A mathematical insight into the size of infant vocabularies

Julien Mayor (julien.mayor@psy.ox.ac.uk)
Department of Experimental Psychology, South Parks Road
Oxford, OX1 3UD, United Kingdom

Kim Plunkett (kim.plunkett@psy.ox.ac.uk)
Department of Experimental Psychology, South Parks Road
Oxford, OX1 3UD, United Kingdom

Abstract

For the last twenty years, many researchers interested in language acquisition have quantified the receptive and productive vocabulary of infants using CDIs – checklists of words filled in by the caregiver. While it is generally accepted that the caregiver can reliably say whether the infant knows and/or produces a given word, we lack an estimate for words that are not listed on CDI. In this study, we provide a mathematical model providing a link between CDI reports and a more plausible estimate of vocabulary size. The model is constrained by statistical data collected from a population of infants and is validated on a longitudinal study comparing diary report with CDI measures.
Keywords: Vocabulary size; early word learning; vocabulary overlap; vocabulary spurt

Introduction

How many words does an infant know? Traditionally, this question has been answered by counting the number of different words an infant produces within a representative period of time. Diary methods and home-based recordings provide an estimate of what the infant knows and offer a rich account of vocabulary knowledge in infants. They are, however, time-consuming and expensive strategies for assessing individual vocabulary sizes. Moreover, infants learn new words every day, while they do not use all their lexicon every day. Such direct measures face a dilemma; short recordings lead to a sub-sampling of the real lexicon whereas longer recordings, over a period of many days mask new acquisitions within that period. A further limitation of these approaches is that they only provide a measure of productive vocabulary which may inadequately reflect an infant's total vocabulary knowledge. Infants may understand many words they do not say and they man say words they do not understand.

As an alternative to diary methods and home-based recordings, one can interrogate parents about their infant's vocabulary knowledge. Parents are useful judges of whether their infant comprehends and/or produces a given word and can be asked to fill in checklists of words likely to be known by their infant. This method provides a snapshot of the infant lexicon on the day the form is completed. In addition, it takes just several minutes for the caregiver to complete the list and so is an efficient and inexpensive mean of assessment.

Many researchers assessing the vocabulary size of infants now prefer this method and rely on Communicative Development Inventories (CDIs). A CDI consists of a list of the most frequent words encountered by infants. The caregiver is asked to indicate whether every word on the list is understood (comprehension) and/or produced (production) by the infant. A straightforward method of estimating the typical vocabulary size at a given age is to average the total number of words known by all infants. This simple process leads to an accurate estimate, provided that the caregiver filled in the CDI reliably (Dale, Bates, Reznick, & Morisset, 1989; Dale, 1991) and that CDI includes a suitable range of words likely to be understood or produced by infants. Experimental validation of an infant's knowledge of items reported by the caregiver has provided further support for the accuracy of the instrument (Dale et al., 1989; Dale, 1991; Styles & Plunkett, 2009) though see (Houston-Price, Mather, & Sakkalou, 2007) for an alternative perspective. In constructing CDIs, researchers have strived to include a representative sample of words that infants know at different ages. However, the CDI is not intended to be an exhaustive listing of all the words that any infant might know.

When vocabulary size reaches a significant proportion of the CDI listing, the likelihood that infants would know words that are not listed on the CDI increases dramatically: "Although the present index might approach the status of an atlas for the younger children, it becomes an increasingly smaller subset of vocabulary for older children" (Fenson et al., 1993, p.40). A simple vocabulary count based on a CDI is therefore unlikely to be an accurate estimate for older infants In order to test the validity of the MacArthur-Bates CDI (Fenson et al., 1993), Robinson et al. (1999) compared the CDI-based productive vocabulary estimate with an exhaustive diary report based on data from a single child. The discrepancy increased dramatically with age, thereby confirming Fenson et al. (2003)'s concerns. Such findings would appear to undermine the utility of the CDI in providing accurate estimates of the number of words an individual infant knows. In particular, the CDI leads to provide a systematic underestimate of infant vocabulary knowledge. The goal of this paper is to demonstrate that this underestimate can be quantified even when the reported vocabulary size is a substantial proportion of the CDI listing. We will show that it is possible to offer a more accurate estimate of vocabulary size given a particular level of performance on the CDI. This new estimate takes into account both the idiosyncracies of an infant's individual vocabulary as well as general omissions of common words from the CDI and aims at providing an estimate of the real vocabulary size of an infant, when her reported vocabulary reaches

a substantial fraction of the CDI size.

Method

We present a simple procedure for estimating more accurately the vocabulary of infants from direct measures on the CDI. The straightforward method is to count the number of word on the CDI that each infant knows and compute the average vocabulary size over all infants of the same age. However, this method has the problem each infant is likely to know words that are not present on the CDI. Consequently, the estimated vocabulary size is the average of inaccurate individual word counts.

CDI forms are compiled so that it is possible to determine the probability that a given word w_i is known by infants at a given age; $p(w_i)$. A basic rearrangement of these calculations shows that the straightforward method for measuring vocabulary – the average of individual vocabulary sizes over all infants – is equivalent to computing the sum, over all words on the CDI, of the probabilities that they are known by an infant.

$$\text{Voc}_{est} = \sum_{i=1}^{W} p(w_i) = \frac{1}{N} \sum_{j=1}^{N} \text{voc(j)} \tag{1}$$

where W is the number of words on the CDI and voc(j) measures the vocabulary size of infant j. The advantage of this formulation is that individual terms in the summation – $p(w_i)$ – approach the "exact" value as the number of infants increases, assuming that the caregivers respond accurately. Any inaccuracy is now the outcome of having a calculation that runs over words on the CDI and not over all words in the language, i.e, for a word not included in the CDI we lack information regarding the probability that it is known by an infant. We can estimate the real vocabulary size in terms of the direct CDI measure, plus a term that corresponds to the underestimate. The underestimate is simply the sum of the probabilities for words that are not listed on the CDI:

$$\text{Voc}_{real} = \text{Voc}_{est} + \sum_{i=W+1}^{W_\infty} p(w_i) \tag{2}$$

with W_∞ being the total number of words in the given language. This calculation is correct provided we can have an accurate estimation of all words that have not been included in the CDI.

We distinguish two sources of underestimation of an infant's vocabulary size. First, individual infant's lexicons are only partly overlapping. For example, an infant whose parent is a mechanic is likely to possess an early knowledge about car related words, otherwise rare among other infants. Idiosyncratic words cannot be listed in a CDI, despite their contribution to overall lexicon size, because they would greatly inflate the size of CDIs and the time taken to complete the form. The second source of underestimation derives from frequent words in the language that are not listed in the CDI. CDIs have been built by listing popular words in the infant's vocabulary. However, it is unlikely that the list is a perfect tabulation of *the W* most frequent words. For example, even highly frequent words have been omitted from the MacArthur-Bates CDI [1]. On the assumption that infants learn highly frequent words before lower frequency words, the CDI would necessarily underestimate vocabulary size in proportion to the number of highly frequent words not included in the CDI. We describe in the next section how to evaluate both effects in order to provide an accurate description of the lexicon size based on MacArthur-Bates CDI reports.

The First Correction and Second Correction: An overview

One can sort words according to the proportion of infants that know the word, in descending order and then plot the probability that infants know a given word as a function of its rank. For example, a word that is know by a vast majority of infants ("daddy") will be ranked high on the list, and a word known by only a small fraction of infants will have a low rank. This probability distribution that a word is know to a given infant is a decreasing function where words of low probability occur in the tail of the distribution. Idiosyncratic words – the first correction – correspond to words that are only known to a small minority of infants. These are words that occur in the tail of the distribution. The size of this underestimate thereby corresponds to estimating the length of the tail. Commonly occurring words absent from the CDI – the second source of underestimation – change the shape of the probability distribution. Quantification of the length of the tail and estimation of the parameters for the correct shape of the probability distribution permits a more accurate estimate of an infant's vocabulary size.

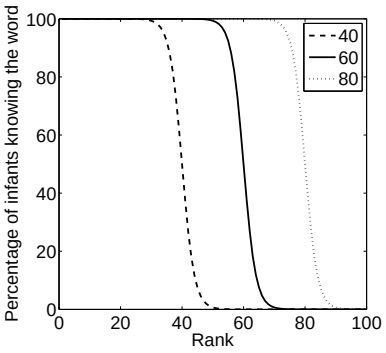
First Correction; Adding idiosyncratic words to the lexicon

We choose to model the distribution of knowledge using a standard sigmoid function that describes the probability $p(w_i)$ that a word is known given its rank i among other words:

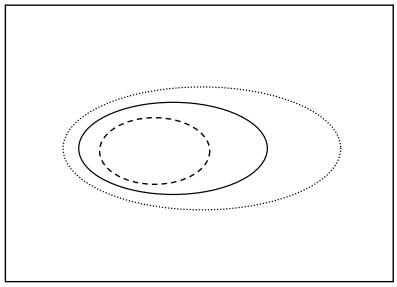
$$f(w_i) = 100(1 - \frac{1}{1 + e^{\frac{-(i-a)}{b}}}) \approx p(w_i) \tag{3}$$

This equation has only two free parameters, a and b. Therefore, finding an optimal value for the parameters is likely to be unique, and the algorithm for finding the solution fast and stable. It also reduces the risk of over-fitting the data, since the number of free parameters is much lower than the number of data points used to constrain the optimisation. Moreover, it provides an intuitive fit of the distribution of knowledge of the words: One expects to have values close to 100% for highly ranked words (very common words, known by everybody) to values closer to 0% for low ranked words, known to only a very small subset of the population. The first free parameter – a – determines overall vocabulary size (see Figure 1a and

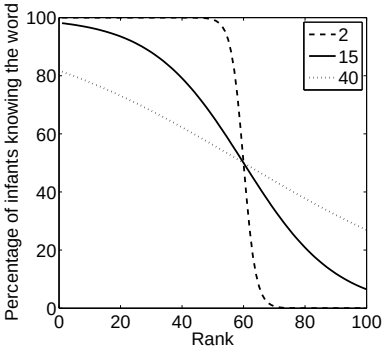
[1] Bag, back, come, computer, digger, down, floor, gate, hole, lift, pigeon, ring, sea, tower, warm, wheel



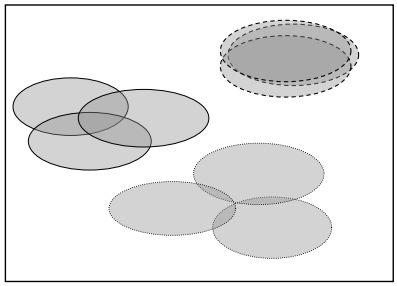
(a) Impact of parameter a on the distribution



(b) Corresponding structures in vocabulary space



(c) Impact of parameter b on the distribution



(d) Corresponding structures in vocabulary space

Figure 1: Examples of curves describing the proportion of infants knowing a word given its rank among other words. The parameter a regulates the overall vocabulary size (top panels) whereas parameter b describes the structure of vocabulary knowledge, i.e., the amount of knowledge overlap in the infant population (bottom panels).

b). More precisely, a determines the location of the rank order of words that are known to 50% of the infants. The second free parameter – b – determines the overlap of word knowledge across the population of infants. A very low value for b corresponds to a steep probability distribution whereas a high value yields a shallow distribution (see Figure 1c). Shallow distributions correspond to low overlap of individual vocabularies whereas low values correspond to high overlap (see Figure 1d). This sigmoidal function possesses another useful property: When the number of words in the language is large and far beyond the sample words included in the CDI, the sum over all words of the function becomes a simple expression of parameters a and b:

$$\text{Voc}_{Corr1} = b \cdot ln(1 + e^{a/b}) \tag{4}$$

where ln is the natural logarithm. Quantification of a and b allows us to determine the value of the first correction.

Second Correction: The role of frequent words missing from the CDI

We assume that the CDI contains the most frequent words known to infants. However, the probability that the CDI lacks a particular word increases with decreasing rank. This is because experienced researchers can reliably list the most common words in the infant's lexicon but will be more prone to error as the word list expands to include less common words. In other words, the more items an infant is reported to understand on the CDI, the greater is the number of potential missing items from the vocabulary estimate. The second underestimate is therefore directly related to the number of words that an infant is reported to know. The fraction of omitted words can be written as:

$$f_{omission} = \alpha \text{Voc}_{Corr1} \tag{5}$$

The only way to quantify this underestimate is by direct comparison with exhaustive word lists that individual infants know such as that reported by Robinson and Mervis (1999).

Results

The validity of the procedure for the first correction is tested by applying it to the Lex2005 database (Dale & Fenson, 1996), based on the American MacArthur-Bates CDI (Fenson et al., 1993), for both production (Words and Sentences) and comprehension (Words and Gestures). Overall, results suggest that the underestimation due to the absence of idiosyncratic words on the CDI increases with age, while the degree of overlap between individual vocabularies remains relatively constant. A simple analytical prediction of the underestimation when vocabulary size reaches a significant proportion of the CDI size is presented. A strong, non-linear, increase of estimated vocabulary size is predicted as the measured vocabulary size becomes large with respect to the CDI.

We investigate the impact of the absence of frequent words in the CDI on the estimated vocabulary size. This second correction is applied to production data on the McArthur-Bates CDI and the number of missing words is estimated, based on a comparison of a diary count and CDI count for a single case study (Robinson & Mervis, 1999). The addition of these two corrections enables us to predict a more accurate estimate of vocabulary size for individual infants as well as typical mean vocabulary sizes from 8 to 18 months of age in comprehension and from 18 to 30 months of age in production.

First Correction: Underestimation due to the absence of idiosyncratic words

The method is applied for both production (16 months to 30 months old) and comprehension (8 months to 18 month olds), based on the MacArthur-Bates CDI (Dale & Fenson, 1996). For each age group, words are sorted in descending order from those that are known by most infants to the least known words. A regression is applied to identify the parameters a and b that minimise the squared error between the data and the model. All fits of the CDI data explain at least 80% of the variance, indicating that a regression using Equation 3 is applicable. The vocabulary size as predicted by the model is obtained by computing Equation 4 with the measured parameters a and b. Figure 2(a) depicts a comparison of the vocabulary in comprehension from the CDI data and from the model (top panel). The model (solid line) predicts a higher vocabulary size for older age groups (14 months and older) than direct CDI measures (dashed line). The model also indicates a smaller vocabulary size than the direct estimate for younger age groups (up to about 11 months). The discrepancy is plotted in Figure 2(b). The negative error at 8-11 months may be due to over-reporting from the caregiver, or to an unsuitable equation to describe these early words or proto-words (in ranked order; mommy, daddy, bye, peekaboo, bottle, no, hi). Note that this source of error is not itself a measure of idiosyncratic words in the infant vocabulary, but an indirect effect of estimating the parameters a and b which quantify the idiosyncratic contribution. As expected, the direct CDI measure underestimates the vocabulary size as age increases. At 15–18 months of age, the model suggests that the CDI underestimates vocabulary size by the order of 8–10%. However, for the ages ranging from 11–13 months, the estimate based on a direct count is reliable.

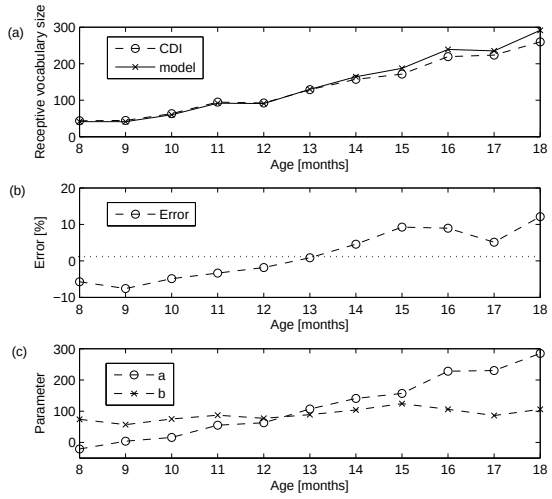


Figure 2: Comprehension. (a) comparison of vocabulary estimates from the CDI and from the model, (b) discrepancy between the two estimate, (c) evolution of the two free parameters with age.

Both parameters a and b are plotted for the different age groups in Figure 2(c). The decomposition of the vocabulary curve into two parameters allows us to disentangle the contribution to overall vocabulary growth, a, and of overlap of vocabulary knowledge across infants, b. Figure 2(c) shows that parameter a mirrors the overall vocabulary growth, approximating the typical vocabulary size of the infants. Parameter b shows that the amount of overlap across infants in the vocabulary space stays relatively constant and that the number of "unique" idiosyncratic words stays approximately constant over the age range under consideration.

The same procedure has applied to the productive vocabulary of toddlers, from 16 months to 30 months of age, using the MacArthur-Bates CDI in English (not shown, due to the limited space). The model predicts a higher vocabulary size than a direct count from the age of 19 months of age, with a discrepancy that increases with age. From about 19 months of age, the underestimate of vocabulary size increases steadily to reach about 18% at 30 months of age. The gradual increase with age suggests that the underestimate of productive vocabulary based on a direct count will be even greater for older toddlers. Parameters a and b can also be computed for productive vocabulary and parameter b remains essentially constant after 19 months of age, indicating that shared productive vocabulary does not change over time.

Having estimated the parameters a and b from the CDI population data, we can estimate the size of a individual infant's vocabulary given a CDI score. We have established that for both receptive and productive vocabulary, there is an increase in the magnitude of underestimation as the total vocabulary increases. The absolute underestimate is determined by the overall size of the CDI and the parameters a and b that we have used to fit group scores to the direct measure of the CDI.

$$\text{Absolute underestimate} = \sum_{W+1}^{W_\infty} f(w_i) = b \cdot ln(1 + e^{\frac{a-W}{b}}) \quad (6)$$

where W is the overall size of the CDI and $f(w_i)$ is the model's estimate that an infant knows the words w_i given the word ranking i.

We have seen that the overlap of knowledge is essentially constant for older age groups. Therefore, we assume that the magnitude of the underestimation can be calculated using the

same value of parameter b for the oldest age groups. Equation 6 predicts the magnitude of the error as the overall vocabulary becomes a substantial proportion of the CDI size W.

Second Correction: Evaluation of the role of omission of frequent words on CDIs

We have assumed so far that the CDI is "perfect", in the sense that *all* the W most frequent words are present on the CDI. Another source of underestimation is the omission of words that should be listed on the CDI that have been left out of the selection of the words known by most infants. The omission of such words may also have an impact on the estimate of parameters a and b, because it changes the shape of the probability distribution, thereby leading to a biased estimate of the contribution of the tail of the distribution of word knowledge to the overall vocabulary. In order to disentangle the impact of the absence of frequent words from the contribution of idiosyncratic words, we randomly deleted a percentage of words on the CDI and compared the direct vocabulary count with the model estimate. If the omission of a fraction of words on the CDI has the same impact on the direct count and the model estimate, this would indicate that missing frequent words do not induce complex biases when estimating the tail of the distribution of word knowledge. In a separate simulation (not shown for space constraints), we showed that both effects do not interfere with each other.

The linearity of the impact of omitting frequent words on the CDI is important. It implies that the two effects – absence of idiosyncratic words and omission of frequent words on the CDI – don't interfere with each other, at least to a first approximation. Consequently, the overall procedure for estimating the correct vocabulary size for infants can be decomposed into two parts. First, based on the CDI, one can estimate the parameters a and b, defining respectively the overall vocabulary size and the overlap of vocabulary knowledge across the population of infants with Equation 3. We can then use Equation 4 to provide the estimated vocabulary size of the infants – Voc_{Corr1} – including the contribution of idiosyncratic words, not listed on the CDI. Finally, an estimate of the fraction of words that should be on the CDI (i.e., among the W most known words) is used to increment the total vocabulary by the same *fraction* rather than the absolute number of words missing from the CDI: $\mathrm{Voc} = \mathrm{Voc}_{Corr1}(1 + f_{omission})$

As mentioned earlier, the fraction of words missing on the CDI is likely to be smaller amongst the most frequent words, where an exhaustive list of well-known words can be established relatively easily, compared to less frequent items, where listing all the better-known words is a difficult task. Therefore, we assume that the fraction of missing words increases linearly with the lexicon size, according to Equation 5. An exhaustive comparison of productive vocabulary based on a detailed diary report with a CDI-based estimation is presented in Robinson & Mervis (1999). They reported the vocabulary production of one male child from about 10 months of age to 2 years of age and identified, on a monthly basis, the total number of different words produced and counted how many of these words were listed on the MacArthur-Bates CDI. As predicted, the underestimation (the words produced that are not on the CDI) increased with vocabulary size. This comparison allows us to constrain the free parameter α of Equation 5 of the second correction by fitting the fully corrected curve to the data provided by Robinson & Mervis.

Note that the first correction for the absence of idiosyncratic words is a mathematical property of computing a definite integral of a sigmoidal curve, and is defined by the measured overlap of vocabulary knowledge and by the size of the CDI form. With this first correction, the shape of the corrected vocabulary size based on direct measurement is defined. The fit to Robinson & Mervis' data is a linear transformation of the first correction. The strong non-linearity predicted by the model derives from the first correction introduced by idiosyncratic words. The omission of frequent words on the CDI serves only to modulate this non-linearity.

Figure 3 depicts corrected vocabulary sizes based on measured productive vocabulary scores, from several different data sources. The triangles correspond to the individual corrections based on the MacArthur-Bates CDI (each triangle corresponds to a different age group with a slightly different overlap parameter b) whereas the dashed line corresponds to the first analytical correction for the role of idiosyncratic words in individual lexicons (only one parameter b for all age groups). The non-linearity of estimated vocabulary size (first correction) based on the number of words reported on the CDI is already apparent. The empty circles are based on the data reported by Robinson & Mervis (1999), where

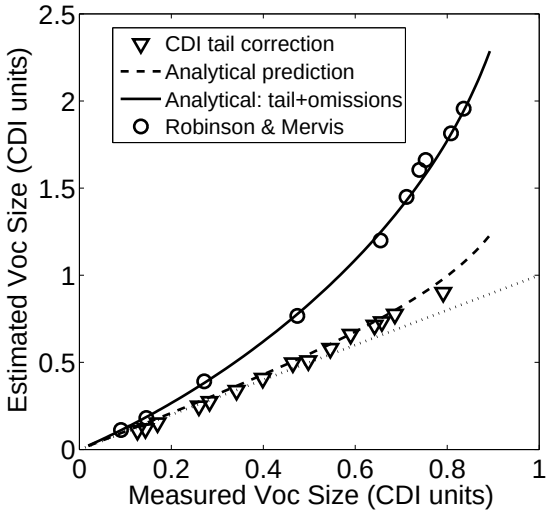


Figure 3: Estimated productive vocabulary size as a function of measured vocabulary, according to the first correction (triangles pointing downwards and dashed line for the analytical prediction) and both corrections (solid line). For comparison, the comparison of diary data with CDI data from Robinson & Mervis (1999) is shown (○). See text for further details.

they compared the lexicon size of an infant based on dairy report to the lexicon size as measured by the CDI. The solid line corresponds to the prediction of vocabulary size based on the fully-corrected model. The fit to the Robinson & Mervis (1999) data confirms the strong non-linearity of underestimation as the number of words in the lexicon increases. This clear agreement between experimental data and the model suggests that the two corrections account for the increasing underestimation of the lexicon as the number of words reported on the CDI increases.

Discussion

We have proposed a mathematical model for estimating the vocabulary size of infants given their CDI scores. Two corrections are applied to the raw CDI measurements: First, the number of idiosyncratic words is estimated via a measure of the overlap of individual vocabularies in the infant population. Second, an estimation of the number of omitted words on the CDI is constrained by a comparison of dairy reports and direct CDI counts.

The model is applied to the MacArthur-Bates CDI database. The model suggests that, as predicted by its creators, the underestimation of vocabulary size increases with the number of words known on the CDI. Moreover, the magnitude of the underestimation is highly non-linear; for small vocabularies the straight CDI count is relatively accurate whereas when infants know 90% of the words on the CDI, they are likely to know about three times as many words.

This non-linearity has important implications. For example, CDIs are a widely used tool for diagnosing language delays. Whereas criteria for identifying delays are not absolute, we advocate against their use based on the raw CDI scores. For example, imagine that a whole data-set consists of just three infants; infant A is reported to know one word on the CDI, infant B six words and infant C nine words. A diagnostic based on the raw data would suggest infant B knows about 10% more words (6.0) than an average infant (5.3). However, after a non-linear transformation, this result does not hold. If, as an example, actual vocabulary sizes are the square of direct CDI scores, infant B would then know about 10% *less* words than an average infant (36 vs. 39.3).

The results can also offer additional information for those interested in characterising the vocabulary spurt often observed at the end of the second year of life. Many researchers attempt to identify an inflection point in the increasing vocabulary size. Often, direct measure from the CDI indicates a slowing-down in the speed of acquisition of new words after the spurt. Again, the non-linearity of the correction suggests more caution in the analysis of vocabulary sizes as the deceleration is likely to disappear after the vocabulary size correction.

Finally, the analysis of the amount of overlap between individual vocabularies suggests that the absolute number of idiosyncratic words does not change with age, over the considered age range. This observation may offer some important boundary conditions in the attempt to apply statistical models of lexicon growth such as preferential attachment (Steyvers & Tenenbaum, 2005) or preferential avoidance (Hills, Maouene, Maouene, Sheya, & Smith, 2008).

References

Dale, P. (1991). The Validity of a Parent Report Measure of Vocabulary and Syntax at 24 Months. *Journal of Speech, Language and Hearing Research*, *34*(3), 565–571.

Dale, P., Bates, E., Reznick, S., & Morisset, C. (1989). The validity of a parent report instrument of child language at 20 months. *Journal of Child Language*, *16*, 239–249.

Dale, P., & Fenson, L. (1996). Lexical development norms for young children. *Behavior research methods, instruments & computers*, *28*(1), 125–127.

Fenson, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., et al. (1993). *Macarthur communicative development inventories: User's guide and technical manual*. San Diego: Singular Press.

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2008). *Categorical Structure in Early Semantic Networks of Nouns.* Austin, TX: Cognitive Science Society.

Houston-Price, C., Mather, E., & Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behaviour in a preferential looking task. *Journal of Child Language*, *34*(04), 701–724.

Robinson, B., & Mervis, C. (1999). Comparing productive vocabulary measures from the CDI and a systematic diary study. *Journal of Child Language*, *26*(01), 177–185.

Steyvers, M., & Tenenbaum, J. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, *29*(1), 41–78.

Styles, S., & Plunkett, K. (2009). What is 'word understanding' for the parent of a one-year-old? Matching the difficulty of a lexical comprehension task to parental CDI report. *Journal of Child Language*, *xxx*, xxx.