KASVATUSTIETEIDEN TUTKIMUSLAITOKSEN JULKAISUJA
REPORTS FROM THE INSTITUTE FOR EDUCATIONAL RESEARCH

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

Sauli Takala

350/1984

Evaluation of Students' Knowledge of English Vocabulary in the Finnish Comprehensive School

EVALUATION OF STUDENTS' KNOWLEDGE OF ENGLISH VOCABULARY IN THE FINNISH COMPREHENSIVE SCHOOL

BY

SAULI JAAKKO TAKALA

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Education
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1984

Urbana, Illinois

EVALUATION OF STUDENTS' KNOWLEDGE OF ENGLISH VOCABULARY IN THE FINNISH COMPREHENSIVE SCHOOL

BY

SAULI JAAKKO TAKALA

B.A., University of Jyvaskyla, 1965
M.A., University of Jyvaskyla, 1970
M.Ed., University of Jyvaskyla, 1980

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Education
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1984

Urbana, Illinois

EVALUATION OF STUDENTS' KNOWLEDGE OF ENGLISH VOCABULARY IN THE FINNISH COMPREHENSIVE SCHOOL

Sauli Jaakko Takala
Department of Educational Psychology
University of Illinois at Urbana-Champaign, 1984

This study estimated the size of students' passive and active English vocabulary knowledge in the Finnish comprehensive school after seven years' study (grades 3 through 9, ages 9-16) with some 450 clock hours. Vocabulary knowledge was measured with the constructed answer technique, in which students wrote the Finnish or English equivalents of decontextualized stimulus words.

A two-stage stratified random cluster sample of 39 schools and 2,415 students was drawn and presented some 950 randomly selected words in 20 rotated test forms. Multiple matrix sampling design meant that each student answered only 40 to 50 words. Interrater agreement for the 115,000 answers scored was above 90%.

There was no reliable difference in the students' passive and active vocabulary knowledge, as they were measured in the study. Also, students' knowledge of simple word-formation rules and their contextual inference ability were poorly developed, in comparison to typical L1 skills. The following reasons were assumed: (1) Finnish and English are not related languages, which may not encourage such skills. (2) The emphasis at this stage is on syntactical patterns, while morphology is largely neglected. (3) The treatment of texts is "intensive", giving students little exposure to English. The estimated average size of vocabulary was about 1,000 words, with great variability in performance. Fast learners knew about 1,500 words,

average students about 900 and slow learners about 450 words. Due to the limited word-formation skills, the estimates ought to be adjusted by up to 45% , by 17%, and by 7% for the three sets, respectively. The relationship between taught and learned vocabulary was 55%, 32%, and 20% for the three sets, respectively.

Variance components analysis showed that words made a greater difference in scores than students and that error of measurement can be lowered more efficiently by increasing the number of word items than by taking a larger student sample. There may also be an optimal size of input in vocabulary learning. Students who used a textbook with a lower input learned less than those whose textbook taught more words. The study is prefaced by an extensive review of vocabulary studies in L1 and L2.

Sauli Takala

Evaluation of Students' Knowledge of English Vocabulary in the Finnish Comprehensive School

Kasvatustieteiden tutkimuslaitoksen julkaisuja 350/1984
ISBN 951-679-158-1
ISSN 0448-0953

Tässä tutkimuksessa arvioitiin oppilaiden aktiivisen ja passiivisen sanaston määrää englannin kielessä peruskoulun päättövaiheessa. Oppilaat olivat opiskelleet englantia luokilla 3 - 9 (9 - 16 vuoden iässä), yhteensä n. 600 oppitunnin ajan eli n. 450 tuntia. Sanaston hallintaa mitattiin niin, että oppilaat kirjoittivat englantilaisten sanojen suomalaiset vastineet (passiivinen hallinta) ja vastaavasti suomalaisten sanojen englantilaiset vastineet (aktiivinen hallinta). Sekä teoreettisin että käytännöllisin perustein sanat esitettiin irrallaan, ilman lause- tai diskurssiyhteyttä.

Kaksivaiheista stratifioitua ryväsotantaa käyttäen otostettiin yhteensä 39 koulua. Kokeisiin osallistui kaikkiaan 2 415 oppilasta, joille esitettiin yhteensä noin 950 sanaa, jotka valittiin satunnaisesti erikseen kahdesta yleisesti käytössä olleesta oppikirjasta. Matriisiotannan soveltamisen ansiosta kukin oppilas vastasi vain 40 - 50 tehtävään. Sanat oli sijoitettu 20 erilaiseen tehtäväversioon (10 kumpaakin oppikirjaa kohti), jotka rotatoitiin satunnaisesti kussakin luokassa. Kaikkiaan n. 115 000 vastausta jouduttiin pisteistämään ja arvioitsijoiden välinen yksimielisyys oli 90 % luokkaa.

Oppilaiden passiivisen ja aktiivisen sanaston välillä ei todettu luotettavaa eroa. Oppilaiden sananmuodotuksen alkeiden hallinta oli varsin puutteellista ja heidän kykynsä päätellä sanojen merkitys lauseyhteydestä oli myös vaatimatonta luokkaa, mikäli niitä verrataan vastaaviin taitoihin äidinkielessä. Tuloksen arveltiin selittyvän mm. seuraavista syistä: (1) Suomi ja englanti eivät ole sukulaiskieliä, joten kielten tietynasteinen samankaltaisuus, mikä rohkaisee analogiapäättelyä sukulaiskielten välillä, ei auta suomalaisoppilaita. (2) Opetuksen painopiste opiskelun tässä vaiheessa on syntaktisten mallien ja sääntöjen opettamisessa, mikä merkitsee sitä, että morfologia jää vähäiselle huomiolle. (3) Tekstin käsittely luokassa on todennäköisesti "intensiivistä", joten vähäinen ekstensiivinen tekstin käsittelykään ei sanottavasti auta oppilaita omaksumaan sanamuodotuksen periaatteita omin päin.

Keskimääräinen sanaston koko peruskoulun päättövaiheessa oli n. 1 000 sanaa. Sanaston määrä vaihteli kuitenkin suuresti. Laajan kurssin oppilaat osasivat n. 1 500 sanaa passiivisesti ja aktiivisesti, keskikurssilaiset n. 900 sanaa ja yleiskurssin oppilaat noin 450 sanaa. Kun otetaan huomioon oppilaiden rajallinen sananmuodostuksen taito ja kyky päätellä sanojen merkitys lauseyhteydestä, todetaan että edellä esitettyjä arvioita tulee korottaa n. 45 % laajan kurssin, n. 17 % keskikurssin ja n. 7 % yleiskurssin osalta. Näin ollen laajan kurssin aktiivisen sanaston koko on n. 2 000 sanaa ja passiivinen sanasto (lauseyhteyden tukemana) n. 2 200 sanaa. Vastaavat luvut ovat 1 025 ja 1 050 keskikurssin osalta, ja 450 molemmissa tapauksissa yleiskurssin osalta.

Laaja kurssi oli oppinut n. 55 % opetetusta sanastosta, keskikurssi vastaavasti 32 %. Yleiskurssi, jolle opetettiin vähemmän sanoja kuin laajalle kurssille ja keskikurssille, oli silti oppinut myös suhteellisesti selvästi vähemmän opetetusta sanastosta (20 %).

Varianssikomponenttianalyysi osoitti, että sanat vaikuttavat enemmän pistemäärissä todettuun vaihteluun kuin oppilaat. Täten mittausvirhettä voidaan alentaa tehokkaammin lisäämällä sanasto-otoksen kokoa kuin kasvattamalla oppilasotoksen kokoa. Tulokset osoittivat, että tietyt sanat ovat suhteellisesti vaikeampia kuin toiset oppilaiden yleisestä tasosta riippumatta. Tuloskäsittelyn tässä vaiheessa ei voida sanoa, mitkä sanat ovat suhteellisesti helppoja ja mitkä vaikeita eikä myöskään miksi vaikeustaso vaihtelee.

Sanaston opetuksessa saattaa vallita tietynlainen optimisuhde opetetun ja opitun sanastomäärän välillä. Toinen yleisesti käytetyistä oppikirjoista opetti n. 350 sanaa enemmän keski- ja laajan kurssin oppilaille. Tämä heijastui positiivisesti myös opitun sanaston määrässä. Sanasto saattaa olla luonteeltaan sellainen, että runsas tarjonta on oppimiselle edullista. Tämä ei ehkä pidä paikkaansa kaikkeen oppimiseen nähden. Tämä optimaalihypoteesi on jossakin määrin ristiriidassa ns. hallintaoppimisen yleisen tulkinnan kanssa. Sitä tulisikin tutkia eri oppiaineissa ja eritasoisissa oppilasryhmissä.

Empiiristen tulosten lisäksi julkaisussa esitetään laaja katsaus sanaston opetuksen ja oppimisen tutkimuksen suuntaviivoista ja tuloksista. Katsaus kattaa sanaston oppimisen sekä äidinkielessä että vieraissa kielissä.

Hakusanat	Descriptors
- sanasto	- vocabulary
- vieras kieli	- foreign language
- äidinkieli	- mother tongue
- arviointi	- assessment
- yleistettävyys	- generalizability
- matriisiotanta	- matrix sampling
- varianssikomponenttianalyysi	- variance components analysis

ACKNOWLEDGEMENTS

I wish to express my deepest appreciation to a number of people: Dr. James L. Wardrop, my advisor and chairman of my doctoral committee, for his help and guidance during this project; Dr. Muriel Saville-Troike for her encouragement and support during my studies; Dr. Richard C. Anderson for his helpful and incisive comments on my thesis and for making a great number of unpublished material available for me; and finally Dr. Alan C. Purves, my thesis advisor, who made it possible for me to come to work with him at the University of Illinois and to finish my degree at the same time, and who has helped me overcome many obstacles in academic and practical life. Dr. and Mrs. Purves have provided me a home away from home, for which I am truly grateful.

When I was planning and carrying out the present investigation, I received a lot of help from several of my collagues at the Institute for Educational Research. I gratefully acknowledge the contributions by Mrs. Paula Bertell, Mrs. Liisa Havola, Mr. Hannu Saari, and Mrs. Anneli Vahapassi. When I was working and studying abroad, they helped me prepare the data for statistical analyses, which would have been impossible without their generous assistance. I am indebted to Mr. Kari Tormakangas for constructing a computer program that accomodated my complicated design and vast data. When I was in a hurry, he quickly and accurately ran the analyses for me and helped me to understand the print-outs. I am indebted to Dr. Wilga Rivers, Harvard University, for the several occasions during international conferences when we discussed L2 research problems in general and vocabulary research in particular. I also wish to thank Dr. Ulla Connor, Purdue University, for

reading the early drafts of my dissertation and for the valuable comments she made to make it more coherent and readable. I am also grateful to Dr. Elaine Degenhart for helping me with my English and checking the references.

Finally, I especially thank Dr. Raimo Konttinen for his help, support and encouragement throughout this project. He taught me about new ideas in test and sampling theory, and advised me in designing the study. He organized the data analyses at the Institute for me and solved many problems that arose during that work. When the data were finally in the print-out form, he gave me useful advice and penetrating comments through letters and phone calls. Our discussions throughout the project have been an on-going source of learning, enrichment and stimulation to me.

PREFACE

I have had a long and keen interest in words. I was first alerted to the importance of vocabulary knowledge when I came into contact with Swedish-speaking people and failed to understand them and, especially, to express myself. I was doing well in Swedish at school but that apparently was not enough. Since that early experience, I have felt that schools do not take vocabulary teaching seriously enough and have taken charge of vocabulary learning myself. I have been and continue to be an avid dictionary browser. That seemed to pay off, since I estimated that I learned some 10,000 Swedish and English words while I was at school.

My interest in vocabulary and semantics continued at the university. My Master's thesis was on the linguistic expressions of the concepts of "happiness" and "unhappiness" in Old English poetry. My extensive involvement in translating scientific articles and dissertations into English further confirmed my belief in the importance of appropriate vocabulary and led to a view that scientific text follows certain set structural patterns and is very formulaic in expression.

When I worked as a substitute teacher, I noticed that students' vocabulary knowledge was an important cause of lack of success in tests. When I instituted intensive vocabulary reviews, I noticed that even the poorest students made excellent progress in vocabulary knowledge and found the experience rewarding and beneficial for their self-image as language learners.

Since the early 1970's, I worked intensively on several L2 syllabuses, and in that work noticed that vocabulary research was a neglected field in L2

research. Thus, when the decision was made to carry out a large-scale assessment of students' progress in several school subjects and I was appointed to be in charge of the ESL part , I had no hesitation in making the estimation of vocabulary learning a major part of that project.

TABLE OF CONTENTS

For words, like Nature, half reveal
half conceal the Soul within.
(Tennyson)

CHAPTER I

INTRODUCTION

Interest in vocabulary has varied both in linguistics and in the study of first and second language learning. Closed systems, like grammar, have tended to attract more attention than more open-ended systems like lexicon. Halliday (1961) has noted that "the grammarian's dream is (and must be, such is the nature of grammar) of constant territorial expansion. He would like to turn the whole linguistic form into grammar, hoping to show that lexis can be defined as 'most delicate grammar'. The exit to lexis would then be closed, and all exponents ranged in systems" (quoted in Kress, 1967, p. 69).

In linguistics, meaning and lexicon were largely ignored since they posed great problems. For Bloomfield (1914; 1933) lexicon was a list of basic irregularities, a depository of all matters that could not be handled systematically within linguistic systems. Similarly, in transformational grammar there was originally relatively little attention given to the lexical compo nent (e.g., Raskin, 1983). This is no longer true as the work done by Chomsky (1972), Halle (1973), Aronoff (1976), Selkirk (1982), Bresnan (1978), and Lieber (1981) indicates. Word-formation and the organization of the lexicon are currently attracting considerable attention as the objects of linguistic research. Recently there has also been growing doubt, even among linguists and psychologists trained within the transformational grammar paradigm, concerning the psychological realism of transformations. This has led to at-

tempts to devise close-to-surface lexical interpretative theories of language and text (e.g., Halle, Bresnan and Miller, 1978; Rozencvejk, 1974). This view has dispelled some of the alleged mystery of syntactic development and processing (but led to talk about children's word wizardry, e.g., Carey, 1978, 1982; Clark, 1983). As Maratsos (1978) points out

> in lightening the burdens of those who hope to explain syntactic development, we complicate the work of those who hope to explain lexical development. Similarities that were previously supposed to be appreciated on the basis of similar deep structures are now said to be appreciated on the basis of similar semantic interpretations. Characterizing lexical information in such a way as to support the appropriate interpretations will complicate the theory of lexical component, and, inevitably raise new problems for those who hope to understand how such lexical information is acquired. (p. 263)

There has been movement towards expanding the role of the lexical component, so that it is assumed to include not only such lexical processes as compounding and derivation but also inflectional processes. Trying to account for the latter syntactically appears to lead to unnecessary complication. Linguists are currently studying questions like what items are to have entries in the lexicon, how lexical items are to be related to one another, what forms are to be derived through generation and what, by contrast, need listing.

In language teaching, there was early an interest in the choice of vocabularies. While in the 17th century Comenius' interest was pedagogical

and he wanted to select the most useful words, in antiquity there had been a greater interest in finding rare words that were fit for elevated literary use (Kelly, 1969). During the centuries since Comenius up to the beginning of the present century, practically oriented language teachers chose the taught vocabulary on the basis of their subjective estimation of the words' importance (e.g., Stern, Wesche & Harley, 1978). While Kaeding (1898) was the first to publish an extensive frequency dictionary (of German), it was the work by E. L. Thorndike in the early part of this century that provided the rationale and methodology for vocabulary selection. He was mainly interested in making school textbooks more appropriate for students and considered their vocabulary load a prime factor in this. Clifford (1978) has made a detailed study of Thorndike's role in the vocabulary research in the United States.

Foreign language educators were quick to pick up Thorndike's lead, since vocabulary selection seems to be even more essential in second language teaching than in mother tongue instruction. Coleman and King (1941) note that 60% of investigations on L2 teaching and learning recorded for the period 1930-1937 dealt with vocabulary selection and learning. This means more than 500 references. This focus on vocabulary selection and teaching is very obvious if one reads back numbers of a journal like Modern Language Journal.

Probably due to the dominance of structuralism in linguistics, behaviorism in psychology, and their pedagogical application in the form of the audio-lingual method in L2 teaching, interest in vocabulary studies waned, especially in the United States, but there was a revival of interest in France (Francais Fondamental) and in Canada in the 1960's and 1970's. With the emergence of transformational grammar, the emphasis shifted again from

vocabulary to syntax and interesting vocabulary studies were carried out mainly in Belgium, the Netherlands and the Soviet Union. Recently there has been a revival of interest in vocabulary in the United States, but mainly among psychologists and mother tongue teachers. Similar interest is singularly missing among second language researchers. A recent synthesis of current research on L2 (Dulay, Burt & Krashen, 1982) does not even mention vocabulary.

It was dissatisfaction with the current emphasis in L2 research and a strong belief in the importance of vocabulary knowledge for discourse comprehension and production that provided the impetus for the present study. From several interesting problems that could be addressed, we have chosen to highlight some quantitative aspects of vocabulary learning. Specifically, the main research task of this investigation is to estimate the size of students' passive and active vocabulary in English after a seven-year course in the Finnish comprehensive school. Before we outline the design of the present study in detail, we will try to set it in context by reviewing how certain central problems in vocabulary research have been addressed in the past and are being addressed at present.

Vocabulary research is occupied with several kinds of topics. An obvious problem is related to the object of research: What is the word? What is vocabulary? Some degree of clarity must be reached on the definition and specification of the research object. There must also be some justification, some rationale, for the study of an object. Why is vocabulary an interesting and/or important topic for research? Since the study has a quantitive approach, several questions appear worth clarifying: What is the total size of

vocabulary in a language? How does people's vocabulary grow? How many words do people know? How many words do professional wordsmiths, authors, use? How common are different words? How many words in a text must we know to understand it? Furthermore, since the study has a pedagogical orientation, the following questions need attention: What is the size of printed school language? How many words are usually taught in L2 courses? What does it mean to know a word? What causes difficulties in vocabulary learning? How is vocabulary taught and how could vocabulary learning be improved? And finally, what do earlier studies say of the size of students' active and passive vocabulary size in L2 learning? It is to such questions that we now turn. We will begin with a discussion of some views of the word.

CHAPTER II

WHAT IS THE WORD?

There are a number of definitions of the word. A useful discussion of these is provided by Karlsson (1976) in his overview of general linguistics. He notes that the concept "word" is ambiguous. Words can be "simple" or "basic" words in contrast to "derived" words or "compound" words. When the basic word form is taken to represent also all word forms derived through inflection, as is frequently done in dictionary entries, it is usually called a "lexeme".

Orthographically, it is a relatively easy task to define an "ortographical word": it is a string of graphemes surrounded by a space on each side. It is not equally easy to define a "phonological word" in many languages. In a language like Finnish, however, this is easier since main stress falls regularly on the first syllable and tells that it begins a new word. Such a fixed stress pattern can be taken as one indication of word boundary.

The word has also been approached grammatically. Thus the Finnish linguistic unit "näin" can be three grammatical words. As an adverb ("in this way", "like this"), it can also be regarded as a lexeme (i.e., it has its own entry in a dictionary). As a verb ("I saw" - the inflected past tense form, first person singular) it is not a lexeme, since the basic form is "nähdä" -"see". As a phrase ("with these" - the plural "instructive" case form for the pronoun "tämä" - "this") it is not a lexeme, either. Its lexeme is the singular nominative case form "tämä".

The word has also been approached semantically, or rather, notionally. The word has been defined as a unit which has independent or unified meaning. Although this often corresponds to the intuitions of "naive" language users, it is problematic linguistically. Its greatest weakness is lack of explicitness. It would be necessary to define what is meant by "independent" and "unified". It has sometimes been sought to overcome the problems of notional definitions of the word by resorting to an operational definition, e.g., by taking the possibility of pausing in front and at the end of a unit as the criterion of the word. Bloomfield used this approach when he defined the word as the smallest free form of language, meaning that it can appear alone in an utterance.

It has also been suggested that the word is a unit which has internal cohesion. Operationally this has been taken to mean that it is not possible to insert any extraneous linguistic material within a word and, similarly, that a word moves as a whole within a sentence. This test seems to work fairly well in many cases but it also has problems. One is circularity: if it is stated, as a generalization, that it is not permissible to insert a word within another one, the concept is defined by resorting to the very same concept used in the definition. On the other hand, units larger than what is considered a word (e.g., a syntagm) can move as whole units within a sentence, but a syntagm would hardly be labeled a word.

Difficulties like the ones described in the above have led many linguists to abandon the whole concept of word and operate with the concept "morpheme". Yet, as Karlsson (1976) points out, the word is not linguistically meaningless or useless only because it is difficult to define it satisfactorily. Also, in some languages - among them Finnish - there are structural

features, which apply to a "word-like" unit. In Finnish, the main stress is regularly on the first syllable of "word-like" units, and there are also certain phonotactic restrictions at the beginning and end of "word-like" units: these positions tend to shun consonant clusters.

People have strong intuitions about words. Even linguistically "naive" people can usually divide an utterance into words (cf. Sapir, 1921; Leontev, 1975). This has led some scholars (e.g., Bloomfield & Newmark, 1963; Moulton, 1970) to suggest that to many people a language is a collection of words. Sapir (1921) has suggested that consciousness of words is shared by speakers of all languages, including those which lack a traditional writing system. Leontev (1975) cites Sapir's work in which he showed that speakers who did not have a linguist's knowledge of language never regarded synsemantic units (e.g., prepositions) as independent words. Luria's studies with aphasics (cited in Leontev, 1975) have demonstrated the same phenomenon. Thus, e.g., ja idu v lec (I go to the forest) is regarded as three words: ja - idu -v lec. Such aphasics can count words correctly when they are fully semantic words (i.e., content words) but start making mistakes when form words (e.g., conjunctions, prepositions) are introduced. Young children have been shown by Luria to do the same. For such persons, words are not the words of school grammar but quant-words (Leontev's term for a psycholinguistic unit of sense). With training, children and even some aphasics can, however, be made to divide speech into words and syllables.

Some twenty years ago Bolinger (1963) suggested that word is unique both in form and in meaning. He asks (Bolinger, 1963) why it is that

> the element of language which the native speaker feels that he knows best is the one about which linguists say the least? To the untutored person, speaking is putting words together, writing is a matter of correct word-spelling and word-spacing, translating is getting words to match words, meaning is a question of word definitions, and linguistic change is merely the addition or corruption of words. Is the reason why the average individual embraces the word but the linguist shuns it, a single coin with two sides ? (p. 113)

Bolinger argues that the word is the source, not the result of phonemic contrasts. He marshals several arguments to support the claim that the phoneme is secondary, as the word "subtends" the phoneme. His conclusion is that "The word commands the phonemes that make it up" (p. 122).

According to Bolinger (1963), the word has other characteristics that set it apart from practically all other linguistic units. Units at other levels can be described distributionally. This is explained to be possible partly because of their combinatorial habits and partly because they are so few in number. The inventory of phonemes and phrase structures is limited, which encourages linguists to adopt total accountability as the ideal of linguistic science. By contrast, the stock of words is open-ended. Bolinger (1963) states that

> Our attempts to deal with it have been discouraging, whether we have tried, with the phonologists, to set up phonological words, only to find that its correspondence with words chosen on other criteria is rough indeed; or, with the generativists, to provide a place for it in the rewrite rules, and too often to be left with no recourse except to list a few examples and add "etc.".

(p. 114)

That the word deserves a special place in the linguistic system is mentioned by some linguists and psycholinguists. Jakobson (quoted in Bolinger, 1963) says that among the linguistic units compulsorily coded "the word is the highest". Maclay and Osgood (1959), in discussing hesitation phenomena, conclude that repeats characteristically involve a single word, occasionally several words, but only rarely units smaller than the word. They suggest that this has clear implications for the nature of encoding units. Householder (1961) claims that the brain does not normally bother to identify phonemes or anything smaller than morphemes. In discussing the uniqueness of semantic mapping, Householder (1962) observes that the mapping of one level to another permits complete freedom in all areas but one. Phonemes can be altered in phonetic shape, graphemes can be assigned to different phonemes or to syllables, morphs can replace morphs. On the other hand, the relationship between morph and meaning is unique: if a text containing a sequence of morphs is altered by having different morphemic values attached to the morphs, it eventually becomes "gibberish". Finally, Pike (1982) states that while his tagmemic theory, which consists of three minimum units (morpheme, phoneme and tagmeme), does not allow any of the three a prior status, if forced to set up a priority, he would take the morpheme as a more logical starting point than the phoneme, since it is more directly related to language as a system of communication.

Bolinger (1963) suggests that hearers have the capacity to discriminate between a large variety of sounds whereas speakers can easily produce only a handful of distinct and relatively instantaneous complexes of sounds. The

hearer apprehends words and larger chunks of sound at a stroke as configurations (as Gestalts), whereas the speaker must build them up element by element. Thus hearing probably does not involve analysis but speaking by necessity involves synthesis. It appears that twenty years later Swain (1983) has independently come to the same conclusion.

Bolinger (1963) also provides a useful discussion of meaning, meaning and context, meaning in a sentence vs. meaning of a sentence, etc. Since that would take us too far from the focus on words, we will not review that discussion but refer an interested reader to the original article. We now turn to some other linguistic discussions of the word.

In stratificational grammar, as it has been formulated especially by Lamb (e.g., Lamb, 1966, 1973), language is described as consisting of a number of strata (4-6 strata have been suggested at different stages in the development of the model). A recent model is shown in Figure 1.

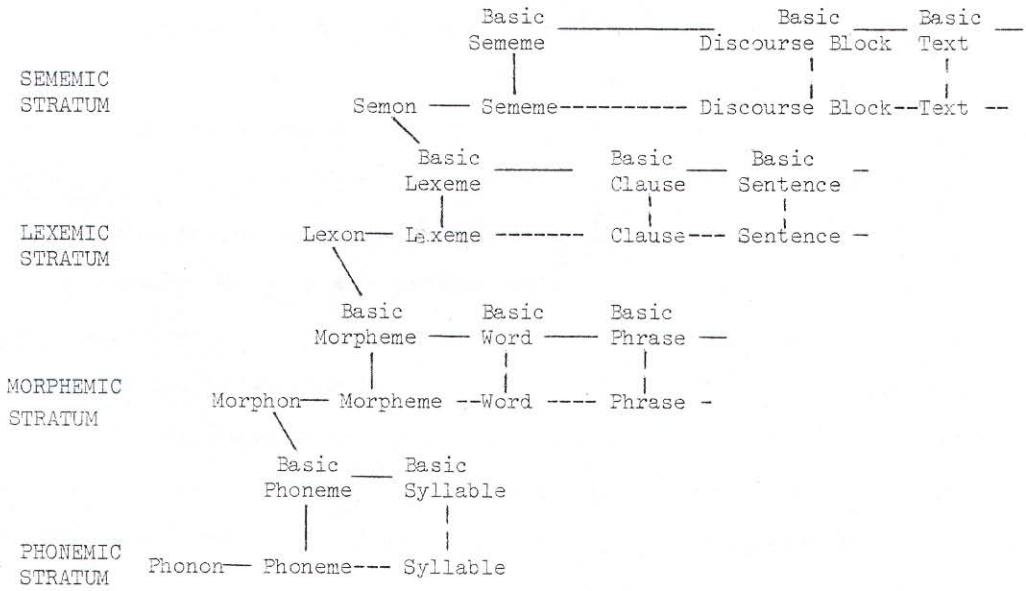

Figure 1. Lamb's model of linguistic strata.

Only a few comments will be made on the model. Stratificational grammar was designed to be a model of cognitive linguistics and it "aims at characterizing the speaker's internal information system that makes it possible for him to speak his language and to understand utterances received from others" (Lamb, 1973, p. 14). Stratificational grammarians found transformations, which featured prominently in the generative grammar, unacceptable since they were process descriptions, involving mutation of forms on the same level. While transformations might be able to account for primary data, they were thought to be unrealistic when applied to encoding and decoding. It is a basic assumption of stratificational grammar that linguistic units are not changed into anything. X as a linguistic unit may be realized as Y on another stratum. Another basic assumption is that the aim of grammar is to generate not sentences, but most texts, of a language. The goal of generating all texts was considered unrealistic.

The model, shown in Figure 1, has three grammatical strata and one phonological stratum for language proper, and a conceptual system containing all of the individual's knowledge (other than language knowledge; actually not shown in the figure but could be placed topmost in it). Each stratum consists of an inventory of its characteristic units or emes and a set of tactic rules that specify how the emes combine with one another on that stratum. Strata are connected with each other by realizational rules, which describe how the emes of one stratum are linked to those of another stratum. Any text, including sentences, has a number of different but simultaneous structures. On the semological level it is a network of relationships between units, at the grammatical level it is a tree of constituent structures, and phonologically it is a string of phonemes.

From the point of view of the present study, it is relevant to note that it has been found that it is not possible to collapse the morphemic and lexemic strata (as it has been to coalesce the phonemic and hypophonemic). An essential characteristic of a stratum is its distinctive tactic pattern. Since languages have syllable structure, a phonotactic system is needed. Similarly, since languages have word stems, prefixes and suffixes, which combine in certain ways to form words, a morphotactic system is needed. The lexotactic system is needed to handle the phenomenon of subjects, predicates, noun phrases and verb phrases. The semotactic level is needed to deal with, e.g., agents, goals, instruments, actives and passives. The morphemic, lexemic and sememic strata have partial coalescence in that they share a common inventory of morphemes (and the latter two also share a common inventory of lexemes). Thus , there must be direct links from morphemes to meaning (i.e., to the hypersememic stratum - or conceptual or gnostemic system as Lamb now prefers to call it) bypassing the lexemic and sememic strata, and the same applies to the lexemic stratum. Thus, some points of the conceptual network connect to the sememic stratum, some to the lexemic stratum, and some to the morphemic stratum.

As the model suggests, word as a linguistic unit plays a distinct role in the language system and language use.

In Halliday's systemic grammar (e.g., Halliday 1961), meaning ("formal meaning" and "contextual meaning") plays an important part. Formal meaning of an item is its operation in the network of formal relations. Contextual meaning is the relation of an item to extratextual features. Grammar is taken to be that level of linguistic form at which closed systems operate. Those

parts of linguistic form that are not concerned with the operation of closed systems belong to the level of lexis. In fact, Halliday suggests that the distinction between the two is not very clear-cut but a cline. A linguistic theory must provide both a theory of grammar and a theory of lexis, and also show how the two are related.

In systemic grammar, the term "unit" refers to those stretches of language that carry grammatical patterns. The units of grammar form a taxonomical hierarchy and thus occupy different ranks. The word as a unit occupies the following rank in the scale from highest to lowest: sentence, clause, group(/phrase), word, morpheme. Each unit consists of one, or more than one, of the unit next below.

According to Halliday (1961), "the grammarian's dream is (and must be, such is the nature of grammar) of constant territorial expansion. He would like to turn the whole lingustic form into grammar, hoping to show that lexis can be defined as "most delicate grammar". The exit to lexis would then be closed, and all exponents ranged in systems" (quoted in Kress, 1976, p. 69). Yet, no description has been so "delicate" that no further systems can be found. Relations at this level of delicacy can only be described statistically. Statistical differences may become so slender that the system eventually becomes an open set, i.e., we have entered the realm of lexis.

Thus Halliday (1961) early argued strongly for the importance of a theory of lexis and lexical relations. According to him, the exit from grammar to lexis tends to be associated predominantly - though probably not exclusively - with one unit, the word. It is fully consistent with this orientation, whose early exponent was Firth, that the study of lexical relations (e.g., collocations) has continued to occupy a prominent position in the work of

linguists trained in the Firthian-Hallidayan tradition. Halliday himself wrote one of the early articles on lexical relations, highlighting various aspects of lexical collocation (Halliday, 1966). More recently, Halliday and Hasan (1976) have studied cohesion in texts, and lexical chains form a prominent part of their treatment of how a cohesive text is created.

In the above, attention has been focused on how some prominent linguists have dealt with the word. What follows is a brief look at the treatment of the word by psychologists and psycholinguists. In the early symposium on psycholinguistics edited by Osgood and Sebeok (1965), considerable attention was devoted to the discussion of the units of language and their psychological status. Instead of citing the discussion included in that classic of psycholinguistics, reference is made to a scheme proposed by Osgood in his Presidential address to the American Psychological Association (Osgood, 1963). It is shown in Figure 2.

		Correlate	Decoding	Encoding	Correlate
MEANINGFUL	Temporal pattern	Interpretations	SENTENCES	SENTENCES	Intentions
		Kernel amalgomation	(PHRASES)	(PHRASES)	Kernel differentiation
	Spatial pattern	Meanings	WORDS	WORDS	Meanings
MEANINGLESS	Temporal pattern	Forms	WORDS	WORDS	Forms
		?	(MORPHEMES)	(MORPHEMES)	?
		Perceptual skill components	PHONEMES	SYLLABLES	Motor skill components
	Spatial pattern	Sensory signals	DISTINCTIVE FEATURES	DISTINCTIVE FEATURES	Motor signals

Figure 2. Osgood's psycholinguistic model.

It is not possible, nor it is needed for the purposes of the present study, to give an account of Osgood's comprehensive discussion of several issues in psycholinguistic theory. The figure indicates what Osgood considered the minimum and sufficient levels of units in language decoding and encoding. Units that in Osgood's opinion were of questionable psychological reality - even if being linguistically real and useful - are bracketed. Psychological correlates of each unit are indicated in the outer columns. Osgood makes a division into meaningful and meaningless levels of organization. In subsequent work, Osgood has made several refinements to the system but they are not crucial for the present discussion.

A few conclusions can be made on the basis of the figure. One is that the units of decoding and encoding are not necessarily identical. Thus the syllable is assumed to play a clear role in speaking but not in listening (cf. Bolinger, 1963). The morpheme, a key concept in linguistic analysis, is regarded to be a psychological non-entity. From the point of view of the present study, the most important observation is, however, that the word - the problem child of linguistics - plays a very interesting <u>double</u> role: at levels below meaningfulness it is the most inclusive unit while at the meaningful levels it is the minimal unit. Osgood (1963) proposes the word as the characteristic unit of <u>perceptual</u> forms in language for the following reason (formulaic sentences and phrases also fit the requirements):

> Given the principles by which the integration systems work, the units must tend toward the largest segments of language that are (a) highly redundant, (b) very frequent in occurrence, and

(c) within the temporal limits of cell-assembly reverberation. (p. 118)

Leontev (1975) has devoted considerable attention to the problems of units in psycholinguistic functioning and analyzed the role of word in it. His views will not be reviewed here. Reference is made to another publication by the author (Takala, in press).

To summarize, what seems to emerge is that the word is essentially a language-specific rather than a universal category. It seems to play a more central role in languages which have a rich inflectional system and in which word paradigms feature prominently. Thus, in a language like Latin, the conjugation of verbs is defined using the word-and-paradigm model. For instance, the lexeme "amo" (love) is conjugated through person, voice, tense and mood, which results in some 120 different word forms. Other verbs following the same conjugation are described by stating that they follow the "amo"-paradigm.

Through this discussion we have seen that while it is not easy to define a word in a way that would satisfy the criteria of a strict linguist, people behave as if the word is a non-problematic and basic unit of language.

CHAPTER III

WHY IS VOCABULARY KNOWLEDGE IMPORTANT?

We pointed out in the introduction that vocabulary has tended to be relegated to a secondary position in linguistic research and in research on language teaching and learning. This is unfortunate, since there are several indications that vocabulary knowledge may be more important than, for instance, the knowledge of grammar. Some of these will be discussed in this chapter.

It has been found that performance on vocabulary tests is highly correlated with reading comprehension. Several studies have yielded factor loadings ranging from .41 to .93 (Botzum, 1951; Clark, 1972; Davis, 1944, 1968). Vocabulary knowledge is, in fact, the best single predictor of discourse comprehension (Anderson & Freebody, 1981). This close relationship is not surprising, since surely comprehending discourse would be difficult or impossible without knowing a substantial part of the words in a text. We will take up this proportion question later (cf. also McKeown, Beck, Omanson & Perfetti, 1983; Mezynski, 1983).

Several considerations suggest that vocabulary is more important than grammar. First, grammar can be almost totally pruned away, but comprehension is still quite unproblematic. Bolinger (1970) gives an example where a deaf person said to another person: You Me Downtown Movie Fun. The message is understandable because of the content words and the knowledge of the world. Leontev (1975) states, in fact, that aphasics show similar patterns and that this kind of speech may resemble the form in which utterances are being planned in human mind, before they are realized in linguistic form. Second,

as Bolinger (1963) and Swain (1983) suggest, hearing and reading may not require detailed and constant grammatical analysis. Sampling of content words may give sufficient clues as to the meaning. If this is approximately correct, it would explain psychologically why a person may understand a foreign language quite a lot without being able to speak or write it. Third, the fact that performance on vocabulary tests is closely connected with assessment of intelligence may be explained, in part, so that intelligence is largely learning from context, and this learning is coded into words (Sternberg & Powell, 1983). On the other hand, it is unlikely that there is a similar close relationship between intelligence and knowledge of grammar.

It has also been shown that the characteristics of a text can play an important part in how the text is comprehended and recalled. The work by Richard Anderson and his colleagues at the University of Illinois has manipulated the stimulus characteristics and studied what effect such experimental control has on a variety of performance measures. In Pittsburgh, Beck and her coworkers have tried to test some hypotheses about the effect of vocabulary on text processing. These studies will be briefly discussed in the following. We will begin with the Illinois research program on the relationship between vocabulary knowledge and reading comprehension.

At the beginning of the 1970's, Anderson (1972) demonstrated in an incisive manner that the assessment of comprehension was a "mess", as he put it. The research program that Anderson has conducted since the early 1970's clearly reflects the concerns voiced in the programmatic article of 1972. There has been a consistent effort to get a better grasp of the stimulus characteristics (how verbal material can vary), to get a better understanding of the nature of comprehension, and to develop methods of measuring

students' response to carefully defined verbal stimuli.

In a series of experiments Freebody and Anderson (1981a, 1981b) attempted to clarify the role of vocabulary knowledge in text comprehension. They asked (Freebody & Anderson, 1981a) what proportion of content words in a text need to be unfamiliar before comprehension reliably deteriorates. A study with 72 sixth-graders showed that a relatively high proportion (1 out of 3) of substance words had to be replaced by a rare synonym before a significant deteriorating effect was detected in a sentence verification measure. On the recall measure, there was a trend for better performance if the vocabulary was easy. The effect of medium difficult vocabulary (every 6th content word replaced by a rare synonum) on comprehension measures was inconsistent. Thus, some 17% of all words and about 30% of all content words had to be rare words before text comprehension appeared affected. In the Soviet Union, Frumkina (1967) found that about 70% of the words in a text must be known for its satisfactory comprehension, and Klychnikova (1973) estimated that a literary text can be understood globally if 75% of words are familiar, all main ideas can be understood if 90% of all words are known, and that 95% of words must be known if most details should also be understood.

Freebody and Anderson (1981a) also asked to what extent the effect of vocabulary difficulty depends upon the location of unfamiliar words in important vs. unimportant ideas. The degree of importance of ideas in texts as determined empirically through student ratings. Three passages were constructed: an easy form with high-frequency words only, a difficult-unimportant form in which at least one rare word was substituted in each of the least important propositions, and a difficult-important form containing rare

synonyms for the original words in each of the propositions rated the most important. Seventy-one sixth- grade students took part in the study. As in the first experiment, the effect was limited. The most salient result was that the passages that contained unfamiliar words in unimportant positions were summarized significantly better than passages containing unfamiliar vocabulary in important positions. The results on the recall and sentence verification measures were less clear due to interaction effects that were difficult to interpret.

The authors (Freebody & Anderson, 1981a) suggest a "minimum effort principle" to account for the the results described in the above. Readers are hypothesized to avoid deep processing of unfamiliar words (i.e., they are skipped) and to proceed without lengthy interruptions. When unfamiliar vocabulary occurs in unimportant positions, little effort is spent on processing them and little disruption of processing ensues. Such strategy might lead to more thorough processing of the other propositions and to their better accessibility. Such effect was found in summaries (for an empirical attempt to test the mimimum effort principle, see Omanson, Beck, McKeown & Perfetti, 1983b, below).

In a third experiment, Freebody and Anderson (1981b) asked to what extent text cohesion interacts with vocabulary difficulty to diminish the negative effect of unfamiliar vocabulary on comprehension. Using the Halliday and Hasan (1976) system of describing text cohesion, three texts were produced: a high-cohesive text, a low-cohesive text, and an "inconsiderate" text which contained eight extraneous propositions. Each of the three texts appeared in an easy vocabulary version and an unfamiliar vocabulary version. The authors predicted that there would be interaction between cohesion and

vocabulary difficulty such that differences between the two vocabulary levels would be minimal in case of high text cohesion but more pronounced when cohesion was lower. The prediction was not confirmed. There were effects for vocabulary difficulty on the recall and summarization tasks, but there was no hypothesized interaction between vocabulary difficulty and cohesion level. There was some indication that reader fatigue might have had some influence on the obtained outcome.

In a fourth experiment Freebody and Anderson (1981b) asked how schema availability interacts with vocabulary difficulty. Is the effect of unfamiliar vocabulary less detrimental to comprehension when the schema of the text is familiar? A game and a visit scheme were used. Four passages were written, keeping syntax controlled, with one displaying a familiar instantiation of the theme and one an unfamiliar instantiation. Each of the four texts had an easy and a difficult vocabulary version. Eighty-two sixth-graders took part in the experiment. There was no significant interaction between vocabulary difficulty and schema availability on any verbal comprehension measure. Both vocabulary difficulty and schema familiarity affected performance in the recall and sentence verification tasks. There were no clear results obtained in the summarization task.

In summary, for all three measures (free recall, sentence verification, and summarization) in all four experiments, vocabulary difficulty was always in the expected direction while the effects were not always significant. Unfamiliar words always tended to decrease performance levels but a relatively high level of them was needed for any appreciable effect to appear in verbal comprehension measures. The fact that the results did not support the

idea of there being an interaction between vocabulary difficulty on the one hand and text cohesion and schema availability on the other was a surprise to the authors. One possible explanation for the cohesion effect is that the texts should be longer for the effect to appear. Also the inconsiderate text might have to have two or more competing "stories", which all make sense, to produce an effect. If the extraneous propositions are totally unrelated (irrelevant), they may be easily ignored. If they compete for attention due to sufficient relatedness, comprehension might be disrupted.

We will now turn to the Pittsburgh team work on text processing. Omanson, Beck, McKeown and Perfetti (1983a) studied the effects of unknown words on text processing. In an earlier study, it had been shown that when fourth-graders read stories in which 11% of the words were unfamiliar, the children recalled less than did children who had been taught and learned the meanings of the words. A model of text processing provided by Kintsch and van Dijk (1978), which claims that "carry-overs" and "reinstatements" mold individual propositions into the micro-structure construction of texts, provided the best fit to recall of stories containing known words. Two models were suggested to account for text processing in cases where texts contain unknown words. The "disrupted" model assumes that imcomplete propositions are ignored after their initial encoding. Incomplete propositions disrupt normal cycling in processing as they are not carried over or reinstated into additional processing cycles. The "suppressed" model assumes that a reader processes incomplete propositions the same way as complete propositions. However, since the meanings of incomplete propositions are vague, they essentially appear only in one cycle, while complete propositions are assumed to be processed in all the cycles in which they appear. For stories containing unknown words,

the best fit was provided by the "suppressed" model, which assumes that unknown words result in incomplete propositions that do not disrupt normal processing but which are poorly recalled.

The results suggest that, when reading stories which contain some 11% unfamiliar words, fourth-graders attempt to maintain normal coherence during processing. They do not seem to simply skip over units that contain unfamiliar words. Yet, readers are unlikely to recall such units. The meanings of words must be known to a sufficient degree in order to allow the propositions critical to the plot to be recalled.

In a related study, using the same data, Omanson, Beck, McKeown and Perfetti (1983b) continued exploring the effects of word knowledge and vocabulary teaching on comprehension. "Unfamiliarity" and "instructional" effects were studied in some detail, which will not be discussed here. Suffice it to note that the unfamiliarity effect was demonstrated (i.e., comprehension was impaired when children were unfamiliar with 11% of all words in a text) and that the "suppression principle" provided a better account of it than the "substitution principle". The latter, called a "minimum effort principle" by Anderson and Freebody (1981), assumes that unfamiliar propositions are skipped over and that familiar propositions are substituted for unfamiliar ones. This would lead to a prediction that the recall of familiar propositions would be better in conditions in which the children do not know a story's target words than in a condition in which the children had been taught the target words in the story. The prediction was disconfirmed.

The instructional effect concerned the phenomenon that instruction on unfamiliar words can enhance the recall of stories containing such words. Two

principles were suggested to account for this result. The "normal" principle assumes that propositions containing instructed words are processed in the same way as are familiar propositions. The "priming" principle suggests that propositions including instructed words receive additional processing. The priming principle was confirmed.

In sum, we have argued that vocabulary deserves increased attention from linguists, psychologists, and language teachers. There are several indicators that it may play a distinctly more important role in discourse comprehension in particular, but also in discourse production that does grammar. The lexical characteristics of text have been shown to be reflected in text recalll, sentence verification, and text summarization. To date, research has not revealed any clear interaction between text cohesion and content schema availability on the one hand and text comprehension on the other. We conjecture, however, that as research is continued and research methodology become more sophisticated, such interaction effects will be reliably detected.

CHAPTER IV

WHAT IS A VOCABULARY?

Overview

In this chapter we will attempt to provide a review of a variety of work that has been carried out in lexicologically oriented research. The emphasis will be on the quantitative analysis and description of the lexicon. First, we will address the question of the total size of lexicon. Then, we will describe the specific nature of the English lexicon. After that, we will look at how many words authors typically use. The discussion then moves to deal with the statistical distribution of words.

Total Size of Vocabulary in a Language

There has been a lot of interest in lexicology and lexicography to estimate various quantitative aspects of languages. One of the basic questions, and also one of the most difficult ones to answer, pertains to the number of words in the lexicon of a language. It has been estimated (Schmidt, 1964) that German has about 5 - 10 milliom words, if technical terms are also included. If the latter are excluded, the size of German vocabulary is estimated to be about 300,000 to 500,000 words.The difference in the estimates depends on whether only German words are counted or also loan words, whether only root words are included or also derived and compounded words, and whether only current words are taken into account or also archaic words.

There has been some work done also on some aspects of the internal structure of the lexicon. Ballmer (1981) has suggested that there is a possibility of characterizing a large set of text structures by means of a lexical analysis. According to him, texts are built up from sentences, sentences in

turn from words. The most important part of a sentence is the verb, since it shows the sequencing of states, events, and actions. Event coherence is denoted by the the verbs in a text and they are assumed to play a basic role for text structure. Since there are only a limited number of verbs in a language (according to Ballmer, about 20,000 in German), text analysis on the basis of verbs and simple sentences promises to provide substantive information about texts. Ballmer and Brennenstuhl have classified 8,000 simple standard German verbs using only two semantic relations: approximate synonymy and presupposition. They have further (Ballmer & Brennenstuhl, 1981) developed a model of speech activities, which distinguishes between expression, appeal, interaction, and discourse. Expression is characterized by an "emotion model", appeal by an "enaction model", interaction by a "struggle model" (which is further divided into an "institutional model" and a "valuation model"), and discourse by "discourse models" (sub-divided into "text models" and "theme models"). The complete verb thesaurus is classified according to the developed system. The work by Ballmer and Brennenstuhl is of interest to anyone working on texts and shows how the analysis of the lexicon can provide new possibilities of investigating larger units of discourse.

Composition of the English Lexicon

It is a well-established fact (eg. Baugh & Cable, 1978; Jespersen, 1905) that English has borrowed extensively from several languages during its recorded history. For a number of reasons, English has borrowed liberally from Latin, Greek, and French, in particular (Welch, 1975).

Baugh and Cable (1978) show that the influence of French was strong in England after the Norman conquest and it is strongly felt in vocabulary. Baugh and Cable (1978) estimate that there were about 900 French words in English before 1250. After that the French influence became very strong: a great number of words related to government and administration; church; law; war; fashion, food, and social life; art, learning, and medicine were incorporated into everyday usage. Baugh and Cable (1978) note that

> So far as the vocabulary is concerned, what we have in the influence of the Norman Conquest is a merging of the resources of the two languages, a merger in which thousands of words in common use in each language became partners in a reorganized concern. English retains a controlling interest, but French as a large minority stockholder supplements and rounds out the major organization in almost every department. (p. 174).

It is estimated that altogether more than 10,000 words were adopted during the Middle English period and 75% of them are still in current use.

Another major influence on English vocabulary was the Renaissance. Baugh and Cable (1978) estimate that about 10,000 words were introduced into the language and about half of them have become a permanent part of the language. A large majority of the words are of Latin origin.

It has often been said that English has three synonymous expressions at three levels - popular, literary, and learned. Similarly, it is often maintained that the native English words are concrete and forceful, the French word expresses refined and polished notions, and the Latin is the recondite synonym. Baugh and Cable (1978) discount this idea and suggest that many French words are equally vivid and concrete as the English words. Still, they

admit that there are clear distinctions between many three-level synonym sets: eg. rise - mount - ascend; goodness - virtue - probity; time - age - epoch.

Given the fact that several languages have had a definite impact on the English lexicon, the question arises how many native Anglo-Saxon words there are in modern English. It has been estimated (Koziol, 1937) that the Old English word hoard which persists in modern English is somewhat less than 35 per cent. This estimate is based on the Oxford New English Dictionary , which is a very large but by no means an exhaustive collection of English words. The figure for the total Germanic element in the NED dictionary is about 35 per cent and that includes compounds and Scandinavian borrowings.

Authors' Vocabularies

There has been a long interest in the study of authors' language. As Williams (1970) points out there may be "fingerprints" in writing, of which the author, and most critics, are not aware. Such characteristics might be amenable to counting and measuring and thus provide an "objective" measure of authors' styles.

The counting of words or letters in manuscripts was applied at least as early as 500 AD, when a school for the study and standardization of the Hebrew Old Testament grew up at Tiberias in Palestine. The members of this Masoretes school were concerned with the preservation of the text of the Old Testament in the exact form in which it came to their hands. They counted the number of letters and the number of words in each Book, and the number of repetitions of certain words, especially those with a more sacred meaning (Williams, 1970). More recently, a group of mid-nineteenth century scholars,

with Furnival as a prominent representative and the New Shakespearean Society (founded in 1874) as their main vehicle, developed the technique further, and it came to known as "stylometrics". While no advanced mathematical methods were used, numbers of word repetitions by classical and English authors were counted as well as variations in verse. These were usually expressed as averages and percentages of the total vocabulary or lines. Comparisons were made between authors and within authors across time (Williams, 1970).

Such studies have shown that the Old Testament has 5,642 words, the New Testament 4,800 words; that Homer used about 9,000 words, Goethe about 20,000 words, Shakespeare about 25,000 words, Milton some 7,000-8,000 words. Most authors typically use 4,000 words at most (de Greve & van Passel, 1971; Jespersen, 1905; Salling, 1958; Schmidt, 1964; Williams, 1970; Yule, 1944).

Given that the native word stock is only about one third of the total English vocabulary, the question arises to what extent authors use native vocabulary in their work. Koziol (1937) estimated that 88 per cent of Chaucer's words are of Germanic origin, 86 per cent of Shakespeare's , and 90 per cent of Tennyson's.

Distribution of Words

Historical Sketch

Bongers (1947), Fries and Traver (1940) and Kelly (1969) have discussed in great detail the development of statistically oriented lexicology ("lexicometrics") from the time of the Alexandrian school's commentaries on Homer to more recent times. Their work shows that in the 17th century Comenius was one of the first in the area of L2 teaching to pay serious attention to vocabulary selection on the basis of its usefulness. Rivers

(1981) notes that Comenius' "Vestibulum" (material for teaching conversation for children) contained a few hundred words arranged in a sentence context. His "Janua Linguarum" was planned to contain some 8,000 most common words. In the 1830's, Pestalozzi's primers provided a great impetus for grading reading vocabularies and he had a clear impact also on foreign language teaching (Rivers, 1981). Some fifty years later, one of the leaders of the Quosque tandem-society (a society founded in 1886 by a number of philologists who wanted to reform the teaching of modern languages), Otto Jespersen came out strongly in favor of vocabulary selection and simplified texts in his classic "Sprogundervisning", which was soon translated into English (How to teach a foreign language, 1904).

Focus on Frequency and Range. Although Kaeding was the first to publish an extensive word count in 1898 (based on a count of an impressive total of 11 million German words), it was Thorndike's work and his "The Teacher's Word Book" (1921) that provided the greatest impetus for vocabulary research for the needs of both L1 and L2 teaching. Studies on various aspects of vocabulary were published at an increasing pace (eg., Arnold, 1932; Blayne & Kaulfers, 1944; Bovee, 1919; Bovee et al., 1934; Broom & Contreras, 1927; Cartwright, 1925; Coleman, 1921, 1931; DeLozier, 1937; Dexter, 1928; Engel, 1931; Fitzgerald, 1931; Fotos, 1931; Grimes, 1933; Handschin, 1933; Haygood, 1933; Henninger, 1944; Holzwarth, 1931; Hubman, 1921; Johnson, 1927; Kaulfers, 1936; Keller, 1923; Kennedy, 1937; Kurath & Stalnaker,1936; Liebesny, 1944; Maronpot, 1930; Morgan, 1925, 1933, 1940; Morgan, 1926; Patterson, 1933; Porterfield, 1934; Reichling, 1916; Rippman, 1906, 1908; Rose, 1933; Roulston, 1929; Russo, 1947; Schobinger, 1934; Sears, 1931; Sharp, 1936;

Simmons, 1929; Skinner, 1935, 1936; Tharp, 1934; West, 1930, 1937; Wilkins, 1924; Wilson, 1939) so that 60% of the bibliography of studies on modern language teaching for the period 1930-37 dealt with vocabulary (Coleman & King, 1941).

A great number of frequency studies were carried out and published. For an exhaustive treatment of these, the reader should consult Fries and Traver (1940), Bongers (1947), Keil (1965), Mackey (1965), and Clifford (1978). As an illustration of early work on lexicometrics, we will cite some figures provided by Godfrey Dewey (1923):

9 words were found to form over 25% of the total words
12 syllables " " " " 25% " " " syllables
4 sounds " " " " 25% " " " sounds
69 words " " " " 50% " " " words
70 syllables " " " " 50% " " " syllables
9 sounds " " " " 50% " " " sounds
723 words " " " " 75% " " " words
339 syllables " " " " 75% " " " syllables
19 sounds " " " " 75% " " " sounds
1027 words occurring over 10 times form 78.6% of the words
1370 syllables " " " " " 93.4% " " syllables
41 plus 1 sounds form all, ie., 100.0% " " sounds

The first modern large-scale word-count of English carried out by Thorndike (Teachers' Word Book, 1921). His work soon found successors in the area of foreign languages. Keniston (1920) published a list of common Spanish words. Henmon (1924) published a frequency book of French based on 400,000 words. Morgan's (1928) frequency word book of German was based on Kaeding's

work. Buchanan (1929) prepared a graded Spanish word book. Vander Beke (1929) compiled another French word book. Keniston (1929) published an idiom list for Spanish, Hauch (1929) for German, and Cheydleur (1929) for French. In 1932 Zipf published his first study of relative frequency in language. The two best known compilations utilizing the work of several earlier frequency counters were by Eaton (1940) and Faucett and Maki (1934).

Several articles appeared in journals reporting on the unsatisfactory degree of word repetition and lack of agreement between different textbooks concerning the number of words taught (eg., Dexter, 1928; Jamieson, 1924; Johnson, 1927; Morgan, 1925; Wadepuhl, 1923; Wood, 1927). This helped to make frequency wordbooks standard tools in textbook making since the 1930's. On the other hand, there was intermittent criticism levelled at the frequency-range approach to vocabulary selection. Engels (1968) drew attention to the "outliers", ie. the importance for comprehension of those words that do not belong to, say, to the 80% most common words. While 80% of the words in a text may be drawn from a pool of 2,000 - 3,000 words, the rest can be taken from the remaining part of the vast total lexicon. In his doctoral dissertation, Richards (1971a) also discusses the limitations of the frequency approach (see also Boot, 1975; Bull, 1950). Sciarone (1979) provides a lengthy discussion of such criticisms and marshals several arguments in favor of the frequency approach.

Focus on coverage. Dissatisfaction with the strictly statistical approach led several vocabulary researchers to search other ways for vocabulary selection. Ogden (1930) discarded the frequency-range approach and used logical and conceptual considerations in arriving at his Basic English. This

led to heated discussions, in which Basic English was hailed as a marvelous tool for international communication and denounced as an atrocity. Much less controversial was the combined objective and subjective approach advocated by Harold E. Palmer and Michael West.

Common to Ogden, Palmer, and West was their interest in coverage, ie., selecting such words that can take the place of as many other words as possible. Such replacement can happen by definition, by combination, by inclusion, and by extension (eg., Mackey, 1973; Mackey & Savard, 1967; Richards, 1971b; Savard, 1970; Walpole, 1937).

For several decades Palmer worked on the general principles of language teaching (Palmer, 1917; 1921) and on vocabulary selection and dictionary making (Palmer, 1938). He also wrote a great number of readers and simplified classical books for foreign language learners. West (1930; 1937; 1956) worked along similar lines, and his "General Service List of English Words" (1953) and "Teaching English in Difficult Circumstances" (1960) have exerted a great influence on the practice of foreign language teaching all over the world. In October 1934, all the most prominent vocabulary researchers (including Faucett, Palmer, Thorndike, and West) met in New York under the auspices of the Carnegie Corporation. Their work resulted in the famed "Interim Report on Vocabulary Selection for the Teaching of English as a Foreign Language", which was published in 1936 in London by King and Son.

Focus on Availability. Michea (1953) was struck by the facts that concrete words are recalled better than structural words, and that people typically report similar concrete nouns when they are asked to tell which words come to their mind when they think of a topic like "going on a trip". Michea drew up 16 such "centers of interest" and then asked students to list

the 20 words that first came to mind, when each was thought of. This work led subsequently to the celebrated basic vocabulary used in the French course "Francais Fondamental" (Binon & Cornu, 1983; Gougenheim, Michea, Rivenc & Sauvageot, 1964). Subsequent works on the availability of words have been published by Savard and Richards (1970), Pfeffer (1964), and Dimitrijevic (1969).

Focus on Familiarity. In his doctoral dissertation, Richards (1971a) pioneered a method, which attempts to measure people's impressions of how familiar words are to them. He took all the concrete nouns in The Advanced Learner's Dictionary and in The New Merriam-Webster Pocket Dictionary (N= 4,495 nouns), prepared a number of rotated test booklets, and presented them to 1,000 college students (mean age 18.7 years). Richards' work is utilized in the LET corpus, but we do not know of any other study in which the methodology has been applied. Yet, given that there are only a limited number of verbs in a language (cf. Ballmer's estimate that there are some 20,000 verbs in German and some 8,000 of them are simple verbs), it would not be too difficult to obtain familiarity estimates for verbs also.

According to Keil (1965), in the mid-1960's there were frequency wordbooks available for 20 languages.

Current View

Thorndike's books were standard works in vocabulary selection and control for decades. The Brown corpus (Kucera & Francis, 1967) was the first large-scale up-to-date frequency count of written English. The corpus consists of about one million words, which were found in 500 samples of about 2,000 words each. These samples were drawn from 15 text categories. The

authors have recently published a new edition of the Brown corpus, which gives detailed lexical and grammatical analyses of the corpus. In 1971, Carroll, Davies and Richman published a word frequency book of written school English as it appears in school textbooks for grades 3 through 9. The Norwegian Computing Center for the Humanities has recently published a frequency book of British and American English (Hofland & Johansson, 1983). It is based on the Brown corpus and a similar LOB (Lancaster-Oslo/Bergen) corpus. Hall, Linn and Nagy (1980) have published a frequency count of the natural conversations of 4,5 - 5 year-olds from 40 families with their families, friends, teachers, and preschool playmates. This amounts to a total of 280 hours of taped conversations. The most useful frequency wordlist for the beginning and intermediate stages of teaching English as a second/foreign language is probably the LET vocabulary list (Leuven English Teaching Vocabulary-list, Engels, van Beckhoven, Leenders & Brasseur, 1981), which combines the information in the LOB corpus, the Leuven Drama Corpus, the noun familiarity list (Richards, 1971a), and high-coverage words (words used to define other words in Longman Dictionary of Contemporary English, 1978 and in West, 1977). A useful handbook for analyzing vocabulary for teaching English for special purposes is provided by Macdonald, Troike, Galvan, McCray, Shaefer and Stupp (1982).

According to Miller (1981), the 50 most frequent words in speech make up 60% of what we say, and the 50 most frequent words in writing make up 45% of what we write. The most commonly occuring words are monosyllables. The most common nouns and verbs are also among the first words learned by children, and such words tend to have many different meanings. If word length is held constant, they are also the easiest to say, read, remember, and think of.

George Kingsley Zipf studied such phenomena in great detail, and his formula for word probabilities is known as the Zipf law. Zipf showed that his law applied to a number of languages.

Table 1, taken from Nagy and Anderson (1982), shows two phenomena in printed school English, which are shared by all frequency counts: most words are in the lower ranges of the frequency spectrum, and morphologically basic words and semantically opaque words occur in the upper end of the frequency distribution. About half of the words in printed school English occur roughly once in a billion words of running text, or less frequently. Also, although there are substantially more transparent derivatives than there are morphologically basic words in printed school English, semantically transparent words are relatively rare among the most frequent words.

Table 1

Cumulative distribution of words by frequency (Source: Nagy & Anderson, 1982, Table 10)

Frequency (in terms of U)	Number of Words in Printed School English at or above a given 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
.00000	609,606	88,500	139,000

Nagy and Anderson (1982) also make some observations on low-frequency words. They point out the fallacy of assuming that words that occur only once in a million words of text are of negligible importance. In fact, words of relatively low frequency in printed school English are of more than marginal utility. It is also mistaken to assume that words of low frequency are necessarily difficult. Finally, since the Word Frequency Book (Carroll, Davies & Richman, 1971), on which the above data are based, gives frequencies by distinct word type and not by word family, the estimates of word frequencies exaggerate the low frequency of many words.

Words in School English

In 1971 Carroll and his colleagues (Carroll, Davies & Richman, 1971) published a pioneering dictionary, which was a word frequency book based on slightly more than five million words of running text from a careful sample of a thousand items of published materials in use in schools. The American Heritage Word Frequency Book is noteworthy not only because it is based on the kind of written language children encounter in school but also because its generalizability extends beyond the vocabulary contained in the frequency book itself to the total vocabulary of the type of materials from which the sample was selected. Grades 3 through 9 are covered by the frequency count.

Nagy and Anderson (1982) have recently used the American Heritage Word Frequency Book (henceforth, the WFB) to conduct a series of studies on printed school English. They claim that determining the absolute size of individual's vocabularies is of both theoretical and practical interest. Whatever the size turns out to be, small or large, theories of learning and language acquisition should be able to posit a mechanism that gives an adequate ac-

count of vocabulary acquisition. Similarly, the size of vocabularies has implications for vocabulary instruction. Direct vocabulary teaching may or may not be a feasible idea depending on the size of the vocabulary that needs to be taught and on the time that is available for it.

Nagy and Anderson (1982) took a sample of 7,260 words from the 86,741 words in the WFB by selecting a random set of 121 chunks of 60 contiguous words. The reason for using chunks of contiguous words was that semantically related words, an important concept in the study, are usually , though not always, close to each other in an alphabetical list.

The sample of 7,260 words was subjected to a detailed analysis. After reviewing earlier work on vocabulary size estimation (eg. Anderson-Inman, Dixon & Becker, 1981; Becker, Dixon & Anderson-Inman, 1980; Dupuy, 1974; Rhode & Cronnell, 1977; Stauffer, 1942; Thorndike, 1941) and noting that they have several problems (e.g., relying too heavily on historical and etymological relationships between words; neglecting compounding; not recognizing that there are varying degrees of relatedness among words), Nagy and Anderson determined that a psychologically more appropriate approach was to see the relatedness between words in terms of the relative ease or difficulty with which a child could either learn the meaning of a given word or infer its meaning from the context while reading. Thus, relatedness among words is a question of both similarity of form and ease of recognition of meaning.

Without going into the details of the method applied by Nagy and Anderson, it can be noted that they coded each word in their sample both in terms of formal relationship between words and in terms of their semantic relatedness. In coding the formal relationships between words, an "immediate ances-

tor" was found for each word and the relationship between the immediate ancestor and the target word was determined. The following relationship categories were employed: morphologically basic word (which obviously has no "immediate ancestor"), simple capitalization, alternate spellings, alternate pronunciations, alternate form of word, alternate form with s, regular inflections, irregular inflections, regular comparatives and superlatives, irregular comparatives and superlatives, suffixation, prefixation, compounds and contractions, truncations, and idiosyncratic morphological relationships.

The semantic relationship between each two compared words was coded in terms of two dimensions. The first represents the semantic relationship between the two most similar meanings of the two words. The second refers to the relationship between the two most similar familiar meanings of the two words. Six levels of semantic relatedness were distinguished on the basis of the following criterion: if a child knows the meaning of the immediate ancestor, but not that 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. The six categories were defined as follows:

SEM 0: the semantic relationship between the target word and the immediate ancestor is transparent and the meaning of the target word is totally predictable. SEM 0 includes most regular inflections, many affixations, and compoundings (eg. red - redness; misinterpret - misinterpretation; plankton, burger - planktonburger)

SEM 1: the meaning of the target word can be inferred from its immediate ancestor with minimal contextual help (eg. represent - misrepresent; wash - washcloth; crowd - crowded)

SEM 2: the meaning of the target word can be inferred from its immediate ancestor with reasonable help from the context making "one exposure learning" possible (eg. therapy - therapeutic; fog - foglights)

SEM 3: the meaning of the target word includes semantic features that cannot be inferred from the meaning of the immediate ancestor without substantial help from the context (eg. pass - password; collar - collarbone; conclusion - conclusive)

SEM 4: the meaning of the target word is only distantly related to its immediate ancestor and,thus, the link between the two needs to be learned (eg. well - farewell; high - high-school; neglect - negligible)

SEM 5: there is no discernible semantic connection between the target word and its immediate ancestor (eg. shift- shiftless; dash - dashboard; fox - foxtrot)

The results of the analysis of the sample of 7,260 words are summarized in Table 2.

Table 2 shows that the size of printed school English in grades 3 through 9 is very large. By several definitions of the "word", the population includes over 200,000 words, and about 100,000 proper names. A total of 170,000 words are derived by suffixation, prefixation and compounding. The number of morphologically basic words, which have to be learned without the help of "immediate ancestors" is considerably smaller, but still amounts to 45,000.

Table 2

Analysis of the Word Frequency Book by Word-Relatedness Categories

(Source: Nagy & Anderson, 1982, Table 4)

Categories	Sample N	Sample %	Corpus N	Population %	Population N
A. Categories that would be included in most definitions of "word"					
Morphologically basic	846	11.6	10,108	7.5	45,453
Idiosync. relation	72	1.0	860	1.0	6,167
Suffixation	722	9.9	8,626	7.6	46,431
Prefixation	233	3.2	2.784	4.0	24,457
Compounding & contract.	1,038	14.3	12,402	17.2	105,044
Truncations	16	0.2	191	0.2	1,144
Abbreviations	12	0.2	143	0.1	897
Subtotal	2,939	40.5	35,115	37.7	229,593
B. Categories with own separate entries in most dictionaries					
Irregular inflections	49	0.7	585	0.3	1,528
Irreg. compar. & superl.	1	0.0	12	0.0	13
Alternate form of words	8	0.1	96	0.2	1,072
Alternate forms with s	8	0.1	96	0.1	693
Semantically irreg. pl.	8	0.1	96	0.0	136
"Scientific plurals"	2	0.0	24	0.0	145
Subtotal	76	1.0	907	0.6	3,587
C. Categories normally without separate entries in dictionaries					
Regular inflections	1,553	21,4	18,555	16.4	99,547
Regular comp. & superl.	46	0.6	550	0.5	3,149
Incorrect regul. infl.	3	0.0	36	0.1	450
Simple capitalization	618	8.5	7,384	8.5	51,906
Alternate spellings	136	1.9	1,625	3.0	18,584
Alternate pronunciation	87	1.2	1,039	1.2	7,381
Subtotal	2,443	33.6	29,188	29.7	181,017
D. Categories relating to proper names					
Basic proper names	929	12.8	11,099	14.8	90,107
Derived proper names	88	1.2	1,051	1.2	7,215
Capitaliz. homogr. of pns	76	1.1	908	0.7	4,114
Variants of proper names	302	4.2	3,608	4.7	28,869
Subtotal	1,395	19.2	16,667	21.4	130,305

table continues

Table 2 (cont.)

Categories	Sample N	Sample %	Corpus N	Population %	Population N
E. Categories not normally counted as words					
Formulae & numbers	339	5.5	4,767	5.9	35,891
Compounds with numbers	41	0.6	490	0.8	4,894
Nonwords	147	2.0	1,756	3.4	20,444
Foreign words	46	0.6	550	0.9	5,618
Subtotal	633	8.7	7,563	11.0	66,847
F. Miscellaneous categories					
Errors in WFB	6	0.1	6	—	—
Ambiguous words	19	0.3	227	0.1	292
Ambiguous proper names	2	0.0	24	0.0	27
Missing ancestors added	203	2.8	2,425	--	--
2nd meanings of ambig. items added	51	0.7	609	--	--

Table 3 shows the estimates of the number of derived words in the total population of words, broken down by type of relationship and by degree of semantic relatedness.

Table 3

Derived words Arranged by Relationship Category and Degree of Semantic Relatedness (familiar meanings) (Source: Nagy & Anderson, 1982, Table 5)

	Relationship Category				
	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

According to the Nagy & Anderson classification of semantic relatedness, SEM 0, SEM 1, and SEM 2 include those cases in which the meaning relationship is transparent. We can see from Table 3, then, that there are an estimated total of 139,020 derived forms in the population of printed school English whose meanings are transparently related to the meanings of their bases ("immediate ancestors"). This figure demonstrates in a convincing manner the importance of knowing the basic principles of word formation. Nagy and Anderson (1982) note that "a reader who cannot take advantage of morphological relatedness among words has in some sense more than twice as many words to deal with as the reader who utilizes these relationships" (p. 26).

If we add up the number of morphologically basic words in the total population of words in printed school English (45,453) and the number of semantically opaque derived words (43,080), we get an estimate of 88,553 words (or rather distinct word families) as an estimate of the total size of printed school English. If students could, in fact, utilize semantic relationships at SEM 3 level, the number of distinct word families encountered by ninth graders would be 61,934. Similarly, if a student cannot perform at SEM 2 level, the estimate of distinct word families would be 135,825.

Nagy and Anderson (1982) have also estimated the number of distinct meanings in morphologically basic words. At level SEM 2, the estimate is 73,417 distinct meanings and at level SEM 3 59,821 distinct meanings. When homophony is incorporated in the estimates, and additional 31,821 distinct derived words is added bringing the estimate up to 105,238 distinct meanings. At SEM 3 level, 7,596 is added making a total of 67,417 distinct meanings.

Taylor (1979) has analyzed about 700,000 words of running text covering high school textbooks in mathematics, science, history, commerce, social

studies, and geography. Thus the textbooks cover so-called content areas and do not include English ("mother tongue") textbooks. The survery covered grades 7 through 10 and was limited to the school system of New South Wales in Australia.

The purpose of the survery was to ascertain what linguistic difficulties are faced by immigrant students at the major secondary levels of education. To achieve this purpose, nine widely used textbooks were analyzed from cover to cover along with randomly sampled extracts from another nine commonly used textbooks. Both manual and computerized analyses were carried out.

Although word frequencies were computed, they are only reported for the letters a-c, by way of example, since the author did not consider a 50-page listing of words of sufficient interest to warrant the use of so many pages. Altogether 615 distinct word families, formed in much the similar way as the word families in the Nagy and Anderson study of printed school English (Nagy & Anderson, 1982) were reported. If the words were evenly distributed across the whole alphabet, that would make a total of 5,330 items in a corpus of 700,000 words and 18 texts as against some 35,000 items in the WFB corpus of about 5 million words and 1000 texts (cf. Table 2, Section A).

Taylor is more interested in reporting predicate patterns. There were altogether 6,937 predicates in the mathematics textbooks, of which 238 were used more than four times. Five verbs (find - 7.55%; use - 5.75%; draw - 3.88%; show - 3.17%; make - 2.47%; write - 2.42%, and give - 2.25%) account for 27.5% of all pedicates in mathematics textbooks, and 27 verbs represent 45.3% of all predicates. There were 3,961 verb occurrences in a science textbook. Of these 195 verbs were used more than four times, 52 occurred four

times, and 502 verbs fewer than four times. Ten verbs account for 21.2% of all predicates (see - 3.51% show - 3.03%; use - 3.00%; make - 2.35%; call - 2.27%; contain - 2.25%; form - 2.12%; give - 1.36%; take - 1.34%; move - 1.31%). Forty verbs represent 43.3% of all content predicates.

In summary, Nagy and Anderson (1982) have demonstrated convincingly that low estimates of basic words in English (e.g., the 12,300 figure given by Dupuy, 1974) are mistaken. They estimated that there are some 88,500 distinct word families in printed school English. The figure rises to 110,000 if homographs and other distinct meanings, abbreviations, etc. are counted as separate words. Some generalizations appear also possible on the basis of Taylor's study. School geograpgy seems to have the most diversified vocabulary, while mathematics has the greatest number of specialized words and history the lowest. Social sciences seem to be more lexically dense than the physical sciences. Some 30 verbs account for almost 50% of all predicates in school mathematics texts and some 40 verbs do the same task in science texts.

Vocabulary Knowledge and Level of Understanding

What percentage of words in a text must be known at various levels of understanding? Frumkina (1967) studied the question of vocabulary minimum in relation to reading comprehension. She states that 1,000 most frequent words cover about 65 to 70% of all words in any text and 2,500 words cover 70 to 80%. The purpose of her experiment was to examine empirically the extent to which familiarity with a certain percentage of words in a text influences the comprehension of texts. If a foreign language were used, it would be impossible or hard to know how many words are, in fact, familiar. Therefore, the experiment was conducted in the mother tongue (i.e., Russian) such that a

part of the words in a text was replaced by quasi-words, which followed the morphological and grammatical characteristics of Russian words. Frumkina concluded that generally speaking 70% of the words in a text must be known for its satisfactory comprehension. This would be reached when the known vocabulary size is 2,000 - 2,500 words.

Johnson (1972) has estimated that 1,300 most common frequent words have a coverage percentage of 74, i.e., a reader who only knows the 1,300 most frequent words is familiar with 74% of the text. Two thousand most frequent words cover 80% of text, while 5,500 most frequent words have a coverage of 91%. Klychnikova (1973) has estimated that a literary text can be understood globally if 75% of words are known, all main ideas can be understood if 90% of all words are familiar, and that 95% of all words in a text must be known if most details should also be understood (cf. the figure given by Freebody and Anderson, 1981a, according to which about 17% of all words and 30% of all content words had to be rare words for comprehension to deteriorate among 6th-graders). Thus, it appears that some 5,000 words is a minimum for a relatively effortless reading of a literary text.