論文 / 著書情報
Article / Book Information

題目(和文)	オントロジを利用したテキスト計量分析フレームワークの構築
Title(English)	
著者(和文)	川島隆徳
Author(English)	Takanori Kawashima
出典(和文)	学位:博士（工学）, 学位授与機関:東京工業大学, 報告番号:甲第9889号, 授与年月日:2015年3月26日, 学位の種別:課程博士, 審査員:猪原　健弘,桑子　敏雄,赤間　啓之,山元　啓史,戦　暁梅
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第9889号, Conferred date:2015/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

オントロジを利用したテキスト計量分析フレームワークの構築

東京工業大学大学院社会理工学研究科
価値システム専攻
平成26年度博士論文

学籍番号　10D41029
川島　隆徳

指導教員　徃住彰文教授・猪原健弘教授

2014 年 1 月 26 日

論文の概要

本研究では，単語カテゴリの集合であるオントロジを利用するテキスト計量分析フレームワークを構築し，またフレームワークを実践するためツールを開発した．

テキスト計量分析は，自然言語処理や統計，ネットワーク解析等の手法を用いてテキストを科学的に分析する方法だが，その方法論自体についての研究は充分になされてこなかった．そこで，本研究では既存の研究を踏まえ，単語を単位とする計量分析においては，オントロジが有効であると仮定した．ここで言うオントロジの定義は，同義語・類似語の集合（カテゴリ）に対してその内容を示すラベルが付与された機械可読なデータである．3つのコーパスを対象として3種類のオントロジ導入方法を検討し，それらを用いた分析を行うことでオントロジの有効性を検討した．

1つめのケーススタディであるデジタルゲームの批評コーパスでは，単語の共起情報を用いて機械的にオントロジを生成するアルゴリズムを開発した．2段階に分けたクラスタリングを行うことで，より精度の高いオントロジを生成することが可能となった．また，このオントロジに含まれるカテゴリ単位での分析を行うことで，デジタルゲームの批評においてビジュアル的な要素等に関する記述が減少し，ゲームの遊び方等に関する記述が増加していることを明らかにした．

2つめのケーススタディである河川文化に関する講演会記録のコーパスは，河川を中心とした多彩な内容を含むため共起情報を用いたオントロジ自動抽出は機能しないことが判った．そこで，河川の専門家の手を借りた半手動のオントロジ構築を行い，専門家の知識をもってコーパスを分析する手法を開発した．構築されたオントロジのカテゴリの単位で計量を行い，統計的検定を行うことで，特定の講演において有意に出現するカテゴリ（話題）を抽出した．さらに，そのカテゴリ同士の組（共起）を数え，カテゴリをノード，共起をエッジとしたネットワークを構築することで，曖昧で多様な概念である「河川文化」の全体像を計量的に示すことに成功した．またこのネットワークを分析することで，河川に関するローカルな問題とグローバルな問題は，「川」や「山」などの土地と，「建築」などの工学を介してつながるということが示された．

3つめのケーススタディでは，文学，映画，演劇，デジタルゲームの4ジャンルに関する批評コーパスを扱った．このコーパスでは，ジャンル間の感性の違いを明らかにするため，他の2つのケーススタディでは扱わなかった形容語を対象として分析を行った．形容語についても共起を用いたオントロジ自動抽出は機能せず，また形容語自体は専門的な知識でカテゴライズできる語でも無いため，汎用的なシソーラスである「分類語彙表」の分類をオントロジとして利用した．ジャンル毎にカテゴリの出現を計量し，統計的に比較することで，それぞれのジャンルにおける形容語の特徴的な出現を明らかにした．具体的には，文芸は多用される形容語というものが無く，映画は直感的・感覚的な語の利用が多く，演劇は比喩の感覚を利用した評価が多いという特徴があった．そしてデジタルゲームは他の

どのジャンルとも大きく異なり，ゲームを攻略するという意識の下での有利不利の評価や，画面上に現れる詳細な内容についての評価が多いことが示された．

以上のケーススタディから，テキストの特性と目的によって，適切なオントロジの導入方法が異なることが明らかとなった．また，単語単位ではなく，オントロジのカテゴリを単位とした分析を行うことで，対象テキストの大局的な特徴を抽出可能であることが明らかとなり，オントロジの有効性が示された．

そこで，オントロジの導入を前提としたテキスト計量分析のフレームワーク OSQTA(Ontology-based Semantic Quantified Text Analysis)を構築した．このフレームワークは，目的の設定，データの収集から実際の分析までをカバーする．特に，利用するオントロジの選択と分析手法の選択に重点を置き，ケーススタディから得られた知見を元にどのような選択を行うべきかを記述した．その他にも，これまでの研究から得られた経験的な規則等を盛り込んだ．OSQTA はテキスト計量分析の全てを網羅したフレームワーク-では無いが，本研究の 3 つのケーススタディだけで無く，複数の既存研究がこのフレームワークの範疇として説明可能であり，汎用的なフレームワークと言える．

さらに，フレームワークを実践するための Text Seer というソフトウェアを開発した．Text Seer を利用することで，OSQTA の基礎的な処理を自動的に行うことが可能となった．

OSQTA 及び Text Seer を利用することで，今後のテキスト計量分析研究を容易に行うことが可能となり，また分析プロセスを標準化することで精度を向上させることが可能となった点が本研究の最大の貢献である．

目次

1. はじめに 11
   1.1. 研究の背景 11
   1.2. 研究の目的と意義 12
   1.3. 研究の方法と論文の構成 13
2. 先行研究 15
   2.1. 本章の目的 15
   2.2. テキスト計量分析の背景 15
      2.2.1. 自然言語処理 15
      2.2.2. プロトコル分析 16
   2.3. テキストを対象とした関連分野の研究 17
      2.3.1. テキストマイニング 17
      2.3.2. 計量言語学 18
      2.3.3. 内容分析 19
      2.3.4. デジタル・ヒューマニティーズ 19
   2.4. テキスト計量分析 20
      2.4.1. テキスト計量分析の定義 20
      2.4.2. テキスト計量分析の先行例 21
3. テキスト計量分析の基礎 24
   3.1. 本章の目的 24
   3.2. テキスト計量分析の流れ 24
   3.3. テキストの収集 24
      3.3.1. テキストの収集 24
      3.3.2. 権利 25
      3.3.3. OCR 25
      3.3.4. 文字コード 25
   3.4. テキストの整形 26
      3.4.1. テキストのサニタイズ 26
      3.4.2. テキストの構造化 26
      3.4.3. メタデータの付与 26
   3.5. 計量単位への分解 26
      3.5.1. 文字 27
      3.5.2. n-グラム 27
      3.5.3. 形態素 27
      3.5.4. オントロジ 28

3.5.5. 命題 29
3.6. 頻度分析 29
3.6.1. 単語頻度 29
3.6.2. 文書頻度 29
3.6.3. TF・IDF 29
3.6.4. 異なり語彙数 30
3.6.5. その他の計量値 30
3.7. パターンの抽出と分析 30
3.7.1. 共起 30
3.7.2. 係り受け 31
3.7.3. 単語のネットワーク 31
3.7.4. ネットワーク分析 32
3.7.5. クラスタリング 35
3.7.6. 統計解析 35
3.8. 特徴の定性的分析 35
3.8.1. KWIC 36
3.9. 分析のためのモデルと基本方針 36
3.10. フレームワークに必要な要素 39
3.11. 以降の分析で使用したソフトウェア 40
4. オントロジの自動生成による概念カテゴリ計量－ゲーム批評の批評対象要素の抽出－ 41
4.1. 本章の目的 41
4.2. 分析の背景と目的 41
4.3. 対象データと手法の選択 43
4.4. オントロジの自動生成 48
4.4.1. 目的 48
4.4.2. 方法 48
4.4.3. 結果 49
4.4.4. 考察 53
4.5. カテゴリの時系列分析 57
4.5.1. 目的 57
4.5.2. 方法 58
4.5.3. 結果 58
4.5.4. 考察 60
4.6. ユーザレビューとの比較 61
4.6.1. 目的 61

4.6.2. データ 61
4.6.3. 方法 61
4.6.4. 結果 61
4.6.5. 考察 65
4.7. 分析の結論 65
5. オントロジの手動構築と概念構造抽出－河川文化における大域的概念構造の抽出－ 67
5.1. 本章の目的 67
5.2. 分析の背景と目的 67
5.3. 対象データと手法の選択 69
5.4. オントロジの手動構築による河川文化要素の抽出 73
5.4.1. 目的 73
5.4.2. 方法 74
5.4.3. 結果 75
5.4.4. 考察 75
5.5. 要素の関連性の分析（全体） 76
5.5.1. 目的 76
5.5.2. 方法 76
5.5.3. 結果 78
5.5.4. 考察 79
5.6. 要素の関連性の分析（講演カテゴリ） 81
5.6.1. 目的 81
5.6.2. 方法 81
5.6.3. 結果と考察 81
5.7. 実践への貢献可能性 85
5.8. 分析の結論 86
6. 既存オントロジの利用による評価語計量比較－4 ジャンルの批評における感性の違いを探る－ 88
6.1. 本章の目的 88
6.2. 分析の背景と目的 88
6.3. 対象データと手法の選択 88
6.4. 評価語の抽出と分析 92
6.4.1. 方法 92
6.4.2. 全テキストでの結果 93
6.4.3. ゲームを除いた結果 98
6.5. 考察 102
6.5.1. 文学の特徴 103

6.5.2. 演劇の特徴 103
6.5.3. 映画の特徴 104
6.5.4. ゲームの特徴 104
6.6. 分析の結論 104
7. 分析手法に関する比較考察 106
7.1. 本章の目的 106
7.2. オントロジに関する手法 106
7.2.1. オントロジを利用しない 106
7.2.2. 自動生成 107
7.2.3. 手動構築 109
7.2.4. 既存オントロジの利用 109
7.2.5. 分析に適切なオントロジ 110
7.3. 分析に関する手法 110
7.3.1. カテゴリ単位の計量比較 110
7.3.2. テキストを単位としたカテゴリネットワーク 111
7.4. 結論 113
8. オントロジを利用したテキスト計量分析フレームワーク 114
8.1. 本章の目的 114
8.2. フレームワーク 114
8.2.1. フレームワークの位置付け 114
8.2.2. 分析フレームワークの全体像 114
8.2.3. 目的の設定 115
8.2.4. テキストの収集 117
8.2.5. 前処理 118
8.2.6. 基本的特徴の把握 119
8.2.7. 分析方法の検討 123
8.2.8. 分析 125
8.2.9. 考察 130
8.3. フレームワークの意義 130
8.4. フレームワークを利用するためのツール 131
8.4.1. Text Seer とは 131
8.4.2. Text Seer の機能 132
8.4.3. 既存のアプリケーションとの比較 138
8.4.4. 実装の詳細 139
8.4.5. インターフェイスの実装 141
9. おわりに 144

9.1. 本研究の成果 144
9.1.1. オントロジの自動生成による概念カテゴリ計量 144
9.1.2. オントロジの手動構築と概念構造抽出 144
9.1.3. 既存オントロジの利用による形容語計量比較 144
9.1.4. フレームワークとツール 145
9.2. 今後の展望 145
9.2.1. フレームワークの整備と発展 145
9.2.2. フレームワークの妥当性を示す検証 146
9.2.3. ツールの拡充 146
謝辞 147
参考文献 148

図表目次

図 1　論文の構造 14
図 2　シンプルなオントロジの構造 28
図 3　テキスト執筆の認知モデル 37
図 4　単語を踏まえたテキスト執筆の認知モデル 38
図 5　出現頻度上位 30%の語による共起頻度 30 回以上の共起ネットワーク 46
図 6　ゲーム批評に対する分析フローの概要図 47
図 7　カテゴリの上位構造 50
図 8　各カテゴリの出現頻度（降順） 52
図 9　各カテゴリに含まれる語の平均使用回数 53
図 10　4 つのカテゴリ上位構造 56
図 11　出現量の相対変化（減少傾向にあるカテゴリ） 59
図 12　出現量の相対変化（増加傾向にあるカテゴリ） 59
図 13　ユーザレビューから生成されたカテゴリの上位構造 63
図 14　出現頻度上位 30%の語による共起頻度 77 回以上の共起ネットワーク 71
図 15　河川文化を対象とした名詞の階層化クラスタリングの結果の一部 72
図 16　出現頻度上位 30%の語による共起頻度 77 回以上の共起ネットワークから【川】【水】【日本】を除外したネットワーク 73
図 17　河川文化に対する分析フローの概要図 74
図 18　カテゴリネットワーク図 79
図 19　土木・空間のネットワーク図 82
図 20　環境・生態のネットワーク図 83
図 21　社会・暮らしのネットワーク図 84
図 22　文化・歴史のネットワーク図 85
図 23　4 ジャンル批評に対する分析フローの概要図 92
図 24　OSQTA の全体像 115
図 25　河川文化テキストにおける単語頻度と累計頻度の関係 121
図 26　ゲーム批評における頻度上位 100 名詞の共起ネットワーク（共起頻度上位 100） 122
図 27　「プレイヤー」を中心としたエゴセントリックネットワーク 123
図 28　Text Seer のアーキテクチャ 140
図 29　Text Seer のデータ構造 140
図 30　Text Seer のメインウィンドウ 142

表 1　関連分野との比較 22
表 2.テキスト計量分析の流れ 24

表 3　共起と係り受けの頻度 31
表 4　ネットワーク描画アルゴリズム 32
表 5　ネットワーク指標一覧 33
表 6　中心性の指標 34
表 7　解析のレベルによる利点と欠点 39
表 8　各ケーススタディで採用されたオントロジと分析方法 40
表 9　対象ゲームとその分類 44
表 10　出現頻度上位 30 位までの名詞の出現頻度と出現テキスト数 45
表 11　カテゴリ一覧 51
表 12　カテゴリと係り受けする形容詞 52
表 13　年代別グループ 58
表 14　時代毎のカテゴリ出現頻度の χ^2 検定結果（p<.01） 60
表 15　ユーザレビューから生成されたカテゴリの一覧 62
表 16　ゲーム批評カテゴリとユーザレビューカテゴリ間の共通語数 64
表 17　河川文化における講演のカテゴリと講演数 70
表 18　出現頻度上位 30 位までの名詞の出現頻度と出現テキスト数 70
表 19　カテゴリ一覧 77
表 20　対象雑誌 89
表 21　各ジャンルにおける頻度上位 10 位までの形容語とその出現頻度 90
表 22　全ジャンルにおける頻度上位 20 の名詞の頻度の χ^2 検定結果（p<.05） 90
表 23　各ジャンルにおける頻度上位 10 位までの名詞とその出現頻度 91
表 24　ジャンル毎の評価語数 93
表 25　中分類毎の評価語数 94
表 26　ゲームを含んだ中項目の χ^2 検定結果（p<.05） 95
表 27　ゲームを含まない中分類の χ^2 検定結果（p<.05） 99
表 28　各ジャンルの特徴 103
表 29　オントロジに関する手法の適用結果 106
表 30　演劇の上位 104 名詞を対象としたクラスタリング 108
表 31　目的と OSQTA の有効性 116
表 32　テキスト収集における留意 118
表 33　テキストにおける基本統計量 120
表 34　オントロジ方針 124
表 35　分析方法の選択 125
表 36　ネットワークの例 129
表 37　OSQTA と Text Seer の対応関係 132
表 38　記述統計量 137

1. はじめに

1.1. 研究の背景

人は芸術をどのように鑑賞しているのだろうか．例えば映画を例に取ってみれば，その鑑賞には，画面に写されている内容物，登場人物の発話，そして物語のプロットの理解，さらには映像や音響から「迫力」を感じること，そういった複数の認知や情動が含まれていることは間違いないだろう．しかし個人の主観では，実際に映画を鑑賞しているときにこういった複数のプロセスを意識することはなく，総合的な「体験」のみがある．

このような複合的な要素を含む「体験」を分析する認知科学の方法の 1 つとして，プロトコル分析がある．この方法では，例えば映画の鑑賞中に感じたこと，考えたことをリアルタイムで発話（プロトコル）してもらい，その内容を分析する．体験中の発話には，無意識的に発話者の内的認知過程が含まれるという前提に立ったプロトコル分析は，認知科学だけでなく製品のユーザビリティテストなどにも広く利用されている．しかし，実際の分析にはデータを取るための心理実験，発話の文字起こしと，発話内容の人手による分類を要する．それゆえに，実験を大規模に行うことは難しく，分析結果の普遍性を説明するためには大きな労力がかかるという欠点がある．

プロトコル法と同様の視点を持ちながらも，その欠点である規模の問題を解決できる方法の 1 つとして，体験者によって書かれたテキストに着目するテキスト計量分析がある．この方法では，例えば映画の批評文を収集し計量的に分析することで，その背後にある認知的要素に迫ろうとする．映画の批評文は映画鑑賞体験をリアルタイムに描写したものではなく，文章の作成という別の高度な脳の働きを要するものではあるが，その体験について語ったものであることには違いない．また，直接的な体験の心理状態を分析することは難しくなるが，体験に当たっての前提となる知識構造や，価値判断の基準といったようなものが含まれているはずであり，より複雑な内容が分析可能となる場合もある．さらに，多数の書き手のテキストを分析することでより普遍的な結果を得ることや，複数のテキスト群の間で比較を行うことなどが容易になるという利点も持つ．一方で，量的な分析のため，得られる理解が浅いという欠点もある．

テキスト計量分析は，多量のテキストを扱う必要があるため，自然言語処理技術を利用した処理の機械化によって実質的に可能になったと言える．今日では日本語形態素解析などの自然言語処理の基礎技術は成熟し，情報工学分野では Web 上の大量データを対象としたテキストマイニングなどへと応用されている．テキスト計量分析のみならず，社会学において長い歴史を持つ内容分析の手法や，計量言語学においても，こういった技術を用いた機械的な処理が取り入れられてきている．さらには，元来計量的な分析手法を持たなかったその他の人文科学の諸ジャンルにおいても，資料デジタル化の推進等を背景として類似の取り組みが始まっている．このような動きは，テキストの分析に留まらず，人文科学における情報技術の活用を目的としたデジタル・ヒューマニティーズと呼ばれる旗印の下，国内でも活発になってきた．

しかしながら，このような技術が一般化することで，弊害とも言える問題が出てきた．テキストの機械的な分析が，その背景の知識もなく手軽に行えるため，分析者がその結果を安易に，また拡大的に解釈してしまうという問題である．筆者は専門家でなくとも使えるテキスト分析のツールを開発し提供する中で，そのユーザである研究者が深く考えることなく，ツールが出力するそれらしい図を見て研究を進めてしまうシーンを目撃してきた．そういった浅い分析で結論づけられた内容では，従来型の研究手法から見た場合，その手法もろとも低い評価を与えられてしまう可能性もある．このような問題が起こる背景としては，分析のルーチンが方法として確立されていない，研究者またはそれを評価する側も分析の特性・限界について把握していない，等があると考えられる．認知科学におけるテキスト計量分析においても，その方法は様々な対象に適用されてきたが，改めて方法自身の位置づけや性質について明確化することは行われていない．

テキスト分析の技術は未だ一般的とは言えず，またその精度や，機械的に処理できる内容に限界もあるが，大規模なデータを扱うためには機械の力を借りることは避けて通れない道である．そういった技術が，正しく扱われないせいで軽んじられる，ないしは普及・発展しないことは，認知科学のみならず他のテキストを扱う学問にとっても損失であると筆者は考える．

1.2. 研究の目的と意義

そこで，本研究では，多分野にわたる複数のケーススタディから得られた知見に基づき，テキスト計量分析を適切に扱うためのフレームワークを構築することを目的とする．さらに，このフレームワークを実際に利用するためのツールとして，専用のソフトウェアを開発する．ソフトウェアを用意することで，研究者はフレームワークを容易に実践することが可能となる．

フレームワーク構築及びツール開発の意義は以下の 2 点である．

(1) フレームワークに則って分析を行うことで，定量的な結果，定性的な結果，考察を分離して，機械的な分析から飛躍しない考察を行うことを可能とし，テキスト計量分析研究の精度を向上させる．
(2) フレームワークとツールを利用することで，分析の手法に明るくない研究者であっても，大規模データを利用した研究を行うことができる．

また，フレームワークの意義の前提となる，テキスト計量分析の意義は以下の 3 点である．

(1) 大規模なデータを扱うことで，経年変化や，テキスト群毎の違いなど，マクロな観点での分析が可能となる
(2) 計量部分については再現性のある研究となるため，科学的な議論が可能となる．
(3) 人間の脳では気がつきにくいような特徴や，無意識のうちに自明としていた内容を発見することが可能となる．

1.3. 研究の方法と論文の構成

本研究では，既存の手法，研究を整理し，既存の方法論では不十分な内容について複数のケーススタディを通じて検討することで，帰納的にフレームワークを構築する方法を取った．

論文の構成は以下の通りである．まず，本章から続く 2 章では，テキスト計量分析の背景としての自然言語処理と認知科学について述べる．また，テキスト計量分析の隣接分野であるテキストマイニング，計量言語学，内容分析と近年の動向であるデジタル・ヒューマニティーズについて述べ，その上でテキスト計量分析の位置付けと先行事例について説明する．

3 章では，テキスト計量分析の基礎的な方法・技術について，既存の研究を踏まえながら整理する．また，分析の前提とする認知的モデルについて仮定を置き，その上で，実用的なフレームワークを構成するためにはオントロジの導入とその利用方法の検討が必要なことについて述べる．

4 章から 6 章にかけては，「ゲーム批評における評価対象の分析」，「河川文化に関する概念構造の分析」，「文学，演劇，映画，ゲームの批評における評価の違いの分析」という 3 つのケーススタディについて述べる．これにより，対象とするテキストの性質と目的によって利用するべきオントロジと分析方法が異なることを示す．なお，4 章から 6 章にかけてはそれぞれが一本の研究論文となっており，目的とする手法の検討以外にも，各分野における分析結果が成果として得られている．しかしながら，そういった個々の成果については本研究の最終的な目的ではないため，章を越えて論ずることはしない．

7 章では，4 章から 6 章で取った分析方法についてまとめ，比較考察する．続く 8 章では，ケーススタディに基づいて構築したテキスト計量分析のフレームワークと，それを利用するために開発したツール「Text Seer」について述べる．

最後に 9 章で本研究の成果をまとめ，課題と今後の展望について述べる．以上の論文の構成を図 1 に示す．

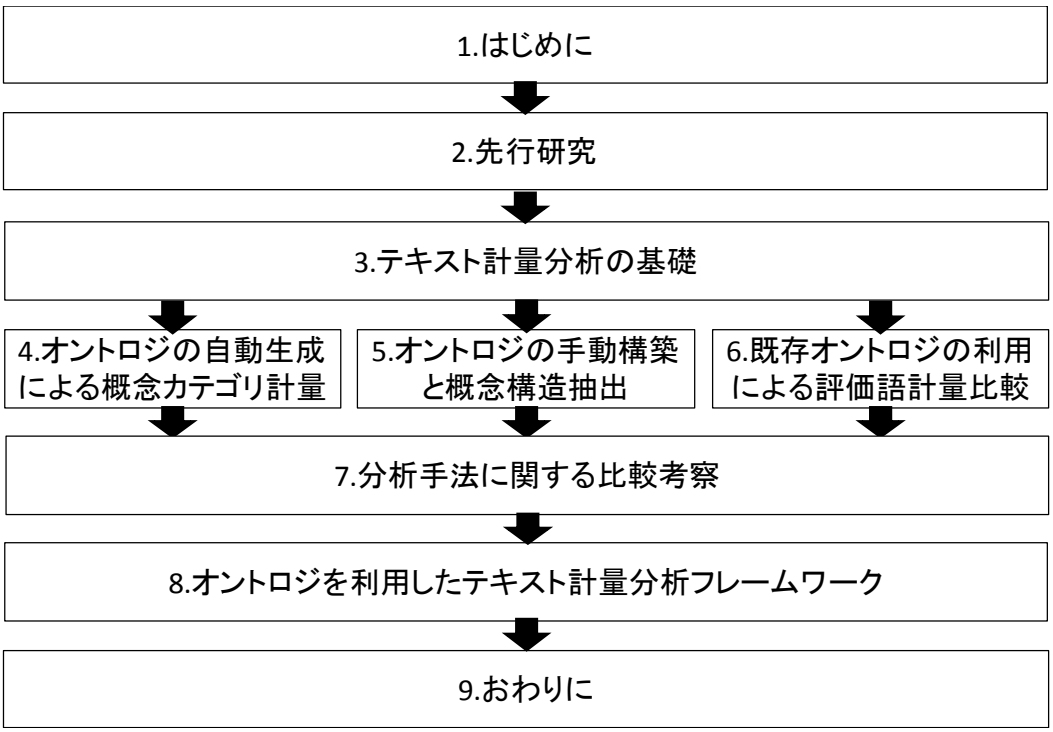
1.はじめに
2.先行研究
3.テキスト計量分析の基礎
4.オントロジの自動生成による概念カテゴリ計量
5.オントロジの手動構築と概念構造抽出
6.既存オントロジの利用による評価語計量比較
7.分析手法に関する比較考察
8.オントロジを利用したテキスト計量分析フレームワーク
9.おわりに

図 1　論文の構造

2. 先行研究

2.1. 本章の目的

本研究の対象とするテキスト計量分析の手法は，1 章で述べた通り認知科学におけるプロトコル分析と自然言語処理に端を発する手法である．そして，類似の手法が言語学や社会学にもあり，テキストを分析する手法は他の分野にも広がってきている．

本章では，テキスト計量分析が依拠する先行研究と，テキストを対象とした類縁の研究，そしてテキスト計量分析の定義と先行例についてまとめる．

2.2. テキスト計量分析の背景

2.2.1. 自然言語処理

自然言語とは，プログラミング言語などの人工的な言語に対し，人間が日常使う言語を意味する．天野によれば[1]，自然言語処理の分野は言語学の分類に従って，形態論，統語論，意味論，運用論の 4 つがある．

形態論では，形態素解析など，文章をその最小単位となる形態素に分割するという研究が行われている．この分野では MeCab というソフトウェアが日本語の形態素分割を約 96%の精度で行うことが可能であり[2]，実用面では十分なレベルにあるといえる．他にも Chasen[3]，JUMAN[4]等が存在する．

統語論は，文が成立する単語間の条件（文法）を論じる分野である．文法を表現する手法はいくつも存在するが，その中でもコンピュータ上で扱い易い構造として句構造文法と格文法が知られている．句構造文法では，文章を名詞句や動詞句など，複数の句で構成されたものとして扱い，最終的にそれぞれの句は単語で構成されることとなる．文章をこのような句構造に分解する方法の 1 つに係り受け解析があり，日本語では CaboCha[5]が 90%程度の精度で解析可能である．なお，CaboCha は句どうしの関係を明らかにするだけで，どの語が主語であるか，どの文節が目的格であるかなどは分からない．一方格文法とは，Fillmore によって提唱された格フレームを元に文を解析する手法である[6]. 格フレームは，例えば「動詞がどのような使われ方をするか」といったような情報を持ったパターンであり，各文章で，動詞の目的語などをパターンに基づいて決定する．日本語では，川原による Web から抽出した 4 万用言の格フレームが存在する[7].

意味論では，文章の意味を解析することを目標とする．この分野は人工知能，および認知科学の分野と深い関わりがあり，基礎となる研究には Quillian の意味ネットワーク[8]，Minsky のフレーム理論[9]，Schank のスクリプト[10]など，知識表現の方法に関するものがある．また，格フレームも，意味解析を目標とした構造でもある．文章の意味を捉えるためには，文章を構成する単語からその意味を引き出す必要がある．人間においてはこれを自分が持っている言語知識を利用して行っているわけだが，コンピュータが意味を理解するためにもそのような知識が必要である．前述した様々な知識表現の方法は，このような言語知識をいかに表現するべきか，という問題へ取り組むための手法の 1 つであるとい

える．現在，このような研究はオントロジ工学と呼ばれる人工知能の一分野で取り組まれており，単語によって表される概念どうしの関係をどのように記述するかに関して議論が行われている．一方で，そのような概念の辞書も実際に作られており，海外のものではCyc[11]やWordNet[12]が知られている．日本では，2009年にNICTが日本語WordNet[13]を公開した．しかし，村田によれば，自然言語処理学会での意味論に関する研究発表の量は減少傾向にあり[14]，その困難さゆえになかなか進展をみない分野であると考えられる．

運用論の分野では，コーパスを利用して，そこから言語の使われ方のデータを抽出するという，計量言語学に近い研究が行われている．しかし，計量言語学が目的とする言語学的な性質よりも，形態論，統語論，意味論で活用するためのデータを抽出することを主たる目的としている．この分野で利用されるコーパスは，計量言語学で用いられるコーパスよりも詳細なタグ付けが行われており，構文情報などを付加してある．Wall Street Journalの記事を収集したPenn Treebankコーパス[15]や，音やイントネーションなども付加した日本語話し言葉コーパス[16]などが知られている．

これら4つの分野は，自然言語処理の基礎というべき分野である．応用分野としては，仮名漢字変換，機械翻訳，情報検索，情報抽出，質問応答などの分野があるが，この中で特に本研究と関連が深いのは，情報検索の分野である．

情報検索では，テキスト情報にどのような索引をつけ，それをどのようにして効率よく検索するかという手法が研究されている．Deerwesterによる LSA（Latent Semantic Analysis）は，語と語の共起頻度をベクトルとし，テキスト内の全単語に関するベクトルを並べた行列を作る手法である．この語の共起行列が，そのテキストの意味を潜在的に表しているというのが，LSAの主張である[17]．LSAはテキストの意味を解釈することよりも，語の共起行列という意味を含んだ指標によって，分類や検索が行えるという点にその重要性を置いている．

2.2.2. プロトコル分析

プロトコル分析は，人間の認知，思考の流れを知るための実験・分析手法である．被験者は，何らかの課題を与えられ，その課題を行いながら頭に浮かんだことのすべてを発話することを求められる．そして，発話，沈黙時間，その他の動作などが書き起こされ，分析対象となる[18]．

多様なデータが得られるため，様々な観点による分析が可能だが，内容的な分析は，発話や発話の中に含まれる命題を単位として，それを分類することで行われる．例えば認知言語学者のChafe[19]は，鑑賞者に短編映画の内容を発話させ，その発話中の命題を分析することで，鑑賞者の背景がその映画の解釈・理解にどのような影響を与えているかを分析した．またConstantinescuらの研究[20]では，ドライブ・シミュレーションと実際のドライブ中の発話を分析し，命題を単位として徃住らが構築した51の感情カテゴリ[21]に分類した．結果として，審美的（景色が美しい，等）な感情はドライブコースが難しくない箇所で発生しやすいこと，背景の変化があった箇所で表出しやすいこと，またそれらが実際

のドライブでも同様の結果となることなどを明らかにし，その認知モデルを構築した．

さらに富岡らは，映画レビューテキストに対して同様の手法を適用し，映画鑑賞における感動体験に着目して鑑賞者の反応を分析した[22]．この分析では発話の代わりに書かれたテキストを対象としたが，命題を単位としてカテゴリに分類し，その特徴を分析するという手法はプロトコル法と共通している．また，齋藤らはアフィリエイト広告の掲載された商品紹介のブログ記事と掲載されていないブログ記事を対象として同様の分析を行い，広告のある記事のほうが対象商品に話題が始終するという傾向を発見した[23]．

上記 2 件のようなテキストの分析は，認知的な内容を対象とし，命題を単位とする点をプロトコル法の分析方法から受け継いだと言えるが，実施内容自体は後述する社会学における内容分析と近いものとなった．また，テキストを対象とすることで，実験が必要不要となり，大量のデータを分析できるようになった．例えば Constantinescu らの研究では，シミュレーションの発話は 13 人，実際のドライブ中の発話は 2 人から取得しているが，齋藤らの研究では，6 商品×12 記事で 72 のテキストを取得できている．

しかし，命題を人手で分類する手法にはコストがかかり，いくら大量のテキストが得られたとしても処理に限界がある．この処理限界に対処する方法の 1 つとして，自然言語処理を活用したテキスト計量分析がある．

2.3. テキストを対象とした関連分野の研究

テキスト計量分析が自然言語処理の普及を受けて具体化してきたのと同様に，情報工学の分野ではデータマイニングが，社会学の分野では内容分析が，言語学の分野では計量言語学やコーパス言語学が自然言語処理の力を利用した分析を行うようになってきた．

2.3.1. テキストマイニング

テキストマイニングとは，テキストデータを対象としたデータマイニングである．初期の研究には，データマイニングの手法をフィンランドの法律文書に適用し，語の共起分析を行った Ahonen らによる研究がある[24]．

テキストマイニングの研究は，2001 年頃にインターネットが普及し始めると，さらに盛んになった．Web を通じて様々なテキストを，書籍や企業組織以外から得られるようになったためであると考えられる．そして，掲示板や Weblog，ショッピングサイト等に商品などの感想が多く書き込まれるようになると，奥村ら[25]のように，そこからユーザの意見を抽出する評判分析という分野も登場するようになった．「評判分析」は，ブログなどの Web 上のデータを対象として，商品などの評判を知る研究分野であり，商業的な利用を目指した研究が多数行われている．乾によれば，この分野における解くべき問題は“ある対象の評価を記述しているテキスト断片に対して，その評価が，肯定的な評価（たとえば「良い」）であるか，あるいは，否定的な評価（例えば「悪い」）であるかを推定する．”として定義されるという[26]．ここで注目しなければならないのは，この分野の研究は「良い」と「悪い」という評価極性を見極める点に注力し，その理由や，評価の対象についてはあまり注

意を払っていない点である．赤木らでは評価極性ではなく，各商品の評価の軸となる属性を自動的に抽出しているが[27]，抽出した属性そのものを調べるわけではなく，それを元に類似したページを紹介することでユーザが商品について調べる際の補助とすることを最終的な目的としている．

2.3.2. 計量言語学

テキストを計量的に調査する研究は長い歴史を持つ．最初の著名な研究である Mendenhall の研究では，「シェイクスピアの戯曲を実はフランシス・ベーコンが書いたのではないか」という説を検証するため，両名の著作を構成する単語の文字数の分布を調べ，結果この説を否定した[28]．このような文章の著者推定，文章の特徴抽出は他にも数多く行われてきており，単語の長さの分布，文の長さの分布[29]，品詞の分布[30]，読点の打ち方[31]など，様々な特徴量に注目した研究が行われてきた．

日本語においては，安本が『源氏物語』の様々な統計量を利用し，最後の 10 巻（いわゆる『宇治十帖』）の執筆者が紫式部である可能性が少ないという結果を示している[32]．

これら一連の研究は，計量文体学という分野で，計量的に言語を調査する計量言語学の一分野といえる．計量言語学の中には，他にも言語使用の実態などを統計的手法で解明することを目的とした研究もある．その代表的な分野が，コーパス言語学と呼ばれる分野である．

コーパスとは，“言語分析のための文字言語，あるいは音声言語の資料の集合体”を意味し[33]，有名なものにブラウン・コーパス[34]がある．これは 1961 年にアメリカのブラウン大学が編纂したコーパスで，アメリカで出版された 15 ジャンル，2000 語のテキストを 500，計 100 万語集めたものである．これが，最初のコーパスといわれている．日本では，国立国語研究所が作成した，Web や書籍からテキストを収集した 1 億語の現代日本語書き言葉均衡コーパス[35]などが知られている．

このようなコーパスを調査することで，最もよく使われる単語や，単語の出現頻度などに関する知見が明らかにされた．例としては，ジップの法則[36]が挙げられる．これは，単語の出現頻度と，出現順位に，反比例の関係があるという法則である．また，KWIC（Keyword in Context）という，単語の前後に出現する語に注目する手法や，単語のコロケーションを調査する手法も，コーパス言語学においてよく利用される手法であり，テキスト計量分析に影響を与えている．

以上のように，計量言語学は，テキスト内にどのような情報が含まれており，それらをどのようにして計量的に扱うかという手法を発展させてきた．しかしながら，計量言語学の分野にはテキストの意味に踏み込んだ研究が少ない．その理由として，「テキストの内容」ではなく「言語的な特徴」の分析を目的としていることが挙げられる．また，計量言語学は統計の学問であるため科学的な厳密さを追求し，結果として意味のような曖昧な部分ではなく，文体の特徴などの明白な部分に研究の焦点を当てているとも考えられる．

2.3.3. 内容分析

Krippendorff によれば，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かつ妥当な推論を行うための 1 つの調査技法である”[37]という．その歴史は古く，例えば 1893 年の Speed らの新聞の内容に関する分析にまでさかのぼることができる[38]. 彼らは, 1881 年から 1893 年までのニューヨークの主要紙を分析し，ゴシップが多くを占めることを示した.

内容分析は歴史的経緯から社会学を中心として用いられ，質問紙調査における自由回答の分析等に用いられているが，そういった回答を機械的に分析するという研究も出てきている[39]. そのような動きの中で, 秋庭らは, 計量テキスト分析という方法を提唱している．これは，“インタビューデータなどの質的データ（文字データ）をコーディングによって数値化し，計量的分析手法を適用して，データを整理，分析，理解する方法”[40]であるという．樋口はこれを拡張し，“計量テキスト分析とは，計量的分析手法を用いてテキスト型データを整理または分析し，内容分析を行う方法である．計量テキスト分析の実践においてはコンピュータの適切な利用が望ましい”と定義した上で，この計量テキスト分析を支援するためのソフトウェアである KHCoder[41]を開発した．計量テキスト分析の特徴は，秋庭らの定義にあるとおり，コーディングによって作られたコードの単位で対象のテキストを数値化することにあると言える．コーディングとは，例えば「A ないしは B という単語が出てきたら，その文に α というコードを与える」といった，分析者が分析の目的に沿って作成する分類ルールを作ることである．コーディングによって，分析者の観点と知識を分析に導入し，分析の精度を向上させるのである．もちろん，このコーディングを分析者が恣意的に行って良いわけではなく，樋口の提唱する計量テキスト分析では，最初に多変量解析を用いることで分析者の理論や問題意識に影響を受けない形でデータを理解し，その後コーディングルールを作成することで理論仮説の検証や問題意識の追求を行う，とある．また，コーディングルールを公開する事で，妥当性を担保する.

計量テキスト分析は，前段として対象テキストの理解から入るものの，それ自体が目的ではない．その出自から，あくまでテキストをデータとして何らかの仮説検証や問題発見を行うための方法と言える.

なお，本論文では，テキスト計量分析という用語を利用しており，計量テキスト分析と混同しやすいため，以降では計量テキスト分析のことを計量内容分析と呼ぶこととする.

2.3.4. デジタル・ヒューマニティーズ

パーソナル・コンピュータが発達し，研究者が高性能な情報環境に容易にアクセスできるようになってきた近年，人文科学においても情報システムを利用した研究が盛んになってきている．デジタル・ヒューマニティーズと呼ばれるこの分野は国内外で徐々に勢いを増しており，国内でも 2012 年に JADH（Japanese Association for Digital Humanities）が設立されるなど，活発な動きが出てきた.

デジタル・ヒューマニティーズの研究は，大別すれば 2 つの方向性を持つ．一方は文化

的資源をアーカイビングするという目的を持って行われるものである．例えば立命館大学の GCOE では，京都や日本文化に関わる文化資源を 100 万件以上デジタル化し，データベース化してきた．もう一方は，人文科学的手法だけでなく，情報工学の手法を人文研究に利用し，新たなる知見を得ようとするものである．これらの研究は地理，歴史，芸術など様々な対象を持ち，手法も情報工学を活用するという点において共通することを除き，多様である．

その中には，当然テキストを対象とした研究も存在する．例えば，Plaisant ら[42]は詩人 Emily Dickinson がその義理の姉に送った手紙を対象としてコンピュータでテキストを分析した．Dickinson の詩に内在する「erotics」が，義理の姉との関係にあるのではないかという既存研究があり，この研究ではそれを手紙から読解することを試みた．具体的には，特別に設計されたシステムを利用し，人文科学者が手紙の一部を「hot」から「not hot」の 5 段階にレーティングする．するとシステムは文書内の単語情報を利用して，全ての手紙を hot と non hot にクラス分けし，それぞれにおいて特徴的な語を抽出する．結果，「mine」などの単語が hot な語として抽出され，これは既存の議論においては注目されていなかった語であることが人文科学者の考察によって明らかになった．しかし，最終的に，義理の姉と erotics の関係を明瞭に示すところまでは行かなかった．この研究は Nora Project という，文学にテキスト分析を施すためのプラットフォーム作成の 1 ケーススタディとなっている．

また，Omar は，小説家 Thomas Hardy の詩 62 作品から主題論的概念の検討をしている[43]．その結果として，複数の作品が含まれた 3 つのクラスターが「隠された死」「田舎の生活」「敵意」等に関係していることを指摘している．

ヒューマニティーズが元来テキストを対象とする以上，デジタル・ヒューマニティーズにおいてもテキストが対象となることは自然な流れであるが，現時点ではまだその手法に対して自覚的ではなく，テキストマイニングや内容分析から言葉を借りて分析方法を説明している状態である．デジタル・ヒューマニティーズは非常に多彩で，領域と言うよりも文理融合の試みの 1 つと言えるため，今後「デジタル・ヒューマニティーズのためのテキスト分析」という手法が出てくるかについては未知数である．

2.4. テキスト計量分析

2.4.1. テキスト計量分析の定義

テキストを計量的に分析することで，その書き手の思想や，知識を抽出しようとする一連の研究がある．先述したプロトコル分析から派生したこれらの研究は，自分たちを直接的に語る名称を持たず，ただテキストの計量的な分析と説明し，具体的な手法の説明を行っていた．

本論文では，これらの研究をテキスト計量分析と名付け，以下のような定義を付与することとする．「テキスト計量分析とは，書き手の頭の中にある知識，命題，信念，感情等の

要素が書き表されたテキストを計量的に処理することで，そのテキストを分析・理解するための方法」である．このテキスト計量分析は，分析の科学的な正当性を高めるために，以下のような特徴を持つ．

(1) 恣意的な判断を極力行わない．
(2) 計量の元となる要素の抽出については，高い精度を要求する．
(3) 計量値ないしは計量値を利用したアルゴリズムによる解析（ネットワーク解析等）を結果とする．
(4) テキストから特定の情報を抽出することではなく，テキスト全体を分析・解釈することを目指す．

これらの特徴について，類似分野であるテキストマイニング，計量内容分析との比較を表 1 に示す．なお，計量言語学は目的が言語学でありその内容ではないため，比較していない．

次節では，具体的なテキスト計量分析の事例を挙げる．

2.4.2. テキスト計量分析の先行例

2.4.2.1. 思想を対象とした研究

赤間はストア派カバニスと動物磁気論者メスメールという二人の思想家のコーパスから共起語を抽出し，因子分析を行った．その結果，「人間の身体組織を巡る思考的考察」及び「医学哲学」と捉えられる共通の因子が抽出され，両者の深い類似性が示唆された[44]．同じく赤間らは，ソシュールのキーワードとして最も有名な「シニフィアン」及びその類義語とされる「聴覚映像」が，どのようにして造られていったのかを，その著書「一般言語学講義」の第三回講義ノートを対象として語の共起ネットワークを作り分析した．「シニフィアン」，「聴覚映像」の登場以前・以後のテキストに対して複数の手法でマルコフクラスタリングをかけることで，「masse」というキーワードが以前・以後で意味合いが変わっており，「シニフィアン」，「聴覚映像」を媒介することを明らかにした[45]．

村井らは，政治家の Web ページに掲載されたテキストを収集し，分析した．結果，そこに使用されている名詞，形容詞の使われ方に一定の傾向が見られることが判明し，また政党によって特徴が異なることも明らかになった．さらに，「美しい国」という政治的スローガンについて，「美しい」という語が係り受けする語を他の語を調べることで，「国」という曖昧な語が実際は何によって構成されているのかについて調べ，やはり政党によってそのニュアンスが異なることを示した[46]．

表 1　関連分野との比較

要素	テキスト計量分析	計量内容分析	テキストマイニング
目的	対象テキスト及びその背後にある認知について理解する.	分析者の観点(問題意識)に基づいて，そのテキストの内容を量的かつ質的に分析する．また，テキストから仮説を説明する.	テキストから有用な情報を抽出する.
テキストの扱い	分析対象．分析したいテキストが選ばれる.	分析対象．問題意識に沿った基準で選ばれる.	データソース．規模があればノイズがあっても構わない.
計量単位	恣意性を廃した分割単位を基本とする.	コード（コーディングルールによって作られた意味単位).	アルゴリズムに依るが，様々.
分析者の関与・観点	分析者の恣意性を極力排除することで，対象の学問であろうとする.	テキストの性質とルールを公開することで客観性を担保した，特定の観点からの分析基準（コーディングルール）を利用して分析を行う.	人間の判断を必要としないことが前提だが，有用性があって労力がかからなければ，判断を取り入れることもある.
主要分野	認知科学	社会学	工学

2.4.2.2. 芸術作品を対象とした研究

青島らは，作曲家武満徹の音楽に関する自著から単語の共起ネットワークを構築し，その中心性から「音楽」,「映画」,「音」といった概念が武満徹の思考において重要であることを明らかにし,「音楽」及び「映画」を含む文の内容分析を行った．結果として，思想・世界観に基づく記述が武満の音楽に重要だったという結論に至っている[47].

工藤らの研究[48]では，村上春樹の初期三部作の本文テキストデータを用いて，単語の出現の変化を調べた．単語の中でも，主人公と同じ扱いを受ける重要な登場人物の名前である「鼠」に着目し，この単語と共起する語が作品を重ねるにつれてどのように変化しているのかを調査した．共起語のネットワークを作り，語の中心性を調べることで,「鼠」という語が徐々にネットワークの中心から外れていく，すなわち,「鼠」の小説内での出現頻度が相対的に減少するということを明らかにした．同じく工藤らは村上春樹の「1Q84」を対象として，並行形式小説の構造分析を行った．二人の主人公のそれぞれの章から高頻度名詞の出現ベクトルを作成して因子分析を行うことでそれぞれに共通する要素を抽出し，またそれぞれの章の頻出語の，章が進んだ際の頻度変化を分析することで物語が対照的な構

造となっていることを示した[49].

滕らはマンガアニメ作品に関連するブログテキストデータから作品名の共起ネットワークを抽出し，ストーリー展開を主要な要素とする作品群とキャラクターを主要な要素とする作品群があることを明らかにした[50].

2.4.2.3. 批評を対象とした研究

村井は，新刊書籍批評である「Web 本の雑誌」を対象とし，本のジャンル毎による特徴的な語を抽出し，語のネットワーク化によって書評で用いられる概念の構造を分析した[51]. 同じく村井は，同様の手法で文芸批評家井口時雄の批評テキストを分析した．形容詞の共起ネットワークから，批評文における「新しさ」,「美しさ」,「深さ」などの語が感性的特徴として現れること，またテキスト中に出現する人物名を計量化することで，批評をおこなう上での背景となる前提的な知識・コンテクストがどのような文学作品と思想家によって主に構成されているのかを示した[52]．さらに村井は，映画と演劇の批評テキストから作品名，人物名を計量し，どのような対象について批評文が言及しているのかを明らかにした．結果として，映画は漫画への言及が多く，演劇は思想への言及が多い事，映画は監督に関する言及が多く，演劇では監督と脚本の両方が同程度言及されていることなどが分かった[53].

また河瀬らの研究では，1987 年から 1993 年までの音楽評論雑誌を計量分析することで，その特徴を抽出した．批評テキストは 20 世紀以前と以降の 2 つのグループに分けられ，それぞれにおいて共起分析とネットワーク化が行われた．ネットワークはさらに中心性の分析を行い，中心的な語が 2 つのグループで異なることを明らかにした[54].

3. テキスト計量分析の基礎

3.1. 本章の目的

テキスト計量分析のフレームワークを構築するため，本章ではテキスト計量分析における基礎的な方法について，分析の流れに沿って述べる．その上で，テキスト計量分析の課題について論じ，フレームワークを構築する上で検討が必要な要素について説明する．

3.2. テキスト計量分析の流れ

テキスト計量分析には様々な事例があるが，テキストを扱うという点において変わりはなく，おおよそ表 2 の流れに沿って行われる．

表 2.テキスト計量分析の流れ

手順	内容
テキストの収集	対象となるテキストを収集する．
テキストのデジタル化	テキストを機械が処理しやすい形に整形する．また，テキストにメタデータを付与する．
計量単位への分解	テキストを計量する単位に分割する．
頻度分析	分解された計量単位の頻度を計算し，比較分析する．
パターンの抽出と分析	計量単位の様々なパターンを抽出し，そのパターンを計量し，またパターンから得られる情報をアルゴリズムで分析する．
特徴の定性的分析	抽出された特徴について，実際のテキストに当たり，特徴の意味合いを明らかにする．

次節以降では，この流れに従って各手順で利用される技術・方法について述べる．

3.3. テキストの収集

計量分析を行うには，対象とするテキストを選定し収集する必要がある．

3.3.1. テキストの収集

対象となるテキストの収集方法には，大別すると以下の 3 種類があると考えられる．それぞれの特性を理解して収集する必要がある．

(1) 書籍から収集する

紙書籍の場合には，スキャンと OCR を伴う．出版されているテキストであるということから品質も高く，新しい書籍であれば OCR の精度や形態素解析の精度も高いことが期待できる．一方で，同じトピックに関する資料を多くは集めにくい場合がある．また，語が洗練されている，あるいは意図的に統制されているため，本質以上に単純化された結果となってしまう可能性があるため，著者について論じたいのでなければ，なるべく複数の著者の複数のテキストを収集する必要がある．とはいえ，3 種類の中で

は最も扱いやすいテキストであると言える．

(2) アンケートなどの方法で被験者にテキストを生成してもらう

過去の蓄積を使う場合は別として，計量分析が有効な規模のデータを集められるかが課題となる．それ以外は，書き方などを統制することで，均質なデータが集まり，有効な分析を行いやすい．

(3) Web 上から収集する（ブログや口コミサイト，政府等のオープンデータなど）

クローラ等のプログラムを使えば数を集めることはたやすいが，特に CGM（Consumer Generated Media）の場合テキストの品質が低いという欠点がある．顔文字や半角カタカナの多用，間違った単語・文法などにより，形態素解析の精度が下がってしまう．

3.3.2. 権利

著作権法第 47 条の 7「情報解析のための複製等」において，情報解析とは「大量の情報から言語，音，映像等を抽出し，比較，分類等の統計的な解析を行うこと」を指す．テキスト計量分析のための紙書籍の電子化及びインターネットからのテキスト収集はこれに当たると解釈できるため，必要と認められる限度において著作権は制限される．本条文は平成 21 年に著作権法が改訂された際に追加されたが，この法律に基づく判例は 0 件であり，実際にどの範囲が「必要と認められる限度」にあたるかについては，今後の課題となっている[55].

なお，Web 上の口コミサイトなどでは，利用規約として分析用途の利用を制限している場合もあるため留意が必要である．

また，論文内でテキストの一部を転記するのは同 32 条の引用にあたるため，これも権利的な問題はない．

3.3.3. OCR

紙の書籍などは，スキャンした後にデジタルテキスト化する必要がある．このために必要な技術が OCR（Optical Character Recognition）である．OCR は成熟した技術と言って良く，市販 OCR ソフトウェアでの最近刊行された本の認識率は 99%を越える．しかし，これを 100%にする技術は存在しないため，OCR したテキストは人手でチェックし精度を確認する必要がある．

3.3.4. 文字コード

コンピュータ上では, テキストは文字コードを用いて符号化されている. Shift-JIS, JIS, EUC, UTF-8 等が一般的なテキストエディタで扱える文字コードだが, 原則として Unicode（UTF-8/UTF-16）を利用するべきである．一般的な日本語テキストであれば他の文字コードでも問題は起きないが，旧字や異体字などを含む場合，後続の処理で文字化けや文字の欠落が発生するためである．ただし，Unicode でも扱えない文字は存在するため，そういった文字は後続の処理で別の記号に置き換えたり，正規化したりする必要がある．

また，Unicode は現在も文字コードに入れる文字種の拡張に向けて議論が続けられており，将来的には古代文字なども含めた網羅的でグローバルスタンダードな文字コードになることが期待できるという利点もある．

3.4. テキストの整形

3.4.1. テキストのサニタイズ

収集したテキストは，その後の処理を行いやすいように整形する必要がある（サニタイズ）．まず，元が Web テキストの場合，HTML からテキストへの変換が必要である．さらに，画像や広告などの不必要な要素は取り除く必要がある．次に，文字種の正規化を行う．具体的には，半角カナなどを全角カナに直し，数字なども半角全角を揃える．また，旧字などがあると形態素解析が一部失敗するため，新字に正規化しておくことが望ましい．古い本でなくとも，OCR ソフトによっては旧字が紛れ込む場合もある．

3.4.2. テキストの構造化

後述する共起の抽出などを行う際には，その単位として「文」をコンピュータが認識できるようにする必要がある．分析に使用するソフトウェアによるが，例えば一行を一文とするなどの方法がある．

「。」や「.」などで機械的に文に分割することも可能だが，例外がないか確認する必要がある．Web テキストの場合，こういったルールでは整形できないこともあるため注意が必要である．

XML を用いることで，章や文の単位を厳密に構造化して記述する事が可能になる．TEI によってテキストを構造化するための汎用的なガイドライン（フォーマット）[56]が提示されている．

3.4.3. メタデータの付与

テキストが一様の群から構成されるのでなければ，群間で計量値を比較することは有力な分析方法となる．テキストの群を作るためには，個別のテキストに種別や年代などのメタデータを付与しておく必要がある．

メタデータのフォーマットとして決まったものがあるわけではないが，汎用的なフォーマットとしては Dublin Core[57]が挙げられる．これは主に WWW 上のリソースのメタデータを記述するために考案されたが，例えばそれを拡張した DC-NDL[58]は書誌データの記述に，またそれを利用する FOAF[59]は人のデータの記述に利用されており，汎用性が高い．他には，先述の TEI もテキストに対するメタデータの仕様を内包している．

3.5. 計量単位への分解

計量する，すなわち数えるためには，文字記号の連続体であるテキストを一定の塊に分解する必要がある．また，自明ではあるが，その分解の方法は恣意的でなく，かつ「同じものを同じもの」と判別できる明確な基準がある必要がある．

3.5.1. 文字

最もシンプルな計量の単位は，文字単位である．テキストをその文字単位で分割するため，分割の方法も自明である．

文字の計量は，文体論や著者推定，言語学的な研究には有用である．漢字やカタカナの量などが，文体的な特徴の指標となる．

3.5.2. n-グラム

n-グラムとは，記号列の中から連続する n 個の記号を取り出したものである．例えば，「あいうえお」から 3-グラムを作ると，「あいう」「いうえ」「うえお」が抽出できる．n-グラムの抽出は文法や意味などを考えずに機械的に行えるため，形態素解析の精度について気にする必要はなくなる．一方で，品詞等の情報は一切なく，また単語ではない断片も出てくるため，意味的な分析は行いにくくなる．形態素解析の精度が非常に低い場合には，有効な場合もある．

n-グラムのデータを大量に集めることで，特定の語の後にどの語が来やすいかを推測することができるため，自然言語処理の分野では発話予測などにも利用されている[60]．

3.5.3. 形態素

形態素(morpheme)とは，言語学において意味を持つ最小の単位を指す．テキスト分析の文脈においては，品詞を付与された単語と同義である．

形態素については，テキストから機械的に抽出する自然言語処理技術がある．形態素解析と呼ばれるこの技術では，テキストを単語単位に分割し（分かち書き），その後品詞や活用を付与する．

形態素解析は，人手で分かち書きされ品詞を付与された教師テキスト（一種のコーパス）を機械学習したプログラムによって，自動的に行うことができる．ソフトウェアとしては，日本語では 2.2.1 で述べた MeCab がデファクトスタンダードとなっている．

形態素解析は，単語や文章の意味を理解して区切っているわけではなく，あくまで機械学習した結果から，どこで切った場合に一番単語らしくなるか，という観点で区切っている．そのため，機械学習の元になっているコーパスに依存した結果となる．MeCab でスタンダードに使われているのは，新聞コーパスである．そのため，口語調の文章や旧漢字等が使われた古い文書では精度が下がる傾向がある．旧漢字については，国立国語研究所が，雑誌「太陽」や青空文庫所収の近代文語作品から構築した，近代文語 UniDic[61]という辞書を利用することで精度を向上させることが可能である．

英語では，Pentree Tagger[62]や，Stanford Parser[63]などがある．英語のような言語の場合，単語と単語の切れ目は自明なので，分かち書きをする必要がないが，連語・熟語がより強く意味に関わってくるため，日本語とは別の難しさがある．

形態素解析の精度は 100%ではない．特に，もともとのコーパスにない単語は，とりあえず未知の名詞として判断されるか，あるいは無理矢理単語分割されてしまう場合がある（例

えば，「ガンダム」→「ガン」+「ダム」)．そのため，特に専門用語が多いテキストの場合，辞書を作って形態素解析の精度を上げる必要がある．

3.5.4. オントロジ

文章では異なる単語が同じ意味を表している場合がある．形態素解析は，意味論を含めて解析するわけではないため，当然そういった区別はつかない．そのためには，何らかの意味論的な知識が必要になってくる．

オントロジとは本来哲学用語で「存在論」を示すが，情報工学の分野では共有された合意内容といったような意味で用いられる．例えば溝口による定義は，“対象とする世界の情報処理的モデルを構築する人が，その世界をどのように「眺めたか」言い換えるとその世界には「何が存在している」と見なしてモデルを構築したかを（共有を指向して）明示的にしたものであり，その結果得られた基本概念や概念間の関係を土台にしてモデルを記述することができるようなもの”となっている[64]．情報工学なので，当然記述された知識は機械可読という前提である．この定義によるオントロジは複雑に構造化された知識の表現を指向するが，その最も単純化されたパーツは同義語の集合とそれに与えられる名前（ラベル）と考えられる．本研究では，テキスト分析での実用性という観点で，オントロジを「同義語・類似語の集合（カテゴリ）に対してその内容を示すラベルが付与された機械可読なデータ」と定義する．具体的には図 2 のようなデータである．

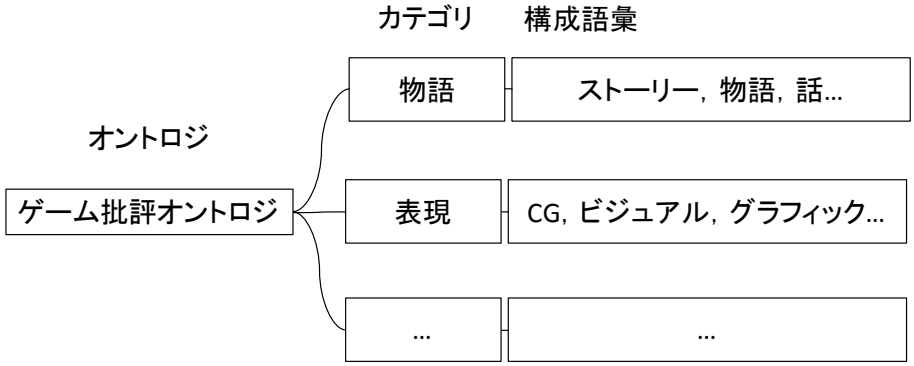

図 2　シンプルなオントロジの構造

テキスト計量分析においては，オントロジがあることで，単語ではなく，そのカテゴリでその後の計量を行うことが可能となる．大規模なデータや複数の書き手がいるようなデータの場合，単一の単語に注目してしまうと，単なる用法や語選択の違いによって数に影響が出て，解釈を誤る可能性があるため，オントロジが有効である．

前述の定義に準ずるなら，汎用的な用語については分類語彙表[65]など既存のシソーラスをオントロジとして利用することが可能である．一方で，その分野特有のオントロジは基本的に人手で作るしか方法がない．本研究では，4 章以降でこのオントロジをなるべく容易に生成する方法についても検討する．

3.5.5. 命題

命題（Proposition）とは，「夏は暑い」など，基本的に主語と述語からなる，文章としての意味を持つ最低限の単位を指す．文章から命題を抽出できれば，さらに精度が高い意味的な分析ができる可能性がある．しかし，命題を機械的に抽出する技術は存在するものの，多種多様なテキストに対して精度が確保されているわけではない．

文章の統語構造の解析は，係り受けについては先述の CaboCha で解析可能であり，またさらなる解析の技術も自然言語処理の分野で研究されている（[66]など）．従って文章としての最小構造を機械的にある程度の精度で抽出することはできるが，命題を計量するには，2 つの命題が同じであることを示さねばならない．そのためには，結局のところオントロジが必要となり，他にも文章構造の同義辞書が必要となってくる．

3.6. 頻度分析

計量単位への分割ののち，その頻度を計算する．最も単純にはその出現回数だが，他の指標もある．なお，様々な計量単位があることは前節で述べたが，以下では単語を例として述べていく．他の計量単位でも同様の分析は可能だが，意味合いは異なってくる．

3.6.1. 単語頻度

単語の頻度（term frequency，TF）を調査する手法は，計量言語学から受け継がれた最も基本的な手法の一部である[67]．頻度とは，テキスト中に出現する単語の回数を示す．

頻度が多い語で，かつ名詞などの非機能語は，そのテキストの意味的な特徴を表していると言える．一方で助詞や前置詞など機能語は文体論で主に用いる．

プログラムでの分析を行う場合，頻度が高い語だけに注目することで，以降の解析の範囲を絞ることができ，計算速度などを向上させることができる．

頻度を利用して専門用語などの特徴的な語を抽出する研究も多数存在する[68]．

3.6.2. 文書頻度

文書頻度（document frequency，DF）は，ある文書中に特定の単語が出現していたら 1 と数え，その中にどれだけの量が含まれているかについては不問とする計量方法である．この方法を取ることで，特定の文書にのみ大量に含まれている語に引きずられて，分析結果が歪むことを阻止することができるが，単語間で差がつきにくく，特徴抽出に失敗する場合もある．

3.6.3. TF・IDF

TF と DF を使って，産出される指標として，TF・IDF がある（式 1）[69]．

$$\mathbf{tf \cdot idf = tf \cdot} \boldsymbol{log} \frac{\boldsymbol{N}}{\boldsymbol{df}} \text{（式 1）}$$

ただし tf はテキスト d における語 t の頻度であり，df は語 t を含むテキストの数，N はテキストの総数．

TF・IDF は，出現頻度が多く，かつ少数のドキュメントにしか出現しない語が高くなるような値である．TF・IDF によって，その文書を特徴づける語が明らかとなるため，特にテキスト分類などの分野を中心に使われている[70].

IDF は，分析対象文書ではなく，「世の中にある一般的な文書」を基準として計算する場合がある（大規模コーパスを利用する）．これは，その単語が普遍的な意味で一般的かどうかということ見るためである．対象文書集合の範囲内で IDF を計算することもできるが，その場合，対象文書集合の偏りによっては，人間の目から見て一般的と言える語であっても，TF・IDF の値が高くなる．どちらが正しいというものではなく，前者は，誰が見てもこのキーワードと呼べる単語が抽出され，後者は対象文書集合という文脈においてキーワードと呼べる単語が抽出されると言える．

3.6.4. 異なり語彙数

文章中に出現する語の豊富さを示す，異なり語彙数という指標もある．これは単語単位の指標ではなくて，テキスト単位の指標であり，テキスト中に出現する語の種類の数によって表される．さらに，異なり語彙数を総語数で割った TTR（type-token ratio）という指標も用いられる．Yule は，K 特性値という TTR を拡張した指標も提唱している[67]．こういった計量値は，著者推定などの文体論で利用される．

3.6.5. その他の計量値

計量単位に基づく計量値ではないが，文体論では句読点の位置（一文の長さ）などを数える場合もある．また，単語の頻度ではなく，品詞のレベルで計量することで，文体的な特徴を捉えようとする場合もある．

3.7. パターンの抽出と分析

頻度分析では，基本的に単語単体での特徴しか見えてこない．意味的に深い分析を行うためには，単語で形作られるパターンを抽出し，さらにそのパターンとパターンの関わりを分析していく必要がある．

3.7.1. 共起

共起とは，特定の範囲内に出てくる 2 つの単語の組を意味する．特定の範囲をどのように設定するかは様々で，文単位や，文書単位，文の区切りを無視して前後一定の語数を見る（ウィンドウニング）こともある．また，コロケーションという形で，直前，直後に登場する語のみを見る場合もある．

文書中に存在する共起パターンを計量することで，何らかの関係がある単語のペアを見つけ出すことができる．共起の範囲を文とするなら，その 2 つの単語を同じ場所で使わなければ説明できない内容があることになり，その単語で表される 2 つの概念が書き手の頭の中で強いつながりをもっていると推測することができる．共起の範囲を文書とすれば，他の文書と比べることで，その文書のスコープを見いだすことができる．

共起は，文レベルの構造や意味を踏まえないため，必ずしも書き手の意図を正確に抽出できるわけではない．ある 1 つの共起だけ見れば，それが偶然の産物ということは十分にあり得る．しかし，計量し頻度が高いことが判れば，それは偶然ではなく，意図的なものにせよ無意識的なものにせよ，書き手の何らかの認知的要素を抽出できていると考えられる．

3.7.2. 係り受け

2.2.1 で述べた係り受け解析を行うことで，係る語と受ける語（厳密には文節）のペアを機械的に抽出することができる．抽出できるペアは，共起がより正確になったものと捉えることができる．単体ではなく，頻度で分析するのも同じである．

共起と比べて正確になった反面，全体としての頻度が減少するという欠点がある．表 3 は 5 章，6 章で分析する文書集合について，名詞を対象とした 1 文中の共起と係り受けの頻度を計量し，その最大頻度のペアの出現回数を示したものである．

表 3　共起と係り受けの頻度

文書集合	総単語数	共起最大頻度	係り受け最大頻度
河川文化テキスト	2323961	760	147
批評テキスト	2401394	560	88

いずれの文書集合でも，係り受けの最大頻度は共起の最大頻度の 20%以下となっている．ネットワーク分析などを行う場合，係り受けで抽出できるペアでは充分なネットワークを作ることが難しくなる．

また，例えば 1 文中に「、」で区切って 2 つの話題が書いてあるような場合，係り受けではその 2 つの関連を抽出できないが，共起であればできる．係り受けのほうが厳密なことに間違いはないが，必ずしも係り受けのほうが優れているとは言えない．

ただし，形容詞のような修飾語は，それが確かに修飾語として使われている場合，その修飾対象の語と強い結びつきがあるため，特定の名詞を修飾している形容詞の一群を抽出する，といった分析は有用であると考えられる．このような係り受けのペアを利用して評判分析を行った研究もある[71]．

また，英語では先述の Stanford Parser に係り受け解析の機能がある他，Malt Parser[72]も存在する．

3.7.3. 単語のネットワーク

共起や係り受けを調べると，語と語のペアが出現する．このペアをエッジ，語をノードとすると，語の関係性を表すネットワークを作ることができる．語のネットワークは語同士の関係を人間が捉えやすいだけでなく，ネットワーク解析の手法を用いることでその性質を数理的に調べることができる．このようなネットワーク解析の手法は，scientometrics

における論文間の引用関係のネットワーク解析[73]や，社会学におけるソーシャルネットワークの解析[74]の手法を，テキストに応用したものと言える．

単語のネットワークにおいては，エッジの重さ（他のノードへの距離）には，共起や係り受けの頻度を用いる場合が多い．

ネットワークは数学的にはただの行列であり直感的ではないため，様々なアルゴリズムを使って図として描画することで理解しやすくなる．ネットワークを描画するソフトウェア Graphviz[75]では，表 4 のようなアルゴリズムで描画することができる．

表 4　ネットワーク描画アルゴリズム

アルゴリズム名	特徴
neato	エッジをバネに近似し，グラフ全体が最小エネルギーとなるようにした描画方式．次数が高い語が中心の方に来るグラフが描画される．エッジの長さを調整しないと，ノードが重なってしまうことがある．
dot	階層的な構造を描くためのアルゴリズム．上の方の語から下の方の語へエッジがつながっていくグラフが描画される．距離は反映されない．
twopi	同心円状のグラフを描画する．中心となるノードが決定され，それから距離が 1（直接エッジを持つ）のノードが 1 つめの円周上，2 のノードが 2 つめの円周上，というように描画されていく．距離は反映されない．
cicro	ノードが円状となるような描画をする．語と語の距離が反映されないが，エッジがあまり重ならないため見易い．
fdp	neato と同様にバネ近似モデルを利用するが，描画されるグラフはノード同士が重ならないように調整される．

いずれにせよ，ネットワーク描画アルゴリズムは，多次元的性質を持つネットワークを二次元に射影して描画しているため，必ずしもその本質を示すものではない．しかし，ネットワーク構造の図は人間の頭が理解しやすいため，その単語の書いてある位置などで恣意的な解釈をしてしまう，という問題がある．そのため，ネットワークの解釈を行う場合には，以下で述べる分析方法を用いて行うことが望ましい．

また，例えば 1000 の共起パターンを描画すると，図としても解釈不能なものが表示されてしまうため，多くの場合ノードの数やエッジの数で制限をかけて図を表示することとなる．そのため，描画の範囲をどのような基準で制限しているかを明記することが重要となる．

3.7.4. ネットワーク分析

ネットワークを分析する方法は無数にあるが，ここでは汎用的にテキストの計量分析で利用可能と考えられる方法について説明する．

まず，ネットワークの基本的な性質として計算できる指標を表 5 に示す.

表 5　ネットワーク指標一覧

指標	計算方法	解釈
ノード数	－	同じ条件でネットワークを作れば，ノード数の違いは話題の豊富さといったような内容として解釈できる
エッジ数	－	エッジの数は，つながりの多さを意味する．多ければ全体として単語間の関係が密接で，少なければ関係が希薄と言える.
密度	$\frac{e}{{}_nC_2}$ eはエッジ数，n はノード数	エッジ数と同様，平均密度が高いネットワークでは単語同士の関わりが密接で，同じような内容を言葉を変えながら語っていると言える.
平均エッジ数	$\frac{\sum_n e_i}{n}$ e_iはノード i に接続するエッジ数，n はノード数	平均してどれだけの単語と関わりを持っているか．解釈可能な特性は密度と同様.

ネットワーク分析を用いることで，単語と単語の間の距離を計算することもできる．これにより，ネットワーク上で直接は接続されていない単語であっても，どこを経由することで，どのような距離で他の単語とつながっているかを調べることができる.

ネットワークにおいて，距離の計算には様々な方法があるが，2 種類に大別できる．1 種類目の指標は，共起関係がある単語を 1 とするものである（重み無し距離）．この場合，直接共起しておらず，ある別の単語と双方共に共起している 2 つの単語の距離は，1+1 で 2 となる．経路が複数ある場合には，最も近い距離をもって 2 単語間の距離とする．もう 1 種類の指標は，共起回数の逆数を 2 単語間の距離とするものである（重み付き距離）．つまり，共起回数が大きい単語のペアほど，その距離が近くなる．直接共起しない単語間の距離は，1 つめの指標と同様にその最短経路を計算する.

さらに，距離のような指標を使って計算できるネットワーク全体の特徴指標として，中心性の指標がある．中心性とは，そのネットワークにおける中心となるノードを見つけるための指標である．図で描画すると，多くの場合，図の中心に描画される語の中心性が高いが，前述の通り図は正確なネットワーク構造を反映していない場合があるため，数値的な指標で解釈を行うべきである．中心性が高い単語は，それがそのままテキストで中心となっている話題に関することがあり[48][54]，テキストの意味理解に有用である．代表的なネットワークの中心性としては，表 6 の指標が挙げられる[76].

表 6　中心性の指標

中心性	計算方法	性質
次数中心性 (degree centrality)	次数中心性は，ノードに接続しているエッジの数に等しい.	最も多くのノードと接続されたノードが中心となる. テキストにおいては，複数の語と関わりを持つ明示的なトピックの中心となる語と考えられる.
近接中心性 (closeness centrality)	$c_i = \frac{n}{\sum_j d_{ij}}$ d_{ij}はノードiからノードjへの重み付き距離，nはノード数	次数中心性はその近隣の語との関係のみを考えるのに対し，近接中心性は全ての語を考えた上で，中心を決めていると言える（全ての語と近いと言える語が中心となる）. テキストにおいては，総体として見た時に暗黙的に中心となる語と解釈可能だが，分断されたネットワークがあるとそれに含まれる語の値が高くなり，あまり有効ではない.
媒介中心性 (betweeness centrality)	$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}$ g_{st}はノードsからノードtへの重み付き最短距離のパスの本数, n_{st}^iはノードsからノードtへの重み付き最短距離のパスのうちノードiを通るパスの本数.	他の語と語を結ぶ経路上にある点を高く評価する指標となる．複数の語のまとまり同士を結びつけているような語（hub）が中心となる. テキストにおいては，複数の概念グループを結びつけている，領域の境界の語と考えられる.
固有ベクトル中心性 (Eigenvector centrality)	$x_i = \kappa_1^{-1} \sum_j A_{ij} x_j,$ $A_{i,j}$は隣接行列（iとjの間にエッジがあれば1なければ0), κ_1は最大固有値．値が収束するまで繰り返し計算を行う.	自分に対してエッジを張っているノードがどれだけの中心性を持っているかを考慮する指標. 他の単語の影響力も加味されたトピックの中心となる語を示す.

最後に，ネットワークを複数のサブネットワークに分割するアルゴリズムがある．ネットワーク全体では解釈が難しい場合や，一見してそのネットワークが複数のグループに分かれることが見て取れる場合，ネットワーク分割アルゴリズムを利用することで解釈可能な単位に恣意的でない方法で分割することができる.

具体的なアルゴリズムとしては, 例えば Girvan-Newman コミュニティ抽出法がある[77]. 名前の通り，元来はソーシャルネットワークからコミュニティを抽出するためのアルゴリズムである．このアルゴリズムでは，ノードではなくエッジの媒介中心性を計算し，それが高いエッジを切断，再度媒介中心性を計算，という処理を繰り返す．これにより，1 つのネットワークから凝集性の高い（＝閉じた）複数のサブネットワークが切り出される．なお，このアルゴリズムは 1 つずつネットワークを切っていくだけであるため，いくつのサブネットワークに分割するかについては，分析者の意図に依存する．ただし， Modularity を計算することで，何分割のネットワークの質が良いかを示すことは可能である．Modularity は 1 以下を取る数字で，分割されたネットワークの各グループ内のノード間のエッジの割合から，エッジがランダムに配置された場合の期待値を引いた値である．従って，分割されたそれぞれのグループ内の結合が密であるほど大きな値を取る．この値は多くの場合 0.3 から 0.7 を取ることが判っており, 0.3 を下回る場合には分割の精度が低い（＝ランダムと変わらない）可能性が考えられる[78].

本研究では，5 章で Girvan-Newman コミュニティ抽出法を利用している．

3.7.5. クラスタリング

クラスタリング(clustering)とは, 対象となるデータ集合を部分集合（クラスター, cluster）に切り分ける手法である．クラスターを作るアルゴリズムは複数あり，またパラメータの取り方は無数にあると言える．例えば，工藤らは村上春樹の複数の作品について，品詞の頻度と語の分類の頻度を合わせて特徴量とし，作品レベルでの階層的クラスタリングを行うことで, 作品の関係性を計量的に示している[79]. 本研究では, 4 章で共起を特徴量とし，ピアソンの相関係数を距離とした名詞の階層的クラスタリング（ウォード法）[69]を行っている．

3.7.6. 統計解析

テキスト計量分析と同じく言語を計量的に扱うコーパス言語学では, χ^2検定, 相関分析, 回帰分析，線形判別法，主成分分析，因子分析等の統計解析が利用されている[69][80]．特にχ^2検定は，コーパス間における単語や単語群の頻度の差に違いがあるかを検証するために利用されており，残差分析も行うことでどのコーパスにおけるどの単語が期待値から外れているか，すなわち有意に違いがあるかを見つけ出すことができる．本研究では，4 章から 6 章において，このχ^2検定を様々な場面で利用している．

3.8. 特徴の定性的分析

単語の頻度や単語のネットワークそれ自体はテキスト計量分析の成果ではない．その結果の意味付けには考察が必要であり，人文的なテキストの場合には人文的な知見に基づく考察を数値的結果と融合させる必要がある．その際には，抽出された特徴が実際の文章中でどのように使われているか，改めて確認する必要がある．というのも，単語 A と単語 B の共起が多いという結果が出てきたものの, 中身を見てみると AB という連語が形態素解析

できずにAとBに分かれていただけだった，あるいは，AとBを含む慣用句の繰り返しが多いことが原因だった，などという場合があり得るからである．また，原文を見ることで，抽出された関係や構造についての考察が可能となる．

3.8.1. KWIC

KWIC は，特定の単語がどのような文脈で使われているかを見ることで，その単語の特徴を明らかにするための手法である．コンコルダンス分析とも言い，もともとは，聖書学などで使い始められた手法である．KWIC は，文書を読むという人文科学の基本的な作業に一番近い分析手法であり，特に意識せずに用いられていることもあると考えられる．しかし，コンピュータを利用して単語を網羅的に調べることで，さらなる知見が得られる可能性もある．

ツールを使う場合には，対象となる単語の前後何十文字かを検索して一覧表示させることで，効率よく，また網羅的に確認ができる．

3.9. 分析のためのモデルと基本方針

まず大前提として，本研究におけるテキスト計量分析フレームワークの目的はテキストの意味を計量的に扱うこととする．テキスト計量分析においては意味ではなく形態を扱うことも可能であるが，形態論は計量言語学で議論されてきた内容であり，改めてフレームワークを検討する必要性はないと考えられるからである．

その上で，フレームワークを検討するに当たって，本研究では徃住による心の計算パラダイム[81]をモデルとする．このモデルでは，人間の認知的な働きを，言語をベースとする記号の計算で説明しようとするものである．認知科学において，このモデルの説明力が高いというわけではないが，“言語，知識，推論，思考という高次認知部門との適合性が高いのはいうまでもないが，これらを構成する諸概念の新しい組み合わせとしての感情，審美，価値，信念といった心的概念において大変な力を発揮できる”ことから，人間の高次認知の目に見える１つの成果たるテキストを対象とするに当たっては適当なモデルと言える．

このモデルに立つ場合，テキストとは，その対象を表す記号（知識，命題，信念，感情）の計算結果ないしはプロセスが，「執筆」という別の計算プロセスを経て，文字列となって出力されたものとなる（図 3）．

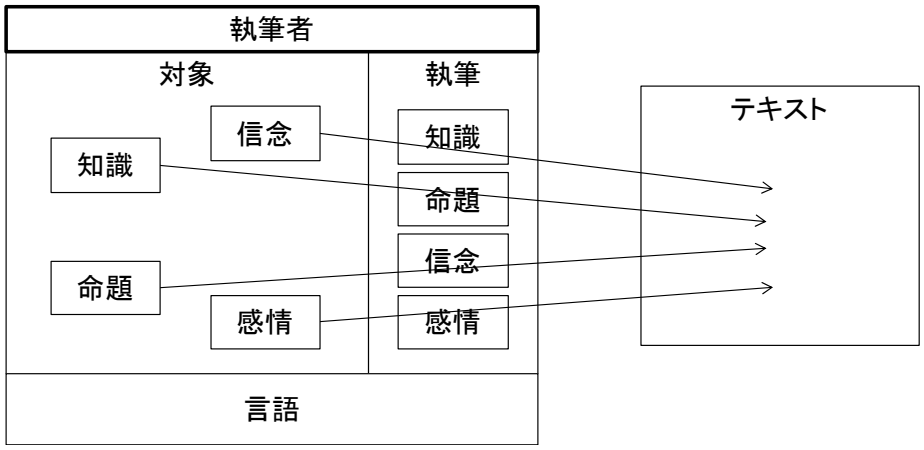

図 3　テキスト執筆の認知モデル

しかしながら，現実問題としては，内部の記号がどのような計算によって文字列に変換されるのか，また，執筆という再計算による影響（例えば，プロトコル法と違って推敲があること）などについて，明確な答えはない．従って，現時点では，以下のような前提・仮定を置くこととしたい．

(1) 執筆というプロセスによる影響については，そのプロセスを含むテキスト産出総体としての分析であることを暗黙の前提とする．すなわち，執筆プロセスの影響を除外することはしない（できない）．ただし，執筆において明らかなバイアス（e.g.思っていることと違うことでも書くことで報酬を得られる，等）がかかることが想定される場合を除く．

(2) 対象を表現する記号は，言語の枠組みに縛られ，何らかの形でテキスト中の表現として表出する．すなわち，「意味のない」表現はなく，逆にテキスト中に表現されていない内容については分析し得ない．

(3) 名詞（句・節）は対象に関する概念を表し，形容詞（句・節）や動詞（句・節）と合わさって，「x は y だ」，「x は y した」等の知識，命題，信念，感情などを示す構造体（以降，全て合わせて命題とする）となる．

モデルとしての前提・仮定は以上の通りだが，ここで技術上の制約が生じる．(3)の仮定に従い，テキストから命題を抽出すれば良いわけだが，3.5.5 で述べた通り，命題の自動抽出は現時点では多種多様なテキストに対して精度が確保されているわけではない．さらに，命題の基礎となり得る係り受けについては，3.7.2 で述べた通り抽出される頻度が少なくなってしまう．この場合，少し複雑な文では，係り受けではその関係性を抽出できないことになってしまう可能性がある．なお，テキストマイニングの分野では，文の意味をその格構造，単語などから推定し同じ意味の文を見つける意味分析などの技術もある[82]．しかしこのような自動化は，精度が 100%ではない上，何らかの辞書データや学習データが必要で，さらにその解釈の妥当性について確たることが言いづらい結果を生み出すと考えられる．

そこで本研究では，単語をベースとして分析を行うこととする．すなわち，命題をさらに分解した名詞や形容詞，動詞の単位での分析とするのである．頭の中における記号としても，言語をベースとしている以上，その裏には名詞や形容詞など単語レベルの意味合いを持っていると考えられるからである（図 4）．

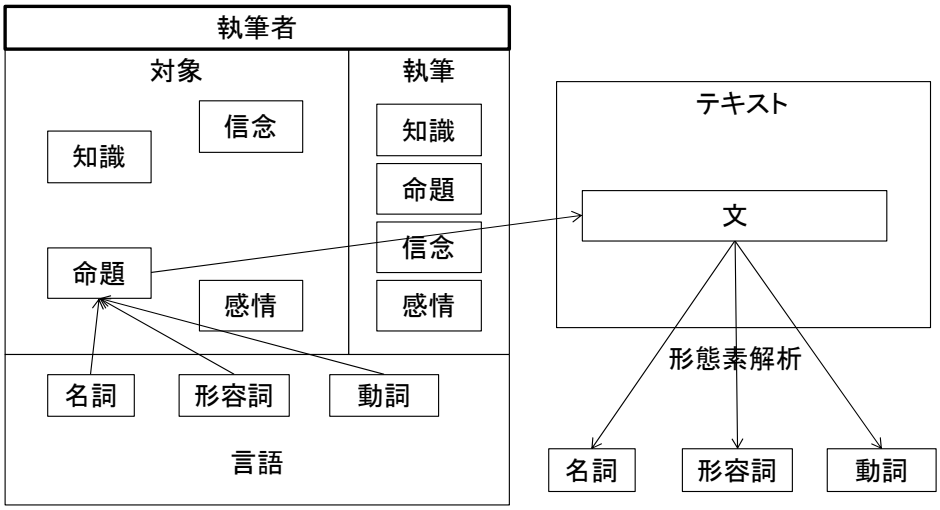

図 4　単語を踏まえたテキスト執筆の認知モデル

単語をベースとすることで，計量単位への分解は形態素解析で機械化でき，頻度についてもほぼ漏れがないことは担保される．しかしながら，意味構造を踏まえなくなるため以下のような制約が生まれる．

(1) 同じ概念を指している単語が複数あっても，それぞれ別のものとして認識してしまう．
(2) 単語の多義性を扱うことができない．出現場所によって意味合いが異なるような単語についても，「同じ」として扱ってしまう．
(3) 句や節などの構造を扱うことができない．複数の単語が集まって明確な概念を示しているような場合に，それを分解してしまう．

こういった制約は命題レベルの分析の利点の裏返しとなっている．これらの特徴をまとめたのが表 7 である．

テキストマイニングの分野においては，データの中から有用と思われる情報一片でも見つけ出すことができれば有用であると言えるため，その過程がなんであろうと大きな問題はない．一方で，テキスト計量分析が目的とするのは対象の理解であるため，不用意に自動化すると，手法としての信頼性が下がってしまう．命題と単語，どちらの分析レベルを選択したとしても曖昧さや不正確さが残ってしまう以上，プロセスの再現性が高い単語レベルの手法を採用することが望ましいと言える．

表 7　解析のレベルによる利点と欠点

分析のレベル		技術	利点	欠点
命題	手動	―	・正確な分類・分析が可能	・大量データを処理するのにコストがかかる. ・分類・分析から主観を排除するのが困難
	機械	意味解析等	・大量なデータに対して，意味的な分析が可能	・精度を確保できない ・数での比較が難しくなる ・自動化された内容について，説明・解釈が難しくなる.
単語		形態素解析	・大量データを処理可能 ・表面的な情報において欠落がない ・数での比較が行いやすい	・抽出できる情報の意味合いは曖昧

3.10. フレームワークに必要な要素

前節でのモデル及び基本的な方法の選択を踏まえると，フレームワークの構築に当たって必要な要素は，単語レベルでの解析における制約を緩和させるための手法となる.

まず，同じ概念を指している単語が複数あり得るという問題については，3.5.4 で述べたオントロジを導入し計量することで解決できる．オントロジは，一般的な類義語情報を導入する以外に，対象のテキストから自動生成する方法や，手動で構築する方法が考えられる．本研究では，これらの 3 つの方法について，3 つの異なるテキスト群に対して適用し，その有効性を検討した．次に，句や同義語などの詳細な意味を扱えないという問題から生じる制約を軽減するため，構築したオントロジを用いて対象テキストを大きな単位・構造で分析する方法について検討した．ただし，分析の方法は個別の研究目的によって様々であるため，網羅的に扱うことはできない．そこで本研究では，3 つのケーススタディを通じて，それぞれにおいて必要な分析方法を検討することとした．単語の多義性の問題については，それを解消するためには人手の作業ないしは意味理解の技術が必要になると考えられるため，本研究では検討していない.

4 章から 6 章にかけてのケーススタディについて，その目的と，どのようなオントロジと分析の方法が利用されたかを表 8 にまとめる．それぞれの内容については，次章以降で詳細に述べる.

表 8　各ケーススタディで採用されたオントロジと分析方法

対象テキスト	目的	有効なオントロジ	分析方法
ゲーム批評	ゲームにおける批評の対象とその変遷を明らかにする	自動生成	オントロジを単位とした計量
河川文化	河川文化の全体像を理解する	手動構築	オントロジの共起を利用した構造の抽出
文学，映画，演劇，ゲームの批評	分野における評価・感性の違いを明らかにする	既存オントロジの利用	オントロジを単位とした計量比較による特徴抽出

3.11. 以降の分析で使用したソフトウェア

形態素解析は，MeCab を IPADic で利用した．χ^2 検定については，js-STAR[83]を利用した．また，階層的クラスタリングについては R[84]を利用した．

4. オントロジの自動生成による概念カテゴリ計量－ゲーム批評の批評対象要素の抽出－

4.1. 本章の目的

本章では，ゲーム批評からその批評の対象を抽出する分析を通じて，オントロジの自動生成と計量が分析として有効であることについて述べる[85].

4.2. 分析の背景と目的

ゲームを 1 つのメディアとして見た時に，それを遊ぶ（プレイする）という体験には様々な要素が含まれている．その中には画像や音声，コントローラーの振動，操作などの感覚的体験だけでなく，プレイに伴う試行錯誤やストーリー理解による感動など，高次認知的な体験も多く含まれていると考えられる．この中で，何がゲームをゲームたらしめ，ゲームを面白くしているのかという点については活発な議論がなされている．

Malone は発達心理学や認知科学などの背景を元に，ゲームをプレイする動機として，「challenge」「fantasy」「curiosity」の 3 つがあると述べている[86].「challenge」と「curiosity」は字義通りの意味である.「fantasy」はそのゲームという環境が与える動機であり，願望の充足などを意味している．また，メディア学の立場から，Vorderer はゲームなどの新しいメディアの根源的な面白さは「interactivity」にあり[87], ゲームにおいては「interactivity」が「competition」に基づく楽しみとなっていると指摘している[88]．プレイヤはインタラクティブに競技的要素を完遂していくことで楽しみを得るのではないかというこの仮説は，アンケート調査法によって得られた，競技的要素を含み，よりたくさんの行動を許容するゲーム状況の方が好まれるという結果によって支持されている．Sweetser では，Csikszentmihalyi によって提唱された「Flow Experience」という体験のモデル[89]にゲームを適合させ，「challenge」だけでなく，プレイヤの状況操作能力を示す「control」，ゲームからのレスポンスである「feedback」，ゲームへの没入を示す「immersion」などの要素を統合した「Gameflow」という評価モデルを構築し，実際に 2 つのゲームの比較を行っている[90].

Song らの研究では，MMORPG（Massively Multiplayer Online Role Playing Game）におけるプレイヤの問題解決の認知的プロセスを記述することを目的とし，既存研究を元にモデルを構築した上でプロトコル分析を行った[91]．結果を元に，Newell による PBG（Problem Behavior Graph）[92]と類似した GBG（Gameplay Behavior Graph）を描き，それによって上級，中級，初級のプレイヤ間での問題解決のパターンの違いを明らかにした.

また，近年では，ゲームを 1 つの芸術や文化として捉え，人文科学的に評価しようという動きもある. しかし既存の文学, 詩や演劇と同様に扱うこともできないため,「Ludology」というゲームの仕組みやデザインにフォーカスした方法論が提唱されている．一方で，ゲームとそのストーリーは密接に関係あるという「Narratology」という方法論を唱える立場

も存在し，論争が行われてきた[93].

馬場はゲームの面白さとして，プレイヤ・ルール・ツールのバランスが最適であることを挙げている[94]．この場合のプレイヤは複数人で遊ぶことも考慮し，主体となるプレイヤだけでなく，プレイヤ同士の関係性も考慮した概念である．ルールとは，ゲームの難易度を決定する概念であり，上述の「challenge」に類似している．一方ツールとはゲーム機を指す．近年のゲームは特に，グラフィックやサウンドでプレイヤに多くの情報を与えるため，これを無視することはできない．馬場は，既存の遊びの理論がツールに対する配慮が少ない点を指摘している．山下らはアンケート調査法によってゲームに対する感情表現を採取し，因子分析を行った．結果，「ゲーム本来の楽しさ」，「感覚運動的興奮」，「設定状況の魅力」，「和みと癒し」，「難解・頭脳型」などの因子を抽出している[95].

これらの既存研究には 2 つの問題点がある．1 つめの問題は，定量的な方法論の欠如である．基本的に理論やモデルの提案や，個別のゲームにおけるケーススタディが根拠となっているため，山下らのようなアンケート調査方式の研究を除いて，どういった要素が支配的なのかについて科学的な議論ができない．2 つめの問題は，ゲームの発展に研究が対応できていない点である．Malone で調査されていたゲームはごく単純なピンボールや，テキストベースのシミュレーションなどであった．一方で Song らの研究で対象とされていた「World of Warcraft」という MMORPG は，世界中で多くの人間が同時に遊ぶ，非常に美麗なグラフィックの RPG である．このような時間と共に大きく変化する対象を扱うためには，時系列での分析が必要となる．しかしながら，ゲームの歴史的推移に関する研究は少なく，蔵の研究のようにハードウェアなど産業的な推移に注目したものがほとんどである[96].

これらの問題を解決するため，本分析では批評テキストに着目する．批評テキストとは，その本来の意味から考えて，ある特定のジャンルや作品に関して，その価値がどこにあるのかを論じたテキストであると考えられるため，その内には対象に関する人間の認知構造や価値判断等の情報が含まれているはずである．もちろん，テキスト中に直接的・明示的に対象についての認知構造が記されているわけではないが，対象に関する描写，用語の使い方，評価などにはその背後にある認知構造が意識的・無意識的に反映されているはずであり，大量のテキストを分析して共通的な特徴を抽出することでその構造に定量的に迫ることが可能となる．また，複数年に渡る均質なテキストを収集することで，批評の時代の変遷を知ることが可能となるはずである．

ゲームにも批評テキストは多く存在する．過去にはゲーム情報誌が主な掲載場所だったが，インターネットの普及に伴い，オンライン上でも多くの批評が見られるようになった．また，商用ゲーム雑誌にもゲームの批評が掲載されている．これらのテキストを大量に分析することで，ゲームの特定ジャンルや特定の書き手の影響を受けにくい，より汎用的な結果を得ることが可能になる．

テキストには種々の要素が含まれるが，「何が」ゲームの面白さであるかという問題に迫

るためには名詞に着目すればよいと考えられる．一般的に，ゲームの批評には「物語」や「ゲームシステム」といった要素についての説明や評価が記されている．批評であるからには，ゲームにとって重要でない，すなわちその面白さ，価値に関係のない記述は少ないはずであり，名詞として記述されるこういった要素の登場の仕方や頻度を調査することで，ゲームにおいては重要なことは何か，ということを明らかにできる．

以上より，本分析の目的は，ゲーム批評テキストに含まれる名詞の計量分析により，そこに込められたゲームの面白さに関わる要素とその構造を抽出することである．複数年に渡って刊行されたゲーム雑誌の批評テキストを計量分析することで，ゲームの評価においてどのような要素が支配的・特徴的であるか，さらにそれが時系列でどのように変化してきたかを明らかにする．

4.3. 対象データと手法の選択

分析対象とするデータは，マイクロマガジン社発行のゲーム雑誌，「ゲーム批評」[97]のゲームソフト批評のコーナーのテキスト vol.1～vol.69（1994 年～2006 年）とした．「ゲーム批評」は 1994 年 9 月に創刊された季刊（96 年 1 月より隔月刊）の雑誌で，2006 年 7 月の休刊までに計 69 冊が発行された．この雑誌の核となるのは，タイトルの通りゲームに関する批評を掲載する記事だが，中立的な視点を保つため，ゲームの広告を打たないという特徴がある．一般的なゲーム雑誌では，新しいゲームの速報と攻略のための情報が主たるコンテンツになっている．そのため，特定のゲーム会社と懇意になり，なるべく早く新しいゲームの情報を手に入れる必要がある[97]．従ってその記事の多くは宣伝を主とする，いわゆる「提灯記事」であり，公正な批評ではない場合が多いと考えられる．一方で「ゲーム批評」では，発売後のソフトを完全に攻略した後に批評している．そして，ゲームの紹介や良い点だけでなく，否定的な評価や意見も見られる．批評記事は読者投稿ではなく専門の記者によって書かれており，計 140 名（うち女性 6 名）の外部の記者と，17 名の編集部内部の記者がいた（編集部の記事は全体の 5.4%）．プロフィールを見ると，記者はゲーム関連のフリーライター，及びゲーム開発に携わっている者が多くを占める．ゲームの批評・レビューテキストは他にも存在するが，「ゲーム批評」を用いることで，コマーシャリズムを排したゲーム内容そのものの批評を調べることができる．加えて，専門的な批評家による執筆のため，一般的なユーザでは自然言語として表現しにくい内容，及び専門家の豊富な知識を背景とした，評価についての詳細な理由や対象の分析といった内容も得られると考えられる．ただし，専門家の知識構造や評価観点が一般ユーザと異なる可能性も考えられるため，4.6 ではユーザレビューとの比較を行った．

コーナー本文を OCR でテキスト化したところ，884 本の記事が得られた．合計 32984 文（「。」区切り），形態素解析の結果 886951 単語となった．なお，形態素解析の際，未知語と認識されるゲーム特有の語については専用の辞書を構築した．また，「対戦」と「格闘」に分解されてしまう「対戦格闘」などの連語については，その出現頻度が 50 以上であれば，複合語として辞書に登録した．50 という数字については，各 n グラムを頻度順に上から見

ながら，「ゲームー自体」や「ゲームー部分」等の一つの語として扱うには不適切と考えられる語が多くなる前の数字で切りのよい数字を選んでいる．

批評テキストが対象とするゲームについては表 9 に示す．なお，ジャンルの分類は「ゲーム批評」本誌の記載，並びに Amazon のゲーム売り場におけるカテゴリをもとに，筆者が決定した．

表 9　対象ゲームとその分類

ジャンル	本数	ハードウェア	本数
アクション	196	ニンテンドーDS	18
ロールプレイング	159	PSP	13
アドベンチャー	134	Xbox360	6
シミュレーション	126	ゲームボーイ	51
シューティング	92	ゲームキューブ	36
その他	38	Xbox	32
スポーツ	35	プレイステーション 2	251
対戦格闘	34	ドリームキャスト	49
レース	31	NINTENDO64	35
パズル	14	プレイステーション	197
テーブル	14	セガサターン	41
音楽	11	ワンダースワン	5
		SFC 以前	49
		パソコン	27
		アーケード	73
		その他	1

4.2 で述べた通り，ゲーム批評の評価の対象となる要素は，批評テキスト中に名詞（一般名詞・固有名詞・サ変接続名詞，以下同様）という形で存在しているはずである．表 10 にテキスト中の高頻度名詞上位 30 語を示す．【部分】や【要素】など，単語のレベルでは解釈し難い語も混じっているが，【敵】，【キャラクター】，【ストーリー】など，ゲームの要素を直接指す専門用語が多くを占めることが判る．しかしながら，上位 30 位以内でも【ストーリー】と【物語】のような同義語が存在するため，その同定が必須である．また全名詞を見るとユニークな名詞が 20605 語あり，頻度合計で 30%をカバーするためでも 143 語，50%をカバーするためには 550 語は分析の対象としなければならない．従って，頻度の高い語のみを対象とした場合，同様の概念や対象を指し示す他の語を無視することになってしまう．3.9 で述べた単語レベルでの分析の欠点である．

そこで，0 で述べた方針に従いオントロジを導入するが，そのための 3 パターンの方法を

検討する．まず，既存のオントロジの利用だが，ゲームは専門用語が多いため，汎用的な類義語データでは単語の網羅性が低い．また，参考にするべきゲームに関する構造化された用語集も存在しない．次に，テキストデータをもとに手動構築する方法だが，本分析の目的が評価の対象がどのような語であり，量がどの程度かということを明らかにすることであるため，構築作業そのものが直接的に結果に影響を与えかねない．ゲームの専門家がオントロジを構築し，それを用いて分析することも考えられるが，それはその専門家から見たゲームの概念構造となってしまう．量については評価できるが，概念構造については評価不可能となってしまう．そこで，ゲーム批評の分析においてはテキストデータからのオントロジの自動生成を試みることとした．生成されたオントロジは，テキストの背後にある認知的構造の一端を表すと考えられるため，本分析ではオントロジの生成そのものが対象の分析となる．さらに，生成されたオントロジのカテゴリを単位として計量することで，高単語レベルでの計量よりも正確にゲームの特徴を捉えることができる．

表 10　出現頻度上位 30 位までの名詞の出現頻度と出現テキスト数

単語	出現頻度	出現テキスト数	単語	出現頻度	出現テキスト数
ゲーム	5409	840	攻撃	686	272
プレイヤー	1676	580	戦闘	677	240
作品	1438	544	自分	619	323
プレイ	1187	514	存在	615	379
本作	1176	370	RPG	607	204
敵	1141	385	物語	600	255
システム	1136	477	ストーリー	580	276
キャラクター	1112	395	キャラ	552	245
世界	920	381	前作	551	204
シリーズ	834	307	発売	527	309
部分	786	404	ソフト	523	251
人	775	417	表現	498	269
主人公	723	339	シナリオ	498	207
ユーザー	704	317	登場	492	319
要素	697	400	魅力	480	288

オントロジの自動生成は，単語のクラスタリングによって実現できる．テキスト中に登場する語を何らかの指標でグルーピングしたカテゴリを作り，カテゴリの名前を付与すれば，3.5.4 で説明したオントロジの定義に合致する構造が得られる．

クラスタリングを行うための指標としては，テキスト単位での出現量や他の単語との共起回数が考えられるが，本分析では後者を採用する．前者の場合はゲームにおける語の意

味というよりも，語と特定のジャンルとの結びつき（RPG で多い語，等）によって単語がグループ分けされてしまう可能性が高い．後者の場合は，共起のウィンドウを文単位とすることでジャンルによる束縛からある程度逃れることが可能であり，さらに単語ベースの分析の欠点である，文の構造を扱えないという課題を擬似的に乗り越えることが可能となる．なぜならば，文中の単語共起をパラメータとして抽出されるカテゴリは，同じような共起の特徴を持つ，すなわち，同じような文，同じような表現で使われている単語の集合であり，カテゴリの出現量と，そのカテゴリの意味に関する記述の量には相関関係があると推定できるからである．

パラメータとして利用する共起の特徴を明らかにするため，共起ネットワークとして可視化を行った（図 5）．頻度 162 回以上（出現頻度で名詞の 30%をカバー）の名詞における文単位の共起について頻度 30 回以上のものを描画し(79 ノード, 202 エッジ, 密度 0.0656, 平均エッジ数 10.2)，描画アルゴリズムはノードが重ならないことを最優先としている．ノードの大きさは単語の出現頻度を示す．なお,【ゲーム】という語を含めた場合，それだけ次数が 100 以上あり（図 5 では最高でも次数 35)，ネットワークがゲームのエゴセントリックネットワークのように見えるため，ここでは除外している．この図を見ると,【敵】,【キャラクター】,【プレイヤー】,【作品】,【攻撃】等を中心として，いくつかのグループがある様子がうかがえる．特定のノードの次数が極端に高い構造ではないため，共起ベクトルはそれほど疎にならず，クラスタリングのパラメータとして適切であることが予想される．

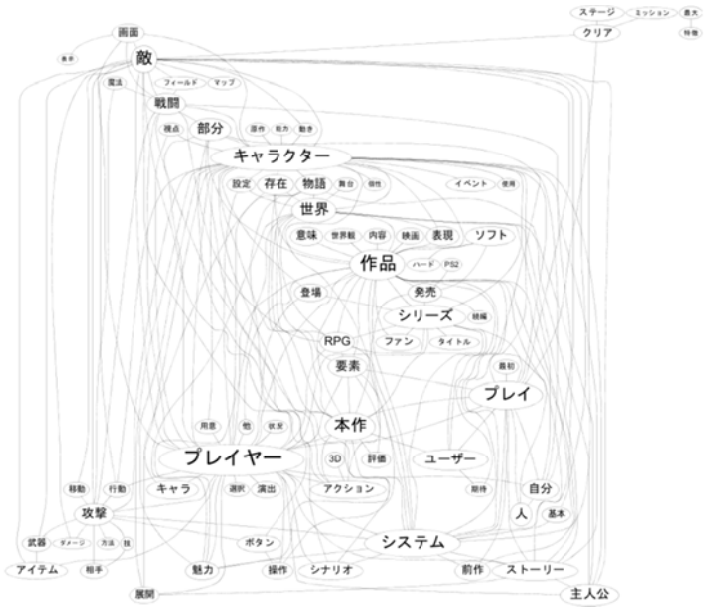

図 5　出現頻度上位 30%の語による共起頻度 30 回以上の共起ネットワーク

以降の流れを説明する．まず次節 4.4 では，オントロジ自動生成を行う．自動生成は，2 段階の階層化クラスタリングによって行った．この際，得られたカテゴリの特徴を分析す

るため，カテゴリに含まれる語と係り受けする形容詞・動詞についても調べる．既存研究の成果や高頻度の名詞，共起ネットワーク図から考えると，「キャラクター」や「ストーリー」などの状況設定に関する要素と，「戦闘」や「システム」等のルール的な要素が抽出されることが想定される．

続く 4.5 では，4.4 で得られたカテゴリが時間と共にどのように増減するかを分析する．ゲームはハードウェアの入れ替わりにより進歩し，グラフィックの向上，通信等の機能追加がなされる．そこで，新しい世代のハードウェアの登場を区切りとしてデータを 4 つの年代に区切り，年代毎の差異を見ることとする．ハードウェアに基づく変化として大きいことが予想されるのは，32 ビット機と呼ばれるプレイステーション・セガサターンの登場によりグラフィックが大幅に向上し，2D から 3D が一般的になる点と，タッチスクリーン・無線通信機能を備え，年齢を問わず普及したニンテンドーDS の登場だろう．前者は画面周りに関するカテゴリの変化があるはずだが，後者に関してはニンテンドーDS のソフトに関するレビューは 18 本と本数が少ないため，その影響は本分析では限定的であると考えられる．

最後に 4.6 では，「ゲーム批評」を対象として得られたカテゴリの普遍性を検証するため，Amazon のユーザレビューを対象として 4.4 と同様の分析を行い，同様のカテゴリが抽出されるかを調べた．

一連の分析の流れを図 6 に示す．

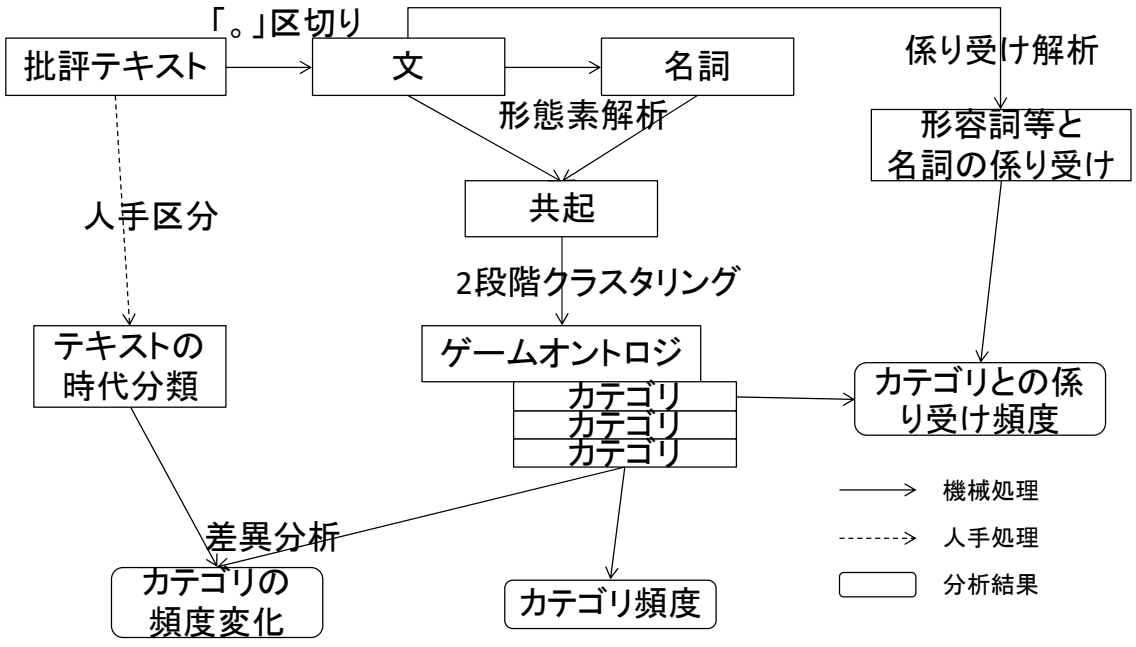

図 6　ゲーム批評に対する分析フローの概要図

4.4. オントロジの自動生成

4.4.1. 目的

本節の目的は，オントロジの自動生成により，ゲーム批評の評価の対象となる要素をカテゴリとして抽出・計量することで，ゲーム批評そしてゲームの面白さにおける中心的な諸概念を明らかにすることである．

4.4.2. 方法

4.4.2.1. オントロジの自動生成手順

4.3 で述べた通り，オントロジの自動生成はクラスタリングによって行う．形態素解析により得られた「ゲーム批評」の語彙 29098 語のうち，出現頻度が 10 以上である一般名詞，固有名詞（全 3077 語，以降 N_{all} とする）を対象とした．頻度 10 以上とすることで，名詞の出現の 84%をカバーできる．これより頻度を下げると，特定のゲーム名など，1 ないしは 2 本程度のレビューにしか出現しない語も多くなってきて（頻度 10 では 18, 9 では 28, 8 では 46，7 では 69…と増えていく），共起のパラメータとして適切では無くなってくる．

パラメータとしては語の共起頻度を選択したが，アルゴリズムについてはウォード法による階層的クラスタリングを採用することとする．非階層的クラスタリングでは原則として分類するクラスターの数を分析者が与えなければならないからである．階層的クラスタリングでも最終的にはクラスター数を指定する必要があるが，上位の構造を把握できるため，分割の弊害を知ることができる．

階層的クラスタリングの計算量は $O(n^2)$であり，全単語を対象としたクラスタリングには時間がかかる．また，最も近い語（群）同士をグルーピングしていく性質から，外れ値の特徴ベクトル（例えば，ほとんどの値が 0 で一部にだけ 1 や 2 が入っているような疎なベクトル）があると，こういった値からグルーピングが始まり，例えば特定のテキストにのみ出現するような語や，ジャンル特有の表現が大きなクラスターとなるような事態が生じてしまう．そのような結果は，目的とするゲーム全般にわたる要素の抽出と合致しないため，本分析では出現頻度を基準として選択した主要な構成語に対して 1 段階めのクラスタリングを行った後，得られたクラスターに残りの単語を分類する 2 段階クラスタリングを行った．これにより，頻度が高い語で充分に密なベクトルを用いたソリッドなクラスターが生成された後，多少精度は下がるが他の単語がいずれかのクラスターに所属し精度と網羅性が確保できる．具体的な手順は以下の通りである．

(1) 1 段階めのクラスタリングのため，N_{all} の中から，出現頻度の高い語から順に，選んだ語の合計出現頻度が，N_{all} の合計出現頻度の 30%以上になるような主要語 N_{top} を抽出した．143 名詞が選出された．
(2) N_{top} をクラスタリングするための特徴量ベクトルを用意した．単語 w の特徴ベクトルの各要素は，w と N_{all} に含まれる各単語との共起頻度である．
(3) 得られた指標を用いて，N_{top} に階層的クラスタリング（ウォード法）を施した．クラス

ター間の距離の計算指標はピアソンの相関係数を利用した．クラスターの分割数は，各クラスターに含まれる単語の数が 1 にならないような最も高い値とし，結果 15 のクラスターに分割された．

(4) 2 段階めのクラスタリングでは残りの単語について N_{top} と同様に共起頻度ベクトルを用意した．そして各クラスターについて，属する単語の共起頻度ベクトルの平均をそのクラスターのベクトルとした．

(5) これらのベクトルの内積を取り，内積が最大となったクラスターにその単語を分類した．このようにして得られたクラスターをオントロジの各カテゴリとした．

(6) 各カテゴリに含まれる単語を基準として筆者がカテゴリの名称を付与した．従って，カテゴリの名称は恣意的である．

4.4.2.2. カテゴリの詳細分析

得られた各カテゴリに属する単語がどのように評価されているのかを調べるため，カテゴリに所属する単語と係り受けする頻度が高い，形容詞（形容動詞語幹名詞を含む）及び全般的なゲーム体験を表すと考えられる動詞（「遊ぶ」，「楽しむ」，「味わう」，「感じる」の 4 種類，以降体験動詞と略す）を，係り受け解析を用いて調べた．

また，各カテゴリがテキスト中にどれだけの頻度で登場するのかを調べるため，各カテゴリに含まれる単語の登場割合をカテゴリ毎に総計した．

4.4.3. 結果

4.4.3.1. オントロジ自動生成の結果

得られたカテゴリをとそれに関する結果を表 11 に示す（なお，「C システム」は「キャラクターシステム」の略称である）．各サイズはその段階でのカテゴリに含まれる語数を示す．カテゴリの内容を示す単語は，各段階で分類された語で，そのクラスターの平均ベクトルとの内積が高い順に上位 10 単語までを記した．

また，クラスタリングによって構築されたクラスターの上位構造をデンドログラムとして図 7 に示す．それぞれのカテゴリは 2～4 個でまとまって上位クラスターを形成することが確認される．

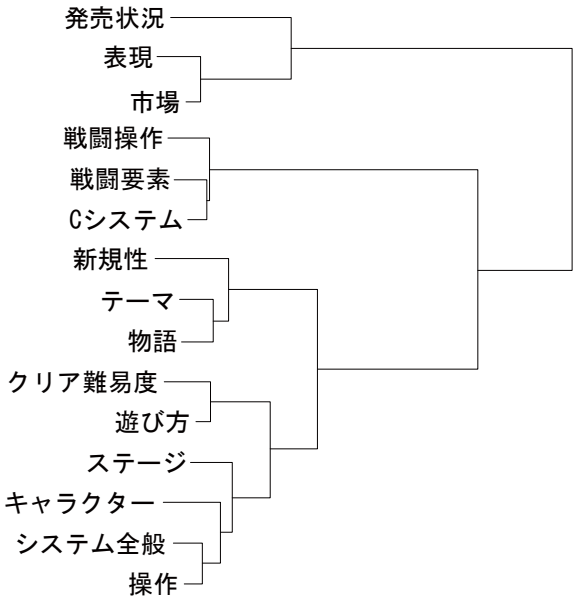

図 7　カテゴリの上位構造

4.4.3.2. カテゴリ詳細分析の結果

表 12 に係り受け解析の結果を示す. 表には各カテゴリと係り受けする形容詞で, TF･IDF（ただし文書数の代わりにカテゴリ数を用いた）が高い順に 3 単語を示してある．また同表の「形」の欄は，カテゴリに所属する各単語あたり平均何回の形容詞と係り受けしているかを示す（形容詞の種類に制限はかけていない）．この値は「新規性」,「物語」,「システム全般」などが非常に高く，よく修飾されている対象であるということが判る．一方「動」の欄は，前述した体験動詞とそのカテゴリの単語あたりの係り受け回数を示す.「新規性」が非常に高いが，その理由の 1 つは【ゲーム】という語が含まれているからであり，【ゲーム】を除いた場合 0.65 に下がる．また，形容詞との係り受け率と体験動詞の係り受け率には強い相関が見られた（r=0.77）

各カテゴリに属する語の出現頻度については図 8 に示す．また，カテゴリに含まれる語彙の平均使用回数（カテゴリの出現頻度÷カテゴリサイズ）を図 9 に示す．出現頻度で「市場」がトップとなるのはそのカテゴリのサイズが 667 と大きいからで, このカテゴリが様々な要素を含んだものであることを示している．しかし次に出現頻度が高い「物語」，そして「新規性」に関しては，カテゴリサイズに依存しているわけではない．これらのカテゴリに含まれる語の平均使用回数は 70 回を超えており，これは全カテゴリでの平均 48 回を大きく上回っている．つまり，この 2 つのカテゴリに関しては，同じ言葉で多く表現されている対象だということが分かる．なお，平均使用回数は「テーマ」が最も低く，多様な言葉で表現されている対象であることが分かる.

表 11　カテゴリ一覧

カテゴリ名	1 段階めで分割された語	サイズ	2 段階めで分類された語	サイズ
発売状況	タイトル, ソフト, 発売, PS2, 期待, ファン, ハード, 続編	8	リリース, 人気, PS, 発表, 良作, 移植, 層, リメイク, 現状, ハード	212
表現	演出, 表現, アニメ, 技術, グラフィック, シーン	6	映像, 手法, ムービー, 再現, ポリゴン, リアル, ビジュアル, 恐怖, CG, 空間	138
市場	印象, 部分, 意味, ユーザ, 要素, 内容, 魅力, 評価, 自体, 作品	30	可能性, 完成度, そのもの, 一般, タイプ, 完成, 本質, メディア, 話題, 感	667
戦闘操作	敵, 移動, 方法, 相手, ボタン	5	通常, 回避, 連続, 防御, 方向, コンボ, タイミング, 戦略性, 手段, 反撃	154
戦闘要素	武器, モンスター, 魔法, 攻撃	4	装備, ダメージ, パターン, 配置, ボス, 出現, 味方, 背後, 剣, 経験値	150
キャラクターシステム	使用, 能力, 動き, 状態, 種類, 個性, 技	7	コマンド, 格闘ゲーム, 特性, 組み合わせ, 駆使, 性能, ヴァリエーション, モーション, 反応, 表情	108
新規性	本作, シリーズ, 前作, RPG, ゲームシステム, ゲーム性, ゲーム, 理解	8	試み, 違和感, 意図, 配慮, 不満, オリジナル, SLG, 既存, 発展, 試行錯誤	190
テーマ	心, 謎, テーマ, 言葉, 舞台, 街	6	冒険, 夢, 町, 架空, 未来, 裏, 完結, 少年, 妖精, 絵本	274
物語	存在, ストーリー, プレイヤ, 他, 展開, システム, シナリオ 目的, 世界, 話	21	進行, ルール, 意識, 自由度, 中心, 想像, スタイル, 反映, 楽しみ, 間	281
クリア難易度	最初, クリア, 気, 感じ, 難易度, 感覚, 筆者, 最後	8	初心者, 攻略, 難度, 開始, 緊張感, モチベーション, 気持ち, かなり, 奥, スタート	204
遊び方	モード, 対戦, 最大, 選手	4	充実, 試合, 対戦, ネット, オンライン, チーム, 練習, ゲームセンター, 多人数, 運営	67
ステージ	マップ, ステージ, ダンジョン, ミッション, 次, フィールド	6	条件, 探索, 場所, 発見, 地形, 一定, 謎解き, トラップ, ボーナス, ルート	90
キャラクター	キャラクター, 主人公, 状況, 設定, 選択, キャラ, 行動, 用意, すべて, 先	11	目標, 判断, 情報, 自身, 幅, 感情移入, 会話, ユニット, 最終, 戦い	266
システム全般	基本, 戦闘, 変化, ポイント, アイテム, レベル, 特徴, 目, 逆, バランス	11	ストレス, 数, 戦略, 採用, 複数, アクション性, 前述, 駆け引き, 機能, 爽快感	167
操作	アクション, アクションゲーム, 画面, 視点, 操作, 3D, ロボット, STG	8	シューティング, 要求, リアルタイム, 方式, 見た目, 格闘, アドベンチャー, パズル, 移行, FPS	109

表 12　カテゴリと係り受けする形容詞

カテゴリ	形容詞	形	動
発売状況	根強い，熱心，ふさわしい	2.26	0.47
表現	美麗な，過激な，忠実な	5.21	0.65
市場	間違いの，ストイックな，正当な	6.25	0.99
戦闘操作	良好な，速い，正確な	2.92	0.14
戦闘要素	強力な，自動的に，必至	2.97	0.06
C システム	速い，豊か，強烈	3.34	0.32
新規性	シンプルな，良質な，おもしろい	7.02	1.75
テーマ	平和な，巨大な，独自の	0.91	0.22
物語	壮大な，ドラマチック，大まかな	6.92	1.39
クリア難易度	率直，易しい，心地よい	3.06	0.84
遊び方	可能な，高い，強い	1.79	0.67
キャラクター	強い，多い，可能	3.68	0.52
システム全般	大幅に，格段に，絶妙に	6.77	0.82
ステージ	安全な，広大な，意外な	2.82	0.27
操作	良好な，見にくい，正確な	4.60	0.70

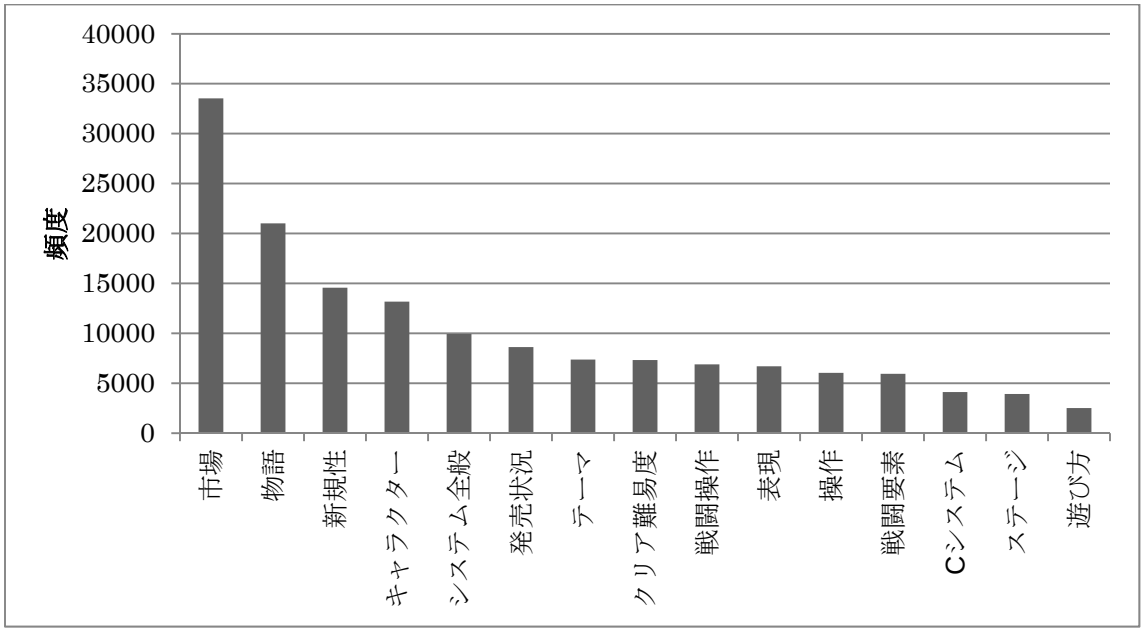

図 8　各カテゴリの出現頻度（降順）

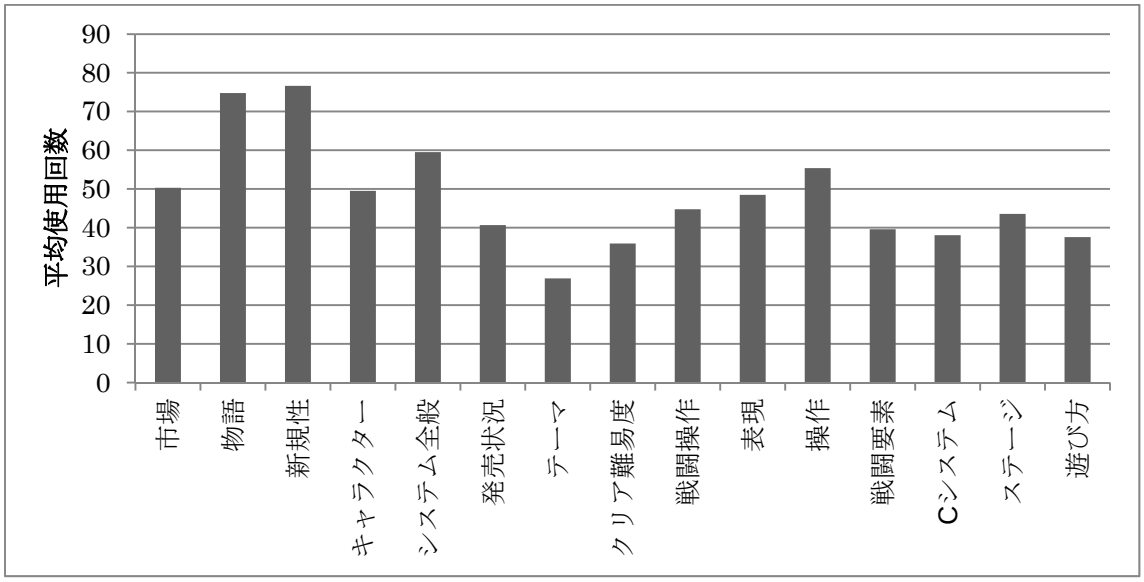

図 9　各カテゴリに含まれる語の平均使用回数

4.4.4.　考察

4.4.4.1. 各カテゴリの解釈

得られたカテゴリについて，それが何を示すカテゴリであるか考察を加える．なお，以降【】で示される語は，各カテゴリに含まれる語を示す．また，「本文」とは，そのカテゴリに含まれる単語を，他のカテゴリと比べて最も多く含む文を指す．

(1)発売状況

表に挙げた単語以外に，各ハードの名前，シリーズ名などが含まれ，そのゲームがどういった状況で発売されたかに関する内容を表していると考えられる．本文に特によく見られるのはシリーズ作品における過去の流れを記述した文である．しかしそれだけではなく，ファンが作品に寄せる期待（【ファン】）なども見受けられる．

(2)表現

ゲームで使われている表現に関する記述．CG やムービーだけではなくて，イベントシーンの演出や音声なども含んでいる（【BGM】,【ボイス】）．本文を見ると，特に演出などが【リアル】【綺麗】であることへの言及が多い．特徴的な形容詞もそれを裏付ける．また画像表示に関連して，座標計算などの技術に対する言及も見受けられる（【処理】,【技術】）．

(3)市場

ゲームのゲーム市場における価値・影響や，そのゲームが出された市場そのものの現状に関する内容だと推測される．以下に本文から特徴的な一文を引用する．“しかし，その試みは高いレベルでまとまり，デジタル・【エンターテイメント】の【可能性】を広く示唆している，できればこうしたヘンテコなソフトが【コンシューマ】の世界の幅を広げると共に，【コンシューマ】【ユーザ】にも【PC】の世界をかいま見てもらいたいと願う．”（ゲーム批評 vol.29 ;【】は本論文筆者による．）

また，デザイナや企業に関する記述や，原作などのゲーム以外の作品に関する記述も含む．どの記述もゲームそのものから一歩引いた視点であることが多い．

(4)戦闘操作

戦闘における操作方法や，操作の種類に関する記述．特にタイミングなどの実際のボタン操作に関する要素となっている（【タイミング】，【コンボ】）．

(5)戦闘要素

「戦闘操作」に対して，システム面の言及を行うときの対象となっている．システムの説明という面が強く，あまり評価を含まない．

(6)キャラクターシステム

キャラクターや敵の性能など，キャラクターのシステマティックな側面についてのカテゴリである（【特性】，【性能】）．レースゲームにおける車の性能などに関する単語も含まれている．

(7)新規性

本文は主にシステム（戦闘システム等を含む）に関する言及があり，それらに対して「おなじみ」「発展」「失望」などの評価が下されている．評価の軸として，【新規性】や【ゲーム性】が用いられている点が特徴的である（【ゲーム性】，【独創】）．ただし，システムに関係なく評価を下す本文も多々見られ，「評価」カテゴリとも言える．

(8)テーマ

そのゲームのテーマやモチーフ，世界観や背景設定，雰囲気などに関する記述．ただし，このカテゴリを含む文が実際には物語を語っていることもあり，物語と完全に分離されたテーマのみを対象とするわけではない．本文にはビジュアルに関する記述も見られる．

(9)物語

ゲームのストーリーに関する記述．ストーリーの説明だけではなく，そこにプレイヤがどう関わっていけるのかというシステム面や（【フラグ】，【自由度】），ストーリーに対して何を感じたかなどを踏まえた評価が含まれる（【感動】，【お約束】）．

(10)クリア難易度

ゲームの難易度に関する説明や，感想の記述（【バランス】，【難易度】）に加えて，ゲームの最終目標である「クリア」に関する記述も含まれる（【クリア】，【攻略】）．また，ゲームをするモチベーションに関する内容もある（【モチベーション】）．ただしゲーム攻略に関する記述だけではなく，操作における感想など他の主観的感想も混じっている．

(11)遊び方

【モード】，【対戦】などが含まれるこのカテゴリは，少なくとも 2 つの部分からなっている．一方は【選手】部分であり，これはスポーツゲーム特有の表現によって生起している．他方は【対戦】部分であり，これはゲームセンターにおける対戦ゲーム特有の表現によって生起している．また他に【ネットワーク】などの語も見受けられ，【モード】とも相まって全体としてはどのような遊び方があるかという主題に関するカテゴリとなっている．

(12)ステージ

プレイヤの操作するキャラクターが動き回る画面である,【マップ】や【フィールド】,【ダンジョン】などの単語だけでなく,【ミッション】や【ステージ】などのフィールドとフィールドをつなぐメタ的構造，仕組みに関する要素も含む（【ルート】,【ロード】）.

(13)キャラクター

前述のキャラクターシステムとは異なり，キャラクターの性格，行動など物語と関連する側面を含んだカテゴリである（【感情移入】,【登場人物】）．また，プレイヤ個人の思い入れに関する記述も含む（【好み】,【思い入れ】）.

(14)システム全般

前述した「新規性」カテゴリと非常に似ており，システムの説明とそれに関する評価が行われている．ただし「システム全般」の方がシステムの説明が多く，また評価の軸としても新規性などではなく，プレイアビリティに関するものが多い（【爽快感】,【ストレス】,【バランス】）.

(15)操作

操作性に関するカテゴリである．プレイヤキャラクターの操作（【操作】,【自機】）に付随して，3D 画面等におけるカメラアングルの問題や（【カメラ】,【視点】），インターフェイスに関する言及（【インターフェース】）も含む．戦闘操作と異なり，移動に関する記述が多い.

4.4.4.2. カテゴリの妥当性と構造

得られた 15 のカテゴリは高確率で内容的に類似した語をグループ化しており，またそれぞれに特徴的な形容詞にも一貫性とカテゴリ毎の違いが見られるため，それぞれが批評における意味的構造を反映したまとまりであると言える.

また，図 10 に点線で示した A～D のカテゴリの上位構造を見ると，4.3 で想定した「状況設定」に関する C の群と,「ルール」に関すると考えられる B, D の群があることからも，オントロジの自動生成によるカテゴリ抽出は成功したと言える.

ただしキャラクターについては，物語のような要素ではなく，操作やシステムと近い位置にあった．これは，ゲームにおけるキャラクターが，物語上のキャラクターというよりも，操作する対象としてのキャラクターとして認識されているということを示していると考えられる．しかし一方では，ルール周りの B と D は離れたクラスターとなっている．これは，戦闘やキャラクターシステムがゲームにおける人工物であり，物語などとの関連が薄いためであると考えられる．戦闘のシステムはルールとして面白ければ，物語や世界設定とアンマッチでも問題ないからである．一方で，キャラクターやステージは，ルール的な内容に寄っているとは言え，物語やテーマと完全に切り離すことはできないと言える.

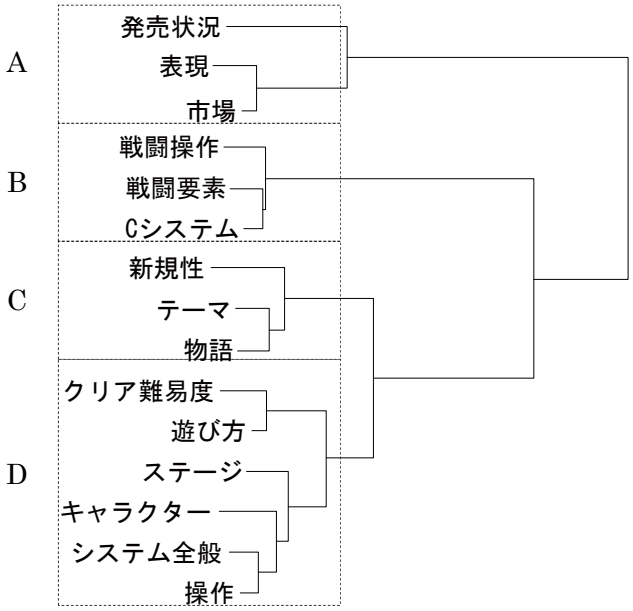

図 10　4 つのカテゴリ上位構造

4.4.4.3. ゲーム批評テキストの特徴

「ゲーム批評」で最も多く登場するカテゴリは「市場」である．具体的には各ハードウェアの状況や，また日本，海外の市場の動向などについて述べているが，これは批評専門誌という立場から来る性質であると考えられる．ゲームそのものに立ち入った内容であることは少なく，ユーザエクスペリエンスから一歩引いた視点での記述である．

また，「新規性」も多い．これは「ゲーム批評」に非常に特徴的な「変わらないことへの批判」から来ていると考えられる．特にシステムに関して批評する時に，「問題はない」が，「目新しくない」といった記述が多く見受けられた．これらが示すのは，「ゲーム批評」はゲーム作品の停滞に強い危機感を持っており，たとえ面白く遊べるものであっても，新規性を要求するような批評が多いということである．そのような傾向は，特にシリーズ作品において顕著なようである．そしてその新規性は単一のカテゴリとして存在し，実際のプレイに関連する要素とは多少切り離して語られているということも，カテゴリの上位構造から推察できる．また「新規性」に含まれる語の平均使用回数が多いことから考えられるのは，新規性に関しては一定の批評しか行えず，新しい方向性の提案などは難しいということである．

4.4.4.4. 面白さとの関係

ゲームの面白さを構成する要因の中では，当然のことながらゲームプレイによる直接的なユーザエクスペリエンスが主要な位置を占めるはずである．しかし，ユーザエクスペリエンス以外にゲームの面白さを構成する要因がないとは断定できない．

本分析で対象としたような批評テキストデータから，ゲームの面白さを構成する要因を計量的に抽出する 1 つの指標として，感性語を分析することが考えられる．一般的には，対象によって強く影響を受け，感情を喚起させられるほど，文章に感性語が増えると考えられるからである．感性語として代表的な形容詞を見てみると，例えば，「新規性」のカテゴリを形容する語は表 12 に含まれる語以外にも「駄目」「珍しい」「斬新」など，感性的な形容詞が特徴的に用いられている．また，形容詞が用いられる頻度も高い．一方，形容詞が係り受けする頻度が少ない「戦闘要素」で用いられる形容詞は，表の語以外に「有効」「有利」「多彩」などがある．「多彩」を除くと，他の形容詞は全般的に非感性的であり，システムの説明的な評価語である．このため，批評文の筆者による評価を伴ってはいるものの，面白さのような対象によって喚起された感情を強くあらわす傾向は小さいと言える．この形容詞の係り受け頻度と感性語の使用傾向はほぼ全てのカテゴリにあてはまり，「新規性」「物語」「システム全般」「市場」「操作」「表現」など，形容詞の係り受け頻度が高いカテゴリは感性的側面が非常に強く，一方頻度が低い「戦闘操作」「戦闘要素」「テーマ」「ステージ」「遊び方」などはどちらかといえば評価的で，ゲームによって喚起された感情に基づく感性的な側面が弱いと考えられる．さらに，ユーザエクスペリエンスを表すために使われると考えられる体験動詞との係り受け頻度と，形容詞の係り受け頻度との間に相関があるという結果も得られた．このことからも，「新規性」などのカテゴリは，評者がゲームの面白さを実際に感じながら，感性的な語を用いて記述したと考えられる．

ところが，この感性的要素を多く含むカテゴリは，必ずしも従来注目されてきたユーザエクスペリエンスの要素とは一致しない．「物語」や「システム全般」，「操作」などは従来通りであるものの，「新規性」や「市場」は従来注目されてこなかった対象である．これらのカテゴリは，プレイ時の体験をそのまま記述した内容だけでなく，背景知識などを合わせた記述だと考えられ，プレイ時ではなくプレイ後に特有の特徴だとも考えられるが，ゲーム体験を測定する上で決して無視できる要素ではない．ユーザエクスペリエンスと「新規性」や「市場」などのカテゴリがどのように相互に影響し合っているのかは明らかではない．しかし，感性語が頻出するカテゴリは感性的な評価の主要な対象であり，これらの総体がゲーム自体の評価をも構成するので，感性語評価語が多用されるカテゴリは「ゲーム」の面白さにとって重要な要素であると言える．

4.5. カテゴリの時系列分析

4.5.1. 目的

本節の目的は，各カテゴリの出現量の変化を見ることで，ゲーム批評における観点の変化を明らかにすることである．なお，出現量の変化は，どのゲームジャンルの記事が多いか，どのジャンルが多く発売されたのかなどに影響されると考えられる．本分析では，これらの影響を除くのではなく，ジャンルの比率なども時代性に含まれると捉えて分析を行う．

4.5.2. 方法

4.4 で得られた各カテゴリの出現量を 4 つの年代別にクロス集計する．年代はハードウェアの発売を基準として筆者が作成した（表 13）．

年代毎の違いを統計的に示すため，集計結果に対してχ^2検定の残差分析を行った．

表 13　年代別グループ

グループ	批評数	特徴
時代 1：94 年～96 年	83	ゲーム批評の創刊から，NINTENDO64(以下 N64)が発売されるまでの 3 年間である．96 年の記事には N64 の批評は存在しないため，スーパーファミコン，セガサターン，PS の批評が中心となる．
時代 2：97 年～99 年	255	97 年から N64 の記事が出てくるようになる．98 年にはドリームキャスト（以下 DC）が発売され，PS，N64，DC の批評が中心となる．
時代 3：00 年～03 年	324	2000 年にプレイステーション 2（以下 PS2），2001 年にゲームキューブ，Xbox が相次いで発売される．これらの批評を中心とするのがこの時代である．
時代 4：04 年～06 年	222	2004 年にニンテンドーDS とプレイステーションポータブルが発売される．さらに，2005 年には Xbox 360 が発売．これらの最新機種と，PS2 の批評が中心となる．

4.5.3. 結果

各カテゴリの量（全体に対する割合）の時代変化を 11，図 12 に示す．各グラフ共，時代 1 における割合を基準として相対化してある．

χ^2検定の結果，各カテゴリの時代による偏りは有意であった（$p<.01$）．残差分析の結果を表 14 にまとめた．表中，▲は有意に多い項目，▽は有意に少ない項目を意味する．

減少傾向が見られるのは「表現」，「物語」，「C システム」，「戦闘操作」である．一方増加傾向が見られるのは，「クリア難易度」，「遊び方」，「発売状況」である．

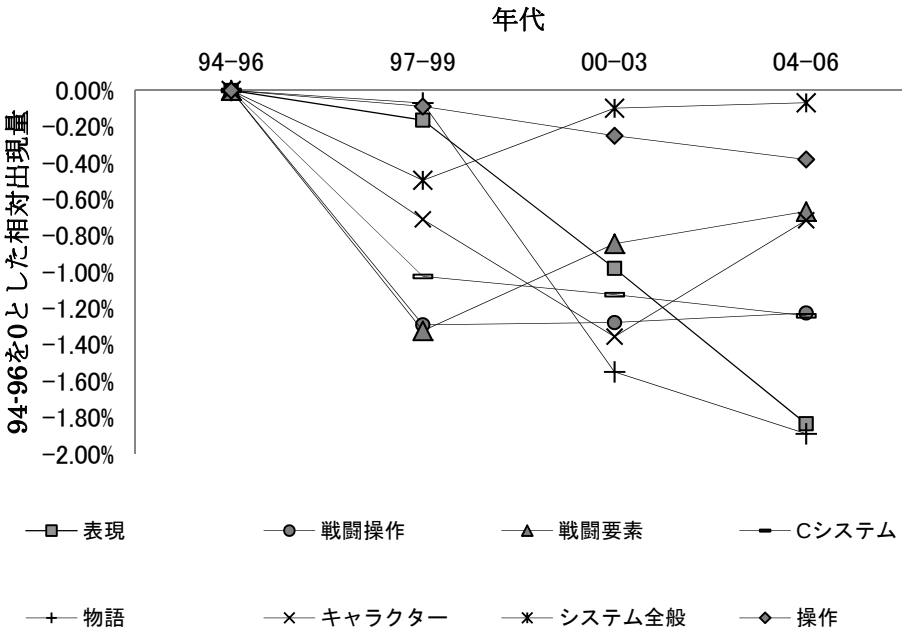

図 11　出現量の相対変化（減少傾向にあるカテゴリ）

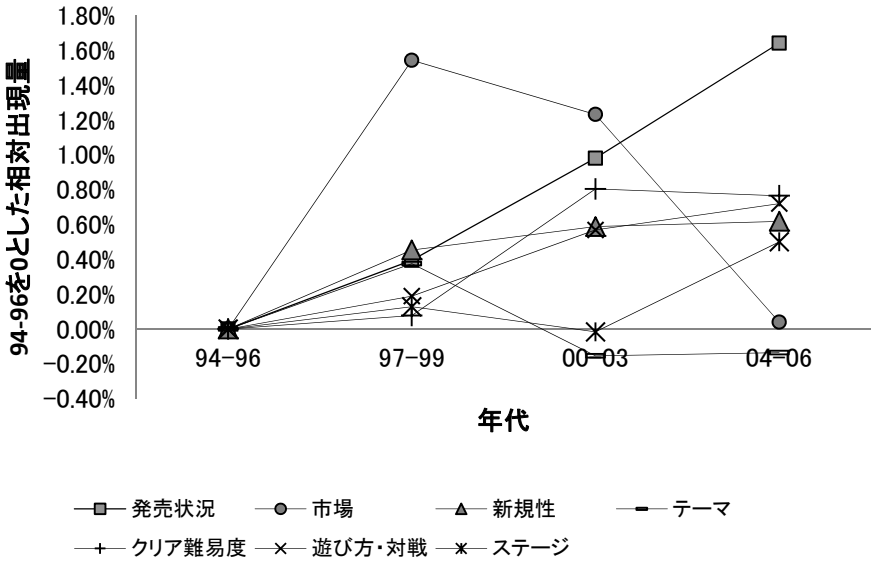

図 12　出現量の相対変化（増加傾向にあるカテゴリ）

表 14　時代毎のカテゴリ出現頻度のχ^2検定結果（p<.01）

時代	表現	操作	新規性	戦闘要素	発売状況
94-96	876▲	667	1401▽	773▲	703▽
97-99	1895▲	1459	3348	1150▽	1754▽
00-03	2427	2196	5393	2156	3187
04-06	1491▽	1717	4418▲	1861▲	2979▲
時代	**クリア難易度**	**システム全般**	**遊び方**	**テーマ**	**市場**
94-96	648▽	1058	167▽	762	3295▽
97-99	1490▽	2157▽	459▽	1878▲	8085▲
00-03	2868▲	3692	993▲	2601▽	12577▲
04-06	2318▲	3030▲	897▲	2133	9581▽
時代	**戦闘操作**	**キャラクター**	**物語**	**C システム**	**ステージ**
94-96	934▲	1530▲	2392▲	625▲	370▽
97-99	1526▽	3122	5342▲	950	888
00-03	2424▽	4488▽	7416▽	1433▽	1304▽
04-06	2008	4029▲	5857▽	1103▽	1360▲

4.5.4. 考察

「表現」は減少傾向にある．時代 1，時代 2 は 32 ビット機の登場によりグラフィックスが美しくなったことで，それが積極的な評価の対象となったと考えられる．しかし，徐々にグラフィックスが美麗であることが当たり前になってくるにつれて減少したと考えられる．もはやグラフィックスが綺麗なだけでは評価されないのである．

一方上昇傾向にある「発売状況」であるが，これはシリーズ作品への言及を含むからであるからと考えられる．例えば，“『ラプソディア』は，『幻想水滸伝Ⅳ』の事実上の＜続編＞である．”（ゲーム批評 vol.66）などの文がこれに当てはまる．取り上げられているシリーズの続編は年々増加している．特に，PS の初期，PS2 の初期に発売されたタイトルの続編が多く，PS2 が発売されて 3 年経った 04 年頃の「発売状況」が有意に多いのはその理由によるものである．

「遊び方」が増加しているのは，ここに含まれる「ネットワーク」によって説明できる．この語は時代 1 には全テキストを通じて 1 回しか登場しないが，時代 2 には 20 回，時代 3 には 66 回登場している．「オンライン」という単語も同様の特徴を持っている．オンラインゲームの登場により遊び方の多様性が広がったことが結果に現われていると考えられる．

「クリア難易度」であるが，このカテゴリと係り受けする形容詞を時代ごとに見ていくと，時代 1 と 2 では係り受けがそもそも非常に少ないことが分かる．多く見られるのは必要（【調整】）や難しい（【クリア】）程度であるのに対し，時代 4 になると，大きい（【快感】，

【苦労】)，簡単（【クリア】)，心地よい（【緊張感】,【感覚】）など多彩な形容詞が登場する．さらにこれらの形容詞が係っている単語として数多く見られるのが，「感覚」という単語である．ここから推察されるのは，ユーザの要求がより感覚を重視したものになってきているのではないか，ということである．この「感覚」の重視も，ハードウェアの進歩に伴って，ゲームがよりリアルになってきたことと深い関係があると考えられる．さらに「表現」の減少は，このような「感覚」の増加に伴い，客観的に「表現」を見るのではなく，より主観的に体験する方向へゲームが進化していることを示していると考えられる．

4.6. ユーザレビューとの比較

4.6.1. 目的

4.4 で抽出されたカテゴリは，ゲームというメディアに対する認知をどれだけ本質的に捉えているのか．本節ではそれについて考察するため，ユーザレビューを対象として同様の分析を行い結果を比較した．

4.6.2. データ

「ゲーム批評」で対象となっていたゲームについて Amazon.co.jp のユーザレビューを採取した（2009 年 8 月 12 日実施）．4780 レビューが取得され，合計 50391 文，881796 単語となった．なお，「ゲーム批評」と異なり，ゲームタイトル毎にユニークレビュー数は異なる．

4.6.3. 方法

4.4.2 と同様の方法でオントロジを自動生成した．その後，4.4.3 のカテゴリと語の重複率を調べ，同様のカテゴリが得られたのか確認した．

4.6.4. 結果

得られたカテゴリに関する結果を表 15 に示す．ユーザレビューで特徴的なカテゴリとしては，他のユーザにゲームを推奨する時に使う語を多く含む「おすすめ」,【感覚】や【感じ】などの非具体的な感想を述べる際に利用される「大まかな感想」，発売前に書かれた，新作への期待などを表す「ソフト（発売前)」がある．また，クラスタリングによって構築されたカテゴリの上位構造をデンドログラムとして図 13 に示す．

表 15　ユーザレビューから生成されたカテゴリの一覧

カテゴリ名	1段階めで分割された語	サイズ	2段階めで分類された語	サイズ
戦闘操作	爽快感，攻撃，動き	2	数，雑魚，ロックオン，ボス，味方，ストレス，ザコ，快感	219
ストーリー	話，世界，物語，キャラ，主人公，感動，キャラクター	7	展開，心，存在，人間，設定，メイン，関係，謎	229
操作	要素，操作性，レベル，アクション，画面，システム，操作，リアル	8	基本，操作，特徴，バランス，謎解き，目，バトル，不満	190
おすすめ	プレイ，ソフト，RPG，作品，星，オススメ，初心者，ファン	8	損，満足，興味，名作，作り，アクションゲーム，おすすめ，経験	196
市場	続編，期待，PS2，発売	4	PS，プレステ，新作，PSP，人気，進化，スーパーファミコン，評判	137
表現･雰囲気	雰囲気，グラフィック，音楽，個人的，魅力，世界観，演出	7	BGM，絵，表現，映像，アニメ，文句，イマイチ，ドラマ	129
大まかな感想	内容，人，プレイヤー，意味，感じ，評価，部分，他	8	自体，感覚，逆，印象，でき，体験，完成度，ホラー	519
キャラクター	仲間，今作，登場	3	個性，オリジナル，魅力的，愛着，同士，思い入れ，感情移入，性格	173
ソフト（発売前）	前作，ゲーム，シリーズ，是非	4	感想，SFC，スパロボ，プレイ時間，安心，環境，体験版，シルバー事件	131
クリア難易度	クリア，気，手，最初，難易度，最後	6	先，かなり，飽き，次，攻略，クリア，気分，攻略本	135
ステージ	ダンジョン，ステージ，モード	3	マップ，ミッション，条件，おまけ，ランク，ストーリーモード，攻略，達成感	81
戦闘要素	戦闘，武器，アイテム，敵	3	序盤，魔法，行動，コンボ，種類，場所，移動，必殺技	228
ストーリー演出	ストーリー，ムービー，エンディング，一つ，シナリオ，イベント	6	あと，台詞，セリフ，声，場面，データ，本編，表情	71
通常演出	3D，シーン	2	ポリゴン，迫力，ロード時間，ボイス，立体，カメラワーク，カット，横	32

図 13　ユーザレビューから生成されたカテゴリの上位構造

表 16 は，ゲーム批評のカテゴリ（横軸）とユーザレビューのカテゴリ（縦軸）において，共通する語の数を集計したものである．濃灰のセルは最大共通数を，薄灰のセルは次点の共通数を示している．またχ^2検定の残差分析の結果，有意に多い項目を▲，有意に少ない項目を▽としている（$p<.05$）．

ゲーム批評の観点から見ると，全 15 カテゴリのうち，最も共通語が多いカテゴリがほぼ同様の意味合いのカテゴリであるような例が 3 件あった（表中，「関係（最多数）」の行で○になっているカテゴリ）．また，ゲーム批評カテゴリのデンドログラムにおいて近いカテゴリ（2 階層上がることで同じクラスターとなるようなカテゴリ）で見ると，5 カテゴリが類似していると言える（同△．() 内は，いくつ上位のクラスターで見れば良いかを示す）．例えば，ゲーム批評における「テーマ」は一階層上がると「テーマー物語」のカテゴリとなる．ゲーム批評における「テーマ」と，ユーザレビューによる物語が最も共通語が多い（ユーザレビューにはテーマに相当するカテゴリは現れなかった）．

さらに，次点で共通語が多いカテゴリを見ると，○が 2 カテゴリ，△が 1 カテゴリ増え，15 カテゴリ中 11 カテゴリは類似のカテゴリと語を多く共通する，と言える．また，いずれも説明に寄与しているセル（最大，次点共）は，全て有意に多いという結果となっている．他にも絶対数が多くはないが検定的には有意と言える数のセルは，例えば「ユーザレビュー：戦闘操作／ゲーム批評：戦闘操作」のセルや「ユーザレビュー：ストーリー演出／ゲ

ーム批評：物語」等，カテゴリがマッチしているものがほとんどである．逆に，絶対数は多かったがカテゴリがマッチしていなかったセル，例えば「ユーザレビュー：大まかな感想／ゲーム批評：クリア難易度」は，有意に少ないという結果だった．

表 16　ゲーム批評カテゴリとユーザレビューカテゴリ間の共通語数

		表現	新規性	操作	テーマ	市場	キャラクタ	物語	システム	ステージ	発売状況	遊び方・対戦	システム全	クリア難易	戦闘要素	戦闘操作
	サイズ	138	190	109	274	667	266	281	108	90	212	67	167	204	150	154
戦闘操作	219	3 ▽	2 ▽	9	5 ▽	5▽	22	8 ▽	11 ▲	9	0 ▽	1	14	13	44 ▲	28 ▲
物語	229	7	11	4	60 ▲	21 ▽	29 ▲	40 ▲	1 ▽	1 ▽	2 ▽	1	5 ▽	4 ▽	2 ▽	1 ▽
操作	190	14 ▲	7	17 ▲	3 ▽	21 ▽	17	7 ▽	6	7	2 ▽	4	20 ▲	18 ▲	2 ▽	11
おすすめ	196	6	11	6	9	50 ▲	7	6 ▽	3	1	13	7	3 ▽	12	0 ▽	1 ▽
市場	137	1	7	3	4	27 ▲	1 ▽	2 ▽	1	0	45 ▲	1	2	2	0 ▽	0 ▽
表現・雰囲気	129	26 ▲	8	3	9	27 ▲	5	8	2	0	3	0	3	1 ▽	0 ▽	1
大まかな感想	519	12 ▽	28	7 ▽	30	161 ▲	26 ▽	53 ▲	5 ▽	5 ▽	29	17 ▲	13 ▽	22	5 ▽	7 ▽
キャラクタ	173	4	13 ▲	5	13	13 ▽	16	17	7	2	3	3	10	7	4	5
ソフト（発売前）	131	0 ▽	10 ▲	2	5	19	2 ▽	3	2	0	18 ▲	4	2	6	1	0 ▽
クリア難易度	135	1	5	1	7	11 ▽	19 ▲	15	4	9 ▲	0 ▽	4	5	21 ▲	2	1
ステージ	81	1	0	0	1	5▽	5	3	0	16 ▲	1	5 ▲	6	10 ▲	2	1
戦闘要素	228	2 ▽	9	6	5 ▽	3▽	24	8 ▽	12 ▲	7	1 ▽	2	34 ▲	8	36 ▲	29 ▲
ストーリー演出	71	8 ▲	3	0	2	8	5	15 ▲	1	1	1	0	2	4	1	3
通常演出	32	5 ▲	0	4 ▲	0	2	0	1	0	1	0	0	1	1	2	3 ▲
関係（最多数）		○	×	○	△ (1)	×	× (4)	×	△ (1)	○	△ (2)	×	× (6)	×	△ (2)	△ (2)
関係（次点）			× (4)			×	×	○				×	△ (1)	○		

4.6.5. 考察

インターネット上のユーザレビューは専門家が書いたものではなく，またその 1 つ 1 つはそれほどの長さを持たないため，大きく性質の異なるテキストであると言える．抽出されたカテゴリも異なるが，分析の結果，語彙的に共通する類似のカテゴリがあるということが判ったと言える．とはいえ，共通する語は 20%程度であり，同様のカテゴリが得られた，とは言えない．また，ユーザレビューにおいては，「大まかな感想」という巨大（519 語）で曖昧なカテゴリがあり，このカテゴリが多くの語を吸収しているという特徴がある．ユーザレビューでは短文で単純な，意味を抽出することが難しい感想のレビューも多くあり，そのようなレビューに結果が引きずられた可能性もある．ゲーム批評において語られる要素にはプロ・アマを問わない本質的な内容があると考えられるが，差異も大きく，その本質を特定することは難しい，という解釈が現状のデータと分析による限界であると考えられる．

では，そのユーザレビューとプロの批評家による批評の差異とは何か．1 つには，ユーザレビューには「おすすめ」「大まかな感想」というカテゴリがあることである．これらのカテゴリは，例えば“是非　【プレイ】しておきたい良作の一本だろう”や，“一度クリアしてもまたプレイしたくなって，そのたびにハマルような【感じ】です.”等の使われ方をする語を含み，先述した通り単純な「感想」を根拠なく述べているような記述が見受けられる．一方で，ゲーム批評では「物語」と「テーマ」に分かれていたカテゴリが「物語」に，「発売状況」と「市場」だったカテゴリが「市場」に結合していることが見て取れる．これらのカテゴリはゲーム批評においても近いカテゴリではあるため，クラスタリングにおける分割単位の問題という可能性はあるが，全体として，ユーザレビューのほうが対象について曖昧で，感覚的・感想的な内容を多く含むという傾向がある．ゲームという対象について理解するという観点では，論ずるべきものが見えているプロの批評家のテキストを分析する方が目的に適う．

4.7. 分析の結論

本章ではゲームの批評テキストに対してオントロジの自動生成を行い，内容，特徴に一貫性のある 15 のカテゴリを得た．カテゴリの上位構造を見ると，既存研究でも指摘されていたような「状況設定」や「ルール」に関する要素が含まれていることから，得られたカテゴリには妥当性があると言える．「ゲーム批評」テキストは，高頻度の語だけで分析すると，【ゲーム】や【プレイヤー】などのあまりに一般的で，内容を捉えがたい語を見ることになってしまう．テキスト中の情報から生成されたカテゴリを単位として集計することで，単語を単位とするより明確で深い分析が可能になったと言える．以上のことから，本分析で採用したオントロジの自動生成は，分析において有効であったと言える．

計量分析による新規の発見としては，従来のユーザエクスペリエンスモデルで注目されてきたシステム，操作，物語等だけでなく，「市場」や「新規性」などの要素が批評には含まれており，かつそれらの記述が体験から受ける影響が強いことがある．ゲームの価値が

プレイによる一次的な直接体験のみでなく，プレイ経験を振り返っての二次的な反芻や評価，過去の他作品のプレイ経験との比較からも生じることを考えると，従来のユーザエクスペリエンスに合わせてこれらの要素をゲームの評価対象として含める必要があると考えられる．ユーザエクスペリエンスがインターネットにおけるレビューなどで他のユーザに伝播していく現代のネット社会においては，テキスト化した時に現れてくる要素が他者にとってのユーザエクスペリエンスに大きな影響を与えうるという点でも重要であり，直接測定したユーザエクスペリエンスとの関係などを明らかにしていく必要があると言える．

また，計量によって従来取り組まれてこなかったゲーム評価の時系列の分析を行い，「表現」や「遊び方」,「難易度」について変化が起きてきたことを計量的に示すことができた．時系列や分野毎の変化，差異の分析には計量化し統計的検定をかける手法が有効であることが示された．

ユーザレビューとの比較の結果，媒体・書き手によって概念構造に違いがあることも判った．雑誌「ゲーム批評」から得られたカテゴリが普遍的であるとは簡単には言えない結果となったが，他媒体の比較で共通語において統計的な有意性があった事を鑑みても，ゲーム批評から得られたカテゴリが対象テキスト特有のもの，ないしはランダムなものであったとは言い難い．そのような観点から考えると，対象となるテキストを増やし，それぞれのテキストの特性を踏まえて分析することで，より普遍的な概念の抽出に近づくことができると考えられる．他の雑誌や Web 上のレビュー，プレイ時のプロトコルなど他の媒体にも同様の手法を適用することにより，さらに比較・検討していく必要がある．

5. オントロジの手動構築と概念構造抽出－河川文化における大域的概念構造の抽出－

5.1. 本章の目的

本章では，土木，環境，歴史文化，暮らしなどの様々な要素が，河川というキーワードを中心として散逸的に存在するような分野である河川文化のテキストを扱う．このようなテキストでは，オントロジの手動構築による分析が有効であることについて述べる．また，文などテキストの微細な構造ではなく，まとまったテキストの単位で関連性を抽出することで，河川文化のような曖昧な対象からも全体構造を抽出できることについても説明する[98].

5.2. 分析の背景と目的

1997 年の河川法改正によって，我が国における河川整備あるいは河川管理の目的に治水・利水のみならず良好な環境の保全・創出が加えられ，国土交通省や地方自治体は，環境配慮型河川整備に積極的に取り組むようになった．また，2006 年に国土交通省が全ての川づくりの基本として示した多自然川づくり基本指針には，「河川が本来有している生物の生息・生育・繁殖環境および多様な河川景観を保全・創出」するとともに，「地域の暮らしや歴史・文化との調和にも配慮」した河川管理を行うことをうたっている．そこで，自然科学的諸条件のみならず河川を取りまく様々な文化的要素をふまえた川づくりの推進が求められるようになった．つまり，「河川文化」が川づくりにおける重要なキーワードになったのである．

我が国において河川は，古来人びとの生活の軸として様々な機能や役割を有していた．近代以降における利水・治水や生態学的見地のみならず，河川を取りまく文化的要素は多岐の分野にわたっており，河川文化は複雑で曖昧な概念である．したがって河川文化をふまえて川づくりを展開していくためにはまず，河川文化を構成する要素を明らかにしなければならない．さらに，具体的な活動を展開するなかで有効であると考えられるのは，川づくりを実施する主体が，自らの活動が河川文化のどの範囲をカバーしているのか，あるいはどのような要素を新たに加えればより網羅的に河川文化を反映できるのかということを把握することである．そのためには，河川文化の概念を構造化し，かつ人びとが直感的に理解できるような概念体系を構築する必要がある．

そのような河川をとりまく多様な要素の体系化は，実際の川づくりのプロセスでも重要なニーズとなっている．本章の共同研究者である桑子，高田は，新潟県佐渡市の天王川自然再生事業に，合意形成マネジメントチームとして携わっている[99]．その事業で重要な課題であったのが，天王川の下流に位置する加茂湖という汽水湖との関係であった．加茂湖はカキやアサリの養殖が盛んであり，多くの人びとが漁業によって生計を立てている．加茂湖の漁業者は，天王川で工事を実施することによって，下流に位置する加茂湖の水産資源に悪影響を及ぼすことを強く危惧していた．つまり，河川における生態系保全と湖沼に

おける水産資源保護が対立的に捉えられていたのである．したがって，天王川再生を地域との合意のもとに進めていくためには，河川の自然再生と漁業の活性化をどのように関連付けていくかが大きな課題であった．このような具体的な合意形成マネジメントに携わる立場として，河川整備事業をとりまく様々な要素の関係性を把握することは，きわめて重要な意味合いを持つ．なぜなら，実際の合意形成の場面においては，ステークホルダーが事業の具体的な課題とその解決の方策を共有することが求められるからである．天王川の例では具体的に，「川づくり」と「漁業」，あるいは「生態系」といった要素がどのようにかかわっているか，その理解を共有し事業の方向性を検討しなければならない．そのような場面において，河川文化概念の諸要素が構造的に明示されていれば，合意形成マネジメントを実施する者は，ステークホルダーとともに，課題解決のためのいくつかの有用なヒントを得ることができる．たとえば，「川づくり」と「漁業」という要素が乖離しているのであれば，その間をつなぐ要素にはどのようなものがあるかということを把握することで，創造的なブレイクスルーを見いだすことが可能となる．つまり，河川文化概念の構造化は，河川整備事業の合意形成プロセスにも大きく貢献すると考えられる．

河川文化にかかわる既往研究は，大きく二種類に分類できる．1 つは，個別事例におけるケーススタディや調査研究である．たとえば，林ら[100]は，京都の鴨川を対象に，明治・大正期の料理屋・貸座敷営業者による河岸地と堤外地の土地利用のしくみについて研究している．中嶋らの研究[101]は，郡上八幡における地域コミュニティと水辺空間の関係について調査を行なったものである．また，竹林[102]は，富士川の歴史的変遷を詳細に調査し，富士川を軸とした様々な工学的・文化的要素を紹介している．富山の研究[103]は，淀川，利根川，木曽川，筑後川のそれぞれの流域における文化史に焦点をあてたものである．

2 つめは，ある特定の分野・観点から河川にかかわる文化的事象について考察しているものである．中村[104]は，景観工学の観点から，生態系と地域文化の一体的再興を通じた水辺空間形成の必要性を説いている．高橋[105]は，琵琶湖疏水や信濃川などの例をあげながら，土木技術と文化の関係について言及している．また，富野[106]も，土木工学の観点から，日本における様々な伝統的河川工法を紹介している．大熊[107]は，洪水の歴史に着目し，治水と地域文化のなかで形成されてきた治水の技術と思想について考察した．

以上のように，個別事例における研究，あるいは特定の分野からの文化に関する考察は実施されているものの，工学・社会・環境・教育・歴史・文学・芸術など多岐の分野にまたがる河川文化を網羅的に取り扱ったものはない．また，広範な領域にまたがる河川文化の概念が体系的に示されていないということは，たとえば第三者がある事例から何かしらのヒントを得ようとした時，それが自分のプロジェクトのなかのどのような問題に対して適応されるべきかわかりにくく，継承されにくいという問題点も存在する．これらの問題を解決するためには，河川にまつわる様々な要素の抽出，分類，関係性等の調査と，対応する事例を検索するためのしくみが必要となる．

河川文化は領域融合的でかつステークホルダー間の価値観の相違の問題を含んでいるた

め，特定の問題のみを扱う既存の方法では全体像は見えてこない．そこで，本分析では河川文化に関する多様かつ大規模なテキストデータに注目する．多様と言っても，河川を中心とした内容である限り，必ず他のテキストと何らかの概念を共有しているはずである．そこで，本分析ではこのような共有概念を抽出し，概念同士の関係を単一のテキストを越えて構造化することで，複雑で捉えがたい河川文化の全体像に迫ることを目的とする．その上で，得られた全体像が河川に関する実際の活動へ適用可能であるかについて検討することとしたい．

5.3. 対象データと手法の選択

分析するべきデータとしては，河川とその文化にかかわる様々な概念について言及した文献や講演記録などのテキストが考えられる．様々な方向から語られる河川に関する言説を分析することで，河川文化概念の領域とその諸要素間の関係性が明らかになる可能性があるからである．そこで本分析では，財団法人日本河川協会が主催する「河川文化を語る会」の講演集「河川文化」その 1〜31（1995-2010）[108]に含まれる全 132 の講演会の書き起こしを分析対象として用いた．選定理由は以下の 3 点である．

(1) この講演会では，河川文化を一貫したテーマとしながらも，様々なバックボーンをもった講演者が，自由に，バラエティに富んだ講演を実施していること．したがって，既成の枠組みにとらわれることなく，河川文化を構成する要素について考察することが可能となる．

(2) 本講演集が河川文化を一貫したキーワードとしている点．「河川」や「河川環境」をテーマにした文献は存在するものの，河川文化を直接のテーマにした文献は他にみられない．

(3) 同様のフォーマットでまとまった量のテキストデータが蓄積されている点．

「河川文化」が河川文化の全てを包括しているわけではなく，講演者の偏りもあるという欠点が考えられるが，そもそも河川文化の概念が曖昧である以上，網羅的にデータを収集することは難しい．まずはまとまった形のある「河川文化」を対象とすることで概念を明確化するための適当な一歩を踏み出すことが可能となる．

PDF 形式のテキストから図表，キャプション，ルビを取り除き，その後形態素解析を行った．全 84861 文（「。」区切り），のべ 2323961 語となった．なお，分析の対象となる名詞の中でも，講演という形態及び文中での図表への言及に由来する語（【自分】，【話】，【お話】，【皆さん】，【写真】，【図】，【拍手】，【先生】，【司会】，【質問】）及び抽象概念などの非具体的な語（【関係】，【意味】，【先】，【下】，【間】，【部分】，【１つ】，【形】，【辺】，【例】，【次】，【最後】）で頻度 500 回以上のものは以降の分析から除外した．頻度 500 未満でも前述のような語はあるが，他の語も多様になってきて見つけづらくなり，恣意性が増すため 500 で切った．頻度 500 以上では，明らかに河川に関係ある語が多く，関係ない語は目立つのである．

さらに，132 の講演をまとまりとして扱うため，講演者の肩書きと内容によって講演を 4

つの大きなまとまり（講演カテゴリ）に分類した．分類は，講演の内容および講演者の専門を基準として人手で行った．分類の結果を表 17 に示す．

表 17　河川文化における講演のカテゴリと講演数

講演カテゴリ	講演数
環境・生態	44
社会・暮らし	20
土木・空間	30
文化・歴史	38

河川文化に関する概念も，ゲーム批評と同様に名詞としてテキスト中に存在していると考えられる．表 18 に，テキスト中の高頻度名詞上位 30 語を示す．上位 30 語には，【水】や【川】，【環境】，【海】などの語が含まれており，河川に関係することは判るが，「ゲーム批評」と比較すると抽象的，ないしは多義的な語が多く，具体的に何について語っているのか判然としない．

表 18　出現頻度上位 30 位までの名詞の出現頻度と出現テキスト数

単語	出現頻度	出現テキスト数	単語	出現頻度	出現テキスト数
水	5835	128	調査	863	95
川	5383	125	東京	855	101
人	3765	132	生物	792	61
日本	3152	132	魚	790	79
河川	1835	123	洪水	762	79
研究	1535	114	技術	755	94
環境	1331	100	木	734	92
海	1211	109	利用	687	104
時代	1174	126	橋	680	60
人間	1161	118	雨	679	79
世界	960	117	ダム	674	75
地域	946	111	状況	651	106
山	939	105	文化	648	97
場所	880	124	生活	635	113
子供	864	83	紹介	635	122

文脈を踏まえた単語の意味を探るため，単語の共起ネットワークを作成した（図 14）．頻度 223 回以上（出現頻度で名詞の 30%をカバー）の名詞における文単位の共起について頻

度 77 回以上のものを描画し（75 ノード，103 エッジ，密度 0.0371，平均エッジ数 5.49），描画アルゴリズムはノードが重ならないことを最優先としている．共起頻度の閾値については，「ゲーム批評」で 30 回としたことを基準に，30×[河川文化の文数]÷[ゲーム批評の文数]で計算した値である．【川】，【水】，【日本】の 3 つのノードに他のノードが集中して接続するようなネットワーク構造となっていることが見て取れる．実際これらの単語のみ次数中心性が二桁（それぞれ 36，35，13）であり，他は一桁という値となっている．【日本】や【研究】を中心としたコミュニティ構造はあるが，この図から河川文化に関係する概念同士の関係を読み取ることは困難であり，このネットワーク図をもって河川文化の全体像とは言い難い．

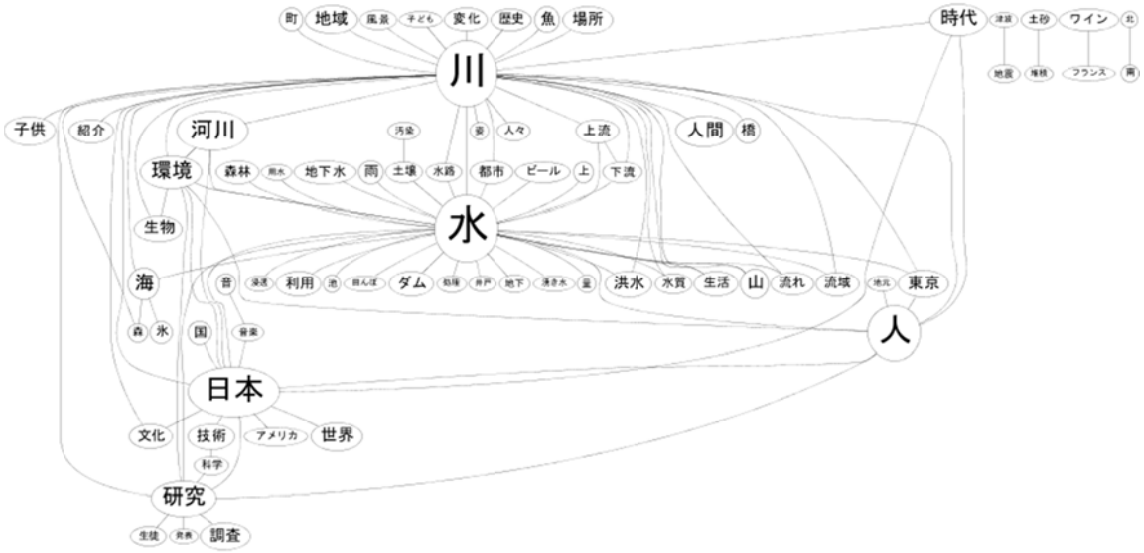

図 14　出現頻度上位 30%の語による共起頻度 77 回以上の共起ネットワーク

河川文化についても，単語のレベルではその構造を捉えられないため，オントロジを導入する．3 つの方法の選択だが，河川文化についても既存のオントロジは利用できないと考えられる．ゲームに比べれば専門用語は少ないが，今度は対象となる語の意味的範囲が広いため，通常のシソーラスではカテゴリとしての収束性が低くなる．例えば，【川】，【河川】，【水路】は同義語だが，これで 1 カテゴリとすると似たような規模の小さなカテゴリが大量にできてしまうことになる．一方で，「水」の関連語，といったカテゴリでくくってしまうと，カテゴリが大きくなりすぎる．

そこで，「ゲーム批評」と同様に，テキスト中の共起情報から自動生成を試みた．対象とする語は共起ネットワークと同様に頻度 223 回以上の 265 語とし，パラメータとする語の頻度は 100 回とした．図 15 は得られたクラスターの一部を抜き出したものである．最も小さな単位である【国】，【ヨーロッパ】や【アメリカ】，【外国】の辺りまではクラスターとしての意味も取れるが，それと結合している語は【音楽】，【ワイン】であり，さらに上位

まで行くと技術や科学が結びついてしまい，クラスターとしての意味づけが難しくなってしまう．それぞれの語の共起を見ていくと，いずれの語も【日本】という語との共起が多く，それが特徴となってクラスターが形成されている様子が見えた（図 14 左下参照）．従って，このクラスターは「【日本】に関連がある，ないしは相対的に語られる語のクラスター」という解釈となるが，そのようなカテゴリが河川文化の代表的な概念の 1 つとして適切とは言い難い．

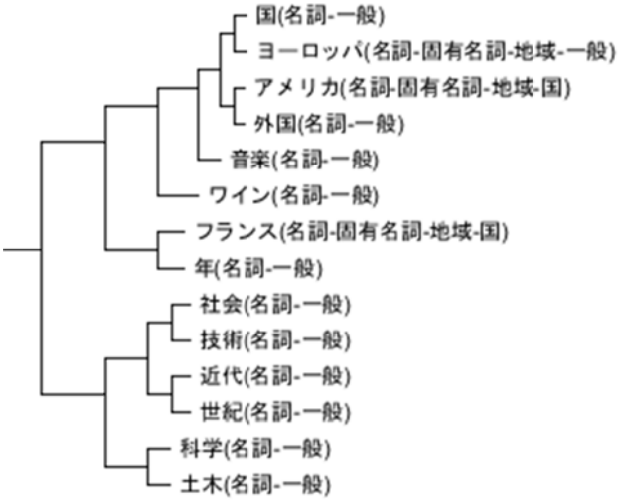

図 15　河川文化を対象とした名詞の階層化クラスタリングの結果の一部

クラスタリングが機能しない原因は何であろうか．図 14 で示したネットワークの密度は「ゲーム批評」の図 5 と比べると半分程度である．共起の閾値を取り払い，全ての共起を含んだネットワークとして計算しても，「ゲーム批評」のネットワーク密度が 0.82（ノード数 154）であるのに対して，「河川文化」では 0.63（ノード数 266）となる．また，図 14 のネットワークについて，【川】，【水】，【日本】の 3 つの中心的なノードを除外して同様の条件でネットワークを作成したところ，共起の数が一気に減少し，77 回以上のエッジは 22 本となった（図 16）．残った共起も，多くはテキスト全体の特徴と言うよりも，特定の少数の講演テキストに由来するものが多い．例えば，【ワイン】－【フランス】の共起回数は 118 回もあるが，そのうち 113 回は「フランス生活を通して体験したワイン文化」という講演に由来する．

これらのことから判るとおり，「河川文化」テキストは名詞の共起情報が乏しく，また数が多い共起は汎用的な高頻度語との共起であるか，特定の講演の内容に依存した共起である．そのため，共起をパラメータとしたクラスタリングによって望ましい構造を抽出できなかったと考えられる．これは，講演の内容が多岐に渡り，またそれぞれの講演は講演会としてまとまった話とするために，特定のトピックを中心とした内容となることに由来する．共通する用語がないことから，河川文化が曖昧でとらえどころがない，という元々の

問題が発生しているとも言えるだろう．

他の 2 つの方法が利用できないため，本分析では人手によるオントロジの構築を行い，それを利用して講演全体の分析を行う方針とする．オントロジ構築の詳細な手順については次節 5.4 で説明する．

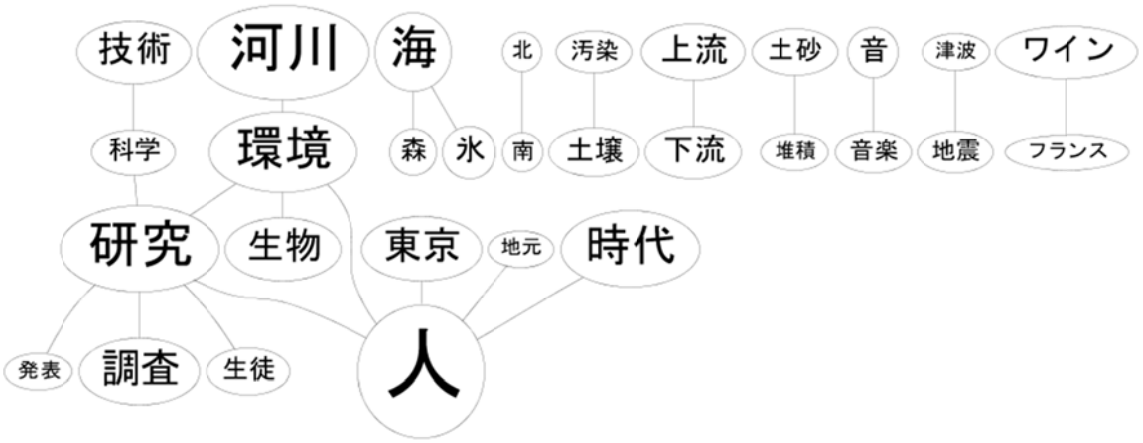

図 16　出現頻度上位 30%の語による共起頻度 77 回以上の共起ネットワークから【川】【水】【日本】を除外したネットワーク

その後，5.5，5.6 では要素の関連性の分析を行う．作成されたオントロジのカテゴリ単位で計量するだけでは全体の構造は見えないため，カテゴリ同士の関係性をテキストから抽出する必要がある．しかし，文単位での共起を見た場合，【ワイン】－【フランス】のように特定の講演の記述に全体が引きずられる可能性があるため，講演を単位とした共起を見る．具体的には，その講演で頻出すると言える複数のカテゴリを明らかにし（これは，講演に対してその内容に応じてタグ付けをする作業を，計量的に行っているということができる），そのカテゴリ同士を共起しているとして扱う．この共起のネットワークを構築し，コミュニティ抽出の分析を行う．5.5 ではテキスト全体に対して，5.6 では表 17 の分類ごとに行った．実際の講演の内容を見ると，川づくりや教育ボランティアなど具体的で身近な場所や対象に根ざした活動に関する内容と，気候変動や環境汚染など規模の大きな内容の 2 つの傾向がある．これらの内容が要素のコミュニティとして抽出されることが予想されるが，それらが分かれた内容となるか，他の概念によって接続されるのかは未知である．

一連の分析の流れを図 17 に示す．

5.4.　オントロジの手動構築による河川文化要素の抽出

5.4.1.　目的

本節では，テキストに基づいてオントロジを手動構築することを目的とする．

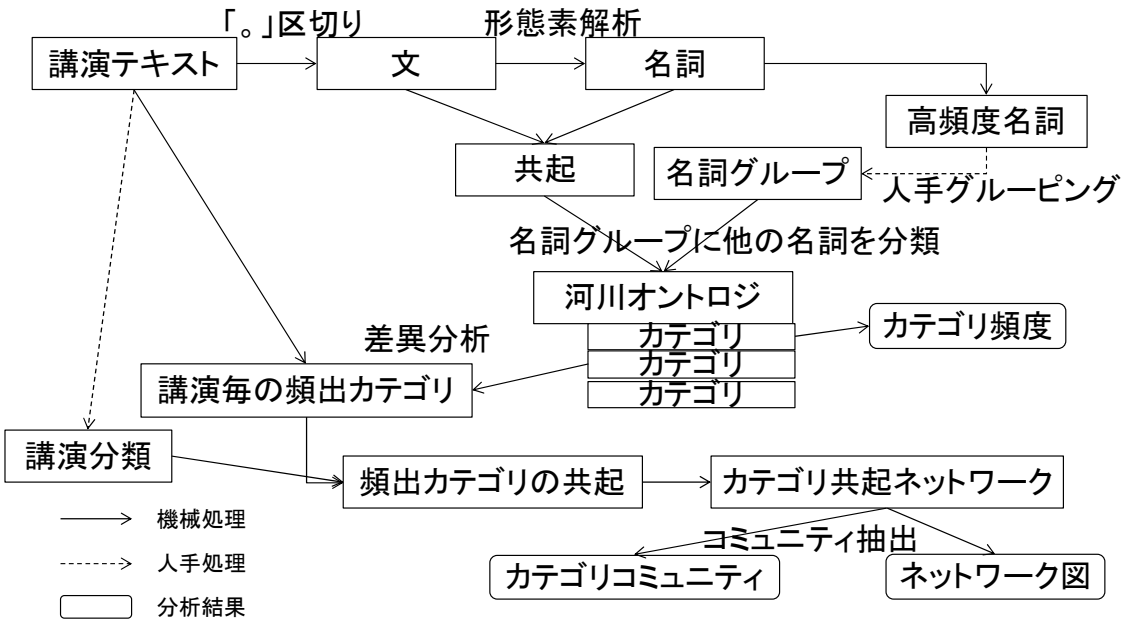

図 17　河川文化に対する分析フローの概要図

5.4.2. 方法

オントロジを手動構築する方法は無数に考えられるが，本分析では質と量を確保するため，ゲーム批評と同様の 2 段階のクラスタリングを人手を含めて行うこととした．1 段階めは，選択された高頻度の少数の単語を人手でクラスタリングし，2 段階めは単語の共起ベクトルを利用して 1 段階めで得られたカテゴリに残りの単語を分類する．5.3 で共起ベクトルによるクラスタリングは有効ではないと述べたが，それは大きな構造レベルの話で，例えば図 15 でも【アメリカ】と【外国】など末端のクラスターは解釈可能なまとまりとなっている．従って，既にあるカテゴリに対して分類していくような使い方であれば大きな問題は生じないと考えられる．ただし，特徴ベクトルが疎なことに代わりはなく，誤分類も多いと考えられるため，最後に人手によるレビューを行うこととする．

人手のクラスタリングは，構築者が自分の知識から自由な発想に基づいて分類を行うのではなく，テキストから頻度の高い名詞を選出し，その名詞の KWIC を見ながら分類を行った．そのため，そこには構築者の持つ知識が対象テキストという文脈を無視してそのままアウトプットされているのではなく，構築者が対象テキストを解釈した判断の基準が反映されていると言える．従って，得られたオントロジは辞書的・汎用的な内容ではなく，ゲーム批評のオントロジと同様，対象テキストに内在する情報から作り出されたテキスト特有のものである．

具体的な手順は以下の通りである．

(1) 品詞が「一般名詞」「固有名詞」「サ変接続名詞」（これら 3 つは接尾詞等とは異なり，他の語の一部となっていない名詞である）である語を抽出した．

(2) それぞれの講演毎に，単語の出現回数を参考に人手によって重要な名詞を選択した（全617 語）．この際，河川文化のどの講演でも頻度が高く，講演毎の特徴を消失させかねない語（【水】，【川】，【河川】，【日本】，【人】，【時代】）は除いた．これらの語はいずれも頻度 1000 以上であり，全講演の 90%以上に登場する語である．

(3) 抽出した名詞において，まず人手で同義語と考えられるものをまとめた（例：【降雨】，【降水】）．次に，人手で意味が近い語をグルーピングしていった．このようにしてできあがった単語の集合をカテゴリと呼ぶ．

(4) 出現頻度が 50 回以上でカテゴリに含まれない名詞（重要だが，見逃した可能性のある名詞と解釈できる）を，頻度 50 以上の名詞との共起ベクトルで各カテゴリに分類した．それぞれの単語の共起ベクトルと，各カテゴリの平均共起ベクトルの内積をとり，それが最大となるカテゴリに単語を分類した．頻度 50 未満となってくると，単一のテキストにしか登場しない語が増えてくる（60 台では 3 語，50 台では 4 語に対して，40 台では 7 語．累乗関数的に増えていく）ため，50 をボーダーとした．

(5) 自動分類の結果について，人手で明らかな誤分類を修正した．

(6) カテゴリには，含まれる頻度の高い単語を基に人手でカテゴリ名を付与した．

(7) カテゴリと分類について，別の専門家が妥当性をレビューした．

(8) このようにして作られたオントロジの各カテゴリについて，その出現頻度を計量した．

(9) さらに，講演毎にカテゴリの出現頻度を計量し，講演間で χ^2 検定の残差分析を行って，カテゴリの出現数が有意（$p<.05$）に多いと判断された講演を求めた．

5.4.3. 結果

テキストに含まれる名詞全 36079 語（合計出現頻度 439408 回）のうち，1064 語（合計出現頻度 102425 回）が要素として抽出され，49 のカテゴリに分類された．分類の結果を表 19 に示す．「含まれる単語の数」は，各カテゴリが内包する語の数を，「カテゴリの出現頻度」はカテゴリに含まれる単語が合計で何回登場したかを，「頻出テキスト数」はそのカテゴリが χ^2 検定の残差分析の結果，有意に多く出現すると見なされた講演数を示す．

出現頻度が最多のカテゴリは，「地理・地形・地質」であり，最少のカテゴリは「機械・装置」であった．一方頻出テキスト数が最大なのは「日本地名」，最小なのは「機械・装置」であった．頻出テキスト数の平均は 18.9 講演となった．

5.4.4. 考察

頻出テキスト数が最大であったのは「日本地名」であった．これは多くの講演が特定の場所に関する話題を扱ったものであるからだと考えられる．類似の「海外地名」なども頻出テキスト数が多く，河川文化に関する活動の多くは特定の川や地域に根ざしたものであるということが示唆される．

「河川一般」，「社会一般」，「科学技術全般」のカテゴリには，様々なコンテクストで使われ，単体では河川文化の特定の一面を示すとは言えないような語が多く分類された．例

えば，「河川一般」の【水辺】という単語は，景観，生物の生息地，護岸等の災害関連，人と川がふれあう場所等，様々な意味で用いられている．これらのカテゴリは，分類としては必要だが，その多義性から解釈が困難であるため，以降の分析対象から除外することとした．

5.5. 要素の関連性の分析（全体）

5.5.1. 目的

本節では，5.4 で抽出した河川文化の要素同士が講演全体の中でどのような関係性を持っているのか明らかにする．

5.5.2. 方法

5.3 で説明したとおり，河川文化における要素の関連性は特定の講演内容に大きく影響される文レベルの共起ではなく，講演単位での共起を見る．本分析では，1 つの講演内で，2 つの要素 A と B が両方触れられている（有意にその要素が出現する）とき，それらに関連性があるとした（以降，この組みをパターンと呼ぶ）．もちろん，時間の制約上講演者が語りたくとも語れなかった部分があると考えられるが，限られた時間の中で選択された話題だからこそ，その間の関連についても語りたかったと考えることができる．

具体的には，以下のような手順で各講演が内包する要素の関連性を抽出した．

(1) 5.4 と同様の方法（χ^2 検定および残差分析）により，特定の講演で有意に多く出現するカテゴリを求める．ただし，カテゴリとして単一の意味づけが困難である「河川一般」,「社会一般」,「科学技術全般」及びカテゴリとして抽象的な意味を持ちにくい（＝固有名詞である）「日本地名」,「海外地名」,「河川名称」の計 6 つのカテゴリは分析対象外とした．

(2) 複数の頻出カテゴリの全ての組み合わせについて，そのカテゴリ間に関連があるとし，パターンを作る．

(3) 全ての講演で，パターンの出現回数を数える．(1)～(3)を具体例で説明する．講演「地球環境問題と河川」では，「環境衛生」,「気候・気象」,「防災」,「政治・政策」,「地球環境」,「農業」,「水利用」,「河川工学」の 8 つのカテゴリの出現頻度が高かった．この講演ではこれらの 8 つに関係性があると考え，${}_8C_2=28$ のパターンが講演から抽出される．各パターンの出現を他の講演でも見てみると，例えば「地球環境」と「気候・気象」の両カテゴリに言及している講演は全部で 11 あったため，このパターンの出現回数は 11 となる．

表 19　カテゴリ一覧

カテゴリ名	代表的な語	含まれる単語の数	カテゴリの出現頻度	頻出テキスト数
地理・地形・地質	地下水，土壌，土砂	35	5022	22
環境・衛生	環境，汚染，栄養	43	4644	28
河川一般	上流，流域，流れ	28	4602	32
生態系	生物，動物，保護	37	4281	25
気候・気象	雨，冬，風	29	3997	22
土木構造物・建設工事	工事，ダム，橋	37	3980	24
防災	洪水，災害，治水	32	3693	21
日本地名	東京，京都，大阪	44	3444	33
森林・植生	木，植物，森	20	3392	19
海外地名	中国，アメリカ，ドイツ	45	3311	29
風習・民俗	文化，伝統，神社	50	3118	25
水中生物	魚，貝，アユ	25	3115	17
地球環境	地球，エネルギー，自然環境	18	3079	19
歴史	歴史，江戸時代，近代	40	3041	28
教育	子供，学校，学生	24	3011	16
科学技術全般	技術，開発，情報	25	3008	26
海	海，海岸，海水	20	2990	21
農業	農業，畑，米	32	2491	20
水利用	水源，用水，水道	25	2163	22
ヒト・カラダ	体，心，女性	30	2095	23
交通	道路，船，舟	12	2090	17
都市・まち	町，都市，市民	11	2026	27
食品・飲料	酒，食べ物，食料	19	1997	9
文学	歌，文学，物語	34	1833	16
学術	研究，学者，論文	3	1793	24
政治・政策	事業，行政，国土交通省	25	1770	19
生活・暮らし	生活，ふるさと，暮らし	32	1625	21
湖沼	湖，池，琵琶湖	19	1536	17
河川名称	利根川，多摩川，隅田川	24	1509	24
山	山，里，山地	15	1495	28
経済・経営	経済，会社，消費	13	1302	14
医療・福祉	命，生命，障害	30	1272	15
河川工学	流量，浸透，流速	22	1227	20
社会一般	社会，人口，ネットワーク	6	1186	16
音楽	音，音楽，琴	11	1154	6
景観	風景，景観，景色	14	1038	12
哺乳類・鳥類	鳥，コウノトリ，カラス	6	1012	10
物質・資源	コンクリート，資源，鉄	12	914	17
爬虫類・両生類・虫	ホタル，トンボ，オオサンショウウオ	11	900	6
水質	水質，塩分，軟水	7	862	13
スポーツ・アメニティ	観光，遊び，旅行	20	846	15
漁業	水産，釣り，漁業	16	813	15
絵画・美術	絵，芸術，デザイン	13	700	15
テレビ・マスメディア	テレビ，映像，映画	9	672	16
造園	庭，桜，栽培	12	621	13
建築	屋根，建築，住宅	8	562	22
物理	空間，光，物理	5	548	19
芸能	能，歌舞伎，芸能	11	368	7
機械・装置	ポンプ，人工衛星，GPS	5	277	5

(4) このようにして得られたカテゴリのパターンを，カテゴリをノード，パターンをエッジとするネットワークとして可視化した．なお，頻度が高いパターンの代わりに少数のパターンに着目して，講演の独自性や講演者の個性などを抽出することも可能であるが，本分析では一般的な河川文化概念の抽出を目的とするため，出現頻度の高いパターンに着目している．

(5) ネットワークの構造を計量的に捉えるため，ネットワークに対して Girvan-Newman コミュニティ抽出法でコミュニティ抽出を行った．

5.5.3. 結果

全 132 の講演は最低 3 カテゴリ，最大 12 カテゴリ，平均 7.0 カテゴリについて言及しているという結果が得られた．パターンの種類の最大論理値は ${}_{43}C_2$＝903 種類であるが，実際に見いだされたパターンは 697 種類であり，カテゴリの出現パターンに一定の偏りがあることが判る．パターンの最大出現回数は 14 回，平均は 2.9 回であった．

全てのパターンを表示したネットワーク図はエッジが多く解釈が困難であるため，閾値を決め値以上の出現回数を持つパターンのみを分析することでネットワークの中心となる構造を明らかにする．そのために，閾値を 2 から 1 ずつ増加させてネットワーク構造の変化を見た．閾値 3 までは，全てのノードがネットワークにつながっていたが，閾値 4 で「音楽」が，閾値 5 で「機械・装置」，「爬虫類・両生類・虫」，「芸能」がメインコンポーネント[1]からそれぞれ単独で分離された．閾値 6 になると，「造園」と「景観」がパターンでメインコンポーネントから外れる．閾値 7 になると，「スポーツ・アメニティ」，「食品・飲料」，「テレビ・マスメディア」，「農業」がさらにメインコンポーネントから外れた．8 回以上のパターンで描画すると，「医療・福祉，ヒト・カラダ，生態系，水中生物，漁業」と「海，気候・気象，地球環境，学術」という 4 つ以上の要素を含むサブネットワークがメインコンポーネントから分離した．閾値 4 から 7 までの分離は，全体として他とつながりの薄いトピックの分離と考えられるが，7 から 8 に上昇させた時の分離は，理解したい概念のつながり方自体の分離と考えられる．そこで，閾値としては 7 を採用することにした（図 18）．

この閾値 7 のネットワークに対して Girvan-Newman コミュニティ抽出法を行った．分割数 2 から 20 までを計算したところ，分割数 4 でネットワーク分割のまとまりの良さを示す Modularity が 0.558 の最高値を取った．図 18 では，この分割数 4 の場合のグループをノードの形で示している．なお，分割数 3 では 0.512（●と▼が 1 グループになる），分割数 5 では 0.555 であった（▲が「ヒト・カラダ」－「生態系」のエッジで 2 つに分かれる）．

[1] ネットワークの最も大きな塊

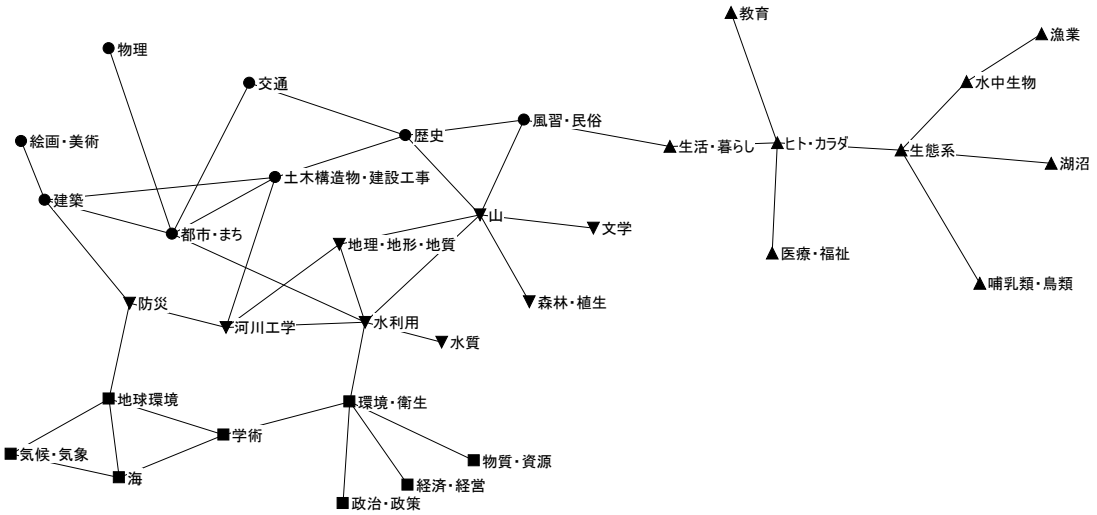

図 18　カテゴリネットワーク図

5.5.4. 考察

Girvan-Newman 法で分割された 4 つの部分ネットワークは，それぞれ以下のように解釈できる．まず 1 つめは，ノードの形が■の環境問題を中心としたカテゴリ群である．講演内容の傾向から出現が予測された，「地球環境」・「気象・気象」・「海」など，グローバルな視点が多く含まれる群である．2 つめのカテゴリ群は，こちらも出現が予測された，ノードの形が▲の，ローカルな要素を含む群である．例えば，ヒト・カラダ，生活・暮らし，教育などが含まれる．ただし，「文学」，「歴史」などのローカルな内容はこのコミュニティと隣接してはいるが，多少離れた位置にある．

これらのグローバルなカテゴリ群とローカルなカテゴリ群を繋ぐ形で，ノードの形が●と▼の 3 つめの群が形成されている．この群は「河川工学」，「水利用」，「水質」など，河川そのものに関わるカテゴリ（▼）と，「都市・まち」や「建築」などの工学的側面を持つカテゴリ（●）が合わさった形になっている（さらに，それぞれの要素に強く結びつく文化的要素のカテゴリ「絵画・美術」や「歴史」が結びついている）．ここから確認できるのは，河川文化を構成する要素において，ローカルな問題とグローバルな問題は，「川」や「山」などの土地と，「建築」などの工学を介してつながるということである．

以下の節では，ネットワークの中で重要な関連を取り上げて考察していく．

5.5.4.1. 河川工学と地球環境問題

地球環境と気候・気象を包含した問題として，気候変動の問題をあげることができる．気候変動の問題は，人類の絶対的生息基盤の存続のために喫緊の課題である．例えば，脱温暖化のための具体的な CO_2 削減の実現にむけて，国家レベルで様々な取り組みを実施している．そのような社会的背景のもとで，河川整備においても地球温暖化問題へ何らかの形で貢献することが求められる．例えば，コンクリートを極力用いない多自然工法による

河川工事は,従来の工法に比してCO_2排出量が大幅に低減できることが示されている[109]. また，再生した草本類をバイオマスとしてエネルギー利用することにより，さらなる削減効果が期待できる．さらに，河川における小水力発電設備の導入によって，火力発電や原子力発電に依らないクリーンエネルギーの生産が可能となる[110]．このように，今後検討しなければならない重要なテーマは，河川整備や河川管理を通してどのように地球規模での気候変動問題の解決に貢献するかということである．しかし，ネットワークでは「地球環境」と「河川工学」は直接接続されておらず（パターンの出現回数は 4 回),「防災」を通じて繋がっている.

ネットワーク図には表示されていない「河川工学」と「地球環境」「気候・気象」のパターンを含む講演を調べると，前者は 4 講演，後者は 6 講演が見つかった．これらの講演では，特に世界における気候変動に起因すると考えられる水害，およびそれらの水害等に対する対処の方法について言及しているという点で，おおむね共通している．例えば,「地球環境問題と河川」の講演者である沖大幹は，河川工学において的確な治水対策を施すためには，降雨や水文データなど広範囲の気候現象を視野に入れ，予測を実施していくことが重要であると述べている．河川の問題についても地球レベルでの視野が必要となるのである．また,「ネパール王国を襲う氷河湖決壊洪水」では，氷河湖の決壊とそれがもたらす災害について論じるなかで，気候変動と氷河変動の問題についても言及している.

上述のようなことは，ネットワーク図においても,「河川工学」と「地球環境」のカテゴリは,「防災」カテゴリを介することでつながりをもっていることからも把握できる.

5.5.4.2. 水系を包括的にとらえる視点

ネットワークから，河川工学と海の要素間のつながりはそれほど強調されていないことが判る．河川と海は連続しており，それらを流域や水系の視点から連続的に捉えることの重要性は，多くの研究者によって指摘されている．川から海への水系を包括的に捉える視点を導入するためのヒントはどのような点にあるのだろうか.

「海」カテゴリと「河川工学」カテゴリのつながりが強かった講演は全部で 4 つ存在した．その 1 つ,「さまよえる湖・消えゆく湖－変貌する中央アジアの水環境－」では，タクラマカン砂漠で頻発するようになった洪水被害とアラル海の環境問題との関連性について,水循環のバランスに着目しながら言及している．アラル海は海ではなく内陸湖であるものの，砂漠，洪水，湖というそれぞれが連関する流域システムを包括的に論じる視点をもっていると言える．この講演では，人為によって水環境にわずかな変化を加えれば，それが時としてより広範囲の水循環のバランスに影響をおよぼす可能性を指摘し，自然の摂理に従った水資源開発や水管理の必要性を説いている．その他 3 つの講演においては，海と河川に関する用語が出てくるものの，それらは別々の物として語られており，川から海までの水系全体の関連性はほとんど強調されていない.

5.5.4.3. 防災と生活

「防災」カテゴリと「生活・暮らし」カテゴリについて，双方を網羅的に語る講演は存

在しなかった．ネットワーク図では，「防災」は「地球環境」や「河川工学」のカテゴリと結びついていることから，これまでの防災については技術的あるいは科学的側面を重視して考えられてきたことが判る．東日本大震災の経験から判るように，科学的技術的な方法で対処することがコスト等の観点も含めて現実的ではない災害が発生することがある．重要なのは，科学技術によって天災からハードに人間を守る姿勢と，それでは防げない災害にソフトに応じていこうとする姿勢をうまく融合させることである．さらに言えば，自然を対立項と捉えて災害を押さえ込もうとする近代の思考的枠組みを超えることである．また，人びとの生活のなかに防災意識を位置づけ，自然現象と良好な関係を築いていくような新たな文化を構築していかなければならない．このことが，これから人間が河川文化を醸成していくうえでの重要な課題となるだろう．

5.6. 要素の関連性の分析（講演カテゴリ）

5.6.1. 目的

本節では，要素の関連性を講演カテゴリ毎に分析し，講演者の所属や専門領域，関心分野などによる視点の違いを探る．

5.6.2. 方法

表 17 に示す講演カテゴリ毎に，5.5 と同様の手法で分析を行った．閾値については，全体のネットワークの時と同様に 2 から閾値を上昇させ，部分ネットワークが生じないようにした．結果として，講演数が少ない「社会・暮らし」,「土木・空間」の講演カテゴリについては閾値 3,「環境・生態」,「文化・歴史」については閾値 4 となった．

5.6.3. 結果と考察

5.6.3.1. 土木・空間のネットワーク

土木・空間の講演カテゴリには，具体的には河川工学，水文学，建築，景観工学，防災などの分野における研究者・実務家の講演が含まれる．

ネットワーク図（図 19）を見ると,「河川工学」（次数中心性 11 で最高値）を中心として，土木，建築，都市・まち，といった要素が多く語られていることが判る．またネットワークの密度が高く，要素同士の関連が比較的まとまっていることから，これらの講演は自分達の領域の要素をしっかりと捉えている，あるいは要素が限られている，ということが推察される．Girvan-Newman 法で分割すると，Modularity の最大値は 2 分割で 0.261 であり,「建築」を中心とした都市的な内容（■）とそれ以外（●）に分かれた．0.3 よりも低いことから，ネットワークの分割性が良いとは言いがたく，1 つの凝集したネットワークと考えられる．

特徴的なのは，絵画・美術に関する言及が多かった（6 講演）ことである．これは，景観や建築物および土木構造物のデザインに関する講演のなかで,【絵】,【デザイン】,【美学】,【アート】といった単語が頻出していることに起因している．また,「風習・民俗」（5 講演）,

「歴史」（7 講演）についても多く語られている．信玄堤に関する講演や伝統的河川工法など，建築史に関わる講演が見受けられた．

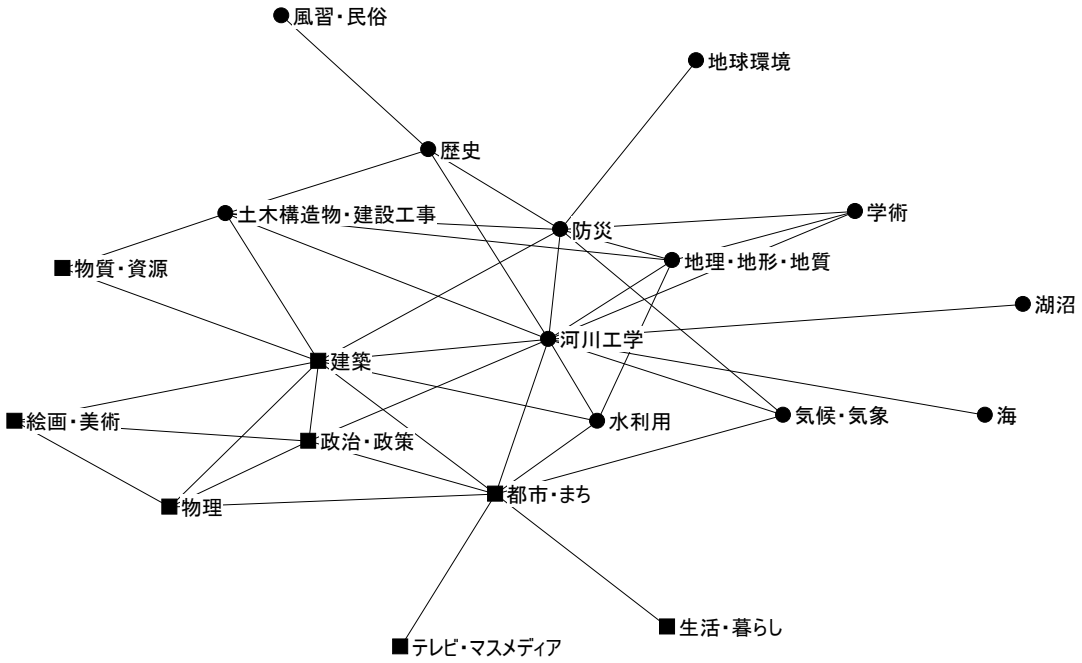

図 19　土木・空間のネットワーク図

5.6.3.2. 環境・生態のネットワーク

環境・生態の講演カテゴリには，衛生工学，地球環境，バイオ・リサイクル，化学，森林，生物，生態系，海洋などの分野における研究者・実務家の講演が含まれる．

環境･生態のネットワーク（図 20）を Girvan-Newman 法で Modularity が最も高い（0.500）4 つに分割すると，「生態系」（▲），「気候・気象」（■），「環境・衛生」（▼），「山」（●）をそれぞれ中心とする群に分かれる．「気候・気象」の群は図 18 と同様グローバルな視点のカテゴリを示すと考えられる．ここに「生態系」を中心とする生物のカテゴリ群と，環境と言っても身の回りの環境を示す「環境・衛生」群が結びついている構造が見てとれる．

「気候・気象」では，図 18 では離れたところにあった「物理」カテゴリが近接している．これは，物理シミュレーションによって気候や環境の問題に取り組むいくつかの講演に起因している．「生態系」の群で注目すべきは，医療・福祉について多く言及している講演が 5 つある点である．これらの講演では，環境と生命の問題を関連付けて医療を論じていることが特徴的である．「環境・衛生」の群では，「歴史」のカテゴリが特徴的である．これには，近代以前の水循環システムを考察しながら，現代の水環境のあり方について言及した講演が影響している．

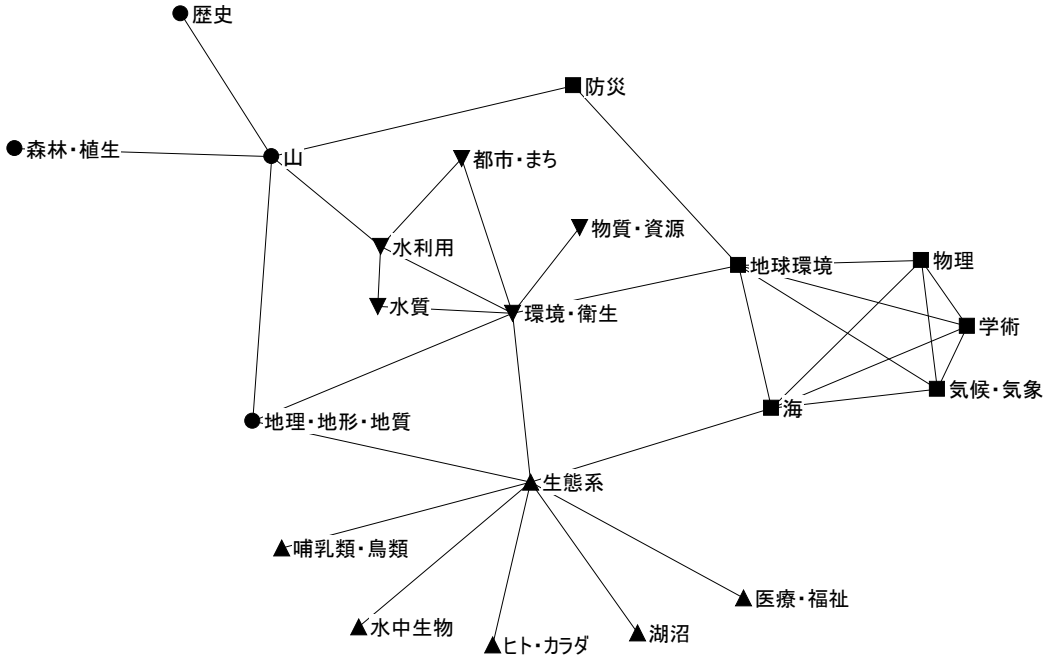

図 20　環境・生態のネットワーク図

5.6.3.3. 社会・暮らしのネットワーク

社会・暮らしの講演カテゴリには，教育，食品，水産業，農業，政治などの分野における研究者・実務家の講演が含まれる．

ネットワーク（図 21）は大きく 2 つの群に分かれる．「教育」をベースにした上半分（ノードの形が●）と，「漁業」，「食品」，「資源」などをベースにした身の回りの資源の群（ノードの形が▲・■）である．Girvan-Newman 法では，3 分割が最も Modularity が高いが（0.407），2 分割（▲と■が結合される）の Modularity も 0.398 であり，グループとしては上半分と下半分と考えて問題ない．

上半分の教育の群では，高校教員などの教育実務者によって，教育現場における先端的科学技術の実践や草花の品種開発等の取り組みが語られていることもあり，「学術」とも関係がある．ここに，「医療・福祉」や「スポーツ・アメニティ」など，人間を対象とした生活の要素が結びつきこの群を構成している．

一方で，下半分のカテゴリは「水中生物」を中心とした漁業や水辺の生き物が，「物資・資源」を経由して「食品」関連のカテゴリにつながっている．これらのカテゴリは，漁業，食品会社，農業などの個別の講演から生じているが，「政治・政策」も踏まえた資源という大きな絵が重要であるということを示唆している．ただし，社会・暮らしの講演カテゴリでは，「水質」に関する言及は 1 講演，「水利用」に関する言及は 0 講演であり，水を資源として捉える観点が不足していると考えられる．「水利用」は土木・建築の講演カテゴリで主に語られている．

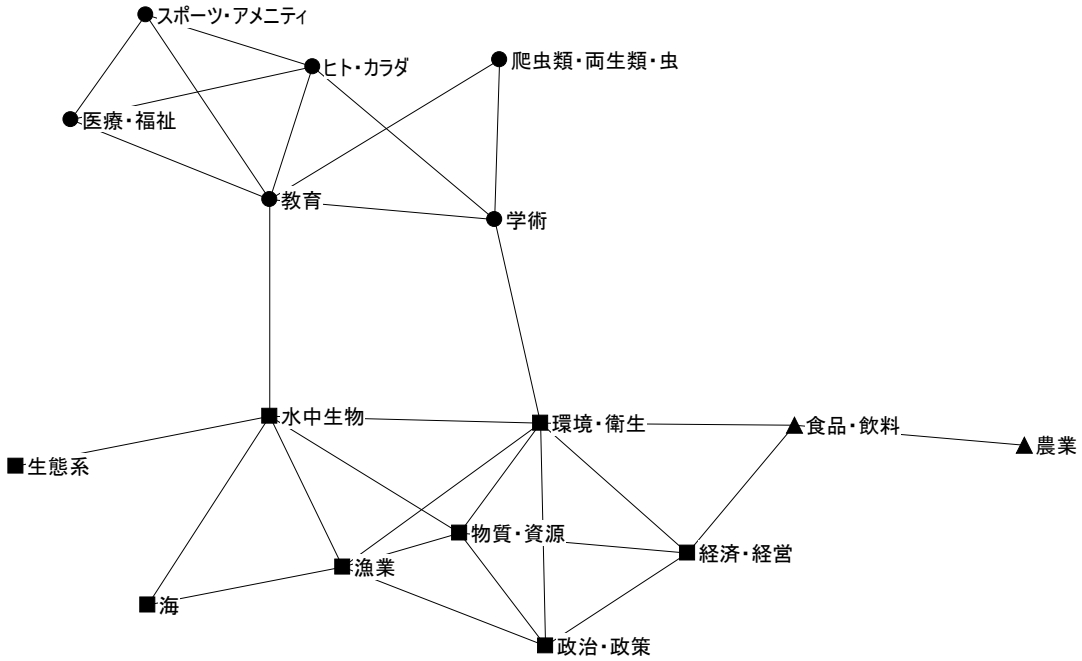

図 21　社会・暮らしのネットワーク図

5.6.3.4. 文化・歴史のネットワーク

文化・歴史系の講演カテゴリには，絵画，文学，映像，音楽，歴史，民俗学，宗教などの研究者および実務家の講演が含まれる．これらの講演では「風習・民俗」を中心として，凝集したネットワークになっている（図 22）．Girvan-Newman 法では，5 分割が最も Modularity が高かったが（0.343），「景観」「造園」や「水中生物」，「漁業」などの端に位置するノードが分割されるのみであり，意味の解釈には役立たなかった．

興味深いのは「交通」で，文化・歴史の中の 12 の講演で，「交通」に関する単語が有意に多く語られていた．これらの講演のほとんどで，川を軸とした交通機能としての舟運について言及している．また，「河川を舞台にした能－殺生の川・恩愛の川－」では，川を舞台にした能の演目のなかで，やはり舟の登場する情景を語っている．ネットワーク上でも他の様々な要素と結びついていることから，河川の交通には様々な歴史的観点が関係していることが示唆される．なお，「交通」は他の講演カテゴリではネットワークに出現していない．現在ではあまり利用されていない河川を使った交通を対象としているため，必然的に歴史的な話になってしまうからであると考えられる．しかし，講演の中にはこのような歴史的な水運を復興させようという動きの話もあり，土木・空間の講演カテゴリ等でも言及されてよい要素である．

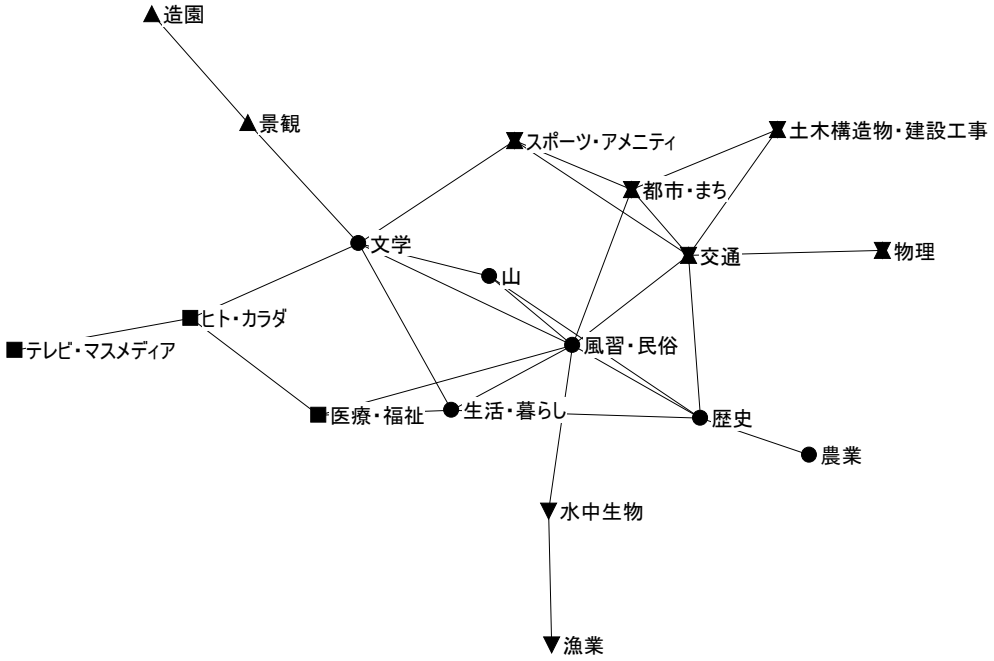

図 22　文化・歴史のネットワーク図

5.7. 実践への貢献可能性

5.2 で述べたように，本章の共同研究者である桑子，高田は，新潟県佐渡市の天王川自然再生事業に，合意形成マネジメントチームとして携わっている．そこでは，河川の下流に位置する湖の漁業者との合意形成が重要なポイントであった．このような川づくりの実践上の課題について,本分析の成果が貢献しうる内容の１つとして講演の検索が挙げられる．

もちろん，適切なキーワードで全文検索することでも必要な講演が見つかる可能性はある．しかし，分野横断的な領域の場合，類似の内容を別の語で示すことは一般的であり，網羅性をもったキーワード検索は困難である．また関連分野数が多いため，周辺分野や関連分野の関係性や，個々の分野が問題としている内容も明示的ではない．これに対して本分析の手法は，ネットワーク構造の視覚的提示を実現し，かつ特定の問題領域を示す語をカテゴリに分類したことで，分野間の関係性を把握した上でその問題点に関わる適切なキーワード（カテゴリ）を用いて論文を検索することを可能にしている．

例えば,「漁業」は,「水中生物」を介して「生態系」とつながっているが,「生活・暮らし」,「環境・衛生」,「河川工学」などのカテゴリとのつながりについては，これまでの講演でそれほど強調されていないことが判る．そこで,「漁業」と「生活・暮らし」,「環境・衛生」,「河川工学」のカテゴリについて有意なパターンを構成している講演があるかどうかを調べてみる．「漁業」と「生活・暮らし」および「河川工学」のカテゴリについて有意に語っている講演はなかった．「環境・衛生」については,「二十一世紀の食料・農業問題―農的循環社会への道―」,「汽水に包まれた国，日本」,「河川の生態系と漁業について－

汽水域のヤマトシジミを中心にー」,「海と水産業の再生―森と川とのかかわり―」の 4 つの講演が抽出できた．講演数としてはわずかであるが，これらの講演のなかから課題解決のための重要な示唆を得られるかどうかみてみよう．

その「汽水に包まれた国，日本」では講演者が，カキ養殖のための良好な環境は，海から森までの環境を一体的にとらえることで担保されるという考えを述べている．つまり，上流側の環境改善は，ひいては下流側の水産資源の品質確保へとつながるというのである．そこで講演者は，山に漁師が広葉樹をつくるという例を紹介している．また，その「海と水産業の再生―森と川とのかかわり―」では，河川を含む水循環全体のシステムの改善が，水産業や農業などの問題解決につながると指摘している．

このように講演記録から，上流側の環境再生は，下流に位置する海，あるいは湖の水産資源にも必要な取り組みであるという情報を得ることができる．河川文化の講演集を構造化したことによって，それぞれの取り組みにおけるトピックに即したテキストの抽出が可能となり，現場での有用性が増したと言える．さらに，直感的には関連性を持っているにもかかわらず，その関係を語った講演がないことは，そのような視点についても意識的に考えなければならないということを示唆している．

5.8. 分析の結論

本分析では，河川文化のテキストからその要素を抽出し，ネットワーク化することでその構造化を試みた．講演毎に異なった特徴を持つためにローカルな構造を抽出する手法の効果が低いコーパスに対して，人手によるオントロジ構築とテキスト単位の特徴抽出を行うことで，全体構造を抽出することが可能となった．人手のオントロジ構築については，2 段階のクラスタリングを採用することで人手による労力を軽減させつつ量を確保することができた．結果として，図 18 で示されたように，河川文化に関わる種々の要素はグローバルな部分，ローカルな部分に分かれており，その間をつなぐ要素として河川工学などの「河川」があるという大きな構造が示唆された．

本分析は講演集「河川文化」を対象としたケーススタディである．しかし，5.7 で述べたようにこの構造を利用して，実際の活動で問題となる要素を抽象化したり，応用可能な講演を探したりすることができる．たとえば，本分析の成果を用いることで河川管理者は，河川にかかわるより多様な視点を得ることが可能となる．このことは，河川環境を多機能的に捉え，河川内の部分的な整備・管理にとどまらず，川を軸とした自然と人間が共生する豊かな地域社会を形成していくためのヒントを見出すことにもつながる．そういった点において，本分析は河川文化という複雑かつ曖昧な概念を，川づくり，あるいは河川にかかわる研究活動のなかに組み込んでいくためのひとつの契機，あるいは具体的手法の提案という意味合いも持つ．

今後の展望としては，今回明らかになった概念構造を元にして，更に多くの河川文化を対象とするテキストを分析することが挙げられる．さらには実践の中で出てきたドキュメントに本分析の手法を適用することで，個別の事例の構造化と全体の構造の比較を行うこ

とも考えられる．また，東日本大震災後には防災関連等で大きな意識の変革があったと考えられるため，同じ「河川文化」を題材として，震災前後の比較を行うことも検討する必要がある．

6. 既存オントロジの利用による評価語計量比較－4 ジャンルの批評における感性の違いを探る－

6.1. 本章の目的

本章では，映画，演劇，文学，ゲームという 4 つのジャンルの批評テキストを対象とした評価語の分析について，既存のシソーラスを利用し計量的に比較する方法が有効であることを述べる．

6.2. 分析の背景と目的

感性をキーワードとして，人間が対象に下す評価を分析・応用する様々な研究が行われてきた．より盛んなのは応用で，感性語（評価語）は非言語的な感覚を言語化することができるため，これを利用した検索システムが画像[111]，音楽[112]，文学[113]等の分野で実現されてきた．

一方で，感性語そのものに関する分析は，映画，ゲームのジャンルでは事例があるが[114][115]，演劇や文学においては存在せず，それらのジャンルを横断した研究もまた存在しない．感性語を利用するだけなら単一ジャンルの分析であっても問題はないが，感性を理解するためには，どのような感性が，なぜそのジャンルにおいて特徴的であるのかを明らかにする必要があり，そのためには比較が重要になってくる．筆者を含む徃住，村井らのグループでは，文学[51]や音楽[54]の分野で批評文の計量的分析を行ってきており，映画と演劇における批評対象の特徴と違いについても研究を重ねてきた[53]．本分析は，それらの研究を受け，批評における感性語の特徴についてジャンル間の違いを明らかにすることを試みるものである．

本分析では，文学，映画，演劇，ゲームという 4 つの異なるメディアにおける批評文の感性について分析を行う．目的は，分野毎に異なった性質をもつ批評テキストについて，各分野がどのような感性語で形容・評価されているかを調査し，その類似点と違いを示し，もって分野毎の感性的構造の違いに関する示唆を得ることである．

6.3. 対象データと手法の選択

対象としたデータは，各ジャンルにおける批評雑誌（ジャンルにつき 1 誌）である．対象雑誌を表 20 にまとめる．

各雑誌は，偏った記事（いわゆる，提灯記事）が少ない，批評誌としての性質の強い雑誌を選択した．それぞれの雑誌から，1 作品に原則数ページを割いて紹介・批評しており，かつ長期にわたって継続しているコーナーを選び出し，その全記事を OCR によって電子テキストとした．

表 20　対象雑誌

ジャンル	雑誌タイトル	コーナー名	期間	テキスト量（語数）	対象作品数	著者数
文学	文藝[116]	Book Review	1995-2009	397279	454	244
演劇	シアターアーツ[117]	劇評	1995-2009	434814	187	98
映画	映画芸術[118]	映画評	1995-2009	664739	575	213
ゲーム	ゲーム批評[97]	ゲームソフト批評	1994-2006	904562	884	140

本分析の目的は批評テキストに内在する感性を探ることであるため，4 章，5 章のように名詞を分析の対象とするのではなく，形容語（自立形容詞及び形容動詞語幹名詞）を対象とする．まず単純に，各ジャンルにおける頻出形容語上位 10 位を列挙した（表 21）．表中，背景が薄灰色の単語は 4 ジャンルの全てで上位 10 位以内に出現する単語であり，濃灰色の単語は 10 位以内ではそのジャンルにしか出現しない単語を示している．この結果からでも，4 ジャンルの批評で使われている語の頻度傾向が異なることや，特徴的な語があることは判る．定量的な分析を行うため，全ジャンルの形容語を合計した際の頻度上位 20 語について，χ^2 検定の残差分析を行った（表 22）．ここから，ゲームにおける語彙特徴は【多い】，【高い】など量的な語が多く，映画では【面白い】，【良い】，【悪い】などの直接評価的な語が多い事などが見て取れる．

しかしながら，例えば【大きい】は数量的な内容ではあるものの演劇で多く，ゲームにおいて「量的な語が多い」と抽象化して良いかについて定かではない．さらに，形容語は 2475 語，合計頻度 70432 であり，上位 20 語では合計頻度 14955 と全体の 21%でしかないため，1 つ 1 つの頻度が低い他の複数の類似の語を合計しても同様の傾向が出現するのかについては定かではない．従って，形容語を分析するにあたってもオントロジを導入し，多くの語を含むカテゴリを単位として分析する必要があると考えられる．

表 21　各ジャンルにおける頻度上位 10 位までの形容語とその出現頻度

	文学		演劇		映画		ゲーム	
	単語	出現頻度	単語	出現頻度	単語	出現頻度	単語	出現頻度
1	ない	1392	ない	1178	ない	2176	ない	2644
2	問題	202	問題	305	いい	347	高い	768
3	いい	165	可能	172	問題	256	面白い	731
4	生	130	新しい	139	面白い	239	多い	676
5	必要	120	若い	131	強い	205	必要	560
6	深い	107	必要	123	必要	175	可能	542
7	自由	99	強い	120	若い	175	楽しい	526
8	可能	94	高い	118	良い	175	良い	515
9	強い	85	大きい	102	悪い	166	強い	513
10	美しい	85	確か	100	多い	162	問題	459

表 22　全ジャンルにおける頻度上位 20 の名詞の頻度の χ^2 検定結果（$p<.05$）

単語	文学	演劇	映画	ゲーム
問題	202▲	305▲	256▲	429▽
面白い	61▽	67▽	239▲	731▲
いい	165▲	87▽	347▲	415▽
多い	81▽	87▽	162▽	676▲
高い	43▽	118	80▽	746▲
必要	120	123	175	560
強い	85	120	205▲	512
可能	94	172▲	132	393▽
良い	48▽	55▽	175▲	513▲
新しい	85	139▲	101▽	419
大きい	49▽	102	80▽	417▲
楽しい	33▽	37▽	53▽	525▲
悪い	50	58▽	166▲	326
非常	27▽	27▽	54▽	447▲
確か	72	100▲	132▲	250▽
深い	107▲	99▲	129▲	205▽
自由	99▲	80▲	97	216▽
重要	58	83▲	92	237▽
難しい	28▽	32▽	45▽	360▲
長い	75▲	83▲	86	216▽

前章までと同様，3 パターンのオントロジ導入について検討する．まず自動生成だが，クラスタリングのパラメータ選定が困難である．形容語同士の共起は，その共起が必ずしも形容語同士の意味の近さを示さないと考えられるため，適切ではない．名詞との共起を使った場合，「類似の対象を評価する語」という観点でクラスタリングが可能と考えられるが，ジャンルを越えた場合には機能しないと考えられる．これは各ジャンルにおける高頻出名詞を示した表 23（背景色については表 21 と同様）を見れば明らかであるが，ジャンルによって名詞頻度の特性が大きく異なるからである．その他，意味語以外や文の表層的特徴を使ってクラスタリングすることも，意味的に同様の形容語をまとめるという観点から適切ではない．次に手動での分類だが，2000 語以上もある形容語について，分析者の直感や感性でクラスタリングを行っていくことは現実的ではない．共起情報が利用できないため，二段階クラスタリングも不可能である．以上より，本分析では既存のオントロジを採用する．専門用語ではないため，既存のシソーラスで対象となる語の多くをカバーできる．具体的には，国立国語研究所によって編纂されたシソーラスである分類語彙表[65]を利用することとした．分類語彙表の感性への適用は，政治テキストにおける感性を対象とした先行事例[46]も存在する．

表 23　各ジャンルにおける頻度上位 10 位までの名詞とその出現頻度

	文学		演劇		映画		ゲーム	
	単語	出現頻度	単語	出現頻度	単語	出現頻度	単語	出現頻度
1	小説	1037	舞台	932	映画	4381	ゲーム	7368
2	世界	1023	作品	611	作品	1199	プレイヤー	1677
3	言葉	795	演劇	606	監督	955	作品	1600
4	自分	780	劇	589	自分	870	システム	1417
5	物語	652	演出	534	男	804	世界	1286
6	作品	648	観客	468	女	728	プレイ	1271
7	本書	524	身体	463	物語	696	本	1230
8	人	511	言葉	453	世界	620	敵	1202
9	本	492	男	394	人	590	キャラクター	1179
10	人間	481	上演	365	日本	535	シリーズ	860

ジャンル毎の比較分析は，単語に対して実施したようにジャンル毎の χ^2 検定の残差分析で行う．ジャンルのテキストを 1 つの塊として扱うことで，それぞれにおける評価対象の違いや文章の書き方の違いといった内容による影響を排除できる．書き手がテキスト中で利用する形容語の選択には対象芸術の感性に関わる何らかの意思決定が反映されているはずであり，単純な計量による比較でもそういった精神活動の断面が抽出されうると考えられる．

分析の手順を図 23 に示す.

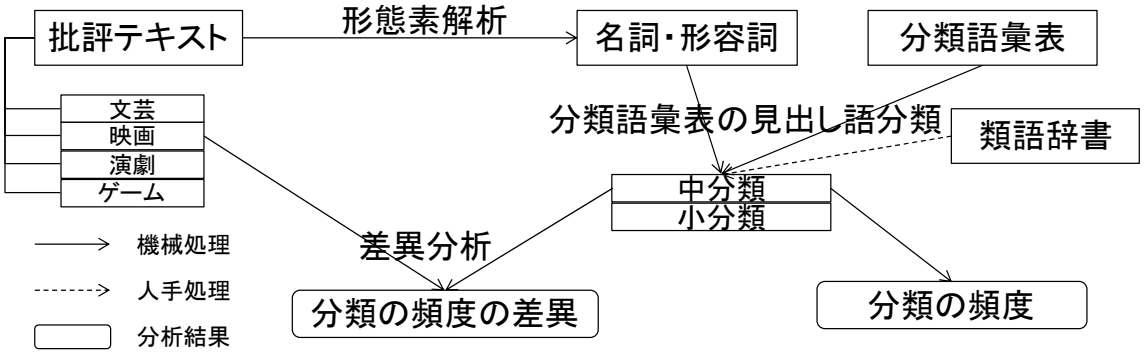

図 23　4 ジャンル批評に対する分析フローの概要図

6.4. 評価語の抽出と分析

6.4.1. 方法

分類語彙表は日本語を 4 つの類（名詞の仲間である体の類，動詞の仲間である用の類，形容詞・形容動詞・副詞等の仲間である相の類，その他）に分けた上で，5 つの部門（抽象的関係，人間活動の主体，精神および行為，生産物および用具，自然物および自然現象）とそれをさらに細分化した中項目，小項目に分類したシソーラスで，7 万 9 千語が収録されている．本分析では，この中でも「相の類」を利用した．各部門の中には，複数の語をまとめた中項目，小項目が立てられているため，その項目を 1 つのカテゴリとすることで，容易にオントロジとしての利用が可能である．

分類語彙表の見出し語は必ずしも形態素解析で得られる単語とは一致せず，複数の語の連続と一致する場合がある．また，見出し語に品詞情報はないため，形容語のみを見出しとマッチさせると，分類語彙表の語彙を生かし切れない可能性がある．一方で，得られる形容語の全てが分類語彙表に掲載されているわけではない．そこで，一部人手を交えながら，単語の n-グラムを見出し語にマッチさせる以下の方法を取った．なお，抽出された語は形容語以外も含むため，以降では評価語と称する．

(1) 単語を形態素解析し，単語へと分割した．
(2) 抽出された単語の 1～5gram について，分類語彙表の相のカテゴリに含まれる見出し語と文字列が一致する単語列を抽出し，評価語とした．
(3) 抽出された評価語について，明らかに評価語として使われていない誤抽出（名詞や形容詞を含まない n-グラムが文字列的にたまたまマッチし，評価語として抽出されているケース等）を人手で除去した．同時に，複数の分類にまたがる評価語について，文脈上最も用法が多い 1 つの分類に決定，同音異義語の判定も行った．
(4) 分類語彙表に見出し語がない単語のうち，形態素解析によって「形容詞」「形容動詞語

幹」とタグ付けされた語について，類語辞典[119]を参照しながら手動で分類語彙表の分類に当てはめ評価語とした．

(5) 得られた評価語群において，実際に評価的な感性とは結びつかないと考えられる分類語彙表の「真偽/こそあど」（e.g.こんな，こういう），「量/程度」（e.g.どんなに，それくらい），「量/限度」（e.g.てんで，まるで）のカテゴリを除いた．

(6) 抽出された評価語は，分類語彙表の相の類の 112 小項目のいずれかに分類されている．さらにこれらの小項目は 23 の中項目に内包される形を取っている．そこで，大局的な分析のために，各ジャンルにおいて中項目に所属する評価語の出現回数を合計し，詳細な分析のために小項目についても同様の値を取得した．

(7) 得られた値について，ジャンル間の比較をするため χ^2 検定を行い，残差分析によって有意に（$p<.05$）多い／少ないカテゴリを求めた．

6.4.2. 全テキストでの結果

4303 種類，93994 語の評価語がテキストから抽出された．ジャンル毎の評価語数は表 24 の通り．また，分類語彙表における中分類毎の合計頻度は表 25 の通り．

表 24　ジャンル毎の評価語数

ジャンル	評価語数	全語数	評価語割合
文学	14796	397279	3.72%
演劇	15219	434814	3.50%
映画	24186	664739	3.64%
ゲーム	39793	904562	4.40%

表 25　中分類毎の評価語数

中分類	分類された語数	語の出現頻度
空間	21	388
形	52	289
経済	67	1813
言語	76	630
交わり	30	152
行為	426	4998
作用	188	2262
時間	264	10394
自然	218	2958
心	1019	20985
真偽	105	3561
身体	37	149
生活	197	2248
生物	8	106
生命	52	706
存在	84	1587
待遇	126	1228
天地	14	67
物質	31	86
様相	560	18493
量	540	13220
力	61	2057
類	127	5617

全データを対象とした場合のχ^2検定の結果を表 26 に示す. 表中, ▲は有意に多い項目, ▽は有意に少ない項目を意味する.

ゲームにおいて多い語はゲームのみが有意に多く（「経済」, 「作用」, 「真偽」等), ゲームにおいて少ない語はゲーム以外が有意に多い（「行為」, 「時間」, 「心」等）という結果となってしまい, 文学, 映画, 演劇の間での特徴が相対的に失われてしまっていることが判る. そこで, 本節では特にゲームに多く見られる「経済」, 「作用」, 「真偽」, 「様相」, 「量」, 「力」について詳細を述べ, 次節においてゲームを除いた場合の結果を分析することとした.

表 26　ゲームを含んだ中項目のχ^2検定結果（p<.05）

中分類	文学	演劇	映画	ゲーム
空間	67	102▲	108	111▽
形	47	71▲	87	84▽
経済	241▽	228▽	339▽	1005▲
言語	107	91	162	270
交わり	32	19	53▲	48▽
行為	871▲	933▲	1622▲	1572▽
作用	371	373	504▽	1014▲
時間	1829▲	1955▲	2971▲	3639▽
自然	530▲	602▲	703▽	1123▽
心	3451▲	3257▽	5792▲	8485▽
真偽	538	464▽	912	1647▲
身体	19	27	71▲	32▽
生活	460▲	465▲	687▲	636▽
生物	36▲	20	42▲	8▽
生命	152▲	180▲	263▲	111▽
存在	307▲	285	498▲	497▽
待遇	212	209	358▲	449▽
天地	20▲	13	18	16▽
物質	25▲	17	29	15▽
様相	2539▽	2571▽	4382▽	9001▲
量	1807▽	1967▽	2700▽	6746▲
力	239▽	280▽	486▽	1052▲
類	896	1090▲	1399	2232▽

6.4.2.1. 経済

「経済」においてゲームが有意に多い小項目は，「経済・収支」であり，【必要】【もったいない】【有利】【必須】【不可欠】【不利】等を含む．これらの表現は，ゲームが攻略するべきものであることから来ている語であり，他のジャンルには見られない特徴である（“システムを細かく解析しながら，どうしたら【有利】になるかを考えていくのは，なかなか楽しい作業だ.”（ゲーム批評 vol.23)）．攻略そのものは批評ではないが，攻略の難易度やゲーム中の要素を評価する上での説明上，不可欠な語である．

また，ゲーム批評では【必要】という言葉が他と比べて極度に多いが（全語数に対する比率で見ても 2.0 倍～2.4 倍ある)，これは，攻略としての必要だけでなく，ゲームに求められる新しい要素などが必要であるといった，改善案の提案的内容として多用されている

ことに由来する（“ですが，それらの持つ「仕組み」の面白さをゲームで伝えるためには，もう少しビジュアル的な気配りが【必要】だったのではないでしょうか.”（ゲーム批評 vol.36)）

6.4.2.2. 作用

「限定・優劣」は，【優れた】【圧倒的】【秀逸】【優秀】等の語を含む．これらの語は評価語としてはありふれた語に思えるが，他のジャンルでは映画で少し多いだけで，出現頻度は低い．また，例えば【優れた】が掛かる語は映像表現，システム等多岐に渡り，特定の語と強く結びついているというよりも，ゲーム批評における一般的な評価語のようである．

「進行・過程・経由」が有意に多い理由は，【スムーズ】という語の頻度が大きいことに依る．【スムーズ】は，主にゲームの操作や画面の遷移がストレスなく行えるか，という評価語であり，操作という観客の主体性が入り込むメディアであるが故の評価語であると言える．

「統一・組み合わせ」で多い語は，【都合】，【そろって】，【併せて】等である．【都合】は“また，会話が最初から全開バリバリで相手の【都合】で進むので，勘所がなかなかつかみにくい．”（ゲーム批評 vol.11）等，プレイヤーの意志と無関係である点を強調する場合に使われるようである．これも，ゲームがプレイという主体性を持つことに起因すると考えられる．なお，【そろって】，【併せて】については，特筆すべき用法は無かった．

6.4.2.3. 真偽

「真偽・是非」では，【実際】，【リアル】，【基本的】，【オーソドックス】等の語が頻出している．【実際】は，“【実際】にプレイしてみると判るのだが，このゲーム，本当に楽しいのはラスト 2 時間のプレイだろう．”（ゲーム批評 vol.56）という文に代表されるように，操作というプレイヤーの関与を実際であると表現しているのが特徴的だった．【リアル】は，“しかも【リアル】なドライビング・シミュレータとして．”（ゲーム批評 vol.33）等，ゲームの映像等が本物らしいかを表現する語で，その他にも【本格的】【正に】など，ゲームの本当らしさを評価する語が多く見られる．一方で，【基本的】や【オーソドックス】は，ゲームの作りを評価するための語で，ゲームの手順の説明（“【基本的】なシーケンスは，情報の取得→人質の探索→脱出を誘導と進む．”（ゲーム批評 vol.23)）や，既存のゲームとの比較（“ゲームの進め方は至って【オーソドックス】．”（ゲーム批評 vol.69)）等の使われ方である．他ジャンルの批評では，【オーソドックス】はどちらかと言えばネガティブな評価の場合が多く，頻度も高くない．ゲームにおいては，その骨格部分（ゲームの仕組み）についても様々な形があり得ることから，こういった評価語が出現し，その上でディティールについて評価が進む構造と考えられる．また，“全般的な印象も，とても【オーソドックス】，正統派.”（ゲーム批評 vol.43）のように，オーソドックスであること自体をポジティブな評価として捉えるような表現も散見された．これは，斬新な部分がなくとも，快適に遊べれば問題ないという観点から出てくる評価のように見受けられる．

「本体・代理」では，【メジャー】【主要な】【典型的な】等の語が多いが，これらも【基本的】等と同様，ゲームの骨格部分を語るための利用が多い．なお，【典型的な】はゲームの物語ついて言及する場合も多く，そちらにも骨格的な部分とディティールがあることが示唆される．

6.4.2.4. 様相

「難易・安危」は，【可能】【不可能】【難しい】などを含み主にゲームの難易度について言及している．「経済」の分類と同様である．

「弛緩・粗密・繁簡」に多い【簡単】，【単純】，【さまざま】等の語は主にゲームのシステムの評価に用いられ，ゲームの要素のヴァリエーションが多い，少ないといった内容に言及するものである（“しかも，こうした【さまざま】な要素が絡んだ戦闘をリアルタイムで行うのであるから，難易度は必然的に上昇する．”（ゲーム批評 vol.36））．

「特徴」に含まれる【オリジナル】【特殊】【通常】等の語もゲーム特有のものである．【特殊】や【通常】はゲーム内部での状態を示すために使われる場合もあるが，他のゲームと比べてその内容に新しいところ，変わったところがあるかないかという評価の視点があることを示すものと考えられる（“確かに前作のシステムを継承してはいるが，桁違いの進化を見せ，もはや【オリジナル】といっても過言ではないレベルに達している．”（ゲーム批評 vol.37））．

「良不良・適不適」については，【良く】という語が多ジャンルと比べて多い．実際に中身を見ていくと，ゲームにおいては【良く】が他の名詞と結合し，様々な感性表現が生じている．例えば，テンポが良い，バランスが良い，効率良いなどである．しかしながら，数の面から言ってこれらの表現だけが多い理由ではなく，単純に様々な文で【良く】という語が多く使われているという事実がある．他にも，ゲームの「良不良・適不適」では【悪い】【うまく】【いけない】などの頻度が高く，映画と同様，直接的な評価が多いことが見受けられる．

6.4.2.5. 量

「量」については「過不足」，「広狭・大小」，「速度」，「多少」，「長短・高低・深浅・厚薄・遠近厚薄・遠近」の 5 小項目が有意に多い．「広狭・大小」，「速度」，「長短・高低・深浅・厚薄・遠近厚薄・遠近」「多少」については，主にゲームの中で展開される画面世界について，【広い】，【速い】，【高い】，（要素・モノが）【多い】などを表現するために使われている．特に「速度」については，ユーザがゲームの進んでいく速度に併せて操作を行わなければならないという分野特有の事情によって多く出現していることが本文から窺える（“最初は展開の【速】さに戸惑うが，慣れれば腕の振るいがいのある骨太なシステムだ．”（ゲーム批評 vol.51），“あまり【遅い】とゲームの流れ自体が止まってしまい場がしらけてしまうが，もう少し入力のスピードは遅くして，プレイヤが考えながらコマンドを入れられるようにして欲しかった．”（ゲーム批評 vol.9））．

「過不足」については，【不足】という語がゲームでは多く見られる．ボリューム不足，

調整不足等，【良く】と同じように要素と合わせて使われていた．

6.4.2.6. 力

【強い】，【弱い】等を含むこの項目は，ゲームにおいては特にプレイ時の有利不利を表現するために使われており，結果として出現が多い．これらは，どちらかと言えば直接的な評価ではなく，ゲームの内容を説明する際に利用されている（“特に，【強い】ボス敵を相手にする場合などは，プレイヤが細かく指示を与えなければ，勝利することは難しい．”（ゲーム批評 vol.37））．

6.4.3. ゲームを除いた結果

ゲームを除いた，中項目でのχ^2検定の残差分析の結果は表 27 の通りである．以下，ジャンルによる出現回数において有意な差がある中項目について，その詳細を特に有意に差がある小項目に注目して述べる．

6.4.3.1. 空間

空間に関する語は演劇で多い．小項目で見ると，その中でも「方向」というカテゴリが有意に多いことが判る．ここに含まれる語としては，【向かって】【斜め】【水平】等がある．【斜め】や【水平】は舞台を形容する語として出てくる（“舞台の約半分が板塀によって【斜め】に区切られ，閉塞した空間が描かれる．”（Theatre Arts vol.39））．一方【向かって】は，目的や未来など抽象的なものに【向かう】がある一方，客に向かって，舞台に向かって等，方向を示す用法が見受けられる．同様の用法は映画にもあるが，相対的な量としては演劇の方が多い．

6.4.3.2. 形

【平板】や【四角い】等，ものの形を示すカテゴリだが，“平板な舞台”（＝単調な演劇）等，ものでなく抽象的な事象を表現することにも用いられているため，単純に演劇が形に関する記述が詳細というわけではない．もちろん，文学よりは映画と演劇のほうが形そのものに言及する場合が多いが，演劇と映画の間では特に大きな違いは無かった．

表 27　ゲームを含まない中分類のχ^2検定結果（p<.05）

中分類	文学	演劇	映画
空間	67	102▲	108
形	47	71▲	87
経済	241	228	339
言語	107	91	162
交わり	32	19▽	53
行為	871▽	933	1622▲
作用	371	373	504▽
時間	1829	1955	2971
自然	530	602▲	703▽
心	3451	3257▽	5792▲
真偽	538	464▽	912▲
身体	19▽	27	71▲
生活	460	465	687
生物	36▲	20	42
生命	152	180	263
存在	307	285	498
待遇	212	209	358
天地	20	13	18
物質	25	17	29
様相	2539	2571▽	4382▲
量	1807	1967▲	2700▽
力	239▽	280	486▲
類	896	1090▲	1399▽

6.4.3.3. 行為

「行為」下には 5 種の小項目があり，映画では「威厳・行為・品行」，「人柄」について有意に多く，演劇では「行為・活動」について有意に多い．

映画の「威厳・行為・品行」で特徴的なのは，【乱暴】【凶暴】【粗暴】【力強い】などの暴力的な語の出現率が他の分野と比べて高いことで，小説や演劇に比べてバイオレンスが表現しやすい映像芸術ならではの観点と考えられる．

同じく映画の「人柄」では【正直】【素直】【優しい】【忠実】【誠実】が多い．このうち，「正直」は他の批評でも，評者の感想を表す文頭語として利用されていることが多く（“【正直】，今はわからない”．（映画芸術 vol.412)），コンテンツの内容評価の要素は少ない．一方，【忠実】については“原作に忠実”など，他作品や歴史的事実の再現性をめぐる評価と

して利用される場合が半数以上を占めている．その他の語も，登場人物を評する語として使われているよりも，制作者の意図や特定のシーンの雰囲気を表する語として利用されている．

演劇の「行為・活動」で多い語は，【激しい】【大胆】【積極的】【強引】等の語である．映画の「威厳・行為・品行」分類と同様のアクティブな要素を表す表現だが，演劇のほうが柔らかい表現が特徴的に使われていることを示すと考えられる．

6.4.3.4. 作用

作用の分類はトータルでは映画において少ない．その 15 小項目を見ると，演劇において「作用・変化」，文学において「近接・接触・隔離」，映画において「進行・過程・経由」，「包み・覆いなど」が有意に多い．

「作用・変化」には，【自然】【自然に】【不安定】【安定】【躍動的】等が含まれる．演劇には身体的な表現が含まれ，また「Theatre Arts」が舞踏などの批評も含むことから，これらの語が特徴的に用いられていると考えられる．

文学の「近接・接触・隔離」は，【ぴったり】という語の出現回数が多いことによる．とはいえ，その出現回数は 16 回で，特に特徴的な使われ方をしているというわけでは無かった．同様に，映画の「進行・過程・経由」，「包み・覆いなど」などもジャンルとしての特徴を表しているような表現は見受けられなかった．

6.4.3.5. 自然

自然には人の五感に関連する 7 つの小分類があり，文学では「音」，「材質」が，演劇では「光」，「色」が有意に多い．

文学での「音」は，他の分野でも多い【静か】という語以外に，【かちかち】【ガサガサ】といった擬音語が非常に多いのが特徴的である．これらは，原文からの引用として登場する場合がほとんどであった．

「材質」も同様で，【柔らかい】のような他の分野でも多い語の他に，【ふかふか】や【ざらざら】等の擬態語が含まれていることに依存する．が，こちらは本文からの引用以外にも，“読み手の気持ちにざらざらと引っかかるそのトゲ”（文藝 2002 夏）といった形で，文章や内容の形容として使われていることも多々あった．

演劇の「光」で多いのは【透明】【暗い】【視覚的】【鮮明】【くっきり】等である．いずれも，単純に舞台の光量を示すだけでなく，それによってもたらされる演劇の雰囲気を表現している点が特徴的であった．

また，【白い】【赤く】等を多く含む演劇の「色」は「光」に比べれば舞台の中の具体的なモノ（“白い顔”）等を示す場合が多かった．なお，「音」，「材質」，「色」については，それらをメディア特性上多用すると考えられる映画では有意に少ないという結果となった．

6.4.3.6. 心

「心」には 20 の小項目が含まれるが，映画では，「快・喜び」を筆頭として「苦悩・悲哀」，「敬意･感謝･信頼」，「好悪・愛憎」，「自信･誇り･恥･反省」，「信念・努力・忍耐」，「心」，

「判断・推測・評価」,「表情・態度」の 9 分類で有意に多く，映画においてこの分類の語が非常に多く用いられていることが示される.

「快・喜び」は映画で有意に多く，他二分野で有意に少ない分類であり，中でも【面白い】という語の出現が群を抜いている（文藝 61 回, Theatre Arts67 回, 映画芸術 239 回）. その他にも,【楽しい】【おかしい】【つまらない】【おもしろい】などの単純な評価の表現が多いことが判る．同様の内容は【好き】や【嫌い】を含む「好悪・愛憎」にも言える.

類似の分類として,【悲しい】【切ない】を含む「苦悩・悲哀」,【恥ずかしい】【みっともなく】を含む「自信・誇り・恥･反省」などの原則として主観的な判断を含む語も映画に多く見られた.

「表情・態度」も映画で多く，これには【クール】,【無表情】,【無愛想】等の表情の形容語が含まれる．が，演劇ではこういった表現は少なかった．これは，アップを利用して役者の表情に迫れる映画ならではの評価要素由来と考えられる.

一方，文学では「意味･問題･趣旨など」,「説・論・主義」が多く，演劇でも「説・論・主義」が多い.「意味・問題・趣旨など」には【重要】,【具体的】,【観念的】等の語が含まれ,「説・論・主義」においては【社会的】,【政治的】,【現実的】等の語が含まれる．いずれもコンテンツそのものを感性的に評価するだけでなく，その背景や関連する言説と紐付けて評価するための語であると考えられる.

なお, 評価として重要な要素の 1 つであると考えられる「感動・興奮」の中分類（【熱い】,【関心】,【劇的】）はどの分野でも使用されており，有意差は無かった.

6.4.3.7. 真偽

真偽の 3 小項目のうち，映画において「真偽・是非」が多かった．これは【実際】【まさに】【リアル】【本当】【正しい】等を含む分類で，様々な使われ方があるため断定はできないが，やはり「心」の部と同様，評者の主観を表す語と考えられる.

6.4.3.8. 身体

「身体」という小項目のみを含むこの分類も，映画で多く出現する.【肉体的】【セクシー】【グラマー】のような，直接人の身体を表現する語である.

6.4.3.9. 生物

生物も 1 つだけ小項目を持つが，こちらは文学で有意に多い．その結果を作り出しているのは【性的】という一語である．とはいえ，この分類に他にも同様の語が多く含まれているわけではなく，どちらかと言えばテーマに依存して出現する語であるため，分野としての評価の特徴とは言いがたい.

6.4.3.10. 様相

「様相」には 7 小項目があり，映画では「趣・調子」「良不良・適不適」，演劇では「弛緩・粗密・繁簡」が有意に多いという結果が得られた.

「趣・調子」には【見事】【素晴らしい】【立派】などの直接的な評価語が多く含まれる.「良不良・適不適」も同様で,【良く】【悪く】【うまく】【ダメ】等．もちろん，映画の中

の要素の描写や，特定要素の善し悪しについて語られるために使われることもあるが，“悪い映画じゃない”(映画芸術 vol.375)のように曖昧な形で使われている例も多く見られる．

演劇の「弛緩・粗密・繁簡」には，【さまざま】【単純】【多様】【複雑】等の語が含まれる．これらはおおむね演劇の全体的な様子を表し，総評的な調子で用いられることが多い（“舞台の上で起きた【様々】なでき事，断片となって引用された夢を縫い綴じるかのようにゆっくりと時間が流れる．”(Theatre Arts vol.24)）．演劇においては，複雑性や多様性が評価の重要な軸としてあることが示唆される．

なお，様相には芸術にとって重要と考えられる【美しい】等を含む「美醜」の小項目があるが，こちらについては，3 ジャンルで特に差異は見られなかった．

6.4.3.11. 量

「量」の小項目では，文学が「一般・全体・部分」及び「軽重」が有意に多く，演劇が「広狭・大小」,「長短・高低・深浅・厚薄・遠近厚薄・遠近」で多かった．

「一般・全体・部分」は【あらゆる】【一般的】等を含む項目で，主に文章の背景や思想に関して述べる際に用いられている．

「軽重」は【重い】【軽い】等を多く含み，文章のテーマや文体について語られる際に用いられている傾向がある．

「広狭・大小」は【大きな】【大きく】等の副詞的な使い方の語以外に，特に演劇では【巨大】や【微細】等の語が多く用いられている．これらは演劇の象徴性などを描写する上で，舞台上の背景や大道具を誇張して形容するために用いられているようである．「長短・高低・深浅・厚薄・遠近厚薄・遠近」も同様で，単純な副詞的使い方に加えて，舞台上の空間的な描写とそれが象徴するものについて述べるために利用されている．

6.4.3.12. 力

「力」は同名の 1 小項目を持ち，映画において多く見られる．【強い】【弱い】【強烈】【凄い】等の語を含んでおり，その使われ方は印象から登場人物の描写まで多岐にわたる．

6.4.3.13. 類

類は 6 つの小項目を持ち，うち「異同・類似」,「関係」,「相対」の 3 つについて演劇で有意に多く,「理由・目的・証拠」については文学で多い．

「異同・類似」は,【同じ】【異なる】【異質】等の語を含む．演劇では複数の役を同じ役者が演じたり，同じ背景で異なるシーンを演出したりと，類似や相似を使った効果や演出が多いようで，それらに関する言及に含まれている．また,「関係」「総体」は【逆】【対照的】【無縁】【直接的】などを含み，全体としてはやはり類似や舞台と現実の関係性などについて言及する際に利用されている．

一方文学の「理由・目的・証拠」は,【もっとも】という語が多いことによって有意差が出ており，特に特徴的な点は無かった．

6.5. 考察

以下の節では，分析結果から得られた各ジャンルにおける有意に多い項目，少ない項目

（表 28）に着目して考察を行う．

表 28　各ジャンルの特徴

ジャンル	有意に多い項目	有意に少ない項目
文学	生物	行為，身体，力
演劇	空間，形，自然，量，類	交わり，心，真偽，様相
映画	行為，心，真偽，身体，様相，力	作用，自然，量，類
ゲーム	経済，作用，真偽，様相，量，力	その他のカテゴリ

6.5.1. 文学の特徴

文学は中項目で見ると有意に多い項目がない（唯一生物は多いが，これは「性的」という一語に依存した結果だった）．加えて，小項目で有意に多かった「説・論・主義」等から考察すると，文学の批評は評者の主観をあまり交えない内容の紹介とその思想や主張について述べていると考えられる．

今回の批評の分析結果から言えば，文学における評価とは，他の芸術から相対的に見れば感性的評価とは関係が薄いと言える．これには，文学は文章すなわち言語と論理から成り立っており，五感に直接訴えかけるものではないため，それを評価する内容も論理的になりがちであるということが理由として考えられる．ただし，これは決して文学が感性的ではないということではなく「小説を読んで感動した」といった場合であっても，その感動の理由を感性語では「語れない」ためであり，文学における感性が高次であることを示唆しているとも言える．

6.5.2. 演劇の特徴

演劇における評価語の使用傾向には大きく 2 つの特徴がある．第一の特徴は，舞台の描写に関する空間・視覚的な語が多いということである．「広狭・大小」,「長短・高低・深浅・厚薄・遠近厚薄・遠近」「空間」「形」等は舞台の上の小道具や大道具の配置や空間の演出方法に関する評価として用いられ，また「色」や「光」などの視覚的な様子を示す語も多い．第二の特徴は，メタファーや多重性に関わる語が豊富ということである．「弛緩・粗密・繁簡」が特徴的だが，空間に関する語などでも，あえて【巨大】という語が用いられるなど，単純な見た目だけでなく，それが象徴するものを示すために利用されているようである．つまり，演劇の評論においては，空間・視覚的なコンテンツを，形容的な隠喩を用いることでその意味するところを拡張し，それを評価しようとする試みが行われていると考えられる．

こういった内容は映画でもあって然るべきだが，演劇においては直接的なリアルな表現が難しいがゆえに，意味の多重性や曖昧性を利用した表現を多用することになり，これが特徴的な評価語の分布につながっていると考えられる．

6.5.3. 映画の特徴

映画においては，【おもしろい】等主観的かつ感想的な評価語が多いという特徴が見受けられた．また，要素の評価としては，「行為・活動」のような人物の表情や感情を評する語と，「威厳・行為・品行」や「力」のような，アクションや映像の迫力といった内容に関する評価が多いことが判る．これらは，映画が人間ドラマや画面の迫力などの一般に広く訴えかける要素を視点として強く持つことを示唆している．

文学や演劇と比較すれば，映画においては比喩ではなく直喩の表現が容易である．従って，我々が日常的に感じることが可能な，情欲や恐怖などのある意味では解釈を必要としない感情を想起させ易い．そのようなメディアの特性が，直接的な評価語の多用を促進したと考えられる．

6.5.4. ゲームの特徴

ゲームを除いた場合には出現しない「経済」にある【有利】【不利】の語は，実際にユーザがゲームとインタラクションし，勝ち負けを決定することに関わっている．ゲーム＝競争の本質は勝ち負けであるため，その内容を説明し評価する際に，これらの語が出てくると考えられる．【強い】，【弱い】を含む「力」，及び「限定・優劣」を含む「作用」についても，同様の理由で頻度が高いと考えられる．

「量」については，ゲームを除けば演劇が多いが，ゲームを含んだ場合はゲームが圧倒的に多い．ゲーム内には，ユーザが操作可能な要素が多く含まれる．演劇や映画，文学では，例えば登場人物といったような要素はあるにしても，背景映像や武器，敵などと細かくは分かれていかない．演劇はそういった部分があるとはいえ，基本的に登場するモノを色々と描写しても，評価にはなりづらいのだと考えられる．逆にゲームにおいては，登場するモノとのインタラクションがあるため，どういったモノが，どのように登場するかを記述することが，ゲームの評価に繋がると考えられる．

「真偽」「様相」の語については，多少の違いはあるが，ゲームがない場合に映画において多く見られる語と共通している部分が多い．これは，映画とゲームが共に娯楽的なニュアンスが強く，評者の主観による「楽しい」「楽しくない」といった情報が批評として重要であるということによると考えられる．

ゲームは映画と同様の直接的に視覚に訴えかけるメディアではあるが，その映像のほとんどは実写ではなくアニメーションである．それゆえ，映画の特徴である「行為」はゲームでは多くない．むしろ，ゲームを攻略するという目的意識をベースにした感性が特徴的に出ている．

6.6. 分析の結論

本章では，分類語彙表を用いて評価語を分類し計量する手法を利用し，各ジャンルの批評の特徴を定量的に示した．結果として見られた傾向は，6.3 で単語レベルの計量比較で見えた傾向と矛盾していないため，分類語彙表を導入したことで対象テキストに対する理解

が歪んだとは考えにくい．マルチジャンル，形容詞となると，分類語彙表などの外部から得たオントロジを使う以外の方法は考えにくい．ジャンルによってそのニュアンスは違うということはあるにしても，形容詞の専門用語は多くはないと考えられるため，外部由来の汎用的なオントロジでも充分に分析が可能であることが示されたと言える．

また，本章での分析においても，5 章と同様の発想でテキストの細かな内容ではなく，ジャンルでまとめた大きなテキストの塊が含有する評価・感性的な傾向に分析の焦点を当て，計量と検定を行った．汎用的なオントロジを使った場合，ジャンル間での差が出ない可能性もあった．しかしながら，結果としてジャンルの間で解釈可能な計量的特徴が抽出できたため，本手法は有効であったと言える．

文学，映画，演劇は互いに相補的に発展してきた芸術ジャンルであり，ゲームもそれぞれから影響を受け，与えている．そして，ビジュアル，物語などの重なる要素で構成されているにも関わらず，評価語の分布は異なった．この違いに大きな影響を与えているのは，各芸術の娯楽としての側面であると考えられる．特に映画やゲームなどでは，その娯楽性に評価の重点が置かれている．これは，対象とした雑誌における「批評」という行為が，その対象そのものがどうあるべき，かくあるべきといった議論ではなく，その対象のどこをどう楽しむ，あるいは楽しめない，と言った内容（いわゆる「レビュー」記事）に偏重しているからであると考えられる．とすれば，それぞれが別の娯楽として成立する以上，観客の感性に訴えかけるポイントが異なるというのは当然であるとも言える．ただし，今回の対象はそれぞれのジャンルにつき 1 雑誌が対象であるため，その点を確定させるためにはさらに対象を増やした分析が必要である．

今回明らかになった評価語の分布は，専門家が注目する評価のポイントであるとも言える．Web 上に多数存在する自然言語の評価を自動的に極性付けする研究は自然言語処理分野で行われているが，多くは語の一般的なポジティブ／ネガティブの意味を取って判断している．本分析の成果を応用することで，各芸術においてフォーカスされる語分類に重みをつけて極性付けを行うことや，登場する語を元にレビューの有用度の自動評価を行うことが可能である．

7. 分析手法に関する比較考察

7.1. 本章の目的

本章では，4 章から 6 章のケーススタディにおいて採用した手法についてまとめ，比較考察を行うことで手法の特徴を明らかにする．

7.2. オントロジに関する手法

本研究では，①オントロジの自動生成，②オントロジの手動構築，③外部汎用オントロジの導入，④オントロジ未使用の 4 パターンについて 3 種類のテキストで検証した．全体の結果を表 29 にまとめた．以下，手法毎に考察する．

表 29　オントロジに関する手法の適用結果

手法		ゲーム批評	河川文化	4 ジャンル批評
テキストの特徴		専門用語の多いまとまったテキスト	同じテーマを持ちつつも複数の専門性を持ったテキスト	4 つの異なったジャンルのテキスト
目的		批評対象概念の理解	河川文化概念全体像の理解	各ジャンルにおける感性の理解
対象品詞		名詞	名詞	形容詞
オントロジの方法	利用しない	単語ベースの分析では，対象となる語が多すぎて把握が難しい		
	自動生成	機能した．	語の共起の特徴がローカルで，機能しなかった．	形容詞については機能しなかった．
	手動構築	手動の場合結果が客観的でなくなる可能性があった．	機能した．	形容詞については困難．
	既存利用	テキストが専門的であり，存在しない．	テキストが専門的であり，存在しない．	分類語彙表が機能した．

7.2.1. オントロジを利用しない

4 章から 6 章のいずれの分析でもオントロジを利用せず，単語ベースで分析することを検討した．しかしながら，上位 20 程度の名詞に着目するだけでは頻度ベースでの網羅性が低く，それを全体的な傾向とは言えないという結論となった．定性的な分析であればともかく，定量的な分析を行う場合には頻度ベースの網羅性は担保する必要がある．

4 章，6 章では，単語ベースでの分析で得られる印象と大きく変わらない結果が得られている．しかし，4 章で得られた「市場」に関する概念は，単語単位であればノイズとして見

逃すか，別の解釈をしていた可能性もある．

一方で 5 章では，単語単位での分析では個別の事象（例えば，【ヤマトシジミ】と【ホタル】; オントロジを使った場合「生物」としてくくられる）について語っている複数のテキストをブリッジすることができず，全体構造の抽出は不可能だったか，5 章で生成された単語の文単位の共起ネットワークのように密度の低いものになったと考えられる．

また，研究を理解・説明するという観点からも，上位の概念で語ることが重要と考えられる．とはいえ，上位概念のラベルは現状では人手で付与するため，「都合の良いラベル」によって実データへのリンクを無視した議論となってしまう恐れがある．それは防がなければならず，まず単語を単位として生のデータを見てみる必要性は高い．

7.2.2. 自動生成

4 章のゲーム批評では，共起情報を利用した自動生成が有効に機能した．ゲーム批評テキストにおいては，暗黙の了解として，そのテキストの内容は「ゲームの批評」というある特定のドメインについて記載したものであるという前提があった．その上で，複数の評者の手には依るものの，ゲーム批評業界において共有されているであろうドメイン知識を抽出したのが 4.4 の分析内容であったと言える. 似たような性質を持つと考えられる「Theatre Arts」の批評テキストを対象として同様のクラスタリングを行うと（頻度上位 105 名詞，頻度 10 以上で 100 名詞のうち最低 25%以上で共起頻度が 0 でない 930 語をパラメータとする），表 30 のような結果が得られる．クラスターは，含まれる語が 2 語以下とならないように最大まで分割している．各カテゴリに含まれる語は共通的な意味を持っていると解釈可能な内容であり，演劇批評を対象としても自動生成は機能していると言える．

表 30　演劇の上位 104 名詞を対象としたクラスタリング

カテゴリに含まれる語	カテゴリの意味
蜷川，野田，俳優，役者，印象，演技，力，場，自身，意味，理解，存在，意識，舞台，想像，空間，観客，自体，イメージ，場所	役者，空間
壁，水，装置，客席	舞台
登場，セリフ，台詞，笑い，芝居，物語，設定，作者，原作，最後，場面，シーン，展開	物語
日常，人々，間，人，関係，自分，人間，死，生活，目，姿，現実，心，ドラマ，話，家族，戦争	テーマ
女性，男，人物，女，手，妻，娘，父	人物
身体，音楽，声，音，記憶，言葉，表現，行為，肉体，動き，ダンス，ダンサー，舞踏	表現
日本，芸術，中心，劇，作品，戯曲，公演，劇場，初演，演出，上演，演劇，東京，劇団，作家，関西	演劇市場
文化，主義，政治，日本人，世界，社会，構造，時代，一つ，現代，方法，歴史，状況	主義

それに対して，5 章のテキストについては，「河川文化」と銘打ってはいるものの，実際は多種多様な関連ジャンルの集合に過ぎない．そこから「河川文化」という共通的なドメイン知識を抽出することは不可能であると考えられるため，オントロジの自動生成は失敗したと考えられる．こうしたテキストの違いは，テキスト中に含まれる名詞の分布の分散によって示されると考えられる．そこで，ゲーム，河川文化，演劇の 3 種類のテキストに対して，頻度 10 以上の単語の長さ 1 に正規化した頻度ベクトルをテキスト全体とテキスト毎に計算し，全体とテキスト毎のベクトル間のピアソンの相関係数を求め，その平均を取った．結果として，河川文化が 0.58 であったのに対して，ゲーム批評は 0.61，演劇は 0.62 と，わずかながら河川文化の値は低く，他のテキストより内容のばらつきが大きい事が示唆される．とはいえその差はわずかであり，テキストの規模や語の頻度分布も異なるため，この数字を指標として自動生成が適当／不適当という事を結論づけることは難しい．定性的に考えれば，内容（特に主題）が多岐に渡るようなテキストについては，自動生成ではなく手動構築の方が望ましいと言える．

6 章の形容詞については，クラスタリングのためのパラメータ選定が困難だった．テキスト中の情報を利用してしまうと，名詞の出現頻度が大きく異なる 4 ジャンルを比較することは難しくなる．また，名詞との共起をパラメータとするクラスタリングでは間接的にジャンル間の名詞を比較することにもなりかねない．

7.2.3. 手動構築

手動構築は原始的ではあるが，その道の専門家がテキストの内容を見て作成し，さらにダブルチェックを行えば最も信頼性が高いと言える．しかしながら，その分野における既存の知識や固定観念が影響を与え，分析が中立的とならない可能性もある．4 章のゲーム批評では，正にそのような懸念から手動構築は避けた．分析方法が比較的単純な計量のため，頻度が大きな単語を移動させるだけで結果の恣意的な操作が行いやすいという懸念もあった．

一方で，5 章の河川文化については，その全体像を誰も把握しているとは言えない以上，その全体像を見越して恣意的にオントロジを作成することは難しい．また，全体像を作成する分析では統計検定処理やネットワーク解析の上で結果を出しているため，望みの結果を得るために試行錯誤をすることは難しい．従って手動構築を問題無く適用できたと言える．

7 章で仮に手動で行う場合には，2000 以上の形容詞を対象として行う必要があった．これは，心理実験等を伴えば不可能ではないが，コストはかかる．河川文化でも頻度を確保するためには大量の名詞を対象に人手の作業が必要であったが，自動構築の手法と同様に 2 段階とし後半は機械化することで，労力を大きく減らすことができた．特に，頻度の少ない語はその一般的な意味合いではなく特定のテキスト・文脈と結びついた形で存在している場合があるため，機械化することで先入観による誤分類を阻止することができる．また，最初の分類もテキスト毎に頻度が高い名詞をグルーピングしていく作業としたことで，対象の選定には恣意性は働いていない．

しかしながら，手動であることには違いなく，自動構築と比べれば恣意性は高いと言える．a)不要単語の選定，b)単語の類似性の判断，c)機械による自動分類のレビューで研究者による恣意性が入り込んでおり，この部分については科学的な正当性も，他の研究者による再現性もない．とはいえ，分類作業は専門家によって行われており，結果については，分類を行った専門家の主観をシステマティックに構造化し，その構造を持って対象テキストを解析したと考えることができる．また，完全に 0 からオントロジを構築した場合と比較すれば恣意性は低く，辞書などの汎用的なオントロジを利用した場合と比較すれば対象に関する知識をベースにしている分だけ内容は深くなる．

7.2.4. 既存オントロジの利用

4 章，5 章では対象に関する充分なオントロジが存在しなかったため，既存オントロジは利用できなかった．6 章については，特定の対象に依存しない語である形容詞を対象としたため，既存のオントロジが利用できた．

形容詞にもジャンル毎に込められた感性の違いはあるはずだが，横断して分析するという観点からは汎用的な辞書を使っても違いを検出できることが判明した．6 章での分析は，言い換えれば，一般的な「感性」から見た時にジャンルがどのように違うのかを調査したものと言える．この一般的な感性とジャンル毎の特化した感性の違いを計量するためには，

また別の方法を検討する必要がある．

7.2.5. 分析に適切なオントロジ

実際にはどのように利用するオントロジを決めれば良いのか．目的に沿って考えるが正道だが，一つの方法としては，まず自動生成を試すのが良いと考えられる．自動生成は試すコストも低く，その結果自体が対象のテキストの性質を知る手がかりになるからである．自動生成により，あまり適合性があると思えない結果となった場合には，手動を試すのが良い．

ここで重要なのは，オントロジには対象テキストに対する適合性がある，と言うことである．当たり前のことであるが，テキスト中に登場する単語をグルーピングする組み合わせは無数にあり，ランダムなもので無かったとしても，そのテキストをより良く表しているオントロジとそうで無いものがある．例えば，一般的な類語辞書に従えば，ゲーム批評に出てくる名詞のほとんどは「ゲーム用語」としてくくられてしまい，巨大なカテゴリが数個できるのみであろう．これでは，ゲーム批評を説明するオントロジとは言えない．この例を逆に考えると，分析に適切なオントロジとは，対象のドメインに対して「解像度の高い」，すなわちなるべく細分化されたオントロジであると考えられる．突き詰めると最上の形態は単語，あるいは文脈を踏まえた個々の単語の用法といったレベルになってしまうが，分割され過ぎると人間には把握できない結果となってしまう．そのため，「適当」な分割を考える（階層的クラスタリングで言えばどの距離でクラスターを切断するか）のが分析者の重要な役割となる．ここには，必ず判断が発生する．その判断は，対象のドメインに関する知識や感覚に基づいて行われるべきである．従って，分析者は対象ドメインについてある程度の知識や直感を持っていることが分析の前提となってくる．

分析者の判断が発生することで，科学性が低減することも確かである．しかし，そもそもテキスト計量分析が対象とするテキストは自然科学的な単一の回答がある対象では無く，分析者の観点によって複数の解釈，回答がある．従って適切なオントロジも複数ある．しかし，計量分析には必ず再現性はあるため，分析者が同様の観点に立っているなら，手法やパラメータに基づく議論が可能である．また，データからオントロジを導き出すことで，データに根ざした議論が可能となり，議論の範囲は無限には拡散しない．従って，個別の知識や判断を必要とは言っても，テキストの計量分析は十分に科学的であると言える．

7.3. 分析に関する手法

7.3.1. カテゴリ単位の計量比較

4 章及び 6 章では，カテゴリ頻度を χ^2 検定の残差分析にかけることで群間の差異を検出する手法を取った．

年代やジャンル等の大きな単位での比較は，0 で述べた通り，微細な意味的内容に定量的に踏み込めないという，単語を計量単位とした分析の欠点を補うためのものである．結論から言えば，大きな単位で集計した語の頻度であっても，群間の意味のある違いが検出で

きるということが 4 章，6 章の結果から言えると考えられる．

また 4 章では，カテゴリ単位での係り受けの集計も行い，カテゴリ毎の違いを見いだすこともできた．組み合わせの数が少なくなってしまう係り受け解析も，カテゴリという大きな単位で数えることで数量的な違いを明示しやすくなる．

この手法は，単純な単語の頻度分析をオントロジのレベルに引き上げたものであると言える．

7.3.2. テキストを単位としたカテゴリネットワーク

5 章では，一般的な単語の共起ネットワークの代わりに，カテゴリのテキスト共起ネットワークでテキスト群の全体構造を示した．

5 章ではオントロジを手動構築しており，その構造は 4 章の自動生成とは異なって，背後にあるのは専門家の認識でありテキストにおける共起の情報ではない．つまり，文脈を捉えていないオントロジであると言える．そこで，5 章ではこのオントロジのカテゴリの登場頻度をテキスト（講演）の単位で計量し，さらに統計的検定をかけることで，テキスト単位で相対的に「良く語られている」カテゴリ（講演のトピックとも言える）を抽出した．句や節，文の構造がどのようになっていて，どのようなことが語られていようとも，単語（群）として言及がないトピックについてそのテキストが語っているということはなく，また仮に語っているとすれば，繰り返しそのカテゴリに関する単語が出てくるはずである．他のテキストでもそのような語が出てくる可能性はあるが，χ^2検定を行うことで他のテキストと比べて有意に頻出するテキストが特定でき，他のテキストでの出現は高い確率で誤差であると判定することができる．これによって，文脈を捉えていないオントロジであってもテキストの中身を捉えることができるようになる．

このテキスト中でのカテゴリの（相対的に多い）出現を共起（「話題が共起している」と言える）と見なし，共起ネットワークを構築した．このネットワークの有効性は，5.3 の事前分析で作成した単語の共起ネットワーク（図 14）と 5.5 の結果のネットワーク（図 18）を比較すれば明白である．前者はほぼ【川】や【水】のエゴセントリックネットワークであり，様々な要素がこれらを中心に集まっているという解釈以外は困難である（もちろんその解釈自体は誤りとは言えない）．また，それら以外とつながっている見えるエッジについても，その多くは KWIC を見ると特定のテキストで頻出している共起に過ぎないことが判る．テキスト全体から出力したものではあるが，テキストの全体，すなわち河川文化を捉えてはおらず，むしろ特定のテキストの内容の断片を全体であるかのように見せる，ミスリーディングな図となっている．一方で，図 18 のネットワークは，そのエッジ 1 つ 1 つが，そのエッジが繋ぐ複数のトピックについて語っている複数のテキストを示している．実際のテキストを調べると 132 テキスト中 119 のテキストが，いずれかのエッジに寄与していることが判る．つまり，図 18 は河川文化というテキスト群のほぼ全体像をマクロな視点で捉えた図と言えるのである．また，5.7 でも触れられているように，図 18 は，要素の関係性から講演を検索する副次的な機能もある．一方で図 14 では，単語の文中の共起をベ

ースとしているため，特定の共起を含む文に飛ぶことしかできない．

5 章の分析は，抽象化して考えると，人間が講演にその内容を示すタグを付与して，そのタグを使って可視化した，と捉えることもできる．この作業をテキストの計量指標を用いて行うことの意義は以下の 3 点である．第一に，人が内容を読んで，分類タグを付与するという方法では，どのようにして分類タグを付けたか，という点について科学的な説明はできない．計量的な指標で説明することにより，その背後にあるオントロジを前提とはするものの，確かにそのテキストにおいて，その内容が語られているということが説明可能となる．第二に，130 講演なら問題なくても，例えば 1000 講演あったときに一人で読んで一貫したタグを付ける作業は労力がかかる．本手法であれば，数が増えても負荷は変わらない．第三に，人の手では付与しにくい，頻度が大きくはないが，相対的には多く語られているようなタグが付く事がある．あるテキストにおいて代表的なタグを付けることは容易だが，2 番めに語られているトピックや，他と比べて相対的に語られているトピックといった内容を人手で付与することは難しい．

この手法は，共起ネットワーク分析をオントロジのレベルに引き上げたものであると言える．5 章の分析では講演をウィンドウとした共起で分析したが，文や章単位の共起を分析することも考えられる．ただし，オントロジは原理的に詳細を捨象する仕組みであり，改めて文に戻っても有意義な結果が得られない場合も考えられる．実際にゲーム批評においてオントロジ同士の文での共起を計測したが，その共起頻度は共起するオントロジ同士のサイズの積と相関した（ピアソンの相関係数 0.94）．つまり，サイズが大きいほど共起しやすいという結果となり，オントロジ同士の特異な関係性は見い出せなかった．このような場合には，オントロジの登場頻度を踏まえた補正が必要であり，河川文化における講演単位で統計的に有意に出現するカテゴリを求めた手順は，このような補正の役割を果たしていたと言えるだろう．

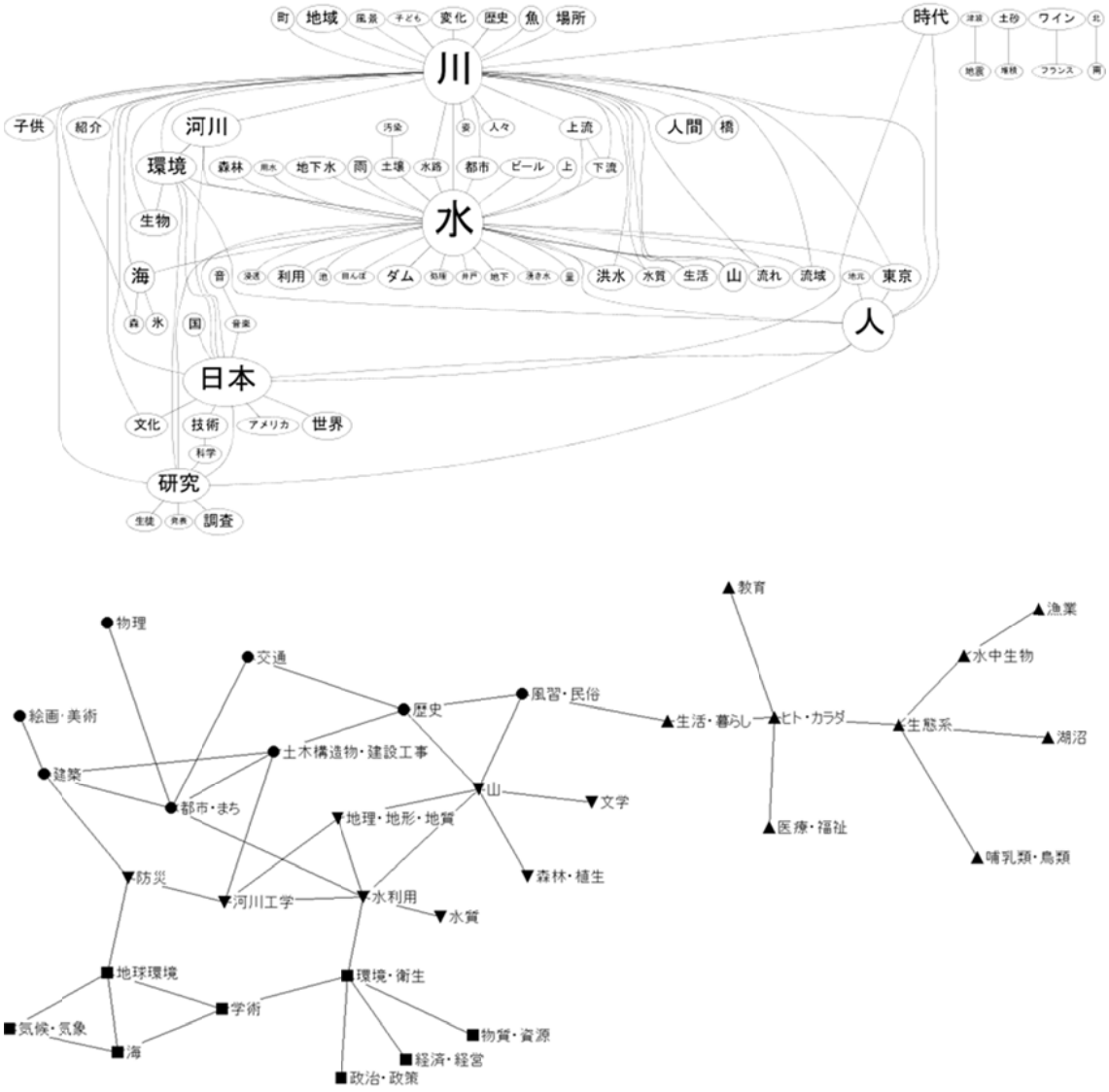

図 14 及び図 18（再掲）

7.4. 結論

本章では，オントロジ生成方法の比較考察を行い，テキストの性質と分析の目的により適切なオントロジが異なることを示した．オントロジの自動生成は演劇批評など他のテキストでも精度が出ると考えられるが，そのようなテキストと河川文化のような自動生成が働かないテキストの計量的な違いは明確にはならなかった．批評では無いテキストなどのデータセットを増やし，比較していく必要があると考えられる．

また，オントロジを利用した 2 種類の分析手法は，いずれも単語レベルで良く行われてきた頻度分析や共起分析をオントロジを利用し，大きな観点で実施したものと言える．このような方法はテキストの詳細な意味を捉えることには適さないが，全体の傾向や構造を把握するためには有効であると言える．

8. オントロジを利用したテキスト計量分析フレームワーク

8.1. 本章の目的

本章では，4 章から 7 章までの内容を踏まえて，オントロジを利用した計量分析のフレームワークを提示する．さらに，そのフレームワークを実際に利用するために開発されたツールである「Text Seer」について説明する．

8.2. フレームワーク

8.2.1. フレームワークの位置付け

本研究では，3.9 で基本方針を決定し，4 章から 7 章にかけてその基本方針の下で生じる欠点を緩和することを試みてきた．従って，フレームワークの前提も基本方針を踏まえて以下の通りとなる．

(1) 意味の分析を行うため，機能語は対象としない．
(2) 機械的な分解による単位は形態素（単語）とする．
(3) 単語レベルでの分析の問題を解決するため，オントロジを導入する．

以上のような前提を置くため，本フレームワークは網羅的なテキスト計量分析のフレームワークとは言い難い．そこで，本フレームワークには OSQTA(Ontology-based Semantic Quantified Text Analysis)という名称を付け，以降ではこの名称を使用する．

8.2.2. 分析フレームワークの全体像

OSQTA は，テーマの設定からデータの収集，分析から考察までをカバーする．フレームワークの全体像を図 24 に示す. 以降の節では, それぞれの内容について詳細を述べていく．

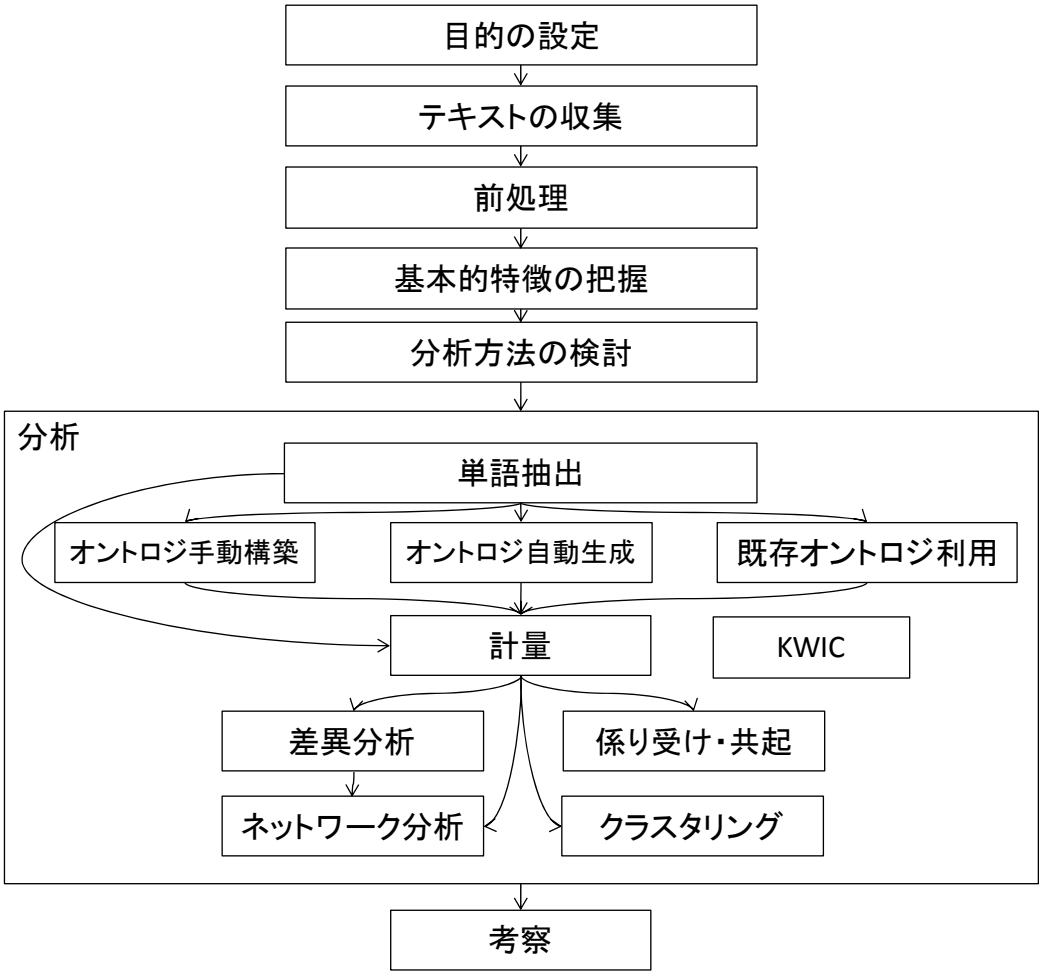

図 24　OSQTA の全体像

8.2.3. 目的の設定

テキストを収集する前に，研究の目的を設定する必要がある．当然だが，目的によって収集するべきテキストが異なるからである．目的を設定するにあたっては，OSQTA の得手不得手とするところを踏まえる必要がある．具体的には以下のような観点について考慮する必要がある．

(1) 大量のテキストを対象とする

大量のデータを対象としない場合，人手による精緻な意味分析，解釈のほうが有意義な分析が行える可能性が高い．客観性の担保という観点はあるが，データが少ない場合，客観的であってもデータのばらつきによる影響が大きくなり，誤った分析となる可能性がある．

(2) 意味的に深い分析は難しい

単語を基礎とした分析では，深い意味の理解は困難である．語の共起を利用することで深層的な意味を擬似的に理解することはできるが，人が読解して理解するレベル

に到達することはない.

(3) 比較によって説明する

テキスト計量分析では，計量値の絶対値で説明・証明できることは少ない．そこで，量の比較が必要となってくる．他の量と比較することで，すなわち相対値によって分析することで，統計的な差異を論ずることが可能になり，科学的な妥当性が増す.

以上の観点を踏まえ，目的に応じた OSQTA の有効性を表 31 にまとめる．目的によっては OSQTA 以外の方法論を求めるべきである.

表 31　目的と OSQTA の有効性

目的	有効性	
横断的，ないしは複雑な対象の特徴を捉える	◎	分野横断的な内容や，専門家でも全体像を把握しきれないような分野を理解する場合，その特徴を抽出することは OSQTA が得意とするところことである（例：河川文化）.
多くのことが判っていない，ないしは証明されていない対象を解釈する	○	対象が未知のものか，科学的な分析がなされていない場合には，深い分析ができなくとも有用な内容が判明する可能性がある（例：ゲーム）.
比較を行う	○	分野間の比較や，年代の比較は計量分析にとって最も証明しやすい内容である．ただし，証明できることは違いがあることまでであり，その違いに関する解釈を証明することはできない.（例：4 ジャンル批評の比較）
既存の理論・解釈を越えた分析を行う	△	既存の理解，理論が高度に組み上がっている領域に適用した場合，それらを打ち壊すほどの確たる結果は得られない可能性が高い. 既存の理論の証明の評価や（例えば，[79]など），既存の解釈に代わるものの探索として利用することは可能である．既存の証明がない理論の証明等に使うことは困難である.
対象について深い意味的な分析を行う	△	先述の通り，単語レベルでは深い分析は難しいが，対象領域に関する信頼性・精度が高いオントロジを外部から与えることができるのなら，深い分析を行うことも可能である.
一冊の本や少量のテキストを理解したい	×	テキスト量が少ない対象に適用した場合，出てくる量的な結果は説得力を持ちにくい．従来通り，テキストを読み理解するべきである．その際に特定の単語やフレーズに注目し数を比較することはあっても，KWIC を利用して説明するものであり，数そのものを論じるべきではない.
テキスト全体の性質の理解を目的としない場合 テキストから何らかの有用な情報を引き出せればよい場合	×	OSQTA はテキストそれ自体が分析・理解の対象であるため，（目的に照らして）意味のないテキストが大量に混じったようなデータから，傾向を抽出したり，何らかの有用な断片を抽出したりするといった目的には沿わない．そのためにはテキストマイニングの技術を利用する方が効率が良い.
分析者の観点・目的に立ってデータを分析・解釈したい場合	×	分析者の特定の意図をもってデータを分析していく場合には，計量内容分析が手法・ツール共に確立しているため，そちらを利用することが望ましい.
言語的な分析を行う	×	OSQTA は，言語的な特徴や言語学的な内容は対象外としている.

8.2.4. テキストの収集

目的を定めたら，テキストを収集する必要がある．テキストの収集にあたっては，分析しようと考えている内容を含む網羅的なテキストであるかどうかが重要となる．テキスト計量分析ではテキストそのものを分析の対象とするため，そのテキストが断片的であったり，関係のない内容が混ざっていたりした場合，有用な結果が得られない可能性が高まる．また，分析の質を高めるためにはテキスト量が重要となるため，その意味でも網羅的な収集が必要である．例えばある詩人の詩の分析を行う場合に，特定のいくつかの詩集を任意に選んで対象にするべきではない（特定の時期に書かれた詩集の分析を目的としている，等，目的に沿えば問題はない）．全数での分析ができない場合には，ランダムサンプリングによる分析とするか，ケーススタディに過ぎないことを自覚した上で考察を行うべきである．

テキストの量が必要という点については再三述べてきたところであるが，具体的にはどの程度あればいいのか．OSQTA では，最終的に単語をいくつかのカテゴリに分割してその単位で計量することを基本としている．そこで，そのカテゴリの出現頻度に対してχ^2検定をかけることを考えると，Cochran の規則[120]より，頻度が 5 未満のカテゴリが，全体の 20%以上になってはいけない．ここでは，単純に，いずれのセルの最小値も 5 以上であるとする．いくつのカテゴリに分けるかでサイズは異なってくるが，本論文のいずれのケーススタディでも，カテゴリの出現頻度の最大値と最小値には 10 倍の差があった．カテゴリの出現頻度の分布は様々だが，ここでは，最小値（5）から最大値（50）まで均一に分布すると仮定すると，全カテゴリを構成する語の出現頻度は，n 個のカテゴリでは27.5nとなる．さらに，このカテゴリを k のグループに分けてグループ間の差異を見ると仮定すれば，全体としては k 倍の量が必要になる．一方で，現代日本語書き言葉均衡コーパス[35]の統計情報によれば，テキスト中に含まれる名詞の割合は 35%，形容詞（形状詞）の割合は 2.7%となっている．ただし，その全てが分析に使える語ではないため，上位 50%の語を分析に利用すると仮定する．以上の仮定を踏まえて，20 カテゴリ，10 グループを想定すれば，名詞の場合で 31420 語，形容詞の場合で 407400 語程度の量が最低ラインとなる．

他，テキスト収集の際の留意点を表 32 にまとめる．

表 32　テキスト収集における留意

留意点	理由
分析に必要なだけの多様性を備える	多様性の中でも最も重要なのは著者の数である．一人の人間が書いたテキストを分析した場合，その人物に関する結果となってしまう恐れがある．特定の人物について分析したい場合は，もちろんその限りではない． その他，コピー&ペーストによる内容の重複がないか，特定の対象に関するテキストのボリュームが突出していないか，等に留意する必要がある．
品質を揃える	テキスト間の品質が異なると，分析が困難になる場合がある．例えば，専門家の書評と素人がブログに上げたブックレビューをまとめて分析してしまった場合，利用される語などの違いで，同様の内容が計量的には大きく異なってしまうという問題が起こりえる．単語を概念の諸元として行うため，言葉が違うグループ（専門度合い，言語，年代等）をまとめて分析することは難しい．ただし，分析をそれぞれのグループ毎に行い，その比較などをすることは可能である．
高品質なテキストを集める	目的がテキストやその書き手ではなく，そのテキストが対象としている領域にあるのであれば，アマチュアよりは専門家，Web 上のテキストよりは出版されたテキストを集めるのが良い．4.6 でも述べた通り，アマチュアのテキストには意味の抽出が難しい断片的なテキストが含まれている場合が考えられるからである．
背後に別の意図があり得るテキストを避ける	例えば雑誌やインターネット上の批評・レビューには，対象を宣伝することを目的とした，いわゆる提灯記事も存在する．こういったテキストは，内容の紹介などに記述が偏っている場合がある[23]ため，特にそういった偏りを分析したいのでなければ避けるべきである．

8.2.5. 前処理

OSQTA では，以下のような前処理を行う．

(1) 機械的なテキストの整形

HTML のタグ，及び文頭文末の空白（スペース・タブ）を除去する．

(2) 正規化（任意）

意味が失われないことが判っている場合，半角カタカナを全角に，また旧漢字を新漢字に変換する．

(3) 文単位の切り出し

1行1文となるようにテキストを修正する(その後の処理プログラムによってはXML等で明示してもよい)．一般的に，文と言えば「。」から「。」（ないしはそれに相当する記号）までである．機械的に処理する場合，テキストファイルから全ての改行を削除し，「。」等の後を改行する，というやり方もある．文の定義を変える場合は，明記する．なお，形態素解析は改行を越えては行われないため，文の途中で改行が入らないように留意すること．

(4) テスト形態素解析

形態素解析を行い語の頻度をみる．そして未知語や頻度の高いエラーを探す．OCRの場合，「“つ”が“っ”になっている」等のエラーがあり得る．全てを見ることは困難であるため，例えば頻度 10 以上と決めて，おかしな単語がないかどうかをチェックすること．頻度が低い単語 2 つを修正したらつながって高頻度の語になる，という場合もあるため，少し低い頻度の語もチェックする．機械的に，国語辞書に載っていない語をチェックするという方法もある．

(5) 手動でのテキストの整形

(4)で明らかになった問題を手動で修正する．正規表現を利用したテキストの一括置換等を活用する．

(6) 同名別品詞語の確認

同じ単語が，形態素解析のアルゴリズム上，別品詞として認定される場合がある．機械的に探査できるので，そういった語において頻度が高い語は統合する．この作業には一般的には恣意性はないと考えられる．

(7) 分割語の連結

形態素解析は，1 つのまとまりとなっている語（連語）についても強制的に分割してしまう場合がある．そこで，連語として頻度が高いものについては，結合した単語として形態素解析辞書に登録し，再度形態素解析を行う．この作業には若干の恣意性があるため，基準等を明記すること．基準となる頻度については，全体の頻度にもよるが，上から 10 ずつ見ていったときに，半分以上が一つの語として認定するには難がある語であれば，その手間を基準値とするやり方が考えられる．

(8) 未知語の登録

形態素解析の辞書にない未知語について，頻度が高い語は品詞を登録する．多くは名詞となるだろう．

(9) 形態素解析

修正及び辞書整備の完了後，あらためて形態素解析をかけて単語の頻度を計量しておく．

8.2.6. 基本的特徴の把握

形態素解析後，以下のようにテキスト全体の基本的な性質を確認しておく．これ自体は通常は手順として明記する必要はないし，逆にこれだけでは OSQTA を使った研究にはな

らない.

(1) 基本統計量を見る

テキストにおける基本統計量を確認する．テキストの規模（テキスト数，語数）は論文において基礎的なデータとして提示する必要がある．基本統計量には表 33 のような値がある．

表 33　テキストにおける基本統計量

統計量	意味
テキスト数	収集したテキストの数
単語数	テキスト中に登場する全単語の出現頻度の合計
語彙数	テキスト中に登場する単語の種類
文数	テキストに含まれる文の合計
平均文長	一文に含まれる単語の数.「。」を打つ回数は書き手によって異なり，平均文長が長いと，文単位での共起をとった場合には意味の薄い共起が多くなってしまう.

(2) 高頻度の語を調べる

注目したい品詞について，頻度の高い語を調べる．累計頻度を調べ，頻度が何回以上の語に注目すれば，対象とする品詞の出現頻度で半数以上をカバーできるか調べる．例えば，図 25 は河川文化のテキストを対象として，その名詞（一般名詞，サ変名詞，固有名詞）を対象として，出現頻度順に語を並び替え，横軸に頻度，縦軸に累計頻度を取ったグラフである（ジップの法則に従っていることが見て取れる）．この場合，頻度 97 以上の語について調べることで，全体の 50%をカバーできる．もちろん，テキストの特性によっては半数をカバーしようとすると出現頻度が非常に小さな値になってしまう場合もあるかもしれないが，頻度が低いと，ノイズに対して弱くなってしまい分析結果を誤る可能性があるため，前処理を徹底して形態素解析等の誤りを減らすか，テキストの量を増やすことが望ましい.

(3) 分析から外すべき語を調べる

高頻度でありながら出現テキストが少ない語は，特定の書き手しか使っていない語だと言える．そのような語は，対象テキスト全体の特性として調べるにはふさわしくない語と言えるので，明記した上で分析対象から外すこともあり得る．逆に，どんなテキストにも何回も出てきて，語の共起の上位がそれで埋まってしまうような語（共起ネットワークを作ると自明となる）も対象から外すことが考えられる．例えば河川文化では【川】,【水】といった語を外したが，これは，このような語が河川文化を語る講演中に出てくることは自明であり，分析の観点からは意味を持たないという判断をしたからである.

さらに，テキスト中に存在するが，その内容に直接関係せずテキストの形態や取得の状況に依存して出現する語で，頻度が高い語も見つけておく．例えば河川文化では，講演であることに由来する【拍手】といった語を分析から除いた．

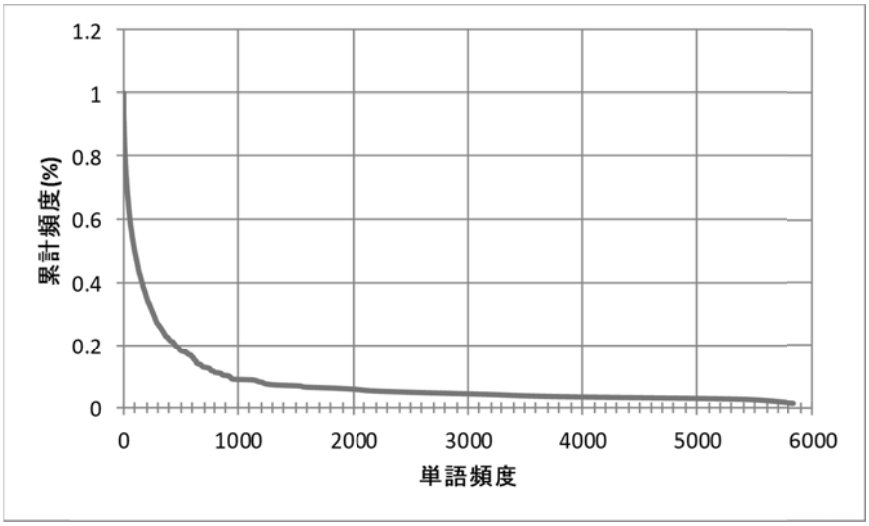

図 25　河川文化テキストにおける単語頻度と累計頻度の関係

(4) 共起ネットワークを見る

テキストにおける語の全体像を見ることは，概要を掴み分析方法を検討する上で役立つ．そこで，(2)で求めた頻度以上の単語を対象として共起を計算し，単語の共起ネットワークを作る．これを見ると，どういった概念と概念のグループがあるのか，概要が把握できる．また，(3)で述べたような不要と思われる語も判るので，それらを除いたネットワークも見てみると良い．参考として，ゲーム批評の単語共起ネットワーク（100 回以上出現する名詞の頻度 100 以上の共起）を図 26 に示す．「攻撃」，「キャラクター」，「作品」など次数中心性の大きな語が複数あり，これらの語がゲーム批評における中心的な概念であることが示唆される．

図 26　ゲーム批評における頻度上位 100 名詞の共起ネットワーク（共起頻度上位 100）

(5) 特定の単語に着目する

テキストの特徴をさらに深く捉える方法の 1 つとして，特定の単語を詳しく調べるというものがある．具体的には，その語の KWIC と共起，係り受けなどを見れば良い．例えば上のネットワークでは，次数中心性が高い語として「プレイヤー」，「システム」，「キャラクター」，「世界」，「作品」等がある．「プレイヤー」に着目して，係り受けのエゴセントリックネットワーク（対象とする語を中心としたネットワーク）を描画すると（上位 30 係り受けまで），図 27 が得られる．頻度最大は「操作」で，次点は「他」である．「操作」はともかくとして，「他」は「他のプレイヤー」つまりゲームを一緒に遊ぶ他のユーザのことで，オンラインプレイなどに関する議論で出てくる．このように，特定の単語に注目することでテキストの中心的な語の特徴について論じる事が可能ではあるが，これだけでは計量からヒント得た定性的な研究となってしまうため，OSQTA においてはこのような方法は分析方法とはしない．ただし，再度考察等で補助的な証拠とする，ないしは解釈のための洞察を得るために，このような手法に立ち返ることはある．

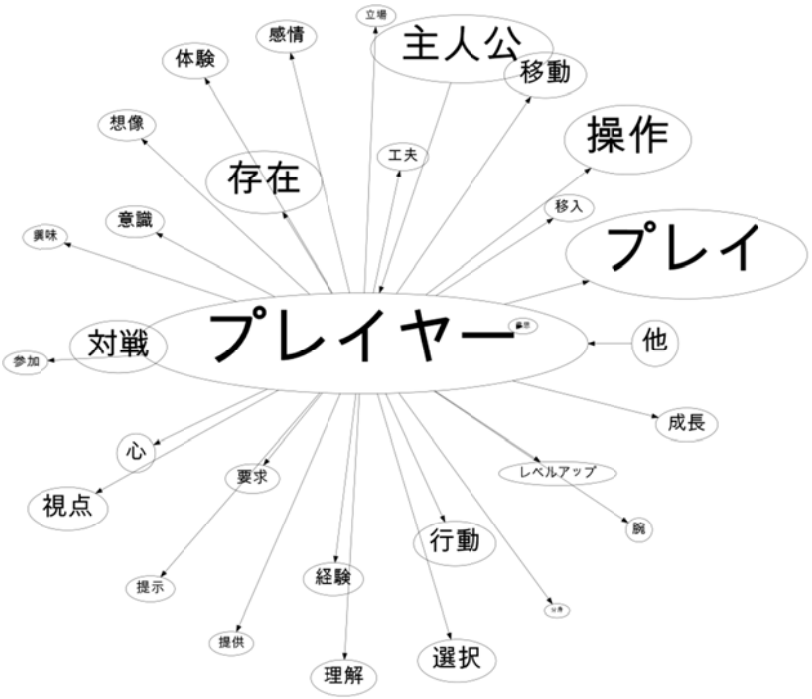

図 27　「プレイヤー」を中心としたエゴセントリックネットワーク

8.2.7. 分析方法の検討

基本的な特徴を把握したら，次に分析方法を決定する．もちろん試行錯誤的に行うことも間違いではない. OSQTA においては, 原則としてオントロジを利用してカテゴリを作り，その計量と，さらに必要があれば差異の分析，係り受け，ネットワーク化と分析，クラスタリングを行う．

まず，オントロジ利用の方針について，表 34 にまとめた．目的とテキストの内容に応じて，方針を定める必要がある．

表 34　オントロジ方針

方針	実施内容	合致する目的・テキスト	留意点
自動生成	テキストにおける語の共起情報から自動的にオントロジを生成する	・特定のジャンルについて語っているテキストで，既存の分析・解釈の枠組みが規定されていないような場合. ・でき上がったオントロジ自体を説明したい場合．恣意性を除きたい場合.	・単語の共起情報等が十分に取得できることが前提になる.
手動構築	機械的なサポートを利用しながら，人手でオントロジを構築する	・複数のジャンルのテキストを，一定の視点で見たい場合. ・既存の分析・解釈の枠組みが規定されている場合.	・オントロジを構築する，対象分野に関する専門家がいることが前提となる. ・でき上がったオントロジは分析の観点に過ぎないため，それ自体を成果にはできない.
既存利用（汎用）	類語辞書等をオントロジとして利用する	・複数ジャンルを見たいが，恣意性を廃したい場合. ・専門家ではなく，一般的な目線で見たい場合．ないしはテキストが専門的でない場合. ・形容詞など，対象ジャンルに依らない特徴が想定される場合.	・分析が表面的となるリスクがある. ・形態素解析の単位と辞書の単位をマッチさせる作業が必要になる. ・辞書にない単語の扱いを決める必要がある.
既存利用（特化）	外的知識を用いて構造化されたオントロジを利用する	・対象のテキストに対して深い分析を行いたい場合 ・既存の解釈が多く存在し，その知識が集積されているテキスト（聖典等）	
利用しない	単語レベルでの分析とする	・単語 1 つ 1 つの意味が強く，単語の種類が少ない場合（詩など） ・テキストが単純な場合 ・オントロジによって意味を曖昧にしたくない場合	・分析が一部だけに注目した内容となるか，数値的に特徴のない発散した形になる可能性がある. ・データの把握・説明が困難になる場合がある.

オントロジを構築し計量した後はそれを元に分析を行う．それぞれの分析方法の特徴と結果を表 35 にまとめる. もちろん，これ以外の様々な分析方法がありえる. いずれにせよ，得られる結果から言えることの限界を念頭に置きながら，各手法を様々なパラメータで試すことで，考察に値する結果が得られる.

表 35　分析方法の選択

分析方法	実施内容	合致する目的	結果
差異分析	群間でカテゴリの量の差異を見る	・テキスト（群）間の違いがどこにあるのか見たい場合 ・そのテキスト群において相対的に特徴的なカテゴリがどれかを示したい場合	特定のテキスト群において，特定のカテゴリが特徴的に多い・少ない
係り受け･共起	カテゴリに対して係り受け，共起する語（群）を調べる	・カテゴリの特徴を詳細に調べたい場合（例えば，カテゴリに対する評価を調べたい場合）	特定のカテゴリを特徴づける語（群）がある・ない
ネットワーク分析	何らかのネットワーク構造を作り調べる	・カテゴリ間やテキスト群間の関係性を調べたい場合	特定のカテゴリとカテゴリ，テキスト群とテキスト群にはどのような関係性がある・ない
クラスタリング	カテゴリ・テキスト群をカテゴリに分ける	・カテゴリやテキスト群をさらにグループ分けしたい場合 ・グループ分けに根拠を持たせたい場合	生成されたグループ

8.2.8. 分析

8.2.8.1. 単語抽出

OSQTA では，要素の最小単位を単語（形態素）とする．ただし，プロセスとしては，前処理の次点で形態素解析を行っているために再度実施する必要はない．

原則として，実際の分析は単語ではなく，オントロジを用いて生成されるカテゴリを単位として行う．

8.2.8.2. オントロジ自動生成

オントロジ生成は，テキスト内の情報を使って自動的にオントロジを生成する手法である．4.4 で取った手法と同様になる．プロセスは以下の通りである．

(1) 対象とする単語を決定する．例えば，頻度合計が 50%となる名詞，等．

(2) 単語を分類する特徴量を決定する．LSA の発想に基づき，原則としては単語共起頻度ベクトルを利用する．助詞等，どのような語とも共起し得る語を特徴量としても有意義なクラスタリングは行えないため，意味語（特に他の名詞）を選ぶ必要がある．また，頻度が低い語は他の語との共起頻度も低く，そういった語をパラメータとすると行列が疎になってしまい，他の頻度の高いパラメータの特徴を相対的に減じさせるため，ある程度の出現頻度を持つ語との共起頻度ベクトルを特徴量とするべきである．

このベクトルの長さは計算量に影響しない．パラメータの選び方で，カテゴリの細部の構造が変わるため，例えばパラメータを変えてクラスタリングを行い，違いを明らかにすることで信頼性が向上する．

(3) 特徴量を使って，階層的クラスタリングを行う．階層的クラスタリングの計算量は$O(n^2)$であるため，対象とする語が多い場合には一般的なコンピュータでは計算が終わらない場合がある上，頻度の低い周辺的な語にカテゴリの構造が左右されてしまう．そういった問題を解決するため，4.4 で行ったように，頻度の大きい一部の単語を階層的クラスタリングし，残りの単語はでき上がったカテゴリにベクトルの内積等で比較して分類していく方法を採ることが可能である．

(4) でき上がった階層を特定のレベルで切り，カテゴリを得る．この判断は恣意的にならざるを得ないため，基準を決めて（例えばカテゴリに含まれる単語が 1 つにならない）どこかで切るしかない．あるいは，切るレベルを変化させ，その後の分析にどのような変化が生じるかについて見ることも有効である．

(5) 生成されたカテゴリの中身を見て，理解のための名称を付ける．この名称は恣意的なものであるため，名称に引きずられてカテゴリを構成する語を忘れないようにすること．

(6) カテゴリの上位構造を意識する．共起によるカテゴリ分析で得られる集合は本質的には明確に切れるものではなく，集合同士を比較する際には，その上位構造についても検討しなければならない．例えば，集合 A と B が異なる特徴を示しているがカテゴリ上は近い位置にあったらどうか．特徴の内容にも依るが，分析ミスの可能性もある．

生成されたオントロジは，それ自体が対象テキストさらにはそのテキストの対象に内在する概念構造の一端を示していると考えられる．従って，得られたそれぞれのカテゴリについて解釈を行い，そのような構造が生まれてきた理由について考察することが可能である．その際には，そのカテゴリにおいて代表的（頻度が高い）な語の KWIC を見ることが重要になる．

8.2.8.3. オントロジ手動構築

オントロジ手動構築は，機械の支援を得ながら，なるべく恣意性を排しつつも人手でオントロジを作り出す手法である．5.4 で取った手法と同様になる．プロセスは以下の通りである．

(1) 最初に人手で分類する語を，テキストの頻度を基準として選択する．基準としては，頻度ベースで全体の 10%を占める語，等．

(2) 選択された語を，文の KWIC 等を見ながらグルーピングする．グルーピングに当たっては，階層的クラスタリングと同じように似た語をグルーピングしていくといった手法が考えられる．グルーピングのプロセスや根拠を記録しておく必要がある．さらには，複数人でグルーピングを行えば，その妥当性を示すことができる．

(3) 人手で構築されたグループに，8.2.8.2 の(3)と同様に共起ベクトルの内積で他の単語を

分類していく．これで，オントロジのカバーする単語の割合を高める．

(4) 自動的な分類の結果を，人手でレビューして修正する．恣意性が生じるが，むしろ分析の精度を高めるために実施するという判断もある．この場合にも，複数人でレビューして確認することで，妥当性を高めることができる．

(5) それぞれのグループに名前を付与する．

この手法は，対象となる領域が横断的だが共通する概念があるような領域には有効である．一方で，文脈や目的を排して半機械的に作っているとは言え，分析者の観点に立った分析であることに違いはない．そのため，この分類結果そのものに対して考察を行うことに OSQTA の観点からは意味はない．ただし，このように作られたオントロジは専門家の頭の中にある認知マップの一端であるということは言えるため，このオントロジを元にしたその後の分析は，恣意的ではなく，「専門家が行う（であろう）テキスト解釈を機械的に行った」と位置付けることができる．

8.2.8.4. 既存オントロジ利用

分類語彙表や Word Net のような既存のオントロジ（シソーラス）を使う方法で，6.4 で取った手法と同様になる．これは手法として複雑なことはないが，以下のような問題を解決する必要がある．

(1) オントロジに含まれる見出し語が，対象としたい語をカバーしていない場合には，オントロジを拡張する必要がある．例えば 6.4 では，分類語彙表に登場しない語については，他のシソーラスを引いて分類語彙表に分類した．他にも共起ベクトルを使って分類するなどの方法もあるが，恣意的にならない方法を選択する必要がある．

(2) 形態素が，オントロジの語と対応しない場合がある．例えば分類語彙表は単語よりも長い内容を含んでいる場合があるため，文字列マッチで対応する内容を探した．

(3) 利用するオントロジの妥当性を示す必要がある．分類語彙表などは，それがベースとするコーパスから汎用的であるため，どのような分野に使っても不適切ということはない（結果が出ない可能性はある）．一方で，より専門的なオントロジを使うという場合にはそのオントロジを使う理由や，オントロジとしての信頼性について説明する必要がある．

8.2.8.5. 計量

いずれの方法にせよ，オントロジができ上がれば，後はそれぞれのカテゴリに含まれる単語を計量し，合計することでそれぞれのカテゴリの計量ができる．基礎的な情報として，それぞれのカテゴリの全体量を明示する．

もしオントロジを利用せず，単語をベースに計量を行う場合には，この時点で対象とする単語とその選定方法について明示しておく必要がある．

8.2.8.6. 差異分析

差異の分析は，OSQTA で最も有力な方法であり，またその適用範囲も広い．

特定のグループを作り，そのグループにおける各カテゴリの頻度を計量する．この結果

は二次元のマトリックスになるが，それに対してχ^2検定及び残差分析をかけることで，有意に多いと言えるセルを見いだすことができる．例えば，ゲーム批評では，テキストを時代毎の群に分けて，各カテゴリの量の変遷を見た．また河川文化では，テキストを単位としてカテゴリの多少を判定し，そのテキストにおいて有意に語られているカテゴリ＝トピックを明らかにした．さらに批評の分析では，ジャンル間における違いを明らかにするためにこの手法を用いた．また，テキスト以外の単位，例えば章の単位等を用いることも可能である．

分析結果からは，特定のカテゴリ（概念，評価等）が，特定のテキストや分野で多いということを主張することが可能になる．これは，そのオントロジを真とし，統計的検定における有意水準を真とする限りにおいて，明確に真と言える内容である．

8.2.8.7. 係り受け・共起

3.7.2 で述べた通り，係り受けは，日本語としての意味をより正確に捉えるには有用なツールであるが，その頻度を計量しづらいという問題があった．しかし，カテゴリを設定することで，そのカテゴリと係り受けする語を調べることが可能となり，その問題を軽減できる．例えば 4.4.3.2 では，そのカテゴリと係り受けする形容詞，及び選定した動詞を調べ，それぞれの群に特有の評価語を明らかにしたり，形容される割合を示したりしている．さらに精緻に行うなら，形容詞とカテゴリの共起頻度のマトリックスを作り，χ^2検定を行うことで，統計的に有意な語を見つけることも可能である．ただし，カテゴリでまとめているとしても係り受けは絶対数が少ないので，データが多くなければ，係り受けの代わりに共起を使うことも考えられる．

係り受けの分析は，カテゴリの意味やそれに纏わる感覚，動作を明らかにすることができると言える．しかし，テキストの全体よりも詳細な意味を捉えた分析となるため，はっきりした特徴がなければ，考察等も曖昧になってしまうという問題点がある．

8.2.8.8. ネットワーク分析

3.7.4 でも述べた通り，ネットワーク分析では様々な手法が可能である．また，何をエッジとして何をノードとするかによっても分析できることが大きく異なる．表 36 に，いくつかのエッジとノードの取り方の例を挙げる．

表 36　ネットワークの例

ノード	エッジ	ネットワークの意味	目的
単語	単語の文中での共起	ぼんやりとしたテキスト全体の意味.	概観の把握，中心となっている単語の理解.
カテゴリ	カテゴリの文中での共起	単語よりはっきりとしたテキストに含まれる概念構造.	中心的な概念の把握，カテゴリどうしの近さ.
カテゴリ	テキスト中での共起（有意に多いカテゴリを対象とし,1テキスト 1 回以上は数えない)	テキストで関連づけて語られている概念の全体構造.	どのような概念については併せて語られ，どうような概念の組み合わせは注目されていないかを知る.

なお，共起をエッジとして重みを付ける場合，その重みは頻度ではなく，そこから計算する相互情報量やシンプソン係数などを用いて付与することもある．相互情報量は式 2，シンプソン係数は式 3 で表される．

$$d_i = log \frac{N|X \cap Y|}{|X||Y|} \quad (式\ 2)$$

$$d_s = \frac{|X \cap Y|}{\min(|X|,|Y|)} \quad (式\ 3)$$

（ただし|X|は単語 X の頻度，|X∩Y|は単語 X と単語 Y の共起頻度，N は全単語頻度）

語の共起数は出現回数が多い語が単純に大きくなってしまうが,「出現する時には必ず一緒に特定の語と出てくる」語，つまり非常に関係性が強い語だとしても，絶対的な頻度が低ければ小さくなってしまう．相互情報量やシンプソン係数では分母に語の頻度が入るため，このような問題を解消できる．一方で，シンプソン係数は全体頻度が低い語ほど値が大きくなってしまうという性質があるため，この係数を使って上位 n 位までをネットワーク化という処理をしてしまうと,頻度が 1 の語のネットワークができてしまう恐れがある．従って，閾値を設けて頻度が低い語は無視するか，ネットワーク自体は共起頻度が高い組で作成し,エッジの重みのみシンプソン係数にする,といった対応が考えられる[121][122]．ただし，どのようなネットワークを構築するとしても，その特徴については見た目ではなく，中心性（3.7.4 を参照）や分割アルゴリズムの適用など，あくまで数値的・アルゴリズム的な観点で分析することが重要である．ネットワークは本来多次元的な構造体であり，

ネットワーク描画アルゴリズムはそれを二次元に射影しているに過ぎないため，そこからその実体を正しく捉えることはできない．

ネットワークの分析では，カテゴリとカテゴリの関係性や，テキストとテキストの関係性を示すことが可能である．ただし，示すことができる内容の詳細については，ネットワークの作り方と，その分析の方法に依存するため，一概には言えない．

8.2.8.9. クラスタリング

階層的クラスタリングはオントロジ生成で利用したが，さらにテキストをクラスタリングすることも可能である．例えば，カテゴリの頻度を特徴ベクトルとして，テキストをクラスタリングする，等である．目的としては，テキストにおけるグループの発見が挙げられる[79]．形容詞との共起頻度ベクトルを特徴量として，名詞のカテゴリをクラスタリングする，などの手法も考えられる．これによって，同じような感性で語られる概念とは何か，といった内容を明らかにできる．

クラスタリングは，グループを作ろうとすると，恣意的にどこかの階層で切らなければならないという欠点がある（階層的でないクラスタリングでも，いくつのグループを作るかをあらかじめ示すことが必要である）．この恣意性を減じつつ同様の結果を得るためには，因子分析を用いることも 1 つの方法となる[49]．

8.2.8.10. KWIC

KWIC 自体は分析の方法ではないが，どのような分析を行っても必ず立ち返るべき方法である．なお，キーワードだけでなく，語の共起や係り受けなども KWIC のように元の文脈を見ることが可能である．

8.2.9. 考察

分析結果を基に考察を行う．重要な事は，計量結果に根ざした，ないしは計量結果そのものに対する考察とすることである．例えば，特定のカテゴリの出現量が多かったからと言って，そのカテゴリについてのみ論じるべきではなく，そのカテゴリが相対的にどのような位置付けで，全体の中でどのような意味を持つから，どうである，といったような内容を論じるべきである．もし，論じたい一部の内容を見つけ，それに注目して計量の範囲を超えて論じたいという目的であれば，それはテキストマイニングの方法を用いた，しかし特に科学的とは言えない方法であって，OSQTA のスコープではない．

8.3. フレームワークの意義

OSQTA は，高度な自然言語処理技術や，対象テキストに関する背景知識の複雑なオントロジを必要としないシンプルなフレームワークである．そのため，このフレームワークから計量的に述べられる内容は，同様にある程度シンプルな内容になる．そのため，目的の設定でも述べたが，このフレームワークが適しているのは既存の深い分析が行われていない分野や，大規模なテキストを対象とした研究である．そういった研究であれば，このフレームワークを利用することで十分な成果を生み出せることは，ケーススタディから実証

されていると考える．また，2.4.2 で述べた先行事例も多くは本フレームワークの範疇に収まる研究である．これらを含めると，以下のような対象・分野・目的での成果が期待できる．

(1) 対象テキスト

批評テキスト[51][54][85]，感想・ブログテキスト[50]，プロパガンダテキスト[46]，講演テキスト[98]，文芸・思想テキスト本体[48][49]

(2) 分野

認知科学[47]，感性工学[46]，メディア論[50][85]，文学[48][51]

(3) 目的

対象テキストが内包する概念構造の抽出[47][48][51][54][85][98]，評価に関する性質の抽出[46]，抽出した性質の変化分析[48][54][85]

もちろん，OSQTA は万能ではなく，テキスト計量分析の真の目的である対象領域の認知的理解や機械によるテキストの理解というレベルに，このフレームワークだけでは力が及ばないということも明白である．しかしながら，そのような目的は一足飛びに実現できるものではなく，OSQTA レベルの研究を積み重ね，オントロジを集積し，言語と認知の関係がテキストとしてどのように表れるのかを徐々に理解して初めて到達できるものである．そのための発展的な議論を行うためにも，個別の研究がその方法論に関して毎回，「MeCab で形態素解析し…」といった基本的な説明をするのではなく，OSQTA と類似の研究について引用するだけで，本論に入れる状態が望ましいと考えるものである．

8.4. フレームワークを利用するためのツール

本研究では，フレームワークを利用するためのツールである「Text Seer」を開発した．本節では，このソフトウェアについて概略を述べる．

8.4.1. Text Seer とは

Text Seer は，OSQTA の基本的なプロセスを GUI を利用して行うことができるアプリケーションである．ただし汎用的なテキスト分析ツールも指向しているため，OSQTA のスコープではない分析機能も一部実装されている．OSQTA のプロセスと，Text Seer の機能の対応を表 37 に示す．次節以降で，機能の詳細を説明していく．

表 37　OSQTA と Text Seer の対応関係

OSQTA のプロセス・手法	Text Seer の機能
前処理	テキストの整形
	テキストの形態素解析
－	テキストの選択
基本的特徴の把握・KWIC	単語解析
	共起解析
	ネットワーク解析
	記述統計量
	係り受け解析
－	n-グラム
オントロジ自動生成	クラスター分析
（オントロジ生成語の）計量	セット解析
－	多言語対応
その他の分析	ライブラリとしての利用

8.4.2. Text Seer の機能

8.4.2.1. テキストの整形

Text Seer が扱えるテキストは日本語と英語のテキストである. どちらの言語であっても, Text Seer は改行を 1 文の区切りとして扱う．共起解析を行う際にはこの文の単位で解析を行うため，改行が文の途中に挟まっていると正しい結果が得られない．そのため，そのような形式になっていないテキストは整形作業を行わなければならない．このテキストの整形のために，いくつかの機能が用意されている.

(1) 検索置換

まず，全テキストを対象として，一括検索，一括置換を行うことができる．正規表現を利用することも可能である．検索・置換の結果は表形式で一覧される．この一覧結果をクリックすることで，テキストの該当箇所を表示することができる.

(2) テキストエディタ

この該当箇所の表示など，実際のテキストの表示を行う簡易テキストエディタが Text Seer には付属している．簡易とは言え，文字コードの指定，正規表現をサポートした検索置換，検索結果のハイライト表示など，テキストエディタに必要な最低限の機能は備わっている．そのため，Text Seer で解析を行っている際に見つけた誤植などを別のアプリケーションを開いて修正する必要はなく，その場で修正することが可能になっている.

(3) 修正マクロ

テキストの修正を行う際，頻繁に行われる作業はマクロとして登録してある．マクロは「句読点での自動改行」，「空白行の削除」，「文頭の空白文字の削除」の 3 種類である．通常のテキストであれば，このマクロを利用することで 1 ステップの作業で利用できるフォーマットに整形することができる．

8.4.2.2. テキストの形態素解析

整形したテキストは，まず形態素解析を行う．形態素解析には 3 つのオプションが存在する．ユーザ辞書の使用の有無，未知語の自動推定をするか否か，そして英語の場合の大文字小文字の区別である．

(1) ユーザ辞書

ユーザ辞書とは，形態素解析で単語へと分解する際に，特殊な単語などを登録しておくことで正しく分解するための辞書である．辞書には，単語名と読み，品詞などを登録する．現時点では，日本語の形態素解析のみユーザ辞書に対応している．

(2) 未知語の自動推定

未知語の自動推定とは，形態素解析アプリケーションの辞書にない固有名詞などの単語を前後等から自動的に品詞判別する機能である（MeCab の機能）．精度は 100%ではないが，このオプションを選択することで未知語はなくなる．

(3) 大文字小文字の区別

これは，大文字の単語と小文字の単語を区別するかどうかのオプションである．このオプションを利用する場合，大文字と小文字は無視され，同じ単語として認識される．ただし，文頭などで大文字になっている場合には，このオプションとは関係なく同じだと認識されるし，また形態素解析アプリケーションの辞書に該当単語が載っている場合にも，正しく認識される．このオプションが有効なのは，固有名詞などの未知語が含まれており，その語が大文字始まりであったり，小文字始まりであったりする場合に限られる．

形態素解析の後，係り受け解析が行われる．係り受け解析に関しては，ユーザが設定できるオプションはない．その後，形態素と係り受けの情報は，今後の検索を素早くするためにデータベースに登録される．データベースに登録されて始めて，単語のカウントなどが行えるようになる．登録が終わると，形態素分解が正常に行われているかを確認することができるようになる．Text Seer には，この確認のための機能が 3 つ付属している．なお，以下の 3 つ機能は，ユーザ辞書の関係上，日本語テキストでのみ機能する．

(1) 未知語リスト機能

この機能では，未知語の自動推定オプションを使用しなかった場合に「未知語」として分類されてしまった語のリストを表示する．そして，一括してユーザ辞書に登録することができる．この作業を行うことで，自動推定ではうまく推定できなかった語を正しく認識させることが可能になる．

(2) 分割された語をつなげる機能

専門用語などは，複数の単語に分割されてしまうことが多い．この機能では，そのような分割されたと思わしき語を検索し，1 つの語としてユーザ辞書に登録することが可能である．

(3) 同名別品詞の登録機能

同じ単語であるにも関わらず，別の品詞として認識されてしまった語を抜き出す機能である．これは，自動指定オプションを有効にした際に良く起きる問題で，細かい名詞の分類などが同一単語で異なってしまう場合がある．この機能では，そのような同じ文字列からなる単語の一覧を作成し，該当単語をユーザ辞書に登録することができる．

データベースへの登録を行うことで，以降，そのデータベースの内容から全ての解析を行うことが可能になる．そこで，このデータベースの内容への接続を保存することで，今後形態素解析などの手間なく同じデータにアクセスできるようになる．Text Seer は「.proj」という拡張子のプロジェクトファイルとしてこの情報を保存することが可能である．保存したプロジェクトファイルは後で開く他に，Java で記述したプログラムから開くことも可能であり，コマンドライン上で繰り返し形態素解析を行う手間を省くことができる．

8.4.2.3. テキストの選択

Text Seer では複数のテキストが含まれるコーパスを一括して扱う．以降で説明する解析においては，基本的にはコーパスを単位とした解析が行われる．しかし，テキスト間の違いを見る場合など，対象を限定したい場合もあり，Text Seer はこれに対応している．GUI の場合であれば，画面左側にテキストの一覧とチェックボックスが表示され，このチェックボックスがチェックされているテキストだけを対象として解析を行うことが可能である．このテキストの選択に関して，2 つの追加機能がある．

(1) グループ機能

複数のテキストをグループとして扱い，グループ単位で解析の対象とするか否かを選択できる機能である．GUI であれば，グループ単位のチェックボックスが表示され，ここをクリックすることでグループ内の全テキストを選択したり，非選択にしたりすることができる．グループはファイルに保存したり，そこから読み込んだりすることも可能である．

(2) 比較機能

テキスト同士を比較したい場合に使う機能である．通常のテキスト選択画面に追加してもう 1 つテキスト選択画面を表示し，全ての解析をそれぞれのテキスト選択に基づいて行うことができる．解析の結果も 2 つの画面で表示されるため，テキスト間での違いを比較する際に有効である．

8.4.2.4. 単語解析

単語解析では，テキスト中に登場する 1 つ 1 つの単語に注目することができる．その最も基本的な機能は，登場する単語を表形式で一覧する機能である．表には単語，その品詞，

出現頻度，全単語数に占める出現割合，いくつのテキストに出現しているか，及び，単語のテキストにおけるユニークさを表す指標である TF・IDF が表示される．表は各指標によるソート，文字列検索による単語の絞り込みが可能である．表は品詞毎に存在し，また，一般名詞，数名詞，固有名詞などいくつかの副分類を統合した「名詞」などの上位分類も存在する．品詞に関わらず全単語を表示したり，逆に複数の多様な品詞を指定して表示したりすることも可能である．

表の各単語をクリックすることで，その単語に関するさらに詳しい情報を閲覧するための新しいウィンドウが開く．このウィンドウには，以下のような 4 つの機能が備わっている．

(1) 単語の出現一覧（KWIC）

選択した単語が，テキスト中のどこ（何行目か，何文字目か）に出現するのかを一覧表示する．また，その前後の文を表示する．出現をクリックすることで，テキストエディタが開き，対象の行がハイライトされる．

(2) 単語の出現をグラフィカルに表示

選択した単語がテキスト中のどこに出現するのかを，画像で表示する．テキスト全体を横長のバーで示し，単語の登場位置を縦線で示す．テキストの位置による単語の出現に偏りがあれば，それが視覚的に明らかになる．

(3) 単語の共起検索

選択した単語と共起する単語と，その共起回数を一覧にする．さらに，検索された「選択語と共起する単語」同士の共起も検索することが可能である．後者の検索により，選択語周辺の単語がどのような関係を持っているのかを明らかにすることができる．検索した結果をネットワークとして表示することも可能である．前者の結果の場合には，選択語を中心とした放射状のネットワークが表示されるだけであるが，後者の場合には選択語を中心としつつも，複雑なネットワークが構成される．さらに後者の場合，オプションで選択語をネットワークから除くことができる．この場合，選択語の影響がない状態で周辺語の関係を可視化することができるため，周辺語の特性がより明確になる．

(4) 単語の係り受け検索

選択した単語と係り受けする語を，共起検索と同様に一覧にする．「係る」語のみや，「受ける」語のみをカウントすることも可能である．また，共起検索と同様にネットワーク表示することも可能であるが，こちらは共起検索ほどのオプションは持たない．

8.4.2.5. 共起解析

共起解析では，条件で指定した単語の共起を検索し，その共起回数を一覧表示する．共起は複数の条件でその範囲を指定できる．

(1) 語の出現頻度閾値

テキストにおける出現回数がこの閾値を越えている語のみを検索する．登場回数が

少ない語は複数の語と共起している可能性が低いため，頻度が高い語に注目した方がより多くの語との関係性を見ることができる．また単純に，頻出語ほど重要である可能性も高い．

(2) 品詞の制限

名詞や動詞といったおおまかな分類と，一般名詞，固有名詞などの細かな分類で合わせて，日本語で 77 種類，英語で 38 種類の品詞指定が可能である．

(3) 範囲の指定

通常，共起は一文における共起を見る．しかし，範囲を指定することで，2 つ以上の連続する文における共起を調べることが可能になる．範囲を増やすことでより多くの共起を得られることになるが，同時にノイズとなる意味が薄い共起も増加する．

(4) 単語別設定

場合によっては，出現頻度が高い語であっても不要である場合，逆に出現頻度が低くても重要な語であることがある．恣意的に語を除いたり加えたりすることは科学的な正当性を低下させるが，論理的に説明可能な場合にはそのような操作を行っても良いと考えられる．Text Seer では，たとえ閾値や品詞の範囲に含まれていなくても，特定の単語を強制的に共起の候補に加えたり，逆に外したりすることができる．このように強制化された単語はリストとして一覧表示することが可能であり，それを保存したり，読み込んだりすることもできる．

検索された共起をクリックすると，その出現を一覧表示する別ウィンドウが開く．一覧にはその位置と，共起している部分のテキストが表示される．さらにそれをクリックすることでテキストエディタが開き，該当箇所がハイライト表示される．

8.4.2.6. 係り受け解析

係り受け解析機能は，基本的には共起解析と同様の機能を持つ．係り受けを調べる単語の範囲についても同じような条件を指定することができる．ただし，係り受けは文という単位で完結しているため，文の範囲を増やすことはできない．また，一覧から係り受けの組みをクリックすることで，その出現を一覧できるウィンドウが出現する．この一覧から見たい文章を指定することで，その文章全体の係り受け構造をグラフィカルに表示する機能もある．

8.4.2.7. n-グラム

文字レベルと単語レベルで n-グラムを検索することができる．具体的には，n 文字，あるいは n 単語の連続をひとつのグループとして捉え，その出現回数を数える．単語の n-グラムの場合には品詞設定によって，特定の品詞のみで構成された連続を検索することができる．品詞を指定しないと，名詞と副詞のパターンなど，意味的には重要でない n-グラムが頻出するため，このオプションは名詞の連語などを見つけ出したい場合に有効である．

8.4.2.8. 記述統計量

表 38 に示すような記述統計量を，テキストごとに表示する．また，選択中のテキストお

よび全テキストにおける平均も表示する.

表 38　記述統計量

記述統計量	説明
文数	テキスト中の文の数.
単語数	テキストの全単語数．語数ではなく，のべ単語数.
文字数	テキストの文字数.
平均文長	文の平均的な長さを，単語単位および文字単位で表示.
文字種別	全角ひらがな，カタカナ，アルファベットなどの，文字種の全体に対する割合.
語彙数	テキスト中に出現する単語の種類数.
TTR	語彙数÷語数．TTR が高いほど，様々な語を使っていることになる.
台詞率	『「』で始まり，『」』で終わる部分の割合（英語の場合は”）．カギ括弧の始まりと終わりが不正確な場合，異常な値が出る.

8.4.2.9. ネットワーク解析

共起解析，および係り受け解析の結果をネットワークとして表示する機能である．共起や係り受けの出現は，二種類の単語の組みと，その出現回数という形式で表すことが可能である．ここで，単語をノード，共起や係り受けをエッジとし，エッジの重みを出現回数とすることで，共起や係り受けをネットワークとして表現できる．なお，共起の場合はエッジに向きがない無向グラフ，係り受けの場合には「係り→受け」という方向に向きがある有向グラフとして表現している.

ネットワーク描画のために設定しなければならないパラメータは，上位何組までの共起を描画するか，である．例えば 100 件の共起が検索された場合，そのすべてを描画すると 100 のエッジを持つネットワークとなる. エッジが多いネットワークは時として見辛いため，共起回数が一定回数以上のペアのみをネットワーク化する，などのように閾値を決めて範囲を狭める必要がある．その時に利用するのがこのパラメータである．また，ネットワーク描画のためのアルゴリズムは 5 種類が利用可能であり，それぞれ異なる見た目のネットワークを表示する．ただし，これはあくまで見た目の問題であり，ネットワークの構造が異なるわけではない．その他，エッジの相対的な長さや太さ，ノードの大きさなどをカスタマイズして見易い表示にすることができる．ズームイン／アウトも可能である.

ネットワークを解析するために，ノード数，エッジ数，平均密度，平均エッジ数，単語間の距離，各種中心性を計算する機能がある．Text Seer が持つネットワーク解析機能は限定されているため，さらにネットワークのクラスタリングや，部分ネットワークの抽出などの高度な解析を行うためのエクスポート機能も搭載している.

8.4.2.10. 多言語対応

Text Seer は現在日本語のテキストと英語のテキストに対応している．日英共に利用可能な機能はほぼ変わらないが，英語の形態素解析を行う場合にはユーザ辞書を利用することができない.

8.4.2.11. クラスター分析

共起情報を利用して，単語のクラスタリングを行う機能がある．ここから，オントロジを生成することも可能である.

8.4.2.12. セット解析

作成されたオントロジを読み込んで，その単位で頻度や共起を数えることができる機能.

8.4.2.13. ライブラリとしての利用

Text Seer はその多くの機能を Java のライブラリとしても提供している. すなわち, Java のプログラムから利用可能である．この機能によって，研究によって詳細が異なる様々な分析を 0 からプログラムを組むことなく利用できる.

最も基本的な機能は，プロジェクトファイルとして保存したコーパスのオブジェクトとしての提供である．テキストや文，形態素分割された単語といったデータを利用することができる．単語データにはその出現頻度などの情報も含まれている.

次に，共起解析や係り受け解析，n-グラムといったアルゴリズムも利用可能である．プログラムから利用する場合には，共起を検索する単語の範囲など，アルゴリズムのパラメータを自由に指定することができるため，より柔軟な検索を行うことが可能になる.

最後に，Text Seer の GUI を構成している様々なコンポーネントも呼び出すことができる．例えば品詞の設定画面や，単語の一覧表，係り受け解析のネットワークなどが利用できる．これによって，オリジナルのアルゴリズムの結果をグラフィカルに表示したり，GUI でないと操作や条件の変更が煩雑になる部分に利用したりすることが可能になる.

8.4.3. 既存のアプリケーションとの比較

研究者がフリーで使うことができる日本語テキスト分析用のソフトウェアとしては，KHCoder を代表として，MTMineR[123]，TTM[124]等がある．ただし，KHCoder を除いては, 統合的なテキスト解析のプラットフォームとしては Text Seer と比べて圧倒的に機能が少ないか，R 等の知識が必要なソフトウェア（RMeCab[125]）となっている.

KHCoder と比較すると，Text Seer 独自の機能としては，係り受け解析の機能やセット解析ができる点が挙げられる．一方で，KHCoder はコーディングのための機能や R を利用した複数の多変量解析の機能を有している．このように，Text Seer と KHCoder は基本的な機能は共通しているものの，目的とする分析の主眼が異なるため，機能の詳細やユーザビリティには様々な差異がある.

なお，商用のソフトウェアが多数あるが，研究者に無料で提供されているものは存在せず，またどちらかと言えばテキスト計量分析ではなくテキストマイニングを指向するものであるため，これらとの詳細な比較は行っていない.

8.4.4. 実装の詳細

8.4.4.1. 実装言語

Text Seer は Java で実装されている．ただし，いくつかの Java ではない外部アプリケーションに依存している．

8.4.4.2. 外部アプリケーション

形態素解析は外部プログラムとして MeCab 及び Pentree Tagger を利用している．また，係り受け解析には CaboCha 及び Malt を利用している．

ネットワークは Graphviz を利用して描画する．5 種類の描画方法を利用できるようにしており，そのパラメータの一部も GUI から操作できる．さらに，ネットワーク解析を Text Seer 内で行うため，JUNG[126]というライブラリを利用している．JUNG はネットワーク構造を扱い，解析し，表示するためのライブラリである．中心性の計算や単語同士の距離を求めるためのダイクストラ法アルゴリズムなどはこのライブラリから利用している．

8.4.4.3. アーキテクチャ

Text Seer のプログラムは図 28 に示すような構造を取っている．Java で記述されている部分は，GUI 部分，データと関数部分，そしてデータベース部分である．GUI は Project クラスなどのデータ型を通じてデータベースにアクセスするようになっている．データベースを直接操作することはない．

MeCab 等の形態素解析外部アプリケーションは，Text クラスから外部アプリケーションとして実行され，テキストファイルを通じてデータのやり取りを行っている．ライブラリの jung は，GUI 上でネットワーク表示を行う機能と，ネットワークを解析する機能を担っているため，GUI 側とデータ側の両方から呼び出せる．GraphViz は単体で実行され，生成された SVG ファイルを batik ライブラリを通じて呼び出し，画面に表示している．

Word クラスなどのデータ部分は，それ単体で CUI からの呼び出しが可能である．また，GUI の一部のパーツも，CUI から呼び出すための静的関数を実装している．

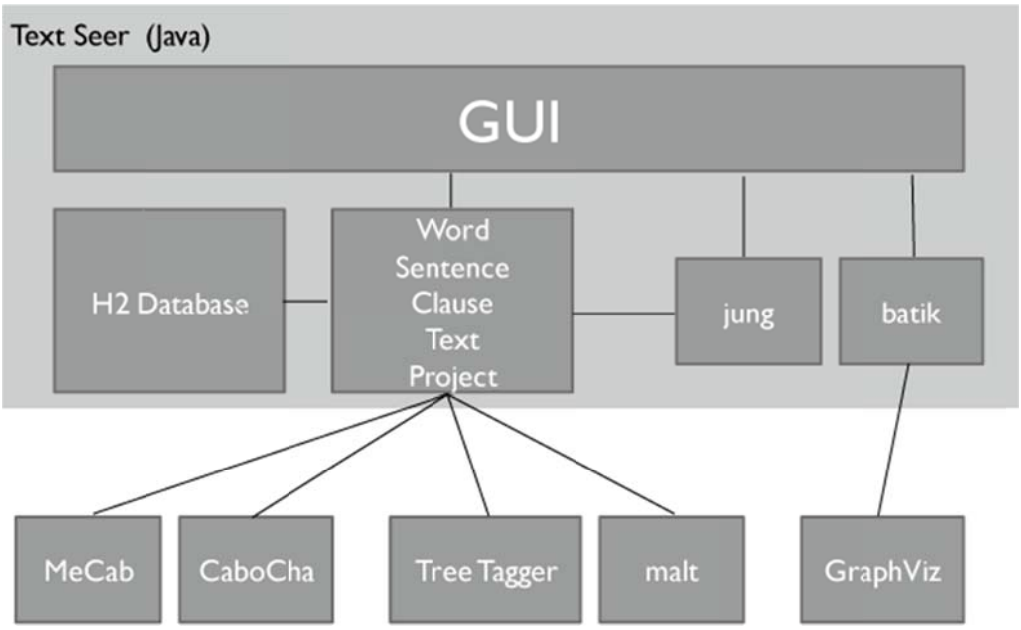

図 28　Text Seer のアーキテクチャ

8.4.4.4. データ型

Text Seer は内部で，図 29 に示すようなデータ型を持っている．これらのデータ型は全てクラス（WordType のみ列挙子）であり，ライブラリとして利用する際にアクセス可能である．

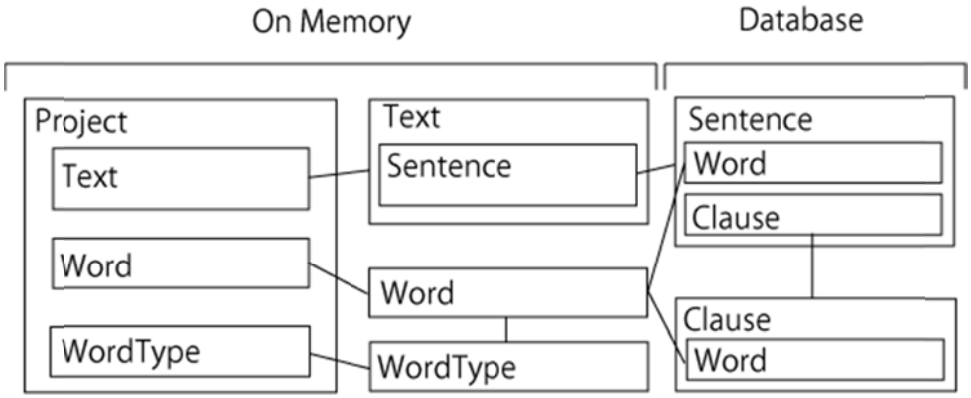

図 29　Text Seer のデータ構造

(1) Project

Project クラスは，現在扱っているテキスト全体を表す．内部には各テキストを表す Text クラスへの参照と，全テキストを総計した単語の情報が格納されている．また，データベースへのコネクションを保持するのも Project クラスであり，他のクラスは Project クラスを通じてデータベースにアクセスする．テキストのグループに関する情

報や，どの品詞が有効であるか，なども保持している．

(2) Text

Text クラスはテキストファイル 1 つを表すクラスである．形態素解析のための処理などを行う．単語がテキスト中のどこにあるかの情報は量が膨大であるため，データベースに Sentence や Clause の形で格納され，必要に応じて Text クラスから検索される．

(3) Sentence

文 1 つを表すクラスである．Sentence はそれを構成する単語の列と，節（Clause）の列を情報として持つ．Sentence の情報はデータベースに格納されており，共起を検索するときなど，必要に応じて Sentence インスタンスとしてメモリ上に展開する．

(4) Clause

文節 1 つを表すクラスである．英語の場合には，文節はないので単語 1 つを表す．その文節を構成する単語列を保持している．Sentence と同様，通常はデータベース上にあり，必要に応じてインスタンス化される．

(5) Word

単語を表すクラスである．Word クラスは，特定の単語 1 つにつき，1 つのインスタンスのみが存在する．つまり，名詞の「ゲーム」という単語を表す Word インスタンスは，メモリ上に 1 つしかない．データベース上の文や文節を表すエントリーには，このインスタンスの ID が記されている．複数のテキストを扱う場合，延べ語数は膨大になるが，異なり語彙数はそこまで大きくならないため，同じ単語を同じインスタンスで表すことで必要なメモリの量が激減する．Word クラスに対象とするテキストの範囲を設定することで，その範囲における出現頻度などの情報を引き出すことができる．

(6) WordType

品詞を表す列挙子である．各単語は，その単語を示す文字列と，品詞を示す WordType への参照で構成されている．

8.4.5. インターフェイスの実装

Text Seer のインターフェイスは，様々な解析をシームレスに行えるように設計した．実装には Java の Swing ライブラリを利用している．Text Seer のメインウィンドウは図 30 のようになっている．

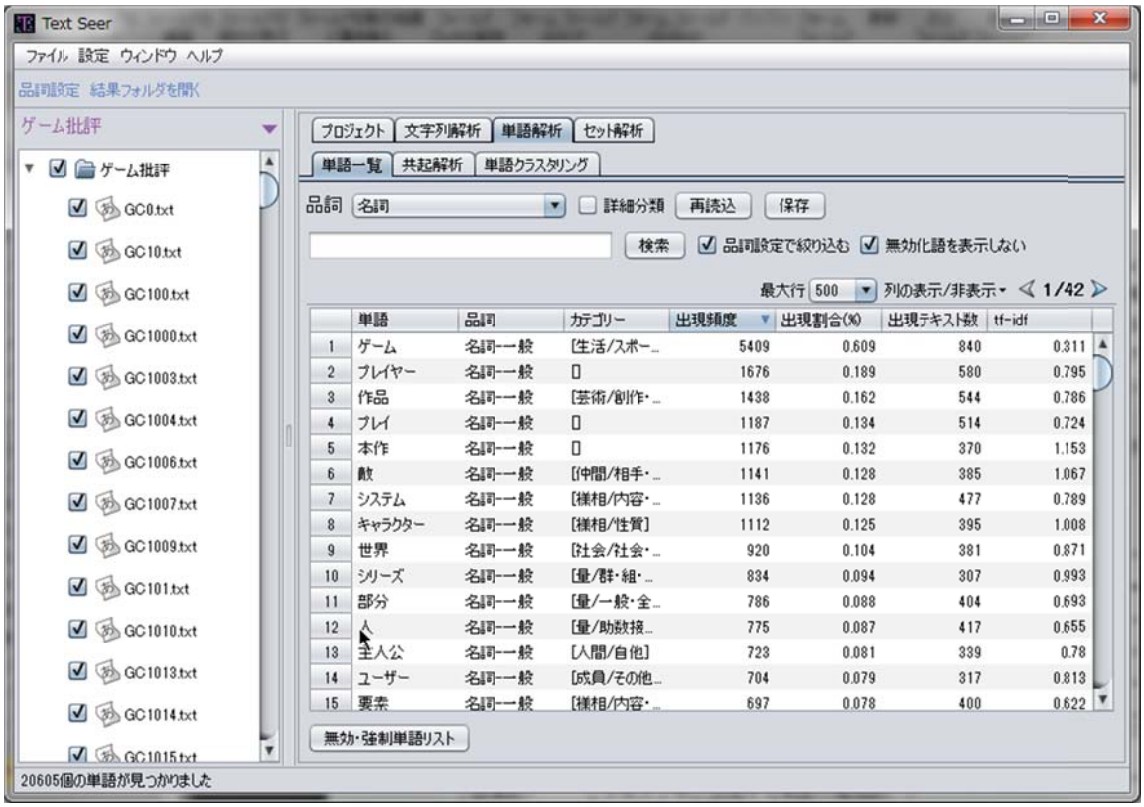

図 30　Text Seer のメインウィンドウ

左側に表示されているのがテキストのリストである．ここで，どのテキストを解析対象にするかを選ぶことができる．どの解析を行っているときでもこのリストは表示されているため，特定のテキストについてだけ見てみたい時に有用である．

右側にあるのが解析の画面で，データを読み込み，修正して形態素解析等を行う「プロジェクト」タブ，文字列レベルでの分析を行う「文字列解析」，単語レベルの分析を行う「単語解析」，オントロジレベルの分析を行う「セット解析」の 4 つのタブに分かれ，さらにその下に個別の機能が配置されている．

単語解析の結果や，共起解析の結果をクリックすることで，この基本画面とは別のウィンドウが出現する．これらのウィンドウには対象をさらに詳細に見ていく機能が備わっており，最終的にはテキストエディタを開いて原文を表示することが可能になっている．このように，注目したい単語などにクローズアップしていくような GUI になっているため，解析中に気になったデータを，その場で即座に追求していくことができる．これは，他のアプリケーションや別の画面に切り替えなければならなかった従来のアプリケーションと比べて，作業をシームレスにする．

表として表示される全てのデータを CSV として保存することができる．その際，保存のパス等を尋ねられることはなく，プロジェクトフォルダ内に日付のフォルダが切られてそこに保存される．上部の「結果フォルダを開く」ボタンを押すとこのフォルダが開き，表

計算ソフト等を使った分析へとつなげることができる．ファイルにもタイムスタンプが押されるため，データの管理が行いやすい．

9. おわりに

9.1. 本研究の成果

本研究では，テキストを分析する手法であるテキスト計量分析において，単語とそのオントロジをベースとしたフレームワークである OSQTA をケーススタディから構築した．以下では，4 章から 8 章までの成果を改めてまとめる．

9.1.1. オントロジの自動生成による概念カテゴリ計量

4 章ではゲーム批評を対象としたケーススタディにおいて，批評の対象となる概念を名詞のオントロジ自動生成手法によって得ることに成功した．また，抽出した概念のカテゴリに対して計量分析，感性語との係り受けの分析，さらに時系列的な計量比較分析を行い，1)ゲーム批評において「新規性」が観点となること（p.53），2)既存のゲームやシリーズの前作等との比較が評価における重要観点となること（p.54），3)時代の変遷と共にグラフィックスに対する評価が中心ではなくなり，オンライン要素など多様な遊び方に関する評価が増えてきたこと（p.57），などの新たな知見を得ることに成功した．

これらの成果から，オントロジの自動生成による概念カテゴリの抽出と計量分析が，ゲームのような専門的な用語を多く有するテキストでは有効であることを示した（p.62）．

9.1.2. オントロジの手動構築と概念構造抽出

5 章では，河川文化を対象としたケーススタディにおいて，専門家の持つオントロジを手動構築する手法と，複雑で多彩な内容を含む一連のテキスト全体から，テキストの話題となるトピックの全体像とその関係性を抽出する手法を開発した．

かつ，示された関係性を分析することで，河川文化を構成する要素において，ローカルな問題とグローバルな問題は，「川」や「山」などの土地と，「建築」などの工学を介してつながるということが示された（p.76）．また専門家があると想定する概念同士の関係性がない（少ない）ことから，河川文化における問題点も抽出された（p.78）．さらに，作られたネットワークを探索することで，実践に必要な示唆を含む講演を見つけ出せることが示された（p.83）．

これらの成果から，オントロジ手動構築の手法と，テキストを単位とした有意に多い概念の共起ネットワークが，多様な概念を持つ対象について有効であることを示した（p.83）．

9.1.3. 既存オントロジの利用による形容語計量比較

6 章では，文学，映画，演劇，ゲームという 4 ジャンルを対象とした分析により，外部から導入したシソーラスを使って評価語をジャンル単位で分析する手法を検討した．

分類語彙表を利用する手法は既存の研究[46]と同様だが，4 つのジャンルが異なるテキストの形容語に対してもその手法が適用可能であり，ジャンル毎の差異が計量的に抽出可能であることを示した．具体的には，1)文学は形容語レベルでの特徴はない（p.100），2)演劇は比喩の感覚を利用した評価が多い（p.100），3)映画は直感的・感覚的な語の利用が多い

（p.101），4)ゲームは他のどのジャンルとも大きく異なって，それを攻略するという意識の下での有利不利の評価や，画面上に現れる詳細な内容についての評価が多い（p.101），という特徴があることを明らかにした．

以上から，既存のオントロジを利用した評価語の分析が複数ジャンルを対象としたテキストで有効であることを示した（p.102）．

9.1.4. フレームワークとツール

7 章では，4 章から 6 章までのオントロジ導入及び分析の手法を比較考察し，各手法がどのように有効あるいは無効であったかをまとめた．

8 章では，それを踏まえ，オントロジを利用したテキスト計量分析フレームワークである OSQTA を提案し，説明した．OSQTA は，単語ベースでの分析の問題点を解決するため，ケーススタディで得られた各種の手法を組み込むことで，広範な分析を可能としたフレームワークである．フレームワークには，目的の設定からデータの収集，分析方法，考察の方法までを網羅することで，どのような研究者でもフレームワークを活用し研究が可能となるような内容とした．さらに，フレームワークの全てをカバーするわけではないが，その手法を実践可能なテキスト計量分析のツールである Text Seer を開発した．7 章では，その機能の概要及び実装の詳細についても説明した．

OSQTA と Text Seer を利用することで，テキスト計量分析に関する知識を持たない研究者でも計量分析を適切に行うことが可能になった．これにより，ケーススタディと目的を同じくする研究や類似の研究において，精度の高い研究を可能にしたことが本論文の成果である．

9.2. 今後の展望

9.2.1. フレームワークの整備と発展

本論文で提案した OSQTA は，特に分析手法については本研究で行ったケーススタディを主とし，既存のテキスト計量分析で利用された全手法を網羅しているわけではない．こういった手法についてもその目的，成果，限界を明らかにすることでフレームワークに組み込み，フレームワークを拡充していく必要がある．

本論文で対象としたオントロジは，単語をカテゴリに分類しただけのシンプルなシソーラスである．しかしながら，対象について深い分析を行うためには，語と語の関連性や，上部構造などを持ったオントロジが必要となり，そのようなオントロジをどのように用意するかについて検討する必要がある．その 1 つの可能性として，筆者はオントロジの共同構築プラットフォームを開発した[127]．これは，オントロジを複数人の専門家が相互レビューしながら構築するためのプラットフォームで，このプラットフォームを使って明治期の貴族及び写真師の人物オントロジを構築する実験も行われ[128]，またそのオントロジを利用した分析成果もあるものの[129]，プラットフォームとしては参加ユーザのモチベーション等について解決できず，汎用的に問題を解決できるレベルには達していない．本研究

の成果である自動生成の手法と人手の手法等の成果を組み合わせ，また Open Linked Data 等の世の中の流れに合わせ，作られたオントロジを流通させることも指向しつつ，深い分析に必要なオントロジの構築方法を模索していく必要がある．

9.2.2. フレームワークの妥当性を示す検証

OSQTA では，様々な場面で研究者による判断を迫る．例えば，オントロジ自動生成におけるパラメータの閾値や，共起ネットワークにおける閾値などである．こういったパラメータについては，これまでの筆者らの研究の成果による経験値を本論文にも記載したが，それが「正しい」のかについて証明はされていない．実際に証明することは難しいが，様々なテキストデータや汎用的なコーパスを対象として分析を行うことで，テキスト，言語のデータとしてどのような値が「普通」で，どのような値が「外れ値」かということについて，知見を積み重ねていく必要がある．

また，OCR や形態素解析には必ず誤差があるため，そのような誤差が分析全体に与える影響についても，エラーを故意に混在させる等の方法で検証し，また複数のパラメータで分析を行ってその共通を自動的に取るようなアルゴリズムを構築し，研究者の判断によらず現れる事象と，研究者の判断によって揺れ得る事象を区別することを可能にしていく必要がある．

9.2.3. ツールの拡充

Text Seer は現時点でも OSQTA の全ての分析手法を単一のソフトウェアで網羅はしていない．研究者の利便性を考えると，全てを Text Seer 上で実行できることが望ましいため，機能の拡充を行っていく必要がある．また，現在の Text Seer は Relational Database をデータストアとして使っており，テキストの量が 10M を越えた辺りから，処理が遅くなる．大規模なデータに対応するためには，Relational Database は効率が悪いため，全文検索等で利用される転置インデックスなどのデータ構造を導入し，数百メガ以上のテキストデータにも対応できるようにしていく必要がある．

謝辞

最初に，指導教員であり，私がこの道を歩むきっかけを下さった徃住彰文教授に深い感謝を捧げます．先生の「認知科学講究」の授業が無ければ，今の私はありませんでした．

徃住研究室の諸先輩方にも感謝を．徃住研究室で先輩方が行った研究の積み重ねが，本論文の基礎となりました．特に村井源先生にはテキスト計量分析の基礎から実践まで多くのことを丁寧にご指導頂きました．ありがとうございました．

第 5 章の共同研究者である、神戸市立工業高等専門学校の高田知紀先生には，河川とそれに纏わる取り組みに関して豊富な知識を共有していただきました．また，文芸評論家の石堂藍氏には，批評の分析について深い洞察を頂きました．対象ドメインの専門家の知見を得ることなしには，この研究は成立しなかったと思っています．ここに感謝を捧げます．

そして，本論文の主査を引き受けて頂いた価値システム専攻の猪原健弘教授を初めとする審査員の先生方にも御礼申し上げます．様々な御指摘を頂いたことで，この論文はあるべき姿にたどり着けました．

最後に，博士課程への進学と長きに渡った論文の執筆を暖かく見守ってくれた家族に，心より感謝します．

参考文献

[1] 天野真家, 石崎俊, 宇津呂武仁, 成田真澄, 福本淳一:『ITText シリーズ 自然言語処理』, 情報処理学会編, オーム社, 2007.

[2] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto:"Applying Conditional Random Fields to Japanese Morphological Analysis", Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.

[3] 黒橋禎夫, 長尾眞:「京都大学テキストコーパス・プロジェクト」, 言語処理学会第 3 回年次大会発表論文集, pp.115-118, 1997.

[4] 黒橋禎夫:日本語形態素解析システム JUMAN version 3.5, http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN (2014.10.27 閲覧)

[5] 工藤拓, 松本裕治:「チャンキングの段階適用による日本語係り受け解析」, 情報処理学会論文誌, Vol.43, No.6, pp.1834-1842, 2002.

[6] Fillmore, C. J.:"The Case for Case, In Universals in Linguistic Theory", Bach and Harms eds., Holt, Rinehart and Winston, New York, 1968.

[7] 河原大輔, 黒橋禎夫:「高性能計算環境を用いた Web からの大規模格フレーム構築」, 情報処理学会 自然言語処理研究会, Vol.171, No.12, pp.67-73, 2006.

[8] Quillian, M. R.:"Semantic Memory", In Semantic Information Processing, Minsky, M. eds., MIT Press, 1968.

[9] Minsky, M.:"A Framework for Representing Knowledge", In The Psychology of Computer Vision, Winston, P. eds., McGraw-Hill, New York, 1975.

[10] Schank, R. and Abelson, R.:"Scripts, Plans, Goals and Understanidng", Erlbaum, 1977.

[11] Lenat, D.:Cyc Project, http://www.cyc.com/ (2014.10.27 閲覧)

[12] Miller, G. A.:Word Net, http://wordnet.princeton.edu/ (2014.10.27 閲覧)

[13] Isahara, H., Bond, F., Kanzaki, K., Uchimoto, k., Kuribayashi, T., Utiyama, M., Cook, D., Sumida, A., Kuroda, K. and Torisawa, K.:日本語 Word Net, http://nlpwww.nict.go.jp/wn-ja/ (2014.10.27 閲覧)

[14] 村田真樹, 一井康二, 馬青, 白土保, 井佐原均:「過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査」, 言語処理学会第 11 回年次大会論文集, pp.77-80, 2005.

[15] Marcus, M.:Penn Treebank Project, http://www.cis.upenn.edu/~treebank/ (2014.10.27 閲覧)

[16] 前川喜久雄:「日本語話し言葉コーパスの設計と実装」, 平成 15 年度 国立国語研究所公開研究発表会論文集, pp.1-8, 2003.

[17] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R.:"Indexing by latent semantic analysis", Journal of the American Society For Information Science, Vol.41, No.6, pp.391-407, 1990.

[18] 海保博之, 原田悦子編:『プロトコル分析入門』, 新曜社, 1993.

[19] Chafe, W.L.:"The Pear Stories", Cognitive, Cultural and Linguistic, 1980.

[20] Ana Constantinescu, Naoko Matsumoto, Daisuke Moriyasu, Hajime Murai, Akifumi Tokosumi, "Transition of aesthetic emotions in interactive environments", Proceedings of the Sixth International Conference on Cognitive Modeling (ICCM 2004), pp.339-340, 2004.

[21] Tokosumi, A. and Matsumoto, N.:"Appraisal components for aesthetic emotions", Abstracts of the Fourth International Conference on Cognitive Science, pp.181, Sydney, 2003.

[22] 富岡美帆, 徃住彰文:「映画レビュー・テキストの分析による感情オントロジーの精緻化」, 日本認知科学会第 23 回大会発表論文集, pp.268-269, 2006.

[23] 斉藤香里, 村井源, 徃住彰文:「心の状態と言語的特徴:ブログにおける商品紹介文の分析」, 情報知識学会第 17 回年次大会, Vol.19, No.2, pp.144-151, 2009.

[24] Ahonen, H., Heinonen, O., Klemettinen, M. Verkamo, I.:"Mining in the Phrasal Frontier", Proc. First European Symp. Principles of Data Mining and Knowledge Discovery (PKDD'97), 1997.

[25] 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕:「blog ページの自動収集と監視に基づくテキストマイニング」, 人工知能学会セマンティックウェブとオントロジー研究会 (SIGSWO-A401-01), 2004.
[26] 乾孝司, 奥村学:「テキストを対象とした評価情報の分析に関する研究動向」, 自然言語処理, Vol.13, No.3, pp.201-241, 2006.
[27] 赤木法生, 大島裕明, 小山聡, 田島敬史, 田中克己:「レビューページ例からの属性抽出に基づくレビューページ検索」, 電子情報通信学会第 17 回データ工学ワークショップ（DEWS2006）論文集, 2C-i10, 2006.
[28] Mendenhall, T.C.:"A mechanical solution of a literary problem", Popular Science Monthly, Vol.60, pp.97-105, 1901.
[29] Wake, W.C.:"Sentence-length distributions of Greek authors", Journal of the Royal Statistical Society, Vol.120, pp.331-346, 1957.
[30] Antosch, F.:"The diagnosis of literary style with the verb-adjective ratio", In Statistics and Style, Doleszel, L. and Bailey R. W. eds., American Elsevier, New York, 1969.
[31] 金明哲:「読点の打ち方と著者の文体特徴, 計量国語学, Vol.19, No.7, pp.317-330, 1994.
[32] 安本美典:「文体統計による筆者推定—源氏物語, 宇治十帖の著者について」, 心理学評論, Vol.2, No.1, pp.147-156, 1958.
[33] 伊藤雅光:『計量言語学入門』, 大修館書店, 2002.
[34] Francis, W. N. and Kucera, H.:"Brown Corpus Manual. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers, Department of Linguistics", Brown University, 1979.
[35] 前川喜久雄:「KOTONOHA 『現代日本語書き言葉均衡コーパス』 の開発 (< 特集> 資料研究の現在)」, 日本語の研究, Vol.4, No.1, pp.82-95, 2008.
[36] Zipf, G. K.:"The psycho-biology of language.", Houghton, Mifflin, 1935.
[37] クラウス・クリッペンドルフ:『メッセージ分析の技法 「内容分析」 への招待』, 三上俊治, 椎野信雄, 橋元良明訳, 勁草書房, 1989.
[38] Speed, Gilmer,J.:"Do newspapers now give the news." Forum. Vol. 15., 1893.
[39] 高橋和子, 高村大也, 奥村学:「機械学習とルールベースの組み合わせによる職業コーディング」, 言語処理学会第 10 回年次大会発表論文集, pp.737-740, 2004.
[40] 秋庭裕, 川端亮:『霊能のリアリティへー社会学, 真如苑に入る』, 新曜社, 2004.
[41] 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』, ナカニシヤ出版, 2014.
[42] Plaisant, C., Rose, J., Yu, B., Auvil, L., Kirschenbaum, M., Smith, M., Clement, T. and Lord, G.:"Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces", Proc. of the 6thth ACM/IEEE-CS Joint Conference on Digital Libraries, pp.141-150, 2006.
[43] Omar, A, A.:"Addressing Subjectivity and Replicability in Thematic Classification of Literary Texts:Using Cluster Analysis to Derive Taxonomies of Thematic Concepts in the Thomas Hardy's Prose Fiction", Journal of the Chicago Colloquium on Digital Humanities and Computer Science 2010, Vol.1, No.2, pp.1-14, 2010.
[44] 赤間啓之:「ベクトル空間モデルに則った, 近代ストア主義とメスメリスムの類似性に関する計量文体論的分析」, 情報処理学会, Vol.50, No.1, pp.1-8, 2001.
[45] 赤間啓之, 鄭在玲, 三宅真紀:「グラフクラスタリングを用いたソシュールの概念ネットワーク解析.」, 情報処理学会研究報告. 人文科学とコンピュータ研究会報告(2007.9), pp.33-40, 2007.
[46] 村井源, 松本斉子, 山本竜大, 徃住彰文:「Web の計量言語学的分析からみた政治的感性の特徴」, 感性工学会研究論文集, Vol.7, No.3, pp.561-569, 2008.
[47] Aoshima, Y. and Tokosumi, A., "Extracting musical concepts from written texts:A case study of Toru Takemitsu", Proceedings of KEER 2007, CD-ROM, 2007.
[48] 工藤彰, 村井源, 徃住彰文:「村上春樹の初期三部作における構造解析」, 情報知識学会第 17 回年次大会, Vol.19, No.2, pp.126-131, 2009.
[49] 工藤彰, 村井源, 徃住彰文:「共通語の布置と変化に基づく並行形式小説の物語構造」, 情報知識学会誌, Vol. 22, No. 3, pp. 187-202, 2012.

[50] 滕文娜, 川島隆徳, 村井源, 徃住彰文:「エンターテイメントコンテンツ作品の相互関係性―大量ブログデータのネットワーク分析―」, 第6回日本感性工学会春期大会予稿集, CD-ROM 16B-04, 2009.
[51] 村井源, 徃住彰文:「テキスト批評の計量化に向けてー書評の計量分析ー」, 情報知識学会第17 回年次大会, Vol.19, No.2, pp.120-125, 2009.
[52] 村井源, 徃住彰文:「文芸批評の計量解析による批評行為の背景的特徴の抽出」, 情報知識学会誌, Vol. 20, No. 2, pp.117-122, 2010.
[53] 村井源, 川島隆徳:「総合芸術の批評における批評対象の特徴―映画と演劇の批評の計量的比較―」, 情報知識学会誌, Vol.22, No.3, pp.203-222, 2012.
[54] 河瀬彰宏, 村井源, 徃住彰文:「音楽評論論文にみる概念構造の変遷―ネットワーク中心性を用いた音楽概念の抽出―」, 情報知識学会第 17 回年次大会, Vol.19, No.2, pp.138-143, 2009.
[55] 末吉亙:「情報解析と著作権」, 情報管理, 55.6, 434-437, 2012.
[56] Text Encoding Initiative:TEI P5 Guidelines, http://www.tei-c.org/ (2014.10.27 閲覧)
[57] Dublin Core, http://dublincore.org/ (2014.10.27 閲覧)
[58] 国立国会図書館:DC-NDL, http://www.ndl.go.jp/jp/aboutus/standards/meta.html (2014.10.27 閲覧)
[59] FOAF, http://www.foaf-project.org/ (2014.10.27 閲覧)
[60] 笹田鉄郎, 森信介, 河原達也:「テキストと音声を用いた単語と読みの自動獲得」, 情報処理学会研究報告, SLP-72, 2008.
[61] 小木曽智信, 小椋秀樹, 田中牧郎, 近藤明日子, 伝康晴:「中古和文を対象とした形態素解析辞書の開発」, 情報処理学会研究報告, 2010(4), pp1-8, 2010.
[62] Schmid, H.:"TreeTagger - a language independent part-of-speech tagger", http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (2014.10.27 閲覧)
[63] Manning, D., et al.:"Stanford Parser", http://nlp.stanford.edu/software/lex-parser.shtml (2014.10.27 閲覧)
[64] 溝口理一郎:『知の科学 オントロジー工学』, オーム社, 2005.
[65] 国立国語研究所:『分類語彙表』, 大日本図書, 2004.
[66] 渡邉陽太郎, 浅原正幸, 松本裕治:「述語語義と意味役割の結合学習のための構造予測モデル.」, 人工知能学会論文誌 25.2, pp252-261, 2010.
[67] Yule, G. U.:"The Statistical Study of Literaty Vocabulary", Cambridge University Press, 1944.
[68] 中川裕志, 森辰則, 湯本紘彰:「出現頻度と連接頻度に基づく専門用語抽出」, 自然言語処理, Vol.10, No.1, 2003.
[69] 金明哲 : 『テキストデータの統計科学入門』,岩波書店,2009
[70] Salton, G.:"Automatic Text Processing", Addison-Wesley, 1989.
[71] 藤村滋, 豊田正史, 喜連川優:「文の構造を考慮した評判抽出手法」, 電子情報通信学会第 16 回データ工学ワークショップ(DEWS2005), pp.60-78, 2005.
[72] Nilsson, J., Nivre, J.:"MaltParser", http://maltparser.org/index.html (2014.10.27 閲覧)
[73] 梶川裕矢, 森純一郎:「ネットワーク指標を用いた学際的な論文の抽出」, 情報知識学会誌, Vol.19, No.2, pp.170-173, 2009.
[74] 金光淳:「社会ネットワーク分析の基礎」, 勁草書房, 2003.
[75] Ellson, John, et al.:"Graphviz-open source graph drawing tools." Graph Drawing. Springer Berlin Heidelberg, 2002.
[76] Newman, M.E.J.:"Networks: An Introduction", OUP Oxford, 2010
[77] Girvan, M and Newman, M.E.J.:"Community structure in social and biological networks", Proc. Natl. Acad. Sci. USA 99, pp.7821-7826, 2002.
[78] Newman, M.E.J. and Girvan, M:"Finding and evaluating community structure". Physical review E, Vol.69, no.2, pp.026113, 2003
[79] 工藤彰, 村井源, 徃住彰文:「計量分析による村上春樹長篇の関係性と歴史的変遷」, 情報知識学会誌, Vol.21, No.1, pp.18-36, 2011.
[80] 石川慎一郎, 前田忠彦, 山崎誠:『言語研究のための統計入門』, くろしお出版, 2010
[81] 徃住彰文:『心の計算理論』, 東京大学出版会, 2007.

[82] 原田実:「意味解析により進化するテキストマイニング (解析手法, 第 1 回テキストマイニング・シンポジウム).」 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol.111, No.119, pp25-30., 2011.
[83] 田中敏 :「js-STAR」, http://www.kisnet.or.jp/nappa/software/star/ (2015.1.24 閲覧)
[84] The R Project for Statistical Computing:http://www.r-project.org/ (2015.1.24 閲覧)
[85] 川島隆徳, 村井源, 徃住彰文:「ゲーム批評から見たゲームの「面白さ」―レビューテキストの計量解析による叙述対象の自動抽出―」, デジタルゲーム学研究, Vol.4, No.1, pp.69-80, 2010.
[86] Malone, T. W.:"Toward a theory of intrinsically motivating instruction", Cognitive Science, Vol.4, pp333-369, 1981.
[87] Vorderer, P.:"Interactive entertainment and beyond", In D. Zillmann & P. Vorderer (Eds.), Media entertainment:The psychology of its appeal (pp. 21-36). Mahwah, NJ:Lawrence Erlbaum Associates, 2000.
[88] Vorderer, Peter, Tilo Hartmann, and Christoph Klimmt:"Explaining the enjoyment of playing video games:the role of competition." Proceedings of the second international conference on Entertainment computing. Carnegie Mellon University, 2003.
[89] Csikszentmihalyi, Mihaly, Mihaly Csikzentmihaly:"Flow:The psychology of optimal experience", Vol.41, New York:HarperPerennial, 1991.
[90] Sweetser, P. and Wyeth, P. GameFlow:"A Model for Evaluating Player Enjoyment in Games", Computers in Entertainment, Vol.3, No.3, Article 3A, 2005.
[91] Song, Seungkeun, Joohyeon Lee, Jun Jo:"The analysis of game playing experiences:focusing on massively multiplayer online role-playing game." Advances in Hybrid Information Technology, Springer Berlin Heidelberg, pp.333-343, 2007.
[92] Newell, A., Simon, H.:"Human problem solving", Prentice Hall, Englewood Cliffs, 1972.
[93] Frasca, G.:"Ludologists love stories, too:notes from a debate that never took place." in Copier & Raessens (Eds.) Level up:Digital Games Research Conference, Utrecht University, 2003.
[94] 馬場章:「ゲーム学の国際的動向～ゲームの面白さを求めて～」, 映像情報メディア学会誌, Vol.69, No.4, pp.491-494, 2006.
[95] 山下利之, 清水孝昭, 栗山裕, 橋下友茂「コンピューターゲームの特性と楽しさの分析」, 日本教育工学会論文誌, 第 28 巻第 4 号, pp.349-355, 2004.
[96] 蔵琢也:「各種計量指標から見るゲーム機ハードの歴史」, ITEC Working Paper Series, Vol.6, No.05-06, 2005.
[97] 小泉俊昭（編）:『ゲーム批評』, マイクロデザイン出版局, 1994～2006.
[98] 川島隆徳, 高田知紀, 桑子敏雄, 村井源, 徃住彰文:「テキスト解析手法を用いた河川文化概念の構造化」, 情報知識学会誌, Vol.24, No.1, pp.3-18, 2014.
[99] 高田知紀:「多自然川づくり事業における合意形成プロセスの評価枠組みに関する研究, 平成 21 年度東京工業大学大学院修士論文」, 2010.
[100] 林倫子, 神邊和貴子, 出村嘉史, 川崎雅史:「明治・大正期の納涼床営業者の鴨川官有地利用に関する研究－先斗町三条・四条間を対象として－」, 土木学会論文集 D, Vol.66, No.2, pp.246-254, 2010.5.
[101] 中嶋伸恵, 田中尚人, 秋山孝正:「水辺空間を基盤とした地域コミュニティの形成に関する研究」, 土木学会論文集 D, Vol.64, No.2, pp.168-178, 2008.
[102] 竹林征三:『甲斐路と富士川－川を守り・道を拓く－』, 土木学会山梨会, 1996.
[103] 富山和子:『水の文化史』, 文藝春秋, 1980.
[104] 中村良夫:「湿性文化の行くへ－生態・文化複合系を再構築しよう－」, 河川, pp.3-6, 2009.
[105] 高橋裕:『現代日本土木史』, 彰国社, 1990.
[106] 富野章:『日本の伝統的河川工法』, 信山社サイテック, 2002.
[107] 大熊孝:『増補・洪水と治水の河川史』, 平凡社ライブラリー2007.
[108] 日本河川協会編:『河川文化』, 日本河川協会, 1995-.
[109] 山田聡宣, 島谷幸宏, 末松吉生:「中小河川の改修手法の工夫による CO2 排出量の削減」, 河川技術論文集, Vol.16, pp.455-458, 2010.

[110] 独立行政法人科学技術振興機構・社会技術研究開発センター「地域に根ざした脱温暖化・環境共生社会」研究開発領域・地域分散電源等導入タスクフォース（編著）：『小水力発電を地域の力で』, 公人の友社, 2010.
[111] 椋木雅之, 田中大典, 池田克夫:「対義語対からなる特徴空間を用いた感性語による画像検索システム」, 情報処理学会論文誌, Vol.42, No.7, pp.1914-1921, 2001.
[112] 池添剛, 梶川嘉延, 野村康雄:「音楽感性空間を用いた感性語による音楽データベース検索システム」, 情報処理学会論文誌, Vol.42, No.12, pp.3201-3212, 2001.
[113] 原田隆史:「書評中の感性キーワードを用いた小説の分類」, 情報知識学会誌, Vol.15, No.2, pp.57-62, 2005.
[114] 菅谷智浩, 杉原敏昭, 佐藤美恵, 春日正男:「映画を対象とした感性的評価手法に関する一検討」, ヒューマンインタフェース学会研究報告集:human interface, Vol.11, No.4, pp.29-34, 2009.
[115] 石崎博之, 杉原敏昭, 佐藤美恵:「映画・TV ゲームにおける感性的評価手法に関する検討（メディア工学)」, 映像情報メディア学会技術報告, Vol.31, No.39, pp.37-40, 2007.
[116] 河出書房新社 [編]:「文藝」, 河出書房新社, 1962-, http://www.kawade.co.jp/np/bungei.html (2014.10.27 閲覧)
[117] 国際演劇批評家協会日本センター:「Theatre Arts」, 晩成書房, 1994-, http://theatreart.exblog.jp/ (2014.10.27 閲覧)
[118] 荒井晴彦:「映画芸術」, 編集プロダクション映芸, 1946-, http://eigageijutsu.com/ (2014.10.27 閲覧)
[119] Weblio 類語辞典:http://thesaurus.weblio.jp/ (2014.10.27 閲覧)
[120] Cochran, William G.:"Some methods for strengthening the common χ 2 tests.", Biometrics Vol.10, No.4, pp.417-451, 1954.
[121] 松尾豊, 友部博教, 橋田浩一, 中島秀之, 石塚満:「Web 上の情報からの人間関係ネットワークの抽出」, 人工知能学会論文誌, Vol.20, No.1, pp.46-56, 2005.
[122] 須永哲矢:「コロケーション強度を用いた中古語の語認定」, 国立国語研究所論集, Vol.2, pp.91-106, 2011.
[123] 金明哲, 張信鵬:「SA6-5 テキスマイニングツール MTMineR のコンセプトと機能（特別セッション コーパスとテキストマイニング II).」, 日本行動計量学会大会発表論文抄録集, Vol.41, pp360-363, 2013.
[124] 松村真宏, 三浦麻子:TinyTextMiner, http://mtmr.jp/ttm/ (2014.10.27 閲覧)
[125] 石田基広:『R によるテキストマイニング入門』, 森北出版, 2008.
[126] Jung the Java Universal Network/Graph Framework:http://jung.sourceforge.net/ (2014.10.27 閲覧)
[127] 川島隆徳, 佐住彰文:「フォークオントロジーの提案と構築環境の開発」, 人工知能学会 知識ベースシステム研究会(2008/6/25), pp17-24, 2008.
[128] 川島隆徳, 研谷紀夫:「著者名典拠情報を拡充するための共同編集プラットフォーム」, 情報知識学会誌, Vol.20, No.2, pp.183-188, 2010.
[129] 研谷紀夫, 川島隆徳:「Digital Cultural Heritage を用いて家族写真の特性を明らかにする方法の提示とその検証に関する研究」, アート・ドキュメンテーション研究, Vol.21, pp.3-21, 2014.