

2019年度学習支援部会企画
学生スタッフによる
学習支援の研究手法

ワークショップを通じたデータ分析

TAによる学習支援を研究する －言語データの収集と分析－

椿本 弥生
(東京大学)

mio@fye.c.u-tokyo.ac.jp

日本リメディアル教育学会 第15回全国大会・総会
学習支援部会ワークショップ

2019年8月26日（月）

於 金沢工業大学 扇が丘キャンパス

自己紹介

● 椿本 弥生（つばきもと みお）

● 職歴

- ・ 東京大学 大学総合教育研究センター・同 情報学環
- ・ 公立はこだて未来大学 システム情報科学部
- ・ 東京大学 大学院総合文化研究科・教養学部附属 教養教育高度化機構
初年次教育部門 特任准教授（現在）

● 関心領域

- ・ 文章の心理学、特に文章評価
- ・ 教授学習過程の支援・教育効果測定

● 主な研究テーマ

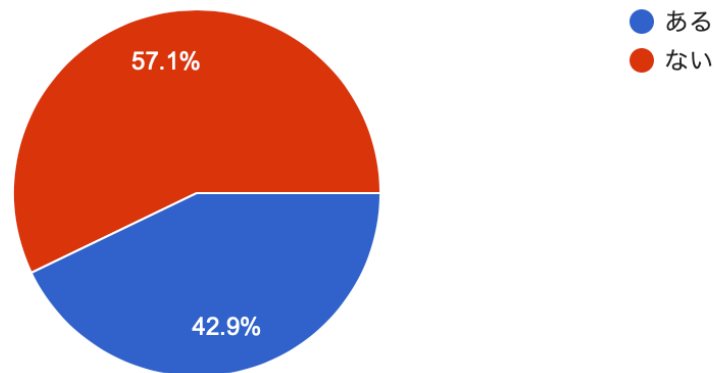
- ・ TAの学習効果の測定
- ・ 初年次教育プログラムの遅延効果の測定



事前アンケートより

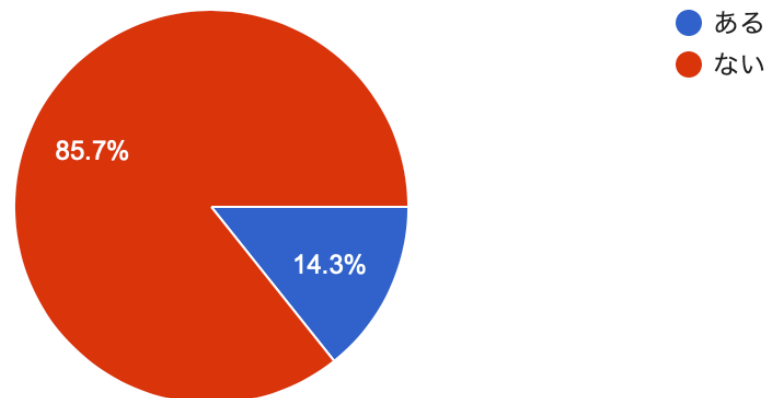
ご自分の研究や業務のために、量的内容分析を行ったことはありますか？

7件の回答



KH Coderを使ったことはありますか？

7件の回答



事前アンケートより（動機）

- テキストマイニングについて知りたい
- （KH Coderの）利用方法を知りたい
- 学生アンケートの記述データを分析する際の
手がかりを得たい
- （KH Coderの）使い方を徹底したい
- マインドマップのワードの分類や傾向分析を
したい
- 量的内容分析がどのように行われるのか
見てみたい

本日の目的・目標

目的

テキストマイニングを用いた研究を始めるための準備をする

目標

- テキストマイニングとは何かを説明できる
- KH Coderでの分析の流れを経験できる
- 仮説、従属変数、独立変数を考えられる

本日の内容

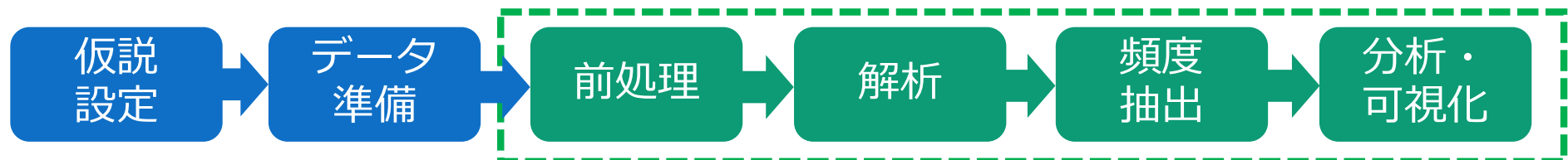
- テキストマイニングとは

- ・研究サイクル
- ・研究のチェックポイント

- KH Coderによる分析事例

- ・独立変数と従属変数
 - ・TAへの質問紙調査の分析
 - ・歌詞の分析

- KH Coderを用いた言語分析（演習）

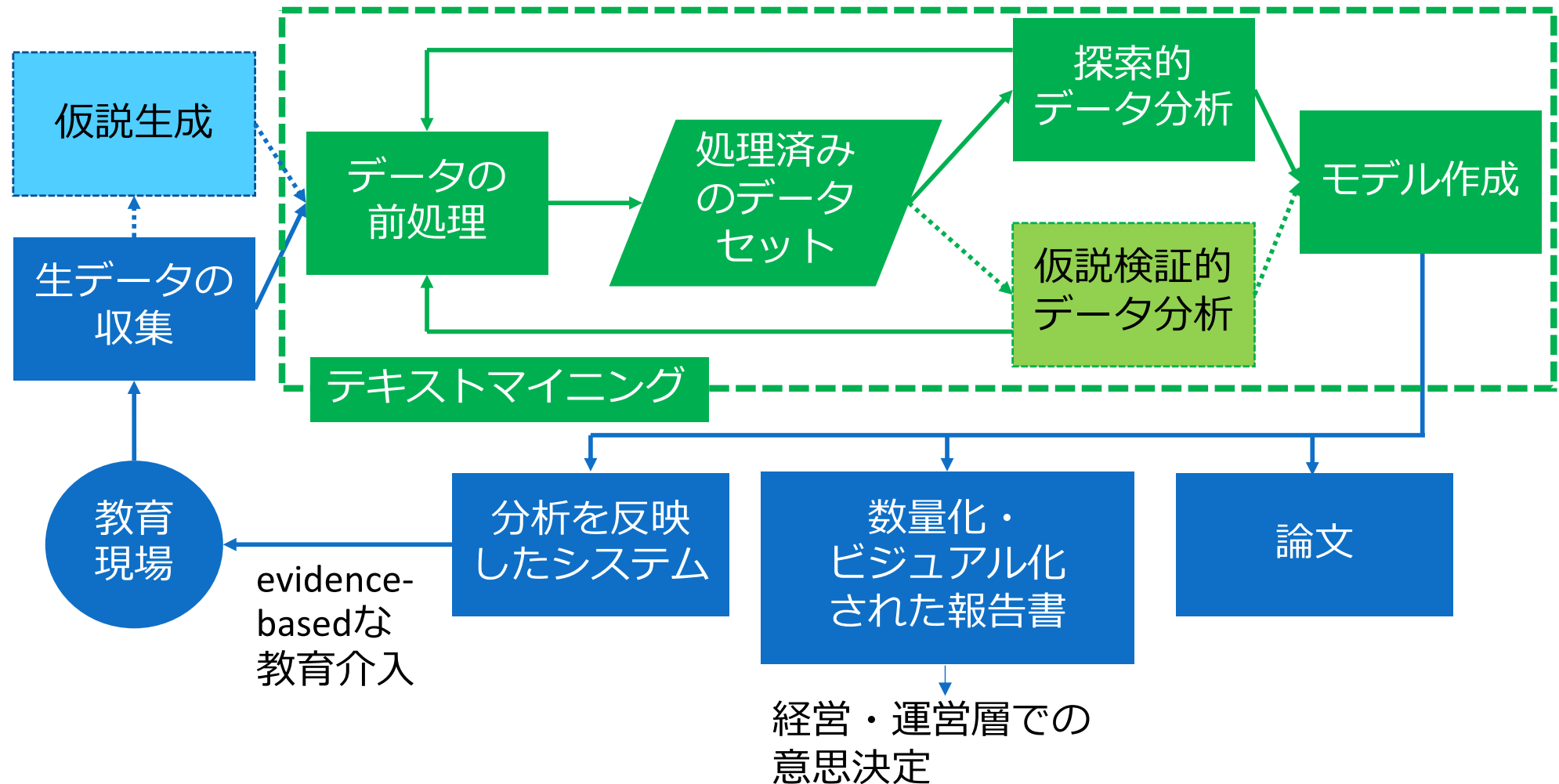


テキストマイニングとは

テキストマイニングとは

- 見たままに読めばわかる記述を、数量データに変換し、言語を量的に処理し、データから知見を発見（mining）する / 仮説を検証する
- 言語データ
 - ・ 自由記述、作文、ミニッツペーパー、教科書、ノート、学習記録、日常会話、議論、インタビュー、プロトコル、SNS・メール・チャット・掲示板などのログ、学習動画のセリフとキャプション（MOOC等）、ラジオ音声...など
- **いずれにも語（word）、文（sentence）、文章（text/document）の単位がある**
 - ・ 語は品詞別に注目する場合もある。また、不要な品詞は削除する場合もある。
 - ・ 文の長さや、文内／文間の係り受けも分析対象になりうる。
 - ・ 文章内の段落ごとに分析する場合もある。
- **文章タイプやメディアで言語データの特徴が異なる可能性**
 - ・ 意見文、生活文、説明文など文章タイプが異なれば、文章の論理構成も異なる
 - ・ 書かせるメディアによっても言語使用が特徴づけられる可能性あり

テキストマイニングを用いた (教育実践) 研究サイクル



データをどう分析するかだけでなく、 分析目的の明確化がまず大事

サイクルに関するチェック項目

- ✓分析の目的は明確か
- ✓仮説生成か仮説検証か
- ✓どんな言語データをどれだけ持っているか
- ✓データの量と質は十分か
 - ✓不十分なら、データを追加で取得できるか
- ✓関連するデータはあるか（量的・質的）
 - ・テスト、アンケート、行動観察の結果など
- ✓全てのデータに関する理解は十分か
 - ・十分でないなら、関連する担当者や対象者にアクセスできるか
- ✓分析のサイクルを回せそうか

KH Coderを用いた 分析事例

問題と目的

- 複数キャンパスに所属する300名以上のTA
- 授業内TAの研修や対面相談の機会が少ない
 - 実務問題に対応できているのか？
 - そもそも、どんな問題があるのか？

◆授業内TA の意見を収集

◆内容分析から授業内TA 育成の課題を発見

方法

- 質問紙調査
 - 2017年度の初ゼミ理科の授業内TA 386名
 - Web上の質問紙に個別回答
 - 回答時間、場所、機器などに制限なし
 - 謝礼あり（1,000円）
- 項目
 1. 楽しかった，やりがいと感じたこと
 2. 辛かった，大変だったこと
 3. 疑問や問題に思ったこと
 4. TAの業務や制度で，継続してほしいこと
 5. TAの業務や制度で，改善してほしいこと
 6. 授業外TAとの連携に必要な仕組み
 7. TA研修にあってよかった・あればよかった内容
 8. その他，TA業務に関する意見
 - 初ゼミ理科TA 担当回数
- 有効回答75件

方法

- 内容分析

- KH Coder v3.Alpha.11a

1. クリーニング

- 誤字、表記揺れ修正 ex. 諸年次 → 初年次

2. 前処理

- 記号などを無視 ex. 中黒（・）、「」、（）
 - 文脈依存の特殊表現を強制抽出 ex. ITC-LMS

3. コーディングルール設定

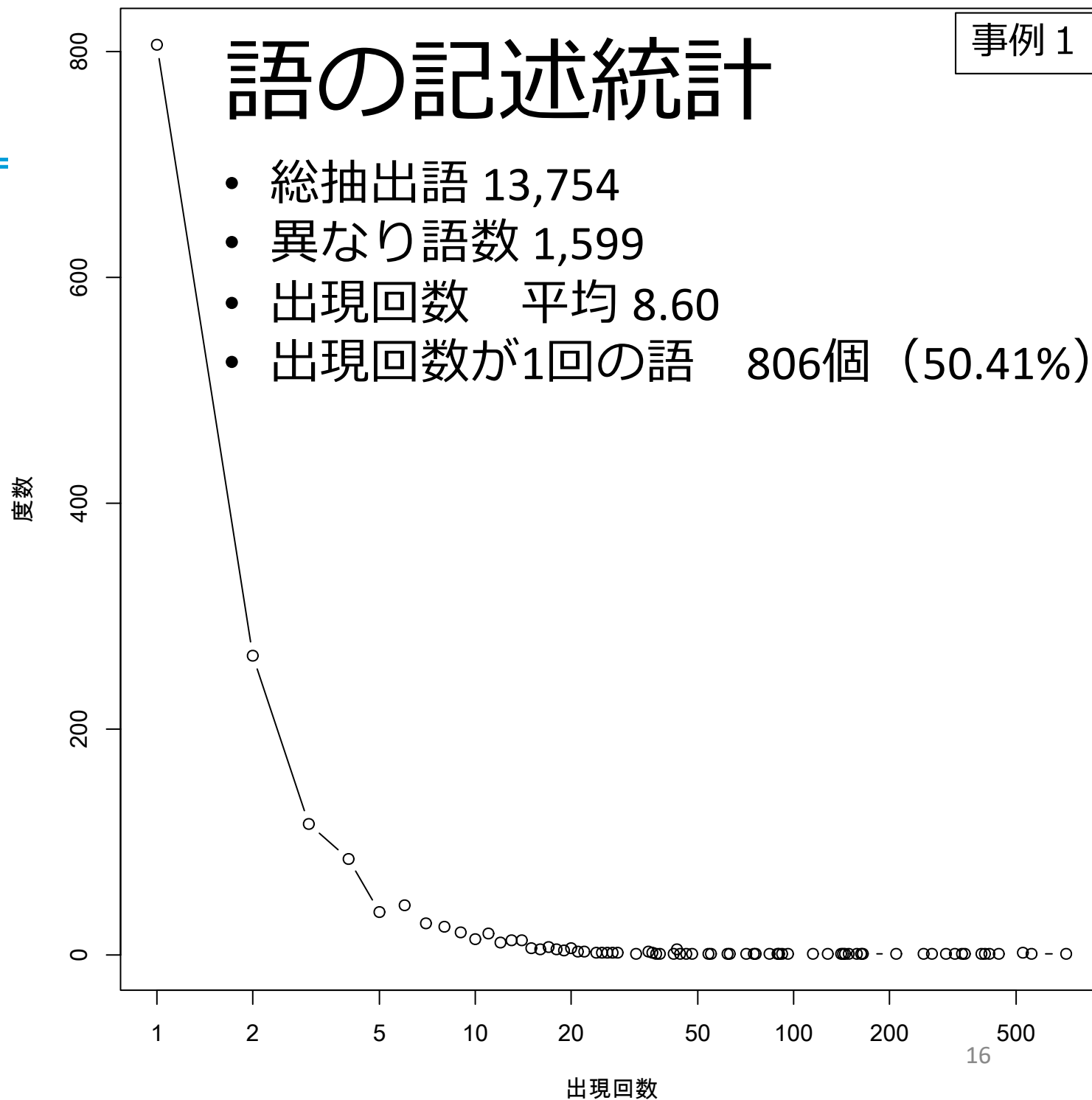
- 特定の記述をいくつかのカテゴリーに分類
 - 類義語をまとめる、上位概念を付与する
 - ex. 先生、教員、教授 → 「教員」
 - ex. 情報共有、情報提示、情報交換など → 「情報の活用」

結果

事例 1

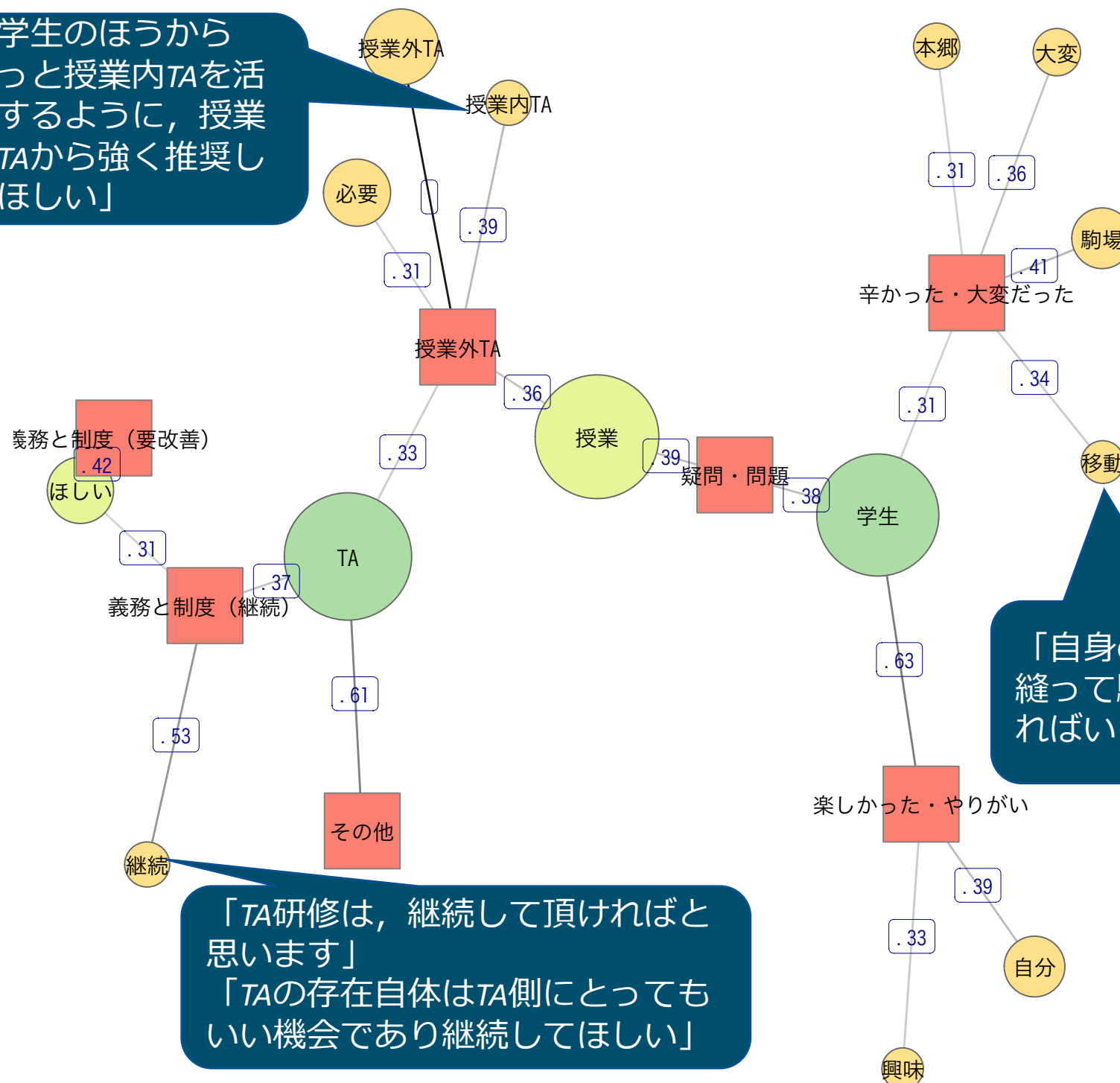
語の記述統計

- 総抽出語 13,754
- 異なり語数 1,599
- 出現回数 平均 8.60
- 出現回数が1回の語 806個 (50.41%)



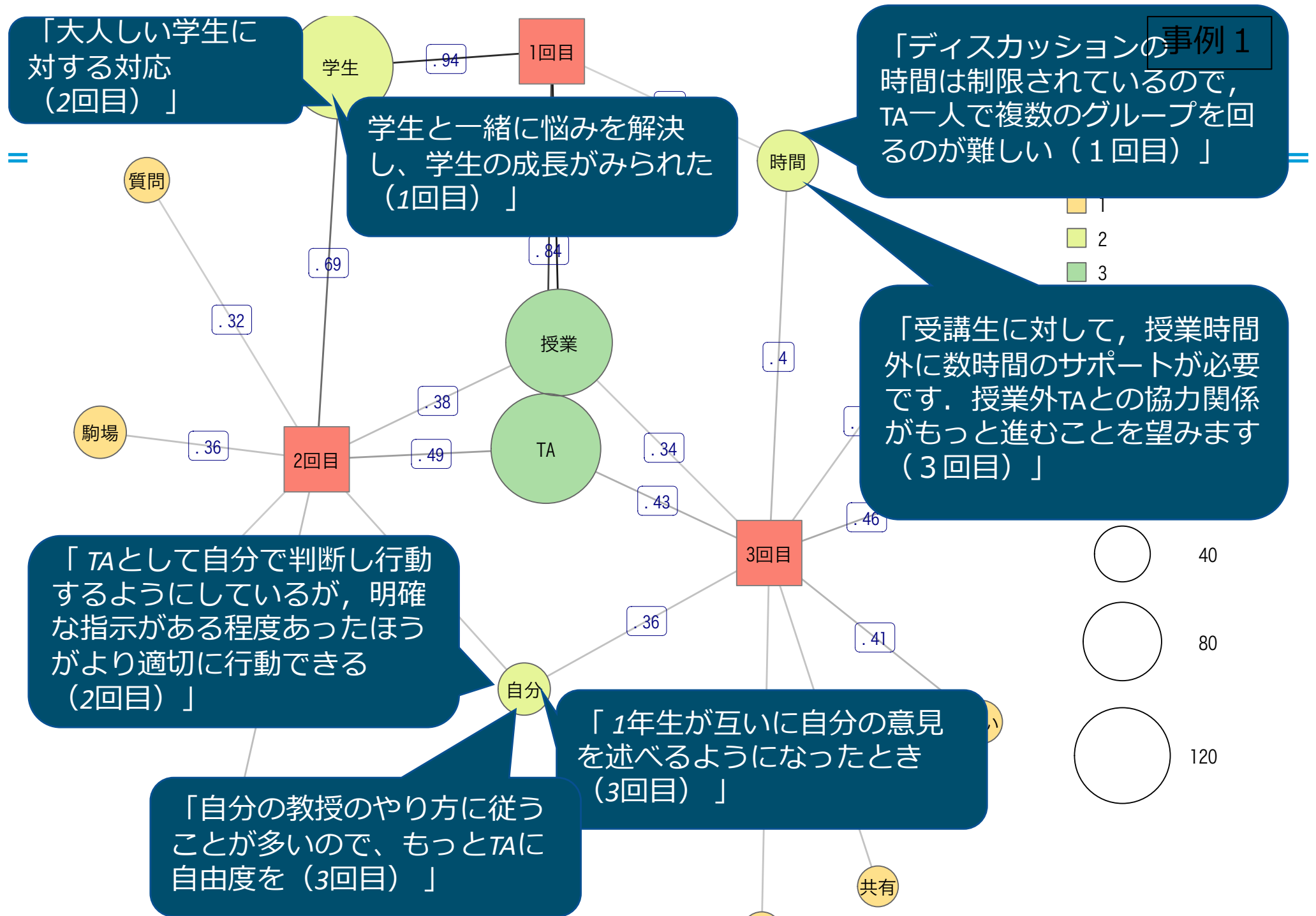
事例 1

「学生のほうから
もっと授業内TAを活用するように、授業
外TAから強く推奨し
てほしい」



「自身の研究の合間を
縫って駒場に移動しなけ
ればいけなかったこと」

「TA研修は、継続して頂ければと
思います」
「TAの存在自体はTA側にとっても
いい機会であり継続してほしい」



分析事例

- 歌詞

RQ(問い)と仮説の関係

	問いのレベル	仮説の有無	研究デザイン
レベル1	これは何であるか？	なし	質的記述研究
レベル2	何が起きているのか？	なし	質的記述研究 または 実態調査に基づく 量的記述研究
レベル3	これらには 関係 があるのか？ これらには 差 があるのか？	あり	量的仮説検証型研究
レベル4	これらには 因果関係 があるのか？	あり	量的因果関係検証型 研究

戈木クレイグヒル（2006）を参照・改変

●NGな仮説の書き方

～を検証する、～を確認する、～の影響 など

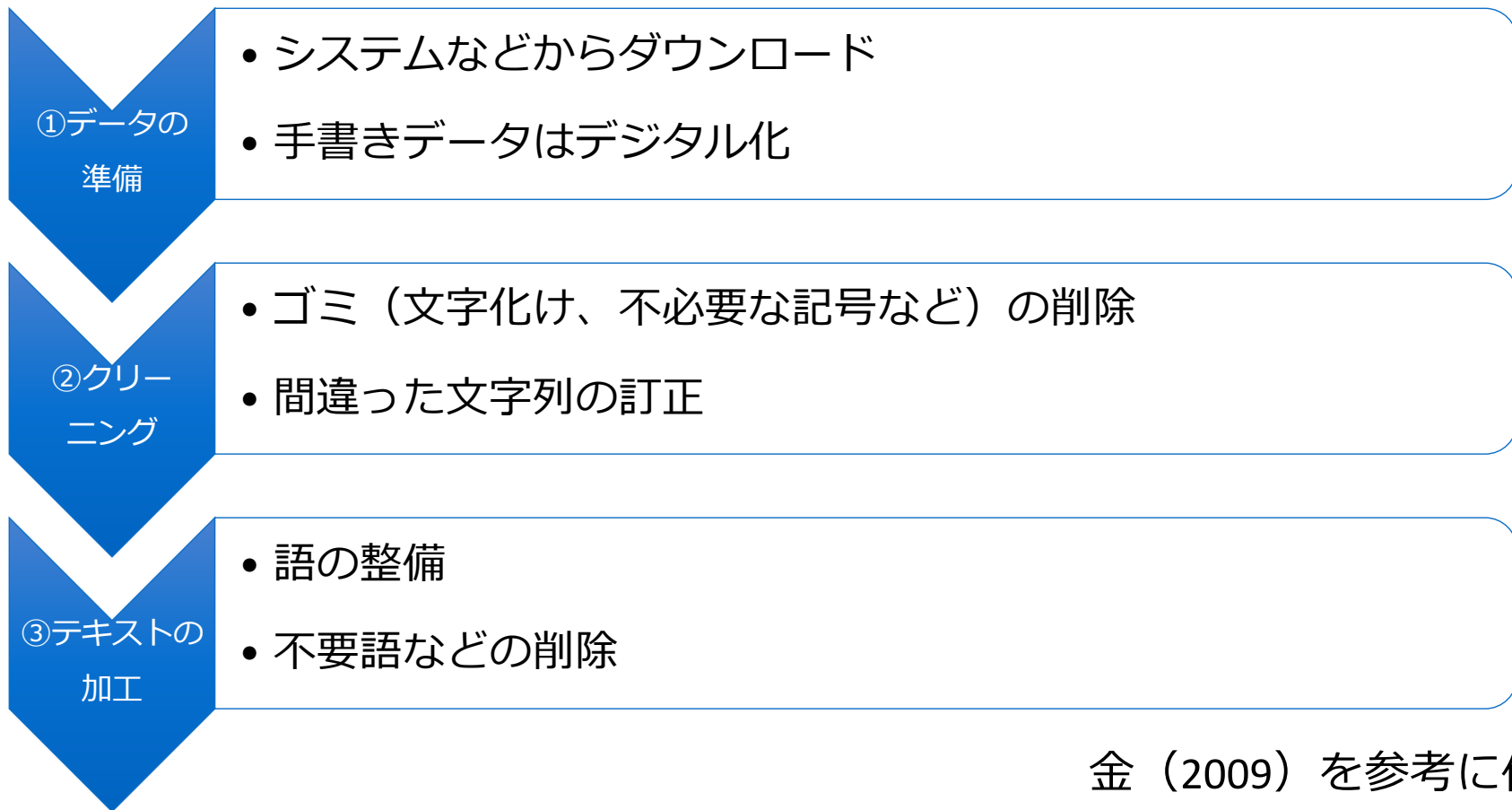
●OKな書き方

例) ～という状況では、●●テストの得点が高い群のほうが、低い群よりも、
▲と■の概念間の共起頻度が有意に高くなる

ここまでで
ご質問はありますか？

KH Coderを用いた 言語分析

TMの一般的な手順（１）



TMの一般的な手順（1の説明）

• 1. 文字型データの準備

- システムからダウンロード／手書きのものはデジタル化
- csvやtxt形式など

• 2. データのクリーニング

- 「、」と「,」や、「。」と「.」の統一
- 不要なスペースや改行（記号）の消去
- 全角・半角の統一
- 誤った文字の修正

【ポイント②】

2と3で何をしたかを記録しながら作業する。後から全てを思い出すのは困難！

→論文の「手続き」で書く。書かなくても、査読で指摘される。

• 3. テキストの加工（1）不要語、複合語、同義語、新語・専門用語の設定

- **不要語の例**：「私」← 意味をなさないことが多い頻出語（ただし研究目的によっては必要になる）
- **複合語**：「反転学習」← 設定しないと「反転」と「学習」として集計される
- **同義語**：「eラーニング」「e-ラーニング」「eLearning」／「コメント」「意見」← 設定しないと全て異なる語として集計される
- **専門用語**：「発達」の最近接領域」← 設定しないと「発達」「の」「最近」「近接」「領域」などとなる

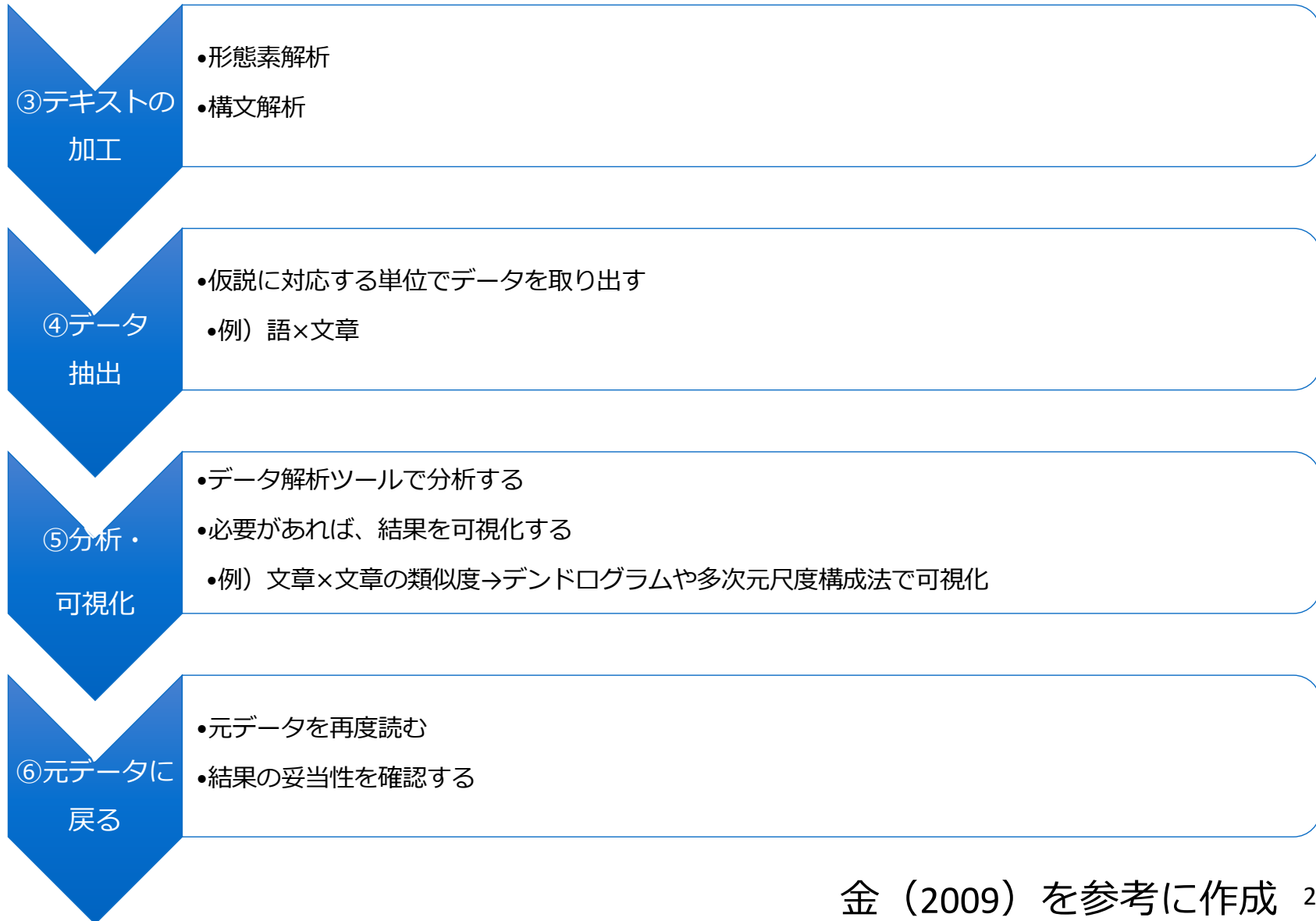
- ツール搭載の類義語辞書で、ある程度整備可能。

しかし、自らの目的や仮説に合わせたコーディング・ルールの整備は必須

- ここまでの実行結果を確認し、2と3を繰り返す

ここまでで、文、文章、句読点の単位での解析は可能

TMの一般的な手順（2 つづき）



TMの一般的な手順（２の説明）

- 3.テキストの加工 （２）各種解析（形態素解析、構文解析）

- 文を語単位に切り分ける & 語に品詞情報を付与する
- 文を文節に切り分ける & 文節の係り受け関係の情報を付与する
 - これ以降、単語や文節単位で分析できるようになる

- 4. データ抽出

例) 列：文章、行：単語 の頻度マトリックスを作成

- 例) 「eラーニング」と「繰り返す」と「宿題」は共起している（＝一緒に用いられている）
- 例) 「eラーニング」と「効果」は、「孤独」よりも共起頻度が高い
- 例) 「eラーニング」と「プレゼンテーション」は共起していない
- 行の単位：文、文節、文章、学習者など

- 5. **必要に応じて & 目的に合った手法で可視化**

- 6. **元の言語データに戻って、結果の妥当性を確認**

- 機械による意味処理と、人間による意味処理の結果には差がある

まとめ

- 何でも質疑応答

こんなツールもあります

- **形態素解析ウェブアプリUniDic-MeCab**

- <http://www4414uj.sakura.ne.jp/Yasanichi1/unicheck/>
- 形態素解析のみ
- 結果をExcelでダウンロードできる
 - すぐ統計検定したい場合、とりあえず形態素解析の結果をすぐ見てみたい場合向け

- **日本語作文の自動採点**

- 森リン
 - <https://www.mori7.net/moririn/moririn1200.php>
- Jess
 - <http://tk2-203-11024.vs.sakura.ne.jp/jess/>
- 採点結果（数値）をExcelなどに移して分析する

参考文献

- KDnuggets (2016) The Data Science Process.
<https://www.kdnuggets.com/2016/03/data-science-process.html> (Accessed 2019.8.24)
- 金明哲 (2009) テキストデータの統計科学入門.
岩波書店
- 戈木クレイグヒル滋子 (2006) グラウンデッド・セオリー・アプローチ ―理論を生み出すまで.
新曜社
- 塚本邦尊, 山田典一, 大澤文孝, 中山浩太郎, 松尾 豊
(2019) 東京大学のデータサイエンティスト養成
講座. マイナビ出版
- 椿本弥生 (2018) 研究を指向した初年次教育におけるTAの育成課題. 日本教育工学会研究報告集
(JSET19-1) , 485-488