the intermittent silence process. "Journal of the American Society for Information Science and Technology", Abril 2009, vol. 60, núm. 4, p. 837-843, which has been published in final form at http://dx.doi.org/10.1002/asi.21033. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

# The frequency spectrum of finite samples from the intermittent silence process

**Ramon Ferrer-i-Cancho[1] & Ricard Gavaldà[1]**
(1) Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya. Campus Nord, Edifici Omega. Jordi Girona Salgado 1-3. 08034 Barcelona, Spain.
E-mail: {rferrericancho,gavalda}@lsi.upc.edu.

Submitted to the *Journal of the American Society for Information Science and Technology*

Please do not circulate

**It has been argued that the actual distribution of word frequencies could be reproduced or explained by generating a random sequence of letters and spaces according to the so-called intermittent silence process. The same kind of process could reproduce or explain the counts of other kinds of units from a wide range of disciplines. Taking the linguistic metaphor, we focus on the frequency spectrum, i.e. the number of words with a certain frequency, and the vocabulary size, i.e. the number of different words of text generated by an intermittent silence process. We derive and explain how to calculate accurately and efficiently the expected frequency spectrum and the expected vocabulary size as a function of the text size.**

## I. Introduction

In a seminal work, Benoît Mandelbrot put forward a theory of word frequencies (Mandelbrot 1951, 1953). A product of his seminal work is the following simple stochastic process. Consider that you generate random words by choosing characters at random from an alphabet made of $N$ letters plus a special character indicating the end of a word (e.g., a space). A popular version (Miller & Chomsky 1963) assumes that letters are equally likely and that the special character has probability $\sigma$ (hence a specific letter has probability $(1-\sigma)/N$). We will refer to this kind of process as the intermittent silence process (ISP), borrowing the term "intermittent silence" from Miller (1957). Although other terms are used for referring to this process we believe that they are not accurate enough. For instance, Li (1992) uses the term random text but random texts can be generated in many ways, not necessarily through his ISP. For instance, one could generate a random text reproducing the long range correlation of real writings using the model by Lacalle *et al.* 2006 (notice that Li's ISP generates a sequence of independent words). Furthermore, Li's ISP generates words by concatenating characters while Lacalle et al.'s (2006) picks already existing words.

Here we aim to study the frequency of words produced by Miller & Chomsky's (1963) ISP, in which letters are equally likely. Although the case of unequal letter probabilities has been considered in the literature (Li 1992, Cohen *et al.* 1997, Ferrer i Cancho & Solé 2002, Wolfram 2002), it is not the focus of our article. In general, the empirical frequency of elements (e.g., words) can be studied by means of the rank distribution, i.e. the relationship between the absolute or relative frequency of a word and its rank (a word has rank $i$ if it is the $i$-th most frequent word of a text) or the frequency spectrum, i.e. the number or the proportion of words of a given frequency

within a given text (Tuldava 1996). Whether actual word frequency counts can be explained by the ISP is the subject of a long controversy concerning the linguistic relevance or utility of word frequency counts (Miller 1957, Miller & Chomsky 1963, Li 1992; Ferrer i Cancho & Solé 2002, Suzuki et al. 2005, McCowan *et al.* 2005; Ferrer i Cancho 2005). We aim to provide some fundamental results for future rigorous comparisons of the ISP versus real words and other situations where the ISP may apply, such as frequency counts from animal communication (McCowan *et al.* 1999) or DNA (Furusawa & Kaneko 2003). The reader should not interpret that the scope of the problem is restricted to communicative units like words. Unit frequency counts and their explanation is the subject of disciplines as different as information science (Bailón-Moreno *et al.* 2005, Egghe 1998), quantitative biology (Gisiger 2001) or network theory (Albert & Barabási 2002). Given a certain frequency count, an ISP is always a possibility, as the process can be formulated in an abstract way by replacing letters with any convenient subunit. This is the spirit of Suzuki *et al.* (2005), who replace letters with faces of a die to make the ISP abstract and general. Hereafter the units of word length are letters.

Miller & Chomsky (1963), restating what Mandelbrot (1953) had calculated previously, showed that, $<i>$, the mean rank of words of the same length generated by an ISP, obeys

$$p(<i>) \sim (b+<i>)^{-a}, \tag{1}$$

where $p(<i>)$ is the probability of the mean rank $<i>$,

$$a = 1 - \frac{\log(1-\sigma)}{\log N}, \tag{2}$$

and

$$b = \frac{N+3}{2(N+1)}. \tag{3}$$

We define $T$ as the text length or the sample size in words. The derivation of Eq. 1 is made assuming that all words of the same length appear at least once in the text, which is only true for any length in the limit of large $T$. Furthermore, the derivation is rough, as it concerns the probability of the mean rank over words of the same length and not the probability of an individual rank. Notice that that it is customary to use Eq. 1 as if it had been derived for individual ranks in different disciplines as information sciences (Bailón-Moreno *et al.* 2005, Egghe 1998) and computational linguistics (Manning & Schütze 1999). In contrast, here we aim to derive the exact expected frequency histogram of words of finite (and not necessarily very large) samples from the point of view of the frequency spectrum.

In the literature, there is no agreement on the minimum word length $L_0$ of the words that an ISP generates. For instance, Li (1992) excludes empty words, i.e. $L_0$=1, whereas Miller & Chomsky (1963) assume empty words (or consecutive spaces), i.e. $L_0$=0. We believe that $L_0$=1 is more reasonable and realistic for human language, but we will embrace all the possibilities with a generalized ISP with three parameters, i.e. $N$, $\sigma$ and $L_0$.

The main goal of this article is deriving the expected frequency spectrum of our three-parameter ISP, i.e. the expected value of $n(f|T)$, the number of words produced by an ISP that occur $f$ times knowing that the text has length $T>0$, with $f\in[1,T]$. In what follows, we assume that $N$ is a strictly positive natural number, $\sigma\in(0,1)$ and $L_0\geq 0$.

## II. Analytical derivation of the frequency spectrum

We define $l(w)$ as the length of the word $w$. Knowing that there are $N^L$ words of length $L$, we can write $n(f|T)$ as

$$n(f\mid T)=\sum_{L=L_0}^{\infty}\sum_{l(w)=L} I(f\mid w,L,T), \tag{4}$$

where $I(f|w,L,T)$ is a Bernoulli variable indicating if the word $w$ has appeared $f$ times knowing that $l(w)=L$ and the text produced by an ISP has length $T$ ($I(f|w,L,T)=1$) or not ($I(f|w,L,T)=0$). Hence, the expected value of $n(f|T)$ is

$$E[n(f\mid T)]=\sum_{L=L_0}^{\infty}\sum_{l(w)=L} E[I(f\mid w,L,T)]. \tag{5}$$

Being $I(f,w,L,T)$ a Bernoulli variable, we have $E[I(f|w,L,T)]=p(f|w,L,T)$, the probability that a word $w$ is produced $f$ times by and ISP knowing that $l(w)=L$ and that the text has length $T$. Since there are $N^L$ words of length $L$, we have

$$E[n(f\mid T)]=\sum_{L=L_0}^{\infty} N^L p(f\mid w,L,T). \tag{6}$$

Now we aim to derive, $p(f|w,L,T)$. On the one hand, the length of a word produced by an ISP is geometrically distributed, i.e.

$$p(L)=(1-\sigma)^{L-L_0}\sigma, \tag{/}$$

with $L=L_0$, $L_0+1$, $L_0+2$,… When $L_0=0$, Eq. 7 defines the typical geometric distribution (Wimmer & Altmann 1999) while it defines a shifted or displaced geometrical distribution when $L_0=1$. On the other hand, there are $N^L$ words of length $L$ and the probability that an ISP produces a word $w$ which has length $L$ is

$$p(w,L)=p(w\mid L)p(L), \tag{8}$$

where $p(w|L)$ is the probability that an ISP produces $w$ knowing that it has length $L$. Being all letters equally likely, it follows that all words of the same length $L$ are equally likely, $p(w|L)=1/N^L$. This way, replacing Eq. 7 into Eq. 8 we finally obtain

$$p(w,L)=\frac{(1-\sigma)^{L-L_0}\sigma}{N^L}. \tag{9}$$

We define $p(w,L|T)$ as the probability that an ISP produces a word $w$ of length $L$ in a text of length $T$. We obviously have that $p(w,L|T)=p(w,L)$ due to independence. Therefore, the frequency of occurrence of a word $w$ which has length $L$ in a text of length $T$ is binomially distributed with parameters $T$ and $p(w,L)$, i.e.

$$p(f \mid w,L,T) = \binom{T}{f} p(w,L)^f \left(1 - p(w,L)\right)^{T-f}, \tag{10}$$

with $f \in [1,T]$. Finally, replacing Eq. 10 into Eq. 6 yields

$$E[n(f \mid T)] = \binom{T}{f} \sum_{L=L_0}^{\infty} N^L p(w,L)^f \left(1 - p(w,L)\right)^{T-f}. \tag{11}$$

We also want to study the vocabulary growth of the ISP with $T$. We define $n(T)$ as the number of different words produced by an ISP in a text of length $T$. Writing $n(T)$ as

$$n(T) = \sum_{f=1}^{T} n(f \mid T), \tag{12}$$

it becomes obvious that $n(T)$ is a statistic of the frequency spectrum. Using Eq. 12, the expected value of $n(T)$ becomes just simply

$$E[n(T)] = \sum_{f=1}^{T} E[n(f \mid T)]. \tag{13}$$

Replacing Eq. 11 into Eq. 13 and knowing that

$$\sum_{f=1}^{N} p(f,w \mid L,T) = 1 - p(0,w \mid L,T) = 1 - (1 - p(w,L))^T, \tag{14}$$

we finally obtain

$$E[n(T)] = \sum_{L=L_0}^{\infty} N^L (1 - (1 - p(w,L))^T). \tag{15}$$

Using the binomial expansion, Eqs. 11 and 15 can be rewritten as an exact finite summation (Appendix A)

$$E[n(f \mid T)] = \binom{T}{f} \sigma^f N^{L_0(1-f)} \sum_{i=0}^{T-f} \frac{(-1)^i}{1 - r(f+i)} \left(\frac{\sigma}{N^{L_0}}\right)^i \binom{T-f}{i} \tag{16}$$

and

$$E[n(T)] = \sum_{i=1}^{T} (-1)^{i+1} \frac{\sigma^i N^{L_0(1-i)}}{1 - r(i)} \binom{T}{i}. \tag{17}$$

## III. Numerical calculation of the frequency spectrum

In Appendix C, it is argued that it is convenient to calculate $E[n(f|T)]$ through Eq. 11 and $E[n(T)]$ through Eq. 13. Thus, the crux of the numerical calculation problem reduces to Eq. 11. Eq. 11 contains a summation from $L=L_0$ to $\infty$. In practice, the summation should be performed in a finite range, i.e. $L\in[L_{min},L_{max}]$ (with $L_{min}\geq L_0$) out of which the contribution of the terms of the summation might be neglected. We actually would like to calculate the relationship between $L_{max}$ and a desired error or actually (for simplicity) an upper bound of the desired error, i.e. $\gamma_{max}$ (e.g., $\gamma_{max}=10^{-10}$), when neglecting the contributions of the terms of lengths above $L_{max}$ in the calculation of $E[n(f|T)]$. Similarly, we would like to do the same for and upper bound of the desired error when neglecting the contribution of the terms below $L_{min.}$ This way, Eq. 11 gives an upper bound of the right error

$$\binom{T}{f}\sum_{L=L_{\max}+1}^{\infty}N^{L}p(w,L)^{f}\left(1-p(w,L)\right)^{T-f}\leq\gamma_{\max}^{+}(f\,|\,T)\,, \tag{18}$$

and the left error

$$\binom{T}{f}\sum_{L=L_{0}+1}^{L_{\min}-1}N^{L}p(w,L)^{f}\left(1-p(w,L)\right)^{T-f}\leq\gamma_{\max}^{-}(f\,|\,T)\,, \tag{19}$$

for $E[n(f|T)]$. In practice, we want to fix the desired maximum error $\gamma_{max}$ and determine $L_{max}$. It can be shown that the upper bound of the right error of $E[n(f|T)]$ gives (Appendix B)

$$L_{\max}=\frac{\log(\gamma_{\max}^{+}(f\,|\,T)G(f))}{\log r(f)}-1 \tag{20}$$

with

$$G(f)=\frac{1-r(f)}{\binom{T}{f}}\left(\frac{(1-\sigma)^{L_0}}{\sigma}\right)^{f}, \tag{21}$$

$$r(f)=N\left(\frac{1-\sigma}{N}\right)^{f}, \tag{22}$$

and $L_{max}+1-L_0\geq 0$ while an upper bound of the left error gives (Appendix B)

$$L_{\min}=\frac{\log(r^{L_0}(f)-\gamma_{\max}^{-}(f\,|\,T)G(f))}{\log r(f)}\,. \tag{23}$$

E[$n(T)$] can be calculated with maximum error $\gamma_{\max}(T)$ from Eq. 12. If we impose that $\gamma^{+}_{\max}(f\,|\,T)$ and $\gamma^{-}_{\max}(f\,|\,T)$ are the same for each frequency, the maximum errors of individual frequencies are related with $\gamma_{\max}(T)$ through

$$\gamma_{\max}(T) = T(\gamma^{+}_{\max}(1\,|\,T) + \gamma^{-}_{\max}(1\,|\,T)). \tag{24}$$

If we impose $\gamma^{-}_{\max}(1\,|\,T) = 0$ for simplicity on Eq. 24, we obtain the necessary maximum error of individual frequencies, i.e.

$$\gamma^{+}_{\max}(1\,|\,T) = \frac{\gamma_{\max}(T)}{T}. \tag{25}$$

Further technical remarks for calculating E[$n(f|T)$] and E[$n(T)$] with a computer are given in Appendix C.

## IV. Some numerical results

Fig. 1 shows E[$n(f|T)$] for increasingly larger values of $T$ in order of magnitude, with $N$=26 and $\sigma$=0.18, which according to Miller (1957) and Miller & Chomsky (1963) are representative of written English. Calculations are based on Eq. 11 with bounded length. The error of the finite interval numerical approximation does not exceed $10^{-40}$. To see it, notice that we obtained $L_{max}$ for each frequency $f$ from Eq. 20 ($\gamma^{+}_{\max}(f\,|\,T) = 10^{-40}$ for each $f$) and used $L_{min}$=$L_0$ for simplicity (thus $\gamma^{-}_{\max}(f\,|\,T) = 0$).

Each mode in the curves of Fig. 1 is mainly due to the contribution of words of the same length $L$ within the range of $L$ that is expected to be observed at least once in a text of length $T$. In Fig. 1, arrows are used to indicate the peaks of the different lengths. The gaps between modes can be explained by the fact that the probabilities of words of length $L$ are smaller than those of words of length $L$-1 ($L$>$L_0$) by a factor of (1-$\sigma$)/$N$. It is well-known from numerical experiments that a more gradual transition between the probabilities of words of different lengths (e.g., by not using equally likely letters) smoothes the frequency spectrum of the ISP (Ferrer i Cancho & Solé 2002, Cohen *et al.* 1997). Similarly, it is known that non-commensurate letter probabilities smooth the rank distribution of the ISP (Li 1992, Wolfram 2002).

Fig. 2 shows the practically linear growth of E[$n(T)$] in logarithmic scale with the same parameters as in Fig. 1. Calculations are based on Eq. 15 with bounded length. The maximum error of the finite interval numerical approximation does not exceed $10^{-35}$, i.e. $\gamma_{\max}(T) \le 10^{-40}$, in Fig. 2. To see it, consider that we calculated E[$n(T)$] through Eq. 13 and for each frequency $f$ and each $T$ we fixed the maximum error to $10^{-40}$ as in Fig. 1 ($\gamma^{+}_{\max}(f\,|\,T) = 10^{-40}$ and $\gamma^{-}_{\max}(f\,|\,T) = 0$ for each $f$ and each $T$). Thus, $\gamma_{\max}(T)$ obeys Eq. 25 in Fig. 2. Defining $T_{\max}$ as the maximum value of $T$, Eq. 25 gives

$$\gamma_{\max}(T) \le T_{\max}\gamma^{+}_{\max}(1\,|\,T). \tag{26}$$

Thus, replacing $T_{\max}$=$10^5$ from Fig. 2 and $\gamma^{+}_{\max}(1\,|\,T) = 10^{-40}$ into Eq. 26, we conclude that the error of E[$n(T)$] in Fig. 2 does not exceed $10^{-35}$.

## V. Discussion

In this article we have derived the frequency spectrum of the ISP and a particular aspect of this spectrum, i.e. the vocabulary growth as the text length increases. We have explained how the expected frequency spectrum and the expected vocabulary growth can be calculated efficiently and accurately with a computer. By doing so, we have provided the basis for evaluating the goodness of the fit of the ISP to empirical histograms (for instance, plots of the number of words with a certain frequency or plots of the authors with a certain a certain number of publications). Imagine that we want to evaluate the goodness of the fit of concrete parameters of the ISP. One possible way of proceeding could be the following three steps (Goldstein *et al.* 2004):

1. Calculating the deviation $\delta$ between the actual frequency histogram and the expected frequency spectrum for an ISP with these parameters.
2. Calculating the probability of obtaining a deviation larger or equal than $\delta$ (e.g., using a Monte Carlo procedure to estimate this probability).
3. If this probability is below a certain (low) significance level one concludes that it is unlikely that the histogram has been generated by an ISP. Otherwise, this possibility cannot be denied.

Our article is crucial for step 1.

Notice that our results are a turning point in the characterization of the distributions generated by the ISP and its applications (e.g., fitting). To see it, consider that using Eq. 1 to evaluate the fit of an ISP to a rank histogram is problematic because this equation:

- Does not define the relationship between a rank and its probability but the relationship between the mean rank (over words of the same length) and its probability. Therefore evaluating the fit of the ISP using Eq. 1 would lack precision.
- It assumes that all words of a certain length have appeared in a text of certain length, while this is not true for sufficiently long words in finite texts.

Therefore Eq. 1 cannot be used for step 1 in a rigorous statistical test of fit. In contrast, we have shown that the expected frequency spectrum of the ISP can be calculated accurately for individual frequencies taking into account the exact length of the text, which are two weak points in the popular derivations of the "rank" distribution of the ISP (e.g., Miller & Chomsky 1963, Li 1992, Suzuki *et al.* 2005). We leave for future work a systematic and rigorous study of the goodness of the fit of the ISP for the frequency spectrum of real words or other units.

## Acknowledgment

We thank three anonymous referees for their valuable comments. We are grateful to Brita Elvevåg for the references on Mandelbrot's seminal work. This
work was partially supported by the FIS2006-13321-C02-01 (RFC) and the project MOISES-TA, TIN2005-08832-C03 (RG) from the Spanish Ministry of Education and Science.

## References

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47-97.

Alvarez-Lacalle, E., Dorow, B., Eckmann, J.P. & Moses, E. (2006). Hierarchical structures induce long-range dynamical correlations in written texts. Proceedings of the National Academy of Sciences USA 103, (21) 7956-7961.

Bailón-Moreno, R., Jurado-Alameda, E., Ruiz-Baños, R. & Courtial, J. P. (2005). (2005). Bibliometric laws: empirical flaws of fit. Scientometrics 63 (2), 209-229.

Cohen, A., Mantegna, R. N. & Havlin, S. (1997). Numerical analysis of word frequencies in artificial and natural language texts. Fractals 5 (1), 95-104. pp. 98 (Fig. 1).

Egghe, L. (1998). On the law of Zipf-Mandelbrot for multi-word phrases. Journal of the American Society for Information Science 50 (3): 233–241.

Ferrer i Cancho, R. & Solé, R. V. (2002). Zipf´s law and random texts. Advances in Complex Systems 5 (1), 1-6.

Ferrer i Cancho, R. (2005). Zipf's law from a communicative phase transition. European Physical Journal B, 47(3), 449-457.

Furusawa, C. & Kaneko, K. (2003). Zipf's law in gene expression. Physical Review Letters 90, 088102.

Gisiger, T. (2001). Scale invariance in biology: coincidence or footprint of a universal mechanism? Biological Reviews 76, 161-209.

Goldstein, M. L., Morris, S. A. & Yen. G. G. (2004). Problems with fitting to the power-law distribution. European Physical Journal B 41 (2), 255-258.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Transactions on information theory 38 (6), 1842-1845.

McCowan, B., Hanser, S. F., and Doyle, L.R (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. Animal Behaviour 57, 409-419.

McCowan, B., Doyle, L. R., Jenkins, J. & Hanser, S. F. (2005) The appropriate use of Zipf's law in animal communication studies. Animal Behaviour 69, F1-F7.

Mandelbrot, B. (1951). Adaptation d'un message à la ligne de transmission I & II. Comptes Rendus des Séances Hebdomadaires de l'Académie des Sciences de Paris 232, 1638-1640 & 2003-2005.

Mandelbrot, B. (1953). An information theory of the structure of language. In: Commnication Theory, W. Jackson (ed.). London: Butterworth., pp. 486-502.

Manning, C. D. & Schütze, H. (1999). Foundations of statistical natural language processing. Cambridge, MA: MIT Press.

Manolopoulos, Y. (2002). Binomial coefficient computation. Recursion or iteration? SIGCSE Bulletin 34 (4), 65-67.

Miller, G. A. & Chomsky, N. (1963). Finitary models of language users. In Luce, R. Duncan, Bush, Robert R. & Galanter, E. (eds.), Handbook of mathematical psychology. New York: Wiley, pp. 419-491.

Miller, G.A., (1957). Some effects of intermittent silence. American Journal of Psychology 70, 311-314.

Suzuki, R. Tyack, P. L. & Buck, J. (2005). The use of Zipf's law in animal communication analysis. Animal Behaviour 69 (1), F9-F17.

Tuldava, J. (1996). The frequency spectrum of text and vocabulary. Journal of Quantitative Linguistics 3 (1), 38-50.

Wolfram, S. (2002). A New Kind of Science. Champain (USA): Wolfram Media Inc.
pp. 1014.

Wimmer, G. & Altmann, G. (1999). Thesaurus of univariate discrete probability distributions. Essen: Stamm.

## Appendix A

*A.1.Binomial expansion of E*[*n*(*f*|*T*)].

Now we will transform Eq. 11 from a summation on the infinite interval [$L_0$,∞] to a summation on [0,*T*-*f*] employing the binomial expansion

$$\left(1-p_x(w,L)\right)^{T-f}=\sum_{i=0}^{T-f}\binom{T-f}{i}\left(-p_x(w,L)\right)^i . \tag{A1}$$

Replacing Eq. 9 and Eq. A1 into Eq. 11 gives

$$E[n(f\mid T)]=\binom{T}{f}\left(\frac{\sigma}{(1-\sigma)^{L_0}}\right)^f\sum_{i=0}^{T-f}(-1)^i\binom{T-f}{i}\left(\frac{\sigma}{(1-\sigma)^{L_0}}\right)^i\sum_{L=L_0}^{\infty}r^L(f+i) . \tag{A2}$$

Before we proceed, we need to pay attention to two issues. First, the fact that

$$r\sum_{x=x_{\min}}^{x_{\max}}r^x=\sum_{x=x_{\min}}^{x_{\max}}r^x+r^{x_{\max}+1}-r^{x_{\min}} \tag{A3}$$

yields

$$\sum_{x=x_{\min}}^{x_{\max}}r^x=\frac{r^{x_{\min}}-r^{x_{\max}+1}}{1-r} . \tag{A4}$$

for $r\neq 1$. Secondly, notice the fact that $\sigma\in(0,1)$, $f\in[1,T]$ and thus (recall Eq. 22) gives

$$r(T)\le r(f)\le r(1)=1-\sigma . \tag{A5}$$

Applying Eq. A4, the inner summation within Eq. A2 becomes

$$\sum_{L=L_0}^{\infty}r^L(f+i)=\frac{r^{L_0}(f+i)}{1-r(f+i)} . \tag{A6}$$

assuming $r(f+i)<1$, which is warranted by Eq. A5 and the fact that $f+i\ge 1$.

Finally, replacing Eqs. 22 and A6 into Eq. A2 we obtain

$$E[n(f\mid T)]=\binom{T}{f}\sigma^f N^{L_0(1-f)}\sum_{i=0}^{T-f}\frac{(-1)^i}{1-r(f+i)}\left(\frac{\sigma}{N^{L_0}}\right)^i\binom{T-f}{i} . \tag{A7}$$

after some algebra.

*A.2. Binomial expansion of E*[*n*(*T*)].

With the same methods employed for deriving Eq. A7, it is possible to transform Eq. 15 into a sum on a finite interval. The steps for the derivation are the following. Firstly, replacing the binomial expansion

$$\left(1-p_x(w,L)\right)^T = 1+\sum_{i=1}^{T}\binom{T}{i}\left(-p_x(w,L)\right)^i \tag{A8}$$

and Eq. 9 into Eq. 15 gives

$$E[n(T)] = \sum_{i=1}^{T}(-1)^{i+1}\left[\frac{\sigma}{(1-\sigma)^{L_0}}\right]^i\binom{T}{i}\sum_{L=L_0}^{\infty} r^L(N,\sigma,i)\,. \tag{A9}$$

Recalling A4 and knowing that $r(i)<1$ (recall Eq. A5), Eq. A9 becomes

$$E[n(T)] = \sum_{i=1}^{T}(-1)^{i+1}\frac{\sigma^i N^{L_0(1-i)}}{1-r(i)}\binom{T}{i} \tag{A10}$$

after some algebra.

**Appendix B**

We aim to derive, $\gamma^{+}_{\max}(f\,|\,T)$, an upper bound of the error of calculating E[$n(f|T)$] approximately to a maximum length $L_{max}$, which is defined in Eq. 18. Knowing that $T$-$f$ ≥0 and using $\left(1-p(w,L)\right)^{T-f}\le 1$, Eq. 18 gives

$$\gamma^{+}_{\max}(f\,|\,T) = \binom{T}{f}\left(\frac{\sigma}{(1-\sigma)^{L_0}}\right)^f \sum_{L=L_{\max}+1}^{\infty} r(f)^L\,. \tag{B1}$$

Applying Eq. A4 on Eq. B1, we obtain

$$\gamma^{+}_{\max}(f\,|\,T) = \binom{T}{f}\left(\frac{\sigma}{(1-\sigma)^{L_0}}\right)^f \frac{r^{L_{\max}+1}(f)}{1-r(f)} \tag{B2}$$

assuming $r(f)<1$, which is warranted by Eq. A5.

Similarly, we aim to derive $\gamma^{-}_{\max}(f\,|\,T)$, an upper bound of the error of calculating E[$n(f|T)$] approximately from a minimum $L_{min}$ (with $L_{min}$≥0), which is defined in Eq. 19. Knowing that $T$-$f$ ≥0 and using $\left(1-p_x(w,L)\right)^{T-f}\le 1$, Eq. 19 gives

$$\gamma^{-}_{\max}(f\,|\,T) = \binom{T}{f}\left(\frac{\sigma}{(1-\sigma)^{L_0}}\right)^f \sum_{L=L_0}^{L_{\min}-1} r(f)^L\,. \tag{B3}$$

Applying Eq. A4 on Eq. B3, we obtain

$$\gamma^{-}_{\max}(f \mid T) = \binom{T}{f} \left( \frac{\sigma}{(1-\sigma)^{L_0}} \right)^{f} \frac{r^{L_0}(f) - r^{L_{\min}}(f)}{1 - r(f)} \tag{B4}$$

assuming $r(f) \neq 1$, which is again supported by Eq. A5.
It is possible to derive $L_{min}$ and $L_{max}$ from the maximum desired left and right error, respectively. On the one hand, Eq. B2 yields

$$L_{\max} = \frac{\log(\gamma^{+}_{\max}(f \mid T) G(f))}{\log r(f)} - 1 . \tag{B5}$$

with

$$G(f) = \frac{(1 - r(f))}{\binom{T}{f}} \left( \frac{(1-\sigma)^{L_0}}{\sigma} \right)^{f} . \tag{B6}$$

On the other hand, Eq. B4 yields

$$L_{\min} = \frac{\log(r^{L_0}(f) - \gamma^{-}_{\max}(f \mid T) G(f))}{\log r(f)} . \tag{B7}$$

Notice that Eq. B7 is useless when $L_0=1$ because $L_{\min} \leq L_0$. To see it, consider that the condition $L_{\min} \leq 1$ on Eq. B7 yields

$$\gamma^{-}_{\max}(f \mid T) G(f) \geq 0 \tag{B8}$$

when $L_0=1$. Eq. B8 is trivially true and thus $L_{\min} \leq 1$ as $\gamma^{-}_{\max}(f \mid T)$ and $G(f)$ are positive numbers. When $L_0=0$, our experience with typical parameters of $N$ and $\sigma$ used in the literature of the ISP (e.g., Miller & Chomsky 1963, Li 1992), is that Eq. B7 provides very little advantage compared to taking $L_{\min}=0$ with small $\gamma^{-}_{\max}(f \mid T)$ (e.g., $\gamma^{-}_{\max}(f \mid T) = 10^{-40}$).

## Appendix C

Here we consider various technical issues that arise when calculating the expected frequency spectrum and the expected vocabulary size. Firstly, when calculating $N^L$ and $1-(1-p_x(w,L))^T$ separately in Eq. 15, one has to be very careful with the divergence of $N^L$ and or the vanishing of $1-(1-p_x(w,L))^T$ beyond computer precision. For instance, if $1-(1-p_x(w,L))^T$ is practically zero according to the computer precision, one may incorrectly conclude that $(1-(1-p_x(w,L))^T) = 0$ although $N^L$ might be huge. A way of avoiding this problem is calculating E[$n(f|T)$] by means of Eq. 11 and Eq. 13 instead of Eq. 15.

Secondly, Eq. 11 needs calculating the binomial coefficient efficiently. This can be done in $\theta(\min(T,T-f))$ time (Manolopoulos 2002). Computation time can be saved in the calculation of binomial coefficients for $f \in [1,T]$ knowing that

$$\binom{T}{f} = \binom{T}{T-f}. \tag{C.1}$$

Thus, if suffices to calculate the binomial coefficient only in the range $f \in [1,\lceil T/2 \rceil]$ using Eq. C.1 when $k > \lceil T/2 \rceil$. Furthermore, it is possible to calculate the binomial coefficients for $f \in [0,\lceil T/2 \rceil]$ in time $\theta(T)$ time through the recurrence

$$\binom{T}{f} = \begin{cases} 1 & if \quad f = 0 \\ \binom{T}{f-1}((T+1)/f-1) & if \quad f \in [1,T]. \end{cases} \tag{C.2}$$

Thirdly, the taking of logarithms for very large and very small quantities and their products is necessary in many cases.

Fourthly, although the binomial coefficient in Eq. 11 is a common factor, it is convenient for safe reasons to move it inside the summation so that very low numbers within the original summation do not give a false zero if their product with the binomial coefficient cannot be neglected.

Finally, notice that although Eqs. 16 and 17 are mathematically equivalent to Eqs. 11 and 15, respectively, it is convenient to use Eqs. 11 and 13 for computer calculation for various reasons. First, direct implementation of Eqs. 16 and 17 allows one to calculate E[$n(f|T)$] and E[$n(T)$] exactly but at the risk of having to calculate more terms of the summation (with regard to Eqs. 11 and 15) than needed for a given numerical precision. Secondly, the binomial coefficients in Eqs. 16 and 17 yield huge natural numbers for sufficiently large $T$. Such large numbers cannot be stored with enough numerical precision (specially if logarithms are taken and the recurrence of Eq. C.2 is employed), which is problematic as the lower weight digits, the ones that are discarded when stored as a floating point real number, are relevant for the results of the alternating sign summation of Eqs. 16 and 17. For this reason, we leave for future work the derivation of lower or upper bounds of the error of E[$n(f|T)$] and E[$n(T)$] using Eqs. 16 and 17, as well as, the study of the convergence of the summations in Eqs. 16 and 17.

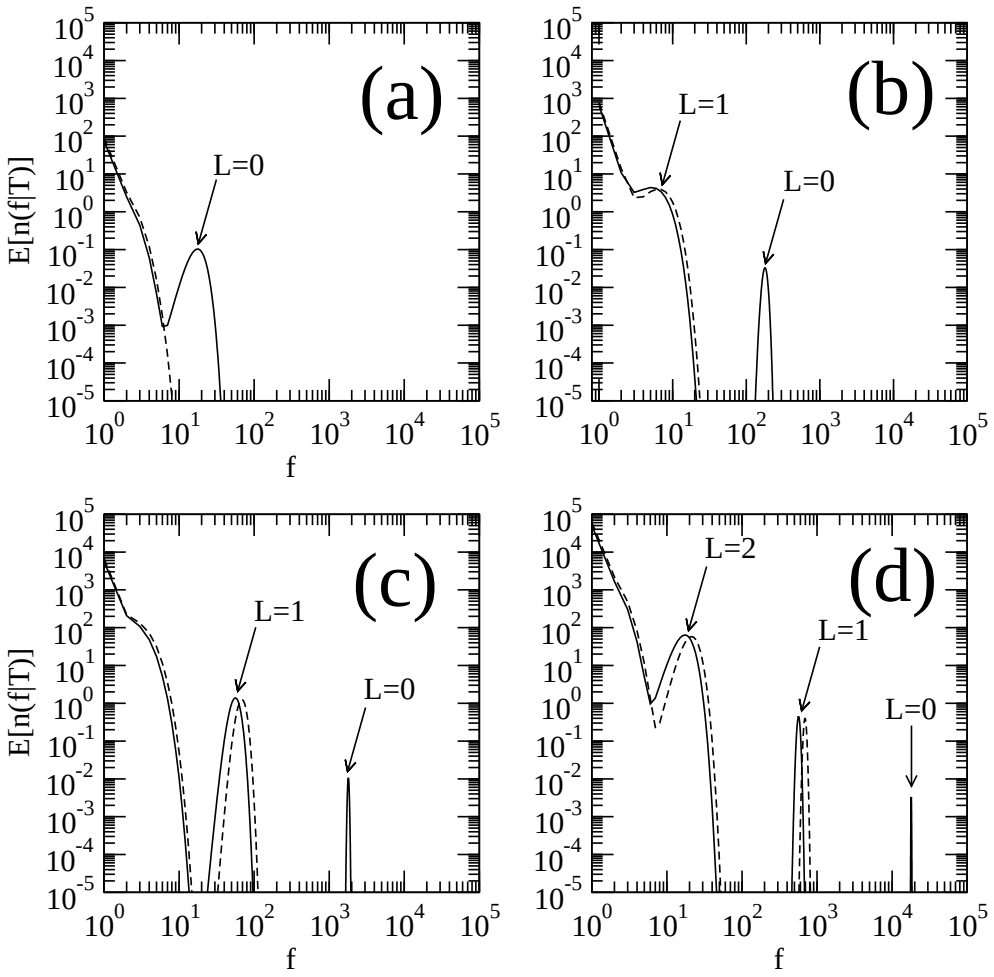


Fig. 1. The expected frequency spectrum of an ISP with $N$=26 and $\sigma$=0.18 and two different minimum lengths $L_0$=0 (solid line) and $L_0$=1 (dashed line), for different text lengths $T$. (a) $T$=$10^2$, (b) $T$=$10^3$, (c) $T$=$10^4$ and (d) $T$=$10^5$. To ease visualization and comparison, the values of $n(f|T)$ below $T$=$10^{-5}$ are not shown.

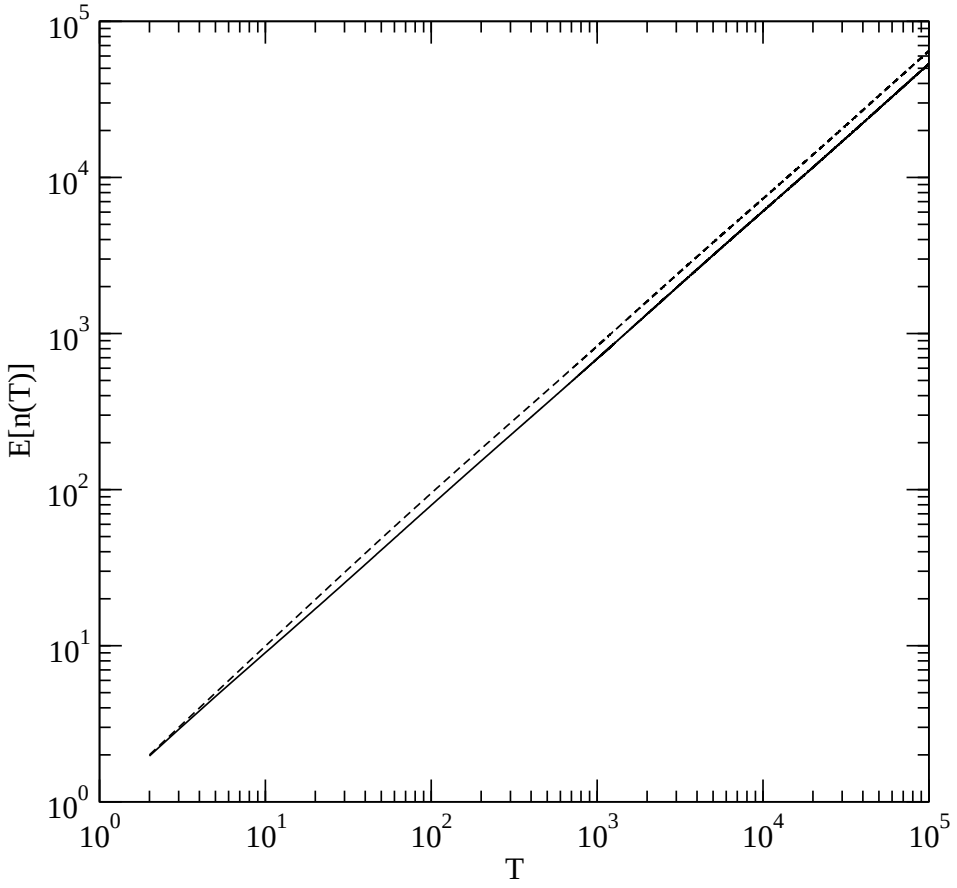


Fig. 2. The expected vocabulary growth over text length $T$ of an ISP with $N$=26 and $\sigma$=0.18 and two different minimum lengths $L_0$=0 (solid line) and $L_0$=1 (dashed line).