Work Papers of the Summer Institute of Linguistics, University of North Dakota Session

Volume 18 Article 9

1974

Text vs. dictionary letter frequencies for primers

Kenneth D. Smith
SIL-UND

Text vs. Dictionary Letter Frequencies for Primers

Kenneth D. Smith

0. Introduction

An essential aspect of primer construction is the preparatory step of making phoneme frequency counts in search of phonemes having greatest productivity. Gudschinsky (1973.84) defines productivity as "what we can do with any given letter or syllable....The most productive letters or syllables are those which can be used in making the greatest number of useful content words (nouns, verbs, adjectives), in order to construct natural, idiomatic sentences." She emphasizes further that "it is exceedingly important to notice that which is wanted is not the absolute frequency (i.e. the exact number of times the letter occurs), but the number of different words in which each letter occurs" (p.85). With regard to the source for making frequency counts she adds "preferably, they would be taken from actual stories and essays that might be used in the primer" (p.149).

In the preparation of primers one is faced with the relative ease of making frequency counts using a dictionary as a source rather than the more tedious method of counting letters in text material where one must be careful not to count the letters of function words or repeated occurrences of content words. Theoretically a text count (especially when using texts of the type to be employed in the primer) is more effective than a dictionary because the text contains a sample of actual words that may be used in a primer story whereas the dictionary lists many words in technical semantic areas that would never be used in primer stories. The number of words in a dictionary may be an inverse measure of the dictionary"s usefulness in this regard, in that a small dictionary--say under 2,000 entries--has few technical words whereas a large dictionary--say over 4,000 entries--may be at the stage where only technical words are being entered.

FN1) With regard to the Sedang language[1] primers I have been concerned about the differences which might result from frequency counts using a dictionary rather than text material.

In Sedang phonology, it is exceedingly important to note the position of letters within the maximal bisyllabic words: $(C6)_p(C_i)\overset{R}{V}(G)(C_f)$ where (C6) represents the consonant and non-contrastive schwa vowel of the unstressed presyllable, C_i represents the initial consonant or consonant cluster C_mC_n, of the stressed main syllable, V represents the simple vowel, R represents the intersecting prosodies of register and/or nasalization, G the vowel glides, and C_f the final consonant. For example, the presyllable consonant /p/ is not assumed to be psycholinguistically equivalent for the Sedang with the initial /p/ in the main syllable, nor is either equivalent with word-final /p/. The inventory of consonants that occur in each of these positions is considered as a somewhat independent consonantal subset. And for purposes of Sedang literacy it is assumed that the consonants of each subset must be

FN2) taught separately and independently for effective teaching.[2]

At the time of primer construction I had available two dictionaries or lists of Sedang words to facilitate the finding of meaningful words for use as keywords or in syllable drills. One such list was my principal Sedang-English typed dictionary, being a listing of all known words listed in alphabetical order, except that, in order to avoid mixing nonequivalent phonemes or phonemic clusters, two-syllable words, are listed apart from one syllable words and initial consonant clusters are listed apart from words beginning with single or simple consonants. For frequency counts this dictionary provides an easy means for counting one-versus two-syllable words, as well as for counting the various initial consonants and consonant clusters since they are grouped together and the number of pages covering each letter or syllable section would be almost as reliable as counting words.

The second listing of Sedang words had been abstracted from this dictionary, but arranged in rhyming groups such that all words with a particular word final R_{-VGC} were grouped together on a separate sheet. This made a convenient source when searching for words of a given word-final type. (The alternative would have required time-consuming thumbing through the entire dictionary for each such word-final type. The time spent preparing this second rhyming list of words was less than the time that otherwise would have been spent searching for the words in the main dictionary. Furthermore, in a language with contrastive vowel laryngealization and/or nasalization, front-, central-, and back vowel glides, each vowel having a different set of final consonants due to defective distribution, the rhyming lists, which were sorted out by the Sedang language teacher, provided a safeguard to correct identification of each word-final type.) For frequency counts, these rhyming lists provide an easy means for counting vowels and final consonants, the vowel glides, laryngealization and nasalization, as well as their various combinations.

It was these dictionary-based frequency counts which were used to rank the Sedang phonemes and clusters for the initial indication of their order of introduction in the Sedang primers. (Adjustments in this order were made, of course, as primer construction proceeded as dictated generally by the semantic domain of the story sequence and corresponding word possibilities.) Since the primers are (apparently) satisfactorily completed, the following enquiry is more academic than practical except as it gives greater insight into Sedang phonemics, may have application to other languages, and may clarify a theoretical point in regard to frequency counts.

I. Enquiry

This question has bugged me: how would the ranking of Sedang phonemes in order of their frequency have differed had text material rather than a dictionary been used for the word counts?

During the spring of 1974 in a Computational Linguistics course at the University of Pennsylvania, I had opportunity to check this

```
A          329 CHAI-4       4 CHUAZ      3 DUOOH        1
AH         285 CHAI-5       2 CHUI       1 D/EENG       3
AI         231 CHAM        43 CHUIH      2 D/ENGZ       1
AIZ        290 CHAMZ       18 CHUNG      5 D/I          1
AMZ        104 CHANG        9 CHUOO     12 D/ONG        1
AP           1 CHANGZ       7 CHUOONG    6 D/OONG       1
AT           1 CHAT        10 CHU-2     26 D/OT         4
AU3          5 CHAU         3 CHU-3     47 D/RA         1
AZ         684 CHAUZ       13 CHU-5      4 D/RUN        2
A6CHE        5 CHAZ         3 CHU-6      1 D/UH         4
A6LAI        2 CHA-2       14 CHU-7      3 D/6D/O-I    10
A6LEP        1 CHA-3       22 CHU/       1 E           39
A6RAIZ       5 CHA-4       32 DAH        2 EE          74
...
...            ...            ...          ...
CHAI-3      17 CHUAT        3 DUIH       1 HENGZ        1
```

Chart 1. Sample page printout of input words with their text frequency.

```
 1 ME        OCCURS 1707 TIMES.  26 MOOI       OCCURS 193 TIMES.
 2 GAZ       OCCURS 1334 TIMES.  27 TUNG       OCCURS 189 TIMES.
 3 AZ        OCCURS  684 TIMES.  28 NEEOOZ     OCCURS 180 TIMES.
 4 KOOZ      OCCURS  593 TIMES.  29 N/AI       OCCURS 179 TIMES.
 5 VAI       OCCURS  558 TIMES.  30 PREEIZ     OCCURS 161 TIMES.
 6 TI        OCCURS  521 TIMES.  31 PANGZ      OCCURS 154 TIMES.
 7 OOH       OCCURS  478 TIMES.  32 KONG       OCCURS 151 TIMES.
 8 KI        OCCURS  404 TIMES.  33 NAH        OCCURS 151 TIMES.
 9 HIANGZ    OCCURS  376 TIMES.  34 PA-2       OCCURS 148 TIMES.
10 TA-2      OCCURS  355 TIMES.  35 HA2        OCCURS 146 TIMES.
11 KA        OCCURS  334 TIMES.  36 CHIANG     OCCURS 145 TIMES.
12 DEEI      OCCURS  333 TIMES.  37 PA-3       OCCURS 142 TIMES.
13 VA        OCCURS  333 TIMES.
14 A         OCCURS  329 TIMES.     ...
15 EH        OCCURS  306 TIMES. 196 RA         OCCURS  22 TIMES.
16 AIZ       OCCURS  290 TIMES. 197 SU         OCCURS  22 TIMES.
17 AH        OCCURS  285 TIMES. 198 TIU        OCCURS  22 TIMES.
18 PIAN      OCCURS  284 TIMES. 199 TIUZ       OCCURS  22 TIMES.
19 U         OCCURS  277 TIMES. 200 HAI-2      OCCURS  21 TIMES.
20 KHEENZ    OCCURS  234 TIMES. 201 KLEEA      OCCURS  21 TIMES.
21 KOO       OCCURS  233 TIMES. 202 KO-3       OCCURS  21 TIMES.
22 AI        OCCURS  231 TIMES. 203 K6TAU      OCCURS  21 TIMES.
23 KIA       OCCURS  231 TIMES. 204 LUA        OCCURS  21 TIMES.
24 LAI       OCCURS  204 TIMES. 205 PREEI      OCCURS  21 TIMES.
25 OOIZ      OCCURS  202 TIMES. 206 R6TEEANGZ  OCCURS  21 TIMES.
```

Chart 2. Sample printout of words in order of frequency (>20)

```
 11 WORDS HAVE A TEXT FREQUENCY OF 20.
  9 WORDS HAVE A TEXT FREQUENCY OF 19.
 10 WORDS HAVE A TEXT FREQUENCY OF 18.
 15 WORDS HAVE A TEXT FREQUENCY OF 17.
 12 WORDS HAVE A TEXT FREQUENCY OF 16.
 13 WORDS HAVE A TEXT FREQUENCY OF 15.
 19 WORDS HAVE A TEXT FREQUENCY OF 14.
 14 WORDS HAVE A TEXT FREQUENCY OF 13.
 22 WORDS HAVE A TEXT FREQUENCY OF 12.
 27 WORDS HAVE A TEXT FREQUENCY OF 11.
 26 WORDS HAVE A TEXT FREQUENCY OF 10.
 30 WORDS HAVE A TEXT FREQUENCY OF  9.
 34 WORDS HAVE A TEXT FREQUENCY OF  8.
 52 WORDS HAVE A TEXT FREQUENCY OF  7.
 59 WORDS HAVE A TEXT FREQUENCY OF  6.
 62 WORDS HAVE A TEXT FREQUENCY OF  5.
101 WORDS HAVE A TEXT FREQUENCY OF  4.
119 WORDS HAVE A TEXT FREQUENCY OF  3.
198 WORDS HAVE A TEXT FREQUENCY OF  2.
370 WORDS HAVE A TEXT FREQUENCY OF  1.
```

Chart 3. Sample printout of number of words having frequency of 20 or less

	No. of words	Total text frequency	Average text frequency
All words	1409	27,437	19
Content words	1379	15,419	11
Sight words	30	12,018	401

Chart 4. Summary of the number, frequency, and average frequency of all words, of content words, and of sight words

matter using, not just a limited text of a few pages which primer makers may endure through in their word counts, but the entire, though modest-sized, 27,437 word corpus of Sedang texts included in the Sedang word-concordance produced by the SIL-Oklahoma Concordance Project. FN3)

The specific part of the concordance which I utilized is the alphabetized list of words extracted from the input texts and total number of occurrences of each word within the entire corpus of texts, which number is printed following the citation of all such occurrences of the word. Each of the 1409 different alphabetized words in the concordance was key-punched in their computer-adapted orthography[4] FN4) onto IBM cards, each word followed by its frequency in the texts. This constituted the data input for each of the following computer programs.

FN5) The first computer program[5] produced an echo printout of the entire input list of 1409 words and their frequency in the alphabetic chart 1) sequence of the input data. See Chart 1. This provides a very compact and convenient list of words occurring in the concordance and their frequency and makes possible a check of the input dat aif question should arise.

The second computer program produced a printout of the 206 words having a text frequency greater than 20, listed in order of their text frequency, the most frequent word listed first. (This has subsequently Chart 2) become a part of the Oklahoma project.) See Chart 2. This provides an informative listing of the words which, because of their higher frequency, should definitely be included in primer stories. Further, it suggests which words, on the basis of text frequency, might better be taught as sight words than left to be read as built words. For a subsequent program, a frequency of 160 was set as the somewhat arbitrary lowest frequency of "sight words". Above 160 the frequency figures start to spread out, averaging a difference of about 10 between each adjacent word; immediately below 160 the average is about 3 or 4. Any word occurring more than 160 times in these texts was then considered a sight word and any word occurring less than 160 times was considered a content word. A grammatical or semantic definition would have required a different type of input to the computer, whereas a definition in terms of frequency could easily be identified by the computer from the input data. Of the 30 words so included as sight words, five are semantically "content" words: KA 'to eat', VA 'to want', AIZ 'to have', KHEENZ 'to say', and KIA 'ghost'. The last has a high frequency because many of the included tales in the text have ghosts as characters. Immediately below 160 fewer words are functors and most are semantically content words.

Another program then summarized the number of words having Chart 3) frequencies of 20 or less. See Chart 3. Thus Charts 2 and 3 account for all 1409 words of the texts.

Another program summarized the number, frequency, and average frequency of all words, of content words, and of sight words based on the frequency-of-160 definition just discussed. See Chart 4. This chart impresses one with the stark frequency difference between functors and content words. The former generally occur 40 times more frequently than the latter. And if this distinction is valid, it emphasizes the importance of the distinction in literacy between sight words (the functors) and built words (the content words).

II. Word count summaries

The primary program entailed the preparation of a series of 20
word count summaries of the phonemes in the various word positions and
Chart 5) in various combinations. See Chart 5.

Each word was analyzed by the computer into three word parts.

Word part 1: the presyllable $(C6)_p$

A "6" as the second or third letter of a word identified for the computer both the occurrence of and the end of a presyllable. A word without a "6" as the second or third letter is a one syllable word.

A two-dimensional presyllable array or matrix (17 x 3, or 51 cells) was established by which the computer sorted and stored the input data for preparation of summaries 1 and 2. The first dimension (or coordinate) provided 17 slots corresponding to the 16 different presyllable types occurring in the data with a 17th slot for words without a presyllable. The second dimension provided three subcategories for each of the 17 slots of the first dimension. The first subcategory was used to count the number of different words having a given presyllable; the second subcategory added the text frequencies of all content words having a given presyllable; and the third subcategory added the text frequencies of all (i.e. both function and content) words having a given presyllable. It is the first subcategory which makes the frequency count in a text described by Gudschinsky as "the number of different words in which each letter occurs." The other two subcategories contrast this with the absolute number of occurrences of the item among all content words and in the entire text.

Chart 6) Chart 6 presents Summary 1 which contrasts the dictionary and text counts of one- and two-syllable words. The first two columns of numbers cite the personally counted dictionary count of each item and the corresponding percentage of the total; the other columns were extracted by the computer from the text utilizing the presyllable array and present both the number of occurrences and their percentage of the total for the three subcategories. This summary shows that a dictionary word count indicates a somewhat lower percentage of one-syllable words than a text count (60% versus 69%); but, further, that in the text the absolute count of one-syllable content words is much higher (84%). Noting from Chart 2 that all function words are one-syllable, it follows that the absolute count of all one-syllable words in the text would be yet higher (91%).

Summary number:	Summary universe	Summary parts	No. Items	No. Words
1	W	$C6=\emptyset$ versus $C6\neq\emptyset$	2	1409
2	$C6\neq\emptyset$	each C6	16	442
3	W	each C_i	50	1409
4	W	$C_i=\emptyset$, $C_i=C_n$, $C_i=C_mC_n$	3	1409
5	$C_i=C_n$	each C_n	10	965
6	$C_i=C_mC_n$	each C_n	14	376
7	$C_i\neq\emptyset$	each C_n	10	1341
8	$C_i=C_mC_n$	each C_mC_n	30	376
9	$C_i=C_mC_n$	each C_m	7	376
10	W	each $\overset{R}{V}(G)(C_f)$	166	1409
11	W	each V	7	1409
12	W	each G	4	1409
13	$G=\emptyset$	each V	7	1135
14	G=A	each VA	4	227
15	G=OO	each VOO	3	30
16	G=E	each VE	2	17
17	W	each C_f	15	1409
18	W	each R	4	1409
19	W	R=laryn versus R≠laryn	2	1409
20	W	R=nasal versus R≠nasal	2	1409

Chart 5 Table of word count summaries

Source:	Dictionary		Text					
Type:	Word Count	%	Word Count	%	Content word frequency	%	Total frequency	%
1-syl	2,854	60	967	69	13,020	84	25,038	91
2-syl	1,914	40	442	31	2,399	16	2,399	9
Total:	4,768	100	1409	100	15,419	100	27,437	100

Chart 6. Summary 1: One versus two-syllable words

K6	455	23	108	24	680	28	680	28
T6	371	19	93	21	508	21	508	21
H6	267	13	66	15	258	11	258	11
R6	273	14	59	13	300	13	300	13
P6	221	11	44	10	253	11	253	11
M6	134	7	38	9	262	11	262	11
L6	53	3	12	3	25	1	25	1
I6	41	2	7	2	52	2	52	2
A6	32	2	6	1	16	1	16	1
B6	9	1	2	0	7	0	7	0
B/6	2	0	2	0	2	0	2	0
S6	19	1	1	0	1	0	1	0
D6	1	0	1	0	6	0	6	0
D/6	2	0	1	0	10	0	10	0
006	1	0	1	0	11	0	11	0
TR6	1	0	1	0	8	0	8	0
14 other types	32	2	-	-	-	-	-	-
Total:	1,914	98	442	98	2399	99	2,399	99

Chart 7. Summary 2: Presyllables

Source:	Dictionary		Text					
Type:	Word Count	%	Word Count	%	Content Word frequency	%	Total frequency	%
T	421	13	119	12	1917	18	2982	15
P	331	10	102	11	1541	14	1825	9
CH	285	8	96	10	1126	10	1126	6
K	261	8	76	8	1073	10	2868	15
L	213	7	72	7	957	9	1161	6
X	260	8	71	7	787	7	787	4
H	202	6	68	7	773	7	1149	6
N	233	7	63	7	523	5	703	4
M	197	6	54	6	542	5	2442	13
D	142	4	51	5	227	2	560	3
R	196	6	36	4	279	3	279	1
NG	97	3	30	3	563	5	563	3
NH	61	2	28	3	94	1	94	0
B	99	3	23	2	145	1	145	1
V	89	3	23	2	123	1	1014	5
J	79	2	22	2	154	1	154	1
G	34	1	16	2	41	0	1375	7
S	59	2	14	1	68	1	68	0
Y	11	0	1	0	20	0	20	0
Total:	3270	99	965	99	10953	100	19315	99

Chart Summary 5: Simple (single) consonants only

Chart 7) Chart 7 presents Summary 2 which contrasts the dictionary and text counts of the presyllables of the 2-syllable words which are classed together in Summary 1. The presyllables are listed here in decreasing order of the text word count. The summary shows only insignificant differences between the dictionary and text orderings; but, significantly, the dictionary includes 14 additional presyllable types that do not occur in the texts. Since no function words are two-syllable, it follows that the content word figures equal those of all words combined.

Word part 2: the main syllable initial consonant or consonant cluster ($C_i = C_mC_n$)

Whatever occurs before a vowel a, e, i, o, or u and after the presyllable "6" if present, identified for the computer the main syllable initial consonant or consonant cluster.

A three-dimensional initial consonant array (20 x 8 x 3, or 480 cells) was established by which the computer sorted and stored the input data for preparation of summaries 3 through 9. The first dimension provided 20 slots corresponding to the number of initial simple consonants charted below:

P	T	CH	K	#
B	D	J	G	
M	N	NH	NG	
V		Y		
	X	S		H
	L R			

The second dimension provided 8 subcategories to discriminate each of the above 20 units by distinguishing the various types of consonant clusters from the above simple consonants: (1) unmodified simple consonant, (2) CH, aspiration, (3) CL, (4) CR, (5) C/, preglottalization, (6) HC, voiceless consonants, (7) and (8) C/L and C/R, complex clusters with preglottalization and following L or R. The third dimension (like the second dimension of the presyllable two-dimensional array) provided three subcategories for each of the above 20 x 8 units, distinguishing a word count, content word frequency, and total frequency.

Summary 3 presented a printout of every initial consonant and consonant cluster occurring in the texts--50 items. The 13 most frequent items--all those having a frequency greater than 2% of the total--are Chart 8) compared with a corresponding dictionary count in Chart 8. There are only 4 discrepancies. L and # (initial glottal, unwritten before vowels) have a somewhat low dictionary count rating, and N and R a high dictionary count rating. The total frequency figures reveal that K, #, and M occur much more frequently in text material than the word count would suggest. Since the content word figures are ordered almost the same as the text word count, the much higher total text frequency of these items is the result of the fact that 16 of the 30 function words start with either K, #, or M. The dictionary count futher supplied 8 forms which did not occur at all in the text material.

Chart 9) Chart 9 shows Summary 4 which contrasts the word count and text frequencies of initial consonants versus constonant clusters. The dictionary and text word counts are identical. The total frequency figures are skewed away from the CC types and toward no initial consonant (no C) inasmuch a only 3 lower frequency functors have initial consonant clusters (KHEENZ, N/AI, PREEIZ) whereas nine others have no initial consonant.

Chart 10) Chart 10 shows Summary 5 which contrast the various simple (single) consonants without including their occurrence in consonant clusters. This list corresponds to the letters probably most sought after for use in the beginning primer lessons since, as Summary 4 indicated, simple consonants are much more productive than consonant clusters. Both dictionary and text word counts give the same order of the four most productive initial consonants (all voiceless stops): T, P, CH, K. Thereafter the order of the dictionary count is slightly different. The order of the content word frequency is almost the same as the text word count whereas the order of the total frequency is considerably different raising both M, G, and V much higher in overall ranking.

Summary 6 indicated the word count and frequency of consonants which occur as the centers or nucleii of consonant clusters. The ordering of the six most productive consonants in consonant clusters by text count are: K, D, P, N, T, B; whereas by dictionary count they are: K, P, D, T, B, N.

Summary 7 indicated the word count and frequency of consonants as simple consonants or as the nucleii of consonant clusters. The most productive by both text and dictionary count are, in order: T, P, K, D, N, L, CH, M. The frequency in content words gives the same ordering as by word count, but the total frequency would raise the ranking of M from 8th to 3rd place.

Summary 8 indicated the word count of consonant clusters. In text the most productive, in order, are: DR, KL, TR, PR, KR, N/, HM, HN. The dictionary count is almost the same except that PL has a high dictionary ranking (6th) compared to the text ranking (14th).

Summary 9 indicated the word count of consonant cluster types. The seven types are ordered identically in all four rankings: CR, HC, C/, CL, CH, C/L, C/R.

Word part 3: the word final -$\overset{R}{V}$(g)(C_f)

Whatever occurs after word part 2 above, except for hyphenated suffixes distinguishing homonyms and presyllable reduplicative vowels, identified for the computer the word final -V(G)(C_f).

A five-dimensional array (7 x 4 x 15 x 4 x 3, or 5040 cells) was established by which the computer sorted and stored the input data

Source:	Dictionary		Text					
Type:	Word Count	%	Word Count	%	Content word frequency	%	Total frequency	%
T	421	9	119	8	1917	12	2982	11
P	331	7	102	7	1541	10	1825	7
CH	285	6	96	7	1126	7	1126	4
L	213	4	72	5	957	6	1161	4
K	261	5	76	5	1073	7	2868	10
X	260	5	71	5	787	5	787	3
#	188	4	68	5	1057	7	4139	15
H	202	4	68	5	773	5	1149	4
N	233	5	63	4	523	3	703	3
M	197	4	54	4	542	4	2442	9
DR	155	3	51	4	514	3	514	2
D	142	3	51	4	227	1	560	2
R	196	4	36	3	279	2	279	1
...								

Chart Summary 3: Initial consonants and consonant clusters

No C	188	4	68	5	1057	7	4139	15
C only	3270	68	965	68	10953	71	19315	70
CC	1310	28	376	27	3409	22	3983	15
Total:	4768	100	1409	100	15419	100	27437	100

Chart Summary 4: Simple consonants versus consonant clusters for preparation of word summaries 10 through 20. The first dimension provided 7 slots corresponding to the number of single vowel positions:

I U
EE OO
E A O

The second dimension provided 4 subcategories to discriminate each of the above 7 units by distinguishing the simple vowels from each of the 3 vowel glides: V, VA, VOO, VE.

The third dimension provided 15 subcategories to discriminate each of the above 7 x 4 units by distinguishing open syllable words from each of the 14 final consonants occurring in the data: # and:

M N NG
P T K
U I
IH H
I/ /
R L

The fourth dimension provided 4 subcategories to discriminate each of the above 7 x 4 x 15 units by distinguishing the two intersecting syllabic prosodies of laryngealization ("vowel register") and nasalization: clear-oral, laryngeal-oral, laryngeal-nasal, and clear-nasal. The fifth dimension provided 3 subcategories for each of the above 7 x 4 x 15 x 4 units to store the input word count, the content word frequencies and total word frequencies.

Chart A indicates the various Summaries prepared. The unavaiaability of the Sedang rhyming dictionary makes a comparison of a dictionary and text word count of these various groups unfeasible at this time.

III. Conclusion

The following seems aparent from this study:

1. A dictionary or thorough phonemic statement must be consulted for preparation of complete lists ofphonemes and their various combinations into consonant and vowel clusters.
2. The Concordance Project provides a good source for determining the most frequently occurring function words.
3. For Sedang and similar languages, a dictionary word count of letters and their combinations is almost as reliable as a text word count of the same letters.
4. For Sedang and similar languages, the text frequency of letters and their combinations in content words only (omitting function words) is almost as reliable as a text word count of the same letters.
5. The total text frequency of letters and their combinations

as they occur in all words (i.e. including function words) is unreliable as an indicator of a text word count of the same letters, because of the distortions caused by high frequency function words.

Therefore, for one who must count letters and their combinations whithout the aid of a computer, note:

6. Making a text word count is difficult because one quickly loses track of which words have or haven't been counted, unless a concordance is used which groups together the various different occurrences of each word.

7. For lack of a concordance an acceptable substitute method may be either (a) a dictionary word count, or (b) a text frequency of the letters in all content words, eliminating the predetermined function words from the count.

FOOTNOTES:

[1] This paper was first presented to the SIL (Norman, Oklahoma) Literacy Forum on July 2, 1974.

Sedang is a Mon-Khmer language whose approximately 40,000 speakers have traditionally lived in central Kontum Province in the South Vietnam highlands, although in recent years have been partially scattered south as refugees to an area near Banmethuot.

[2] For this reason the symbol count of letters (including digraphs and trigraphs) computed by the SIL/-Oklahoma Concordance Project (see Footnote 3) was not useful for Sedang word counts.

[3] The concordance was compiled by the University of Oklahoma Computer Laboratory on their IBM410 computer by the Linguistic Information Retrieval Project of the Summer Institute of Linguistics and the University of Oklahoma Research Institute, and sponsored by Grant CS-934 of the National Science Foundation. This concordance, in additon to typical text material--dialogues, folklore, personal experiences, etc.--included Sedang songs. The content the repetitious and special vocabulary and the dialect variants of the songs seemed inappropriate for primers; the songs, however, more so coded within the concordance that words which occurred only or primarily in the songs were easily and appropriately deleted from inclusion in this study of frequency counts. Similarly borrowings from other languages and personal and village manes were also excluded.

[4] The following adaptations of Sedang orthography were made

(C, V represent Consonant, Vowed):

Sedang orthography	Computer orthography
Cǒ (presyllable)	C6
'C (preglottalized consonant)	C/
..	
V (final glottal stop)	V/
V (laryngealization)	V_z
V (naso-laryngealization)	V_2
V (nasalization)	V_3
e e	ee
o o	oo
Homonyms	-2, -3,
Presyllable vowel reduplication	-i, -u

5. The programs are written in the ALGOL-W computer language and were run on the Univac Spectra 70/46 computer at the University of Pennsylvania. The two introductory programs are probably applicable for any (tribal) language input, but the main program preparint the various summaries is unique for Sedang because of the unique Sedang phonemic system for which it was prepared.

BIBLIOGRAPHY:

Bruns, Paul C. "The use of a basic computer concordance in the preparation of literacy materials." Notes on Literacy (SIL), No. 4, January, 1969, p. 1-4.

Gudschinsky, Sarah C. A manual of literacy for preliterate peoples. Edited by Ramona Lucht, Jacqueline Firchow, and Eunice Loeweke. Ukarumpa, Papua New Guinea: SIL, 1973. ix, 180 p.