MRC Psycholinguistic Database:
Machine Usable Dictionary,
Version 2.00.

Michael Wilson
Rutherford Appleton Laboratory, Oxfordshire, England.

Running Head: MRC Machine Usable Dictionary.

All correspondence should be addressed to:

Michael Wilson
Informatics Division
Science and Engineering Research Council
Rutherford Appleton Laboratory
Chilton,
Didcot,
Oxon, OX11 0QX
U.K.

FOOTNOTE

I am grateful to Professor Coltheart, Philip Quinlan and Roger Mitton for making available their version of the MRC database (produced under Grant Number 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; labelled 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 Road, Oxford OX2 6NN England.

ABSTRACT

The MRC machine usable dictionary contains 150837 words with up to 26 linguistic and psycholinguistic attributes for each. The attributes are from sources that are publicly available, but difficult to obtain and structure into a single dictionary. Three utility programs are described which permit the selection of words defined by a set of specified attribute values, and the attribute values for a set of specified words. These programs permit the construction of word sets for psycholinguistic experiments which 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, of which this paper is an update) to provide such control in selecting word sets. The service drew on 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 and Piper, 1973). The service has now 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 which can be used either to access the dictionary, or as examples on which to model programs which match users' specific needs. The changes from the first version of the database include the addition of 52299 new entries, the inclusion of data on written word capitalisation and spoken word frequency, and an expansion of the categorisations 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 which 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; for use by researchers in artificial intelligence as a source of information required for natural language processing and cognitive simulation, and for use by computer scientists who wish to use the word lists and syntactic information in the design of text processors.

The file contains 150837 words and provides information about 26 different linguistic properties, although it is not the case that information about every property is available for every one of the 150837 words: nobody, for example, has yet collected imagery ratings on such a large set of words, and thus only 9240 of the words possess an imagery rating. The dictionary file does not contain any information which is original to it, but was assembled by merging a number of smaller databases of limited availability.

The dictionary file currently occupies 11 Mbyte as a sequential UNIX[1] file. Each entry occupies one line of the dictionary. The composition of the dictionary file is summarised 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 fourteen properties are stored in the file as numerical values. For these properties, the occurrence count refers to the number of non zero entries.

Table 1 about here

The first three properties 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 The first of these refers to a word's written frequency of occurrence as given in the norms of Kucera 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. Kucera 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 the Thorndike -Lorge book. For example, the frequency value of a singular word which 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 from the London-Lund Corpus of English Conversation (Svartvik and Quirk, 1980) by Brown (1984). There are 14529 entries for 8985 different strings in the WORD field.

FAM CONC and IMAG These stand for subjective ratings of printed words for 'familiarity',

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

'concreteness' and 'imageability' respectively. These were derived from merging three sets of norms: Paivio (unpublished, these are an expansion of the norms of Paivio, Yuille and 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 and 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 (40810 occurrences), this word is a derivational variant of another word in the dictionary file (e.g. baptist, from baptism). When TQ2 has the value 2 (4166 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) which was created by taking all the left justified bold faced 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 found in the third edition of Webster's New International Dictionary. There are ten different syntactic categories, coded as shown in Table 2. When you are interested in syntactic category, WTYPE can sometimes be unsatisfactory. For example, the words FREEZE and HARASS are Nouns according to WTYPE (as well as verbs); and indeed when these are looked up in the Shorter Oxford English Dictionary or Webster's, they are described as nouns. If you want to avoid such esoteric usages, PDWTYPE may be useful. It refers to the syntactic categories given in

Jones' Pronouncing Dictionary (Jones, 1963), and very unusual uses of words are not considered. However PDWTYPE uses only four categories, not ten: these four are noun (N, 22061 occurrences), verb (V, 6333 occurrences), adjective (J, 8817 occurrences) and other (O, 1179 occurrences).

Table 2 about here

ALPHSYL If this = A, then the word is an abbreviation (130 occurrences); if S, the word is a suffix (282 occurrences); if P, a prefix (1374 occurrences); if H, the word is hyphenated (13716 occurrences); if T, a multi-word 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 perusal of Table 3 should make the meanings of these categories sufficiently clear.

Table 3 about here

VAR This refers to words which 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 (1233 occurrences). Either or both of these groups of words may be classed as homographs by some definitions.

CAP If this = 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 (17441 occurrences), this can be used in conjunction with TQ2 to select irregular forms; where IRREG = Y, the word is a singular form (1024 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 (4407 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 150837 entries in the dictionary, there are only 115331 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 the Edinburgh Associative Thesaurus (Kiss et al, 1973) 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 2500 proper names from the Machine Usable Version of the Oxford Advanced Learner's Dictionary (Mitton, 1986), which were added to the version of the dictionary published by the Oxford University Press (Hornby, 1974).

PHON and DPHON The 12th edition of the Jones' 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 to 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 150837 words which 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 which meet them. For example, to select nouns (+PS N) which are of standard usage according to the Shorter Oxford Dictionary (+STATUS S), with Kucera and Francis frequencies between 100 and 500, with between 3 and 6 phonemes and a meaningfulness 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 which fall within values for specified properties.

PSYCHDICT The complete dictionary is large at 11MByte. This program reduces it to contain only those entries for which psychological measures are available. This program can produce a smaller dictionary which 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 3MByte sequential UNIX file and contains entries for 39300 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. American Society for Information Science (ASIS); Knowledge Industry Publications, Inc.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavioural Research Methods Instrumentation and 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. and 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. and Logie, R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and 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 appliquee,* **4,** 48-64.

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

Kiss, G.R., Armstrong, C., Milroy, R. and Piper, J (1973). An associative thesaurus of English and its computer analysis. In Aitkin, A.J., Bailey, R.W., and Hamilton-Smith, N. (Eds.), *The computer and Literary Studies*. Edinburgh: University Press.

Kucera, H. and Francis, W.N. (1967). *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Jones, D. (1963). *Everyman's English Pronouncing Dictionary, 12th edition.* London, England: Dent.

Mitton, R. (1986). A Partial Dictionary of English in Computer Usable Form. *Literary and Linguistic Computing,* **1,** 214-215.

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

Paivio, A., Yuille, J.C. and Madigan, S.A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplement,* **76** (3, part 2).

Quinlan, P. (1986). *Description of machine-readable dictionary files.* Report. Dept. of Psychology, Birkbeck College, London.

Svartik, J. and Quirk, R. (1980). *A Corpus of English Conversation.* Lund: Gleerup.

Thorndike, E.L. and Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words.* New York: Teachers College, Columbia University.

Toglia, M.P. and Battig, W.F. (1978). *Handbook of Semantic Word Norms.* New York: Erlbaum.

Waley, C.P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behaviour,* **17,** 143-154.

Wells, J.W. (1986). A standardised machine-readable phonetic notation. In *Proceedings of the IEE conference on speech input/output: techniques and applications.* London, Easter 1986.

	name	property	occurrences
1	NLET	Number of letters in the word	150837
2	NPHON	Number of phonemes in the word	38438
3	NSYL	Number of syllables in the word	89402
4	K-F-FREQ	Kucera and Francis written frequency	29778
5	K-F-NCATS	Kucera and Francis number of categories	29778
6	K-F-NSAMP	Kucera and Francis number of samples	29778
7	T-L-FREQ	Thorndike-Lorge frequency	25308
8	BROWN-FREQ	Brown verbal frequency	14529
9	FAM	Familiarity	9392
10	CONC	Concreteness	8228
11	IMAG	Imagery	9240
12	MEANC	Mean Colorado Meaningfulness	5450
13	MEANP	Mean Paivio Meaningfulness	1504
14	AOA	Age of Acquisition	3503
15	TQ2	Type	44976
16	WTYPE	Part of Speech	150769
17	PDWTYPE	PD Part of Speech	38390
18	ALPHSYL	Alphasyllable	15938
19	STATUS	Status	89550
20	VAR	Variant Phoneme	1445
21	CAP	Written Capitalised	4585
22	IRREG	Irregular Plural	23111
23	WORD	the actual word	150837
24	PHON	Phonetic Transcription	38420
25	DPHON	Edited Phonetic Transcription	136982
26	STRESS	Stress Pattern	38390

Table 1. Properties described in the dictionary file.

Syntactic Category	Code	occurrences
Noun	N	77355
Adjective	J	25547
Verb	V	30725
Adverb	A	4243
Preposition	R	230
Conjunction	C	108
Pronoun	U	134
Interjection	I	352
Past Participle	P	5939
Other	O	6136

Table 2. Syntactic Category codes for WTYPE.

Status of Word	Code	occurrences
Dialect	D	2780
Alien	F	6003
Archaic	A	959
Colloquial	Q	405
Capital	C	2
Erroneous	N	0
Nonsense	E	62
Nonce Word	W	33
Obsolete	O	10549
Poetical	P	183
Rare	R	2756
Rhetorical	H	22
Specialised	$	7731
Standard	S	58065
Substandard	Z	0

Table 3. The possible values of STATUS.