言語資料間の語彙の類似度
—— 被爆手記をデータとして ——*

松 尾 雅 嗣
広島大学平和科学研究センター

DEGREE OF SIMILARITY IN THE VOCABULARY BETWEEN VERBAL DATA

Masatsugu MATSUO
Institute for Peace Science, Hiroshima University

SUMMARY

The present paper discusses how to measure the degree of similarity or difference in the vocabulary between the two language data, and proposes a new measure of vocabulary similarity, which is defined as :

$$S = \Sigma \left\{ \sqrt{p_i(X)\,p_i(Y)} - \frac{1}{2} \left| p_i(X) - p_i(Y) \right| \right\}$$

where $p_i(X)$ and $p_i(Y)$ are the rates of occurrence of the ith vocabulary type (or word type) in the two laguage data.

Since S is the sum of s_i which is defined as :

$$s_i = \sqrt{p_i(X)\,p_i(Y)} - \frac{1}{2} \left| p_i(X) - p_i(Y) \right|$$

s_i can be regarded as a measure showing how much a given vocabulary type contributes to the overall similarity or dissimilarity of the two language data.

It is also shown how S and s_i are applied to the analysis of a set of actual data, taking hibakusha's memoirs of the A-bomb experiences as example.

* この研究の一部には文部省科研費（課題番号 56530021）の補助を受けた。

目　　次

1. 問題提起と前提

ふたつの比較可能な言語資料があるとしよう。このとき，言語資料の親近性あるいは異質性とは何ぞやという問題はひとまず措くとして，ふたつの言語資料の親近性が高ければ，それはふたつの資料において使用される語彙の親近性として反映されるであろう。また逆にふたつの言語資料の異質性が高ければ，これもまた両資料における語彙の異質性として反映されるであろう。言うまでもないことながら，このような命題が常に成立するという保証はないし，逆もまた常に成立するとは限らない。しかし，この命題をひとつの作業仮説と見るならば，これが十分検討に価する作業仮説であることは否めない。[1)]言語資料の親疎遠近は，様々な要因あるいは見る者の視点，立場によって定まる，言わば未知の母数であり，語彙という側面における親疎遠近は，この未知母数に接近するひとつの指標，極限的な状態にあってはこの未知母数のひとつの代表値，と見なしうるであろう。より具体的に言うならば，ふたつの言語資料の語彙における親疎遠近を測ることができれば，これを当該資料間の親疎遠近のひとつの指標として用いることができるということになろう。

このように考えるならば，問題は，ふたつの言語資料の語彙の親疎遠近，即ち語彙の類似度を，如何にして測りうるかということになる。本稿の目的は，この問題に対してひとつの解を提示することにある。即ち，ふたつの言語資料の語彙の類似度を測る新たな指標を提示することにある。

本論に入るに先立ち，ここで若干の仮定と用語，記号の定義をしておく。

本稿で問題とする限りにおいて，言語資料は有限個の語彙トークンの集合である。ここで言う語彙トークンは，日常的な意味における「単語」，計量国語学に

おける「単位語」にほぼ相当するが，熟語，語形（word-form），形態素等をも包含するものとする。言語資料をこのように語彙トークンの集合と規定することは，念のために言えば，自然言語資料における語彙トークンの線状性（linearity）や構造の問題を捨象することを意味する。

語彙トークンという概念は当然のことながら，語彙タイプという概念と対を成す。語彙タイプの具体的言語資料における個々の現われが語彙トークンである。[2)] 本稿では以下特に断わらないが，語彙の類似度を云々するときには，ひとつの語彙タイプが実際には他の複数の語彙タイプの集合と規定されることもありうるものとしておく。[3)] 以下の議論はこのような場合についても成立する。

ここで，互いに異なる語彙タイプのみから成る集合をWとし，これを語彙タイプ集合Wと呼ぶことにする。また比較可能な言語資料をXとYとする。XとYの少なくとも一方に出現する（あるいはトークンを有する）語彙タイプ集合をW（X，Y）としよう。XとYの語彙の類似度を測ることは，W（X，Y）に関してXとYの類似度を測ることにほかならない。一般にはXとYが与えられたとき，語彙タイプをどのように定めるかという厄介な問題が介在するため，W（X，Y）は必ずしも一義的には定まらない。しかし，本稿では以下，W（X，Y）は定めることができ，かつ既に定められているものとする。

今，語彙タイプ集合W（X，Y）の空でない部分集合をW_k（X，Y）とする。現実には，W（X，Y）に関してではなく，その部分集合W_k（X，Y）に関して，ふたつの言語資料XとYの類似度を測ることも少なくない。そこで，X，Yそれぞれの部分集合のうち，W_k（X，Y）のすべての要素のすべてのトークンから成る部分集合をX_k，Y_kとする。このとき，W_k（X，Y）が決まれば，X_k，Y_kは一義的に定まる。そして，W_k（X，Y）に関してX_k，Y_kの類似度を測ることになる。逆に考えるならば，W（X，Y）に関してXとYの類似度を測ることは，W_k（X，Y）に関してX_kとY_kの類似度を測る極限的なケースと言うことができる。W_k（X，Y）がただひとつの語彙タイプから成る場合も極限的なケースであるが，これについては後述する。

W(X，Y)，W_k（X，Y）をこのように定義することに多少の問題がないわけではない。上述の定義は言語資料X，Yのいずれにも出現しない語彙タイプはW

(X, Y)から排除されている。このことは，ふたつの言語資料のいずれにも出現しない語彙タイプは，語彙の類似度には何ら影響しないと仮定することを意味する。しかし，W(X, Y)，W_k(X, Y)の定義を拡張して，このような語彙タイプを含めても以下の議論に特に影響はないので，ここでは煩雑を避けるため，このような語彙タイプはW(X, Y)，W_k(X, Y)の定義には含めない。

以上の記号を整理し，さらに後の議論に必要な記号を定義しておく。

X, Y：言語資料。

W(X, Y)：言語資料X, Yの少なくとも一方に出現する語彙タイプのみから成る語彙タイプ集合。

W_k(X, Y)：語彙タイプ集合W(X, Y)の空でない部分集合。

N：語彙タイプ集合W_k(X, Y)の要素数，W_k(X, Y)に含まれる語彙タイプ数。

w_i(K)：W_k(X, Y)の任意の要素である語彙タイプ。W(X, Y)の要素でもある。

X_k, Y_k：X, Yそれぞれの部分集合のうち，W_k(X, Y)の要素であるすべての語彙タイプのすべてのトークンから成る部分集合。X_k, Y_kがともに空であることはないが，一方のみが空であることはありうる。

f_i(X), f_i(Y)：語彙タイプw_i(K)のX, Yそれぞれにおけるトークン数(出現度数)。w_i(K)のX_k, Y_kそれぞれにおけるトークン数(出現度数)を$f_i(X_k)$, $f_i(Y_k)$とすれば，X_k, Y_kの定義からして，

$$f_i(X_k) = f_i(X)$$
$$f_i(Y_k) = f_i(Y) \qquad (1.1)$$

T(X), T(Y)：言語資料X, Yそれぞれの要素数，即ち語彙トークン総数。今，X_k, Y_kの語彙トークン総数をそれぞれ$T(X_k)$, $T(Y_k)$とすれば，1.1式より，

$$T(X_k)=\sum^{N} f_i(X_k)=\sum^{N} f_i(X)$$
$$T(Y_k)=\sum^{N} f_i(Y_k)=\sum^{N} f_i(Y) \qquad (1.2)$$

また，1.2より，

$$T(X)\geqq T(X_k)$$
$$T(Y)\geqq T(Y_k) \qquad (1.3)$$

であり，等号は$W(X, Y)=W_k(X, Y)$，即ち，$X=X_k$ かつ $Y=Y_k$のときである。

$p_i(X)$，$p_i(Y)$：語彙タイプ$w_i(K)$のX，Yにおける使用率。

$$p_i(X)=f_i(X)/T(X)$$
$$p_i(Y)=f_i(Y)/T(Y) \qquad (1.4)$$

また，

$$\sum^{N} p_i(X)\leqq 1$$
$$\sum^{N} p_i(Y)\leqq 1 \qquad (1.5)$$

であり，等号が成り立つのは，1.3式と同様$W(X, Y)=W_k(X, Y)$のときである。

本稿では，ふたつの言語資料全体の比較を優先するので，$w_i(K)$のX_k，Y_kにおける使用率，例えば，

$$p_i(X_k)=f_i(X_k)/T(X_k)$$

なる量$p_i(X_k)$は考えない。[4)]

記号をこのように定め，X，Y，$W_k(X, Y)$が定まれば，次の表1，表2の

ようなクロス表ないしはデータ行列が得られる。表1は出現度数をデータとしたものであり，表2は使用率をデータとしたものである。以下この表をもとに語彙の類似度を考えてみよう。

表1

	$w_1(K)$	$w_2(K)\cdots\cdots w_i(K)\cdots\cdots w_N(K)$	計
X_k	$f_1(X)$	$f_2(X)\cdots\cdots f_i(X)\cdots\cdots f_N(X)$	$T(X_k)=\sum^N f_i(X)$
Y_k	$f_1(Y)$	$f_2(Y)\cdots\cdots f_i(Y)\cdots\cdots f_N(Y)$	$T(Y_k)=\sum^N f_i(Y)$

表2

	$w_1(K)$	$w_2(K)\cdots\cdots w_i(K)\cdots\cdots w_N(K)$	計
X_k	$p_1(X)$	$p_2(X)\cdots\cdots p_i(X)\cdots\cdots p_N(X)$	$\sum^N p_i(X)$
Y_k	$p_1(Y)$	$p_2(Y)\cdots\cdots p_i(Y)\cdots\cdots p_N(Y)$	$\sum^N p_i(Y)$

2. 語彙の類似度

宮島達夫（1970）は，ふたつの言語資料の語彙の類似度を測るにあたって，単なる語彙タイプ数や語彙タイプの頻度を用いるのは適切でないことを例証している。宮島も指摘する通り，言語資料中の語彙タイプ数や語彙タイプの出現度数は，一般に言語資料の大きさに多分に左右される。このように，言語資料の大きさに大きく影響される値を用いることは好ましくない。[5] そこで，宮島は語彙の類似度は語彙タイプの使用率にもとづいて測るべきであるとして，次のような測度を提案する。

$$C=\Sigma \min\{p_i(X),\ p_i(Y)\}^{6)} \qquad (2.1)$$

（$\min(a,b)$は，a，bの大きくないほうの値を表わす。）

宮島（1970）ではCは言語資料に含まれるすべての語彙タイプ，つまりW(X，Y)について定義されているが，任意の部分集合W_k(X，Y)についても定義することができる。このとき，

$$C = \sum_{\Sigma}^{N} \min\{p_i(X),\ p_i(Y)\} \qquad (2.2)$$

である。

宮島のCは，言語資料間の語彙の類似度に関し従来提案された唯一の指標であると言ってよい。[7] それゆえ，ここでCを少し詳しく検討してみよう。まず，Cは次のように表わすことができる。

$$C = \Sigma \left\{ \frac{p_i(X) + p_i(Y) - |p_i(X) - p_i(Y)|}{2} \right\} \qquad (2.3)$$

Cは，2.1式から明らかなように，すべての語彙タイプについて

$$p_i(X)\, p_i(Y) = 0 \qquad (2.4)$$

のとき，最小値C＝0をとる。また，2.3式から明らかなように，すべての語彙タイプについて，

$$p_i(X) = p_i(Y) \qquad (2.5)$$

が成り立つとき，1.5式より，最大値C＝1をとる。言うまでもなく，Cの値が大きければ，類似度は大きく，Cの値が小さければ，類似度は小さい。

ここで，

$$c_i = \frac{p_i(X) + p_i(Y)}{2} - \frac{|p_i(X) - p_i(Y)|}{2} \qquad (2.6)$$

とすれば，2.3式より，

$$C = \Sigma\, c_i \qquad (2.7)$$

であり，c_i は，語彙タイプ $w_i(K)$ についてのXとYの類似度，あるいは $w_i(K)$

がXとYの類似度にどの程度寄与しているかを示す量である。[8] c_iは2.6式より，使用率の相加平均から，使用率の差の1／2を減じた値として定義されている。[9] さらにc_iは次の要請を満たす。

i）使用率の差が同じであれば，使用率の和が大きいほうが値が大きい。
（2.8）

ii）使用率の和が同じであれば，使用率の差が小さいほうが値が大きい。
（2.9）

現実の言語資料においては語彙タイプの使用率にはきわめて大きなバラツキがあり，ある語彙タイプの使用率が他の語彙タイプの使用率の数千倍，数万倍に達することも珍らしいことではない。[10] 上掲のi，iiは，このような言語資料の性質に鑑みて，使用率の和によってウェイトをつけようとするものであると言える。

c_iは，2.5式が成立するとき，つまり$w_i(K)$のX，Yにおける使用率が等しいとき最大値

$$c_i = \frac{p_i(X) + p_i(Y)}{2} \qquad (2.10)$$

をとり，2.4式が成り立つとき，つまり$w_i(K)$がX，Yのいずれか一方にのみ出現するとき，最小値

$$c_i = 0 \qquad (2.11)$$

をとる。ここで，問題になるのは，c_iの最小値である。2.4式，つまり，

$$p_i(X)\ p_i(Y) = 0$$

が成立すれば，c_iの値は，使用率の差（この場合和としても同じ）の如何にかか

わらず，常に $c_i = 0$ となる。このことは，宮島のCでは，例えば，

$$p_i(X) = 0.01, \quad p_i(Y) = 0 \qquad (2.12)$$

という場合と，

$$p_i(X) = 0.0001, \quad p_i(Y) = 0 \qquad (2.13)$$

という場合の区別ができないことを意味する。前者のほうが違いが大きいという事実を反映して，類似度の測度としては，値が小さくなるほうが好ましい。

c_i は，前述のように，使用率の相加平均がら，使用率の差の 1／2 を減じたものである。ここで，上述の c_i の欠陥を修正するため，使用率の相加平均の代りに，相乗平均を用いて，s_i なる量を次のように定義しよう。

$$s_i = \sqrt{p_i(X)\, p_i(Y)} - \frac{|p_i(X) - p_i(Y)|}{2} \qquad (2.14)$$

s_iでは，2.12，2.13が区別できることはこの式から明らかである。また，2.14式を，

$$s_i = \frac{1}{2}\sqrt{\{p_i(X) + p_i(Y)\}^2 - \{p_i(X) - p_i(Y)\}^2} - \frac{|p_i(X) - p_i(Y)|}{2} \qquad (2.15)$$

と変形すれば，2.8，2.9がともに成り立つことも明らかである。また相加平均と相乗平均の関係から，

$$s_i \leqq c_i \qquad (2.16)$$

s_iは，$p_i(X)\,p_i(Y)=0$のとき最少値

$$s_i=\frac{1}{2}\left|p_i(X)-p_i(Y)\right|=-\frac{1}{2}\left\{p_i(X)+p_i(Y)\right\} \qquad (2.17)$$

をとり，$p_i(X)=p_i(Y)$のとき最大値

$$s_i=\sqrt{p_i(X)\,p_i(Y)}=\frac{1}{2}\left\{p_i(X)+p_i(Y)\right\} \qquad (2.18)$$

をとる。ここで語彙タイプ集合$W_k(X, Y)$についてs_iの総和を求め，これをSとすれば，Sは，$W_k(X, Y)$で測ったXとYの類似度と考えることができる。2.17式，2.18式より$S=\sum^{N} s_i$の値の範囲は

$$-\frac{1}{2}\sum^{N}\left\{p_i(X)+p_i(Y)\right\}\leqq S\leqq\frac{1}{2}\sum^{N}\left\{p_i(X)+p_i(Y)\right\} \qquad (2.19)$$

$W_k(X, Y)=W(X, Y)$であれば，1.5式より

$$-1\leqq S\leqq 1 \qquad (2.20)$$

となる。Sが最小値をとるのはすべての語彙タイプについて$p_i(X)\,p_i(Y)=0$のとき，つまりXとYに共通して出現する語彙タイプが存在しないときであり，最大値をとるのはすべての語彙タイプについてXとYにおける使用率が等しいときである。

表3

資料＼語彙	w_1	w_2	w_3
X	0.1	0.4	0.5
Y	0	0.5	0.5
Z	0.1	0.5	0.4

（数字は使用率）
（宮島(1970)による）

ここで宮島自身がCの難点を示す例として挙げている仮想データを用いて，SとCを比較してみよう。データを表3に掲げる。[11)]

言語資料X，Yの類似度をS(X，Y)，C(X，Y)として示すことにする。表3から，S，Cを求めると次のようになる。

$$C(X, Y) = C(Y, X) = 0.9$$
$$C(Y, Z) = C(Z, Y) = 0.9$$
$$C(X, Z) = C(Z, X) = 0.9$$

$$S(X, Y) = S(Y, X) = 0.8472$$
$$S(Y, Z) = S(Z, Y) = 0.8472$$
$$S(X, Z) = S(Z, X) = 0.8944$$

宮島のCによれば，言語資料X，Y，Zは互いに等距離にある。しかし，宮島も指摘するように，常識的には，語彙タイプw_1の不在により，XとZの類似度は，YとXあるいはYとZの類似度より大きいと考えるべきであろう。表3のような言語資料に関しては，この意味では，CよりもSのほうが好ましいと言えよう。

以上宮島達夫によって提案された語彙の類似度の指標Cの欠陥を修正するという形でSを導いた。これまで指摘した点についてはCよりSのほうが好ましいとは言えるが，現在のところでは決定的な判断材料に乏しい。具体的な言語資料に適用して検討すべき問題であろう。以下，論じ残した2，3の問題について簡単に触れておこう。

本稿ではこれまで，類似度の測度として既存の指標，とくに2変数や2属性の相関係数について吟味してこなかった。語彙の類似度の測度としては不適当だからである。理由は次の通りである。[12)]

まず順位相関にもとづく係数，例えばスピアマンのρ，ケンドールのτは，使用率や使用率の差のもつ情報が大幅に失なわれるという欠点がある。この意味ではピアソンの相関係数γのほうが好ましい。ピアソンの相関係数γは，W(X，Y)

については，次の式で与えられる。

$$r=\frac{1}{N}\Sigma\frac{\left\{p_i(X)-\frac{1}{N}\right\}\left\{p_i(Y)-\frac{1}{N}\right\}}{\sqrt{\left[\frac{1}{N}\Sigma\left\{p_i(X)\right\}^2-\left(\frac{1}{N}\right)^2\right]\left[\frac{1}{N}\Sigma\left\{p_i(Y)\right\}^2-\left(\frac{1}{N}\right)^2\right]}} \qquad (2.21)$$

rでは2.21式から明らかなように語彙タイプw_iのXとYにおける使用率が，ともに平均使用率1／Nより大きいか，ともに1／Nより小さければrの正の方向即ち類似度を高める方向に寄与する。一般にある程度以上の大きさの言語資料を比較するとき，多くの語彙タイプの使用率は，ふたつの資料のいずれにおいても平均使用率より大きいか，いずれにおいても平均使用率より小さい。このような語彙タイプは，rの値の正の方向，つまり類似度を高める方向に寄与し，実際rの値は予想外に高くなることが多い。語彙の類似度を測るためには，このようなふたつの言語資料のいずれにおいても使用率が平均以下か，または平均以上であるような語彙タイプについても，その使用率の差（の大小）をより直接に反映しうることが望ましい。この意味でrは語彙の類似度の測度としては好ましくない。

次の問題としてはふたつの言語資料における特定の語彙タイプの差異の問題がある。これには，w_i(K)のXとYにおける使用率に有意差があるか否かを統計的に検定するという方法もある。しかし，この方法では，個々の語彙タイプの使用率の差異と，XとYの全体として類似度との関係が説明できない。これに対して本稿で定義したs_i，c_iは，XとYの全体としての類似度，あるいはW_k(X，Y)で測った類似度との関係はきわめて明快である。s_i，あるいはc_iは，個々の語彙タイプが類似度に寄与する度合を示す量であり，その意味で類似度への寄与率と称してよいであろう。寄与率は，2.14式あるいは2.6式で定義されるが，W_k(X，Y)が極限的なケースとしてただひとつの語彙タイプから成るときのSあるいはCの値と定義することもできる。

寄与率という概念を導入することにより，単にふたつの言語資料の全体としての語彙の類似度のみならず，個々の語彙タイプがそれにどのような寄与をしてい

るか，あるいはしていないかを明らかにすることができる。こう考えれば語彙の類似度の指標は，SやCのように寄与率と関係づけることのできるものでなくてはならないことは明らかであろう。

3. 被爆体験記への応用例

本節では，前節で提案した語彙の類似度の測度Sを現実の言語資料に適用した例を示す。資料として用いたのは，被爆体験記『原爆に生きて』[13]の第1部「生きる」所収の次の6編である。

短かき夜の流れ星
生命の河
白血病と闘う
ヌートリアの思い出
真如の心
母子抄

上記6編をテクスト処理プログラムLEX[14]に入力し，処理したが，[15]入力に際しては，

i) 助詞，助動詞，代名詞，固有名詞は除く，
ii) 用言は終止形に統一する，

といった便法をとった。i)により，類似度は大幅に低下することが予想される。体験記6編の語彙タイプ（この場合見出語）数とトークン数（のべ語数）を表4に示す。

LEXによって，語彙タイプの体験記ごとの頻度をデータとする，表5に示すようなデータ行列を作り，このデータ行列にもとづいて体験記6編相互についてSを算出した。

表4　体験記6編の語彙タイプ数とトークン数

	見出語数（タイプ数）	のべ語数（トークン数）
短かき夜の流れ星	1,622	2,360
生　命　の　河	891	1,335
白血病と闘う	652	794
ヌートリアの思い出	768	1,075
真　如　の　心	765	985
母　子　抄	684	846
全　体	4,346	7,395

表5　類似度算出のためのデータ（部分）

ｱｲｺｸ ﾌｼﾞﾝｶｲ	1	0	0	0	0	0
ｱｲｻﾂｽﾙ	0	0	0	3	0	0
ｱｲｼｭｳ	1	0	1	0	0	0
ｱｲｿ	1	0	0	0	0	0
ｱｲﾀｲ	0	0	0	1	0	0
ｱｲﾂｲﾃﾞ	0	1	0	0	0	0
ｱｲﾃﾆ	0	0	0	0	0	1
ｱｲﾇ ｼﾞﾝ	1	0	0	0	0	0
ｱｳ	3	0	2	1	2	0
ｱｵﾑｹ ﾆ ﾅﾙ	0	0	0	0	0	1
ｱｶｱｶ	0	0	1	0	0	0
ｱｶｲ	0	1	0	0	0	0
ｱｶｲﾝｷ	1	0	0	0	0	0
ｱｶｸ ﾔｹﾙ	0	0	1	0	0	0
ｱｶｼﾞ ｵ ｵｷﾞﾅｳ	0	0	1	0	0	0
ｱｶｽ	0	0	0	1	0	0
ｱｶﾁﾔﾝ	0	0	0	0	0	1
ｱｶﾐ ﾆ ﾅﾙ	0	0	0	0	0	1
ｱｶﾑﾗｻｷｲﾛ	0	1	0	0	0	0
ｱｶﾘ	1	0	0	0	0	0
ｱｶﾙｲ	0	0	0	0	4	0
ｱｶﾙﾐ	1	0	0	0	0	0
ｱｶﾞﾂﾃｸﾙ	0	0	0	0	1	0
ｱｶﾞﾙ	1	0	0	0	0	0
ｱｷ	3	0	0	0	1	2
ｱｷﾏﾂﾘ	1	0	0	0	0	0
ｱｷｭｳｼｮｳ	0	2	0	0	0	0
ｱｷﾗｶ	0	3	0	0	0	0
ｱｷﾗﾒﾙ	1	0	0	1	3	0
ｱｷﾚﾙ	0	0	0	1	0	0
ｱｸｴｲｷｮｳ	0	2	0	0	0	0
ｱｸｼﾞｮｳｹﾝ	0	0	0	0	0	1
ｱｸﾏ	0	0	0	0	1	0

ｱｸﾑ	0	0	0	0	2	0
ｱｹｶﾞﾀ	1	0	0	0	0	0
ｱｹﾞｸ ﾉ ﾊﾃ	0	0	0	0	1	0
ｱｹﾞﾙ	1	0	0	0	0	0
ｱｻ	3	0	1	4	1	2
ｱｻﾏｼｻ	1	0	0	0	0	0
ｱｻﾘ	1	0	0	0	0	0

（数値は各体験記ごとの単語の頻度を示す）

この結果を類似度行列の形で表６に示す。

表６　体験記６編の語彙の類似度行列

	短かき夜の流れ星	生命の河	白血病と闘う	ヌートリアの思い出	真如の心	母子抄
短かき夜の流れ星	1.0000	−0.7631	−0.6791	−0.7356	−0.7261	−0.7930
生命の河	−0.7631	−1.0000	−0.7178	−0.7686	−0.6686	−0.8055
白血病と闘う	−0.6791	−0.7178	1.0000	−0.7025	−0.5967	−0.7714
ヌートリアの思い出	−0.7356	−0.7686	−0.7025	1.0000	−0.6733	−0.7432
真如の心	−0.7261	−0.6686	−0.5967	−0.6733	1.0000	−0.7236
母子抄	−0.7930	−0.8055	−0.7714	−0.7432	−0.7236	1.0000

$S = \Sigma \{ \sqrt{p_i(X)\, p_i(Y)} - \frac{1}{2} | p_i(X) - p_i(Y) | \}$ の値を示す。
小数第５位以下切捨て。

類似度Ｓはふたつの言語資料に関して対称であるから，表６の類似度行列も対角要素に関して対称である。また対角要素は定義からして$S = 1$となる。

同様の方法によって算出した宮島のＣを，参考までに表７に示す。

被爆体験記の比較，分析それ自体は本稿の目的を越えるので，表６ないしは表７の分析はここでは差控えるが，表６のような類似度行列が得られたときの分析法のひとつの例として因子分析の例を挙げておこう。[16] 表６の類似度行列をデータとして因子分析を行なうと，表８の因子負荷行列が得られる。[17] 第１因子，第２因子の負荷量をプロットしたのが図１である。

表7　宮島のCによる類似度行列

	短かき夜の流れ星	生命の河	白血病と闘う	ヌートリアの思い出	真如の心	母子抄
短かき夜の流れ星	1.0000	0.0943	0.1333	0.1100	0.1122	0.0820
生命の河	0.0943	1.0000	0.1159	0.0962	0.1407	0.0776
白血病と闘う	0.1333	0.1159	1.0000	0.1268	0.1792	0.1030
ヌートリアの思い出	0.1100	0.0962	0.1268	1.0000	0.1428	0.1064
真如の心	0.1122	0.1407	0.1792	0.1428	1.0000	0.1202
母子抄	0.0820	0.0776	0.1030	0.1064	0.1202	1.0000

$C = \Sigma \min\{p_i(X),\ p_i(Y)\}$ の値を示す。
小数第5位以下切捨て。

表8　因子負荷行列

	第1因子	第2因子	第3因子
短かき夜の流れ星	0.2800	0.8915	−0.6948
生命の河	0.5467	−0.8508	−0.4154
白血病と闘う	0.3974	0.2025	0.3332
ヌートリアの思い出	−0.2811	0.2888	0.7839
真如の心	0.2025	−0.2730	0.4541
母子抄	−1.0860	−0.2499	−0.3845
固有値	1.8349	1.7802	1.7352
説明率	30.6	29.7	28.9

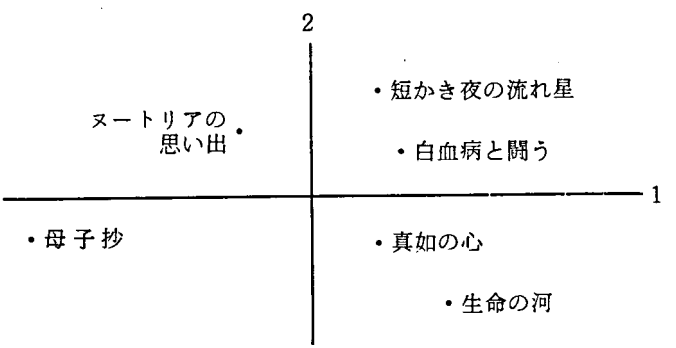

図1　因子分析にもとづく体験記6編の相互的位置

この因子分析の妥当性の吟味あるいは因子の解釈等は対象とした被爆体験記各編の詳細な検討を経て行なわなければならないが，図1に示した結果が，表6に示した体験記間の類似度の大小，即ち体験記間の距離をかなり忠実に反映していることは看取されるであろう。

この例のように因子分析もひとつの応用例ではあるが，グルーピングという観点からするならば，類似度を使った一種のクラスター分析も可能であろう。例えば，表6で最も類似度の高い「白血病と闘う」と「真如の心」をまず最初のクラスターとし，このクラスターと，他の4編との類似度を算出し，次のクラスターを作るといったやり方で順次クラスターを作ることも可能であろう。

次に，個々の語彙タイプが体験記6編相互の類似度にどの程度寄与しているか，あるいはいないかを考えてみよう。

6編を合計して頻度10以上の語彙タイプ（この場合は見出語）を選んで寄与率を計算し，寄与率の大きい語彙タイプと小さい語彙タイプ，具体的には，

$$|s_i| \geqq 0.003$$

となる語彙タイプを表にしたのが次の表9である。

表9では，対角要素の右上側には寄与率が正となる語彙タイプ，左下側には寄与率が負となる語彙タイプが与えられている。第i行第j列の桝目には，i番目の体験記とj番目の体験記に関し，$|s_i| \geqq 0.003$ となる語彙タイプがその寄与率とともに列挙されている。

表9に現われる語彙タイプは，少なくとも出現度数ないしは使用率に関しては，対象とした体験記6編においては大きな比重を占める。そして，ここに現われた語彙タイプは，ごく大ざっぱに言って次の3つのカテゴリーに分けることができよう。[18)]

家に関する語彙……「家」，「家に帰る」

家族（の構成員）……「妻」，「長男」，「次男」，「姉」，「兄」，「お父さん」，「母」，「家族」

表9　体験記6編の類似度に対する語彙タイプの寄与率

（　）内は $s_i = \sqrt{P_i(X)\,P_i(Y)} - \frac{1}{2}\,|\,p_i(X) - p_i(Y)\,|$ の値を示す）

	短かき夜の流れ星	生命の河	白血病と闘う	ヌートリアの思い出	真如の心	母子抄
短かき夜の流れ星			妻 (0.0098) 長男 (0.0082) 家 (0.0062) 帰る (0.0037) 見る (0.0033) 行く (0.0033)	帰る (0.0037) 家 (0.0034) 見る (0.0033)	家 (0.0038) 見る (0.0033)	家 (0.0058) 夜 (0.0033)
生命の河	妻 (−0.0072) 症状 (−0.0063) 工場 (−0.0061) 長男 (−0.0042) 原爆症 (−0.0041) 治療 (−0.0033)			ABCC* (0.0058) 病院 (0.0032)	病院 (0.0070) 治療 (0.0066) ABCC* (0.0039) 原爆症 (0.0036) 傷 (0.0030)	医師 (0.0059) 傷 (0.0030) 苦しい (0.0030)
白血病と闘う	医者 (−0.0044) 働く (−0.0031)	症状 (−0.0063) 長男 (−0.0062) 妻 (−0.0050) 医者 (−0.0044) 病院 (−0.0041) 起る** (−0.0037)		医者 (0.0087) 病気 (0.0046) 見る (0.0036) 帰る (0.0036) 家 (0.0035) 入院する (0.0035)	医者 (0.0081) 見る (0.0050) 治療を受ける (0.0040) 家 (0.0039) 家族 (0.0037) 入院する (0.0030)	家 (0.0059) 医者 (0.0044) 食べる (0.0037) 顔 (0.0035)
ヌートリアの思い出	妻 (−0.0072) 工場 (−0.0061) 次男 (−0.0057) 医者 (−0.0051) ABCC* (−0.0046) 長男 (−0.0042) 入院する (−0.0032)	症状 (−0.0063) 医者 (−0.0050) 原爆症 (−0.0041) 苦しい (−0.0041) 起る** (−0.0037) 治療 (−0.0033) お父さん (−0.0032)	長男 (−0.0062) 妻 (−0.0050)		医者 (0.0080) 治る (0.0040) 家 (0.0037) 見る (0.0036) ABCC* (0.0035) 病院 (0.0034)	医者 (0.0042) 家 (0.0035) 兄 (0.0034)
真如の心	工場 (−0.0061) 次男 (−0.0057) 治療 (−0.0045) 長男 (−0.0042) 医者 (−0.0040)	医者 (−0.0040) 起る** (−0.0037)	長男 (−0.0062) 病院 (−0.0035)	治療 (−0.0045) お父さん (−0.0032)		体 (0.0047) 医者 (0.0045) 家 (0.0039)
母子抄	姉 (−0.0074) 妻 (−0.0072) 工場 (−0.0061) 次男 (−0.0057) 傷 (−0.0047) 長男 (−0.0042) 医師 (−0.0041) 兄 (−0.0035)	姉 (−0.0106) 症状 (−0.0063) 母 (−0.0060) 原爆症 (−0.0041) 起る** (−0.0037) 兄 (−0.0035)	母 (−0.0130) 姉 (−0.0106) 長男 (−0.0062) 妻 (−0.0050) 傷 (−0.0047) 家に帰る (−0.0035) 働く (−0.0031)	姉 (−0.0106) ABCC* (−0.0046) 入院する (−0.0032) お父さん (−0.0032) 母 (−0.0030)	母 (−0.0130) 姉 (−0.0106) 医師 (−0.0041) 兄 (−0.0035)	

*　Atomic Bomb Casualty Commission（原爆傷害調査委員会）

**　（症状が）発生する

被爆後遺症に関するもの……「症状」，「治療」，「治療を受ける」，「原爆症」，「ABCC」[19]，「入院する」，「医者」，「医師」，「傷」，「病院」，「起る」[20]，「病気」，「治る」

第1の『家族』というカテゴリーの頻度は他のふたつに比べて小さいけれども，ここで扱った体験記に共通するのは，『家』，『家族』，『被爆後遺症』を表わす語彙であると言ってよいであろう。実際，このようなカテゴリー自体を，本稿で定義した意味での語彙タイプと見なして類似度Sを計算するならば，Sの値は表6に示した値よりも相当に大きくなるであろう。「共通する」というのはこの意味においてである。しかし，表9に着目すれば，単にこの3のカテゴリーに属する語彙が6編の手記に共通して大きな比重を占めるというだけでは済まされないことは明らかである。前述のように，表9の右上半分に現われる語彙タイプは，寄与率の高い，従って複数の体験記で共通に多用されている語彙タイプである。これに対して，左下半分に現われるのは寄与率の低い，従って，多用されているが，体験記ごとのバラツキの大きい語彙タイプである。換言すれば，体験記相互の区別に寄与している語彙タイプである。このことを念頭に置いて表9を検討するならば，次の点は明らかであろう。

i) 『家』はほとんどの体験記に共通して多用されている。

ii) 『家族』は寄与率がきわめて低い。

iii) 『被爆後遺症』の寄与率は体験記間で乱高下する。

ii)，iii)から，『家族』，『被爆後遺症』は，このような抽象的レベルでは，被爆体験記を特徴づけると言ってよいほど各体験に共通して多用されているが，具体的レベルにおいては，むしろ各体験記を区別する方向に作用していると言える。とくに『家族』については，恐らくは体験記執筆者の家族内における位置を反映して，例えば夫であるか下の娘であるかを反映して，記述の対象が，家族の誰であるかは体験記ごとに異なる。例えば，「短かき夜の流れ星」では，妻，長男，次男への言及が多く，これが本編と他の体験記を分つひとつの要因となっている。

また，「母子抄」では表題の示すとうり，母，姉への言及が多いことがひとつの弁別的特徴となっている。「姉」の寄与率は，「妻」などに比べても異様に低い。[21]

以上，語彙の類似度Sと寄与率の具体的な言語資料への適用例を手短かに示した。Sや寄与率の適用可能範囲の吟味，適用範囲の拡張等は今後の課題である。

4. 結び

言語資料間の語彙の類似度を測るという問題に関し，本稿では，

$$S=\Sigma\{\sqrt{p_i(X)\,p_i(Y)}-\frac{1}{2}|p_i(X)-p_i(Y)|\} \qquad (4.1)$$

なるSを提案し，一応の解決を与えた。本稿ではさらに，

$$s_i=\sqrt{p_i(X)\,p_i(Y)}-\frac{1}{2}|p_i(X)-p_i(Y)| \qquad (4.2)$$

によって定義される寄与率なる概念を導入し，S，s_iの具体的資料への適用を試みた。語彙の類似度の測度に関する一般的問題は上述の過程でほぼ論じ尽したと思うが，今後なお検討を要する問題も多い。以下そのうちの2，3を今後の課題として挙げておく。

まず第1に，宮島のCも同様であるが，Sは，統計的に検定できない。これは相関指標としては決定的む難点であるが，他方検定可能な既存の統計指標は語彙の類似度の測定として適当と言えないのは前述の通りである。

第2に，Sは対称な指標であり，本稿では非対称な指標は考察の対象としなかった。しかし，ふたつの言語資料間に何らかの方向性，例えば一方が他方の影響を受けていること，が明らかなとき，Sのような対称な指標が，このような非対称な言語資料間の語彙の類似度の指標として果して適切であるかという問題が残る。

最後に，本稿では語彙タイプは互いに独立である，即ち相関関係も排反関係もないと仮定してきた。しかし，語彙タイプの決め方によっては強い相関あるいは排反関係にある語彙タイプが存在しないとは限らない。このような語彙タイプが存在するとき，Sは適当な指標と言えるであろうか。これも今後の検討課題である。

註

1) 実際，用語等にもとづく著作者の推定はこのような仮定にもとづいたものと解してよいであろう。

2) 例えば，

雨は降る降る人馬は濡れる

という資料は，「雨」，「は」，「降る」，「人馬」，「濡れる」を語彙タイプと考えれば，5個の語彙タイプ，7個の語彙トークンから成る。タイプとトークンの説明にこの例を用いるのは水谷（1977）に拠る。

3) 例えば，「原爆症」，「原子症」，「原爆病」，「原爆関連症」を，ひとつの語彙タイプ「原爆症」としてグルーピングするなど。

4) X_k，Y_k における w_i（k）の使用率を用いて後述のCやSを求めることは不可能ではない。
指標としては，w_k（X，Y）に関する限り，X_k，Y_k における使用率を用いるほうが鋭敏であるとは言える。しかしこの場合後述のSや，s_i との関係がつかない。

5) 本稿で語彙の類似度の指標として，χ^2 値や χ^2 系統の関連係数，例えば ϕ 係数やクラマーのV，を採り上げないのもこの理由による。

6) 以下宮島のCについては原論文で用いられた記号ではなく，本稿で定義した記号を用いる。

7) 水谷静夫による非対称な指標Dがあるが，本稿では非対称な指標についての考察は割愛する。水谷のDについては，水谷（1982）に紹介がある。

8) c_i なり，個々の語彙タイプの寄与率という概念は，宮島（1970）ではまったく触れられていない。

9) 宮島は，出現度数の差が同じであれば，c_i の値は一定であるという趣旨のことを述べている。しかし，これは宮島の誤解であって，出現度数の差が一定であれば，使用率の差も一定ではあるが，2.6式から明らかなように c_i は一定にはならない。この場合，使用率の和（従って出現度数の和）の大きいほうが c_i の値も大きい。

10) 英語の例では，例えば Carrol *et al.*（1971）の頻度順リスト（rank list）参照。

11) 宮島（1970）pp 46－47。表3ではもとのデータの記号を変え，出現度数を使用率に改めた。

12) χ^2 値や χ^2 系統の指標については註5参照。

13) 書誌事項は「引用文献」の項に示す。

14) LEXについては松尾（1982）参照。

15) 被爆体験記のこのようなコンピューター処理は，栗原登，宇吹暁，渡辺正治，大牟田稔，河合幸尾と筆者による共同研究のための予備作業として行われたものである。

16) 同様に語彙の類似度行列に適用可能な方法としては，林の数量化の e_{ij} 型解析（所謂第Ⅳ類）がある。具体例については，水谷（1982）pp64－82参照。

17) 主成分分析の初期因子負荷量である。回転後の結果は芳しくない。計算はＳＰＳＳによる。

18) 「工場」，「働く」，「苦しい」，「帰る」，「見る」，「行く」，「夜」，「食べる」，「顔」，「体」はこの分類にあてはまらない。

19) Atomic Bomb Casualty Commission（原爆傷害調査委員会）。

20) 「生命の河」で「（症状が）発生する」の意味で用いられている。

21) 筆者は，ここで家族構成や執筆者の家族構成中の位置といった，ある程度偶然的な要因が，被爆体験記の分類や比較にとって本質的に重要な要因であると主張しているのではない。表9のようなデータは予備作業としては不可欠であり，表9からは本文で述べた結論（しかも唯一のではない）を引き出すことができることを示したにすぎない。

引　用　文　献

Carrol, J. B. *et. al. (1971) Word Frequency Book,* American Heritage, New York

原爆被害者の手記編纂委員会（編）(1953)，原爆に生きて ― 原爆被害者の手記，三一書房，東京．

松尾雅嗣(1982)，テクスト語彙処理プログラムＬＥＸ，広島大学平和科学研究センター．

宮島達夫(1970)，「語いの類似度」，国語学82，pp42－64.

水谷静夫(1977)，「語彙の量的構造」，岩波講座日本語　9，pp43－86.

_____(1982)，数理言語学，培風館，東京．