# 助数詞「本」のカテゴリー化をめぐる一考察

JCLA5（2005）ワークショップ
「コーパス利用とこれからの認知言語学」
2005/9/18


濱野　寛子
京都大学大学院

李　在鎬
情報通信研究機構

# 0. お品書き

1. **本研究の概要** **<2枚>**
2. **記述的問題と先行研究** **<3枚>**
   - 問題の根底にあるもの、先行研究の問題点
3. **代替案（補強案）** **<4枚>**
4. **本研究の方法論** **<7枚>**
   - データ作成手順とクラスタアルゴリズムの紹介
5. **結果と考察** **<7枚>**
6. **最後に（今後の課題など）**

# 1. 本研究の概要 （1/2）

1. **目的**
   i. 多変量解析の手法を導入し、助数詞「本」をコーパスベースに記述・分析。
   ii. 認知言語学におけるカテゴリー化モデルの精緻化。同時にコーパス利用の有効性を示す。
2. **主張・提案**
   i. 記述レベル：助数詞「本」の使用には、対象の外的属性に基づく制約のみでは不十分。身体的経験(限定された意味での「視点」)に基づく特徴づけが必要
   ii. 方法論レベル：多変量解析は大量の言語使用のデータの隠れた性質を効率よく発見できる強力なツール

# 1. 本研究の概要 (2/2)

3. **論証**
   i. 記述レベル→二つのクラスタ分析結果を比較
      - 外的な属性の変数(e.g. 形状)での分析結果と身体論的特徴づけの変数(e.g. 振回す)を取り入れた分析結果を比較。後者が助数詞「本」の事実により忠実であることを示す。
   ii. 方法論レベル→K-means法による解析結果を報告
      - プロトタイプを計算論的に推定し、カテゴリー化の実態を可視化する
4. **意義・示唆**
   i. コーパス言語学から認知言語学へ：UBMの実質化
   ii. 認知言語学からコーパス言語学へ：理論の提供

# 2. 記述的問題と先行研究（1/3）

1. **問題の根底にあるものは？**
   i. 数える対象が明示されてない
      1. まずは一本の電話から・・・。「駅前の歯医者さん」
      - 「電話」ではなく「通話の回数」を数える。

      論証）* 一本の電話を{壊した,落とした,分解した,破壊した}。

   ii. 規範的制約のみでは記述しきれない
      2. 敗者復活戦で一本の魚しか検量してもらえなかった。
         「Sports Fishing Japan」
      - 「匹」でも「本」でも数えられる。

*生成語彙論 (小野 2005) の考え方によれば、「一台の電話」と「一本の電話」の違いは「電話」という名詞の形式クオリア単位での数えか目的クオリア単位での数えかの違いによるものである。この種のメトニミー/活性領域の影響は本論で扱う助数詞に限ったことではなく、語の多義性の全般において観察される問題

# 2. 記述的問題と先行研究（2/3）

## 2. 従来のアプローチ

### i. 拡張に基づく一般化（プロトタイプカテゴリー）

- 形式的特徴からの拡張：「細長さ(e.g. 鉛筆)」「棒状のものの握り方(e.g. バット)」のような物理的な特徴の一部が、軌跡・作品・情報など抽象的領域へ拡張（西光 2004）
- 「際立った一次元的なもの」(プロトタイプ) からの拡張：巻かれているもの、容器の形状のメトニミ的拡張、経験的一次元性(e.g. 小説、論文)や軌道形成(e.g. 電話の通話、ホームラン)の拡張（Matsumoto 1993）

### ii. 記述的研究

- 現象の網羅的記述：辞書的考察（飯田 2004）

# 2. 記述的問題と先行研究（3/3）

3. **問題点**
   i. 制約のあり方をめぐって
      i. 制約が一般的すぎて解釈次第ではどうにでもなる
         - e.g., 一続きの経験でなければならない。
         - e.g., 細長さ（絶対的な属性にならず）
      ii. 相互排他的変数でない以上、どの制約がどれだけ関与しているかを明確にされていない。
   ii. 予想される問題点
      i. 過剰般化の危険性（経験事実面での反例は濱野2005参照）
      ii. 応用性の欠如（分析結果を共有できる形にすべき）

# 3. 代替案（補強案） （1/4）

1. **理論的方向付け（濱野 2005）**
   i. 認知主体の主観性に関する問題
      i. 助数詞には我々の**話者としての主観性**が多分に反映される(cf. 井上1999)
      ii. 助数詞の使用は、話者や状況によって捉えられた対象の側面応じて変わり、また、そうした用法は、**人間と環境のインタラクション**によってもたらされるものである(cf. Denny1976, 1979)
   ii. 認知主体の視点ベースの分析モデル
      i. 認知主体が**どのような視点で対象を捉え**ているのか。
         3. 猟師が5本の秋刀魚を釣った。
         4. ?お母さんが5本の秋刀魚を買った

# 3. 代替案（補強案）　（2/4）

iii. しかし！！・・先行研究同様の問題点

i. 「視点」とは何か：単なるメタファ。定義次第で何もかもが視点になる。一般化が困難。抽象的すぎる。
宅どうにでも判断できてしまう

ii. 具体的・限定的に捉えなおす必要がある

iv. アフォーダンスからのヒント（p.c. 仲本）

i. 身体を介した具体的な経験で捉え直す

1. 大きさや太さ：握れるか, 人が入れるか, 振り回せるか
2. 容器性：ものを詰められるか、立体か
3. 専門性：対象への制御可能性やレスポンスビリティ等

ii. 複数の変数の競合と創発的効果として位置づける

# 3. 代替案（補強案） (3/4)

## 2. 分析手法（定量的分析）（李 2004）

### i. コーパス利用

- コーパスからデータを収集することで、
  - 安定したサンプルの収集。観察バイアスを軽減。
- データセットを明確にすることで分析精度のレベルを明確化（○○データで調べた結果において○○だ）
  - すべての事実に対して万能である必要はない
  - 次の分析へ結果を継承させ、比較可能、という利点あり

### ii. 統計的手法を利用

- 文脈情報に基づく一般化。完全なボトムアップ方式であり、データ駆動型の記述
- 多変量による解析。直感をうまくコントロール

# 3. 代替案（補強案） (4/4)

## 3. 先行研究に対して

### i. 身体論的視点ベースのモデルの採用による利点

- 物理的属性基盤モデルに対してよりダイナミックな分析が可能（言語事実面での詳細な議論は濱野(2005)参照）
- 抽象概念(大きさ、細さ、長さ)を具体概念(握れる、入れる)化することで特徴づけの判断を容易にさせる

### ii. 定量的分析モデルの採用による利点

- 制約のあり方を明確化（どの要素にどの制約がどれくらい関与しているのか）
- 拡張における相互関係を明らかにする（ある要素とある要素がどれくらい似てるか、どれくらい拡張してるか）

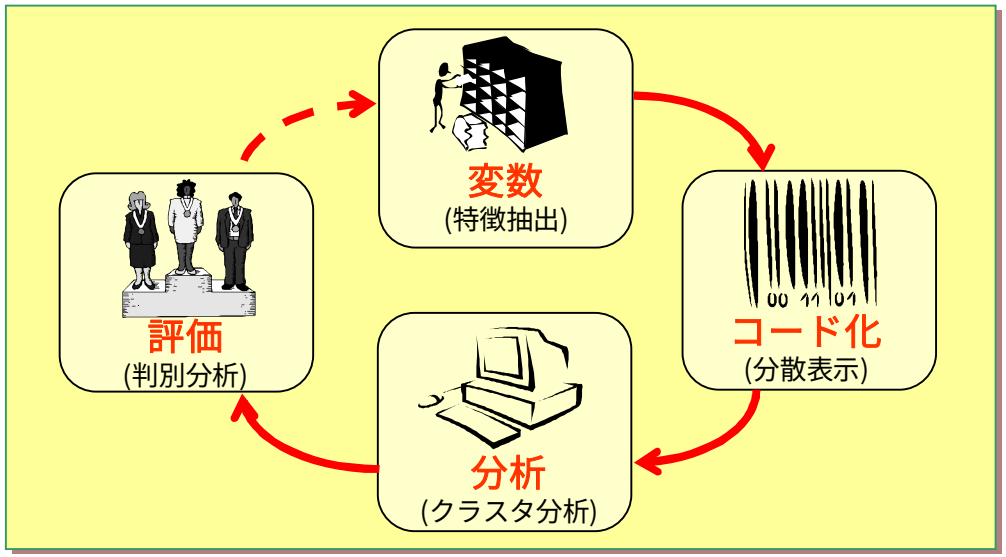
# 4. 方法論 （1/7）

## 分析モデル（手順など）

**Stage1　サンプル収集**

1. コーパスの決定
2. KWIC検索
3. ノイズの除去
4. コンコーダンス

**Stage2　分析サイクル**



# 4. 方法論 (2/7)

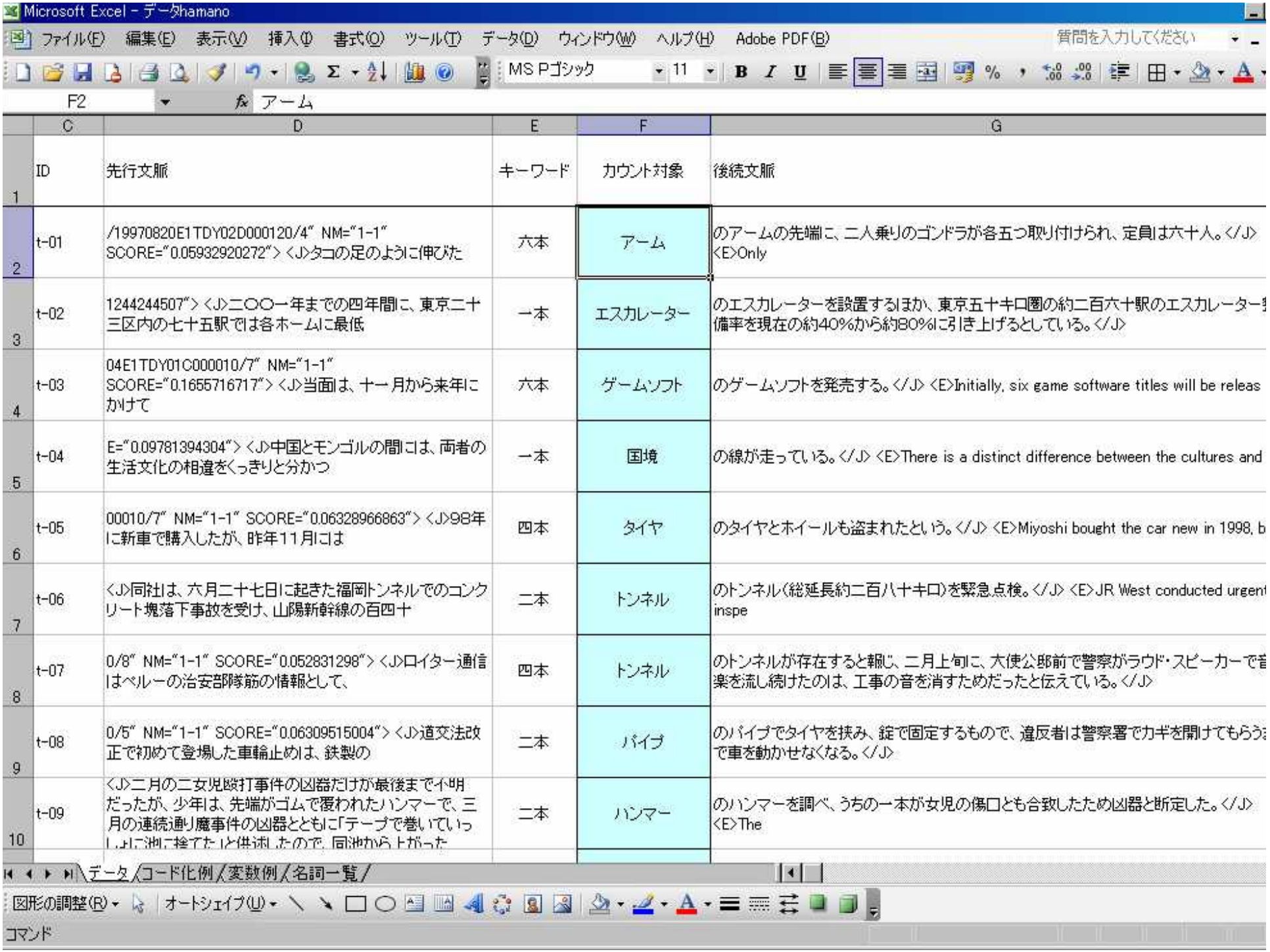
## 1. サンプル収集

### i. コーパス：日英新聞記事対応付けデータ

- 読売新聞と The Daily Yomiuri から自動作成された日英対応付けコーパス。
- Web上で利用可能。無料

### ii. KWIC検索

- 構文のバイアスを抑える目的で統語パターンを限定
  宅　「X本のY」（e.g., 3本の鉛筆）
- KWIC Finder 3.21（http://www31.ocn.ne.jp/~h_ishida/KWIC.html）を利用
- ノイズ（日本の花火）除去後、96例のサンプルを得る


Microsoft Excel - データhamano

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H) Adobe PDF(B) 質問を入力してください

MS Pゴシック 11

F2 アーム

| | C | D | E | F | G |
|---|---|---|---|---|---|
| 1 | ID | 先行文脈 | キーワード | カウント対象 | 後続文脈 |
| 2 | t-01 | /19970820E1TDY02D000120/4" NM="1-1" SCORE="0.05932920272"> <J>タコの足のように伸びた | 六本 | アーム | のアームの先端に、二人乗りのゴンドラが各五つ取り付けられ、定員は六十人。</J> <E>Only |
| 3 | t-02 | 1244244507"> <J>二〇〇一年までの四年間に、東京二十三区内の七十五駅では各ホームに最低 | 一本 | エスカレーター | のエスカレーターを設置するほか、東京五十キロ圏の約二百六十駅のエスカレーター整備率を現在の約40%から約80%に引き上げるとしている。</J> |
| 4 | t-03 | 04E1TDY01C000010/7" NM="1-1" SCORE="0.1655716717"> <J>当面は、十一月から来年にかけて | 六本 | ゲームソフト | のゲームソフトを発売する。</J> <E>Initially, six game software titles will be releas |
| 5 | t-04 | E="0.09781394304"> <J>中国とモンゴルの間には、両者の生活文化の相違をくっきりと分かつ | 一本 | 国境 | の線が走っている。</J> <E>There is a distinct difference between the cultures and |
| 6 | t-05 | 00010/7" NM="1-1" SCORE="0.06328966863"> <J>98年に新車で購入したが、昨年11月には | 四本 | タイヤ | のタイヤとホイールも盗まれたという。</J> <E>Miyoshi bought the car new in 1998, b |
| 7 | t-06 | <J>同社は、六月二十七日に起きた福岡トンネルでのコンクリート塊落下事故を受け、山陽新幹線の百四十 | 二本 | トンネル | のトンネル(総延長約二百八十キロ)を緊急点検。</J> <E>JR West conducted urgent inspe |
| 8 | t-07 | 0/8" NM="1-1" SCORE="0.052831298"> <J>ロイター通信はベルーの治安部隊筋の情報として、 | 四本 | トンネル | のトンネルが存在すると報じ、二月上旬に、大使公邸前で警察がラウド・スピーカーで音楽を流し続けたのは、工事の音を消すためだったと伝えている。</J> |
| 9 | t-08 | 0/5" NM="1-1" SCORE="0.06309515004"> <J>道交法改正で初めて登場した車輪止めは、鉄製の | 二本 | パイプ | のパイプでタイヤを挟み、錠で固定するもので、違反者は警察署でカギを開けてもらうまで車を動かせなくなる。</J> |
| 10 | t-09 | <J>二月の二女児殴打事件の凶器だけが最後まで不明だったが、少年は、先端がゴムで覆われたハンマーで、三月の連続通り魔事件の凶器とともに「テープで巻いていっしょに池に捨てた」と供述したので、同池から上がった | 二本 | ハンマー | のハンマーを調べ、うちの一本が女児の傷口とも合致したため凶器と断定した。</J> <E>The |

データ / コード化例 / 変数例 / 名詞一覧 /

図形の調整(R) オートシェイプ(U)

コマンド

# 4. 方法論 (3/7)

## 2. コード化

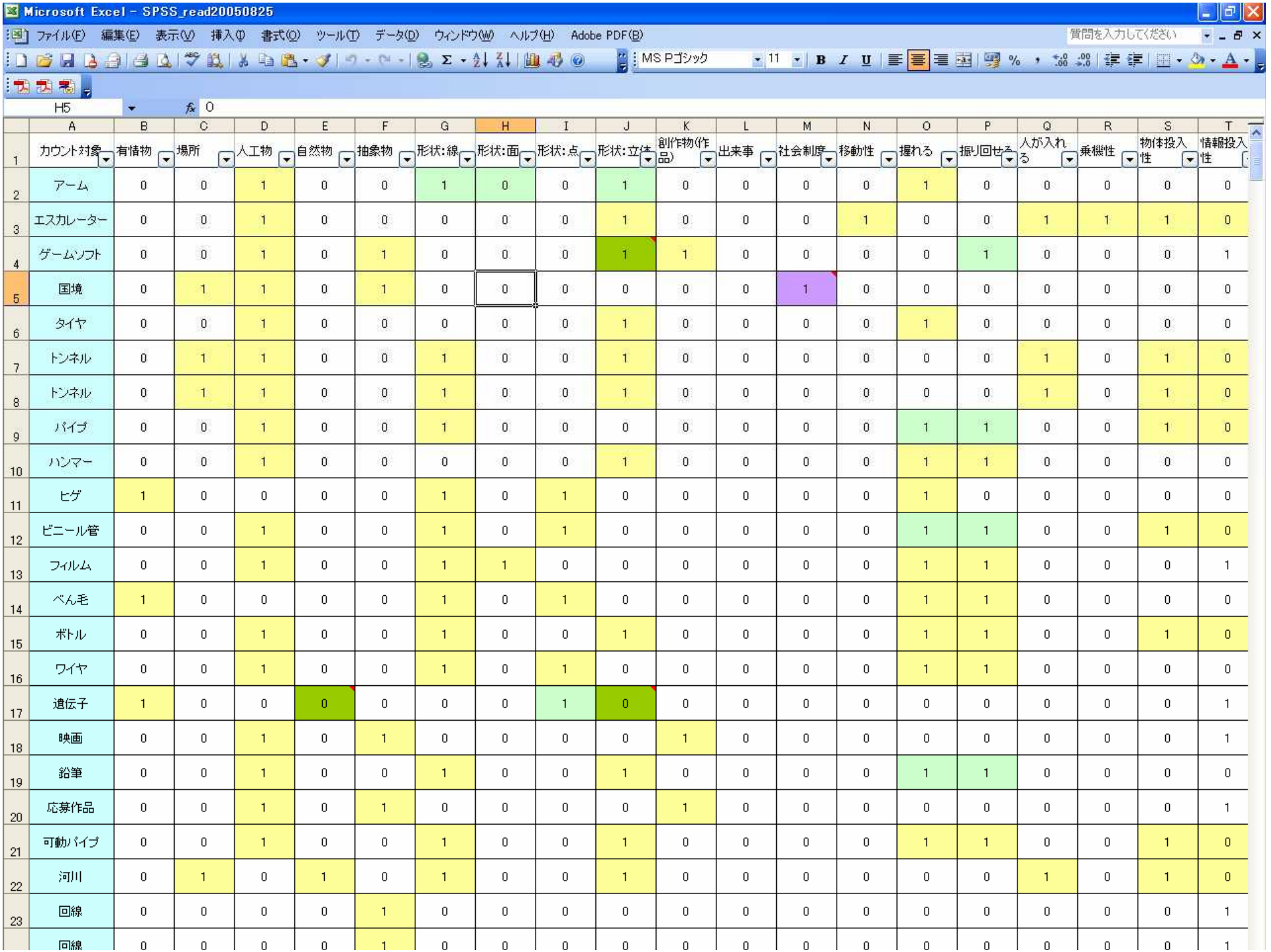
### i. 変数の特定

- 最適な変数セットの条件：種差を捉えつつ、類似関係が定義可能な特徴づけ。

### ii. 三種類の変数セットを利用

- カウント対象の外的な属性に基づく特徴づけ
  - 人工物、自然物、抽象物、形状1 [線]、形状2 [面] 等
- カウント対象に対する主体の関わり方に基づく特徴づけ
  - 握れる、振り回せる、人が入れる、乗機性、物体投入性等
- 文脈情報（主語との関係、統語的環境など）
  - 主語の制御可能性、格関係(が・を・に・で)

### iii. 1/0で判定し、分散表示する


Microsoft Excel - SPSS_read20050825

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H) Adobe PDF(B)

質問を入力してください

MS Pゴシック 11

H5 0

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | カウント対象 | 有情物 | 場所 | 人工物 | 自然物 | 抽象物 | 形状:線 | 形状:面 | 形状:点 | 形状:立体 | 創作物(作品) | 出来事 | 社会制度 | 移動性 | 握れる | 振り回せる | 人が入れる | 乗機性 | 物体投入性 | 情報投入性 |
| 2 | アーム | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | エスカレーター | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| 4 | ゲームソフト | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 5 | 国境 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | タイヤ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | トンネル | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 8 | トンネル | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 9 | パイプ | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 10 | ハンマー | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 11 | ヒゲ | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 12 | ビニール管 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 13 | フィルム | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 14 | べん毛 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 15 | ボトル | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 16 | ワイヤ | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 17 | 遺伝子 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 18 | 映画 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 19 | 鉛筆 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 20 | 応募作品 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 21 | 可動パイプ | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |
| 22 | 河川 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 23 | 回線 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 回線 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

# 4. 方法論 （4/7）

## 3. 統計解析

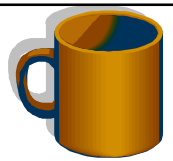
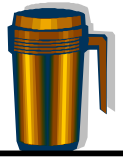
### i. 多変量解析

- 複数の変数を同時に分析する統計分析手法
- 因子分析、主成分分析、多次元尺度分析（Croft 2005)など
- データの隠れた性質を発見するのに良い方法

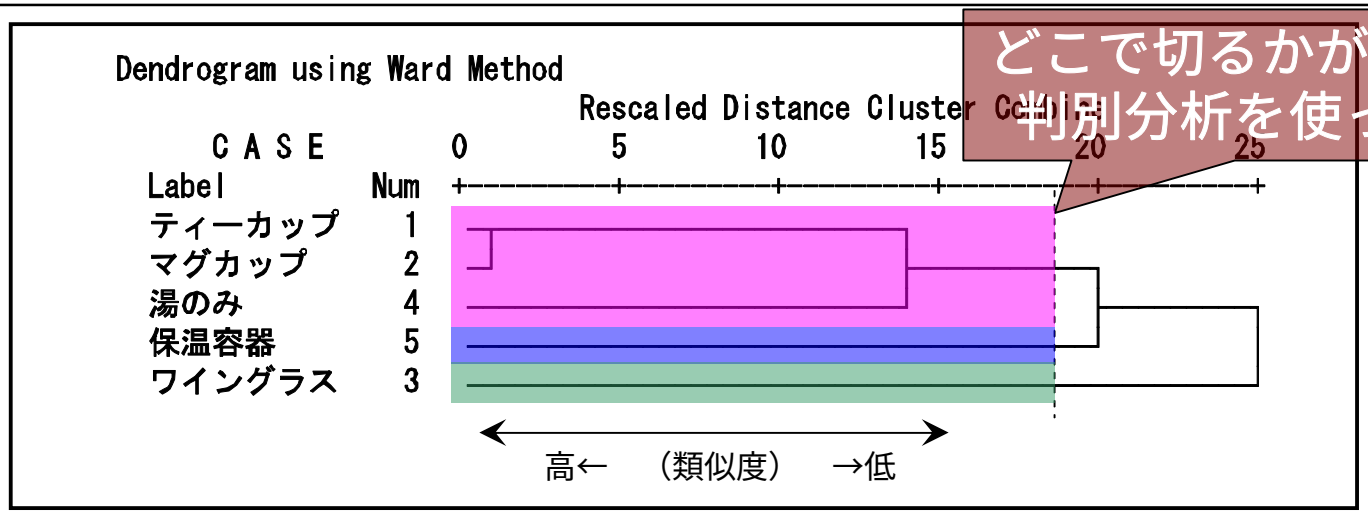
### ii. クラスタ分析

- 似通った個体グループ化を行うための分析手法：階層的クラスタ法と非階層的クラスタ法で大別できる。
- 探索的手法で、言語事例の分析に適している（李 2004）
- SPSS（Win）Ver. 13.0を使用

# 4. 方法論 （5/7）

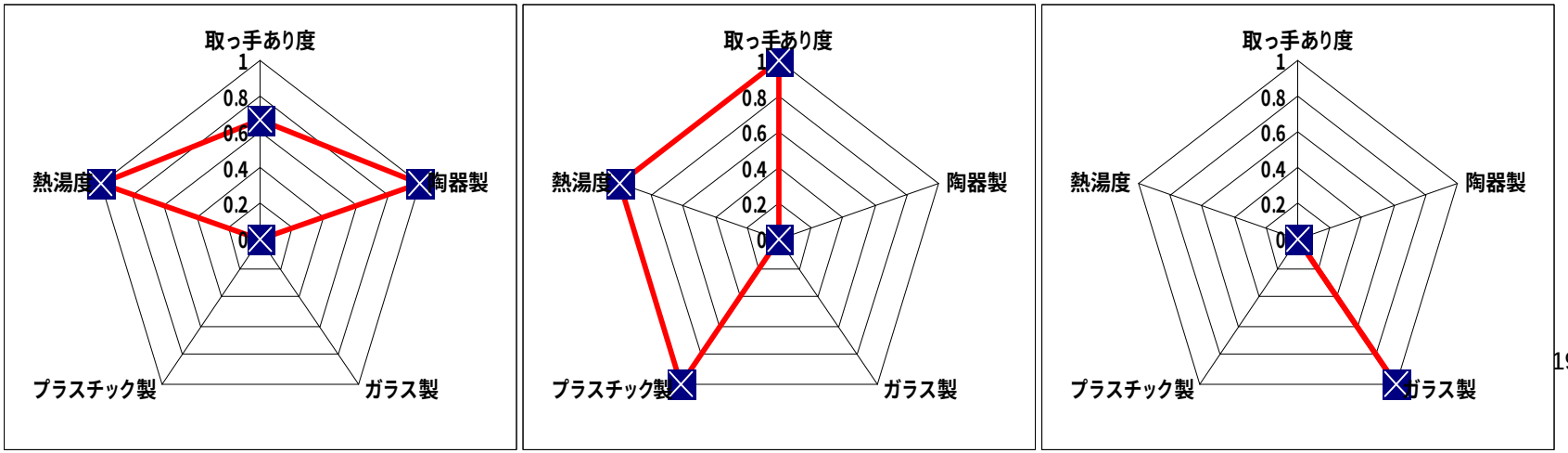
| | 持ち方 | 材質 | 液体の特徴 |
|---|---|---|---|
| 1  | 取っ手あり | 陶器 | 熱い |
| 2  | 取っ手あり | 陶器 | 熱い |
| 3  | 取っ手なし | ガラス | 冷たい |
| 4  | 取っ手なし | 陶器 | 熱い |
| 5  | 取っ手あり | プラスチック | 熱い |

# 4. 方法論 （6/7）




どこで切るかが重要だ！
判別分析を使って検証

K-means法によるクラスタ中心



# 4. 方法論 （7/7）

## 4. 適用の仕方

### i. 比較検討

a. 外的属性に関わる変数郡のみでクラスタ解析した場合**（分析A）**と、身体論的変数や文脈情報などもすべての変数を入れてクラスタ解析した場合**（分析B）**を比較

b. クラスタ化の最適度の評価には「判別分析*」を用いた。

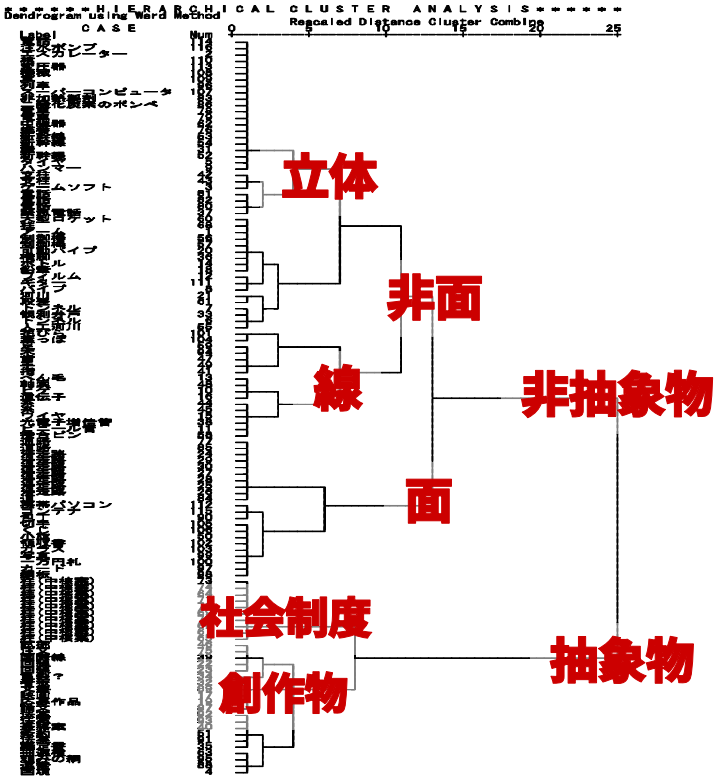
c. K-means**法**（分析C）**を使ってプロトタイプを求める。カテゴリー化の動機付けを明らかにする。拡張の度合を数値化する。

* 個体（対象者）の特性（変数）から、その個体（対象者）がどの群に属するかを判別する手法です。

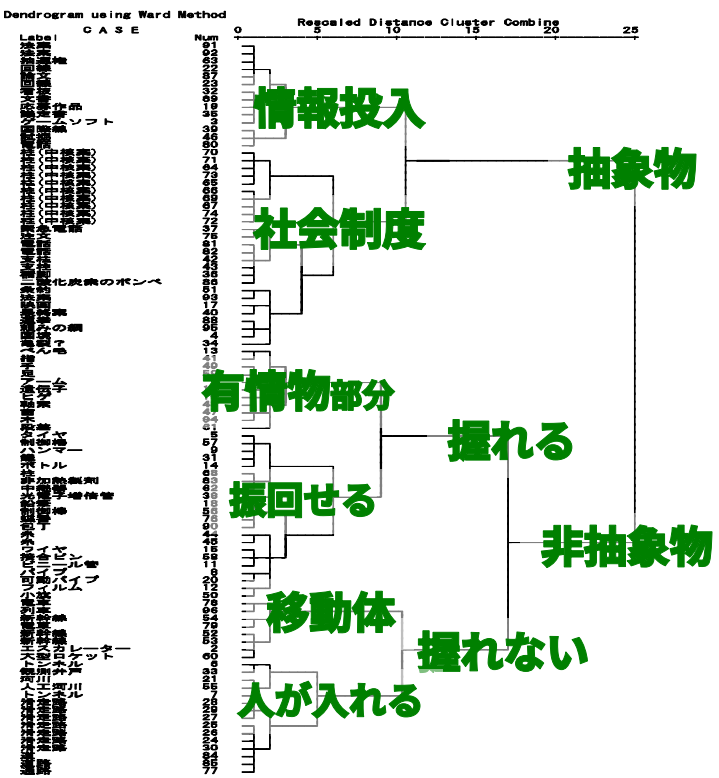
** Partitioning Methodと呼ばれるクラスタリングの一種で、データを与えられたk個のクラスターに分割する。K-Means法は、クラスターの中心値をそのクラスターを代表する値とする。

# 5. 結果(階層的クラスタ法) （1/8）

1. 分析A



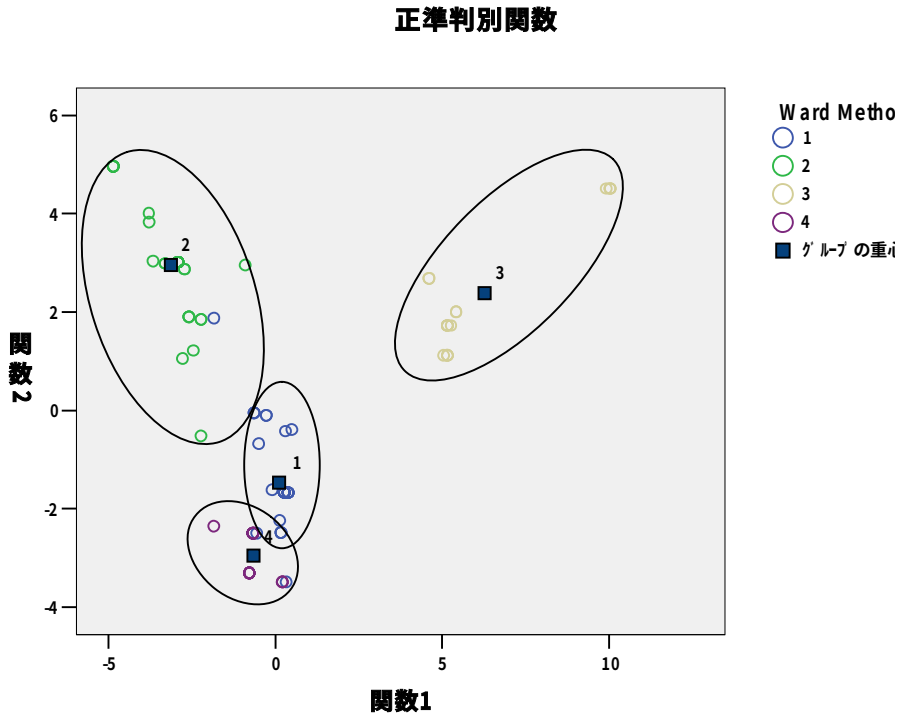

2. 分析B




Ward法、ユークリッド距離によるデンドログラム

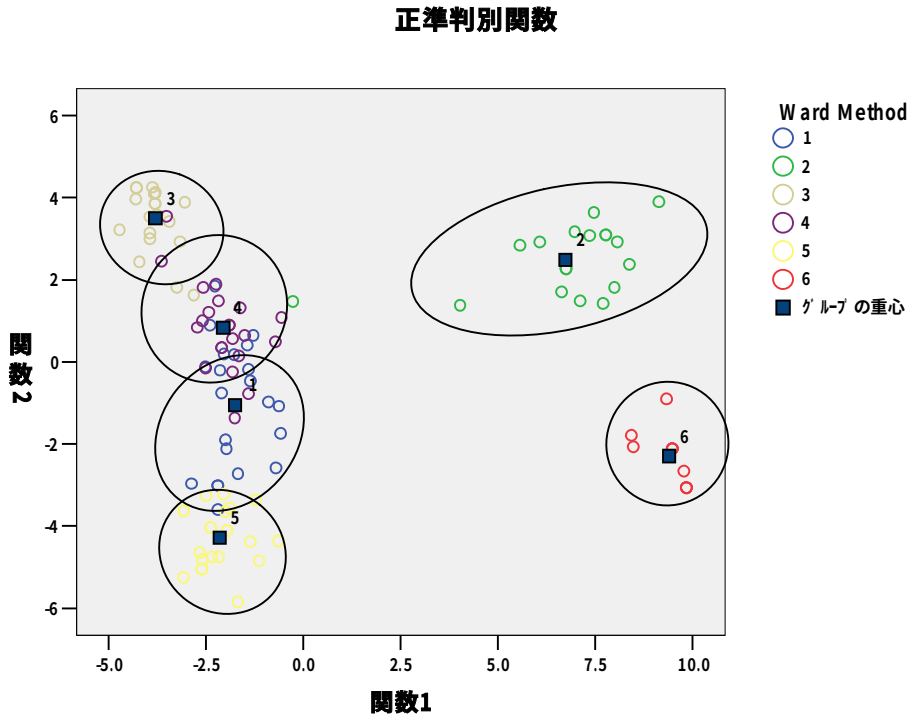
# 5. 結果(判別分析)　(2/8)

## 1. 分析A




判別率 98.6%

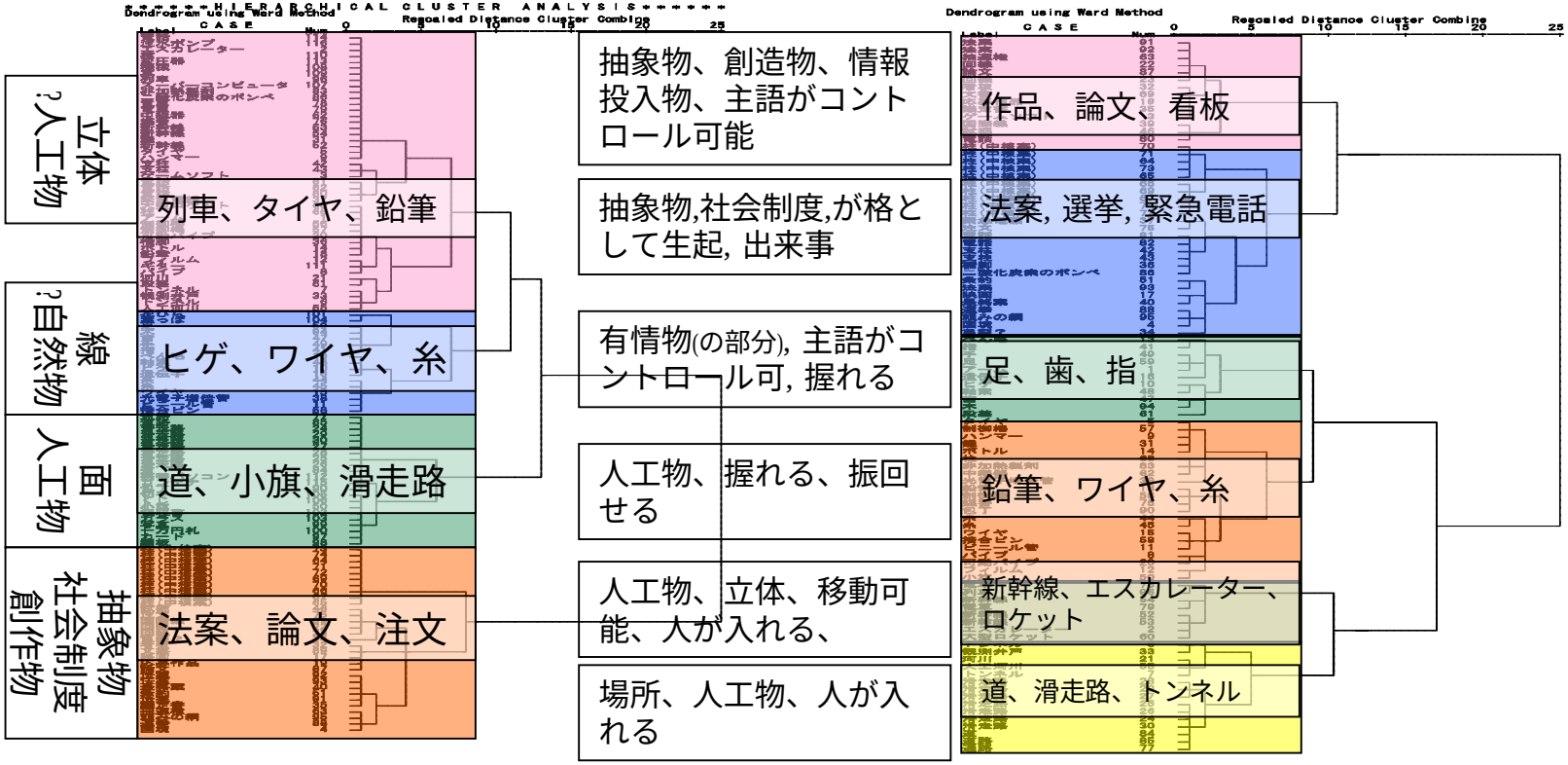
## 2. 分析B




判別率 94.6%

# 5. 結果(階層的クラスタ法)　(3/8)

1. **分析A**




2. **分析B**

# 5. 結果 (4/8)

## 3. 比較１

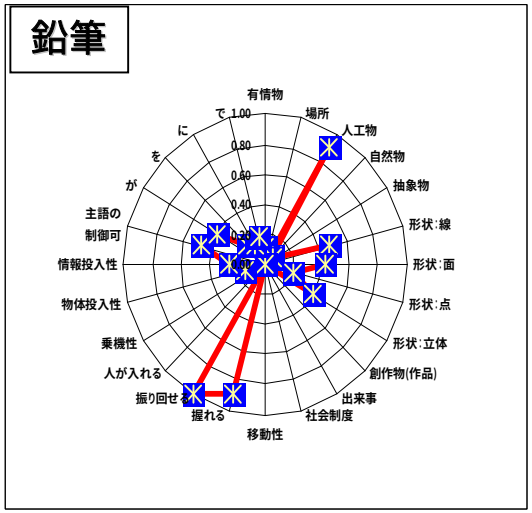
- 階層的クラスタ分析の結果から
  a. 分析A、Bに共通してみられる特徴
     - 抽象物か否か大きく関与
     - 立体物か否かが重要な要因
  b. 分析Aは、物の分類としてはエレガント。だが、本の動機付けを反映していない（容器性や軌道性、専門性など）。
  c. 分析Bは、細長い物とそうでない物、容器と非容器、専門的なものとそうでない物を近似によってうまくカテゴリー化している。
  d. 物の形状は関わり方の変数で擬似的に表現可能（むしろ、それによって生じるものと捉えるべきではないか？）

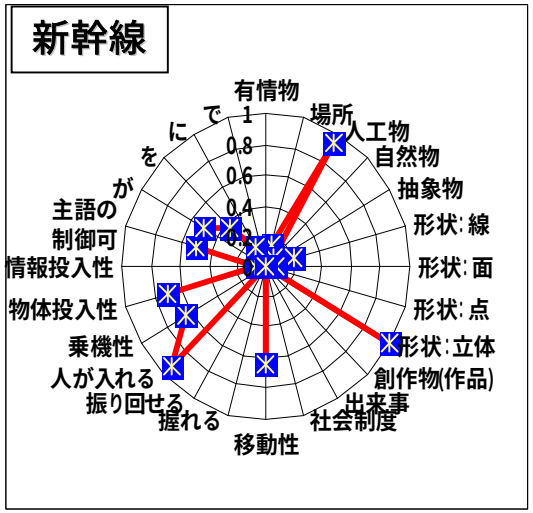
# 5. 結果 (5/8)

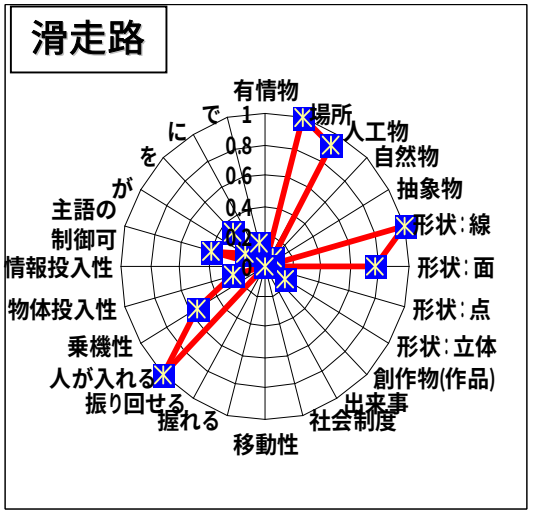
4. **比較２**

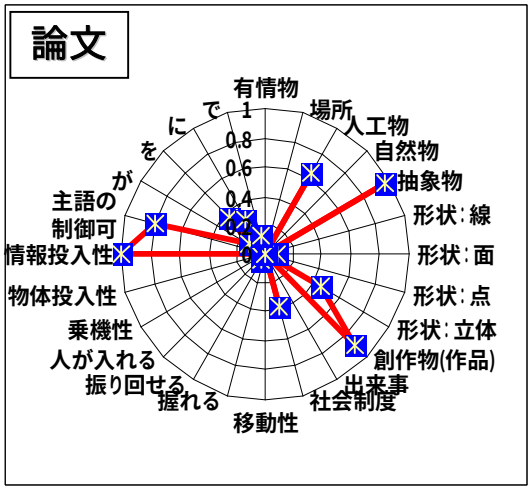
- 判別分析の結果から
  a. 分析Aの場合：三つの合成関数によって判別。一つの合成関数によって7割強(76.70%)が分類されている宅**単純**
  b. 分析Bの場合：五つの合成関数によって判別。宅**複雑**
  c. 変数間の相関（構造行列から）
     - 分析A：「線・自然物（*抽象物は負の相関）」
     - 分析B：「人が入れる・乗れる」、「線・握れる・振回せる（*抽象物は負の相関）」
     - 分析A・B：「立体」はいずれにおいても独立性が強い

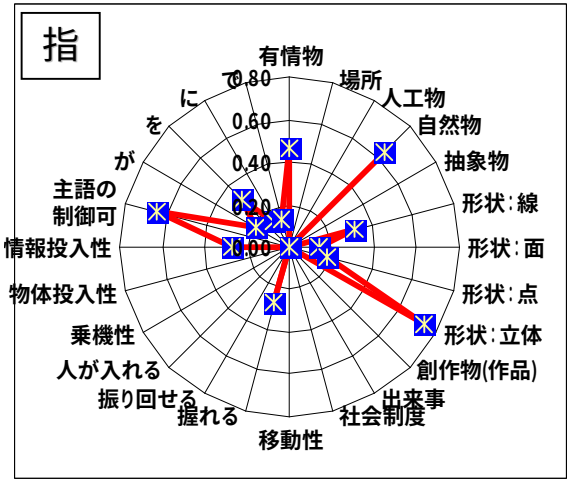
# 5. 結果（分析C: K-means法）　（6/8）

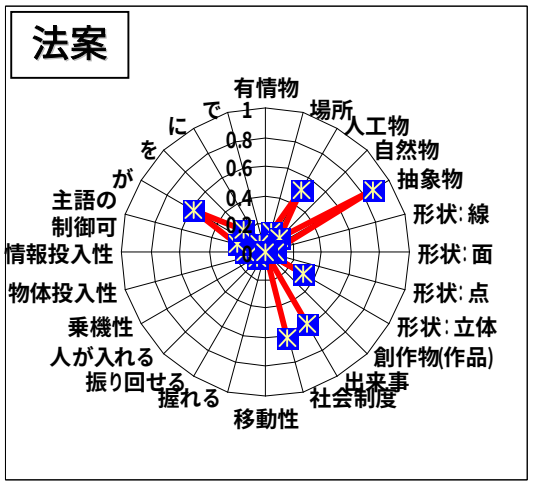


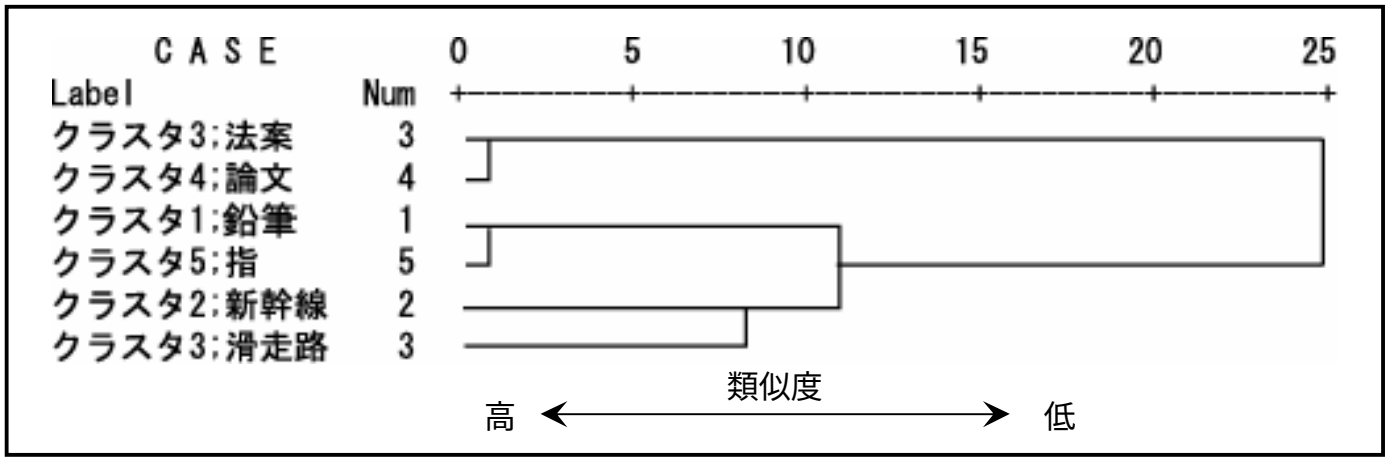



# 5. 結果（分析C: K-means法）　（7/8）

5. **カテゴリー内の相互関係(遠近による拡張度, 中心メンバーを特定)**

| 対象 | 中心からの距離 | | 対象 | 中心からの距離 |
|---|---|---|---|---|
| 鉛筆 | 1.160 | | ワイヤ | 1.500 |
| 制御棒 | 1.160 | | アーム | 1.516 |
| ハンマー | 1.349 | | 橋脚 | 1.531 |
| 鎌 | 1.349 | | 包丁 | 1.547 |
| ボトル | 1.451 | | 小旗 | 1.636 |

6. **カテゴリー間の相互関係（類似関係）**



# 5. 結果 （8/8）

## 1. プロトタイプカテゴリーモデルに対する示唆

① プロトタイプの存在を仮定しない分析の妥当性

- ➔ クラスター化は拡張によって形成されるのではない
- ➔ 個々の事例の類似度の大小によって形成され，各々のクラスターの中心がプロトタイプとして特徴づけられる。
- ➔ プロトタイプは派生的に定義される（プロトタイプは解析の結果であって、前提ではない）

## 2. 理論的帰結

① 「プロトタイプからの拡張」という説明は本当に妥当な説明なのか

② 「プロトタイプは妥当な記述のために不可避な概念」ではない

# 6. 最後に （1/2）

## 1. まとめ

- 記述面：

① 六つのクラスタで最適な分離が得られた。(新聞コーパスでは)

② 抽象であるが否かの対立はいずれの解析結果にも保持

③ 統語的特徴はあまり顕著ではない

- 方法論面：

① カテゴリー化の内部構造を見ることが可能

② カテゴリー間の関係を明示的に示すことが可能

③ 隠れた性質を発見することが可能

④ 抽象概念を前提にせず事実を分析

# 6. 最後に (2/2)

## 2. 今後の課題と方針

① データの拡大（テキスト間の比較なども）

② 負例の導入による検証

- 現在は記述レベルに留まっており使い分けの議論がない。
  宅根本問題2(匹・本)には答えてない
- 自己組織化マップで検証予定

# 謝辞


1. 中本 敬子（京都大学教育学部）
2. 仲本 康一郎（情報通信研究機構）
3. 山梨 正明（京都大学人間環境学研究科）
4. 井佐原 均（情報通信研究機構）


**ご清聴 ありがとうございました!!!**