Published in final edited form as:

Lab Phonol. 2011 May ; 2(1): 157–183. doi:10.1515/labphon.2011.005.

Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults

Peter T. Richtsmeier

University of Kansas

Abstract

Phonotactics—the permissibility of sound sequences within a word—correspond to lexical statistics, but controversy persists over which statistics are being tracked. In this study, lexical type and token counts were compared as they contributed to phonotactic extraction from an artificial lexicon. Young-adult participants were familiarized with a set of CVCCVC nonwords contextualized as a lexicon of Martian animal names. The type and token frequencies of word-medial consonant sequences within those names were varied systematically. Participants then rated new nonwords, containing the same medial sequences, on a 7-point scale for similarity to the Martian animal names. Higher ratings only followed high type frequency familiarization conditions, suggesting that word-types drove phonotactic extraction. Additionally, participants reversed the typical preference for high frequency English sequences, likely because they rated nonwords according to their membership to an unknown language. This finding suggests cognitively separable tracking of artificial language statistics and preexisting representations.

Keywords

language learning; phonotactics; type frequency; token frequency; talker variability; wordlikeness ratings

Phonotactic Probabilities

Phonotactics, or the restrictions on what sound sequences can compose the words of a language, have lately been the focus of research on language experience. For example, phonotactics vary from language to language (e.g., the biphone /kn/ is an allowable sequence at the beginning of Dutch words but not English words), suggesting that the subset of sequences employed by a language must be learned. Sequences also vary in frequency within languages (e.g., /st/ is frequent in English words but /fp/ is not), suggesting that learners must learn the degrees to which phonotactics are more or less typical of a language. These “phonotactic probabilities” have been observed across a number of languages, including Cantonese (Kirby & Yu, 2007), Dutch (Kager & Shatzman, 2007), Jordanian Arabic (Frisch & Zawaydeh, 2001), Mandarin (Myers & Tsay, 2005), and in Spanish-English bilinguals (Frisch & Brea-Spahn, 2010). In fact, phonotactic probabilities

have been documented for several decades (e.g., Greenberg & Jenkins, 1964; Hammond, 2004; J. J. Ohala & Ohala, 1986).

Probabilistic phonotactics have also been demonstrated in the linguistic knowledge of infants and young children. In a study of infant perception, Jusczyk, Luce, and Charles-Luce (1994) showed that English learning infants prefer words like /tʃʌn/, which are composed of high frequency sequences, over words like /giθ/, which are composed of low frequency sequences. In child speech production, three- to five-year-old children produce high frequency English phonotactic sequences (e.g., /sn/ in /busnədi/) more accurately than they produce infrequent sequences (e.g., the /bn/ in /subnədi/; Beckman & Edwards, 1999). Similarly, two-year-old children produce coda consonants more or less accurately depending on the frequency of the preceding consonant-vowel sequence (Zamuner, Gerken, & Hammond, 2004).

Language learning studies examine how phonotactics are subject to learning, thereby revealing what makes phonotactics probabilistic. For example, infants can learn categorical consonantal restrictions to onsets and codas within an experiment (Chambers, Onishi, & Fisher, 2003; Seidl & Buckley, 2005), as can adults (Dell, Reed, Adams, & Meyer, 2000; Warker, Xu, Dell, & Fisher, 2009). Goldrick (2004) found that participants who produced strings of words from an artificial language learned both consonantal and featural phonotactic restrictions related to word onsets and codas. The featural restrictions were relative, however. For example, a place of articulation feature like labio-dental was slightly more frequent as an onset than as a coda. Participants' production errors reflected both this relative frequency, indicating phonotactic probabilities, as well as the overall acceptability of labio-dental consonants. In sum, studies of learning, acquisition, and cross-linguistic consistency provide a detailed picture of how linguistic experience shapes phonotactic knowledge.

The Evidence for Phonotactic Probabilities

Based on these findings, researchers have begun to ask what evidence learners rely on to learn phonotactic probabilities. Language learners must be tracking statistics in the language input, but it is not yet clear which statistics they are tracking. One hypothesis has come out of the speech development literature. To explain why children with larger vocabularies have more accurate speech, Edwards, Beckman, and Munson (2004) propose that phonotactics are generalizations across 'word-types', or word representations that have unique phonological and semantic properties. Types are necessarily abstract entities because they are defined by abstract qualities such as phonemic composition. Elsewhere, philosophical and theoretical work includes arguments for the existence of word-types (Pierrehumbert, 2003; Wetzel, 2009), and some empirical work also favors types for phonotactic learning. For example, Bailey and Hahn (2001) asked adults to rate a set of nonwords for their similarity to real English words. Regression analyses indicated that type frequencies were the best predictors of ratings (cf. Albright, 2009, for a similar analysis and Bybee, 1995, for evidence that type frequencies dominate morphological pattern learning).

However, other research has focused on ‘word-tokens’, or unique instances of a word in time and space. These studies provide strong evidence that we store and recall word-tokens (Goldinger, 1996, 1998; Nygaard, 2008). Such studies often look at repetition priming, or the finding that responses to a stimulus are faster when that stimulus has been encountered previously. For example, participants are faster to identify a word (the target) that they have heard previously in the experiment (the prime). Importantly, repetition priming is modulated by token-specific details of the prime: In both recognition memory and perceptual identification tasks, Goldinger (1996) observed that adults were fastest when the same talker produced the prime and the target. Talker-specific priming was also robust across long time periods, and Goldinger’s participants were sensitive to a particular talker’s prime for at least a week. Repetition priming is also sensitive to a prime’s token frequency, and faster responses to a target are a function of the number of primes that have been previously encountered (Grant & Logan, 1993). Clearly, word-tokens play an important role in language processing.

Logically, types imply tokens; any instance of a type is also a token. Furthermore, a developmental perspective requires token-based learning because word-types are language-specific entities. Therefore, word-types must be formed from concrete, language-specific tokens. In fact, some research suggests that word-token effects subsume word-type effects. A review of priming studies by Tenpenney (1995) found little evidence for priming of abstract, type-based representations. There is also evidence that token frequencies contribute to phonotactic probabilities, at least marginally (Bailey & Hahn, 2001; Frisch & Zawaydeh, 2001). Furthermore, the wordlikeness rating studies above (Albright, 2009; Bailey & Hahn, 2001), which appear to support the existence of types, quantified the effect of token frequency using a somewhat marginal parameter compared to the more central word-type statistic. Perhaps a more central role for tokens, or the incorporation of token-specific properties such as talker characteristics, would yield even larger token effects. To this end, Goldinger (1998) proposes and tests a model of the lexicon that consists only of tokens, eliminating the need for type-based representations. Returning to the question of what statistics underlie phonotactic probabilities, both types and tokens are implicated.

Learning Phonotactics from Types and Tokens in an Artificial Lexicon

Unfortunately, little has been published in the way of direct comparisons between type- and token-based statistical learning, and methodological differences across the wordlikeness and priming literatures inhibit progress. Still, some methodologies seem sensitive to both types and tokens. For example, artificial language learning studies have reported effects of both types and tokens. Regarding tokens, Richtsmeier, Gerken, Goffman, and Hogan (2009) asked how a manipulation of a word’s token frequency during familiarization would influence production accuracy in four-year-olds. Although token frequency did influence production speed and accuracy, it only did so when children were familiarized with phonetically variable words (i.e., words produced by different talkers. The finding suggests that phonetic variability was the key component to learning from token statistics). Facilitative effects of phonetic variability have also been found in word monitoring studies with infants (Houston & Jusczyk, 2003; Singh, 2008).

Word-types have also been studied in an artificial language learning paradigm. In a follow-up to the study above, Richtsmeier, Gerken, and Ohala (in press) familiarized four-year-old children with sets of words containing target word-medial consonant sequences (henceforth sequences) that also controlled type and token frequencies. Following familiarization, children produced the same sequences in new words. Accuracy improved when children were familiarized with the sequences in multiple word-types spoken by multiple talkers,¹ but not when children were familiarized with multiple talker-tokens of a single word-type, or multiple word-types each spoken by a single talker. The results point to an interaction between types and tokens: Types clearly supported phonotactic generalization, but tokens also played a role, possibly helping children to establish robust word representations and enabling higher-order generalizations (cf. Pierrehumbert, 2003).

One value of artificial language learning studies is their ability to simulate frequency effects in a real lexicon. For example, the combined type and token frequency effect observed in Richtsmeier et al. (in press) are similar to the phonotactic probability effects seen in child speech accuracy (Beckman & Edwards, 1999; Edwards et al., 2004). That is, children were more accurate to produce sequences they heard in multiple words spoken by multiple talkers, just as they produce high frequency English sequences more accurately. Furthermore, the study reveals the relative contributions of particular statistics (types and tokens) to learning.

To summarize, artificial language learning studies can detail how real-language frequency effects arise, such as phonotactic probabilities. They can also address the evidence language learners use for learning a particular frequency effect (Goldrick, 2004; Richtsmeier et al., in press). In the present study, adults were tested for sensitivity to type and token frequencies in an artificial grammar. The study combines the familiarization-followed-by-testing paradigm common in the infant and child literature (cf. Jusczyk & Aslin, 1995) with a wordlikeness rating task, in which adults consistently show phonotactic frequency effects (Albright, 2009; Bailey & Hahn, 2001). Using a similar set of words and manipulations employed by Richtsmeier et al. (in press), wordlikeness ratings were evaluated as a function of familiarization to type statistics, token statistics, or both.

Method

The present experiment takes up the question of how type and token frequencies influence extraction of phonotactic sequences from an artificial lexicon. Participants were first familiarized with a randomized set of word-types and/or -tokens that instantiated the target sequences. They were then presented new words containing the same sequences and were asked to rate how much the new words sounded like the familiarization words. If participants were influenced by the familiarization, it was expected that they would generalize and extract the target sequences, influencing subsequent ratings of test words containing those sequences.

¹Richtsmeier et al. (in press) only observed generalization for low English frequency sequences, which they attribute to a ceiling effect affecting the high English sequences.

Before moving to the methodological details, it may be helpful to establish some operational definitions. In this study, *phonotactic sequences* are word-medial consonant sequences, for example, the /fp/ in /mæfpəm/. *Token frequency* refers to different talker-tokens of a word. Prior research shows that infants are better able to track a word they have heard spoken by multiple talkers (Houston & Jusczyk, 2003), children's production accuracy improves as a result of hearing multiple talkers (Richtsmeier et al., 2009, in press), and adults store, remember, and access talker-specific word tokens in priming studies (Goldinger, 1996). Therefore, defining token frequency with respect to talkers captures important qualities of that statistic. *Type frequency* refers to phonologically and semantically distinct words. Phonologically, all words were composed of unique sets of phonemes, although they shared a CVCCVC shape and trochaic stress. Semantically, familiarization words were described as the names of Martian animals, and test words were described as potential Martian animal names.

The term *extraction* refers to the participants' sensitivity to the type and token manipulations of the target sequences as measured by changes to wordlikeness ratings and response times. 'Extraction' required generalization because different words were used for familiarization and testing. The term is also meant to situate the procedure somewhere between a priming study and a learning study. In fact, many researchers interpret priming as a form of implicit learning (Goldinger, 1996; Grant & Logan, 1993). I assume that a metalinguistic task like rating wordlikeness is sensitive to previously stored information, and will therefore involve some degree of priming (cf. the discussion of neighborhood density in Bailey & Hahn, 2001; Goldinger, Luce, Pisoni, & Marcario, 1992), although priming experiments to date have not considered the kinds of manipulations at work in this experiment. I also assume that wordlikeness ratings are sensitive to learning because they—and grammaticality judgments in general—change as a function of repetition (Nagata, 1988; Schütze, 1996).

Participants

One hundred thirty-nine adults were recruited from a pool of native-English-speaking undergraduates enrolled at the University of Kansas. Participants completed a short questionnaire regarding their language background, and those with extended experience with a second language (one or more years of undergraduate courses, $N = 4$) or who reported a personal history of speech, language, or cognitive difficulties ($N = 8$) were removed from the analysis. Nine participants were removed due to computer malfunction or experimenter error, and as explained in detail in the analysis section, 17 participants were removed for failing to meet criteria for attending to the task. The remaining 101 participants (46 female) were entered into the analysis.

Materials

Participants were familiarized with and asked to rate sets of nonsense words. Across familiarization and testing, the words shared word-medial consonant sequences. These sequences were the extraction targets, so discussion of the materials below generally refers to the sequences, rather than the words. Eight sequences were selected, four frequent in English: /kt/, /mp/, /sp/, and /st/; and four infrequent: /pk/, /mk/, /ʃp/, and /fp/. Medial

consonant sequences were chosen following studies by Munson (2001) and Richtsmeier et al. (in press).

Forty-eight nonwords were used in the experiment, or six words per sequence, and are given in Table 1. The words were divided into two sets. For half of the participants, the first set served as the familiarization words and the second set as the test words. Assignments were then reversed for the other participants so that a given word served for both familiarization and testing. Presentation of these words varied according to three factors of interest: English frequency, experimental frequency, and exposure group.

English frequency—High English frequency (HiEng) and low English frequency (LoEng) consonant sequences were chosen based on their occurrence in the middle of English words. HiEng and LoEng words were matched such that the same word frame was used for both. For example, the frame /dɪ---əs/ was combined with the sequences /mp/ and /mk/ to create the words /dɪmpəs/ and /dɪmkəs/. Because words were matched in this way, they were not considered randomly selected members of the population of nonwords, and by-items analyses were not conducted (cf. Raaijmakers, Schrijnemakers, & Gremmen, 1999, for criteria for selecting a by-items analysis).

Sequence frequencies were computed using the web-based Phonotactic Probability Calculator (Vitevitch & Luce, 2004, <http://www.people.ku.edu/~mvitevitch/PhonoProbHome.html>, henceforth the PPC). The PPC algorithm calculates phonotactics with respect to the word's left edge, but not syllable structure or stress. Therefore, PPC probabilities represent the likelihood of that sequence appearing as the third and fourth phonemes of a six-phoneme word. For example, /st/ is the most frequent sequence and has a probability of .0232 of occurring word-medially, whereas /ʃp/ and /fp/ never occurred in the corpus and have probabilities of .0. The full set of sequence probabilities is given in Table 2.

The PPC calculates the probabilities of phonotactic sequences with respect to token frequency. According to the PPC, the token frequencies of words containing /st/ are what determine the probability of /st/ in English. The exact method of calculating English phonotactics is not critical here, however. A type-based calculator would likely group the eight sequences in the same manner, but an observation of English syllable-based phonotactics will do just as well: The HiEng clusters /mp/, /kt/, /st/, and /sp/ can all appear as English onsets, or codas, or both, whereas the LoEng clusters cannot. Regardless of whether the English frequencies were calculated using token statistics or type statistics, the English frequency factor represents prior knowledge on the part of the subject. Here, we are most interested in the statistics that participants track during the experiment. Those statistics are relevant to the experimental frequency and exposure group factors that are discussed below. In sum, the English frequency factor represents a relatively straightforward categorization of word-medial consonant clusters, but it is not critical for interpretation of the two other factors.

Previous experiments (Albright, 2009; Bailey & Hahn, 2001) have found that participants rate words composed of high English frequency phonotactic sequences as being more wordlike than words composed of low frequency sequences. Therefore, participants might

be expected to rate words containing HiEng sequences as more wordlike than words containing LoEng sequences. However, because participants made decisions with respect to an unfamiliar language, the usual English frequency effect may not have applied. In fact, novice second language learners are occasionally biased toward novel categories. De Jong, Hao, Park, and Silbert (2008) discuss a number of examples, including evidence that Korean learners of English exhibit a novelty bias for /f/ and /ð/, which do not occur in Korean. Thus, because the participants were asked to rate the sequences based on their membership to the unknown language Martian, it was possible that the typical English frequency effect would disappear altogether or reverse.

Experimental frequency—In addition to manipulating the frequency of sequences as they occurred in English, the frequency that participants heard sequences within the experiment was also manipulated, or experimental frequency hereafter. Participants heard half of the familiarization sequences 12 times, that is, in the high experimental frequency condition (ExpHi); they heard the other half 4 times each, that is, in the low experimental frequency condition (ExpLo). A sequence's experimental frequency was manipulated during familiarization only, not test. Furthermore, experimental frequency was manipulated on the basis of sequence frequency (not word frequency), to which both type and token frequency could contribute. The number of words associated with a given sequence was a function of exposure group, which is described below.

The purpose of the experimental frequency manipulation was to simulate real-language frequency effects and thereby determine whether familiarization statistics would influence subsequent wordlikeness ratings. In Richtsmeier et al., experimental frequency resulted in more accurate consonant sequence production when children were exposed to those sequences in multiple words spoken by multiple talkers (that is, from a combination of type and token frequencies). Thus, the experimental frequency manipulation created an ExpHi frequency advantage similar to the HiEng advantage observed elsewhere (e.g., Beckman & Edwards, 1999; Edwards et al., 2004). Extrapolating from the child production data, adult participants were expected to rate ExpHi sequences as being more wordlike than the ExpLo sequences, depending on exposure group.

Exposure Group—Exposure group was the only between-subjects factor and was a manipulation of the number of talker-tokens and/or word-types that participants heard for each consonant sequence in the ExpHi familiarization condition (i.e., when participants heard a sequence 12 times during familiarization). Talker-tokens and word-types were crossed to create three conditions: one emphasizing tokens (tokens only group), one emphasizing types (types only group), and one that combined properties of both (types + tokens group). A visual representation of the three exposure groups is given in Figure 1 below.

For the tokens only group, the number of talker-tokens that participants heard for a single familiarization word varied (recall that tokens were operationalized as talker-tokens). During familiarization, participants heard ExpHi sequences in a single word spoken by 12 talkers and heard ExpLo sequences in a single word repeated from a single talker's production four times (four acoustically identical tokens, cf. Figure 1a). The words in italics in Table 1 were

used in the tokens only condition. That is, only one of the three familiarization words per sequence in a set was used in the tokens only condition.

For the types only group, the number of familiarization words corresponding to a sequence was manipulated. Participants heard ExpHi sequences in three words and heard ExpLo sequences in a single word (Figure 1b, the same words from the tokens only condition were used for the ExpLo sequences). Each word was spoken by a single talker four times. Eight words occurred within a familiarization block (2 ExpHi sequences $\times$ 3 words + 2 ExpLo sequences $\times$ 1 word), so participants heard a total of eight unique talkers during the familiarization. This is slightly fewer than the twelve different talkers heard in the tokens only group (ExpHi condition), but matched the eight talkers appearing in the types + tokens condition, as described below.

For the types + tokens group, the number of words associated with a sequence was manipulated, but multiple talkers produced each word. ExpHi sequences were heard in three words, and each word was spoken by four different talkers; ExpLo sequences were heard in a single word spoken by four talkers (Figure 1c, words from the tokens only condition were used for the ExpLo sequences).² To match the total number of talkers to the types only condition, only eight talkers were heard during a given familiarization phase. Thus, each talker produced four of the eight words in a familiarization block. Following Richtsmeier et al. (in press), phonotactic extraction, defined as increased wordlikeness ratings for sequences in the ExpHi condition, was only predicted for the types + tokens exposure group.

Participants heard a maximum of 12 unique talkers during the familiarization and a thirteenth talker during the test. Thirteen women were recruited to record the words. All talkers were native speakers of American English with a Midwestern accent. Only women were recorded to make the results comparable to previous research that varied exposure to talkers (Richtsmeier et al., in press). The talkers were recorded in an anechoic chamber at a sampling rate of 44.1 KHz. They were given a prompt by a trained phonetician (the author) and asked to produce three or more tokens of each word using careful speech with clear cues for the component phones, particularly the two phones forming the medial sequence. For example, talkers were asked to release the first stop of the /pk/ and /kt/ sequences and aspirate the second stop of the /sp/ and /st/ sequences.³ The words were produced with emphasis on the first syllable, or trochaic stress. From these recordings, a single production of each word from each talker was chosen for use in the experiment. The selected productions were extracted from the original recording with 100 ms of silence on

²Because four talkers were heard for the ExpLo types + tokens condition, it differs from the single talker used in the ExpLo types only and tokens only conditions. This difference was carried over from the experiments by Richtsmeier et al. (2009a), who wanted to emphasize the importance of word-types compared to talker-tokens. In other words, for the types + tokens condition, any differences between the ExpHi and ExpLo would be most appropriately attributed to the difference in word-types given that both ExpHi and ExpLo sequences were spoken by multiple talkers. Unfortunately for the present study, this puts the exposure group factor slightly out of balance. Given that the tokens in question represented just a third of the overall exposure phase, I did not expect that this difference contributed substantially to the results.

³Phonetically, stops following fricatives, as in the sequences [st] and [sp], are often unaspirated and indistinguishable from [d] and [b], respectively, to native English speakers. Although aspirating these stops is somewhat unnatural for English, talkers produced them in this way to ensure maximum clarity for the component phones of the target sequences. In this and in a similar, unpublished experiment, participants were randomly and informally polled about their opinion of the experiment. They were asked if they found the task to be highly unusual. No participant reported finding the task or the words unnatural. The possibility remains that the results reflect phonetically unnatural forms, but such an effect would presumably apply to all exposure groups.

either side, and were normalized for root-mean-square amplitudes. In order to make the sound files compatible with the experimental software, they were resampled at 16 KHz.

Several steps were taken to eliminate confounding factors, including differences between familiarization and test words, associations between words and exposure groups, order effects based on blocking, and the neighborhood density of the words. A full description of these steps can be found in Appendix A.

Procedure

As a context for learning the artificial lexicon, participants were told that they would be completing a language learning experiment involving a language called “Martian”, and the words were characterized as the names of Martian animals. The experimenter explained that there were two parts to the experiment. The first part was the familiarization during which participants heard a series of Martian animal names; the task was to watch and listen. The second part was the wordlikeness rating task during which participants heard potential Martian animal names; the task was to rate how much the potential names sounded like the Martian animal names from the familiarization. Within a test block, participants rated test words containing the same sequences in the familiarization words. Ratings were made on a Likert scale (cardinal numbers 1 to 7), where 1 meant “Definitely not a Martian animal name”, 4 meant “Neutral”, and 7 meant “A great Martian animal name”. Intervening numbers were also given labels for reference. In addition to the verbal description, the participants also saw a written description of the experiment when they arrived at a computer station. Ratings and response times were collected, although participants were not told to respond quickly or within any given time period.

The procedure was repeated across two experimental blocks, or two blocks of familiarization then testing. Four sequences appeared in each experimental block. The purpose of this design was to avoid strain on participants’ attention spans and to separate presentation of words with the same frame, such as /dɪmpəs/ and /dɪmkəs/.

Consider the procedure for a participant in the types-only exposure group who received List 1A. During the first familiarization block, the participant heard the sequence /mp/ in the words /nʌmpət/, /gumpən/, and /ʃampən/, the sequence /kt/ in the words /zerktəs/, /saʊktəs/, and /pʌktəm/, the sequence /fp/ in /mæfpəm/, and the sequence /ʃp/ in /keʃpəs/ (eight words total). One talker-token of each word was repeated four times, but each of the eight words was spoken by a different talker. All 32 familiarization tokens (eight words × four tokens) were presented in random order. Following familiarization, the participant rated 12 test words (four sequences × three words). For the /mp/ sequence, participants rated the words /dɪmpəs/, /koɪmpət/, and /sæmpəf/; for /kt/ they rated /bɪktəm/, /ləktəf/, and /tuktən/; for /fp/ they rated /neɪfpən/, /poɪfpən/, and /ʃeɪfpək/; and for /ʃp/ they rated /zaʃpəm/, /foʃpəm/, and /lɪʃpət/. Test words were also presented in random order. The second experimental block was structured identically but focused on the sequences /st/ and /sp/ (both ExpLo), and /mk/ and /pk/ (both ExpHi).

The presentation was controlled by Psyscope software, version 1.2.2 (Cohen, MacWhinney, Flatt, & Provost, 1993). Familiarization ordering was randomly generated by the software,

with no pause between presentations, although the 100 ms of silence on either side of the sound files effectively created a 200 ms pause between each word. During familiarization, words were paired with colorful, hand-drawn, make-believe animals (D. K. Ohala, 1999) whose presence on the screen was time-locked to the onset and offset of the audio file. Test words were presented once each in random order, but with a blank white screen. Once the test word was played, the rating scale appeared at the bottom of the screen and stayed visible until the participant pressed one of the relevant number keys. Participants were given as much time to make their decisions as they needed. After participants made their selection, the next test trial began. Following the test, the experiment proceeded to the second set of familiarization and test blocks and then ended. Participants took about 8 minutes on average to finish the experiment.

Analysis

To ensure that the data came from participants who were actively engaged in the experiment, a combination of two exclusionary criteria were applied to the results. If participants made the same rating five or more times in a row ($N = 7$), if they made four or more ratings less than 300 ms after the end of the test word ($N = 6$), or if they had both short rating times and repetitive ratings ($N = 4$), they were excluded from the analysis. For the remaining 101 participants, ratings made in under 300 ms were removed from the analysis. A total of 33 data points were removed in this way, with no more than two points removed per participant.

Results

Ratings were collapsed across words to derive four data points per participant, one for each of the within-subjects conditions, HiEng + ExpHi, HiEng + ExpLo, LoEng + ExpHi, and LoEng + ExpLo. Ratings ($M = 4.35$, $SD = .85$) were then entered into a 2 English frequency $\times$ 2 experimental frequency $\times$ 3 exposure group mixed design ANOVA. A graph of the results is given in Figure 2.

As expected, there was a main effect of English frequency, $F(1, 98) = 30.65$, $p < .001$, $\eta_p^2 = .24$ (partial eta squared values of .2 or less reflect relatively small effect sizes; Cohen, 1988), although the direction of the effect was somewhat unexpected: Participants consistently rated words containing LoEng sequences as more Martian-like than words with HiEng sequences (compare the HiEng and LoEng column groups across all three exposure groups in Figure 2). There was also a significant effect of exposure group, $F(2, 98) = 3.28$, $p = .05$, $\eta_p^2 = .06$. Probing contrasts (with $\alpha = .05$) among the tokens only, types only, and types + tokens groups ($M_s = 4.18$, 4.47, and 4.47, respectively), there were significant differences between the tokens only and types only groups ($p = .02$), the tokens only and types + tokens groups ($p = .05$), but not between the types only and types + tokens groups ($p = .79$). Thus, participants rated test words as more Martian-like if they were familiarized with Martian in the types only or the types + tokens groups. It is not clear at present why this difference should exist.

Although there was an effect of experimental frequency, $F(1, 98) = 4.20$, $p = .04$, $\eta_p^2 = .04$, there was also a significant experimental frequency $\times$ exposure group interaction, $F(2,$

98) = 3.42, $p = .04$, $\eta_p^2 = .07$.⁴ Looking at each exposure group individually, experimental frequency did not influence participants' ratings in the same way. Words with ExpHi sequences were rated as more Martian-like in the types only group, $F(1, 33) = 7.17$, $p = .01$, $\eta_p^2 = .178$, and in the types + tokens group, $F(1, 33) = 4.22$, $p = .05$, $\eta_p^2 = .113$, but not in the tokens only group, $F(1, 32) = .93$, $p = .34$, $\eta_p^2 = .03$ (compare Figure 2b and 2c to 2a).

Figure 2 also suggests that ExpLo ratings were made relative to ExpHi ratings because the ExpLo ratings were not equivalent across exposure groups. However, when the ExpLo ratings were tested alone, there was not a significant effect of exposure group ($F < 2$). Relative ratings are nevertheless sensible given the relatively subjective task and exposure only to the clusters to be rated. In other words, there was not an anchor point common to all three exposure groups which would allow participants to make uniform ExpLo ratings.

Returning to the experimental frequency $\times$ exposure group interaction, participants only extracted phonotactic sequences when those sequences were supported by high type frequency. The null result in the tokens only exposure group should not be interpreted directly, but the interaction indicates that participants in the tokens only group were rating words differently than participants in the types only and types + tokens groups. Additional support for this conclusion appears in Figure 3, which provides the results for individual sequences. Generally, higher ratings were given to ExpHi sequences in the types only and types + tokens groups but not in the tokens only group. The results therefore suggest that type frequency, not token frequency, allowed participants to generalize and extract the target sequences.

Interestingly, the direction of the experimental frequency effect (for the types only and types + tokens exposure groups) is in the opposite direction of the English frequency effect. If we consider the 'high frequency' conditions—HiEng and ExpHi—high English frequency led to lower ratings whereas high experimental frequency led to higher ratings. This difference in directionality is described in greater detail in the Discussion section.

There was no English frequency $\times$ experimental frequency interaction or a three-way interaction ($F_s < 1$).⁵

Richtsmeier et al. (in press) found that four-year-old children produced consonant sequences more accurately if they heard them in multiple words spoken by multiple talkers (similar to the types + tokens exposure group here), but not if they heard the words spoken by a single talker (similar to the types only group). The child production data make the results

⁴In an analysis that included participants with short response times, this interaction was only marginally significant, $F(2, 104) = 2.45$, $p = .09$. The effects of experimental frequency for individual levels of exposure group were essentially the same.

⁵Response times were also subjected to an ANOVA. The results included a significant main effect of English frequency, $F(1, 98) = 4.79$, $p = .03$, $\eta_p^2 = .05$, but not of experimental frequency or of exposure group ($F_s < 1$). As in the ratings analysis, the English frequency effect was also in the opposite direction of what is normally expected—participants were slower to rate words containing high English frequency sequences ($M = 1940.0$ ms, $SD = 843.5$) compared to words containing low English frequency sequences ($M = 1839.8$ ms, $SD = 812.2$). There was a marginal interaction of experimental frequency $\times$ exposure group, $F(2, 98) = 2.59$, $p = .08$, $\eta_p^2 = .05$. Looking at the individual exposure groups, there were no significant effects of experimental frequency (all $p_s > .1$). Participants were numerically faster to respond to ExpHi words in the types only (70 ms ExpHi advantage) and types + tokens groups (142 ms ExpHi advantage) but slower in the tokens only group (74 ms ExpHi disadvantage). All other interactions in the main analysis failed to reach significance ($F_s < 1$).

for the types only exposure group here somewhat surprising, so a post-hoc analysis was conducted to directly compare the types only and types + tokens groups. As in the full analysis, there were main effects of English frequency, $F(1, 66) = 25.03$, $p < .001$, $\eta_p^2 = .28$, and of experimental frequency, $F(1, 66) = 11.11$, $p < .01$, $\eta_p^2 = .14$, but not of exposure group ($F < 1$). The results therefore suggest that the two exposure groups were equivalent in their ability to facilitate generalizations about Martian phonotactics, and contrary to expectations based on the child study, phonetic variability was not necessary for adults to make those generalizations.

Discussion

The present study replicated previous findings of phonotactic learning within an experiment (e.g. Chambers et al., 2003; Goldrick, 2004; Richtsmeier et al., in press). Here, participants tracked the experimental frequencies of eight word-medial consonant sequences during a familiarization, and rated test words with those same sequences according to the experimental frequencies. However, not all experimental frequencies resulted in a learning/extraction effect. Experimental frequency manipulations in the types only and types + tokens exposure groups significantly influenced participants' wordlikeness ratings, but experimental frequency in the tokens only exposure group did not. That is, participants rated test words as more Martian-like if those test words shared a sequence with three familiarization words. Thus, the artificial Martian type frequencies simulated findings from previous wordlikeness rating studies: Participants gave higher ratings to sequences with greater word-type support (Albright, 2009; Bailey & Hahn, 2001). Participants did not significantly vary their ratings based on hearing 1 versus 12 tokens of a word, however, and token frequency alone was unable to simulate a typical natural language frequency effect. The implication is that natural language may, in general, be more sensitive to type frequency than to token frequency.

Second, and simultaneous to the effect of type frequency, participants' ratings were influenced by English frequency, but in the opposite direction of what is typically reported. That is, words with low English frequency sequences were rated as more Martian-like than words with high English frequency sequences. Taken together, the findings are relevant to several research areas: theories of linguistic representations, online cognitive processes, and language development; all of which are discussed below.

Phonotactic Representations

How does evidence of type-based learning influence our understanding of the lexicon? I follow Richtsmeier et al. (in press) in arguing that the generalizations specific to the types only and types + tokens exposure groups require some form of abstraction, namely, the use or formation of phonotactic categories (Pierrehumbert, 2003). I propose that participants in this study made *abstract* generalizations because they exhibited extraction effects only in reference to an inherently abstract statistic, namely word-type frequency (Wetzel, 2009), and because participants were relatively unaffected by talker-tokens in isolation (the tokens only group). Furthermore, there were no differences between learning from types alone and types

combined with talker-tokens. Thus, the experimental frequency $\times$ exposure group interaction is best explained by recourse to abstraction.

I would also argue that models which eschew abstraction, such as exemplar models, have difficulty explaining the results. Exemplar models might model the present procedure with a distance algorithm that compares test words to previously stored, phonologically related words. This comparison is essentially priming, similar to phonological and repetition priming (Goldinger et al., 1992; Grant & Logan, 1993). Thus, the extraction effects observed here would be interpreted as priming effects.

With respect to phonological priming, phonologically similar words should prime each other (Goldinger et al., 1992), so, based on exemplar models, we would expect a familiarization word like /mæfpəm/ to prime a test word like /neɪfpən/. This expectation matches a description of the tokens only condition, however, where no priming was observed.

Furthermore, exemplar models do not always store exemplars based on abstract categories, such as phonemes (Goldinger, 1998; Kirchner & Moore, 2008; but cf. Bailey & Hahn, 2001), so it is unclear why word-types *would* lead to priming in an associative learning model. Thus, to explain the experimental frequency $\times$ exposure group interaction, exemplar learning models may require some degree of abstraction.

The present data do not rule out many appealing qualities of exemplar models, however. The priming and exemplar literatures provide strong evidence that highly detailed exemplars are stored in memory, and I do not debate that claim. Rather, I follow Pierrehumbert (2003) in arguing that exemplar storage must be supplemented by a system of abstract representations, such as those that could be provided by Bayesian learning.

Online Cognitive Processes

The directionality of English and experimental frequency effects in the present study is worth exploring in detail, particularly as it relates to cognitive representations of words and sublexical structure. Recall that participants rated test words as less Martian-like if they contained frequent English sequences like /kt/ and /st/, but they rated test words as more Martian-like if they contained infrequent English sequences like /pk/ and /fp/. It is a common result for participants to rate nonwords higher when they are composed of high native language frequency phonotactic sequences (Albright, 2009; Bailey & Hahn, 2001).

Why were participants in the present study reversing this trend? From a pragmatic perspective, the flip seems reasonable. Participants were asked to rate the test words according to their status as Martian words, not as English words, so it is sensible that they might rate the words that were least like English as being most like an alien language. Similarly, L2 research suggests that novice L2 learners often exhibit novel category biases (De Jong et al., 2008). By this logic, participants may have regarded infrequent sequences such as /fp/ and /ʃp/ as novel categories and, therefore, as better examples of Martian phonotactics. In the context of the significant effect of experimental frequency, however, this reversal becomes more intriguing.

Participants were inverting their knowledge of English phonotactics at the same time that they were giving higher ratings to test words with high type frequency support. That is, test words with the same sequences as high experimental frequency familiarization words in the types only and types + tokens exposure groups were rated as more Martian-like, regardless of whether they contained high or low English frequency sequences. Thus, 'high frequency' conditions led to both higher ratings (ExpHi) and lower ratings (HiEng) within the same dependent variable. How is such a result possible?

One possibility is that participants were performing two separate computations. One computation was to invert previously acquired knowledge of English phonotactics, the second computation tracked and extracted statistical information across the familiarization words. The claim of separate cognitive computations has some existing empirical support, including evidence that artificial language statistics may be calculated independently of a participant's native language (Magnuson, Tanenhaus, Aslin, & Dahan, 2003, but note that the present data provides evidence of the integration of artificial and real-language statistics). However, the timing of the English and experimental frequency computations is unclear. Was the English frequency reversal happening simultaneously to the experimental frequency calculation, or prior to it? If the reversal was simultaneous to the experimental frequency calculation, participants must have been performing separate but parallel computations. However, if the reversal was made prior to the experimental frequency calculations, perhaps as a result of pragmatic expectations established at the start of the experiment, then participants may have established the reversed English probabilities immediately and updated those probabilities based on the familiarization statistics. Certainly, future research is necessary to understand how and when participants incorporated English and experimental frequencies into their ratings. Additional behavioral work, such as a recognition memory task (Goldinger, 1996), can begin to address the time course these computations. The response time data hint that such research will be fruitful.

Phonetic Variability in Language Development

Talker-tokens were studied here because they embody phonetic variability and its facilitative effects. Phonetic variability has facilitated production accuracy in children (Richtsmeier et al., 2009, in press) and improved word monitoring in infants (Houston & Jusczyk, 2003; Singh, 2008). It was therefore surprising that phonetic variability did not appear to influence adult wordlikeness ratings. Based on a comparison of this study to the child and infant studies, the lack of a phonetic variability effect could be attributed to task differences. That is, infant and child studies may have used more sensitive methods for probing phonetic variability. The fact that English frequency effect sizes were much larger than experimental frequency effects is in keeping with this claim.

Nevertheless, the present study was similar to both the infant and child studies in its use of a familiarization-followed-by-testing paradigm. Additionally, the materials used here largely overlapped with the materials used in the child studies by Richtsmeier and colleagues. Both this study and those child studies focused on the same eight word-medial consonant sequences, and they involved virtually identical familiarization statistics, although two new words per sequence were added here to create two sets of words per sequence.

Another difference across the studies was the population studied—adults versus children. If the population difference is responsible for the different results, the studies may reflect a developmental change in sensitivity to phonetic variability. Given its consistent and facilitative role in learning studies with infants and children, I propose that phonetic variability has a diminishing role in phonological development, first serving to promote phonological category formation during infancy and childhood, but later having minimal relevance to phonological learning and instead acting more like noise. However, it should be noted that the issue of a developmental change in sensitivity to phonetic variability is separate from the issue of phonological representations. It may be that the change in sensitivity relates more to changes in attention than to formation of abstract phonological categories (cf. Rost & McMurray, 2010, for further discussion).

Future lines of research may improve our understanding of phonetic variability in phonological development. For example, future studies should compare age groups using a more consistent methodology, allowing age-specific-task differences to be ruled out. Similarly, using multiple methodologies to study adult learning from phonetic variability will provide a clearer picture of adult sensitivity to variability. Additionally, studies of second language (L2) learners will help us to understand phonetic variability's role regarding learning, rather than maturation. Existing L2 acquisition work suggests that phonetic variability does facilitate the acquisition of non-native phonetic contrasts (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Wang, Jongman, & Sereno, 2003).⁶ Thus, phonetic variability may be necessary for the initial formation of phonological categories in children but also relevant to new category formation in adult L2 learners.

In sum, it is surprising that the results of this study were not more similar to the infant and child studies, and that, here, phonotactic extraction seemed insensitive to talker variability. Because of several differences in the respective methodologies, it is unclear what factor or set of factors best account the discrepancy, but age differences provide one compelling explanation. Thus, a more comprehensive examination of type and token learning across age groups is warranted.

Conclusion

This study investigated the evidence adult English speakers use to learn, or “extract” phonotactic probabilities from an artificial lexicon. Adults appeared to rely on abstract word representations, or word-types, to learn the probabilities, but the results have important implications for both type-based and token-based learning. Ultimately, although participants were asked to learn “Martian”, the results reflect an agile and powerful human language faculty, able to apply previous knowledge to an unfamiliar task while tracking and learning new knowledge.

⁶Because the present study could also be regarded as an L2 learning study, the null effects of phonetic variability seem to contradict the work by Bradlow et al. (1997) and Wang et al. (2003). There are at least two explanations for this discrepancy. First, it is possible that the present study is not really an L2 learning study, and that participants were aware on some level that Martian is really English. Second, the studies by Bradlow et al. and Wang et al. had participants complete multiple familiarization sessions over multiple days, which can be contrasted with the brief familiarizations used here, and by Richtsmeier et al. (2009) and Richtsmeier et al. (in press).

Acknowledgments

These data were collected while the author was a Postdoctoral Researcher supported by funds from the National Institutes of Health to the University of Kansas through the Schiefelbusch Institute for Life Span Studies (R01 DC 006472). The author would like to thank Brad Kessell and Erin Williams for help with data collection, and Alex Francis, LouAnn Gerken, Lisa Goffman, and two anonymous reviewers for comments and suggestions. The author is currently affiliated with Purdue University.

Appendix A

Three steps were taken to balance the English frequency, experimental frequency, and exposure group factors, and to avoid confounds. First, consider the phones and biphones for the word /mæfpəm/ in Table 3.

Because /mæfpəm/ was used as an ExpHi word in the tokens only exposure group, it was important to ensure that its phonotactic properties were similar to the average properties of the three ExpHi words in the types only and types + tokens exposure groups: /mæfpəm/, /bæfpəm/, and /gɪfpək/. Thus, the phone and biphone frequencies in /mæfpəm/ would be, ideally, somewhere between the phone and biphone frequencies in /bæfpəm/ and /gɪfpək/. It was also important to insure that the properties of /mæfpəm/, /bæfpəm/, and /gɪfpək/ from the first word set were similar to the properties of /neɪfpən/, /pɔɪfpən/, and /ʃeɪfpək/ from the second word set. At the same time, it was important to make the words sound different, so a large set of consonants and vowels were used in word creation. These two demands were often at odds, but Table 4 shows that the phonotactic properties of the single words appearing in the tokens only exposure group (scores in columns labeled “1 WD”) were reasonably similar to the average phonotactic properties of the three words appearing in the types only and types + tokens exposure groups (scores in columns labeled “3 WDs”).

As a second step to balance factors, the same words were used for both familiarization and testing, with half of the participants hearing any given word during familiarization and the other words during testing. This ensured that the results reflected familiarization with the widest possible number of words. Third, four word lists per exposure group were created to counterbalance order of presentation, talker-to-word associations, picture-to-word associations, and picture-to-talker associations. Counterbalancing for /mæfpəm/ of experimental frequency, blocks, and talkers can be seen in greater detail in Table 5 below.

A few example words may be helpful. The words /mæfpəm/ and /neɪfpən/ were used across all three exposure groups, and both appeared in two familiarization lists per exposure group. In one tokens only list, the word /mæfpəm/ was ExpHi, appeared in the second familiarization block, and was paired with Picture 1 (pictures are not included in Table 5). In another list, /neɪfpən/ was ExpHi, appeared in the first familiarization block, and was paired with Picture 2. In two other lists, /mæfpəm/ and /neɪfpən/ were ExpLo, appeared in the opposite block (first block for /mæfpəm/), with a different talker producing the word, and a different associated picture (Pictures 3 and 4, respectively). In the types only and types + tokens groups, /mæfpəm/ and /neɪfpən/ were also distributed across four lists, balanced for experimental frequency, block order, and picture associations.

Table 4:

Phone and biphone scores for experimental words, broken up by word sets, consonant sequences, and exposure groups. Scores for sequences in the tokens only condition are in columns labeled “1 WD”. Scores for sequences in the types only and types + tokens conditions are in columns labeled “3 WDs”.

		Phone 1		Phone 2		Phone 3		Phone 4		Phone 5		Phone 6		Phone Sum		
		1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	
Set 1	/mp /	.024	.020	.039	.041	.049	.049	.036	.036	.082	.082	.101	.089	.331	.317	
	/mk /	.024	.020	.039	.041	.049	.049	.042	.042	.082	.082	.101	.089	.337	.323	
	/kt/	.003	.063	.029	.024	.054	.054	.089	.089	.082	.082	.055	.049	.312	.361	
	/pk/	.003	.063	.029	.024	.037	.037	.042	.042	.082	.082	.055	.049	.248	.297	
	/st/	.057	.045	.079	.070	.079	.079	.089	.089	.082	.082	.036	.038	.422	.403	
	/fp/	.057	.045	.079	.070	.020	.020	.036	.036	.082	.082	.036	.038	.310	.291	
	/sp/	.093	.063	.073	.052	.079	.079	.036	.036	.082	.082	.055	.061	.418	.372	
	/ʃp/	.093	.063	.073	.052	.008	.008	.036	.036	.082	.082	.055	.061	.347	.301	
Set 2	/mp /	.052	.082	.096	.060	.049	.049	.036	.036	.082	.082	.055	.057	.371	.366	
	/mk /	.052	.082	.096	.060	.049	.049	.042	.042	.082	.082	.055	.057	.377	.372	
	/kt/	.051	.043	.032	.042	.054	.054	.089	.089	.082	.082	.036	.044	.343	.354	
	/pk/	.051	.043	.032	.042	.037	.037	.042	.042	.082	.082	.036	.044	.280	.291	
	/st/	.024	.039	.029	.021	.079	.079	.089	.089	.082	.082	.083	.070	.386	.380	
	/fp/	.024	.039	.029	.021	.020	.020	.036	.036	.082	.082	.083	.070	.274	.268	
	/sp/	.003	.028	.061	.047	.079	.079	.036	.036	.082	.082	.036	.057	.295	.329	
	/ʃp/	.003	.028	.061	.047	.008	.008	.036	.036	.082	.082	.036	.057	.224	.258	
		Biphone 1		Biphone 2		Biphone 3		Biphone 4		Biphone 5		Biphone Sum				
		1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs			
Set 1	/mp /	.001	.001	.005	.004	.009	.009	.004	.004	.008	.014	.028	.032			
	/mk /	.001	.001	.005	.004	.000	.000	.002	.002	.008	.014	.017	.021			
	/kt/	.000	.001	.002	.002	.004	.004	.006	.006	.008	.009	.019	.021			
	/pk/	.000	.001	.002	.001	.000	.000	.002	.002	.008	.009	.012	.013			
	/st/	.010	.005	.007	.008	.023	.023	.006	.006	.012	.009	.058	.051			
	/fp/	.010	.005	.001	.002	.000	.000	.004	.004	.012	.009	.027	.019			
	/sp/	.001	.002	.007	.004	.008	.008	.004	.004	.008	.009	.028	.027			
	/ʃp/	.001	.002	.000	.000	.000	.000	.004	.004	.008	.009	.014	.015			
		/mp /	.015	.007	.007	.004	.009	.009	.004	.004	.008	.006	.044	.030		

	Phone 1		Phone 2		Phone 3		Phone 4		Phone 5		Phone 6		Phone Sum	
	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs	1 WD	3 WDs
/mk/	.015	.007	.007	.004	.000	.000	.002	.002	.008	.006	.032	.019		
/kt/	.002	.003	.002	.003	.004	.004	.006	.006	.012	.011	.025	.026		
Set 2 /pk/	.002	.003	.002	.002	.000	.000	.002	.002	.012	.011	.018	.018		
/st/	.002	.001	.003	.002	.023	.023	.006	.006	.017	.012	.051	.044		
/fp/	.002	.001	.000	.000	.000	.000	.004	.004	.017	.012	.023	.018		
/sp/	.000	.003	.002	.002	.008	.008	.004	.004	.012	.011	.027	.028		
/ʃp/	.000	.003	.000	.000	.000	.000	.004	.004	.012	.011	.016	.018		

Table 5:

The distribution of talker-tokens and word-types across the two blocks of familiarization, the English frequency and experimental frequency factors, the two stimuli lists, and the three exposure groups. Exposure groups are listed in the final three columns, with the blocks, English frequency, and experimental frequency grouped according to rows. Word sets are provided in different columns; the two familiarization lists that are not shown (1B and 2A) were created by flipping the order of the blocks and the assignment of experimental frequency to sequences. T1 through T12 represent the 12 different talkers present across the three experiments. Examination of the ‘Exper Freq’ and ‘English Freq’ columns will verify that these factors were fully crossed.

	Exper Freq	English Freq	Set 1 (List 1A)	Set 2 (List 2B)	Tokens Only Group	Types Only Group	Types + Tokens Group
Block 1	ExpHi	HiEng	nʌmpət	neɪstən	T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12	T1, T1, T1, T1	T1, T2, T3, T4
			gumpən	poɪstən	-	T2, T2, T2, T2	T5, T6, T7, T8
			ʃampən	ʃeɪstək	-	T3, T3, T3, T3	T1, T2, T3, T4
			zeɪktəs	zaspəm	T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12	T4, T4, T4, T4	T5, T6, T7, T8
	ExpLo	LoEng	saʊktəs	fospəm	-	T5, T5, T5, T5	T1, T2, T3, T4
			pəɪktəm	lɪspət	-	T6, T6, T6, T6	T5, T6, T7, T8
			mæfɹəm	dɪmkəs	T1, T1, T1, T1	T7, T7, T7, T7	T1, T2, T3, T4
			keɪpəs	bɪpkəm	T6, T6, T6, T6	T8, T8, T8, T8	T5, T6, T7, T8

^aThe ‘-’ indicates that a word was not used in the tokens only exposure group.

Table 6:

This is a continuation of Table 5.

	Exper Freq	English Freq	Set 1 (List 1A)	Set 2 (List 2B)	Tokens Only Group ^a	Types Only Group	Types + Tokens Group
Block 2	ExpHi	LoEng	nʌmkət	neɪfən	T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12	T5, T5, T5, T5	T5, T6, T7, T8

Exper Freq	English Freq	Set 1 (List 1A)	Set 2 (List 2B)	Tokens Only Group ^a	Types Only Group	Types + Tokens Group
		gumkən ʃamkən	porɪpən ʃerɪpək	- -	T6, T6, T6, T6 T7, T7, T7, T7	T1, T2, T3, T4 T5, T6, T7, T8
		zeɪpkəs	zaʃpəm	T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12	T3, T3, T3, T3	T1, T2, T3, T4
		saʊpkəs paɪpkəm	foʃpəm liʃpət	- -	T4, T4, T4, T4 T8, T8, T8, T8	T5, T6, T7, T8 T1, T2, T3, T4
ExpLo	HiEng	mæstəm kɛspəs	dɪmpəs bɪktəm	T9, T9, T9, T9 T3, T3, T3, T3	T2, T2, T2, T2 T1, T1, T1, T1	T5, T6, T7, T8 T1, T2, T3, T4

For the types only group, a unique talker was assigned to a word for each of the four lists. For the types + tokens group, where four talkers produced each word, a different four-talker set was assigned to the words in different lists.

Neighborhood density

Neighborhood density effects have been noted in studies of artificial lexicons (Magnuson et al., 2003), so it was important to ensure that the experimental words did not create unintended lexical neighborhoods. To make this possible, each word began with a unique CV sequence, and, to the extent possible, both the initial consonant and vowel were unique to a given word in a familiarization block. Although there was a significant amount of overlap between the ends of words, Magnuson et al. found that neighbor competition is greatest when words have overlapping onsets. Therefore, significant neighborhood effects among experimental words were not anticipated.

Additionally, Magnuson et al. found little influence of real English words on the neighborhood effects in their artificial lexicon. It was therefore not expected that the similarity of these words to real English neighbors would have much of an effect on the results. Regardless, similarity to English was minimized for all words. Neighborhoods based on a one-phoneme edit distance were calculated using the Washington University Neighborhood Database (<http://128.252.27.56/Neighborhood/Home.asp>). Edit distance is difficult to apply to multisyllabic words, but none of the words had English lexical neighbors within one phoneme, so all experimental words likely had comparably sparse neighborhoods.

References

- Albright A (2009). Feature-based generalization as a source of gradient acceptability. *Phonology*, 26(1), 9–41.
- Bailey TM, & Hahn U (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44, 568–591.
- Beckman ME, & Edwards J (1999). Lexical frequency effects on young children's imitative productions. In Broe MB & Pierrehumbert JB (Eds.), *Papers in Laboratory Phonology V: Acquisition and the lexicon* (p. 208–218). Cambridge, MA: Cambridge University Press.
- Bradlow AR, Pisoni DB, Akahane-Yamada R, & Tohkura Y (1997). Training Japanese listeners to identify English /r/ and /l/: Iv. some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, 101 (4), 2299–2310.

- Bybee J (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10, 225–255.
- Chambers KE, Onishi KH, & Fisher C (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, 87, B69–B77. [PubMed: 12590043]
- Cohen J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen J, MacWhinney B, Flatt M, & Provost J (1993). Psyscope: An interactive graphic system for defining and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, and Computers*, 25, 257–271.
- De Jong K, Hao Y-C, Park H, & Silbert N (2008). Novel-category biases in second language perception and production. *IULC Working Papers - Online*, 8. Retrieved February 23, 2010 from <https://www.indiana.edu/~iulcwp/pdfs/08-DeJong.pdf>.
- Dell GS, Reed KD, Adams DR, & Meyer AS (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 1355–1367.
- Edwards J, Beckman ME, & Munson B (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, 47, 421–436.
- Frisch SA, & Brea-Spahn MR (2010). Metalinguistic judgments of phonotactics by bilinguals. *Laboratory Phonology*, 1 (2), 345–360.
- Frisch SA, & Zawaydeh BA (2001). The psychological reality of OCP-Place in Arabic. *Language*, 77 (1), 91–106.
- Goldinger SD (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1166–1183.
- Goldinger SD (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 105 (2), 251–279. [PubMed: 9577239]
- Goldinger SD, Luce PA, Pisoni DB, & Marcario JK (1992). Form-based priming in spoken word recognition: The roles of competition and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18 (6), 1211–1238.
- Goldrick M (2004). Phonological features and phonotactics constraints in speech production. *Journal of Memory and Language*, 51, 586–603.
- Grant SC, & Logan GD (1993). The loss of repetition priming and automaticity over time as a function of degree of initial learning. *Memory and Cognition*, 21 (5), 611–618. [PubMed: 8412713]
- Greenberg JH, & Jenkins JJ (1964). Studies in the psychological correlates of the sound system of American English. *Word*, 20, 157–177.
- Hammond M (2004). Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies*, 4 (2), 1–24.
- Houston DM, & Jusczyk PW (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29 (6), 1143–1154. [PubMed: 14640835]
- Jusczyk PW, & Aslin RN (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29 (1), 1–23. [PubMed: 7641524]
- Jusczyk PW, Luce PA, & Charles-Luce J (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33, 630–645.
- Kager R, & Shatzman K (2007). Phonological constraints in speech processing. In Los B & van Koppen M (Eds.), *Linguistics in the Netherlands 2007* (p. 99–111). Retrieved May 1, 2009, from <http://www.let.uu.nl/~Rene.Kager/personal/publications.htm>.
- Kirby JP, & Yu ACL (2007). Lexical and phonotactic effects on wordlikeness judgments in Cantonese. In *Proceedings of the ICPhS* (p. 1389–1392). Retrieved May 1, 2009, from http://home.uchicago.edu/~aclyu/papers/icphs_gaps.pdf.
- Kirchner R, & Moore RK (2008). Computing phonological generalization over real speech exemplars. Retrieved January 15, 2010 from <http://www.ualberta.ca/~kirchner/kirchnermoore.pdf>.

- Magnuson JS, Tanenhaus MK, Aslin RN, & Dahan D (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, 132 (2), 202–227. [PubMed: 12825637]
- Munson B (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research*, 44, 778–792.
- Myers F, & Tsay J (2005). The processing of phonological acceptability judgments. In *Proceedings of the Symposium on 90-92 NSC Projects* (p. 26–45). Taipei, Taiwan.
- Nagata H (1988). The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *The Journal of Psycholinguistic Research*, 17 (1), 1–17.
- Nygaard LC (2008). Perceptual integration of linguistic and nonlinguistic properties of speech. In Pisoni DB & Remez RE (Eds.), *The handbook of speech perception* (p. 390–413). Malden, MA: Blackwell.
- Ohala DK (1999). The influence of sonority on children's cluster reductions. *Journal of Communication Disorders*, 32, 397–422. [PubMed: 10560714]
- Ohala JJ, & Ohala M (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In Ohala JJ & Jaeger JJ (Eds.), *Experimental phonology* (p. 239–252). Orlando, FL: Academic Press.
- Pierrehumbert JB (2003). Probabilistic phonology: Discrimination and robustness. In Bod R, Hay J, & Jannedy S (Eds.), *Probabilistic linguistics* (p. 177–228). Cambridge, MA: MIT Press.
- Raaijmakers JGW, Schrijnemakers JMC, & Gremmen F (1999). How to deal with “the language as fixed-effect fallacy”: Common misconceptions and alternative solutions. *Journal of Memory and Language*, 41 (3), 416–426.
- Richtsmeier PT, Gerken L, & Ohala DK (in press). Contributions of phonetic token variability and word-type frequency to phonological representations. *Journal of Child Language*.
- Richtsmeier PT, Gerken LA, Goffman L, & Hogan T (2009). Statistical frequency in perception affects children's lexical production. *Cognition*, 111, 372–377. [PubMed: 19338981]
- Rost GC, & McMurray B (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, 15 (6), 608–635.
- Schütze C (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.
- Seidl A, & Buckley E (2005). On the learning of arbitrary phonological rules. *Language Learning and Development*, 1 (3), 289–316.
- Singh L (2008). Influences of high and low variability on infant word recognition. *Cognition*, 106, 833–870. [PubMed: 17586482]
- Tenpenney PL (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review*, 2 (3), 339–363. [PubMed: 24203715]
- Vitevitch MS, & Luce PA (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, 36, 481–487. Retrieved May 1, 2009, from <http://www.people.ku.edu/~mvitevitch/PhonoProbHome.html>.
- Wang Y, Jongman A, & Sereno JA (2003). Acoustic and perceptual evaluation of mandarin tone production before and after perceptual training. *Journal of the Acoustical Society of America*, 113 (2), 1033–1043.
- Warker JA, Xu Y, Dell GS, & Fisher C (2009). Speech errors reflect the phonotactic constraints in recently spoken syllables, but not in recently heard syllables. *Cognition*, 112 (1), 81–96. [PubMed: 19398099]
- Wetzel L (2009). *Types and tokens: On abstract objects*. Cambridge, MA: MIT Press.
- Zamuner TS, Gerken L, & Hammond M (2004). Phonotactic probabilities in young children's speech production. *Journal of Child Language*, 31, 515–536. [PubMed: 15612388]

Learning /fp/ from tokens only

Experimental High	Experimental Low
<i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i>	<i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i>

Learning /fp/ from types only

Experimental High	Experimental Low
<i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i> <i>baɪfpəm baɪfpəm</i> <i>baɪfpəm baɪfpəm</i> <i>ɡɪfpək ɡɪfpək</i> <i>ɡɪfpək ɡɪfpək</i>	<i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i>

Learning /fp/ from types + tokens

Experimental High	Experimental Low
<i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i> <i>baɪfpəm baɪfpəm</i> <i>baɪfpəm baɪfpəm</i> <i>ɡɪfpək ɡɪfpək</i> <i>ɡɪfpək ɡɪfpək</i>	<i>mæfpəm mæfpəm</i> <i>mæfpəm mæfpəm</i>

Test
Rate "neɪfpən" "poɪfpən" "jeɪfpək"

Figure 1.

A visual representation of the three exposure groups in which the type and token properties of ExpHi sequences varied. The sequence /fp/ is used as an example, but the differences between exposure groups were consistent across all eight sequences. Note that a given participant would hear /fp/ as either ExpHi or ExpLo but not both. To represent talker differences, each talker is represented by a unique combination of font, shading, and size. Note that a unique talker produced the test words, however.

Figure 2.
Wordlikeness ratings, with separate graphs given for each of the three exposure groups.
Error bars reflect one standard error.

Figure 3.

Wordlikeness ratings by sequence. Separate graphs are given for each of the three exposure groups. Sequences are ordered according to English frequency, with HiEng sequences on the left side of each graph. Results for the ExpHi conditions correspond to dark gray bars; results for ExpLo conditions correspond to light gray bars. Note that the sequence /ʃp/ is labeled “shp”.

Table 1:

Experimental words, organized by set, English frequency, and exposure group (described below). Words sharing a consonant sequence are grouped by columns. Words used in the tokens only exposure group are presented in the top row of each column and in italicized font. For the the types only and types + tokens exposure groups, all three words within a column were used.

Set 1				
	/mp/	/kt/	/st/	/sp/
High	<i>nʌmpət</i>	<i>zeɪktəs</i>	<i>mæstəm</i>	<i>keʃpəs</i>
English	gumpən	saʊktəs	bəɪstəm	tuspən
	ʃampən	pɑɪktəm	ɡɪstək	daspək
	/mk/	/pk/	/fp/	/ʃp/
Low	<i>nʌmkət</i>	<i>zeɪpkəs</i>	<i>mæfpəm</i>	<i>keʃpəs</i>
English	gumkən	saʊpkəs	bəɪfpəm	tʃspən
	ʃamkən	pɑɪpkəm	ɡɪfpək	dəʃpək
Set 2				
	/mp/	/kt/	/st/	/sp/
High	<i>dʌmpəs</i>	<i>bɪktəm</i>	<i>neɪstən</i>	<i>zaspəm</i>
English	kɔɪmpət	lektəf	pɔɪstən	fɔspəm
	sæmpəf	tuktən	ʃeɪstək	lɪspət
	/mk/	/pk/	/fp/	/ʃp/
Low	<i>dʌmkəs</i>	<i>bɪpkəm</i>	<i>neɪfpən</i>	<i>zəʃpəm</i>
English	kɔɪmkət	lepəkəf	pɔɪfpən	fɔʃpəm
	sæmkəf	tupkən	ʃeɪfpək	hɪʃpət

Table 2:

Biphone probabilities computed by the web-based Phonotactic Probability Calculator (PPC) for the eight sequences used in the experiment.

sequence	HiEng CC				LoEng CC			
	/mp/	/kt/	/st/	/sp/	/mk/	/pk/	/fp/	/ʃp/
Biphone Probability	.0091	.0036	.0232	.0081	.0002	.0002	0.0	0.0

Table 3:

Individual phone and biphone scores, as well the the phone and biphone sums calculated by the Phonotactic Probability Calculator (PPC), for the familiarization word /mæfrəm/.

Phone 1	Phone 2	Phone 3	Phone 4	Phone 5	Phone 6	Phone Sum
m	æ	f	p	ə	m	mæfrəm
.0572	.0794	.0197	.0362	.0816	.0355	.3096
Biphone 1	Biphone 2	Biphone 3	Biphone 4	Biphone 5	Biphone Sum	
mæ	æf	fp	pə	əm	mæfrəm	
.0101	.0013	.0000	.0042	.0117	.0272	