DOCUMENT RESUME

ED 277 977 CS 008 597

AUTHOR	Goodman, Kenneth S.; Bird, Lois Bridges
TITLE	On the Wording of Texts: A Study of Intra-Text Word Frequency. Program in Language and Literacy Occasional Paper, No. 6.
INSTITUTION	Arizona Univ., Tucson. Coll. of Education.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Mar 82
CONTRACT	NIE-C-00-3-0087
NOTE	52p.
AVAILABLE FROM	Program in Language and Literacy, College of Education, Room 504, University of Arizona, Tucson, AZ 85721 ($3.00 including postage).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; *Language Patterns; Language Research; *Language Usage; *Lexicology; *Reading Comprehension; Reading Instruction; Reading Research; *Vocabulary Development; *Word Frequency
IDENTIFIERS	*Textual Analysis

ABSTRACT

Analyzing word frequency in six complete texts, a study investigated how vocabulary can be used to define texts. The texts included three stories from 5th and 6th grade readers, selections from literature anthologies for 8th grade and 12th grade students, and a magazine essay for adults. Results indicated that if particular words occur frequently in a text, they do so because the language requires it. Few words were found to occur in all of the texts; those that did were almost all function words, "be" forms, or pronouns. Analyses of word frequency lists revealed that some content words occurred more than others simply because they referred to common concepts. Overall, findings suggested that the wording of any text is not random and that, in fact, texts "self-control" their vocabulary. These results suggested that (1) readers can build dependable strategies for dealing with words common in a text but less common in the language as a whole; (2) authors and editors should concentrate on relating the content of texts to the audience rather than focus on controlling vocabulary through the use of word lists; and (3) teachers concerned with vocabulary development should focus on functional use of words in the context of real texts rather than resort to decontextualization lists or dictionary exercises. (JD)

On the Wording of Texts: A Study of Intra-text Word Frequency

Kenneth S. Goodman
Lois Bridges Bird
University of Arizona

A Research Report
March 1982 No. 6

Occasional Papers
Program in Language and Literacy
Arizona Center for Research and Development
College of Education
Univerisity of Arizona

Co-Directors:
Yetta M. Goodman
Kenneth S. Goodman
402 Education, Bldg. 69
University of Arizona
Tucson, Arizona 85721

Research reported herein was supported in part by NIE (NIE C-00-3-0087) and by the Center for Expansion of Language and Thinking. Such support does not constitute an endorsement of the opinions or conclusions in the report.

Statement of Purpose

This series of working papers will provide a report of our current thinking and make available the work of our program to those who may be interested. It is our intent to stimulate an on-going dialogue with other professionals who share similar interests in educational theory and practice. We welcome responses from readers. Comments may be directed to the author of the paper or to the directors of the program.

Some but not all of the papers may appear in other publications in modified form. We are making this publication available at cost.

Word Frequency in Texts and In General

The topic of this research report is really the wording of texts; Halliday considers "wording" the folk term for what he describes as the lexico-grammar of the language (Halliday, 1981). As such the wording is both the final written representation of the meaning and the process by which the final selection is made. Which words constitute the visible text is certainly determined by the writer but only within strong lexico-grammatical constraints that the structure of language, meaning and social communication provide.

But in the field of reading and in the history of reading research the focus on the wording of language has concentrated on WORD FREQUENCY, the frequency with which words occur in general in the language.

As with so many popular notions in education, study of the issue of word frequency is dominated by the original reasons it was considered important and we have not looked objectively at the realities of wording as the result of text characteristics. Unless we examine how characteristics of coherent, functional, meaningful texts relate to choice and frequency of words in those texts we cannot truly understand the significance of relative frequency of words in use. So, to put our study in an educational context and to examine current educational belief and practice, we are making word frequency WITHIN TEXTS the focus of this paper.

Purposes for Studies of Word Frequency

One of the most deeply rooted notions among teachers of reading is that controlled vocabulary, based on studies of relative word frequency, is necessary in instructional materials for developing readers. This emphasis on using word frequency lists to build controlled vocabulary materials was not the original purpose for constructing such lists, however.

Counts of word frequency were first compiled as a means of determining the readability of existing texts. The researchers were operating on the premise that words which occur most frequently in the language are more easily recognized, learned, and processed in reading than less common words. So it seemed reasonable that the more high frequency words a text contains proportionate to low frequency words the easier it would be to read.

While using word frequency as a means of DETERMINING readability in existing texts may seem logical, there is a gap in the logic of manipulating word frequency to CONTROL readability, by restricting the vocabulary of new texts or rewriting old texts by substituting high frequency words for low f quency words. Research has demonstrated a moderate correlation between the proportion of unusual words and text difficulty even if it has also shown that this correlation is insufficient to determine readability. Most readability formulas still include some measures of unusual words. But artificially reducing vocabulary to create

texts with low proportions of uncommon words tampers with the very factors that may contribute both to word frequency and text difficulty. There are good reasons why particular words occur in particular places in particular texts and author's choice is only one of these reasons. Tampering with the wording of texts without understanding why words occur in texts the way they do may make texts less readable rather than more so.

Over the years a mystique has grown up around the significance of controlling vocabulary to control comprehensibility.

A major reason for accepting the importance of controlling vocabulary has been that many teachers and researchers have operated from a view of reading as getting words and of learning to read as learning to get words.

John Carroll, author of a comparatively recent study that used a base of 5,000,000 words, argues that the size of the reader's vocabulary is the most important causative factor in comprehension. (Carroll, 1981)

Ironically the methodology of counting word frequency has itself contributed to misconceived applications.

Early frequency studies showed that words vary so much in frequency from one context to another that it's necessary to examine a very large corpus of language to be able to support the assertion that the frequency found is representative of the whole language. Later studies were

based on awareness that the corpus must be a broader representation than a single source such as the Bible. Such grand scale studies may or may not provide a list truly representative of "all" language. But such a large corpus of language containing many texts eliminates the very factors that constrain the choice and frequency of vocabulary in a single coherent text. So when the frequency list is used to construct or rewrite texts it may make them strange and unpredictable. When the list is used to judge readability of a text it imposes the assumption that words are of equal difficulty regardless of where they occur and what kind of contextual support is provided.

Word Frequency as a Feature of Text

In this study we have put our focus on what grand scale word frequency studies could not shed light on: what does word frequency mean in the context of a single coherent and cohesive text not written or adapted for the purpose of the study? We want to know what it is about language in use that produces variable word frequency. Such knowledge will help put the issue of vocabulary in its proper context (no pun intended). But it will also help to define a text in terms of its use of vocabulary. This will provide knowledge of the relative importance of any particular word to the text and text comprehension. It will also suggest how vocabulary is developed through reading.

John Carroll (1981) reasons that since good compre-

hension correlates with good vocabulary, then vocabulary development is essential to comprehension. A more logical conclusion is that people who read a lot develop large vocabularies. So an important corrolary question is, "what is there about word use and frequency in texts that builds vocabulary during reading?".

Our study is rooted in a psycholinguistic theory of reading and it draws on data from past miscue studies (Goodman and Burke, 1973, Goodman and Goodman, 1978).

In this study we've examined word frequency in six complete texts which have been used in miscue research. Our purpose is to determine not only the frequency with which words occur in each text but also to determine why. We are seeking to understand how the vocabulary of the text relates to its other characteristics, and what constraints a complete text imposes on its vocabulary. Such understanding may call into question the use of word frequency lists in judging readability and in structuring controlled vocabulary basal readers.

A Historical Summary of Word Frequency Research

Over a long period research efforts have centered on proving which word frequency list or readability formula was the most effective and why. The word frequency variables selected to measure text difficulty were numerous: number of running words, percentage of different words, percentage of different infrequent, uncommon, or "hard"

words, percentage of polysyllabic words, vocabulary difficulty, vocabulary diversity, number of abstract words, number of affixed morphemes, and so on. (Lorge, p. 22, 1938)

Vocabulary control is not a new idea. Lorge traces word and idea counts back to the Talmudists in 900 A.D. who used frequency of occurrence to distinguish usual from unusual meanings. Nor is interest in word frequency comparatively recent in the United States; in 1840 the McGuffey Readers were claimed to contain words carefully selected for "ease of understanding," though criteria for selection were not made explicit.

N. A. Rubakin and F. W. Kaeding compiled word lists in 1889 and 1898 respectively. Kaeding set the precedent for using actual word counts to produce lists of words in order of frequency of occurrence. (Lorge, 1944)

While the early word counts all produced frequency lists they varied considerably in terms of the language sample from which they were drawn. Relying primarily on Bible passages, Knowles produced a three hundred and fifty word basic vocabulary for the blind. Eldridge's (1911) list of SIX THOUSAND COMMON ENGLISH WORDS, was drawn from four issues of the Buffalo, New York Sunday papers dated July and August 1909. (Klare, 1963)

Ernest Horn's A BASIC WRITING VOCABULARY (1925), a list of approximately five million words, was based on

personal and business correspondence. Next, Horn tackled the job of counting the spoken vocabularies of young children, ages one to six, and in 1926 published "The Commonest Words in the Spoken Vocabulary of Children up to and Including Six Years of Age." (Horn, 1926)

It was Thorndike's THE TEACHER'S WORD BOOK, published in 1921, however, that heralded the dawn of readability formulas. Thorndike's ten thousand most frequent words served as the basis for the first significant readability formulas and controlled vocabulary readers. (Thorndike, 1921)

Lively and Pressey in 1923, calculated the vocabulary burden of a book selecting a thousand-word sampling, assigning each word an index of difficulty that corresponded with the Thorndike list, and then computing the weighted median index number for the passage. Thus, the index numbers were based strictly on the frequency of the use of words. (Lively and Pressey, 1923)

Lively and Pressey's work sparked the interest of other researchers including Washburn and Vogel who in 1928 declared that the number of different words in a thousand was the most reliable indicator of passage-difficulty because of its close correlation to median reading scores obtained from the paragraph-meaning section of the Stanford Achievement Test. (Washburn and Vogel, 1926)

William S. Gray, father of the basal reader, cautioned: "It is reasonable to assume that the number of differ-

ent words used is a fair measure of difficulty, because it indicates the range of concepts involved. It fails, however, to consider whether the words used represent relatively simple or difficult concepts" (Gray, p. 492, 1947).

This is a serious oversight of the early vocabulary lists. Not only did they fail to consider the difficulty or simplicity of the concepts represented by the words, but they overlooked meaning and meaning variation completely.

The vocabulary studies that followed in rapid-fire succession employed various methods of compiling words.

Another 1928 study conducted by Dolch analyzed textbook series according to five indices of difficulty: percentage of different words; percentage of difficult words (using his combined word study list); degree of difficulty of words; median frequency of difficult words; and degree of difficulty for supplementary reading. Difficulty was equated with infrequency. (Dolch, 1928)

Next was Lewerenz's somewhat unorthodox 1929 study in which he focussed on words beginning with w, h, b, i, or e. He reported that words that start with w, h, or b occurred with relative frequency and could be classified as easy words, while words that begin with i or e were relatively few and were therefore, "hard" words. (Lewerenz, 1929)

Johnson relied on still another index of vocabulary difficulty. In 1930, he reported that the percentage of polysyllabic words in a passage is a reliable indicator of

the reading difficulty that children will experience. (Johnson, 1930)

One year later, in 1931, Patty and Painter presented a modified version of the Lively-Pressey method by listing all words located on the third complete line of each fifth page, multiplying the corresponding Thorndike index numbers of the selected words by the frequency of the use of the respective words, and finally, calculating the average-word-weight value by dividing by the total number of words in the sample. (Patty and Pointer, 1931)

In keeping with the vocabulary analysis tradition, Thorndike himself produced still another technique in the thirties based on his own word list. Using a sample of ten thousand words from the book to be analyzed, he counted the number of words it contained that were in the various categories of the TEACHER'S WORD BOOK, and then calculated the norms for each grade (Thorndike, 1932).

Klare characterizes this early period of readability research as follows:

1) primary attention paid to vocabulary (frequency) as a basis for predicting readability; 2) dependence upon Thorndike's TEACHER'S WORD BOOK as the basis for determining vocabulary difficulty; 3) use of "relatively crude criteria of reading difficulty." (Klare, 1963 p.44,)

Meanwhile, new approaches to vocabulary sampling were turning out still more word lists. In 1936, Buckingham and

Dolch utilized a free association technique in which 21,000 children in grades two through eight were asked to write all the words which they thought of during a fifteen-minute period. The result was a pool of two and a half million words which were then tabulated by lexical unit according to Thorndike's procedures.

In 1938, Rinsland and Moore, after collecting nearly six million written words from school children, announced their proposal for a list; "to assemble all data and words...into a consolidated list of approximately 15,000 different words with eight columns of frequencies for the eight grades after each word." (Lorge, p. 547, 1946)

It was Dale, however, who came closest to fitting the bill for a graded word list. He based his work on the premise that ninety percent of the children entering fourth grade would know some meaning of a vocabulary selected from previous word counts including Gates' A READING VOCABULARY FOR THE PRIMARY GRADES, all words from A STUDY OF THE VOCABULARY OF CHILDREN BEFORE ENTERING THE FIRST GRADE, and the first thousand most frequent words in Tidyman's A SURVEY OF THE WRITING VOCABULARIES OF PUBLIC SCHOOL CHILDREN. His testing procedure was quite simple and straightforward. He asked almost 8,000 children in grades 4, 6 and 8 to state whether or not they knew a given word. In 1943, Dale published a list of easy words based on the study. (Dale 1943)

Another major word count was conducted by Lorge in an effort to obtain an estimate of the frequency of the occurrence of words in adult reading material. Drawing from five adult magazines with high circulation: SATURDAY EVENING POST, LADIES' HOME JOURNAL, WOMAN'S HOME COMPANION, TRUE STORY and READERS DIGEST, Lorge pooled approximately five million running words. These were then combined with the Thorndike 20,000 Word Book, The Thorndike Juvenile Literature Count, and the Lorge-Thorndike Semantic Count and published as THE TEACHER'S WORD BOOK OF 30,000 WORDS. (Thorndike and Lorge, 1944)

Lorge himself suggested that word lists can be used most effectively in establishing a core vocabulary for children. He warned, however, that they "cannot be the ONLY basis of selection...," as they fail to account for the meanings of words. Even the most frequent words commonly have more than one meaning. For example, Lorge says the 850 WORDS OF BASIC ENGLISH represent 12,425 listed meanings in THE OXFORD DICTIONARY with approximately 5,991 additional senses that are not separately listed. Furthermore, Lorge said, each reader brings his or her own background of experience to the text, and the meaning the writer intended for a particular word or passage may not be the meaning the reader receives. (Lorge, 1944)

Everyone seemed to agree right from the start that

frequency of occurrence was important due to the common sense notion that common words are easier to recognize. So the next concern was the nature of the language sample from which the words were drawn. Everything from the Bible to popular adult magazines to the Buffalo Sunday newspaper were used. While some researchers swore by samples collected from written language, others (e.g. Ernest Horn) insisted that oral language was the best source. Still another variable was the age of the subjects from which the language was collected. The subjects ranged in age from preschool to adult.

Although there was wide variation in the nature of the sample, the basic approach to the collection was the same. A large number of words were generated over a range of related texts from a particular source.

The validity of grand-scale word counts was essentially assumed and never seriously questioned. When people began to realize that word counts alone were not adequate for readability, attention gradually shifted to the development of readability formulas that did incorporate other criteria besides vocabulary such as complex versus simple sentences, sentence length and qualitative factors including obscurity and incoherence in expression. But basal text continued to focus strongly on two factors: controlled vocabulary and repeated exposure, a contribution from behavioral psychology.

The essential weakness of all this word counting is that word frequency is treated as a phenomenon that exists independently of the text in which it occurs. Word frequency has been treated as a cause of text difficulty but not as a result of characteristics of the text itself.

Early research established that very large amounts of text must be used to get some sense of the relative frequency of words in general. But, as we said earlier, using huge bodies of language with millions of running words and thousands of different words blots out the characteristics of a text which determine the choice of words and their frequency.

Though authors have some choice in the words they use in creating a text, there is always a considerable amount of constraint on that choice. Some syntactic features of the language are extremely constraining. Common nouns particularly in the singular in English almost always require determiners. So THE and A are going to be very frequent in all English texts. THE will be more frequent than A because THE has an anaphoric quality: it is used with nouns already introduced in the text. Some semantic features of a text serve an essential and repeated purpose. If the text contains dialogue, SAID will occur very often. This explains why it is often the most common verb in a text.

But other semantic constraints derive from the message

or meaning being represented. So a story about a sheep dog defending her flock against predatory coyotes will make frequent use of some words not likely to be common in even several million words from school texts, several Sunday editions of the Buffalo Sunday paper, or many other sources. Such frequency doesn't make the text hard to read.

This is why in our study we've focussed on a very different set of questions. We have examined word frequency in the context of connected discourse, looking at the choice and frequency of words in relation to other text characteristics.

We have asked: What is REALLY happening in a text? Why are some words in a text more frequent than others? How are the words related to each other; semantically, syntactically? How do the words function syntactically? Do some words serve more than one syntactic function? How does word frequency relate to text cohesion?

Description of the texts selected for this study

The texts we have selected include two middle grade basal stories, S51,"Freddy Miller Scientist", (fifth grade) and S53, "My Brother is a Genius", (6th grade). S59, "Sheep Dog",is a selection from an eighth grade literature book.

S60, POISON, is a story by Roald Dahl published in unabridged form in a twelfth grade literature anthology. S70, "Ghost of the Lagoon," written by Armstrong Sperry, appears in its original form in a sixth grade reader. S61 "Why We Need the Generation Gap" is an adult magazine essay.

Operating from a theory that language controls its own vocabulary, we have examined word frequency in these texts. All have similar characteristics but are different too, depending on the author's purpose and style. We chose these particular stories because we have lots of miscue data on subjects reading them, allowing us to compare across texts with some degree of sophistication. We have purposely avoided using beginning reading material as it tends to employ rigidly controlled vocabulary.

Table One: Word Frequency: Types and Tokens.

Story Number	Running Words	Different Words	Used Once	Typ/Tok Ratio	% Words Used once
S51	1369	466	263	2.94	56.44
S53	2030	604	336	3.36	55.63
S70	2775	809	457	3.43	56.49
S59	3667	952	507	3.85	53.26
S60	4208	883	499	4.77	56.51
S61	1318	608	459	2.17	75.49

Table One provides some general data about the word frequencies in the six stories. Story length in terms of total running words (tokens) is proportional to the grade

level of the school selections. S51, the fifth grade story, has only 1369 total running words. The sixth grade stories, S53 and S70, have 2030 and 2775 respectively. The story from an eighth grade text, S59, has 3667 words while S60, the adult short story from a 12th grade anthology has 4208 words. The magazine essay is shorter with 1318 words.

The number of different words (types) also increases in materials for more advanced readers. So types increase from 466 in S51 to 604 in S53 to 809 in S70 and 952 in S59. But in S60 number of types actually is lower. This relates to a steady increase in the type/token ratio for successively more advanced stories from less than 3 uses per type in S51 to almost 5 in S60. We might expect the most effect of vocabulary control and deliberate repeated use of vocabulary in the two stories from the Betts readers. In fact the rising type-token ratio suggests that in less consciously controlled narrative material some words occur very frequently, in fact more frequently than in more controlled texts.

S61, the magazine essay, shows a very different pattern however with 608 types for 1318 tokens and a ratio of only 2.17. This probably represents the difference between this non-narrative text and the others which are all narrative.

We must examine another aspect of word frequency in these texts to get a more complete picture. In every word

study, no matter how large the corpus, many words will be found to occur only once. In all of these stories more than half the tokens occurred only once. In fact all of the five narrative stories show similar per-cents of word types used only once, 53.26% to 56.51%. It's not surprising that better than 75% of the types in S61, the essay, occur only once considering the low type/token ratio. One dimension of the wording of these narratives, then, is that more than half of the types occur only once. But an opposite dimension is that a few words occur extremely often.

Table Two: Words Representing % of Total Tokens

Stories		Total Words Tokens	Diff. Words Types	Cumulative Percent of Running Words 10%	20%	30%	40%	50%
S51	Words	1369	466	3	8	16	31	53
	% Types			0.64	1.72	3.43	6.65	11.37
S53	Words	2030	645	3	7	14	26	52
	% Types			0.47	1.09	2.17	4.03	8.06
S70	Words	2775	809	2	6	13	30	64
	% Types			0.25	0.74	1.61	3.71	7.91
S59	Words	3667	952	1	5	11	27	64
	% Types			0.11	0.53	1.16	2.84	6.72
S60	Words	4208	883	2	6	12	27	58
	% Types			0.23	0.68	1.36	3.06	6.57
S61	Words	1318	608	3	8	17	38	85
	% Types			0.49	1.32	2.80	6.25	13.98

To have a complete picture of relative frequency of these very frequent words in these six texts, we need to look at the number of different words (types) it takes to

account for cumulative percents of the running words (tokens). This information is indicated in Table Two. It takes only from 1 to 3 words in any of these six texts to account for 10% of the tokens. The most common word, THE, accounts for between 3.9%(S53) and 9.9%(S59) of all tokens.

To account for 20% of the tokens takes only 5 to 8 different words. That's from .51% to 1.71% of the types. It takes 11-17 types to account for 30% of the tokens. This is only 1.2 to 3.4% of the types. The latter figure is for the fifth grade text(S51). To account for 40% of the running words takes only 2.8 - 4% of the types except for S51(6.7%) and the essay, S61,(6.3%). To understand what this means consider that for 27 words in S60 to account for 40% of the total of 4208 words each of these 27 words occurs an average of 63 times.

Half of the tokens in each text are represented by 6.6 to 8.1% of the types except for S51 which requires 11.4% and S61 requiring 14%. Clearly these two selections are getting into lower frequency words than the other four. Each of the 58 words that account for the first 50% of S60 occurs an average of 36 times. Each of the 53 words that account for the first 50% of S51 occurs an average of 12.9 times. And for S61 each occurs only 7.8 times.

What all this illustrates is the extreme variability of word frequency within a text. More than half the individual words (types) occur only once in any of these six

texts while small numbers of types account for huge proportions of the total running words (tokens).

So far, however we have not looked carefully at which words appear so frequently and what text characteristics might account for their frequency. Figures 1a and 1b show the 25 most common words in each story (in four cases we include more due to matching frequencies).

The words on the list of the 25 most frequent words in each story represent from 35 to 39% of the running words of each text.

Figure 1a: Most Frequent Words in Frequency Order

	S51				S53				S70			
No.	Word	N	%	Cum. %	Word	N	%	Cum. %	Word	N	%	Cum. %
1.	the	78	5.6	5.6	the	82	3.9	3.9	the	269	9.6	9.6
2.	he	40	2.8	8.4	I	80	3.9	7.8	of	86	3.0	12.6
3.	Freddie	37	2.6	11.0	a	65	3.1	10.9	his	67	2.4	15.0
4.	to	36	2.6	13.6	and	52	2.5	13.4	a	63	2.2	17.2
5.	a	29	2.0	15.6	he	51	2.4	15.8	and	59	2.1	19.3
6.	was	28	2.0	17.6	said	51	2.4	18.2	he	57	2.0	21.3
7.	it	26	1.8	19.4	to	48	2.3	21.5	to	55	1.9	23.2
8.	his	25	1.8	21.2	you	31	1.5	23.0	was	46	1.6	24.8
9.	in	19	1.3	22.5	Mr.	28	1.3	24.3	in	38	1.3	26.1
10.	that	18	1.3	23.8	my	28	1.3	25.6	Mako	35	1.2	27.3
11.	and	18	1.3	25.1	of	28	1.3	26.9	canoe	33	1.1	28.4
12.	I	17	1.2	26.3	baby	26	1.2	28.1	on	26	.9	29.3
13.	you	16	1.1	27.4	Barnaby	25	1.2	29.3	it	25	.8	30.1
14.	had	15	1.0	28.4	at	24	1.1	30.4	that	24	.8	30.9
15.	of	14	1.0	29.4	was	24	1.1	31.5	boy	24	.8	31.7
16.	Elizabeth	14	1.0	30.4	Andrew	23	1.1	32.6	Tupa	24	.8	32.5
17.	Miller	14	1.0	31.4	in	22	1.0	33.6	him	23	.8	33.3
18.	with	13	.9	32.3	his	20	.9	34.5	Afa	22	.7	34.0
19.	uncle	12	.8	33.1	it	19	.9	35.4	with	21	.7	34.7
20.	mother	10	.7	33.8	on	17	.8	36.2	had	20	.7	35.4
21.	at	10	.7	34.5	as	14	.6	36.8	into	18	.6	36.0
22.	for	10	.7	35.2	but	14	.6	37.4	as	16	.5	36.5
23.	father	9	.6	35.8	for	14	.6	38.0	water	16	.5	37.0
24.	then	9	.6	36.4	that	13	.6	38.6	from	15	.5	37.5
25.	this	9	.6	37.0	typical	13	.6	39.2	out	15	.5	38.0
	Mrs.	9	.6	37.6								
	like	9	.6	38.2								
	said	9	.6	38.8								
	she	9	.6	39.4								

Figure 1b: Most Frequent Words in Frequency Order

	S59			Cum.	S60			Cum.	S61			Cum.
No.	Word	N	%	%	Word	N	%	%	Word	N	%	%
1.	the	370	9.9	9.9	the	259	6.0	6.0	the	73	5.2	5.2
2.	and	116	3.1	13.0	and	166	3.8	9.8	to	40	2.8	8.0
3.	to	110	2.9	15.9	he	137	3.1	12.9	and	35	2.5	10.5
4.	she	105	2.8	18.7	I	123	2.8	15.7	we	32	2.3	12.8
5.	her	105	2.8	21.5	to	117	2.7	18.4	will	32	2.3	15.1
6.	of	100	2.6	24.1	it	90	2.0	20.4	of	29	2.1	17.2
7.	a	76	2.0	26.1	his	85	1.9	22.3	in	24	1.7	18.9
8.	was	62	1.6	27.7	was	82	1.9	24.2	a	20	1.4	20.3
9.	Peggy	40	1.0	28.7	a	78	1.8	26.0	our	19	1.3	21.6
10.	it	36	.9	29.6	of	73	1.6	27.6	that	17	1.2	22.8
11.	sheep	34	.9	30.5	in	58	1.3	28.9	us	17	1.2	24.0
12.	in	33	.9	31.3	Harry	45	1.0	29.9	have	16	1.1	25.1
13.	for	33	.9	32.1	on	36	.8	30.7	for	15	1.0	26.1
14.	had	31	.8	32.9	you	36	.8	31.5	is	15	1.0	27.1
15.	as	31	.8	33.7	at	33	.7	32.2	they	12	.8	27.9
16.	from	27	.7	34.4	that	31	.7	32.9	when	12	.8	28.7
17.	on	27	.7	35.1	me	30	.6	33.5	be	10	.7	29.4
18.	coyote*	24	.6	35.7	but	29	.6	34.1	I	10	.7	30.1
19.	that	22	.5	36.2	him	29	.6	34.7	on	10	.7	30.8
20.	at	21	.5	36.7	up	29	.6	35.3	with	10	.7	31.5
21.	were	21	.5	37.2	there	28	.6	35.9	children	9	.6	32.1
22.	he	20	.5	37.7	said	28	.6	36.5	it	9	.6	32.7
23.	his	20	.5	38.2	now	26	.6	37.1	all	8	.5	33.2
24.	down	19	.5	38.7	for	25	.5	37.6	their	8	.5	33.7
25.	into	18	.4	39.1	Ganderbai	25	.5	38.1	are	7	.5	34.2
	coyotes*	18	.4	39.5	my	25	.5	38.6	from	7	.5	34.7
	band	18	.4	39.9	not	25	.5	39.1	can	7	.5	35.2
					out	25	.5	39.6	who	7	.5	35.7
									one	7	.5	36.2

The words on these lists account for better than a third of the running words in each text. Only eight words appear on all six lists. These are (with their mean rank):

THE(1), AND(4.5), TO(4.5), A(6.3), OF(8.3), IN(10.8) IT(12.8), THAT(15.5)

Of these eight words all are function words. IT and THAT can also function as pronouns. Even THE, the most common word in all 6 texts ranges from 3.9% to 9.9% of the running words in each text. This illustrates a key feature of word frequency in connected texts: VARIABILITY WITHIN CONSTRAINT. The language requires the use of THE but it permits sufficient variation to allow considerable range.

The words comprising these most common words may be divided for purpose of analysis into these main kinds:

1. Function words
2. Copula
3. Pronouns
4. Content words.

Function words include:

determiners (the, a),

verb markers (was, had, were, will, are, is, can),

conjoiners (and, as, that, but, when),

prepositions and particles (to, in, of, with, at, for, into, from, on, up, out,)

others (it, there, not)

One simple reason for the frequency of many function words is that, while the grammar of the language requires their functions, there are only a few words in the language which can fulfill each function. Only a few words can be determiners. There are few conjunctions and other conjoining elements in the language. There are more

prepositions but they still represent a finite set of words. Furthermore, while the language adds to its store of content words it does not add to its store of function words. Yet they are the binding material which makes the language cohesive and coherent.

To illustrate this, Table 3 shows the percentage of each type of function word in each of the six texts.

Table 3 Percent of Total Running Words

Function Word Type	S51	S53	S70	S59	S60	S61
Noun Marker	8.7	7.7	12.8	12.3	8.2	7.3
Verb Marker	3.5	3.2	2.6	2.7	3.4	6.2
Verb Particle	3.1	2.8	2	2.3	3.8	2.8
Question Marker	.3	.2	.1	.1	.3	-0-
Clause Marker	3.3	2.1	2.5	2.5	2.3	4.7
Phrase Marker	7.8	7.2	11.3	11.6	8.9	10
Intensifier	1.7	1.6	1.8	.6	1.5	1.2
Conjunction	2	3.5	2.6	3.8	5	3.5
Negative	.7	.4	.4	.7	.9	.7
Quantifier	1.1	1.3	1.2	1.6	1.2	1.5
Other	.4	1.2	.3	.5	.7	.7
Total	32.7	32.1	37.6	38.7	36.4	38.9

From 32.1 to 38.9% of each text's running words are functions words. The terms we use here to describe the various functions are those of C.C.Fries. We prefer them for this purpose because of their descriptive reference to what they do.(Fries, 1952)

The noun markers are few, mostly THE and A(AN) but they represent from 7.3 to 12.8% of the running words. The phrase markers(prepositions) are more common but still represent a small set of words. These words also serve as

verb particles. Contrast "he ran up the street" with "he ran up the flag." The former is a phrase marker, marking a prepositional phrase. The latter is a verb particle, part of the verb, RAN UP. In these combined functions, this set of words represents from 10 to 14% of the running words in each text.

There is substantial variation from text to text in use of conjunctions; S60 uses two and a half times as many as S51. But together with clause markers, which introduce subordinate clauses, conjunctions account for 5.1% to 8.2% of each text's running words. Again, a very small set of words in the language carries a big part of the running text.

The words which serve as copula are the BE forms. BE, WAS, WERE, IS, ARE show among the most common in these six texts. These words also can serve as VERB MARKERS. Which BE forms appear as COPULAS or VERB MARKERS depends very much on the prevailing tenses in the text which in turn is determined by whether the text is about the past, present or future. So S51 and S70 show only WAS and HAD among their most common words. S53 and S60 list just WAS. S59 has WAS, HAD and WERE. But the essay S61 shows WILL, HAVE, IS, BE, ARE among its most common words.

Pronouns are clearly common among the most frequent words in each text. That's because the language requires the use of pronouns for recurrent nouns. IT is common in

all our texts, but which other pronouns are used depends on characteristics of the text. This is well illustrated in S59, Sheep Dog. The central character is a female dog, Peggy. SHE and HER occur 105 times each and tie for fourth and fifth most common word in the text. HE and HIS occur, but only 20 times each.

S53, My Brother is a Genius, has predominately male characters and is told in the first person. So among its most common words are: I, HE, YOU, MY, HIS. S70 also has male characters but is told in third person so its common words include these pronouns: HIS, HE, HIM. S51 has both male and female main characters and quite a bit of dialogue. Its common words include: HE, I, YOU, SHE, HIS. The essay S61 uses a great deal of first person plural to represent a generalized society: "When we..." So it's not surprising that these pronouns are among the most common words: WE, OURS, US, THEY, I, THEIR, WHO, ONE.

To sum up, pronouns are important cohesive elements but which ones are common in any text depends on text characteristics such as cast of characters, dialogue, and whether it's first person or third person narration.

Possessive pronouns are actually the most common noun modifiers. In fact function words acting as "pro" elements can take the place of any of the content words, not just nouns. Verb phrases may be replaced by verb markers: "Will you get it? Yes, I will." Adverbials may be replaced by

prepositions: "He walked in and looked around."

All these text characteristics explain the frequency of function words. But they also explain the surprising infrequency of content words.

Nouns are the only content words to appear in any number among the lists of most frequent words in Figures 1a and 1b. Here are the nouns that appear:

S51- FREDDIE, ELIZABETH, MILLER, UNCLE, MOTHER, FATHER.

S53- BARNABY, BABY, ANDREW.

S70- MAKO, CANOE, BOY, TUPA, AFA, WATER.

S59- PEGGY, SHEEP, COYOTE, COYOTES, BAND.

S60- HARRY, GANDERBAI.

S61- CHILDREN.

It's not surprising that in each of the narrative texts the most common noun is the name of one of the characters. In three of them it's the principal character but in S53 and in S60 it is not the main character because these are first person stories. In fact, in S53 the main character is never named. What is more surprising is that the most common nouns in these stories are not necessarily common in the language. Only BABY, BOY, and CHILDREN could be considered truly common. And some really uncommon nouns appear among these most frequent words: CANOE (to cross the lagoon), SHEEP, BAND (the group of sheep) and COYOTE and COYOTES (pair of adversaries).

The essay, S61 had only one noun, CHILDREN, among its most common words. Only three nouns occurred more than four times in the entire text: CHILDREN (9 times), GENERATION (6), and AGE (5). It is apparently possible to write an essay without using the same nouns very often, particularly since there are main ideas but no main characters.

What about other content words? S51 has a verb modifier, THEN, a kind of noun modifier, MRS., and the verb SAID. UNCLE actually appears fairly often as a noun modifier. Mrs. Miller keeps telling Freddie he's "just like UNCLE...". S53 has SAID, a verb and TYPICAL, a noun modifier. The story centers around whether Andrew is a typical baby. HAD, used as a verb, is the only non-noun content word among the most frequent list in S70 and S59. THERE and NOW as verb modifiers are among the most common words in S60 and SAID is the only verb. HAVE, sometimes a verb is the only non-noun content word on the S61 list.

Only five verbs in S51 occur five times or more in the entire text. These are SAID, THOUGHT, GET, KNEW, and CALLED. In S53 the five most common verbs (Six times or more) are SAID, THINK, SEE, KNOW, and GO. The five most common verbs in S70 (occurring five times or more) are SAW, COME, LEAPED, HEARD, and ROSE. The contrast between this more active set of verbs and those in S51 and S53 also shows in S59. The most common verbs in that are TURNED, SAW, LEAPED, LOOKED, MADE (6 times or more). S60, with much tension

but little action has these five most common verbs (13 times or more): SAID, WENT, MOVE, LOOKED, STOOD. These verbs occur three times or more in S61: FIND, SUSPECT, KNOW, BECOME, DO, JOIN, SEEN.

While these verbs provide interesting insight into the content of each text they show also that few verbs are frequent across texts and few verbs are frequent within texts. SAID, of course, will be common where there is dialogue.

Few verb modifiers occur with any great frequency in any of the texts. THEN is relatively frequent in all texts except S61. THERE, sometimes a verb modifier, is also found with moderate frequency in most but not all of the texts. NOW is found several times in three of the texts. Beyond that, the verb modifiers that occur more than two or three times are specific to the text. The five most common verb modifiers in S70 involving the killing of a shark are: THEN, AWAY, AGAIN, BEFORE, QUICKLY. S59 with the fighting of the dog and coyotes has a similar list: THEN, AGAIN, SLOWLY, FORWARD, CAREFULLY. And the very suspenseful S60 shows: SLOWLY, AGAIN, CAREFULLY, QUICKLY, SHARPLY.

Noun modifiers other than possessive pronouns are even more varied. Few occur more than five times even in the longer texts. Not all of the more common noun modifiers are adjectives. In S59 COYOTE and SHEEP are used five or more times as noun adjuncts. BEDDING, verb derived, occurs

five times (THE BEDDING SHEEP). Again the lists of more common noun modifiers show their particularness to each text: S51 shows Freddie's problem experiments: DARK, SMALL, BAD, PROUD, QUEER. S70's list shows the shark fight theme: GREAT, WHITE, OLD, GREEN, DEAD. S61 has HUMAN, POLITICAL, VIETNAM, and GOLD (that's the Generation Gap).

Table 4 Percent of Running Words in each Grammatical Category

Grammatical Category	S51	S53	S70	S59	S60	S61
Pronouns*	9.3	11.6	4.9	6.7	11.8	6.9
Other Nouns	21.5	17.9	24.5	22.8	16.1	20.6
Total Nouns	30.8	29.5	29.4	29.5	27.9	27.5
Verbs	17.6	18.3	15.3	15.4	18.4	17.5
Noun Modifiers*	10.2	10.7	10.7	10.2	8.8	11.6
Verb Modifiers	4.6	4	4.8	4.1	5.8	3.1
Function Words	32.7	32.1	37.6	38.7	36.4	38.9
Indeterminate	0	.7	.2	.1	.3	0
Contractions	2.3	4.2	.6	.6	2.2	.6

*Possessive pronouns are included as noun modifiers.

To put this information about the relative frequency of different grammatical categories of content words into perspective, Table 4 presents the distribution of each category in each entire text.

It's interesting to note that the total percent of noun positions in these six texts only varies from 27.9% to 30.8%. Yet the texts vary considerably in what part of those noun positions are filled by pronouns, from 4.9% to 11.8%. The two first person stories, S53 and S60 have similar high pronoun percents, 11.6 and 11.8 respectively. These two stories have sharply lower percents of other

nouns. The rest of the variation in use of pronouns and nouns seems to reflect amount of dialogue and other stylistic factors. English clauses and sentences require nouns as subjects, direct and indirect objects, objects of prepositions, etc. The proportion, at least in these texts, seems to vary little. But other factors, some of which the author may control, appear to decide how many nouns are replaced by pronouns.

Verbs show less variation, from 15.3 to 18.4%. S70 and S59, the two texts with the lowest rate of verbs, have little dialogue because of text factors. In S59 there are no human characters in much of the story. In S70 a considerable part of the story involves only Mako, a boy, his dog, Afa, and Tupa, a great white shark. So, whereas SAID occurs 51 times in S53 and ties for fifth most common word, it occurs only five times in S59 and twice in S70. Representation of oral dialogue in written text requires a special grammar which includes an extra clause representing at least the speaker and some representation of the verb SAID.

The amount of dialogue present also seems to explain the variation in the relative amounts of contractions in each text since most of the contractions appear in dialogue. S53 with the most dialogue has 4.2% contractions. S70, S59, and the essay S61 with little or no dialogue have only .6% contractions each.

There is a very specific textual reason for the percent of words with indeterminate grammatical function in S53. The central plot of the story is about an 8-month old baby learning to say big words by listening to his older brother read words from the dictionary. So words like PHILOSOPHICAL and INTELLECTUAL occur as word names out of syntactic context and are classified as indeterminate. That adds up to 7% of the running words of the text, in contrast to negligible proportions for the other texts.

Noun modifiers and verb modifiers vary moderately in proportion from text to text, apparently for stylistic reasons. Possessive pronouns, included in noun modifiers, range from 2.2 to 3.5% of each text. The grammar of English requires neither noun modifiers nor verb modifiers to produce grammatical sentences. The meaning the author is representing may require a good deal of describing and qualifying but how much is clearly a function of the author's purpose and style. S60, contains a lot of terse dialogue. One central character, Harry Pope, thinks he has a poisonous snake resting on his abdomen so he's minimizing his speech and movements in order to avoid startling the snake. This leads to fewer noun modifiers. In the essay, S61, there are more noun modifiers because the author uses a lot of embedding transformations to produce long, complex clauses and sentences. He also uses more adverbial clauses than adverbs. So he has a higher proportion of

noun modifiers and a lower proportion of verb modifiers. His text is at the high end in use of function words, which also reflects its syntactic complexity. Table 3 shows this text has the highest percents of clause markers and verb markers among the function words.

To summarize this discussion of the distribution of grammatical categories in these texts we can make the following statements.

The syntax requires some proportional distribution of these grammatical categories within the texts but other text characteristics including semantic structure of the story and the author's purpose and style produce some variations among the texts in these proportions. Some very common grammatical functions can only be filled by a relatively small set of words, so these words are likely to be common in any text. Function words and pronouns (including possessive pronouns) are the principal examples.

On the other hand the categories of content words, nouns, verbs, noun modifiers, and verb modifiers, are much larger classes of words often called "open" classes because the language is continually adding to them. Still, the characteristics of particular texts exercise some constraint on the choice of words to fill these grammatical

slots. Particularly proper nouns, the names of characters in the story, are likely to be among the most frequent words. There is a similar but more moderate influence on verb frequency. Narratives with lots of action will select verbs of movement while suspenseful texts will choose another set. Still, SAID is the only verb likely to become very frequent.

In the case of all content words there is a counter pressure to the factors causing some words to occur more frequently than others. That's the rhetorical value that authors in the English language place on using varied terms and alternate ways of representing the same referents. We don't like to keep using the same nouns, verbs, adjectives or adverbs over and over and we'll even avoid using the same sentence patterns repeatedly.

MULTIPLE MEANINGS

Lorge criticized word lists for their failure to account for multiple meanings of words. This criticism does appear to be a major shortcoming, especially when you consider that the many meanings of even a common word such as RUN fill a dictionary page. However, within the confines of the single texts we examined multiple meanings for particular words seldom occur. In fact, after examining our six texts, we were able to find only one word, ALLOWANCE in S51 that has two clearly different meanings in the story itself. In one instance, Mrs. Miller, chiding Freddy for

ruining his sister's doll says, "I want you to save half of your ALLOWANCE for it each week." In the other, after Freddy has used his scientific ingenuity to free his sister from a dark closet, Mrs. Miller says proudly, "After this we must make some ALLOWANCE for experiments that do not turn out so well."

Our finding is somewhat surprising in light of the fact that these stories do make use of controlled vocabulary. Authors of controlled vocabulary texts often use words over and over again without regard for a possible change in meaning.

While the multiple meanings of a given word may not occur in a single text, nevertheless the meaning of a word in a particular text may not be a common one and the reader may be unfamiliar with the unusual meaning. In S59, for instance, the author repeatedly refers to a BAND of sheep. Called upon to define BAND out of context, you might think of "band of gold," "rubber band," "brass band," and so forth, before naming BAND as a term for a group of animals. Likewise, you might be hard pressed to come up with the meanings out of context for AIR and LIVE that appear in S53 in relation to television. Mr. Barnaby bemoans the fact that in five minutes they are going "on the air," "with a live show." In S60 DRAW and variations DREW and DRAWING appear four times but never in the way one would probably think of first, to DRAW a picture. Rather, we find the

following examples:

1. "He...DREW his breath sharply through his teeth."
2. "...he stuck the needle through the rubber top of the bottle and began DRAWING a pale yellow liquid up into the syringe by pulling out the plunger."
3. "Shall we DRAW the sheet back quick...?"
4. "Slowly he DREW out the rubber tube from under the sheet."

Within a given text, an author may use words in unusual ways either frequently (BAND of sheep in S59) or infrequently (BODY of the island in S70). But the meaning of any word is always derived from the context in which it is embedded.

Almost any word can be used metaphorically. The authors of our six stories employ metaphor to greater or lesser extents. The metaphorical uses of common words, BODY (of the island), FACES (of the cliffs), ARMS (of the island) in the opening passage of S70 are descriptively powerful but textually unpredictable. S59 begins with a string of vivid metaphors:

> The rays of the setting sun lingered over the high Arizona desert, touching the rocky tip of Badger Mountain and tinting the bold face of Antelope Rim.

What is clear is that the particular meaning of a word in a text, whether literal or metaphoric may not be predicted from the word's general frequency. Common words may

be used in quite uncommon ways.

Text Cohesion

The wording of a text is strongly influenced by the need for the text to be syntactically and semantically cohesive, that is to have a unifying structure.

The information we have presented so far shows that syntactic cohesion requires some proportionate distribution of grammatical functions and that some words will be common simply because there are few words to fill very common syntactic functions. Determiners, prepositions, pronouns are some examples.

We've also seen some evidence of the influence that maintaining semantic cohesion has on text wording and word frequency. But this is more complex as it relates to choice of content words, synonyms, and "pro" elements.

We can illustrate semantic cohesion by looking at approximately 20 opening lines of each text. Each author needs to accomplish a good deal in these opening lines to set up a cohesive text and create a semantic structure.

S51 starts with a lament: "Poor Freddie was in trouble again". The author, in the opening 22 lines, focuses on creating Freddie's character, his experimenting and the constant trouble this gets him into. Freddie's family is also introduced and a sub-theme, his mother's comparing Freddie with his Swiss uncles is also established.

In these opening lines we find these cohesive chains

(frequency in parentheses):

FREDDIE (30): FREDDIE(4), HE(5), HIS(5), FREDDIE'S, YOU'VE, I(4), YOU(4), YOUR(2), HIM(2), TINKER(2)

TROUBLE(7): TROUBLE, TURNED GREEN, POOR, WRECKED, QUEER, BAD, SADLY.

CHEMISTRY(4): CHEMISTRY SET, EXPERIMENT, MIXTURE, CHEMICALS

ELIZABETH(5): ELIZABETH(2), LITTLE, SISTER, HEARTBROKEN

MOTHER(10): MOTHER(2), SHE(3), I(2), MRS. MILLER(2), ANGRY

UNCLES(7): UNCLE AUGUST, UNCLES(2), SWITZERLAND, ONE, THEM, LIKE(FREDDIE)

There are 30 references to Freddie that use 10 different words in these 22 opening lines. The abundance of dialogue in S51 results in an interesting pattern: Freddie is referred to by name, nickname and by first, second, and third person pronouns. The semantic cohesion in S51 results in some words being repeated while at the same time the author achieves variety by using alternatives and related terms.

S53 opens with a statement of the problem:

"If it bothers you to think of it as baby sitting," my father said, "then don't think of it as babysitting. Think of it as homework...."

In the opening 22 lines the author creates the problem. A school age boy does his homework while caring for his baby brother. The older brother, who is the un-

named narrator, and the baby brother are established as is the task.

NARRATOR(22): YOU(2), MY(4), YOUR(3), I(11), FELLOW, ME

HIS MOOD(6): FOOLISH, ASHAMED, YELLED(2), SHOUTED, STAY

BABY(12): BABY(4), BROTHER(2), YOU(2), ANDREW(2), ANDREW'S, HIM

BABY'S CHARACTER(10): SILLY, SOUNDS, CRY(3), DISTURB, FAULT, SLEEPING, WANT, TRIED, HOLD

BABYSITTING(8): BOTHERS, IT(2), BABY SITTING(2), DISTURB, STAY, HOME

HOMEWORK(19): HOMEWORK, PART, EDUCATION(2), STUDYING, DICTIONARY, WORD(3), PHILOSOPHICAL(3), STUDY(2), MEANINGS(3), DEFINITIONS(2)

The main character here is referred to 22 times, almost all in first and second person, requiring only 6 words and no name. HOMEWORK, a key event throughout the story, has 20 references and 10 different words in these 22 opening lines.

S70 begins by establishing the setting: "The island of Bora Bora, where Mako lived, is far away in the South Pacific". The author concentrates on the setting and on Mako, his young hero in the opening 24 lines. There are these cohesive chains:

SETTING(9): ISLAND(3), BORA BORA, SOUTH PACIFIC, IT(3), MAIN BODY

ISLAND CHARACTERISTICS(13): FAR AWAY, RISES, HIGH, (LIKE)

CASTLE, WATERFALLS, FACES, CLIFFS, UPWARD, CRAG(2), EDGE, ARMS, REEF

WATER(6): SOUTH PACIFIC, SEA(2), WATER, SURF, LAGOO

MAKO(13): MAKO(4), HIS(4), HE(2), THEY, COMPANIONS, TWO

AFA: AFA(6)

MAKO'S CHARACTER(6): CLEVER, MADE, SPENT, BORN, HANDS, HEIGHT

HARPOON(6): HARPOON, STRAIGHT, ARROW, TIPPED, SPEARS, POINTED

CANOE(12): CANOE(2), LARGER, OUTRIGGER, SIDE, BOAT, TIPPING, LARGE, HOLD, HOLLOWING, TREE, LONGER

Thirteen references to Mako, single or with his dog Afa require six words. Half the references use pronouns. The characteristics of Mako and the island use many references with only one word, CRAG, used twice.

559 also begins with the setting: "The rays of the setting sun lingered over the high Arizona desert, touching the rocky tip of Badger Mountain and tinting the bold face of Antelope Rim." In the first 25 lines, the author creates both mood and setting while introducing sheep dogs at work. We find these cohesive chains:

EVENING(10): RAYS, SUN, TINTING, SETTING, DARKNESS, BEDDING DOWN, NIGHT, DROWSINESS, DARK.

PLACE(10): DESERT, ROCKY TIP, BADGER MOUNTAIN, BOLD FACE, ANTELOPE RIM, BASIN, SALT CREEK WASH, POOL, PATCH

SHEEP(14): BAND(3), SHEEP(3), 800, LAMB(S)(2), BLEATING,

MASS, EWE, HER, FAR-SIDE

DOG(S)(7): DOG(S)(3), TWO, EARS, HER, MATE

PEGGY'S CHARACTER(6): PATROLLING, URGED, ALERT, LARGER, TURNED, ASSURED

These chains use many words once. Only DOG(S), BAND, SHEEP, LAMB(S) occur more than once. The main character, a sheep dog, is not named until line 27 of the story.

The author devotes the next 27 lines to creating Peggy as the central character. In that sequence this chain occurs:

DOG(S)(3), PEGGY(2), SHE(6), HER(8), BREED, COLLIES, COAT, HEAD, EYES(2), DESCENDANT, FOREPAW, TOES, FOOT. There are 29 references then to PEGGY, more than one per line, yet even after her name is introduced the author only uses it twice in this sequence, using 14 pronouns instead.

S60 begins with creating the mood and establishing two main characters in the setting.

"It must have been around midnight when I drove home...," There are these chains in the 26 lines:

TIMBER WOOD (Narrator)(16): I(12), ME (2), TIMBER (2)

HARRY POPE(11): HARRY POPE, HIS(2), HE(5), HE'D, HARRY'S, HIM

SETTING(20): HOME(2), GATES, BUNGALOW, WINDOW, SIDE BEDROOM, DRIVE, STEP(S)(2), BALCONY(2), ONE, TOP, DOOR(S)(2), HOUSE, ITSELF, HALL, ROOM, IT

DARK(9): MIDNIGHT, SWITCHED OFF, BEAM, SWING IN, LIGHT(2),

STILL ON, DARK, SWITCHED ON

SLEEP(9): WAKE, AWAKE(2), DROPPED OFF, QUIETLY, LYING, BED, MORE, TURN

ATTITUDE(6): APPROACHED, BOTHERED, NOTICED, CAREFULLY, QUIETLY, LOOKED IN

MOVEMENT(11): DROVE, APPROACHED, OPENED, COMING, PARKED, WENT UP, TAKE, GOT TO, CROSSED, PUSHED THROUGH, WENT ACROSS.

Three words are enough to represent Harry Pope 16 times. But 11 different verbs are used to show movement of the main character with no one verb used twice. This shows again the text characteristics that make words both common and diverse in a connected text.

S61 starts with establishing two age groups. "Recently, I spoke with a man twice my age who expressed great faith in the future of American youth:"

Cohesion chains advance the two groups and set a tone for the essay.

YOUTH(16): YOUTH(S)(2), THEM, YEARS, YOUNG(2), TROUBLEMAKERS, WRONG, I, MY, AMERICAN, THEY(2), AGE, MILLIONS, SONS,

ELDERS(6): MAN, TWICE, HE, MATURITY, FATHERS, CYNICISM

ARGUMENT(8): SPOKE, EXPRESSED, ENVISIONS, THINKING (WISHFUL), WANT, ACCEPT, CYNICISM, PRE-CONCEPTION

THE GAP(4): DIVIDED, GAP(GENERATION), FUTURE(2)

ACTIONS(5): MARCHING, FIGHTING, RISK, DROPPING, SHAVING

SYMBOLS OF MATURITY(7): SHAVING, DROPPING (HEMS),
ACCULTURATING, FAMILY, MORTGAGE, PAYMENTS, YES

Each of these opening sequences illustrates how the need for semantic cohesion limits the writer's choices but how writers achieve such cohesion while also achieving stylistic diversity and a richness of wording. By producing a mix of function words, pronouns and varied content words, the author builds a cohesive text and builds literary style at the same time.

In each text, the author achieves the semantic and pragmatic purposes of the opening lines by staying within text constraints while still making use of the rich language resources. In S60 few different words are needed to refer to the two characters but 14 terms establish the house and 11 different verbs impel the reader into the story as Timber progresses to Harry's room.

The author can, to some extent, choose to use fewer terms, more common terms, or less varied terms. But authors seem to be aware that, as context builds, variety adds depth to the comprehensibility without making the whole less comprehensible.

So in S61, when the author uses BABY FOOD, WEED KILLER, and CONVERTIBLE DEBENTURES as examples of how youth will be acculturated, he knows that his readers can get his point without exactly knowing what a debenture is. In fact the term may have been chosen deliberately to sound tech-

nical and boring.

Armstrong Sperry in S70 prefers variety over repetition in representing the CANOE and ISLAND and does not avoid unfamiliar terms such as CRAG, REEF, PANDANUS, SURF, LAGOON, OUTRIGGER when they seem appropriate. His purpose is to create a sense of setting, not to teach an island vocabulary. But, in fact, through he use of synonyms and related terms in cohesive chains, the author creates a context which makes it possible for readers to infer meaning and build vocabulary.

CONCLUSION

Our study of the wording of texts or, if you prefer, word frequency as a text characteristic has demonstrated that general word frequency lists can at best tell only part of the story.

If words are frequent across texts it's because the language requires them to be. But such words that are very frequent in all texts are very few in total number and almost all are function words, be forms, and pronouns.

In the word frequency lists some content words will be found considerably higher than others. That's because they are used in common ways to refer to common concepts and experiences. But in particular coherent, cohesive texts, which content words are common depends on the content of the text. In narratives the common content words usually involve characters' names, some other important nouns and a

few others. But authors avoid using content words repetitiously for stylistic reasons. So cohesive chains are built of common pronouns, key content words, and a varied set of terms all somehow semantically related in the context of the story.

The wording of any text is thus by no means random. In fact texts self-control their vocabulary. For readers that means they will build dependable strategies for dealing with words common in the text but less common in the language as a whole. They will also build strategies for knowing the relative importance to text comprehension of particular words and terms. And of course they will build strategies for expanding their vocabularies through the reading of naturally worded texts.

Authors and editors would do better to focus on relating the content of texts to the audience than to focus on controlling vocabulary through use of word lists. A sense of audience and use of the natural constraints of the language will result in text wordings which are in keeping with the backgrounds of the intended readers and the strategies readers develop.

Teachers concerned about vocabulary development would do better to focus on functional use of words and terms in the context of real texts than to resort to decontextualized lists or dictionary exercises. A text, after all, is considerably more than the sum of its words.

Bibliography

Buckingham, B. R. and E. W. Dolch, A COMBINED WORD LIST, Boston: Ginn and Company, 1936.

Carroll, J. B., P. Davies and B. Richmond, THE AMERICAN HERITAGE WORD FREQUENCY BOOK. Boston: Houghton, Mifflin, 1971.

Carroll, J. B., NEW ANALYSIS OF READING SKILLS, presentation, International Reading Association, New Orleans, April, 1981.

Dale, E. EASY WORD LIST. Columbus: Bureau of Educational Research, Ohio State University, 1943.

Dolch, E. W., "Vocabulary Burden", JOURNAL OF EDUCATIONAL RESEARCH, March 1928, pp 170-183.

Fries, C. C., THE STRUCTURE OF ENGLISH: AN INTRODUCTION TO THE CONSTRUCTION OF ENGLISH SENTENCES. New York: Harcourt, Brace and World, 1952.

Gates, A., A READING VOCABULARY FOR THE PRIMARY GRADES. New York: Teachers College, Columbia University, 1935.

Goodman, Y. M. and C. L. Burke, READING MISCUE INVENTORY, New York: Macmillan, 1972.

Goodman, K. S. and Y. M. Goodman. READING OF AMERICAN CHILDREN WHOSE LANGUAGE IS A STABLE RURAL DIALECT OF ENGLISH OR A LANGUAGE OTHER THAN ENGLISH. Final Report, NIE-C-00-3-0087. August, 1978.

Gray, W. S. "Progress in the Study of Readability," THE ELEMENTARY SCHOOL JOURNAL, May, 1947, pp 491-499.

Halliday, MAK. LANGUAGE AS SOCIAL SEMIOTIC. Baltimore: University Park Press, 1978, p. 40.

Horn, E. A. A BASIC WRITING VOCABULARY. Iowa City: University of Iowa Monographs in Education, First Series, No. 4, April, 1926.

Horn, E. A. "The Commonest Words in the Spoken Vocabulary of Children Up to and Including Six Years of Age" in TWENTY-FOURTH YEARBOOK of the National Society for the Study of Education, Part I, Chapter VII, 1925.

Johnson, G. R. "An Objective Method of Determining Reading

Difficulty", JOURNAL OF EDUCATIONAL RESEARCH, April, 1930.

Klare, G.R., THE MEASUREMENT OF READABILITY, Iowa City: Iowa State University Press, 1963.

Lewerenze, A.S. "Measurement of the Difficulty of Reading Materials." LOS ANGELES EDUCATIONAL RESEARCH BULLETIN, March, 1929, pp. 11-16.

Lively, B.A. and S.L. Pressey, "A Method for Measuring the 'Vocabulary Burden' of Textbooks." EDUCATIONAL ADMINISTRATION AND SUPERVISION, October 1923, pp. 389-398.

Lorge, I., "Predicting Reading Difficulty of Selections for Children." THE ELEMENTARY ENGLISH REVIEW, 1938.

Lorge, I., "Word Lists as Background for Communication" TEACHERS COLLEGE RECORD, May, 1944, pp. 453-52.

Patty, W.W. and Painter, W.I. "Improving our Method of Selecting High School Textbooks," JOURNAL OF EDUCATIONAL RESEARCH, June, 1931, pp. 23-32.

Rinsland, H.D., and J.H. Moore. THE VOCABULARY OF ELEMENTARY SCHOOL CHILDREN IN THE UNITED STATES. Norman, Oklahoma: Works Progress Administration, 1938.

Thorndike, E.L. THE TEACHER'S WORD BOOK. New York: Teachers College, Columbia University, 1921.

Thorndike, E.L. A TEACHER'S WORD BOOK OF 20,000 WORDS. New York: Teachers College, Columbia University, 1932.

Thorndike, E.L. AND I. Lorge, THE TEACHER'S WORD BOOK OF 30,000 WORDS. New York: Teachers College, Columbia University, 1944.

Washburn, C. and M. Vogel, "What Books Fit What Children?", SCHOOL AND SOCIETY, January, 1926. pp. 22-24.

TEXTS ANALYSED

S51: Moore, L. "Freddie Miller, Scientist," in ADVENTURES HERE AND THERE, Editors, E.A. Betts and C.M. Welch, New York: American Book Company, 1965.

S53: Hayes, W.D. "My Brother is a Genius," in ADVENTURES NOW AND THEN, Editors, E.A. Betts and C.M. Welch, New York: American Book Company, 1965.

S59: Stovall, J.C. "Sheep Dog," in WIDENING VIEWS, Editors, W.D. Sheldon and R.A. McCracken, Boston: Allyn and Bacon, 1966.

S60: Dahl, R. "Poison" in ADVENTURES IN ENGLISH LITERATURE. New York: Harcourt, Brace, and World, 1958.

S61: Rapoport, R. "Why We Need a Generation Gap", LOOK January, 13, 1970, P. 14

S70: Sperry, A. "Ghost of the Lagoon", in OPEN HIGHWAYS, Editors, H. Robinson et al. Glenview: Scott, Foresman, 1967.

PROGRAM IN LANGUAGE AND LITERACY
ARIZONA CENTER FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
COLLEGE OF EDUCATION
UNIVERSITY OF ARIZONA

The Program in Language and Literacy is an innovative effort to provide a center for a variety of activities dedicated to better knowledge of development in language and literacy and more effective school practice. The program is concerned with language processes as well as learning and teaching of language.

Activities of the program have several main concentrations:

- Research on oral and written language –
  - on development of oral and written language.
  - on teaching for effective use of oral and written language.
  - on curriculum for language growth and use.
  - on bilingual, bicultural, biliterate language development, language instruction.
  - on issues of adult basic literacy.
- Theory development in oral and written language processes.
- Acquisition and instruction of oral and written language processes.
- Development of curriculum and methodology for effective monolingual and bilingual school programs.
- Support for language and literacy components of pre-service teacher education programs.
- In-service programs to help teachers, curriculum workers, and school administrators to achieve more effective programs in language and literacy.
- Consultation to school systems and other agencies to plan and evaluate more effective programs in language and literacy.
- Graduate courses, seminars, minors and combined majors in educational linguistics to help educators become more effective as teachers, curriculum workers material developers and teacher educators.
- Conferences, workshops, symposia to provide dialogue among researchers, disseminators and practitioners.
- Publications including working papers, position papers and research reports.

The program focuses on written language. Written language is a receptive and productive process in a literate society where people have the alternative of using oral language in face-to-face situations or written language over time and space.

The program is cross-disciplinary. It draws on a wide variety of bases-sociology, sociolinguistics, psycholinguistics, and areas of psychology-so that we can understand the learning of language and cognition and see the relationship of thought and language. We draw from other disciplines as well on neurology, physiology, and of course pedagogy, the study of education itself. The Program in Language and Literacy is a program in educational linguistics.

Staff:

Dr. Kenneth S. and Yetta M. Goodman, Co-directors
Faculty, Elementary Education, University of Arizona, Tucson

Lois Bird, Research Assistant, Series Editor
Suzanne Gespass, Research Assistant
Myna Haussler, Research Assistant
Wendy Kasten, Research Assistant
Sherry Vaughan, Research Assistant
Sandra Wilde, Research Assistant
Diana Ybarra, Secretary
Dorothy McCormack, Secretary

Order Form
Occasional Papers
Program in Language and Literacy
Arizona Center for Research and Development
402 Education, Bldg. 69, College of Education
University of Arizona, Tucson, AZ 85721

Cost $2 per copy — No. of copies

No. 1: Goodman, K.S. and Goodman, Y.M. A WHOLE-LANGUAGE COMPREHENSION CENTERED VIEW OF READING DEVELOPMENT. A position paper, February, 1981. _____

No. 2: Goodman, K.S. and Gollasch, F.V. WORD OMISSIONS IN READING: DELIBERATE AND NON-DELIBERATE. A research report, March, 1981. _____

No. 3: Altwerger, B. and Goodman, K.S. STUDYING TEXT DIFFICULTY THROUGH MISCUE ANALYSIS. A research report, June, 1981. _____

No. 4: Goodman, Y.M. and Altwerger, B. A STUDY OF LITERACY IN PRESCHOOL CHILDREN. A research report, September, 1981. _____

No. 5: Milz, Vera E. YOUNG CHILDREN WRITE: THE BEGINNINGS. February, 1982. _____

No. 6: Goodman, K.S. and Bird, L.B. THE WORDING OF TEXTS: INTRA-TEXT WORD FREQUENCY. A research report. April, 1982 _____

No. 7: Goodman, K.S. and Gollasch, F.V. RECONCEPTUALIZING INSERTIONS IN ORAL READING. A research report, in preparation. _____

No. 8: Goodman, K.S. and Gespass, S. TEXT FEATURES AS THEY RELATE TO MISCUES: PRONOUNS. A research report, in preparation. _____

No. 9: Goodman, K.S. and Gespass, S. TEXT FEATURES AS THEY RELATE TO MISCUES: DETERMINERS, THE and A/AN. A research report, in preparation. _____

No. 10: Goodman, K.S. and Gespass, S. TEXT FEATURES AS THEY RELATE TO MISCUES: DIALOGUE AND DIALOGUE CARRIERS. A research report, in preparation. _____

Total amount ($2 each) _____

N.I.E. Final Report: Reading of American Children Whose Language is a Stable Rural Dialect of English or a Language Other Than English. Director: Kenneth S. Goodman (1978) _____
($12 per copy)

$1.00 handling per order* _____

TOTAL: _____

Please make check payable to Program in Language and Literacy
*Actual shipping charges will be added for large orders or overseas orders.

END

U.S. DEPT. OF EDUCATION

OFFICE OF EDUCATIONAL RESEARCH AND IMPROVEMENT (OERI)

ERIC®

DATE FILMED

JUNE 9 1987