質的研究法

テキストマイニング

いとうたけひこ(伊藤武彦）
2015年

テキストマイニングとは

- テキストマイニングはデータマイニングの一種、
- 蓄積された膨大なテキストデータを何らかの単位(文字，単語，フレーズ)に分解し，これらの関係を定量的に分析すること（金　明哲：テキストデータの統計科学入門(第1版)，第1版，pp. 2-4，pp. 64-72，岩波書店，東京都，2009．金　明哲：テキストデータの統計科学入門(第1版)，第1版，pp. 2-4，pp. 64-72，岩波書店，東京都，2009．）
- つまり、テキストマイニングは質的なテキストデータを対象にして、統計的手法を用いる量的な分析であるといえる。

★表１　典型的な質的研究と量的研究とテキストマイニングとの比較

方法	データ	分析方法
量的研究	数値（量的データ）	量的分析（統計）
質的研究	文字（質的データ）	質的分析
データマイニング	数値（量的データ）	量的分析（統計）
テキストマイニング	**文字（質的データ）**	**量的分析（統計）**

テキストマイニングは質的方法と量的方法との両方の特徴をもつ。
この特性を生かした研究への活用法が期待される。

※小平朋江・伊藤武彦・松上伸丈・佐々木　彩（2007）　テキストマイニングによるビデオ教材の分析：精神障害者への偏見低減教育のアカウンタビリティ向上をめざして　マクロ・カウンセリング研究, 6, 16-31.

服部(2010)による説明:資料集

服部兼敏 2010 テキストマイニングで広がる看護の世界 ナカニシヤ出版

- 自然言語処理
- 日本語の特徴:形態素と分かち書き

孤立語:単語が独立　中国語

屈折語:語尾変化　印欧語

膠着語:後置詞、助詞　日本語

膠着語は分かち書きに工夫が必要

	日本	美味しい	優秀	看護師	神戸	米	パン	車	の	さん	は	だ
日本の看護師さんは優秀だ	1	0	1	1	0	0	0	0	1	1	1	1
日本の車は優秀だ	1	0	1	0	0	0	0	1	1	0	1	1
日本の米は美味しい	1	1	0	0	0	1	0	0	1	0	1	0
神戸のパンは美味しい	0	1	0	0	1	0	1	0	1	0	1	0

対象となる文字データ

- （1）公開、公共化されているもの：著作権に配慮すれば、倫理委員会を通さなくて良い。

例：Cinii、医中誌などのデータベース

闘病記、小説などの著作物

website, twitterなど

公開されているデータの2次利用

- （2）（自分で他者から）収集したもの：ヒトを扱う研究であるので倫理委員会・倫理的配慮が必要

例：インタビュー記録、質問紙調査（自由記述など）、等

研究対象となるテキストの例

- 論文データベース（Cinii,　医中誌など）
  （孫,2008；久木田,2008；井上・伊藤,2009；小沢2011）
- 質問紙の自由記述回答 （佐々木,2007）（八城,2007）
- 調査報告書の自由記述再分析　（井川, 2009）
- 試験の答案 （松上,2006）
- 記述式の質問紙 （伊藤・川島,2009; 野中, 2009）
- Webサイト、ブログ （大高他,2008；　松上・伊藤,2008）
- 新聞記事データベース （目黒、2009）
- 単行本 （闘病記；コミックエッセイ（小平・伊藤、2012））

テキストマイニングでできること
服部本の目次

服部兼敏 2010 テキストマイニングで広がる看護の世界 ナカニシヤ出版

- 原文参照機能
- 単語頻度分析
- 単語検索
- 係り受け（共起）分析
- 注目語分析
- 特徴語分析
- ネットワーク分析
- 評判分析
- 対応（バブル）分析
- クラスター分析（文章分類）
- 時系列分析
- ★TMSのメニュー

1．単語検索（Word, Adobe Reader）

3.1.3　テキストマイニングソフトの検索（原文参照）機能の活用

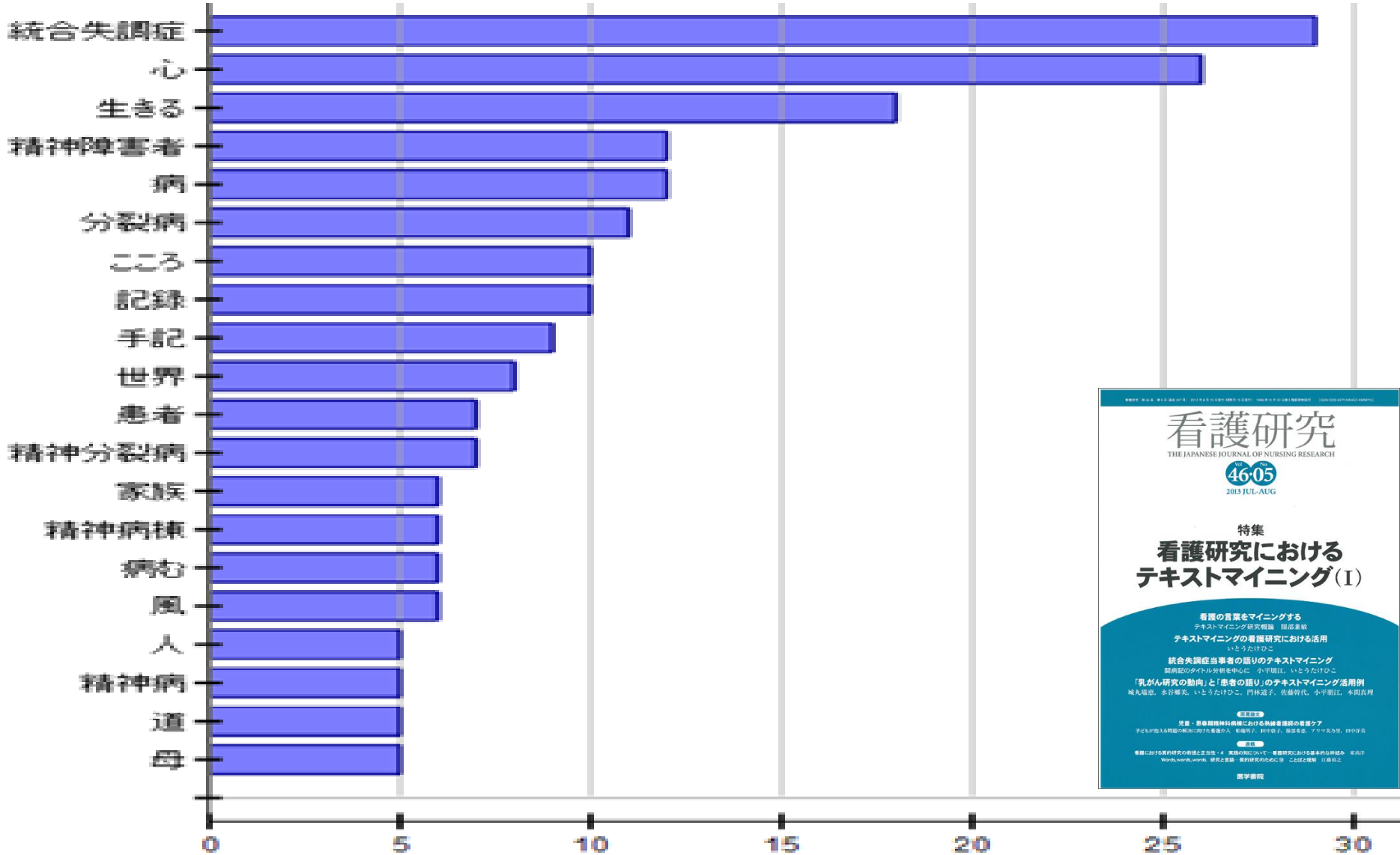
テキストマイニングソフトの中には原文の文字列や単語や類義語群，さらにはグループ化された一連の単語群を検索し，リストを作成し，そのリストを一覧表として出力する機能があり，質的研究においてきわめて有用である。この機能は，検索機能あるいは原文参照機能などと呼ばれている。いまや検索・原文参照機能はワープロソフトやPDF読み取りソフトにも標準装備されており，例えば本稿執筆のために使用しているソフトでは，文書ファイルの中に「いとう」「伊藤」「Ito」という文字列がそれぞれ，35，8，12件あったということが簡単にカウントされ，本文内のその文字列にジャンプすることができる（例えば，Word 2010以降のバージョンの「ナビゲーション」による「文書の検索」機能）。また，テキスト認識済みのPDF形式の文書ファイルでも同様に，文字列のカウントと検索による原文参照ができる（例えば，Adobe Reader XIにおける「高度な検索」機能）。その原理は自然言語処理における*n*-gram（エヌグラム）の集計機能である（金，2009，p.38）。しかし，これらの非テキストマイニングのソフトの機能は簡素で，文書内の文字列の総数のみをカウントできるだけである。

2. 原文参照機能（Text Mining Studio）

ある種のテキストマイニングソフト〔例えば，「Text Mining Studio」(数理システム)〕では，総数，出現箇所だけでなく，テキスト以外の属性(例えば，性別やがんの種類ごと)も一覧表に表示され，また属性に限定した原文参照も可能である。加えて，単語，類義語，グループ化された一連の単語群などの意味的なレベルでの検索と原文参照が可能であるので，質的データ分析におけるツールとして有益性が高い。検索と原文参照の機能の支援を受けて，能率的な質的分析をテキストマイニングソフト単体で行なうことができるのである。

３　単語頻度分析

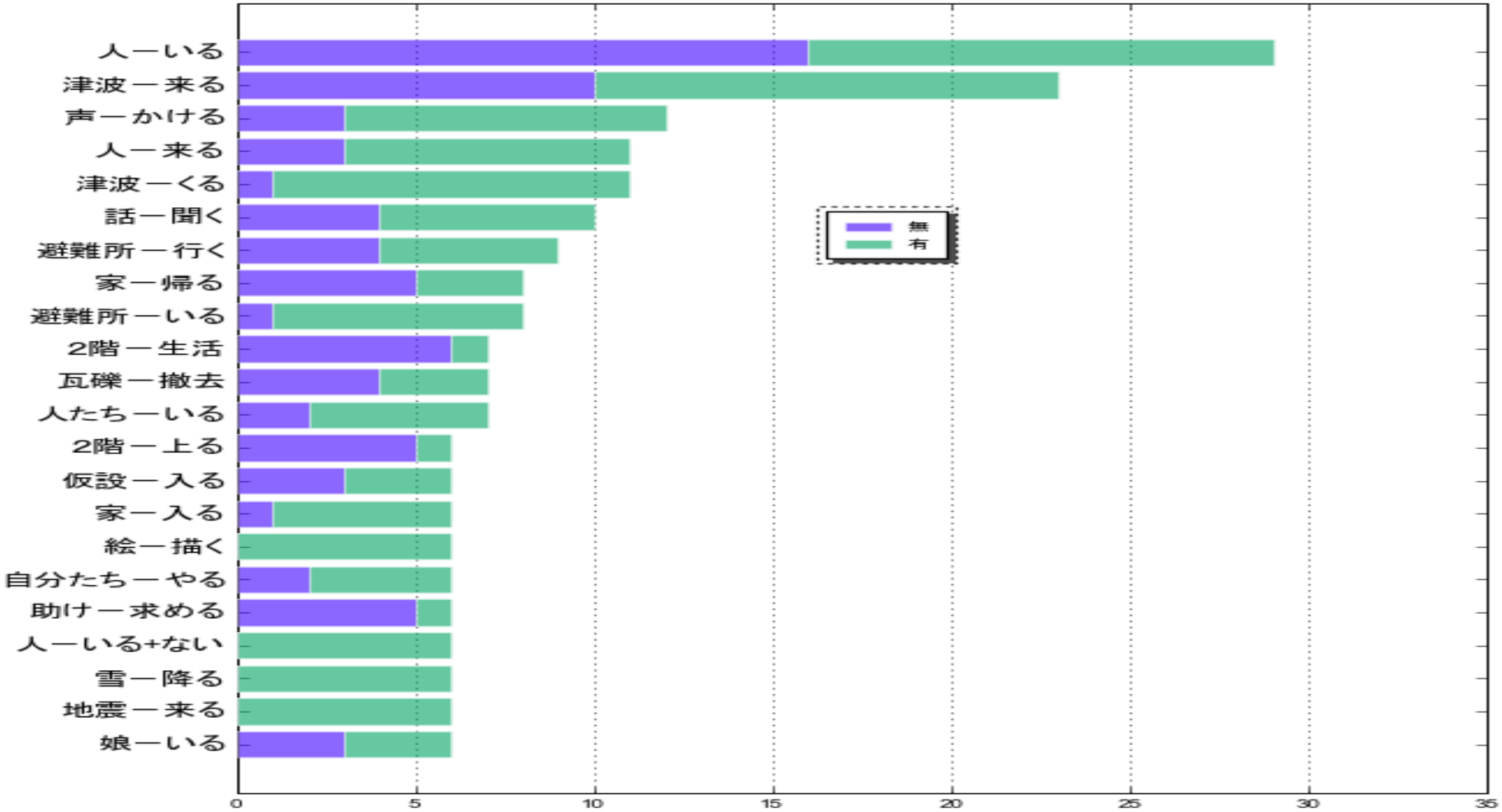
統合失調症闘病記217冊の表題における単語頻度分析（上位20件）



４　係り受け頻度解析(死について言及しているかの有無)

「『東北の声』のテキストマイニング分析ー死に関する表現の有無を手がかりにしてー」 下条 照世(2014)

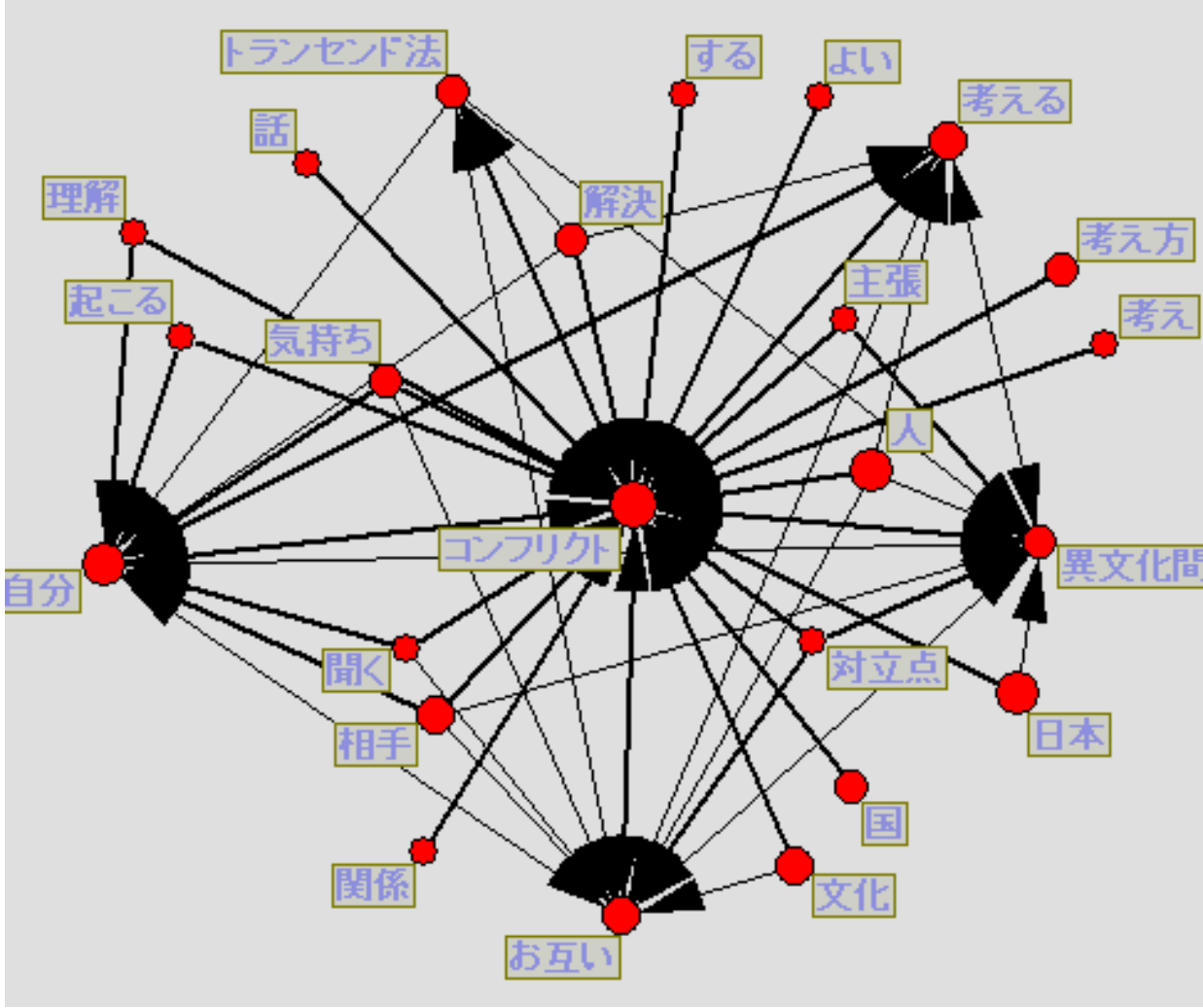
http://www.msi.co.jp/tmstudio/stu14result.html



5. 注目語分析

松上伸丈2006　異文化間コンフリクトの解決方法について

分析データとして、東京都内A 大学の2006 年度前期期末試験課題の回答者192 人のうち33 人分（17%）をサンプリングにて使用した。



6. 特徴語分析

児童養護施設職員の児童養護施設の改善に関するニーズと職務継続意思に関する一研究

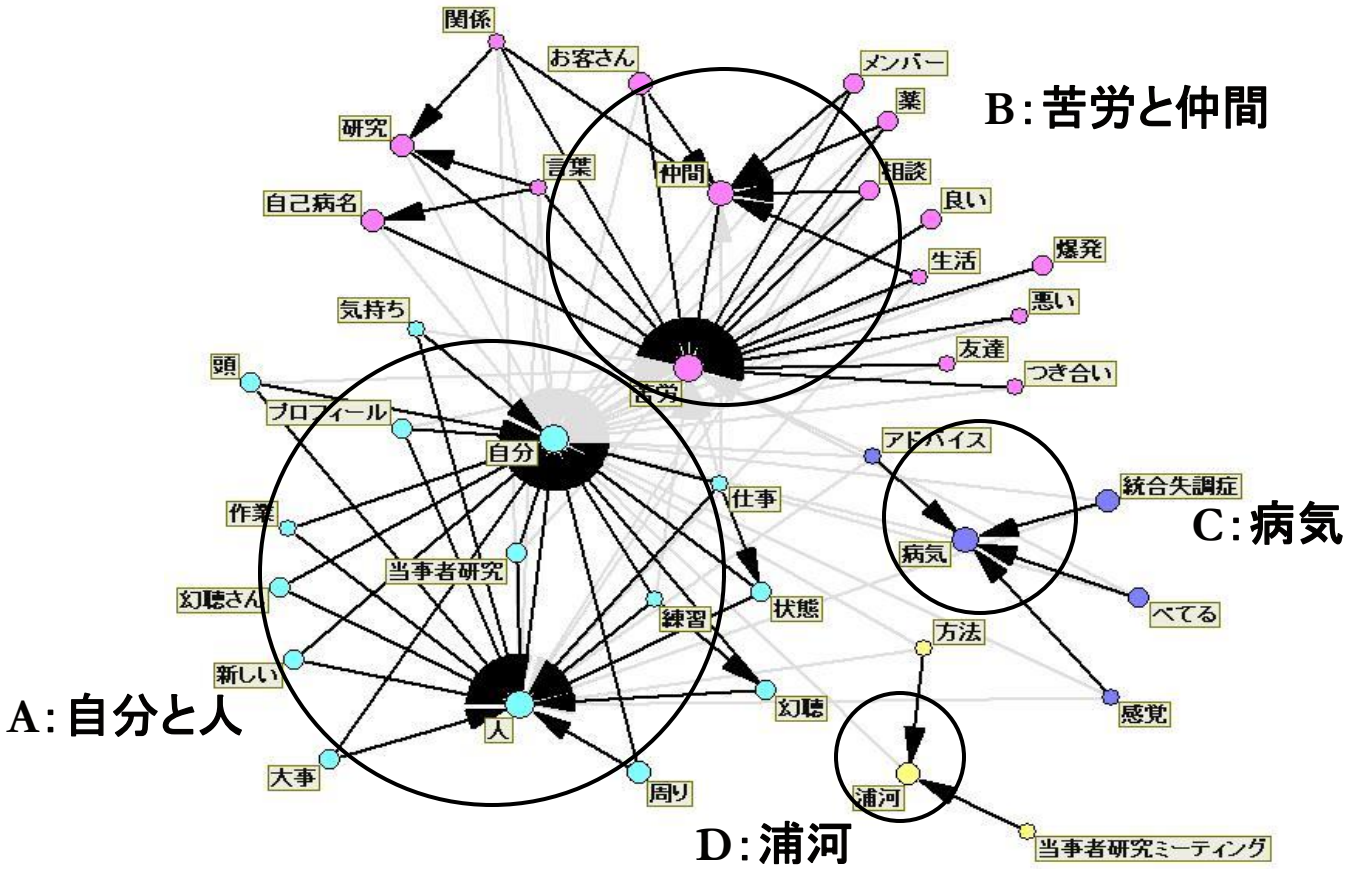
八城 真里 2007

児童養護施設で直接処遇に従事する職員を対象に質問紙調査をおこない、826人の「現状をさらによくしていくために何が必要ですか」の質問の自由記述回答を分析した。

表5-1. 男女別特徴語抽出［上位20］

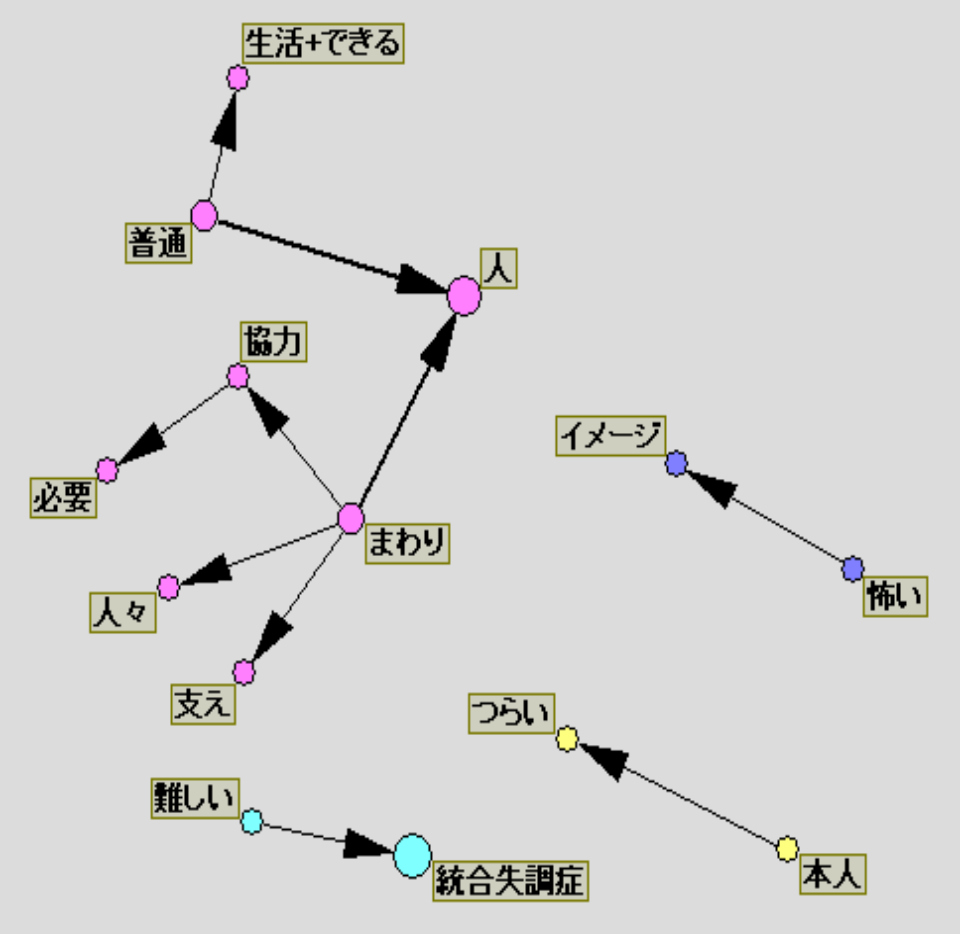
男性		女性	
単語	指標値	単語	指標値
職員配置基準	44.976	子ども	25.111
見直し	27.388	大切	16.049
向上	18.58	支援	13.5
職員	17.194	理解	12.112
給料	16.672	知識	10.734
改善	15.842	現状	9.913
予算	13.557	良い	9.827
措置費	12.736	数	9.639
小規模化	9.8	職員同士	8.638
増やす	9.564	見る	8.346
質	7.779	心	7.892
確保	7.232	親	7.156
少ない	6.316	連携	6.523
改正	6.137	十分	6.42
人材	5.768	意見	6.146
人件費	5.589	大人	5.872
処遇	5.459	女性	5.778
解消	4.947	定期的	5.778
統一	4.947	入所	5.778
増える	4.853	上司	5.598

7. ネットワーク分析

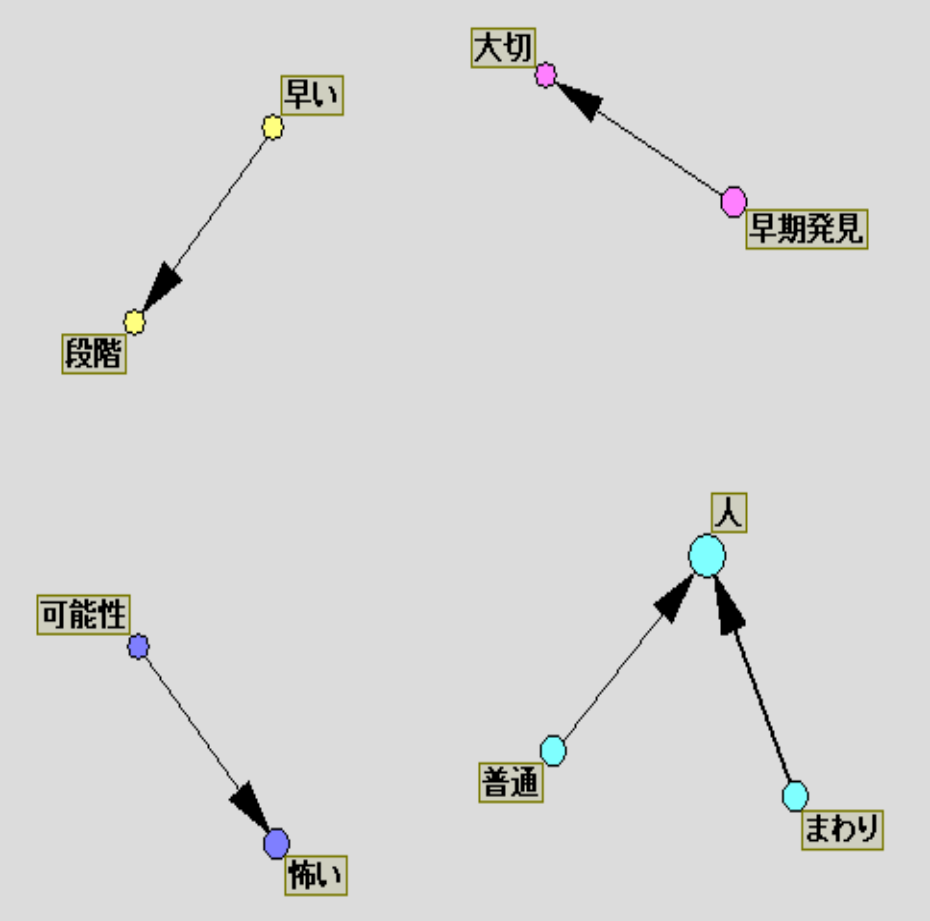


精神障害者のためのWebsite『当事者研究の部屋』の記述の分析　大高 庸平2008

大学生の統合失調症に対するビデオ視聴後のテキスト分析
(小平・伊藤・松上・佐々木, 2007)



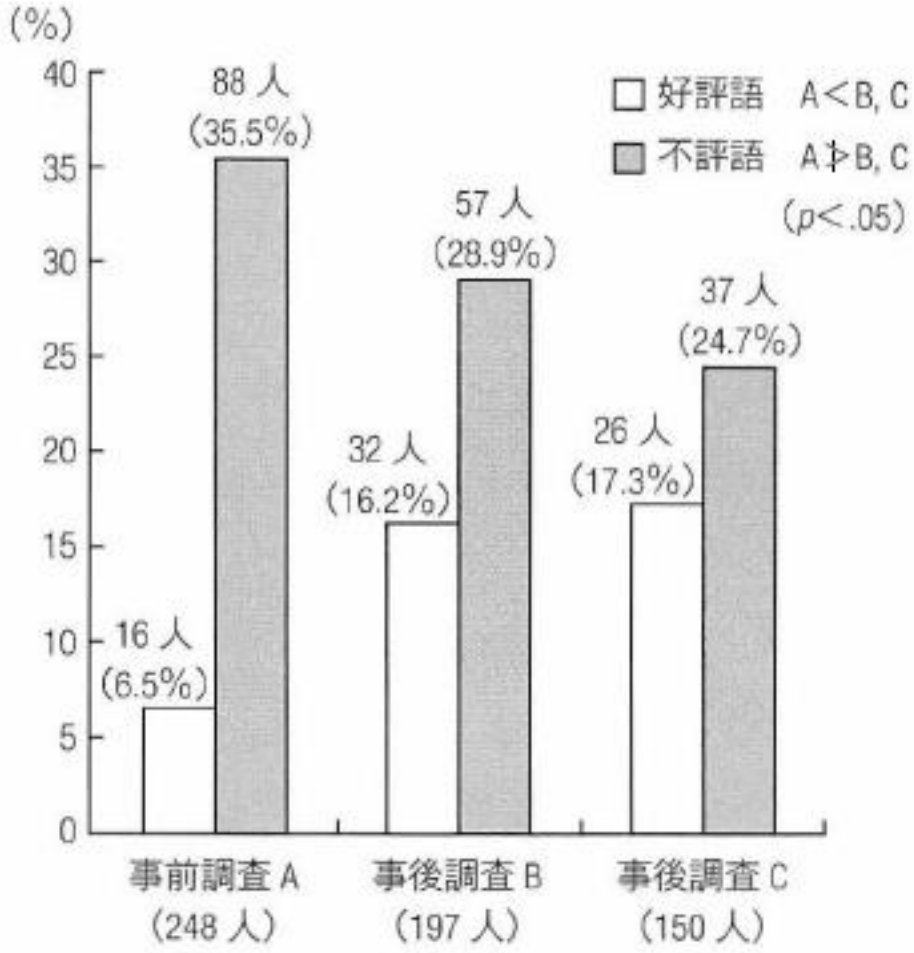
ことばネットワーク　患者談話ビデオ



医者説明ビデオ

８　評判分析

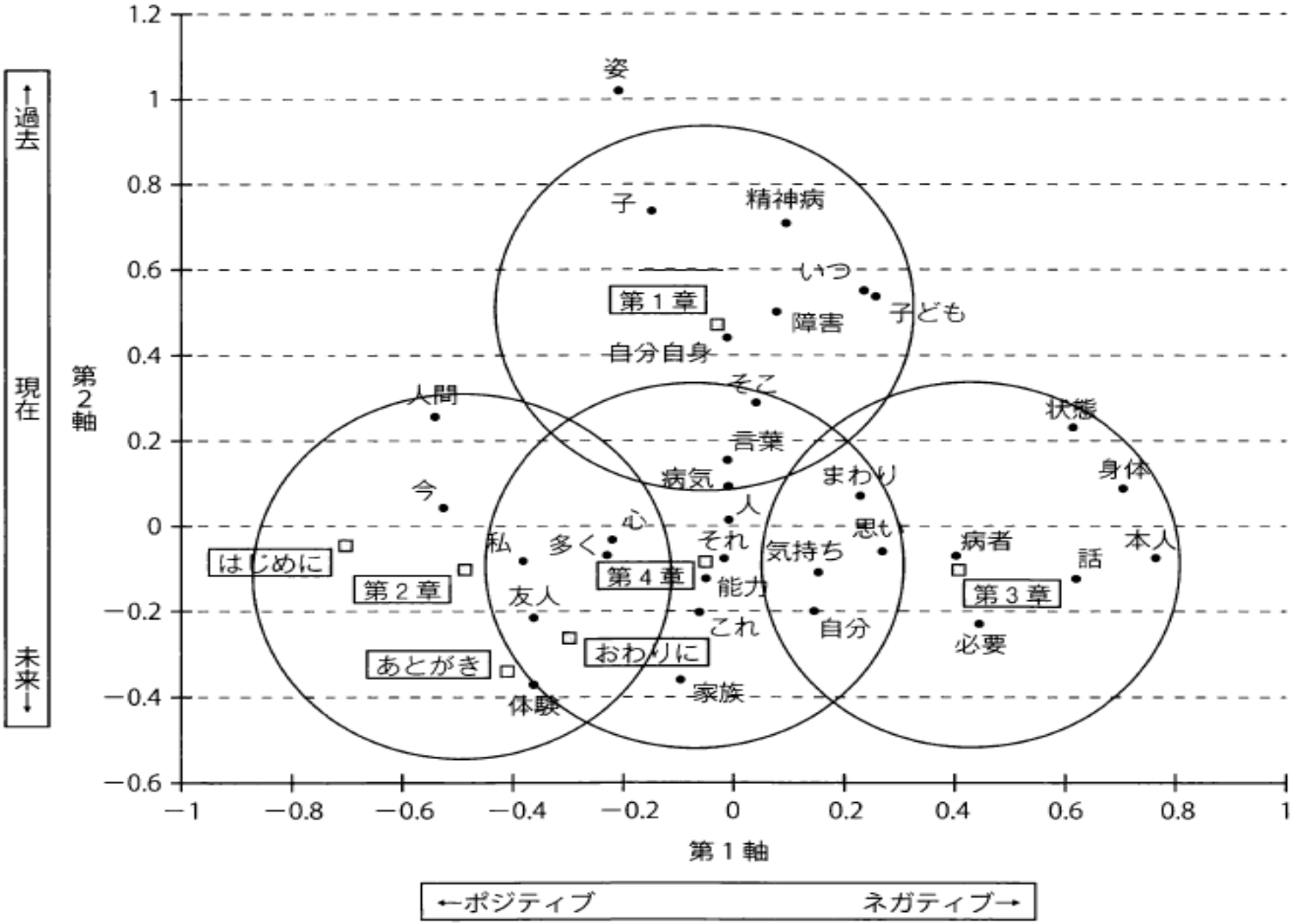
図２　ナラティブ教材を活用した授業での
統合失調症に対する好評語と不評語の使用の
人数割合



(小平，いとう，in press)

９　対応（バブル）分析

図３　ある統合失調症の闘病記の章立ての構造



小平，いとう，大高（2010）

大高庸平・城丸瑞恵・いとうたけひこ　2010　手術とホルモン療法を受けた乳癌患者の心理：テキストマイニングよる語りの分析から　昭和医学会雑誌，　70（4），302-314.

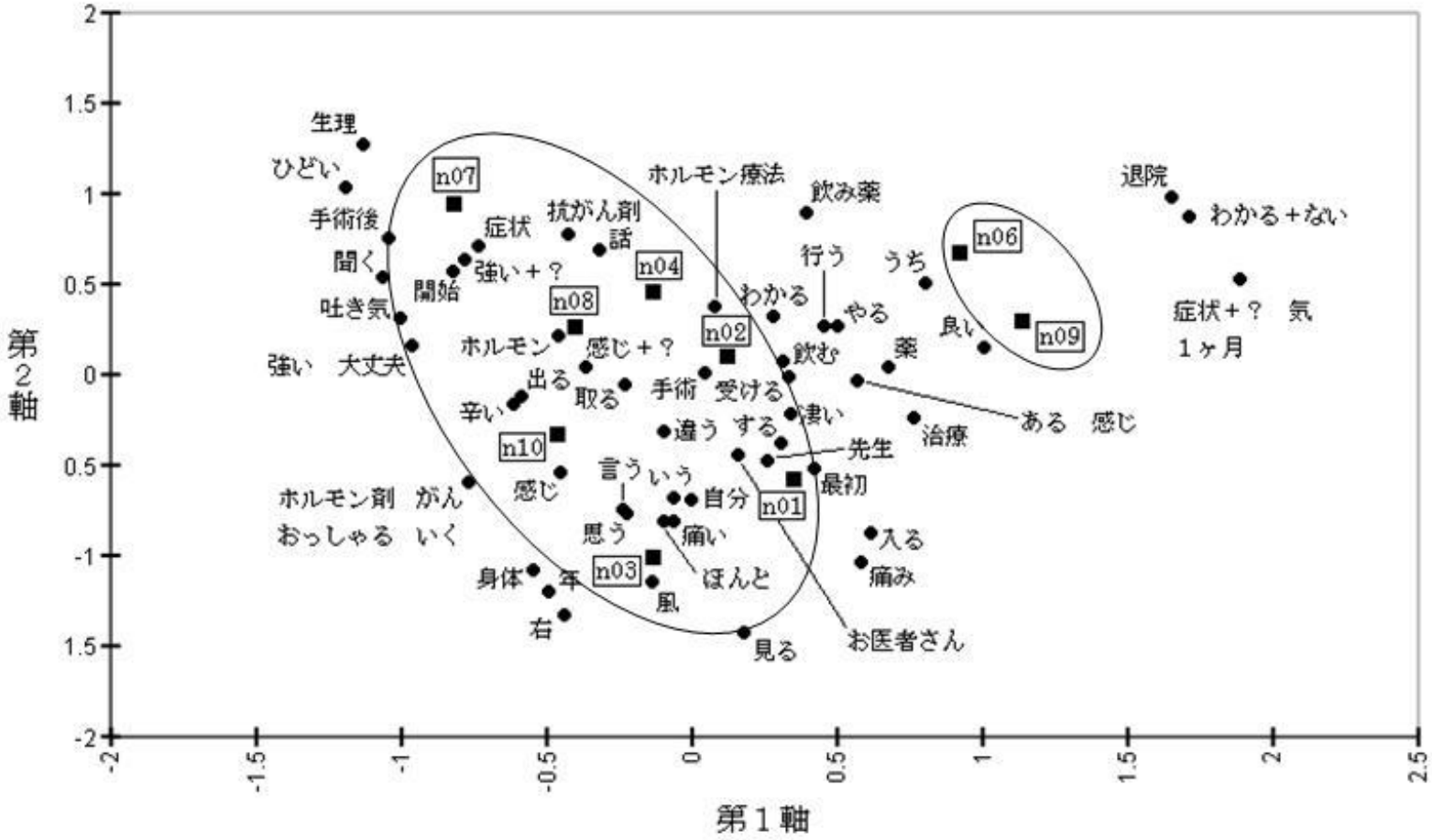
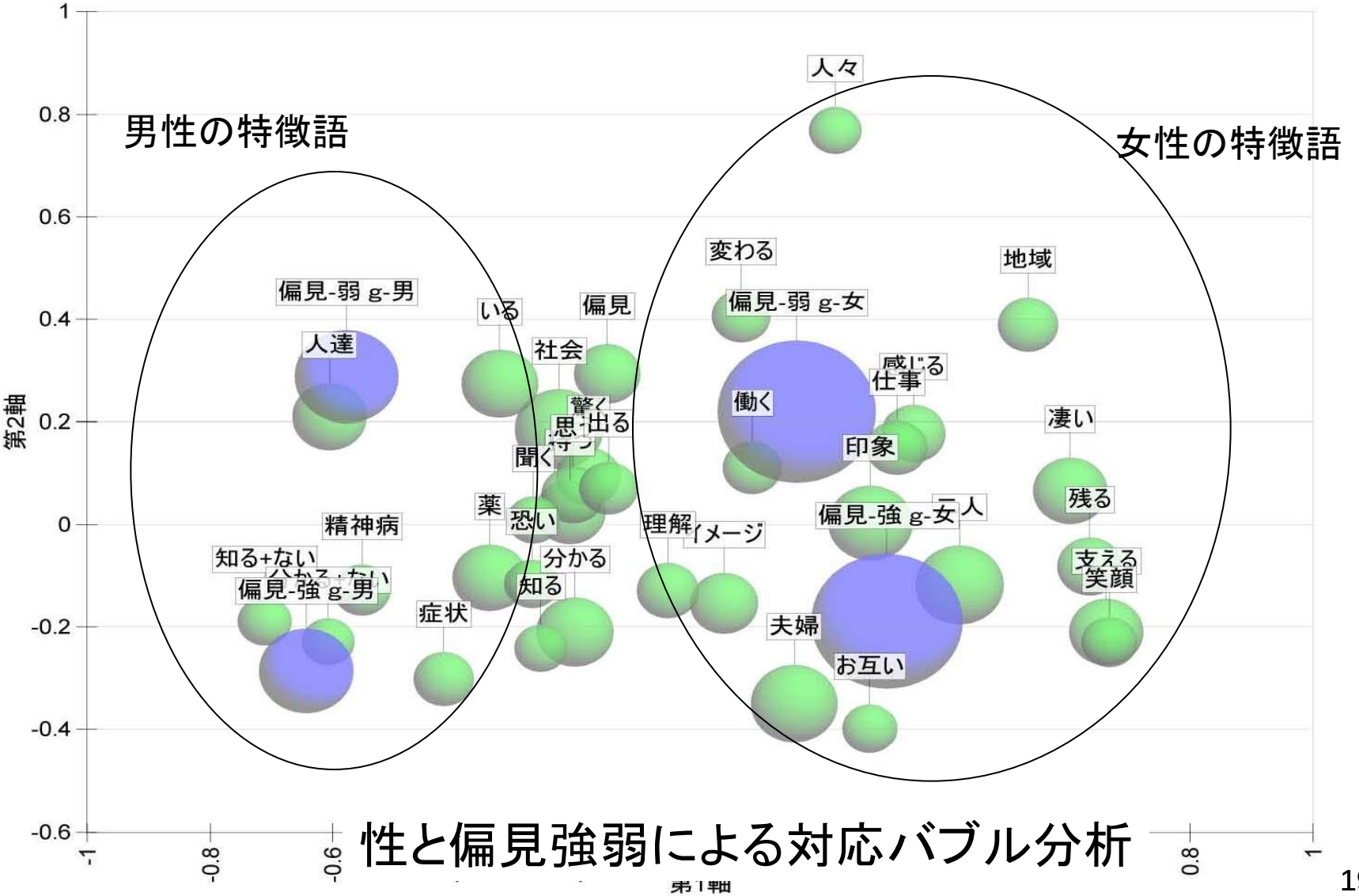


Fig. 1　カテゴリ“ホルモン療法”と“手術”内における個人(n)と単語による対応分析

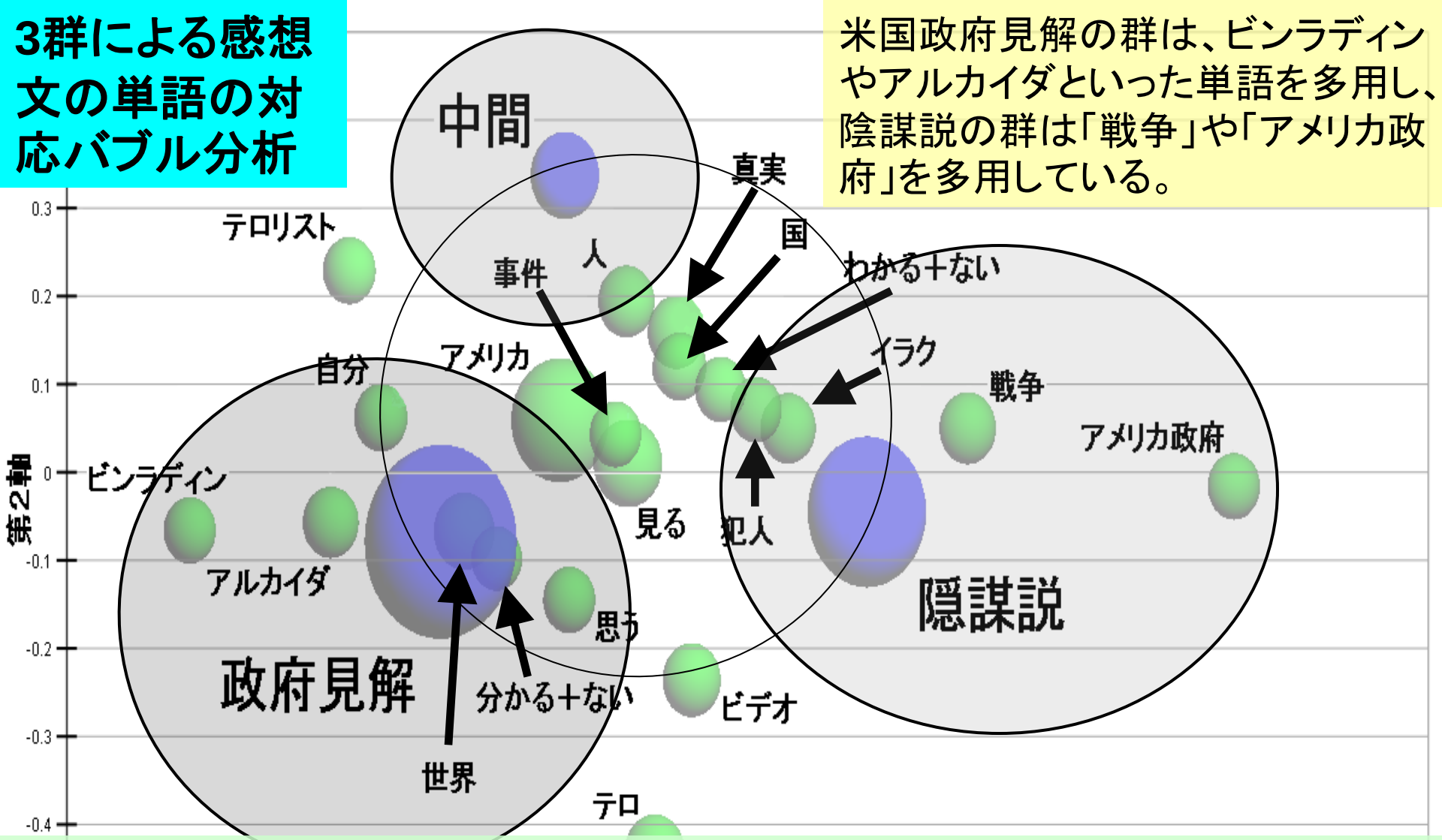
テレビ番組視聴による統合失調症を持つ人に対する偏見低減の効果　小平・伊藤(2009)



性と偏見強弱による対応バブル分析

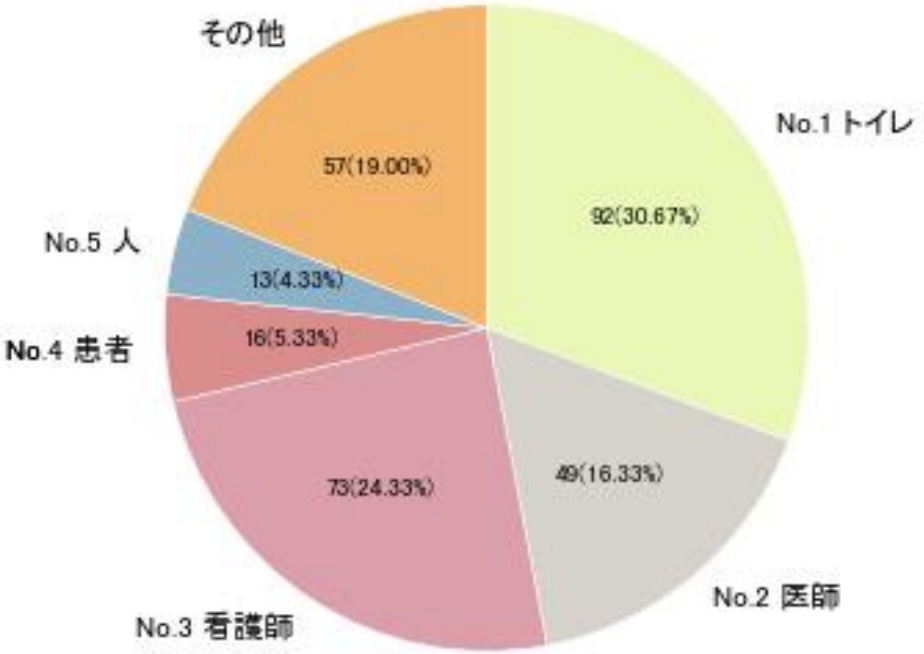
3群による感想文の単語の対応バブル分析

米国政府見解の群は、ビンラディンやアルカイダといった単語を多用し、陰謀説の群は「戦争」や「アメリカ政府」を多用している。



ビデオ視聴による政治的態度の変容:『911ボーイングを捜せ』視聴前から視聴後への「陰謀説」支持の増加はなぜおこったか？　伊藤・川島(2009)

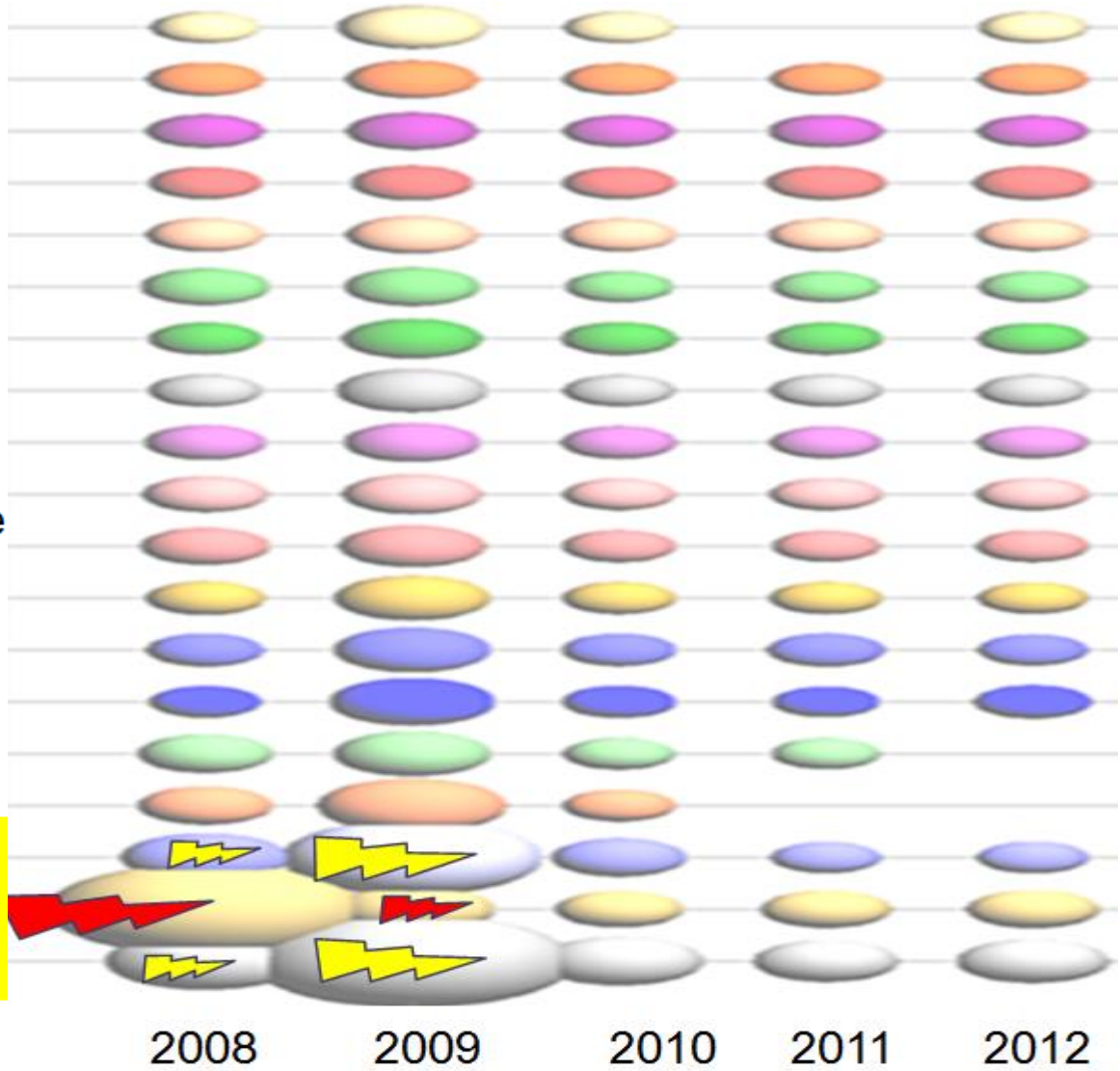
10　クラスター分析（文章分類）



11　時系列分析

Ito, T., Sueyoshi, E. & Inoue, T. (2013). Text mining analysis of the narratives of a patient with fibromyalgia: Focusing on expressions of pain.
16th East Asian Forum of Nursing Scholars (EAFONS 2013) Forum Proceedings, 348.

疼痛	throbbing
感じ	feeling
ブログ	blog
ベッド	bed
良い	good
ドクター	doctor
言葉	words
身体	body
薬	medication
人	person / people
激痛	severe pain
自分	myself
オピオイド	opioid
命	life
線維筋痛症	fibromyalgia
副作用	side effects
痛い	painful
痛み	pain



テキストマイニングの役割

- テキストマイニング中心の研究
- 混合研究法による補完的使用
- 質的研究との連携
- 探索的研究vs.仮説検証的研究（仮説生成的研究）

看護学においては量的研究として
心理学においては質的研究として

★（1）テキストマイニング単独の（中心の）研究

１　情報的文字資料の分析：

例：Ciniiなどの文献データベース、公式的なWebサイト、ガイドブック、ハンドブック・・、等の情報中心の記事；　※医師説明条件のビデオ

２　物語り的（ナラティブ）文字資料の分析

例：闘病記、感想文、ブログ、小説、体験談、DIPExなどのインタビューサイトなど、話者の語りを分析する場合　※患者談話条件のビデオ

★（2）混合研究法①質的研究との連携

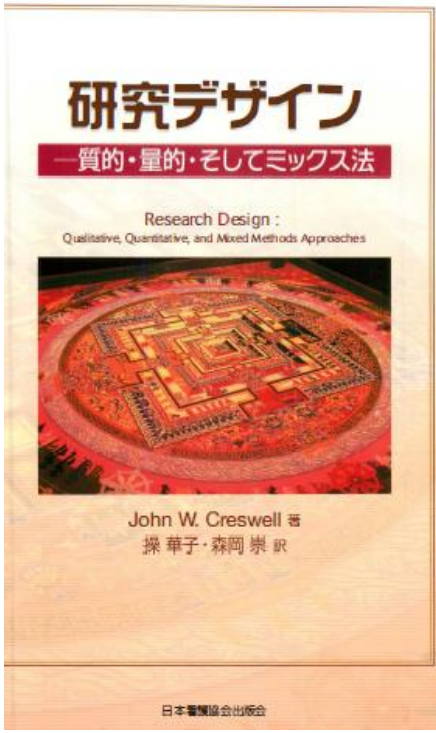
質的心理学や文化人類学や質的研究中心の看護学研究においては、テキストマイニングは量的研究として、エスノグラフィーやグラウンデッドセオリーなどによる患者・当事者の物語り（ナラティブ）の質的研究と連携することにより混合研究法を実現できる。

①順次的研究方略：質的データ収集と分析の後に、テキストマイニングによる分析を行ない、解釈の段階で統合する。質的分析で仮説生成的に得られた仮説をテキストマイニングにより量的な裏付けをもって検証することができる。質的研究データの再分析としてのテキストマイニングである。

例：精神障害者の居場所の研究（小平、未公刊）

②並行的トライアンギュレーション：テキストマイニング研究を質的研究と独立して研究；並行する2つの方法で得た結果を分析の段階で統合する。質的研究とテキストマイニングの分析結果は対等のものとして統合される。質的研究者との共同研究プロジェクト。

Creswell 2003 操・森岡訳　2007　研究デザイン　質的・量的・そして混合研究法　日本看護協会出版会



★（3）混合研究法②量的研究との連携

- 心理学では（質的心理学・臨床心理学以外では）量的資料による量的分析が主流である。そのような量的研究と連携してテキストマイニングは次のように位置付く。

①並行的入れ子方略：量的データと質的データを同時に収集する。質問紙の評定値と自由記述の結果のように、量的データ・量的分析が優位で支配的である。補足的に行なわれるテキストマイニングによる文字データ分析により、量的分析結果を裏付けたり、補足したり、矛盾点を指摘したりできる。

②並行的トライアンギュレーション：テキストマイニング研究を量的研究と独立して研究；並行する2つの方法で得た結果を分析の段階で統合する。量的研究とテキストマイニングの分析結果は対等のものとして統合される。量的研究者との共同研究プロジェクト。

★テキストマイニングと研究目的のタイプ

探索的　仮説生成的　仮説検証的

個性記述的

法則定立的

- 論理科学的思考モード
- 物語的思考モード

★「水平社宣言」のテキストマイニング

長い間虐められて來た兄弟よ、過去半世紀間に種々なる方法と、多くの人々によってなされた吾らの爲の運動が、何等の有難い効果を齎らさなかった事實は、夫等のすべてが吾々によって、又他の人々によって毎に人間を冒涜されてゐた罰であったのだ。そしてこれ等の人間を勦るかの如き運動は、かえって多くの兄弟を墮落させた事を想へば、此際吾等の中より人間を尊敬する事によって自ら解放せんとする者の集團運動を起せるは、寧ろ必然である。

兄弟よ、吾々の祖先は自由、平等の渇迎者であり、實行者であった。陋劣なる階級政策の犠牲者であり、男らしき産業的殉教者であったのだ。ケモノの皮を剥ぐ報酬として、生々しき人間の皮を剥ぎ取られ、ケモノの心臓を裂く代價として、暖かい人間の心臓を引裂かれ、そこへ下らない嘲笑の唾まで吐きかけられた呪はれの夜の惡夢のうちにも、なほ誇り得る人間の血は、涸れずにあった。そうだ、そして吾々は、この血を享けて人間が神にかわらうとする時代にあうたのだ。犠牲者がその烙印を投げ返す時が來たのだ。殉教者が、その荊冠を祝福される時が來たのだ。

吾々がエタである事を誇り得る時が來たのだ。

吾々は、かならず卑屈なる言葉と怯懦なる行爲によって、祖先を辱しめ、人間を冒涜してはならなぬ。そうして人の世の冷たさが、何んなに冷たいか、人間を勦る事が何であるかをよく知ってゐる吾々は、心から人生の熱と光を願求禮讃するものである。

水平社は、かくして生れた。

人の世に熱あれ、人間に光りあれ。

綱領

一、特殊部落民は部落民自身の行動によって絶対の解放を期す

一、吾々特殊部落民は絶対に経済の自由と職業の自由を社会に要求し以て獲得を期す

一、吾等は人間性の原理に覚醒し人類最高の完成に向って突進す

大正十一年三月三日　全國水平社創立大會

「水平社宣言」テキストのパラドックス

歴史的な部落差別批判の声明であるが、「差別」「偏見」の文字は無い。

具体物としての現象（＝単語）がないからといって、本質が存在しないということではない。

効楽安近短モデル（省エネモデル）（伊藤, 2008）によるソフトウェアの選定

- 効楽安近短モデルとは教材選定、会議進行理念、会社経営、教育方法、サプリメント選定、ダイエット方評価、アンチエイジング法評価、**卒論執筆**などの分野で適用可能なモデルである
- 効果：目標達成の度合
- 楽しい、続けたい、充実した時間
- 安価・安全・安心
- 近接可能性（アクセス可能性）
- 短時間

★効楽＝達成的/時間的ベネフィットの最大化、
安近短＝心理的/時間的/金銭的/資源的コストの最小化

看護学における混合研究法の例

Text Mining Studio

トップ > 参考書籍・論文 > 2014度 VMStudio & TMStudio 学生研究奨励賞 結果発表

2014年度 学生奨励賞の結果発表

今年度も多数のご応募誠にありがとうございました。奨励賞の結
た。
※同一賞内の表示は学生研究奨励賞への応募日付の順に掲載

優秀賞

◆「医療専門職連携 実習 における学生の経験と学びのプロセスを可視化する試み」

加藤 尚子 様
国際医療福祉大学 医療福祉・マネジメント学科
発表論文 pdf

▸▸受賞コメント

この度は、優秀賞に選出していただき、誠にありがとうございました。ご指導をいただいた服部兼敏先生と、株式会社NTTデータ数理システムの皆様に感謝申し上げます。TM Studioに巡り合い暗中模索的に研究を進めておりましたが、この賞をいただいて大変励みになりました。

◆「不安発 出来た感経由 出来た行き」

樺島 稔 様
東京医科歯科大学大学院 保健衛生学研究科
発表論文 pdf

▸▸受賞コメント

この度は、優秀賞という光栄な賞に選出して頂き誠にありがとうございます。TMStudioは、文章の奥に隠れている情報を目に見える形で示してくれる面白い分析ツールで、初心者の私も気が付くとのめり込んでいました。これからもTMStudioを自らの研究に役立てていけたらと考えております。最後になりましたが、熱心に指導して頂いた先生方、応援して頂いた先輩方、本当にありがとうございました。

審査員特別賞

◆「外国人看護師の理解のための情報提供の検討」

菅野貴文 様
広島国際大学 看護学部 看護学科
発表論文 pdf

▸▸受賞コメント

この度は、特別賞を選出していただき、誠にありがとうございました。数量的なデータだけでなく、Text Mining Studio で自由記述を分析することで、その数値が示す裏側まで知ることができ、より深い分析をすることができました。この賞を励みに、今後も精進していきたいと思います。

◆「国立看護大学系のアドミッションポリシーの分析」

北谷 幸寛 様
富山大学大学院 医学薬学教育部

2015年度学生研究奨励賞に応募しましょう（1）

http://www.msi.co.jp/tmstudio/stu14result.html

2014年度 学生研究奨励賞のご案内

2014年度の学生研究奨励賞は終了しました。
2015年度の応募をお待ち申し上げております。(7月頃に応募要項を発表します)
2014 年度の学生研究奨励賞の結果はこちら

弊社のマイニングツール（ Visual Mining Studio / Text Mining Studio ）を用いた、学生の学術研究の支援・啓蒙および発表の場の提供を目的として、学生奨励賞の制度を開催しております。

募集要項詳細はこちら pdf

受賞研究は、毎年行われる 数理システムユーザーコンファレンス にて、ポスターセッション形式で発表されます。また、最優秀賞を受賞された研究につきましては、可能な限り当日の口頭発表をお願いしております。

本年度も学生の皆様のご応募を心よりお待ちしております。
なお、昨年度の受賞者の皆様の研究内容は、2013年度 VMStudio & TMStudio 学生研究奨励賞 結果発表 のページでご覧いただけます。

※7月1日よりお申込みを開始しました

学生研究奨励賞の内容

2014 年度の審査は、株式会社NTTデータ数理システム にて行います。

最優秀賞 1 件
賞金 10 万円
優秀賞 1 ～ 2 件
賞金 5 万円
佳作 数件
賞金 1 万円

受賞者は、数理システムユーザーコンファレンスの当日（ 2014 年 11 月 21 日 ）に会場で発表いたします。

また、最優秀賞を受賞された方には、事前に連絡を差し上げますので、可能な限り数理システムユーザーコンファレンスでの口頭発表をお願いいたします。この場合、国内に限り、別途旅費を支給いたします。

応募資格

- 応募者 が 数理システムユーザーコンファレンス開催時（ 2014 年 11 月 21 日 ）に、学生（ 大学院生・社会人学生も含む ）であること。
- Visual Mining Studio または Text Mining Studio を利用した研究であること。

応募の流れ

お手続き	締切日	内容
エントリ	~~2014 年 10 月 17 日 (金) 午前10 時まで~~	お申込みは締め切りました。
研究内容提出	2014 年 10 月 31 日 (金) 午前10 時まで	研究内容を電子媒体でNTTデータ数理システムまでご提出いただきます。 詳細は下記 研究内容提出方法 をご覧下さい。
受賞研究発表	2014 年 11 月 21 日 (金)	数理システムユーザーコンファレンスにて受賞者を発表いたします。 最優秀賞受賞者には、事前に連絡を差し上げますので可能な限りコンファレンスでの口頭発表をお願いいたします。

もし、エントリ後に研究内容が提出できない事態が生じてリタイアを希望される場合、リタイアされる旨 と その理由、及び 当初予定していた研究の内容 を提出していただきます。

製品の貸与

貸与期間:エントリー時から、2014年11月30日まで

エントリ方法

エントリ受付期限
~~2014 年 10 月 17 日 (金) 午前10時~~ **お申込は締め切りました**

以下を明記の上、vtc14@msi.co.jp までお送りください。

- 予定発表タイトル
- 発表者名
- 所属（ 学校・学部・研究室・指導教官のお名前 ）
- E-Mail アドレス
- Visual Mining Studio・Text Mining Studio 貸与希望の方はその旨ご明記ください。

2015年度学生研究奨励賞に応募しましょう（2）

http://www.msi.co.jp/tmstudio/stu14result.html

折り返し、エントリ受理のご連絡をお送りいたします。発表者が複数の場合は代表発表者をご明記ください。

研究内容提出方法

提出期限

2014 年 10 月 31 日（金）午前10時

フォーマット

電子媒体 MS-Word・MS-PowerPoint・PDF のいずれかの形式
提出物だけで（口頭の説明無しに）研究内容がわかるものであること
論文の総ページ数：なるべく MS-Word の場合は 10 ページ以内、
MS-PowerPoint の場合は 40 スライド以内としてください。

提出方法

E-Mail もしくは 郵送

E-Mail による提出の場合は、vtc14@msi.co.jp までご送付ください。郵送の場合は、上記提出期限必着で、弊社までお送りください。紙ベース でのご提出はご遠慮ください。

その他

- 研究にご利用になるデータは、結果を**公開可能であるものをご利用下さい**。特に企業からの提供を受ける場合、NTTデータ数理システムの学生研究奨励賞に応募する旨の確認を必ず取ってください。結果を公開できないデータを利用された研究は、選外とさせていただきます。
- 提出された研究内容に関して、入賞分原稿を複製配布し、また弊社 Web 上での掲載を行う以外に、NTTデータ数理システムが権利を主張することはありません。
- 連名での応募も受け付けておりますが、受賞は研究単位となります。
- 応募者の数理システムユーザーコンファレンス参加は義務としませんが、他の発表内容を閲覧するなど非常に有意義と思われるため、ぜひ参加をおすすめします（参加無料）。
- 2013年度受賞者の方々の研究内容は 2013年度 VMStudio & TMStudio 学生研究奨励賞 結果発表のページでご覧いただけます。