MultiMRC

USER MANUAL

Interface for the MRC Psycholinguistic Database

UNC-CH Cognitive Science Program
Harry Halpin
hhalpin@email.unc.edu
Rough Draft: 12/8/01

Table of Contents

1. Introduction to MultiMRC

1.1 What is MultiMRC?

MultiMRC is a software program that allows cognitive science researchers, such as psycholinguists, to use the powerful MRC Pscyholinguistic Database on a standalone computer on large amounts of text. This allows the retrieval of large amounts of psycholinguistic statistical knowledge on words in the English language such as number of syllables, written word frequency, concreteness, and so on. With this information in an readily accessible format, making hypotheses about the properties of various types of natural language corpora become less a matter of qualitative guess-work and more a matter of quantitive knowledge. This product, like the MRC Psycholinguistic Database, is available for research purposes.

1.2 What is the MRC Psycholinguistic Database?

The MRC Psycholinguistic Database is one of the largest databases of psycholinguistic information available. Psycholinguistics, of the psychological study of language, often has to take into account the human usage or linguistic performance – not only linguistic competence. However, statistics regarding word frequency and other psychological measures require an immense amount of information gathering and theoretical considerations. Luckily, many psycholinguistic researchers have spent a considerable amount of time and effort to compile these statistics, and these have been put into together by into one large database.

The database has 150837 words and provides information about 26 different psycholinguistic properties, although not every property is available for every word. For example, only 9240 of the words contains imagery ratings. For information on the compilation of the MRC, please visit the following web-site:
http://www.psy.uwa.edu.au/MRCDataBase/mrc2.html

The original MRC Psycholinguistic Database was available as an online service, and that has now been discontinued. However, researchers soon created a web-enable version that is accessible via the web at:
http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm

Also, the original UNIX programs that the MRC consists of can be mailed to you for a nominal handling charge from:

```
Oxford Text Archive
Oxford University Computing Service
13 Banbury Road,
Oxford OX2 6NN
U.K.

Tel: Oxford (0865) 56721
```

MultiMRC was created to facilitate the usage of MRC. In particular, the web-site had limited capabilites in processing large amounts of text and then producing the results in a format readily accessible. The UNIX utilities are also not easily accessible to psycholinguists not familiar with UNIX. So, MultiMRC was created to fill this gap by providing a standalone, Windows compatible interface to the MRC Psycholinguistic Database.

1.3 Installing MultiMRC

MultiMRC currently is available as *multimrc.zip* from the http://lcb.unc.edu website under the "Software" section.

1) Before dowloading MultiMRC, you should create a directory to hold it in, like *C:\multimrc*.

2) When *multimrc.zip* is downloaded use WinZip or some other "unzip" program to extract MultiMRC to the directoy you created earlier.

3) Often it is easier to use Windows Programs from the Start Menu than the DOS Command Prompt. So right-click on the Start Button and click on the "Open" option. A window will appear. In that window, click on the Programs icon. Then right click anywhere inside the new "Programs" window and then click on the "New->Folder" option. Create the folder, labeling it "MultiMRC". Then right-click inside your newly created folder to then click on the "New->Shortcut" option. When the "Create Shortcut" window comes up, click on the "Browse" option and go to your directory where MultiMRC is stored and then select *multimrc.exe*. Once this has been done, you can select an icon for MultMRC. Now MultiMRC should be accessible from the Start Menu.

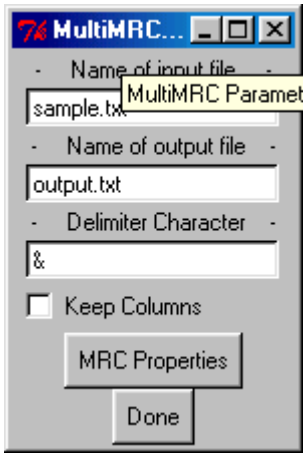
1.4 Files Included with MultiMRC

All of these files are created when you unzip *multimrc.zip* and should be put in the same directory. If any of these files (except the User Manual and sample text) are missing MultiMRC will not work. Also note that the input file and the outputfile should be in the same directory as MultiMRC.

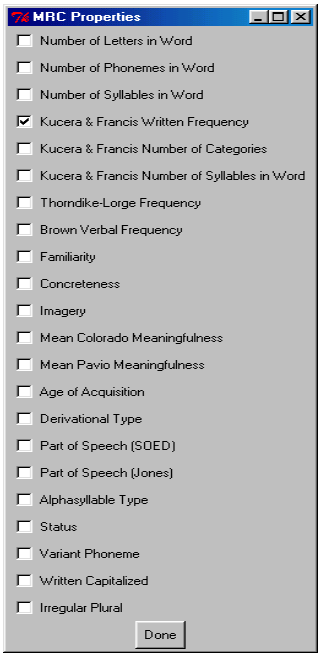
multimrc.exe	File that actually runs MultiMRC
usermanual.pdf	User Manual
getentry.exe	MRC utility - Used by MultiMRC
mrc2.dct	Dictionary file – Used by MultiMRC
sample.txt	sample text for processing

2. Quick Start with MultiMRC

1) Open MultiMRC. A window will appear asking you for the name of various input and output files. For the name of the input file enter *sample.txt* and for the name of the output file enter *output.txt.* It will also ask you for a delimiting character. A "delimiting character" is what separates sentences, if you have sentences, in your input file. For *sample.txt* just enter in &.



2) Then click on the "MRC Parameters" Button. A new screen will appear. This screen will have the various types of linguistic properties. Click on the property to get that information for your text from the MRC Database In this example, it is always useful to find out what the Kucera Written Word Frequency is, so check that one.



3) Then, since you are only finding one property, don't click on the "Keep Columns" check button.

4) Then press the "Done" Button. MultiMRC will freeze up for a few minutes. *The database access is slow – MultiMRC is working.* Then MultiMRC will exit and disappear from your screen.

5) Go to Windows explorer and open *output.txt* with Notepad or some other text editor. The first column will have the sentence number, the second column the word nubmer, the third column the actual word itself, and the next column the Kucera-Francis Written Word Frequency.

3. Using MultiMRC

3.1 Format of Input File

The input file used by MRC can be a collection of words or a collection of sentences. MultiMRC parses this file according to whatever is the "delimiting character" specified in the window. A "delimiting character" separates or "delimits" sentences. For example, if you had a collection of sentences you would probably want the delimiting character to a period. You are free to use other characters though. Also note that the MRC can only take words that are clearly separated by spaces, and will think anything separated by spaces is a word. If there is a word with some punctuation, such as "war," it will not be able to retrieve the proper property values for "war" since it will consider "war," to be one word and look it up in the MRC. So, punctuation such as commas should be removed before being processed through MultiMRC. See Section 3.4 to see how the output looks like.
An example input file would be, with delimiting character '&':

Tom and Mary went to the school yesterday morning &

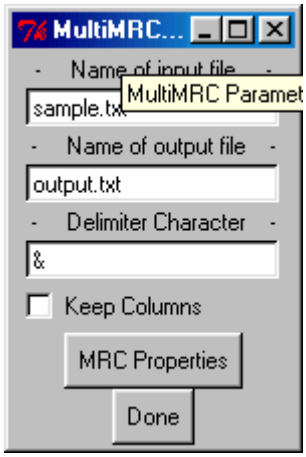
There they will be taught about John Brown and the incident at Harpers Ferry &

The MRC will most likely not be able to find any information for proper names such as "Tom" and "John Brown" or places such as "Harper's Ferry".

3.2 MultiMRC Parameters Window

The MultiMRC Parameters window allows one to enter the three parameters needed for MultiMRC to run properly – the name of the input file with the words or sentences to be processed, the delimiting character used in those sentences, and the name of an output file. The input file should be in the same directory as MultiMRC, and the output file will be created in the same directory as MultiMRC. The delimiting character can be an ASCII text character. For an explanation of delimiting characters see section 3.1. If a delimiting character is not entered into the program **will not work properly** and will delimit every letter. Also, a check-box allows you to "Keep Columns" if you check it. For an explanation of this, please go to Section 3.4 to see how this influences the format of your output files. Generally, if one wants more than one or two properties for

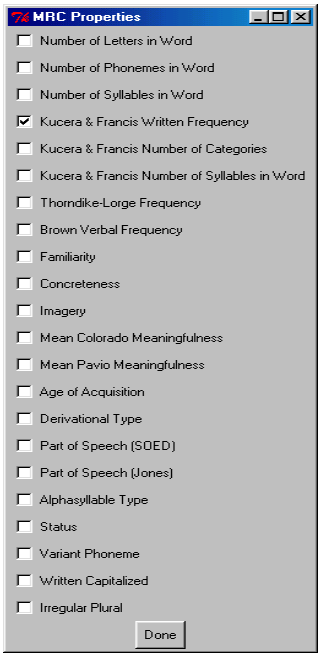
your words you should check “Keep Columns” so the columns and names are kept in the output file. Once these parameters are entered, please press the “MRC Properties” button. Once you have selected the various psycholinguistic properties you wish to investigate with MRC, move on to click the “Done” button.



MultiMRC Parameters Window

3.3 MRC Properties

The MRC Properties button displays a list of all the psycholinguistic properties that can be accessed by MRC. To have a particular property investigated for all of the words in your input file, click the check box next to the name of the property. Once all the names have been selected, click the “Done” button. The program will then begin processing. This will take a minute or two for sentences, up to several hours for a large text corpus. While this is happening *MultiMRC will apparently be frozen, but do not worry; it is working*. An explanation of what these properties mean is currently beyond the scope of this manual, for references please check the MRC web-site at: http://www.psy.uwa.edu.au/MRCDataBase/mrc2.html.



MultiMRC Properties Window

3.4 Format of Output File

The format of the output file can take two main forms, depending on whether or not one clicked on the "Keep Columns" check-box in the MRC (See Section 3.2). The format of the output is in columns to allow easy processing by a variety of software. If the "Keep Columns" check-box was checked, the columns will also have a heading on top that has the MRC abbreviations for the various properties. The number of the sentence or other category defined by the delimiter will be the first column. The number of the word within that category will be the second column. All the properties will be listed and blank columns will appear where they were supposedly. If "Keep Columns" was not checked then there will be no heading and the properties that were checked in the MRC Properties screen during this usage of MultiMRC will be listed in the order that they appear on the MRC Properties Screen. Also note that if all the zeros are present in the entry for a column, the entry for the column is blank in MRC and so this word should be excluded from any analysis. Here is an example of an output file:

```
                                    NLET      NPHON       NSYL  K-F-FREQ
   1     1                 Welcome                 06          2     00050
   1     2                      to                 02          1     26149
   1     3                     the                 00          0     69971
   1     4                     MRC                 00          0     69971
   1     5     Psycholinguistic                    00          0     69971
   1     6                Database                 00          0     69971
   2     1                       A                 01          1     23237
   2     2                  wealth                 04          1     00022
   2     3                      of                 02          1     36411
   2     4             information                 09          4     00269
   2     5                      at                 00          1     05378
```

2	6	your	00	1	00923
2	7	fingertips	00	0	00000
3	1	So	02	1	01984
3	2	begin	00	2	00084
3	3	your	00	1	00923
3	4	research	06	2	00171