Perception of Wordlikeness: Effects of Segment Probability and Length on Subjective Ratings and Processing of Nonword Sound Patterns

Nathan R. Large\textsuperscript{1}, Stefan Frisch\textsuperscript{2} and David B. Pisoni

\textit{Speech Research Laboratory}  
\textit{Department of Psychology}  
\textit{Indiana University}  
\textit{Bloomington, Indiana 47405}

\textsuperscript{1} This work was supported by NIH Training Grant DC00012 and NIH Research Grant DC00111 to Indiana University. We would like to thank Janna Carlson for her patient assistance in recording the stimuli used in this study. Thanks also to Janet Peirrehumbert, Richard Shiffrin and Mike Vitevitch for comments and suggestions.

\textsuperscript{2} Now at Department of Psychology, University at Buffalo, 365 Park Hall, Buffalo, NY 14260 (large@acsu.buffalo.edu)

\textsuperscript{3} Now at Program in Linguistics, University of Michigan, 1064 D Freize Building, 105 S. State St., Ann Arbor, MI 48901
Perception of Wordlikeness: Effects of Segment Probability and Length on Subjective Ratings and Processing of Nonword Sound Patterns

Abstract. A stochastic phonological grammar based on positional constituent probabilities of words in the mental lexicon was used to generate nonword sound patterns. The subjective wordlikeness of these patterns was evaluated by naïve listeners using a 7-point rating scale. Subjective wordlikeness ratings were lower for nonwords containing low-probability constituents and were also lower for nonwords with more syllables, as predicted by a log product probability metric. However, subjective ratings exhibited a floor effect for very low probability items, demonstrating that there is a limit to the discriminability of low probability patterns. Differences in the processing of these same nonwords which reflected their wordlikeness were also found in a recognition memory task. Nonwords with high-probability constituents yielded better recognition performance, suggesting that participants were able to use their familiarity with frequently occurring lexical patterns to improve recognition for these stimuli. Finally, individual differences in the processing of these nonwords were examined using a test of word familiarity as a measure of lexical knowledge. These differences suggest that participants with greater lexical knowledge (as measured by their familiarity with uncommon real words) were better able to analyze the nonwords into sub-word constituents. Together, these results demonstrate the contribution of the mental lexicon as the foundation of phonological competence, which is reflected in linguistic and metalinguistic tasks involving the processing of novel nonword stimuli.

Introduction

In the process of studying the human ability to recognize spoken words, investigators have described many factors which affect the way linguistic stimuli are perceived and processed. Some of these are based on ‘lexical’ properties of a particular item, including its semantic relationship with other words, its frequency of use in text or speech (Treisman, 1978), its familiarity to the listener (Chalmers, Humphreys, & Dennis, 1997), or its confusability based on the set of similar-sounding words in the lexicon (its ‘lexical neighborhood’; Luce & Pisoni, 1998). However, other properties may be due to the linguistic structure of the item: morphological considerations such as the presence of bound grammatical morphemes, embedded words, or compounding; and phonological features such as length, stress pattern, and acoustic-phonetic composition. Descriptions of the acoustic-phonetic information can focus on the frequency of constituent sounds (Landauer & Streeter, 1973), or the frequency of the combinations of these sounds (most commonly, transitional and biphone probabilities, e.g., Treisman, Kessler, Knewasser, Tincoff, & Bowman, in press).

In order to remove most lexical and morphological factors from spoken stimuli, many studies have employed ‘nonsense words’, ‘nonwords’, or ‘pseudowords’: linguistic stimuli which have no meaning within a particular language. Nonword stimuli have been used extensively in situations where actual words would be inappropriate. For example, nonwords have been used to examine novel word learning (Dollaghan, 1985; Gathercole, 1995), and to elicit subjective judgments of linguistic acceptability (Greenberg & Jenkins, 1964). Nonword stimuli have also been used as controls against semantically meaningful stimuli with matching phonetic composition to examine the effects of lexical status on perception (Hulme, Roedents, Brown, & Mercer, 1995). The process of lexical access has been studied using a mix of word and nonword stimuli in a ‘lexical decision’ task (Rubenstein, Garfield,
& Milliken, 1970; Marslen-Wilson, 1984; Forster, 1978) to discover how words are organized in long-term memory and retrieved from the 'mental lexicon' (Oldfield, 1966).

Nonword stimuli vary in their similarity to actual words in a listener's native language; this is referred to as 'wordlikeness'. Wordlikeness, which we will show has a strong relationship to linguistic acceptability, is an important factor in the perception and processing of nonword stimuli (Gathercole, 1995; Jusczyk, Luce, & Charles-Luce, 1994; Vitevitch & Luce, 1998; Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997). For example, differences in wordlikeness have been used by Vitevitch and Luce (1998) to demonstrate the facilitary effects of greater structural probability on repetition naming for nonword stimuli, and by Gathercole (1995) to show that memory span for low-wordlike stimuli is a better predictor of later vocabulary development and reading skills for young children than memory span for high-wordlike stimuli (see also Baddeley, Gathercole, & Papagno, 1998). It is hypothesized that high-wordlike stimuli are easier to process because they activate a greater number of familiar phonological patterns in the mental lexicon. Thus, nonwords provide an important tool for investigating the processing and organization of words in the lexicon while avoiding some lexical confounds.

Despite the importance of wordlikeness, relatively few studies have tried to examine what factors are considered when listeners make judgments of wordlikeness and how these judgments are related to differences in processing. Some studies have assumed relatively simple metrics of wordlikeness based on segment probability and transitional probability between segments (Jusczyk, Luce, & Charles-Luce, 1994) or the degree of overlap between a given nonword and the words in the lexicon (Greenberg & Jenkins, 1968). Other studies bypass the question of a priori classification entirely, by simply asking subjects for acceptability ratings for a set of stimuli, then dividing the stimuli into groups based on their average rating (Gathercole, 1995; see also Gathercole, Frankish, Pickering & Peaker, submitted). Two lines of research have investigated the basis for judgments of wordlikeness. Studies of sound pattern similarity seek to describe what information listeners employ to evaluate the similarity between a nonword and the words in their native language (e.g., Vitz & Winkler, 1973; Sendlimier, 1987). In addition, a growing body of psycholinguistic research on the linguistic concept of phonotactics, which describes the constraints on well-formed words/morphemes, provides explicit hypotheses about what properties make an item more acceptable to listeners (e.g., Vitevitch & Luce, 1998; see also Glushko, 1979; Treiman, Goswami, & Bruck, 1990; for precursors to this work in visual word recognition).

While the factors which influence judgments of wordlikeness may be the same as those which lead to processing differences in other tasks, previous studies of wordlikeness judgments have not examined this question. Only a few studies have intentionally manipulated properties of nonwords to influence wordlikeness judgments and also investigated whether these same manipulations affect performance in other psycholinguistic tasks. Thus, it is not yet clear whether any particular manipulation of wordlikeness will influence performance on any particular task, or if some factors which influence metalinguistic judgments of wordlikeness do not affect performance on certain linguistic tasks.

Greenberg and Jenkins (1964) carried out one of the earliest investigations of wordlikeness judgments. They examined the 'distance' of four-phoneme (CCVC) monosyllabic nonsense words from English by computing a score based on the number of combinations of positions where any phoneme substitution would not yield a real English word. This score correlated highly with both closed and open-scale judgments of 'distance from English' elicited from participants. Greenberg and Jenkins theorized that there is a 'phonological space' in which words are assembled in memory according to their sound structure (see also Treisman, 1978). Words differ according to the number and frequency of other words which occupy the 'similarity neighborhood' in this space. In this view, combinations of acoustic-phonetic information which are more common reside in more 'dense' neighborhoods in the lexical space,
while uncommon segments and combinations occupy 'sparse' neighborhoods. This hypothesis was subsequently confirmed by Landauer and Streeter (1973) whose computational analysis of a dictionary (as an approximation of a native speaker's mental lexicon) demonstrated that words in larger neighborhoods are more common in everyday usage than words in smaller, sparser neighborhoods (see also Eukel, 1980). They also confirmed that common items (from larger neighborhoods) are composed of different, more common sounds than rare items (in sparser neighborhoods). More recently, Luce and Pisoni (1998) have summarized a number of theoretical findings based on this concept of 'lexical neighborhood', uncovered over the past 30 years. Their own findings and interpretation of past results provide considerable support for the view that the sound patterns of spoken words are organized in the mental lexicon in a phonetic similarity space.

In related research, Vitz and Winkler (1973) and later Sendlemeier (1987) sought to describe the information used when subjects make specific comparisons between pairs of words or nonwords. Vitz and Winkler analyzed similarity judgments based on a model which compared the overlap of phonemes between a pair of words, finding a clear relationship between degree of overlap and similarity ratings. Sendlemeier (1987) compared listeners' abilities to classify nonwords based on similarity and concluded that individuals with more phonetic training could make similarity judgments along more dimensions (phonetic features). These two-item comparison tasks may involve processes analogous to those used in wordlikeness judgments. Thus, we might expect that words constructed with more common phonemes would be judged to be more wordlike, and that there will be individual differences in the amount of information used in making these wordlikeness judgments.

In linguistic theory, the term phonotactics is typically used to refer to the conditions within a language which specify its well-formed words. The traditional approach to phonotactic constraints is to outline the 'legal' versus the 'illegal' patterns for a language. Sound patterns which are composed of phonemes in legal combinations, but have no lexical entry, have been termed 'pseudowords', in contrast to impossible nonword patterns (Brown & Hildum, 1956). Phonotactic constraints apply over linguistic constituents such as the morpheme, syllable, onset, or rime (Davis, 1988; Treiman, 1988; Pierrehumbert, 1994). The inventory of constraints which denote possible and impossible patterns is a component of the phonological grammar, which ideally can produce only acceptable sound-strings for a particular language, but never an illegal string.

More recently, it has been noted that certain sound patterns, while not impossible, occur rarely enough in the language that their acceptability is questionable. In a landmark study, Pierrehumbert (1994) examined restrictions between word-internal (medial) consonant clusters. She claimed that particularly improbable combinations are equivalent to illegal pairings. Combining this idea of probabilistic, or stochastic, constraints with a phonological grammar results in a 'stochastic grammar', which can both: (1) specify the statistical likelihood of a particular nonword, relative to other nonwords, and (2) be used to generate nonwords of a given relative probability.

Recently, Kessler and Treiman (1997) used the idea of a stochastic grammar to examine onset-vowel (CV) and vowel-coda (VC) cooccurrences in real words using computational techniques (see Pisoni, Nusbaum, Luce, & Slowiaczek, 1985). They found that many vowel-coda pairs occurred more or less frequently than would be expected by chance, while onset-vowel combinations were random. For example, the rime /æl/ occurs less frequently than expected based on the high probabilities of /æ/ and /l/ independently appearing as the vowel and coda in a syllable. They argued that since constraints exist between VC but not CV pairings in real words, the stochastic grammar provides converging evidence for the rime as a syllable-internal constituent (see also Treiman, 1988; Treiman, Kessler, Knewasser, Tincoff, & Bowman, in press).
In support of the idea of a probabilistic grammar, Coleman and Pierrehumbert (1996) note that nonwords vary in the number of subjects who accept them as potential words in their native language. Using a simple probabilistic grammar, they generated probabilities (the log product of the nonwords’ constituent probabilities) for a set of nonwords which had been judged for acceptability in a previous study on consonant clustering. They compared the log product probability for each nonword to the number of participants who accepted it and found the two measures to be highly correlated. A non-probabilistic, categorical grammar, which would only allow a string to be acceptable or unacceptable, could not account for this continuous measure of ‘percentage rejecting’.

Since Coleman and Pierrehumbert (1996) performed a post-hoc study using data designed for a different experiment, we became interested in how stimuli constructed explicitly from such a stochastic grammar would compare when rated by subjects. The primary goal of our study is to replicate and extend their findings with nonwords using the same stochastic grammar. We propose that this grammar can be used as a productive method of generating nonword stimuli which vary predictably in wordlikeness. This same grammar also provides an account of the information participants employ in the subjective rating task, as well as the structural factors that influence performance in memory and naming tasks using nonwords.

We generated a corpus of nonwords that varied in the probability of their constituents, with either High or Low probability constituents in each position. We also varied the length of these stimuli between 2, 3, and 4 syllables. A stochastic grammar based on log product probability predicts that nonwords with high-probability constituents should be rated higher than nonwords with low-probability constituents, and that longer nonwords will be rated lower than shorter nonwords with constituents of comparable probability. These nonword stimuli were used in four experiments using two different behavioral tasks: subjective ratings of wordlikeness or acceptability, and recognition memory. To gain further insight into the information which is used by participants in performing these tasks, we also considered the effect that individuals’ mental lexicons might have on their perception and memory for these nonword sound patterns. To examine this factor, participants in three experiments also completed a short test of word familiarity with a corpus of actual English words. Their rating and/or recognition memory performance with the nonword stimuli was compared to this lexical familiarity data to assess any existing relationships.

For purposes of presentation the paper is organized primarily by theoretical topic. Group performance on the subjective wordlikeness and acceptability tasks are presented first under the heading Subjective Ratings of Nonwords. Group performance in the recognition memory experiment is then presented under the heading Recognition Memory for Nonwords. Finally, the analyses of individual differences across all of the experiments are presented in the last section under the heading Individual Differences in Lexical Knowledge.

**Subjective Ratings of Nonwords**

**Experiment 1**

**Method**

*Stimuli.* The nonword stimuli used in these experiments contained high or low probability constituents, according to the Coleman and Pierrehumbert (1996) grammar. This grammar considers the likelihood of each constituent (onset-rime) of each syllable in terms of its prosodic position within the word. The prosodic positions reflect the stress of the syllable (strong-weak) and the constituent’s position within the word (initial, final, or medial). Since only onsets can be initial and only rimes can be final, this yields eight different distributions of constituents-by-position, such as strong medial onsets (Osm), weak
final rimes (Rwf), and so on. Each constituent distribution consists of the inventory of potential constituents occurring in a given position, as well as the probability of that item's occurrence in that position, computed from a parsed dictionary of English. The probability of each constituent is equal to the number of words which contain that particular constituent at the specified position, divided by the total number of constituents at that position. Coleman and Pierrehumbert multiplied the probabilities of the constituents of each syllable in a nonword to yield that nonword's expected probability in the grammar. They found that the 'log product of probabilities' for a nonword was a better predictor of subjective acceptability than the product of the probabilities, the probability of the worst syllable, or the probability of the best syllable.

To construct our nonword stimuli, we selected constituents of high and low probability for each prosodic position from probability distributions derived from an on-line version of Webster's Pocket Dictionary (Nusbaum, Pisoni, & Davis, 1984). We used only single-consonant onsets (C), all medial rimes consisted only of a single vowel or diphthong (V), and all final rimes were a vowel-consonant (VC) pair, so that all of our nonwords had an alternating CV pattern. We constructed words of 2, 3, and 4 syllables in length, using the most common stress pattern of that word length in the dictionary (see Figure 1). High and low probability constituents were combined to create syllables for each position of high and low probability. Vowel-only stressed medial rimes (for constructing both stressed initial and stressed medial syllables) were a limiting factor, allowing 24 words per probability group and length. Each syllable occurred six times across the stimuli. Nonword stimuli were constructed from these syllables so that no two syllables occurred in the same word more than once within a group. The resulting stimulus set contained 144 items, 72 high probability items and 72 low probability items. During construction, we were careful to avoid actual English words (which sometimes occurred as high-probability strings). To avoid semantic confounds, high probability final rimes which might be interpreted as suffixes were not used (e.g. /-ik/, /-us/, /-ad/). The complete set of stimuli is given in Appendix 1.

Each nonword's expected probability, based on the log product of its constituent probabilities, was also computed. We confirmed the predicted differences of Length and Probability Group with a 2-factor ANOVA. As expected, there was a main effect of constituent probability ($F(1,143) = 2708, p < 0.0001$) and a main effect of length ($F(2,143) = 703, p < 0.0001$). There was also a divergent constituent probability by length interaction ($F(2,143) = 91.1, p < 0.0001$) indicating that length decreased the log product probability more for nonwords with low probability constituents. Figure 2 shows the mean log probability for the nonword stimuli in each probability and length group.

A single talker produced spoken versions of the 144 nonwords in a randomized order, recorded over several sessions using the SAP digital signal recording program (Hernandez, 1995). The talker was a female research assistant with some phonetic training. The talker sat in an IAC booth and recordings were made using a Shure SM98 microphone. Transcriptions of the nonwords were presented individually on a CRT monitor. The recorded stimuli were screened for accuracy and fluency by the first author. Each stimulus was then digitally leveled to 64 dB SPL. The talker was not aware of the methods used to generate the stimuli or the purposes of the experiment.
Figure 1: Structured phonological representation of nonword sound patterns created from the 6 constituent distribution for each length (2, 3, or 4 syllables).
Figure 2: Mean expected log probability of the constructed nonword stimuli average within cells by Probability group (chosen from High or Low probability constituents from each distribution) and Length in Syllables (2, 3, or 4).
Participants. Twenty undergraduate students earning experimental credit for introductory psychology courses at IU participated in this experiment. All reported that English was their first language and that they had no recent history of speech or hearing disorders.

Procedure. Participants were instructed to rate the nonwords based on their wordlikeness. The instructions included descriptors for a seven-point scale to be used for rating the “nonsense words”. A rating of 1 was described as “Low- Impossible- this word could never be a word of English”, a rating of 4 as “Medium- Neutral- this word is equally good and bad as a word of English”, and 7 as “High-Possible- this word could easily be a word of English”. Ratings of 2 or 3 represented “unlikely”, or items which fit between 1 and 4; 5 or 6 was “likely”, items which were better than 4 but not as acceptable as 7. We also instructed subjects to work “as quickly as possible without sacrificing accuracy”. After reading instructions for the rating task, subjects entered the testing room.

Stimuli were presented by computer over Byerdynamic DT100 headphones at 74 dB SPL. Participants sat in individual sound-absorbent testing booths, facing a computer monitor. Participants responded using custom-designed 7-button response boxes. The rating response and latency were recorded for each item. The response boxes were covered by printed cards indicating the digits 1-7 and the guidelines “Low Wordlikeness”, “Medium Wordlikeness”, and “High Wordlikeness” above the digits 1, 4, and 7, respectively. Stimuli were randomized by the computer for each listener.

Length and probability group of the stimuli were within-subjects variables. A between-subjects variable, number of presentations (1 or 2), divided the subjects into two groups. This factor was varied for methodological reasons to see if ratings would be affected by a second presentation of each item. Note that each stimulus was rated only once by each participant for both groups; the difference between groups was whether the stimulus was played once or twice per trial. There was a 2-second interval between successive trials. For the two-presentation condition, the first and second presentations were separated by a 1-second interval. Before each trial began, the word ‘***READY***’ was flashed on the computer screen.

Results and Discussion

Ratings. Analysis of variance was performed with Probability Group, Length in Syllables, and Number of Presentations as factors across subjects and across stimuli. Figure 3 displays the mean rating for each length and probability group. The hypothesized effect of constituent probability was present across subjects, $F(1,108) = 142, p < 0.001$, and across stimuli, $F(1,276) = 610, p < 0.001$. Patterns with higher-probability constituents received higher ratings. An effect of length was also present across subjects, $F(2,108) = 37.9, p < 0.001$, and across stimuli, $F(2,276) = 162, p < 0.001$. Longer stimuli were rated lower, as expected based on log product probability. There was an interaction present between probability and length across stimuli, $F(2,276) = 5.9, p < 0.005$, but not across subjects, $F(2,108) = 1.4$, ns. This marginal interaction was not the interaction originally predicted based on the log product probability. Instead of becoming more different as item length increased, subjective ratings tended to converge between high and low probability stimuli as length increased (compare Figures 2 and 3). This convergence appears to be due to a floor effect in the ratings for very low probability stimuli. Mean ratings for each stimulus, as a function of log product probability, are shown in Figure 4. Despite the floor effect, expected log probability and average ratings were highly correlated, $r(143) = +0.87, p < 0.001$, replicating Coleman and Pierrehumbert’s (1996) earlier findings.

--------------------------------------------------------
Insert Figures 3 and 4 here
--------------------------------------------------------
Figure 3: Mean subjective ratings across subjects for the nonword stimuli divided by Probability Group (High vs. Low) and Length in syllables (2, 3, or 4).

Figure 4: Mean subjective ratings for each nonword as a function of the log product constituent probability for that sound string. Probabilities were generated by a version of Coleman and Pierrehumbert's (1996) grammar.
Figure 5: Mean subjective ratings for nonword stimuli divided by Probability Group (High vs. Low), Length in syllables (2, 3, or 4), and number of Presentations (1 or 2).

Figure 6: Mean percentage of implicit rejections (ratings of '1') for nonword stimuli in each condition divided by Probability Group and Length in syllables.

Figure 7: Percentage of subjects responding with an implicit rejection (rating of '1') for each stimulus item as a function of the log product constituent probability of that stimulus.
Finally, number of presentations also had an effect on ratings across subjects, $F(1,108) = 3.58, p = 0.06$, and across stimuli, $F(1,276) = 15.37, p < 0.001$. There were no interactions between number of presentations and any of the other factors. Figure 5 shows average ratings across length and constituent probability groups, for each presentation group. Participants who heard each stimulus twice before choosing a rating responded with slightly higher ratings for all types of sound patterns. We hypothesize that hearing an extra presentation of each stimulus enhanced the activation of similar sounding familiar patterns in long-term memory, making the nonwords seem more familiar and thus more wordlike. Rather than changing the distribution of their ratings, listeners receiving an extra presentation of each stimulus only moved their range upward slightly. This shift in ratings was confirmed by transforming each subject’s ratings to z-scores based on individual means and variance. The effect of presentations was removed across subjects, $F(1,108) = 0.002$, ns., and across stimuli, $F(1,276) = 0.002$, ns., while the effects of probability, length, and the probability by length interaction remained.

Post-hoc Analysis: Implicit Rejections. Examination of the ratings revealed a clear floor effect in subjective wordlikeness ratings for nonwords with probability below 10^{-12}. Coleman and Pierrehumbert (1996) found no such floor effect, but our studies differ from theirs in two important ways. First, nearly all of their stimuli had probabilities above 10^{-12}. Second, they used a two-alternative forced-choice accept/reject task rather than a 7-point wordlikeness scale. With this in mind, we carried out a post-hoc analysis of the number of ratings equal to ‘1’, either for each stimulus across subjects or by each subject across stimulus categories. Our instructions described a rating of ‘1’ as equivalent to “Low-Impossible-this word could never be a word of English”, making such a rating potentially analogous to an ‘unacceptable’ response in a 2-choice task. We call this response an ‘implicit rejection’, signifying that listeners were not giving a rating of ‘impossible’ to contrast with ‘possible’, but rather were giving a response of the lowest rating on a continuous scale. An ANOVA with mean number of ‘implicit rejections’ as the dependent measure showed a significant effect of constituent probability across subjects, $F(1,108) = 65.8, p < 0.001$, and across stimuli, $F(1,276) = 339, p < 0.001$, as well as a significant effect of length across subjects, $F(2,108) = 15.45, p < 0.001$, and across stimuli, $F(2,276) = 77.7, p < 0.001$. Most importantly, there was a significant divergent interaction between these two factors across subjects, $F(2,108) = 4.7, p < 0.05$, and across stimuli, $F(2,276) = 23.5, p < 0.001$. Figure 6 shows implicit rejections for each group. The pattern displayed here strongly resembles the means for expected probability shown in Figure 2. Figure 7 shows ‘implicit rejections’ for each stimulus as a function of log probability. Once again, we see there is a very strong negative correlation with log probability, $r(143) = -0.90, p < 0.001$, as the floor effect is no longer found.

The present findings show that the Coleman and Pierrehumbert (1996) grammar can be successfully employed to generate nonword stimuli of relatively high or low wordlikeness, as determined by subjective ratings. Individual participants are unable to discriminate differences in probability for very low probability strings using a rating scale, but these stimuli are still differentiated in the number of floor-level ratings they receive across participants. To determine if these implicit rejections are substantially different from rejections on a two-choice accept/reject task, as well as to more closely replicate Coleman & Pierrehumbert’s study, we conducted a second experiment using the same set of stimuli. Instead of using our 7-point scale of wordlikeness, however, we asked subjects to respond either
“Yes” or “No” to the question “Could this word be a possible word of English?” The number of true rejections on this task can then be compared to our measure of implicit rejections, and to the rating data, to see if differences in wordlikeness judgments arise under different response conditions.

Experiment 2

Methods

Participants. Twenty-four undergraduates enrolled in psychology courses at Indiana University participated in this study for class experiment credit. All 24 were native speakers of English, between the ages of 18 and 30, and reported having no speech or hearing disorder at time of testing.

Stimuli & Apparatus. The nonword stimuli used in this experiment were the same as those used in Experiment 1. Responses were made with the same 7-button boxes, with labels reading “Yes” and “No” placed above the ‘2’ and ‘6’ buttons, respectively. The instructions for this task were changed from the previous experiment. Subjects were told to accept or reject each sound pattern as a possible word of English and to respond only “Yes” or “No” by pressing the appropriate button.

Procedure. The testing procedure and presentation of stimuli were identical to Experiment 1, except that each stimulus was presented only once per trial. In addition, after the nonword acceptability task, we administered a lexical familiarity (FAM) task to obtain an estimate of the number of real words an individual participant knew. A detailed description of this task and the results of the individual differences analysis is presented below in a separate section under the heading Individual Differences in Lexical Knowledge.

Results and Discussion

The mean number of “Yes” responses across subjects and across stimuli was used in an ANOVA. This analysis showed a significant effect of constituent probability across subjects, $F(1,138) = 343, p < 0.001$, and across stimuli, $F(1,138) = 506, p < 0.001$, as well as a significant effect of length across subjects, $F(2,138) = 49.1, p < 0.001$, and across stimuli, $F(2,138) = 72.5, p < 0.001$. The interaction between these two factors was significant across stimuli, $F(2,138) = 3.3, p < 0.05$, but not across subjects, $F(2,138) = 2.2, ns.$, and appears to be convergent. That is, like the average ratings obtained using the 7-point scale, the mean number of “Yes” responses between high and low constituent probability stimuli was smaller for longer stimuli. Mean “Yes” responses for stimuli in each group are shown in Figure 8. Figure 9 shows the percentage of “Yes” responses for each stimulus as a function of log probability. As shown here, there is a floor effect near probability $10^{-12}$ as in Experiment 1. However, despite this floor effect, percentage of “Yes” ratings and log probability for the stimuli were highly correlated, $r(143) = +0.86, p < 0.001$. Thus, we find that the 2-choice acceptability task and the 7-point wordlikeness rating task produce very similar results, although the 7-point scale allows scores to be averaged within individual subjects and the variance appears much smaller. By contrast, a comparison between the “No” responses and the implicit rejections measure from Experiment 1 shows that these are not equivalent rejections. With only two choices available, participants used the “No” choice more often than the ‘impossible’ rating of ‘1’ on the 7-point task. The implicit rejection is used only for the most unacceptable items.

Insert Figures 8 and 9 here

107
Figure 8: Mean number of "Yes" responses (acceptances) out of 24 items possible across subjects separated by Probability Group and Length.
Figure 9: Percent of "Yes" responses (acceptances) out of 24 subjects possible by stimulus item displayed as a function of the log product probability of that stimulus.
Discussion: Ratings of Nonwords

Taken together, Experiments 1 and 2 replicate and extend the original study by Coleman and Pierrehumbert (1996), providing support for their phonological grammar as a method of generating and describing nonword sound patterns. However, we found limits to our participants' ability to discriminate very low probability nonword patterns. This suggests that there are limitations in the usefulness of subjective wordlikeness as a classification tool for nonword stimuli. At the very lowest end of the probability scale, subjective ratings of acceptability or wordlikeness no longer discriminated among the stimuli. It may be that the lowest-probability patterns (below $10^{-12}$) are not even considered at all wordlike by listeners. They are too improbable to be considered acceptable patterns in the language. Although these very low probability patterns were not discriminated in terms of wordlikeness, they may show differences in other behavioral tasks. In addition, very high probability stimuli should resemble English words enough to show lexical influences from highly similar words. We undertook the next study to provide some data for comparison between subjective ratings of nonwords collected in the first two experiments and a language processing task (i.e., recognition memory) that requires participants to encode, store, and retrieve these nonword patterns from memory.

Recognition Memory for Nonwords

Experiment 3

Introduction

The third experiment employed a discrete recognition memory task to study the representation and processing of our nonwords in memory. If participants are using long-term knowledge about the words in their language to perceive and evaluate the wordlikeness of strings, then items which are more wordlike should have stronger activation or more associations with real words in the language and thus be more distinctive in memory (i.e., more discriminable and more easily recognized). Thus, we expect that participants would recognize the stimuli with high probability constituents more accurately than stimuli with low probability constituents and would display sensitivity to the regularities inherent in these nonword patterns.

Note that this prediction is the reverse of the pattern for real words. High frequency real words, which are composed of high probability constituents (Landauer & Streeter, 1973), are more difficult to recall in recognition memory than low frequency words (see Chalmers & Humphreys, 1998; Guttentag & Carroll, 1997). The frequency effect in recognition memory is apparently due to listeners' difficulty separating recent presentations of high frequency words from their previous presentations. Studies comparing low frequency words with very low frequency words find that low frequency words actually show better memory recognition performance than very low frequency items (which occur less than once per million words of text), analogous to the nonword pattern. Chalmers, Humphreys, and Dennis (1997) demonstrated that individual recognition performance for a set of words can be changed by manipulating listeners' familiarity with those words.

In a recent study, Vitevitch and Luce (1998) presented data from a naming task which demonstrated a similar pattern. They found that nonwords with high probability segments and high transitional probability between pairs of segments were repeated more quickly than low probability nonwords. For real words, words containing high probability segments were repeated more slowly than words containing low probability segments. Goldinger (1989, 1998) also showed that naming of low frequency words is facilitated by greater neighborhood density, again indicating the use of long-term lexical knowledge to improve performance for low frequency items.
Predicting the effect of nonword length on recognition memory is more problematic. In spoken word recognition, longer words are more easily identified than shorter words, presumably due to the additional acoustic-phonetic cues for lexical access contained in the longer words (Cluff & Luce, 1990; Kirk, Pisoni, & Osberger, 1995; Pisoni, Nusbaum, Luce, & Slowiaczek, 1985). Since longer nonwords contain more acoustic-phonetic information which may provide additional cues for recognition and recall, they should also be easier to remember in recognition memory than shorter words. This has been demonstrated for visually presented words (Zechmeister, 1972). However, as with the effects of constituent probability discussed above, nonwords may not behave similarly to words. It is equally plausible that longer nonwords, which are less wordlike, would activate fewer real words and therefore may be more difficult to recognize in a recognition memory task.

Method

Participants. Thirty undergraduate students at Indiana University participated in this experiment for credit in an introductory psychology course. All reported that English was their first language. None of the participants reported any history of speech or hearing disorders at the time of testing.

Stimuli and Procedure. The first part of this experiment, the study phase, was identical to the nonsense word rating task used in Experiment 1, with the exception that each participant received only half of the stimulus list (72 items). The stimuli in each probability and length group were randomly divided into four quarters {a, b, c, d}, which we then recombined into six different pairings for the rating task {ab, ac, ad, bc, bd, cd}. Thus, half of the participants (15 individuals total) rated each nonword stimulus and each participant rated half of the stimuli. We collected these ratings in the same manner as Experiment 1 and evaluated them as a measure of replication. In each trial, each item was presented twice for every participant as in the two presentation condition in Experiment 1. The rating task thus served as the ‘study phase’ for the recognition memory experiment; however, the participants were not informed that there would be a surprise recognition memory test at the end of the rating task. Between the study phase and test phase, we gave participants a paper-and-pencil arithmetic activity with 20 simple addition and subtraction questions to eliminate recency effects. For the recognition memory test, participants were asked to determine if each item was “old” or “new”. They entered their responses using the same 7-button boxes, but with the label “Old” placed over the second button and “New” over the fifth. Participants were now presented with the full stimulus list (144 items) and responded to each item, depending on whether they recognized that item from the earlier study phase. In the final part of the experiment, participants also took the real word familiarity (FAM) test, described in the Individual Differences in Lexical Knowledge section below.

Results and Discussion

Ratings. A 2 by 3 ANOVA with constituent probability and length as independent factors showed an effect of constituent probability across subjects, $F(1,174) = 149, p < 0.001$, and across stimuli, $F(1,138) = 119, p < 0.001$, with an effect of length across subjects, $F(2,174) = 51.9, p < 0.001$, and across stimuli, $F(2,138) = 81.6, p < 0.001$. The constituent probability by length interaction was converging but did not reach statistical significance. Except for the lack of an interaction between constituent probability and length across stimuli, this pattern replicates the main findings obtained in Experiment 1. It is likely that the lack of significant interaction is due to a reduction in statistical power because in this experiment only half as many ratings were collected during the study phase as compared to Experiment 1. Mean ratings for each constituent probability and length group in each experiment were nearly identical.
Once again, we find the 'implicit rejections' measure to be more reflective of the log product probability metric. Mean number of ratings of '1' showed an effect of constituent probability across subjects, $F(1,174) = 75.9, p < 0.001$, and across stimuli, $F(1,138) = 175, p < 0.001$, and an effect of length across subjects, $F(2,174) = 21.5, p < 0.001$, and across stimuli, $F(2,138) = 49.8, p < 0.001$. A divergent interaction between constituent probability and length was present as well across subjects, $F(2,174) = 5.3, p < 0.01$, and across stimuli, $F(2,138) = 12.1, p < 0.001$.

**Recognition Memory.** Three measures of recognition memory performance were examined: hit rate, false alarm rate, and $d'$. An ANOVA on hit rate (the percentage of previously heard items to which participants correctly responded "old") showed a significant effect of constituent probability across subjects, $F(1,174) = 18.0, p < 0.001$, and across stimuli, $F(1,138) = 16.0, p < 0.001$, but there was no effect of length across subjects or across stimuli. We also failed to observe any interaction between these factors. Figure 10 displays the mean hit rate for stimuli in each constituent probability and length category.

-------------------------
Insert Figure 10 here
-------------------------

The second dependent variable, false alarm rate (the percentage of new items to which participants mistakenly responded "old") showed no effects of constituent probability or of length. None of the interactions between these factors was significant.

Finally, the measure $d'$, a composite of hit rate and false alarm rate which normalizes for an individual's decision criteria (the tendency to respond "old" more often than the actual probability of old items in the test phase), gave predictable results based on these first two findings. Average $d'$ revealed a significant effect of constituent probability across subjects, $F(1,174) = 16.8, p < 0.001$, and across stimuli, $F(1,138) = 13.4, p < 0.001$, but there was no effect of pattern length in either the subject or item analysis. There were also no interactions between constituent probability and length, across subjects or stimuli, as expected from the results for hit rate and false alarms.

Our hypothesis concerning the effects of constituent probability on recognition memory performance was borne out: nonword patterns with high-probability constituents were consistently recognized more accurately than nonword patterns with low-probability constituents. The lack of any length effect was somewhat puzzling and may be due to the method used to construct the nonword stimuli. Recall that a limited inventory of constituents was used as input for the phonological grammar, so a variety of words in the stimulus set shared the same constituents. In addition, nearly all possible combinations of onset and rime appeared as syllables in the nonwords, and groupings of syllables were similarly combinatorial. Thus, there was very little redundancy between constituents in the words, regardless of length. This artifact of the design process for the nonwords could negate the advantage of additional cues for longer stimuli found with real words. In addition, the repeated use of constituents may have equalized the amount of overlap between our stimuli and actual words within each probability group. Since, in general, the long and short nonwords shared the same constituents with actual words of English, they may have had equal benefit from the activation of familiar words and therefore no advantage would be gained from longer patterns.

**Discussion**

We found in Experiment 3 that nonwords with high probability constituents are recognized more accurately than nonwords with low probability constituents. However, since there was no length effect,
Figure 10: Mean hit rate (% hits responding "Old" to previously presented stimuli out of 12 possible) within each Probability Group and Length in syllables.
the performance in recognition memory corresponded only in a general way to phonotactic probability, or to ratings of wordlikeness. For example, although 2- and 3-syllable nonwords with high probability constituents had different log product probabilities and were rated differently in our wordlikeness task, recognition performance was the same for these items. Thus, this experiment suggests that factors influencing wordlikeness judgments and factors affecting the processing of nonwords (i.e., recognition memory) may not entirely overlap. Differences in wordlikeness due to length in our stimuli did not have the same effect on recognition memory as differences in wordlikeness due to constituent probability. We suspect that similar effects may emerge in other experimental tasks which have shown effects of both wordlikeness and length, namely working memory span and repetition naming.

Individual Differences in Lexical Knowledge

Since the stochastic grammar consists of probability distributions based on the occurrence of words in the lexicon, there is good reason to believe that differences between subjects with differing vocabulary knowledge may also play a role in these measures. If native speakers have a grammar that is based on constituent probabilities from the words they actually know, would a larger or smaller mental lexicon result in different ratings for our nonsense sound patterns? Would subjects with more experience with rare and unusual English words use that knowledge in language tasks employing nonword stimuli? There is now some evidence that the extent of an individual’s lexical knowledge is an important individual factor in spoken word recognition and memory. Lewellen, Goldinger, Pisoni, and Greene (1991, 1993) used lexical knowledge, as measured by a word familiarity test along with several other scores measuring verbal ability, to differentiate between high and low-verbal subjects. These subjects performed differently on several language tasks involving real word stimuli. Their results suggested that individual differences in lexicon size may play a role in a variety of language tasks that require participants to access and manipulate real words.

To examine the role of lexical knowledge in the processing of nonwords, we obtained a measure of lexical knowledge from our participants based on their relative familiarity with a variety of words. The familiarity test we used to measure lexical knowledge in our participants was a shortened version of the word familiarity test (FAM test) used by Lewellen et al. (1991). The FAM test is based on comparing an individual’s familiarity with a variety of English words to the norms obtained by Nusbaum, Pisoni, and Davis (1984). Nusbaum et al. presented the 20,000 entries from the Webster’s Pocket Dictionary to groups of participants in a large study. Each word was rated for familiarity by 10 undergraduate psychology students on a 7-point familiarity scale. Each individual participant rated 500 items. The Lewellen et al. FAM test used the same task with a subset of 450 words, 150 from each of three broad categories: High familiarity (mean familiarity ≥ 6), Mid familiarity (3 < mean familiarity < 6), and Low familiarity (mean familiarity ≤ 3). Lewellen et al. compared their participants’ ratings to the Nusbaum et al. norms to assess the lexical knowledge of their participants against the general undergraduate population. Our modified version of the FAM test used 150 items, 50 items at each of the three familiarity levels. A reduced list was used due to time constraints on the experimental session. Appendix 2 contains the list of the actual words used, separated by familiarity level, along with the mean familiarity scores for each level from Nusbaum et al.

The FAM portion of each experiment was always the final section. In this phase, the participants were asked to rate each real word, using a 7-point scale based on their ‘familiarity’ with that item. A rating of ‘1’ corresponded to “I have never seen this word before”, a ‘4’ rating corresponded to “I have seen this word, but do not know its meaning”. A ‘7’ rating corresponded to “I know this word and know at least one meaning”. Ratings of 2, 3, 5, and 6 were described individually, as intermediate points between these extremes. The instructions emphasized that this was a different task from the previous nonsense word studies, and that the task was to rate real words according to a different scale.
Responses were made using the same 7-button boxes with new labels above buttons 1, 4, and 7. A program otherwise identical to the one used for the spoken nonword stimuli presented the words visually in random order on a CRT monitor in 2-cm green lettering. There was a 1s inter-trial interval between the participants’ response and the presentation of the next word.

Following Lewellen et al. (1991), FAM scores were computed using the mean familiarity rating for each 50-word subset, Fam High, Fam Mid, and Fam Low. Replicating their earlier results, we found that Fam High did not produce much variance to differentiate between individuals, so this score was discarded. In order to use the maximum number of stimuli and still retain some variance, we averaged the Medium and Low-familiarity items together to produce a composite index score, Fam ML. This Medium and Low-familiarity score was used as a measure of an individual’s lexical knowledge. Correlations between Fam ML and performance on the recognition memory task (Experiment 3) are discussed in the next section. Correlations between Fam ML and performance in the subject rating tasks are discussed below in the section Individual Differences in Wordlikeness Judgments.

Individual Differences in Recognition Memory

As noted earlier, participants in Experiment 3 followed the recognition memory task with the modified FAM test just described. Table 1 shows correlations between hit rate, false alarm rate, and d’ for each participant against Fam ML. Correlations are shown for all stimuli together, as well as for stimuli in each probability and length group. Table 1 shows that participants with greater lexical knowledge actually tended to have lower d’ scores overall. More specifically, this difference is due to a strong relationship between Fam ML and false alarm rate – subjects with greater lexical knowledge tended to respond “old” to new stimuli more often than those with less lexical knowledge ($r = +0.51, p<0.01$). The increase in false alarm rate was greatest with the low constituent probability nonword stimuli. Also, shorter stimuli showed larger differences in false alarm rates between participants. Thus, it appears that participants with greater lexical knowledge displayed a shift in criteria relative to participants with less lexical knowledge, especially for low constituent probability stimuli.

Table 1

<table>
<thead>
<tr>
<th>Measure</th>
<th>High 2</th>
<th>High 3</th>
<th>High 4</th>
<th>Low 2</th>
<th>Low 3</th>
<th>Low 4</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hits</td>
<td>0.17,ns.</td>
<td>0.01,ns.</td>
<td>-0.12,ns.</td>
<td>0.15,ns.</td>
<td>0.15,ns.</td>
<td>0.02, ns.</td>
<td>0.13, ns.</td>
</tr>
<tr>
<td>FAs</td>
<td>0.24,ns.</td>
<td>0.26,ns.</td>
<td>0.08,ns.</td>
<td>0.37,***</td>
<td>0.38,***</td>
<td>0.27,ns.</td>
<td>0.51, ***</td>
</tr>
<tr>
<td>d’</td>
<td>-0.12,ns.</td>
<td>-0.06,ns.</td>
<td>-0.17,ns.</td>
<td>-0.18,ns.</td>
<td>-0.14,ns.</td>
<td>-0.27,ns.</td>
<td>-0.29, ns.</td>
</tr>
</tbody>
</table>

* $p < .10$  ** $p < 0.05$  *** $p < 0.01$

These results suggest that participants differed in the extent to which they analyzed the nonword stimuli into their constituents (as opposed to treating the stimuli holistically). Since the stimuli contained several instances of each onset and rime constituent, a more analytic approach would lead to a greater false alarm rate overall, as familiar constituents would be falsely recognized in new stimuli (Guttensttag & Carroll, 1997). In general, participants with greater lexical knowledge appear to be more analytic, while participants with less lexical knowledge appear to be more holistic in their judgment of these nonword patterns.
The extent to which stimuli are treated analytically is also a function of the nature of the stimuli. Sendmeier (1987) found that longer nonwords are judged for similarity based on more holistic features. This trend was observed in our nonwords as well. There were fewer individual differences for longer (4-syllable) stimuli, suggesting that all participants were treating these stimuli more holistically. The extent to which there were individual differences was influenced by the probability of the constituents, as larger differences were found with the nonwords containing low probability constituents. This suggests that all of the participants treated the stimuli with high probability stimuli analytically, due to the commonality of the constituents, while only the participants with greater lexical knowledge treated the low probability stimuli more analytically. The participants with greater lexical knowledge presumably have greater familiarity with these low probability constituents since they know more rare words (Landauer & Streeter, 1973).

**Individual Differences in Wordlikeness Judgments**

Additional evidence for individual differences in nonword processing can be found in the subjective rating tasks. In Experiment 2, where participants made accept/reject judgments of nonword acceptability, the FAM test was given after the acceptability task. There were significant individual differences between participants in this task which we present below. The FAM test was not given to the participants in Experiment 1. While the pattern of ratings were very similar between the 7-point wordlikeness task and the two alternative accept/reject task, it is possible that these tasks differ in their sensitivity to individual differences. In order to examine individual differences in the 7-point wordlikeness task, we conducted a fourth experiment in which participants had the same wordlikeness judgment task as in Experiment 1 and also took the FAM test.

**Experiment 4**

**Method**

*Participants.* Twenty-five undergraduate students earning experimental credit for introductory psychology courses at IU participated in this experiment. All reported that English was their first language and that they had no history of speech or hearing disorders. We removed one subject’s data from the final analysis because of a failure to follow instructions during the rating task. This left 24 individuals for analysis.

*Stimuli and Procedure.* This experiment consisted of two parts: the previously described nonword rating task, followed by the modified FAM test described above. For the rating part of the experiment, the stimuli, apparatus, and testing procedure were identical to Experiment 1 with the exception that all 24 subjects were presented with each stimulus only one time per trial.

**Results**

*Ratings.* Mean ratings in Experiment 4 were examined as a replication measure for Experiment 1. The ANOVA revealed significant effects of constituent probability across subjects, $F(1,138) = 185, p < 0.001$, and across stimuli, $F(1,138) = 426, p < 0.001$, and length across subjects, $F(2,138) = 49.5, p < 0.001$, and across stimuli, $F(2,138) = 103, p < 0.001$. There was a significant interaction between constituent probability and length across stimuli, $F(2,138) = 3.6, p < 0.05$, but not across subjects $F(2,138) = 2.3$, ns., as in Experiment 1. The correlation between log probability for each stimulus and that pattern’s mean rating was $r = +0.87, p < 0.001$. Correlation with the number of ‘implicit rejections’ was even stronger, $r = -0.90, p < 0.001$, replicating the earlier results obtained in Experiment 1.


**Lexical Familiarity.** The individual differences in the 7-point wordlikeness task used in Experiment 4 and the two-alternative accept/reject task in Experiment 2 are very similar, so the results are presented together here. Table 2 shows correlations between mean familiarity ratings for Medium and Low familiarity items on the modified FAM test (Fam ML) with the mean ratings and mean number of implicit rejections from Experiment 4 and mean number of “Yes” responses from Experiment 2. For mean wordlikeness ratings in Experiment 4, significant or near-significant positive correlations were found for all except high probability constituent 2-syllable items. Participants with greater lexical knowledge gave higher wordlikeness ratings on the 7-point scale, $r = +0.39$, $p < 0.05$, especially for items of lower probability. These higher ratings are also reflected in negative correlations between Fam ML and the number of implicit rejections. Participants with greater lexical knowledge gave fewer implicit rejections in Experiment 4, $r = 0.29$, ns. Since low-probability nonwords received more implicit rejections overall, the decrease in implicit rejections related to lexical knowledge is more noticeable for low-probability stimuli. In the two-alternative accept/reject task in Experiment 2, participants with greater lexical knowledge gave relatively fewer “Yes” responses to stimuli with high probability constituents, and relatively more “Yes” responses to stimuli with low probability constituents.

**Table 2**

<table>
<thead>
<tr>
<th>Measure</th>
<th>High 2</th>
<th>High 3</th>
<th>High 4</th>
<th>Low 2</th>
<th>Low 3</th>
<th>Low 4</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wordlikeness</td>
<td>0.06,ns.</td>
<td>0.30,ns.</td>
<td>0.42,**</td>
<td>0.43,**</td>
<td>0.22,ns.</td>
<td>0.41,**</td>
<td>0.39,*</td>
</tr>
<tr>
<td>Implicit Rejections</td>
<td>-0.02,ns.</td>
<td>-0.18,ns.</td>
<td>-0.14,ns.</td>
<td>-0.31,ns.</td>
<td>-0.30,ns.</td>
<td>-0.30,ns.</td>
<td>-0.29,ns.</td>
</tr>
<tr>
<td>Acceptability</td>
<td>-0.16,ns.</td>
<td>-0.18,ns.</td>
<td>-0.11,ns.</td>
<td>0.09,ns.</td>
<td>0.23,ns.</td>
<td>0.11,ns.</td>
<td>-0.02,ns.</td>
</tr>
</tbody>
</table>

* $p < .10$ ** $p < 0.05$ *** $p < 0.01$

Just as in the recognition memory task, individual differences in the subjective rating tasks reveal a shift in criteria for judgments of wordlikeness for participants with greater lexical knowledge. Once again, our results suggest that participants with greater lexical knowledge are treating the nonword stimuli more analytically. That is, participants with greater lexical knowledge are giving relatively higher wordlikeness ratings or accepting some of the low probability stimuli, indicating that these participants are able to discriminate differences between these nonword stimuli. The participants with less lexical knowledge are unable to discriminate differences among these stimuli, giving the lowest rating or rejecting all of the low probability stimuli.

Further support for this account can be found by comparing the degree to which participants’ ratings in Experiment 4 agree with the log product probability of the stimuli generated from the stochastic grammar. If the subjects with greater lexical knowledge are better able to discriminate differences in the low probability stimuli, then there should be higher correlations between probability and ratings for these stimuli. Table 3 shows the relationship between lexical knowledge (Fam ML) and the correlation between log product probability and ratings for each participant in Experiment 4. As predicted, the participants with greater lexical knowledge (larger lexicons) tended to give ratings which were more reflective of the log product probability for low probability stimuli. Among shorter, higher probability stimuli, this pattern was reversed. For the shorter, higher probability stimuli, the participants with less lexical knowledge (smaller lexicons) gave ratings which were more reflective of the log...
product probability. The participants with smaller lexicons used the entire rating scale for the highest probability items and gave implicit rejections to almost all low probability stimuli. These participants displayed a strong floor effect in their ratings overall. The participants with larger lexicons were better able to differentiate the low probability stimuli leaving less of the rating scale for differentiating the high probability stimuli. Thus, it appears that the participants with larger lexicons analyzed the longer, low-probability stimuli into their constituents, while the participants with smaller lexicons rejected these stimuli outright based on some global, holistic judgment. It is important to note here that the low probability stimuli were also the ones treated more analytically by participants with greater lexical knowledge in the recognition memory task (Experiment 3), so these two tasks provide converging evidence for the importance of systematically examining individual differences in the processing of nonwords.

Table 3

<table>
<thead>
<tr>
<th>Measure</th>
<th>High 2</th>
<th>High 3</th>
<th>High 4</th>
<th>Low 2</th>
<th>Low 3</th>
<th>Low 4</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>P/R Corr.</td>
<td>-0.53,***</td>
<td>-0.37,*</td>
<td>0.30,ns.</td>
<td>-0.56,***</td>
<td>0.10,ns.</td>
<td>0.20,ns.</td>
<td>-0.27,ns.</td>
</tr>
</tbody>
</table>

* p < .10  ** p < 0.05  *** p < 0.01

Our findings are consistent with two other studies of individual differences in the processing of nonwords which drew very similar conclusions. In a nonword similarity rating task, Sendmeier (1987) compared the performance of na"ive participants with participants who had some academic training in phonetics. Not surprisingly, he found that participants with phonetic training were able to discriminate the nonword stimuli along more perceptual dimensions than na"ive subjects, and concluded that the participants with phonetic experience were treating the stimuli more analytically than the na"ive participants.

More recently, Treiman, Fowler, Gross, Berch, and Weatherston (1995) compared the performance of students from two different universities (Dartmouth and Wayne State University) on novel word games. These games involved transposing various sets of segments between two words to produce novel strings. In some games, the sets of segments corresponded to onset and rime constituents, and in other games, the segments were syllable codas (a sub-part of the rime) or syllable initial consonant-vowel sequences which did not correspond to any structural constituent. Treiman et al. found that students from Dartmouth who had higher mean verbal SAT scores were better at the games overall and also showed fewer effects of the type of constituent which was to be manipulated than students from Wayne State. The authors concluded that the Dartmouth students, with better verbal skills, were better able to analyze the words into segments and manipulate the internal segmental units. The Wayne State students who had poorer verbal skills were able to manipulate the larger constituents of onset and rime but had more difficulty with the individual segments.

Together, these studies of individual differences in the processing of nonwords provide evidence that participants may vary greatly in their skill in perceiving and processing nonword sound patterns. These differences have important theoretical consequences in psycholinguistics. The phonological grammar employed in this study is based on probabilistic generalizations made across the inventory of lexical items and so can accommodate individual differences in our participants’ responses in the metalinguistic wordlikeness task. Traditional phonological theories predict no differences based on knowledge of individual lexical items, as the productive component of linguistic competence is not the
lexicon but a store of very general rules or constraints. Evidence that an individual’s experience is an important component of processing suggests that learning and environment play an important role in shaping cognitive abilities. Clearly, individual differences are an important topic for further research as such differences may provide important new insight into fundamental theoretical questions in cognitive science.

**General Discussion**

Across four experiments we have obtained consistent and clear support for a relationship between phonotactic probability, in terms of a stochastic grammar, and subjective ratings of wordlikeness or acceptability for nonword sound patterns. The stochastic grammar incorporates prosodic features such as stress pattern, syllable structure, and syllable position into the probability of constituents, and therefore makes specific claims about the use of this information in direct (rating) and indirect perception of nonwords. Stimuli generated with such a grammar can be used to examine whether effects of length (as in our study), lexical competition, morphemic composition, or stress pattern, are separate, or separable, from the effects of constituent probability. We see this as a useful experimental tool, both for the creation of novel nonword stimuli with specific properties (to be contrasted against real words) and for the examination of how those properties affect nonword perception and memory performance in a range of psycholinguistic tasks.

Subjective judgments of wordlikeness were examined in several different tasks. Beyond a lowest limit, wordlikeness judgments failed to discriminate between very low probability nonsense sound patterns. Participants reached the bottom of the 7-point scale in Experiment 1, and, in Experiment 2, very few subjects would accept any of the lowest probability patterns. These floor effects were replicated in Experiments 3 and 4. In the post-hoc analysis we referred to as implicit rejections (the number of ‘1’ ratings from the 7-point scale given to a particular stimulus or group of stimuli), we found that these very low probability nonwords can be differentiated if a sufficient amount of data is aggregated. This suggests to us that listeners may perform more than one task when judging wordlikeness on a continuous scale; they must decide whether an item is even acceptable at all and they must decide to what extent the acceptable items are wordlike. In the accept/reject acceptability task in Experiment 2, the same two processes appear to be at work. With only two possible responses, “No” was used both for ‘unacceptable’ and for ‘not very wordlike’, resulting in a similar aggregate pattern for subjective wordlikeness judgments regardless of the available response options.

The consistency of the wordlikeness judgments across all these experiments provides converging evidence that these judgments are made by comparing nonwords to similar, familiar word patterns in the participants’ mental lexicons. In other words, wordlikeness judgments are based on the participants’ actual experience with words. Not only were words containing more frequent, more familiar constituents judged as more wordlike, but there is clear evidence for individual differences in the rating task. In Experiment 1, participants who were given an extra presentation of each stimulus gave the stimuli slightly higher ratings than participants who heard the stimuli only once. The extra presentation increased the sense of wordlikeness, perhaps by briefly enhancing the amount of activation of similar, familiar lexical items for those participants. In addition, individual differences in lexical knowledge influenced ratings. In Experiment 4, participants who knew more uncommon real words, compared to participants who knew fewer uncommon words, rated all of the nonword patterns as more wordlike, especially those with the most uncommon constituents.

The results of Experiment 3, comparing these same nonword stimuli in a recognition memory task, provide further evidence that the perception of nonword stimuli makes use of knowledge about familiar word patterns. Nonwords with high probability constituents were more easily recognized than
nonwords with low probability constituents. The differences in estimated string probability between items of the same probability group that are caused by length did not have the same effect as the differences between constituent probability groups. The fact that there was no effect of length on recognition memory in this study may simply have been an artifact of our stimulus construction method, but it may also represent a dissociation between the stochastic grammar used to rate stimuli in metalinguistic wordlikeness judgments, and the actual perception and processing of those same items in linguistic tasks. Further research is needed to determine whether this is a genuine effect of the task or whether it is stimulus-dependent. Individual differences in false alarm rates also demonstrate that the perception of nonwords utilizes knowledge about actual words. Participants with greater lexical knowledge had higher false alarm rates for low probability stimuli, an effect that is probably due to greater familiarity with actual words which contain some of the low probability constituents. In short, the present set of findings demonstrate that participants judge nonwords against a background of knowledge about sound patterns in their language (Goldinger, 1998; see also Glushko, 1979, for similar conclusions regarding visually presented nonwords).

Conclusions

The results of these experiments demonstrate that metalinguistic judgments of wordlikeness are heavily influenced by the relative frequency of occurrence of sounds in the lexicon. A stochastic grammar like that proposed by Coleman and Pierrehumbert (1996) can be used to generate nonwords of relatively high and low subjective wordlikeness, as well as to predict the subjective wordlikeness of a novel sound pattern, using the dictionary as an approximation of the lexicon. However, we also found there is a point where human listeners can no longer differentiate individually between very low-probability nonwords, suggesting that a gradient stochastic grammar may only be appropriate for more wordlike patterns. As noted by Pierrehumbert (1994), there may be a cut-off point where items below a certain probability are ungrammatical and therefore not wordlike at all. The location of this cutoff varies among individuals, as we found there were individual differences in the location and degree of a floor effect in wordlikeness ratings. Above this cutoff, participants produce consistent, reliable ratings of wordlikeness.

We also found that phonotactic knowledge is reflected in linguistic tasks such as recognition memory, where nonwords with highly probable constituents are recognized more accurately from previous presentations. Note that these findings are precisely the opposite of the results obtained with high frequency real words, but are consistent with other research on the processing of nonwords (Chalmers & Humphreys, 1998; Vitevitch & Luce, 1998). It is important to consider both the processing of words and the processing of nonwords in behavioral tasks, as differences between lexical and non-lexical aspects of language processing can provide important insights into the underlying source for an observed effect (Frisch, 1996; Glushko, 1979; Treiman, 1988; Treiman et al., in press; Vitevitch & Luce, submitted).

Our findings are also consistent with a growing body of research on the interaction between probabilistic phonotactics and language processing (Aslin, Saffran, & Newport, 1998; Frisch, in press; Jusczyk et al., 1994; Treiman et al., in press; Vitevitch et al., 1997; Vitevitch & Luce, submitted; inter alia). We have extended the domain of inquiry for this line of research by showing that individual differences in lexical knowledge influence familiarity with the constituents which are found in novel words and affect performance on tasks employing this knowledge. Greater lexical knowledge in this case, greater knowledge of rare, low-familiarity real words (i.e., a larger lexicon) was associated with a more analytic mode of perception for nonword stimuli. Taken together, the present results demonstrate that the lexicon contains many sources of information about the sound patterns and sequences of spoken words in the language. This information appears to be accessed and used not only in tasks employing real
words that are known to the subject but also in tasks that employ novel nonwords that contain patterns, constituents, and regularities that reflect the probabilistic phonotactics of the grammar.

References


