JOURNAL OF MEMORY AND LANGUAGE 32, 781–804 (1993)

Neighborhood Density and Frequency across Languages and Modalities

U. H. FRAUENFELDER, R. H. BAAYEN, AND F. M. HELLWIG

Max-Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

AND

R. SCHREUDER

Interfaculty Research Unit for Language and Speech, Nijmegen, The Netherlands

This research exploits the English and Dutch CELEX lexical database to investigate the form similarity relations between words. Lexical statistics analyses replicate and extend the findings of Landauer and Streeter (1973) concerning the relation between a word's frequency and the density and frequency of its similarity neighborhood. The results for both Dutch and English reveal only a weak tendency for high-frequency written and spoken words to have more neighbors than rare words and for these neighbors to be more frequent than those of rare words. However, the number of neighbors was found to correlate more highly with bigram frequency than with word frequency. To clarify the relations between these properties, a stochastic model is presented which captures the relevant effects of phonotactic structure on neighborhood similarities. The implications of these findings for models of language production and comprehension are considered.

Research in lexical processing suggests that the identification of any given word-form in the mental lexicon depends not just on the evidence in the signal for the word itself, but also on existence in the lexicon of other words that are similar in form to the target. The influence of these lexical competitors or neighbors upon word recognition is expressed in one way or another in most models of word recognition. Indeed, in theoretical accounts of word recognition for both the spoken (Luce, (1986); Marslen-Wilson & Welsh, (1978); Marslen-Wilson, (1987) and the visual (Grainger, (1990); Humphreys, Evett, & Quilian, (1987)) domain, the perceptual choice made in recognizing a target word is assumed to be conditioned by the target's competitor environment. The exact definition of the members of the competitor set and the competitor's influence upon word recognition—which both may differ across modality—is still under intense experimental investigation.

If indeed, the timing of the word recognition processes can only be understood with reference to the set of lexical choices which a listener (or a reader) must discriminate, then research into these processes will have to be based on a proper description and understanding of this set. The description of the on-line lexical search space requires the statistical analysis of lexical databases that accurately represent the current state of the language. Analyses of computerized lexical databases can thus provide us with detailed characterizations of the structural and distributional properties of the words in a language. Such characterizations are essential for assessing the con-

The authors thank William Marslen-Wilson, Guus Peeters, and Eric Schils for their participation in earlier phases of this research. They are indebted to Tom Carr and two anonymous referees for their constructive criticism of an earlier version of this paper. Correspondence and reprint requests should be addressed to U. H. Frauenfelder, FPSE, University of Geneva, 9 route de Drize, CH-1227 Carouge, Switzerland. email: frauenfe@uni2a.unige.ch.

0749-596X/93 $5.00

sequences of various hypotheses and experimental results concerning lexical processing and representation.

Landauer and Streeter (1973) were among the first to study the distribution of words in the lexicon as represented in a lexical database. They defined a similarity neighborhood within the lexicon as that set of words or neighbors from which a given target word cannot be distinguished when there is a specific loss of information (a single letter substitution in their study) about this target word. Two different properties of the similarity neighborhoods were examined: neighborhood density—the number of neighbors of a word—and neighborhood frequency—the frequency of these neighbors. In their analysis, Landauer and Streeter (L & S) compared the similarity neighborhoods of common (frequency greater than 75 occurrences per million in Kučera & Francis (1967)) and rare (one occurrence per million) written English words to test the assumption that these two classes of words do not differ in their similarity neighborhoods. Neighbors were defined as those words that differed from the target by a single letter in any position (e.g., *fun*, *sin*, and *sum* are all neighbors of *sun*). Using this definition, they contrasted the similarity neighborhoods of 50 common and rare English four-letter words. They found that common and rare words were not equivalent in either the density or in the frequency of their neighborhoods; common words had more neighbors than rare words, and the mean frequency of the neighbors of common words was higher than that of the neighbors of rare words.

This finding that common and rare words occupy, on average, different similarity neighborhoods has some important consequences for word recognition research. As L & S pointed out, these results argue against the perceptual equivalence hypothesis, according to which high- and low-frequency words do not differ from each other in any dimensions other than word frequency itself. Common words were shown to have different similarity neighborhoods than rare words. These observed differences in the similarity neighborhoods are hard to reconcile with the well-known word frequency effect (Gardner, Rothkopf, Lapan, & Lafferty, 1987; Rubenstein & Pollack, 1963; Whaley, 1978). According to most accounts of word recognition, the efficiency and speed with which a given word can be recognized is a function of how easily it can be discriminated from its neighbors. This discrimination has been argued (Luce, 1986) to depend upon the number and frequency of the word's neighbors; the more neighbors and the more frequent these neighbors, the slower and more difficult the recognition process is. If, as the results of L & S suggest, high-frequency words have more neighbors that are themselves high frequency, then we would incorrectly predict that high-frequency words are harder to identify than low-frequency words. This apparent conflict between word frequency and the properties of the similarity neighborhoods constitutes an apparent paradox as has been pointed out by L & S and by Nusbaum (1985). It also raises the more fundamental question of how neighborhood density and neighborhood frequency influence the time course of lexical processing.

Given the important implications of these results for current theorizing about the similarity structure of the lexicon and the processes of word recognition, it is essential that these facts about lexical neighborhoods be established clearly. Closer examination of the L & S study reveals several shortcomings that may undermine its reliability. First, it was based upon relatively small samples of words; L & S arbitrarily limited their analysis of 50 randomly selected words of the 260 possible in the frequency classes they defined. Second, their analysis was restricted to only a single length (four letters). Third, the neighborhoods of words in only two frequency ranges were examined. By restricting the analyses in this fashion, the authors were

unable to document the changes in similarity neighborhoods across the entire frequency range and therefore could not shed light on the more general relationship between frequency and similarity neighborhoods. Fourth, the two properties of similarity neighborhoods (neighborhood density and neighborhood frequency) were analyzed separately. An analysis combining the two measures into one global measure may provide more insight into similarity neighborhoods, as we will show below. Lastly, they restricted their analysis to written words.

More recently, however, Pisoni, Nusbaum, Luce, and Slowiaczek (1985) reported on a study that compared the similarity neighborhoods of common and rare spoken words. Their study improved on the original L & S study in two ways. First, it expanded on this study by analyzing words of different lengths (ranging from one to eight phonemes) rather than using one single length. Second, it characterized lexical neighborhoods with an additional measure based on the Luce-choice rule that incorporated both neighborhood frequency and density. One major defect of the study was its use of a very restricted high-frequency range (above 1000 per million). This definition had the negative consequence of producing extremely small samples (2, 36, 39, 14, and 1 item(s) for common words with lengths 1 to 5, respectively).

The results of the Pisoni et al. study only partially replicated the pattern obtained by L & S. On the whole, their results for neighborhood density were quite different. High-frequency words did not have more neighbors than low-frequency words for any of the lengths examined—if anything, the difference went in the other direction (although none of the differences were tested statistically). The mean neighborhood frequency of common words was higher than that of rare words for the two shortest lengths examined (two and three phoneme words).

The divergent pattern of results obtained for visual and auditory neighborhoods may be due in large part to differences in the methodology adopted in the two respective studies. Most importantly, the frequency ranges examined in the two studies were very different. The frequency ranges of rare words in the two studies did not overlap and the set of common words included by Pisoni et al. (1985) represented only a small subset of the words analyzed by L & S. If the properties of similarity neighborhoods of frequent and infrequent words do depend on the specific frequency ranges adopted, then some serious doubts about the generality and robustness of the original L & S findings must be raised.

The first objective of this paper is test the validity of the L & S findings concerning the relation between a written word's frequency and its similarity neighborhood. We first replicate the L & S study with the CELEX database. To deal with the limitations of their study, we then conduct a more complete and systematic analysis of the similarity neighborhoods of English written words of different lengths across the entire frequency distribution. This analysis provides a test of the generality of the L & S findings which were restricted to a single length and two frequency ranges.

Our second objective is to establish whether the neighborhood properties of written and spoken words actually differ as the Pisoni et al. results suggest. We present an analysis of spoken words, repeating the procedure used by these authors. Later we compare the similarity neighborhoods of English phonological forms of different lengths across the entire frequency range using the same procedure as for orthographic neighborhoods.

Under Similarity Neighborhoods of Dutch Words we examine the cross-linguistic generality of the findings concerning the similarity neighborhoods of English spoken and written words, turning our attention to another language in the CELEX database, Dutch. An analysis of the similar-

ity neighborhoods of orthographic and phonological forms in Dutch allows us to compare the two languages.

The goal of Probabilistic Aspects of Lexical Density is to get a better picture of the similarity neighborhoods of words the listener or reader normally encounters. To do so we move from the type-based analyses adopted in the previous sections to a token-based analysis. This allows us to compute the probability that a listener will be confronted with words with a particular similarity neighborhood and, more generally, to assess the probability of identification of words in the lexicon. Furthermore, we present a unifying analysis in which the relation between word frequency, number of neighbors, neighborhood frequency and bigram frequency are examined in terms of a Markov model. The implications of these findings for lexical processing are discussed under the General Discussion.

The analyses reported here all exploited the CELEX lexical databases (Baayen, 1991b; Burnage, 1988) for English and Dutch. Here is a short description of each database.

English: The English database (V1.0) is drawn from several different sources. It constitutes the overlap of the Ascot version of the Longman dictionary of Contemporary English and the Oxford Advanced Learner's Dictionary. It includes 80,531 word forms with 29,967 lemmas. The token frequencies are those of the Collins/Cobuild frequency count of the 17,979,343-word corpus of current English compiled at the University of Birmingham over the past few years.

Dutch: The Dutch database (V3.1) contains 124,136 Dutch lemmas and 381,292 word forms. Its frequency counts are derived from the 42,380,000-word token corpus of the Institute for Dutch Lexicology in Leiden. This corpus was taken from 835 different contemporary texts. Note that, since compounds in Dutch are written without intervening blanks, the number of lemmas in the Dutch database is much larger than the number of lemmas in the English database.

Similarity Neighborhoods of English Words

In this section, we examine the neighborhood density and frequency of English written and spoken words.

L & S Replication for Orthographic Neighborhoods

Method

The original L & S study was based on the 1,000,000-token frequency counts of Kučera and Francis (1967). The CELEX frequency counts, in contrast, are based on 18,000,000 English tokens. In order to make a more direct comparison with the L & S study possible, 1,000,000 tokens were sampled (without replacement) from the 18,000,000 tokens of the English corpus.[1] For each type in this smaller sample, the token frequency was obtained. The resulting scaled-down English database was used to replicate the L & S results.

All the four letter words falling into the common (frequency $\geq$ 76/1,000,000) and rare (frequency = 1/1,000,000) frequency classes used by L & S were selected from the reduced database along with their frequency. This database produced 278 rare words and 285 common words. It should be noted that the size of the resulting frequency classes corresponds closely to that obtained by L & S (260 rare and 260 common words).

The procedure used was essentially the same as that employed in the L & S study.

[1] The reduction procedure used here is virtually identical to Muller's (Muller, 1977) reduction method. The process of sampling without replacement can be viewed as an attempt to construct a smaller corpus on the basis of a random selection of the texts which constitute the original corpus. There is some evidence (see e.g., Brunet (1978)) that these methods tend to overestimate the number of types in the reduced sample. This is undoubtedly due to the fact that the method builds on the incorrect assumption that words occur independently in texts. Nevertheless, this is the most reliable method available for scaling down corpora (see e.g., Khmaladze & Chitashvili, 1989).

The N-count neighborhood definition was used. Thus, all words that differed from a given word in a single letter in any position were counted as neighbors. However, all the words in the two frequency ranges were analyzed, unlike the L & S study which only used a random sample of 50 words from each class. For each word in the two frequency classes, the number of neighbors and the mean of the frequencies of these neighbors were computed. From these values both the mean and median number of neighbors and the mean and median of the (mean) neighborhood frequencies were computed for the two frequency classes.

TABLE 1
MEAN AND MEDIAN NEIGHBORHOOD DENSITY AND FREQUENCY FOR ENGLISH FOUR-LETTER WORDS IN THE ORIGINAL AND REPLICATION STUDIES

	L&S analysis		Our analysis	
	Rare	Common	Rare	Common
	Number of neighborhood			
Mean	4.85	8.64	5.96	7.63
Median	4	9	5	8
	Frequency of neighborhood			
Mean	100.36	116.21	93.24	163.33
Median	20.21	55.12	19.30	69.78
Number of words	50	50	265	285

Results

The results of our replication study with the CELEX database are displayed in Table 1. We can see that common and rare words do differ in their similarity neighborhoods as the original study suggested. Common four-letter words have more neighbors than do rare words. This difference between the two word classes was significant by the Mann–Whitney U test ($Z = 4.26$, $p < .001$). Further, the neighbors of common words were more frequent than those of rare words.[2]

Discussion

While this analysis provides a useful replication, it necessarily suffers from the same defects as the original study. By restricting the analysis to two frequency ranges, we cannot assess the overall relation between word frequency and the properties of the similarity neighborhood. Furthermore, the choice of a broad frequency range (≥75) for common words makes it impossible to determine how neighborhood density varies within this large class. To determine more systematically how neighborhood density and neighborhood frequency vary with word frequency, we conducted a neighborhood analysis for the full frequency distribution. By comparing the properties of the similarity neighborhoods across the complete frequency range, we hoped to get a more precise picture of the relationship between these important variables. In addition, we decided to study words of different length (three to eight letters) to investigate how the relation between word frequency and neighborhood properties varied with word length.

[2] If, like L & S, we use the same nonparametric test we find that this difference is again highly significant ($Z = 8.48$, $p < .001$). Unfortunately, however, the Mann–Whitney U test is not appropriate here. The problem is that the test presupposes that the observations for the dependent variable (neighborhood frequency) are independent of the observations for the independent variable (word frequency). This condition is not met: the frequency of some type X may appear both in the independent variable (as word frequency) and in the dependent variable (as the neighbor frequency of other words Y, Z, . . .).

English Orthographic Neighborhood Analysis

Method

All English word forms, ranging in length from three to eight letters, were selected from the CELEX database for English. These words were of diverse morphological structure: base, inflected, and derived forms. For each word the number of neighbors and the mean of the logarithmic transform of the frequencies of these neighbors were calculated. The frequencies were based on an 18,000,000 English token count. Neighbors were again defined as those words that differed from the target by

TABLE 2
FREQUENCY CLASSIFICATION BASED ON MARTIN (1983)

Class	Characterization	Definition	Cobuild frequency ranges
1	Very frequent/common	$f \geq$ 1:10,000	$f \geq 1800$
2	Frequent/common	1:100,000 $\leq f <$ 1:10,000	$180 \leq f \leq 1800$
3	Upper neutral	1:500,000 $\leq f <$ 1:100,000	$36 \leq f < 180$
4	Lower neutral	1:1,000,000 $\leq f <$ 1:500,000	$18 \leq f < 36$
5	Rare	1:10,000,000 $\leq f <$ 1:1,000,000	$2 \leq f < 18$
6	Extremely rare	$f <$ 1:10,000,000	$f = 1$

only a one letter substitution in any position.[3] In order to gain insight into the relation between word frequency and the similarity neighborhood structure for the full frequency range, we assigned each word to one of the six frequency classes listed in Table 2. This frequency classification is based on that proposed by Martin (1983, 1988). Martin's frequency classification is especially useful here because it also takes into account some psycholinguistic results on the subjective perception of frequency differences (Shapiro, 1969), building on work by Carroll (1967, 1969) on the lognormal model for word frequency distributions. As such it is the best motivated frequency classification that has come to our attention. However, since the corpora underlying the present investigations are very much larger than the 1,000,000-word corpora for which Martin developed his classification, we have subdivided Martin's lower two frequency classes into upper neutral and lower neutral and rare and extremely rare, respectively. This allows us to study the lower frequency ranges in more detail.

Results

The analysis of the neighborhood properties of four-letter words will be presented first to allow a comparison with the replication study just reported. The results of the analysis of neighborhood density are summarized in Fig. 1. The results revealed a gradual increase in both the mean and the median number of neighbors as a function of word frequency. The lowest frequency class showed the lowest density values (median 5.0; mean 6.16) and the two upper frequency ranges showed the highest values (median: 9.0, 8.5; mean 9.15, 8.68), with the intermediate ranges falling somewhere in between (median: 6.0, 8.0, 8.0; and mean: 7.18, 7.88, 7.95). The boxplot also reveals a large amount of variance in the separate classes. Although the highest frequency class appears with the higher median (or mean) number of neighbors when compared with the lowest frequency class, many low-frequency words exist with a high number of neighbors and vice-versa. This suggests that the correlation between word frequency and neighborhood density is rather weak.

To gain further insight into how frequency and number of neighbors are correlated, we made use of the WARPing approximation to the Nadaraya–Watson estimate for nonparametric kernel regression smoothing (Haerdle, 1991, 123–143). This nonparametric regression technique was used in light of the decided nonnormality of the marginal distributions.[4] Figure 2 shows

[3] The logarithmic transform of the neighbor frequencies was used to minimize the effect of the skewness of the neighborhood frequency distributions. Without this correction the mean neighborhood frequency would be determined almost completely by the high frequency outliers, whereas we were interested primarily in the distributional pattern of the whole neighborhood.

[4] Using the χ^2 test to evaluate the goodness-of-fit of the normal distribution it was found that the word frequency, logarithmically transformed, is highly unlikely to be normally distributed ($\chi^2 = 76.78$, $df = 22$, $p < .0001$). The same holds for the number of neighbors ($\chi^2 = 193\ 02$, $df = 23$, $p < .0001$).

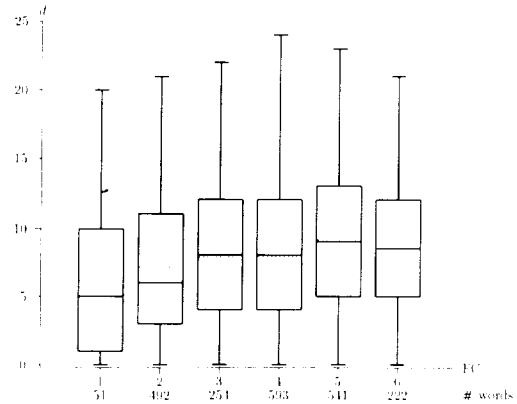

FIG. 1. Boxplots showing the number of neighbors d for six frequency classes (FC) of Table 2 for English four-letter word forms. The plot lists the median, the upper and lower quartiles, and the most extreme data points within 1.5 of the IQR of the upper and lower quartiles by means of horizontal lines. Outliers are represented by dots. The arithmetic mean is represented by a dotted line.

the scatterplot and the regression curve corresponding to Fig. 1, with the independent variable, word frequency, on the horizontal axis and the dependent variable, number of neighbors, on the vertical axis.[5] The word frequencies were transformed logarithmically to reduce the effect of outliers.

Figure 2 reveals the same trend as Fig. 1, although the downward curvature for the higher frequency word appears to be more pronounced than suggested by the slight lowering of mean and median observable for the highest frequency class in Fig. 1. Evidently, the positive correlation reported by L & S for frequency and number of neighbors is reversed for the highest-frequency types. Because L & S compared means and medians for extreme frequency classes (logarithmically transformed: [0.0; 2.89] and [7.22, ∞]), they failed to observe this downward trend for their class of common words, probably because the majority

[5] The regression curve of Fig. 2 was obtained using an Epanechnikov Kernel. For bin width $\delta = .5$ the optimal cross-validation score was obtained for window width $h = 2.5$ (see Haerdle, 1991, pp. 151–171). This is the window width underlying Fig. 2.

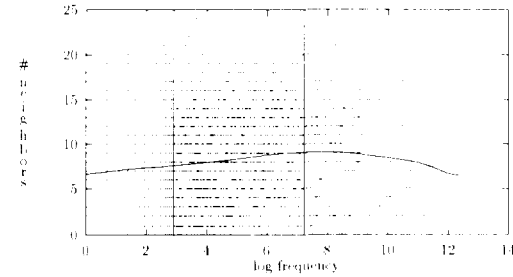

FIG. 2. Neighborhood frequency nonparametric regression curves for Dutch orthographic word forms of lengths 3–8.

of their common types appeared at the lower end of this frequency range (see Fig. 2).

We can test whether the main trend for the number of neighbors to increase with type frequency is significant using the Spearman rank correlation coefficient r_s. Interestingly, word frequency and neighborhood density emerge as being significantly correlated ($r_s = .16$, which is statistically highly significant ($p < .0001$) for 2151 degrees of freedom). However, the huge scatter shown in Fig. 2 considerably reduces the potential importance of the neighborhood density effect. In fact, the amount of variance explained for the rank-based transform of the data is only $r_s^2 = .026$.[6] We are forced to conclude that (i) the neighborhood density effect reported by L & S, even though it is significantly present, is extremely weak, and (ii) that the positive correlation is reversed for the 300-odd highest-frequency types.

[6] It is also possible to use the Pearson correlation coefficient r. For the present data r equals .1370. The estimated slope of the associated regression line is .026. Note that the amount of variation explained by the linear regression of neighborhood density on (log) target frequency is even less ($r^2 = .019$) than in the case of the rank-coded transform of the data ($r_s^2 = .026$). Also note that the slope of the linear regression line suggests a weaker neighborhood density effect (an increase from $E[Y|X = 1] = 7.98$ to only $E[Y|X = 12] = 8.30$) than the nonlinear effect which emerges from Fig. 2. However, due to the nonnormality of both marginal distributions, we cannot use r to ascertain whether the observed correlation is significant. Hence we have opted for carrying out our analyses in terms of r_s.

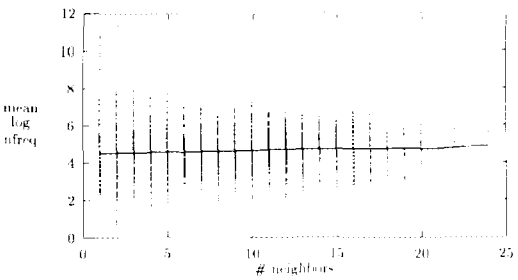

FIG. 3. Scatterplot and corresponding nonparametric regression curve for the mean log neighbor frequency as a function of the number of neighbors for English four-letter word forms.

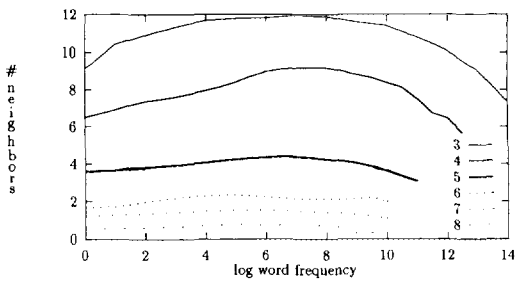

FIG. 4. Neighborhood density nonparametric regression curves for English orthographic word forms of length 3–8.

The neighborhood frequency effects were analyzed in a slightly different way than was done in the L & S study. To avoid the problem of nonindependence between word frequency and (mean) neighborhood frequency (see footnote 2), we focused on the relation between the number of neighbors of words and the mean of the logarithmically transformed frequencies of their neighbors.

Figure 3 presents the scatterplot and the corresponding nonparametric regression line for English orthographical word forms of length 4.[7] Figure 3 suggests that neighborhood frequency increases slightly as a function of the number of neighbors ($E[Y|X = 1] = 4.50$, $E[Y|X = 24] = 4.89$). As in the case of a neighborhood density effect, we can test whether this slight increase is significant. The Spearman rank-order correlation coefficient for these neighborhood frequency data equals $r_s = .115$ ($df = 2067$, $p < .0001$). As before, we are dealing with a very weak but statistically significant correlation.[8]

We performed the same analyses of the similarity neighborhoods for words of other lengths (three, five, six, seven, and eight letters). The results of these analyses are summarized in Figs. 4 and 5 in terms of the nonparametric regression curves. Figure 4 shows that the neighborhood density effects were the strongest for words with three and four letters. The number of neighbors first increases with log word frequency and then decreases. Note that the curves for words with five or more letters are nearly flat, showing that there is no density effect here. Turning to the neighborhood frequency effect as plotted in Fig. 5, we find that only four-letter words evidence a small but steady increase in mean neighborhood frequency as the number of neighbors increases. Both the density and frequency effects become weaker as word length increases.

Table 3 summarizes the corresponding Spearman rank-order correlation coefficients. What we find is that there is no significant density correlation for word length 3, and that the density correlations for lengths 5–8 are so weak that their relevance becomes highly doubtful, even though they are statistically significant. The four-letter words show up with a significantly stronger

[7] The regression curve of Fig. 3 was obtained using an Epanechnikov Kernel. For bin width $\delta = 1$ the optimal cross-validation score was obtained for window width $h = 6$. This is the window width underlying Fig. 3.

[8] The Pearson correlation coefficient for these data equals .083, the amount of variance being explained by the regression analysis being .007. The estimated slope of the regression line is slightly less than that of the nonparametric regression line of Fig. 3 (.0035 versus .0172). Note that the heteroskedasticity in the data renders the interpretation of r somewhat problematic. Since both marginal distributions are decidedly nonnormal ($\chi^2_{(19)} = 149.16$, $p = .0001$ for the number of neighbors (hermits excluded), $\chi^2_{(7)} = 411.24$, $p = .0001$ for the mean log neighborhood frequency), again no significance testing can be carried out on the basis of r.

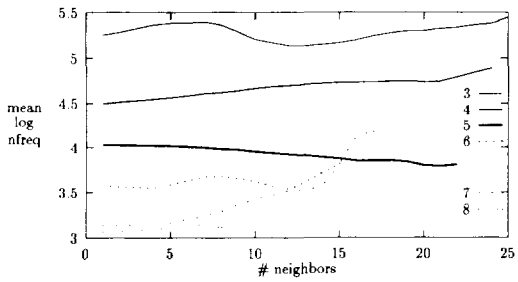

FIG. 5. Neighborhood frequency nonparametric regression curves for English orthographic word forms of lengths 3–8.

correlation than the six- and seven-letter words.[9]

The downward curvature at the right hand side of the regression curve of orthographic four-letter word forms requires further investigation. We examined whether this downward curvature is due to closed class words which dominate the highest frequency range. Figure 6 shows that the closed class words, represented by small circles, dominate the high frequency ranges and are responsible for the downward curvature of the regression line observable for words with log frequency exceeding 8. Although there are not too many open class types with higher frequencies, the regression line for open class words only, also plotted in Fig. 6, has roughly the same slope for the whole frequency range. Not surprisingly, the corresponding rank-order correlation coefficient is slightly higher (r_s = .2110).[10]

Turning to the neighborhood frequency effect, the only significant correlations at the .01 level appear for word lengths 4 and 7. Again the four-letter words show up with higher correlations than the three-, five-, and six-letter words.[11] Interestingly, the upward curvature of the regression line of words of length 7 is entirely due to the presence of large numbers of words with the extremely productive suffix *-ing*. Of all seven-letter words with more than five neighbors, 70% end in *-ing* and more than 90% of all types with more than eight neighbors similarly contain this suffix. In fact, when all words ending in *-ing*, which can only be neighbors among themselves, are removed from the set of seven-letter words, no significant neighborhood frequency effect remains (r_s = .0037, p = .093). The same holds for the neighborhood density effect for seven-letter words, where r_s assumes its lowest value: removal of targets ending in *-ing* lowers r_s from .046 to .026, which is no longer significant at the 1% level (p = .0347 instead of .0001). This suggests that the neighborhood effects found for seven-letter words are due to the presence of these effects for the four-letter verbs to which *-ing* attaches. Note that the only other significant neighborhood fre-

TABLE 3

SPEARMAN RANK-ORDER NEIGHBORHOOD DENSITY (WORD FREQUENCY AND NUMBER OF NEIGHBORS) AND NEIGHBORHOOD FREQUENCY (NUMBER OF NEIGHBORS AND MEAN NEIGHBORHOOD FREQUENCY) CORRELATION COEFFICIENTS, SIGNIFICANCE LEVELS p AND NUMBER OF OBSERVATIONS n FOR ENGLISH ORTHOGRAPHICAL WORD FORMS OF LENGTH 3–8

	Density			Frequency		
Length	r_s	p	n	r_s	p	n
3	0.052	0.056	554	−0.012	0.787	542
4	0.156	0.000	2153	0.115	0.000	2069
5	0.086	0.000	3936	−0.003	0.850	3343
6	0.078	0.000	5846	0.036	0.026	3920
7	0.046	0.000	7381	0.072	0.000	3996
8	0.085	0.000	7217	0.043	0.024	2767

[9] In general it is impossible to use r_s to test whether two samples have been drawn from the same population. However, it is possible to ascertain whether two τ correlation coefficients are significantly different, using Kendall's conservative significance test (Lienert, 1986). The difference in density correlations measured in terms of Kendall's τ for lengths 4 and 6 is minimally significant at the 10% level (Z = 1.426) minimally. For the lengths 4 and 7 it is significant at the 5% level.

[10] The contribution of the closed class items to the downward curvature for three- and five-letter words is weaker than for the four-letter words.

[11] The differences are again evaluated by means of Kendall's τ. The significant ($p < .10$) Z scores for the comparisons ((4,3), (4,5), and (4,6)) are 1.32, 2.33, and 2.40, respectively.

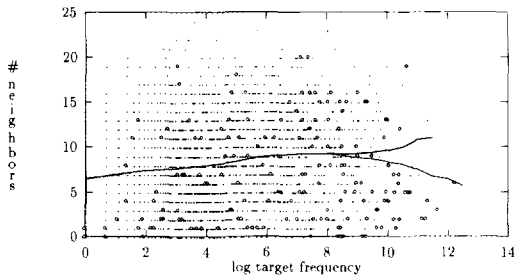

FIG. 6. English orthographical word forms, length 4: regression curves with and without closed-class words.

quency effect was found for precisely the four-letter word length.[12]

Finally we checked the effect of closed class items on the neighborhood frequency effect. A comparison of the neighborhood frequency effect with and without closed class items showed only a negligible and nonsignificant effect of this word class on the shape of the regression line.

Discussion

The preceding analyses of the similarity neighborhoods of all the English four-letter words in the CELEX database only partially confirm the pattern initially identified by L & S. Although neighborhood density increased with increasing word frequency, it decreased for the highest frequency words (some 9% of all types). The mean log frequency of the neighbors also increased as a function of word's neighborhood density. However, both correlations were very weak. Given our rank-coding of the data, word frequency explained less than 2.5% percent of the variance for both neighborhood density and frequency. The correlations are very much weaker or even nonsignificant for the other lengths examined, except for the cases in which the neighborhood effects depended on effects for shorter words via morphological structure. These analyses also revealed that lexical properties, other than word frequency, also influence the similarity neighborhood effects. For example, the highest frequency English closed-class words tend to have fewer neighbors than open-class words in the same frequency range.

These findings put the relation between a word's frequency and its similarity neighborhood into a clearer perspective. The results suggest that L & S were fortunate in their choice of both word length (four-letters) and the particular frequency ranges used. Had they examined other word lengths or other frequency ranges, they would not have found any effects of neighborhood density or frequency.

Next, we turn our attention to phonological neighborhoods. Our goal here is to replicate with the CELEX database the results obtained by Pisoni et al. (1985) for spoken words.

Pisoni et al. (1985) Replication of Phonological Neighborhoods

In undertaking an analysis of phonological neighborhoods, we must ask whether the N-count neighborhood definition adopted for orthographic neighborhoods is actually appropriate for spoken words.[13] It may be objected that spoken and written words do not have the same neighbors. Indeed, it is widely accepted that the neighbors of spoken words are defined sequen-

[12] Hence we predict no significant neighborhood frequency effect for eight-letter words in *-ing*. This is indeed what we find. Although r_s is somewhat higher (.0815), the correlation fails to reach significance at the 1% level ($p = .026$). Conversely, the significant neighborhood density correlation observed for five-letter words allows the eight-letter complex words ending in *-ing* to just reach significance ($r_s = .080$, $p = .0095$).

[13] It is also unclear how to best define lexical neighborhoods for written words. The N-count definition is an extremely crude measure that does not take into account a number of possibly important factors. For example, the degree and the position of mismatch between words do not play a role in this definition. Important properties such as the CV-structure and the letter or sound similarity between words are simply ignored. Moreover, under the N-count definition neighbors are matched for length. Fortunately, when we performed additional lexical statistical analyses for words with another neighborhood definition—counting words differing in length by one letter or phoneme also as neighbors—we did not obtain a substantially different pattern of results.

tially given the properties of the speech input. The most explicit sequential account of spoken word neighborhoods is provided by the cohort model (Marslen-Wilson, 1987) in which the neighbors or so-called cohort members are those words that share their initial part with the target word. A word is assumed to be recognized when it no longer has any neighbors. Thus, a target word's recognition point is assumed to correspond to the uniqueness point or the moment at which this word diverges from all other words in the lexicon.

Although there is some empirical support (Marslen-Wilson, 1984; Radeau & Morais, 1990) for the uniqueness point as a predictor of spoken word recognition performance, other definitions of neighbors have received some empirical support as well. Luce and his colleagues (Luce, Pisoni, & Goldinger, 1988) have taken the similarity neighborhoods to include all words that differ in a single phoneme in any position. They have conducted a number of experiments that support the predictions of the N-count definition. Thus, the problem of the proper neighborhood definition for spoken words remains unresolved. For the present purposes of comparing modalities and evaluating the Pisoni et al. study, we will continue to exploit the N-count definition.

Method

The neighborhood analysis in the Pisoni et al. study, like the L & S analysis, was based upon the one-million-token frequency count of Kučera and Francis (1967). Again, to allow a more direct comparison with this study, we used the frequency counts of the resized database as discussed above. All three- and four-phoneme words falling into the common (frequencies $\geq$ 1000/1,000,000) and rare (frequencies ranging between 10 and 30) frequency classes were selected from this database. The selection process produced 315 and 604 rare words and 55 and 16 common words that were three- and four-phonemes in length, respectively. The procedure used for computing neighborhood frequency and density was the same employed above.

Results

The results of the analysis are shown in Table 4 along with those of Pisoni et al. An analysis using the Mann–Whitney U test revealed that only the difference in neighborhood frequency between three phoneme common and rare words was significant (Z = 5.48).

Discussion

The present replication presents a pattern similar to the one obtained by Pisoni et al. However, our results contrast rather strikingly with those reported for written words both by L & S and by us here. For example, the neighborhood density and frequency effects obtained for four-letter words do not emerge for the phonological word forms of the same length. The most obvious explanation for this discrepancy is that different words went into the phonological and orthographic analyses. In English there is no regular correspondence be-

TABLE 4
MEAN AND MEDIAN NEIGHBORHOOD DENSITY AND FREQUENCY FOR ENGLISH THREE- AND FOUR-PHONEME WORDS IN THE ORIGINAL AND REPLICATION STUDIES

	Pisoni analysis		CELEX analysis	
	Length 3			
	Rare	Common	Rare	Common
Number of neighborhood				
Mean	22.64	19.97	16.67	16.36
Median	—	—	17	17
Frequency of neighborhood				
Mean	119.29	501.22	175.15	322.25
Median	—	—	83.37	263.61
Number of words	278	39	315	55
	Length 4			
	Rare	Common	Rare	Common
Number of neighborhood				
Mean	7.88	6.21	7.72	6.44
Median	—	—	7.0	6.0
Frequency of neighborhood				
Mean	69.35	69.91	37.18	41.71
Median	—	—	18.80	26.39
Number of words	441	14	604	16

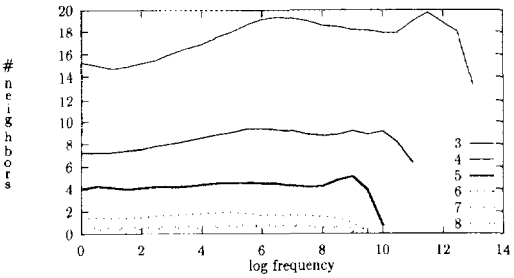

FIG. 7. Neighborhood density nonparametric regression curves for English phonological word forms, lengths 3–8.

tween either the identity of graphemes and phonemes (i.e., English is not orthographically shallow) or the number of symbols defining these two units (e.g., four-letter words often become three-phoneme words). Furthermore, very different frequency ranges were used in the two studies.

To sort out these differences and to compare more systematically the similarity neighborhoods of spoken and written words, we conducted analyses of phonological word forms of different lengths for the entire frequency range.

English Phonological Similarity Neighborhoods

Method

We followed exactly the same procedure as for written word forms, computing the neighborhood density and frequency of all English words ranging in length from three to eight phonemes.

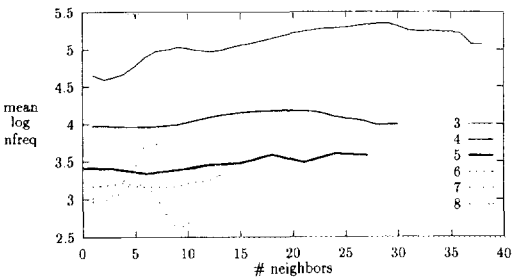

FIG. 8. Neighborhood frequency nonparametric regression curves for English phonological word forms, lengths 3–8.

TABLE 5
SPEARMAN RANK-ORDER NEIGHBORHOOD DENSITY (WORD FREQUENCY AND NUMBER OF NEIGHBORS) AND NEIGHBORHOOD FREQUENCY (NUMBER OF NEIGHBORS AND MEAN NEIGHBORHOOD FREQUENCY) CORRELATION COEFFICIENTS, SIGNIFICANCE LEVELS p AND NUMBER OF OBSERVATIONS n FOR ENGLISH PHONOLOGICAL WORD FORMS OF LENGTH 3–8

	Density			Frequency		
Length	r_s	p	n	r_s	p	n
3	0.169	0.000	1886	0.218	0.000	1882
4	0.129	0.000	4443	0.094	0.000	4244
5	0.037	0.002	7323	0.035	0.006	6132
6	0.087	0.000	7848	0.051	0.000	4888
7	0.097	0.000	6919	0.085	0.000	2919
8	0.138	0.000	5728	0.087	0.000	2015

Results

The results of the analyses across the different lengths for phonological neighborhood density and frequency are plotted in Figs. 7 and 8, respectively. The corresponding Spearman rank-order correlation coefficients are given in Table 5, together with the associated p-values. Table 5 shows weak but significant effects for both neighborhood density and frequency across the different lengths examined. The density effect is significantly stronger for four-phoneme words than for five-phoneme words but not for the other lengths, where the differences fail to reach significance with the conservative test used here.[14] The neighborhood frequency effect is stronger for three-phoneme words than for four-phoneme words and stronger for four-phoneme than for five-phoneme words.[15]

Discussion

This analysis of the English spoken words revealed a significant, but weak relation between a word's similarity neigh-

[14] According to Kendall's τ-based test of significance, the four- and five-phoneme words differ significantly at the 5% level.

[15] The differences between the τ-values corresponding with the (4,3) and (4,5) comparisons are significant at the 5 and 10% levels, respectively.

borhood and its frequency. For the shorter words (lengths 3–5), there is a gradual increase in neighborhood density with increasing word frequency but then a decrease for the highest-frequency words. The pattern emerging is roughly similar to that obtained for orthographic word forms: a gradual increase in density followed by a decrease. However, for spoken words this downward curvature tends to be steeper and its onset tends to begin at higher frequencies than for the orthographic regression curves. As for written words, the presence of high-frequency closed-class items contributes to the downward trend observed. For these shorter words, this decrease disappears when the closed-class items are eliminated.

The present analyses suggest that the discrepancy between the previously obtained results for spoken and written words is in large part due to the specific methodology used. The failure of Pisoni et al. to find effects of frequency upon neighborhood density can be understood by looking at Fig. 7. Their rare and common words were located in the frequency ranges in which lexical density rose (log frequency in the range 2.3–3.4) and fell (log frequency > 6.9), respectively. The problems associated with sampling in restricted frequency ranges again demonstrate the importance of examining the entire frequency range.

Similarity Neighborhoods of Dutch Words

To examine the generality of the findings for English, we repeated our analyses for the CELEX database of Dutch word forms.

Method

All Dutch word forms, ranging in length from three to eight letters or phonemes, were selected from the CELEX database for Dutch. Neighbors were again defined as those words that differed from the target word by only one character in any position. The number of neighbors and the mean log frequency of the neighbors was computed for each word. The nonparametric regression curves were estimated, and the Spearman rank-order correlations were calculated.

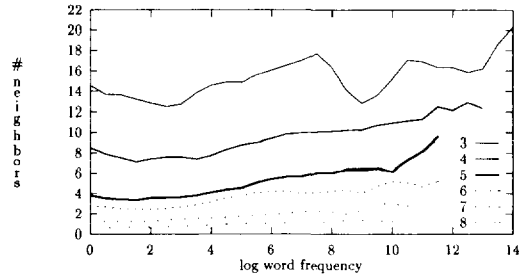

FIG. 9. Neighborhood density nonparametric regression curves for Dutch orthographic word forms of lengths 3–8.

Results

The nonparametric regression curves are shown in Figs. 9–12. The corresponding Spearman rank-order correlation coefficients appear in Table 6. The correlations for orthographical and phonological word forms were significant for all lengths for both orthography and phonology. The shorter word lengths, especially the lengths 4 and 5, showed up with the larger values of r_S.[16] For these lengths, the correlations for the phonological word forms were slightly higher than those for the corresponding orthographical word forms, suggesting that the effects were stronger in the phonological word forms. The differences failed to reach significance by the highly conservative test used here, however.

Discussion

The results of this analysis revealed a weak but significant correlation between word frequency and neighborhood density and neighborhood frequency in Dutch. These effects appear to be stronger than those in English, where some correlations

[16] Measured in terms of differences in Kendall's τ, four-phoneme words show up with significantly stronger effects ($p < .05$) than the longer words. For orthographic words, the differences between four-letter words and the lengths 7 and 8 reach significance ($p < .10$).

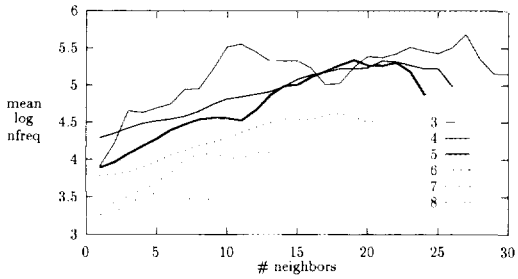

FIG. 10. Neighborhood frequency nonparametric regression curves for Dutch orthographic word forms of lengths 3–8.

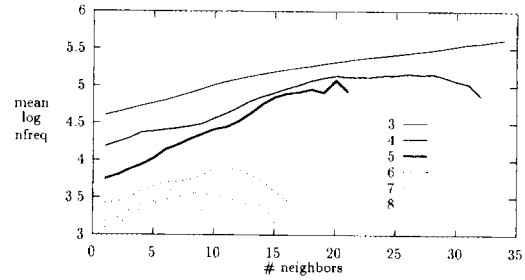

FIG. 12. Neighborhood frequency nonparametric regression curves for Dutch phonological word forms of lengths 3–8.

failed to reach significance. Some caution is required in making this cross-language comparison, however, as the corpora underlying the present analyses are different in size (18 million for English, 42 million for Dutch). Since the Dutch corpus contained more types than the English corpus, types can have more neighbors. Similarly, the word frequency range is greater, perhaps allowing the density effects to emerge somewhat more clearly.

One important difference between English and Dutch is the virtual absence in Dutch of an inverse relation between target frequency and neighborhood density for the highest frequency types. We observed earlier that this trend in English was caused by the closed-class items. In Dutch, closed-class items do not behave differently with respect to the neighborhood effects than open-class words. In fact, the regression lines are virtually identical for open- and closed-class items combined on the one hand and only open-class words on the other, the only difference being that the regression line of open-class items is slightly shorter. Interestingly, the neighborhood frequency effect for English orthographical word forms of length 4, with the closed-class items excluded, is of the same order of magnitude as for the Dutch orthographical word forms, closed-class items included (r_s = .21 (E), r_s = .20 (D)). It appears that the difference in the strength of the effects between the two languages is in part due to

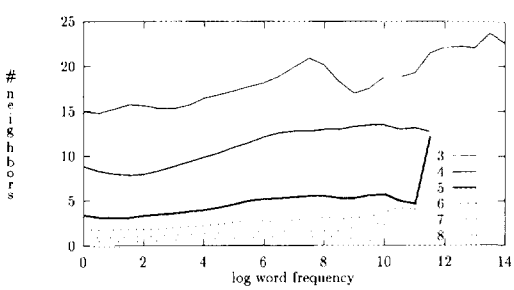

FIG. 11. Neighborhood density nonparametric regression curves for Dutch phonological word forms of lengths 3–8.

TABLE 6
SPEARMAN RANK-ORDER NEIGHBORHOOD DENSITY (WORD FREQUENCY AND NUMBER OF NEIGHBORS) AND NEIGHBORHOOD FREQUENCY (NUMBER OF NEIGHBORS AND MEAN NEIGHBORHOOD FREQUENCY) CORRELATION COEFFICIENTS, SIGNIFICANCE LEVELS p AND NUMBER OF OBSERVATIONS n FOR DUTCH ORTHOGRAPHICAL AND PHONOLOGICAL WORD FORMS OF LENGTH 3–8

	Density			Frequency		
Length	r_s	p	n	r_s	p	n
			Orthography			
3	0.161	0.000	685	0.199	0.000	675
4	0.204	0.000	2132	0.262	0.000	2032
5	0.205	0.000	3818	0.239	0.000	6432
6	0.155	0.000	7182	0.189	0.000	5265
7	0.114	0.000	10602	0.154	0.000	6046
8	0.121	0.000	13546	0.109	0.000	5569
			Phonology			
3	0.215	0.000	1374	0.279	0.000	1362
4	0.266	0.000	3620	0.321	0.000	3424
5	0.196	0.000	6488	0.243	0.000	5382
6	0.137	0.000	11313	0.123	0.000	7578
7	0.141	0.000	15058	0.125	0.000	7442
8	0.139	0.000	17084	0.109	0.000	5874

closed-class items. Note that the differential properties of closed class items across the two languages show that it is inappropriate to draw general conclusions about the possible functional role of the low density of closed-class words from one language (English) to another (Dutch). Another interesting cross-linguistic difference can be found in comparing the regression lines for orthographic and phonological forms in the two languages. The regression lines for lexical density for Dutch phonological and orthographic forms are more similar to each other than those for English. This is caused by the closer correspondence between orthography and phonology in Dutch.

It is instructive on the basis of the preceding analyses to reconsider the claims of L & S. These authors used their results to argue against the perceptual equivalence hypothesis and to suggest that the observed difference in the similarity neighborhoods of common and rare words constitutes a potential processing paradox. The tendency for higher frequency words to have more higher frequency neighbors suggests that, contrary to fact, high-frequency words should be harder to identify than low-frequency words. They resolved this paradox by showing that common and rare words are not equivalent on a second dimension. The frequency distributions of phonemes and letters in common and rare words were found not to be equivalent; certain phonemes were found to occur more frequently in rare words or in common words. They also reported the results of a perceptual identification experiment involving two classes of test stimuli that were matched in word frequency but made up of phonemes typically found in common and rare words, respectively. The words made up of phonemes primarily found in high-frequency words were identified more easily in noise than those words made up of low-frequency phonemes. They took the word frequency effect to be consistent with their phoneme distribution data, but inconsistent with the structure of similarity neighborhoods.

The results reported here weaken the empirical foundation for the proposed paradox for lexical processing considerably. Neighborhood density and neighborhood frequency increased only slightly across the entire frequency range. Moreover, the variance was huge; a large number of low-frequency types showed up with as many neighbors as the high-frequency types. At the same time, these neighborhood effects did not generalize to words of all lengths. Thus it is doubtful whether the similarity relations in the lexicon really give rise to a processing paradox. Moreover, the relations between phoneme frequency, word frequency, and neighborhood density and frequency are closely interrelated, as we will demonstrate below.

Probabilistic Aspects of Lexical Density

Having characterized the neighborhood density and frequency effects in English and Dutch in a type-based analysis, we now turn to consider the consequences of a token-based approach. Then we will consider what factors give rise to the observed lexical density effects.

Types and Tokens

A primary objective of statistical analyses of large lexical databases such as those presented here is to constrain theories of lexical processing and representation. It is important to note that—although informative—the preceding analyses of lexical neighborhoods are not optimally suited for evaluating the processing consequences of lexical similarity relations in the Dutch and English lexica. This is because these analyses are type-based. Type-based analyses consider each word to be equiprobable. More realistic from a processing point of view is a token-based analysis in which lexical entry has a probability of being encountered and processed that is proportional to its token frequency. Since higher-

frequency words have significantly more neighbors than low-frequency words, we can expect the probability of having to process word types with more neighbors to be underestimated in the type-based analysis.

Figure 13 contrasts the estimated type-based and token-based probability density functions of Dutch orthographic word forms of length 4. Note that given a particular interval on the horizontal axis, the area under the curve represents the probability of encountering words with a number of neighbors falling in the specified range. The total area under each curve equals unity. The highly skewed density function for the type-based analysis would appear to suggest that the lexicon is so composed as to avoid words with very many neighbors; roughly half the types have less than 10 neighbors. However, the token-based probability density function is moved to the right with respect to the type-based distribution such that some 20% of the probability mass is shifted from the lowest numbers of neighbors to the higher number of neighbors. This shift is a direct consequence of the neighborhood density effect. It raises the question why lexical usage tends to favor precisely those types that occupy densely populated neighborhoods, if an increasing number of neighbors is indeed detrimental to processing efficiency. Perhaps these facts suggest that there are no processing costs associated with the presence of neighbors.

Luce (1986) and Luce, Pisoni, and Goldinger (1988) have developed an alternative method of estimating the processing cost of neighbors which simultaneously takes into account both neighborhood density and neighborhood frequency. They assumed that each lexical competitor has a probability of being activated that depends on its frequency. They defined the probability of identifying the target in its neighborhood in terms of the choice rule of Luce (1959)

$$p(i) = \frac{f_i}{f_i + \sum_j f_{j(i)}}, \qquad (1)$$

where f_i is the frequency of the target and $f_{j(i)}$ the frequency of its j^{th} neighbor. The relation between log word frequency and the probability of identification is shown in Fig. 14. What we see is that both in a type-based and in a token-based analysis higher-frequency words tend to have higher probabilities of identification (cf., rank effect, Frauenfelder, 1990). However, the concentration of types with medium frequency (range 2 to 8) and low probability of identification seen in Fig. 14 suggests that, on average, the probability of identification in the lexicon is quite low.

This pattern raises the more general question how often words with a high probability of identification are encountered in the course of language processing. To answer this question, we computed the probability of encountering a word with a par-

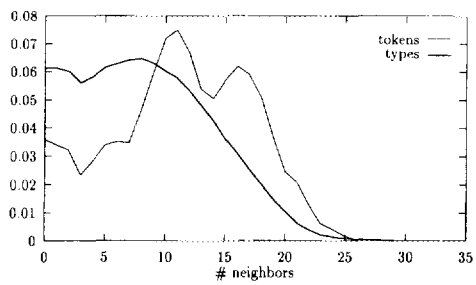

FIG. 13. Type- and token-based density estimates for neighborhood density (Dutch four-letter words).

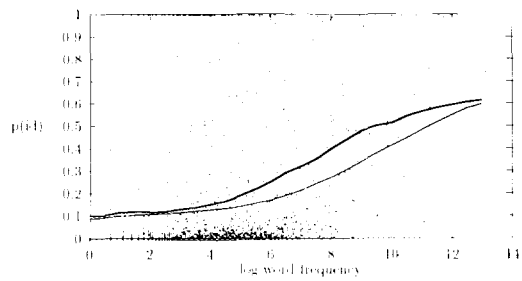

FIG. 14. Probability of identification $p(id)$ as a function of log word frequency for Dutch four-letter words. The type-based curve is represented by the lower line and the token-based curve by the upper line.

ticular probability of identification as shown in Fig. 15. The type-based analysis in which each type is assumed to occur equiprobably shows that the majority of types has a very low probability of identification ($\Pr(p(id) < .1) \approx .67$). However, when we take into account the fact that some words are more likely to be used than others, a substantially different and more uniform distribution is obtained. As shown in this figure words with lower probabilities of identification do not have the highest probabilities of being encountered in this token-based distribution ($\Pr(p(id) < .1) \approx .12$). The flat distribution suggests that there are many word tokens which are confusable with other words. A probability distribution suggesting even a considerably greater processing efficiency is obtained when density and word frequency are uncorrelated. This is represented in Fig. 15 by the "random" density function, estimated on the basis of 20 resampled distributions in which the empirical token frequencies were assigned at random to the types. In this hypothetical distribution, where no density effects are present, most word tokens have relatively high probabilities of identification. Assuming that Luce et al.'s analysis of the processing consequences of lexical density is correct, we are forced to conclude that the presence of density effects in the lexicon is not optimally adapted to lexical processing. Alternatively, we can conclude that lexical processing is not aversely affected by lexical competition.

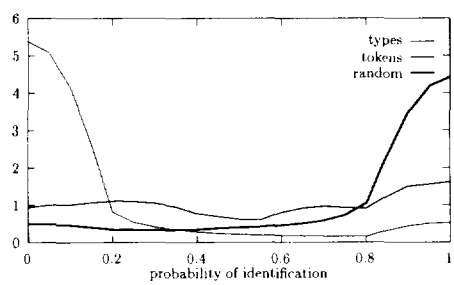

FIG. 15. Density estimation for the probability of identification according to Luce's choice rule for Dutch orthographic four-letter words.

Language Structure, Language Use, and Lexical Density

Here we consider how regularities in language structure contribute to the neighborhood relations observed in the lexicon. Next we examine the crucial role played by phonology (phonotaxis) and morphology (internal constituent structure of words) in creating lexical neighborhoods. Later we discuss a stochastic model that generates word frequency distributions with density effects that are qualitatively similar to the density effects in natural language. The model predicts that the number of neighbors is more closely correlated with diphone frequency than with word frequency. An examination of the English and Dutch data shows that this is indeed the case.

Phonology and Morphology

The contribution of phonological and morphological structure to creating neighborhood structure can be gauged by comparing the observed density for a linguistically structured lexicon with the expected density of a lexicon sampled at random from a linguistically unstructured set of "words." To estimate the mean number of neighbors for both structured and unstructured theoretical lexica, it is convenient to start with what we will call saturated lexica, lexica that contain all possible "words" of that length. In the baseline unstructured saturated lexicon, if there are k different phonemes, any string of length m will have $m(k - 1)$ neighbors. For an arbitrary selection of n word types from the complete set of k^m words of length m, the number of neighbors will on average be

$$\overline{N} = \frac{n}{k^m}[m(k - 1)]. \qquad (2)$$

Given 2, we can compute the expected mean number of neighbors ($\overline{N}$) for the unstructured Dutch saturated lexicon consisting of 3620 words (n) with four phonemes in length (m) made up from a set of 40 pho-

nemes ($k = 40$). Here n is fixed at 3620 in order to compare the theoretical mean density with the observed mean density of the 3620 Dutch four-phoneme words in our database. The resulting density of 0.22 neighbors is much less than the actual observed mean number of neighbors: 8.5. For longer words such differences become even more pronounced. For instance, we calculate a mean of 0.01 for an arbitrary selection of the 6488 five-phoneme words from the saturated lexicon, while the mean observed for the 6488 five-phoneme words in our database equals 4.07.[17]

To evaluate the density of a lexicon with phonotactic restrictions, we used a restricted saturated lexicon in which the only legal segmental structure for four phoneme words was a CVCV pattern. With 20 consonants (c) and 20 vowels (v) the average number of neighbors for a random selection of n words from this lexicon is given by

$$\frac{n}{c^2v^2}[2(c-1)+2(v-1)]. \qquad (3)$$

With $n = 3620$ the average number of neighbors is now found to be 1.72, which is substantially higher than 0.22. The phonotactic restrictions in force for languages such as Dutch and English are much more varied, of course, greatly reducing the number of legal strings in the language. Their effect is to restrict severely the number of possible strings in such a way that these strings are highly similar to each other.

Like phonotaxis, morphological structure also increases the density of similarity neighborhoods. For instance, the 3620 Dutch four-phoneme words contain a subset of roughly 1150 CVCV words with a word-final inflectional schwa. With a saturated lexicon that is similarly structured and without any assumptions about the phonotaxis of the 2470 remaining nonschwa final words, an average number of 2.6 neighbors is obtained for an arbitrary selection of 3620 words from the saturated lexicon. These calculations show how neighborhood relations emerge thanks to phonological and morphological structure. To this we should add that for the longer words the observed neighborhood relations are almost entirely due to morphological structure. This structure leads to a large numbers of types which are neighbors on the basis of their stems, as discussed above for the neighborhood frequency effect in English seven-letter words.

Neighborhood Effects and Bigram Frequencies

We have shown that phonological and morphological structure bring about the clustering of formally similar words in the lexicon. We now consider how this clustering interacts with the frequency properties of words and their neighbors. An important study that deals with the relation between word frequency and neighborhood density is Nusbaum (1985). Nusbaum studied the behavior of a zero-order[18] Markov approximation of English orthography, where the transitional probability into a given letter was defined as its relative frequency in English, independent of the preceding letter. What he found was that for fixed word lengths the neighborhood density effect critically depended on the letters having different relative frequencies. On the basis of this result he hypothesized that phonotactic rules play an important role in constraining the process by which words appear in language.

In what follows we will first clarify why the neighborhood density effect appears in more general first-order Markov approximations of natural language, and extend the

[17] Equation [2] shows why the mean number of neighbors decreases rapidly as words become longer. The numbers of observed types m for the lengths 4 to 8 are roughly of the same order of magnitude, differing by maximally a factor 5. On the other hand, the number of possible types increases by a factor k with each additional phoneme. Hence the outcome is a rapid decrease in the expected numbers of neighbors.

[18] We follow the convention that an nth order Markov process denotes a process that is conditioned on n preceding states of the system.

analysis to cover the neighborhood frequency effect. We will then discuss a particularly striking prediction of this approach, namely that diphone frequency and neighborhood density should be more closely correlated than word frequency and neighborhood density.

To see why the neighborhood density and frequency effects emerge in a first-order Markov approximation of the lexicon, consider the probability p_y of some word y with length m:

$$p_y = p_{oi_0} p_{i_0 i_1} \cdot \cdot \cdot p_{i_{m-2} i_{m-1}} p_{i_{m-1} o}. \quad (4)$$

Here p_{ij} is the probability of encountering the j^{th} element of the alphabet after having processed the i^{th} element. The zero represents the word boundary. The expected token frequency N_y of a word y in a sample of size N is known to be

$$E[N_y] = N_{p_y}. \quad (5)$$

Consider the conditions under which a target may obtain a high token frequency. A high token frequency requires that the probability p_y be high. Hence the transitional probabilities p_{ij} should be large. Since a neighbor of a high-frequency word y is obtained by replacing a single phoneme and hence two adjacent transitional probabilities, we find that for longer word lengths m, the remaining $m - 2$ relatively high transition probabilities will generally prevent the resulting neighbors from assuming probabilities that are very different from that of the original target word y. In other words, the probabilities of the neighbors of some target y will on average be similar to that of y itself, with the proviso that for extremely high or low p_y the likelihood of higher-frequency neighbors decreases and increases, respectively (regression toward the mean).

The consequences are twofold. If a target has a high probability, p_y, the probabilities of its neighbors will therefore on average be high too, hence the expected number of neighbors will be large. For target words with low probabilities, p_y, the reverse holds. This is how the neighborhood density effect arises in a first-order Markov approximation of language. Second, since high word probabilities give rise to high token frequencies by [5], we expect high-frequency words to have high-frequency neighbors. For low-frequency words the reverse holds: the few neighbors that are found will, on average, have low token frequencies. What we find, then, is that the neighborhood frequency effect is brought about by the same mechanism that underlies the neighborhood density effect.

Perhaps one of the most interesting predictions of this Markovian approach is that the correlation between word frequency and neighborhood density should, at least in part, be due to both word frequency and neighborhood density being correlated with diphone frequency. In other words, the model predicts that the correlation between neighborhood density and word frequency is substantially weaker when the correlation between word frequency and bigram frequency is factored out. Table 7 shows that this is in fact the case. This table lists the correlation coefficients obtained for a second-order Markov approximation of Dutch words of length 4 (see Baayen, 1991a), where a second-order Markov approximation was chosen in order to model the phonotactics of Dutch more precisely. Neighborhood density and target frequency are strongly correlated, but the correlation between neighborhood density y and di-

TABLE 7
CORRELATION RESULTS IN TERMS OF KENDALL'S τ FOR THE MARKOV MODEL (MM), THE HYBRID MODEL (HM), DUTCH PHONOLOGICAL WORD FORMS OF LENGTH 4 (DPL), AND ENGLISH PHONOLOGICAL WORD FORMS OF LENGTH 4 (EPL)

	MM	HM	DPL	EPL
τ_{xy}	0.525	0.227	0.182	0.090
τ_{xz}	0.507	0.245	0.213	0.160
τ_{yz}	0.604	0.622	0.457	0.325
$\tau_{xy \cdot z}$	0.298	0.079	0.088	0.038

Note. x, word frequency; y, number of neighbors; z, (geometric mean) diphone frequency.

phone frequency z is even stronger. When the correlation with diphone frequency is factored out, the neighborhood density effect emerges in a much reduced form, the partial correlation coefficient being only .298.

When we turn to our English and Dutch data sets, we find a highly similar state of affairs, as shown by the last two columns of Table 7. For both languages, the correlation between target frequency x and diphone frequency z is stronger than that between target frequency x and neighborhood density y, while the strongest correlation of all is that between neighborhood density y and diphone frequency z. Since the partial correlation coefficients $\tau_{xy \cdot z}$ are substantially smaller than the corresponding coefficients τ_{xy}, it is likely that the observed correlation between neighborhood density and target frequency is brought about mainly by diphone structure.

Interestingly, the empirical differences in the strengths of the correlations are underestimated by the Markov model. For instance, $\tau_{xy} = .87 \cdot \tau_{yz}$ in the Markov model whereas for English $\tau_{xy} = .28 \cdot \tau_{yz}$. This suggests that the Markov model is not quite correct, the density effects emerging in too strong a form. This problem is discussed in detail in Baayen (1991), who shows that a severe mismatch in the type–token ratio is characteristic of such Markov approximations. The solution proposed there is to combine a Markovian source for words with a model of lexical usage inspired by Simon's (1955) model for word frequency distributions. The addition of this second component serves a dual purpose. First, it forces words to be reused to a far greater extent than in the purely Markovian approach. This has the effect of correcting the type–token imbalance and of reducing the strengths of the density effects, as required. Second, it allows words to be assigned token frequencies that are not a direct function of their probability. This is a desirable property of lexical models, because word frequencies are also influenced by factors such as fashion and language change. The Markovian "front-end" of this hybrid model can be thought of as defining a probability distribution that reflects the relative ease with which words can be pronounced by the human vocal tract. The second component of the model can be thought of as simulating the random effects on word frequency of factors pertaining to language use.

As shown in Table 7, the correlations obtained with the hybrid model are more in line with empirical data, although the correlation between neighborhood density and diphone frequency is still too strong. This is undoubtedly due to the inability of this model to take the effects of morphological structure on lexical density and diphone sequencing into account. In spite of this defect, the partial correlation $\tau_{xy \cdot z}$ produced by the hybrid model is of the right order of magnitude, suggesting that the uncoupling of the strict link between word frequency and diphone structure is an important step toward the correct modeling of the way neighborhood structure, word frequency, and diphone frequencies are correlated in actual lexica.

To conclude, we have seen that the neighborhood density and frequency effects are brought about by differences in diphone frequencies. We return to the possible consequences of this result for lexical processing below.

General Discussion

This paper reports on the findings of several analyses of the similarity neighborhoods of words across different modalities (written and spoken) and languages (English and Dutch). We replicated the original findings of L & S for comparable restricted sets of four-letter words. An extensive analysis of all words with length ranging from three to eight letters and phonemes showed, however, that both the neighborhood density effect and the neighborhood

frequency effect are very weak. The variance is huge, and for longer words, these effects are either absent or too weak to be of interest. Nevertheless, we were able to observe some noteworthy differences between English and Dutch. In English, but not in Dutch, function words have relatively few neighbors given their high frequencies. Furthermore, the regression curves for the neighborhood density of orthographic and phonemic representations are more similar in Dutch than in English, due to the more shallow orthography of Dutch. In addition to type-based analyses, we also conducted token-based analyses, which produced rather different patterns of results. For instance, the density function obtained for a type-based analysis is skewed to the left, suggesting that most words will have relatively few neighbors. In a token-based analysis, however, a density function is obtained that is shifted towards the higher numbers of neighbors. Since it is the higher frequency words that tend to have the larger neighborhoods, listeners will encounter more words with many neighbors than a type-based analysis suggests.

In their study, L & S used the density effects they uncovered to argue against the perceptual equivalence hypothesis, according to which high- and low-frequency words do not differ from each other in any dimensions other than word frequency itself. They identified what they thought were two independent dimensions along which high- and low-frequency words differed. The first dimension concerns the properties of the similarity neighborhoods, the second pertains to the segmental make-up of the two word classes. However, we have seen that the two dimensions are causally linked. Common phonemes will generally be found in the more common words of the language. By extension, these common words will have the more common diphone transitions, hence they will appear in larger neighborhoods and have higher-frequency neighbors. The Markov model discussed under Language Structure, Language Use, and Lexical Density allows us to formulate a unitary explanation for the distributional phenomena discussed by L & S.

L & S also attempted to draw the implications of their findings for lexical processing, and, more specifically, for the word-frequency effect. They construed their phoneme distribution data as being consistent with the word-frequency effect and the structure of similarity neighborhoods as being at odds with it. However, the results we have presented here raise doubts about whether the lexical similarity relations are strong enough to attenuate the word-frequency effect. The increase with frequency of the number of neighbors is small, the variance on the other hand is huge, as the scatterplots (Fig. 2 and 3) show. Thus our analyses have demonstrated that the density structure of the lexicon cannot serve as evidence against the perceptual equivalence hypothesis. We conclude that it is hazardous to advance sweeping claims about lexical organization on the basis of restricted samples.

What are the consequences of our findings for theories of lexical organization and processing? We have argued that the similarity effects that have emerged in our more encompassing analyses arise due to phonotactic constraints on segmental sequences, constraints that appear to be imposed on language by the mechanical properties of the human vocal tract (Lindblom, 1983). This finding pushes the source of the weak but significantly present density effects into the domain of production. The question then arises how the resulting neighborhood structure of the lexicon affects word recognition. The obvious way to answer this question is to determine the effect of neighbors on word recognition experimentally. Unfortunately, no consensus on the role of neighbors has emerged in the experimental literature, despite considerable effort. In the absence of conclusive evidence for one

of the three logically possible types of neighborhood influence, no influence, facilitation, and inhibition, it is worthwhile to try to assess by means of lexical statistics what the global consequences of such processing assumptions are.

The main body of the present paper has, in fact, been concerned with the most interesting processing assumption, namely the one made by L & S that neighbors slow down word recognition. Their claim has received some experimental support both in the auditory and the visual domain. Luce (1986, 1988) argues that the influence of neighbors in auditory word recognition is best described by the choice rule 1. If we now ask the question whether the similarity structure of the lexicon is optimal for auditory word recognition, the answer is clearly negative. A comparison of the probabilities of identification of lexica with and without density effects showed that lexical organization would be better from a processing perspective if no density effects were present at all (see Fig. 15).[19]

Turning to the visual domain, Grainger (1993) and Jacobs and Grainger (1992) also found an inhibitory effect of neighbors. More specifically, they claim that the existence of a single substantially higher frequency neighbor is crucial for inhibition to take place. It can be shown that especially the lower-frequency words have a neighborhood structure that would slow down their recognition and that the number of such tokens is extremely small (less than 0.5%).[20] Thus Grainger's assumptions appear to imply that lexical processing is only minimally affected by neighborhood density and frequency. Unfortunately, lexical statistics alone cannot provide us with a means of evaluating the functionality of lexical structure given Grainger's model of word recognition, as they did for Luce's auditory word recognition model. The reason is that Grainger explains the effects he observes in terms of an interactive activation model for which the predictions cannot be translated into a simple formula. However, one may use his computational model to obtain recognition scores for every word in two contrasting lexicons: a lexicon in which words appear with their proper frequencies of use and a lexicon in which the empirical word frequencies are assigned at random. By comparing not individual recognition times but the resulting distributions of recognition times, the functionality of lexical structure may be gauged just as has been done above for Luce's approach. An analysis along these lines can be used to evaluate the consequences of the assumption of a positive effect of neighborhood size on word recognition (Andrews, 1989) as well. In fact, this methodology is more generally applicable to the study of the consequences of other properties of the lexicon for lexical processing like morphological structure (see Schreuder and Baayen, (1993)). We believe that language statistics are essential in that they allow us to evaluate processing models by confronting them with the language data which these models are designed to handle.

Finally, it is important to take into account that the lexicon is shaped by both perception and production factors and that it is impossible to evaluate consequences of lexical relations for the efficiency in perception without taking production into account and vice versa. In our study, we have attributed the similarity relations in the lex-

[19] Although Fig. 15 is based on orthographical representations, a highly similar pattern of results can be obtained for phonological representations.

[20] Tentatively operationalizing Grainger's claim, we defined a substantially higher-frequency neighbor as a word with a frequency that is higher by a factor 2 on the 10 log scale. For Dutch four-letter words, we counted 284 word types with exactly one such neighbor on a total of 2437 word types. The mean log frequency of the words supposedly suffering inhibition equals 3.13. The complementary set has a mean log frequency of 4.74 ($p < .001$, Welch Modified Two-Sample t test). For the range of word lengths studied here (3–8), a token-based analysis shows that 0.36%, less than four in one thousand, of the tokens are involved

icon to phonotactic constraints. Since these constraints are presumably motivated by articulatory factors, minor processing disadvantages in perception, if they exist (Coltheart, Davelaar, Jonasson, & Besner, 1977), may well be counterbalanced by greater efficiency in production.

REFERENCES

ANDREWS, S. (1989). Frequency and neighborhood size effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 802–814.

BAAYEN, R. H. (1991). A stochastic process for word frequency distributions. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (pp. 271–278). Berkeley, CA.

BAAYEN, R. H. (1991b). De celex lexicale database. *Forum der Letteren*, 32, 221–231.

BRUNET, E. (1978). *Le vocabulaire de Jean Giraudoux* (Vol. 1 of *TLQ*). Genève: Slatkine.

BURNAGE (1988). *CELEX: A guide for users*. Nijmegen: Centre for Lexical Information.

CARROLL, J. B. (1967). On sampling from a lognormal model of word frequency distribution. In H. Kučera & W. N. Francis (Eds.), *Computational analysis of present-day American English* (pp. 406–424). Providence: Brown Univ. Press.

CARROLL, J. B. (1969). A rationale for an asymptotic lognormal form of word frequency distributions. In *Research bulletin*. Princeton: Educational Testing Service.

COLTHEART, M., DAVELAAR, E. T., JONASSON, J., & BESNER, D. (1977). Access to the internal lexicon. In S. Dornick (Ed.), *Attention and performance* (Vol. VI). Hillsdale, NJ: Erlbaum.

FRAUENFELDER, U. H. (1990). Structure and computation in the human mental lexicon. In H. Haken (Ed.), *Synergetics of cognition*. Berlin: Springer-Verlag.

GARDNER, M. K., ROTHKOPF, E. Z., LAPAN, R., & LAFFERTY, T. (1987). The word frequency effect in lexical decision: Finding a frequency based component. *Memory and Cognition*, 15, 24–28.

GRAINGER, J. (1993). Orthographic neighborhoods and visual word recognition. In R. Frost & L. Katz (Eds.), *Orthography, Phonology, Morphology, & Meaning*. Amsterdam: Elsevier.

GRAINGER, J., & SEGUI, J. (1990). Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies. *Perception and Psychophysics*, 47, 191–198.

HAERDLE, W. (1991). *Smoothing techniques with implementation in S*. Berlin: Springer-Verlag.

HUMPHREYS, G. W., EVETT, L. J., QUINLAN, P. T., & BESNER, D. (1987). Orthographic priming: Qualitative differences between priming from identified and unidentified primes. In M. Coltheart (Ed.), *Attention and performance XII*. London: Erlbaum.

JACOBS, A. M., & GRAINGER, J. (1992). Testing a semi-stochastic variant of the interactive activation model in different word recognition experiments. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 1174–1188.

KHMALADZE, E. V., & CHITASHVILI, R. J. (1989). Statistical analysis of large number of rare events and related problems. In *Transactions of the Tbilisi Mathematical Institute* (Vol. 91).

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

LANDAUER, T. K., & STREETER, L. A. (1973). Structural differences between common and rare words: Failure or equivalence assumptions for theories of word recognition. *Journal of Learning and Verbal Behavior*, 12, 119–131.

LIENERT, G. A. (1986). *Verteilungsfreie Methoden in der Biostatistik* (Vol. 1). Königstein: Anton Hain.

LINDBLOM, B. (1983). Economy of speech gestures. In P. MacNeilage (Ed.), *The production of speech* (pp. 217–245). New York: Springer-Verlag.

LUCE, P. A. (1986). *Neighborhoods of words in the mental lexicon*. (Research on speech perception technical report 6.) Bloomington, IN: Indiana University.

LUCE, P. A., PISONI, D. B., & GOLDINGER, S. D. (1988). *Similarity neighborhoods of spoken words*. (Research on speech perception progress report 14.) Bloomington, IN: Indiana University.

LUCE, R. D. (1959). *Individual choice behavior*. New York: Wiley.

MARSLEN-WILSON, W. D. (1984). Function and process in spoken word recognition, a tutorial overview. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 125–150). Hillsdale: Erlbaum.

MARSLEN-WILSON, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25, 71–102.

MARSLEN-WILSON, W. D., & WELSH, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10, 29–63.

MARTIN, W. (1983). On the construction of a basic vocabulary. In S. Burton & D. Short (Eds.), *Proceedings of the 6th International Conference on Computers and the Humanities* (pp. 410–414). Comp. Science Press.

MARTIN, W. (1988). Lexical frequency. In K. van Reenen-Stein (Ed.), *Distributions spatiales et temporelles, constellations des manuscrits: Etudes de variation linguistique offertes à An-*

thonij Dees à l'occasion de son 60me anniversaire (pp. 139–152). Amsterdam: Benjamins.

MULLER, C. (1977). *Principes et méthodes de statistique lexicale*. Paris: Hachette.

NUSBAUM, H. C. (1985). *A stochastic account on the relationship between lexical density and word frequency*. (Research on speech perception, progress report 11.) Bloomington, IN: Indiana University.

PISONI, D. B., NUSBAUM, H. C., LUCE, P. A., & SLOWIACZEK, L. M. (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, **4**, 75–95.

RADEAU, M., & MORAIS, J. (1990). The uniqueness point effect in the shadowing of spoken words. *Speech Communication*, **9**, 155–164.

RUBENSTEIN, H., & POLLACK, I. (1963). Word predictability and intelligibility. *Journal of Verbal Learning and Verbal Behavior*, **2**, 147–158.

SCHREUDER, R., & BAAYEN, R. H. (1993). Prefix-stripping re-revisited. Submitted for publication.

SHAPIRO, B. J. (1969). The subjective estimation of word frequency. *Journal of Verbal Learning and Verbal Behavior*, **8**, 248–251.

SIMON, H. A. (1955). On a class of skew distribution functions. *Biometrika*, **42**, 435–440.

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

(Received July 15, 1992)
(Revision received March 31, 1993)