Idea Density — A Potentially Informative Characteristic of Retrieved Documents

Michael A. Covington
Institute for Artificial Intelligence
The University of Georgia
Athens, GA 30602-7415 U.S.A.
mc@uga.edu

October 31, 2008

Abstract

Idea density, or number of propositions divided by number of words, is a well-known psycholinguistic measurement which can now be estimated reliably by software. Preliminary tests indicate that idea density distinguishes between documents about the same subject written for specialist and nonspecialist audiences, and that it does not correlate with lexical diversity or Flesch-Kincaid readability.

1 Introduction

The success of information retrieval is almost always judged by whether the retrieved documents match the subject of the query (Demartini and Mizzaro 2006), but the type or genre of documents retrieved is also important.

In this paper I show that idea density, a measurement with several well-known uses in psycholinguistics, is a promising criterion for distinguishing documents written for specialist vs. nonspecialist audiences. More generally, idea density may indicate the accessibility of the content of a document to a nonspecialist. Documents written for nonspecialists have appreciably lower idea density.

2 Idea density

2.1 Defined

In a long line of psycholinguistic research beginning with Kintsch and Keenan (1973) and Kintsch (1974), the *proposition* or *idea* is taken to be the basic unit involved in the understanding and retention of text. *Idea density* is the number of propositions in a text sample divided by the number of words.

A proposition, in turn, is whatever can be true or false. In *The old gray mare has a very large nose* (following Brown et al. 2008), there are five propositions:

$\langle 1 \rangle$ *has*(*mare*,*nose*)
$\langle 2 \rangle$ *old*(*mare*)
$\langle 3 \rangle$ *gray*(*mare*)
$\langle 4 \rangle$ *large*(*nose*)
$\langle 5 \rangle$ *very*($\langle 4 \rangle$)

978-1-4244-3978-2/09/$25.00 ©2009 IEEE

In Kintsch's system, unlike formal semantics, common nouns are not predicates (propositions). Further, information about verb tense, modality, or coreference is not counted as separate propositions.

Much could be said for bringing this system in line with current semantic theory, but on the other hand, Kintsch's system has been standard in psycholinguistics for decades and has been validated by a large number of experiments. It is not a theory of logic or knowledge, but rather a theory of how the human mind encodes information.

2.2 Psycholinguistic significance

Kintsch and Keenan (1973) showed that written texts with lower idea density are easier to understand, a result confirmed by a long line of subsequent work (Miller and Kintsch, 1980; Kintsch, 1998). Roughly speaking, each proposition or idea requires a certain amount of processing effort, and high idea density makes for slower processing. That is, idea density, as affecting readability, is a matter of pace; it is the rate at which material is being given to the reader to process.

On the other hand, low idea density in language production can indicate mental impairment (Covington, Riedel et al. 2007, 2008). In a study that brought idea density to the attention of clinicians, Snowdon et al. (1996) successfully predicted Alzheimer's disease from low idea density in autobiographies written 50 years before onset of symptoms.

2.3 Measurement by computer

Propositions correspond roughly to verbs, adjectives, adverbs, prepositions, and subordinating conjunctions (not nouns or pronouns). Exploiting this fact, the computer program CPIDR (Brown et al., 2008) measures the idea density of text by using a part-of-speech tagger, then counting the appropriate parts of speech and applying corrective rules to adjust the count in certain situations. For example, auxiliary verbs are removed; *either...or* becomes one conjunction rather than two; *appears* + adjective is one proposition, not two; and so forth. Following standard practice, appositive or modifying nouns are not counted as propositions. Figure 1 shows examples of high and low idea density.

CPIDR was developed using Turner and Greene (1977) as a guide and successfully replicated their results. Brown et al. (2008) report that CPIDR agrees with the consensus of a panel of trained human idea-density raters better than the raters agree with each other.

3 Experiment with retrieved documents

3.1 Method

For this experiment, fourteen documents were retrieved from the World Wide Web, all on the subject of inflation and U.S. monetary policy. Ten of these were chosen from the links returned by the Google query "predict U.S. inflation rate." In addition, four speeches or reports by Federal Reserve chairmen (Bernanke and Greenspan) were included.

Bibliographies, footnotes, section headings, tabular material, displayed formulae, displayed quotations, and the introductory and closing remarks in speeches were excluded. So were passages containing such a high proportion of displayed formulae that the English style was seriously disrupted. All decisions about choice of material and excluded sections were made before any part of the analysis had been performed. All documents were converted to UTF-8 plain text.

Idea density of each document was measured with CPIDR 3.2 (Covington 2007). Lexical diversity was measured as a average of the type-token ratio in a moving 300-word window, and Flesch-Kincaid reading level was computed by Microsoft *Word 2003*.

3.2 Results

The results are shown in Table 1. In the table, the documents are sorted in order of increasing idea density, and each has been informally characterized as “popular” (news reporting), “introductory” (for serious nonspecialists), “scholarly” (research papers), or “technical” (addressed to experienced economic decision makers rather than researchers).

By these criteria, popular or introductory documents always have an idea density below 0.5, while technical documents are always above 0.5. Indeed, the speeches of Alan Greenspan — actually used mostly in written form, and notorious for their information-packed style — are both above 0.525.

Scholarly research papers are scattered across the whole range. This probably reflects the fact that a scholarly paper that breaks new ground can and should be written as an introduction to its (new) subject, while one that continues an existing line of work is more like a technical paper.

More importantly, idea density does not correlate significantly with lexical diversity or with Flesch-Kincaid readability, nor does either of these appear to distinguish technical from nontechnical writing. In the sample, the lowest Flesch-Kincaid level and the highest are both technical documents with high idea density. The lexical diversity of one of Greenspan’s speeches is the same as that of a Bloomberg News report.

3.3 Why it works

One possible explanation of the results is the following. The idea density of a text determines the amount of work a reader must do in order to understand it (Miller and Kintsch, 1980). If an idea is familiar to the reader, it does not require as much work to process as if it were new. Accordingly, readers already familiar with a subject can comfortably process text with higher idea density than would be suitable for newcomers to the field. Thus, idea density is an indirect measure of the amount of presupposed knowledge.

4 Conclusions

The results show that idea density is a promising tool for distinguishing introductory from advanced-level treatments of subjects, and that it is distinct from lexical diversity and the Flesch-Kincaid readability index (which is based on word length and sentence length). Further investigation of its potential usefulness in information retrieval is warranted.[1]

References

Brown, Cati; Snodgrass, Tony; Kemper, Susan J.; Herman, Ruth; and Covington, Michael A. (2008) Automatic measurement of

[1] I want to thank Livia Polanyi (of *www.powerset.com*), who suggested to me in early 2007 that idea density might be useful for evaluating retrieved documents in some way. She is of course not responsible for any of the content of this paper.

High idea density:

When investors are **familiar with** the environment, they **perceive less** risk **than** they **do for objectively comparable** investment opportunities **in far distant**, **less familiar** environments.

Alan Greenspan, speech, 2005/12/02

Low idea density:

An increase **in** the factory workweek **made** the **biggest** positive contribution **to** the July **leading** indicators, **adding 0.12** percentage point.

Bloomberg News, 2006/08/17

Figure 1: Examples of high and low idea density (using boldface to indicate words counted as ideas by CPIDR).

Table 1: Comparison of 14 documents on the same subject.

Document (sorted by ascending idea density)	Genre	Idea Density	Lexical Diversity[a]	Flesch-Kincaid Reading Level[b]
Bloomberg News, "U.S. Leading Indicators" (2006/08/17)	Popular	0.434	0.572	12.9
Kitov, "Exact Prediction" (U. Munich working paper)	Scholarly	0.481	0.525	10.8
Associated Press, "Fed Revises..." (2008/02/21)	Popular	0.482	0.572	12.4
Wikipedia, "Monetary Policy"[c]	Introductory	0.485	0.531	15.5
Hyclak & Ohn, "Wage Inflation" (*Econ. Letters*)	Scholarly	0.486	0.503	17.2
USA Today, "Greenspan Predicts" (2007/09/14)	Popular	0.489	0.569	11.5
Investopedia, "Trying to Predict Interest Rates"	Introductory	0.493	0.500	12.4
Wikipedia, "Inflation"	Introductory	0.498	0.523	14.6
Bernanke, speech, 2008/01/10	Technical	0.504	0.581	16.3
Stockman, "Dollar Depreciation" (SOMC working paper)	Technical	0.509	0.479	9.3
Wright, "Forecasting U.S. Inflation" (FRB working paper)[c]	Scholarly	0.516	0.514	14.0
Bernanke, report to Congress, 2008/02/27	Technical	0.519	0.563	16.8
Greenspan, to congressional committee, 2005/06/09	Technical	0.528	0.596	15.0
Greenspan, speech, 2005/12/02	Technical	0.533	0.572	17.4
Correlation with idea density		r =	0.053	0.356
(not significant in either case)		P =	0.85	0.21

[a] Type-token ratio measured as a moving average over a 300-word window.

[b] Computed by Microsoft *Word 2003*.

[c] Only the first part of the document was analyzed; later portions contained too many formulae to be effective samples of English style.

propositional idea density from part-of-speech tagging. *Behavior Research Methods,* in press.

Covington, Michael A. (2007) *CPIDR 3 User Manual.* CASPR Research Report 2007-03, Artificial Intelligence Center, The University of Georgia. Available, with software, at *http://www.ai.uga.edu/caspr.*

Covington, Michael A.; Riedel, Wim J.; Brown, Cati; He, Congzhou; Morris, Eric; Weinstein, Sara; Semple, James; and Brown, John (2007) Does ketamine mimic aspects of schizophrenic speech? *Journal of Psychopharmacology* 21 (3) 338-346.

Covington, Michael A.; Riedel, Wim J.; Brown, Cati; He, Congzhou; Morris, Eric; Weinstein, Sara; Semple, James; and Brown, John (2008) Ketamine and schizophrenic speech: more difference than originally reported. *Journal of Psychopharmacology,* in press.

Demartini, Gianluca, and Mizzaro, Stefano (2006) A classification of IR effectiveness metrics. In M. Lalmas et al. (eds.), *ECIR 2006* (Lecture Notes in Computer Science, 3936), pp. 488–491. Berlin: Springer.

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

Kintsch, Walter A. (1998) *Comprehension.* Cambridge: Cambridge University Press.

Kintsch, Walter A., and 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.

Miller, James R., and Kintsch, Walter (1980) Readability and recall of short prose passages: a theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory* 6:335-354.

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* 275:528532.

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.