香 川 大 学 経 済 論 叢
第93巻 第3号 2020年12月　45－71

テキスト分析の理論的基盤：頻出語分析の記号学的考察(1)

繁　本　知　宏

Ⅰ．本研究の目的と構成

近年，テキスト分析が多くの研究領域において広く用いられるようになっている。もともとテキスト分析は欧米でマスコミ研究を中心に発展してきた分析手法であり，わが国でも経営学や経済学，社会学，看護学など幅広い分野で活用が進んできた。会計学の分野においても，有価証券報告書などの文書を分析対象としたテキスト分析が増加の兆しを見せつつある。

テキスト分析が普及するにつれて分析手法も多様化し，さらには発達が著しい人工知能の力を借りた分析も行われるようになっている。もっとも，複雑な分析手法であっても，テキスト分析の出発点は形態素解析であり，その次に行われるのは形態素（意味を持つ最小の言語単位）ごとの出現頻度の分析（頻出語分析）である。すなわち，テキスト分析を用いた研究では必ずと言って良いほど頻出語分析が行われる。

ところで，頻出語分析の結果は何を表すのだろうか。換言すれば，ある語の出現頻度が高いことは何を意味するのだろうか。テキスト分析を行った多くの先行研究において，出現頻度の高い語をもとに分析対象文書の主題を推測することが行われているが，なぜ頻出語が文書の主題を表すと言えるのか。非常に素朴な疑問であるが，これらの疑問に対し理論的な回答を提示している研究は意外なほど少ない。

（1）　本稿はJSPS科研費（課題番号「19K02014」）の助成を受けた研究成果の一部である。JSPSの助成に対し感謝の意を表する。

しかし，テキスト分析が理論に根差した科学的な分析手法としての地位を強固なものにするためには，分析の出発点である頻出語分析の理論的基盤を明確化する必要があろう。本稿はこうした問題意識に基づき，頻出語分析の理論的基盤を明らかにすることを研究の目的に据える。考察を進めるに当たっては，頻出語分析の理論的基盤に触れている数少ない先行研究が共通して言及している記号学を分析枠組みとして用いる。

本稿の構成は次のとおりである。はじめに第Ⅱ章ではテキスト分析の定義とテキスト分析において広く用いられる分析手法を整理する。次に第Ⅲ章では本稿の研究目的に関連する先行研究を整理した上で，リサーチ・クエスチョンを設定する。これを受けて第Ⅳ章では分析枠組みとなる記号学の理論を整理する。なかでもコードモデルと呼ばれるモデルに焦点を当てる。続く第Ⅴ章ではコードモデルに依拠しつつ頻出語分析の理論的基盤を明らかにし，第Ⅵ章では文書中に出現する語の重要性を表し頻出語分析でもしばしば用いられる TF-IDF の理論的基盤を明確化する。さらに第Ⅶ章ではコードモデルの限界を指摘した上で，それを克服するための理論を提示し，定量的な頻出語分析に続けて質的分析を追加実施することの妥当性を論じる。最後に第Ⅷ章では本稿の結論と限界ならびに残された課題を述べて本稿を締めくくる。

Ⅱ．テキスト分析の定義と分析手法

1．内容分析

古くから「内容分析」（Content Analysis）という分析手法がある。内容分析という言葉が最初に用いられたのは 1940 年のことであった（Kuckartz［2014］訳本 42 ページ）。Krippendorff［1980］訳本 7 -20 ページに依拠して内容分析の歴史を簡潔に振り返ると，内容分析の原型は 18 世紀頃まで遡ると言われるが，20 世紀になると新聞の大量印刷化が進み，世論の掌握に対する関心が高まってきたことを背景に，新聞の定量的分析に関心が寄せられるようになった。初期の定量的分析は話題のカテゴリー別の記事量を計測するというアプローチであり，この種の分析は今日でも行われている。その後，ラジオ放送

など新聞以外のマスメディアの発達や，マスメディアの社会的，政治的影響力の増大などの要因から内容分析は第二の発達局面を迎えた。さらに第二次世界大戦におけるプロパガンダ分析に内容分析が適用される中で洗練されていった。そして戦後になって，Berelson［1952］が内容分析に関する最初の統合的な指針を提示したことを契機として，多くの研究分野において内容分析が用いられるようになっていった。加えて 1950 年代後半から 1960 年代以降は内容分析にコンピュータを利用することが一般的となった。

内容分析の定義は論者によって様々である。Berelson［1952］p. 18 は「表明されたコミュニケーション内容を客観的，体系的，そして量的に記述するための研究技法」とする一方，Stone *et al.*［1966］p. 35 は「テキスト内の特定の特徴を体系的かつ客観的に推論するための研究技法」としている。また，Weber［1990］p. 9 は「テキストから妥当な推論を行うための一連の手続である。その推論はメッセージの送り手，メッセージ自体，メッセージの受け手について行われる」とし，Neuendorf［2002］p. 1 は「メッセージの特徴の体系的，客観的，量的な分析」と定義している。そのほか，Riffe *et al.*［2014］訳本 4 ページは「ルールに従いコミュニケーションのコンテンツをカテゴリへ体系的に分類し，そして，統計的な手法を用いてこれらのカテゴリの関係を分析する研究方法」と述べる。さらに Krippendorff［1980］訳本 21 ページは「内容分析とは，データをもとにそこから（それが組み込まれた）文脈に関して再現可能で（replicable）かつ妥当な（valid）推論を行うための一つの調査技法」とする。

これらの定義を俯瞰すると「客観的」「体系的」「再現可能」といった言葉が目立つが，これは内容分析を科学的な分析手法として位置付けようとしていることの現れであると解釈できる。他方，「量的」「統計的な手法」という表現については，定義に含む論者と含まない論者に分かれる。つまり，すべての論者が内容分析に定量的分析を求めている訳ではない。特に Krippendorff［1980］は，Berelson の定義に含まれる「量的」という要請は制約が大きすぎるとして自らの定義から意図的に除外したと述べている。その理由としては，質的分析も，プロパガンダ分析や精神的疾患の治療をはじめとする多くの分野で成功を

収めていることに加え，定量的な説明を行うとしても解釈は質的なプロセスであることを挙げている。

もっとも，その Krippendorff［1980］も強調しているように，多くの科学的営為において定量化は重要である。内容分析においても定量的分析を行えば「客観的」「体系的」「再現可能」といった要件を満たしやすくなるであろう。また近年は，高性能なコンピュータと操作性に優れたソフトウェアが急速に普及し，大量のデータの解析が容易化してきた。こうしたこともあって，内容分析の遂行に際してはコンピュータの利用が一般的となった。内容分析におけるコンピュータの利用場面としては，まず，膨大なデータに含まれる語や文の数を数えたり，語の相対的な出現頻度を計算したりといった，データの全体的な特徴を定量面から把握するための作業が挙げられる。また，データのサンプリングや統計的な解析においても複雑かつ正確な処理が可能となる。こうした場面におけるコンピュータ利用は研究結果に対する「偏った，不完全な，そして非常に選択的な印象」（Krippendorff［1980］訳本 188 ページ）の回避にも役立つ。次の利用場面としては，研究者が予め設定したカテゴリーにデータを分類するコーディングにおいて，コンピュータは威力を発揮する。カテゴリー設定とコーディング・ルール作成は研究者自身の理論視角や仮説を基に自ら実施する必要があるが，その後のコード別の振り分け作業は曖昧さや恣意性を排除した形でコンピュータに自動的に行わせることができる。このことは Krippendorff［1980］が強調する内容分析における再現可能性と妥当性の確保に大きく寄与する。

本稿においては，後述するように定量的分析と質的分析の循環的，相乗的な利用を重視する立場を採るため，Krippendorff［1980］の見方を支持したい。

2．テキストマイニング／テキスト分析

近年，「テキストマイニング」（Text Mining）あるいは「テキスト分析」（Textual Analysis）という言葉を頻繁に目にするようになった。テキストとは文字・記号列の集合をいい（金［2018］ 1 ページ），文書は典型的なテキストである。

マイニングとは価値のある情報を掘り出すことを意味する。したがってテキストマイニングとは，テキスト情報から価値ある情報を掘り出す行為である。テキストマイニングは1990年代中頃から用いられるようになった用語であり，データマイニングから派生している（金［2012］1ページ）。特に最近ではビッグデータを活用したデータマイニングあるいはテキストマイニングに注目が集まっている。

データマイニングとテキストマイニングはいずれも分析素材となるデータに潜在する価値あるパターンを試行錯誤しながら見つけ出し，有用な情報を得ることを目指している点で共通している（Feldman and Sanger［2007］訳本1ページ）。しかし，データマイニングの分析素材が構造化されたデータであるのに対し，テキストマイニングは構造化されていないデータ（テキスト情報）である点に違いがある。例えば企業の開示情報でいえば，会計数値は組織的に記録された情報が，会計基準という社会的な規則体系に従って加工され，財務諸表という標準化された様式で表示される。つまり財務諸表の数値は構造化されたデータである。しかし，定性的な記述情報については，開示ルールが最低限開示すべき事項を定めているだけであり，文章の形式や具体的内容は企業によって多種多様である。つまり定性的な記述情報すなわちテキスト情報は構造化されていない（より正確には構造化の度合いが低い）データであるといえる。

ところで，内容分析，テキストマイニング，テキスト分析という3つの用語の相違については，現状では厳密に使い分けられている訳ではなく，論者によって区々というのが実情であろう。この点について，特に区別する必要がある場合を除き，本稿では以下，テキスト分析という用語を用いる。その理由は，内容分析という用語を用いると研究の焦点が文書の意味内容の分析にあると受け止められるおそれがあり，他方のテキストマイニングという用語を用いると知識発見のためのデータ処理技術重視との印象を与える可能性があると考えたからである。(2)

3．テキスト分析における分析手法

⑴ 潜在的テキスト分析と顕在的テキスト分析

喜田［2018］50-51 ページによると，テキスト分析は大別すると潜在的テキスト分析と顕在的テキスト分析の２つに分けられる。潜在的テキスト分析とは，分析対象の文書を引用したり文章例を挙げたりしつつ，研究者が当該文書を主観的に解釈する分析手法である。潜在的テキスト分析は研究者の多様な関心を満たす方法である反面，客観性に欠ける面があると指摘されている。他方の顕在的テキスト分析とは，研究者が設定したキーワードの出現頻度や，キーワードで著わされるような特性に注目し，ある言語全体の中での出現頻度や，同一文書における他の語の出現頻度との比較などを中心に研究がなされる。定量的分析が行われることが特徴であり，一般にテキスト分析といえば顕在的テキスト分析を指す。

⑵ 頻出語分析

テキストの定量的分析では，研究者が自らの分析視角から設定したキーワードの出現頻度（TF : term frequency）を調べる頻出語分析が最も基本である。ある語が文書の中で多数出現すれば，その文書はその語に強く関連すると考えるのである（黒橋・柴田［2016］125 ページ）。その中でも最も初歩的な頻出語分析はキーワードの出現回数を単純に数えることだが，語の出現回数は文書の長さに影響を受けるため文書間の比較が難しい。そこで文書中の語の総数（すなわち文書の長さ）で正規化した出現頻度を用いることもある（前田・西原［2017］33-34 ページ）。

ただ，出現頻度が高くても多くの文書に出現する語は一般的に使われる語であって重要性はそれほど高くない，とも考えられる。そこで，少数の文書に

（２） 無論，これは筆者の主観に基づく見方に過ぎない。Krippendorff［2019］pp. 408-412 は，内容分析はテキストが用いられる文脈から再現可能で妥当な推論を行うための研究方法，テキスト分析は語数と KWIC（Keyword in Contents）を基本技術としてテキストの文字列や構成，一致，構文構造，割り付けを説明する方法，テキストマイニングはコンピュータを用いた大量テキストの分析技術と説明している。

しか出現しない語の重み付けを大きくし，多数の文書に出現する語の重み付けを小さくする逆文書頻度（IDF：inverse document frequency）を用い，TF と IDF を掛け合わせて語の相対的な重要性を示す重み付け指標（TF-IDF）がしばしば用いられる。IDF の計算式は，文書集合に含まれる全文書数を N とし，ある語 t が出現する文書数を df(t)とすると，一般に $IDF(t) = \log \frac{N}{df(t)}$ と計算される[(3)]。対数で表すのは N が非常に大きく，かつ df(t)が非常に小さい場合に IDF(t)が過大になることを抑制するためである（岸田［1998］84 ページ）。

こうした語の出現頻度に着目した分析は，文書の主題を推測したり，他の分析手法を用いる際に分析対象とする語を限定したりするために使われることが多い。また，一般的な辞書には格納されていない専門用語を探し出し，マニュアルで辞書を拡充する際にも役立つ。もっとも，頻出語分析は語を文脈から切り離すため，文章の書き手がどのような意味で当該語を用いているかは判別しにくいという欠点がある。例えば会計学の論文に出現する「業績」という語は，文書から切り離した状態だと，その意味が「売上高」なのか「利益」なのか，それとも「著者の学術的成果」なのか判別できない。

⑶　共起ネットワーク分析とクラスター分析

そこで，文書中における語の意味を探求するための分析手法として，共起ネットワーク分析やクラスター分析がある[(4)]。共起ネットワーク分析とはどの語

（3）　IDF の算出方法は複数存在する。例えば全ての文書に出現する語の重みが 0 とならないよう $IDF(t) = \log \frac{N}{df(t)} + 1$ とする方法や，全文書数 N の代わりに最も多くの文書に出現する語の文書数を用いる方法もある（岸田［1998］83-84 ページ）。また，自然対数が用いられることが多いものの，対数の底として 10 や 2 が用いられることもある。なお，第Ⅵ章で述べるように，2 を底とする IDF は情報理論の観点から特別な意味を持つ。

（4）　テキスト分析では知識発見プロセスにおけるデータの対話的分析が重視される。大規模な文書集合を対象とした分析では表示すべきパターンや特徴の数が極めて多くなるため，洗練されたグラフィカルインターフェースによって，パターンを認識する研究者の視覚的能力を刺激・活用する方法が重視される（Feldman and Sanger［2007］訳本 245-246 ページ）。こうした観点から共起ネットワーク分析とクラスター分析は情報量が比較的縮減されているため，研究者が解釈しやすいという特徴があるためよく用いられている。もっとも，情報量が少ないことはこれらの手法の弱点でもある。

とどの語が同一文書内で結びついていたのかを分析する方法であり，相互に結びついている語のグループから文書中に多く出現していた主題を探索することができる（阪口・樋口［2015］190ページ）。分析結果はグラフ理論に基づき描画されたネットワーク図で表示される。(5)

他方のクラスター分析とは，一群の対象のどれとどれが類似しているかを見つけ出すために用いられる様々な数学的方法をいい，階層的クラスター分析が中心的な方法である。階層的クラスター分析による分析結果は類似度の階層を示す樹形図（デンドログラム）によって表示され，複数のクラスターがどの階層で併合されているかを読み取ることができる（Romesburg［1989］訳本1-4ページ）。

共起ネットワーク分析とクラスター分析は，いずれも語と語の共起関係に着目し，文書の主題を探求するために用いられることが多い分析方法である。共起ネットワーク分析は文書作成者が使用している語の共起関係のネットワークを一覧することに重きを置く一方，クラスター分析は語の結びつきを樹形図で表すことにより語と語の近さをより明確にし，語の含まれるクラスターを明らかにする点に重きを置いている（高木［2016］85-86ページ）。

Ⅲ．先行研究とリサーチ・クエスチョン

1．先行研究

前章ではテキスト分析における定量的な分析方法である頻出語分析，共起ネットワーク分析，クラスター分析についてそれぞれの特徴を概観した。確かにこれらの分析方法は多くの先行研究で用いられており，多くの有益な知見をもたらしてきた。このことから帰納的に考えれば，これらの分析手法は，非構造データであるテキスト情報を分析するための手法としての地位を確立しているようにみえる。

しかし，テキスト分析の出発点である頻出語分析に関し，なぜ「ある語が

（5） 共起ネットワーク図の詳細は鈴木［2017］を参照されたい。またグラフ理論については宮崎［2015］が詳しい。

文書の中で多数出現すれば，その文書はその語に強く関連する」といえるのか。換言すれば，なぜ頻出語がその文書の主題を表すと言えるのか。経験的，直感的には考えるに値しない疑問かもしれない。しかしながら学術的には，経験や直感をもって頻出語分析を正当化する根拠にはなり得ない。多数出現する語と文書の主題の関連性を理論的に説明する必要がある。また同様に，語と語の共起関係を分析することが，なぜ文書の主題の推測に有効なのか。これについても共起度と文書の主題との関連性を理論的に説明する必要がある。

こうした点を明らかにしなければ，テキスト分析は理論的基盤を欠いた単なる経験的な分析手法だと評価されても仕方がない。ところが，テキスト分析を用いた先行研究やテキスト分析に関する学術書をみると，個別の分析手法の特徴に関する説明（例えば共起ネットワーク分析とはどのような手法か）や，分析結果を視覚的に表示するためのグラフィティカルインターフェースの理論的基盤に関する説明（例えば，共起ネットワーク分析における図はグラフ理論に基づいて描画されている）は詳細に加えられていても，テキスト分析の理論的基盤まで踏み込んで考察しているものは意外なほど少ないのが現状である。以下では数少ない先行研究として Krippendorff［1980］と井上［2016］の考察を整理する。

Krippendorff［1980］はテキスト分析が特定の理論に根差していると明確に述べている訳ではないものの，「メッセージのシンボリックな意味を探る手段だという点に内容分析の特徴がある」（訳本 23 ページ），「一般的にメッセージやシンボルを介したコミュニケーションが関与するのは直接に観察され得ないような現象に関してである」（訳本 24 ページ）といった記述がみられることから，記号学が意識されていることが推察できる。しかし同時に「メッセージのもつ意味は唯一ではない」（訳本 23 ページ），「意味は必ずしも共有されない」（訳本 24 ページ），「受け手は知覚したデータから，経済的環境の一部に関する特定の推論を行う」（訳本 24 ページ），「内容分析が行う作業は，シンボルを介したコミュニケーションを理解しようと試みる際，受け手が行う作業と同様の推論である」（訳本 24 ページ）といった記述もみられており，メッセージ発信

者の意図が受信者に必ずしも正確に伝わらないことを前提とした推論の意義を強調している。このことは，受け手によるメッセージの解読プロセスに注目し，そこに推論プロセスが組み込まれた伝達モデルが想定されていることを示唆している。このようなモデルを「推論モデル」[(6)]と呼び，伝達は証拠の提示と解釈からなるとする（Sperber and Wilson［1995］訳本３ページ）。

他方，井上［2016］は，テキスト分析において共起に着目することの理論的基盤について考察を加えている。井上［2016］はテキスト分析を「記号表現の分析手法」と位置付け，記号表現と記号内容が結合した記号は[(7)]，新たな記号表現となって別の記号内容と結合し新たな記号となる，というプロセスが繰り返されるとするBarthes［1957］の所説を取り上げる[(8)]。特定の記号表現に充当される記号内容は，他の記号との関係によって補足・修正され相対的に規定されるのである。その上で，記号表現間の関係は，特定の記号表現を参照する別の記号表現の頻度と，参照する記号表現間の関係の構造という２つの視点から観察でき，前者は共起率，後者は共起ネットワークで表されると論じている。記号表現間の共起率は，他の記号表現でも表せる記号内容が特定の記号表現でしか表現されていないという偏りを示しており[(9)]，記号表現間の関係に書き手が何らかの意味を込めていると解釈できる。共起ネットワークはテキスト全体における記号表現の参照構造をネットワーク図で示したものである。共起ネットワーク図を用いれば共起率の高い記号表現のクラスターを把握でき，クラスター間の結節点となる記号表現も浮き彫りにできる。

（６） 推論モデルの詳細はGrice［1989］を参照。とりわけ同書に再録されているGrice［1957］が推論モデルの核をなす。

（７） 井上［2016］は記号表現，記号内容をそれぞれ，Saussureがいうシニフィアンとシニフィエの意味で用いている。

（８） Barthesの所説はSaussureの記号学の影響を受け，発展させたものと位置付けられる（土田・青柳・伊藤［1996］14-15ページ）。

（９） この見方の背景には，記号内容と記号表現は結合関係にあるものの，ある記号内容は特定の記号表現と必ず結びつくという排他的な対応関係があるのではないとするSaussureの理論がある。当該理論は第Ⅴ章にて検討を加える。

2．リサーチ・クエスチョン

これら2つの先行研究は，濃淡の違いはあるにせよ，いずれもテキスト分析の背後に記号学の支えがあることを示唆している。とりわけ共起ネットワーク分析は，井上［2016］によってSaussure派の記号学に依拠していることが指摘されており，本章の序盤で提示した「語と語の共起関係を分析することが，なぜ文書の主題の推測に有効なのか」という疑問に対し一定の理論的回答を提示している。

しかしもう一つの疑問である「なぜ頻出語がその文書の主題を表すと言えるのか」という疑問は未解決のままである。すなわち，この疑問は語が出現する文書自体，あるいはその文書を作成する情報の発信者に向けられるものであるが，Krippendorff［1980］が依拠する推論モデルは情報の受信者に重点を置いている。また井上［2016］が依拠するBarthes［1957］は発信者の意図を考慮しているものの，記号の共起関係に焦点を当てている。したがって，いずれの研究もこの疑問に対する直接的な回答を与えていない。そこで以下では，「頻出語がその文書の主題を表す」という命題を支える理論的根拠を明らかにすることを本稿のリサーチ・クエスチョンとして，記号学の視点からアプローチすることを試みる。

Ⅳ．コードモデル

1．伝達と意味作用

記号学（semiology）とは「社会生活の内部で諸記号がどのような働きをしているのかを研究する学問」であり，記号の本質や記号を支配する法則を明らかにすることを目指している（Saussure［1916］訳本35ページ）。ここでいう「記号」とはその定義自体が記号学におけるひとつの論点であるが，ここでは「他の何物かの代わりに置かれたもの」（Eco［1973］訳本25ページ），つまり他の何かを代理して表すものとしておく。この「何か」には，人の外界に存在する有形無形の事物に限らず，人の内面にある心情や思考も含まれる。また，代理して表すものとしては，意図的に作られた人工的表現（例えば発話や文字

表記）が該当する点は異論がないものの，非意図的な表現を含むか否かは論者によって見解が分かれる。前者に限るとする代表は Saussure であり，後者の代表は Peirce である。[10] このような記号学の視点に立って，人間の伝達行為の基本的なプロセスを概観すると次のように説明できる（池上［1984］37-39 ページ）。

伝達（communication）とは，発信者が頭の中に描いている抽象的な思考内容のコピーを受信者の頭の中にも作り出す行為と言える。伝達を行う場合，発信者はまず，何らかの形で思考内容の存在を知覚できるよう表現しなければならない。こうして表現されたものがメッセージである。メッセージは，伝達の役目を果たすためには何かを意味するもの，すなわち記号（signe）により構成されなければならない。ここであるものが他のものを意味する働きを意味作用（signification）という。記号学でいう伝達には意味作用が伴わなければならない。伝達の目的を正確に達成するためには，メッセージを構成する記号とその意味は，発信者と受信者の間で共通の了解に基づいた決まりに従っていなければならない。この決まりがコード（code）であり，記号とその意味ならびに記号の結合の仕方を規定する。発信者はコードを参照しながら伝達内容を記号化してメッセージを作成し，何らかの経路を伝って受信者に到達する。そして受信者はコードを参照しながらメッセージを解読し，伝達内容を自らの思考に取り込む。

2．コードモデル

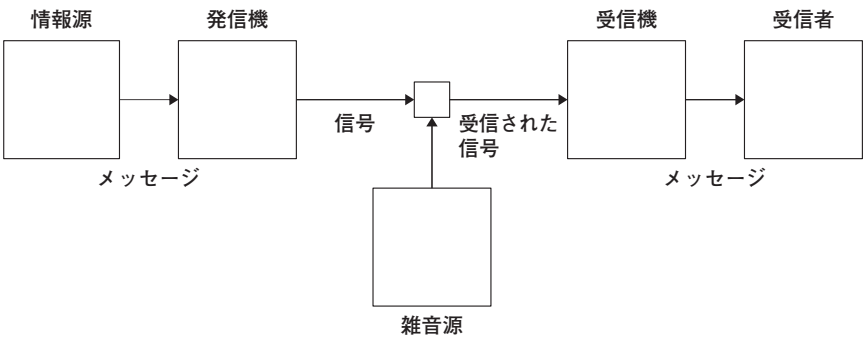
このように，伝達内容のコード化とコード解読によって伝達が達成されるとするモデルを，Sperber and Wilson［1995］にならい「コードモデル」と呼ぶ。コードモデルは模式図（図表１）を使い電気通信に例えて説明されることがあ

(10) Peirce を源流とする理論体系は一般に記号論（semiotics）と呼ばれる。Hawkes［1977］によれば，記号学という言葉は Saussure と同じヨーロッパ人が多く用い，記号論は Peirce と同じ英語圏国民が好む傾向があるという点が唯一の違いとされる。また Eco［1976］は，記号学は言語学に依存する一方，記号論は非言語的な現象も扱う点に違いがあると指摘すると同時に，記号論（学）の世界においても両者が厳密に使い分けられている訳ではないと指摘している。こうした点に鑑み，言語を扱うテキスト分析を念頭に置く本稿では記号学という表記を用いる。

る。図表１は，発信者が発信機という機械を使って発出したメッセージが受信側で受信され複製されるプロセスを表している。こうした一連の伝達行為が完全に行われるならば，換言するとメッセージの移動プロセスの中で一切の雑音（noise）[11]が発生しなければ，発信者の思考内容は受信者に完全に複製される。

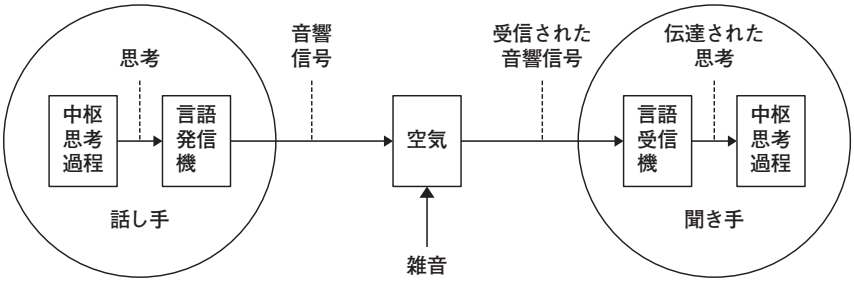
Sperber and Wilson［1995］訳本５-６ページはこれを人間の発話に当てはめて次のように修正した（図表２）。図表１と図表２を対比させると，図表１の

（図表１）



（出所）Shannon and Weaver［1949］訳本 64 ページ

（図表２）



（出所）Sperber and Wilson［1995］訳本６ページ

(11)　雑音は音に限らず完全な伝達を阻害する全ての要素を指す。

情報源と受信者はそれぞれ，図表２の話し手の中枢思考過程と聞き手の中枢思考過程に相当する。また図表１の発信機と受信機はそれぞれ，図表２の話し手の言語発信機と聞き手の言語受信機に相当する。さらに図表１の信号（電気信号）は図表２では音響信号（音）に当たる。なお，Sperber and Wilson［1995］は図表２の前提として，人間の言語はコードであること，およびコードが思考を音に結びつけることの２つを挙げている。この点について，図表２を文字を用いた伝達に置き換えれば，コードが思考を文字表記に結びつけることとなる。

図表２の伝達行為が完全に行われるためには，前述したように雑音が発生しないことが必要である。しかし，より重要なことは，話し手の思考はいかにして音や文字で表される言語（すなわち記号）に結びつけられるのかということ，ならびに話し手と聞き手が共有するコードは一体どのようなものなのか，という点である。次章では Saussure の所説に基づいてこれらを検討した後，テキスト分析における頻出語分析の理論的説明を試みる。

Ⅴ．頻出語分析の理論的説明

１．Saussure のコードモデル

人間の伝達行為のプロセスを Saussure の理論に従って分解していくと，まず人間が認識する事実を概念と呼ぶ。次に，概念を表現する働きをする聴覚映像[(12)]に概念が結合する。こうした結合によって語が脳内で喚起される（Saussure［1916］訳本 27-28 ページ）。すなわち，概念と聴覚映像の結合体である語が記号となる。そして Saussure は概念をシニフィエ，聴覚映像をシニフィアンと呼んだ。シニフィエとシニフィアンは結合関係にあるものの，あるシニフィエは特定のシニフィアンと必ず結びつくという排他的な対応関係があるのではな

(12) 聴覚映像は物理的な音に限らず，音の心的な刻印，つまり人間の感覚によってその存在が証拠づけられる表示を意味する（Saussure［1916］訳本 101 ページ）。さらに聴覚映像は音素に対応する視覚映像すなわち文字に置き換えることができ，短時間で消え去る音を文字として固定的に表示することが可能となる（Saussure［1916］訳本 34 ページ）。ただし Saussure は文字表記よりも音の方が言語にとって重要だと述べている（Saussure［1916］訳本 48-51 ページ）。

く，恣意的に結びつく（Saussure［1916］訳本 100-104 ページ）。ただし，ここでいう恣意的とは，特定の意図（時に悪意）を持った主観的な態度という，一般的な言葉遣いとしての「恣意的」とは意味が異なる点に注意しなければならない。Saussure がいう「恣意的」とは，現にそうなっているからそうした結びつき方をしているに過ぎない，つまり結びつき方に必然性がある訳ではないという意味である。しかし，シニフィエとシニフィアンの結合関係は個人が自由に選択できる訳ではなく，ラング(13)と呼ぶ規則体系に基づいて成立する。ラングは言語共同体が共有する一種の社会制度であり，個々人の脳内に存在する。個人はラングに従ってシニフィエとシニフィアンを結合させ，さらに発話や文字表記といったパロール（parole）を実行して具体的な状況における伝達行為を行う。

文書による伝達においては語が記号の役割を果たし，人間の知覚は記号の一側面であるシニフィアンに向けられる。このため，シニフィアンと記号が同義語として扱われることも少なくない。他方，記号のもう一つの側面であるシニフィエは直接知覚できず，シニフィアンの背後にその存在を窺い知ることができるに過ぎない。前述したとおり，シニフィエとシニフィアンの結合関係は恣意的であるため，ある文書で用いられている語の意味と，別の文書で用いられている同じ語の意味が同一である保証はない。しかしながら，文書の書き手と読み手が言語に関する社会制度（ラング）を共有していれば，完全に同一ではなくとも，読み手は書き手によるシニフィエとシニフィアンが結合した記号である語を解読して書き手の思考を自らの頭の中に複製できる。

前出の図表 2 に Saussure の理論を当てはめれば次のように説明できる。まず発信者の思考は，中枢思考過程においてラングに従いシニフィエとシニフィアンを結合させて記号化される。次に言語発信機でパロールを実行する。これが発話であれば声すなわち空気の振動が音響信号となって受信者の言語受信機

(13)　ラングもコードの一つであり，慣用によって打ち立てられる平均値といえる（Eco［1973］訳本 87-88 ページ）。この結果，ラングに従えば，完全に同じではないにせよ，ほぼ同じ形で，同じ概念に結びついた同じ記号を，すべての人間が再現できることになる（Saussure［1916］訳本 30 ページ）。

に到達する。雑音が全くなければ音響信号が完全な形で受信者に伝わるが，雑音が混じれば音響信号の完全な伝達が阻害される。そして受信者は受信した記号をラングに従って解読し，発信者の思考を自身の中枢思考過程に取り込む。こうして一つの伝達が完了する。

2．頻出語分析の理論的基盤

Saussure のモデルによれば，シニフィエとシニフィアンがラングに基づいて結合し，これによって語が脳内で喚起される。シニフィエとシニフィアンは結合関係にあるものの，あるシニフィエは特定のシニフィアンと必ず結びつくという排他的な対応関係があるのではなく，恣意的に結びつくとされる。つまり，ラングに基づけば，あるシニフィエが複数のシニフィアンと結合できるにもかかわらず，特定のシニフィアンと多く結びついているとすれば，それは偶然そうなったのではなく，発信者が意図的に行った結果と解釈できる。

したがって，一つの文書中に特定の記号（語）が多く出現することは，書き手の脳内において特定のシニフィエが頻繁に喚起されていることに加え，書き手がその語を使うことに何らかの意味を込め意図的に引き起こした結果だと解釈できる。このように考えると，文書中における特定の語の出現頻度の分析は，単なる語の数え上げではなく，文書作成において書き手の思考が強く向けられている概念，すなわち文書の主題を明らかにするための，記号学に根差した分析手法と考えることが可能である。

続いて次章では，文書中に出現する語の重要性を表し頻出語分析でもしばしば用いられる TF-IDF の理論的基盤について考察する。TF-IDF については，実務的に有用だが理論的根拠がない（Jagadeesh and Wu［2013］p. 714）ともいわれるが，本当にそうなのか。以下では Saussure のコードモデルを発展させた Eco の s コード理論とそれを支える情報理論を基に考察する。

Ⅵ．TF-IDF の理論的基盤

1．Eco の s コード理論

Saussure はシニフィエとシニフィアンの結合に関しラングという言語共同体が共有する規則体系がコードの役割を果たすとし，ラングこそが言語研究の中心と考えた。実際，Saussure［1916］はラングの性質の解明を中心的な目的に据えている。そしてコードモデルの系譜に属する研究者も同様に，ラング（あるいはコード）に焦点を当てた研究成果を残している。そうした研究者のうち Eco［1976］はコードの精緻化を試みた。Saussure はラングを社会的な規則体系の総体と考えたが，伝達プロセスの局面ごとに細分化されたラングが存在するとは明言していない。これに対し Eco［1976］は，統語，意味，音韻のそれぞれ[14]に体系（s コード）が存在し，音韻体系が統語体系を通じて意味体系と結びつけられる段階の規則体系をコードと考えた。つまり，コードは s コードという部分によって構成されているとしたのである。

さらに，s コードの存在を前提として，s コードがどの程度の情報を伝達し得るかという情報理論的な観点から考察が加えられた。具体的には，伝達可能なすべての情報／s コードが許容する範囲で伝達可能な情報，および発信点において発信可能な情報／実際に発信され受信された情報という 2 つの軸を用いて 4 つに場合分けし（図表 3），記号学の立場からは③と④が関心領域だと指

（図表 3）

①伝達可能なすべての情報 発信点において発信可能な情報	③ s コードが許容する範囲で伝達可能な情報 発信点において発信可能な情報
②伝達可能なすべての情報 実際に発信され受信された情報	④ s コードが許容する範囲で伝達可能な情報 実際に発信され受信された情報

（出所）Eco［1976］訳本 I 巻 64-66 ページをもとに筆者作成

(14)　言語学における統語，意味，音韻とはそれぞれ，文の構造，語や文の意味（ただし意味の厳密な定義は難しい），言語における音の機能を指す（風間ほか［2004］）。

摘した。(15)

ここでいう情報とは，出現確率がすべて等しい要素から成る体系の中で，ある事象が生起する確率を意味する。換言すれば，情報とはメッセージを選択する時の選択の自由度である（Shannon and Weaver［1949］訳本 24 ページ）。生起確率が等しい n 個の事象のうち，ある特定の事象が生起する確率は $\frac{1}{n}$ である。情報理論では 1 つの事象が持つ情報量を $\log_2 n$ ビットと定義する。

伝達では情報量が多ければ良いとは限らない。情報量が多いことは，裏を返せば多数の記号からの選択が必要なことを意味する。現実の伝達においては，あまりに多数の選択を限られた時間内で実行することは困難である。このため，記号の選択肢を予め絞っておくことが望ましい。s コードは，等確率の事象がランダムに選ばれる状況に対して，ある種の制限を加える機能を持つ。すなわち s コードは，ある結合は可能だが他の結合は認めないという制限を加えることによって情報量を減らす機能を果たす。s コードが機能すれば n 個の事象は m 個（$n > m$）に絞られ，ある特定の事象が生起する確率は $\frac{1}{m}$ となる。$n > m$ であるから，$\log_2 n > \log_2 m$ であり，s コードが機能することによって特定の事象が持つ情報量は減少する。この結果，伝達が成功する可能性が高まると考えられる。

2．TF-IDF の理論的基盤

ところで前述したとおり，TF-IDF は語の出現頻度（TF）と逆文書頻度（IDF）の積として表され，一般に IDF は $IDF(t) = \log \frac{N}{df(t)}$（N：文書数，df(t)：語 t が含まれる文書数）として計算される。IDF の右辺は自然対数が用いられることが多いが特に決まりはない。ここで，N 個の文書から無作為に一つの文書を選んだ時に語 t が含まれる確率を $p = \frac{df(t)}{N}$ とし，IDF の右辺の対数の底を 2 とすれば，$IDF(t) = \log_2 \frac{1}{p} = -\log_2 p$ と表すことができる。これは前述した

(15) ①は③との関連においてのみ，②は④との関連においてのみ，辛うじて記号学と関連し得ると Eco は述べている。

情報理論における情報量であり，さらに言えば自己情報量（自己エントロピー）である。マイナス符号がついていることは，生起確率が低い事象ほど情報量が大きくなることを意味する。

こう考えると，TF-IDF は，語の出現頻度すなわち書き手の思考が強く向けられている概念の表出濃度と，情報量を掛け合わせた指標である。そしてこのことから，TF-IDF は文書中における語の重要度を表す，理論的基盤を有する指標と位置付けることができる。

Ⅶ．コードモデル的アプローチの限界と克服

１．コードモデルの限界

これまで多くの研究において頻出語分析は，理論的基盤が必ずしも明らかにされないまま広く用いられてきたが，前章までの考察によって，理論的基盤を有する分析手法であることが明らかとなった。ただ，その基盤であるコードモデルは，一連の伝達行為が完全に行われる理想的な状況を想定している。すなわちコードモデルでは，メッセージの移動プロセスの中で一切の雑音が発生することなく，発信者の思考内容が受信者に完全に複製されることが仮定されている。発信者と受信者が機械であればこうした理想的な姿が実現するかもしれない。

しかし人間の伝達行為は不完全であるし，柔軟でもある。例えば同じ日本語を母国語とする者どうしであっても，それぞれの生来的，社会的，文化的背景が違えば，それぞれの頭の中にあるコードが異なることはあり得る。他方，発信者のメッセージに欠損があったり文法的な誤りがあったりしても，受信者はコードから敢えて逸脱した解釈を行って発信者の意図を読み取ることができる。このように現実の世界では，コードモデルでは説明できないような事象が多々観察される。換言すれば，コードモデルは理想的な状態を前提とした純粋な理論としての性格が強いがゆえに，現実に対する説明力に弱さがあるのも事実である。

このことは，コードモデルを基礎に据える頻出語分析にも弱点をもたらす。

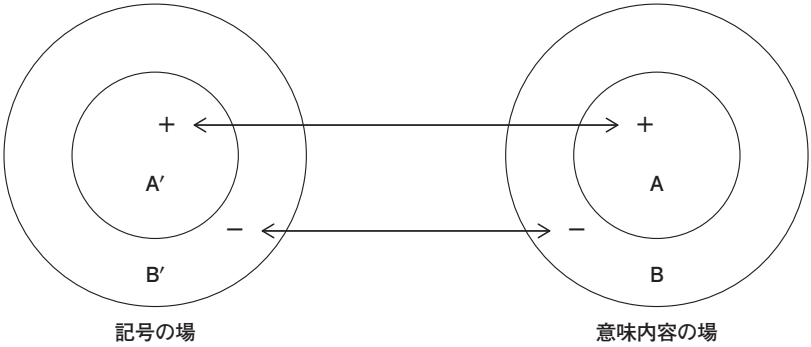
その弱点とは，語をコンテクストから切り離す結果，書き手が込めた語の意味を的確に捉えた分析に失敗する可能性があることである。とりわけ「行間を読む」ことが求められる小説や比喩の多い詩などコンテクストに強く依存するテキストを分析対象とする時に大きな問題となる。では，頻出語分析を行っても現実の世界から遊離した結論しか導き出せないのだろうか。こうした疑問に対する回答のヒントを与えてくれるのが Prieto のモデルであろう。

2．Prieto のモデル

Prieto［1972］は，シニフィエ／シニフィアンについて，Saussure が述べたようなシニフィエは概念，シニフィアンは聴覚映像という単純な構造に留め置かず，次のように集合概念を用いたやや複雑な構造を提示した。すなわち，特定の信号[16]およびこれと同じコードに属する他の信号によって許容されるメッセージ群の全体集合を意味内容の場（champ noétique）と呼び，信号はこの場を２つに分割するとした。その上で，１つは当の信号が許容するメッセージ群から成るクラス（シニフィエクラス）であり，他方は当の信号によって排除されるクラスであるとした。こうした全体集合とその分割という関係は指示するものとしての信号にも当てはまる。特定のコードに属する全ての信号群の全体集合を記号の場（champ sématique）と呼ぶ。記号の場はある信号が発信されることにより，信号が属するシニフィアンクラス（同一のシニフィエクラスを有する全信号群のクラス）とその補集合に分割される。これらの点について図表４を用いて説明すると，受信者は，発信者によって発信された信号が特定のシニフィアンクラス（円 A′）に属していて，その補集合（環形 B′）に属していなければ，発信者が伝達しようとしているメッセージは対応するシニフィエクラス（円 A）に属していて，その補集合（環形 B）には属していないことを知る。これは記号の場の分割と意味内容の場の分割との間の対応関係があってこそ成

(16) Prieto は，ある事象に関する不確定性を全てもしくは部分的に除去するものを指標と呼び，そのうち指示を与えるために意図的に作り出された人工的な指標を信号と定義している。

（図表 4）



（出所）Prieto［1972］訳本 48 ページの図を一部修正

立する。そしてこのシニフィアンクラスとこれに対応するシニフィエクラスが一体となって統合記号（séme）[17]と呼ぶ実体を作る。

ところで，信号が許容するのはメッセージ群であり，複数のメッセージが選択肢として含まれる。したがって，受信者は複数の選択肢から特定のメッセージを選択して信号に帰属させることが求められるが，この際に補助的な指示を与えるのが情況であると Prieto は述べる。情況は記号行為に関係なく受信者が知っているあらゆる事実を指す[18]。つまり，Prieto のモデルはコードモデルの中でも完全なコード依存型ではなく，Saussure らが捨象していたコンテクストをモデルに取り込んだ点に大きな知的展開が見出される。そしてこのことがコードモデルの現実に対する説明力の向上に寄与するのである。

(17)　シニフィエとシニフィアンが結合している点において統合記号は記号と同義のように見えるが，Prieto によると統合記号はあくまで言語的なものであり，例えば進入禁止標識のような非言語的実体は統合記号に含まれない点で記号と異なることが強調されている。

(18)　受信者に焦点を当てる点において Prieto のモデルは第Ⅲ章で触れた推論モデルと共通する。しかし，推論モデルは発話の背後にある発信者の意図について受信者が仮説を立てて推論していくとみなす（西山［1999］54 ページ）。つまり推論モデルはコードの存在を仮定していない。これに対し Prieto のモデルは統合記号の作成と解読において許容する選択肢に幅があるコードの存在を含意している。そして受信者だけでなく発信者もモデルの中で重要な役割を演じている。ここに両モデルの相違がある。

3．コンテクストの回復

Prieto のモデルによれば，コードモデルとコンテクストは背反するものではない。むしろメッセージの解読に当たっては情況，すなわちコンテクストが介在することを指摘している。そうであればコードモデルを基盤とする頻出語分析を行った後，コンテクストを参照しつつ書き手の意図を解釈するという，二段構えの方法は妥当性を有するだろう。

コンテクストを分析に取り込むための具体的な方法としては，語と語の共起関係を観察，分析することを通じて，ある語がどのようなコンテクストの下で使用されることが多いかを確かめることが考えられる。すなわち前述したクラスター分析や共起ネットワーク分析を行うのである。これらの分析手法を用いれば，視覚的に洗練されたグラフを通じて，膨大な非構造化データの中から語のクラスターを抽出することが可能となる。クラスター分析におけるデンドログラムや共起ネットワーク分析におけるネットワーク図の描画はいずれも数理的処理を行う定量的なプロセスであり[(19)]，適切なソフトウェアが搭載されたコンピュータにデータを入力すれば自動的に作画される。しかし，コンピュータがグラフの解釈まで自動的に行ってくれる訳ではなく，研究者自身が解釈を加えなければならない。的確な解釈を行うためには，研究者は分析対象の文書内容に関する専門知識を具備している必要がある。いわば研究者が文書の書き手と「コード」を共有していなければならない。ただ，こうした分析では文書中における共起の全体的な傾向は把握できても，個々の出現事例におけるコンテクストとの繋がりまでは摑み切れない。

そこで，分析対象の文書に立ち戻り，定量的分析で抽出した語が文書の中で実際にどう用いられていたのかを直接確認することは効果的であろう。ただ，膨大な文書の中で個々の使用事例を手作業で検索することは容易でない。そこで有用なのが KWIC（keyword in contents）である。文書が然程多くなければ Word や Excel の検索機能を利用した分析もできなくはないが，KWIC 機能を

(19) もっとも，描画に当たってはクラスター化法や類似度尺度，中心性などの選択に研究者の判断が介入するため，完全に定量的なプロセスにはならない。

装備したソフトウェアを用いれば検索対象の語がその前後の文章とともに一覧表で表示され，実際の文脈中における使われ方の比較検討も容易となる。また，ソフトウェアによっては条件付きの検索機能を備えたものもあり[20]，研究者の関心に沿った柔軟な検索を可能としている。こうしたプロセスはコンピュータの助けを借りるものの，ほとんど質的な分析作業である。そしてその結果は，定量的分析の結果解釈に役立つだけでなく，定量的分析から得られた新たな発見をもとに再び質的なコーディング作業に立ち戻って追加分析を行う契機にもなり得る。

テキスト分析の定義でみたように，テキスト分析ではもともと，定量的分析だけでなく質的分析も重要な位置を占める。また，樋口［2020］が指摘するように，可能な限り厳格な形で定量的分析を行ったとしても，研究の様々な段階で質的な作業が必要となる。文書という質的なデータを量的に扱える形に変換する作業は，決して純粋に量的な作業ではない。出現頻度の計測対象とするキーワードの設定において研究者の主観的な分析視角を排除することはできないし，キーワードをいくつかのカテゴリーに分類していくコーディング作業は研究者の専門的想像力が発揮されるべき極めて重要かつ高度に質的な作業である。つまり，定量的分析と質的分析は断絶した排他的な関係にあるのではない。テキスト分析の研究実施プロセスにおいては，両者を循環的，相乗的に活かすことが不可欠である。

このように，定量的な頻出語分析を行った後，コンテクストを参照しながら追加分析するという流れは木に竹を接ぐ訳ではなく，理論的にも研究実践的にも妥当な分析手法の組合せ方だと言える。こうした二段構えの方法は，コンテクストから一旦切り離された語の意味を的確に捉えた分析を可能とするであろう。

---

(20)　例えば樋口［2020］が詳説するテキスト分析用ソフトウェアである KH Coder では，「抽出したい語の直前または直後に特定の語が出現していること」という条件を付けた抽出が可能である。

Ⅷ．結　　論

本稿では記号学における一つのモデルであるコードモデルに依拠しつつ，テキスト分析の中で広く用いられている頻出語分析が依って立つ理論的基盤を明らかにすることを試みた。その結果，頻出語分析は，単なる語の数え上げではなく，文書作成において書き手の思考が強く向けられている概念，すなわち文書の主題を明らかにする，理論的基盤を有する分析手法だと位置付けることができた。また，頻出語分析でしばしば用いられる TF-IDF についてはコードモデルだけでなく情報理論の視点も踏まえて検討した結果，語の出現頻度すなわち書き手の思考が強く向けられている概念の表出濃度と情報量を掛け合わせて文書中における語の重要度を表す，理論的基盤を有する指標であることが明らかとなった。頻出語分析は，テキスト分析を用いた研究では多くの学術領域において用いられてきた実績を有する。しかしながら，その理論的基盤まで踏み込んで考察した研究はほとんど見当たらないことから，本稿の考察はテキスト分析の理論的基盤の強化に貢献し得るであろう。

もっとも，本稿には多くの限界があることも事実である。まず，本稿では主として頻出語分析に焦点を当てて考察したが，テキスト分析では多様な分析手法が採られる。クラスター分析と共起ネットワーク分析については本稿で幾分述べたものの踏み込んだ考察は行っていない。さらに，近年の人工知能の発達に伴い，自己組織化マップが注目されることがある。自己組織化マップは Kohonen［1995］が提唱した教師なし学習モデルであり，クラスター分析や共起ネットワーク分析と同様に語のクラスターを把握するために有効な方法とされるが，本稿では触れていない。このように，テキスト分析で用いられる分析方法を網羅的に考察していない点は本稿の一つの限界である。

また，本稿が依拠したコードモデルには，記号学の分野では多くの批判もある。特に，コードモデルはコンテクストを捨象しているため現実に対する説明力が弱いという弱点がある。この点について，Prieto はコードモデルにコンテクストを取り込む形で解決しようとした。しかし Sperber and Wilson［1995］

訳本 17 ページが指摘するように，コードモデルを前提とするならば，読み手が解読する際に用いる暗黙の前提や推論規則は書き手と共有されており，かつそれ以外の暗黙の前提や推論規則は用いられないという条件を満たさなければ正しい解読は行えないが，特に後者を立証することは難しい。Sperber and Wilson［1995］はむしろ，コンテクストを参照する推論プロセスとコード解読プロセスをひとつのモデルの中で融合させるのではなく，コードモデルと推論モデルという 2 つのモデルの存在を認めるべきと主張する（訳本 32 ページ）。コードモデル自体の理論的基盤が揺らぐとすれば，頻出語分析の理論的基盤も再構築する必要性に迫られる可能性がある。この点については今後の研究課題としたい。

参 考 文 献

池上嘉彦『記号論への招待』岩波書店，1984 年。

井上祐輔「『イノベーションが普及する』とは，どういうことなのか－テキストマイニングの利用可能性－」（竹岡志朗・井上祐輔・高木修一・高柳直弥『イノベーションの普及過程の可視化　テキストマイニングを用いたクチコミ分析』日科技連，2016 年），36-54 ページ。

風間喜代三・上野善道・松村一登・町田健『言語学（第 2 版）』東京大学出版会，2004 年。

岸田和明『情報検索の理論と実務』勁草書房，1998 年。

喜田昌樹『新テキストマイニング入門』白桃書房，2018 年。

金明哲「コーパスとテキストマイニング」（石田基広・金明哲編著『コーパスとテキストマイニング』共立出版，2012 年），1 -14 ページ。

金明哲『テキストアナリティクス』共立出版，2018 年。

黒橋禎夫・柴田知秀『自然言語処理概論』サイエンス社，2016 年。

阪口祐介・樋口耕一「震災後の高校生を脱原発へと向かわせるもの－自由回答データの計量テキスト分析から－」（友枝敏雄編『リスク社会を生きる若者たち　高校生の意識調査から』大阪大学出版会，2015 年），186-203 ページ。

鈴木勉『ネットワーク分析（第 2 版）』共立出版，2017 年。

高木修一「テキストマイニングを用いた基礎的な分析とその限界」（竹岡志朗・井上祐輔・高木修一・高柳直弥『イノベーションの普及過程の可視化　テキストマイニングを用いたクチコミ分析』日科技連，2016 年），78-89 ページ。

土田知則・青柳悦子・伊藤直哉『現代文学理論　テクスト・読み・世界』新曜社，1996 年。

西山佑司「語用論の基礎概念」(田窪行則・西山佑司・三藤博・亀山恵・片桐恭弘『談話と文脈』岩波書店，1999 年)，1 -54 ページ。

樋口耕一『社会調査のための計量テキスト分析　内容分析の継承と発展を目指して(第 2 版)』ナカニシヤ出版，2020 年。

前田亮・西原陽子『情報アクセス技術入門　情報検索・多言語情報処理・テキストマイニング・情報可視化』森北出版，2017 年。

宮崎修一『グラフ理論入門　基本とアルゴリズム』森北出版，2015 年。

Barthes, Roland, *Mythologies*, Les Editions du Seuil, 1957.(篠沢秀夫訳『神話作用』現代思潮新社，1967 年)。

Berelson, Barnard, *Content Analysis in Communication Research*, Free Press, 1952.

Eco, Umberto, *Il Segno*, ISEDI, 1973.(谷口伊兵衛訳『記号論入門　記号概念の歴史と分析』而立書房，1997 年)。

Eco, Umberto, *A Theory of Semiotics*, Indiana University Press, 1976.(池上嘉彦訳『記号論Ⅰ・Ⅱ』岩波書店，1980 年)。

Feldman, Ronen, and James Sanger, *The Text Mining Handbook : Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2007.(辻井潤一監訳『テキストマイニングハンドブック』東京電機大学出版局，2010 年)。

Grice, Paul, "Meaning," *The Philosophical Review*, Vol. 66, No. 3, 1957, pp. 377-388.

Grice, Paul, *Studies in the Way of Words*, Harvard University Press, 1989.(清塚邦彦訳『論理と会話』勁草書房，1998 年)。

Hawkes, Terence, *Structuralism and Semiotics*, Methuen, 1977.(池上嘉彦・荻野蔵平・小倉敏博・鏡味治也・加藤泰・久慈洋子・栗田博之・佐伯泰樹・三宮郁子・高橋時男・村山和行・米山三明訳『構造主義と記号論』紀伊國屋書店，1979 年)。

Jegadeesh, Narasimhan, and Di Wu, "Word power : A new approach for content analysis," *Journal of Financial Economics*, Vol. 110, No. 3, 2013, pp. 712-729.

Kohonen, Teuvo, *Self-Organizing Maps*, Springer, 1995.(徳高平蔵・岸田悟・藤村喜久郎訳『自己組織化マップ』シュプリンガー・フェアラーク東京，1996 年)。

Krippendorff, Klaus, *Content Analysis : An Introduction to Its Methodology*, Sage Publications Inc., 1980.(三上俊治・椎野信雄・橋元良明訳『メッセージ分析の技法「内容分析」への招待』勁草書房，1989 年)。

Krippendorff, Klaus, *Content Analysis : An Introduction to Its Methodology, fourth edition*, Sage Publications Inc., 2019.

Kuckartz, Udo, *Qualitative Text Analysis*, Sage Publications Inc., 2014.(佐藤郁哉訳『質的テキスト分析法　基本原理・分析技法・ソフトウェア』新曜社，2018 年)。

Prieto, Luis J., *Messages et Signaux*, Presses Universitaires de France, 1972.(丸山圭三郎訳『記号学とは何か－メッセージと信号－』白水社，1998 年)。

Neuendorf, Kimberly A., *The Content Analysis Guidebook*, Sage Publications Inc., 2002.

Riffe, Daniel, Stephen Lacy, and Frederik Fico, *Analyzing Media Messages : Using Quantitative Content Analysis in Research*, Taylor & Francis, 2014.（日野愛郎監訳『内容分析の進め方－メディア・メッセージを読み解く』勁草書房，2018 年）。

Romesburg, Charles H., *Cluster Analysis for Researchers*, Robert E. Krieger Publishing Company, 1989.（西田英郎・佐藤嗣二共訳『実例クラスター分析』内田老鶴圃，1992 年）。

Saussure, Ferdinand de, *Cours de linguistique générale*, Payot, 1916.（町田健訳『新訳　ソシュール一般言語学講義』研究社，2016 年）。

Shannon, Claude E., and Warren Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, 1949.（植松友彦訳『通信の数学的理論』筑摩書房，2009 年）。

Sperber, Dan, and Deirdre Wilson, *Relevance : Communication and Cognition, second edition*, Blackwell Publishing, 1995.（内田聖二・中逵俊明・宋南先・田中圭子訳『関連性理論（第２版）－伝達と認知－』研究社，1999 年）。

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie, *The General Inquirer : A Computer Approach to Content Analysis*, MIT Press, 1966.

Weber, Robert P., *Basic Content Analysis, second edition*, Sage Publications Inc., 1990.