様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　８　日現在

機関番号：　３４３１５
研究種目：　若手研究(A)
研究期間：　2014～2017
課題番号：　２６７０５００６

研究課題名（和文）社会調査のための計量テキスト分析と実践に必要なソフトウェアの開発

研究課題名（英文）Proposal of Quantitative Text Analysis Method for Social Researchers and Development of Required Software

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）　　8,000,000円

研究成果の概要（和文）：　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記述のほか、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。本研究では、こうしたデータをコンピュータによって計量的に分析する方法を提案し、その実践に必要な分析用ソフトウェアとして「KH Coder」の開発を行なった。また本方法とソフトウェアをどのように使用すれば学術的価値の高い研究につながるかという点について検討を行なった。

研究成果の概要（英文）：Various textual data are obtained by social researches, such as responses to open ended questions in questionnaires, transcripts of interview records, conversation records, and newspaper articles. In this research, we propose a method to quantitatively analyze such data using a computer. And developed "KH Coder" as analysis software necessary for its practice. We also examined how to use this method and software to carry out high academic value research.

研究分野：　社会学

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

１．研究開始当初の背景
(1)　分析方法の提案
　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記述のほか、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のが困難だったこともあって、国内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必ずしも進んでいなかった。こうした分析方法を確立できれば、例えば新聞記事から流行歌の歌詞まで、現在整理と蓄積が進みつつある様々なテキスト型データについて、その社会学的な利用価値がいっそう高まるものと考えられる。

(2)　分析用ソフトウェア開発
　こうした分析のためには適切なコンピュータ利用が不可欠であるが、従来は分析に適したソフトウェアが無かったために、研究者が各自でソフトウェアを開発せねばならなかった。あるいは、テキストマイニング用として販売されているソフトウェアの機能を一部援用することは可能であったが、こうしたソフトウェアの価格は少なくとも数十万円から数百万円と高価であり、新しい方法を試すために容易に調達できるものではなかった。

(3)　上手く活用するための方策
　国内では内容分析を用いた研究が少ないこともあり、こうした方法・ソフトウェアが、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かった。

２．研究の目的
(1)　分析方法の提案
　社会科学の分野において日本語テキスト型データを計量的に分析するための、実用的で利用しやすい分析方法を提案する。

(2)　分析用ソフトウェア開発
　この方法による分析を行なうために必要なソフトウェアを開発し、フリーソフトウェア（自由ソフトウェア）として広く一般に公開する。

(3)　上手く活用するための方策
　この方法とソフトウェアの利用に適したデータと研究領域を探索し、方法とソフトウェアを上手く活用するための方策を示す。

３．研究の方法
(1)　分析方法の提案
　テキスト型データを統計的に分析する方法は、社会学の分野では「内容分析（content analysis）」と呼ばれることが多い。この内容分析の方法を土台にして、近年急速な発達を見せる統計処理技術、データ可視化技術、自然言語処理技術などを取り入れることで、方法の発展を目指す。たとえば係り受け解析のような自然言語処理と、高次元判別分析（HDDA）のような統計解析など、新しい技術を評価する。
　また技術の進化を取り入れることで自動化を目指す部分と、人間による判断を支援する部分とを区別することが重要である。社会科学分野では、分析を行なう研究者の独自の着想やひらめきを活かすことができなくてはならない。そのために、人間がデータを閲覧して判断する一連の作業を、簡単に行なえるような仕組みを検討する。

(2)　分析用ソフトウェア開発
　手軽なマウス操作で、誰でも容易に分析を行なうことができるよう、独自の分析用ソフトウェアを開発する。そしてこのソフトウェアを、処理内容をすべて開示し、さらに必要に応じて利用者が機能を修正できるフリーソフトウェア（自由ソフトウェア）として公開する。
　ここで開発する分析用ソフトウェアは、日本語データだけでなく英語データも分析できるように開発する。これによって応用範囲を英語データにも拡大するとともに、国内外にソフトウェアを公開して国際的な研究交流を促進する。このためには、ソフトウェアの画面に加えて、チュートリアル・マニュアル類についても英語版を準備する必要がある。

(3)　上手く活用するための方策
　まずは筆者自身がこの方法とソフトウェアを利用した応用研究を示すことで、活用法を提示することを目指す。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。
　そうした応用研究の実施に加えて、似通った方法を用いた研究事例のレビューを併せて行う。さまざまな研究事例を調べ上げる中で、どのようにこの方法・ソフトウェアを活用すれば、学術的意義のある研究となりやすいのかを探索する。

４．研究成果
(1)　分析方法の提案
　まずコンピュータを用いた内容分析について、先行する研究を収集・検討した。主として英語圏のテキストマイニングや内容分析に関する文献に加えて、実際の分析に用いられているソフトウェアについても収集・検討した。内容分析ソフトウェア「Word Stat 7」、質的データ分析ソフトウェア「Atlas.ti 7」

「Nvivo 10」、コーパス分析ソフトウェア「WordSmith 6」、さらに統計ソフトウェア「JMP 11」を調達し、これらのソフトウェアを用いて試験的な分析を行なった。

こうしたレビューを通じて、第一に本研究で提案する手法「計量テキスト分析」でも活用すべき分析機能、あるいは取り入れるべき分析手順がないかどうかを探索した。第二に、本研究で提案する手法を、日本語以外のテキストデータに適用することが可能かどうかについて検討を行なった。その結果、本研究で提案する「計量テキスト分析」を、日本語データだけでなく英語・中国語・韓国語・ロシア語等の文章データに対しても適用する目処が立った。

その上で、実際に本提案手法を分析に使用しつつ、この方法の有効性についての検証と評価を行った。分析結果の妥当性を多くの人が判断しやすいように、内容が多くの人によく知られているデータとして小説『赤毛のアン』原文（英語データ）を取り上げ、試験的な分析を行った。この分析を通じて、①データ量が多くても判断基準の揺らがない分析が可能かどうか、②事例的な解釈にとどまらず、データの全体像の把握や、特徴的な箇所はどこかという探索、データの潜在構造の発見など、計量的手法の利点を活かした分析が可能かどうか。また③質的データをいかに分析し結論を導いたかという分析プロセスを明示しうるかどうかを検討した。この結果として、いずれの点に関しても一定の水準に達していると判断された。

(2)　分析用ソフトウェア開発

①国際化

分析用ソフトウェア「KH Coder」の国際化を進め、日本語・英語に加えて、中国語・韓国語・ロシア語・カタロニア語・フランス語・ドイツ語・イタリア語・ポルトガル語の文章データを分析する機能を備えた。またソフトウェア上に表示されるメニューやボタンの言語についても、日本語・英語に加えてスペイン語・中国語・韓国語を選択できるようになった。チュートリアル・マニュアルについても日本語のものだけでなく、英語版を整備した。

こうした国際化によって本ソフトウェアが多様な言語データの分析に利用されるようになれば、より多くの研究者から、より多彩な改善要望や問題提起などのフィードバックが得られるようになると期待できる。こうした研究交流がいっそうの進展、活性化につながることを期待したい。

②人間による判断を支援する機能

本ソフトウェアは文章中から自動的に言葉を切り出して分析を行なうが、これは言葉が利用されていた文脈を無視してよいという考えにもとづく機能ではない。それぞれの言葉がもとの文章中でどのように使用されているかを分析者が把握しておくことが、分析結果を理解するためには重要である。

この観点から、抽出語リストであれ、対応分析のような多変量解析であれ、分析結果中の言葉をクリックするだけで、文脈の一覧表示（KWIC コンコーダンス）を実行する機能を備えた。統計分析の結果だけを見て判断するのではなく、分析結果と元の文章とを循環的に行き来しながら分析者が判断を下すことを支援する機能である。

また、こうした元データの確認以外にも、多様な使い方をされる中で、「こうした形の支援が欲しい」というニーズが生じる場合もあるだろう。そこで、ユーザー自身のニーズにもとづいて、ユーザーが独自の機能の追加やカスタマイズを行ないやすい仕組みを準備し、使用方法を書籍にまとめた。

③各分析機能の高度化

さまざまな形で分析機能の高度化を図った。テキスト型データを計量的に分析しようとする際に、障害となりうる主要な問題の 1 つは、同じ言葉であっても使われる場所によって意味が変わることである。この問題があるので、テキストデータから機械的に言葉を切り出して分析することには、一定の限界があると考えられてきた。この問題に対する 1 つの対処として、次のような機能を加えた。言葉の意味が変化した場合には、一緒に出現する言葉（共起語）も変化しがちである。そこで、共起語の変化を描くことを通じて、意味の変化を可視化する機能を備えた。なおこの機能の考案にあたっては、大阪大学大学院人間科学研究科教授・川端亮先生のご示唆を得た。ただし本方法やソフトウェアに何らかの瑕疵があったとすれば、それはすべて筆者の責によるものである。

また、言語学分野における分析の方法に学び、当該分野で多く用いられている統計指標を取り入れた。具体的には、特定の言葉と一緒に使われる（共起する）ことが多い語をリストアップする機能で、共起の程度をあらわす指標として「Mutual Information」「T Score」「Log Likelihood」などを使用できるようにした。

この他、一定の処理時間を要するものの、これまでよりも容量の大きいデータの分析に耐えられるよう見直しを行い、125 万件の文書群を対象として多変量解析を行えることを確認した。

以上の改善に加えて、さまざまな細かな改善を継続的に行なってきた結果、本報告書を執筆している 2018 年 5 月現在、本ソフトウェアを利用した研究事例の数は 2,000 件を数えている。

(3)　上手く活用するための方策

第一に筆者自身がこの方法とソフトウェアを利用した応用研究を示すことで、活用方法を提示した。1 つ 1 つは短い文章を多数集

めたデータの例としてアンケート自由回答データの分析を行なった。また長く続く文章の中で文脈が形作られていくデータの例として、英文の小説「赤毛のアン」の分析を行なった。

自由回答データの分析においては、アンケート中に自由回答型の質問を置くことで、ごく小規模なインタビュー「マイクロ・インタビュー」を実施できること、また、それによってより良いアンケート（質問紙調査）の実施につながりうることを示した。アンケート（質問紙調査）は社会学に限らずさまざまな学問分野で広く利用される調査法であるから、そのより良い実施法を提案することには意義があると考えられる。

長文データ「赤毛のアン」の分析においては、物語の中で言葉の意味が変わっていくことに注目した。テキストデータの計量的な分析においては、同じ言葉であっても文脈によって意味が変わることが、しばしば問題・欠点として指摘される。しかし、そうした意味の変化を計量的に捉えることが可能であり、それによってデータの特徴を明らかにしうることを示した。具体的には、登場人物「マリラ」が変化していく様子を計量的方法で記述した。

第二に、似通った方法を用いた研究事例のレビューを通じて、本方法およびソフトウェアを活用する方法についての検討を行なった。本ソフトウェアを最初に公開したのは2001年で、これを利用した研究は現在2,000件に達している。したがって現在は，ただ応用研究を増やすのではなく，KH Coderがいっそう上手く利用され，優れた応用研究が生み出されることを企図しての努力が重要な段階にあると考えられる。そこで，現在の応用研究を概観的に整理することを通じて，どのように KH Coder を利用すればデータから社会学的意義のある発見を導きやすいのかを探索した。

ここではなるべく優れた応用研究を取り上げて，方法やソフトウェアをどのように利用しているかを記述した。また，なるべく多様なデータを分析対象とした研究を取り上げることで，応用研究を概観することを目指した。以上のような整理をもとに，この方法やソフトウェアを上手く利用するための方策や，今後の展開について論文にまとめた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計8件）

① 樋口耕一、「計量テキスト分析および KH Coder の利用状況と展望（特集号 テキストマイニングをめぐる方法論とメタ方法論）」『社会学評論』、査読あり、68(3)、2017、334-350

② 樋口耕一、「言語研究の分野における KH Coder 活用の可能性」 『計量国語』、査読なし、31(1)、2017、36-45

③ Koichi Higuchi、“A Two-Step Approach to Quantitative Content Analysis: KH Coder Tutorial Using Anne of Green Gables (Part II)” 『立命館産業社会論集』、査読なし、53(1)、2017、137-147

④ 樋口耕一、「フリーソフトウェア「KH Coder」の文章データ分析機能」『統計』、査読なし、68(4)、2017、42-47

⑤ Koichi Higuchi、“A Two-Step Approach to Quantitative Content Analysis: KH Coder Tutorial Using Anne of Green Gables (Part I)” 『立命館産業社会論集』、査読なし、52(3)、2016、77-91

⑥ 星野崇宏・荘島宏二郎・樋口耕一・富田英司、「教育心理学研究のためのテキストデータの計量分析」 『教育心理学年報』、査読なし、55、2016、313-321

⑦ 樋口耕一、「フリーソフトウェア『KH Coder』による計量テキスト分析――手軽なマウス操作による分析からプラグイン作成まで」 『研究報告人文科学とコンピュータ（CH）』、査読なし、2015-CH-107(9)、2015、1-2

⑧ Ozawa, W., Makita, Y., Higuchi, K., Nishimura, K., Ishikawa, K., Ogawa, H. & Kato, H. “The Local Community Volunteer Social Worker System in Japan: Analysis of Survey Data”『立命館産業社会論集』、査読なし、50(3)、2014、1-20

〔学会発表〕（計4件）

① Koichi Higuchi、“Statistical analysis of Japanese textual data using PC: developing free software KH Coder” The 28th European Association of Japanese Resource Specialists Conference（国際学会）、2017

② 樋口耕一、「質的データのより幅広い活用を目指しての方法的検討とソフトウェア開発」、日本教育心理学会第56回総会（招待講演）、2015

③ 樋口耕一、「KH Coder による計量テキスト分析――アンケート自由回答の分析を中心に」、第17回日本水環境学会シンポジウム、2014

④ 樋口耕一・阪口祐介、「現代の高校生を

脱原発へと向かわせるもの」、第 65 回関西社会学会大会、2014

〔図書〕（計 4 件）

① 李在鎬・石黒圭・伊集院郁子・河原大輔・久保圭・小林雄一郎・長谷部陽一郎・樋口耕一、ひつじ書房、『文章を科学する』、2017、197（82-101）

② 友枝敏雄・浜日出夫・山田真茂留・樋口耕一・ほか多数、有斐閣、『社会学の力』、2017、301(246-249)

③ 友枝敏雄・平野孝典・杉村健太・小籔明生・山田真茂留・多田隈翔一・森康司・平松誠・久保田裕之・阪口祐介・樋口耕二、大阪大学出版会、『リスク社会を生きる若者たち――高校生の意識調査から』、2015、245(186-203)

④ 石田基広・神田善伸・樋口耕一・永井達大・鈴木了太、共立出版、『R のパッケージおよびツールの作成と応用（シリーズ Useful R・金明哲編）』、2014、199（73-128）

〔その他〕

分析用ソフトウェア「KH Coder」のホームページ

http://khcoder.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