Towards a Standard for Measuring the Accuracy of Any Computer-hyphenation Program

Dwight D. Brown

No standard of measurement yet exists to measure computer hyphenation accuracy. The author discusses the need for such a standard; among considerations discussed are: word frequency, hyphenation probability, inter-word spacing, and line expansion factors. Very high hyphenation accuracy can be obtained if the computer program can select the words it chooses to hyphenate without being "chastised" for failing to hyphenate where hyphenation is possible. The author presents a series of formulas for arriving at hyphenation accuracy ratings in different publishing environments and for measuring "positive" and "negative" hyphenation errors.

Due to the widespread application of digital computers to text processing applications and the concomitant requirement for the division of words at the end of a line, a number of computer techniques for the division of words have been developed. With the advent of high-speed photo composition devices and more powerful and inclusive typesetting computer programs, high quality computer hyphenation becomes an absolute necessity.

At present there is no standard for measuring the accuracy of computer hyphenation. Selection between various techniques and the effects of changes on a particular technique cannot be effectively measured.

If one were to adopt the convention of dividing correct hyphenations by the total of attempted hyphenations to arrive at the accuracy percentage, 100% accuracy could be obtained by storing one word in the computer and not attempting to hyphenate any others.

While no one has approached this degree of statistical flagrancy, measurement figures have been quoted which did not deduct from a program's accuracy rating for failure to hyphenate

when hyphenation was indeed possible. If accuracy ratings are directed toward the measurement of a program's accuracy for typesetting, there are several things which should be taken into consideration. Among these are:

Frequency. Any attempt to measure hyphenation accuracy without taking usage frequency of the words into consideration would be misleading to a typesetter. Correct hyphenation of such words as "medium," "likely," or "picture" is much more important than the proper hyphenation of words such as "languor," "maxixe," or "epicene." Any accuracy measurement technique should take frequency of use into consideration.

Word Length as a Function of the Average Number of Characters Per Line. Consider a set of characters, all of which have the same width, as on a computer printer. Also assume that a column width equal to 40 character positions is to be set. Since over 99.98% of all words are less than 20 characters in length, let's consider only words of fewer than 20 characters. Neglecting certain fringe effects at the beginning of a line, the chances are approximately equal that a word will begin in any of the 40 character positions, except for the second one. This is, of course, for lines as presented to the computer before hyphenation or justification have been performed.

A word can never be a candidate for hyphenation unless one or more of its characters extend past the fortieth character position. Therefore, the probability of a word being a possible candidate for hyphenation can be expressed as a function of the number of characters in the word and the number of characters in the line. In a 40-character line, a word could begin in any of 39 positions (in any position except the second one). A five-character word could become a candidate for hyphenation by beginning in any of the last four positions of the line. Therefore, a five-character word would have four chances out of 39 of being a candidate for hyphenation provided it appeared once, and only once, in the line. Figure 1 shows the probability of a word being a contender for hyphenation as a function of the number of characters in the word and the number of characters in the line. The graphs in Figure 1 illustrate one reason for placing greater emphasis on longer words when measuring hyphenation accuracy.

FAMILY OF LINES PLOTTING P_c vs WORD LENGTH FOR VARIOUS LENGTH LINES

Where:

X = Number of characters in word.

C = Number of characters in line.

P_c = Probability of a word being a possible contender for hypenation. (extending over the end of the line)

This probability P_c (neglecting the case of a word appearing more than once in the line) is given by:

$P_c = (X - 1)/(C - 1)$ except where C is less than or equal to X, in which case $P_c = (X - 2)/(C - 1)$.

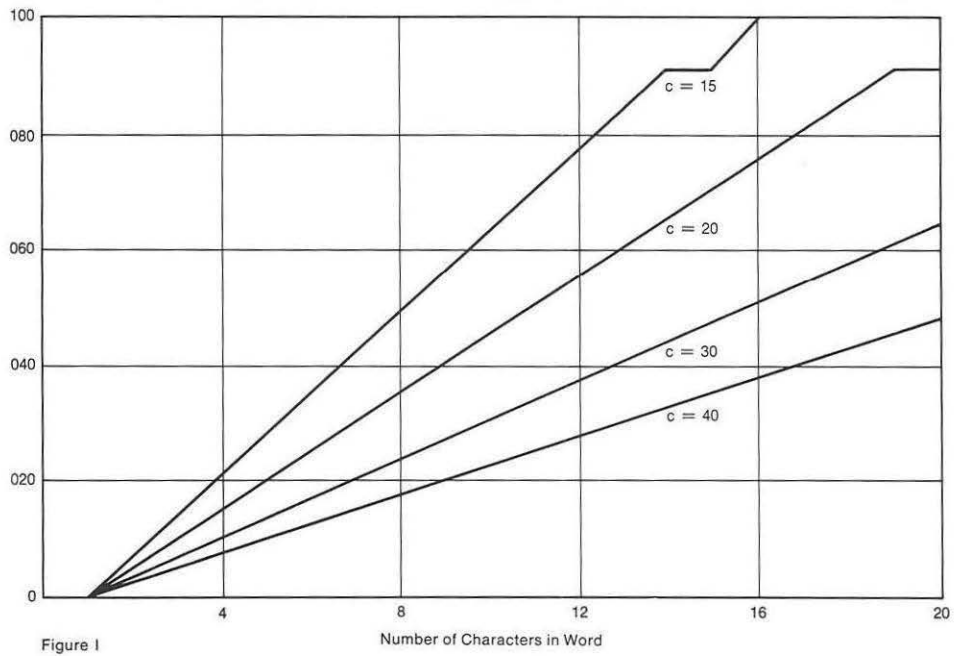

Figure I

Figure 1. Word hyphenation probability as a function of number of word characters and number of line characters.

Number of Spaces in the Lines and Spaceband Expansion. Hyphenation becomes a typesetting requirement when the inter-word space expands past some graphic quality value.

The fundamental reason for syllabification, in a typesetting application, is to avoid excessive inter-word spacing. This spacing can be expressed as the expansion of the minimum inter-word space. For example: an expansion factor of 1.5 merely states that each inter-word space must be expanded to 1.5 times its minimum value to obtain justification. Some expansion limit is normally reached before hyphenation will be attempted. That is, if the over-extending word is carried to the next line and the expansion factor is still less than the value set up (usually 1.7), no hyphenation need be attempted. This expansion factor is given by the formula

$$E = \frac{S + (X + 1)}{S} = 1 + \frac{X + 1}{S} \qquad [1]$$

E = expansion factor

S = minimum width of all spaces in the line

(X+1) =Line deficit (amount of excessive space remaining in the line after the over-extending word has been taken to the next line.) X is the minimum number of characters which must be carried to the next line to exceed E. The space which would have preceded the character group taken to the next line is accounted for by the "1" added to X.

The application of this formula results in a series of graphs (shown as Figure 2). These graphs illustrate the effect of the type of work being set on the importance of word lengths as pertains to hyphenation.

For example, a job requiring 10 spaces per line and permissible inter-word spaces of 1.7 will never require hyphenation of words fewer than seven characters in length.

Since about 63% of all words greater than four characters in length, by frequency of use, are between five and seven characters in length, an accuracy rating which included words in this length range would be meaningless to a typesetter who specializes in wide-measure work. For example, a hyphenation program could

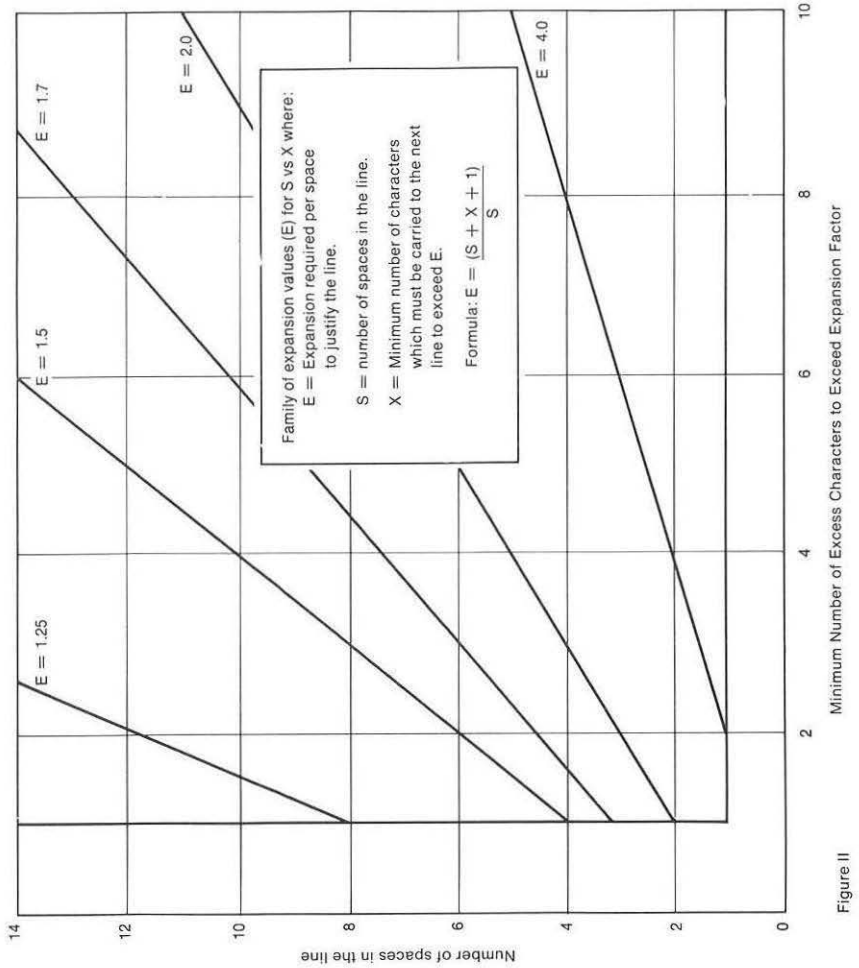

Figure II

Figure 2. The effect of the type of work being set on the importance of word lengths as pertains to hyphenation.

correctly hyphenate all the words of less than seven characters and two-thirds of the others for an accuracy rating of 87.42%, but the true accuracy rating would be 66.6%. A meaningful accuracy rating must take into consideration the type of work being set.

A combination of the technique used in generating Figure 1 and the result of Figure 2 produces the probability of the word being hyphenated as the function of the type of work being set and the length of the word. This is shown for an expansion value of 1.7 in Figure 3.

Figure 4 shows the word "accomplished" falling into the latter positions of a 40-character line. If expansion of less than 1.7 is permissible, and a column width of 40 characters is to be set, excess space equal to at least five character positions must be distributed among the spaces to equal or exceed the 1.7 limit. Therefore, "accomplished" would be hyphenated if—and only if—it began in positions 30-36. If it begins past position 36, less than five character positions will have to be distributed, resulting in an expansion factor of less than 1.7. If it begins prior to position 30, the entire word will remain on the line.

The word "accomplished" has seven chances in 39 of being hyphenated under the graphic conditions given. This is equal to a probability of about 18%.

The results shown in Figure 3 were derived by application of the formula:

$$P_h = \frac{L - X}{C - 1} \qquad [2]$$

P_h= probability of a word being hyphenated, given that it appears once and only once, in a line.

L = length of word

C = number of characters in the line

X = number of excess characters required to exceed the expansion factor (from Figure 2)

This formula holds, providing the line length is long in relation to the word length. The formula does not hold true as the word length approaches the line length.

E = Expansion factor = 1.7

$$P_h = \frac{L - X}{C - 1} = \frac{L + 1 - S(E - 1)}{C - 1}$$

WHERE: P_h = probability of a word being hyphenated, given that it appears once and only once, in a line.
L = length of word.
C = number of characters in the line.
X = number of excess characters required to exceed the expansion factor.
S = Number of spaces in the line.
E = allowable expansion factor value before hyphenation need be attempted.

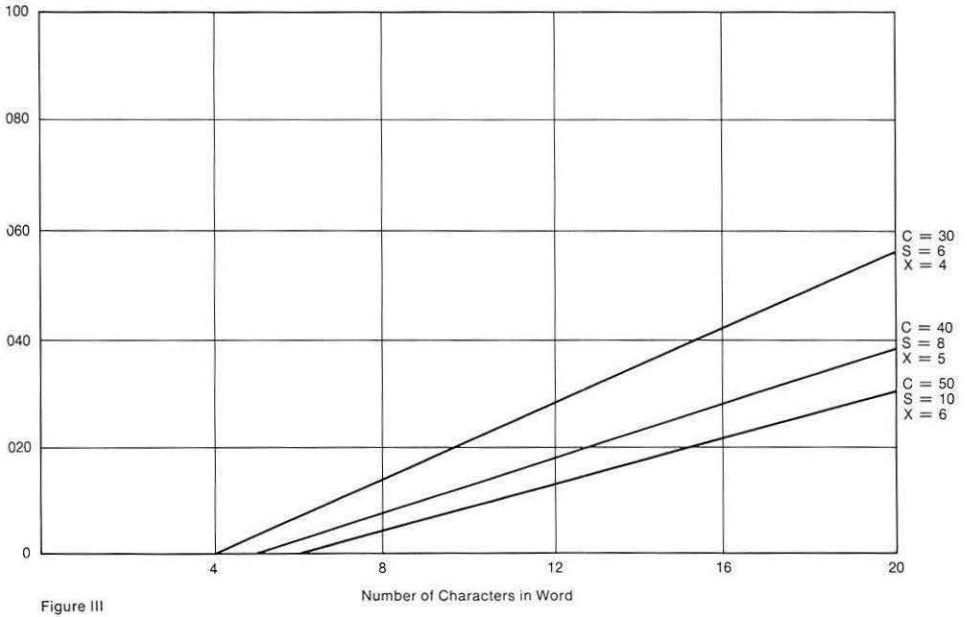

Figure III

Figure 3. Word hyphenation probability as a function of type of work being set and word length.

1	2	3	.....	29	30	31	32	33	34	35	36	37	38	39	40	(positions on 40-character line)
				A	C	C	O	M	P	L	I	S	H	E	D	
					A	C	C	O	M	P	L	I	S	H	E	D
						A	C	C	O	M	P	L	I	S	H	E D
							A	C	C	M	M	P	S	H	E H	E D
								A	C	C	P	M	L	S	H S	E D
									A	C	C	O	P	I	H S I	E D
										A	C	C	M	L	H S I L	E D
											A	C	C	P	H S I L P	E D
												A	C	C	H S I L P M	E D
													A	C	H S I L P M O	E D
														A	H S I L P M O C	E D
															H S I L P M O C C	E D
															H S I L P M O C C A	E D

Figure 4. Given an expansion factor of 1.7, the word "accomplished" has a hyphenation probability of 18% in a 40-character line.

The relationship between X and the graphic parameter E and S was given by [1]. This formula may be rearranged so that X is expressed as:

$$X = S(E - 1) - 1 \qquad [3]$$

This formula can be substituted into [2] yielding

$$P_h = \frac{L + 1 - S(E - 1)}{C - 1} \qquad [4]$$

This formula expresses the probability of a word being hyphenated as a function of its length and the parameters of the job to be performed.

The only variable, once the graphic parameters of the job have been defined, is the length of the word. This formula can easily be applied and may be expressed as a function of length plus constants.

This formula results in a weighting factor which should be applied to the word being hyphenated. The frequency of use of a word gives a fairly accurate picture of the number of times a word would occur in a given job. The probability P_h gives a representation of the number of times, per 100 occurrences, that the word will be hyphenated.

Therefore, the writer proposes that the hyphenation accuracy rating should use the frequency of use of a word times P_h (for the length word in that typesetting environment) to arrive at the weight given to the word as far as hyphenation accuracy is concerned.

Type of Hyphenation Errors

Two types of hyphenation errors commonly occur. These have been characterized as "positive" and "negative." A positive hyphenation error is said to occur when the computer program picks an incorrect point. A negative hyphenation error occurs when the computer program does not hyphenate a word even though a correct breaking point exists.

Hyphenation accuracy measurements should take both types of errors into consideration. Many newspapers feel that the positive errors are the most serious since they cause the resetting of two or

more lines. The introduction of negative errors results in poor graphic quality and excessive letterspacing, but need not require corrective measures. Very high accuracy figures are possible, with mediocre program techniques, if negative errors are not counted. It may also be possible to row, row, row your boat down rivers of white space.

Book publishers and job shop printers cannot, however, tolerate poor graphic quality. Negative errors are just as pernicious, in this environment, as positive ones.

Why Measure Accuracy More Accurately?

This article has sought to explain some of the considerations which should go into hyphenation accuracy measurements. If a standard encompassing these "on the job" considerations could be agreed upon, it would serve at least three purposes:

1. Give typesetters a meaningful yardstick with which to measure a computer program's performance.

2. Allow computer programmers to determine the effect of hyphenation program changes as they affect the actual job to be performed, rather than measuring these changes in terms of such irrelevant criteria as percentage of total words hyphenated correctly. A program change could easily increase the accuracy figure attained against a dictionary while decreasing the program's actual on-the-job accuracy.

3. Allow the "tailoring" of hyphenation (rules) to the type of job to be performed. This will allow hyphenation programs to be tailored to a specific type of publisher according to his individual needs, rather than supplying one general program trying to be all things to all people.

Toward these objectives the following method of determining the hyphenation accuracy of any computer program is proposed:

1. Obtain a representative sample of the words being set in the typesetting environment in which the program is to be used. This may be accomplished by simply collecting TTS tape, over a period of time, and then putting the words (text between spacebands) on magnetic tape or some other form of computer storage.

2. Next, sort these words, deleting duplicates but maintaining a count of the number of duplicates found for each word. This count then becomes "raw" frequency count for that word.

3. Determine the typesetting environment in terms of:
(a) Average number of characters in the lines to be set,
(b) Average number of spaces in the lines to be set,
(c) Expansion factor allowed before attempting hyphenation.

4. Substitute the values found in 3(a), 3(b), and 3(c) into formula [4] to arrive at a value of P_h for various length words.

5. Multiply the frequency count obtained in [2] by the value of P_h (depending upon the length of the word) found in (number 4). If P_h is negative, give it a value of 0.

6. This now provides a word sample which is weighted by frequency as well as typesetting environment.

Steps 1 through 6 could be performed by the typesetter wishing to obtain hyphenation accuracy information as it pertains to his particular job, or better still by some agency of the graphic arts industry such as a research council.

If an agency were to perform this function, it could collect samples of words from various classes of typesetting jobs and maintain these by types of jobs such as "medical," "news magazines," "newspaper," etc. If these magnetic tapes were kept with "raw frequency" information rather than "environmentally weighted frequency," they could be used to produce tapes for any typesetting environment. This could be accomplished by simply adding the values of those considerations mentioned in 3(a), 3(b), and 3(c) above and running this data against a program which would apply formula [4] to produce a new tape. This new tape would then contain the "environmentally weighted frequency" count.

The "environmentally weighted" tape, of the types of words to be analyzed, would form the input sample for the hyphenation program and hyphenation accuracy measurements.

Let's call this environmentally weighted frequency F_{ew}.

This master dictionary, properly hyphenated and with F_{ew} for each word, would then be hyphenated by the program technique under consideration. If the following statistics were maintained, the program accuracy, attributes, and shortcomings could be better evaluated.

Counters should be maintained for the following:

1. The number of hyphens in each master dictionary word is

multiplied by F_{ew} for that word and this quantity is then added into a counter. Let's call this counter "A."

2. The computer hyphenated word is compared to the master dictionary hyphenation and a number of points which match are counted. This number is multiplied by F_{ew} of that word and the resulting quantity is added into another counter. Let's call this counter "B."

3. The number of points in the computer hyphenated word which did not match the master dictionary hyphenation are counted. This number times F_{ew} are added into a third counter. Let's call this counter "C."

4. The number of hyphens chosen by the computer and the number of hyphens in the master dictionary word are compared.

a. If the number of hyphens in the computer hyphenated word exceeds the number in the master dictionary word, the difference times F_{ew} is placed into a counter. Let's call this counter "D."

b. If the number of hyphens in the master dictionary word exceeds the number the computer program chose, the difference times F_{ew} is added into another counter. Let's call this counter "E."

These counters, after the sample words have been hyphenated by the computer, will allow the calculation of several ratios to pinpoint a program's characteristics (Figures 5 and 6). The following ratios should be computed:

$$\text{Ratio}_1 = \frac{\text{Total B}}{\text{Total A}} \times 100\% = \text{percentage of total correct}$$

hyphens chosen (weighted by F_{ew}). This ratio, taken alone, is in no way indicative of a program's performance if the percentage is high. If it is low, it indicates poor performance but does not show why.

$$\text{Ratio}_2 = \frac{\text{Total C}}{\text{Total A}} \times 100\% = \text{percentage of positive hyphena-}$$

tion errors. This percentage may exceed 100% if the program chooses too many points.

$$\text{Ratio}_3 = \frac{\text{Total D}}{\text{Total A}} \times 100\% = \text{percentage over-hyphenation.}$$

COMPUTER HYPHENATED	MASTER DICTIONARY	F_{ew}	COUNTERS				
			A	B	C	D	E
COMPU - TER	COM - PU - TER	100	200	100	0	-	100
COM - MON	COM - MON	300	300	300	0	-	-
AC - CEPT	AC - CEPT	500	500	500	0	-	-
SUPPER	SUP - PER	100	100	0	0	-	100
A - BA - BA	A - BA - BA	10	20	20	0	-	-
MA - XI - XE	MA - XIXE	10	10	10	10	10	-
REG - IMENT	REG - I - MENT	20	40	20	0	-	20
FEB - RU - ARY	FEB - RU - ARY	300	600	600	0	-	-
TOTAL			1770	1550	10	10	220

$RATIO_1 = 87.6\% = \frac{\text{Total B}}{\text{Total A}} \times 100\%$ $\quad RATIO_3 = 0.6\% = \frac{\text{Total D}}{\text{Total A}} \times 100\%$

$RATIO_2 = 0.6\% = \frac{\text{Total C}}{\text{Total A}} \times 100\%$ $\quad RATIO_4 = 12.4\% = \frac{\text{Total E}}{\text{Total A}} \times 100\%$

Figure 5. Sample computer hyphenation program technique number 1.

Figure 6. Sample computer hyphenation program technique number 2.

COMPUTER HYPHENATED	MASTER DICTIONARY	F_{ew}	COUNTERS				
			A	B	C	D	E
C-O-M-P-U-T-E-R	COM-PU-TER	100	200	200	500	500	-
C-O-M-M-O-N	COM-MON	300	300	300	1200	1200	-
A-C-C-E-P-T	AC-CEPT	500	500	500	2000	2000	-
S-U-P-P-E-R	SUP-PER	100	100	100	400	400	-
A-B-A-B-A	A-BA-BA	10	20	20	40	40	-
M-A-X-I-X-E	MA-XIXE	10	10	10	40	40	-
R-E-G-I-M-E-N-T	REG-I-MENT	20	40	40	100	100	-
F-E-B-R-U-A-R-Y	FEB-RU-ARY	300	600	600	1500	1500	-
TOTAL			1770	1770	5780	5780	

$R_1 = 100\%$ $\quad R_3 = 326\%$

$R_2 = 326\%$ $\quad R_4 = 0\%$

$$\text{Ratio}_4 = \frac{\text{Total E} \times 100\%}{\text{Total A}} = \text{percentage under-hyphenation.}$$

Knowledge of these ratios for various hyphen techniques will allow a meaningful comparison of those techniques.

For example, a high value for both 1 and 2 indicates very poor program performance resulting from over-hyphenation. Good program performance is indicated by a high value of 1 and a low value of 2. If ratio number 1, alone, were used, as has been the case with certain measurements taken in the past, a perfect score should be obtained by simply placing hyphens after every letter in the word except the last one. This problem would have been easily pinpointed by examination of ratios 2 and 3.

One of the more prevalent problems in most hyphenation techniques will be spotted by ratio number 4. This pinpoints a program's "reluctance" to hyphenate words it is not "sure" of. Very high hyphenation accuracy can be obtained if the program can select the words it chooses to hyphenate without being "chastised" for failing to hyphenate where hyphenation is possible.

The accompanying sample word lists, showing simulated computer hyphenation as well as master dictionary hyphenation for various hyphenation technique shortcomings, should pinpoint the importance of knowing these ratios when evaluating a hyphenation program's accuracy.