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，各種のコーパスを利用する方法を紹介
  - 『ひまわり』（ver.1.6_ls2 = ver.1.6b02+実習資料・開発版）
  - 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
  - 応用例
  - テキストデータのインポート

『ひまわり』とは

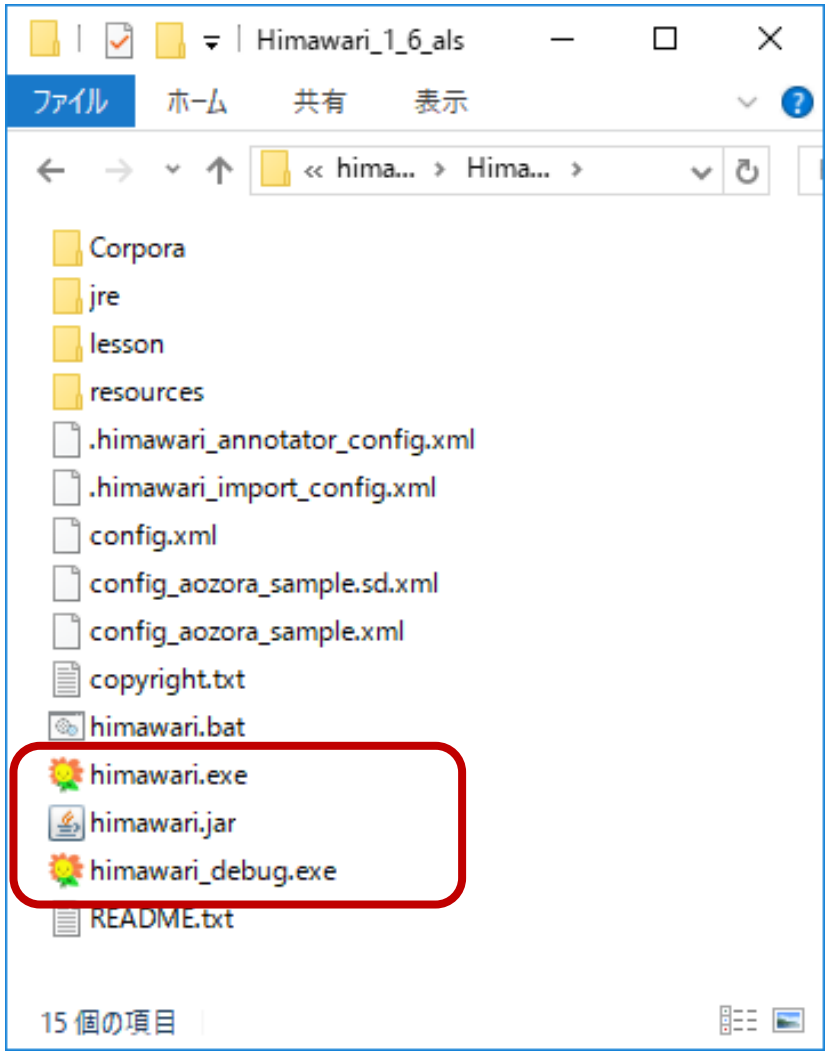
- 言語研究用の全文検索システム
  - 指定された文字列を網羅的に検索して，前後文脈付きで結果を表示します（コンコーダンサ）
  - 『太陽コーパス』（20世紀初頭の総合雑誌『太陽』）用の検索システムとして構築しました

- 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

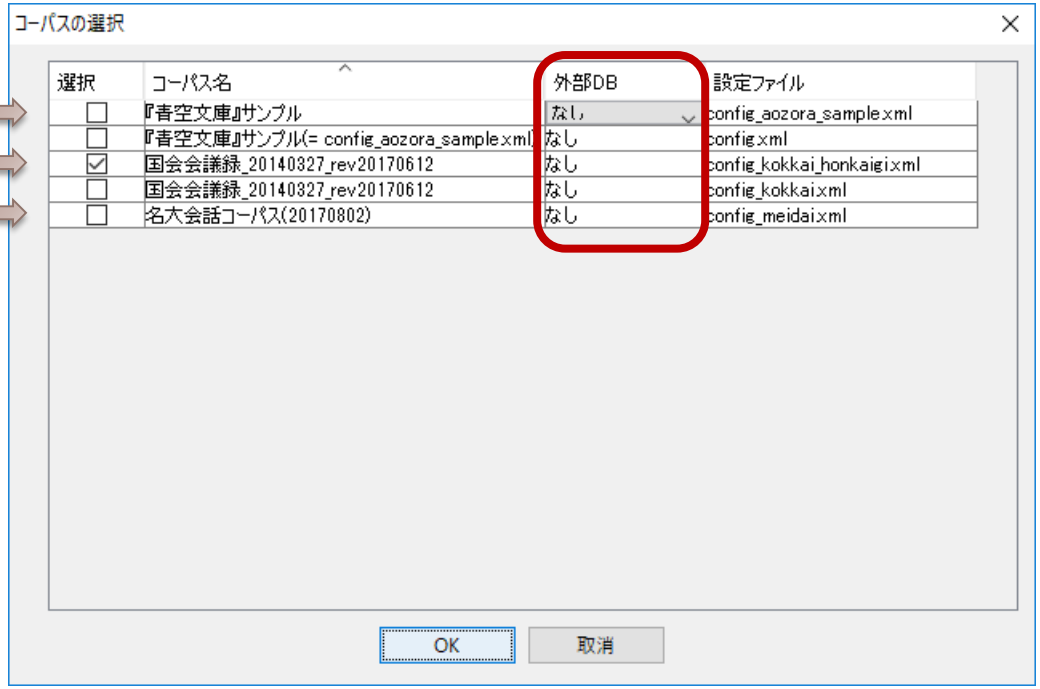
コーパスを作るとき
検索の途中経過を見たいとき
（Windows 専用）
himawari_debug.exe

himawari.jar

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

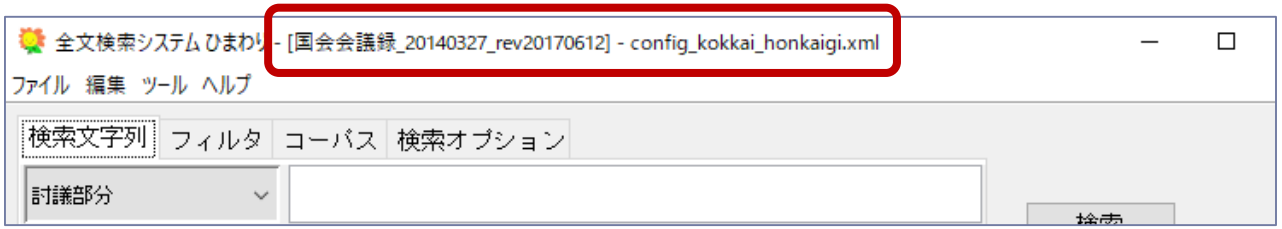
コーパスの選択

- [ファイル]⇒[コーパス選択]



- 本日使うコーパス(⟹)
- 「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

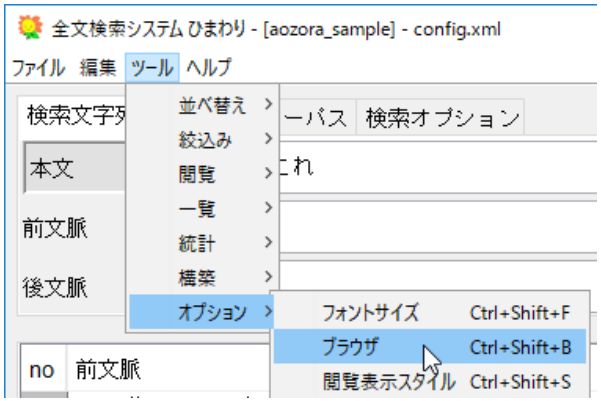

「検索文字列」欄では
右クリックで履歴表示
全文検索システム ひまわり - [aozora_sample] - config.xml
ファイル 編集 ツール ヘルプ
検索文字列
フィルタ
コーパス
検索オプショ
検索文字列
検索の実行
本文
これ
前文脈
で終る
後文脈
で始まる
検索
字体変換
クリア
no
前文脈
キー
後文脈
Path
タイトル
著者
検索結果
検索総数:597
検索総数
途中経過の表示

ブラウザでの閲覧

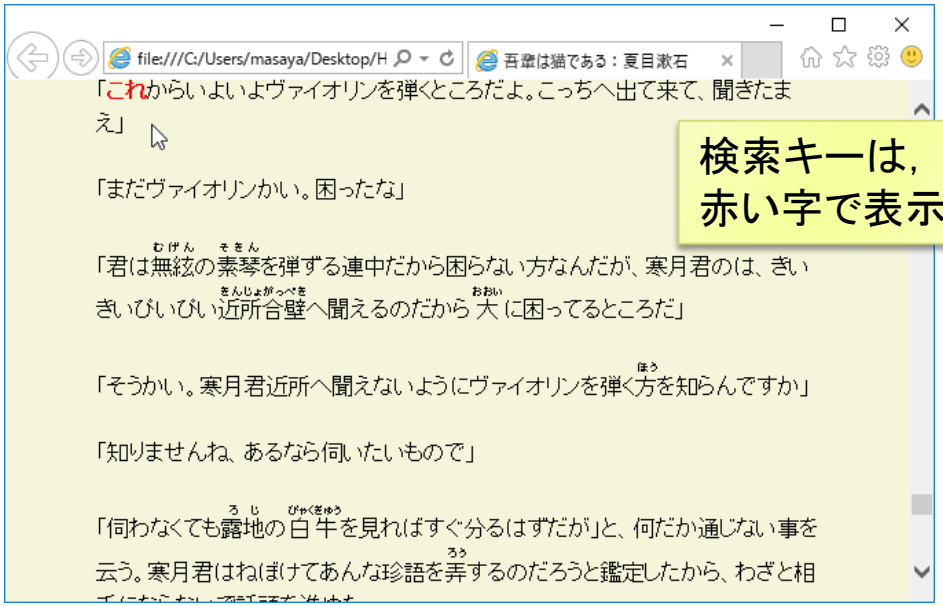
no	前文脈	キー	後文脈	Path
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...
3	弾くところです」「	これ	からいよいよヴァイオ	/aozora_s...
4	い話があるかい」「	これ	からいよいよヴァイオ	/aozora_s...
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aozora_s...
6	見当がつかない」「	これ	からいよいよ弾くとこ	/aozora_s...
7	めちゃなお困ります。	これ	からがいよいよ佳境に	/aozora_s...

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

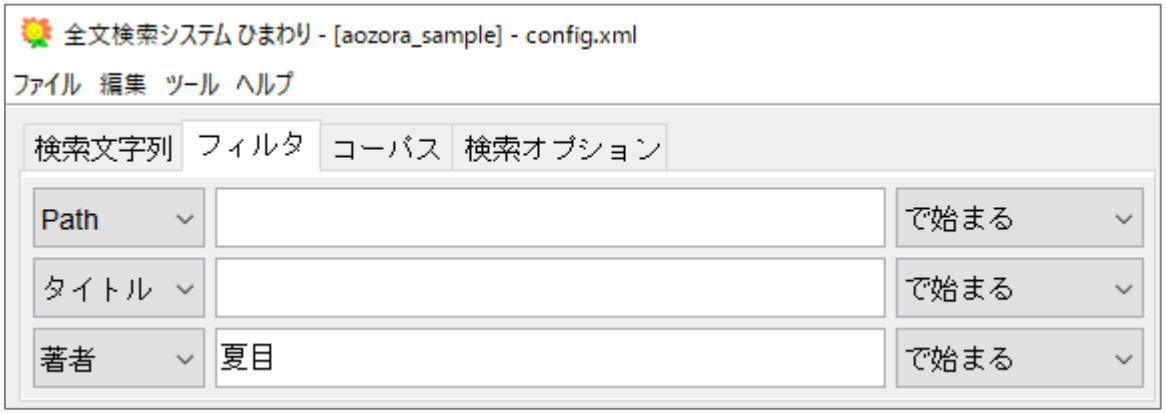
- 昇順
  - 列タイトルをクリック
- 降順
  - シフトキーを押しながら
    列タイトルをクリック

- 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

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

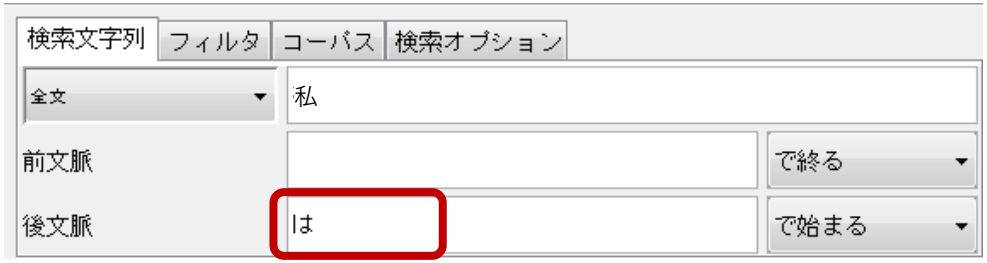
[文字列指定]
[置換]
夏目漱石
芥川龍之介

絞り込みたい値を選択
⇒右クリック
⇒フィルタでもOK

さまざまな検索と各種機能

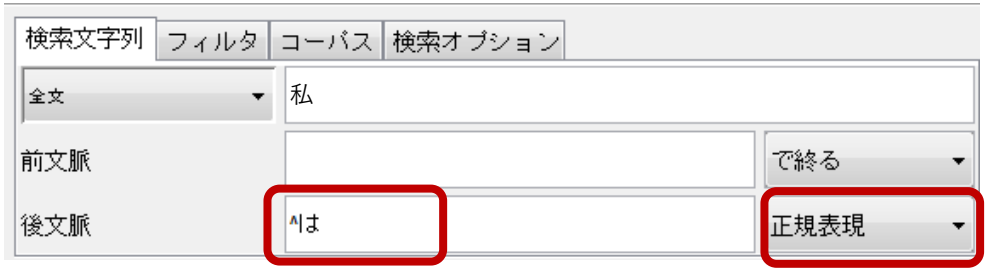
前後文脈の制限

A) 後文脈を
「は」で「始まる」に制限

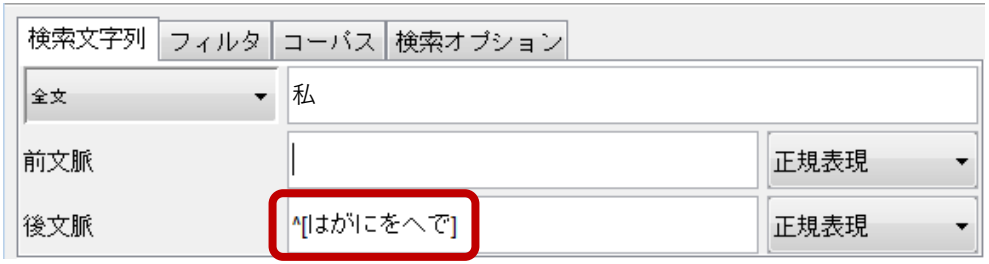


B) 正規表現で
「は」で「始まる」を表現

^ … 文字列の先頭を表す



C) 「私」
＋格助詞，係助詞

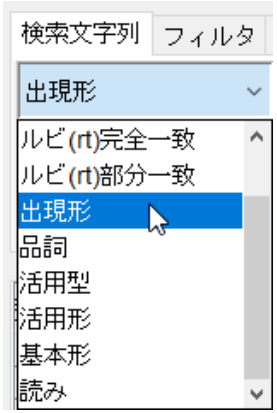


（「検索オプション」の「後文脈を含む」をチェックしてみましょう）

単語での検索(1)

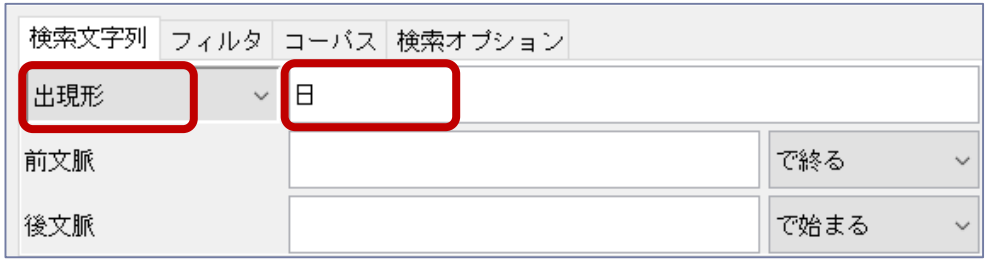
青空文庫サンプル
(形態素解析結果付き)を対象に
config_aozora_sample.sd.xml

- 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
  - 名大会話コーパスについては，HPを参照



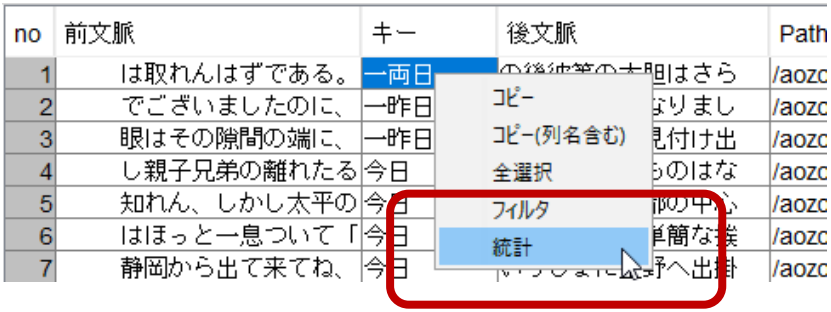
A) 「日」を含む単語

「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



単語での検索(2)

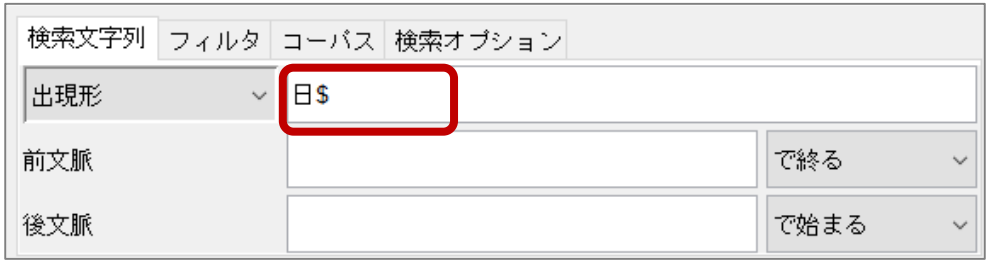
青空文庫サンプル
(形態素解析結果付き)を対象に
config_aozora_sample.sd.xml

C) 先頭が「日」の単語

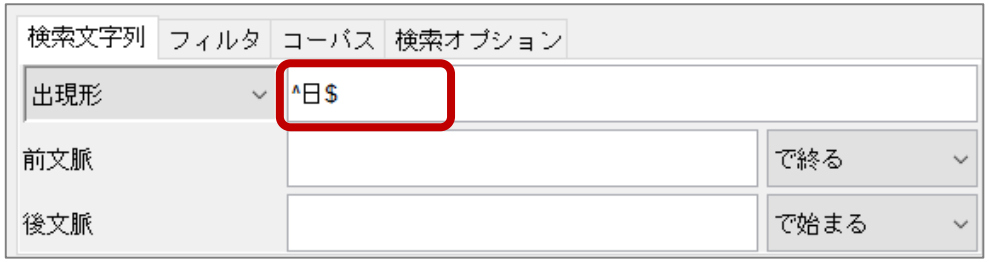
正規表現の「^」
(文字列の先頭)

D) 末尾が「日」の単語

正規表現の「$」
(文字列の末尾)

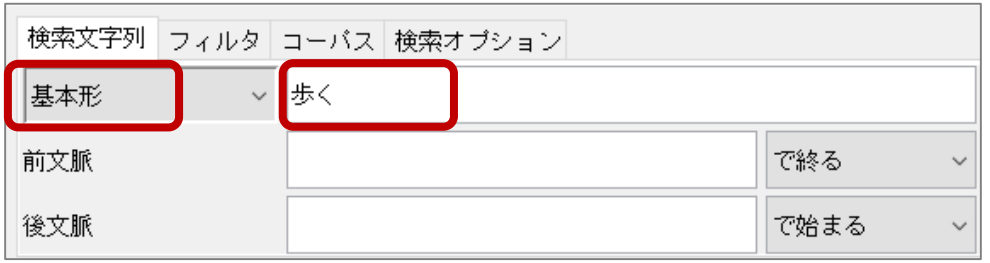


E) 単語「日」のみ



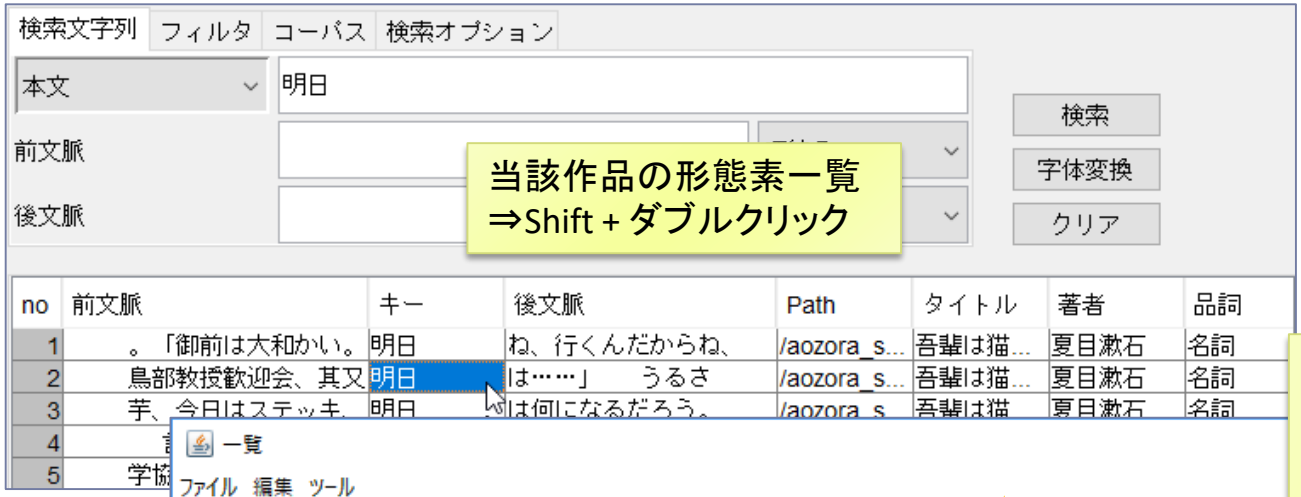
F) 活用語の基本形

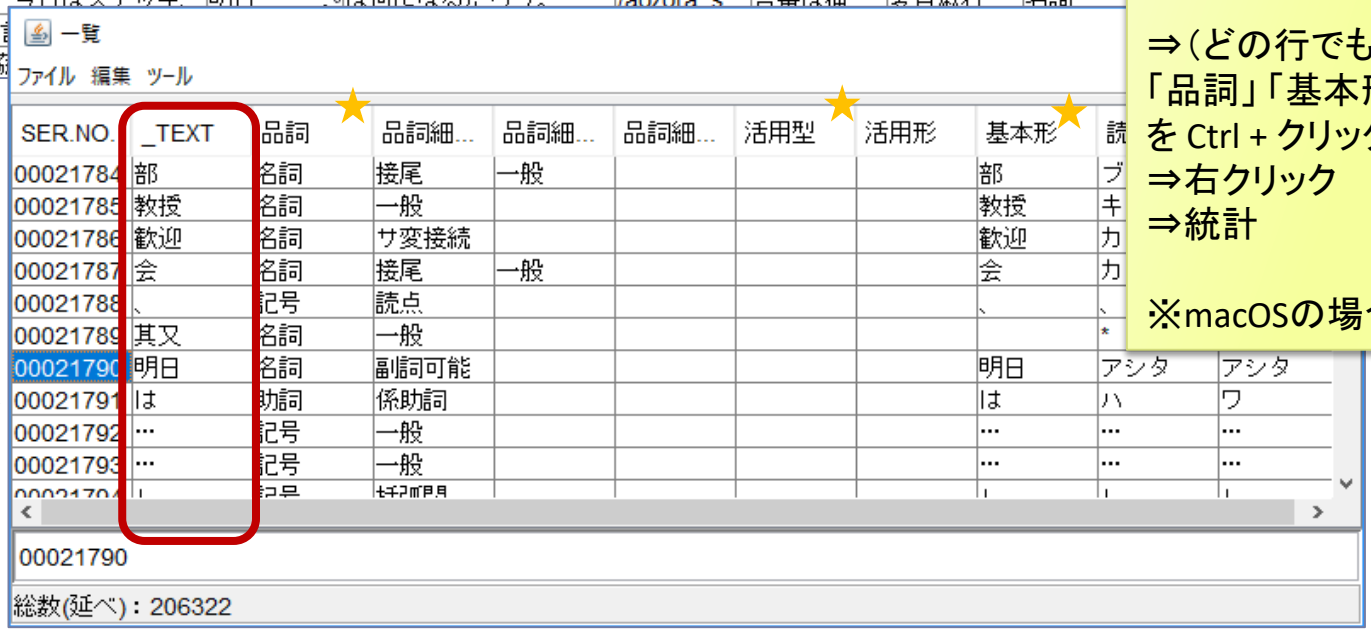
すべての語形を
一括して検索



形態素解析結果の閲覧

この機能は，config_xxx.sd.xml タイプの資料のみ実行可能





テキスト進行方向

語彙一覧（頻度付き）

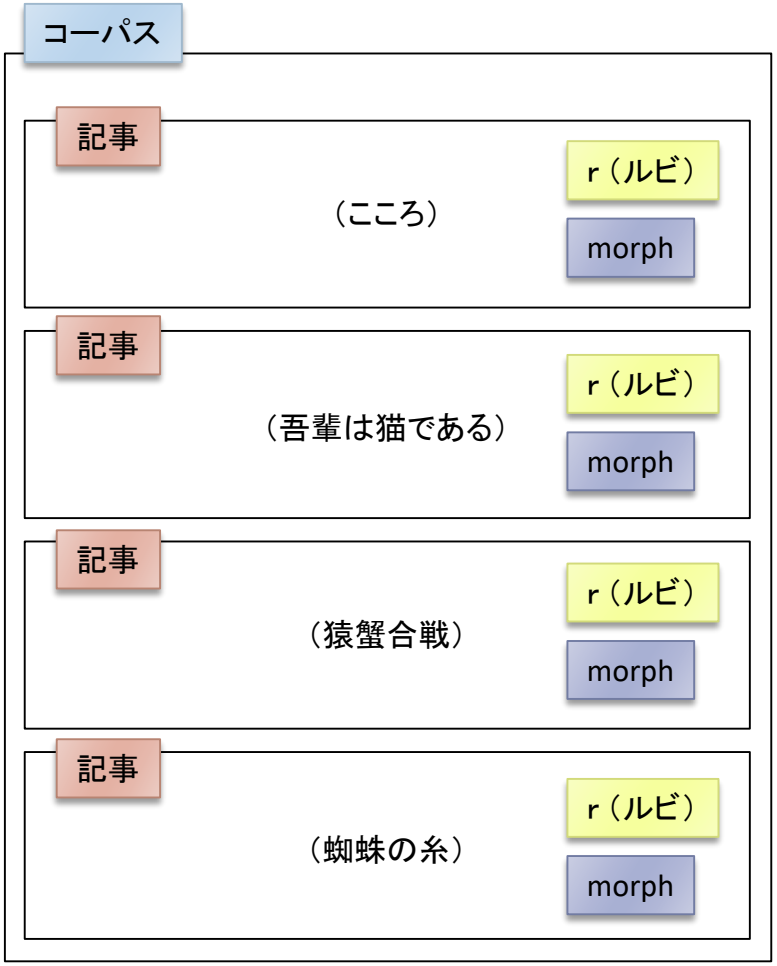
⇒（どの行でもよい）
「品詞」「基本形」「活用型」を Ctrl + クリックで選択
⇒右クリック
⇒統計

※macOSの場合は，command

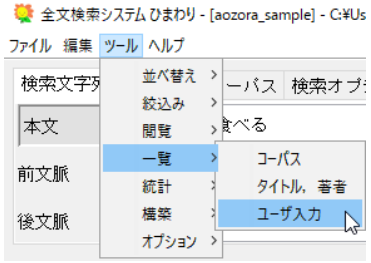
コーパスの構造と検索

コーパスの構造と検索（青空文庫サンプル）

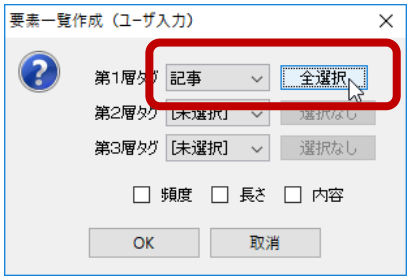
コーパス全体の構造

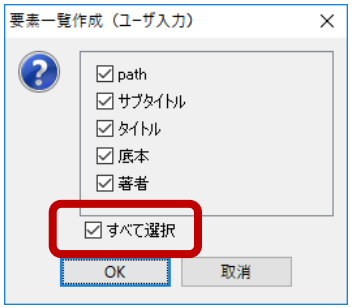


一覧機能（ユーザ入力）で付与情報を閲覧

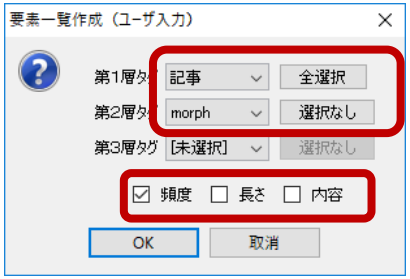


記事





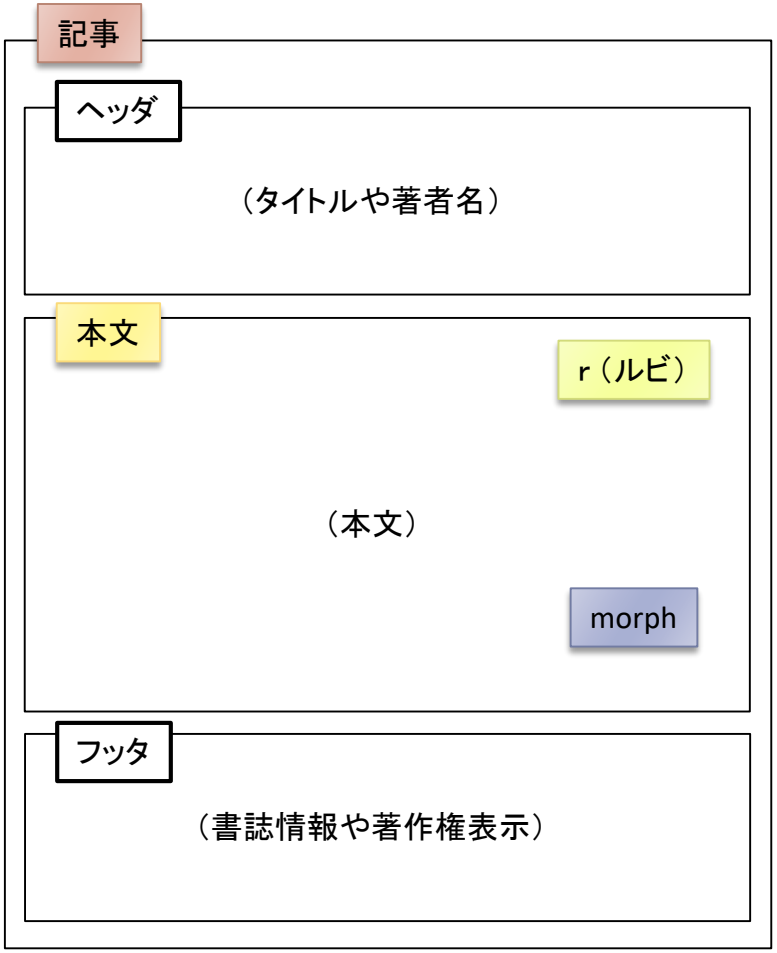
記事＋morph

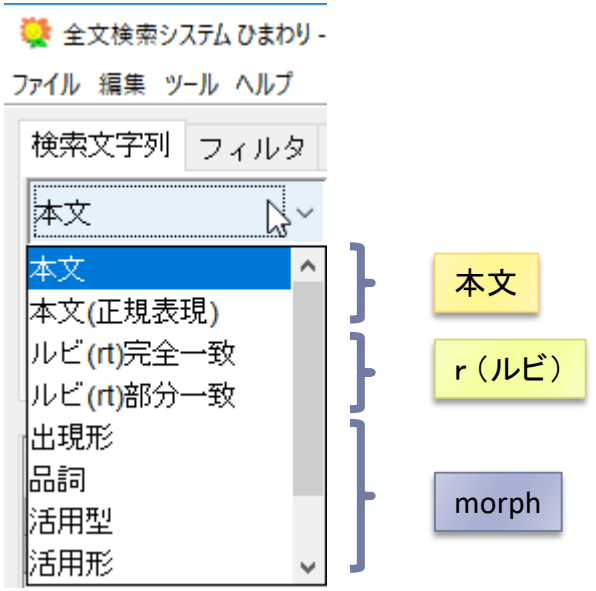


各「記事」に含まれる単語数を計測

コーパスの構造と検索（青空文庫サンプル）

- 記事部分





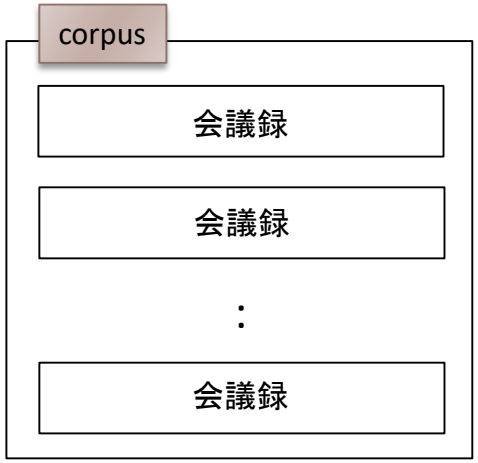
コーパス本体を見たい場合
Corpora/aozora_sample/corpus.xml

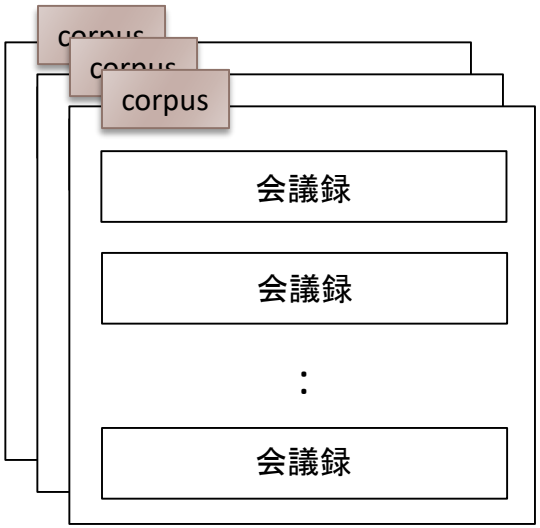
ブラウザで閲覧した記事
Corpora/aozora_sample/xslt/__searched_tmp.xml

※「秀丸」などのテキストエディタを利用のこと

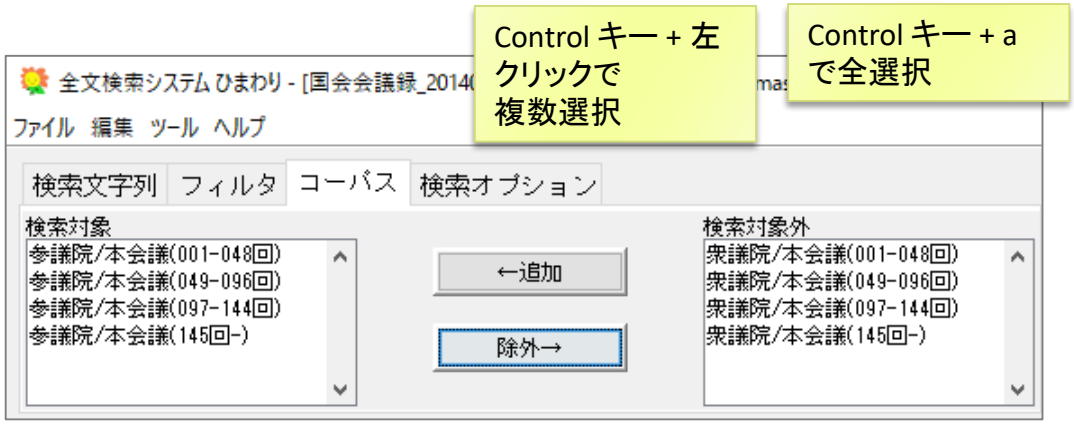
config_kokkai_honkaigi.xml

コーパスの構造と検索（国会会議録）





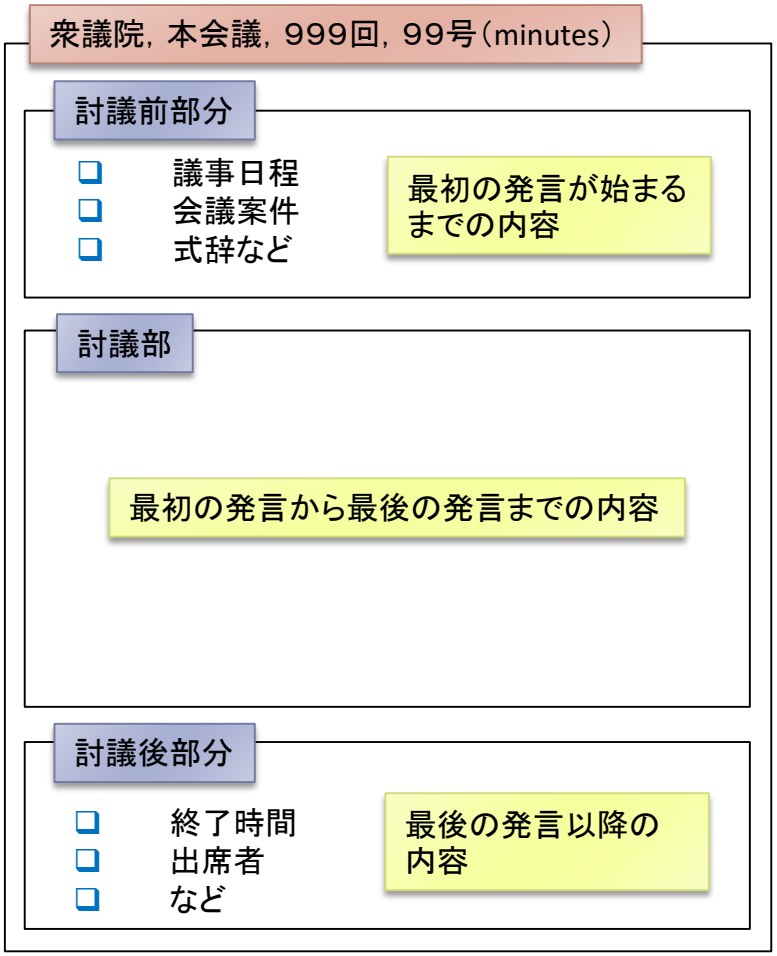
- 計8個のサブコーパス
  - 参議院/本会議
    - 001–048回
    - 049–096回
    - 097–144回
    - 145回–
  - 衆議院/本会議
    - 001–048回
    - 049–096回
    - 097–144回
    - 145回–

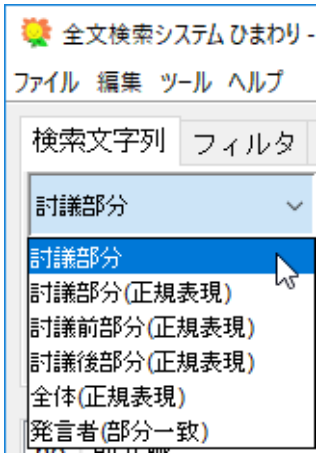


- サブコーパスにしている理由
  - ファイルサイズ・検索速度などシステム上の制約
  - テキストの品質（145回より古い会議はOCRによるテキスト入力）

コーパスの構造と検索（国会会議録）

- 会議録全体





- 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

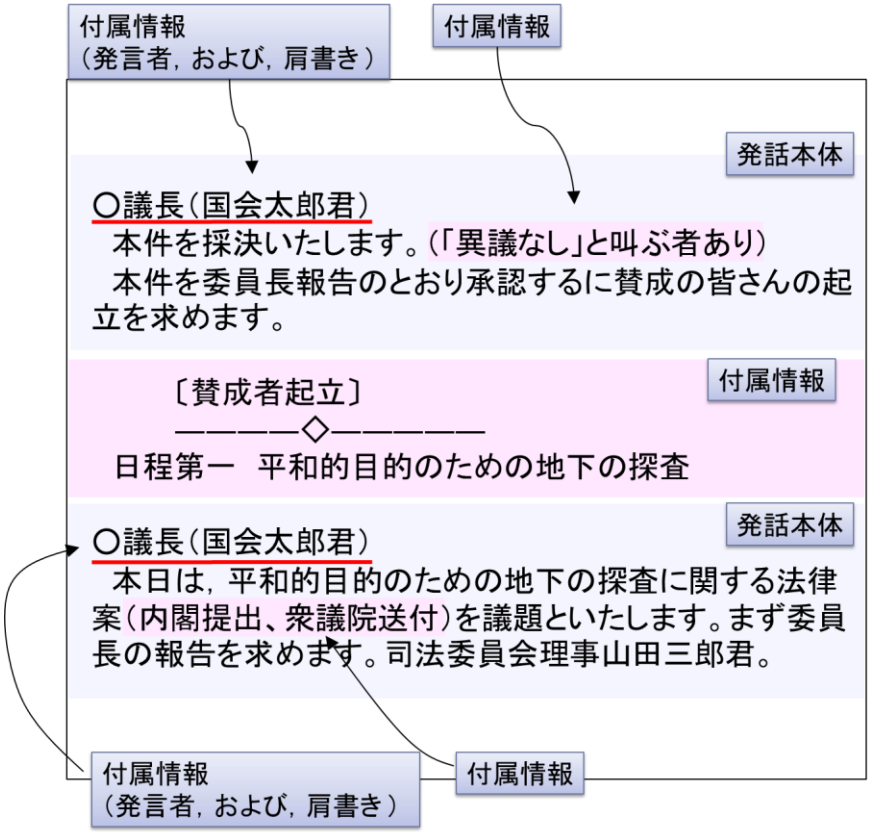
ブラウザで閲覧した記事
Corpora/Kokkai/xslt/__searched_tmp.xml

※「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

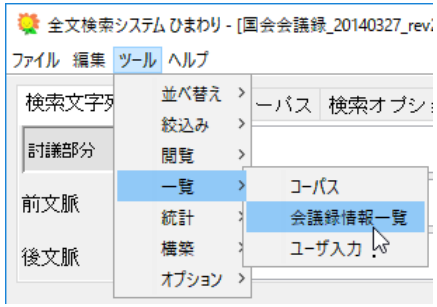
付属情報は本文の全文検索から除外

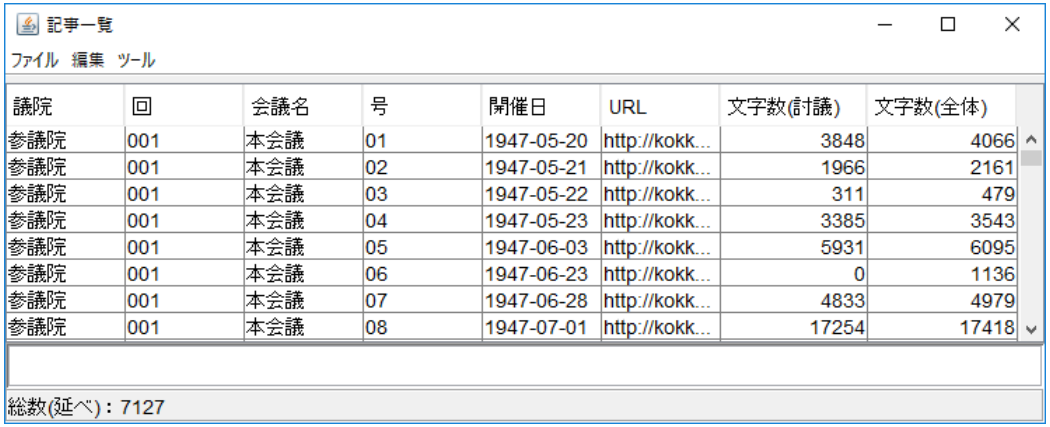


- 発言者の「国会太郎」を全文検索してもマッチしない
- ブラウザ表示では，付属情報も含めて表示される
- 「法律案を」にマッチする
  「（内閣提出、衆議院送付）」は読み飛ばされる

「国会会議録」パッケージを概観する

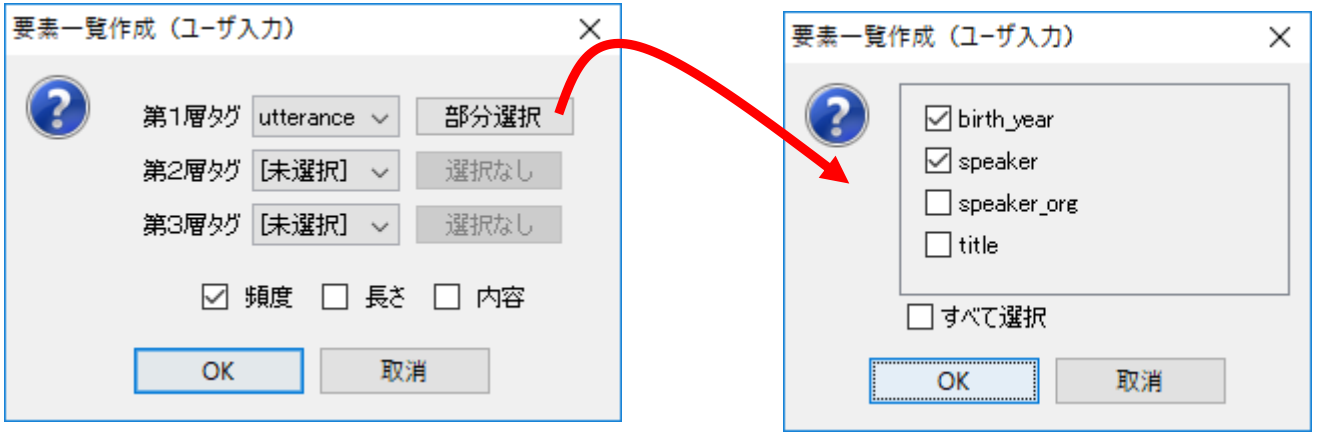
- 収録されている会議録一覧





議院	回	会議名	号	開催日	URL	文字数(討議)	文字数(全体)
参議院	001	本会議	01	1947-05-20	http://kokk...	3848	4066
参議院	001	本会議	02	1947-05-21	http://kokk...	1966	2161
参議院	001	本会議	03	1947-05-22	http://kokk...	311	479
参議院	001	本会議	04	1947-05-23	http://kokk...	3385	3543
参議院	001	本会議	05	1947-06-03	http://kokk...	5931	6095
参議院	001	本会議	06	1947-06-23	http://kokk...	0	1136
参議院	001	本会議	07	1947-06-28	http://kokk...	4833	4979
参議院	001	本会議	08	1947-07-01	http://kokk...	17254	17418

- 発言者一覧



応用例

表記の経年変化

国会会議録
config_kokkai_honkaigi.xml

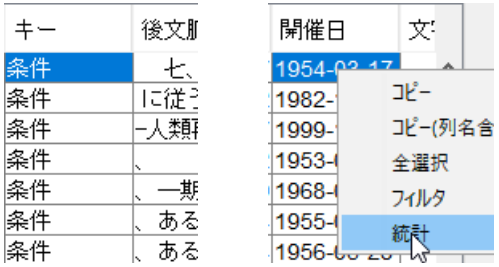
参議院/本会議(001-048回) を対象に

1 「条件」と「條件」の検索

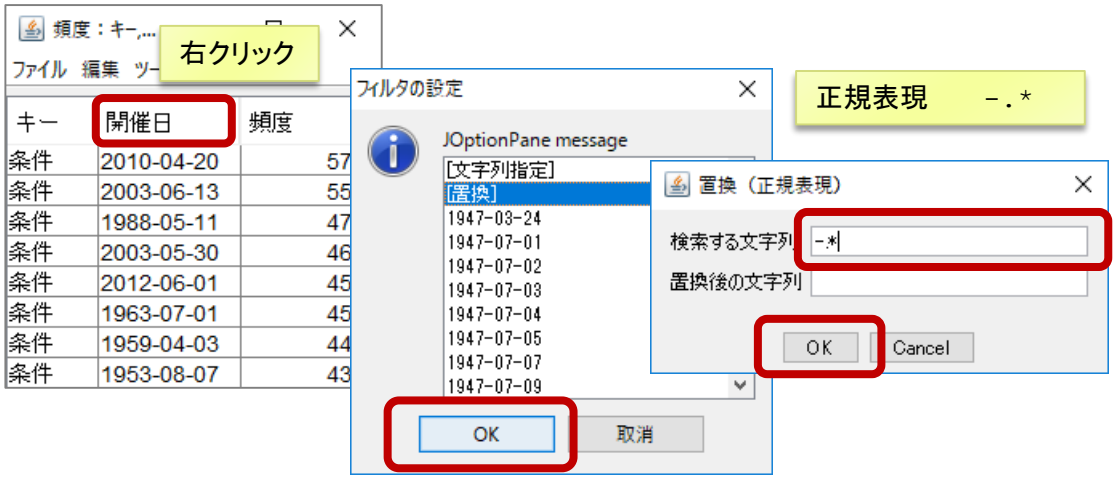
「条件」と入力して，
「字体変換」ボタン

※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



4 再集計

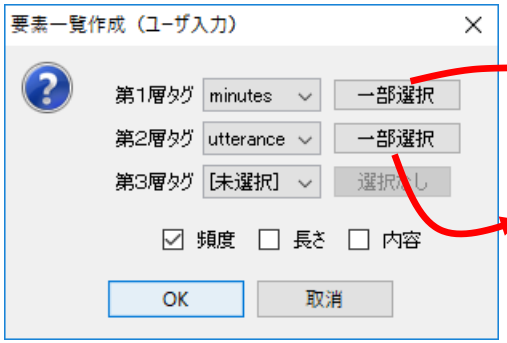
「キー」「開催日」を選択

現在の頻度欄の値を考慮

発言者の年齢分布の経年変化

1 発言者の一覧

[ツール]⇒[一覧]⇒ [ユーザ入力]



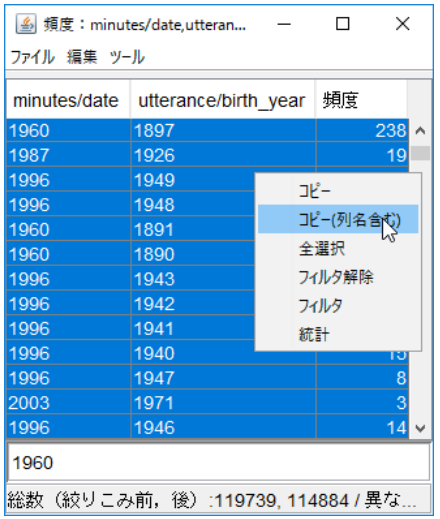
date 属性
（開催年月日）

birth_year
属性
（発言者生年）

3 エクスポート

Excel などの
外部プログラムに
「貼り付け」

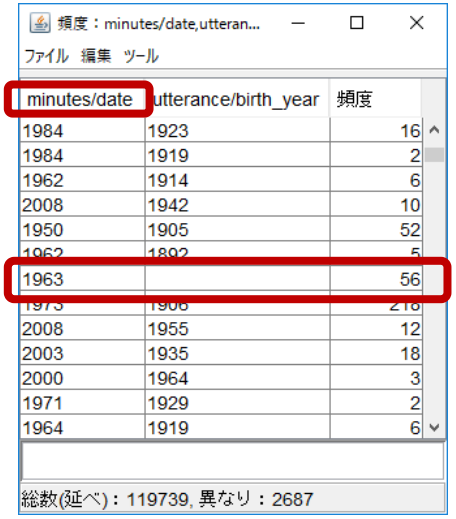
下図は
ピポットテーブルで
クロス集計し，
グラフ表示したもの

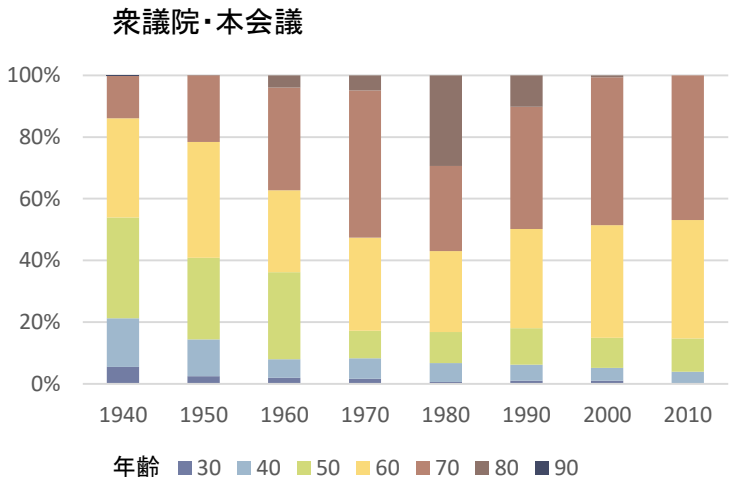


2 整形

- 年月日⇒年変換
- 生年が空白の
データを
フィルタリング

⇒列名を右クリック
⇒フィルタ
⇒文字列指定
（正規表現の「.」）



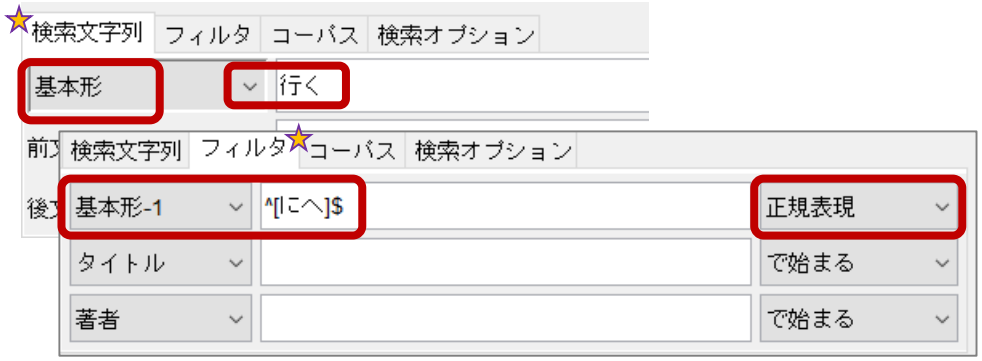


各種応用例

青空文庫サンプル
（形態素解析結果付き）を対象に
config_aozora_sample.sd.xml

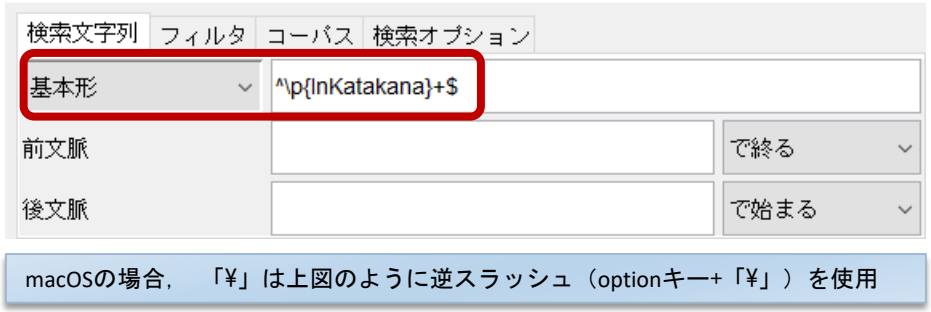
A) 共起語の集計
（「～へ行く」,「～に行く」）

「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）

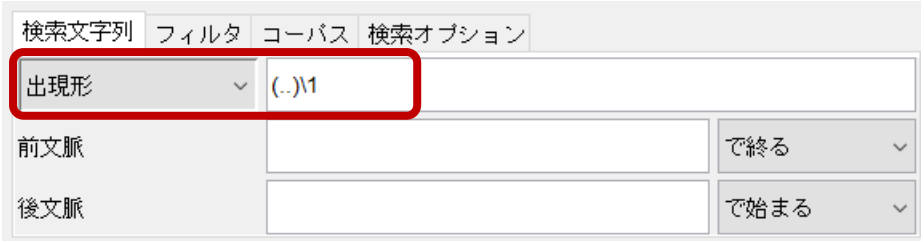
¥p{InHiragana}　…　ひらがな
¥p{InKatakana}　…　カタカナ
¥p{InCJKUnifiedIdeographs}　…　漢字
+　　…　直線の文字の繰り返し



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

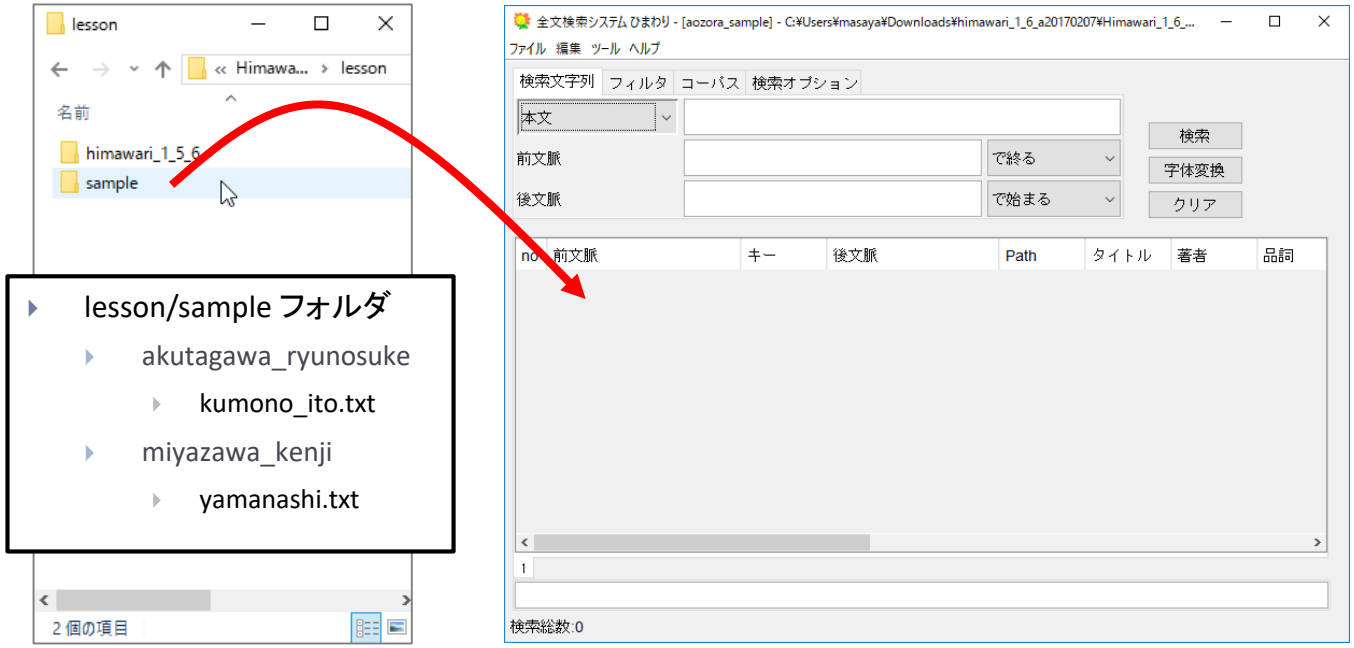
C) 繰り返し表現の抽出

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



テキストファイルのインポート

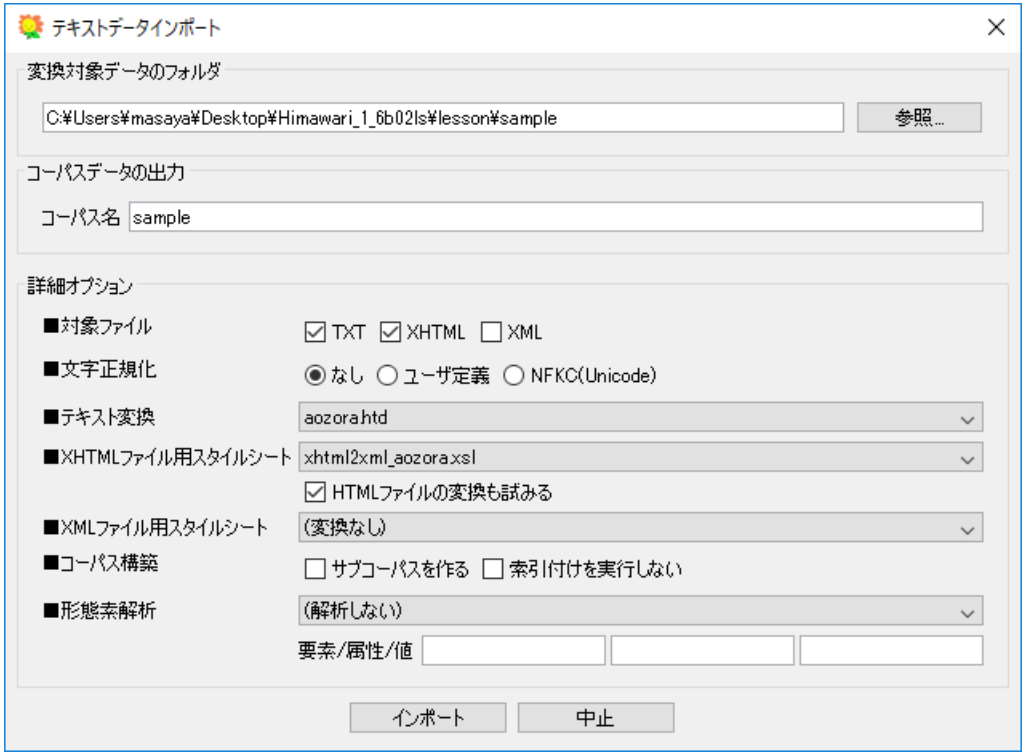
- sampleフォルダを，起動している『ひまわり』にドラッグ &ドロップ



- 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
- 生テキストやHTML，XMLをインポート可能

- 文字コードは自動判別
- 詳細オプションで形態素解析も実行可能
  （MeCab もしくは JUMAN をインストールのこと）
- インポート時の文字列変換規則も定義可能

インポート時のオプション



- 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 で定義
    - 改行位置に，
を挿入
    - 注記，ルビをタグに変換

変換前

［＃8字下げ］ー［＃「ー」は中見出し］

　ある日の事でございます。御釈迦様《おしゃかさま》は極楽の蓮池《はすいけ》のふちを、独りでぶらぶら御歩きになっていらっしゃいました。

変換後

```
<br />
<注 内容="＃8字下げ" 付与="" 種別="注記" />ー<注 内容="＃「ー」は中見出し" 付与="" 種別="注記" /><br />
<br />
　ある日の事でございます。<r rt="おしゃかさま">御釈迦様</r>は極楽の<r rt="はすいけ">蓮池</r>のふちを、独りでぶらぶら御歩きになっていらっしゃいました。
```

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
  - 応用例

- さらに詳しく知るには
  - 『ひまわり』ホームページ（ver.1.6は本年度安定版になる予定）
  - 『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
  - 簡単な検索用データの作成方法2

- 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/)