EFFECTIVENESS OF VARIOUS AUTOMATED READABILITY MEASURES FOR THE COMPETITIVE EVALUATION OF USER DOCUMENTATION

James R. Lewis
IBM Conversational Speech Solutions
Boca Raton, FL

I examined samples from a number of companys' user publications using several automated reading measures and a graphics/text ratio. The goal was to answer two questions: Were there reliable differences in writing style among the competitors? If so, were these differences related to their rank position in published surveys of user satisfaction with documentation? Of the measures included in the study, only the Cloudiness Count had any significant relationship to rank position in the surveys. A second evaluation, focused on the components of the Cloudiness Count, indicated that both of its components (passive voice and 'empty' words – a type of infrequent word) contributed equally to its effectiveness. 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.

INTRODUCTION

In 1991, two Dataquest surveys indicated that differences existed in user ratings of user documentation (such as setup and user guides) for systems manufactured and sold by various computer companies. 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. This paper describes part of the effort to improve the competitive position of IBM user documentation – the part focused on the use of readability formulas. Admittedly, these data are not recent, but they provided a rare opportunity for assessing the effectiveness of a number of readability measures for the purpose of the competitive evaluation of documentation because the Dataquest surveys provided independent data on user satisfaction for documents that were the subject of a set of readability analyses.

The history of the development of readability formulas shows that the most successful formulas include two components, one syntactic and one semantic (Collins-Thomson and Callan, 2005; Zakaluk and Samuels, 1988). Virtually all readability 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 improved 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.

For example, consider the following two sentences:

(1) The back button should be utilized to delete characters.

(2) Use the back button to delete characters.

The first sentence is cloudier than the second because it (1) has a passive structure and (2) has the word 'utilized' rather than 'used'. Research in psycholinguistics and human factors has consistently shown 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).

According to trace theory (Garrett, 1990), a passive sentence is harder to process because a reader or listener must process a trace that is in the passive version of the sentence, co-indexing the passive verb with its object (which is a noun phrase moved from its normal position following the verb). For example, consider the active sentence:

(3) The dog chased the cat.

The corresponding passive sentence is:

(4) The cat(*i*) was chased(*ti*) by the dog.

In these sentences, "cat" (tagged with (*i*) for "index") is the direct object of "chased" (tagged with (*ti*) for "trace associated with *i*"). The word "cat" appears in the normal position for a direct object in (3), but in (4), a reader must process the trace (*ti*) to recover the relationship between the verb and its object.

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). The "empty" words of the Cloudiness Count are a special type of infrequent word. 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.

The goal of this study was to answer two questions. First, were there detectable differences in writing style between IBM and its competitors in the personal computer market of 1991? Second, were any of these differences related to the rank positions of publications in the 1991 surveys?

METHOD

I used READABLE (an IBM internal document analysis tool available at the time of the study) to evaluate text samples from the publications for seven competitive products (two IBM, five IBM competitors, labeled Competitor A-E in this paper). It provided the following measures:

- o Reading Grade Level (low score is better)
- o Cloudiness Count (low score is better)
- o Flesch Index (high score is better)
- o Fog Index (low score is better)
- o British Reading Age (low score is better)
- o Kincaid Index (low score is better)

Bailey (1989) recommends that text analysts use at least five 100- to 150-word samples to estimate readability. I used a stratified random selection procedure to select ten text samples containing at least 200 words from each system's

user publications, then divided each document set into ten equal parts and used a random number table to select a random sample within each part. Next, I analyzed each sample with READABLE. For each sample I also measured the square inches of text and the square inches of graphics and calculated the graphics/text ratio.

RESULTS

Rank Position in Surveys

Table 1 shows the rank position for each document in the three surveys.

Table 1. Rank Positions in the Surveys

Product	Dataquest, 1991a	Dataquest, 1991b, HW	Dataquest, 1991b, SW
Competitor A	2	1	1
Competitor B	NA	3	2
Competitor C	3	5	4
Competitor D	1	2	3
Competitor E	4	4	5
IBM	5	6	6

Correlations among the Readability Measures

Pearson product-moment correlations among the readability results for the seven documents showed that the Reading Grade Level, Flesch Index, Fog Index, British Reading Age and Kincaid Index had high correlations (absolute value of all r>.86, p<.013). These formulas essentially measure the same thing because they all use average words per sentence and average syllables per word.

Neither the Cloudiness Count (based on frequency of occurrence of "empty" words and passive structures) nor the graphics/text ratio (G/T ratio – the ratio of page areas devoted to graphics and text) had a high correlation with the other measures or with each other (absolute value of all r<.36, p>.42).

Because the readability measures based on sentence and word length had high correlations, I conducted additional analyses using only the two most common readability measures: the Reading Grade Level and the Fog Index, along with the less-commonly-used but uncorrelated Cloudiness Count and G/T Ratio.

Analyses of Variance for Selected Readability Measures

Analyses of variance indicated significant main effects of publications for the Cloudiness Count (F(6,63)=5.5, p<.0001), Fog Index (F(6,63)=4.3, p=.001) and Reading Grade Level (F(6,63)=3.2, p<.008). Differences in the G/T Ratios among the document sets were not statistically significant (F(6,63)=1.3, p=.26). These results (see Figure 1) showed that there were detectable readability differences among the publications.

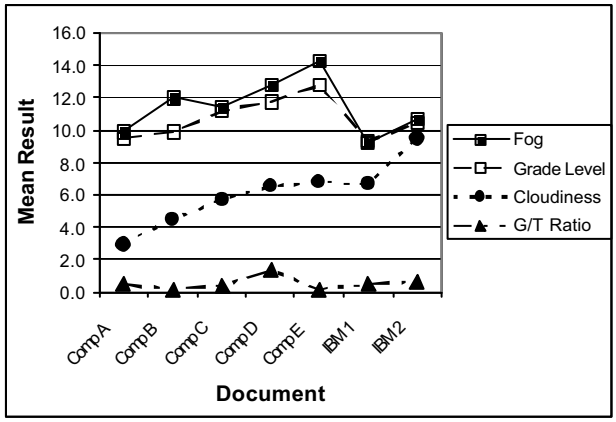

Figure 1. Publications' Automated Readability Scores

Relationship among Measures and Rank Survey Position

Because the data for these analyses included ranks, the most appropriate statistic for the analysis of correlation was the Spearman rank method. For the general survey (Dataquest, 1991a), the largest correlation was between survey position and the Cloudiness Count (r=.72, p=.10). All other correlations were much too low (given the relatively small sample size) to indicate a relationship between the variables (Reading Grade

Level and Fog Index, r=-.38, p=.46; G/T Ratio, r=-.22, p=.67).

For the hardware component of the second survey (Dataquest, 1991b), the correlation between Cloudiness Count and survey position was also .72 (p=.10), and for the software component, the correlation was .86 (p<.01). All other correlations with rank position in the surveys were nonsignificant.

Examination of the Components of the Cloudiness Count

Given the unexpected success of the Cloudiness Count, I conducted an additional analysis 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 were 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 correlated highly with ranking in external user surveys and the percentage of empty words did 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 accounted for the "cloudiness" effect, the appropriate strategy would be to replace empty words with more frequently occurring words (or simply eliminate them) wherever possible in publications. If both passive voice and empty words contributed to cloudiness, this would indicate that 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 infrequent (empty) words (supporting the construct of "cloudiness").

Spearman rank correlations between frequency of passive voice and rank position in the surveys (general, hardware, and software, respectively) were .64 (p=.17), .59 (p =.16), and .74 (p =.06), and for empty words were .38 (p =.46), .58 (p =.18), and .76 (p =.05). These results indicated that the contribution of both components of the Cloudiness Count were of about equal magnitude. The results also indicated that the combination of the components into the Cloudiness Count led to stronger correlations with survey position than either component alone.

DISCUSSION

These results answered the questions posed in the introduction. There were detectable differences in writing style among the documents. IBM publications had low Reading Grade Levels and Fog Indices (indicative of good readability due to short words and sentences), but had high Cloudiness Counts (indicative of poor readability due to too many passive structures and "empty" words). However, the only readability measure related to rank position in the survey was the Cloudiness Count. Thus, these results also support the probable effectiveness of a class of readability metrics that count passive sentence structures and use word-frequency information (such as the Cloudiness Count).

These results did not prove that cloudiness was the sole cause of the rank position in the surveys, but did indicate that cloudiness might have some influence, certainly more than Reading Grade Level, the Fog Index, or the G/T Ratio. The results of further analysis of the Cloudiness Count showed that both of its components, percentage of "empty" words and use of passive voice, contributed equally to the measurement, indicating that writers should strive to reduce their use of both of these elements as much as possible for this type of technical writing (user documentation).

It is unfortunate that there has been so little published work on the Cloudiness Count. Its inventor, Gerald Cohen, has never published a description of the measure or any empirical evaluation of its effectiveness. He has continued to promote good writing practices with his Writing Style Checker (Cohen, 2004). It isn't clear, however, whether the Writing Style Checker provides the Cloudiness Count as a readability metric (G. Cohen, personal communication, June 12, 2006). I have attempted to obtain a copy of the program (which does not appear to be commercially available), but at the time of writing this paper, I had not yet acquired it.

It is fairly easy in English to automatically detect passive voice. Any time there is a version of "to be" followed two or three words later by a past tense verb (for example, "was chased"), it's likely to be a passive sentence.

It also should be fairly easy to develop frequency measures for the words used in a sentence (for a simple example, checking words against a list of frequently-used words to develop an "infrequency" percentage – a method which is similar to the "empty word" percentage of the Cloudiness Count).

The effectiveness of the Cloudiness Count relative to the better-known readability measures suggests that additional research in the development of more passive/frequency types of readability measurements (in contrast to sentence/word length measurements) would be worthwhile.

These data are consistent with the general rule that writers should not use passive voice or infrequent words unless their use serves a specific purpose (for example, using passive voice to deliberately obscure the actor for social or political purpose, or fronting the object in a sentence to achieve a more natural flow of discourse). It is important to keep in mind, however, that the primary goal of writers (especially, technical writers) must be to write clearly. If, in a writer's judgment, replacing a passive verb with an active or intransitive verb, or replacing an empty or infrequent word with a more frequent one, obscures or changes the meaning of a sentence, then he or she should not make the change – but this should be a deliberate rather than a haphazard decision.

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.

Cohen, G. (2004). *The need for a writing style checker*. Paper presented at the 51st Annual Conference of the Society for Technical Communications, Baltimore, MD, May 9-12, 2004.

Collins-Thompson, K., and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, *56*, 1448-1462.

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.

Garrett, M. F. (1990). Sentence processing. In Osherson, D. N. and Lasnik, H. (Eds.), *Language* (pp. 133-175). Cambridge, MA: MIT Press.

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.