

Normative data on the familiarity and difficulty of 196 Spanish word fragments

CARMEN DASÍ, MARÍA JOSÉ SOLER, and JUAN CARLOS RUIZ
University of Valencia, Valencia, Spain

In this article, normative data on the familiarity and difficulty of 196 single-solution Spanish word fragments are presented. The database includes the following indices: difficulty, familiarity, frequency, number of meanings, number of letters given in the fragment, first and/or last letters given, and ratio of letters to blanks. A factor analysis was performed on difficulty, and two factors were obtained. Frequency, familiarity, and number of meanings loaded highly on the first factor, which we consider to measure lexical processes, whereas number of letters in the fragment, first and/or last letters given, and ratio of letters to blanks loaded highly on the second factor, which we judge to be determined by perceptual information. Regression analyses using factor scores as predictors showed that both factors accounted for a significant part of the completion probability scores. The full set of these norms may be downloaded from the Psychonomic Society Web archive at www.psychonomic.org/archive/.

A word fragment completion test is an experimental paradigm used in cognitive psychology research (e.g., in research on implicit memory, verbal learning, visual memory, word recognition, bilingual memory, and clinical and neuropsychological assessment). In memory research, word fragment completion tests are one of the more frequently used ways of eliciting nonconscious memory (Toth, 2000). In these implicit memory tests, subjects are told to complete a word fragment by supplying deleted letters [e.g., A P _ _ _ O T becomes APRI-COT in English (Erickson, Gaffney, & Heath, 1987), and G I G _ _ _ E becomes GIGANTE in Spanish].

Subjects benefit from having had a prior encounter with the stimulus (see, e.g., Rajaram, Srinivas, & Travers, 2001; Roediger & Blaxton, 1987; Tulving, Schacter, & Stark, 1982; Weldon, 1993). This benefit is called *priming*. When word fragments are used, the study of the effect of different variables on priming requires the control of variables such as word frequency, familiarity, and concreteness, as well as the specific control of fragment completion difficulty, or *baseline rate of completion*, in order to avoid interpretation problems.

Two relevant word fragment normative databases for the English language were collected in the late 1980s (Erickson et al., 1987; Gibson & Watkins, 1988).

Erickson et al. (1987) presented a corpus of 192 English single-solution word fragments six to eight letters in length. Some of the indices included in this database are frequency from the Kučera and Francis (1967) norms, familiarity (on a 7-point scale), and difficulty of the fragments measured as the proportion of correct responses. For this corpus, the variables of familiarity, first and/or

last letters given, and ratio of letters to blanks in the fragments were the best predictors of completion difficulty.

The Gibson and Watkins (1988) word database was made up of 1,086 items three to eight letters in length. Each item is uniquely specified by a two-letter fragment. The pool is very extensive, but its only index for each word is frequency, as described in the Kučera and Francis (1967) norms.

The word completion task has been used in research in other languages as well, including Japanese (Kitao & Inoue, 1998), Chinese (Li & Lin, 1998), Korean (Park, 1998), and Spanish (Ruiz-Vargas & Cuevas, 1999; Soler, Ruiz, & Dasí, 2002), but published word fragment norms are few. For example, Nyberg and Olofsson (1991; Olofsson & Nyberg, 1992) collected a pool of 2,159 single-solution Swedish word fragments. These authors included the difficulty and familiarity of 104 fragments and some factors related to the structure of the fragments, such as the number of letters given and whether or not the first and last letters were given. A stepwise regression analysis of the data, including some structural factors based on Erickson et al. (1987), revealed that the ratio of letters to blanks, familiarity, and the presence of unusual letters were the determinant factors for predicting fragment completion probability.

A recent review of the published normative studies in the Spanish language (from Spain and Latin American countries) on materials for psychological experiments shows the absence of normative data in Spanish word fragments (Pérez, Campoy, & Navalón, 2001). The main purpose of this article is to aid in the selection of appropriate stimuli by providing normative data on one-solution fragments for some Spanish words. This pool has been completed with the most relevant indices described in the published norms (Erickson et al., 1987; Gibson & Watkins, 1988; Olofsson & Nyberg, 1992).

Correspondence concerning this article should be addressed to C. Dasí, Facultad de Psicología, Avda. Blasco Ibáñez, 21, 46010 Valencia, Spain (e-mail: carmen.dasi@uv.es).

METHOD

Participants

The participants were 602 undergraduate students (18–44 years of age) from the University of Valencia, who received bonus points toward their grades for taking part in the experiment. All were native Spanish speakers and had normal or corrected-to-normal vision.

Materials and Procedure

A pool of 220 word fragments was prepared by randomly deleting two to four letters from each word, following Rajaram and Roediger (1993). Two letters were deleted from five-letter words, two or three letters were deleted from six-letter words, and three or four letters were deleted from seven-letter words. If a fragment had more than one solution, another random deletion of letters was performed until the fragment had a single solution. The words were selected from the University of Valencia's computerized word pool (Algarabel, Ruiz, & Sanmartín, 1988) and from the Royal Spanish Academy's (2000) *Corpus de Referencia del Español Actual*, which includes words from the dictionary of the Royal Spanish Academy (2001) as well as technical words from the different sciences. The words were morphologically simple nouns and were five to seven letters in length. The word fragments were normed over a 2-year period.

In order to obtain the difficulty and familiarity norms, a two-phase word completion task was used.

In Phase 1, a list of words was presented in lowercase letters at the center of a computer screen for 8 sec per word. The participants were instructed to judge the familiarity of each presented word on a scale ranging from 1 (*a very little-known word*) to 7 (*a very well-known word*), following Erickson et al. (1987).

In Phase 2, the participants were given a list of word fragments presented in lowercase letters at the center of the screen for 12 sec per word. Half of the words corresponded to those presented in Phase 1, and the other half were new words. The participants were asked to complete each fragment with the first word that came to mind and to write the solution on a sheet of paper. No participant was required to complete more than 120 word fragments. The set of critical fragments not presented in Phase 1 served as a measurement of the difficulty of the fragments. *Difficulty* refers to the proportion of participants who successfully completed a fragment with no prior exposure to the word represented by the fragment in Phase 1 (Erickson et al., 1987).

We decided to use a two-phase word fragment completion task because this design provides the familiarity index, the word fragment difficulty index, and another index not included in this normative study but relevant for researchers who need to select word fragments for their experiments. This new index could be called the *word fragment primability index*, or the probability of a word fragment's correct completion after it has been studied minus the probability of its correct completion when it has not been studied.

All the fragments were constructed with the idea that they should have single solutions. However, 24 of the fragments were found to have more than one solution and therefore were not included in the set. The final pool consisted of 196 words and their corresponding fragments, and the following indices were specified: number of meanings of the words, word frequency, number of letters given in the fragments, ratio of given letters to blanks, first and/or last letters given, word familiarity, and difficulty.

The number of meanings was obtained from the dictionary of the Royal Spanish Academy (2001) and represents the number of accepted common-noun meanings of each word.

Frequency represents the number of occurrences of a word per million. In this study, only the word frequency estimates provided in Sebastián, Martí, Carreiras, and Cuetos's (2000) computerized lexical pool (LEXESP) have been included, because it has been claimed that a word frequency effect is more likely to emerge when

recent frequency measures are used (Bonin, Peereman, Malardier, Méot, & Chalard, 2003; Ellis & Lambon Ralph, 2000). LEXESP is a pool of 5,020,930 Spanish words.

The number of letters given in the fragment varies from two to four, depending on word length.

The ratio of letters to blanks refers to the quotient between the number of letters given in the fragment and the number of missing letters.

First and/or last letters given is a 3-point scale on which 0 means that neither the first nor the last letter is given in the fragment, 1 means that only the first or only the last letter is given, and 2 means that both are given.

Familiarity is usually defined as the estimated frequency of the occurrence of a word, in written or spoken form, in the participants' own daily lives or in the daily lives of people like themselves (Connine, Mullennix, Shernoff, & Yelen, 1990). This index was obtained using a 7-point scale (through the procedure described above). The average sample size for the 196 words was 132.42 subjects (*SD*, 30.68; range, 93–225).

Difficulty is in fact a measure of "easiness," in that it indicates the proportion of correct fragment completion. In our database, it was determined using the procedure described above. The average sample size was 139.68 subjects (*SD*, 35.30; range, 93–256).

Results and Discussion

In the analysis of difficulty, 4 participants were eliminated because their performance was 2 *SDs* below the mean. In fact, their proportion of correct responses was near zero.

Table 1 shows the indices included in the database and some descriptive statistics.

The mean familiarity rating was 4.49 (*SD* = 1.85), with values between 1.02 and 6.75. The distribution was slightly negatively skewed (–.63), but this is not a good descriptive measure because the distribution does not have a unique mode. There is a major mode on the right side around the value 6 ($n = 23$) and a minor mode on the left side at around 1.25 points ($n = 14$). This is due to the fact that we have included a pool of 41 very low-frequency words (no occurrences per million in Sebastián et al.'s, 2000, norms). For these very low-frequency words, familiarity is close to 1 (i.e., the correlation between familiarity and frequency is very high; see Table 2). These data are similar to those of Erickson et al. (1987). They have a highly skewed familiarity distribution because they included a large number of very unfamiliar words (modal frequency of 0 in the norms of Kučera & Francis, 1967).

Table 1
Difficulty and Descriptive Statistics of the Variables Included in the Factor Analysis

Variable	<i>M</i>	<i>SD</i>	Skewness	Minimum	Maximum
Difficulty	.37	.32	.58	.00	1.00
FLG	1.12	.64	–.12	.00	2.00
RLB	1.41	.41	.10	1.00	2.00
NLG	3.53	.52	–.34	2.00	4.00
NM	4.20	4.12	1.90	.00	27.00
Familiarity	4.49	1.85	–.63	1.02	6.75
Frequency	28.17	37.67	1.61	.00	166.00

Note—FLG, first and/or last letters given; RLB, ratio of given letters to blanks; NLG, number of letters given in the fragment; NM, number of meanings.

Table 2
Correlation Among Predictor Variables

Variable	1	2	3	4	5	6
1. NM	–					
2. LogF	.58‡	–				
3. Familiarity	.52‡	.83‡	–			
4. NLG	–.07	–.01	.01	–		
5. RLB	–.05	.02	.06	.63‡	–	
6. FLG	.04	.12	.18*	.29‡	.31‡	–
7. Difficulty	.15*	.39‡	.54‡	.11	.37‡	.19†

Note—NM, number of meanings; LogF, log-transformed frequency; NLG, number of letters given in the fragment; RLB, ratio of given letters to blanks; FLG, first and/or last letters given. * $p < .05$. † $p < .01$. ‡ $p < .001$.

Difficulty shows a mean of .37 ($SD = .32$). This variable has a positively skewed distribution (.58) with probabilities ranging from 0 to 1. The findings are similar to Tulving et al.'s (1982) baseline rate of completion (.30), and it is clearly higher than the probability of .13 obtained by Erickson et al. (1987), who attributed their low success level to population differences.

The matrix of Pearson correlations among mean values of the numerical variables for the 196 fragments is reported in Table 2. Frequency values were transformed using the formula $\log(1 + x)$ (Cuetos & Alija, 2003).

There is a substantial degree of colinearity among the variables, since some of the correlations reflect the relationship between closely related measures of similar constructs, such as frequency and familiarity (.83, $p < .001$), number of letters given and ratio of letters to blanks (.63, $p < .001$), and frequency and number of meanings (.58, $p < .001$). Number of meanings is probably related to frequency and familiarity because words with more than one meaning are probably used more often, and in different contexts, than words with single meanings; therefore, their frequency and familiarity are also higher.

From the correlations matrix, we can deduce that our data require a multivariate approach in which the contribution of each predictor can be compared when variance from overlapping variables is controlled. Our multivariate strategy is similar to that which has been described in studies on other languages, such as English (Gilhooly & Logie, 1981, 1982; Paivio, Clark, Digdon, & Bons, 1989; Rubin, 1980; Whaley, 1978), Italian (Bates, Burani, D'Amico, & Barca, 2001), and French (Bonin et al., 2003).

In order to reduce the correlation matrix to an interpretable summary, the variables were submitted to a factor analysis (i.e., a principal components analysis) with orthogonal (varimax) rotation. The standard criterion of eigenvalues greater than or equal to 1 was used to determine the number of factors. Variables that were highly intercorrelated consequently loaded on a common factor.

Table 3 presents the factor solution, referring to the numbers of correlations, or loadings, among the variables and factors.

Two factors with eigenvalues greater than 1.0 emerged from this analysis. The first factor accounted for 38.9% of the variance and loaded heavily onto the measures of

frequency, familiarity, and number of meanings. This factor can be described as a lexical factor because it is based on the three variables that presumably represent the accessibility of a lexical representation in memory. The second factor accounted for 30.8% of the variance and loaded heavily onto number of given letters, ratio of given letters to blanks, and first and/or last letters given. This factor is determined by the amount of (perceptual) information given to the participants in a fragment.

The two factors together account for 69.7% of the common variance. When Table 3 is examined, it can be seen that the factors are quite clean, with three variables loading onto each factor. With these two latent variables, each with a relatively straightforward interpretation, we conducted a stepwise multiple regression analysis with the proportion of correct responses (i.e., difficulty) as the dependent variable. This was done in order to assess the contribution of each factor when it was entered into the equation during the last step. The problems of colinearity are not present now because the factors are orthogonal, and therefore they are not correlated. Together, the factor scores accounted for 25.39% of the variance in the dependent variable ($F = 32.84$, multiple $r = .50$, $p < .001$). Both factors contributed significantly to explain the variance of the dependent variable (Table 4). The lexical factor accounted for 16.62% and the amount-of-information factor accounted for 8.76% of the variance. As Table 4 shows, the significance of the coefficients regression for the two factors is $< .001$, which minimizes the risk of Type I errors.

Our findings are similar to those of Erickson et al. (1987) and Olofsson and Nyberg (1992), although in both normative studies the data analyses consisted of a multiple regression on the independent variables. We agree with Erickson et al. that “for fragments that contain first and last letters, the word search process should be easier . . . , and having fewer blanks in the fragment would rather obviously influence difficulty. Also, of course, if subjects are not familiar with a word, they are not apt to use it to complete a word fragment” (p. 372).

The present results give support to the conclusions reached by Ostergaard (1998) that different sources of information are available when word completion tasks are used: (1) the information accumulated in memory from all prior encounters with the item (e.g., familiarity, frequency, and number of meanings), and (2) perceptual

Table 3
Results of Factor Analysis Across Predictor Variables

Variable	Factor 1	Factor 2
NM	.78	–.07
NLG	–.08	.85
RLB	–.03	.86
FLG	.16	.61
LogF	.92	.05
Familiarity	.90	.11

Note—NM, number of meanings; NLG, number of letters given in the fragment; RLB, ratio of given letters to blanks; FLG, first and/or last letters given; LogF, log-transformed frequency.

Table 4
Results of the Multiple Regression Analysis with Difficulty as the Dependent Variable

Variable	B	SE	β	t	p
Constant	.371	.020		18.662	.000
Factor 1	.131	.020	.408	6.558	.000
Factor 2	.095	.020	.296	4.761	.000

Note—Factor 1 measures lexical processes. Factor 2 measures perceptual information.

information from the stimulus, which depends on how the item is presented (e.g., number of given letters, ratio of given letters to blanks, and first and/or last letters given).

Finally, we are in agreement with Shaw (1997) that the difficulty of word completion has to be estimated using normative data, since collecting data during experiments (as in implicit memory research, for example) could pose some problems. This is due to the possibility of obtaining widely varying probabilities of correct completion within the set of words. Thus, it is interesting to develop more consistent materials for selection of appropriate stimuli when word fragment completion tasks are used in cognitive research. Furthermore, the normative data presented in this article include the variable of difficulty, not previously available for the Spanish language. We think that future normative studies in Spanish are necessary to amplify this database with other interesting indices in order to know other characteristics of the verbal materials.

REFERENCES

- ALGARABEL, S., RUIZ, J. C., & SANMARTÍN, J. (1988). The University of Valencia's computerized word pool. *Behavior Research Methods, Instruments, & Computers*, **20**, 398-403.
- BATES, E., BURANI, C., D'AMICO, S., & BARCA, L. (2001). Word reading and picture naming in Italian. *Memory & Cognition*, **29**, 986-999.
- BONIN, P., PEEREMAN, R., MALARDIER, N., MÉOT, A., & CHALARD, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, **35**, 158-167.
- CONNINE, C. M., MULLENNIX, J., SHERNOFF, E., & YELEN, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychological: Learning, Memory, & Cognition*, **16**, 1084-1096.
- CUETOS, F., & ALJIA, M. (2003). Normative data and naming times for action pictures. *Behavior Research Methods, Instruments, & Computers*, **35**, 168-177.
- ELLIS, A. W., & LAMBON RALPH, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 1103-1123.
- ERICKSON, J. R., GAFFNEY, C. R., & HEATH, W. P. (1987). Difficulty and familiarity norms for 192 single-solution word fragments. *Behavior Research Methods, Instruments, & Computers*, **19**, 370-376.
- GIBSON, J. M., & WATKINS, M. J. (1988). A pool of 1,086 words with unique two-letter fragments. *Behavior Research Methods, Instruments, & Computers*, **20**, 390-397.
- GILHOOLY, K. J., & LOGIE, R. H. (1981). Word age-of-acquisition, reading latencies and auditory recognition. *Current Psychological Research*, **1**, 251-262.
- GILHOOLY, K. J., & LOGIE, R. H. (1982). Word age-of-acquisition and lexical decision making. *Acta Psychologica*, **50**, 21-34.
- KITAO, N., & INOUE, T. (1998). The effects of spaced repetition on explicit and implicit memory. *Psychologia*, **41**, 114-119.
- KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- LI, J., & LIN, Z. (1998). A comparative study of the effects of word frequency in direct and indirect measures of memory. *Psychological Science*, **21**, 196-200.
- NYBERG, L., & OLOFSSON, U. (1991). *Materials for word completion: Subject generated completions of 198 word stems and a pool of 2,159 single-solution word fragments*. Umeå: University of Umeå, Department of Psychology.
- OLOFSSON, U., & NYBERG, L. (1992). Swedish norms for completion of word stems and unique word fragments. *Scandinavian Journal of Psychology*, **33**, 108-116.
- OSTERGAARD, A. L. (1998). The effects on priming of word frequency, number of repetitions, and delay depend on the magnitude of priming. *Memory & Cognition*, **26**, 40-60.
- PAIVIO, A., CLARK, J. M., DIGDON, N., & BONIS, T. (1989). Referential processing: Reciprocity and correlates of naming and imaging. *Memory & Cognition*, **17**, 163-174.
- PARK, T. J. (1998). Relation between implicit memory and lexical processing. *Korean Journal of Experimental & Cognitive Psychology*, **9**, 95-118.
- PÉREZ, M. A., CAMPOY, G., & NAVALÓN, C. (2001). Índice de estudios normativos en idioma español [Index of normative studies in the Spanish language]. *Revista Electrónica de Metodología Aplicada*, **6**, 85-105.
- RAJARAM, S., & ROEDIGER, H. L., III (1993). Direct comparison of four implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 765-776.
- RAJARAM, S., SRINIVAS, K., & TRAVERS, S. (2001). The effects of attention on perceptual implicit memory. *Memory & Cognition*, **29**, 920-930.
- ROEDIGER, H. L., III, & BLAXTON, T. A. (1987). Effects of varying modality, surface features, and retention interval on priming in word-fragment completion. *Memory & Cognition*, **15**, 379-388.
- ROYAL SPANISH ACADEMY (2000). *Corpus de referencia del español actual* [Reference corpus of current Spanish]. Retrieved Nov./Dec. 2001 from <http://corpus.rae.es/creanet.html>.
- ROYAL SPANISH ACADEMY (2001). *Diccionario de la Lengua Española* [Dictionary of the Spanish Language] (22nd ed.). Madrid: Espasa.
- RUBIN, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. *Journal of Verbal Learning & Verbal Behavior*, **19**, 736-755.
- RUIZ-VARGAS, J. M., & CUEVAS, I. (1999). Priming perceptivo versus priming conceptual y efectos de los niveles de procesamiento sobre la memoria implícita [Perceptual versus conceptual priming and effects of levels of processing on implicit memory]. *Psicothema*, **11**, 853-871.
- SEBASTIÁN, N., MARTÍ, M. A., CARREIRAS, M. F., & CUETOS, F. (2000). *LEXESP, léxico informatizado del Español* [LEXESP: A computerized lexicon of the Spanish language]. Barcelona: Ediciones de la Universitat de Barcelona.
- SHAW, R. J. (1997). Unprimed stem completion is only moderately predicted by word frequency and length. *Behavior Research Methods, Instruments, & Computers*, **29**, 401-424.
- SOLER, M. J., RUIZ, J. C., & DASÍ, C. (2002). More on magnitude of priming in implicit memory tasks. *Perceptual & Motor Skills*, **95**, 889-894.
- TOTH, J. P. (2000). Nonconscious forms of human memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 245-261). Oxford: Oxford University Press.
- TULVING, E., SCHACTER, D. L., & STARK, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 336-342.
- WELDON, M. S. (1993). The time course of perceptual and conceptual contributions to word fragment completion priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 1010-1023.
- WHALEY, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning & Verbal Behavior*, **17**, 143-154.

ARCHIVED MATERIALS

The following materials associated with this article may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, <http://www.psychonomic.org/archive/>.

To access the following files, search the archive for this article using the journal (*Behavior Research Methods, Instruments, & Computers*), the first author's name (Dasi), and the publication year.

FILE: Dasi-BRMIC-2004.zip.

DESCRIPTION: The compressed archive file contains one file: `dasi2004norms.txt`, containing the norms developed in the present arti-

cle as a 50K tab-delimited document file generated by Word 98 for the Macintosh. Each row represents one of 196 fragments; each column, one of seven indices (number of meanings, frequency, number of letters given in the fragment, ratio of given letters to blanks, first and/or last letters given, familiarity, and difficulty).

AUTHOR'S E-MAIL ADDRESS: carmen.dasi@uv.es.

AUTHOR'S WEB SITE: <http://www.uv.es/~dasic>.

(Manuscript received June 6, 2003;
revision accepted for publication June 18, 2004.)