計　算　意　味　論　の　実　験*

西　村　恕　彦**　岩　坪　秀　一**

Abstract

The distances between words in natural sentences were measured by the cooccurrence frequency and were represented by a matrix. Eigenvectors, obtained by solving the equation $Ay=\lambda By$, gave coordinate values of words and sentences placed in a semantic space obtained. The constellation ot these points quite agreed with intrinsic properties of the linguistic phenomena sampled here. The approach promises well for the future of the computational semantics.

1. はじめに

言語事象を要素の連鎖としてとらえる理論はチョムスキらによってよく解明されている．われわれもまた自然語や人工語の翻訳システムを作って，文法的処理の実験をいくつか行なった．それに反して，要素の連鎖としてではない言語事象の模型化，文法論を越えた言語情報の処理方法については，これといって見るべき理論も実験もなかったようである．ここに報告するのは，文や単語の**意味**を純粋に計算機的な手続きによって数値化して表現する理論模型と，その一応用例である．

2. 結果の図示

一群の英語の単語の意味をこの手法によって数値化して，2次元の意味空間に投影した最終結果が**Fig. 1**である．これは計算機関係の論文からの文章と，現代小説からの文章とを抽出して処理した例であるから，現われた単語も特異なものが多いが，それはそれなりに納得できる結果が得られている．

この意味空間においては，類似した単語どうしは互いに集まりあい，異質な単語どうしは離れる．原点は

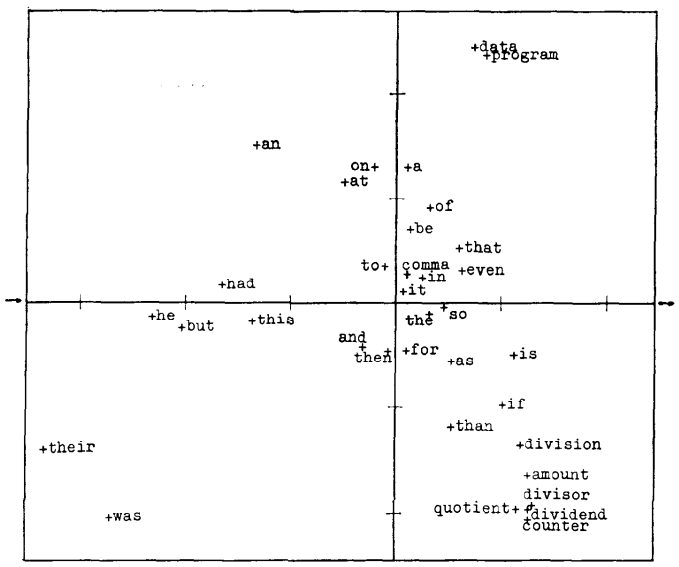

Fig. 1　Configuration of Words in Semantic Space.

* An Approach to Computational Semantics of Natural Languages, by Hirohiko Nisimura and Shuichi Iwatsubo (ETL)

** 通商産業省・電気試験所・情報基礎研究室

中性点であって，単語などの座標は意味のベクトルとして解釈できる．つまり，ベクトルの方向が意味のカテゴリーに相当し，左右，上下などの次元がそれぞれの意味分類に対応する．

ベクトルの長さ，つまり原点からのへだたりが，その方向への意味の強さ，あるいは純粋さを表わす．これによってこの意味空間を観察してみると，空間の左右（第1軸）の方向は their, he, but などの単語に対する，divisor, program など計算機技術の術語の対立を示している．

この意味空間の上下(第2軸)の方向は，右象限（第1軸の正方向）においてだけはっきりしており，それは，data, program など計算機ソフトウェアの術語に対する，counter, dividend など計算機ハードウェアの術語の対立である．

この意味空間の手前奥行（第3軸）の方向は，左象限(第1軸の負方向)において分化しており，それは，then, their, and などに対する，on, this, was などであって，二つの小説の文体の相違を示唆していると考えられる（第3軸は図示してない）．

これらの単語群における意味分化の投影に対応して一つ一つの文もまた同様の意味空間[注1]に投影できる．それが **Fig. 2** である．この図では一つ一つの文について算出された数値が意味空間中の点としてそれぞれ表示されている．各文はその属する層（後述）別に違う形の点で示してある．

ここでも第1軸は小説対計算機技術，第2軸は計算機ソフトウェア (s) 対計算機ハードウェア (f)，第3軸（図示してない）は二つの小説（o と x）の文体の対立などをよく表現している．

このように，ここで報告する数値化の手続きの特徴は，単語の集合としての自然語の文章をそのまま入力として，まったく計算機的な処理だけによって多次元の空間を設定し，その空間中の点として単語や文を位置づけることである．そしてそれらの単語や文を表わす点の位置が，われわれの直観的に認識している単語や文などの**意味**の類似や分化，対立とよく合致して納得できるということである．

3. 相伴の概念

言語事象についてその意味を測定するための，原始データの単位としては，いろいろの大きさのものを考えることができるが，ここではとりあえず「単語」と「文」とを選ぶことにした．単語は辞書的な意味をになう単位としては手ごろなものである．文章中では文字の列として表記されていて，欧文の文章では前後を空白で仕切られていることも，形式的な取扱いに便利である．

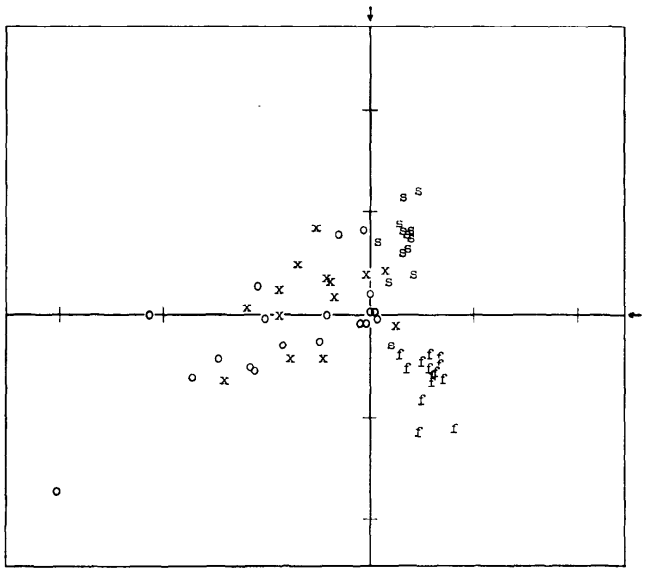

Fig. 2 Configuration of Sentences in Semantic Space.

注 1) ただし回帰の影響をうけて，文の座標値の分散は単語の座標値の分散よりも小さくなっているので，二つの図を直接に重ね合わせることはできない．分散比は横軸では 0.430，縦軸では 0.341 である．

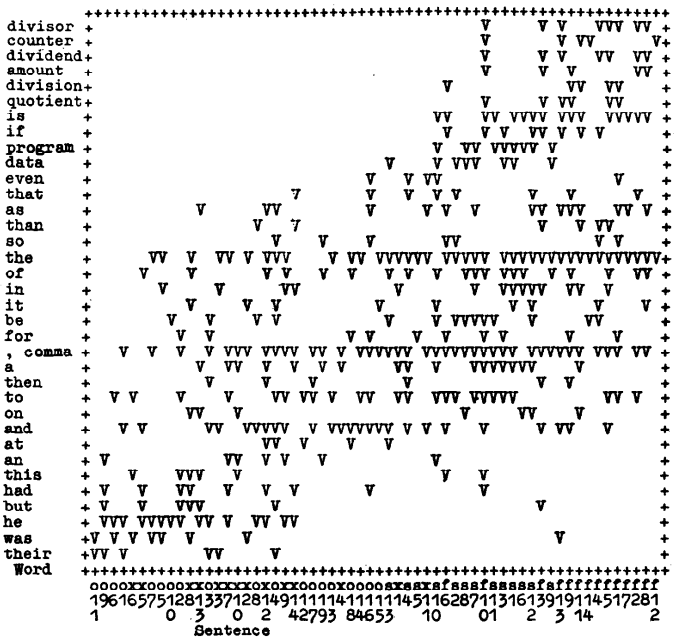

Fig. 3 Joint Distribution Matrix of Words and Sentences.

次の単位としていくつかの単語からなる集合をとらなければならないのだが，その単位として「文」をとった．すなわち情報を伝達するための文はある単語群を好んで選択する，と考える．そして普通に表記された文を検査して，ある単語がどの文とどの文とに含まれているか（V）を表示した相伴表（**Fig. 3**）を作る．

ただし実際の処理では，計算機の記憶容量の制約から単語数をいくらでも大きくとることはできない．そこである任意の基準にしたがって，少数個の標本単語をあらかじめ決めておき，それらが各標本文に含まれているかどうかを相伴表に記入する．一つの文のなかに相伴して使われることの多い単語群は互いに類縁な関係にあると考える．

ある一群の単語がある文のなかに同時に含まれたり，あるいはまったく含まれなかったりするふるまいのパターンが，この一群の単語について共通であるならば，それらの単語は互いに類縁な一集団を形成するものと考えられる．またある単語が，ほかのどの単語とも好き嫌いなく相伴してどの文にも含まれるならば，その単語は無性格な中性的なものとみなせよう．このようにして，ある語彙のなかの単語どうしに，文のなかでの相伴出現という観点からの類似度を想定し，語彙の意味的構造を解析し，単語を類別することができる．

相伴表の一方の軸には単語，他方の軸には文があるのを，それぞれ適当な順序にならべかえて，相伴する単語どうしは近く，相伴しない単語どうしは遠くになるようにする．その結果は Fig. 3 に示すとおり，相伴の印（V）が対角線状に集中する．こうして得られた配列順序は，単語および文の類別の第1因子の値の大小を反映していることになる．

この処理手続きは，実際には全く計算機によって実行されるのであり，しかも単語や文は配列順序が定められるのではなくて，数値を算出されるのである．Fig. 3 の縦軸の単語は実は Fig. 1 の第1軸の値の大小順に配列したものであり，同様に Fig. 3 の横軸の文は Fig. 2 の第1軸の値の大小順に配列したものである．

したがってこの相伴表における単語や文の配列順序がさきに述べた小説対計算機技術の対立をそのまま再現していると同時に，小説の文がそれぞれに固有な単語を含み，計算機関係の文がその術語を含むという，単語と文との相伴関係をもその第1因子の観点から表現していることになる．

4. 文の模型と標本抽出

言語事象の一例としての文章を次のように形式化する．

（1） 前後を空白で囲まれた，空白以外の文字（記号類を含む）からなる文字列を単語とする．単語は語頭からの 18 字までの綴りと文字列の長さだけで識別される．

（2） いくつかの単語がならんで，終止符で区切られたものを文とする．

（3） すべての単語（異なり語）のうち，任意に選ばれたいくつかを標本単語とする．

（4） 現在のプログラムでは，標本単語の異なり語数の上限を 100 語（計算機の記憶容量としては 100×100×2 語となる）とする．文の個数には実質上の制限はない．

（5） 一つの文のなかの単語を次のように分ける．

（a） その文のなかで1回だけ出てくる標本単語

（b） その文のなかに2回以上重出する標本単語

（c） 標本単語以外の単語

標本単語以外の単語はすべて空とされ，この数量化の手続きにおいてはまったく評価されない．一つの文のなかでの標本単語の重出の情報も切り捨てられる．すなわちある標本文のなかに，それぞれの標本単語が含まれているかいないかの情報だけが利用される．

言語事象から標本としてとる文章はある母集団をよく代表するものでなければならない．Osgood[2]は適当な母集団からとられた語については，国籍や分野を越えて共通の要因が見出だされると主張し，その要因のことを「意味」とよんでいる．しかしここでは最初の実験であるので，要因が出現しやすいようにあらかじめ人為的にかたよりを設定した標本文を選ぶことにした．外部的にあるかたよりを与えられた一群の文のことを「層」とよぶことにする（層の設定は，この数量化の手法においては必須のものではない．結果の解釈を容易にするために層を設定したが，一般的な言語事象を論ずるためには，こういうかたよりはないほうがよい．）．

対象とする国語としてはいちおう日本語と英語とを考慮した．多義語の機械翻訳などの応用性を考えると日本語の語彙の構造解析のほうが必要度は高い．しかし鍵盤穿孔や計算機入力がめんどうなので，この実験ではとりあえず英語の文章を選んだ．英語の文章は，**Table 1** に示すとおり，四つの異なった層[注2)]からそれぞれほぼ同じ大きさになるように連続的にとった．それらの層には記号 (s, f, o, x) を与えた．この記号は本報告中では統一して用いられる．

Table 1. Sampled Sentences

Group	Subject	Number of Sentences	Running Words	Different Words
s	Computer software	13	291	247
f	Computer hardware	14	375	293
o	Steinbeck's novel	19	285	261
x	Bellow's novel	14	303	259
Total		60	1,254	483

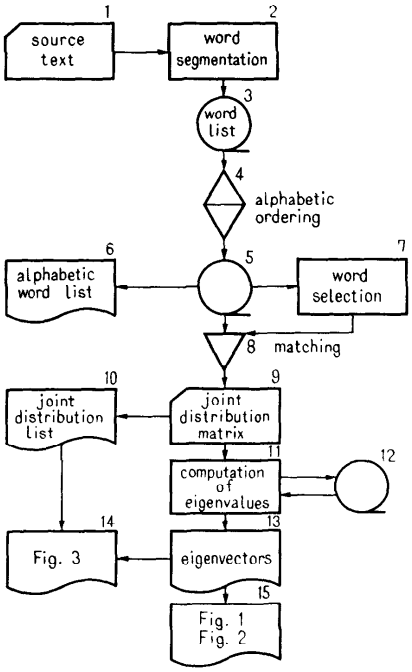

Fig. 4 Flowchart.

5. 処理手続き

選定された英語の標本文を 80 欄カードに穿孔して計算機の入力とした．そして以下の手順 (**Fig. 4**) にしたがって処理し，結果を各標本単語および各標本文の固有ベクトルの値(意味の因子負荷量)として得た．これらの値はさきに示した意味空間での座標値である．

注 2)
s……Computer software
P. Sherman: Programming and Coding Digital Computers, from page 407.
f……Computer hardware
W. Buchholz, ed.: Planning a Computer System, from page 117.
o……Steinbeck's novel
J. Steinbeck: Molly Morgan, from page 53.
x……Bellow's novel
S. Bellow: The Gonzaga Manuscripts, from page 53.

Table 2. Program Statistics (by Facom 230-50)

	Function	Language	Source Program Lines	Object Program Words	Time*
(2)	Word Segmentation	COBOL	185	8,322	5m 30s
(4)	Alphabetic Ordering	SORT	5	?	0m 19s
(6)	Alphabetic Listing	COBOL	41	3,802	3m 49s
(7, 8)	Selection and Matching	COBOL	273	10,582	6m 01s
(11)	Computation of Eigenvalues	FORTRAN	236	?	5m 44s
				Total	21m 23s

* Compiling, loading and execution

(1) 標本文の鍵盤穿孔　文章をカードに穿孔する．各単語は空白で区切っておく．文末の終止符は取り除き，文の区切りとして通貨記号 ($) を入れた．

(2) 単語の切り出し　文のなかの単語を切り出して，各単語の綴りとそれが含まれていた文の番号とを出現順に単語テープ(3)に書き出す (COBOL)．

(4) 単語のならべかえ　単語を ABC 順にならべかえる (SORT)．

(7) 標本単語の選定　ABC 順になった単語テープ (5) を読み込んで同一文中での重出単語を除く．各単語の拡がり (いくつの文に含まれたか) を累計する．拡がりの大きいものから順に標本単語として選ぶ (COBOL)．標本単語は必ずしも拡がりの順に選定する必要はない．しかし拡がりの小さい単語は出現度数も小さいので標本誤差が大きいと考えられる．この実験で選んだ 35 語のそれぞれの拡がりは 41～5 であった．

(8) 照合　標本単語のリストを，文番号の入っているテープ (5) と照合し，各標本単語がどの標本文に含まれているかの相伴表 (9) を作って書き出す (COBOL)．

(11) 固有値の計算　回転法の改良形[7] によって固有値を求める．正規化された固有ベクトルを各標本単語にたいしてまず算出し，次にこれを相伴表のテープ (12) に掛けて各標本文の正規化された固有ベクトルを得る．

これらに要した時間その他の資料を **Table 2** に示す．

6. 結果と解析

こうして算出された意味空間に各標本単語や各標本文を位置づけたのが，前出の Fig. 1 および Fig. 2 である．その具体的な視察はすでに述べたが，以下に全体的な考察を述べる．

(1) 意味空間の原点の近くにくる単語は，使用度数が高く，かつどの文にも平等に含まれるものである．品詞としては冠詞，前置詞，代名詞などであって，構文的・形式的な性質が強く，文や層の性格を特徴づけない．ここで二つの be 動詞，was と is とが分離していることに注目せよ．前者は小説 x，後者は計算機技術の文を特徴づけている．

(2) 原点から遠い位置にある単語は，文によって使用の度合に著しいかたよりのあるものである．したがって文や層の性格を強く特徴づける．じっさい，ある層に属する文に集中的に出現する単語ほど周辺部に位置している．このことは，これらの単語があるかたよりをもった情報，すなわち意味を伝達していることを示すものである．これらの単語およびその属する文や層がどの座標軸の方向に沿って原点から離れているかを吟味すれば，それぞれの座標軸の性格も推定されよう．

(3) さきに述べたように，各層の文および単語の性格は，定性的には第 1，第 3 軸の値の正負の組み合わせで示される注3)．これを **Table 3** に示す．

(4) 意味空間 (Fig. 1, Fig. 2) 中の狭い領域に集中している単語 (たとえば dividend, divisor, quotient) は，今回の標本において類似のふるまいかたをして各文中に出現しているものであって，したがって

注 3) 各層を特徴づけている単語を多く含んでいる文をそれぞれの層から一つずつ選んで例示する．太字は特徴的な標本単語，斜体は他の層を特徴づける標本単語である (文頭の記号は各文に与えた識別番号)．

s 8 Computer software

Therefore, once the test **data** have successfully **passed** through the **program,** revealing the absence **of** gross errors, variations *on* these data should also **be** used.

t 13 Computer hardware

If the left-zeros **counter** contents are zero, the **dividend** *was* shifted **as** far as the **divisor,** the **quotient** did not overflow, *and* no scaling is required.

o 3 Steinbeck's novel

For two years they waited, **and then their** mother said he must *be* dead.

x 8 Bellow's novel

This limousine probably had run **on** the boulevards *of* Madrid before Clarence **was** born **but** it was mechanically still beautiful.

Table 3. Characteristic of Each Group

Group	Subject	Factor		
		1	2	3
s	Computer software	+	+	
f	Computer hardware	+	−	
o	Steinbeck's novel	−		+
x	Bellow's novel	−		−

Table 4. Variances

Factor	Variance	Accumulated Variance
1	.109	.109
2	.086	.195
3	.076	.271
4	.070	.341

Table 5. Eigenvalues and Analysis of Variance

Factor	Correlation Coefficient	Eigenvalue	Between Group Variance	Within Group Between Sentence Variance	Within Sentence Variance
1	.656	.430	.274	.156	.570
2	.584	.341	.245	.096	.659
3	.546	.298			
4	.527	.277			

この模型に関するかぎり一つの意味的カテゴリーに属する，きわめて縁の近い単語であるといえる．

(5) 算出された数値を **Table 4, 5** に示し，数値的な吟味を行なう．まず Table 4 では各標本単語の相伴のふるまいかたが，わりあいに少ないパラメータで表現できることが示されている．すなわち各単語の全変動のうち 34% までが，はじめの4つの因子（固有ベクトル）だけで説明できることがわかり，これは満足できる数値である．

つぎに Table 5 では相関係数と分散分析とが示されている．相関係数 .656 は文と単調とを第1ベクトルの値で関係づけたときの値で，これは Fig. 3 における相伴の印（V）の分布の相関係数である．これが非常に高い値注4) をとっていることは，各文がそれぞれに特有の単語を含んでいることを示すものである．

これをさらに分解してみると（第1ベクトルの値に関しては），単語出現のすべてのばらつきのうち，層による違いが 27%，一つの層のなかの文によるばらつきが 16%，一つの文のなかの単語のばらつきが残りの 57% ということになる．

注 4) 別の実験（文献 8）では，単一の層からとったデータについて，さらに高い相関係数 .948 が得られた．

7. 計 算 式

ここで用いた模型と計算方法とは，林知己夫[1] の数量化理論のうち，外的基準のない場合によっている．この手法は従来，行動科学(心理学や社会学)の分野で要因の解析に使われてきたものである．これは Fig. 3 のような相伴表（もちろん文や単語の配列順序は任意でよい）を与えられたときに，文と単語とのあいだの相関係数が最大になるように，一つ一つの文，一つ一つの単語にそれぞれの数値を付与することに相当する．ただし文と単語とのあいだの同時分布は，その文がその単語を含む場合には，1/(拡がり語数) とし，その文がその単語を含まない場合には 0 とする．

この相関係数はさきにも述べたように，相伴の印(V)を相伴表で対角線状にならべたときに最大になるのであって，次の方程式を解けばよい．

$$A\boldsymbol{x}=\lambda B\boldsymbol{x}$$

ここで，Aは単語どうしの相関を表わす対称行列である．各文について標本単語の語数の逆数を計算しておいて，その逆数をある二つの標本単語を共に含んでいる文について拾って累計したものを，その二つの標本単語の相関の強さとしてAの要素にする．

Bは計算のために導入された，単語に関する対角行列であって，その対角要素は各標本単語の拡がり語数である（非対角要素はすべて 0 である）．

この方程式を適当な計算手法[7] で解くと固有ベクトル$\boldsymbol{x}$が求められる．これが前述した相関係数を最大にする標本単語の数値であり，同時に，意味空間に標本単語を位置づける座標値でもある．一つ一つの標本文の数値は，それに含まれている標本単語の数値を合計して平均した値で与えられる．

固有値λは相関係数の二乗になっている注5)．

注 5) 相伴表が入力として与えられたときに，それぞれの標本単語 W_i に数値 $x_i(i=1\sim l)$，それぞれの標本文 S_j に数値 $y_j(j=1\sim m)$ を与えれば，相伴表を $P(X=x_i, Y=y_j)$ なる確率をもつ同時分布と想定することができる．確率 $P(X=x_i, Y=y_j)$ は $\delta_j(i)/N$ と考える．ここで $\delta_j(i)$ は

$$\begin{cases} \text{標本文 } S_j \text{ が標本単語 } W_i \text{ を含むとき，すなわち相伴の印 V があるとき} & 1 \\ \text{標本文 } S_j \text{ が標本単語 } W_i \text{ を含まないとき，すなわち相伴の印 V がないとき} & 0 \end{cases}$$

となるような関数，N は拡がり語数，すなわち相伴の印 V の総数である．

以上から X と Y との間の相関係数 ρ_{XY} を記述し，さらにこの相関係数が最大になるように $x_i(i=1\sim l)$, $y_i(j=1\sim m)$ の値を決定するのである．

(次ページ脚注につづく)

8. 計算意味論について

計算言語学の立場からみて，意味論の研究課題は三つある．

（1） 意味的な分類の算法の研究．本報告はもっぱらこの点について実験したものである．

（2） ある言語系をカバーする語彙表または分類体系の作成．たとえば文献（6）がある．

（3） 意味的情報の応用手法の研究．たとえば後出の「情報の簡約」の項に述べた．

この報告は林知己夫[1]たちが開発してきた数量化理論のうちの一つを言語事象の処理，とくに語彙の意味的構造の解析に利用してみた試験的な一例である．

従来の意味論の多くは人文科学系の人たちによって思弁的な方法で考察されてきた[6]．それとは対照的な研究方向として，意味論の客観的操作的な測定手法の開発という課題があり，たとえば Osgood[2] の意味解析法 注6) があった．しかしこれも人間（被験者）の主観的な判断をいったん経由しなければならないことにいささかの疑点があった．

われわれの試みの目的の一つは，人間の思惟の介在をできるだけ排除した操作的な言語処理の可能性を模索することでもあった．電子計算機の全面的な利用は，処理手続きの完全な透明性・客観性を保証するものであると考える．そしてまた計算機による言語情報処理の一例としての意義をもっている．

今回の実験では，このような言語事象のモデル化がいちおう妥当すると思われる結果を得たが，さらに問題点を示しておこう．

言語標本の抽出法　本実験は試行的なものであるので，小規模でしかも故意にかたよりをもたせた標本について解析したが，より一般性・信頼性のある結果を得るためには大標本を処理しなければならない．ことに言語事象は非常にゆれの大きいものなので，小さい標本では一般的な結論は得られない．どういう標本をとるにしても抽出の客観的な基準が必要である．プログラムとしては標本文の個数に制限はないにしても，どういう層からどういう割合で標本文を抽出するかということは，母集団を明瞭にしなければならないので困難な問題となる．

標本単語の選択　このプログラムにおける標本単語の異なり語数の上限はさしあたり 100 語であって，言語事象を特徴づけることができるというにはあまりに少ない語数である．この数量化法は要するに語彙の構造を解析するためのものであるから，語彙といえる程度の大きさ，つまり少なくとも数百語の標本単語の処理が必要である．現在のところこのように大きい次元の固有方程式を解くための技術は，数値解析的にも計算機の記憶容量としても知られていない．

標本単語を選択するのに拡がりの大きい単語から順に 35 語をとったが，このやりかたでは全体として使

（前ページ脚注のつづき）

$$\rho_{XY}=\frac{\sum_{i=1}^{l}\sum_{j=1}^{m}x_i y_j P(X=x_i, Y=y_j)}{\sqrt{\sum_{i=1}^{l}x_i^2 P(X=x_i)-\left(\sum_{i=1}^{l}x_i P(X=x_i)\right)^2}\ *\ \sqrt{\sum_{j=1}^{m}y_j^2 P(Y=y_j)-\left(\sum_{j=1}^{m}y_j P(Y=y_j)\right)^2}}$$

$$-\left(\sum_{i=1}^{l}x_i P(X=x_i)\right)\left(\sum_{j=1}^{m}y_j P(Y=y_j)\right)$$

を最大にするという条件

$$\frac{\partial\rho_{XY}}{\partial x_i}=0,\quad \frac{\partial\rho_{XY}}{\partial y_j}=0\quad (i=1\sim l, j=1\sim m)$$

より，前記の確率の値をあてはめて計算，変形を行なうと

$$Ax=\lambda Bx$$

なる形の方程式が得られる．

A は要素が

$$a_{ik}=\sum_{j=1}^{m}\frac{\delta_j(i)\delta_j(k)}{\nu_j}\quad (i,k=1\sim l)$$

なる $l\times l$ 対称行列である．ここで ν_j は $\nu_j=N\times P(Y=y_j)$ $(j=1\sim m)$ であり，各標本文に含まれる標本単語の語数を表わす．

B は要素が

$$b_{ii}=\mu_i\ (i=1\sim l)$$

なる $l\times l$ 対角行列である．ここで μ_i は $\mu_i=N\times P(X=x_i)$ $(i=1\sim l)$ で，各標本単語の拡がり語数を表わす．

x は $x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_l\end{pmatrix}$ なる l 次元ベクトルである．

A, B は相伴表から簡単につくることができる．この方程式を解いて1未満の最大固有値に対する（X の分散が1となるように）正規化された固有ベクトル $x^*=\begin{pmatrix}x_1^*\\x_2^*\\\vdots\\x_l^*\end{pmatrix}$ を求める．

各標本文の数値は

$$y_j=\frac{1}{\nu_j}\sum_{i=1}^{l}\delta_j(i)x_i^*\quad (j=1\sim m)$$

で与えられる．

第2固有値以下に対する固有ベクトルを用いれば，別のかたちの単語の数量化が得られ，要因の分析に利用することができる．

注 6) Osgood の意味解析法は要するに言語事象にたいして，(1) 主観的な評定法（rating method）によって原始データを集め，(2) 相関係数によって単語のあいだの距離を表現し，(3) 因子分析法によって意味空間の座標を抽出するものである．

著者の一人はかつてこの方法を日本語について追試して興味ある結果[3][4]を得たが，その方法の全体が機械化に適していないこと，ことばを両極的な対（たとえば「大きい—小さい」）にしなければならないこと，相関係数を距離の測度とすることの妥当性などに難点があると考えた[5]．ここに報告した実験はそれらの難点を排除した方法であるが，それにもかかわらず Osgood の方法からまったく離れたものではなく，その直接の小さな改良であると考える．もちろんこの方法もことばの意味の測定法としては完全に近いものでさえもない．

用度数の高い単語しか選ばれず，林の数量化理論の利点の一部を生かしていないきらいがある．すなわちこの理論では使用度数の低い単語は低いなりに正当な評価をうけるようになっていて，たとえば使用度数の低い単語どうしが強く相伴していればそのことは重要な情報になるはずである注7)．しかし一方，使用度数の低い単語は標本誤差が大きくなる．だから標本誤差をある限度内におさえておいて，いろいろな使用度数の単語をまんべんなく選ぶ手法を考えなければならない（選択規準の立て方が問題である）．

情報の簡約 この試みでは文は異なった単語からなる集合として扱われた．各文の有する情報を，それぞれの文がそれぞれの標本単語を含んでいるかいないかの指標だけにまで切り縮めてから数量化に移った．したがって，文中での標本単語以外の単語の存在，標本単語の重出，文中の語順や構文，文と文との関係など，言語事象を特徴づける重要な情報が多分に捨てられている（たとえば否定文）．

また単語は文字列の形態だけで識別した．だから同綴異義語，多品詞語などが標本単語中にあると，数量化の結果の解釈が困難になることもありうると考えられる．逆に，単数形と複数形，現在形と過去形など綴りの変化した語がまったく別の単語とされることも，問題によっては欠点となろう（文字列の形態や，多義語については，いろいろな解決方法が考えられる）．

この定量化の手法は，単語や文や文章の意味の解析，語彙の構造の解析，未知の言語の形式的処理，あいまいさを含んだ情報の検索，文章スタイルの定量的指標の設定などに利用でき，さらに実用的には，シソーラスの自動作成，専門用語辞典の自動作成，ドキュメントの自動分類[8]，多義語の自動翻訳注8)などにも応用できる．そのためにも上述した問題点の，標本抽出の手法，言語事象に適合したモデルと数量化法との開発，大きな次元の固有方程式の数値計算法などの研究が必要である．

またこのように小規模の実験でさえも，標本文を機械媒体に入力してやること，および得られた数値をグラフにうつして視察することの労力は非常なものであった．これらを自動化することはたいへん望ましい．

最後に日ごろいろいろ有益な助言をしてくださる末包室長ならびに室員の方たち，数量化理論の勉強の便宜をはかってくださった電通(株)電子計算室の柳井次長，森本，羽賀両氏，初期のプログラム作成をしてくださった中央大学管理工学科の金子，和田両君，図表作成にあたった岩坪恵子嬢の皆さんに感謝したい．

注 7) 拡がりの大きい単語から順に 100 語をとって解析してみたところ，35 語の場合よりも大きい値の固有値が得られた．このことは使用度数の低い単語のほうが強くかたよった情報（意味）をになっているという予想を裏書きするものである．

	35 語	100 語
第 1 固有値	.430	.541
第 2 固有値	.341	.452
第 3 固有地	.298	.397

注 8) たとえば climb, high, mountain, buy, expensive, car などの単語が数量化されていれば，和文英訳を機械でやらせるときに，「高い山に登った」と「高い車を買った」とにおける多義語「高い」の訳語を判別させることは容易であろう．

参考文献

1) 林知己夫：数量化理論とその応用例 (II)，統数研彙報 **4**-2, 1956.
2) Charles E. Osgood, George J. Suci, Percy H. Tannenbaum: The Measurement of Meaning, University of Illinois Press, 1957.
3) 山本和郎，西村恕彦，野村健二，飽戸 弘，岡部蓉子：S. D 法による日本語の意味構造の研究，市場調査 82 号，1960-8.
4) Moriji Sagara, Kazuo Yamamoto, Hirohiko Nishimura, Hiroshi Akuto: A Study on the Semantic Structure of Jasanese Language by the Semantic Differential Method, Japaness Psychological Research, **3**-3, 146-156, 1961-9.
5) 西村恕彦：適性検査による予測，東京大学大学院人文科学科 1961 年修士論文.
6) 国立国語研究所：分類語彙表，国語研究所資料集，6, 1964.
7) 駒沢 勉：固有値解法の一工夫について，統数研彙数，**12**-1, 1964.
8) 西村恕彦，岩坪秀一：計算機による文献の自動分類，第 6 回情報科学技術研究集会論文集，1969-10.
9) 西村恕彦：計算機による言語の意味の解析，言語生活，1969-10.

（昭和 44 年 7 月 14 日受付）