GPAT: A Genre Purity Assessment Tool

Philip M. McCarthy

Department of English
University of Memphis, Memphis, TN, USA
pmmccrth@memphis.edu

Abstract

This study introduces a Genre Purity Assessment Tool (GPAT). GPAT calculates *genre purity* by using SIF n-graphs (statistically improbable graph strings) to identify genre characteristics in text. The study describes the tool and assesses it across five experiments that feature a variety of text types and text lengths. The results demonstrate that GPAT is at least as effective as a system that uses a combination of 30 complex textual analysis indices. The results further demonstrate that GPAT is informative on texts as short as three words. The study is of value to discourse psychologists, psycholinguistics, and any researchers for whom the genre of texts is a component of the analysis.

Introduction

A genre is composed of an underlying and diverse set of (unconsciously) agreed upon characteristics (Downs 1998; Hymes 1972). However, McCarthy et al. (2009) demonstrate that although the "presence, prevalence, and prominence" of these characteristics allow a text (T) to be recognized as a member of a genre (G), any given text is only about 83% homogeneous in terms of genre. The composition of the remainder of that text (i.e., the other 17%) will vary depending on the genre to which it is originally assigned. For instance, the remainder of a narrative text tends to be composed mostly of history-like-structures; the remainder of a history text is composed mostly of narrative-like-structures, and the remainder of a science text is composed of an equal division of narrative- and history-like structures. McCarthy and colleagues reasoned that the internal variation in the heterogeneity of the text (relevant to its genre) may benefit some readers more than others in relation to their knowledge or skill level. This variation can be attributed to readers' text comprehension being influenced by their familiarity with the characteristics (or structure) of the text (Bhatia 1997; Graesser et al. 2002; Zwaan 1993). That is, research suggests that skilled readers utilize different comprehension strategies relative to the text genre that they identify (van Dijk & Kintsch 1983; Zwaan 1993). Thus, readers' memory activations, expectations, inferences, depth of comprehension, evaluation of truth and relevance, pragmatic ground-rules, and other psychological mechanisms depend upon the readers' interpretation of the text's genre. For instance, readers are less likely to need to evaluate *the truth of events* for a narrative, although world knowledge has importance for expository texts (Gerrig 1993). Similarly, unfamiliar topics in a text (more likely in expository texts) cause a greater cognitive burden for readers than familiar topics (Otero, Leon, & Graesser 2002). Empirical evidence of differences between the processing of narrative and expository texts is also evident, from tests of recall (Graesser et al. 1980) to reading time (Graesser, Hoffman, & Clark 1980), demonstrating that narratives are recalled approximately twice as well as and read twice as fast as their expository counterparts.

Switching a text from being largely expository to being largely narrative (or visa-versa) is either impossible or fraught with difficulty. However, modifying the degree to which a text is of one genre or another is presumably possible. With this in mind, researchers such as McCarthy et al. (2009) have hypothesized that increasing the incidence of narrative structures in an expository text might facilitate lesser skilled/knowledge readers, even if the modifications are not always central to the content of the text. That is, higher incidences of narrative structures, wherever they might occur, are likely to be more familiar to the reader and, therefore, more easily processed. At the very least, such modifications may reduce the cognitive burden on the reader, freeing up resources for processing more content relevant information. Given this supposition, an evaluation of the degree to which a text is *of a genre* is useful information for researchers to factor into their text assessments.

If psychological studies do not fully evaluate the degree to which a text is of a genre (i.e., its genre purity level), then there is the danger that researchers will continue to assume that they have used narrative texts or used expository texts, when it is very unlikely that such texts are homogenous to one genre (e.g., Trabassao & Batolone 2003; Radvansky et al. 2001; and Zwaan, Magliano, & Graesser 1995). Such a homogeneity assumption not only runs counter to McCarthy and colleagues' findings, but also to those of Rouet et al. (1996) and Hendersen and Clark (2007). For instance, Rouet and colleagues show that expository texts can have several variations, while Hendersen and Clark demonstrate numerous effects based on varying the type of narrative. These studies raise questions as to experimental analyses that do not provide a clear indication of the *genre purity* of the texts. Indeed, a genre purity assessment of the texts is vital so that researchers may more clearly understand the implications of their findings. With this in mind, the *genre purity assessment tool* (GPAT) was developed. This paper describes that tool and a series of five experiments that assess it.

Genre Purity Assessment Tool (GPAT)

GPAT (http://tinyurl.com/5bwo64) is a freely available computational tool designed to assess genre in terms of *expository* (represented by science) and *narrative* (represented by literature). The tool provides five primary indices including values for *expository purity*, *narrative purity*, *shared expository/narrative*, *unknown*, and a binary categorization value of either *expository* or *narrative*. GPAT assesses genre purity without the requirement of resources such as taggers, parsers, or semantic spaces; instead, using *statistically improbable n graph features* (SIF n-graphs). The *graphs* of "SIF n-graphs" refers to keyboard characters (e.g., letters, symbols, numbers, punctuation marks, and the space bar) together with the single symbol of @, which is used to represent all function words. The n-*graphs* of "SIF n-graphs" refer to strings of such graphs, with GPAT using *quadgraphs*, or strings of four characters such as *e d @ t*. SIF n-*graphs* are n-graphs that occur with above average frequency in one corpus (e.g. science texts), and at the same time do *not* occur with above average frequency in a sister corpus (e.g. narrative texts). All SIF n-graphs were obtained using a modified version of the Gramulator (Min & McCarthy 2010), which is typically used to identify SIF n-*grams* (i.e., words) as opposed to SIF n-*graphs*. We refer to examples of SIF n-graphs as *genremes*. For instance "t + y + [comma] + [space]" is a narrative genreme, as is "[quotation mark] + [space] + [function word] + [space]", whereas "s + e + s + [period]" is an expository genreme, as is "[function word] + [space] + p + h." The percentage of the text that is composed of science genremes is the percentage of the text that is of *science purity*. The percentage of the text that is composed of narrative genremes is the percentage of the text that is of *narrative purity*. N-graphs that occur with above average frequency in both corpora are *shared*, and the (typically small) remainder of the text is *unknown*. These values were used to form predictor variables for a discriminant analysis statistical procedure, from which, a model for genre prediction was derived. The resultant coefficients were automated within GPAT for the categorization evaluation (e.g. *science* or *narrative*).

In and of itself, the categorization of texts into genres of science and narrative is a relatively straightforward process (see McCarthy et al. 2009). For instance, the present tense will be common to science texts, and the past tense to narratives. Science texts are also more likely to features low-frequency words. And because science texts require greater explanations, they are more likely to feature sentence to sentence word co-reference (e.g., argument overlap, see Graesser et al. 2004). However, each of these assessments presents problems, even if we assume their accuracy is high. For instance, assessing verb tenses and content word overlap requires a parser, which slows processing; and assessing word frequency requires maintaining a large database of terms. In contrast, the SIF n-graphs approach of GPAT is computationally lighter. Strings of characters do not require maintenance, and are likely to persist with similar accuracies from text to text and corpus to corpus. This flexibility can be attributed to sequence uniqueness such as with narratives featuring past tense endings followed by a closed class word (e d [space] @) or nominal morphemes in sentence ending positions (i t y [period]).

GPAT parameters

An endless variety of options for SIF n-graph parameters were available. At this stage of GPAT development, the following major parameters are incorporated. These parameters should be considered a *starting point*, with the numerous alternatives to be tested in future research.

Textual punctuation is retained because absent a good reason to delete anything from a text, it is general policy to retain it, especially when such aspects may themselves be meaningful (Jurafsky & Martin 2009; p. 193). Furthermore, narrative texts are likely to contain greater numbers of characteristic punctuation such as commas (because of embedded clauses and sentence modifiers that provide textual esthetics as a means of varying the speed of reading).

Function words are converted to a single character representation (the @ is used). The approach uses 219 function words, identified from lists on various websites (e.g. myweb.tiscali.co.uk/wordscape/museum/funcword.html).
There were three major reasons for converting function words to single character representations. First, the problematic high frequency of function words often means that they are excluded from consideration in computational tools (e.g. the LSA approach either *stops* function words or devalues them). Second, research has shown that function words are limited in their ability to distinguish text types (Conway 2008; Homes & Forsyth 1995), presumably because their grammatical (rather than content) role makes them common to most text types. And third, the SIF approach would mean that for any function word to be included, it would have to be highly common to one genre and highly uncommon to the other; in the case of the function words, such an outcome is unlikely. However, excluding function words is also problematic. Simply removing function words means that some graphs would appear to be next to each other when in reality they are separated by one or more function words. As such, GPAT currently retains function words as a single character symbol.

Quad graphs, or *four character long SIF n graphs* are used. The longer a sequence of characters (or words), the more unique it is; and therefore, the less likely it is to be repeated elsewhere. Thus, the longer an n-graph, the fewer examples there will be, and the larger the training corpus must be to find a suitable representation. As such, shorter n-graphs may provide more examples from relatively few texts. But although shorter n-graphs have the potential of providing greater n-graph diversity, longer n-graphs are potentially more diagnostic. In addition, the longer the n-graph, the less likely it is that minor textual changes can affect the evaluation of the text, meaning that materials designers and researchers could make a number of edits without being overly concerned that they are significantly altering the genre purity value of the text. Also needing to be considered are suffixes and prefixes. These morphemes are potentially highly diagnostic, and where such features combine (at the end of one word, over a word boundary [space], and to the beginning of the next word), a potentially highly diagnostic feature is likely to occur. Indeed, 11.6% of narrative genremes and 8.3% of science genremes in this study featured word boundaries. For such a feature to be possible, the n-graphs would have to be a minimum length of three characters. Taking all of these considerations into account, we set the n-graphs length in this study at four characters, acknowledging that more research is needed to optimal and/or better validate these parameters.

Narrative and science genremes were identified using a sub-sample of texts randomly selected from the Touchstone Applied Science Associates (TASA) corpus. The domains of interest were science and narrative texts. Along with an assigned genre, all TASA texts are provided with a Degrees of Reading Power (DRP) value, which is a reading comprehension index developed and administered by TASA. Because the primary area of interest was high-school students, the texts were divided by their DRP value into 13 groups (representing the 12 grades of school and one college grade). The derived grades 7 through 12 were retained and from these grades, 100 texts were randomly sampled (where 100 texts were available), culminating in a corpus of 1192 texts in total (mean length = 282.439, SD = 26.069). We hereafter refer to this corpus as *TASA7 12*. The 1192 texts were then randomly divided into three data sets (*n graph derivation* = 625, *model derivation* = 379, *testing* = 188), with set sizes guided by Witten and Frank (2005) according to the computational requirements of deriving and testing signals from corpora. The *n graph derivation* set was used to identify SIF n-graphs; the *model derivation* set was used in the statistical discriminant analysis procedure to derive coefficients from the narrative and science genre purity values; testing was conducted on the remaining 188 items (see Experiment 1).

Using GPAT

GPAT is available in single file or multiple file processing versions. For the multiple file version, the user browses to the desired folder and then clicks "process". All results are saved to the clipboard and can be pasted to Excel, SPSS, or similar software. For the single file version, the user has the option of browsing to a folder or pasting directly to the tool. Two forms of output are provided: the genre prediction (e.g., narrative or science) as well as values of science and narrative purity in the form of percentages. Shared genre values and the remaining "unknown" percentage are also provided. Two lower windows show the text assessed as a narrative and assessed as a science. The user can toggle between viewing the original text (non-processed) and the text assessed as either genre. All genremes are bolded.

Assessment of GPAT: Method

The validity of any genre purity assessment approach can be gauged by the ability of that approach to categorize examples of texts that are members of the genres in question. Thus, evidence for the validity of GPAT purity values can be proffered by using those purity values as predictors in a model that categorizes a range of narrative/science texts over a number of lengths of texts, and over a number of corpora. In this study, we use five such data sets, with each set described in its relevant experiment.

For GPAT comparison purposes, an alternative categorization model is also considered and accuracy results compared. The alternative model stems from a combination of 193 discourse, sentence, and word indices (cohesion, word frequency, and syntax) from Coh-Metrix (Graesser et al. 2004). Following standard procedures to avoid collinearity issues (see Jurafsky & Martin 2009; McNamara et al. in press.), a final set of 30 variables were culled. These variables included three cohesion (e.g., noun overlap), two frequency (*written frequency logarithm all words [mean]* and *written frequency in sentence [SD]*), and 25 syntax (e.g., *frequency of past tenses*, *lexical diversity*, and *frequency of 3rd person singular*). As previously discussed, cohesion, word frequency, and syntax, especially in combination, are theoretically well suited to categorizing science and narrative texts. Thus, although such a Coh-Metrix model is computationally more expensive (including numerous word lists and parsing), its potential for high accuracy serves as a good guide for the accuracy of the GPAT approach. To maximize the effectiveness of the Coh-Metrix model, a discriminant analysis was conducted on the combined *n graph derivation* and *model derivation* sets (*n* =1034).

Experiment 1

For Experiment 1, the accuracy of categorization for GPAT and Coh-Metrix were compared against the test set data from TASA7-12 (*n* = 188). The results for GPAT were encouraging, moderately outperforming Coh-Metrix (see Table 1). The results suggest that GPAT is at least as accurate as a combination of 30 discourse variables.

Experiment 2

In Experiment 1, the GPAT results compared favorably (93.5%) to the Coh-Metrix model (89.4%). However, a robust genre purity index needs to be extendable (i.e., able to provide accurate results of text analysis well beyond its original data source). As such, the Experiment 1 models (GPAT and Coh-

Metrix) were retested using 400 new randomly selected TASA texts, all of which were *below* the grade 7 level used to create the models (i.e. grades 1-6). Although these new texts (*TASA 1 6*) are also from TASA, the lower DRP values indicate that the texts are for a very different audience.

The results of Experiment 2 were again encouraging (see Table 2). For GPAT, 368 of the 400 texts were correctly allocated to their respective genres. Thus, the accuracy of the model for GPAT on TASA1-6 was 92.3% (compare with GPAT for TASA7-12 at 93.5%). For the Coh-Metrix model a slightly lower figure of 358 texts were correctly allocated. Thus, the accuracy of the Coh-Metrix model for TASA1-6 was the same as in Experiment 1 (89.5%).

Summary of Experiments 1 and 2

The results of Experiments 1 and 2 establish that GPAT is at least as accurate as a combined Coh-Metrix model, which features 30 powerful indices. In Experiments 3, 4, and 5, GPAT is further tested for extendibility; first, by going beyond science and narrative to include history texts; and second, by examining sentence and sub-sentence level texts: lengths that features such as cohesion cannot assess because multiple sentences are required for overlap evaluations.

Table 1. Recall, Precision, and F1 for GPAT and Coh-Metrix on Test Set Data

Model		Recall	Precision	F1
GPAT	Narrative	0.951	0.914	0.932
	Science	0.920	0.955	0.937
Coh-Metrix	Narrative	0.905	0.887	0.896
	Science	0.882	0.901	0.891

Table 2. Recall, Precision, and F1 for GPAT and Coh-Metrix on TASA1-6 data

Model		Recall	Precision	F1
GPAT	Narrative	0.909	0.919	0.914
	Science	0.930	0.921	0.925
Coh-Metrix	Narrative	0.936	0.854	0.893
	Science	0.859	0.938	0.897

Experiment 3

The GPAT results for both TASA7-12 and TASA1-6 were impressive. However, to further demonstrate the accuracy and extendibility of the approach, Experiment 3 features the analysis of 150 new texts (50 * science; 50 * narrative; and 50 * history) derived from the MetaMetrics corpus (Duran et al. 2007). The texts in the Duran and colleagues corpus (hereafter *D corp*) are all from high school text books. The average length of texts in words is 409.878 (SD = 17.687), making them considerably longer than the TASA7-12 corpus M = 282.439, SD = 26.069). Thus, the difference in source and length provides a stern test for the robustness of GPAT.

D-corp is of particular interest because it includes history texts, thus providing a further test of GPAT. Because history texts are expository (inasmuch as the text tends to relate factual information) but also narrative (inasmuch as the texts relate events that are typically in chronological order; see Duran et al. 2007) we predicted that GPAT results for the history texts would be split between science and narrative categorization. However, we predicted an unequal split favoring narratives because both the studies of Duran and colleagues' and McCarthy et al. (2009) reported that history texts were more narrative in nature than science.

The results for GPAT on D-corp were once more encouraging. Of the 50 narrative texts, 49 (98%) were correctly assigned; of the 50 science texts, 44 (88%) were correctly assigned. In total, 93% of the texts (narrative/science) were correctly assigned, a result very much in line with TASA7-12 (93.5%) and TASA1-6 (92.3%). To better understand the GPAT errors (7 cases), the individual *genre prediction evaluations* were assessed. These evaluations were based on the coefficients generated from the discriminant analysis performed on the TASA data. For a discriminant analysis, there are two evaluations, one for narrative and one for science; and, following typical discriminant analysis procedure, the higher evaluation is deemed to be the genre to which the text belongs. We can presume that the greater the difference in the evaluation (between the genres), the more *confidence* we can have in the genre prediction. The confidence can be said to occur because higher values can only stem from higher incidence of genremes. Similarly, when the difference between the two evaluations is low, we can argue that the confidence in the evaluation is low. The average difference in the evaluations between *correct* narrative and science assessment was 4.690 (SD = 1.482), whereas the one misaligned text had a value difference of just 0.262, meaning that the one erroneous assignment was low in confidence. The average difference between correct science assessment and narrative assessments was similar to the narrative genre difference, 4.538 (SD = 2.060). The six misaligned texts had an average value difference of 1.232, well below the average for correct assessment, the largest difference being 2.730. As such, the argument is that these errors in assignment were low in confidence. Thus, the accuracy of the findings using D-corp provides more compelling evidence for GPAT. In addition, the low value differences for the misaligned texts suggest that a confidence value could be a useful addition to the index such that larger differences indicate greater confidence.

Turning to the history texts, the results were largely as predicted. Of the 50 texts, 35 were assigned as narrative and

15 were assigned as science. In terms of confidence values, the average absolute difference between narrative and science assessments was 1.550 (SD = 1.143), well below the average differences for science and narrative texts, suggesting that where the texts were aligned to one of the two genres, the confidence levels were low. Taken as a whole, the results for D-corp provide evidence that GPAT can assess and distinguish science and narrative texts and that meaningful results can also be provided for text types such as history that fall somewhere between narratives and science.

Experiment 4

Because GPAT only requires 4-character strings, it can assess texts as short as sentences (or sub-sentences). Naturally, cohesion indices (such as overlap indices) cannot assess single sentence length text because a minimum of two sentences are required for overlap to be evaluated. Thus, if GPAT can accurately assess sentence length texts then it has a further advantage over comparable approaches.

To assess sentence length text, we used a corpus of 210 sentences (science, narrative, history) used in McCarthy et al. (2009). The corpus (hereafter M-corp) comprises 210 sentences, equally representing the genres of science, history, and narrative (average sentence length in terms of number of words = 15.437; SD = 7.113). Our predictions for history texts were the same as those for Experiment 3: a split between narrative and science that favors narrative.

The results were in line with our predictions. Of the 70 narrative sentences, 58 (82.9%) were correctly assigned. Of the 70 science sentences, 56 (80.0%) were correctly assigned. And of the 70 history sentences, 48 (68.6%) were assigned to narrative and 22 (31.4%) were assigned to science. Considering that GPAT was trained on considerably longer texts (TASA7-12; mean length words = 282.439, SD = 26.068), where even the sentence length average differs considerably (mean length words = 20.428, SD = 7.856), the accuracy of the results here are impressive. The results of the history sentence analysis are very similar to that of the full text examples from D-corp in Experiment 3 with the full history texts providing 42.9% science assignments and the single sentence version providing 45.8%.

Experiment 5

Although the results of M-corp in Experiment 4 are impressive, McCarthy et al. (2009) report that participants in their experiment were able to correctly identify genre using as few as *the first three words of sentences* to an accuracy of 80%. As such, a further GPAT analysis was conducted using the first three words of each sentence from M-corp. These sentence fragments are referred to as M-frag.

Of the 210 *fragments* (i.e., 3-word sub-sentences) in M-frag, 70 items (33%) registered 0% genremes for science and for narrative, and so could not be assigned to any genre (narrative = 26; history = 27; science = 17). Of the remaining fragments, for narratives, 38 out of 54 (86.364%) were correctly assigned; and for science, 34 out of 53 (64.151%) were correctly assigned. Six narrative fragments were incorrectly assigned to science and a 19 science fragments were incorrectly assigned to narrative. As predicted, the history results produced more fragments assigned to narrative (*n* = 31) than to science (*n* = 12).

Discussion

The findings of this study demonstrate that GPAT is simple yet highly effective genre evaluation tool. GPAT provides instantaneous evaluations of genre for texts with an accuracy at least at high as a combination of 30 cohesion, frequency, and syntax variables. Furthermore, GPAT is able to maintain such accuracy across a variety of text sources, and text lengths, including sentence and sub-sentential levels.

Although the accuracy of GPAT is noteworthy, readers are once again reminded that genre categorization is not in and of itself problematic. Rather, the goal is to provide a computationally inexpensive assessment of *genre purity* for texts (i.e., degree of narrative and degree of expository.) Thus, the results of these five experiments serve to provide confidence in the GPAT purity values because those values are the basis for the categorizations, and those categorizations are at least as accurate as a combination of 30 leading Coh-Metrix indices.

Taken as a whole, the results reported here suggest that GPAT provides a fast and highly accurate assessment system that is of value to discourse psychologists, psycholinguistics, and any researchers for whom the genre of texts is a component of the analysis. Future GPAT research will assess potential benefits of shorter and longer SIF n-graphs, and will also broaden the genre evaluation system into various registers such as conversational speech, essay types, and legal language. In addition, human evaluations of the genremes identified for GPAT will be assessed to establish whether (and the degree to which) the character sequences have recognizable values. Thus, although much remains to be done, this study benefits cognitive science research and artificial intelligence approaches by demonstrating a tool that can evaluate the genre purity of texts with a very high degree of accuracy, and across a wide range of texts lengths.

Acknowledgement

This research was supported in part by the Institute for Education Sciences (Grants R305GA080589, R305G020018 02, and R305G040046), by the National Science Foundation (Grant IIS 0735682), and by the Department of Defense/CIFA (H9C104 07 C0019). The author acknowledges the contribution made to this project by Danielle S. McNamara, Hyunsoon C. Min, and Nick D. Duran.

References

Bhatia, V. 1997. Applied genre analysis and ESP. In *Functional Approaches to Written Text: Classroom Applications*, ed. Tom Miller. Washington, DC: USIA.

Conway, M. 2008. Mining a corpus of biographical texts using keywords. Literary and Linguistic Computing. 28:327–43.

Downs, W. 1998. *Language and Society*. Cambridge University Press.

Min H.C. and McCarthy, P.M. 2010. Identifying varietals in the discourse of American and Korean scientists. In C. Murray & H. W. Guesgen (Eds.), *Proceedings of the 23rd Annual Conference of the Florida Artificial Intelligence Society*. Menlo Park, California: AAAI Press.

Duran, N.D., McCarthy, P.M., Graesser, A.C., and McNamara. D.S. 2007. Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods* 39: 212-223.

Gerrig, R. 1993. *Experiencing narrative worlds: On the psychological activities of reading*. Cambridge, MA: MIT Press.

Graesser, A.C., Hauft-Smith, K., Cohen, A.D., and Pyles, L.D. 1980. Advanced outlines, familiarity, text genre,and retention of prose. *Journal of ExperimentalEducation*: 48, 209-220.

Graesser, A.C., Hoffman, N.L., and Clark, L.F. 1980. Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior* 19: 131-151.

Graesser, A.C., Olde, B. A., and Klettke, B. 2002. How does the mind construct and represent stories? In M. Green, J. Strange, and T. Brock (Eds.), *Narrative Impact: Social and Cognitive Foundations*. Mahwah, NJ: Erlbaum.

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

Hendersen D.J.O and Clark H. 2007. Retelling narratives as fiction or nonfiction. In D. S. McNamara and G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society* (pp. 353-358). Cognitive Science Society.

Homes D. and Forsyth, R. 1995. The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing* 10: 111-127.

Hymes, D. 1972. Models of interaction of language and social life. In *Directions of Sociolinguistics: The Ethnography of Communication* (Eds.) J.J. Gumperz & D. Hymes. New York: Holt, Rinehart and Winston.

Jurafsky, D., and Martin, H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 2nd Ed. Prentice-Hall.

McCarthy, P.M., Myers, J.C., Briner, S.W., Graesser, A.C., and McNamara, D.S. 2009. A psychological and computational study of sub-sentential genre recognition. *Journal for Language Technology and Computational Linguistics* 24: 23-55.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M. and Graesser, A.C. in press. Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*.

Otero, J., Leon, J.A., and Graesser, A.C. (Eds.), 2002. *The psychology of science text comprehension.* Mahwah, NJ: Erlbaum.

Radvansky, G. A., Zwaan, R. A., Curiel, J. M., and Copeland, D. E. 2001. Situation models and aging. *Psychology & Aging* 16: 145-160.

Rouet, J.F., Levonen, J., Dillon, A.P. and Spiro, R.J. (eds.) 1996. *Hypertext and cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Trabasso, T., and Bartolone, J. 2003. Story understanding and counterfactual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 904-923.

Van Dijk, T. A., and Kintsch, W. 1983. *Strategies of discourse comprehension*. New York: Academic Press.

Witten, I.H., and Frank, E. 2005. *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Zwaan, R.A. 1993. *Aspects of literary comprehension. Amsterdam*. John Benjamins.

Zwaan, R.A., Magliano, J.P., and Graesser, A.C. 1995. Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21: 386-397.