

国立国語研究所学術情報リポジトリ

全文検索システム『ひまわり』講習会

メタデータ	言語: Japanese 出版者: 公開日: 2022-08-24 キーワード (Ja): キーワード (En): 作成者: 山口, 昌也 メールアドレス: 所属:
URL	https://doi.org/10.15084/00003654



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。

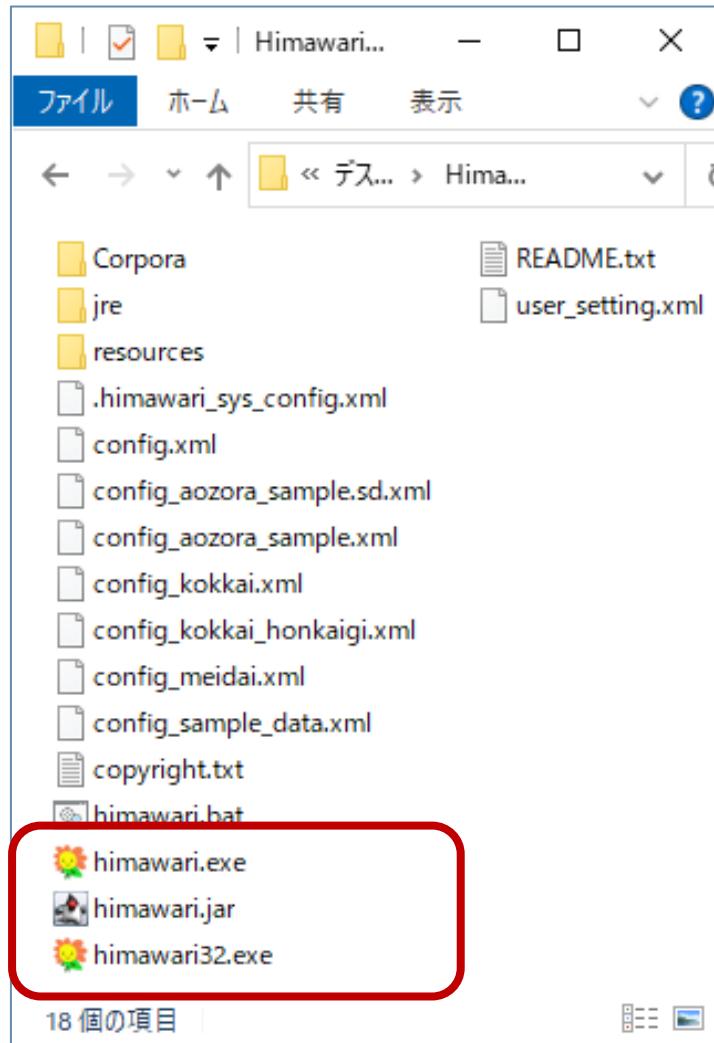
『ひまわり』とは

- ▶ 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

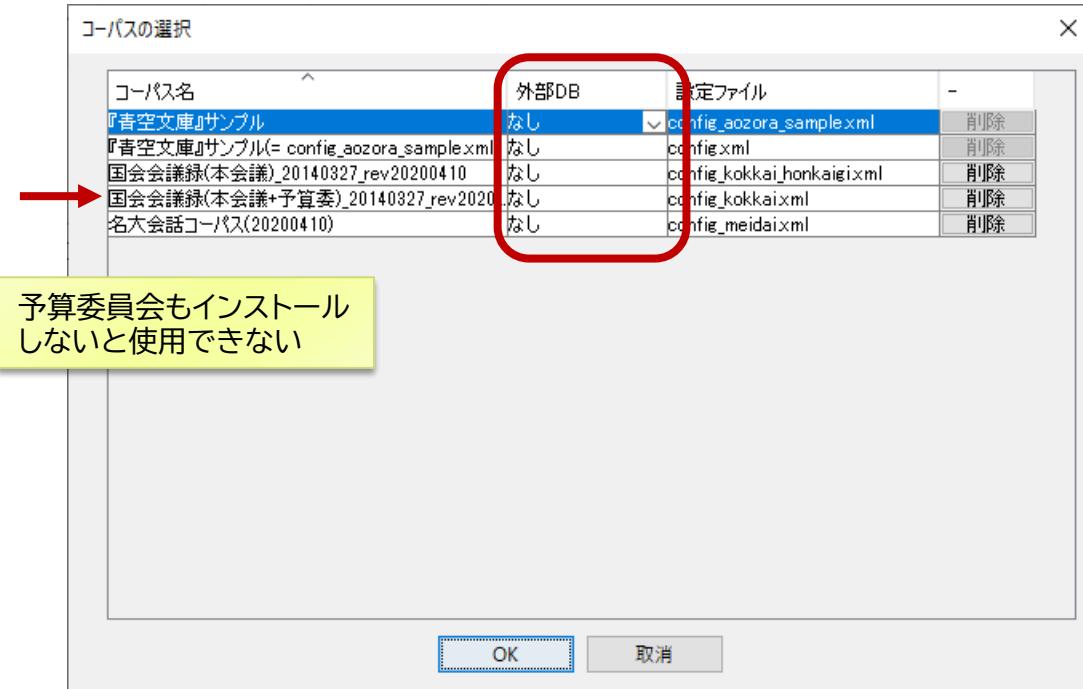


himawari.jar

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

```
C:\WINDOWS\system32>  
C:\Users\masaya\Desktop\ProgramData\Oracle\em32\Wbem;C:\WINDOWS\System\Microsoft\Windows  
C:\Users\masaya\Desktop\t.jar  
osname: windows 10  
osname: windows 10  
osname: windows 10  
info:  
info: _sys _pred  
info: _sys _key  
info: _sys _full  
info: 記事 path  
info: 記事 タイ  
info: 記事 著者  
corpusname aozora_sample  
info(available memory  
history added: これ  
doSearch:start  
corpusname aozora_sample  
open_eix  
eix:記事  
n:597  
info(available memory  
time[milisec]: 65  
time[milisec]: 65
```

「検索文字列」欄では右クリックで履歴表示

検索の実行

検索結果

途中経過の表示

検索総数

検索文字列

本文

これ

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	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	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	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	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	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	はすこぶる不憧だよ。これ	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	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	これ	からが結論だぜ。一	/aozora_s...	吾輩は猫...	夏目漱石
12	とーと息ついた。「これ	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	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	これ	からさき彼を相手にす	/aozora_s...	こころ	夏目漱石

検索総数:597

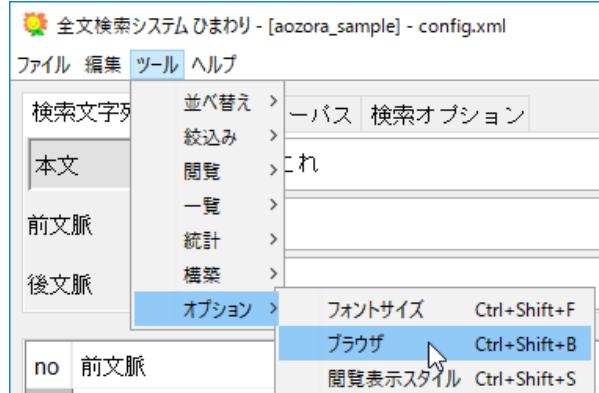
ブラウザでの閲覧

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	これ	からいよいよヴァイオ	/aozora_s...
4	い話があるかい」「これ	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	これ	からいよいよ会津領へ	/aozora_s...
6	見当がつかない」「これ	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	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



検索キーは、赤い字で表示

吾輩は猫である：夏目漱石

「これからいよいよ弾くところです」

「**これ**からいよいよヴァイオリンを弾くところだよ。こっちへ出て来て、聞きたまえ」

「まだヴァイオリンかい。困ったな」

「君は無絃の素琴を弾ずる連中だから困らない方なんだが、寒月君のは、きいきいぴいぴい近所合壁へ聞えるのだから 大に困ってるところだ」

「そうかい。寒月君近所へ聞えないようにヴァイオリンを弾く方を知らんですか」

「知りませんね、あるなら伺いたいもので」

「伺わなくとも 露地の白牛を見ればすぐ分るはずだが」と、何だか通じない事を云う。寒月君はねばけてあんな珍語を弄するのだろうと鑑定したから、わざと相手にならないで話題を進めた。

「ようやくの事で一策を案出しました。あくる日は天長節だから、朝から

[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

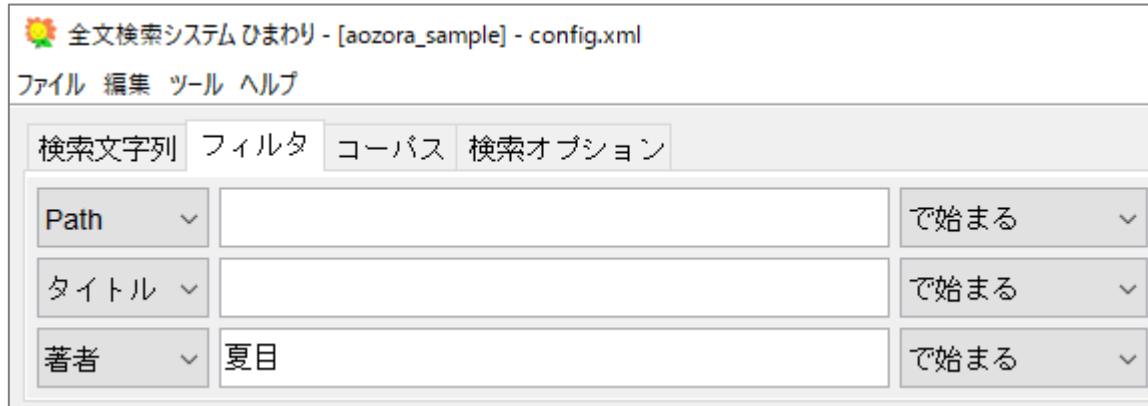
▶ 複数列を考慮したい場合

▶ 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

no	前文脈	キー	後文脈	Path	タイトル	著者	操作
1	として、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	[文字列指定] [置換]
	これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目漱石	夏目漱石
	」「これ		からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目漱石	夏目漱石
	」「これ		からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目漱石	夏目漱石
	て、これ		からいよいよ会津領へ	/aozora_s...	吾輩は猫...	夏目漱石	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	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	夏目漱石

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

I. 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	くらいで充分淨土へ行	/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

検索文字列 フィルタ コーパス 検索オプション

本文 明日

前文脈

後文脈

検索

字体変換

クリア

当該作品の形態素一覧
⇒Shift + ダブルクリック

no	前文脈	キー	後文脈	Path	タイトル	著者	品詞
1	。「御前は大和かい。明日	明日	ね、行くんだからね、	/aozora_s...	吾輩は猫...	夏目漱石	名詞
2	鳥部教授歓迎会、其又	明日	は……」 うるさ	/aozora_s...	吾輩は猫...	夏目漱石	名詞
3	芋、今日はステッキ。明日	明日	は何になるだろう。	/aozora_s	吾輩は猫	夏目漱石	名詞
4	言						
5	学協						

一 視 ツール

一覧 ファイル 編集 ツール

テキスト進行方向

SER.NO. _TEXT 品詞 品詞細... 品詞細... 品詞細... 活用型 活用形 基本形 読み 発音

SER.NO.	_TEXT	品詞	品詞細...	品詞細...	品詞細...	活用型	活用形	基本形	読み	発音
00021784	部	名詞	接尾	一般				部	ブ	ブ
00021785	教授	名詞	一般					教授	キョウジ	キョージ
00021786	歓迎	名詞	サ変接続					歓迎	カンゲイ	カンゲイ
00021787	会	名詞	接尾	一般				会	カイ	カイ
00021788	、	記号	読点					、	、	、
00021789	其又	名詞	一般					*	*	*
00021790	明日	名詞	副詞可能					明日	アシタ	アシタ
00021791	は	助詞	係助詞					は	ハ	ワ
00021792	…	記号	一般					…	…	…
00021793	…	記号	一般					…	…	…
00021794	トヨ	記号	トヨ	トヨ				トヨ	トヨ	トヨ

00021790

総数(延べ): 206322

検索文字列 フィルタ

出現形

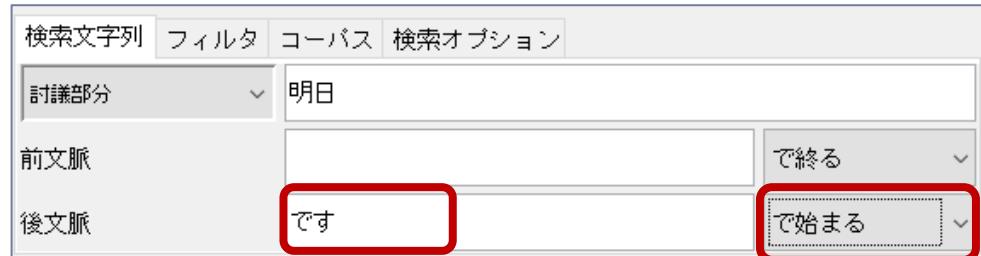
ルビ(rt)完全一致
ルビ(rt)部分一致
出現形

品詞
活用型
活用形
基本形
読み

さまざまな検索と各種機能

前後文脈の制限

- 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

The screenshot shows a search interface with the following settings:

- 検索文字列: いただく
- 検索オプション: 前文脈を含む (selected), 後文脈を含む (unchecked)
- 前後文脈長: 3 文字
- 検索範囲: 10 文字

The results table displays 9 entries where the word 'いただく' is preceded by three characters. The first entry (no. 4) is highlighted with a blue background.

no	前文脈	キー ^	後文脈	議院	回	会議名
4	入れて	いただく	、ある	参議院	150	本会議
5	力して	いただく	、ある	衆議院	001	本会議
6	あけで	いただく	、ある	衆議院	093	本会議
7	審議を	いただく	、かよ	参議院	034	本会議
8	申して	いただく	、かよ	参議院	058	本会議
9	かって	いただく	、かよ	参議院	051	本会議

▶ 方法2: 前後文脈の検索条件

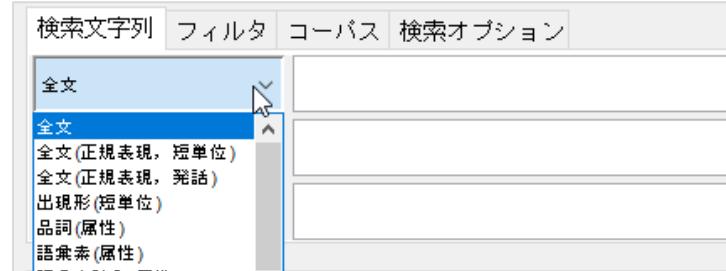
- ① 前文脈に正規表現「...」(任意の3文字)
- ② 「前文脈を含む」をチェック (検索オプション)

The screenshot shows a search interface with the following settings:

- 検索文字列: いただく
- 検索オプション: 前文脈を含む (selected)
- 前文脈: ...
- 後文脈: で始まる

本文の正規表現検索

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

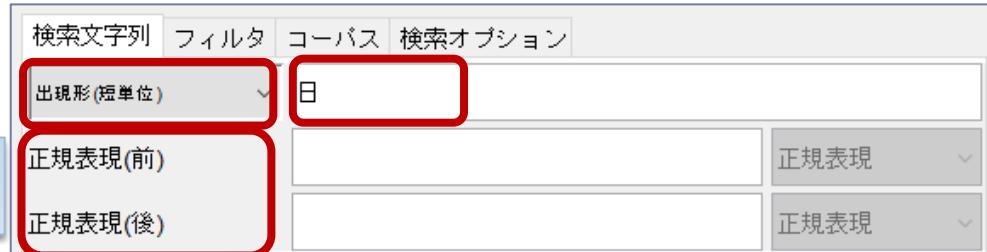
.	(ピリオド)	… 任意の1文字
[わいうえ]	… 「わ」「い」「う」「え」のいずれか	
[^。！]	… 。！以外の任意の1文字	
*	… 直前要素の0個以上の繰り返し	
+	… 直前要素の1個以上の繰り返し	
+?	… 直前要素の1個以上の繰り返し(最短)	
{0,5}	… 直前要素の0～5個の繰り返し	
()	… マッチした範囲を記録	
¥1	… 1個目の記録した要素	

(詳細は本資料末の参考文献参照)

単語(タグ)での検索

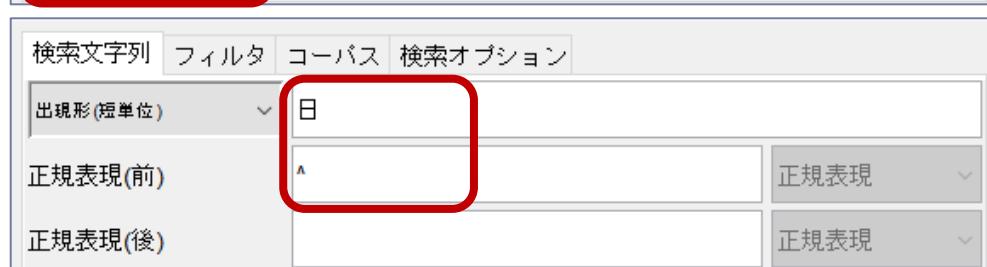
A) 「日」を含む単語

インターフェイスが
変わることに注意



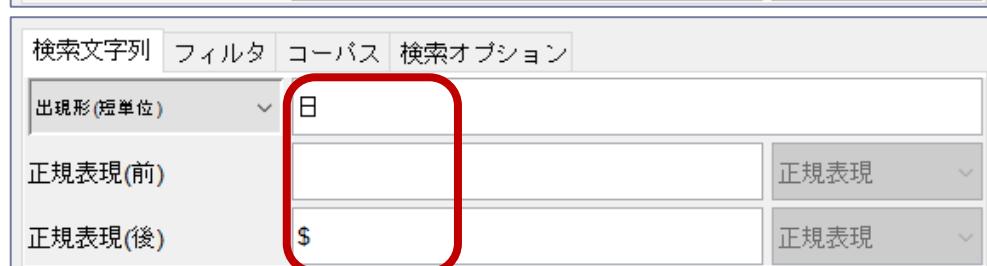
B) 先頭が「日」の単語

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



各種応用例

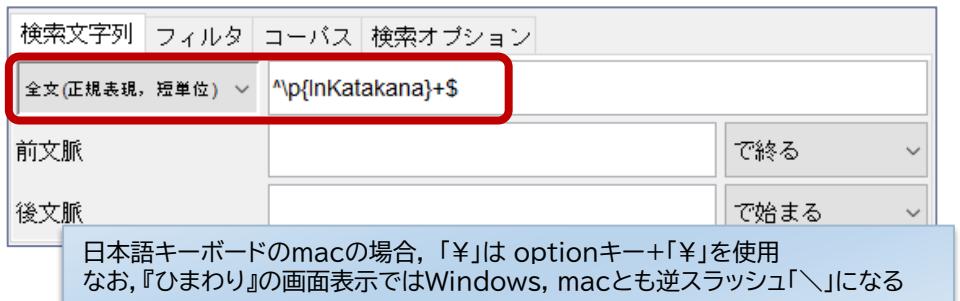
A) 共起語の集計(「～に行く」)

- 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



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）

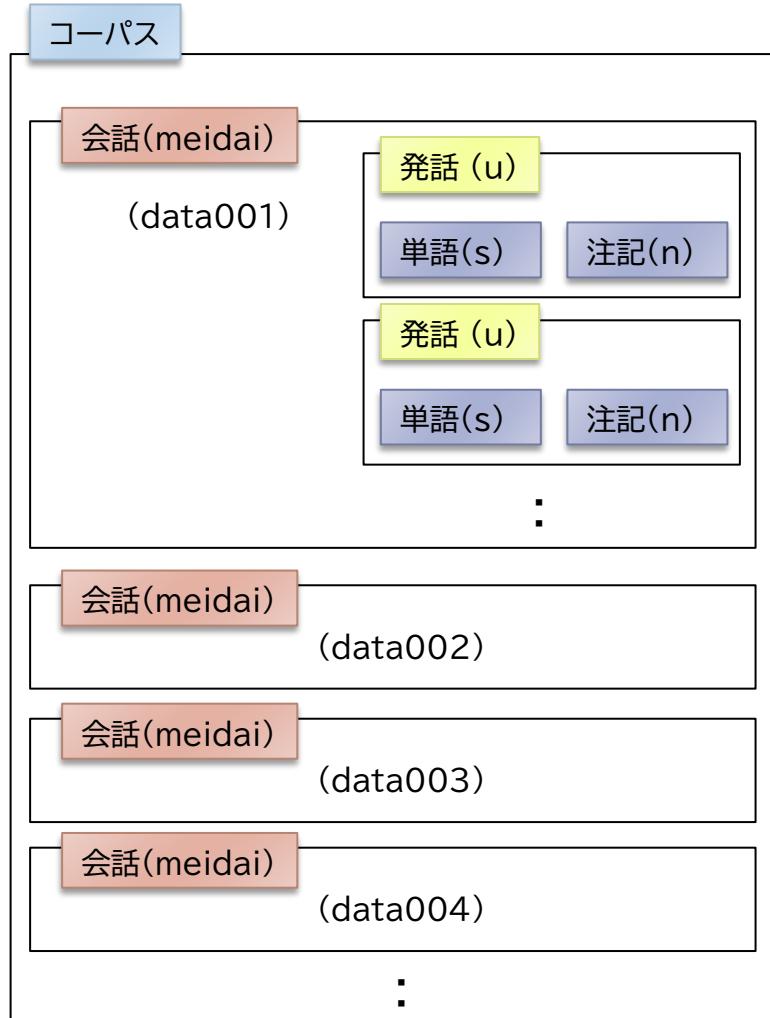
The screenshot shows a search interface with tabs: 検索文字列, フィルタ, コーパス, 検索オプション. The '検索文字列' tab is active. In the search input field, the word 'なんか' is selected. Below the search bar, there are tabs for '全文(正規表現, 発話)', '抽出', and '字体'. Under '全文(正規表現, 発話)', a red box highlights the search input field. To the right, a results table lists words and their frequencies. At the bottom, there are search options: '全数' (All), 'ラベル', '抽出数上限' (Number of Extracted Items), '頻度計測のみ' (Frequency Measurement Only) (which is selected), '表示方法' (Display Method), '一覧' (List), and '総合' (Overall). A note at the bottom says '結果が出るまで数分かかります！！'.

キー	頻度
なんか	8163
うん。	7890
うん、	6687
、そう	5996
ってい	5582
だから	5436
ていう	5399

総数(延べ): 1918742, 異な...

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

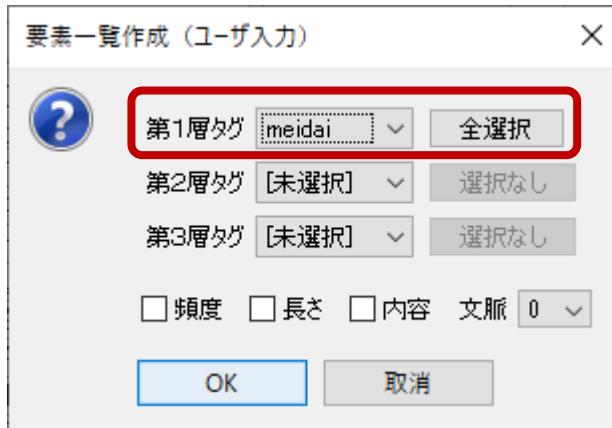
タグの集計

▶ [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

□ 会話中(meidai)の発話数(u)

要素一覧作成 (ユーザ入力)

第1層タグ: meidai ▾ 一部選択

第2層タグ: u ▾ 選択なし

第3層タグ: [未選択] ▾ 選択なし

頻度 長さ 内容 文脈 0

OK 取消

要素一覧作成 (ユーザ入力)

データ名 収録日 収録時間

短単位数 収録場所 話者関係

補足情報 全話者 すべて選択

OK 取消

[46] 一覧: meidai ファイル 編集 ツール

	頻度
meidai/@データ...	895
data001	1583
data003	977
data004	1048
data005	1758
data006	1636
data007	916
data008	2617

総数(延べ): 173296 異なり : 129

□ 発話者一覧(u)

要素一覧作成 (ユーザ入力)

第1層タグ: u ▾ 一部選択

第2層タグ: [未選択] ▾ 選択なし

第3層タグ: [未選択] ▾ 選択なし

頻度 長さ 内容 文脈 0

OK 取消

要素一覧作成 (ユーザ入力)

話者年齢 話者出生地 相づち

[id] 話者情報 話者居住地 話者 話者性別

すべて選択

OK 取消

[49] 一覧: u ファイル 編集 ツール

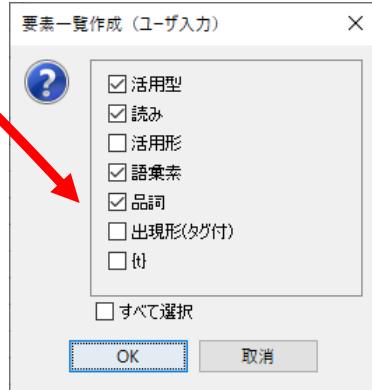
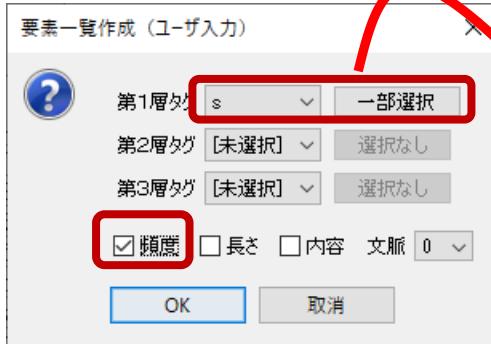
u/@話者	u/@話者出...	u/@話者性別	頻度
F001	山梨県北巨摩...	女性	2746
F002	兵庫県伊丹市	女性	1393
F003	栃木県宇都宮...	女性	435
F004	兵庫県西宮市	女性	10366
F005	愛知県名古屋...	女性	1263
F006	神奈川県横須...	女性	172
F007	東京都	女性	603
F008	愛知県知多市	女性	461
F009	愛知県名古屋...	女性	429
F010	千葉県市川市	女性	139
F011	大阪府堺市	女性	1338
F013	岡山県邑久郡	女性	619

総数(延べ): 173296 異なり : 204

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

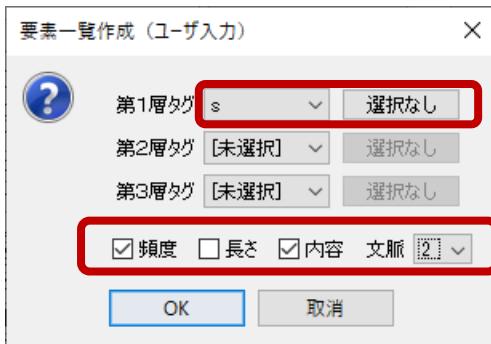
□ 単語一覧(s)



[50] 一覧: s

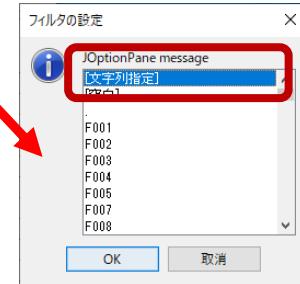
s/@品詞	s/@活用型	s/@読み	s/@語彙素	頻度
himawari_発話末				173296
補助記号-読点		、		138009
補助記号-句点		。		115036
助動詞	助動詞-ダ	ダ	だ	53667
感動詞一般		ウン	うん	41010
助動詞	助動詞-タ	タ	た	30117
助動詞-接続助詞		テ	て	29685
助動詞-終助詞		ネ	ね	29584
助詞-準体助詞		ノ	の	26394
助詞-副助詞		カ	か	23664
助詞-格助詞		ト	と	20931
助詞-格助詞		ノ	の	20699
助詞-係助詞		モ	も	20034
代名詞				
総数(延べ): 1595687, 異なり: 18530				

□ 単語 n-gram



[58] 一覧: s

s%内容	s[1]%内容	s[2]%内容	頻度
よ	ね	。	3705
*	*	*	3447
ん	なん	か	3156
ん	だ	けど	2725
な	あの	、	2342
で	ん	だ	2048
なん	も	、	2031
た	か	、	2005
た	ん	だ	1916
だ	で	も	1901
うん	よ	ね	1821
だ	、	うん	1768
	から		1747
総数(絞りこみ前, 後): 1595687, 1114062 / 異なり (絞り ...)			

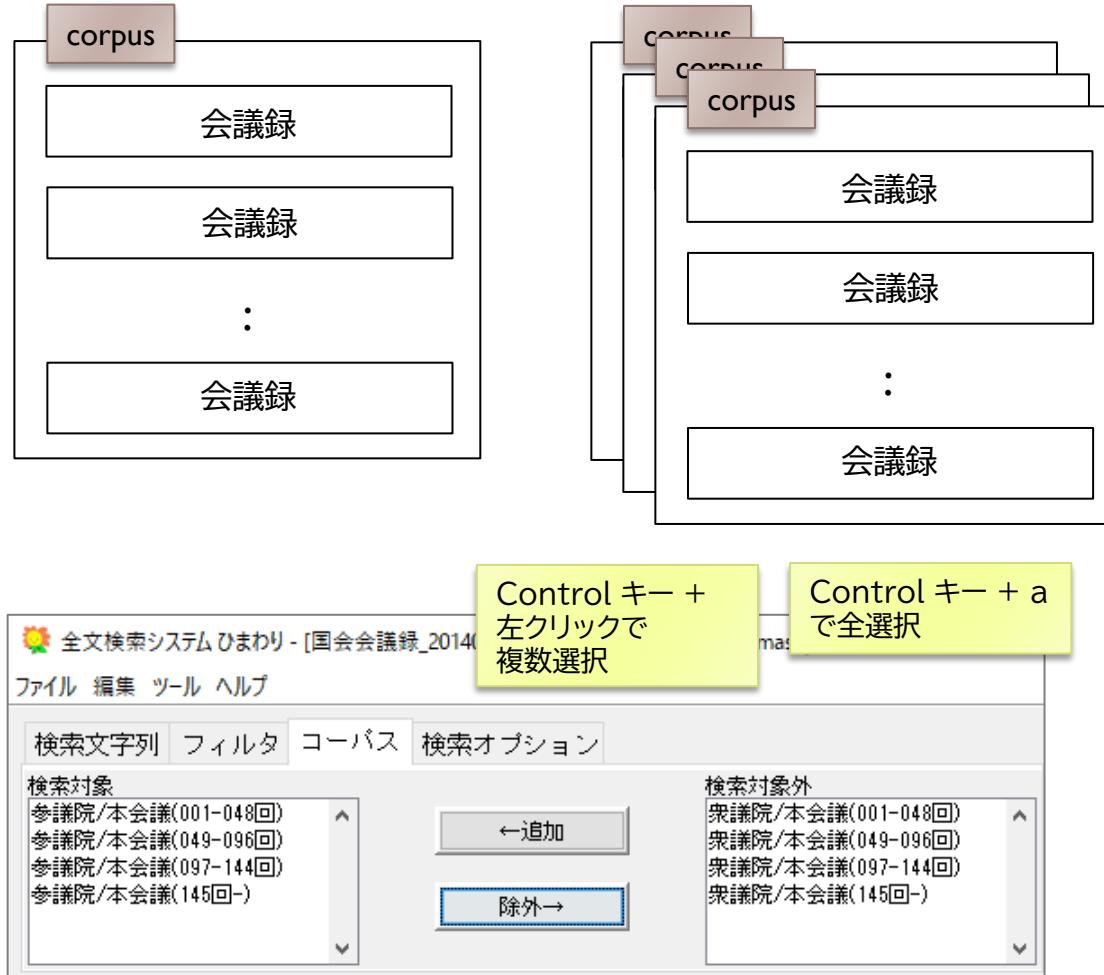


- 3-gram の場合、「文脈」は2
- 「内容」オプションにより出現形が計測対象となる

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 → 「.」
- 「himawari_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

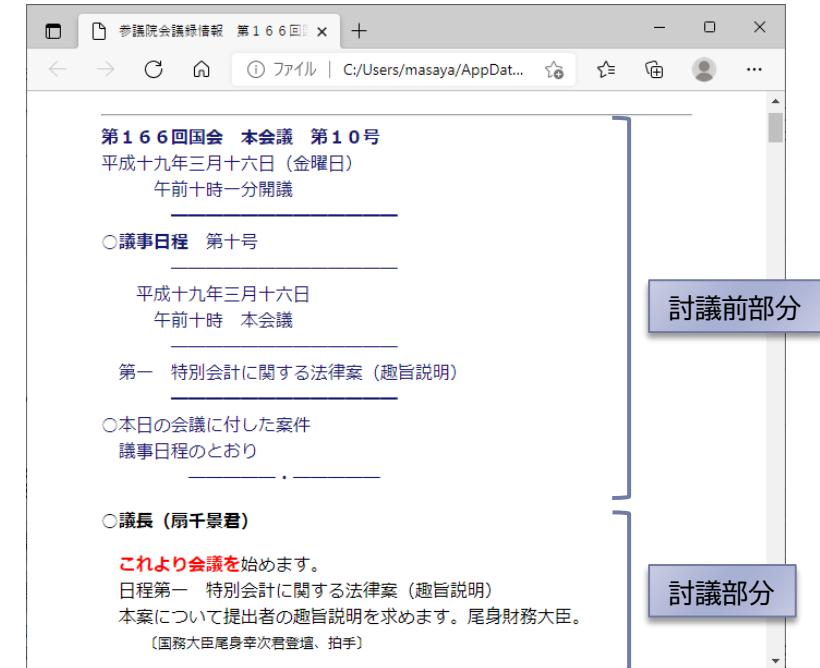
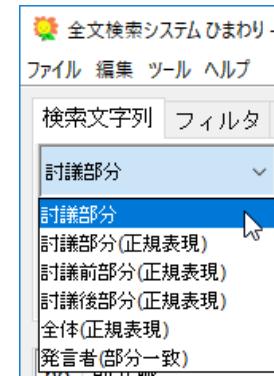
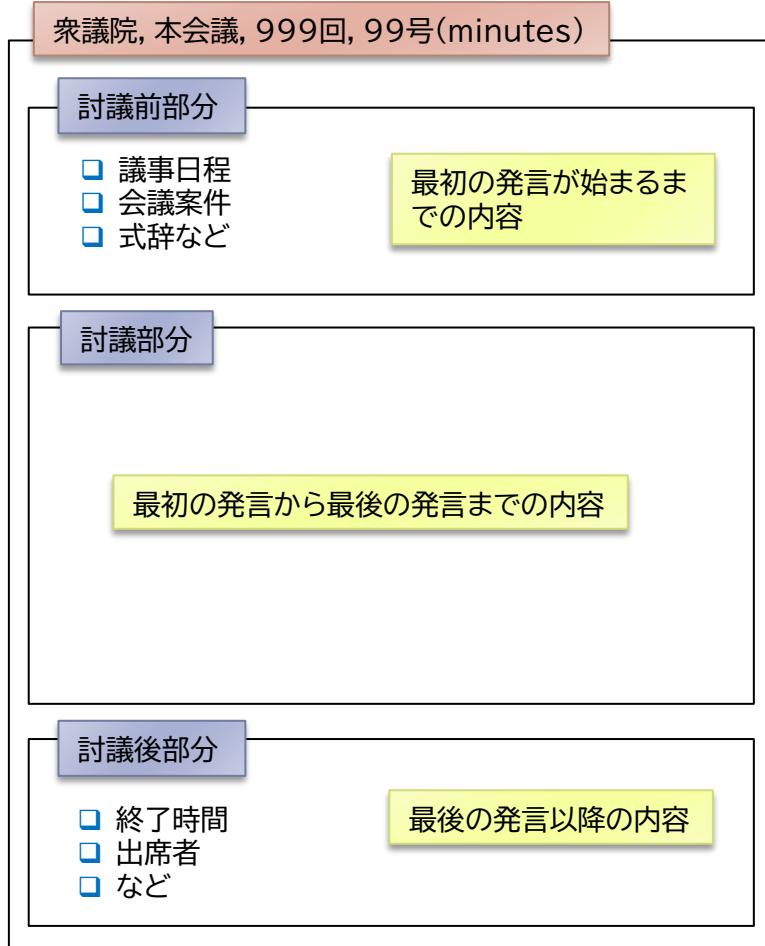
- ▶ 参議院/本会議
 - 001-048回
 - 049-096回
 - 097-144回
 - 145回-
- ▶ 衆議院/本会議
 - 001-048回
 - 049-096回
 - 097-144回
 - 145回-

▶ 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



コーパスの構造と検索(国会会議録)

▶ 討論部分

討論部(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
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。

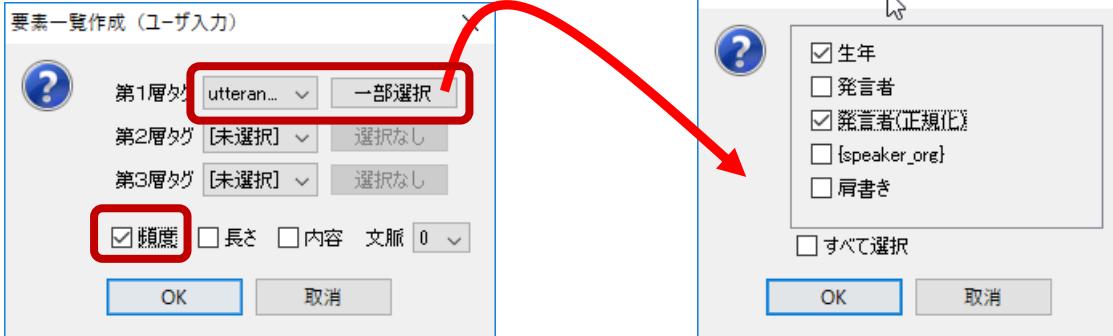
赤下線、赤枠部分は、付属要素として、
本文の全文検索から除外

- ▶ 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)

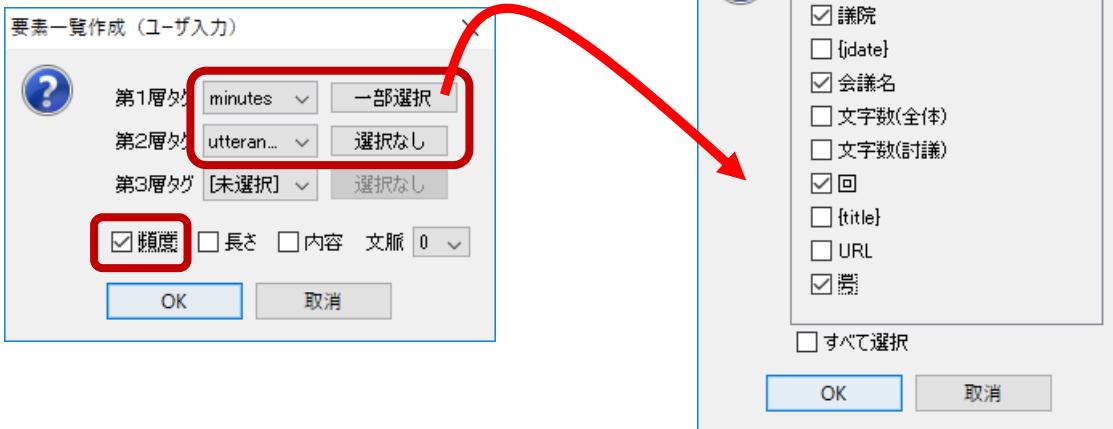
国会

□ 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)

□ 議事録ごとの発言数



応用例：表記の経年変化

1 「条件」と「條件」の検索

全文検索システム hinomari - [国会会議録_20140327_rev20170201] - C:\Users\masay

検索文字列: [條件]

検索オプション: 検索, 字体変換 (highlighted)

※常用漢字の旧字体に変換

キー	後文脈
条件	七、
条件	に従う
条件	人類類
条件	、
条件	、一期
条件	、ある
条件	、ある

2 「キー」「開催日」のセルを選択し、「統計」

開催日 文

開催日	文
1954-02-17	コピー
1982-	コピー(列名含む)
1999-	全選択
1953-1	フィルタ
1968-4	統計
1955-1	統計
1956-6	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)

置換 (正規表現) dialog:

- 検索する文字列: -.* (highlighted)
- 置換後の文字列: (empty)
- OK button (highlighted)

キー	開催日	頻度
条件	2003-06-13	55
条件	1959-04-03	44
条件	1963-07-01	42
条件	1953-08-07	41
条件	1988-05-11	16
条件	1992-03	11
条件	1984-06	10
条件	1951-1	9
条件	1955-07	8
条件	1950-11	7
条件	1985-06	6
条件	1974-05	5

4 再集計

「キー」「開催日」を選択

キー	開催日	頻度
条件	2010	57
条件	2003	44
条件	1988	43
条件	2003	43
条件	2012	43
条件	1963	43
条件	1959	43
条件	1953	43
条件	2009	43
条件	1956	43

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】

青空文庫の独自タグ
(3種類)

《》: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 | 歪《ひき》の

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)

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 | 歪《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。』
『クラムボンはかぶかぶわらったよ。』
『クラムボンは跳《は》ねてわらったよ。』
『クラムボンはかぶかぶわらったよ。』

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 | 歪《ひき》の

ルビ, 注記は, 本文とは区別され,
全文検索の対象外となる

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]

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)

二 | 歪《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。』

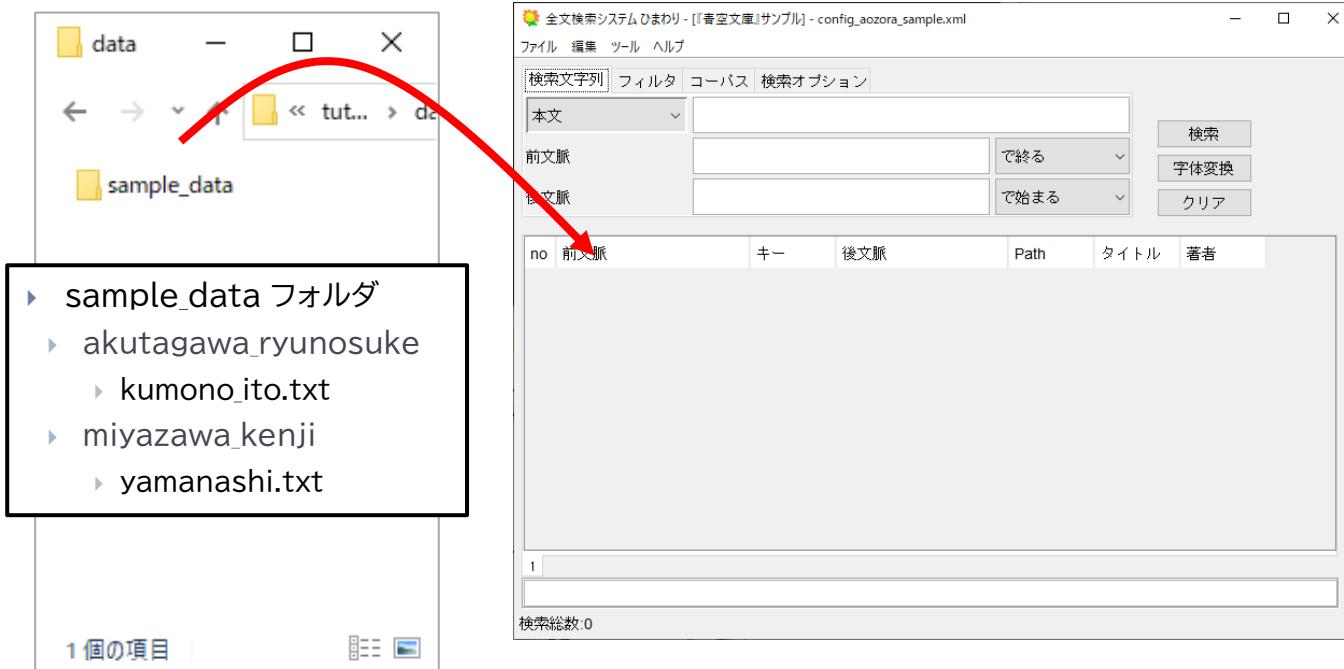
『クラムボンはかぶかぶわらったよ。』

『クラムボンは跳《は》ねてわらったよ。』

『クラムボンはかぶかぶわらったよ。』

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

全文検索システムひまわり - [sample_data] - config_sample_data.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 検索オプション

本文 蟹の子供

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	一、五月	二疋の蟹の子供	らが青じろい水の底で	/sample_d.../yamanashi	yamanashi	
2	て来るだけです。	蟹の子供	らは、あんまり月が明	/sample_d.../yamanashi	yamanashi	
3	二、十二月	蟹の子供	らはもうよほど大きく	/sample_d.../yamanashi	yamanashi	
4	泡が流れて行きます。	蟹の子供	らもぼっぽっぽっとつ	/sample_d.../yamanashi	yamanashi	

1 蟹の子供
検索総数: 4

- 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

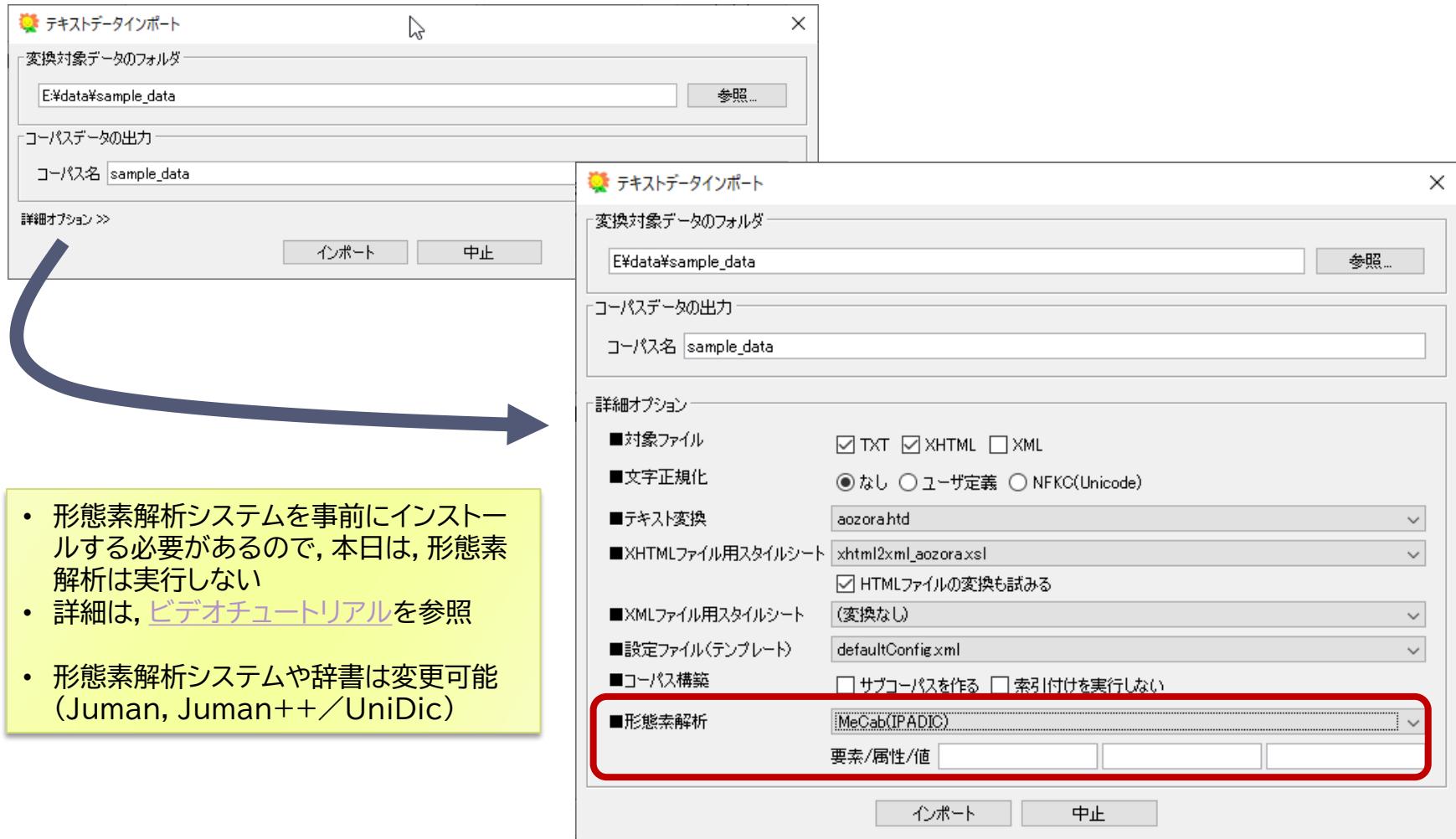
yamanashi :

小さな谷川の底を写した二枚の青い 幻燈 です。
げんとう

3字下げ 一、五月 # 「一、五月」は中見出し

二疋の蟹の子供らが青じろい水の底で話していました。
ひき かに 二疋の蟹の子供
『クラムボンはわらったよ。』
『クラムボンはかぶかぶわらったよ。』
『クラムボンは跳ねてわらったよ。』
『クラムボンはかぶかぶわらったよ。』
上の方や横の方は、青くくらく 銅 のように見えます。そのなめらかな
てんじょう 天 井 を、つぶつぶ暗い 泡 が流れて行きます。
つぶつぶ泡が流れて行きます。蟹の子供らもぼっぽっぽっとつづけて五
つぶつぶ泡を吐きました。それはゆれながら水銀のように光って 斜めに上
の方へのぼって行きました。
つうと銀のいろの腹をひるがえして、一疋の魚が頭の上を過ぎて行きました。

インポート時の形態素解析



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](#)