Running head: COMPUTER UNDERSTANDING

A Computer's Understanding of Literature

Arthur C. Graesser, Nia Dowell, and Cristian Moldovan

University of Memphis

Send correspondence to:

Art Graesser
Psychology Department
202 Psychology Building
University of Memphis
Memphis, TN, 38152-3230
901-678-4857
901-678-2579 (fax)
a-graesser@memphis.edu

Abstract

Everyone agrees that a computer could never understand and appreciate literature, but the fields of computational linguistics and discourse processing have made important advances in automatic detection of language and discourse characteristics. We have analyzed literary texts and political speeches with two computer tools, namely Coh-Metrix and Linguistic Inquiry Word Count (LIWC). Coh-Metrix provides hundreds of measures that funnel into 5 principal components: word concreteness, syntactic simplicity, referential cohesion, deep cohesion, and narrativity. LIWC classifies words on 80 categories, such as first person pronouns, negative emotions, and social words. This paper illustrates how computer tools can unveil new insights about literature and can empirically test claims by literary scholars and social scientists. Our approach offers a computational science of literature.

.

A Computer's Understanding of Literature

The vision of a computer understanding literature would be blasphemous in most circles. No one seriously believes that a computer could understand Shakespeare. The intuitions of most people, both in and outside academia, are that computers are too rigid, formulaic, and cold to pick up the subtle nuances, emotions, and creative brilliance manifested in literature. In today's literary circles, the formalists and structuralists are quite out of vogue after their systematic attempts to extract literary interpretations from the linguistic features and explicit text ideas. They have been pummeled for over five decades by the reader response theorists (Tompkins, 1980), constructivists (Schmidt, 1982), and other literary movements that emphasize how meaning grows in the reader's mind through processes beyond the words and within a sociocultural context. There also is a cottage industry of philosophers who have thoroughly dismissed computer systems as lacking the intelligence, knowledge, experience, flexibility, and body to adequately extract meaning (Searle, 1980).

We nevertheless argue in this article that much can be learned about literature by using recently developed computer tools. During the last two decades there have been revolutionary advances in computational linguistics (Jurafsky & Martin, 2008), corpus linguistics (Biber, 1988), discourse processes (Graesser, Gernsbacher, & Goldman, 2003; Kintsch, 1998), and statistical representations of world knowledge and meaning (Landauer, McNamara, Dennis, & Kintsch, 2007). These advances empower researchers to quickly analyze large bodies of literary texts on many characteristics of language and discourse. Researchers can discover new characteristics of particular literary works, categories of texts, and writing styles of authors. They can empirically test claims that are explicitly or implicitly endorsed by experts in literature,

education, media, and social sciences. In essence, much can be learned from a computational science of literature.

Perhaps the best way to illustrate the value of a tool is through examples. This article will show how some recent automated text analyses have been meaningfully applied to literary texts. The hope is that other researchers will adopt these systems and perform analyses of their own in systematic scientific inquiry or satisfaction of personal curiosities. One system, *Coh-Metrix* (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, McCarthy, & Graesser, 2010; http://cohmetrix.memphis.edu) provides hundreds of measures that funnel into 5 principal components: word concreteness, syntactic simplicity, referential cohesion, deep cohesion, and narrativity (Graesser & McNamara, in press). Another system, called *Linguistic Inquiry Word Count*, LIWC (Pennebaker, Booth, & Francis, 2007) classifies words on 80 psychological categories. We also have recently developed a speech act classifier (SAC) that automatically assigns sentences or clauses to categories of speech acts, such as question, assertion, request, and expressive evaluation.

The computer tools are more sophisticated than traditional computer facilities, such as on-line lexicons, word counters, spell checkers, and syntactic analyzers. They are not the metrics of overall text difficulty, such as Flesch-Kincaid Grade Level (Klare, 1974-5) or Lexile scores (Stenner, 2006), which primarily depend on word length, word frequency, and sentence length. Instead, the tools we have in mind analyze texts on more subtle linguistic, discourse, and psychological constructs. The systems are aligned with multilevel theoretical frameworks that identify the representations, structures, strategies, and processes at multiple levels of language and discourse (Graesser & in press; Kintsch, 1998; Snow, 2002). These multilevel frameworks have typically proposed the following levels: *words*, *syntax*, the explicit *textbase*, the referential

situation model (sometimes called the mental model), the discourse *genre and rhetorical structure* (the type of discourse and its composition), and the *pragmatic communication* level (between speaker and listener, or writer and reader).

Coh-Metrix, LIWC, and the Speech Act Classifier

It is beyond the scope of this article to specify precisely how the computer tools measure or annotate the texts automatically. However, we will briefly mention some highlights of the advances and scholarship of each system.

Coh-Metrix. Coh-Metrix (Graesser et al., 1994; in press) has hundreds of measures that can be grouped into those related primarily to words, sentence structure, versus connections between sentences. The *word* measures include many categories of parts-of-speech (POS), which can be segregated into content words (e.g., nouns, verbs, adjectives, adverbs) and function words (e.g., prepositions, determiners, pronouns). We compute the *relative frequency* of these POS word categories, as well as other elements in Coh-Metrix, by counting the number of instances per 1000 words of text. Connectives comprise an important category because they are important in establishing situation model cohesion (Louwerse, 2001; Halliday & Hasan, 1976). Word frequency is the frequency (per million words) of a word appearing in a representative set of published documents. Coh-Metrix measures words on characteristics in an established Psycholinguistic Database (Coltheart, 1981), a collection of human ratings of several thousands of words along several psychological dimensions: age of acquisition, meaningfulness, concreteness, imagability, and familiarity. The semantic content of words is analyzed by semantic categories of *WordNet* (Miller et al.,1990), a lexicon with psychologically relevant features, such as the animate nouns, causal verbs, and the number of different senses of a word.

The *sentence* measures compute information load (words per sentence) and syntactic composition. A syntactic parser developed by Charniak (2000) assigns syntactic tree structures to sentences and computes different measures of syntactic complexity, such as the number of modifiers per noun-phrase, number of words before the main verb of the main clause, and passive constructions. Syntactic similarity is the similarity of syntactic structures among sentences in a paragraph.

The measures of *connections between sentences* are particularly relevant to the cohesion of the explicit textbase and the deeper situation model (Kintsch, 1998). Some metrics involve pairs of adjacent sentences in the text whereas others span all pairs of sentences within each paragraph. Referential cohesion is important for the *textbase* level of discourse. Coh-Metrix tracks different types of lexical co-reference by considering the overlap in content words and morphological stems (e.g., table/tables, run/running/runner). Indices of lexical diversity are related to cohesion because greater lexical diversity means that new words need to be encoded and integrated into the discourse context; type-token ratio is the number of unique words in a text (i.e., types) divided by the overall number of words (i.e., tokens) in the text. Coh-Metrix measures conceptual overlap between sentences by a statistical model of word meaning called *Latent Semantic Analysis* (LSA; Landauer et al., 2007). LSA is an important method of computing similarity because it considers implicit knowledge in addition to the explicit words. Connectives are important words that explicitly link events, actions, and states expressed in the text with respect to the cohesion of the situation model (Kintsch, 1998), a deeper level meaning that taps causality, intentionality, temporality, and spatiality. Coh-Metrix computes the ratio of cohesion particles (mainly connectives) to the relative frequency of semantic idea units reflected in main verbs. Two Coh-Metrix temporal measures involve a repetition score that tracks the

consistency of tense (e.g., *past* and *present*) and aspect (*perfective* and *progressive*) across a passage of text.

Graesser and McNamara (in press) recently performed a Coh-Metrix analysis on 37,520 texts in a corpus provided by Touchstone Applied Science Associates (TASA). This corpus represents the texts that a typical senior in high school would have encountered from kindergarten through 12th grade. This large corpus has served as a normative base when we scale other texts with Coh-Metrix. Moreover, the TASA researchers have classified the texts on different genres, most being in language arts, science, and social studies/history. A principal components analysis was performed on the TASA texts, based on several dozen of the Coh-Metrix measures. We discovered that a small number of dimensions accounted for an impressive 67.3% of the variability among texts. Below are the five major dimensions:

Narrativity. The extent to which the text is in the narrative genre that conveys a story, a procedure, or a sequence of episodes of actions and events with animate beings. Informational texts on unfamiliar topics are at the opposite end of the continuum.

Deep cohesion. The extent to which the ideas in the text are cohesively connected at a deeper conceptual level that signifies causality or intentionality.

Referential cohesion. The extent to which explicit words and ideas in the text are connected with each other as the text unfolds.

Syntactic simplicity. Sentences with few words and simple, familiar syntactic structures. At the opposite pole are structurally embedded sentences that require the reader to hold many words and ideas in working memory.

Word concreteness. Many content words that are concrete, meaningful, and evoke mental images as opposed to abstract words.

Linguistic Inquiry Word Count (LIWC). LIWC (Pennebaker et al., 2007) is an automated word analysis tool that reports the percentage of words in a text that are in grammatical (e.g. *articles, pronouns, prepositions*), psychological (e.g. *emotions, cognitive mechanisms, social*), or content categories (e.g. *home, occupation, religion*). For example, "crying", and "grief" are words in the *sad* category, whereas "love" and "nice" are words in the *positive emotion* category. LIWC provides roughly 80 word categories, but also groups these word categories into broader dimensions. The broader dimensions are linguistic words (e.g. pronouns, past tense), psychological constructs (e.g. causations, sadness), personal constructs (e.g. work, religion), paralinguistic dimensions (e.g. speech disfluencies), and punctuations (e.g. comma, period). The coding of these words is based on human judgments and ratings. The distribution of function words often has interesting psychological consequences. For example, first-person singular pronouns (e.g., *I, me, my*) have higher usage among women, young people, and people of lower social classes. Pronouns have been linked to psychological states such as depression and suicide across written text, natural conversations, and in published literature (Rude, Gortner, & Pennebaker, 2004).

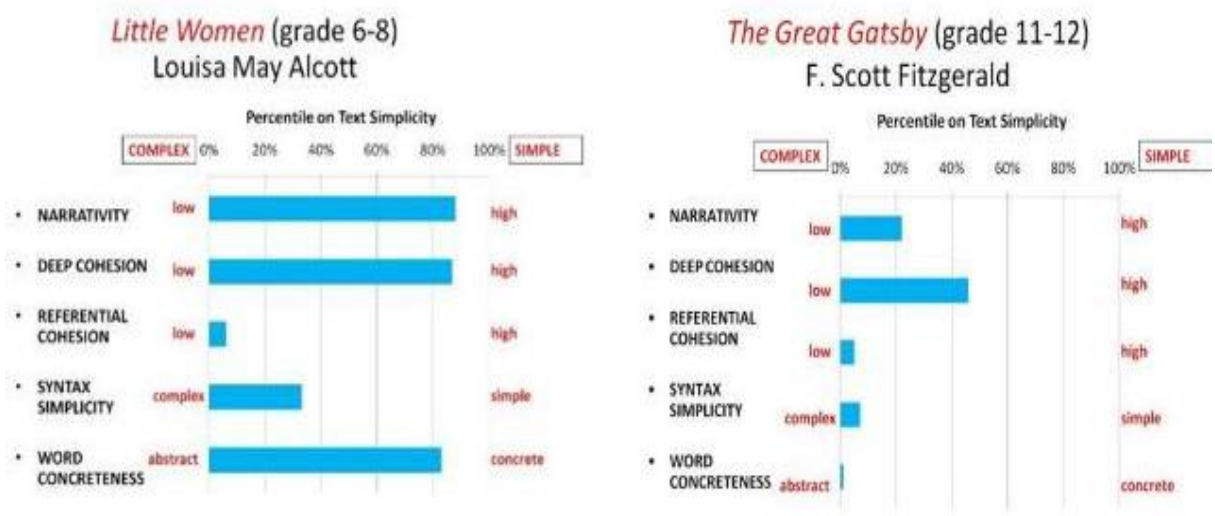
Speech Act Classifier. Utterances, sentences, and clauses are frequently classified into one or more speech act categories, such as greetings, assertions, declarations, questions, answers, direct requests, indirect requests, short responses, expressive evaluations, promises, and so on (Austin, 1962; D'Andrade & Wish, 1985). The patterns of speech act categories are expected to be important in literature, particularly when there is embedded dialogue between characters in narrative. The automatic classification of speech acts is emerging in computational linguistics (Jurafsky & Martin, 2008; Olney et al., 2003). Most of these systems attempt to classify speech acts by training classifiers with context features, syntax features, and statistical features. The

context features identify the current state of the dialogue or rhetorical structure that is set up by previous speech acts (e.g., a question is followed by an answer). Syntax features are heavily language-dependent and use the grammar rules that differentiate speech act classes, such as parts-of-speech, the order of words in the sentence, and punctuation. For instance, a question is marked by the inverse order of subject and predicate and ends with a question mark. Statistical features are discovered from machine learning algorithms that sometimes, but not always, are aligned with theory.

Examples in Drama, Poetry, Literature, and Political Speeches

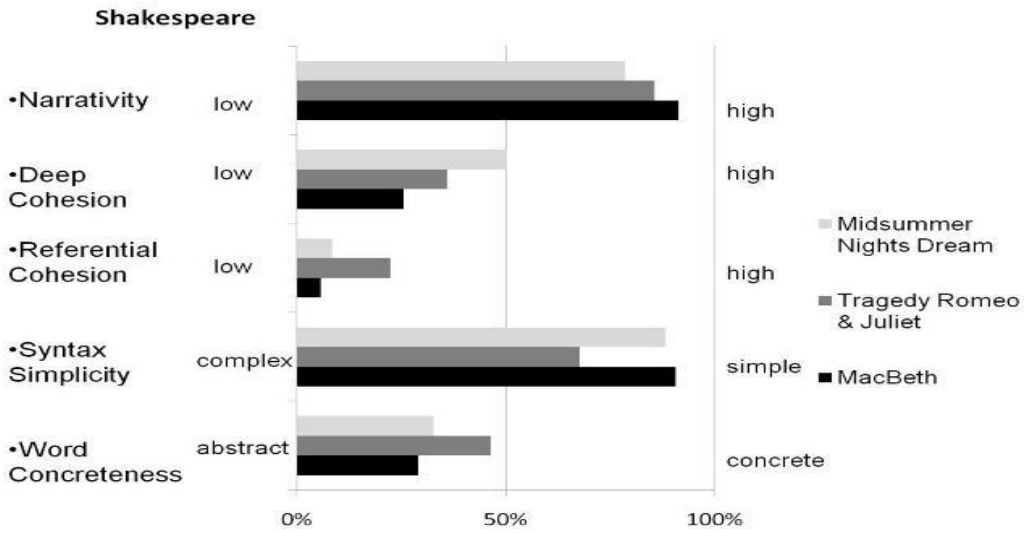
We have used Coh-Metrix to analyze the linguistic and discourse characteristics of literary texts. As an example, consider the difference between Louisa May Alcott's *Little Women* (a text in the grade band of 6-8) and F. Scott Fitzgerald's *The Great Gatsby* (in the grade band of 11-12). Figure 1 compares the texts on the 5 Coh-Metrix dimensions. The X-axis is a percentile score that varies from 0 (complex) to 100% (simple) for each of the 5 dimensions when compared with the TASA norms. All five dimensions reveal that *Little Women* is less complex than *The Great Gatsby*, which as a gratifying confirmation that the grade bands adopted by school systems make sense. *Little Women* is more narrative and cohesive, with a simpler syntax and more concrete words. It should be noted, however, that referential cohesion is not much different and leans to the complex end of the spectrum, so the picture is not entirely in harmony. Indeed, as we have inspected hundreds of texts, we find very different profiles of scores on these dimensions for literary texts, and the grade bands are frequently out of kilter with the 5 dimensions of Coh-Metrix.

Figure 1. Coh-Metrix Percentile Scores for Five Dimensions on Two Literary Texts.



Consider Shakespeare, for example. Figure 2 shows the profiles of three texts by Shakespeare. *Midsummer Night's Dream, Romeo and Juliet,* and *Macbeth* are in grade bands 6-8, 9-10, and 11-12, respectively, according to the school systems. However, this gradient in grade complexity is not at all reflected in the Coh-Metrix percentile scores. The profiles are indistinguishable among the three dramas. The language and discourse of Shakespeare is indeed Shakespeare. The differences apparently reside in the age appropriateness of the themes: the first is a comedy, the second a tragedy for adolescents, and the third a more sophisticated tragedy for adults. It is the theme that matters rather than the language and discourse features. The importance of theme is of course undeniable and has been studied in the interdisciplinary arena of IGEL (Louwerse & van Peer, 2001).

Figure 2. Coh-Metrix Percentile Scores for Five Dimensions on Three Shakespeare Dramas



Consider next the use of LIWC to analyze poetry. The claim is sometimes made by biographers that there is a link between mood disorders and artistic creativity, as in the case of poets being more depressed and suicidal than authors of other literary forms (Jamison, 1993). LIWC was used to investigate a large sample of poems written by suicidal versus non-suicidal poets (Stirman & Pennebaker, 2001). Indeed, there were systematic differences in first-person pronouns and social references associated with suicidal poets and these indicators become more pronounced over time. Pennebaker examined the works of 10 predominant British and American novelists, playwrights and poets throughout their career (Pennebaker & Stone, 2003). Previous research in this area has shown the use of first-person pronouns to be associated with depression (Rude, Gortner, & Pennebaker, 2004) and suicidal processes (Stirman & Pennebaker, 2001). An increased use of 'I' suggests that an individual is more self involved, and in the same way, more psychologically involved in their writing topic. The British and American writers had a

significant decrease in the use of I and over time individuals had a more healthy detachment from writing topics. Another trend is that there was a decrease in negative affect and neuroticism with increasing age, a finding that is compatible with research in aging (Carstensen, Pasupathi, Mayr, & Nesselroade, 2000).

The line between fiction and non-fiction is not easily discernable, especially in literary research (Barone, 2008). More recently, literary scholars embraced a new genre of literature, known as creative non-fiction. This includes biographies, history, essays, speeches, and narrative non-fiction; it recognizes that it is not necessary to tell a story to be considered meaningful contributions to literature (e.g., Martin Luther King's *I have a dream*). Political leaders often use rhetorical devices, carefully selected words, and stylistic devices to persuade the citizens.

We have investigated the language and discourse patterns of Arabic, Chinese, and Spanish leaders, such as Mubarak, Mao, and Castro. These speeches occur in multiple media, such as newspapers, speeches, and interviews. The leadership style, personality, and social status of leaders is expected to be manifested in language and discourse. It should systematically vary over time for a leader, across different cultures and languages, and for different parameters of history (economy, war, domestic uprising). We have found leaders' cohesion to decrease over time. This would be expected to the extent that there is greater common ground (shared knowledge) among people in a culture, but an alternative explanation is that aging yields a decrease in cohesion. We anticipate cohesion to increase during moments in history when the leader wants to convince the populace to adopt a new plan or agenda. We are currently investigating how the political leader's discourse aligns with key cultural and historical events. The LIWC and speech act profiles are also expected to be systematically related to the speeches of leaders.

In closing, we believe that there is so much to be learned from computer analyses of literature. Computers may never understand and fully appreciate Shakespeare. But humans don't either. Meanwhile we can learn from computer analyses just as we learn from the insights of literary scholars. A computational science of literature is a worthy player in the interdisciplinary arena.

Acknowledgements

This research was supported by the National Science Foundation (BCS 0904909, DRK-12-0918409), the Institute of Education Sciences (R305G020018, R305A080589), The Gates Foundation, and U.S. Department of Homeland Security (Z934002/UTAA08-063). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies.

References

Austin, J.L. (1962). How to do things with words. Oxford University Press.

Barone, T. (2008). *Creative nonfiction and social research*. In A. Cole & G. Knowles (Eds.), *Handbook of the arts in social science research*. Thousand Oaks, CA: SAGE.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.

Carstensen, L. L., Pasupathi, M., Mayr, U., & Nesselroade, J. (2000). Emotion experience in everyday life across the adult life span. *Journal of Personality and Social Psychology, 79,* 644–655.

Charniak, E. (2000). A maximum-entropy-inspired parser. *Proceedings of the First Conference on North American Chapter of the Association for Computational Linguistics* (pp. 132-139). San Francisco: Morgan Kaufmann Publishers.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A,* 497-505.

D'Andrade, R., & Wish, M. (1985). Speech act theory in quantitative research on interpersonal behavior. *Discourse Processes, 8*, 229-259.

Graesser, A. C., Gernsbacher, M. A., & Goldman, S. (Eds.). (2003). *Handbook of discourse processes*. ?Mahwah, NJ: Erlbaum.

Graesser, A.C., & McNamara, D.S. (in press). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*.

Graesser, A.C., & McNamara, D.S. (in press). Technologies that support reading comprehension. In C. Dede and J. Richards (Eds.), *Digital teaching platforms*. Cambridge, MA: Harvard University Press.

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

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Jamison K. (1993). *Touched with fire: manic-depressive illness and the artistic temperament*. New York: Free Press.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing*. Englewood, NJ: Prentice Hall.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Klare, G. R. (1974–1975). Assessing readability. *Reading Research Quarterly, 10,* 62-102.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2008)(Eds.). *Handbook of Latent Semantic Analysis.* Mahwah, NJ: Erlbaum.

Louwerse, M.M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics, 12,* 291-315.

Louwerse, M.M., & W. Van Peer (Eds.) (2001). *Thematics: Interdisciplinary studies*. Amsterdam: John Benjamins.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*, 292-330.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, & K. J. Miller (1990). Introduction to wordnet: An on-line lexical database. *Journal of Lexiography, 3*, 235-244.

Olney, A., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A. (2003). Utterance classification in AutoTutor. In J. Burstein & C. Leacock (Eds.), *Proceedings of the*

HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing. Philadelphia: Association for Computational Linguistics.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX: LIWC.net (www.liwc.net).

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology, 85,* 291-301.

Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18,* 1121-1133.

Schmidt, S.J. (1982). *Foundations for the empirical study of literature: The components of a basic theory*. Hamburg, Germany: Helmut Buske Verlag.

Searle, J.R. (1980). Mind, brains, and computers. *Behavioral and Brain Sciences, 3*, 417-457.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND Corporation.

Stenner, A.J.(2006). *Measuring reading comprehension with the Lexile framework*. Durham, NC: Metametrics, Inc. presented at the California Comparability Symposium, October 1996. Retrieved January 30, 2006 from http://www.lexile.com/DesktopDefault.aspx?view=re.

Stirman S. W. & Pennebaker, J W. (2001). Word use in poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine, 63,* 517-522.

Tompkins, J. P. (Ed.) (1980). *Reader-response Criticism: From Formalism to Post-structuralism*. Johns Hopkins University Press.