Writing Tools - The STYLE and DICTION Programs

L. L. Cherry

W. Vesterman

Livingston College

Rutgers University

readability

ABSTRACT

Text processing systems are now in heavy use in many companies to format documents. With many documents stored on line, it has become possible to use computers to study writing style itself and to help writers produce better written and more readable prose. The system of programs described here is an initial step toward such help. It includes programs and a data base designed to produce a stylistic profile of writing at the word and sentence level. The system measures readability, sentence and word length, sentence type, word usage, and sentence openers. It also locates common examples of wordy phrasing and bad diction. The system is useful for evaluating a document's style, locating sentences that may be difficult to read or excessively wordy, and determining a particular writer's style over several documents.

August 3, 1987

1. Introduction

Computers have become important in the document preparation process, with programs to check for spelling errors and to format documents. As the amount of text stored on line increases, it becomes feasible and attractive to study writing style and to attempt to help the writer in producing readable documents. The system of writing tools described here is a first step toward such help. The system includes programs and a data base to analyze writing style at the word and sentence level. We use the term ‘‘style’’ in this paper to describe the results of a writer’s particular choices among individual words and sentence forms. Although many judgements of style are subjective, particularly those of word choice, there are some objective measures that experts agree lead to good style. Three programs have been written to measure some of the objectively definable characteristics of writing style and to identify some commonly misused or unnecessary phrases. Although a document that conforms to the stylistic rules is not guaranteed to be coherent and readable, one that violates all of the rules is likely to be difficult or tedious to read. The program STYLE calculates readability, sentence length variability, sentence type, word usage and sentence openers at a rate of about 400 words per second on a PDP11/70 running the UNIX† Operating System. It assumes that the sentences are well-formed, i. e. that each sentence has a verb and that the subject and verb agree in number. DICTION identifies phrases that are either bad usage or unnecessarily wordy. EXPLAIN acts as a thesaurus for the phrases found by DICTION. Sections 2, 3, and 4 describe the programs; Section 5 gives the results on a cross-section of technical documents; Section 6 discusses accuracy and problems; Section 7 gives implementation details.

2. STYLE

The program STYLE reads a document and prints a summary of readability indices, sentence length and type, word usage, and sentence openers. It may also be used to locate all sentences in a document longer than a given length, of readability index higher than a given number, those containing a passive verb, or those beginning with an expletive. STYLE is based on the system for finding English word classes

---

† UNIX is a trademark of Bell Laboratories.

or parts of speech, PARTS [1]. PARTS is a set of programs that uses a small dictionary (about 350 words) and suffix rules to partially assign word classes to English text. It then uses experimentally derived rules of word order to assign word classes to all words in the text with an accuracy of about 95%. Because PARTS uses only a small dictionary and general rules, it works on text about any subject, from physics to psychology. Style measures have been built into the output phase of the programs that make up PARTS. Some of the measures are simple counters of the word classes found by PARTS; many are more complicated. For example, the verb count is the total number of verb phrases. This includes phrases like:

> has been going
> was only going
> to go

each of which each counts as one verb. Figure 1 shows the output of STYLE run on a paper by Kernighan and Mashey about the UNIX programming environment [2].

```
box; l1l.  programming environment readability grades:        (Kincaid) 12.3  (auto) 12.8  (Coleman-Liau)
11.8  (Flesch) 13.5 (46.3) sentence info:        no. sent 335 no. wds 7419        av sent leng 22.1 av word
leng  4.91        no.  questions  0  no.  imperatives  0        no.  nonfunc  wds  4362   58.8%   av  leng  6.38
        short sent (<17) 35% (118) long sent (>32)  16% (55)        longest sent 82 wds at sent 174; shortest
sent 1 wds at sent 117 sentence types:        simple  34% (114) complex  32% (108)        compound  12%
(41) compound-complex  21% (72) word usage:        verb types as % of total verbs        tobe  45% (373)
aux  16% (133) inf  14% (114)        passives as % of non-inf verbs  20% (144)        types as % of total
        prep 10.8% (804) conj 3.5% (262) adv 4.8% (354)        noun 26.7% (1983) adj 18.7% (1388) pron
5.3% (393)        nominalizations   2 % (155) sentence beginnings:        subject opener: noun (63) pron
(43) pos (0) adj (58) art (62) tot  67%        prep  12% (39) adv   9% (31)        verb   0% (1)  sub_conj
6% (20) conj   1% (5)        expletives   4% (13)
```

Figure 1

As the example shows, STYLE output is in five parts. After a brief discussion of sentences, we will describe the parts in order.

2.1. What is a sentence?

Readers of documents have little trouble deciding where the sentences end. People don’t even have to stop and think about uses of the character ‘‘.’’ in constructions like 1.25, A. J. Jones, Ph.D., i. e., or etc. . When a computer reads a document, finding