Technical Report

IBM

Relative Importance of the Components of the Cloudiness Count: Passive Voice and Infrequent Words

James R. Lewis

August 5, 1993

TR54.759

Relative Importance of the Components of the Cloudiness Count: Passive Voice and Infrequent Words

James R. Lewis
Human Factors Group
IBM Boca Raton, FL

Abstract

Previous research has shown that the Cloudiness Count, a readability measure based on the relative frequencies of passive voice and empty words (a type of infrequent word), correlates highly with the ranking of computer user publications in external user surveys. The purpose of the current research was to examine the relative importance of the two components of the Cloudiness Count. If one component tended to be more important than the other, it might be more parsimonious to consider only that component rather than the combination of the two. This result would also have implications for using readability analyses based on the Cloudiness Count to improve the readability of text. The components in combination (the Cloudiness Count) correlated significantly with ranking of publications for three different user ratings (general satisfaction with user publications, clarity of hardware publications, and clarity of software publications). In isolation, the components correlated significantly with ranking of publications only for clarity of software publications. The correlations of percentage of empty words with ranking of publications in the user surveys did not differ significantly from the correlations of percentage of passive voice with ranking of publications in the user surveys. In two cases, the correlations differed by no more than .02. The components correlated equally with the Cloudiness Count. Therefore, it is important to consider both variables in text analysis. This is consistent with psycholinguistic research that indicates that it is harder for people to extract the meaning from a passive sentence relative to its active counterpart, and that word frequency is the variable with the most influence on the speed of lexical access.

Table of Contents

Introduction

The history of the development of readability formulas shows that the most successful formulas include two components, one syntactic and one semantic (Zakaluk and Samuels, 1988). Virtually all the formulas (for example, the Fog Index and the Reading Grade Level) use sentence length to estimate syntactic difficulty and word size to estimate semantic difficulty (or, more specifically, word frequency). Certainly there is a general correspondence between sentence length and syntactic difficulty, and between word size and word frequency, but this correspondence may not be very strong. Despite the potential weakness of the correspondence, readability formulas work well for some purposes. Their predictive ability is as strong as most other psychoeducational measures (Klare, 1984). Numerous studies have shown that such readability formulas correlate with reading comprehension assessed by traditional multiple choice questions or cloze passages, oral reading errors, how many words a typist continues to type after the copy page is covered, and other similar readability measures (Fry, 1989).

Given the success of readability formulas based on sentence length and word size, it might be possible to devise different formulas that contain both a syntactic and semantic component, but to choose components that have a stronger relationship to syntactic and semantic difficulty than sentence length and word size. An alternative readability formula, the Cloudiness Count, does exactly this. The Cloudiness Count is the number of verbs in passive voice plus the number of words that are in a lexicon of "empty" words, divided by the number of words in the passage and expressed as a percentage. Psycholinguistic and human factors research consistently show that it is harder for people to extract the meaning from a passive sentence relative to its active counterpart (Broadbent, 1977; Miller, 1962). Some research indicates that, on the average, it takes people 25% longer to understand a sentence expressed in passive voice (Bailey, 1989). Psycholinguistic research also shows that the variable that most influences the speed of a reader's lexical access is the frequency with which a word appears in the language (Forster, 1990; Whaley, 1978). "Empty" words are a special type of infrequent word, also referred to as "abstract". They often appear in business and technical writing as filler words without substantial meaningful content (such as "system" and "documentation"), but appear rarely in general English speaking and writing.

Use of the Cloudiness Count in competitive evaluations has shown that it correlates highly with the ranking of computer user publications (such as quick references and user guides) in external user surveys. The first survey (Dataquest, 1991a) investigated user satisfaction with publications in general. The second survey (Dataquest, 1991b) asked respondents to rate the clarity of hardware and software publications, respectively. Researchers selected ten random text samples from user publications for seven computer systems. A computerized text analysis program calculated the Cloudiness Count, the Reading Grade Level, and the Fog Index for each sample. The detailed results of the competitive evaluation of the publications are confidential, but rank correlations between the Cloudiness Count and ranking in the external user surveys on satisfaction with publications were significant. The surveys correlated significantly with one another, and the Reading Grade Level and Fog Index

also showed significant agreement. The rank correlations between the readability measures based on sentence length and word length (Reading Grade Level and Fog Index) and ranking in the surveys were not significant. Table 1 summarizes the relevant, non-confidential results of the evaluation.

Table 1. Spearman Correlations Among Measures and Ranking in Surveys

	General Rating	Hardware Rating	Software Rating	Cloud. Count	Fog Index	Reading Grade Level
General Rating:	*1.00*	*0.88* *0.00*	*0.94* *0.00*	**0.72** *0.00*	-0.38 **0.10**	-0.38 0.46
Hardware Rating:	*0.88* *0.00*	*1.00*	*0.93* *0.00*	**0.72** **0.06**	-0.32 0.47	-0.13 0.78
Software Rating:	*0.94* *0.00*	*0.93* *0.00*	*1.00*	**0.90** **0.00**	-0.18 0.69	0.05 0.90
Cloudiness Count	**0.72** **0.10**	**0.72** **0.06**	**0.90** **0.00**	*1.00* *0.00*	0.14 0.76	0.36 0.43
Fog Index:	-0.38 0.46	-0.32 0.47	-0.18 0.69	0.14 0.76	*1.00* *0.00*	*0.89* *0.01*
Reading Grade Level:	-0.38 0.46	0.13 0.78	0.05 0.90	0.36 0.43	*0.89* *0.00*	*1.00* *0.00*

Table notes: The first number in a pair is the Spearman correlation, and the second number is the observed significance level of the correlation (n=7). Italics indicate correlations that are large but expected, such as the correlations among the Dataquest surveys and between Reading Grade Level and the Fog Index. The correlations between the Cloudiness Count and the survey rankings are in bold type.

The purpose of this paper is to examine the relative importance of the two components of the cloudiness count: percentage of passive voice and percentage of empty words. If one component is more important than the other, it would be more parsimonious to consider only that component rather than the combination of the two. This result would also have implications for using readability analyses based on the cloudiness count to improve the readability of text. For example, if the percentage of passive voice alone correlates highly with ranking in external user surveys and the percentage of empty words does not, it would be better to describe the attribute "cloudiness" as "passiveness," and to focus on eliminating passive voice rather than

empty words from publications. Alternatively, if empty words account for the "cloudiness" effect, the appropriate strategy would be to replace empty words with more frequently occurring words wherever possible in publications. If both passive voice and empty words contribute to cloudiness, writers of user publications should attempt to achieve both the goal of limiting the use of passive voice and the goal of using frequent rather than empty words.

Method

I modified the database that produced Table1 by separating each Cloudiness Count into the percentage of passive voice and the percentage of unfamiliar words, relative to the number of words in the sample. The researchers who built the database had evaluated text samples from the user publications for seven different computer systems. Following the recommendations of Bailey (1989), they used a stratified random selection procedure to select ten samples from each system's user publications (such as user's guides and quick references). Each sample contained at least 200 words.

Results

Table 2 shows the rank correlations between the survey items, the components of the Cloudiness Count, and the Cloudiness Count.

Table 2. Spearman Correlations Among Cloudiness Components and Ranking in Surveys

	General Rating	Hardware Rating	Software Rating	Cloud. Count	Percent Passives	Percent Empty Words
General Rating:	*1.00*	*0.88*	*0.94*	**0.72**	0.64	0.38
		0.00	*0.00*	*0.00*	**0.10**	0.17	0.46
Hardware Rating:	*0.88*	*1.00*	*0.93*	**0.72**	0.59	0.58
	0.00	*0.00*	*0.00*	**0.06**	0.16	0.18
Software Rating:	*0.94*	*0.93*	*1.00*	**0.90**	**0.74**	**0.76**
	0.00	*0.00*	*0.00*	**0.00**	**0.06**	**0.05**
Cloudiness Count:	**0.72**	**0.72**	**0.90**	*1.00*	*0.86*	*0.86*
	0.10	**0.06**	**0.00**	*0.00*	*0.01*	*0.01*
Percent Passives:	0.64	0.59	**0.74**	*0.86*	*1.00*	0.57
	0.17	0.16	**0.06**	*0.01*	*0.00*	0.18
Percent Empty Words:	0.38	0.58	**0.76**	*0.86*	0.57	*1.00*
	0.46	0.18	**0.05**	*0.01*	0.18	*0.00*

Table notes: The first number in a pair is the Spearman correlation, and the second number is the observed significance level of the correlation (n=7). Italics indicate correlations that are large but expected, such as the correlations among the Dataquest surveys and between the Cloudiness Count and its components. The significant correlations between the Cloudiness Count or its components and the survey rankings are in bold type.

Discussion

The results of the correlation analysis in Table 2 indicate that the two components of the Cloudiness Count are of almost equal value, and that removing either one from the cloudiness formula would have an adverse effect on the effectiveness of the Cloudiness Count. The components in combination (in other words, the Cloudiness Count) correlated significantly with ranking of publications in the external user surveys. In isolation, the components correlated significantly only with ranking software publications for user ratings of clarity. The correlations of percentage of empty words with ranking of publications in the user surveys did not differ significantly from the correlations of percentage of passive voice with ranking of publications in the user surveys. In two cases (hardware rating and software rating), the correlations differed by no more than .02. The correlations of the components with the Cloudiness Count were equal.

Therefore, it is important to consider both variables in text analysis. Writers of computer user publications should limit their use of passive voice and empty words. This strategy is consistent with psycholinguistic research that indicates that it is harder for people to extract the meaning from a passive sentence relative to its active counterpart, and that word frequency is the variable with the most influence on the speed of lexical access. In most cases, following this strategy will result in clearer text. However, the primary goal of writers must be to write clearly. If, in a writer's judgment, replacing a passive verb with an active verb or an empty word with a frequent word obscures or changes the meaning of a sentence, then he or she should not make the change.

Future research in the Cloudiness Count should focus on expanding the two components of the measure in a way consistent with psycholinguistic research. Other syntactic structures that are similar to passive voice (because the object of the verb does not follow the verb) are object-relative and object-cleft sentences. The Cloudiness Count should include counts of these structures in addition to counts of passive voice. Also, rather than tracking the occurrence of empty words, the Cloudiness Count should track the occurrence of all words that are infrequent in English, using a reference source such as the Living Word Vocabulary (Dale and Chall, 1981). Although the current Cloudiness Count is an effective readability measure, these changes could improve it.

References

Bailey, R. W. (1989). Human performance engineering: Using human factors/ ergonomics to achieve computer system usability. Englewood Cliffs, NJ: Prentice-Hall.

Broadbent, D. E. (1977). Language and ergonomics. Applied Ergonomics, 8, 15-18.

Dale, E., and O'Rourke, J. (1981). The living word vocabulary. Chicago, IL: World Book - Childcraft International.

Dataquest, Inc. (1991a). Score report: Customer satisfaction -- personal computers. Third Quarter 1990 through First Quarter 1991. San Jose, CA: Author.

Dataquest, Inc. (1991b). Score report: Customer satisfaction -- personal computers. Fourth Quarter 1990 through Third Quarter 1991. San Jose, CA: Author.

Forster, K. I. (1990). Lexical processing. In Osherson, D. N. and Lasnik, H. (Eds.), Language (pp. 95-131). Cambridge, MA: MIT Press.

Fry, E. B. (1989). Reading formulas -- maligned but valid. Journal of Reading, 32, 292-297.

Klare, G. R. (1984). Readability. In Pearson, P. D. (Ed.), Handbook of Reading Research. New York, NY: Longmans.

Miller, G. A. (1962). Some psychological studies of grammar. American Psychologist, 17, 748-762.

Whaley, C. P. (1978). Word-nonword classification time. Journal of Verbal Learning and Verbal Behavior, 17, 143-154.

Zakaluk, B. L., and Samuels, S. J. (1988). Readability: Its past, present, and future. Newark, DE: International Reading Association.