全文検索システム『ひまわり』講習会

全文検索システム『ひまわり』講習会

山口昌也（国立国語研究所／東京外国語大）

本日の内容

- 全文検索システム『ひまわり』を使って，既存コーパスの検索と分析用基礎データの集計する方法を紹介
  - 『ひまわり』（ver.1.7.1）
  - 『国会会議録』パッケージ
  - 『名大会話コーパス』
  - サンプルデータ（青空文庫2作品）

- 全体的な流れ
  - 『ひまわり』の紹介と基本的な使い方
  - さまざまな検索と各種機能
  - コーパスの構造と検索・集計
  - テキストデータのインポート

ツール・資料などの確認

- ツール・資料のインストール
  - 『ひまわり』
  - 『国会会議録』パッケージ
  - 『名大会話コーパス』パッケージ
  - サンプルデータ

- 当日配布資料(この資料)

macOSの「プレビュー」では, PDF中のハイパーリンクが機能しない場合があるので,
ChromeやFirefoxに本ファイルをドロップして閲覧してください。

『ひまわり』とは

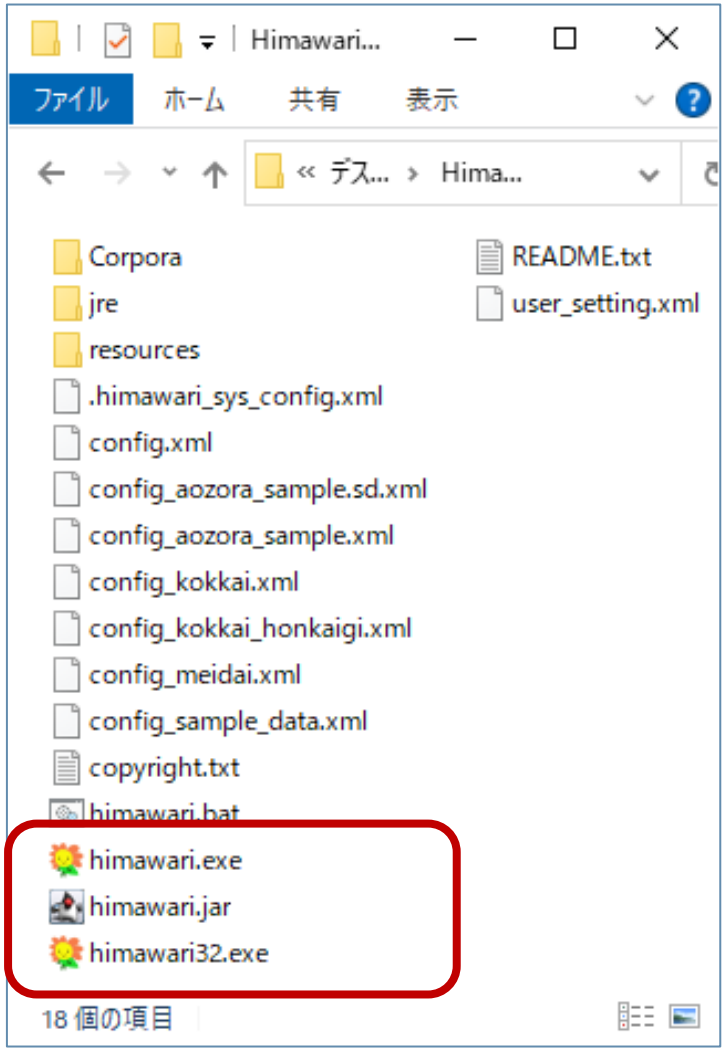
- 言語研究用の全文検索システム
  - 指定された文字列を網羅的に検索して，前後文脈付きで結果を表示します（コンコーダンサ）
  - 『太陽コーパス』（20世紀初頭の総合雑誌『太陽』）用の検索システムとして構築しました

- 特徴
  - XMLでタグづけされたコーパスを全文検索できます
  - Windows, Mac OS, Linux など，多くのOS上で動作します
  - 無料です

分析用基礎データの集計機能やコーパス自作支援機能を強化
（例：総文字数，総単語数）

『ひまわり』の基本的な使い方

『ひまわり』の起動と『ひまわり』フォルダの確認（Windowsの場合）



himawari.exe

普段使うとき
（Windows 専用，64ビット版）
himawari.exe

himawari32.exe

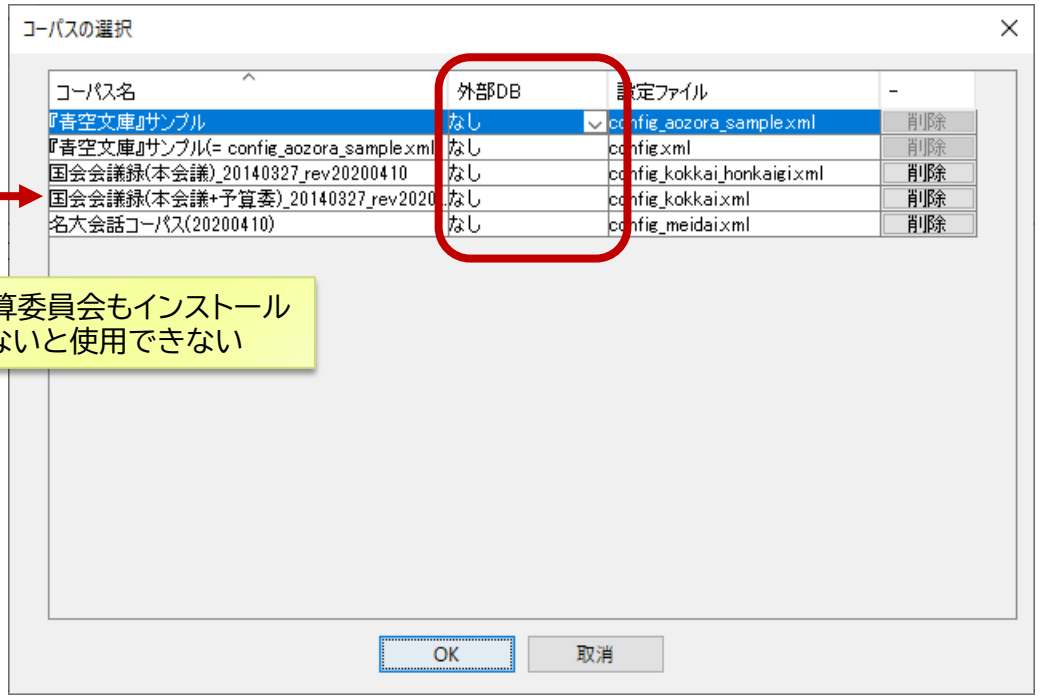
himawari.exeが動かないとき
（Windows 専用，32ビット版）
himawari32.exe

himawari.jar

汎用
（Windows, Mac, Linux など）
himawari.jar

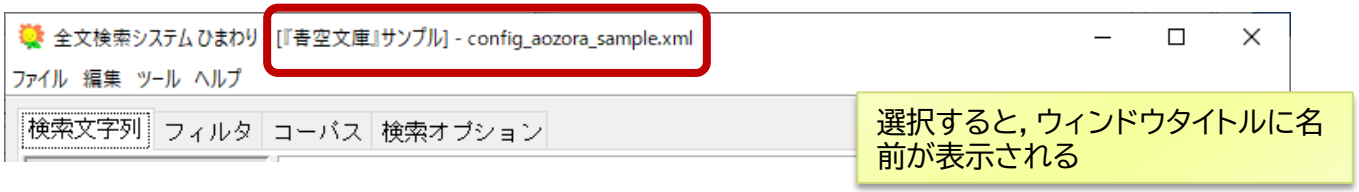
コーパスの選択

- [ファイル]⇒[コーパス選択]



- 「外部DB」
  - コーパスファイルに直接記述していない付与データを格納
  - 『青空文庫』サンプルの場合は，形態素解析結果

- 従来どおり，設定ファイルを『ひまわり』にドロップする方法でもOK



検索する

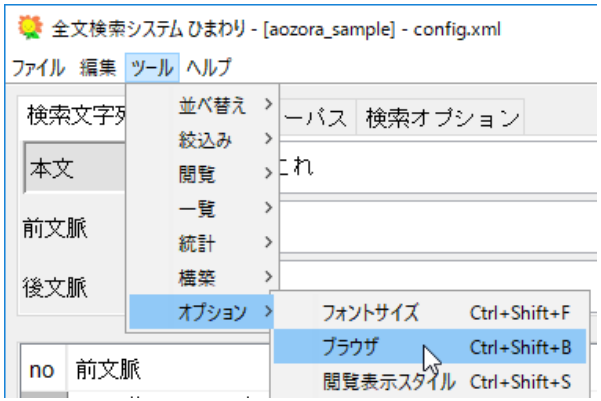

「検索文字列」欄では右クリックで履歴表示
全文検索システム ひまわり - [aozora_sample] - config.xml
ファイル 編集 ツール ヘルプ
検索文字列
検索の実行
検索文字列
本文
これ
前文脈
で終る
後文脈
で始まる
検索
字体変換
クリア
no 前文脈 キー 後文脈 Path タイトル 著者
検索結果
検索総数:597
検索総数
途中経過の表示

ブラウザでの閲覧

no	前文脈	キー	後文脈	Path
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...
3	弾くところです」「	これ	からいよいよヴァイオ	/aozora_s...
4	い話があるかい」「	これ	からいよいよヴァイオ	/aozora_s...
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aozora_s...
6	見当がつかない」「	これ	からいよいよ弾くとこ	/aozora_s...
7	めちゃなお困ります。	これ	からがいよいよ佳境に	/aozora_s...

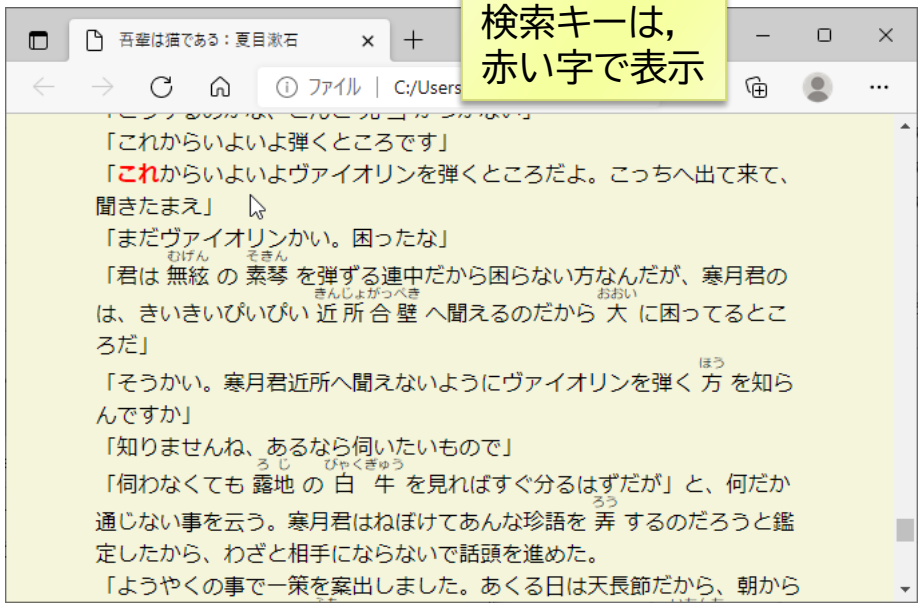
閲覧したい用例をダブルクリック

■閲覧用のブラウザの変更



[ツール]⇒[オプション]⇒ [ブラウザ]

検索キーは，赤い字で表示



検索結果のソート

列名を左クリック

no	前文脈	キー	後文脈	Path	タイトル	著者
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...	こころ	夏目漱石
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目漱石
3	弾くところです」「	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目漱石
4	い話があるかい」「	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目漱石
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aozora_s...	吾輩は猫...	夏目漱石
6	見当がつかない」「	これ	からいよいよ弾くとこ	/aozora_s...	吾輩は猫...	夏目漱石
7	めちゃなお困ります。	これ	からがいよいよ佳境に	/aozora_s...	吾輩は猫...	夏目漱石
8	うと云うんです。さあ	これ	からがいよいよ失恋に	/aozora_s...	吾輩は猫...	夏目漱石
9	はすこぶる不憫だよ。	これ	からがいよいよ巧妙な	/aozora_s...	吾輩は猫...	夏目漱石
10	充分あらわれている。	これ	からが化物の記述だ。	/aozora_s...	吾輩は猫...	夏目漱石
11	か両君能く聞き給え、	これ	からが結論だぜ。—	/aozora_s...	吾輩は猫...	夏目漱石

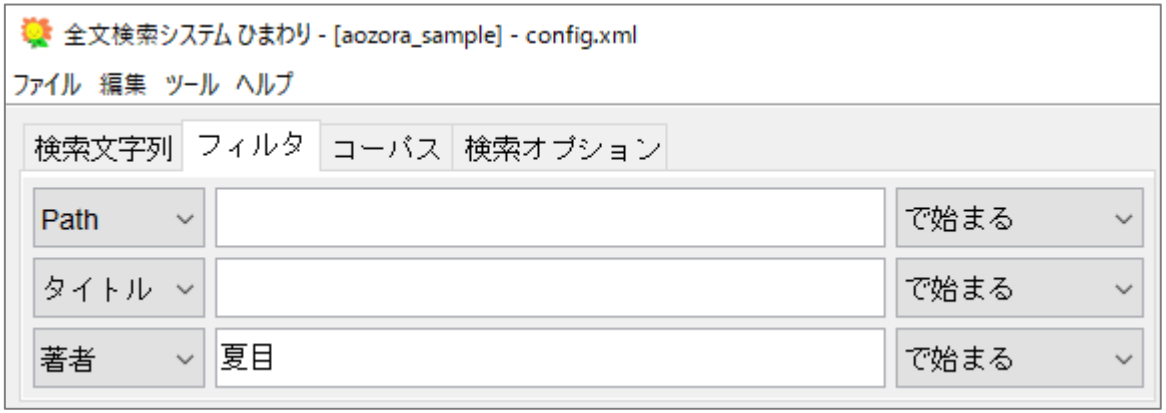
- 昇順
  列タイトルをクリック
- 降順
  シフトキーを押しながら
  列タイトルをクリック

- 複数列を考慮したい場合
  - 優先順位の逆順でソートを実行

例:「タイトル」ごとに「後文脈」でソート
→ 「後文脈」「タイトル」の順

検索結果の絞り込み

- 検索時に指定



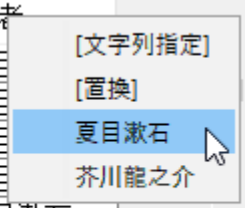
「著者」欄が「夏目」で始まる結果のみに絞り込まれる

- 検索後に絞り込み

列名を右クリック

絞り込みたい値を選択
⇒右クリック
⇒フィルタでもOK

no	前文脈	キー	後文脈	Path	タイトル	著者
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...	こころ	夏目
	これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目
	「	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	「	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	て、	これ	からいよいよ会津領へ	/aozora_s...	吾輩は猫...	夏目漱石
	「	これ	からいよいよ弾くとこ	/aozora_s...	吾輩は猫...	夏目漱石
	す。	これ	からがいよいよ佳境に	/aozora_s...	吾輩は猫...	夏目漱石



検索結果の頻度集計

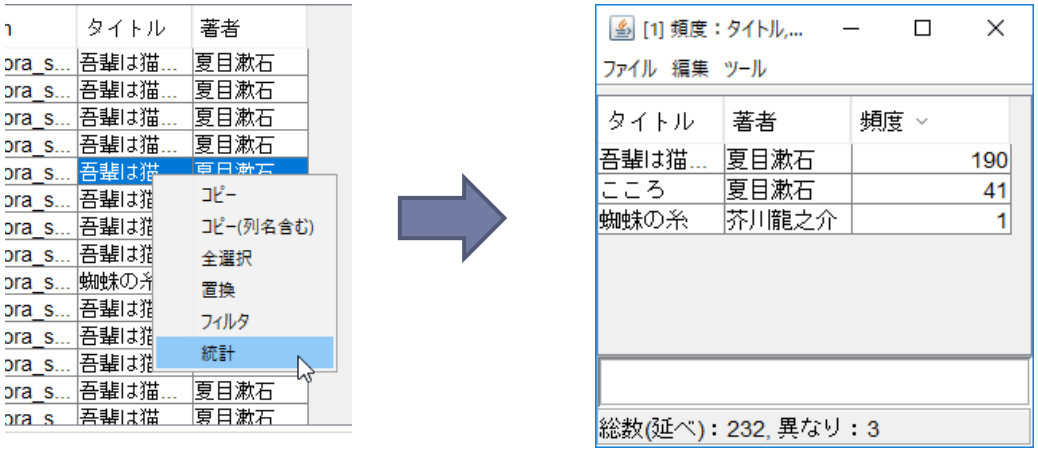
1. 集計したい列を選択

no	前文脈	キー	後文脈	Path	タイトル	著者
1	これは本当の嘶だと、	あの	うそつきの爺やが申し	/aozora_s...	吾輩は猫...	夏目漱石
2	ました、なに猫だから	あの	くらいで充分浄土へ行	/aozora_s...	吾輩は猫...	夏目漱石
3	が来ましたぜ。月並も	あの	くらいになるとなかな	/aozora_s...	吾輩は猫...	夏目漱石
4	まで随分ひきましたが	あの	くらい美しい音が出た	/aozora_s...	吾輩は猫...	夏目漱石
5	なら、立町は豚仙さ、	あの	くらい食い意地のきた	/aozora_s...	吾輩は猫...	夏目漱石
6	ますまい」と云う。「	あの	ちょっとくらい外出致	/aozora_s...	吾輩は猫...	夏目漱石
7	雪江さんが聞く。「	あの	ね。あとでおならは御	/aozora_s...	吾輩は猫...	夏目漱石
8	さんは謙遜した。「	あの	ね。坊たん、坊たん、	/aozora_s...	吾輩は猫...	夏目漱石

複数の列を
選択することも可

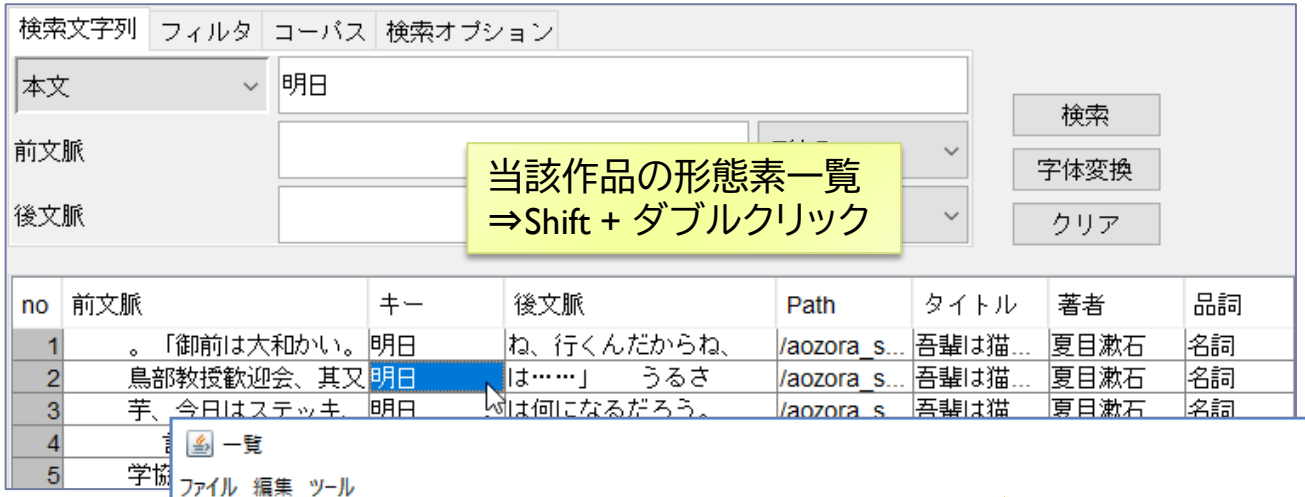
離れた列の選択
- WindowsはCtrlキー
- macOSはcommandキー

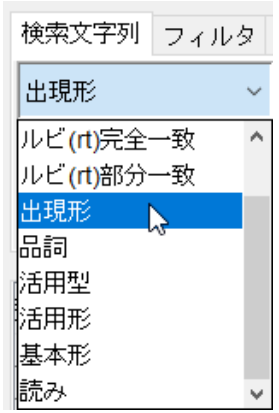
2. 右クリック⇒「統計」

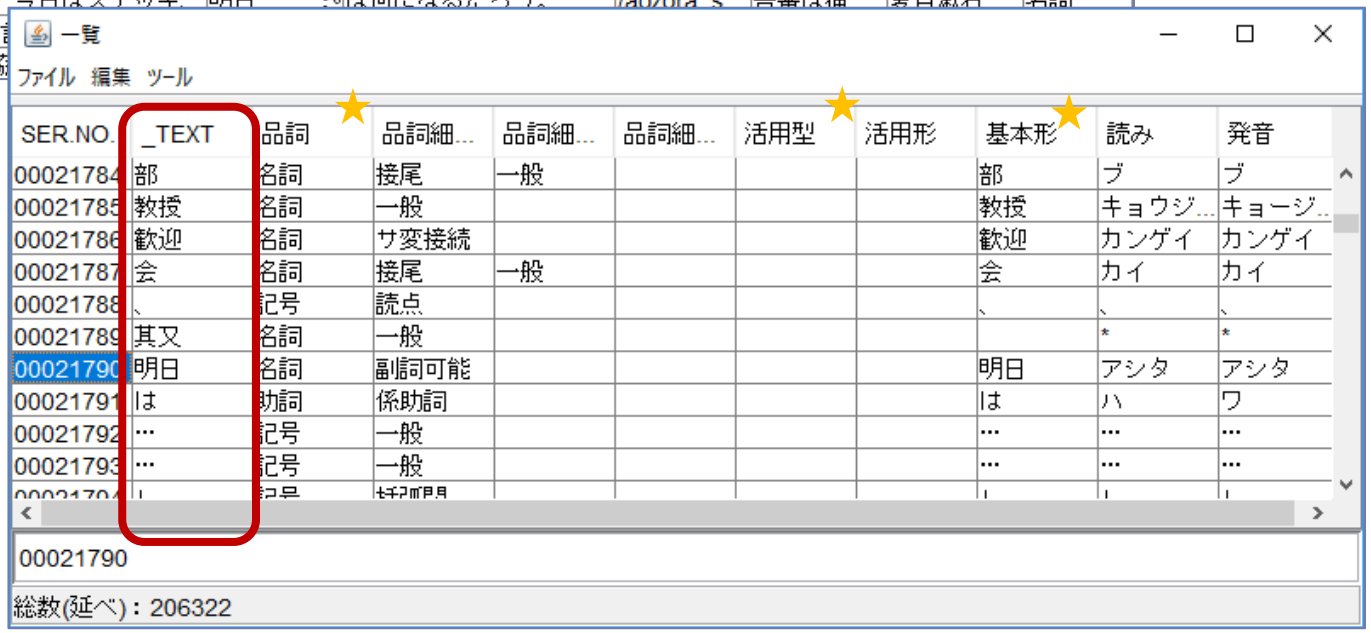


形態素解析結果の閲覧

この機能は，
外部DB「sd」の資料のみ実行可能







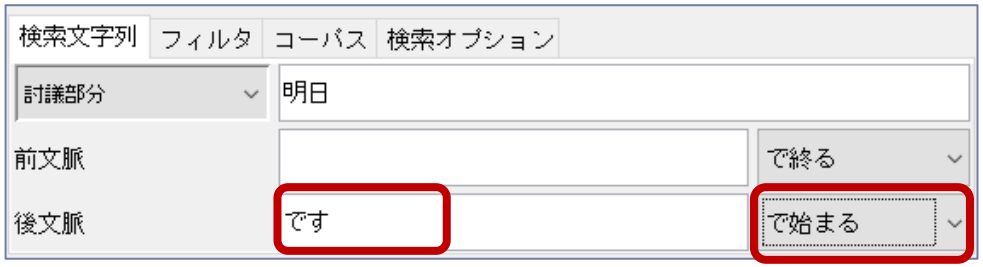
テキスト
進行方向

さまざまな検索と各種機能

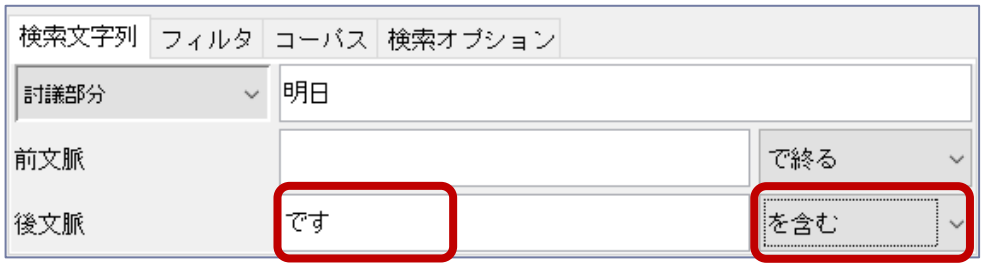
前後文脈の制限

国会

A) 後文脈を
「です」で「始まる」に制限

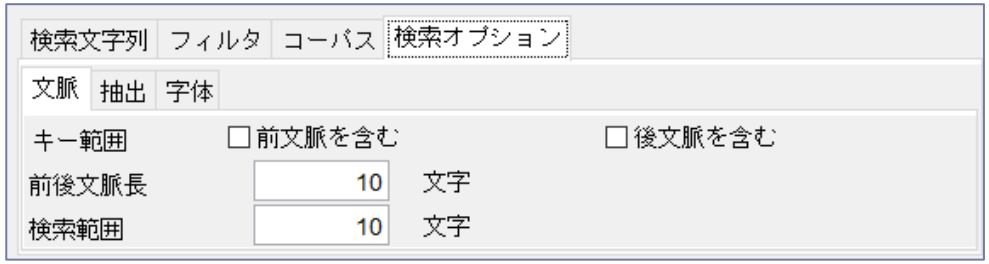


B) 後文脈の制限を
「です」を「含む」にする
(「正規表現」と同一)



- 検索オプション(文脈)
  - キー範囲: 一致した前後文脈をキーに統合
  - 前後文脈長: 表示用
  - 検索範囲: 検索用

基本的に同じ値にする

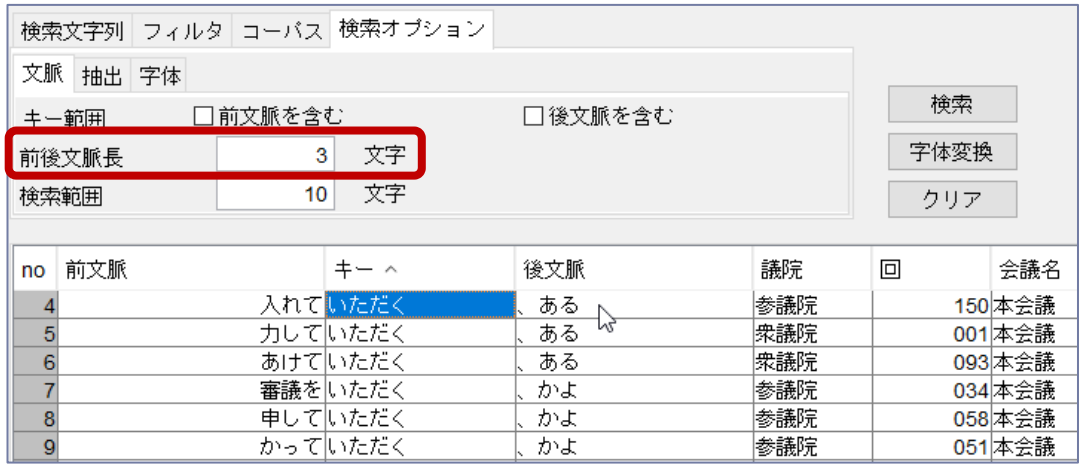


文脈情報の集計
(「いただく」に前接する文字列の集計例)

国会

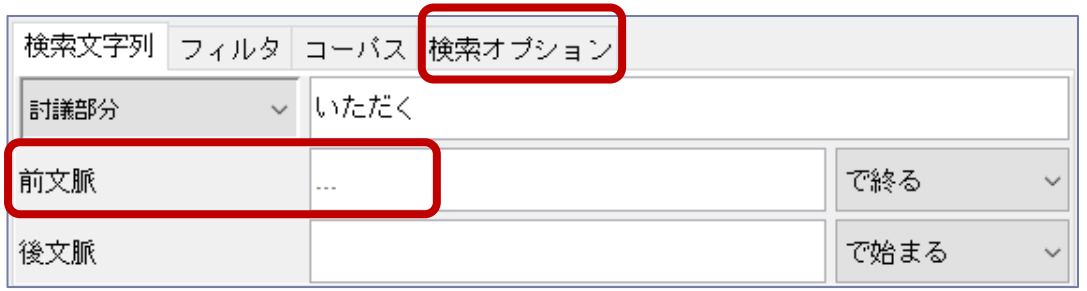
- 方法1:
  前後文脈長を調節

① 前後文脈長の
オプション指定
② 「全文」で「いただく」
を検索
③ 前文脈で集計



- 方法2:
  前後文脈の検索条件

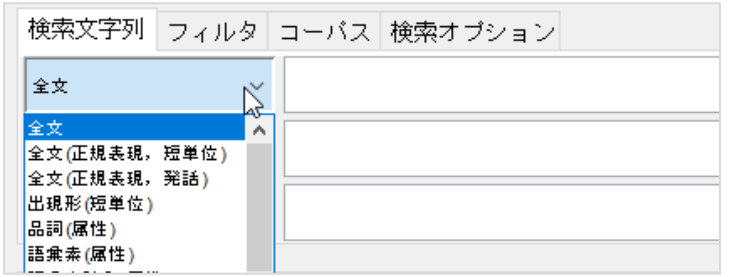
① 前文脈に正規表現
「...」(任意の3文字)
② 「前文脈を含む」を
チェック
(検索オプション)



本文の正規表現検索

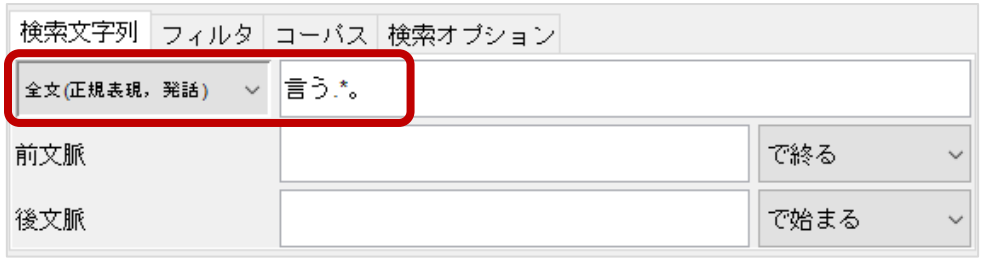
名大

- 検索速度は，通常の全文検索よりも低速
- 検索範囲は，「発話」「短単位」の2種類
  （名大会話コーパスの場合）



A) 「言う」で始まり，「。」で終わる文字列
（「言う。」「言うんです。」など）

正規表現 「言う.*。」



B) 雑多な例
- 言う.{0,5}。
- 言[わいうえおっ]
- ようやく[^。！]*。
- (..)¥1

.（ピリオド）	… 任意の1文字
[わいうえ]	… 「わ」「い」「う」「え」のいずれか
[^！。]	… 。！以外の任意の1文字
*	… 直前要素の0個以上の繰り返し
+	… 直前要素の1個以上の繰り返し
+?	… 直前要素の1個以上の繰り返し(最短)
{0,5}	… 直前要素の0～5個の繰り返し
()	… マッチした範囲を記録
¥1	… 1個目の記録した要素

（詳細は本資料末の参考文献参照）

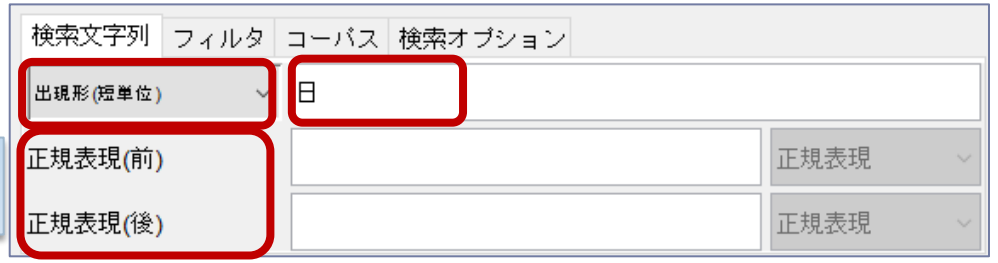
日本語キーボードのmacの場合，「¥」は optionキー+「¥」を使用
なお，『ひまわり』の画面表示ではWindows，macとも逆スラッシュ「＼」になる

単語(タグ)での検索

名大

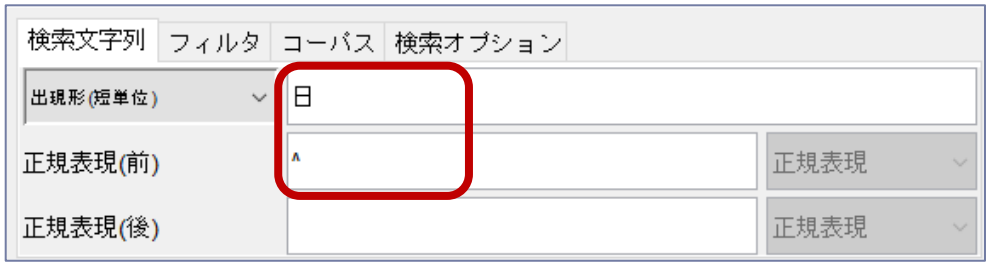
A) 「日」を含む単語

インターフェイスが
変わることに注意



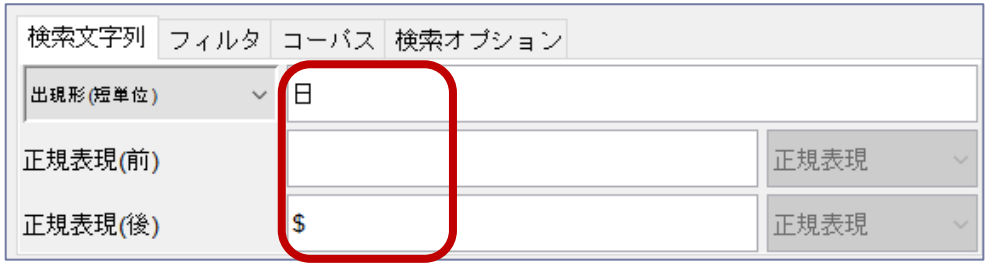
B) 先頭が「日」の単語

正規表現の「^日」と同義
(先頭の文字が「日」)



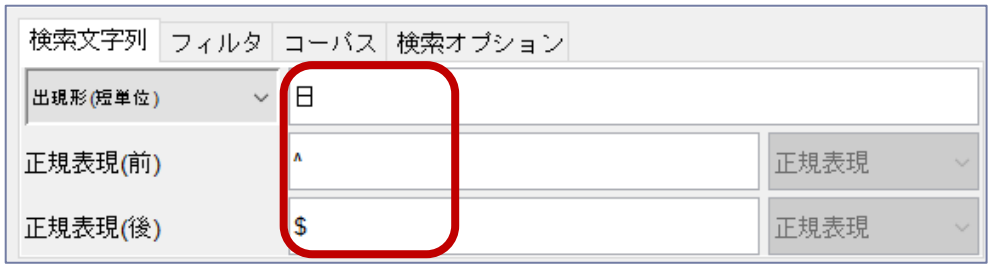
C) 末尾が「日」の単語

正規表現の「日$」と同義
(末尾の文字が「日」)



D) 単語「日」のみ

正規表現の「^日$」と
同義

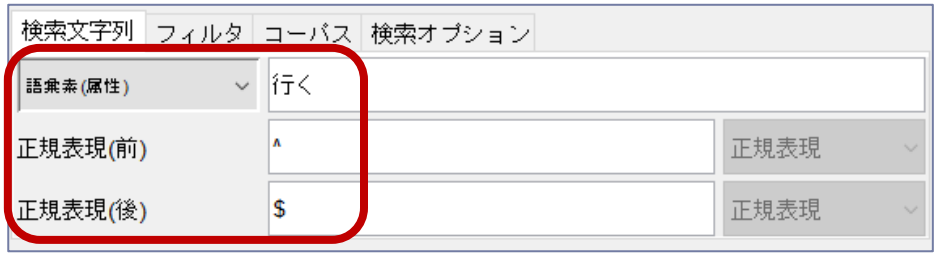


各種応用例

名大

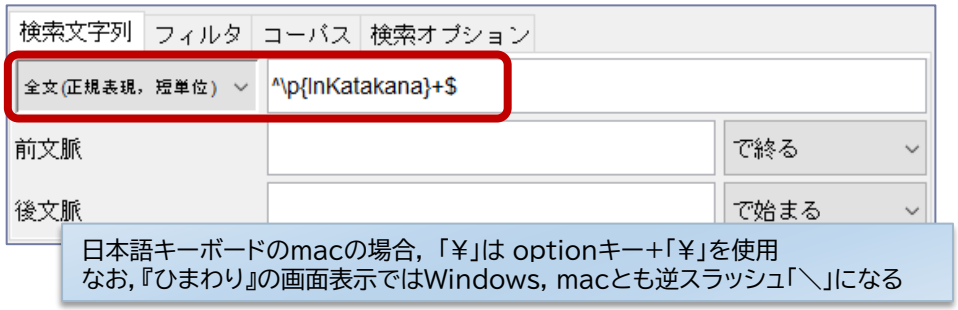
A) 共起語の集計(「～に行く」)

① 語彙素「行く」を検索
② 「語彙素－1」欄を「に」でフィルタ
③ 「語彙素－2」欄に「統計」機能を適用



B) 文字種の指定 (例:カタカナ列の単語)

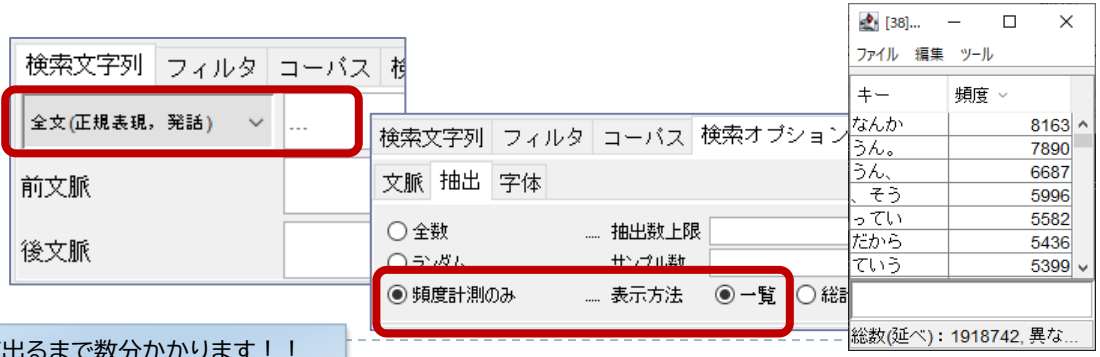
¥p{InHiragana} … ひらがな
¥p{InKatakana} … カタカナ
¥p{InCJKUnifiedIdeographs} … 漢字



日本語キーボードのmacの場合，「¥」は optionキー+「¥」を使用
なお，『ひまわり』の画面表示ではWindows，macとも逆スラッシュ「＼」になる

C) 文字 n-gram

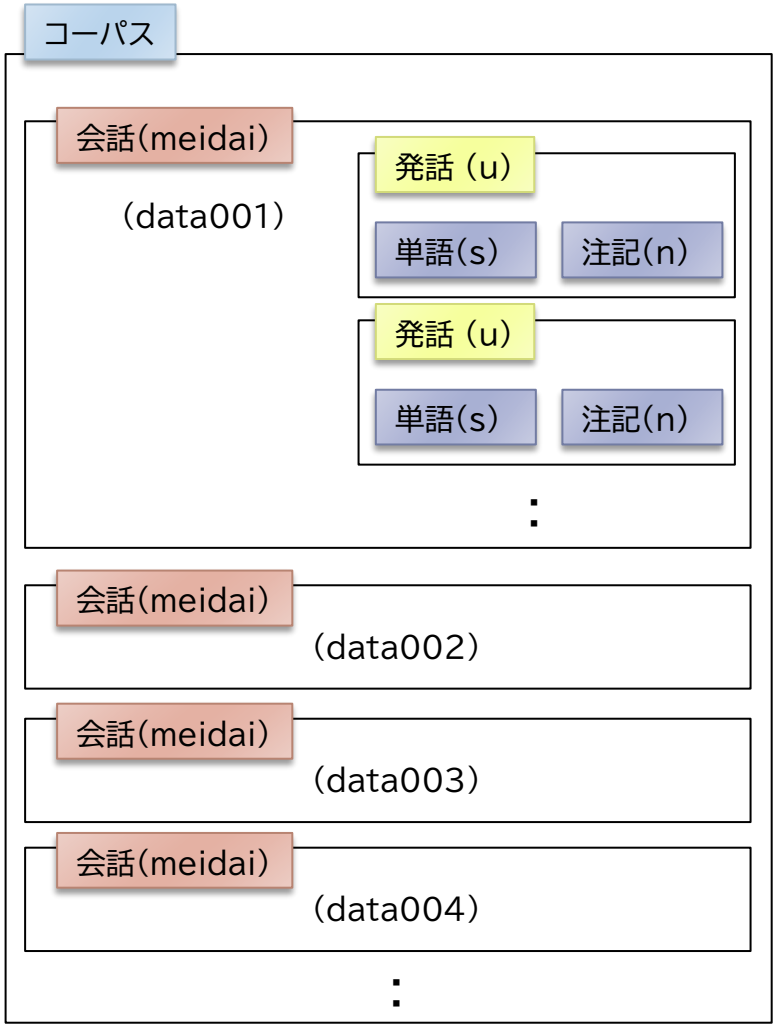
① 何か検索して，「キー」列のセルを選択 (どこでもよい)
② 検索オプションを設定 ([抽出]・頻度計測のみ・一覧)
③ 正規表現「...」を検索 (3-gramの場合)



結果が出るまで数分かかります！！

コーパスの構造とタグの集計

コーパスの構造と検索(名大会話コーパス)

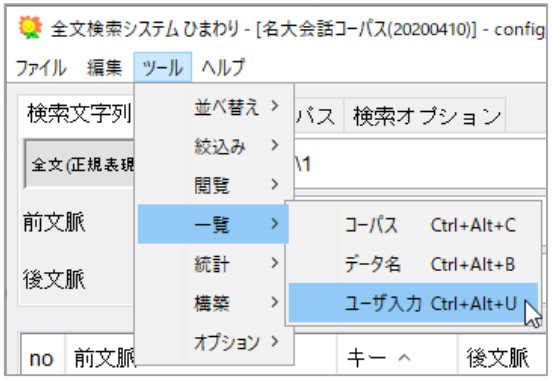


- 会話データ(meidai)
  - 発話(u)の集まり
  - 収録日, データ名などの属性
- 発話(u)
  - 単語(s)と注記(n)の集まり
  - 発話の名前, 性別などの属性
  - 発話末尾には, 発話末を表す, 長さ0のダミー単語が挿入されている
- 単語(s)
  - 短単位で記述
  - 品詞などの属性
- 注記(n)
  - ＜笑い＞などの雑多な情報
  - 本文ではなく, 属性として記述

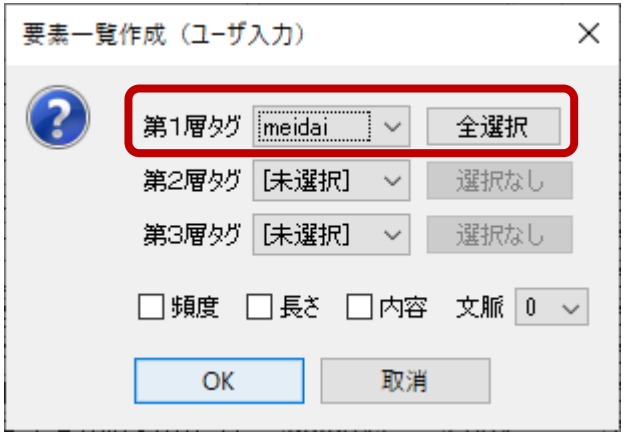
タグの集計

名大

- [ツール]⇒ [一覧] ⇒ ユーザ入力
  - タグの階層構造を利用しつつ,
    タグの数や属性を集計する
  - 例1： 発話数
  - 例2： 単語一覧



- meidaiタグの情報

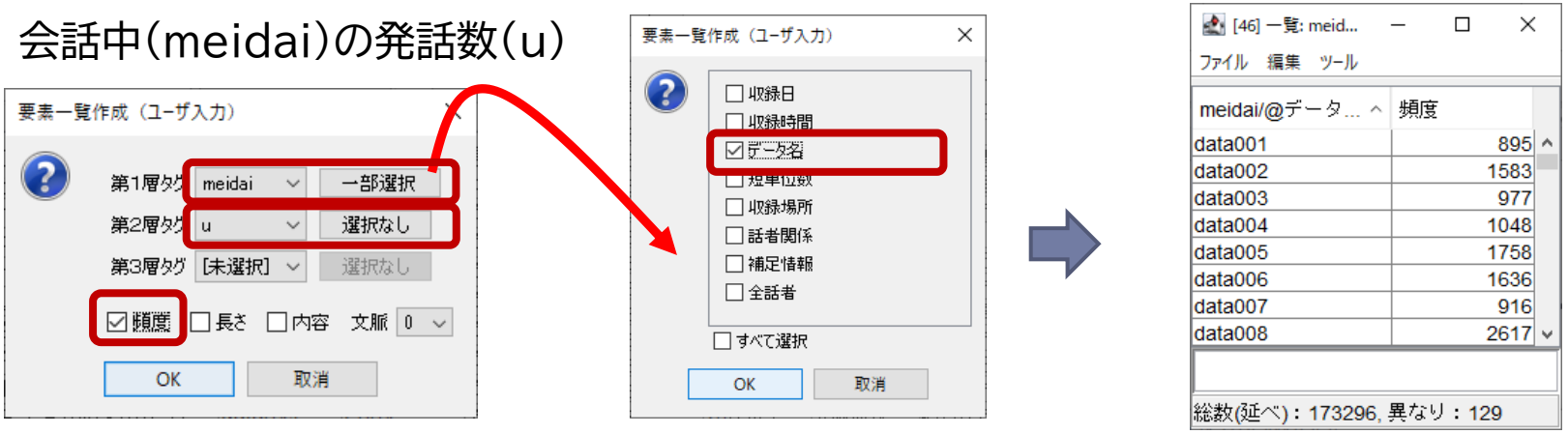


- 頻度： 指定したタグの頻度
- 長さ： マークアップされている文字列の長さ
  （空白やXMLタグは除く）
- 内容： マークアップされている文字列
- 文脈： 後続する同種のタグの属性をn個表示
  （単語の場合, n+1 gramになる）

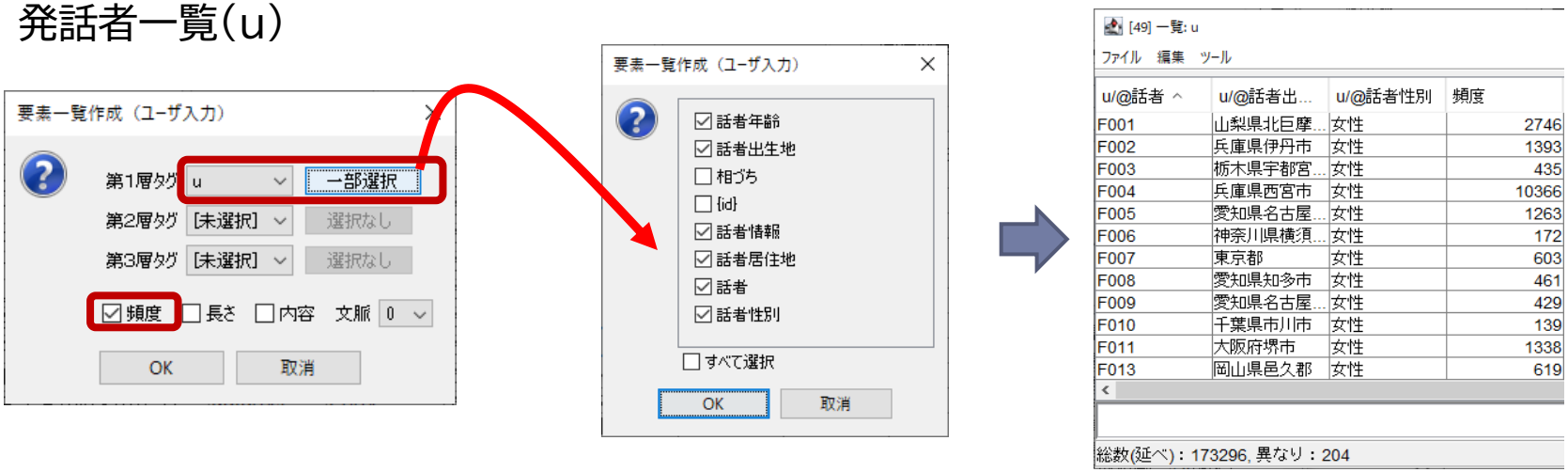
タグの集計の例1

名大

- 会話中(meidai)の発話数(u)



- 発話者一覧(u)

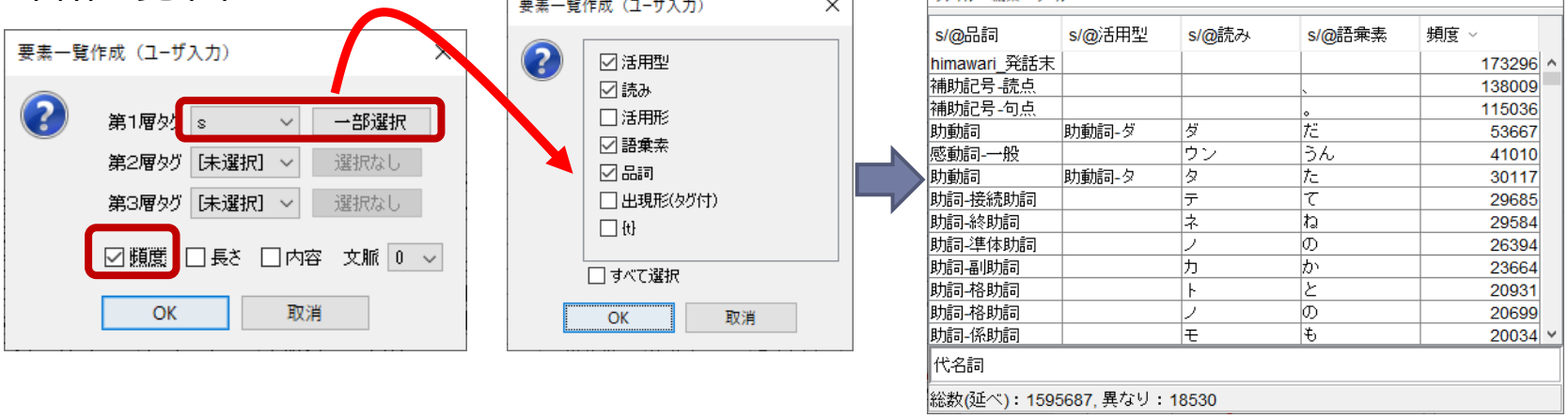


集計結果をさらに集計することも可能
(例：男女別の発話数)

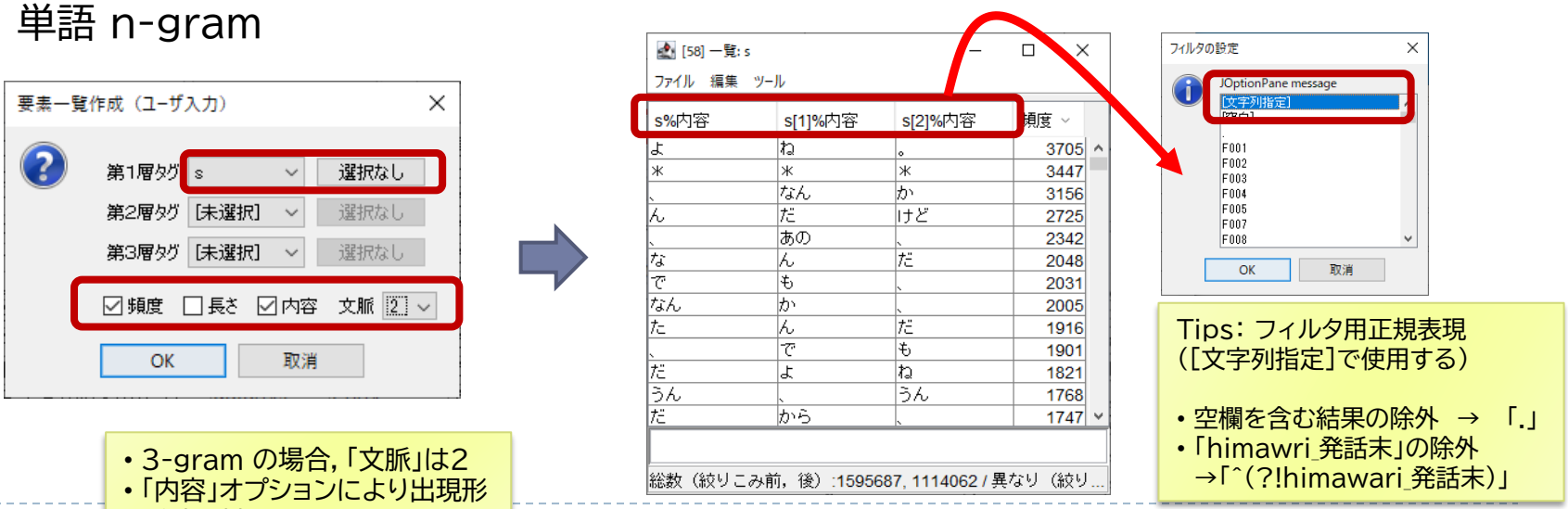
タグの集計の例2

名大

単語一覧(s)



単語 n-gram

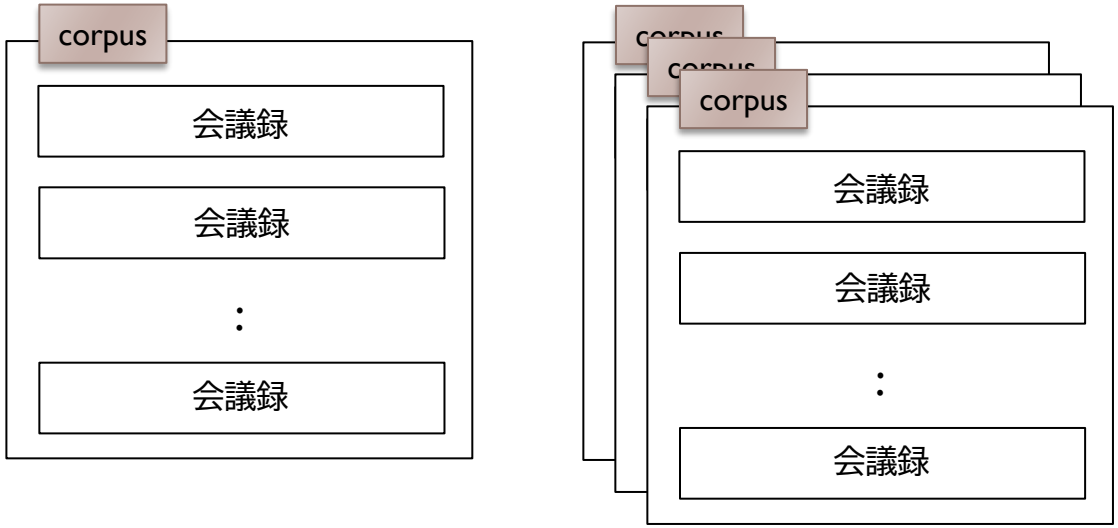


- 3-gram の場合,「文脈」は2
- 「内容」オプションにより出現形が計測対象となる

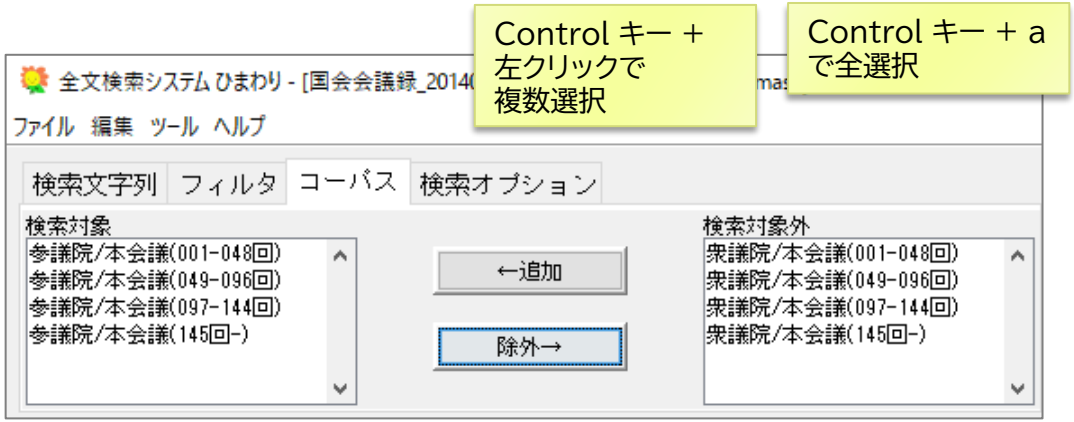
Tips: フィルタ用正規表現
([文字列指定]で使用する)

- 空欄を含む結果の除外 → 「.」
- 「himawri_発話末」の除外
  →「^(?!himawari_発話末)」

コーパスの構造と検索(国会会議録)



- 計8個のサブコーパス
  - 参議院/本会議
    - 001-048回
    - 049-096回
    - 097-144回
    - 145回-
  - 衆議院/本会議
    - 001-048回
    - 049-096回
    - 097-144回
    - 145回-



- サブコーパスにしている理由
  - ファイルサイズ・検索速度などシステム上の制約
  - テキストの品質
    (145回より古い会議はOCRによるテキスト入力)

コーパスの構造と検索(国会会議録)

- 会議録全体

衆議院, 本会議, 999回, 99号(minutes)

討議前部分

- 議事日程
- 会議案件
- 式辞など

最初の発言が始まるまでの内容

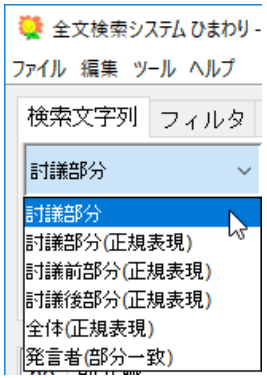
討議部分

最初の発言から最後の発言までの内容

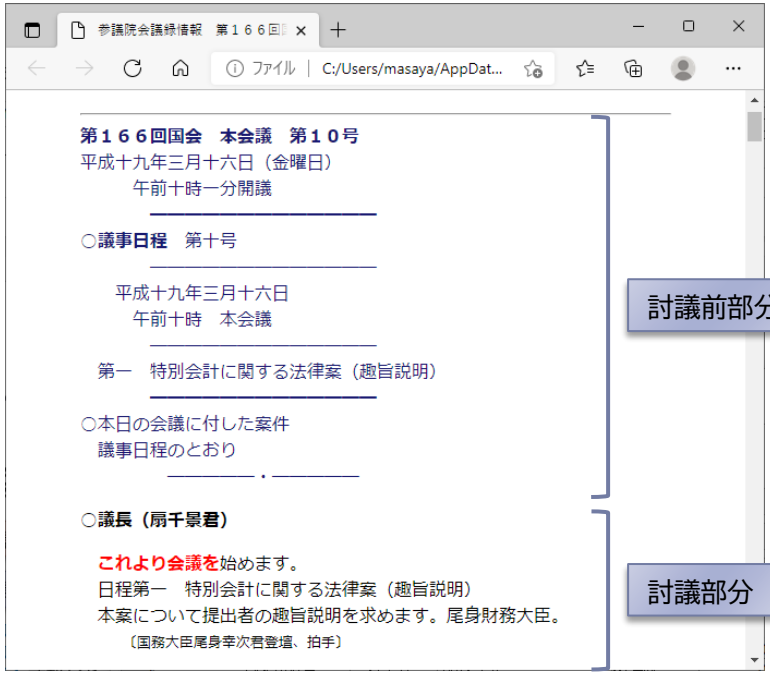
討議後部分

- 終了時間
- 出席者
- など

最後の発言以降の内容



- 通常の全文検索は, 討議部分のみ
- 正規表現検索では, 部分ごとの検索が可能
- 発言者検索では, 結果の発言全体が「キー」欄に入る



コーパスの構造と検索(国会会議録)

- 討論部分

赤下線, 赤枠部分は, 付属要素として, 本文の全文検索から除外

討議部(minutes)

発話(utterance)

○議長(国会太郎君)
本件を採決いたします。(「異議なし」と叫ぶ者あり)
本件を委員長報告のとおり承認するに賛成の皆さんの起立を求めます。

〔賛成者起立〕
―――◇―――
日程第一　平和的目的のための地下の探査

発話(utterance)

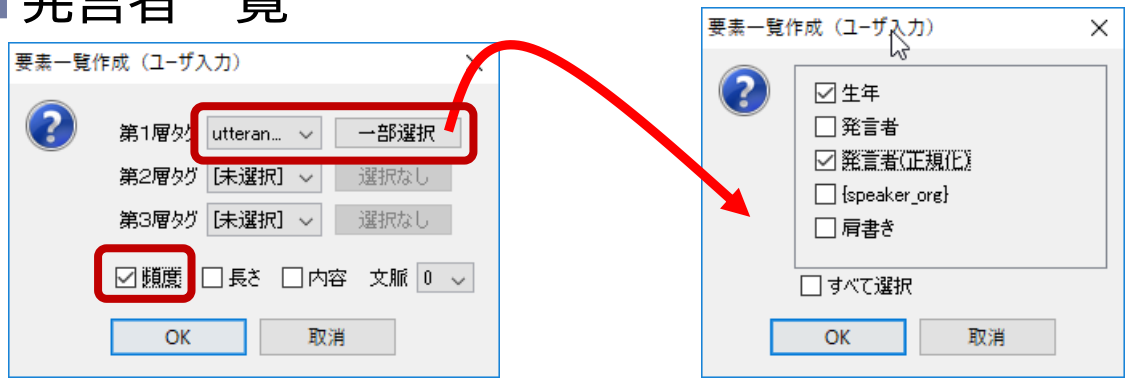
○議長(国会太郎君)
本日は, 平和目的のための地下の探査に関する法律案(内閣提出, 衆議院送付)を議題といたします。まず, 委員長の報告を求めます。司法委員長山田三郎君。

- 発言者の「国会太郎」を全文検索してもマッチしない
- ブラウザ表示では, 付属情報も含めて表示される
- 付属要素の認識は機械的に行っているため, 間違いも含む

タグの集計(minutes, utterance)

国会

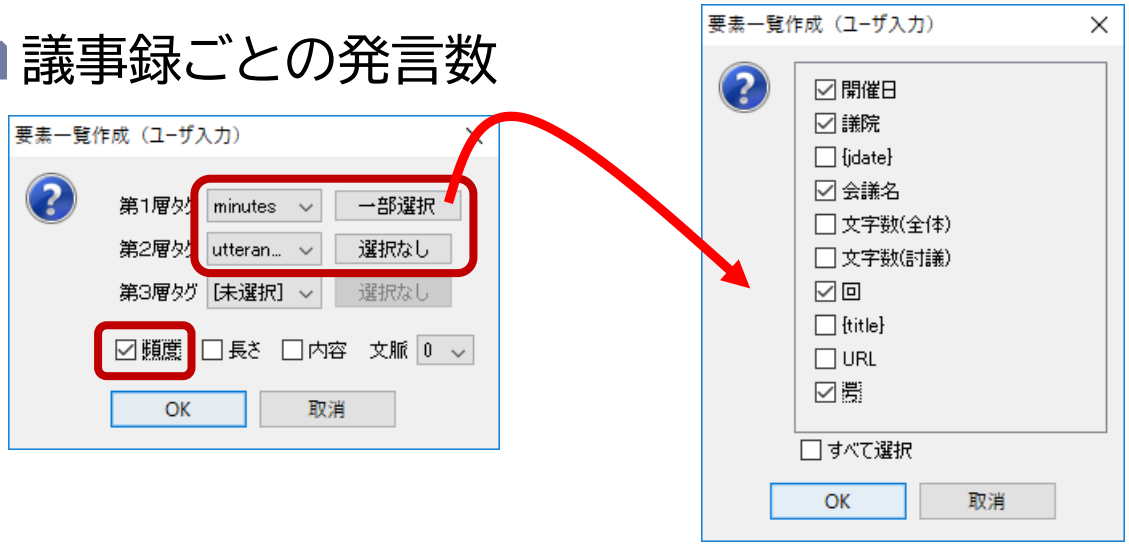
❑ 発言者一覧



発言者名の正規化例
佐藤榮作　⇒　佐藤栄作
（常用漢字の旧字体⇒新字体）

[ツール]⇒ [一覧] ⇒ ユーザ入力

❑ 議事録ごとの発言数



応用例:表記の経年変化

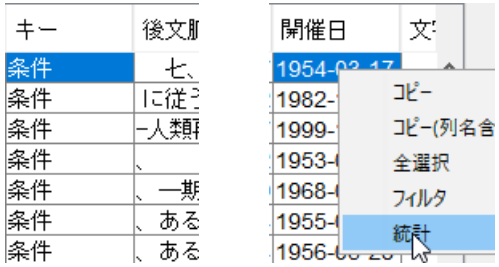
国会

1 「条件」と「條件」の検索

「条件」と入力して,
「字体変換」ボタン

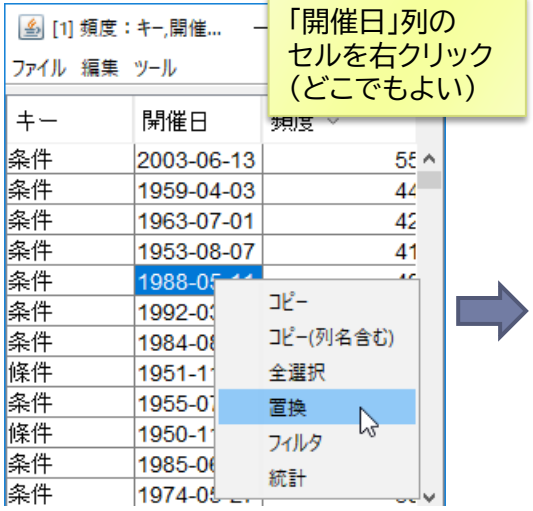
※常用漢字の旧字体に変換

2 「キー」「開催日」のセルを選択し,「統計」

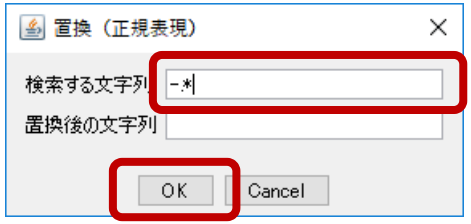


3 年月日を年に置換

「開催日」列の
セルを右クリック
(どこでもよい)

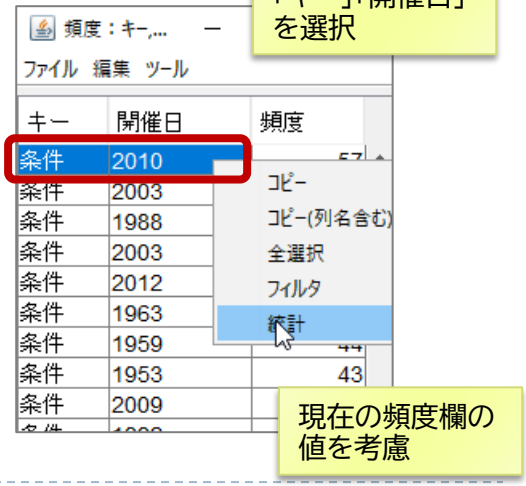


正規表現　-.*
(「-」以降の文字列を削除)



4 再集計

「キー」「開催日」
を選択



現在の頻度欄の
値を考慮

テキストファイルのインポート

テキストファイルのインポート
―青空文庫のテキストデータを例に―

sample/miyazawa_kenji/yamanashi.txt

やまなし
宮沢賢治

【テキスト中に現れる記号について】

《》:ルビ
(例)幻燈《げんとう》

［＃］:入力者注　主に外字の説明や、傍点の位置の指定
(例)［＃3字下げ］一、五月［＃「一、五月」は中見出し］

｜:ルビの付く文字列の始まりを特定する記号
(例)二｜疋《ひき》の

　小さな谷川の底を写した二枚の青い幻燈《げんとう》です。

［＃3字下げ］一、五月［＃「一、五月」は中見出し］

　二｜疋《ひき》の蟹《かに》の子供らが青じろい水の底で話していました。
『クラムボンはわらったよ。』
『クラムボンはかぷかぷわらったよ。』
『クラムボンは跳《は》ねてわらったよ。』
『クラムボンはかぷかぷわらったよ。』

青空文庫の独自タグ
(3種類)

生テキストをインポートする際,
青空文庫のタグは,『ひまわり』用
のタグに変換される(デフォルト)

テキストファイルのインポート
―青空文庫のテキストデータを例に―

sample/miyazawa_kenji/yamanashi.txt

やまなし
宮沢賢治

【テキスト中に現れる記号について】

《》:ルビ
(例)幻燈《げんとう》

[＃]:入力者注　主に外字の説明や、傍点の位置の指定
(例)[＃3字下げ]一、五月[＃「一、五月」は中見出し]

｜:ルビの付く文字列の始まりを特定する記号
(例)二｜疋《ひき》の

小さな谷川の底を写した二枚の青い幻燈《げんとう》です。

[＃3字下げ]一、五月[＃「一、五月」は中見出し]

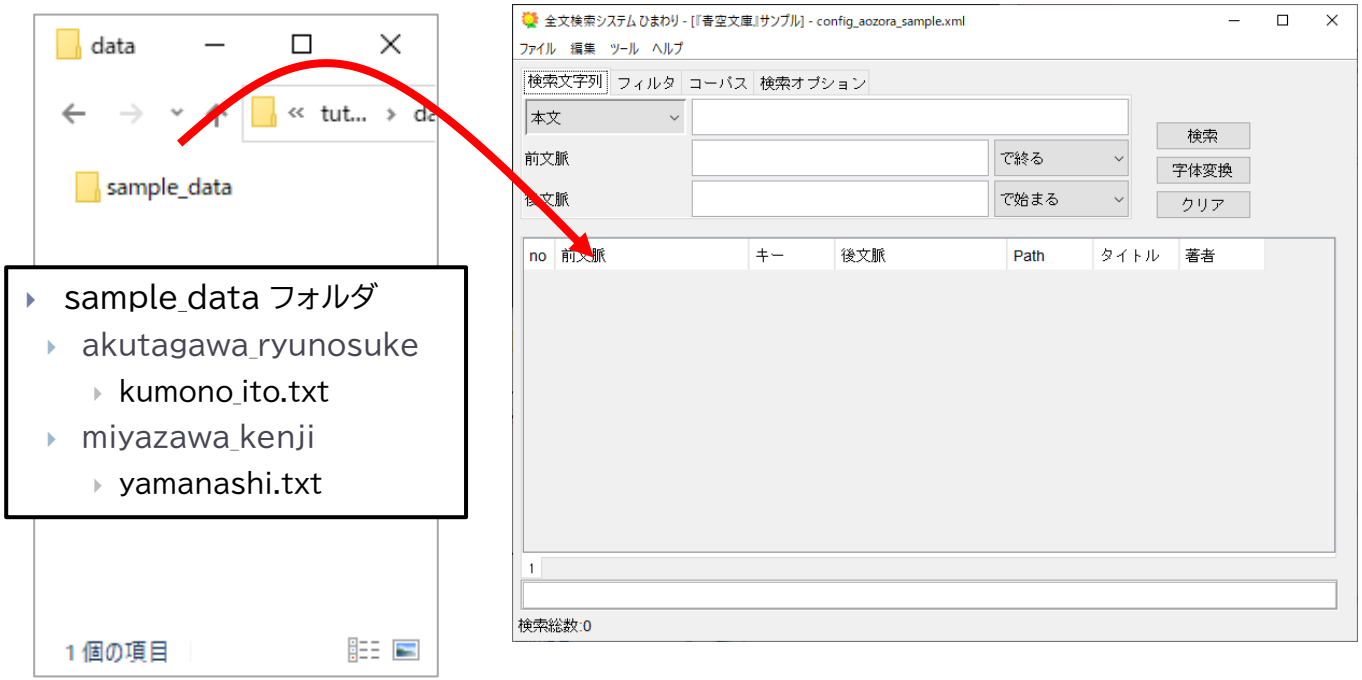
二｜疋《ひき》の蟹《かに》の子供らが青じろい水の底で話していました。
『クラムボンはわらったよ。』
『クラムボンはかぷかぷわらったよ。』
『クラムボンは跳《は》ねてわらったよ。』
『クラムボンはかぷかぷわらったよ。』

ルビ，注記は，本文とは区別され，全文検索の対象外となる

文字列の照合時は，タグは無視される
（例：「蟹の子供」にも照合）

インポートの実行

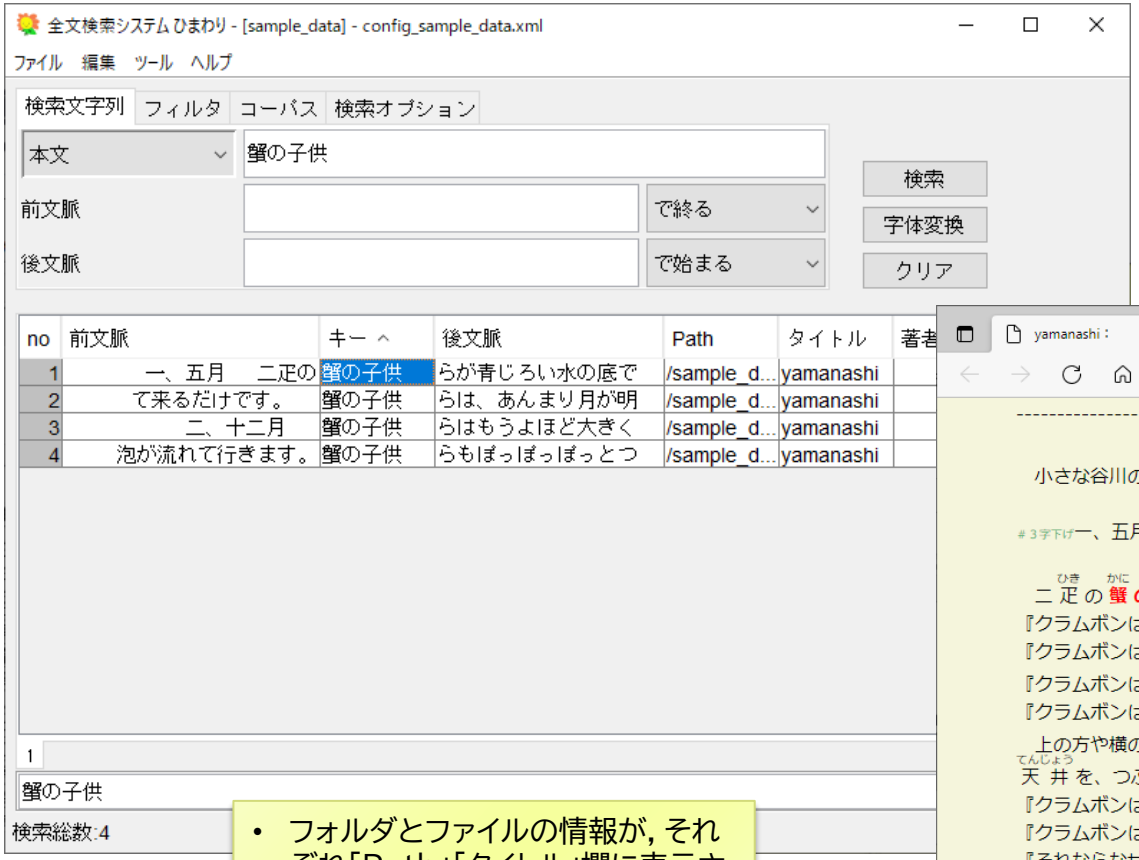
- sample_dataフォルダを，起動している『ひまわり』にドラッグ＆ドロップ



- フォルダの情報をインポート時に利用
  - フォルダ階層 ⇒ Path 欄
  - ファイル名 ⇒ タイトル欄
- ドロップしたフォルダ名がコーパス名になる

- HTML, XMLもインポート可能
- 文字コードは自動判別
- 詳細オプション（文字列変換，形態素解析など）

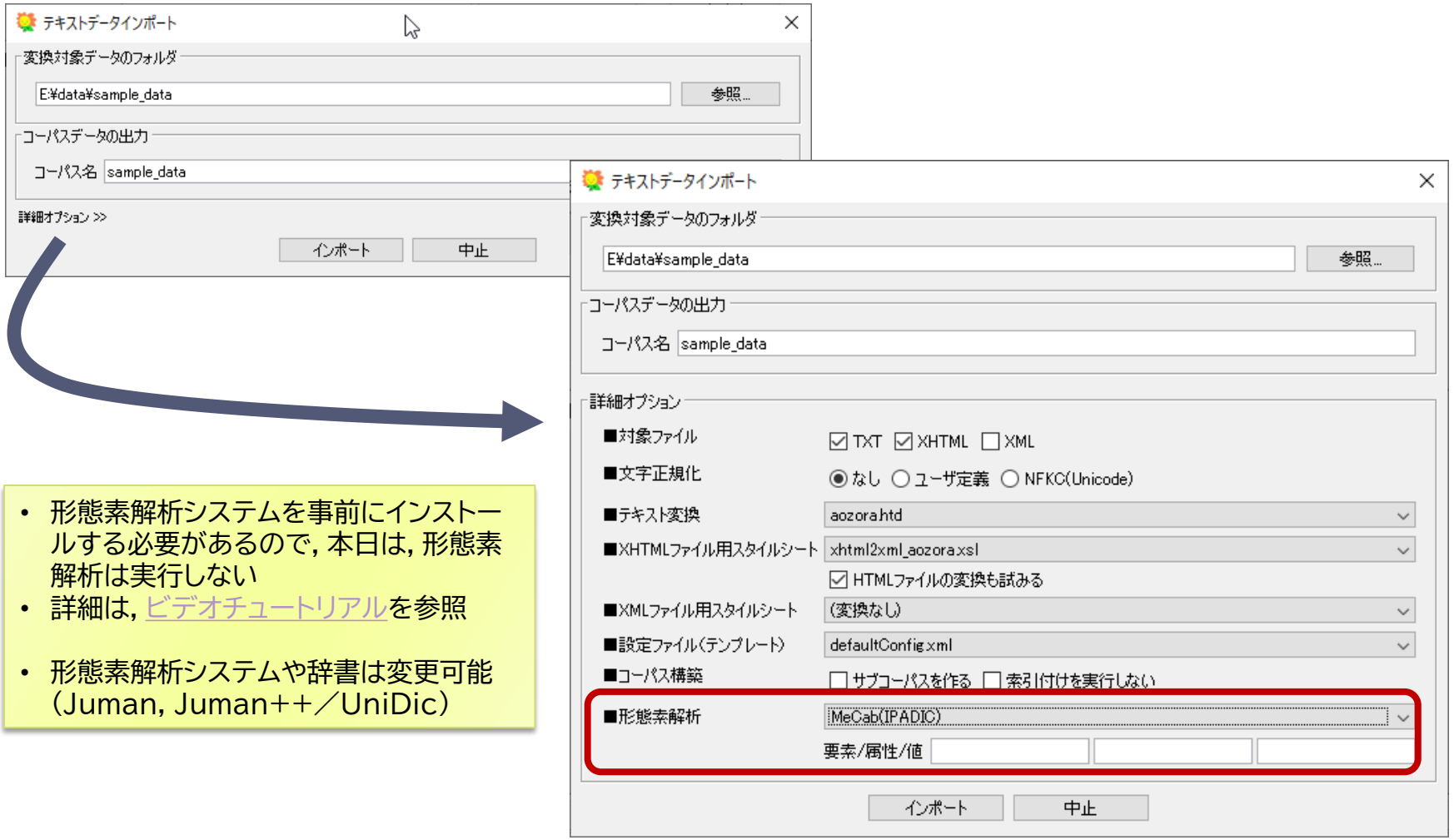
検索例



- フォルダとファイルの情報が，それぞれ「Path」「タイトル」欄に表示される
- 「著者」欄は空欄

- ルビ，注記が変換されていることに注目
- ルビ，注記自体はタグの属性として記述されているため，「本文」検索ではマッチしない

インポート時の形態素解析



- 形態素解析システムを事前にインストールする必要があるので，本日は，形態素解析は実行しない
- 詳細は，ビデオチュートリアルを参照

- 形態素解析システムや辞書は変更可能（Juman，Juman++／UniDic）

おわりに

- 全文検索システム『ひまわり』を使った，既存コーパスの検索と分析用基礎データの集計
  - 『ひまわり』の紹介と基本的な使い方
  - さまざまな検索と各種機能
  - コーパスの構造とタグの集計
  - テキストデータのインポート

- さらに詳しく知るには
  - 『ひまわり』ホームページ
  - 『ひまわり』用各種パッケージ
    - 『名大会話コーパス』パッケージ
    - 『国会会議録』パッケージ

参考資料

- 『ひまわり』関連
  - 利用者マニュアル
  - ビデオチュートリアル
  - 研究発表

- 正規表現関連
  - Java正規表現の使い方
  - Java Pattern クラス