Commentary

# 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――

樋口　耕一（立命館大学産業社会学部准教授）

## 1　社会調査におけるテキスト型データの利用

経験を積んだ研究者であっても，質問紙調査を計画するときには，どのような質問文を盛り込むべきかと頭を悩ませる場合が多いだろう。このときには，数値データとして入力される選択肢型の設問だけでなく，回答者の考えをそのまま表現してもらう自由回答型の設問にも魅力を感じるかもしれない。限られた選択肢から選ぶことを強制せずに自由に答えてもらえば，調査を行う側が予想していなかった結果を得られることもあるだろう。自由に書き込まれた内容からは，別の設問でなぜ特定の選択肢を選んだのかという考え方の筋道が見えるかもしれない。

ただし，多くの自由回答が得られても，単にそれを目で読むだけでは分析が難しい。データ量が多い場合には，全体にどんな回答が多かったということすら，正確に把握できるとは限らない。まして，学歴・性別・職業などの属性によって回答内容がどう変化したかといったことは，ますます判然としなくなる。こうした問題を解決するために，筆者らは計量テキスト分析という方法を提案してきた（川端，2009；樋口，2006）。これはテキストマイニングと総称される方法と似通ったものであるが，内容分析（content analysis）の考え方に依拠している点に1つの特徴がある。本稿では，拙作のフリーソフトウェアKH Coderを用いて，自由回答項目の計量テキスト分析を行う手順を実践的に紹介したい。

なお，本稿では自由回答項目の分析を例として取り上げているが，それ以外の種々のテキスト型データに対しても同じ手順を適用できる。たとえば新聞記事から社会意識を探索することも考えられるし（樋口，2011），インターネット上にもさまざまなデータが蓄積されつつある。一例として，ネット上でのやりとりの記録から，社会的ネットワークの動態とコミュニケーション内容とを組み合わせて分析することが提案されている（Bohn et al., 2011）。またKH Coderは日本語のほかいくつかの西ヨーロッパ言語の分析を行える。社会調査の領域において計量テキスト分析を適用しうる対象は，種類・量ともに拡大しつつあると言えるだろう。

## 2　自由回答データ分析の目的と準備

ここからは「高校生とその母親の教育意識に関する全国調査」（川端，2002）に含まれる自由回答項目を用いて，分析の手順を紹介する。この調査では母親に対して，もしも自分の子どもが特に勉強したいこともなしに「なんとなく進学する」ことになった場合，賛成できるかどうかをたずねている。さらに，その賛否の理由を自由回答として記入するよう依頼している。なお，「なんとなく進学」への賛否と母親自身の学歴とを見ると，学歴が高い方が賛成する傾向があった。

それでは，高い学歴をもつ母親は，自身の経験からどのような理由づけを得て，「なんとなく進学」に賛成するようになったのだろうか。賛否の根拠となる語彙や論理が，学歴の高低によってどのように変化したために，賛否の違いが生じたのだろうか。もちろんこうした問題については，先行研究にもとづく理論的解釈や，利用できる意識項目を用いた構造方程式モデリング等による探求を行うべきである。それらを最大限に行ったうえで，さらに自由回答項目があれば可能な分析として，以下の例をご覧いただきたい。

まず分析を始めるための準備として，自由回答データを整理したうえでKH Coderに登録する。

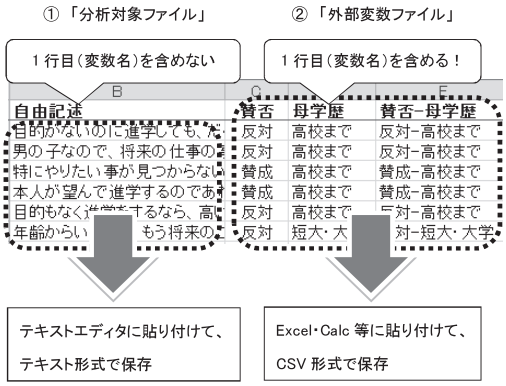



図 1　データファイル 2 種類の準備

表 1　頻出語リスト（頻出 30 語）

| 抽出語 | 頻度 | 抽出語 | 頻度 | 抽出語 | 頻度 |
|---|---|---|---|---|---|
| 進学 | 345 | 賛成 | 57 | 見つかる | 36 |
| 思う | 278 | 今 | 53 | 道 | 36 |
| 自分 | 161 | 生活 | 53 | 学ぶ | 34 |
| 目的 | 115 | 行く | 52 | 人生 | 31 |
| 大学 | 102 | 就職 | 51 | 仕事 | 30 |
| 勉強 | 100 | 時間 | 50 | 進む | 30 |
| 考える | 95 | 持つ | 49 | 人 | 30 |
| 目標 | 94 | お金 | 47 | 良い | 28 |
| 将来 | 83 | 決める | 45 | 社会 | 27 |
| 見つける | 76 | 学校 | 38 | 必要 | 26 |

そのための手順として，**図 1** に示すように「分析対象ファイル」と「外部変数ファイル」の 2 種類を作成する必要がある。この際には，自由回答が記入されていないケース（行）はあらかじめすべて削除しておくとよい。ファイルの準備ができたら KH Coder を起動し，メニューから「プロジェクト」「新規」とたどって，分析対象ファイルを指定する。次に「前処理の実行」をしてから，外部変数ファイルの「読み込み」を行う。以上の操作によって，文章中から自動的に語を取り出す処理が行われ，分析の準備が完了する。

ここまでの操作の結果を確認するとともに，本格的な分析に入る前の準備として，「抽出語リスト」（表 1）を確認しておくとよいだろう。**表 1** に示したのはわずか上位 30 語だが，それでも「進学」する「目的」「目標」について，また「将来」の「就職」「仕事」「人生」について，また「時間」「お金」といったコストについての言及が多かったことをうかがえる。その他にも，子どもへの期待をあらわす語であろう「見つける」や「見つかる」が見られる。

以上の手順の留意点として，もしも前処理やデータの読み込み時にエラーが表示される場合は，「分析対象ファイルのチェック」を行うと原因について手がかりが得られるかもしれない。また，Excel を使ってデータ入力・作成を行った場合は，セル内に改行文字が含まれていないかどうか確認しておくとよいだろう。Excel では Clean 関数を使うと改行文字を一括して取り除くことができる。

## 3　自由回答に多く出現した主題

次に，どの語とどの語が一緒に使われていたのかという共起に注目することで，データ中にどのような主題が多く出現していたのかを探る。共起する語を線で結んだネットワークを描き（Osgood, 1959），ネットワーク分析の手法を用いることで（Danowski, 1993），データ中の主題を効率的に探索することができる。

そのための手順として，KH Coder の「共起ネットワーク」コマンドを利用する。オプション画面にはさまざまな選択肢が並んでいるが，単に「OK」をクリックすれば**図 2** が得られる。ただしデフォルトの設定では，もっとも強い共起関係から順に 60 が辺（edges）として描画されるが，図 2 ではこの描画数を 80 に増やしている。また，お互いに強く結びついている部分を取り出してグループ分けした結果が，KH Coder の画面上では色分けによって示されているので，図 2 では色分けの代わりに点線で区分している。

得られた結果である図 2 の右上部分を見ると，前節で確認した「お金」と「時間」がどちらも「無駄」と一緒に用いられていたことが分かる。他にも比較的大きなグループを見ていくと，中央上部には「進学」「大学」「生活」など，賛否の理由を述べるのに一般的に用いられた語が集まっている。その下には「人」との「出会い」や「いろいろ」「学ぶ」「経験」など，「なんとなく進学」に賛成する理由を表現したと見られる語が集まっている。次に左上の部分からは，総じて「進路」

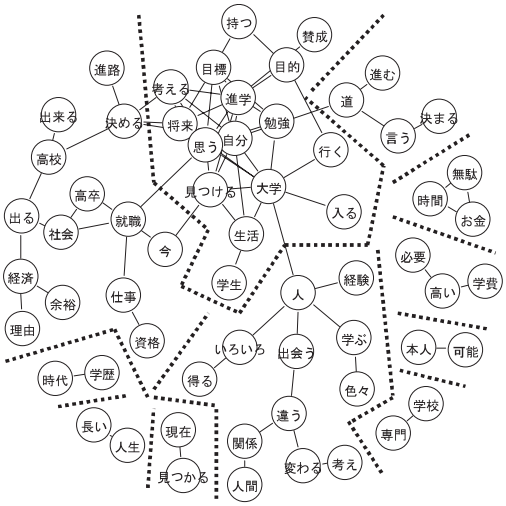



図２　頻出語の共起ネットワーク

「就職」に関わる記述の様子がうかがえる。具体的な回答としては「高卒では，就職も期待できない」や「仕事をして自立してもらったほうが経済的にも助かります」などがあった。「仕事」や「就職」については，賛成の理由の中でも，反対の理由の中でも言及されていたことが分かる。

ここでの留意点として，分析結果の解釈を行う際には，もとのテキストにあたっての確認を欠かさずに行う必要がある。分析結果の中に，もとの文章中でどのように使われていたのか分かりにくい語や，気になる語があれば「KWIC コンコーダンス」コマンド（**図 3**）で１つずつ確認していくとよいだろう。その他，共起ネットワーク作成時に語の配置を決める方法や，グループ分けの方法について詳しくは KH Coder のマニュアルを参照されたい。また，語の共起関係を探索する方法としては，共起ネットワークに限らず，クラスター分析や多次元尺度構成法の機能も KH Coder に備えている。こうした，より広く用いられ，素性がよく知られている方法を用いるのも１つの選択であろう。

## 4　自由回答と他の変数の関連

いよいよ，「なんとなく進学」への賛否によって，また学歴の違いによって，自由回答項目に記入された理由がいかに異なっているかを探索する。

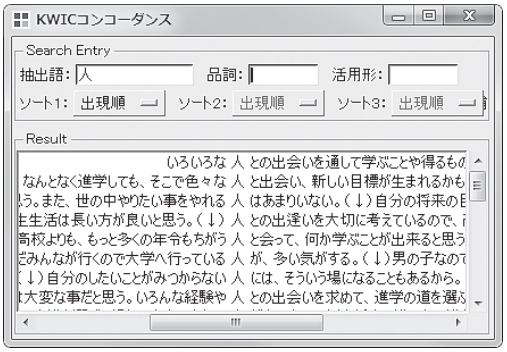



図３「人」という語の使われ方

このように，自由回答の内容が他の変数といかに関連しているかを見る方法の１つとして，対応分析がある。

対応分析を行うための手順として，もし **2** 節で読み込んだ外部変数ファイルに欠損値が含まれる場合には，あらかじめ欠損値の設定が必要である。KH Coder のメニューから「ツール」「外部変数と見出し」「リスト」をクリックする。そして，たとえば「9」や「99」が欠損値を表している場合には，それらの値のラベルとして半角ピリオド「.」を入力しておく。

次に KH Coder の「対応分析」画面で，「分析に使用するデータ表の種類」として「抽出語 x 外部変数」を選ぶ。そして，分析に用いる外部変数を選んで「OK」をクリックすれば**図 4** が得られる。図４では「なんとなく進学」への賛否と，母親学歴を組み合わせた変数を分析に用いた。これは「賛成－高校まで」「賛成－短大・大学」「反対－高校まで」「反対－短大・大学」という４種類の値を持つ変数である。なお図４の点線については，筆者が解釈の際に注目した語を「手作業」で強調したものである。

得られた結果である図４を見ると，長方形で囲んである外部変数の布置から，成分１（x 軸）に賛否の違いがあらわれ，成分２（y 軸）に学歴の違いがあらわれたことが分かる。各成分の寄与率から，学歴よりも，賛否による回答内容の違いの方が顕著であったようだ。そして，「なんとなく進学」に反対した人については，学歴間のプロットの位置が比較的近いのに対して，賛成した人では「高校まで」と「短大・大学」とが比較的大きく上下に離れている。ここから，賛成した人の場

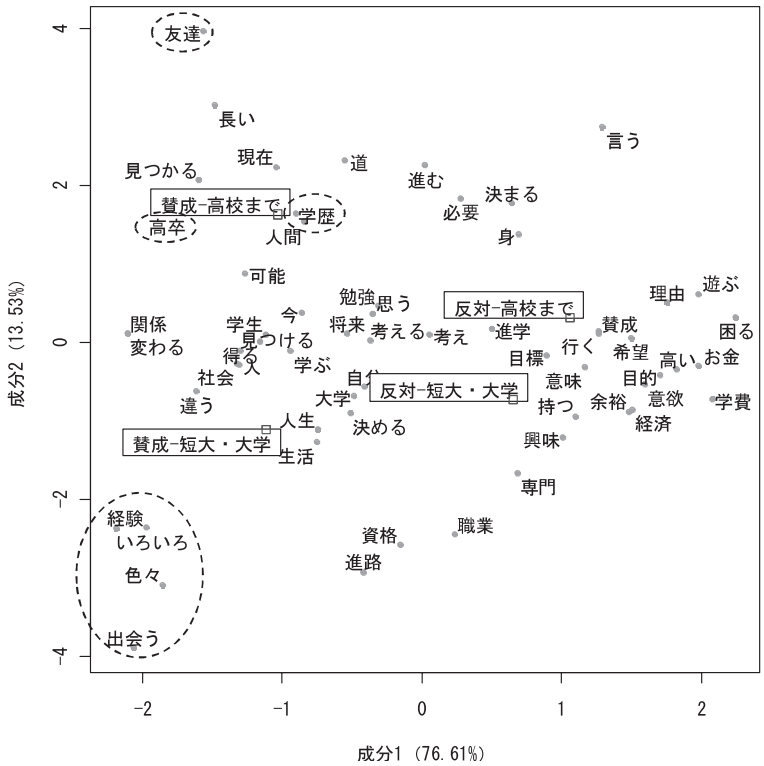


図４　賛否と学歴による自由回答の変化

合には学歴による理由の違いが比較的大きかったのに対して，反対した人ではそれほど学歴による違いが見られなかったことが分かる。

それでは賛成した人の場合，学歴によるどのような理由の違いがあったのだろうか。まず高校までの学歴をもつ人については，特徴的な語として「高卒」と「学歴」があり，低い学歴で社会に出ることへの危惧がうかがえる。また最上部に布置されていることから「友達」も非常に特徴的な語であり，実際の回答には「友達との出会いがあるので」といった表現があった。

それに対して大卒学歴をもつ母親，すなわち自らも大学生活を経験した母親の賛成理由には，「いろいろ」「経験」「出会う」などの語が特徴的である。実際の回答には「色々な経験をする中で自分の道を見つけてくれれば」「色々な人と出会い，新しい目標が生まれる」といった表現があった。必ずしも同世代の「友達」作りに限定されない，多様な経験を通じて，自分の道を見出していく過程への評価ないしは期待を読み取れる。多様な経験を得る過程の中で，何か１つでも強く触発されるものがあれば，進路を選ぶ手がかりになるかもしれない。大学生活の経験をもつ人の場合，この過程への評価や期待から「なんとなく進学」に賛成する傾向があったことがうかがえる。以上が本稿の分析例における主な分析結果である。

分析全体を通じての留意点として，本稿では図２・図４などの分析結果について部分的にしか解釈していないが，これは必ずしも紙幅の都合だけが理由ではない。これらの分析では，自動抽出された語の中で出現回数の多いものを機械的に選んで分析している。この方法には，恣意的になり得る作業を含まないという利点があるのだが，一方で，分析結果として意味がない語，あるいは分析者の問題意識と直接関係ない語も分析に含まれてしまう。たとえば，因子分析に耐えるよう設計・確立された選択肢型質問項目のセットであれば，分析モデルに含めたすべての項目について結果を確認・解釈することに意味があるだろう。しかし，自由回答から機械的に取り出した語の集まりでは

そうはいかない。よって，解釈できる箇所をなるべく多く見つけたうえで，分析結果として得られた図表全体の中で，どの部分に注目し，いかに解釈したのかを明確に記述するのが穏当な手順であろう。これによって，分析者による解釈の妥当性や，分析者とは異なる解釈の可能性について，第三者の判断が可能な状態にしておくことが，現在の技術的状況では最善の方法と考えられる。

## 5 さらに分析を進めるために

以上のような，自動抽出した語による分析は手軽に行える一方で，限界があることも否めない。たとえば「経験」という概念は，1つの語「経験」だけであらわされるわけではない。「体験」や「自分の目で見る」についても，同じ「経験」という概念として数えるべきかもしれない。あるいは「経験」という同じ語であっても，本稿の分析の場合には，「社会に出て経験を積む」は除外すべきかもしれない。こうした操作をコーディングと呼ぶ。自動抽出された語による分析を行ったうえで，必要に応じてKH Coderによるコーディングを行うという2段階での分析を筆者は提案している。

またKH Coderによる分析はすべてマウス操作で行えるが，本誌の読者には手慣れた方も多いであろうから，マウス操作では融通がきかないという不満が生じるかもしれない。その場合には図2・図4を「R Source」形式で保存すれば，これらのプロットを作成するためのRのコマンドが得られる。KH Coderによる多変量解析やプロット作成はすべてRを用いて行っているので，Rコマンドを編集すればいかなるカスタマイズも可能である。あるいは解析用のデータをCSV形式で取り出して，R以外の任意の統計ソフトウェアを使うこともできる。またKH Coderはすべてのデータを MySQLデータベースに保存しているので，直接MySQLにアクセスすれば任意の検索やデータ抽出を行える。

KH Coderではこうした柔軟性を重視しており，お仕着せの分析しかできない既製品ではなく，利用者による創意工夫を発揮できる分析環境の製作を，目標の1つとしている。利用して下さる方の必要に迫られての工夫があってこそ，分析方法も発展していくものと考えるからである。

［付記］「高校生とその母親の教育意識に関する全国調査」（川端，2002）のデータ利用にあたっては，調査代表者ならびに大阪大学大学院人間科学研究科社会環境学講座経験社会学研究室の許可を得ました。記して感謝いたします。

文献

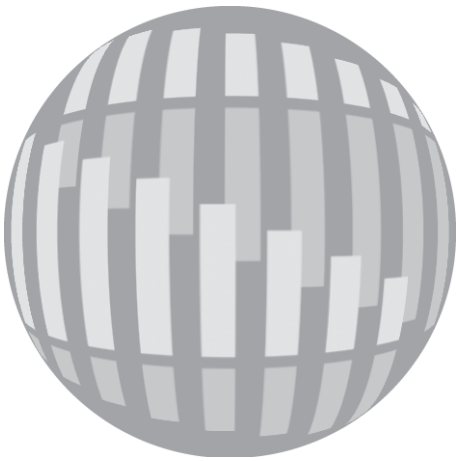
Bohn, A., I. Feinerer, K. Hornik and P. Mair, 2011, “Content-Based Social Network Analysis of Mailing Lists,” *The R Journal*, 3 (1): 11–18.

Danowski, J. A., 1993, “Network Analysis of Message Content,” in W. D. Richards Jr. and G. A. Barnett eds., *Progress in Communication Sciences, XII*, Norwood, NJ: Ablex, 197–221.

樋口耕一，2006，「内容分析から計量テキスト分析へ――継承と発展を目指して」『大阪大学大学院人間科学研究科紀要』32：1–27。

――――，2011，「現代における全国紙の内容分析の有効性――社会意識の探索はどこまで可能か」『行動計量学』38（1）：1–12。

川端亮，2002，「高校生とその母親の教育意識に関する全国調査」SRDQ事務局編『SRDQ――質問紙法にもとづく社会調査データベース』（http://srdq.hus.osaka-u.ac.jp/2011年11月18日アクセス）。

――――，2009，「質的データのコンピュータ・コーディング」谷富夫・芦田徹郎編『よくわかる質的社会調査 技法編』ミネルヴァ書房，134–47。

Osgood, C. E., 1959, “The Representational Model and Relevant Research Methods,” I. de S. Pool ed., *Trends in Content Analysis*, Urbana, IL: University of Illinois Press, 33–88.