

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(コンコーダンス)
- ▶ 『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

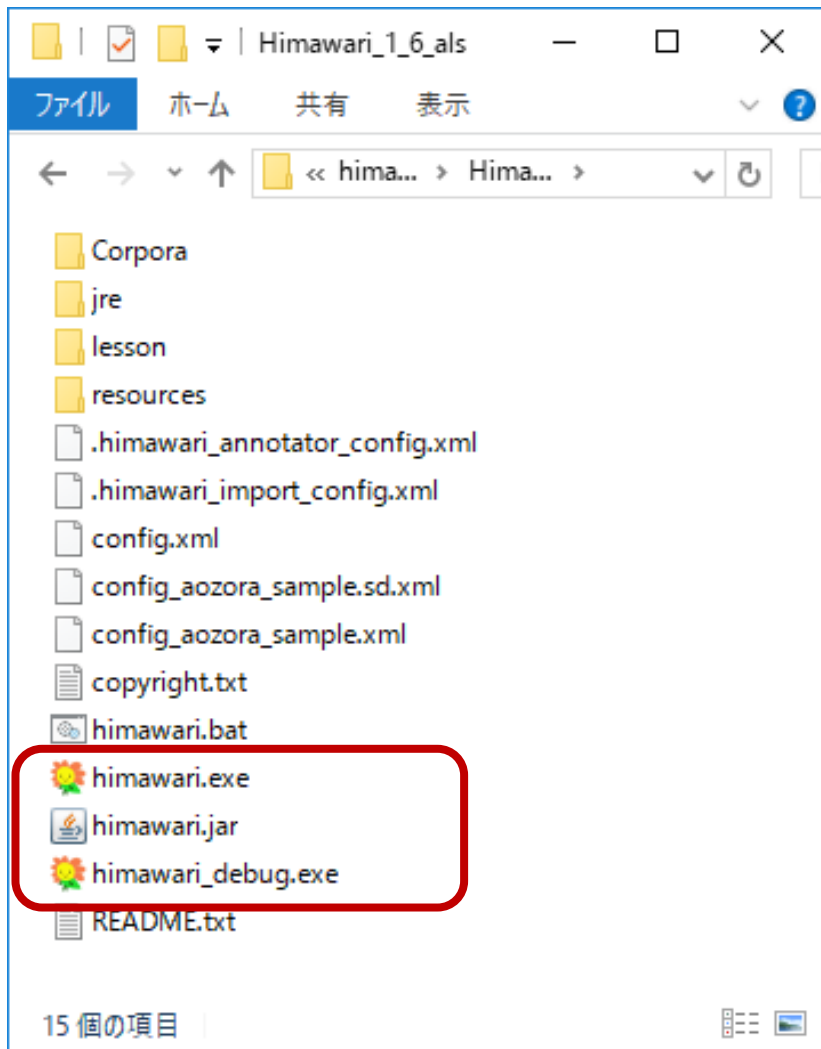
- ▶ XMLでタグづけされたコーパスを全文検索できます
- ▶ Windows, Mac OS, Linux など、多くのOS上で動作します
- ▶ 無料です

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

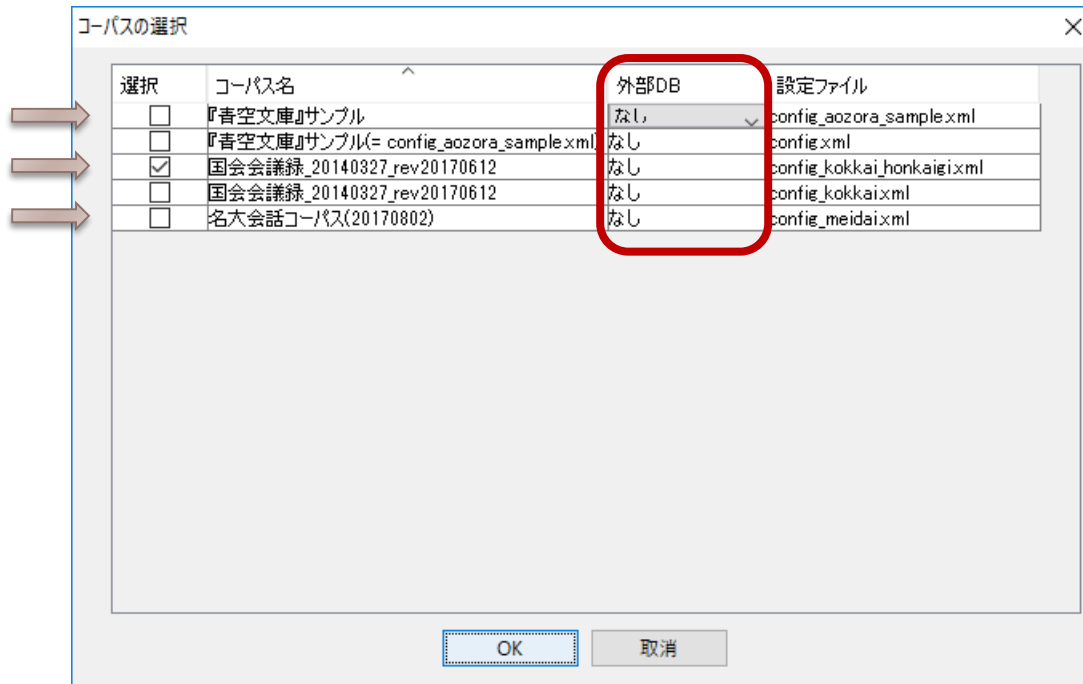


himawari.jar

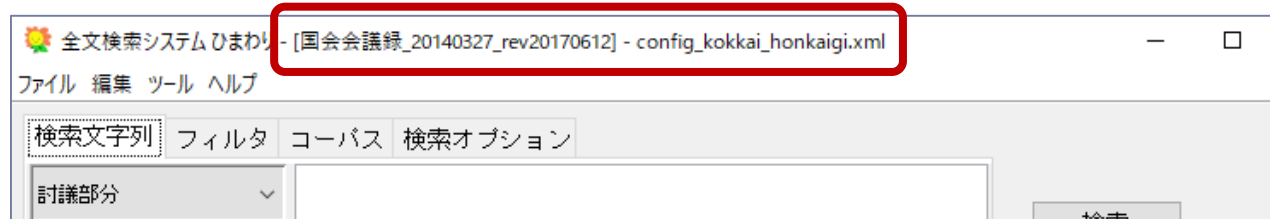
汎用
(Windows, Mac, Linux など)
himawari.jar

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

検索文字列 フィルタ コーパス 検索オプション

本文 検索文字列

前文脈 検索の実行

後文脈

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	弾くところですよ」	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	めちゃんお困ります。	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。一	/aozora_s...	吾輩は猫...	夏目漱石
12	と一と息ついた。「	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

検索総数: 597

途中経過の表示

検索総数

検索結果

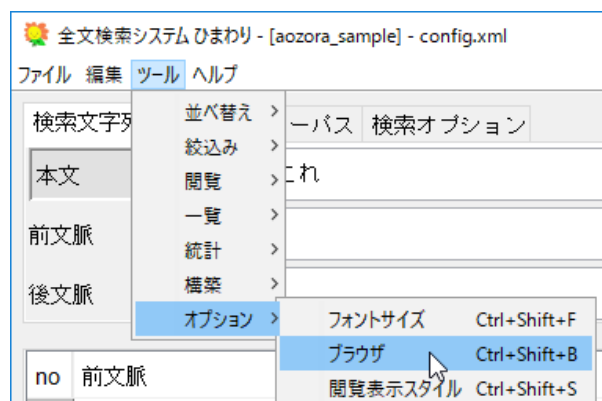
ブラウザでの閲覧

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	めちゃんお困ります。	これ	からがいよいよ佳境に	/aozora_s...

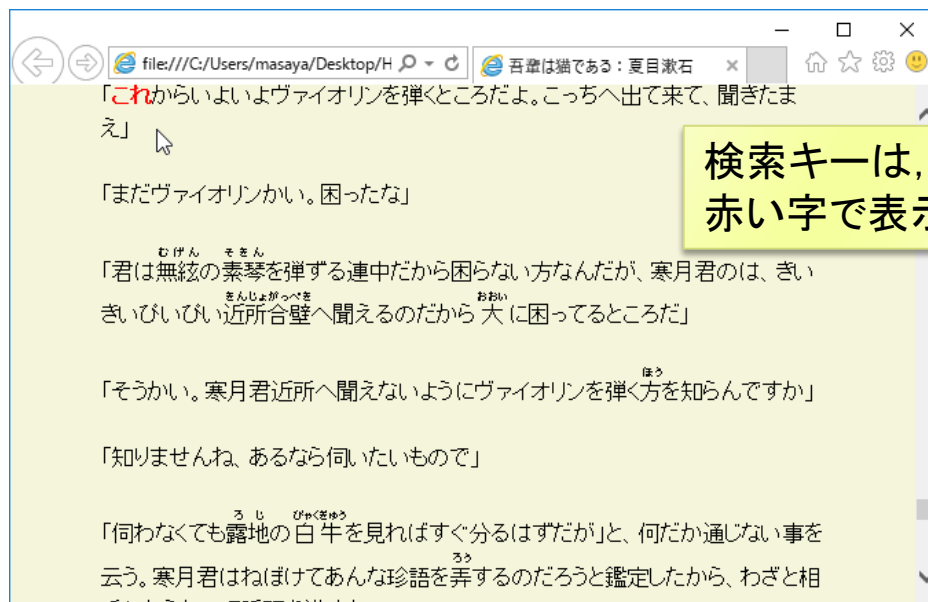
閲覧したい用例をダブルクリック



■ 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	めちゃんお困ります。	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。一	/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

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 検索オプション

Path		で始まる
タイトル		で始まる
著者	夏目	で始まる

「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

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	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	かい」「これ	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	これ	からいよいよ会津領へ	/aozora_s...	吾輩は猫...	夏目漱石
	ない」「これ	これ	からいよいよ弾くところ	/aozora_s...	吾輩は猫...	夏目漱石
	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

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

[文字列指定]
[置換]
夏目漱石
芥川龍之介

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

検索文字列	フィルタ	コーパス	検索オプション
全文		私	
前文脈			で終る
後文脈		は	で始まる

- B) 正規表現で
「は」で「始まる」を表現

^ ... 文字列の先頭を表す

検索文字列	フィルタ	コーパス	検索オプション
全文		私	
前文脈			で終る
後文脈		^は	正規表現

- C) 「私」
+ 格助詞, 係助詞

検索文字列	フィルタ	コーパス	検索オプション
全文		私	
前文脈			正規表現
後文脈		^[はが]にをへで	正規表現

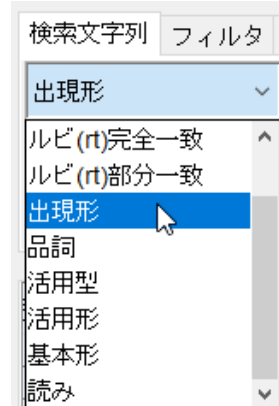
(「検索オプション」の「後文脈を含む」をチェックしてみましょう)

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

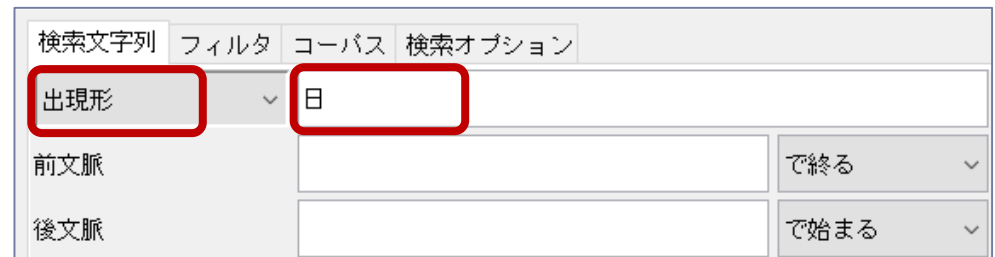
▶ 単語単位で正規表現検索

- ▶ 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-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

no	前文脈	キー	後文脈	Path
1	は取れんはずである。	一両日	の後続節の本胆はさら	/aozc
2	でございましたのに、	一昨日	コピー	なりまし /aozc
3	眼はその隙間の端に、	一昨日	コピー(列名含む)	見付け出 /aozc
4	し親子兄弟の離れたる	今日	全選択	ものはな /aozc
5	知れん、しかし太平の今	今日	フィルタ	部の中心 /aozc
6	はほっと一息ついて「今	今日	統計	単純な様 /aozc
7	静岡から出て来てね、今	今日	いっしょにたべ	へ出掛 /aozc

単語での検索(2)

C) 先頭が「日」の単語

正規表現の「^」
(文字列の先頭)

検索文字列	フィルタ	コーパス	検索オプション
出現形			<input type="text" value="^日"/>
前文脈			で終る <input type="button" value="v"/>
後文脈			で始まる <input type="button" value="v"/>

D) 末尾が「日」の単語

正規表現の「\$」
(文字列の末尾)

検索文字列	フィルタ	コーパス	検索オプション
出現形			<input type="text" value="日\$"/>
前文脈			で終る <input type="button" value="v"/>
後文脈			で始まる <input type="button" value="v"/>

E) 単語「日」のみ

検索文字列	フィルタ	コーパス	検索オプション
出現形			<input type="text" value="^日\$"/>
前文脈			で終る <input type="button" value="v"/>
後文脈			で始まる <input type="button" value="v"/>

F) 活用語の基本形

すべての語形を
一括して検索

検索文字列	フィルタ	コーパス	検索オプション
基本形			<input type="text" value="歩く"/>
前文脈			で終る <input type="button" value="v"/>
後文脈			で始まる <input type="button" value="v"/>

形態素解析結果の閲覧

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

この機能は、
config_XXX.sd.xml タイプ
の資料のみ実行可能

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	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	は何になるだろう。	/aozora_s	吾輩は猫...	夏目漱石	名詞
4							
5	学協						

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



一覧

ファイル 編集 ツール

SER.NO.	_TEXT	品詞	品詞細...	品詞細...	品詞細...	活用型	活用形	基本形	読
00021784	部	名詞	接尾	一般				部	ブ
00021785	教授	名詞	一般					教授	キ
00021786	歓迎	名詞	サ変接続					歓迎	カ
00021787	会	名詞	接尾	一般				会	カ
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
00021791	は	助詞	係助詞					は	ハ
00021792	…	記号	一般					…	…
00021793	…	記号	一般					…	…
00021794	!	記号	感嘆符					!	!

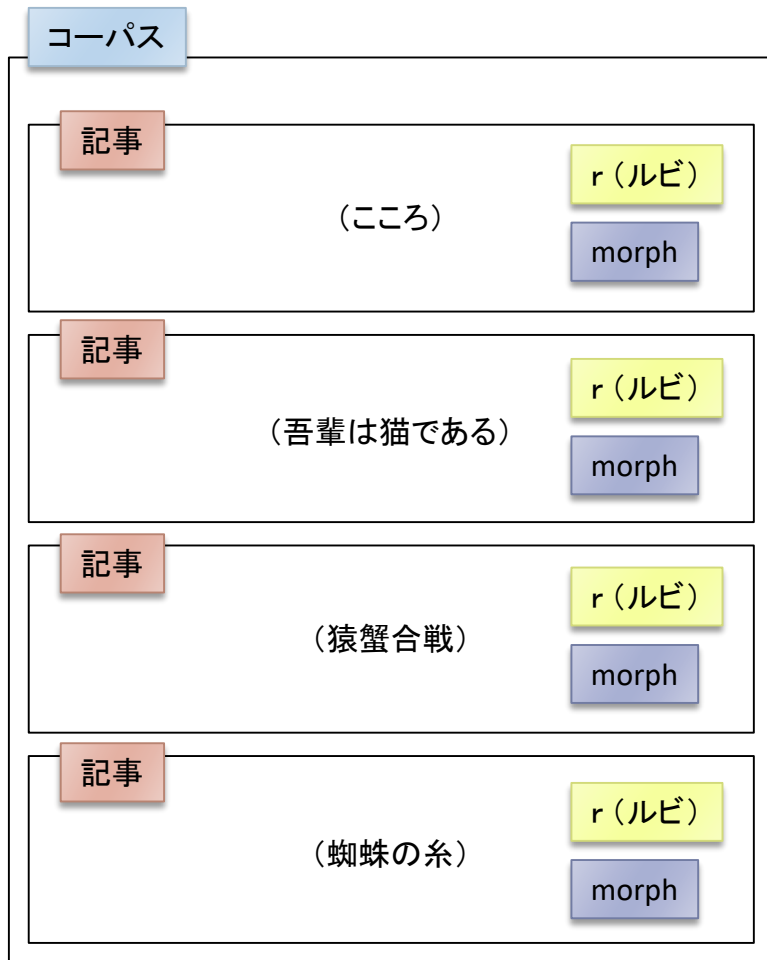
総数(延べ) : 206322

コーパスの構造と検索

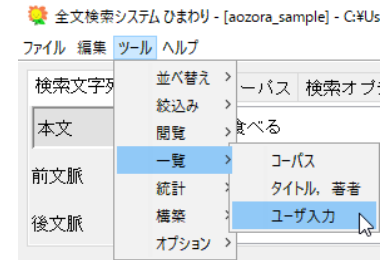


コーパスの構造と検索 (青空文庫サンプル)

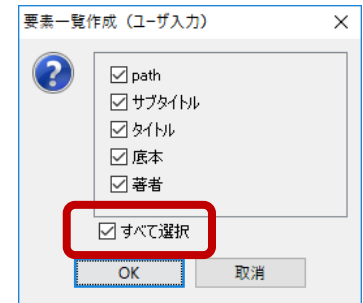
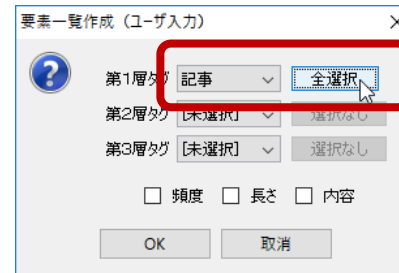
▶ コーパス全体の構造



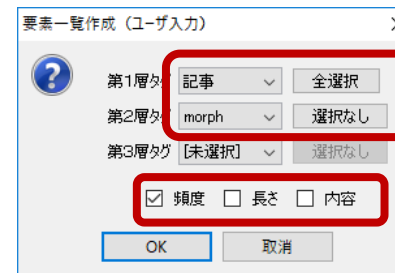
▶ 一覧機能 (ユーザ入力) で付与情報を閲覧



■ 記事



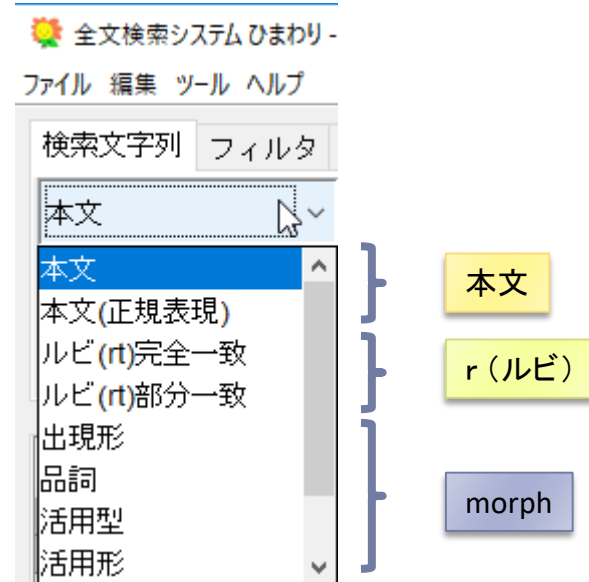
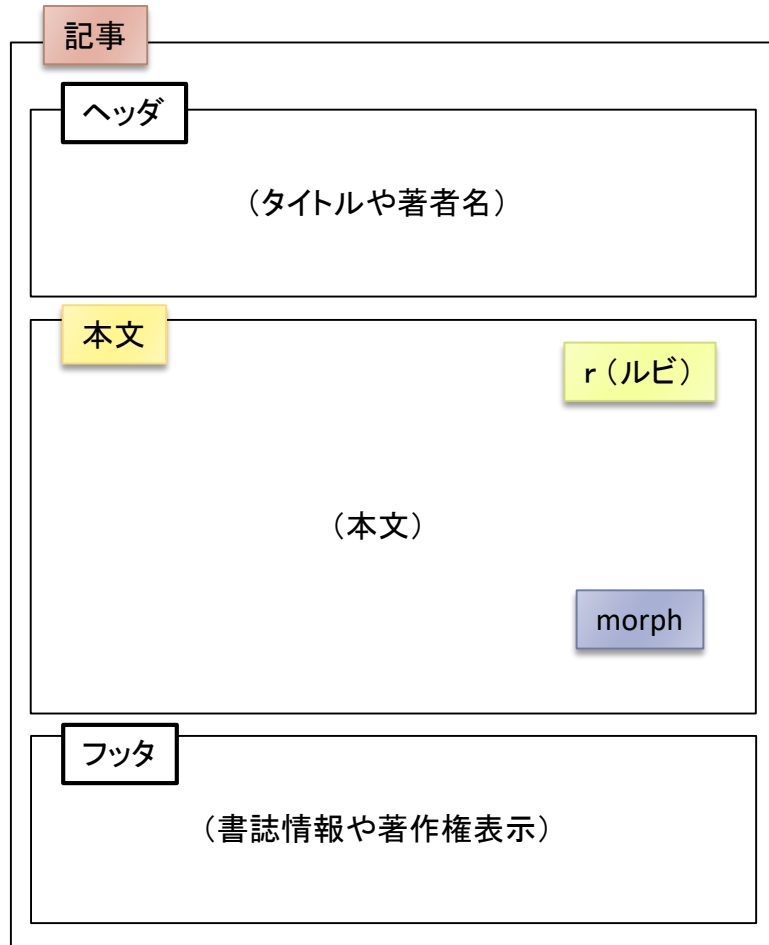
■ 記事 + morph



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

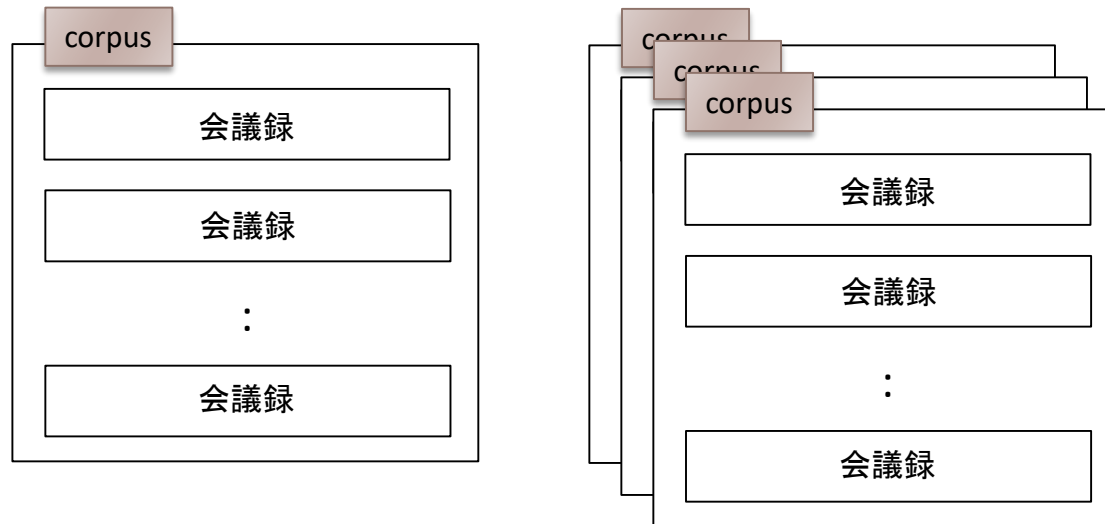
Corpora/aozora_sample/corpus.xml

ブラウザで閲覧した記事

Corpora/aozora_sample/xslt/__searched_tmp.xml

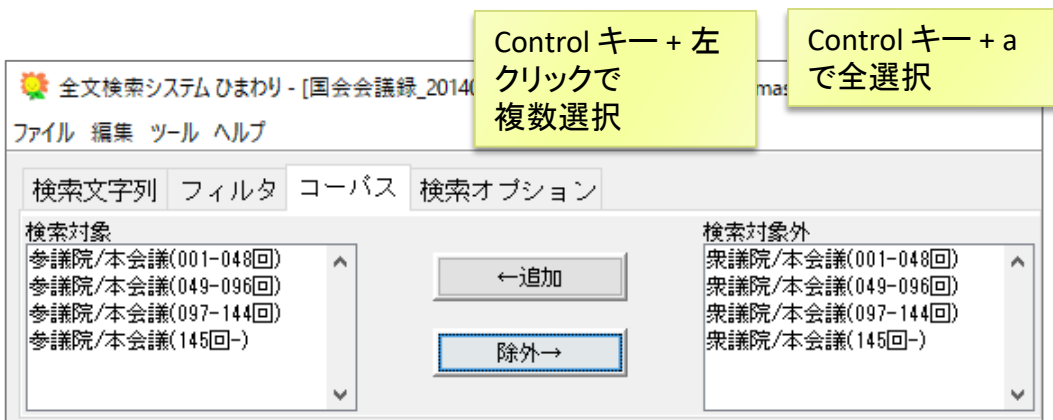
※「秀丸」などのテキストエディタを利用のこと

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-

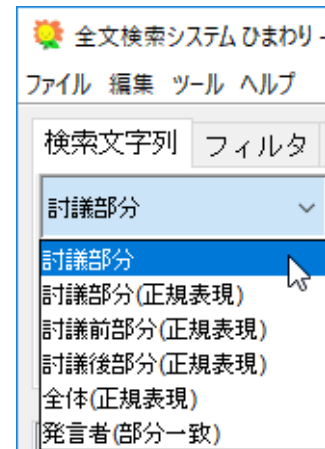
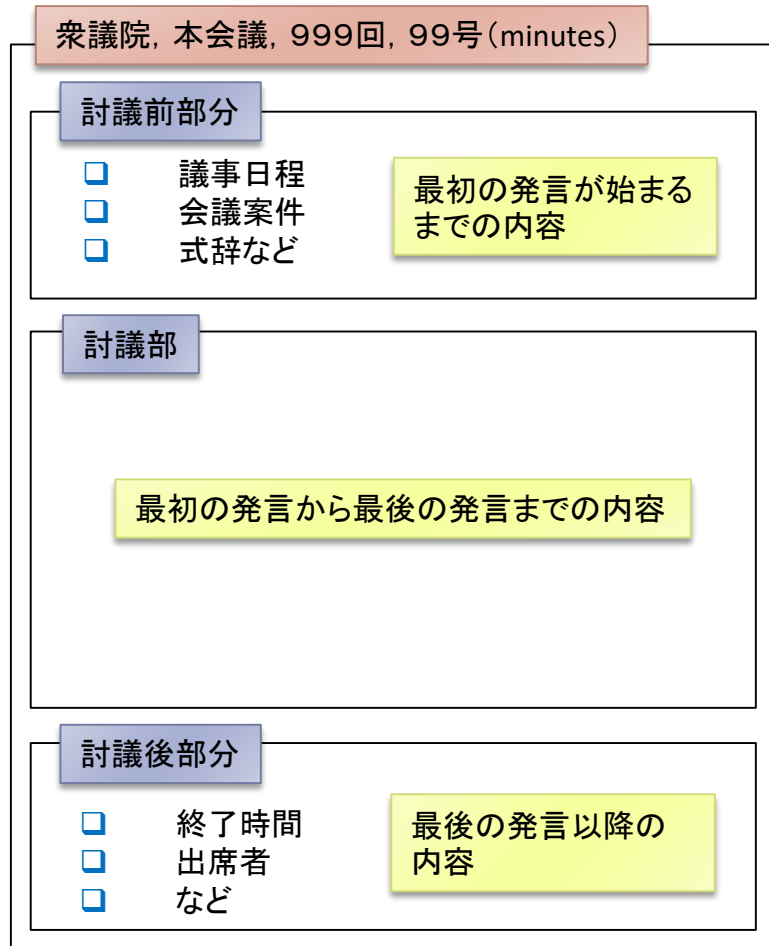


▶ サブコーパスにしている理由

- ▶ 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

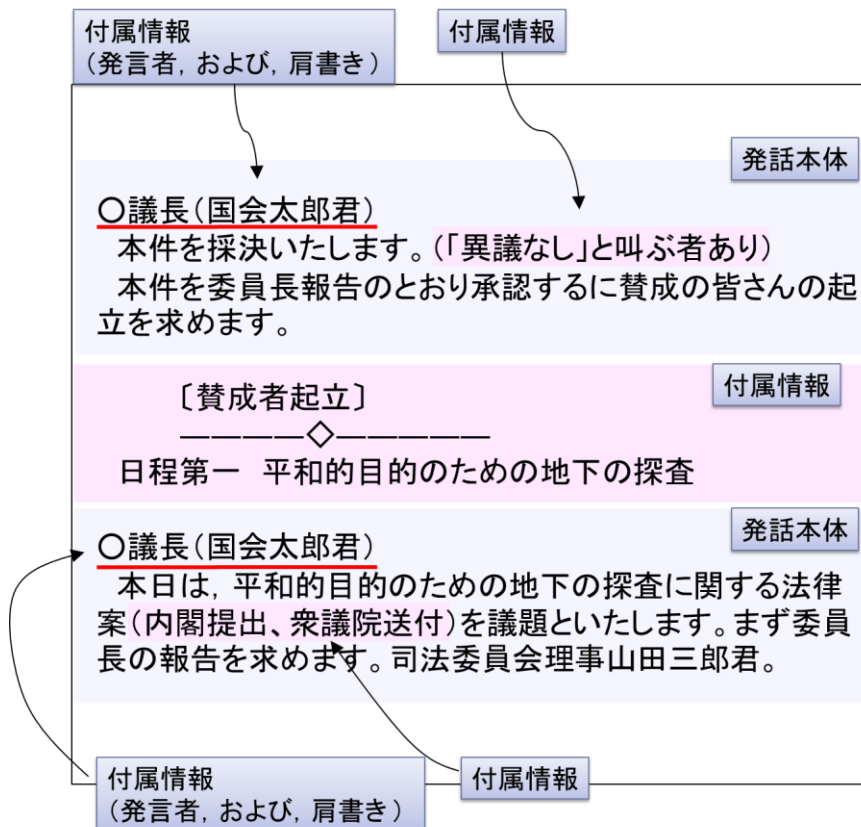
Corpora/Kokkai/xslt/__searched_tmp.xml

※「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

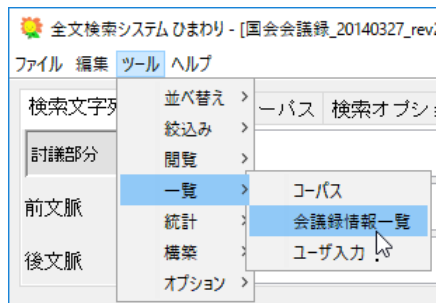
付属情報は本文の全文検索から除外



- ▶ 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

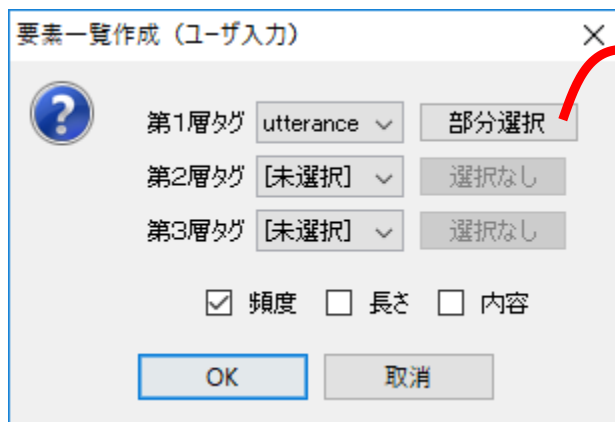


記事一覧

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

総数(延べ): 7127

▶ 発言者一覧



応用例



表記の経年変化

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

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

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

討議部分 [条件]件

前文脈 []で終わる

後文脈 []で始まる

検索

字体変換

クリア

2 「キー」「開催日」のセルを選択し、「統計」

キー	後文脈	開催日	文
条件	七、	1954-03-17	
条件	に従	1982-	コピー
条件	-人類	1999-	コピー(列名含)
条件	、	1953-	全選択
条件	、一期	1968-	フィルタ
条件	、ある	1955-	統計
条件	、ある	1956-	

3 年月日を年に置換

頻度: キー...

右クリック

キー	開催日	頻度
条件	2010-04-20	57
条件	2003-06-13	55
条件	1988-05-11	47
条件	2003-05-30	46
条件	2012-06-01	45
条件	1963-07-01	45
条件	1959-04-03	44
条件	1953-08-07	43

フィルタの設定

JOptionPane message

[文字列指定]
[置換]

置換 (正規表現)

検索する文字列: -.*

置換後の文字列: []

OK Cancel

正規表現: -.*

OK 取消

4 再集計

頻度: キー...

「キー」「開催日」を選択

キー	開催日	頻度
条件	2010	57
条件	2003	55
条件	1988	47
条件	2003	46
条件	2012	45
条件	1963	45
条件	1959	44
条件	1953	43
条件	2009	
条件	1988	

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]

要素一覧作成 (ユーザ入力)

第1層タグ minutes 一部選択

第2層タグ utterance 一部選択

第3層タグ [未選択] 選択なし

頻度 長さ 内容

OK 取消

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

minutes/date	utterance/birth_year	頻度
1960	1897	238
1987	1926	19
1996	1949	
1996	1948	
1960	1891	
1960	1890	
1996	1943	
1996	1942	
1996	1941	
1996	1940	
1996	1947	8
2003	1971	3
1996	1946	14

1960

総数 (絞りこみ前, 後): 119739, 114884 / 異な...

2 整形

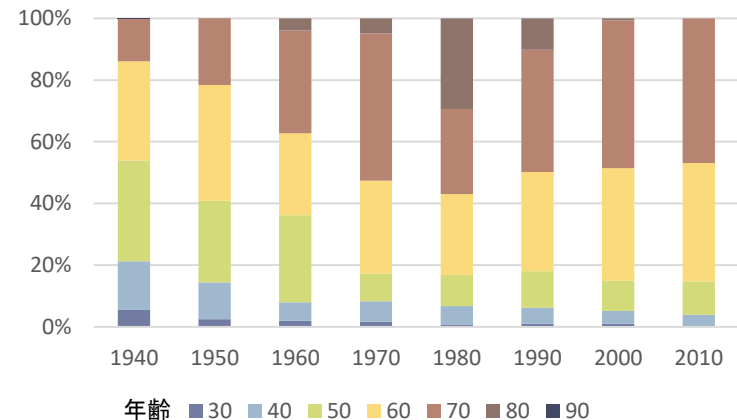
- ▶ 年月日⇒年変換
- ▶ 生年が空白のデータをフィルタリング

⇒列名を右クリック
⇒フィルタ
⇒文字列指定
(正規表現の「.」)

minutes/date	utterance/birth_year	頻度
1984	1923	16
1984	1919	2
1962	1914	6
2008	1942	10
1950	1905	52
1962	1892	5
1963		56
1973	1908	218
2008	1955	12
2003	1935	18
2000	1964	3
1971	1929	2
1964	1919	6

総数(延べ): 119739, 異なり: 2687

衆議院・本会議



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

検索文字列 フィルタ コーパス 検索オプション
基本形 行く
前文脈 検索文字列 フィルタ コーパス 検索オプション
基本形-1 [にへ]\$ 正規表現
タイトル で始まる
著者 で始まる

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

検索文字列 フィルタ コーパス 検索オプション
基本形 ^p{InKatakana}+\$
前文脈 で終る
後文脈 で始まる

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

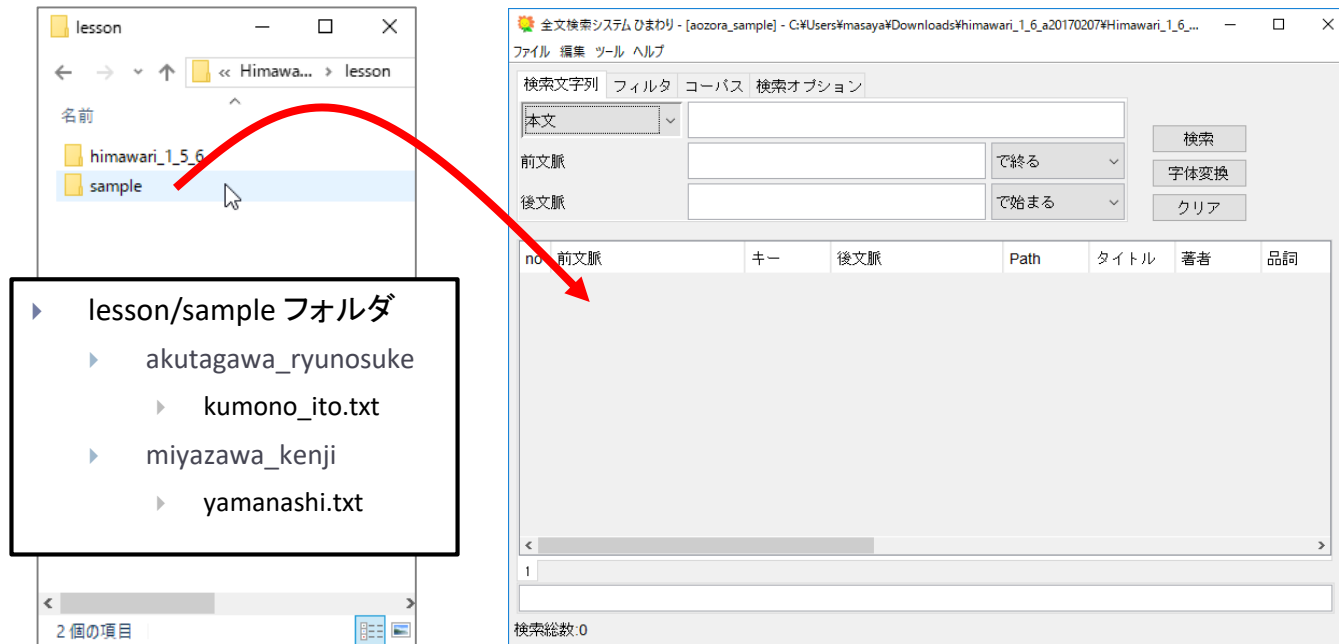
C) 繰り返し表現の抽出

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

検索文字列 フィルタ コーパス 検索オプション
出現形 (..)\1
前文脈 で終る
後文脈 で始まる

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

テキストデータインポート

変換対象データのフォルダ
C:\Users\masaya\Desktop\Himawari_1_6b02ls\lesson\sample 参照...

コーパスデータの出力
コーパス名 sample

詳細オプション

- 対象ファイル TXT XHTML XML
- 文字正規化 なし ユーザ定義 NFKC(Unicode)
- テキスト変換 aozorahtd
- XHTMLファイル用スタイルシート xhtml2xml_aozora.xsl
- HTMLファイルの変換も試みる
- XMLファイル用スタイルシート (変換なし)
- コーパス構築 サブコーパスを作る 索引付けを実行しない
- 形態素解析 (解析しない)

要素/属性/値

インポート 中止

▶ 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

[#8字下げ]—[#「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。

変換前



<注 内容="#8字下げ" 付与="" 種別="注記" />—<注 内容="#「ー」は中見出し" 付与="" 種別="注記" />

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。

変換後

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/>)