Annotation of Children's Oral Narrations: Modeling Emergent Narrative Skills for Computational Applications

Rebecca J. Passonneau and **Adam Goodkind**
Columbia University
(becky@cs.|adam@)columbia.edu

Elena T. Levy
University of Connecticut
elena.levy@uconn.edu

Abstract

We present an annotation method for developing a model of children's comprehension that differentiates between their recall for the objective content of a story and inferred content. We apply the annotation method to a corpus of retellings, in which children retell the same story on three successive days. Our results indicate differences over time: on Day three, children have a more evenly distributed recall of events throughout the story, and include significantly more inferences. The results suggest a cognitive bootstrapping effect. We discuss the potential for application to diagnostic assessment of children's narrative skills and tutorial applications.

Introduction

Our goal is to develop an annotation method for corpora of children's narrations that can support educational or diagnostic applications. We show how the annotation we propose provides a model of children's comprehension that differentiates between their recall for the objective content of a story and inferred content. We apply the annotation method to a corpus of retellings, in which children retell the same story on three successive days. Our results indicate differences over time: on Day three, children have a more evenly distributed recall of events throughout the story, and include significantly more inferences.

We describe our corpus in the next section, our content annotation method in the following section, and the application of the annotation method to seven of the ten retellings in our corpus in the fourth section. In the fifth section, we present the results of a pilot test of applying the model to the three remaining retellings; while extremely provisional due to the small sample size, it illustrates the direction we aim to pursue. We follow with a discussion section on how our findings buttress much earlier results from the educational literature pertaining to assessing and improving older children's comprehension of written material. Based on the connection we draw with this earlier work, we discuss the prospect for developing applied NLP approaches for assessment of children's narrative skills, or tutorial applications. In the conclusion, we describe our next steps, and point very briefly to the many questions our results raise, with particular reference to the interdependence between linguistic, narrative and cognitive competence (Levy 2003; Levy & Fowler 2005).

Corpus of Narrative Retellings

The corpus of narrative retellings was collected by Elena Levy for use in research on cognitive development. (Levy 2003) presents a theory of cognitive bootstrapping, in which the process of telling a narrative serves as a *scaffold* for a more coherent perspective on the described events. The methodology for collecting the retellings derives from (Chafe 1980) and (Bartlett 1932). As in (Chafe 1980), the narrative stimulus for creating the corpus is a silent movie. Each subject who retells the narrative chooses what to say and how to say it. As in (Bartlett 1932), subjects retell a story on subsequent days, without re-viewing the movie, to provide a means to observe *constructive* aspects of recall.

Participants were ten children between the ages of five and seven, evenly balanced by gender, from public schools in Connecticut who were told that they were helping Dr. Levy and her students in their research. They were shown an abridged version of Lamorisse's film, *The Red Balloon*. It was shortened primarily to make the data collection more convenient for the participating schools. Levy had already used the full film for similar research on other age groups, and the abridgement preserved the main storyline. Researchers filled the listener role, and provided backchannel utterances. The children were told that the researchers did not know the story. After viewing the movie on Day 1, subjects met one-on-one with a researcher to retell the story in their own words. After the first retelling, the same subjects were asked to retell the story on two successive days, without re-viewing the film. There was a different listener for each of the three retellings. The narrations were recorded and transcribed using conventions described in (Levy 2003).

We randomly selected seven retellings for constructing a comprehension model for each day that we refer to as a pyramid (model or training set). Table 1 indicates the average number of utterances and words in each retelling for the model set versus the test set, and overall. Within each set, lengths of retellings vary widely, presumably following a normal curve.

	Day 1		Day 2		Day 3		Avg.	
Spkr	utt	words	utt	words	utt	words	utt	words
model avg	33	255.1	25.7	206.9	26.9	210.3	28.5	224.1
test avg	34.7	259.3	57.0	446.3	53.3	410.0	48.3	371.9
Overall Avg	33.5	256.4	35.1	278.7	34.8	270.2	34.5	268.4

Table 1: Lengths of narrations in utterances and words

Annotation of Narrative Content Units

We designed an annotation method to address three criteria: 1) to facilitate quantitative and qualitative assessment of the content that individuals choose to express in a narrative; 2) to allow comparison of the individuals' narrations with the objective event line of a story; and 3) to distinguish elements of the story that can be objectively identified in the source movie from story elements that the speaker has inferred. To address the first criterion, we use the pyramid method (Nenkova & Passonneau 2004), then we extend the pyramid representation to capture the distinction between the *referential* and *evaluative* dimensions of narrative (Labov & Waletzky 1967).

Pyramid annotation

Pyramid content annotation capitalizes on an observation seen in human summaries of news sources: summaries from different humans always have partly overlapping content. It is difficult to quantify the overlap because the same content is typically expressed in different ways. Pyramid annotation is a manual procedure to identify Summary Content Units (SCUs) by abstracting over a set of model summaries. As with factoid annotation (Teufel & van Halteren 2004), SCU annotation results in *emergent* semantic units: the discovered units depend on the content expressed in a sample of model summaries. Each SCU has a weight to indicate the number of models that express the represented content.

Figure 1 illustrates the SCU-like elements of a Narrative Content Unit (NCU) created from Day 3 retellings: a label, a weight (W), and contributors consisting of utterances, utterance fragments, or utterance sequences from narrations that express similar content. This NCU has contributors from four retellings. An NCU contributor can be smaller or larger than an utterance; three of the contributors in Figure 1 have two utterances each.

A major difference between the summary data for which the SCU annotation method was developed and these narrations is that a target length was imposed on the summaries. Length largely defines what a summary is, in contrast to a précis or report. In contrast, a story is an independent semantic object, and in principle can have any length. Thus one of the additional representational elements of NCUs not present in SCUs is a positional index.

We number utterances sequentially, but in addition, each utterance in a contributor has a positional index indicating how far through the narration it occurs. Where N is the number of utterances in a narrative and N_i is the sequential index of a given utterance, we define the positional index for an utterance to be $\mathrm{N}_i \times \frac{999}{N}$. For a narrative of twelve utterances, the positional indices would be 83, 166, 249, ..., 830, 913, 999. The first utterance constitutes approximately 8.3% of the narrative, the first two constitute 16.6%, and so on. For narrators 4, 6 and 7 in Figure 1, the corresponding contributor occurs relatively late in their narrations, whereas for narrator 3 it spans roughly the midpoint (555) and the two-thirds point (666). The narration narration from speaker 3 is much shorter (9 utterances compared with 44, 41 and 46).

All narrative utterances go into at least one NCU, and the NCUs for a given day constitute a pyramid. The weights form a partition over a pyramid. There tend to be fewer NCUs that have the maximum weight, and increasingly many NCUs at each lower weight, with the most NCUs at weight=1. Table 2 shows the number of NCUs of each weight (or each tier) for the three pyramids. Notice the Zipfian distribution: there is a power law increase in the number of SCUs per tier as the weight decreases. It is because of this *bottom-heavy* distribution that we refer to the content

Label (W=4): Kid steps on balloon

narr	position	contributor
3	(555)	and um it it just gets popped
	(666)	because somebody steps on it
4	(864)	then like for ten minutes later,
		um a boy came
	(886)	and stepped on it
6	(877)	and then there was this boy
	(901)	who popped the balloon with his foot
7	(912)	and then this kid stepped 'n the balloon

Figure 1: An example NCU of weight 4 from the Day 3 pyramid

NCU	Number of NCUs per tier		
weight	Day 1	Day 2	Day 3
7	5	0	0
6	0	2	2
5	3	4	1
4	2	1	7
3	10	6	9
2	21	18	15
1	89	79	62
Total	130	110	96
avg wt	1.68	1.54	1.71

Table 2: Cardinality of NCUs at each weight

models as pyramids. On Day 1, there were five NCUs that appeared in all seven narrations, so we predict that these are more likely than NCUs of lower weights to recur in a new Day 1 retelling. While the set of NCUs of weight 1 has the highest cardinality, each NCU in this set represents content that only one out of seven children expressed; we hypothesize that NCUs of weight 1 have a low probability of being re-expressed.

Two advantages to pyramid annotation that carry over to analysis of narrations is that the annotation method has been found to be reliable and reproducible with naive annotators (see subsection *Reliability of annotation method*), and that a pyramid serves as a predictive model of the content of new narrations produced under the same conditions (e.g., same input story, same day's retelling). A summarization pyramid models the hypothesis that new summaries are more likely to contain information from the higher weighted *tiers*. By assigning different weights to different SCUs (or NCUs), a pyramid accounts for the observation that not all content is equally relevant to a news topic (or narrative), and it also accounts for the fact that summaries with different information content can be qualitatively equivalent.

A test summary is annotated against a pyramid to determine which SCUs in the pyramid are expressed in the new summary, and what the weights of the re-expressed SCUs are. The sum of the weights of a test summary are used to rate its content on a scale from zero to 1 by normalizing the sum in one of a number of ways (see (Passonneau *et al.* 2005)). For rating novel narrations, we will normalize the summed weights of the NCUs in a new narration that re-express NCUs from a corresponding pyramid using a ratio of the observed sum of NCU weights in a new narration to the maximum sum for the average number of NCUs per model narration in each pyramid. For example, the Day 1 average number of NCUs per narration is 31. The maximum sum that can be assigned to 31 NCUs from the Day 1 pyramid is given by taking 31 NCUs from the topmost tiers and computing the sum of their weights: $(7 \times 5) + (5 \times 3) + (4 \times 2) + (3 \times 10) + (2 \times 11) = 110$.

Recalled NCUs

In contrast to SCUs, NCUs can be aligned with an independently arrived at set of semantic elements. As part of a separate project, Levy, her students, and Passonneau developed guidelines for creating a scene structure for *The Red Balloon*.

We defined a scene in terms of the physical location and time of a story element. All observable events taking place within the same continuous time period and at the same location or along the same continuous physical path went into the same scene. Scenes were numbered sequentially. Within a scene, distinct events were listed sequentially. All scenes and scene events (SEvents) were given ordinary language labels. Two annotators worked independently to create the scene structure, and differences were adjudicated. Differences had mainly to do with granularity, not with actual content. Figure 2 illustrates scene 1.

Inferences and psychological motivations are not part of the scene structure. For example, in SEvent 1.1 where the boy looks up the lamppost, the balloon is not visible on screen, so the SEvent label for 1.3 does not mention that the boy climbs up the lamppost because he sees a balloon and wants to retrieve it, although these are inferences that many observers would make retroactively once the balloon comes into view, and the boy takes it from the lamppost.

After annotating NCUs, we align them with scene events (SEvents). The NCU in Figure 1 aligns with SEvent 17.1 "Kid steps on the balloon and pops it."

Inferred NCUs

To capture the *evaluative* dimension (Labov & Waletzky 1967) of the narrations, we created five categories during our pilot study of narrations produced by older children; in addition we use a sixth *none of the above* category. Along with each definition, we give one or more examples from the narrations. Although we created the categories independently, they have a fairly direct correspondence with those created by (Tannen 1980) for the Pear stories and by (Donaldson 1986) for children's explanations; we believe this coincidence of categorization schemes supports their validity.

1. **Thematic (T):** utterance refers to an overarching theme (e.g. *the movie is about a boy*), or describes a thematic pattern *the balloon follows the boy everywhere*
2. **Psychological Inference (P):** utterance that expresses an inference about the psychological state of a character (e.g. *his nanny didn't like having the balloon in the house*)
3. **Interpolated Event (I):** utterance about an objective action that was not represented in the movie, but where commonsense knowledge supports the inference that the action occurred (e.g., *he went to bed, he woke up* at transition to a new day in the film)
4. **Causation (C):** utterance that explains why an action happened, whether it be a concrete action or an interpersonal one, such as teasing (e.g., *the balloon followed the principal to get the key to unlock the door*
5. **Desire (D):** inferring a character's positive or negative desires (e.g., *because they wanted to touch his balloon; he didn't want him to go on (the bus)*)

Reliability of annotation method

A pyramid of a set of narrations is directly analogous to a pyramid of a set of summaries. The annotation reliability of pyramid annotation has been tested in several ways so we do not report a new reliability assessment of narrative NCUs. Interannotator agreement on pyramid creation for five pyramids comprised of seven summaries ranged from 0.68 to 0.81 (Passonneau 2006; In submission), using Krippendorff's α (Krippendorff 1980) combined with a set-based distance metric that assigns partial credit when annotator's choices overlap. A more grounded method of assessing annotation validity is to measure the differences in results when a different annotation is substituted in the context of an application of the annotation (Passonneau 2006), such as the use of a summarization pyramid to score the content quality of unseen summaries (Nenkova, Passonneau, & McKeown To Appear). A test of each of the five pairs of pyramids

Scene id	Scene name	SEvent id	SEvent name
1	street scene moving towards and reaching descending steps	1	boy carrying school satchel walks up street towards stairs
		2	boy stops to pet cat
		3	boy goes down steps
2	at bottom of steps	1	boy looks up lamppost
		2	boy puts satchel down behind lamppost
		3	boy climbs up lamppost

Figure 2: The first six SEvents in the scene structure of *The Red Balloon*

on the impact of average scores of sixteen summarization systems over eight document sets was found to be negligible. Using analysis of variance of average system scores per document set, combined with Tukey's Honest Significant Difference method to identify significant differences among systems, only 1.7% of the $\binom{16}{n} = 120$ system comparisons differed (Passonneau In submission) when a different annotator's pyramid was used.

For the retellings, we measured interannotator reliability on the task of matching NCUs to the independently created event list, or on recalled versus inferred NCUs (RNCUs versus INCUs), and on the task of categorizing INCUs, using a set of thirty one base NCUs. We presented an independent annotator who was not previously familiar with the narrations, the content of the film, or the annotation method with a list of NCUs, the event list, and a description of the six types of inferred events. The annotator's task was to identify the subset of NCUs that matched items in the event list, then to sort the remaining NCUs into the six categories of INCUs. Note that the fact that the new annotator had not seen the movie handicaps this annotator to some degree, because the event list is essentially a short hand for the actual events in the movie. Using both Krippendorff's α and Cohen's κ, interannotator agreement was 0.81 on the RNCU/INCU distinction, and 0.75 on the classification of INCUs.

Day-by-Day Model of Childrens' Narrations

Comparison of referential content across days

Figures 3 show which SEvents are mentioned by how many children across the three days. The x-axis of each chart is the linear sequence of SEvent ids, and the y-axis is the number of children who mentioned the corresponding SEvent. A bar of height y at x indicates that y children mentioned SEvent x. Superimposed on each chart is a trend line. All days have a parabolic distribution, with the highest points at the beginning and end of the narrative. Note, however, that the curve flattens on Day 3.

On Day 1, eight SEvents are mentioned by most of the children; this decreases to seven on Days 2 and 3. The identity of the frequently mentioned SEvents changes from day to day. The proportion of frequently mentioned SEvents in the middle of the narrations goes down from 25% (2/8) on Day 1 to zero on Day 2; it increases to 43% on Day 3.

We can test for the significance of the difference in distributions across days by comparing the frequency means for each pair of days. As the total number of SEvents is 103, we use a method for comparing the means of small samples where we compute the t-statistic for each comparison of means. For all three comparisons, we can reject the null hypothesis that there is no difference (p=0).[1]

Recalled versus inferred content across days

Table 3 shows the distribution of recalled versus inferred contributors across the three days, meaning contributors to

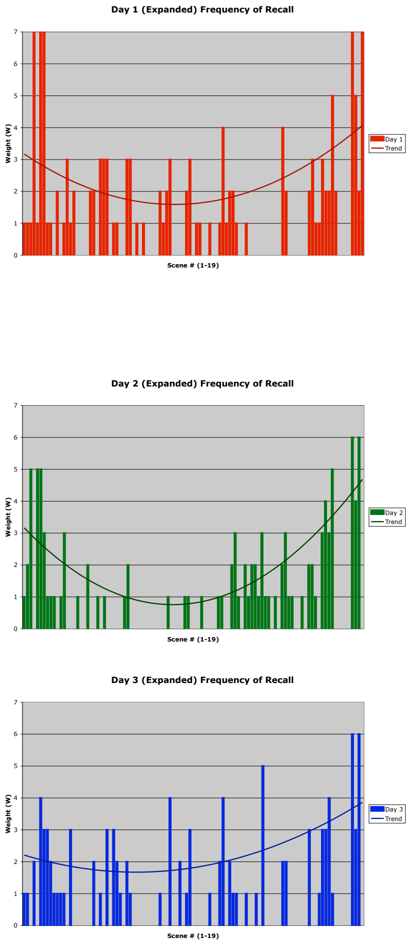

Figure 3: Frequency Distribution of SEvents on Days 1-3

[1] Where w is the weight (frequency), we used $\frac{1}{w}$ to transform the parabolic distribution to a normal one for the means test.

Contributor type	day 1	day 2	day 3	Row totals
Recalled	189	137	125	451
Inferred	30	28	42	100
Col Total	219	165	167	551

Table 3: Recalled versus inferred contributors

narrator	day 1	day 2	day 3
A	0.40	0.50	0.54
B	0.81	1.07	1.05
C	0.62	0.99	0.99

Table 4: Content scores for three reserved re-tellings

RNCUs versus INCUs. (Because contributors can be greater or smaller than an utterance, column totals do not correspond to the utterance counts in Table 1.) Each day, there is a decrease in the number of contributors to RNCUs, and a large increase in the proportion of contributors to INCUs. A chi-squared test indicates a statistically significant distribution (p=0.0137; df=2).

Applying the Model to the Test Set

One of the functions of a pyramid content model has been for use in scoring automated summarizers based on average performance on large numbers of news clusters. The score for a single summary is based on summing the weights of the content units observed in the test summary, then normalizing the sum. The normalization method that we use here is the one referred to as the *modified* pyramid score in (Passonneau *et al.* 2005). It is a ratio of the observed sum of weights of the NCUs expressed in the retelling to the maximum sum that could be achieved using the average number of NCUs (N) per model retelling in a pyramid: $\frac{Sum_{Obs}}{Sum_{Max}}$. Sum_{Max} for a pyramid with N tiers is defined as the sum given by taking a sample of N NCUs from the pyramid, without replacement, such that there is no NCU left in the pyramid that is of higher weight than any of the N NCUs in the sample, and summing the N weights. If the number of NCUs in a test retelling is never greater than the average number of NCUs per model in a pyramid, then the modified pyramid score ranges from 0 to 1. The higher the score, the closer is the content to an ideal retelling predicted by the pyramids.

Table 4 gives the modified pyramid scores for the test set. In the original application of pyramid scoring, the scored summaries and model summaries were always the same length. Here, we have novel retellings that can be longer than the pyramid average, thus the scores for B on Day 2 and Day 3 are somewhat higher than 1. The trend in Table 4 is for the score on Day 1 to be lower than the other two days.

Discussion

We hypothesize a connection between children's ability to retell a story and their reading comprehension skills, based on studies that address the oral narratives of poor readers when children are presented with non-written materials to describe. Two of the more recent studies used Labov and Waletzky's (1967) *high point* analysis: Norris and Bruning (1988) found that when kindergarten and first grade children were asked to describe a series of photographs, low achieving readers produced fewer propositions that maintained the story's theme, and also produced a greater number that were tangential to the theme, as well as comments that did not refer to the story at all. Recently, Celinska (2004) found that when the personal experience narratives of learning disabled fourth graders were compared to nondisabled children, learning disabled girls produced fewer narratives with high points.

Given a connection between oral narrative skills and reading ability, we see a potential for both diagnostic and tutorial applications of content-based analysis of children's narrations. There has been scattered evidence from prior work that children's comprehension of written material can be improved by revising their readings, thus a systematic understanding of the revisions that are most helpful could be useful in developing tutorial software. A 1980 study of adaptive software that adjusted to children's paragraph-by-paragraph reading of on-line essays provided evidence that text revision can improve reading comprehension in eleventh graders (L'Allier 1980). It used manually rewritten passages that simplified the syntax and vocabulary. (Beck *et al.* 1991) presented evidence that revision improves comprehension for fifth graders on history texts. In contrast to (L'Allier 1980), (Beck *et al.* 1991) attempted to clarify the discourse structure by making causal relations explicit, and by ensuring that the reader would be given any necessary background knowledge to understand the causal relations. (L'Allier 1980) relied on standardized reading comprehension questions to measure children's comprehension. (Beck *et al.* 1991) applied a knowledge- and labor-intensive, manual content annotation method (Omanson 1982), and provided little information on interannotator agreement.

One obstacle to applying revision more widely is that these earlier methods for analyzing and revising texts were difficult to reproduce. The annotation method we present here is derived from a summarization evalutation method that has been used for the past two Document Understanding Conferences, and that has been shown to have high interannotator reliability. An automatic summarizer based on the ideas encapsulated in the pyramid method was presented in (Nenkova 2006), and the same techniques could be adapted for revision.

(Halpin, Moore, & Robertson 2004) applied information extraction techniques to identify features for an automated classifier applied to childrens' written stories. Children read a story and rewrote it in their own words, then teachers rated the stories on a 4-point scale. The classifier was able to identify stories that reproduced the sequence of events. In the authors' view, the features they used did not distinguish narratives demonstrating an understanding of the *point* of the story, above and beyond the event sequence. Their conclusion is reminiscent of Labov and Waletzky's (1967) distinction between *referential* and *evaluative* content.

Conclusions and Future Work

We have found that children's narrations on the last day of three retellings evidence more evenness in recall of events from the beginning, middle and end of the story, in contrast to day one or two retellings. Previous work on children's and adults' recall of stories using story grammar methods posited a *story category effect* in which recall is best for the setting, and for the beginning and consequence of the protagonist's attempt to reach a goal, while recall is worse for motivating states and for reactions (McCabe, Capron, & Peterson 1991). This resembles the graphs for the day one and two retellings in Figure 3. What we may be finding is that with repeated retellings, children are continuing to extract more meaning from the sequence of events as they retell the story. This accords with Levy's (2003) theory of *scaffolding*, which posits a conceptual bootstrapping effect similar to, but less conscious than, revision of written material. Her findings indicate that the same individual re-uses lexical and phrasal material from earlier retellings, but that via a process Levy refers to as *narrative compression*, the subsequent re-statements involve fewer clauses, combined with greater syntactic complexity per clause.

In future work, we plan to investigate whether the same pattern of increased recall for the middle of a story and more frequent inferences on day three retellings holds for other stories and other age groups. We are currently applying the same methodology presented here to a corpus of retellings of Buster Keaton's *Sherlock Jr.* produced by ten year olds.

References

Bartlett, F. C. 1932. *Remembering: A Study in Experimental and Social Psychology*. Oxford: Macmillan.

Beck, I.; McKeown, M.; Sinatra, G.; and Loxterman, J. 1991. Revising social studies text from a text-processing perspective: evidence of improved comprehensibility. *Reading Research Quarterly* 251–276.

Celinska, D. K. 2004. Personal narratives of students with and without learning disabilities. *Learning Disabilities Research and Practice* 83–98.

Chafe, W., ed. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, New Jersey: ABLEX Publishing Corporation.

Donaldson, M. 1986. *Children's Explanations: A psycholinguistic study*. New York: Cambridge University Press.

Halpin, H.; Moore, J. D.; and Robertson, J. 2004. Automatic analysis of plot for story rewriting. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Krippendorff, K. 1980. *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Labov, W., and Waletzky, J. 1967. Narrative analysis. In Helm, J., ed., *Essays on the Verbal and Visual Arts*. University of Washington Press. 12–44.

L'Allier, J. J. 1980. *An evaluation study of a computer-based lesson that adjusts reading level by monitoring on task reader characteristics*. Ph.D. Dissertation, University of Minnesota.

Levy, E. T., and Fowler, C. A. 2005. How autistic children may use narrative discourse to scaffold coherent interpretations of events: A case study. *Imagination, Cognition and Personality* 207–244.

Levy, E. 2003. The roots of coherence in discourse. *Human Development* 169–188.

McCabe, A.; Capron, E.; and Peterson, P. 1991. The voice of experience: The recall of early childhood and adolescent memories by young adults. In McCabe, A., and Peterson, C., eds., *Developing Narrative Structure*. Hillsdale, NJ: Erlbaum. 137–174.

Nenkova, A., and Passonneau, R. J. 2004. Evaluating content selection in summarization: the pyramid method. In *Proceedings of HLT/NAACL*.

Nenkova, A.; Passonneau, R. J.; and McKeown, K. To Appear. The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*.

Nenkova, A. 2006. *Understanding the process of multi-document summarization: content selection, rewrite and evaluation*. Ph.D. Dissertation, Department of Computer Science, Columbia University.

Norris, A., and Bruning, R. H. 1988. Cohesion in the narratives of good and poor readers. *Journal of Speech and Hearing Disorders* 416–424.

Omanson, R. C. 1982. An analysis of narratives: Identifying central, supportive and distracting content. *Discourse Processes* 119–224.

Passonneau, R.; Nenkova, A.; McKeown, K.; and Sigelman, S. 2005. Applying the pyramid method in 2005. In *Proceedings of the 2005 Workshop of the Document Understanding Conference (DUC)*.

Passonneau, R. J. 2006. Measuring agreement for set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Passonneau, R. J. In submission. Formal and functional assessment of the pyramid method.

Tannen, D. 1980. A comparative analysis of oral narrative strategies. In Chafe, W., ed., *The Pear Stories*. Norwood, New Jersey: ABLEX Publishing Corporation. 9–50.

Teufel, S., and van Halteren, H. 2004. Evaluating information content by factoid analysis: human annotation and stability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.