Measuring Propositional Idea Density through Part-of-Speech Tagging

Contact: mc@uga.edu, www.ai.uga.edu/caspr

Cati Brown *Tony Snodgrass* *Michael A. Covington*
University of Georgia

Ruth Herman *Susan J. Kemper*
University of Kansas

Overview

We present a computer program, CPIDR, for measuring propositional idea density in English text.

Propositional idea density is the amount of information (as a proposition count) divided by the number of words. It is a fundamental measurement in the study of discourse comprehension and is also a clinically useful diagnostic indicator.

Snowdon et al. (1996) found reduced idea density in the writing of Alzheimer's Disease victims 50 years before the onset of symptoms.

CPIDR (Computerized Propositional Idea Density Rater, pronounced "spider") counts propositions by counting verbs, adjectives, adverbs, prepositions, and conjunctions (as suggested by Snowdon et al. 1996). It then applies adjustment rules to make the count more accurate.

Propositions

A proposition is a piece of information, an idea or belief that can be true or false.

Proposition theory was introduced by Kintsch (1974) and has become a standard part of the methodology of applied psycholinguistics.

Unlike later formal semantics, Kintsch's theory does not count common nouns as propositions (predicates), nor does it count verb tense or modality separately from the verb itself.

Propositions in a text are normally identified by trained human raters following the handbook of Turner and Greene (1977) and measurement is subject to local and personal variation. We present a technique for counting propositions objectively through part-of-speech tagging.

How CPIDR Works

CPIDR uses the Java edition of MontyTagger (Liu, 2004) to tokenize the input text and tag the parts of speech. MontyTagger is based on the earlier Brill tagger and the Penn Treebank.

Then CPIDR adjusts its proposition count by applying a set of adjustment rules. For example, a linking verb is not counted separately from its adjective; *seems old* is counted as one proposition, not two.

CPIDR's adjustment rules are implemented as a Java program that scans the tagger's output using a 4-item moving window, deciding whether to count the last item of the four. If it is discovered that an earlier item should not have been counted, its count is decremented.

The final output consists of the proposition count, the word count, and the quotient of the two (idea density).

The adjustment rules in CPIDR were constructed to handle all the example sentences in Turner and Greene (1977) and obvious generalizations of them.

Basic Algorithm

Proposition Count =
- \+ Verbs
- \+ Adjectives
- \+ Adverbs
- \+ Prepositions
- \+ Conjunctions
- \+ Determiners (except *a, an, the*)
- \+ Modals (only if negative)
- – Auxiliary verbs
- – Linking verbs

Adjustment Rules
(examples)

Do not count *the, a,* or *an* as a determiner.

Count a verb in the set *is, seems, looks, smells, becomes...* followed by an adjective or adverb as a linking verb. (Actual set is larger.)

Count a verb immediately followed by *not* as an auxiliary verb.

Count *either...or* as one conjunction, not two.

Do not count modals unless they end in *n't*.

(Etc., for a total of about 20 rules)

Example

Sentence:

I called my dad because I was scared.

Propositions:

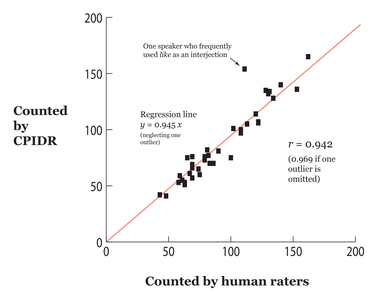
(1) called(I,dad)
(2) my(dad)
(3) because(1,4)
(4) scared(I)

Tagger output	Preliminary proposition count = 5	Adjusted proposition count = 4
I / pronoun		
called / verb	called	called
my / determiner	my	my
because / sub. conj.	because	because
I / pronoun		
was / verb	was	
scared / adjective	scared	scared

Algorithm initially counts parts of speech likely to signify propositions (verbs, adjectives, conjunctions, etc.).

Adjustment rules remove the linking verb *was* from the count.

Proposition counts of 40 text samples



Testing CPIDR

CPIDR was tested on 40 samples of spontaneous speech previously collected and analyzed into propositions by co-authors Kemper and Herman.

Language samples were elicited from 40 volunteers in response to the question, "What do you remember about the morning of 9/11?"

The final 10 sentences of each sample were analyzed for grammatical complexity as detailed by Kemper et al. (1989), then analyzed into propositions following Turner and Greene (1977). Five different human raters analyzed propositions; their agreement exceeded $r = 0.82$.

These transcripts, widely differing in length, were then analyzed with CPIDR and the proposition counts are shown in the accompanying scatterplot.

Results

CPIDR agreed with the group of human raters appreciably better than the raters agreed with each other ($r = 0.942$, or 0.969 if one outlier is excluded, vs. $r \geq 0.82$ for human vs. human).

CPIDR's proposition counts ran about 5% lower, on average, than those from the human raters (see regression line in scatterplot). This is probably due mainly to a stricter interpretation of Turner and Greene (1977). The human raters sometimes counted auxiliary and linking verbs.

One text, shown as an outlier on the graph, was greatly overcounted by CPIDR. The explanation is that this speaker frequently used *like* as an almost meaningless interjection, which the human raters correctly skipped. CPIDR, however, using a tagger trained on the Penn Treebank, treated *like* as a verb or preposition.

The regression line on the graph excludes this outlier.

What to do next

Refine the set of adjustment rules and capture all relevant linguistic generalizations

Package CPIDR as a shareable software package

Test CPIDR on other corpora that have been analyzed into propositions by human raters *(collaborators welcome!)*

Apply CPIDR to our ongoing studies of language in:
- schizophrenia
- learning disorders
- Alzheimer's disease

Factor propositional idea density into its components (verb density, adjective density, etc.) and determine the neuropsychological relevance of each

References

Kemper, S.; Kynette, D.; Rash, S.; Sprott, R.; and O'Brien, K. (1989) Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics* 10:49–66.

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

Liu, Hugo (2004) *MontyLingua: An end-to-end natural language processor with common sense.* http://web.media.mit.edu/~hugo/montylingua.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and 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.

Turner, A., and Greene, E. (1977) *The construction and use of a propositional text base.* Technical report 63, Institute for the Study of Intellectual Behavior, University of Colorado, Boulder.

This research was supported in part by grants from the National Institutes of Health to the University of Kansas through the Center for Biobehavioral Neurosciences in Communication Disorders, grant number P30 DC005803, as well as by grant RO1 AG025906 from the National Institute on Aging.

Linguistic Society of America
Anaheim, California
January 2007