

テキストマイニングの実習

— 2日目 —

2018/7/12

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

- 1日目: 7/5
 - 説明 – データ分析の手順
 - 演習 – データの理解 (Excel)
- 2日目: 7/12
 - 説明 – テキストマイニングツールの使い方 (KHCoder)
 - 練習 – テキストマイニングツールの使い方 (KHCoder)
- 3日目: 7/26
 - 演習 – データ分析の実践 (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,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
- 論文検索サービスも提供 →
<http://khcoder.net/bib.html?year=2018&auth=all&key=>
- 研究事例リスト**
- KH Coderを用いたご研究の成果を発表された際には、書誌情報をフォームにご記入いただけますと幸いです。
- 出版年 :
- 著者名 :
- キーワード :
- ヒット件数 : 076 / 2042
- KH Coderを用いた研究事例のリスト 2042件**
- ※ 2018/6/16 現在 (961件→1206件→昨年1646件)

KH Coder の情報

ホームページ <http://khcoder.net/>

The screenshot shows the official website for KH Coder. At the top right, there are language options for Japanese and English. Below the header, there's a large image of the book '社会調査のための計量テキスト分析' (Quantitative Text Analysis for Social Research). To the left of the book is a blue banner with the text 'KH Coder'. The main content area includes:

- Index**: A section with a message about an upcoming seminar.
- 概要**: A brief introduction to KH Coder, mentioning it's a text-based (documentary) data analysis software used for quantitative text analysis, survey自由記述, interview records, news articles, etc. It's designed for social research data analysis.
- 主な機能と分析の手順**: A list of features and steps for analysis.
- KH Coderを用いた研究事例のリスト**: A link to a list of research examples (2042件).
- 機能紹介 (スクリーンショット)**: Screenshots of the software interface.
- KH Coderの入手**: Download links for KH Coder 3 (最新アルファ版), KH Coder 2 (古い安定版), 必要なソフトウェア / ハードウェア, バージョンアップ履歴, and 使用許諾.

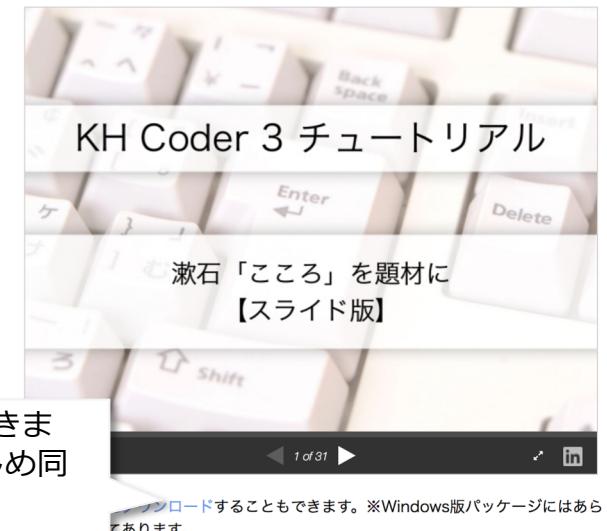
参考書



PDFファイルをダウンロードすることもできます。※Windows版パッケージにはあらかじめ同梱してあります。

チュートリアル
http://khcoder.net/kh_tuto.html

チュートリアル & ヒント



チュートリアル用データ

チュートリアルの実行に必要なデータファイルです。
※Windows版パッケージには同梱してありますので、別途ダウンロードする必要はありません。

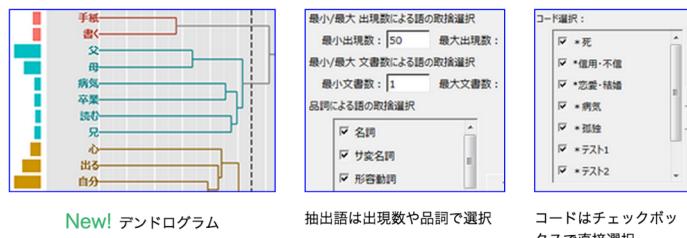
KH Coder の主な分析手法

分析手法	解説
階層的クラスター分析	<ul style="list-style-type: none">・出現パターンの似た単語をクラスタリングしたもの・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・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
多次元尺度構成法(MDS)	<ul style="list-style-type: none">・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・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・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
対応分析	<ul style="list-style-type: none">・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・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・外部変数も同時にプロット可能
共起ネットワーク	<ul style="list-style-type: none">・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・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・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
自己組織化マップ	<ul style="list-style-type: none">・出現パターンの似た単語を近くに集めて図示したもの・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
文書のクラスター分析	<ul style="list-style-type: none">・似た文書同士をクラスタリングしたもの・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・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を作成

KH Coder –スクリーンショット

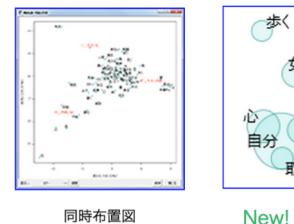
階層的クラスター分析

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。



対応分析

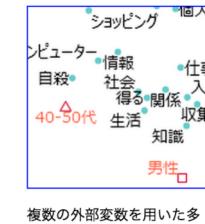
同じく抽出語またはコードを用いての、対応分析です。



同時布置



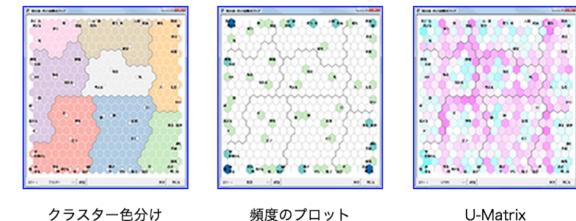
New! バブルプロット



複数の外部変数を用いた 重対応分析

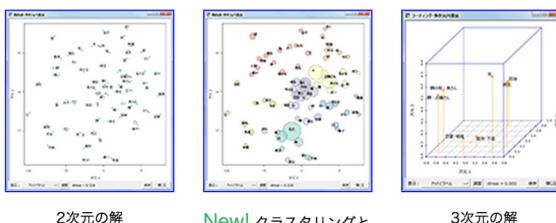
自己組織化マップ

抽出語またはコードを用いての、自己組織化マップです。



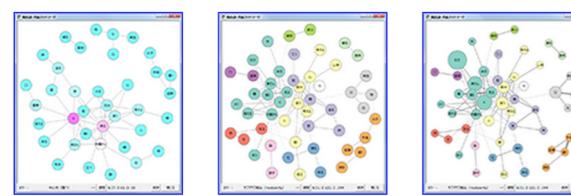
多次元尺度構成法 (MDS)

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。



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、すなはち共起関係を線（edge）で表したネットワークを描く機能です。



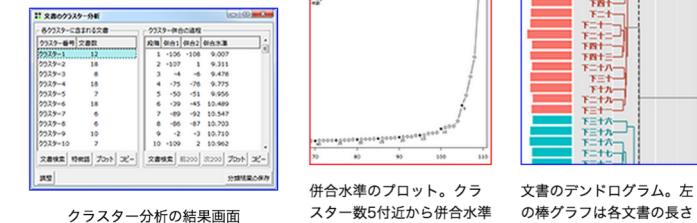
共起の程度が非常に強い
ものだけを線で結んだ図

やや弱い共起関係を
に含め、自動的にセ
ブ分け（色分け）

出現数が多い語ほど大きく、また共起の程度が強いほど太い線で描画

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。



併合水準のプロット。クラスター数5付近から併合水準が急上昇。10でも少し上がっているので、この場合クラスター数は11が良いか。

単語の出現パターン

- ある文に各単語が含まれて「いる=1」「いない=0」かを考える
→全文中に含まれるすべての単語を要素(次元)とするベクトル

文ID	2	3	4	6	7	8	9	10	11	13	14	15	16	18	19	20	21	23	24	25	26	27	28	30
部屋	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0
ホテル	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
風呂	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
温泉	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
お部屋	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
立地	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
スタッフ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
フロント	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
最高	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
感じ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

KH Coder で使われる距離尺度

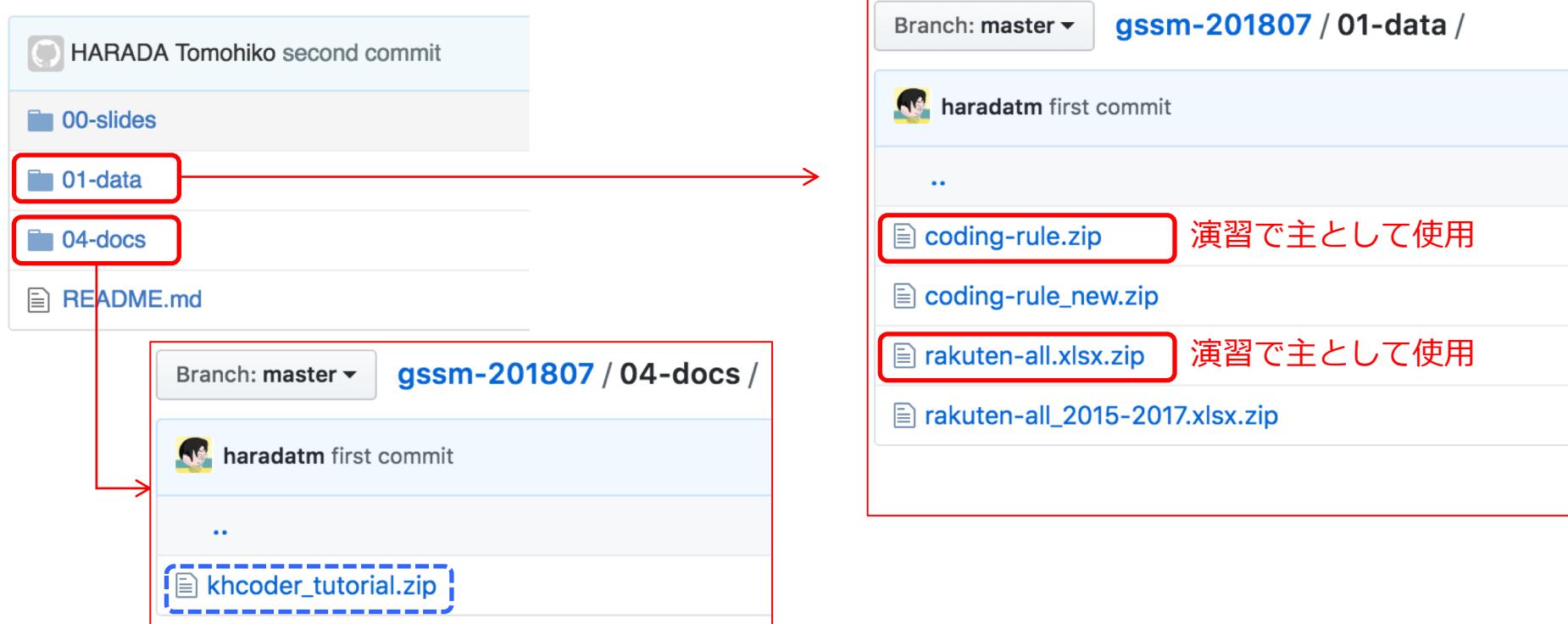
- KH Coder では Jaccard 距離を多用
 - 語Aと語Bのどちらも出現していない文書(0-0対)が沢山あっても語Aと語Bが類似しているとは見なさない → スパースなデータ分析向き

Jaccard 距離	コサイン距離	ユークリッド距離									
<ul style="list-style-type: none">• 1つ文書に含まれる語が少なく、各語が一部の文書中にしか含まれていないスパースデータ向き• 1つの文書の中に語が1回出現した場合も10回出現した場合も単に「出現あり」と見なしてカウントした語と語の共起数を計算	<ul style="list-style-type: none">• 1つひとつの文書が長く、多数の文書に含まれている語が多いデータ向き(各文書中での語の出現回数の大小が重要な場合)• 文書中における語の出現回数(1,000語あたりの出現回数に調整)を計算	<ul style="list-style-type: none">• 増減傾向が似ているかどうかだけを見る場合向き• サイズの差までも見る場合向き									
<table border="1"><tr><td></td><td>1</td><td>0</td></tr><tr><td>1</td><td>n_{11}</td><td>n_{10}</td></tr><tr><td>0</td><td>n_{01}</td><td>n_{00}</td></tr></table> $J\text{ }S = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$		1	0	1	n_{11}	n_{10}	0	n_{01}	n_{00}	$\text{cos } S(\mathbf{x}, \mathbf{y}) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$	$E d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum (x_i - y_i)^2}$
	1	0									
1	n_{11}	n_{10}									
0	n_{01}	n_{00}									

<http://mjin.doshisha.ac.jp/R/68/68.html>

使用するデータ

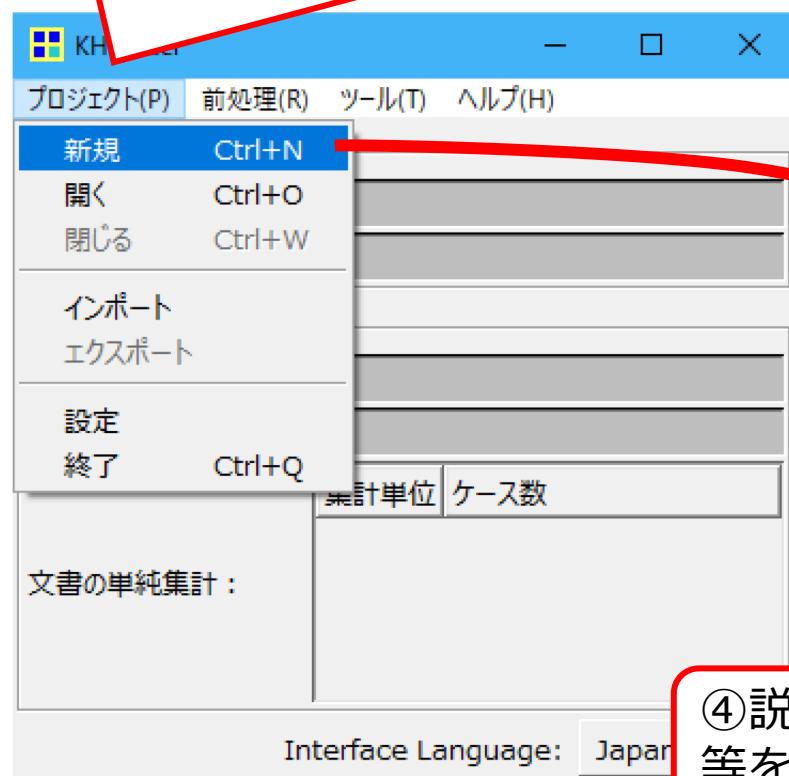
- <https://github.com/haradatm/gssm-201807>



KH Coder のチュートリアル (Windows版に同梱)

操作説明 —プロジェクトの作成

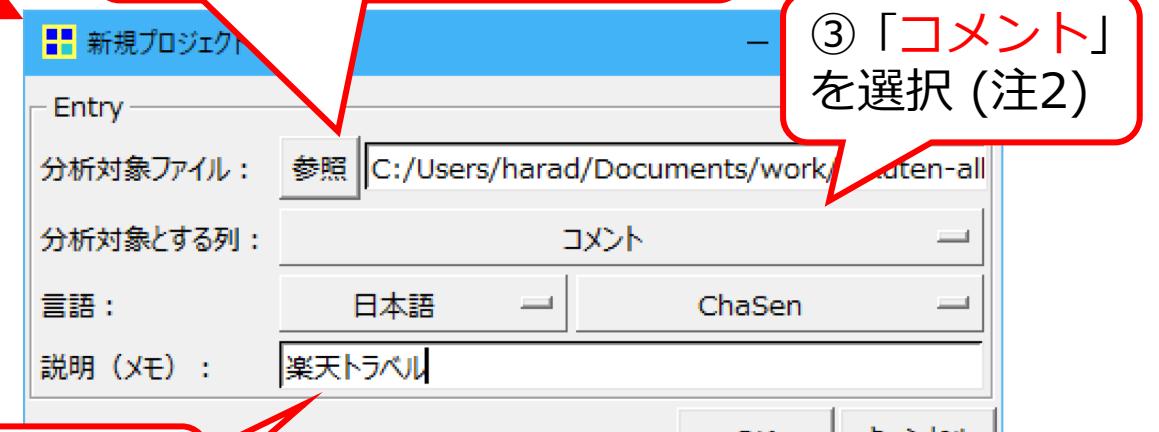
①メニューから「プロジェクト」「新規」を選択 (注1)



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

②「参照」をクリックして
「rakuten-all.xlsx」を開く

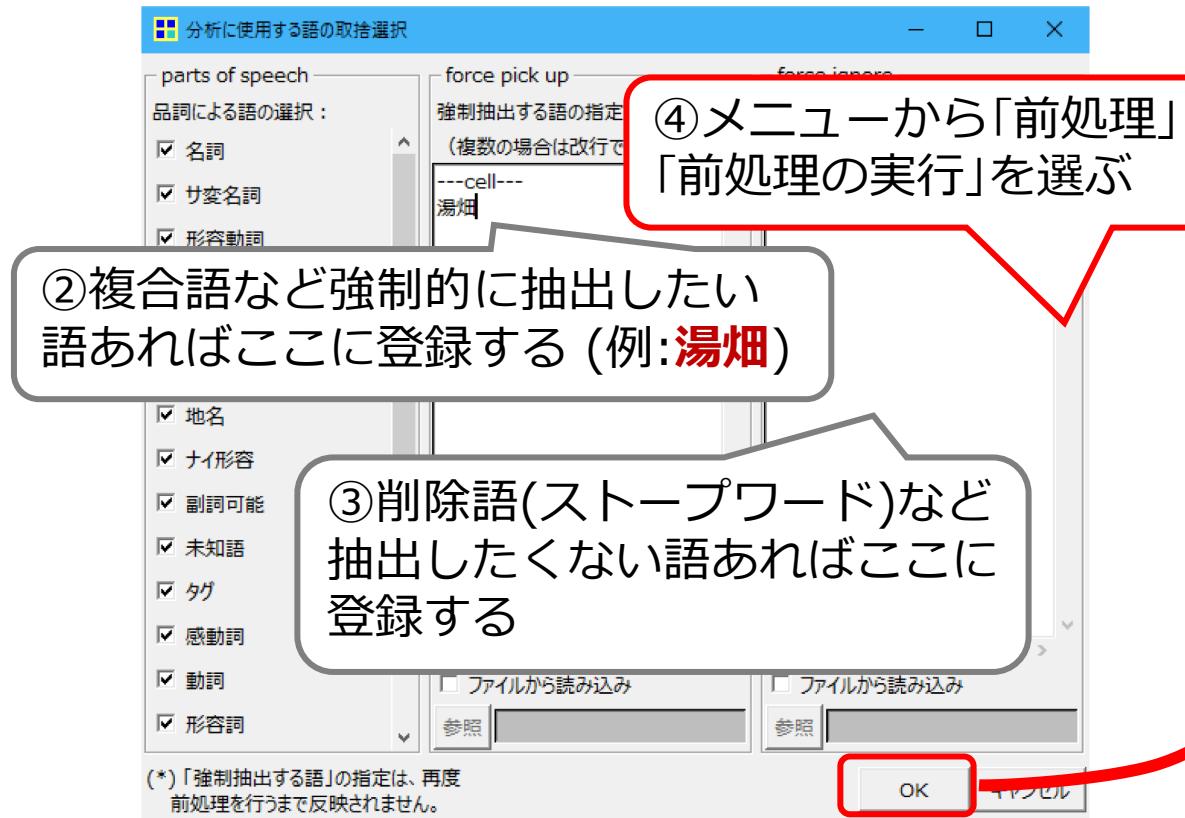


④説明「楽天トラベル」
等を入力

⑤「OK」をクリック

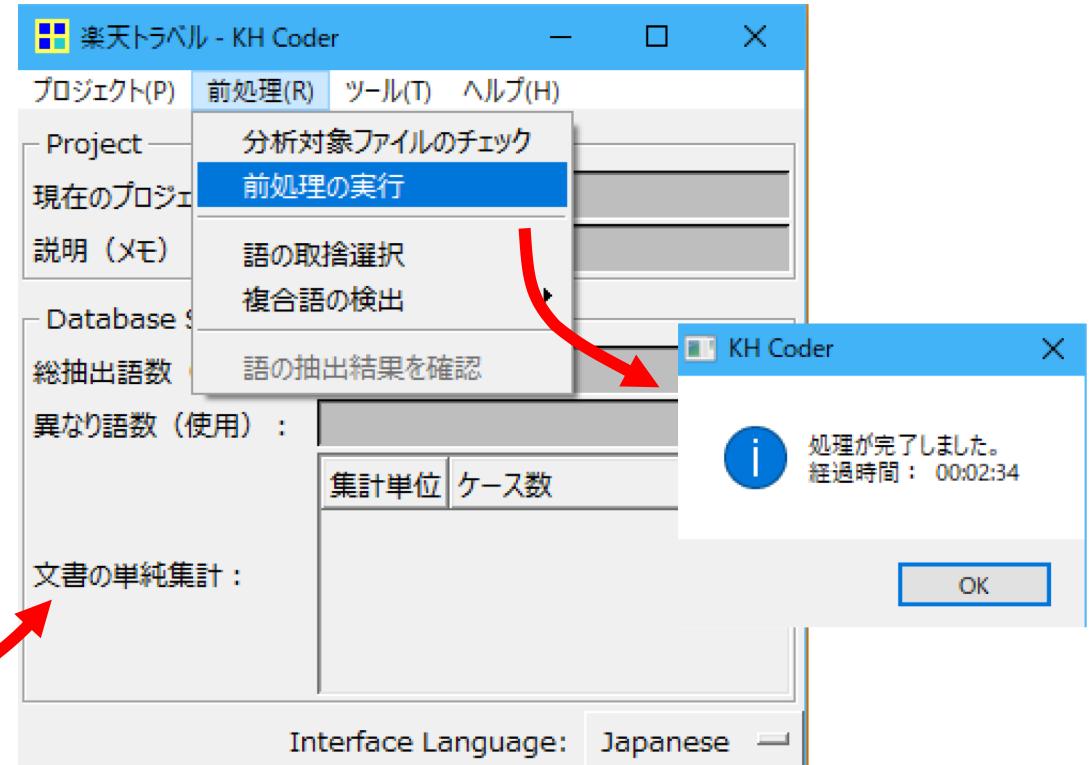
操作説明 – 前処理 (形態素解析)

①メニューから「前処理」「語の取捨選択」を選ぶ



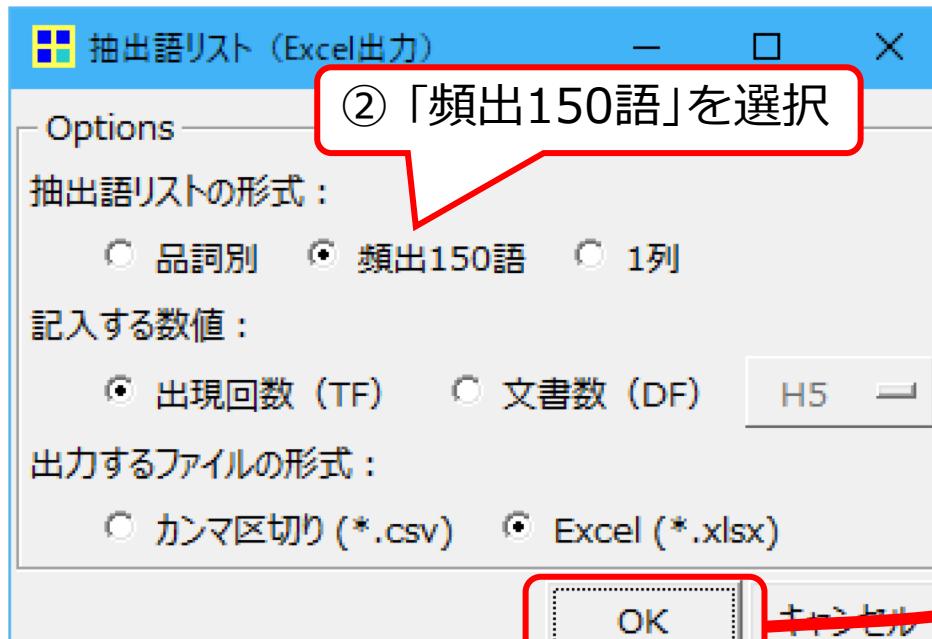
注1: EXCELファイルを読み込んで分析する場合、あらかじめ「---cell---」が入力されています

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**



操作説明 – 頻出語を確認する

- ①メニューから「ツール」「抽出語」「抽出語リスト」を選択



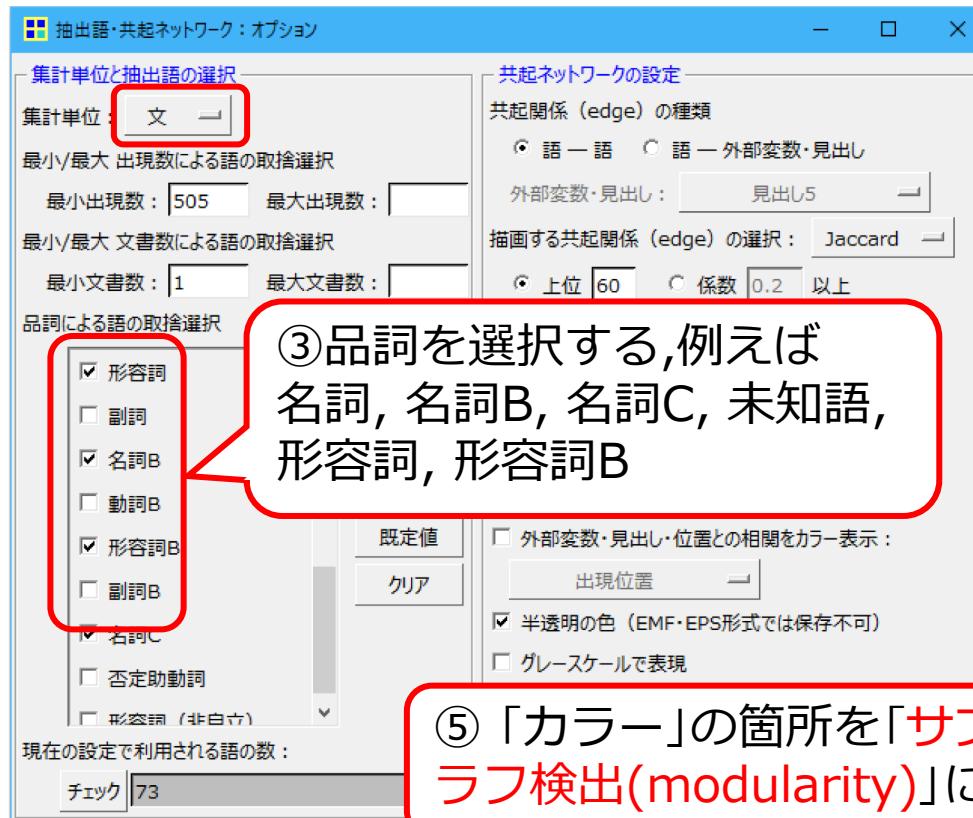
- ③「OK」をクリック

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	4879	バス	657	問題	422		
3 思う	4125	月	647	歩く	414		
4 良い	3884	バイキング	614	無料	401		
5 利用	3662	清潔	612	気持ちよい	400		
6 ホテル	2893	お世話	597	料金	389		
7 宿泊	2817	初めて	595	設備	387		
8 風呂	2659	家族	588	接客	385		
9 食事	2441	狭い	575	静か	378		
10 朝食	2149	旅行	553	お願い	375		
11 満足	2044	コンビニ	551	置く	375		
12 温泉	1753	入れる	548	女性	373		
13 美味しい	1647	人	543	湯畑	372		
14 行く	1386	過ごす	533	従業	371		
15 お部屋	1377	子供	531	清掃	361		
16 対応	1358	場所	527	次回	359		
17 立地	1302	古い	526	ビジネス	354		
18 広い	1279	値段	524	お湯	353		
19 スタッフ	1216	特に	523	掃除	352		

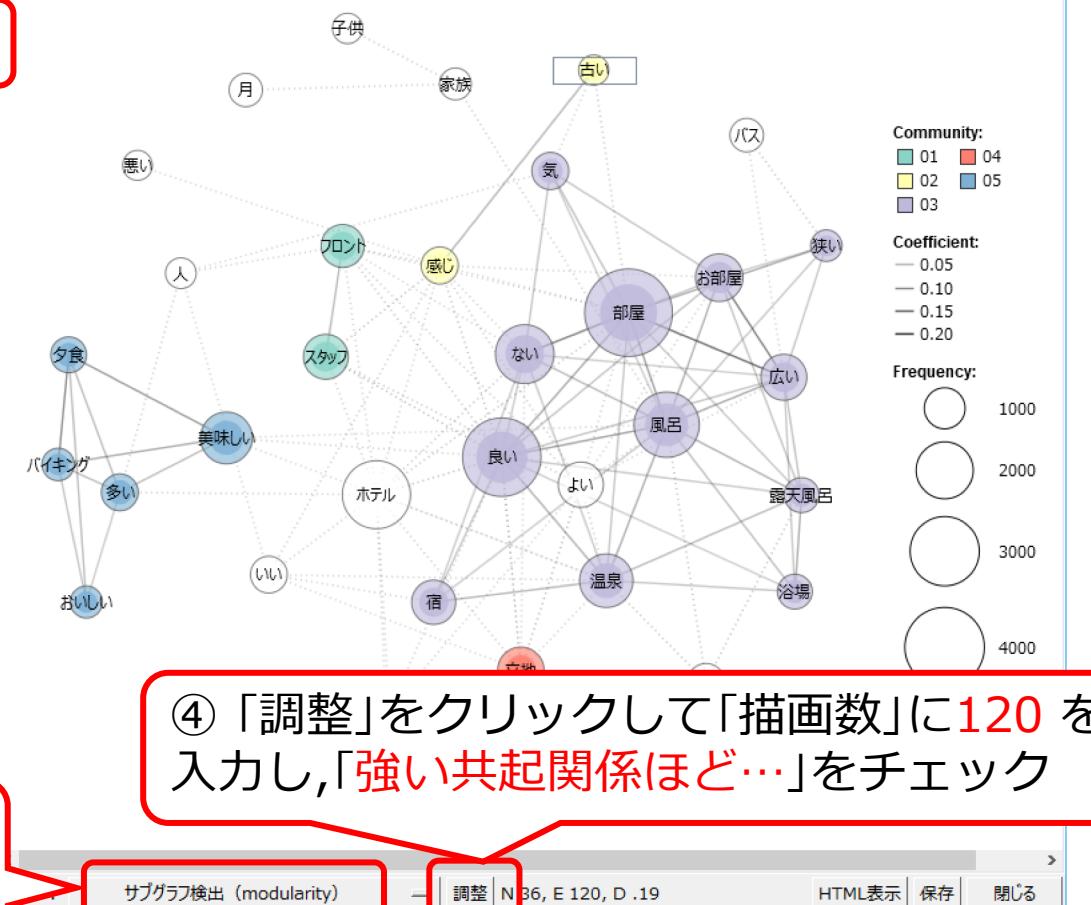
操作説明 —共起ネットワークの作成

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



KH Coder の品詞体系

表 A.1 KH Coder の品詞体系

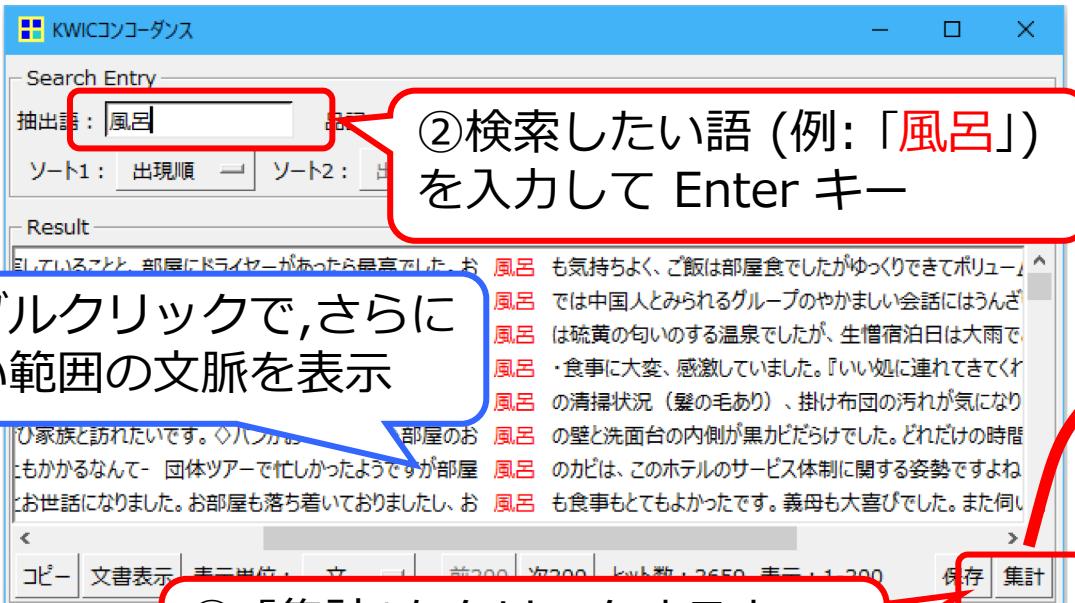
KH Coder 内の品詞名	茶筌の出力における品詞名
名詞	名詞一般（漢字を含む 2 文字以上の語）
名詞 B	名詞一般（平仮名のみの語）
名詞 C	名詞一般（漢字 1 文字の語）
サ変名詞	名詞-サ変接続
形容動詞	名詞-形容動詞語幹
固有名詞	名詞-固有名詞一般
組織名	名詞-固有名詞-組織
人名	名詞-固有名詞-人名
地名	名詞-固有名詞-地域
ナイ形容	名詞-ナイ形容詞語幹
副詞可能	名詞-副詞可能
未知語	未知語
感動詞	感動詞またはフィラー
タグ	タグ
動詞	動詞-自立（漢字を含む語）
動詞 B	動詞-自立（平仮名のみの語）
形容詞	形容詞（漢字を含む語）
形容詞 B	形容詞（平仮名のみの語）
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 3 リファレンス・マニュアル」
P.11 より

注：どの品詞を選択すべきかは、分析対象のデータ
や分析目的により異なります。分析結果を確認
しながら、適宜、適切な品詞選択を検討すること
が重要です

操作説明 – 語句の前後文脈を表示する

- ①メニューから「ツール」「抽出語」「KWICコンコーダンス」を選ぶ



The screenshot shows the Collocation Statistics application window. The search entry field also contains '風呂'. A red box highlights the '右合計' (Right Total) button at the bottom of the window. To its right, another red box highlights the 'ソート' (Sort) button. A blue box highlights the '右1' column header in the table with the text '「右1」は右側の1つ目(=直後)に出現していた回数' (The count of occurrences immediately after the target word). The table lists various words along with their part of speech and co-occurrence counts. A blue box highlights the row for '広い' with the text '「広い」は「風呂」の2語後に74回出現' (Appears 74 times two words after '風呂').

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	スコア
1	良い	形容詞	222	73	149	35	17	10	11	0	6	42	32	26	43	72.8
2	広い	形容詞	141	37	104	11	9	10	7	0	1	74	16	8	5	57.6
3	よい	形容詞B	64	22	42	17	2	2	1	0	11	9	11	10	19.3	
4	狭い	形容詞	52								10	2	2	20.7		
5	ない	形容詞B	5								8	8	7	18.0		
6	気持ちよい	形容詞	3								9	7	2	12.6		
7	大きい	形容詞	3								4	3	4	16.8		

- ③「集計」をクリックすると
コロケーション統計(右)を開く

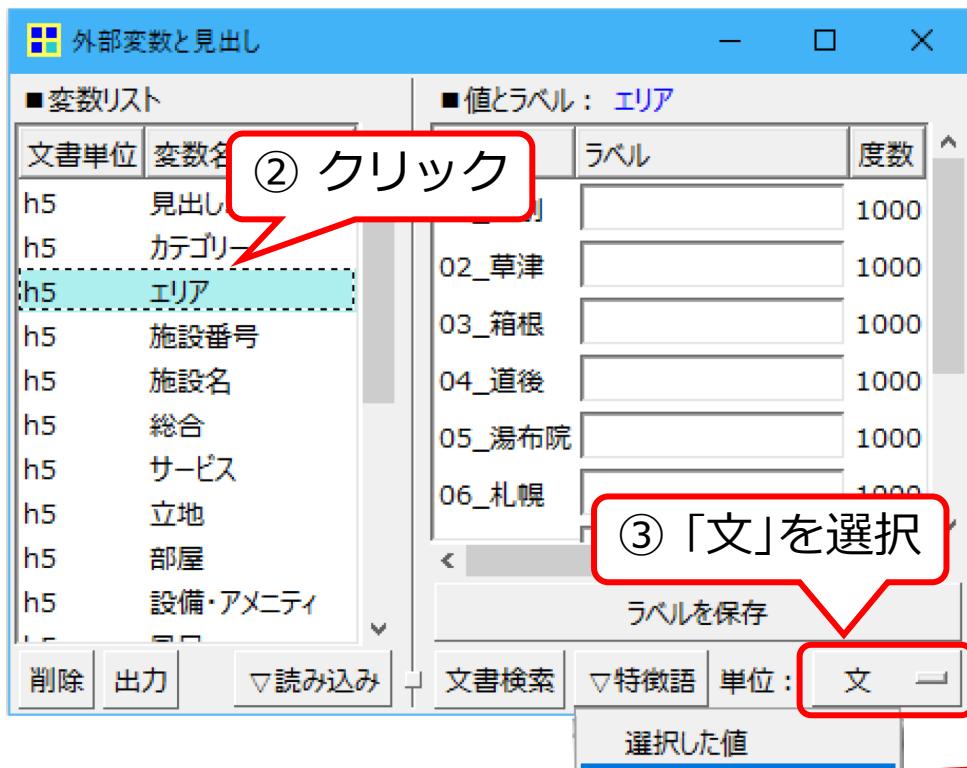
- ④表示する語の品詞を選択
(例: 形容詞, 形容詞B)

- ⑤「右合計」でソート

注: 共起ネットワーク上で「風呂」をクリックすると①②と同じ操作となります(V3以降)

操作説明 – 外部変数(エリア)を利用する

- ①メニューから「ツール」「外部変数と見出し」「リスト」を開く



- ④「特徴語」「一覧(Excel形式)」を選択

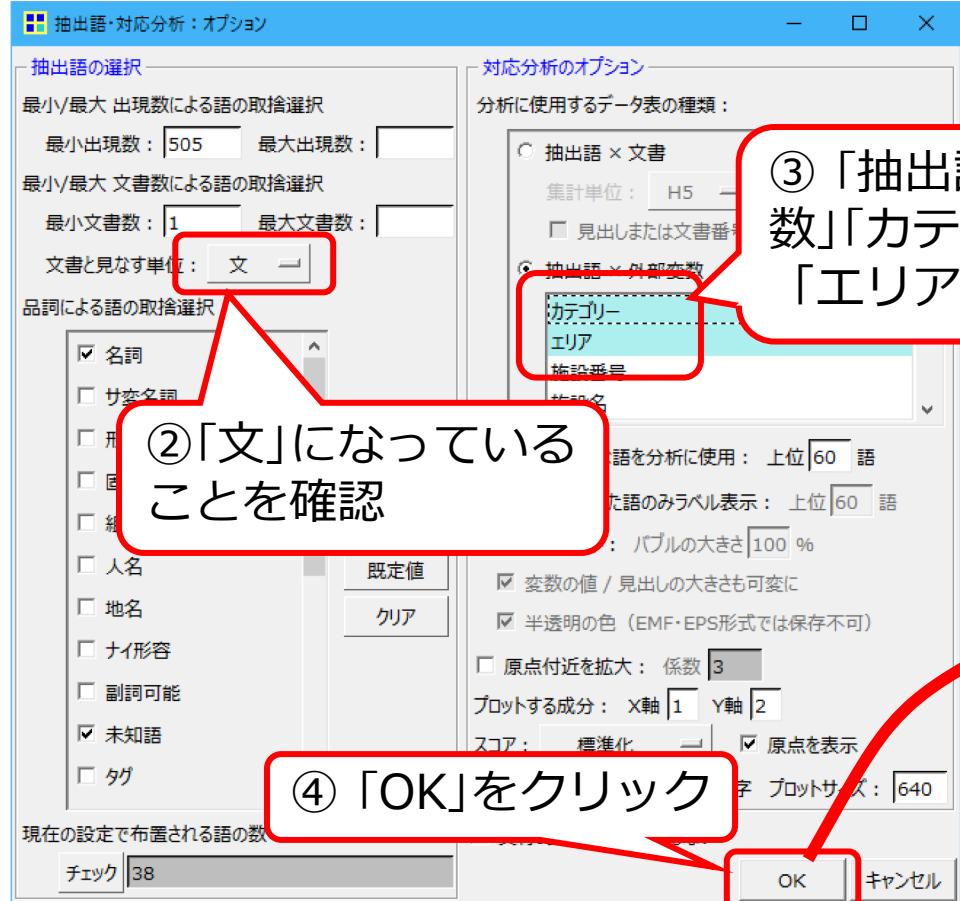
注: Jaccard係数は共起尺度のひとつで、
共通要素の数を少なくとも一方にある数で割ったもの

	A	B	C	D	E	F	G	H	I	J	K
1											
2	01_登別		02_草津			03_箱根			04_道後		
3	食事	.058	温泉		.067	食事		.070	温泉		.059
4	部屋	.057	湯畑		.065	良い		.059	部屋		.053
5	風呂	.055	風呂		.063	風呂		.054	道後		.052
6	思う	.053	草津		.061	美味しい		.049	良い		.050
7	宿泊	.052	思う		.061	満足		.041	利用		.050
8	温泉	.042	食事		.061	温泉		.037	ホテル		.047
9	満足	.039	良い		.060	料理		.035	朝食		.043
10	美味しい	.035	宿		.043	お部屋		.033	宿泊		.039
11	バギング	.033	満足		.042	サービス		.033	満足		.035
12	行く	.030	美味しい		.040	露天風呂		.031	立地		.033
13	05_湯布院		06_札幌			07_名古屋			08_東京		
14	宿	.071	朝食		.064	名古屋		.065	利用		.069
15	食事	.066	札幌		.055	利用		.059	ホテル		.056
16	思う	.060	利用		.055	部屋		.056	部屋		.054
17	風呂	.059	部屋		.053	朝食		.052	駅		.044
18	美味しい	.052	ホテル		.053	ホテル		.051	便利		.042
19	宿泊	.048	良い		.050	思う		.047	宿泊		.039
20	料理	.048	立地		.039	便利		.034	近い		.036
21	満足	.048	広い		.032	フロント		.033	朝食		.035
22	温泉	.043	便利		.032	駅		.032	立地		.030
23	お部屋	.041	フロント		.032	立地		.031	快適		.029
24	09_大阪		10_福岡								
25	利用	.059	利用		.061						
26	部屋	.059	ホテル		.058						
27	ホテル	.058	部屋		.050						
28	思う	.049	博多		.048						
29	大阪	.041	朝食		.042						
30	宿泊	.039	宿泊								
31	朝食	.037	駅								
32	立地	.035	立地								
33	駅	.035	便利								
34	便利	.032	近い								

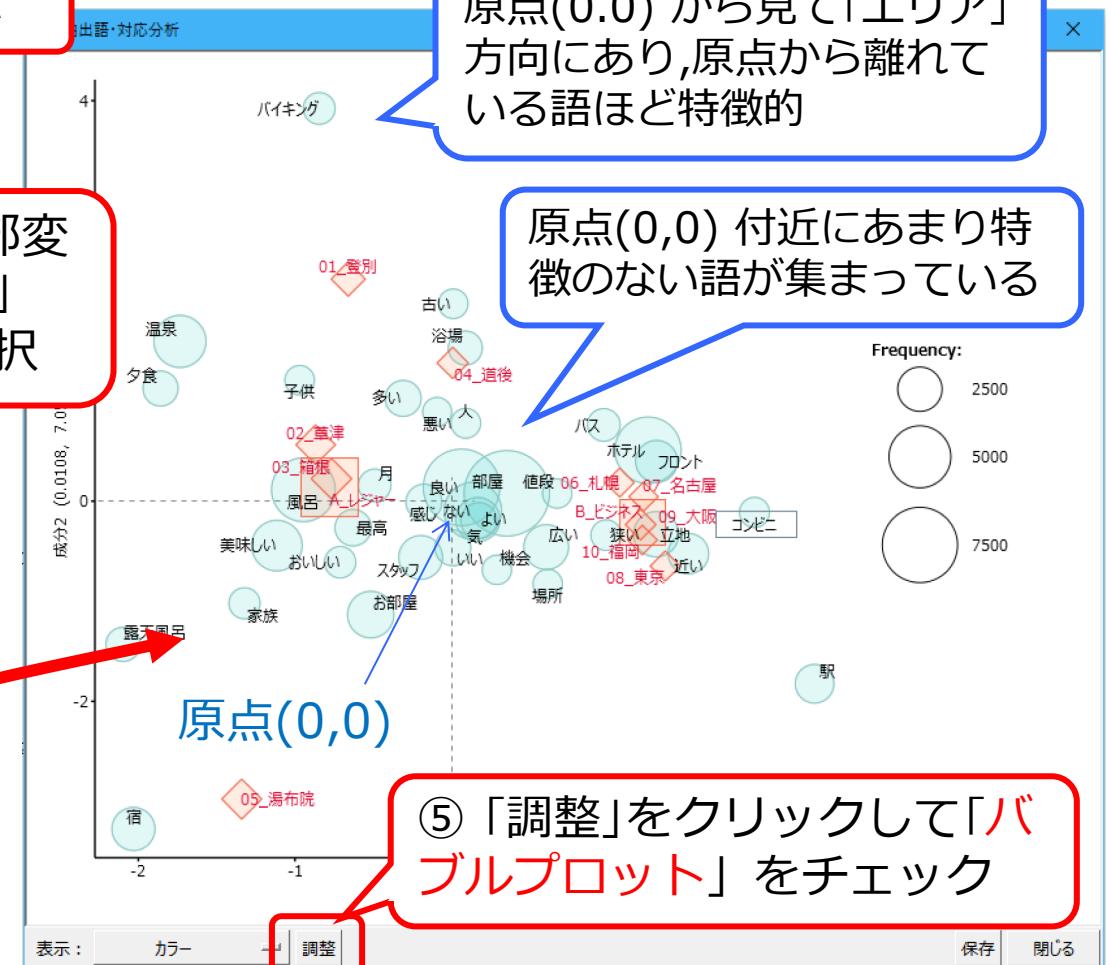
各エリアの特徴語を10件ずつ
一覧 (数値は Jaccard係数)

操作説明 – 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



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



- ③「抽出語×外部変数」「カテゴリー」「エリア」を選択

- ②「文」になっている
ことを確認

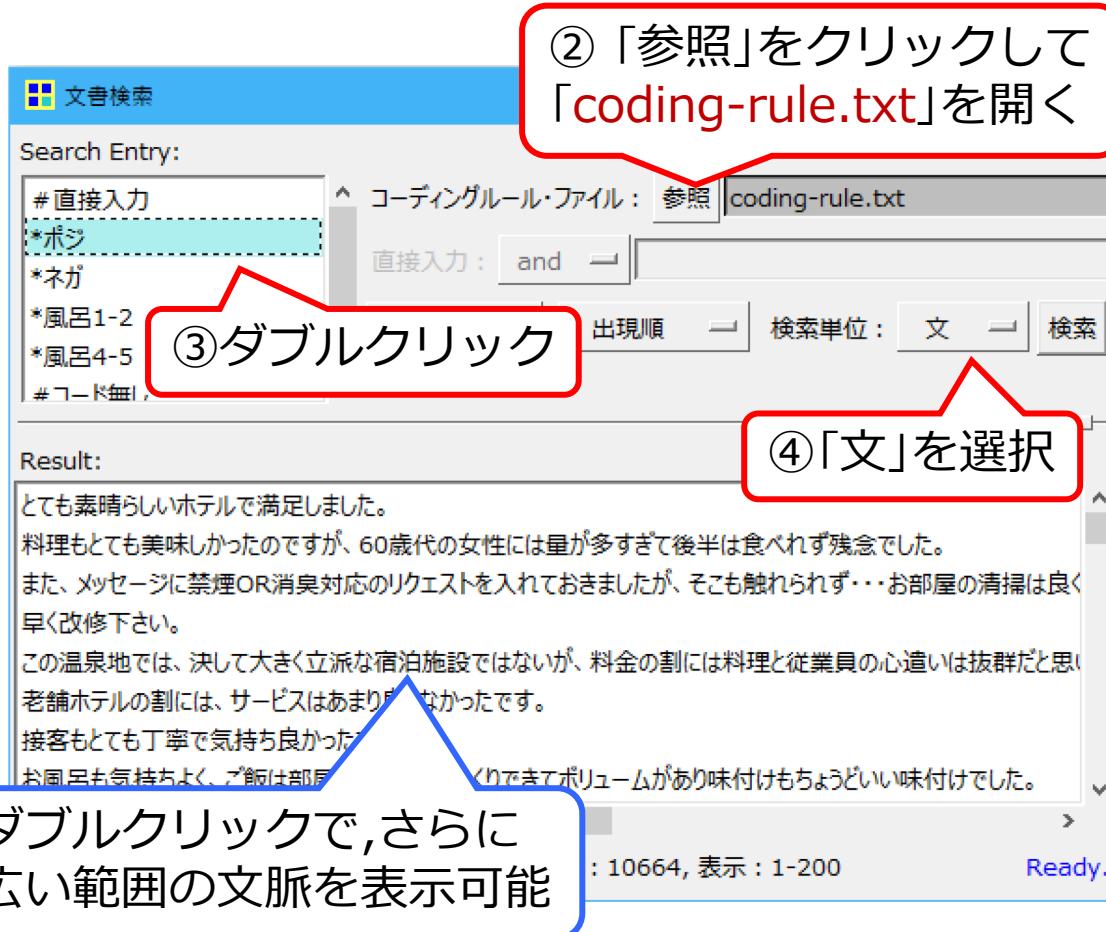
- ④「OK」をクリック

原点(0,0) 付近にあまり特
徴のない語が集まっている

- ⑤「調整」をクリックして「バ
ブルプロット」をチェック

操作説明 - コーディングルール

①メニューから「ツール」「文書」「文書検索」を選ぶ



ダブルクリックで、さらに
広い範囲の文脈を表示可能

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

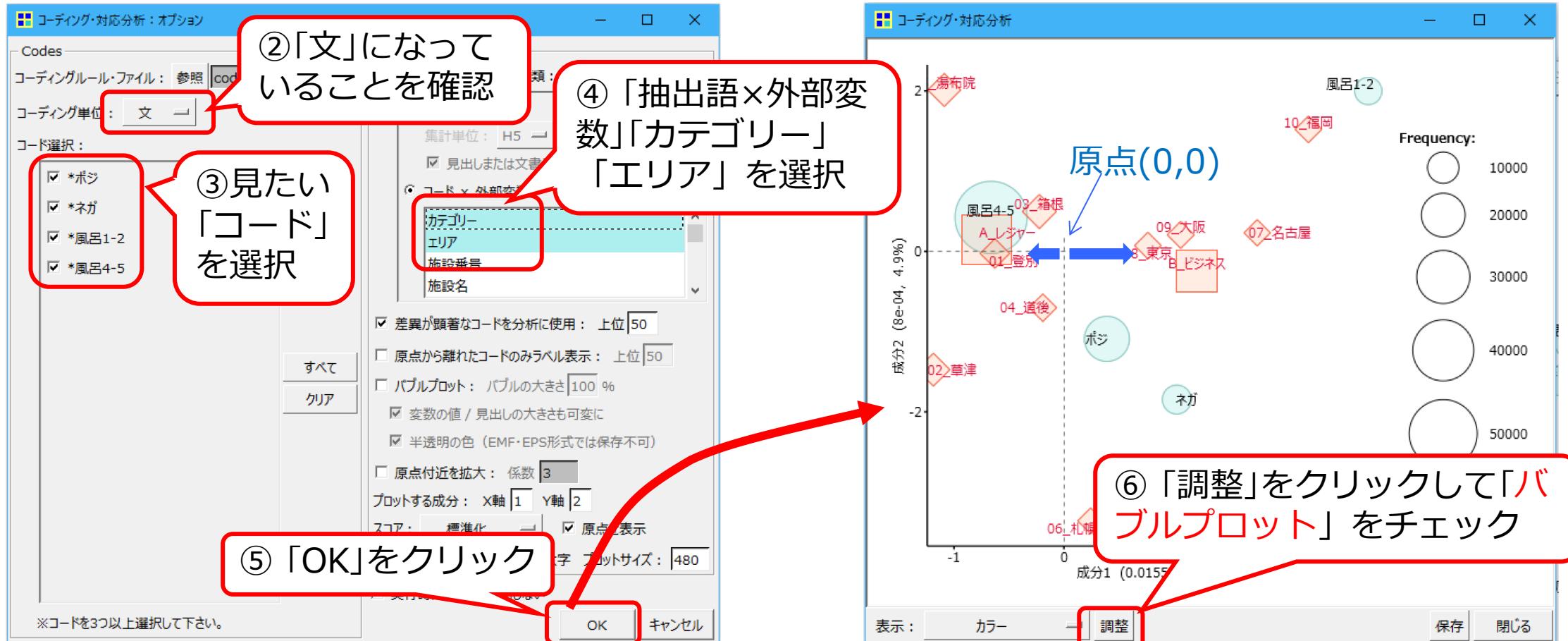
*風呂4-5

<>風呂-->4 | <>風呂-->5

外部変数

操作説明 – 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



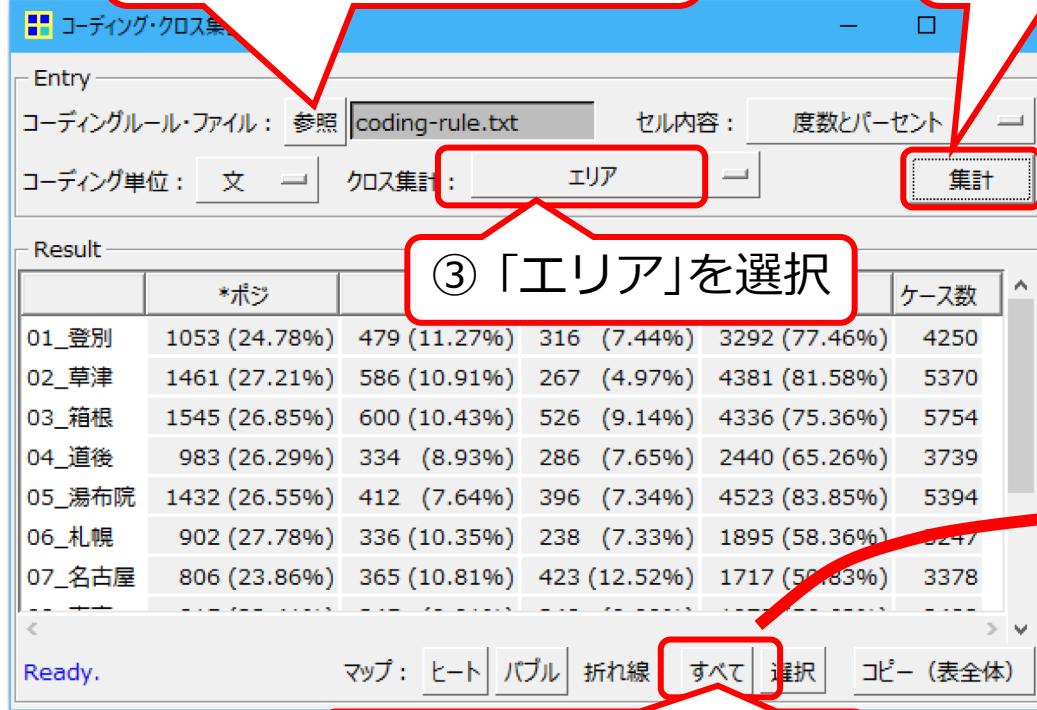
操作説明－クロス集計1

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ

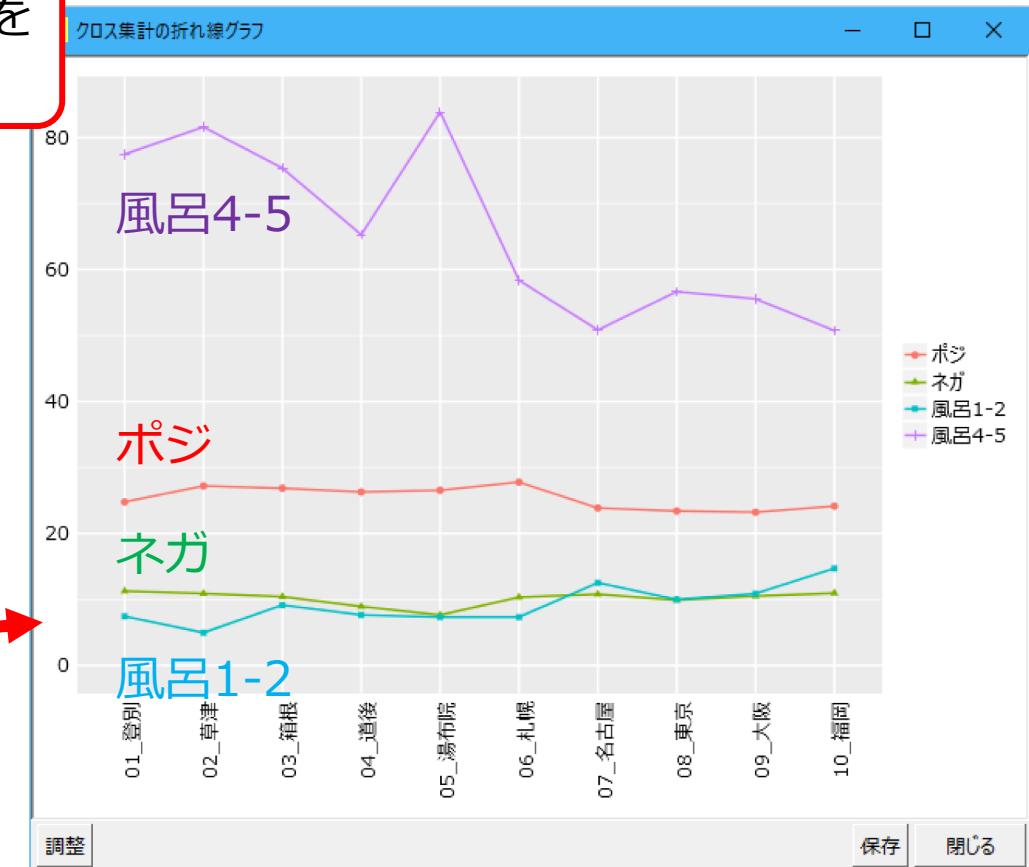
②「参照」をクリックして
「coding-rule.txt」を開く

④「集計」を
クリック

③「エリア」を選択



⑤「すべて」をクリック



練習 一数値評価と口コミの傾向比較

- ・コーディングルール 「coding-rule.txt」 中の「風呂1-2」「風呂4-5」を参考に、「総合1-2」「総合4-5」のルールを追加したコーディングルール 「coding-rule_new.txt」 を作成する
- ・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

操作説明－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

④「集計」を
クリック

③「エリア」を選択

Entry

コーディングルール・ファイル: 参照 coding-rule_new.txt セル内容: 度数とパーセント

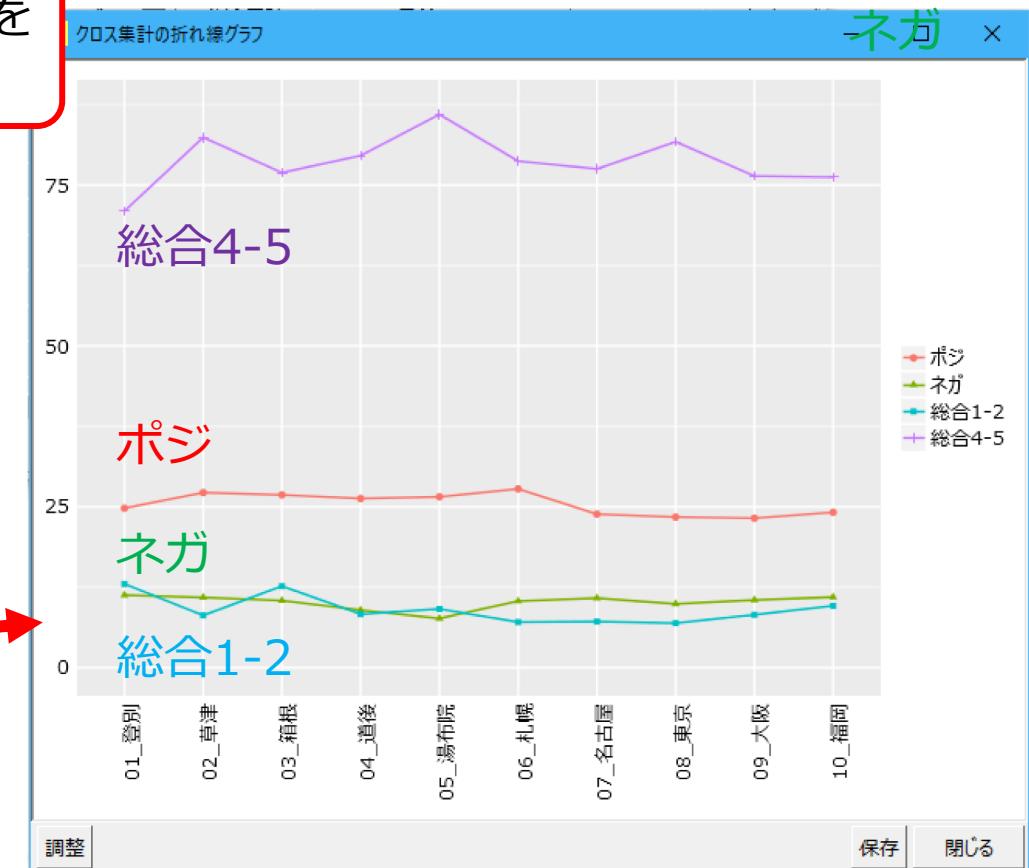
コーディング単位: 文 クロス集計: エリア 集計

Result

	*ポジ					
01_登別	1053 (24.78%)	479 (11.27%)	553 (13.01%)	3015 (70.94%)	316 (7.09%)	*風呂 1
02_草津	1461 (27.21%)	586 (10.91%)	436 (8.12%)	4425 (82.40%)	267 (4.61%)	
03_箱根	1545 (26.85%)	600 (10.43%)	729 (12.67%)	4426 (76.92%)	526 (9.41%)	
04_道後	983 (26.29%)	334 (8.93%)	311 (8.32%)	2975 (79.57%)	286 (7.61%)	
05_湯布院	1432 (26.55%)	412 (7.64%)	492 (9.12%)	4637 (85.97%)	396 (7.61%)	
06_札幌	902 (27.78%)	336 (10.35%)	230 (7.08%)	2556 (78.72%)	233 (7.08%)	
07_名古屋	806 (23.86%)	365 (10.81%)	242 (7.16%)	2618 (75.50%)	423 (12.11%)	
08_東京						
09_大阪						
10_福岡						

Ready. マップ: ヒート バブル 折れ線 すべて 選択 コピー(表全体)

⑤「すべて」をクリック



参考書

(KH Coder)

- [1] 樋口耕一. 社会調査のための計量テキスト分析 –内容分析の継承と発展を目指して–. ナカニシヤ出版, 京都, 2014.
- [2] 樋口耕一. テキスト型データの計量的分析 –2つのアプローチの峻別と統合–. 理論と方法, 数理社会学会, 2004, 19(1): 101-115.

(Windows環境によるCGM収集の参考に)

- [3] テキストマイニングソフトを利用した新未来洞察手法の研究. 第10分科会, (財)市場創造研究会. http://www.shijo-sozo.org/news/%E7%AC%AC10%E5%88%86%E7%A7%91%E4%BC%9A_1.pdf

参考書

(Rを使った参考書)

- [4] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [5] 石田基広. "RMeCabによるテキスト解析. Rによるテキストマイニング入門." 森北出版, 2008, 51-82.

(他のツールを使った参考書)

- [6] 那須川哲哉. "テキストマイニングを使う技術/作る技術: 基礎技術と適用事例から導く本質と活用法." 東京電機大学出版局, 2006.
- [7] 上田隆穂, 黒岩祥太, 戸谷圭子. "テキストマイニングによるマーケティング調査." 講談社, 2005.

(統計解析を中心とした参考書)

- [8] 前田忠彦; 山崎誠. 言語研究のための統計入門. くろしお出版株式会社, 東京, 2013.