

**RESEARCH ON SPOKEN LANGUAGE PROCESSING**  
Progress Report No. 22 (1998)  
*Indiana University*

**Subjective Familiarity of Words: Analysis of the Hoosier Mental Lexicon<sup>1</sup>**

**Nathan R. Large<sup>2</sup> and David B. Pisoni**

*Speech Research Laboratory  
Department of Psychology  
Indiana University  
Bloomington, Indiana 47405*

---

<sup>1</sup> This work is supported by NIH-NIDCD Research Grant DC00111 to Indiana University.

<sup>2</sup> Now at the Department of Psychology, University at Buffalo, Buffalo, NY.

## Subjective Familiarity of Words: Analysis of the Hoosier Mental Lexicon

**Abstract.** We conducted a statistical analysis of several subsets of words from the Hoosier Mental Lexicon in order to examine some factors underlying the subjective familiarity ratings collected by Nusbaum, Pisoni, and Davis (1984). In this analysis, we grouped words into High-FAM (average familiarity rating greater than 6 on a 7-point scale), Mid-FAM (between 4.5 and 3.5), or Low-FAM (less than 2) sets, with primary interest being placed on the small set of Low-familiarity words (502 total), which were unknown to the majority of native listeners. This Low-FAM set consisted mainly of non-English borrowings, technical terms, and archaic words. These items received much higher ratings from Speech Research Laboratory staff, who we assume have greater language experience, showing that the familiarity scale employed is not continuous for very low ratings, especially in terms of the properties of Low-FAM words. As expected, High-FAM words were the majority and had greater frequency of occurrence, greater neighborhood density and frequency, and were shorter on average than Mid or Low words. However, when these sets of words were controlled for length, we found that High- and Mid-FAM words did not differ significantly in neighborhood density and frequency, though High-FAM words were still significantly more frequent on average than Mid-FAM words, which were only slightly more frequent than the Low-FAM items. With word length equated, Low-FAM words still came from much sparser and lower-frequency neighborhoods. This finding indicates that the difference in neighborhood density between High- and Mid-FAM items is due primarily to the generally shorter length of common, highly familiar words, whereas the lower density for Low-FAM words is genuinely due to differences in constituent structure. Words with Mid-range familiarity appear to comprise low frequency, mid- to high-density lexical items, which are most difficult to recognize according to the neighborhood activation model of spoken word recognition.

### Introduction

“Frequency effects” are integral to any processing account of language perception or production. The relative frequency of words within a language influences how quickly and accurately those words can be recognized, classified, repeated, or recalled from memory (for a review of previous research and theory, including interactions of frequency with other lexical variables, see Lively, Pisoni, & Goldinger, 1994; Marslen-Wilson, 1984; and Pisoni, Nusbaum, Luce, & Slowiaczek, 1985). As with other psychological studies of perceived frequency, the *magnitude* of a lexical entry’s relative frequency, as a log of the number of occurrences in a particular corpus, has been shown to be a better predictor than linear frequency. But although frequency is well established as an important variable, many important questions remain about its role: is frequency information actually stored directly in the mental lexicon, as an attribute of each lexical item? If so, is this value stored as a relative value or weight (i.e., the “counter in the head”), or does it arise from the number of (episodic) traces associated with each entry in the lexicon? Proposed alternatives to the common lexical-variable assumption include the possibility that “estimated frequency” is, in part or whole, computed from the similarity of an item’s sound-structure to that of other words in the lexicon (Eukel, 1980), or that words are stored in memory by frequency, so that a word’s location in a lexical “file” indicates its relative frequency of occurrence (Forster, 1978). All of these theories and the experimental results that lead to them, rest on a necessarily abstract idea of frequency, typically computed from a count of words in text (though word counts from spoken language have become available, notably the Celex database, Centre for Lexical Information, 1993). The

assumption, then, is that speakers will not differ much in the relative proportion of words they consider highly frequent, moderately frequent, somewhat rare, and so forth, no matter how they are storing this frequency information. This seems to imply that the estimated familiarity of listeners with the words in their native language will also be comparable. This assumption obviously cannot be intended to apply to listeners of differing linguistic backgrounds, since education or environment may cause significant differences in individual familiarity with linguistic items, both by increasing the frequency with which a word is encountered, as well as providing semantic information that makes the word more useful, more strongly active in memory, or simply more likely to be noticed later (attention-related). Thus, text counts cannot apply equally well to different populations that differ in their linguistic backgrounds. However, concern has also arisen from the fact words that have low frequency, as measured by number of occurrences in text, may nonetheless be quite familiar to listeners; a common example is *violin*. Although the reverse case, an unfamiliar yet highly frequent item, is unlikely, there may still be significant differences in the proportion of words used regularly in one corpus that are well-known to all speakers. If *subjective* familiarity, and not objective frequency, is the source of “frequency effects” in word recognition (and this seems entirely plausible), then any discrepancy between objective frequency and subjective familiarity should be evident when both are compared to task performance. Evidence for such a dissociation has been reported by Gernsbacher (1984), who showed that subjective familiarity ratings collected from native speakers were better predictors of naming performance (by another group of speakers) than a published set of frequency counts. Recently, Chalmers, Humphreys and Dennis (1997) have shown that perceived frequency can be manipulated by training. They demonstrated that normally rare technical (computer science) terms were treated like high-frequency items in a recognition memory task by listeners who had become familiar with those terms (computer science majors). These findings also imply that items like *violin*, that are used in highly specific, relatively rare situations, may still be highly familiar to many listeners, and show effects normally attributed to high-frequency words.

The Hoosier Mental Lexicon was created by Nusbaum, Pisoni, and Davis (1984) to examine the relationship between word frequency, as measured by published norms such as Thorndike and Lorge (1944) and Kucera and Francis (1967), and subjective ratings of word familiarity collected from native speakers of English. In addition to determining how closely related these two measures were, the authors also wanted to provide a large database of word familiarity ratings for future use, as well as use the relative proportions of words at each familiarity level to produce an estimate of native speakers’ lexicon size. Familiarity was expressed by each participant using a 7-point scale, ranging from a low rating of “1,” indicating that a word was completely unknown, through “4,” meaning that a word was recognized but the meaning unknown, to a maximum of “7,” indicating that a word was very well known, that is, highly familiar. The large corpus of words encompassed 19,750 words from Webster’s Pocket Dictionary, and was divided into smaller groups (see original paper for details); each word was rated by 12 participants in all, and these ratings were then averaged to give that word’s familiarity score. Only 11,750 of these words were also in the Kucera and Francis (1967) corpus of text, and only these items (with frequency  $\geq 1$ ) were used in the comparison between average familiarity and frequency counts. The remaining items were still included in the HML with frequency = 0.

For these 11,750 words, the correlation between mean familiarity ratings and log frequency of words was .43 ( $p < .01$ ), a small but still significant relationship. Nusbaum, Pisoni, and Davis note that this is a surprisingly low correlation and that log text frequency manages to account for only 17.5% of the variance in the familiarity ratings. They divided this set of words into three groups based on average familiarity rating: (a) words with FAM between 7 and 5.5 were considered “familiar,” (b) FAM scores between 2.5 and 5.5 indicated that a word was “recognized,” and (c) words with FAM below 2.5 were considered “unknown.” These groups were displayed by the log number of words from each group that fell into a particular range of log word frequency values (increments of .5). This analysis revealed that no “unknown” word had log frequency greater than 4 (the highest was 55 occurrences per million words),

and no “recognized” word was higher than 5.5 (maximum of 245 occurrences per million words). “Familiar” words continued along the frequency range to its maximum possible. Clearly, words selected from these different ranges would, on average, differ in their frequency in text, but the difference was much greater between the “familiar” words and the other two groups than between “recognized” and “unknown.” The “unknown” words were most sharply limited in average frequency, with the majority well below three occurrences per million, whereas the other two groups showed a more gradual decrement in number of words at each frequency level. Added to this analysis was a comparison of the percentage of total familiarity rating responses across all subjects and stimuli. This latter analysis showed that fully 60.7% of the responses made were “7,” or “very familiar” responses, and 73% of total responses were 5 or greater. Although this meant that most of the words in Webster’s Pocket Dictionary were considered “familiar,” as might be expected, Nusbaum et al. were more concerned that the percentage was not greater. If only ~14,400 of 19,750 words could be considered “familiar,” then prior estimates of the size of native speakers, lexicons, sometimes as high as 200,000 words (Hartman, 1946) appeared excessive. They concluded that prior estimates of lexical size have been misled by the use of smaller sample word sets, and that unless another large-corpus study produced results differing from the HML study, estimates of average lexicon size should be scaled down.

Also included in the HML corpus was information about the similarity neighborhood of each word (see Luce, 1986; Luce & Pisoni, 1998; Luce, Pisoni & Goldinger, 1990 for accounts of neighborhood effects on spoken word recognition). Of interest here were the variables Neighborhood Density, or the number of words that are similar to a target word using a one-phoneme substitution metric, and Neighborhood Frequency, expressed as the average of the log-weighted frequencies of the target word’s neighbors as determined by the same similarity metric. Data were provided in the HML for two methods of computing neighborhood metrics: Method A counts only the number of items that could be generated from the target word by the replacement of exactly one phoneme, for example, *cap* /kap/ compared to *cat* /kat/. Method B, in addition, counts as neighbors all items that could be generated by adding or deleting one phoneme from the target word, for example, *at* /at/ and *scat* /skat/ would also be allowed as neighbors of *cat* /kat/. The neighborhood frequency value for a target word is expressed as the average of the log frequencies of that word’s neighbors, as determined by either method. Thus, Density B will always be greater to or equal to Density A for any given item, though Log Frequency B may be greater or smaller depending on the additional neighbors’ frequencies. However, no analyses were ever conducted to examine how differences in familiarity might be related to differences in these neighborhood properties.

As previously mentioned, Eukel (1980) proposed that listeners might be able to estimate relative frequency of words based on their similarity to other words. He collected behavioral data using nonword stimuli that were estimated for “expected frequency” by listeners according to their distance from actual English words. His results could be explained in terms of the phonotactic regularity of rated words; the sounds and patterns present in more common words tend to be more frequent themselves (Landauer & Streeter, 1973), and so words containing those sounds and patterns would be rated as more frequent. These findings yield ratings with a good degree of accuracy, even if a target word or non-word pattern is unfamiliar to the participant(s). Thus, we need to examine the relationship not only between objective frequency and subjective familiarity, but also between these variables and the relative similarity of words, their neighborhood characteristics. It may be that these variables are all closely related, in which case our task is to determine which sources of information are actually being used, both in these subjective familiarity rating tasks and in the actual perception of spoken words. As a first step, we conducted an analysis of the HML corpus, both to extend the original descriptive work in Nusbaum et al. (1984), and to examine the effects of neighborhood Density and Frequency in this analysis.

We were especially interested in the “recognized” and “unknown” items in the HML for several reasons. First, these items are apparently much less likely than “frequent” items to be in an individual’s mental lexicon, and thus their ratings will differ widely between individuals. Second, this gap in familiarity may be due to differences in exposure to these words (low vs. high frequency), but it might also be due to structural factors that make these words less “learnable” than other more “familiar” words; that is, words with more similarity to already learned words might be easier to remember. Third, the difference in familiarity ratings may be due to a perceptual bias, such that words in less dense neighborhoods (with few or no neighbors) are thought of as less common even though they may not actually differ in objective frequency from words with a few more neighbors or more common neighbors. To determine which of these might be possible, we divided the HML corpus more narrowly (FAM score range = 1) into subgroups and studied the distribution of word frequency along with neighborhood density and frequency across these subgroups. Item length was explored separately but later became an important consideration and these results are also included in this analysis.

### Analysis I: High, Mid, Low

#### Method

Because our interest was on Very Low Familiarity words, the first subgroup of words isolated from the HML corpus consisted of all items with Familiarity ratings (FAM scores) less than two. These words should be the items to which a majority of listeners responded with a rating of one or two, indicating that those items were completely or nearly unknown to them. This group of words consisted entirely of content (as opposed to function) words, and totaled 523 items out of 19,750, or 2.6% of the total lexicon examined. Twenty-one items (proper names and abbreviations) were removed from this set on the grounds that these items were obviously idiosyncratic, either known well or not known at all depending on a listener’s background (e.g., Algerian, TNT) leaving 502 items. This was called the Low-FAM Word group. Next, to contrast with this group, two more subgroups of words were selected from the top and exact center of the familiarity scale, from a range proportionate to the one used to select Low-FAM items ( $1 \leq \text{FAM} < 2$ ). All items with FAM Scores greater than six were labeled High-FAM. This group comprised 11,031 words (55.9% of the total lexicon), 358 of which were removed as abbreviations, proper names, or function words. Function words were removed because of their disproportionate frequency counts (~65,000 per million words of text for *the* alone), and because evidence indicates that these words are processed and stored in a different fashion from other, “content,” words (Bradley, 1978; Bradley & Garrett, 1983; but see Chiarello & Nuding, 1987; Besner, 1988; see also Surprenant, et al., 1998, for data on proper names as statistically unique lexical items). This procedure left us with 10,673 words in the High-FAM group. The Mid-FAM group consisted of items in the very center of the FAM scale with ratings from 3.5 to 4.5 inclusive. This group of words initially contained 1,990 entries (10%) of which 44 were discarded as abbreviations, proper names, or function words, leaving 1,946 words.

#### Results and Discussion

The HML provided descriptive statistics for each word and these values are summarized as means and standard deviations for each descriptor across all accepted words in each group. Table 1 presents a summary of these results for each of the three groups of words. “Length” represents the average length of words in that group measured by number of phonemes, determined from a phonetic transcription of each item. Low- and Mid-FAM items did not differ in Length (6.66 vs. 6.68 phonemes) but High-FAM items (6.1 phonemes) were significantly shorter than either the Mid-FAM or Low-FAM words as determined by an unpaired  $t$  test between groups,  $t(12617) = -9.63$ ,  $p < .001$ . High and Mid item Length,  $t(11173) = -4.99$  between High- and Low-FAM item Length, respectively.

Raw frequency counts, based on values given in the Kucera and Francis corpus (1967), derived from number of occurrences of the target in one million words of printed text followed a similar pattern. Low- and Mid-FAM items showed no difference in frequency  $t(2446) = 0.997$ , n.s., but High-FAM items were much more frequent than either Mid- (H/M:  $t(12617) = .44$ ,  $p < .001$ ) or Low-FAM words (H/L:  $t(11173) = 3.9$ ,  $p < .001$ ). When the log of these frequency counts was used (Log Frequency), High-FAM items were still greater in relative frequency, on average, than Mid- (H/M:  $t(12617) = 41.06$ ,  $p < .001$ ) or Low-FAM items (H/L:  $t(11173) = 22.89$ ,  $p < .001$ ). However, the difference between Mid- and Low-FAM item frequencies was now significant using the log transformation  $t(2446) = 6.09$ ,  $p < .001$ . However, the difference in mean log frequencies for these groups is extremely small when compared to the increase from Mid- to High-FAM items, .06 versus .67.

The average neighborhood characteristics of words in each group were also computed and compared using the metrics of Neighborhood Density (listed here as Density A and B) and Log Neighborhood Frequency (Log Frequency A and B). Unless some qualitative difference exists in the analysis of these factors, significance levels will be provided only for Method B for the sake of brevity. For the three groups, neighborhood Density appeared to increase by a factor of 2 for each higher FAM group and all pair-wise differences were statistically significant: (H/M:  $t(12617) = 11.54$ ,  $p < .001$ ; M/L:  $t(2446) = 4.66$ ,  $p < .001$ ; H/L:  $t(11173) = 9.41$ ,  $p < .001$ ). Log Frequency also increased steadily with FAM group. All differences between groups were significant: (H/M:  $t(12617) = 11.95$ ,  $p < .001$ ; M/L:  $t(2446) = 5.54$ ,  $p < .001$ ; H/L:  $t(11173) = 11.94$ ,  $p < .001$ ).

Table 1

## Descriptive Statistics for High-, Mid-, and Low-FAM Words in the HML.

	High	Mid	Low
<i>N</i>	10673	1946	502
Familiarity	6.79 ( <i>SD</i> = 0.268)	4.00 (0.310)	1.55 (0.299)
Length	6.13 (2.35)	6.68 (2.23)	6.66 (2.23)
Frequency	30.60 (169.33)	2.00 (21.23)	1.05 (0.285)
Log Frequency	1.74 (0.715)	1.07 (0.217)	1.01 (0.072)
Density A	3.22 (5.97)	1.67 (4.09)	.829 (2.77)
Density B	4.08 (7.14)	2.14 (4.87)	1.07 (3.20)
Log Frequency A	0.799 (0.939)	0.550 (0.846)	0.361 (0.723)
Log Frequency B	0.945 (0.961)	0.665 (0.887)	0.425 (0.778)

None of these results are surprising, particularly in light of past studies such as Landauer and Streeter's (1973) analysis of a text lexicon. Their analysis indicated that high-frequency words tended to have more similar items than low-frequency items and that these "neighbors" tended to be of high frequency themselves. With the small but significant relationship between frequency and familiarity present in the HML, we should expect comparable results with our FAM groupings. This analysis does add a median group, however, and further emphasizes that although density appears to increase steadily average item frequency and familiarity increase very rapidly as the items in the most dense neighborhoods of the lexicon are reached.

In carrying out this analysis, we noted almost immediately that the difference in phoneme length observed for High-FAM words versus the other two groups contributes significantly to the observed differences in neighborhood characteristics among the three groups of words. Computationally, it follows that as items in a set are reduced in number of features (fewer phonemes), more items will be closely similar to a given item (similarity neighborhoods will increase). Indeed, the correlation between word length and density was consistently high across groups. Length and density B were negatively correlated ( $r(10673) = -.62$ ) for High-FAM items by a Pearson's  $r$  score and subsequent  $r$ -to- $z$  transformation, indicating that as length decreased, density increased. This correlation was  $-.56$  for Mid-FAM words and  $-.45$  for Low-FAM words. All correlations were significant at  $p < .001$ . Thus, differences in length among the three groups were contributing to the differences in density. However, there was no significant difference in average length between the Mid- and Low-FAM groups, so length may be only a contributing factor for the High-FAM items' density mean. A second analysis was designed to remove the influence of mean length (in phonemes) on average density to see how much this factor was contributing to differences in density between groups.

## Analysis II: High, Mid, Low Equated for Length – 500 word sets

### Method

To balance the collected FAM groups for average word length in phonemes, we decided to make the smallest group, Low-FAM words, the standard and then randomly select items from the other two groups in order to obtain equal Ns with equivalent proportions of words at each phoneme length. One Low-FAM word, *antivivisectionist*, was discarded as an outlier on the length distribution, with 17 phonemes. A second item, *baht* /bat/ was removed because it was phonetically identical to *bought*, a much more common and familiar English word; it was also the Low-FAM item with greatest density (34 word [Method B]), but was not an outlier along the distribution of item density scores. We were left with 500 words, divided by phoneme length in the following proportions: (3 words of 2 phonemes, 21 of 3 phonemes, 58 of 4, 89 of 5, 83 of 6, 86 of 7, 57 of 8, 46 of 9, 32 of 10, 14 of 11, 9 of 12, and 2 of 3 phonemes in length;  $M = 6.65$  phonemes over 500 words). Equal numbers of words from each of the other two groups were selected to yield High- Mid-FAM sets of 500 words each. These sets were then analyzed in the same manner as before. Summaries of the descriptive statistics are provided in Table 2. At the time of this publication, the two remaining FAM “groups” had also been separated and reduced to 500-word sets with equivalent proportions of words at each length. These are High-Mid-FAM, representing all items with FAM between 6 and 4.5 ( $6 \geq \text{FAM} > 4.5$ ; 3581 items) and Low-Mid-FAM, all words with FAM between 3.5 and 2 ( $3.5 > \text{FAM} \geq 2$ ; 2131 items). We originally intended to make all the groupings of equivalent range; however, too few items remained to further divide these Mid-groupings. In a future analysis, we may want to lump all three Mid-groups together as was done in Nusbaum et al. (1987) when comparing familiarity levels to frequency ranges. Although summations of the two new full sets were not yet available, and neither the full nor 500-word versions of these two sets have been analyzed for significant differences, statistics for the 500-word sets have been added to Table 2 to provide a comparison with the previously compiled data.

**Table 2**

**Descriptive Statistics for High, High-Mid, Mid, Low-Mid and Low-FAM Words,  
Groups Balanced for Word Length  
(in phonemes; Length = 6.65 phonemes and N = 500 words for all groups).**

	High	High-Mid	Mid	Low-Mid	Low
Familiarity	6.78 ( <i>SD</i> = .271)	5.33 (.436)	4.01 (.315)	2.78 (.419)	1.55 (.299)
Frequency	20.92 (58.80)	1.94 (2.77)	1.69 (4.95)	1.27 (1.85)	1.05 (.285)
Log Frequency	1.67 (.681)	1.15 (.280)	1.07 (.227)	1.04 (.156)	1.01 (.072)
Density A	1.70 (3.95)	1.50 (3.70)	1.55 (3.86)	1.35 (3.60)	.770 (2.42)
Density B	2.24 (4.75)	1.93 (4.43)	1.94 (4.52)	1.75 (4.40)	1.00 (2.85)
Log Frequency A	.586 (.854)	.536 (.836)	.579 (.870)	.557 (.872)	.358 (.719)
Log Frequency B	.732 (.909)	.644 (.879)	.693 (.911)	.621 (.877)	.422 (.774)

## Results and Discussion

As anticipated, the creation of the length-balanced stimulus sets did not have any qualitative effects on average familiarity scores or (log-) frequency counts between groups. However, it is interesting that frequency counts for the High-FAM items dropped an average of 10 occurrences per million, no doubt as a result of higher frequency counts for shorter words. An examination of the original High-FAM set indicates a small but significant inverse correlation between Log Frequency and Length for High-FAM words,  $r(10672) = -.28, p < .001$ , supporting this idea. The differences between groups were still significant, however, albeit at slightly lower  $t$  values due to the reduction in power because the  $N$ s were smaller.

Several changes in focus occurred when we began to look at neighborhood Density and Log Frequency for these length-balanced subsets. Although the relationship between Mid- and Low-FAM groups was largely unchanged, neighborhood density and frequency for words in the High-FAM group were much lower and were no longer significantly different from density and frequency values for the Mid-FAM group. This was confirmed by an unpaired  $t$  test comparing Neighborhood Density (B), H/M:  $t(998) = 1.02, p > .05$ . High- and Mid-FAM words still both came from more dense neighborhoods than Low-FAM words (H/L:  $t(998) = 4.99, p < .001$ ; M/L:  $t(998) = 3.92, p < .001$ . Mean Log Frequency of neighborhoods for High-FAM versus Mid-FAM words showed no significant difference (H/M:  $t(998) = .680, p > .05$ ), but again, High- and Mid-FAM words were both from higher-frequency neighborhoods than words in the Low-FAM group (H/L:  $t(998) = 5.80, p < .001$ ; M/L:  $t(998) = 5.06, p < .001$ ). Figures 1 and 2 display mean neighborhood Density and Log Frequency values (both Methods A and B) for each of the three main FAM groups for the full sets and then for the 500-word length-balanced sets demonstrating the change in values after length was balanced.

-----  
 Insert Figures 1 & 2 about here  
 -----





With the proportion of words at each phoneme length held constant across groups, we found that the relationship between Low-FAM words and the other groups was not changed but High-FAM and Mid-FAM words essentially became indistinguishable in terms of their neighborhood characteristics (i.e., density and average neighborhood frequency). Thus, the differences in density and neighborhood frequency observed in our first set of analyses between these groups were due almost entirely to the correlation between length and density and the tendency of short words to receive high FAM ratings (or for High-FAM words to be shorter on average). Mid-FAM words, when length is not a factor, are from equally dense and frequent neighborhoods as High-FAM words, yet their frequency (from text counts) is much lower, nearly as low as for the Low-FAM words. This pattern indicates that these words are lower on the familiarity scale due to their less common usage, rather than from bias or from perceptual difficulty due to structural properties. Perceptual difficulties, however, may be present. According to the Neighborhood Activation Model (Luce & Pisoni, 1998), these are precisely the kinds of words that are the most difficult to recognize, producing lower accuracy rates and longer response times in several behavioral tasks. The interaction of low frequency and thus lower activation or response bias, and competition from many high-frequency neighbors should make these words the most difficult to recognize out of the three groups. High-FAM words may be aided by their high frequency and Low-FAM items have little lexical competition, as long as they are at least present in the listener's mental lexicon.

### **Analysis III: Very Low-FAM Words and Linguistic Expertise**

The results obtained from the previous two analyses indicated that the Low-FAM words differed in several substantial ways from the rest of the words in Webster's Pocket Dictionary. Visual examination of these Low-FAM words and anecdotal accounts indicated that the bulk of these 500+ words were either foreign loanwords, technical terms, or archaic English words. By *foreign loanwords*, we mean that the word is a more common word in a non-English language and has been borrowed into English in the recent past. Examples from the HML include *obi*, a Japanese word for a belt or sash, and *zloty*, a Polish coin denomination. Words like these had very low neighborhood density since they originated in languages with phonetic and phonological structures that are very different from English; that is, they are low-probability phonotactic patterns in English. In fact, the z onset is impossible according to English phonotactic convention and occurs *only* in loanwords. *Technical terms* were words used by specific professions, new coinages used for technology or scientific discourse and not in common use in the language, or other jargon that might be specific to a particular environment. Examples of technical terms from the HML corpus were *eczematous* (medical) and *jingoism* (literary or political). These words would have extremely low frequency in text, usually zero, and were considered very unfamiliar by the average subject, but individuals with a particular background would be extremely familiar with these terms. The Chalmers et al. (1997) experiment using computer science students demonstrates jargon as highly familiar to those individuals, but of low familiarity otherwise. Thus, there will be little middle ground on a familiarity scale of this kind. Finally, *archaism* was used to refer to words that were definitely English but have fallen out of common use. Literate native speakers may know them to varying degrees but the average undergraduate at Indiana University may never have heard or seen the word, despite its phonotactic regularity. These words tended to have greater average neighborhood density than the other two types of words. Some notable examples of archaisms were *tress*, a strand or lock of hair and *ell*, a measurement of distance. Other archaisms occurred within the Mid-FAM group. Some of these items may be familiar to listeners from poetic, historical, or other literary sources but they are largely unused in their everyday discourse.

To provide a contrast with the original HML participants (undergraduate psychology students), as well as confirm these assumptions about relative familiarity and the different types of Low-FAM words, we presented the 502 Low-FAM words in the original set from Analysis I (with proper nouns and

abbreviations removed) to 10 individuals working in the Speech Research Laboratory (SRL). These individuals ranged from professors and students at the Ph.D. level of education to undergraduates (upperclassmen, however, in contrast to the freshman and sophomore participants in the original study). All of these participants can reasonably be assumed to have a higher level of education and linguistic expertise than the participants in the original HML study.

### Procedure

A randomized paper-copy list of the 502 Low-FAM words was presented to each of the 10 SRL personnel individually. They were requested to write down a rating for each of the words in the list corresponding to their familiarity with the given item. This was done on the participants' own time and the amount of time spent on the task was not controlled. The 7-point rating scale given to participants was identical to the one used by the original HML subjects. Participants in this task were not informed that all the items had very low HML familiarity scores but some individuals may have learned of this prior to list presentation. The list was collected from each participant and the average rating across all 10 staff members was computed for each Low-FAM item and compared with its HML FAM rating.

### Results and Discussion

Overall, the average SRL staff ratings for the 502 Low-FAM words were much higher than the HML ratings, with a mean of 3.05 versus the HML mean of 1.55. This difference was shown to be significant by a paired  $t$  test across stimulus words,  $t(998) = -18.28$ ,  $p < .001$ . When ratings for each stimulus item were directly compared, initial agreement on ratings between groups was very low, and in fact, the correlation was negative between SRL and HML ratings for the same words,  $r(501) = .19$ . However, by separating the words into groups, the negative correlation was found to be due to the fact that the items all HML participants had rated with a "1," the completely unknown words, were rated across the scale by the SRL staff, and in fact received higher SRL staff ratings than words that had originally received HML rating between 1 and 2, noninclusive. The mean SRL rating for words with mean HML ratings of "1" was 5.25, whereas the mean SRL rating for all other low-HML-FAM words was 2.77. When the words that had originally been rated "1" were removed, the correlation between SRL and HML ratings was positive and significant,  $r(56) = +.18$ ,  $p < .01$ . The basic result, that native speakers with greater educational and language backgrounds compared to the original HML participants would have higher ratings for these Low-FAM items, was obtained as predicted. In addition, these results appear to support our intuitions about the makeup of the "unknown" group. Some of these Low-FAM items are unknown to most speakers, but very well known to certain speakers with different backgrounds, either educational or linguistic. Anecdotal evidence from the SRL staff indicated that knowing the language from which a loanword originated usually placed the rating for that word within the "familiar" range. Technical terms and archaisms were more widely distributed, with some individuals knowing them well, others able to recognize but not define them, and some participants rating them as entirely unknown with no discernable pattern.

This investigation does have obvious faults as an empirical study of familiarity. Most notably, subjects were given only a subset of the words, one chosen by its low familiarity in the HML corpus. Thus, subjects may have adjusted their rating scale across the 502 items in this task in a different manner that was different from the original HML subjects who had to "reserve" the higher end of their scale for very common words. This is more likely to occur across the mid-range of the 7-point scale than toward the top, since each level of the scale is given a specific description of what that rating means, e.g., "7"- I am very familiar with this word and know at least one meaning well, versus "4"- I have seen this word before but do not know its meaning. Also, participants in this study were given the words in a complete list rather than responding to them serially on a computer terminal as was done by the original subjects.

This may have prompted comparisons between former responses and latter ones though subjects were instructed not to change a previous rating once they had written one down. Finally, subjects were not totally naïve to the expectations of this study having been recruited by fellow researchers for a study they may have previously learned about. Clearly, researchers desiring a strong investigation of the effects of expertise on lexical familiarity would have to control these factors and present a much more valid experimental paradigm. However, our goals were much more modest: to investigate the above assumptions about how different kinds of words might be rated differently depending on the listener's background and show that some differences in familiarity are entirely possible on the basis of linguistic expertise (see also Lewellen, Goldinger, Pisoni, & Greene, 1991, 1993).

## General Discussion

The present set of results present a clearer picture of the lexicon as arranged by relative familiarity. The most familiar words, High-FAM, include most function words but also most shorter words (monosyllabic or fewer than 5 phonemes). These High-FAM items tend to have very high frequency in text counts but may not all be high-frequency. They may not be in very dense lexical neighborhoods though it is true that words in the most dense neighborhoods are highly familiar. In general, High-FAM words are also highly frequent from dense and highly frequent neighborhoods and tend to be shorter than other words. All of these conclusions are well established (Zipf, 1935) and come as no surprise to anyone working on word recognition or lexical access.

Low-FAM items which were rated on average below “2” come from unique and generally limited categories which we described as recent foreign loanwords, technical terminology or jargon, and archaisms. Each of these categories is included in the HML because these words are present in some speakers' lexicons but these words are largely unknown across much larger populations for various reasons related to their origins. Comparisons can be drawn between these items and non-word stimuli. Loanwords may or may not be similar to English phonotactic structure but if they are completely unknown to a listener, these might be considered non-words of relatively high or low “wordlikeness.” The same applies to archaisms, though these tend to have higher “wordlikeness” as they have been accepted and used in the language for a long period. Technical terms, however, tend to have internal structure and lexical morphemes that distinguish them (these may be English, Greek, Latin, or others). Thus, even if a given listener has never heard a specific term, they will be more certain that it probably means something, possibly even constructing an appropriate meaning on the spot. This jargon will also tend to be more sharply divided between “those who know” and those who don't, rather than showing a graded response.

Mid-FAM items, however, represent the real point of divergence in these analyses. Some higher-frequency items are present in this middle range, but on average, words rated at the middle of the familiarity scale are uncommon in everyday use. Mid-FAM items tend to be longer words, content words, and used for more specific situations. It might also be interesting to compare the contextual properties of these items to High-FAM words, as we might expect to find that Mid-FAM words enter into fewer syntactic roles (*travel* vs. *go*, for example). Despite their lower frequency of use, Mid-FAM words do not come from less dense neighborhoods, or less frequent neighborhoods, than High-FAM words when these groups are matched for length. That is, the only reason High-FAM words occupy more dense and frequent neighborhoods in the lexicon is that there are more short High-FAM words. This presents us with a view of Mid-FAM words as relatively low-frequency words occupying similarity spaces of moderate to high density, the types of words predicted to be the most difficult to recognize by the Neighborhood Activation Model. Again, we emphasize that these groups are overlapping; some High-FAM words are of low frequency and/or density, and a number of Mid-FAM words are from either higher frequency or less dense neighborhoods (but rarely both). However, the general characteristics of words rated at these levels seem clear from our analyses.

Finally, this description of words in the Hoosier Mental Lexicon, a large corpus of lexical items, re-emphasizes that frequency, familiarity, and lexical organization by sound structure are all closely related to perceptual similarity. There are very few high-frequency words with very few neighbors. Another method of describing sound similarity, the “regularity” of a sound-pattern when compared to all other words within that language (or a native speaker’s lexicon), is tied to linguistic ideas of phonotactic structure, and more recently, probabilistic constraints on word structure (Pierrehumbert, 1994; Treiman, Kessler, Knewasser, Tincoff & Bowman, in press; Vitevitch & Luce, 1998). This dimension is also closely tied to the similarity relations between words in the lexicon, and should be considered as a covariate with the other characteristics we examined in this study. A future examination of the HML corpus will include a description of word “regularity” or the probability of that word occurring as a word of English given the sequential structure of all other words in the language. Depending on how this metric is formulated, it may match quite closely with neighborhood structure (high for High-FAM words, slightly less for Mid-FAM, and very slight for Low-FAM), or there may be theoretically interesting divergences between the two measures. For example, it may be the case that although density is affected by even a .5 phoneme difference in average word length between High- and Mid-FAM words, the average phonotactic regularity of those words might still be indistinguishable. Or, when length is controlled, Mid-FAM words might still be less probable patterns than High-FAM words, though the average neighborhoods of the two groups are roughly comparable.

### Conclusions

The present analysis of words in the HML corpus uncovered several important discoveries about the composition of the corpus as well as the subjective familiarity of words in different subgroups. Our expectations about the relative proportions of very familiar, moderately familiar, and very unfamiliar words in the lexicon were confirmed. The relative frequency of words at each of these familiarity levels was also as expected, very high for High-FAM (“familiar”) words, considerably lower for Mid-FAM (“recognizable”) words, and slightly lower again for the lowest-FAM (“unknown”) word group. Neighborhood density was highest overall for High-FAM words as was neighborhood frequency and Mid-FAM words showed greater density and neighborhood frequency than Low-FAM. However, word length was a possible confound in these results because High-FAM words were shorter by about .5 phonemes on average than either Mid- or Low-FAM words. When subsets of these FAM groups were created so that the proportion of words at each phoneme length was equivalent across groups, the difference in neighborhood density and frequency between High- and Mid-FAM words virtually disappeared indicating that this difference was primarily due to the greater number of short High-FAM, high frequency words, a conclusion supported by the drop in frequency for the reduced High-FAM set. Low-FAM words, however, appear to genuinely come from more sparse neighborhoods and are very rare in the language. An examination of this subset indicated some possible reasons for this wide gap since Low-FAM words were apparently from very specific sets with different reasons for being outside the realm of common usage. When presented to listeners with more extensive linguistic and educational backgrounds than the original HML participants, these Low-FAM items were rated considerably higher and anecdotal evidence supported some of our descriptions of these words. An empirically sound, more complete study would be necessary to confirm these hypotheses regarding very rare words, familiarity, and the effects of lexical expertise. We are currently expanding and detailing the statistical description of the HML corpus by including the remaining words, either as two new groups (High-Mid and Low-Mid-FAM), or by widening the range of the original three groups in some specific manner, possibly using the ranges employed in the original study (Nusbaum et al., 1984). Finally, we would like to conduct a study comparing the variables studied here with a probabilistic account of phonotactic “regularity” to see how this may differ from word similarity based solely on similarity neighborhoods using the one phoneme substitution metric.

## References

- Besner, D. (1988). Visual word identification: Special-purpose mechanisms for the identification of open and closed class items? *Bulletin of the Psychonomic Society*, 26, 91-93.
- Bradley, D.C. (1978). Computational distinction of vocabulary type. Unpublished doctoral dissertation, Massachusetts Institute of technology.
- Bradley, D.C. & Garrett, M.E. (1983). Hemispheric differences in the recognition of closed and open class words. *Neuropsychologia*, 21, 155-159.
- Centre for Lexical Information. (1993). *The Celex lexical database*. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics.
- Chalmers, K.A., Humphreys, M.S., & Dennis, S. (1997). A naturalistic study of the word frequency effect in episodic recognition. *Memory & Cognition*, 25, 780-784.
- Chiarello, C. & Nuding, S. (1987). Visual field effects for processing content and function words. *Neuropsychologia*, 25, 539-548.
- Eukel, B. (1980). A phonotactic basis for word frequency effects: Implications for automatic speech recognition. *Journal of the Acoustical Society of America*, 68, S33.
- Forster, K.I. (1978). Assessing the mental lexicon. In E. Walker (Ed.), *Explorations in the biology of language* (pp. 139-174). Montgomery, VT: Bradford.
- Gernsbacher, M.A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113, 256-281.
- Hartman, G.W. (1946). Further evidence on the unexpected large size of recognition vocabularies among college students. *Journal of Educational Psychology*, 37, 436-439.
- Kucera, 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.
- Lewellen, M.J., Goldinger, S.D., Pisoni, D.B., & Greene, B.G. (1991). Word familiarity and lexical fluency: Individual differences in serial recall of spoken words. In *Research on Speech Perception Progress Report 17* (pp. 229-239). Bloomington, IN: Indiana University.
- Lewellen, M.J., Goldinger, S.D., Pisoni, D.B., & Greene, B.G. (1993). Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General*, 122, 316-330.

- Lively, S.E., Pisoni, D.B. & Goldinger, S.D. (1994). Spoken word recognition: Research and theory. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 265-301). San Diego, CA: Academic Press.
- Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception Technical Report No. 6*. Bloomington, IN: Speech Research Laboratory, Department of Psychology, Indiana University.
- Luce, P.A., Pisoni, D.B. & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives* (pp.122-147). Cambridge, MA: MIT Press.
- Luce, P.A. and Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Speech and Hearing, 19*, 1-36.
- Marslen-Wilson, W.D. (1984). Function and process in spoken word recognition. A tutorial review. In H. Bouma & D. G. Bouwhis (Eds.), *Attention and performance X: Control of language processes*. Hillsdale, NJ: Erlbaum.
- Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. In *Research on Speech Perception, Progress Report 10*, (pp. 357-376). Bloomington, IN: Speech Research Laboratory, Indiana University
- Pierrehumbert, J.B. (1994). Syllable structure and word structure: A study of triconsonantal clusters in English. In P. Keating (Ed.), *Phonological structure and phonetic form: Papers in Laboratory Phonology III* (pp. 168-188). Cambridge: Cambridge University Press.
- Pisoni, D.B., Nusbaum, H.C., Luce, P.A. & Slowiaczek, L.M. (1985). Speech perception, word recognition, and the structure of the lexicon. *Speech Communication, 4*, 75-95.
- Suprenant, A.M., Hsueh, T.-H., Hura, S.L., Harper, M.P., Jamieson, H.H., Long, G., Thede, S.M., Rout, A., Hockema, S.A., Johnson, M.T., Laflan, J.B., Srinivasan, P., & White, C. (1998). 4aSC5: Familiarity and pronounceability of nouns and names: The Purdue proper name database. Poster presented at the 135<sup>th</sup> Meeting of the Acoustical Society of America, Seattle, WA. [Abstract in *Journal of the Acoustical Society of America, 103*, 2980.]
- Thorndike, E. & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.
- Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (in press). English speakers' sensitivity to phonotactic patterns. In M. Broe and J.B. Pierrehumbert, (Eds.), *Papers in laboratory phonology V: Language acquisition and the lexicon*.
- Vitevitch, M.S. & Luce, P.A. (1998). When words compete: Levels of processing in spoken word recognition. *Psychological Science, 9*, 325-329.
- Zipf, G. K. (1935). *The psycho-biology of language*. New York, NY: Houton-Mifflin.

## Figure Captions

**Figure 1.** Mean Neighborhood Density (in number of words, computed for each word using either Method A or B) of each FAM group, either as a full set, or a 500-word length-balanced subset.

**Figure 2.** Mean Log Neighborhood Frequency (in average log of frequency counts for all neighbors of a given word, as computed by either Method A or B) of each FAM group, either as a full set, or a 500-word length-balanced subset.