MMS 27−2
(1977. 1. 25)

語彙量予測法について

松岡　潤，　絹川博之，　筒井健嗣，　尾本　健
（日立製作所　システム開発研究所）

1．はじめに

電子計算機によって、自然語を処理する場合に、大抵の場合において、標準となる言葉や文字列の集合を、予め計算機システムの内部に蓄えておき（これを機械辞書と呼ぶ）、入力される処理対象自然語と、この機械辞書のデータとの比較を行ない、その結果に基づく処理が行なわれる。このようなシステムの性能を評価するには、機械辞書に含めているデータが、一般の自然語全体に対してどれだけの部分を覆っているか、またデータを新たに追加したとき、覆率がどれだけ増加するか、が推定できることが必要となる。

自然語の中から、機械辞書に収容する予定の語を抽出・選択するには、いわゆる語彙調査が必要である。語類（言葉や漢字など）の出現率に関する調査は、国立国語研究所をはじめ、多くの機関や、人々によって今までに行なわれて来ているが[1)〜14)]、数万〜数千万の言葉や文字を扱うため、厖大な作業量を必要とする。それ故、個々の調査対象語類に対し、語彙調査に先立って、それに必要となる作業量を推定できることが望まれる。

語彙調査結果に対する要求覆率（自然語の中から任意に1語を採って来たとき、その語が調査結果内にすでに入っている確率を「覆率」と呼ぶ）を与えられたとき、それを満足するための語彙の必要調査量を推定する方法についての検討を、本稿は述べるものである。

2. 語類の出現率の近似式

2.1　用語の説明

以下に、本稿で用いる用語について、予め説明する。

語類：自然語の中で、着目している単位語の類別を表わす総称である。たとえば、日本人の姓に着目して考えるとき，対象語類は日本人の姓であるといい，ある文献中の漢字に着目するとき，対象語類は漢字であるという。

語　：1つの語類を構成する単位要素をいう。例えば、日本人の姓を対象語類とするとき、「鈴木」は1つの語であり、ある文献中の漢字を対象語類とするとき、文字「木」は1つの語である。

覆率：1つの語類に属する語の集団αを考えるものとし、αの覆率とは，対象語類に属する任意の1つの語を採ったとき、その語がαに入っている確率をいう。

非覆率：語の集団αを考えるものとし，1−（覆率）　をαの非覆率という。

2.2　出現率の近似[16)]

1つの語類において、個々の語による出現率の変化の様子の例を、図1.に示す。同図は、文献4)による雑誌の語彙についてのもので、同文献の表1をグラフにプロットしたものである。横軸は、出現率の高い語から順に1，2，…とつけた語の順位番号を表わし，たて軸は，順位1から順位xまでの語の出現率の和（すなわち、累積出現率）を表わしている。

この曲線は

$$y = 1 - e^{-ax} \qquad (1)$$

と似た形をしているので、

$$Y = 1 - y \qquad (2)$$

とおき、(x, Y) を片対数グラフにプロットしてみたところ、図2.のようで，直線とはならない。しかし、図2.は横倒しの放物線に近い形をしているから、直線化するために、さらに両軸の対数をとりプロットする。すなわち、横軸には $\log n$ をとり、たて軸には、$\log\{-\log(1-y)\}$ をとる。これを行なったものが、図3.である。同図においては、上記のことを能率的に行なうため，横軸は対数目盛であり、たて軸は、原点から $\log\{-\log(1-y)\}$ の所に $y \times 100$ (%) と目盛ったものを用いており、一種のワイブル確率紙である。これをCFグラフと呼ぶことにする。

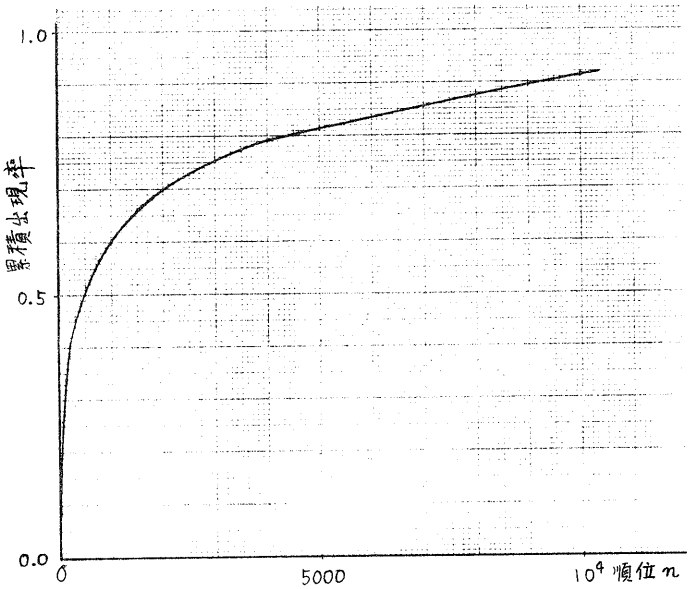

図1. 雑誌の語彙の累積出現率

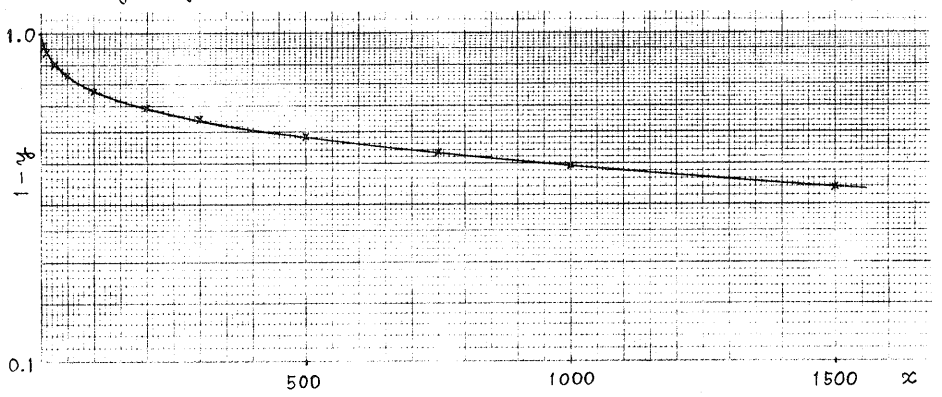

図2. 雑誌の語彙における累積出現率の補数の変化

この図では、曲線はほぼ直線に近い2つの部分と、それを滑らかに結ぶ曲線部分から成っている。このことから、この曲線は次の式で近似できることがわかる。

$$f(n) = \phi_1(n) \cdot \phi_2(n) \qquad (a)$$
$$\phi_i(n) = 1 - \exp(-p_i n^{q_i}) \qquad (i = 1, 2) \qquad (b) \qquad (3)$$

ここに、n：語の順位番号

$f(n)$：順位1から n までの語の出現率の和

p_i, q_i $(i = 1, 2)$：定数

式(3)では、$f(n)$ は $\phi_1(n)$ と $\phi_2(n)$ とに関して対称である。そこで次のように ϕ_1 と ϕ_2 とを区別することとする。CFグラフは、n の値が約50以上の部分では、ほぼ直線に近い。したがって $\phi_1(n)$ または $\phi_2(n)$ のどちらか一方のもの（仮にこれをAとする）は0.99以上となり、他方のもの（仮にBとする）が $f(n)$ の値を殆んど決める状態となっている。このBを $\phi_1(n)$ と命名し，Aの方を $\phi_2(n)$ と命名することとする。

筆者らの調査した会話の語彙，および他の発表されている語類についても、同様にCFグラフを示すと、図4.，図5.のようになる。これらの図のデータの出典

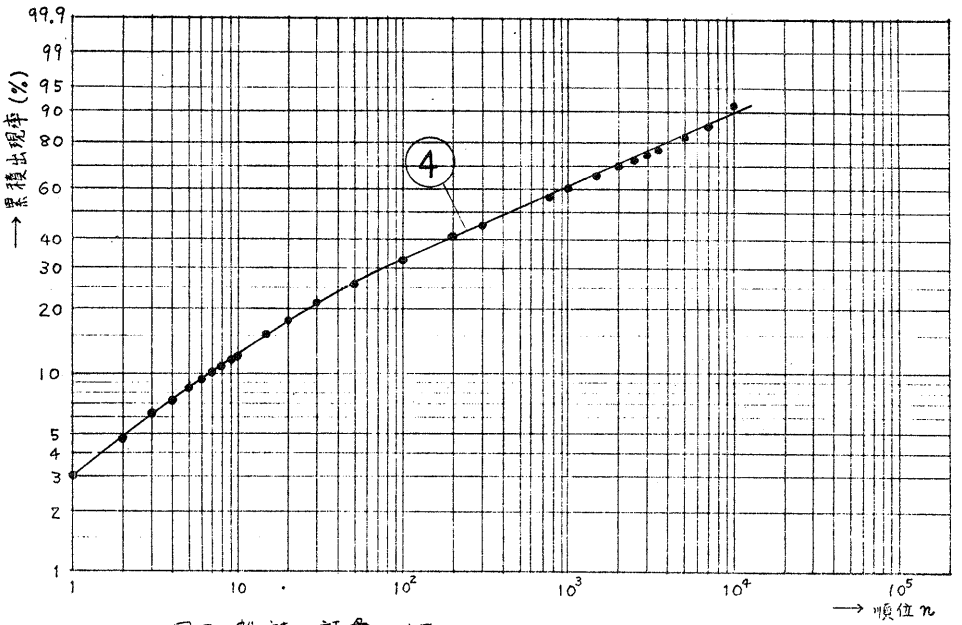

図3. 雑誌の語彙のCFグラフ

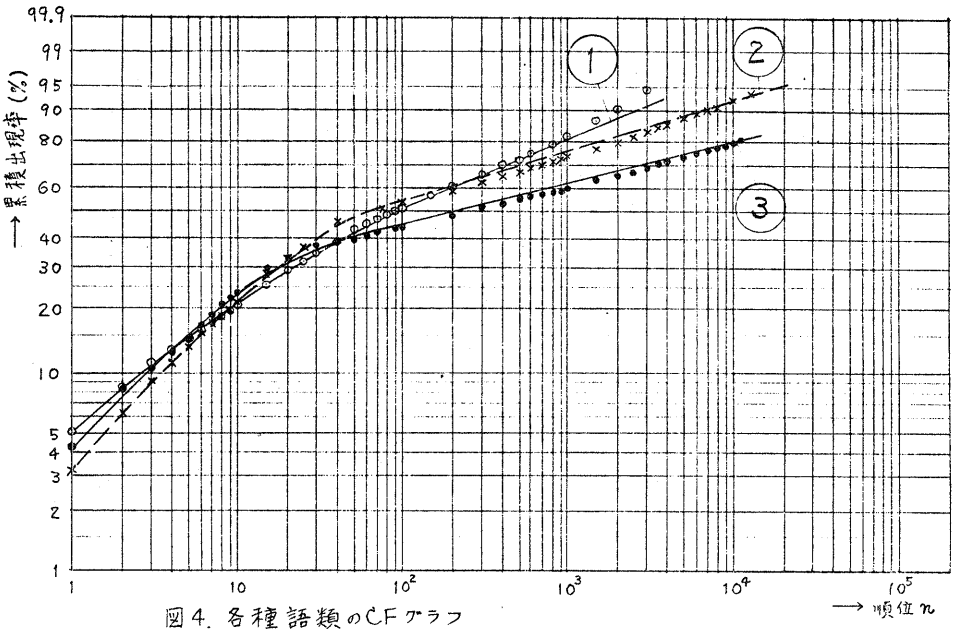

図4. 各種語類のCFグラフ

等は表1.に示すとおりである。図で、線によって示してあるものさ、各語類に対して、定数 p_i, q_i ($i = 1, 2$) を表1.のように定め，式(3)によって計算された値である。これらの図から、$f(n)$ が1に近い部分を除き，式(3)で近似できることがわかる。相対誤差は5%以下である。

定数を定める方法は、次のようにする。まず，実測値グラフの中央部分のほぼ直線に近い部分に着目し、これを近似する直線を求める。これにより $\varphi(n)$ が定まる。次にれの小さい部分の適当なm個の実測値から

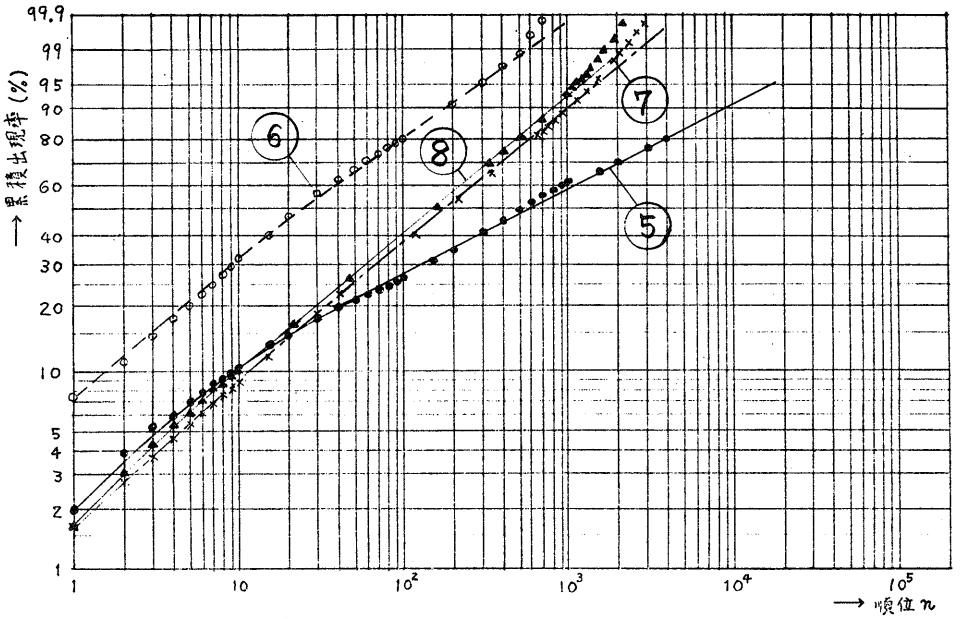

図5. 各種語類のCFグラフ

表1. 各種語類の出現率調査データと近似式の定数

項目＼語類名		(1) 会話の語彙	(2) 新聞の語彙 (短単位)	(3) 新聞の語彙 (長単位)	(4) 雑誌の語彙	(5) 日本人の姓	(6) 姓における漢字	(7) 雑誌の漢字	(8) 新聞の漢字
実測データ $f_{実}(n_j)$	(a) 調査対象	お通記録	3大新聞 (54年分)	同左	雑誌90種 (53年分)	都道府県教職員名簿	文献9)	雑誌90種 (53年分)	3大新聞 (54年分)
	(b) 報告文献 (参考文献番号)		7)	同左	4)	9)	10)	5)	8)
	(c) 調査機関	日立製作所システム研	国立国語研	同左	同左	佐久間英	日本ユニバック 田中康仁	国立国語研	同左
	(d) 標本語数	39588	940533	679342	532795	約35万	約1.2万**	280094	630313
	(e) 異なり語数	5118	47805	101081	約4万	約10万*	911	3328	2879
	(f) グラフ番号	①	②	③	④	⑤	⑥	⑦	⑧
近似式 $f_{計}(n)$	(g) p_1	0.114	0.217	0.221	0.069	0.040	0.096	0.018	0.021
	(h) q_1	0.397	0.266	0.210	0.378	0.452	0.625	0.706	0.699
	(i) p_2	0.647	0.179	0.241	0.593	0.715	1.658	2.229	1.574
	(j) q_2	0.469	0.801	0.798	0.445	0.706	0.194	0.215	0.658
備考									

＊約35万の中での異なり姓ではなく、日本全国でのもの

＊＊原文にないので筆者が推定

$$h_j = \frac{f_{実}(n_j)}{\phi_1(n_j)} \qquad (j = 1, 2, \cdots, m) \tag{4}$$

を求め，この m 個の点 (n_j, h_j) について，これらを近似する直線 $\phi_2(n)$ を求める。ただし，$f_{実}(n)$ は実測の累積出現率である。この方法は最良近似ではないが，簡単であり，実用上の精度が損なわれない。表1.に示した各定数は，この方法で，相対誤差を最小とするような最小二乗法で求めたものである。

$f(n)$ が1に近い部分では，実測値グラフはれの増大とともに急激に上昇し，式(3)からはずれてゆく。このことを仮に「グラフの急上昇」と呼ぶこととし，次節で考察する。

2．3　グラフの急上昇の理由[17]および異なり語数

有限の語の集合である、1つの標本をとれば、その中の異なり語の数も有限である。異なり語の数をいま k とすれば、標本の世界では、当然 $f(k)=1$ であり、CF グラフ用紙では、その点は上方無限遠に位置するから、グラフの急上昇は当然ということができる。しかし、その近辺におけるグラフの状況を、次のように考察することができる。

先づ、語類の母集団を、無限要素をもつものと考え、そこにおいて式(3)が成り立っているものと仮定する。語類の出現確率密度関数 $g(n)$ は

$$g(n)=f(n)-f(n-1) \tag{5}$$

と定義される。

いま、この母集団からランダムに、大きさ N の標本 S を抽出したとする。この標本の中に、順位 n の語が ν 個含まれている確率を $P(N,n,\nu)$ と書くことにすれば

$$P(N,n,\nu)=\binom{N}{\nu}\{g(n)\}^{\nu}\{1-g(n)\}^{N-\nu} \tag{6}$$

のような2項分布に従うものと考えられる。この分布に従うバラツキによって、母集団で順位 n であった語が、標本 S では順位が n となるとは限らない。S での順位番号の期待値を $n^*(N,n)$ と書くこととすれば

$$n^*(N,n)=n-\sum_{n'=1}^{n-1}P(N,n',0) \tag{7}$$

と書ける。これは、語 n が標本に現われるとすれば、母集団において順位1から $n-1$ までの語で標本 S に現われなかったものの分だけ、番号がくり上るからである。母集団での n の近辺の順位の語が、バラツキによって相互に順位が入り乱れる現象も起るが、それらは多数回の試行ではバランスし、期待値としては式(7)のようになることがわかる。

式(6)を用いると、式(7)は

$$n^*(N,n)=1+\sum_{n'=1}^{n-1}\left[1-\{1-g(n')\}^{N}\right] \qquad (n>2) \tag{7'}$$

と書ける。

さらに、以上のことから、異なり語数の期待値 $k(N)$ は

$$k(N)=\lim_{n\to\infty}n^*(N,n) \tag{8}$$

によって表わされることがわかる。

$n^*(N,n)$、および $k(N)$ の数値計算に当っては次のようにした。

(a) 出現率関数 $g(n)$ の計算に当っては、n の小さい所では定義式(5)によって計算し、n の大きな所では、次式によった。

$$\left.\begin{aligned} g(n) &= \frac{df(n)}{dn}=(A_1+A_2)\frac{f(n)}{n} && (a)\\ A_i &= \frac{q_i t_i}{\exp(t_i)\sinh(t_i)} \qquad (i=1,2) && (b)\\ t_i &= \frac{p_i n^{q_i}}{2} \qquad (i=1,2) && (c) \end{aligned}\right\} \tag{9}$$

(b) 式(8)の極限を計算機で求めるに当っては、次のようにした。

$$k(N)\fallingdotseq 1+\sum_{n'=1}^{n_t}\left[1-\{1-g(n')\}^{N}\right]+N\{1-f(n_t)\} \tag{10}$$

n_t としては $Ng(n_t)\leqq 0.13$ となるようにとることにより、相対誤差を1%以下とした。

数値計算の結果の例を図6.に示す。対象語類は前出の会話の語彙である。図で、筆者らの調査した2つの標本サイズの場合についての実測結果と、上記の方法に

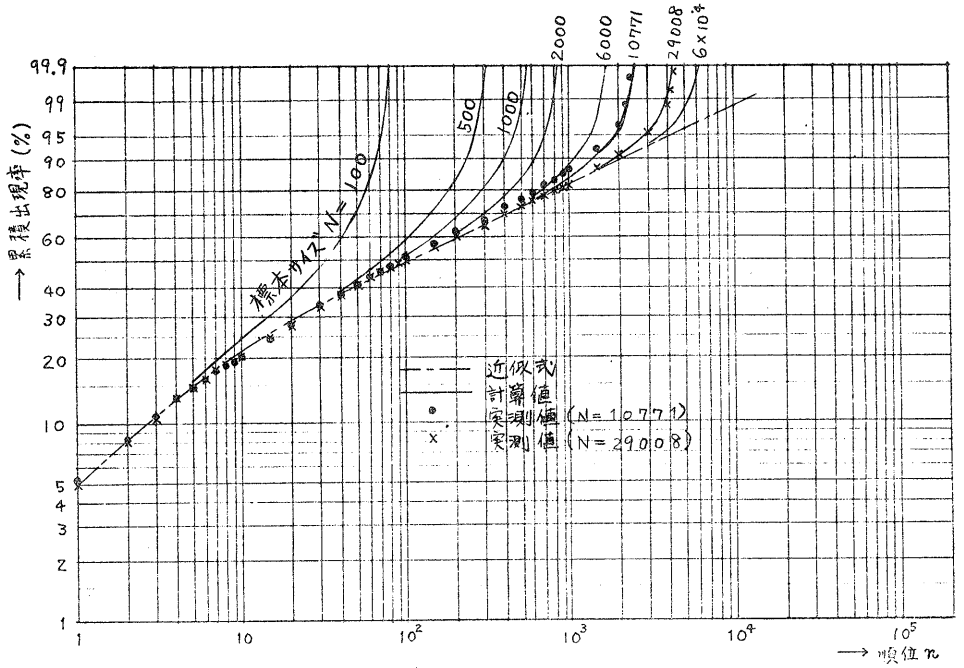

図6. 会話の語彙の標本サイズとCFグラフ

よる計算結果の対照をも示した。また、他の例について、報告されている標本サイズNから異なり語数f(N)を、式(10)によって計算したものと、報告されているf(N)の実測値との比較を、表2.に示した。

表2. 異なり語数f(N)の実測値と計算値との比較

項目＼語類		会話の語彙			新聞の語彙 (長単位)	雑誌の語彙
標本の大きさ N		10771	29008	39588	679342	532795
f(N)	実測値 f_1	2342	4391	5118	101021	約40000
	計算値 f_2	2460	4232	4951	104860	42811
相対誤差 (%) $(f_2-f_1)/f_1$		5.0	-3.6	-3.3	3.7	7.0

以上のことから、「グラフの急上昇」は母集団から有限の標本を抽出したことによって起るものであり、無限要素をもつ母集団ではNの大きい所まで式(3)によって表わされるものと考えられる。

3. 語の大量集団の確率の推定

いま、調査対象母集団から、ランダムに大きさNの標本を抽出したとし、この標本の確率の期待値を$S^*(N)$と書くこととする。母集団における順位nの語が標本に出現しない確率は、式(6)による$P(N,n,0)$であるから

$$S^*(N) = 1 - \sum_{n=1}^{\infty} P(N,n,0)\cdot g(n) \qquad (11)$$

と書ける。再び式(6)と

$$1 = \sum_{n=1}^{\infty} g(n)$$

とを用いると

$$S^*(N) = \sum_{n=1}^{\infty} g(n)\left[1 - \{1-g(n)\}^N\right] \qquad (11')$$

とも書ける。この和は次のようにして近似的に求めることができる。

$$S^*(N) = \sum_{n=1}^{n_t} g(n)\left[1 - \{1-g(n)\}^N\right] + \sum_{n=r_t+1}^{\infty} g(n)\left[1-\{1-g(n)\}^N\right] \qquad (11'')$$

であるが、n_tが充分大きければ、

$$0 < g(n) << 1 \tag{12}$$

であるので、式(11″)右辺第2項は

$$第2項 \fallingdotseq N \sum_{n=n_t+1}^{\infty} \{g(n)\}^2 \tag{13}$$

である。いま

$$I \equiv \sum_{n=n_t+1}^{\infty} \{g(n)\}^2 \tag{14}$$

$$\eta(n) \equiv -\frac{p_1 q_1}{2 n^{1-q_1}} \exp\left(-2 p_1 n^{q_1}\right) \tag{15}$$

とおくと

$$0 < I < -\eta(n_t) \tag{16}$$

であることが証明できる。このことにより

$$S^*(N) \fallingdotseq \sum_{n=1}^{n_t} g(n)\left[1 - \{1 - g(n)\}^N\right] - N \cdot \eta(n_t) \tag{17}$$

によって $S^*(N)$ は計算でき，その誤差は

$$|\varepsilon| < -N \cdot \eta(n_t) \tag{18}$$

である。

この式によって計算した例を図7.に示す。ただし、この図においては、覆率の期待値 $S^*(N)$ の代りに、非覆率の期待値 $S(N)$ をたて軸にとって示してある。なお曲線上には標本サイズ N に対する異なり語数の期待値 $f_K(N)$ を目盛った。この形式のグラフをNKSグラフと呼ぶことにする。この図の利用法を次の例に示す。

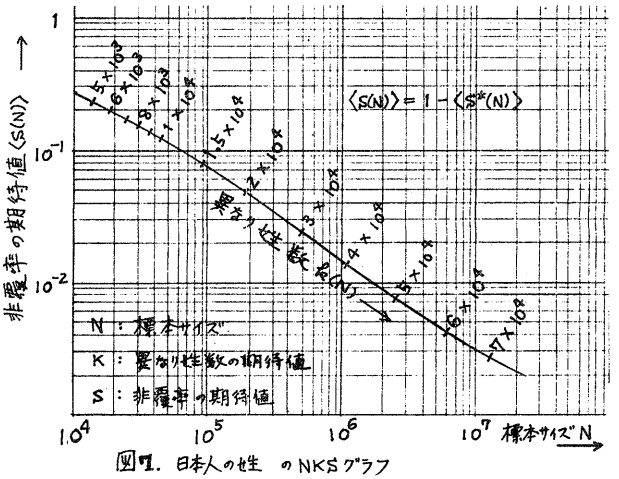

図7. 日本人の姓 のNKSグラフ

(例) 宛名のカナ漢字変換システムのような，姓を扱う情報システムで，処理対象データの99%以上を処理可能にしたいという場合。図において $S(N) = 1 - 0.99 = 0.01$ の所を見ると、

$$N \fallingdotseq 1.7 \times 10^6$$

$$f_K(N) \fallingdotseq 4.5 \times 10^4$$

が得られる。それ故，170万個の姓を集めて来て調査し，そこに現われるはずの約4万5千個の異なる姓全部を機械辞書に入れる必要がある。(例終り)

以上のことにより、語類の大量調査により得られる語集団の覆率の推定を行なうには、次のようにすればよいと言える。

(a) 1万～2万程度の標本を対象とする予備調査を行なう。

(b) 調査結果をCFグラフ化する。

(c) 近似式(3)の係数を求める。

(d) 定まった近似式により、適当にとった幾つかのNに対し、異なり語数 $f_K(N)$ を計算する。(式(8))

(e) 適当にとった幾つかのNに対し、覆率 $S^*(N)$ を計算する。(式(11))

(f) 上記(d)，(e)の結果をNKSグラフに描く。

なお，予備調査での標本の大きさを，あまり小さくすると，係数 p_1, q_1 として誤った値を得ることになるので注意を要する。1万～2万は経験的な値である。

4. 最適な語集合とランダムな語集合

母集団における各語の出現率順位番号がわかっているとし，各語を $\alpha(n)$ と表わすことにする。任意の正の整数 m を与えられたとき，語の集合 $\{\alpha(n) \mid n=1,2,\cdots,m\}$ を大きさ m の最適な語集合と呼ぶことにする。自然語処理システムの機械辞書を作るには，特に他の条件がない限り，最適な語集合を辞書に格納することが，能率のよいシステムのための必要条件となる。それは辞書用の記憶場所の大きさを一定とするとき，最適な語集合が覆率の期待値を最大とするからである。

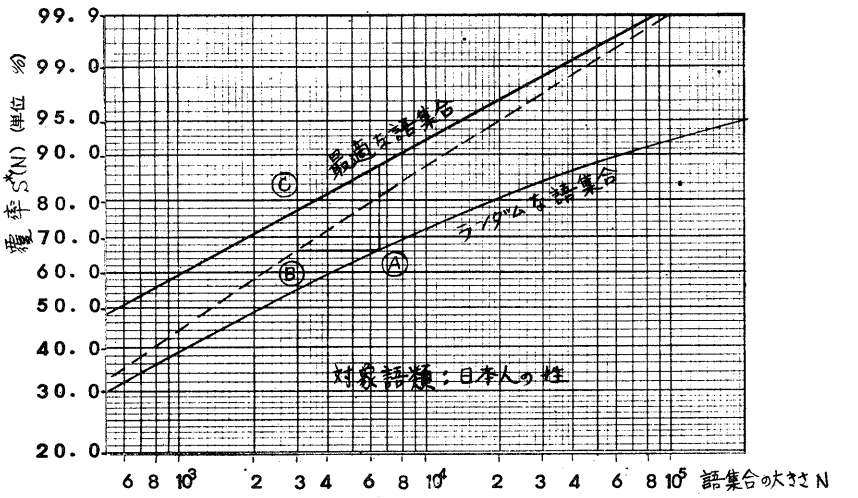

図8. 語集合の覆率の比較　日立製作所システム開発研究所

図8.に姓のランダムな語集合と，最適な語集合とでの覆率の比較を示す。破線で示した補助線は，ランダムな語集合の曲線上の各点（例えば $N = N_0$ の点）を，水平に $N = f(N_0)$ の位置まで移動したものである。例えば，ランダムに6500個の語集合をとると，それはⒶ点で示される覆率をもつ。この語集合に含まれる異なり語は点Ⓑの横座標で示される数（すなわち3000個）だけある。しかし，この異なり語の集合も最適な語集合ではないために，点Ⓒより約10%低い覆率しかもっていない。

図8.から次のことがわかる。ランダムな語集合は，最適な語集合にくらべて，覆率は小さいが，その差において，

(1) 語集合の大きさ N が大きくなると，N と異なり語数 $f(N)$ とのちがいに起因する差が大部分を占める。

(2) N が小さい所では，逆に異なり語の組合せがランダムであることによる差が大部分を占める。

5. おわりに

本稿の提案する推定法は，予備調査として1万～2万程度の語彙調査を必要とする。この予備調査自身，そう簡単ではない。しかし，数10万～数100万の語彙調査を行なうことは，10数人が数年を要する作業となり，かつ調査の過程で誤りの混入を防ぐことも至難の業となろう。したがって，調査結果のもつ価値と，所要コストとの明確な見通しがない限り，実施が困難である。本方法は，①対象語類の性格のちがいが，どのように作業量に影響してくるか，②得られる結果が，母集団に対してもつ覆率はどうか，について回答を与える試みである。

今後，さらに次の検討を行なってゆきたいと考えている。

(1) 本方法は，予備調査に基づく累積出現率を外挿する方法である。その妥当性

は、間接的には確認されているが、機会をとらえ、直接的確認を行ないたい。

(2) 本推定の各段階、すなわち、予備調査のサンプル抽出、統計過程、モデル式の係数決定、などの段階でもち込まれる誤差が、推定結果にどの程度の誤差を与えるかについての吟味。

参考文献

1) 国立国語研究所：婦人雑誌の用語：国研報告4（1953）
2) 〃：総合雑誌の用語前編：国研報告12（1957）
3) 〃：〃 後編：国研報告13（1958）
4) 〃：現代雑誌九十種の用語用字(1)総記および語彙表：国研報告21（昭37-9）
5) 〃：現代雑誌九十種の用語用字(2)漢字表：国研報告22（昭38-3）
6) Paul H. Klingbiel：Multimillion Word Data Bases - A Preliminary Report, Volume 1：Defence Document Center, Distributed by NTIS, AD-777-200（April 1974）
7) 国立国語研究所：電子計算機による新聞の語彙調査：国研報告37（昭45-2）
8) 〃：現代新聞の漢字調査（中間報告）：国研資料8（昭46-2）
9) 佐久間 英：日本人の姓：六芸書房
10) 田中 康仁：日本人の姓と名に使われる漢字：言語生活，267（昭48-12）
11) 〃：日本人の姓と名の統計：言語生活，254（昭47-11）
12) 江連 隆：漢文教科書の漢字調査：計量国語学，No.52（1970）
13) 田中 靖康：全国地名の漢字使用頻度調査：Computer Report（1973-10）
14) 国立国語研究所：電子計算機による新聞の語彙調査（IV）：国研報告48（1973）
15) Herdan, G.：Language as Choice and Chance：P. Nordhoff N.V., Groningen（1956）
16) 松岡：語彙出現率曲線の近似法について：昭50年度情処学16回大会（昭50-11）
17) 松岡：日本語の語彙統計に関する一考察：昭51年度電子通信学会全国大会（昭51-3）