@島根大学人間科学部
2020.07.26

応用心理学Ⅰ
データ収集後探索的解析（テキストマイニング）　その３

テキストマイニング
実習編

明治大学　研究・知財戦略機構
佐藤浩輔

位置づけ

- 7/25 午後　Rブートキャンプ
- 7/26 午前　質的データの量的コーディング*
- 7/26 午後　（前半）テキストマイニング講義編
  （後半）テキストマイニング実習編←これ

*中分遥先生（高知工科大学）担当

第２部
実習：Rを使ったテキスト分析

実習編のアウトライン

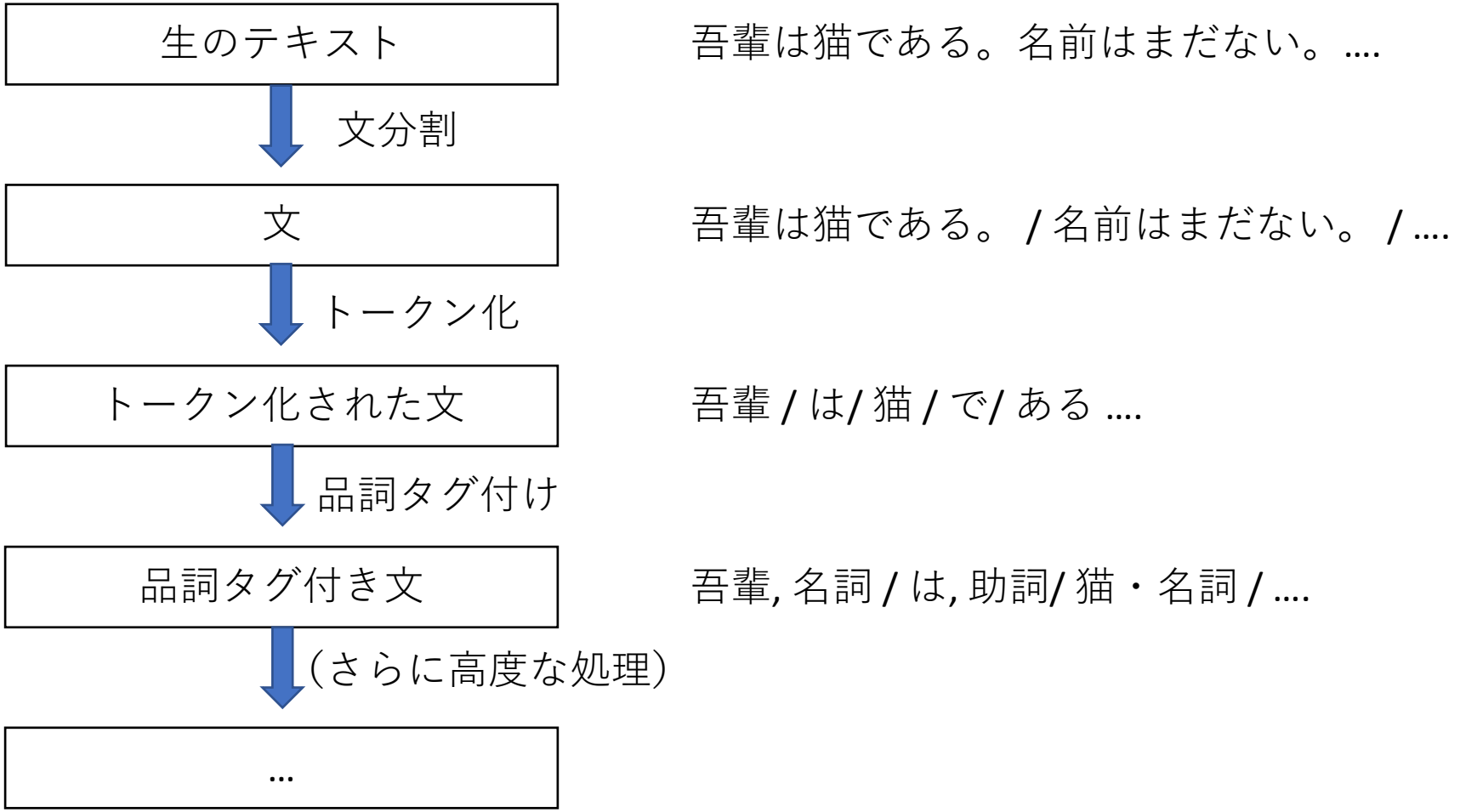
- 前半
  - 実習：データ前処理
    - 前処理の流れ
    - クリーニングの実際
    - データハンドリング入門
      - 単語の頻度
      - キーワード抽出
- 後半
  - 実習：データ分析
    - 主成分分析
    - ネットワークグラフによる共起の可視化
    - クラスター分析（時間があれば）
  - 結果の報告と解釈
    - 結果の書き方に関するヒント
  - クロージング
    - 講義全体のまとめ
    - 課題について

第 2 部
実習①データ前処理編

なぜ前処理が必要か

- 通例心理学で扱うようなデータ
  →きれいに整形されている（べき）
    - 表形式・構造化データ
    - ノイズは少ない
    - 変数は少ない
    - すぐに分析できる
- テキストデータ
  →dirty
    - 整形されていない・非構造化データ
    - ノイズがたくさん
    - 変数が膨大
    - そのままでは扱えない

情報抽出アーキテクチャの例



Bird, Klein, Loper(2010) を改変

生のテキスト（平テキスト）

吾輩は猫である。名前はまだ無い。

どこで生れたかとんと見当がつかぬ。何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。吾輩はここで始めて人間というものを見た。しかもあとで聞くとそれは書生という人間中で一番獰悪な種族であったそうだ。

…

――夏目漱石『吾輩は猫である』

- 何のタグもつけられていないテキスト
- このままでは分析できない

文に分割

吾輩は猫である。/ 名前はまだ無い。/ どこで生れたかとんと見当がつかぬ。 / 何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。 / 吾輩はここで始めて人間というものを見た。/しかもあとで聞くとそれは書生という人間中で一番獰悪な種族であったそうだ。 /

…

- 文ごとに分割されたテキスト

単語(トークン)に分割

吾輩/は/猫/で/ある/。
名前/は/まだ/無い/。
どこ/で/生れ/た/か/とんと/見当/が/つか/ぬ/。
…

- 文を単語ごとに分割
  →こうやって分析できる単位に分割していく

クリーニングと前処理

データ前処理

- クリーニング
  - ノイズを取り除く
  - 辞書を整備する
- 前処理
  - テキストの分割
    - 文書→文→単語に分割
  - 単語の処理
  - 様々なタグをつける
- データハンドリング・構造化
  - 様々な分析手法が可能なようにデータを整える

クリーニング

- 分析にかけられるよう、データを整備する
  - 誤字脱字のチェック
  - 辞書の整備
    - 専門用語
    - 新語・死語
    - 方言
- ノイズが多く混じっていると結果が歪む

クリーニングと前処理

形態素解析

- 英語などの言語：
  - 単語と単語の間に切れ目がある
    →スペースで分割できる
- 日本語のような言語
  - 単語と単語との間に切れ目がない
  - 分割できるように「分かち書き」する必要がある
    →「形態素解析」という技術を使う

  - 形態素解析ソフトを使うことで、
    - 分かち書き
    - 品詞タグ付け

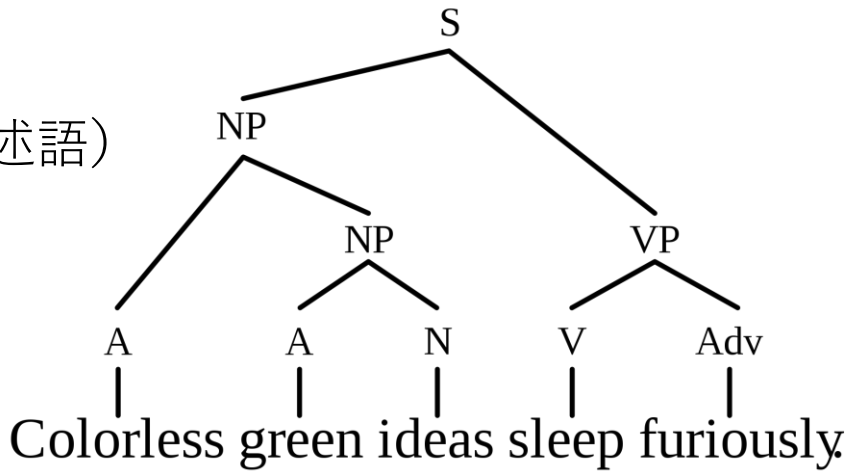
    を、一定の精度でまとめて処理することができる

単語の処理

- ストップワード stopword
  - 話題の種類と関連を持たない語(e.g. a, 冠詞、助詞)
  - 分析に必要なければ除外する
- ステミング stemming
  - 派生語などを含めて同じ素性とみなす処理
    - e.g. operate→oper, operational→oper
  - 単語の形状をみて一律に処理する
    - e.g. Porter's stemmer
- 見出し語化 lemmatization
  - 単語を基本形に戻す処理
  - 文脈を考慮して処理

タグ付け

- 付加情報をつける作業
  - 品詞 part of speech (POS)
    - 名詞
    - 動詞
    - 形容詞
    - 副詞
    - 助詞
  - 構文情報
    - 文法的な構造（e.g. 主語, 述語）



クリーニングと前処理

構造化

- 構造化の例
  - bag of words
    - 単語の頻度
    - 語順の情報は無視される
  - n-gram
    - 連続するn語の組み合わせ
      - 連続する 2 語→bigram
      - 連続する 3 語→trigram
  - 共起
    - 同じ文内に出現した語の組み合わせ

bag of words

this	1
is	1
a	1
pen	1

bigram

this-is	1
is-a	1
a-pen	1

共起

(this, is)	1
(this, a)	1
(this, pen)	1
(is, a)	1
(is, pen)	1
(a, pen)	1

→構造化されたデータを分析・可視化する

実習

事前準備

- ソフトウェア
  - R version 3.4.3以上
  - MeCab
    - Windowsの場合はShift-JIS版がインストールされていること
    - IPA辞書がインストールされていること
- Rパッケージ
  - RMeCab
  - tidyverse
    - dplyr(tidyverseに含まれる)
    - stringr(tidyverseに含まれる)

RMeCabの使い方

- RMeCab
  - MeCab: 日本語の形態素解析エンジン
  - RMeCab: R上でMeCabを使うためのパッケージ(石田, 2017)
    - 作者のページに各関数の詳しい解説がある
      - http://rmecab.jp/wiki/index.php?RMeCabFunctions
    - RMeCabC(): 文字列を形態素解析して返す
    - docMatrixDF(): フォルダ内のファイルごとに解析する
    - docDF(): データフレームの行ごとに解析する
    - 解析時に辞書ファイルを指定できる

クリーニングの実際

- 例文：2019年のインタビューより
  - Q:「やまいり？やまはいって木を丸める？」
  - A:「切ってきて、丸めて、家の屋根にしげておくん。それはてんかごめんけどどこのやまいってきってきてもよいということになっちょった。」

- これを形態素解析にかけてみる

1. **動詞:** '切っ'
2. **助詞:** 'て'
3. **動詞:** 'き'
4. **助詞:** 'て'
5. **記号:** '、'
6. **動詞:** '丸め'
7. **助詞:** 'て'
8. **記号:** '、'
9. **名詞:** '家'
10. **助詞:** 'の'
11. **名詞:** '屋根'
12. **助詞:** 'に'
13. **形容詞:** 'しげ'
14. **助詞:** 'て'
15. **動詞:** 'おく'
16. **助動詞:** 'ん'
17. **記号:** '。'
18. **名詞:** 'それ'
19. **助詞:** 'は'

方言または
誤字・聞き取りミス

20. **名詞:** 'てんか'
21. **感動詞:** 'ごめん'
22. **接続詞:** 'けど'
23. **名詞:** 'どこ'
24. **助詞:** 'の'
25. **助詞:** 'や'
26. **動詞:** 'まいっ'
27. **助詞:** 'て'
28. **動詞:** 'きっ'
29. **助詞:** 'て'
30. **動詞:** 'き'
31. **助詞:** 'て'
32. **助詞:** 'も'
33. **形容詞:** 'よい'
34. **助詞:** 'という'
35. **名詞:** 'こと'
36. **助詞:** 'に'
37. **動詞:** 'なっ'
38. **名詞:** 'ちょ'
39. **動詞:** 'っ'
40. **助動詞:** 'た'
41. **記号:** '。'

「てんかごめん」
（天下御免）の誤認識

誤字・聞き取りミス？

「やま　いって」が
「や　まいって」に

方言

クリーニングの実際

クリーニング方略

- ノイズを減らす
  - 誤字・脱字をチェックする
  - 句読点を入れる
  - 漢字に変換する
  - 表記揺れを減らす
- 固有名詞への対応
  - 辞書を整備する
- 方言への対応
  - 辞書を整備する
  - 標準語に変換する
  - 方言に関する知識を持つ

クリーニングの実際

修正の例

「切ってきて、丸めて、家の屋根にしげておくん。それはてんかごめんけどどこのやまいってきってきてもよいということになっちょった。」

「切ってきて、丸めて、家の屋根にすげておくん。それは**天下御免**でどこの**山**に**行**って**切**ってきてもよいということになっていた。」

Before

1. **動詞:** '切っ'
2. **助詞:** 'て'
3. **動詞:** 'き'
4. **助詞:** 'て'
5. **記号:** '、'
6. **動詞:** '丸め'
7. **助詞:** 'て'
8. **記号:** '、'
9. **名詞:** '家'
10. **助詞:** 'の'
11. **名詞:** '屋根'
12. **助詞:** 'に'
13. **形容詞:** 'しげ'
14. **助詞:** 'て'
15. **動詞:** 'おく'
16. **助動詞:** 'ん'
17. **記号:** '。'
18. **名詞:** 'それ'
19. **助詞:** 'は'

After

1. **動詞:** '切っ'
2. **助詞:** 'て'
3. **動詞:** 'き'
4. **助詞:** 'て'
5. **記号:** '、'
6. **動詞:** '丸め'
7. **助詞:** 'て'
8. **記号:** '、'
9. **名詞:** '家'
10. **助詞:** 'の'
11. **名詞:** '屋根'
12. **助詞:** 'に'
13. **動詞:** 'すげ'
14. **助詞:** 'て'
15. **動詞:** 'おく'
16. **助動詞:** 'ん'
17. **記号:** '。'
18. **名詞:** 'それ'
19. **助詞:** 'は'

Before

20. **名詞:** 'てんか'
21. **感動詞:** 'ごめん'
22. **接続詞:** 'けど'
23. **名詞:** 'どこ'
24. **助詞:** 'の'
25. **助詞:** 'や'
26. **動詞:** 'まいっ'
27. **助詞:** 'て'
28. **動詞:** 'きっ'
29. **助詞:** 'て'
30. **動詞:** 'き'
31. **助詞:** 'て'
32. **助詞:** 'も'
33. **形容詞:** 'よい'
34. **助詞:** 'という'
35. **名詞:** 'こと'
36. **助詞:** 'に'
37. **動詞:** 'なっ'
38. **名詞:** 'ちょ'
39. **動詞:** 'っ'
40. **助動詞:** 'た'
41. **記号:** '。'

After

20. **名詞:** '天下'
21. **名詞:** '御免'
22. **助詞:** 'で'
23. **名詞:** 'どこ'
24. **助詞:** 'の'
25. **名詞:** '山'
26. **助詞:** 'に'
27. **動詞:** '行っ'
28. **助詞:** 'て'
29. **動詞:** '切っ'
30. **助詞:** 'て'
31. **動詞:** 'き'
32. **助詞:** 'て'
33. **助詞:** 'も'
34. **形容詞:** 'よい'
35. **助詞:** 'という'
36. **名詞:** 'こと'
37. **助詞:** 'に'
38. **動詞:** 'なっ'
39. **助詞:** 'て'
40. **動詞:** 'い'
41. **助動詞:** 'た'

★MeCabの辞書の設定

- 追加辞書用csvを作る
  - MeCab/dic/ipadic配下に辞書のcsvがたくさん入っているので参考にする
- コンパイルする
  - MeCab/bin配下のmecab-dict-index.exeを使う

- RMeCabの関数実行時に辞書ファイルを指定

クリーニングの実際

★辞書CSVの作成

指定不要　　IPA 品詞体系を参考にする

表層形	左文脈ID	右文脈ID	コスト	品詞	品詞細分類1	品詞細分類2	品詞細分類3	活用型	活用形	原形	読み	発音
けん	*	*	1000	助詞	接続助詞	*	*	*	*	けん	ケン	ケン

同じカテゴリの単語を参考にする

けん,*,*,1000,助詞,接続助詞,*,*,*,*,けん,ケン,ケン

クリーニングの実際

★辞書のコンパイル

MeCabフォルダに移動し、

```
¥bin¥mecab-dict-index.exe -d デフォルト辞書フォルダ
-u 出力ファイル名
-f 文字エンコーディング（入力ファイル）
-t 文字エンコーディング（出力ファイル）
入力ファイル名
```

を実行（行を分けずに入力する）

・実行例

```
¥bin¥mecab-dict-index.exe -d dic¥ipadic -u
c:¥Users¥satoc¥projects¥nlp2019¥example.dic -f shift-jis -t
shift-jis c:¥Users¥satoc¥projects¥nlp2019¥shimane.csv
```

```
RMeCabC('まあそういうことで、余分な話はいいけん。')
```

1. **副詞:** 'まあ'
2. **連体詞:** 'そういう'
3. **名詞:** 'こと'
4. **助動詞:** 'で'
5. **記号:** '、'
6. **名詞:** '余分'
7. **助動詞:** 'な'
8. **名詞:** '話'
9. **助詞:** 'は'
10. **形容詞:** 'いい'
11. **助詞:** 'け'
12. **助詞:** 'ん'
13. **記号:** '。'

```
RMeCabC('まあそういうことで、余分な話はいいけん。', dic='example.dic')
```

1. **副詞:** 'まあ'
2. **連体詞:** 'そういう'
3. **名詞:** 'こと'
4. **助動詞:** 'で'
5. **記号:** '、'
6. **名詞:** '余分'
7. **助動詞:** 'な'
8. **名詞:** '話'
9. **助詞:** 'は'
10. **形容詞:** 'いい'
11. **助詞:** 'けん'
12. **記号:** '。'

どれくらいクリーニングすればよいか？

- テキストデータは一般に膨大
  - 人手で完璧にチェックすることは不可能
    - 自動で大量に処理できることのメリットが失われる
  - 事前に何が問題になるかはわかりづらい
    - 探索的なプロセス / データセットそれぞれの固有の問題
- 無難なアプローチ：分析しながら、漸進的に改善
  - ノイズになっている個所を見つけたら対応する
    - 固有名詞や方言 / 形態素解析の誤認識
  - 単語の頻度表をチェックする
    - 当然多くなるべき単語が多くなっているか
    - 不可解な言葉が多くなっていないか
  - 分析
    - いつでも元データから最新のデータを作れるようコードを保存する
    - コアとなる分析に関係する部分は入念にチェックする
    - 日ごろから様々なエラーの可能性を検討しておく

クリーニングのヒント

- 辞書に載っている表記に変える
  - ひらがな・カタカナを漢字にする
  - 伸ばし棒なのか母音なのか
- 方言に対応する
  - MeCabの辞書に追加する
    - 追加する際の情報は標準語の対応する単語を参考にする
  - 辞書で対応できそうになければ、一括で置換する
    - Googleスプレッドシートは正規表現での検索・置換に対応している

データハンドリング入門：「夢十夜」の分析

- 夏目漱石「夢十夜」
  - 1908年に朝日新聞紙上で連載
  - 夢を題材にした10話からなる小説
  - パブリックドメイン

夏目漱石
(1867-1916)

- データ
  - 青空文庫で公開されているデータ(新字新仮名)を用いた
  - ルビの削除等の処理はAozora()関数(石田, 2017) を用いた
    - 著者のホームページでも公開されている
    - URL
  - 今回はtsv形式に加工したものを使う

分析①：「夢十夜」

データセットの構造

話ID	段落ID	本文
section_id	paragraph_id	content
1	1	こんな夢を見た。
1	2	腕組をして枕元に坐っていると、仰向に寝た女が、静かな声でもう死にますと云う。女は長い髪を枕に敷いて、輪郭の柔らかな瓜実顔をその中に横たえている。真白な頬の底に温かい血の色がほどよく差して、唇の色は無論赤い。とうてい死にそうには見えない。しかし女は静かな声で、もう死にますと判然云った。自分も確にこれは死ぬなと思った。そこで、そうかね、もう死ぬのかね、と上から覗き込むようにして聞いて見た。死にますとも、と云いながら、女はぱっちりと眼を開けた。大きな潤のある眼で、長い睫に包まれた中は、ただ一面に真黒であった。その真黒な眸の奥に、自分の姿が鮮に浮かんでいる。
1	3	自分は透き徹るほど深く見えるこの黒眼の色沢を眺めて、これでも死ぬのかと思った。それで、ねんごろに枕の傍へ口を付けて、死ぬんじゃなかろうね、大丈夫だろうね、とまた聞き返した。すると女は黒い眼を眠そうに※たまま、やっぱり静かな声で、でも、死ぬんですもの、仕方がないわと云った。
1	4	じゃ、私の顔が見えるかいと一心に聞くと、見えるかいって、そら、そこに、写ってるじゃありませんかと、にこりと笑って見せた。自分は黙って、顔を枕から離した。腕組をしながら、どうしても死ぬのかなと思った。
1	5	しばらくして、女がまたこう云った。
1	6	「死んだら、埋めて下さい。大きな真珠貝で穴を掘って。そうして天から落ちて来る星の破片を墓標に置いて下さい。そうして墓の傍に待っていて下さい。また逢いに来ますから」

分析①：「夢十夜」

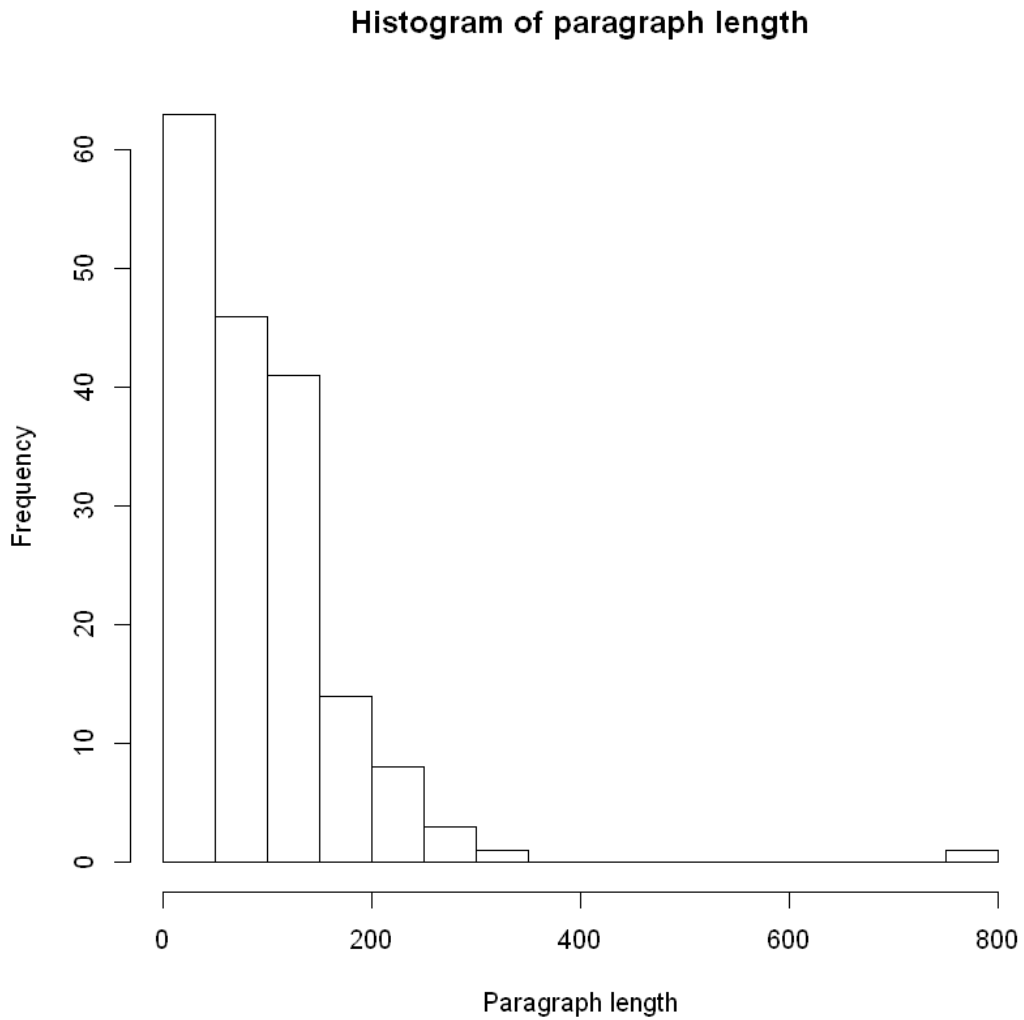
データハンドリング

- テキスト：様々な分析単位がありえる
  - 文字・単語・文・段落・章・文書全体・文書の集合…
    →様々な分析単位を自由に行き来できる形式が望ましい

  - インタビューデータ：話者の発話ごと
    - 特定の話者の発言のみを抜き出す
    - 特定の話題を含む発言のみ抜き出す

分析①：「夢十夜」

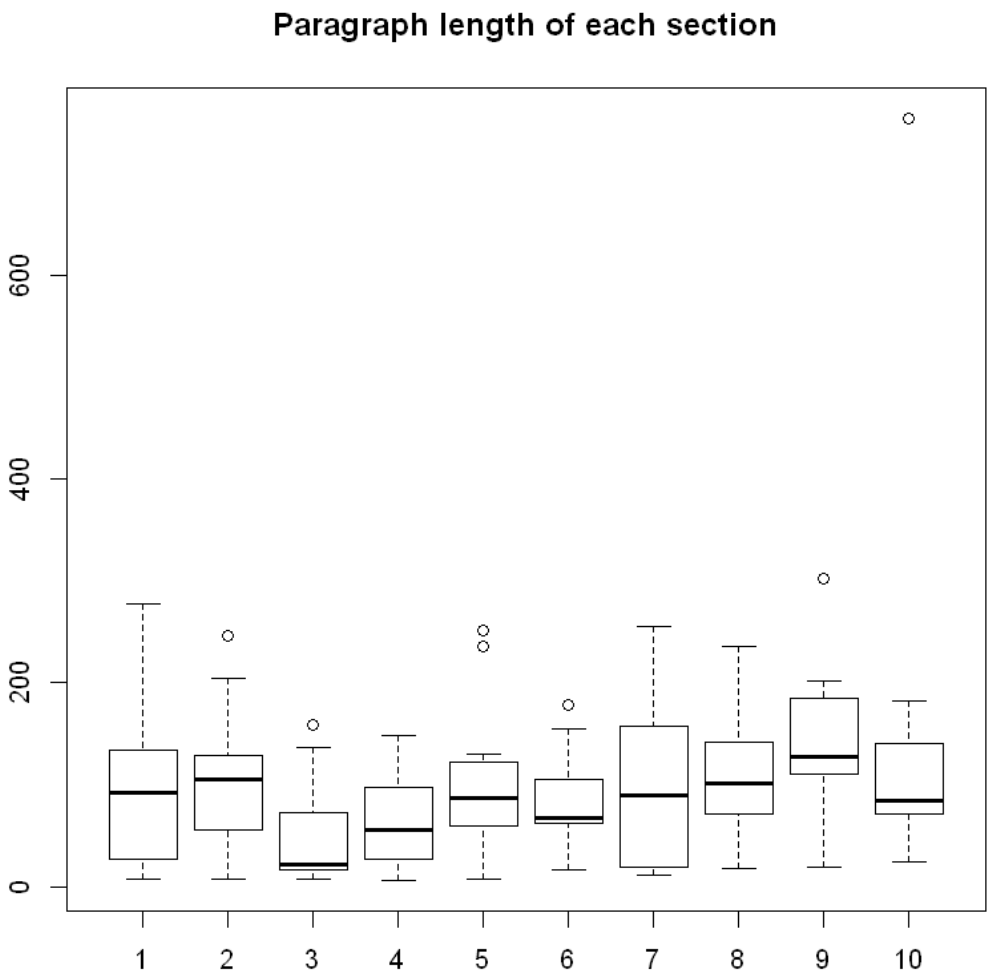
段落の長さの分布



- ほとんどが400字以下だが、非常に長い段落がごく少数ある

分析①：「夢十夜」

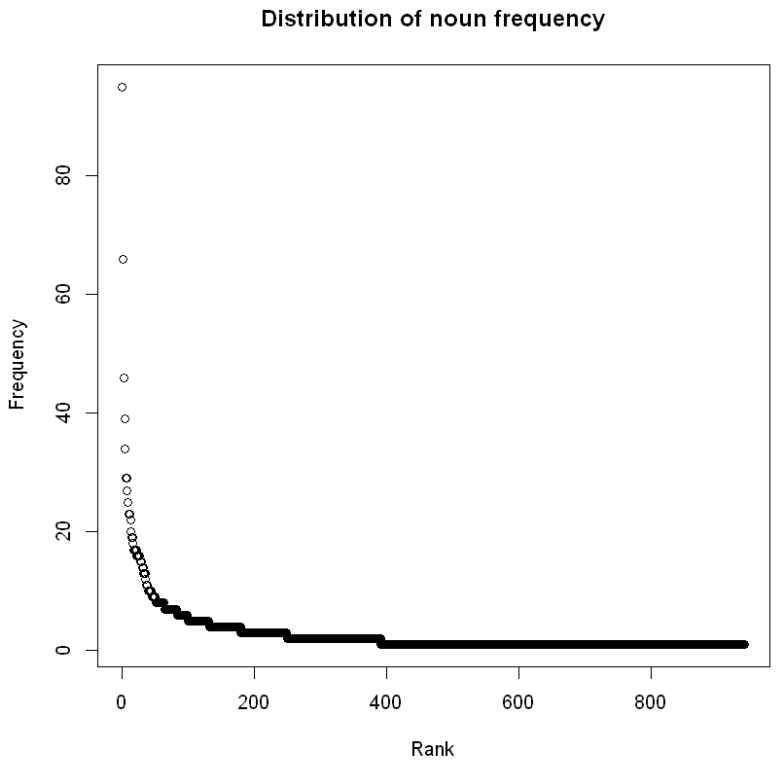
段落の長さ

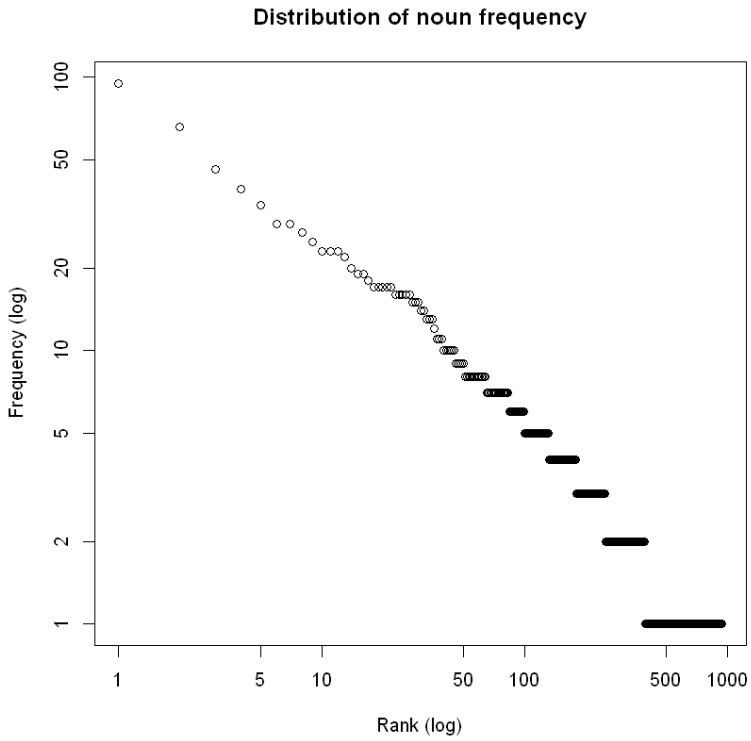


- 第十夜に非常に長い段落があるのがわかる
- 他の話でも、1,2段落ほど外れ値的に長い段落がある

分析①：「夢十夜」

名詞の分布





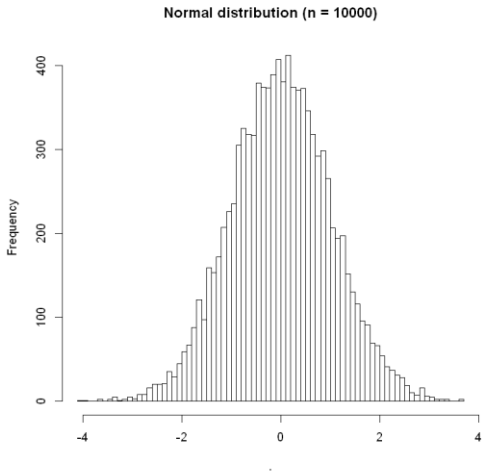
- ごく少数の単語が多く出現するが、大多数の単語はほとんど出現しない
- 両対数グラフにとると、直線のようにみえる
  →単語の分布一般の特徴（Zipfの法則）

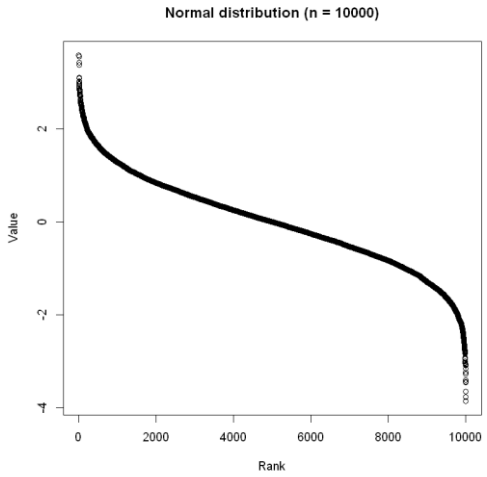
★頻度データの特徴

- 頻度：0以上の値をとる
  - 正規分布とは異なる確率分布に従う
    - e.g. Poisson分布
- べき分布 power distribution
  - べき則　$f(x) = ax^k$　の形に従う分布
  - 突出した頻度をもつ少数のアイテムと、頻度が小さい膨大な数のアイテムからなる（long-tailな分布）
  - 自然現象や社会現象の一部で観察される
    - Paretoの法則(Pareto, 1896)
      - 所得分布は上位20%が全体の80%を占める
    - Zipfの法則(Zipf, 1949)
      - 文書内の単語の頻度は順位に反比例する
    - Gutenberg-Richter則(Gutenberg & Richter, 1941)
      - 地震の頻度は規模に反比例する

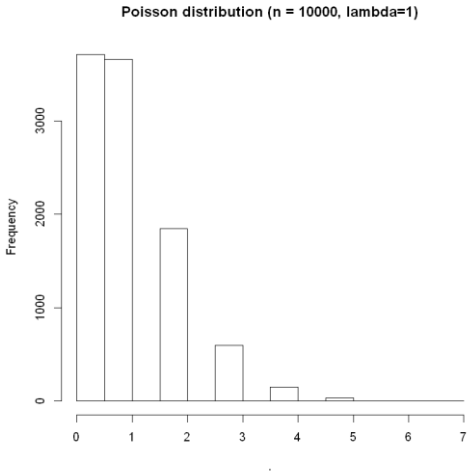
★べき分布：他の分布との比較

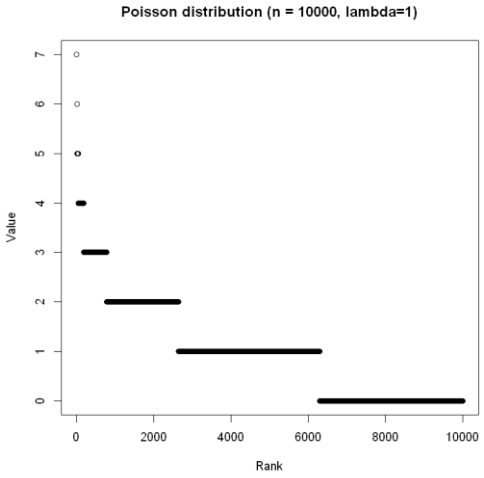
正規分布

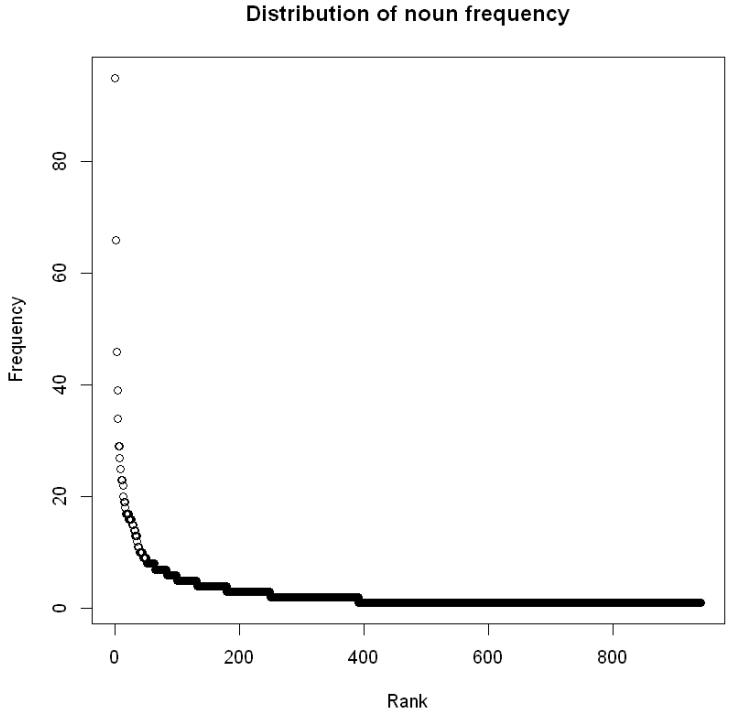




Poisson分布

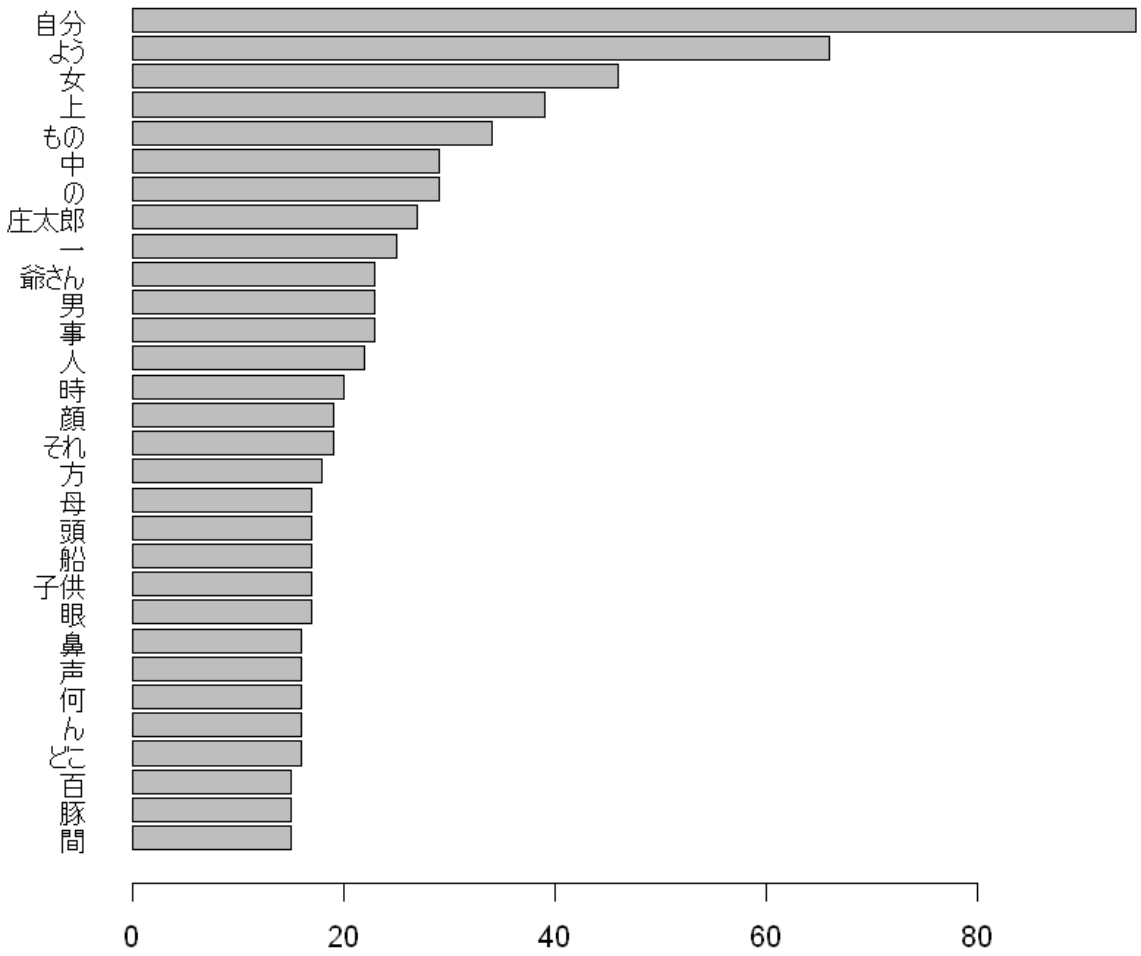






分析①：「夢十夜」

頻度の多い単語



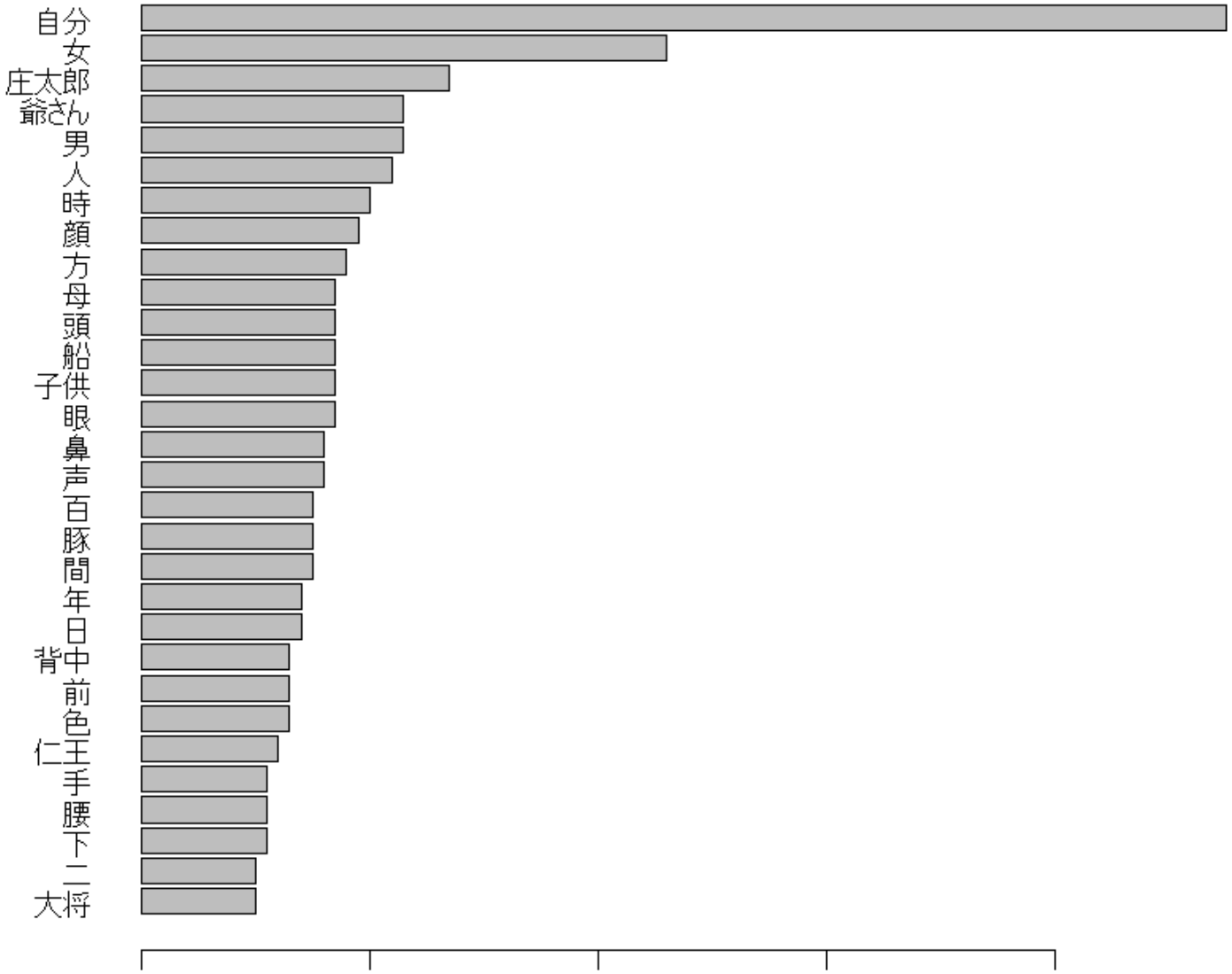
- 一部の単語は内容を理解するのに寄与しない
- 例：よう・上・もの・中・の・一・それ・事・それ・etc.
  →ストップワードとして分析から除外する

分析①：「夢十夜」ストップワードの除外

Before

自分
よう
女
上
も
の
中
の
庄太郎
一
爺さん
男
事
人
時
顔
それ
方
母
頭
船
子
供
眼
鼻
声
何
んど
こ
百
豚
間

After



分析①：「夢十夜」

TF-IDF

- TF-IDF: Term Frequency – Inverse Document Frequency
  - 文書群について、単語がどれくらい特徴的かを表す指標
    →文書のキーワードを抜き出すために使える
    - TF:Term Frequency 単語頻度
      - それぞれの文書について、その単語が出てくる程度
    - IDF:Inverse Document Frequency 逆文書頻度
      - 全体の文書のうち、その単語を含む文書の程度（の逆数）
      - 複数の文書に出現する単語ほど特徴的でない
    - TF-IDF
      - TFとIDFの積
      - 少数の文書に頻出する単語ほど強く重みづける
  - 経験的な指標：理論的な基礎ははっきりしないが、**有用**なためテキストマイニングや検索エンジンなどで幅広く使われている
  - 共通する単語は低く重みづけられるので、同系統の文書を分析するときには注意が必要

分析①：「夢十夜」

TF-IDFの計算方法

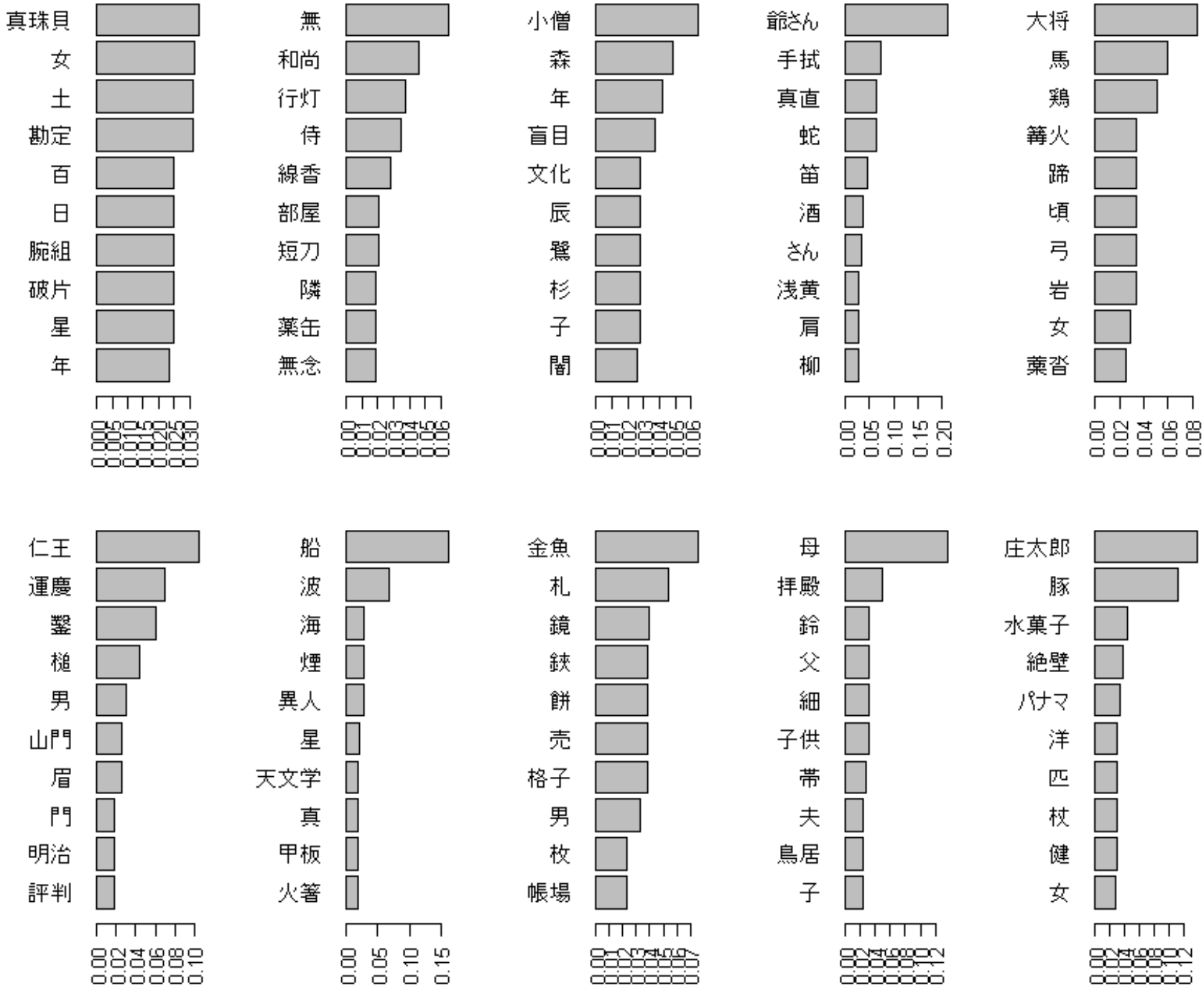
$$TF = \frac{\text{当該の単語の出現回数}}{\text{文書内の総単語数}}$$

$$IDF = \log \frac{\text{総文書数}}{\text{当該の単語を含む文書の数}}$$

$$= \log\left(1 / \frac{\text{当該の単語を含む文書の数}}{\text{総文書数}}\right)$$

$$= \log\left(\frac{1}{\text{文書頻度}}\right)$$

- すべての文書に当該の語が含まれる場合、IDFはゼロになる
- 「当該の単語を含む文書の数」がゼロになると計算できないので、実用上１を足して計算することがある


真珠貝
女
土
勘定
百
日
腕組
破片
星
年
0.00
0.05
0.10
0.15
0.20
0.25
0.30
無
和尚
行灯
侍
線香
部屋
短刀
隣
薬缶
無念
0.00
0.01
0.02
0.03
0.04
0.05
0.06
小僧
森
年
盲目
文化
辰
鷺
杉
子
闇
0.00
0.01
0.02
0.03
0.04
0.05
0.06
爺さん
手拭
真直
蛇
笛
酒
さん
浅黄
肩
柳
0.00
0.05
0.10
0.15
0.20
大将
馬
鶏
篝火
蹄
頃
弓
岩
女
薫沓
0.00
0.02
0.04
0.06
0.08
仁王
運慶
鑿
槌
男
山門
眉
門
明治
評判
0.00
0.02
0.04
0.06
0.08
0.10
船
波
海
煙
異人
星
天文学
真
甲板
火箸
0.00
0.05
0.10
0.15
金魚
札
鏡
鋏
餅
売
格子
男
枚
帳場
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
母
拝殿
鈴
父
細
子供
帯
夫
鳥居
子
0.00
0.02
0.04
0.06
0.08
0.10
0.12
庄太郎
豚
水菓子
絶壁
パナマ
洋
匹
杖
健
女
0.00
0.02
0.04
0.06
0.08
0.10
0.12

小まとめ

- データのクリーニング
  - ノイズを減らす、辞書を整備する
  - 分析しながら適宜データを洗練させていく
  - データ固有の知識（e.g.方言）が役に立つ
- 前処理/データハンドリング
  - テキスト：様々な単位で分析したい
    →柔軟に扱えるデータ構造にする
- 簡単な分析でも色々なことがわかる
  - 頻度の分析
  - TF-IDF

第２部
実習②テキスト分析編

テキスト分析：『こころ』の分析

- 夏目漱石『こころ』
  - 1914年に朝日新聞で連載
  - 三部構成
    - 「先生と私」「両親と私」「先生と遺書」
  - 新潮文庫版：発行718万部
    - 新潮文庫で最も売れている小説(日本経済新聞, 2016)
  - パブリックドメイン

- データ
  - 青空文庫で公開されているデータ(新字新仮名)を用いた

分析②：『こころ』

分析の方略

- 『こころ』：三部構成
  - 各部によって登場人物・主題が異なる
    →各部の特性の違いを可視化する

- 単語の分布の違いの分析
  - 主成分分析を用いて、各部の単語（名詞）の分布を可視化する
- 単語同士の共起関係を分析する
  - 全体の単語の共起関係を分析
  - 登場人物による共起する単語の違いを可視化する

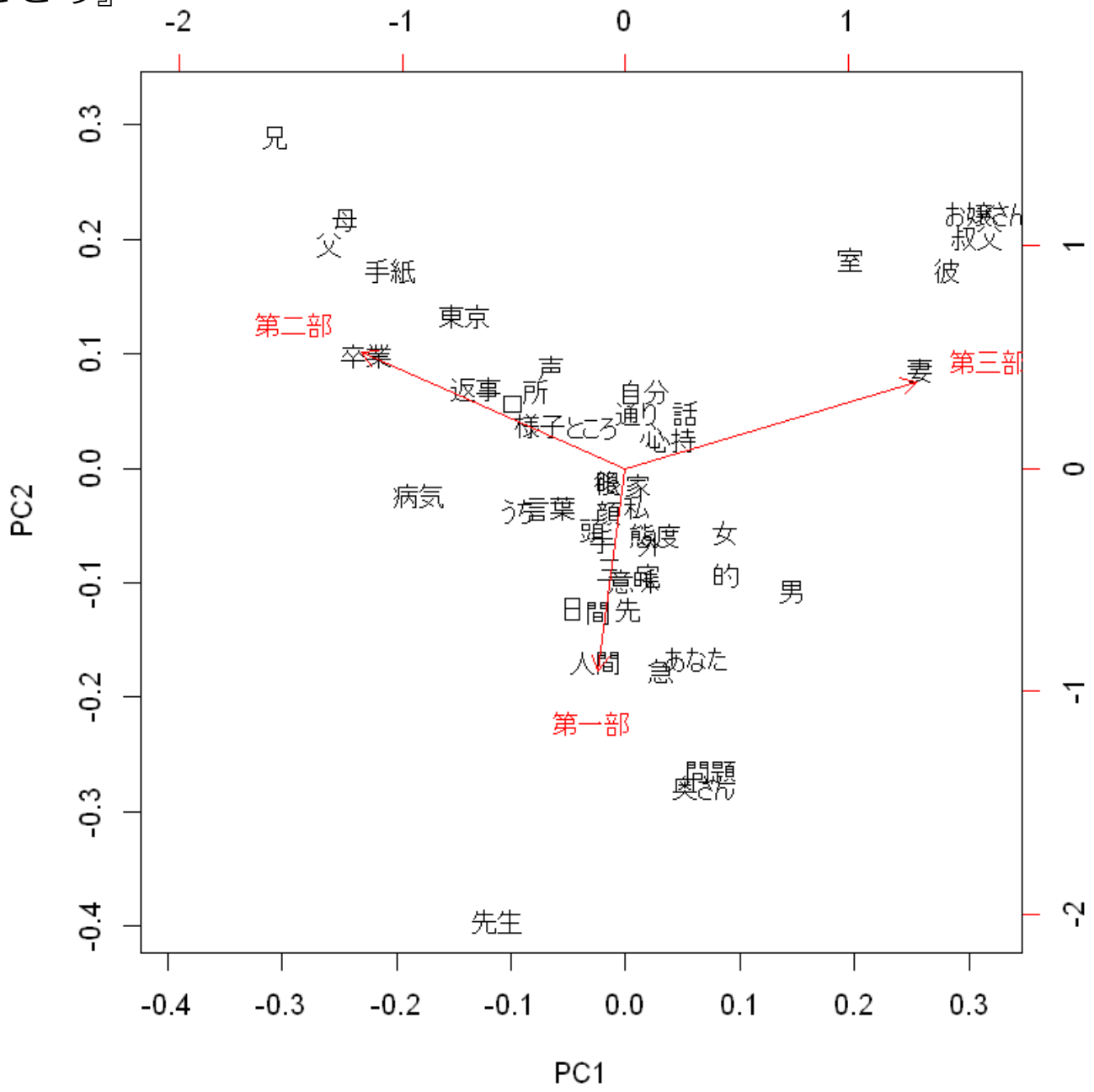
分析②：『こころ』

主成分分析

- データのばらつきをよく説明する「主成分」を大きいものから順に抜き出していく
  - 任意性の少ない手法
- データを少ない変数で要約できる
- 各軸は直交する（相関がない）ように選ばれる

- バイプロット biplot
  - 多次元データを**二次元**空間に射影したものが得られる
  - 合成スコアなので軸の解釈は特にする必要はない
  - 多次元のデータを図示するのに使える
    - 二次元空間への射影なので、軸が多数あるようなデータの場合はうまく図示できない

分析②：『こころ』



『こころ』各部と名詞同士の共起関係を用いた主成分分析

分析②：『こころ』

共起分析 co-occurrence analysis

- 共起をカウントする範囲
  - 文書
  - 章・節
  - 文
  - 窓関数
    - 前後n単位

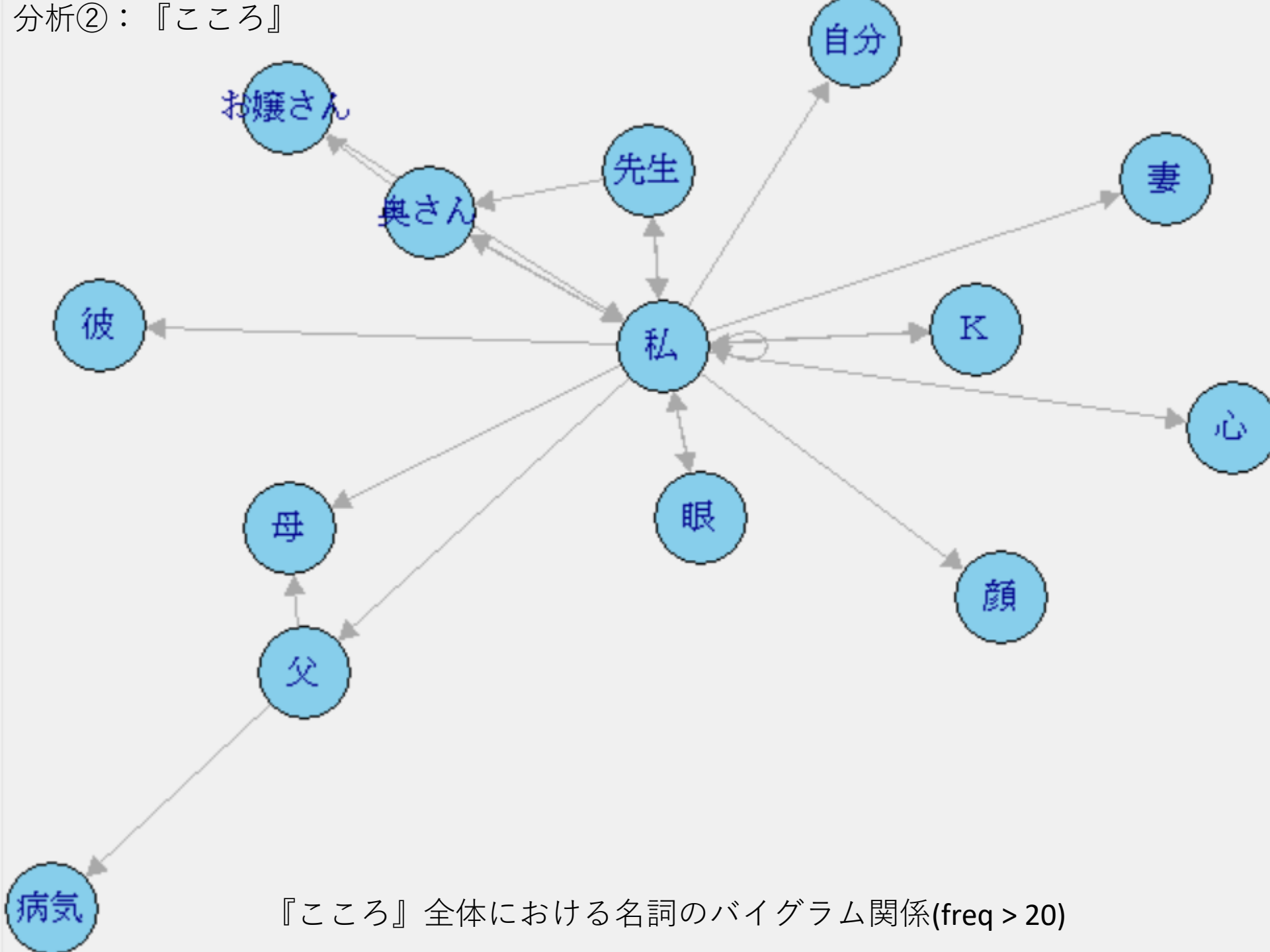
  ※範囲が広くなるほどデータサイズが膨大になる
- カウントの仕方
  - Bag of words: 出現のみを考慮する
  - N-gram: 語順も考慮に入れる

分析②：『こころ』

N-gramを用いた共起分析の例

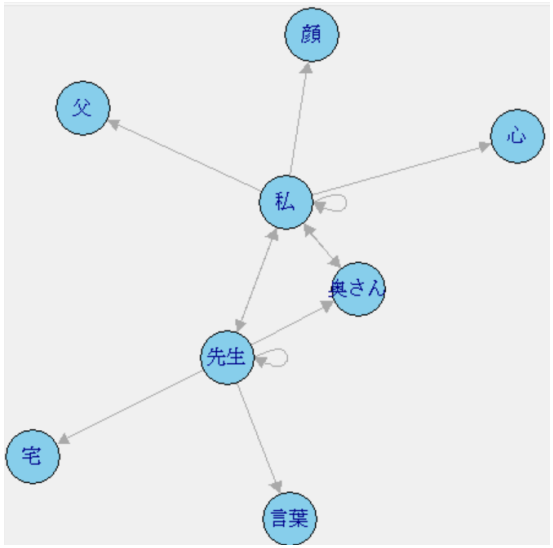
- 手続き
  - 対象とする単語群を決める（e.g. 品詞）
  - N-gramを抽出する
  - カウントする
  - 分析・可視化する
- 特徴
  - 順序・距離が保持される
    - 文章の空間的構造が比較的反映されている
      →距離的に近い組み合わせを抽出できる
  - 数が少なくすむ
  - 直近の共起関係しかわからない

分析②：『こころ』

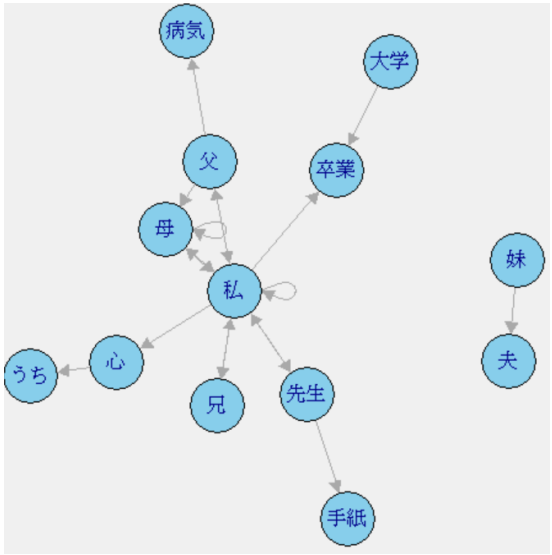


『こころ』全体における名詞のバイグラム関係(freq > 20)

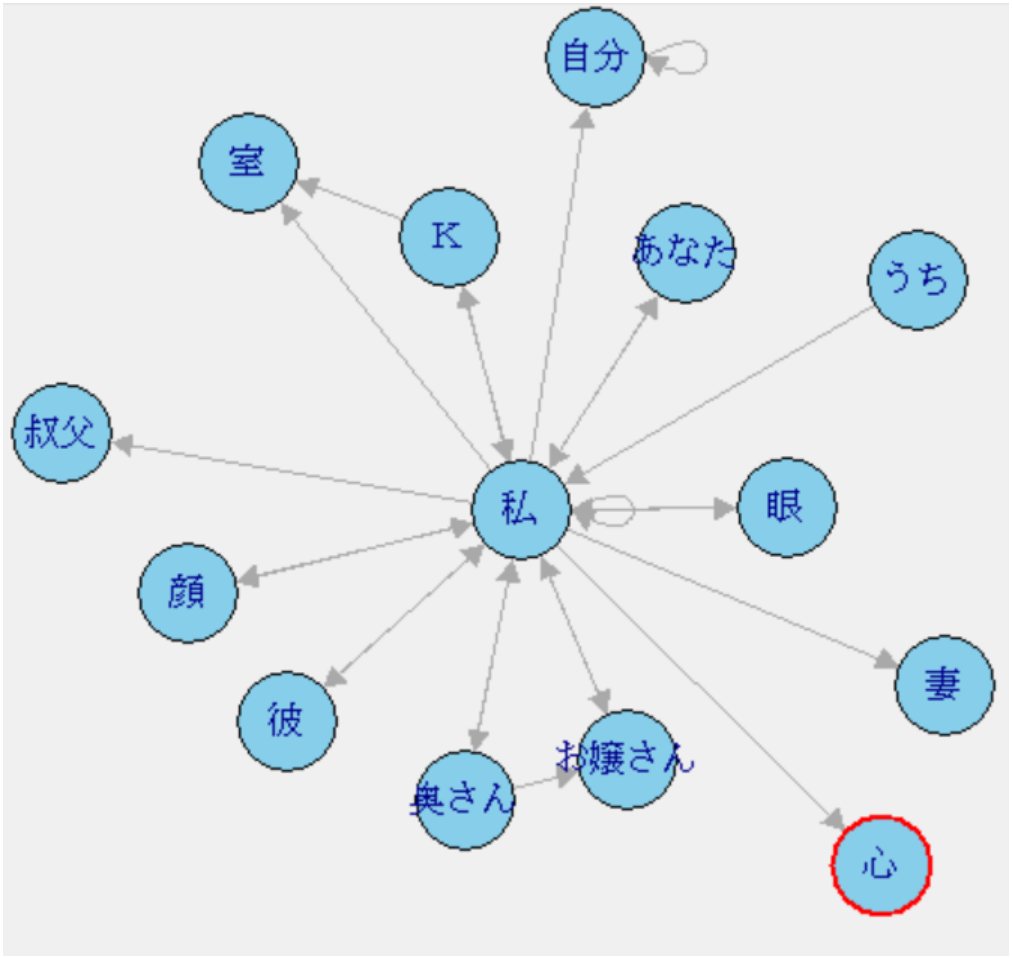
分析②：『こころ』



第一部(freq > 10)



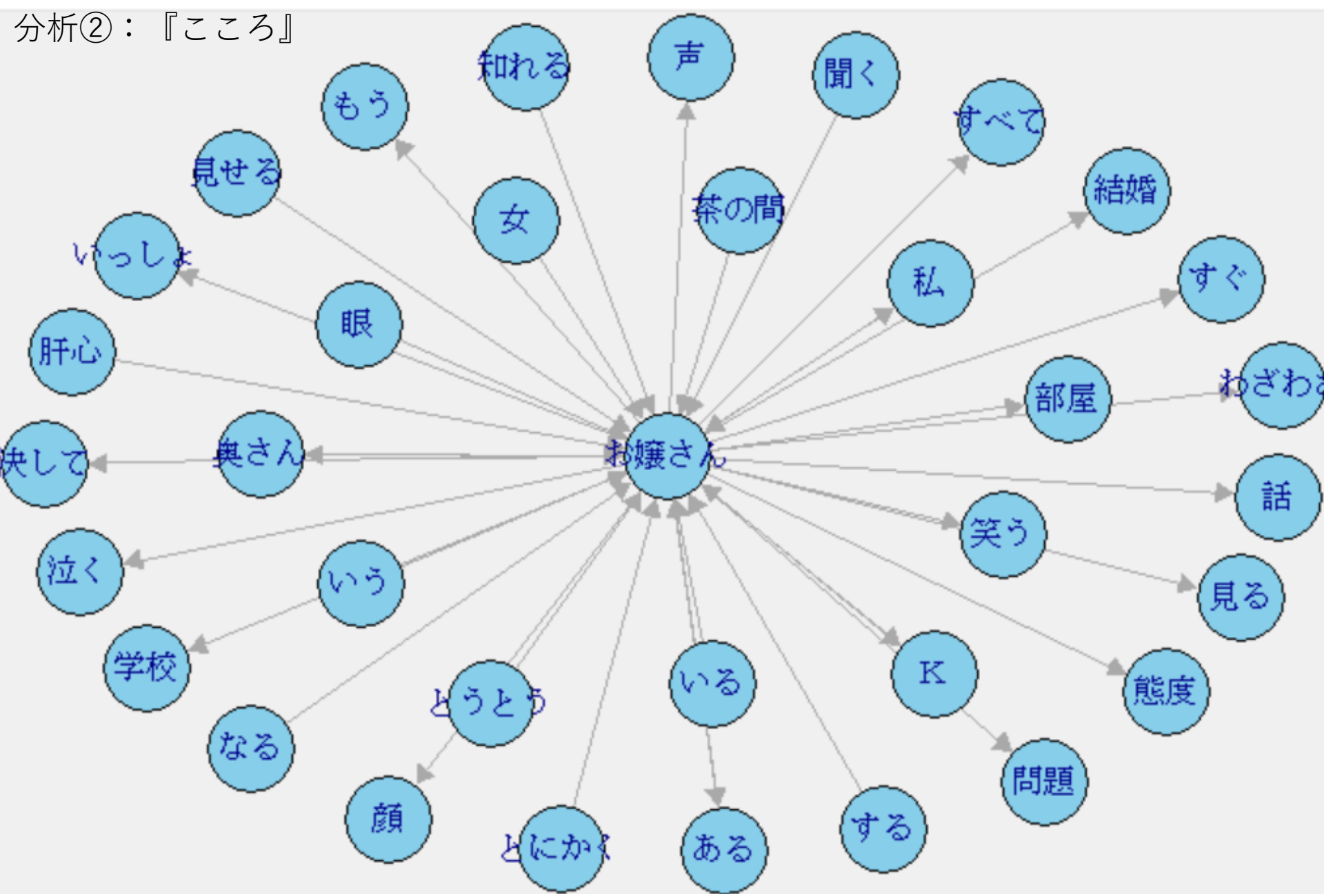
第二部(freq > 5)



第三部(freq > 10)

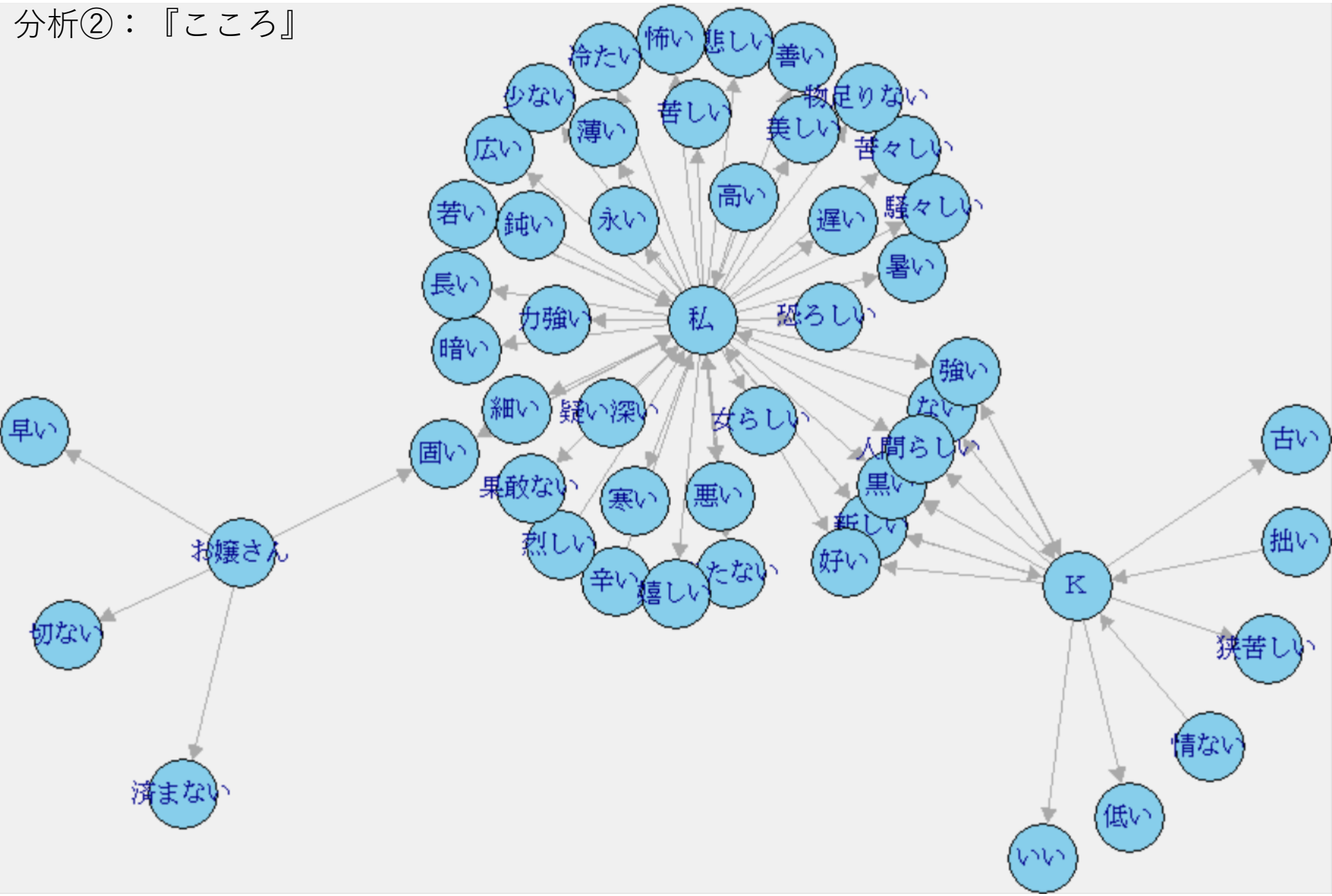
『こころ』各部における名詞のバイグラム関係

分析②：『こころ』



第三部における単語「お嬢さん」のバイグラム関係
（動詞・名詞・形容詞・副詞, freq > 1）

分析②：『こころ』



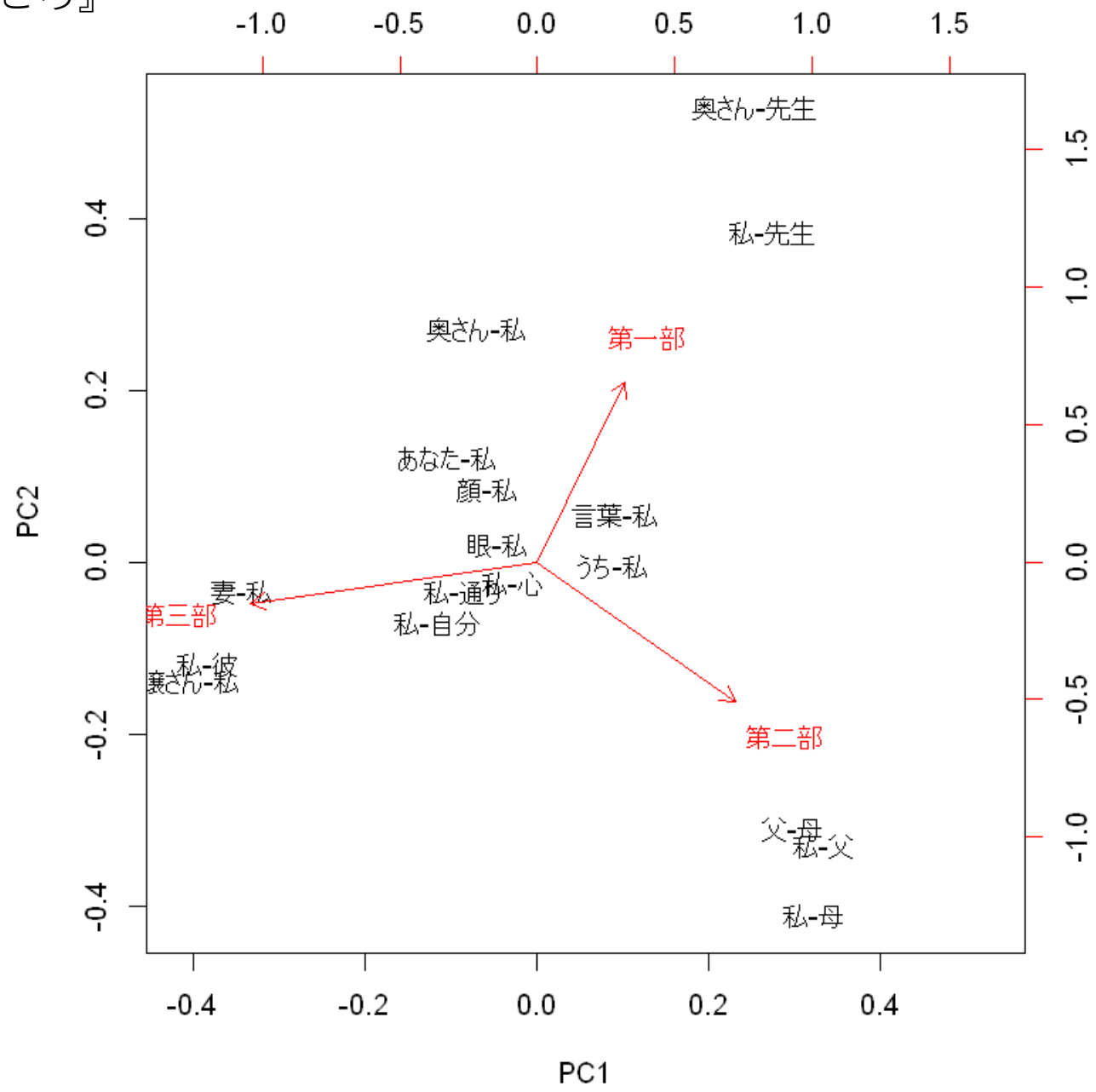
第三部における単語「私」「お嬢さん」「K」と形容詞とのバイグラム関係（freq > 0）

分析②：『こころ』

Bag of wordsを用いた共起分析

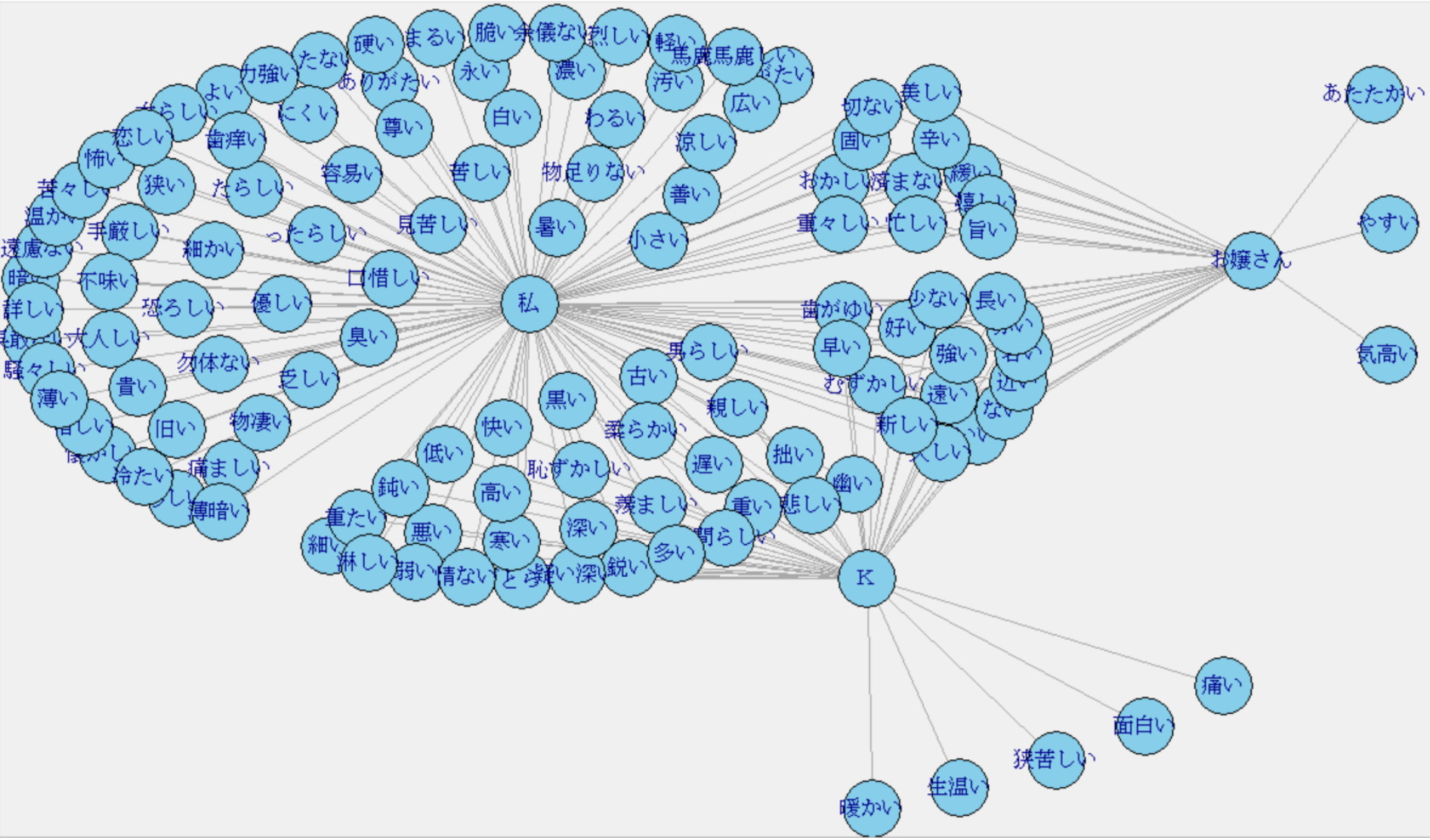
- 手続き
  - 対象とする単語群を決める(e.g. 品詞)
  - 文ごとに単語の共起を抽出する
    - Bag of wordsに基づいて、総組み合わせを抽出する
  - 集計する
  - 分析・可視化
- 特徴
  - 文内の距離・順序を保持しない
    - 空間的構造は反映されない
      →文内の距離に影響されない
  - 組み合わせが膨大になる
    - Bag of wordsの長さの二乗に比例する
  - 意味的なつながりの薄い組み合わせも拾ってしまう

分析②：『こころ』



『こころ』各部と名詞同士の共起関係を用いた主成分分析

分析②：『こころ』



第三部における単語「私」「お嬢さん」「K」の形容詞との共起関係（freq > 0）

分析②：『こころ』

分析の落とし穴

- 人物の名前
  - 「K」が固有名詞でなく記号として分類
  - 第一部・第二部に登場する「私」と第三部に登場する「私」は違う人物
    - 会話文の中も考えれば他の人物も「私」と発言している可能性がある
  - 第一部・第二部と、第三部の「奥さん」は違う人物

    ⇒領域/データ固有の知識が必要

- 否定表現
  - 単純に共起語を抜き出しただけではそれがどのように使われているかまではわからない
  - 「Aをした/Aである」ではなく「Aをしなかった/Aでない」かもしれない
  - 否定表現まで含めて分析したいなら構文解析などといった処理を使う必要がある

    ⇒データの性質/分析の限界を理解する

小まとめ

- データを可視化する様々な手法がある
  - 主成分分析：少ない次元に可視化する
  - 共起ネットワーク：結びつきを可視化する
  - 統計的検定に載せられないような「質的な」性質も表現できる可能性がある

⇒どうすればうまくデータの性質を可視化できるかを考える

- テキスト分析の罠
  - 否定表現
  - 単語の誤認識
  - アーティファクトではないか(cf. Back et al.)

⇒データの内容および背景知識、手法に関する知識を持つことで、罠にはまる可能性を減らせる

第 2 部
結果の報告と解釈

論文の形式

- 心理学論文はおおむねIMRAD形式(APA, 2008, 2013 ;日本心理学会, 2015)
  - **I**ntroduction
    - 緒言または序論。研究の意義と目的について説明する
  - **M**aterials and **M**ethods
    - 方法。研究の方法について説明する
  - **R**esults
    - 結果。方法により得られた結果について説明する
  - (And)
  - **D**iscussion
    - 議論と考察。得られた結果がどのような意味を持つかについて議論し、結論を述べる。

序論

- なぜその研究を行う必要があったのか
  - 先行研究のレビュー
  - 新たにその研究を行う意義
  - 研究を行うための理論的根拠(rationale)
- どのような目的でその研究を行うのか
  - どのようなデータが得られるのか
  - なぜその分析をするのか
- 何を明らかにするのか
  - 理論仮説
  - 作業仮説
  - 予測

目的があるということ

- 例①
  - 「本研究では夏目漱石の『こころ』を分析した。」

- 例②
  - 「本研究では、夏目漱石の『こころ』について、各部における表現の違いを調べるために、計量テキスト分析の手法を用いた。」
- 例③
  - 「本研究では、計量的な手法でテキストの内容を可視化できるか検討するために、夏目漱石の『こころ』を題材に分析を行った。」

方法

- どのようにその研究を行ったのか
  - 研究に用いた材料と方法論について述べる
  - 研究の妥当性の根幹にかかわる部分
- 何をどのように測定したか
  - 特に、理論的な概念と材料（データ）がどのように対応するのか→構成概念妥当性
    - どんなに立派な序論を書いてもここがダメだとダメ
    - 対応が取れていないと総じて無価値な研究になりがち
  - 他の人がそれを読んで研究結果を再現できるように

報告すべき内容

- 分析対象（データセット）
  - データセットの概要
  - データセットの入手方法
  - データセットの構造
- データ処理
  - クリーニング
  - 前処理
  - 除外データの有無
  - 分析に用いる変数・尺度
- 使用ツール
  - プログラム言語
  - 形態素解析エンジン

- その他特記事項
  - 研究倫理（個人情報の保護など）

方法の書き方の例

- データセット
  - 分析には『こころ』本文を用いた。
  - テキストは青空文庫(https://www.aozora.gr.jp/)にて公開されているものを用いた。データの取得には石田(2017)のAozora関数を用い、同時に元のHTMLにあるルビ等のタグは削除した。
  - テキストは一文を一行のレコードとし、部・節の番号、並びに登場した順に段落・文の番号を１から連番のIDとして付した。このIDは、第二部第一節第三段落第四文であれば2,1,3,4となる。このように一文を単位にしたデータをタブ区切りのファイルとし、分析に用いた。

方法の書き方の例

- クリーニング・前処理
  - クリーニング作業として、誤字や脱字が含まれていないかどうかをチェックした。具体的には形態素解析の結果分割した内容を集計した頻度表を作成し、登場人物などの固有名詞が認識されているか、また、誤認識の結果が集計されていないかを確認した。
  - 確認作業の結果、うまく認識されていなかった単語については形態素解析ソフトの辞書に追加した。
  - 漢字の送り仮名などの表記ゆれについては同じ単語として集計されるよう統一した
- 除外データの有無
  - 今回の分析にはすべての文を分析の対象とした
  - 頻度の高い語のうち、「もの」「こと」などのように、内容の理解に寄与しない単語群をストップワードとして集計から除外した

方法の書き方の例

- 分析に用いる尺度・変数
  - 分析に用いる変数は以下のように集計した。
  - 頻度：出現した単語をそのまま集計した。
  - バイグラム：文から抽出した対象の品詞の連続をバイグラムとしてカウントした。すなわち、「吾輩は猫である」から、名詞のみを抽出した場合、「吾輩-猫」がバイグラムとなる
  - 共起：文から単語をbag of wordsとして抽出し、そのすべての単語の組み合わせを共起としてカウントした。ただし、同じ単語が一文に２回以上出現しても１単語として組み合わせを求めた。

方法の書き方の例

- 使用ツール
  - 分析にはR(4.0.2)を用いた。
  - 日本語の形態素解析には工藤ら(2004)のMeCab(ver. 0.996)を用いた。Rからの操作にはRMeCabパッケージ(石田, 2017)を用いた。

結果

- 何がデータから得られたか
  - データの特徴に関する記述
    - 基本的な記述統計→基本的な事実の確認
    - 実験であれば操作の妥当性に関するチェックなど
  - 主要な分析結果とその説明

研究の結果を，内容の重要度に従って事実に即して忠実に述べる。自分の予期に反した事実も省略しない。

心理学における研究では統計的仮説検定が分析にしばしば用いられるが，仮説検定はデータ分析の一側面に限られる。必要に応じて仮説検定に限らず適切な分析手法を用いるのが望ましい。特に，仮説検定の適用にあたっては，前提とするデータの性質（データの分布の正規性や，標本相互の独立性など）が成立していることを確認する。

分析結果の記述においては，研究結果の重要性を評価できるよう効果量とその信頼区間も示す。元来の測定単位・尺度によって表された効果量は理解が容易であるが，必要に応じて尺度に依存しない標準化された効果量の指標（Cohen の d や標準化回帰係数等）を示す。

データの欠測は分析の結果に大きな影響をしばしば与える。欠測を伴うデータを分析する場合には，欠測の頻度や件数を示すとともに，欠測の発生について経験的あるいは理論的な説明を記述する。分析において採用した欠測モデルの性質（MCAR，MAR，NMAR の区分）や，欠測に対応するために採用した方法（多重埋め合わせなど）について記述することが望ましい。

日本心理学会(2015) 『執筆・投稿の手びき』

結果の報告と解釈

- 報告：事実＝得られたデータについて述べる
  - ○「データは～である」→事実
  - ×「人々は～である」→推論

- 解釈：事実＝得られたデータについての説明
  - ○「実験の結果～であった。そのため、参加者は～であったと考えられる」
  - ×「実験の結果～であった。そのため、人は～であると考えられる」

- 一般化可能性については議論のセクションで述べる

議論と考察

- 何が明らかになったのか
  - 結果のまとめ
  - そこから何が言えるのか
- どのような意味を持つか
  - 結果が仮に正しく、一般化したとすれば、どのような意味があるか
  - 先行研究との比較
- どこまで一般化できるか
  - 研究の適用範囲・限界
  - 結果の妥当性・信頼性
- 結び
  - 今後の展望など

小まとめ

- 序論
  - なぜその研究や分析をしたのかをわかるように書く
  - 特に目的は明瞭に書く
- 方法
  - 何をどうやって扱ったのかを明瞭に書く
  - 読んだ人が同じ手続きを取れるように書く
- 結果
  - データから事実としていえることを書く
  - そのことをわかりやすく言い換える
  - データは概念そのものでないことに注意する
- 考察
  - 結果から導き出される結論を書く
  - 序論で言及したことに対して結果がどういう意味を持つか議論する

全体のまとめ

- テキストマイニング/計量テキスト分析
  - 手近なものが分析対象になる
    - インタビュー
    - 文学作品
    - スピーチ・演説
    - SNS

  ⇒研究の幅が広がる
- 研究デザインは大事
  - 測定に常に注意を払うこと
  - テキスト分析固有の罠もある
  - 手法の可能性と限界を理解しつつ活用するのがよい

References

- American Psychological Association. (2013). *Publication Manual of the American Psychological Association, Sixth Edition*. American Psychological Association.
- APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839.Reporting
- Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media.
- Gutenberg, B., & Richter, C. F. (1941). Seismicity of the Earth.
- Pareto, V. (1982). First course in applied political economy, academic year 1893--1894. *Premier Cours d'économie Appliquée, Année Académique 1893--1894*.
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10
- Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press.
- 奥村学. (2010). 自然言語処理の基礎. コロナ社.
- 高村大也. (2010). 言語処理のための機械学習入門. コロナ社.
- 石田基広, & 金明哲(編). 2012コーパスとテキストマイニング. 共立出版.
- 石田基広. (2017). *Rによるテキストマイニング入門 第2版*. 森北出版.
- 工藤拓, 山本薫, & 松本裕治. (2004). Conditional Random Fieldsを用いた日本語形態素解析. *情報処理学会研究報告自然言語処理（NL）*, *2004*(47), 89–96. Retrieved from https://ci.nii.ac.jp/naid/110002911717/
- 日本経済新聞. (2016). 漱石没後100年、人気衰えず 書店で文庫フェア :日本経済新聞. Retrieved July 9, 2019, from https://www.nikkei.com/article/DGXLASDG08H0C_U6A211C1CR0000/