自然言語処理 65－1
（1988. 3. 25）

係り関係による単語のクラスタリングの試行

仲尾　由雄　　籾山　陽子

（富士通研究所）

現在，単語を上位・下位などの意味的関係により体系化し，意味処理に必要な情報を圧縮して記述しようという試みが広く行われている．本稿は，こうした意味の体系の構築の難しい形容詞等の概念に対し，単語の係り受け機能を記述するための体系を作ることを目指した試行について述べる．サンプルテキストより抽出された係り受け頻度を確率的手法により正規化し，数量化Ⅲ類と呼ばれる統計手法を用いて単語の特徴を示す値を計算した．統計的手法により得られた値を単語間の意味的相対距離と解釈し，原データとの関わりで定性的な評価を行った．その結果，総係り受け数が10^2 ～10^3 の小規模・中規模なデータに対し，安定に単語間の相対的距離を計算できる見通しが立った．単語間相対距離の定量的な評価は今後の課題となっている．

Word Clustering by Word Bindings

Yoshio Nakao　Youko Momiyama

FUJITSU LABORATORIES LTD

1015, Kami-kodanaka, Nakahara-ku, Kawasaki, 211 Japan

This paper reports experiments on word clustering. Firstly it proposes a method to evaluate the similarity of words by using the frequency of modification between words. Secondly it reports the result of two experiments on word classification; one is about 12 verbal adjectives and 17 verbs, another is about 208 nouns and 217 verbs. In these experiments semantic distances of words were calculated by the statistic method "Suuryouka-3rui" using matrixes of the similarity mentioned above. The results show the efficiency of the method in evaluating the similarity. This suggests the possibility of automatic calculation of general semantic distances. The quantitative estimation of semantic distance calculated in the experiments is not included.

1. はじめに

単語の意味的性質を上位・下位などの関係によりうまく体系化すれば，上位から下位に意味的性質を継承することで，大幅に情報記述量を削減することができる．また，こうしたシソーラス体系を作ることは，人間の自然言語処理を模した高度な処理の可能性を示唆する．

シソーラスの作成は，名詞概念のように現実世界とのマッピングによりかなりの部分が可能なものや，一部の動詞概念のように意味素の抽出により可能となるものがあるが，形容詞的概念や副詞的概念などは階層化が困難である．

例えば，「鳩」という名詞概念に対し，「鳥」という上位概念を与え，「翔ぶ」ことができる等という意味的性質を継承することは比較的明白である．これにより，「鳩がとぶ」という文の解析において，「とぶ」の意味として「翔ぶ」を「跳ぶ」に優先して扱うことが可能になる．移動の性質を持つ動詞群を抽出し，さらに移動するものに従い細分し，「移動」概念の下位に位置づけることも妥当性が高い．

一方，形容詞においては，「赤い」「青い」のように，色に関する性質を示すものとして色に関する形容詞群にまとめるて処理することが可能なものと，「寒い冬」とは言えるが「暑い冬」とは言い難いもののように，気温に関する主観的な評価として同じような性質をもちながらもその程度・強度の違いで特殊な処理を必要とするものがある．（「あつい冬の下着」の常識的解釈を得るためにはこういった情報が必要となる）さらに，「暑い」「寒い」に関する処理は否定が加わると逆転するという複雑さをも持っている．

「暑い」「寒い」のような概念に対する記述を単純な上位・下位のシソーラスで表現することは難しく，このような概念に対する被修飾関係を単継承の形に体系化して記述するためには，例え名詞概念であってもほとんど単語レベルまで細分化した体系が必要となる．

今回の試行は，こういった体系化・整理の困難な意味特性に対して，係り関係の成立の可能性を定量的に扱うことで，問題の所在を明らかにし，分類・記述への糸口をみつけようとしたものである．

もし，「暑い」「寒い」のような概念に対し，その意味特性を客観的に評価可能な特性値で表すことが出来れば，特性値の関係として意味を一貫して記述することが可能になる．例えば，「暑い度」を設定し「暑い」に1.0「寒い」に-1.0「涼しい」に0.5「あたたかい」に-0.5を与え，「冬」に-1.0「夏」に1.0の「夏度」を与えることでこれらの単語処理をまとめることが出来る．その特性値の評価法を状況により変えることで，取り扱う分野による変化に対しても，柔軟な対処が可能となる．

そこまで行かなくとも，係り受け関係を基本にした分類で，意味的な考察からは気が付き難い分類が見えてくる可能性がある．

以上の観点から，本稿では，技術分野の新聞記事約6000文より係り関係の明確なものを抽出し，その成立頻度により単語を分類する試行について述べる．以下，2章で今回用いた分類法の概要を述べたあと，3章で実際に得た結果の検討を行う．

2. 係り受け頻度による単語の分類の試行の概要

2.1 原データ

科学技術関係の新聞記事約6000文から，係り受けの関係が明確な形容動詞・動詞のペア，名詞＋「を」・動詞のペアを抽出し，表記別の係り関係頻度を原データとして用いた．

形容動詞・動詞のペアとしては，約550 件抽出されたもののうち成立頻度が極端に少ないペアを除いた173 件を用いた．名詞＋「を」・動詞のペアとしては，同様に，約4100件抽出されたもののうち2279件を用いた．

2.2 分類法

2.2.1 分類法の概要（文献１，文献２）

今回の試行で用いた分類の方法は，アンケート調査や語彙論（文献３）等の分野でよく用いられる統計的手法である数量化Ⅲ類を拡張したものである．

アンケート調査を例にとると，数量化Ⅲ類を使うと，アンケートの質問項目に対する回答パターンにより回答者・質問項目双方を同時に分類することができる．○×式のアンケートでは，分類された回答者のグループと質問項目のグループを照合することで，どういった回答者がどういった傾向の質問項目に○をつけるのかという関係が明らかになる．

今回の試行は，上の例の回答パターンに，形容動詞（名詞）が動詞に係る度合いのパターンが対応し，形容動詞（名詞）と動詞を同時に分類することで，どういう形容動詞（名詞）の群がどういう動詞の群に係りやすいのかを調査したことになる．

今回の試行は，これを一歩進めて，分類の際に算出される値（2.2.2，2.2.3参照）の差をそのまま同一グループ内の単語の意味距離として用いることをも考慮している．

2.2.2 数量化Ⅲ類の数学的内容

数量化Ⅲ類を数学的に言うと，ある関係マトリクスに対し，１でない相関係数が最大になるようなベクトル対

を求めるものである．そして求まったベクトルの値により行・列に対応する要素を並べ換え，性質の似た要素を見出すことになる．

数量化Ⅲ類は，0-1 マトリクスを基本とし，あるいはもっと多レベルに離散化したマトリクスを扱うことが普通であるが，今回の試行では，浮動小数で表された連続値域をもつマトリクスをも扱った．

2.2.3　分類法の解釈

数量化Ⅲ類の別の定式化を行うと，関係マトリクスをF，係り元要素に与える得点ベクトルをａ，係り先要素に与える得点ベクトルをｖとして

$$\Sigma F_{ij} (a_i - v_j)^2$$

が

$$\Sigma F_{ij} (a_i^2 + v_j^2) = \mathrm{const}$$

のもとで最小となるようなａ，ｖをFに関する固有値問題として解くことになる．

今，係り元・係り先がそれぞれある性質を強度ａ，ｖをもつと仮定し，強度が近い程係り受けを形成しやすいと考えると，Ｆは係り受けのし易さの重みのマトリクスと解釈され，係り元と係り先単語の親近性を示す関係マトリクスとなる．

2.3　係り受け関係マトリクスの作成法

2.2.2の解釈によれば，表１に示されるような係り受け頻度マトリクスは，そのままでは関係マトリクスとして適さない．すなわち，係り受けを多く形成した単語はどの相手単語に対しても大きい係り受け頻度が予想されるため，頻度マトリクスそのままでは，単語間の親近性を表していないと考えられる．

今回は，単語の親近性を表現するために，係り受けのし易さの度合いとして，全体の係り受け頻度から期待される係り受け頻度からのずれを用いた．具体的には，使用した係り受けの全体数をｔ，係り元の登場頻度をｆ，係り先の登場頻度をｇとおいて

$$\text{期待値} = f * g / t$$
$$\text{分散} = \sqrt{\text{期待値}}$$

と仮定し（ポアッソン分布で近似），期待値からのずれを分散により正規化した値を用いた．

表１の形容動詞・動詞の係り受け頻度マトリクス（原データ）を，上記の方法で正規化した係り受け関係マトリクスを表２に示す．表２の値は，浮動少数で計算されたものを四捨五入して整数化したものである．なお，表１・表２中の動詞・形容動詞は係り受けの頻度に応じてソートしてある．

表１　係り受け頻度マトリクス（形容動詞・動詞）

動詞＼形容動詞	ように	可能に	自由に	大幅に	簡単に	急速に	自動的に	独自に	容易に	気軽に	着実に	大量に	合計
なる	5	4	2	3	1	1	0	0	0	0	0	0	83
使う	4	1	1	0	0	0	0	0	0	0	0	1	18
利用する	1	6	0	5	0	0	0	0	0	0	0	0	10
翻訳する	1	0	0	2	0	0	0	0	0	0	0	0	8
進む	2	0	0	4	0	0	0	0	0	0	0	0	8
拡大する	0	0	0	0	0	0	0	0	0	0	0	0	7
向上する	0	0	0	0	0	0	0	0	0	0	0	0	7
構築する	0	0	0	0	0	0	0	0	0	0	0	0	7
翻訳する	0	0	0	0	0	0	0	0	0	0	0	0	7
発展する	0	0	0	0	0	0	0	0	0	0	0	0	7
減少する	0	0	0	0	0	0	0	0	0	0	0	0	7
成長する	0	0	0	0	0	0	0	0	0	0	0	0	7
減る	0	0	0	0	0	0	0	0	0	0	0	0	7
展開する	0	0	0	0	0	0	0	0	0	0	0	0	7
縮む	0	0	0	0	0	0	0	0	0	0	0	0	7
短い	0	0	0	0	0	0	0	0	0	0	0	0	7
合計	59	13	10	8	7	7	6	4	4	3			173
	39	13											

表２　係り受け関係マトリクス（形容動詞・動詞）

動詞＼形容動詞	ように	可能に	自由に	大幅に	簡単に	急速に	自動的に	独自に	容易に	気軽に	着実に	大量に	合計
なる	10	8	0	0	0	0	0	0	0	1	1	2	28
使う	0	8	2	2	5	3	3	3	3	4	4	5	43
利用する	1	1	7	3	4	4	4	4	6	5	5	6	51
翻訳する	1	1	7	4	7	6	4	5	5	5	5	5	54
進む	2	2	4	4	4	4	4	5	5	5	5	5	55
拡大する	4	3	4	5	5	5	6	5	5	5	5	5	55
向上する	1	2	5	6	5	5	5	5	5	5	5	5	52
構築する	2	3	5	5	5	5	5	5	5	5	5	5	56
翻訳する	2	3	5	5	5	5	5	5	5	5	5	5	56
発展する	2	3	5	5	5	5	5	5	5	5	5	5	58
減少する	2	3	5	5	5	5	5	5	5	5	5	5	56
成長する	2	3	5	5	5	5	5	5	5	5	5	5	56
減る	2	3	5	5	5	5	5	5	5	5	5	5	58
展開する	2	3	5	5	5	5	5	5	5	5	5	5	56
縮む	2	3	5	5	5	5	5	5	5	5	5	5	56
短い	2	3	5	5	5	5	5	5	5	5	5	5	56
合計	39	53	83	75	80	77	81	82	85	83	81	83	902

2.4　数量化Ⅲ類による計算法と結果の表現法

数量化Ⅲ類の算法は，関係マトリクスから相関係数の自乗を固有値とする計算用マトリクスを作成し，行列の固有値問題として解くものである．

計算により得られた結果は，本稿では，図３～図10に形容動詞・動詞，名詞・動詞をそれぞれ得点ベクトルに従いプロットし，散布図として示した．各散布図でそれぞれの品詞内での分類結果の外観をみることができる．

2.5　分類結果の評価と解釈

係り受け関係マトリクスの計算により得られた結果の

評価は，(1)計算値の数値自体の有効性の評価と，(2)計算値の解釈可能性の評価を相互に参照し，総合的に行った．

(1)の計算値の有効性の評価は，関係マトリクスの固有値計算により得られた相関係数の値と，得られた全固有値から計算される寄与率に基づいて行った．

寄与率とは，行列全体のもつ情報をどれだけその分類が説明しているかの指標となるもので，相関係数の自乗である行列の固有値をλとしたとき

$$寄与率_i = \lambda_i * \Sigma \lambda_j$$

但しλ_jは1でない固有値．

なる値である．

(2)の計算値の解釈は，計算結果のうち相関係数の大きいもの3個に対応する分類（単語分布）を散布図に表し，その分布の傾向と単語のもつ意味の関係を考察することで行った．

3. 試行結果と評価

3.1 形容動詞・動詞の係り受けによる分類

この試行は，今回用いた方法の具体的挙動を調査することを目的として行った．以下では，今回の分類法の特徴を抽出すべく，詳細な検討を行う．

試行の結果得られた，相関係数と寄与率を，図1・図2に示す．また，関係マトリクスを浮動小数のまま計算した結果の相関係数の大きい分類軸3つによる形容動詞・動詞の散布図を図3～図6に示す．

3.1.1 係り受け関係マトリクスの評価

この項では，今回の係り受け関係マトリクスの作成法が不用な情報をマトリクスに付加していないかどうかを確かめるため，離散化したマトリクスの収束性と振動性を検討する．

表1の係り受け頻度の正規化により，表2の関係マトリクスに不用な情報（雑音）が混入しているならば，離散化する値域により雑音の影響が異なるため，収束性が悪くなると考えられる．以下では，このような雑音の混入をみるため，値域〔0，1〕～〔0，10〕に離散化したマトリクスの計算結果と浮動少数のまま計算した計算結果の比較を行う．

(1) 計算値の収束性

まず，図1の相関係数・図2の寄与率は共に値域を拡げるに従い，値域内の整数値の偶奇による振動を弱めながら，浮動少数によるものに収束していくことが分かる．偶奇による振動の原因については(2)で述べるが，その振動をならした仮想的収束曲線はきれいに収束へ向かっている．

これは，今回の関係マトリクスに混入する不規則な雑音が少ないことを示し，ある幅以上の値域をとれば安定した結果が得られることが予想される．本稿には載せていないが，得られた値による散布図の目視からも，〔0，4〕値域以降ではほぼ浮動小数計算のものと同様の傾向が確認された．

あまり値域を狭くすることは，表1に示される係り受け頻度の情報を落とすことになるので，値域が〔0，1〕や〔0-2〕で相関係数が高くとも，結果は係り関係の特徴の微細な構造を反映しない．特に，〔0，1〕の値域の値は，寄与率が極端に低く，ほとんど無意味な結果となっている．

(2) 計算値の振動性

次に，図1・図2の細かい変化に注目すると，離散関係マトリクスの取りうる値の数が偶数か奇数かにより，相関係数・寄与率ともに異なる傾向を示すことが見て取れる．

相関係数・寄与率ともに，取りうる値の数が奇数のものの方が偶数のものより，全体として高い値をとっている．（仮想的収束曲線との比較による）これは，今回採用した方法によると，余り特徴的でない係り受け頻度のものが，関係マトリクス上で平均的な5.0前後の値をとる（表2で5の値をとる）ことと関係している．

関係マトリクスを偶数個の値域を持つ離散マトリクスに変形すると，特徴的でないグループの係り受け関係の評価値が2つに分裂し，本来不要な情報が関係マトリクスに混入する．結果として相関係数の値が落ち，分類の寄与率が下がることになる．

(3) 計算で用いるべきマトリクス

(1)(2)をもとに，原文中の係り受け関係により単語の分類を行うのに適当なマトリクスを考えると，奇数個の比較的大きい値域を持つ離散マトリクスを採用するか，浮動小数のままのマトリクスを採用することがよいと結論される．

図1・図2の収束性のよさから観て，浮動小数計算の結果と比較的大きい値域をとった離散マトリクスの結果の違いは小さいと考えられる．また，離散マトリクスを採用すると，計算の際に落ちる情報中に意味の微細構造に係わる重要なものが含まれている可能性もあるので，以下では，浮動小数のまま計算した結果を中心に議論を進めることとする．

3.1.2　相関係数の大きさ

相関係数という用語の意味を字義通り解釈すると，図1の値はかなり低い値しか示していない．特に，第三相関係数についてみると，ほとんど無相関といえる値となっている．しかしながら，後で述べるように，この第三相関係数に対応する分類においても，かなり有効な意味特性が見て取れる．

数量化Ⅲ類において，(1)相関係数は計算マトリクスの固有値の平方根となること，(2)必ず1つは相関係数1の解が存在すること，また，単語の意味構造において，(3)微細な構造まで辿れば単語の数のオーダーのほとんど無限次数の分類が存在すること，(4)意味の構造は必ず例外が存在し登場単語数の増加に伴い例外の観測数が増加することなどがこのことを説明すると思われるが，本稿ではこれ以上の深入りは避ける．

マトリクスの次数の増加がもたらす影響については，3.2の名詞＋「を」・動詞による分類で検討するが，この試行内に限れば，相関係数（固有値）の値に係わらず，有為な結果が出ている．固有値計算の誤差に注意すれば，相関係数の絶対値ではなく，寄与率として計算された値を重視すべきであるといえる．

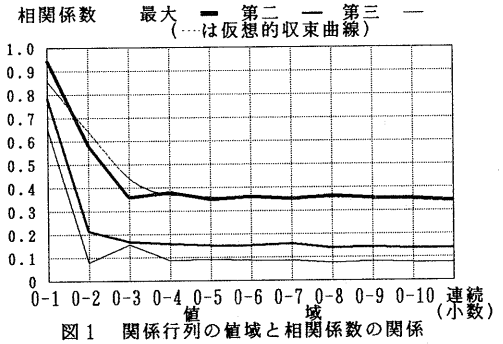

図1　関係行列の値域と相関係数の関係

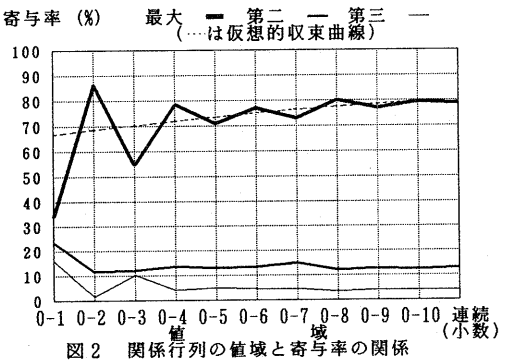

図2　関係行列の値域と寄与率の関係

3.1.3　分類結果の解釈

相関係数の大きい方から3つの分類結果を図3～図4に示す．図で1軸とあるのは最大相関係数に対応して得られた分類軸である．以下2番目の相関係数に対応するのが2軸，3番目のものに対応するのが3軸である．

(1)　分布の外観

ここでは，図3・図5に示される最大相関係数・第二相関係数に対応する分類結果の散布図について全体の傾向をみる．

まず，図5の動詞の散布図を参照すると，1軸においては，「なる」とその他の動詞に大きく分離され，2軸では，「する」とその他の動詞に大きく分類されていることが分かる．

さらに，微細な構造を探ると，2軸においては，「する」と「進む」を両極端とする構造が見出される．

これに対応している図3の形容動詞の分類を参照すると，1軸では，「なる」と同方向に「可能に」「ように」の2単語が挙動を共にし，「容易に」を除く形容動詞群から分離している．「容易に」はその中間に位置している．

2軸においては，「可能に」と「ように」を両極とする構造が現れ，その他の形容詞においてもきれいな系列がみられる．

原データにおける出現頻度（表1）を参照すると，「なる」と「ように」「可能に」の組み合わせが異常に多く，また，上で個別に名前の挙がった単語は比較的頻度が大きいものばかりであることが見て取れる．

(2)　解釈

(1)の観察で気がつくことは，原データにおける出現頻度に応じて軸の上で特徴的な単語となることである．これを単語による軸の支配と考えると，軸を支配する単語の意味属性に近い（反対に遠い）単語が，その出現頻度に応じて軸の方向（意味属性が近い時は支配単語の方向・遠い時は逆方向）に分離されていることが分かる．

さらに，ある軸で支配的に働いた単語は，次の軸では一般的な単語として機能する傾向があり，次に出現頻度の高い単語に支配権を譲ることが多いことも観察される．

この見地から軸の解釈を行うと以下のようになる．

①　1軸（図3・図5）

動詞は「なる」に支配され，形容動詞は「ように」と「可能に」の共通特性に支配されている．動詞においては，「なる」の意味的特性のうちここで支配的であったものは状態的概念やモノ的な概念の自動詞化という特異なものであったため「なる」のみが他から分離した．ま

た，形容動詞においては，「ように」「可能に」という異なる強い特性をも有するものに共同支配されたため，他の形容動詞においてははっきりした分離が起きなかった．（状態的な意味と動作的意味の分類と見えなくもない）

② 2軸（図3・図5）

形容動詞は「可能に」と「ように」の特性のうち異なる部分の特性に支配され，動詞においては「する」の特性のうち「可能に」と「ように」を区別する要素による支配が行われている．

動詞・形容動詞とも，これらの意味特性は一般的な面があったため，動詞においては「進む」「使う」「利用する」が，形容動詞においては「簡単に」「自由に」をはじめとする大部分のものが，その特性の強さと出現頻度に応じて展開された．

③ 3軸（図4・図6）

3軸においては，「なる」「する」・「ように」「可能に」の影響が1軸・2軸の分類によって薄められ，次に頻度の高い「使う」「できる」「利用する」「進む」や「自由に」「大幅に」「簡単に」「急速に」等の意味特性が支配力を強めている．

これらの単語群は，同一基準で分類し易いものであるため，かなり理解し易くはっきりとした系列が現れた．動詞においては，「向上する」—「削減・軽減する」—「進む」…「作成する」—「利用する」「使う」という自動的・変化的概念と意図的・作用的概念の系列が現れ，それに対応して形容動詞では，「大幅に」—「急速に」—「自動的に」…「気軽に」—「簡単に」—「自由に」という量的・状態修飾的概念と意志的な動作概念を修飾する概念の系列が並んでいる．

⑶ 軸の交替支配に関する考察

⑵で観察された原データの出現頻度に応じて軸を交替して支配する傾向は，関係マトリクス作成法と深い関わりがある．

出現頻度に応じた値（2.3で分散と定義した値）で係り受け頻度マトリクスを正規化したため，関係マトリクス上では，出現頻度が大きいもの程弁別性の高い値を示している．この弁別性により，単語の持つ意味的特異性が強く現れ，優先的に軸を支配すると考えられる．

単語の持つ意味的特異性は，軸を支配したことにより次軸以降での支配力を弱めるため，次の軸では優先度の高い別の意味的特異性を持つ単語が支配的に働くことが多くなる．

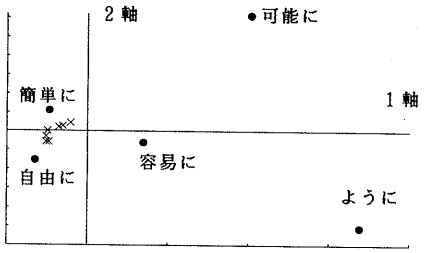

図3 形容動詞の散布図（1-2 軸）

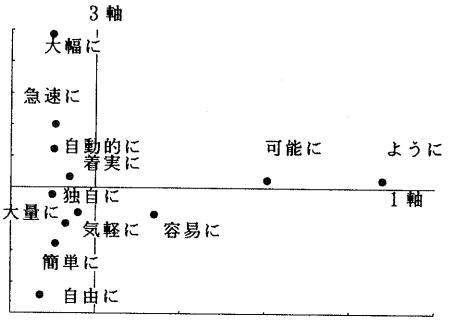

図4 形容動詞の散布図（1-3 軸）

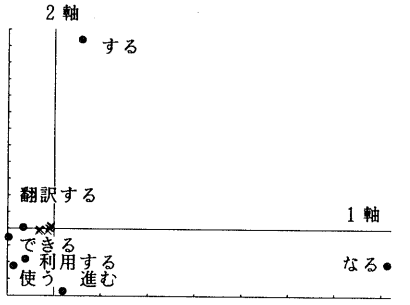

図5 動詞の散布図（対形容動詞 1-2 軸）

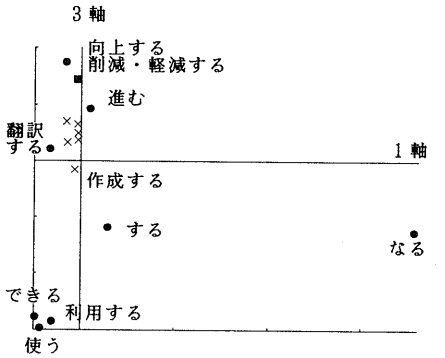

図6 動詞の散布図（対形容動詞 1-3 軸）

軸支配の優先順位は，単語の出現頻度だけでなく，単語の意味的特異性の強さにも関係する．すなわち，単語の持つ意味的特異性が特に強ければ，多少出現頻度が低くても，軸支配の優先度が高くなることが予想される．また，単一では軸支配するだけの情報を関係マトリクス上で示さない単語でも，共通な意味特性を持つ単語が強い支配力を持つか，そういった単語数が多ければ，軸上で大きい値をもつものとして登場することになる．

以上のように考えると，出現頻度が極端に高く，かつ非常に特殊な特性を持つ単語があった場合，1軸や比較的上位の軸で，多数の単語が分類されないままであることが予想される．このときには，もっと相関係数の低い軸で解釈可能な分類が行うわれている可能性が高く，例え相関係数が低くとも，計算誤差を許容できる範囲内なら検討する価値があると言える．

3.2　名詞＋「を」・動詞の係り関係による分類

名詞＋「を」・動詞の試行は，3.1の小規模の試行で得られた知見について，中規模なデータに対して検証することを目的として行った．

固有値計算の結果，相関係数は，最大.029，以下.024，.016，.012，.011と続き，それぞれの寄与率は，33.5%，23.7%，10.8%，6.0%となっている．尚，固有値算出に要した計算量は，確率ヤコビ法の繰り返し数（回転数）で594 回であった．

原データにおける係り受け頻度が多いものとしては，名詞では，「システム」250 頻度，「機能」84頻度，「開発」81頻度等であり，動詞では，「開発する」115頻度，「使う」109 頻度，「進める」98頻度「持つ」75頻度（「もつ」35頻度を合わせると110 頻度）等となっている．

相関係数の大きい3軸による名詞・動詞の散布図を図7～図10に示す．

3.2.1　3軸までの全体的傾向

図7～図10の各軸を支配した単語・意味特性は，3.1の知見通り原データの係り受け頻度を反映していた．各軸で支配的であった単語を1軸から順に述べると，

1軸：名詞　「システム」
　　動詞　「開発する」・「進める」で異なる特性
2軸：名詞　「機能」
　　動詞　「開発する」・「進める」の共通特性
3軸：名詞　「開発」
　　動詞　「もつ」「持つ」「使う」に共通で「開発する」「進める」とは異なる特性

となっている．

まず名詞について概観する．

1軸は（図7），全体として「システム」とその他の単語に分離した．その他の単語群の微細構造をみると，「ソフト」「辞書」「技術」等の人工具体物が「システム」側に位置し，「開発」「作業」「販売」等の動作概念や「意味」「知識」等の抽象概念が反対方向に分布した．

2軸は（図7），「機能」と「販売」「サービス」「開発」「情報」「意味」などが正の方向に分布した．大体抽象物が正の方向に分離したようにみえるが，「計画」等は原点付近に他の名詞と共に位置している．

3軸は（図8），「機能」「技術」…「販売」—「開発」の系列が現れた．モノ概念と動作的概念の系列と判定できる．

次に動詞について概観する．

1軸は（図9），「開発する」「使う」が正の極に，「進める」が負の極となり，「使う」「作る」の単語群と「進める」「始める」「持つ」の単語群に連続的に分解した．具体物に対する動作概念の極と，抽象概念や動作的概念を受けることが可能なグループとみえる．

2軸は（図9），「開発する」「使う」「進める」が負の方向（名詞の「機能」の位置に対して逆方向）に分布し，「持つ」「開始する」「始める」「目指す」「採用する」や「つくる」グループが負の方向の中間あたりまで分離している．この軸は，「機能」に関する軸としては納得できるものであるが，他の解釈は難しい．

3軸は（図10），「持つ」「もつ」および「通過」「利用する」のグループを正方向の極とし，「進める」を負方向の極とする系列に分解した．正方向には「持つ」や「利用する」「使う」といったモノ概念を受けるグループが分布し，負の方向には「進める」「開始する」「目指す」といった動作概念を受けるアスペクトマーカ的な動詞群が分布している．

3.2.2　相関係数と分類の考察

この試行における相関係数は最大でも0.029 と通常は無相関と見る値であった．しかし，上に見たように分類としては，名詞においては具体物概念と動作概念・抽象物概念という構造を表し，動詞においては，動作的名詞を「を」格に取りうる概念・状態的概念・動作的概念の構造を示した．また，求まった解の挙動も，比較的相関係数の高かった3.1の試行と同様であった．

相関係数が低い理由は，係り受け頻度マトリクスにおける総頻度数が登場単語数に比べ少ないことによる．頻度が低いため，関係マトリクスが，係り受け機能の特徴が抽出出来なかったことを示す値 5.0前後の値が多く，

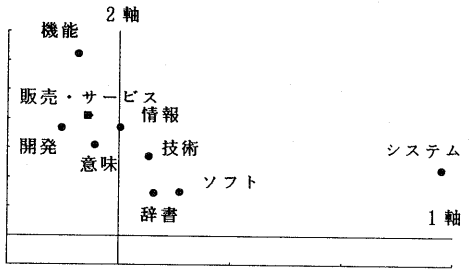

図7　名詞の散布図（対動詞　1-2 軸）

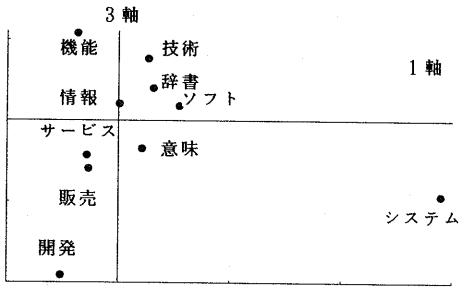

図8　名詞の散布図（対動詞　1-3 軸）

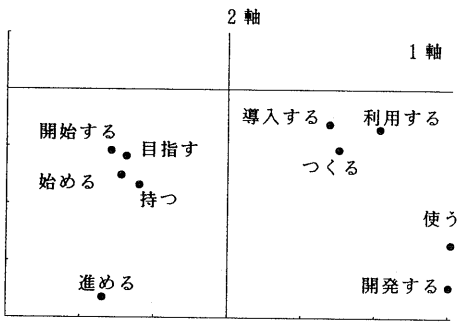

図9　動詞の散布図（対名詞　1-2 軸）

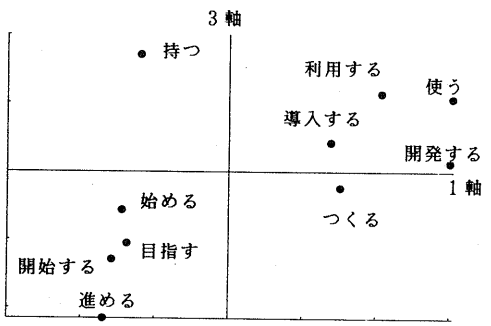

図10　動詞の散布図（対名詞　1-3 軸）

ほとんど一様な分布となったことが原因である．

しかしなから，相関係数の低い3軸において最も解釈の容易な分類が現れたことは，頻度の低い単語の情報が少なくとも分類を阻害しないことを示すと考えられ，膨大なデータを待たずにも安定した単語の意味距離計算が可能なことを示唆している．

4.　まとめ

本稿では，係り受け頻度の正規化により単語の意味構造に関する情報を安定して取り出す見通しが立ったことを報告し，テキスト中で強い働きを持つ意味的特性から順に抽出されるという今回の分類法の性質を述べた．

これらにより，サンプルテキストの入力により，分野等による特徴に応じた意味分類のチューニングの可能性が示される．また，意味を意味特性と特性値のペアによる意味処理の可能性についても，定性的には示したと考えている．

後者に関しては，計算により単語に与えられる値が，そのままでは意味特性の強度としては用いることができないことが課題となっている．値が単語の原データにおける頻度と意味特性の強さの両方の成分を含むところに問題がある．今後は，この値から意味特性に関する成分と頻度に関する成分を分離し，意味特性の値とその確かさという形に表現する方法を，理論・実験の両面から検討し，単意味特性値の計算法を目指したいと考えている．

謝辞

本稿を書くにあたり，貴重な助言を頂いたソフトウェア第一研究室の方々に深く感謝いたします．

参考文献

〔1〕岩坪：「数量化の基礎」　朝倉書店　1987
〔2〕西里：「質的データの数量化」朝倉書店　1982
〔3〕水谷：「語彙」　朝倉日本語新講座　1983