原著論文

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―

樋口 耕一
（大阪大学）

【要旨】

新聞記事や質問紙調査における自由回答など、社会調査において計量的な分析の対象となるテキスト型データには、様々なものが挙げられる。これらのテキスト型データを計量的に分析する際、従来は Correlational アプローチか Dictionary-based アプローチのうち、いずれかが用いられることが多かった。前者は多変量解析の応用、例えば、クラスター分析を用いて頻繁に同じ文書の中にあらわれる言葉のグループを見つけだすといった方法で、データ中の主題を探索するアプローチである。それに対して後者のアプローチでは、分析者の指定した基準にそって言葉や文書が分類され、計量的な分析が行われる。本稿ではこれらのアプローチを検討し、それぞれに一長一短を持つこれら 2 つを、互いに補い合う形で統合したアプローチを提案する。そして、その実現に必要なシステムを作製・公開するとともに、本アプローチ・システムを用いて自由回答データの分析を行った例を示す。その上で、従来のアプローチに対する本アプローチの有効性について若干の検討を加える。

キーワード： 内容分析、テキストマイニング、質的データ、 Correlational アプローチ、Dictionary-based アプローチ

1. はじめに

社会調査において計量的な分析の対象となるテキスト型（文章型）データには、実に様々なものがある。例えば、新聞記事はコミュニケーション研究において古くから取り上げられ、世論や価値についての分析が行われてきた（Woodward 1934, Lasswell 1941）。また質問紙調査では、網羅的で完全な選択肢を提示することが難しいという選択型の設問が持つ問題（安田 1970）を、自由回答型の設問で補うことができる。

現在、新聞記事をはじめとする各種データベースの整備やインターネット調査の方法論開発などによって、これらテキスト型データの収集こそ容易になりつつあるものの、その分析については未だ決して容易とは言い難い。というのも、テキスト型データのような質的データを計量的に分析するためには、統計処理が可能となるように、何らかの形でデータをコード化せねばならない点がまず問題になるからだ。例えば、自由回答項目にどんな回答がいくつ記述されたのかを調べるためには、何らかの基準にそって回答を分類するというコーディング作業が必要になる。

この作業を含めて、テキスト型データの分析にコンピュータを利用する利点として、Seale（2000）は大量のデータを扱えること、信頼性があること、共同研究が可能となること、サンプルの選択に役立つことの 4 つを挙げている。すなわち、コード化作業をコンピュータによって自動化することで、大量のデータであってもコード化の基準に揺らぎや恣意性が生じない。そして、その堅固な基準を第三者に明示できるため、検証や検討、ひいては共同研究も可能になる。また、大量のデータの中から、ある条件を満たすデータを発見することもコンピュータの得意とするところであり、これによって、データの典型例や特異例を抽出することが容易になる。

もっとも、コンピュータを利用することで問題が全て解決するわけではない。むしろ、コンピュータやそれを用いた自然言語処理・統計解析が進歩するにつれて、分析プロセスのどの程度までをコンピュータに任せてしまうことが可能なのか、あるいは好ましいのかといった、別種の問題が生じている。例えば、不要な広告メールを選り分けるために、届いた電子メールの内容を「分析」するという場合、現在ではほぼ完全な自動化が可能である[1)]。しかし、社会学者が抱くであろう様々な理論仮説や問題意識にそった分析の場合はどうだろうか。自動要約ではなく分析を目指す以上、この場合は自動化にも限度があり、テキスト型データに含まれる様々な側面の中から、分析者が自ら特定の側面に焦点をあてる必要があるだろう。本稿ではこのような問題意識のもとに、半世紀以上にわたって研究がなされてきた内容分析（content analysis）の方法論に特に依拠しつつ、テキスト型データを計量的に分析する新たなアプローチを提案したい。

2. 新たな計量的分析アプローチの提案

2.1 先行研究 ―2 つのアプローチ―

分析プロセスのどの程度までを自動化すべきかという問題は、決して新しい問題ではない。というのも、コンピュータを用いたテキスト型データの計量的分析は、内容分析の一手法として英語圏では非常に早くから行われており、1960 年代の後半には既に、2 つの異なるアプローチが登場している。1 つは、分析者が作成した基準にしたがって言葉や文書を分類するためにコンピュータを用いるアプローチである。もう 1 つは、頻繁に同じ文書の中にあらわれる言葉のグループや、あるいは、共通する言葉を多く含む文書のグループを、多変量解析によって自動的に発見・分類するためにコンピュータを用いるアプローチである。これらのアプローチはいずれもコンピュータを利用して言葉や文書を分類し、その結果を計量的に分析するという点では似通っているが、分類の基準を分析者が自ら指定するのか、それとも多変量解析に分類を全面的に任せてしまうのかという点で大きく異なっている。

前者のアプローチは現在 Dictionary-based アプローチと呼ばれており、ここでは、分類基準を作成することで分析者の持つ理論や問題意識を操作化することが目指された（Osgood et al. 1957）。それに対して、現在 Correlational アプローチと呼ばれている後者のアプローチでは、分類をコンピュータに任せてしまうことで、分析者の理論仮説や問題意識によって「汚染（contaminate）」されていない状態で、データを分析することが目指された（Iker & Harway

1969）。1967 年にフィラデルフィアで開催された内容分析に関する会議（Annenberg conference）では、これら 2 つのアプローチの間に際だった乖離が見られたという（Stone 1997）。以下に述べるように、これら 2 つのアプローチはそれぞれに独自の発展をとげており、その結果として 4 半世紀を経た現在でも、両者の間に著しい乖離を見て取ることができる。

まず、分析者の作成した基準にそって言葉や文書を分類する Dictionary-based アプローチでは、例えば「aimless、anarchy、chaos などの言葉をアノミーの現れと見なす」といったコーディングの規則（dictionary）が作成された。当初の研究では、このような規則を数多く作成してコーディングを行い、アノミーのような概念がデータ中に出現した回数を数え上げるという分析が行われた。このように概念の出現頻度に注目する方法は Thematic text analysis と呼ばれ、現在では、概念間の関係を詳細に分析する Semantic text analysis や、とりわけ多数の概念間の関係に注目する Network text analysis といった手法が、適宜組み合わせて用いられるようになっている。また、このアプローチにおける重要な進歩をもう 1 つ挙げるとすれば、コーディング規則の複雑化である。上述のアノミーの例では、単に 1 つ 1 つの言葉を何らかの概念に結びつけていたわけだが、これでは「デフレ対策の会議」も「井戸端の会議」も、同じ「会議」と見なしてしまうことになる。現在では、こういった言葉の曖昧さの問題に対処するために、例えば「同じ文の中に井戸端という言葉が無ければ」という条件を追加するなど、より複雑なコーディング規則による"disambiguation"が図られている（Popping 2000, Roberts ed. 1997）。

次に、分析者がコーディング規則を作成するのではなく、多変量解析を用いて言葉や文書を分類しようとするCorrelationalアプローチでは、データから自動的に言葉が切り出され、その結果を用いて因子分析やクラスター分析などの多変量解析が行われた。このアプローチの進歩は、社会調査というよりも、純粋な情報処理の分野における研究成果に負うところが大きい。具体的には、文書の自動分類や自動要約、効率的な検索などを行うための研究が盛んに行われており、近年ではこれらの方法と、これらの方法を用いた探索的解析のことを、総じてテキストマイニングと呼ぶこともある（Renz & Franke 2003, 那須川ほか 2001）。そして、このテキストマイニングを社会調査に活用するための研究としては、テキスト型データの収集法まで含めて見直しを行い、通常の選択肢型項目との併用など、質問紙調査における自由回答項目の分析法・活用法を追究した大隅・Lebart（2000）の研究が挙げられる[2)]。

2.2 統合アプローチの提案

現在でも著しく乖離しているとはいえ、これら 2 つのアプローチは根本的に異なるものというよりも、むしろそれぞれに一長一短があり、互いに補い合うべきアプローチと見なすことができよう。まず Dictionary-based アプローチの利点としては、コーディング規則を作成することで、分析者の理論や問題意識を自由に操作化し、テキスト型データの様々な側面に自由に焦点を絞れるということが挙げられる。その反面、かつて Berelson（1952）が戒めたように、意図的ないしは無意識のうちに、理論や仮説に都合の良いコーディング規則ばかりが作成・利用されてしまう危険性も完全には否定できない。

この客観性に関わる問題は、多変量解析によってデータを要約する Correlational アプロー

チを併用することで、補うことができる。すなわち、多変量解析によってデータ全体を要約・提示した上で、コーディング規則を公開するという手順を踏めば、データ全体の中から、どの部分、あるいはどの側面がコーディング規則によって取り上げられたのかを、第三者が把握できるようになる。これによって、研究手法が批判や検討・検証に耐えるオープンさを有しているという意味での客観性（Phillips 1990）を、大きく向上させることができよう。

また、一方の Correlational アプローチでは、多変量解析に大きく依存する以上、理論や問題意識を自由に操作化し追究することは困難である。なぜなら、研究者が抱きうるあらゆる理論・仮説にもとづいた分類を、多変量解析によって自動化できるとは考えにくいからだ。この点についても、2 つのアプローチを併用すれば、Dictionary-based アプローチの利点によって補われることとなる。そこで、本稿では次の 2 段階からなる統合アプローチを提案したい。

段階 1: Correlational アプローチに倣い、多変量解析を用いることで、分析者の持つ理論や問題意識の影響を極力受けない形で、データを要約・提示する。

段階 2: Dictionary-based アプローチに倣い、コーディング規則を作成することで、理論仮説の検証や問題意識の追究を行う。

3. 分析用システムの作製指針

従来の研究において開発されてきた分析用システムとしては、以下のようなものが挙げられる。まず、Dictionary-basedアプローチにおいては、当初のGeneral Inquirer（Stone et al. 1966）に始まり、日本語を扱うことができるシステムとしてAutocode（佐藤 1992, 田中・太郎丸 1996）が現在無償で公開されている[3)]。次に、Correlationalアプローチにおいては、当初のものとしてWORDS（Iker & Harway 1969）があり、日本語を扱えるものとしてはWordMiner（大隅・Lebart 2000）が現在販売されている。その他にも、データから自動的に言葉を切り出し、量的な分析の準備をするための汎用的なツールとしてKT2 システム（谷口 1999）が無償公開されており、さらに日本語に対応したテキストマイニング・システムとしても様々なものが販売されている。

しかし、これら既存のシステムを用いて、本稿で提案する分析アプローチを実現することには難点も多い。第一に、唯一コーディング規則を扱うことができる Autocode にしても、単純な文字列によるコーディング規則の指定しかできないという問題がある。この点は高橋（2000）に倣い、あらかじめ自動的に言葉を切り出しておくことで、例えば「父」という言葉を探そうとして「秩父」という地名を見つけてしまうようなことが起こらないシステム、すなわち「父」と「秩父」を自動的に区別できるシステムの作成が望ましい。第二に、既存のシステムを利用する場合、多変量解析によってデータを要約する段階と、コーディングを行う段階とで、それぞれ異なるシステムを用いなければならない。しかし、言葉の切り出し方が違うような、基本的な仕様に違いがあるシステムを併用することは困難である。

よって、本稿で提案する分析アプローチに適した分析システムとしてKH Coderを作製し、これをフリー・ソフトウェアとして公開する[4)]。以下にKH Coderの主な機能とその設計指針

について述べる。

3.1 多変量解析によるデータ要約のための機能

まず、多変量解析によるデータの要約を行うためには、それぞれの文書や回答の中にどんな言葉が何回出現していたのかを調べ上げねばならない。そして、その結果を表 1-（b）に示す形に整理すれば、因子分析やクラスター分析など各種の多変量解析が可能になる。よって、表 1-（a）に示す形のデータを、自動的に表 1-（b）に示す形に整理する機能をKH Coderに備えた[5)]。これによって、多変量解析によるデータの要約を行う段階では、SASやSPSSなどの統計システムで解析可能な形にテキスト型データを変形するという、いわば橋渡しの役割をKH Coderが果たすことになる[6)]。

表 1　多変量解析のためのデータ整理

（a）　素データ

文書 1　データが多い
文書 2　データを読むのが難しい
文書 3　データの把握も難しい

（b）　語の出現数を整理

	データ	難しい	読む	把握	多い
文書 1	1	0	0	0	1
文書 2	1	1	1	0	0
文書 3	1	1	0	1	0

このデータ整理の際に、出現していた語を全て用いると、語の種類が数千・数万を越えてしまい、解析が難しくなる場合も多い。そこで、多変量解析に用いる語の数をコントロールできるように、以下のような機能をKH Coderに加えた。まず、KH Coderは助詞・助動詞などを省いてデータ中から語を切り出すので、例えば「データが多い」という文からは、「データ」と「多い」の 2 語が切り出される。また、活用を持つ語は基本形に直して抽出されるので、例えば、「多ければ」「多くて」「多い」「多かった」といった記述がデータ中にあった場合、KH Coderは「多い」という語が 4 回出現していたものと見なす[7)]。さらに、一定の回数以上出現している語だけを、統計システムに渡すデータに含めたり、分析の目的に応じて名詞だけ、あるいは形容詞だけをデータに含めたりといった品詞による語の選択も可能である。同様に、平仮名だけからなる語には一般的なものが多く、分析では扱いにくいことも多いので、名詞や動詞であっても平仮名だけからなる語は省くといった選択も可能である。

KH Coder では、このような方法で語を選択することは容易だが、例えばデータに含める語を逐一選択するといった「手作業」は、非常に困難な仕様とした。この仕様は、多変量解析によるデータの要約を行う段階では、恣意的なものとなりうる「手作業」を廃することで、分析者の持つ理論や問題意識によるバイアスを極力排除するための仕様である。

3.2 コーディング規則を扱う機能

次に、KH Coder には当然、「aimless、anarchy、chaos などの言葉をアノミーの現れと見なす」といったコーディング規則を扱うための機能が備わっている。KH Coder が扱うコーディング規則とは、例えば次のようなものである。

＊仕事

仕事 or 会社 or (会議 and not 井戸端)

この例では、「仕事という語が含まれるか、あるいは会社という語が含まれるか、あるいは会議という語が含まれていてなおかつ井戸端という語が含まれない」という条件を満たす文書に「＊仕事」というコードが与えられる。KH Coderでは、この例で示したような抽出語の有無だけでなく、表 2 に示すような条件を自由に組み合わせて指定することができる。また、複数の条件を組み合わせる際には、上に挙げた例のように、and、or、and not、or notなどの論理演算子を利用できる[8)]。

表 2　コーディング規則における条件指定

指定できる条件	条件の具体例
抽出語の有無	「仕事」という語が出現していれば
抽出語の出現数	「仕事」という語が 3 回以上出現していれば
他に作成したコード	「＊交渉」または「＊相談」というコードが与えられていれば
外部変数	（データが自由回答の場合）女性の回答であれば
文書の長さ	（データが自由回答の場合）1 語のみからなる回答であれば
文字列	（基本形に直された抽出語ではなく）「多ければ」という文字列が出現していれば

いったんコーディング規則を作成すれば、そこから先はKH Coderがコーディングを行い、表 1-（b）のような形にコーディング結果を整理して出力する。表 1-（b）には「データ」「難しい」といった語の出現数が記述されているのに対して、ここでは「＊仕事」「＊アノミー」といったコードの有無が 1-0 の 2 値変数で記述されることになる[9)]。

上述の仕組みによって、様々な理論仮説や問題意識に応じた柔軟なコーディングを、再現可能な形で行えるようになる。さらに、通常のプログラミング言語を用いてコーディングを行った場合のことを考えれば、記述されたコーディング規則も格段に短くて読みやすい、理解しやすいものとなる。これによって、コーディング規則を開示した際に、第三者がそれを理解し、検討・検証することが容易になる。

3.3 データ検索の機能

データを要約する段階では、抽出された語が、元のテキストデータ中でいかに用いられているのかを容易に確認できることが好ましい。またコーディング規則の作成に際しては、指定した条件によって妥当なコーディングが行われているか否かを確認する必要が生じる。さらに、計量分析によってえられた知見が顕現している事例を、適切に選択・引用するためにも検索機能が役立つ。これらの点で、データ検索の機能は分析プロセス全般において利用されるべき重要な機能である。よって KH Coder には、いかなる語が抽出されているかを検索する機能や、元のテキストデータ中で抽出語が用いられている文脈を確認するためのコンコ

ーダンス機能（図 1)、また、特定のコードが与えられた文書、あるいは複数のコードが共に与えられた文書を検索するための機能などを備えた。

図 1　コンコーダンス機能

なお、必要に応じて SAS や SPSS などの統計システムを併用しつつ、以上に述べてきた KHCoder の機能を用いて、本稿で提案するアプローチにそった分析を行う手順の概略を図 2 に示した。

4. 分析例

本稿で提案するアプローチにそったテキスト型データの分析を、KH Coder を用いて行った例として川端・樋口（2003）の研究がある。これは、インターネットや情報技術のことを考えるときに思い浮かぶ事柄を 3 つまでたずねた自由回答項目を分析することで、インターネットに対する人々の意識を明らかにした探索的な研究である。方法論的な注釈を加えつつ川端・樋口（2003）が行った分析の概略を示すことで、分析の事例としたい。

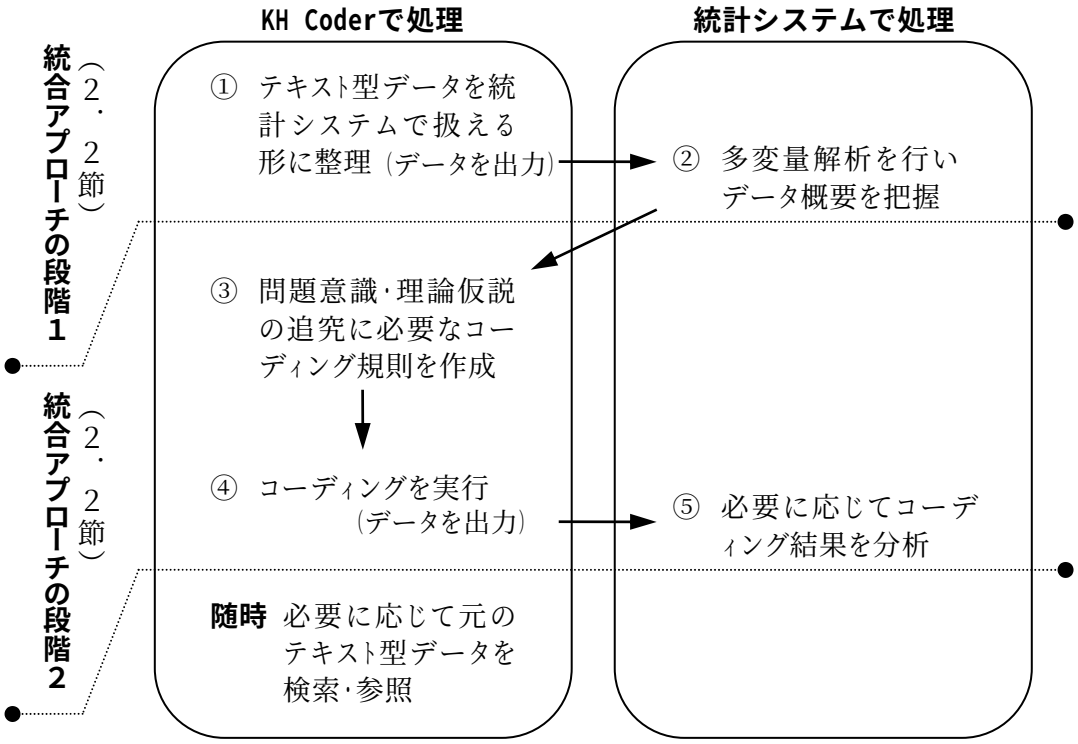
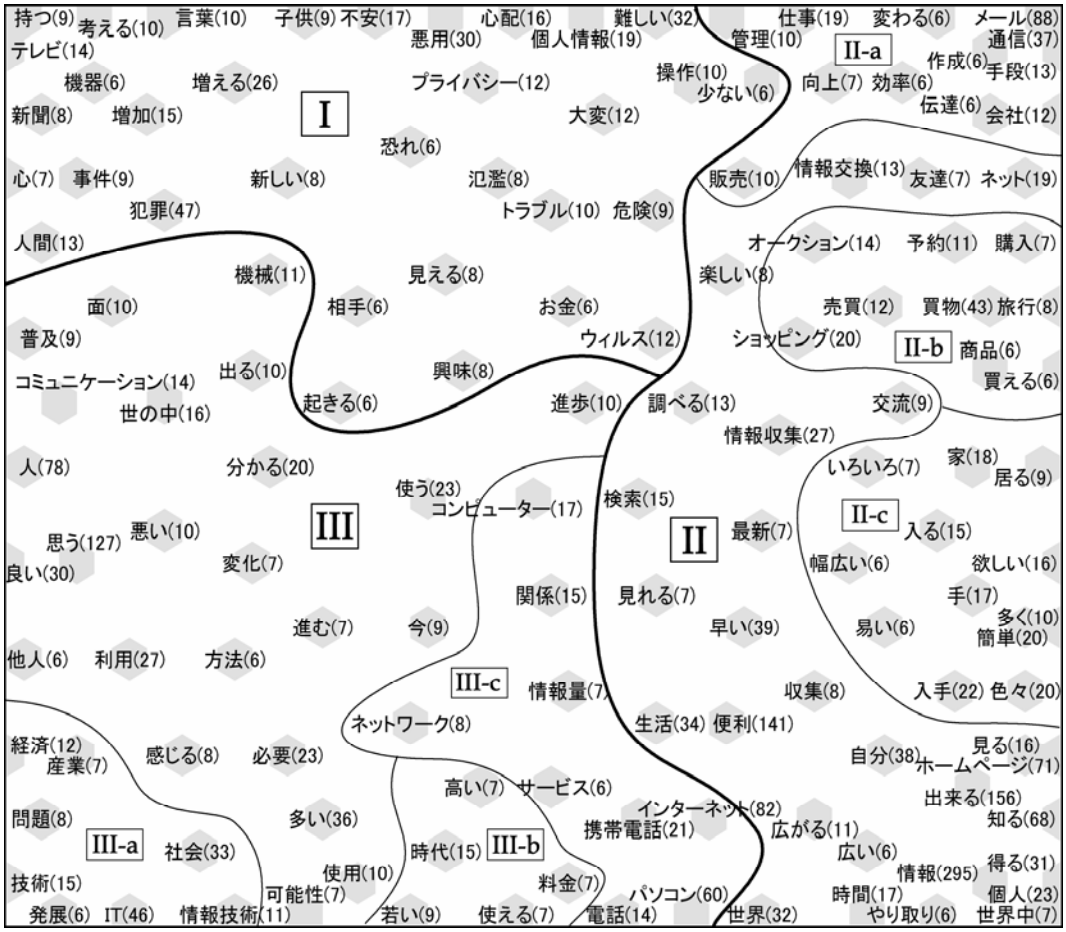


図 2　分析手順の概略

4.1　データ概要の把握と提示

2.2 節で述べた統合アプローチの段階 1 にあたる手順として、データ概要を把握・提示するために、頻出していた語を用いて自己組織化マップ[10]を作成した（図 3）。この結果からは、いかなる言葉が回答中に多く見られたのかということをまず確認することができる。また、近くに布置されている言葉の組み合わせを見ることで、いかなる言葉が似通った文脈で使われていたのかを読みとることができる。図 3 を作成するにあたっては、KH Coderを用いてテキスト型データを表 1-（b）に示したような形、すなわち統計システムで扱える形に整理し、統計システム上で自己組織化マップの作成を行った。したがって、特定の語を選んだりといった「手作業」を一切まじえずに、語の布置が行われている。

ただし図 3 における図中の線については、似通った主題をあらわす語が集まったと思われる部分を、分析者の手によって区切ったものである。この際には、必要に応じて図 1 に示した検索機能を用いることで、元のテキストを参照しながら図の解釈を行った。よって図中の線は、分析者らの解釈の過程をあらわすものと言える。このような図 3 の解釈から、自由回答データは大きく分けて（I）悪用や犯罪への不安、（II）便利さへの評価・期待、（III）社会・経済が変化しつつあるという漠然とした印象ないし変化の予感という 3 つの主題から成っていたことが分かった。



884 nodes, 31(columns) x 29(rows). Quantization error = .01543

カッコ内は語の出現数。線引きは筆者による。（出典：川端・樋口 2003）

図 3　頻出 140 語のマッピング（Self-Organizing Map）

4.2　問題意識の追究

次に、2.2 節で述べた統合アプローチの段階 2 にあたる手順として、図 2 から発見された 3 つの主題をさらに詳しく追究するために、コーディング規則が作成され、探索的に分析が進められた。例えば同じ（I）悪用や犯罪への不安にしても、漠然とした犯罪の増加への不安なのか､それとも個人情報の漏洩などへのより具体的な不安なのかを区別するといったように、この段階では分析者らの問題意識を反映させつつ、コーディング規則が作成された。

この結果、例えば同じ（I）悪用や犯罪への不安であっても、3 つ用意された回答欄の中で増減を示すコードと、そうでないコードが存在することが分かった。さらに、増減を示すコードに注目した結果、1 つ目の回答欄でもっとも出現率が高く、2 つ目、3 つ目と出現率が減少するコードには、（II）便利さへの評価・期待をあらわすものが多く、逆に増加するものには（I）悪用や犯罪への不安をあらわすものが多いことが分かった（表 3)。ここから、インターネットや情報技術のことを考えたとき、まず始めに想起されるのは便利さへの評価・期待であり、その後に文字通りの"second thought"として、悪用や犯罪への不安が想起される傾向が発見されている（川端・樋口 2003)。

表 3　回答欄によるコード出現率の変化

コード名	コードが与えられた回答の例	回答欄ごとの出現率		
		1 つ目	2 つ目	3 つ目
パソコン	「パソコン」	6.60%	3.10%	3.81%
便利さ	「便利」	16.13%	6.01%	6.09%
情報収集	「分からない事を調べる」	18.91%	9.69%	7.11%
スピード	「速い」「早い」	7.48%	4.65%	3.05%
犯罪・トラブル・悪用	「犯罪」「ハッカー」	3.96%	8.33%	14.21%
人間関係-コミュニケーション(-)	「人のふれあいが減る」	0.88%	0.97%	3.81%
個人情報・プライバシー	「個人情報が漏れやすい」	1.17%	2.52%	3.55%
日常生活-その他	「旅行案内」「遠隔医療」	1.32%	2.91%	4.75%

回答欄ごとに明らか（$p < .05$）な差異があったコードのみを記述。
パーセントは各回答欄の中でのコード出現率である。（出典：川端・樋口 2003 の表 5・表 6 より作成）

5. 統合アプローチの意義

本稿では、テキスト型データを計量的に分析する従来の 2 つの方法、すなわち分析者の作成したコーディング基準にそって言葉や文書を分類する Dictionary-based アプローチと、多変量解析によって言葉や文書を分類する Correlational アプローチとを、互いに補い合う形で統合することを提案した（2 節）。そして、この統合アプローチによる日本語テキスト型データの分析に適したシステムとして KH Coder を作製・公開し（3 節）、本アプローチ並びにシステムを用いた分析の事例を示した（4 節）。

分析事例では、分析の最初の段階において、恣意的なものとなりうる「手作業」をまじえずにデータを要約したものとして、自己組織化マップによる言葉の布置が提示された（図 3）。これは、分析者と第三者が共有しうる資料としての意味を持つものである。そして分析の次の段階では、分析者の解釈ないし問題意識を追究するために、コーディング規則の作成と、コーディング結果の計量的な分析が行われた（4.2 節）。

5.1　2 つのアプローチの峻別

分析例で示したように、分析の最初の段階では、多変量解析によってデータを要約するという点で、一見すると従来の Correlational アプローチに似通った作業が行われている。だが、従来の Correlational アプローチとは大きく異なる部分もある。

というのも、テキスト型データからそのまま言葉を取り出して表 1-（b）に示した形に整理すると、多くの種類の言葉が取り出され、そのほとんどが数回程度しか出現していないという、一般的な多変量解析には不向きなデータ分布となることが多い。そこで、従来の Correlational アプローチではデータ分布の問題に対処するために、分析に用いる言葉を逐一選んだり、「以前」と「これまで」のような似通った言葉を同じものとして扱うよう指定するなどの、「手作業」を行うことが多かった（例えば Miller & Riechert 2001, 大隅・Lebart 2000）。

いくつかの言葉を同じものとして扱うという指定は、当初の Dictionary-based アプローチで用いられた、「aimless、anarchy、chaos などの言葉をアノミーの現れと見なす」といった単純なコーディング規則と、似通ったものに見えないだろうか。異なる点は、「アノミー」のような命名を行っていない点だけである。この観点から見ると、Dictionary-based アプローチと似通った作業が、従来の Correlational アプローチには混入していたと言える。そして、こういった作業を行っている間には、Correlational アプローチにおける当初の主張に反して、分析者が自らの理論や問題意識を、知らず知らずのうちにデータに対して押しつけてしまうことも当然起こりうる。

よって本アプローチでは、このような「手作業」に訴えることはせず、そのかわりに、伝統的な手法ほどデータ分布に関して厳しい前提を持たない、ニューラルネットワークや連関規則などの多変量解析手法を用いることとした[11)]。この点が、従来のCorrelationalアプローチとは大きく異なる点である。また、これによって（1）恣意的なものとなりうる「手作業」を一切まじえずに、データを要約・提示する段階と、（2）コーディング規則作成によって理論仮説ないし問題意識を明示的に操作化し追究する段階とを、明確に区別した点に本アプローチの特徴がある。言葉を換えれば、従来の 2 つのアプローチを峻別の上で統合したものが、本アプローチである。

5.2　従来のアプローチとの比較

従来の Correlational アプローチと比較して本アプローチでは、データを要約・提示する際に「手作業」を省くことで、分析者の持つ理論や問題意識によるバイアスをより明確に排除できるようになった。また、意図的に理論や問題意識を分析に反映させるにしても、単にいくつかの似通った言葉を同じものと見なすだけでなく、例えば「デフレ対策の会議」と「井戸端会議」とを区別できるようなコーディング規則を採用したことで、理論や問題意識をより正確に操作化し測定できるようになった。さらに本アプローチでは、データ全体にわたって似通った言葉の組み合わせを探しだし、同じものとして扱うよう指定するといった作業が不要になったので、必要な労力が軽減されうる。分析の最初の段階でえられたデータ概要を参照しつつ、理論の検証や問題意識の追究に必要な部分だけをコーディング規則によって取り上げれば良いためだ。

次に、従来のDictionary-basedアプローチと比較して本アプローチでは、多変量解析によってデータを要約・提示するという手順を加えたことで、分析の客観性が向上している（2.2 節）。さらに、前もって多変量解析によるデータ要約を行っているため、データに即したコーディング規則の作成が容易になったことも、利点として挙げられよう[12)]。というのもデータの量が一定以上になると、データ全体を記憶すること、言葉を換えれば、データを読み進めながら理解を積み重ねてデータの全体像を把握することが難しい。当然この状態では、データに即したコーディング規則を作成して分析を行うことは難しかったであろうからだ。なお、理論や問題意識を自由に操作化し追究できるというDictionary-basedアプローチの利点については、本アプローチでもそのまま継承している。つまり本アプローチでは、従来と同様に、テキスト型データに含まれる様々な側面に自由に焦点を絞ることができ、さらに従来よりも

客観的かつ容易に、それを行えるようになった。以上のような点に、本アプローチの意義を見出すことができよう。

【謝辞】
執筆にあたり、匿名審査員の先生方より有益なコメントをいただきました。記して、感謝いたします。また本稿は科学研究費補助金による研究成果の一部です（基盤研究A(2)13301007「情報通信技術（IT）革命の文化的・社会的・心理的効果に関する調査研究」研究代表者：直井優、特別研究員奨励研究「コンピュータ・コーディングを用いた電子コミュニティの分析」研究代表者：樋口耕一）。

【注】

1) 例えば、現在 http://popfile.sourceforge.net にて公開されているソフトウェア「POP File」の場合、ベイズ推定によるメールの選別を行っており、利用者の 6 割以上が、自動選別において 95%を超える精度をえたとされている。
2) 大隅・Lebart（2000）の場合、Correlational アプローチを提案した Iker & Harway （1969）の研究に言及・依拠しているわけではない。だが Correlational か Dictionary-based かという、本稿で取り上げた観点から見れば、主に対応分析という多変量解析の手法によって言葉や文書（回答）を分類、ないし布置する大隅・Lebart（2000）の手法は、明らかに Correlational アプローチに含まれるものである。
3) システムが公開ないし販売されているわけではないが、その他に興味深いものとして、SSM 職業コーディングを自動化する高橋（2000）の研究が挙げられよう。
4) KH Coder は現在 http://hey.to/KO-ichi にて公開している。公開にあたっては、マニュアルの中で処理内容を詳細にわたって開示し、さらにフリー・ソフトウェアとして公開することで、自由に利用できる、開かれたシステムとしての公開を目指した。フリー・ソフトウェアとは、「無料」というよりも「自由なソフトウェア」の意であり、処理の内容を確認したり、変更したりといったことが可能な状態（ソースコード）で配布される点に特徴がある。これによって、万一必要とあらば処理内容をチェックしたり、あるいは KH Coder に新たな機能を付け加えたりといったことが自由に行える。この点は、商品としての性格上、公開できないノウハウがあるために利用者が処理内容を正確に理解できない場合があったり、システムの操作こそ非常に容易ではあるが、あらかじめ準備された、極めて限られた種類の分析しか行えない場合がある市販のテキストマイニング・システムと大きく異なる点である。
5) データ中から語を切り出す際には、形態素解析システムとして「茶筌」（松本ほか 2003）ないしは「JUMAN」（黒橋・長尾 1998）を利用する。これらのシステムが切り出すのは厳密には語ではなく形態素であるが、本稿では便宜的にこれを語と見なしている。
6) KH Coder にはクラスター分析や因子分析などの多変量解析を行う機能は無く、それらの解析を行うためには SAS や SPSS などの統計システムが別途必要となる。基本的に統計計算は統計システムで行うという指針で KH Coder は作製されているが、例外的に、（i）その都度、統計システムを用いていたのでは分析手順が煩雑になり、試行錯誤を伴う分析が困難になるものや、（ii）一般的な統計システムでは実行が難しいものについては、計算機能を KH Coder に備えることとした。例えば、コー

ディング結果を集計するための一連の機能や、語の連関規則（association rule）を算出するための機能を備えている。

7) さらに、例えば「製造」「製する」「作る」などを同じ語として扱うといった指定を無数に集めた既製のシソーラスを適用すれば、扱う語の数をさらに減らすことができる。しかし、分析時に、非常に一般的な意味の解釈を目指すのでない限り、大規模なシソーラスの利用は難しい。分析の目的によっては、明らかに異なる語と見なされるべき語が、同じ語と見なされてしまいかねないためだ（Krippendorff 1980=1989:195）。もちろん、語を基本形に直して抽出することで同様の弊害が生じる危険性はあるが、大規模シソーラスと比べればその程度は軽微であり、多変量解析によるデータの要約を行う段階においては、これを許容しうるリスクと見なした。ただし、コーディング規則を作成する段階では、この危険性は回避されるべきものと考え、基本形に直されていない状態の言葉による条件指定も可能なように、KH Coder を設計した。

8) KH Coder は現時点では、構文解析、係り受け解析、ないし格フレーム等の技術を用いたさらに複雑な条件指定には対応していない。各種のデータを対象とした場合にどの程度正確に自動解析が可能なのか、また、様々な研究関心にもとづくコーディングを行う際、これらの技術をどの程度有効に用いることができるのかといった点の検討を含めて、これらの技術の利用については今後の課題としたい。

9) KH Coder におけるコーディングは、基本的には文書を分類するのではなく、文書から要素を抽出するという考え方なので、1 つの文書が複数のコーディング規則に合致すれば、当然、複数のコードが与えられることになる。なお、他のコードが与えられていないことという条件指定を加えつつ、コーディング規則を作成することで、排他的な分類を行うことも可能である。

10) 自己組織化マップとはニューラルネットワークの一種で、中間層を持たない 2 層型の教師無し競合学習モデルであり、高次元空間の複雑で階層的な関係を 2 次元平面に表現可能であるとされている（Kohonen 1988=1993）。文書空間は明らかにそのような高次元空間であり（Doszkocs et al. 1990）、自由回答データもそれに準じるものであろうと考えて自己組織化マップを用いた。

11) 「手作業」を一切まじえずに多変量解析を行い、それによってデータ概要をできうる限り把握することが、本アプローチにおける重要な点である。よって、4 節では一例として自己組織化マップが利用されているが、必ずしも自己組織化マップという手法に固執するものではない。より適切な手法が無いか否かの検討を常に行うべきであろう。

12) この点は、仮説検証というよりも、4 節で例示したような探索的ないし帰納的な分析を行う場合に、とりわけ重要な利点となろう。

【文献】

Berelson, B. 1952. *Content Analysis in Communication Research*. New York: Hafner Press.

Doszkocs, T. E., J. Reggia & X. Lin. 1990. "Connectionist Models and Information Retrieval." *Annual Review of Information Science and Technology* 25:209-60.

Iker, H. P. & N. I. Harway. 1969. "A Computer Systems Approach toward the Recognition and Analysis of Content." in G. A. Gerbner, O. R. Holsti, K. Krippendorff, W. J. Paisly & P. J. Stone (eds.) *The Analysis of Communication Content: Developments in Scientific Theories and Computer Techniques*. New York: Wiley & Sons.: 381-486.

川端亮・樋口耕一. 2003. 「インターネットに対する人々の意識 ―自由回答の分析から―」『大阪大学大学院人間科学研究科紀要』29:163-81.

Kohonen, T. 1988. *Self-Organization and Associative Memory*. New York: Springer-Verlag. =1993. 中谷和夫（監訳） 『自己組織化と連想記憶』シュプリンガー・フェアラーク.

Krippendorff, K. 1980. *Content Analysis: an Introduction to its Methodology*. London: Sage. =1989. 三上俊治ほか（訳）『メッセージ分析の技法-「内容分析」への招待』勁草書房.

黒橋禎夫・長尾真. 1998. 『日本語形態素解析システム JUMAN version 3.61』京都大学大学院情報学研究科.

Lasswell, H. D. 1941. "The World Attention Survey: An Exploration of the Possibilities of Studying Attention Being Given to the United States by Newspapers Abroad." *Public Opinion Quarterly* 5(3):456-62.

松本裕治・北内啓・山下達雄・平野善隆・松田寛・高岡一馬・浅原正幸. 2003. 『形態素解析システム「茶筌」version 2.3.1 使用説明書』奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座.

Miller, M. & B. P. Riechert. 2001. "Frame Mapping: A Quantitative Method for Investigating Issues in the Public Sphere." in M. D. West (ed.) *Theory, Method, and Practice in Computer Content Analysis*. London: Ablex: 61-75.

那須川哲哉・河野浩之・有村博紀. 2001.「テキストマイニング基盤技術」『人工知能学会誌』16(2):201-10.

大隅昇・L. Lebart. 2000. 「調査における自由回答データの解析 ―InfoMiner による探索的テキスト型データ解析―」『統計数理』48(2):339-76.

Osgood, C. E., G. J. Suci & P. H. Tennenbaum. 1957. *The measurement of Meaning*. Urbana: University of Illinois Press.

Phillips, D. C. 1990. "Subjectivity and Objectivity: An Objective Inquiry." in E. W. Eisner & A. Peshkin (eds.) *Qualitative inquiry in education: The continuing debate*. New York: Teachers College Press.: 19-37.

Popping, R. 2000. *Computer-assisted Text Analysis*. London: Sage.

Renz, I. & J. Franke. 2003. "Text Mining." in J. Franke, G. Nakhaeizadeh & I. Renz. (eds.) *Text Mining: Theoretical Aspects and Applications*. Heidelberg: Physica-Verlag. 1-19.

Roberts, C. W. ed. 1997. *Text Analysis for the Social Sciences*. Mahwah: Lawrence Erlbaum.

佐藤裕. 1992. 「自由回答のコンピュータコーディング」『第 14 回数理社会学会大会研究報告要旨集』.

Seale, C. 2000. "Using Computers to Analyze Qualitative Data." in D. Silverman (ed.) *Doing Qualitative Research: A Practical Handbook*. London: Sage.: 154-74.

Stone, P. J. 1997. "Thematic Text Analysis: New Agendas for Analyzing Text Content." in C.W. Roberts (ed.) *Text Analysis for the Social Sciences*. Mahwah: Lawrence Erlbaum: 35-54.

Stone, P. J., D. C. Dunphy & D. M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press.

高橋和子. 2000. 「自由回答のコーディング支援 ―格フレームによる SSM 職業コーディング自動化システム―」『理論と方法』15(1):149-65.

田中重人・太郎丸博. 1996.「文章データのコンピュータコーディング ―プログラムとその応用―」『第 22 回数理社会学会大会研究報告要旨集』: 20-23.

谷口敏夫. 1999. 「全文からの『位置情報付き用語』の抽出」川端亮（編著） 『非定型データのコーディング・システムとその利用』平成 8 年度～10 年度科学研究費補助金(基盤研究(A)(1))(課題番号 08551003) 研究成果報告書, 大阪大学. 31-58.

Woodward, J. L. 1934. "Quantitative Newspaper Analysis as a Technique of Opinion Research." *Social Forces* 12:526-37.

安田三郎. 1970. 『社会調査の計画と解析』東京大学出版会.

（受稿 2003 年 10 月 7 日／掲載決定 2004 年 2 月 19 日）

Quantitative Analysis of Textual Data: Differentiation and Coordination of Two Approaches

Koichi HIGUCHI

Faculty of Human Sciences

Osaka University

1-2 Yamadaoka, Suita, Osaka 565-0871, JAPAN

In the field of social research, there have been quantitative analyses of various kinds of textual data, such as newspapers or open-ended survey questions. In most of these cases, either the correlational approach or the dictionary-based approach has been employed. In the former approach, multivariate analyses are utilized to examine word correlations or co-occurrences and discover themes. For example, cluster analysis of words is used to find groups of words that appear frequently in the same document. In the latter approach, words or documents are assigned to preexisting categories following rules that are created by the researchers, and categorization results are analyzed quantitatively. In this paper, the author evaluates these two approaches and proposes a third approach in which they are complementarily coordinated. In order to perform analyses that follow this new approach, original software was developed and distributed as free software. The author presents an analysis of responses to open-ended questions that uses the new approach and this software. In the conclusion, the author highlights the advantages of the new approach in comparison with the traditional two approaches.

Keywords and phrases: content analysis, text mining, qualitative data, correlational approach, dictionary-based approach

（Received October 7, 2003 / Accepted February 19, 2004）