Knowing the Unseen: Estimating Vocabulary Size over Unseen Samples

Suma Bhat
Department of ECE
University of Illinois
spbhat2@illinois.edu

Richard Sproat
Center for Spoken Language Understanding
Oregon Health & Science University
rws@xoba.com

Abstract

Empirical studies on corpora involve making measurements of several quantities for the purpose of comparing corpora, creating language models or to make generalizations about specific linguistic phenomena in a language. Quantities such as average word length are stable across sample sizes and hence can be reliably estimated from large enough samples. However, quantities such as *vocabulary size* change with sample size. Thus measurements based on a given sample will need to be *extrapolated* to obtain their estimates over larger unseen samples. In this work, we propose a novel *nonparametric* estimator of vocabulary size. Our main result is to show the *statistical consistency* of the estimator – the first of its kind in the literature. Finally, we compare our proposal with the state of the art estimators (both parametric and nonparametric) on large standard corpora; apart from showing the favorable performance of our estimator, we also see that the classical Good-Turing estimator consistently underestimates the vocabulary size.

1 Introduction

Empirical studies on corpora involve making measurements of several quantities for the purpose of comparing corpora, creating language models or to make generalizations about specific linguistic phenomena in a language. Quantities such as average word length or average sentence length are stable across sample sizes. Hence empirical measurements from large enough samples tend to be reliable for even larger sample sizes. On the other hand, quantities associated with word frequencies, such as the number of *hapax legomena* or the number of distinct word types changes are strictly sample size dependent. Given a sample we can obtain the seen vocabulary and the seen number of *hapax legomena*. However, for the purpose of comparison of corpora of different sizes or linguistic phenomena based on samples of different sizes it is imperative that these quantities be compared based on similar sample sizes. We thus need methods to extrapolate empirical measurements of these quantities to arbitrary sample sizes.

Our focus in this study will be estimators of vocabulary size for samples larger than the sample available. There is an abundance of estimators of population size (in our case, vocabulary size) in existing literature. Excellent survey articles that summarize the state-of-the-art are available in (Bunge and Fitzpatrick, 1993) and (Gandolfi and Sastri, 2004). Of particular interest to us is the set of estimators that have been shown to model word frequency distributions well. This study proposes a nonparametric estimator of vocabulary size and evaluates its theoretical and empirical performance. For comparison we consider some state-of-the-art parametric and nonparametric estimators of vocabulary size.

The proposed non-parametric estimator for the number of unseen elements assumes a regime characterizing word frequency distributions. This work is motivated by a scaling formulation to address the problem of unlikely events proposed in (Baayen, 2001; Khmaladze, 1987; Khmaladze and Chitashvili, 1989; Wagner et al., 2006). We also demonstrate that the estimator is strongly consistent under the natural scaling formulation. While compared with other vocabulary size estimates, we see that our estimator performs at least as well as some of the state of the art estimators.

2 Previous Work

Many estimators of vocabulary size are available in the literature and a comparison of several non

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 109–117,
Suntec, Singapore, 2-7 August 2009.

parametric estimators of population size occurs in (Gandolfi and Sastri, 2004). While a definite comparison including parametric estimators is lacking, there is also no known work comparing methods of extrapolation of vocabulary size. Baroni and Evert, in (Baroni and Evert, 2005), evaluate the performance of some estimators in extrapolating vocabulary size for arbitrary sample sizes but limit the study to parametric estimators. Since we consider both parametric and nonparametric estimators here, we consider this to be the first study comparing a set of estimators for extrapolating vocabulary size.

Estimators of vocabulary size that we compare can be broadly classified into two types:

1. *Nonparametric estimators*- here word frequency information from the given sample alone is used to estimate the vocabulary size. A good survey of the state of the art is available in (Gandolfi and Sastri, 2004). In this paper, we compare our proposed estimator with the canonical estimators available in (Gandolfi and Sastri, 2004).

2. *Parametric estimators*- here a probabilistic model capturing the relation between expected vocabulary size and sample size is the estimator. Given a sample of size n, the sample serves to calculate the parameters of the model. The expected vocabulary for a given sample size is then determined using the explicit relation. The parametric estimators considered in this study are (Baayen, 2001; Baroni and Evert, 2005),

   (a) Zipf-Mandelbrot estimator (ZM);
   (b) finite Zipf-Mandelbrot estimator (fZM).

In addition to the above estimators we consider a novel non parametric estimator. It is the nonparametric estimator that we propose, taking into account the characteristic feature of word frequency distributions, to which we will turn next.

3 Novel Estimator of Vocabulary size

We observe $(X_1, \ldots, X_n)$, an i.i.d. sequence drawn according to a probability distribution $\mathbb{P}$ from a large, but finite, vocabulary Ω. Our goal is in estimating the "essential" size of the vocabulary Ω using only the observations. In other words, having seen a sample of size n we wish to know, given another sample from the same population, how many unseen elements we would expect to see. Our nonparametric estimator for the number of unseen elements is motivated by the characteristic property of word frequency distributions, the *Large Number of Rare Events* (LNRE) (Baayen, 2001). We also demonstrate that the estimator is strongly consistent under a natural scaling formulation described in (Khmaladze, 1987).

3.1 A Scaling Formulation

Our main interest is in probability distributions $\mathbb{P}$ with the property that a large number of words in the vocabulary Ω are unlikely, i.e., the chance any word appears eventually in an arbitrarily long observation is strictly between 0 and 1. The authors in (Baayen, 2001; Khmaladze and Chitashvili, 1989; Wagner et al., 2006) propose a natural scaling formulation to study this problem; specifically, (Baayen, 2001) has a tutorial-like summary of the theoretical work in (Khmaladze, 1987; Khmaladze and Chitashvili, 1989). In particular, the authors consider a *sequence* of vocabulary sets and probability distributions, indexed by the observation size n. Specifically, the observation $(X_1, \ldots, X_n)$ is drawn i.i.d. from a vocabulary Ω_n according to probability $\mathbb{P}_n$. If the probability of a word, say $\omega \in \Omega_n$ is p, then the probability that this specific word ω does not occur in an observation of size n is

$$(1-p)^n .$$

For ω to be an unlikely word, we would like this probability for large n to remain strictly between 0 and 1. This implies that

$$\frac{\check{c}}{n} \leq p \leq \frac{\hat{c}}{n}, \tag{1}$$

for some strictly positive constants $0 < \check{c} < \hat{c} < \infty$. We will assume throughout this paper that $\check{c}$ and $\hat{c}$ are the same for every word $\omega \in \Omega_n$. This implies that the vocabulary size is growing *linearly* with the observation size:

$$\frac{n}{\hat{c}} \leq |\Omega_n| \leq \frac{n}{\check{c}}.$$

This model is called the *LNRE zone* and its applicability in natural language corpora is studied in detail in (Baayen, 2001).

3.2 Shadows

Consider the observation string $(X_1, \ldots, X_n)$ and let us denote the quantity of interest – the number

of word types in the vocabulary Ω_n that are not observed – by $\mathbb{O}_n$. This quantity is random since the observation string itself is. However, we note that the distribution of $\mathbb{O}_n$ is unaffected if one relabels the words in Ω_n. This motivates studying of the probabilities assigned by $\mathbb{P}_n$ without reference to the labeling of the word; this is done in (Khmaladze and Chitashvili, 1989) via the *structural distribution function* and in (Wagner et al., 2006) via the *shadow*. Here we focus on the latter description:

Definition 1 *Let X_n be a random variable on Ω_n with distribution $\mathbb{P}_n$. The* shadow *of $\mathbb{P}_n$ is defined to be the distribution of the random variable $\mathbb{P}_n(\{X_n\})$.*

For the finite vocabulary situation we are considering, specifying the shadow is *exactly equivalent* to specifying the unordered components of $\mathbb{P}_n$, viewed as a probability vector.

3.3 Scaled Shadows Converge

We will follow (Wagner et al., 2006) and suppose that the scaled shadows, the distribution of $n \cdot \mathbb{P}_n(X_n)$, denoted by Q_n converge to a distribution Q. As an example, if $\mathbb{P}_n$ is a uniform distribution over a vocabulary of size cn, then $n \cdot \mathbb{P}_n(X_n)$ equals $\frac{1}{c}$ almost surely for each n (and hence it converges in distribution). From this convergence assumption we can, further, infer the following:

1. Since the probability of each word ω is lower and upper bounded as in Equation (1), we know that the distribution Q_n is non-zero only in the range $[\check{c}, \hat{c}]$.

2. The "essential" size of the vocabulary, i.e., the number of words of Ω_n on which $\mathbb{P}_n$ puts non-zero probability can be evaluated directly from the scaled shadow, scaled by $\frac{1}{n}$ as

$$\int_{\check{c}}^{\hat{c}} \frac{1}{y}\, dQ_n(y). \tag{2}$$

   Using the dominated convergence theorem, we can conclude that the convergence of the scaled shadows guarantees that the size of the vocabulary, scaled by $1/n$, converges as well:

$$\frac{|\Omega_n|}{n} \to \int_{\check{c}}^{\hat{c}} \frac{1}{y}\, dQ(y). \tag{3}$$

3.4 Profiles and their Limits

Our goal in this paper is to estimate the size of the underlying vocabulary, i.e., the expression in (2),

$$\int_{\check{c}}^{\hat{c}} \frac{n}{y}\, dQ_n(y), \tag{4}$$

from the observations $(X_1, \ldots, X_n)$. We observe that since the scaled shadow Q_n does not depend on the labeling of the words in Ω_n, a *sufficient statistic* to estimate (4) from the observation $(X_1, \ldots, X_n)$ is the *profile* of the observation: $(\varphi_1^n, \ldots, \varphi_n^n)$, defined as follows. φ_k^n is the number of word types that appear exactly k times in the observation, for $k = 1, \ldots, n$. Observe that

$$\sum_{k=1}^{n} k\varphi_k^n = n,$$

and that

$$V \stackrel{\text{def}}{=} \sum_{k=1}^{n} \varphi_k^n \tag{5}$$

is the number of *observed* words. Thus, the object of our interest is,

$$\mathbb{O}_n = |\Omega_n| - V. \tag{6}$$

3.5 Convergence of Scaled Profiles

One of the main results of (Wagner et al., 2006) is that the scaled profiles converge to a deterministic probability vector under the scaling model introduced in Section 3.3. Specifically, we have from Proposition 1 of (Wagner et al., 2006):

$$\sum_{k=1}^{n} \left| \frac{k\varphi_k}{n} - \lambda_{k-1} \right| \longrightarrow 0, \quad \text{almost surely,} \tag{7}$$

where

$$\lambda_k := \int_{\check{c}}^{\check{c}} \frac{y^k \exp(-y)}{k!}\, dQ(y) \quad k = 0, 1, 2, \ldots. \tag{8}$$

This convergence result suggests a natural estimator for $\mathbb{O}_n$, expressed in Equation (6).

3.6 A Consistent Estimator of $\mathbb{O}_n$

We start with the limiting expression for scaled profiles in Equation (7) and come up with a natural estimator for $\mathbb{O}_n$. Our development leading to the estimator is somewhat heuristic and is aimed at motivating the structure of the estimator for the number of unseen words, $\mathbb{O}_n$. We formally state and prove its consistency at the end of this section.

3.6.1 A Heuristic Derivation

Starting from (7), let us first make the approximation that

$$\frac{k\varphi_k}{n} \approx \lambda_{k-1}, \quad k = 1, \ldots, n. \tag{9}$$

We now have the formal calculation

$$\begin{aligned}
\sum_{k=1}^{n} \frac{\varphi_k^n}{n} &\approx \sum_{k=1}^{n} \frac{\lambda_{k-1}}{k} && (10)\\
&= \sum_{k=1}^{n} \int_{\check{c}}^{\hat{c}} \frac{e^{-y} y^{k-1}}{k!}\, dQ(y)\\
&\approx \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \left(\sum_{k=1}^{n} \frac{y^k}{k!} \right) dQ(y) && (11)\\
&\approx \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \left(e^y - 1\right)\, dQ(y) && (12)\\
&\approx \frac{|\Omega_n|}{n} - \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y}\, dQ(y). && (13)
\end{aligned}$$

Here the approximation in Equation (10) follows from the approximation in Equation (9), the approximation in Equation (11) involves swapping the outer discrete summation with integration and is justified formally later in the section, the approximation in Equation (12) follows because

$$\sum_{k=1}^{n} \frac{y^k}{k!} \to e^y - 1,$$

as $n \to \infty$, and the approximation in Equation (13) is justified from the convergence in Equation (3). Now, comparing Equation (13) with Equation (6), we arrive at an approximation for our quantity of interest:

$$\frac{\mathbb{O}_n}{n} \approx \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y}\, dQ(y). \tag{14}$$

The geometric series allows us to write

$$\frac{1}{y} = \frac{1}{\hat{c}} \sum_{\ell=0}^{\infty} \left(1 - \frac{y}{\hat{c}}\right)^{\ell}, \quad \forall y \in (0, \hat{c})\,. \tag{15}$$

Approximating this infinite series by a finite summation, we have for all $y \in (\check{c}, \hat{c})$,

$$\begin{aligned}
\frac{1}{y} - \frac{1}{\hat{c}} \sum_{\ell=0}^{M} \left(1 - \frac{y}{\hat{c}}\right)^{\ell} &= \frac{\left(1 - \frac{y}{\hat{c}}\right)^M}{y}\\
&\leq \frac{\left(1 - \frac{\check{c}}{\hat{c}}\right)^M}{\check{c}}. && (16)
\end{aligned}$$

It helps to write the truncated geometric series as a power series in y:

$$\begin{aligned}
&\frac{1}{\hat{c}} \sum_{\ell=0}^{M} \left(1 - \frac{y}{\hat{c}}\right)^{\ell}\\
&= \frac{1}{\hat{c}} \sum_{\ell=0}^{M} \sum_{k=0}^{\ell} \binom{\ell}{k} (-1)^k \left(\frac{y}{\hat{c}}\right)^k\\
&= \frac{1}{\hat{c}} \sum_{k=0}^{M} \left(\sum_{\ell=k}^{M} \binom{\ell}{k} \right) (-1)^k \left(\frac{y}{\hat{c}}\right)^k\\
&= \sum_{k=0}^{M} (-1)^k\, a_k^M\, y^k, && (17)
\end{aligned}$$

where we have written

$$a_k^M := \frac{1}{\hat{c}^{k+1}} \left(\sum_{\ell=k}^{M} \binom{\ell}{k} \right).$$

Substituting the finite summation approximation in Equation 16 and its power series expression in Equation (17) into Equation (14) and swapping the discrete summation with the integral, we can continue

$$\begin{aligned}
\frac{\mathbb{O}_n}{n} &\approx \sum_{k=0}^{M} (-1)^k\, a_k^M \int_{\check{c}}^{\hat{c}} e^{-y} y^k\, dQ(y)\\
&= \sum_{k=0}^{M} (-1)^k\, a_k^M\, k! \lambda_k. && (18)
\end{aligned}$$

Here, in Equation (18), we used the definition of λ_k from Equation (8). From the convergence in Equation (7), we finally arrive at our estimate:

$$\mathbb{O}_n \approx \sum_{k=0}^{M} (-1)^k\, a_k^M\, (k+1)!\ \varphi_{k+1}. \tag{19}$$

3.6.2 Consistency

Our main result is the demonstration of the consistency of the estimator in Equation (19).

Theorem 1 *For any $\epsilon > 0$,*

$$\lim_{n \to \infty} \frac{\left| \mathbb{O}_n - \sum_{k=0}^{M} (-1)^k\, a_k^M\, (k+1)!\, \varphi_{k+1} \right|}{n} \leq \epsilon$$

almost surely, as long as

$$M \geq \frac{\check{c} \log_2 e + \log_2 (\epsilon \check{c})}{\log_2 (\hat{c} - \check{c}) - 1 - \log_2 (\hat{c})}. \tag{20}$$

Proof: From Equation (6), we have

$$\begin{aligned}\frac{\mathbb{O}_n}{n} &= \frac{|\Omega_n|}{n} - \sum_{k=1}^{n} \frac{\varphi_k}{n} \\ &= \frac{|\Omega_n|}{n} - \sum_{k=1}^{n} \frac{\lambda_{k-1}}{k} - \\ &\quad \sum_{k=1}^{n} \frac{1}{k} \left(\frac{k\varphi_k}{n} - \lambda_{k-1} \right). \quad (21)\end{aligned}$$

The first term in the right hand side (RHS) of Equation (21) converges as seen in Equation (3). The third term in the RHS of Equation (21) converges to zero, almost surely, as seen from Equation (7). The second term in the RHS of Equation (21), on the other hand,

$$\begin{aligned}\sum_{k=1}^{n} \frac{\lambda_{k-1}}{k} &= \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \left(\sum_{k=1}^{n} \frac{y^k}{k!} \right) \, dQ(y) \\ &\rightarrow \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \left(e^y - 1 \right) \, dQ(y), n \rightarrow \infty, \\ &= \int_{\check{c}}^{\hat{c}} \frac{1}{y} \, dQ(y) - \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \, dQ(y).\end{aligned}$$

The monotone convergence theorem justifies the convergence in the second step above. Thus we conclude that

$$\lim_{n \rightarrow \infty} \frac{\mathbb{O}_n}{n} = \int_{\check{c}}^{\hat{c}} \frac{e^{-y}}{y} \, dQ(y) \quad (22)$$

almost surely. Coming to the estimator, we can write it as the sum of two terms:

$$\begin{aligned}&\sum_{k=0}^{M} (-1)^k \, a_k^M \, k! \lambda_k \quad (23) \\ &\quad + \sum_{k=0}^{M} (-1)^k \, a_k^M \, k! \left(\frac{(k+1)\, \varphi_{k+1}}{n} - \lambda_k \right).\end{aligned}$$

The second term in Equation (23) above is seen to converge to zero almost surely as $n \rightarrow \infty$, using Equation (7) and noting that M is a constant not depending on n. The first term in Equation (23) can be written as, using the definition of λ_k from Equation (8),

$$\int_{\check{c}}^{\hat{c}} e^{-y} \left(\sum_{k=0}^{M} (-1)^k \, a_k^M y^k \right) \, dQ(y). \quad (24)$$

Combining Equations (22) and (24), we have that, almost surely,

$$\begin{aligned}&\lim_{n \rightarrow \infty} \frac{\mathbb{O}_n - \sum_{k=0}^{M} (-1)^k \, a_k^M \, (k+1)! \, \varphi_{k+1}}{n} = \\ &\int_{\check{c}}^{\hat{c}} e^{-y} \left(\frac{1}{y} - \sum_{k=0}^{M} (-1)^k \, a_k^M y^k \right) \, dQ(y). \quad (25)\end{aligned}$$

Combining Equation (16) with Equation (17), we have

$$0 < \frac{1}{y} - \sum_{k=0}^{M} (-1)^k \, a_k^M y^k \leq \frac{\left(1 - \frac{\check{c}}{\hat{c}}\right)^M}{\check{c}}. \quad (26)$$

The quantity in Equation (25) can now be upper bounded by, using Equation (26),

$$\frac{e^{-\check{c}} \left(1 - \frac{\check{c}}{\hat{c}}\right)^M}{\check{c}}.$$

For M that satisfy Equation (20) this term is less than ϵ. The proof concludes.

3.7 Uniform Consistent Estimation

One of the main issues with actually employing the estimator for the number of unseen elements (cf. Equation (19)) is that it involves knowing the parameter $\hat{c}$. In practice, there is no natural way to obtain any estimate on this parameter $\hat{c}$. It would be most useful if there were a way to modify the estimator in a way that it does not depend on the unobservable quantity $\hat{c}$. In this section we see that such a modification is possible, while still retaining the main theoretical performance result of consistency (cf. Theorem 1).

The first step to see the modification is in observing where the need for $\hat{c}$ arises: it is in writing the geometric series for the function $\frac{1}{y}$ (cf. Equations (15) and (16)). If we could let $\hat{c}$ along with the number of elements M itself depend on the sample size n, then we could still have the geometric series formula. More precisely, we have

$$\begin{aligned}\frac{1}{y} - \frac{1}{\hat{c}_n} \sum_{\ell=0}^{M_n} \left(1 - \frac{y}{\hat{c}_n} \right)^{\ell} &= \frac{1}{y} \left(1 - \frac{y}{\hat{c}_n} \right)^{M_n} \\ &\rightarrow 0, \quad n \rightarrow \infty,\end{aligned}$$

as long as

$$\frac{\hat{c}_n}{M_n} \rightarrow 0, \quad n \rightarrow \infty. \quad (27)$$

This simple calculation suggests that we can replace $\hat{c}$ and M in the formula for the estimator (cf. Equation (19)) by terms that depend on n and satisfy the condition expressed by Equation (27).

4 Experiments

4.1 Corpora

In our experiments we used the following corpora:

1. The *British National Corpus* (BNC): A corpus of about 100 million words of written and spoken British English from the years 1975-1994.

2. The *New York Times Corpus* (NYT): A corpus of about 5 million words.

3. The *Malayalam Corpus* (MAL): A collection of about 2.5 million words from varied articles in the Malayalam language from the Central Institute of Indian Languages.

4. The *Hindi Corpus* (HIN): A collection of about 3 million words from varied articles in the Hindi language also from the Central Institute of Indian Languages.

4.2 Methodology

We would like to see how well our estimator performs in terms of estimating the number of unseen elements. A natural way to study this is to expose only half of an existing corpus to be observed and estimate the number of unseen elements (assuming the the actual corpus is twice the observed size). We can then check numerically how well our estimator performs with respect to the "true" value. We use a subset (the first 10%, 20%, 30%, 40% and 50%) of the corpus as the *observed sample* to estimate the vocabulary over twice the sample size. The following estimators have been compared.

Nonparametric: Along with our proposed estimator (in Section 3), the following canonical estimators available in (Gandolfi and Sastri, 2004) and (Baayen, 2001) are studied.

1. Our proposed estimator $\mathbb{O}_n$ (cf. Section 3): since the estimator is rather involved we consider only small values of M (we see empirically that the estimator converges for very small values of M itself) and choose $\hat{c} = M$. This allows our estimator for the number of unseen elements to be of the following form, for different values of M:

M	$\mathbb{O}_n$
1	$2(\varphi_1 - \varphi_2)$
2	$\frac{3}{2}(\varphi_1 - \varphi_2) + \frac{3}{4}\varphi_3$
3	$\frac{4}{3}(\varphi_1 - \varphi_2) + \frac{8}{9}\left(\varphi_3 - \frac{\varphi_4}{3}\right)$

Using this, the estimator of the true vocabulary size is simply,

$$\mathbb{O}_n + V. \tag{28}$$

Here (cf. Equation (5))

$$V = \sum_{k=1}^{n} \varphi_k^n. \tag{29}$$

In the simulations below, we have considered M large enough until we see numerical convergence of the estimators: in all the cases, no more than a value of 4 is needed for M. For the English corpora, very small values of M suffice – in particular, we have considered the average of the first three different estimators (corresponding to the first three values of M). For the non-English corpora, we have needed to consider $M = 4$.

2. Gandolfi-Sastri estimator,

$$V_{\mathrm{GS}} \stackrel{\mathrm{def}}{=} \frac{n}{n - \varphi_1}\left(V + \varphi_1\gamma^2\right), \tag{30}$$

where

$$\gamma^2 = \frac{\varphi_1 - n - V}{2n} + \frac{\sqrt{5n^2 + 2n(V - 3\varphi_1) + (V - \varphi_1)^2}}{2n};$$

3. Chao estimator,

$$V_{\mathrm{Chao}} \stackrel{\mathrm{def}}{=} V + \frac{\varphi_1^2}{2\varphi_2}; \tag{31}$$

4. Good-Turing estimator,

$$V_{\mathrm{GT}} \stackrel{\mathrm{def}}{=} \frac{V}{\left(1 - \frac{\varphi_1}{n}\right)}; \tag{32}$$

5. "Simplistic" estimator,

$$V_{\mathrm{Smpl}} \stackrel{\mathrm{def}}{=} V\left(\frac{n_{\mathrm{new}}}{n}\right); \tag{33}$$

here the supposition is that the vocabulary size scales linearly with the sample size (here n_{new} is the new sample size);

6. Baayen estimator,

$$V_{\mathrm{Byn}} \stackrel{\mathrm{def}}{=} V + \left(\frac{\varphi_1}{n}\right) n_{new}; \tag{34}$$

here the supposition is that the vocabulary growth rate at the observed sample size is given by the ratio of the number of *hapax legomena* to the sample size (cf. (Baayen, 2001) pp. 50).

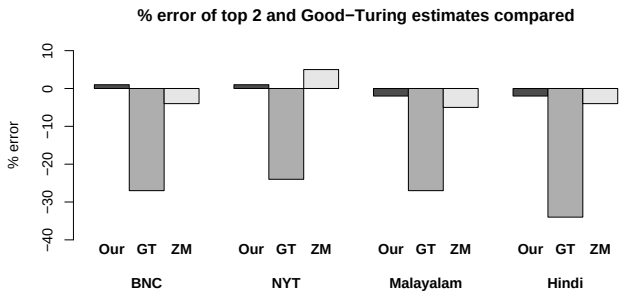

Figure 1: Comparison of error estimates of the 2 best estimators-ours and the ZM, with the Good-Turing estimator using 10% sample size of all the corpora. A bar with a positive height indicates and overestimate and that with a negative height indicates and underestimate. Our estimator *outperforms* ZM. Good-Turing estimator widely *underestimates* vocabulary size.

Parametric: Parametric estimators use the observations to first estimate the parameters. Then the corresponding models are used to estimate the vocabulary size over the larger sample. Thus the frequency spectra of the observations are only *indirectly* used in extrapolating the vocabulary size. In this study we consider state of the art parametric estimators, as surveyed by (Baroni and Evert, 2005). We are aided in this study by the availability of the implementations provided by the `ZipfR` package and their default settings.

5 Results and Discussion

The performance of the different estimators as percentage errors of the true vocabulary size using different corpora are tabulated in tables 1-4. We now summarize some important observations.

- From the Figure 1, we see that our estimator compares quite favorably with the best of the state of the art estimators. The best of the state of the art estimator is a parametric one (ZM), while ours is a nonparametric estimator.
- In table 1 and table 2 we see that our estimate is quite close to the true vocabulary, at all sample sizes. Further, it compares very favorably to the state of the art estimators (both parametric and nonparametric).
- Again, on the two non-English corpora (tables 3 and 4) we see that our estimator compares favorably with the best estimator of vocabulary size and at some sample sizes even surpasses it.
- Our estimator has theoretical performance guarantees and its empirical performance is comparable to that of the state of the art estimators. However, this performance comes at a very small fraction of the computational cost of the parametric estimators.
- The state of the art nonparametric Good-Turing estimator wildly underestimates the vocabulary; this is true in each of the four corpora studied and at all sample sizes.

6 Conclusion

In this paper, we have proposed a new nonparametric estimator of vocabulary size that takes into account the LNRE property of word frequency distributions and have shown that it is statistically consistent. We then compared the performance of the proposed estimator with that of the state of the art estimators on large corpora. While the performance of our estimator seems favorable, we also see that the widely used classical Good-Turing estimator consistently underestimates the vocabulary size. Although as yet untested, with its computational simplicity and favorable performance, our estimator may serve as a more reliable alternative to the Good-Turing estimator for estimating vocabulary sizes.

Acknowledgments

This research was partially supported by Award IIS-0623805 from the National Science Foundation.

References

R. H. Baayen. 2001. *Word Frequency Distributions*, Kluwer Academic Publishers.

Marco Baroni and Stefan Evert. 2001. “Testing the extrapolation quality of word frequency models”, *Proceedings of Corpus Linguistics , volume 1 of The Corpus Linguistics Conference Series*, P. Danielsson and M. Wagenmakers (eds.).

J. Bunge and M. Fitzpatrick. 1993. “Estimating the number of species: a review”, *Journal of the American Statistical Association*, Vol. 88(421), pp. 364-373.

Sample (% of corpus)	True value	% error w.r.t the true value							
		Our	GT	ZM	fZM	Smpl	Byn	Chao	GS
10	153912	1	-27	-4	-8	46	23	8	-11
20	220847	-3	-30	-9	-12	39	19	4	-15
30	265813	-2	-30	-9	-11	39	20	6	-15
40	310351	1	-29	-7	-9	42	23	9	-13
50	340890	2	-28	-6	-8	43	24	10	-12

Table 1: Comparison of estimates of vocabulary size for the **BNC corpus** as percentage errors w.r.t the true value. A negative value indicates an underestimate. Our estimator *outperforms* the other estimators at all sample sizes.

Sample (% of corpus)	True value	% error w.r.t the true value							
		Our	GT	ZM	fZM	Smpl	Byn	Chao	GS
10	37346	1	-24	5	-8	48	28	4	-8
20	51200	-3	-26	0	-11	46	22	-1	-11
30	60829	-2	-25	1	-10	48	23	1	-10
40	68774	-3	-25	0	-10	49	21	-1	-11
50	75526	-2	-25	0	-10	50	21	0	-10

Table 2: Comparison of estimates of vocabulary size for the **NYT corpus** as percentage errors w.r.t the true value. A negative value indicates an underestimate. Our estimator *compares favorably* with ZM and Chao.

Sample (% of corpus)	True value	% error w.r.t the true value							
		Our	GT	ZM	fZM	Smpl	Byn	Chao	GS
10	146547	-2	-27	-5	-10	9	34	82	-2
20	246723	8	-23	4	-2	19	47	105	5
30	339196	4	-27	0	-5	16	42	93	-1
40	422010	5	-28	1	-4	17	43	95	-1
50	500166	5	-28	1	-4	18	44	94	-2

Table 3: Comparison of estimates of vocabulary size for the **Malayalam corpus** as percentage errors w.r.t the true value. A negative value indicates an underestimate. Our estimator *compares favorably* with ZM and GS.

Sample (% of corpus)	True value	% error w.r.t the true value							
		Our	GT	ZM	fZM	Smpl	Byn	Chao	GS
10	47639	-2	-34	-4	-9	25	32	31	-12
20	71320	7	-30	2	-1	34	43	51	-7
30	93259	2	-33	-1	-5	30	38	42	-10
40	113186	0	-35	-5	-7	26	34	39	-13
50	131715	-1	-36	-6	-8	24	33	40	-14

Table 4: Comparison of estimates of vocabulary size for the **Hindi corpus** as percentage errors w.r.t the true value. A negative value indicates an underestimate. Our estimator *outperforms* the other estimators at certain sample sizes.

A. Gandolfi and C. C. A. Sastri. 2004. “Nonparametric Estimations about Species not Observed in a Random Sample”, *Milan Journal of Mathematics*, Vol. 72, pp. 81-105.

E. V. Khmaladze. 1987. “The statistical analysis of large number of rare events”, *Technical Report, Department of Mathematics and Statistics.*, CWI, Amsterdam, MS-R8804.

E. V. Khmaladze and R. J. Chitashvili. 1989. “Statistical analysis of large number of rate events and related problems”, *Probability theory and mathematical statistics* (Russian), Vol. 92, pp. 196-245.

. P. Santhanam, A. Orlitsky, and K. Viswanathan, “New tricks for old dogs: Large alphabet probability estimation”, in Proc. 2007 *IEEE Information Theory Workshop*, Sept. 2007, pp. 638–643.

A. B. Wagner, P. Viswanath and S. R. Kulkarni. 2006. “Strong Consistency of the Good-Turing estimator”, *IEEE Symposium on Information Theory*, 2006.