

Naturalistic and Experimental Analyses of Word Frequency and Neighborhood Density Effects in Slips of the Ear*

Michael S. Vitevitch

University of Kansas

Key words

Mondegreens

neighborhood density

speech errors

Abstract

A comparison of the lexical characteristics of 88 auditory misperceptions (i.e., slips of the ear) showed no difference in word-frequency, neighborhood density, and neighborhood frequency between the actual and the perceived utterances. Another comparison of slip of the ear tokens (i.e., actual and perceived utterances) and words in general (i.e., randomly selected from the lexicon) showed that slip of the ear tokens had denser neighborhoods and higher neighborhood frequency than words in general, as predicted from laboratory studies. Contrary to prediction, slip of the ear tokens were higher in frequency of occurrence than words in general. Additional laboratory-based investigations examined the possible source of the contradictory word frequency finding, highlighting the importance of using naturalistic and experimental data to develop models of spoken language processing.

1 Introduction

Naturalistic speech error corpora (Cutler, 1982; Fromkin, 1980; MacKay, 1972) and laboratory-based production tasks—such as SLIP (Baars, Motley, & MacKay, 1975), tongue twisters (Shattuck-Hufnagel & Klatt, 1979), and picture naming (Oldfield & Wingfield, 1965)—have provided complementary pieces of information about the processes and representations involved in the rapid and fluent production of speech. For example, Vitevitch (1997) analyzed a published speech error corpus and found evidence to suggest that *neighborhood density* (i.e., the number of words that sound similar to a

* *Acknowledgments:* This research was supported by Research Grant DC 04259 from the National Institute of Deafness and Other Communicative Disorders, National Institutes of Health. I would like to thank Zinny Bond, Steven B. Chin, Allard Jongman, Mitch Sommers, Holly Lynn Storkel and the anonymous reviewers for many helpful suggestions and comments on an earlier version of this manuscript. I would also like to thank Jonna Armbrüster, Revital Berkovith, and Stacy Greenbaum for their assistance in collecting the data in Experiments 1 and 2.

Address for correspondence: Michael S. Vitevitch, Ph.D., Spoken Language Laboratory, Department of Psychology, 1415 Jayhawk Blvd., University of Kansas, Lawrence, KS 66045; or via e-mail: <mvitevitch@ku.edu>.

target word) may influence speech production as well as speech perception (Luce & Pisoni, 1998). Several experimental techniques, including the tip-of-the-tongue elicitation task, tongue twisters, and picture naming (Vitevitch, 2002a; see also Harley & Bown, 1998; Vitevitch & Sommers, in press) further demonstrated that neighborhood density influences the speed and accuracy of speech production.

In contrast to the use of naturalistic and laboratory-based techniques in studies of speech production, research on speech perception has relied primarily on laboratory methods—employing tasks such as perceptual identification, lexical decision, shadowing, gating, AX-matching, phoneme monitoring, and word spotting, to name but a few—to develop models of spoken word recognition. Compared to research in speech production, there is a paucity of systematic analyses of naturalistically collected speech perception errors (for exceptions see e.g., Bond & Garnes, 1980; Bond & Robey, 1983; Garnes & Bond, 1980). The small number of such analyses in the area of speech perception may be due to the intense commitment of time and other resources that is required to compile a corpus of sufficient size so that statistically significant differences in the data can be observed. However, the publication of a relatively large corpus of “slips of the ear” (Bond, 1999) may provide a readily available means for researchers to reinforce the conclusions of laboratory-based behavior with ecologically valid data from more naturalistic settings (Stemberger, 1992).

1.1

What is a “slip of the ear”?

Slips of the ear are misperceptions of an intended speech signal (Bond, 1999). That is, “[a] listener reports hearing, as clearly and distinctly as any correctly perceived stretch of speech, something that does not correspond to the speaker’s actual utterance.” (Bond, 1999; p. 1). Slips of the ear should not be confused with slips of the tongue in which the speaker intends to utter one thing, but erroneously produces another. Speech production errors, for example, may result in phonological segments being misordered, such as saying “darn bore” while intending to say “barn door,” in the substitution of whole words, or malapropisms, such as saying “monotonous” for “monogamous” (Fay & Cutler, 1977; Vitevitch, 1997), or several other types of blends, reversals, or errors made by the speaker (e.g., Bock, 1996). In slips of the ear, the utterance is produced correctly (i.e., as intended), but an error is made by the perceiver.

Slips of the ear are also referred to as mondegreens. The term mondegreen was coined by Sylvia Wright when she confessed in a column written in the *Atlantic* in 1954 that she had heard the lyric “Oh, they have slain the Earl o’ Morray and laid him on the green” from the Scottish folk song “The Bonny Earl of Morray” as “Oh, they have slain the Earl o’ Morray and Lady Mondegreen.” Typing the word ‘mondegreen’ into an internet search engine will provide the reader with a long list of web sites that contain collections of misheard musical lyrics (see also Carroll, 2002; Edwards, 1998). Some well-known examples of mondegreens include hearing “the ants are my friends” instead of “the answer my friends” (Dylan, 1963) and “excuse me, while I kiss this guy” instead of “excuse me, while I kiss the sky” (Hendrix, 1967).

Collections of musical mondegreens can provide the reader with hours of amusement, however, the reliability of such collections for scientific analysis is questionable.

That is, some contributions may be intentional distortions of lyrics generated for maximal comic effect. Furthermore, it is unclear whether other factors in a song, such as the musical instruments in the background or the temporal characteristics of the music, influence perception in some way. For example, the sounds of certain instruments may mask certain auditory frequencies relevant for speech perception, or the rhythm of a song may induce an unnatural word segmentation strategy (Cutler & Carter, 1987) which could systematically affect the perception of certain phonemes or words. Fortunately, the corpus found in Bond (1999) contains errors in perception that were collected from conversational speech rather than musical lyrics (and with more scientific rigor; see Chap. 1 of Bond, 1999).

1.2

Why are Slips of the Ear important?

Stemberger (1992; pg. 211) stated in a comparison of naturalistic and experimentally induced speech production errors that "... naturalistic data show that the experimental data are ecologically valid, that the results are not due to task-specific strategies, and that the experimental... techniques constitute a reasonable facsimile of normal language processing." Although naturalistic and experimental data have been used to study speech production, theories of speech perception have been developed primarily from data collected in laboratory situations. Typically, experimental investigations of spoken language employ stimuli that are carefully selected and balanced across several variables, recorded with pristine sound quality, and presented to participants under conditions that minimize extraneous noise. Furthermore, listeners in the laboratory participate in rather peculiar tasks, such as the lexical decision task. In the lexical decision task a listener must decide if the sequence of phonological segments heard over a set of headphones was a real word or a nonsense word. Deciding whether you heard a real word or nonsense word differs greatly from what one typically does with spoken input, namely interpret the idea or intention that was conveyed auditorily by an interlocutor.

Although naturalistically collected data has inherent ecological validity, the constraints and limitations of human memory and perception often raise concerns regarding the reliability of error corpora. For example, was the "error" that was recorded a misproduction or a misperception? Are certain types of errors more likely to be detected or remembered than other types of errors? Several studies comparing naturalistic error corpora and controlled laboratory tasks involving the elicitation or detection of errors suggest that naturalistically collected language data can indeed be quite reliable. In an examination of speech production errors, Stemberger (1992) found similar types of errors in naturalistically collected corpora and experimentally elicited speech errors. In speech perception, Voss (1984) found that listeners reported the same kinds of errors for casual conversation that they did for controlled laboratory materials, further increasing our confidence in the reliability and value of naturalistic error corpora.

Despite the artificiality of laboratory tasks and the limitations (and the limited number) of corpora of naturalistic errors, these two methodologies can be used together to provide complementary and converging evidence for factors that affect the recognition of spoken words. For example, the work of Cutler and Norris (1988; see also Cutler & Carter, 1987) suggested that syllable stress is an important cue used to segment words

from continuous speech. In English, a strong syllable (one containing a full vowel) is more likely to be the initial syllable of a lexical item or word. Weak syllables (syllables containing central or reduced vowels) are less likely to be word initial syllables, or if they are word-initial syllables, they are grammatical words such as *the* or *of*. Cutler and Butterfield (1992) found in a collection of naturalistic misperceptions and in laboratory induced misperceptions that listeners erroneously inserted word boundaries before strong syllables to produce lexical words, and before weak syllables to produce grammatical words. The results of Cutler and Butterfield (1992) show a nice correspondence between the results collected through naturalistic observation and through an experimental task to demonstrate that syllable stress is an important factor in the segmentation of words from continuous speech.

Naturalistic observations and experimental data also suggest that the initial part of a word is important for the correct identification and recognition of spoken words. In a laboratory setting, Grosjean (1980; see also Marslen-Wilson, & Zwitserlood, 1989) presented listeners with fragments of words (starting with the initial portion of the word) that gradually increased in size; a task commonly referred to as the gating task. Grosjean (1980) found that listeners proposed candidate words and were also able to correctly identify the word portions, even though they had not heard the entire word, suggesting that the initial portion of a word may play an important role in generating lexical candidates during spoken word recognition. (See Vitevitch, 2002b, for the influence of the initial phoneme on the speed of word recognition.)

Complimentarily, Bond (1999, p. 59) found in her collection of slips of the ear that consonant misperceptions tended to occur in the initial position more than anywhere else in a word by a ratio of about 2:1. This finding suggests that the initial portion of a word is important for spoken word recognition in the following way: If a listener misperceives the initial portion of a word, an incorrect lexical candidate will be retrieved because the set of possible lexical candidates is based on incorrect information about the beginning of the word. The resulting and erroneous lexical selection made by the listener will also be recorded in a corpus of slips of the ear (if a particular psycholinguist is in earshot). Because a listener will most likely have correctly recognized a word before the entire word has been heard (Grosjean, 1980), any incorrect perception of phonemes near the end of the word will have relatively inconsequential effects on the correct recognition of the word. With the correct recognition of the word (despite the misperception of phonemes near the end of the word) there is no “slip of the ear” to observe. Together these naturalistic and experimental findings suggest that the initial part of a word is important for quickly and accurately recognizing a spoken word.

1.3

The present Slip of the Ear analysis

Experimental work has shown that several other factors influence the speed and accuracy of spoken word recognition, including the frequency with which a word occurs in the language (*word-frequency*), the number of words, or neighbors, that sound similar to that word (*neighborhood density*), and the mean word-frequency of those neighbors (*neighborhood frequency*). The influence of word-frequency on spoken word recognition was demonstrated experimentally some time ago (Brown & Rubenstein, 1961). In a

variety of laboratory tasks, it has been shown that common words in the language are recognized more quickly and accurately than rare words in the language.

Work by Luce and colleagues (e.g., Luce & Pisoni, 1998; Luce, Goldinger, Auer, & Vitevitch, 2000; Luce, Pisoni, & Goldinger, 1990) shows that neighborhood density and neighborhood frequency also influence spoken word recognition. Specifically, words that activate few neighbors (i.e., a *sparse neighborhood*) are recognized more quickly and accurately than words that activate many neighbors (i.e., a *dense neighborhood*). Words with neighbors that are low in frequency (i.e., *low neighborhood frequency*) are recognized more quickly and accurately than words with neighbors that are high in frequency (i.e., *high neighborhood frequency*). These factors have been found to influence word recognition in a number of laboratory tasks including perceptual identification, auditory shadowing, and lexical decision (Luce & Pisoni, 1998). These factors also influence word recognition in many different listener populations, including normal hearing adults (Luce & Pisoni, 1998), elderly adults (Sommers, 1996; Sommers & Danielson, 1999) and adults with cochlear implants (Kirk, Pisoni, & Miyamoto, 1997).

Although the variables of word-frequency, neighborhood density, and neighborhood frequency have received much experimental attention, little work has examined how these factors may influence the perception of words under naturalistic conditions outside of the laboratory. Can complementary and converging evidence for the influence of word-frequency, neighborhood density and neighborhood frequency on spoken word recognition be found in a corpus of naturalistically collected slips of the ear? To answer this question, an analysis of the corpus of slips of the ear found in Appendix B of Bond (1999) was conducted. This corpus contains almost 900 tokens of misperceptions that were collected over a period of several years (see Chap. 1 of Bond, 1999, for details of how the corpus was compiled).

As demonstrated in the laboratory, words with low word-frequency, dense neighborhoods, and high neighborhood frequency are perceived more slowly and less accurately than words with high word-frequency, sparse neighborhoods, and low neighborhood frequency (Luce & Pisoni, 1998). The present examination of a corpus of errors (Bond, 1999) used two different types of analyses to examine the slips of the ear. These analyses paralleled those previously used to examine the speech production error known as a malapropism (Vitevitch, 1997). In the first analysis, the lexical characteristics of the actual utterances were compared to the lexical characteristics of the misperceived utterances. It was predicted that the words that were misperceived would have higher word-frequency, sparser neighborhoods, and lower neighborhood frequency than the actual word that was produced. That is, the lexical characteristics of the actual utterance make that item particularly difficult to perceive. Instead, what will be (mis-) perceived is a lexical item with characteristics that make it more easily retrieved than the target word.

In the second analysis, the values for the actual utterances and the erroneous perceptions were compared to the word frequency, neighborhood density, and neighborhood frequency values of words in general (i.e., words randomly selected from the lexicon) to further examine the influence of these variables on the recognition of spoken words in a naturalistic setting. In this analysis it was predicted that the slip of the ear tokens (i.e., the actual and misperceived utterances) would have lower word-frequency, denser neighborhoods, and higher neighborhood frequency than words in general. That

is, the words that are involved in slips of the ear have lexical characteristics that make them more difficult to perceive than words in general. The outcome of these two analyses could demonstrate that the experimental data are ecologically valid and further guide the development of models of spoken word recognition.

2 Method for lexical analyses

The Bond (1999) corpus contains misperceptions reported by children (106 tokens) and adults (784 tokens). Only adult misperceptions were analyzed because of the relatively small number of children's misperceptions, the limited information in Bond (1999) regarding the age ranges of the children, and the limited availability of word frequency counts for children at various ages. Furthermore, only misperceptions that Bond classified as misperceptions of vowels or of consonants were examined (decreasing the number of tokens that could be analyzed to 241). The misperceptions that were not included in the present analysis, which Bond classified as "complex errors" or errors with "extensive mismatch between utterance and perception," contained errors that often involved the incorrect parsing of adjacent words. Incorrectly parsing adjacent words sometimes resulted in rather large discrepancies in the number of words in the actual and perceived utterances, making a comparison of the lexical characteristics difficult.

Finally, tokens were only analyzed if the words in the actual and perceived utterance pair were both found in the computerized database. If lexical characteristics could not be found for the words in either the actual or perceived utterance, that slip of the ear token was not included in the analysis. Items that were not found in the computerized database could be generally classified as proper nouns, foreign words or phrases, acronyms or auditory spellings, and domain-specific technical terms. Examples from the Bond (1999) corpus that were and were not included in the analysis are listed in Table 1. A total of 88 auditory misperception tokens (consisting of the actual and perceived utterance) were subsequently analyzed (and are listed in Appendix A).

In addition to assessing word-frequency, neighborhood density and neighborhood frequency in the actual and perceived utterances, the number of phonemes, the number of syllables, and familiarity ratings for the words were also analyzed. *Word familiarity ratings* were based on a seven-point scale. A rating of 1 corresponded to an item that was unfamiliar and whose meaning was unknown, whereas a rating of 7 corresponding to an item that was very familiar and whose meaning was well known (Nusbaum, Pisoni, & Davis, 1984). *Word frequency* was measured by using log-transformed values of the Kučera and Francis (1967) word counts.

Neighborhood density, or the number of similar lexical items for each word, was estimated with a simple computational metric. Using the computational metric, neighborhood density is estimated by counting the number of words formed by the substitution, addition or deletion of a single phoneme into any position of a target word (Landauer & Streeter, 1973; Luce & Pisoni, 1998). If a real word is formed, that word is considered a neighbor of the target word. For example, in the target word /sæd/ (*sad*), the substitution of a single phoneme will form the neighbors /bæd/ (*bad*), /sid/ (*seed*), and /sæk/ (*sack*). Note there are more neighbors for the word *sad*, but only a few are listed for illustrative purposes. A word that has many neighbors formed by the substitution,

TABLE 1

Examples of tokens that were and were not analysed

<i>Included in analysis:</i>	Misperception of vowel stir this \rightleftharpoons store this Misperception of consonant it looks like it's carved of teak \rightleftharpoons teeth
<i>Not included in analysis:</i>	Complex errors a purse and a billfold \rightleftharpoons a personal billfold Extensive mismatch between utterance and perception A linguini is a noodle \rightleftharpoons A lean Wheatie is a noodle Proper nouns I just talked to her and saw Maria \rightleftharpoons Marina Foreign words or phrases Kamasutra \rightleftharpoons Karmasutra Wie geht's \rightleftharpoons i gates Acronyms or auditory spellings He's got a CB too \rightleftharpoons CV It's D R O I N O \rightleftharpoons B R O I N O Domain-specific technical terms He's going to write a paper on tonology \rightleftharpoons tenology

Note: Examples are derived from Appendix B of Bond (1999)

addition, or deletion of a single phoneme is said to have a dense neighborhood, whereas a word that has few neighbors formed in this manner is said to have a sparse neighborhood. *Neighborhood frequency* was assessed by taking the mean word frequency (based on log-transformed values of the Kučera & Francis, 1967, word counts) of the items that were considered neighbors via the computational metric described above.

All of the lexical characteristics (familiarity, frequency, density, and neighborhood frequency) were assessed using a database containing computer readable transcriptions of approximately 20,000 words from *Webster's Pocket Dictionary* (see Luce & Pisoni, 1998). This database is accessible via the Internet on a website maintained by the Speech and Hearing Laboratory (directed by Mitch Sommers, Ph.D.) in the Psychology department of Washington University (<<http://128.252.27.74/neighborhood/Home.asp>>).

To compare the actual and perceived utterances to words in general, 10 samples of 88 words were randomly drawn from the computerized lexicon. (See Vitevitch, 1997 for a similar analysis of malapropisms.) In a given sample, words were drawn without replacement, but after a sample of 88 words had been drawn, those words were replaced and eligible to be drawn in the next sample. To ensure an equitable comparison between words in general and the slip of the ear tokens with regards to word frequency, neighborhood density, and neighborhood frequency several constraints were placed on the

sampling procedure. First, because the slips of the ear tokens involved only content words, only content words were sampled from the computerized lexicon.

Second, given the relationship between word length and word frequency (Zipf, 1935) comparing the lexical characteristics of words that were of different length would not be a fair comparison. Therefore, words randomly drawn from the database were included in the sample for further analysis if they were approximately equal in length (as measured by the number of syllables and the number of phonemes) to the slip of the ear tokens. Independent-ANOVAs with number of syllables and number of phonemes as the dependent variables confirmed that there was no statistical difference in word length between the actual and perceived utterances, and the randomly drawn samples of words, $F(11,1044) < 1$ for both dependent variables.

Finally, the randomly sampled words were as well known as the words in the actual and perceived utterances. That is, the word familiarity values were equated. Independent-ANOVAs confirmed that there was no statistically significant difference in word familiarity ratings between the actual and perceived utterances, and the randomly drawn samples of words, $F(11,1044) < 1$. Together, these results suggest that the sampling procedure did indeed select words that were comparable in word class, word length, and familiarity, allowing for an equitable comparison between words in general and the slip of the ear tokens with regard to word frequency, neighborhood density, and neighborhood frequency. Thus, any influence of long, relatively unfamiliar words with idiosyncratic lexical characteristics on the differences between the slip of the ear tokens and the randomly drawn samples of words in word frequency, neighborhood density, and neighborhood frequency should be greatly attenuated.

3 Results

3.1

Comparing the actual utterance to the perceived utterance

The lexical characteristics of 88 slip of the ear tokens (actual and perceived utterance pairs) were assessed. A comparison of the actual utterance to the perceived utterance was carried out with separate independent-ANOVAs using number of syllables, number of phonemes, familiarity, word frequency, neighborhood density, and neighborhood frequency as the dependent variables. For all of the dependent measures, no significant differences were found between the actual and the perceived utterances, all F 's (1,174) < 1 . The mean values for each dependent measure for the actual and the perceived utterances are listed in the two left-most columns of Table 2.

3.2

Comparing the actual and perceived utterances to words in general

For word frequency, an omnibus-ANOVA showed a significant difference among the sets of words, $F(11, 044) = 10.12$, $p < .001$. Post hoc analyses using weighted means comparisons (the actual and perceived utterances—recall the first analysis showed that these conditions were statistically equivalent—vs. the 10 random samples) confirmed that slip-of-the-ear tokens were *more* common in the language than words in general, $F(1, 044) = 102.79$, $p < .001$. Finally, a post hoc weighted means comparison between

TABLE 2

Mean values for the actual and perceived utterances and the 10 random samples of words

	Act.	Per.	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
Phonemes	3.56	3.52	3.68	3.60	3.64	3.77	3.64	3.68	3.78	3.65	3.69	3.69
Syllables	1.20	1.21	1.29	1.33	1.39	1.37	1.36	1.35	1.21	1.33	1.33	1.35
Familiarity	6.84	6.84	6.80	6.82	6.84	6.85	6.83	6.80	6.82	6.88	6.87	6.86
Frequency	1.58	1.68	.84	.94	.99	.94	.90	1.06	.93	1.10	1.03	.91
Density	12.3	13.1	10.6	10.4	9.7	9.4	10.6	10.1	10.9	10.8	10.3	10.1
NHF	1.22	1.28	.85	.84	.78	.68	.85	.81	.86	.85	.81	.71

Notes: **Act.** = Actual utterance. **Per.** = Perceived utterance. **R** = Random sample of comparable words. **Phonemes** = length of the word in number of phonemes. **Syllables** = length of the word in number of syllables. **NHF** = neighborhood frequency. Frequency and NHF are in \log_{10} values. For the random samples, the number of phonemes, the number of syllables, and the familiarity ratings were equated to those values for the actual and perceived utterances; these values are provided only for comparison.

the slip of the ear tokens and the two sample means from the randomly sampled words that were closest to the word frequency values for the slip of the ear tokens (in this case random samples R6 and R8) showed a significant difference between these two groups of means, $F(1, 044) = 44.31$, $p < .001$, suggesting that the difference in word frequency between the slip of the ear tokens and the randomly selected words in general was not due to an unbalanced weighting scheme in the previous analysis. The mean frequency values for the actual utterance, the perceived utterance, and the 10 random samples of comparable words are presented in Table 2. Note that the finding that slip-of-the-ear tokens are more common in the language than words in general contrasts with the prediction derived from laboratory-based research (e.g., Brown & Rubenstein, 1961). This contradictory result will be explored in greater depth in the experiments that follow.

For neighborhood density, an omnibus-ANOVA showed a significant difference among the sets of words, $F(11, 044) = 2.01$, $p < .05$. Post hoc analyses using weighted means comparisons (the actual and perceived utterances vs. the 10 random samples) confirmed that slip-of-the-ear tokens had denser neighborhoods than words in general, $F(1, 044) = 17.38$, $p < .001$. Another post hoc weighted means comparison between the slip of the ear tokens and the two sample means from the randomly sampled words that were closest to the density values for the slip of the ear tokens (in this case random samples R7 and R8) again showed a significant difference between those two groups of means, $F(1, 044) = 6.06$, $p < .05$. As predicted from laboratory studies (e.g., Luce & Pisoni, 1998), words that were misperceived had denser neighborhoods than words in general. That is, the words that were involved in a slip of the ear tended to have more similar sounding words, or neighbors, than words randomly selected from the lexicon. As suggested by Luce and Pisoni (1998) a word with many neighbors is subjected to more competition from neighboring lexical items than a word with few neighbors, accounting for the lower probability that a word in a dense neighborhood will be quickly or correctly selected from the lexicon. The mean neighborhood density values for the actual utterance, the perceived utterance, and the 10 samples of words are presented in Table 2.

Finally, for neighborhood frequency an omnibus-ANOVA showed a significant difference among the sets of words, $F(11, 044) = 13.16, p < .001$. Post hoc analyses using weighted means comparisons (the actual and perceived utterances vs. the 10 random samples) confirmed that slip-of-the-ear tokens had higher neighborhood frequency than words in general, $F(1, 044) = 130.65, p < .001$. A post hoc weighted means comparison between the slip of the ear tokens and the two sample means from the randomly sampled words that were closest to the neighborhood frequency values for the slip of the ear tokens (in this case random samples R7 and R8) again showed a significant difference between those two groups of means, $F(1, 044) = 61.56, p < .001$. As predicted from the laboratory data (Luce & Pisoni, 1998), words that are likely to be misperceived have higher neighborhood frequency than words in general. Said another way, a word that has neighbors that are low in frequency will more likely be recognized quickly and correctly than a word that has neighbors that are high in frequency. The mean neighborhood frequency values for the actual utterance, the perceived utterance, and the 10 random samples of comparable words are presented in Table 2

4 Discussion

To obtain a more complete (i.e., ecologically valid) understanding of the processes and representations involved in spoken word recognition, two analyses of naturally occurring slips of the ear were performed using tokens from the Bond (1999) corpus. The first analysis compared the number of syllables, number of phonemes, familiarity, word frequency, neighborhood density, and neighborhood frequency values of the actual and (mis-) perceived utterances. It was predicted that the words that were misperceived would have higher word-frequency, sparser neighborhoods, and lower neighborhood frequency than the actual word that was produced. However, for all of the dependent measures, no statistically significant differences were found between the actual and perceived utterances. Several other studies that used either a larger number of errors resulting in more statistical power (Cutler & Butterfield, 1992) or different statistics with different distributional assumptions (Bond, 1999) have similarly found no difference between the actual and perceived utterances with regards to word-frequency. The fact that similar (null) results have been found in analyses with more errors and different statistics weakens claims that a statistical artifact may account for this somewhat counterintuitive finding in the present analysis.

Given the repeated failure to find a significant difference in word frequency between the actual and misperceived utterances and a similar failure to find a significant difference in a number of other variables (neighborhood density, neighborhood frequency, familiarity, and word length), perhaps a general design characteristic of the language processing system rather than a statistical artifact is responsible for these results. One general design characteristic of cognitive systems that may account for the lack of a significant difference between the actual and perceived utterances is the processing principle of *graceful degradation*. Graceful degradation is the ability of a processing system to not catastrophically halt processing when given incomplete or incorrect information, but to retrieve the representation that “best matches” the input (McClelland, Rumelhart, & Hinton, 1986).

By way of illustration, consider the example of quickly looking at the food displayed

on a picnic table. In looking at the display quickly, the visual input contains incomplete information. If one were to see a red, somewhat-spherical object about the size of one's fist sitting on the table one might "perceive" what is actually an apple to be a tomato. Note that tomatoes and apples have several similar characteristics or features, such as size, shape, color, appropriateness as picnic-fare, and so forth. Also note that one most likely would not misperceive the red, somewhat-spherical object about the size of one's fist to be a watermelon. Although a watermelon is considered appropriate picnic-fare, there is not enough similarity between the incomplete input and the stored representation of a watermelon (with regard to size, shape, color, etc.) for the representation of a watermelon to be sufficiently activated and erroneously retrieved or perceived.

Now consider graceful degradation in the context of words or lexical representations. One might mistake the incomplete input (due to noise or inattention) of /_æt/ in a discussion of animals to be the word 'rat' rather than the uttered word 'cat'. Note that "cat" and "rat" have several features that are quite similar, such as syllable and phoneme length, phonological segments, frequency of occurrence and familiarity values, and so forth. Further note that one most likely would not mistake "cat" for "pterodactyl," (even though it too is an animal) because there is not enough similarity between the two lexical representations (with regard to word length, phonology, neighborhood density, etc.) for "pterodactyl" to be sufficiently activated and erroneously retrieved or perceived. Viewed in terms of graceful degradation, the lack of a significant difference in all of the dependent variables between the actual and the perceived utterances is not surprising at all. Recall that a processing system that gracefully degrades selects the representation that best matches incomplete or erroneous input rather than catastrophically halting. "Best match" may be defined by any or all of the lexical variables, hence the similarity between actual and misperceived utterances on all of the lexical dimensions examined. In the context of general processing principles such as graceful degradation, the lack of a significant difference in all of the dependent variables between the actual and the perceived utterances is actually to be expected.

A second analysis of naturally occurring slips of the ear compared the lexical characteristics of those tokens to a comparable set of items (with regard to word class, length, and familiarity) that were randomly selected from the lexicon. In this analysis it was predicted that slip of the ear tokens would be lower in word frequency, have denser neighborhoods, and higher neighborhood frequency than words randomly sampled from the lexicon.

The predictions for neighborhood density and neighborhood frequency were supported by the results of the analyses. In general, the utterances that were misperceived tended to contain words that had more phonological neighbors (i.e., dense neighborhoods) and neighbors that were higher in frequency (i.e., high neighborhood frequency) than words of comparable length and familiarity that were randomly selected from the lexicon. The results of the present slip of the ear analysis regarding neighborhood density and neighborhood frequency generalize the influence of these lexical characteristics on spoken word recognition previously obtained only under laboratory conditions. That is, neighborhood density and neighborhood frequency influence the recognition of words in natural, casual conversation as well as the recognition of words in laboratory tasks.

In contrast, the results for word frequency were in the opposite direction of that which was predicted. Based on previous research (e.g., Brown & Rubenstein, 1961), it was predicted that slip of the ear tokens would be lower in word frequency than words randomly sampled from the lexicon. The results of the present analysis showed that slip of the ear tokens were actually higher in word frequency than words randomly sampled from the lexicon. Not only is this finding inconsistent with previous laboratory-based research on spoken word recognition, but it also is inconsistent with a comparison of naturalistic and experimentally elicited speech errors made by Stemberger (1992). Stemberger found that in almost all cases (there were a few cases in which there were null results in one task, but significant differences in the other) the pattern found in naturalistically collected errors was similar to the pattern found in experimentally elicited errors, differing only in magnitude. Stemberger further demonstrated that in most cases those differences in magnitude were due to task differences. In no case that Stemberger examined was the pattern obtained from naturalistically collected errors the *opposite* of the pattern obtained from experimentally elicited errors, making the present finding regarding word frequency puzzling indeed.

A simple solution to this puzzling and counterintuitive finding would be to hypothesize that there was a perceptual bias that influenced the collection of the slips of the ear. Unfortunately there is little evidence to support this hypothesis. Bond (1999; p. 2) states that “[t]he corpus of perceptual errors undoubtedly leans toward the more obvious or noticeable, errors which were significant enough to make listeners puzzled about what they heard. The errors which were reported are also probably the most memorable.” But, what makes an error obvious, noticeable, or memorable? Consider the work of Mandler, Goodman, and Wilkes-Gibbs (1982) who showed that high frequency words are more accurately recalled than low frequency words. If one is not able to immediately record a misperception at the time of observation (i.e., pen and paper are not in hand), one might expect that the error corpus may contain more errors for high frequency words, which are more “memorable” or more readily recalled at a later time, than errors for low frequency words. Alternatively, a word with low frequency of occurrence (i.e., a lexical isolate) may seem more obvious or noticeable than a word with a higher frequency of occurrence. This linguistic example of the von Restorff effect suggests that errors for low frequency words should be more “memorable” than errors for high frequency words and, therefore, low frequency words should be over-represented in the corpus. If one is to hypothesize that a perceptual bias is the cause of the contradictory word frequency effect, then one must also account for why one type of bias is systematically favored over the other.

Furthermore, if a perceptual bias was operating to influence and reverse the robust and ubiquitous effect of word frequency typically found in experimental studies of spoken word recognition, why didn't the same perceptual bias also reverse the neighborhood density and neighborhood frequency effects observed in the present analysis? (See Goh & Pisoni, in press, and Roodenrys, Hulme, Lethbridge, Hinton, Nimmo, 2002, for evidence that neighborhood density affects STM.) Finally, the work of Stemberger (1992) and Voss (1984) found high rates of correspondence for error elicitation/detection results obtained under naturalistic and laboratory conditions, with differences being found only in magnitude *not direction*, providing even less support for the hypothesis that listener bias accounts for the opposite effect of frequency observed in the

present investigation of an error corpus. Instead, the fact that the present analysis found, not a difference in magnitude, but a complete reversal of the robust and ubiquitous word frequency effect demands further investigation.

One possible explanation for the reversal in word frequency observed in the present analysis relates to the types of discursive repair initiated by the listener when a message is not clearly received (e.g., Gagne, Stelmachovich, & Yovetich, 1991; Tye-Murray, 1991; Tye-Murray, Knutson, & Lemke, 1993; Tye-Murray & Witt, 1997; Tye-Murray, Witt, & Castelloe, 1996). It is possible that listeners may use different strategies to clarify a partially perceived message depending on the frequency of occurrence of the word when comprehension breaks down. For example, when a low frequency word is not clearly perceived, the listener may request a repetition of the utterance because not enough information was received to even partially activate lexical candidates. However, when a high frequency word is not clearly perceived, the listener may have received enough information to partially activate several lexical candidates (though perhaps not the right one). The listener may then select a lexical candidate that is "close enough" to the incomplete information received as input (see the earlier discussion of graceful degradation) resulting in a slip of the ear that is later clarified and recorded in a corpus. Low frequency words may indeed be occasionally misperceived in conversation, but these misperceptions may not be recorded in a corpus of slips of the ear because such misperceptions may be resolved in a different way (i.e., a request to repeat the utterance) than the misperception of a high-frequency word (i.e., process the representation that best matches the incomplete input). Such differences in how listeners respond to high and low frequency words that have been incompletely perceived may have been overlooked in the laboratory because many experimental tasks, such as the lexical decision task, do not have an option to repeat a stimulus item (or some other analog of common conversational repair strategies). Unfortunately, to the best of my knowledge, there is no empirical evidence that even suggests that different conversational repair strategies may be used as a function of word frequency, so this hypothesis is speculative at best.

Another possible explanation for why slips of the ear were higher rather than lower in word frequency compared to words in general—one that is partially supported by previous research and will be further pursued in the experiments that follow—relates to the way that high and low frequency words are produced. Wright (1979) found that speakers in several laboratory tasks—including reading a list of words, repeating a short list that was memorized, and reading words embedded in short contexts (e.g., "to ___ at" or "go ___ stop")—produced low frequency words approximately 24% slower than high frequency words (measuring both word and segment durations). Other research has shown that words spoken at a faster rate of speech are recognized less accurately than words spoken at a slower rate of speech (cf., Bradlow, Torretta, & Pisoni, 1996; Kirk et al., 1997). If high frequency words are spoken at faster rates than low frequency words (Wright, 1979), and words spoken at faster rates of speech are recognized less accurately than words spoken at slower rates (Kirk et al., 1997), then the combination of these findings suggests that high frequency (i.e., faster) words should indeed be recognized less accurately than low frequency (i.e., slower) words; a pattern similar to that observed in the present analysis of slips of the ear.

The conclusion that high frequency words should be recognized less accurately than low frequency words, like the results of the present analysis, contradicts several decades worth of laboratory demonstrations of the word frequency effect (e.g., Brown & Rubenstein, 1961). Note, however, that in laboratory studies of speech perception and word recognition, the duration with which the stimuli are produced is commonly held constant across conditions. Thus, a higher rate of incorrectly perceived high frequency words might be observed in naturalistic speech where differences in speaking rate for high and low frequency words might occur, but not in the laboratory where stimulus duration is held constant across conditions. The experiments that follow will more directly explore the differences in word frequency and word duration to provide evidence that may account for the counterintuitive finding regarding word frequency obtained in the present analysis of slips of the ear.

5 Experiment 1

Cutler and Butterfield (1992) described several ways that misperceptions could be induced in the laboratory including filtering or noise-masking the speech signal. Because these techniques do not affect all speech sounds equally, distorting some speech sounds more than others (e.g., Miller & Nicely, 1955), Cutler and Butterfield decided to use faint speech (i.e., speech presented at a level which allowed participants to hear about 50% of the presented input). Although each of these techniques may be used effectively to elicit misperceptions in the laboratory, they may not necessarily mimic the conditions under which the naturalistically observed slips of the ear were obtained. (Note, Bond, 1999, did not include descriptions of the ambient environment for each slip of the ear; e.g., occurred along a busy street, heard over a cell-phone, etc.) Furthermore, techniques to elicit misperceptions that rely on distorting the signal with noise would not adequately address the hypothesis currently being investigated: High frequency words with short durations will be recognized less accurately than low frequency words with long durations. Fortunately, Cutler and Butterfield (1992; p.234) alluded to an experiment that used time-compressed speech to elicit misperceptions; a technique that seems ideally suited to evaluating durational influences on perception as a function of word frequency.

Digitally compressing the stimulus words by 25% to simulate the duration differences observed by Wright (1979) proved useful for two reasons. First, digital compression allowed for very precise control over the duration of all of the words. Repeated attempts by the author to produce natural versions of the words spoken slowly and rapidly were unsuccessful at reaching the 24–25% difference for all of the words. Interestingly Wright (1979) also observed variability in the duration of individual words varying in frequency of occurrence; some rare words (e.g., ‘clef’) had consistently shorter durations than some common words (e.g., ‘song’). By using digital compression, the duration of every word could be reduced by 25%. Had naturally produced tokens been used, it would not have been clear whether the duration of every word was 25% shorter or if the 25% difference was the result of a statistical artifact (e.g., some stimuli spoken very, very quickly and some stimuli spoken very, very slowly, when averaged together may show a 25% difference).

Second, and perhaps most importantly, digital compression enabled the duration

of each word to be decreased without significantly decreasing the intelligibility of each word. Naturally produced changes in speaking rate result in a number of phonetic changes, such as vowel reduction, that occur in order to produce a word more quickly (e.g., Bond & Moore, 1994; Lieberman, 1963; Moon & Lindblom, 1994; Picheny, Durlach, & Braida, 1986; van Bergem, 1995). An example of vowel reduction and the potential decrease in intelligibility that may result from it can be seen in the words 'cat' (/kæt/) and 'cot' (/kɑt/). If the full vowels in these two words are reduced to a schwa, the nondistinct form /kəʔ/ would result. Although /kəʔ/ might be correctly recognized as either "cat" or "cot" in fluent conversational speech with the assistance of contextual cues, there is a significantly decreased likelihood that /kəʔ/ would be accurately recognized if it were edited from that context and presented in isolation (e.g., Bard & Anderson, 1994; Drager & Reichle, 2001; Fowler & Housum, 1987). Furthermore, Haan (1977) found that compressed speech was actually more intelligible than speeded speech, providing additional motivation to use digitally compressed rather than naturally produced tokens. By using digital compression the duration of all of the stimuli could be uniformly decreased without decreasing their intelligibility (relative to naturally produced versions of the same words at a faster speaking rate), thereby attenuating the potential confound between word duration and intelligibility that would occur with naturally produced tokens.

In the present experiment, an identification task was used; participants heard a word presented in the clear over a set of headphones and had to type on a computer keyboard the word that they heard. Note that no noise of any kind (e.g., white, pink, bit-flipped, envelope modulated, etc.) was added to the signal as is traditionally done in perceptual identification tasks (e.g., Brown & Rubenstein, 1961). Although still a somewhat unnatural task, the identification task more closely approximates what people do in natural discourse (i.e., hear speech sounds and map them onto a word-form stored in the lexicon) than many other traditional laboratory tasks such as the lexical decision task (i.e., hear a sound and decide if those sounds are in your lexicon; i.e., decide if it is a real word in English or not).

Half of the words participants heard had a high frequency of occurrence, and the rest of the words had a low frequency of occurrence. Half of the high frequency words were presented in their originally recorded form (referred to as "slow") and half of the high frequency words were presented in their digitally modified form (referred to as "fast"). The same manipulation was made for the low frequency words. It was predicted that a main effect of word frequency would be observed such that high frequency words would be more accurately identified than low frequency words (e.g., Brown & Rubenstein, 1961). A main effect of word duration was also predicted such that slow-words would be more accurately identified than fast-words (Kirk et al., 1997).

Most important for demonstrating that the counterintuitive frequency effect observed in the present analysis of slips of the ear is, in fact, not counterintuitive is the prediction that slow-low frequency words would be more accurately identified than fast-high frequency words. (Note, the difference between these two conditions may occur whether word frequency and duration interact or not.) Recall Wright (1979) found that high frequency words tend to have shorter durations than low frequency words and Kirk et al. (1997) found that words produced at a faster rate (i.e., shorter duration)

were recognized less accurately than words produced at a slower rate (i.e., longer duration). If the results of both of these laboratory-based studies co-occur in natural contexts, then one would predict that slow-low frequency words would be more accurately identified than fast-high frequency words, suggesting that the word frequency effect observed in the present slip of the ear analysis is not counterintuitive, but should in fact be expected. Given that fast-low frequency words and slow-high frequency words rarely occur (Wright, 1979), and that differences in these conditions would neither support nor detract from the hypothesis being investigated, no predictions were postulated for the remaining conditions.

5.1

Method

Participants: Twelve students from the Introductory Psychology pool of research participants at the University of Kansas took part in the experiment in exchange for partial course credit. All the participants were native speakers of English, and reported no history of speech or hearing disorders.

Materials: A set of 60 English words varying in frequency of occurrence as measured by log-transformed values of the Kučera & Francis (1967) word counts was used in this experiment (and are listed in Appendix B). Words classified as high frequency items had a mean log-frequency of occurrence of 1.58, whereas words classified as low frequency items had a mean log-frequency of occurrence of 0.57, $F(1, 58) = 92.65$, $p < .001$. Familiarity, neighborhood density, and neighborhood frequency were calculated using the same database as that used in the analysis of the slip of the ear tokens. The high frequency words had a mean familiarity of 6.9, a mean neighborhood density of 20 neighbors, a mean neighborhood frequency of .9, a mean phoneme frequency of .148, and a mean biphone frequency of .005. (As in Vitevitch & Luce, 1998; 1999 phoneme and biphone frequency constituted the two measures used to assess phonotactic probability.) The low frequency words had a mean familiarity of 6.9, a mean neighborhood density of 19 neighbors, a mean neighborhood frequency of 1.0, a mean phoneme frequency of .143, and a mean biphone frequency of .005. There were no significant differences for any of these variables between the two conditions, all $F(1, 58) < 1$.

Furthermore, the same number of words in each condition had the same initial phonemes (3 words in each condition started with each of the following phonemes /b, d, f, k, l, p, r, s, t, w/). All of the stimuli were spoken in isolation and recorded by the author in an IAC sound attenuated booth using a high-quality microphone. The stimuli were digitized at a sampling rate of 20 kHz using a 16-bit analog-to-digital converter. All words were edited into individual digital files, leveled at 70 db SPL, and stored on computer disc for later playback. The durations of the naturally produced tokens were equivalent between the two conditions, $F(1, 58) < 1$. High frequency words had a mean duration of 704 ms, and low frequency words had a mean duration of 718 ms.

Copies of the sound files were modified with the Tempo function in Sound Edit 16 (Macromedia, Inc.) to digitally decrease the duration of each stimulus item by 25%. Although the overall duration of the tokens was decreased, the pitch of each stimulus file was not altered. More precisely, the Tempo function corrects for any changes in pitch that typically accompany changes in duration. Thus, there was the original set

of 60 words with the durations they were recorded at, and a set of the same 60 words that had their duration digitally altered. Further note that this was the only change that was made to the stimuli. No noise of any kind (e.g., white, pink, bit-flipped, envelope modulated, etc.) was added to the signal; that is, all words were presented “in the clear.”

Two counterbalanced lists each containing 60 words were prepared. Each list had 15 high frequency and 15 low frequency words recorded at the original duration (i.e., slow), and 15 high frequency and 15 low frequency words that had the duration digitally altered (i.e., fast). The second list contained the same words as the first list, but had the opposite duration of that word on the first list.

Procedure: Participants were seated in front of an iMac running PsyScope 1.2.2 (Cohen, MacWhinney, Flatt, & Provost, 1993) that controlled stimulus randomization and presentation. A trial proceeded as follows: The word ‘READY’ appeared in the center of the computer screen for 500ms to indicate the beginning of a trial. Participants were then presented with one of the randomly selected words at 70dB SPL over a pair of Beyerdynamic DT-100 headphones. Participants were instructed to type, using the computer keyboard, as accurately as possible the word that they heard over the headphones. Responses appeared on the screen as they were being typed allowing participants to correct any errors before pressing the “enter” or “return” key and initiating the next trial. Participants were not allowed to hear any of the words a second time. Prior to the experimental trials, each participant received 10 practice trials. These trials were used to familiarize the participants with the task and were not included in the final analysis. Equal numbers of participants received only one of the counterbalanced lists.

5.2

Results and Discussion

The author scored the typed responses for accuracy. Correct responses were those typed responses that exactly matched the target word. Responses that did not exactly match the target word but were nonetheless scored as correct responses included homophonus responses (e.g., “seem” for *seam*), incorrect/phonologic spellings (e.g., “toade” for *toad* or “tuff” for *tough*), and typographical errors. Typographical errors were defined as the insertion or substitution of a letter that was one key away from a target letter, or the deletion of target letters. For example the response “rioce” for the target word *rice* (*i* and *o* are neighboring keys), the response “duit” for the target word *suit* (*d* and *s* are neighboring keys), the response “coug” for the target word *cough*, and the response “pag” for the target word *page* were all scored as correct responses. Perceptual errors were those responses that did not match the target word and did not fit into any of the alternative categories for a correct response. An example of a perceptual error would be the response “lull” for the target word *wall*.

A repeated measures ANOVA was used to examine the influence of word frequency and duration on the accuracy of word identification (no difference was found between the two lists in a mixed ANOVA, $F(1,10) < 1$, so additional analyses collapsed over this between subjects factor). Because a set of highly controlled stimuli, selected to match several criteria (same initial consonant, neighborhood density value, neighborhood frequency value, etc.), that almost exhausted the pool of possible items was used in this experiment, only ANOVAs with participants as a random factor were conducted. That

is, the stimulus items were not selected randomly so ANOVAs treating stimulus items as a random variable are not appropriate (Cohen, 1976; Hino & Lupker, 2000; Keppel, 1976; Raaijmakers, Schrijnemakers, & Gremmen, 1999; Smith, 1976; Wike & Church, 1976).

The results showed a main effect of duration, $F(1, 11) = 14.44 < .01$, such that words spoken at the original duration (i.e., slow) were identified correctly more often (2.8% error rate) than words at the digitally modified duration (i.e., fast; 8.1% error rate), replicating the results of Kirk et al. (1997). No significant main effect of word frequency was observed in the present experiment, High frequency words = 5.8% error rate, Low frequency words = 5.0% error rate; $F(1, 11) < 1$. Issues related to statistical power may be responsible for the statistically nonsignificant difference. Alternatively, tasks that rely on reaction time measures, such as the lexical decision task, may be more sensitive to the time course of cognitive processes than tasks that rely on accuracy rates, such as the identification task employed in the present experiment (Levelt, Roelofs, & Meyer, 1999; p. 2). Also note that the identification task used in the present experiment differed from the identification task used by Brown and Rubenstein (1961). Although they used monosyllabic words like those used in the present experiment, they embedded their words in noise. Recall that the words used in the present experiment were presented in the clear. Without stimulus degradation, performance may be at ceiling, making it difficult to observe a word frequency effect. Indeed, the worst performance in this experiment was the 8.3% error rate for the fast-high frequency words, which is still a high level of performance; the fast-low frequency words had a 7.8% error rate, the slow-high frequency words had a 3.3% error rate, and the slow-low frequency words had a 2.2% error rate.

Although there was no significant interaction between duration and word frequency, $F(1, 11) < 1$, theoretical interest warrants further analyses of the individual conditions. Recall the hypothesis for the contradictory word frequency effect observed in the analysis of slips of the ear: high frequency words with shorter durations should be identified less accurately than low frequency words with longer durations. A post hoc means comparison between the low frequency words with the original duration (slow) and the high frequency words with the digitally modified duration (fast) shows a significant difference between the two conditions, $F(1, 11) = 6.74, p < .05$, such that slow-low frequency words (2.2% error rate) were more accurately identified than fast-high frequency words (8.3% error rate). This result supports the hypothesis that the contradictory frequency effect observed in the corpus of naturally occurring recognition errors may be due to differences in the production of words varying in frequency of occurrence in naturalistic settings. Specifically, high frequency words may be spoken more quickly than low frequency words in naturalistic settings, accounting for the higher prevalence of high frequency words in a corpus of slips of the ear. This prediction, the results of the slip of the ear analysis, and now the results of the present experiment all run counter to the word-frequency effects typically observed in previous laboratory studies. Note, however, that typical laboratory studies use stimuli that have equivalent durations between high and low frequency conditions. The next experiment attempted to demonstrate the word-frequency effect using the same words as were used in the present study, but in a more traditional laboratory task, namely the lexical decision task, under typical laboratory conditions (i.e., the words had equal durations).

6 Experiment 2

In the present experiment, the same 60 words that were used in Experiment 1 (i.e., the original recordings) were presented to participants in a classic laboratory task under traditional laboratory conditions; a lexical decision task was used with words of equal duration. In the lexical decision task a listener is presented with either a word or nonsense word, then presses a button on a response box to indicate whether they heard a real word or a nonsense word. The lexical decision task was also selected because it allows the use of reaction times as a dependent measure. Time-based dependent measures may be more sensitive to cognitive processes than dependent measures that simply measure the end-state, such as the accuracy rates used in the identification task in Experiment 1 (Levelt et al., 1999; p.2). By replicating the word frequency effect typically found in studies of spoken word recognition with the traditional laboratory constraints imposed on the same words that produced the opposite word frequency effect observed in Experiment 1 we can provide further evidence that the differential influence of word frequency and word duration may have caused the opposite finding we observed in the analysis of naturalistically collected slips of the ear.

6.1

Method

Participants: Ten students from the Introductory Psychology pool of research participants at the University of Kansas took part in the experiment in exchange for partial course credit. All the participants were right handed, native speakers of English, and reported no history of speech or hearing disorders. None of the participants took part in Experiment 1.

Materials: The original recordings of the 60 real words used in Experiment 1 were also used in the present experiment. A set of 60 nonwords was created by substituting a phoneme into the last position of real words that were not part of the stimulus set. The nonwords were recorded and treated in the same way as the real words used in Experiment 1.

Procedure: Participants were seated in front of an iMac running PsyScope 1.2.2 (Cohen et al., 1993) that controlled stimulus randomization and presentation. Response latencies were collected with millisecond accuracy via a New Micros button box interfaced to the computer. The response box had the label 'NONWORD' on the left button and the label 'WORD' on the right button (under the dominant hand). A trial proceeded as follows: The word 'READY' appeared in the center of the computer screen for 500 ms to indicate the beginning of a trial. Participants were then presented with one of the randomly selected spoken stimuli at 70 dB SPL over a pair of Beyerdynamic DT-100 headphones. Reaction times were measured from the onset of the stimulus to the button press response. If the maximum reaction time (3 s) expired, the computer automatically recorded an incorrect response and presented the next trial. Participants were instructed to respond as quickly and as accurately as possible.

Half of the trials consisted of real words in English, half of the trials consisted of the nonwords. Prior to the experimental trials, each participant received 10 practice trials. These trials were used to familiarize the participants with the task and were not included in the final analysis.

7 Results and Discussion

A repeated measures ANOVA was used to examine the influence of word frequency independently on the mean reaction times and accuracy rates. For the accuracy rates there was no significant difference, $F(1, 9) < 1$, between high frequency (93%) and low frequency (93%) words, suggesting that there was no speed-accuracy trade-off. For the reaction times, a significant difference was found, $F(1, 9) = 16.56, p < .01$. Participants responded more quickly to high frequency words (958 ms) than low frequency words (1001 ms), demonstrating the classic word frequency effect under traditional laboratory conditions. That is, when the duration of the words is equivalent, high frequency words are responded to more quickly than low frequency words.

The same set of words was used in Experiments 1 and 2 to demonstrate the conditions that result in the traditional word frequency effect (i.e., when the words have equal durations) and the conditions that result in a reversal of the word frequency effect (i.e., when high frequency words have shorter durations than low frequency words). The results of these two experiments directly support the hypothesis that the opposite word frequency effect observed in the analysis of naturalistically collected slips of the ear may have been due to differences in the way speech is produced in naturalistic versus laboratory settings. This hypothesis was derived from previous studies that separately investigated the influence of word frequency on word duration (Wright, 1979), and of speaking rate on word recognition accuracy (Kirk et al., 1997). Although logic would suggest that the combination of these previous studies would confirm this prediction, Experiment 1 provided direct empirical evidence to support this hypothesis. These results cast further doubt on the simple hypothesis that the opposite word frequency effect obtained in the slip of the ear analysis was the result of a perceptual bias on the part of the error collector. Rather, these results suggest that there may be a "production bias" on the part of the speaker to produce high frequency words more rapidly than low frequency words, leading to specific perceptual consequences for the listener.

8 General Discussion

The present results from an analysis of naturalistically collected slips of the ear and from two laboratory-based experiments exemplify two points made by Norman (1981) in his analysis of errors in action:

The collection and analysis of naturally occurring errors forces us to consider behavior that is not constrained by the limitations and artificiality of the experimental laboratory. By examining errors, we are forced to demonstrate that our theoretical ideas can have some relevance to real behavior. (Norman, 1981; p. 13.)

The finding from the corpus analysis that slips of the ear tend to have denser neighborhoods and neighborhoods with higher frequency than words in general provides consistent and ecologically valid evidence in support of the laboratory studies that have previously demonstrated influences of neighborhood density and neighborhood frequency on spoken word recognition (e.g., Luce & Pisoni, 1998). To the best of my knowledge, the present study is the first demonstration of an influence of neighborhood density and

neighborhood frequency on spoken word recognition in a naturalistic rather than a laboratory setting.

Norman (1981; p. 14) also stated, "... to validate what has been theoretically postulated as the cause of errors, laboratory tests are useful." The results of Experiments 1 and 2 provided direct, laboratory-based evidence for the cause of the opposite word frequency effect observed in the corpus analysis; slips of the ear tended to be higher in frequency than words in general. Wright (1979) observed that high frequency words tended to have shorter durations than low frequency words. Kirk et al. (1997) observed that words spoken at a faster rate of speech were identified less accurately than words spoken at a slower rate of speech. Experiments 1 and 2 directly demonstrated that the same words varying in word frequency had different perceptual consequences depending on whether the duration of those words varied (Experiment 1) or not (Experiment 2). Broadly speaking, the results from Experiments 1 and 2 are consistent with other research that suggests that temporal properties (i.e., rate) of speech can influence the perception of phonetic categories (e.g., Miller, Aibel, & Green, 1984; Miller & Dexter, 1988; Miller & Grosjean, 1981; Miller & Volaitis, 1989; Wayland, Miller, & Volaitis, 1994). Thus, it is not implausible that the word 'bat' may be misperceived as the word 'pat' depending on the duration with which the word is produced.

It is important to note here that the variable that was manipulated in the present experiments was the duration of the words, not the "intelligibility" of the words. Intelligibility is a psychological, subjective assessment of a spoken word based on multiple dimensions. One of those dimensions is word duration (cf. Haan, 1977), however other dimensions include vowel dispersion (Bond & Moore, 1994) and gender (Bradlow, Torretta, & Pisoni, 1996), to name only a few. Although intelligibility is often assessed with the dependent measure of accuracy (or error) rates, one may just as easily use a subjective rating scale to assess intelligibility (e.g., Clarke, 1960; De Bodt, Hernandez-Diaz-Huici, & Van De Heyning, 2002; Metz & Schiavetti, 1994; Tye-Murray, Barkmeier, & Folkins, 1991). It is important not to confuse the construct with the measurement; intelligibility does not necessarily equal accurate identification. For a simple illustration of this point one need only use any commercial speech recognition product; utterances rated as intelligible by humans may still not be correctly recognized by computer software. The opposite situation—accurate identification of stimuli that are not perfectly intelligible—is one of the hallmarks of human performance.

The results of the current studies provided important insight into the processing of spoken language, and emphasized the importance of using multiple research methods (e.g., naturalistic observation and laboratory experiments) to investigate psychological phenomena. The slip of the ear analysis and the experiments that followed join a large body of literature that suggests that frequency acts as a bias in the word recognition system (Broadbent, 1967; Catlin, 1969; Goldiamond & Hawkins, 1958; Luce & Pisoni, 1998; Newbigging, 1961; Pollack, Rubenstein, & Decker, 1960; Savin, 1963; Solomon & Postman, 1952; Triesman, 1971, 1978a, b). *Ceteris paribus* the word recognition system is biased to select an item that has a high frequency of occurrence, even if no stimulus is actually presented (Goldiamond & Hawkins, 1958). However, when a difference in word frequency occurs along with variation in phonological similarity (Luce & Pisoni, 1998), word duration (Wright, 1979), and so forth, the traditional processing advantage for high

frequency words may be attenuated (or indeed reversed as demonstrated in the corpus analysis and Experiment 1).

The trade-off between word frequency and word duration observed in the present study may be similar to other trade-offs observed in speech research (e.g., Denes, 1955) and in psychology in general (e.g., trade-offs between speed-and-accuracy, risks-and-gains, etc.). Models of spoken word recognition (as well as models of speech production and other cognitive processes in general) must properly account for the influence of frequency-based biases. Cognitive models that treat frequency as a static, inherent component of the activation or threshold levels of words (e.g., McClelland & Elman, 1986; cf. Luce & Pisoni, 1998) may not be able to account for the malleability of frequency effects observed in the present study. By using data obtained under naturalistic and laboratory conditions, we can better guide the development of models of spoken language processing.

*Received: August 31, 2001; revised manuscript received: December 18, 2002;
accepted: February 10, 2003*

References

- BAARS, B. J., MOTLEY, M. T., & MACKAY, D. (1975). Output editing for lexical status from artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, **14**, 382–391.
- BARD, E. G., & ANDERSON, A. H. (1994). The unintelligibility of speech to children: Effects of referent availability. *Journal of Child Language*, **21**, 623–648.
- BOCK, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin & Review*, **3**, 395–421.
- BOND, Z. S. (1999). *Slips of the ear: Errors in the perception of casual conversation*. New York: Academic Press.
- BOND, Z. S., & GARNES, S. (1980). Misperceptions of fluent speech. In R. A. Cole (Ed.), *Perception and production of fluent speech*. (pp. 115–132.) Hillsdale, NJ: Lawrence Erlbaum Associates.
- BOND, Z. S., & MOORE, T. (1994). A note on acoustic-phonetics of inadvertently clear speech. *Speech Communication*, **14**, 325–337.
- BOND, Z. S., & ROBEY, R. R. (1983). The phonetic structure of errors in the perception of fluent speech. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (pp. 249–283). New York: Academic Press.
- BRADLOW, A. R., TORRETTA, G. M., & PISONI, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, **20**, 255–272.
- BROADBENT, D. E. (1967). Word-frequency effect and response bias. *Psychological Review*, **74**, 1–15.
- BROWN, C. R., & RUBENSTEIN, H. (1961). Test of response bias explanation of word-frequency effect. *Science*, **133**, 280–281.
- CARROLL, J. (2002). Still Mondegreened, still loving it. *San Francisco Chronicle*, August 29, 2002.
- CATLIN, J. (1969). On the word-frequency effect. *Psychological Review*, **76**, 504–506.
- CLARKE, F. R. (1960). Confidence ratings, second-choice responses, and confusion matrices in intelligibility tests. *Journal of the Acoustical Society of America*, **32**, 35–46.

- COHEN, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior*, **15**, 261–262.
- COHEN, J. D., MACWHINNEY, B., FLATT, M., & PROVOST, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, **25**, 257–271.
- CUTLER, A. (1982). *Slips of the tongue and language production*. Berlin: Mouton.
- CUTLER, A., & BUTTERFIELD, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory & Language*, **31**, 218–236.
- CUTLER, A., & CARTER, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, **2**, 133–142.
- CUTLER, A., & NORRIS, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113–121.
- DE-BODT, M. S., HERNANDEZ-DIAZ-HUICI, M. E., & VAN-DE-HEYNING, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, **35**, 283–292.
- DENES, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, **27**, 761–764.
- DRAGER, K. D. R., & REICHLE, J. E. (2001). Effects of discourse context on the intelligibility of synthesized speech for young adult and older adult listeners: Applications for AAC. *Journal of Speech, Language, and Hearing Research*, **44**, 1052–1057.
- DYLAN, B. (1962). Blowin' in the wind. On *The Freewheelin' Bob Dylan* [CD]. New York: Columbia Records (1963).
- EDWARDS, G. (1998). *Deck the halls with Buddy Holly: And other misheard Christmas lyrics*. Harper Perennial.
- FAY, D., & CUTLER, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, **8**, 505–520.
- FOWLER, C. A., & HOUSUM, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, **26**, 489–504.
- FROMKIN, V. (1980). *Errors in linguistic performance: Slips of the tongue, ear, pen and hand*. New York: Academic Press.
- GAGNE, J. P., STELMACOVICH, P., & YOVETICH, W. (1991). Reactions to requests for clarification used by hearing-impaired individuals. *The Volta Review*, **93**, 129–143.
- GARNES, S., & BOND, Z. S. (1980). A slip of the ear: A snip of the ear? A slip of the year? In V. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen and hand* (pp. 231–239). New York: Academic Press.
- GOH, W. D., & PISONI, D. B. (in press). Effects of lexical competition on immediate memory span for spoken words. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*.
- GOLDIAMDON, I., & HAWKINS, W. F. (1958). Vexiersuch: The logarithmic relationship between word-frequency and recognition obtained in the absence of stimulus words. *Journal of Experimental Psychology*, **56**, 457–463.
- GROSJEAN, F. (1980). Spoken word recognition and the gating paradigm. *Perception and Psychophysics*, **28**, 267–283.
- HAAN, H. J. (1977). A speech-rate intelligibility threshold for speeded and time-compressed connected speech. *Perception and Psychophysics*, **22**, 366–372.
- HARLEY, T. A., & BOWN, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighborhood effects in speech production. *British Journal of Psychology*, **89**, 151–174.
- HENDRIX, J. (1967). Purple Haze. On *Are you experienced?* [CD]. New York: Experience Hendrix/MCA Records.

- HINO, Y., & LUPKER, S. J. (2000). Effects of word frequency and spelling-to-sound regularity in naming with and without preceding lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, **26**, 166–183.
- KEPPEL, G. (1976). Words as random variables. *Journal of Verbal Learning and Verbal Behavior*, **15**, 263–265.
- KIRK, K. I., PISONI, D. B., & MIYAMOTO, R. C. (1997). Effects of stimulus variability on speech perception in listeners with hearing impairment. *Journal of Speech and Hearing Research*, **40**, 1395–1405.
- KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- LANDAUER, T. K., & STREETER, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, **12**, 119–131.
- LEVELT, W. J. M., ROELOFS, A., & MEYER, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, **22**, 1–38.
- LIEBERMAN, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, **6**, 172–187.
- LUCE, P. A., GOLDINGER, S. D., AUER, E. T., Jr., & VITEVITCH, M. S. (2000). Phonetic priming effects in spoken word shadowing. *Perception and Psychophysics*, **62**, 615–625.
- LUCE, P. A., & PISONI, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, **19**, 1–36.
- 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*. (pp. 142–147) Cambridge, MA: MIT Press.
- MACKAY, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology*, **3**, 210–227.
- MANDLER, G., GOODMAN, G. O., & WILKES-GIBBS, D. L. (1982). The word-frequency paradox in recognition. *Memory & Cognition*, **10**, 33–42.
- MARSLÉN-WILSON, W. D., & ZWITSERLOOD, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, **15**, 576–585.
- McCLELLAND, J. L., & ELMAN, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1–86.
- McCLELLAND, J. L., RUMELHART, D. E., & HINTON, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Volume 1: Foundations. MIT Press. (pp. 3–44.)
- METZ, D. E., & SCHIAVETTI, N. (1994). Current and future directions in research on speech intelligibility assessment of persons who are deaf. *Journal of the Academy of Rehabilitative Audiology*, **27**, 237–249.
- MILLER, J. L., AIBEL, I. L., & GREEN, K. (1984). On the nature of rate-dependent processing during phonetic perception. *Perception and Psychophysics*, **35**, 5–15.
- MILLER, J. L., & DEXTER, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 369–378.
- MILLER, J. L., & GROSJEAN, F. (1981). How the components of speaking rate influence perception of phonetic segments. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 208–215.
- MILLER, G. A., & NICELY, P. E. (1955). Analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338–353.

- MILLER, J. L., & VOLAITIS, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, **46**, 505–512.
- MOON, S. J., & LINDBLOM, B. (1994). Interaction between Duration, Context, and Speaking Style in English Stressed Vowels. *The Journal of the Acoustical Society of America*, **96**, 40–55.
- NEWBIGGING, P. L. (1961). The perceptual redintegration of frequent and infrequent words. *Canadian Journal of Psychology*, **15**, 123–132.
- NORMAN, D. A. (1981). Categorization of action slips. *Psychological Review*, **88**, 1–15.
- NUSBAUM, H. C., PISONI, D. B., & DAVIS, C. K. (1984). *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words*. Research on Speech Perception, Progress Report no. 10. Speech Research Laboratory, Psychology Department, Indiana University, Bloomington, Indiana.
- OLDFIELD, R. C., & WINGFIELD, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, **17**, 273–281.
- PICHENY, M., DURLACH, N., & BRAIDA, L. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech, *Journal of Speech and Hearing Research*, **29**, 434–446.
- POLLACK, I., RUBENSTEIN, H., & DECKER, L. (1960). Intelligibility of known and unknown message sets. *Journal of Acoustical Society of America*, **31**, 273–259.
- RAAIJMAKERS, J. G. W., SCHRIJNEMAKERS, J. M. C., & GREMMEN, F. (1999). How to deal with “The language-as-fixed-effect fallacy”; Common misconceptions and alternative solutions. *Journal of Memory and Language*, **41**, 416–426.
- ROODENRYS, S., HULME, C., LETHBRIDGE, A., HINTON, M., & NIMMO, L. M. (2002). Word frequency and phonological neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28**, 1019–1034.
- SAVIN, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of Acoustical Society of America*, **35**, 200–206.
- SHATTUCK-HUFNAGEL, S., & KLATT, D. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, **18**, 41–55.
- SMITH, J. E. K. (1976). The assuming-will-make-it-so fallacy. *Journal of Verbal Learning and Verbal Behavior*, **15**, 262–263.
- SOLOMON, R. L., & POSTMAN, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, **43**, 195–201.
- SOMMERS, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology and Aging*, **11**, 333–341.
- SOMMERS, M. S., & DANIELSON, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context. *Psychology and Aging*, **14**, 458–472.
- STEMBERGER, J. P. (1992). The reliability and replicability of naturalistic speech error data: A comparison with experimentally induced errors. In B. J. Baars (Ed.), *Experimental slips and human error: Exploring the architecture of volition*. New York: Plenum Press. (pp. 195–215.)
- TRIESMAN, M. (1971). On the word frequency effect: Comments on the papers by J. Catlin and L. H. Nakatani. *Psychological Review*, **78**, 420–425.
- TRIESMAN, M. (1978a). A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review*, **85**, 525–570.
- TRIESMAN, M. (1978b). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior*, **17**, 37–59.
- TYE-MURRAY, N. (1991). Repair strategy usage by hearing-impaired adults and changes following communication therapy. *Journal of Speech and Hearing Research*, **34**, 921–928.
- TYE-MURRAY, N., BARKMEIER, J., & FOLKINS, J. W. (1991). Scaling and transcription measures of intelligibility for populations with disordered speech. *Journal of Speech and Hearing Research*, **34**, 697–699.

- TYE-MURRAY, N., KNUTSON, J. F., & LEMKE, J. H. (1993). Assessment of communication strategies use: Questionnaires and daily diaries. *Seminars in Hearing*, **14**, 338–353.
- TYE-MURRAY, N., & WITT, S. (1997). Communication strategies training. *Seminars in Hearing*, **18**, 153–165.
- TYE-MURRAY, N., WITT, S., & CASTELLOE, J. (1996). Initial evaluation of an interactive test of sentence gist recognition. *Journal of the American Academy of Audiology*, **7**, 396–405.
- Van BERGEM, D. R. (1995). Perceptual and acoustic aspects of lexical vowel reduction, a sound change in progress. *Speech Communication*, **16**, 329–358.
- VITEVITCH, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, **40**, 211–228.
- VITEVITCH, M. S. (2002a). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28**, 735–747.
- VITEVITCH, M. S. (2002b). The influence of onset-density on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **28**, 270–278.
- VITEVITCH, M. S., & LUCE, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, **9**, 325–329.
- VITEVITCH, M. S., & LUCE, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, **40**, 374–408.
- VITEVITCH, M. S., & SOMMERS, M. (in press). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*.
- VOSS, B. (1984). *Slips of the ear: Investigations into the speech perception behavior of German speakers of English*. Tübingen, Germany: Gunter Narr Verlag.
- WAYLAND, S. C., MILLER, J. L., & VOLAITIS, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America*, **95**, 2694–2701.
- WIKE, E. L., & CHURCH, J. D. (1976). Comments on Clark's "The language-as-fixed-effect-fallacy." *Journal of Verbal Learning and Verbal Behavior*, **15**, 249–255.
- WRIGHT, C. E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition*, **7**, 411–419.
- ZIPF, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. New York: Houghton Mifflin.
-

Appendix A

The 88 slip of the ear tokens used in the reported analyses

<i>Actual Utterance</i>	<i>Perceived Utterance</i>	<i>Actual Utterance</i>	<i>Perceived Utterance</i>
bike	back	death	deaf
wet	white	guide	guy
math	mouth	he	she
beach	bitch	cable	table
bell	bill	street	string
cattle	kettle	grape	grate
stir	store	raise	race
shirt	short	move	mood
king	can	plant	plan
and	in	savor	sabre
system	sister	hat	cat
went	want	lift	list
better	bitter	fry	fly
barn	born	cart	car
grass	grasp	drape	grape
cool	cruel	wrong	long
van	man	coke	coat
teak	teeth	cup	cuff
porpoise	corpus	humid	human
node	nose	text	test
ear	rear	van	fan
hall	whore	long	lawn
honor	otter	move	moo
insufficient	inefficient	life	lie
lime	line	he	she
fritter	critter	trench	french
nasal	naval	pace	face
snip	sniff	ear	year
slip	snip	ship	ship
rice	ice	wrong	long
internal	eternal	car	card
coke	coat	deaf	death
air	hair	trap	track
traitor	trader	train	tray
trap	track	league	leave
breath	breed	corpus	porpoise
apple	ample	their	air
poor	whore	service	circus
grew	threw	cap	cat
part	park	who	goo
noon	nude	class	glass
cook	hook	air	hair
witch	wish	plain	play
fad	bad	face	mace

Appendix B

Words varying in frequency of occurrence used in Experiments 1 and 2

<i>High Frequency of Occurrence</i>	<i>Low Frequency of Occurrence</i>
beam	bib
boil	bead
book	bean
dash	dial
doll	dime
dose	duck
fog	fade
foam	fern
fine	fuss
cape	cough
case	cage
come	cove
lease	lamb
loop	ledge
lime	lace
path	peach
page	pig
pipe	pearl
rope	robe
rose	rhyme
rice	rung
seam	sock
cease	soak
suit	soothe
tape	tease
tip	toad
tough	tug
wall	web
wet	wig
wire	worm
