*The Neighborhood Characteristics of Malapropisms**

MICHAEL S. VITEVITCH

State University of New York at Buffalo

KEY WORDS

lexical density

malapropisms

speech production

ABSTRACT

This study examined the phonological neighborhood characteristics (frequency, density, and neighborhood frequency) of 138 malapropisms. Malapropisms are whole word substitutions that are phonologically, but not semantically, related. A statistical analysis of a speech error corpus suggests that neighborhood density and word frequency differentially affected the number of malapropisms. Specifically, a greater number of malapropisms were found among high frequency words with dense neighborhoods than with sparse neighborhoods. Exactly the opposite pattern was found among low frequency words. That is, more errors were found among low frequency words with sparse neighborhoods than with dense neighborhoods. More malapropisms resided in low frequency neighborhoods than in high. The average word frequency, average neighborhood density, and average neighborhood frequency of the malapropisms were significantly lower than the same averages computed from randomly sampled control words. Finally, more target words were replaced by error words that had relatively higher frequency than by error words that had relatively lower frequency. The implications of these findings for models of lexical representation and processing are discussed.

INTRODUCTION

Psycholinguists have attempted to use speech errors to answer many questions regarding the structure and organization of the lexicon. Speech errors are often classified into two main types: phonological errors and word substitution errors. Phonological speech errors are characterized by the addition, deletion, or transposition of one or more phonemes or phoneme clusters. Word substitution errors are characterized by a different word being used in place of an intended word. Within each type of speech error there are a number of subtypes.

Phonological speech errors and word substitution errors are thought to occur at distinct levels of processing. Word substitution errors are believed to occur during the recoding of a semantic representation into a phonological word form. Phonological errors are believed

* Acknowledgments: This research was supported (in part) by research grant number 1 R01 DC 0265801-A1 and T32 DC 00036 from the National Institute on Deafness and Other Communicative Disorders, National Institutes of Health. Thanks are due to Paul A. Luce, Trevor A. Harley, and Joerg D. Jescheniak, as well as the Slips of the Tongue Research Group at the University at Buffalo for helpful comments on an earlier version of this manuscript.

Address for correspondence: Michael S. Vitevitch, Speech Research Laboratory, Department of Psychology, Bloomington, Indiana 47405-1301. E-mail: mvitevit@indiana.edu

to occur during the recoding of a phonological word form into subphonological units (Garrett, 1988; Levelt, Schriefers, Vorberg, Meyer, Pechmann, & Havinga, 1991a; Mowrey & MacKay, 1990). Although these two main types of speech errors seem to operate at two different levels, they may both be influenced by the organization and structure of the lexicon (see Dell & O'Seaghdha, 1991; Levelt, Schriefers, Vorberg, Meyer, Pechmann, & Havinga, 1991a, 1991b for an extended discussion on this issue).

A great deal of work has led to suggestions about the locus of phonological speech errors (see Fromkin, 1973, and Levelt, 1992, for brief reviews of the literature and a summary of the issues related to phonological speech errors). However, compared to the work on phonological speech errors, relatively few psycholinguistic studies have examined the whole word speech error known as a *malapropism* (see Fromkin, 1973; Nooteboom, 1969; and Tweney, Tkacz, & Zaruba, 1975; for some early exceptions). Fay and Cutler (1977) examined several linguistic factors in malapropisms. They defined a malapropism as a real word that erroneously intrudes on an intended, or target, word. Target and error words are not semantically related but share a close relationship in their pronunciations.

It should be noted that these word substitutions are true speech errors. The producers of a malapropism know the correct meaning and usage of both the target and the error words (unlike Sheridan's (1775) character Mrs. Malaprop from *The Rivals*). This competence is often attested by the spontaneous correction of the utterance by the speaker. Thus, malapropisms are the result of a mis-selection made somewhere in the speech production system, and not the result of misusing a word.

Fay and Cutler's analysis of 183 malapropisms from a speech error corpus found that the target and error were in the same grammatical category 99% of the time. Additionally, the target and error had the same number of syllables 87% of the time. The target and error also shared the same stress pattern most of the time (98%).

They further analyzed 156 malapropisms that had the same stress pattern. Fay and Cutler believed that the distinctive feature hypothesis (Jakobson & Halle, 1956) and the nature of temporal processing would influence the degree of similarity between the target and error words. Specifically, they predicted that the target and error words would have similar features marked at the point the two words departed from being identical (starting from the beginning of the word).

Fay and Cutler performed a feature counting analysis on these words by looking at the transcriptions of the words and comparing the segments in which the transcriptions differed. Their predictions were supported by the data. In general, they found that many words had a small number of differences in the features at the point at which the two words diverged. Additionally, they found that few words had many different features at the point at which the two words diverged.

Fay and Cutler (1977) found many interesting results regarding the stress pattern, grammatical category, and number of syllables in their analysis of intended and error words. Other researchers (Harley & MacAndrew, 1992, 1995; del Viso, Igoa, & Garcia-Albea, 1991) have further examined a number of other characteristics of words that are involved in malapropisms. These characteristics include imageability, word length, and word frequency.

Harley and MacAndrew (1992, 1995) found that whole word speech errors that are phonologically related are unaffected by imageability. That is, the intruding word is as likely

to be of higher imageability as it is to be of lower imageability as measured by ratings from the Oxford Psycholinguistic Database (Quinlan, 1992). They also found that phonologically related speech errors tended to be longer words compared to semantically related speech errors or to "control" words.

Harley and MacAndrew (1992; 1995) also examined the influence that word frequency may have on producing word substitutions. Much previous work in speech *perception* has demonstrated that word frequency influences the accuracy and speed with which words are perceived (see for example Howes, 1957; Newbigging, 1961; Savin, 1963; Solomon & Postman, 1952). Work in speech production has also demonstrated that word frequency influences the speed with which a picture can be named (Oldfield & Wingfield, 1965). The role of word frequency in speech errors has only recently been demonstrated in phonological speech errors (Dell 1990; Stemberger & MacWhinney, 1986).

Harley and MacAndrew's analysis of whole word speech errors found that the target words of phonologically related whole word speech errors tended to be lower in frequency than the control items. In addition they found that target words were replaced by words of similar frequency, and that the targets of the phonological substitutions were of relatively low frequency overall. This set of findings regarding the influence of word frequency on whole word errors seems to contradict the findings of del Viso et al. (1991) on Spanish speech errors. Del Viso et al. (1991) examined 275 malapropisms in Spanish and found that the error word was more often higher in frequency than the target word, contrary to the null finding by Harley and MacAndrew (1992; 1995).

These discordant results may be better understood by examining the statistical measures that each researcher used. Del Viso et al. (1991) used a chi-square analysis to measure the likelihood that the percentage of words that went from a low frequency target to a higher frequency error was greater than the percentage of words that went from a high frequency target to a lower frequency error. While this analysis proved to be statistically significant, it only took into account the relational differences of the frequency of the targets and the errors.

Del Viso et al. (1991) did not determine if the differences in word frequency between the target words and the error words were statistically significant differences. The only information the chi-square analysis provided was that the percentage of error words higher in frequency than their corresponding target word was greater than the percentage of error words lower in frequency than their corresponding target word. On the other hand, Harley and MacAndrew (1992, 1995) employed a *t*-test in their analysis of an error corpus. This standard parametric test, however, fails to examine frequency differences in a relational manner. Harley and MacAndrew's (1992, 1995) analysis compared the variance around the means of the two frequency categories, but did not examine the proportion of items in each frequency category as del Viso et al.'s analysis did.

In addition to the different languages and the different error corpora being examined, the different methods of assessment used in each investigation may also have contributed to the contradictory results. These results also raise the question of how great the magnitude of the difference between lexical items must be in order for them to influence one another. That is, how different must two lexical items be for them to affect the language production system? Is the language production system affected by statistically significant differences

(as measured via conventional parametric statistics) or merely relative differences between lexical items?

The influence of word frequency on whole word errors can be contrasted with the influence of word frequency on phonological speech errors. Phonological speech errors are those errors in which one or more phonemes or phoneme clusters are added to, deleted from, or substituted for, an intended phoneme or phoneme cluster. Stemberger and MacWhinney (1986) and Dell (1990) found that high frequency words tend to be immune to *phonological* speech errors. That is, the constituent phonemes of low frequency words tend to be switched with phonemes of other words, dropped from the word, or added to another word more often than the phonemes of high frequency words.

A number of speech production models within the connectionist framework (Berg, 1988; Dell, 1986, 1988, 1990; Harley, 1984; Levelt, 1989; MacKay, 1982, 1987; Stemberger, 1985a) have been proposed to account for these findings in the phonological speech error literature. These interactive models suggest that activation spreads through multiple levels of representation. The numerous levels of representation and the activation processes that occur within and between levels of representation act as multiple sources for speech errors. A further test of these models would be to examine their ability to account for the data from other types of speech errors, such as malapropisms.

Although a number of characteristics of words (such as word frequency, word length, imageability of a word, number of syllables in a word, etc.) have been examined for the potential role they may play in speech production, one characteristic of words has not been extensively investigated in speech production—namely, the composition of phonological similarity neighborhoods. The work on the neighborhood activation model (NAM) summarized in Luce, Pisoni, and Goldinger (1990) suggests that the characteristics of the phonological similarity neighborhood of a word influence spoken word *recognition.* However, little work has examined the role of phonological similarity neighborhoods in speech production (see Goldinger & Summers (1989) for an exception).

The notion of similarity neighborhoods implies that the number and frequency of words that are similar to a target word can act to facilitate or impair the recognition of that target word (Luce, Pisoni, & Goldinger, 1990). Evidence from studies using processing times and accuracy rates as dependent measures have found that phonological similarity influences recognition in predictable and demonstrable ways (Goldinger, Luce, & Pisoni, 1989; Luce, 1986; Luce, Pisoni, & Goldinger, 1990).

One important factor is the density of the neighborhood, or the number of words that are similar to a particular word (see Figure 1 for an example of a dense and sparse neighborhood). The more words resembling a particular word, the denser its neighborhood. The predictions of NAM regarding density have been verified experimentally (Goldinger, Luce, & Pisoni, 1989; Luce, 1986; Luce, Pisoni, & Goldinger, 1990). The results show that words with dense neighborhoods require more time to be recognized than words with sparse neighborhoods. In addition, recognition accuracy is higher for words with sparse neighborhoods than for words with dense neighborhoods.

The frequency of the target word (i.e.,word frequency) and the frequency of the words in the neighborhood (i.e., neighborhood frequency) are other factors that influence word recognition. The predictions of NAM regarding the influence of frequency have also been

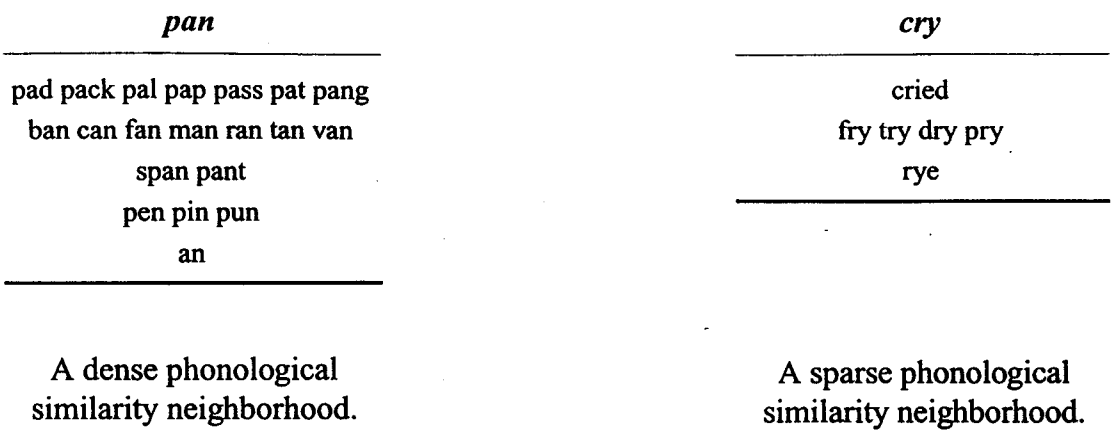

Figure 1

A graphic representation of a dense phonological neighborhood for the word "pan" (left panel) and a sparse phonological neighborhood for the word "cry" (right panel). (N.B. The words selected for this example are only meant to explicate the concept of neighborhood density and should not be thought of as an exhaustive listing of all the possible neighbors for those target words.)

verified experimentally (Goldinger, Luce, & Pisoni, 1989; Luce, 1986; Luce, Pisoni, & Goldinger, 1990). Like earlier studies of word recognition, these experiments show that a high frequency word will be recognized more quickly and more accurately than a low frequency word. Furthermore, a word with more frequent neighbors (i.e., a high neighborhood frequency) will be recognized more slowly and less accurately than a word with less frequent neighbors (i.e., a low neighborhood frequency).

Evidence suggests that word frequency affects both spoken word recognition and spoken word production in demonstrable ways. Do other factors that influence speech perception also influence speech production? More specifically, do the characteristics of phonological similarity neighborhoods — word frequency, neighborhood frequency, and neighborhood density — influence word production? To explore this possibility, the current analysis of a malapropism error corpus was conducted. The analyses were relational like the analysis used by del Viso et al. (1991), and nonrelational like the analysis used by Harley and MacAndrew (1992; 1995). The use of both methods of assessment on the same corpus may produce a more complete view of the influence of phonological similarity neighborhoods on word production.

Given that word frequency affects perception and production in similar ways, it seems plausible that the effects of similarity neighborhood characteristics will also be found in both domains. Because high density-high neighborhood frequency neighborhoods make perception difficult, it was predicted that more malapropisms should be found among low frequency words in such neighborhoods than among high frequency words with sparse neighborhoods and low neighborhood-frequency.

The converse prediction regarding neighborhood density and neighborhood frequency is equally plausible if one considers the intrinsic processes involved in speech perception

and speech production. During speech perception and word recognition, discrimination among the multiple candidates that have been activated must occur. During word production, conspiracies of representations may be essential to sufficiently activate the desired representation. From this perspective, the opposite outcome is predicted. That is, more malapropisms should be found among low frequency words in sparse neighborhoods than among high frequency words with dense neighborhoods. However, either hypothesis is acceptable, because the main goal of this investigation is to assess the possibility that phonological similarity neighborhoods affect production as well as perception.

METHOD

The malapropisms used in the statistical analyses were obtained from the appendix of Fay and Cutler (1977). As described in Fay and Cutler, these words came from a corpus compiled by David Fay that contained 2000 speech errors. From that collection of 2000 errors, Fay and Cutler omitted any errors that involved the addition, deletion, or substitution of a *single* phoneme, since these might be phonological errors. Also, malapropisms involving prepositions, adverbs, and compound nouns, were eliminated due to the higher incidence of disagreement on number of syllables. This left 183 phonologically related word substitution errors involving nouns, verbs, and adjectives. The 156 malapropisms that had the same stress pattern, number of syllables, and grammatical category, were selected for analysis to determine their neighborhood characteristics. In order to maximize the number of words that could be found in the computerized lexicon, all derivational and inflectional suffixes were stripped from the words.

Of the 156 root word pairs from Fay and Cutler (1977), 138 root word pairs were found in the online version of the approximately 20,000 word Webster's Pocket Dictionary. A pair was discarded from the analysis if one or both words were not found in the computerized lexicon. The neighborhood characteristics for these 138 words were then calculated. The computations to derive the neighborhood characteristics were performed on the phonetic transcriptions contained within the computerized lexicon.

There are several ways to compute similarity among lexical items. One way is to use Coltheart's N (Coltheart, Davelaar, Jonasson, & Besner, 1977), which changes a single letter to find similar lexical items. This metric, however, may underestimate the number of similar lexical items because it fails to include the lexical items produced when a letter is added or deleted from the target word. While Coltheart's N, a measure of orthographic similarity, is highly correlated with measures of phonological similarity (Harley & Brown, 1997), the *addition*, *deletion*, or *substitution* of a single *phoneme* is a more appropriate and more direct measure of similarity among phonologically similar items than is the *changing* of a single *orthographic* unit.

Another measure of similarity among lexical items is the cohort count (Marslen-Wilson & Welsh, 1978). The cohort measurement indicates the number of words that have the same phonemes, counting from the beginning of the word in a left-to-right manner, prior to the divergence point of the word. The divergence, or uniqueness, point is the point in which a word differs from all other words. While this measure is an improvement over the N-count in that it uses phonemes rather than orthographic units, the cohort count does not take into account the potential similarity that may occur among words after the divergence

point. An analysis of the Fay and Cutler corpus by Hurford (1981) revealed that malapropisms tended to resemble their targets at both extremities of the lexical item. Because the cohort count does not take into account the similarity among words after the divergence point, this metric may greatly underestimate the number of similar lexical items. For these reasons the neighborhood count of similarity, in which a single phoneme is added, deleted, or replaced, was used to assess similarity in the analysis of the targets and errors.

The neighborhood characteristics that were assessed included word frequency, neighborhood density, and neighborhood frequency. Word frequency is the frequency with which that word appears in the language. Because log frequencies tend to be better correlated with performance measures than raw frequencies, log frequencies were used in this analysis. The log frequency estimates were based on the Kučera and Francis (1967) word frequency counts. To make the present comparisons commensurate with those in the phonological speech error literature, values previously employed in that literature were used to partition lexical items into low and high frequency categories. Thus, a word was considered to be a low frequency word if its log frequency was below the log value of 35. A word with a log value above 35 was considered to be a high frequency word. This criterion was used in Experiment 3 of Stemberger and MacWhinney (1986). Separate analyses were also carried out on the sample using a criterion point of the log value of 85, as in Experiment 1 of Stemberger and MacWhinney (1986).

Neighborhood density is the number of words in the lexicon that can be produced by adding, deleting, or replacing a single phoneme in the target word. The median value of the density characteristic from the 138 selected targets was used to partition the target words into dense neighborhoods (i.e., the target has many similar words) and sparse neighborhoods (i.e.,the target has few similar words). The median neighborhood density value for the targets was a value of 1.

Neighborhood frequency is the mean log frequency of occurrence of all the words in the neighborhood of a target word. Like words, low neighborhoods had mean frequencies below log 35, or, in the alternate analysis, below log 85. Neighborhood frequencies above these values were classed as high.

Using these three factors, the 138 target words from the error corpus were partitioned into the appropriate cells (high vs. low frequency, sparse vs. dense neighborhood, high vs. low frequency neighborhoods). The distribution across conditions was then submitted to statistical analysis using a chi square. The values for the three neighborhood characteristics were also analyzed using analysis of variance. Thus, absolute and relational differences among and between targets and errors could be assessed.

RESULTS

A relational and nonrelational analysis of the error corpus obtained from Fay and Cutler (1977) was performed in order to examine the pattern of neighborhood characteristics in malapropisms. As seen in Figure 2 and Tables 1 and 2, there was an antagonistic interaction of word frequency and neighborhood density. A one-way chi square shows that this cross-over interaction is significant for both frequency criterion points; at log 35 $\chi^2(1,$

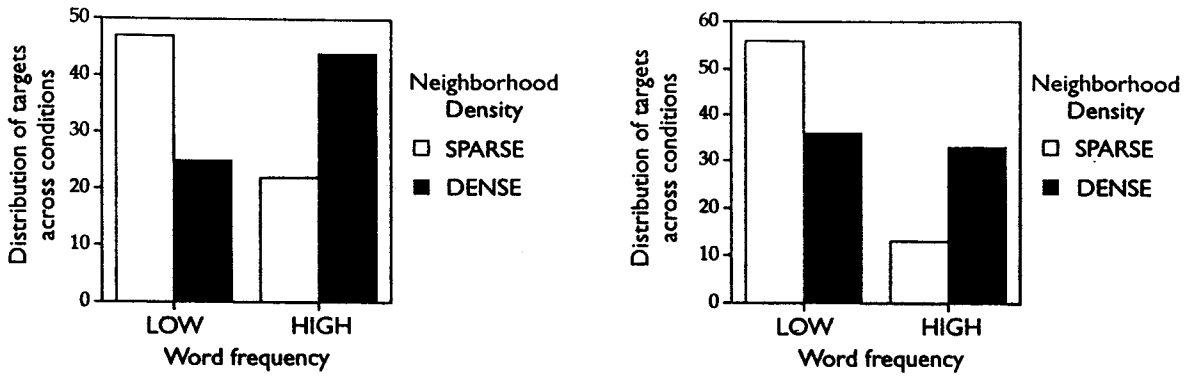

Figure 2

The number of errors found in the corpus of 138 malapropisms as a function of neighborhood density and target word frequency. The lefthand panel displays the results using the log value of 35 word frequency criterion, while the righthand panel displays the results using the log value of 85 word frequency criterion.

$n = 138) = 14.05, p < .001$; at log 85 $\chi^2(1, n = 138) = 13.04, p < .001$. Specifically, there were more errors in low frequency words with sparse neighborhoods than with dense neighborhoods. Conversely, there were more errors among high frequency words with dense neighborhoods than with sparse neighborhoods. This pattern was the same for both word frequency criterion points.

A one-way chi square was performed to see if the apparent overall difference between low frequency words and high frequency words was significant. With the log 35 criterion point, the difference between low frequency words and high frequency words was not significant, $\chi^2(1, n = 138) = .26, p > .10$. With the log 85 criterion point a one-way chi square shows that the difference between low and high frequency words is significant, $\chi^2(1, n = 138) = 15.33, p < .001$: there are more errors among low frequency targets than high frequency targets.

In addition, a one-way chi-square analysis of the distributions of neighborhood-frequency values shows significantly more target words came from low frequency neighborhoods than from high frequency neighborhoods. With the log 35 criterion point for neighborhood frequency, 135 targets had low frequency neighborhoods and three targets had high frequency neighborhoods, $\chi^2(1, n = 138) = 126.26, p < .001$. With the log 85 criterion point, all the targets had low frequency neighborhoods, $\chi^2(1, n = 138) = 138.00, p < .001$.

Further examination of the 138 words from the corpus revealed that more words tended to "slip" from a relatively low frequency word to a word of relatively higher frequency, $\chi^2(2, n = 138) = 53.51, p < .001$. That is, 75 target words had error words of a relatively higher frequency, 56 target words had error words of a relatively lower frequency, and seven target words had error words of equal frequency. One should note that this "slip" is not a slip from a low-frequency word to a high-frequency word, hence frequency criterion points are not relevant for this analysis. Rather, the errors that were produced were of a relatively higher frequency than the frequency of the intended word, replicating the findings of del Viso et al. (1991).

One of the assumptions of this analysis is that there is a normal distribution of lexical items around a target word. That is, there is an equal chance that an error will be higher than, or lower than the target word. A normal distribution of words around a target may not reflect the true distribution of items in the lexicon. Rather, there may be a skewed distribution of lexical items around a target word. Indeed, a number of investigations have demonstrated abnormality in the distribution of items in the lexicon (see, e.g., Bard & Shillcock, 1993; and Zipf, 1935/1965). If the distribution of the lexicon is negatively skewed, the results obtained may only be a reflection of the statistical properties inherent in the language rather than a reflection of psychological mechanisms and processes. To more accurately assess the levels of chance, and therefore, distinguish between a statistical and a psychological explanation, several additional analyses were conducted using items from the lexicon to establish a baseline.

The average length of the target malapropism words (in number of phonemes) was used in order to find a set of words from the lexicon that were comparable. The mean length of the malapropism targets was 5.6, or 6, phonemes. A search of the online version of the approximately 20,000 word Webster's Pocket Dictionary resulted in 1,717 six-phoneme content words.

The mean log-frequency value of the malapropism targets (1.39 occurrences per million) was used to partition the "population" of six-phoneme content words in a type and token analysis. The mean of the malapropism targets was used because it was assumed that the distribution of lexical items around this point would provide an accurate estimate of mis-selecting a lexical item by chance (i.e., a distribution of potential errors around a target word).

A type analysis examined the number of words above and below the mean value of the malapropism targets. There were 1,453 words below the mean (85%), and 264 words above the mean (15%). The positive skew in this distribution suggests that a *lower frequency item* should be selected by chance alone. This prediction is the opposite of that just observed — more targets slipped to words that were relatively higher, not lower, in frequency.

A token analysis examined the log-frequency values of those words above and below the mean log-frequency value of the malapropism targets (1.39 occurrences per million). The sum of the log-frequency values for words below the mean was 7,214 and the sum of the log-frequency values for words above the mean was 20,923. From the total token frequency distribution, the words below the mean accounted for 25% of the distribution, while the words above the mean accounted for 75% of the distribution.

There were 75 targets out of 138 malapropisms that slipped to a word higher in frequency (54%) and 56 targets out of 138 malapropisms that slipped to a word lower in frequency (41%). A chi-square analysis using the values from the six-phoneme-words distribution as expected frequencies and the values from the malapropisms as the observed frequencies was performed. The analysis found that the distributions were significantly different, $\chi^2(1, n=95)=16.12, p<.001$, suggesting that the malapropisms are slipping in the *direction* that chance might predict, but not to the *extent* that chance might predict.

To further assess whether the distribution of words among the frequency-density factors is exhibited in the lexicon in general, ten random samples of 138 content word tokens were drawn from the online version of the approximately 20,000 word Webster's Pocket

TABLE 1

The observed frequencies of 138 malapropisms as a function of neighborhood density and target frequency for two different target frequency criteria

	Partition point of log 35		*Partition point of log 85*	
Target Frequency	*Low*	*High*	*Low*	*High*
Low Density	47	22	56	13
High Density	25	44	36	33

Dictionary. The samples were limited to content words because of the well documented differences between content and function words regarding the rate of involvement in speech errors and other lexical characteristics. While other researchers have used random selections of targets and nontargets from other types of errors as control items (Harley & MacAndrew, 1995), a random selection of items from the lexicon as a whole was deemed a better method of evaluating whether the patterns exhibited by the malapropisms are the same as those patterns exhibited by words in the lexicon in general. If the patterns displayed by the malapropism targets and the random samples from the lexicon are similar, then the results from the earlier analyses may only be a reflection of the statistical properties of the lexicon. If the patterns are dissimilar, the obtained results would suggest that certain words are more prone to mis-productions as a function of their lexical characteristics.

If the pattern obtained in the malapropism corpus is the same as the pattern that exists in the lexicon on the whole, then one would expect to find the same frequencies of occurrence for the randomly selected words as for the words from the malapropism corpus. To measure this likelihood, 10 two-way chi-squares were performed using the distributions across conditions for the malapropism targets as the observed frequencies and the distributions across conditions for the random samples as the expected frequencies.

The random samples were partitioned in a manner analogous to the methods used for the malapropisms. The log values of 35 and 85 were used to classify items as high or low frequency items. The median density value of all the random samples (a value of 8) was used to partition each sample into dense and sparse neighborhoods.

All of the randomly drawn samples of words differed significantly from the malapropisms at the .001 level (see Table 2 for the chi-square values. All $df = 1$; all $n = 138$). Thus, the pattern of neighborhood characteristics found in malapropisms is different than the pattern of neighborhood characteristics found in items randomly selected from the lexicon.

To examine the 138 malapropisms and their relation to the lexicon as a whole, a one-way ANOVA was carried out for each neighborhood characteristic (word frequency, density, and neighborhood frequency). Each ANOVA compared the target words, the error words, and the ten random selections of 138 word tokens from the lexicon. The ANOVA on word frequency revealed a significant difference among word types, $F(11, 1644) = 17.06$, $p < .001$. A planned contrast revealed that there was no difference between the frequency of the target words and the frequency of the error words, $F(1, 274) < 1$. A planned contrast

TABLE 2

The number of malapropism targets and control words from ten sets of randomly selected words as a function of word frequency (at the log value of 35 and 85 criterion points) and neighborhood density. (All χ^2 have $df = 1$, $n = 138$)

			Set of randomly selected words									
Word freq.	*N'd Density*	*Targets*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Log value of 35 criterion points												
Low	Low	47	16	21	25	22	26	23	19	21	27	21
Low	High	25	7	4	9	2	1	5	8	3	5	12
High	Low	22	42	51	41	47	46	49	58	44	52	41
High	High	44	73	62	63	67	65	61	53	70	54	64
χ^2			127.39	164.15	62.33	314.10	612.26	124.65	101.25	214.18	113.97	61.32
Log value of 85 criterion points												
Low	Low	56	30	40	40	39	46	44	43	35	42	38
Low	High	36	19	15	13	9	10	7	11	6	12	20
High	Low	13	28	32	26	30	26	28	34	30	37	24
High	High	33	61	51	59	60	56	59	50	67	47	56
χ^2			58.63	53.43	65.04	110.19	85.72	142.90	79.49	189.48	72.40	35.81

between the words involved in speech errors (i.e., both target and errors) and the average of the ten sets of 138 words randomly selected from the lexicon revealed a significant difference, $F(1, 1644) = 179.387$, $p < .001$: both targets and errors were of lower frequency than the randomly selected items from the lexicon. (See the top-left panel of Figure 3.) These planned contrasts replicate the findings of Harley and MacAndrew (1992, 1995) regarding word frequency. They found that target words were lower in frequency than control items.

The ANOVA on neighborhood density revealed a significant difference among word types, $F(11, 1644) = 9.05$, $p < .001$. Planned contrasts revealed that there was no difference between the neighborhood densities of the target and error words, $F(1, 274) < 1$). A planned contrast between the words involved in speech errors (i.e., both target and errors) and the average of the ten sets of 138 words randomly selected from the lexicon did reveal a significant difference, $F(1, 1644) = 81.05$, $p < .001$: items involved in malapropisms (i.e., the target and errors) had lower density neighborhoods than the randomly selected items from the lexicon. (See the top-right panel of Figure 3.)

The ANOVA on neighborhood frequency also revealed a significant difference among word types, $F(11, 1644) = 7.121$, $p < .001$. Planned contrasts revealed no difference between the neighborhood frequency of the target words and error words, $F(1, 274) < 1$. A planned contrast between the words involved in speech errors (i.e., both target and errors) and the

TABLE 3

The mean value for each neighborhood characteristic and the standard deviations (in parenthesis) for targets and errors in 138 malapropisms

	Target words	*Error words*
Log-frequency (of occurrences per million)	1.39	1.48
	(.83)	(.76)
Density (number of similar words)	5.44	5.13
	(7.96)	(5.13)
Log-neighborhood-frequency (of occurrences per million)	.10	.08
	(.99)	(.97)

randomly selected words from the lexicon revealed a significant difference, $F(1, 1644) = 66.94$, $p < .001$: items involved in malapropisms (i.e., the target and errors) had a lower average neighborhood frequency than the randomly selected items from the lexicon. (See the bottom panel of Figure 3.)

DISCUSSION

The results of the chi-square analysis of 138 target words at two different word frequency criterion points demonstrated that significantly more malapropisms were low frequency words with sparse neighborhoods than were low frequency words with dense neighborhoods. The opposite effect was found for high frequency words at each criterion point. More malapropisms were high frequency words with dense neighborhoods than with sparse neighborhoods. Furthermore, more targets had low frequency neighborhoods than high frequency neighborhoods (at both frequency criterion points).

Additional chi-square analyses and ANOVA between random samples of words from the lexicon and words involved in malapropisms demonstrated that words prone to the whole word speech error known as a malapropism are different from words contained in the lexicon as a whole. Specifically, malapropisms seem to be lower in word frequency, neighborhood density, and neighborhood frequency than items in the lexicon in general.

Although malapropisms seem to differ from words in the lexicon on a number of characteristics, malapropisms are similar to words in the lexicon in the way that word frequency affects performance on these items. In speech perception it has been shown that high frequency words are perceived more quickly and accurately than low frequency words. Superior performance of high frequency words also seems to be found in speech production. High frequency words tend to be "immune" to phonological speech errors (see Dell, 1990 and Stemberger & MacWhinney, 1986), and are named faster than low frequency words (see Oldfield & Wingfield, 1965; and Jescheniak & Levelt, 1994).

The results of the current investigation found a similar frequency advantage for malapropisms. Significantly more malapropisms were low frequency words than high frequency words when the log 85 criterion point was used. At the log 35 criterion point,

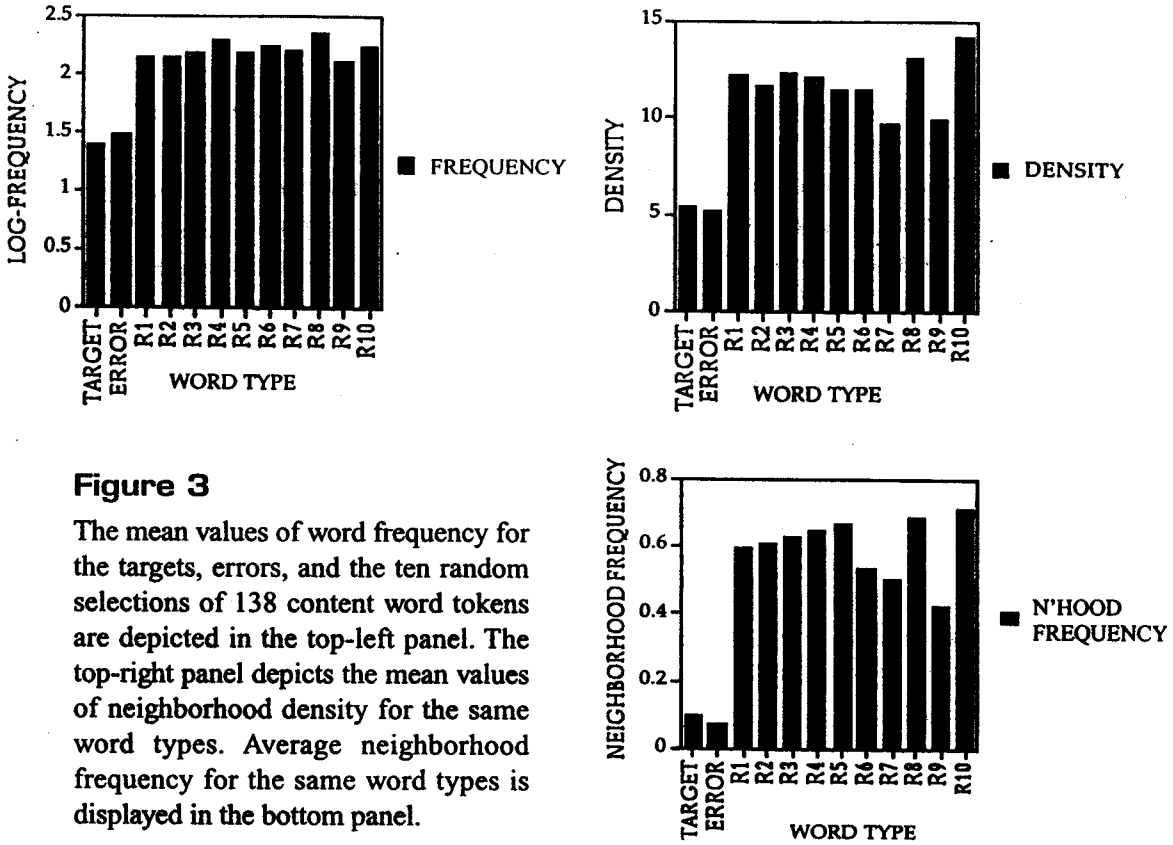

Figure 3

The mean values of word frequency for the targets, errors, and the ten random selections of 138 content word tokens are depicted in the top-left panel. The top-right panel depicts the mean values of neighborhood density for the same word types. Average neighborhood frequency for the same word types is displayed in the bottom panel.

the difference between the number of malapropisms that were high or low frequency words was not significant. The difference was, however, in the same direction as the log 85 criterion point; more malapropisms were low frequency words rather than high frequency words. This finding replicates, to some degree, the findings of Dell (1990) and Stemberger and MacWhinney (1986) regarding the influence of word frequency on the prevalence of *phonological* speech errors.

The influence of word frequency in malapropisms should be viewed with caution. Although the significant difference at the log 85 criterion point may suggest a similarity between whole word and phonological speech errors, this similarity may depend on criteria used to determine frequency categories. The means of assessment and the criterion points set in an investigation, as well as the language and representativeness of a particular error corpus, must all be kept in mind when evaluating the results of a particular examination.

In addition, the problems surrounding the collection of speech errors have been well described (Cutler, 1982; Ferber, 1991) and should not be taken lightly. However, Stemberger (1985b) suggests that converging measures can be used to investigate a speech production phenomenon more reliably than any single measure. Further experimentation, statistical analyses, and simulations must be performed to clarify this frequency issue.

Caution is also called for in interpreting the difference in frequency between target words and errors. A chi-square analysis found that more target words tended to slip to words that were of a relatively higher frequency than to words of a relatively lower, or equal frequency. These findings replicate the results of del Viso et al. (1991) for a Spanish error corpus. An ANOVA, however, found no difference between the frequency of the target words

and the frequency of the errors. This analysis replicated the null-finding by Harley and MacAndrew (1992, 1995).

The replication of both sets of results using the same error corpus further stresses that one must be conscious of the methods of assessment used in an investigation. These findings also raise an interesting question regarding the sensitivity of the speech production system. Is the speech production system sensitive to small, relative differences in word frequency among lexical items? If so, how does this affect the performance of the speech production system? The chi-square analysis would suggest that even minor differences in word frequency influence the accuracy with which an item is selected for production.

Furthermore, chi-square analyses using chance estimates derived from type and token frequency distributions obtained from the lexicon suggest that the results regarding the frequencies of targets and errors are not due to statistical properties of the lexicon. Rather, psychological mechanisms and processes appear to constrain the degree to which target words slip to an error.

Most importantly, the results from the current investigation suggest that the density of phonological similarity neighborhoods affect speech production as well as speech perception. The influence of neighborhood density on speech production is almost exactly the opposite of the effects found in speech perception (Goldinger, Luce, & Pisoni, 1989; Luce, 1986; Luce, Pisoni, & Goldinger, 1990).

It was predicted that the neighborhood density influences on speech production should parallel those on speech recognition. This prediction was based on the parallel influence of word frequency in both perception and production (i.e., the phonological error results of Dell 1990; and Stemberger & MacWhinney, 1986). Specifically, it was predicted that more malapropisms should be found among low frequency words with dense neighborhoods, rather than among high frequency words with sparse neighborhoods. Exactly the opposite pattern of neighborhood density was found for the malapropisms: more malapropisms were low frequency words with sparse neighborhoods or high frequency words with dense neighborhoods.

The density effect found in the current study contrasts with the predictions derived from NAM, a model of spoken word recognition. There may be several reasons for the inability of NAM to predict which items are prone to speech errors. One reason is that any model of speech perception may not accurately describe the processes used in speech production. In production the initial activation is top-down from a semantic representation to a phonological representation. In this case, "conspiracies" (Taraban & McClelland, 1987) of similar lexical items may be needed to sufficiently activate the target item. The larger the set of similar items, the more supportive activation the target item will receive.

In speech perception the initial activation is bottom-up from phonological representations to semantic representations. In this case, a single item must be discriminated from numerous similar candidates which have been partially activated. As the number of similar items increases, so does the level of difficulty in discriminating the target from the other activated items.

Another reason that NAM was unable to predict which items are prone to speech errors may be that NAM, in particular, is lacking some element(s) that would make it useful in understanding both perception and production. Specifically, NAM contains only one level

of representation in which the phonological form (analogous to the lexemes found in speech production models) of similar words compete in order to be selected. Many models of speech production contain multiple levels of representation (see Dell, 1986; 1988; 1990; Garrett, 1988; and Harley, 1984; for examples). Perhaps if NAM were to include a sublexical segmental level (much like the individual phoneme representations in many models of speech production), it would be able to make more accurate predictions regarding phonologically related speech errors based on how activation spreads between and within these levels.

How models of speech production, such as Dell (1986, 1988, 1990) or Jescheniak and Levelt (1994), would account for the findings of whole word speech errors, such as the malapropism results of the current study, is not certain. Dell (1986, 1988) describes a pathway which is meant to account for the lexical bias in speech errors. This pathway begins at a lemma and proceeds to the lexeme representation and then to the appropriate phonological units. A phonological speech error may occur when activation spreads back up from a phonological unit, to a different lexeme, and then to the lemma connected to it. Although Dell proposed this pathway to account for the lexical bias that is found in phonological speech errors, the same mechanism may be used to explain some of the effects found here for malapropisms. However, predicting how an interactive model would perform without actually conducting the simulation is a hazardous practice. Exactly how this model would account for the neighborhood density effects found in the malapropisms in this investigation is unclear. Converging evidence from additional statistical analyses, experimental investigations and computer simulations would further clarify the influence of phonological similarity neighborhoods in speech production.

It is also unclear if the speech production model proposed by Jescheniak and Levelt (1994) could account for the density effects of the current investigation. Jescheniak and Levelt (1994) reported that density (as measured by Coltheart's N and the cohort count) did not seem to influence naming latency. The contrast between this finding and the current results makes it difficult to formulate predictions based on their model.

One possible reason for the differences could be the metrics used in the two experiments. As discussed in the methods section, the cohort count is insensitive to similarity beyond the divergence point. Coltheart's N is technically a measure of orthographic similarity, and is not a direct measure of phonological similarity like the phonological neighborhood metric used in this investigation. Even if a phonological N count was used, the number of similar lexical items produced by *changing* one phoneme may have been underestimated. The addition, deletion, or substitution of a phoneme, as in the neighborhood metric, may be a more accurate assessment of the number of similar lexical items. In addition, the regression analysis used by Jescheniak and Levelt to analyze the influence of density on production may have been insensitive to density effects. Thus, the measures and analyses used by Jescheniak and Levelt (1994) may have been insensitive to any density effects that may have existed in their data.

Finally, the results regarding neighborhood frequency also contradict the predictions derived from NAM. More errors were found in low frequency neighborhoods than in high frequency neighborhoods. It was expected that more errors would be found in high frequency neighborhoods because there would be more items with increased activation (as a function of their higher frequency) competing to be selected by the speech production system. The

opposite findings further suggest that "conspiracy effects" (Taraban & McClelland, 1987) between a word and its neighbors may be at work in the speech production system. That is, a target word may receive facilitation from its neighbors. Those targets that receive sufficient facilitation from their neighbors (i.e., those targets in high frequency neighborhoods) are correctly selected by the speech production system. However, those targets in low frequency neighborhoods may not receive sufficient facilitation from their neighbors, and may, therefore, be more prone to speech errors.

A series of analyses by Bard and Shillcock (1993) found a positive correlation between competitor set size and competitor set frequency. This correlation suggests that the frequency of the competitor set rather than the size of the competitor set may be responsible for some of the "density" findings in the word recognition literature. However, the neighborhood frequency effect, taken together with the density interaction found in the current study, suggests that such a correlation does not exist for malapropisms. Most malapropisms came from low frequency neighborhoods whether these neighborhoods were sparse or dense. These findings also suggest that the characteristics of malapropisms are different from the characteristics of the lexicon in general.

The current investigation clearly demonstrates that phonological similarity neighborhoods influence speech production. At the very least this investigation demonstrates that neighborhood density influences *mis*productions or speech errors. The likelihood of a word being involved in a speech error, specifically a malapropism, is influenced to some degree by the frequency of the word *and* by the sparseness of its phonological neighborhood. Further investigations of phonological similarity neighborhoods are needed to clarify their role in speech production, as well as speech perception.

Received: February 27, 1996; revised manuscript received: June 11, 1997; accepted: June 20, 1997.

REFERENCES

BARD, E. G., & SHILLCOCK, R. C. (1993). Competitor effects during lexical access: Chasing Zipf's tail. In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting* (pp. 235–275). Hillsdale, NJ: Lawrence Erlbaum Associates.

BERG, T. (1988), *Die abbildung des Sprachproduktionprozesses in einem Aktivationsflussmodel.* Tübingen: Max Niemeyer.

COLTHEART, M., DAVELAAR, E., JONASSON, J. T., & BESNER, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

CUTLER, A. (1982). The reliability of speech error data. In A. Cutler (Ed.), *Slips of the tongue and language production* (pp. 7–28). Berlin: Walter de Gruyter/Mouton (also appeared in *Linguistics* (1981) **19**, 561–582).

DELL, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, **93**, 283–321.

DELL, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory & Language*, **27**, 124–142.

DELL, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, **5**, 313–349.

DELL, G. S., & O'SEAGHDHA, P. G. (1991). Stages of lexical access in language production. *Cognition*, **42**, 287–314.

FAY, D. A., & CUTLER, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, **8**, 505–520.

FERBER, R. (1991). Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue. *Journal of Psycholinguistic Research*, **20**, 105–122.

FROMKIN, V. A. (Ed.) (1973). *Speech errors as linguistic evidence.* The Hague: Mouton.

GARRETT, M. F. (1988). Processes in language production. In F. J. Newmeyer (Ed.). *Linguistics: The Cambridge survey: Vol. III Language: Psychological and biological aspects* (pp. 69–96). Cambridge, U.K.: Cambridge University Press.

GOLDINGER, S. D., LUCE, P. A., & PISONI, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, **28**, 501–518.

GOLDINGER, S. D., & SUMMERS, W. V. (1989). *Lexical neighborhood in speech production: A first report.* Research on Speech Production. (Progress report No. 15) Bloomington: Indiana University, Department of Psychology.

HARLEY, T. A. (1984). A critique of top-down independent level models of speech production: Evidence from nonplan-internal speech errors. *Cognitive Science*, **8**, 191–219.

HARLEY, T. A., & BROWN, H. E. (in press). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology.*

HARLEY, T. A., & MACANDREW, S. B. G. (1992). Modeling paraphasias in normal and aphasic speech. Proceedings of the *14th Annual Conference of the Cognitive Science Society*, (pp. 378–383). Hillsdale, NJ: Lawrence Erlbaum.

HARLEY, T. A., & MACANDREW, S. B. G. (1995). Interactive models of lexicalization: Some constraints from speech error, picture naming, and neuropsychological data. In J. P. Levym D. Bairaktaris, J. A. Bullinaria, P. Cairns (Eds.), *Connectionist models of memory and language* (pp. 311–331). London: UCL Press.

HOWES, D. H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, **29**, 296–305.

HURFORD, J. R. (1981). Malapropisms, left-to-right listing, and lexicalism. *Linguistic Inquiry*, **12**, 419–423.

JAKOBSON, R., & HALLE, M. (1956). *Fundamentals of language*, The Hague: Mouton.

JESCHENIAK, J. D., & LEVELT, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**, 824–843.

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

LEVELT, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA.: MIT Press.

LEVELT, W. J. M. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition*, **42**, 1–22.

LEVELT, W. J. M., SCHRIEFERS, H., VORBERG, D., MEYER, A. S., PECHMANN, T., & HAVINGA, J. (1991a). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, **98**, 122–142.

LEVELT, W. J. M., SCHRIEFERS, H., VORBERG, D., MEYER, A. S., PECHMANN, T., & HAVINGA, J. (1991b). Normal and deviant lexical processing: Reply to Dell & O'Seaghdha (1991). *Psychological Review*, **98**, 615–618.

LUCE, P. A. (1986). *Neighborhoods of words in the mental lexicon.* Unpublished doctoral dissertation, Indiana University, Bloomington, Indiana.

LUCE, P. A., PISONI, D. B., & GOLDINGER, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann, (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

MACKAY, D. G. (1982). The problem of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review*, **89**, 483–506.

MACKAY, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills.* New York: Springer-Verlag.

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

MOWREY, R. A., & MACKAY, I. R. A. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America*, **88**, 1299–1312.

NEWBIGGING, P. L. (1961). The perceptual reintegration of frequent and infrequent words. *Canadian Journal of Psychology*, **15**, 123–132.

NOOTEBOOM, S. G. (1969). The tongue slips into patterns. *In Leyden Studies in Linguistics and Phonetics*, Mouton: The Hague. (Also reprinted in Fromkin, 1973).

OLDFIELD, R. C., & WINGFIELD, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, **17**, 273–281.

QUINLAN, P. T. (1992). *The Oxford Psycholinguistic Database.* Oxford: Oxford University Press.

SAVIN, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America*, , **35**, 200–206.

SOLOMON, R. L., & POSTMAN, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, **43**, 195–201.

STEMBERGER, J. P. (1985a). An interactive activation model of language production. In A. Ellis (Ed.) *Progress in the psychology of language, Vol 1*, (pp. 143–146). London: Lawrence Erlbaum.

STEMBERGER, J. P. (1985b). *The reliability and replicability of speech error data: A comparison with experimentally induced errors.* (Progress Report No. 11). Bloomington: Indiana University, Research on Speech Perception.

STEMBERGER, J. P., & MACWHINNEY, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, **114**, 17–26.

TARABAN, R., & McCLELLAND, J. L. (1987). Conspiracy effect in word pronunciation. *Journal of Memory and Language*, **26**, 608–631.

TWENEY, R., TKACZ, S., & ZARUBA, S. (1975). Slips of the tongue and lexical storage. *Language and Speech*, **18**, 388–396.

VISO, S. DEL., IGOA, J. M., & GARCIA-ALBEA, J. E. (1991). On the autonomy of phonological encoding: Evidence from slips of the tongue in Spanish. *Journal of Psycholinguistic Research*, **20**, 161–185.

ZIPF, G. K. (1935/1965). *The Psycho-biology of language: An introduction to dynamic philology.* Cambridge, MA: MIT Press.