科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月 １２ 日現在

機関番号：３４３１５
研究種目：若手研究(B)
研究期間：2011～2013
課題番号：２３７３０５０２

研究課題名（和文）社会調査におけるテキスト型データ分析支援システムの開発

研究課題名（英文）Development of a Textual Data Analysis System for Social Researches

研究代表者
樋口　耕一（HIGUCHI, Koichi）

立命館大学・産業社会学部・准教授

研究者番号：００４５２３８４

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：アンケートの自由回答やインタビュー記録など、社会調査によって得られるさまざまなテキスト型データを分析する方法を検討し、分析用ソフトウェア「KH Coder」を開発した。KH Coderはフリー（自由）ソフトウェアとしてインターネット上で公開し、その使用法を単著書籍として刊行した。また、この分析方法とKH Coderを用いることが、質問紙調査（アンケート）のより良い実践につながる可能性を示した。

研究成果の概要（英文）：In social researches, various textual data such as answers for open-ended questions in surveys and transcripts of interviews are collected. In this research, a quantitative analysis method has been developed for such textual data. And a free software "KH Coder" has been developed to perform the analysis based on the method. The software is freely distributed on the Internet. Also, a book on the method and the usage of the software was published. Finally, a possibility to improve questionnaire surveys by using the method and the software is demonstrated through the actual applied research.

研究分野：社会科学

科研費の分科・細目：社会学・社会学

キーワード：社会調査法　質的研究　内容分析　計量テキスト分析　テキストマイニング　KH Coder

１．研究開始当初の背景

(1) 分析方法の提案と支援システムの開発

社会調査によって得られるテキスト型データには、質問紙調査における自由回答項目のほか、インタビュー記録や会話記録のトランスクリプト、新聞記事など様々なものがある。しかし日本語をコンピュータで扱うことが困難だったこともあって、我が国ではこうしたデータをコンピュータで扱ったり、計量的に分析したりする方法についての研究が進んでいない。

また分析用ソフトウェア、すなわち分析支援システムについても、研究者が自ら開発するか、高価な「テキストマイニング用」ソフトウェアの機能を一部援用するしかない状況である。

(2) 応用に適したデータと研究領域の探索

国内では内容分析を用いた研究が少ないこともあり（三上俊治，1988,「放送メディアの内容分析」『放送学研究』38: 101-18)、本研究で提案する計量的な分析方法がどのような種類のテキストに適しているのか、また、どのような領域のどのような理論的背景を持つ研究に適しているのかということが明確になっていない。他の研究者にとって利用しやすい分析方法として提案するために、これらの点を探索することも重要な課題である。

２．研究の目的

(1) 分析方法の提案と支援システムの開発

分析の信頼性を維持しつつ、各研究者が持つ独自の観点や切り口を活用できるような分析方法の提案を目指す。こうした分析方法は内容分析と呼ばれることが多く、英語圏ではコンピュータ利用が始まってからでも半世紀にわたる方法論の蓄積がある（例えばStone P. J. et al., 1966 The General Inquirer, Cambridge: MIT Press)。そこで英語圏の内容分析に依拠しつつ、近年急速な発展が見られる日本語のコンピュータ処理技術、すなわち自然言語処理技術を取り入れることで、日本語の分析方法を提案する。また、この方法を実現するための支援システム（分析用ソフトウェア）を開発する。

(2) 応用に適したデータと研究領域の探索

応用に適したデータの種類や研究領域を探索するにあたっては、まずは申請者自身が応用研究を示すことを目指す。現在では新聞・雑誌の記事をはじめとして、小説や流行歌の歌詞など様々なテキストの整理と蓄積が進みつつある。これら種々のデータから 1 種類か 2 種類を取り上げて、方法の紹介のための分析例ではなく、実質的な社会学的認識課題にこたえる応用研究を目指す。一般に内容分析を行うことで、分析対象データを生産したコミュニケーションの送り手側や、データを消費する受け手側のおかれた社会的文脈を推論できるとされる（Riffe, D. et al. 1998 Analyzing Media Messages, London: Lawrence Erlbaum)。上述のような応用研究からは、当該の種類のデータを分析することで、いかなる社会的文脈を推論しうるのかを示すことができよう。

３．研究の方法

(1) 分析方法の提案と支援システムの開発

既存研究・関連技術の集約としてまず英語圏におけるコンピュータを用いた内容分析について、文献だけでなく支援システムを収集し評価する。国内でテキストマイニング用として販売されているプログラムについても可能な範囲で購入・評価する。その上で、新しい技術を取り入れることで自動化を目指す部分と、人間・研究者の判断が欠かせない部分、すなわち自動化というよりもむしろ研究者が比較的少ない負担で確認を行えるような支援機能を考える部分とを区別しつつ、新たな分析方法を提案する。こうした区別は決して容易なものではないが、内容分析の方法論・考え方を助けとして判断を行う。そして、分析支援システム（分析用ソフトウェア）KH Coder の開発を進め、フリー（自由）ソフトウェアとして公開する。

(2) 応用に適したデータと研究領域の探索

各種データの分析からいかなる社会的文脈を推論しうるのかを探索するために、内容分析やそれに準じる方法を用いた近年の研究事例のレビューを行う。それに加えて、本研究で提案した方法を用いて、新しい技術の普及過程に関する応用研究を自ら実施する。普及過程論においては、単に新しい技術や製品が広がる過程だけでなく、① 新しい技術が解釈され、その技術に付随する新しい考え方・概念が広がるコミュニケーションの過程が重要とされているが、その過程の定量的な分析は進んでいない。また、② インターネットという新しいコミュニケーション・チャンネルの登場によって普及過程が変質した可能性が示唆されているものの、その具体的な変化の内容は明らかになっていない（Rogers, E. M. 2003 Diffusion of Innovations 5th ed. New York: The Free Press)。これらの点について、本研究で提案する方法によって大量のテキストの分析が可能になるという、独自の強みを活かした貢献を試みる。そのために、特定の新技術に注目し、その技術についての現在の意識を質問紙調査から分析する。

４．研究成果

(1) 分析方法の提案と支援システムの開発

先行研究や市販ソフトウェアをもとに分析方法について検討を進め、独自の分析支援システム（分析ソフトウェア）KH Coder に以下のような機能を追加した。

①　日本語現代文だけでなく、英語や、日本

語の中古和文・近代文語データの分析を行えるようになった。これによって分析できるテキストデータの幅が大きく広がった。なお、英語データを扱う際には、Stanford POS Tagger または Snowball Stemmer を利用することで、文章中から単語を取り出している。また中古和文や近代文語は、MeCab と中古和文 UniDic・近代文語 UniDic を用いることで分析が行える。

② 分析を支援するための機能として、以下のような機能追加と改良を行った。第一に、抽出語やコードの自己組織化マップを作成する機能を追加した。図 1 に自己組織化マップを作成し、U-Matrix を可視可した例を示す。この図で近くに付置された語同士は、よく一緒に使用されていたことを示している。よって、近くに付置された語のグループから、データ中の主題を読み取ることができる。

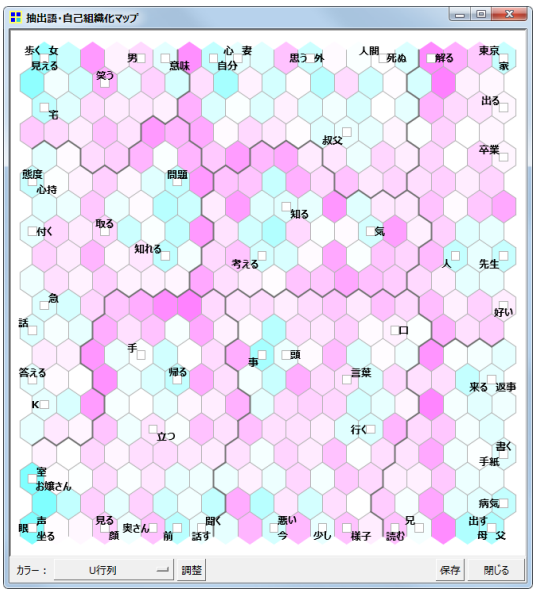

図 1　抽出語の自己組織化マップ

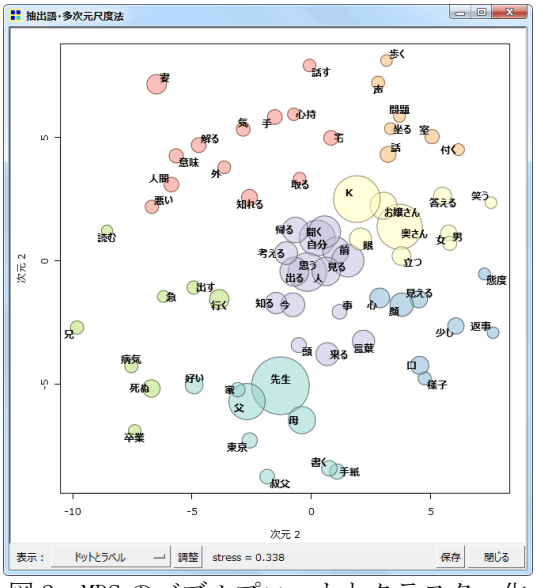

図 2　MDS のバブルプロットとクラスター化

第二に、多次元尺度構成法の結果を示す際に、多く出現していた語ほど大きな円で表示するバブルプロットの機能を追加した（図 2）。こうしたバブルプロットでは、円と円とが重なってしまい、図を読み取りにくくなる場合がある。この点については、それぞれの円を半透明の色で塗りつぶす工夫によって、問題を緩和している。また、50 から 100 といった多くの語（コード）を用いて分析を行うと、プロットの解釈が難しくなる場合がある。そこで解釈の参考のために、クラスター化を行い、その結果を色分けによって示す機能を追加した（図 2）。

第三に、データ中で一緒に出現することが多かった語同士を結ぶ「共起ネットワーク」を描く機能について、以下の改良を行った。1 つは、特定の語に注目し、その語と関連が強い語をピックアップして、それらの語の共起をネットワークとして描く機能である。この機能の使用例を図 3 に示す。図 3 では、はじめに注目した語「父」を他の語と区別するために 4 角形で描画している。

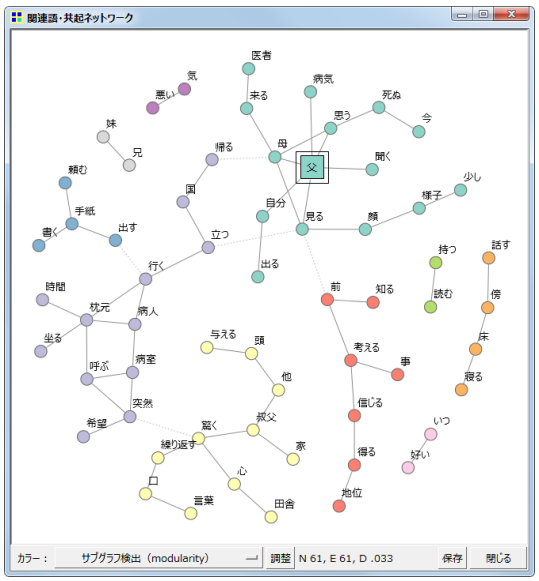

図 3　特定の語を中心とするネットワーク

またネットワークを描く際に、最小スパニングツリーを表示する機能を加えた（図 4）。語と語を結ぶ線（edge）が多くなった場合には、どの edge が重要かを示す手がかりがあった方がプロットを解釈しやすい場合がある。そこで解釈の参考のために、最小スパニングツリーの一部かどうかという観点から、重要とみられる edge を選び、強調する（色の濃い太線で示す）機能を追加した。具体的には、Prim の方法で共起の強さを考慮した最小スパニングツリーをもとめ、この最小スパニングツリーを構成する edge をすべて強調している。

さらに、コミュニティ検出アルゴリズムによるグループ分けを行った際に、グループ分けの結果をグレーの濃淡で表現する機能を追加した（図 5）。これはグレースケール印刷

で論文を刊行する際でも、グループ分けの結果を読み取りやすくするための工夫である。

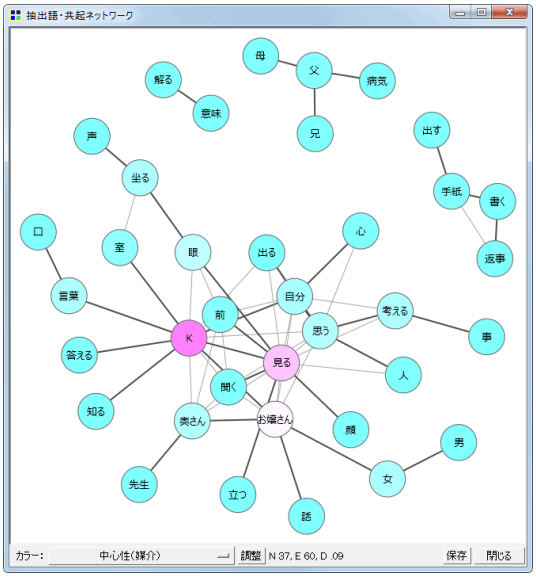

図 4　最小スパニング・ツリーの表示

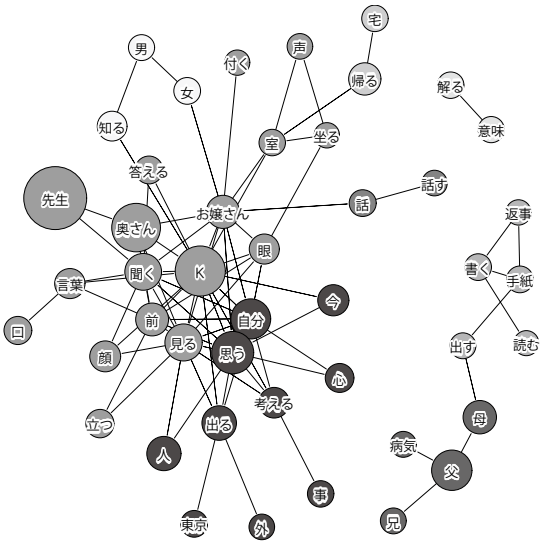

図 5　グレーの濃淡によるグループ分け表現

第四に、抽出語やコードのクラスター分析の結果を示す際に、クラスターを色分けによって示すとともに、抽出語やコードの出現数を棒グラフによって示す機能を追加した（図 6）。

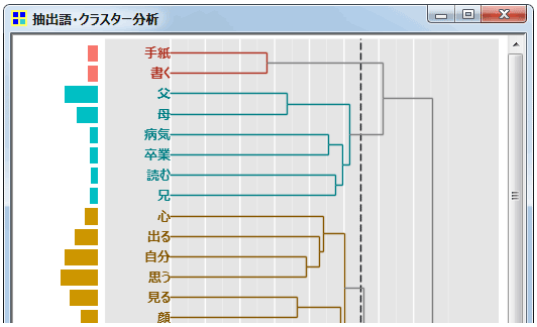

図 6　クラスター分析の機能

第五に、コーディングを行った際に、各コードがデータ中のどの部分に多く出現していたかを可視化するための機能として、ヒートマップ（図 7）およびバブルプロット（図 8）を作成する機能を準備した。

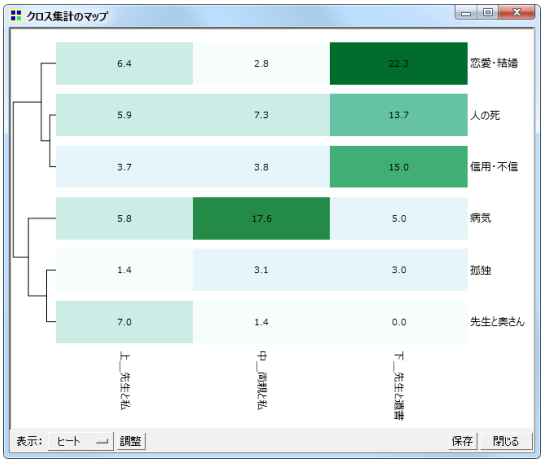

図 7　コーディング結果のヒートマップ

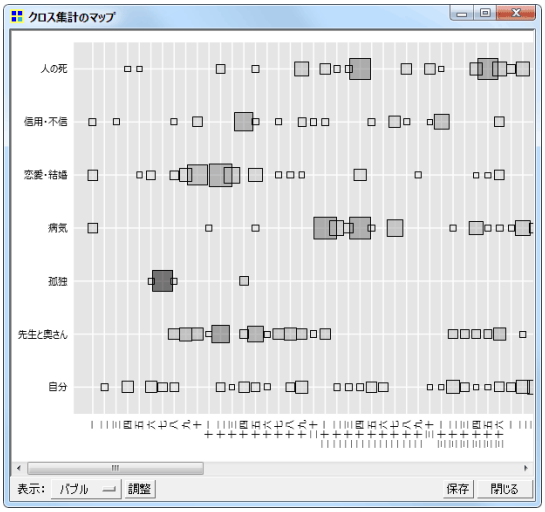

図 8　コーディング結果のバブルプロット

③　さらに、これらの機能を活用して計量テキスト分析を行う方法について、単著書籍にまとめて公表した（樋口耕一，2014，『社会調査のための計量テキスト分析』ナカニシヤ）。なおソフトウェア「KH Coder」そのものはインターネット上でフリー（自由）ソフトウェアとして公開している。

(2) 応用に適したデータと研究領域の探索

計量テキスト分析の方法と KH Coder を用いて、新しい技術の普及課程に関する応用研究に自ら取り組んだ。この成果は 2013 年に学術雑誌『ソシオロジ』に公表した。

この応用研究を通じて、計量テキスト分析と KH Coder を活用することが、質問紙調査（アンケート）のよりよい実践につながる可能性を示した。質問紙調査はさまざまな分野で頻繁に利用される調査法であるため、この調査法の改善は重要な課題である。

従来の質問紙調査において，回答を統計的に処理するためには，限られた選択肢の中から答えを選ぶよう回答者に求めざるをえない場合が多かった。しかし計量テキスト分析を用いれば，回答者の自由な言葉で答えてもらう自由回答についても，統計処理を通じた分析が容易に行える。自由回答型の質問項目を用いる利点の１つは，網羅的で完全な選択肢を準備できないような場合にも，質問を行えることである。どんな選択肢を準備して良いかわからないときも、自由回答項目を使えば探索的に研究を進められる。

次に，自由回答データの計量テキスト分析は，通常の選択肢型質問の分析と組み合わせることで，より大きな利点を生じうる。たとえば，通常の質問を使った分析結果を見て，その解釈に悩んでしまうことは決して珍しくないだろう。なぜ学歴の高い人は，大学の授業で直接的に習ったわけでもないのに，一定の行動をとる傾向があるのか。なぜ特定の社会的属性が，ある従属変数に対して強い効果を示すのか。このような解釈が難しい部分について，言わばごく小規模なインタビューである自由回答項目の分析が大きな手がかりとなる。

こうした利点を実際の研究事例を通じて提示し、質問紙調査における計量テキスト分析・KH Coder の活用を提案した。

５．主な発表論文等

〔雑誌論文〕（計６件）

① 樋口耕一，2013,「自由回答項目によるマイクロインタビューと通常項目による線形モデルとの連携 ： 質問紙調査の分析事例から」『日本行動計量学会大会発表論文抄録集』 41: 288-289. http://ci.nii.ac.jp/naid/110009731796 査読無
② 樋口耕一，2013,「情報化イノベーションの採用と富の有無――ウェブの普及過程における規定構造の変化から」『ソシオロジ』 57(3): 39-55. 査読あり
③ 樋口耕一・中井美樹・湊邦生，2012,「Web 調査における公募型モニターと非公募型モニターの回答傾向――変数間の関連に注目して―」『立命館産業社会論集』 48(3): 95-103 査読無 http://www.ritsumei.ac.jp/acd/cg/ss/sansharonshu/483pdf/48-3_03-01.pdf
④ Wataru Ozawa, Yukifumi Makita, Koichi Higuchi, Kuniko Ishikawa, Hiroko Yamada, Martha Mensendiek, Eiji Ogawa, Hiroshi Kato 2012, “Volunteer Support Network for Elderly Foreigners: A New Movement of Korean Residents in Kyoto” 『立命館産業社会論集』 48(3): 19-40. 査読無 http://www.ritsumei.ac.jp/acd/cg/ss/sansharonshu/483pdf/48-3_02-02.pdf
⑤ 樋口耕一，2012,「質問紙調査における自由回答の分析――KH Coder による計量テキスト分析の手順と実際―」『社会と調査』8: 92-96. 査読無
⑥ 酒井佐枝子・稲垣由子・樋口耕一・加藤寛，2011,「児童養護施設内における子ども間暴力の内容と対応の分析」『子どもの虐待とネグレクト』 13: 115-124. 査読あり

〔学会発表〕（計３件）

① Wataru Ozawa, Yukifumi Makita, Koichi Higuchi, Kuniko Ishikawa, Hiroko Yamada, Martha Mensendiek, Eiji Ogawa & Hiroshi Kato “Voluntary Support Network for the Elderly Foreigner: A New Movement of Korean Old Comers in Kyoto” 10th International Conference of the International Society for Third-Sector Rsearch (Siena, Italy: July 12). 2012
② Koichi Higuchi “`The Internet’ in Newspaper Articles and People’s Minds: A Corpus-Based Exploratory Approach to Social Consciousness in Japan” 4th International Conference on Corpus Linguistics (Jaen, Spain: 3/22). 2012
③ 樋口耕一・中井美樹・湊邦生 「Web 調査における公募型モニターと非公募型モニターの回答傾向」 数理社会学会第 53 回大会，於・鹿児島大学，2012 年 3 月 14 日

〔図書〕（計２件）

① 樋口耕一，2014,『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. 223
② 樋口耕一，2012,「社会調査における計量テキスト分析の手順と実際(第 10 章)」「今日から始めるテキストマイニング――計量テキスト分析の環境『KH Coder』」 石田基広・金明哲編著 『コーパスとテキストマイニング』共立出版 119-128, 204-209.

〔その他〕

① 分析支援システム「KH Coder」の Web サイト： http://khc.sourceforge.net

６．研究組織

(1)研究代表者

樋口　耕一（HIGUCHI ,Koichi）
立命館大学・産業社会学部・准教授
研究者番号：00452384

(2)研究分担者

(3)連携研究者