Behavior Research Methods
2008, 40 (2), 540-545
doi: 10.3758/BRM.40.2.540

Automatic measurement of propositional idea density from part-of-speech tagging

Cati Brown
H5, San Francisco, California

Tony Snodgrass
University of Georgia, Athens, Georgia

Susan J. Kemper and Ruth Herman
University of Kansas, Lawrence, Kansas

and

Michael A. Covington
University of Georgia, Athens, Georgia

The Computerized Propositional Idea Density Rater (CPIDR, pronounced "spider") is a computer program that determines the propositional idea density (P-density) of an English text automatically on the basis of part-of-speech tags. The key idea is that propositions correspond roughly to verbs, adjectives, adverbs, prepositions, and conjunctions. After tagging the parts of speech using MontyLingua (Liu, 2004), CPIDR applies numerous rules to adjust the count, such as combining auxiliary verbs with the main verb. A "speech mode" is provided in which CPIDR rejects repetitions and a wider range of fillers. CPIDR is a user-friendly Windows .NET application distributed as open-source freeware under GPL. Tested against human raters, it agrees with the consensus of two human raters better than the team of five raters agree with each other [$r(80) = .97$ vs. $r(10) = .82$, respectively].

Computerized Propositional Idea Density Rater, third major version (CPIDR 3, pronounced "spider three") is a computer program that determines the propositional idea density of an English text automatically on the basis of part-of-speech tags.[1] As far as we know, CPIDR is the only extant software that makes this measurement.

It is well known that propositional idea density (proposition density, P-density)—in the sense of Kintsch (1974) and Turner and Greene (1977)—can be approximated by the number of verbs, adjectives, adverbs, prepositions, and conjunctions divided by the total number of words (Snowdon et al., 1996). In an earlier study (Brown, Snodgrass, Covington, Herman, & Kemper, 2007), we refined this technique and used a part-of-speech tagger plus readjustment rules to obtain accurate idea density measures. CPIDR 3 is the latest product of this research program. Tested against human raters, it agrees with them better than they agree with each other ($r = .97$ vs. .82, respectively).

Implementation

CPIDR 3 runs on any Windows 2000, XP, or Vista system with Microsoft .NET Framework 2.0 installed. As input, CPIDR 3 accepts ASCII or Unicode text files or input typed on the keyboard or pasted from the Windows clipboard. Normal punctuation is expected (though not highly critical), and, in addition, ^ can be used to indicate the end of an unfinished sentence.

During initialization, CPIDR displays a splash screen giving the exact version number and date and time of compilation. For scientific integrity, research done with CPIDR should always cite the exact version number, since different versions will give slightly different proposition counts.

CPIDR includes two open-source components, MontyLingua (Liu, 2004), which performs part-of-speech tagging, and IKVM (Frijters, 2004) for Java-to-C# interoperability (needed by MontyLingua). CPIDR 3 is distributed as open-source freeware under the general public license (GPL), which it inherits from MontyLingua. A future version of CPIDR will be self-contained, not relying on MontyLingua or IKVM.

Usage

Figure 1 shows the main CPIDR screen, which is largely self-explanatory. The user can type sentences into

M. A. Covington, mc@uga.edu

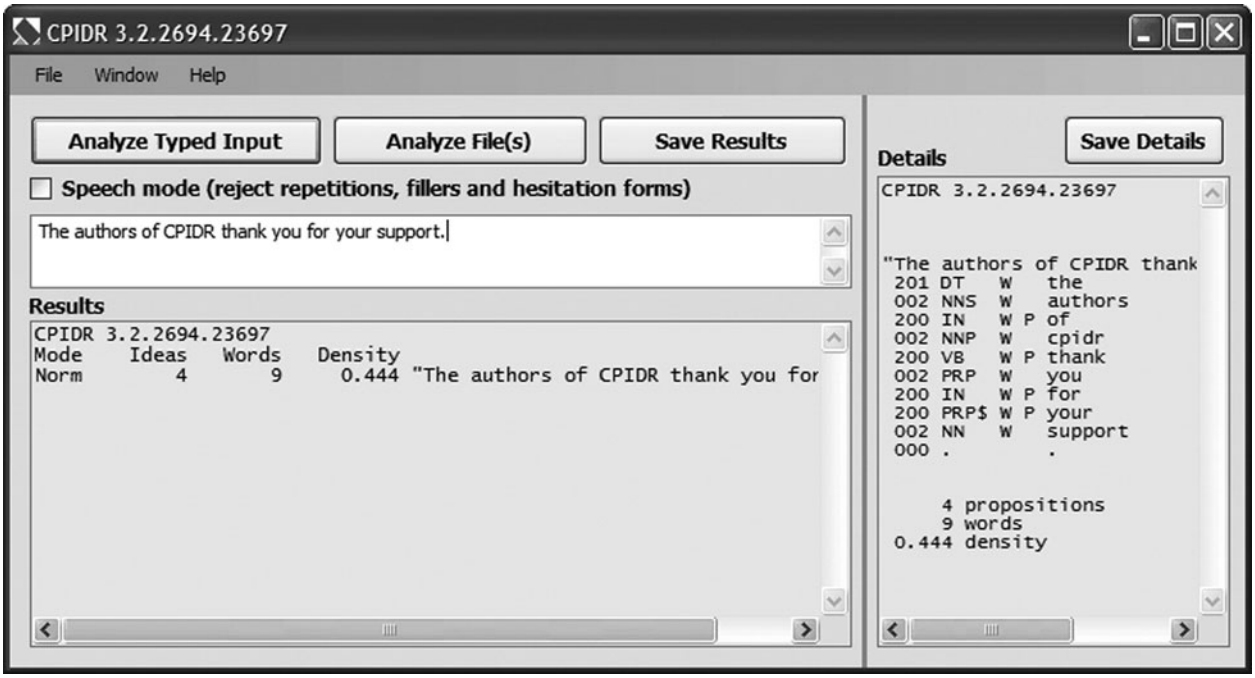

Figure 1. Main screen of the Computerized Propositional Idea Density Rater (CPIDR).

the white box or paste them from the clipboard and then choose "Analyze Typed Input." The alternative is to place the input in text files and choose "Analyze File(s)."

As Figure 1 shows, the output of CPIDR is displayed in two windows—the main results on the left and the details on the right. Each of these can be saved to a file. The details window consists of data such as:

```
"This is an example."
 054 PRP W   this
 200 VBZ W P is
 201 DT  W   an
 002 NN  W   example
 000 .       .
```

where the first column indicates which rule most recently applied to each word (054, 200, 201, etc.), the second column is the tag (PRP for pronoun; see Table 1), the third column is W if the item is a word, and the fourth column is P if the item is a proposition.

There is no limit to the length of text that can be analyzed. As consecutive files are analyzed in a single session, the results window accumulates the results in a single, concise table.

The Algorithm

Propositions. A long line of research started by Kintsch and Keenan (1973) and Kintsch (1974) assumes—with good experimental support—that propositions are the units involved in the understanding and remembering of texts. In Kintsch's system—elaborated by Turner and Greene (1977)—the main verb and all of its arguments (subject, object, indirect object, etc.) are one proposition. Additional descriptive elements, such as adjectives, adverbs, and qualifier phrases are additional propositions. Thus,

The old gray mare has a very large nose.

breaks up into:

(HAS, MARE, NOSE)
(OLD, MARE)
(GRAY, MARE)
(LARGE, NOSE)
(VERY, (LARGE, NOSE))

Each of these could be true or false separately from the others; for instance, the nose could be large, but not very large. In addition, connectives such as *and*, *if . . . then*, and *because*, and hedges such as *unfortunately*, are separate propositions.

Kintsch's propositions differ from those in logic or logical semantics in at least two ways. First, most information about verb tense, aspect, and modality is omitted from Kintsch's propositional structure so that (for instance) *Steve eats chocolate cake* and *Steve was to have eaten chocolate cake* are the same (Turner & Greene, 1977, p. 15).

Second, common nouns are not propositions in Kintsch's system. To a logician, *dog*, *brown*, and *barks* are one-place predicates, denoting the property of being a dog, the property of being brown, and the property of barking, respectively. In most human languages, however, *dog* is encoded as a common noun, which is syntactically like a name (cf. *Snoopy*) except that it can refer to any dog, not just a particular one.

Because this method of propositional analysis is now a widely used standard, we have not attempted to critique it or introduce input from newer theories of seman-

Table 1
The Main Part-of-Speech Tags Used by MontyLingua and CPIDR

Tag	Interpretation
.	sentence-ending punctuation
CC	coordinating conjunction
CD	cardinal number
DT	determiner
IN	preposition, except *to*
JJ, JJR, JJS	adjective (positive, comparative, superlative)
MD	modal verb
NN, NNS	noun (singular, plural)
PDT	predeterminer
POS	possessive *'s*
PP$, PRP$	possessive pronoun
RB, RBR, RBS	adverb (positive, comparative, superlative)
TO	to (preposition or infinitive)
VB, VBZ, VBD, VBN, VBG, VBP	verb (various forms)
WDT, WP, WPS, WRB	interrogatives and relatives (e.g., *which*)

Note—For the full set used by MontyLingua and CPIDR, see Santorini (1995).

tics. The purpose of CPIDR is simply to make the same measurements that psycholinguists have been making for a long time. Although we acknowledge the value of alternative proposals such as those of Bovair and Kieras (1985) and Perfetti and Britt (1995), we have aimed simply to replicate the counts prescribed by Turner and Greene (1977).

Idea density. Idea density is the number of expressed propositions divided by the number of words. In terms of semantics, idea density is a measure of the extent to which the speaker is making assertions (or asking questions) rather than just referring to entities.

Numerous psychological experiments have related idea density to readability (Kintsch, 1998; Kintsch & Keenan, 1973), memory (see, e.g., Thorson & Snyder, 1984), the quality of students' writing (e.g., Takao, Prothero, & Kelly, 2002), aging (Kemper, Marquis, & Thompson, 2001; Kemper & Sumner, 2001), and prediction of Alzheimer's disease. Snowdon et al. (1996) found reduced idea density in essays written by individuals who were to develop Alzheimer's disease 50 years later.

Part-of-speech tagging. Propositions correspond to certain parts of speech. Snowdon et al. (1996, p. 529) remarked in passing that each proposition is "typically a verb, adjective, adverb, or prepositional phrase" and that logical connectives between sentences are also propositions.

This idea led us to measure idea density in an earlier study (Covington et al., 2007) by counting verbs, adjectives, adverbs, prepositions, and subordinating conjunctions in the output of a computer program that identifies parts of speech (a part-of-speech tagger).

This approach was successful, but further investigation led us to refine it. In CPIDR 3, part-of-speech tagging is followed by numerous readjustment rules that adjust the proposition count. CPIDR 3 does not understand every sentence in full and therefore does not produce perfect proposition counts; however, in our tests, it agreed with the consensus of human raters better than the humans agreed with each other.

Part-of-speech tagging in CPIDR is done by MontyLingua (Liu, 2004), which uses the part-of-speech tags of the Penn Treebank (Santorini, 1995; not later versions). The most important tags are shown in Table 1.

Rules. The full set of proposition-counting rules is documented in the file *IdeaDensityRaterRules.cs*, which is installed with CPIDR 3 (in the *src* folder). This file is copiously commented so that nonprogrammers can read it; users with C# programming capability can even alter the rules. The rules have identifying numbers that are not always consecutive. In CPIDR 3.2, there are a total of 37 rules, 7 of which are specific to speech mode (see next section).

The first few rules determine which tokens should count as words and which should also count as propositions. For example, punctuation marks are not words.

Initially, every token with the tag CC, CD, DT, IN, JJ, JJR, JJS, PDT, POS, PP$, PRP$, RB, RBR, RBS, TO, VB, VBD, VBG, VBN, VBP, VBZ, WDT, WP, WPS, or WRB—that is, every conjunction, numeral, (pre)determiner, preposition, adjective, adverb, possessive, verb, relative, or interrogative—is flagged as a proposition.

Later rules adjust the proposition count and occasionally the word count. For example, *either . . . or* counts as one proposition, not two; *to verb* is one proposition, not two; and so forth. Following Turner and Greene (1977), the determiners *a*, *an*, and *the* are not propositions, and modals are not counted as propositions unless they are negative (thus, *can't* is a proposition, but *can* is not). The copula (*is*, *are*, *was*, *were*) is a proposition when it introduces a noun phrase (e.g., *is a dog*) but not an adjective phrase; that is, the copula does not add a proposition to the one already signified by the adjective.

Many of the rules condense complicated verb phrases into single propositions. For example, *may have been singing* is just one proposition (following Turner & Greene, 1977). *May not have been singing* is two propositions (*not* and *sing*), not five.

Subject–auxiliary inversion is undone in order to handle questions correctly. For example, *Has he resigned?* is changed to *he has resigned* so that subsequent rules that handle *has resigned* will apply. In the Details window, this is displayed as:

```
"Has he resigned?"
 002          has/moved
 002 PRP W    he
 402 VBZ W    has
 200 VBD W P  resigned
 000 .        ?
```

indicating the original and moved positions of *has*. In some cases, an auxiliary verb moves too far; for example, *Is he president?* is changed to *he president is*, but the proposition count is still correct.

Speech mode. CPIDR has a "speech mode"—selectable by a checkbox on the main screen—for analyzing transcripts of minimally edited speech. In speech mode, additional rules are activated to remove repeated

words from the proposition count (although not from the word count) and to reject lexical fillers more extensively. In speech mode, *like* in some contexts, and *you know* in all contexts, are considered propositionless.

Validation

Validation against Turner and Greene (1977). CPIDR 3 was designed to replicate the proposition counts given by Turner and Greene (1977, chapter 2) for their 69 examples. It does so (with speech mode turned off), with the following exceptions.

Turner and Greene's (1977) Example 17—showing coreference across three sentences—was not used, since the example sentences are not complete. In the examples in which multiple paraphrases are given (e.g., 18, 54, 55, 56), only the first version of each sentence was used.

CPIDR 3 always counts verb + preposition + noun phrase as two propositions (treating *come to/from Colorado* and *eaten by Steve* exactly like *sing in Colorado* and *eaten in Colorado*, respectively). Turner and Greene (1977) usually did the same, but they did not count *to* as a proposition in their Examples 2 (*Fred went to Boulder*), 53 (. . . *refusing to come to the party*), and 64 (. . . *returned from work*), nor did they count passive *by*-phrases as propositions separate from the verb (18j–18k).

In Turner and Greene's (1977) Example 46 (*Jimmy ate an orange and a banana*), the MontyLingua tagger mistakenly tags *orange* as an adjective, leading CPIDR 3 to count an extra proposition.

Validation against human raters. CPIDR was tested on 80 samples of spontaneous speech that had been previously collected and analyzed into propositions by coauthors S.J.K. and R.H. (see Kemper, Schmalzried, Leedahl, Mohankumar, & Herman, 2007).

Language samples were elicited from 80 volunteers in two age groups in response to the spoken question, "What do you remember about 9/11—where were you and what were you doing that morning?" Further prompting was used as needed to elicit at least 50 utterances from each speaker.

The samples were analyzed following the procedures described by Kemper, Kynette, Rash, Sprott, and O'Brien (1989). The samples were transcribed and broken into utterances (pause-delimited units, not necessarily complete sentences). Lexical fillers, such as *and*, *you know*, *yeah*, and *well* were included in the transcript, but nonlexical fillers, such as *uh*, *umm*, and *duh* were excluded. Also excluded were utterances that repeated or echoed those of the examiner.

The final 10 sentences of each speech transcript were then selected for analysis. Each sample was transcribed by one trained coder who identified all sentences and fragments; a second coder verified the transcription.

Five different trained human raters counted propositions; working separately, each analyzed 10 transcripts and on the set of 10, their agreement exceeded $r = .81$. Then, on the full set of 80 transcripts, two coders jointly analyzed each sample to ensure consensus.

The same 80 transcripts were then analyzed by CPIDR 3.2 with speech mode turned on, and the proposition counts were compared and plotted using Microsoft Excel 2002 SP3. As Figure 2 shows, CPIDR's proposition counts correlated very closely with the consensus of two human raters ($r = .97$); CPIDR's counts were about 5%

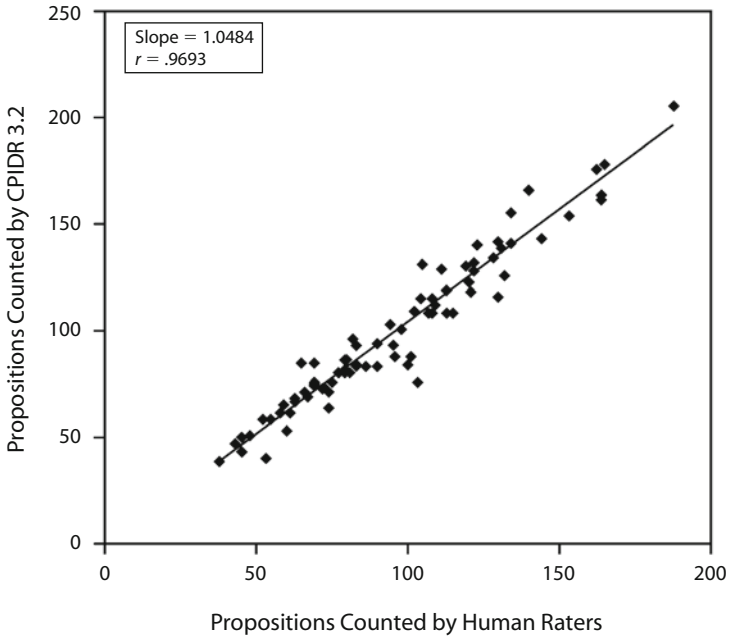

Figure 2. A comparison of proposition counts of 80 speech samples by human raters and the Computerized Propositional Idea Density Rater, version 3.2 (CPIDR 3.2).

higher. Much of the remaining inconsistency is probably attributable to the humans rather than to CPIDR.

An important property of CPIDR is that, even when in error, it is always consistent; the same sentence always gets the same rating. Thus, by using CPIDR to count propositions, an element of nonreproducibility is eliminated.

Potential Applications

Until now, almost all measurement of idea density has relied on manual raters. Rapid, reproducible automatic measurement will make existing uses of idea density more practical and will lead to new applications.

Readability and reading comprehension. Propositional density has been recognized as a major source of reading comprehension problems (Kintsch, 1998; Kintsch & Keenan, 1973); yet, efforts to assess "comprehendability" rather than "readability" have been hampered by computational challenges. Readability is typically assessed by counts of word length, sentence length, and the like (Flesch, 1948; Kincaid, Fishburne, Rogers, & Chissom, 1975). In addition, readability—in this very superficial sense—is now commonly computed by word processors and grammar checkers. Automated propositional analyses will open up the possibility of developing style guides for struggling writers as well as applications to critical domains, such as the analysis of text factors affecting health literacy, the improvement of technical documents, and the development and standardization of basal readings and standardized reading assessments (Anderson, 1982; Embretson & Gorin, 2001; Freedle & Kostin, 1991).

Aging and Alzheimer's disease. Idea density of speech and writing is well known to decline in old age, particularly in the presence of Alzheimer's disease (Kemper et al., 2001; Snowdon et al., 1996). Kemper and Sumner (2001) showed that, in a multifactorial ANOVA in language ability, idea density correlates with other measures of vocabulary and of processing efficiency (speed and fluency), but not of working memory.

Neuropsychological tests—such as the story recall test included on the Wechsler Logical Memory Scale (Wechsler, 1945)—are very sensitive to subtle cognitive deficits associated with mild cognitive impairment and the onset of Alzheimer's disease and other neuropathologies (Johnson, Storandt, & Balota, 2003; Storandt & Hill, 1989). Yet, these tests are of limited utility for broad-based screening of older adults at risk for such diseases, since both their interpretation requires extensive training to ensure reliability, and the analysis is time consuming. Automated analysis of transcribed speech can enable clinicians and researchers to perform annual screenings, community-based assessments, and epidemiological studies, and it would assist with the early detection and differential diagnosis of disabling conditions.

A potential, negative impact of the global increase in life expectancy is the aging of political leaders. British Prime Minister Ramsay MacDonald (1866–1937) most likely suffered from Alzheimer's disease, and U.S. President Ronald Reagan may have also been experiencing the early stages of Alzheimer's disease while in office (L'Etang, 1995). The speech of political leaders is widely available to the public, and computer-aided screening for subtle changes can provide an early warning of cognitive impairment.

Other uses. Other applications are also possible. Idea density is a potentially useful stylometric measurement for author identification and other forensic purposes (cf. the other measures discussed by Olsson, 2004). It is also likely to be useful for judging the informativeness of texts retrieved by search engines.

Future Refinement

Automatic replication of the proposition counts of Turner and Greene (1977) is, of course, not the last word. After developing applications for CPIDR, we can refine CPIDR in the light of them. One of the biggest questions is how much each part-of-speech tag or each CPIDR rule actually contributes to accurate measurement. For instance, neurological impairments that reduce propositional density may turn out to act mainly on verbs rather than, say, on adjectives or conjunctions (Covington et al., 2007). It may well be possible to split the proposition count per se into multiple factors that are better indicators of the things to be measured.

AUTHOR NOTE

Coauthors S.J.K. and R.H. were supported by Grant P30 DC005803 from the National Institutes of Health to the University of Kansas through the Center for Biobehavioral Neurosciences in Communication Disorders, as well as by Grant RO1 AG025906 from the National Institute on Aging. The authors' Web site is www.ai.uga.edu/caspr. We thank three anonymous reviewers for helpful comments. Correspondence concerning this article should be sent to M. A. Covington, Artificial Intelligence Center, The University of Georgia, Athens, GA 30602-7415 (e-mail: mc@uga.edu).

REFERENCES

ANDERSON, R. C. (1982). How to construct achievement tests to assess comprehension. *Review of Educational Research*, **42**, 145-170.

BOVAIR, S., & KIERAS, D. E. (1985). A guide to propositional analysis for research on technical prose. In B. K. Britton & J. B. Black (Eds.), *Understanding expository text* (pp. 315-362). Hillsdale, NJ: Erlbaum.

BROWN, C., SNODGRASS, T., COVINGTON, M. A., HERMAN, R., & KEMPER, S. J. (2007, January). *Measuring propositional idea density through part-of-speech tagging*. Poster session presented at the meeting of the Linguistic Society of America, Anaheim, CA. Available at www.ai.uga.edu/caspr.

COVINGTON, M. A., RIEDEL, W. J., BROWN, C., HE, C., MORRIS, E., WEINSTEIN, S., ET AL. (2007). Does ketamine mimic aspects of schizophrenic speech? *Journal of Psychopharmacology*, **21**, 338-346.

EMBRETSON, S. E., & GORIN, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, **38**, 343-368.

FLESCH, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32**, 221-233.

FREEDLE, R., & KOSTIN, I. (1991). *The prediction of GRE reading comprehension item difficulty for expository prose passages* (ETS Research Report No. RR–91–29). Princeton, NJ: Educational Testing Service.

FRIJTERS, J. (2004). IKVM, an implementation of Java for Mono and the .NET Framework [Computer software and documentation]. Retrieved March 27, 2008, from www.ikvm.net.

JOHNSON, D. K., STORANDT, M., & BALOTA, D. A. (2003). Discourse analysis of logical memory recall in normal aging and in dementia of the Alzheimer type. *Neuropsychology*, **17**, 82-92.

KEMPER, S., KYNETTE, D., RASH, S., SPROTT, R., & O'BRIEN, K. (1989).

Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics*, **10**, 49-66.

KEMPER, S., MARQUIS, J., & THOMPSON, M. (2001). Longitudinal change in language production: Effect of aging and dementia on grammatical complexity and propositional content. *Psychology & Aging*, **16**, 600-614.

KEMPER, S., SCHMALZRIED, R., LEEDAHL, S., MOHANKUMAR, D., & HERMAN, R. (2007). *Aging and effects of dual task demands on language production*. Unpublished manuscript.

KEMPER, S., & SUMNER, A. (2001). The structure of verbal abilities in young and older adults. *Psychology & Aging*, **16**, 312-322.

KINCAID, J. P., FISHBURNE, R. P., JR., ROGERS, R. L., & CHISSOM, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel* (Research Branch Report 8–75). Millington, TN: Naval Technical Training Command.

KINTSCH, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

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

KINTSCH, W., & KEENAN, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, **5**, 257-274.

L'ETANG, H. (1995). *Ailing leaders in power: 1914–1994*. London: Royal Society of Medicine Press.

LIU, H. (2004). MontyLingua: An end-to-end natural language processor with common sense [Computer software and documentation]. Retrieved March 27, 2008, from http://web.media.mit.edu/~hugo/montylingua.

OLSSON, J. (2004). *Forensic linguistics: An introduction to language, crime, and the law*. London: Continuum.

PERFETTI, C. A., & BRITT, M. A. (1995). Where do propositions come from? In C. A. Weaver III, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 11-34). Hillsdale, NJ: Erlbaum.

SANTORINI, B. (1995). *Part-of-speech tagging guidelines for the Penn Treebank Project* (3rd rev. ed.). Philadelphia: University of Pennsylvania. Retrieved March 27, 2008, from ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.

SNOWDON, D. A., KEMPER, S. J., MORTIMER, J. A., GREINER, L. H., WEKSTEIN, D. R., & MARKESBERY, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *JAMA*, **275**, 528-532.

STORANDT, M., & HILL, R. D. (1989). Very mild senile dementia of the Alzheimer's type: II. Psychometric test performance. *Archives of Neurology*, **46**, 383-386.

TAKAO, A. Y., PROTHERO, W. A., & KELLY, G. J. (2002). Applying argumentation analysis to assess the quality of university oceanography students' scientific writing. *Journal of Geoscience Education*, **50**, 40-48. Retrieved March 27, 2008, from www.nagt.org/files/nagt/jge/abstracts/Takao_v50n1p40.pdf.

THORSON, E., & SNYDER, R. (1984). Viewer recall of television commercials: Prediction from the propositional structure of commercial scripts. *Journal of Marketing Research*, **21**, 127-136.

TURNER, A., & GREENE, E. (1977). *The construction and use of a propositional text base* (Tech. Rep. No. 63). Boulder: University of Colorado, Institute for the Study of Intellectual Behavior.

WECHSLER, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology*, **19**, 87-95.

NOTE

1. The name CPIDR has been applied to several programs: a prototype implemented in Prolog by coauthor C.B., a Java program implemented by coauthor T.S. using a more sophisticated rule set (Brown et al., 2007), the same program ported to C# by the same author and using the same rule set (CPIDR 2), and the current program, coded in C# by coauthor M.A.C. and using a considerably revised rule set (CPIDR 3).

(Manuscript received August 13, 2007;
revision accepted for publication September 18, 2007.)