LEXGUIDE-2003

A GUIDE TO THE LEXICAL ANALYSIS OF NATURAL TEXTS USING QLEX or QANALYSIS

Donald P. Hayes

CORNELL UNIVERSITY

SOCIOLOGY TECHNICAL REPORT SERIES
# 2003-1

` MAY, 2003

TABLE OF CONTENTS

I. THE CONCEPT OF LEX: A SCIENTIFIC MEASURE

LEX describes an English text's accessibility, lexical difficulty and comprehensibility.

LEX shows (a) how the 10,000 most common grammatical and content words in the English language were used in a specific text, and (b) the text's use of the uncommon and rare terms, i.e., those ranked beyond the first 10,000. Each text's word use is compared to a common Reference Lexicon (Carroll, Davies and Richman, Word Frequency Book, 1971). In that lexicon, the most commonly used word is: 'the' (occurring ~73,000 times per million words in natural texts), the 10,000th ranked word is 'tournament' (occurring 2.8 times per million).

LEX meets five requirements for a natural science measure: validity, reliability, robustness, precision and stability over time. It is based on a model of word choice–one characterized by the lognormal statistical distribution.

.

LEX scores have both a sign and magnitude, as in LEX = +18. When the word choice in a text is skewed toward common words, LEX scores are negative., as in mother talk to her child in the home. When word choice is skewed toward the rare words of English, LEX signs are positive, as in research articles published by Nature – all have positive LEX levels. The magnitude of the LEX score describes the degree to which word choice is skewed, when compared with (a) a model of word choice, (b) the highly correlated Carroll et al's corpus, and (c) the empirical linear pattern of word choice found in English-language newspapers since 1665.

LEX's precision is conditional on the text's sample size and on how the sample was derived: by stratified simple random sampling or by some less systematic sampling procedure. LEX is suitable for both old and modern texts and should be suitable for analyzing English texts well into the next century -- provided they are edited to LEX standards For texts derived by 24 stratified SRS passages of >150 words each, the standard error of measurement is 1.3 LEX--i.e., 1.3 units where natural texts vary in LEX from +58 to -81. Normally, LEX scores are based on text samples of 1,000 or more words, in ten or more sub-samples of 100 or more consecutive words, in full sentences, taken by stratified random sampling methods. In 1000-word such samples, the standard error is closer to 2 LEX. While not yet proven, it is believed that the substitution of French, German or other Reference Lexicons for the English lexicon should yield equally strong measures of a text's accessibility and lexical difficulty.

LEX 's relationship to 'Readability Indices'. LEX is inappropriately linked to 'readability' measure (e.g., the Flesch-Kincaid, or Gunning Fog indices). Flesch developed his pragmatic tool to aid primary school teachers in making quick decisions on the suitability of books for their students--taking into consideration their reading skills, level of comprehension and interests. His 'readability' measure was a composite of two measures: (a) the fraction of words with six or more letters, and (b) the text's average length of sentence--in words. LEX and Readability scores for the same text are positively correlated (r= ~+.70) but 'Readability' indices are not grounded in a theoretical model, have not been well validated by normal scientific standards , nor are they correlated as highly with multiple criteria---as is LEX . Furthermore, 'readability's' sentence length component is uncorrelated certain types of conversation texts nor is it strongly correlated with newspaper texts–whose lengths are long, but not difficult. LEX is supported by a large, carefully constructed, comparative data base of over 5000 natural texts selected to represent the full spectrum of texts -- the Cornell Corpus-2000.

Another reason LEX and readability scores produce different estimates for the same text's difficulty is that 'readability' texts are *not edited*--- thus they are not comparable with each other nor with LEX texts. All LEX texts are edited according to a comprehensive set of transcription rules--a necessity step if two or more texts are to be compared and their differences interpreted correctly. Text difficulty' is a multi-dimensional concept with important grammatical, metric, semantic and lexical components. LEX measures the lexical component. It has proved to be difficult to compare the relative contributions of these several components in overall text comprehension.

LEX has been validated in numerous ways: by experiments in which predictions were madeas to the extent and direction of changes in LEX levels under experimental conditions; by comparisons with other less well-validated measures; by its compliance with the model of word choice; and by several substantive studies. These include: (a) the 'dumbing down' of American basal readers, beginning in 1947, when compared with pre World War II basals; (b) the growing inaccessibility of science journals to the educated reader, since World War II; (c) the comparative richness of the natural language experiences of children growing up in underclass through upper-middle class families; and (d) by predictions of the same witness's LEX level when testifying under direct and cross-examination.

LEX's interpretation as a measure of a text's 'accessibility', ‘lexical difficulty’ and ‘comprehensibility' has also been validated by internal comparisons among the 5000+ texts edited and analyzed in a file named Cornell Corpus-2000. This Corpus includes every major text category: broadcasts, print, and conversation, both formal and informal.

II. LEX is based on a model of word choice patterns.

The abstract generalizations formulated in mathematics, in physics or in syntax are among mankind's most powerful intellectual achievements. The power of those generalizations is due, in part, to their essential independence from specific content. For Galileo's laws of motion, it does not matter (theoretically) whether the falling body is a planet, feather or cannonball, so long they are in a vacuum. In mathematics, the numeric units may be dollars, weights, distances--which specific unit of those concepts does not matter, since the fundamental relations refer to a higher order of abstraction. Similarly in syntax, which specific noun is used in a phrase or sentence is unimportant--what does matter for syntax is that word's grammatical role in the sentence.

The model underlying the LEX statistics also describes an abstract feature of all texts: its pattern of word usage. One cannot detect these patterns by simply reading a text because ***the***

order in which words appear, so important to syntax and semantics, is irrelevant to the text's pattern of word choice. The words of any large text may be randomized in a thousand different ways, yet the pattern of word choice and the LEX scores will remain the same. Pattern of word usage is a higher order concept.

The model used here for patterns of word usage states that the probability of a word's choice in a text is conditional on the log of that word's general frequency of usage. That implies that each person's lexicon reflects their differential experience with words in context, in the capacity for its retention and in their generalization of those experiences..

A package of programs named **QLEX** (before 2003) or **QANALYSIS** (since 2003) identify these patterns of word choice. From them, the programs produce the LEX statistic from a cumulative proportion distribution. This is done by first determining what fraction of a text's words is 'the'--English's most common word. To that proportion is added the proportion of English's second most common type--'of'; the fractional use of the third type 'and', and so on through all the 10,000 most commonly used word-types in English. The resulting cumulative proportion distribution supplies the text's pattern of word choice. **Figure 1** describes the use French journalists made of their lexicon in a full year's issues of the Paris newspaper, *Le Monde.* The X axis ranks those first 10,000 French terms (by their logs). The Y axis is simply the cumulative proportion of all words in *Le Monde*. The first ten word-types account for ~23 percent of all words used.

FIGURE 1 -- word choice distribution in *Le Monde* (1995)

LEX and its related measures describe these patterns of word choice. In every QLEX/QANALYSIS analysis, the ranking of word types in the Reference Lexicon (always represented along the X-axis by the log of each word's rank) is based on two bodies of evidence: the largest and currently the most diverse corpus of English word usage, Carroll, Richman and Davies' The Word Frequency Book (1971); and the pattern of word choice in English-language newspapers (an early finding in the course of this line of research). These first 10,000 most commonly used word types are **a *constant*** for every analysis, i.e., their ranking of the first 10,000 most common English word-types serves as the **Reference Corpus** against which every sample text's actual choice of words is compared. When a new, well-designed contemporary, far larger, equally or even more diverse in domain coverage corpus becomes available, Carroll, et al's corpus will be replaced and that new Corpus would serve as this Reference Lexicon. Presumably its ranking and estimated frequency of usage per million words will be more precise–but will remain highly correlated in rank with that in Carroll, et al. and older corpuses (e.g., Thorndike and Lorge, 1944)---well above r=.97.

How well do actual texts fit this model? **Figure 2** describes the empirical pattern of word choice in 63 newspaper samples drawn from all over the world, between 1665 and 1992 (once their archaics and spellings are modernized, as required by LEX editing rules). The patterns of word choice in these newspapers (based on ~1200 word samples), were then compared against the expected linear pattern of the model. In these sample newspapers, the mean, the Q1 and the Q3 papers all have approximately a linear pattern across the range word rank 1 to 10,000. Each sample was comprised of ten or more sub-samples of 100 or more words. Not only do those three newspapers' distributions approximate the model, they are parallel–showing that their small but differential use is primarily due to a single word--'the' . When their differential uses of 'the' are equated, their three word choice distributions fall virtually on top of one another -- despite the immense diversity of subject matter.

FIGURE 2 -- word choice distribution in 63 international newspapers

The standardized educational and occupational testing industry capitalized on these powerful regularities in word choice by constructing 'word knowledge' tests. Empirically, children were found to rank the words of the English lexicon in largely the same order as those in Thorndike and Lorge, Dale, Brown University and the Carroll, Richman and Davies corpora. Independent tests showed that a test word's 'difficulty' was highly correlated with its general frequency of usage--'hard' words were usually uncommon or rare, 'easy' words were nearly always common. Those standardized tests were validated by the discovery that students who scored well on word knowledge tests generally know more of the uncommon words of the lexicon, had higher levels of reading comprehension, academically performed well, and achieved at a higher levels on other texts of verbal skills. Research by cognitive psychologists also found that acquisition of, access to, and retrieval times for words from a lexicon were strongly related to a word's general frequency of usage. .

Figure 3 describes three word choice patterns representing much of the effective range of LEX patterns. The topmost distribution represents the pattern of word choice of 32 mothers' while speaking with their child at home (all the children were 30 months of age at the time). The mean distribution for the mothers' speech is strongly skewed toward common words (relative to the model's distribution, and that of newspapers). The middle distribution describes a single 1000+ word sample taken from the New York Times (1987). It too approximates the average newspapers' linear pattern. The bottom of these three distributions is the pattern of word choice in main research articles in Nature (1994). Every article in that general scientific journal was skewed toward rare words, and that skew has grown in each decade since World War II.

FIGURE 3 -- the range of word choice distributions

III. LEX uses: theoretical and applied.

The first use of LEX was to test a niche-like theory of how text difficulty shaped the dynamics of the science publishing industry since World War II. As the number of scientists grew, adaptations were made by the science journals and magazines. The principal hypothesis tested was that science magazines have to dispersed themselves (lexically) so as to minimize competition for readers and advertisers.. When magazines sought to occupy the same lexical niche (defined by its LEX score) , two outcomes become apparent: either the magazine shut down or the publisher moved his journal into another lexical niche by changing the mean LEX level of its principal articles.

For example, a series of musical chairs-like moves was initiated in the science journal industry when, in 1947, Nature vacated its historical niche (LEX = ~0.0)--the mean level of a newspaper. By 1950, Nature's major articles had risen to LEX = +17, reflecting a decision by the editor permit researcher/authors to write specifically for fellow scientists and to largely ignore the interested, but not technically educated, reader. Science held to its traditional niche (LEX = +6) until 1960, when it too abandoned its old niche and followed Nature's lead, now. publishing its major articles at higher LEX levels. Whether Science moved because Nature was getting the better papers is not known.. Science's time series reached LEX = +15 by 1970 [by which time, Nature's articles were averaging LEX = +25].

Sensing Science's now-vacated niche, the publisher and editors of Scientific American changed its LEX level (from ~0.0 to LEX = +5 by 1970), effectively moving into Science's former niche. By 1980, Scientific American's average article had risen to LEX = +10. That must have exceeded many of its subscribers' ability to understand the articles because within a short period, well over a hundred thousand subscribers failed to renew their subscriptions. That, in turn, induced many advertisers to withdraw. That loss of resources weakened the magazine's financial position, and made it a target of a corporate acquisition.

By vacating its traditional LEX = 0.0 niche (which Scientific American had long shared with Popular Science for decades), that now-empty niche was viewed as an opportunity by the editors of four new science magazines, who felt there was room for a magazine designed for the educated reader interested in science. Those four were Science Digest, SciQuest, Science-80 and Discover. Each sought and occupied that same LEX = 0.0 niche. Unfortunately, there were insufficient subscribers and advertisers to sustain them all . The first three ceased operations by 1986. By vacating that niche, still another science magazine filled it , New Scientist--whose texts rose to LEX = +7 by 1990 (Hayes, 1991). Similar changes and outcomes were experienced by several other science/technology magazines, including the weekly Science News. LEX levels shape audiences and writers–which has substantial effects on publication success.

Every major professional science journal’s time series shows a rapid expansion of specialized, technical language and higher positive LEX scores since 1950. In consequence, professional science is now largely inaccessible to the general college-educated reader. That increases the public’s reliance on intermediaries to select and translate developments in science forus (e.g. the New York Times and Washington Post’s science sections and the rapid growth in the coverage of science and medicine by television networks).

Finally, high school science textbooks have grown far more difficult than the textbooks which are used in the rest of the 12th grade curriculum, e.g., history, English and social science texts. Science texts’ relatively high LEX levels probably contributes to student avoidance of the elective advanced science courses -- which contributes to the declining fraction of scientists, engineers and software specialist trained in the United States which increases our dependence upon the foreign-trained.

A second illustrative theory-test using LEX measurements was to help resolve an empirical dispute between behavior geneticists (BG) and most social and developmental scientists. At issue is the relative importance of children's environments and experiences (particularly their natural language experiences) in shaping children's verbal achievement. Recent BG research on adopted children in families with other siblings finds that such language experiences are 'essentially' the same for children growing up in well-off and relatively poor families. In contrast, social and developmental scientists point to their own voluminous research showing language experiences are different for children growing up in underclass or upper-middle class families.

LEX was used to measure the difficulty of the natural language experiences children have with: (a) the texts of the television programs they choose to watched 'regularly'; (b) the texts of the books and magazines they chose to read for themselves, and (c) the texts of household conversations children have with their mothers. Exceptional for this kind of research, and unlike the samples used in the behavior genetics research, the children in these studies were statistically representative samples of all British children, all public school California 6th graders and families in the Bristol England metropolitan area born within a week of one another.

Differences in the richness of children's language experience of those growing up in different social class backgrounds were found (consistent with social and developmental research), but all those differences were small (several LEX only). Differences in the quantity of language experience was also small. However, those small differences were pervasive and persisted throughout childhood in our cross-sectional and time-series data. Differential quantity and difficulty of language experience was found at every age from 30 months through 14 years of age. One implication is that small, persistent differences in language experience, may produce cumulative effects on childrens' general domain knowledge (e.g., their knowledge of baseball, genetics or finance), affecting their verbal skills (which include the extent of their conceptual knowledge and lexical development). Such differences would doubtless contribute to their reading comprehension and academic performance and may help to account for the world-wide pattern of higher verbal skills among children with richer language experience (Hayes, Wolfer, Roschke, Tsay and Ahrens, 1999).

A final example of LEX's theoretical use is in testing possible mechanisms affecting a person's pattern of word choice. Two have been explored--personal stress and lack of preparedness. In one test of courtroom testimony, witnesses often must answer questions put to them by two attorneys: one under direct and the other under cross-examination. LEX analyses document that virtually every witness in the Patricia Hearst trial gave lexically simpler testimony in response to cross-examination (where the interrogator is the opponent's attorney and whose job it is to challenge the witness' credibility, and if possible, discredit that testimony with the jury and judge (Hayes and Spivey, 1999). Both cognitive scientists and trial lawyers predict this finding and attribute it to the distracting effect of high stress and lack of preparation on the witness' word choice under the pressure of realtime text production

LEX has also been used in applied research. One study sought to answer this question: What can explain the nationwide decline in SAT-verbal scores in the mid 1960-late 1970s? Mean nationwide scores had remained at near-constant levels between the mid-1950's through the early 1960s despite rising numbers of those tested (the highest mean level recorded was 1963), but then declined in each of the next 16 consecutive years, stabilizing in 1978. Mean SAT-verbal scores have remained low ever since. To explain this SAT-verbal time series, Chall (1967, 1977) hypothesized that one major contributing factor was the lowering of academic standards--including the 'dumbing down' of the schools' curriculum, reduction in homework and difficulty of the principal school reading materials used by students throughout the nation.

In this application, LEX was used to measure the difficulty of nearly 800 basal readers used across the United States during three time periods: 1919-1945, 1946-1962, and 1963-1991. We found that LEX levels of basal readers published in the US between 1946-1962 period declined sharply from their pre-war and wartime II levels. (First grade texts were not simplified in the UK--so the American decline in LEX level was a choice made by educational leaders and their school publishing colleagues). Since the major schoolbook publishers marketed their basal series nationwide, one effect of these new basals may have been to reduce the range and depth of domain and conceptual knowledge which student derive from their basic texts. A primary task of schooling is to broaden the range of a child's domain knowledge. Simplifying texts narrows the breadth and depth of conceptual and lexical resources available to students which may have contributed to the reduced verbal achievement on the SAT-verbal tests when these students took the SAT-verbal test some years later toward the end of their high school experience.

The conventional wisdom in education on this question attributes this decline to changes in the composition of those taking the SAT-- many more and less-select students began to take the SAT. That explanation fails to explain why the SAT scores began to fall in 1963 when it had remained stable even while the number of less-qualified students taking the test increased.
It also fails to explain why the scores declined for those 16 consecutive years, since the number taking the SAT had stabilized at or about 1963. Nor can that explanation explain the low level of mean verbal achievement since 1978 to the present. Nor can it explain why verbal achievement declined especially among those at the highest levels of verbal achievement. Despite many more students taking the SAT, there was an absolute decline in the number scoring over 600, and an even greater proportional decline in those scoring over 700. Something affected the achievement level of even the most able students--those who had always taken the SAT.

LEX analyses supplies evidence compatible with Chall's hypothesis: schoolbooks were shown to have become much less difficult after 1947 and a side-effect (doubtless unintended) of that ‘ dumbing down’ of textbooks was to lower the students' domain/conceptual knowledge and skills. When, years later, the students took the SAT, those cohorts whose verbal SAT scores declined were the ones who had first encountered the simplified texts (Hayes, Wolfer & Wolfe, 1996). Cross-lagged correlation of basal reader levels throughout their elementary and secondary schooling with the national SAT-verbal scores is generally consistent with this hypothesis. Since LEX levels of contemporary basal readers and major textbooks have remained unchanged since 1978, there are no grounds to expect verbal scores to rise -- and they have not risen in the last 25 years -- according to the SAT and NEAP testing.

Worth of note is that when students, teachers and parents complained vigorously about the low verbal SAT verbal scores, ETS finally gave in and "re-calibrated" the verbal scores upwards, especially for those at the upper range. They did this knowing full well that their own research showed the comparability of tests with those in use in the late 1950s and early 1960s. The decline in verbal aptitude was real and persists to this day. The scores were changed to end the uproar against ETS when they revealed an unpalatable truth.

IV. CONDITIONS FOR USING QLEX/QANALYSIS

The QLEX/QANALYSIS package of programs provides researchers with tools for measuring important facets of any English text--its accessibility, comprehensibility or lexical difficulty and for its patterns of use for the principal grammatical/closed class terms.

These programs were developed across several generations of mini-computers and PCs, using different operating systems, several programming languages, and file formats. Most of this work was done under the DOS operating system.

You are free to use QANALYSIS *so long you agree to these conditions*: this program is copyrighted by Marty White . QANALYSIS is for research purposes, is not intended for commercial purposes, under this or any other name. The programs were designed to provide scientists with a scientific measure for text analysis. There is no guarantee the programs will run on your machine; nor are they guaranteed to be free of errors (though QLEX 5.1–the predecessor program) has been used thousands of times without problems. Scientists are encouraged to continue LEX's development as a scientific tool.

I would like to be notified of changes you make, the justification for them, and the effects those changes have on LEX measurements. My e-mail address is **dph1@cornell.edu**.

V. PREPARING TEXTS FOR QANALYSIS.

QLEX/QANALYSIS are the names for software packages which carry out a systematic lexical analyses on any LEX-edited English text. It may be used on any text: spoken, broadcast or printed. First developed and put to substantive used in 1980, QLEX/QANALYSIS continues to evolve as more we learned about the underlying theoretical model, the mechanisms, measures and their interpretation. This is a work in progress. First reports on LEX's uses appeared in technical papers in the Cornell Sociology Technical Reports series catalogued in the computer data base for research universities called RLIN; in multiple presentations at the AAAS annual meetings; and in papers given at psychological and sociological professional meetings. Recent descriptions of models and research are reported in Hayes, 1988; Hayes and Ahrens, 1988; Hayes, 1992; Hayes, Wolfer & Wolfe, 1996; Hayes, 2000).

A. Text files must be in ASCII (have the **.txt** extension) To be run on QANALYSIS, your text must be an ASCII file.

What should a text file look like after it has been edited? The following is an example of an EDITED FILE taken from an Internal Revenue Service publication:

> &&000 IRS1040.ASC
> This publication was sent by the IRS to all taxpayers who had filled out Form 1040 in 1986. It describes some changes in the tax law which affect the 1987 tax. A stratified SRS from the first 19 pages is shown.
>
> &&111 This year for the first time the =TaxReformAct of =1986 will have major impact on the preparation of your tax return. This has important consequences for you and for us. Changes made by the new =Act are summarized on the next page. You can learn more about the ones that affect you by getting one of the publications listed near the end of this booklet. Learning about the changes now will make it easier for you to prepare your return when you start working on it.
>
> Increased standard deduction. The standard deduction (formerly the zero bracket amount), has increased for most individuals.
>
> Alternative minimum tax. The tax rate has been increased to =21 percent and several tax preferences have been added or deleted.
>
> Allocation of interest expense. Whether your interest expense is subject to the new limits that apply to personal and investment interest depends on how and when the loan proceeds were used. Special rules apply in determining the type of interest on loan proceeds deposited in a personal account, such as a checking account.
>
> Joint or separate returns. Generally, married couples will pay less tax if they file a joint return because the tax rate for married persons filing jointly is lower than the tax rate for married persons filing separately. However, as a result of some of the changes in the tax law, such as the increased income limit that applies to medical and dental expenses and the new individual retirement arrangement deduction rules that apply to certain individuals, you may want to figure your tax both ways to see which filing status is to your tax benefit.
>
> Married persons who live apart. Some married persons who have a child and who do not live with their spouse may file as head of household and use tax rates that are lower than the rates for single or for married filing a separate return. This also means that you can take the standard deduction even if your spouse itemizes deductions. You may also be able to claim the earned income credit. Children of Divorced or Separated Parents. The parent who has custody of a child for most of the year, the custodial parent, can generally take the exemption for that child if the child's parents together paid more than half of the child's support. This general rule also applies to parents who do not live together at any

time during the last six months of the year.

1. ID (Identification) Codes The actual words of a Text prepared for QLEX/QANALYSIS fall into two types: (a) words QLEX/QANALYSIS is to include in its analysis, and (b) words QLEX /QANALYSIS is ***to ignore***. To distinguish these two ID types: **&&000** is placed at the beginning of any passage which QLEX /QANALYSIS is **to ignore** (e.g., information about this text file, its date, how the text was produced, when it was edited, by whom, and details of the sample, its context, etc); and **&&???** (any number between 1 and 9, **but not zero**) for any words you want included in QLEX /QANALYSIS ' analysis. In the IRS document (above), &&111 is the ID number used for analyzing all text and words except those in the Header, which begins with &&000 -- to designate that those words are not to be analyzed. .

If the expression &&000 begins a line or a passage, QLEX /QANALYSIS will continue to ignore everything thereafter until it comes across another ID symbol, i.e., QLEX /QANALYSIS assumes that all words after the &&xxx symbol (e.g., &&111) should be analyzed--until it comes across the &&000 symbol. It will resume analysis when it encounters the next &&xxx.

2. Header and Comments The ID code &&000 must appear at the top left edge of the text, as shown in IRS1040.ASC. Normally, a text file is described by identifying its source (e.g. stratified simple random sample taken from the New York Times, July 5, 1998) or, if it is a conversation, the cast of characters (e.g. Mother and her daughter, Jennifer, age 4 years, 3 months, in their home). The context of that conversation could also be described (Jennifer is tired after having had no nap. This conversation took place in the kitchen. No one else was present). You may place such comments anywhere you like in your text, so long as those comments appear after the &&000 symbol – which is always placed at the far left of a line, followed by a space. After you complete a comment, don't forget to use the ID &&xxx (e.g. &&111) at the beginning of the line where QLEX /QANALYSIS is to resume its analysis.

For example, &&121 could be a mother (speaker identified as #1), talking to her child (the principal target listener – identified as #2), as in: this brief conversation segment:

```
&&121 Bring me some butter from the refrigerator.
    You have to have yummy ingredients when you make
    cookies.
&&211  Do you want the whole box or just one of the pieces?
    Oh, there's only one piece left in the box!
```

The first number is always the speaker, the second digit is always the principal target of that remark, and the third digit may be used for designating the context for the conversation (e.g. #1 is mother-child talk while in the yard, #2 is while they are talking at the grocery; #3 is while they waited for a brother to get out of school), or this third digit may be used to keep track of the different segments in a time-series. For example, in that mother-child conversation, the third digit could represent segment 3 of a multi-part conversation. For analyzing segment #3, one would use ID &&123 to instruct QLEX/QANALYSIS to search for text spoken by the mother to her child, in segment #3 only; &&215 would ask QLEX/QANALYSIS to search for the child talking with her mother, in segment #5 only.

Most analyses of conversation do not distinguish between the several speakers of a conversation or different segments of the same text. In that case, simply put &&111 in front of the first word of their conversation--that will suffice for the entire text (as in the IRS sample). In some conversations, however, you may wish to distinguish what each person said to another, as in what President Nixon said to John Dean as opposed to his principal adviser, Robert Haldeman.

VI. EDITING RULES.

A. ***Quick-and-dirty text editing***. Use any transcription rules, including none--so long as you use the *Bare-Minimum* editing rules . QLEX/QANALYSIS will run on texts with bare minimum editing, you will have output, and can get the LEX statistics. Those values will be in the general ballpark of genuine LEX scores calculated with a fully edited text. Unedited text LEX scores will not be comparable to the LEX values of the 5000+ texts in the Cornell Corpus-2000 -- important for interpreting your findings.

If comparability and precision is needed, you should use LEX standard text editing procedures below. Editing takes time, close attention to detail, and it can be lugubrious--the cost of obtaining a scientific measurement.

B. **Standard LEX** text editing rules**.** Texts should be transcribed according to normal conventions of spelling and orthography--as found in an unabridged dictionary.

There are two exceptions: (1) QLEX/QANALYSIS analyses work on **word-types**, i.e. every uniquely spelled variant of a term is treated as a distinct type in the English lexicon (e.g., while 'boat' and 'boats' share a common stem, they are different word-types in lexical analysis); and (2) the distinction between capital and lower case versions of the same type is not retained in QLEX or in QANALYSIS analyses . Consequently 'the', 'The' and 'THE' are combined as ‘the’. Misspellings are corrected under the LEX editing protocol.

All terms in a text should be transcribed, even if they cannot be found in an unabridged dictionary, e.g. new words, word-fragments and filled pauses. Word-fragments and filled pauses are often disregarded by transcribers making their transcription unreliable unless done by trained, conscientious workers. Some substantive areas of psychology, linguistics and sociology, consider such information useful for some analytic and interpretive purposes.

PROPER NAMES, PLACES, PRODUCTS AND SCIENCE TERMS. Dictionaries are not fully consistent in their treatment of proper names, so it is necessary to impose a common practice for all such terms. Before every proper name, place, or product is inserted an equal sign (e.g. =Mary, =FBI, =Chicago; =Coke). Multi-term names (e.g. =RiodeJaneiro) should be run together since they form a single unit -- otherwise it would be treated as three 'words'. Care must be taken with such terms as '=honey', '=love', '=darling' and '=dear' (when the name refers to a specific individual-- as opposed to the generic 'dearie' waitresses might use for their customers.

Science makes heavy use of contractions and technical terms to avoid the repeated use of long phrases. Chemical compounds, e.g. NaCl, should be =NaCl, but Na by itself has no = sign because it is a recognized dictionary entry. Biology and chemistry texts pose complex transcription problems with their vast sets of technical terms and names. The general rule is: if it is a proper name (often a Latinate species or family name), place the equal sign before such terms. Normal dictionary entries, words like hormone or proteins, however, do not have the = sign before them.

NUMBERS. Some numbers (e.g. 'one' through 'ten') are dictionary entries so the equal sign is not used for them. However, Arabic and roman numbers (except when written out as one, two, .., ten) do have an equal sign placed before them since they are not ordinarily treated as part of the lexicon: e,g, =forty, =1492; =$28; =43 years old, =IX). If the number is an entry in a dictionary, no = sign is used.

DECIMALS, LARGE NUMBERS, EQUATIONS AND COLONS. Dictionaries are inconsistent in their handling of numbers, so they require special handling. The number 9.95 should be transcribed as =9'95 because periods are reserved as the sentence-ending symbol in the QLEX /QANALYSIS's sentence sub-routine. The comma in =360,000 is converted to =360'000, otherwise the comma would divide that number into two terms, 360 and 000.

Equations seldom appear as entries in dictionaries. They are transcribed with the expression =equa. The colon with time (e.g. =11:23) is dropped, for the same reasons.

SOUNDS AND SPECIAL COLLOQUIAL EXPRESSIONS. Many common terms used to communicate a message or mood are not included as entries in unabridged dictionaries. They are given the equal sign. Examples include =meow, =Phew! =Zot! =Gosh, =bang, =pop, =Ha, =Ouch!, =Whee, and =Hurrah!

PUNCTUATION. Use normal punctuation for printed texts, but punctuation of natural conversation is more difficult since run-on sentences are common and one often needs a good recording to detect the shifts of intonation at the end of a sentence. Punctuate texts from the intonation, pauses, substance and common sense. The ability to go over the same passage again and again is essential to reduce the arbitrariness of these decisions. Since reliability of punctuation from spontaneous speech is not as high as from print, statistical differences between two text's MLU (mean length of utterance--based on words) may be due, in part, to unreliability and arbitrariness in punctuation.

CONTRACTIONS. As a general rule, type words as they appear in print or as spoken (unless otherwise specified above or below). The exceptions include such terms as 'cause, which dictionaries treat as because, and terms which are contracted in casual speech, e.g. 'whatever's', meaning whatever is, should be decomposed into 'whatever is'. There are many such contractions, and all should be treated this way.

REGIONAL EXPRESSIONS. Regional accents may alter a word's form or phonology to such an extent that the referent may not be clear to persons from outside the region. In such cases, use the common referential term to replace the exaggerated or local form. For example, while working-class Scots use English, sometimes they substitute Scottish terms which require transcription to their near-equivalents in English (e.g. the word 'fit' in Scottish is normally translated as 'make', in English). Extreme working-class accents in London and in the southern United States require similar treatment.

HYPHENATION. "Cost-of-living" and "first-aid" are complex expressions which appear in dictionaries in that form. They should be transcribed with their hyphens in place. Hyphenation is sometimes missing in similar expressions, such as 'thank you' and 'ice cream'. Transcribe these terms/expressions with the missing hyphen: as thank-you and ice-cream. This prevents QLEX /QANALYSIS from treating such expressions as two (or more) separate words. Since language is dynamic, the time may come when this and other rules may have to be changed.

SPACING. There must be a space between every word, unless there is (or should be) a hyphen, an apostrophe, etc. A space separates an end-of-sentence symbol (period, qustion mark, exclamation mark and the special interruption symbol [@] from the first word of the next sentence. For example: "Stop doing that!You know that!" is wrong because 'You' was not separated from the previous exclamation mark. In QLEX /QANALYSIS, the colon and semi-colon are not considered sentence ending marks.

SPLITTING WORDS AT THE END OF LINES. QLEX/QANALYSIS are a capable set of programs but they do not handle instances of wrap-around and split words at the end of lines. Be sure that words and expressions are not broken at the end of a line.

QUOTES. Use the symbol ", not the symbol ' for quotes, because ' is reserved in QLEX /QANALYSIS for contractions and numerical expressions. Actual quotation marks (“) are ignored by QLEX/QANALYSIS programs.

INTERRUPTIONS. Aural interruption of one person's speech by another is handled in many different ways in psycholinguistics. Under QLEX/QANALYSIS, when one speaker is interrupted by another, instead of ending that person's passage with the usual sentence ending, use the special symbol (@), meaning that this is the final word of a passage spoken by a person who had been interrupted–this is in place of the period, question mark or !. It is not uncommon for the interrupter, in turn, to be interrupted. If that happens, the interrupter's turn should also be ended by the @ symbol. QLEX /QANALYSIS can determine who was interrupted and how often each speaker from their unique ID (e.g. &&231). A new ID code must be used before the interrupter's first word to show the person holding the 'floor' has shifted to a new speaker. This convention does not show the precise point in the text where the interruption took place, nor how many words were spoken by the two speakers simultaneously. That requires closer inspection of the text.

PERIODS. Under QLEX/QANALYSIS, periods are reserved exclusively to mark the end of sentences. In the case of abbreviations like Mrs., Dr., or in i.e., and e.g., the periods must be omitted, otherwise the sentence length measures would be invalid.

Also, in some printed texts, unfinished sentences are sometimes expressed by a string of periods--a convention suggesting that the voice tailed off. Each period would be incorrectly interpreted as a sentence of zero word length, unless all but the final period is omitted.

MISSING OR UNDECIPHERABLE WORDS AND THE =*ZZZZ* SYMBOL. In spontaneous conversations recorded in their natural contexts, background noise or poor recording quality can

make it difficult or impossible to detect a missing word or phrase. A missing word is transcribed by the unique expression: =zzzz. When a whole passage is missing, use the Comment symbol (&&000) at the beginning of a line to indicate the nature of the problem and its approximate length. One indicator of a text's quality is the frequency of the =zzzz symbol's use.

BRITISH vs AMERICAN ENGLISH. Numerous words are spelled differently in the UK and USA, e.g. grey and gray; practise and practice. Use the American spelling convention, since the Reference Lexicon used for these analyses (Carroll, Richman and Davies, 1971) was derived from American English texts.

FILLED-PAUSE SPELLINGS. Use these conventions for filled pauses: =uh-huh (i.e. I acknowledge, agree--usually spoken with rising inflection); =un-huh (I disagree, no--with falling inflection); =um (I follow you, I'm listening); =uh (hold it, I'm groping for the right word or what to say next) ; =oh-oh (a problem)..

FALSE STARTS and FRAGMENTS. All false starts, incomplete phrases and repetitions should be typed, as produced. Word fragments should be preceded by the equal sign since they are not recognized as words in dictionaries.

PRINT CONVENTIONS REPRESENTING CONVERSATION. There is a publishing convention about conversation which requires a change to avoid invalid sentence measures. Publishers do this: "Where is it, Mom?" said Jane. The problem is that there are two sentence endings in that one sentence. Transcribed, it should be: "Where is it, =Mom," said =Jane?
A comma is substituted and the question mark shifted to the end of the sentence.

FOREIGN LANGUAGE TERMS AND EXPRESSIONS. Use comparable Americanisms where possible, unless there is no suitable expression--in which case, type the foreign expression with the words run together and with one equal sign in front.

MONEY. Convert money (e.g. pounds, marks, yen, and francs) into near-dollar equivalents. The numeric values will be wrong but the concept is correct.

TECHNICAL TERMS. Type--as shown or spoken.

RARE WORD REPETITIONS. On occasion, a rare term is repeated many times. For example, in the 1,000+ word text sample taken from several Woody Woodpecker cartoons, the word 'woodpecker' appeared 18 times. That one rare word had the effect of making the text's 'lexical difficulty' appear greater than it was. To prevent rare word repetitions from giving false estimates for a text's lexical difficulty, an arbitrary rule was adopted--no single rare word (i.e. a term not listed among the 10,000 most common word-types in English) may appear more than five times per thousand tokens. Every additional instance of that term gets an equal sign before it. Use a Comment header (&&000) to describe that this rare word rule was invoked, how often the = sign was used, and why it was necessary.

DIRTY WORDS. Contemporary speech and some texts contains frequent 'dirty' \words. Modern unabridged dictionaries contain some such words, but most are omitted. All such terms should be included, though most get the = sign.

WORDS REQUIRING SPECIAL TREATMENT. To reduce the most serious distortions of word use, my colleague Margaret Ahrens has compiled a list which occur principally in conversation. For each, the transcription convention is:

1. Words without equal signs:

   holidays (e.g. Christmas); bye; bye-bye (a hyphenated word) good-bye--a hyphenated word). Carroll, et al. report that the 'good-bye' form is 3 times more common than 'goodbye'. gonna becomes going to; and wanna becomes want to.

2. Words requiring the = sign:

   =Mom, =Dad, =honey, =sweetie, =darling (when referring only to a specific person), letters of the alphabet, standing alone. **Important exceptions: I** and **a**.

3. Special contractions: All of these terms are used in informal speech, but rarely in formal print. The preferred solution is to decompose them.

   =how'd =that'd =there'd =what'd =when's

=how's =that'll =there're =what'll =where'd
=how're =there'll =what're =who'd =there've
=what've and =why's

4. Terms generally omitted by dictionaries but whose informal use is common. Such terms get the = sign.

For example: =eh =ouch =wow =yuck
=yup =ick =ow =yep =yucky

C. Print's distorting effects on the Reference Lexicon.

Carroll, Richman and Davies' Word Frequency Book is based on word use in printed texts–which is normally written in formal style. Printed texts distort the relative frequency of certain words, and particularly under-represent words used in casual conversation (Hayes, 1988). Especially under-represented are household words, consequently, terms which are commonplace in a pre-schooler's experience (e.g. 'pajamas', 'diaper', 'potty' and 'bottle') or informal words like 'gonna' and 'where'd', appear as relatively rare words in the Carroll, et al. corpus--the Reference Lexicon common to all QLEX/QANALYSIS analyses.

The most under-represented words in print include '**I**' and '**good-bye**'. While pervasive in actual use, the Carroll, et al. list reports frequencies much lower than they appear in natural conversation, e.g. 'good-bye' in print occurs only 4 times per million tokens, far rarer than in the Cornell Corpus of natural conversations.

The most over-represented word in print is '**said**'. This term appears in print to help the reader keep track of the speaker. So common is the use of 'said' that it is the 43rd most common word on Carroll, et al's list--placing it amidst all the high frequency function words of English. 'Said' seldom appears in the natural conversations in the Cornell Corpus.

D. SEMI-AUTOMATIC TEXT EDITING. A program named **REPLACE2** serves to reduce, but not eliminate, editing. Furthermore, in fixing some editing problems, **REPLACE2 sometimes introduces errors of its own**. This program saves time in preparing a text for QLEX/QANALYSIS analysis, but all texts must be examined, *word for word*, to ensure compliance with the transcription rules--a lugubrious but necessary process.

FINDING ERRORS IN THE TEXT. Once a text has been edited, it is inevitable that editing errors will remain in your text. To find these errors, perform the QANALYSIS , then examine its output file for RESIDUALS --- this is a list of all words QANALYSIS could not find in its REFERENCE LEXICON of the 10,000 most common English types. These may be misspellings, archaics, and mistakes of editing will be apparent. Repair the text file and rerun QLEX /QANALYSIS.

MODIFYING THE REPLACE2 UTILITY. You can add/delete words in the REPLACE-LIST.TXT file. Use your word-processor to make your changes and then convert that file back into an ASCII file. Keep a backup of that file since your changes may not work as planned.

No set of transcription rules can serve all purposes. For this reason, one design goal was to keep QLEX /QANALYSIS texts as close to their original form as possible. It should be relatively simple to delete most LEX editing amendments to texts. For example, a 'macro' can be written to eliminate all instances of = signs before proper names, etc; or to insert periods for contractions (like Mrs and Dr); or to remove other of these QLEX/QANALYSIS editing amendments.

VII. INTERPRETING THE LEX STATISTIC.

LEX (C) (i.e., LEX for content or open class words only) is interpreted as measuring a text's accessibility, lexical difficulty, or comprehensibility. This interpretation is based on large body of research on the 'difficulty' and recognition of single words.

A. Word 'difficulty' and regularities in word choice. Several decades of research has shown that text difficulty is a concept of considerable complexity (Carr and Levy, 1990; Levy and Carr, 1990; Holland, 1981). This section describes what is probably the single most powerful component of a text's difficulty--its lexical difficulty, as represented by LEX (c) and by its related statistics.

Powerful statistical regularities in word choice were summarized and formed the basis for Herdan's (1960, 1966) lognormal model for word choice: when the types of a lexicon are ranked by their frequency in general use, and those ranks are expressed by their common log. He proposed that the resulting word choice distribution fits the lognormal statistical distribution (next to the normal distribution–one of nature's most common distributions). When represented as shown in Figure 1, actual word usage is essentially linear across the range from rank 1 through ~10,000. Put another way, the probability of a word's choice is based on the log of its frequency in general usage. Carroll (1971), using a specially-designed five million word corpus of texts sampled from school books from 17 domains for children age 9 to 15, found that English word choice fits Herdan's model well for most of its distribution, but the test was constrained by the fact that his five million word sample contained only 86,000+ of the estimated 609,000 (as estimated by Carroll, 1971) word types in the English lexicon.

Choice of word is conditional on its frequency in general usage. The rarer a term, the longer the period required to retrieve it from memory, and the longer the time required for its recognition when presented tachistoscopically (Just & Carpenter, 1987). Consequently, the rarer the word, the less likely it is to be used if the text is produced in realtime.

Consequently, spontaneous conversations should be (and are) less difficult than rehearsed text and most printed texts (typically written off-line).

Despite the 609,000 word-types in English, something on the order of just 100 act as function or closed-class. Those grammatical terms accounts for about half of all the words in natural texts.

B. The REFERENCE LEXICON: word choice in newspapers. The Newspaper mean LEX (c) is approximately 0.0 because its pattern of word choice is nearly linear in the text samples taken from major English language newspapers published in Africa, Asia, Australia, Europe, India and in North America, and as far back as 1665. LEX's interquartile range in these 63 newspaper samples is relatively narrow (LEX = -3.8 to +2.6).

Each feature section of a newspaper, however, has a characteristic pattern of word choice and difficulty. Samples from comics were written at LEX = -25; advice columns at LEX = -21; sports at -13.8, science and medicine at +3.7; and business news at + 4.7.

Several factors shape a text's pattern of word choice. These include the target audience's effect on the phrasing of a message, and on the choice of domain (e.g. technical vs mundane). Texts addressed to foreigners who are barely acquainted with English, to young children and to animals are generally heavily skewed toward common terms, showing that speakers and writers self-consciously manipulate LEX levels--though seldom with much accuracy--according to our research. In experiments designed to measure subjects success at hitting specific LEX targets, few could consistently hit near their intended level of text difficulty.

Several determinants of the pattern of word choice may occur simultaneously. Texts written by scientists for fellow specialists peers are heavily skewed toward rare terms--a combined audience and domain selection effect.

Other determinants of the pattern of word choice include realtime versus offline text production, and the level of personal stress/distraction while producing texts (as when witnesses testify under high or low stress levels).

C. Domains and LEX diversity. *Texts in widely diverse domains may have the same LEX score, and texts within the same domain may nonetheless have a wide range of LEX scores*. A more general model for word choice than Herdan's is the theoretical spectrum of LEX values found in the CORNELL CORPUS-2000. Every LEX score represents a unique pattern of word choice.

Speech clearly violates Herdan's lognormal pattern of word choice, but little is known about the specific mechanism(s) which allow the lognormal pattern to be overridden. Speaking to knowledgeable or ill-informed audiences induces complex adjustments of domain, topic and lexical choice (cf. Cornell PhD dissertation, Margaret G Ahrens). Witness testimony under direct and cross-examination also suggests that immediate personal stress and the necessity to produce answers in realtime are among the many constraints which shape LEX levels.

D. LEX's stability over the past 334 years. While the English lexicon is growing rapidly with new words or words take on new meanings and lose other meanings, still other words become archaic, change their spelling or sense. The patterns of word choice (particularly the use of the 10,000 most common types) has changed little in over the past 335 years. How do we know that? When this QLEX software (with its built-in modern lexicon) was applied to edited samples of English and colonial American newspapers published in the middle 1600's and 1700's (in London, Ipswich, Philadelphia, Richmond, and Charleston), their level (LEX = -3.5) were well within the range of contemporary newspapers. The 1791 Times (London) was written at LEX = -3.1; in 1850; at +3; and -1.7 in 1992. The New York Times was written at LEX = +2.0 in its first year (1852) and at -0.85 in 1987. Overall, newspaper levels appear to have risen at the rate less than 1 LEX per century, and perhaps, not at all.

E. LEX's Validation: experiments. There are multiple grounds for believing that LEX validly estimates a text's accessibility and difficulty. Decades of research on reading ability show that knowledge of the meanings and uses of uncommon words strongly predicts reading comprehension levels (Thorndike, 1973; Saarnio, et al, 1990). LEX's validity has also been confirmed in a series of experiments in which speakers address different audiences on different topics (Hayes and Ahrens). Subject's spontaneous speech and writing for these audiences on those topics contained the predicted directional changes in LEX levels but variances within the same experimental condition were wide, indicating a lack of precision in the subject's adjustments for audience or disagreements as to the extent of the necessary adjustments.

.

F. LEX Validation: comparisons within the 5000+ text Cornell Corpus. Another way to validate the interpretation that LEX measures the lexical difficulty of a text is to compare texts against one another. **Table 1** contains LEX and related lexical statistics on a spectrum of texts sampled from the 5000+ texts in the Cornell Corpus.(Hayes, 1988; Hayes and Ahrens, 1988).

TABLE 1 -- the spectrum of natural texts

G. LEX validation using US basal readers: Grades 1 to 8. Still another standard for validating LEX scores is the level at which American basal readers were pitched at different periods of American education. These were the texts used as the principal teaching materials between 1919 through 1991. These basal readers were used as primers and in first grade through middle-school. LEX levels were higher before 1947. Beginning in 1947, nearly all schoolbook publishers (except Xerox) simplified their basal reader series. This was the era of the controversial Dick, Jane and Spot basal readers produced by the Scott, Foresman company. Strong parent and teacher reaction to that level of simplification forced the publishers to restore their first through third grade basal readers to their pre-WW II and 1946 levels, but they left their LEX levels for grades 4 through 8 simplified -- lower than comparable grade basals used between World War I through WW II.. All the nearly 800 text samples are included in the Corpus.

In simplifying these texts, the publishers also shortened their sentence lengths (e.g. these declined from an average of 20 (near the level in newspapers) down to about 12 words per sentence for 5th grade basals published after 1946.

VIII. INTERPRETING QLEX /QANALYSIS'S OTHER STATISTICS.

If every analysis and printing option is chosen, QLEX /QANALYSIS output can be a bewildering mass of words and numbers--initially. This section identifies and interprets these sentence and lexical measures.

A. SENTENCES. At the top of each text's output file is the name of the file (e.g., IRS1040.out) and certain information about this analysis, including where in the text the analysis began, ended, and the ID code used in that analysis (e.g. 111). Beneath those statistics is a HISTOGRAM of the text's sentence lengths. A vast literature about sentence lengths serves as a framework for interpreting these numbers. In the Cornell Corpus--2000, the mean sentence length (**MLU-- in words**) of spontaneous conversations between adults and school-age children is 6.6 words, but lower (5.1) when adults talk with pre-school children. MLU in popular TV show conversations (usually scripted) are comparable in length to parent-older child speech (6.6).

In the Cornell Corpus-2000, the most widely read texts (e.g. newspapers have MLU's around 22 words, but magazines vary widely (top ten magazines have MLU's around 16. Research articles in Nature have MLU's exceeding 27 words, while books chosen and read by elementary school children have mean MLU's of 11.3. Books read to pre-school children average 12.2 words.

The first page of the output (e.g. IRS1040.OUT) shows that IRS text contained 56 full sentences, and 1028 tokens of which 41 were coded with the equal sign, leaving 987 non-name tokens for that full sentence analysis. The median length of sentence was 16 words, the MLU was 18.4 (with a large SD--12.7 words). There was one one-word sentence (2% of all sentences), and all sentences in this sample ended with a period.

In a representative sample of 101 texts from every major class of text in the Cornell Corpus, the linear correlation between text MLU (in words) and LEX was r =.763, but there are numerous anomalies where the association is much lower, leaving the interpretation open. The correlations within the 200+ sub-directories of the CORNELL CORPUS-2000 are normally much smaller.

It should be noted that sentence length measures based on spontaneous conversations of several participants are normally untrustworthy since such transcriptions require extraordinary care in punctuation, and excellent recordings--a situation seldom found in the general literature.

B. TOKENS AND THEIR FREQUENCIES 'Tokens' refer to the 'words' in a text, i.e.

terms separated by spaces. 'Word-types' refer to uniquely spelled tokens. **NOTE**–the use of word-types *ignores polysemy* (relatively high among common terms, but falls off rapidly among less common types--Hayes, 1988). Terms ranked beyond type ranked 1,000 seldom have alternative meanings or major differences in senses.

As a general rule, the more tokens in a text, the smaller the proportion of word-types in that text. In their five million word sample corpus, Carroll, Richman and Davies (1971) found only 86,741 word-types–all the rest were duplicates. For this reason, the ratio of word-types to tokens (TTR) in two texts cannot be compared--unless both texts have exactly the same number of tokens. All printed text samples and samples of transcripts from television shows in the Cornell Corpus are based on ~1,000-word samples (ten stratified sub-samples), making their TTRs roughly comparable. They would be exactly comparable only if one designates the exactly 1000 types be used in carrying out a LEX analysis.

The second page of the IRS1040.ASC output is the listing of all that text's word-types, alphabetically (on the left) and by their frequency of occurrence (on the right). This output file may be examined from your monitor. The full listing may not be of interest to most investigators. In the IRS case, note that the term 'deduction' is the highest ranked in frequency of use) of the words --- indicating something about the content in this IRS sample text. The eleven most commonly used word-types (function terms) in the IRS1040.ASC text convey virtually nothing about its content, yet those eleven accounted for 280 of the 1000 tokens. The listing of the uncommon and rare terms (>rank 10,000–the RESIDUALS) alone does a pretty good in conveying what the text is all about.

Among the many uses for these word lists, a child's use of specific grammatical or self/other reference terms during its first months and years of speech can be of use in testing theories of a child's syntactic, lexical development and self development.

C. TABLE OF RESIDUALS Page 8 of the IRS1040.out output file contains all the uncommon and rare words used of that text, i.e. those which QLEX /QANALYSIS did not find among the 10,000 most common word-types of English. A few are uncommon inflections of common words, but most are simply uncommon or rare names for objects, places, events and relations. Most typographical errors turn up in this list, making it essential that you examine this list and correct the text before running QLEX /QANALYSIS for the final time. To help determine whether the word is a genuine uncommon or rare word-type (not an inflection or derivation of a common word), the closest preceding and following words in QLEX /QANALYSIS's 10,000 word REFERENCE LEXICON are printed to the right of each residual term.

D. THE CUMULATIVE PROPORTION DISTRIBUTION. *This is the most important table in a QLEX /QANALYSIS analysis*. This Table contains the information necessary for calculating the text's accessibility/lexical difficulty. It describes the author or speaker's pattern of word choice--a concept closely related to the text's 'accessibility' or 'difficulty'.

The IRS1040.ASC output shows that lexical analysis was based on exactly 1,000 tokens; the number of types in those 1000 terms was 340, making the ratio of types to tokens (the TTR) .340); and its mean content word (after excluding instances of the first 75 most common, function terms) occurs with a frequency of 206.7 per million in the Carroll, et al Reference Lexicon. The words at the 10%, Q1, Q2, Q3 and 90% positions in the distribution of usage and their word ranks in the Reference Lexicon are also included as indicators of the text's dispersion in word choice.

The left hand column of this table represents the first and most common 10,000 word-types in English. In this IRS publication (IRS1040.ASC) 'the' alone accounted for 7.1 percent of all the text's 1,000 words. In Carroll, et al.'s Reference Corpus, 'the' alone accounts for 7.2% of all terms used. When combined with the second most common word in English ('of'), those two words account for 10.1 percent of all words in the IRS text. The first ten most common word-types account for 24.5 percent of all tokens used in the IRS text; the first 75 (function words) accounted for 46.5 percent of this sample text's tokens. The first 1,000 types on Carroll's list account for 65.8%; the first 5,000 account for 84.7%; and the first 10,000 most common English words account for 90.0% of the terms in the IRS text, leaving exactly 100 rare words from the original 1,000 as RESIDUAL words, i.e., the words QLEX /QANALYSIS could not find in its Reference Lexicon--these are the text's uncommon/rare words.

In spontaneous conversations recorded in their natural contexts within the Cornell Corpus, 'the' alone accounts for between 2 and 3 percent of all tokens--not the 7 percent in print. The most common 75 words (virtually all grammatical words) account for 45% of newspaper words, 44% in popular magazines, and 43% in adult books, but 51% of all words in adult-adult conversations--a bit less when adults talk with children.

The top 1,000 most common word-types account for 55% of all words in abstracts of articles in Science; 68% of all words in newspaper texts; 69% of all words in general magazines; but 84% of all words used in adult speech to children under age 2; 85% to children between age 2 and 6, and 85% for children of primary school age.

The top 5,000 word-types account for 74% of all words used in scientific abstracts, 84% of words in newspapers, 85% in popular magazines but 94% in adult-to-adult conversations, and nearly 96% in adult-with-child texts.

Again, the ranking of those 10,000 most commonly used words English comes from Carroll, et al's (1971) analysis. While their corpus is the most modern, largest, and most comprehensive of its kind, it would be desirable to have a REFERENCE LEXICON based on a far larger and more diverse data base. When one becomes available, it will be substituted for the Carroll, et al list. The pattern of word choice in Carroll's corpus is virtually identical to that in newspapers, general news magazines and encyclopedias. The Carroll corpus is representative of adult word use, in print, because the publishers of those schoolbooks use the first 10,000 most common word-types in virtually the same way as do authors writing for adults. The Carroll corpus is dissimilar in pronounced ways from word use in spontaneous conversation between family members, formal conversations and technical writing (cf. Hayes, J. of Memory and Language, 1988).

IX. RECOMMENDED MEASURES OF LEXICAL ACCESSIBILITY/DIFFICULTY.

Work on scientific measures for assessing a text's accessibility, difficulty and comprehensibility is a on-going process, just as the fundamental measures of science continue to be refined and developed. So far as is known, the most comprehensive, and best validated measure of a text's difficulty is **LEX (c) (open class word-types only)**--based on QLEX /QANALYSIS's analysis of edited natural texts.

Nearly as good a measure of a text's difficulty is the 'MeanU' statistic--i.e., the frequency per million tokens (in the Carroll, et al corpus) of that word which lies at the mean of all open class words in a sample text. The larger that word's meanU value, the more common the word choice in that text. *MeanU is the statistic of choice when texts samples are small* (e.g. <500 words). One advantage of MeanU is its interpretation--the higher the value, the simpler the text; the smaller the value, the less accessible the text.

LEX and MeanU are highly correlated (r = -.976) in the 101 texts which form the 1st Universal Sample of texts taken from the Cornell Corpus (cf. also Table 1). One reason LEX is considered a better measure of a text's lexical difficulty is that the entire cumulative proportion distribution can be closely approximated when one knows a text's LEX scores--something which cannot be done with a text's 'MeanU' statistic. LEX supplies far more information about the full pattern of word choice at all points from word rank 76 through 10,000 than does 'MeanU'.

A text's 'MedianU' statistic has proved to be a less valid measure than either LEX or MeanU because it bears a nonlinear relation to them both.

X. REFERENCES.

Carr, T. H. and B. A. Levy. (Eds.) Reading and Its Development 1990, Academic Press, San Diego.

Carroll, J. B. in Word Frequency Book, J. B. Carroll, P. Davies and B. Richman 1971, Houghton Mifflin, Boston).

Hayes, D. P. Polysemy in the lexicon, 1987 (unpublished ms.)

Hayes, D. P. Speaking and writing: distinct patterns of word choice. J. of Memory and Language, 1988, 27, 572-585.

Hayes, D. P. and M. G. Ahrens, Vocabulary simplification for children: a special case of 'motherese'. J. of Child Language, 1988a, 15, 395-410.

Hayes, D. P. The Growing Inaccessibility of Science. Technical

Report Series, Department of Sociology, Cornell University, Ithaca, NY. (1991); also in Nature, 1992, 356, 739-40.

Hayes, D. P. Wolfer, L. T. and M. F. Wolfe, Schoolbook simplification and it relation to the decline in SAT-verbal scores. Amer. Educational Research Journal, 1996, 33, 489-508.

Hayes, D. P. and M. Spivey. A general model for word choice: behavior under stress. Sociology Technical Report #4, Cornell University, 1999.

Hayes, D. P. LEX — A scientific measure of a text's accessibility. 5th International Conference on Social Science Methodology, Univ. of Cologne, Germany, Oct 3-6, 2000.

Herdan, G., Type-token mathematics 1960, Mouton, The Hague.

Herdan, G., The Advanced Theory of Language as Choice and Chance. 1966, Springer-Verlag, Berlin.

Holland, M. V., Technical Report No. 12, 1981, American Institutes for Research, Washington D. C.

Just, M. A. and P. A. Carpenter, The Psychology of Reading and Language Comprehension. 1967, Boston, Allyn and Bacon, Inc.

Levy, B. A. and T. H. Carr, in Carr, T. H. and B. A. Levy (Eds) Reading and Its Development 1990, Academic Press, New York.

Saarnio, D. A., Okla, E. R. and S. G. Paris, in Carr, T. H. and B. A. Levy, 1990, Reading and Its Development Academic Press, New York, 57-79.

Thorndike, R. L. Reading comprehension in 15 countries in International Studies in Education, III. 1973, Halstead Press, New York. 1-179.

Watson, J. D. and F. H. C. Crick, Nature, 1953, 117, 737-8.

Wells, G. Language Development in the pre-school years. 1985 Cambridge University Press, Cambridge.

XI. ADDRESSES:

Donald P. Hayes, 382 Uris Hall, Sociology, Cornell University, Ithaca, New York, 14853.
Phone # is 607-255-1425.
E-mail = dph1@cornell.edu.
Web address: www.soc.cornell.edu/hayes-lexical-analysis

:

1. To cite LEX, use this reference:

Donald P. Hayes, "The Growing Inaccessibility of Science", Nature, (1992) 356, pp 739-40.

2. To test your text for its LEX level:

Consult: www,soc,cornell.edu/hayes–lexical-analysis

3. To learn more about LEX measurement::

Consult the same www site for README-FIRST 2003

XII. CREDITS: software, and others' assistance

Version 1.0 of QLEX /QANALYSIS was written by Peter Bond in the spring of 1980 in BASIC for a DEC PDP-11-34 mini-computer. Version 2.0 through version 5.0 were written to run under DOS. Version 2.0 was written by Scott McAllister in 1982. Version 3.0 was written in PASCAL for IBM PCs by David Post in 1985. The principal version --the one most extensively tested and used–-was written in TurboPASCAL by Domingo Bernardo in 1988. Mignon Belongie wrote several utilities for the QLEX system in the early-1990s. The most recent version (6.0) named QANALYSIS, was written by Marty White in Perl. (Cf. Technical Details in the **README–FIRST file**). QANALYSIS remains in the testing stage., but can be used now. The numbers are correct (coincide with those from QLEX 5.1) but further development is needed on making the program more user-friendly, improve the ease of transport of output files to spreadsheet, improve the graphics (like Figures 1-3), etc. This program's development is driven by a research agend, not a commercial goal. Anyone can download this LEXGUIDE, the CORNELL CORPUS-2000, README-FIRST and the other files from my website. All these programmers were or became interested in text analysis, and each contributed ideas beyond simply writing the programs. Their contributions continue to be appreciated.

While many Cornell undergraduates assisted in this line of research--too many to name--Margaret G Ahrens for several years devoted her considerable talents as a graduate student as my collaborator and is the co-author of several papers. Aside from her skill as an experimenter and analyst in over a dozen validation experiments (some reported in her dissertation), she made countless intellectual contributions and proposed numerous alternative ways of editing and analyzing these texts which strengthened the analyses. Those efforts are most appreciated.

Credit is also due to the many colleagues, researchers, librarians, public officials and student subjects of experiments who generously allowed me to use their data (especially Gordon Wells at the Univ. of Toronto, Frank Whitehead of Sheffield University in the UK, and Brian MacWhinney and Catherine Snow at the Carnegie-Mellon consortium on child language development--CHILDES), who gave me access to their primary data, provided access to subjects or participated in the many experiments--to all--many thanks.

Filename: LEXGUIDE2003.WPD ON C:\LEXGUIDE2003 & ZIP WSMASTER#3
MAY 13, 2003 11:30