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）

本日の内容

- 全文検索システム『ひまわり』を使って，既存のテキストデータを利用する方法を紹介
  - 『ひまわり』（ver.1.6ls04 = ver.1.6+実習資料）
  - 国会会議録（本会議）
  - 名大会話コーパス
  - 青空文庫（サンプル）

- 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
  - テキストデータのインポート

『ひまわり』とは

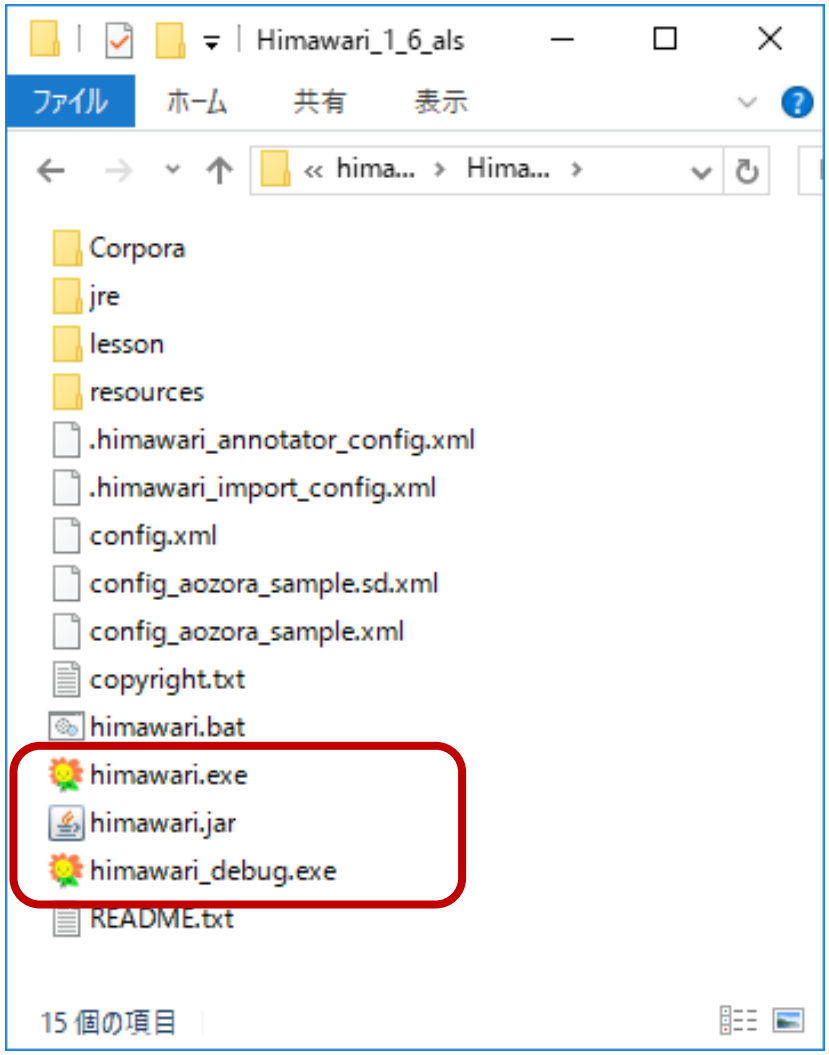
- 言語研究用の全文検索システム
  - 指定された文字列を網羅的に検索して，前後文脈付きで結果を表示します（コンコーダンサ）
  - 『太陽コーパス』（20世紀初頭の総合雑誌『太陽』）用の検索システムとして構築しました

- 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

Ver.1.6 ⇒　統計的な分析に必要なデータの収集支援機能を強化
（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』を起動する



himawari.exe

普段使うとき
（Windows 専用）
himawari.exe

himawari_
debug.exe

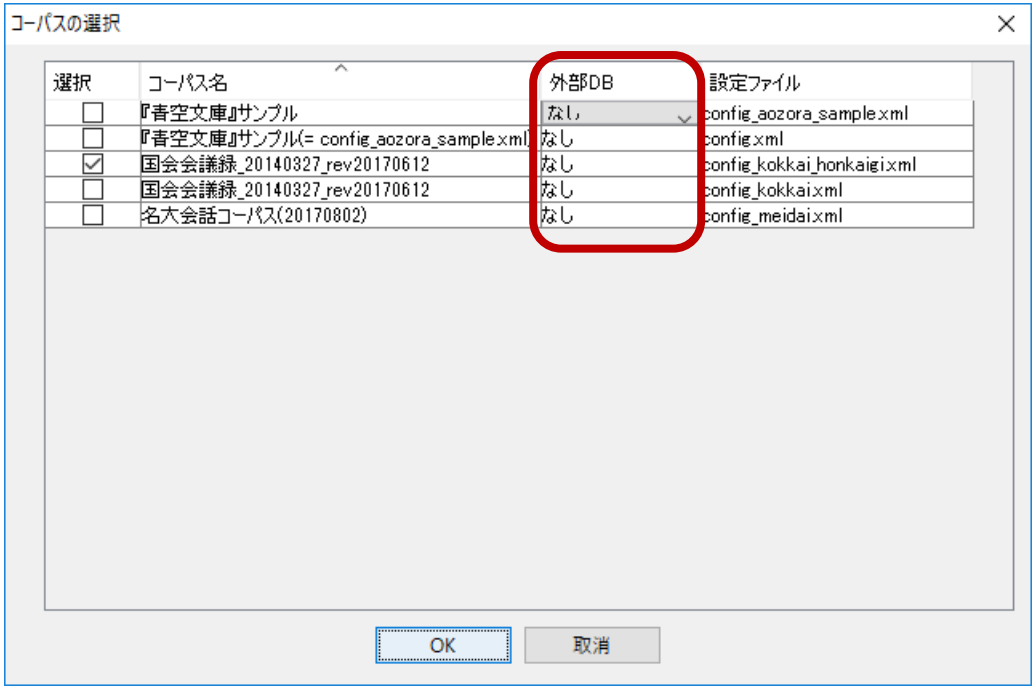
コーパスを作るとき
検索の途中経過を見たいとき
（Windows 専用）
himawari_debug.exe

himawari.jar

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

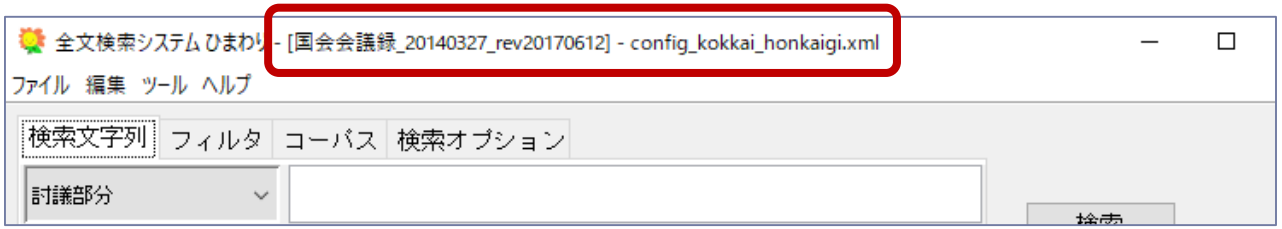
コーパスの選択

- [ファイル]⇒[コーパス選択]



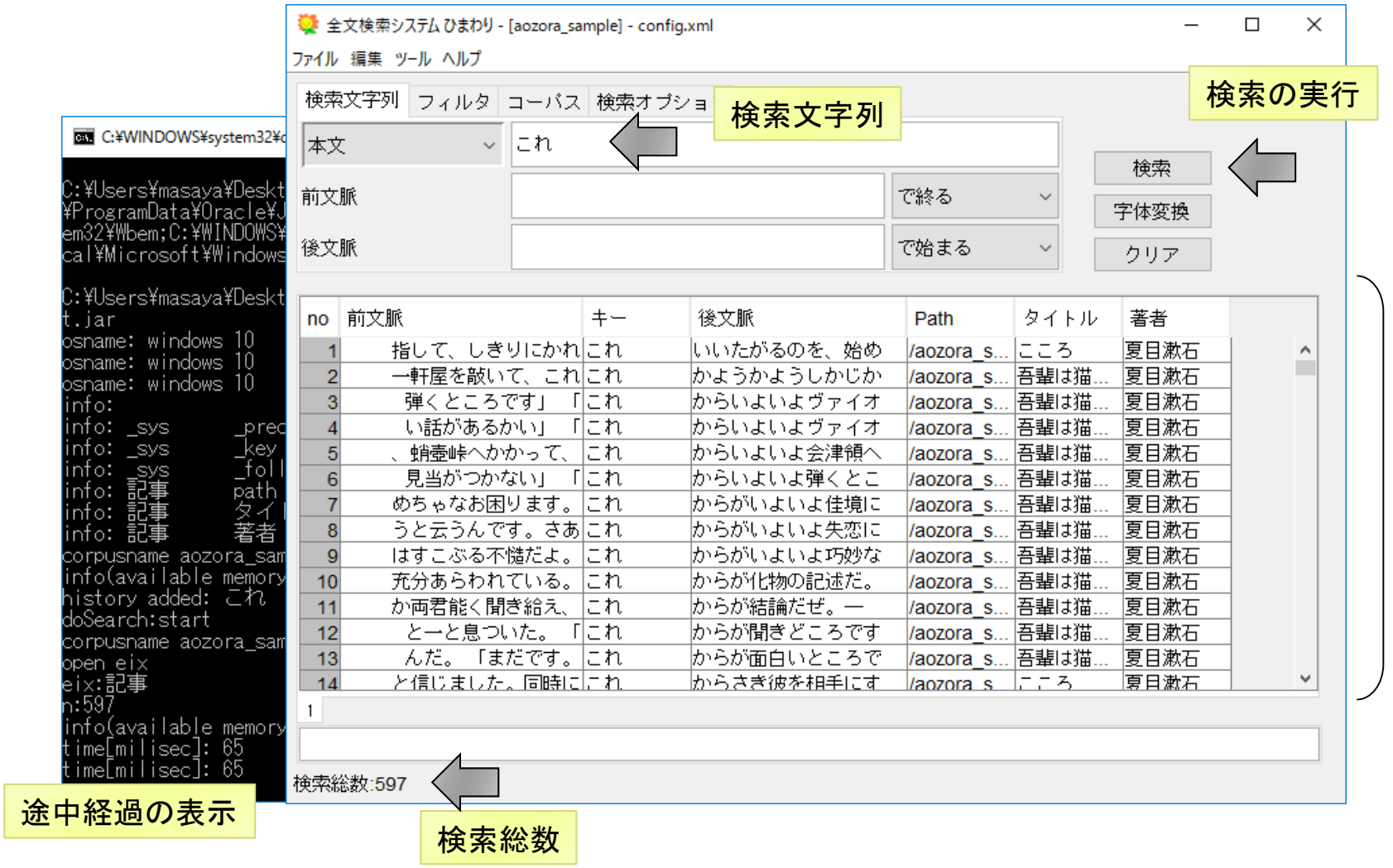
- 「外部DB」
  - コーパスファイルに直接記述していない付与データを格納
  - 『青空文庫』サンプルの場合は，形態素解析結果

- 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



検索結果

途中経過の表示

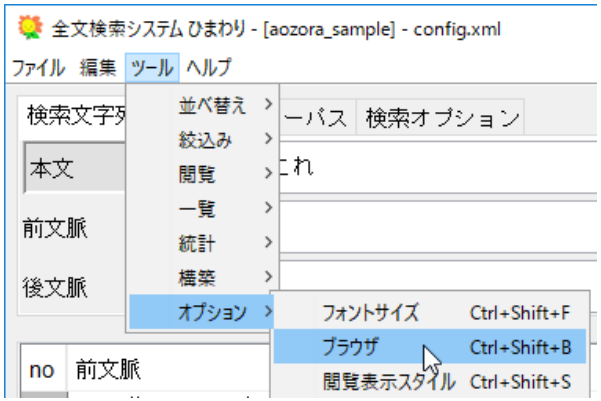
検索総数

ブラウザでの閲覧

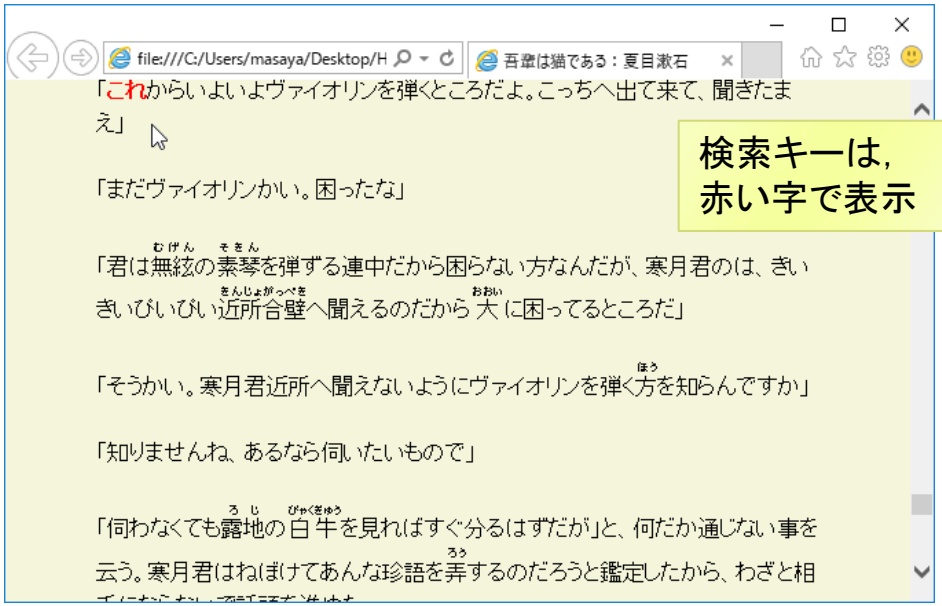
no	前文脈	キー	後文脈	Path
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...
3	弾くところです」「	これ	からいよいよヴァイオ	/aozora_s...
4	い話があるかい」「	これ	からいよいよヴァイオ	/aozora_s...
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aozora_s...
6	見当がつかない」「	これ	からいよいよ弾くとこ	/aozora_s...
7	めちゃなお困ります。	これ	からがいよいよ佳境に	/aozora_s...

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	はすこぶる不慥だよ。	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

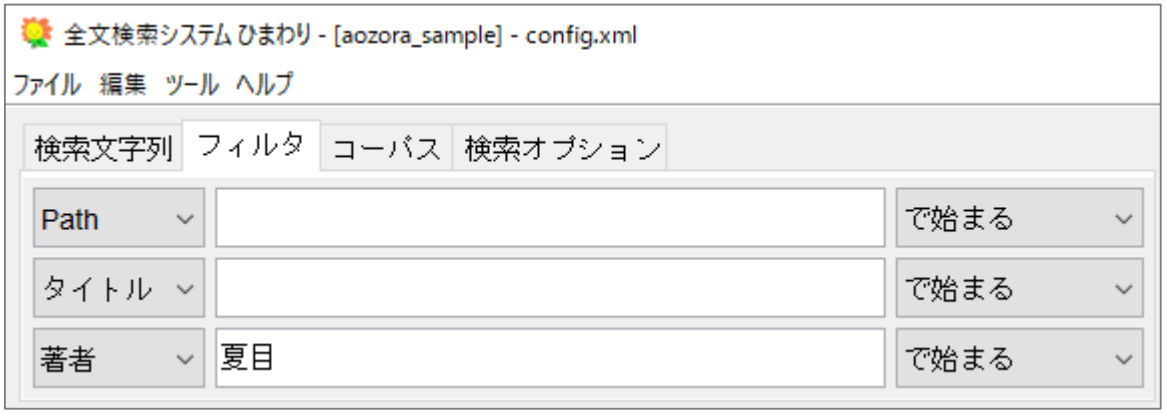
- 昇順
  - 列タイトルをクリック
- 降順
  - シフトキーを押しながら
    列タイトルをクリック

- 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:「話者」ごとに「後文脈」でソート
→ 「後文脈」「話者」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

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
	て、これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目
	です」「	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	い」「	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	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
	い」「	これ	からいよいよ弾くとこ	/aozora_s...	吾輩は猫...	夏目漱石
	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

[文字列指定]
[置換]
夏目漱石
芥川龍之介

絞り込みたい値を選択
⇒右クリック
⇒フィルタでもOK

検索結果の頻度集計

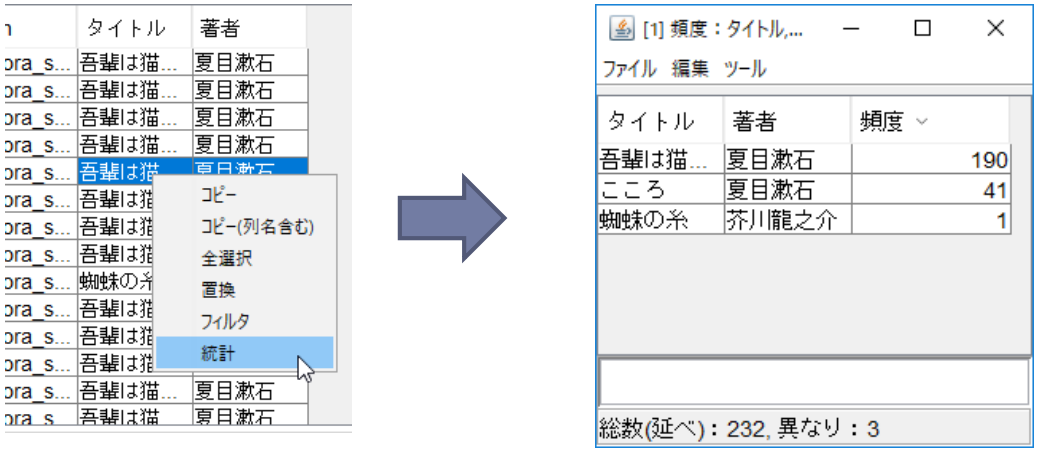
1. 集計したい列を選択

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

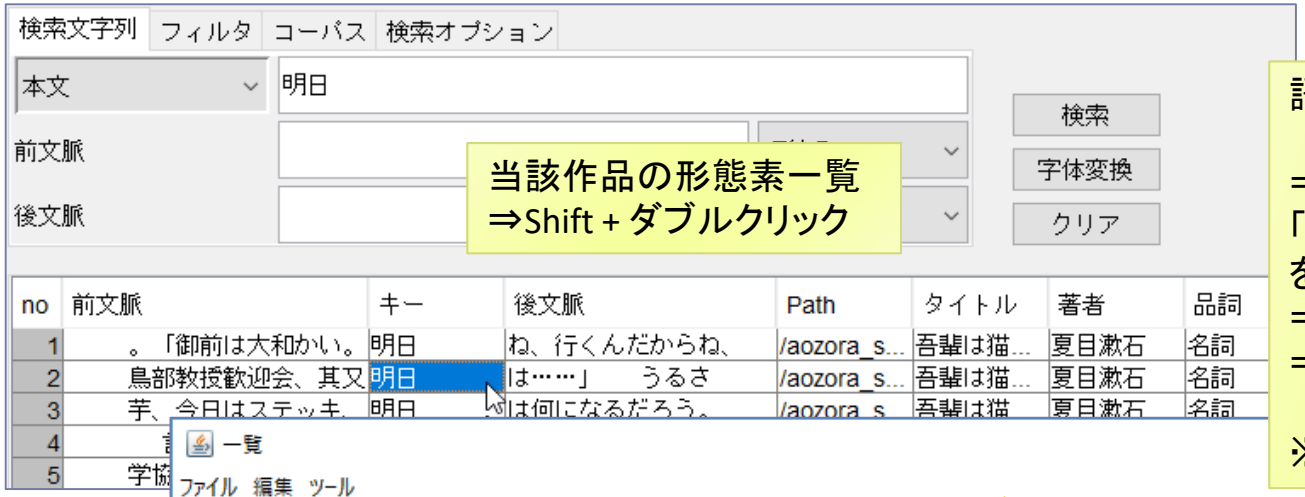
離れた列の選択
- WindowsはCtrlキー
- macOSはcommandキー

2. 右クリック⇒「統計」



形態素解析結果の閲覧

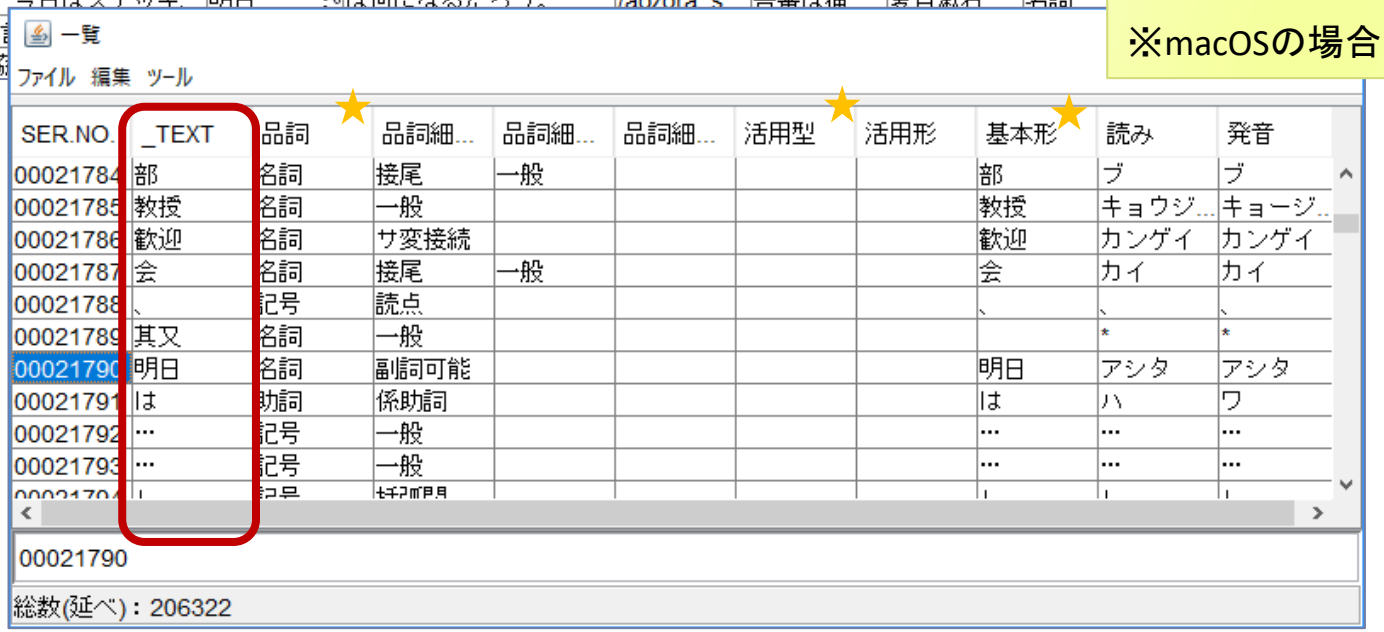
この機能は，
外部DB「sd」の資料のみ実行可能



語彙一覧（頻度付き）

⇒（どの行でもよい）
「品詞」「基本形」「活用型」
を Ctrl + クリックで選択
⇒右クリック
⇒統計

※macOSの場合は，command

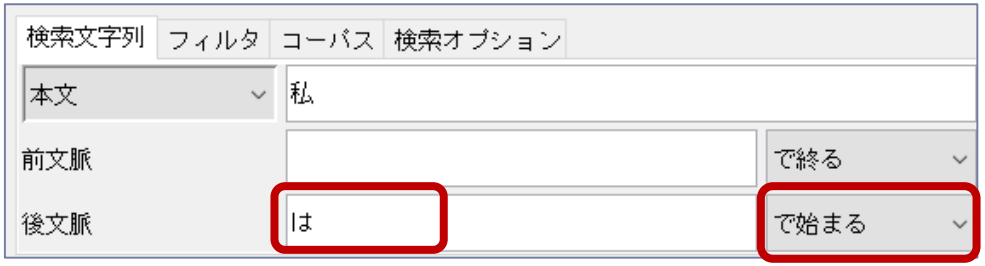


テキスト
進行方向

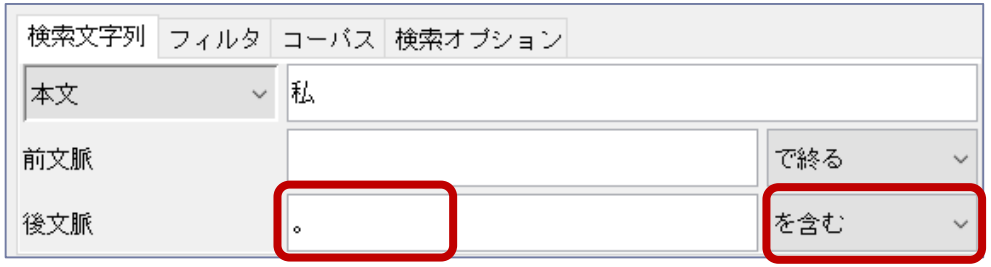
さまざまな検索と各種機能

前後文脈の制限

A) 後文脈を
「は」で「始まる」に制限

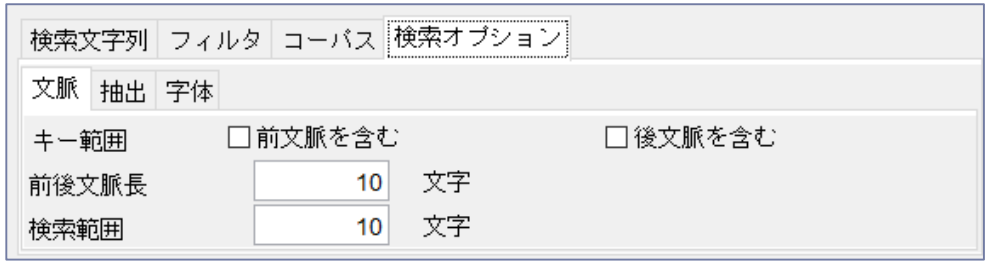


B) 後文脈を
「。」を「含む」に制限
（「正規表現」と同一）



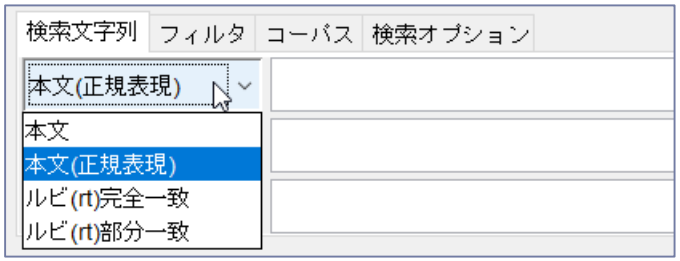
- 検索オプション（文脈）
  - キー範囲： 一致した前後文脈をキーに統合
  - 前後文脈長： 表示用
  - 検索範囲： 検索用

基本的に同じ値にする



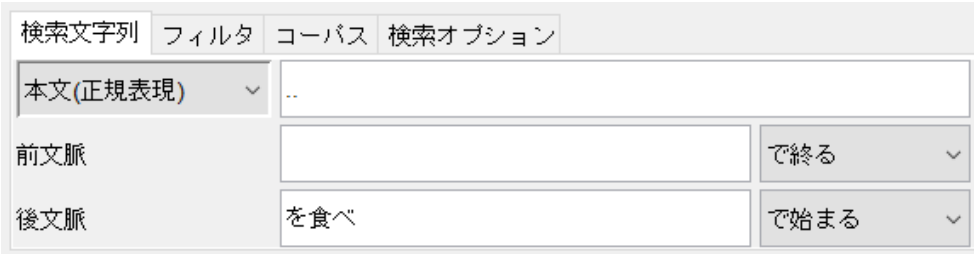
本文(正規表現)

- 本文検索に正規表現が利用可能
- 検索速度は低速
- マッチングの範囲は，行内(資料に依存)



A) 「～を食べ」

. (ピリオド) … 任意の1文字



B) 雑多な例

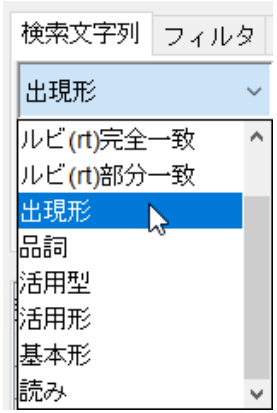
- 私[がをにへ]
- ようやく[^。！]*。
- 私が.+?を.+?のは
  (改良版) 私が.{1,5}を. {1,5}のは
- (..)¥1

[がをにへ]	… 「が」「を」「に」「へ」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)
{1,5}	… 直前要素の1～5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

単語(外部DB)での検索(1)

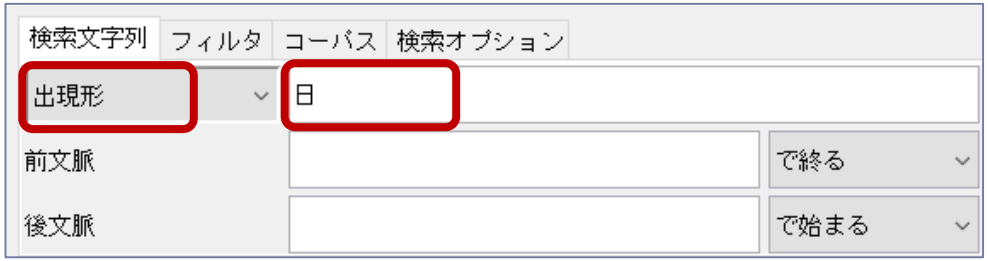
青空文庫サンプル
(外部DBあり)を対象に
config_aozora_sample.sd.xml

- 単語単位で正規表現検索
  - 単位をまたいだ検索はできない
  - 青空文庫サンプルは，MeCab (ver.0.996)で解析
  - 名大会話コーパスは外部DBではない
    ⇒検索方法異なるので注意



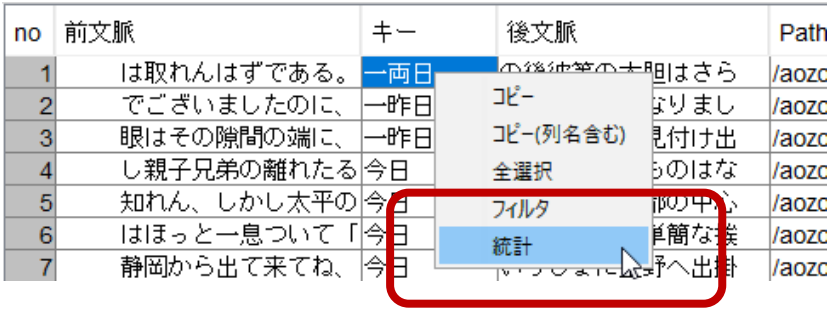
A) 「日」を含む単語

「基本形ー1」「基本形1」欄は，それぞれ前後の単語の基本形



B) 「キー」欄(出現形)の一覧を求める

「キー」欄のどれかを選択
⇒右クリック
⇒統計



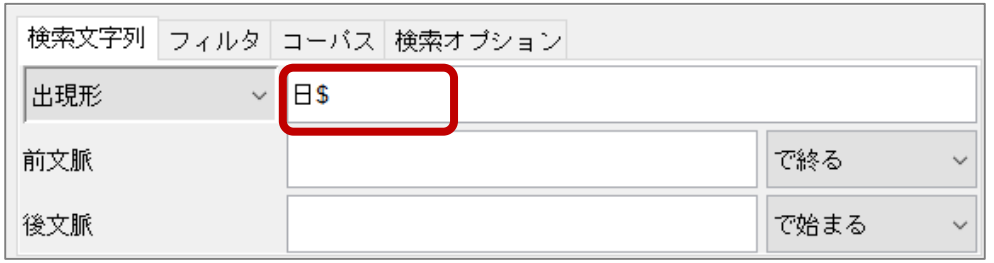
単語（外部DB）での検索（2）

C) 先頭が「日」の単語

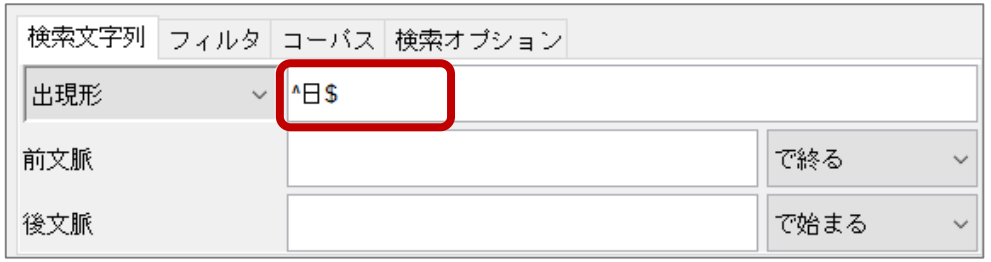
正規表現の「^」
（文字列の先頭）

D) 末尾が「日」の単語

正規表現の「$」
（文字列の末尾）

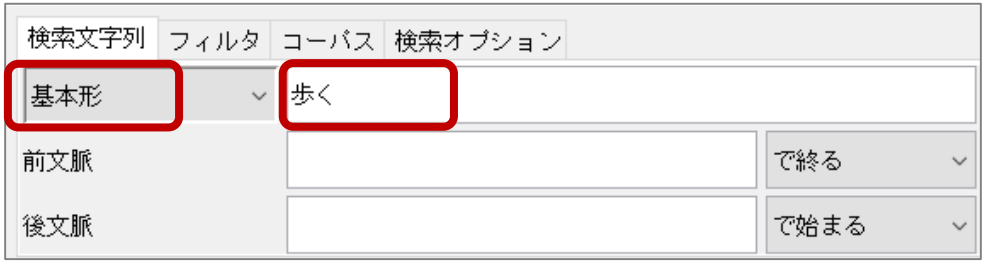


E) 単語「日」のみ



F) 活用語の基本形

すべての語形を
一括して検索

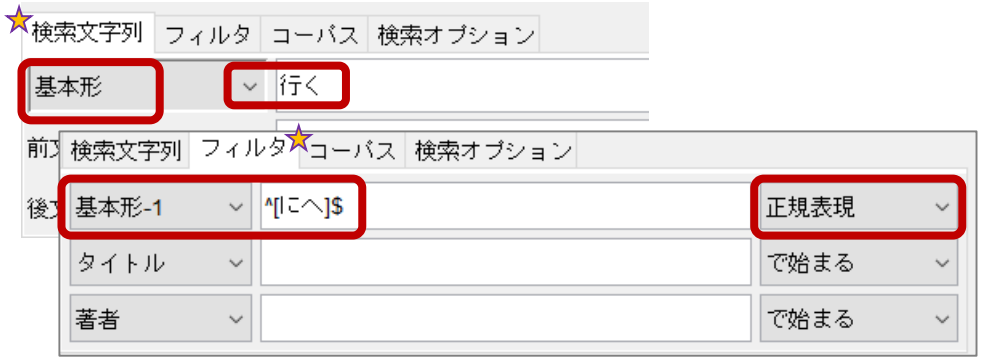


各種応用例

青空文庫サンプル
（外部DBあり）を対象に
config_aozora_sample.sd.xml

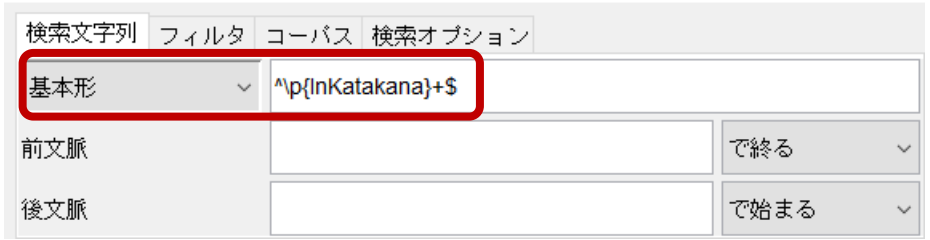
A) 共起語の集計
（「～へ行く」,「～に行く」）

「基本形－2」欄に対して，
「統計」機能を適用



B) 文字種の指定
（例：カタカナ列の単語）

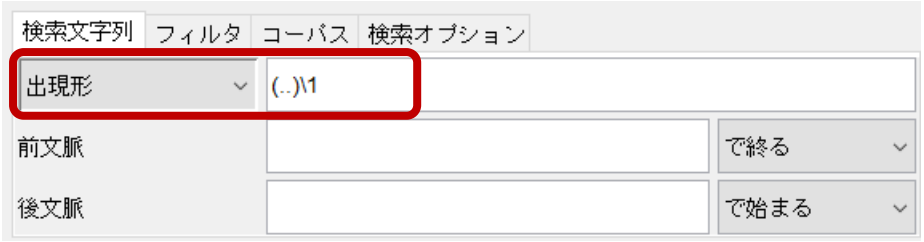
¥p{InHiragana}　…　ひらがな
¥p{InKatakana}　…　カタカナ
¥p{InCJKUnifiedIdeographs}　…　漢字
+　　…　直線の文字の繰り返し



macOSの場合，　「¥」は上図のように逆スラッシュ（optionキー+「¥」）を使用

C) 繰り返し表現の抽出

()　…　範囲を定義
¥n　…　n番目の範囲
(. .)¥1　…　1番目の範囲の繰り返し



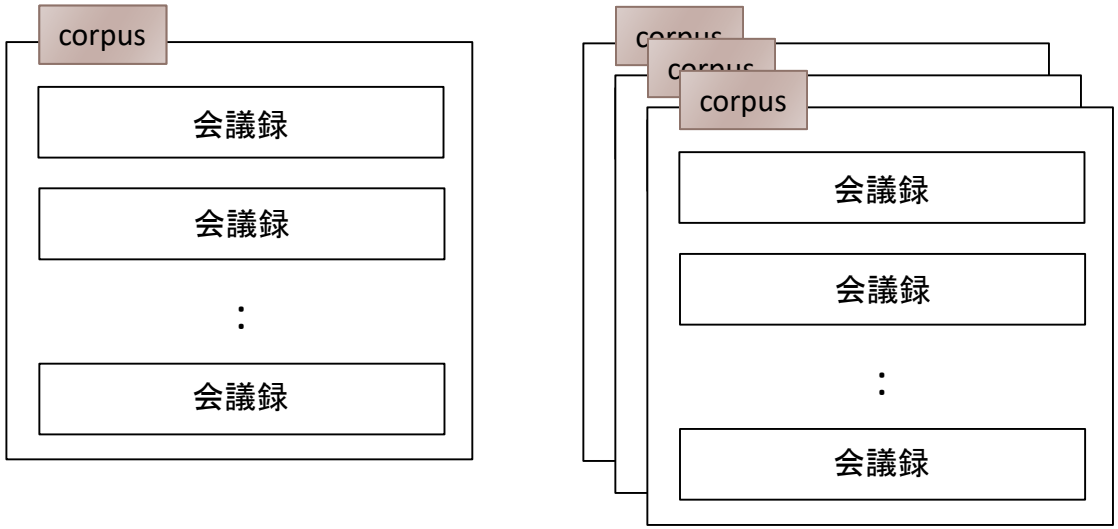
練習

- 「夏の日差し」のような，「～の～」型の名詞句を検索し，それぞれの出現頻度を求める

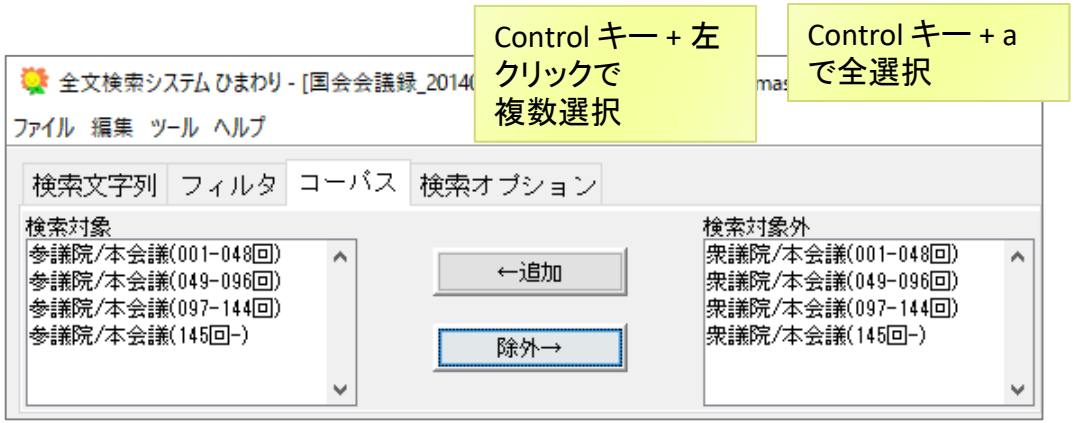
コーパスの構造と検索

config_kokkai_honkaigi.xml

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

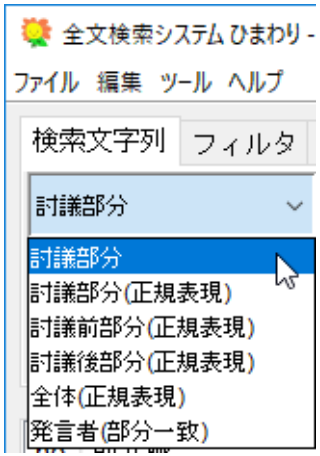
討議部

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



- 正規表現検索は，通常の検索よりも低速
- 発言者検索では，結果の発言全体が「キー」欄に入る

コーパス本体を見たい場合
Corpora/Kokkai/honkaigi/corpus_sangiin_hon0x.xml

ブラウザで閲覧した記事
Corpora/Kokkai/xslt/__searched_tmp.xml

※「秀丸」などのテキストエディタを利用のこと

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

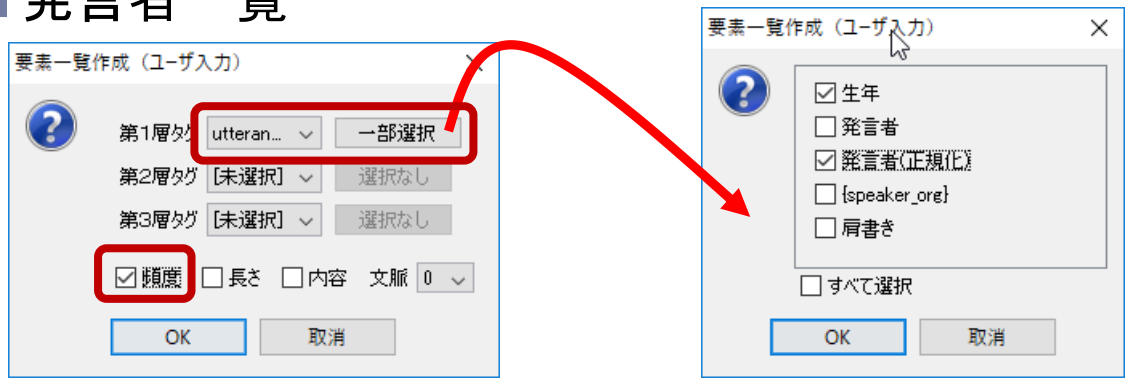
発話（utterance）

○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）

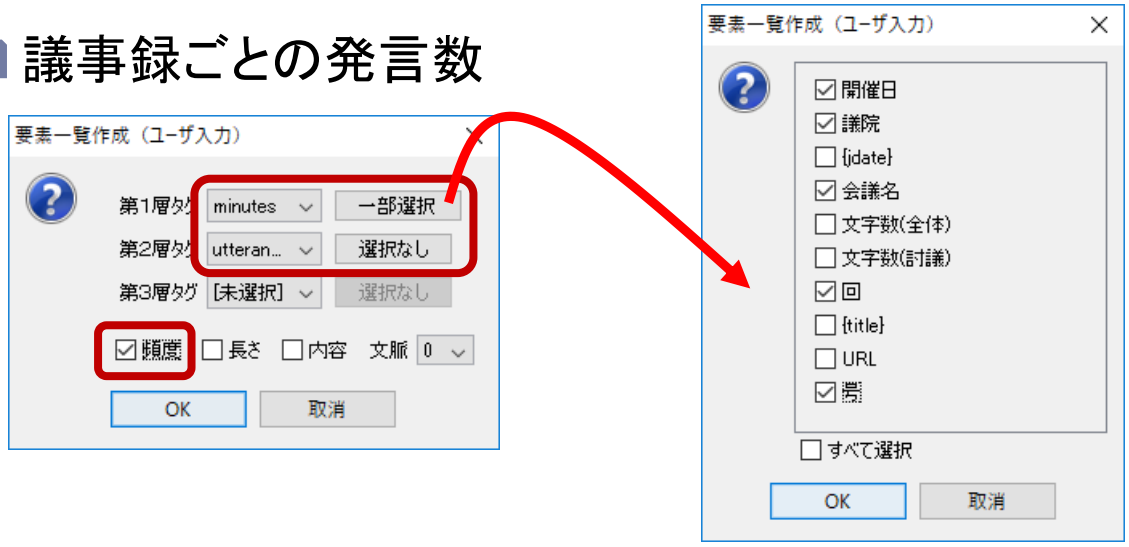
❑ 発言者一覧



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



- 「頻度」「長さ」「内容」「文脈」は，最下位層のタグが対象
- 長さ： マークアップされている文字列の長さ
- 内容： マークアップされている文字列
- 文脈： 後続する同種の要素の属性をn個表示（単語の場合n+1 gramになる）

応用例：表記の経年変化

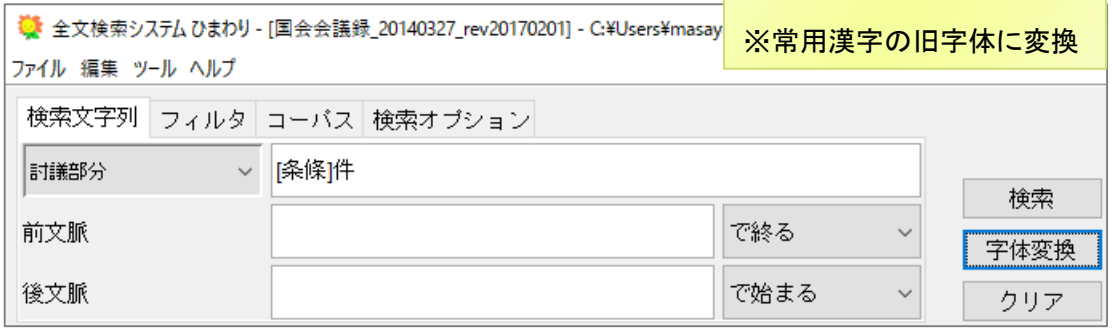
国会会議録
config_kokkai_honkaigi.xml

参議院/本会議(001-048回) を対象に

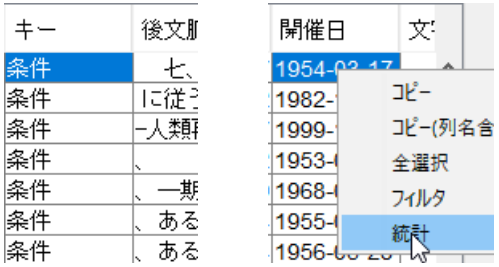
1 「条件」と「條件」の検索

「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

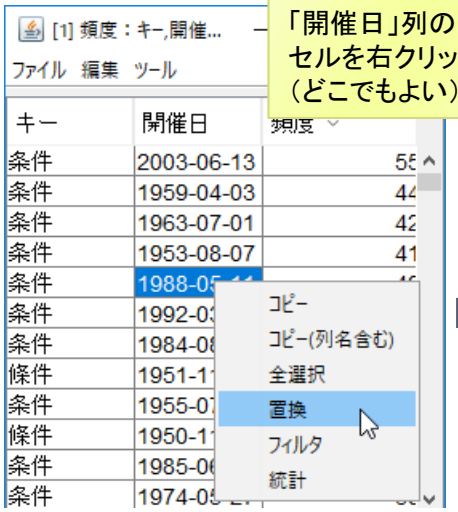


2 「キー」「開催日」のセルを選択し，「統計」

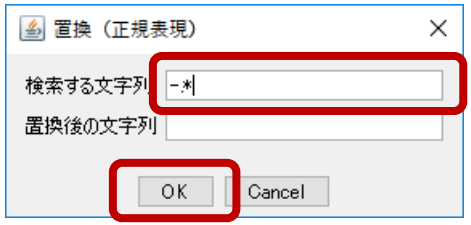


3 年月日を年に置換

「開催日」列の
セルを右クリック
（どこでもよい）

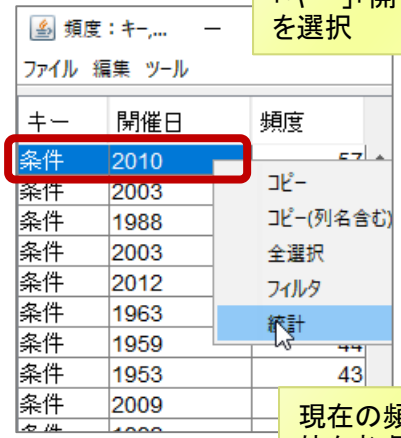


正規表現　-.*



4 再集計

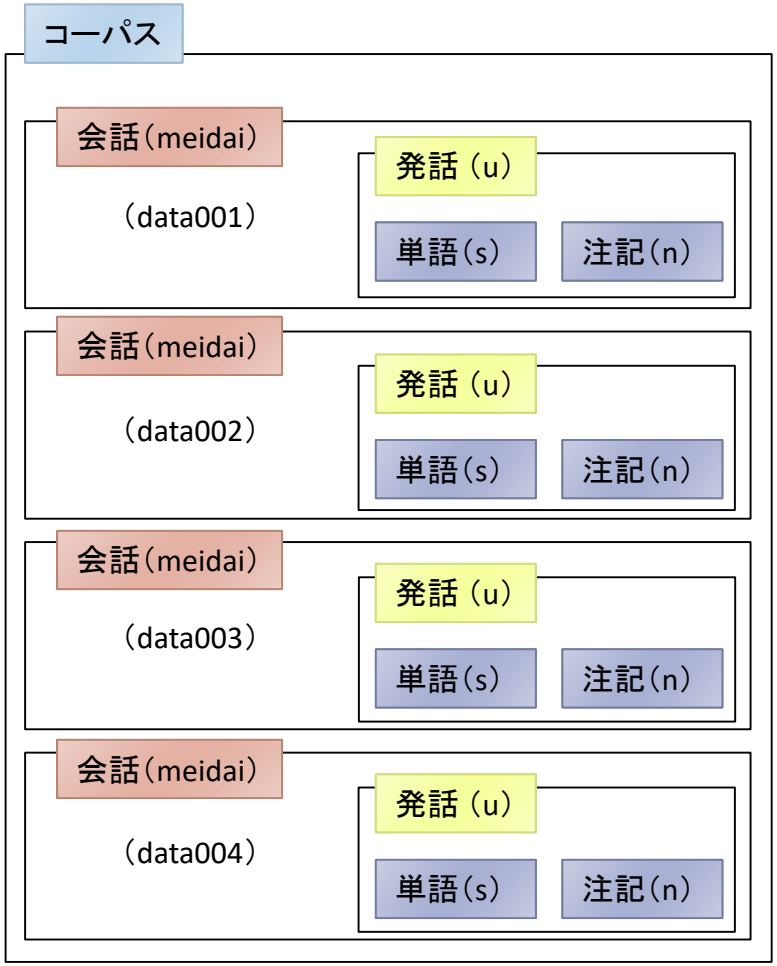
「キー」「開催日」
を選択



現在の頻度欄の
値を考慮

config_meidai.xml

コーパスの構造と検索（名大会話コーパス）

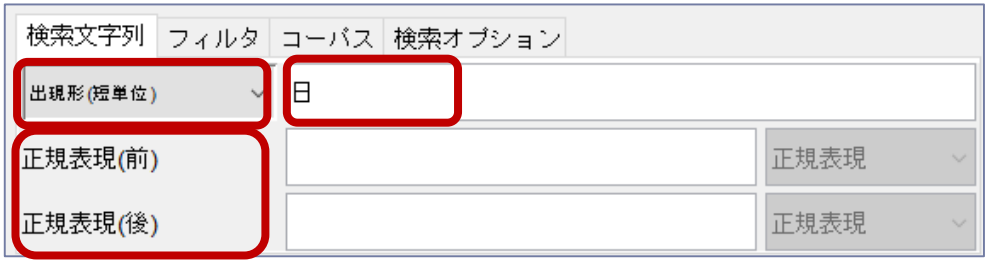


- 会話データ（meidai）
  - 発話（u）の集まり
  - 収録日，データ名などの属性
- 発話（u）
  - 単語（s）と注記（n）の集まり
  - 発話の名前，性別などの属性
- 単語（s）
  - 短単位で記述
  - 品詞などの属性
  - 外部DBではないので検索方法に注意
- 注記(n)
  - ＜笑い＞などの雑多な情報
  - 本文ではなく，属性として記述

単語(タグ)での検索

A) 「日」を含む単語

インターフェイスが変わることに注意

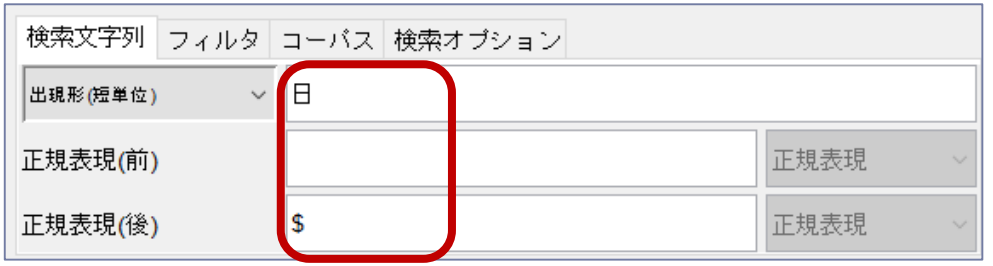


B) 先頭が「日」の単語

正規表現の「^日」と同義
(先頭の文字が「日」)

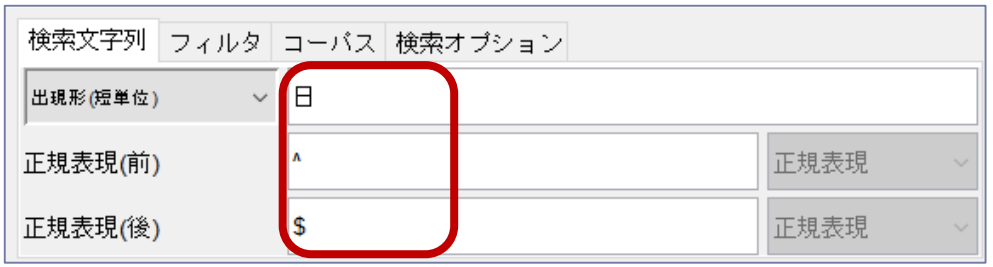
C) 末尾が「日」の単語

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



練習

- 発話データごとの発話数を求める

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―

やまなし
宮沢賢治

3種類の独自タグ

【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。』

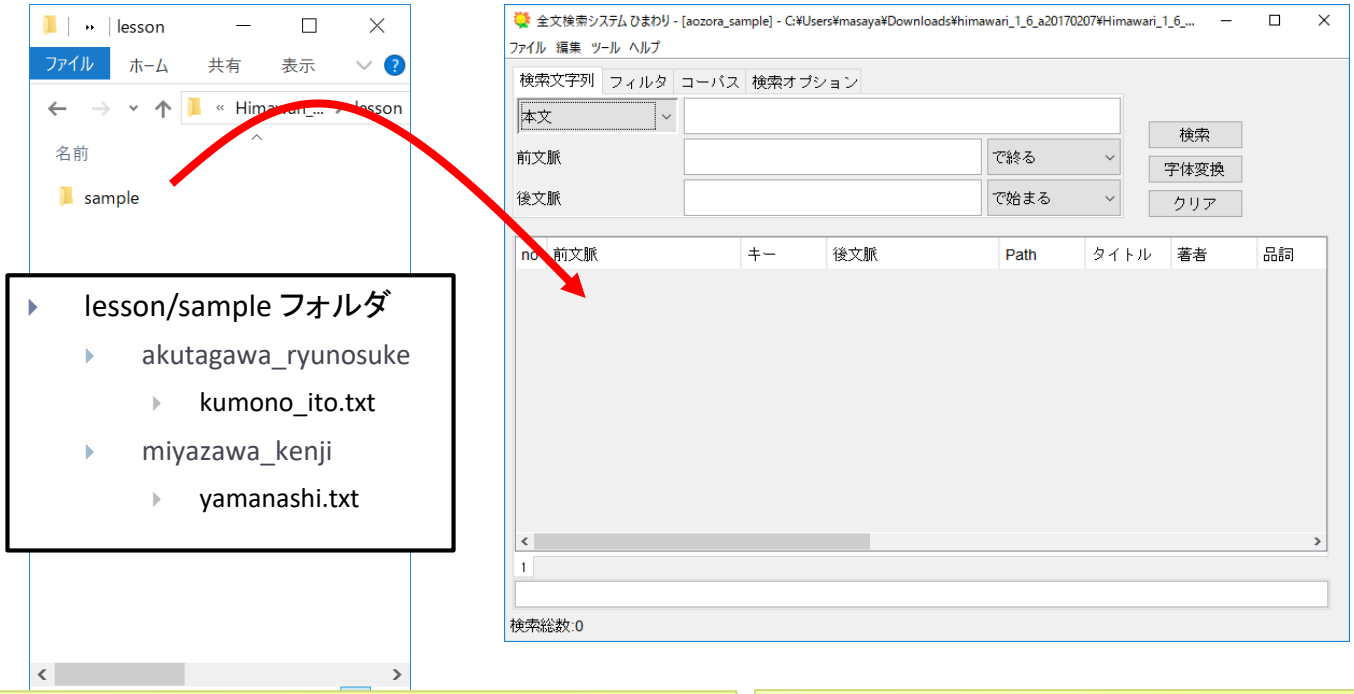
- タグは検索の障害になる
  - 例：「跳ねて」が検索できない
  - 例：注記の中の「五月」が検索される
- 本文とそれ以外の区別ができない
  - 例：ルビ
  - 例：資料の注記（左のタグの説明など）

『ひまわり』はタグを解釈して全文検索

- 本文と付与情報（属性）は区別して検索
- 「跳ねて」も検索OK
- ルビでの検索OK

インポートの実行

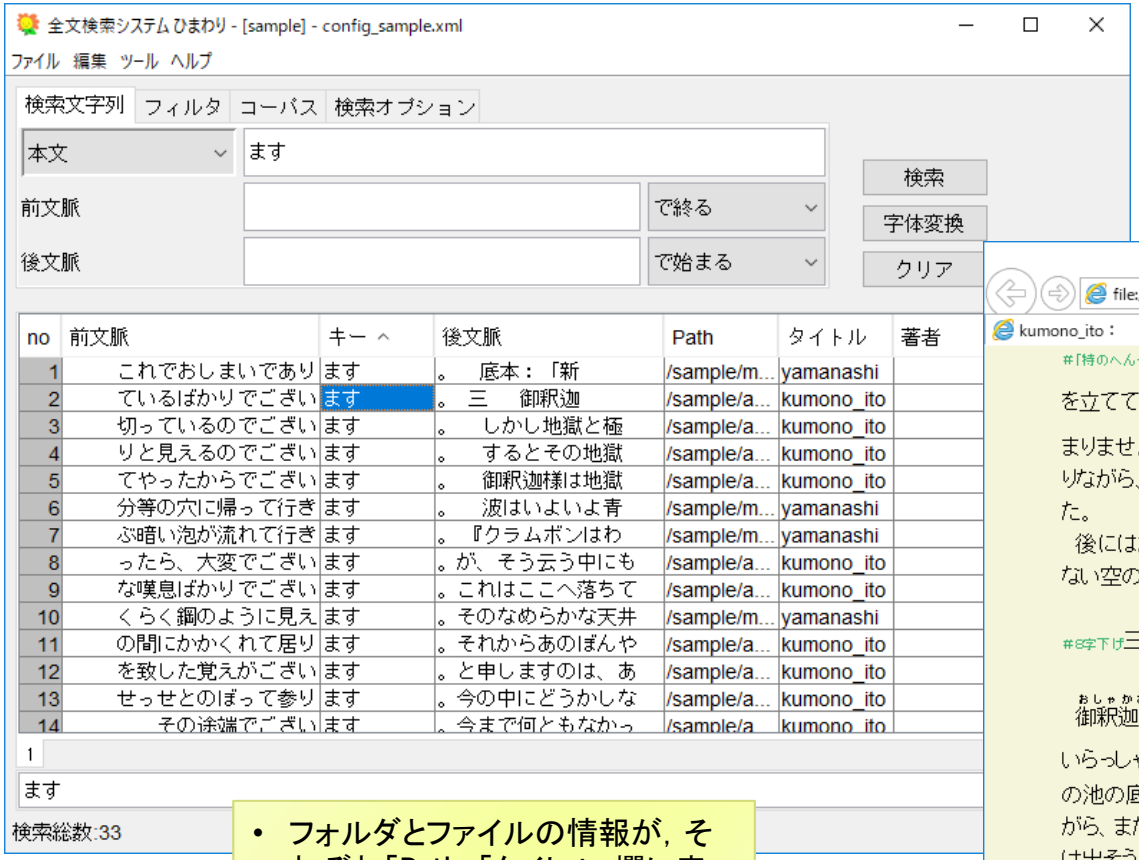
- sampleフォルダを，起動している『ひまわり』にドラッグ &ドロップ



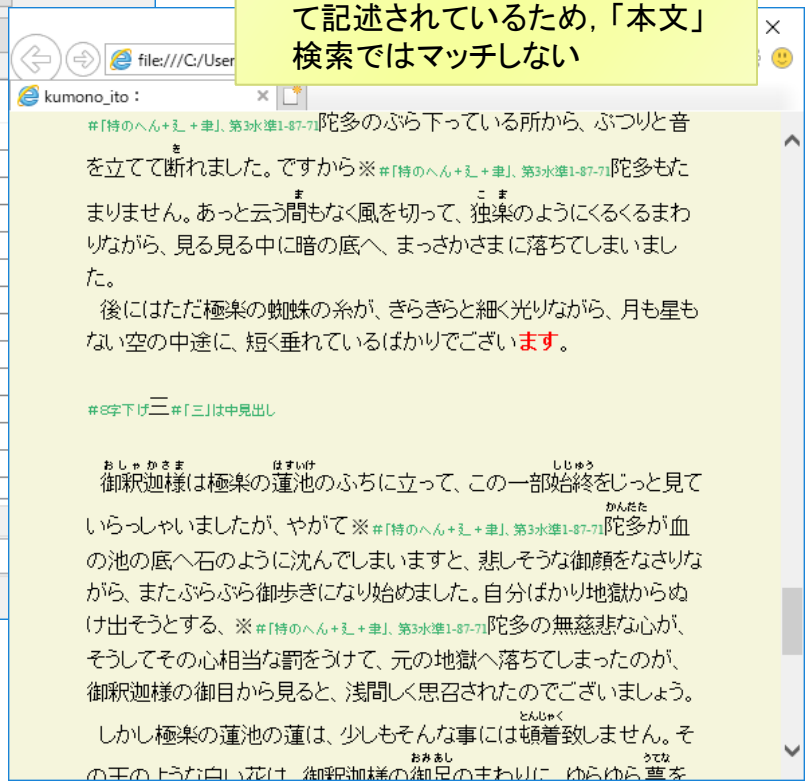
- フォルダの情報をインポート時に利用
  - フォルダ階層 ⇒ Path 欄
  - ファイル名 ⇒ タイトル欄
- ドロップしたフォルダ名がコーパス名になる

- 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

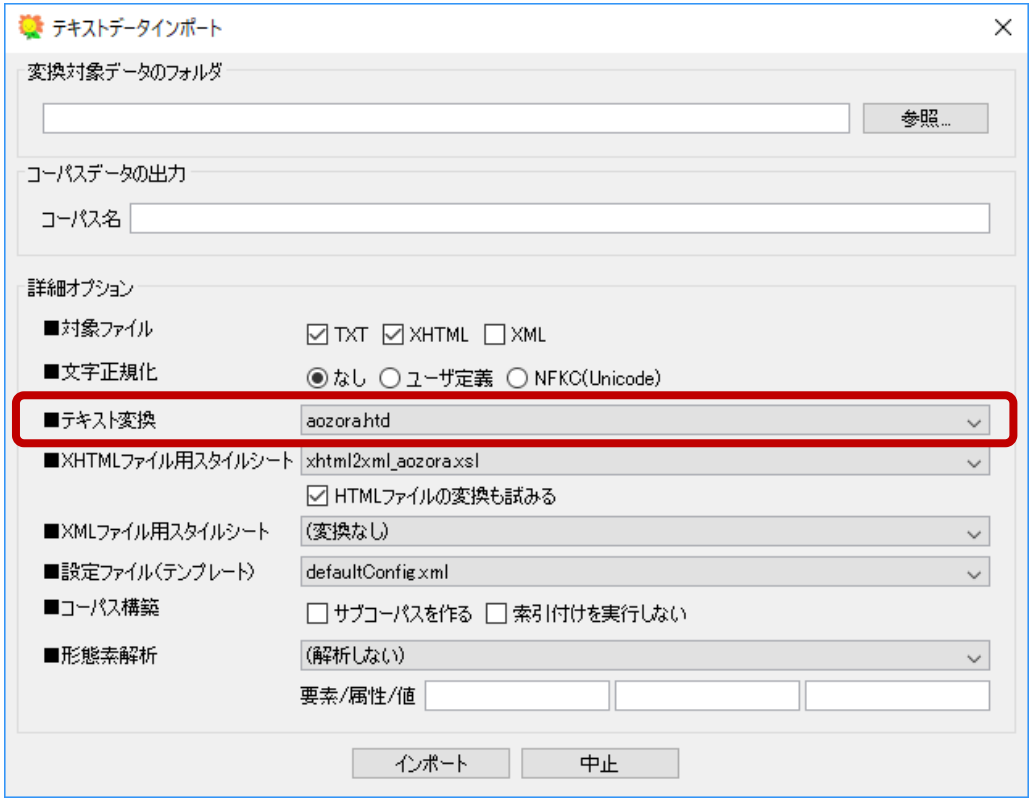


- フォルダとファイルの情報が，それぞれ「Path」「タイトル」欄に表示される
- 「著者」欄は空欄



- 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

インポート時のオプション



- 本資料では，対象ファイルTXT（テキスト変換）のみを扱う
- XHTML，XML（スタイルシート）については，一般的な規格なので，適宜資料を参照のこと。また，文字正規化，形態素解析などの処理はTXTと同様に適用される

- 文字正規化
  - ユーザ定義： 半角英数字⇒全角
    (.himawari_import_config.xml参照)
  - NFKC: Unicodeで規定される正規化
    - 例： 全角英数字 ⇒ 半角英数字
    - 例： 半角カタカナ ⇒ 全角カタカナ
- テキスト変換
  - resources/htd/aozora.htd
    - 改行位置に，
を挿入
    - 注記，ルビをタグに変換
  - resources/htd/diy.htd
    - 自作コーパス用
    - 汎用タグでテキストにタグ付け可能
- 形態素解析
  - MeCab, Jumanなどで解析し，結果を「外部データベース」に格納
  - 解析対象の要素を指定できる

おわりに

- 全文検索システム『ひまわり』チュートリアル
  - 『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
  - テキストファイルのインポート

- さらに詳しく知るには
  - 『ひまわり』ホームページ
  - 『ひまわり』用各種パッケージのWebページ
    - 青空文庫
    - 名大会話コーパスなど

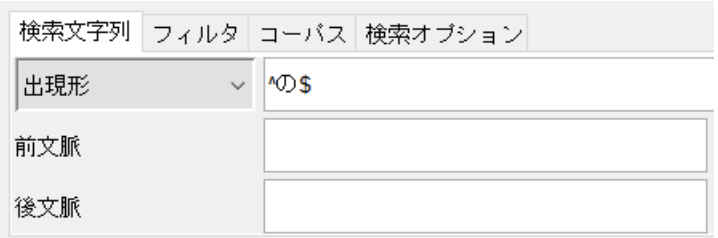
参考資料

- 全文検索システム『ひまわり』
  (http://www2.ninjal.ac.jp/lrc)
  - 『国会会議録』パッケージ
  - 『名大会話コーパス』パッケージ
  - 『ひまわり』で『日本語話し言葉コーパス』を利用する方法
  - マニュアル（インポートについては，7章）

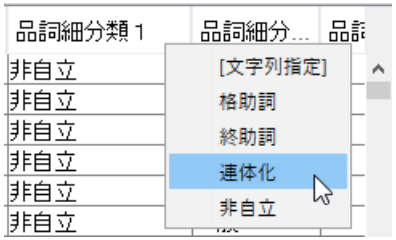
- 正規表現
  - Java Pattern クラス（『ひまわり』で利用できる正規表現の仕様）
    (https://docs.oracle.com/javase/jp/8/docs/api/java/util/regex/Pattern.html)
  - 「Java正規表現の使い方」
    (http://www.javadrive.jp/regex/)

練習問題(p.19)の解答例

1. 出現形で「の」を検索



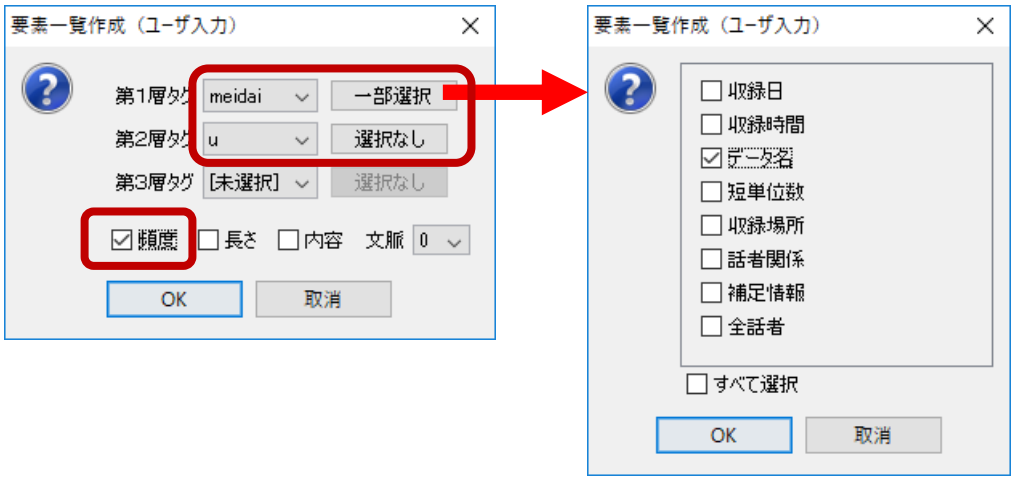
2. 「品詞再分類1」列のフィルタで「連体化」を指定



3. 「基本形ー1」,「基本形1」列中のセルを選択して，右クリック⇒「統計」

練習問題(p.28)の解答例

1. [ツール]⇒[一覧]⇒ユーザ入力
2. meidai，uタグを次のように指定



実行結果