Neighborhood density and word frequency in child German

Aleksandra Zaba and Thomas Schmidt
Research Center on Multilingualism at the University of Hamburg

Words from dense neighborhoods in the mental lexicon, such as *cat* (with many phonological neighbors, that is, phonologically similar words, e.g., *mat*, *at*, *cab*, *rat*, *pat*) are produced more accurately and quickly by adult native speakers of English than words from sparse neighborhoods, such as *wolf* (with fewer neighbors, e.g., *woof*, *wooly*, *wool*). High frequency of words, such as in *dog* (frequent) vs. *molecule* (less frequent) also contributes positively to the accuracy and speed of word production. This facilitative effect of high density and frequency was demonstrated, for example, by Vitevitch & Sommers (2003) in their study on speech production in English adults. Contrasting results were provided by Vitevitch & Stamer (2006), who found that high density neighborhoods inhibited lexical production in a picture naming task in Spanish adults. This outcome was attributed to morphological differences between Spanish and English. Specifically, Spanish has richer morphology than English, and its phonological neighborhoods may contain more morphologically and semantically similar neighbors than does English (e.g., *niño* 'male child' – *niña* 'female child'), which are likely to introduce competition during word retrieval and, therefore, may inhibit production.

Studies that focused on the influence of lexical variables on speech production in children established that the role of lexical variables is largely identical to the one played in adult speech (e.g., Stemberger 1984; Gierut 2001; Storkel & Morrisette 2002; German & Newman 2004; Storkel 2004). For example, based on data from two English databases, Storkel (2004) showed that early words were more frequent and from denser neighborhoods (in an adult lexicon) than later words, suggesting that these two lexical factors are of facilitative influence in the early production of English.

The present study aimed at contributing to an understanding of the influence of density and frequency on speech production by looking at data from a new group of subjects, namely German children. Similarly to Spanish, German has richer morphology than English, which may cause words from denser neighborhoods to be produced relatively late. Alternatively, as child speech tends to contain simplified morphology, problems associated with rich morphology as assumed for adult speech may not hold in the case of children. Research question 1 was correspondingly formulated in two ways (see (1)).

(1) a. Do German children produce words from dense neighborhoods later than words from sparse neighborhoods?
b. Do German children produce words from dense neighborhoods earlier than words from sparse neighborhoods?

Further, we investigated the influence of word frequency on production in this group of children. Assuming that frequency is not subject to the effect of rich morphology (there is no corresponding literature), the influence of frequency was hypothesized to be similar to the one found in English. That is, early produced words

were expected to be of high frequency. The second research question was therefore as in (2).

(2) Do German children produce more frequent words before less frequent ones?

Neighborhood density was defined in the present study as the number of neighbors possessed by a word produced by a child, as determined by reference to a phonetic dictionary. Neighbors were words that differed based on one phoneme, either by addition, deletion, or substitution (Luce & Pisoni 1998). For example, some neighbors of *cat* are *mat, at, cab, rat,* and *pat,* while the neighbors of *wolf* are *woof* and *wooly. Cat* has higher neighborhood density than *wolf* as the former has more neighbors. Word frequency simply referred to the number of occurrences of a given word in a given language, as determined based on a frequency dictionary.

Our participants were four monolingual German children from the corpus of project PAIDUS at the Research Center on Multilingualism at the University of Hamburg in Germany. Their aliases were Bernd, Johannes, Marion, and Thomas. The ages in focus were 2;00 (phase 1), 2;05-2;06 (phase 2), and ca. 3;00 (phase 3). The children had been audio- and video-recorded at their homes in play situations while interacting with one of the parents and a researcher, for approximately 30 minutes per recording. After the recordings, data were transcribed for the analysis. We only used 1- and 2-syllable words, from various word classes (N=1203). The list of words analyzed with respect to both density and frequency was the same.

The calculation of density in the present study was based on the Levenshtein distance (Levenshtein 1966), which is the minimum number of insertions, deletions, or substitutions of single characters necessary to transform string A (in our case, a word) into string B (another word). In our case, characters represented phones in a phonetic dictionary of German. To calculate neighbors, we first extracted a list of orthographic tokens from our selection of transcripts for a given child. We then normalized the tokens in that list by mapping inflected forms (such as *komm* 'come!' for *kommen* 'to come') onto their base forms (based on a suggestion in Anderson (2007)). For each item in this list of normalized orthographic forms, we determined its phonetic form. We then compared this phonetic target form to each of its neighbor candidates, i.e., to each phonetic form in a phonetic dictionary of German, and calculated the Levenshtein distance between each pair of forms (target and neighbor candidate). Since the Levenshtein distance calculation operates on individual characters, but some phonological entities are represented by a combination of two or more characters (such as /øː/), which we wanted to treat as a single entity in the Levenshtein calculation, we replaced such character combinations with a single symbol before performing the comparison. If the Levenshtein distance between a target and a given neighbor candidate was 1, we put the corresponding candidate into the list of neighbors for the target word. The process was implemented in a Java program. We used the list of neighbors for statistical analyses of mean neighborhood differences among the phases. No density lexica exist for German, neither for adult nor child speech. Also, to our best knowledge, no dictionaries of German child language have been compiled yet. Therefore, we conducted our own density calculations, based on a dictionary of adult speech, the 141,490 word phonetic dictionary of German by Portele, Krämer & Stock (1995).

The frequency of these same (normalized) words was taken to be as in the Frequency Dictionary of German (Jones & Tschirner 2006), which is based on 4.2 million words of adult spoken and written German, and ranks words on a scale from 1 to 4034 (based on their number of occurrences per one million words). For each item in the list of normalized orthographic forms, we looked up in the dictionary its relative frequency rank (from 1 to 4034), and added the number to the list. This list of relative frequency ranks was used for the statistical analyses, shown next.

Table 1 presents the results for neighborhood density for each of the 4 children.

	Bernd	**Johannes**	**Marion**	**Thomas**
Phase 1 (2;00)	17.29	14.78	15.64	15.39
Phase 2 (2;05-2;06)	15.72	16.52	16.71	16.25
Phase 3 (ca. 3;00)	N/A	16.29	15.41	19.55

Table 1: Average neighborhood densities for each German child by age phase

Based on nonparametric tests, the only instances in which there were statistically significant differences between the phases were Johannes phase 1 (14.78) vs. 2 (16.52), p=.097 (marginal), and Thomas phases 1 (15.39) vs. 3 (19.55), p<.001, and 2 (16.25) vs. 3 (19.55), p<.001, respectively. In each case, the difference involved lower density at the earlier phase and higher density at the later phase. Based on this result, the response to research question 1 (in (1)) seems to be that German children produce words from dense neighborhoods later than words from sparse (or less dense) neighborhoods. However, this is a tentative result since the difference in relative density between the phases was significant only in the case of a few children and phases.

Table 2 provides a summary of the frequency results for each child.

	Bernd	**Johannes**	**Marion**	**Thomas**
Phase 1 (2;00)	828	806	828	939
Phase 2 (2;05-2;06)	626	638	396	703
Phase 3 (ca. 3;00)	N/A	567	452	204

Table 2: Average word frequencies for each German child by age phase, based on scale from 1-4034 (Jones & Tschirner 2006)

As nonparametric tests reveal, significant differences between the phases, with higher frequency at the earlier than at the later phases, were detectable in most cases: Bernd phase 1 (828) vs. 2 (626), p<.001, Johannes 1 (806) vs. 2 (638), p=.074 (marginal), and 1 (806) vs. 3 (567), p=.010, Marion 1 (828) vs. 2 (396), p=.014, and 2 (396) vs. 3 (452), p=.083 (marginal), and Thomas 1 (939) vs. 3 (204), p=.001. The few instances in which the differences were not significant were: Johannes phase 2 (638) vs. 3 (567) , Marion 1 (828) vs. 3 (452), and Thomas 1 (939) vs. 2 (703), and 2 (703) vs. 3 (204). Thus, the response to our research question concerning word frequency (in (2)) is that in general, German children produce more frequent words before less frequent ones.

Based on the present results, one can make the cautious claim that the effect of neighborhood density on speech production depends on the richness of morphology of a given language. In this case, the relatively rich morphology of German may contribute to an inhibition of production in high density words, similarly to Spanish. Further,

frequency effects seem to be independent of morphology as high frequency words were produced early on.

To exclude the possibility that the present results are a reflection of the particular characteristics of the lexicon used in the present study, it would be interesting to calculate density and frequency of the words also based on different sources, for example, as a proportion of the transcripts/lexica that the words were extracted from (i.e., each particular child lexicon), and/or use more adult lexica as a base. Also, we are planning further analyses: More phases for each child will be looked at (ages in between the phases already examined, as well as beyond that), and words will be divided into different word classes for the analysis, as density and frequency may have various impact on various word classes. We will also look for possible interactions between frequency and density, as previous research (e.g., Storkel 2004) found different effects of density on low than on high frequency words (in English adults). Finally, we are planning to address more variables, such as neighborhood frequency (which refers to the frequency of a target word's neighbors), as these also play a role in the development of speech production.

References

Anderson, J.D. 2007. Phonological neighborhood and word frequency effects in the stuttered disfluencies of children who stutter. *Journal of Speech, Language and Hearing Research* 50. 229-247.

German, D.J. & R.S. Newman. 2004. The impact of lexical factors on children's word-finding errors. *Journal of Speech, Language, and Hearing Research* 47. 624-636.

Gierut, J.A. 2001. A model of lexical diffusion in phonological acquisition. *Clinical Linguistics & Phonetics* 15. 19-22.

Jones, R.L. & E.P. Tschirner. 2006. *A frequency dictionary of German: Core vocabulary for learners*. New York: Routledge.

Levenshtein, V. 1966: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10. 707-710.

Luce, P.A. & D.B. Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear & Hearing* 19. 1-36.

Portele, T., J. Krämer & D. Stock. 1995. *Symbolverarbeitung im Sprachsynthesesystem Hadifix*. Proc. 6. Konferenz Elektronische Sprachsignalverarbeitung. Wolfenbüttel. 97-104.

Stemberger, J.P. 1984. Structural errors in normal and agrammatical speech. *Cognitive Neuropsychology* 1. 281-313.

Storkel, H.L. 2004. Do children acquire dense neighborhoods? *Applied Psycholinguistics* 25. 201-221.

Storkel, H.L. & M.L. Morrisette. 2002. The lexicon and phonology: Interactions in language acquisition. *Language, Speech, and Hearing Services in Schools* 33. 24-37.

Vitevitch, M.S. & M.S. Sommers. 2003. The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition* 31. 491-504.

Vitevitch, M.S. & M.K. Stamer. 2006. The curious case of competition in Spanish speech production. *Language and Cognitive Processes* 21. 760-770.