Program Abstracts/Algorithms

MRC Psycholinguistic Database search program

STEPHEN FEARNLEY

Oxford Brookes University, Oxford, England

This paper describes a computer search program based on the Medical Research Council Psycholinguistic Database of English words. The program allows words to be extracted from that database according to word length, number of syllables or phonemes, and various psycholinguistic criteria such as frequency of use, imageability, concreteness, meaning, and so forth. Thus it is possible to create, for example, lists of two-syllable words of high and low familiarity. It is also possible to examine properties of given sets of words created by the researcher. Lists of these words with or without their properties and with or without a statistical analysis of those properties may be produced. Particular spellings but not particular phonemes may be searched for.

It is frequently useful for behavioral scientists to use words as stimuli. When they do so, it is important for them to know the characteristics of the words—for example, the words' frequency of usage, imageability, and so on. Rudell (1993), for example, used the frequency counts generated by Kučera and Francis (1967) in examining the relationship between frequency of use and word difficulty.

Coltheart (1981) described a computerized database—the MRC Psycholinguistic Database—developed with the aid of a Medical Research Council grant and available on a UNIX platform. Wilson (1988) produced a more recent version of the database (Version 2) as a “computer-usable resource.” He made some corrections to erroneous entries and added spoken word frequencies (Brown, 1984). The database consists of around 100,000 words, together with information about their length, spelling, and syntactic category (Noun, Verb, Adjective, etc.)

Many of the words in the database also are accompanied by information about their linguistic properties, such as frequency of use in written English, familiarity, imageability, and so forth, as found in investigations by various researchers. This subset of 39,300 words that includes information on their linguistic properties is held in a second database (PSYCHOL.DCT). This makes the search faster when sets of words with known linguistic

properties are required. Not all the words have information about all linguistic properties. That subset is quite small—around 100. Many words appear two or three times: WORK, for example, is entered as a Noun and a Verb; CONTENT is entered as an Adjective, Noun, and Verb; and STAIN is entered as a Noun and a Verb. Out of the 39,300 words, 24,244 are unique, so there are some 15,000 duplications of one kind or another. Nevertheless, it is a considerable body of words.

Wilson (1988) supplied a search program for selecting words from the database according to these properties. To search the database with Wilson's search program, one adds the option to the command line for the program. He describes (p. 9) searching for Nouns with Kučera–Francis written frequencies between 100 and 500, with three to six phonemes, and with a Pavio meaningfulness value between 500 and 700, outputting the words only to a file TEST1 by using the following command:

```
dict +PS N -kffreqmin 100 -kffreqmax 500 -nphonmin 3  
-nphonmax 6 -meanpmin 500 -meanpmax 700 -W  
>TEST1
```

A comparable but enhanced facility has now been developed for IBM-PC-compatible computers in a much more accessible form than the command line options provided originally by Wilson, which students found difficult to use. The current facility is menu driven, with on-line help being provided about the characteristics being selected. This help includes the range and standard deviation of each characteristic and the original source papers for the research. Not only can lists of words with particular characteristics be created, but summary statistics of those words, or of words selected by other criteria, can be generated. These summaries include means, standard deviations, and range and quartiles of the characteristics being used. An example using color words will appear later.

The Program

The program MRCDATA (MRC Psycholinguistic Database Search Program) has been written in Turbo Pascal and should run on all 286, 386, and upwardly compatible computers but not on Macintoshes. It is an MS-DOS program but will run under Windows (3.x and '95) and in a networked environment. It is menu driven for nearly all options (see Exhibit 1 for the Main screen), which makes it easy for novices to use. Help is available for most of the facilities. This is provided in a separate ASCII file that may be altered easily for local needs. The help not only covers the running of the program, but also includes references to original material for the informa-

I wish to acknowledge the help of R. Lindsay from the Oxford Brookes Psychology Unit in the creation of the help file associated with this program. Correspondence concerning this article should be addressed to S. Fearnley, Psychology Unit, Oxford Brookes University, Gipsy Lane, Oxford OX3 0BP, England (e-mail: sfearnley@brookes.ac.uk).

```

M R C Psycholinguistic Database Search Program (s.fearnley)(3.x)

Current DataBase :PSYCHOL.DCT

MRC MAIN MENU

0 Enter Variables
1 Run the search
2 Load Search Variables from file
3 Save Search Variables to file
4 Choose DataBase for Search
5 Analyse Stats for Database
6 Make new DataBase
7 Word Search
8 Quit from program

Select Option Number or Letter and then press the <Enter> key

Use the <F1> key for help  Q + <Enter> or <Alt>X to exit Program

```

Exhibit 1. Main menu screen.

tion about the linguistic properties. This may be viewed or printed out as ASCII text to a file or printer. There are two databases available. The main one (MRC.DCT) contains around 100,000 words. In practice the smaller one (PSYCHOL.DCT) containing the 39,300 words with linguistic properties is more likely to be used.

It is possible to search for combinations of properties and output the words to a file as ASCII text or to the screen together with the values on each of the properties. The output may include headings, or may include only the selected words for use elsewhere. Search criteria may be saved for use on other occasions, and new databases consisting of subsets may be generated. Once a set of criteria

has been entered it is possible to: (1) create a list of the words that satisfies the criteria, to a file, printer or screen with or without their linguistic properties; (b) create a new database that consists of only the words that match the criteria; and (3) analyze the subset to obtain descriptive statistics of the word properties.

The first requirement is to Enter Variables for Search; here the criteria for searching are entered. It is possible to specify any joint combination of linguistic properties. An entry screen allows the lower and upper limits of the search within each linguistic category to be specified. (See below for the categories.) The limits of each linguistic category are shown and checked on input to en-

```

M R C Psycholinguistic Database Search Program (s.fearnley) (3.0)

Enter the values for your search - Press the <Esc> key to finish
Only enter the values you wish to search for - leave the others blank

Variable      Minimum  Maximum  Variable      Minimum  Maximum
Word Length           5          Brown Freq
Phoneme Number           Brown Freq
Number of Syllables           Familiarity
Kucera Francis Freq           Concreteness
Kucera Francis Cats           Imageability
Kucera Francis Samp           Colorado Meaning
Thorndike Lorge Freq           Paivio Meaning
                                Age of Acquisition

Number of letters in word (mean = 7 sd = 2.5 range 1 to 23)

Part of Speech  Noun  adJective  Verb  Adverb  Other :
Currently selected: Adverb Adjective Noun Verb

Dictionary used for part of speech [SOED ] [Jones]: SOED

If you wish to limit the search to parts of the alphabet, fill in the next:
Start Search at:          Finish Search at:

Use the <F1> key for topic help.  Move with cursor ↑ ↓ → ← and <Tab> keys

```

Exhibit 2. Example of entry screen.

sure that they are within range. See Exhibit 2 for an example of the main entry screen.

The following criteria may be used:

Word Length, Number of Phonemes, and Number of Syllables. Most words have these, though some phoneme counts are missing.

Brown Spoken Frequency. This value is available for 8,096 words (Brown, 1984).

Familiarity, Concreteness, and Imageability. Around 4,900 words are included, which have been set up as a combination of three scales (Gilhooly & Logie, 1980; Toglia & Battig, 1978; and Paivio, Yuille, & Madigan, 1968).

Kučera–Francis Written Frequencies, Category Frequencies, and Number of Samples. Around 17,790 words are included (Kučera, & Francis, W. N., 1967).

Thorndike Lorge Frequency. 14,830 words (Thorndike & Lorge, 1944).

Colorado Meaningfulness. 2,628 words (Toglia & Battig, 1978).

Paivio Meaning. 693 words (Paivio, Yuille, & Madigan, 1968).

Age of Acquisition. 1,904 words (Gilhooly & Logie, 1980; Carroll & White, 1973).

Spelling. The search may be limited to a range of words—for example, all words between (say) CAD and COD, or beginning with, including, or ending with a sequence such as CARE.

Part of Speech. The search may also be restricted to one or more parts of speech (e.g., Nouns or Verbs). The database includes two separate definitions for parts of speech. One is based on the *Shorter Oxford English Dictionary* (SOED) and one on Jones's *Pronouncing Dictionary* (1963). The SOED definitions, assembled by Dolby, Resnikoff, and MacMurray (1963), include 10 different categories: Noun, Adjective, Verb, Adverb, Pronoun, Preposition, Conjunction, Interjection, Past Participle, and Other. In the program, these are reduced to the first four and Other (which includes all the others, including "Other").

According to Wilson (1988) these definitions "can sometimes be unsatisfactory. For example the words

freeze and *harass* are nouns (as well as verbs) according to WTYPE." (WTYPE is the part of speech derived from the *Shorter Oxford English Dictionary*.) In order to avoid such esoteric usage, he suggests using the classifications based on Jones (1963). Jones uses only four categories: Noun, Verb, Adjective, and Other. Many fewer words are classified by this method. The current version allows either the SOED or Jones to be used to define parts of speech. The parts of speech are limited to the four categories specified by Jones, as well as Adverbs.

Performing the Search

Once the criteria have been entered, the next option is to Run the Search; this creates a list of words using the criteria entered by the earlier option. The words may have their linguistic properties included with the printout, or they may be printed on their own. After the words are output, it is possible to obtain an analysis of the words selected. This analysis may also be achieved without creating a list of the words. An analysis may be made of a whole database or of a subset as specified by the criteria entered by the first option. The statistics include minima and maxima, quartiles, and mean and standard deviation for all the linguistic properties of the words selected by the search process. Thus it is possible to know reasonably well what the properties of a stimulus set are. The criteria, words, and analysis may be saved to a file as ASCII text. The complete analysis of the PSYCHOL.DCT dictionary is shown in Exhibit 3. This summary of the characteristics of the dictionary shows that there are 39,300 words, but that only 24,244 of them are unique. This means that many occur two or more times (as, e.g., nouns and adjectives). Detail is given on the range (minima, maxima, quartiles, averages, and standard deviations) for the various properties held for the words in the database. For example, the maximum word length is 19 letters (mean length, 7), and the most syllables in a word, 8 (mean number, 2).

It is clear from Exhibit 3 that not all words have all properties listed. As well as to create lists of words, the criteria may be used to create new dictionaries which are subsets of the original. These may be searched in the same way as the original.

M R C Psycholinguistic Database Search Program (s.fearnley) (3.x)									
39300 words analysed from PSYCHOL.DCT (24244 unique entries)									
Name:	Min	Q1	Median	Q3	Max	Mean	N	S.D.	
Word Length	1	6	7	9	19	7	24244	2.5	
Phoneme Number	1	5	6	8	17	7	12381	2.5	
Number of Sylls	1	2	2	3	8	2	21130	1.1	
Kucera Francis Freq	1	1	4	13	28852	42	17791	467.7	
Kucera Francis Cats	1	1	2	6	15	4	17793	4.0	
Kucera Francis Samp	1	1	3	9	500	15	17793	43.8	
Thorndike Lorge Freq	1	8	21	82	236472	281	14831	1274.4	
Brown Freq	1	1	2	5	6833	23	8096	198.0	
Familiarity	43	423	498	547	657	472	4924	105.8	
Concreteness	158	331	436	556	670	438	4295	121.1	
Imageability	129	366	449	542	667	447	4829	109.5	
Colorado Meaning	127	372	421	464	617	410	2628	81.6	
Paivio Meaning	192	521	592	664	922	589	893	110.1	
Age of Acquisition	125	329	419	519	697	426	1904	124.2	
There were 16863 Nouns, 1421 Adverbs, 9564 Verbs, 7386 Adjectives,									
and 4066 Other words.									

Exhibit 3. Output of analysis of the PSYCHOL.DCT dictionary.

Current File: COLOURS.TXT 11-May-1995 page 1
 M R C Psycholinguistic Database Search Program (s.fearnley)(3.1)
 Searching for Adverb Adjective Noun Verb using SOED definitions

	/Brown Freq	/Concreteness	/Imageability	/Age of Acquisition	/Part Of Speech
BLACK	12	470	589		Adjectiv Noun Verb
BLUE	7	459	569		Adjectiv Noun Verb
BROWN	2				Adjectiv Noun Verb
CRIMSON					Adjectiv Noun Verb
DUN					Adjectiv Noun Verb
EMERALD		613	602		Noun
GOLD	5	576	594		Adjectiv Noun
GREY		471	541		Verb Adjectiv Noun
GREEN	9	460	609	225	Adjectiv Noun Verb
MAROON		486	503		Adjectiv Noun Verb
MAUVE					Adjectiv Noun
ORANGE	3	601	626	203	Adjectiv Noun
PINK	3				Adjectiv Noun Verb
PURPLE					Adjectiv Noun Verb
RED	9	501	585		Verb Adjectiv Noun
SCARLET		519	587		Noun Verb Adjectiv
WHITE	19	472	566		Adjectiv Noun Verb
YELLOW	2	537	598		Noun Verb Adjectiv

There were 49 words found (18 unique)
 39181 words analysed from PSYCHOL.DCT using COLOURS.TXT for selection
 49 words matched the selection criteria (18 unique entries)

Name:	Min	Q1	Median	Q3	Max	Mean	N	S.D.
Word Length	3	4	5	6	7	5	18	1.3
Phoneme Number	3	3	3	4	7	4	8	1.4
Number of Sylls	1	1	1	2	3	1	18	0.6
Kucera Francis Freq	1	3	48	116	365	83	18	100.0
Kucera Francis Cats	1	3	12	13	15	9	18	5.1
Kucera Francis Samp	1	3	31	69	156	44	18	45.1
Thorndike Lorge Freq	3	61	364	989	2663	619	18	661.2
Brown Freq	2	2	6	10	19	7	10	5.4
Familiarity	428	492	555	590	607	546	12	58.8
Concreteness	459	470	493	566	613	514	12	55.8
Imageability	503	566	588	601	626	581	12	32.9
Colorado Meaning	394	433	446	464	502	452	12	28.9
Paivio Meaning	580	580	580	668	668	628	3	44.5
Age of Acquisition	203	203	217	225	225	214	2	

There were 18 Nouns, 0 Adverbs, 14 Verbs, 17 Adjectives,
 and 0 Other words.

All entries written to COLOURS.PRN

Exhibit 4. Example of output from a search of color words.

Finally, the program may be used to determine the properties of stimuli used in a research exercise. The words are simply collected into a file (e.g., COLOURS.TXT) and the Word option selected. The words will be arranged into alphabetical order and printed out together with their properties (if found). This list may be followed by an analysis of the words as described above.

Example of Word Search

Exhibit 4 illustrates the output from a search through PSYCHOL.DCT for a list of color words. The words were prepared as an ASCII text file (COLOURS.TXT), and the Word Search option was selected. Detail was requested on Brown Frequency, Concreteness, Imageability, Age of Acquisition, and Part of Speech. The output was sent to a file (COLOURS.PRN) and this output is reproduced in Exhibit 4. This shows the value for each word *where there is an entry in the dictionary*. The phrase "There were 49 words found (18 unique)" indicates that there were 49 entries of the words, but that some occurred as (e.g.) Nouns *and* Adjectives. There were 18 different words in the list. Finally there is a summary show-

ing, for all properties, the minimum and maximum values found, the quartiles (Q1, Median, and Q3), and the mean, standard deviation, and number of values for each characteristic. It may be seen that Brown spoken frequencies for BROWN (2) and YELLOW (2) are low (equal to the first quartile of 2), whereas BLACK (12) and WHITE (19) are high (above the third quartile of 10). SCARLET has no Brown frequency entry.

Hardware Specifications

The program will run on any IBM-compatible computer (286 and above) with MS-DOS 3+ and around 500K of memory. The program itself is around 120K, and the PSYCHOL.DCT dictionary, around 3 Mb. The full MRC dictionary is around 11 Mb. The program may be run under Windows, though it is not a Windows program.

Availability

The program is available for a fee of £30 sterling from the author, S. Fearnley. It can be provided with the PSYCHOL.DCT dictionary if required. The main MRC dictionary is available by anonymous FTP from the Oxford

Archives at ota.ox.ac.uk, where it is held as a public domain document.

REFERENCES

- BROWN, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instruments, & Computers*, **16**, 502-532.
- CARROLL, J. B., & WHITE M. N. (1973). Age-of-acquisition norms for 220 picturable nouns. *Journal of Verbal Learning & Verbal Behavior*, **12**, 563-576.
- COLTHEART, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.
- DOLBY, J. L., RESNIKOFF, H. L., & MACMURRAY, F. L. (1963). A tape dictionary for linguistic experiments. In *Proceedings of the American Federation of Information Processing Societies: Fall Joint Computer Conference* (pp. 419-423). Baltimore, MD: Spartan Books.
- GILHOOLY, K. J., & LOGIE, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, **12**, 395-427.
- JONES, D. (1963) *Everyman's English pronouncing dictionary* (12th ed.). London: Dent.
- KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence RI: Brown University Press.
- PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 words. *Journal of Experimental Psychology Monographs*, **76**(3, Pt. 2).
- RUDELL, A. P. (1993). Frequency of word usage and perceived word difficulty: Ratings of Kučera and Francis words. *Behavior Research Methods, Instruments, & Computers*, **25**, 455-463.
- THORNDIKE, E. L., & LORGE, I. (1944). *The teacher's word book of 30,000 words*. New York: Columbia University, Teachers College Press.
- TOGLIA, M. P., & BATTIG, W. R. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.
- WILSON, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, **20**, 6-10.

(Manuscript received June 14, 1995;
revision accepted for publication May 21, 1996.)