既存の作文コーパスを『ひまわり』で活用する

全文検索システム『ひまわり』講習会

「作文コーパスの活用」

本日の内容

- 第一部　既存の作文コーパスを『ひまわり』で活用する
（担当：山口昌也・国語研）
  - 全文検索システム『ひまわり』の基本的な使い方
  - 『ひまわり』による『小中高大生による日本語絵描写ストーリーライティングコーパス』(JASWRIC)の利用方法
  - JASWRICを使った簡単な分析

『ひまわり』の紹介＋生テキストの活用

- 第二部　『日本語学習者作文コーパス』を『ひまわり』で検索する
（担当：森篤嗣先生・京都外国語大）
  - 『日本語学習者作文コーパス』を『ひまわり』で検索する意義
  - 誤用の集計
  - 本来使用されるべき正用に基づく誤用の集計

作文コーパスの誤用分析

第一部
既存の作文コーパスを『ひまわり』で活用する

- 全文検索システム『ひまわり』の基本的な使い方
- 『ひまわり』による『小中高大生による日本語絵描写ストーリーライティングコーパス』(JASWRIC)の利用方法
- JASWRICを使った簡単な分析

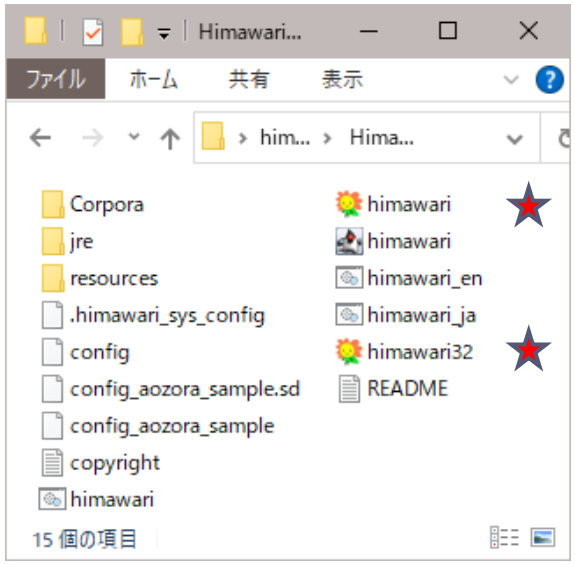
『ひまわり』とは

- 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
  - 『日本語日常会話コーパス』，『日本語話し言葉コーパス』，『分類語彙表』，『青空文庫』など多数の言語資料に対応

- 特徴
  - XMLでタグづけされたコーパスに対する全文検索，単語検索（格納しているXMLデータは他のシステムでも利用可能）
  - 検索結果，アノテーション結果の集計（例：総単語数）
  - 資料の特徴に合わせた資料の閲覧（例：横書き／縦書き表示）
  - コーパス自作支援機能

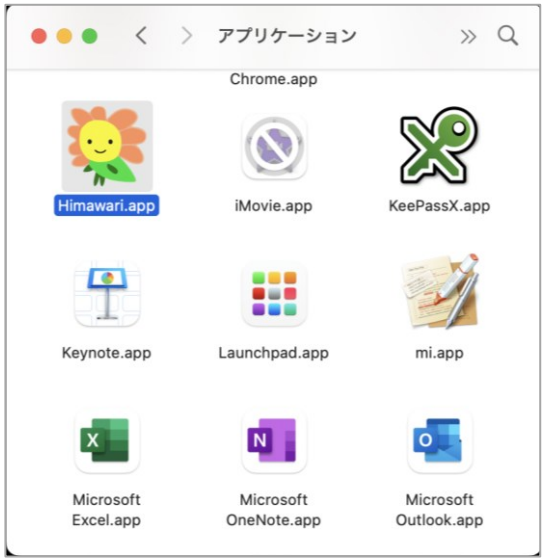
『ひまわり』の起動と『ひまわり』フォルダの確認

- Windowsの場合



  - 通常は  のhimawariを使用
  - 32bit版のWindowsの場合はWindows32を使用
  - OSの設定によらず日本語を使いたい場合は, himawari_ja

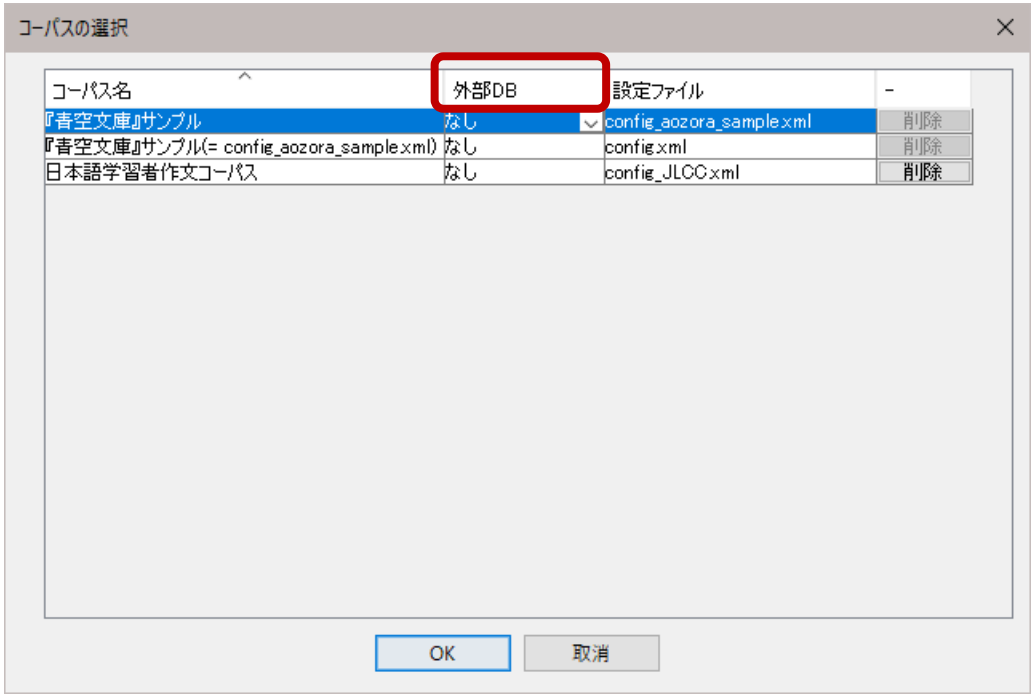
- macOSの場合



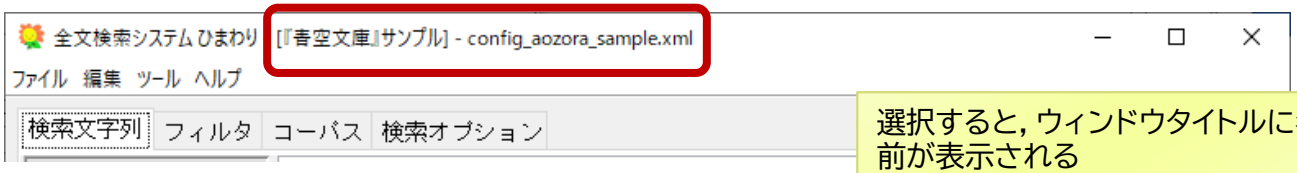
  - Himawari.appを使用
  - 右クリック→「パッケージの内容を表示」でフォルダを参照可能
  - Contents → Resources フォルダにコーパスなどを格納

コーパスの選択

- [ファイル]⇒[コーパス選択]



- 「外部DB」
  - 『青空文庫』サンプルなど，『ひまわり』のインポート機能で形態素解析を行った場合は，「あり」が選択可能
  - 本日のJASWRICも「あり」で使用する
  - 『日本語学習者作文コーパス』の形態素解析結果は原資料のデータを利用しているので「なし」



選択すると，ウィンドウタイトルに名前が表示される

検索する

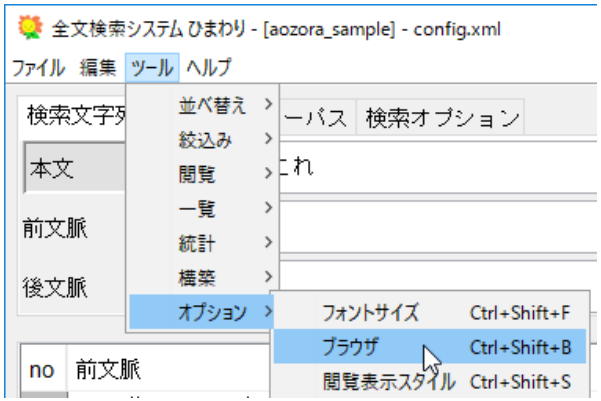

「検索文字列」欄では右クリックで履歴表示
全文検索システム ひまわり - [aozora_sample] - config.xml
ファイル 編集 ツール ヘルプ
検索文字列 フィルタ コーパス 検索オプショ
検索文字列
検索の実行
本文
これ
前文脈
で終る
後文脈
で始まる
検索
字体変換
クリア
no 前文脈 キー 後文脈 Path タイトル 著者
1 指して、しきりにかれ これ いいたがるのを、始め /aozora_s... こころ 夏目漱石
2 一軒屋を敲いて、これ これ かようかようしかじか /aozora_s... 吾輩は猫... 夏目漱石
3 弾くところです」「 これ からいよいよヴァイオ /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 はすこぶる不愉だよ。 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
12 とーと息ついた。「 これ からが聞きどころです /aozora_s... 吾輩は猫... 夏目漱石
13 んだ。「まだです。 これ からが面白いところで /aozora_s... 吾輩は猫... 夏目漱石
14 と信じました。同時に これ からさき彼を相手にす /aozora_s こころ 夏目漱石
検索結果
検索総数:597
検索総数
C:¥WINDOWS¥system32¥
C:¥Users¥masaya¥Deskt
¥ProgramData¥Oracle¥J
em32¥Wbem;C:¥WINDOWS¥
cal¥Microsoft¥Windows
C:¥Users¥masaya¥Deskt
t.jar
osname: windows 10
osname: windows 10
osname: windows 10
info:
info: _sys _prec
info: _sys _key
info: _sys _foll
info: 記事 path
info: 記事 タイ
info: 記事 著者
corpusname aozora_sam
info(available memory
history added: これ
doSearch:start
corpusname aozora_sam
open eix
eix:記事
n:597
info(available memory
time[milisec]: 65
time[milisec]: 65
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...

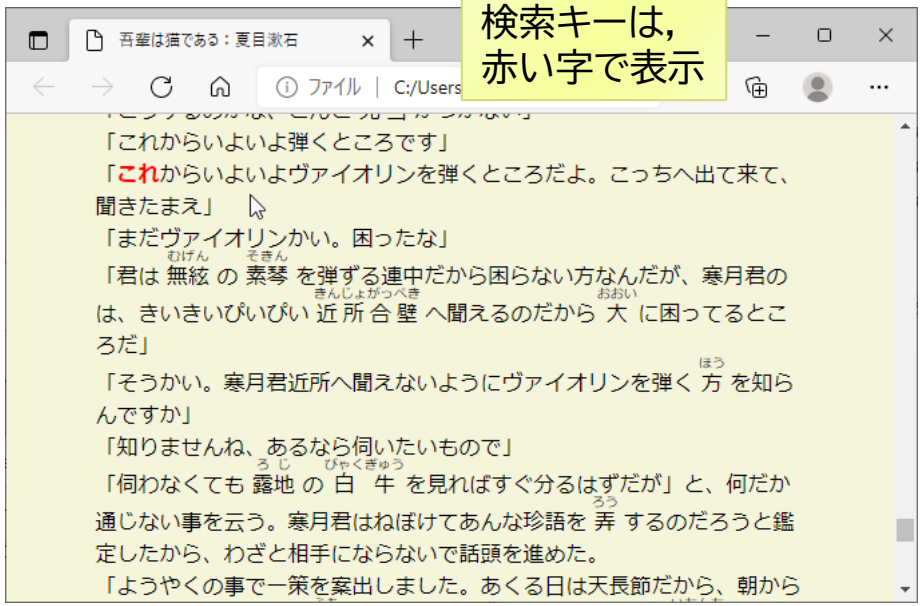
閲覧したい用例をダブルクリック

■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

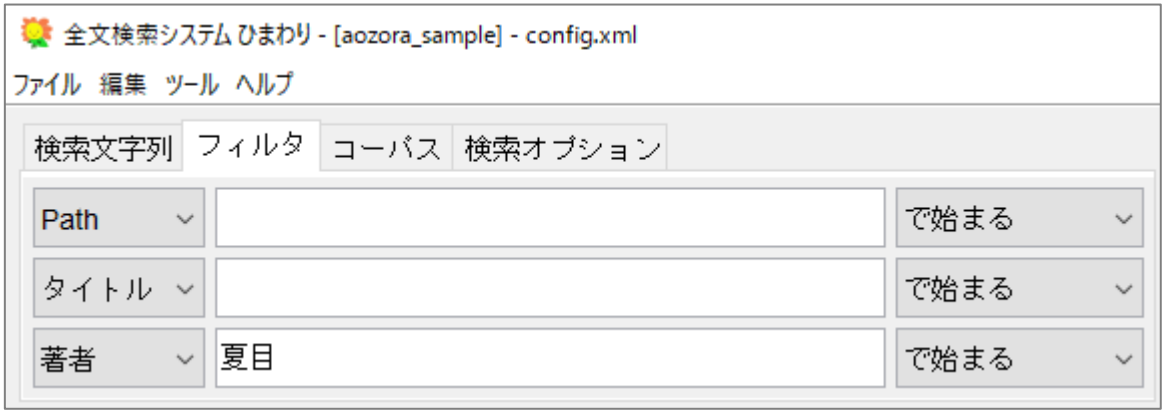
- 昇順
  列タイトルをクリック
- 降順
  シフトキーを押しながら
  列タイトルをクリック

- 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:「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

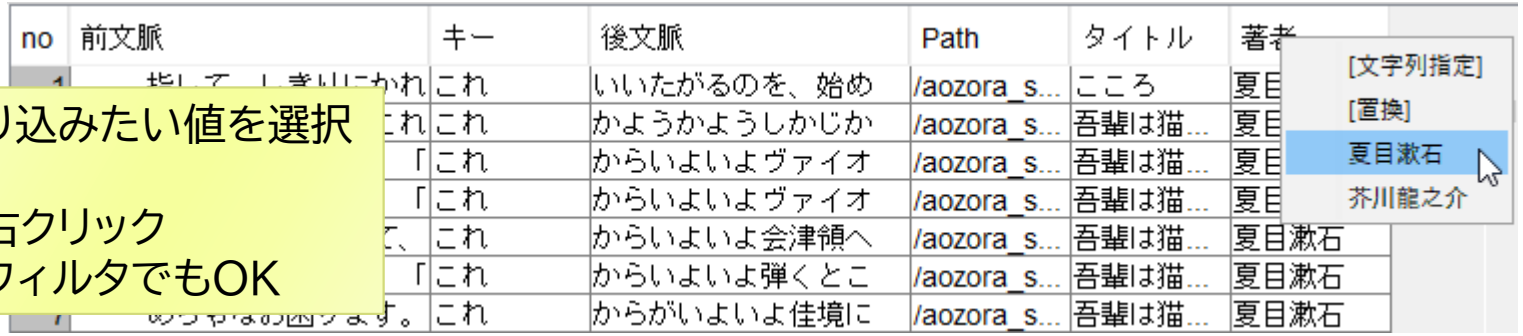
- 検索時に指定



「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

検索結果の頻度集計

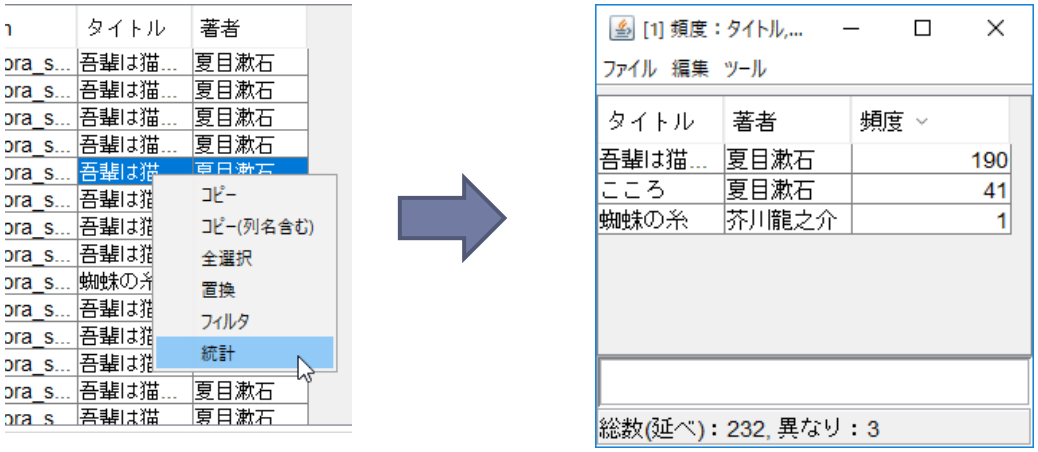
1. 集計したい列を選択

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」



『小中高大生による日本語絵描写ストーリーライティングコーパス』（JASWRIC）のインポート

JASWRICの概要

- 収録データ
  - 2種類の連続イラストに基づき，日本の小・中・高・大学生，合計700名が書いた作文を集めたもの
  - 作文数は1400件，総語数は約13万6千語

（https://language.sakura.ne.jp/jaswric）

- データ構成
  - Edited Data
    - 「形態素解析用に校閲を加えたデータ」
  - Raw Data
    - 手書き作文のスキャンデータ，書き起こしテキスト
  - JASWRIC_Participant Survey.xlsx
    - 著者情報
  - JASWRIC_Tagged.xlsx
    - 形態素解析済みデータ（校閲済みのデータ）

本講習で利用
（書き起こしテキスト）

本家サイトの検索システムで使用されている

『ひまわり』で利用する方法

- Raw Dataをインポート
  - 本講習
  - 『ひまわり』標準のテキストデータインポート機能を利用

- JASWRIC_Tagged.xlsxをインポート
  - 『ひまわり』のHPでインポート方法を公開
  - タブ区切りテキストを『ひまわり』用XMLデータに変換

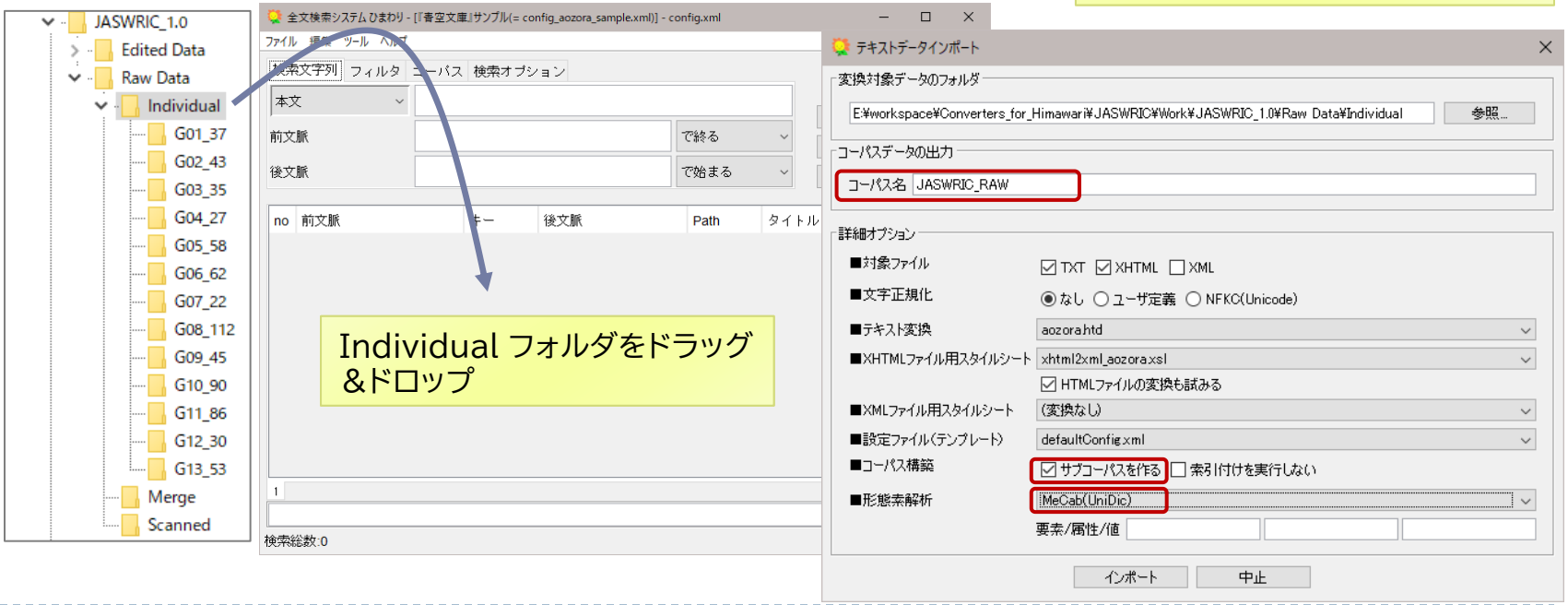
利点

■ 『ひまわり』の各種機能を利用できること
（例： 全文検索, 単語検索, 集計機能など）
■ 検索システム中の元データを確認・検証できること
■ 原資料に対するアノテーション・編集ができること
（例： 形態素解析システム JUMANでアノテーション）

インポートの実行

- インポート時の処理
  - 学年ごとにサブコーパスとしてインポート
  - テキスト自体には変更は加えない
  - MeCab(UniDic)で形態素解析

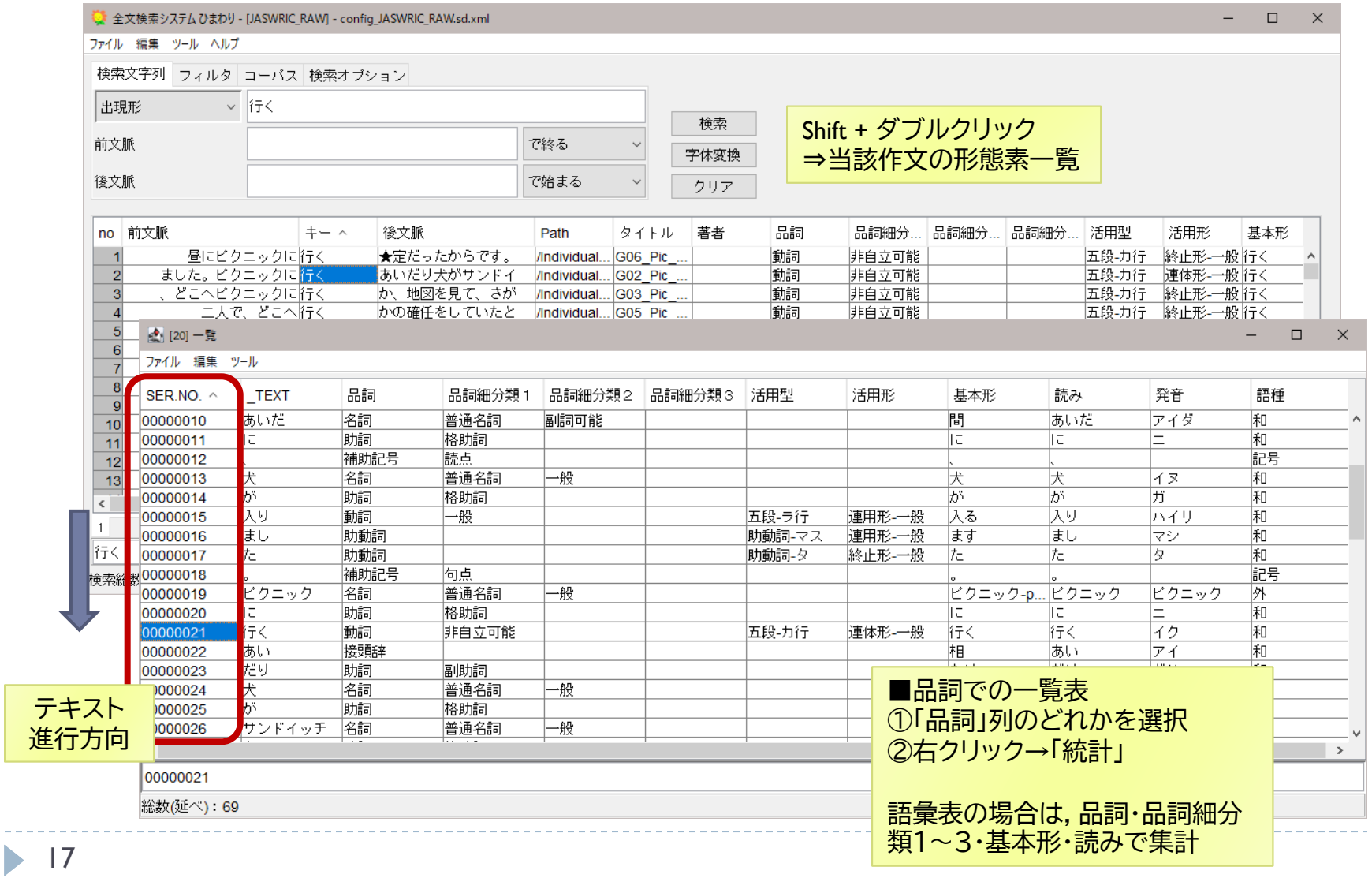
- コーパス名: JASWRIC_RAW
- 「サーブコーパスを作る」
- MeCab(UniDic)



動作の確認： 検索と作文全体の閲覧(1)

- 「行く」(出現形)で検索

動作の確認： 検索と作文全体の閲覧(2)



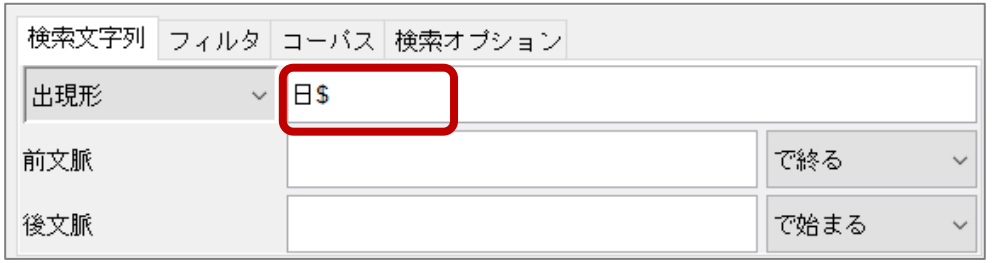
補足：単語での検索（本講習会のJASWRICなど）

C) 先頭が「日」の単語

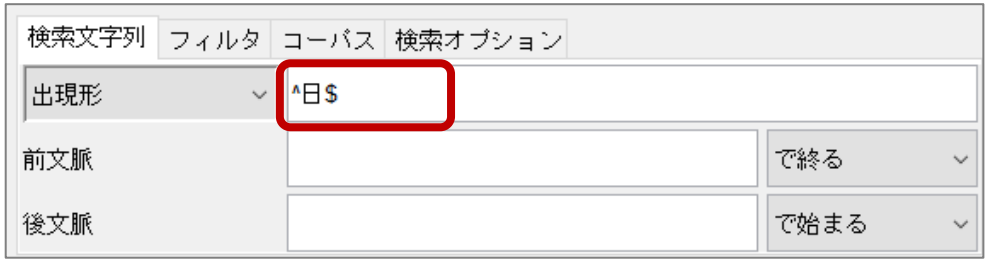
正規表現の「^」（文字列の先頭）

D) 末尾が「日」の単語

正規表現の「$」（文字列の末尾）

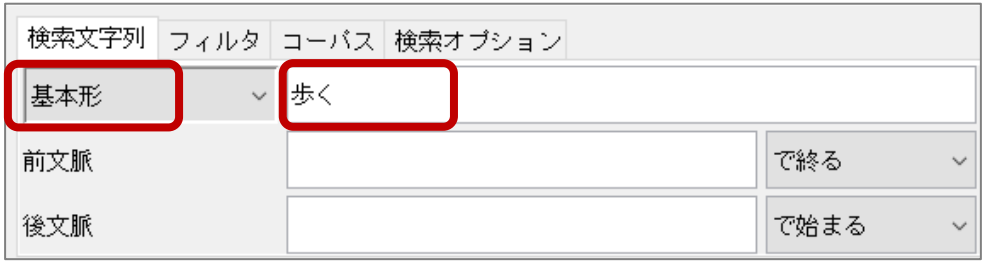


E) 単語「日」のみ



F) 活用語の基本形

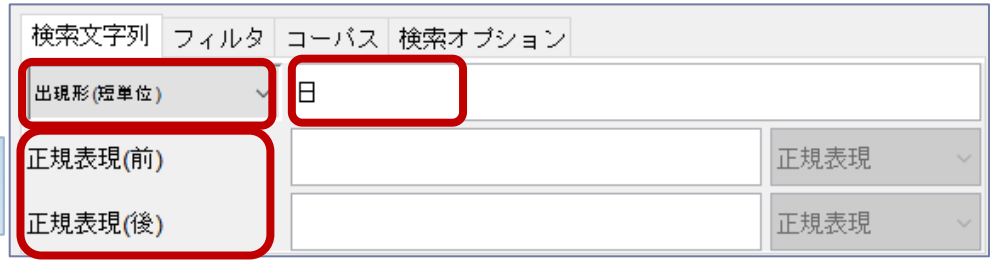
すべての語形を一括して検索



補足：単語での検索
(「正規表現(前)」「正規表現(後)」となる場合)

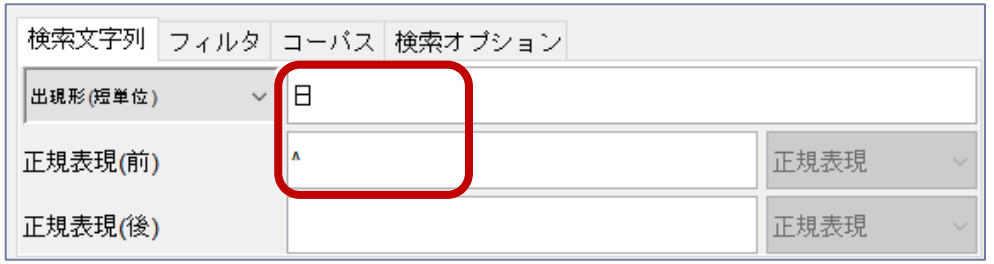
A) 「日」を含む単語

インターフェイスが
変わることに注意



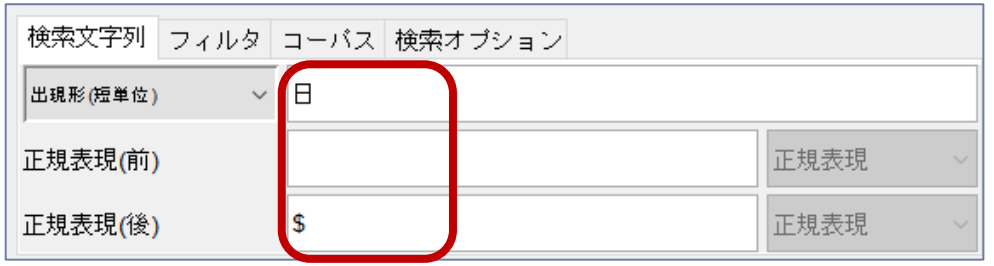
B) 先頭が「日」の単語

正規表現の「^日」と同義
(先頭の文字が「日」)



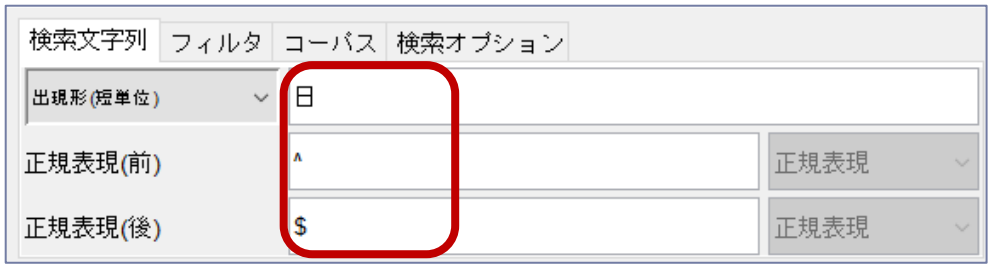
C) 末尾が「日」の単語

正規表現の「日$」と同義
(末尾の文字が「日」)



D) 単語「日」のみ

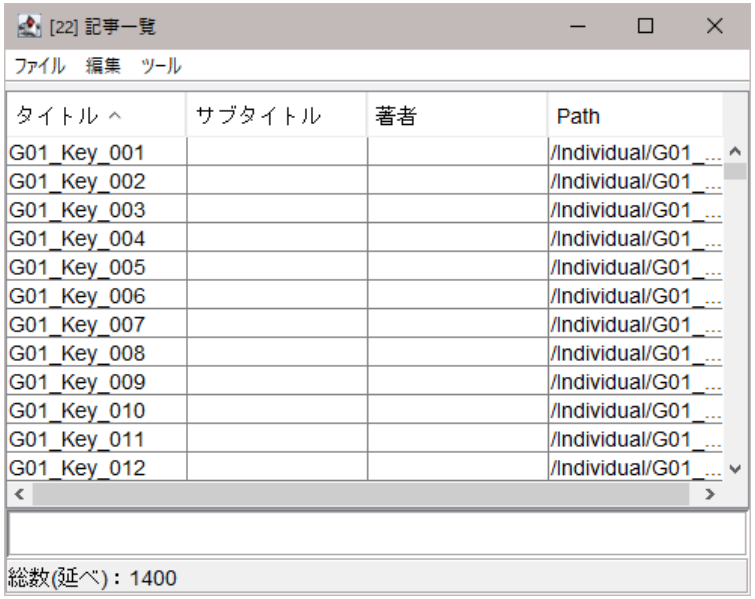
正規表現の「^日$」と
同義



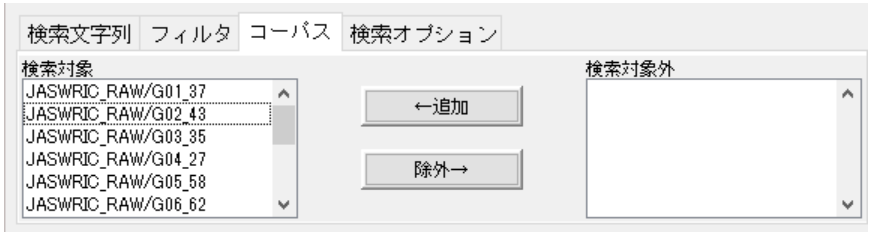
収録内容の確認

❑ 作文一覧

（[ツール]→ [一覧]→ [タイトル・著者]



❑ サブコーパス

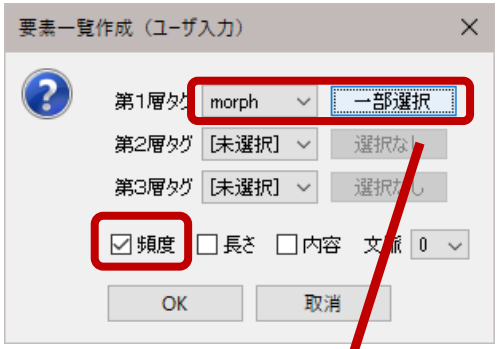


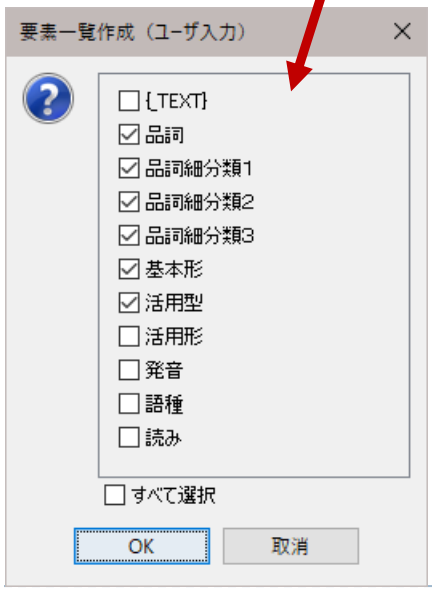
❑ XMLデータ

- インポート結果のXMLファイルの場所
  ⇒Corpora/JASWRIC/サブコーパス名/copus.xml
- 個々の作文をXMLに変換した結果を連結したもの
- テキストエディタで閲覧することが可能
- 形態素解析結果は，外部データベースに保持し，XMLでは記述されない

収録内容の確認

- 単語一覧(s)　　([ツール]→ [一覧]→ [ユーザ入力])





チェックした属性を組として，
頻度が集計される

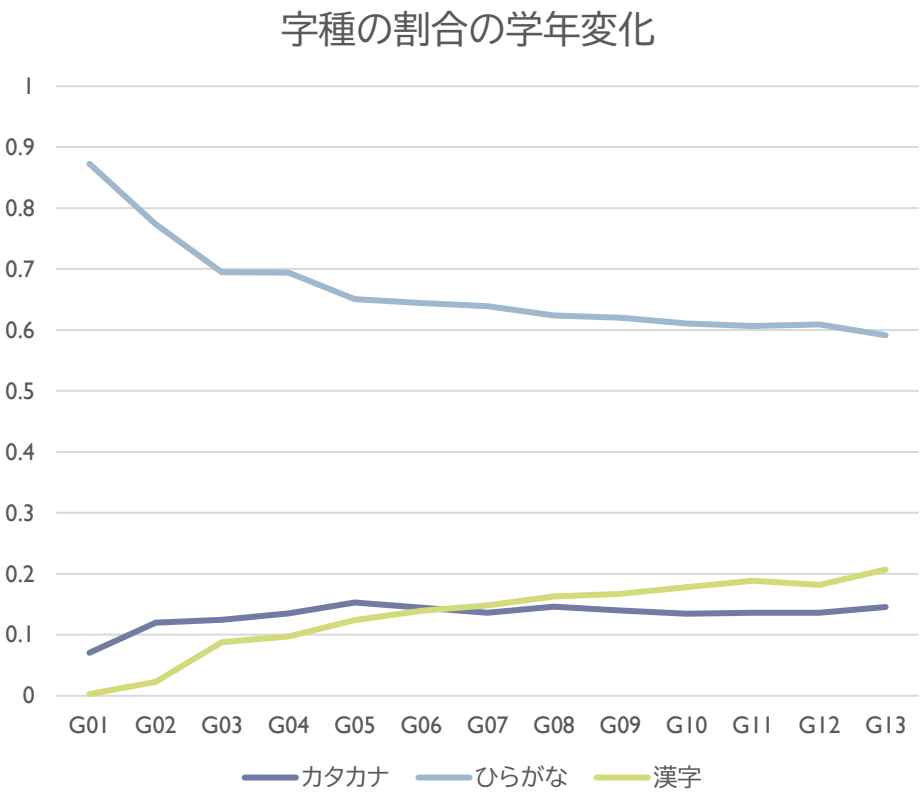
- 集計する属性にチェックを入れる
- 「活用形」ごとには集計しないので，チェックはしない
- {_TEXT}は，出現形

JASWRICを用いた簡単な分析

― 学年ごとに字種の割合を調べてみる ―

目標

- 次のようなグラフを作ります

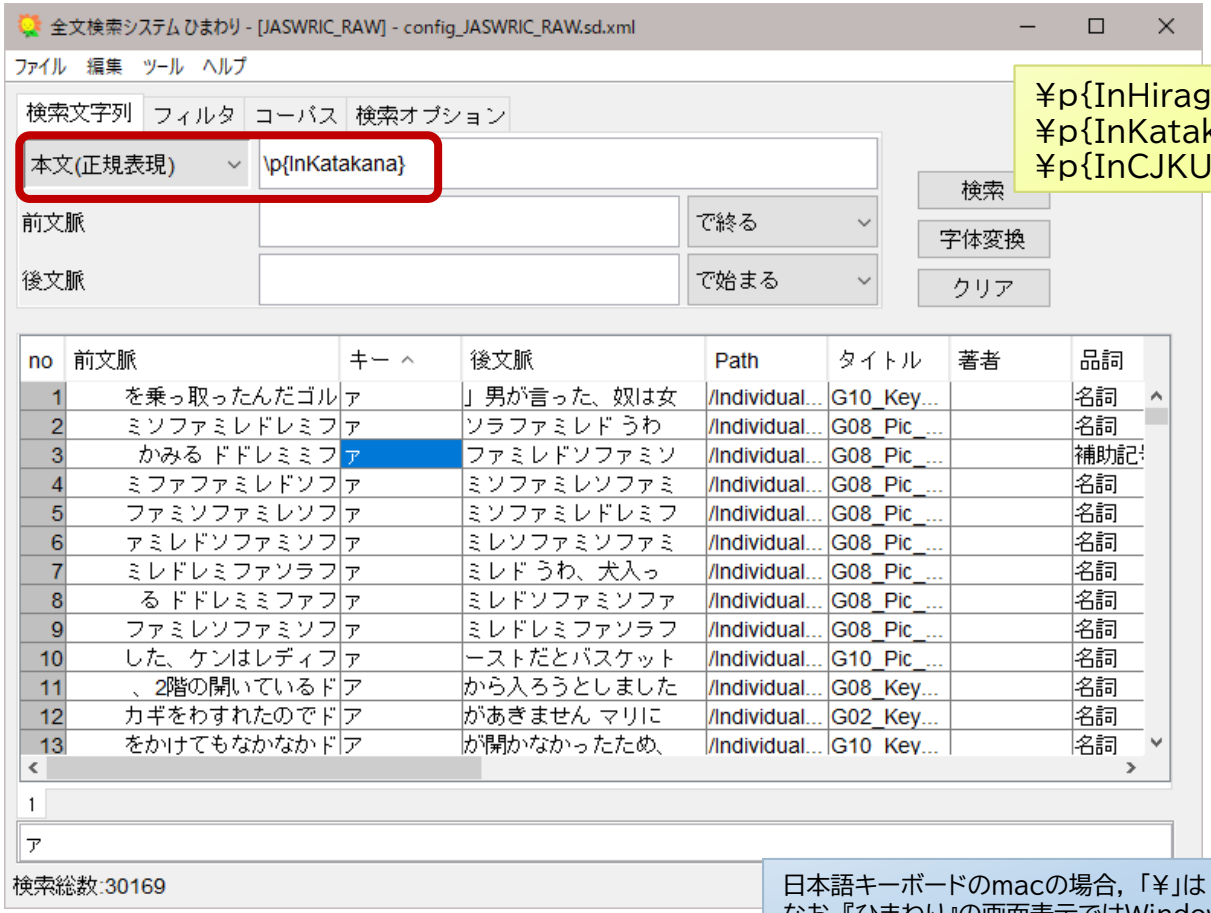


手順

1. 字種Aを指定して全文検索(漢字, ひらがな, カタカナ)
2. 学年別(G1～G13)に頻度 f_{Gi} を集計
3. 学年別に総文字数 f_s を集計
4. 学年別に字種Aの割合を計算(f_{Gi} / f_s)
5. 1～4をすべての字種に対して行う

手順1　字種を指定して全文検索

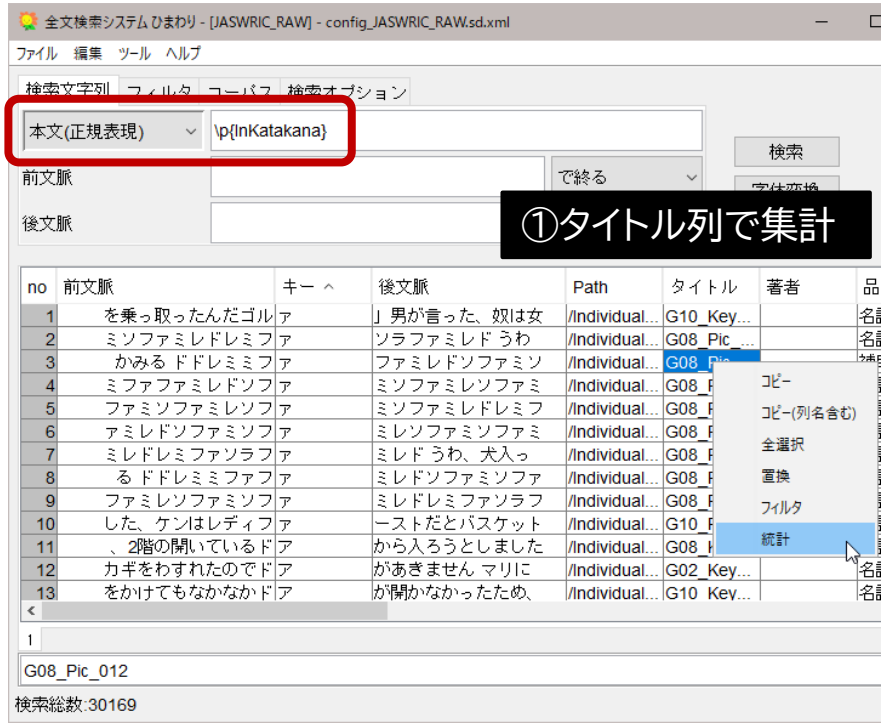
- 正規表現で字種を指定して全文検索(各字種ごとに個別に検索)

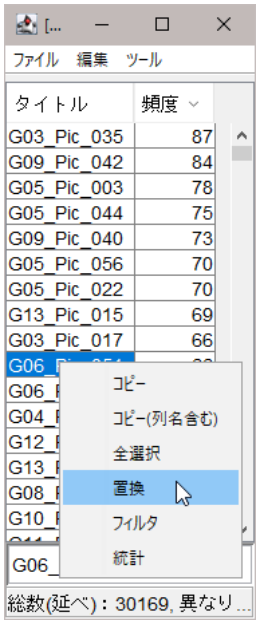


¥p{InHiragana}　… ひらがな
¥p{InKatakana}　… カタカナ
¥p{InCJKUnifiedIdeographs} … 漢字

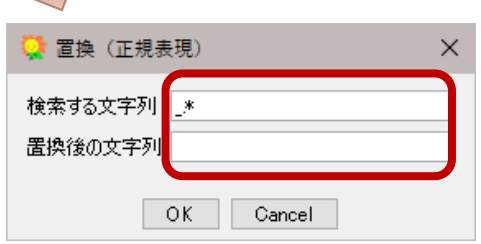
日本語キーボードのmacの場合，「¥」は optionキー+「¥」を使用
なお,『ひまわり』の画面表示ではWindows，macとも逆スラッシュ「＼」になる

手順2　学年ごとに頻度を計測



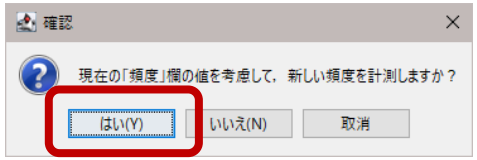


②学年情報(G01～G13)以外を置換で削除

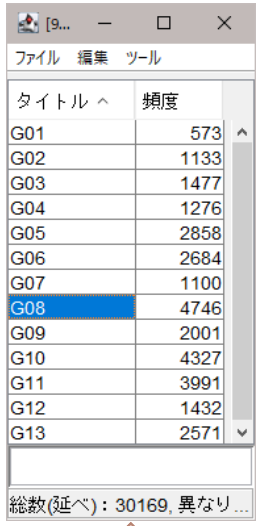


正規表現 「_.*」
　…　「_」+長さ0以上の文字列

⇒「_」以降を削除

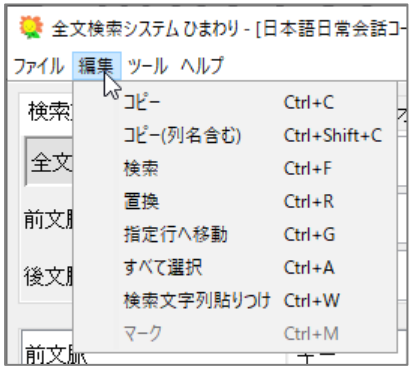


③再集計

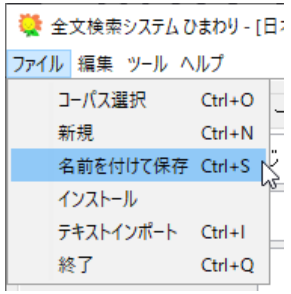


補足： 結果のエクスポート

- 方法1：クリップボードを使用する方法
  1. 結果を選択
     - 全選択したい場合は，Ctrl キー ＋ A
  2. 選択範囲をコピー
     - 通常のコピー：
       Ctrl キー ＋ C
     - 列名を含めたコピー：
       Ctrl キー ＋ Shift キー ＋ C
  3. Excel などにペースト



- 方法2：［ファイル］ ⇒ ［名前を付けて保存］
  - タブ区切りのテキストとして保存



Excelは，日付や数値などを自動変換するので，
必要に応じて，コピー先のセルの書式を「文字列」にしておく

補足：「抽出」オプション

- 全数の検索が不可能な場合などに利用

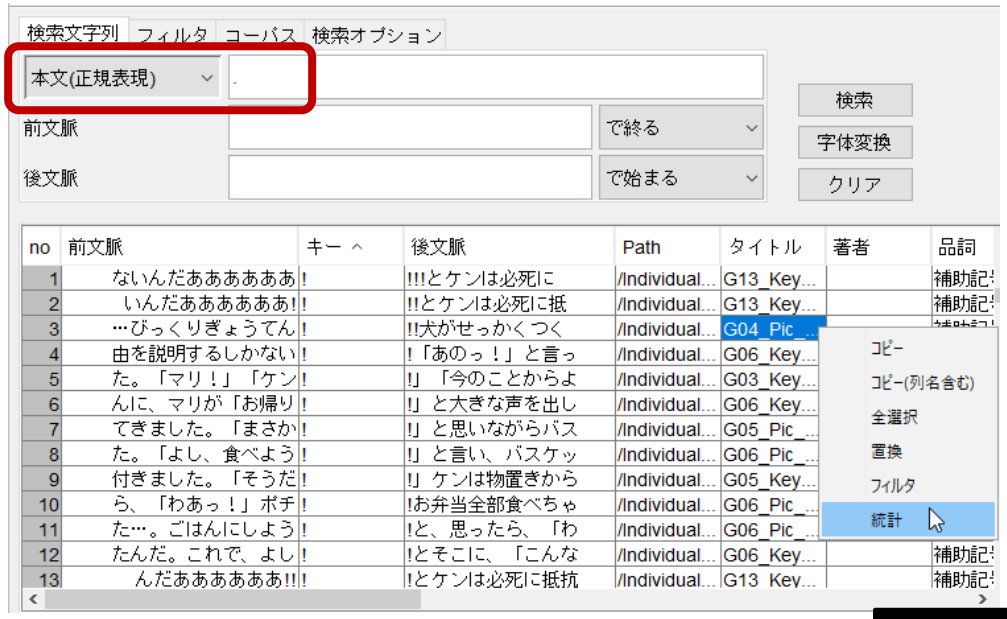
- 「頻度計測のみ（一覧）」
  - 指定した列（の組み合わせ）で頻度を計測
  - 手順
    1. 「全数」などで検索総数の少ない文字列を検索
    2. 検索結果（どの行でもよい）で，集計する列を選択
    3. 頻度計測のみ（一覧）を選択し，希望の条件で検索を実行（フィルタも使用可）

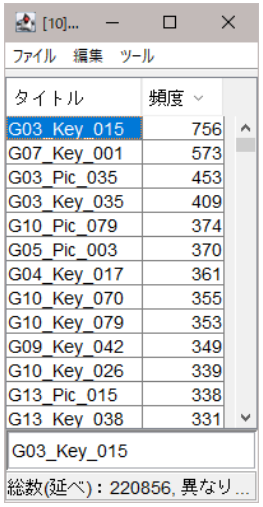
手順3　学年別に総文字数を集計

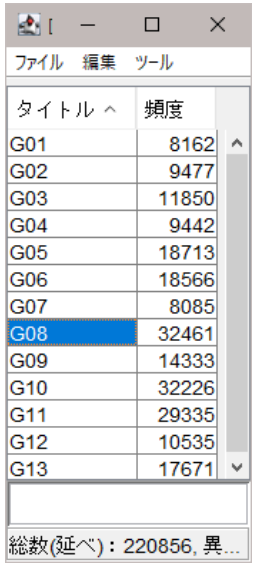
3. 学年ごとの総頻度

⇒　正規表現 「.」ですべての文字を検索し，学年ごとに集計

（指定する正規表現が異なるだけであとは同じ処理）







正規表現 「.」（半角のピリオド）
…　任意の1文字

①タイトル列で集計

②学年情報（G01～G13）以外を置換で削除
③再集計

手順4　字種の割合を計算

- 粗頻度

	カタカナ	ひらがな	漢字	総文字数
G01	573	7121	22	8162
G02	1133	7329	217	9477
G03	1477	8233	1038	11850
G04	1276	6554	917	9442
G05	2858	12170	2317	18713
G06	2684	11962	2586	18566
G07	1100	5166	1197	8085
G08	4746	20244	5291	32461
G09	2001	8891	2395	14333
G10	4327	19675	5737	32226
G11	3991	17796	5524	29335
G12	1432	6417	1913	10535
G13	2571	10447	3653	17671

②選択して，「形式を選択して貼り付け」(除算)

①選択して，コピー

- 総文字数に対する比率

	カタカナ	ひらがな	漢字	総文字数
G01	0.070203	0.872458	0.002695	8162
G02	0.119553	0.773346	0.022898	9477
G03	0.124641	0.694768	0.087595	11850
G04	0.135141	0.694133	0.097119	9442
G05	0.152728	0.65035	0.123818	18713
G06	0.144565	0.644296	0.139287	18566
G07	0.136054	0.638961	0.148052	8085
G08	0.146206	0.623641	0.162996	32461
G09	0.139608	0.620317	0.167097	14333
G10	0.13427	0.610532	0.178024	32226
G11	0.136049	0.606647	0.188307	29335
G12	0.135928	0.609112	0.181585	10535
G13	0.145493	0.591195	0.206723	17671

おわりに

- 既存の作文コーパスを『ひまわり』で活用する方法を紹介
  - 全文検索システム『ひまわり』の基本的な使い方
  - 『ひまわり』によるJASWRICの利用方法
  - JASWRICを使った簡単な分析

- さらに詳しく知るには
  - 『ひまわり』ホームページ
  - チュートリアルビデオ
  - 利用者マニュアル
  - 研究発表
  - 過去の講習会

補足：申込み時のコメントを受けて

- 発話データの集計
  - 『ひまわり』で利用可能なパッケージ
    - 名大会話コーパス, 昭和話し言葉コーパス, 国会会議録
    - 日本語日常会話コーパス, 日本語話し言葉コーパス
    - 自作したい場合は, 第13回コーパス利用講習会の資料を参照
  - タグの集計機能の利用
    - [ツール]→ [一覧]→ [ユーザ入力] （本資料p.21）
    - 第14回コーパス利用講習会の資料 （p.20～）

- 文脈情報を利用した絞り込み
  - 前後文脈欄の利用
    - 第14回コーパス利用講習会の資料p.15
  - 前後の単語の利用
    - 検索結果の「基本形-2」「基本形-1」「基本形1」「基本形2」列で絞り込み