How many words are there in printed school English?

WILLIAM E. NAGY
RICHARD C. ANDERSON
University of Illinois at Champaign-Urbana

THE PURPOSE of this research was to determine the number of distinct words in printed school English. A detailed analysis was done of a 7,260 word sample from the Carroll, Davies and Richman, *Word Frequency Book*. Projecting from this sample to the total vocabulary of school English, our best estimate is that there are about 88,500 distinct words. Furthermore, for every word a child learns, we estimate that there are an average of one to three additional related words that should also be understandable to the child, the exact number depending on how well the child is able to utilize context and morphology to induce meanings. Based on our analysis, a reconcilation of estimates of children's vocabulary size was undertaken, which showed that the extreme divergence in estimates is due mainly to the definition of "word" adopted. Our findings indicate that even the most ruthlessly systematic direct vocabulary instruction could neither account for a significant proportion of all the words children actually learn, nor cover more than a modest proportion of the words they will encounter in school reading materials.

Combien de mots y-a-t-il dans l'Anglais scolaire imprimé?

LE BUT de cette recherche était de déterminer le nombre de mots distincts dans l'Anglais scolaire imprimé. Une analyse détaillée a été faite d'un échantillon de 7260 mots à partir du *Livre de fréquence de mots* Carroll, Davies et Richman. En projetant à partir de cet échantillon jusqu'au vocabulaire total de l'Anglais scolaire, notre estimation meilleure est qu'il y a environ 88500 mots distincts. De plus, pour chaque mot qu'un enfant apprend, nous estimons qu'il y a une moyenne d'un à trois mots additionnels en rapport qui devraient être aussi compris par l'enfant, le nombre exact dépendant de combien l'enfant est capable d'utiliser le contexte et la morphologie pour produire des significations. Basée sur notre analyse, on a entrepris une réconciliation d'estimations de la quantité de vocabulaire des enfants, qui a montré que l'extrême divergence en estimations est due principalement à la définition du "mot" adopté. Nos découvertes indiquent que même l'instruction de vocabulaire directe la plus systématique ne pourrait ni compter pour une proportion significative de tous les mots que les enfants apprennent vraiment, ni couvrir plus qu'une proportion modeste de mots qu'ils rencontreront dans le matériel de lecture scolaire.

¿Cuántas palabras hay en el inglés escolar escrito?

EL OBJETIVO DE ESTA INVESTIGACION era determinar el número de distintas palabras en el inglés escolar escrito. Se hizo un análisis detallado de una muestra de 7.260 palabras del *Word Frequency Book* (Libro de Frecuencia de Palabras) de Carroll, Davies y Richman. Reflejando desde esta muestra al vocabulario total del inglés escolar, nuestro mejor cálculo indica que hay aproximadamente 88.500 palabras distintas. Además, por cada palabra que un niño aprende, calculamos que hay un promedio de una a tres palabras relacionadas adicionales que también deberían comprenderse por el niño, dependiendo el número exacto hasta qué punto el niño es capaz de utilizar contexto y morfología para inducir significado. Basado en nuestro análisis, se procedió a una conciliación de cálculos sobre la extensión del vocabulario de los niños, que demostró que la extrema discrepancia en los cálculos se debe principalmente a cuál definición se adopta sobre lo que es "palabra". Nuestros resultados indican que ni siquiera la más rigurosa y sistemática instrucción directa de vocabulario podía justificar una proporción significativa de todas las palabras que los niños realmente aprenden, ni cubrir más que una modesta proporción de las palabras con las que se enfrentarán en materiales de lectura escolares.

Determining the absolute size of individuals' vocabularies is of more than purely theoretical interest. If students must learn 8,000 words by their senior year in high school, an ambitious program of direct instruction might hope to cover every word. If, on the other hand, the number of words to be learned were closer to 80,000, this goal would be beyond the reach of even the most intensive direct instruction that could be accomplished in the time available. The absolute size of vocabularies also has implications for theories of learning and language acquisition. If some seventh-grade students have vocabularies of over 50,000 words, as is estimated by some researchers, a theory of language acquisition must include mechanisms that account for this phenomenal accomplishment.

There is, in fact, a substantial lack of agreement among researchers as to the absolute size of vocabulary at any given age or level of development (see Anderson & Freebody, 1981). For example, estimates of average total vocabulary size at third grade range from 2,000 words (Dupuy, 1974) to 25,000 words (Smith, 1941). The same two researchers estimate the vocabularies of seventh-grade students to be around 4,760 and 51,000 words, respectively. Some of the reasons for such large disparities between estimates are the source of words (e.g., what dictionary or corpus to take as representing English vocabulary, and how to choose a representative sample), testing methods (disagreements about when a word can be counted as "known," and how to test such knowledge), and the definition of "word" adopted (disagreements about, for example, whether to include proper names, or under what conditions to count derived words as separate items).

It is with the third of these issues that we are primarily concerned here. Our goal has been to answer the question "How many different words are there?" in a number of ways, for a variety of criteria for defining "distinct words." The answer to this question is important because estimates of people's vocabulary size hinge critically on assumptions about the number of words in the language. Based on what we have judged to be a linguistically and psychologically sensible definition of a "distinct word," we attempted to reconcile previous estimates of vocabulary size that have applied different and sometimes questionable criteria for defining words. Our technique was to recalibrate previous estimates using benchmarks derived from a corpus that we analyzed in depth. No original data on the number of words children know is reported in this paper; instead, data that others have collected are reinterpreted.

Constructing or evaluating a test which attempts to measure absolute vocabulary size depends on the answer to three questions: What source of words should be used, what types of words should be included or excluded, and under what conditions related words should be grouped together or treated as separate items. The goal of our work has been to provide a basis for estimates of vocabulary size that are interpretable in terms of their implications for vocabulary instruction.

A Corpus of Words Representative of Printed School English

Dictionaries are often used as a starting point for building tests to estimate vocabulary size, although, as Carroll (1964) pointed out, this is a questionable practice. The organization and inclusion or exclusion of items in a dictionary will reflect not only linguistic principles, but also diverse practical demands such as page format and limitations on overall size. The estimates of vocabulary size that a given test produces are related to the size of the dictionary that was used in constructing the test (Hartman, 1941; Lorge & Chall, 1963). Further variation is introduced in the selection of items from the dictionary. Researchers differ in whether categories such as proper names, technical terms, or scientific names of flora and fauna should be included, and in the criteria for determining which derived words are to be counted as separate items.

We have chosen as our source of words Carroll, Davies, and Richman's (1971) *American Heritage Word Frequency Book* (*WFB*). This book is based on the American Heritage Intermediate Corpus, which contains, 5,088,721 words of running text from over a thousand items of published materials in use in schools. The materials sampled included textbooks, workbooks, kits, novels, poetry, general nonfiction, encyclopedias, and magazines chosen "to represent, as nearly as possible, the range of required and recommended reading to which students are exposed in school grades three through nine in the United States" (p. xxi). Carroll, Davies, and Richman have been able to use the corpus to determine properties not just of the vocabulary contained in the *WFB*, but of the total vocabulary of the type of materials from which the sample was collected. This total vocabulary is a theoretical construct, but its overall size (and several other properties) can be predicted with a substantial degree of confidence. Thus, our analysis can be generalized not just to the vocabulary in the *WFB*, but to the entire population of published material of which the *WFB* constitutes a representative sample. Because of the way that the American Heritage Intermediate Corpus was collected, we can justifiably refer to this population as "printed school English" (with the restriction to Grades 3 through 9 understood).

Printed school English, in this sense, gives us the basis for an operational definition of the total number of words of English. A vocabulary test based on this material could not be taken as a measure of a child's *oral* vocabulary, but would certainly be appropriate as a measure of a child's *reading* vocabulary.

On Defining the Concept "Word"

Absolute vocabulary size can only be discussed in terms of some theory of relatedness among words. For example, the *WFB* is described as containing 86,741 different words, or types. However, since the corpus was sorted by computer, "word" is defined as a graphically distinct sequence of characters bounded right and left by a space. By this definition, *doctor, Doctor,* and *DOCTOR* are counted as three different words. Obviously, a psychologically more realistic definition of word will count these three types as instances of the same word.

Dictionaries have traditionally treated regular inflectional variants, such as *walk, walks,* and *walked,* as forms of the same word. This is pedagogically justifiable; by the time children reach first grade, they have normally learned the basics of English inflection. If a child has learned the word *antelope,* no separate instruction about the plural *antelopes* is needed; children can automatically apply the rules of regular pluralization to new forms (Berko, 1958).

Some dictionaries take other types of relatedness into account when grouping words into entries. Many list semantically transparent derivatives as subentries. For example, the *American Heritage School Dictionary* gives *meekness* and *meekly* as subentries under *meek* without further definition. Along similar lines, Thorndike (19[illegible]) grouped abverbs ending in *-ly* under their base forms, thus counting *sadly* and *sad* as one word. From a theoretical perspective, Aronoff (1976) argued that words derived by totally productive word formation processes (e.g., *-ness, -ly*) should not be given separate entries in the lexicon.

However, there is a great variety of types and degrees of relatedness among words that might be taken into consideration when estimating vocabulary size, ranging from the transparent cases just mentioned to more obscure relationships such as that between *quiet* and *acquiesce*. And there has been little agreement among vocabulary researchers as to how different types of relatedness among words should be treated. The extremes run from counting inflectional variants as separate words on the one hand, to a radical grouping such as in Dupuy (1974), who excluded from his count of "Basic Words" almost all suffixed, prefixed, and compound items, since these could in some sense be considered to be derived from more basic words, and hence at least partially redundant. It should be clear that such decisions concerning how words should be counted is a major factor in determining how large estimates of absolute vocabulary size turn out to be.

Previous analyses of relatedness among words have not provided an adequate basis for meaningful measures of absolute vocabulary size; each suffers from at least one of a number of weaknesses. Many take an etymological approach to relationships among words, positing relationships based on historical information not available to the normal language learner. Some statistical analyses of word formation have been limited to prefixes, or to suffixes, or perhaps both of these, while neglecting compounding. Previous studies have usually adopted a single criterion of relatedness among words, without distinguishing types or degrees of relatedness. Some studies are based on wordlists such as Thorndike and Lorge (1944), which are now outdated.

Becker, Dixon, and Anderson-Inman (1980) have perhaps come closest to our purposes in their analysis of a vocabulary list derived by modifying and updating Thorndike and Lorge. They analyzed a list of 25,782 words into morphographs (minimal "meaningful" units of written English), and assigned each word a root word which represents the smallest word from which a given word can be "semantically derived." This root word analysis defines patterns of interrelatedness among words. For example, *divide, divided, dividend, dividers, dividing, divisible, division, divisional,* and *divisor* are related in that all have been assigned the same root word *divide.*

However, in their analysis, there are no distinctions made between possible types of degrees of relatedness. Also, relatedness is defined on a etymological basis. For example, *millenium* was assigned the root word *annual,* on the grounds that both are derived from the Latin *annus*, year. A historical linguist can certainly see the relationship in form between these two words, but it is dubious that the normal speaker of English, armed only with such knowledge of morphology as can be gained from words currently in the language, would find any but a semantic relationship between them. *Animism* and *animosity* were assigned the root word *anima*; in this case, the relationship in form may be obvious, but the semantic relationship is rather distant. In the case of *polynomial* and its root word *name*, both the formal and semantic relationships are tenuous.

Analyses of affixes (Thorndike, 1941; Stauffer, 1942) have also typically been done on an etymological basis, e.g., segmenting *fragile* into a root *frag-* and the suffix *-ile*, or *deceive* into the prefix *de-* and the root *-ceive.* An exception to this is found in Harwood and Wright (1956) who specify in their counts which suffixed forms have a free base (e.g., *acceptable*) and which do not (e.g., *amiable*). However, while these analyses do give an indication of the extent to which some suffixes account for a portion of the overall vocabulary, they do not provide a basis for estimating the overall size of vocabulary, that is, they do not tell us what number of words actually are derivable using a given suffix.

Rhode and Cronnell (1977) have analyzed a set of vocabulary items especially compiled to cover words used in grades K-6. However, their analysis, while including much useful information, focuses on types of letter-sound correspondence, so that their definitions of prefix and suffix are not in terms of productive word-formation processes in today's English. For example, their list of suffixes includes the *-om* of *bottom* and *-il* of *peril.*

We analyzed relatedness among words, not in terms of their historical derivations, but in terms of the similarity of their current meanings. For example, the relationship of a derivative word to its base (e.g., *business* to *busy* or *darkness* to *dark*) was viewed in terms of the relative ease or difficulty with which an individual who knew the meaning of only one of the words could guess or infer the meaning of the other when encountering it in context while reading. Also, we defined different types and degrees of relatedness among words, so that we could adjust our definitions of *related* and *distinct* to match the knowledge of word-relatedness of children at a given age or ability level.

Method

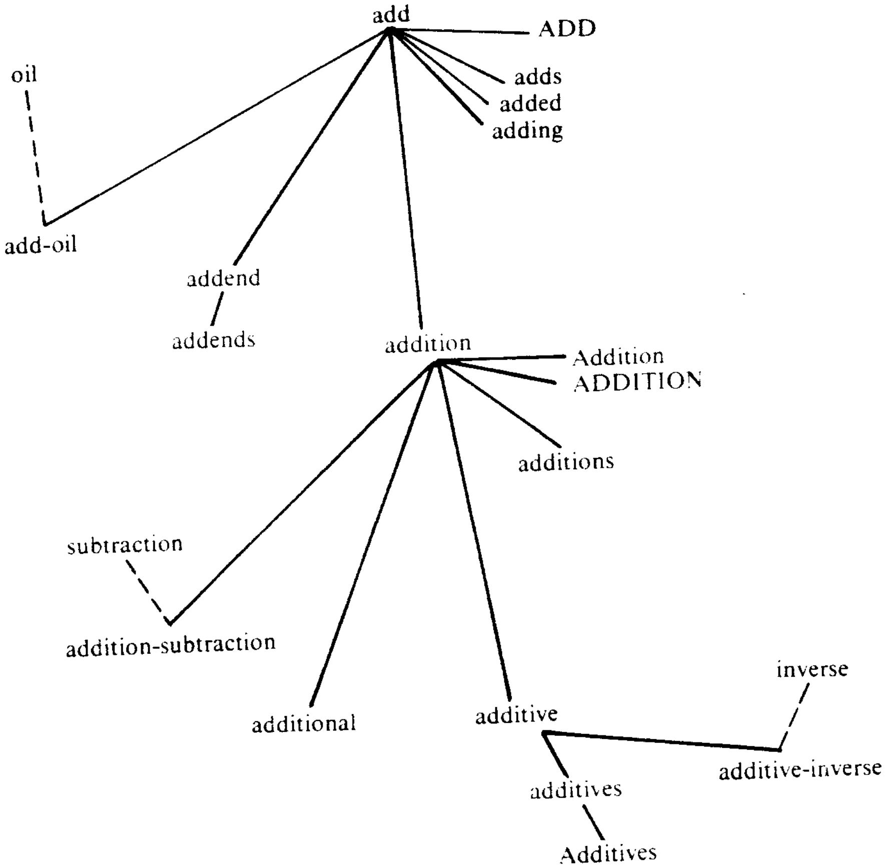
The data and statistical analyses in the *WFB* provide a reliable starting point for investigating the vocabulary of printed school English. However, the definition of "word" adopted for the purpose of compiling the

WFB is, as the authors would freely admit, inappropriate for any linguistic or pedagogical estimate of vo-^bulary size; some system must be adop-. 1 for grouping together graphically distinct types that constitute instances of the "same word." Our goal, then, is to categorize the different types of words in the *WFB*, and how they are related to each other, in order to arrive at a meaningful estimate of the number of different words in printed school English.

A random sample of 7,260 words was drawn from the 86,741 words in the *WFB*. This sample consisted of 121 chunks of 60 alphabetically contiguous words taken from the computer tape version of the *WFB*, including the *hapax legomena* (types found only once in the corpus) that appear at the bottom of the page in the printed version. Contiguous groups of words were taken because related words are usually (but not always) close to each other in an alphabetical listing.

Figure 1 gives a representation of the pattern of interrelationships in a group of related words, or word family, found in a chunk. All of the words in this family are related to the word *add*, some directly and some through intervening words. Compounds have multiple relationships, so there will frequently be overlap between word families.

Figure 1
Graphic representation of relationships among words



A more complete representation of word family structure would have to distinguish different types and different degrees of relatedness. This can be done by breaking down the tree structure in Figure 1 into pairs of words that are adjacent within the tree, and characterizing the type and degree of relatedness for each pair. Table 1 gives the family structure of the words from Figure 1 represented in this form. The pairs were constructed by assigning to each word in the family an "immediate ancestor," that is, the word most closely related to that word which is also closer to the root word of that tree.[1] The basic categories of relationship types used in our analyses are listed and exemplified in Table 2. (We will use the term suffixation to cover only derivational suffixes; inflectional suffixes are categorized under regular and irregular inflections.)

Coding Semantic Relatedness

In addition to distinguishing among different types of *formal* relationships between a word and its immediate ancestor (e.g., suffixation, prefixation, compounding), our coding system categorizes the semantic relationship between the two. For some pairs, e.g., *tranquil/tranquility*, the semantic relationship is fairly direct. For other pairs of words, it is more distant, e.g., *fun/funny*, *live/lively*, or *fix/prefix*.

An immediate problem in trying to characterize the semantic relationship between two words is the fact that one or both of them may have a number of meanings; before one can describe the semantic relationship between the two, one must first decide which two meanings are to be compared. The analyses reported here are based on the most similar familiar meanings of each pair of words. For example, the words *carry* and *carriage* have some meanings (concerning posture and bearing) that are very closely related. But the more familiar meanings *carry* = "to hold while moving" and *carriage* = "vehicle" are somewhat more distant. It is the latter pair of meanings that would be characterized by our semantic relatedness code.

Table 1 Relationships among members of word family in terms of target words and "immediate ancestors"

Target Word	Immediate Ancestor	Affix	Relationship
add	- - -	- - -	Morphologicaly basic word
Add	add	- - -	capitalization
add-oil	add	- - -	compound (first member)
add-oil	oil	- - -	compound (second member)
added	add	- - -	regular inflection
addend	add	end	suffixation
addends	addend	- - -	regular inflection
adding	add	- - -	regular inflection
Adding	adding	- - -	capitalization
addition	add	ition	suffixation
Addition	addition	- - -	capitalization
ADDITION	addition	- - -	capitalization
addition-subtraction	addition	- - -	compound (first member)
addition-subtraction	subtraction	- - -	compound (second member)
additional	addition	al	suffixation
additions	addition	- - -	regular inflection
additive	addition	ive	suffix replacement
additive-inverse	additive	- - -	compound (first member)
additive-inverse	inverse	- - -	compound (second member)
additives	additive	- - -	regular inflection
Additives	additives	- - -	capitalization
adds	add	- - -	regular inflection

Table 2 Categories of relationships among words

	Examples	
	Target Word	Immediate Ancestor
Morphologically basic word	add	- - -
Simple capitalization	Think	think
Alternate spellings	cart-horse	carthorse
Alternate pronunciations	fishin'	fishing
Alternate form of word	soya	soy
Alternate form with *s*	towards	toward
Regular Inflections	walks	walk
Irregular inflections	went	go
Regular comparatives & superlatives	taller	tall
Irregular comparatives & superlatives	best	good
Suffixation	frustration	frustrate
Prefixation	unknown	known
Compounds and contractions	farmhand	farm, hand
	can't	can, not
Truncations	rhino	rhinoceros
Idiosyncratic morphological relationships	prophesy	prophecy

Degrees of Semantic Relatedness

The *American Heritage School Dictionary* (1977) was used as the primary reference for determining the meanings of words, since this dictionary is based on the corpus we have analyzed, thus reflecting meanings occurring in that corpus. The code for semantic relatedness was defined in terms of the following question: Assuming that a child knew the meaning of the immediate ancestor, but not the meaning of the target word, to what extent would the child be able to determine the meaning of the target word when encountering it in context while reading? Six levels of semantic relatedness were distinguished:

SEM 0. The semantic relationship between the target word and immediate ancestor is semantically transparent. The possible exception is any semantic feature that would be totally predictable from a change in part of speech. For example, if a child knows the word *red* and has any grasp of the suffix *-ness*, that child should be able to determine the meaning of *redness* without help from the context. This level of semantic transparency is associated with almost all regular inflections, many compounds (if one knows the meaning of *plankton* and *burgers*, the meaning of the rather novel word *planktonburgers* is easy to compute), and many derivational affixes (knowledge of the word *misinterpret* should almost guarantee that a person would understand the word *misinterpretation*).

SEM 1. The meaning of the target item can be inferred from the meaning of its immedicate ancestor with minimal help from context. Any semantic components in the target word beyond those in the immediate ancestor, or different from them, would be trivial and predictable. For example, the word *hunter* may have some connotations of occupational role or habitual behavior beyond the simple meaning "one who hunts," but these are usually associated with the suffix *-er*, and therefore could be inferred by a reader without contextual information.

SEM 2. The meaning of the target item can be inferred from the meaning of its immediate ancestor with reasonable help from the context; "one exposure learning" would be possible. The target word may contain nontrivial semantic features different from or in addition to the semantic features in the immediate ancestor, but these would require only general contextual information to be inferred. For example, the word *gunner* means not just anyone who uses a gun, but normally is used for military personnel with the specific assignment of using or operating guns. Presumably the semantic components specifying "military personnel" would be inferrable from the general context in which the word was used, ruling out, for example, the interpretation "gunfighter."

SEM 3. The meaning of the target item includes semantic features that are not inferrable from the meaning of the immediate ancestor without substantial help from the context. For example, the meanings of the words *copper* and *head* definitely contribute to the meaning of the word *copperhead*. One could infer that it might mean "something with a head made out of copper, or resembling copper, or of the color of copper." Even with a context like "While walking through the woods I almost stepped on a copperhead," however, one could not be sure whether the object in question was a snake, an insect or spider, or perhaps some rare antique copper coin. Even a phrase such as "bitten by a copperhead" would not distinguish between snakes and spiders.

SEM 4. The meaning of the target word is related to the meaning of its immediate ancestor, but only distantly. The relationship is probably not apparent without being pointed out, and one is unlikely to guess the exact meaning of the target word if one knows only the meaning of the immediate ancestor. Examples of pairs of words with this degree of semantic relatedness are: *vicious/vice, farewell/well, motley/mottle, inertia/inert,* or *saucer/sauce*.

SEM 5. There is no discernible semantic connection: the meaning of the immediate answer is of no use in learning or remembering the meaning of the target word. Examples of such relationships are *clerical/cleric, groovy/groove, dashboard/dash*. (Remember that we are considering only relatively familiar meanings of each of these words.)

Table 3 contains some additional examples of words and their immediate ancestors illustrating each level of semantic relatedness.

A second rater trained in the semantic relatedness scale used in this study coded a 20% subsample of the derived words (including

Table 3 Target word—immediate ancestor pairs

	Target Word	Immediate Ancestor
Illustrating SEM 0	senselessly	senseless
	desegregation	desegregate
	cleverness	clever
	decentralization	decentralize
Illustrating SEM 1	elfin	elf
	misrepresent	represent
	litigant	litigate
	enthusiast	enthusiasm
	crowded	crowd
	lower-class	lower
	geneticist	genetic
	fragmentary	fragment
	sunbonnet	sun
	washcloth	wash
	various	vary
	wily	wile
Illustrating SEM 2	therapeutic	therapy
	gunner	gun
	uncountables	uncountable
	cow-hand	cow
	knowledge	know
	stringy	string
	gunnery	gun
	foglights	fog
	additional	addition
	mainly	main
	everyday	every
	theorist	theory
Illustrating SEM 3	password	pass
	collarbone	collar
	tweeter	tweet
	washroom	wash
	noblesse	noble
	pasteurize	Pasteur
	planetarium	planet
	collinear	linear
	handspring	hand
	airfoil	air
	visualize	visual
	hookworm	hook
	ominous	omen
	percentile	percent
	chloride	chlorine
	conclusive	conclusion
Illustrating SEM 4	crowbait	crow
	fender	fend
	saucer	sauce
	apartment	apart
	condescend	descend
	saucepan	sauce
	high-school	high
	artificial	artifice
	colleague	league
	impregnable	impregnate

(cont'd on p. 312)

Table 3 (cont'd)

Illustrating SEM 5	dog-days	dog
	prefix	fix
	shiftless	shift
	Burma-Shave	Burma
	peppermint	pepper
	foxtrot	trot

those derived by prefixation, derivational affixation, compounding, contraction, and certain irregular morphological relationships). Although exact agreement on the 6-point scale was not achieved, over half of the differences between the two raters involved only a one-step difference on this scale. In terms of the basic distinction between derived words coded as semantically transparent (SEM 0 - SEM 2) and semantically opaque (SEM 3 - SEM 5), there was 76.6% agreement.

This level of reliability is more than adequate as a foundation for the arguments put forward in this paper. Although the two raters did disagree in individual cases about the semantic relationship of a derived word to its component parts, there was a close agreement about the overall number of derived words that were considered semantically transparent or opaque. The means of the two raters differed by only 0.3 on the 6-point scale. This amounted to a 4.4% difference in the total number of derived words rated as transparent or opaque. Since the implications for vocabulary learning are based on the general order of magnitude of our results, only large disagreement among raters would affect the force of the points made.

Types of Words

Estimates of the total number of words in English differ not only in how words are counted—e.g., whether derived forms are counted as separate from their bases or not—but also in terms of whether certain classes of words are counted at all. The *WFB* contains various special categories of words that are often excluded from counts of words: proper names, numbers, formulae, compounds containing numbers, abbreviations, and nonwords (strings of characters that clearly do not represent vocabulary items). Each item in our sample was marked as to whether it belonged in any of these categories.

Unlike some vocabulary researchers, we did not mark words as rare, archaic, obsolete, or technical; nor did we mark the scientific names of flora or fauna. If a word actually occurs in the *WFB*, children encounter it in their school reading; we consider this a justifiable operational criterion for defining the boundaries of printed school English. Rather than trying to come up with criteria for specialized or technical vocabulary, we feel that such distinctions, if they become necessary, could be best defined operationally in terms of the actual distribution of words in the corpus.

Results

Table 4 presents the results of our study in terms of the relationship and word class categories of our coding system. For each category, five different figures are given. *Sample N* is the number of items in our sample falling into this category: *Sample %* is the percent of our sample which this category constitutes, i.e., 100 x Sample N/7,260. The *Corpus N* is the estimated number of items in this category that would be found in the whole *WFB*. The *Population N* is the number of words in the total vocabulary of printed school English (Grades 3 through 9) that would fall into this category. *Population %* is the percentage of words in this category in the population, i.e., 100 x Population *N*/609,606.[2]

Table 4 is organized as follows: First, the different coding categories are arranged approximately according to how they relate to possible definitions of "word." The first group of coding categories are those which would be counted as constituting "separate words" in many definitions of "word," and which would appear as separate entries in most dictionaries. The second group of coding categories are those that might not be considered separate words for some purposes but would often have separate entries in dictionaries. For example, *mice* might not always be considered to be a separate word from *mouse*, for the purpose of counting words, but it would occur as a separate entry in most dictionaries. The third group of

Table 4 Analysis of the word frequency book by word-relatedness categories

Category	Sample N	Sample %	Corpus N	Population %	Population N
A. Categories that would be included in most definitions of "word."					
Morphologically basic	846	11.65	10,108	7.46	45,453
Idiosyncratic relation	72	1.00	860	1.01	6,167
Suffixation	722	9.94	8,626	7.62	46,431
Prefixation	233	3.21	2,784	4.01	24,457
Compounding & contractions	1,038	14.30	12,402	17.23	105,044
Truncations	16	0.22	191	0.19	1,144
Abbreviations	12	0.17	143	0.15	897
Subtotal	2,939	40.48	35,115	37.66	229,593
B. Categories that would have their own separate entries in most dictionaries.					
Irregular inflections	49	0.67	585	0.25	1,528
Irregular comparative & superlative	1	0.01	12	0.002	13
Alternate forms of words	8	0.11	96	0.18	1,072
Alternate forms with *s*	8	0.11	96	0.11	693
Semantically irregular pl.	8	0.11	96	0.02	136
"Scientific plurals"	2	0.03	24	0.02	145
Subtotal	76	1.05	907	0.59	3,587
C. Categories that would not normally occur as separate dictionary entries.					
Regular inflections	1,553	21.39	18,555	16.37	99,547
Regular comparative & superlative	46	0.63	550	0.51	3,149
Incorrect regular inflections	3	0.04	36	0.07	450
Simple capitalization	618	8.51	7,384	8.51	51,906
Alternate spellings	136	1.87	1,625	3.05	18,584
Alternate pronunciations	87	1.20	1,039	1.21	7,381
Subtotal	2,443	33.65	29,188	29.69	181,017
D. Categories relating to proper names.					
Basic proper names	929	12.80	11,099	14.78	90,107
Derived proper names	88	1.21	1,051	1.18	7,215
Capitalization homographic with p.n.'s	76	1.05	908	0.67	4,114
Inflectional and other variants with p.n.'s	302	4.16	3,608	4.74	28,869
Subtotal	1,395	19.21	16,667	21.38	130,305
E. Categories not normally counted as words.					
Formulae & numbers	339	5.50	4,767	5.89	35,891
Compounds containing numbers	41	0.56	490	0.80	4,894
Nonwords	147	2.02	1,756	3.35	20,444
Foreign words	46	0.63	550	0.92	5,618
Subtotal	633	8.80	7,563	10.97	66,847

(cont'd on p. 314)

Table 4 (cont'd)

F. Miscellaneous categories.

Errors in WFB (duplpicated entries)	6	0.08	6	—	—
Ambiguous words (excluding proper names)	19	0.26	227	0.05	292
Ambiguous proper names	2	0.03	24	0.004	27
Missing ancestores added	203	2.80	2,425	—	—
2nd meanings of ambiguous items added	51	0.70	609	—	—

categories contains those such as regular inflections that would not normally occur as separatel items in dictionaries.

The fourth group contains categories of proper names, which are excluded from some, but not all, dictionaries and estimates of vocabulary size. Proper names were further subdivided as follows: Basic proper names are those proper names which were also categorized as morphologically basic. Derived proper names are words derived from proper names by some word-formation process, i.e., by suffixation, prefixation, compounding, or some morphologically idiosyncratic relationship. Inflectional and other variants of proper names include plurals and other variants of proper names that would not be given separate entries in a dictionary. Capitalizations homographic with proper names are those forms, such as *Cliff*, which might be either a proper name or the capitalization of a nonproper name. Since the noncapitalized form *cliff* has already been counted elsewhere, we have counted these as constituting proper names. In answer to the question "How many distinct proper names are there?", one would probably want to include all of these categories except for "inflectional and other variants of proper names."

The remaining categories in Table 4 are those which would not normally be counted as separate words or be listed as words in a dictionary. Note that the categories of special types of words—proper names, formulae and numbers, compounds containing numbers, nonwords, and foreign words—are not included in the relationship categories in the first three groups.

Even without further analysis, certain things are already clear about the estimated vocabulary of printed school English. Most importantly, it is very large,—over 200,000—words, and another 100,000 proper names. While over 170,000 words are derived by suffixation, prefixation, and compounding, there are still 45,000 which are basic, that is, which cannot be derived from any other word.

Semantic Transparency of Derived Words

In Table 5, estimates of the number of derived words in the population are broken down according to relationship type—suffixation, prefixation, compounding, and idiosyncratic relationships—and by degree of semantic relatedness. We can divide the degrees of semantic relatedness into two classes: Semantically transparent (SEM 0 - SEM 2) and semantically opaque (SEM 3 - SEM 5). From Table 5, we see that there are an estimated 139,020 semantically transparent derived forms in the population. This suggests strongly that knowledge of word-formation processes opens up vast amounts of vocabulary to the reader. There are also 43,080 derived forms that are relatively opaque semantically. The majority of these, 26,599 words, are at the level SEM 3, which means that although the meaning of the derived form is not completely predictable from the meanings of its component parts, the meanings of the component parts do in fact contribute something to the derived meaning. Even in these cases, the knowledge of word formation processes can be helpful to the reader trying to figure out the meaning of words in context. On the other hand, the semantic opacity of these words is sufficient that many readers—especially poor readers—will not be able to figure out their meanings, and thus will have to learn them individually.

Table 5 Derived words arranged by relationship category and degree of semantic relationships

	Relationship Categories				
	Suffix	Prefix	Compound	Idiosyncratic	Total
SEM 0	26,840	12,999	21,773	519	62,131
SEM 1	6,289	4,051	28,591	666	39,597
SEM 2	6,904	3,476	26,033	879	37,292
SEM 3	3,717	2,630	17,817	2,435	26,599
SEM 4	1,413	636	4,675	1,162	7,886
SEM 5	1,269	666	6,155	505	8,595
SEM 0-2	40,033	20,526	76,397	2,064	139,020
SEM 3-5	6,399	3,932	28,647	4,102	43,080

How Many Words Are There in English?

To answer the question "How many words are there in English?" one has to determine what is the appropriate definition of word to use. We feel that the best way to approach the counting of words is in terms of distinct "word" families, i.e., a group of morphologically related words such that if a person knows one member of the family, he or she will probably be able to figure out the meaning of any other member upon encountering it in text.

Counting as distinct word families all morphologically basic words and semantically opaque (SEM 3, SEM 4, and SEM 5) derived words, we have estimated that there are 88,533 distinct word families in printed school English. However, some substantial qualifications must be made before this number can be interpreted correctly.

First of all, how words are to be counted depends on why you are counting them. Our interest in estimating the number of words in printed school English is to determine the size and nature of the task that children face in learning the vocabulary of school texts. Whether we should count *understand* and *misunderstand* as one word or two depends on how children actually deal with them. If children who know the meaning of *understand* can learn the word *misunderstand*, or interpret it in context, with little or no additional effort, then we would want to count these two words as being members of a single word family.

Therefore, any criterion for counting words must be relative to some level of morphological knowledge. For this reason, a truly meaningful estimate of the number of words in printed school English will require empirical studies of children's knowledge of morphology. Our system of coding different degrees of semantic relatedness is an attempt to approximate what we believe the results of such studies would be; but it remains speculative until these coding categories can be tied to particular age and ability levels.

Our estimates of 88,533 distinct word families assumes that children in Grades 3 through 9 would not be helped much by morphological relatedness among words if the degree of semantic relatedness were SEM 3, SEM 4, or SEM 5. For example, knowing the meanings of *hook* and *worm* would not provide sufficient information for the child to guess the full meaning of *hookworm* unless the context were rich enough to give unmistakable clues for the remaining semantic components (e.g., parasitic, causing disease). Therefore, *hookworm* and similar derived forms were counted as constituting separate word families. However, if we could somehow establish that ninth-grade students were able to make use of SEM 3 relationships in learning or interpreting new word meanings, our estimate of the number of distinct word families for ninth-grade students would have to be reduced to 61,934. Conversely, if we were to find that children at a certain grade level were less adept than we expected at seeing and utilizing relationships among words, our estimate of the number of distinct word families for children at that grade level would be revised upwards.

Other categories of nonredundant words.

Another way to talk about word families is in terms of redundant versus nonredundant words. If a child who knows the meaning of *estimate* can automatically intepret or learn *overestimate*, the latter word is redundant; it does not contribute to the child's vocabulary learning task, or add to the vocabulary load of a text the child is reading. Our figure for the total number of distinct word families is supposed to reflect the number of nonredundant words in printed school English. However, there may be several types of words not included in this count which also should probably be counted as nonredundant in terms of the effort they would require to learn or interpret.

For example, abbreviations were not included in our count of distinct word families, because they do not constitute distinct words in the prototypical sense. One might consider them to be redundant in that an abbreviation has the same meaning as the word for which it stands. However, the relationship of an abbreviation to its unabbreviated form, and hence its meaning, is not at all obvious in some cases; most often, an abbreviation must be learned as a separate item. On similar grounds, one might want to include in the count of distinct word families other categories in our coding system such as truncations, irregular inflections, irregular comparatives and superlatives, some alternate forms of words, and semantically irregular plurals. For each category, it could be argued that many or most of the items were not redundant—that is, that knowledge of other, related forms would not guarantee the reader a fair chance of understanding that item when encountering it the first time in reading.

All the categories just mentioned would add only an estimated 4,935 words to the population, bringing our total vocabulary estimate up to 93,468 distinct word families. However, if we want to estimate the total number of words in printed school English in terms of nonredundant items to be learned, several other categories of items might be added which would increase this overall figure substantially.

Proper names. Both Dupuy (1974) and Lorge and Chall (1963) exclude proper names from their count of basic words. This exclusion is presumably based on the fact that proper names are functionally distinct from other vocabulary items in a number of ways. In the context of reading, it might be argued that a child only has to recognize a proper name as being such, and that any information about the individual associated with that name either will be supplied in the story itself, or should be considered knowledge about the world, and not vocabulary knowledge as such.

This is a complex issue, more so than we could do justice in the scope of this paper. One could argue, however, that there is at least a subset of proper names that should be counted as part of general vocabulary. To be sure, the names of characters are usually assigned a referent within the context of a story, so that the reader often needs little, if any, prior knowledge about that name to successfully comprehend the text. But there are some proper names which are most often not explained within texts. Lack of knowledge of familiar geographical names such as *Washington, Florida, Alaska,* or *Panama,* for example, could contribute to comprehension failure in exactly the same way that ignorance of the meaning of other words in the text might.

A related point is that the line between proper names and other areas of vocabulary—for example, names of flora and fauna, or technical terms—is not clearly defined. For example, *eagle* is counted by Dupuy as a basic word, but *Megaloceros* as a proper name. There are differences between these two words, in terms of usage and frequency, but it is not clear that these differences bear directly on the classification of an item as a common or proper noun.

Determining which or how many proper names should be included in an estimate of vocabulary size would require some more detailed work on the role of proper names in reading comprehension. A rough estimate, however, was made in the following fashion: First, a count was made of those proper names in our sample which intuitively seemed to be "important," i.e., knowledge of them would be likely to be assumed in at least a large proportion of school texts. A second count, of those proper names that were listed in the *American Heritage School Dictionary*, produced an almost identical list. Among the important proper names were *American, Boston, Canaan, December, Hanoi, Iran, Lithuania, Maine, Olympus, Passover, Rome,* and *Yugoslavia*. The unimportant names included both lesser-known geographical place names (*Amesbury, Champaign-Urbana, Muskegon*) and many personal names (*Fenwick, Jethro, Ryan, Sylvia*).

This count indicated that there would be about 1,000 important names in printed school English. This number is certainly conservative; especially in the higher grades, one would expect that an increasing number of proper names would be assumed rather than explained in school texts, and thus should be counted as part of the demands on the child's vocabulary knowledge.

Homographs. Most estimates of vocabulary size, and all of those we have been discussing, lump together all homographs. But a child who knows only the noun *bear* (= animal), when confronted with the verb *bear* (= carry) in a text for the first time, is encountering a brand new word. Knowledge of the one meaning of *bear* is no help in figuring out the new meaning. If an estimate of vocabulary size attempts to reflect the number of nonredundant items a child would have to learn, it would have to count distinct meanings of homonyms as separate items. Even related, but somewhat different, meanings of the same word may present difficulties to young readers.

An analysis of polysemy among the words in our sample was performed (Nagy & Anderson, 1982) to determine the number of distinct meanings represented in printed school English. The *American Heritage School Dictionary* was used as the primary source for determining the number of meanings for a word, since this dictionary was based on the same corpuse as was the *WFB*. Meanings were grouped according to the six levels of semantic relatedness already described.

We estimated the total number of meanings at a given level of semantic relatedness by taking the number of distinct meanings for morphologically basic words, plus the number of derived words with a meaning or meanings distinct from any meaning of their immediate ancestor. This sum is undoubtedly an underestimate, first, because the *American Heritage School Dictionary* would not contain every meaning of a word that would occur in the full range of usage for printed school English, and second, because a maximum of only one meaning was added for each semantically obscure derived word. Our finding was that if we count distinct meanings at the SEM 3 level or greater, there are more than 105,000 distinct meanings underlying printed school English.

Total count of nonredundant items. Given the 1,000 proper names and 4,000 "other" nonredundant words, the additional polysemy criterion raises our overall estimate to 110,000 distinct words in printed school English. This estimate assumes that individuals are only able to utilize SEM 0, SEM 1, and SEM 2 relationships in learning or interpreting new words. For someone who is able to utilize SEM 3 relationships as well, the number of distinct words would be 72,000.

Comparison with Other Estimates of English Vocabulary

We have made some detailed comparisons (Nagy & Anderson, 1982) of our results with several other estimates of the number of words in English in order to see what differences exist, and to determine, as far as possible, the sources of the differences.

First of all, we compared our calculations for printed school English with *Webster's Third New International Dictionary*, unabridged (1961). Dupuy (1974) made a thorough analysis of the number and types of main entries in this dictionary as a basis for his Basic Word Vocabulary Test, and determined that there were about 240,000 main entries, excluding multiple meanings of homographs and entries which consisted solely of prefixes, suffixes, or single letters other than one-letter words. If one also excludes compound entries—entries with internal spaces, which were also excluded from our count due to the way the *WFB* was compiled—there are only about 171,000 main entries in *Webster's Third*.

By translating the criteria for main-entry status in *Webster's Third* into the categories of our coding system, we determined that there would be 243,000 main entries in a dictionary based on printed school English. Thus there are slightly more words in printed school English than there are in a rather large unabridged dictionary.

One might wonder how this could be. Part of the answer lies in the fact that books in these grade levels sample from a very broad range of topics. Part of the explanation must also lie in the large number of derived words in printed school English. As Table 5 showed, there are about 139,000 semantically transparent derived words, a little more than half of which are compounds. Many of these derived forms, especially the compounds, are low-frequency words coined for specific purposes or contexts, and are not likely to be found in any dictionary. Examples of such words are *essayist-poet, European-owned, ex-florist,* and *everlengthening*. The existence of large numbers of such words in school texts makes knowledge of word-formation processes an important factor in dealing with low-frequency words.

Dupuy's estimate of the number of words in English. Dupuy (1974) undertook to construct a vocabulary test that would provide a meaningful estimate of an individual's absolute vocabulary size. Any measure of absolute vocabulary size presupposes a definition of "word;" Dupuy based his test on the construct "Basic Words." Dupuy estimated the total number of Basic Words in English by analyzing a 1% sample of *Webster's Third*, and excluding from his count of Basic Words main entries that fell into any of the following categories:

(1) compound and hyphenated entries,
(2) proper names,
(3) abbreviations,
(4) items which are not main entries in three other dictionaries: *The Random House Dictionary of the English Language* (*1966*), *The World Book Dictionary* (*1969*), and *Funk and Wagnalls New Standard Dictionary of the English Language* (*1965*);
(5) items listed as foreign, archaic, slang or informal, or technical in the *Random House Dictionary;* and
(6) "derived, variant, or redundant" words.

Dupuy's criteria for defining Basic Words, are in most points similar to our definition for distinct word families; yet Dupuy estimated that there are only 12,300 Basic Words in English, whereas our results indicate that there are around 88,500 distinct word families.

Most of the difference in these estimates stems from the fact that Dupuy excluded from his count of Basic Words all main entries in *Webster's Third* that did not appear as main entries in three other large dictionaries. Some of the items excluded on this basis could have been excluded on other grounds as well, but a substantial number apparently failed to make it inot one or more of the three dictionaries simply because of their low frequency. Even fairly large dictionaries cannot contain the full lexical resourses of the language, and we cannot be sure that the inclusion or exclusion of low frequency items in a dictionary was based on linguistically or pedagogically relevant criteria, or, for that matter, that it consistently followed any principled criteria at all. Therefore, Dupuy has not succeeded in finding a non-arbitrary way of dealing with low frequency words, and has, in fact, excluded many which would actually appear in printed school materials.

Another source of difference between our results and Dupuy's is that Dupuy placed greater weight on morphological relatedness than we did, and considered words to be redundant which we would judge to have only distant semantic relationships to their base forms. For example, Dupuy excluded *coloratura* from his count of basic words, presumably because it is related to the word *color*. A semantic relationship does exist, but one so tenuous that it would give a reader no help at all in trying to infer the meaning of the derived form. A final source of difference is that the total vocabulary of *Webster's Third* is somewhat smaller than that of printed school English.

Seashore and Eckerson's estimate. Like Dupuy, Seashore and Eckerson (1940) attempted to construct a test which would measure not only relative vocabulary knowledge, but also give an indication of the absolute size of an individual's vocabulary. Their test has been the basis for much subsequent research in vocabulary (e.g., Smith, 1941; Templin, 1957). They also used the method of selecting a random sample of items from an unabridged dictionary, taking as their population of words the entries in *Funk and Wagnalls New Standard Dictionary of the English Language* (two volumes edition of 1937). This dictionary was chosen because it was large enough to represent the full range of adult vocabulary without including extremely rare words.

This dictionary contains two types of entries: "Basic" words, or main entries, printed in heavier type and next to the left margin, and "derivative" terms, which are indented under the basic term. Seashore and Eckerson estimated that the dictionary contains 166,247 basic words, and an additional 204,018 derivative words, excluding multiple meanings and variants in spelling. Unlike Dupuy, Seashore and Eckerson did not further categorize the entries. A major difference is that they included proper names in their estimate of total English vocabulary.

We attempted to translate Funk and Wagnall's criteria for basic and derivative entry status into the categories of our coding system. On this basis, printed school English appears to have more basic entries than the dictionary (192,000 versus 166,000), but somewhat fewer total words (344,000 versus 370,000). In general, though, we get the same picture as we did from our comparison with *Webster's Third*: The number of words in printed school English is of the same order of magnitude as a fairly large unabridged dictionary.

Several problems arise as to the interpretation of Seashore and Eckerson's estimates. First, as is true for Dupuy's estimates, it is questionable whether choices made by dictionary editors about the inclusion and exclusion of words can serve as the basis for research in vocabulary. This is especially problematic in the investigation of absolute vocabulary size, where low frequency words play such a crucial role.

Second, the distinction between basic and derived entries in the Funk and Wagnalls dictionary is inconsistently drawn. To some extent, this distinction can be stated in terms of word formation processes; derivative entries are words derived from their basic entries by suffixation or compounding. Seashore and Eckerson give the example of the basic word *loyal* and its derivatives *Loyal Legion, loyalism, loyalize,* and *loyally*. However, there are numerous exceptions to this principle. There are many semantically transparent derivatives (e.g., *evaporation*) that appear as separate main entries, and also semantically opaque derivatives (e.g., *stayplow*) that appear as derived entries. In many cases, whether a compound is listed as a basic or derived entry is determined by alphabetical order. Furthermore, because prefixed forms are alphabetically removed from their bases, they always constitute separate main entries.

Third, Seashore and Eckerson's space-sampling method has been shown to result in a disproportionate number of high-frequency words being included in the test. Also, test scores would be inflated by the fact that different meanings of homographs were counted as separate words in the estimate of the number of words in the language, whereas in the vocabulary test, a word is counted as known if any common meaning is recognized. These and other problems are discussed at length in Lorge and Chall (1963).

Thus, the two major attempts to estimate the number of words in English are seriously flawed. Dupuy's estimate is based on unrealistic assumptions about morphological and semantic relatedness, leading to a substantial underestimation of the number of distinct words in English. Seashore and Eckerson's estimate, while in the right ballpark, suffers from several other weaknesses which make it impossible to have much confidence in scores on their vocabulary test.

The Distribution of Words by Frequency

So far, we have shown that printed school English includes at least as many words as a fairly large unabridged dictionary. It would be useful to know how many of these words students will actually encounter as they pass through Grades 3 through 9 as well as

Table 6 Cumulative distribution of words by frequency

Frequency (in terms of *U*)	Number of Words in Printed School English At or Above that Frequency		
	Graphically Distinct Types	Morphologically Basic Words and Semantically Opaque Derivatives	Semantically Transparent Derivatives
100.00	890	555	55
31.623	2,305	1,225	175
10.000	5,480	2,450	455
3.1623	11,980	4,330	1,290
1.0000	24,108	6,700	3,300
.31623	44,743	10,400	7,150
.10000	76,757	15,350	13,400
.03162	122,045	21,700	23,000
.00132	304,803	46,300	65,000
.00003	512,886	75,000	116,000
0.0000	609,606	88,500	139,000

how many of them would actually be useful.

One way to approach this question is to look at word frequency distributions. Table 6 shows how the words in printed school English are distributed by frequency. Frequencies are given in terms of *U*, or estimated frequency per million words of text. A word with *U* = 10.0, for example, would be expected to occur on the average about ten times in a million words of text. Details of how *U* is calculated are found in the *WFB* (p. xl).

The number of graphically distinct types with a frequency equal to or greater than a given value are interpolated from tables in the *WFB*. These numbers are predicted on the basis of the lognormal model; according to this model, if frequencies are expressed logarithmically, words will be found to occur in a normal distribution along the frequency continuum.[3]

At least two things are clear about the distribution of words by frequency. First of all, most words are in the lower ranges of the frequency spectrum. About half the words in printed school English, no matter how one counts them, occur roughly once in a billion words of text or less. Second, semantically transparent derivatives are skewed towards the low end of the frequency distribution to a greater degree than are morphologically basic words and semantically opaque derivatives. Among the most frequent words, semantically transparent derivatives are relatively rare. As one moves downward in frequency, however, the proportion of semantically transparent derivatives increases constantly, until the total number of semantically transparent derivatives is greater than the number of morphologically basic words and semantically opaque derivatives.

This difference in distribution has some distinct implications for instruction. If a child were exposed only to vocabulary controlled carefully by frequency, there would be both relatively little opportunity to learn, and little necessity to make use of, the word-formation processes that relate derived words to their component parts. The relatively few transparent derived words that do occur in the higher frequency ranges are likely to be learned, at least at first, as unanalyzed wholes (cf. Kuczaj, 1977; Silvestri & Silvestri, 1977). On the other hand, it is clear that as one's exposure to the language expands into the lower frequency ranges, knowledge of word-formation processes becomes an increasingly necessary skill.

One might be tempted to argue that words occurring once in a million words of text are really not worth much consideration. If the student encounters such words on the average of once a year or less (for any individual word), there would not seem to be a need to include them in any program of vocabulary instruction. But before jumping to any conclusions about words in the lower ranges of the frequency continuum, it might be useful to look at what words are actually involved. Many of them appear to be of little general use, but there are some seemingly useful words there as well. Among the words occurring less than once in 100 million words of text ($U = 0.008$) are ones such as:

amnesty
appall
assimilate
busybody
cheeseburger
contemporary
eczema
elevate
evict
expound
flex
fluent
fume
furor
gnome
hornswoggle
ignoramus
jellybean
liturgy
mediate
papaya
persecute
racoon
rambunctious
rote
shamrock
stenographer
syncopate

Among the even rarer words, occurring less than three times in a billion words of text ($U = 0.0025$) are:

ammeter
cyanide
solenoid
calliope
nettle
spinnaker
anneal
deform
template
emanate
pidgin
fresco
billfold
hex
unwieldy
extinguish
saturate
inflate
cloverleaf
orthographic
ventilate
flippant
seagull
sacrament

This is not a representative sample of low-frequency words, to be sure, but these examples do demonstrate that just because a word has a relatively low frequency in printed school English does not mean that it is not worth learning.

Since a word's frequency does correlate with the probability that an individual will know that word, it is easy to confuse frequency with difficulty. But almost any book by Dr. Seuss will serve as proof that utterly novel words do not necessarily disrupt comprehension. Yet many such words occur only once in a single story, and thus would have astronomically low frequencies in any large scale survey of word frequency.

The frequency of a word reflects a number of factors; one of them is often the conceptual difficulty of the word. But in general, it might be said that a word's frequency reflects the range of contexts in which the word might appear. A "rare" word such as *sacrament* is important within a certain set of contexts, but this set of contexts is very small compared to the universe of contexts that are covered in printed school English.

It should also be noted that frequency studies such as the *WFB* that involve very large samples of written language are not representative of an individual student's exposure to the language. Because choice of words will be more consistent within a given author's works or a given subject category, any individual student will not get a random sample of vocabulary containing a wide range of low frequency words occurring once each. Rather, in a given student's reading, most low frequency words will not occur at all, and of those that do, many may occur a number of times.

There is an important sense in which the frequencies listed in the *WFB* underestimate the true frequency of occurrence for a given word. A student's exposure to the word *drive*, for example, is not a function of the frequency of that graphically distinct type alone, but rather, a function of the sum of the frequencies of all members of the family. In this case, one would certainly want to include forms such as *Drive, driven, driver, Driver, driver's, Driver's, drivers, drivers', drives,* and *drove*. The frequency of this entire family is over three times greater than the frequency of the morphologically basic word *drive*. This particular family is more extensive than many, but family frequency is necessarily greater than or equal to the frequency of any individual member. In this sense, students may encounter some of the low-frequency words in printed school English more often than one would gather from the frequencies reported in the *WFB*.

Finally, it should be noted that the materials on which the *WFB* is based tend to have a higher proportion of high frequency words than does printed matter written for adults. This means that the frequencies reported for rare words in the *WFB* will in general be lower than the reported frequencies for the same words in adult materials.

The distribution of words by frequency shows that a large portion of the many words in the vocabulary of printed school English have very low frequencies. Nevertheless, one must be careful in interpreting this fact. It would be a mistake to suppose, for example, that all words occurring once in a million words of text were so technical or specialized as to be of no pedagogical significance.

How Many Different Words Do Children Actually Encounter

To get an accurate picture of the vocabulary that students actually encounter in printed school materials would require both information on the amount and type of reading done by children in and out of school, and a reanalysis of our data by grade level. Our plans for future research include both these steps; at present, however, we can get an approximate idea of the number of words students have to deal with in school reading. At the lower end of the spectrum, one might imagine a less able reader at one of the lower grade levels reading as few as ten pages a day from books with large print and frequent pictures, averaging 100 words per page. If this rate were maintained through 100 days of the school year, 100,000 running words of text would be covered. On the other hand, an average reader in seventh grade might spend 50 minutes a school day in actual reading, at a rate of 100 to 200 words per minute. In 100 school days, 500,000 to 1,000,000 running words of text would be covered. This is certainly not a maximum; given a higher reading speed, a little more time spent in reading, and more consistent reading during the year, and a child might cover 10,000,000 running words.

The foregoing estimates may be conservative. Carroll (1964) has conjectured that college students may be exposed to as many as a million running words a week in their reading, lectures, and conversations. Our own conjecture is that there are avid readers from the middle grades who approach this figure.

From the *WFB* (see Table B-9, p. xxxvii), it appears that a student who goes through Grades 3 through 9 reading 500,000 to 1,000,000 running words of text a year will be exposed to between 20,000 and 40,000 graphically distinct types. From our analyses of the *WFB*, this would mean that somewhere between 4,000 and 10,000 distinct word families might be encountered. More precise estimates will require analysis of our data by individual grade levels. In the meantime, we can be fairly confident that an average reader in Grades 6 through 9 would encounter at least 5,000 distinct word families in a year (perhaps as many as 10,000). At least 1,000 of these would be families that had not been encountered in the previous year, and it is quite possible that an active reader in these grades could come across 3,000 or 4,000 totally new vocabulary items in the course of a school year. These rough estimates demonstrate that direct instruction could not cover more than a small fraction of the words that a student will actually encounter in school reading.

Word Families in School English

How much interrelatedness is there among words in printed school English? One way to approach this question is in terms of the size of the average word family. If there are 609,606 graphically distinct types in printed school English, and only 88,533 distinct word families, one would expect there to be 6.88 members per family. This figure is inaccurate, however, because there are several kinds of words (e.g., numbers and proper names) which are not included in any family at all.

Table 7 represents the average composition of a word family in printed school English. Since the concept "word family" can be defined only with respect to some level of morphological ability, we have decided to give figures based on two different definitions.

Definition A adopts a conservative estimate of the number of distinct word families in printed school English. Assuming, in this case, that some individuals might make effective use of even SEM 3 and SEM 4 relatedness in learning derived words, we count as distinct word families only morphologically basic words and derived words with a semantic relatedness level of SEM 5. By this definition, there are about 54,000 distinct word families. Since people frequently learn

Table 7 The average composition of a word family

Number of Words		
Definition A	Definition B	Type of Words
1.00	1.00	Base word (a morphologically basic word or semantically opaque derivative)
.15	---	SEM 4 derivatives
.49	---	SEM 3 derivatives
.65	---	Total semantically obscure derivatives (SEM 3, SEM 4)
.69	.42	SEM 2 derivatives
.73	.45	SEM 1 derivatives
1.15	.70	SEM 0 derivatives
2.57	1.57	Total semantically transparent derivatives (SEM 0-SEM 2)
.04	.02	Truncations and abbreviations
.07	.02	Irregular inflecions, comparatives and superlatives; alternate forms of words; semantically irregular plurals
1.90	1.16	Regular inflections, comparatives and superlatives
2.00	1.22	Total inflections, abbreviations and truncations
.94	.58	Simple capitalizations
.34	.21	Alternate spellings
.14	.08	Alternate pronunciations
1.42	.87	Total minor variations in form
7.64	4.66	Total family size in graphically distinct types

words without perceiving relationships that do exist between them (e.g., *basement* and *base*), we consider this to be an underestimate of the true number of distinct word families; however, it can serve as a reasonable lower limit.

Definition B is the definition of word family we have adopted up to now; it includes morphologically basic words and derivatives at levels SEM 3, SEM 4, and SEM 5. By this definition, there are around 88,500 distinct word families. This is by no means an upper limit; as discussed above, the number would be raised considerably if, for example, distinct meanings were counted as separate word families, or if even a small portion of proper names were included. But given that we want a figure comparible to Definition A (i.e., one that excludes proper names and does not consider problems of polysemy) this can be taken as our best estimate of the number of distinct word families.

Table 7 shows that for each word known most people will readily interpret .87 to 1.42 words that differ only in minor details of form, and from 1.16 to 1.90 words which are inflections of the base word. It can also be seen that in the average word family, for each base word, there are between 1.57 and 2.57 additional semantically transparent derivatives. For the child who is able to make use of SEM 3 and SEM 4 derivatives, for each word learned there are more than three derived words with meanings recognizably related to that of the base, and at least two of these involve fairly transparent relationships. This demonstrates that the ability to utilize morphological relatedness among words puts a student at a distinct advantage in dealing with unfamiliar words.

Implications

Programs of Vocabulary Instruction

Our basic finding is that when a psycholinguistically and pedogogically justifiable way of counting words is employed, the number of words in printed school English is extremely large. Advocates of direct vocabulary instruction have leaned heavily on the assumption that the number of distinct words in school English is small, that unaided year to year growth in vocabulary is modest, and that the total number of word meanings known by a typical child at any age is not large. Notably, Becker, Dixon, and Anderson-Inman (1980), accepting Dupuy's estimate that the average high school senior knows approximately 7,800 words, have attempted to lay out a program of systematic instruction for a core vocabulary of 8,000 words.

A program of systematic instruction for a core vocabulary of 8,000 words is not necesssarily a bad idea. As Table 6 shows, if 8,000 words were correctly chosen, they could cover all distinct word families found among words that occur at least once in a million words of text. But the theoretical foundation of this program—taking Dupuy's Basic Words as a benchmark for the number of items to be learned—is questionable.

Our findings suugest that high school students may actually know far more words, perhaps somewhere between 25,000 and 50,000, or even more. Dupuy (1974) estimates that third-grade students know only 2,000 words, but estimates by others are higher. Cuff (1930) places third grade vocabularies at around 7,425 words, and Smith (1941), using vocabulary tests based on Seashore and Eckerson (1940), set the figure at 25,000 basic words. It is quite possible, then, that the average third-grade student already knows 8,000 words.

There is reason to worry that Becker, Dixon, and Anderson-Inman did not find the right set of 8,000 words and that they made unreasonable assumptions about semantic relatedness. They culled their set of 8,000 words from a list of 26,000 based on the Thorndike and Lorge (1944) *Teacher's Word Book of 30,000 words*, with some adjustments to bring the list up to date. The list of 26,000 "object words" was collapsed to 8,000 "root words," where a root word was defined as "the smallest word from with the other words can be semantically derived. . . . In designating a root word for any given object word a search was made for the *smallest word within the object word that contains the core meaning of the object word*" (p. 21; emphasis in the original). The assignment of root words was frequently the same as in the present analysis; for example, the root word of *helpless* was *help*. However, in our judgment, Becker and his associates often stretched the criteria of semantic and morphological relatedness beyond reason. For example, all of the following words were assigned the root word *judge* on the basis of their semantic relatedness: *juror, juridical, jurisdiction, jurisprudence, jury, judicious, judicature, prejudice, prejudicial, unprejudiced, judicial, judiciary, judge,* and *judgment.*

The problem with this grouping is the assumption that direct instruction on the root words and on affixes would automatically result in a child knowing the meanings of the whole set of words. Becker, Dixon and Anderson-Inman (1980) admit that "providing systematic instruction for even 8,000 root words is a monumental undertaking" (p. 7). We consider it even more monumental for a student, having been taught only the meaning of *judge*, to be able to identify what words were in fact related to it, and then to figure out their meanings. How could a child encountering words such as *Judaic, judicious, judo, juggernaut, juggle, jugular, Julian, junta,* and *jury* for the first time in text, know which were historically related to *judge*? Furthermore, the most important part of the meaning of a word such as *jury* is not what it has in common with the root word *judge* (this much of its meaning would probably be pretty obvious from the context), but how it differs from it. Furthermore, since the root words were usually chosen to be one of the more frequent members of a set of related words, it may well be that children already know many or most of the 8,000 root words, and that it is the "derived" words such as *judicial, jury,* and *judiciary*, rather than root words like *judge*, for which they really need instruction.

Of course, many of the derived words were in fact transparently related to their root words. But because no distinctions were made among different degrees of relatedness or different types of relatedness, Becker and his colleagues underestimated the number of words that are functionally distinct as far as vocabulary learning is concerned.

Beck, McCaslin, and McKeown (1980)

have formulated an intensive program of vocabulary instruction which has as a major aim increasing student's reading comprehension. One motivation for their program was the unexpected failure of several previous experimental studies to produce significant increases in reading comprehension via vocabulary instruction (e.g., Jenkins, Pany, & Schreck, 1978). Beck and her associates hypothesize that vocabulary instruction can facilitate reading comprehension only if the words are learned thoroughly—to the point where the word's meaning can be accessed quickly or automatically, and where a fairly rich network of semantic connections between that word and others has been developed. Because of this, their program involved repeated exposure to words. Children in their study were exposed to each word 10-18 times in a variety of tasks. There was also a subset of words in their study which were repeated 26-40 times, to see if the additional repetition would result in even greater learning. Results from an application of this program in a fourth-grade classroom are described in detail in Beck, Perfetti, and McKeown (1982). Even with the intensive instruction and repetition, children learned 77.6% of the words that were repeated 10-18 times, and 86.5% of the words repeated 26-40 times. So it does not appear that the program was unnecessarily repetitive.

How much ground did the program cover? Just 104 words were taught over a 5-month period, with one half hour per day devoted exclusively to this vocabulary program. At this rate, 208 words could be covered in a school year. If the program were streamlined by having all words repeated only 10-18 times (that is, dropping the extra repetition of the special subset of words), one might be able to cover a little over 400 words per year. Note that Becker, Dixon, and Anderson-Inman's program to cover 8,000 words in 10 years would have to progress at twice this rate, either by spending more total time on vocabulary or less time on each word.

According to Beck, McCaslin, and McKeown (1980) it takes "an extended series of fairly intensive exposures [to a word]. . .before it can be quickly accessed and applied in appropriate contexts" (p. 8). It may well be, of course, that automaticity of access is the key factor in the relationship of word knowledge to reading comprehension; but the puzzle that must be solved by those who propose to produce automatically using word drills is how to do it in the available time, not just 400-500 words per year but for thousands of new words that children actually encounter in their school reading.

The schools have never had programs of vocabulary instruction as extensive as that proposed by Becker or as intensive as that proposed by Beck. The question that naturally arises is, up to now, how have readers acquired their vocabulary knowledge.

Vocabulary Acquisition

A basic implication of our study is that, because of the sheer volume of vocabulary that students will encounter in reading, any approach to vocabulary instruction must include some methods or activities that will increase children's ability to learn words on their own. Any attempt to do this could be based on one or more of three possible emphases: Motivation, inferring word meanings from word parts (morphology), and inferring word meanings from context.

There is basically no experimental literature that confirms the success of any of these in facilitating children's learning of words on their own. But we can speculate on the implications of our findings as to the effectiveness of such approaches.

Motivation. Motivation may be as important as any other aspect of vocabulary instruction. To quote from Petty, Herold, and Stoll (1968),

> [M]any researchers considering vocabulary development pass over motivation without mention. No classroom teacher genuinely attempting to teach vocabulary makes that mistake. . . .[T]eachers reporting on favorite techniques begin with discussions of how student interest in word study was created. (p. 19)

Beck's program does include a strong motivational component. For instance, some of the learning activities take the form of competitive games, and there are incentives for children to report instances of instructed words they found outside the classroom. Beck and her colleagues feel motivation may be a reason for the apparent increase in the experimental children's performance on tests of words not covered in the instruction. However, this effect may have been the result of improved vocabulary test taking skills, or an artifact of experimental design, rather than a real generalized increase in word learning.[4] A replication of this study (McKeown, Beck, Omanson, & Perfetti, 1983) failed to produce a significant increase in experimental children's knowledge of noninstructed words.

Morphology. Our findings suggest an important role for morphology in the learning of vocabulary. Semantically transparent derived words are relatively rare among the most frequent words, but constitute an increasingly greater proportion of the vocabulary as one goes towards the lower end of the frequency continuum. For this reason, frequency cannot be the only criterion by which words are chosen for vocabulary instruction. If the students only encountered words of fairly high frequency, there would be little opportunity to learn the productive word-formation processes in English that constitute the key to understanding the bulk of lower-frequency words.

The introduction of new words should be determined by family relationships as well as by frequency (cf. Dale, O'Rourke, & Bamman, 1971). For example, *drama* and *dramatic* are fairly frequent words (with *U*'s of 11 and 18, respectively), but the other derivative forms are fairly rare in printed texts, e.g., *dramatist* (*U* = .02), *dramatize* (*U* = .40), and *dramatization* (*U* = .50). Teaching words together as a family has a number of advantages. First, if the most frequent words in the family are already known, this procedure builds a bridge from familiar to new. In any case, once the meanings of *drama* were instructed, the meanings of the derivatives could be covered with little additional effort. What additional time is devoted to the derivatives would also function to reinforce the learning of the base word.

Another benefit of teaching words in families would be to call the students' attention to the word-formation processes that relate the different members of the family, so that they would be more likely to take advantage of such relationships when learning other words on their own. In addition, covering a family of words would familiarize students with the types of changes in meaning that often occur between related words, thus preparing them to deal with cases in which the semantic relationships among morphologically related words are not so transparent.

It should be remembered, however, that our definition of word family is based on relationships among existing words in English, not on historical roots, and on semantic relationships that are transparent enough for students to perceive on their own. We remain highly skeptical of approaches to vocabulary that proceed on an etymological or historical approach to word meanings, approaches which feign that words such as *dialect, collect,* and *intellect* have some basic meaning in common. There may be some perceptual or mnemonic value to analyzing words into historically-based components, but this remains to be established. Shepherd (19[illegible]) found that knowledge of Latin roots (e.g., *-ceive, lect*) is not strongly related to the knowledge of the meanings of words containing such roots (e.g., *receive, collect*), whereas knowledge of stems which themselves are English words (e.g., *sane*) is strongly related to knowledge of the meanings of related derived forms (e.g., *sanity*).

Context. That word meanings are learned from context is an inescapable fact. It is hard to conceive how words such as *fi, and* or *the*, for example, could be learned in any other way than from verbal context. Pointing to something in the world that corresponds to the concept of hypotheticality would be difficult to say the least, and any child old enough to understand a noncircular definition of *if* is surely already able to use the word fluently.

Many older students and almost all educated adults would be able to read through any school materials for Grade [illegible] through 9 with a high level of comprehension. This presumably would require knowing a large proportion of 88,500 distinct word families. This many words could not be acquired from direct instruction nor from looking them up in a dictionary. There is only one other possible source of knowledge: inference based on context. Thus, logic forces the conclusion that successful readers must learn large numbers of words from context, in most cases on the basis of only a few encounters.

Good readers may acquire large vocabularies because they are better at inferring word meanings from context. One indication of this is the fact that a cloze test is a satisfactory measure of reading ability. While a cloze test is taken as indicating overall reading ability, the skill it measures most directly is the ability to use contextual information to supply the meanings of words missing from text—a task analogous to that of identifying the meanings of unknown words in context.

Knowledge of morphological relatedness among words probably makes an important contribution to learning word meanings from context. Indeed, Anderson and Freebody (1983) have shown that good readers in the middle grades aggressively apply morphological principles in attempting to deduce the meanings of unfamiliar words. Our findings here show that a large number of infrequent words are transparent derivatives of other words, in many cases of words the student is likely to know already. While context often is not sufficient to determine the meaning of an unfamiliar word, it may provide enough information to permit a guess at the appropriate meaning of a word whose semantic content is partially determined by its morphology. A child who knows the meaning of *drama* and the function of the suffix *-ist* will need only minimal help from context to determine the meaning of *dramatist*. A hypothesis that should be explored in future research is that joint utilization of contextual and morphological information is a strategy employed by children who develop large vocabularies.

The only thing problematical about attributing an important role to context in the acquisition of vocabulary is that experimental studies have often indicated that children do not seem to learn word meanings very well from context. For example, Jenkins, Pany, and Schreck (1978) found that exposure to words in written context produced little increase in knowledge of their meanings, and no measurable increase in the comprehension of text containing those words. There are at least three ways to account for these findings. First, there is reason to doubt whether the contexts used in this and other experiments were really suitable for learning the meanings of the new words. Second, as Jenkins and his colleagues point out, it may be that readers can encounter a substantial number of unfamiliar words in a text with little loss of comprehension (see also Freebody & Anderson, 1981). A third way to reconcile the apparent importance of learning from context with its poor showing in experimental studies is by recognizing it as a gradual process. Experiments investigating learning word meanings from context have often used only one or a few encounters with a word, whereas the learning of many types of words from context may require multiple exposures in a variety of contexts (Deighton, 1959). Whatever the explanation, the failure to find experimental evidence for contextual learning of word meanings ought to be regarded as a conundrum for experimentalists rather than the basis for educational policy.

We hypothesize that the principal force driving vocabulary growth is volume of experience with language. Oral language experience is important, of course, particularly for the young child; oral language allows for interaction and feedback, and is itself embedded in a rich context of extra-linguistic information. Oral language typically contains a lower proportion of difficult or low-frequency words than written language. While this may make it easier to learn those new words that do occur, it also diminishes the number of exposures to new vocabulary items. We judge that beginning in about the third grade, the major determinant of vocabulary growth is amount of free reading.

It is a surprising fact that there are no satisfactory estimates of the number of words read per year by children of different ages. Earlier we guessed that the least able and motivated children in the middle grades might read 100,000 words a year while average children at this level might read 1,000,000. The figure for the voracious middle grade reader might be 10,000,000 or even as high as 50,000,000. If these guesses are anywhere near the mark, there are staggering individual differences in volume of language experience, and therefore, opportunity to learn new words. Notice also that variation of this magnitude could readily explain differences between good and poor readers in automaticity of word access.

In conclusion, any program of direct vocabulary instruction ought to be conceived in full recognition that it can cover only a small fraction of the words that children need to know. Trying to expand children's vocabularies by teaching them words one by one, ten by ten, or even hundred by hundred would appear to be an exercise in futility. Vocabulary instruction ought, instead, to teach skills and strategies that would help children become independent word learners. The challenge to those who would advocate spending valuable instructional time with individual words is to demonstrate that such instruction will give the child an advantage in dealing with the ocean of words not instructed.

REFERENCES

THE AMERICAN HERITAGE SCHOOL DICTIONARY (1977). Boston: Houghton Mifflin.

ANDERSON, R.C., & FREEBODY, P. (1981). Vocabulary knowledge. In J.T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (77-117). Newark, DE: International Reading Association.

ANDERSON, R.C., & FREEBODY, P. (in press). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutton (Ed.), *Advances in reading language research, a research annual.* Greenwich, CT: JAI Press.

ARONOFF, M. (1976). *Word formation in generative grammar.* Cambridge, MA: M.I.T. Press.

BECK, I., MCCASLIN, E., & MCKEOWN, M. (1980). *The rationale and design of a program to teach vocabulary to fourth-grade students.* Pittsburgh: University of Pittsburgh, Learning Research and Development Center.

BECK, I., PERFETTI, C., & MCKEOWN, M. (1982). The effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74,* 506-521.

BECKER, W., DIXON, R., & ANDERSON-INMAN, L. (1980). *Morphographic and root word analysis of 26,000 high frequency words* (Tech. Rep. 1980-1). Eugene, OR: University of Oregon Follow Through Project.

BERKO, J. (1958). The child's learning of English morphology. *Word, 14,* 150-177.

CAMPBELL, D.T., & BORUCH, R.F. (1975). How regression artifacts can mistakenly make compensatory education look harmful. In C.A. Bennett & A.A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs* (195-296). New York: Academic Press.

CARROLL, J.B. (1964). *Language and thought.* Englewood Cliffs, NJ: Prentice-Hall.

CARROLL, J.B., DAVIES, P., & RICHMAN, B. (1971). *The American Heritage word frequency book.* Boston: Houghton Mifflin.

CUFF, N.B. (1930). Vocabulary test. *Journal of Educational Psychology, 21,* 212-220.

DALE, E., O'ROURKE, J., & BAMMAN, H. (1971). *Techniques of teaching vocabulary.* Palo Alto, CA: Field Educational Publications.

DEIGHTON, L.C. (1959). *Vocabulary development in the classroom.* New York: Bureau of Publications, Teachers College, Columbia University.

DUPUY, H.P. (1974). *The rationale, development and standardization of a basic word vocabulary test.* Washington, DC: U.S. Government Printing Office. (DHEW Publication No. HRA 74-1334)

FREEBODY, P., & ANDERSON, R.C. (1981). *Effects of differing proportions and locations of difficult vocabulary on text comprehension* (Tech. Rep. No. 202). Urbana: University of Illinois, Center for the Study of Reading. (ERIC Document Reproduction Service No. ED 201 992)

FUNK AND WAGNALLS NEW STANDARD DICTIONARY OF THE ENGLISH LANGUAGE. (1937). New York: Funk & Wagnalls Co., 2 Vol. unabridged edition.

FUNK AND WAGNALLS NEW STANDARD DICTIONARY OF THE ENGLISH LANGUAGE. (1965). New York: Funk and Wagnalls Co.

HARTMAN, G.W. (1941). A critique of the common method of estimating vocabulary size, together with some data on the absolute word knowledge of educated adults. *Journal of Educational Psychology, 32,* 351-358.

HARWOOD, F.W., & WRIGHT, A.M. (1956). Statistical study of English word formation. *Language, 32,* 260-273.

JENKINS, J.R., PANY, D., & SCHRECK, J. (1978). *Vocabulary and reading comprehension: Instructional effects* (Tech. Rep. No. 100). Urbana: University of Illinois, Center for the Study of Reading. (ERIC Document Reproduction Service No. ED 160 999)

KUCZAJ, S.A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior, 16,* 589-600.

LORGE, I., & CHALL, J. (1963). Estimating the size of vocabularies of children and adults: An analysis of methodological issues. *Journal of Experimental Education, 32,* 147-157.

MCKEOWN, M.G., BECK, I.L., OMANSON, R.C., & PERFETTI, C.A. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Reading Behavior, 15,* 3-18.

NAGY, W.E., & ANDERSON, R.C. (1982). *The number of words in printed school English* (Tech. Rep. No. 253). Urbana: University of Illinois, Center for the Study of Reading.

PETTY, W.T., HEROLD, C.P., & STOLL, E. (1968). *The state of knowledge about the teaching of vocabulary.* Urbana, IL: National Council of Teachers of English.

RANDOM HOUSE DICTIONARY OF THE ENGLISH LANGUAGE. (1966). New York: Random House.

RHODE, M., & CRONNELL, B. (1977). *Compilation of a communication skills lexicon coded with linguistic information* (Tech. Rep. No. 58). Los Alamitos, CA: SWRL Educational Research and Development.

SEASHORE, R.H., & ECKERSON, L.D. (1940). The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology, 31,* 14-38.

SHEPHERD, J.F. (1974). Research on the relationship between meanings of morphemes and the meanings of derivatives. In P.L. Nacke (Ed.), *23rd N.R.C. Yearbook* (115-119). Clemson, SC: National Reading Conference.

SILVESTRI, S., & SILVESTRI, R. (1977). Developmental analysis of the acquisition of compound words. *Language, Speech, and Hearing Services in the School, 8,* 217-221.

SMITH, M.K. (1941). Measurement of the size of general English vocabulary through the elementary grades and high school. *Genetic Psychology Monographs, 24,* 311-345.

STAUFFER, R.G. (1942). A study of prefixes in the Thorndike list to establish a list of prefixes that should be taught in the elementary school. *Journal of Educational Research, 35,* 453-458.

TEMPLIN, M.C., (1957). Certain language skills in children: Their development and interrelationships. *The Institute of Child Welfare, Monograph Series No. 26.* Minneapolis: University of Minnesota.

THORNDIKE, E.L. (1921). *The teacher's word book of 10,000 words.* New York: Teachers College Press.

THORNDIKE, E.L. (1941). *The teaching of English suffixes.* New York: Teachers College Press.

THORNDIKE, E.L., & LORGE, I. (1941). *The teacher's word book of 30,000 words.* New York: Teachers College Press.

WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY. (1961). Springfield, MA: Merriam Co.

THE WORLD BOOK DICTIONARY. (1969). Chicago: Chicago Field Enterprises Educational Corp.

Footnotes

The research reported herein was supported by the National Institute of Education under Contract No. US-NIE-C-400-76-0116.

[1] In the majority of cases, the identity of the immediate ancestor is not problematic. For an inflected form, e.g., *adds*, the immediate ancestor is the uninflected stem or infinitive, *add*. For plurals, the immediate ancestor is the singular. For forms with a prefix, e.g., *unknown*, the immediate ancestor is the unprefixed form, *known*. For forms with a derivational suffix, *additional*, the immediate ancestor is the form without the suffix, *addition*. (Note that the immediate ancestor is found by removing only *one* suffix). For compounds, e.g., *addition-subtraction*, there are two immediate ancestors, one for each part, in this case, *addition* and *subtraction*.

In some relationships, for example, that between *multiple* and the verb *multiply*, it is difficult to say which item is more "basic" than the other. We recognize the dangers and complications of saying that one word is "derived from" another. For the purposes of analysing patterns of interrelatedness among the words in the corpus, it is necessary to break down the relationships into assymetrical dyads; however, we assign no theoretical weight to the directionality of the relationship.

In some cases, the immediate ancestor of a given item was not found in the corpus. For example, *abatement* and *abates* are both found, but not *abate*. In this case, the item *abate* was added to the list, and flagged as a "missing ancestor." Sometimes intermediate forms were missing. In the group of words in Table 1, e.g., if the word *addend* had not occurred in the corpus, the relationship between *addend* and *add* would have involved two steps, suffixation and pluratization. In our analysis we supplied such "missing links" wherever necessary, flagging them to mark that they were not in the original list of words from the *WFB*.

[2] Since our sample is essentially a random sample of the *WFB*, we can assume that the percentage of items in a category in our sample will be approximately the percentage of items in that category for the entire *WFB*. However, there is an important sense in which the *WFB* (and hence our sample of it) is not representative of the population of words from which it is drawn. As the analyses by Carroll, Davies, and Richman (1971) indicate (see Table B-8 on p. xxxvi) all of the roughly 14,000 words in printed school English with frequencies greater than 2.5 per million would be expected to occur at least several times in the *WFB*. On the other hand, of the more than 200,000 words with a frequency of less than two per billion, less than 100 would be expected to show up in a corpus this size. Thus, in extrapolating from any corpus to the total vocabulary, a very high frequency word represents only itself, so to speak, whereas a low frequency word must be taken as representative of a large number of low frequency words which did not actually appear in the corpus.

Our estimates of the composition of the population have taken this into account by assigning a weight to each word, which is an inverse function of its frequency. This is why the Population % is often substantially different from the Sample %. For example, 11.65% of the words in our sample are morphologically basic. However, it turns out that morphologically basic words are not evenly distributed by frequency. Among the most frequent words in our sample (those that would occur on the average twice or more in a million running words of text), almost 28% were morphologically basic. However among the less frequent words, this percentage decreased, averaging around 6% in the lower frequency ranges. The percentage of morphologically basic words in the population (7.46%) reflects the fact that the population of words in printed school English has a higher proportion of low frequency words than does the *WFB* or our sample.

The details of how the projections were made from the sample to the population can be found in the full report of this research (Nagy & Anderson, 1982). Suffice it to say here that we are confident of the projections; alternative approaches for making the projections did not produce figures that varied much from those in Table 4.

[3] The number of morphologically basic words and semantically opaque derivatives (included here are SEM 3, SEM 4, and SEM 5 derived forms) gives us an approximate idea of the number of distinct word families among the words above any given frequency level. It should be cautioned that the number of distinct word families at any given level is underestimated somewhat, since the most frequent member of a word family is sometimes a regular inflection or transparent derived form. The word *inch*, for example, has a *U* of 79.48, whereas the *U* of the plural *inches* is 131.53. Thus, the word family containing *inch* and *inches* is not included in the count of 555 morphologically basic words and semantically opaque derivatives that have a *U* of 100.00 or greater. However, among the words in that frequency range, one does encounter a representative of the *inch* family, so that more than 555 word families are actually represented.

[4] Beck, Perfetti, and McKeown (1982) matched children from different intact classes on the basis of pretest scores. Some of the control subjects were drawn from a combined third- and fourth-grade class. This class may have had lower reading attainment than the other classes. It is well known that matching does not eliminate preexperimental differences when the populations sampled are different (cf. Campbell & Boruch, 1975).

IRA Research Grants

Guidelines for applying for research grants from the IRA Research Fund are now available. A maximum of US$5,000 has been place on each grant. Proposals must be postmarked by December 15, 1984. Grants will be awarded at the IRA Annual Convention in 1985. For guidelines and criteria, write to Research Grants, Research Department, International Reading Association, 800 Barksdale Road, PO Box 8139, Newark, Delaware 19714-8139, USA.