

# テキストマイニングの実習

## －2日目－

2016/7/13

ビジネス科学研究科  
経営システム科学専攻

# スケジュール

- 7/6
  - 説明 データ分析の手順
  - 演習 データの理解 (Excel)
- 7/13
  - 説明 ツール (KHCoder)
  - 練習 ツール (KHCoder)
- 7/20
  - 演習 データの分析 (KHCoder)

# KH Coder –立命館の樋口先生が開発

社会調査データを分析するために開発された  
フリーのテキストマイニングツール

- 高機能でも商用可能でフリー
- Rを用いた多変量解析と可視化
- 実装されている分析手法
  - 階層的クラスター分析
  - 多次元尺度構成法(MDS)
  - 対応分析
  - 共起ネットワーク
  - 自己組織化マップ
  - 文書のクラスター分析

[KH Coderを用いた研究事例のリスト](#) < 1646件

論文検索サービスも提供 →  
[http://khc.sourceforge.net/bib.html?  
year=2017&auth=all&key=](http://khc.sourceforge.net/bib.html?year=2017&auth=all&key=)

[\[ KH Coder \]](#)

KH Coderを用いた研究事例のリストです。※KH Coderを用いたご研究の成果を発表された際には、書誌情報をお送りいただけますと幸いです。

出版年： -2005 06 07 08 09 10 11 12 13 14 15 16 2017-  
著者名： あ か さ た な は ま や ら わ A-Z  
キーワード：  
ヒット件数： 090 / 1646

飯村諭吉・時得紀子 2017 「初等教員養成課程の音楽指導法をめぐる実践的考察 — アクティブラーニングによる身体表現活動に焦点を当てて—」『教育実践学論集』 18: 163-171

※ 2017/6/25 現在 (一昨年961件→昨年1206件)

# KH Coder の情報

## • ホームページ

<http://khc.sourceforge.net/>

The screenshot shows the official website for KH Coder. At the top, there's a large blue banner with the 'KH Coder' logo. Below it is a navigation bar with links for Japanese and English. The main content area has several sections: 'Index', '概要' (Overview), '機能紹介' (Function Introduction), 'KH Coderの入手' (How to Get KH Coder), and 'サポート' (Support). A sidebar on the right contains a list of tweets from the account @khcoder and a link to a presentation slide titled 'KH Coder の本 ご好評発売中！'.

## • スライドも充実

<http://www.slideshare.net/khcoder/presentations>



## • 参考書



# KH Coder の主な分析手法

分析手法	解説
階層クラスタリング	<ul style="list-style-type: none"><li>出現パターンの似た単語をクラスタリングしたもの</li><li>出現パターンは,ある単語がどの文書に出現したかといった単語ベクトルで表現</li><li>類似度計算には Jaccard, ユークリッド, コサイン距離を用い,いわゆる Ward法, 群平均法, 最遠隣法で樹形図を作成</li></ul>
多次元尺度構成法	<ul style="list-style-type: none"><li>出現パターンの似た単語を近くに置くよう図示したもの</li><li>出現パターンは,ある単語がどの文書に出現したかといった単語ベクトルで表現</li><li>類似度計算には Jaccard, ユークリッド, コサイン距離を用い, クラシカル, Kruskal, Sammon 法のいずれかで2次元にプロット</li></ul>
対応分析	<ul style="list-style-type: none"><li>出現パターンの似た単語や外部変数を近くに置くよう図示したもの</li><li>単語と単語または外部変数が同時に出現した頻度をクロス集計し, それぞれの相関が最大になるような2変数で数値化し, 2軸上にプロット</li><li>外部変数も同時にプロット可能</li></ul>
共起ネットワーク	<ul style="list-style-type: none"><li>同時に出現した単語間をネットワークで結んで図示したもの</li><li>同時に出現したかといった共起の有無を集計し, ネットワークを作成</li><li>関係の強さ Jaccard 係数で評価, サブグラフは媒介性, クラスタリング精度(エッジ内の密度の高さ)を使って検出</li></ul>
自己組織化マップ(SOM)	<ul style="list-style-type: none"><li>出現パターンの似た単語を近くに集めて図示したもの</li><li>ニューラルネットワークを利用して近い単語を集める方法で, 距離にはユークリッド距離を使い, クラスタリングは Ward法</li></ul>
文書のクラスター	<ul style="list-style-type: none"><li>似た文書同士をクラスタリングしたもの</li><li>各文書は, 文書中に出現する単語の有無でベクトル化した文書ベクトルで表現</li><li>類似度計算には Jaccard, ユークリッド, コサイン距離を使い, いわゆる Ward法, 群平均法, 最遠隣法で階層クラスターを作成</li></ul>

# KH Coder –スナップショット

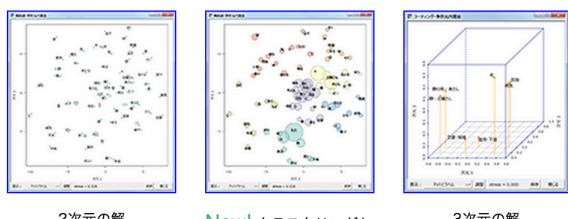
## 階層的クラスター分析

抽出語の階層的クラスター分析を行い、デンドログラムを表示します。抽出語だけでなくコーディング結果（コード）についても、同じように分析を行えます。



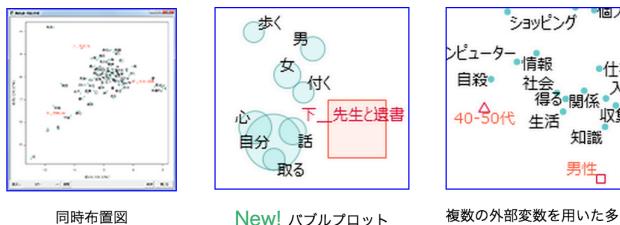
## 多次元尺度構成法 (MDS)

同じく抽出語またはコードを用いての、多次元尺度構成法です。



## 対応分析

同じく抽出語またはコードを用いての、対応分析です。



H29年度 01KA168

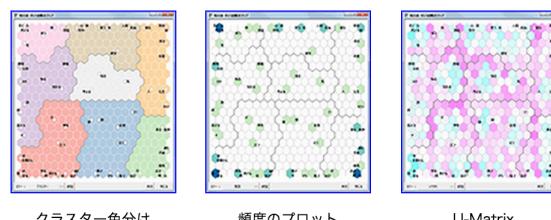
## 共起ネットワーク

抽出語またはコードを用いて、出現パターンの似通ったものを線で結んだ図、すなわち共起関係を線（edge）で表したネットワークを描く機能です。



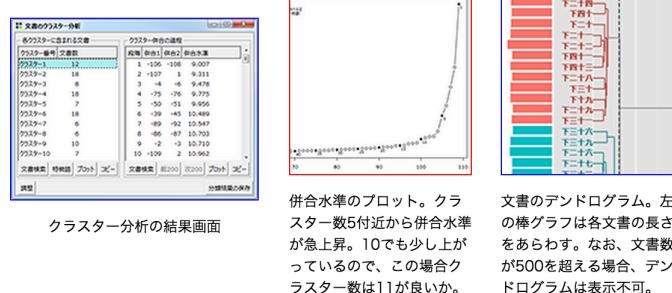
## 自己組織化マップ

抽出語またはコードを用いての、自己組織化マップです。



## 文書のクラスター分析

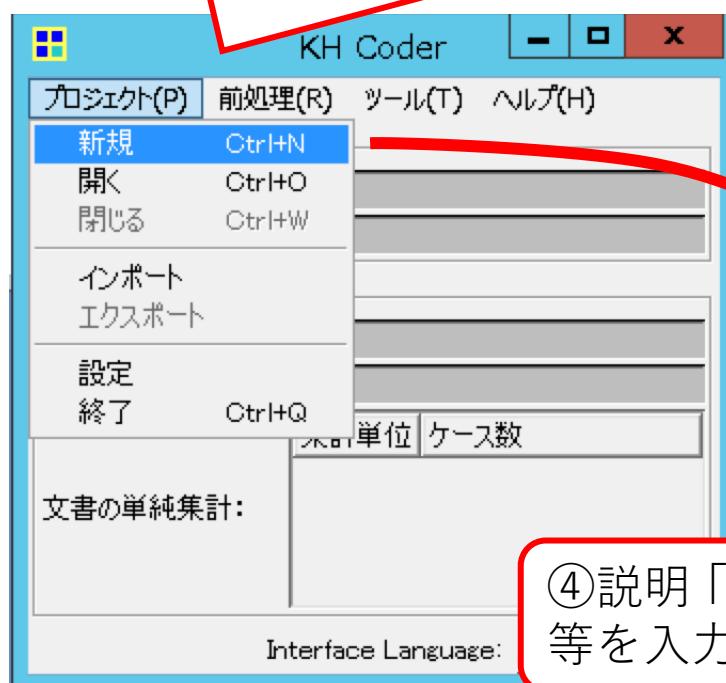
文書の分類を行うクラスター分析です。



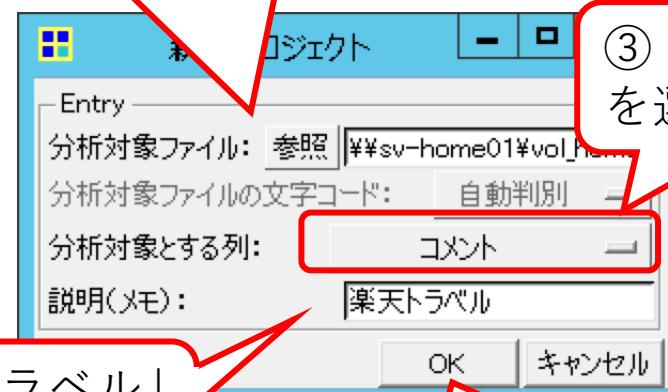
# 練習 — プロジェクトの作成

- ・ファイル rakuten-all.xlsx を開く

①メニューから「プロジェクト」「新規」を選択 (注1)



②「参照」をクリックして  
「rakuten-all.xlsx」を開く



③「コメント」  
を選択 (注2)

④説明「楽天トラベル」  
等を入力

⑤「OK」をクリック

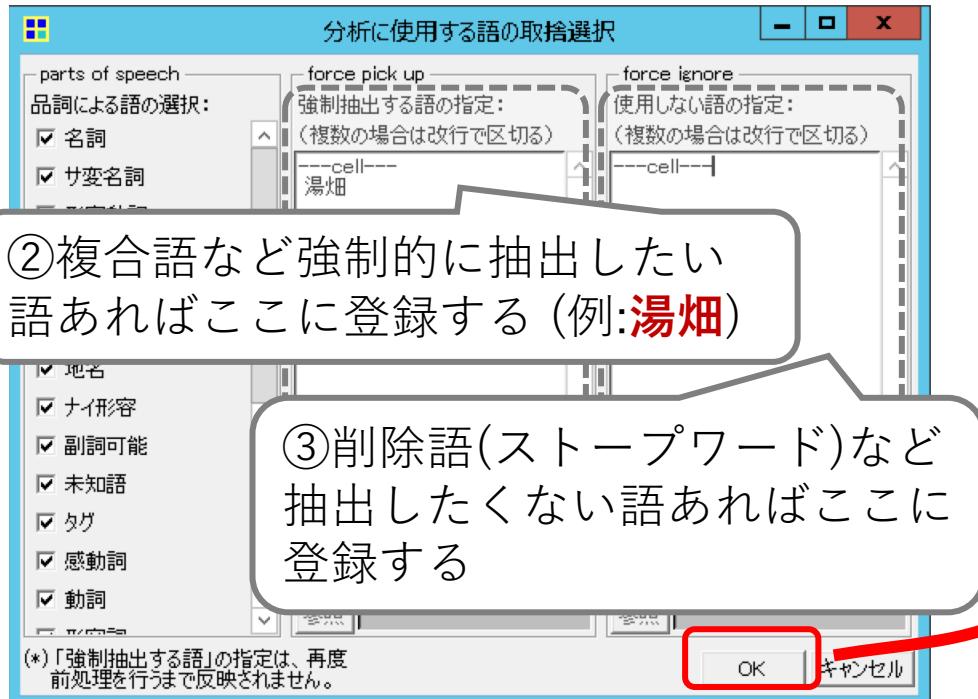
注1: 次回 KH Coderを起動した時は「新規」ではなく「開く」を選択します

注2: ②のファイル選択後、ここに「テキスト」等の選択項目が表示されるまで数分がかかります

# 練習 - 前処理

## ・形態素解析を行う

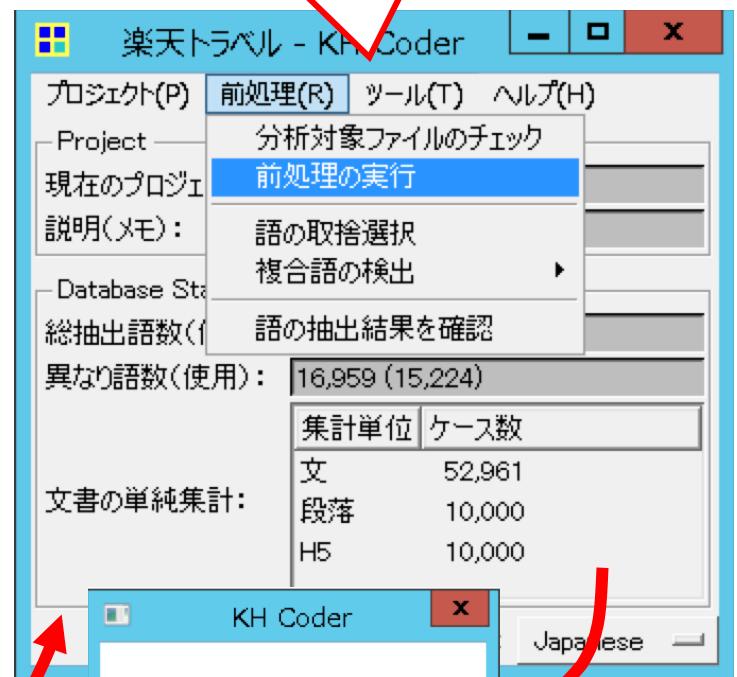
①メニューから「前処理」「語の取捨選択」を選ぶ



②複合語など強制的に抽出したい語あればここに登録する(例:湯畑)

③削除語(ストップワード)など抽出したくない語あればここに登録する

④メニューから「前処理」「前処理の実行」を選ぶ

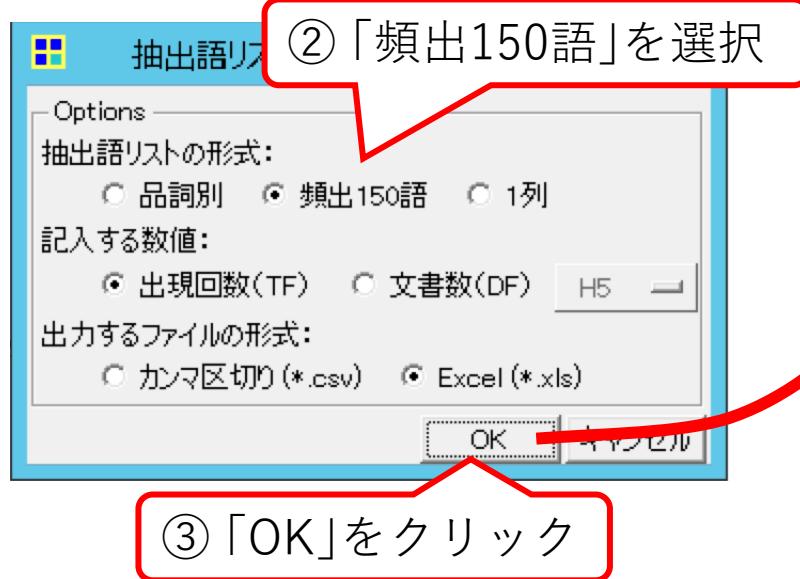


注1: EXCELファイルを読み込んで分析する場合,あらかじめ「---cell---」が入力されています  
注2: メニューから「前処理」「複合語の検出」を選ぶと,複合語候補の一覧を出力できます

# 練習 一頻出語を確認する

- ・頻出語リストを出力する

①メニューから「ツール」「抽出語」「抽出語リスト」を選択

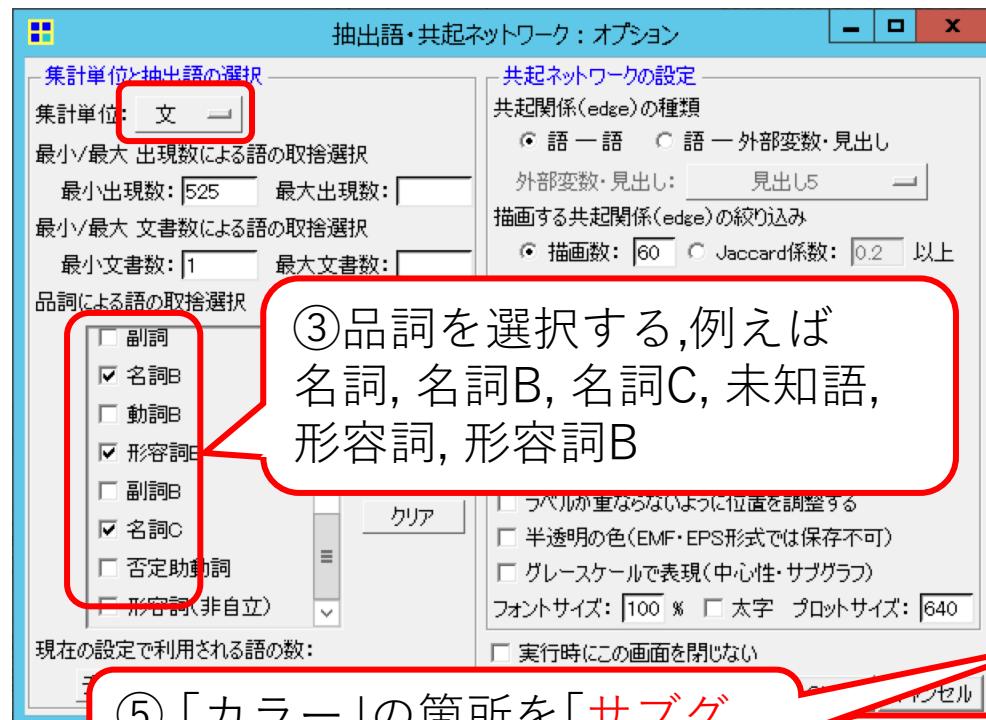


A	B	C	D	E	F	G	H
1 抽出語	出現回数	抽出語	出現回数	抽出語	出現回数	抽出語	出現回数
2 部屋	5090	バイキング	667	旅館	420		
3 思う	4562	チェックイン	661	湯畑	412		
4 良い	4007	バス	650	チェック	407		
5 利用	3883	清潔	619	歩く	404		
6 ホテル	3193	値段	618	アウト	402		
7 宿泊	2859	旅行	615	無料	396		
8 風呂	2817	お世話	570	問題	387		
9 食事	2473	過ごす	568	アメニティ	386		
10 朝食	2248	古い	564	期待	384		
11 満足	2120	場所	561	もう少し	378		
12 温泉	1879	入れる	559	次回	375		
13 美味しい	1699	素晴らしい	558	ビジネス	373		
14 対応	1492	使う	556	お湯	372		
15 行く	1481	子供	549	掃除	370		
16 立地	1398	人	545	客	365		
17 大変	1327	狭い	541	従業	364		
18 お部屋	1313	月	541	接客	363		
19 スタッフ	1294	コンビニ	540	施設	362		
20 広い	1252	プラン	540	価格	361		
21 宿	1195	駐車	536	雰囲気	360		
22 フロント	1184	夜	535	出張	354		
23 残念	1138	過ごせる	525	置く	349		
24 サービス	1115	特に	523	料金	343		
25 便利	1074	来る	519	十分	342		
26 時間	1070	悪い	518	ルーム	339		
27 ホキス	1057	非常	510	吉祥	335		

# 練習 一起ネットワークの作成

①メニューから「ツール」「抽出語」「共起ネットワーク」を選ぶ

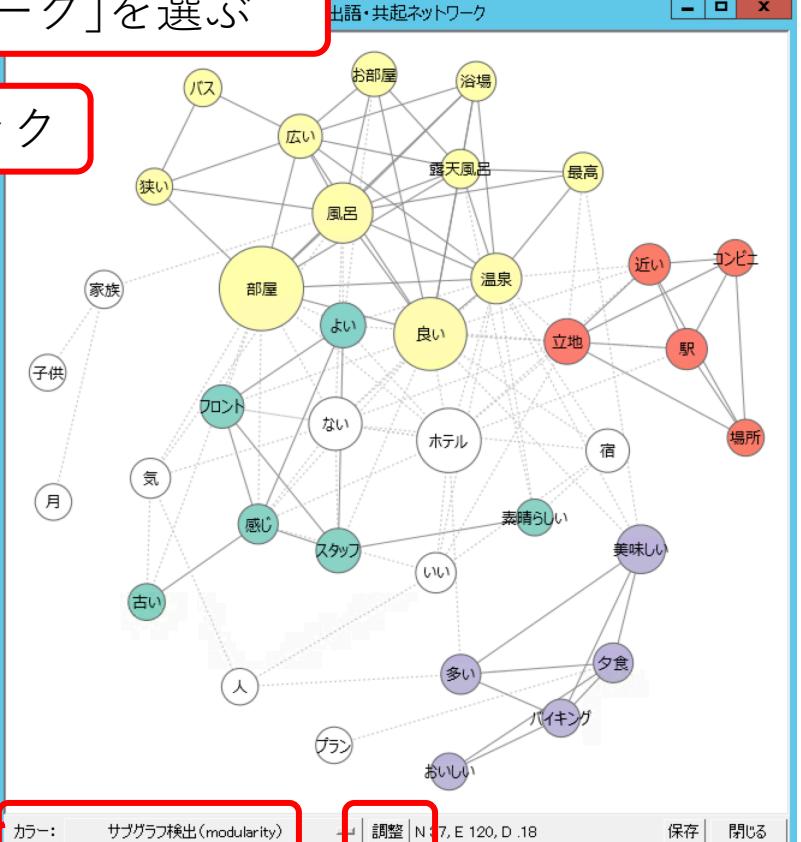
②「集計単位」として「文」を選んで「OK」をクリック



③品詞を選択する,例えば  
名詞, 名詞B, 名詞C, 未知語,  
形容詞, 形容詞B

⑤「カラー」の箇所を「サブグ  
ラフ検出(modularity)」に変更

④「調整」をクリックして「描画数」に120 を  
入力し、「出現数の多い語ほど…」をチェック



# KH Coder の品詞体系

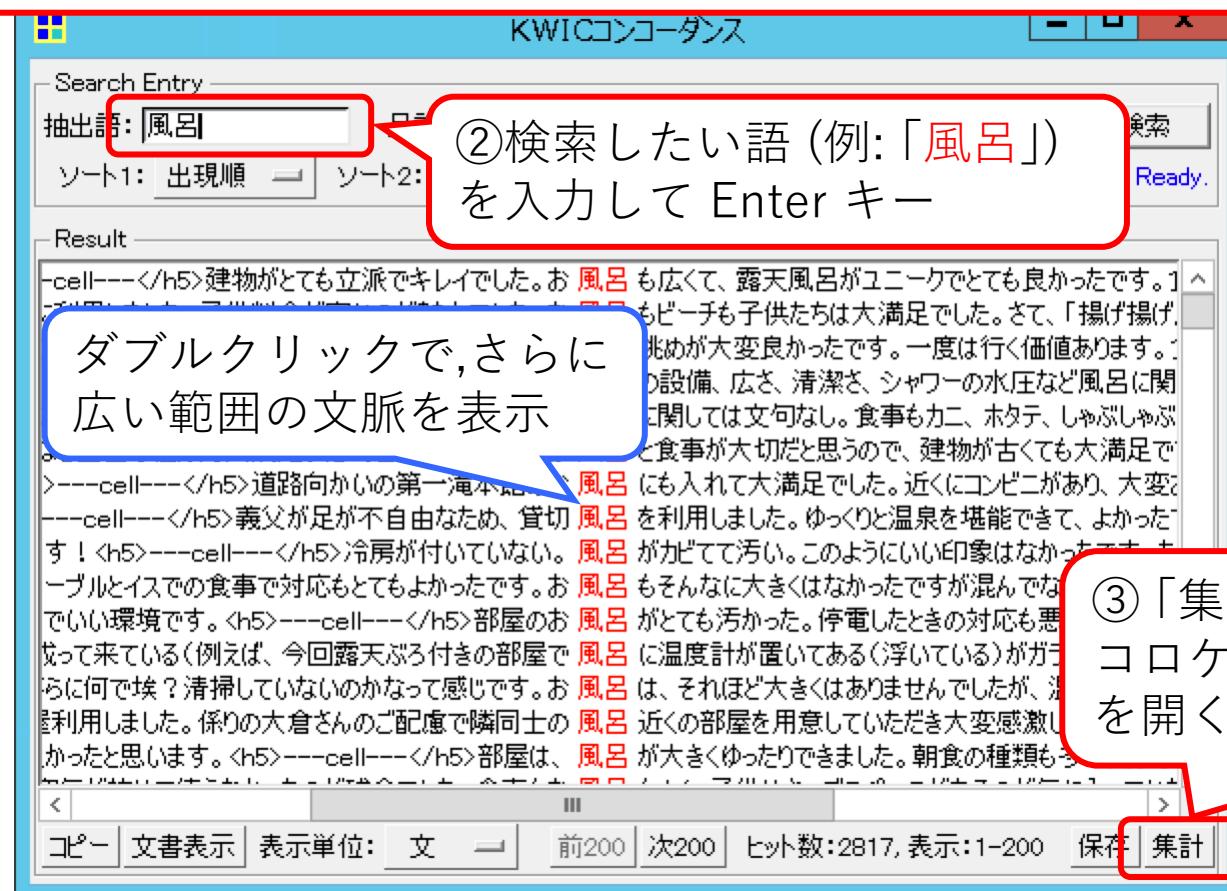
KH Coder 内の品詞名	茶筌の出力における品詞名
名詞	名詞—一般（漢字を含む 2 文字以上の語）
名詞 B	名詞—一般（平仮名のみの語）
名詞 C	名詞—一般（漢字 1 文字の語）
サ変名詞	名詞—サ変接続
形容動詞	名詞—形容動詞語幹
固有名詞	名詞—固有名詞—一般
組織名	名詞—固有名詞—組織
人名	名詞—固有名詞—人名
地名	名詞—固有名詞—地域
ナイ形容	名詞—ナイ形容詞語幹
副詞可能	名詞—副詞可能
未知語	未知語
感動詞	感動詞またはフィラー
タグ	タグ
動詞	動詞—自立（漢字を含む語）
動詞 B	動詞—自立（平仮名のみの語）
形容詞	形容詞（漢字を含む語）
形容詞 B	形容詞（平仮名のみの語）
副詞	副詞（漢字を含む語）
副詞 B	副詞（平仮名のみの語）
否定助動詞	助動詞「ない」「まい」「ぬ」「ん」
形容詞（非自立）	形容詞—非自立（「がたい」「つらい」「にくい」等）
その他	上記以外のもの

「KH Coder 2.x リファレンス・マニュアル」P.11 より

# 練習 - KWICコンコーダンス1

- ・テキスト中でその語がどう使われているか

- ①メニューから「ツール」「抽出語」「KWICコンコーダンス」を選ぶ



# 練習 - KWIC コンコーダンス2

- ①前のページの手順でコロケーション統計を開く

コロケーション統計

Node Word  
抽出語: 風呂 品詞: 活用形:

Result

N	抽出語	品詞	合計	左合計	右合計	左5	左4	左3	左2	左1	右1	右2	右3	右4	右5	ス
1	良い	形容詞	234	77	157	38	17	12	10	0	0	64	43	19	31	78
2	広い	形容詞	152	44	108	14	7	10	10	3	0	74	16	9	9	62
3	狭い	形容詞	62	17	45	5	7	1	4	0	0	26	11	2	6	23
4	よい	形容詞B	60	3	57	1	1	1	1	1	1	13	8	8	19	19
5	ない	形容詞B	67	3	64	5	6	5	5	5	5	6	12	6	23	23
6	大きい	形容詞	35	3	32	0	0	0	0	0	0	4	2	3	15	15
7	気持ちよい	形容詞	38	3	35	0	0	0	0	0	0	8	4	7	12	12
8	熱い	形容詞	29	4	25	1	1	0	1	1	0	7	6	6	0	10
9	小さい	形容詞	23	3	20	1	1	1	1	0	0	10	3	5	2	8
10	素晴らしい	形容詞	25	6	19	1	2	3	0	0	0	10	2	2	5	8
11	ぬるい	形容詞B	17	3	14	0	2	1	0	0	0	5	5	2	2	5
12	遠い	形容詞	10	3	13	1	1	1	0	0	0	6	5	2	0	5
13	いい	形容詞B	19	7	12	1	2	0	0	1	0	5	0	0	1	7

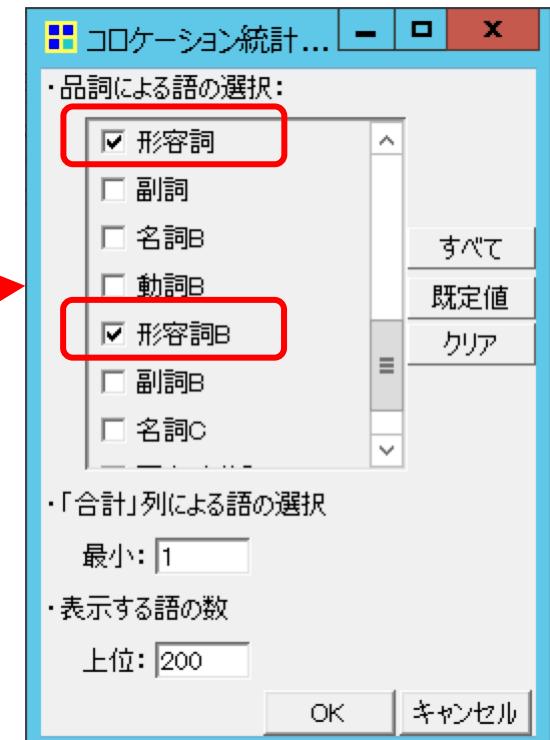
コピー フィルタ設定 ソート: 右合計

「右1」は右側の1つ目(=直後)  
に出現していた回数

「広い」は「風呂」の  
2つ後に 74 回出現

②「右合計」でソート

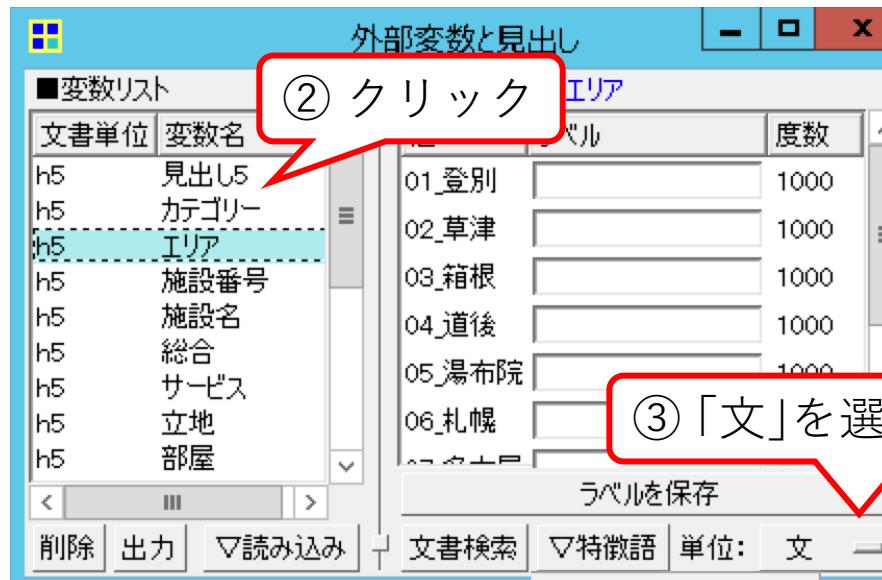
③表示する語を品詞(例: 形容詞,  
形容詞B)とともに選択



# 練習 - KWICコンコーダンス3

- 外部変数(エリア)ごとの特徴語を確認する

- メニューから「ツール」「外部変数と見出し」「リスト」を開く



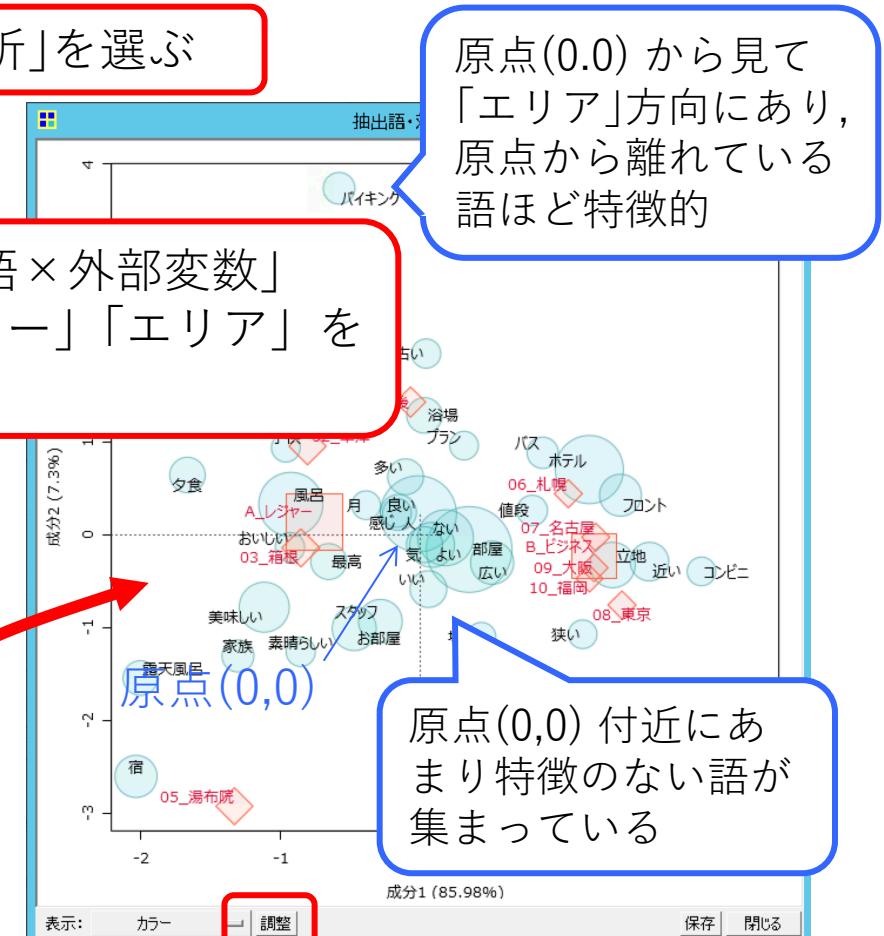
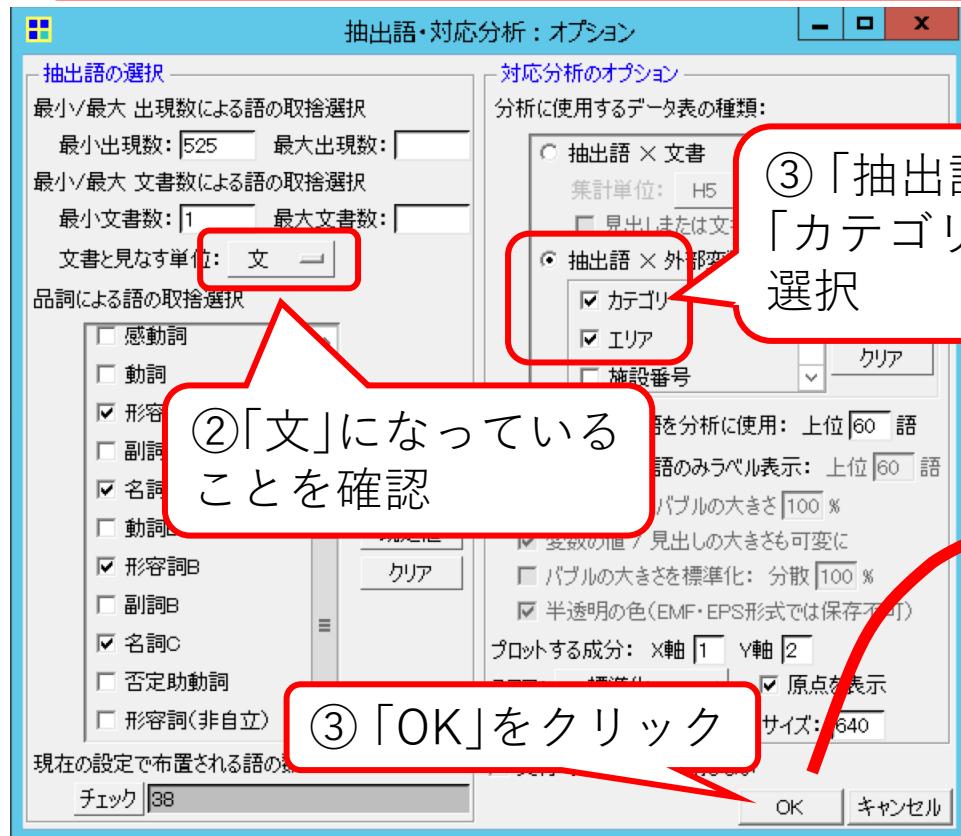
	01 豊かな	02 早起き	03 箱根	04 道後
3	食事 .124	湯畑 .326	食事 .166	道後 .196
4	バイキング .115	草津 .263	風呂 .135	温泉 .128
5	風呂 .106	温泉 .152	美味しい .130	松山 .122
6	宿泊 .096	食事 .150	箱根 .121	朝食 .095
7	温泉 .095	風呂 .149	温泉 .117	本館 .092
8	良い .093	良い .119	露天風呂 .116	ホテル .086
9	部屋 .090	宿 .114	思う .115	対応 .074
10	思う .090	思う .106	料理 .113	立地 .070
11	満足 .088	美味しい .102	夕食 .111	フロント .068
12	残念 .086	満足 .101	良い .108	美味しい .067
13	05 湯布院	06 札幌	07 名古屋	08 東京
14	宿 .180	札幌 .175	名古屋 .200	駅 .113
15	食事 .159	ホテル .097	ホテル .090	便利 .102
16	料理 .159	朝食 .097	利用 .089	ホテル .097
17	美味しい .158	利用 .094	朝食 .088	利用 .090
18	湯布院 .157	フロント .074	駅 .085	立地 .080
19	露天風呂 .142	広い .068	部屋 .084	朝食 .073
20	風呂 .126	便利 .065	立地 .080	フロント .072
21	家族 .108	立地 .063	フロント .071	近く .071
22	スタッフ .107	駅 .060	近く .069	近く .070
23	ありがとう .107	すすきの .059	便利 .064	東京 .066
24	09 大阪	10 福岡		
25	大阪 .133	博多 .139		
26	駅 .094	便利 .090		
27	利用 .090			
28	便利 .089			
29	ホテル .088			
30	立地 .085			
31	近く .080			
32	コンビニ .079			
33	快適 .078			
34	フロント .061	近く .064		

各エリアの特徴語を10件ずつ  
一覧 (数値は Jaccard係数)

注: Jaccard係数は共起尺度のひとつで,共通要素の数を少なくとも一方にある数で割ったもの

# 練習 一対応分析による探索1

①メニューから「ツール」「抽出語」「対応分析」を選ぶ



④「調整」をクリックして「バブルプロット」をチェック

原点(0.0) から見て  
「エリア」方向にあり,  
原点から離れている  
語ほど特徴的

原点(0,0) 付近に  
あまり特徴のない語が  
集まっている

# 練習－コーディングルール1

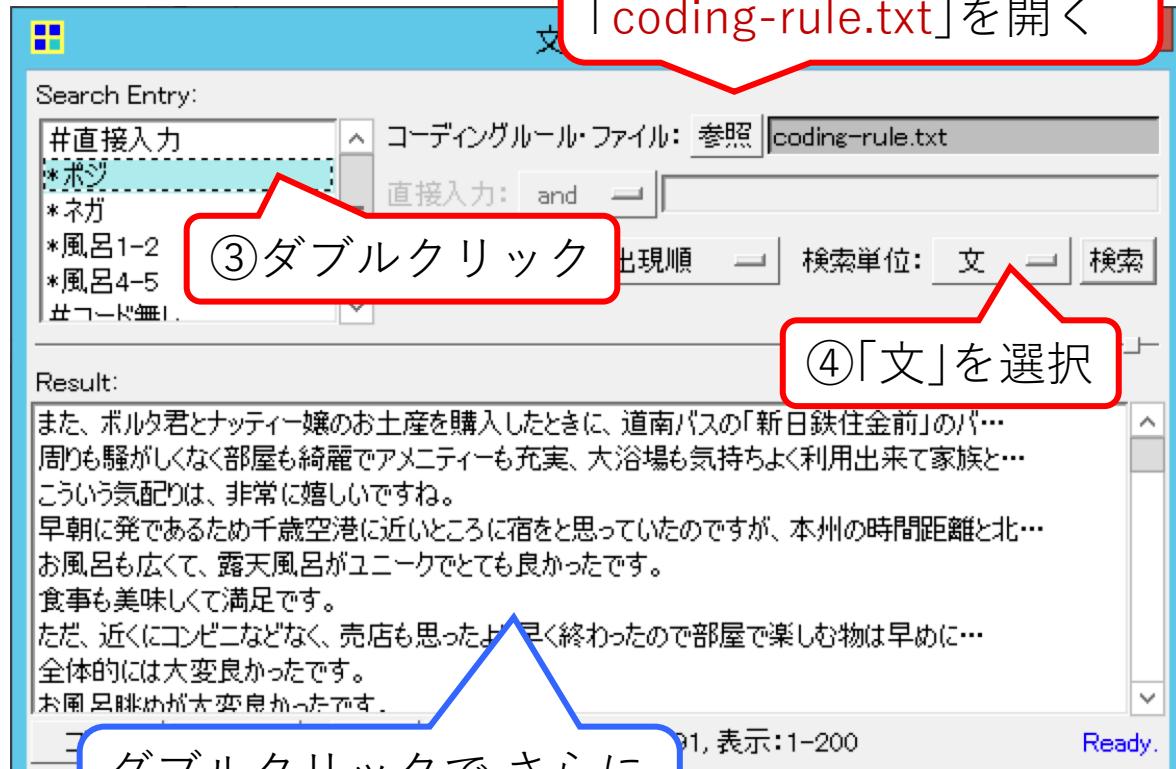
①メニューから「ツール」「文書」「文書検索」を選ぶ

②「参照」をクリックして  
「coding-rule.txt」を開く

③ダブルクリック

④「文」を選択

ダブルクリックで、さらに  
広い範囲の文脈を表示



coding-rule.txt の中身

\*ポジ

良い or 美味しい or 広い or 多い or 素晴らしい or 嬉しい or 気持ちよい or 楽しい or 近い or 大きい or 気持ち良い or 温かい or 早い or 優しい or 新しい or 暖かい or 快い or 明るい or 美しい or 可愛い

\*ネガ

古い or 無い or 高い or 悪い or 小さい or 狹い or 少ない or 寒い or 遅い or 熱い or 欲しい or 暑い or 冷たい or 遠い or 臭い or 暗い

\*風呂1-2

<>風呂-->1 | <>風呂-->2

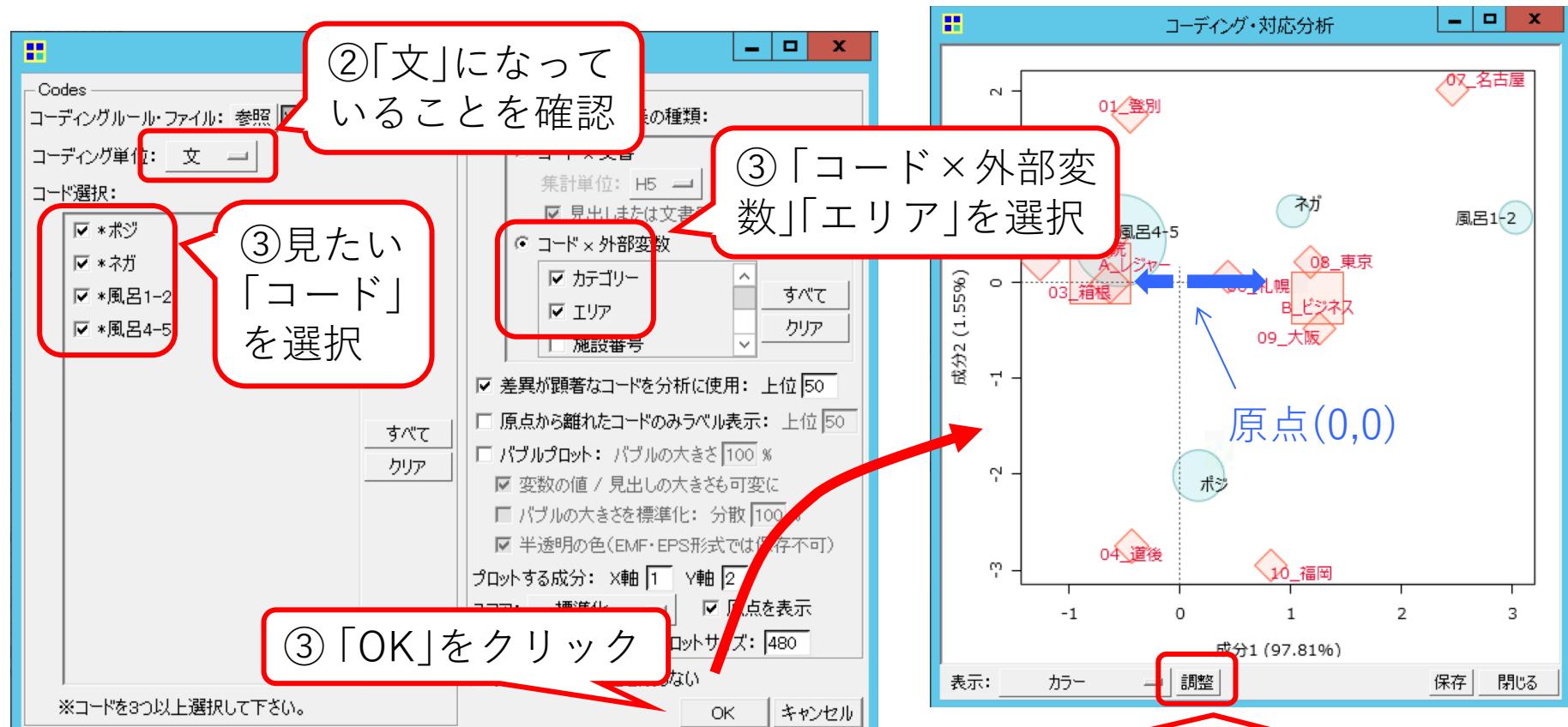
\*風呂4-5

<>風呂-->4 | <>風呂-->5

外部変数

# 練習 一対応分析による探索2

①メニューから「ツール」「コーディング」「対応分析」を選ぶ

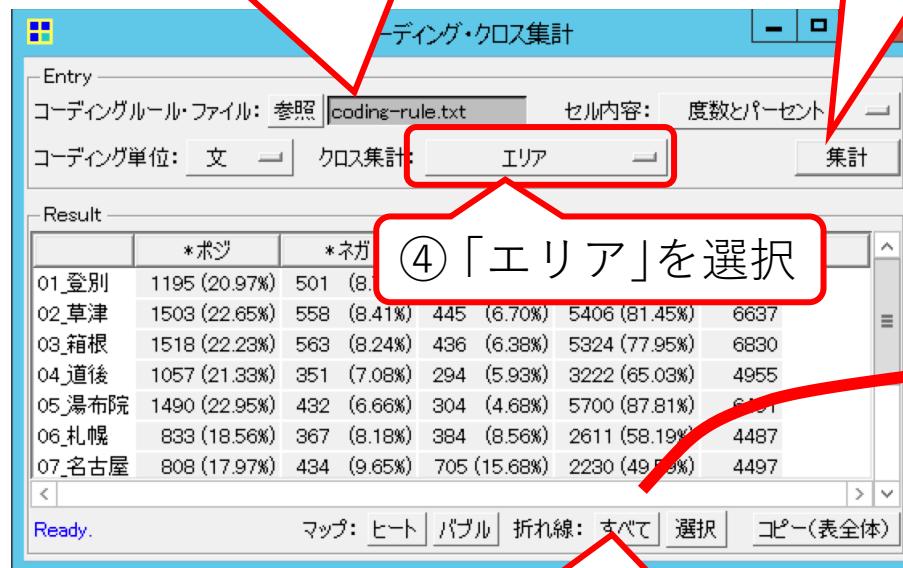


# 練習－クロス集計1

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

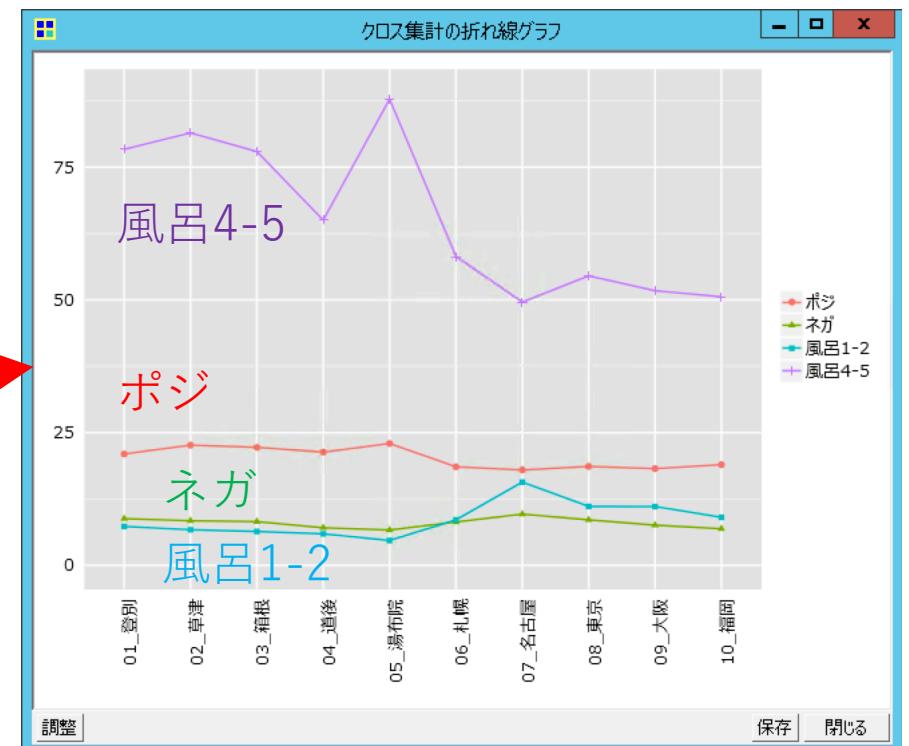
②「参照」をクリックして  
「coding-rule.txt」を開く

⑤「集計」を  
クリック



④「エリア」を選択

⑥「すべて」をクリック



# 練習－コーディングルール2

- 数値評価と口コミの傾向比較
  - 前ページで紹介したクロス集計を用いて,エリアごとの  
**ポジ・ネガ意見の傾向**と,**数値評価の総合点**を比較し,違  
いについて考察する

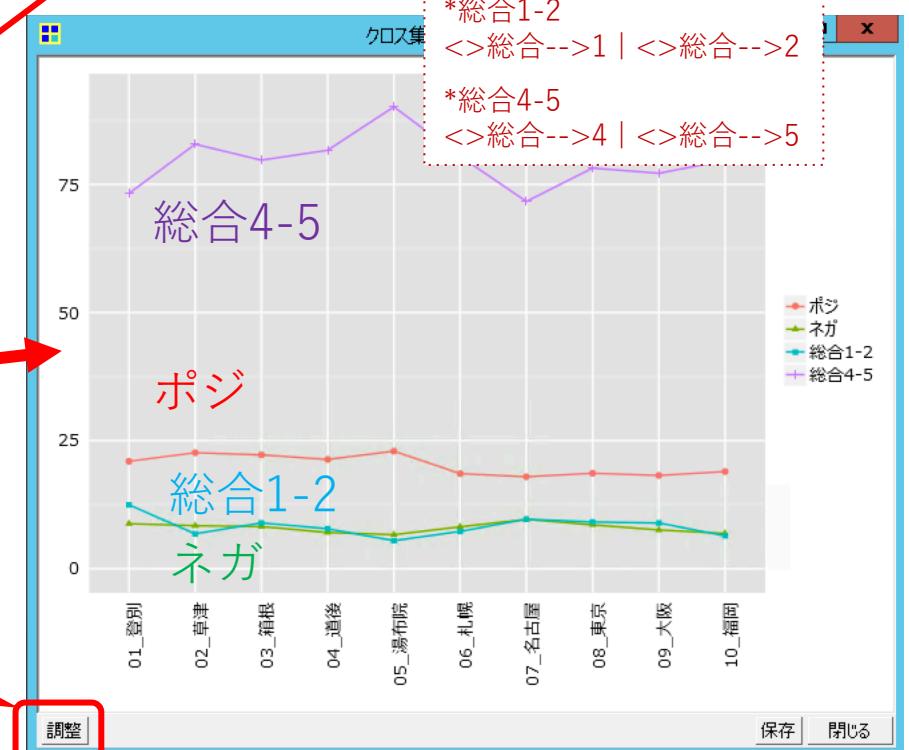
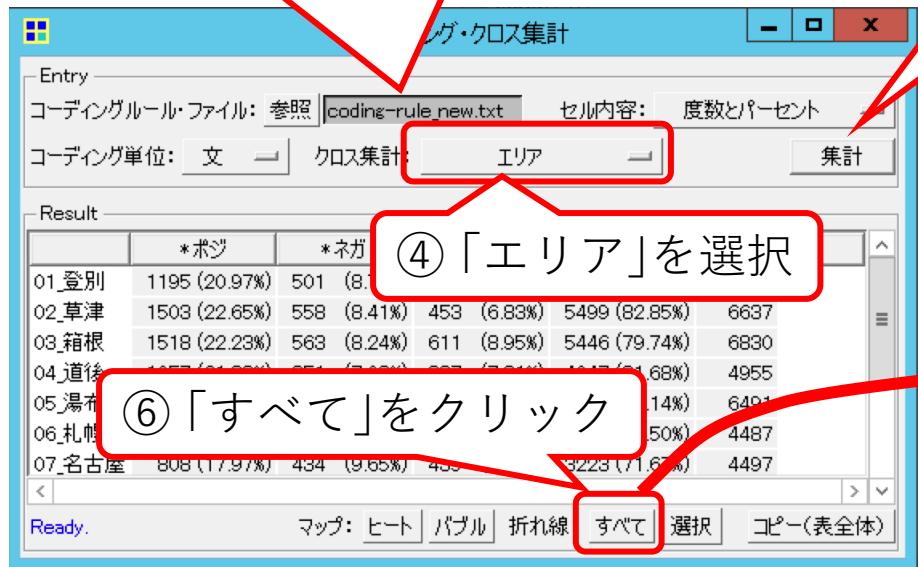
ヒント: 「風呂1-2」「風呂4-5」を参考に「総合1-2」「総合4-5」を追加し,クロス集計する

# 練習－クロス集計2

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

②「参照」をクリックして  
「coding-rule\_new.txt」を開く

⑤「集計」を  
クリック



⑦「ポジ」「ネガ」「総合1-2」「総合4-5」の4つを選択