情 報 学 基 礎 57－9
自 然 言 語 処 理 136－9
（2000．3．21）

# 多項関係としての格標識共起知識の獲得

永井 秀利, 中村 貞吾, 野村 浩郷

九州工業大学 情報工学部 知能情報工学科
{nagai,teigo,nomura}@ai.kyutech.ac.jp

格パターン知識は非常に重要な知識の一つである．従来のほとんどの獲得知識は，一つの格要素と動詞との間の単項関係に基づく知識だが，そのような単項関係知識をいくら集めたとしても，複数の格要素の間の相関性を含めた特徴を捉えることはできない．動詞と格要素の集合との共起性を捉える多項関係知識が必要である．本研究では，特に格標識に着目し，動詞と格標識の集合との間での多項関係知識の獲得を目指す．このような多項関係知識を純粋に統計的に得ることは，組み合わせの膨大さから不可能に近い．そこで本稿では，動詞の用法をモデル化することにより，少ないサンプルからできるだけ妥当性の高い知識を獲得するための手法を提案する．

**キーワード**：格標識，共起知識，文型，動詞の用法，クラスタリング，構文解析

# Acquisition of Knowledge of Cooccurrence as the Polynomial Relation between a Verb and Case Markers

Hidetoshi NAGAI, Teigo NAKAMURA, Hirosato NOMURA

Department of Artificial Intelligence
Kyushu Institute of Technology
{nagai,teigo,nomura}@ai.kyutech.ac.jp

Usual knowledge of case pattern is the relation between a verb and a case element. However, the collection of such monomial relations cannot represent a characteristic based on a correlation between case elements. We focus on case markers, and aim to acquire knowledge of cooccurrence between a verb and a set of case markers. It is extremely difficult to acquire such polynomial relation with a purely statistical method. Therefore we propose a modeling of a usage of a verb, and a method to obtain the most proper possible frequency of the set of case markers from an insufficient set of sample data.

**keywords**: case marker, knowledge of cooccurrence, sentence pattern, usage of verb, clustering, syntactic analysis

# 1 はじめに

格文法で語られることが多い日本語の構文解析において，格パターンに関する知識は非常に重要かつ有用なものである．従来獲得が進められてきた格パターン知識は，「ある動詞において，どのような属性の格要素が，どのような格標識を伴って，いかなる格関係で，どの程度の頻度で共起するか」というものであると言える．このような知識は，構文解析における曖昧性絞り込みや意味解析において活用される．

ある動詞を主動詞とする文は，その動詞によって指し示される一つの状況を表すものであり，その文を構成する格要素は，その状況の構成要素である．従来の研究で獲得が進められてきた知識は，一つの格要素あるいは格標識と動詞との間の関係，すなわち，単項関係に過ぎない．状況の特定に十分な要素が与えられているかの判断には，格要素の存在を個々に見るだけでは不十分である．格要素の組み合わせとして，状況記述の充足性を見る必要があると言えよう．そこで本研究では，ある動詞と共起している格要素あるいは格標識の組み合わせとの間の関係，すなわち，多項関係としての性質に着目する．

本稿では特に格標識に注目し，動詞と共起する格標識の組み合わせについて，どのような組み合わせがどの程度出現しやすいかという言語知識を獲得することを目的とする．この知識は格関係名であるとか格要素の属性とかの情報に欠けるため，多項関係としての格パターンに関する知識としては不足である．しかしこの知識は，ある動詞が表層的に取りやすい文型の情報を与えることができ，用法面から見た動詞の分類や，構文解析時の曖昧さの解消などに有用であると考える．

# 2 多項関係知識の必要性

本研究が目的とする多項関係知識の獲得の意義を明確にするには，従来の研究で求められてきた単項関係知識と多項関係知識との差異を明らかにし，多項関係知識が単項関係知識を単に寄せ集めたものではないことを示す必要がある．

今，2 個の動詞 $V_1$，$V_2$ と 2 個の格標識 $M_1$，$M_2$ とが存在したとする．これらの動詞，格標識をあるテキストコーパスで調査したところ，いずれの動詞と格標識との組み合わせにおいても，その格標識がその動詞を主動詞とする一文中に出現する平均回数が 0.5，分散が $\sigma$ であったと仮定する．このとき，2 個の動詞 $V_1$，$V_2$ は同等の性質を持つと解釈することは正しい判断と言えるだろうか．

図 1 は，上記仮定を満足するという条件下において，動詞 $V_1$，$V_2$ についてのサンプル分布の可能例の一つを示したものである．縦軸は各サンプルにおける格標識 $M_1$ の出現回数，横軸は格標識 $M_2$ の出現回数とする．図に描かれた楕円は，サンプルのおおよその分布傾向を表すものとする．

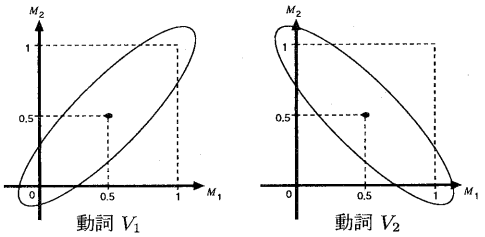


図 1: 単項関係知識では違いを扱えない例

図では，$M_1$ と $M_2$ の出現に関して，$V_1$ においては同時性が強く，$V_2$ においては排他性が強いというように，動詞 $V_1$ と $V_2$ との間に明らかな性質の違いが見られる．これは，上記仮定のように単項関係での知識で見ている限りにおいては，現れてこないような性質の違いが存在しうることを示している．

単項関係での知識として得た各格標識の平均出現回数は，サンプルの分布における重心位置 (図の●) を示しているに過ぎない．分布の形状についても，単項関係での分散の知識を集めただけでは，座標軸に平行した軸を持つ超楕円体としてしか捉えることができず，この図のような分布の差異を表現することはできない．よって，動詞の持つ用法的な性質の違いを的確に捉えるためには，多項関係での知識を獲得することが必要である．

多項関係での知識によって新たに得られる情報は，ある動詞の文脈下における格標識間の相関性についての情報である．例として，この共起性に関する知識を，構文解析における係り先の曖昧性解消のための優先順位付けに用いる場合を考える．単項関係による知識しか持たない場合は，「$\sim M_1 \sim M_2 \sim V_1$」と「$\sim M_1 \sim M_2 \sim V_2$」とには同じ評価値しか与えることができない．それに対し，多項関係による知識を持つ場合には，「$\sim M_1 \sim M_2 \sim V_1$」には「$\sim M_1 \sim M_2 \sim V_2$」よりも高い優先順位となる評価値を与えることが可能となる．したがってこのような場合においては，多項関係での共起知識を用いるこ

とによって初めて，有効性を持つヒューリスティックスを与えることができると言える．

この知識は，例えば「$\sim M_1 \sim M_2 \sim V_1 \sim M_2 \sim V_2$」のような文における係り受け構造を，荒いながらも高速に推定する場合や，確率文法において「格は互いに独立」という仮定をしないモデル化を行う場合，あるいは，擬人化などの可能性があるために格要素の属性による曖昧性の絞り込みが難しいような状況において，候補の絞り込みを行いたい場合などに活用できるものになると考える．

# 3 格標識共起知識の獲得

## 3.1 共起知識獲得の方針

前章で多項関係としての格標識共起知識の必要性については述べた．このような知識の獲得には，何らかのコーパスが知識元としては必要となろう．とはいえ，近年，様々なコーパスが利用可能となってはいるものの，まだまだ単項関係としての知識ですら，純粋に統計的に獲得するには分量不足と言える．多項関係としての知識の場合は，組み合わせによるパターンの数が膨大になるため，純粋に統計的に獲得することを目指した場合，単項関係の場合とは比較にならないほどの膨大なサンプルが必要となる．近い将来まで含めて，それほどの巨大なコーパスの存在に期待するのは非現実的である．

しかし，何らの知識元にも依らずに知識を獲得することはできない．そこで既存のコーパスを知識元とし，サンプルは常に不足状態にあるという前提を置く．その上で，サンプル量が多ければできるだけ素直にそのサンプル集合に従うように，サンプル量が少なければ例外の可能性をより大きく評価するようにして，できるだけ妥当性の高い知識を獲得することを目指す．

本研究では，動詞の用法に関して次の仮定を置く．

——[仮定]——

動詞には，その動詞が表すことができる状況があり，その状況を特定するに必要十分と言えるだけの状況要素集合というものが存在する．ここで言う状況要素とは，状況を構成する基本単位となる情報で，動詞によってその種類は異なり，例えば「区間」という状況要素に対して「起点格と終点格の組」が対応するというように，必ずしも格関係とは一致しない．また，各状況要素の必須性の度合も異なる．

ある動詞を述語とする文としては，必要十分と言える状況要素集合を過不足なく満たすだけの情報を含めた文が最も頻繁に生起する．それに対し，省略などにより含まれない情報の量が多くなるほど，状況記述としての不完全性が増大するために生起確率は低下する．また，付加要素として追加される情報の量が多くなればなるほど，状況記述としての冗長性，特殊性が増大するために生起確率は低下する．

この情報の省略，付加は，状況要素を基本として行われる．そのため，例えば先の例の「区間」という状況要素の必須性が比較的低かった場合，「起点格」と「終点格」とが同時に生起したり，省略されたりする可能性が高く，いずれか一方のみに言及するような文は，状況要素の一部のみを述べることになるために特殊性が強くなり，生起確率が低くなる可能性が生じる．

---

本研究では，各動詞には，状況要素集合を記述する際に頻繁に用いられる「文型」が存在すると考える．ここで言う「文型」は，文を構成する格標識の組み合わせで表すものとする．ある状況要素集合を記述する文型は必ずしも一つではない．例えば一部の状況要素の記述に用いられる格標識が二種類あるとすれば，それらの格標識が排他的となるような二種の文型が出現するであろう．特にその状況要素の必須性が高ければ，二種類の格標識の一方だけを含むような文の生起確率が高くなる．

こうした仮定に基づき，本研究では，各格標識の一文中における出現個数を一つの座標軸とするとするような多次元空間を考え，この空間上でのサンプルの分布傾向を超楕円体の形で捉える．サンプルの分布における第 1 主成分がこの超楕円体の最長軸であり，以下，順に分布の各主成分が超楕円体の軸となる．その重心は必須性を考慮した基本の文型を表し，超楕円体の軸，すなわち，分布の方向性と分散の大きさとを考慮した重心からの距離が遠いほど，出現頻度が低下する．本研究で求める「動詞の用法」とは，この超楕円体で捉えた格標識の組み合せの出現傾向のことを指すものとする．

## 3.2 共起知識の獲得手法

動詞の用法を前節で述べたように捉える場合，文を構成するある格標識の組み合わせがその文の動詞の文としてどの程度妥当であるかは，超楕円体の重心

からの距離の大小で表される．この距離は超楕円体の軸の方向性を考慮する必要があるため，Manhattan 距離や Euclid 距離ではうまく表現することができない．これらの距離では，単に単項関係知識を集約したものに過ぎず，扱うことができるのは各軸が座標軸に平行な超楕円体形の分布となる．多項関係知識が表現する格標識間の相関性を扱えない．

分布の軸と分散とを考慮した距離の一つとして，Mahalanobis 距離と呼ばれるものがある．今，分布の重心を $\bar{\boldsymbol{x}} = (\bar{M}_1, \bar{M}_2, \ldots)^t$，分散共分散行列を $S$ とするとき，分布の重心からのサンプル $\boldsymbol{x} = (M_1, M_2, \ldots)^t$ の Mahalanobis 距離 $D_M(\boldsymbol{x})$ は次のように表される．

$$D_M(\boldsymbol{x})^2 = (\boldsymbol{x} - \bar{\boldsymbol{x}})^t \, S^{-1} \, (\boldsymbol{x} - \bar{\boldsymbol{x}})$$

本研究では，距離評価としてこの Mahalanobis 距離を用いるものとする．すなわち，本研究における格標識共起知識とは，各格標識の出現数を座標軸とする空間上において，各動詞が一文で取り得る格標識の組み合せの分布における重心と分散共分散行列との情報のことである．

この重心と分散共分散行列からなる格標識共起知識を単純にサンプルから求め，それにより Mahalanobis 距離を求めようとした場合，重大な問題が生じる．Mahalanobis 距離の定義式の通り，計算には分散共分散行列の逆行列が必要である．しかしながら，分散共分散行列を単純にサンプルから求めた場合，逆行列を求めることができない，すなわち $|S| = 0$ となる場合が非常に多く発生する．このような状況は次の場合に発生する．

**Case 1 :** ある特定の格標識において，分散が 0 となる．すなわち，すべてのサンプルにおいて，ある特定の格標識が必ず同じ回数だけ出現するか，もしくは，全く出現しない．

**Case 2 :** ある特定の二つの格標識間で相関係数が 1 または $-1$ となる．すなわち，各サンプルにおけるこれら格標識の出現が，完全に同時であるか，完全に排他であるかのいずれかである．

一般に，ある動詞が絶対に取り得ない格標識は存在しうると考えられる．そのような格標識が一つでも存在すれば，上記 1 の場合が発生するし，複数存在すれば上記 2 の場合も発生する．また，サンプルが少ない場合，偶然的要因で，必須性が高い複数の格標識がすべてのサンプルで出現することは十分にありうるし，必須性が低い複数の格標識が一度も出現しないということも十分にありうる．このような場合も上記のような状況となり，Mahalanobis 距離を求めることができない．

一つの対策として考えられるのは，分散が 0 であるような格標識や相関係数が 1 であるような格標識の組を事実上切捨ててしまい，距離計算に関与させないような方法である．第 $k$ 主成分までというように次元数を落してしまう方法もこれに含まれる．しかし，サンプルが極めて多量に存在する場合ならともかく，そのような格標識における状況が偶然生じたものではないと言い切ることはできない．この点を無視して，次元の縮退を強行した場合，サンプル集合に対する例外に対応することができなくなる．ましてサンプル集合は不足状態にあるという前提であるので，そのように信頼性が低いサンプルに基づく次元の縮退など，許されることではない．

無論，格標識の集合によっては，相関係数が必ず 1 ないし $-1$ になるような格標識の組が含まれうる．十分な調査と分析の上で真にそのような関係にあると確認されれば縮退可能となるが，この点はあくまでも対象とする格標識集合の選定の問題であるため，本稿における議論の対象とはしない．

次元の縮退を行わない対策として，本研究では以下に述べるような方法を取ることにする．

まず，サンプル集合は不足状態にあるという前提から，必ず例外が存在するという仮定を置き，その例外に相当するダミーサンプルをサンプル集合に加えた上で評価する．これにより，上記 1 の場合，すなわち分散が 0 になるような状況を回避することが可能となる．本研究では，ダミーサンプルとして，すべての格標識の出現回数を 0.5 としたもの 1 個を加えるようにしている．どのようなサンプル集合に対してもこの 1 個のダミーサンプルを加えるようにすることで，サンプル集合が小さい場合は例外の可能性を大きく評価し，サンプル集合が大きい場合は例外の可能性を小さく評価するということを，統一された手法で扱うことができる．

しかし，ダミーサンプルを用いるという手法では，上記 2 の場合にうまく対処することはできない．問題となる格標識の組にだけ対応しようとすると，分散への影響が均一とはならないし，かと言って，分散への影響が均一となるようにすべてに対応しようとすると，一つや二つのダミーサンプルでは対処できない．多くのダミーサンプルを使わねばならないようでは，その影響が強すぎるということになる．

相関係数が 1 になるような状況が問題なのであるから，本研究では，そのような状態こそが特別であり，相関係数が 1 にならないような例外が存在すると仮定する．そこで，分散共分散行列の共分散成分に $(N-1)/N$ (ただし，$N$ はサンプル総数) をかけることによって，相関係数の絶対値を 1 より小さく押える方法を取る．この方法は数学的には正しくはないが，サンプル集合に影響されず統一的に行うことができ，サンプル数が大きくなればなるほど影響が小さくなる点で望ましいと考える．

### 3.3 共起知識の獲得実験

本研究では，知識獲得のためのサンプルデータとして，京都大学テキストコーパス [2] を用いた．コーパス規模としては大きなものではないが，係り受け関係が記述され，かつ，それが単に機械的に与えたものではなく，人手によって修正がなされている点が，今回の実験に都合が良いためである．実験には能動形のみを用いた．受動形，使役形については文型が変化するため，今回は除外している．

今回の実験では，京都大学テキストコーパス Version 1.0 に出現している動詞の内，能動形での出現頻度の多いものから 60 個を知識獲得実験の対象とした．単文に直した総サンプル数は 2799 個である．

サンプルとした文に出現している表現に基づき，格助詞，副助詞からなる次の 17 個を，今回の実験における格標識集合とした．

> に，が，は，で，と，も，を，
> まで，から，には，では，にも，とも，とは，
> からは，にまで，までは

格標識として扱うべき表現は他にも存在するが，今回のサンプル中に出現しなかったものについては含めていない．また，含めているものは文型に影響すると判断したものである．例えば「には」を「は」や「に」に集約してしまわなかったのは，いずれに置換えたとしても文として奇妙な表現となる可能性が高く，文型の把握に悪影響を及ぼすと考えたためである．

実験により獲得された格標識共起知識がうまくサンプルの分布を捉えているかを評価するため，今回の実験で用いた 17 次元の空間における格子点 (各座標軸の値が整数値) のサンプル数を調べた．実際の文における各格標識の出現数は整数であるため，必ず格子点に出現する．

例として，図 2 に動詞「話す」におけるサンプルの分布をグラフ化したものを示す．これは，Mahalanobis 距離が小さい格子点 (サンプル数 0 を含む) から順に存在するサンプル数を示したものである．

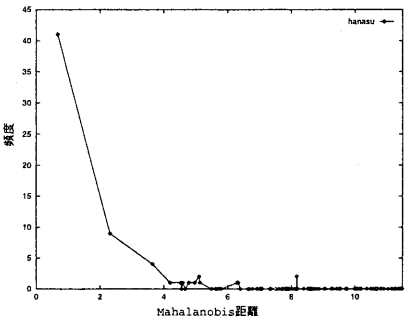


図 2: 「話す」におけるサンプルの分布

動詞「話す」については，Mahalanobis 距離が小さいほどサンプルが多く，距離が大きくなるほど少なくなるという傾向が見て取れ，用法的特徴をうまく捉えていると言えよう．

多くの動詞は，動詞「話す」と同様に，その用法的特徴をある程度うまく表現できていると考えることができる．しかし，一部にはあまりうまくは表現できていないと考えられる動詞も存在する．例として，図 3 動詞「ある」におけるサンプルの分布を示す．

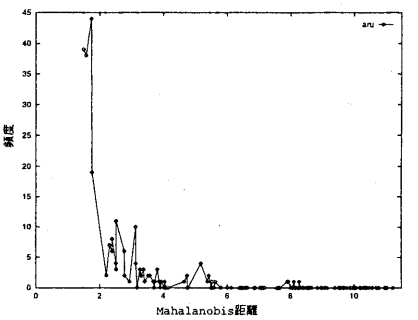


図 3: 「ある」におけるサンプルの分布

動詞「ある」の場合，重心から最近傍の格子点がサンプル数最大の格子点というわけではなく，全体的には距離が大きくなるほど少なくなる傾向はあるものの，距離増大につれてのサンプル数の増減が激しくて減少傾向の単調性に乏しい．

このようにうまく捉えることができない原因は，表層的には同一でありながら用法の傾向が異なる動詞が存在するためと考えられる．複数の用法から出現したサンプルが混在する分布を一つの超楕円体で捉えようとしたために，分布の重心付近が必ずしも密ではなくなってしまったと思われる．

# 4 動詞の多義性のクラスタリング

## 4.1 動詞の用法的多義性

動詞は一般的に多義性を持つ．語義が異なれば，その語義における中心的用法にも違いがありうると考えられるが，3.3 節で述べた共起知識獲得実験では動詞の同一性を表層的な文字列だけで判断しており，語義に基づく区別は行われていない．これは，コーパス上で語義の情報までは付与されていないためである．しかし，この点を完全に無視していたのでは，前章の「ある」の例のようにうまく用法を捉えることができないものも生じてしまう．

一般に多義と言う場合は意味的な違いを指すが，構文解析時を考えた場合には，意味的な差異よりは用法的な差異を重視して捉える方がより有用と思われる．本研究では，表層的に同一である動詞が取り得る用法的な差異を「用法的多義性」と呼ぶ．

意味的多義性と用法的多義性とは必ずしも一対一に対応するものではない．しかし，用法的な差異は人手によって多義の識別を行う場合の判断基準の一つにもなるため，用法的多義性を荒い意味的多義性の一形態として扱うこともできよう．

本研究では，この用法的多義性に基づく動詞の識別をクラスタリングによって行う．

## 4.2 クラスタリング手法

動詞の用法的多義性の識別をクラスタリングによって行う場合，次の点に注意を払う必要がある．

- いくつのクラスタを構成すべきかは不明．ただし，格子点毎に一つのクラスタとなるほどの細分化は避けねばならない．
- サンプルは空間における一部の格子点上にしか存在しない．格子点にはサンプルが多数集中する可能性がある．
- 構成したいクラスタは，その重心付近にサンプルが集中し，重心から離れるほど疎らになるような性質を持つ．

クラスタリングアルゴリズムとしては様々なものが提案され，用いられているが，これらの条件にうまく適合させることは難しい．構成するクラスタ数を仮定しない手法は多く存在するが，重心付近へのサンプル集中を保証できない．重心付近の密度が高いクラスタを構成する手法としてはモード法があるが，サンプルの存在が格子点に限られ，しかもどれだけ集中するか不明のため，うまく機能しない．

そこで本研究では，対象とするサンプルの性質を鑑み，次のアルゴリズムを提案する．

—[クラスタリングアルゴリズム]————————

**[Step 0] 初期化**

サンプルが存在するすべての格子点について，$\langle d, k, c\rangle$ なる 3 項系列を生成する．ただし，$d$ は密度評価値，$k$ は属するサンプル総数を表し，この時点においては $d = k =$ [格子点 $c$ におけるサンプル数] である．この 3 項系列の集合を $CL$ とする．また，集合 $ZC = \phi$ とする．これは，サンプルが存在しない格子点の内，クラスタに所属させられたものの管理に用いられる．

**[Step 1] 処理対象クラスタの選択**

$CL$ から密度評価値が最大の要素 $max\_d = \langle d_m, k_m, c_m \rangle$ を取り除く．$c_m$ が格子点集合なら，これはクラスタであるので，これを処理対象とする．$c_m$ が単一の格子点なら $max\_d = \langle d_m, k_m, \{c_m\} \rangle$ として新たなクラスタを生成し，処理対照とする．

密度評価値最大の要素が複数存在する場合，

(a) クラスタと格子点とでは，過度の細分化回避のため，クラスタを優先する．

(b) クラスタ同士の場合，含まれるサンプル数が多い方を優先する．それも同じであるなら，含まれる格子点数が少ないものを優先する．( 格子点同士である場合，これらの値は等しい．)

(c) $CL$ に登録されているすべてのクラスタについて，各クラスタからの距離の総和が最も大きいものを優先する．距離はクラスタ重心間のユークリッド距離で測るものとする．

(d) サンプル全体の重心からクラスタ重心位置へのマハラノビス距離が小さいものを優先する．

(e) 以上すべてで同じであるならば順位は同じとし，無作為に一つを選択する．

**[Step 2] 終了条件**

$CL$ において，$\langle d, k, c\rangle$ (ただし，$c$ は単一の格子点) なる要素が存在しなければ，$CL + max\_d$ をクラスタの集合 (要素が $\langle d, k, C\rangle$ であるとき，$C = \{c_1, c_2, \cdots\}$ で表される格子点集合を一つのクラスタとする) として，アルゴリズムを終了する．

**[Step 3] 処理対象クラスタへの格子点追加**

$max\_d$ で示されるクラスタに隣接 ( Manhattan 距離が 1 ) する格子点 $c$ を調べ，

(a) $\langle d, k, c\rangle \in CL$ か，または

(b) $c$ におけるサンプル数が 0 ( $\langle 0, 0, c\rangle$ とする ) で，かつ，$c \in ZC$ ではない

ようなもので，$d \leq d_m + \alpha$ であるものを選出する．ただし，$\alpha$ はサンプルの揺らぎを吸収するためのパラメータであり，クラスタの不必要な細分化を避ける目的で導入されるものである．

クラスタに加えられるべき格子点は，これら選出された格子点の内，クラスタに加えた際のクラスタの密度評価値の変化量が最小のものとする．密度評価値の変化量が同じ場合，対象となっているクラスタの重心からのユークリッド距離が最小のものを選択する．それも同じなら，それら複数の格子点を同時に選び，それぞれを同ステップで加えるものとする．

選択された格子点 $c$ に対し，

(a) $\langle d, k, c\rangle \in CL$ ならこの要素を $CL$ から削除し，

(b) $c$ におけるサンプル数が 0 なら $c$ を $ZC$ に加えて，$max\_d = \langle d'_m, k_m + k, c_m + c\rangle$ とする．

ただし，密度評価値 $d'_m$ は次のように求める．

(a) $d > d_m$ のとき，$d'_m = d_m$ とする．
これは，密度評価値の上昇によって，サンプル数が増す格子点が順次加えられ，分割されるべきクラスタが分割されなくなることを避けるための処置である．

(b) $d \leq d_m$ のとき，

$$d'_m = d + (((d_m - d) * |c_m|)/(|c_m| + 1))$$

とする ($|c_m|$ は $max\_d$ で示されるクラスタに属する格子点数)．
単純にサンプル数を格子点数で割ったものとしないのは，上記 (a) を考慮したためである．

**[Step 4] 反復処理**

$max\_d = \langle d_m, k_m, c_m\rangle$ において，

(a) Step 1 と同手順で選択した $max\_d' = \langle d, k, c\rangle$ に対し，$d_m \geq d - \beta$ 以上のとき，または

(b) $k_m < \gamma$ のとき，

上記 Step 2 に戻り，繰り返す．

そうでない場合は，$max\_d$ を $CL$ に加えた後，Step 1 に戻り，繰り返す．

ただし，$\beta(\geq 0)$ は，先に選択されたクラスタをどの程度優先的に扱うかに寄与するパラメータであり，$\gamma$ は，あまりに小さなサンプル数からなるクラスタが多数生成されるのを避けるためのパラメータである．

本アルゴリズムは，サンプルが多く存在する (密度が高い) 格子点から出発して，次第に疎らになっていくように周辺の格子点をクラスタに加えていく方法である．そのため，サンプルが多く存在する格子点が Manhattan 距離 1 で連続しない場合，例えば，第 2 章の動詞 $V_2$ のように格標識間に排他性がある場合，これを一つのクラスタとして捉えることができず，二つのクラスタを生成してしまう可能性が高い．しかし，単純に Manhattan 距離 2 以上の格子点間でのクラスタ生成を許すアルゴリズムとすると，交差したクラスタやドーナツ状クラスタといった問題の多いクラスタを生成する可能性がある．この点に関するアルゴリズムの改良は今後の課題としたい．

## 4.3 クラスタリング結果の評価

前節で述べたアルゴリズムを適用した結果，実験対象とした 60 個の動詞の内，17 個でクラスタが二つに分かれた．第 5 章でうまく捉えられなかった例として示した動詞「ある」もその一つである．各クラスタにおけるサンプルの分布を図 4，図 5 に示す．いずれのクラスタにおいても，重心から最近傍の格

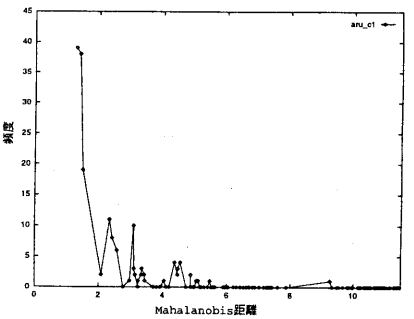


図 4: 「ある」のクラスタ 1 におけるサンプルの分布

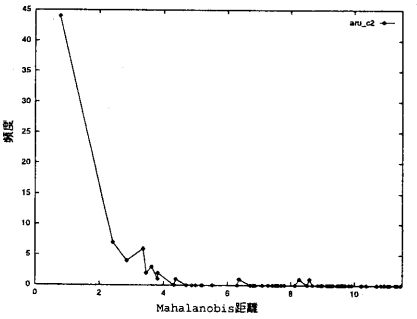


図 5: 「ある」のクラスタ 2 におけるサンプルの分布

子点がサンプル数最大の格子点となっている．距離増大につれての減少傾向は，クラスタ 2 については単調減少傾向が強くなっているが，クラスタ 1 についてはまだまだ単調性には乏しくみえる．格子点単位でクラスタ分割をしているため，クラスタ間での重複ということもありうるが，多様な用法を分離しきれていない可能性が高い．この点もアルゴリズムの改良を要する部分である．

クラスタリングによって得られた用法的多義性を評価するために，各クラスタの中心的文型 (クラスタの重心から Mahalanobis 距離的に最も近い格子点に相当する文型) と計算機用日本語基本動詞辞書 IPAL (Basic Verbs)[3] における多義分類との比較を行った．結果は全体的におおよそ良好なものであった．いくつかの動詞について，その結果を示す．

**「考える」** クラスタリングの結果,『～ を』を中心的文型とするクラスタと『～ と』を中心的文型とするクラスタとが得られたのに対し，IPAL 辞書には『$N_1$ が $N_2$ を』,『$N_1$ が $S$ と』,『$N_1$ が』,『$N_1$ が $N_2$ を $S$ と』の 4 種の文型が登録されている．
クラスタリングの結果は，格標識「が」については省略されるケースが多いことを示し，これを除外した場合，IPAL 辞書の分類と類似した分類がなされていることがわかる．

**「ある (有る/在る)」と「いる (居る)」** クラスタリングでは，いずれの動詞も『～ も』を中心的文型とするクラスタと『～ が ～ に』を中心的文型とするクラスタとが得られた．IPAL 辞書では「ある」には「が」格を二つ持つ文型も存在するが，これは『～ が ～ に』の用法の一部とされ，単一の用法としては切り出されていない．
これらの動詞はいずれも存在を表す性格を持つため，似た用法的性格を持つとされたことも自然なことと言えよう．また，格標識「も」が用いられる場合，前出の文を受けて用いられることが多く，それゆえ他の格要素が省略されやすいという性質が現れていると見なせることが興味深い．

**「続く」** 『～ が』を中心的文型とするクラスタと『～ は』を中心的文型とするクラスタとが得られた．これは格標識「が」に代って格標識「は」が用いられていると考えられるため，過度にクラスタ分割が行われたものと言えよう．これは，4.2 節で述べたアルゴリズムの問題点が現れた例である．

**「見つかる」** 『～ が ～ で』を中心的文型とするクラスタと『～ が ～ から』を中心的文型とするクラスタとが得られたのに対し，IPAL 辞書には『$N_1$ が』と『$N_1$ が $N_2$ を』とが登録されている．
これは，IPAL 辞書では重視されていなかった発見場所に関する情報が，比較的強い必須性を持つことを表していると言える．

**「思う」** IPAL 辞書では『$N_1$ が $N_2$ を』と『$N_1$ が $S$ と』とが登録されているが，クラスタリングでは,『～ と』を中心的文型とする単一のクラスタが得られた．格標識「を」を含むサンプルも存在したが，格標識「と」を用いる文型が圧倒的に多かったために，明確なモード (高密度点) を形成できなかったものと考えられる．

## 5 おわりに

本稿では，動詞と格要素との共起知識を多項関係で獲得することの必要性を示し，その一例として格標識の組と動詞との共起知識を獲得するための手法を提案した．本手法は，格標識の集合と係り受け関係付きのコーパスとが与えられれば人手に寄らず実行可能であり,「 ～ に対する」のような格標識相当語句の追加などの変更を行っても容易に新たな知識を獲得することができる．また，クラスタリングで得る用法的多義性は，従来あまり扱われていなかった側面から動詞の多義性を捉えたものと言える．

本稿で獲得された共起知識は，数少ないサンプルから人手をかけずに求めたものであるにもかかわらず，多大なる人的資源を投入して獲得された IPAL 日本語基本動詞辞書と比較しても，なかなか良好と言えるものになっている．これは，本手法の有効性を示す一例と考えることができるだろう．

## 参考文献


[1] 奥野忠一 他:多変量解析法《改訂版》, 日科技連出版社,1986

[2] 黒橋禎夫 他:京都大学テキストコーパス 作業マニュアル, 京都大学テキストコーパス Version1.0 付属ドキュメント,1997

[3] 計算機用日本語基本動詞辞書 IPAL (Basic Verbs) 説明書, 情報処理振興事業協会技術センター,1988