This bibliography provides citations to studies of word/letter frequencies and vocabulary applications based on issues of word frequency. While the original intent was to gather information pertinent to developing communication board vocabularies for argumentative communication systems, these resources should be useful for a variety of word frequency applications, including developing social resources and textbooks, estimating relative difficulty of sets of words, developing core vocabulary lists, etc. Its contents include: "Spoken Production Data," "Text Analysis—Original Word Frequency Data," "Comparison/Compilation of Other Frequency Studies," "Vocabulary Frequency Information for Special Populations," "Technical/Specialized Word Frequency Studies," "Diary-Based Vocabulary Studies," "Letter/Syllable Frequency Analysis," "Semantically-Based Frequency Analysis," "Methods and Theories of Vocabulary Frequency Analysis," "Differences in Word Frequency and Use by Communicative Mode," "Bibliographies/Reviews of Vocabulary Frequency Studies," "Selected Master's Theses Related to Word Frequency Analysis," and "Ph.D. Dissertations and Unpublished Papers Related to Vocabulary Frequency." (JDD)
BIBLIOGRAPHY OF VOCABULARY FREQUENCY AND WORDSET ANALYSIS STUDIES

Cress, C. J.

1986

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
M. Brady ______________

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Trace Research and Development Center
on Communication, Control, and Computer
Access for Handicapped Individuals
University of Wisconsin-Madison
Bibliography of Vocabulary Frequency
and Wordset Analysis Studies

Version 2.0
June, 1986

compiled by
Cynthia J. Cress
Trace Research and Development Center
Waisman Center
University of Wisconsin-Madison
BIBLIOGRAPHY UPDATE REQUEST SHEET

The maintenance and distribution of this bibliography is part of a series of on-going vocabulary projects conducted at the Trace Center. We are therefore interested in any further resources and references in the area of wordset or vocabulary analysis. We are particularly interested in:

- additional references concerning the topic areas listed in this bibliography;
- additional topic areas that should be covered in the bibliography, as well as any references in those areas;
- practical application tips and other research implications for information of this type;
- any other comments or suggestions.

This bibliography, and future revisions of it, will be available from the Trace Center's Reprint Service. If you would like to be informed of new revisions as they are published, please check the box at the bottom of this page and return it to the Trace Center.

Thank you for your interest and help.

Cynthia Cress
Project Coordinator

___ Please let me know as revised versions of this bibliography are published.

NAME: ____________________________________________
ADDRESS: _________________________________________
ADDRESS: _________________________________________
CITY/STATE/ZIP: ___________________________________________________________________

My suggestions for revision or addition to this bibliography are:

____________________________________________________________________________________
____________________________________________________________________________________
____________________________________________________________________________________

Send this form, and any correspondence or questions in this area, to Cynthia Cress, Project Coordinator, Trace R&D Center, Room S-151 Waisman Center, 1500 Highland Avenue, Madison, Wisconsin 53705, phone 608/263-1156.
Introductory Notes to Bibliography

Version 2.0
June 1986

This bibliography is a collection of studies of word/letter frequencies and vocabulary applications based on issues of word frequency. The original purpose was to gather together a range of information pertinent for the development of communication board vocabularies for augmentative communication systems. Often, the choice of what vocabulary to code for a communication device depends in part on what vocabulary is used most frequently by other groups of persons in similar contexts. The relevance of a particular vocabulary for a user may depend largely on the accuracy of match between predicted vocabulary characteristics and actual communicative needs of the user. The more specific our information on typical vocabulary production in all of the settings in which the communicative system will be used, the better our ability to apply normative data to predict actual vocabulary use.

While the topic areas for this bibliography were chosen to pertain to the development of communication vocabularies, these resources should be useful for persons interested in any application of word frequency information, including development of school resources and textbooks, estimating relative difficulty of sets of words or reading materials, developing core vocabulary lists for specific topics or populations, or prediction of changes in word use with various adaptations of the communicative setting. Topics include surveys of vocabulary studies of both normal and disabled populations for various population samples (children, college students, adults, professionals), communication modes (speech, writing, printed texts), settings (school, home, office, public places), and tasks (free speech/writing, conversation, formal themes or speeches). The bibliography includes frequency studies by special topics, such as technical vocabularies, special population vocabulary use, letter and syllable frequencies, effects of communicative mode of frequency, and cognitive and semantic issues related to frequency of word use.

Some of the sections are not intended to be exhaustive surveys of available references on a given topic, and are noted instead as 'selected references' in the explanatory paragraph. Other bibliographic sources, such as those listed in Section XII, may be useful for additional references in areas not addressed by this bibliography. Further references will be added periodically as available and appropriate to topic areas covered.

This bibliography document was prepared with support from grant G008300 045 from the National Institute of Handicapped Research, Office of Special Education and Rehabilitative Services, United States Department of Education.
TABLE OF CONTENTS

I. Spoken Production - Original Frequency Data .................................................. 1
II. Written Production - Original Data .................................................................. 3
III. Text Analysis - Original Word Frequency Data ............................................... 5
IV. Comparison/Compilation of other frequency studies - not original reports ........ 9
   A. Comparison/Compilation with individual word frequency reported ............... 9
   B. Compilation/Core Vocabularies including rank/group data only .................... 10
   C. Compiled/Core Vocabularies with no relative frequency/rank reported: Includes wordsets not derived from production or text analysis .............................................. 11
   D. Compilation/Core Wordsets reporting number of words used/known only ....... 12
V. Vocabulary Frequency Information for Special Populations .............................. 14
VI. Technical/Specialized Word Frequency Studies .............................................. 17
VII. Diary-Based Vocabulary Studies ................................................................... 19
VIII. Letter/Syllable Frequency Analysis ................................................................ 21
IX. Semantically-Based Frequency Analysis ......................................................... 24
    A. Wordlists/Frequency Studies organized by word meanings ......................... 24
    B. Basic English and Semantic Simplification Wordlists ................................ 25
    C. Readability or Word Difficulty Analysis ....................................................... 25
X. Methods and Theories of Vocabulary Frequency Analysis
   A. Methods of Content/Vocabulary Analysis
   B. Theories of Vocabulary Frequency Distribution and Language Prediction

XI. Differences in Word Frequency and Use by Communicative Mode
    A. Qualitative/Quantitative Differences in Communication by Modality
    B. Information Processing Considerations for Different Communicative Modes

XII. Bibliographies/Reviews of Vocabulary Frequency Studies
     A. Bibliographies
     B. Reviews of Vocabulary Studies

Selected Master's Theses
Related to Word Frequency Analysis

Ph.D. Dissertations and Unpublished Papers Related to Vocabulary Frequency
I. **Spoken Production - Original Frequency Data**

This section lists reports of original surveys of the relative frequency of words/phrases in connected speech. Contexts for collecting samples of speech include face-to-face conversations, telephone conversations, picture descriptions, and formal speeches. Studies also vary according to speaker age, topic of communication, sampling and analysis techniques, and extent of data reported. References in this section are assigned a letter code based on the type of frequency information available in the report cited (although more detailed information may be available in another source), as follows:

- **A** - study reports absolute frequency of occurrence in sample for each word listed in the paper
- **A** - study reports absolute frequency of all words included in sample, including words of low frequency of occurrence
- **B** - study reports only relative rank of words, or relative frequency by groups of words, (e.g. by 100's) instead of individual frequency
- **C** - study reports selected or core (most frequent or important) wordlists with no report of relative frequency or rank of words
- **D** - study reports only number of words produced in sample, with no wordlists or individual frequency data reported


II. Written Production - Original Data

This section lists reports of original surveys of the relative frequency of words/phrases in samples of written communication. Written samples collected in these studies include personal letters, business letters, essays, and other school papers. Studies also vary according to subject age, topic of communication, sampling and analysis techniques, and extent of data reported. This section includes only written communication by individuals, one sample per person; studies analyzing printed texts and newspapers are listed in Section III. References in this section are assigned a letter code based on the type of frequency information available in the report cited (although more detailed information may be available in another source), as follows:

A - study reports absolute frequency of occurrence in sample for each word listed in paper
A* - study reports absolute frequency of all words sampled, including words of low frequency of occurrence
B - study reports only relative rank of words, or relative frequency by groups of words, (e.g. by 100's) instead of individual frequency
C - study reports selected or core (most frequent or important) wordlist, with no report of relative frequency or rank of words
D - study reports only number of words produced in sample, with no wordlists or individual frequency data reported


(D) Clarke, W.F. (1921) Writing vocabularies. Elementary School Journal, 21, 349-351


(A) Horn, E. (1926) A basic writing vocabulary: 10,000 words commonly used in writing. *University of Iowa Monographs in Education, Series 1(4)*.

(B) Horn, E. (1926) The 3009 commonest words used in adult writing. *Fourth Yearbook of the Department of Superintendence, 146-172*.


(C) Jones, W.F. (1915) *Concrete Investigation of the Material of English Spelling*, Vermillion: University of South Dakota.


III. Text Analysis - Original Word Frequency Data

This section lists reports of original surveys of the relative frequency of words/phrases in samples of printed texts and newspapers. Types of text samples collected in these studies include pre-primers and grade school reading books, textbooks, novels, magazine articles, newspaper articles, poetry, comic strips, spellers and vocabulary lists. Studies also vary according to target audience of text, topic of materials, sampling and analysis techniques, and extent of data reported. References in this section are assigned a letter code preceding the citation, based on the type of frequency information available in the report cited (although more detailed information may be available in another source).

- **A** - study reports absolute frequency of occurrence in sample for each word listed in paper
- **A** - study reports absolute frequency of all words sampled, including words of low frequency of occurrence
- **B** - study reports only relative rank of words, or relative frequency by groups of words, (e.g. by 100's) instead of individual frequency
- **C** - study reports selected or core (most frequent or important) wordlist, with no report of relative frequency or rank of words
- **D** - study reports only number of words produced in sample, with no wordlists or individual frequency data reported

Additional bracketed letter codes following some entries indicate whether the information reported is also organized according to difficulty levels by grade level recommendations (G) and/or for use in spelling lists (S).

(B) Ayer, F.C. (1945) *A Study of High School Spelling Vocabulary*, Austin TX: Steck Co. (S)


(B) Betts, E.A. (1940) *Spelling Vocabulary Study: Grade Placement of Words in Seventeen Spellers*, N.Y.: American Book Co. (G, S)


(A) Eldridge, R.C. (1911) *Six Thousand English Words: Their Comparative Frequency and What Can Be Done With Them*, Niagra Falls, N.Y.


(A) Knowles, Rev. J. (1911) The 353 most frequently occurring words in an aggregate of 100,000 words, made from passages from the English Bible and various authors. In Eldridge, R.C. *Six Thousand Common English Words*, Niagra Falls, N.Y.

(B) Krantz, L.L. (1945) The Author's Word List for the Primary Grades, Minnesota: Curriculum Research Co.


(A*) Thorndike, E. (1921) *The Teacher's Word Book of 10,000 Words*, N.Y.:Teacher's College, Columbia University.

(A*) Thorndike, E. (1931) *The Teacher's Word Book of 20,000 Words*, N.Y.:Teacher's College, Columbus University.


IV. Comparison/Compilation of other frequency studies - not original reports

This section lists articles which discuss the relative frequency of words/phrases based on data derived from other sources than direct analysis of original data. These other sources of information include reports of spoken, written, or textual vocabulary frequency analysis cited from another primary research article, as well as judgements of the relative importance of vocabulary items based on personal experience. Some articles in this section are side-by-side comparisons of relative frequencies or ranking of words across different studies, others compile information from several studies to form a single wordlist, while others recommend core vocabularies (based at least in part on frequency considerations) of important words for a certain purpose. Studies vary according to purpose of comparison/compilation, topics and types of materials utilized, sampling and analysis techniques, and extent of data reported. References in this section are divided into four subsections based on the type of frequency information available in the report cited, as follows:

A - studies report absolute frequency of occurrence in sample for each word listed in paper
B - studies report only relative rank of words, or relative frequency by groups of words, (e.g. by 100's) instead of individual frequency
C - studies report selected or core (most frequent or important) wordlist, with no report of relative frequency or rank of words
D - studies report only number of words produced in sample, with no wordlists or individual frequency data reported

Additional bracketed letter codes following some entries indicate whether the information reported is also organized according to difficulty levels by grade level recommendations (G) and/or for use in spelling lists (S).

A. Comparison/compilation with individual word frequency reported


B. Compilation/Core Vocabularies including rank/group data only

Ashbaugh, E.J. (1919) *The Iowa Spelling Scales: Their Derivation, Use and Limitations*, Bloomington, Ill.: Public School Publishing Co. (G, S)


Greene, H.A. (1955) *The New Iowa Spelling Scale*, Iowa City: State University of Iowa. (G, S)


C. Compiled/Core Vocabularies with no relative frequency/rank reported: includes wordsets not derived from production or text analysis

Buckingham, B.R. (1913) Spelling Ability: Its Measurement and Distribution, N.Y.: Teacher's College, Columbia. (S)


Knipp, H.B. (1952) *Basic Vocabulary, Phrases, and Sentences for Early Reading Instruction*, Meadsville, PA: Keystone View Co. (G)


Nicholson, A. (1915) A Speller for the Use of the Teachers of California, State Board of Education: California State Printing Office. (G, S)


Stone, C.R. (1950) Progress in Primary Reading, St. Louis: Webster Publishing Co. (G)


D. Compilation/Core Wordsets Reporting Numbers of Words Used/Known Only


V. Vocabulary Frequency Information for Special Populations

This section lists reports of wordlists and vocabulary frequency information for specified populations (other than groups of non-disabled children or adults). Sources of information for articles include original data of spoken or written production, recommended core wordlists, or principles of wordlist selection based on frequency data for a given population. Special populations discussed include persons who have cerebral palsy, mental retardation, deafness, psychological disorders, aphasia, physical disabilities, and cultural disadvantage. Studies vary according to ages of populations sampled, form and topics of communication, sampling and analysis techniques, and type of data presented.


VI. Technical/Specialized Word Frequency Studies

This section lists reports of frequencies of words commonly used in limited contexts, such as technical or educational settings. Articles may include reports of original frequency data or compiled/created lists of words particularly common for a given setting, and vary in topics, sources of information, and data collection and analysis methods. Examples of specific sources of information analyzed in these reports include professional and military letters, scientific and technical papers, radio broadcasts, literature, encyclopedias, textbooks and materials for various subjects addressed in school classrooms. For additional information on relative frequencies of vocabulary common to specific educational subjects (such as social studies or mathematics), see:


Struck, F.T. (1943) 102 key words. *Industrial Arts and Vocational Education*, 32, 57.
VII. Diary-Based Vocabulary Studies

This section lists reports of wordlists and frequencies derived from samples of spoken or written production from a single child. Most of these studies are compilations of information on children's speech gathered by the investigators in a speech diary, either at a particular age or compared over a longer period of time. Ages of children sampled in these studies range from eighteen months to over six years of age, with some studies comparing a child's productive vocabulary development at different ages. Many of these studies list all words produced by the child, but only a few include individual frequency data for each word.


Nice, M.M. (1917) The speech development of a child from eighteen months to six years. Pedagogical Seminary, 24, 204-243.

Nice, M.M. (1920) Concerning all day conversations. Pedagogical Seminary, 17, 166-177.


Uhrbrock, R.S. (1936) Words most frequently used by a five-year-old girl. Journal of Educational Psychology, 27, 155-158.

VIII. Letter/Syllable Frequency Analysis

This section lists reports of original surveys of the relative frequency of letters, sounds, syllables and combinations of letters. Various studies report letter and syllable frequencies according to word position (initial, medial, final), word length, relationship to other letters or words, or overall letter frequencies or combinations. Often, letter/syllable frequencies are reported as units of various lengths, such as frequencies of bigrams (2-letter units), trigrams (three-letter units), up to ngrams (units of unspecified length). Information reported for these units may be absolute frequencies, relative rank orders of occurrence, or versatility of positions in which they occur. Sources from which reported information is gathered includes conversational speech, written production, text or dictionary analysis.


Topper, G., Macey, W., & Solso, R. (1973) Bigram versatility and bigram frequency. *Behavior Research Methods and Instrumentation, 5*, 51-52.


IX. Semantically-Based Frequency Analysis

This section includes selected references concerning the relationship of word meaning and vocabulary frequency. Some studies provide original frequency data; others compile wordlists according to grammatical or functional properties of words, or discuss theories and potential wordlists based on principles of vocabulary organization and simplification. Topics in this section include the following:

A. Wordlists/Frequency Studies Organized by Word Meanings (including some separate frequency lists for various dictionary definitions of words)
B. Basic English and Semantically Simplified Wordlists (i.e. using a limited set of words to represent multiple concepts)
C. Readability or Word Difficulty Analysis (including formulas and principles to analyze semantic difficulty of words and written text)

A. Wordlists/Frequency Studies Organized by Word Meanings


B. Basic English and Semantic Simplification Wordlists


C. Readability or Word Difficulty Analysis


X. Methods and Theories of Vocabulary Frequency Analysis

This section includes selected references for articles addressing methods, principles, and theories related to vocabulary frequency analysis. Topics discussed include statistical principles of sample analysis, principles of word-frequency distribution, mathematical models of language, and computer analysis of word frequency and content of conversation, texts, and written output. Categories within this section include:

A. Methods of Content/Vocabulary Analysis
B. Theories of Vocabulary Frequency Distribution and Language Prediction

A. Methods of Content/Vocabulary Analysis


B. Theories of Vocabulary Frequency Distribution and Language Prediction


Zipf, G.K. (1935) Observations of the possible effect of mental age upon the frequency-distribution of words, from the viewpoint of dynamic philology. *Journal of Psychology*, 4, 239-244.

XI. Differences in Word Frequency and Use by Communicative Mode

This section lists selected references that discuss potential effects of communicative modes on the aspects of linguistic production. Different types of communicative modes discussed include spoken, handwritten or typewritten conversation; problem solving using speech, writing, typing or a variety of communicative devices; and interacting with a computer or machine by a variety of input methods. Potential modality effects on language discussed in these articles include changes in length and complexity of language, linguistic or problem solving style, speed of communication, ease of operation of communicative mode, and cognitive processing load of using various arrangements within communicative modes. Topic headings in this section include:

A. Qualitative/Quantitative Differences in Communication by Modality
B. Information Processing/Cognitive Limitations for Different Communicative Modes

A. Qualitative/Quantitative Differences in Communication by Modality


**B. Information Processing Considerations for Different Communicative Modes**


XII. Bibliographies/Reviews of Vocabulary Frequency Studies.

This section lists resources which reference or review studies of vocabulary frequency and word set analysis. Studies included in the review listing either discuss word frequency studies singly, or compare relative merits and problems of different approaches to word frequency analysis.

A. Bibliographies


B. Reviews of Vocabulary Studies


Lindsay, W.B. (1932) The Coleman Spelling Vocabulary and How it Was Obtained, Chicago: Hall & McCreary Co.

Lorge, I. (1944) Word lists as background for communication. Teachers College Record, 45, 543-552.


Selected Master's Theses Related to Word Frequency Analysis


Oberman, M.H. (1933) A study of the degree to which the written vocabulary of children has been determined. Master's thesis, University of Iowa.


Ph.d Dissertations and Unpublished Papers Relevant to Vocabulary Frequency


Ashbaugh, E.J. Vocabulary of high school pupil's letters. Unpublished study, University of Iowa (in education office), Iowa City.


Fennell, E.L. (1924) An analytical study of the vocabularies used in primary readers. Unpublished study, University of Chicago, Dept. of Education.


