# 看護研究のためのテキストマイニング（質的/量的研究現場での利用事例）

2015.11.20

東京医科歯科大学

大河原知嘉子

# テキストマイニングとは

- 質的データである文字テキストを
  量的分析手法である統計や多変量解析などに
  よって分析する手法

| 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分析と量的分析との両方の特徴を持つ。
- ✓ Mix Method(混合研究法)としても有用。

# 分析対象

1. 情報的文字資料
   - Ciniiなどの文献データベース、webサイト
   - ガイドブック、ハンドブック
   - 情報中心の記事
   - 新聞記事

2. 物語(ナラティブ)的文字資料
   - 闘病記、感想文、レポート、小説、書籍
   - ブログ、ツイッター
   - 体験談、インタビューの逐語録

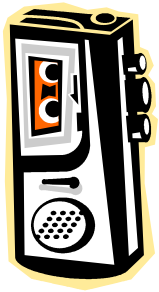
# テキストマイニング

**自由記載の**
**アンケート**

ご自由に
お書きください

---
---
---
---

**面接での逐語録**



**今までは**

**人の手でテキストを集計**
**人がテキストを読み、**
**内容を理解し、情報を得る**

利点：人が理解した意味内容を
　　分析に反映できる
欠点：大量のデータの分析困難
　　分析が主観的
　　情報処理に人手と時間がかかる

# テキストマイニング

**自由記載の**
**アンケート**

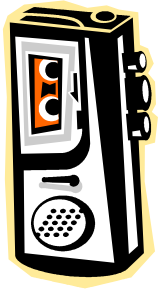
ご自由に
お書きください

---

---

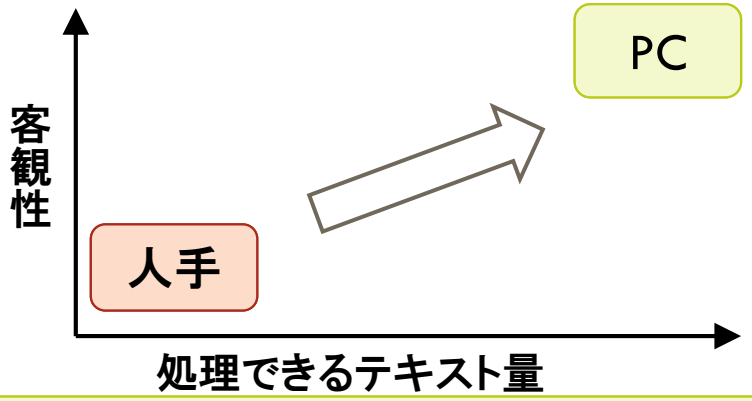
---

**面接での逐語録**



テキスト
マイニング

**パソコンを利用して、**
**テキストを数値データと**
**同じように処理できる**




利点：分析が客観的、再現可能
　　大量のデータの分析が可能
　　時間短縮が出来る
欠点：意味内容にまで踏み込んだ
　　分析が出来ない
　　ことばの落とし穴

# テキストマイニングの効果

- **大量のテキストデータを分析できる**
- **テキストデータの特徴や傾向を量的に示せる**
  →文章基本情報、単語頻度分析など
- **テキストデータの特徴を視覚的に示せる**
  →言葉ネットワーク分析、対応バブル分析など
- **文章と属性との関係を定量的に説明できる**
- **再現性があり誰でも同じ基準で分析できる**
- **分析時間を大幅に削減できる**

人が読むだけでは得られない情報を
獲得できる

# テキストマイニングの限界

- 文章の意味内容や行間を加味した分析が困難
- カテゴリーを作成してもことばの繋がりや使われ方、頻度などにより文章が分類される
  - ✓ ことば自体の意味で文章を分類するため、文章全体の意味内容での分類は困難
- 日本語の落とし穴
  1. 英語と異なり、ことばの切れ目が不明確
  2. 同じことばでも意味がさまざま→同音異義語の存在
  3. 会話分析ではことばの置換がおきるため関係性の分析が出来ない・・・意味的分類、会話の落とし穴

**辞書作りが分析の精度を左右する**

# ①不明瞭なことばの切れ目

- 英語と異なりことばの切れ目が不明確

E)It is my special day in my life.

E)It/is/my/special/day/in/my/life.

J)優秀な日本人の看護師は美しい

J)優秀な/日本人の看護師は/美しい

J)優秀な日本人の/看護師は/美しい

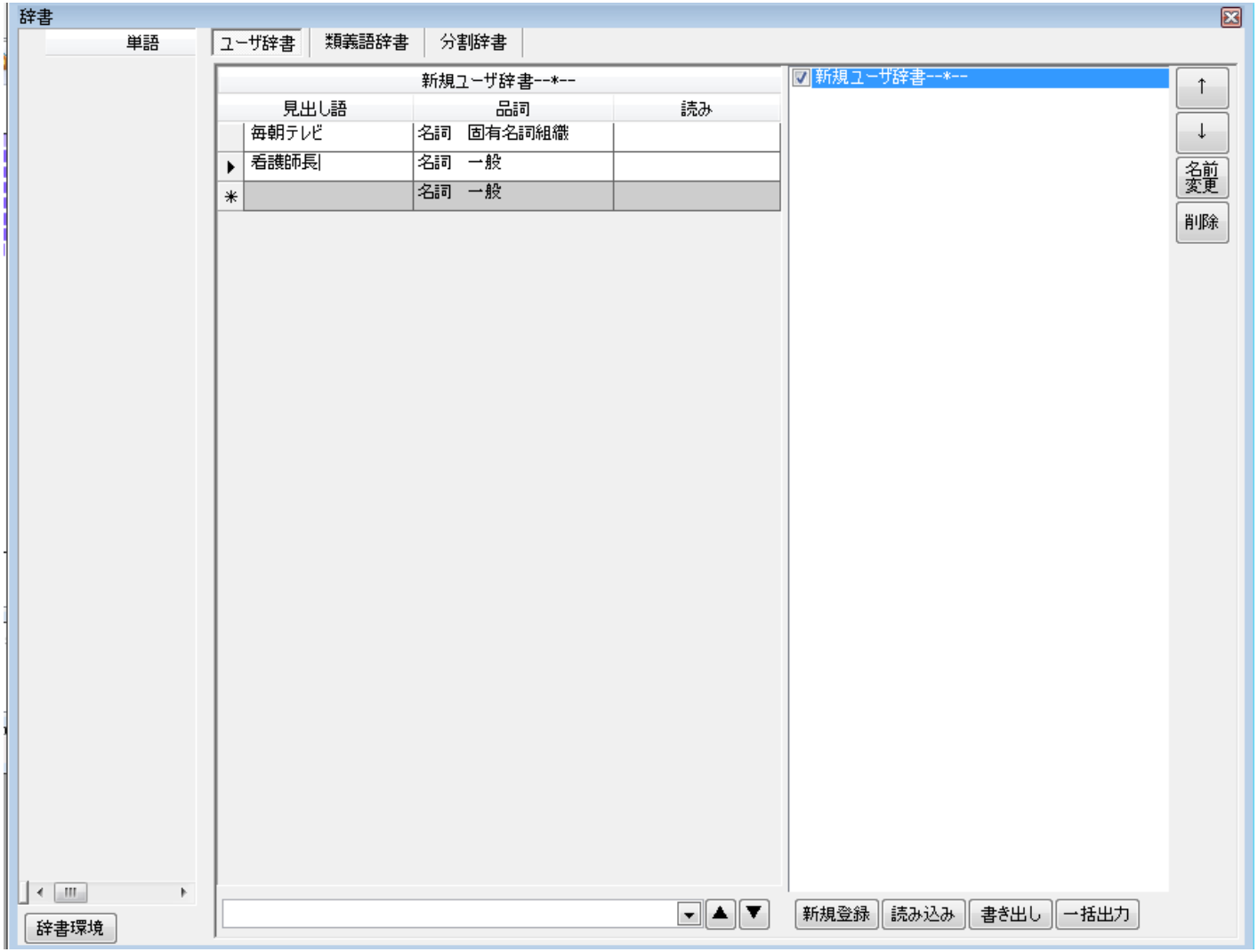
優秀なのは誰？
日本人の看護師？
日本人？

**文脈の中での理解が必要**

# どこで区切りますか？

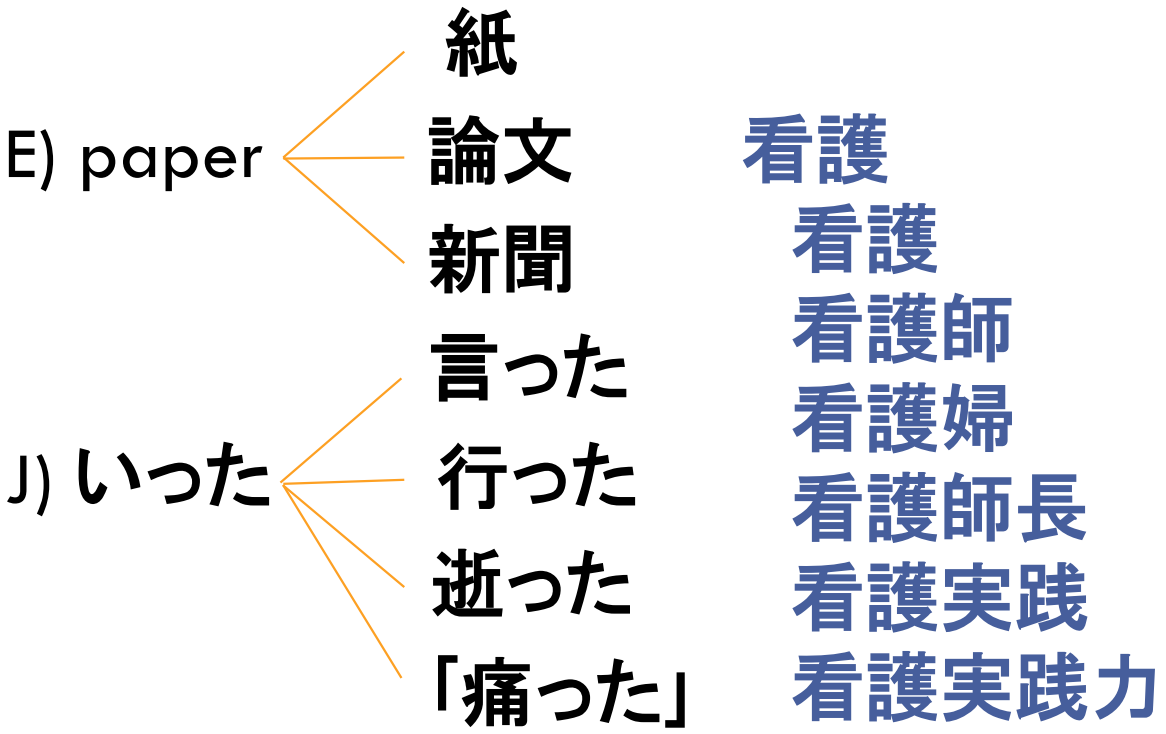
- わかちがき＝文章を意味のあるまとまりに区切ること
  ex）私は毎朝テレビで天気予報を確認する。
  - 私は毎朝テレビで天気予報を確認する。・・・形態素解析
  - 私は毎朝テレビで天気予報を確認する。・・・分かち書き

# ユーザ辞書登録


辞書
単語
ユーザ辞書
類義語辞書
分割辞書
新規ユーザ辞書--*--
見出し語
品詞
読み
毎朝テレビ
名詞　固有名詞組織
看護師長
名詞　一般
名詞　一般
新規ユーザ辞書--*--
↑
↓
名前変更
削除
辞書環境
新規登録
読み込み
書き出し
一括出力

# ②同音異義語

- 同じことばでも意味や構成がさまざま



# 辞書登録

- 類義語辞書


- ユーザ辞書

# ③-1 意味的分類

- 「看護師の対応がよい」
- 「対応した人の態度が良かった」
- 「気持ちの良い対応だった」

看護師について聞いている調査であればどれも
看護師の対応の仕方が良かったと言っている

ある程度の分類は自動的にできるものの簡単には
いかず、分析過程で文章をどう扱うかがポイントとなる

# ③-2 会話の落とし穴

- 会話の録音データから逐語録を作成し、それをデータとして分析する場合におこる落とし穴
  - 会話には句読点がないため、“、”や“。”をつける位置は研究者
  - 聞いた会話は、漢字になっていないため、漢字変換をするのは研究者
  - 「あー」、「えーっと」等、意味をなさないつなぎ言葉の扱い
  - 「ハハハ」「あはは」「ウフフ」等笑い声の扱い

# 録音データ

- 録音データ
  - えーまあでもわたしにそのかんがえなさいっていううんきかいとじかんをちゃんとあたえてくれてこたえがでたときもちゃんとそれをきいてくれるひとだったみたいなあはは

- テープ起こし・・・「素起こし」「ケバ取り」「整文」
  - でも私に考えなさいっていう機会と時間をちゃんと与えてくれて、答えが出た時もちゃんとそれを聞いてくれる、その聞いてくれる人だったみたいな。

# ③-2 会話の落とし穴：ことばの置換

- 会話分析ではことばの置換がおきるため
  関係性の分析が出来ない

  例：もう嫌になっちゃって…私

  ↓

  私はもう嫌になった

  私はもう嫌になっちゃって〇〇

# 分析

# 文章の基本情報

| 項目 | 値 |
| --- | --- |
| 総行数 | 300 |
| 平均行長(文字数) | 14.2 |
| 総文数 | 80 |
| 平均文長(文字数) | 53.3 |
| 延べ単語数 | 1,634 |
| 単語種別数 | 814 |

- 分析対象となっている文章の特徴を量的にとらえるために最初に行う分析
- その文章の行数や行長、延べ単語数、単語種別数などから、文章の長さや、使用していることばの豊富さなどを明らかにする

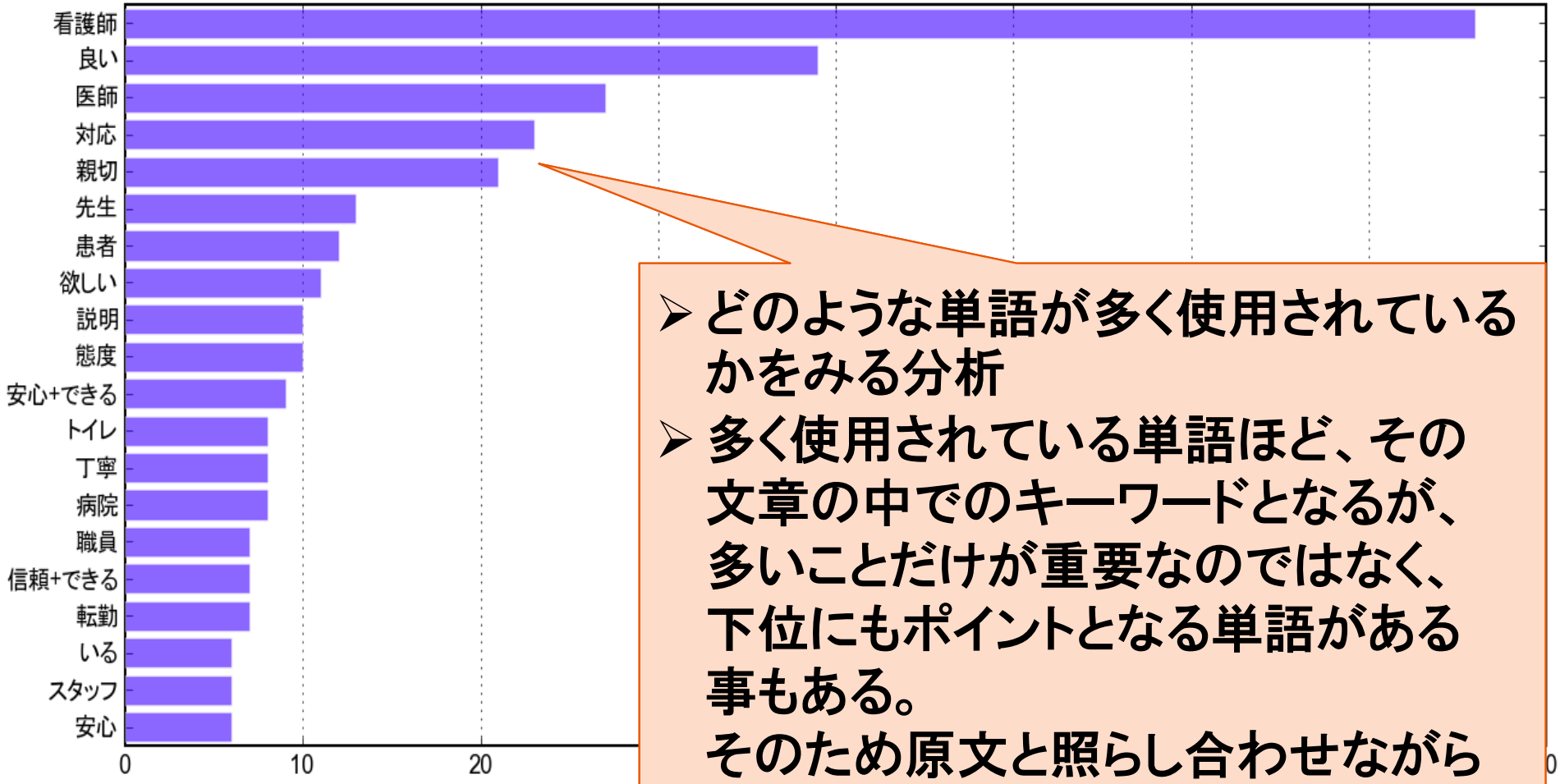
# 品詞の出現頻度

| 品詞 | 出現回数 |
|---|---|
| 名詞 | 4,406 |
| 動詞 | 3,318 |
| 形容詞 | 790 |
| 副詞 | 298 |
| 連体詞 | 245 |
| 接続詞 | 45 |

- その文章内で、どのような品詞の単語が多く使用されていたかを表す
- この文章だと名詞、動詞が多く使用されているが、それ以外の品詞はあまり使用されていない

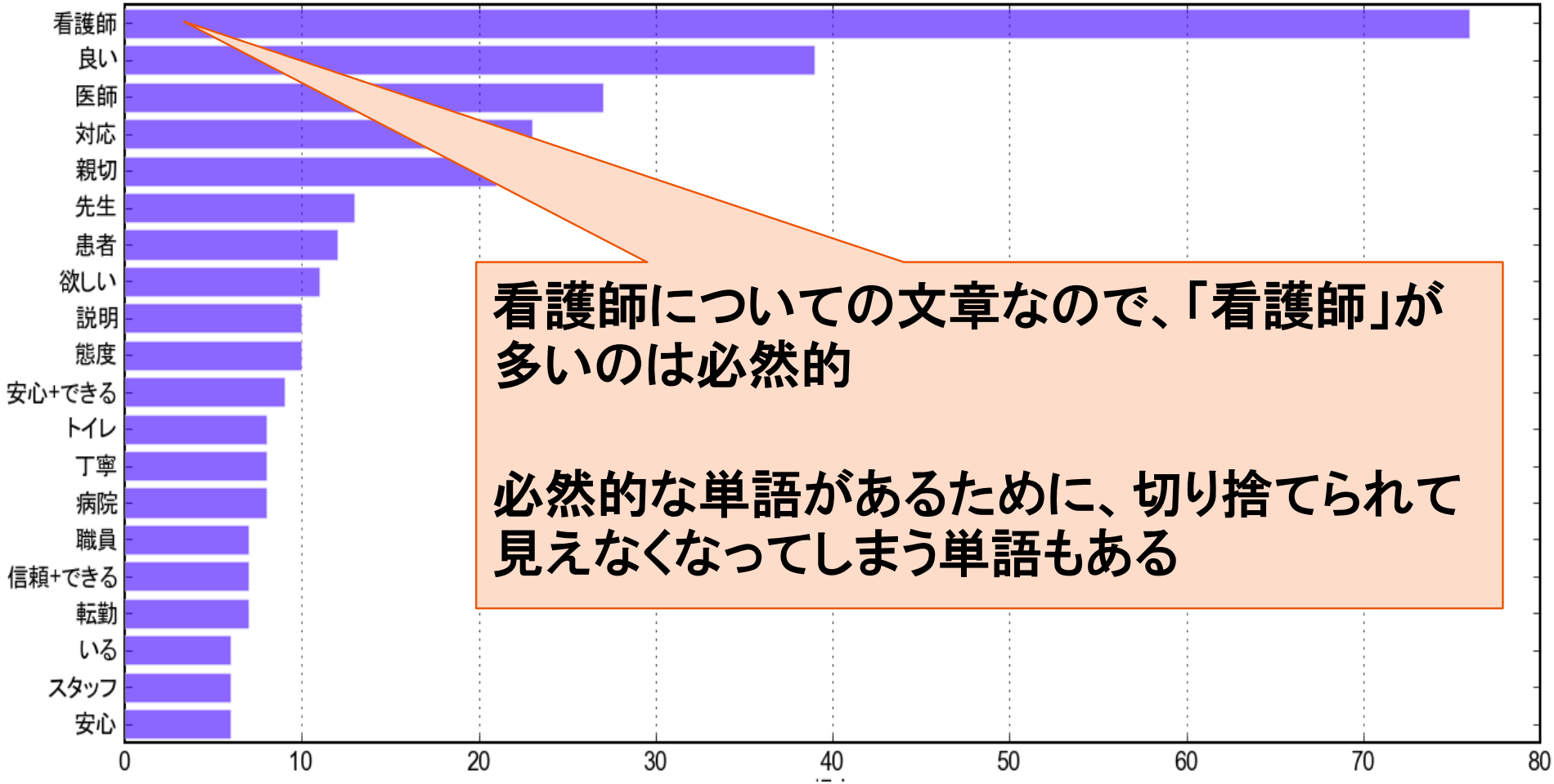
分析対象を決定するのに重要となる分析

# 単語頻度解析

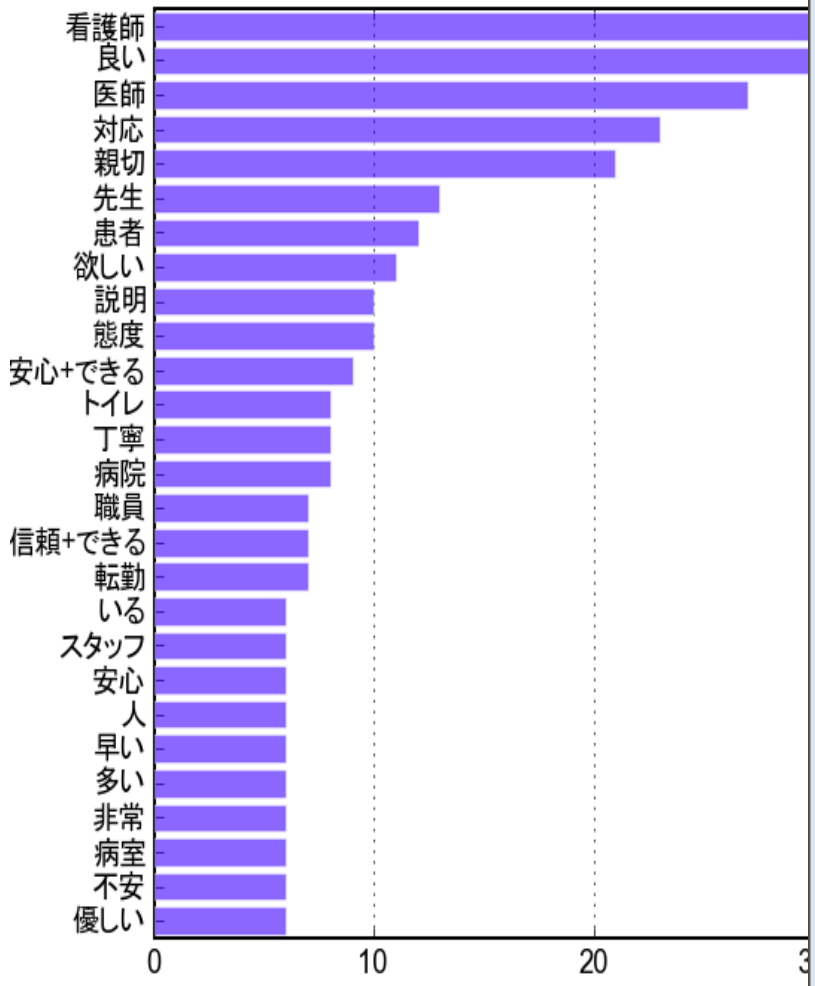


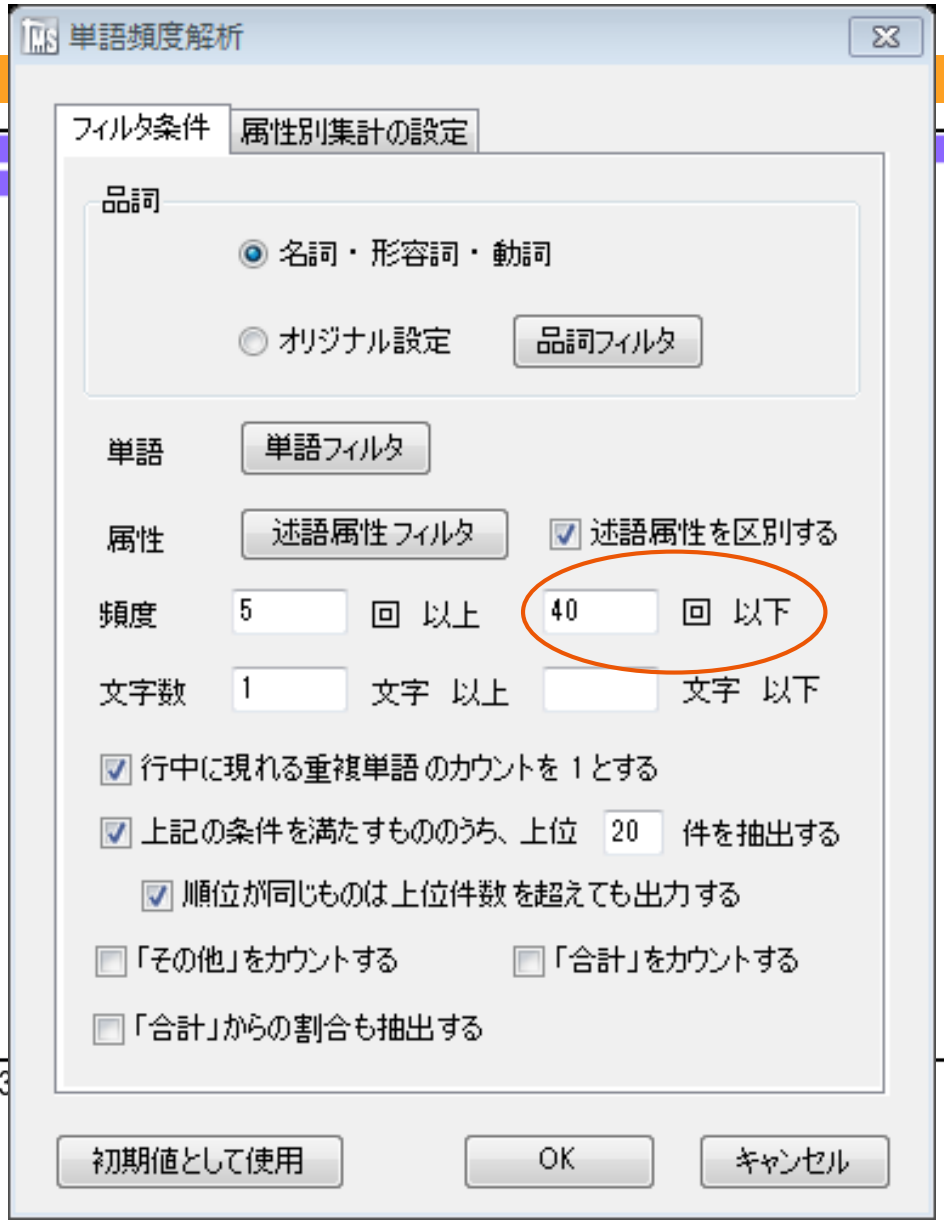

- どのような単語が多く使用されているかをみる分析
- 多く使用されている単語ほど、その文章の中でのキーワードとなるが、多いことだけが重要なのではなく、下位にもポイントとなる単語がある事もある。
  そのため原文と照らし合わせながら注目する

# 単語頻度解析

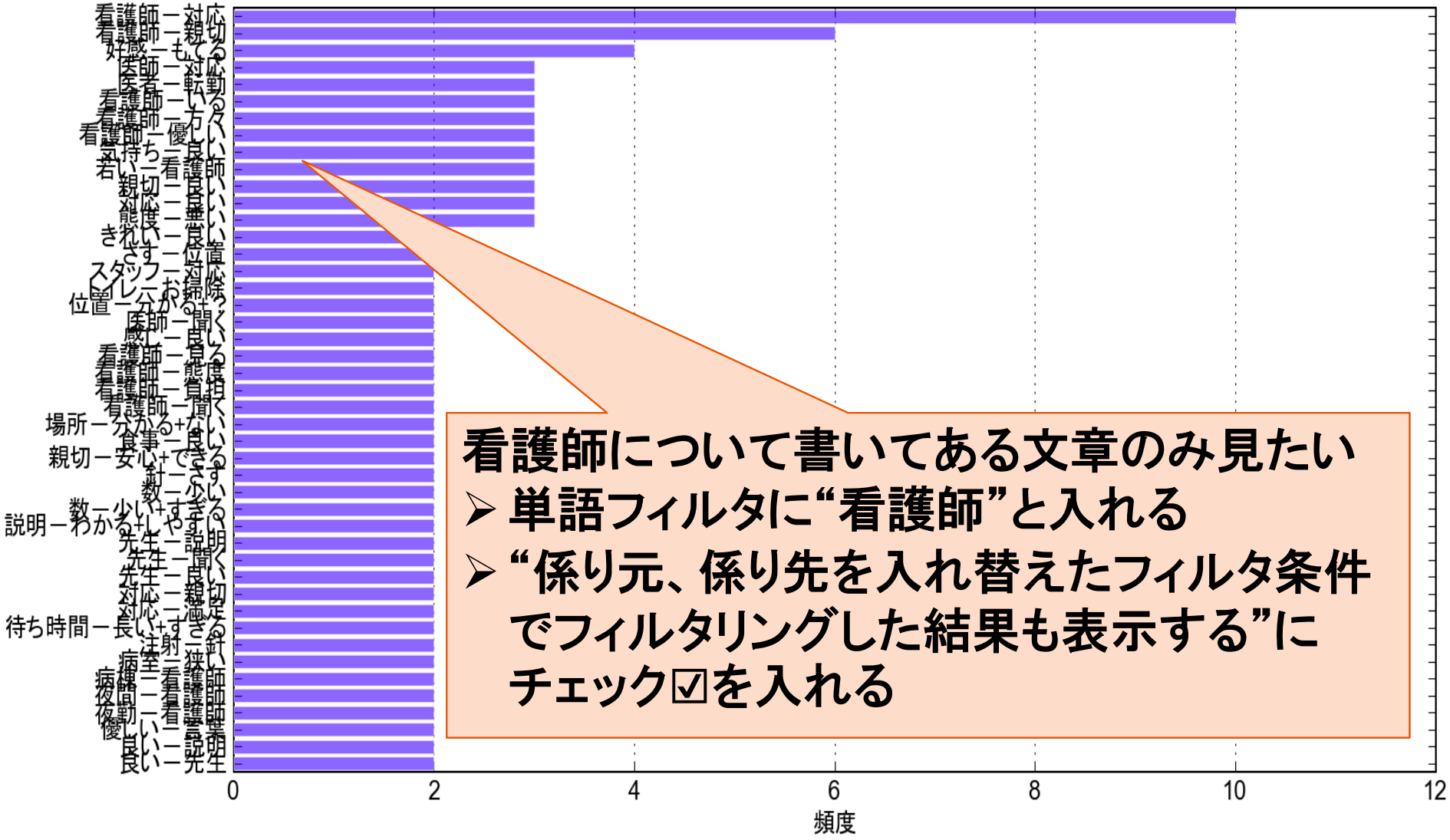

看護師
良い
医師
対応
親切
先生
患者
欲しい
説明
態度
安心+できる
トイレ
丁寧
病院
職員
信頼+できる
転勤
いる
スタッフ
安心
0
10
20
30
40
50
60
70
80
看護師についての文章なので、「看護師」が多いのは必然的
必然的な単語があるために、切り捨てられて見えなくなってしまう単語もある

# 単語頻度解析


看護師
良い
医師
対応
親切
先生
患者
欲しい
説明
態度
安心+できる
トイレ
丁寧
病院
職員
信頼+できる
転勤
いる
スタッフ
安心
人
早い
多い
非常
病室
不安
優しい
0
10
20


単語頻度解析
フィルタ条件
属性別集計の設定
品詞
名詞・形容詞・動詞
オリジナル設定
品詞フィルタ
単語
単語フィルタ
属性
述語属性フィルタ
述語属性を区別する
頻度
5
回 以上
40
回 以下
文字数
1
文字 以上
文字 以下
行中に現れる重複単語のカウントを 1 とする
上記の条件を満たすもののうち、上位
20
件を抽出する
順位が同じものは上位件数を超えても出力する
「その他」をカウントする
「合計」をカウントする
「合計」からの割合も抽出する
初期値として使用
OK
キャンセル

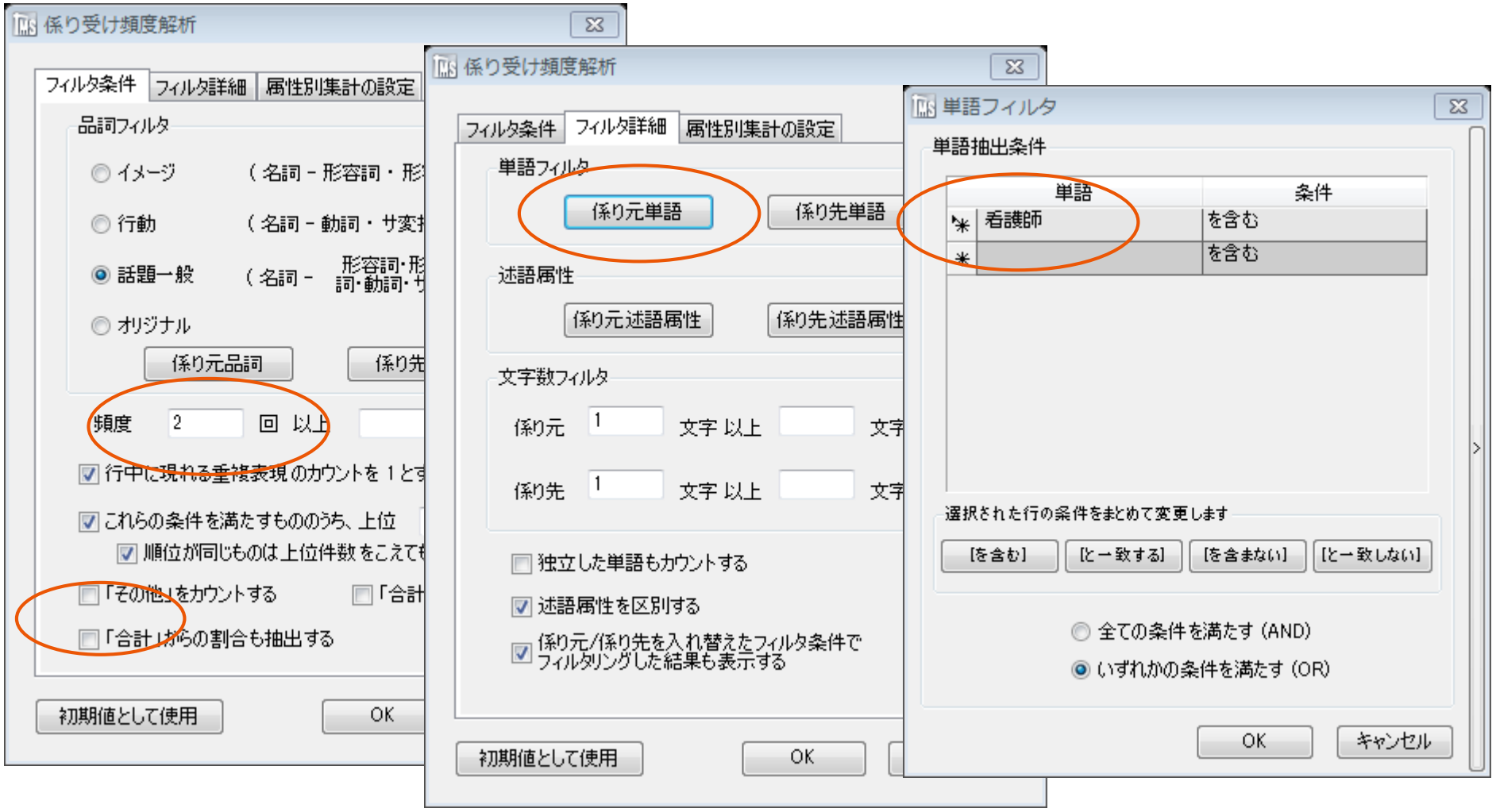
# 係り受け頻度解析



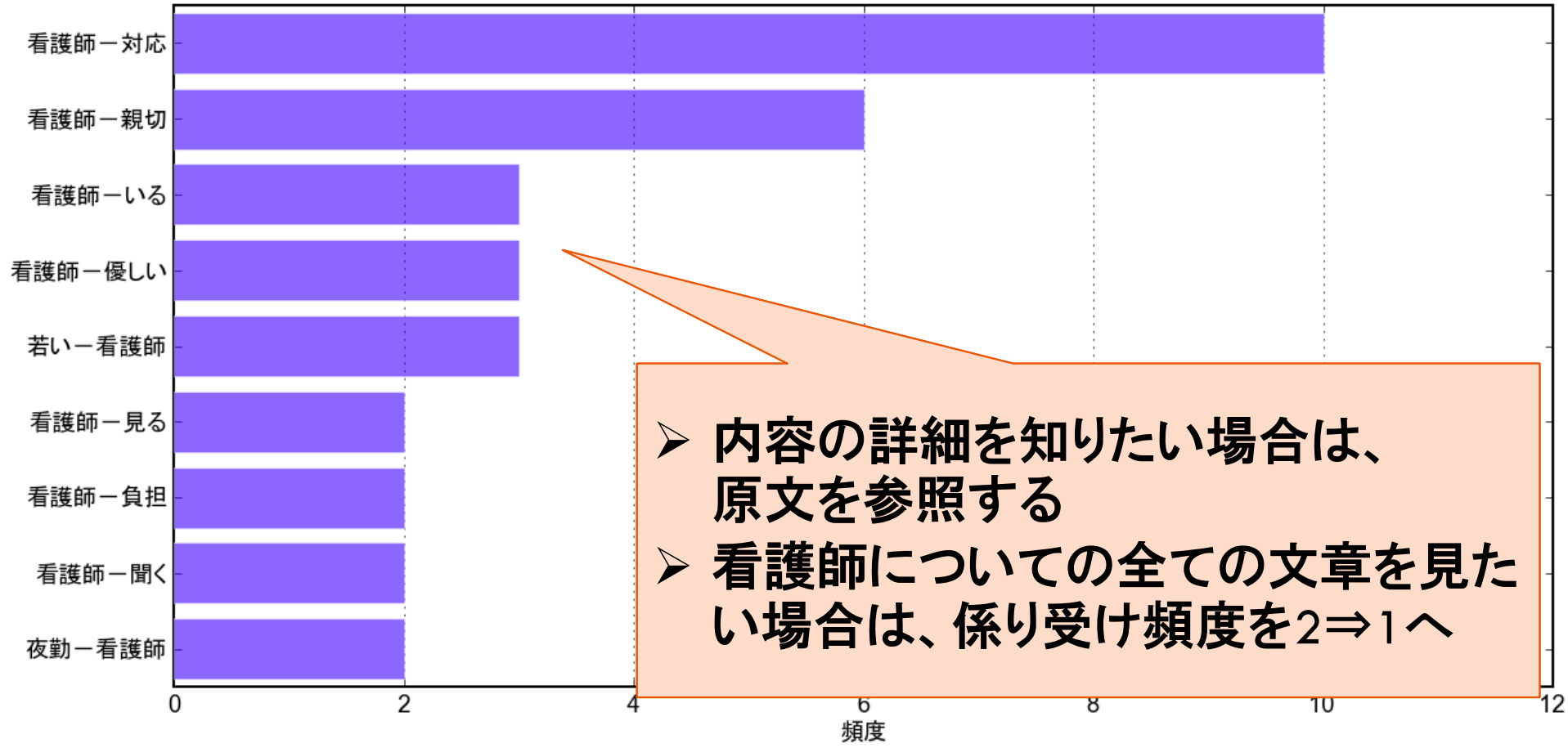

看護師について書いてある文章のみ見たい

- 単語フィルタに“看護師”と入れる
- “係り元、係り先を入れ替えたフィルタ条件でフィルタリングした結果も表示する”にチェック☑を入れる

# 係り受け頻度解析～看護師～



# 係り受け頻度解析～看護師～




- 内容の詳細を知りたい場合は、原文を参照する
- 看護師についての全ての文章を見たい場合は、係り受け頻度を2⇒1へ

# 共起関係と係り受け関係

## 共起関係

- 文章中にある言葉が同時に出現するか否かという確率をもとにした言語現象

ことばの方向性はない

## 係り受け関係

- 主語述語の関係、修飾と被修飾の関係、補助の関係、　並立の関係など文章の中で単語と単語がどのようにつながっているかを示す関係

ことばの方向性がある

**テキストマイニングの基本は共起関係**

# 身近な共起関係


Google
東京医科歯科大学
東京医科歯科大学
東京医科歯科大学病院
東京医科歯科大学 偏差値
東京医科歯科大学 図書館
東京医科歯科大学 歯学部
東京医科歯科大学大学院
東京医科歯科大学歯学部付属病院
東京医科歯科大学 教養部
検索
すべて
画像
地図
動画
ニュース
ショッピング
もっと見る
東京都文京区
場所を変更
他のキーワード: 東京医科歯科大学偏差値 東京医科歯科大学図書館
東京医科歯科大学教養部 東京医科歯科大学生協
東京医科歯科大学難治疾患研究所
国立大学法人 東京医科歯科大学
www.tmd.ac.jp/ - キャッシュ
東京医科歯科大学 TMDU の公式ホームページです。大学案内、入学案内、学部・大学院・附属病院等の紹介、研究活動、産学連携、国際交流など、東京医科歯科大学に関する情報をご覧頂けます。
Google+ ページ
東京
特徴

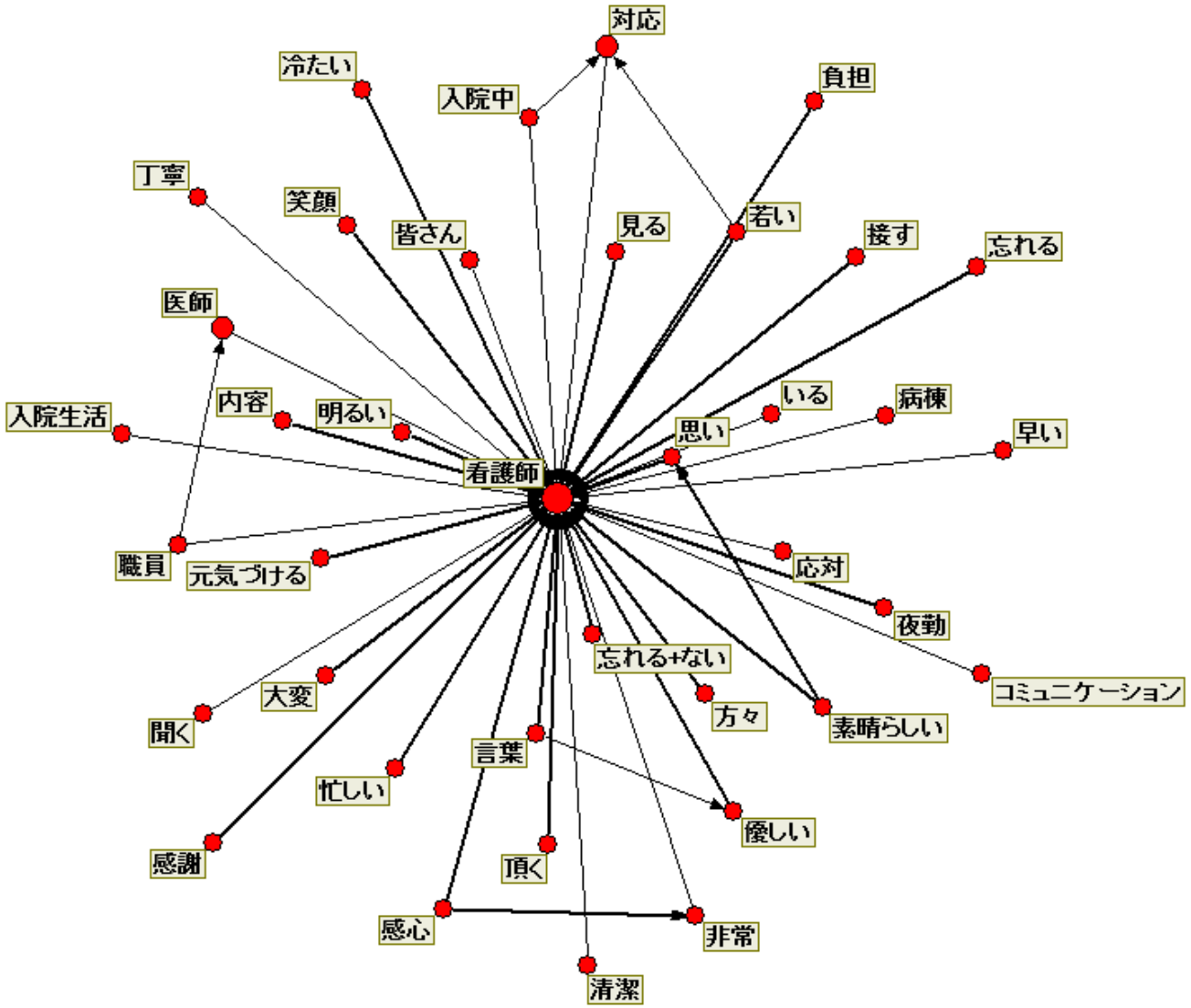
# 注目語情報

- ある特定の単語に注目し、出現頻度や使われ方等の情報から、その単語がテキストの中でどのように使われているかを調べることが出来る

  - 他のどのような単語・属性と同時に出現しているか
    →共起関係⇒**ネットワーク図**で表示

  - 注目した単語がどのような表現の中で使われているか
    →係り受け関係⇒**注目語表現情報**で表示

# 注目語情報～ネットワーク図～

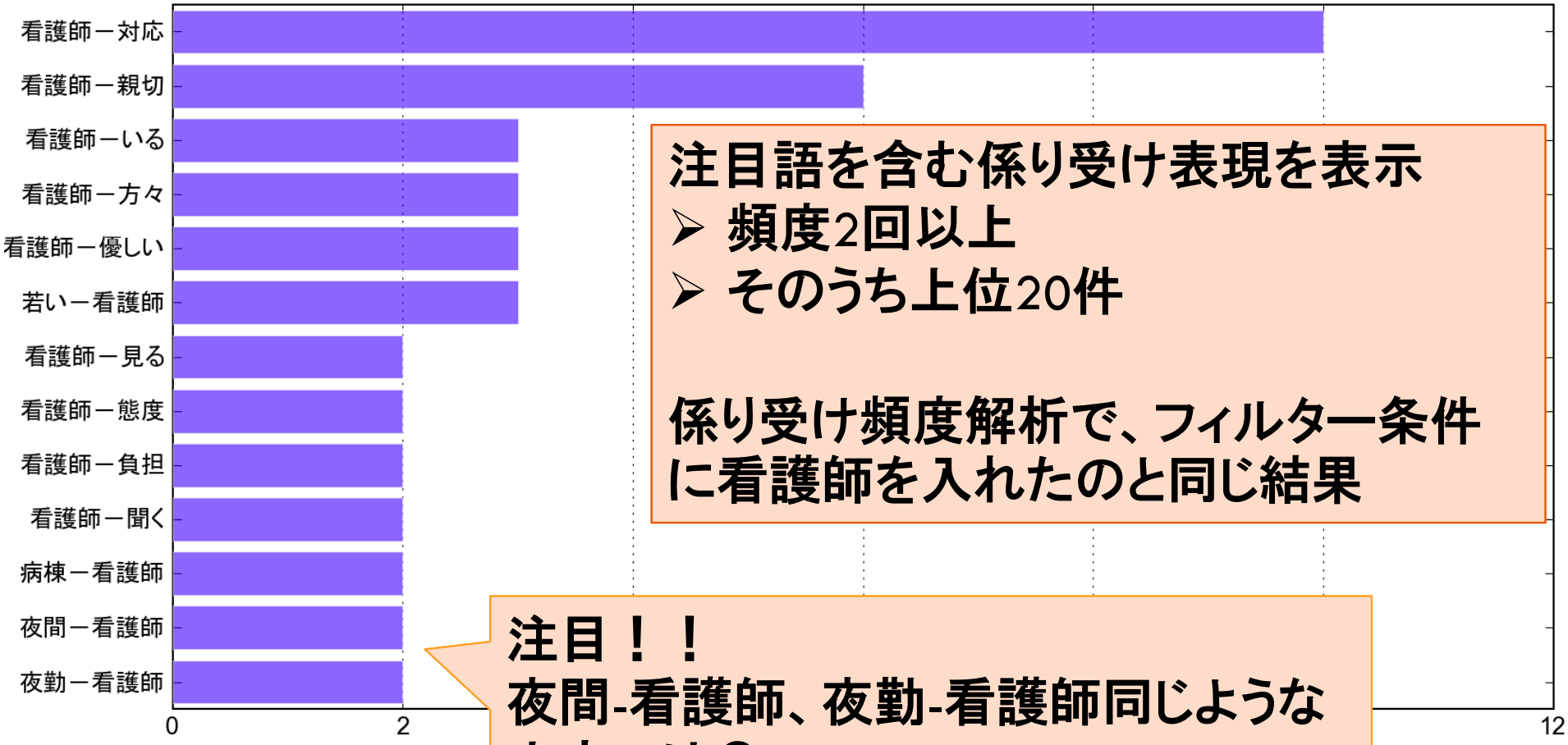



下記の共起ルールに
当てはまるものを
最大100ルール図示

- 注目語：看護師
- 出現頻度2回以上
- 行単位での共起

看護師と共起性が
強い単語は太い線で、
弱い単語は細い線で
結ばれる

# 注目語情報～注目語表現情報～




注目語を含む係り受け表現を表示

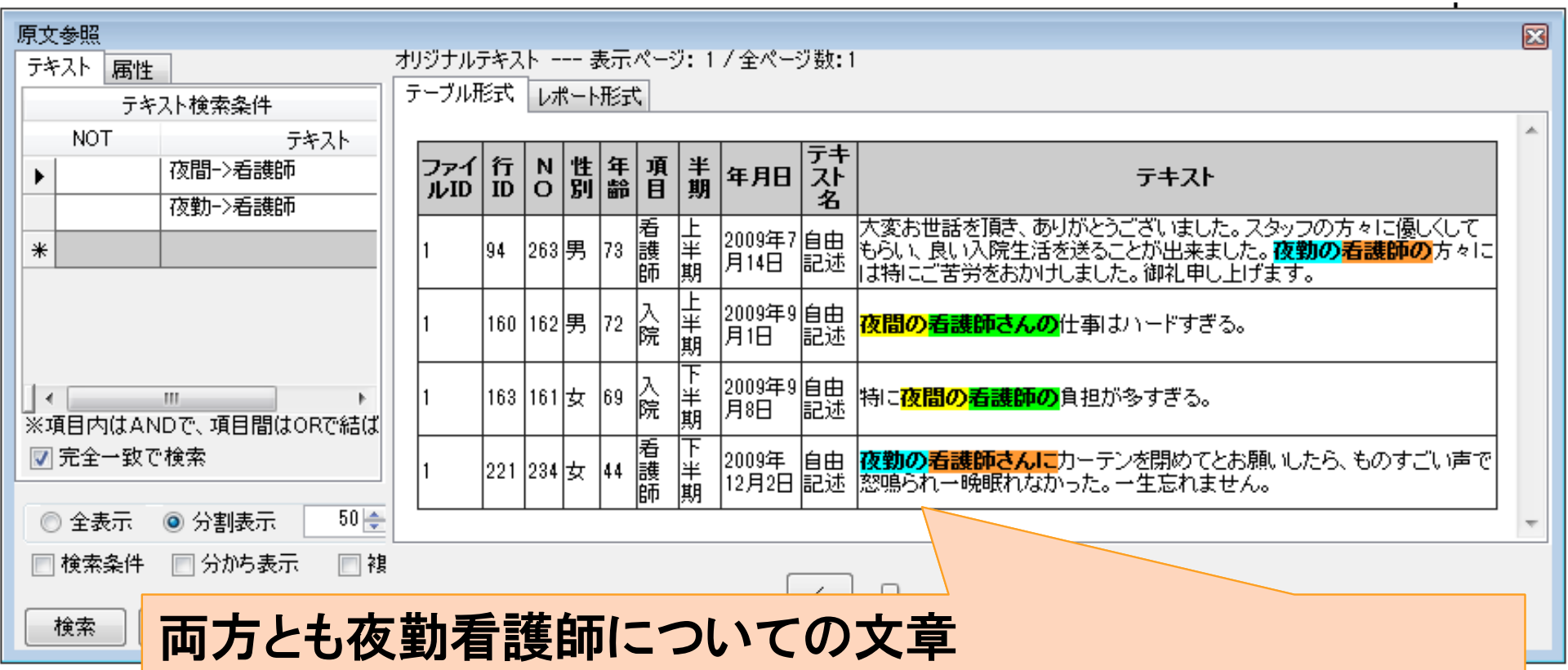
- 頻度2回以上
- そのうち上位20件

係り受け頻度解析で、フィルター条件に看護師を入れたのと同じ結果

注目！！
夜間-看護師、夜勤-看護師同じような内容では？

- 原文を検索

# 原文参照



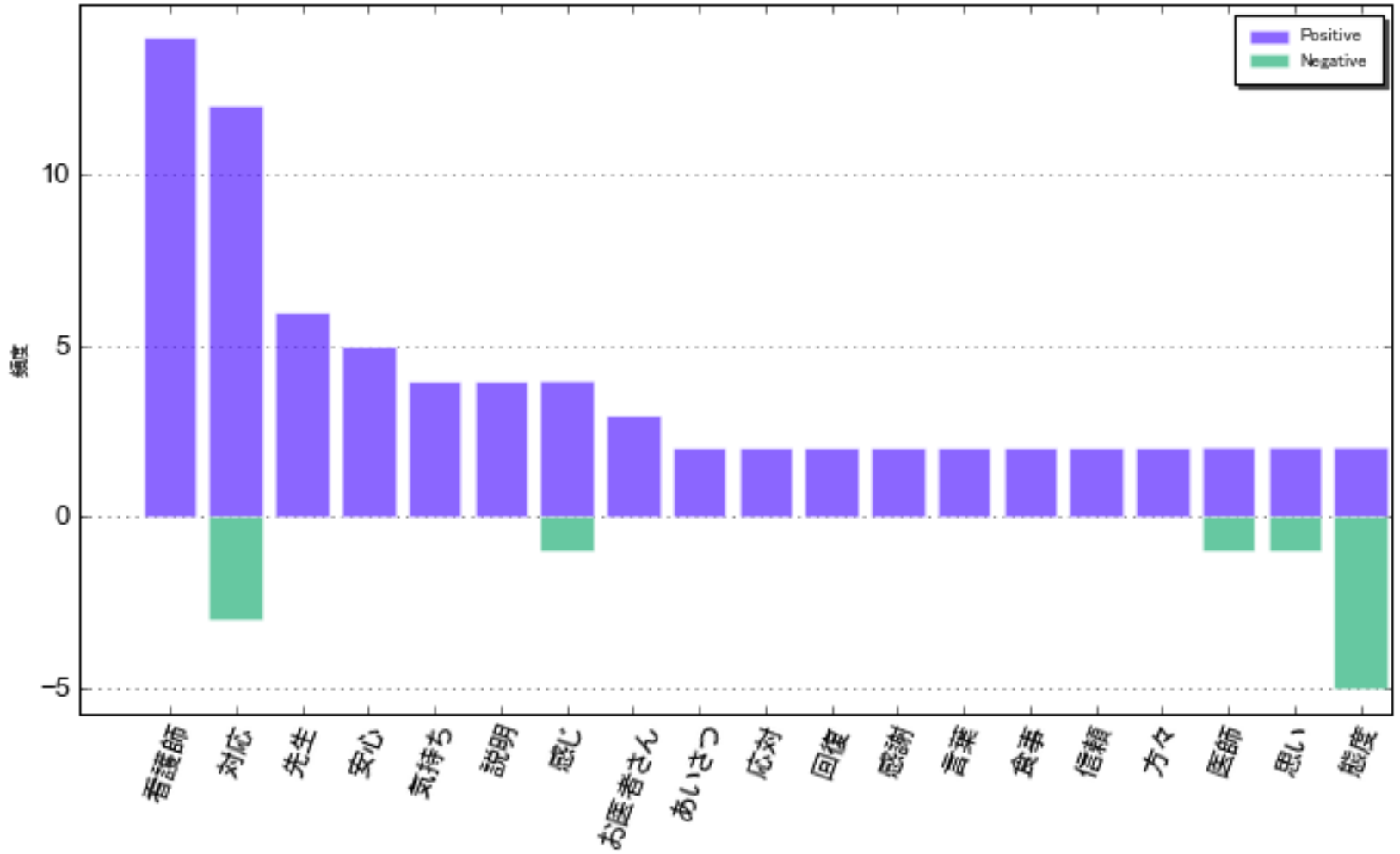

**両方とも夜勤看護師についての文章**

- **“夜勤”と“夜間”を類義語辞典で一つにまとめて登録することもできる**
- **夜間、夜勤のみの原文も参照することが重要！**

# 評判分析

- 分析対象となる単語を、その単語に係り受けする好評語（ポジティブなイメージを持つ単語）と不評語（ネガティブなイメージを持つ単語）の頻度から単語の評価を分析する。
- 好評語と不評語はもともとTMSのなかで割り当てられている。
- ただ、文章が複雑になると、正確に評価されない場合があるため、原文に戻って確認、追加する必要がある。

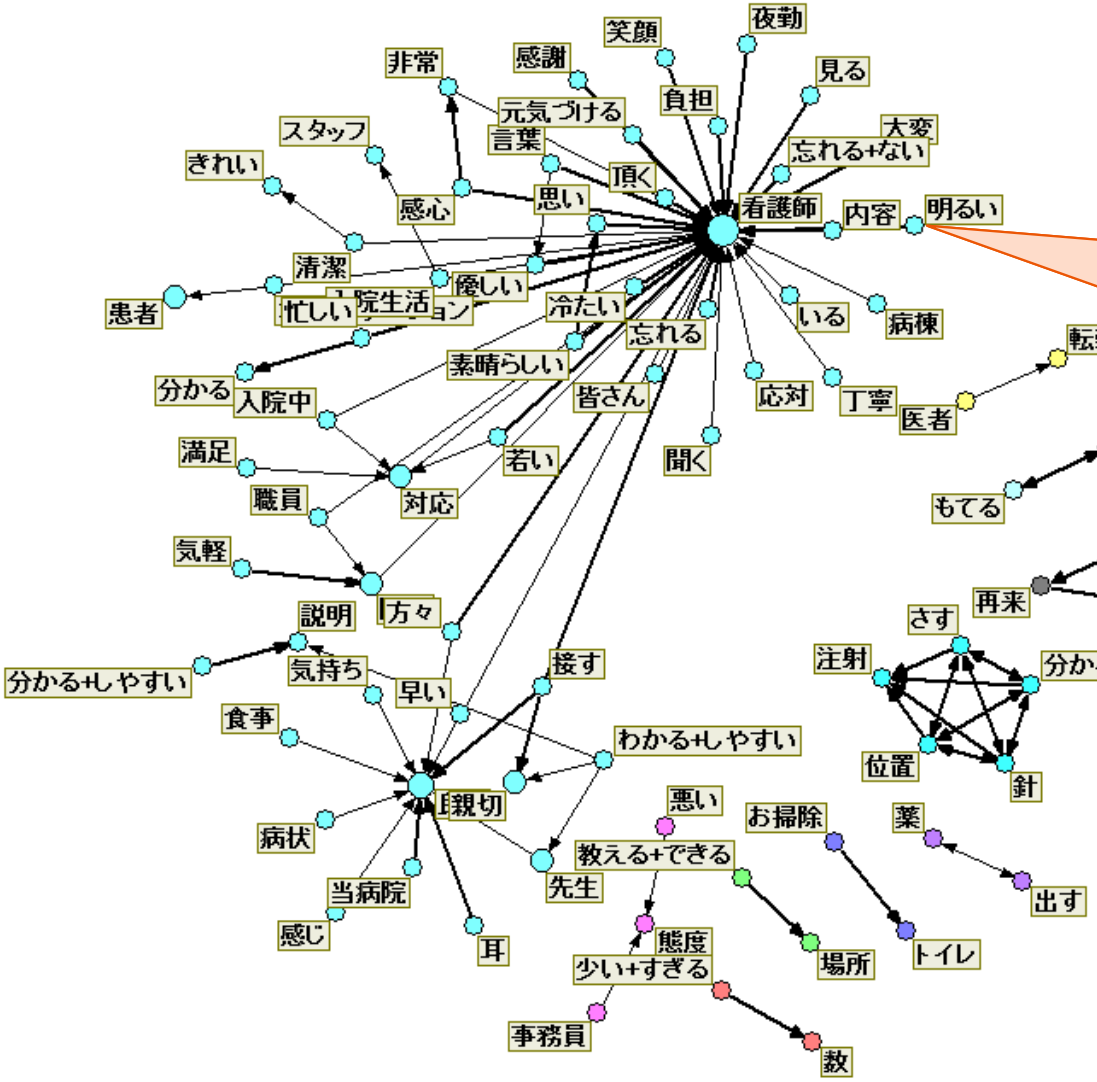
# 評判分析


Positive
Negative
頻度
10
5
0
-5
看護師
対応
先生
安心
気持ち
説明
感じ
お医者さん
あいさつ
応対
回復
感謝
言葉
食事
信頼
方々
医師
思い
態度

# 言葉ネットワーク分析

- ことばとことばの関連をグラフを使用して可視化する分析方法
- 共起関係や係り受け関係をもとに、ことば同士の関連の強さを表す方法がある
- 頻度が高いほど関連が強いという考えをもとした分析

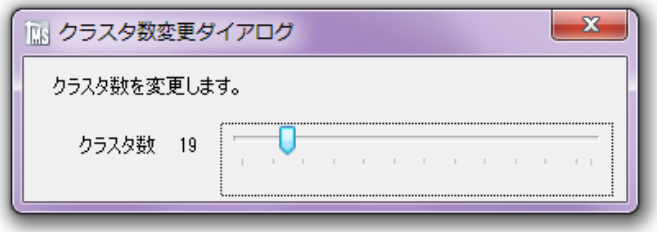
# ことばネットワーク

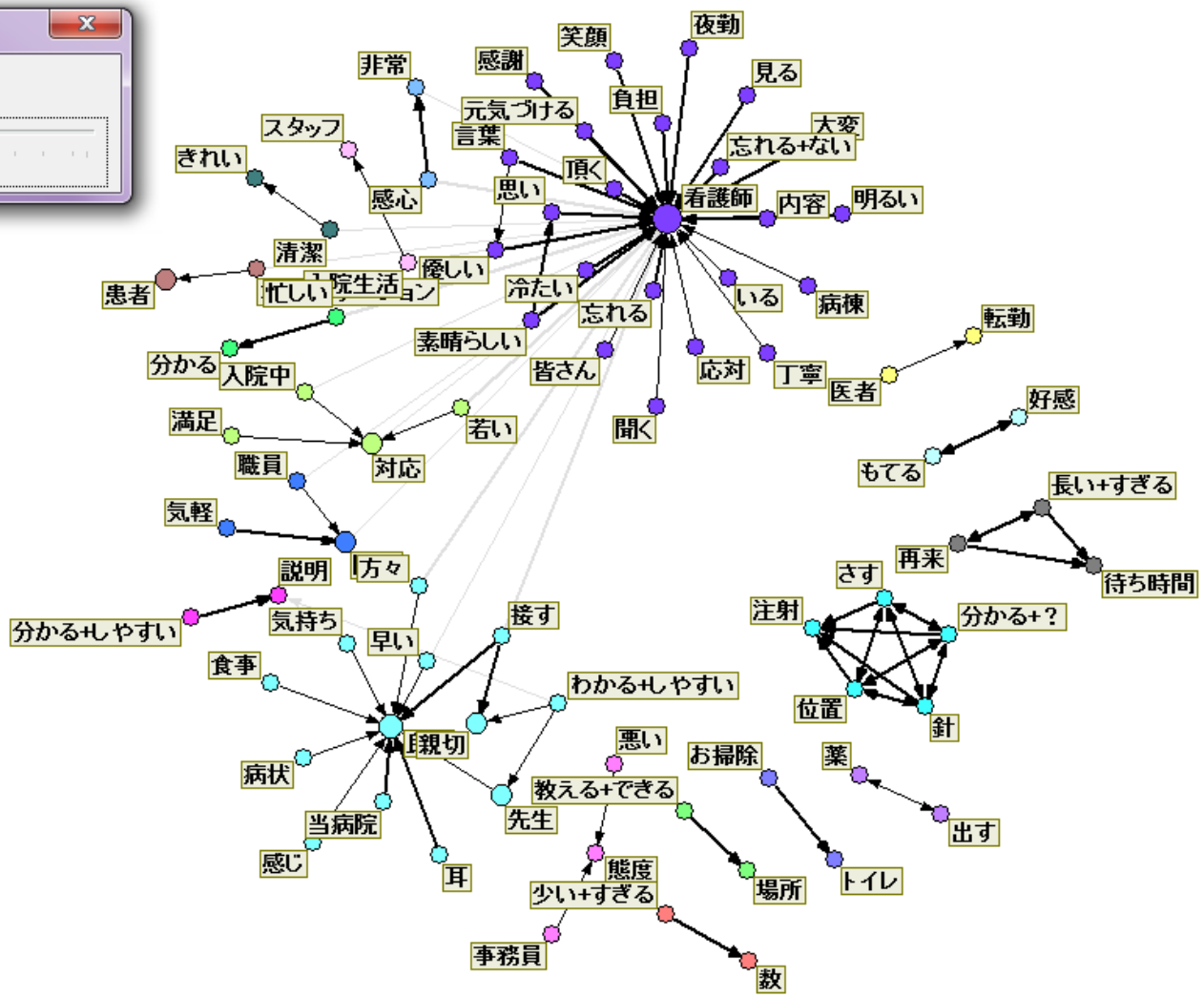



ノード
・頻度を表している
・大きければ大きいほど頻度が高い

エッジ
・関連の強さを表している
・太ければ太いほど関連が強い

# クラスタ数変更


クラスタ数変更ダイアログ
クラスタ数を変更します。
クラスタ数 19


笑顔
夜勤
非常
感謝
見る
負担
元気づける
言葉
大変
忘れる+ない
スタッフ
きれい
頂く
思い
看護師
内容
明るい
感心
清潔
優しい
患者
忙しい
入院生活
冷たい
いる
病棟
忘れる
転勤
素晴らしい
分かる
入院中
皆さん
応対
丁寧
医者
好感
満足
若い
聞く
もてる
対応
職員
長い+すぎる
気軽
再来
方々
さす
待ち時間
説明
注射
分かる+?
分かる+しやすい
気持ち
早い
接す
食事
わかる+しやすい
位置
針
親切
悪い
お掃除
薬
病状
教える+できる
出す
先生
当病院
態度
トイレ
感じ
耳
場所
少い+すぎる
事務員
数

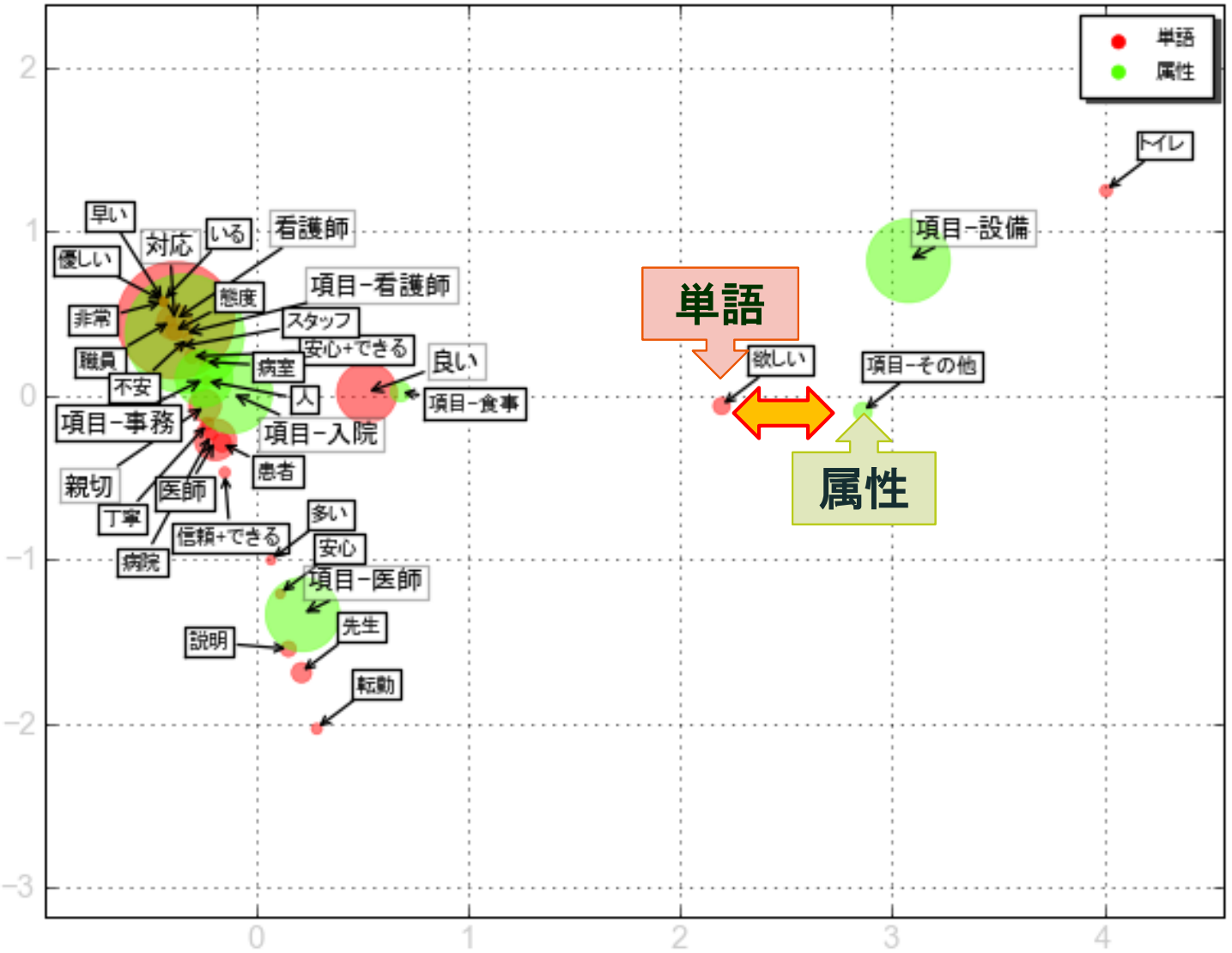
# 対応バブル分析

- 対応(コレスポンディング)分析をもとにした分析手法。
  文中の単語と属性との関係を、単語頻度をもとに
  2次元のバブルで視覚化する。

- 単語の使われ方が似ている属性は近くに表示される。

- バブルの大きさは単語や属性の頻度に対応する。

# 対応バブル分析


単語
属性
トイレ
項目−設備
欲しい
項目−その他
単語
属性
早い
対応
いる
看護師
優しい
態度
項目−看護師
非常
スタッフ
安心+できる
職員
病室
良い
不安
人
項目−食事
項目−事務
項目−入院
親切
患者
医師
丁寧
多い
信頼+できる
安心
病院
項目−医師
説明
先生
転勤
2
1
0
−1
−2
−3
0
1
2
3
4

# 注意して分析したい点

- 今まで困難であったテキストデータの分析が
  パソコンで容易にできる夢のツール ・・・ではない
- コンピュータにすべてを任せるのではなく、
  研究者の目によるアナログな確認が重要
- 結果から何を読み取るかを常に考えて分析しないと
  ことばの確率論に終始する

# 最後に・・・

- 壊れることはほぼありません！
  - まずは、触りながらいろいろな分析を試してみましょう!
- 分析のOKボタンを押す時に、自分で問いかけましょう。この分析は何の目的で、何を知りたくてするのか。ただ押すだけでは、カチカチクリックの中身のない結果になってしまいます。
- 結果を読むのは人間です。他の統計分析と同じで、出てきた結果をどう読み取るかは、研究者のセンスや知識、経験が重要になります。