Minimal Narrative Annotation Schemes and Their Applications

Elahe Rahimtoroghi, Thomas Corcoran, Reid Swanson, and Marilyn A. Walker
Natural Language and Dialogue Systems Lab
University of California Santa Cruz
Santa Cruz, CA, USA
{reid, elahe, maw}@soe.ucsc.edu, tcorcora@ucsc.edu

Kenji Sagae and Andrew S. Gordon
Institute for Creative Technologies
University of Southern California
Los Angeles, CA, USA
{sagae, gordon}@ict.usc.edu

Abstract

The increased use of large corpora in narrative research has created new opportunities for empirical research and intelligent narrative technologies. To best exploit the value of these corpora, several research groups are eschewing complex discourse analysis techniques in favor of high-level minimalist narrative annotation schemes that can be quickly applied, achieve high inter-rater agreement, and are amenable to automation using machine-learning techniques. In this paper we compare different annotation schemes that have been employed by two groups of researchers to annotate large corpora of narrative text. Using a dual-annotation methodology, we investigate the correlation between narrative clauses distinguished by their structural role (orientation, action, evaluation), their subjectivity, and their narrative level within the discourse. We find that each simple narrative annotation scheme captures a structurally distinct characteristic of real-world narratives, and each combination of labels is evident in a corpus of 19 weblog narratives (951 narrative clauses). We discuss several potential applications of minimalist narrative annotation schemes, noting the combination of label across these two annotation schemes that best support each task.

Introduction

At the intersection of computer science and narrative studies there has been an increasing interest in empirical methods, where large corpora of narratives are the subject of analysis (Elson, 2012a; Finlayson, 2012; Gordon et al, 2011; Sagae et al, 2013). This trend parallels progress in computational linguistics, where multiple layers of rich syntactic, semantic and discourse annotation have been applied to large corpora. Rhetorical Structure Theory (Mann & Thompson, 1988) and the Penn Discourse Treebank (Prasad et al., 2008) are two such annotation schemas that have been used successfully to analyze many genres of discourse. These theories of discourse codify the relationships between elementary discourse units (e.g., clauses, phrases, etc.), such as *elaboration*, *contrast* or *justification*. Both of these annotation schemes offer researchers a principled approach to the structural analysis of text, along with software tools that provide some capacity for automated annotation (Soricut and Marcu, 2003; Sagae, 2009; Pitler et al., 2009; Louis et al., 2010).

Much of the discourse analysis in the computational linguistic community has focused on newswire text. In this genre, and other expository discourse, meaning is often conveyed through the use of specific discourse relations between elementary discourse units (EDU), which make these theories highly informative (Mann & Thompson, 1988). However, it has been argued that meaning in narrative is conveyed in a significantly different fashion.

As Elson argues (2012b), meaning in a narrative is usually conveyed through the attitudes, goals and intentions of the characters. He proposes an alternative representation for narrative annotation and analysis (*Story Intention Graph*) that is grounded in literary and psychological theories of narrative. In this framework, textual elements, similar to EDUs are annotated at several layers of analysis that capture the causal and temporal relationships between events, their interpretation in terms of the goals and motivations of the characters, and the affective impact these events have on them.

The detailed structure and richness of these frameworks enable many types of discourse analysis and formal representations useful for automated reasoning, such as narrative generation (Montfort, 2011; Rishes et al, 2013). However, the decision to adopt such a framework has several drawbacks. Many of the existing tools that can produce these analyses automatically are trained on newswire text and do not perform adequately on out-of-domain text. Annotating new data within the target domain can also be problematic. Achieving adequate levels of

L&W	Wiebe	Genette	Story Clause
Orientation	Subjective	Extradiegetic	So this has been a crazy trip already.
Orientation	Objective	Diegetic	We left yesterday morning;
Action	Objective	Diegetic	everything was on time
Action	Objective	Diegetic	and we made it to Houston without any problems.
Action	Subjective	Diegetic	We spent a good chunk of time there,
Action	Subjective	Diegetic	then boarded a flight to Miami which ultimately arrived late…
Orientation	Objective	Diegetic	well, the gate was occupied whenever our plane pulled in.
Action	Objective	Diegetic	So we had to wait for about 20 minutes,
Evaluation	Subjective	Diegetic	and were expecting that the flight to Recife would be a little delayed since there were quite a few people on our plane also going there.
Evaluation	Objective	Diegetic	We were wrong.
Action	Subjective	Diegetic	And the jerk lady at the gate said we couldn't board the flight because she's a big stinky turd..
Evaluation	Objective	Extradiegetic	Well, not really.
Evaluation	Objective	Diegetic	So basically, we missed the flight to Brazil.

Table 1. An excerpt of a personal story annotated using the Labov and Waletzky (L&W) categories from Rahimtoroghi et al. (2013) and according to subjectivity and diegetic levels from Sagae et al. (2013).

inter-rater reliability requires significant expertise and may not always be possible given the complexity of the domain and annotation schema. Even when these hurdles are overcome, it is often expensive and time consuming to acquire a sufficient amount of data useful for machine learning methods. Although annotation tools such as the *Story Workbench* (Finlayson, 2011) and *Scheherazade* (Elson, 2012b) improve the efficiency and accuracy of the annotation process, the burden is still quite large.

Several pieces of recent work on narrative has opted to use annotation schemes that are much simpler for many of the reasons stated above. Rahimtoroghi et al. (2013) describe an analysis of fables based on Labov & Waletzkey's (1997) theory of oral narratives. This schema employs just three labels for narrative discourse units: *evaluation*, *orientation*, and *action*, with no annotation of relational structure. Likewise, Sagae et al., (2013) analyze personal narratives in weblogs using a simple scheme inspired by Wiebe et al. (2004) and Genette (1980). This typology assigns two binary classes to discourse elements: *diegetic* vs. *extradiegetic* discourse and *subjective* vs. *objective* discourse, and also eschews annotation of relational structure. By focusing on a small set of narrative-specific labels, Rahimtoroghi et al. (2013) and Sagae et al. (2013) greatly reduced the costs of annotation with high inter-rater agreement, and provide sufficient quantities of annotated data to train automated annotation software.

The software tools created by these two research groups may be more broadly applicable to narrative studies and applications. However, it is unclear how the different annotation schemes relate to one another, or what sorts of studies and applications each annotation scheme supports. In this paper, we compare and contrast the two annotation schemes used by these research groups. We describe a dual annotation exercise using a set of personal narratives from weblogs, such as the one in Table 1. These were annotated using both schemes by trained annotators in order to calculate the correlation across annotation labels. We then consider the broader applicability of these two annotation schemes in empirical narrative research.

Evaluation, Orientation, and Action

Rahimtoroghi et al. (2013) describes the annotation of 20 narratives from Aesop's Fables using an annotation scheme derived from a typology of narrative clauses from Labov and Waletzky's (1997) theory of oral narrative. In their original formulation, narrative clauses could be classified along three dimensions: temporal, structural, and evaluative points. Rahimtoroghi et al. focused solely on the structural types and evaluations:

Orientation. An *orientation* clause gives information on the time and place of the events of a narrative, the identities of the participants and their initial behavior.

Action. A complicating *action* clause is a sequential clause that reports a next event in response to a potential question, "And what happened then?"

Evaluation. An *evaluation clause* provides evaluation points in the story and information on the consequences of the events as they pertain to the goals and desires of the participants. They also describe events that did not occur, may have occurred, or would occur in the future.

The left column of Table 1 provides examples of each of these structural labels applied to a personal narrative.

Labov and Waletzky defined two additional structural types that were not used in the work of Rahimtoroghi et al. An *abstract* is an initial clause in a narrative that reports the entire sequence of events. Likewise, a *coda* is a final clause which returns the narrative to the time of speaking, precluding a potential question, "And what happens then?" As justification for their omission, Rahimtoroghi et al. argued that the *abstract* label was too ambiguous, and the *coda* label was only applicable to oral narrative. Rahimtoroghi et al. annotated 20 of Aesop's fables (315 clauses) using this annotation scheme, achieving an inter-rater agreement of 89.6% (Cohen's kappa = 0.816). Using this data to train an automatic labeller using machine learning they achieved precision and recall values near 0.90 for all categories except *orientation* (recall 0.45).

Subjectivity and Diegetic Levels

Sagae et al. (2013) describe the annotation of 40 personal stories drawn from public weblogs using an annotation scheme that encodes both subjectivity and the diegetic level of narrative clauses. Annotation involves labeling along two orthogonal dimensions. First, subjective narrative clauses are distinguished from objective, based on Wiebe et al.'s (2004) definition of subjective language.

Subjective clauses express private states, which include emotions, opinions, evaluations and speculations that are not open to external observation or verification by others.

Objective narrative clauses express states that can be externally observed and verified by others.

The second dimension attempts to characterize the different levels at which narratives are told. A narrator provides information about a set of events that transpired in the world of the story and provides information about the world as it is at the time of the narration. Here Sagae et al. adopted terms proposed by Genette (1980) to indicate diegetic levels, simplified to a binary distinction.

Diegetic narrative clauses give information about events as they occurred in the world of the story.

Extradiegetic narrative clauses give information about the world in which the narrator is addressing the reader.

These two dimensions, each with binary classes, yield four distinct labels for narrative clauses. Columns 2 and 3 (Wiebe and Genette) of Table 1 provide examples of these labels applied to the same personal story. Sagae et al. (2013) applied a modified version of this annotation scheme to 40 personal stories drawn from public weblogs (571 segments), achieving an inter-rater agreement of 84% (Krippendorf's alpha = 0.73). Using these annotations to train an automated system, Sagae et al. were able to achieve 78% accuracy on the binary subjective labeling task and 81% accuracy on the binary diegetic task.

Dual Annotation Study

Looking only at the definitions of the labels used by these two research groups, one might expect substantial overlap between certain terms across sets. The *action* label used by Rahimtoroghi et al. shares meaning with the *diegetic* label used by Sagae et al., as both implicate events in the story world. Likewise, the *evaluation* label seems to relate to the *subjective* label, as both are associated with private mental states. However, to really understand the relationship between these terms, it is necessary to consider how they are applied in practice, on real narrative texts.

To investigate the relationship between these terms, we conducted a dual annotation study using human annotators trained by each of these two research groups, each analyzing the same narrative texts. In this study, we selected 19 narratives drawn from public weblogs. Each narrative was authored by a different blogger, and consisted of nonfiction narrative text describing an event from their personal life, e.g. a wilderness camping trip, an automobile collision, and a trip to an amusement park.

An annotator from Sagae et al.'s (2013) original study assigned labels for subjectivity and narrative level, where the inter-rater agreement had already been established with Krippendorf's α of 73%. For the structural labels of Labov and Waletzky, three annotators from Rahimtoroghi et al.'s original study annotated each text, so that inter-rater agreement could be established for this genre. Chance-corrected agreement (Krippendorff's alpha) was 0.574. Although acceptable, this level of agreement was markedly lower than seen when applying these labels to Aesop's fables, and lower than labels of subjectivity and narrative level. In calculating the correlation across labels, the label selected by the simple majority of annotators was used. When no annotators agreed on a label, one of the selected labels was chosen randomly.

Each set of annotators segmented these 19 texts into narrative clauses independently, and then aligned. The more fine-grained segmentation was used in cases of disagreement, replicating annotations for any coarse-grained segment that was divided by this method. In all, these 19 narratives were segmented into 951 clauses.

Table 2 presents examples from the annotated texts of each of the twelve possible combinations of these labels, along with their counts. In considering this table, our assessment is that each combination represents a structurally distinct type of narrative clause, and that these sets of labels are indeed orthogonal.

Table 3 presents the correlation among labels using chi-squared, Pearson's *r*, and mutual information statistics. The chi-squared tests for all pairs yield p-values of less than 10^{-4} indicating significantly low probability of independence. As expected, the highest correlation observed across labels is between *action* and *diegetic* clauses, and between *evaluation* and *subjective* clauses. Strong (negative) correlation is also seen between *evaluation* and *diegetic*, indicating that *evaluation* labels are often assigned to *extradiegetic* narrative clauses.

More surprising is that none of these labels completely subsumes one from the other set. Every combination of *orientation/evaluation/action*, *diegetic/extradiegetic*, and *subjective/objective* labels are represented in the set of 951

Label and Percentage				Id	Examples
Diegetic 59.20%	Orientation 14.93%	Subjective	6.1%	1 2	*I was hiding from writing.* *And one of them was at the meeting last night, cute as a freakin' button.*
		Objective	8.8%	3 4	*Their last show was on Friday, the first day of the festival.* *There was no bus driver around.*
	Evaluation 14.20%	Subjective	11.9%	5 6	*so at first I was a little reluctant* *That was really hard.*
		Objective	2.3%	7 8	*So basically, we missed the flight to Brazil.* *They could be down in an hour.*
	Action 30.07%	Subjective	12.5%	9 10	*He was so insistent, that I decided to give him a bit of a run.* *I found myself looking at him a lot.*
		Objective	17.6%	11 12	*In the middle of the night, I sat up in the futon bed* *the woman behind the counter showed me a book that has the recipe in it.*
Extradiegetic 40.80%	Orientation 15.35%	Subjective	8.6%	13 14	*and it is time for me to get back into the swing of things.* *My brother and I haven't had much of a relationship as adults.*
		Objective	6.7%	15 16	*I've written six pages,* *because it was one of the only sources of sugar in the desert.*
	Evaluation 23.87%	Subjective	22.4%	17 18	*It's been fun trying to use my Portuguese too.* *Life wouldnt be the same without our little man.*
		Objective	1.5%	19 20	*So that's where I am.* *Well, not really.*
	Action 1.37%	Subjective	1.1%	21 22	*However I'm learning.* *In the meantime, I'm doing a lot of praying.*
		Objective	0.3%	23 24	*She would alternate between saying hello to me and saying hello to Andy* *but she'd always do it with a wagging tail and a big grin.*

Table 2. Percentage of each combination of the labels across two annotation schemes in the dataset, along with an example

narrative clauses. This is somewhat unintuitive because one might expect an *evaluation* clause would always be *subjective*, an *orientation* clause would always be *diegetic*, an *action* clause would always be *diegetic* and *objective*. We believe this diverse distribution is primarily a combination of subtle distinctions in the definitions of the L&W categories, the interpretive nature of the annotation process, and the expressivity of language.

As the name suggests, *evaluation* clauses often provide the opinions and emotions of the characters in relation to the actions in the story, which lead one to believe they are always subjective. However, due to subtle nuances in Labov & Waletzkey's definitions, as well as clauses that do not clearly fit their typology (Swanson et al, 2014), this is not always the case. In addition to revealing the characters mental states and attitudes, Labov and Waletzky also consider several other types of functionality to be *evaluative*. These include: *direct speech*, *alternative*, *conditional* or *future timelines*, *inner monologues*, *questions* and some types of *statives* (e.g., the physical or mental consequences of an action). For example, clause (8) in Table 2 is clearly not an actual action in the narrative timeline, but it is an externally verifiable fact.

We also observe orientation clauses from both the *diegetic/extradiegetic* and the *subjective/objective* labels. For example, clause (14) provides background information about the characters, i.e. the narrator and her brother, but it is not stating the events within the world of the story. Alternatively, clause (1) expresses objective background facts within the event structure of the narrative. *Actions* can be expressed objectively, or include opinions such as clause (9) in Table 4. These often occur because an action clause not only expresses *what* happened, but also a subjective interpretation of *how* it happened.

Finally we also notice a small percentage of *extradiegetic actions* despite our expectation that these would be a proper subset of the diegetic clauses. We believe these are likely to be errors in annotation that arise from difficulties in interpretation of the story, which often requires a complete understanding of the entire narrative and the author's intended purpose. For example, clause (21) provides information on the consequences of the previous actions and relates to the desires and goals of the

		χ^2	r	$I(X;Y)$
Orientation	Diegetic	16.747	-0.133	0.013
	Subjective	34.906	-0.193	0.027
Evaluation	Diegetic	116.145	-0.349	0.089
	Subjective	186.146	0.442	0.158
Action	Diegetic	239.912	0.502	0.220
	Subjective	71.082	-0.275	0.054

Table 3. Correlation between labeling schemes using Chi-squared, Pearson's r, and Mutual Information

narrator, and could legitimately be labeled as an *evaluation*. This conclusion is reinforced by the fact that most of these clauses had low agreement between annotators using the Labov and Waletzky schema.

Applications

In comparison to prominent discourse annotation schemes such as Rhetorical Structure Theory or that of the Penn Discourse Treebank, the simple annotation schemes employed by Rahimtoroghi et al. and Sagae et al. enable the quick labeling of large amounts of narrative text with reduced costs for annotation labor. However, this advantage can only be realized in certain applications for which the label sets match the functional needs of the task. In this section, we consider six applications of our annotation schemes, and discuss the particular combination of labels necessary to support the task.

Comparative Analyses. The most straightforward utility of narrative annotation schemes such as these is in the comparative analysis of two narrative corpora. How do narratives written in weblogs compare with those handwritten in personal diaries? How has the structure of narrative in novels changed over time? Questions like these can be efficiently answered by annotating samples of each narrative population, and computing statistical differences. For this purpose, all seven of the labels compared here may be relevant, depending on investigators' research question.

This approach may be most useful when comparing corpora that are written in different languages, where other forms of structural comparison, e.g. syntax, are not applicable. For example, minimalist narrative annotation schemes could be used to efficiently investigate cultural difference between Chinese and American narrative conventions, e.g. those evident in large corpora of personal stories extracted from weblogs (Gordon et al., 2013; Gordon & Swanson, 2009). Native speakers of Mandarin and English could each annotate narrative samples from each population using the same, abstracting over the substantial grammatical differences between languages.

Narrative Schema Induction. There has recently been a growing interest in computational methods for extracting narrative schemas that capture the causal and temporal relationships between events (Manshadi et al. 2008; Chambers & Jurafsky, 2009; Hu et al., 2013). However, to our knowledge, all of these approaches treat every verb in the narrative as an event. Our annotation analysis indicates that less than half of the clauses in a personal narrative are actions that are causally and temporally related to the experience being described. Our simple annotation schemes provide a method for separating these clauses. This will help eliminate extraneous verbs (phrase, clauses, etc.) from being included, reducing the noise and improving the overall accuracy of the learned schemas. Separating these clauses also highlights the fact that narratives are not only descriptions of causally related events, but also contain emotional reactions and embedded analysis of the events that occur. To our knowledge no automated schema induction system incorporates this knowledge into its representation. However, we believe these elements are fundamental to the essence of a narrative and what makes them such a powerful method of discourse.

Case-based Interactive Narrative. One of Rahimtoroghi et al.'s motivations in applying Labov & Waletzky's theory to narrative text was to advance technologies for case-based interactive narrative. In their research, they investigated the structure of the text-based interactive narratives generated by the Say Anything system (Swanson & Gordon, 2012), where the computer advances an unfolding fictional narrative by selecting the next sentence from a corpus of millions of non-fiction narratives drawn from public weblogs. Rahimtoroghi et al. found that Say Anything narratives over-represented *evaluation* sentences, compared to *orientation* and *action* sentences. By explicitly labeling sentences in the case repository by their structural function in a narrative, an improved version of Say Anything could tailor its selection of sentences to produce structural distributions that are judged higher by users.

We believe that the *diegetic* / *extradiegetic* distinction can also be exploited to improve the sentence-selection algorithm of Say Anything, e.g. by favoring *diegetic* sentences that describe the world of the story events over *extradiegetic* commentary. By focusing more exclusively on story-world events, the resulting narratives downplay the perspectives of the multitude of disparate narrators who contribute to an emerging narrative, which may be impossible to combine into a coherent discourse.

Commonsense Reasoning. Roemmele et al. (2011) presented a novel evaluation for automated commonsense reasoning, the Choice Of Plausible Alternatives (COPA), consisting of a thousand binary choice causal reasoning questions written as English sentences. Leading systems (Goodwin et al., 2012; Gordon et al., 2011) each succeed by computing average pointwise mutual information between question and answer words, calculated from extremely large text corpora. Gordon et al., in particularly, achieved the highest score on this evaluation by using a large corpus of narrative drawn from public weblogs. The implication is that using a corpus that best overlaps in content with the COPA questions achieves the best results.

In analyzing COPA questions, however, we see that not all structural classes are represented. COPA questions are overwhelmingly *diegetic* and *action* sentences, with a good mix of both *subjective* and *objective* content. Accordingly, we expect that significant improvements in COPA scores could be achieved by computing pointwise mutual information values using only the *diegetic action* portions of large narrative corpora, or at least weighting this content higher than *extradiegetic orientation* and *evaluation* text.

Sentiment Analysis. Much of the research on sentiment analysis over the last decade has been driven by the interest in online product reviews, including movies, restaurants, etc. (Pang & Lee, 2004; Blitzer et al., 2007). As sentiment-expressing text is often paired with quantitative judgments (e.g. star ratings), customer reviews scraped from commercial websites are often used as a convenient source of training and test data in this line of research.

Subjective content is certainly the central focus of sentiment analysis, but the inclusion of other labels across the two high-level annotation schemes may allow for a more nuanced approach to the analysis task. For example, one might expect that *extradiegetic evaluation* sentences are most indicative of a reviewer's opinion of a product, e.g. "*I can't think of anything redeemable about this movie.*" Likewise, recognizing diegetic text may help systems to discount positive or negative subjective text that describes only the context of a user's experience, e.g. "*I sprayed it on the nasty grime at the bottom of the tank.*"

Narrative Retrieval. In parallel with the development of large corpora of narrative text has come new information retrieval methods that are specifically tuned to genre of discourse. For example, several researchers have focused on the task of activity-based retrieval: finding examples from a collection where the narrated events are part of a desired activity context, e.g. car crashes and heart attacks (Campbell et al., 2012; Wienberg & Gordon, 2012). A representative example approach is seen in Gordon et al. (2012), where relevance feedback is used to learn a topic model for an activity consisting of weighed term lists, e.g. narratives of people having a stroke exhibit words such as *stroke*, *triage*, *speech*, *ambulance*, and *emergency* with certain frequencies. Highly-weighted terms are those that participated in the expected *script* of the activity, in the Schank and Abelson (1977) sense of word, and would be among the same as those included in other approaches to script specification, e.g. Li et al. (2013), Jung et al. (2010).

These terms describe events that occur at the *diegetic* level of discourse, i.e. in the sequence of events of the story world that the narrator is describing. Accordingly, we believe that activity topic models could be more efficiently learned (higher retrieval accuracy, fewer training examples) by considering only the *diegetic* narrative clauses in the source narratives.

Conclusions

The increased use of large corpora in narrative research has created new opportunities for empirical research and intelligent narrative technologies. To best exploit the value of these corpora, some researchers are eschewing complex discourse analysis techniques in favor of high-level narrative annotation schemes that can be quickly applied, achieve high inter-rater agreement, and are amenable to automation using machine-learning techniques. In this paper we explore the similarities between different schemes used by two different research groups with different analysis goals. Through a dual-annotation study, we have found that the orientation/action/evaluation scheme employed by Rahimtoroghi et al. (2013) is orthogonal to the subjectivity and narrative levels analysis seen in Sagae et al. (2013). Each scheme captures a structurally distinct characteristic of real-world narratives, and each combination of labels is evident in a corpus of 19 weblog narratives (951 narrative clauses).

We expect that these types of schemes will see increased use in future research, where applications will drive the selection of terms that are used. We reviewed six applications where various combinations of terms could bring improvements to state-of-the-art systems: comparative analyses, narrative schema induction, case-based interactive narrative, commonsense reasoning, sentiment analysis, and narrative retrieval. The broad applicability of these terms across tasks justifies future work along these lines, particularly where it leads to the dissemination of annotated corpora and automated annotation tools that can be more broadly used by the narrative research community.

Acknowledgments

This research was sponsored by the NSF grant number 1002921, NSF grant number 1238553, and Defense Advanced Research Projects Agency, grant number D12AP00069. The content of this publication does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, Bollywood, boom boxes and blenders: Domain adaptation for sentiment classification. In Proceedings of the Association for Computational Linguistics, Prague, Czech Republic.

Campbell, A., Wienberg, C., and Gordon, A. (2012) Collecting Relevance Feedback on Titles and Photographs in Weblog Posts. Proceedings of the 2012 International Conference on Intelligent User Interfaces, Lisbon, Portugal.

Chambers, N. and Jurafsky, D. (2009) Unsupervised learning of narrative event chains. Proceedings of the Joint Conference of the ACL and the International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore.

Elson, D. K. (2012a) DramaBank: Annotating Agency in Narrative Discourse. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Elson, D. K. (2012b) Modeling Narrative Discourse. Ph.D. Thesis, Columbia University, New York, NY.

Finlayson, M. A. (2011) The Story Workbench: An Extensible Semi-Automatic Text Annotation Tool, Intelligent Narrative Technologies Workshop, Stanford, CA.

Finlayson, M. A. (2012) Learning Narrative Structure from Annotated Folktales, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Genette, G. (1980). Narrative Discourse: An Essay in Method. Cornell University Press, 1980.

Goodwin, T., Rink, B., Roberts, K., and Harabagiu, S. (2012) UTDHLT: COPACETIC System for Choosing Plausible Alternatives. Proceedings of the International Workshop on Semantic Evaluation, Montreal, Canada.

Gordon, A. and Swanson, R. (2009) Identifying Personal Stories in Millions of Weblog Entries. International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA.

Gordon, A., Bejan, C., and Sagae, K. (2011) Commonsense Causal Reasoning Using Millions of Personal Stories. Conference on Artificial Intelligenc, San Francisco, CA.

Gordon, A., Huangfu, L., Sagae, K., Mao, W., and Chen, W. (2013) Identifying Personal Narratives in Chinese Weblog Posts. Intelligent Narrative Technologies Workshop, Boston, MA.

Gordon, A., Wienberg, C., and Sood, S. (2012) Different Strokes of Different Folks: Searching for Health Narratives in Weblogs. International Conference on Social Computing, Amsterdam, The Netherlands.

Hu, Z., Rahimtoroghi, E., Munishkina L., Swanson R., and Walker, M. A. (2013) Unsupervised Induction of Contingent Event Pairs from Film Scenes. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Seattle, Washington.

Jung, Y., Ryu, J., Kim, K., & Myaeng, S. (2010): Automatic Construction of a Large-Scale Situation Ontology by Mining How-to Instructions from the Web, Web Semantics: Science, Services and Agents on the World Wide Web, vol. 8, no 2-3, pp. 110-124.

Labov, W. and Waletzky, J. (1997) Narrative analysis: Oral versions of personal experience. Journal of Narrative and Life History 3–38.

Li, B., Lee-Urban, S., Johnston, G., and Riedl, M. (2013). Story Generation with Crowdsourced Plot Graphs. Proceedings of the AAAI Conference on Artificial Intelligence, Bellevue, Washington.

Louis, A., Joshi, A., Prasad, R. and Nenkova, A. (2010). Using entity features to classify implicit discourse relations. Proceedings of the Annual SIGDIAL meeting on Discourse and Dialogue. Tokyo, Japan.

Mann, W. C. & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8 (3). 243-281.

Manshadi, M., Swanson, R., and Gordon, A. S. (2008) Learning a probabilistic model of event sequences from internet weblog stories. In Proceedings of the FLAIRS Conference. Coconut Grove, Florida.

Montfort N. (2011) Curveship's automatic narrative style. Proceedings of the International Conference on Foundations of Digital Games, New York, NY.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics, pages 271–278, Stroudsburg, PA.

Pitler, E., Louis, A. and Nenkova, A. (2009) Automatic sense prediction for implicit discourse relations in text. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP). Suntec, Singapore.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. and Webber, B. (2008) The Penn Discourse Treebank 2.0. In Proceedings of the International Conference on Language Resources and Evaluation. Marrakech, Morocco.

Rahimtoroghi, E., Swanson, R., Walker, M. A., and Corcoran, T., (2013) Evaluation, Orientation, and Action in Interactive StoryTelling. Proceedings of the Intelligent Narrative Technologies Workshop, Boston, MA.

Rishes, E., Lukin, S. M., Elson D. K., and Walker M. A. (2013) Generating Different Story Tellings from Semantic Representations of Narrative, in International Conference on Interactive Ditigal Storytelling. Istanbul, Turkey.

Roemmele, M., Bejan, C., and Gordon, A. (2011) Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University.

Sagae, K. (2009). Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In Proceedings of the International Conference on Parsing Technologies, pages 81--84.

Sagae, K., Gordon, A., Dehghani, M., Metke, M., Kim, J., Gimbel, S., Tipper, C., Kaplan, J., and Immordino-Yang, M. (2013) A Data-Driven Approach for Classification of Subjectivity in Personal Narrative. Workshop on Computational Models of Narrative, Hamburg, Germany.

Schank, R. C., and Abelson, R. (1977) Scripts Plans Goals. Lawrence Erlbaum Associates.

Soricut, R. and Marcu, D. (2003) Sentence Level Discourse Parsing using Syntactic and Lexical Information. Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), May 27-June 1, Edmonton, Canada.

Swanson, R. and Gordon, A. (2012) Say Anything: Using Textual Case-Based Reasoning to Enable Open-Domain Interactive Storytelling. ACM Transactions on Intelligent Interactive Systems 2(3).

Swanson, R., Rahimtoroghi, E., Corcoran, T., and Walker, M. A. (2014) Identifying Narrative Clause Types in Personal Stories. To appear in Proceedings of the SIGDIAL 2014 Conference, Philadelphia, PA.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., and Martin, M. (2004) Learning subjective language. Computational Linguistics, 30(3):277–308.

Wienberg, C. and Gordon, A. (2012) PhotoFall: Discovering Weblog Stories Through Photographs. Proceedings of the ACM Conference on Information and Knowledge Management, Maui, HI.