言語要素間の共出現の指標について
自然言語データ分析の一手法として

松　尾　雅　嗣
広島大学平和科学研究センター

MEASUREMENT OF COOCCURRENCE RELATIONSHIP BETWEEN LINGUISTIC ITEMS – AS A TOOL FOR ANALYSES OF NATURAL LANGUAGE DATA

Masatsugu MATSUO
Institute for Peace Science, Hiroshima University

SUMMARY

In various research fields of the humanities and the social sciences, cooccurrence relationships between two or more linguistic items have been used as a measure of the interrelatedness (distance, intimacy, etc.) of the linguistic items, and many measures, indexes, coefficients of cooccurrence relationship have been proposed and applied to natural language data. In the past researches, however, methodological examination of the proposed measure of cooccurrence has been neglected. For example, such aspects of the proposed measure as their applicability to different data and the limitation of their applicability have been given little attention.

In order to create a measure of coocurrence between two lingustic items which is universally applicable, the present paper first discusses various

* この研究の一部は文部省科学研究費補助金（課題番号 56530021）の援助を受けた。

methodological aspects of possible measures of cooccurrence, and then proposes a new measure of cooccurrence of two liguistic items, called *coefficient of cooccurrence* (C).

目　次

0. はじめに.
1. 理論的諸前提
　1.1　言語資料の構造
　1.2　実測度数と出現度数
　1.3　共出現度数
2. 共出現度数の値域と確率
　2.1　関連，独立，排反
　2.2　共出現度数の確率分布
3. 共出現の指標
　3.1　指標の満たすべき条件
　3.2　従来用いられた指標
　3.3　四分表に用いられる指標
　3.4　共出現係数
結　び

0. はじめに

ふたつ，あるいはそれ以上の言語要素（ないしは言語シンボル）の間の親疎，遠近，距離の指標として言語要素間の共出現関係（cooccurrence）を利用することは，Osgood（1959），水谷(1977），松尾ほか（1979）など，既に多くの研究者の手によって行われており，研究分野としても，人文科学，社会科学の多くの分野にわたっている。言語要素の共出現関係に着目した従来の研究の多くは，言語要素間の相互的親疎関係の究明，例えば分類やグルーピングを主たる目的とす

るものであった。このような多要素間，より一般的には多変数間の関係が問題になるとすれば，クラスター分析，林の数量化理論といった所謂多変量解析の手法が多用されるのは当然のことである。[1] しかし，多くの言語要素間の関係の分析を目的として多変量解析を行うにしても，ふたつの言語要素間の関係が基礎になっていることは否定できない。また，ふたつの言語要素間の共出現関係だけが問題となる場合も勿論ありうる。従って，言語要素間の共出現を論ずる基礎となるのは，ふたつの言語要素間の共出現関係であるということができよう。

言語要素間の共出現の度合いをどのように測り，判定するかについては，一般には，言語要素の何らかの意味での実測（観測）度数が基礎となっている。しかしながら，従来の研究においては，対象となるデータに即した指数，係数，あるいは相関度といった指標が用いられたとしても，その一般性普遍性，逆の立場からすれば適用限界といった観点からの方法的吟味はほとんど行われていないと言っていい。

本稿は，以上ふたつの意味から，ふたつの言語要素の共出現関係に着目し，それを測る一般的普遍的指標の方法的吟味を目的とする。このため，本稿ではまず共出現関係測定の基礎となる出現度数，共出現度数について考察し，それにもとづいて旧来の諸指標を比較検討する。そして，結論として厳密な確率にもとづいた新たな指標である共出現係数（coefficient of cooccurrence）を提案する。

1. 理論的諸前提

1.1 言語資料の構造

ふたつの言語要素の共出現関係を論ずるに際しては，予め大まかな議論と分析の枠組を設定しておく必要がある。特定のデータ，特定の分野に関してはともかく，共出現という概念が一般的な形で述べられたことがないという事情を考慮すれば，この必要性は尚更大である。言うまでもなく，言語要素間の共出現関係を論ずるにあたって大前提となるのは，対象となるデータ，具体的に言えば，対象となる言語資料（corpus）である。言語資料（以下単に資料と称する）は一般に，自然言語から成る素材を何らかの形で処理・加工したものである。[2] 共出現関係を論じうるためには，対象となる言語資料は複数の構成要素から構成されていなけ

ればならない。この構成要素をここではユニットと称することにしよう。ユニットは，資料の唯一の直接構成要素である必要はないが，一定の観点からしたとき，資料の直接構成要素と見なしうるものでなくてはならない。図式的に示せば，対象となる言語資料の構造は次の図1のようになる。

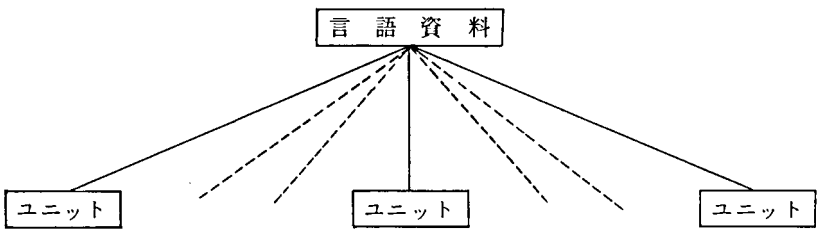


図1. 言語資料の構造

図1の樹構造図式において，ユニットは，直ちに想起される，章，段落，文といった所謂テクスト構造上の単位である場合から，書簡集の一通の書簡，一回の議会演説や記者会見，一編の演歌である場合まで含まれる。あるいは極端な例としては，常に2語からなる単語の連鎖や対でもありうる。[3)]

ふたつの言語要素の共出現を論じうるためには，ユニットは何らかの言語要素の連鎖もしくは列として定義できるものでなくてはならない。即ち，任意のユニット i は，言語要素 E_{ij} の連鎖または列として，

$$E_{i1} - E_{i2} - E_{i3} - \cdots\cdots - E_{in}$$

と表わせるものでなくてはならない。E_{ij} は具体的に，言語上の単位，例えば形態素（morpheme），語，句あるいはこれらの連鎖という形をとることになる。但し，このとき，言語要素 E_{ij} のすべてが言語学的に同一レベルにある単位である必要はない。このように表わされたユニットにおいて，任意の言語要素Xが，ユニットを構成する連鎖の，部分連鎖（部分列）であるとき，言語要素Xがユニット i に出現すると言う。他の言語要素，例えば，Yについても，Yがユニット i の部分連鎖であれば，Yはユニット i に出現する。[4)] ふたつの言語要素XとYがユニット i にともに出現するとき，XとYは（ユニット i において）共出現すると

言う。[5]

ユニットは更に，適当な単位によってその大きさを決定できるものでなければならない。ユニットの大きさがまったく問題にならないときには，ユニットの大きさはユニット自体によって測りうるものとすれば，ユニットの大きさが問題にならない場合も，この要請を満たすものと解釈できる。

共出現が問題にされる言語要素は，従来の研究においては，大部分の場合，単語（もしは，その集合）であるか，特定の意味なり価値を表現する言語表現であるといった形で規定されている。しかし，本稿では，問題の言語要素が十分明示的に定義され，かつ最終的には自然言語によって実現されるものとだけ仮定しておく。問題の言語要素をどのように定義するかは実際の分析なり研究においては極めて重要な問題であるが，ここではこの点には立入らず，上のように仮定して論を進める。[6]

1.2 実測度数と出現度数

ふたつの言語要素間の共出現関係を測る基礎となるのは，任意のユニットにおいて，この言語要素がそれぞれ実際に出現する回数（度数）である。これを実測度数と呼ぶことにする。ここで，問題となるふたつの言語要素をXとYとし，資料中の i 番目のユニットにおけるXとYそれぞれの実測度数を x_i，y_i で表わすことにしよう。ふたつの言語要素の共出現関係を論ずるとき，実測度数 x_i，y_i をそのまま用いることも少なくないが，実測度数を何らかの形で修正した値を用いることもある。それゆえ，ここでは，出現度数という語を，実測度数と明確に区別して，実測度数をそのまま用いる場合も含め，何らかの形で実測度数を加工した，言語要素の出現の度合を示す尺度という意味で用いることにする。言語要素XとYの第 i ユニットでのこの意味での出現度数をそれぞれ x'_i，y'_i とすれば，一般に出現度数は実測度数の関数として，

$$x'_i = f(x_i),\quad y'_i = f(y_i)$$

と表わされる。ここで，以下の議論のために，記号を次のように定義する。

N：資料中のユニットの総数

F_X, F_Y：資料全体における，言語要素XとYそれぞれの総出現度数。

t_i：i番目のユニットの大きさ（あるいは長さ）

T：資料全体の大きさ。　$T=\Sigma t_i$

実測度数と区別した意味での出現度数に関しては，主として単語の頻度や使用率に着目する研究において論じられているが，共出現を問題にするときには，次のような方法がある。

まず，最も簡明な方法は，実測度数をそのまま出現度数とする方法である。このとき，明らかに，

$$x'_i=x_i,\quad y'_i=y_i \qquad (1\cdot 1)$$

である。この方法を方法Ⅰと呼ぶことにしよう。方法Ⅰは成程簡明ではあるが，ユニットの大きさに極端な差のある資料には一般に適用できないという欠点がある。この方法は，ユニットの大きさがほぼ均一であると仮定できる資料か，何らかの理由でユニットの大きさを無視しうる資料でなければ適用できない。また，ふたつの言語要素の共出現関係を考えるとき，この方法であれば，$k\times l$分割表（多分表，$k\times l$クロス表）に用いられる多くの指標を適用できるように思われる。しかし，XとYがともに非常に多くの値をとり，離散変量というより連続変量に近い値をとるときには，そのひとつひとつをカテゴリーとすることにも問題があり，また多分表のほとんどの枡の度数が0と1になってしまうこともありうる。従って，実測度数をそのまま出現度数とするこの方法では，一般に多分表による分析は不可能と言わざるをえない。勿論，適当に級分けすれば，多分表を用いることは可能であるが，そのときには級の数，級の幅等の問題が生じ，級分けするのであれば，後述の方法Ⅲ～Ⅴのように，2値に級分けするほうが簡明である。

次の方法としては，ユニットの大きさを考慮に入れて，ユニットごとの出現比率を出現度数とするやり方がある。この方法では，適当な単位，例えば単語や文（sentence）で測ったi番目のユニットの大きさ（あるいは長さ）t_iを用いて，

$$x'_i = x_i / t_i$$
$$y'_i = y_i / t_i \qquad (1.2)$$

を出現度数とする。このとき，x'_i，y'_i はそれぞれXとYの第 i ユニットにおける出現比率（生起率）である。他の資料との比較のために，資料全体の大きさ，あるいはユニットの大きさを統一する必要があれば，一般に適当な定数 α を定めて，

$$x'_i = \alpha x_i / t_i$$
$$y'_i = \alpha y_i / t_i \qquad (1.3)$$

とすればよい。例えば，$\alpha = 1,000,000 / T$ とすれば，α は，資料全体の大きさ，例えば単語総数，を百万としたときの，各ユニットの平均化された大きさ（長さ）を示す。また x'_i，y'_i はこのとき，資料の大きさ百万，ユニットの大きさ α に標準化された資料の第 i ユニットにおける出現度数となる。定数 α をどのように定めるかは主として操作性という規準による。

このように出現比率を用いるとき，これを確率とみなして，XとYの出現度数を情報理論の観点からエントロピーとして把え，

$$-\Sigma \left[(x_i / t_i) \log (x_i / t_i) \right]$$

といった量を考えることもできる。資料全体での出現度数については，この考え方を適用した Carroll（1970）などの例があるが，ユニットごとの出現度数は問題にされていないので，ここではエントロピーについてはこれ以上触れない。

ユニットにおける出現比率を実測度数とするこの方法を差当り方法Ⅱとしておこう。方法Ⅱが，ユニットの大きさにばらつきのある資料に適用できることは言うまでもない。また，この方法にもとづくとき，XとYのユニット i での出現度数 x'_i，y'_i は他の方法と異なり，連続変量となる。従って，XとYの関係を測るのに使用できる既成の指標としては，ピアソンの相関係数 γ，スピアマンの順位相関係数 ρ，それに相関比の3つが挙げられようが，特定の分布を前提にしな

いこと，XとYに関して対称であることを考慮に入れれば，実際に使えるのは順位相関係数ρだけである。級分けして，多分表に縮約する場合については，方法Iに関して述べたと同じ議論が，この方法についても妥当する。

以上ふたつの方法では，実測度数を用いるにせよ，出現比率を用いるにせよ，出現度数x'_i，y'_iのとる値は原理的に多値である。これに対して以下に述べる方法は，何らかの閾値を用いて，出現度数を2値に標準化するものである。多値ないしは連続量のもつ情報が失なわれるという犠牲を仮に払ったとしても，出現度数が2値であれば処理や解決は格段に容易である。出現度数を2値にするとすれば，これを0と1に定め，

0であれば，「出現せず」，
1であれば，「出現する」

と解釈できる形にするのが最も適当であろう。そこで，問題は，どのような閾値を設定するかということになる。閾値に関して，まず関数$b(u, v)$を次のように定義する。

$$\begin{aligned} &u \geqq v \text{であれば，} b(u, v) = 1 \\ &u < v \text{であれば，} b(u, v) = 0 \end{aligned} \qquad (1 \cdot 4)$$

閾値を用いる最も簡明な方法は，実測度数と閾値1を使うことである。このとき，X，Yの出現度数は，

$$\left.\begin{aligned} x'_i &= b(x_i, 1) \\ y'_i &= b(y_i, 1) \end{aligned}\right\} \qquad (1 \cdot 5)$$

で与えられる。これを方法Ⅲとしよう。この方法は実際にはよく用いられる方法であるが，[7] X，Yの実測度数x_i，y_iがほとんどすべてのユニットで1以上となる資料，換言すれば，X，Yがともにほとんどすべてのユニットに出現するよう

な資料では利用価値がない。このような資料ではふたつの言語要素の出没よりも，むしろすべてのユニットに出現することを前提にして，出現の度数の大小を問題にしなければならないからである。このような資料では，閾値として1ではなく，XとYのユニットごとの実測度数の平均値を用いるほうがよい。今，$\overline{X}$，$\overline{Y}$を，

$$\left.\begin{aligned}\overline{X}&=\frac{1}{N}\Sigma x_i\\ \overline{Y}&=\frac{1}{N}\Sigma y_i\end{aligned}\right\} \qquad (1.6)$$

とすれば，$\overline{X}$，$\overline{Y}$は，それぞれX，Yのユニットごとの実測度数の平均値である。この平均値を用いる方法を方法Ⅳとしよう。方法Ⅳでは，ユニットごとの出現度数は，

$$\left.\begin{aligned}x'_i&=b\,(x_i,\ \overline{X})\\ y'_i&=b\,(y_i,\ \overline{Y})\end{aligned}\right\} \qquad (1.7)$$

で与えられる。この方法Ⅳでは，X，Yが，第iユニットにおいて，その平均実測度数$\overline{X}$，$\overline{Y}$以上に出現したとき，X，Yが出現したと解釈するということになる。この方法では，閾値を1とした方法Ⅲにおける適用限界の問題は解決されているが，1．7式からも明らかなように，ユニットの大きさという情報が用いられていないので，方法Ⅰ（あるいは方法Ⅲも同様）に関して述べたように，ユニットの大きさに極端なばらつきのある場合には適用できない。この点を配慮するならば，出現比率を用いた方法Ⅱにもとづいて，ユニットごとの出現比率を，平均出現率と比較するという方法が考えられる。これを方法Ⅴとしよう。今，$\overline{X}_P$，$\overline{Y}_P$をそれぞれ，

$$\left.\begin{aligned}\overline{X}_P&=\frac{1}{N}\Sigma\,(x_i/t_i)\\ \overline{Y}_P&=\frac{1}{N}\Sigma\,(y_i/t_i)\end{aligned}\right\} \qquad (1.8)$$

とするならば，$\overline{X}_P$，$\overline{Y}_P$はそれぞれ，XとYのユニットごとの出現率の平均値である。この方法によれば，XとYの第iユニットでの出現度数は，それぞれ，

$$\left.\begin{aligned} x'_i &= b\,(\,x_i / t_i,\ \bar{X}_P\,) \\ y'_i &= b\,(\,y_i / t_i,\ \bar{Y}_P\,) \end{aligned}\right\} \qquad (1.9)$$

と表わされる。この方法で，ユニット単位の平均出現率$\bar{X}_P$，$\bar{Y}_P$の代りに，資料全体での平均出現率，

$$\Sigma x_i / \Sigma t_i = \Sigma x_i / T$$

を用いることも考えられないではないが，ユニットの大きさがほぼ一定でない限り問題がある。

以上，幾つかの方法を比較したが，これまでの議論から，ふたつの言語要素について，ユニットごとの出現度数を決める方法としては，方法Vが最も制約が少なく適用範囲が広いことは明らかであろう。勿論このことは，方法Vが最も一般性，普遍性が大であるということであって，個々の具体的資料に関して，常に最も簡明な方法であることを意味するわけではない。

方法Vが最も一般性があるとはいえ，方法V自体に問題がないわけではない。ひとつは，閾値として用いた平均出現率$\bar{X}_P$，$\bar{Y}_P$が，X，Yのユニットごとの出現比率の分布のそれぞれどのような位置にあるかということである。$\bar{X}_P$，$\bar{Y}_P$が，分布の上の極端に偏った位置にあるときには，むしろ中央値を閾値として用いることも考えなければなるまい。しかし，本稿では，この点についてはこれ以上立入らず一応平均出現率を閾値として用いることに問題はない，また問題があったとしても別の適切な閾値を定めうるものとして議論を進めることにする。

方法Vについて更に一言しておけば，方法Vは，出現比率x_i / t_i，y_i / t_iを，平均0，分散1となるように所謂Zスコアに標準化しておいて，これを閾値0と比較することと同値である。[8] また方法Vによるとき，XとYの資料全体での出現度数F_X，F_Yは1．9式より次のように表わされる。

$$\left.\begin{aligned} F_X &= \Sigma x'_i = \Sigma b\,(\,x_i / t_i,\ \bar{X}_P\,) \\ F_Y &= \Sigma y'_i = \Sigma b\,(\,y_i / t_i,\ \bar{Y}_P\,) \end{aligned}\right\} \qquad (1.10)$$

以下，ふたつの言語要素XとYの出現度数が方法Vによって定められる場合を

中心に議論を進めることにしよう。

1.3 共出現度数

前項で述べた方法のいずれかによって，任意のユニットにおける言語要素X，Yの出現度数x'_i，y'_iが定められたとき，これにもとづいて，そのユニットにおけるXとYの共出現度数c_i（x'_i，y'_i）が定義される。言語要素の共出現関係を論ずるにあたって共出現度数という概念をまったく必要としないという可能性もないとは言えまいが，一般には何らの形で共出現度数の概念が用いられていると言ってよい。

第iユニットにおける言語要素XとYの共出現度数，c_i（x'_i，y'_i）としては，XとYの出現度数x'_i，y'_iのうちの大きくないほうの値をとるのが最も適当であろう。[9] 関数$min(u, v)$を，uとvのうちの大きくないほうの値をとるものと定義すれば，任意のユニットにおけるXとYの共出現度数$c_i(x'_i, y'_i)$はこの関数を用いて，前項の方法Ⅰ～Ⅴの如何に拘らず，

$$c_i(x'_i, y'_i) = min(x'_i, y'_i) \qquad (1 \cdot 11)$$

と表わされることになる。方法Ⅴのように，各ユニットにおける出現度数x'_i，y'_iが，0と1の2値に標準化されていれば，任意のユニットにおける共出現度数c_i（x'_i，y'_i）もまた0と1の値しかとりえないことは明らかである。また，このときには，明らかに，

$$\begin{aligned} c_i(x'_i, y'_i) &= min(x'_i, y'_i) \\ &= x'_i y'_i \end{aligned} \qquad (1 \cdot 12)$$

である。

資料全体におけるXとYの共出現度数をC_{XY}とすれば，C_{XY}は一般に，

$$C_{XY} = \Sigma c_i(x'_i, y'_i) = \Sigma min(x'_i, y'_i) \qquad (1 \cdot 13)$$

即ち，各ユニットにおける共出現度数の総和として定義される。資料全体におけるXとYそれぞれの出現度数F_X，F_Yについては，一般に，

$$\left.\begin{aligned} F_X &= \Sigma x'_i \\ F_Y &= \Sigma y'_i \end{aligned}\right\} \qquad (1 \cdot 14)$$

であるから，C_{XY}は常に次の関係式を満たす。

$$0 \leqq C_{XY} \leqq min\,(F_X,\ F_Y) \qquad (1 \cdot 15)$$

またユニットごとの出現度数x'_i，y'_iが0と1の2値に標準化されていれば常に次の関係が成り立つ。

$$0 \leqq F_X \leqq N \qquad (1 \cdot 16)$$

$$0 \leqq F_Y \leqq N \qquad (1 \cdot 17)$$

$$max\,(0,\ F_X + F_Y - N) \leqq C_{XY} \leqq min\,(F_X,\ F_Y) \qquad (1 \cdot 18)$$

（但し，$max\,(u,\ v)$は，u, vのうちの小さくないほうの値を表わす。）

1・18式は，x'_i，y'_iの値が0と1の2値に標準化されているとき，XとYの資料全体における共出現度数C_{XY}のとりうる値の範囲に関する制約を表現したものである。

方法Vに即して，$c_i\,(x'_i,\ y'_i)$，C_{XY}を定義すれば，1・9式，1・12式，1・13式よりそれぞれ次のようになる。

$$c_i\,(x'_i,\ y'_i) = min\left[b\,(x_i/t_i,\ \overline{X}_P),\ b\,(y_i/t_i,\ \overline{Y}_P)\right] \qquad (1 \cdot 19)$$

$$C_{XY} = \Sigma\, min\left[b\,(x_i/t_i,\ \overline{X}_P),\ b\,(y_i/t_i,\ \overline{Y}_P)\right] = \Sigma\ b\,(x_i/t_i,\ \overline{X}_P)\,b\,(y_i/t_i,\ \overline{Y}_P) \qquad (1 \cdot 20)$$

2. 共出現度数の値域と確率

2.1 関連，独立，排反

本節では，資料全体におけるXとYの共出現度数C_{XY}のとりうる値の範囲とその確率を考察する。これは，共出現の指標を論ずるため不可欠の予備作業である。

XとYのユニットごとの出現度数x'_i，y'_iの値が0と1に標準化されている

とき，次の表1のような四分表（２×２分割表，２×２クロス表）が得られる。また他の方法で出現度数を定めたとしても必要に応じこのような四分表に縮約することができる。[10] 以下表1の四分表を使って議論を進めることにするが，この四分表に関しては，

$$\left.\begin{aligned} 0 < F_X < N \\ 0 < F_Y < N \end{aligned}\right\} \qquad (2.1)$$

が常に成立するものとする。２．１式は，１．16式と１，17式に更に制約を加えたものであり，ＸとＹそれぞれについて，ユニットごとの出現度数が常に１または常に０となることはないという仮定である。

表1　四　分　表

X ＼ Y	0	1	Xの周辺度数
0	$N + C_{XY} - F_X - F_Y$	$F_Y - C_{XY}$	$N - F_X$
1	$F_X - C_{XY}$	C_{XY}	$F_X \left(= \sum_{i=1}^{N} x'_i\right)$
Yの周辺度数	$N - F_Y$	$F_Y \left(= \sum_{i=1}^{N} y'_i\right)$	N

一般にふたつの言語要素については，次の３種類の関係が想定できる。

関　連
排　反
独　立（無関係）

通常の統計学的用法からすれば，関連は正の相関，排反は負の相関と言い換えることができよう。ここで，これまで言語要素としたＸとＹを変数としよう。そして変数ＸとＹが，統計学的な意味で独立であるとしたときの，共出現度数 C_{XY} の期

待値をE（C_{XY}）と表記する。C_{XY}の期待値E（C_{XY}）については次項で詳述するが，この期待値E（C_{XY}）を用いれば，上の三つの関係は次のように定式化できる。

$C_{XY} > E(C_{XY})$ であれば，関連
$C_{XY} = E(C_{XY})$ であれば，独立
$C_{XY} < E(C_{XY})$ であれば，排反　　（2．2）

Osgood（1959）のように，期待値を閾値として，ふたつの言語要素の関係を，2，2式のように，関連と無関連に二分する考え方もあるが，[11] 共出現の指標は，関連なり排反の度合いをも表わしうるものが望ましい。しかし，そのためには，共出現度数C_{XY}がどのような値をとりうるかを，予め考察しておく必要がある。

ふたつの言語要素XとYの共出現度数C_{XY}のとりうる値の範囲は1，18式に与えたように，出現度数が標準化されていれば，

$$max(0, F_X + F_Y - N) \leqq C_{XY} \leqq min(F_X, F_Y) \qquad (2．3)$$

である。ここで，

Xが任意のユニットにn回出現すれば，Yも同一のユニットにn回出現する。
（但し，$n > 0$）　　（2．4）

Yが任意のユニットにm回出現すれば，Xも同一のユニットにm回出現する。
（但し，$m > 0$）　　（2．5）

というふたつの命題を考える。[12] このふたつの命題が常に成立するとき，これを完全関連と呼び，少なくとも一方が常に成立するとき，これを最大関連と呼ぶことにする。[13] 完全関連のときには，2．4，2．5から明らかに，

$$C_{XY} = F_X = F_Y \qquad (2．6)$$

が成立する。これは，2．3式でC_{XY}が最大となる場合の特殊なケースである。これに対して，最大関連は，これよりも緩やかで，

$$C_{XY} = min\,(F_X,\ F_Y) \qquad (2.7)$$

が成立すればいい。もっとも，命題2．4と2．5あるいは式2，6による完全関連の定義は，言語要素XとYの出没が，他の言語的，言語外的要因によって制約される可能性を無視したものであると言うこともできる。しかし，理論的には，そのような制約・影響要因がゼロに近い極限状態としての完全関連が定義できることが望ましい。従って本稿では，2．6式が成立すれば完全関連，2．7式が成立すれば最大関連と称することにする。

2．2に示したふたつの言語要素の関係のうち，「関連」の場合の極限状態としての完全関連，最大関連は上述のように定義される。「排反」の場合も同様に，極限としての完全排反，最大排反が定義できるはずであるが，この場合問題は多少複雑である。常識的に考えるならば，

任意のユニットにXとYがともに出現することはない。　（2．8）

という命題が常に成り立てば，これを完全排反と称してよいように思われる。しかし，表1のような四分表で考えるならば，これは実は最大排反でしかない。[14] 理論的な完全排反であれば，命題2．8のみならず，

任意のユニットにおいては，XとYのいずれか一方が出現する。　（2．9）

という命題も常に成立しなければならない。この意味での完全排反のときには，

$$C_{XY} = F_X + F_Y - N = 0 \qquad (2.10)$$

が成立する。これに対して，最大排反は一般には命題2．8が成立すればよいから，共出現度数C_{XY}がゼロとなればよいが，実際には，C_{XY}は，制約式2．3のもとでは，最小値がゼロとならぬこともある。そこで，命題2．8に拘らず，これを緩めて，

$$C_{XY} = max\,(0,\ F_X + F_Y - N) \qquad (2.11)$$

が成立すれば最大排反とする。このような完全排反，最大排反の定義，とくに完全排反の定義は，言語データに関しては場合によっては相当に強引な定義であるが,[15] 本稿では次に述べるような理論的整合性に鑑みて，この定義に従うことにする。

上述の完全関連，最大関連，最大排反，完全排反は，別の観点からすれば，次のように定義することもできる。表1の四分表において，X，Yの周辺度数F_X，F_Yを固定しないで考えてC_{XY}が最大となるときが完全関連であり，C_{XY}が最小となるときが完全排反である。これに対して，X，Yの周辺度数F_X，F_Yを固定して考えて，C_{XY}が最大となるときが最大関連であり，C_{XY}が最小となるときが最大排反である。即ち，「完全」は，周辺度数を固定しないときの極限であり，「最大」は周辺度数を固定したときの極限である。理論的には，共出現の指標は，この4つの状態に対し，異なった値を示す指標であることが望ましい，即ち完全関連と完全排反にもとづく指標であることが望ましいと言えよう。しかし，周辺度数F_X，F_Yは言語要素の出現度数であり，これを固定しないで何らの制約も加えないことを前提とすることが，対象が言語要素であるとき果して適切かどうかという問題がある。この点も含め，完全関連，完全排反の定義にもとづく指標と，最大関連と最大排反にもとづく指標との比較は後に改めて論ずる。

2.2 共出現度数の確率分布

本項では，共出現の指標の示す値の解釈に欠くことのできない，共出現度数C_{XY}が任意の値をとる確率を考察する。

今，XとYを確率変数とすることにしよう。（本稿では，X，Yを言語要素としても，また変数としても用いている。横着な用法だが，誤解の問題は生じないだろう。）このとき，XとYの関数である共出現度数C_{XY}もまた（離散型の）確率変数である。XとYの分布の型が明らかであれば，その関数C_{XY}の分布の型も明らかになるが，XとYいずれについても分布の型はおろか，出現比率ないしは確率変数としての期待値も明らかではないというのが普通であろう。従って，現実には，XとYの出現比率としては，資料での出現比率，F_X/N，F_Y/Nを用い

ざるをえない。このことは，四分表で考えれば，XとYの周辺度数を固定したものと仮定することを意味する。

この仮定のもとで，確率変数C_{XY}の確率分布を考えるとすれば，最も簡明な方法は二項分布による近似であろう。任意のユニットにXが出現する確率，あるいは，変数Xが，X＝1となる確率は，F_X/Nであり，任意のユニットにYが出現する確率，あるいはY＝1となる確率はF_Y/Nである。XとYが独立であるとすれば，XとYが任意のユニットにともに出現する確率，即ちX＝Y＝1となる確率は，

$$F_X F_Y / N^2 \qquad (2.12)$$

である。従って共出現度数C_{XY}が二項分布に従うものとすれば，N個のユニットにおいて，$C_{XY}=k$となる確率は，

$$P_r\{C_{XY}=k\} = {}_NC_k\left(\frac{F_X F_Y}{N^2}\right)^k\left(1-\frac{F_X F_Y}{N^2}\right)^{N-k} \qquad (2.13)$$

で与えられる。また，このときのC_{XY}の期待値$E(C_{XY})$と分散$V(C_{XY})$は，

$$\left.\begin{aligned} E(C_{XY}) &= F_X F_Y / N \\ V(C_{XY}) &= \frac{F_X F_Y}{N}\left(1-\frac{F_X F_Y}{N^2}\right) \end{aligned}\right\} \qquad (2.14)$$

となる。F_X，F_YがNに比べて十分に小さければ，二項分布ではなく，ポアッソン分布に近似して，

$$P_r\{C_{XY}=k\} = \frac{\left(\frac{F_X F_Y}{N}\right)^k e^{-\left(\frac{F_X F_Y}{N}\right)}}{k!} \qquad (2.15)$$

とすることもできる。このときの期待値も，二項分布の場合と同じである。あるいは，近似の適用限界の問題はひとまず措いて，

$$Z=(C_{XY}-F_XF_Y/N)/\sqrt{(F_XF_Y/N)(1-F_XF_Y/N^2)}$$

(2．16)

なる量Zを考えれば，平均0，分散1の規準型正規分布で近似することもできる。

ポアッソン分布，正規分布については，適用の制約の問題があるが，それを別にしても，上記三つの分布による近似では，共出現度数C_{XY}がXとYの関数であり，2．3式のような制約があるということが無視されているという点が問題である。簡単な例を挙げてみよう。今，$N=100$，$F_X=50$，$F_Y=2$ とすれば，このときC_{XY}のとる値の範囲は，2．3式より，

$$0\leqq C_{XY}\leqq 2$$

となる。ところが，2．13式，2．15式，2．16式においては，この制約は考慮されていないから，$C_{XY}\geqq 3$ なる場合も許容され，その確率$P_r\{C_{XY}\geqq 3\}$は，ほぼ次のようになる。

二項分布	$P_r\{C_{XY}\geqq 3\}=0.079$
ポアッソン分布	$P_r\{C_{XY}\geqq 3\}=0.080$
正規分布	$P_r\{C_{XY}\geqq 3\}=0.158$

従って，一般には，共出現度数C_{XY}の確率分布を，二項分布，ポアッソン分布，正規分布によって近似することはできないと言える。限られた条件のもとではこのような近似が可能なこともありえようが，一般には，2．3式の制約を考慮したより厳密な確率を与える分布を探る必要がある。

周辺度数F_X，F_Yが固定されているときのC_{XY}の厳密な確率は，Hartleyら(1979)の提案するように超幾何分布によらねばならない。Hartleyらの提案する超幾何分布は，本稿の対象としている共出現そのものではなく，その下位概念である言語要素間の順序のある共出現を対象としている点と確率計算の基礎となるサンプル・スペース，即ち，起りうるすべての場合の数のとり方に問題があるこ

との二点により，直ちにはこれを援用するわけにはいかない。[16]

Hartleyらは，言語要素XとYに関し，X＝ classe, Y＝ouvrièreとすると，資料中で$F_X=12$，$F_Y=33$，$C_{XY}=7$，$N=1399$である[17]ことから，次の方法で確率を計算している。今，X′なる要素がF_X個，Y′なる要素がF_Y個，S′なる要素が（N－1）個[18]あるとし，これらすべての要素からなる総順列を考えれば，その数は，

$$(F_X+F_Y+N-1)!/(F_X!\,F_Y!\,(N-1)!) \quad (2.17)$$

である。これは，言語要素X′とY′，そしてユニットの境界S′とが，資料中で完全にランダムに分布しうると仮定したときの順列の総数に他ならない。ここで，2.17式に与えられた数の順列のうち，例えば，

$$X'-Y'-S'$$

なる部分列（ないしは連鎖）をk個だけ含む順列の数を求めれば，$C_{XY}=k$ となる確率が計算できる。Hartleyらの場合には，C_{XY}は，XがYの前にあるという制約が満たされた場合だけを数えているので，本稿の議論との比較は直接にはできないが，例えば，'classe'と'ouvrière'が，この順で7回以上出現する確立は，実に10^{-9}程度と計算されている。

Hartleyらの方法で問題となるのは，2．17式で与えられた数の総順列のうちには，

$$\overbrace{X'X'X'\cdots\cdots}^{F_X\text{個}}\quad\overbrace{Y'Y'Y'\cdots\cdots}^{F_Y\text{個}}\quad\overbrace{S'S'S'\cdots\cdots}^{(N-1)\text{個}}$$

$$\overbrace{X'X'X'\cdots\cdots}^{F_X\text{個}}\quad\overbrace{S'S'S'\cdots\cdots}^{(N-1)\text{個}}\quad\overbrace{Y'Y'Y'\cdots\cdots}^{F_Y\text{個}}$$

といった言語要素の配列なり連鎖としては到底考えられないような順列が含まれているということである。

このような欠点は，次のようにすれば改めることができる。今，F_X個の要素

1と（N−F_X）個の要素0から成る集合X′があり，F_Y個の要素1と，（N−F_Y）個の要素0から成る集合Y′があるとする。X′のすべての要素から成る順列の総数は，

$$ {}_{N}C_{F_X} = N!/(F_X!(N-F_X)!) \qquad (2.18) $$

Y′のすべての要素から成る順列の総数は，

$$ {}_{N}C_{F_Y} = N!/(F_Y!(N-F_Y)!) \qquad (2.19) $$

従って，X′の順列から任意のひとつをとり，Y′の順列から任意のひとつをとったときの対の総数は，2．18式，2．19式より，

$$ {}_{N}C_{F_X}\ {}_{N}C_{F_Y} = \frac{N!N!}{F_X!(N-F_X)!F_Y!(N-F_Y)!} \qquad (2.20) $$

これは，N個のユニットから成る資料において，言語要素XとYがそれぞれF_X回，F_Y回出現するという制約のもとで，XとYが完全に独立に出没すると仮定したときの，可能な現われ方の総数である。

ここでX′の順列のうちの任意のひとつをとれば，要素1の占めるF_X個の位置（あるいは，Xの出現するユニット）は一意的に定まっている。このとき，Y′の任意の順列において，X′の任意のひとつの順列における上述の意味でのF_X個の定まった位置に対応する位置は当然F_X個あり，一意的に定まる。Y′の任意の順列におけるこのF_X個の定まった位置のうちのk個を，Y′の要素1が占めるとするならば，kは，言語要素XとYの共出現度数に他ならない。議論を簡単にするために，今，$F_X \leqq F_Y$とすると，Y′の任意のひとつの順列において，前述の意味での定められたF_X個の位置のうちのk個をY′の要素1が占める場合の数は，

$$ {}_{F_X}C_{k} = \frac{F_X!}{k!(F_X-k)!} \qquad (2.21) $$

また，この順列において，残りの（N−F_X）個の位置には，要素1が（F_Y-k）

個，要素0が（$N-F_Y-F_X+k$）個あるから，

$$ {}_{(N-F_X)}C_{(F_Y-k)} = \frac{(N-F_X)!}{(F_Y-k)!(N-F_X-F_Y+k)!} \qquad (2.22) $$

通りの部分順列がある。従って，$C_{XY}=k$となる場合の数は，X′の任意のひとつの順列について，2．21式，2．22式より，

$$ {}_{F_X}C_k \ {}_{(N-F_X)}C_{(F_Y-k)} = \frac{F_X!(N-F_X)!}{k!(F_X-k)!(F_Y-k)!(N-F_X-F_Y+k)!} \qquad (2.23) $$

通りある。X′の順列の総数は2．18式に与えられているから，これを用いれば，$C_{XY}=k$となる場合の数は，結局，

$$ {}_{N}C_{F_X} \ {}_{F_X}C_k \ {}_{(N-F_X)}C_{(F_Y-k)} = \frac{N!}{k!(F_X-k)!(F_Y-k)!(N-F_X-F_Y+k)!} \qquad (2.24) $$

となる。この2．24式と，2．20式を用いれば，$C_{XY}=k$となる確率$P_r\{C_{XY}=k\}$は，

$$ P_r\{C_{XY}=k\} = \frac{{}_{N}C_{F_X} \ {}_{F_X}C_k \ {}_{(N-F_X)}C_{(F_Y-k)}}{{}_{N}C_{F_X} \ {}_{N}C_{F_Y}} $$

$$ = \frac{{}_{F_X}C_k \ {}_{(N-F_X)}C_{(F_Y-k)}}{{}_{N}C_{F_Y}} \qquad (2.25) $$

または，

$$P_r\{C_{XY}=k\}=\frac{F_X!F_Y!(N-F_Y)!(N-F_X)!}{k!(F_X-k)!(F_Y-k)!(N-F_X-F_Y+k)!N!}$$

（2．26）

2．25式は，特性Xをもつ要素をF_X個を含む大きさNの有限母集団からF_Y個を選ぶとき，特性Xをもつ要素がk個含まれる確率を示す超幾何分布にほかならない。[19] これまで，$F_X \leqq F_Y$と仮定したが，$F_Y < F_X$とするならば，2．25 式において，F_XとF_Yを入れ替えればよい。2．26式は，こうしても同じである。

また，2．26式に着目して，この式の右辺のkをC_{XY}で置換すれば，この方法が四分表におけるフィッシャーの直接確率計算法と同一のものであることは明らかである。

共出現度数C_{XY}の確率密度が，2．25式もしくは，2．26式で与えられるとき，C_{XY}の期待値と分散は次の通りである。[20]

$$E(C_{XY})=F_X F_Y/N$$
$$V(C_{XY})=F_X F_Y(N-F_X)(N-F_Y)/N^2(N-1)$$

（2．27）

ここで，$C_{XY}=k$となる確率を，二項分布，ポアッソン分布，超幾何分布にもとづいて計算し，比較してみよう。N＝100として一定にし，$F_X=50$，$F_Y=2$，$F_X=20$，$F_Y=5$，$F_X=10$，$F_Y=10$とした3つの場合について，次の表2に示す。

表2から明らかなように，XとYの共出現度数C_{XY}の確率密度の分布は，二項分布，ポアッソン分布による場合と，超幾何分布による場合とでは相当に異なる。従って，厳密な確率が問題になるときには，特にNがあまり大きくなければ，共出現度数C_{XY}の確率分布は超幾何分布によって求めなければならない。ただ，実

際問題としては，超幾可分布にもとづく確率計算を手作業で行うのは非常に煩雑であり，Hartleyらのようにコンピューターを使うことになろう。しかし，超幾何分布にもとづく確率計算のプログラム自体はむしろ容易な部類に属し，筆者の使っているものでも，PL／Ⅰでせいぜい数十ステップである。

3. 共出現の指標

3.1 指標の満たすべき条件

本節では，ふたつの言語要素間の共出現の度合いを示す係数，指標そのものを検討する。その前提として，まず，共出現の指標が満たすことが望ましい条件を考察すれば，次のような条件が挙げられよう。

資料独立性，一般的適用可能性

資料，ユニット，言語要素の質，性格に拘わらず適用できること。

比較可能性

異なった資料間，言語要素間での比較が可能であるように，値が標準化されていること。

両方向性

関係の密接さ（関連の度合い）だけでなく，稀薄さあるいは排反の度合をも表現できること。

対 称 性

指標のとる値が，ふたつの言語要素に関して同一であること。

検定可能性

指標のとる値について，任意の有意水準を定めたとき，検定が容易であること。

表 2. 二項分布，ポアッソン分布，超幾何分布の比較

C_{XY}	二項分布* $F_X=50,\ F_Y=2$ $F_X=20,\ F_Y=5$ $F_X=10,\ F_Y=10$ いずれの場合も同じ	ポアッソン分布* $F_X=50,\ F_Y=2$ $F_X=20,\ F_Y=5$ $F_X=10,\ F_Y=10$ いずれの場合も同じ	超幾何分布 $F_X=50,\ F_Y=2$	超幾何分布 $F_X=20,\ F_Y=5$	超幾何分布 $F_X=10,\ F_Y=10$
0	0.366032	0.367879	0.247474	0.319309	0.330476
1	0.369730	0.367879	0.505050	0.420144	0.407995
2	0.184865	0.183940	0.247474	0.207343	0.201509
3	0.060999	0.061313	0	0.047848	0.051793
4	0.014942	0.015328	0	0.005148	0.007553
5	0.002898	0.003066	0	0.000205	0.000639
6	0.000463	0.000511	0	0	0.000005
7	0.000063	0.000073	0	0	**
8	0.000007	0.000009	0	0	**
9	0.000001	0.000001	0	0	**

* 二項分布と，ポアッソン分布の数値は北川・稲葉（1960）によった。

** 10^{-6} 以下であることを示す。

研究対象，研究目的によっては，このような条件は厳しすぎることもあろうし，不要であることもあろう。しかし，ここではそのような場合を個別的に論ずることはしないで，以上の５つの条件を，共出現の指標が一般的に満たすべき条件として論を進めることにする。

3.2 従来用いられた指標

従来の研究においては，様々な指標が共出現の指標として用いられているが，既に述べたように，厳密な定式化，方法的吟味は等閑に付されていると言っても過言ではない。そのなかで，比較的明示的に定式化されているものとしては，次のものがある。

Osgood（1959）が，共出現度数C_{XY}の期待値を用いて，ふたつの言語要素の関係を，「関連あり」と「関連なし」に二分していることは既に述べたが，Osgoodが実際に期待値として用いているのは，共出現度数そのものの期待値ではない。Osgoodは実際には，

$$F_X F_Y / N^2$$

という共出現比率の期待値を，実際の共出現比率（C_{XY}/N）と比較している。[21] 言うまでもなく，これは，実際の共出現度数をその期待値と比較することと同値である。Osgoodはこれとは別に，ふたつの言語要素の距離を測る方法も用いている。Osgoodによれば，XとYの距離Dは，

$$D=\sqrt{\Sigma d_i^2}=\sqrt{\Sigma(x'_i-y'_i)^2} \qquad (3.1)$$

で与えられる。出現度数が0と1に標準化されていれば，表1の四分表を用いて，

$$D=F_X+F_Y-2C_{XY} \qquad (3.2)$$

Dのとる値の範囲は，次のようになる。

完全関連	$D=0$	
最大関連	$D=F_Y-F_X$	$(F_X<F_Y)$
	$D=F_X-F_Y$	$(F_X>F_Y)$
最大排反	$D=F_X+F_Y$	$(F_X+F_Y-N<0)$
	$D=2N-F_X-F_Y$	$(F_X+F_Y-N>0)$
完全排反	$D=N$	

Dは，XとYが独立であるときの値が明確でないことと，前項で挙げた指標の条件のうち，比較可能性と検定可能性を満たさないことの2点が問題である。

Dは共出現度数C_{XY}を必ずしも必要としない指標であるが，次に共出現度数C_{XY}を用いる指標について考察することにしよう。このうちで最も簡明なものは，松尾・森・阿部（1979）等で用いられた非対称な共出現率である。まずXに対するYの共出現率$cor^*(YX)$は，

$$cor^*(YX)=C_{XY}/F_X \qquad (3\cdot 2)$$

またYに対するXの共出現率$cor^*(XY)$は，

$$cor^*(XY)=C_{XY}/F_Y \qquad (3\cdot 3)$$

で表わされる。類似のものとしては，McKinnon（1977）が用いている関連指数（association index）があり，これは，次のように定式化される。[22)]

$$\text{関連指数}=\frac{\alpha\Sigma C_i(x'_i,\ y'_i)}{\Sigma x'_i}=\frac{\alpha C_{XY}}{F_X} \qquad (3\cdot 4)$$

（αは定数で，$\alpha=10000\ T/2613$）

McKinnonは，出現度数として実測度数をそのまま用いている。3・4式で$\alpha=1$とすれば，これは3，2式に与えたXに対するYの共出現率に一致する。

このような非対称な指標もそれなりに限定された適用範囲を有するが，一般性を欠くので，ここではこれ以上触れない。これに対し，XとYに関して対称な共出現率は，日本電信電話公社電気通信研究所の「ＣＩＲＥＳシソーラス」(1970)，佐藤ら(1981)などで用いられており，上述のふたつの非対称な共出現率の幾何平均をとったものである。この対称な共出現率 cor_{XY} は，

$$cor_{XY} = C_{XY} / \sqrt{F_X F_Y} \qquad (3.5)$$

で表わされる。cor_{XY} の極値は次のようになる。

完全関連　$cor_{XY} = 1$

最大関連　$cor_{XY} = \sqrt{F_X / F_Y}$　(< 1)

$(F_X < F_Y)$

$cor_{XY} = \sqrt{F_Y / F_X}$　(< 1)

$(F_Y < F_X)$

最大排反　$cor_{XY} = 0$

$(F_X + F_Y - N < 0)$

$cor_{XY} = (F_X + F_Y - N) / \sqrt{F_X F_Y}$　(> 0)

$(F_X + F_Y - N > 0)$

完全排反　$cor_{XY} = 0$

共出現率 cor_{XY} は，出現度数をどのように決めても適用できるという利点があるが，他方，排反の度合いが明確に表示できないこと，直接に対応する検定法がないなど，前項に挙げた条件を満たさない欠陥がある。また最大排反でも，完全排反でも0になりうる点も問題である。なお，対称な共出現率に類似の指標としては，McKinnon (1977) が次のような指標を用いている。[23)]

$$\alpha C_{XY} / F_X F_Y \qquad (3.6)$$

(α は定数)

以上，従来用いられた共出現の指標の幾つかについて考察したが，本項で考察した諸指標は，もっぱら関連の度合いを測ることを主眼とするものであり，指標の両方向性という点での配慮に欠ける。また指標の比較可能性，検定可能性という点からしても問題のあるものが多い。従って，次項では従来統計処理に用いられてきた二変数の相関を示す指標のうち，共出現の指標として適切なものがあるかどうかを検討する。

3.3 四分表に適用される指標

二変数の関係を示す指標は非常に多いが,[24] XとYの出現度数が0と1に標準化されていれば，表1の四分表を使えばいいから，四分表に適用される指標を考察すればよい。これに対して，実測度数や出現比率をそのまま出現度数とする方法Ⅰ，Ⅱでは，四分表に適用される指標は原則的には適用できない。方法Ⅰ，ⅡによってFX，FY，CXYを求めたときに適用可能な指標は，ピアソンの相関係数r，スピアマンの順位相関係数ρ，相関比の三種であろう。ピアソンの相関係数rは，Charpentier(1978)がIris Murdochの作品 *Unicorn* における，prison, freedom, guilt 等々の単語の相互関係の指標として用いているが,[25]言語要素の出現度数や実測度数のように分布の型の明らかでない変量に適用することには問題があろう。この意味では相関比のほうが好ましいが，相関比はふたつの変数に関して非対称であることと，排反の定義ができないことにより，共出現の指標としては問題がある。このような消去法的議論に従えば，この三者のうちでは，スピアマンの順位相関係数ρが最も好ましいと言えよう。ρは，指標の満たすべき条件を満たし，検定も既成の数表を使うなり，t分布，正規分布に近似することにより行うことができる。またSPSSのような既成の統計パケッジを利用すればρの算出，検定とも極めて容易である。更に，スピアマンの順位相関係数ρは，後述のように，四分表に適用すれば四分点相関係数ϕに一致する。

方法Ⅰ，ⅡにもとづいてXとYの出現度数を求めたときには，このようにスピアマンの順位相関係数ρが，最も適切な指標と言えるが，方法Ⅰ，Ⅱ自体に問題があることは既に述べた通りである。

議論を四分表に戻そう。四分表に適用され，両方向性と対称性というふたつの条件を満たす指標としては，ユールの関連係数(coefficient of association)Qと，

四分点相関係数ϕが挙げられる。[26] なお，多分表で用いられるグッドマンとクラスカルの順序関連係数γは，四分表に適用すればQに一致し，クラマーの関連係数V，ケンドールの順位相関係数τ_b，スピアマンの順位相関係数ρ，ピアソンの相関係数γは，四分表に適用したとき，すべてϕに一致する。[27]

ユールのQ，四分点相関係数ϕは，表1の四分表を用いれば，次のように表わされる。

$$Q=\frac{C_{XY}(N-F_X-F_Y+C_{XY})-(F_X-C_{XY})(F_Y-C_{XY})}{C_{XY}(N-F_X-F_Y+C_{XY})+(F_X-C_{XY})(F_Y-C_{XY})}$$

$$=\frac{NC_{XY}-F_XF_Y}{NC_{XY}-F_XF_Y+2(F_X-C_{XY})(F_Y-C_{XY})} \qquad (3.7)$$

$$\phi=\frac{C_{XY}(N-F_X-F_Y+C_{XY})-(F_X-C_{XY})(F_Y-C_{XY})}{\sqrt{F_XF_Y(N-F_X)(N-F_Y)}}$$

$$=\frac{NC_{XY}-F_XF_Y}{\sqrt{F_XF_Y(N-F_X)(N-F_Y)}}$$ [28] $$(3.8)$$

まずユールのQについて，その極値を調べてみよう。Qの極値は次のようになる。

完全関連　$Q=1$
最大関連　$Q=1$
最大排反　$Q=-1$
完全排反　$Q=-1$

Qは明らかに最大関連，最大排反において定義される指標である。また，3.7式より，Qは，

$$C_{XY}=F_XF_Y/N$$

のとき，$Q=0$となる。2．14式または2．27式より，右辺はC_{XY}の確率分布の期待値であるから，独立もこれにより定義できる。

これに対して，ϕの極値は，

完全関連　$\phi=1$

最大関連　$\phi=\sqrt{F_X(N-F_Y)}\ /\ \sqrt{F_Y(N-F_X)}$　(<1)

$(F_X<F_Y)$

$\phi=\sqrt{F_Y(N-F_X)}\ /\ \sqrt{F_X(N-F_Y)}$　(<1)

$(F_Y<F_X)$

最大排反　$\phi=-\sqrt{F_X F_Y}\ /\ \sqrt{(N-F_X)(N-F_Y)}$　(>-1)

$(F_X+F_Y-N<0)$

$\phi=-\sqrt{(N-F_X)(N-F_Y)}\ /\ \sqrt{F_X F_Y}$　(>-1)

$(F_X+F_Y-N>0)$

完全排反　$\phi=-1$

となる。このことから明らかなように，ϕでは完全関連と最大関連，完全排反と最大排反が区別できる。この意味では，ϕはこの区別のできないQに比べて理論的には勝れていると言えよう。独立の定義は，ϕもQの場合と同様である。検定可能性という観点から両者を比較すれば，一般的検定法との対応のないQに対し，ϕは，

$$\chi^2=N\phi^2$$

なる関係を利用してχ^2検定を行うことが考えられるが，χ^2検定の制約条件を考慮すれば，必ずしもϕのほうが便利とも言い切れない。

ユールのQも，四分点相関係数ϕも，本節第1項に掲げた共出現の指標が満たすべき条件を，これまで検討した諸指標のうちでは最もよく満たす指標である。そして，ϕとQを比較すれば上述のようにϕのほうが好ましいと言うことができる。ここで，実用という観点から，ϕとQをもう一度比較してみよう。一般に指

標を実際の言語資料に適用した場合，指標は最大排反と最大関連の間の値しかとりえない。ユールのQは，常に最大排反で最小値－1，最大関連で最大値1をとるから，この点では問題はない。しかし，ϕのように，完全関連と完全排反の間の値をとるように定義された指標では，最大排反と最大関連の値の幅が場合により，極めて狭いものになることがある。次の表3は，N，F_X，F_Yを変えて，ϕが実際にとりうる値の範囲を示したものであるが，表3からも明らかなように，ϕの値の範囲は極めて小さいものになることがある。

表3. 四分点相関係数ϕの最小値と最大値

$N=50$, $(F_X, F_Y)=(5, 5)$ … $-0.111 \leqq \phi \leqq 1.000$
$N=50$, $(F_X, F_Y)=(5, 15)$ … $-0.218 \leqq \phi \leqq 0.509$
$N=50$, $(F_X, F_Y)=(5, 25)$ … $-0.333 \leqq \phi \leqq 0.333$
$N=50$, $(F_X, F_Y)=(15, 15)$ … $-0.428 \leqq \phi \leqq 1.000$
$N=50$, $(F_X, F_Y)=(15, 25)$ … $-0.654 \leqq \phi \leqq 0.654$
$N=50$, $(F_X, F_Y)=(25, 25)$ … $-1.000 \leqq \phi \leqq 1.000$
$N=100$, $(F_X, F_Y)=(5, 5)$ … $-0.052 \leqq \phi \leqq 1.000$
$N=100$, $(F_X, F_Y)=(5, 30)$ … $-0.150 \leqq \phi \leqq 0.350$
$N=100$, $(F_X, F_Y)=(5, 50)$ … $-0.229 \leqq \phi \leqq 0.229$

$(F_X, F_Y)=(m, n)$は$F_X=m$, $F_Y=n$または$F_X=n$, $F_Y=m$であることを示す。

この結果，表3の例からも明らかなように四分点相関係数ϕについては，一般には，ϕの値からだけでは，XとYの間にどの程度の関連があるか，あるいはどの程度の排反関係があるかは直ちには判断できないと結論せざるをえない。この意味ではユールのQのほうが共出現の指標としては遙かに適切であるということになる。これは言うまでもなく，ϕが完全関連，完全排反にもとづく，即ち，周辺度数F_X，F_Yを固定しない場合の指標であり，逆にQは最大関連，最大排反にもとづく，即ち，周辺度数F_X，F_Yを固定した場合の指標であることを別の角度から論じたものにほかならない。上述のϕとQの比較から，共出現の指標としては，最大関連，最大排反にもとづく，即ち周辺度数F_X，F_Yを固定した指標のほ

うが望ましいことは明らかである。このことは，また，2節2項において，C_{XY}の確率分布を，周辺度数F_X，F_Yを固定して超幾何分布によって求めたことと理論的に整合するものである。

ここで，ユールのQと四分点相関係数ϕの値をまたひとつ別の角度から検討してみよう。ふたつの言語要素の共出現の指標が関連，排反の様々な度合いを示しうるものでなくてはならぬことは言うまでもないが，その度合いにもとづいて，ふたつの言語要素の関係を幾つかのカテゴリーに分けることも場合によっては必要である。2・2のように，期待値を境に三分，ないしは二分するのは最も簡明ではあるが，実際には偶然の入り込む余地が大きく実用性はない。一般にはα，β（$\alpha<\beta$）ふたつの値を定めて，指標の値iと次のような形で比較することになろう。

$i<\alpha$　であれば　排反
$\alpha<i<\beta$　であれば，独立（ないしは無関係）
$i\geqq\beta$　であれば　関連

このとき，問題になるのは，α，βをどのように定めるかである。一般には，有意水準という形で，$P_r\{i\leqq\alpha\}$，$P_r\{i\geqq\beta\}$を，0.1なり0.05なり，0.01に設定する方法が行われる。所謂有意性の検定がこれである。ϕとQを用いて，具体例を検討してみよう。今，$N=1000$，$F_X=100$，$F_Y=300$とし，有意水準を0.01としよう。$P_r\{C_{XY}\leqq k\}\leqq 0.01$となる$k$を2.26式の超幾何分布によって求めると，$k=19$である。$C_{XY}=19$のときのQと$\phi$の値は，

$Q=-0.318$
$\phi=-0.008$

である。逆に，$P_r\{C_{XY}\geqq k\}\leqq 0.01$となる$k$は，超幾何分布によれば$k=41$である。$C_{XY}=41$のとき，Qと$\phi$の値は，

$Q = 0.264$

$\phi = 0.080$

である。有意水準0.01であるから，$k \leqq 19$ あるいは $k \geqq 41$ のとき，ふたつの言語要素XとYの間には他の要因が特に明白な場合は別として，強い排反関係，あるいは強い関連があると判断される。ところが，排反，関連いずれの場合も，Q，ϕの値は，通常であれば，XとYの間に特別の関連，排反はないと判断される値を示している。$Q \leqq -0.318$ となる確率，$\phi \leqq -0.008$ となる確率を求めてみなければこの場合のXとYの間に強い排反関係の存在することは判らない。もはや他の例を挙げるまでもなく，Qやϕの値のみからでは，XとYの関係については直ちには判断できないことは明らかであろう。そしていずれにせよ確率を求めるとすれば，直接確率にもとづいた指標のほうが合理的，労力節約的であることもまた明らかであろう。この意味で，次項において確率に直接もとづいた新たな指標である共出現係数が提案される。

3.4 共出現係数

ふたつの言語要素XとYの共出現度数C_{XY}の確率分布は超幾何分布にもとづいて求めなければならぬことは既に明らかにした。ここで，mを次のように定義する。

$$m = max\ (0,\ F_X + F_Y - N) \qquad (3.9)$$

また，pを，

$$p = P_r\{C_{XY} < k\} = P_r\{C_{XY} \leqq k-1\} \qquad (3.10)$$

とする。この確率は，超幾何分布によって与えられるから，2．26式より，

$$p=\sum_{i=m}^{n}\left(\frac{F_X!(N-F_X)!F_Y!(N-F_Y)!}{i!(F_X-i)!(F_Y-i)!(N-F_X-F_Y+i)!N!}\right)$$

（但し，$n=max$（m，$k-1$）とする。）　　（3．11）

ここで，

$$C'=p-\frac{1}{2}$$　　（3．12）

なる量C'を考えれば，$0\leqq P\leqq 1$より，

$$-\frac{1}{2}\leqq C'\leqq\frac{1}{2}$$　　（3．13）

$C=2C'$とおけば，3．12，3．13式より，

$$C=2P-1$$
$$-1\leqq C\leqq 1$$　　（3．14）

3．11，3．14式より，Cは，

$$C=2\sum_{i=m}^{n}\left(\frac{F_X!(N-F_X)!F_Y!(N-F_Y)!}{i!(F_X-i)!(F_Y-i)!(N-F_X-F_Y+i)!N!}\right)-1$$

（但し，$n=max$（m，$k-1$）とする。）　　（3．15）

このCを共出現係数と呼ぶことにする。共出現係数Cは，3．14式より，

$P<\frac{1}{2}$ であれば，$C<0$

$P=\frac{1}{2}$ であれば，$C=0$

$P>\frac{1}{2}$ であれば，$C>0$

となり，$P=\frac{1}{2}$，即ち$P_r\{C_{XY}\leqq k-1\}=\frac{1}{2}$ のとき，独立と定義される。$C=0$のときは，しかし，必ずしも$C_{XY}=F_XF_Y/N$（2・27）とはならないので，共出現係数Cによる独立の定義は，ユールのQ，四分点相関係数ϕによる独立の定義とは一致しない。また，Cの極値は，

最大関連　$C=1$
最大排反　$C=-1$

となり，ユールのQと同様，最大排反と最大関連で定義される指標である。

共出現係数Cは，本節1項で掲げた指標の満たすべき条件をすべて満たす。しかも，検定も確率を基礎としているのできわめて容易である。共出現係数Cが与えられたとき，$P_r\{C_{XY}\geqq k\}$は，

$$P_r\{C_{XY}\geqq k\}=\frac{1-C}{2} \qquad (3\cdot 16)$$

として容易に求められる。また，排反が問題になるときには，

$$P_r\{C_{XY}<k\}=P_r\{C_{XY}\leqq k-1\}=\frac{C+1}{2} \qquad (3\cdot 17)$$

を利用すればよい。またϕ，Qの場合と異なり，共出現係数Cでは，その値によって，ふたつの言語要素の関連と排反の度合を直ちに知ることができる。Cの値と有意水準の関係は次のようになり，極めて簡明である。

$C\geqq 0.98$　有意水準0.01で関連
$C\geqq 0.90$　有意水準0.05で関連
$C\geqq 0.80$　有意水準0.1　で関連
$C<-0.98$　有意水準0.01で排反
$C<-0.90$　有意水準0.05で排反
$C<-0.80$　有意水準0.1　で排反

排反の場合，等号が付かないのは，共出現係数Cのもとになる確率が，$P_r\{C_{XY} < k\}$であることによる。従って，有意水準αを定めたとき，$C < \alpha$となる共出現度数kが決まるから，実用的には，$C_{XY} \leqq k-1$であれば，有意水準αで排反と考えておけばよい。

また，$-0.8 < C < 0.8$であれば，ふたつの言語要素の間にとくに有意味な関連，排反はないと考えることができる。

ここで，実際のデータが与えられたとき，共出現係数Cがどのような値をとるか，実例を検討してみよう。次の表4は，$N = 1000$，F_XとF_Yの対が，50, 300のときの，C_{XY}のとりうる値のすべてについて，共出現係数Cの値と，比較のため，共出現率cor_{XY}，ユールのQ，四分点相関係数ϕ，それにx^2の値を併せ示したものである。また図2は，同じデータについて表4をグラフにしたものである。

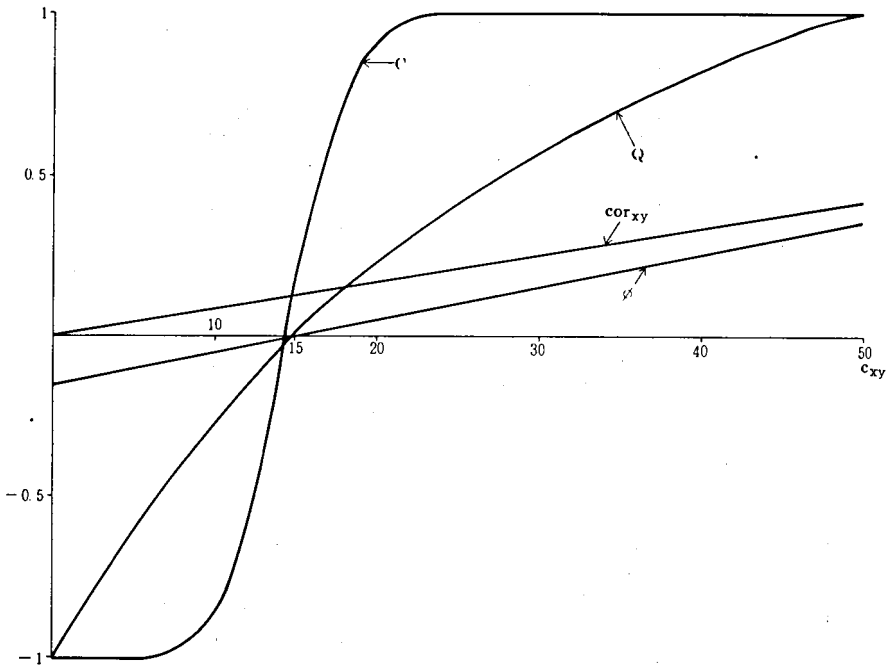
表4，図2から，資料によってはcor_{XY}，ϕは極めて限られた範囲の値しか示しえず，共出現の指標としては適当でないことが明らかであろう。残るユールのQと共出現係数Cの値を比較すると，$C_{XY} \leqq 2$または$C_{XY} \geqq 27$でCの絶対値は1にきわめて近くなるが，Qは，この範囲でも，関連，排反の度合いの相違，例えば$C_{XY} = 30$と$C_{XY} = 40$との相違を示すことができる。しかしユールのQは前述のように，有意水準あるいは確率との直接の対応関係がないという欠陥がある。この意味では，共出現係数Cのほうが勝れていると言えるが，共出現係数Cとユールの Qのいずれをとるかは実際には研究目的によって決まるであろう。

表 4. 共出現指標の値

(N = 1000 , (F_X, F_Y)=(50, 300))

C_{XY}	cor_{XY}	Q	ϕ	x^2	C
0	0.000	− 1.000	− 0.150	22.56	− 1.000
1	0.008	− 0.914	− 0.140	19.65	− 1.000
2	0.016	− 0.832	− 0.130	16.94	− 1.000
3	0.024	− 0.753	− 0.120	14.44	− 0.999
4	0.032	− 0.677	− 0.110	12.13	− 0.998
5	0.040	− 0.604	− 0.100	10.03	− 0.995
6	0.048	− 0.533	− 0.090	8.120	− 0.987
7	0.057	− 0.465	− 0.080	6.416	− 0.967
8	0.065	− 0.399	− 0.070	4.912	− 0.926
9	0.073	− 0.335	− 0.060	3.609	− 0.852
10	0.081	− 0.274	− 0.050	2.506	− 0.734
11	0.089	− 0.215	− 0.040	1.604	− 0.566
12	0.097	− 0.158	− 0.030	0.902	− 0.354
13	0.106	− 0.104	− 0.020	0.401	− 0.110
14	0.114	− 0.051	− 0.010	0.100	0.140
15	0.122	0.000	0.000	0.000	0.375
16	0.130	0.049	0.010	0.100	0.575
17	0.138	0.096	0.020	0.401	0.730
18	0.147	0.145	0.030	0.902	0.841
19	0.155	0.186	0.040	1.604	0.912
20	0.163	0.229	0.050	2.506	0.955
21	0.171	0.270	0.060	3.609	0.979
22	0.179	0.310	0.070	4.912	0.990
23	0.187	0.348	0.080	6.416	0.996
24	0.196	0.385	0.090	8.120	0.998
25	0.204	0.421	0.100	10.03	0.999
26	0.212	0.455	0.110	12.13	0.999
27	0.220	0.488	0.120	14.44	1.000
28	0.228	0.520	0.130	16.94	1.000
29	0.236	0.551	0.140	19.65	1.000
30	0.244	0.581	0.150	22.56	1.000
31	0.253	0.610	0.160	25.66	1.000
32	0.261	0.637	0.170	28.97	1.000
33	0.269	0.664	0.180	32.48	1.000
34	0.277	0.690	0.190	36.19	1.000
35	0.285	0.715	0.200	40.10	1.000
36	0.293	0.739	0.210	44.21	1.000
37	0.302	0.762	0.220	48.52	1.000
38	0.310	0.785	0.230	53.03	1.000
39	0.318	0.806	0.240	57.74	1.000
40	0.326	0.827	0.250	62.66	1.000
41	0.334	0.848	0.260	67.77	1.000
42	0.342	0.867	0.270	73.08	1.000
43	0.351	0.886	0.280	78.60	1.000
44	0.359	0.904	0.290	84.31	1.000
45	0.367	0.921	0.300	90.23	1.000
46	0.375	0.938	0.310	96.34	1.000
47	0.383	0.954	0.320	102.7	1.000
48	0.391	0.970	0.330	109.2	1.000
49	0.400	0.985	0.340	115.9	1.000
50	0.408	1.000	0.350	122.8	1.000

図2. ϕ, Q, cor_{XY}, Cの値の変化
(N=1000, (F_X, F_Y)=(50, 300))



結　び

本稿では，ふたつの言語要素の共出現の指標について，具体的資料から離れて，一般的方法論的吟味を加え，共出現係数という新たな指標を提案した。個々の具体的資料にもとづく検討は今後の課題である。また，ふたつの言語要素の共出現の指標に関する方法論的問題はほぼ尽しえたと思うが，三つ以上の言語要素間の関係については論じえなかった。二要素間の関係から，多要素間の関係への拡張は，今後の検討課題のひとつである。

註

1) 具体的な適用例としては，水谷(1976 a)，水谷(1976 b)，松尾ほか(1979)などを参照。

2）現実の文献や文書をそのままの形で，即ち自然言語のままの形で，対象とすることも皆無とは言えまいが，実際問題としては，この場合にも，言語要素の何らかの規準化（lemmatization），カテゴリー化等は，避けて通るわけにはいかない。註6参照。

3）所謂連語（collocation）や慣用句が2語から成るときが前者の例であり，韻文の二連行節（couplet）における脚韻語の対は後者の例である。前者の例としては，Vikis-Freibergs & Freibergs(1978)による，ラトヴィアの太陽神話における2語からなる定型表現の研究がある。後者については，Phelan（1979）（p. 68）が簡単な例を挙げている。更に脚韻における共出現関係に着目したものとしては，Kumamoto（1981）pp. 16－17に Tennyson の *In Memorium* における豊富な例がある。ただこれらの論文においては，言語要素の共出現関係は，暗黙裡に前提にされているが，共出現という概念が明示的な形で用いられているわけではない。

4）厳密には，「Xなり Yなりのトークン（token）が，ユニット i の部分列を成す」等と，トークンの概念を用いて定義すべきであろう。

5）同様の定義については水谷（1977）註20を参照。なお，水谷（1976 a）は，言語が単語であるときの共出現の厳密な定義を与えている。

6）問題となる言語要素がどのように定義されるか，またそのときどのような問題があるかについては，必ずしも共出現に着目しないにせよ，所謂内容分析（content analysis）の手法を用いる研究が参考になる。このような問題に関しては，決定ないしはコーディングの信頼性の問題も含め，Charpentier（1978）pp. 23－24，Manheim（1979）p. 59，武者小路（1977）pp. 152－157，Riggs（1977）pp. 538－539など参照。なお，Charpentier（1978）は，内容分析の手法とは関係なく，主として言語学，テクスト言語学の観点からの議論である。

7）例えば武者小路（1972）（pp. 183－185）は，出現度数をこの方法で算定している。

8）関数 $b(u, v)$ については，一般に，

$$b(u, v) = b(u-\beta, v-\beta) \qquad (1')$$

また，$\alpha > 0$ であれば，

$$b(u, v) = b(\alpha u, \alpha v) \qquad (2')$$

が成り立つ。従って(1′)(2′)より，

$$b(u, v) = b(\alpha(u-\beta), \alpha(v-\beta)) \qquad (3')$$

今，$u = x_i / t_i$，$v = \bar{X}_P$，$\beta = \bar{X}_P$ とすると，まず，1．9式と(1′)より，

$$x_i' = b(x_i/t_i, \bar{X}_P) = b(\{x_i/t_i - \bar{X}_P\}, 0)$$

ここでαを

$$\alpha = 1/\sigma_X \text{（但し，} \sigma_X = \sqrt{\frac{1}{N}\Sigma(x_i/t_i - \bar{X}_P)^2}\ )$$

とすれば，$\alpha > 0$ だから，

$$x_i' = b\left(\frac{x_i / t_i - \bar{X}_P}{\sigma_X}\ ,\ 0\right)$$

ここで，$(x_i / t_i - \bar{X}_P) / \sigma_X$ は，平均0，分散1に規準化された所謂Zスコアにほかならない。y_i' についても同様である。

9) McKinnon (1977) は，キエルケゴールにおける 'systemet' と他の語の共出現度数を求めるとき，X = systemet, Y = 他の任意の単語，とすれば，第 i ユニット（この場合，文）において，

$x_i = 0$ であれば，$C_i(x_i', y_i') = 0$
$x_i \geqq 1$ ならば，$C_i(x_i', y_i') = y_i$

という方法を用いている。しかし，McKinnon のデータ（p. 149）から判断する限り，$x_i \geqq 1$ のときには，ほとんどの場合 $x_i = y_i = 1$ と推測されるから，実質的には1．11式を用いたものと言える。

10) 方法Ⅰ，Ⅱによる場合も，α を適当に定めて，

$x_i' < \alpha$ ならば $x_i' = 0$
$x_i' \geqq \alpha$ ならば $x_i' = 1$

等とすればよい。

11) Osgood (1959) pp. 62－64.

12) 2．4，2．5のいずれかの一方で，$m = 0$ あるいは $n = 0$ であるとすれば，常に

$$x_i' = y_i'$$

となり，一方だけで完全関連の定義となる。

13) 安田・海野 (1977) p. 20 参照。但し，同書で「負の関連」と呼ばれている状態に対し，本稿では言語要素間の関係という意味から，「排反」という用語をあてている。

14) 安田・海野 (1977) pp. 20－21 参照。

15) 完全排反であれば，2．10式より，

$$F_X + F_Y = N$$

であるが，これは本来出現比率のあまり大きくない言語要素間に課する条件としては，相当に無理な条件である。

16) 記号を次のように定めるならば，共出現関係は，順序の有無，隣接性によって，4つの場合に下位区分される。

X，Y……言語要素
V_i ……不在でもありうる任意の可変的言語要素もしくはその連鎖
#……ユニットの境界

4つの下位区分は次の通りである。

順序あり，隣接
$\#-V_1-X-Y-V_2-\#$
順序なし，隣接
$\#-V_1-X-Y-V_2-\#$
または，
$\#-V_1-Y-X-V_2-\#$
順序あり，非隣接
$\#-V_1-X-V_2-Y-V_3-\#$
順序なし，非隣接
$\#-V_1-X-V_2-Y-V_3-\#$
または，
$\#-V_1-Y-V_2-X-V_3-\#$

本稿で論じているのは，最後の順序なし，非隣接の場合であり，これは前三者をすべて含む最も広い規定である。Hartleyら(1979) が対象としているのは二番目，即ち，順序あり，非隣接というカテゴリーである。

17) Hartley(1979) pp. 243－244 の例による。なお，資料はフランスの労働組合の大会宣言であり，ユニットは文(sentence)である。但し，この文は，文法的言語学的カテゴリーとしての文(structural sentence)というより，むしろ記法上，表記上の文(orthographic sentence)である。この点については，同論文 p. 242 参照。

18) Nではなく，(N－1)とするのは，S′はユニットの境界であるから，常にひとつのS′が順列の最後の位置になくてはならないという制約による。

19) 例えば， Fraser(1958) pp. 53－55，芝(1976) pp. 27－28 参照。

20) 芝(1976) p. 28.

21) Osgoodは更に比率の標準誤差を用いて検定を行っている。

22) McKinnon自身は定式化していない。 McKinnon(1977) p. 147 参照。

23) McKinnon (1977) pp. 149－150.

24) 池田(1976) pp. 156－157，安田・海野(1977) pp. 54－55など参照。

25) Charpentier (1978)は，スピアマンの順位相関係数とピアソンの相関係数の両方に言及しているが，例として挙げられたデータ(p. 25，p. 27)については，どちらを使用しているのか明らかでない。

26) 池田(1976)第5章，安田・海野(1977) 第1章3節による。なお，四分点相関係数の記号としてϕを用いることには学者により用法の違いが見られるが，ここではϕを用いることにする。

27) 但し，正確には絶対値が一致すると言うべき場合もある。

28) 3・8式は，四分点相関係数の定義からして，相関係数 r の定義式に，表1の度数を代入しても導くことができる。r は，

$$r=\frac{N\Sigma x_i' y_i'-(\Sigma x_i')(\Sigma y_i')}{\sqrt{(N\Sigma x_i'^2-(\Sigma x_i')^2)(N\Sigma y_i'^2-(\Sigma y_i')^2)}}$$

XとYが0と1の2値のとき，1．12，1．13式より

$$\Sigma x_i' y_i' = \Sigma C_i (x_i' , y_i') = C_{XY}$$

また，

$$\Sigma x_i'^2 = \Sigma x_i' = F_X , \quad \Sigma y_i'^2 = \Sigma y_i' = F_Y$$

従って，

$$\gamma = \frac{N C_{XY} - F_X F_Y}{\sqrt{(N F_X - F_X^2)(N F_Y - F_Y^2)}}$$

$$= \frac{N C_{XY} - F_X F_Y}{\sqrt{F_X F_Y (N - F_X)(N - F_Y)}}$$

四分点相関係数は，二変数がともに0と1の2値のときの相関係数 r にほかならないから，

$$\gamma = \phi = \frac{N C_{XY} - F_X F_Y}{\sqrt{F_X F_Y (N - F_X)(N - F_Y)}}$$

安田・海野（1977）pp. 23－24参照。

引　用　文　献

Carroll, John B. (1970) 'An Alternative to Juilland's Usage Coefficient for Lexical Frequencies, and a Proposal for a Standard Frequency Index (SFI)', *Computer Studies in the Humanities and Verbal Behavior,* Vol. 3, No. 2, pp. 61–65.

Charpentier, Colette (1978) 'Thematic Analysis: A Linguistic View of the Methodological Aspects of Quantification', *Ass. for Literary and Linguistic Computing Bulletin,* Vol. 6, No. 1, pp. 23–27.

Fraser, D.A.S. (1958) *Statistis: An introduction,* John Wiley & Sons.

Hartley, Anthony. F, Lafon, Pierre, and Tournier, Maurice (1979) 'A New Lexicometric Approach to Co-occurrences in a Text', *Ass. for Literary and Linguistic computing bulletin,* Vol. 7, No. 3, pp. 238–247.

池田央（1976）．統計的方法Ⅰ：基礎，新曜社。

北川敏男・稲葉三男（1960）．統計学通論，共立出版。

Kumamoto, Sadahiro (1981) The Structures of Rhyme Words in Tennyson's *In Memorium,* 熊本大学修士論文。

Manheim, Jaral. B. (1979), 'The Honeymoon's Over: The News Conference and the Development of Presidential Style', *Journal of Politics,* Vol. 41, No. 1, pp. 55–74.
松尾雅嗣・森祐二・阿部耕一朗(1979)「文献情報にみる軍事問題研究」 平和研究, Vol. 4, pp. 153－164.
McKinnon, Alastair (1977) 'From Co-occurrences to Concepts', *Computers and the Humanities,* Vol. 11, No. 3, pp. 147–155.
水谷静夫(1976ａ)「共出現関係に拠る語彙分類の試み」, 計量国語学, No. 77, pp. 1－13.
〃 (1976ｂ)「語の共出現に拠る語彙構造探求の諸法」, 計量国語学, No. 79, pp. 1－18.
〃 (1977) 「語彙の量的構造」, 岩波講座日本語９, pp. 43－86.
武者小路公秀(1972) 行動科学と国際政治, 東京大学出版会。
Osgood, C.E. (1959) 'The Representational Model and Relevant Research Methods', *Trends in Content Analysis (ed.I. de Sola Pool,)* Univ. of Illinois Press (1959)
Phelan, Walter S. (1978) 'The Study of Chaucer's Vocabulary', *Computers and the Humanities,* Vol. 12, No. 1/2, pp. 61–70.
Riggs, Robert E. (1977) 'One Small Step for Functionalism: UN Participation and Congressional Attitude Change', *International Organization,* Vol. 31, No. 3, pp. 515–539.
佐藤雅之他(1981)「キーワード自動選択システムの開発」, 第18回情報科学技術研究集会予稿集
芝祐順(1976) 統計的方法Ⅱ 推測, 新曜社。
Vikis-Freibergs, Vaira and Freibergs, Imants (1978) 'Formulaic Analysis of the Computer Accessible Corpus of Latvian Son-songs', *Computers and the Humanities,* Vol. 12, No. 4, pp. 329–339.
安田三郎・海野道郎(1977) 社会統計学(改訂２版), 丸善。