Following this model, given two matrices, *Ai,j,k* and *Bp,q,r*, for comparison, we construct a 3*D* matrix, *X*, of dimensions *p*×*q* ×*φ* where *φ* is the smallest of *k* and *r*, initialized to all 1s. We then proportionally shrink the larger timeline to match the smaller one so that *k* = *r*.

For each *φ*, as in the HIT-inspired algorithm (Fackler, 2005) proposed by Kollias et al., we compute

*X* ← *BXAT* + *BT XA* (1)

and normalize *X* after each iteration. The even iterations converge to a final similarity matrix. To simplify and speed this process, we use the Kronecker product and the vec(.) operator. This results in

*x* ← (*A*⊗*B* + *AT* ⊗*BT* )*x* (2)

where *x* = *vec*(*X*). These sets of equations give a similarity score frame per scene, which is then aggregated to produce a final similarity score.

**Discussion**

For the purposes of testing our methodology, we used 10 of the 58 known iterations of the Aarne-Thompson type 333 story (ATU333), ‘Little Red Riding Hood’ (Lang, 1891; Grimm and Grimm, 1812; Wratislaw, 1889; Schneller, 1867; Ashliman, 2007; Marelles, 1895; Potter, 1908; Bates, 1883; Thurber, 1983; Carter, 1993; Tehrani, 2013). The purpose of using this corpus was to strengthen the possibility of narrative overlap and focus the method on fine-grain distinctions between re-tellings. The primary purpose of the current work is to test our method’s detection of event and character similarities and differences across a given corpus; however, we do not anticipate that our method is dependent on such homogeneous data. By utilizing categories of variables associated with elements of the fabulaic and discourse levels of narrative, we anticipate our method will function similarly with larger, more heterogeneous corpora.

Via this method, 1,384 events were extracted across 10 story iterations. Numbering 8,450 tokens, including titles and authorship information, the overall density of extracted events to tokens is high. Contrasted to event detection methods reliant on temporal expressions (SUTime identified two events), this density of event detection provides a good basis on which to compare narrative structure. Generalizing event types from specific verb tokens to hypergrams of those verbs (e.g., event 41 from Carter, 1993): ‘armed’ lemmatized to ‘arm’ hypergrammed to ‘supply’) generally retains the function of each event within the story, but abstracts sufficiently to allow for variation such as is introduced by translation to not prevent the recognition of correspondences. The automatically produced matrices for this work are exemplified by Table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| undergo,EVENT104 | grandmother | wolf | lrrh |
| bed | 1 | 1 | 1 |
| perceive,EVENT105 | grandmother | wolf | lrrh |
| bed | 1 | 1 | 1 |
| undergo,EVENT106 | grandmother | wolf | lrrh |
| bed | 1 | 1 | 1 |
| perceive,EVENT107 | grandmother | wolf | lrrh |
| bed | 1 | 1 | 1 |
| undergo,EVENT108 | grandmother | wolf | lrrh |
| bed | 1 | 1 | 1 |
| seize,EVENT109 | grandmother | wolf | lrrh |
| bed | 1 | 1 | 1 |
| undergo,EVENT110 | grandmother | wolf | lrrh |
| bed | 1 | 1 | 1 |
| consume,EVENT111 | grandmother | wolf | lrrh |
| bed | 1 | 1 | 1 |

Table 2. Eight matrix layers from 3*D* stack of event matrices.

Table 2 shows eight layers from the 3*D* event matrix stack. It, and the following list, a one-line extract, highlight the generalized form and granular detail of the events this method extracts: 10; prospered; prosper; change state; OCCURRENCE; child, grandmother’s house; person, location; 45. The stack and list correspond to the ‘Oh, grandmother, what big ears you have!’ to ‘[a]nd with that he jumped out of bed, jumped on top of poor Little Red Cap, and ate her up’ sequence from Lang (1891).

Results of functions (1) and (2) on the adjacency matrices are exemplified below in Table 3. Column headings correspond to entities from Grimm and Grimm (1812) for event 3, and row headers correspond to entities from Lang (1891) for event 4. Table 3 shows that the measure of similarity between Little Red Riding Hood (‘lrrh’) and Little Red Cap (‘lrc’), is 0*.*32. Although low, that score was calculated only based on entity-entity connections, the sequence of those connections, and event hypergrams. These preliminary results are encouraging, as the correlation between those characters is high relative to all correlations found. More importantly, as shown by the diagonal symmetry of the measures, the correlation between events is high relative to correlations between that event and others within the stories. Overall correlation between event 3 from one variation and event 4 from another variation suggests that this method can align scenes across variations of the same ATU. Table 4 shows the potential for this method to align characters from different versions based upon their position within the story.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | lrc | wolf | grandmother | huntsman | home | woods | gm house |
| lrrh | .32 | .25 | 0 | .25 | 0 | .32 | 0 |
| wolf | .32 | .25 | 0 | .25 | 0 | .32 | 0 |
| grandmother | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| woodcutters | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| home | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| forest | .32 | .25 | 0 | .25 | 0 | .32 | 0 |
| old woman’s house | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 3. Character similarity across ‘Little Red Riding Hood’ and ‘Rothkppchen’.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | lrrh | wolf | gm | wc | home | forest | owh |
| lrc | **.67** | .76 | .31 | .14 | .14 | .48 | .37 |
| wolf | .79 | **.94** | .42 | .14 | .14 | .56 | .5 |
| gm | .35 | .47 | **.31** | 0 | 0 | .16 | .37 |
| hunts | .23 | .28 | .18 | **0** | 0 | 0 | .26 |
| home | 0 | 0 | 0 | 0 | **0** | 0 | 0 |
| woods | .48 | .53 | .16 | .14 | .14 | **.48** | .16 |
| gmh | .39 | .52 | .34 | 0 | 0 | .16 | **.42** |

Table 4. Character alignment across all events for iterations Lang (1891) and Grimm and Grimm (1812) of ATU 333.

Table 4 sums all event matrices for two variations of the tale. Version 1 occupies the columns (Little Red Riding Hood, Wolf, Grandmother, Woodcutters, Home, Forest, and Old Woman’s House) and version 2 the rows (Little Red Cap, Wolf, Grandmother, Huntsman, Home, Woods, Grandmother’s House). Name-independent character alignment is demonstrated by the 0*.*96 correspondence between the two wolves. Interestingly, the event matrix suggests that certain characters function dissimilarly between variations: most notably, Grandmother. The corresponding value between the Grandmother characters is only 0*.*31, suggesting that they share some event associations but not as many as are held by other cross-document pairings. That assessment is accurate as, in version 1, the story concludes upon the wolf’s consumption of both Little Red Riding Hood and Grandmother. In story version 2, both survive to boil a second hungry wolf.

**Conclusion and Further Work**

This preliminary work resulted in a viable method for automated narrative stylometrics across iterations of stories, for narrative alignment, and for the cross-document coreference of characters bearing different names but similar story functions. The extraction and matrix comparison methods are implemented and tested. Our next stage of this research is to refine the comparison algorithm, apply it to a corpus of dissimilar narratives, explore the fit of structuralist models of narrative to non-fiction, and assess the method’s ability to cluster stories by narrative similarity.

**Acknowledgements**

This work is supported in part by NSF award 1209172.

**References**

**Ashliman, D. L.** (2007). Conte de la mre-grand. Website of Bibliothque nationale de France.

**Bal, M.** (1997). Narratology: Introduction to the Theory of Narrative. University of Toronto Press.

**Bates, C. D.** (1883). *Little Red Riding-Hood.* D. Lothrop and Co.

**Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P. and Van Dooren, P.** (2004). A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching. *SIAM Review,* **46** (4): 647–66.

**Carter, A.** (1993). The Werewolf. In *The Bloody Chamber*. Penguin Books, pp. 108–10.

**Dumais, S. T.** (2004). Latent Semantic Analysis. *Annual Review of Information Science and Technology,* **38**(1): 188–230.

**Fackler, P. L.** (2005). *Notes on Matrix Calculus.* North Carolina State University.

**Finlayson, M. M. A.** (2012). *Learning Narrative Structure from Annotated Folktales.* PhD dissertation, Massachusetts Institute of Technology.

**Fisseni, B. and Löwe, B.** (2014). What Makes Stories Similar? Report on a Research Project, 2011–2014 (Invited Report). In Finlayson, M. A., Meister, J. C. and Bruneau, E. G. (eds), *2014 Workshop on Computational Models of Narrative*, OpenAccess Series in Informatics (OASIcs) no. 41. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 9–12, http://drops.dagstuhl.de/opus/volltexte/2014/4640.

**Genette, G. and Lewin, J. E.** (1983). *Narrative Discourse: An Essay in Method.* Cornell University Press, Ithaca, NY.

**Grimm, J. and Grimm, W.** (1812). Rothkppchen. In Realschulbuchhandlung (ed.), Ashliman, D. L. (trans.), *Kinder-und Hausmrchen*, vol. 1, no. 26, pp. 113–18.

**Hearst, M. A.** (1997). Texttiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics,* **23**(1): 33–64.

**Jørgensen, M. W. and Phillips, L. J.** (2002). *Discourse Analysis as Theory and Method*. Sage.

**Lang, A.** (1891). None Given. In *The Blue Fairy Book*, 5th ed. Longmans, Green, and Company, pp. 51–53.

**Magerman, D. M.** (1995). Statistical Decision-Tree Models for Parsing. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1995, pp. 276–83.

**Marelles, C.** (1895). The True History of Little Golden-Hood. In *The Red Fairy Book*. Longmans, Green, and Company, pp. 215–19.

**Miller, B., Shrestha, A., Derby, J., Olive, J., Umapathy, K., Li, F. and Zhao, Y.** (2013). Digging into Human Rights Violations: Data Modelling and Collective Memory. In *Big Data, 2013 IEEE International Conference on Big Data*, IEEE, pp. 37–45.

**Potter, B.** (1908). *The Tale of Jemima Puddle-Duck*. Frederick Warne and Company.

**Propp, V. I.** (1968). *Morphology of the Folktale.* Vol. 9. University of Texas Press.

**Saurí, R., Knippen, R., Verhagen, M. and Pustejovsky, J.** (2005). Evita: A Robust Event Recognizer for QA Systems. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 700–707.

**Schneller, C.** (1867). Mrchen und Sagen aus Wlschtirol: Ein Beitrag zur deutschen Sagenkunde. In Ashliman, D. L. (trans.), *Verlag der Wagner’schen Universitts-Buchhandlung*, ch. Das Rothhtchen, pp. 9–10.

**Shrestha, A., Zhu, Y., Miller, B., and Zhao, Y.** (2013). Storygraph: Telling Stories from Spatio-Temporal Data. In *Advances in Visual Computing*. Springer, pp. 693–702.

**Tehrani, J. J.** (2013). The Phylogeny of Little Red Riding Hood. *PloS One,* **8**(11): e78871.

**Thurber, J.** (1983). The Little Girl and the Wolf. In *Fables for Our Time and Famous Poems Illustrated*. Harper Collins, p. 3.

**Wratislaw, A.** (1889). Little Red Hood. In Stock, E. (ed.), *Sixty Folk-Tales from Exclusively Slavonic Sources*, no. 15, pp. 97–100.