A Sample-size-invariant Estimation of Lexical Diversity

Shohei Hidaka (shhidaka@indiana.edu)
Department of Psychology and Brain Sciences,
1101 E. Tenth Street Bloomington, IN 47405-7007 USA

Abstract

In a long history of studies on language development, the lexical diversity has been used as one of measure of vocabulary growth. A typical analysis, type-token ratio, which is ratio of sample size of words to the number of unique types of words, has been used as a measure of lexical diversity. Because the type-token ratio is not valid for vocabulary sets with different token sample size, recently, an improved measure has been proposed. In the present study, however, we show that the diversity estimated by the past proposed measures are not robust for corpus with different sample sizes. Accordingly, we propose a new formal model of the lexical diversity, which distinguish the latent number of types and property of frequency distribution robustly for different token size. The proposed model is compared to other measures of lexical diversity. We also applied the method to an actual vocabulary development data.

Keywords: Lexical diversity; Small sample size; Number of latent types; Power-law distribution

Formal Measures of Lexical diversity

Lexical diversity has been considered as one of central characteristics of a corpus. In corpus linguistic literatures, there are many proposals for a measure of the lexical diversity. However, most of these proposals have suffered from the problem that these measures of lexical diversity vary depending on the sample size. For example, the type-token ratio is one of the most frequently used measure of lexical diversity. The type-token ratio is the ratio of types, the number of unique words, relative to tokens, the number of sampled words. It has been still used nowadays despite of its well-known flaw. An apparent flaw of this method is that the type-token ratio depends on the sample size of tokens. A corpus with a larger number of tokens might have a relatively smaller number of types, because no types would appear after all possible types are sampled. Due to this nature of the type-token ratio, one must compare the type-token ratio of the same sample size. This is well known in the literatures (Javis, 2002; Malvern & Richard, 2002). Since measures of lexical diversity vary on the sample size more or less in general, many studies have proposed different ones in order to normalize the sample size effect.

The first kind of proposal is to use a transformed function of type-token ratio, such as type-token ratio with squared sample size (Guiraud, 1954), logarithm (Herdan, 1960; 1964), power function (Dugust,1979), etc. However, these transformations has the same problem as the type token ratio has (for example, Weitzman, 1971).

The second kind of proposals is based on frequency spectrum, using the number of types sampled i-times in sample size M, $V(i, M)$. For example, Good (1953) propose the entropy of frequency distribution as a measure (Good, 1953), however, this measure also depends on sample size (Johnson, 1979).

The third kind which has been proposed relatively recently is a curve fitting method using the type-token ratio as function of sample size. Malvern & Richard (1997, 2002) have proposed the slope parameter of the type-token ratio curve as a measure of lexical diversity. If the curve of type-token ratio as function of sample size follows a particular class of function with some free parameter, this curve would be estimated by fitting the theoretical curve to multiple sampling points of an actual data set. Their basic idea seems attractive, because we do not need a large sample size if the curve can be estimated with a small number of samples.

The Approach in the present study We follow the basic idea of this recent theoretical proposal, a curve fitting method. However, our approach is supposed to be a combination of the second and third type of proposals. That is our model estimates both the number of latent types and property of frequency distribution based on the curve regression. This method, estimating both types and property of frequency distribution at the same time, has two significant benefits. First, it is robust to different size of samples, because both number of types and property of the frequency distribution are theoretically sampling-invariant properties. Obviously, all the past efforts on measurement of lexical diversity aim to this, robustness of measurement to different sample size. However, as we show in later section, even the measure using curve regression is not robust. The reason is also relevant to the second point.

Second, the estimation of the method which we propose has a smaller bias, because it separates number of latent types and property of frequency distribution which both influence the curve of sampled types in different ways. The basic idea of curve regression is that the slope of the type-sample curve should be robust to independently from the sample size. However, the slope of curve depends on both the latent number of types and tail-heaviness of frequency distribution (See also Theoretical Formulation). Accordingly, even if two given sets of corpora have the same numbers of latent types, this does not mean that the sampled results from these two corpora would be the same. For example, the corpus A has the uniform distribution of types in which any type is sampled in the same rate, in contrast, the corpus B might have a heavy-tail distribution such as the Zipf distribution in which a few types are sampled in high rate but most types are sampled in low rate. The corpus with the uniform distribution seems to have high diversity than those with other skewed distributions even in a same sample size, because the sampled types to the number of samples would be larger in that with the uniform distribution.

Therefore, when we measure the diversity by fitting the sample-type curve, which is compound of at least two separable properties of the corpus such as the latent number of types and the property of word frequency. Without separating these properties, a measure cannot avoid bias in estimator of type-sampling curve. As a result of using a biased estimator, two corpora may be estimated as if they had the same diversity even when they have different set of properties.

In the present study, we allow us to focus on a particular class of frequency distribution. That is the Zipf (Zeta) distribution or that known as "Zipf's law" in which frequency of the type follows the power of ranks (Zipf, 1949). Since the Zipf's law is observed in many empirical corpora, it is useful to evaluate the lexical diversity of a broader range of corpora. We derive the average sampled types as function of sample size in case that frequency follows the uniform distribution or Zipf distribution. The theoretical type-sample curve was confirmed in a numerical simulation, and we compare this with the other standard methods estimating lexical diversity on analysis of empirical data.

Theoretical Formulation

Here we derive the theoretical number of types for given sample size. In the special case, when word frequency follows uniform distribution, the expected number of types is precisely obtained in an analytic form. Moreover, we approximately evaluate the expected number of types in an extended form when the word frequency follows the Zipf distribution. We call this the *type sampling model*, because it assumes independent sampling of each type. The lexical diversity in the model is defined with the two factors, the latent number of types in a corpus, which is the maximum number of types in sampling an infinite number of tokens, and the exponent of power-law frequency distribution.

Types for the uniform frequency distribution

We first derive the probabilistic distribution $P(n,m,k)$ of the number of sampled types k for a given sample size (the number of types) m when there are *n-latent types* and their frequency follows the uniform distribution. That is $P(n,m,k) =_n C_k n^{-m} \sum_{i=0}^{k} (-1)^i \, {}_kC_i (k-i)^m$ where ${}_kC_i = k!\{i!(k-i+1)!\}^{-1}$ is the number of possible combination of drawing i types from k types. The number of possible combination of sampling is n^m, and the possible combination of sampling (k-i)-or less types of words $(k\text{-}i)^m$. Since the number of combination that one sample just k-types of words is subtract that of (k-1)-types of words from $(k\text{-}i)^m$ (i.e., inclusion and exclusion principle), after recurrent calculation, we obtain $\sum_{i=0}^{k} (-1)^i \, {}_kC_i (k-i)^m$. This result can be rewritten with the Stirling number of the second kind $S(m,k) = (k!)^{-1} \sum_{i=0}^{k} (-1)^i \, {}_kC_i (k-i)^m$ which is the number of possible combination of partitioning m elements into k groups. Using the Stirling number of the second kind, the probability is rewritten as follows: $P(n,m,k) = (n)_k n^{-m} S(m,k)$ where $(n)_k = n!/(n-k)!$ is the falling factorial.

Next we derive the expectation of sampled types $E[k \mid m,n]$ when m tokens are sampled from corpus with uniform-distributed n latent types. According to the property of the Stirling number, the recurrence formula $S(m,k) = S(m-1,k-1) + kS(m-1,k)$ and $\sum_{k=1}^{m} (n)_k n^{-m} S(m,k) = 1$ are available. It leads the following recurrence equation: $E[k \mid m,n] = E[k \mid m-1,n](1-n^{-1}) + 1$. By solving this with $E[k \mid 1,n] = 1$, the following is derived

$$E[k \mid m,n] = n\left(1-\left(1-n^{-1}\right)^m\right) \approx n\left(1-\exp\left(-mn^{-1}\right)\right) \qquad (1).$$

This result suggests that the number of sampled types follows the sum of cumulative probability that each type is independently sampled until m tokens are sampled. Note that the $\left(1-\exp\left(-mn^{-1}\right)\right)$ is the cumulative exponential distribution with the sampling rate $P(k) = n^{-1}$. This insight is used for the extension of this result.

Types for the power-law distribution

Next we derive the number types for given sample size when the word frequency follows a power-law distribution. The Zipf distribution is $P(k) = k_0^{-1} k^{-a}$ where $k_0 = \sum_{k=1}^{N} k^{-a}$ and k is the rank of type in token frequency (i.e., k=1 for the most frequent type, and k=2 for the second most frequent one) and a is the exponent of power function. In fact, the uniform distribution is considered as the special case of the Zipf distribution with the exponent a=0. In general Zipf distribution, the probability distribution of the number of sampled types $P(n,m,k)$ or the expected number of sampled types $E[k \mid m,n]$ is hard to derive in a closed form. However, the expected number of sampled types can be calculated by assuming that sampling of each type is independent which is approximately true mentioned for the uniform distribution above. We use this insight in order to extend it to the number of types with a non-uniform distribution

By assuming sampling of each type is independent, the calculation of the sum of number of types is simple. Replacing the probability of the uniform distribution $P(k)=n^{-1}$ in equation (1) with the probability of the Zipf distribution $P(k) = k_0^{-1} k^{-a}$, we obtain the following Equation.

$$E[k \mid m,n] = \sum_{k=1}^{n} \left(1-\exp\left(-mk_0^{-1}k^{-a}\right)\right)$$

This summation does not have a closed analytic form, but we approximate this with integral as follows.

$$E[k \mid m,n] \approx \int_1^{n+1} \left(1-\exp\left(-tk_0^{-1}k^{-a}\right)\right)dk \qquad (2)$$
$$= N + G\left(a, tk_0^{-1}\right) - G\left(a, tk_0^{-1}(N+1)^{-a}\right)$$

where $G\left(a, tk_0^{-1}\right) = {}_1F_1\left(-a^{-1}; 1-a^{-1}; -tk_0^{-1}\right)$ is the confluent hypergeometric function of the first kind.

Monte Carlo simulation Since we approximately derive the expected number of types, we evaluate how much the theoretical function is dissociated from the numerical values. Figure 1 shows the theoretical number of types as function sample size in five cases, {N,a}={10^2,1.2}, {10^2,1.5}, {10^2,2.0}, {10^3,1.2}, {10^4,1.2}. The numerical values were calculated by the average across samples in the Monte Carlo simulation. The theoretical values fit well to the numerical values in all five cases (R>0.99).

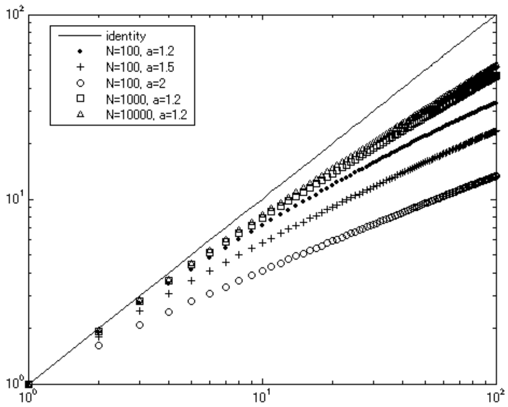

Figure 1: The theoretical expected number of types as function of sample size.

Simulation

In order to evaluate how precise the model can estimate the latent number of types, we ran Monte Carlo simulations of the number of sampled types. We compared the number of latent types estimated by the type sampling model with that of estimated by the Good-Turning method. The Good-Turning estimation is derived upon a more general assumption, and thus it is distribution free estimator. Our model has a strong and less general assumption compare to the Good-Turning estimation, and thus the goal of the simulation is how this additional assumption gains the precise estimation on the number of types. We also compare it with the other curve-fitting method proposed by Malvern & Richards (1997, 2002) and its superordinate model (Sichel, 1986).

Method

Procedure In the simulation, for given the number of types and exponent of distribution, a set of randomly generated types are sampled m times. The set of types follow the Zipf distribution with the exponent -*a* and number of latent types *n*=1000. The sample size *m* and the exponent of the Zipf distribution *a* are manipulated. Specifically, the sample size varies *m*={300, 500, 1000, 2000} and the exponent varies *a*={0, 0.5, 1.0, 1.5} in the artificial corpora. 10 different sets of artificial corpora for each condition are used for calculating average and standard deviation of estimated values.

Type sampling model The number of sampled types as function of the number sampled words is fitted with the equation (2) in Theory Formulation. The parameters *n* and *a* are estimated by minimizing the least square error $L = \sum_{m=1}^{M} \{\log(k(m)) - \log(E[k \mid m,n])\}^2$ where $k(m)$ is the average number of sampled types when *m* words are sampled which is calculated from 50 reordered sets of the same sampled words (i.e., for the same set of sampled words, the sampled order *m* is randomized and averaged across the 50 duplicated sets).

Exponent of Zipf distribution by regression
The log-log plot of the Zipf frequency distribution as function of rank *n* is linear, and its slope is the exponent *a*. The regression of the log-log plot is used in order to estimate the empirical exponent of distribution for given samples.

Good-Turing estimation
According to Good (1953), the expected proportion of unknown types to all possible types is the proportion of types sampled only one time to all types sampled. Thus, the expected number of latent types *n* is estimated by $n = \bar{n} V(1,m)^{-1} \sum_{k=1}^{m} V(k,m)$ where $\bar{n}$ is the number of sampled types in *m* total token samples and $V(k,m)$ is the number of types with k tokens in *m* total token samples.

Malvern & Richard's *vocd* model
Malvern & Richard (1997, 2002) have proposed a curve regression method to evaluate lexical diversity using a model fitting a type-token ratio curve. This is implemented as *vocd* function in CLAN system of the CHILDES project (MacWhinney & Snow, 1990), and we call this the *vocd* model hereafter. According to the vocd model, the expected number of sampled types is described with the following equation: $E[k \mid m,D] = D\sqrt{1+2mD^{-1}} - D$ where the *D* is a free parameter indicating the slope of the curve. Note that the original equation is for type-token ratio, but this is modified by multiplication with the sampled token size *m*. The vocd model does not have an explicit ceil of number of sampled types (i.e., $\lim_{m\to\infty} E[k \mid m,D] = \infty$) The diversity parameter *D* in this model is originally derived from the inverse-Gauss-Poisson distribution (Sichel, 1986). We also analyzed the original equation (hereafter called the Sichel model) $E[k \mid m,D] = 2(bc)^{-1}\{1-\exp(-b(\sqrt{1+cN}-1))\}$ where *b* and *c* are parameters and $2(bc)^{-1}$ is the theoretical number

of latent types. The parameter *D* in the vocd model or *b* and *c* in Sichel's model is estimated in the same way (the maximum likelihood) as parameters in the type sampling model are.

Results and Discussion

The two methods estimating the number of latent types are compared for the four artificial corpora following different frequency distributions with exponent=0, -0.5, -1.0, -1.5. The Good-Turning estimation succeeds in estimation of the number of types when the word frequency follows uniform distribution (exponent=0). However, the Good-Turning estimation is poor when the word frequency follows the Zipf distribution with exponent -0.5, -1, and -1.5. Good-Turing estimation in all conditions but exponent=0 did not have the true value (the number of latent type *n*=1000) in its average estimated number of types or range of one standard deviation (the middle row in Table 1). The Good-Turning estimation for the Zipf distribution is biased toward the sampled number of sampled types. For instance, the estimated number of types is 190.2 for the number of observed types 180.4 in case the exponent is -1.5.

In contrast, the type sampling model predicts better than the Good-Turning estimation does when the word frequency follows the Zipf distribution. The number of latent types estimated by the model is close to the true value (the bottom row in Table 1). For both methods, estimated values are more inaccurate when the exponent of the Zipf distribution is small (large negative value). This would be due to the smaller number of sampled types. However, the type sampling model have the true value of number of types in its one standard deviation from the mean the even in a difficult case when the exponent is -1.5. One the other hand, the Good-Turning estimation does not predict the true value within the range of its one standard deviation.

It is interesting that the exponents estimated by the type sampling model are more precise than the exponents estimated by regression using the empirical frequency (the sixth row in Table 1). The empirical exponents are estimated by linear regression of double logarithmic plot of probabilities and ranks. Although this is a simple and easy method often used, it is not as good as the estimation of type sampling model (the second lowest row in Table 1). Probably, this may be due to difference between the utilized information in two methods. The empirical exponents are estimated from frequency distribution of *m*-time sampled tokens. On the other hand, The type sampling model estimates it from the entire curve of the number of sampled types as function the sampling from 1 to *m* times.

Next, with different sample sizes, the two methods are compared for the corpora whose word frequency follows the Zipf distribution with the exponent -1. Since the common aim in both vocd model and type sampling model is to estimate lexical diversity independently to the number of sample size, the result of this simulation is significant to evaluate the models. Table 2 shows the results of the simulation. In the Good-Turning estimation, the estimated numbers of types in any sample size from 300 to 2000 are all biased, and the estimate is worth with a smaller sample size (the middle row group in Table 2). As well as the Good-Turning estimation, the empirical exponents estimated by regression analysis are also biased toward smaller values, and the bias is bigger with a smaller sample size (the second top row group in Table 2). In addition to these two above, the curve fitting methods using the vocd model (Malvern & Richard, 2002) and Sichel's model (Sichel, 1986) were also compared (the two bottom row groups in Table 2). The result shows the parameters in both two models varies along sample size. Also the number of latent types estimated by Sichel's model has huge deviation from the true value. This suggests that the vocd and Sichel's model are not robust estimator of lexical diversity when the frequency of types follows the Zipf distribution.

In contrast, the type sampling model estimates both number of types and exponents fairly well in any sample size from 300 to 2000. In particular, the estimation of exponents seems good independently to the sample sizes of tokens (the bottom rows in Table 2). The standard deviation of estimated exponents is quite small even in the sample size 300. The mean number of latent types estimated by the type sampling model seems to have a small bias, and its standard deviation increase with the larger sample size of tokens. However, from the sample size 500 which gives only 215.2 sampled types (i.e., 20% of the true number of latent types 1000), the model can estimate a closer value (n=1177) than the other methods do. Moreover, the type sampling model is fitted to the actual number of types better than the alternatives. The summed square error (SSE) of the vocd model and Sichel's model are significantly larger than that of the type sampling model. It means the type sampling model fits the empirical curves given by Monte Carlo sampling better than others do.

Table 1: The number of latent types (denoted by "Types") estimated by the Good-Turing method and the type sampling model. The asterisk indicates the closed value to the true value in all methods.

True Values	Exponents	0	−0.5	−1	−1.5
	Types	1000	1000	1000	1000
	Sample size	2000	2000	2000	2000
Empirical	Types	**864.7**	**790.8**	**494.9**	**180.4**
	S.D.	10.96	9.7804	9.7804	6.132
	Exponents	**−0.483**	**−0.626**	**−0.890**	**−1.221**
	S.D.	0.0082	0.011	0.011	0.028
Good–Turning	Types	**1001.9***	**929.2**	**569.2**	**190.2**
	S.D.	19.76	20.44	12.78	7.16
Type sampling	Types	**1007.5**	**1018.1***	**1093.9***	**1554***
	S.D.	18.51	24.65	96.15	1125
	Exponents	**−0.083***	**−0.503***	**−0.996***	**−1.52***
	S.D.	0.073	0.012	0.011	0.021

Accordingly, the simulation suggests that the type sampling model is more robust to various sample size than other methods are. In particular, a typical corpus obtained in developmental studies does not offer a large sample size, but they rather have small numbers of sample sizes for individual children. The results in this simulation suggest that the type sampling model is suitable for such situations. For instance, the number of latent types, which is supposed to be the potential vocabulary size, might be estimated accurately based on a relatively small size of corpus.

Table 2: The number of latent types estimated by five methods for different sample sizes. The asterisk indicates the closest value to the true value in all methods.

True Values	Exponents	−1	−1	−1	−1
	Types	1000	1000	1000	1000
	Sample size	300	500	1000	2000
Empirical values	Types	**145.1**	**215.2**	**337.7**	**495**
	S.D.	6.30	6.07	10.96	9.78
	Exponents	**−0.655**	**−0.694**	**−0.788**	**−0.890**
	S.D.	0.0276	0.0142	0.0214	0.011
Good–Turning	Types	**224.9**	**307.4**	**428.8**	**569**
	S.D.	16.60	15.46	17.00	12.78
Type sampling	Types	**1115.7***	**1177***	**1144***	**1094***
	S.D.	400.08	212.88	157.19	96.15
	Exponent	**−1.003***	**−0.999***	**−1.005***	**−0.996***
	S.D.	0.041	0.024	0.020	0.011
	SSE	**0.014***	**0.010***	**0.018***	**0.032***
Vocd	D	**56.56**	**64.03**	**76.86**	**79.42**
	S.D.	8.37	6.01	6.14	2.84
	SSE	**0.42**	**0.88**	**1.57**	**2.87**
Sichel	Types	**1.83 × 10^{14}**	**6.63 × 10^{11}**	**8.71 × 10^{11}**	**5.48 × 10^{11}**
	S.D.	5.71 × 10^{14}	1.39 × 10^{12}	1.17 × 10^{12}	5.36 × 10^{11}
	SSE	**0.34**	**0.76**	**1.78**	**2.58**

Analysis 2: Application to Vocabulary Development

Next we applied the type-sampling model to a real development data. We used corpora from a longitudinal study on conversations of three child-caregiver pairs in free play situation (Brown, 1973) in CHILDES database (MacWhinney & Snow, 1990). The corpora include a short time conversation (30 to 60 minutes) of for each age from 27 months to 61 months old. For each corpus, the model is applied to child's and caregiver's word frequency separately. Since the used data is a brief conversation at a particular age, the estimated number of words would not reflect the entire number of words but that in a particular context. Although the data source is limited, it can describe the lower bound of the latent word acquired until the age.

Method

Data The brief sessions of conversations in mother-child pairs are analyzed. Three children analyzed are Adam (55 sessions from 27.1 to 60.4 month of age, 60 minutes long each), Eve (20 sessions from 18 to 27 month of age, 60 minutes long each), and Sarah (139 sessions from 27.2 to 61.2 month of age, 30 minutes long each).

Analysis The set of frequency of all types are submitted to the type-sampling model and the vocd model. Frequency of a child and his or her caregiver in each session is analyzed separately. The procedure is the same as that in analysis of artificial datasets. Two sessions in Sarah's corpora were excluded from analysis due to too small numbers of tokens.

Results and Discussion

For each conversation session, the type-sampling model and vocd model are applied. First we compared the model fitting of type-sampling and vocd model. Both models reasonably fit the sampled-type curve in overall (R>0.98). However, for all sessions of both child and caregiver's separately sampled words, the type sampling model fits better than the vocd model does. The summed square error of the type sampling model in all sessions is 4.01, and that of vocd model is 19.53. This suggests the empirical frequency in child-caregiver conversation is well described with the Zipf distribution rather than inverse-Gauss-Poisson distribution that the vocd model assumes.

Next we focus on the number of latent types and exponent of frequency distribution estimated by the type sampling model. The analysis shows that the numbers of latent types in all three children increase along their age. In contrast, the caregivers' number of latent types does not increase very much along the children's age except for Eve's caregiver. Sarah and her caregiver's numbers of latent types estimated by the type sampling model are shown in Figure 2. Adam and Sarah's number of latent types start from a smaller point and tend to converge to their caregivers' number of latent types. Although the number of latent types might depend on the context of each session, the correlation between mother and child is not high enough to explain the entire pattern of child's number of types. Thus, this increment pattern of number of types might be due to child's development.

The patterns in the estimated exponents of frequency distribution show correlation to child's age except for Sarah. Adam's and his caregiver's estimated exponents are shown in Figure 3. Figure 3 shows Adam's exponent grows higher along his age. Sarah's exponents vary from -0.6 to -1.1 at her 27 to 30 months of age, but it converges to -0.8 which is close to her caregiver's average exponent -0.77. From the estimated exponents, we could not observe a common and consistent child-caregiver pattern among three pairs.

Due to the small number of children, we cannot draw a strong and general conclusion, but the analysis reveals a common growth pattern of the number of latent types in all three children, and it is more correlated to children's age than those of caregivers do.

Table 3: Correlations among the latent number of types and exponents of age, the child, and caregiver. *: p<0.05, **: p<0.01.

		Child-Age	Caregiver-Age	Child-Caregiver
Adam	Types	0.789**	-0.181	0.151
	Exponent	0.451**	-0.365**	0.125
Eve	Types	0.759**	0.697**	0.827**
	Exponent	0.474*	0.207	0.329
Sarah	Types	0.704**	0.302**	0.436**
	Exponent	0.008	-0.306**	0.250*

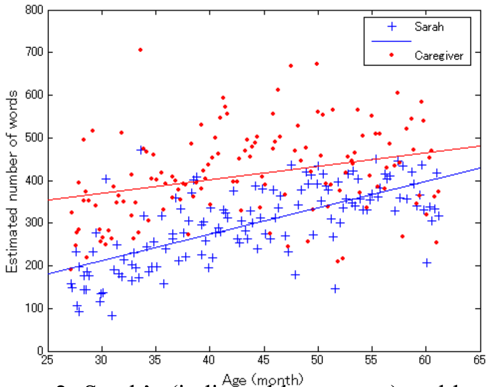

Figure 2: Sarah's (indicated by crosses) and her caregiver's (indicated by dots) number of latent types estimated by the model.

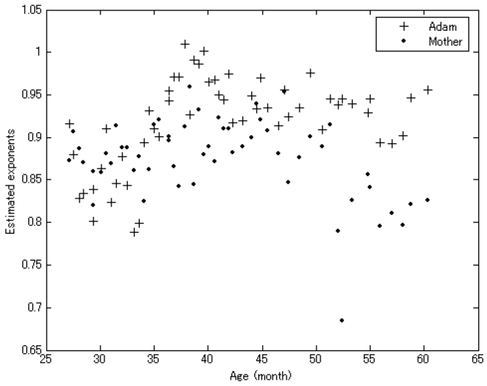

Figure 3: Adam's (indicated by crosses) and her caregiver's (indicated by dots) exponents estimated by the model.

General Discussion

In the present study, we propose a new method for estimation of the number of latent types and property of frequency distribution. Both Monte Carlo simulation and application to an empirical data suggest robustness of the type sampling method compared to its alternatives. The type sampling method may be applicable to multiple corpora with smaller sizes of tokens which are expected in language developmental studies. Typically, the number of words acquired by a child has been examined by parental reports (Fenson et al., 1993) or longitudinal diaries. These methods have their own advantage and disadvantage. The check list is easy but underestimates the actual number of words (Robinson, 1999). On the other hand, the diary study is costly but more accurate. Our proposal may be the third alternative in between the two methods.

Acknowledgments

This study was supported by grants from NIH MH60200.

References

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard, University Press.

Fenson, L., Dale, P., Reznick, J. S., Bate, E., Hartung, J., Pethick, S., et al. (1993). *Macarthur communicative development inventories*. San Diego: CA: Singular Publishing.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40* (3-4), 237-264.

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. Hague, Netherlands: Mouton & Co.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19* (1), 57-84.

Johnson, R. (1979). Measures of vocabulary diversity. In F. E. K. D. E. Ager & M. W. A. Smith (Eds.), (chap. Advances in Computer-aided Literary and Linguistic Research).

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, *17* (2), 457-472.

Malvern, D. D., & Richard, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language: papers from the annual meeting of the British association for applied linguistics*.

Malvern, D. D., & Richard, B. J. (2000). A new method of measuring lexical diversity in texts and conversations. *TEANGA*, *19* , 1-12.

Malvern, D. D., & Richard, B. J. (2002). Investing accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19* (2), 85-104.

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

Sichel, H. S. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist*, *11* , 45-72.

Weitzman, M. (1971). How useful is the logarithmic type-token ratio? *Journal of Linguistics*, *7* , 237-243.

Zipf, G. K. (1949). *Human behavior and the principle of least-effort*. Cambridge, MA: Addison-Wesley.