Behavior Research Methods, Instruments, & Computers
1988, 20 (4), 398-403

The University of Valencia's computerized word pool

SALVADOR ALGARABEL, JUAN CARLOS RUIZ, and JAIME SANMARTÍN
University of Valencia, Valencia, Spain

This paper presents the University of Valencia's computerized word pool. This is a database that includes 16,109 Spanish words, together with 11 psychological variables for limited groups of items. The purpose behind the creation of this database was to have available a large quantity of verbal stimuli in a well-controlled system, ready for automatic selection. The description includes a summary of statistics on each of the 11 psychological variables, together with a correlational and factor analysis of them. This statistical analysis produces results close to those obtained for equivalent English material.

The purpose of the present paper is twofold. First, a description of the University of Valencia's word pool is provided. This is the first computerized word pool available to researchers who need to select Spanish verbal material. Spanish is a language of widespread use as a first or second language in many English-speaking countries, particularly the United States. This fact, together with the relevance of variables associated with verbal material (Rubin, 1980; Rubin & Friendly, 1986), justified the development of the database. Second, the present paper includes a statistical analysis of the variables in the word pool to make the database comparable to similar databases available in the English language.

General Description of the Word Pool

The University of Valencia's computerized word pool is a computer database composed of verbal stimuli classified by their predominant grammatical role—substantives, verbs, and adjectives—as defined by the Real Academia Española's (1970) *Diccionario de la lengua Española*. The main reason for the existence of the database is to have a large controlled pool of words available for computerized selection for verbal learning and memory experiments, similar to word pools existing in the English language (Coltheart, 1981; Logie, 1984; Murdock, 1968).

Table 1 provides a general overview of the number of items in each grammatical category according to 11 variables of psychological significance, together with the number of words that have been presented as stimuli in free association norms and therefore have normative data in this additional dimension. Additionally, we have carried out a letter × position count of all letters in the substantive pool (Algarabel, 1987), producing a table with the probabilities of occurrence of every letter of the Spanish alphabet according to position.

The word pool has been organized, using the Microsoft File database (Microsoft Corporation, 1985), in a simple and straightforward manner to be used by novice computer users, and has been implemented on a Macintosh SE microcomputer with a 20-MB hard disk. The database is customized in 11 numeric fields for the first 11 indices specified in Table 1, and a text field for each word definition. Researchers with other types of computers can easily transfer the database to their particular formats from an ASCII version of the word pool, easily generated within the Microsoft File program.

The computer selection of verbal material is carried out in two steps. First, the database is searched for specific parameter values, resulting in a file that includes the specific sampling universe. The specification of the search is simple and in line with the user-friendly characteristic of the Macintosh interface. Second, a simple computer program samples the number of required items for specific applications. We are constantly expanding the database, and in the near future we will implement it with a more sophisticated database program.

Table 1
General Composition (Items per Category and Variable) of the Database

	Substantives	Adjectives	Verbs	Total
No. of Letters	10,206	3,505	2,398	16,109
No. of Syllables	10,206	3,505	2,398	16,109
Frequency (per 500,000)	2,024	576	649	3,249
No. of Meanings	10,206	3,505	2,398	16,109
Imagery	1,742	0	0	1,742
Meaningfulness	1,742	0	0	1,742
No. of Attributes	1,742	0	0	1,742
Concreteness	1,742	0	0	1,742
Categorizability	1,742	0	0	1,742
Familiarity	1,742	0	0	1,742
Pleasantness	1,742	0	0	1,742
Free Association Norms	307	0	0	307

This research was supported by Grant PB86-0311, from The Dirección General de Investigación Científica y Técnica del Ministerio de Educación y Ciencia. Requests for reprints or information on how to obtain the material described in this paper should be sent to Salvador Algarabel, Departamento de Psicología Experimental, Universidad de Valencia, Blasco Ibáñez, 21, 46010 Valencia, Spain.

Variables

The variables included in the database are of an objective and subjective nature. The objective variables are simply counts along a dimension (letters, dictionary entries, etc.), whereas the subjective variables were obtained from a sample of subjects who provided ratings along psychological dimensions.

The objective variables include number of letters, number of syllables, written frequency, and number of dictionary meanings. The subjective variables include imagery, meaningfulness, number of attributes, concreteness, categorizability, familiarity, and pleasantness. The following are descriptions of the objective variables as defined in the database:

Number of letters. This is simply a computer count of the number of letters that compose each word. The Spanish alphabet is similar to the English alphabet, with the addition of two special composite letters "ch" and "ll," which are considered single letters, and the specific letter "ñ." Given that shorter words are associated with faster or more accurate psychological dependent variables, number of letters has been reflected (by multiplying by −1) in correlational analysis.

Number of syllables. To compute the number of syllables of each word, a judge used Spanish rules. For the same reason given for number of letters, number of syllables has also been reflected.

Frequency. This is a text frequency count of words as they occur in language. The Spanish equivalent to Thorndike and Lorge's (1944) count is Juilland and Chang-Rodriguez's (1964) frequency count, which has provided the printed word frequency. This dictionary gives the frequency of occurrence of a word per 500,000 counts. In correlation analysis, we have taken the log10 of raw frequency because of the skewed nature of its distribution (Lfrequency).

Number of dictionary meanings. This variable has been defined as the number of different entries given to a word without qualifiers. The source for the count is the dictionary of the Real Academia Española (1970). In correlation analysis, we have taken the log10 of the number of meanings because of the skewed nature of its distribution (Lmeanings).

All subjective variables—imagery, meaningfulness, number of attributes, concreteness, categorizability, familiarity, and pleasantness—were taken from Bernia and López (1985). They were obtained in the usual manner (see Paivio, Yuille, & Madigan, 1968; Toglia & Battig, 1978). They used a sample of 2,000 subjects, who rated on a 7-point scale their judgments concerning different samples of words.

Additionally, a restricted set of free association norms was obtained (Algarabel, Sanmartín, García, & Espert, 1986) for a number of words included in the database. The word pool simply indicates whether an individual word is or is not in the norms handbook. The free association norms were obtained following a single-response discrete procedure from a sample of 250 university students, who responded to all stimulus words with the first response that came to mind. The norms list all responses given to each stimulus word in terms of percentage, and separately by sex.

Statistical Analysis

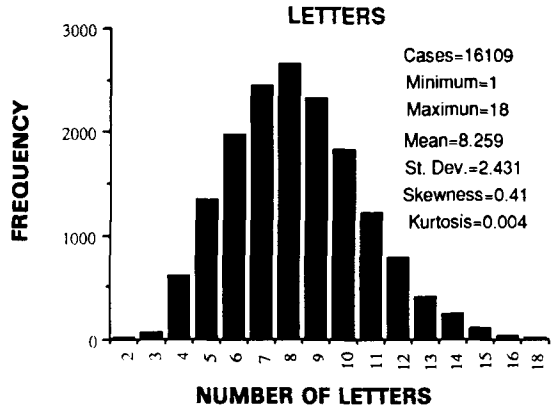
We carried out two types of analysis: summary statistics were computed for every variable, followed by a correlational and factor analysis of the same variables.

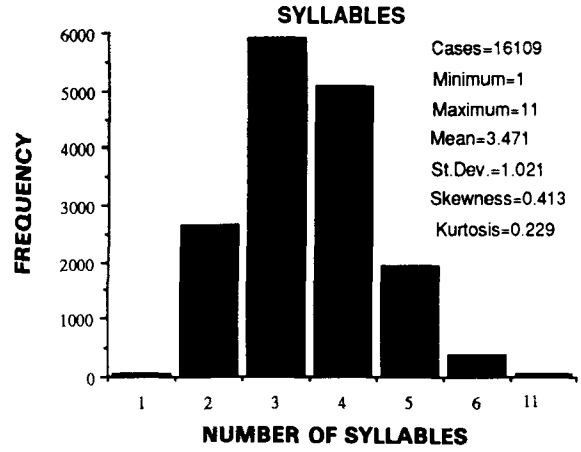
Summary statistics. Figures 1 and 2 present histograms for the variables, together with summary statistics. In general, the summary statistics for the variables coincide with those published for the English language, even those obtained with a more restricted set of words (Gernsbacher, 1983).

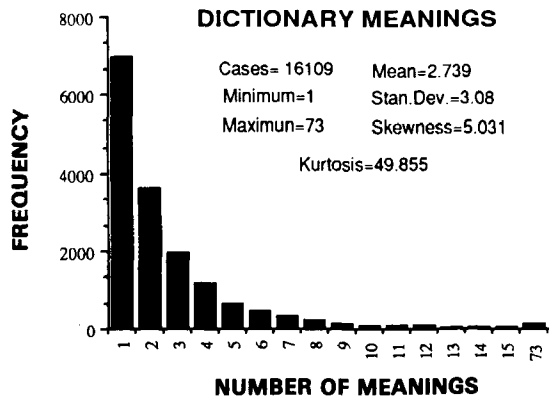
The analysis shows that the distributions for pleasantness, number of attributes, meaningfulness, and familiarity were quite well-shaped; this was also reported in other norms (e.g., Friendly, Franklin, Hoffman, & Rubin, 1982; Paivio et al., 1968). On the other hand, distribution of the length of words in letters is almost perfectly normal, although a bit positively skewed. Number of meanings and written frequency show very skewed, close to exponential, distributions, with a large number of words having very few meanings and low frequency.

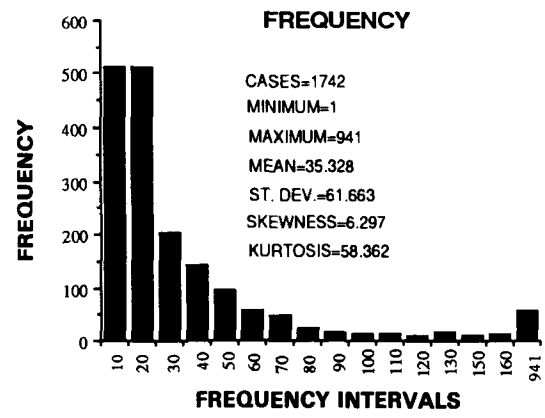
Correlational and factor analysis. Table 2 presents the Pearson intercorrelation matrix for the entire set of variables. Note that the correlations in this analysis are not independent because they are calculated on the same groups of subjects. For this reason, we keep the correlation analysis at a descriptive level. In general, the correlations shown are in the range reported for the same variables in English language studies (Brown, 1984; Brown & Watson, 1987; Friendly et al., 1982; Rubin, 1980; Rubin & Friendly, 1986).

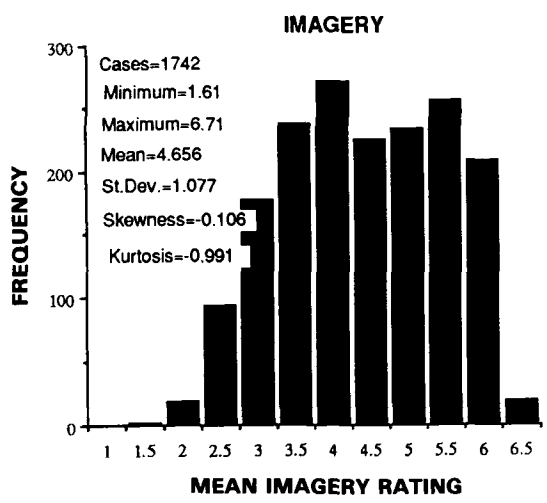
The number of syllables and the number of letters correlate highly (.86), as do concreteness and categorizability (.89). High correlations are also found between imagery-concreteness (.66) and imagery-categorizability (.66). Some of the remaining correlations are of moderate size (Lfrequency-familiarity = .50; imagery-number of attributes = .47; meaningfulness-number of attributes = .50; number of attributes-categorizability = .41), with the rest close to zero. It is interesting to note that the correlation between Lfrequency and familiarity is modest (.50) compared with what the experimental literature would lead us to expect. However, recent reports (Brown & Watson, 1987) seem to show that familiarity ratings do not correlate strongly with written frequency, because they tap psychologically different mechanisms. In fact, the correlation reported by Brown and Watson between the two variables (.365) is lower than the one reported here. On the other hand, Gernsbacher (1984) showed that variability in experiential familiarity with words of low written frequency was responsible for some of the inconsistencies found in the literature involving these and other











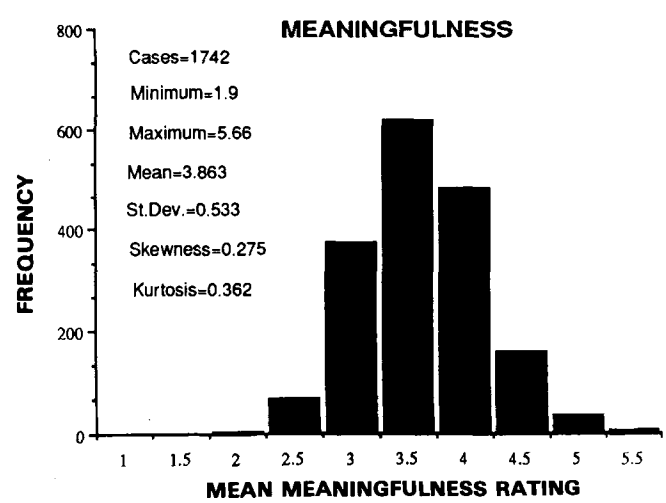


Figure 1. Histogram plots of length in letters, number of syllables, number of dictionary meanings, written frequency per 500,000 words, imagery (7-point scale), and meaningfulness (7-point scale), together with statistical summaries of available samples of words in the University of Valencia's word pool.

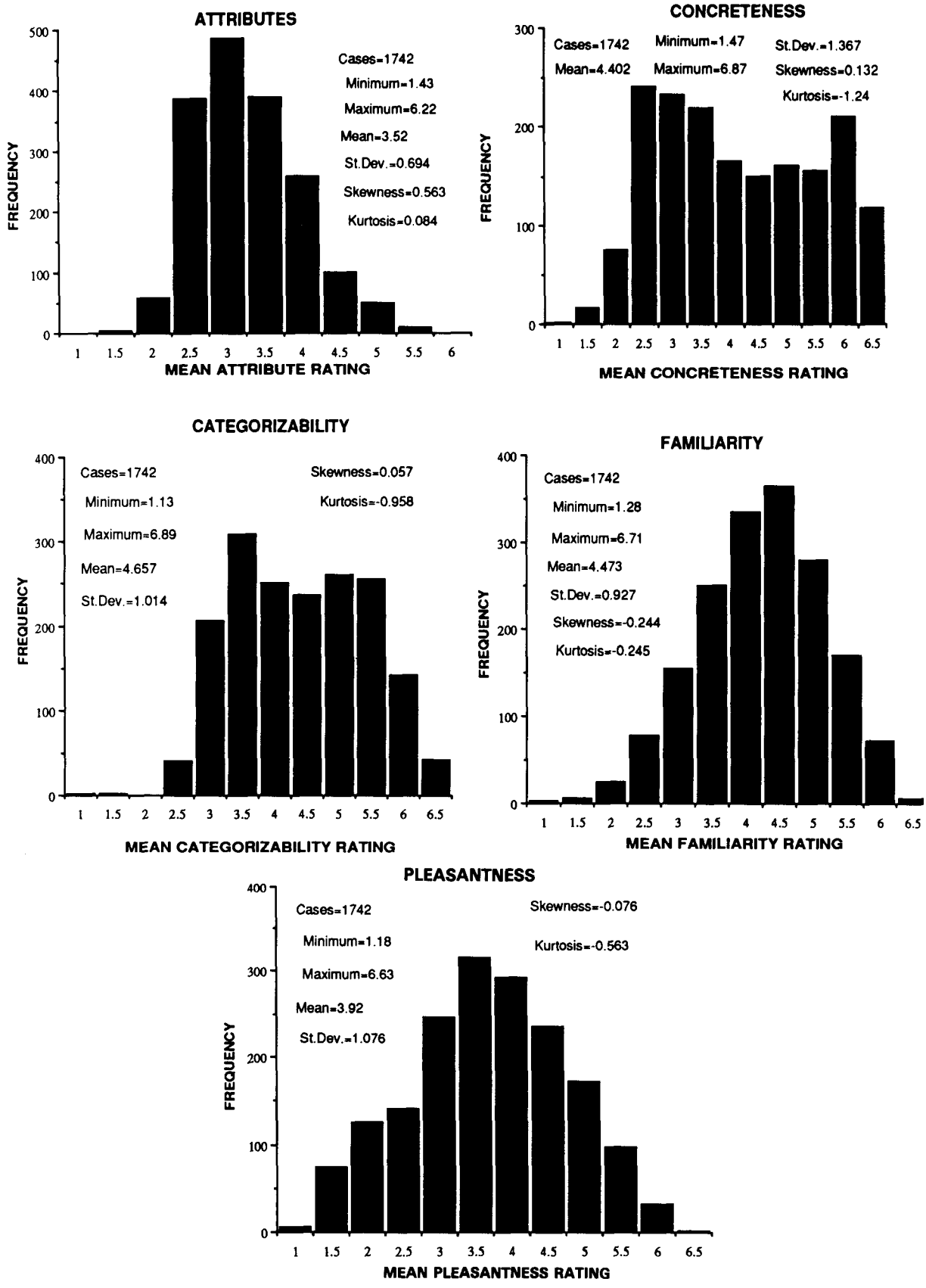


Figure 2. Histogram plots of attributes, concreteness, categorizability, familiarity (7-point scale), and pleasantness (7-point scale), together with statistical summaries of available samples of words in the University of Valencia's word pool.

Table 2
A Correlation Matrix

	1	2	3	4	5	6	7	8	9	10
1 Lmeanings										
2 Lfrequency	38									
3 Letters*	27	21								
4 Imagery	13	09	32							
5 Meaningfulness	04	31	−08	18						
6 Attributes	10	29	05	47	50					
7 Concreteness	13	−01	38	66	−11	32				
8 Categorizability	13	09	36	66	0	41	89			
9 Familiarity	16	50	07	19	34	34	11	22		
10 Pleasantness	02	23	00	13	26	32	02	07	26	
11 Syllables*	27	23	86	25	−09	01	31	28	07	−01

Note—Decimal points have been omitted. *Variable has been reflected by multiplying by −1.

variables. Both sets of data help us to understand why the correlation between written frequency and rated familiarity is lower than expected, although Gernsbacher's claim that written frequency be substituted for familiarity is not supported.

The factor analysis was calculated according to the principal components solution, followed by varimax rotation. The analysis of the present variables (see Table 3) replicates for the most part some of the factors found in the American literature (Rubin, 1980; Rubin & Friendly, 1986). Eleven factors were obtained, although familiarity and pleasantness explained only about 2% of the variance and were dropped. The variance explained by letters, syllables, Lfrequency, Lmeanings, imagery, meaningfulness, attributes, concreteness, and categorizability was 19.29, 17.11, 9.41, 9.27, 9.21, 9.19, 8.85, 8.52, and 6.96, respectively. In the present study, there are two clearly identifiable factors: the first, loaded in imagery (.50), concreteness (.93), and categorizability (.93); the second, loaded on length in letters (.93) and syllables (.95). The other factors in the analysis are loaded in unique variables. This is a logical fact, given that no cluster of variables was input to the analysis. The argument put forward by Brown and Watson (1987), about the relative independence of familiarity and written frequency, is supported in this factor analysis. The factor

Table 3
Factor Analysis

	1	2	3	4	5	6	7	8	9
Letters*	20	93	−03	00	01	10	07	00	09
Syllables*	13	95	−04	−01	02	10	09	−02	05
Lfrequency	−01	14	13	10	25	19	92	11	01
Lmeanings	06	15	01	01	05	97	16	03	03
Imagery	50	15	11	06	06	06	01	20	81
Meaningfulness	−06	−06	94	11	15	01	12	22	07
Attributes	25	−02	27	15	13	02	12	88	16
Concreteness	93	18	−08	−01	02	05	−04	10	18
Categorizability	93	15	00	03	10	04	03	15	16
Familiarity	09	02	15	12	94	05	23	11	04
Pleasantness	02	−01	10	98	11	01	08	12	04

Note—Decimal points have been omitted. *Variable has been reflected by multiplying by −1.

loaded in frequency (.92) is modestly loaded in familiarity (.23), and vice versa (.25 and .94). Looking to past studies, what Rubin (1980) calls the spelling and sound factor is loaded in length in letters (.93), syllables (.81), number of meanings (.38), the Thorndike and Lorge (1944) frequency count (.40), and the Kučera and Francis (1967) frequency count (.19). These are similar to the equivalent factors from this analysis: number of meanings (.14), Lfrequency (.14), letters (.93), and syllables (.95). There are more differences in the second factor (what Rubin called imagery and meaning), which was loaded in imagery, concreteness, categorizability, and meaningfulness, respectively, .90, .91, .91, and .73 in comparison with .50, .93, .93, and .063. The most recent study by Rubin and Friendly (1986) supported a similar conclusion: their spelling factor showed a .95 load on length in letters, and the imagery and meaningfulness factor was loaded .88, .81, and .76 in imagery, concreteness, and meaningfulness, respectively.

In conclusion, the statistical analyses presented here produced results very similar to the ones reported in the English literature. From this point of view, the relationships between variables commonly used in psychological experimentation, such as frequency, familiarity, imagery, and concreteness, maintain the same relative importance to each other as word dimensions. We hope, on the other hand, that this Spanish database is useful in and of itself, and can be related to other English databases in learning, memory, and psycholinguistic studies.

REFERENCES

ALGARABEL, S. (1987). *Letter counts of the University of Valencia's substantive database*. Unpublished manuscript.

ALGARABEL, S., SANMARTÍN, J., GARCÍA, J., & ESPERT, R. (1986). *Normas de asociación libre para investigación experimental*. Unpublished manuscript, Universidad de Valencia, Departamento de Psicologia Experimental.

BERNIA, J., & LÓPEZ, L. (1985). *Estudio normativo de vocabulario en siete dimensiones*. Unpublished manuscript.

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.

BROWN, G. D. A., & WATSON, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition*, **15**, 208-216.

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

FRIENDLY, M., FRANKLIN, P. E., HOFFMAN, D., & RUBIN, D. C. (1982). The Toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, **14**, 375-399.

GERNSBACHER, M. A. (1983). *The experiential familiarity of low frequency words*. Unpublished manuscript.

GERNSBACHER, M. A. (1984). Resolving 20 years of inconsistent interaction between lexical familiarity and orthography, concreteness, and polisemy. *Journal of Experimental Psychology: General*, **113**, 256-281.

JUILLAND, A., & CHANG-RODRIGUEZ, E. (1964). *Frequency dictionary of Spanish words*. London: Mouton.

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

LOGIE, R. H. (1984). Computer selection of verbal research materials. *Behavior Research Methods, Instruments, & Computers*, **16**, 59-60.

MICROSOFT CORPORATION. (1985). *Microsoft File manual*. Bellvue, WA: Author.

MURDOCK, B. B., JR. (1968). Modality effects in short term memory: Storage or retrieval. *Journal of Experimental Psychology*, **77**, 79-86.

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

REAL ACADEMIA ESPAÑOLA. (1970). *Diccionario de la lengua Española*. Madrid: Espasa-Calpe.

RUBIN, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. *Journal of Verbal Learning & Verbal Behavior*, **19**, 736-755.

RUBIN, D. C., & FRIENDLY, M. (1986). Predicting which words get recalled: Measures of free recall, availability, goodness, emotionality, and pronounceability for 925 nouns. *Memory & Cognition*, **14**, 79-94.

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

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

(Manuscript received October 9, 1987;
revision accepted for publication February 12, 1988.)