調査におけるテキスト型データの解析

大隅　昇
テキスト・マイニング研究会代表
統計数理研究所・名誉教授

１．調査におけるテキスト型データのマイニング　–考え方–

ここでは, 調査におけるテキスト型データの分析, 定性情報の分析という枠の中で議論する. 典型的な例が，自由回答質問を用いた調査とそこで取得のデータの分析がある．ここで調査とは，社会調査や市場調査などで用いられる調査方法論を適用することをいう.

もちろん，第 I 部で述べたように，テキスト・マイニング（TM）あるいはテキスト型データのマイニング（TDM）は，自由回答に限らず，テキスト型データの現れる様々な場面に適用される一般的な方法論である．しかしここで，従来の言語学や言語情報処理における研究，あるいはそれらの延長線上にある種々の方法論とは若干視点を変えた方向で考えてみる. ここでは, 日本語のテキスト型データの解析に**「探索的統計データ解析手法」**を積極的に取り入れることで, どこまで客観的な分析が可能であるのか, 適用上の問題や限界は何かを実験検証的に進め, また我々が主張してきたデータ科学の方向で解決を図るための一つの方向付けを行うことである.

断るまでもなく，各種調査の実査環境が大きく変化する中で，定性調査や定性的アプローチで取得した日本語文字情報あるいは文章型テキスト・データの取得法や解析方法を求める声が高い．実際，国内ではテキスト・マイニングというと，市場調査，社会調査などにおける自由回答質問，グループ・インタビューなどで記録したテキスト型データの分析，あるいはコーディングといったことへの関心が主となっている．筆者も主にマーケティング・リサーチの分野において様々なテキスト型データの解析を体験してきた．こうした経験や実験調査を通じて得た知識に基づいて，ここでは以下のような事項について概観する.

- データをどう考えるか
- とくに，データ解析におけるデータの分類区分とテキスト型データの関係
- 調査におけるデータ取得方法のあり方
- 調査における自由回答の分析，その位置づけ
- 日本語の特徴と統計的データ解析との関係
- 日本語の特性を活かした多次元データ解析接近法
- テキスト型データ解析システム：WordMinerr の紹介と設計指針
- 分析例の紹介 –記述統計的な分析，その重要性–

山本夏彦（2000）は「完本文語文」の中で「口で語って耳で分るのが言葉である」と述べ, また近年の日本語の乱れを指摘している．また新聞紙面，雑誌等でも「日本語のあり方」が改めて（あるいは今もってと言うべきか），あれこれと議論されている．国語審議会における議論や多くの研究報告を待つまでもなく，日本語自体が言語として完成された形にあるものではない．むしろ，言語は流動的にたえず変化するものであり，またそれであるからこそ言語であるとの指摘も多々ある.

話題を絞り込んで，いわゆる調査（態度，意見等）あるいはアンケート（enquéte）と総称される分野で用いられる自由回答方式（open-ended question : OA，free answer : FA 等）の調査に限って考えてみても，定性的アプローチの決定的な方法論があるわけでなく，未だ模索の中にあると考えられる.

２．データをどう考えるか

古典的な統計学では，取得した測定値（実測値）を連続的か離散的かに分けて考える．これはその背景に統計的分布を連続的分布と考えるか離散的分布と見ることに関係する．前者の例が正規分布や指数分布であり，後者の例として二項分布やポアソン分布などがある．具体的な測定例でいうと，身長や体重を測定したデータは連続量とみなし，また電話の呼数や車の台数などは離散量と考える．また，品質管理などの分野では，これを計量的，計数的と対応させて考える．

しかし，こうした数理的な考え方だけでは現実のデータの説明は十分にはできない．そこで多くの場合，**尺度**（scale）によるデータの分類区分を用いる．これはまず，**質的データ**（qualitative data）と**量的データ**（quantitative data）とに分けて考え，質的データはさらに**名義尺度・名目尺度**（nominal scale）と**順序尺度・順位尺度**（ordinal scale）に，また量的データは**区間尺度**（interval scale）と**比例尺度**（ratio scale）とに分けて考える．この考え方に従うと，定性情報，定量情報に関わりなく，多くの調査データの解釈が容易となる．

この量的データ，質的データの枠組みと，連続的，離散的と考える見方との対応は，表１のように二元表の形で考えると分かり易い．

しかし，最近はデータの様相が多様化し，こうした枠組みだけでは，必ずしも十分な説明ができなくなっている．つまり別の視点からのデータの分類区分も必要となってきた．たとえば，画像（イメージ：静止画，動画），音声など，見かけ上は計量化せずに扱うことがある．また，テキスト型データを含む文字情報も上のような枠組みでは必ずしもうまくは説明できないこともある．

このようなことで，以下のような区分を考えておくことも時として必要かもしれない．

○数値的か非数値的か

数値的（numerical）

数量，数値，計数として表記されるもの

非数値的（non-numerical）

文字，記号，イメージ（静止画，動画），音声など

○構造的か非構造的か

構造的データ（structured data）

カテゴリー化，タグ化，コード化などを介しデータベース化など整備されたデータ

非構造的データ（unstructured data）

とくに何も措置されない裸のままのデータ

しかしどのような区分分類を行っても，それで一意的に区分して考えるのではなく，状況に応じて解釈や分析上の操作に都合のよい形で用いると考えればよい．

表１　データの分類区分

		質的データ		量的データ	
		名義尺度	順序尺度	区間尺度	比例尺度
連続量		（この組み合わせは考えられない）	音の強さの段階的区分 色度、光沢度	温度（℃） 硬度 比重	単位を持つ測定値データの大部分（長さ，重さなど）
離散量	多値	機械名 作業者名 工場名 原産地名，など	段階的評価の成績データ 調査票の選択式質問における選択肢（「満足」「やや満足」「満足でない」,）など	ＴＶのチャンネル 体育館の利用日数 車の故障台数，など	車の走行台数 都市内人口 参加者数 家の戸数
	二値	性別（男、女） 「あり，なし」（有，無） スイッチの状態（「入，切」） など	物体の大きさ（大きい，小さい） 濃度（濃い，薄い） 硬さ（硬い，柔らかい） など	旅行経験の有無（回数を考慮に入れれば多値データとなる）	瓶入りと缶入りのジュース単価（二値の分類区分で層化）

（注）データあるいは測定値の特性をどう分類区分するかということと，それを何らかの加工を経て，別の情報に置き換えること（情報の変換）とは分けて考えた方がよい．ここでいうデータの分類とは，その操作上の一つの目安とする考え方である．

３．定性調査における自由回答の役割

３．１　定性情報と定性調査

調査環境の急速な変化，とくに調査環境の悪化が指摘されるようになって久しい．調査の質の低下が深刻な問題とされるように，様々の原因で満足できる内容の調査がきわめて困難になってきた．とくに従来からの定量的調査の実施困難性や様々な問題の提起，たとえば，回収率の低下，非標本誤差や無回答の増大，そして貴重な標本抽出枠（サンプリング・フレーム）であった住民基本台帳や選挙人名簿等の閲覧制限，情報公開法の実施等に関連した調査情報取得環境の変容がある．そして，インターネット調査（電子メール調査，Web 調査）などの新たな調査法が登場し，従来の調査法の見直し，たとえばクオータ・サンプリング，エリア・サンプリング，郵送法，電話調査，面接調査等のあり方が改めて問われている．とくに調査費用面の負荷が増大し，まともな調査を実施することが困難となり，いきおきインターネット調査など安易な方向に向かう傾向にある．

このようなことで，従来とは異なる意味で，あるいは従来にもまして質的・定性的調査への関心が高まっている．サンプル数がそれなりに大きく，また伝統的な標本調査法に従ったサンプリング操作を経て行われる量的な調査（たとえば従来型調査の中心であった面接調査，留置調査，郵送調査等）が，経済的にも労力の面からも負担が大きく，一方それに見合った成果が次第に期待できない状況にあることから（たとえば回収率の低下），質的調査や定性調査に関心が移行する傾向が見られる．

もっとも，市場調査分野等では，早くからグループ・インタビュー（GI）やフォーカス・グループ（FG），モチベーション・リサーチなどが利用されてきたが，ここで取得のデータ解析のあり方が改めて注目されている．また，少数のサンプルや，条件を限定した回答者を相手としたモニター調査や，インターネット調査（とくに Web 調査）などの調査方式では，自由回答や自由記述の質問を多用し，ここで取得したデータの質的解析を試みることが多くなってきた．

とくに，インターネットの普及により，**電子的調査情報取得手法**（CASIC：Computer Assisted Survey Information Collection）や**コンピュータ支援によるデータ収集**（CADAC：Computer Assisted Data Collection）の研究や実用化が進み，自由回答に代表される**テキスト型データ**（textual data）の取得が内容の質の適否に関わりなく，容易に，しかも大量取得が可能となった［Couper and others (1998)］．このようなことで，自由回答質問を多用する調査（とくに消費者動向調査，インターネット・マーケティング）が多くなった．

同時に，こうした自由回答・自由記述のデータを解析するための方法論の研究も見られるようになった．またこれとは別に，従来からの定性調査手法として，面接法（深層，集団など），投影法（言語連想，文章完成，略画完成，絵画解釈）等があったが，こうした方法でも，取得データが電子化されてテキスト型データとしての利用が容易になってきた．このほか，第 I 部でも述べたように，CRM（Customer Relationship Management）との関連で，企業のコール・センター，コンタクト・センターや顧客相談窓口における取得データの定性情報解析など多種多様な試みがあり，また具体的方法論や解析システムの開発への期待も高い．このように，今後は，調査環境の多様化に伴う，文章型・文字型によるデータ取得や解析の機会の増大が考えられる．

３．２　通常の調査法の条件と特徴

ところで，従来型の調査あるいは一般的な調査（面接法や留置法）で，たとえば質問紙調査票による選択肢型質問を用いた調査法による調査実施における要件として，「**妥当性，信頼性，客観性，再現性**」等の保証が挙げられる．

一方，定性情報データ，とくに自由記述文・テキスト型データの取得・分析では，これらの保証が得られにくいとされてきた．たとえば，再現性を考えてみても，これをどう扱うのかがあまり明らかではない．

しかし，自由回答方式による質問形式を定性情報取得の有力手段として確立するには，自由回答の取得環境，調査設計・抽出，妥当性をどう考えるか，という調査法や調査方式としての基本的な問題の検討が重要なことは言うまでもない．同時に，現時点における研究成果が未成熟であることから，自由回答のみから得られる情報やその分析処理にも限界があることも自明である．とくに，自由回答データの分析だけでは意味解釈が恣意的となりやすいことは自明であるから，従来の科学的な調査手法に関連した方法論の援用，支援が必要となる．（→これについては後述）

３．３　データ取得環境の変化　–調査法の考え方–

調査における**データ収集**（data collection）の技術面，すなわち**調査方式**（調査モード：survey mode）に目を向けると，コンピュータ利用が一般化したことや，膨大な量の情報の電子化，データベース技術の進歩に伴い，メタ・アナリシス（集積情報の横断的相互利用等），データ・アーカイヴ，マイクロ・リンケージ，そしてデータ・ウェアハウス等の関連技法が登場してきた．これら技術的要素を背景にして，急速にデータ・マイニング（DM：data mining）の方法論が登場し，これと関連して多数のテキスト・マイニング手法が登場している（第 I 部で述べた）．コンピュータの処理能力に期待し，データベース上の大量データ処理を通じて，**知識の組織化**（knowledge organization）や**知識発見**（knowledge discovery）を図るという構想であるが，かけ

声程に見合った高い成果が上がっているかは，今後を見ないと即断はできない.

ここで注意すべきこととして，調査方式の多様化に伴い，調査法とくに**標本調査法**（sampling survey methods）と**調査方式**（data collection mode，survey mode）との関係が曖昧かつ複雑になってきたことがある．もちろん両者は不可分の関係にはあるが，調査を考える上では，意識的に分けて考える方がよい.

ここで標本調査法の研究とは，いわゆる調査法の基礎的な諸事項に関わる研究をいう．たとえば，標本抽出法（サンプリング）として，個々の確率的サンプリング技法（probability-based で考える無作為，系統，確率比例，層化比例など）や非確率的サンプリング技法（non-probability-based のクオータ法，スノーボール・サンプリング，インターセプト法など）のように分ける.

一方，調査方式（調査モード，データ収集方式）とは，「いかにデータを集めるか」の方式をいう．面接法，郵送法，訪問留置法，電話調査，インターネット調査（電子メール調査，Web 調査）等々の調査方式のタイプをいう.

国内では，標本抽出法として，住民基本台帳や選挙人名簿などの優れた標本抽出枠が利用できたことから，ほとんど理想に近い確率的な方式（probability-based）の適用が可能であったことで，調査方式と標本調査法との関係をそれほど厳密に区分して考える必要がなかった.

同様な理由から，**調査誤差**（survey errors）を考える場合も，国内では，**標本誤差**（sampling error）と**非標本誤差**（non-sampling error）に区分し議論することで，大方の説明をつけることが従来は可能であった［社会調査ハンドブック（2002)］．しかし，新しい調査方式，たとえばインターネット調査や RDD 方式（Random Digit Dialing method）に依拠する電話調査（CATI など）の普及により，この調査誤差の考え方も見直しが必要となっている，欧米，とくに米国の調査法研究では，この誤差を Groves による分類区分に従って考えることが多い［Groves（1989），Groves and others (2002)］．これは，インターネット調査に限らず他の調査についての誤差の議論や調査法研究を進める上で有用な枠組みである．ここでは Groves にならって，調査における誤差を，その発生源によって以下の４つに大分類する.

① **カバレッジ誤差**（coverage error）

調査したい対象と標本抽出枠のズレに起因，ここでは抽出確率がゼロである個体（unit，observation）の存在が問題になる

② **標本誤差**（sampling error）

標本抽出に伴う誤差のこと

③ **無回答による誤差**（non-response error）

標本中に回答のない個体が存在することによる誤差，その理由の調査，非標本誤差の一部

④ **測定誤差**（measurement error）

調査票・質問形式の設計，調査員のスキルの違いなどで生じる偏り，回答者の性質・回答傾向，調査方式等，測定に関わる誤差

この分類に従うと、標本抽出（サンプリング）段階においては主に①や②の誤差が問題となる．また，調査方式，データ収集過程においては③や④の誤差を主に問題とせねばならない．それぞれの誤差を小さくすることを目標に調査法や調査方式の改善を図って，つまり，標本抽

出法の研究，無回答や欠測値処理の研究，調査方式や調査票設計等の研究が体系的に行われている．

たとえば，インターネット調査では，計画標本がインターネット・ユーザをどの程度代表するかというカバレッジ誤差の問題，調査票の設計や質問形式（様々な様式が使えること）の差異で回答結果がどう異なるか，偏りが生じるかと言った測定誤差の問題などがある．

このようなことを指摘した理由は，インターネット調査という調査方式を用いることが一般的なこととなってきた中で，調査法研究の議論が十分でないまま，あたかも情報が豊富で信頼できるデータが容易に取得できるといった安易な考えが広まっていることへの警鐘としたいからである．インターネット調査に限らず，調査環境全体に生じている不具合，環境悪化や抱える諸問題がこのような観点で整理され、各研究が方向づけられることが肝要である．

［インターネット調査の定義］

⑴単純には，

電子メール調査（EMS: Electronic mail survey）と，

Web 調査（Web-based survey，Web survey）

と分ける．かって，様々な試みがあったが，現在はこの区分で説明できるであろう．

⑵より具体的に記述すると，以下のようになる．

① **電子的データ取得法**（CASIC，CADAC）の一つであって，Web 調査，電子メール調査など，ネットワークや WWW 環境を用いたインターネット調査システム下で実施される調査のこと

② “主として”インターネット・ユーザを調査対象者とすること

③ コンピュータ支援の**自記式調査**（Computer-assisted self-administered）

④ ［CAI：Computer-assisted interviewing の一つとも考えられる］

⑤ とくに，Web 調査では，ブラウザで閲覧する形式の**電子調査票**を用いる（HTML，XML などの言語で記述）

⑥ 回答（取得）がネットワーク上で**リアルタイムに転送記録**される

インターネット調査とは，以上の仕組みを利用するためのインターネット調査システムの構築を始め，登録者集団（リソース，パネル）の構築と登録運用管理，（電子的な）調査票や質問の設計，サンプリング操作，調査依頼から回収，データ処理と集計分析，調査報告など，調査の全過程にわたって**科学的な調査方法論**を適用して行う総合的な活動をいう．

［インターネット調査の特徴］

以上から Web 調査の特徴として，

- 電子的にデータ取得が可能であること，**電子的な自記式調査**（Computer-assisted self-administered）である
- **画面設計**（電子調査票の質問レイアウト）の自由度があること
- **双方向性**（interactive）があること

・ 面接調査などと違って面接員の誘導，対応の影響が排除できること
・ 画面上に回答に戸惑ったときのガイドや警告を示せること
・ 画像（静止画，動画），音声など**マルチメディア対応**とできること
・ 従って，調査票のデザイン，修飾が豊富であること
・ プログラミングとして**回答の制御や抑制が可能**であること
・ 回答者の**回答行動を電子的に追跡**できること（トラッキング，ロギング）

等々が挙げられる．

自由回答質問について考えると，明らかに測定誤差や無回答誤差などは看過できないわけで，安易に自由回答質問枠を設ければよいとはならない．とくに，インターネット調査では，テキスト・ボックスやテキスト・フィールドの設計（大きさやレイアウト），ポップアップ方式採用の有無などの影響があることが知られている．同様なことは，従来の郵送調査や留置自記式調査でも考慮する必要があったが，インターネット調査では電子的操作・処理を行うことで，状況はより複雑となっている．

なおここで，調査方式（調査モード，データ収集方式）を，その手段が電子的か否かで整理すると理解が容易であろう．吉村（2003）の要約をもとに，これを表２のように要約した．こうした分類を行うと，調査方式を（標本）調査法との関係で議論するうえでも都合がよい．

表２　調査方式・データ収集方式の分類区分

		コンピュータ支援の有無（CA：computer- assisted）			
		なし		CASIC, CADAC	
		面接員方式の有無		面接員方式の有無	
		あり	自記式	あり（CAI）	自記式
調査方式（調査モード）	面　接	面接	訪問留置	CAPI	CASI
	郵　送	－	郵送	－	DBM (Disk by Mail)
	電　話	電話	Facsimile	CATI	－
	ｲﾝﾀｰﾈｯﾄ	－	－	OFG	電子メール調査, Web 調査

（注 1）CATI：computer-assisted telephone interviewing
（注 2）CAPI：computer-assisted personal interviewing
（注 3）CASI：computer-assisted self-interviewing
（注 3）OFG：online focus group

４．日本語文章・テキスト型データの解析の方向

第 I 部で述べたように，テキスト・マイニング（TM），テキスト型データのマイニング（TDM）は，様々な研究分野の研究要素，技術要素の集合体である（第 I 部の図２，図３参照）．これをさらに，自由回答・自由記述の取得に関連する調査データの分析と関連付け，筆者の視点から

図１のように整理した．ここにみるように，調査における自由回答の取得法や取得データの解析は，**従来の（標本）調査法や調査方式の研究との接合面を保持しながら，しかし従来の言語情報処理の諸研究とは異なる方向からアプローチ**すべきことと考えている．もちろん，これは筆者の考え方であり，それがすべてではないが，現時点で調査における自由回答・自由記述のデータを分析する方法としてはそう的はずれでないと考える．

これを事前情報として，調査における自由回答文の解析を考えるとき，以下の特徴や事項に留意した客観的な方法論が必要である．これはまた，調査における自由回答文・自由記述文の取得にあたって，従来から指摘されてきたことでもある．

① 考えたことがないことには答えにくい，また，いきなり質問を受けても答えることが難しい，いわば「白紙を出されて何か絵を書くように」といわれてもなかなか思いつかないことと同様である．
② 一方，予想しなかった回答や知見がえられるという期待がある．つまり，選択肢型質問（定量的調査）では予期できなかった情報が得られるチャンスがある．
③ 無記入が多くなる傾向があるとされてきた．
④ 標本調査法や標本抽出法との関連性が明らかにできないと言われている（妥当性の問題）．
⑤ 他の質問（選択肢型質問）の質問文や選択肢の影響を受けるのではないか（回答誘導や迎合の懸念）．
⑥ 得られた自由回答の適切なデータ解析法がないと言われてきた．
⑦ 客観的な統計解析手法の確立が難しいとされる．
⑧ 内容の再現性がない，あるいは信頼性に欠けるとされてきた．
⑨ 集計の手間がかかると考えられている．
⑩ 同一の質問に対する回答に均一性を欠くと言われている，あるいは再現性がない．
⑪ 質問文の意図がどう反映されたか，自由回答のみからは読みとり難い．

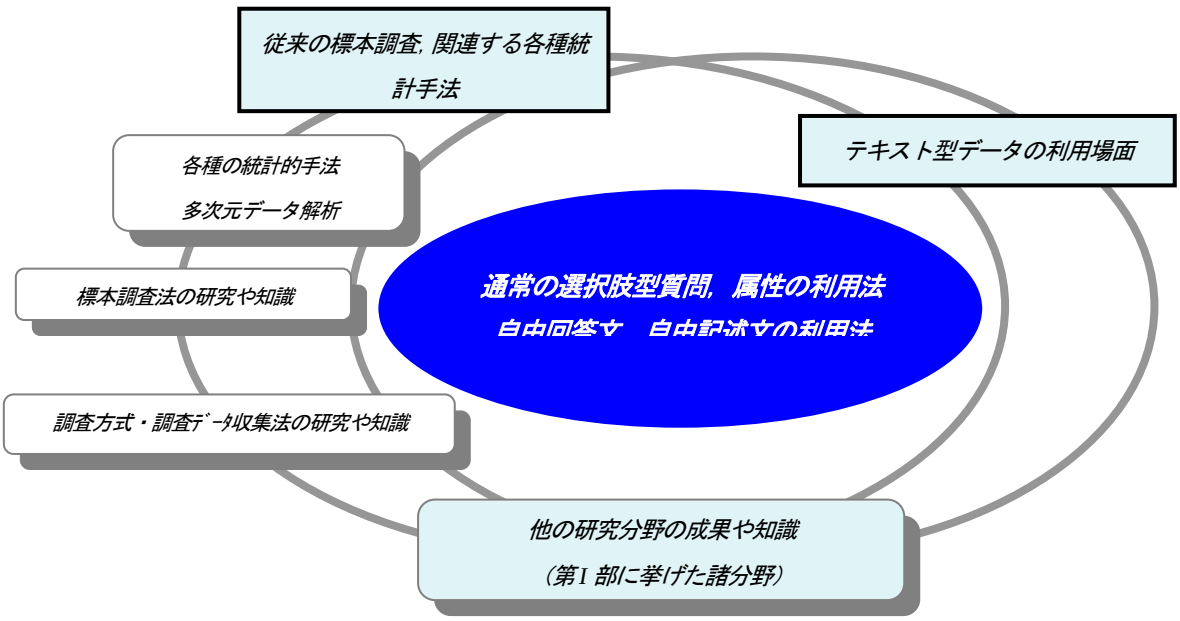

図１　調査における自由回答取得とデータ解析の関係　–概念図–

こうした指摘はもっともであり，この種の研究課題の複雑さ，困難性を示すものである．とくに，調査における自由回答データ取得上の重要な考慮事項は，選択肢質問型調査と異なり，数量として定量的かつ客観的な評価が困難であるとされてきたことにある．たとえば，回答比率で比較する，標本誤差を推定するなどの操作に相当する方法が未だない．また，標本抽出との関連や（サンプル数は適切かなど）調査票や質問の設計やその回答への影響評価をどう考えるか，さらには，調査の反復可能性や再現性の問題をどう捉えるか（通常の回答変動とは異なる事象が現れるであろう）等々，様々の検討要件がある．

一方，今まで述べたように，従来型の調査の実施環境の急激な変化から，とくに実査の困難性が指摘される中で，従来以上に，定性調査への期待が高い．とくに，インターネット調査等の電子的情報取得手段の普及により自由回答の電子的取得がきわめて容易となった事から，その利用可能性の十分な検証のないままに，急速に利用されるようになっている．こうした事情までを考慮して，筆者等は自由回答の取得法や分析法の研究，さらにはその解析システムの開発を進めることが必要と考えてきた．もちろん，ここでも，**調査方式と調査法との関連を意識したアプローチ**が重要であることには変わりはない．このため，以下の視点からの研究展開が重要と考えられる．

(1) “実験調査”による事例検証の積み重ねが必要であること
(2) とくにテキスト型データの異なるタイプの事例検証が重要と思われること
(3) つまり様々な調査方式（調査モード）の比較研究が必要であること
(4) 現場の要請，たとえば市場調査における利用範囲，適用可能性を整理し，システム開発に必要な要件の要因分析を行うこと

たとえば，これを見るための一つの試みとして，今までの筆者の経験例の中から，分析の難易度を“主観的に”ではあるが要約してみた（表３）．これに見るように，体験的情報であるとはいえ，多くのテキスト型データ解析ツールがうたっているように，汎用的に何でも利用できるものではない．これは時として目的に応じたカスタマイズが必要であることを意味している．

理由は様々あろうが，その大元は日本語分析の困難性にあると考えられる．換言すると，TM を有効に活用するには，それなりのデータ取得法を考えるべきということである．また，既に集積化されたテキスト型データの分析を行う場合には，後述するように，また第 I 部で述べたように，その対象データが，どのような段階，様相にあるかを見極めたうえで対処すべきである（6．2 節；第 I 部，3．3 節）．

こうした課題の多くは，従来の自然言語処理，言語情報処理技法だけではカバーできないことは明らかであり，別の観点からのアプローチが必要となる．また，最近の傾向として，データベース検索との関連で，文書，書籍等の電子化の加速化やオンデマンド出版，多様なテキスト型データの蓄積，テキスト・データベースの普及に伴うテキスト・コーパスの利用環境の変化，従来は分析を諦めていた大量の文字型データ・アーカイヴの電子化（の実現容易性），これらのコンピュータ処理の可能性の増大（全文検索，キーワード検索などのツール）等がある．しかし私見ではあるが，こうした方向への TM の適用はあまり進んでいるようには思われない．

また，定性調査を定量調査との優劣を比較するということはあまり意味があるとは思われな

い．とくに従来は，定性調査を定量調査とは別の視点で捉える傾向もあったが，これを改めて，調査における“自由回答データ”にもとづく定性情報の取得においては，従来の選択肢質問形式による調査との“併用”が妥当であるとの観点から議論を進めることを提案したい．つまり，従来の選択肢質問形式を，蓄積のある調査手法に関連した**科学的方法論に裏付けされた定量的評価**に用い，一方，自由回答形式はその**定性情報の計量化**を図る方法論を新たに考え，それら**両者の併用**を工夫することで，より客観的な情報要約を図る方向からデータ解析を考える．また，テキスト型データ解析の統計システム開発に際しても，こうした発想が反映された設計指針が重要と考える．これは，林や大隅が主張してきた**「データ科学」の概念**を具体化するという意味をも含んでいる［林（1998），林（2000），Ohsumi（2000）］．

表３　TM の適用範囲の例の一覧（筆者の過去の分析の経験から）

適用の場面	分析の難易度
調査における自由回答データ	比較的容易
消費者行動調査，自由回答方式の研究	比較的容易
インターネット調査（E-mail調査，Web調査）	比較的容易 調査票・質問設計が重要
Webページ上での製品ユーザの意見聴取	適する
電子メールによるモニターからの回答収集分析	比較的容易
コール・センター，コンタクト・センターでの収集情報	やや面倒
製品に添付の意見葉書の自由記述分析	比較的容易
面接法による録音データの解析（記録を書き起こしてこれを解析）	難しい
自治体「市民の声」の分析	やや面倒，取得環境構築が鍵
患者の聞き取り調査	難しい
被爆者体験日記の解析	難しい
唇口蓋裂患者の矯正治療医師への意識調査	やや面倒，質問設計の専門性高い
論述形式，記述問題の解答の分析	ほぼ適用可
書籍，小説・文芸作品などの文章解析	やや面倒 分析の目的に依存
新聞・雑誌記事など	適用可，分析の目的に依存
発想法，KJ法等の文字データ解析（意見の類型化）	やや面倒
グループインタビュー，フォーカスグループで取得データの分析	難しい
インターネット上でのチャット，対話データ	難しい
日記形式の記述文	かなり難しい
映画・映像のシナリオ分析	難しい
調査票の質問のデータベース化（質問のストック化，共有，メタ解析的）	難しい
コーパス（corpus-corpora，語彙群）生成と辞書化	かなり難しい
メタ・テキスト	かなり難しいこれからの研究

５．自由回答質問の設計

５．１　どのような質問形式があるか

自由回答質問方式を用いることは，いまに始まったことではなく，昔から利用されてきた．とくに，郵送調査や留置自記式調査では，回答者の書き込み時間（回答時間）の余裕があることや，面接調査と異なり回答者の本音を聞き取り易いなどの理由で，用いられてきた．

しかしながら，調査票回収後に，自由回答記入・記述文を何らかの形で整理せねばならないこと，たとえば，各回答を読み込んでアフター・コーディング（ポスト・コーディング）を行い，要約項目別に整理するとか，電子的にコンピュータで処理するために，それら回答をファイル化するなどの操作を必要とした．とくに，コンピュータによる日本語処理機能も十分ではなかったので，アフター・コーディング処理は重要な操作であり，ある種の職人的技を必要とされ，高いスキルや経験が求められてきた．もちろん，欧米語に比べて，日本語処理を行うコンピュータ・プログラム環境も十分でなく，せいぜい内容分析用ソフトを用いて，時にはローマナイズしたデータを用いた処理などが行われた．

しかし，1990 年代半ばから，急速に普及した電子メールなどを使った**電子調査**（electronic survey）などが登場し，電子化されたテキスト・ファイルなどの利用が可能となった．さらに 1995 年頃からいわゆるインターネットの普及により Web 上での調査（**Web 調査**：Web-based survey，Web survey）が始まり，ここでは HTML 言語等を用いた電子的な調査票が利用できる環境が整い，また急速に電子調査環境の整備が進んで，結果として紙形式の調査票による **P&P 方式**（paper and pencil procedure）から，自由回答質問方式を電子的に用意した Web 調査による**電子調査票**（electronic questionnaire）に移行する傾向がある．

つまり，**調査方式（調査モード）に大きな変化が生じた**わけである．しかし，このことは自由回答質問方式の調査設計を行う労力の軽減には寄与しているが，自由回答質問を用いた調査データ収集の過程で生じる諸問題の検討や考慮を軽減するものではない．むしろ，電子的に自由回答を取得できる仕組みが簡単に装備できることから，従来の P&P 方式の調査に比べて，検証すべき問題が加速的に増えたといってよい．

よく耳にする言葉に，

- Web 調査では自由回答の豊富なデータが得られる
- Web 調査では自由回答への書き込み率（記入率）が高い
- 結果として表現内容が豊かで情報量が多いデータが得られる
- （自記式であることから）回答者が本音を書きやすい
- よって，予期しないあるいは思いがけない意見の聴取ができる

といったようなことがある．しかし，これらはいずれも科学的な検証を得た結果から述べられたことではなく，これらを示すには十分な論証，論拠が必要である．そもそもインターネット調査については，調査方式として様々な問題を抱えており，これらを順次解決しながら，慎重に利用すべき調査法なのである．ここらについては．筆者らのグループが永年にわたり実験調査を通じて検証してきたことでもある．とくに，自由回答質問を Web 調査により行う場合のテクニカルな注意事項としては，以下の節に述べるようなことがある[「社会調査ハンドブック」，Dillman（2000），Grossnicle and Raskin（2001）などを参照]．

５．２　インターネット調査における質問形式の留意事項

ここではとくに，**インターネット調査**（電子メール調査，Web 調査）を利用した自由回答取得における留意事項について述べる．自由回答取得手段として，前述のようにインターネット調査，とくに Web 調査を用いることへの期待は高いが，同時に根拠なく安易に利用することは慎まねばならない．

Web 調査の特徴の一つは，**質問形式の多様性**にあるとされている．すなわち，調査票の作成時に，様々な様式の質問形式を取り入れることが可能であるとされてきた．また，現在利用されている Web 調査の回答方法は，原則として**自記式（self-administered）**である．これは，一見すると従来調査における質問紙（P&P 方式）を用いた自記式回答（たとえば，郵送調査や訪問留置自記式調査）に類似している．

しかし，インターネット調査における調査票の作成方法や設計方法については，適切な指針やガイドがあるようで実はあまりない．従来型の調査においても，調査票の設計や質問作成方法は基本事項とされ，これを適切に行うにはそれなりのリテラシーを必要とされてきた．インターネット調査における調査票の設計や質問作成においても，従来型調査における諸事項，留意事項への配慮は当然のことであるが，加えて**インターネット調査特有の手当，対処が必要**とされる．しかし，国内におけるこの分野の研究はあまり進んでいるとは言えない．一方，実査場面では，既にマルチメディア対応を含めて様々な質問形式を取り入れた調査票が用いられている．

一方，回答者側から見ると，Web 調査の調査票は，PC の画面上を見る限りは回答の仕組みが従来の質問紙による方法（P&P 方式）と同じように見える．しかし実査側から見ると，調査票の質問設計や回答取得の仕組みはかなり異なると考えねばならない．

筆者等が行ってきたインターネット調査に関する実験調査でも，調査票質問の形式として，様々な様式を取り入れた試みを行ってきた．たとえば，コンボボックス，プルダウン・メニュー，イメージ（静止画，動画），あるいは他の Web ページへのリンクボタンの設定などを用いた．しかしながら，インターネット関連技術の水準や，調査回答者側のコンピュータ環境とリテラシーのバラツキ，また接続回線環境のバラツキなどから，必ずしも望ましい調査結果ばかりではなく，多様な質問様式を用いることには，慎重であらねばならないとの結論を得ている［たとえば，松田・大隅（2003）］．

もちろん，Web 調査の質問作成の特徴である，ラジオ・ボタン，チェック・ボックス，自由回答記入のテキスト・ボックスやテキスト・フィールドなどは積極的に用いてきた．しかし，調査票の基本的な形式の保持，つまり，回答者が調査票を視認したときに，**従来の質問紙による調査票に比して極端な違和感がないような設計方法**を用いることが，とりあえずはリスクが少ないと考える多数の根拠を実験調査の結果は示している．

たとえば，

① 従来の質問紙型の調査票に近い形式を用いること
② なるべく質問全体が画面内で一覧できること
③ 回答者の利用するブラウザの種類やバージョンを考慮すること
④ 回答者の利用マシーンの画面サイズや解像度を考慮すること
⑤ 質問の選択肢の選択状況がはっきりと視認できること
⑥ 再回答，選択肢の選択変更が可能なこと

⑦ 強制的には回答誘導を行わないこと
⑧ 調査票のダウンロードや回答の負荷があまり大きくならないこと

等々である．こうしたことをあえて列記した理由は，Web 上での調査票設計では，これらの諸事項に抑制を課して，**強制的に回答を誘導することも可能**であることを意味している．たとえば，分岐質問で，条件選択の後，分岐の条件が論理的に一致しないから，回答の再エントリを要求する，回答を行わなかったとき，それを警告し必ず回答が得られるようにし向ける（制御する），などが可能である．このように，Web 調査特有の機能として，実査側で**「回答が制御可能である」**ということがある．これをどう用いるかは慎重であらねばならない．

従って，Web 調査では，画面上で見る調査票の形式が，従来の紙形式のそれに似てはいても，その背景で電子的に回答の取得を制御する仕組みが組み込まれているという点で，従来型質問紙による自記式調査とは異なるものである．

つまり，調査票の設計，質問文の設計などの研究を進めるにあたっては，こうしたインターネット調査特有の技術的要素を十分に念頭において，対処せねばならない．インターネット調査の世界しか知らない者にとっては，従来型調査との間に見られるこの大きな乖離，ギャップが，回答結果にどう影響するかが良く理解されていない（つまり**測定誤差の評価**に大きく影響する要因であるこれを看過している）．

多くの場合，テクニカルな方法論にばかり注目し，調査票への回答者の回答行動をどう考えるかの本質的な議論が欠落している．たとえば，質問形式として，コンボボックスやプルダウン・メニューを用いると画面内に占める調査票のサイズが低減できるとか，マトリクス形式を用いると，多数の選択肢を用意できて情報が増える，イメージ（静止画，動画）や音声を用いて訴求力や関心を高める，テキスト・エリアを多用して豊富な自由回答が得られる等々が喧伝される．しかし，これらを用いた際の回答結果に及ぼす影響評価については，ほとんど議論されることはない．

しかし，従来型調査における自記式の場合を考えても，回答者の様々な回答行動が指摘されてきたわけで，インターネット調査だからといって，これらが生じないということにはならない．ましてや，回答を調査実施側で意図的に抑制や制御する等の方法には疑問こそあれ，これを正当な方法として受け入れる客観的な論拠はない．ともあれ，一般に，Web 調査における質問作成には，以下のような特徴があるとされてきた．

①様々の表示形式が利用できる
　例：ラジオボタン，チェックボックス，プルダウン・メニュー，コンボボックス
②レイアウトとしても様々な様式がある
　例：マトリクス形式，一括表示，段落形式，…
③とくに**自由回答入力の設定が容易**とされる
　例：自由回答入力用のテキスト・ボックス，テキスト・フィールド，
　　この様式も様々ある，文字数制限あり/なし，スクロールあり/なし，
　　別画面でテキスト・ボックスをポップアップ等々
④改ページ処理の有無
　例：改ページを行わない，いわゆる巻物方式か，改ページ方式とするか

⑤マルチメディア対応
例：イメージ（静止画，動画），音声，双方向通信の利用

こうしたことを考えると，回答者が Web 上の調査票を閲覧したときに，仮に見た目が従来の質問紙型とそっくりに見えたとしても，回答者の回答行動・動作に応じて取得される回答，つまり，実査者側からみた調査への回答行動には，非常に異なるものがあると考えるのが自然である．

このような調査票の設計や調査質問の設計は，**「測定誤差」「無回答誤差」の評価**に関わる問題である．とくに Web 調査では，調査票の発信者側（調査実施者側）と受け手側（回答者側）とで，調査票の認識・認知の条件が異なると考えねばならない．

調査実施者が作成したある一つの調査票も，回答者側のコンピュータ環境，通信接続環境，あるいは用いる PC の条件によって，見え方（スクリーン上のサイズ，フォント・サイズ、色、質問や選択肢のレイアウトなど）は様々であり，「同じ」調査票で調査をしたつもりが，実はそうはなっていない状況が起こる．こうした視認・見え方の差違が回答の違いとなって現れることが予想され，実際にこうした調査票設計が回答結果に及ぼす影響・効果測定の研究結果が多数報告されている［Couper(（2003），Couper 他，Dillman（2000)］.

たとえば，筆者等の実験調査では，一つの質問について，質問形式を，選択肢の長さ（長い，短い），選択肢形式（ラジオボタン，プルダウンメニュー，コンボボックス）が異なる状況を設定し，同一の回答者集団（同じパネル）に対して質問を行ない回答への影響（類似，差異）を検証したが、統計的に明らかに違いがあることが分かっている．また，同様の指摘は多数の研究報告にある（日本国内ではほとんど見られない；いくつかの例を後で示す）．ここらについては，以下を参考とするとよい．

［参考］

（1） 大隅昇（2002），インターネット調査，「社会調査ハンドブック」，pp200-240，朝倉書店.

（2） Couper, M.P. (2003)， The Internet and Other Survey Opportunities, JMRA 第 33 回トピックスセミナー「，2003 年 10 月 23 日開催，配布資料.

（3） JMRA 研修セミナー,「インターネット調査を検証する−質の評価と標準化に向けて−」, 配布資料, 2003 年 6 月 10 日～12 日.

（4） 参考文献に挙げた Dillman （2000），Grossnickle and Raskin (2001).など.

いずれにしても Web 調査研究の進んでいる欧米では，調査票はできるだけシンプルなものとすべきであるとされている．測定誤差，無回答誤差の低減という観点から，装飾の多い，構造が複雑な調査票を用いた調査結果の質は疑うべきであり，またどうしてもそうした設計方式を使うなら，事後の評価方式までを考慮した調査設計とすべきである．

６．データ解析上の課題

６．１　日本語の特徴と形態素解析

これについては，既に第 I 部で若干触れたが，ここでもう一度述べる．

言語類型学では，言語を孤立語（語形変化せずに，文法関係が語順で示される言語），膠着語（文法関係が助辞や接辞によって示される言語），屈折語（文法関係が語形変化によって示され

る語）などと分類する．これによると日本語は「膠着語（膠着言語)」である．また，主な欧米語は屈折語である．膠着語の特徴は「自立語」と「付属語」が膠着していることにある．「付属語」（あるいは「辞」）とは，助詞，助動詞，接尾辞，用言の活用語尾をいう．また，「自立語」（あるいは「詞」）とは「付属語」に付くものをいう．これはいわゆる「詞–辞」構造という（時枝誠記による「詞–辞」論；加賀野井（1995， 1998)）．

さらに，日本語が欧米の言語と大きく異なることの一つに，文章・テキスト型データが「べた書き」であって「分かち書き」されていないということがある．ただこの点では中国語など，アジア圏で利用される言語にもべた書きという共通した特徴がある．また，日本語は漢字，カタカナ，ひらがな，それに外来語（やそれに充てられた漢字や仮名等）が混在しているという特徴もある．言語による表記を「表意文字」と考えるか「表音文字」と考えるかという見方もあるだろう．実際に仮名漢字の混用である日本語は，この両者が現れる．

古くは万葉仮名，カタカナの誕生，漢字の読み替え・当て字による造語，そして明治期（以降）における造語現象（主として日本にはなかった諸概念の表現のための新造語，たとえば社会，自由，経済，生産，科学，真理等々多くの現代用語)，明治以降の言文一致体の登場等々，歴史的にみてもきわめて流動的である．特に，我々が現在日常的に用いている語句の多くが，**明治期以降の造語が多いという事実**に注意せねばならない．

別の問題として「べた書き」にどう対応するかがある．欧米語が「単語」という単位で仕切られた活用する言語（屈折語）であることから，その処理系がこれを単位として扱うことができ，活用や変化はあっても個々の単語を抽出する容易性がある．一方日本語はこれが困難であるだけでなく，複数の語が連結されて複合語を形成することが多い．

このために，日本語の処理を行うには，まずある単位に文章を分解する**分かち書き処理**が必要となる．これを含めて幾つかの処理を**形態素解析**（morphological analysis）という．形態素（morpheme，morphology）とは，表記された文章（日本語とは限らない）を「最小の有意義な意味ある単位，意味を持つ最小の単位」と定義されている（池上，1993)．これは池上によるとBloomfield（1933）によって唱えられた概念であるという．また，長尾他（1998）によると，形態素とは「単語や接辞など，文法上，最小の単位となる要素のこと」としている．ここで，両者には明らかに考え方（解釈）にわずかな相違がある．このように形態素とは絶対的な概念ではなく，あくまでも一つの便宜的な約束事である．たとえば，池上によれば，形態素が示す矛盾を説明する概念として「語彙素」(lexeme）を挙げて両者の特徴を指摘している．いずれにしても，形態素，語彙素ともに「語」(word）とは必ずしも一致はしないし，言語学的にも確定的な概念があるわけではない．語句を扱ううえでの必要の目安と考えればよい．

しかし，日本語のデータ解析処理においては，何らかの意味である単位に分けねばならない．そこで，通常は分かち書き処理で得た単位要素の候補を辞書（形態素辞書）と照合し，次にそれを解釈可能な候補に絞り込み，文字・文章の文法的な接続関係（word connection）を検証する．そのうえで，その分かち書き単位についての**品詞の同定，特定化**を行い，続いて辞書にはない語の処理を行う等の手当をする．このような一連の過程が形態素解析である．従ってその処理にはかなり発見的あるいは経験的な要素が含まれることとなり，実際にいろいろな解析方式が提案されてきた．

たとえば，**最長一致法（N 文節最長一致法)，字種区切り法，文節数最小法，コスト最小法，接続規則法**等がある（全文検索システム協議会編（1997）ほか)．いずれにせよ，日本語処理

の初めの処理過程として「分かち書き処理」と「辞書照合」の操作が不可欠となる．ここらの技法は，日本語ワードプロセッサの「かな漢字変換ソフト」にも関係する．

また，どのような方式を用いても**ノイズの混入**は避けられない．最近はコンピュータの処理機能の進歩のおかげで，力仕事でこの処理がかなり可能となってきてはいるし，様々な工夫がなされている．しかしながら，調査における自由回答の場合は，質問（説明文）の内容や主題をかなり絞り込んでも，得られる記述の内容や記法・表記が乱れることが一般的であり，現状の技術力では単語や語句を確実に同定できるか十分には期待できない．

なお，自然言語処理・計算機言語処理などでは，まず表記の構造を形態素解析，構文解析により確認し，続いて意味的なアプローチから意味解析，意味理解といった操作が行われる．いずれにしてもコンピュータ処理の支援は避けられない．しかも一般には相当量の計算処理時間を要する．

このように，いわゆる自然言語処理的な観点に立つと，その要素技術は言語学的というよりも，きわめて工学的な考え方や研究が多い．またこのような処理形態が，実際に人が行う言語処理行動（回答行動）に合っているか否かは，現時点の研究だけでは説明できるものではないし，ここで述べる考え方とは少し方向が異なるものである．

加えて，言語学的観点からは，日本語は未熟あるいは流動的な変化や変容が日常的である，その意味で言語（学）研究そのものが発展過程にある．同時に，欧米型のテキスト・マイニングの考え方やツールをそのまま取り入れても，適切な（日本語テキスト型データの）解析ができるとは限らない．

この他，**日本語の曖昧性**（本当に曖昧かどうかの議論もある），デノテーション（語の明示的な意味，表向きの意味）とコノテーション（語の言外の意味，含意，言外の意味），助詞「テニヲハ」の考慮，カテゴリー論との関係，最近話題となっている認知科学的なアプローチからのメタファの重要性等，「日本語の構造的な特徴」を巡る諸研究や議論が無数にある．さらに，単純に電子的操作・処理法との関係で見ても，ワードプロセッサの登場による表記法の変化やインターネットの利用下における E-mail 用語（専門語），E-mail 語，チャット語，フェイスマークさらに携帯電話用語（ケータイ語）の登場と，日本語の様相は目まぐるしく変化している．

６．２　統計的データ解析の観点からの接近法

こうした日本語の言語的な特質を考え，調査における自由回答データの分析を行うには，自然言語処理や言語情報処理で行われてきたような発見的，計算アルゴリズム的なアプローチによる処理方法から少し離れて，より実用的な観点から統計的データ処理のパラダイムの中で考える．つまり，従来からある**形態素解析と統計解析（とくに多次元データ解析）の諸要素技術の部品（手法）**を適当に組み合わせることで，従来の個々の方法論では解決できなかった調査分野のデータ解析手法としての新たな接近法を構築することを考える．また問題を単純化して，次のように考える．

(1)　テキスト・ファイル化した自由回答文・テキスト型データ等を**構成要素**（fragments）に分解する．これを「分かち書き」処理により，たとえば「単語や文節」（のようなもの）に分解する．つまり，形態素解析の要素技術のうちの「分かち書き処理」の機能だけの援用を受ける．ここで構成要素と断るわけは，分かち書き処理で得た結果の個々の単位

要素が，仮に言葉として意味を持たない場合でも，（統計）解析処理の対象としては，そのまま扱うこともありうる（可能）との観点から考えようというものである．つまり，ここで，構成要素とは単語・語と区別するために用いるある曖昧な概念，データ処理の単位をいう．

(2) これから導かれる構成要素の出現頻度のパターン等の解析を行う方法として考える．通常は，「出現頻度の高い語は重要である」あるいは「頻度の近い位置にある語は関連性が高い」といった経験的なルールを用いることが多い．しかしここでは，分かち書き処理で得た「構成要素」の並びという程度に考える．

① 前述のように日本語には「分かち書き」の考え方はない．また形態素解析の確定的な方式は未だあるとは言えず，それだけに流動的である．そこで，この操作はむしろ事前処理・中間処理として利用する．

② テキスト化された日本語文章を何かの意味で「分かち書き」した各単位，つまり「構成要素」に適当に分解するという程度の緩やかな約束でよいと考える．

③ しかし，分かち書き処理をどう行ったかの過程が明示的に分かるように努める．

④ 構成要素を複数結合した場合を「文節」と呼ぶことにする（文法で言う文節より緩やかな意味）．

分かち書きを緩やかな決まりとする理由は，元々の取得データ自体が曖昧かつ多様な表現であるから，むしろそれを許容して，厳密な定義や拘束を避ける方向で分析を進めるという視点に立つという意味である．たとえば，以下のような理由がある．また，このことが具体的な解析システム（WordMiner）開発時の設計指針に反映されている．

（1）日本語の精密な言語学的研究が目的ではないこと

研究対象とする内容は，元来ノイズが多い自由回答・自由記述等の解析を目標としている（**非構造的なデータ**を対象とする）．**自由回答をいかに科学的に取得し統計的な処理を可能とするか**が重要であり，個々の記述内容の意味論的な分析や言語学的な構造の研究が主たる目標ではない．

現時点でこれに関わることは，そもそも「（近代）日本語」とは何かの解釈の根幹に関わることでありきわめて難しい課題である．しかも，近代日本語の歴史自体が浅く，言語学的，日本語文法的にも明らかでないことが多すぎる．加えて，外来語や新造語の混用が特徴であり，こうした範囲までを言語学的手法によるアプローチで可能とは限らない．

（2）得られるデータに曖昧性があること

更に重要なこととして，アンケート等で取得される自由回答データは，そもそもその**表記法や記述内容に曖昧性**（ambiguity）があり，整った文章が得られるとは限らない．これを表現が豊か，柔軟性があるとする言い方もあるが，それだけ表記内容に自由度・曖昧性が高く意味を捉えにくいとも考えられる．とくに，Web 調査などではカタカナ語，欧米語の氾濫現象がある．また，続々と新語やカタカナ語が増える傾向にある．回答者の表記法・表現法もまちまちである．

（3）他の利用方法との関連

分かち書き処理を行った文章の解析だけではなく，単純なキーワード抽出で得られた「語句の列記」のデータ等も扱うことがある．また，従来から自由回答処理の方法として利用されてきたアフター・コーディング処理などとの併用や比較検証も必要となることがある．さらに，“意図的に”（目的に応じて）テキスト・データを再編集して，解析結果を相互に比較するという利用方法も考えられる（あるデータセットから得られる答え，解は一つとは限らない）．

（4）類型化による規則性の探査と個別意見・回答別意味の把握

集積した自由回答・自由記述データの中に潜在する構造の類似性・差異性や規則性等を知ることは重要な目的である（マイニングの目標である）．このために，探索的な多次元データ解析手法が有効である．とくに，**個々の回答・記述の意味内容や意見の規則性や類型**を知ることが必要となる．しかし同時に，解析から得た「類型」に含まれる「個々の回答データの特徴」を読みとることや，類型で得られた典型や大勢の回答傾向だけでなく，**少数例・少数意見の特徴**も知りたい．つまり，単なる文章要約や分類だけでは十分ではなく，意見の類型化とその内容分析（解釈）が必要となる．

（5）従来の定量的調査法の理論の援用を受けること

自由回答データの特徴の一つに，選択肢型質問や属性などで得た数値データのように定量的に統計値として評価できないということがある．通常の選択肢型質問を例にとれば，回答比率データを算出したり統計的な検定の操作により標本誤差を検討したり質問間の差異を比較検証することが可能である．しかし，自由回答データの場合，こうした操作が難しい．しかし調査結果に何らかの保証を与えるためには，間接的ではあっても従来の標本調査の理論や知識の援用を得たいこと，あるいは比較可能となっていること，つまり定量的操作との併用が，自由回答の解析に妥当性を付加する措置として必須であると考えられる．このことから，従来型の選択肢型質問項目や属性項目などと自由回答質問とを併用し，これらのデータの相互関連性の分析が重要である．自由回答の分析結果に加えて，これらの項目との相互検証を可能にする集計評価機能が重要と考えられることがある．つまり，**定性的調査と定量的調査の併用**が必要と考えられる．

（6）調査法，調査方式の類似・差異の検証が必要であること

さらに重要な事項として，繰り返し指摘するが，**調査法，調査方式（調査モード）等の影響評価を考慮した自由回答取得の調査設計**が重要である．すなわち，得られた自由回答の分析結果，見られた特徴から，どのような意味解釈ができるかは，同時に，「用いた調査方式に依存」しているということである．標本設計・標本抽出はどう行ったか，調査方式として何を用いたのか（Web 調査，郵送調査，面接調査等々），調査票作成，質問形式はどう設計したのか（電子調査票，P&P 方式，面接員の記入，録音等々），何をどう用いたかの影響評価が可能な調査計画（実験計画）を工夫せねばならない．

ここで，調査データの解析，**自由回答質問の分析にとっては，とくに（4），（5），（6）は重要**である．またこれが調査データのテキスト・マイニングが当面目指す方向と考える（現状

では，こうした接近法が合理的かつ無駄がないだろうということ）．

７．テキスト型データ解析システム：WordMiner®

以上のような目的（解析・分析の指針）を達成するためにはそれに適したコンピュータを用いた統計システムの開発が必要とされる．これのために開発されたシステムが WordMiner である．これは，日仏他の研究者を中心に開発された SPAD.T（Système Portable pour l'Analyse des Données，Donnée Textuelles）を基本エンジン部とし，これに分かち書き処理，辞書編集機能他の日本語解析に必要な機能を追加した統計システムであり，長期にわたる日仏共同研究の成果の一つである．

（注）WordMiner の前身である InfoMiner は商標登録第 4387759 号（第 9 類：電子応用機械器具及びその部品）を取得している．WordMiner も同じく登録を取得している．

従来の多くの類似ソフトが形態素解析に始まる一連の言語情報処理的な視点から開発されてきたことと異なり，調査環境下において取得したテキスト型データに発生しうる状況を考えたデータ重視・実践型機能を実装した**記述統計解析・探索的解析**を設計指針とすることが特徴である．とくに，選択肢型質問・属性データを併用する自由回答型を含めた調査データ解析に適している．なお，WordMiner と類似の機能を備えた，とくに調査データの分析に特化したソフトウェアはあまり例がない．一つの例として，フランスで開発された“Sphinx Survey: Plus2 & Lexica”（Sphinx Développement (1998)）があるがこれには当然日本語処理機能は含まれない．

ただ，最近の傾向（ここ２，３年の傾向）として，国内のテキスト・マイニング・ツールも，多次元データ解析などの手法を前向きに取りいれる傾向にある．（第 I 部を参照）．

７．１　WordMiner の設計指針

WordMiner の設計指針はきわめて単純である．SPAD.T から WordMiner に至る開発経緯とそのシステムの詳細を述べることは別の機会に譲るが，現状の WordMiner となるまでに十数年を経ていることだけを指摘しておこう．ここでは前述の日本語の特徴（現時点での利用体系）を考慮したうえで，分かち書き処理機能と統計解析機能を違和感なく接続利用できる利用環境の実現を目指している．回答間，構成要素間それぞれの間の回答パターンの類似性や，回答と構成要素の間の関連性（対応）の理解に役立つ知見を提供するためにはどうあるべきかという考え方が背景にある．

（１）システムの主な特徴

基本操作は，元のテキスト型データを構成要素に分解し，構成要素（単語や文節）と回答（たとえば被験者，回答者，著者，検体），その類型化情報との相互の関係をパターンとして表現し分析することにある．これに対して，計算処理上の工夫が必要とされるので，これへの手当を行う．

① 日本語独自の事前処理（分かち書き処理，その初等統計処理），**初動探査の機能**が含まれる．
② 一般の文章データの分析も可能である．

③ 平易な**多次元データ解析手法**を使っている．既に統計的方法論としての実績のある，対応分析法，クラスター化法およびそれに関連した**基本的な統計処理**を用いる．これは解析内容の透明化を図るためである．
④ 通常の**選択肢型質問・属性データとの併用分析**を行う．
⑤ 数値計算処理上の工夫が必要となる．
⑥ 扱うデータ行列の要素がきわめて「疎」となる事（次元数の大きいデータ表の処理），はずれ値への手当が必要であること，データ表の寸法が不定であること等への対策（構成要素数が確定しないと行列の大きさが確定しない），大量データの分類操作が必要となることなど
⑦ その他，辞書編集機能（不要文字の削除，類似文字，類語・同義語など）や登録辞書の再編集機能等の手当を要すること

（2）システムの動作環境の概要

全文検索等を行うソフトの大半が，ワークステーションや汎用大型コンピュータ，あるいは場合によっては（たとえばデータ・マイニングのツール），並列コンピュータなどの利用を必要とする．しかし，調査データの解析ではなるべく可搬性を考慮して PC で利用できることが重要である．WordMiner は OS が Windows 対応の既存の PC で十分に利用可能である．最近は PC レベルで利用可能なテキスト・マイニング・ツールの利用環境も整いつつあるが，WordMiner は当初からデスクトップ上での利用を前提として開発された．

７．２　WordMiner の主な機能

（1）基本的な特徴

① データの基本形式は，**（回答・サンプル）×（変数・項目）の多変量型データ**である
② 変数（項目）は，テキスト型データ（自由回答質問など）の他，文字データ，数値データを扱うことが可能
③ 変数名の長さと扱える変数の個数（変数の数）はほぼ無制限
④ テキスト型データの長さも原則として無制限
⑤ 英語，仏語などの欧米語が扱える
⑥ アフター・コーディングなどの加工データの分析も可能
⑦ 別の分かち書き処理ツールで分かち書き処理を行ったデータの分析も可能
⑧ ユーザが辞書作成可能，また一度作成済みの辞書を共有化できる（辞書の使い回しが可能）
⑨ 分析，解析結果の多くは，テキスト・ファイルとして外部出力が可能である，これを用いた他の統計ソフトウェアや表計算ソフトなど（(JMP，エクセルなど）による再分析，二次分析が可能である

（2）分かち書き処理機能

日本語文章・テキストを電子化したファイルに基づき，分かち書き処理を行い，統計解析に必要なデータセット・ファイルを生成する．WordMiner では分かち書き処理機能として Happiness（平和情報センター）を WordMiner 用に改良した WinAiBASE を採用している．これを用いて自動的に分かち書き処理が行われる．このとき，分かち書き情報，キーワード情報，

場合によりそれらの計数ファイル（回答別の分かち書き数，キーワード数の計数ファイル）を得る．また，パラメータ変更による処理条件の設定，検索・置換，編集，更新等の機能がある．

（3）辞書の作成管理機能

①辞書作成機能：構成要素辞書生成

②辞書の編集機能

- 除外文字・記号の編集
- 構成要素の最小出現度数の閾値設定の制御
- 解析から外す構成要素の指定
- 置換・読み替え等の単語の指定編集（類語・同義語の編集）

③辞書の更新

- 編集指示に従って辞書を更新

④使用する構成要素辞書の指定や他の課題での再利用（共有化）

（4）解析部機能

①「(回答）×（構成要素)」表の分析

尤も標準的な分析を行うモジュールである．抽出した構成要素（文字列）と回答（たとえば回答者，サンプル）の関連表の多変量解析による情報の縮約を行う．

（注 1）ここで「構成要素」とは，分かち書きで得た緩やかな意味の語句・単語などの集まりのことを言う．

（注 2)「追加処理」(supplementary treatments) を指定するオプションがある．これについては大隅他 (1994) を参照.

②「(構成要素）×（生成クラスター)」のクロス表の有意性テスト

クラスター化で生成したクラスター番号は一種のカテゴリー変数である．これを用いた有意性テストを行う．またクラスター化で，構成要素のクラスターあるいは回答者のクラスターが生成される．

（注）自動分類による「教師なし分類」に相当する．分類結果の情報から二次分析を行うこともある．

・「構成要素出現頻度」を用いたクラスター別の出現構成要素頻度　の有意検定機能（クラスターに有意な単語を知る）

・「カイ二乗距離」を用いた有意検定の機能
（生成した各クラスターに寄与する回答パターンを知る）

・有意として選出の回答（回答者，サンプル）のリストを出力する機能

③「(構成要素）×（質問・属性)」表の解析

抽出した構成要素と，予め用意した「選択肢型質問・属性」データセットの個々の変数との

クロス表の出力，有意性テスト，生成クロス表の対応分析，単語および回答のクラスター化等を行う機能である．つまり**「構成要素と質問・属性間との関連を知りたいとき」**に用いる機能である．たとえば，抽出構成要素と「性別・年齢区分」の変数とがあって，これらがどう対応するかを知りたいとき等に用いる．比較する質問がそれぞれ類似の内容であれば，回答間の相似性や回答の均一性の有無についての探査も行う..

④「(構成要素) × (質問・属性)」クロス表の有意テスト
・抽出した構成要素と「選択肢型質問・属性」データセットの個々の変数との関連表の出力，有意検定，生成データ表の多変量解析等を行う機能
・「(構成要素) × (質問・属性)」表の有意テスト，どの質問・属性が構成要素の説明に寄与しているかの検証.

(注) これは，自動分類に対して，選択肢や属性を分類尺度として利用する，一種の「教師あり分類」と考えることができる.

⑤構成要素の用語検索（コンコーダンス；concordance）機能
コンテンツ・アナリシス（内容分析）の基本操作である KWIC/用語検索（**コンコーダンス**：ある指定した単語が与えられた文章・回答の中でどう使われたか）の一覧を，指定した単語を基準に検索・ソートし出力する．これを用いることで，語句の簡単な共起関係を視認できる.

(注 1) KWIC: Key-Word-In-Context の略
KWIC リストの一部としてコンコーダンスを用いる．この他，クロス参照（cross-reference），KWOC(key-word-out-of-context)，出現単語頻度表などを用いる.
(注 2) ②，④でいう有意性テストについては，事例ならびに Lebart 他（1998），大隅他（1995）を参照.

７．３　多次元データ解析の特徴

前述の機能に合わせて種々のデータ表を対象とした**多次元データ解析（対応分析，クラスター化）**が可能であるが，出発時のデータ表の基本形は**二元のデータ表**である．たとえば“クロス表”を考えればよい．このとき，このデータ表の表側と表頭に何を対応させるかで，様々な解析が可能である．たとえば，以下あるような分析をオプションとして利用可能である.

① 回答者の回答パターンと構成要素群（単語群）の関係を分析
② 回答者のクラスター化情報と構成要素の関係を分析
③ 回答者のクラスターを意味づける構成要素群を知り，典型的な回答者例を有意性チェックのもとに表示する
④ 選択肢型質問や属性情報のうち，どれが構成要素と関連しているかを知る．また，その選択肢別の構成要素の有意性を知り，その典型的な回答者例を有意性チェックのもとに表示する
⑤ 回答者のクラスター化情報と構成要素のクラスター化情報の関連を知る（二次分析）

（注）以上の解析方法の概略については「補足資料」の部を参照のこと．

８．簡単な初等統計分析の例

ここでは，初等的な記述統計解析を通じて見られる自由回答の分析結果の例をいくつか示すことから始める．記述統計解析は，実はテキスト・マイニングの初期段階で，重要な役割を果たすと筆者は考えている．しかし，残念なことに多くのテキスト・マイニング・ツールはここらの処理系が弱い．また，何が行われたかが不透明であることもある．たとえば，その典型的な例が，**分かち書き処理結果の集計・分析処理**である．「どのように分かち書きが行われたのか」は，以後の解析の出発点であり，その内容を明示的に知ることが必須であるが，ここだけをみても，多くのソフトは曖昧である．

以下に，筆者が関連してきた調査，あるいは我々研究グループが産学協同研究として進めてきた一連の実験調査の自由回答質問の中から幾つかの質問を取り出して，これに対する回答データの傾向を調べる．我々の各種実験調査では，これらを含む様々な結果が得られているが，ここではその一部を要約する．

８．１　調査方式，調査票・質問方式の検討

まず，調査方式（調査モード）や用いる調査票，そして質問内容により，回答者の自由回答への回答行動がどう現れるかを簡単な例でみる．

（1）Web 調査と郵送調査の比較実験調査

始めに，Web 調査と郵送調査の比較検証をあげる．

① 調査目的
- Web 調査と郵送調査の自由回答質問の比較検証，とくに調査方式間と質問方式間の比較
- メーカーのブランド・イメージを選択肢型質問（定量的評価）と自由回答質問で取得し比較分析する

② 調査実施機関
- （株）東京サーベイ・リサーチ

③ 調査方式サンプル数
- Web 調査と郵送調査の混合方式（mixed-mode）
- 計画サンプル数：Web 調査（873 サンプル），郵送調査（576 サンプル）
- 回収サンプル数：Web 調査（228 サンプル；），郵送調査（351 サンプル）
- 回収率（有効回答率）：Web 調査（26.1%），郵送調査（60.9%）

（*）Web 調査の回収率がきわめて低いことに注意

④ 調査票と質問
- 調査票は Web 調査と郵送調査で，できるだけ書式／レイアウトを揃えた
- 選択肢型質問，属性項目を含めて 20 問
- 自由回答質問，21 問

（*）自由回答質問数が非常に多いことに注意

(*) 質問形式としては，テキスト・ボックス，連記型（5 項目，10 項目連記）を用いた．また，テキスト・ボックスにイメージ（ブランド名ロゴ）を添付した例も用いた．

⑤ 実験計画（自由回答質問について）

- 調査方式：Web 調査，郵送調査
- ブランド・イメージ：3 社，企業ブランド名／SONY，NEC，Apple；商品ブランド名／Vaio，iMac）
- 質問形式：テキスト・ボックス，連記型
- 以上の組み合わせ：（調査方式）×（ブランド・イメージ）×（質問形式）

⑥ 調査実施時期

- 1998 年 12 月～1999 年 1 月にかけて，Web 調査，郵送調査をほぼ同時期に実施した．

（２）分析例–その１　総単語数，異なり単語数，異なり率（異なり単語率）の観察

まず，21 の自由回答質問の自由回答質問の形式を要約する．なおここで，企業ブランド名，商品ブランド名の割り当て情報は省略した．

表４　質問の内容と質問形式（要約）

質問の内容	質問形式
Q1-1：ブランド名の印象	テキスト・ボックス（A）
Q1-2：商品／事業想起	連記型（5 項目）
Q1-3：ふさわしい言葉（形容詞などを用いて挙げる）	連記型（10 項目）
Q6：企業の長所	テキスト・ボックス（B） (*) ボックス並置／一部イメージ貼付
Q7：自宅のパソコンの購入目的と購入の経緯	テキスト・ボックス（A）
Q3-8：自宅のパソコンの使用状況	テキスト・ボックス（A）
Q6：パソコンの達人のイメージ	テキスト・ボックス（A）

(i)　総単語数と異なり単語数の関係

21 の自由回答質問について，それぞれについて分かち書き処理を行った後，総単語数，異なり単語数，異なり率（異なり単語率）を算出する．これを図に表すと図２のようになった．これを見ると，以下の特徴がある．

① 調査方式（Web 調査と郵送調査）の違いは見かけ上はあまりみられない．
② テキスト・ボックスと連記型の違いは明らかにある．これはその記入方式の違いから明らかである．また，テキスト・ボックスのバラツキが大きい．
③ とくに，左下から，連記型（5 項目），連記型（10 項目），テキスト・ボックスと書き込み量が増える．予想される特徴である．
④ ブランド名間の差異は，あまり見られない（図は省略した）．

⑤ 総単語数が増えると異なり単語数も増えるが，次第に伸び方が低減する（指数的に変化する）．つまり（サンプル数が同じ状況で）書き込み量・記入量が増えると，それは頭打ちとなると傾向にある．これもある程度予想されることで，書き込み量が増えても，ある質問についての回答内容は，いずれはある量に収まることを示している（次の異なり率との関係で考える方がよい）．

ここで⑤の特徴は，様々な調査方式を用いた結果から求めた集計結果からも，同じ様な傾向を示すことが分かってきた．［別例を呈示できる］

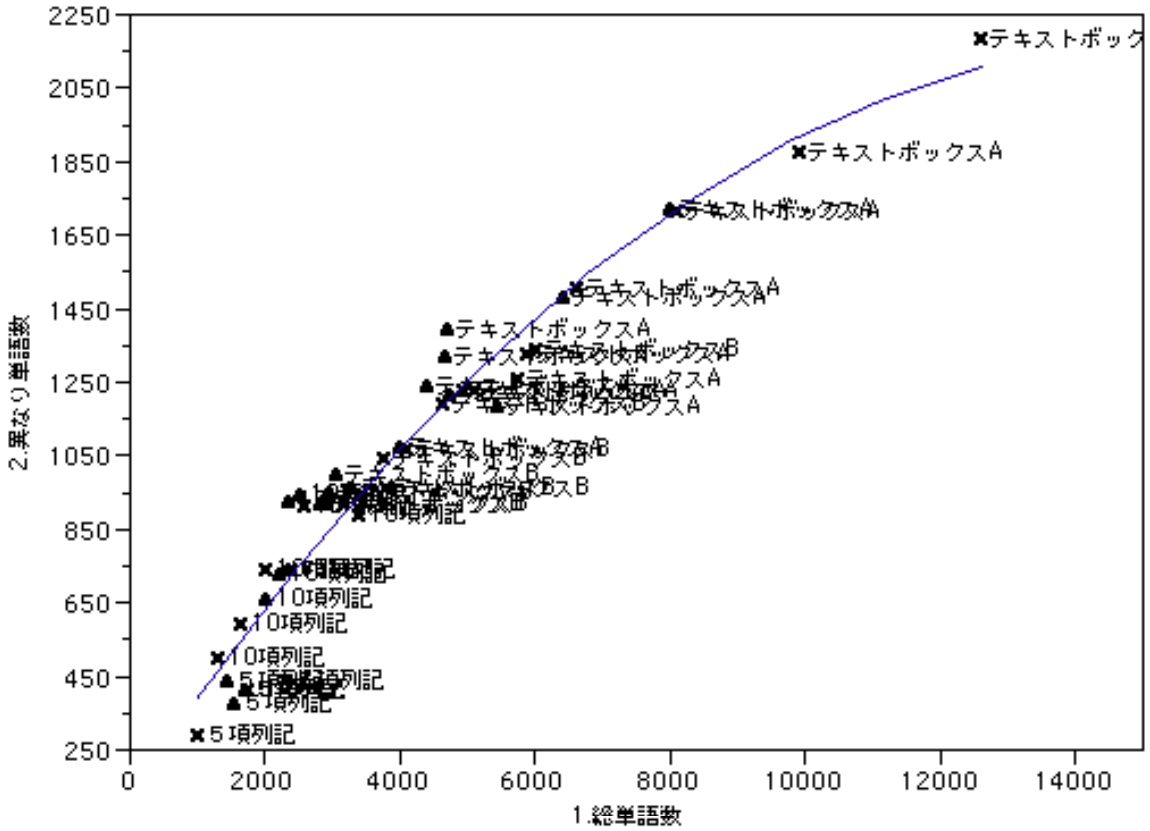

図２　総単語数と異なり単語数の関係（▲：Web 調査，×：郵送調査）

(ii)　総単語数と異なり率（異なり単語率）の関係

では，総単語数と異なり単語率［＝（異なり単語数）÷（総単語数）の%値］の関係はどうであろうか．これを図３で示した．

これを見ると，連記型とテキスト・ボックス形式との関係がより明らかになる．

① こちらは調査方式の違いが顕著である．
② 連記型は，総単語数の分布範囲が限定され（書き込み枠の制限がある），しかし異なり率には変動（バラツキ）がある．つまり，回答の内容に変化がある．
③ Web 調査の方が，郵送調査よりも，若干であるが総単語数の分布範囲が広く，また異なり率のちらばりもわずかに大きい．

④ 異なり率は，自由記述の内容のまとまりの程度（意見の発散度）をみる指標として，あるいは**語彙の潤沢度**（richness of vocabulary），**語彙量**のように言われてきたが，必ずしもそうはならない．つまり，意味内容の潤沢さには関係なく，総単語数が増えれば必然的に異なり率が下がるという性質があると思われる（そう考えることが自然である）．語彙量との関係については，さらなる検討が必要である．

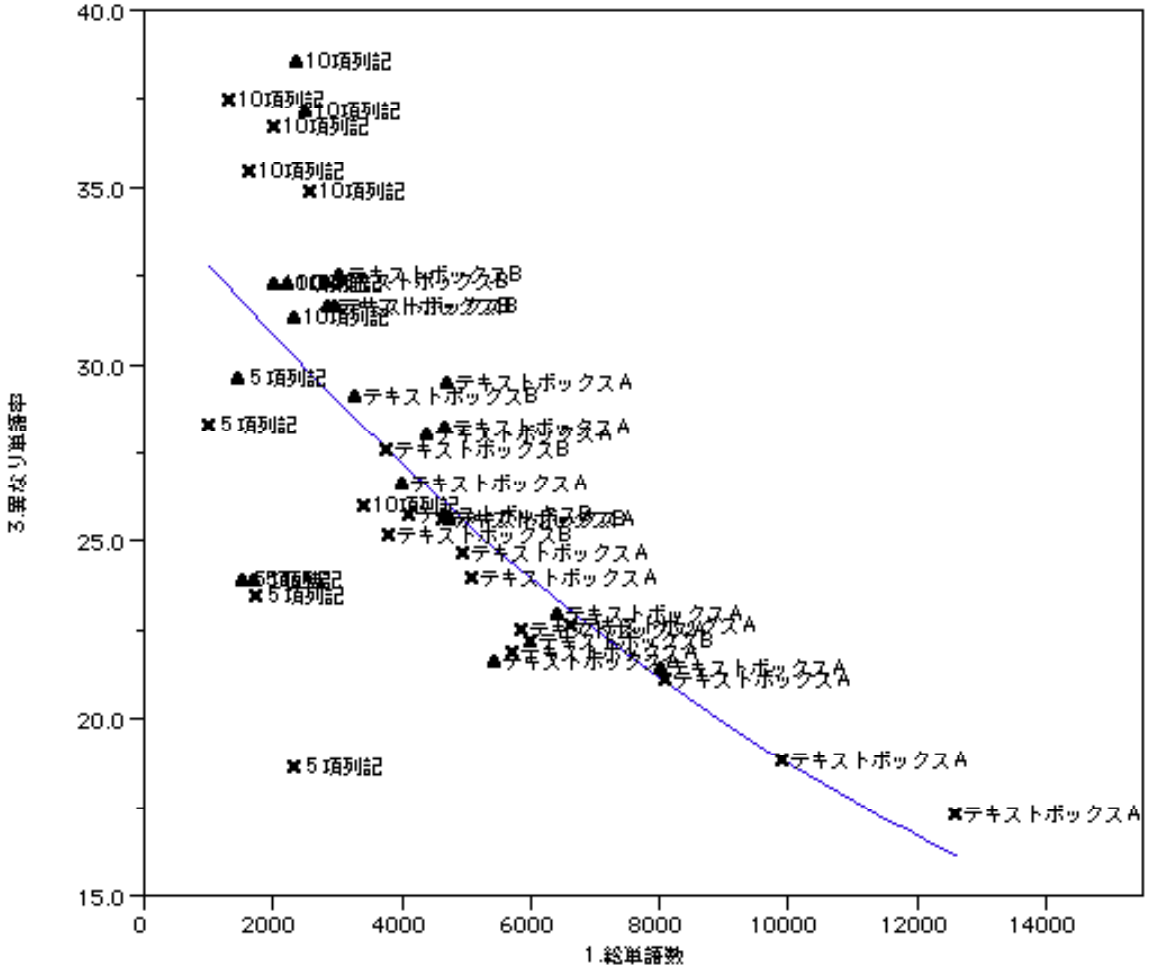

図３　総単語数と異なり単語率の関係

（３）分析例−その２　単語語長の分布の特徴

調査方式の異なる２種の調査（Web 調査，郵送調査）のサンプル数の違い，回収率の差違，自由回答の質問別の回答数の違いなどを考慮して，また，質問内容によって回答の長さ（分布）が異なることなどを考えると，別の統計量を求める必要もある．ここでは，回答者別の単語数の平均単語数（平均語長）とその分布の変動係数（C.V.%）を算出してみた．

まず図４は，平均単語数（平均語長）について，図５はその変動係数（C.V.）について，そして図６は異なり率について，それぞれ **Web 調査と郵送調査で比較**したものである．テキスト・ボックス（２種），連記型（２種）をそれぞれ楕円で括ってみたが，明らかに顕著な特徴がある．

① テキスト・ボックスについては，質問形式ではなく質問内容によって回答記入の量が異なる（自明なことがそのまま見える）．
② 調査方式の間の平均単語数の差異は，あまり見られない．
③ しかし，変動係数の分布をみると，明らかに Web 調査が郵送調査よりもはるかに変動係数が小さい．つまり，回答の集中度が違うと思われる．書き込み方の相対的な変動が，郵送調査は Web 調査に比べて大きく，明らかに特性の違いがある（調査方式の差異がある）．
④ また異なり率の図からは，Web 調査の方が高い傾向にあることが分かる．
⑤ 以上から見るように，調査方式間の差異と，質問形式別の回答の出方（傾向）には，明らかな違いがあることが見えてくる．

このようなごく簡単な例からでも，始めに指摘したような Web 調査が他の調査法と比べて，自由回答取得の方式として優れているといった論証は簡単には得られないし，また調査方式，質問形式に依存して状況が変わるという常識的な結果が見えるのである．

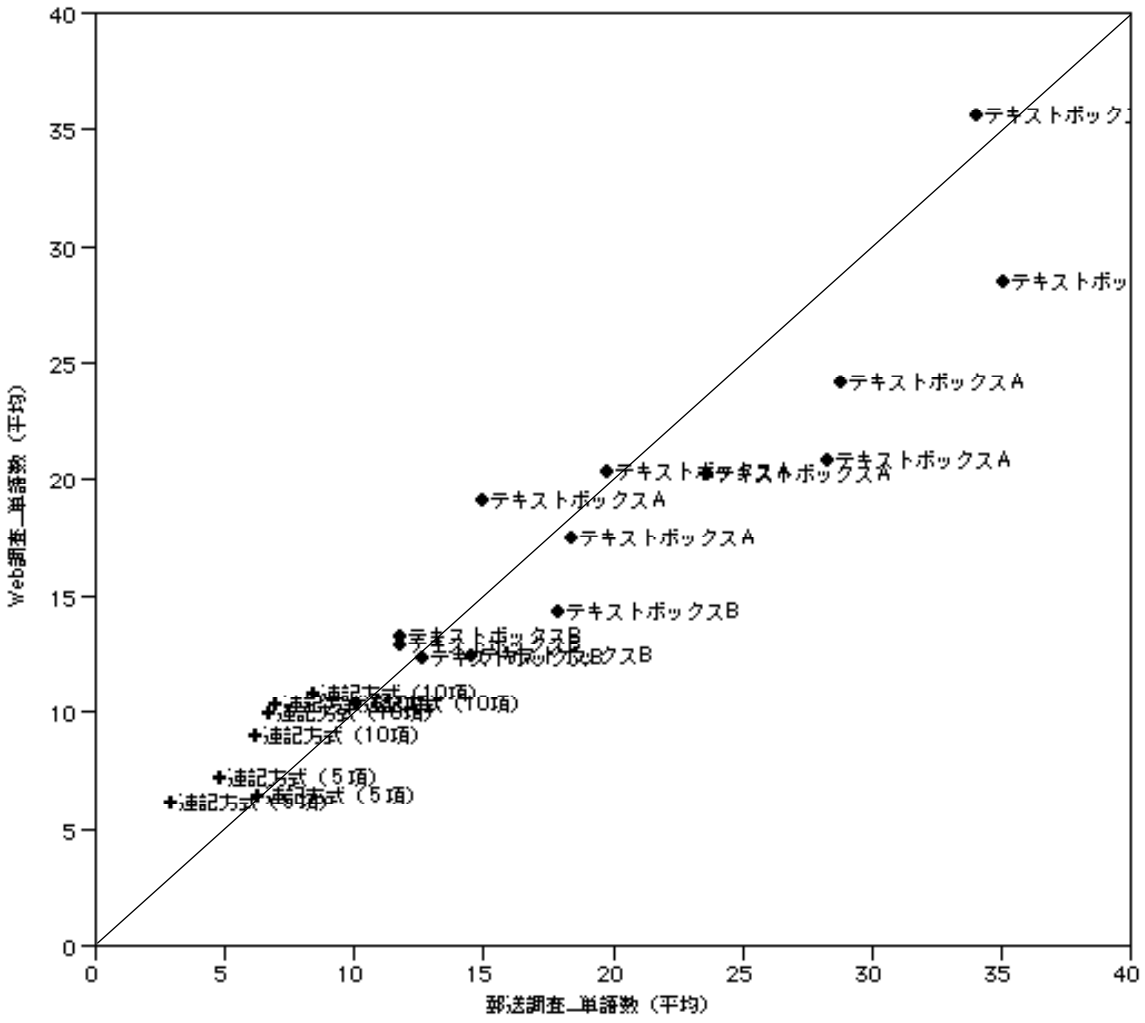

図４　Web 調査と郵送調査の平均単語数（平均語長）の比較

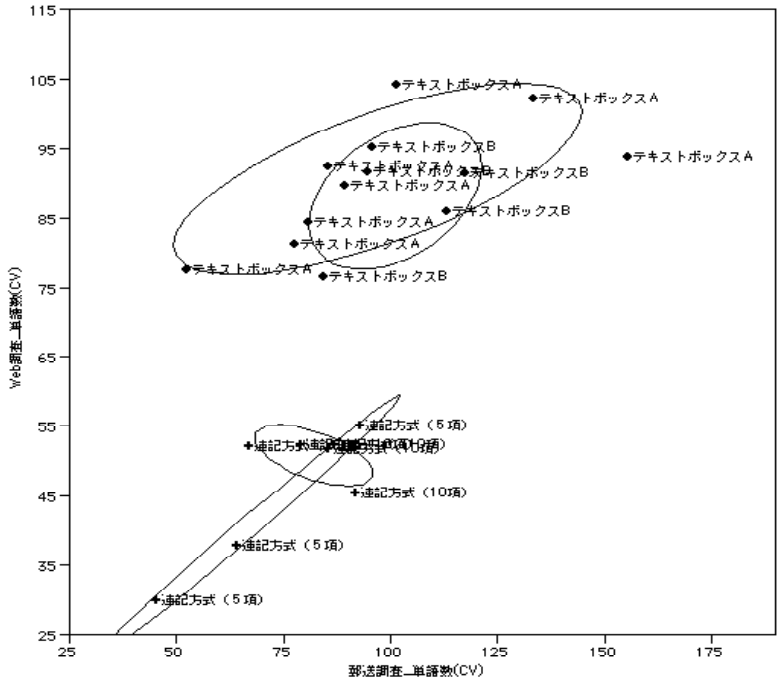

図５　Web 調査と郵送調査の平均単語数（変動係数）の比較

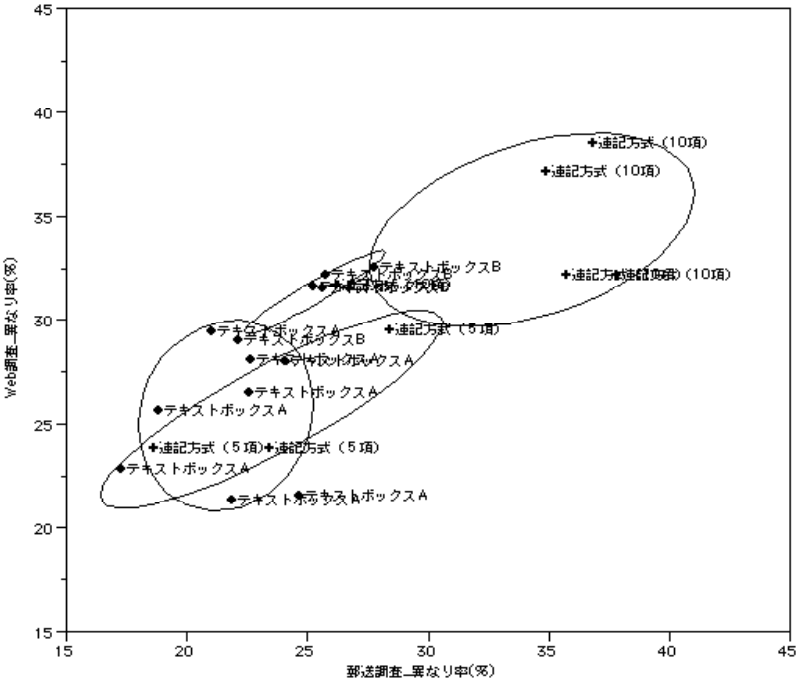

図６　Web 調査と郵送調査の異なり率（%）の比較

８．２　記述的な統計分析による探査

（１）用いる調査データの概要

以下で用いる調査データは，我々が進めてきたインターネット調査に関連した一連の実験調査で得た「第４次実験調査」の結果を引用する．これは以下の調査計画で進められた産学協同研究である．

① 筆者等（研究者グループ）と複数の調査機関との協同実験調査であること．
② 参加調査機関は，電通リサーチ，博報堂，東京サーベイ・リサーチ（博報堂-TSR グループ），日本リサーチセンターである．
③ 調査方式（調査モード）としては，Web 調査の他，オムニバス調査（面接方式，訪問留置方式），郵送調査を用いた．
④ 調査実施上の特徴として，それぞれの調査方式において調査票形式を揃えたこと（なるべく同じ調査票），調整質問をほぼ同じとしたこと，調査実施時期をなるべく揃えたこと（ほぼ同時期に実施），など事前の実験計画を徹底したこと．
⑤ 調査は２回行われ，それぞれ「生活意識編」「インターネット編」として実施した．他の類似調査（日本人の国民性他）との比較分析が可能なように，同一の質問を用いた他，独自の質問も用意した．

このように，**調査票や質問内容，実施時期を揃えたことで，調査方式の違いが回答結果にどう影響するかを客観的に評価できる**であろうとの意図がある．調査の詳細については大隅他（2003）を参照してほしい．なお，以下で用いる質問は，この第 II 部の後ろに［資料］としてまとめて添付した．

ここに示した実験調査結果の範囲でとの制約付きではあるものの，以下のような特徴が読みとれる．要は，このような実験を繰り返すことで，調査結果に及ぼす調査の誤差（とくに測定誤差，無回答誤差）の影響評価とその低減を図るような方向で客観的に検証することが望まれる．同時にこのことが，自由回答の分析結果の解釈や知見取得に影響することは言うまでもない．

（２）分析例−その１　自由回答質問への記入率の観察

Web 調査は，従来型の調査に比較して，自由回答への書き込み率（記入率）が高いという意見があるは，果たして本当であろうか．ここでまず，実験調査の結果からこの書き込み率を要約し，調査方式間でどのような関係があるかを観察する．

なお，自由回答質問は，その内容から Web 調査のみで用いる質問と，他の調査方式でも利用できる質問がある．この２回の調査について得られた結果を表５，表６に要約した．

表５　記入率の比較（生活意識編）

［第１回調査：生活意識編］

問 1-5. 一番大切なもの

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	939	98.7	1.3
	Hot Panel	Web 調査	3,392	98.3	1.7
博報堂-TSR	e-HABIT	Web 調査	931	98.8	1.2
日本リサーチセンター	Cyber Panel	Web 調査	716	99.9	0.1
電通リサーチ	DRPS	訪問面接	630	99.4	0.6
日本リサーチセンター	NOS	訪問留置	1,336	88.5	11.5
博報堂-TSR	HABIT	郵送	965	97.9	2.1

問 1-6. 他に大切なもの

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	939	96.6	3.4
	Hot Panel	Web 調査	3,392	94.4	5.6
博報堂-TSR	e-HABIT	Web 調査	931	96.0	4.0
日本リサーチセンター	Cyber Panel	Web 調査	716	98.0	2.0
電通リサーチ	DRPS	訪問面接	630	92.5	7.5
日本リサーチセンター	NOS	訪問留置	1,336	73.0	27.0
博報堂-TSR	HABIT	郵送	965	91.2	8.8

調査への感想

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	939	21.0	79.0
	Hot Panel	Web 調査	3,392	25.9	74.1
博報堂-TSR	e-HABIT	Web 調査	931	18.2	81.8
日本リサーチセンター	Cyber Panel	Web 調査	716	39.0	61.0

表６　記入率の比較（インターネット編）

［第２回調査：インターネット編］

問 1-1.「インターネットと聞いて思い浮かべる事柄」

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	894	98.7	1.3
	Hot Panel	Web 調査	2,587	99.0	1.0
博報堂-TSR	e-HABIT	Web 調査	896	98.7	1.3
日本リサーチセンター	Cyber Panel	Web 調査	642	99.4	0.6
電通リサーチ	DRPS（第２回）	訪問面接	630	100.0	-
日本リサーチセンター	NOS（第２回）	訪問留置	1,389	67.5	32.5
博報堂-TSR	HABIT（第１回のみ）	郵送	965	79.1	20.9

問 3-1. インターネット活用方法（自分自身）

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	894	98.2	1.8
	Hot Panel	Web 調査	2,587	98.3	1.7
博報堂-TSR	e-HABIT	Web 調査	896	98.1	1.9
日本リサーチセンター	Cyber Panel	Web 調査	642	99.4	0.6

問 3-2. インターネット活用方法（一般的）

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	894	92.2	7.8
	Hot Panel	Web 調査	2,587	92.2	7.8
博報堂-TSR	e-HABIT	Web 調査	896	90.7	9.3
日本リサーチセンター	Cyber Panel	Web 調査	642	96.1	3.9

問 6-7.「インターネット調査は、本音で答えやすいか」の理由

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	894	94.3	5.7
	Hot Panel	Web 調査	2,587	92.7	7.3
博報堂-TSR	e-HABIT	Web 調査	896	92.9	7.1
日本リサーチセンター	Cyber Panel	Web 調査	642	96.1	3.9

表６　記入率の比較（インターネット編）［つづき］

問 7-3.「インターネットはその時々によっていろいろな自分になれることが魅力である」の理由

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	894	92.1	7.9
	Hot Panel	Web 調査	2,587	91.8	8.2
博報堂-TSR	e-HABIT	Web 調査	896	92.2	7.8
日本リサーチセンター	Cyber Panel	Web 調査	642	95.8	4.2

調査への感想

実施機関	パネル名	調査方式	総数	回答あり	回答なし
電通リサーチ	DENTSU_R-net	Web 調査	894	30.6	69.4
	Hot Panel	Web 調査	2,587	33.9	66.1
博報堂-TSR	e-HABIT	Web 調査	896	22.1	77.9
日本リサーチセンター	Cyber Panel	Web 調査	642	48.0	52.0

さて，以上を観察するといくつかの特徴が読みとれる．

① 調査方式の違いが現れる，とくに Web 調査，留置など「自記式」であるか「面接」（面接員記入式）であるかの違いが顕著である．

② とくに，Web 調査とオムニバス（DRPS：面接方式）との差違はあまり見られない（いずれも高い）．

③ 一方，オムニバス（NOS：留置自記式）は，かなり異なる傾向にある．

④ Web 調査における書き込み率は確かに高い傾向にあるが，他の調査方式と比較すると必ずしも内容の豊富さを意味してはいない，記入率だけが情報量の評価には使えない．

⑤ むしろ，質問内容により構成要素数と異なり構成要素数の出方が類似している，つまり回答者は質問内容を見て，それに見合った回答行動を取るという常識的な傾向を示す．（別の集計から得た結果，後述のように構成要素他の初動探査が重要）

⑥ とくに，Web 調査「調査への感想」を何でも記入とした調査票の最後の置かれた質問については，いずれも記入率が少なく，しかもサイト間の特徴が顕著である．

ほとんど同じ調査票（質問）を用い，ほぼ同時期に実施した調査であるということを考慮すると，ここに見える特徴はなかなか興味ある結果であろう．もちろん，ここで得た自由回答データは，他の定量型質問（選択肢型）質問とクロスさせることで，別の情報が得られることは言うまでもないし，それがマイニングの目標でもある．

たとえば，「あなたにとって大切なものは」（一番大切なものは，次に大切なものは）を分析すると何が見えるのか，どのような属性や選択肢型質問が有意に関係するのであるおかといった問への解答がテキスト・マイニングにより得られるのである（後述の解析例 I を参照）．

（２）分析例–その２　–構成要素の記述統計的な観察–

前の例で，構成要素数，異なり構成要素数，異なり構成要素率（異なり率）等については，記述統計的な初動探査が必要であると述べた．これにならって，この調査で得られた情報について，構成要素数と異なり構成要素数の関係および構成要素数と異なり構成要素率の関係を調べる．

構成要素数と異なり構成要素数の関係は図７となる．前の例でみたようにここでも，（断定的にはいえないものの）構成要素数の増加に伴い異なり構成要素率は指数的に増加し，頭打ちとなるようにみえる（単純に比例的には増えない）．

次に，構成要素数と異なり構成要素率については，図８のようになる．こちらは質問名のラベルを入れてみた．図が煩雑になるが，同じ質問群がおおむね近い位置にあること，また質問の特徴が見えること，たとえば，「大切なもの」は構成要素数も異なり率も低い，「調査への意見」は書き込み語数が少なく異なり率が高い，インターネットへの意見はバラツキが大きく，調査サイトの差違があるらしいなどの特徴が見える．また，（図８と合わせて考えると）構成要素数の増加に伴い，異なり率は低減するという傾向が見られる．

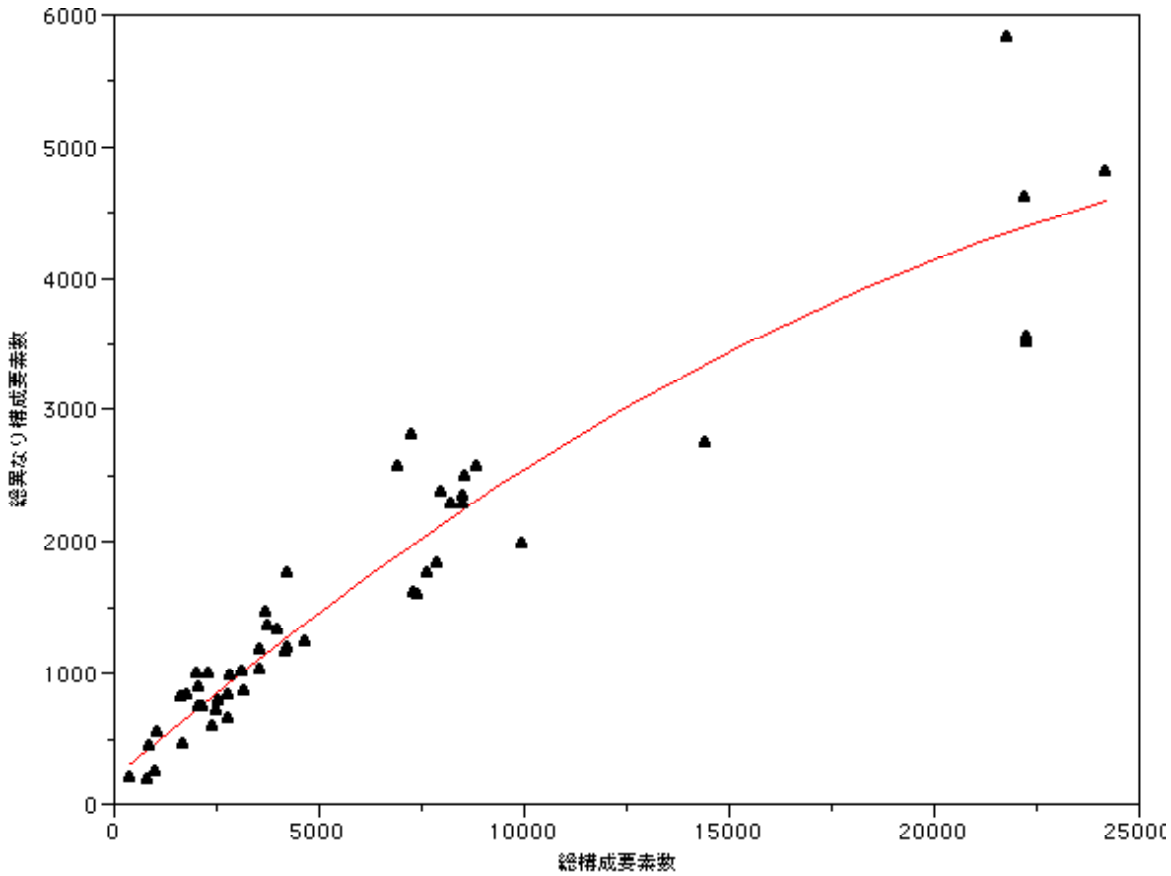

図７　構成要素数と異なり構成要素数の関係

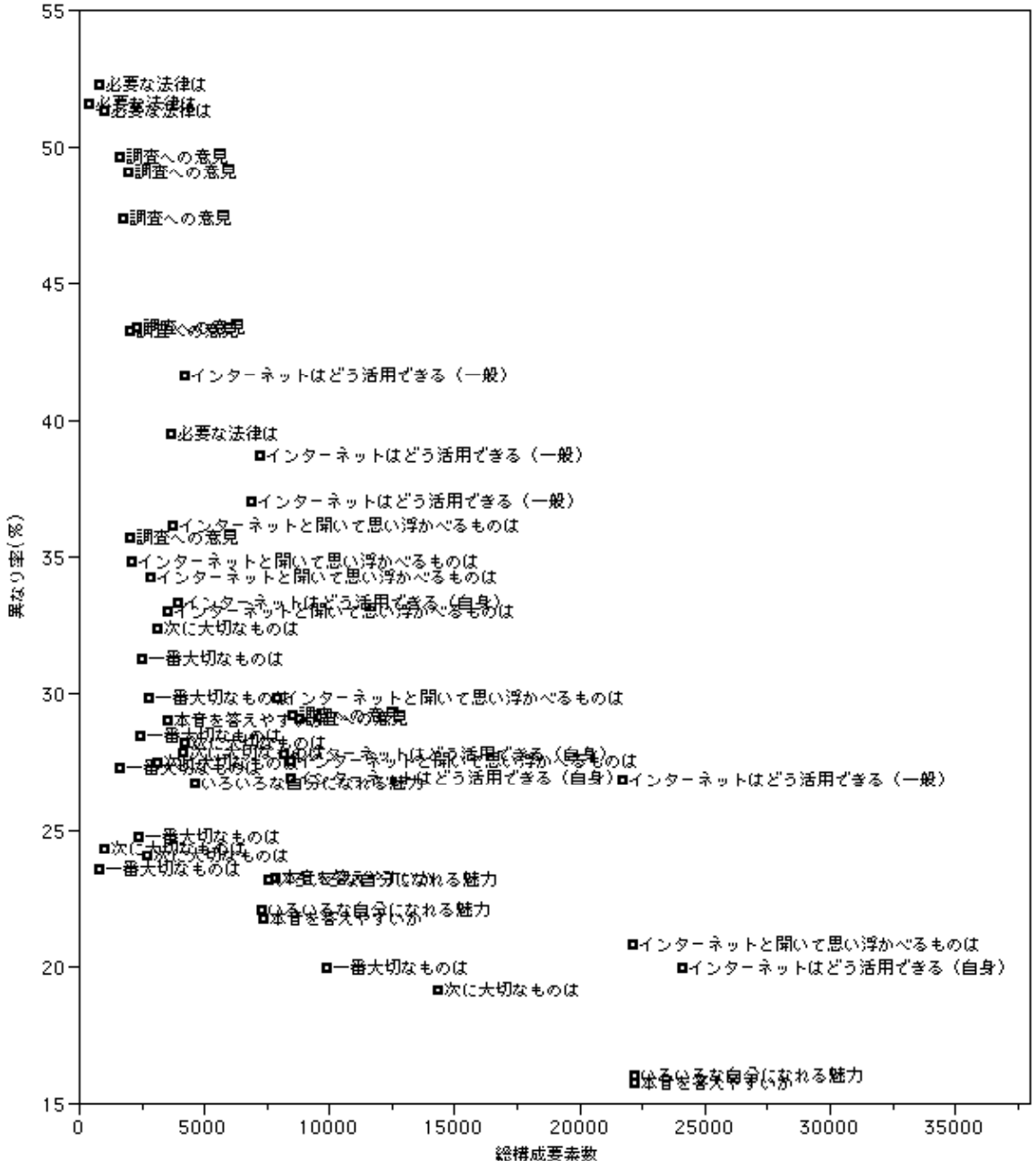

図８　構成要素数と異なり構成要素率（%）の関係

実は，（前の例と）ここで見た特徴は，一般にある規則性をもっていることが，多数の例を調べることで類推される．要は，ここでも調査方式，質問内容と質問形式の影響が，それぞれある特徴として如実に現れることが実証的に示されるということである．**自由回答はただ質問枠を用意して適当に聞けばよいとはならない**ということを示している．

８．３　分かち書き処理ツールの比較分析　–簡単な例示–

分かち書き処理の結果は，用いるソフトや分かち書き処理の方式により異なることは指摘した．これは換言すると，当該データを用いた分析の出発点が異なるから，それから先の分析いかんに関わりなく，同じ解（分析結果）とはならないということに他ならない．しかし，テキスト・マイニングでこのことに触れることや，実際に異なるソフトを相互に比較した結果を確認する機会はほとんどないといってよいだろう．ここでは，共通の自由回答質問を用いて，次の２つのことを試みることとした．

（その１）複数の分かち書きソフトを用いて，その結果がどう異なるかを観察する．
（その２）複数のテキスト・マイニング・ツールを用いて，分かち書き処理を行い，その結果得られた構成要素（単語，語句）の頻度集計を行い，結果の比較を行う．

まず，ここで用いた調査内容と２つの質問を挙げる．

［調査の内容］

調査内容：前述のインターネット調査のうち，第２回調査（インターネット編）
実施機関：電通リサーチ
登録集団（リソース）：DENTSU_R-net
構成：計画サンプル数（1,542 サンプル），回収サンプル数（894 サンプル；分析対象とするサンプル数；有効回答率は 58.0%），うち 878 サンプル（98.2%）に自由回答書き込みあり（表６を参照）

［用いた質問］

Ｑ３．次に、あなたと「インターネット」とのかかわりについてお伺いします。
Q3-1. あなたご自身にとって「インターネット」は、どのようなことがらに活用できると思いますか。どんなことでも結構ですので、以下になるべく具体的にご記入ください。
Q3-2. では、一般的に「インターネット」は、どのようなことがらに活用できると思いますか。なるべく、他にはないような活用法を、どんなことでも結構ですので、以下になるべく具体的にご記入ください。

この２つの質問は，回答者と「インターネット」とのかかわりについて問うたものである．Q3-1 は，「回答者自身」にとってインターネットがどのように活用できるかを，また，Q3-2 は，「一般的に」インターネットがどのように活用できるか，「他にはないような活用法の提案」を求める質問となっている．

（注）以下の比較分析を行うために数社の調査機関のご協力を仰いだ．ご協力いただいた調査機関には厚く謝意を表したい．なお，ここでその社名を記すと，用いたソフトが判明するので，ここではあえてこれを伏した．

（1）分析例–その１　–分かち書きソフトの比較–

まず，以下に挙げる３種の分かち書き処理ソフトを用いて，実際に分かち書きを行う．

① ソフトＷ：これはあるテキスト・マイニング市販製品に搭載のツールである．
② ソフトＱ：これは単体の分かち書きツールとして知られているあるソフトである．
③ 茶筌：これはフリーウェアで公開されている，おそらく国内でもっとも普及している分かち書きソフトである．

この３種のソフトによる上記の自由回答質問への回答データの分かち書き処理結果の中から適当に５サンプルを取りだして表６に一覧とした．少々見づらいが，区切り記号「｜」で分かち書き部分が示されている．**それぞれの結果が微妙に異なる**ことが分かるであろう．

一般には，ここに挙げた例のように短文ばかりとは限らず長短取り混ぜて様々であるから，分かち書きの結果の差違もそれなりにばらつくのである．また，そのソフトにとっての得手不得手もあるであろう．ここでは，**分かち書き処理の結果は同じにはならないこと，従ってその後の解析結果も同等とはならないこと**，を指摘するに留める．

また通常は，形態素解析を行うと，品詞や活用形の情報，分かち書きの情報などが得られる．下に「茶筌」を使った短い例文を挙げておく．

［例］茶筌の結果

＜原文＞

仕事と趣味の情報源。

＜品詞分解＞

仕事	シゴト	仕事	名詞–サ変接続
と	ト	と	助詞–並立助詞
趣味	シュミ	趣味	名詞–一般
の	ノ	の	助詞–連体化
情報	ジョウホウ	情報	名詞–一般
源	ゲン	源	名詞–接尾–一般
。	。	。	記号–句点
EOS			

＜分かち書き結果＞

仕事｜と｜趣味｜の｜情報｜源｜。（記号「｜」で区切った）

（注）詳しくは，本講座の「第 III 部 形態素解析と分かち書き処理」の部を参照されたい．

表６　分かち書き処理の例（３つのソフトの比較）

通番	サンプル ID	ソフトＷの結果	ソフトＱの結果	茶筌の結果
1	100004465	コミュニケーション|手段|として|複数|の|人間|と|同時|に|作業|を|する|場合|に|進行状況|を|常に|ＣＣ|し|て|おく|と|現状|が|**把握|し|やすく**|スムーズ|に|行|える|。|同様|に|、|「|言|った|」|「|言わ|ない|」|など|と|言う|トラブル|が|**無くな|っ|て|い|う|こと|な|し**|。	.コミュニケーション.手段.と|し.て|複数.の|人間.と|同時.に|作業.を|する|場合.に|進行状況.を|常に|CC.し.て|おく.と|現状.が|**把握.し.やす-く**|スムーズ-に|行-え-る.。同様-に.、|「.言-っ.た.」|「.言-わ.ない.」.など.と|言-う|トラブル.が|**無くな-っ.て.い.う|こと|な-し.**。〃	コミュニケーション|手段|として|複数|の|人間|と|同時|に|作業|を|する|場合|に|進行|状況|を|常に|CC|し|て|おく|と|現状|が|**把握|し|やすく**|スムーズ|に|行える|。|同様|に|、|「|言っ|た|」|「|言わ|ない|」|など|と|言う|トラブル|が|**無くなっ|て|いう|こと|なし**|。|」|
2	100006353	・|仕事|上|で|、|社内|外|との|連絡|、|ファイル|交換|・|サークル|内|の|連絡|（|**めー|リング|リスト**|）|、|ファイル|交換|・|**サーチ|エンジン**|による|、|様々|な|検索	.・|仕事.上.で.、|社内外.と.の|連絡.、|ファイル.交換.・|サークル.内.の|連絡|（.**め|ーリングリスト**.）.、|ファイル.交換.・|**サーチエンジン**.に.よ-る.、|様々-な|検索.〃	・|仕事|上|で|、|社内外|と|の|連絡|、|ファイル|交換|・|サークル|内|の|連絡|（|**め|ー|リング|リスト**|）|、|ファイル|交換|・|**サーチ|エンジン**|による|、|様々|な|検索|
3	100006445	オークション|買い物|興味|を|持|っ|た|もの|の|詳細|を|調べ|る|ニュース|など|の|最新情報	オークション.　|買-い+物.　|興味.を|持-っ.た.ものの|詳細.を|調-べ-る.　|ニュース.など.の|最新情報	オークション|　|買い物|　|興味|を|持っ|た|ものの|詳細|を|調べる|　|ニュース|など|の|最新|情報|
4	100007404	趣味|の|情報|取得|、|友人|との|連絡|、|レジャー|情報|の|収集|、|業務|情報|の|収集|、|ショッピング	趣味.の|情報取得.、|友人.と.の|連絡.、|レジャー.情報.の|収集.、|業務情報.の|収集.、|ショッピング	趣味|の|情報|取得|、|友人|と|の|連絡|、|レジャー|情報|の|収集|、|業務|情報|の|収集|、|ショッピング|
5	100012767	・|知|ら|ない|・|または|不明確|な|問題|の|情報|収集|・|多角的|な|情報|から|自分|なり|の|考え|を|確立|・|テレビ|で|見|ら|れ|ない|海外|の|番組|視聴|（|多|チャンネル|化|とも|言え|る|）|・|コミュニケーション|の|道具|・|情報|の|共有	.・|知-ら.な-い.・|または|不明確-な|問題.の|情報収集.・|多角-的-な|情報.から|自分.なり.の|考-え.を|確立.・|テレビ.で|見.ら-れ.な-い|海外.の|番組視聴|（|多|チャンネル-化.と.も|言-え-る.）.・|コミュニケーション.の|道具.・|情報.の|共有.〃	・|知ら|ない|・|また|は|不明確|な|問題|の|情報|収集|・|多角|的|な|情報|から|自分|なり|の|考え|を|確立|・|テレビ|で|見|られ|ない|海外|の|番組|視聴|（|多|チャンネル|化|と|も|言える|）|・|コミュニケーション|の|道具|・|情報|の|共有||

（２）分析例–その２　–分かち書きの構成要素の分布–

ここで，総構成要素数，異なり構成要素数，異なり構成要素率の分布をみる．実は，高度なデータ解析に入る前に，分かち書き処理で得た構成要素数（単語，語句など）の分布の記述的な分析，**初動探査が重要**である．（後述するように）どのような単語・語句が登場したか，その出現単語の頻度分布はどのようになるかについては注目されることが多いが，その基礎情報である，

① 構成要素数，あるいは単語数の分布
② 異なり構成要素数あるいは異なり単語数
③ 異なり構成要素数率あるいは異なり単語率

については，総合的に観察することが少ないようである．

登場した単語頻度の観察は②に相当するが，これに加えて③の指標が重要であることは，今までの例で見たとおりである．ここらの基礎情報の記述的分析の理解が，後の高度な分析に影響を及ぼすことは明らかであるが，テキスト・マイニングのツールをみても，これらはほとんど考察されることはない．しかし，**コーパス言語学や内容分析**などの分野では，こうした情報が適宜利用されている．ここでは，その簡単な例として，上の２つの質問に WordMiner を適用して得られた基礎情報の一部を示す．

まず，図 9～図 12 は，は質問 3–1「インターネットの活用方法（自身について）」，質問 3–2「インターネットの活用方法（一般に）」について，構成要素数，異なり構成要素数，異なり構成要素率をグラフとしたものである．

図 9，図 11 の横軸の「頻度 1，2，…」は，頻度 1 が全構成要素数，つまり分かち書き処理で得た「分かち書きの総数」である．以下，頻度 2 は頻度 2 以上の構成要素数，…となる．なお，ここでは頻度 20 までを表示した．図の各頻度の上にある棒グラフの左側が**構成要素数**（単語数）を表している．また，棒グラフの右側が**異なり構成要素数**（異なり単語数）を表す．また，これら２種の構成要素数の頻度が図の左側の目盛りで示してある．図の折れ線は**異なり構成要素数率**（異なり単語率）を表し，右側の目盛りがその割合（%）を示している．

一方，図 10，図 12 は，上の情報から，各頻度別に構成要素数と異なり構成要素数を再集計した結果である（WordMiner で出力）．図 9，11 が各頻度の累積情報であったのに対して，こちらは頻度別の出現した異なり語数となっている．たとえば，図 10 の頻度 3 とは「3 回だけ出現した構成要素数と異なり構成要素数」である（この場合，291 語の構成要素と 97 語の異なり構成要素があった）．

この 2 例だけからは，これらの指標の用い方が即座には理解できないであろうが，筆書の過去の分析体験から，なによりもまずこの指標を用いて，調査方式，質問形式などの影響を総合的に比較考察することが必要である．この例も含めて，一般に以下の特徴が見られる．

① （かなり自明のことであるが）頻度１（１回しか登場しない語句）の構成要素が圧倒的に多い．

② 頻度２から，急速に構成要素数が低減する（これも多くの場合，共通した特徴）．分析上は，頻度２以上の構成要素の分布に注目すべきである．どの程度の語句の出現頻度が現れるかの内訳を知ることが必要になる（後述）．
③ 異なり構成要素率（異なり単語率）の変化に注目すべきである．これの推移を観察することは，自由回答の内容のまとまりの程度（あるいは意見の発散度），語彙量の目安と考えられているが，前に指摘のように，調査方式などの影響でこれは変化する．
④ 分析上は，異なり構成要素のすべて，つまり総異なり構成要素を用いることは，通常は膨大な語数を扱うこととなり，明らかに分析処理の負荷量が増える．
⑤ しかし，ここに挙げたような情報を観察せずに，勝手に語数をフィルタリングして用いることも，大事な情報を棄てる可能性があり，また，あまりに恣意的である．
⑥ **「稀な言葉」が重要**という言い方がある．この情報源を構成要素・単語に求めるなら，頻度１を棄てることは，この情報を得ることを阻害･困難とすることを意味する．
⑦ 出現頻度の計数の仕方がいろいろあるようだが，原則として**「分かち書き処理で得た全情報，計数情報」が明示的に分かること**が重要である．

繰り返しになるが，自由回答データの結果には，そこで用いた**調査方式，質問形式や質問内容**が反映することは明らかなことであり，この影響を評価するための手がかりの一つが，ここに挙げた構成要素数，異なり構成要素数，異なり構成要素率の分布を記述的に探査することである．もちろん，この情報だけですべてが語れるわけではないが，これらの考察も行わずに先走った分析を行うことは慎むべきことである．このような指摘を行う理由は，既存の多くの商用ソフトが，これらの基礎情報が正確に取得できるとは限らず手当が十分とはいえないという実状がある．

- 抽出した構成要素数，単語数・語句数が幾つか
- 異なり構成要素数（率），異なり単語数（率）はどのような分布か
- また事後の分析にどの程度の構成要素，単語を残すべきか
- その根拠をどこに求めるのか
- それらが，用いた調査方式，質問形式や質問内容とどのように関係するか

こうした基礎情報が見えないまま，高度な分析を行ったからといって，その結果の信憑性ははなはだ怪しいものと考えるのが妥当である．また正しいマイニングの方向とは，どんな初等的，基礎的な分析結果をもないがしろにせず，あくまでも科学性のある客観的なアプローチを心掛けるべきである．

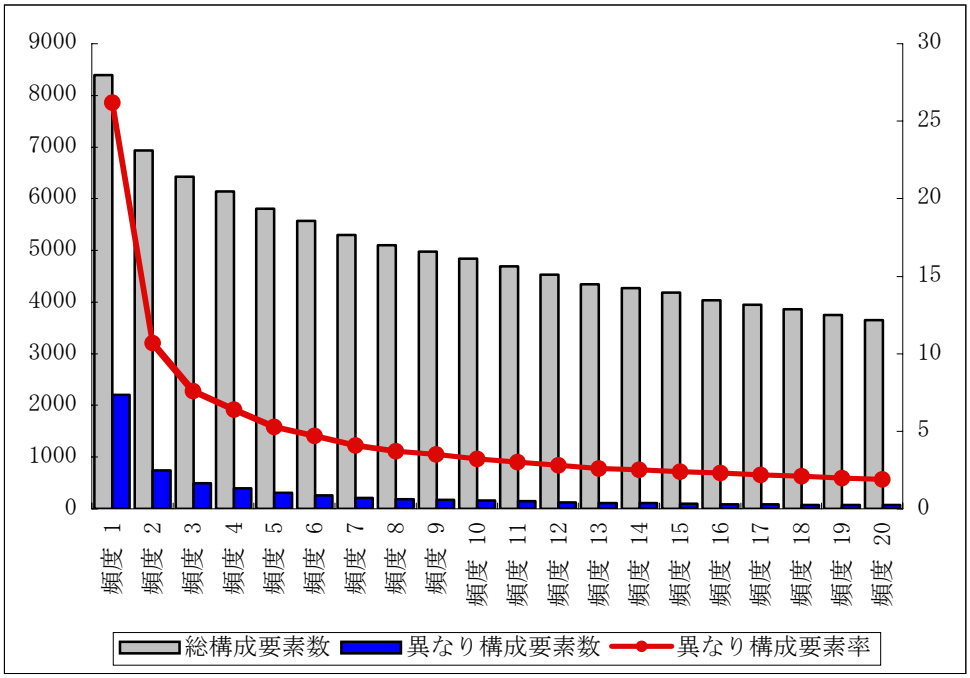

図９　総構成要素数，異なり構成要素数，異なり構成要素率の分布（Q3-1 の場合）

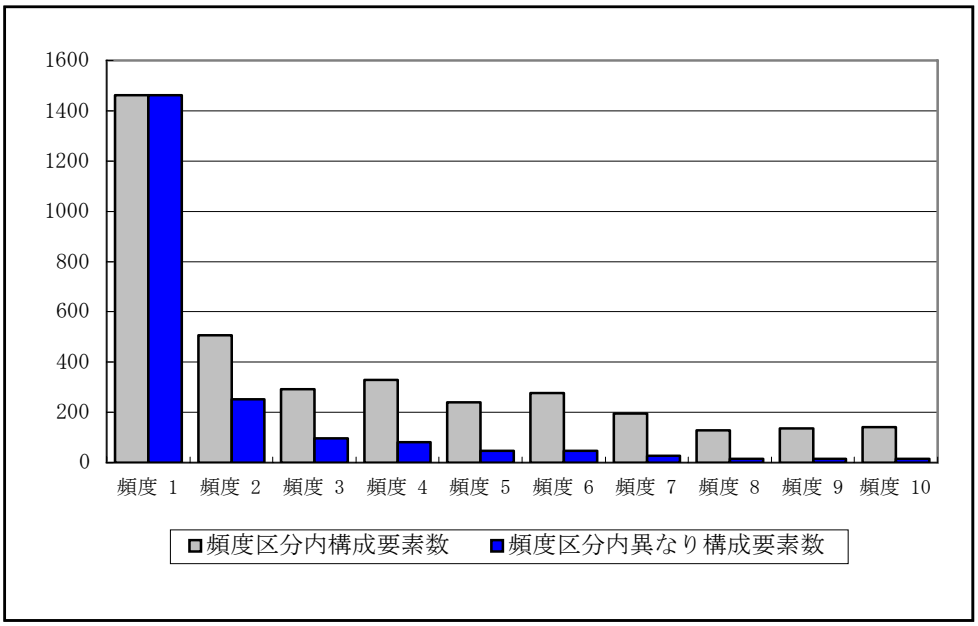

図 10　異なり構成要素数の閾値別分布（Q3-1 の場合）

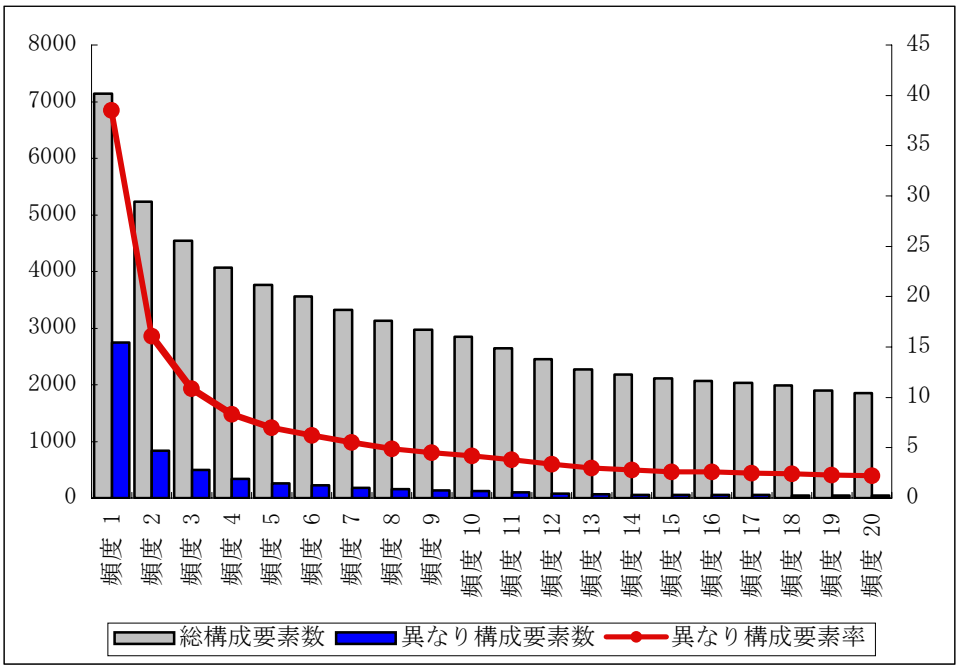

図 11　総構成要素数，異なり構成要素数，異なり構成要素率の分布（Q3-2 の場合）

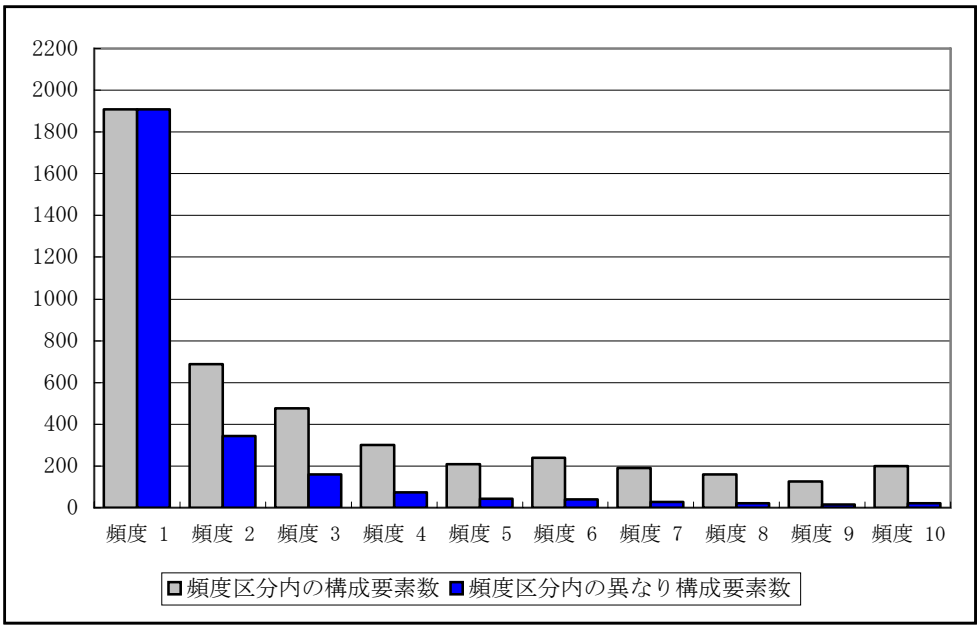

図 12　異なり構成要素数の閾値別分布（Q3-2 の場合）

（３）分析例–その３　–分かち書きの単語頻度集計の比較–

単語集計の結果だけから，２つの質問への回答傾向への特徴（類似や差異）を客観的に読みとること，あるいは高度な知見を求めることは，実はそう容易ではないだろう．しかしこれは，テキスト・マイニングの分析過程の初動探査として必須の手順である．とくに，複数の分析ツールの利用が可能であれば，それらの出力結果を相互比較する意味は大いにある．ここでは，以下の４つのソフトウェアによる単語頻度集計結果を比べてみた．

① WordMiner：分かち書きツールは WinAiBASE
② ソフト D：大手広告会社とソフト会社との共同開発ソフト
③ ソフト V：テキスト・マイニング専用ソフト，分かち書きは茶筌を利用とある
④ ソフト S：調査データ分析用をうたった市販の製品

なお，**WordMiner** と**ソフト D** については，分かち書き処理後に若干の単語編集を行っている．たとえば，

- WordMiner では，「情報」「検索」「収集」などについて，「情報検索」「情報収集」は複合語としてつなげ，「情報の検索」「情報の収集」は，「情報」「検索」「収集」は分けて扱うなどとした．
- また，「電子メール」と「メール」を同義として括るなどの作業を行った．
- ソフト D についてもまた，簡単な単語編集を行った．
- ただし，原則として，誤記，同義語編集などはなるべく行わないで出力するよう努めた（しかし，オプションの指定や単語選択機能等が明らかでないソフトもあった）．

さて，結果を比較すると，仮に若干の編集操作を行った影響を考慮しても，各ソフト間の単語・語句の出現頻度，頻度順には微妙な差違があることが見える．

こうした情報を示すと，そもそも各ソフトの設計指針，思想が異なるのであるから，比較することには意味がないという指摘を受けることがある．しかし，ここに挙げた操作（分かち書き処理の結果から出現単語を要約整理すること）は，**分析の入り口，導入部の基本的な情報**を提供するものであり，テキスト・マイニングの基礎情報である．出発点でこれだけ異なる結果となるということは，それ以降の分析結果や，そこから何らかの解釈を行うマイニング過程がすべて異なることを意味するものであり，ひいてはマイニングの目的である知識発見，知見取得，その解釈に影響を及ぼすものである．

- 分かち書きがどう行われたか，単語編集がどう行われたかがよく見えない．
- ソフトによっては，なぜ一覧に列記の単語が選出・抽出されたかが明かでない．
- 一覧を比較すると，同じ単語・語句の出現頻度数がソフトによりかなり異なることが見える．
- そのような結果となる背景にある処理過程．手順がよく見えない．
- ここでは，紙幅の都合で頻度上位 50 位を示したが，**下位頻度の単語**（稀な出現

単語）が重要であるが，この情報が見えないソフトがある（自動的に切り捨てたようだが，その根拠が分からない）.

- 誤記，同義語句，削除語句などの（編集）情報が明示的に示されない.
- よって，解析結果を対比することに困難性がある（出発点となる単語群が異なるから）.
- ソフトによっては，単語数がかなり絞り込まれる場合がある．とくに視覚化ツール利用の段階でこれが多いが，そのような語句の絞り込みが行われた理由や処理過程が明らかでない.
- つまり「TM による大量データからの知識発見」とはならないケースがある.
- ノイズが多く，しかもボリュームの多い複雑，煩雑なもやもやしたデータから有用情報を探査することになっていない.
- さらに「希少な例，まれな語句の傾向，特徴」はどう読めるのか見えない（よくいわれることだが）.
- そして，定量的質問項目（選択肢型質問，属性など）の対比をどう読み取るのか，あるはどのような操作が可能か.

とくに，実際の分析では，誤記の訂正，同義語や類語の括りや置換，さらには関連語をどう括るのか否かなど，分析作業者が主観的に行う**手作業の部分**も多くある．換言すれば，複数の分析者がおれば，複数の編集済みの単語集合ができて，複数の解析結果があることになる（そして，これがテキスト・マイニング利用現場の実態であろう）．同時に，分析者が，こうした**煩雑な作業からどの程度解放されるか**が重要である．多少はいい加減に分析を行っても，まずデータ構造の全体像を把握すること，精緻な分析を行いたいときにも，要求に合った対応が可能であること，そして何よりも，どんな処理分析を行っても，その分析経緯が追跡可能であって，分析者が何を行ったかが具体的な情報として明示的に示されることが求められる．しかしながら，現状の多くのテキスト・マイニング・ツールは，あまりに暗箱化されていていずれもが曖昧である.

始めに，

（A1）２つの質問の差異を表す特徴的な語句，単語は何か，
（A2）属性，とくに性別，年齢区分などの違いが回答に現れるのか，

といった易しい課題目標を掲げてみたが，果たしてこれが実現できるのだろうか．この設定は特殊なものではなく調査データの分析では日常的に求められることである．通常は，これに加えて，所与の自由回答が調査票内の他の質問とどう関係するか，どの程度有意に関係するかといった情報も求められるであろう.これは単に相関パターンをみる，分類を行う，概念を知る，…といった表層的な言葉だけでは解決できないことである.

あるデータセットを手にすると，様々な要求事項が次々に表れる．これに応えることがデータ・マイニングやテキスト・マイニングの役割だが，以上に述べたことに対応するだけでも，かなりの労働や工夫を強いられるわけで，果たして真の意味でのマイニングの道しるべはどこに，といった印象は否めない.

表 7－1　４つのソフトの単語頻度集計結果の比較（Q3-1 について，上位 50 位までを表示）

Q3-1：インターネットの活用（自身）

WordMiner（WinAIBASE）の場合

頻度順位	構成要素	構成要素数	サンプル度数
1	情報	319	260
2	情報収集-情報集め	203	186
3	事-こと	200	144
4	できる	141	109
5	する	132	113
6	趣味	113	110
7	電子メール-メール	106	99
8	検索	104	97
9	仕事	103	95
10	して	100	82
11	友達-友人	84	82
12	もの	76	68
13	調べる	75	61
14	入手	74	67
15	等	70	61
16	いる	64	56
17	コミュニケーション	56	55
18	収集	52	48
19	買い物	51	50
20	連絡	51	47
21	ショッピング	50	50
22	自分	50	47
23	ある	47	42
24	活用	45	42
25	旅-旅行	44	44
26	ニュース	43	43
27	人	41	38
28	知りたい	41	39
29	時	39	34
30	得る	37	34

ソフト D の場合

頻度順位	単語	頻度	ユニット数
1	情報	762	715
2	収集	265	265
3	調べる	188	180
4	仕事	145	144
5	検索	144	141
6	入手	137	134
7	ショッピング	135	133
8	メール	114	111
9	趣味	114	113
10	友人	91	90
11	連絡	70	68
12	旅行	69	65
13	得る	62	58
14	手段	58	58
15	コミュニケーション	57	57
16	買う	57	56
17	自分	54	54
18	活用	51	51
19	ニュース	50	50
20	予約	49	48
21	時間	46	44
22	ホームページ	45	45
23	人	44	42
24	交換	43	42
25	知りたい	42	42
26	商品	38	35
27	利用	38	38
28	見る	38	38
29	発信	37	37
30	わからない	34	33

（表７−１つづき）

頻度順位	構成要素	構成要素数	サンプル度数	頻度順位	単語	頻度	ユニット数
31	予約	36	33	31	インターネット	33	33
32	利用	36	32	32	使う	32	30
33	ホームページ	34	34	33	やり取り	31	31
34	手段	34	28	34	必要	30	30
35	時間	33	32	35	欲しい	30	29
36	欲しい	33	28	36	生活	29	28
37	では	31	27	37	探す	28	28
38	インターネット-ネット	31	28	38	知る	28	27
39	情報入手	31	29	39	外国	28	28
40	色々な	31	30	40	代わる	27	27
41	仕事上	30	30	41	チケット	27	27
42	購入	29	27	42	事柄	27	25
43	必要	29	29	43	簡単	26	26
44	ため-為	28	25	44	便利	26	26
45	また	28	28	45	オークション	25	25
46	よる	28	25	46	知識	25	25
47	商品	28	25	47	テレビ	25	24
48	調べ	28	28	48	興味	25	25
49	調べ物	28	28	49	銀行	24	22
50	インターネットショッピング-オンラインショッピング	27	27	50	場所	24	24

表 7−2　４つのソフトの単語頻度集計結果の比較（Q3-1 について，上位 50 位までを表示）

ソフト V の場合（茶筌利用，サンプル数の表記なし）

頻度順位	単語	頻度
1	情報	703
2	収集	265
3	でき	157
4	調べ	150
5	検索	131
6	入手	118
7	趣味	116
8	仕事	113
9	メール	110
10	知り	89
11	ショッピング	78
12	連絡	70
13	友人	63
14	旅行	59
15	手段	58
16	なり	54
17	コミュニケーション	54
18	あり	51
19	ニュース	50
20	活用	50
21	自分	50
22	購入	49
23	予約	49
24	より	45
25	インターネット	45
26	買い物	45
27	ネット	43
28	事	41
29	時間	41
30	見	40

ソフト S の場合（WordMiner で再集計）

頻度順位	単語	単語数	サンプル度数
1	情報	545	545
2	収集	221	221
3	仕事	134	134
4	趣味	112	112
5	検索	107	107
6	調べる	103	103
7	メール	92	92
8	ショッピング	78	78
9	入手	72	72
10	友人	62	62
11	連絡	62	62
12	できる	61	61
13	する	60	60
14	上	58	58
15	知る	58	58
16	得る	55	55
17	コミュニケーション	53	53
18	買い物	51	51
19	旅行	50	50
20	ニュース	47	47
21	自分	47	47
22	手段	44	44
23	予約	44	44
24	ある	40	40
25	なる	39	39
26	インターネット	39	39
27	人	39	39
28	時	38	38
29	ネット	35	35
30	よる	34	34

（表７−２つづき）

頻度順位	単語	頻度
31	人	40
32	調べ	39
33	とき	38
34	思い	38
35	商品	38
36	利用	38
37	発信	37
38	得り	36
39	ページ	33
40	欲し	33
41	関し	32
42	仕事上	32
43	時	32
44	ため	31
45	必要	31
46	ホーム	30
47	出来	30
48	探し	28
49	つき	27
50	チケット	27

頻度順位	単語	単語数	サンプル度数
31	活用する	33	33
32	思う	33	33
33	時間	33	33
34	ため	32	32
35	関する	32	32
36	見る	32	32
37	商品	32	32
38	探す	32	32
39	発信	31	31
40	ホームページ	30	30
41	検索する	30	30
42	入手する	30	30
43	調べ	29	29
44	また	28	28
45	調べ物	28	28
46	友達	28	28
47	欲しい	28	28
48	購入	27	27
49	必要だ	27	27
50	として	26	26

表８－１　４つのソフトの単語頻度集計結果の比較（Q3-2 について，上位 50 位までを表示）

Q3-2：インターネットの活用（一般）

WordMiner（WinAiBASE）

頻度順位	構成要素	構成要素数	サンプル度数
1	できる	156	126
2	事-こと	137	110
3	情報	114	103
4	して	110	94
5	する	100	84
6	ない	70	58
7	人	66	57
8	いる	60	53
9	等	59	51
10	インターネット-ネット	55	44
11	思う	55	45
12	ある	53	46
13	では	48	40
14	情報収集-情報集め	47	46
15	には	44	41
16	もの	38	37
17	です	35	31
18	ように	34	30
19	他	34	33
20	活用	33	31
21	なる	32	30
22	コミュニケーション	31	31
23	検索	30	27
24	電話	28	27
25	出来る	27	25
26	思います	26	24
27	電子メール-メール	26	26
28	活用法	25	24
29	いい	24	20
30	その	24	23

ソフト D の場合

頻度順位	単語	頻度	ユニット数
1	情報	238	222
2	人	76	69
3	活用	66	64
4	収集	60	59
5	ネット	56	53
6	インターネット	54	53
7	電話	49	44
8	調べる	49	45
9	テレビ	48	48
10	ショッピング	45	45
11	家	40	39
12	検索	38	36
13	他	37	37
14	使う	36	35
15	コミュニケーション	35	35
16	見る	35	33
17	思いつく	34	33
18	世界	33	33
19	利用	29	29
20	入手	28	27
21	メール	28	28
22	個人	25	24
23	連絡	24	22
24	わからない	23	23
25	商品	23	22
26	ホームページ	23	22
27	買う	23	20
28	自分	23	23
29	発信	22	21
30	交換	22	22

表８−１（つづき）

頻度順位	構成要素	構成要素数	サンプル度数
31	自分	24	23
32	たとえば	24	24
33	利用	23	23
34	した	22	21
35	特に	22	22
36	自宅	21	21
37	色々な	21	21
38	リアルタイム	20	19
39	今	20	20
40	調べる	20	18
41	必要	20	18
42	ホームページ	19	19
43	買い物	19	19
44	いう	18	15
45	だ	18	15
46	ネット上	18	16
47	仕事	18	15
48	時間	18	16
49	家	17	16
50	場所	17	17
51	便利	17	17

頻度順位	単語	頻度	ユニット数
31	仕事	22	19
32	リアルタイム	22	21
33	旅行	21	17
34	行なう	21	20
35	投票	20	20
36	代わる	20	19
37	配信	20	18
38	生活	20	18
39	チャット	19	19
40	オンラインゲーム	19	17
41	場所	19	19
42	通信	19	18
43	手段	18	18
44	医療	18	18
45	学校	18	16
46	時間	18	17
47	便利	17	17
48	意見	17	17
49	可能	17	17
50	お店	17	15

表８−２　４つのソフトの単語頻度集計結果の比較（Q3-2 について，上位 50 位までを表示）

ソフトＶの場合（茶筌利用，サンプル数の表記なし）

頻度順位	単語	頻度
1	情報	215
2	でき	180
3	思い	104
4	な	83
5	なり	63
6	人	61
7	ネット	60
8	収集	60
9	インターネット	59
10	あり	56
11	出来	54
12	電話	53
13	活用	43
14	思いつき	42
15	調べ	42
16	知り	40
17	事	38
18	テレビ	37
19	他	37
20	見	36
21	コミュニケーション	34
22	検索	34
23	上	34
24	わかり	31
25	メール	29
26	いい	28
27	利用	28
28	より	26
29	活	24
30	行き	24

ソフトＳの場合（WordMiner で再集計）

頻度順位	単語	単語数	サンプル度数
1	情報	190	190
2	思う	74	74
3	ない	70	70
4	できる	68	68
5	する	64	64
6	人	58	58
7	収集	56	56
8	なる	49	49
9	インターネット	49	49
10	ネット	49	49
11	ある	46	46
12	思い付かない	44	44
13	出来る	42	42
14	電話	40	40
15	活用	36	36
16	他	36	36
17	コミュニケーション	34	34
18	上	33	33
19	見る	32	32
20	調べる	32	32
21	検索	29	29
22	ＴＶ	28	28
23	使う	28	28
24	活用する	25	25
25	知る	25	25
26	として	24	24
27	わからない	24	24
28	法	24	24
29	ショッピング	23	23
30	自分	23	23

表８−２（つづき）

頻度順位	単語	頻度
31	発信	24
32	用法	24
33	連絡	24
34	ショッピング	23
35	個人	23
36	自宅	23
37	商品	23
38	使い	22
39	自分	22
40	特に	22
41	仕事	21
42	い	20
43	い	20
44	投票	20
45	入手	20
46	配信	20
47	旅行	20
48	とき	19
49	ゲーム	19
50	チャット	19
51	ホーム	19
52	手段	19
53	場所	19
54	生活	19

頻度順位	単語	単語数	サンプル度数
31	いる	22	22
32	メール	22	22
33	個人	22	22
34	自宅	22	22
35	特に	22	22
36	商品	21	21
37	買い物	20	20
38	いい	19	19
39	その	19	19
40	手段	19	19
41	場所	19	19
42	連絡	19	19
43	よる	18	18
44	チャット	18	18
45	利用する	18	18
46	いろいろな	17	17
47	ゲーム	17	17
48	リアルタイム	17	17
49	医療	17	17
50	仕事	17	17

【参考文献】

[1] Couper, M.P., Baker, R.P. and others (1998). *Computer Assisted Survey Information Collection*, John Wiley.

[2] Dillman, D.A. (2000), *Mail and Internet surveys: The Tailored Design Method*, second edition, John Wiley.

[3] Grossnickle, J. and Raskin, O. (2001). *Handbook of Online Marketing Research*, McGraw-Hill.

[4] Groves, R, Dillman, D.A. and others (2002). *Survey Nonresponse*, John Wileys.

[5] Groves, R.M. (1989). *Survey Errors and Survey Costs*, John Wiley.

[6] Hayashi, C.（1998）. What is Data Science? Fundamental Concepts and Heuristic Examples, in *Data Science, Classification, and Related Methods*, p.40 - 51.

[7] Iversen, G.R. (1991). *Contextual Analysis*, Sage Publications, USA.

[8] Lebart, L, Salem, A. and Berry, L.（1998）. *Exploring Textual Data*, Kluwer Academic Publishers.

[9] McEnery, T. and Wilson, A.（1997）. *Corpus Linguistics,* Edinburgh University Press, Edinburgh, United Kingdom.

[10] Oakes, M.P. (1998). *Statistics for Corpus Linguistics*, Edinburgh University Press, Edinburgh, United Kingdom.

[11] Ohsumi, N. (2000), From Data Analysis to Data Science, in Data Analysis (invited paper), Data Analysis, Classification, and Related Methods, 329-334, Springer-Verlag Heiderberg.

[12] Salant, P. and Dillman, D.A. (1994). *How To Conduct Your Own Survey,* John Wiley.

[13] Sphinx Devéloppement (1998). Sphinx Survey: Plus2 & Lexica Editions, Software for Surveys, Statistics and Text Analysis, Reference Manual version 2 for Windows, SCOLARI, Sage Publication Software.

[14] Couper, M.P. (2003). The Internet and Other Survey Opportunities, JMRA 第 33 回トピックスセミナー「インターネット調査とそれを巡る諸調査法の可能性」，配布資料，2003 年 10 月 23 日，東京.（†）

[15] JMRA 研修セミナー「インターネット調査を検証する−質の評価と標準化に向けて−」，2003 年 6 月 10 日～12 日，東京.（‡）

[16] NTT コミュニケーション科学基礎研究所監修，池原悟，宮崎正弘，白井諭他編集（1999）．日本語語彙大系，岩波書店.

[17] ㈱博報堂インタラクティブカンパニー編 (2000). インターネットマーケティング，日本能率協会マネジメントセンター.（82-82 ページで紹介）

[18] 加賀野井秀一（1995）．20 世紀言語学入門，講談社現代新書 1248.

[19] 加賀野井秀一（1999）．日本語の復権，講談社現代新書 1459.

[20] 佐藤武義(1997)．概説日本語の歴史，朝倉書店.

[21] 山本夏彦（2000）．完本文語文，文藝春秋社.

[22] 小池清治(1993)．日本語はいかにつくられたか，ちくまライブラリー25.

[23] 小池清治他編集(1997)．日本語学キーワード事典，朝倉書店.

[24] 松本祐治，今井邦彦他（1997）．言語の科学入門，岩波講座「言語の科学」第 1 巻.

[25] 杉山明子（1984），社会調査の基本，現代人の統計-３，朝倉書店.
[26] 西平重喜（1979），統計調査法，補訂版，培風館.
[27] 西本一志，角康之，門林理恵子，間瀬健二，中津良平（1998）．マルチエージェントによるグループ思考支援，電子情報通信学会論文集，D-I Vol. J81-D-I No.5 ，478-487.
[28] 全文検索システム協議会編（1999）．全文検索システムとは何か？，第１部，1-63.
[29] 大隅昇，Lebart, L.他（1997）．テキスト型データの統計解析システム−SPAD.T/J−，第 11 回日本計算機統計学会シンポジウム.
[30] 大隅昇，Lebart., L. 他(1995)．記述的多変量解析法，日科技連出版社.
[31] 大隅昇，丸岡吉人他(1997)．自由回答データの解析法についての提案−実験調査におけるいくつかの試み−，第 25 回日本行動計量学会大会.
[32] 大隅昇（1989），統計的データ解析とソフトウェア，NHK 放送出版.
[33] 大隅昇（1997）．第 17 回 JMRA トピックスセミナー資料，1997.6.25.
[34] 大隅昇（2000）．「調査環境の変化に対応した新たな調査法の研究」報告書，文部省科学研究費特定領域研究, ミクロ統計データ, 公募研究（研究課題番号：09206117）.
[35] 大隅昇（2000）．定性情報のマイニング−自由回答データの解析−，ESTRELA，第 74 号，5 月号，14-26.
[36] 大隅昇他(1996)．テキスト型データの解析について，第 10 回日本計算機統計学会シンポジウム.
[37] 長尾真，黒橋禎夫他（1997）．言語情報処理，岩波講座「言語の科学」第 9 巻.
[38] 長尾真編（1996）．自然言語処理，岩波講座「ソフトウェア科学」第 15 巻.
[39] 渡部勇，三木和男，新田清，杉山公造（1995）．ハイブリッド発想支援システム：HIPS，計測自動制御学会第 17 回システム工学部会研究会資料.
[40] 統計数理研究所，ISM シンポジウム 2003 予稿集「インターネット調査の現状を検証する−調査法としての評価方法と標準化をどう考えるか−」, 2003 年 3 月 25−26 日.
[41] 北原保雄(1997)．概説日本語，朝倉書店.
[42] 林知己夫（2000）．これからの国民性研究−人間研究の立場と地域研究・国際研究から計量的文明論の構築へ−，統計数理，48，1，33-66.
[43] 鈴木達三，高橋宏一（1998），標本調査法，シリーズ＜調査の科学 ２＞，朝倉書店.

（注１）第 I 部で挙げた参考文献と重複している文献がある.
（注２）（†）（‡）にあげた資料には，インターネット調査に関する実験調査の結果が詳しく報告されている.

■ 資料 ■　用いた質問の例

ここにあげた質問群は，8．2 節で示した実験調査で用いた２種の調査（生活意識編，インターネット編）の自由回答質問の一部である．調査票の中から引用したので質問番号などはそろえていない．

○生活意識編から

問１．はじめに、あなたの普段の生活での「気持ち」についてお伺いします。

［１～４は省略］

5.　あなたにとって、一番大切だと思うものは何ですか。
１つだけあげてください。（どんなことでもかまいません。）

6.　ではこの他に大切なものとして何がありますか。いくつでもあげてください。
［上記，２問が自由回答］

○インターネット編から

問１．まず、あなたの「インターネット」に対する印象や知識についてお伺いします。

1.　「インターネット」と聞いてあなたが思い浮かべる印象や事柄を、以下に具体的にご記入ください。

問３．次に、あなたと「インターネット」とのかかわりについてお伺いします。

1.　あなたご自身にとって「インターネット」は、どのようなことがらに活用できると思いますか。どんなことでも結構ですので、以下になるべく具体的にご記入ください。
2.　では、一般的に「インターネット」は、どのようなことがらに活用できる と思いますか。なるべく、他にはないような活用法を、どんなことでも結構ですので、以下になるべく具体的にご記入ください。

５．あなたが登録している「メールマガジン（メルマガ）」の数をお知らせください。
※「メールマガジン」とは電子メールを媒体とした登録制の雑誌のことです。
（あてはまるものを１つ）
［ここで，選択肢から選択］

6.　では、その登録されている「メールマガジン」を具体的にお知らせください。
（いくつでも列記する）
［自由回答］

3. インターネットに、直接あるいは間接に関連した法規や法律の議論がさかんになり、すでに制定されたものもあります。次にあげるもののうち、あなたが必要だと思うものをお知らせください。
［以下の選択肢について,「必要だと思う」「とくに、必要とは思わない」「名称は知っているがわからない」「知らない」から選択］

プロバイダー法（プロバイダー責任法）
不正アクセス禁止法
インターネットによる情報の流通の適性化に関する法律
迷惑メール商法
情報公開法
住民基本台帳法
個人情報保護法（個人情報の保護に関する法律）
盗聴法（犯罪捜査のための通信傍受に関する法律）
人権擁護法
青少年有害社会環境対策社会法

4. 上記以外に、必要だと思う法律があれば、具体的にお知らせください。
［自由回答］

4. ところで「インターネットを利用した調査」では、「ひとりで何人もの回答者に『なりすます』ことができる」という意見があります。これについて、あなたの体験をお聞きします。(あてはまるものを１つ)

○調査の回答で「なりすまし」を行ったことがある
○調査の回答で「なりすまし」にちかいことを行ったことがある
○「なりすまし」は行ったことがない
○そもそもこのことに関心がない
○答えたくない

※ここで、「調査の回答で『なりすまし』を行ったことがある」または「調査の回答で『なりすまし』にちかいことを行ったことがある」を選んだ方に伺います。

5. 調査の回答で「なりすまし」を行った、あるいはそれに近いことを行った理由はなんでしょうか。なるべく、具体的にお答えください。
［自由回答］

6.　「インターネットを利用した調査」では、
「対象者（回答者）は本音で答えやすい」という意見があります。あなたはこの意見についてどのようにお感じになりますか。（あてはまるものを１つ）
［「全くそう思う」「ややそう思う」「あまりそう思わない」「全くそう思わない」から選択］

7.　あなたがそのようにお答えになった理由を、具体的にお知らせください。
［上を選んだ理由を自由回答］

2.　ところで、「インターネットはその時々によっていろいろな自分になれることが魅力である」
という意見があります。あなたはこの意見についてどのようにお感じになりますか。
（あてはまるものを１つ）
［「全くそう思う」「ややそう思う」「あまりそう思わない」「全くそう思わない」から選択］

3.　あなたがそのようにお答えになった理由を、なるべく具体的にご記入ください。
［上を選んだ理由を自由回答］

◆なお，２つの調査のいずれにおいても，「Web 調査」については調査票の終わりに次の質問を用意した.

Q：この調査に関するご意見、ご感想などがあれば、下記の欄にご自由にお書き下さい。
［自由回答のテキスト・ボックスを用意］