

テキストマイニングの実習3

— 形態素解析を利用した集計と分析 —

2015/7/16
ビジネス科学研究科
経営システム科学専攻

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 (特に,経済向き)
- 主な機能
 - 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](#)

961件

※ 2015/7/9 現在

論文検索サービスも提供 →
[http://khc.sourceforge.net/bib.html?
year=2015&auth=all&key=](http://khc.sourceforge.net/bib.html?year=2015&auth=all&key=)

[KH Coderを用いた研究事例](#)

[KH Coderに戻る]

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: すべて 2005 06 07 08 09 10 11 12 13 14 2015-

著者名: すべて あ か さ た な は ま や ら わ A-Z

キーワード: クリア

ヒット件数: 058 / 961

荒瀬雅子 2015 「災害時の『やさしい日本語』を教室教材として使用する方法を探る
—ラジオ放送用災害時音声素材を中心に—」 『龍谷大学国際センター研究年報』
24: 21-34

KH Coder の情報

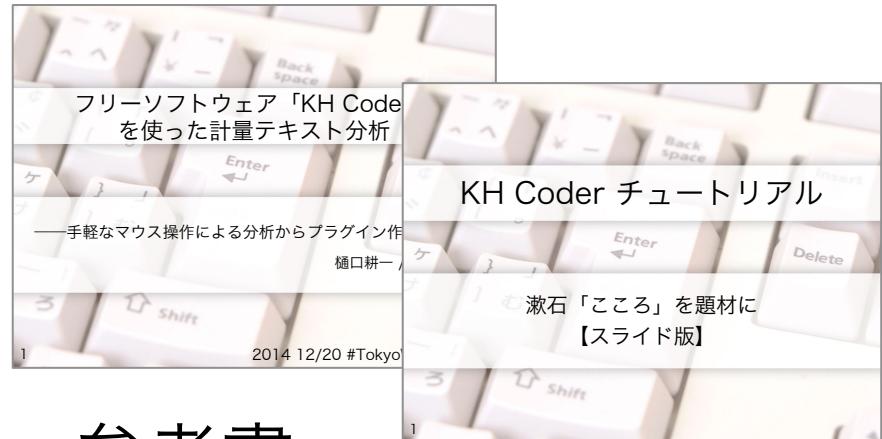
・ ホームページ

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

The screenshot shows the KH Coder Index Page. It features a large blue banner at the top with the text "KH Coder". Below the banner, there's a section titled "概要" (Overview) which describes KH Coder as a tool for statistically analyzing text data. It includes a link to "主な機能と分析の手順". Another section, "機能紹介" (Feature Introduction), lists various features such as "データ中の言葉を探索する", "文書の検索とコーディング", and "複数の言語・環境に対応". A "KH Coderの入手" (How to Get KH Coder) section provides download links and version history. A "サポート" (Support) section includes a forum link and a FAQ page. On the right side, there's a sidebar with a "ツイート" (Twitter) feed from the official account (@khcoder) and a "スライド (slideshare)" section featuring a presentation titled "社会調査のための計量テキスト分析".

・ スライドも充実

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



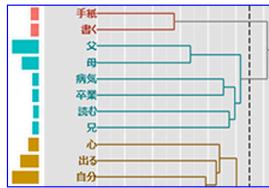
・ 参考書



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。



New! デンドログラム

抽出語は出現数や品詞で選択
最小/最大 出現数による語の取扱選択
最小出現数: 50 最大出現数:
最小/最大 文書数による語の取扱選択
最小文書数: 1 最大文書数:
品詞による語の取扱選択
 名詞
 サ変名詞
 形容動詞

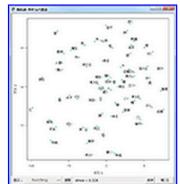
抽出語は出現数や品詞で選択

コード選択:
 *死
 *借用・不信
 *恋愛・結婚
 *病気
 *孤独
 *テスト1
 *テスト2

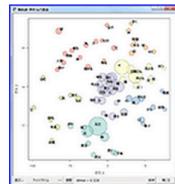
コードはチェックボックスで直接選択

多次元尺度構成法 (MDS)

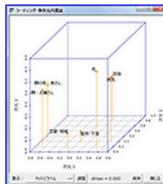
同じく抽出語またはコードを用いての、多次元尺度構成法です。



2次元の解



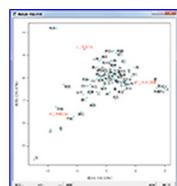
New! クラスタリングと色分け



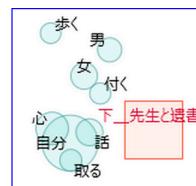
3次元の解

対応分析

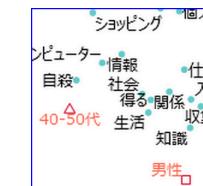
同じく抽出語またはコードを用いての、対応分析です。



同時布置図



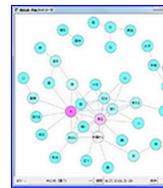
New! バブルプロット



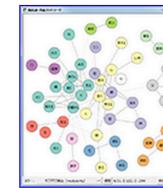
複数の外部変数を用いた多重対応分析

共起ネットワーク

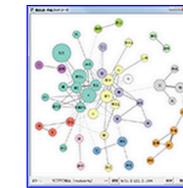
抽出語またはコードを用いて、出現パターンの似通ったものを線で結んだ図、すなわち共起関係を線（edge）で表したネットワークを描く機能です。



共起の程度が非常に強いものだけを線で結んだ図



やや弱い共起関係も描画に含め、自動的にグループ分け（色分け）



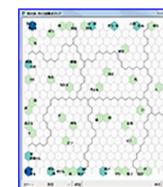
出現数が多い語ほど大きく、また共起の程度が強いほど太い線で描画

自己組織化マップ

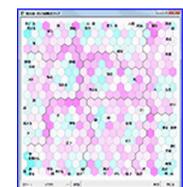
抽出語またはコードを用いての、自己組織化マップです。



クラスター色分け



頻度のプロット



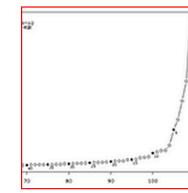
U-Matrix

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。

文書のクラスター分析	クラスター数	各クラスターの文書
文書のクラスター分析	12	1 - 106 - 121 9.007 2 - 127 - 1 9.311 3 - 128 - 1 9.311 4 - 73 - 76 9.375 5 - 50 - 56 9.456 6 - 49 - 52 9.456 7 - 99 - 92 10.347 8 - 65 - 87 10.320 9 - 9 - 10 10.320 10 - 120 - 2 10.320 11 - 121 - 2 10.320 12 - 122 - 2 10.320
文書表示	特徴語	プロト
文書表示	特徴語	プロト
文書表示	特徴語	プロト

クラスター分析の結果画面



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。

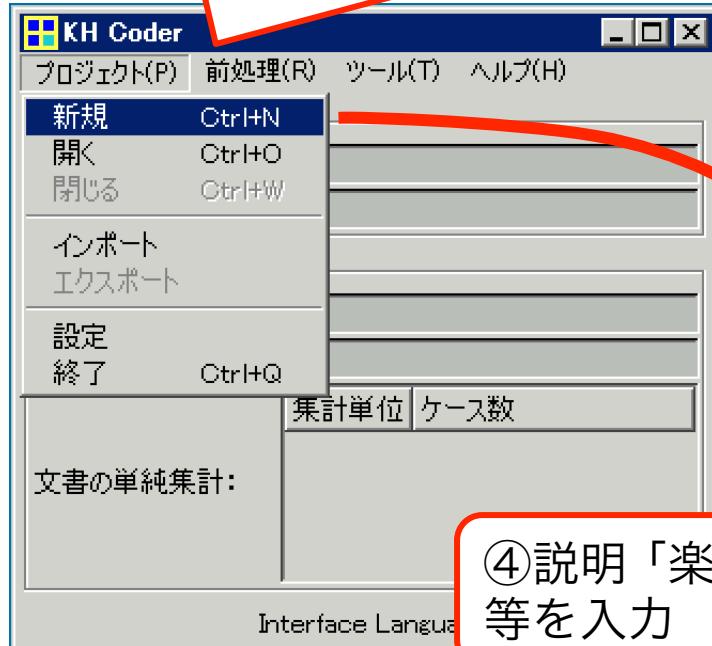


文書のデンドログラム。左の棒グラフは各文書の長さをあらわす。なお、文書数が500を超える場合、デンドログラムは表示不可。

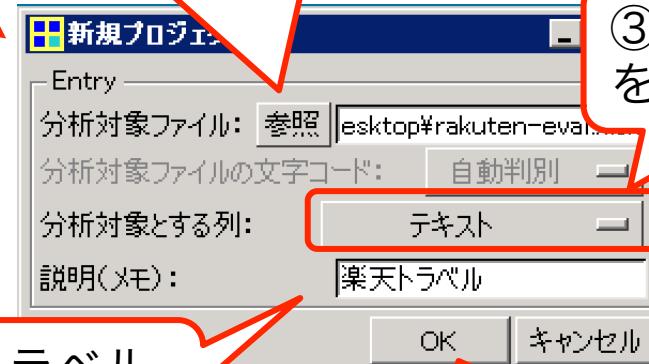
例 — プロジェクトの作成

- ファイル rakuten-eval.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-eval.xlsx」を開く



④説明「楽天トラベル」
等を入力

③「テキスト」
を選択 (注2)

⑤「OK」をクリック

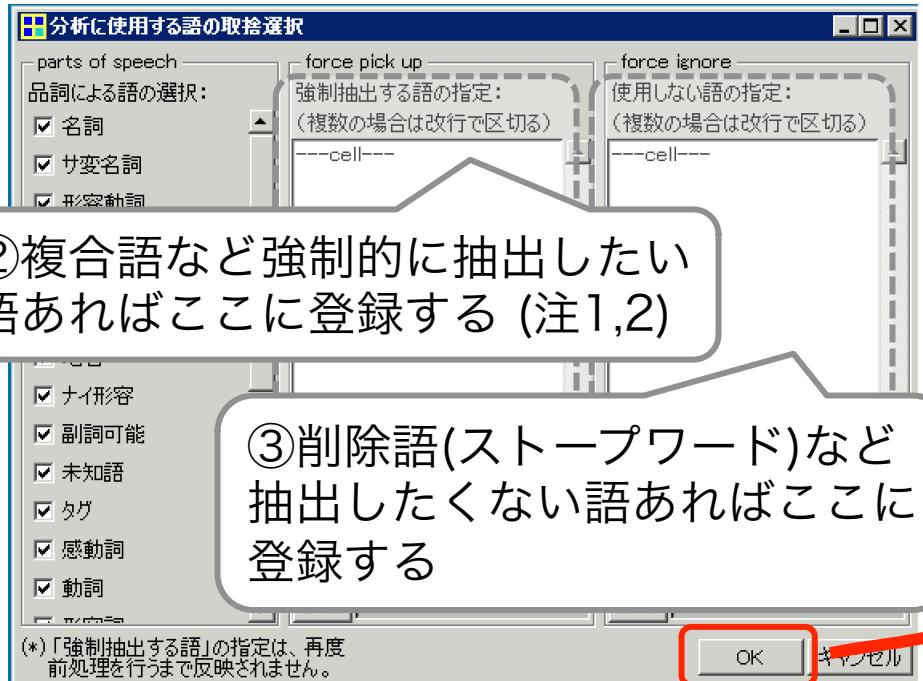
注1: 次回 KH Coderを起動した時は「新規」ではなく「開く」を選択します

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時間がかかります

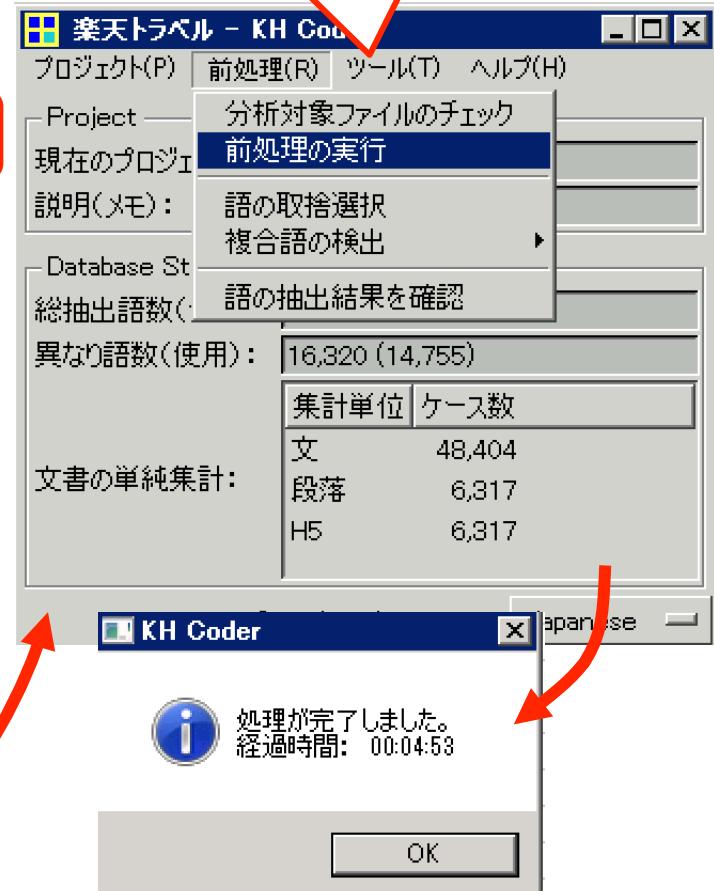
例 — 前処理

• 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



④メニューから「前処理」「前処理の実行」を選ぶ

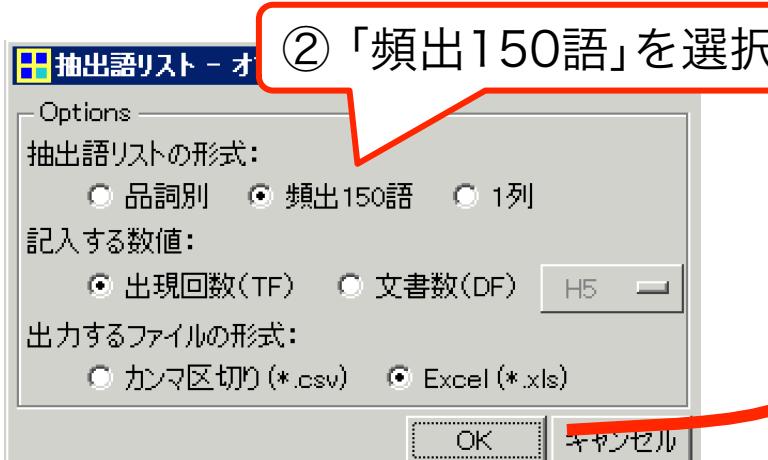


注1: EXCELファイルを読み込んで分析する場合,あらかじめ「---cell---」が入力されています
注2: メニューから「前処理」「複合語の検出」を選ぶと,複合語候補の一覧を出力できます

例 — 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



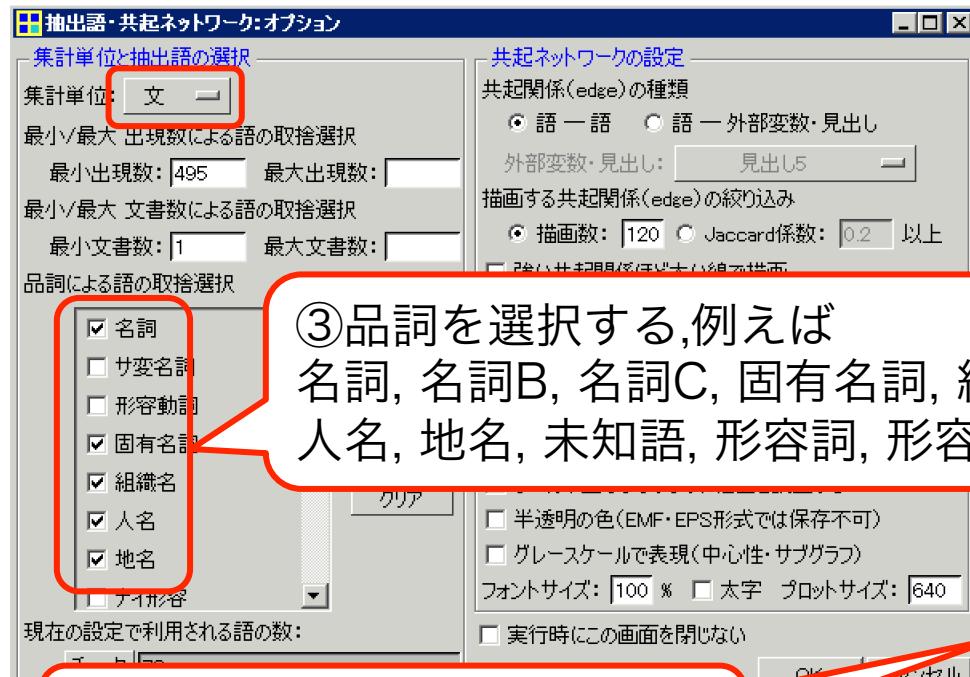
③「OK」をクリック

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		
2 思う	4738	従業	662	近く	390		
3 部屋	4538	刺身	649	到着	389		
4 良い	3759	旅館	643	ご飯	388		
5 風呂	3585	出る	637	プール	388		
6 食事	3503	特に	637	丁寧	388		
7 利用	2875	予約	626	高い	386		
8 宿泊	2776	入れる	601	場所	383		
9 満足	2749	綺麗	595	施設	378		
10 料理	2468	素晴らしい	563	窓	377		
11 美味しい	2255	景色	561	皆さん	376		
12 宿	1993	期待	557	十分	375		
13 海	1843	古い	557	行き届く	372		
14 ホテル	1831	過ごす	546	悪い	371		
15 お部屋	1674	初めて	530	建物	371		
16 行く	1654	貸切	527	接客	370		
17 子供	1581	朝	525	夫婦	370		
18 食べる	1573	清潔	511	もう少し	366		
19 大変	1542	嬉しい	506	助かる	363		
20 露天風呂	1372	味	505	妻	362		
21 温泉	1347	気持ちよい	498	眺め	359		
22 夕食	1330	楽しい	496	種類	358		
23 朝食	1310	目	495	快適	357		
24 サービス	1264	きれい	494	設備	356		
25 残念	1260	夜	494	犬	350		
26 家族	1201	親切	492	料金	348		

例 — 共起ネットワークの作成

①メニューから「ツール」「抽出後」「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, 固有名詞, 組織名,
人名, 地名, 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

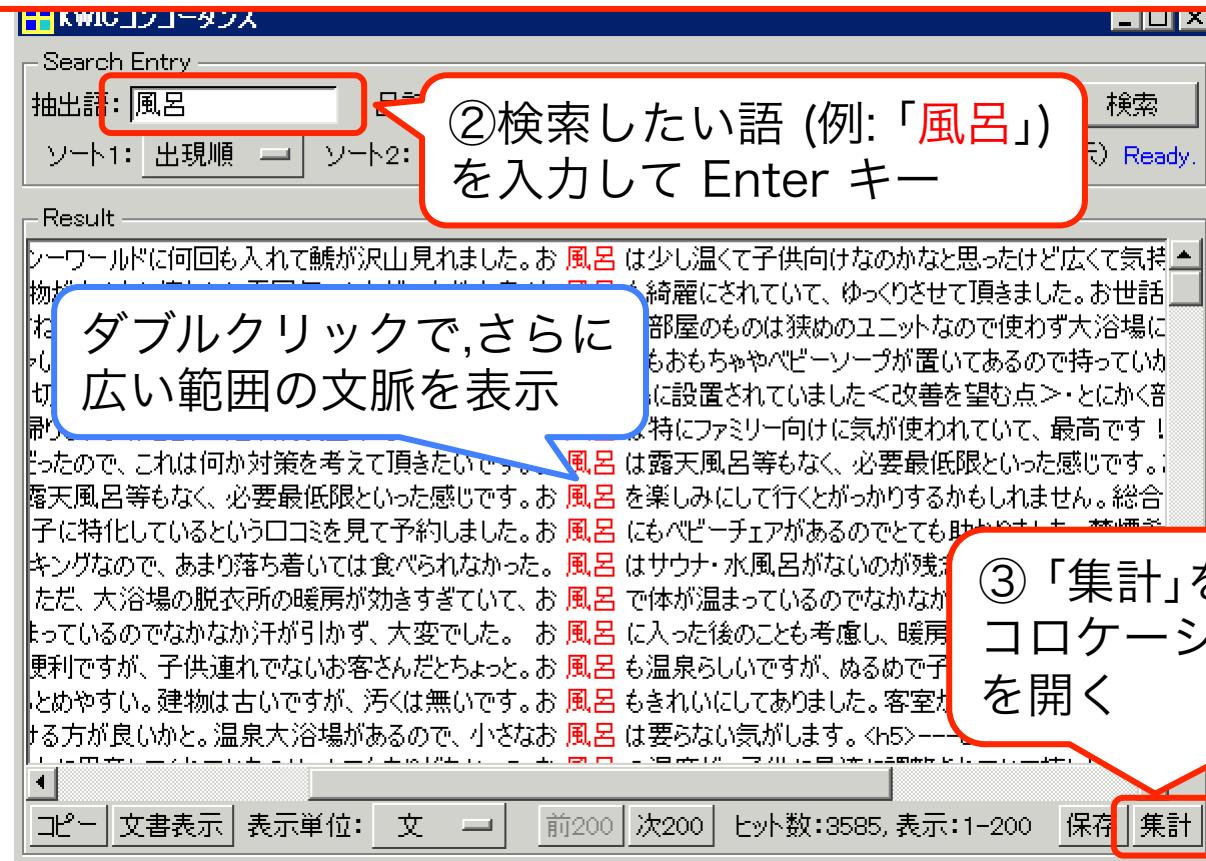
KH Coder の品詞体系

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）	「KH Coder 2.x リファレンス・マニュアル」P.11 より
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	

例 — 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

- メニューから「ツール」「抽出後」「KWICコンコーダンス」を選ぶ



例 — 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

「右1」は左側の1つ目(=直前)に出現していた回数

「広い」は「風呂」の2つ後に 63 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

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	216	59	157	30	15	5	5	4	5	51	27	38	36	74.117
2	広い	形容詞	154	34	120	12	6	9	6	1	0	63	24	24	9	58.200
3	よい	形容詞B	94	33	61	20	6	3	4	0	17	16	12	15	29.333	
4	気持ちよい	形容詞	54	5	5	5	1	1	1	1	13	7	8	19.283		
5	ない	形容詞B	70	21	21	18	18	18	18	18	18	11	10	10	21.283	
6	大きい	形容詞	48	5	5	5	5	5	5	5	6	4	6	6	19.950	
7	狭い	形容詞	50	10	10	10	10	10	10	10	11	4	8	19.033		
8	熱い	形容詞	34	3	31	0	0	0	1	2	1	15	5	7	3	15.017
9	小さい	形容詞	28	2	26	1	1	0	0	0	18	5	1	2	11.767	
10	ぬるい	形容詞B	24	4	20	1	1	0	2	0	0	9	4	1	6	8.733
11	気持ち良い	形容詞	30	10	20	7	1	1	0	0	7	7	2	4	10.117	
12	素晴らしい	形容詞	29	9	20	5	0	3	1	0	1	7	3	2	7	9.900
13	いい	形容詞B	37	19	18	5	3	1	9	1	0	4	8	3	3	13.600

コピーフィルタ設定 ソート: 右合計

KWIC Co-occurrence Statistics - Filter Settings

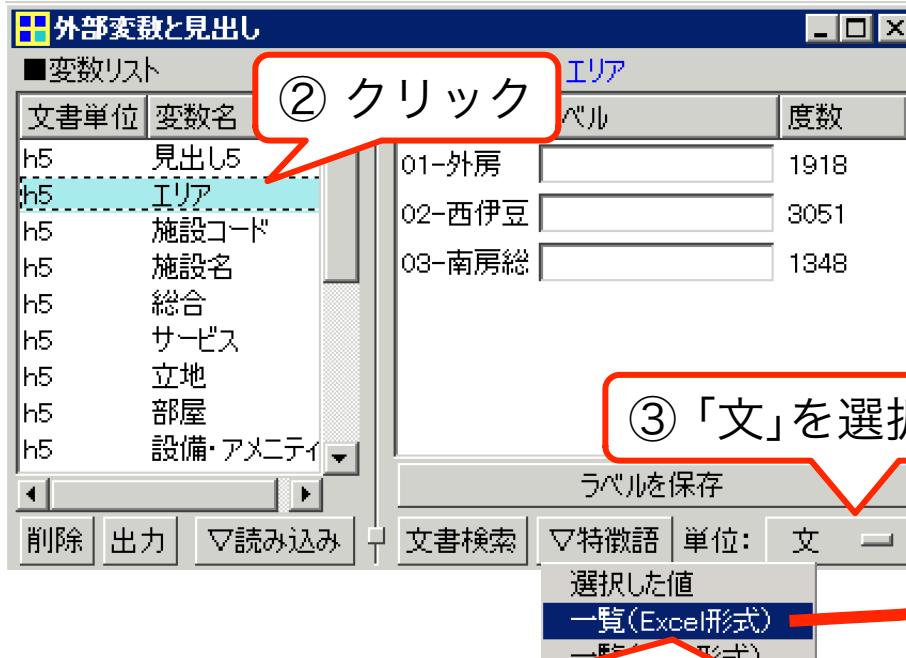
- 品詞による語の選択:
 - 名詞
 - サ変名詞
 - 形容動詞
 - 固有名詞
 - 組織名
 - 人名
 - 地名
- 「合計」列による語の選択:
 - 最小: 1
- 表示する語の数:
 - 上位: 200

OK キャンセル

例 — 外部変数を使う

- 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



- 「特徴語」「一覧(Excel形式)」を選択

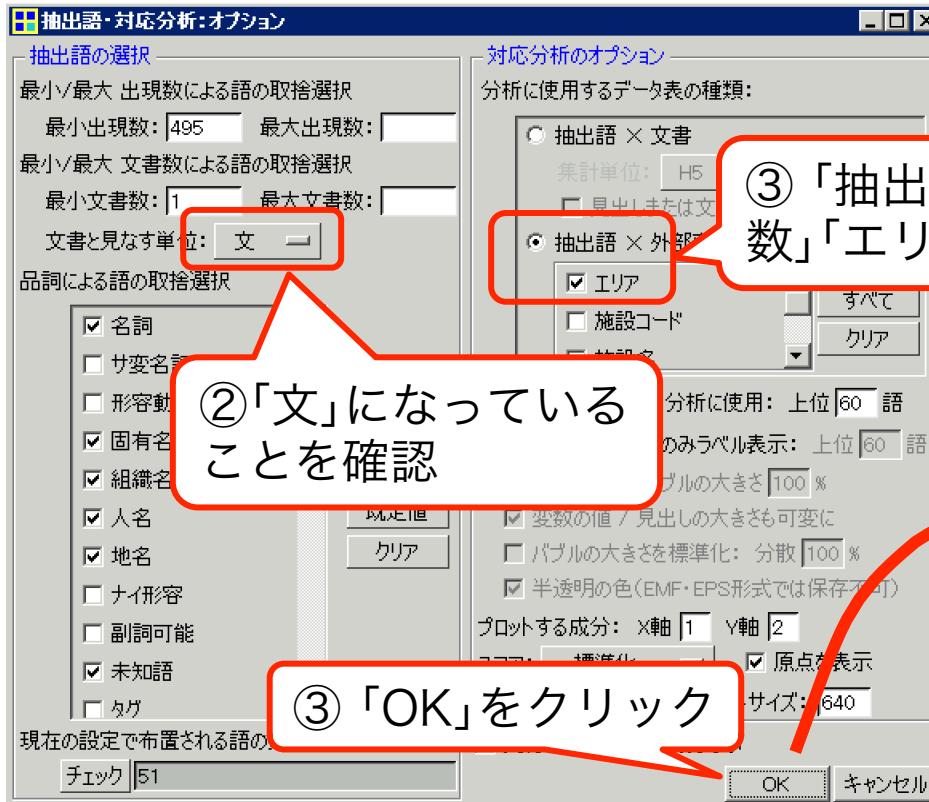
	A	B	C	D	E	F	G	H
1								
2		01-外房		02-西伊豆		03-南房総		
3	部屋	.070	良い	.071	思う	.078		
4	利用	.058	食事	.066	部屋	.066		
5	宿泊	.051	風呂	.064	利用	.052		
6	子供	.042	満足	.056	ホテル	.044		
7	ホテル	.039	料理	.053	海	.033		
8	行く	.031	美味しい	.046	スタッフ	.030		
9	大変	.029	宿	.045	お部屋	.029		
10	朝食	.027	露天風呂	.036	夕食	.028		
11	家族	.025	温泉	.033	朝食	.026		
12	広い	.022	行く	.033	今回	.026		

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

例 — 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



②「文」になっていることを確認

③「抽出語×外部変数」「エリア」を選択

③「OK」をクリック

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)

原点(0,0)付近にあまり特徴のない語が集まっている

④「調整」をクリックして「バブルプロット」をチェック

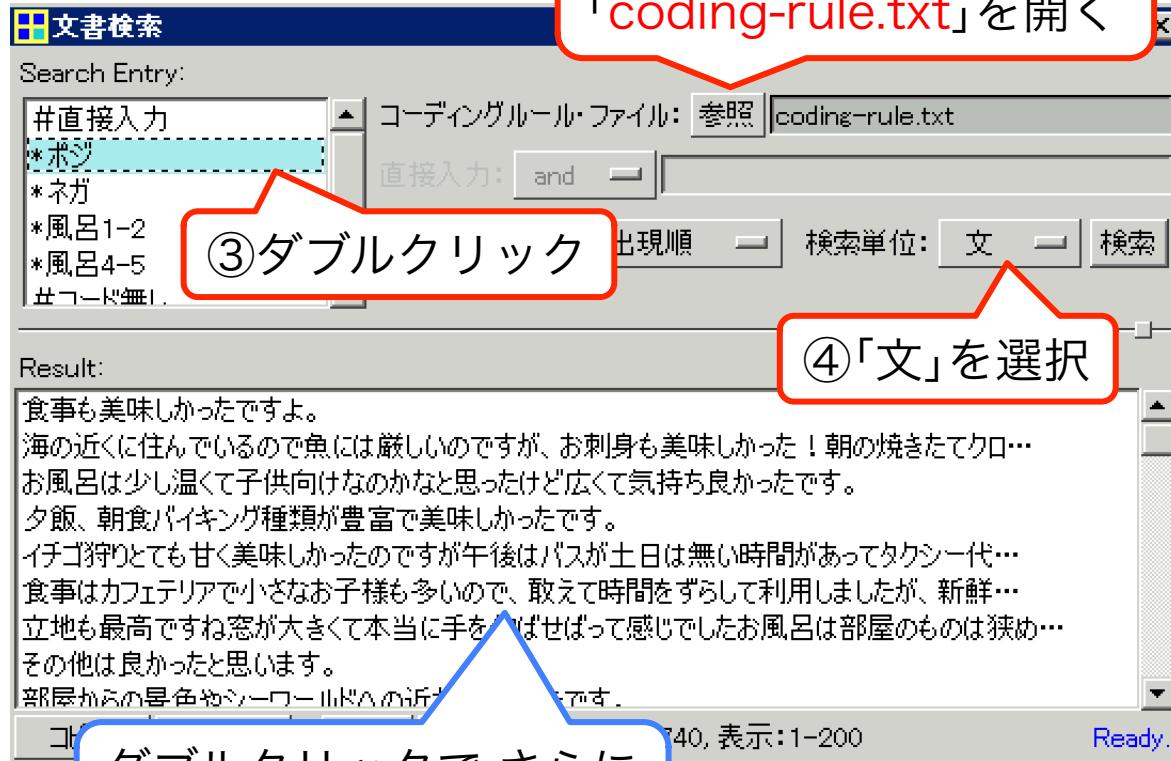
例 — コーディングルール

①メニューから「ツール」「文書」「文書検索」を選ぶ

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



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

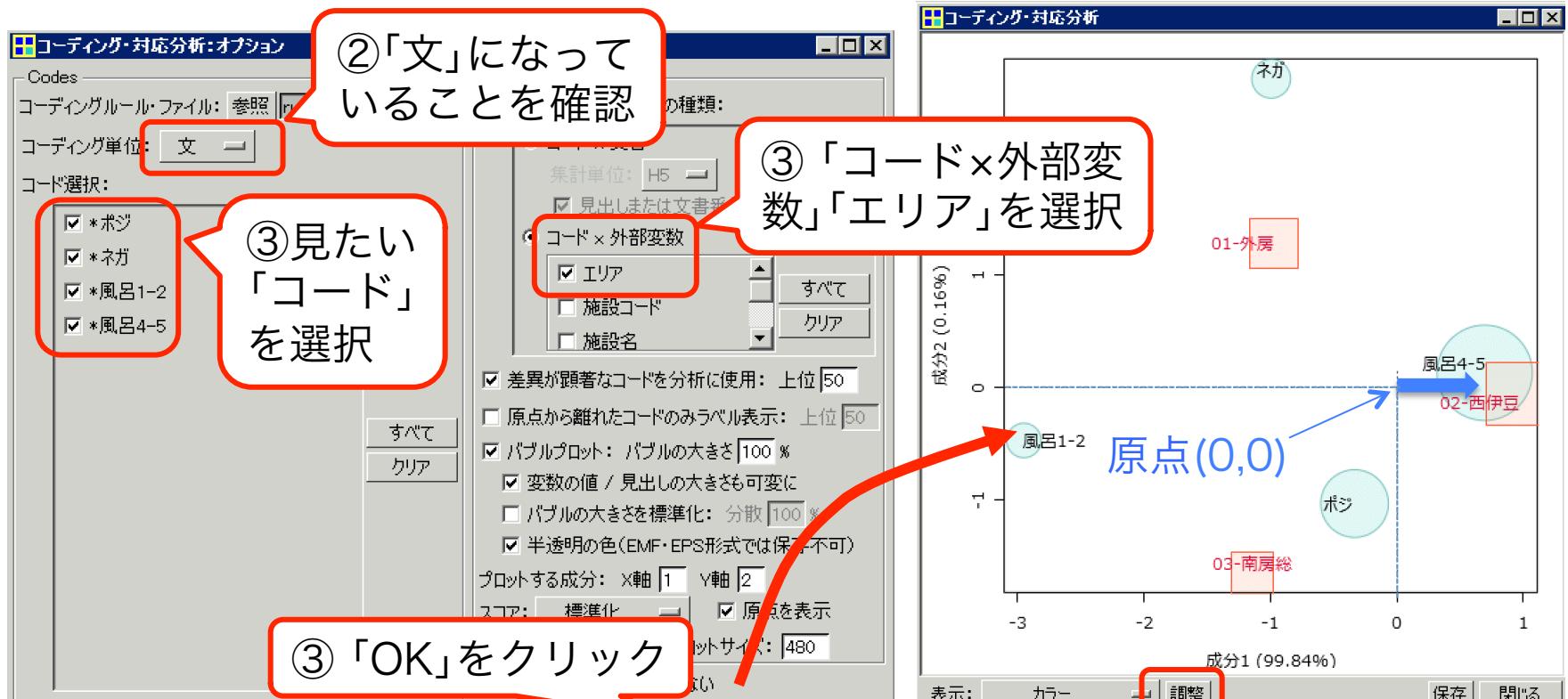
*風呂4-5

<>風呂-->1 | <>風呂-->5

外部変数

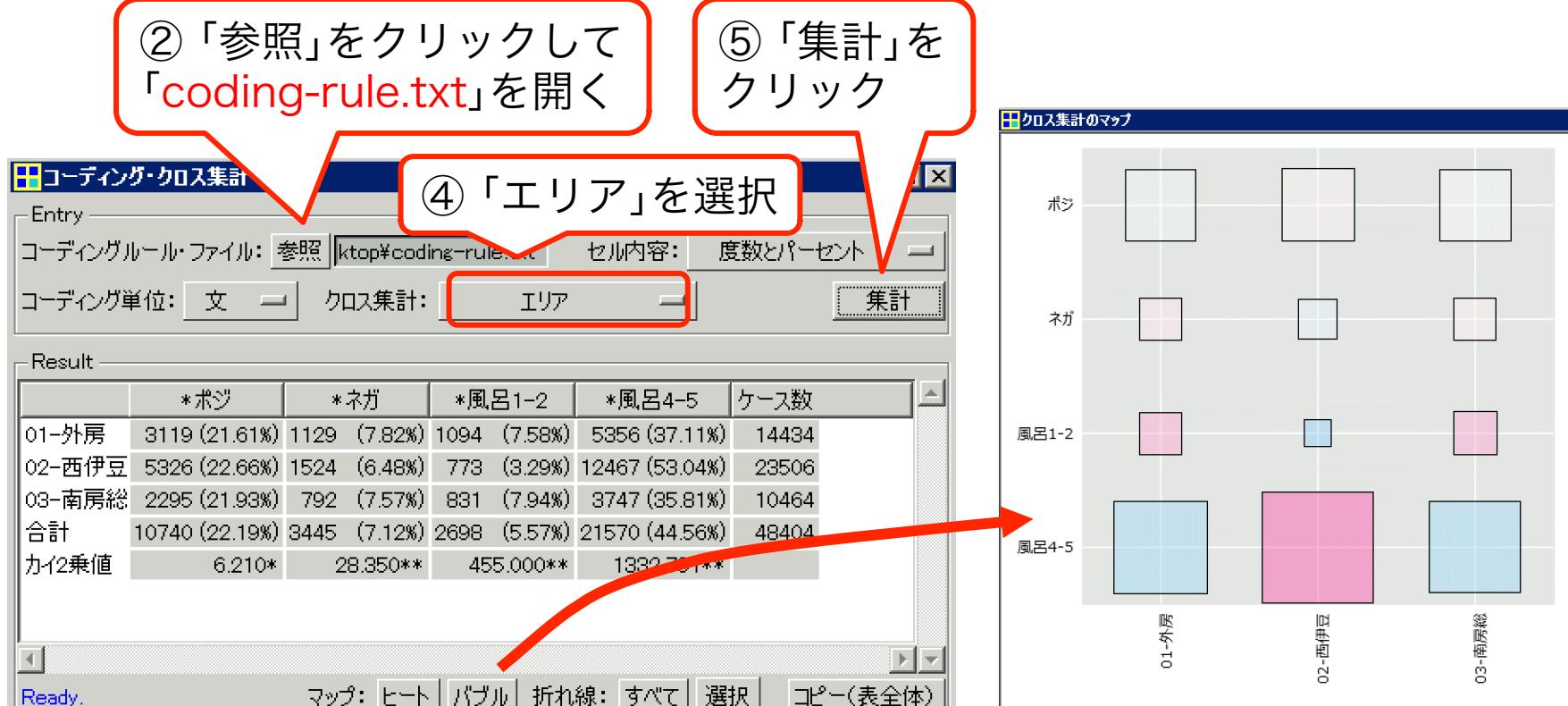
例 — 対応分析による探索2

①メニューから「ツール」「コーディング」「対応分析」を選ぶ



例 — クロス集計

①メニューから「ツール」「コーディング」「クロス集計」を選ぶ



演習課題 (KH Coder)

- 課題1
 - 各エリアの口コミデータ中に発生する単語の出現頻度を集計し、エリアによって高頻度の単語がどのように違うかを比較する
 - 3つのエリアの特徴が分かるか否かを考察する
- 課題2
 - 各エリアの口コミデータ中に発生する名詞「風呂」と形容詞の組みの出現頻度を集計し、エリアによって高頻度の組みがどのように違うかを比較する
 - 3つのエリアの特徴が分かるか否かを考察する
- 課題3
 - 3エリア全体で「風呂」のユーザー評価が低い(評価点1-2)口コミと高い(評価点4-5)口コミを課題1,2の方法を用いて比較する
 - 両者の特徴が分かるか否かを考察する

データの公開場所

- <https://github.com/haradatm/gssm-201507>

The image shows three screenshots of GitHub repository pages:

- Repository: gssm-201507 / +**
 - branch: master
 - first commit by Tomohiko HARADA (9 minutes ago)
 - File structure:
 - 00-slides
 - 01-data** (highlighted with a red box)
 - 02-tools
 - 03-samples
 - 04-docs** (highlighted with a red box)
 - .gitignore
 - README.md
 - Commit history: first commit for each file.
- Repository: gssm-201507 / 01-data / +**
 - branch: master
 - first commit by Tomohiko HARADA (6 minutes ago)
 - ..
 - File structure:
 - coding-rule.txt.zip** (highlighted with a red box)
 - rakuten-eval-sjis.txt
 - rakuten-eval-utf8.txt
 - rakuten-eval.xlsx** (highlighted with a red box)
 - Text annotations:
 - KH Coder演習で使用(要解凍)
 - 演習用データ(変更なし)
- Repository: gssm-201507 / 04-docs / +**
 - branch: master
 - first commit by Tomohiko HARADA (12 minutes ago)
 - ..
 - File structure:
 - khcoder-slides.zip** (highlighted with a dashed blue box)
 - Text annotation:
 - KH Coder のチュートリアル
(from SlideShare)

演習 — KH Coderを使う

- ・ダウンロードとインストール
 - <http://khc.sourceforge.net/dl.html>



- ① ここをクリックすると遷移先のページからダウンロードが始まります
- ② 指示に従いインストール

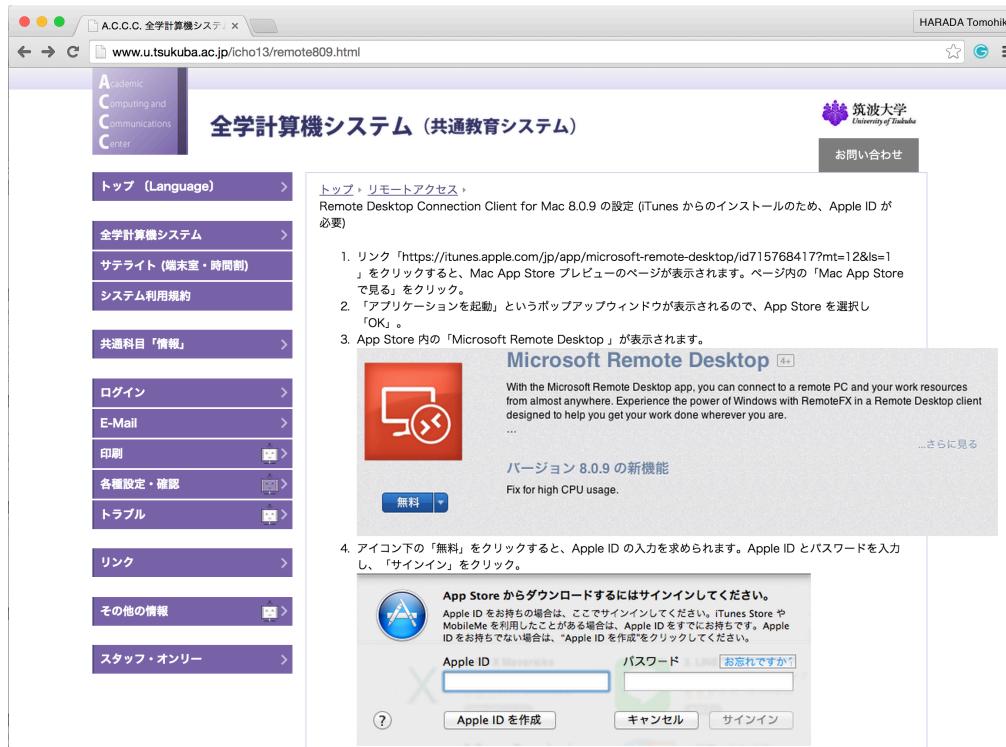
自己解凍ファイルです。このファイルを実行（ダブルクリック）し、開いたWindow の「Unzip」ボタンをクリックすると、（特に変更しなければ）「C:\khcoder」というフォルダにすべてのファイルが解凍されます。解凍されたkh_coder.exeを実行すると、KH Coderが起動します。

注意:

全学のRDPの場合は、ログイン後のデスクトップ上に「khcoder」というフォルダを作成して、その中に解凍してください

演習 — Windows環境のないMac

- 全学計算機システム(RDP)を使います
 - <http://www.u.tsukuba.ac.jp/icho13/remote809.html>



左記のページにある説明に従って、

① ツール (MS Remote Desktop) のインストール

② 全学計算機システム (Windows)へのログイン

を行ってください

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一. 理論と方法, 数理社会学会, 2004, 19(1): 101-115.