Phonological Neighborhood Density in the Trevor Corpus: Perception and Production Factors in Lexical Acquisition

Journal
UC Berkeley PhonLab Annual Report, 6(6)

ISSN
2768-5047

Author
Woodley, Melinda

Publication Date
2010

DOI
10.5070/P784n009j6

Phonological Neighborhood Density in the Trevor Corpus: Perception and Production Factors in Lexical Acquisition

Melinda Woodley
University of California, Berkeley

Since at least the 1980's, phonological neighborhood density (henceforth PND) has been recognized as an important factor in speech processing (e.g. Goldinger, Luce, & Pisoni, 1989). One of the reasons it has provoked such interest is that in contrast to many other factors relevant for speech processing (e.g. word frequency, imageability, age of acquisition, etc.) PND is known to have opposite effects on speech perception and production; words in dense neighborhoods are more difficult to perceive, but easier to produce than words in sparse neighborhoods (see e.g. Dell & Gordon, 2003 for a review).

More recently, phonological neighborhood density has been invoked as a possible driving factor in the development of segment-based lexical representations. Researchers in phonological development have argued that the lower overall density of young children's lexicons allows them to initially store and access words in a more holistic manner (Charles-Luce & Luce, 1990, 1995; Edwards, Beckman, & Munson, 2004; Logan, 1992; Vihman & Croft, 2007, among others). As words are added to the lexicon, however, the potential for confusion between similar sounding words increases, forcing language learners to pay attention to finer grained phonetic details in order to make the necessary contrasts between lexical items.

One potential shortcoming in most studies of developing phonological neighborhoods is that phonological development is considered in the aggregate; in many studies (e.g. Charles-Luce & Luce, 1990, 1995; Logan, 1992) the estimated time course for adding words to the lexicon is based on large corpora of many children. While this approach may gain something in its generalizability, it may also be missing some of the more interesting details that emerge from examining the development of individual children's phonological systems; since each child approaches the learning task based on his or her own experience with language, and his or her own inductive processes, cases studies of individual children's phonological development can shed light on the process by which early phonological generalizations are made and how they eventually develop into a full-fledged phonological system.

Following this logic, Yao (2009) traced the PND development of two individual children from the Manchester corpus of the CHILDES database (MacWhinney, 1991). Because PND is known to have opposite effects on perception versus production, Yao reasoned that if new lexical items tended to be added to sparse neighborhoods, this would indicate that ease of perception drives lexical acquisition, while if they tended to be added to dense neighborhoods, this would indicate ease of production as the driving force. Yao's study therefore aimed to provide an in-depth look at the development of PND over time for two children who had demonstrated different language (i.e. syntactic) learning strategies in a previous study (Ke & Yao, 2008).

Ultimately, Yao found no systematic effects of PND on the acquisition of lexical forms. However, the choice of the Manchester corpus may have been a limiting factor in this type of case study. For one, the amount of speech transcribed for each time period was low enough that many words seemed to disappear and reappear over time; in other words, the sample at each time period was not large enough to constitute an accurate representation of the child's knowledge at that point in time, rendering it difficult to make generalizations about the child's developing system.

Secondly, Yao's study assumed that children's phonological representations were equivalent to adult representations (in her case drawn from the CELEX corpus of British English), a reasonable assumption given previous work that showed little difference in PND depending on the size of units used for computation. Previous work notwithstanding, however, it seems reasonable that the child's mental lexicon would be organized based on his or her phonological generalizations at that point in time, and one possibility is that Yao's study found no systematic effects of PND because it assumed adult-like phonological representations. While the units used for computing PND may not make much difference in the aggregate, they may be relevant for individual children. Thus by computing PND based on the phonological generalizations that emerge and change as an individual child adds words to his or her lexicon, it may become clear whether PND has any systematic effects on lexical acquisition.

One way to build the child's own emerging phonological system into the computation of PND is to use the child's *own* pronunciations, rather than the adult norms, as lexical representations. For example, a child who develops regular consonant harmony appears to have made the generalization that there may be only one consonant place of articulation per word. While it could be argued that consonant harmony and/or other child phonological processes develop out of difficulty in production – and this is almost certainly a contributing factor – there is also reason to believe that such processes exist on a level of abstract phonological analysis as well. There are many documented cases of children who start out with some accurate (adult-like) pronunciations, and then as a given process takes hold, formerly accurate words are subsumed by the new pattern (see Becker & Tessier, 2010 for a computational study of the expansion of such patterns in the Trevor corpus). For the purposes of this paper, then, we will make the simplifying assumption that child pronunciations predominantly reveal something about the child's phonological system at that point in time, and the role of production difficulty will not be considered further.

The present question, then, is whether computing PND based on an individual child's pronunciation at a given point in time can shed any light on the role of PND in lexical acquisition. An excellent corpus for addressing this question is therefore the Trevor corpus (Compton & Streeter, 1977; Pater, 1997), since it contains a very large number of forms: several hours of speech per week phonetically transcribed by Trevor's mother, a speech pathologist, over a period of two years and four months, comprising a total of 13,668 utterances. Trevor also demonstrates several interesting phonological processes over the course of his development, including reduplication, complex onset simplification, and velar, alveolar, and labial consonant harmonies that are productive over slightly different time periods (Pater, 1997; Becker & Tessier, 2010).

In the first part of this paper, I will present several analyses of PND that build in Trevor's own phonological system by using his own pronunciations. We will consider several different ways of calculating PND, including both token- and type-based analyses, as well as segment- vs. syllable-based calculations. In the second part of the paper, I present a follow-up analysis using adult pronunciations of the words in the Trevor corpus in an attempt to disentangle the roles of production and perception in Trevor's PND development.

Corpus Study 1.1: Trevor's Pronunciations (Coded Manually)

Method

By way of exploring the amount of phonological variation present in the corpus, and the extent to which different types of analysis would affect the measures of PND, I first conducted a careful manual analysis of the portion of the Trevor corpus that covers Trevor's first birthday until the end of his 18th month of age – a total of 7 months' worth of data. Each month was taken to be a snapshot in time of Trevor's lexicon; for each month, I extracted all of the unique forms from that month and coded them for whether they were a new or old word, and whether they were a new or old form that month. For example, the word *book* makes its first appearance during month 16, and is transcribed /gʊ/ throughout the entire month. Thus for month 16, book would be would be entered as a new word with one new form, /gʊ/. In month 17, *book* is transcribed either /gʊ/ or /gʊk/, meaning that month 17 contains entries for *book* as an old word, /gʊ/ as an old form, and /gʊk/ as a new form.

In addition, since the original Trevor corpus is transcribed quite narrowly, for this and all subsequently presented studies, several simplifications were made in the phonological coding. The symbols /a/ and /ɔ/ were collapsed into a single phoneme category based on the fact that these phones are generally not considered contrastive in the variety of American English spoken on the west coast, where Trevor was born and raised. The sounds /ə/ and /ʌ/ were also collapsed into one category, partially because they seemed to be used inconsistently in the transcriptions, and also because the present analysis does not take stress into account. Affricates, diphthongs, and syllabic /r/ were considered single phonemes, rather than sequences of two phonemes. This is based on the intuitive feeling that, for example, *toy* /to$^{\text{i}}$/ and *boy* /bo$^{\text{i}}$/ sound more similar than *toy* /to$^{\text{i}}$/ and *toad* /tod/, and therefore the former pair should be considered phonological neighbors, while the latter pair should not.

For the segment-based analysis, phonological neighbors were considered to be forms that differed by a one phoneme substitution, addition, or deletion. The only difference in what I have called the "syllable-based" analysis presented below is that reduplicated syllables were considered to be neighbors of singleton syllables, but segment-based neighbors were still counted; that is, /ka.ka/ would be considered a neighbor of /ka/, /ka.kak/, and /ki.ka/.

For each month of analysis, the "active lexicon" was taken to be all of the new and old items used within the present month and the one preceding it. In the token-based analysis, this meant that for each month, all unique forms from that month and the one preceding it were included as potential neighbors in the lexicon. It should be noted that "unique forms" were subjectively judged to be those that differed in more than just phonetic implementation or transcription variation, e.g. multiple alternations between /kʊk/ and /kʌk/ during one month might be collapsed into a single form /kək/ for that month. This method of including all significantly different phonological forms can be seen as taking something like pattern frequency, or "variations on a theme" into account; by including every unique form Trevor produced, frequently reproduced patterns occur more often in the lexicon, and the contrast between dense and sparse phonological neighborhoods is slightly amplified.

For the type-based analysis, only one form of each word per month was included in the active lexicon. For words with multiple forms in one month, the form that most closely approximated the adult norm was chosen, under the assumption that it was the most accurate representation of what Trevor "knew" at that time; this generally entailed choosing a form with one more segment than the other forms, e.g. for the *book* example given above, choosing the form /gʊk/ as a better representation of Trevor's phonological system at that point in time than /gʊ/. When there were multiple forms that were equally complex and/or accurate, the one that showed evidence of persisting over time was chosen for that month. The type-based analysis was intended to further reduce the effects of noisiness in motor implementation; if Trevor did have one form in mind and he simply had trouble producing it, or it got transcribed in slightly different ways because its phones didn't precisely conform to adult category norms, the type-based analysis might provide a more accurate picture of Trevor's “true” phonological knowledge over time.

The main question at hand is whether new words (and/or forms) tended to be added to sparse or dense phonological neighborhoods. In order to address this question, we must have a way of evaluating “sparseness” vs. “denseness”. Therefore, for each month, I calculated the average number of neighbors (taken from that month's “active lexicon”, described above) for all of the new items introduced that month and compared it to the average number of neighbors for all of the items produced that month that had already been produced in a previous month (the “old” items). In other words, the potential neighbors for each month's analysis were *exactly the same in type and number* for the old vs. new items. In this way, it will be clear whether new items tend to have higher, lower, or the same PND relative to the rest of the lexicon.

Results of Corpus Study 1.1

Token-Based Analysis

Figure 1 below displays the results for the token-based analysis (all unique forms from the present and preceding month included in the active lexicon). The mean number of neighbors for old vs. new forms is plotted over time, with old forms in red and new forms in black. The segment-based analysis is plotted using solid lines, while the syllable-based analysis is plotted with dashed lines.

It is clear that new forms tend to have fewer neighbors than old forms; that is, if a new form is going to be attempted, it will be more than minimally different from other forms already in the lexicon. A two-tailed *t*-test reveals that new forms indeed have significantly fewer phonological neighbors on average than old forms (t(653) = -6.8, p < .001 for the segment-based analysis, t(665) = -8.1, p < .001 for the syllable-based analysis).

Type-Based Analysis

Figure 2 below displays the results for the type-based analysis (only one form per word per month included in the active lexicon). The mean number of neighbors for old vs. new words is again plotted over time, with old words in red and new words in black, and the segment-based and syllable-based analyses in solid vs. dashed lines, respectively.

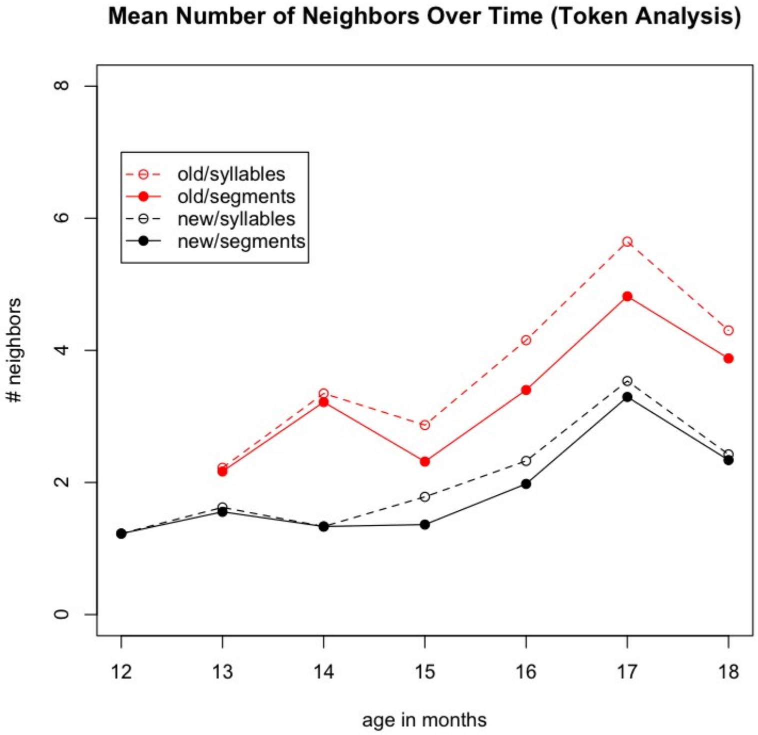

Figure 1: Phonological neighborhood density over time in the Trevor corpus, comparing new forms vs. old forms. See text for details.

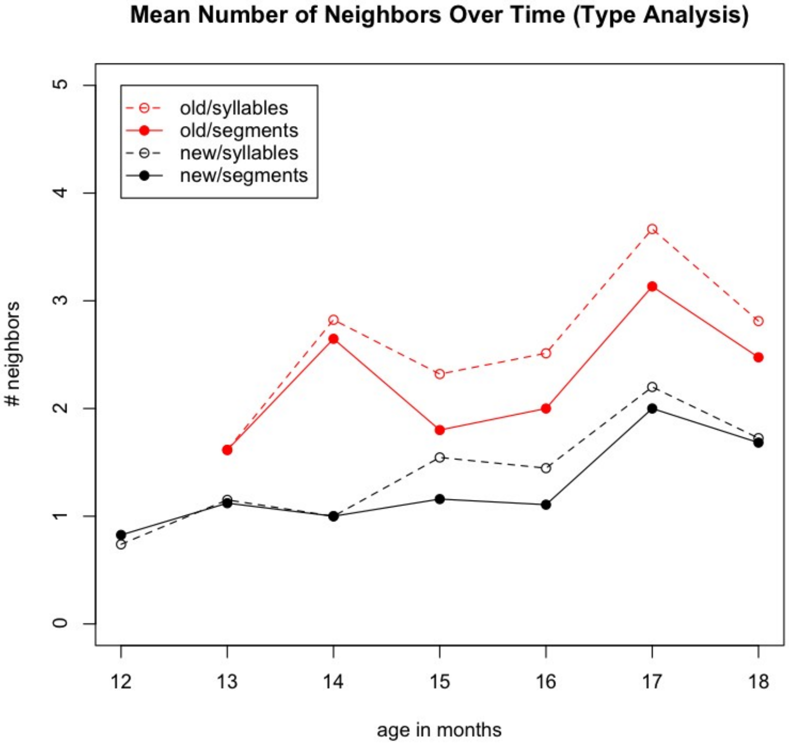

Figure 2: Phonological neighborhood density over time in the Trevor corpus, comparing new words vs. old words. See text for details.

There are no apparent qualitative differences between the token-based and type-based analyses; while words in the type-based analysis have fewer neighbors overall (due to the lower number of items included in the lexicon), new words still tend to have fewer neighbors than old words, and this difference is significant (t(445) = -4.9, p < .001 for the segment-based analysis, t(463) = -6.0, p < .001 for the syllable-based analysis).

Lexical Development Statistics

While the main purpose of this paper is to address the question of PND for new vs. old lexical items, the present analyses have also yielded some interesting statistics on the development of the lexicon over time. In the interest of completeness, I have summarized these findings in Table 3 below.

	12 mos.	**13 mos.**	**14 mos.**	**15 mos.**	**16 mos.**	**17 mos.**	**18 mos.**
# unique words (total vocab size)	24	46	39	70	95	183	225
# unique forms	31	63	53	107	157	306	313
# new words	24	32	10	35	34	86	78
# new forms	31	45	30	69	92	207	174
% new forms	100.0%	71.4%	56.6%	64.4%	58.5%	67.6%	55.5%
% new words (vocab growth)	100.0%	69.5%	25.6%	50.0%	35.7%	46.9%	34.6%
# unique forms/word (~ variability)	1.29	1.36	1.35	1.52	1.65	1.67	1.39

Table 3: Lexical development in the Trevor corpus, based on the manual coding done in Study 1.1.

It is interesting to note that while the total vocabulary size appears to grow exponentially near the end of the period of measurement, the proportion of new forms introduced each month seems to remain relatively stable over time (hovering around the 60% of all unique forms mark). This would seem to indicate that at least during this early stage of lexical development, even as more and more words are added to the lexicon, the amount of variability in forms (perhaps thought of as "phonological experimentation") seems to stay about the same. Of course, more data will be needed to see if this apparent trend is robust over time.

Discussion for Corpus Study 1.1

The results of Study 1.1 suggest that both novel words (type analysis) and forms (token analysis) tend to be added to relatively sparse neighborhoods of Trevor's productive lexicon. Additionally, a more syllable-based analysis that considered reduplicated syllables to be neighbors of singleton syllables produced the same result.

However, the data for Study 1.1 were culled and coded by hand. It is possible that the process of making somewhat subjective judgments about which forms should be considered new vs. old may have biased the “new” forms to be more unique, causing them to contain lower frequency segments and patterns and to have fewer neighbors in the lexicon. Furthermore, this method of culling data is time consuming and limits the amount of analysis that can be done.

In Study 1.2, therefore, the assignment of forms to the new vs. old category was done automatically by a computer script, such that forms were only considered old if they had the *exact* same transcription as in the previous month. While this may have unrealistically inflated the number of true phonological neighbors somewhat, it should have done so across the board, such that old and new forms would be affected to the same extent. Importantly, this approach carries with it the considerable benefit of allowing a larger amount of data to be analyzed.

Corpus Study 1.2: Trevor's Pronunciations (Coded Automatically)

Method

Because the results of Study 1.1 revealed no qualitative differences between the token and type analyses, or between the segment-based and syllable-based analyses, for the sake of simplicity Study 1.2 was carried out on all unique tokens in the corpus, and syllable reduplication was not taken into account. As mentioned above, for Study 1.2, “unique forms” were those whose transcriptions differed in any way; so while in Study 1.1, alternations between /kʊk/ and /kʌk/ over the course of one month might be collapsed into one form, in Study 1.2, these forms would be considered neighbors of one another. Automating the assignment of forms to the “new” vs. “old” categories allowed much more data to be analyzed; Study 1.2 considered all of Trevor's transcribed pronunciations from the age of 12 months through the age of 24 months.

As in Study 1.1, the active lexicon for each month was taken to be all of the forms that appeared in the corpus for either that month or the preceding month; thus the lexicon for age 20 months, for example, contained all of the forms transcribed during month 20 or month 19. The average number of neighbors was then calculated and compared for all of the forms that were new in a given month, versus those that appeared in a given month but had also already appeared in the previous month. Again, this means that for the analysis for any given month, the potential phonological neighbors were exactly the same for both the new and old forms.

Results for Corpus Study 1.2

Figure 4 below summarizes the results for Study 1.2. Once again, novel forms tended to be added to sparse neighborhoods of the lexicon, and the difference in number of neighbors between old and new forms is highly statistically significant (t(15760) = 64.4, p < .0001). The average number of neighbors for old forms (all months collapsed together) was 5.83, while the average number of neighbors for new forms was 2.02. Interestingly, in this analysis, the mean for new forms appears quite stable over time, while the mean for old forms appears to increase; this suggests that as the overall density of the lexicon increases over time (as expected), new forms tend to be added to neighborhoods that are increasingly sparse as compared to the rest of the lexicon.

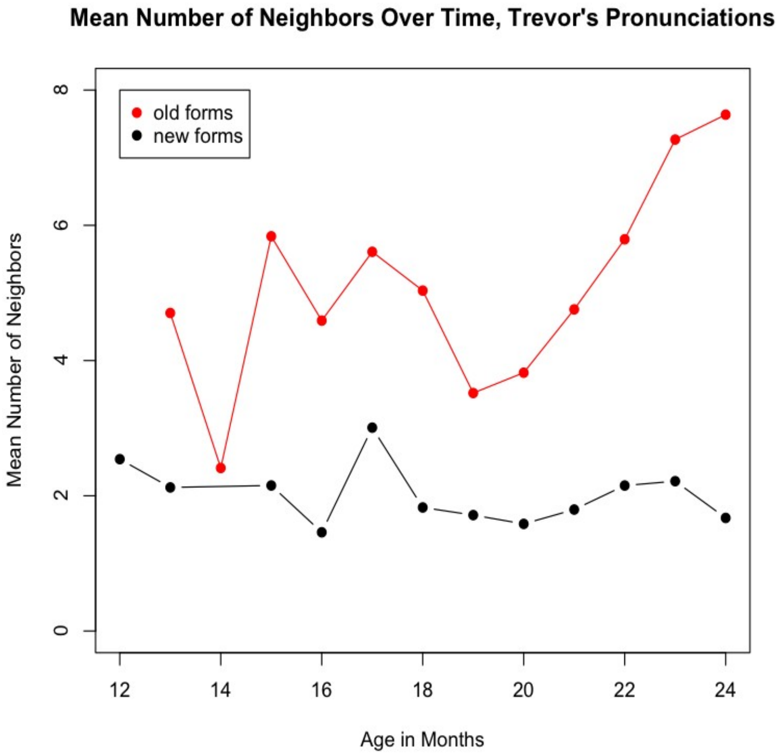

Figure 4: Phonological density over time in the Trevor corpus, coding done automatically by computer script. See text for details.

There appears to be a precipitous drop in the mean number of neighbors for old forms during month 14. This is because the number of forms transcribed during month 14 was considerably lower than in the months preceding and following it. In month 13, there were 81 new forms and 212 old forms in the corpus, for a total of 293 forms transcribed that month. In month 15, there were 289 new forms and 191 old forms, for a total of 480 forms transcribed. In month 14, by comparison, there were only 163 forms transcribed in total, none of which happened to be new. This is why there is no data point for new forms during month 14, and also why the mean for old forms in month 14 is so close to the mean for new forms in month 13; many of the forms that were old in month 14 *are* those that were new in month 13.

In Table 5 below, I have provided a month-by-month analysis of the unique phonological forms in Trevor's lexicon, including the total size of the "active lexicon" for each month, the number of new and old forms, the mean number of neighbors for new and old forms, and the results of a two sample, two-tailed t-test comparing these means. All of the t-tests included a Welch correction for unequal variances (this is why the degrees of freedom are lower than otherwise expected), and all were significant at well below the p < .01 level.

age in months	**total # unique forms**	**# new forms**	**mean # neighbors**	**# old forms**	**mean # neighbors**	**different means?**
12	120	120	2.54	0	--	--
13	293	81	2.12	212	4.7	t(251) = 7.0
14	163	0	--	163	2.41	--
15	480	289	2.15	191	5.83	t(236) = 10.3
16	678	150	1.46	528	4.59	t(605) = 13.5
17	1216	387	3	829	5.6	t(1112) = 11.6
18	1573	295	1.82	1278	5.03	t(997) = 19.0
19	1660	357	1.71	1303	3.51	t(862) = 14.1
20	1976	380	1.58	1596	3.82	t(999) = 17.3
21	2350	432	1.79	1918	4.75	t(1464) = 21.1
22	3086	508	2.15	2578	5.79	t(1529) = 22.3
23	4064	602	2.21	3462	7.26	t(1942) = 29.0
24	3766	406	1.67	3360	7.63	t(1912) = 34.8

Table 5: Summary of Trevor's unique forms over time, with t-tests comparing the mean number of neighbors for old vs. new forms. All t-tests used a Welch correction for unequal variances and were significant at well below the p < .01 level.

Discussion for Corpus Study 1.2

The results of Study 1.2 are consistent with Study 1.1, even though forms were assigned to the "new" vs. "old" categories manually in the first study and automatically by computer script in the second. In Study 1.2, as in Study 1.1, new forms tended to be added to phonological neighborhoods that were relatively sparse as compared to the rest of the lexicon, and the difference in mean PND for the new vs. old forms was found to be highly statistically significant, as confirmed by a series of t-tests. Additionally, it was found that mean PND remained relatively constant over time for the new forms, while mean PND increased for the old forms, especially at the end of the time period under investigation.

Overall Discussion for Corpus Study 1

At least for the portion of the Trevor corpus examined here, it seems clear that both new phonological forms (tokens) and new words (types) tended to be added to comparatively sparser areas of the lexicon. One interpretation of these results is that speech *perception* is driving the addition of items to the lexicon, since items in sparse neighborhoods are easier to perceive. Under this interpretation, Trevor tends to learn words that sound different from words he already knows.

A slight but significant twist on this interpretation would be that Trevor tends to produce new items that he can more easily *make distinct from* the other items already in his lexicon. Under this interpretation, speech *production* could be driving the addition of items to sparse neighborhoods, but it would not be processing facilitation in production, but rather avoidance of homonymy. Because many of Trevor's pronunciations are drastically simplified as compared to adult norms, it may be the case that Trevor is biased (either consciously or not) to attempt new forms that – given his own pronunciation patterns – will sound different from forms he already has in his repertoire.

An important simplifying assumption throughout this paper has been that Trevor's productions are an accurate and sufficient indicator of his phonological organization. In reality, it is possible (perhaps even likely) that there is some separation between Trevor's "personal" or "production" phonological system and his ability to perceive adult contrasts that are not present in his production system. In order to examine whether the "sparseness effect" observed here is more likely due to perceptual distinctiveness or avoidance of homonymy, Study 2 was conducted using adult pronunciation norms for the words in Trevor's production lexicon. If new words in Trevor's lexicon still tend to have fewer neighbors based on adult pronunciations, then the sparseness effect may be entirely due to processing facilitation in speech perception, or to a combination of perceptual distinctiveness and homonymy avoidance. If new words tend to have the same number or more neighbors based on adult pronunciations, then we can reasonably conclude that the addition of items to Trevor's production lexicon is being influenced by his attempt to maintain distinctiveness in his own pronunciations.

Corpus Study 2: Adult Pronunciations

Method

In Study 1, it was assumed that Trevor's production lexicon was organized solely according to the phonological system he exhibited in his own productions. However, as discussed above, it seems unlikely that Trevor had no stored knowledge about the adult forms of the items in his lexicon, even if he did not produce many of the relevant contrasts faithfully. In Study 2, therefore, I examined the mean PND for new vs. old items based on adult pronunciation norms drawn from the Carnegie Mellon University Pronunciation Dictionary (Weide, 1998), a corpus containing pronunciations for more than 125,000 words of North American English.

Because the CMU dictionary contains multiple pronunciation variants for many of the words, for this analysis I extracted only the first (most common) pronunciation for each of the words appearing in Trevor's lexicon from age 12 months through age 24 months. As in all previous analyses, I assumed that /a/ vs. /ɔ/ and /ʌ/ vs. /ə/ were non contrastive, and stress was not taken into account. Affricates, diphthongs, and syllabic /r/ were again considered single segments.

Contrary to previous analyses presented here, in Study 2 the active lexicon was taken to be only those words that were produced during the month under analysis. The rationale for including the present and immediately preceding month in each active lexicon in Study 1 was that children's pronunciations show great variation over time; a variant pronunciation may enter the lexicon and co-exists with several alternate forms for several weeks or months at a time before falling out of usage or overtaking the other variants, so for any given month it is difficult to know which, if any, of the active pronunciations should be considered primary. This variation and corresponding lack of certainty about which of the forms might be considered the most basic was handled by including the present and previous month's set of pronunciations in the active lexicon for each stage of analysis. For the adult forms, however, there would be no comparable change in forms over time. Additionally, the Trevor corpus is quite extensive, containing hundreds of transcriptions for each month. I therefore reasoned that if a word went an entire month without being transcribed at all, it would not be unreasonable to exclude it from that month's active lexicon for the purposes of this analysis. Thus for each month, I computed the mean PND for all words that were new that month and compared it to the mean PND for words that appeared in both the current month and the previous month.

Results for Corpus Study 2

The results of Study 2 are shown in Figure 6 below. As in Study 1, the mean PND for new forms was consistently less than the mean PND for old forms. A series of t-tests confirmed that many of the differences in means were statistically significant. The distribution of new vs. old forms, their mean PND, and the results of the t-tests are provided in Table 7.

Also as in Study 1, there is a large apparent drop in PND for one measurement period (at age 22 months). Again, this appears to be due to relatively fewer forms being transcribed that month, as there were fewer unique words transcribed during month 22 than either month 21 or month 23, and only 5 new words were recorded in month 22.

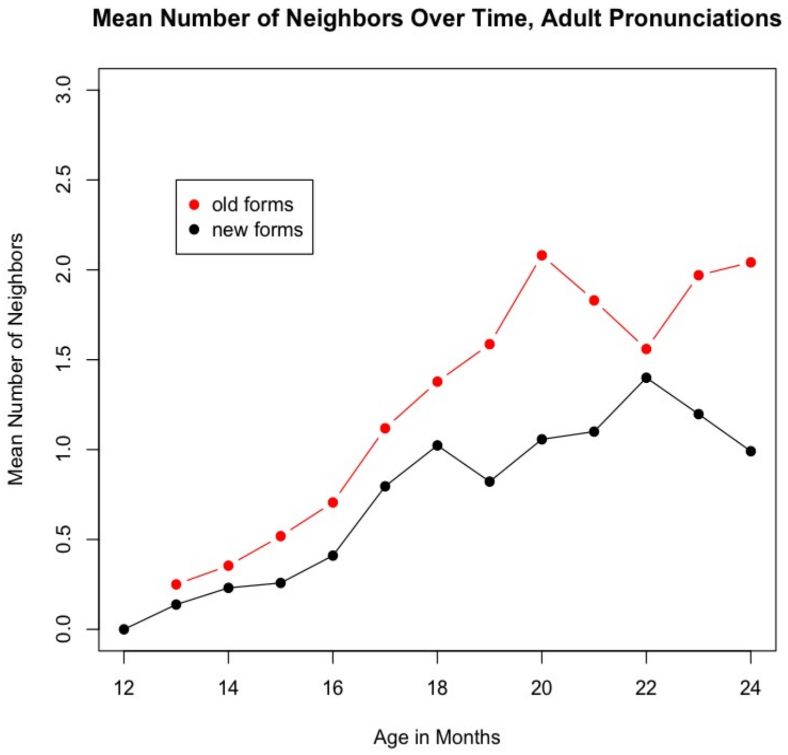

Figure 6: Comparing phonological neighborhood density for new vs. old words over time in the Trevor corpus, based on adult pronunciation norms.

While the differences in mean number of neighbors were quite small, many of the t-tests returned highly significant p-values due to the large amount of data being analyzed. Thus it seems safe to conclude that at least for the latter half of the period under study, Trevor tends to add new words to areas of his production lexicon that are relatively more sparse. That is, Trevor tends to attempt new words whose adult pronunciations are relatively more distinct from the other words in his production lexicon.

age in months	total # unique forms	# new forms	mean # neighbors	# old forms	mean # neighbors	different means?	p-value
12	23	23	0	0	--	--	--
13	53	29	0.13	24	0.25	t(41) = .75	n.s.
14	61	13	0.23	48	0.35	t(32) = .76	n.s.
15	85	31	0.25	54	0.51	t(82) = 1.6	0.09
16	124	39	0.41	85	0.7	t(105) = 1.6	0.1
17	211	93	0.79	118	1.11	t(207) = 1.7	0.07
18	295	86	1.02	209	1.37	t(230) = 1.7	0.07
19	339	73	0.82	266	1.58	t(235) = 4.23	< .0001
20	526	192	1.05	334	2.08	t(522) = 6.25	< .0001
21	622	110	1.1	512	1.83	t(214) = 4.3	< .0001
22	566	5	1.4	561	1.55	t(4) = .17	n.s.
23	693	157	1.19	536	1.97	t(363) = 4.7	< .0001
24	734	111	0.99	623	2.04	t(212) = 5.8	< .0001

Table 7: Summary of Trevor's PND over time, using adult pronunciation norms as input. All t-tests used a Welch correction for unequal variances. Light gray shading indicates significance at the p < .10 level. Darker gray shading indicates significance at the p < .01 level.

Discussion for Study 2

Both Study 1 and Study 2 indicate that Trevor tends to add new words to relatively sparser areas of his production lexicon. Because PND in Study 1 was calculated based on Trevor's own pronunciations, this means that he attempted new forms that sounded relatively more distinct given the pronunciation constraints imposed by his own developing phonological system and/or his difficulties in motor command implementation. Moreover, because the follow-up analysis in Study 2 computed PND based on adult pronunciations, we also know that the new words Trevor attempted tended to sound relatively distinct (in the input to his perceptual system) from the other words in his lexicon.

Given these findings, it is difficult to say whether distinctiveness in perception or production may be playing a more important role in the addition of words to Trevor's production lexicon. However, what does seem clear is that *distinctiveness* is the driving factor; whether Trevor tends to add words that he can more easily tell apart from the other words he already knows in perception, or he adds words that he can more easily make distinct from other words in production, there is certainly no evidence that processing facilitation in production is relevant to Trevor's lexical acquisition. Since previous work has found that words in sparse neighborhoods are more difficult to produce, we can conclude that perceptual facilitation and/or homonymy avoidance may be at play in the expansion of Trevor's production vocabulary.

Overall Discussion

While the present results cannot speak to the relative importance of perception versus production in lexical acquisition, they do suggest that processing facilitation in production does not drive the addition of new items to the lexicon, at least in Trevor's case. The fact that words tended to be added to sparser areas of the lexicon based on adult pronunciations also indicates that competition between lexical items in perception may be important in lexical development; children may be more likely to remember and reproduce words that are more perceptually distinct from other words in their lexicon. Indeed, this has recently been suggested by real time word-learning experiments. Hoover, Storkel, and Hogan (2010) provided evidence that 3- to 5-year-old children were more likely to remember nonwords with lower PND and phonotactic probability.

One obvious potential drawback to the present study is that it has considered PND in complete isolation from other factors that influence lexical acquisition. While significant differences were found between new and old forms in Trevor's developing lexicon, it is simply not the case that PND is *the* driving factor in word learning. Future studies should consider other relevant factors, such as word and segment frequency, phonotactic probability, concept imageability, and others in conjunction with PND to determine the extent to which PND may play a role in the addition of new items to children's lexicons.

References

Becker & Tessier (2010). “Trajectories in faithfulness in child-specific phonology.” Unpublished manuscript, accessed October 10, 2010 from http://becker.phonologist.org/papers.shtml.

Charles-Luce & Luce (1990). “Similarity neighbourhoods of words in young children's lexicons.” *Journal of Child Language* 17, p. 205 – 215.

Charles-Luce & Luce (1995). “An examination of similarity neighbourhoods in young children's receptive vocabularies.” *Journal of Child Language* 22, p. 727 – 735.

Compton & Streeter (1977). “Child Phonology: Data Collection and Preliminary Analyses.” In *Papers and Reports on Child Language Development* 7, Stanford University, Stanford, California.

Dell & Gordon (2003). “Neighbors in the lexicon: Friends or foes?” In *Phonetics and phonology in language comprehension and production: Differences and similarities.* Ed. Niels O. Schiller and Antje S. Meyer.

Edwards, Beckman, & Munson (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(2), p. 421-436.

Goldinger, Luce, & Pisoni (1989). “Priming lexical neighbors of spoken words: Effects of competition and inhibition.” *Journal of Memory and Language* 28(5), p. 501 – 518.

Hoover, Storkel, & Hogan (2010). “A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children.” *Journal of Memory and Language* 63, p. 110 – 116.

Ke & Yao (2008). “Analyzing language development from a network approach.” *Journal of Quantitative Linguistics* 15(1), p. 70 – 99.

MacWhinney, B. (1991). The Childes Project: Tools for Analyzing Talk. Lawrence Erlbaum Associates.

Logan, J. (1992). “A computational analysis of young children's lexicons. Ph.D. dissertation, Indiana University.

Pater, J. (1997). “Minimal Violation and Phonological Development.” *Language Acquisition* 6(3), p. 201 – 253.

Vihman & Croft (2007). “Phonological development: toward a 'radical' templatic phonology.” *Linguistics* 45(4), p. 683 – 725.

Weide, R. (1998). The CMU pronunciation dictionary (Release 0.6). Carnegie Mellon University. Available online at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Yao, Y. (2009). “Phonological Neighbourhood Development in Children's Lexicons.” *UC Berkeley Phonology Lab Annual Report,* p. 44 – 63.