Using Coh-Metrix Temporal Indices to Predict Psychological Measures of Time

Nicholas D. Duran (nduran@memphis.edu)
Phillip M. McCarthy (pmccarthy@memphis.edu)
Arthur C. Graesser (a-graesser@memphis.edu)
Danielle S. McNamara (dsmcnamr@memphis.edu)

Institute for Intelligent Systems
Department of Psychology
202 Psychology Building
Memphis, TN 38152 USA

Abstract

Situation model theories of text comprehension consider temporality to be one of the critical dimensions for building a coherent mental representation of described events. Using this framework, three continuous scale measures were developed to assess temporal coherence based on *tense*, *aspect*, and *adverbial* relations. Experts in discourse processing evaluated 150 texts, excerpted from *science*, *history*, and *literature* textbooks, to establish a *gold standard* of temporality. We then demonstrated that Coh-Metrix, a computational tool that measures textual cohesion on over 200 indices of discourse features, could significantly reflect these human interpretations by incorporating five indices of local, temporal cohesion. We conclude our paper with a discussion of our current research into developments of more sophisticated global temporal indices.

Introduction

Three grammatical devices primarily establish temporal relations in text: *tense*, *aspect*, and *adverbial elements* (Klein, 1994). These linguistic markers are not only important for the structure of discourse, but also facilitate the mental representations of situations described in language (Zwaan & Radvansky, 1998). The markers function as instructions for integrating the interaction of entities, properties, and actions into a coherent mental model of comprehension. Temporal cues, along with other dimensions of situation models (e.g., *space*, *causation, intentionality,* and *protagonist*), also foreground relevant information in a reader's interpretation of described events (Zwaan, Langston, & Graesser, 1995). While such research has contributed significantly to our understanding of situation model construction, computational limitations have restricted empirical research to relatively short passages of manipulated text. Advances in technology, however, now allow large corpora to be indexed to mark the presence of grammatical and lexical features assumed to play important roles in situation model construction.

Graesser et al. (2004) have integrated these text-based linguistic features, as well as other indices of readability and vocabulary, into a web-based software tool called Coh-Metrix (for additional information, visit cohmetrix.memphis.edu). One of the benefits of the tool has been its ability to assess textual cohesion. Cohesion explicitly connects linguistic constituents, propositions, conceptual themes and sub-themes, thereby assisting the reader in generating inferences and bridging conceptual gaps (e.g., McNamara, 2001). Differences in cohesion within a text can be correlated with the coherence of a reader's interpretation of a text. In other words, cohesion as a textual construct can be mapped onto coherence as a psychological construct. However, while the effects of coreference cohesion indices such as argument overlap (Kintsch & van Dijk, 1978) and LSA (Foltz, Kintsch, & Landauer, 1998) have been well-tested, there has not been the same focus on whether temporal indices of cohesion are able to distinguish relevant temporal themes and sub-themes that contribute to coherence. The purpose of this study is to extend the work of Coh-Metrix into the temporal aspect of cohesion. Specifically, we analyze a corpus of *narrative*, *history*, and *science* texts to ascertain the degree to which temporal cohesion indices predict human derived psychological measures of temporality.

Temporal Measures

To establish a psychological *gold standard* for temporal cohesion, we developed three distinct, continuous scale measures. The measures are motivated by theories of situation models that propose that comprehenders make use of linguistic cues, such as *tense*, *aspect*, and *adverbs* to construct temporal dimensions within their mental models (Graesser, McNamara, & Louwerse, 2003; Graesser, Singer, & Trabasso, 1994; Zwaan & Radvansky, 1998).

Grammatical tense, for instance, assists a reader in organizing events along a timeline. Specifically, tense establishes the time of occurrence for an event around a referential point, such as time of utterance. The resulting temporal structure, that places events and actions within the text along a continuum, affects the activation of information in working memory. For example, Carreiras et al. (1997) manipulated tense to allow associations between a character and a job description to be applied in the present (e.g., *Marta now works as an economist*) or separated by a lapse in time (e.g., *Marta in the past worked as an economist*). The recall of the character's occupation was faster and more accurate when the association was depicted in the condition with close temporal proximity.

Whereas tense relates lexical events to a certain point in time, the use of *aspect* in temporal processing conveys the dynamics of the point itself (Klein, 1994). The use of the present participle aspect, such as *painting* in the sentence *Sam is painting the house,* distinguishes the content as ongoing, whereas the use of the perfective aspect, such as the word *painted* in the sentence *Mary has painted the house*, distinguishes the content as completed with an effect remaining in the present.

Magliano and Schleich (2000) emphasized the importance of *aspect* as a cue for maintaining information in working memory. Perfective events are processed as completed and decay faster in working memory than ongoing present participle events. Aspect cues the reader to tag information in their current mental representation that might be relevant for connecting subsequent discourse.

Ohtuska and Brewer (1992) have also shown that a sequence of events is best comprehended when the order of mention in the text corresponds to true chronological order. This psychological default, referred to as the *iconicity assumption*, is modified by grammatical temporal markers. Verb aspect and tense cue the reader to when the *iconicity assumption* should or should not be followed. A passage written in the past perfective (i.e., *had jumped* in a sentence such as *Jimmy had jumped the fence*) sequences the events to fit iconicity. A present participle passage violates the assumption by allowing different events to occur simultaneously. Dowty (1986) argues that in order for the iconicity assumption to be fully incorporated into a situation model construction, events in the text must be both concurrent and contiguous. Adverbial phrases serve the function of modifying this representation. Temporal adverbs and connectives convey time in language by explicitly stating the chronological distance between events (e.g., *before, after, then).* Along these lines, Zwaan (1996) found that when time shifts were manipulated in a passage by the use of temporal adverbs (e.g., *in a moment / five minutes later / the next day*), the mental representation of events was adversely affected by adverbs that imposed a greater gap in time.

The organizational influences of *tense*, *aspect*, and *adverbs* are the crux of our measures for interpreting temporal coherence in texts. Each measure has been implemented on three separate scales that can be assessed for degree of importance. We assumed this would be the most effective approach for capturing inferential generalizations about time. The psychological measures that act as our *gold standard* are presented below.

Measure 1 The *temporal marker salience measurement* is the extent to which temporal word markers (e.g., *adverbials, connectives, particles, dates*) present in the text establish possible event orders on a timeline. Events can follow, precede, or overlap one another, jump in time (e.g., *flashbacks* and *flashforwards*) and have large lapses between them. Temporal word markers assist the reader in properly establishing and ordering these events (Zwaan, Madden, Stanfield, 2001).

Measure 2 The *timeline proportion measurement* captures the level of difficulty in translating the words and sentences into a coherent flow of temporal events. This measure will approximate a reader's ability to distinguish a moment of occurrence relative to its past and future occurrences. In doing so, the measure establishes the proportion of the text that can be easily reconstructed on any timeline structure. (Klein, 1994).

Measure 3 The *iconicity measurement* is the extent to which the order of mention in the text corresponded to the underlying order of events, as reflected in the linguistic features of tense and aspect. When a chronological sequence of events matches the perceptual experience of a reader, the text should be more easily integrated into a coherent representation (Zwaan, 1996).

Coh-Metrix

Coh-Metrix is a computational tool that incorporates over 250 lexical and discourse indices (Graesser et al., 2004). Coh-Metrix harnesses the most recent developments in computational linguistics and discourse processing, featuring advanced syntactic parsers (Charniak, 1997), part-of-speech taggers (Brill, 1995), and Latent Semantic Analysis (LSA, Landauer & Dumais, 1997). Word relationship indices are derived from the WordNet lexical database (Miller, 1990), and psycholinguistic information from the MRC database (Coltheart, 1981). A variety of shallow metrics such as Flesh-Kincaid Grade Level (Klare, 1974, 1975) are also added for purposes of comparison.

These modules are integrated into the automated computational tool, Coh-Metrix, for generating comprehensive cohesion *profiles* of text (Graesser et al., 2004). Coh-Metrix has been involved in many research endeavors, ranging from learning assessment to text identification. For instance, McNamara and colleagues examined the cohesion of textbooks and the resulting benefit for high or low knowledge readers (Best, Ozuro, & McNamara, 2004). Louwerse et al. (2004) investigated cohesion in written and spoken texts, finding six dimensions of relationships between the two modes. McCarthy, Briner et al. (2006) used Coh-Metrix to distinguish segments of texts by their functional relationship. And McCarthy, Lewis et al. (2006) used Coh-Metrix to distinguish texts of different authors even while the individual styles of the authors was shown to significantly shift as the author's style developed.

Temporal Indices

For the present purpose of investigating psychological, temporal coherence measures, we focus on the computationally derived, Coh-Metrix temporal indices. In total, this study used nine temporal indices: six indices are available on the current online version of Coh-Metrix, and a further three were developed for this study. The method of calculation for the current indices is via a *density score* that measures the incidence of a particular category per 1,000 words. Such indices serve a global purpose, assessing the overall substantive content in the text as a whole (Graesser et al. 2002).

The six indices currently available on the online version of Coh-Metrix are delineated into three categories that are on based on grammatical function: *part of speech, connectives,* and *ambiguous elements*. The temporal part of speech indices include; *incidence of past participles* (e.g., *awoken, begun, become*), *incidence of past tense* (e.g., *awoke, began, looked*), and *incidence of present tense* (e.g., *look, move, talk*). The connective indices include; *incidence of positive temporal connectives* (e.g., *before, after, while*), and *incidence of negative temporal connectives* (e.g., *until, from, since*). The ambiguous elements score include *temporal adverbial phrases,* which consist of non-explicit linguistic features (e.g., *at this time, sooner or later*).

As mentioned above, we developed three additional cohesion indices to be incorporated into the Coh-Metrix tool. These additional indices broaden the scope of accounting for the various temporal relations in a text by capturing all explicit adverbs in a text. The first index features all explicit, textual adverbs (e.g., *now, then, yesterday*), as well as numerical dates (e.g., *1997, 435 B.C.)* and nominal dates and time periods (e.g., *Monday, summer, October*). The remaining indices are derivations of these explicit elements. The first includes a score of combined numerical dates, nominal dates, and time periods, and the second combines only numerical dates and time periods. The method of calculation is a ratio score that takes the instances of a category divided by all words in a text. These additional three indices are necessary to understand the relative importance of text features that explicitly place events on a timeline.

Methods

The criteria describing our psychological *gold standard* measures were given to three experts working on discourse processing at the Cognitive Science Educational Practice (CSEP) lab at the University of Memphis. The three experts assessed a corpus of 150 narrative and expository texts to establish agreed upon benchmarks. The Coh-Metrix temporal indices were then used to predict these human interpretations of time.

Corpus Selection

A total of 150 texts, including 50 texts from each of three prominent domains (i.e., *science*, *history*, and *narrative*), were selected from an electronic corpus of academic textbooks provided by MetaMetrics Inc. For reasons of continuity, each text was shortened by randomly selecting paragraph-to-paragraph slices of approximately 400 words. Several selection constraints were applied to ensure uniformity and representation across grade levels and authorship: Within each representative domain, 25 texts from the high-school grades (10^{th} - 12^{th}) and 25 texts from the junior-high grades (7^{th} - 9^{th}) were sampled. Within each grade level, three or more unique textbooks were sampled (see Table 1). In addition, all texts were assessed to guarantee that paragraph breaks and sentences were properly located. All captions, headings, maps, and figures were removed.

To provide confidence in the generalizabilty of our planned statistical analysis, we ensured that a normal distribution of general cohesion for each of the three domains was present. This was achieved through checks of distributions of two major measures of cohesion: *argument overlap* (Graesser et al., 2004) and *Latent Semantic Analysis* (LSA, Landauer et al., 1997).

Table 1: Distribution of unique textbooks and text segments for High School (10th-12th) and Junior High (7th-9th).

	History		Narrative		Science	
	High	Jr.	High	Jr.	High	Jr.
Unique Books	5	5	4	5	3	5
Total Texts	25	25	25	25	25	25

Both argument overlap and LSA are robust in assisting a reader to relate ideas and fill conceptual and structural gaps across text (Graesser et al., 2004). Specifically, argument overlap tracks arguments and word-stems in adjacent sentences for assessing similarity (McCarthy, Lewis, et al. 2006). LSA, on the other hand, is a high-dimensional semantic network that represents words by their shared contextual history in the language environment (e.g., *street* and *road* appear in similar lexical contexts). A composite score of words in sentences can be compared to adjacent composite sentences and all possible combinations for evaluating global semantic relationships. A normal distribution of global cohesion for each domain was obtained using these two measures.

Experimental Design

Inter-rater reliability The human measurements of temporal coherence were assessed using a Likert-type scale ranging from 1 (minimum) to 6 (maximum). A bivariate Pearson correlation for each question was conducted between all possible pairs of raters' responses. Additionally, agreement was analyzed using Cohen's weighted kappa statistic (Altman, 1990). This test is beneficial in compensating for the disagreement that is more likely to occur with a continuous numerical scale. If any two raters were below the *good* threshold (kappa < .06) established by Landis and Koch (1977) and/or correlations were not significant at the $p < .05$ level, ratings were then reexamined and scores were agreed upon by the three raters.

The initial assessment of inter-rater reliability for the psychological *gold standard* measure of *temporal marker salience measurement* indicates that scores ranged from *moderate* to *good* agreement. Judgments between the three possible pairs of raters were significantly correlated, with one of the kappa scores above the 0.6 threshold and the other two slightly below (see Table 2). The psychological *gold standard* measure of *timeline proportion measurement* received consistent scores with all pairs of raters establishing *good* to *very good* kappa agreement, as well as significant correlations (see Table 3). The results for the psychological *gold standard* measure of *iconicity measurement* suggested that raters were not as certain in interpreting texts for chronological order (see Table 4). Though these results were lower than the previous two measures, they remained significant.

The mean score of all three raters per text was taken as the final *gold standard* rating of the 150 texts. After discussions to correct for discrepancies, reevaluated scores resulted in significant correlations and kappa scores that were all in good to excellent agreement (above .6 thresholds). The final scores constitute the empirically established benchmarks for analysis in the linear regression.

Table 2: Inter-rater reliability for the *temporal marker salience measurement.*

Rater Comparison		Kappa	Pearson *r*
Rater 1	Rater 2	.744	.674*
Rater 1	Rater 3	.498	.541*
Rater 2	Rater 3	.512	.588*

*Correlation is significant at $p < .001$.

Table 3: Inter-rater reliability for the *timeline proportion measurement.*

Rater Comparison		Kappa	Pearson r
Rater 1	Rater 2	.736	.754*
Rater 1	Rater 3	.699	.700*
Rater 2	Rater 3	.758	.783*

*Correlation is significant at $p < .001$.

Table 4: Inter-rater reliability for the *iconicity measurement.*

Rater Comparison		Kappa	Pearson r
Rater 1	Rater 2	.412	.363*
Rater 1	Rater 3	.321	.342*
Rater 2	Rater 3	.439	.383*

*Correlation is significant at $p < .001$.

Prediction equation Based on the size of the current dataset, we estimated that five indices would be the maximum number of variables available before problems with overfitting occurred. To provide an objective test of the analysis, a training set of 100 randomly chosen texts (selected from the 150 texts) was established for use in building a prediction equation. We selected variables from the three categories of grammatical function: *part of speech, connectives, and ambiguous elements*, as well representatives from the new indices of *explicit temporal elements.* The variable with the highest correlation to the *gold standard* human measures were selected from each group as predictor variables. Other variables were added provided they passed a co-linearity check (Hair et al., 1998) and that they correlated at $r < .7$.

As a result, the five following predictors variables were used: *temporal elements score*, *incidence of positive temporal connectives*, *past tense parts of speech score*, *present tense part of speech score*, and *ambiguous temporal incidence score*.

Results

A series of forward-entry linear regressions were conducted with each of the three human temporal measurements as the dependent variable. The linear regression produced a set of unstandardized b-weights based on the five Coh-Metrix predictor variables. Any b-weight that was not significant was discarded.

The remaining b-weights were multiplied by their corresponding Coh-Metrix scores in the 50-text test set and added together with the constant to create prediction scores. These scores were then correlated with the actual human scores to determine the degree to which Coh-Metrix temporal indices mirrored human performance.

For each of the three temporal benchmarks, distinct combinations of b-weights were used to predict human scores (see Table 5). The correlations comparing the predicted scores and actual scores were all highly significant (see Table 6).

For the *temporal marker salience measurement,* the most predictive b-weights corresponding to the Coh-Metrix temporal indices were (in decreasing order of significance): 1) *all temporal elements* score, 2) *present tense part of speech* score, and 3) *past tense part of speech* score. For the *timeline proportion measurement,* the most predictive b-weights corresponding to the Coh-Metrix temporal indices were (in decreasing order of significance): 1) *all temporal elements* score, and 2) *past tense part of speech* score. For the *iconicity measurement,* the most predictive b-weights corresponding to the Coh-Metrix temporal indices were (in decreasing order of significance): 1) *ambiguous temporal incidence* score and 2) *past tense part of speech* score. As such, the cohesion indices that emerged from this analysis as being most indicative of the gold standard were incidence scores for *part of speech* and *ambiguous elements*, as well as ratio scores for *explicit elements*. The two connective indices, *incidence of positive temporals* and *incidence of negative temporals*, were not significantly predictive.

Table 5: b-weights of temporal coherence measurements regressed on Coh-Metrix indices for 100-text training set.

Indices	Temporal marker salience	Timeline proportion	Iconicity
All temporal elements score	.032**	.025**	.002
Ambiguous temporal incidence score	.013	.036	.066**
Present tense part of speech score	-.023**	-.014	-.007
Past tense part of speech score	.009*	.023**	.016**
Incidence of positive temporal connectives	.002	-.011	-.028
Constant	2.748	2.143	2.297

*b-weight is significant at $p < .05$, **b-weight is significant at $p < .001$.

Table 6: Correlations between predicted scores and actual scores for temporal coherence measurements.

Measurement	Correlation
Temporal marker salience	0.766*
Timeline proportion	0.850*
Iconicity	0.495*

*Correlation is significant at $p < .001$

Discussion

The present study investigated the organization of temporal relations in text and the influence of these relations on interpretations of temporal coherence. The results suggest that distinctions made by human raters can be successfully identified by Coh-Metrix cohesion indices. Consequently, the utility of Coh-Metrix indices address the cognitive processing of lexical cues of temporality that, according to Zwaan et al. (2001), have received little attention in language comprehension. The ability of Coh-Metrix to reflect human performance highlights the importance of text-based cues for comprehension.

While the results of this study are significant, suggesting that local-level cohesion cues can accurately predict human interpretations of time, we are currently working to develop even more sophisticated temporal indices. These new indices will serve to capture both the grammatical mode and global structure of cohesion. These new *global indices* will provide further discriminatory power by offering representations tense and aspect consistency across individual paragraphs as well as the entire text. Such new indices will be particularly beneficial to our assessment of the iconicity assumption.

A further goal for our temporal investigations is to address the concern that situation model dimensions are studied in isolation while coherence is dependent on the interaction of dimensions (Zwaan et al., 1998). Temporal relationships, for instance, are also influenced by *causal links* between events. As CohMetrix also provides a wide variety of causal and intentional indices, we believe that representation of temporal and causal cohesion can be assessed simultaneously. This may provide insight concerning the co-relevance of these dimensions for comprehension of a text.

The final avenue of interest includes cataloguing characteristics of text domain with indices of cohesion. McCarthy, Lightman et al. (2006) have been successful in showing that cohesion rates in academic textbooks are consistent with the grade level of difficulty, but vary according to domain (e.g., science and history). Future research will build on this approach by examining the situational dimensions that are most relevant in different domains. It is possible that a reader in situation construction may be influenced not only by world knowledge and text-based cues, but feature-specific qualities of genre. The profiles of Coh-Metrix temporal indices from our current study will certainly contribute to this endeavor, as well as exploring further issues of temporal coherence in natural language processing.

While much work remains to be done, this initial investigation contributes to the field by demonstrating that Coh-Metrix derived temporal indices can accurately reflect human evaluations of temporality.

Acknowledgments

The research was supported by the Institute for Educational Sciences (IES R3056020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We would also like to thank Stephen Briner and Erin Lightman for their contributions to this study.

References

Best, R., Ozuru, Y., & McNamara, D. S. (2004). Self-explaining science texts: Strategies, knowledge, and reading skill. In Y. B. Yasmin, W. A., Sandoval, N. Enyedy, A. S. Nixon, F. Herrera (Eds.), Proceedings of the Sixth International Conference of the Learning Sciences: Embracing Diversity in the Learning Sciences (pp. 89-96). Mahwah, NJ: Erlbaum.

Brill., E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics, 21*, 543-565.

Carreiras, M., Carriedo, N., Alonso, M. A., & Fernandez, A. (1997). The role of verbal tense and verbal aspect in the foregrounding of information in reading. *Memory & Cognition, 25*, 438-446.

Charniak, E. (1997). Statistical techniques for natural language processing. *AI Magazine, 18*, 33-44.

Dowty, D. R., (1986). The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics? *Linguistics and Philosophy, 9*, 37-61.

Foltz, P., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*, 285–307.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*, 193-202.

Graesser, A.C., McNamara, D.S.,& Louwerse, M.M (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A.P. Sweet and C.E. Snow (Eds.), Rethinking Reading Comprehension (pp. 82-98). New York: Guilford Publications.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101,* 371-395.

Kintsch W. & van Dijk T A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-94.

Klein, W. (1994*). Time in Language*. London: Routledge.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Louwerse, M.M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, & T. Regier (Eds.), Proceesings of the 26th Annual Meeting of the Cognitive Science Society (pp. 843-848). Mahwah, NJ: Erlbaum.

Louwerse, M. M. (2002). An analytic and cognitve parameterization of coherence relations. *Cognitive Linguistics, 12*, 291-315.

Magliano, J. P., & Schleich, M. C. (2000). Verb aspect and situation models. *Discourse Processes, 29*, 83-112.

McCarthy, P.M., Briner, S.W., Rus, V., & McNamara, D.S. (2006). Textual Signatures: Identifying text-types using Latent Semantic Analysis to measure the cohesion of text structures. In: A. Kao, S. Poteet (Eds.). Natural Language Processing and Text Mining. UK: Springer-Verlag.

McCarthy, P.M., Lewis, G.A., Dufty, D.F., & McNamara, D.S. (2006). Analyzing writing styles with Coh-Metrix. In Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS), Melbourne, Florida.

McCarthy, P.M., Lightman, E.J., Dufty, D.F. & McNamara, D. S. (submitted). Using Coh-Metrix to assess distributions of cohesion and difficulty: An investigation of the structure of high-school textbooks. *28th Annual Conference of the Cognitive Science Society, 2006.*

McNamara, D. S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51-62.

Miller, G. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3,* 235-312.

Ohtuska, K., & Brewer, W. F. (1992). Discourse organization in the comprehension of temporal order in narrative texts. *Discourse Processes, 15*, 317-336.

Radvansky, G. A., Zwaan, R. A., Fedrico, T., & Franklin, N. (1998). Retrieval from temporally organized situation models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1224-1237.

Zwaan, R. A., Madden, C. J., & Stanfield, R. A. (2001). Time in narrative comprehension: A cognitive perspective. In D. Schram & G. Steen (Eds.), *The Psychology and Sociology of Literature.* Amsterdam/Philadelphia: John Benjamins.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.

Zwaan, R. A. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1196-1207.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*, 292-297.