Annotating Story Timelines as Temporal Dependency Structures

Steven Bethard[1], **Oleksandr Kolomiyets**[2], **Marie-Francine Moens**[2]

[1]University of Colorado Boulder
Campus Box 594, Boulder, Colorado 80309, USA
Steven.Bethard@colorado.edu

[2]Katholieke Universiteit Leuven
Celestijnenlaan 200A, B-3001 Heverlee, Belgium
{Oleksandr.Kolomiyets,Sien.Moens}@cs.kuleuven.be

Abstract

We present an approach to annotating timelines in stories where events are linked together by temporal relations into a *temporal dependency tree*. This approach avoids the disconnected timeline problems of prior work, and results in timelines that are more suitable for temporal reasoning. We show that annotating timelines as temporal dependency trees is possible with high levels of inter-annotator agreement – Krippendorff's Alpha of 0.822 on selecting event pairs, and of 0.700 on selecting temporal relation labels – even with the moderately sized relation set of BEFORE, AFTER, INCLUDES, IS-INCLUDED, IDENTITY and OVERLAP. We also compare several annotation schemes for identifying story events, and show that higher inter-annotator agreement can be reached by focusing on only the events that are essential to forming the timeline, skipping words in negated contexts, modal contexts and quoted speech.

Keywords: timelines; dependency trees; annotation

1. Introduction

Understanding the order of events along the timeline of a story is crucial for making sense of the text. Educators evaluate such understanding with reading comprehension questions like "What did Jack do before he looked out of his window?" or "Did Jack look out of his window before or after he buckled his seat belt?" To generate such questions automatically – e.g. for interactive tutoring systems – stories must be annotated with their event timelines.

Thus an important issue in linguistic annotation is establishing reliable methods for annotating the order of events in the plot of a story. The TimeML annotation scheme (Pustejovsky et al., 2003a) provides a starting point for such work, describing what kinds of words and phrases should be annotated as events and times, and describing how temporal relations between events and times can be annotated when they are found. However, TimeML gives no specific guidelines as to when or where a temporal relation should be annotated. Sometimes only temporal relations that were prominent or had explicit cue words were annotated, as in TimeBank (Pustejovsky et al., 2003b), while other times only specific syntactic constructions were annotated, such as matrix verbs in adjacent sentences as in TempEval 2007 (Verhagen et al., 2007), or verbs and events in subordinate clauses as in Bethard et al. (2007) or as in TempEval 2010 (Verhagen et al., 2010). The result of such approaches is often a disconnected timeline - some events are ordered with respect to each other, but many events are not.

In this article, we present an approach to annotating the order of the events in a children's story that guarantees that all events in a plot are connected along the timeline, and thus provides the temporal structure of the plot. We treat event ordering annotation as a type of dependency structure annotation, asking annotators to link events in a chain or tree, labeling each link with a temporal relation such as BEFORE or AFTER. We evaluate several different approaches for selecting events to include in the timeline, and several different sets of temporal relation labels, and show that high levels of agreement are achievable both on which events play a role in the timeline, and which pairs of events should be linked by a temporal relation.

2. Children's Stories and Event Timelines

We draw our children's stories from the set of Aesop's fables collected by (McIntyre and Lapata, 2009)[1]. As an example story, consider:

> Two Travellers were on the road together, when a Bear suddenly appeared on the scene. Before he observed them, one made for a tree at the side of the road, and climbed up into the branches and hid there. The other was not so nimble as his companion; and, as he could not escape, he threw himself on the ground and pretended to be dead. The Bear came up and sniffed all round him, but he kept perfectly still and held his breath: for they say that a bear will not touch a dead body. The Bear took him for a corpse, and went away. When the coast was clear, the Traveller in the tree came down, and asked the other what it was the Bear had whispered to him when he put his mouth to his ear. The other replied, "He told me never again to travel with a friend who deserts you at the first sign of danger." `[37.txt]`

A variety of events occur in this story, such as *climbing*, *pretending*, *sniffing* and *asking*. There are also a variety of linguistic signals about the order of such events, for example, the word *Before* in *Before he observed them*, or the word *and* in *came up and sniffed*. Such cues can be used by an annotator to identify the temporal relations in the text.

Following these intuitions about event and temporal structure, coupled with the guidelines discussed below, Figure 1 shows a graph of the event and temporal relation annotations we expect our annotators to identify in this story. The nodes

[1] Data available from the authors at http://homepages.inf.ed.ac.uk/s0233364/McIntyreLapata09/

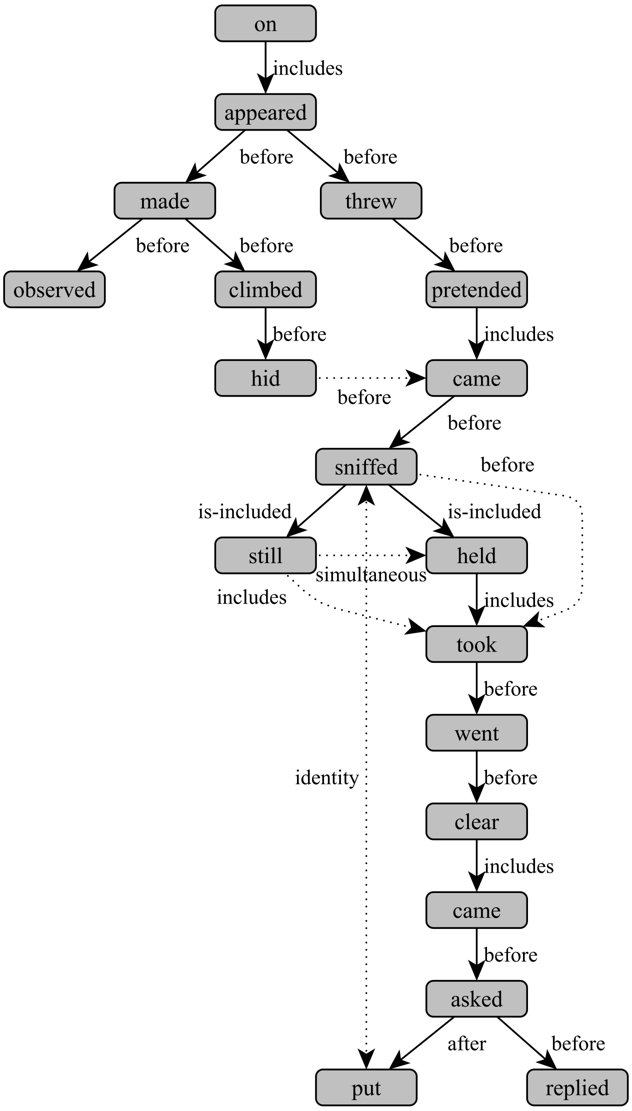


Figure 1: Event timeline for the story of the Travellers and the Bear. Nodes are events and edges are temporal relations. Solid edges are temporal relations signaled by linguistic cues in the text (annotators are expected to identify these). Dashed edges denote other temporal relations, including those that require deeper world knowledge (annotators are not expected to annotate these). Temporal relations that can be inferred via transitivity are not shown (or annotated).

are the events from the story, and the edges are the temporal relations between those events. Note that the solid edges – the edges most motivated by linguistic cues in the story, and the only edges we expect annotators to identify – form a tree structure, with *on (the road)* as the root node.

3. Which events should be annotated?

To annotate story timelines as these temporal dependency structures, we must provide guidelines for annotating both the nodes (events) and the edges (temporal relations). For annotating events, we consider three different annotation schemes:

TimeML Annotators follow the standard TimeML guidelines which annotate all "situations that happen or occur" that are "punctual . . . or last for a period of time" or that are "predicates describing states or circumstances in which something obtains or holds true." TimeML events may appear as verbs, nouns, adjectives or prepositional phrases. When events are phrasal, only the head is annotated, for example:

- *has been* [$_{\text{event}}$ ***scrambling***]
- [$_{\text{event}}$ ***set***] *up*
- *the industry's rapid* [$_{\text{event}}$ ***growth***]

Both light verbs and aspectual verbs are tagged as independent events, for example:

- [$_{\text{event}}$ ***demonstrations***] *have* [$_{\text{event}}$ ***taken***] *place*
- *private sector* [$_{\text{event}}$ ***began***] [$_{\text{event}}$ ***establishing***]

The TimeML guidelines aim to be fairly exhaustive - identifying all words that could potentially play any role in the temporal structure of the story.

No speech or modal Annotators follow TimeML guidelines except that they skip events in direct speech and negated, modal or hypothetical events. Events in direct speech often pose difficulties for novel readers and poor comprehenders, and tend to be less essential to the plot in children's stories For example:

- *"He told me never again to* ***travel*** *with a friend who* ***deserts*** *you at the first sign of* ***danger****"*
- *The Fuller* [$_{\text{event}}$ ***thanked***] *him, but* [$_{\text{event}}$ ***replied***]*, "I couldn't* ***think*** *of it, sir: why, everything* ***I take*** *such pains to* ***whiten*** *would be* ***blackened*** *in no time by your charcoal."*
- *"That's* ***awkward****,"* [$_{\text{event}}$ ***said***] *the Cat to herself: "the only thing to do is to* ***coax*** *them out by a* ***trick****."*

Negated, modal and hypothetical events are often hard to place along a timeline. For example:

- *. . . but he* [$_{\text{event}}$ ***kept***] *perfectly* [$_{\text{event}}$ ***still***] *and* [$_{\text{event}}$ ***pretended***] *to be* ***dead****, for they* ***say*** *that a bear will not* ***touch*** *a dead body.*
- *Imagining the bird must be* ***made*** *of gold inside, they* [$_{\text{event}}$ ***decided***] *to* ***kill*** *it in order to* ***secure*** *the whole store of precious metal at once.*
- *An Old Woman* [$_{\text{event}}$ ***made***] *an* [$_{\text{event}}$ ***agreement***] *with him in the* [$_{\text{event}}$ ***presence***] *of witnesses that she should* ***pay*** *him a high fee if he* ***cured*** *her, while if he* ***failed*** *he was to* ***receive*** *nothing.*

Essentially, this set of guidelines aims to simplify the annotation process by focusing on just the events that are most essential to constructing the timeline.

Paraphrasing Annotators follow the **No speech or modal** guidelines, and additionally, when they encounter phrasal events, select the word that best paraphrases the meaning. For example:

- *. . .* ***kept*** *perfectly* [$_{\text{event}}$ ***still***]*. . .*

- . . . ***used** to* [$_{\text{event}}$ ***snap***]. . .
- . . . ***managed** to* [$_{\text{event}}$ ***scramble***]. . .
- . . . ***did** his **best** to **reach** them by* [$_{\text{event}}$ ***jumping***]
- . . . [$_{\text{event}}$ ***went***] *and **began** to* [$_{\text{event}}$ ***fell***] *a tree. . .*
- . . . ***let** herself* [$_{\text{event}}$ ***hang***] *down. . .*

Using light and aspectual verbs for comprehension questions (e.g. "Where did he manage?" or "How did she let") makes little sense – only the main events contain the necessary semantics (e.g. "Where did he scramble?" or "How did she hang?"). The paraphrasing guideline is thus intended to focus on the events with the greatest semantic content, that is, the events that are most essential for understanding the story and its timeline. Note that this guideline means that, unlike the standard TimeML scheme, light verbs and aspectual verbs are rarely tagged as story events.

These different approaches to event annotation are evaluated in Section 6..

4. Which relations should be annotated?

Having identified several potential guidelines for annotating the nodes (events) in a temporal dependency tree, we now turn to the annotation of edges (temporal relations). Instead of looking for specific temporal signals or particular syntactic constructions, we ask the annotators to link each event in the story to a single nearby event, similar to what has been observed in reading comprehension studies (Johnson-Laird, 1980; Brewer and Lichtenstein, 1982).

For example, consider the events in the following passage:

> Two Travellers were [$_{\text{event}}$ **on**] the road together, when a Bear suddenly [$_{\text{event}}$ **appeared**] on the scene. Before he [$_{\text{event}}$ **observed**] them, one [$_{\text{event}}$ **made**] for a tree. . .

The first event, *on*, becomes the root of the tree. The second event, *appeared*, is related to *on* by the cue word *when*, so we add the relation *on* INCLUDES *appeared*. The third event, *observed*, could potentially be linked to any of the nearby events: *on*, *appeared* or *made*. However, when there are several reasonable nearby events to choose from, the annotators are instructed to choose the temporal relation that is easiest to infer from the text. In this case, the presence of the cue word *Before* suggests that we should add the relation *made* BEFORE *observed*. The result of this annotation process will be a chain or tree of events, linked by temporal relations, much like the one shown in Figure 1.

Note that the temporal dependency trees formed by this annotation process guarantee that all pairs of events are connected by a path of temporal relations through the tree. The resulting annotated timelines are therefore more amenable to temporal reasoning processes such as temporal closure, where additional temporal relations are inferred using, e.g., transitivity properties of temporal relations (Allen, 1983). For example, if event A is BEFORE B and B is BEFORE C, then we can conclude that A is BEFORE C. Such temporal reasoning is difficult to apply when the temporal relation graph is largely disconnected, but much easier to apply when all events are connected as in our temporal dependency trees.

Given this temporal dependency annotation procedure, we consider two annotation schemes, defined in terms of the temporal relation labels that annotators are allowed to use:

Before/Overlap Annotators use the three primary relations from the TempEval challenges (Verhagen et al., 2007; Verhagen et al., 2010): BEFORE, AFTER and OVERLAP. This provides a small set of coarse relations, attempting to keep the task simple by minimizing the number of labels the annotators can choose from.

Before/Includes/Identity/Overlap Annotators select from six temporal relations: BEFORE, AFTER, INCLUDES, IS-INCLUDED, IDENTITY or OVERLAP. This scheme adds some relations from TimeML that were not included in TempEval, with the goal of increasing the expressive power of the resulting annotations. For example, the inclusion relations improve the ability to reason over the resulting annotations. Consider the passage:

> . . . he [$_{\text{event}}$ **threw**] himself on the ground and [$_{\text{event}}$ **pretended**] to be dead. The Bear [$_{\text{event}}$ **came**] up and [$_{\text{event}}$ **sniffed**] all round him. . .

In the **Before/Includes/Identity/Overlap**, these relations would be labeled as:

- *threw* BEFORE *pretended*
- *pretended* INCLUDES *came*
- *came* BEFORE *sniffed*

And we can then use temporal logic to conclude *threw* BEFORE *came*. However, when using the **Before/Overlap** annotation scheme, the INCLUDES relation would instead be labeled OVERLAP, and concluding *threw* BEFORE *came* would no longer be possible.

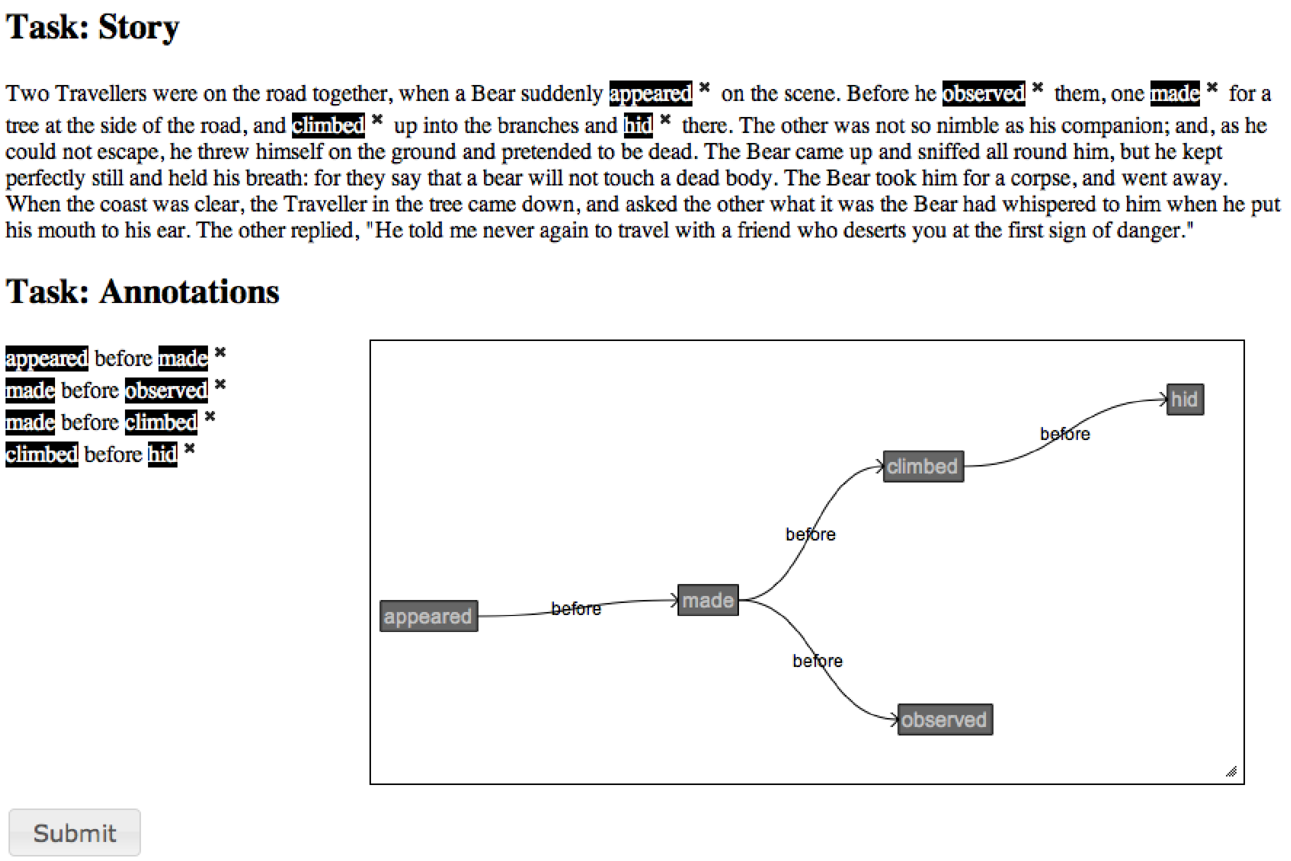
These two temporal relation annotation schemes are evaluated in Section 6..

5. Annotating events and relations jointly

To speed the annotation process, annotators were allowed to annotate events and the temporal relations between them at the same time. A JavaScript/HTML interface presented them with the text of the story and asked them to click on events. Every time two events were clicked, a dialog prompted them for the relation between these two events. Thus the annotators could annotate two events and one temporal relation with only three mouse clicks. Figure 2 shows a screenshot of the annotation interface the annotators used. As annotators linked the events via temporal relations, a graph of the timeline that they were constructing was displayed.

6. Evaluation of annotation schemes

To evaluate the different annotation schemes, two annotators were given a set of 20 stories and asked to annotate



Task: Story

Two Travellers were on the road together, when a Bear suddenly appeared ✖ on the scene. Before he observed ✖ them, one made ✖ for a tree at the side of the road, and climbed ✖ up into the branches and hid ✖ there. The other was not so nimble as his companion; and, as he could not escape, he threw himself on the ground and pretended to be dead. The Bear came up and sniffed all round him, but he kept perfectly still and held his breath: for they say that a bear will not touch a dead body. The Bear took him for a corpse, and went away. When the coast was clear, the Traveller in the tree came down, and asked the other what it was the Bear had whispered to him when he put his mouth to his ear. The other replied, "He told me never again to travel with a friend who deserts you at the first sign of danger."

Task: Annotations

appeared before made ✖
made before observed ✖
made before climbed ✖
climbed before hid ✖

Figure 2: The interface for annotating events and temporal relations.

Number of annotators	2
Type of material	children's stories (Aesop's fables)
Annotator background	computational linguistics
Annotator expertise	expert; schema developers
Annotator training	several days with sample stories
Annotation purpose	corpus for machine learning
Agreement index	Krippendorff's nominal Alpha

Table 1: Overview of the annotation and its evaluation.

Stories	Scheme	Events
1-20	TimeML	0.729
1-20	No speech or modal	0.833
1-20	Paraphrasing	0.876
1-100	Paraphrasing	0.856

Table 2: Annotator agreement (Krippendorff's Alpha) for the event annotation schemes, on which words should be identified as events.

them with each scheme. Table 1 gives some summary information about the annotation, following the suggestions for annotation studies in Bayerl and Paul (2011).

Krippendorff's Alpha for nominal data (Krippendorff, 2004) was used to measure their inter-annotator agreement. For agreement on which words were events, Alpha was calculated considering a binary decision for each word: is it an event or not? For agreement on which event pairs participated in temporal relations, Alpha was calculated considering a binary decision over each possible pairing of two events in the story: is there a relation here or not? For agreement on which temporal relation to assign to an event pair, Alpha was calculated considering a multi-class decision over each event pair where some relation was annotated: which temporal relations holds between these two events?

Table 2 shows the results of these evaluations for event annotation. Agreement was higher using the **No speech or modal** annotation scheme (0.833) than the **TimeML** annotation scheme (0.729), and highest when using the **Paraphrasing** annotation scheme (0.876). This suggests that restricting the sets of events to be tagged (e.g. excluding speech events, modal events and aspectual events) made it easier for annotators to agree on which words in the story represented events.

Table 3 shows the evaluation results for temporal relation annotation. Two types of agreement were measured for temporal relations: agreement on which pairs of events to link, and agreement on the temporal relation labels to assign a given pair of events. Agreement on which pairs of events to annotate was similar regardless of whether the **Before/Overlap** or **Before/Includes/Identity/Overlap** schemes were used (0.856 vs. 0.854), though agreement on which relation to

Stories	Scheme	Event pairs	Relation labels
1-20	Before/Overlap	0.856	0.653
1-20	Bef/Inc/Ide/Ove	0.854	0.629
1-100	Bef/Inc/Ide/Ove	0.822	0.700

Table 3: Annotator agreement (Krippendorff's Alpha) for the temporal link annotation schemes, both on which pairs of events participate in temporal relations, and which temporal relation label should be assigned to each pair.

assign to a pair of events was slightly lower for the scheme with additional relations (0.629 vs. 0.653). This suggests that the annotation scheme with six relation types, which allows for better temporal reasoning (via more transitive inferences through INCLUDES and IDENTITY) can generally be substituted for the three relation annotation scheme without much loss in inter-annotator agreement.

Using the **Before/Includes/Identity/Overlap** and **Paraphrasing** annotation schemes, 100 stories were then annotated. The last rows of Table 2 and Table 3 show that agreement over the whole corpus was similar to that of the original 20 documents. Agreement was slightly lower for events and event pairs, and slightly higher for temporal relation labels.

7. Discussion

In this article, we have shown that in children's stories, the plot timeline can be accurately annotated as a form of temporal dependency structure, where all events in the plot are connected by a single spanning tree. Under this approach, we have observed the highest annotator agreement when direct speech, modal, negated, hypothetical and aspectual events are omitted from the timeline, and have also observed that agreement is still high with an expanded relation set of BEFORE, AFTER, INCLUDES, IS-INCLUDED, IDENTITY and OVERLAP. Our annotated corpus is available at `http://www.bethard.info/data/fables-100-temporal-dependency.xml`.

The primary goal of this annotation effort was to build a corpus that will enable the training of better models for extracting timelines. Using our corpus, Kolomiyets et al. (2012) has already made some promising steps towards better timeline extraction using shift-reduce dependency parsing. We hope that our annotation effort here will inspire other research along these lines.

One of the limitations of the current study is that we focused on children's stories, in part because they typically have simpler temporal structures. The temporal ordering of event words in the story text often follows the true ordering of the events: over 80% of events in our corpus are linked to adjacent events, and nearly 60% of links are of type BEFORE. In other domains, the timeline of a text is likely to be more complex. For example, in clinical records, descriptions of patients may jump back and forth between the patient history, the current examination, and procedures that have not yet happened.

In future work, we plan to investigate how best to apply the dependency structure approach to such domains. One approach might be to first group events into their *narrative containers* (Pustejovsky and Stubbs, 2011), for example, grouping together all events linked to the time of a patient's examination. Then within each narrative container, our dependency tree approach to annotation could be applied. Another approach might be to join the individual timeline trees into a document-wide tree via discourse relations or relations to the document creation time. Work on how humans incrementally process such timelines in text may help to decide which of these approaches holds the most promise.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This research was carried out in the framework of the TERENCE European project (EU FP7-257410).

8. References

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, November.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725, Dec.

Steven Bethard, James H. Martin, and Sara Klingenstein. 2007. Finding temporal structure in text: Machine learning of syntactic temporal relations. *International Journal of Semantic Computing (IJSC)*, 1(4):441–458, December.

William F. Brewer and Edward H. Lichtenstein. 1982. Stories are to entertain: A structural-affect theory of stories. *Journal of Pragmatics*, 6(5-6):473 – 486.

P.N. Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive Science*, 4(1):71–115.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, July. Association for Computational Linguistics.

K. Krippendorff. 2004. *Content analysis: An introduction to its methodology*. Sage Publications, Inc.

Neil McIntyre and Mirella Lapata. 2009. Learning to tell tales: A data-driven approach to story generation. 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 217–225, Suntec, Singapore, August. Association for Computational Linguistics.

J. Pustejovsky and A. Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5 Fifth International Workshop on Computational Semantics*.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TimeBank corpus. In *Corpus Linguistics*, volume 2003, page 40.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic, June. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, July. Association for Computational Linguistics.