



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.1.6ls3 = ver.1.6b06+実習資料+UniDic ver.2.2.0)
 - ▶ 青空文庫(サンプル)
 - ▶ 名大会話コーパス

- ▶ 全体的な流れ
 - ▶ 『ひまわり』の紹介と基本的な使い方
 - ▶ もっとも簡単なインポート
 - ▶ テキストファイル形式の青空文庫 ⇨ ほぼアノテーション(情報付与)なし
 - ▶ すこし複雑なインポート
 - ▶ 5種類のタグを使って、アノテーション

『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

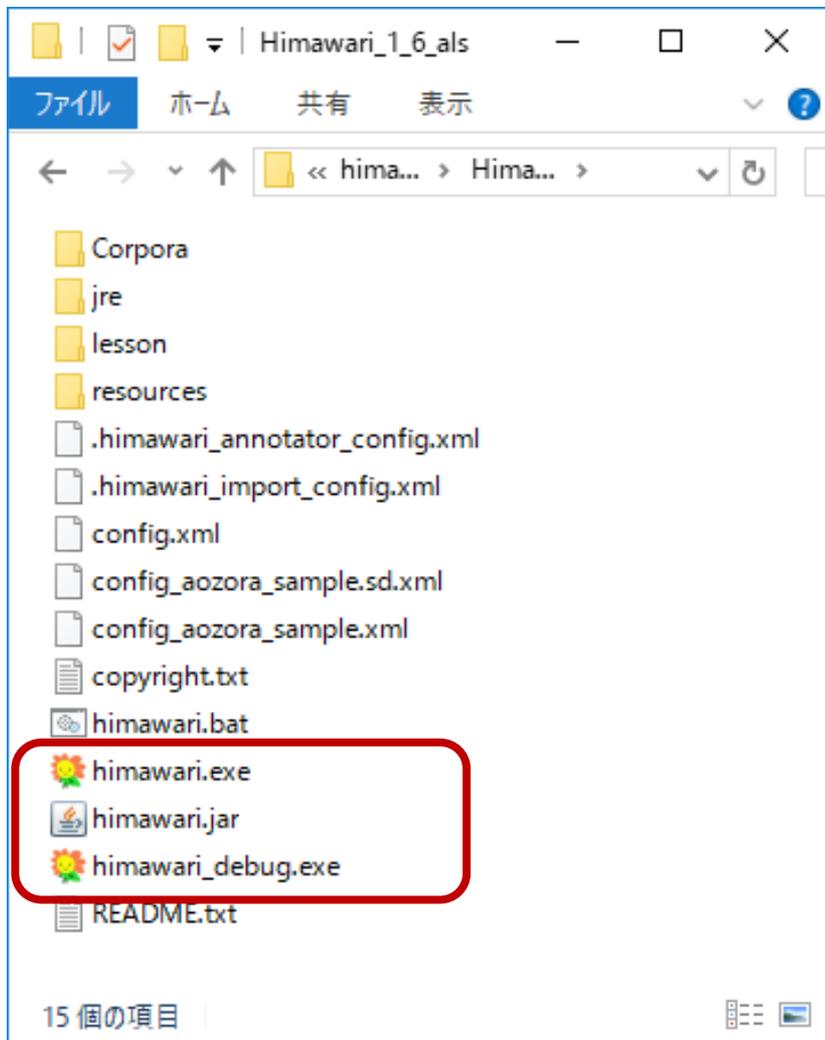
- ▶ XMLでタグづけされたコーパスを全文検索できます
- ▶ Windows, Mac OS, Linux など、多くのOS上で動作します
- ▶ 無料です

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

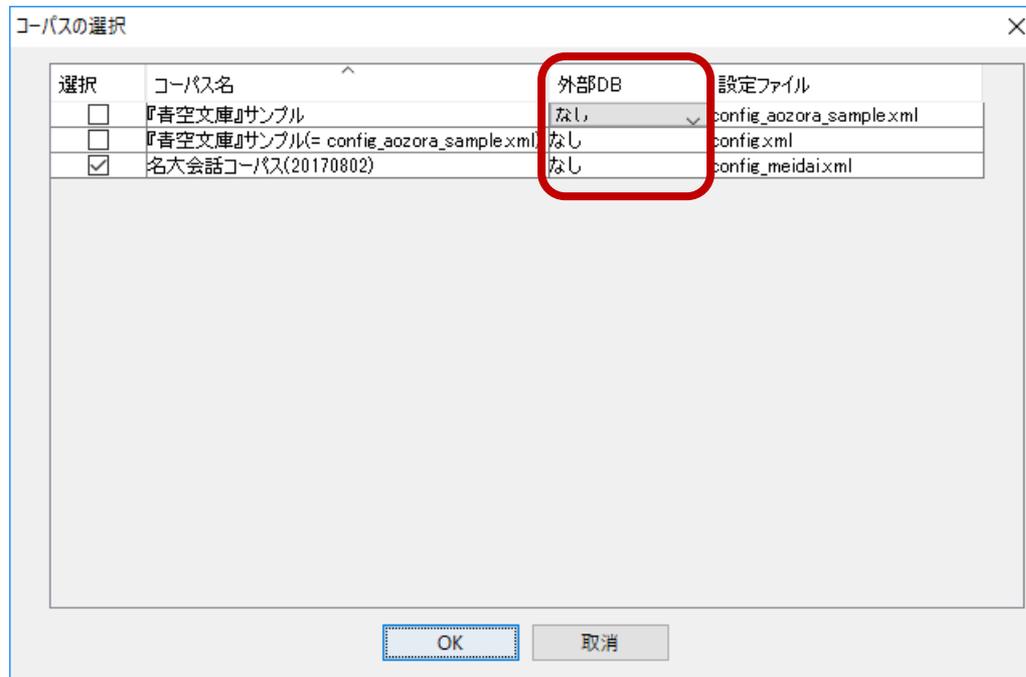


himawari.jar

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

コーパスの選択

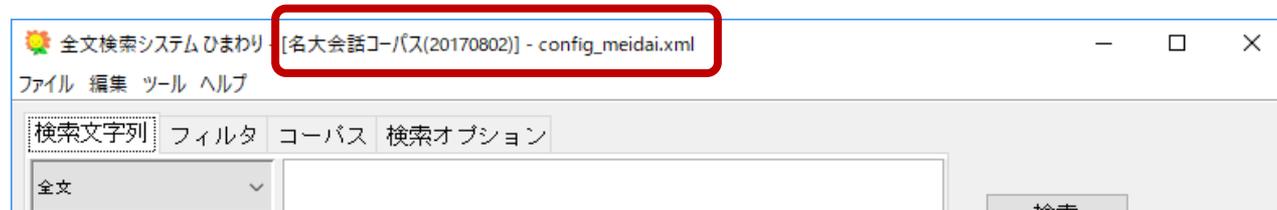
▶ [ファイル]⇒[コーパス選択]



▶ 「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 これ

前文脈 以終る

後文脈 以始まる

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	弾くところですよ」「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	めちゃんお困ります。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	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	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	と一と息ついた。「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

途中経過の表示

検索の実行

検索文字列

検索結果

検索総数

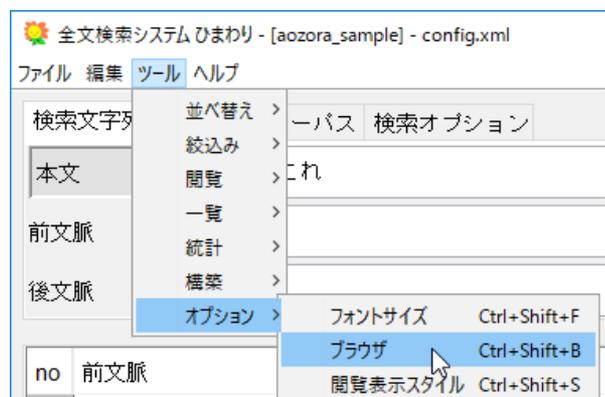
ブラウザでの閲覧

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ろ	/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...

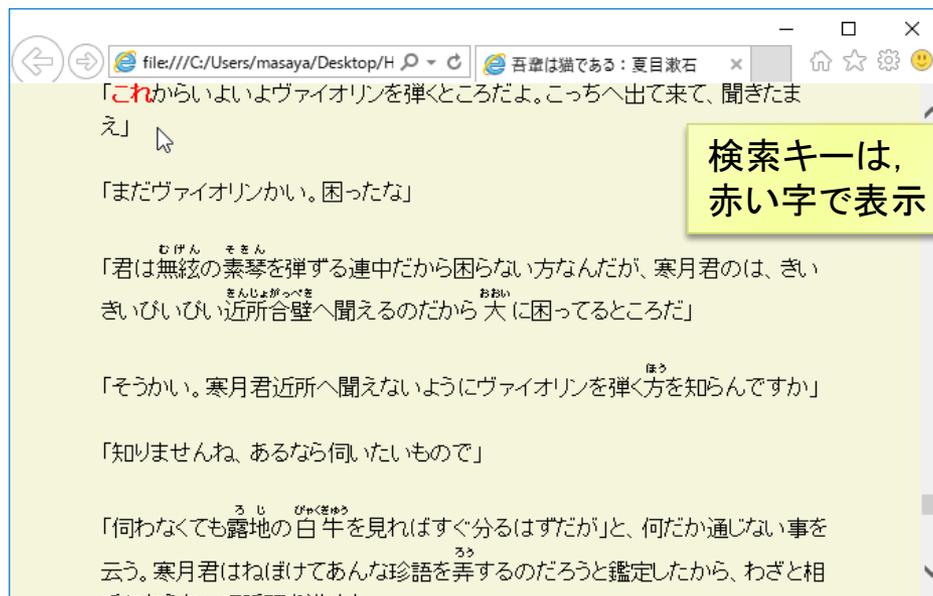
閲覧したい用例をダブルクリック



■ 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

検索結果の絞り込み

▶ 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	<input type="text"/>	で始まる
タイトル	<input type="text"/>	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

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	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

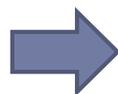
複数の列を
選択することも可

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」

1	タイトル	著者
ora_s...	吾輩は猫...	夏目漱石



タイトル	著者	頻度
吾輩は猫...	夏目漱石	190
こころ	夏目漱石	41
蜘蛛の糸	芥川龍之介	1

総数(延べ): 232, 異なり: 3

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	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						

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

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	!	記号	感嘆符					!	!	!

総数(延べ) : 206322

テキスト
進行方向

もっとも簡単なインポート



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。』

『クラムボンが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

▶ 「跳ねて」も検索OK

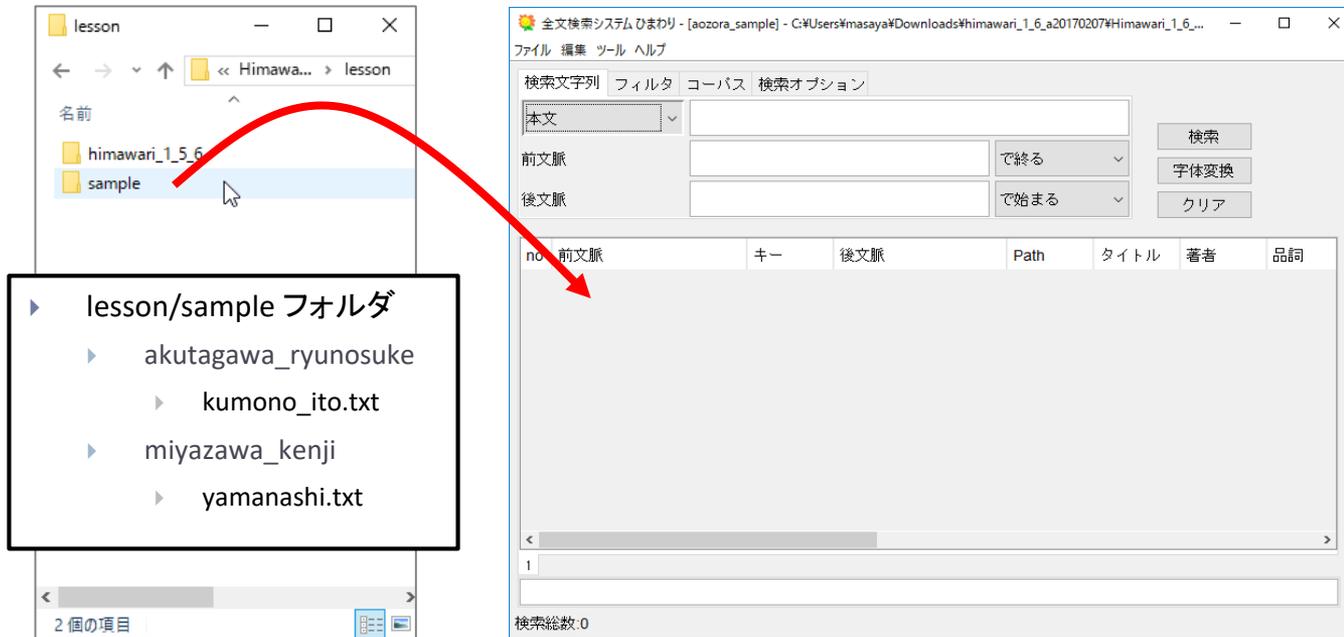
▶ ルビでの検索OK

▶ 検索範囲の指定OK

▶ 本文と付与情報(属性)は区別して検索

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

全文検索システムひまわり - [sample] - config_sample.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 検索

前文脈 で終る 字体変換

後文脈 で始まる クリア

no	前文脈	キー	後文脈	Path	タイトル	著者
1	これでおしまいであり	ます	。 底本：「新	/sample/m...	yamanashi	
2	ているばかりでござい	ます	。三 御釈迦	/sample/a...	kumono_ito	
3	切っているでござい	ます	。しかし地獄と極	/sample/a...	kumono_ito	
4	りに見えるのでござい	ます	。するとその地獄	/sample/a...	kumono_ito	
5	てやったからでござい	ます	。御釈迦様は地獄	/sample/a...	kumono_ito	
6	分等の穴に帰って行き	ます	。波はいよいよ青	/sample/m...	yamanashi	
7	ぶ暗い泡が流れて行き	ます	。『クラムポンはわ	/sample/m...	yamanashi	
8	ったら、大変でござい	ます	が、そう云う中にも	/sample/a...	kumono_ito	
9	な嘆息ばかりでござい	ます	。これはここへ落ちて	/sample/a...	kumono_ito	
10	くらく鋼のように見え	ます	。そのなめらかな天井	/sample/m...	yamanashi	
11	の間にかかくれて居り	ます	。それからあのぼんや	/sample/a...	kumono_ito	
12	を致した覚えがござい	ます	。と申しますのは、あ	/sample/a...	kumono_ito	
13	せっせとのぼって参り	ます	。今の中にどうかしな	/sample/a...	kumono_ito	
14	その途端でござい	ます	。今まで何ともなかっ	/sample/a...	kumono_ito	

1

ます

検索総数:33

- 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

file:///C:/User

kumono_ito :

#[特のへん+し+聿]、第3水準1-87-71 陀多のぶら下っている所から、ぶつりと音を立てて断れました。ですから※#[特のへん+し+聿]、第3水準1-87-71 陀多もたまりません。あっと云う間もなく風を切って、独楽のようにくるくるまわりながら、見る見る中に暗の底へ、まっさかさまに落ちてしまいました。

後にはただ極楽の蜘蛛の糸が、きらきらと細く光りながら、月も星もない空の中途に、短く垂れているばかりでござい**ます**。

#8字下げ三#[三]は中見出し

おしゃかさま(は)極楽(は)蓮池(の)ふちに立って、この一部(始終)をじっと見ていらっしやいましたが、やがて※#[特のへん+し+聿]、第3水準1-87-71 陀多が血の池の底へ石のように沈んでしまいますと、悲しそうな御顔をなさりながら、またぶらぶら御歩きになり始めました。自分ばかり地獄からぬけ出そうとする、※#[特のへん+し+聿]、第3水準1-87-71 陀多の無慈悲な心が、そうしてその心相当な罰をうけて、元の地獄へ落ちてしまったのが、御釈迦様の御目から見ると、浅間く思召されたのでございましょう。

しかし極楽の蓮池の蓮は、少しもそんな事には頓着致しません。その王(の)ような白い花は、御釈迦様の御足のまわりに、ゆらゆら萼を

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

参照...

コーパスデータの出力

コーパス名

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 aozora.htd

■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)

■設定ファイル(テンプレート) defaultConfig.xml

■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
 - ▶ 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

- 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

- 本日は, 「MeCab(UniDic)」を使用
- 辞書は, resource/unidicにインストール

インポート処理の概要



処理の流れ

① テキスト変換

- ▶ インポートするファイルを一定のルールでXMLファイルに変換

② ファイルの統合

- ▶ 変換したXMLファイルを一つのファイルに統合

③ 形態素解析(オプション)

[#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。

変換前



<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" />

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

XMLタグの基本

- ▶ 一定範囲に意味づけ
 - ▶ 開始タグ: <記事>
 - ▶ 終了タグ: </記事>
- ▶ 特定の位置に意味付け(範囲がない場合)
 - ▶ 空要素タグ:

「記事」タグは、「タイトル」、「著者」、「path」属性を持つ

開始タグ ➡ <記事 タイトル="吾輩は猫である" 著者="夏目漱石" path="/sample/テキスト1.txt">

<r rt="わがはい">吾輩</r>は猫である。名前はまだ無い。

どこで生れたかとんと見当がつかぬ。……

終了タグ ➡ </記事>

『ひまわり』は検索時、タグを読み飛ばして、文字列照合する

テキスト変換 (anozora.htdの場合)

- ▶ 改行位置に
を挿入
- ▶ 半角の&<>を全角に変換

- ▶ 青空文庫の注記を「注」タグへ
凡例: [#ここに注記を書<]

[#8字下げ]
[#「一」は中見出し]

- ▶ 青空文庫のルビをrタグへ
凡例1: 漢字列《ここにルビを書<》
凡例2: | 文字列《ここにルビを書<》

御釈迦様《おしゃかさま》
蓮池《はすいけ》
人間中で一番 | 獰悪《どうあく》な種族

[#8字下げ]—[#「一」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。

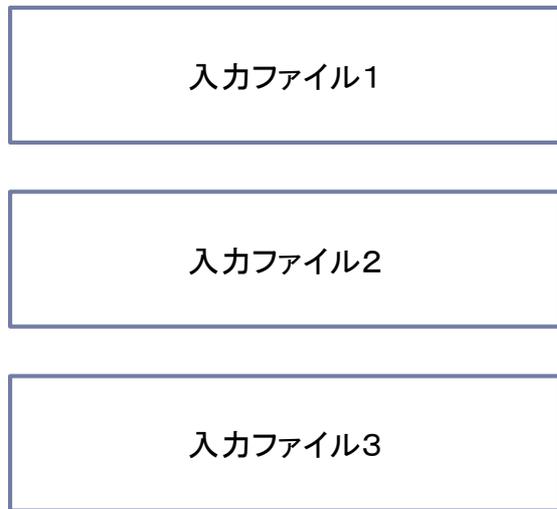
変換前

<注 内容="#8字下げ" 付与="" 種別="注記" />—<注 内容="#「一」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

ファイルの統合



:

インポート

corpus.xml

```
<コーパス>  
<記事>  
<テキスト>  
ここに、入力ファイル1の変換結果が置かれる)  
</テキスト>  
</記事>
```

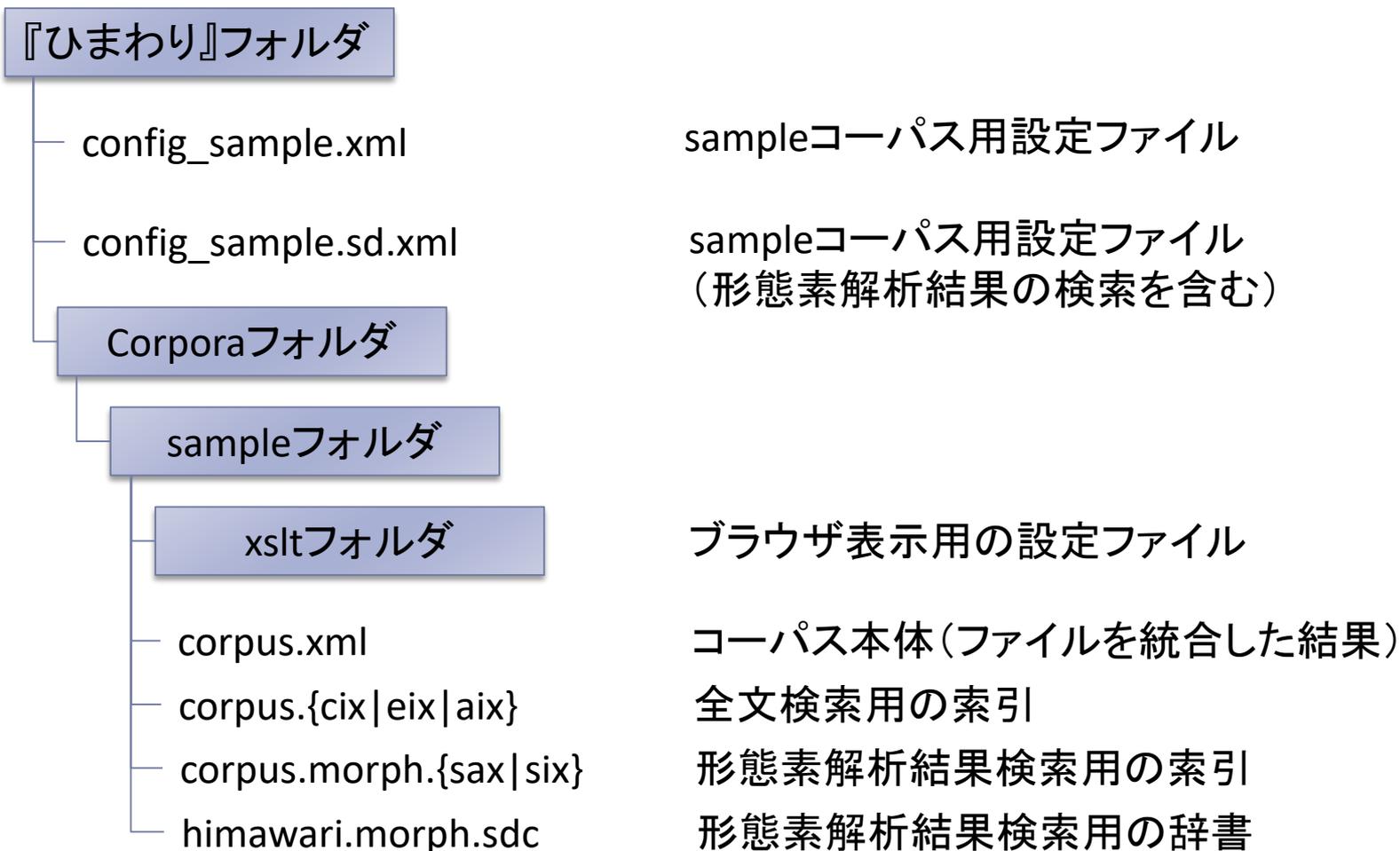
```
<記事>  
<テキスト>  
ここに、入力ファイル2の変換結果が置かれる)  
</テキスト>  
</記事>
```

```
<記事>  
<テキスト>  
(ここに、入力ファイル3の変換結果が置かれる)  
</テキスト>  
</記事>
```

: (入力のファイルの分だけ繰り返す)

```
</コーパス>
```

生成されるファイル(コーパス名sampleの場合)



少し複雑なインポート



概要

- ▶ 5種類の汎用タグを使ったアノテーションを行う
 - ▶ aozora.htd ⇒ ルビと注記のみ
 - ▶ diy.htd ⇒ 自分で意味づけ(資料全体, 作品本体など)

原資料

蜘蛛の糸
芥川龍之介

作者名やタイトルの情報
を利用できない

【テキスト中に現れる記号について】

:

[#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。

:

底本:「芥川龍之介全集2」ちくま文庫
1986(昭和61)年10月28日第1刷発行
1996(平成8)年7月15日第11刷発行

作品以外の部分も検索されてしまう

タグ付け後

t1(蜘蛛の糸,芥川龍之介)

開始タグ

【テキスト中に現れる記号について】

:

開始タグ

t2()

[#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。

終了タグ

:

/t2

底本:「芥川龍之介全集2」ちくま文庫、筑摩書房
1986(昭和61)年10月28日第1刷発行
1996(平成8)年7月15日第11刷発行

/t1

終了タグ

資料全体
(タイトル, 著者)

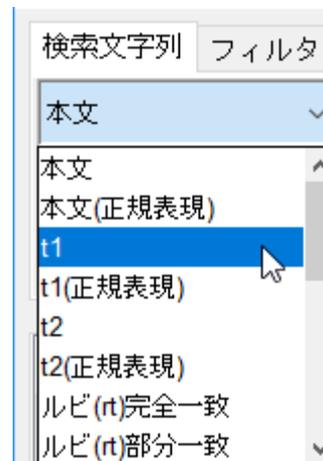
作品本体の
範囲指定

ブロックレベル要素用タグ t1, t2

▶ 機能 (t1, t2 同一)

- ▶ 作品全体, 章や節など, 行以上の範囲をアノテーションするのに使う
- ▶ 開始タグは三つまで属性を持てる
 - 0個 t1()
 - 1個 t1(夏目漱石)
 - 2個 t1(夏目漱石, 吾輩は猫である)
 - 3個 t1(夏目漱石, 吾輩は猫である, 1905)

□ 検索対象



t1タグの範囲のみを検索

□ 検索結果

no	前文脈	キー ^	後文脈	Path	タイトル	t1:属性1	t1:属性2
1		吾輩	は猫である。 名前	/annotatio...	wagahai	夏目漱石	吾輩は猫...

□ 変換結果のXML

タグは, すべて半角文字

```
t1(夏目漱石, 吾輩は猫である)
  吾輩は猫である。 名前はまだ無い。
  :
/t1
```

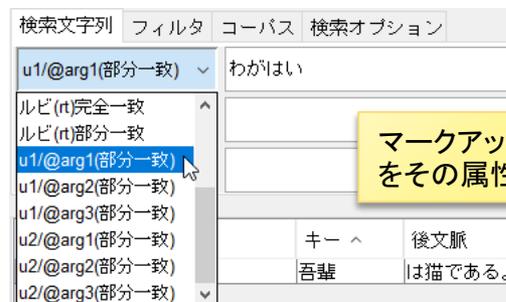


```
<t1 arg1="夏目漱石" arg2="吾輩は猫である">
  吾輩は猫である。 名前はまだ無い。
  :
</t1>
```

インライン要素用タグ u1, u2

- ▶ 機能 (u1, u2 同一)
 - ▶ 行内の範囲をアノテーションするのに使用する。
 - ▶ 開始タグは三つまで属性を持てる (属性なしは不可)

□ 検索対象



マークアップした文字列
をその属性値で検索

□ 青空文庫タグのu1, u2での記述

ルビ

ニ | 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。



ニu1(ひき)疋/u1のu1(かに)蟹/u1の子供らが青じろい水の底で話していました。

注記

この※[#「特のへん+互+聿」、第3水準1-87-71]陀多には蜘蛛を助けた事があるのを御思い出しになりました。



このu2(特のへん+互+聿,第3水準1-87-71) ※/u2陀多には蜘蛛を助けた事があるのを御思い出しになりました。

空要素タグ e1

▶ 機能

- ▶ 原資料のページ番号や行位置など, 位置を表すのに使う
- ▶ 三つまで属性を持てる(属性なしは不可)
 - ▶ e1/(動詞), e1/(動詞,五段), e1/(動詞,五段,未然形)
- ▶ 検索時は, マッチした文字列の先頭文字から見て, 文進行方向の最も近いタグの属性値を表示
 - ▶ 「吾輩」「吾輩は」「輩」の場合 ⇒ 「名詞」
 - ▶ 「猫である」の場合 ⇒ 「名詞」

□ 単語の区切り例

原資料

吾輩は猫である。

タグ付け後

吾輩e1/(名詞)はe1/(助詞)猫e1/(名詞)でe1/(助動詞)あるe1/(助動詞)。e1/(記号)

* (機能の説明用なので, 実用には少し無理がある)

タグを使ってみよう

lesson/annotation_sample_results
フォルダはアノテーション例なので、
一通り終えたあと、参照のこと

▶ lesson/annotation_sample フォルダ

wagahai.txt

【夏目漱石 吾輩は猫である】

吾輩《わがはい》は猫である。【1文目】

名前はまだ無い。【2文目】

どこで生れたかとんと見当《けんとう》がつかぬ。【3文目】

- 全体をt1タグ
- ルビをu1タグ
- 文番号をe1タグ

yamanasi.txt

【宮沢賢治 やまなし】

【子供】『お父さん、いまおかしなものが来たよ。』

【お父さん】『どんなもんだ。』

【子供】『青くてね、光るんだよ。はじがこんなに黒く尖ってるの。
それが来たらお魚が上へのぼって行ったよ。』

- 全体をt1タグ
- 1発話をt2タグ

【】内は属性値とし、本文ではないものとする

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_6ls03\lesson\annotation_sample 参照...

コーパスデータの出力
コーパス名 annotation_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 diyhtd

■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)

■設定ファイル(テンプレート) diyConfig.xml

■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 MeCab(UniDic)
要素/属性/値

インポート 中止

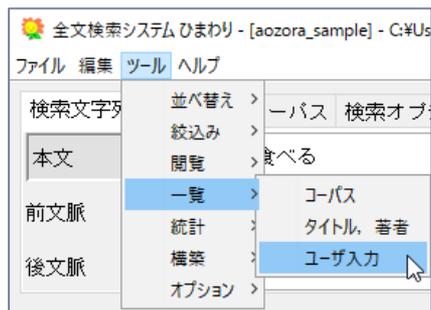
diy.htd

diyConfig.xml

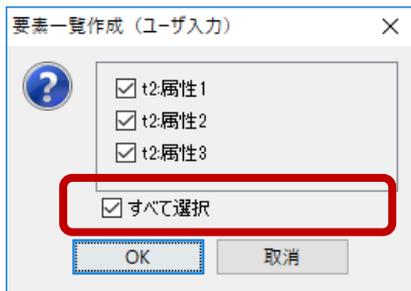
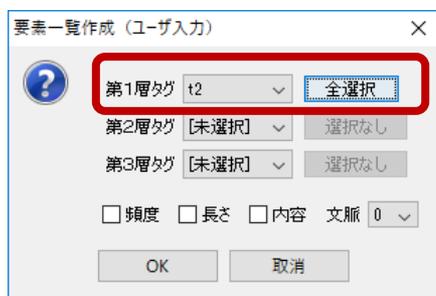
MeCab(UniDic)

アノテーション結果の集計

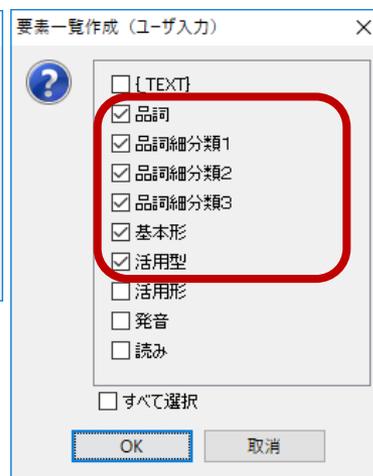
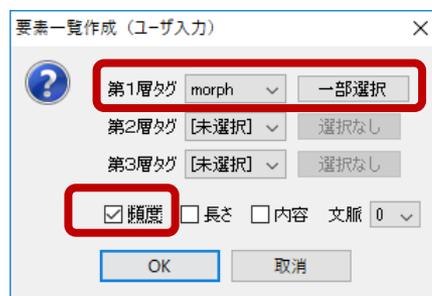
- ▶ 一覧機能(ユーザ入力)で付与情報を閲覧



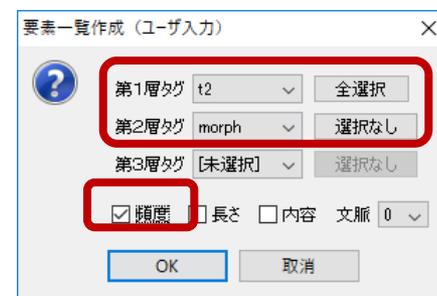
■ t2一覧



■ morph(形態素解析結果)一覧



■ t2中の形態素数(話者別)



タグの包含関係を用いる

補足

設定ファイル(config_*.xml)の調整

▶ 例：列名を変える

```
<!-- 結果レコードのフィールド定義 -->
<field_setting>
  <li align="RIGHT" name="no" type="index" width="30"/>
  <li align="RIGHT" attribute="_preceding_context" element="_sys" name="前文脈" sort_direction="R"
type="preceding_context" width="180"/>
  <li attribute="_key" element="_sys" name="キー" sort_order="1" type="key" width="80"/>
  <li attribute="_following_context" element="_sys" name="後文脈" sort_order="2" type="following_context"
width="160"/>
  <li attribute="path" element="記事" name="Path" type="argument" width="80"/>
  <li attribute="タイトル" element="記事" name="タイトル" type="argument" width="80"/>
  <li attribute="arg1" element="t1" name="t1:属性1" type="argument" width="80"/>
  <li attribute="arg2" element="t1" name="t1:属性2" type="argument" width="80"/>
  <li attribute="arg3" element="t1" name="t1:属性3" type="argument" width="80"/>
  <li attribute="arg1" element="t2" name="t2:属性1" type="argument" width="80"/>
  <li attribute="arg2" element="t2" name="t2:属性2" type="argument" width="80"/>
  <li attribute="arg3" element="t2" name="t2:属性3" type="argument" width="80"/>
```

name属性の値
を変更

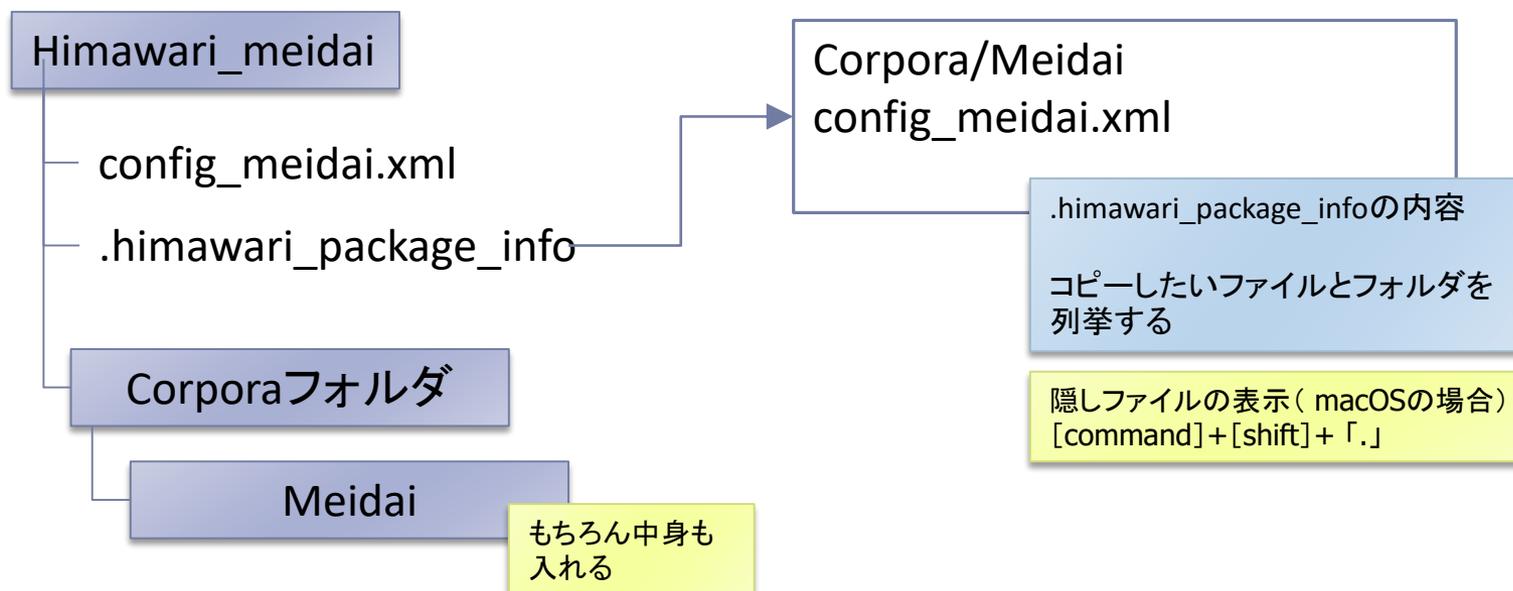
▶ 参考資料

- ▶ 『ひまわり』ホームページ⇒「文書」
⇒「設定ファイルリファレンスマニュアル」

作成した資料の配布

▶ 名大会話コーパスの場合

- ▶ 次の構造のフォルダを作成し, zipで圧縮



▶ 参考資料

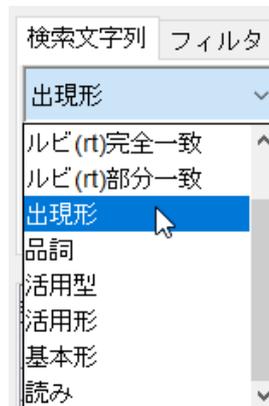
- ▶ 「設定ファイルリファレンスマニュアル」の「パッケージ設定ファイル」

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"/>

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
 - ▶ 青空文庫形式テキストのインポート
 - ▶ 5種類の汎用タグを用いたアノテーション

- ▶ さらに詳しく知るには
 - ▶ 『ひまわり』ホームページの各種資料
(ver.1.6は本年度安定版になる予定)
 - ▶ 『ひまわり』用各種パッケージが実例として使える
 - ▶ テキスト処理の知識があれば、直接XML形式に変換する方法もあり

各種設定ファイル & 参考資料

- ▶ 形態素解析システム
 - ▶ 『ひまわり』フォルダの.himawari_annotator_config.xml
 - ▶ 『ひまわり』フォルダの resources/unidic/dicrc (辞書情報)
 - ▶ 「設定ファイルリファレンスマニュアル」の「アノテーション関連」
- ▶ インポート関連
 - ▶ 『ひまわり』フォルダの.himawari_import_config.xml
 - ▶ 「設定ファイルリファレンスマニュアル」の「インポート関連」
- ▶ テキスト変換規則
 - ▶ 「設定ファイルリファレンスマニュアル」の
import / text_transformation_definition 要素
- ▶ 『ひまわり』用データ作成一般
 - ▶ 『ひまわり』HP ⇒ 「簡単な検索用データの作成方法1」
 - ▶ 『ひまわり』HP ⇒ 「簡単な検索用データの作成方法2」