情 報 学 基 礎 12－3
（１９８９．２．２０）

語彙調査におけるν回出現語の出現率分散に関する考察

ν回出現語の覆内率推定誤差の分散計算法

松岡 潤
日立マイクロコンピュータエンジニアリング（株）

語彙調査で得られるν回出現語の集合の、母集団に対する覆内率$D_{\nu,N}$をＧｏｏｄの式[1]によって推定するときの推定誤差の分散の算定法を提案する。提案した方法は、$\nu+1$回出現語数$C_{\nu+1,N}$及び$\nu+2$回出現語数$C_{\nu+2,N}$から算定するものであり、母集団における語の出現率分布関数の形に依存しない。母集団の１例を用いて本方法を適用し、$C_{\nu+i,N}$（ｉ＝１，２）が２桁以上の数値である場合は１０％以内の相対誤差で算定できることを示した。

A Calculation Method of Variance of ν Times Appearing Words' Cover Ratio Estimation Error.

An Examination about Variance of Cover Ratio of Words appearing ν times in the Sample of Vocabulary Survey.

Hiroshi MATSUOKA
Hitachi Microcomputer Engineering, Ltd.
５－２２－１ Josuihon-cho Kodaira-shi,
Tokyo 187 JAPAN

I propose in this report a calculation method of variance of error which accompanies to estimation $D_{\nu,N}$ by Good's formula.[1] $D_{\nu,N}$ denotes population cover ratio of the words group,those are represented ν times in the sample sized N in a vocabulary survey. The proposed method is the calculation from $C_{\nu+i,N}(i=1,2)$, the number of words appearing $\nu+i$ times, and they are obtained in the vocabulary survey. Further the method does not depend on population frequencies distribution function.

The method was applied to some experimental samplings, and showed that we can get the aimed variances within 10% relative errors, when $C_{\nu+i,N}(i=1,2)$ are greater than 10.

1．まえがき

計算機システムによる自然言語処理には、用語、用字等の辞書が不可欠である。ここにいう辞書は計算機内部に格納された辞書であり、自然言語処理プログラムが自由にアクセスできる形態のものである。たとえば仮名漢字変換システムにおける単語や文節単位の仮名文字列とそれの漢字混じり表現との対応表などである。

そして自然言語処理システムの処理精度、処理性能は、辞書の良否に大きく依存する。

辞書を作成するためには、着目する自然言語の原始データを対象として、語彙調査が必要となる。語彙調査によって、辞書に収納すべき具体的な語と、それらの出現頻度、異なり語数等のデータが得られる。

しかし、このように語彙調査に基づいて作成した辞書を用いても、未登録の語に遭遇することがしばしば起こる。この遭遇の確率をその辞書の覆外率 (non-cover ratio)と名づける。覆外率は辞書の不完全性の一つの重要な指標と見られる。

語彙調査で得られる語全部を辞書に登録した場合の、辞書の覆外率の推定法及びその推定値の誤差の分散の推定法については、すでに報告がなされている。[1)2)3)] その主な内容は、次のとおりである。無作為抽出で得られた延べ語数Nの標本においてν回(ν=0,1,…)出現している語の集合が母集団に対してもつ覆内率$D_{\nu,N}$は

$$E[D_{\nu,N}]=\frac{\nu+1}{N+1}E[C_{\nu+1,N+1}] \tag{1}$$

である。ここにE[x]は変量xの期待値を、$C_{\nu,N}$は大きさNの標本にν回出現している語の数である。[1)3)] また、式(1)を用いて、標本全体の覆外率$D_{0,N}$を推定するときの誤差

$$w_N=\frac{C_{1,N+1}}{N+1}-D_{0,N} \tag{2}$$

の分数V[w_N]は、

$$(N+1)^2V[w_N]\fallingdotseq C_{1,N}+2C_{2,N}-\frac{1}{N+1}(C_{1,N})^2 \tag{3}$$

によって推定できる。[3)]

辞書作成に当っては、語彙調査で得られた語を必ずしも全部登録するとは限らず、所要メモリの都合その他の理由でν回出現語(νは1，2など比較的小さい値をもつものとする)を登録しないという場合もあり得る。そこでν回出現語の集合を考え、この集合の母集団に対してもつ覆内率の推定誤差の分散を簡易に推定する方法について検討した。本報告でこの検討について報告する。

2．覆内率推定誤差の分散推定式

一般のνにおいて、式(1)を用いて$D_{\nu,N}$を推定したときの誤差を$w_{\nu,N}$と書くことにする。これは前記のw_Nが$D_{0,N}$の推定誤差を表わしていたのに対し、これを一般のνへ拡張したものでもある。w_Nの場合と同様に

$$w_{\nu,N}=\frac{\nu+1}{N+1}C_{\nu+1,N+1}-D_{\nu,N} \tag{4}$$

によって$w_{\nu,N}$は表わされる。この分散V[$w_{\nu,N}$]は、

$$\begin{aligned}(N+1)V[w_{\nu,N}]=&\sum_{i=1}^{L}\binom{N}{\nu}g_i^{\nu+1}(1-g_i)^{N-\nu}\{\nu+1+(N-2\nu-1)g_i\}\\&-\sum_{i=1}^{l}\sum_{j\neq i}\binom{N}{2\nu}\binom{2\nu}{\nu}g_i^{\nu+1}g_j^{\nu+1}(1-g_i-g_j)^{N-2\nu}\end{aligned} \tag{5}$$

によって計算できる。その証明を付録2に示す。

さらに、V[$w_{\nu,N}$]は、次の式によって表現することができる。証明を付録3に示す。

$$V[w_{\nu,N}]=\frac{(\nu+1)^2}{(N+1)^2}E[C_{\nu+1,N+1}]+\frac{(N-2\nu-1)(\nu+1)(\nu+2)}{(N+1)^2(N+2)}E[C_{\nu+2,N+2}]$$

$$-\frac{1}{N+1}\binom{N}{2\nu}\binom{2\nu}{\nu}\sum_{r=0}^{N-2\nu}\frac{\binom{N-2\nu}{r}(-)^r}{\binom{N-\nu+1}{\nu+r+1}^2}\{E[C_{\nu+r+1,N-\nu+1}]\}^2 \tag{6}$$

$$+\frac{1}{N+1}\binom{N}{2\nu}\binom{2\nu}{\nu}\sum_{r=0}^{N-2\nu}\frac{\binom{N-2\nu}{r}(-)^r}{\binom{2N-2\nu+2}{2\nu+2r+2}}E[C_{2\nu+2r+2,2N-2\nu+2}]$$

式(6)において条件

(a) $N\gg 2\nu+1$ (7)

(b) $E[C_{\nu+1,N}]\geqq E[C_{\nu+1+i,N}]$ $(i=1,2\cdots)$ (8)

が成り立っているときは、微小項を省略して表現すると

$$(N+1)^2V[w_{\nu,N}]=(\nu+1)^2E[C_{\nu+1,N+1}]$$

$$+(\nu+1)(\nu+2)E[C_{\nu+2,N+2}] \tag{9}$$

$$-\frac{(\nu+1)^2}{N+1}\{E[C_{\nu+1,N-\nu+1}]\}^2+o\left(\frac{E[C_{\nu+1,N}]}{N}\right)$$

とすることができる。証明は紙数の関係で省略する。

標本の大きさNの変動がN自身の大きさと比べて小さい範囲であれば、$E[C_{\nu,N}]$ の変動も小さい。実際

$$\delta C_{\nu+1,N}\equiv E[C_{\nu+1,N+1}]-E[C_{\nu+1,N}] \tag{10}$$

と置くと、

$$|\delta C_{\nu+1,N}|\leqq\frac{\nu+2}{N-\nu-1}E[C_{\nu+1,N}] \tag{11}$$

が成り立つことが証明される。そして、式(7)により、$\nu/N\ll 1$であるので、

$$E[C_{\nu+1,N+1}]=E[C_{\nu+1,N}]+o\left(\frac{\nu}{N}E[C_{\nu+1,N}]\right) \tag{12}$$

が成り立つ。よって式(9)の右辺で、$E[C_{\nu+1,N+1}]$や$E[C_{\nu+1,N+2}]$などを$E[C_{\nu+1,N}]$で置きかえることができる。また実適用に当って、$E[C_{i,j}]$を実測値$C_{i,j}$で置きかえて、

$$(N+1)^2V[w_{\nu,N}]=(\nu+1)^2C_{\nu+1,N}+(\nu+1)(\nu+2)C_{\nu+2,N}$$

$$-\frac{(\nu+1)^2}{N+1}(C_{\nu+1,N})^2 \tag{13}$$

として用いることができる。この式が$V[w_{\nu,N}]$の推定式である。

3．精度の検証および検討

3.1 検証の方法

前章で示した式(5)、および式(13)が妥当であるかどうかを検証する。その方法として下記の方法をとった。

(1)予め各語の出現率g_i(i=1,2,…L)のわかっている充分大きな語集合を用意し、これを母集団とみなす。

(2)予め定めた適当なN，νに対し、式(5)を用いて$V[w_{\nu,N}]$を計算する。この方法で算出した$V[w_{\nu,N}]$または$\sigma[w_{\nu,N}]$を「理論式による計算値」と呼ぶ。

(3)次に、モンテカルロ法によって、無作為語抽出の実験を行ない、得られる$C_{\nu+1,N}$等から式（4）による$w_{\nu,N}$を計算する。この操作を１００回程度繰り返し行ない、その値から$V[w_{\nu,N}]$を計算する。この方法によって算出された$V[w_{\nu,N}]$, $\sigma[w_{\nu,N}]$を「語抽出シミュレーションからの統計値」と呼ぶ。

(4)次に、無作為語抽出の実験を項目(3)と同様の方法で行ない、得られる$C_{\nu+1,N}$, $C_{\nu+2,N}$などを用いて、式(13)によって$V[w_{\nu,N}]$を算出する。この方法によって算出された$V[w_{\nu,N}]$, または$\sigma[w_{\nu,N}]$を「推定式による推定値」と呼ぶ。

(5)以上で得られる3つの値

(i)理論式による計算値

(ii)語抽出シミュレーションからの統計値

(iii)推定式による推定値

を比較検討する。

3.2 検証と検討

母集団としては、座間市民の姓の集合を用いた。これは座間市の電話帳における各姓をかぞえることによって作成したものであり、すでに文献3)表1として示したものと同じである。

この母集団に対して、ν=3,4について算出された上記３種の$\sigma[w_{\nu,N}]$の値をグラフで示す(図3-1,3-2参照)。３者はほぼ等しい値となっていることがわかる。

推定式による推定値の、理論式による計算値からの誤差は、表3-1に示すとおり相対誤差が最大のところで約２５％程度である。この誤差の原因として次のものが考えられる。

(a)推定式(13)は、右辺において$o(\nu^3 E[C_{\nu+1,N}]/N)$程度の微小量を省略していることによるもの。

(b)$C_{\nu+1,N}$および$C_{\nu+2,N}$のある１回の実現値を用いているが、それらは当然$E[C_{\nu+1,N}]$, $E[C_{\nu+2,N}]$からの偏りをもっている。この偏りによるもの。

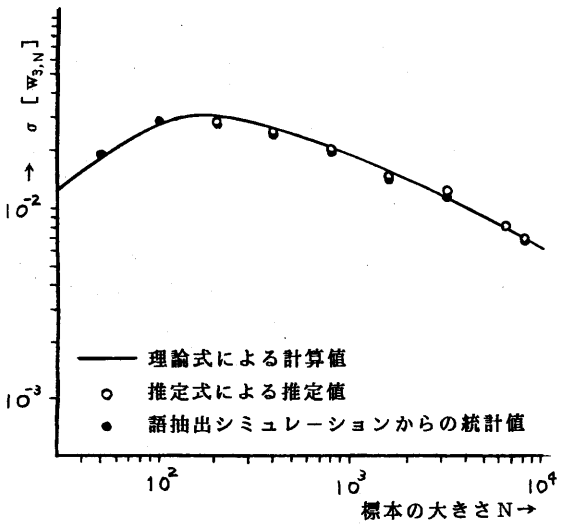

図３－１　$\sigma[w_{3,N}]$のNによる変化

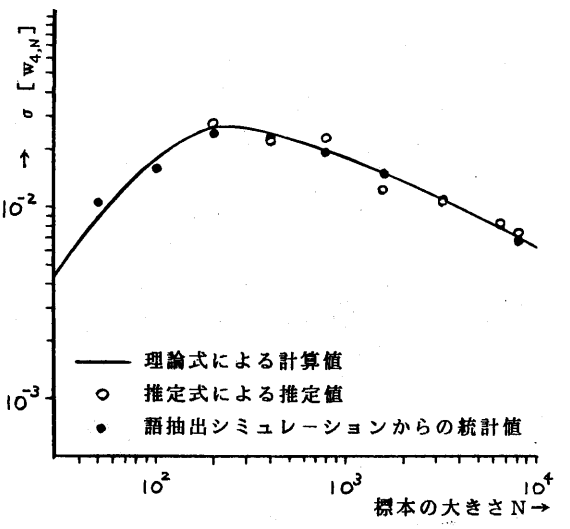

図３－２　$\sigma[w_{4,N}]$のNによる変化

表３－１　推定式による推定値と理論式による計算値との比較

項目＼標本の大きさ N		200	400	800	1600	3200	6400	8000
(a) ν 回出現語数 $C_{\nu,N}$	$C_{1,N}$	148	251	403	643	960	1360	1438
	$C_{2,N}$	16	30	71	132	226	396	484
	$C_{3,N}$	2	9	22	41	82	157	189
	$C_{4,N}$	2	4	5	21	62	91	102
	$C_{5,N}$	0	2	9	11	28	60	69
	$C_{6,N}$	1	1	4	4	17	43	62
(b)推定式 $10^4\,V[w_{3,N}]$		7.8418	6.4279	4.0446	2.1520	1.4959	0.6432	0.4673
による推定値 $10^2\,\sigma[w_{3,N}]$		2.8003	2.5353	2.0111	1.4670	1.2231	0.8020	0.6836
(c)推定式 $10^4\,V[w_{4,N}]$		*	4.9596	5.3378	1.5337	1.1749	0.6775	0.5577
による推定値 $10^2\,\sigma[w_{4,N}]$		*	2.2270	2.3104	1.2384	1.0839	0.8231	0.7468
(d)理論式 $10^2\,\tilde{\sigma}[w_{3,N}]$		3.0781	2.5806	2.0151	1.5798	1.1266	0.7763	0.6862
による計算値 $10^2\,\tilde{\sigma}[w_{4,N}]$		2.6132	2.4309	1.8775	1.5167	1.1181	0.7748	0.6869
(e)相対誤差 δ_ν (単位 %)	δ_3	-9.0	-1.8	-0.2	-7.1	8.6	3.3	-0.4
	δ_4	*	-8.4	23.1	-18.3	-3.1	6.2	8.7

(注1) ＊印は推定式(13)右辺第1項が0となる場合であり、本推定式適用は不適切であるので算出しない。

(注2) $\delta_\nu=(\sigma[w_{\nu,N}]-\tilde{\sigma}[w_{\nu,N}])/\tilde{\sigma}[w_{\nu,N}]$

表3-1からわかるように、式(13)の右辺において用いる$C_{\nu+1,N}$および$C_{\nu+2,N}$の値が小さい所では推定値の相対誤差が大きい。$C_{\nu+1,N}$および$C_{\nu+2,N}$の値が２桁(すなわち１０以上) の値をもつ範囲では相対誤差１０％以下であると言うことができる。

４．むすび

語彙調査で得られるν回出現語の集合の、母集団に対する覆内率$D_{\nu,N}$を式(1)によって推定するときの推定誤差$w_{\nu,N}$の分散$V[w_{\nu,N}]$を、語彙調査で得られているデータを用いて簡便に推定する方法について検討した。その結果、式(7)、および式(8)が成り立っているときは、式(13)によって推定できることを明らかにした。提案した方法は母集団の語彙の出現率分布関数の形によらない方法である。

参考文献

1)Good,I.J.:The Population Frequencies of Species and the Estimation of Population Parameters,Biometrica,Vol.40,Parts 3&4,pp.237-264(1953).

2)Robbins,H.E.:Estimating the Total Probability of the Unobserved Outcomes of an Experiment,Ann Math.Stat.,Vol.39,No.1,pp.256-257(1968).

3)松岡:収集語彙の母集団覆内率推定値の誤差の分散推定法の改善、情処論文誌、Vol.25,No.4,pp.560-569(1984).

[付録1] 用語・記号の説明及び変量間の基本的な関係式

本報告で用いる2・3の用語・記号について説明を記す。

覆内率：語の母集団と、語の集合αとがあるとする。母集団から無作為に語をとりだしたとき、その語がすでにαに含まれている確率を、その母集団に対するαの覆内率という。

覆外率：＝1－覆内率

$C_{\nu,N}$：母集団から無作為に抽出した、大きさNの標本において、ν回出現している語の数。（$\nu \geqq 1$，$N \geqq 1$）

$D_{\nu,N}$：母集団から無作為に抽出した、大きさNの標本において、ν回出現している語の集合の、母集団に対する覆内率。（$\nu \geqq 0$，$N \geqq 1$）この定義により$D_{0,N}$はその標本の、母集団に対する覆外率を表わす。

g_i：母集団における第i番の語の出現率。番号iは個々の語を区別するために適当な順につけるものとする。

L：母集団の語彙量（＝異なり語の総数）。

o（ε）：εの程度以下の微小量。

本報告で扱う変量の間に成り立つ基本的関係を下記する。

$$\sum_{i=1}^{L} g_i = 1 \tag{A1.1}$$

$$\sum_{\nu=0}^{N} E[D_{\nu,N}] = 1 \tag{A1.2}$$

$$E[C_{\nu,N}] = \sum_{i=1}^{L} \binom{N}{\nu} g_i^{\nu} (1-g_i)^{N-\nu} \tag{A1.3}$$

各式の説明は省略する。

[付録2] 式(5)の証明

A2.1 確率の計算式

分散の公式より

$$V[w_{\nu,N}] = E[w_{\nu,N}^2] - \{E[w_{\nu,N}]\}^2 \tag{A2.1}$$

であるが、式(1)および式(4)によって、上式右辺第2項は0である。すなわち、$V[w_{\nu,N}] = E[w_{\nu,N}^2]$である。

語を無作為に母集団から1つづつ抽出する操作を繰り返す思考実験を行なう。まず、N回の語抽出で第i語がy_i個抽出されたとる。次の変量ζ_iを考える。

$$\zeta_i = \begin{cases} 1 & (y_i = \nu \text{ のとき}) \quad \text{(A2.2a)} \\ 0 & (y_i \neq \nu \text{ のとき}) \quad \text{(A2.2b)} \end{cases}$$

さらに、もう1回、1語抽出を行ない、このN＋1回の操作で、第i語が抽出された個数をx_iとする。次の変量を考える。

$$\xi_i = \begin{cases} 1 & (x_i = \nu+1 \text{ のとき}) \quad \text{(A2.3a)} \\ 0 & (x_i \neq \nu+1 \text{ のとき}) \quad \text{(A2.3b)} \end{cases}$$

明らかに

$$\sum_{i=1}^{L} x_i = N+1 \qquad \sum_{i=1}^{L} y_i = N$$

である。また、同様に、

$$C_{\nu+1,N+1} = \sum_{i=1}^{L} \xi_i \tag{A2.4a}$$

$$D_{\nu,N} = \sum_{i=1}^{L} g_i \zeta_i \tag{A2.4b}$$

が成り立つ。よって

$$u \equiv \frac{\nu+1}{N+1} \sum_{i=1}^{L} \xi_i \tag{A2.5a}$$

$$v \equiv \sum_{i=1}^{L} g_i \zeta_i \tag{A2.5b}$$

と置くと、式(4)により

$$w_{\nu,N} = u - v \tag{A2.6}$$

である。ここで

$$\eta_i \equiv \frac{\nu+1}{N+1} \xi_i - g_i \zeta_i \tag{A2.7}$$

とおけば、式(A2.6)から

$$w_{\nu,N}^2 = \sum_{i=1}^{L} \sum_{j=1}^{L} \eta_i \eta_j \tag{A2.8}$$

であり、

$$E[w_{\nu,N}^2] = \sum_{i=1}^{L} \sum_{j=1}^{L} \eta_i \eta_j P(\eta_i \eta_j) \tag{A2.9}$$

が得られる。ただし、ここに$P(\eta_i \eta_j)$は$\eta_i \eta_j$という組み合せが起る確率を表わす。式(A2.9)はさらに

$$E[w_{\nu,N}^2] = \sum_{i=1}^{L} \sum_{j \neq i} \eta_i \eta_j P(\eta_i \eta_j) + \sum_{i=1}^{L} \eta_i^2 P(\eta_i^2) \tag{A2.10}$$

と書けるが、この右辺第1項をA，第2項をBとおくこととする。すなはち、

$$A \equiv \sum_{i=1}^{L} \sum_{j \neq i} \eta_i \eta_j P(\eta_i \eta_j) \tag{A2,11a}$$

$$B \equiv \sum_{i=1}^{L} \eta_i^2 P(\eta_i^2) \tag{A2.11b}$$

A2.2 $P(\eta_i \eta_j)$を求める

(1) $i \neq j$の場合

いま語抽出試行のN回の繰り返しで第i語がa回、第j語がb回出現する確率を$Q_N(a,b)$と書くことにすれば、

$$Q_N(a,b) = \binom{N}{a+b}\binom{a+b}{a}(1-g_i-g_j)^{N-a-b} g_i^{a} g_j^{b} \tag{A2.12}$$

である。そして、第N＋1回めの語抽出で第i語が出現する確率はg_i、第j語が出現する確率はg_j、第i語・第j語以外の語か出現する確率は$1-g_i-g_j$であるので、

$$Q_{N+1}(a,b) = (1-g_i-g_j) Q_N(a,b) + g_i Q_N(a-1,b) + g_j Q_N(a,b-1) \tag{A2.13}$$

が成り立つ。

一方、式(A2.11a)からAを算定するのに、$\eta_i \eta_j$が0であるような項を考えるのは無意味であるので、$(\zeta_i, \xi_i) \neq (0,0)$かつ$(\zeta_j, \xi_j) \neq (0,0)$の場合だけを拾い出して考えることとする。

$(\zeta_i, \xi_i, \zeta_j, \xi_j)$の組み合せが起る確率を

$$P\begin{pmatrix} \zeta_i & \zeta_j \\ \xi_i & \xi_j \end{pmatrix}$$

と書くことにする。この記法は式(A2.9)での$P(\eta_i \eta_j)$と本質的に同一のものであり、ただη_iの構成要素であるζ_i, ξ_iによってη_iを置きかえた記法に過ぎない。拾い出した各場合の確率を表A-1に示す。

表A-1の作成方法について説明する。項番1を例にとる。この欄は、$P\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$の場合、すなわち、$\zeta_i=0, \xi_i=1, \zeta_j=0, \xi_j=1$ の場合に着目している。すなわち、第N回までの語抽出で$y_i \neq \nu$, $y_j \neq \nu$であり、かつ第N+1回までの語抽出で$x_i=\nu+1$, $x_j=\nu+1$となる場合である。第N+1回の語抽出で、x_i, x_jともに$\nu+1$となる確率の総計は$Q_{N+1}(\nu+1, \nu+1)$であるが、そのうち第N回で$y_i \neq \nu, y_j \neq \nu$であったものだけを考慮しなければならない。式(A2.13)にあてはめて考えると、

$$Q_{N+1}(\nu+1, \nu+1)=(1-g_i-g_j)Q_N(\nu+1, \nu+1)$$
$$+g_i Q_N(\nu, \nu+1)+g_j Q_N(\nu+1, \nu)$$

と分解できるが、右辺第2項は$y_i=\nu$の場合であり、第3項は、$y_j=\nu$の場合であるので除外する必要がある。よって今着目している確率は右辺第1項だけであり、これを表の(a)欄に示してある。次に同表の(b)欄及び(c)欄は、式(A2.7)から求まるη_iおよびη_jをそれぞれ記載したものである。また(d)欄は(b),(c)欄の積である。

表A-1の(a)欄と(d)欄との積和をとることにより、式(A2.11a)によってAが算定される。なお式(A2.12)を用いてQの表現をg_i, g_jによる表現へ置きかえる。以上によって

$$A=-\frac{1}{N+1}\binom{N}{\nu}\binom{N-\nu}{\nu}\sum_{i=1}^{L}\sum_{j\neq i} g_i^{\nu+1} g_j^{\nu+1}$$
$$*(1-g_i-g_j)^{N-2\nu} \qquad (A2.14)$$

が得られる。

(2) i=jの場合

$\zeta_i=1$である確率は$\binom{N}{\nu} g_i^{\nu} (1-g_j)^{N-\nu}$である。そしてN+1回めの語抽出で第i語が選ばれる確率はg_iであり、そのとき$\xi_i=1$となる。また第i語でないものが選ばれる確率は$(1-g_i)$でそのときは$\xi_i=0$となる。

$\zeta_i=0$である確率は$1-\binom{N}{\nu} g_i^{\nu} (1-g_i)^{N-\nu}$であるが、N+1回めの語抽出で$\zeta_i=1$となるのはN回めまでに既に第i語が$\nu+1$回出現していた場合だけである。その確率は$\binom{N}{\nu+1} g_i^{\nu+1} (1-g_i)^{N-\nu}$ である。以上の事柄をまとめて、表A-2に示す。

表A-2から$\eta_i \neq 0$となるのは、表の1(a),1(b),2(b)の各欄に対応する場合だけであることがわかる。それぞれの場合の$\eta_i^2 P(\eta_i^2)$は表A-3のように計算される。この表の(ニ)欄と(ホ)欄との積和をとることにより

$$B=\frac{1}{N+1}\binom{N}{\nu}\sum_{i=1}^{L} g_i^{\nu+1} (1-g_i)^{N-\nu}$$
$$*\{g_i (N-2\nu-1)+\nu+1\} \qquad (A2.15)$$

が得られる。

A2.3 計算結果

式(A2.14)及び式(A2.15)を式(A2.10)に代入することにより、$E[w_{i,N}^2]$が得られ、これは直ちに$V[w_{i,N}]$である。結果は式(5)に一致する。

(証明終り)

[付録3] 式(6)の証明

紙数の関係で証明のあらすじだけを示す。式(5)の右辺は$\sum_{i=1}^{L} g_i^{\alpha}(1-g_i)^{\beta}$ の形の項の多項式として表現することができる。そして式(A1.3)を用いて$E[C_{i,N}]$ の系による表現に置き換えることによって、式(5)が得られる。 以上

表Ａ－１　確率Ｐ（$\eta_i \eta_j$）計算表

項番	(a) 確率Ｐ（$\eta_i \eta_j$）	(b) η_i	(c) η_j	(d) $\eta_i \eta_j$
1	$P\begin{pmatrix}0&0\\1&1\end{pmatrix}=(1-g_i-g_j)\,Q_N(\nu+1,\ \nu+1)$	$\frac{\nu+1}{N+1}$	$\frac{\nu+1}{N+1}$	$\left(\frac{\nu+1}{N+1}\right)^2$
2	$P\begin{pmatrix}0&1\\1&0\end{pmatrix}=(1-g_i-g_j)\,Q_N(\nu+1,\ \nu)$	$\frac{\nu+1}{N+1}$	$-g_j$	$-\frac{\nu+1}{N+1}g_j$
3	$P\begin{pmatrix}0&1\\1&1\end{pmatrix}=g_j Q_N(\nu+1,\ \nu)$	$\frac{\nu+1}{N+1}$	$\frac{\nu+1}{N+1}-g_j$	$\left(\frac{\nu+1}{N+1}\right)^2-\frac{\nu+1}{N+1}g_j$
4	$P\begin{pmatrix}1&0\\0&1\end{pmatrix}=(1-g_i-g_j)\,Q_N(\nu,\ \nu+1)$	$-g_i$	$\frac{\nu+1}{N+1}$	$-\frac{\nu+1}{N+1}g_i$
5	$P\begin{pmatrix}1&1\\0&0\end{pmatrix}=(1-g_i-g_j)\,Q_N(\nu,\ \nu)$	$-g_i$	$-g_j$	$g_i g_j$
6	$P\begin{pmatrix}1&1\\0&1\end{pmatrix}=g_j Q_N(\nu,\ \nu)$	$-g_i$	$\frac{\nu+1}{N+1}-g_j$	$-\frac{\nu+1}{N+1}g_i+g_i g_j$
7	$P\begin{pmatrix}1&0\\1&1\end{pmatrix}=g_i Q_N(\nu,\ \nu+1)$	$\frac{\nu+1}{N+1}-g_i$	$\frac{\nu+1}{N+1}$	$\left(\frac{\nu+1}{N+1}\right)^2-\frac{\nu+1}{N+1}g_i$
8	$P\begin{pmatrix}1&1\\1&0\end{pmatrix}=g_i Q_N(\nu,\ \nu)$	$\frac{\nu+1}{N+1}-g_i$	$-g_j$	$-\frac{\nu+1}{N+1}g_j+g_i g_j$
9	$P\begin{pmatrix}1&1\\1&1\end{pmatrix}=0$	$\frac{\nu+1}{N+1}-g_i$	$\frac{\nu+1}{N+1}-g_j$	$\left(\frac{\nu+1}{N+1}-g_i\right)\left(\frac{\nu+1}{N+1}-g_j\right)$

表Ａ－２　確率Ｐ（η_i^2）計算表

項番	(1) 1語抽出Ｎ回完了時の結果			(2) 1語抽出第Ｎ＋１回めの事象			
	y_i	ζ_i	y_iが左の値となる確率	(a)第ｉ語が出現		(b)第ｉ語以外が出現	
1	ν	1	$\binom{N}{\nu}g_i^{\nu}(1-g_i)^{N-\nu}$	$x_i=\nu+1$	$\xi_i=1$	$x_i=\nu$	$\xi_i=0$
2	$\nu+1$	0	$\binom{N}{\nu+1}g_i^{\nu+1}(1-g_i)^{N-\nu-1}$	$x_i=\nu+2$	$\xi_i=0$	$x_i=\nu+1$	$\xi_i=1$
3	ν，$\nu+1$以外	0	$1-\binom{N}{\nu}g_i^{\nu}(1-g_i)^{N-\nu}-\binom{N}{\nu+1}g_i^{\nu+1}(1-g_i)^{N-\nu-1}$	$x_i\neq\nu+1$	$\xi_i=0$	$x_i\neq\nu+1$	$\xi_i=0$
			確率⇒	g_i		$1-g_i$	

表Ａ－３　$\eta_i^2 P(\eta_i^2)$の計算表

項番	(イ) ζ_i	(ロ) ξ_i	(ハ) η_i	(ニ) η_i^2	(ホ)確率Ｐ（η_i^2）
1(a)	1	1	$\frac{\nu+1}{N+1}-g_i$	$\left(\frac{\nu+1}{N+1}-g_i\right)^2$	$\binom{N}{\nu}g_i^{\nu+1}(1-g_i)^{N-\nu}$
1(b)	1	0	$-g_i$	g_i^2	$\binom{N}{\nu}g_i^{\nu}(1-g_i)^{N-\nu+1}$
2(b)	0	1	$\frac{\nu+1}{N+1}$	$\left(\frac{\nu+1}{N+1}\right)^2$	$\binom{N}{\nu+1}g_i^{\nu+1}(1-g_i)^{N-\nu}$

(注) 本表の項番は、表Ａ－２における項番と(2)での事象(a),(b)とを組み合わせて示したものであり、表Ａ－２の該当欄と対応していることを表わす。上表に記載のない、項番2(a),3(a),3(b)では$(\zeta_i,\xi_i)=(0,0)$である。