Published in final edited form as:
Clin Linguist Phon. 2007 June ; 21(6): 423–433.

Comparability of Lexical Corpora: Word frequency in phonological generalization

JUDITH A. GIERUT and **RACHEL A. DALE**
Indiana University, Bloomington, IN, USA

Abstract

Statistical regularities in language have been examined for new insight to the language acquisition process. This line of study has aided theory advancement, but it also has raised methodological concerns about the applicability of corpora data to child populations. One issue is whether it is appropriate to extend the regularities observed in the speech of adults to developing linguistic systems. The purpose of this paper is to establish the comparability of lexical corpora in accounting for behavioural effects of word frequency on children's phonological generalization. Four word frequency corpora were evaluated in comparison of child/adult and written/spoken sources. These were applied post-hoc to generalization data previously reported for two preschool children. Results showed that the interpretation of phonological generalization was the same within and across children, regardless of the corpus being used. Phonological gains were more evident in low than high frequency words. The findings have implications for the design of probabilistic studies of language acquisition and clinical treatment programmes.

Keywords

Word frequency; lexical corpora; phonological generalization

Introduction

Recent studies of language acquisition have focused on the statistical regularities of words in the input as a possible contributor to the rate, course or strategy of children's language learning. Following from the adult literature, the general hypothesis is that the inherent structure of words in a language promotes behavioural effects that differentially serve to facilitate or inhibit the language learning process. Thus far, a number of statistical variables have been explored in childhood as influencing a range of linguistic domains. To illustrate, probabilistic phonotactics, or the likelihood of occurrence of certain sounds or sound sequences, have been reported as facilitating infants' perception of spoken words (Jusczyk, Luce, & Charles-Luce, 1994), toddlers' production of sounds that are typically later acquired (Zamuner, Gerken, & Hammond, 2004), and preschoolers' ease of learning new words (Storkel, 2001). This one variable alone minimally impacts lexical, phonological and morphological learning in perception and production, starting in infancy and continuing through the critical period of language learning. Statistical variables have likewise been shown as being relevant for children with linguistic disorders, including those with specific language impairments (Munson, Kurtz, & Windsor, 2005), reading disabilities (Metsala & Walley, 1998), and phonological disorders (Gierut, Morrisette, & Champion, 1999). Morrisette and Gierut (2002), for example, reported that a word's frequency and neighbourhood density impact the extent of generalization learning

Correspondence: Judith A. Gierut, Department of Speech and Hearing Sciences, 200 South Jordan Avenue, Indiana University, Bloomington, IN 47405-7002, USA. E-mail: gierut@indiana.edu.

for children with phonological disorders receiving clinical treatment, where word frequency is the number of occurrences of a given word in the language and neighbourhood density, the number of phonetically similar counterparts to a word based on one phoneme substitutions, deletions, or additions. The findings that are emerging are provocative, but they also present an interesting set of methodological questions with theoretical consequences.

One question is associated with how the statistical properties of words are established in the first place relative to how they are applied in studies of children. Typically, statistical regularities of language are documented in published or electronic corpora of words generated by adults. For instance, Kuèera and Francis (1967) compiled a comprehensive count of the frequency of words used in American English based on 15 genres of printed text (e.g. press, editorials, reviews, books) that appeared in first publications dated 1961. The database consists of 1,014,312 entries and is among the more widely used resources in statistical studies of language, albeit of the child or adult. At issue, however, is whether the statistical patterns deriving from an adult corpus (presumably representative of a fully developed lexicon) are compatible with the properties of a child's developing lexicon. There are at least two schools of thought. Some claim that children know fewer words and therefore, when adult corpora are applied to their data, the result is an overestimation of behavioural effects (Dollaghan, 1994). By this view, adult counts may inflate the effects of statistical regularities associated with a child's lexicon. Others argue that because children know fewer words, the patterns of a smaller lexicon are even more robust than would be the case in the adult system (Coady & Aslin, 2003). From this vantage, applications of adult corpora might underestimate the impact of statistical variables on children's language learning. To circumvent the issue, a recommendation is to simply use database counts of child utterances in lieu of adult corpora (Dollaghan, 1994). While this has an advantage of pairing child data with the apparent structure of a developing lexicon, its disadvantage is that it limits insights about developmental trajectories. If different corpora are applied in research involving infants versus children versus adults, it is not possible to determine whether particular statistical properties operate uniformly across the lifespan (Charles-Luce & Luce, 1995), or whether shifts in the salience of lexical variables occur at specific points in development (Storkel, 2002). Developmental portraits of this type can be obtained only by holding statistical corpora constant.

A second methodological question bears on the type of data that are typically entered into the construction of lexical corpora, with these being drawn from either printed or spoken sources. As a complement to Kuèera and Francis (1967) cited above, the corpus assembled by Brown (1984) reports the frequency of 190,000 words used by adults in conversation. When these sources are used in conjunction, they afford a dual perspective on the written versus spoken structure of word frequency, respectively, in the fully developed lexicon. Dollaghan (1994) and others (Charles-Luce & Luce, 1995; Coady & Aslin, 2003) have suggested that the differences between written and spoken corpora may afford insight to the receptive versus expressive structure of the lexicon. Presumably, written words approximate a comprehension lexicon as these are drawn from words that are stored in memory; spoken words approximate an expressive lexicon as these are extracted from output samples. If such a distinction is correct, then in studies of children, an issue is which structure is most pertinent. On the one hand, children's early language learning takes place in the receptive domain given that they cull linguistic regularities from the input that they hear. On the other hand, children may need the experienced frequency that comes with (practised) productive use of language in order to recognize and extract its patterned regularities. Indeed, when receptive versus expressive databases have been applied to children's data, they have yielded different interpretations (cf. Charles-Luce & Luce, 1990; 1995; Dollaghan, 1994; Coady & Aslin, 2003). One research need is an examination of a single set of language learning data, using multiple lexical corpora and applying the same methods and operational definitions throughout; this is a goal of the present study.

Specifically, to address concerns about the comparability of lexical corpora in language acquisition research, we asked whether a given set of child data would lead to a common developmental interpretation of results, independent of the database being used in analysis. Data were selected from *Clinical Linguistics & Phonetics*, as an extension of work by Morrisette (1999) who reported the longitudinal course of lexical diffusion for two children with phonological disorders. Lexical diffusion refers to the gradual infusion of sounds in a grammar, which takes place on a word-by-word basis. The frequency of the words that evidenced sound change for these children was of particular interest herein, due to its longstanding and robust effects as reported in the psycholinguistic (Howes, 1957; Landauer & Streeter, 1973), linguistic (Fidelholz, 1975; Hooper, 1976), and acquisition (Leonard & Ritterman, 1971; Cirrin, 1984) literatures. Word frequency corpora were also readily available, have substantial numbers of entries, and afforded comparisons of adult versus child and written (receptive) versus spoken (expressive) lexical counts. The primary intent of the current study was methodological in potentially establishing a “best match” of lexical corpora to children’s phonological generalization learning. It is necessary to begin with a brief recapitulation of Morrisette (1999) before turning to a description of the corpora and methods of analysis used in this study.

Morrisette (1999)

Two children diagnosed with phonological disorders were followed longitudinally during and following clinical treatment designed to improve their production of target sounds in words. Child 2 was 5;2 and had in error the target sounds /f v θ ð s z ʃ tʃ d_3 r/, whereas Child 4 was 3;11, with errors in targets /f θ ð s z ʃ tʃ d_3 l r/. Both children produced the errored sounds in words with 0% accuracy prior to treatment. In a single-subject experimental study, treatment was provided, serving as the independent variable and phonological generalization of sounds in error, the dependent variable (Gierut & Morrisette, 1996). Generalization was measured using the Phonological Knowledge Protocol (PKP; Gierut, 1985), which samples target English sounds in multiple word positions in a spontaneous picture-naming task. In the Morrisette study, each word of the PKP was labelled a priori as being either a high or low frequency item, following from a dictionary database consisting of 20,000 words coded for frequency based on Kuèera and Francis (Nusbaum, Pisoni, & Davis, 1984). (Neighbourhood density was also coded, but is set aside for purposes of the present study; cf. Gierut & Storkel, 2002; Morrisette & Gierut, 2002.) High frequency words were operationalized as having a count of 100 or more occurrences per million (Luce, 1986). Then, for each child, the PKP items that evidenced lexical diffusion were identified for each sound acquired. Specifically, these were the words that changed from incorrect to correct production and remained accurate over time; thus, accuracy and stability were both required for the data to be included in the analysis. The frequency of the words that did versus did not change were noted, and from this, a longitudinal trajectory of lexical diffusion was plotted for each target sound that was acquired. The dependent variable was the proportion of high frequency words that generalized relative to the full set of high frequency items, compared to the proportion of low frequency words that generalized relative to the full set of low frequency items. Thus, all relevant PKP items were evaluated to derive a percentage of generalization by frequency category. The goal was to determine whether lexical diffusion was predictable on the basis of a word’s frequency. Results showed differences between the two children in terms of number and order of sounds acquired, as well as the effects of word frequency on lexical diffusion. Morrisette attributed the individual differences to the complexity of the phonological features being acquired and to the salience of the lexical variable as construed by a given child. Still another possible source of the individual differences may be traced to the lexical corpus, the a priori coding scheme, and the interpretation that these rendered. These differences notwithstanding, it appeared that lower frequency words of the PKP were perhaps more amenable to lexical diffusion. This did not fully accord with prior reports, where high frequency words were shown to be facilitative in adult spoken word

recognition (Luce, 1986) and in phonological generalization learning (Leonard & Ritterman, 1971; Morrisette & Gierut, 2002), thereby further motivating the present study.

Methods

Corpora

Four lexical corpora of word frequency were used in the present study. These were chosen for their accessibility, scope of sample, and shared characteristics of interest. As the adult-receptive source, we relied on Kuèera and Francis (1967) as based on adults' written use of approximately 1 million words. For the parallel child-receptive source, Rinsland (1945) was used. This corpus is based on the frequency of words used in the written compositions of elementary school children. It consists of 25,632 unique words extracted from over 6 million running word strings; 14,571 of these were included in the corpus, having occurred 3 or more times in any grade level. On the expressive side, Brown's (1984) count of 190,000 words spoken by adults was consulted in parallel to Kolson's (1960) count obtained from kindergarteners. The child corpus consists of 590,000 spoken words sampled at school and in the home.

Data and analysis procedures

The sounds and words reported by Morrisette (1999) to have generalized for the two children were used in this post hoc comparison. With respect to sounds generalized, Child 2's production of /f v θ s z ʃ/ was evaluated, as was Child 4's use of /f s z/. For each of these sounds, the PKP words evidencing change and stability in accuracy over time were extracted. Child 2 evidenced gains in 74 PKP items, and Child 4, 30 items. Thus, 104 words were submitted to analysis; these were the same data and procedures reported by Morrisette.

There were three methodological departures. First, raw frequency values of the generalized words were identified independently from each of four databases, consistent with the comparative aim of this study. The raw frequencies were used to compute the mean frequency of generalized words for each sound acquired by each child for each lexical database. Second, the a priori coding of PKP words as either high or low frequency (based on 100 or more occurrences per million) was altered. This was necessary due to inherent differences in the corpora, both in the number of entries and in the frequencies reported. For the coding scheme used herein, grand means were computed for the generalized words, incorporating all sounds acquired. Grand means were specific to each corpus, and used to define a word as high or low frequency. If the raw frequency of a generalized word fell above the grand mean for a given database, then that word was coded as a high frequency item. Similarly, if a generalized word had a raw frequency below the grand mean for a given database, it was coded as a low frequency form. Grand means were also used in correlational analyses. A third methodological departure was the dependent variable. In this study, we reported the proportion of generalized words that were coded as high frequency relative to the grand mean versus those coded as low frequency relative to the grand mean of a corpus. Only words that generalized were included in the proportional analyses.

An example is needed to illustrate the methodological differences. We consider Child 4, who acquired the sound /z/, evidencing accuracy and stability in production. /z/ was the last sound to emerge in the longitudinal tracking of change. Moreover, /z/ was implemented in 5 of 11 PKP words: "zebra", "zipper", "raisin", "rose", and "buzz". These were the data reported by Morrisette (1999). New to the present study, the raw frequency values of the five /z/ words that generalized were independently extracted from each of the four lexical corpora. The raw frequencies were then compared word-by-word to the grand mean of each given corpus. Continuing the example, the raw frequency of the generalized word "zebra" was 1 occurrence per million, using the Kuèera and Francis database. "Zebra" was coded as a low frequency

word because its frequency (i.e. 1) was less than the grand mean of 118 for Kuèera and Francis. Likewise, the remaining generalized words “zipper”, “raisin”, “rose”, and “buzz” were coded as low frequency items, having the respective frequencies of 1, 1, 86, and 13 relative again to the grand mean of 118. Thus, of the five /z/ words that generalized, all were low frequency items under Kuèera and Francis. For Child 4, 100% of the /z/ words that generalized (i.e. 5 of 5) were low frequency items and 0%, high frequency items based on this particular corpus. The same procedures were followed in examining the generalized /z/ words using the remaining corpora.

The coding procedures were applied independently to the data from each child for sounds and words that generalized using each of the four lexical corpora. This resulted in four longitudinal trajectories of lexical diffusion per child. The trajectories reflected the proportion of high versus low frequency words that generalized by corpus, plotted in order of sounds learned by each child. Trajectories were used to describe the effects of word frequency on phonological generalization in order to determine the uniformity of interpretation across children and lexical corpora.

Results and discussion

Table I lists the mean frequency of generalized words for each sound acquired by each child. This is presented for each corpus, along with the corresponding grand means that resulted. Descriptively, it can be seen that the child counts of Rinsland (1945) and Kolson (1960) yielded consistently greater frequency values for generalized words than did the adult counts of Kuèera and Francis (1967) or Brown (1984). Similarly, receptive databases (albeit child or adult) tended to result in greater estimates of word frequency than expressive. Correlational analyses revealed that these descriptive observations were not exact. Specifically, Table II shows that the child-expressive database of Kolson was positively correlated with each of the other corpora, adult-receptive, adult-expressive, and child-receptive. Interestingly, the two adult databases were not significantly correlated. Perhaps less surprising, the adult-expressive corpus was not significantly correlated with the child-receptive.

In view of the correlational results, we predicted that the trajectories of generalization would also reflect differences in behavioural interpretation by corpus. Figures 1 and 2 (panels a–d) display the data from Child 2 and 4, respectively. The X-axes show the longitudinal order of sound emergence and y-axes, the proportion of generalized words coded as high versus low frequency. These data were used to describe the words that generalized and the strategy of generalization over time within and across children. As can be seen, the trajectories of learning for both children looked strikingly similar. Generalization occurred in proportionally more low than high frequency words. This was true across the different corpora, lending a uniform description of the effects of word frequency on generalization for these two children.

Despite the general similarity, there were some differences in the shape of the learning curves associated with particular corpora. For Child 2, the curves based on Rinsland (child receptive, Figure 1b) and Brown (adult expressive, Figure 1c) appeared distinct from the others. The departures were further associated with specific sounds: /s/ in the Rinsland curve and /v/ in the Brown curve. Likewise, for Child 4, the shape of the learning curve associated with the Rinsland corpus (Figure 2b) was somewhat different, most notably for the sound /f/. The reasons for these differences may be traceable to specific entries in the corpora. For instance, in the Rinsland corpus, the words “seven”, “ice”, and “fat” were all coded as high frequency items, with raw frequency values exceeding those cited in the comparison corpora. In fact, in the comparison corpora, the exact same words were coded as low frequency items. This is perhaps due to differential use in the elementary classroom, as the Rinsland corpus was based

on children's written work. Consequently, the proportion of high frequency words that generalized appeared to be greater for the Rinsland database in these isolated cases.

These item-by-item cases aside, the results of the behavioural comparison demonstrate that lexical corpora from adult/child and receptive/expressive sources are largely compatible. When taken together, the four databases yielded a consistent description of word frequency effects on phonological generalization for the two children. Thus, it was not necessary to use a child corpus for insights to children's phonological learning. Likewise, it was not necessary to appeal to an expressive corpus for insights about accuracy of production. Lee (2003) reached essentially the same conclusion in his correlational comparisons of adult databases. Kelly and Martin (1984) have argued on logical grounds that the sheer number of entries in a lexical corpus is bound to smooth the range of variation across databases. As sample sizes increases, variability decreases; hence, lexical corpora may be largely equivocal (Burgess & Livesay, 1998). Based on these observations then, the recommendation that emerges is two-pronged: any type of lexical corpus (child, adult, expressive/spoken, receptive/written) may be appropriate for use in the study of children's language acquisition, but only if it is comprised of a substantial number of entries. It might be further recommended that consideration be given to the existing literature and the corpora that have been utilized in prior research if continuity across the developmental period from infancy through adulthood is a goal. In this regard, Kuèera and Francis (1967) may be the most appropriate resource for examinations of word frequency effects. In future work, it will be necessary to repeat such demonstrations for other lexical factors that have been shown to shape children's language learning, such as neighbourhood density or phonotactic probability (Storkel & Morrisette, 2002). Word familiarity may need to be systematically explored, given that the generalization probe utilized in this study included forms rated as familiar but not unfamiliar (Gierut, in press). It is not yet clear how word familiarity facilitates or inhibits generalization learning in children with phonological disorders. It will also be necessary to expand the scope of study beyond examinations of generalization learning by children with phonological disorders. Perceptual word recognition, morphological and lexical acquisition may be worth evaluating in typical and atypical populations from comparative perspectives. The comparative bases themselves may need to be further explored to determine their true identity and role in shaping the structure of the lexicon. Written and spoken corpora were compared herein as potentially revealing of the receptive and expressive complements of lexical structure. However, corpora that are derived from caregivers' input to children (van de Weijer, 1998) or children's experienced use of words (Garlock, Walley, & Metsala, 2001) may provide alternative perspectives.

While the primary purpose of this research was methodological, the data bear on related theoretical and clinical issues that warrant mention. A first is the finding that lexical diffusion occurred in low frequency words for children with phonological disorders. In contrast, prior research has shown that greater phonological generalization occurred when high frequency words were used in treatment (Gierut et al., 1999; Morrisette & Gierut, 2002): a greater number of sounds is added to a child's inventory when high frequency words are presented as training items. This highlights an apparent dichotomy between the kinds of words that induce generalization and the kinds of words that undergo generalization for children with phonological disorders. The role of word frequency thus appears to vary by learning task. High frequency words are the triggers of generalization, but low frequency words are the targets (or recipients) of that change.

Why should low frequency words be susceptible to change? Some have suggested that low frequency items may be recent additions to a child's lexicon. As such, these words may have less robust phonological representations, rendering them more malleable (Walley, 1993). Another possibility that has been offered is that low frequency words are novel, perhaps on syllabic, segmental, prosodic or semantic grounds (Peters & Strömqvist, 1996). The novelty

of low frequency items may make them more noticeable, attracting a child's attention in the learning task.

There are nonetheless important clinical implications to be gained from the findings. Consistent with the literature, high frequency words may be optimal as treatment stimuli (Gierut, 2001); yet, probes may be most representative of generalization learning if they are balanced in structure. High and low frequency words should both be sampled because if not, the resulting data may give a false impression of a child's generalization learning. A probe consisting of mostly high frequency words may reveal a lack of generalization learning, thereby giving the sense that little progress has made and continued treatment, warranted. A probe of predominantly low frequency words may suggest the reverse, skewing generalization in favour of dismissal from treatment. Thus, in clinical applications, care should be taken in the selection of treatment and probe items based on word frequency corpora.

Continued research is needed to fully explicate the role of word frequency in phonological learning, particularly since there seem to be differential effects associated with the nature of the task. The clinical setting may be an ideal testing ground to establish the role of word frequency because of the varied tasks that are often used in treatment of children with phonological disorders. Through this line of investigation, it may be possible to uniquely differentiate word frequency effects in conditions of auditory bombardment, production practise with or without the use of phonemic contrasts, and metalinguistic tasks. In this way, new insights may be gained about the relevance of statistical variables to the language acquisition process, with applied clinical research serving as the segue to studies of typical development.

Acknowledgements

This research was supported in part by a grant from the National Institutes of Health DC001694 to Indiana University, Bloomington. We thank Michele Morrisette, Holly Storkel, and Kristin Selfridge for their input.

References

Brown GD. A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. Behavior Research Methods, Instruments, and Computers 1984;16:502–532.

Burgess C, Livesay K. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kuèera and Francis. Behavior Research Methods, Instruments and Computers 1998;30:272–277.

Charles-Luce J, Luce PA. Similarity neighbourhoods of words in young children's lexicons. Journal of Child Language 1990;17:205–215. [PubMed: 2312642]

Charles-Luce J, Luce PA. An examination of similarity neighbourhoods in young children's receptive vocabularies. Journal of Child Language 1995;22:727–735. [PubMed: 8789521]

Cirrin FM. Lexical search speed in children and adults. Journal of Experimental Child Psychology 1984;37:158–175.

Coady JA, Aslin RN. Phonological neighbourhoods in the developing lexicon. Journal of Child Language 2003;30:441–469. [PubMed: 12846305]

Dollaghan CA. Children's phonological neighbourhoods: Half empty or half full? Journal of Child Language 1994;21:257–272. [PubMed: 7929681]

Fidelholz J. Word frequency and vowel reduction in English. Chicago Linguistic Society 1975;11:200–213.

Garlock VM, Walley AC, Metsala JL. Age-of-acquisition, word frequency and neighborhood density effects on spoken word recognition by children and adults. Journal of Memory and Language 2001;45:468–492.

Gierut, JA. On the relationship between phonological knowledge and generalization learning in misarticulating children. Bloomington, IN: IULC; 1985.

Gierut JA. Complexity in phonological treatment: Clinical factors. Language, Speech and Hearing Services in Schools 2001;32:229–241.

Gierut, JA. Phonological disorders and the Developmental Phonology Archive. In: Dinnsen, DA.; Gierut, JA., editors. Optimality theory, phonological acquisition and disorders. London: Equinox; in press

Gierut JA, Morrisette ML. Triggering a principle of phonemic acquisition. Clinical Linguistics & Phonetics 1996;10:15–30.

Gierut JA, Storkel HL. Markedness and the grammar in lexical diffusion of fricatives. Clinical Linguistics & Phonetics 2002;16:115–134. [PubMed: 11987493]

Gierut JA, Morrisette ML, Champion AH. Lexical constraints in phonological acquisition. Journal of Child Language 1999;26:261–294. [PubMed: 11706466]

Hooper, J. Word frequency in lexical diffusion and the source of morphophonological change. In: Christie, WM., Jr, editor. Current progress in historical linguistics. Amsterdam: North-Holland; 1976. p. 95-105.

Howes DH. On the relation between the intelligibility and frequency of occurrence of English words. Journal of the Acoustical Society of America 1957;29:296–305.

Jusczyk PW, Luce PA, Charles-Luce J. Infants' sensitivity to phonotactic patterns in the native language. Journal of Memory and Language 1994;33:630–645.

Kelly, MH.; Martin, S. Domain-general abilities applied to domain-specific tasks: Sensitivities to probabilities in perception, cognition, and language. In: Gleitman, L.; Landau, B., editors. The acquisition of the lexicon. Cambridge, MA: MIT Press; 1994. p. 105-140.

Kolson, CJ. The vocabulary of kindergarten children. University of Pittsburgh; 1960. Unpublished doctoral dissertation

Kuèera, H.; Francis, WN. Computational analysis of present-day American English. Providence, RI: Brown University; 1967.

Landauer TK, Streeter LA. Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. Journal of Verbal Learning and Verbal Behavior 1973;12:119–131.

Lee CJ. Evidence-based selection of word frequency lists. Journal of Speech-Language Pathology and Audiology 2003;27:170–173.

Leonard LB, Ritterman SI. Articulation of /s/ as a function of cluster and word frequency of occurrence. Journal of Speech and Hearing Research 1971;14:476–485. [PubMed: 5163881]

Luce, PA. Neighborhoods of words in the mental lexicon Technical Report 6). Bloomington, IN: Speech Research Laboratory, Indiana University; 1986.

Metsala, JL.; Walley, AC. Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In: Metsala, JL.; Ehri, LC., editors. Word recognition in beginning literacy. Hillsdale, NJ: Erlbaum; 1998. p. 89-120.

Morrisette ML. Lexical characteristics of sound change. Clinical Linguistics & Phonetics 1999;13:219–238.

Morrisette ML, Gierut JA. Lexical organization and phonological change in treatment. Journal of Speech, Language, and Hearing Research 2002;45:143–159.

Munson B, Kurtz B, Windsor J. The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. Journal of Speech, Language, and Hearing Research 2005;48:1033–1047.

Nusbaum, HC.; Pisoni, DB.; Davis, CK. Speech Research Laboratory progress report 10. Bloomington, IN: Speech Research Laboratory, Indiana University; 1984. Sizing up the Hoosier mental lexicon; p. 357-376.

Peters, AM.; Strömqvist, S. The role of prosody in the acquisition of grammatical morphemes. In: Morgan, JL.; Demuth, K., editors. Signal to syntax: Bootstrapping from speech to grammar in early acquisition. Mahwah, NJ: Lawrence Erlbaum; 1996. p. 215-232.

Rinsland, HD. The basic vocabulary of elementary school children. New York: Macmillan; 1949.

Storkel HL. Learning new words: Phonotactic probability in language development. Journal of Speech, Language, and Hearing Research 2001;44:1321–1337.

Storkel HL. Restructuring of similarity neighborhoods in the developing mental lexicon. Journal of Child Language 2002;29:251–274. [PubMed: 12109371]

Storkel HL, Morrisette ML. The lexicon and phonology: Interactions in language acquisition. Language, Speech and Hearing Services in Schools 2002;33:24–37.

van de Weijer, J. Language input for word discovery. Wageningen: Ponson & Looijen; 1998.

Walley AC. The role of vocabulary development in children's spoken word recognition and segmentation ability. Developmental Review 1993;13:286–350.

Zamuner TS, Gerken L, Hammond M. Phonotactic probabilities in young children's speech production. Journal of Child Language 2004;31:515–536. [PubMed: 15612388]

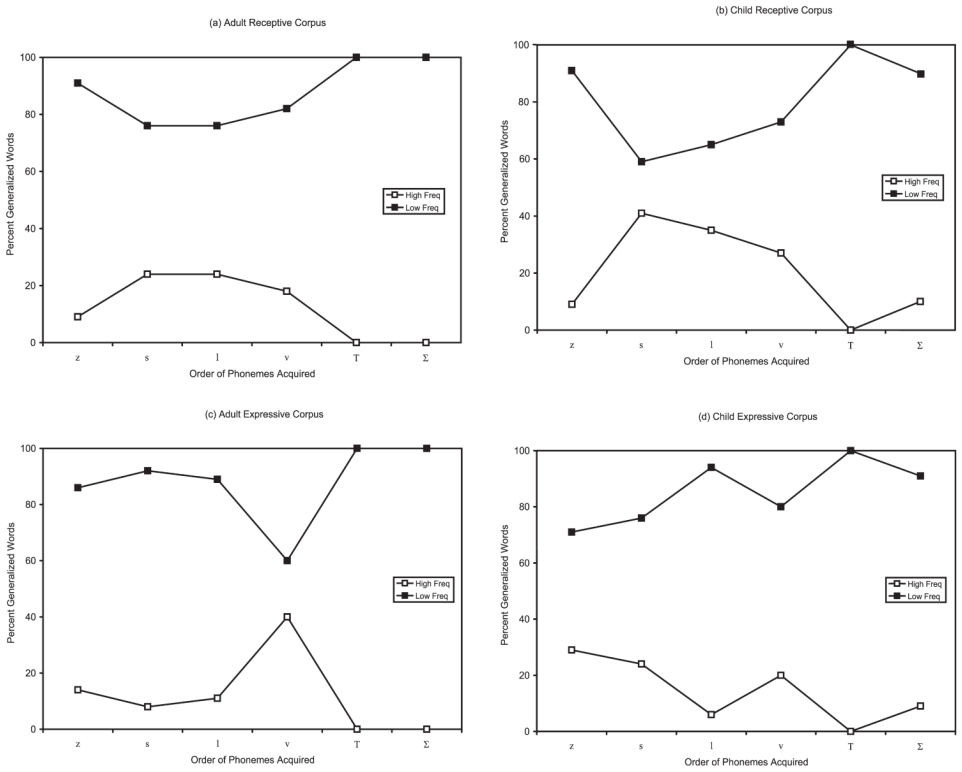

Figure 1.
(panels a–d). Trajectories of phonological generalization for Child 2 as derived from each of four lexical corpora.

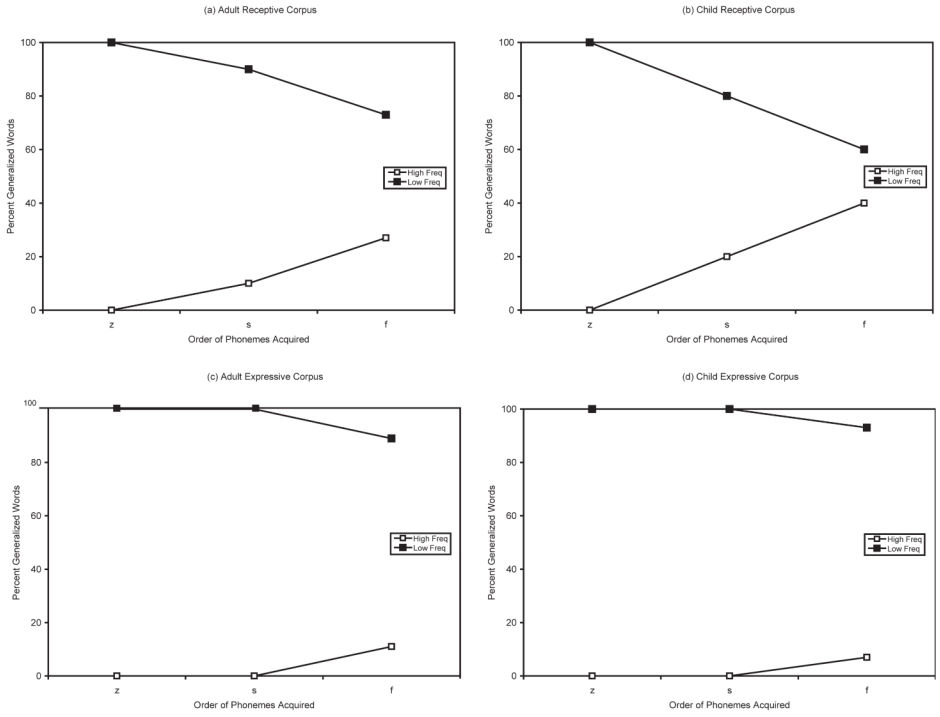

Figure 2.
(panels a–d). Trajectories of phonological generalization for Child 4 as derived from each of four lexical corpora.

Table I

Mean frequencies of generalized words for each child, using each of four lexical corpora. Frequencies are reported for sounds in the order in which they were acquired. For coding purposes, the grand means of the generalized words are also shown.

		Lexical Corpora			
		Receptive		Expressive	
Child	Sounds Acquired	Adult[a]	Child[b]	Adult[c]	Child[d]
2	z	164	554	41	282
	s	210	1546	236	464
	f	81	1196	23	82
	v	165	1603	70	334
	θ	40	393	4	61
	ʃ	21	535	5	94
4	z	20	99	2	11
	s	227	664	10	72
	f	92	1348	23	93
Grand Means		118	1067	93	231

[a] *Notes*: Kuèera and Francis (1967),

[b] Rinsland (1949),

[c] Brown (1984),

[d] Kolson (1960).

Table II

Pearson correlations (where p<.01, two-tailed) for the four lexical corpora.

	Adult-receptive[a]	**Adult-expressive**[b]	**Child-receptive**[c]	**Child-expressive**[d]
Adult-receptive[a]		.052	.472*	.383*
Adult-expressive[b]	.052		.111	.777*
Child-receptive[c]	.472*	.111		.639*
Child-expressive[d]	.383*	.777*	.639*	

[a] *Notes*: Kuèera and Francis (1967),

[b] Rinsland (1949),

[c] Brown (1984),

[d] Kolson (1960).