Behavior Research Methods, Instruments, & Computers
1988, 20 (1), 6-10

MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00

MICHAEL WILSON
Rutherford Appleton Laboratory, Oxfordshire, England

The MRC machine-usable dictionary contains 150,837 words and up 26 linguistic and psycholinguistic attributes for each. The attributes are from sources that are publicly available but are difficult to obtain and structure into a single dictionary. Three utility programs are described that permit the selection of words defined by a set of specified attribute values and the selection of attribute values for a set of specified words. These programs permit the construction of word sets for psycholinguistic experiments that control for the attributes specified in the dictionary. The dictionary may also be of use to researchers in artificial intelligence and computer science who require psychological and linguistic descriptions of words.

Those wishing to construct word sets as stimuli for psycholinguistic experiments must take into account a large number of characteristics of the words (see Cutler, 1981; Whaley, 1978). The Medical Research Council (MRC) Psycholinguistic Database, version 1, was provided as an on-line service (see Coltheart, 1981a; this paper describes an update) to provide control in selecting word sets. The service made available three files and several access programs. The first file was a dictionary of words; the second and third files were sets of word-association norms from the Edinburgh thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973). The service has been discontinued.

The second version of the MRC Psycholinguistic Database is being provided as a computer-usable resource rather than as a service. An updated version of the dictionary file from the database is being provided for public research purposes along with some programs that can be used either to access the dictionary or as examples on which to model programs that match users' specific needs. The changes from the first version of the database include the addition of 52,299 new entries, the inclusion of data on written word capitalization and spoken word frequency, and an expansion of the categorizations used for several properties. Corrections have also been made to erroneous entries discovered during the use of version 1. The entries for reversed spelling and reversed phonetic transcription that were included in version 1 have been removed, since their role can also be filled by the entries for forward spelling and phonetic transcription.

The MRC Psycholinguistic Database dictionary differs from other machine-usable dictionaries in that it includes not only syntactic information, but also psychological data for the entries (see Amsler, 1984, for a review of other machine-readable dictionaries). It also differs from most conventional dictionaries in that it does not currently attempt to provide any semantic information. It is designed to be of use to psycholinguists in selecting stimulus materials for testing, to researchers in artificial intelligence as a source of information required for natural language processing and cognitive simulation, and to computer scientists who wish to use the word lists and syntactic information in the design of text processors.

The file contains 150,837 words and provides information about 26 different linguistic properties, although information about every property is not available for every word: nobody, for example, has yet collected imagery ratings on such a large set of words, and thus only 9,240 of the words possess imagery ratings. The dictionary file does not contain any information that is original to it; it was assembled by merging a number of smaller databases of limited availability.

The dictionary file currently occupies 11 MB as a sequential UNIX[1] file. Each entry occupies one line of the dictionary. The composition of the dictionary file is summarized in Table 1, which specifies the linguistic properties described in an entry. The first column indicates the numbered name of the data field used elsewhere in programs and documentation. The second column specifies the identity of the linguistic property, and the third column indicates the number of words in the database for which information about a particular linguistic property is available. The first 14 properties are stored in the file as numerical values. For these properties, the occurrence count refers to the number of nonzero entries. The first 3 proper-

I am grateful to Professor Coltheart, Philip Quinlan, and Roger Mitton for making available their version of the MRC database (produced under grant SPG 977/912 from the Medical Research Council) and to those who constructed each of the data sets included in the present version. Copies of the dictionary, full documentation, and the utility programs are available for research purposes on magnetic tape in a variety of formats (any of: 800, 1600, 6250 BPI densities; ISO/ASCII, EBCDIC, BCD character codes; labeled for ANSI, ICL VME, none; formatted as fixed, variable, formatted). A modest charge will be made to cover mailing and the cost of the tape. The database can be obtained from Oxford Text Archive, Oxford University Computing Service, 13 Banbury Rd., Oxford OX2 6NN, England.

The author's mailing address is Informatics Division, Science and Engineering Research Council, Rutherford Appleton Laboratory, Chilton, Didcot, Oxon OX11 0QX, U.K.

Table 1
Properties Described in the Dictionary File

	Name	Property	Occurrences
1	NLET	Number of letters in the word	150,837
2	NPHON	Number of phonemes in the word	38,438
3	NSYL	Number of syllables in the word	89,402
4	K-F-FREQ	Kučera and Francis written frequency	29,778
5	K-F-NCATS	Kučera and Francis number of categories	29,778
6	K-F-NSAMP	Kučera and Francis number of samples	29,778
7	T-L-FREQ	Thorndike-Lorge frequency	25,308
8	BROWN-FREQ	Brown verbal frequency	14,529
9	FAM	Familiarity	9,392
10	CONC	Concreteness	8,228
11	IMAG	Imagery	9,240
12	MEANC	Mean Colorado meaningfulness	5,450
13	MEANP	Mean Paivio meaningfulness	1,504
14	AOA	Age of acquisition	3,503
15	TQ2	Type	44,976
16	WTYPE	Part of speech	150,769
17	PDWTYPE	PD part of speech	38,390
18	ALPHSYL	Alphasyllable	15,938
19	STATUS	Status	89,550
20	VAR	Variant phoneme	1,445
21	CAP	Written capitalised	4,585
22	IRREG	Irregular plural	23,111
23	WORD	The actual word	150,837
24	PHON	Phonetic transcription	38,420
25	DPHON	Edited phonetic transcription	136,982
26	STRESS	Stress pattern	38,390

ties refer to counts based on the entries in the WORD and PHON fields. The other properties require some explanation.

K-F-FREQ, K-F-NCATS, K-F-NSAMP

K-F-FREQ refers to a word's written frequency of occurrence as given in the norms of Kučera and Francis (1967). K-F-NCATS gives the number of categories of text in which the word was found, and K-F-NSAMP gives the number of samples found when constructing the norms. Kučera and Francis (1967) should be consulted if these are to be used.

T-L-FREQ

This is the written frequency of occurrence as given in the L count of Thorndike and Lorge (1944). If you plan to use this frequency count, you are advised to read details about it in Thorndike and Lorge's book. For example, the frequency value of a singular word that has a regular plural *includes* the frequency of the plural form, and this is true for other kinds of derivations too.

BROWN-FREQ

This stands for the frequency count of spoken English derived by Brown (1984) from the London-Lund Corpus of English Conversation (Svartvik & Quirk, 1980). There are 14,529 entries for 8,985 different strings in the WORD field.

FAM CONC and IMAG

These stand for subjective ratings of printed words for familiarity, concreteness, and imageability, respectively, and were derived from merging three sets of norms: Paivio (unpublished; these are an expansion of the norms of Paivio, Yuille, & Madigan, 1968), Toglia and Battig (1978), and Gilhooly and Logie (1980). These are expressed as integer values between 100 and 700 (in the original norms, the equivalent range was 1.00 to 7.00). The three sets of norms correlated highly and were merged by adjusting both the means and standard deviations before averaging. The exact method used is described in detail in Appendix 2 of Coltheart (1981b).

MEANC and MEANP

These are the meaningfulness ratings from the Colorado norms of Toglia and Battig (1978) and the norms of Paivio (unpublished) multiplied by 100 to produce a range from 100 to 700. The two sets of meaningfulness ratings were not merged because their correlations were low (only +.529) and the mean values for a set of words common to the two sets of norms were very low (see Toglia & Battig, 1978, Table 2). These differences are due to differences in the instructions to subjects. Thus, the two sets of meaningfulness ratings are not comparable, and so were kept separate.

AOA

This is age of acquisition from the norms of Gilhooly and Logie (1980), multiplied by 100 to produce a range from 100 to 700.

TQ2

When TQ2 has the value Q (40,810 occurrences), this word is a derivational variant of another word in the dic-

tionary file (e.g., *baptist* from *baptism*). When TQ2 has the value 2 (4,166 occurrences), the word ends in the letter R and this R is not pronounced except when the next word begins with a vowel. When an entry should have both values 2 and Q for this attribute, Q is given in this field, and both values are given in DPHON.

WTYPE and PDWTYPE

WTYPE is the syntactic category, as represented in the database assembled by Dolby, Resnikoff, and MacMurray (1963), that was created by taking all the left-justified boldfaced words from the *Shorter Oxford English Dictionary* (Onions, 1933) together with the parts of speech given by that dictionary. In addition, words were taken from the Cornell University tape of 20,000 commonly used words, and the parts of speech for all these words were found in the third edition of *Webster's New International Dictionary*. There are 10 different syntactic categories, coded as shown in Table 2. For determining syntactic category, WTYPE can sometimes be unsatisfactory. For example, the words *freeze* and *harass* are nouns (as well as verbs) according to WTYPE; indeed, when these are looked up in the *Shorter Oxford English Dictionary* or *Webster's*, they are described as nouns. To avoid such esoteric usages, PDWTYPE may be useful. It refers to the syntactic categories given in Jones's (1963) *Everyman's English Pronouncing Dictionary*, and very unusual uses of words are not considered. However, PDWTYPE uses only 4 categories, not 10: these 4 are noun (N, 22,061 occurrences), verb (V, 6,333 occurrences), adjective (J, 8,817 occurrences), and other (O, 1,179 occurrences).

ALPHSYL

If ALPHSYL = A, then the word is an abbreviation (130 occurrences); if S, the word is a suffix (282 occurrences); if P, a prefix (1,374 occurrences); if H, the word is hyphenated (13,716 occurrences); if T, a multiword phrasal unit (436 occurrences). For all of these categories, NSYL = 0. For all other words, ALPHSYL is blank.

STATUS

The 15 possible categories of STATUS are listed in Table 3; these are as given in the Dolby database (Dolby et al., 1963), derived from the *Shorter Oxford English Dictionary*, and the perusual of Table 3 should make the meanings of these categories sufficiently clear.

Table 2
Syntactic Category Codes for WTYPE

Syntactic Category	Code	Occurrences
Noun	N	77,355
Adjective	J	25,547
Verb	V	30,725
Adverb	A	4,243
Preposition	R	230
Conjunction	C	108
Pronoun	U	134
Interjection	I	352
Past Participle	P	5,939
Other	O	6,136

Table 3
The Possible Values of STATUS

Status of Word	Code	Occurrences
Dialect	D	2,780
Alien	F	6,003
Archaic	A	959
Colloquial	Q	405
Capital	C	2
Erroneous	N	0
Nonsense	E	62
Nonce	W	33
Obsolete	O	10,549
Poetical	P	183
Rare	R	2,756
Rhetorical	H	22
Specialized	$	7,731
Standard	S	58,065
Substandard	Z	0

VAR

This refers to words that have the same spelling but different pronunciation and syntactic classes. When the pronunciations differ only in respect of stress (e.g., *object*, *insult*), VAR = O (212 occurrences). When the pronunciations differ phonemically (e.g., *moderate*, *abuse*), VAR = B (1,233 occurrences). Either or both of these groups of words may be classed as homographs by some definitions.

CAP

If CAP = C, then the word is normally written with an initial capital letter. This can be used as an indicator of proper nouns, such as the names of people, towns, states, and countries.

IRREG

This refers to the plurality of words. Where IRREG = Z, the word is plural (17,441 occurrences), and this can be used in conjunction with TQ2 to select irregular forms; where IRREG = Y, the word is a singular form (1,024 occurrences); where IRREG = B, the word is both the singular and the plural form (151 occurrences); where IRREG = N, the word has no plural form (4,407 occurrences); where IRREG = P, the word is plural but acts singular (88 occurrences).

WORD

The dictionary is ordered by the ASCII sequence of these strings. Although there are 150,837 entries in the dictionary, there are only 115,331 different strings, since strings can hold different parts of speech, each of which has a separate entry. The entries in the WORD field were taken from Kiss et al.'s (1973) associative thesaurus and the database of Dolby et al. (1963) based on the *Shorter Oxford English Dictionary* and the Cornell University tape

of 20,000 commonly used words, with the addition of 2,500 proper names from Mitton's (1986) machine-usable version of the *Oxford Advanced Learner's Dictionary*, which were added to the version of the dictionary published by the Oxford University Press (Hornby, 1974).

PHON and DPHON

The 12th edition of *Everyman's English Pronouncing Dictionary* (Jones, 1963) was transferred to magnetic tape by Guierre (1966). This was used as the basis of the phonetic transcriptions in the PHON field. These include a marker for the syllable boundaries, which is not included in the edited phonetic transcription of the DPHON field. The DPHON entry also includes the entry for the TQ2 value. The phonetic symbols used in this database were adjusted following suggestions from Mitton (1986) by Quinlan (1986) to conform the U.K. ALVEY standard for machine-readable phonetic transcription (see Wells, 1986).

STRESS

The STRESS field includes numerical values representing the stress of each syllable in the PHON field.

UTILITY PROGRAMS

There are three utility programs available to access and modify the dictionary. These are written in the C language for the UNIX operating system, but should be usable on any system with a C compiler.

DICT

This program acts as a filter on the MRC database dictionary file. A subset of words can be selected from the total set of 150,837 words that fall within ranges specified by the user for the properties of words classified in the database. The filter can output either the entire record for a word, or any set of the properties. A flag may be used on the command line to specify the desired range or characteristics of each property in the database. If a property is not of interest, then no flag need be used and the value of that property for entries will be ignored. When constructing sets of experimental stimuli, the conditions on each relevant property can be specified to deliver the words that meet them. For example, to select nouns (+PS N) that are of standard usage according to the *Shorter Oxford English Dictionary* (+STATUS S), with Kučera and Francis frequencies between 100 and 500, with between 3 and 6 phonemes, and a meaningfulness value on the Paivio measure of between 500 and 700, and then to output only the words (−W) to a file called test1.materials, the command to DICT would be:

```
dict +PS N +STATUS S −kffreqmin 100
−kffreqmax 500 −nphonmin 3 −nphonmax 6
−meanpmin 500 −meanpmax 700 −W >
test1.materials
```

GETENTRY

This tool complements the DICT filter in that it selects the linguistic properties from the dictionary for a given set of words, rather than the words that fall within values for specified properties.

PSYCHDICT

The complete dictionary is large at 11 MB. This program reduces it to contain only those entries for which psychological measures are available. This program can produce a smaller dictionary that will be sufficient for the construction of psycholinguistic stimuli, but may not serve other purposes that the whole dictionary could. The smaller dictionary is a 3-MB sequential UNIX file and contains entries for 39,300 words.

REFERENCES

AMSLER, R. A. (1984). Machine-readable dictionaries. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST)*, **19**, 161-209.

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.

COLTHEART, M. (1981a). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, **33A**, 497-505.

COLTHEART, M. (1981b). *MRC Psycholinguistic Database user manual: Version 1*. [Available from Professor Coltheart, Birkbeck College, London WC1, U.K.]

CUTLER, A. (1981). Making up materials is a confounded nuisance. *Cognition*, **10**, 65-70.

DOLBY, J. L., RESNIKOFF, H. L., & MACMURRAY, F. L. (1963). A tape dictionary for linguistic experiments. *Proceedings of the American Federation of Information Processing Societies: Fall Joint Computer Conference*, **24**, 419-23. 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.

GUIERRE, L. (1966). Un codage des mots anglais en vue de l'analyse automatique de leur structure phonetique. *Etudes de linguistique appliquée*, **4**, 48-64.

HORNBY, A. S. (1974). *Oxford Advanced Learner's Dictionary of Current English*. Oxford, England: Oxford University Press.

JONES, D. (1963). *Everyman's English pronouncing dictionary* (12th ed.). London, England: Dent.

KISS, G. R., ARMSTRONG, C., MILROY, R., & PIPER, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitkin, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies*. Edinburgh: Edinburgh University Press.

KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

MITTON, R. (1986). A partial dictionary of English in computer usable form. *Literary & Linguistic Computing*, **1**, 214-215.

ONIONS, C.T. (1933). *Shorter Oxford English Dictionary*. London: England: Oxford University Press.

PAIVIO, A., YUILLE, J. C., & MADIGAN, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplement*, **76**(3, Pt. 2).

QUINLAN, P. (1986). *Description of machine-readable dictionary files* (Report). London: Birkbeck College, Department of Psychology.

SVARTVIK, J., & QUIRK, R. (1980). *A corpus of English conversation*. Lund, Sweden: Gleerup.

THORNDIKE, E. L., & LORGE, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

TOGLIA, M. P., & BATTIG, W.F. (1978). *Handbook of semantic word norms*. New York: Erlbaum.

Webster's New International Dictionary (3rd ed.) (1961). Springfield, MA: Merriam-Webster.

WELLS, J. W. (1986). A standardised machine-readable phonetic notation. In *Proceedings of the IEE Conference on Speech Input/Output: Techniques and Applications*. London: IEE.

WHALEY, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning & Verbal Behavior*, **17**, 143-154.

NOTE

1. UNIX is a registered trademark of AT&T in the USA and other countries.

(Manuscript received July 14, 1987;
revision accepted for publication October 9, 1987.)