



# 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.1.6ls05 = ver.1.6.4+実習資料+UniDic ver.2.2.0)
  - ▶ 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
  - ▶ 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』用パッケージの作成

# 本講習会のねらい

---

- ▶ 既存の資料を『ひまわり』で利用できるようになること
  - ▶ 資料の『ひまわり』へのインポート
  - ▶ 『ひまわり』で利用できる形式のタグ付け
  - ▶ 作成した資料の配布(『ひまわり』用パッケージ)
- ▶ 利点
  - ▶ 『ひまわり』の各種機能を利用可能
  - ▶ 資料を他の研究者と共有可能
    - ▶ 研究結果を検証してもらえる
    - ▶ 古い版の資料も利用できる
  - ▶ 誰でも(知識さえあれば)資料全体を検証可能

# 『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

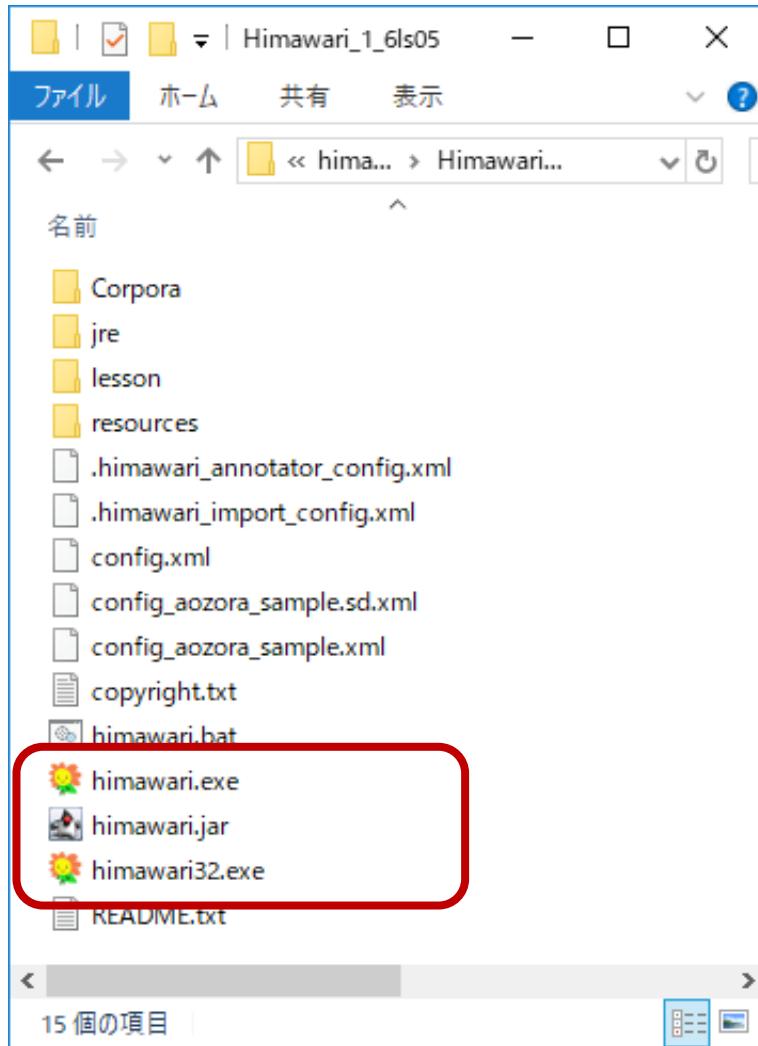
## ▶ 特徴

- ▶ 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(64ビット版)  
(Windows 専用)  
himawari.exe



himawari32.exe

普段使うとき(32ビット版)  
(Windows 専用)  
himawari32.exe

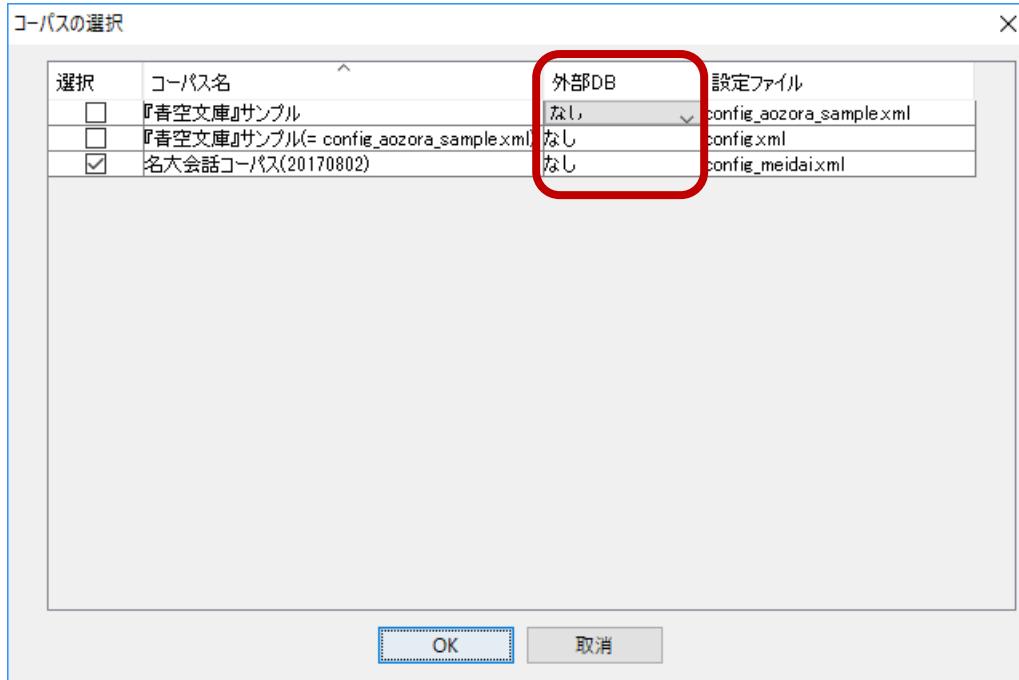


himawari.jar

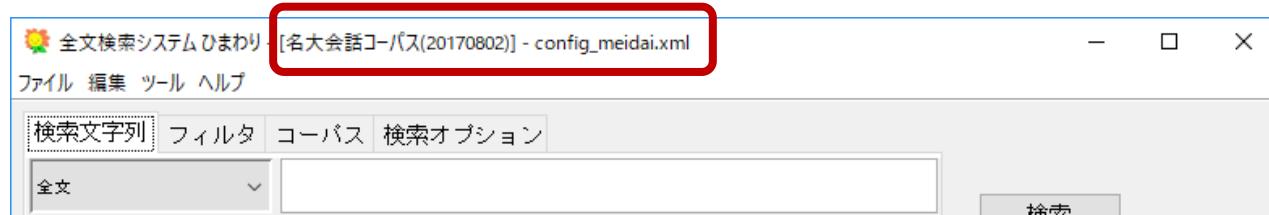
汎用  
(Windows, Mac, Linux など)  
himawari.jar

# 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

「検索文字列」欄では  
右クリックで履歴表示

The screenshot shows the user interface of the全文検索システムひまわり application. The main window title is "全文検索システムひまわり - [aozora\_sample] - config.xml". The menu bar includes ファイル, 編集, ツール, ヘルプ. The toolbar includes 検索文字列, フィルタ, コーパス, 検索オプション. The search interface has a dropdown for "本文" (Entire document) and a search input field containing "これ". Below the search interface are sections for "前文脈" (Previous context) and "後文脈" (Next context). To the right of the search interface are buttons for "検索" (Search), "字体変換" (Font conversion), and "クリア" (Clear). A large yellow callout box labeled "検索の実行" (Search execution) points to the "検索" button. Another yellow callout box labeled "検索結果" (Search results) points to the results table. A third yellow callout box labeled "検索総数" (Total search count) points to the status bar at the bottom. A fourth yellow callout box labeled "途中経過の表示" (Intermediate process display) points to the terminal-like output on the left. A blue callout box labeled "「検索文字列」欄では右クリックで履歴表示" (Right-click on the search text input field to display history) points to the search input field.

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

途中経過の表示

検索総数

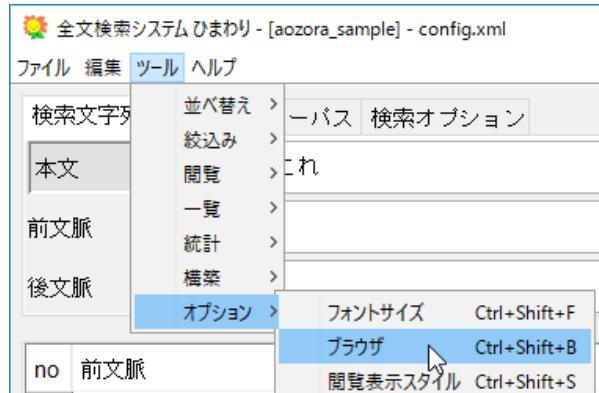
# ブラウザでの閲覧

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	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



「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、きいきいひいひ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

検索キーは、赤い字で表示

[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

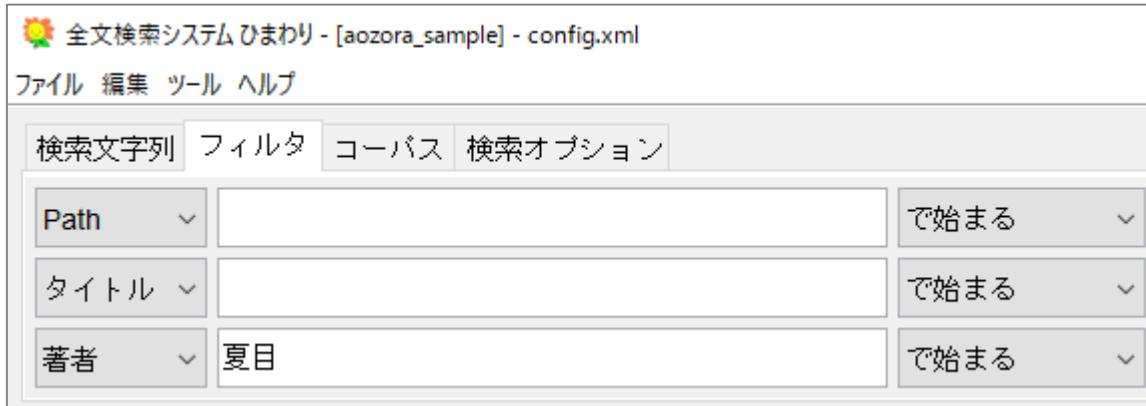
no	前文脈	キー	後文脈	Path	タイトル	著者
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...	こころ	夏目漱石
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目漱石
3	弾くところです」「これ		からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目漱石
4	い話があるかい」「これ		からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目漱石
5	、蛸壺峠へかかって、これ		からいよいよ会津領へ	/aozora_s...	吾輩は猫...	夏目漱石
6	見当がつかない」「これ		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	はすこぶる不憮だよ。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

# 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	[文字列指定] [置換]
	て、これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目	夏目漱石
	です」「これ	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目	芥川龍之介
	いい」「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	
	よい」「これ	これ	からいよいよ弾くとこ	/aozora_s...	吾輩は猫...	夏目漱石	
	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	

絞り込みたい  
値を選択  
⇒右クリック  
⇒filtration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」

The image shows two windows. On the left, a context menu is open over the word '吾輩は猫' in the 5th row of the table. The menu items are: コピー (Copy), コピー(列名含む) (Copy with column name), 全選択 (Select All), 置換 (Replace), フィルタ (Filter), 統計 (Statistics), and 吾輩は猫 (Wagahai wa neko). The '統計' item is highlighted with a blue arrow pointing to the right. On the right, a separate window titled '[1] 頻度 : タイトル,...' displays the frequency statistics for the selected words. The table shows:

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

At the bottom of the window, it says '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	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

ファイル 編集

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	トヨ	トヨ				トヨ	トヨ	トヨ
00021790									
総数(延べ): 206322									

テキスト進行方向

➡

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

# 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

【テキスト中に現れる記号について】

### 3種類の独自タグ

《》: 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 | 圮《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 | 圮《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。』

『クラムボンはかふかふわらったよ。』

『クラムボンは跳《は》ねてわらったよ。』

『クラムボンはかふかふわらったよ。』

### 通常の文字列検索の問題点

▶ 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』の全文検索

▶ 「跳ねて」も検索OK

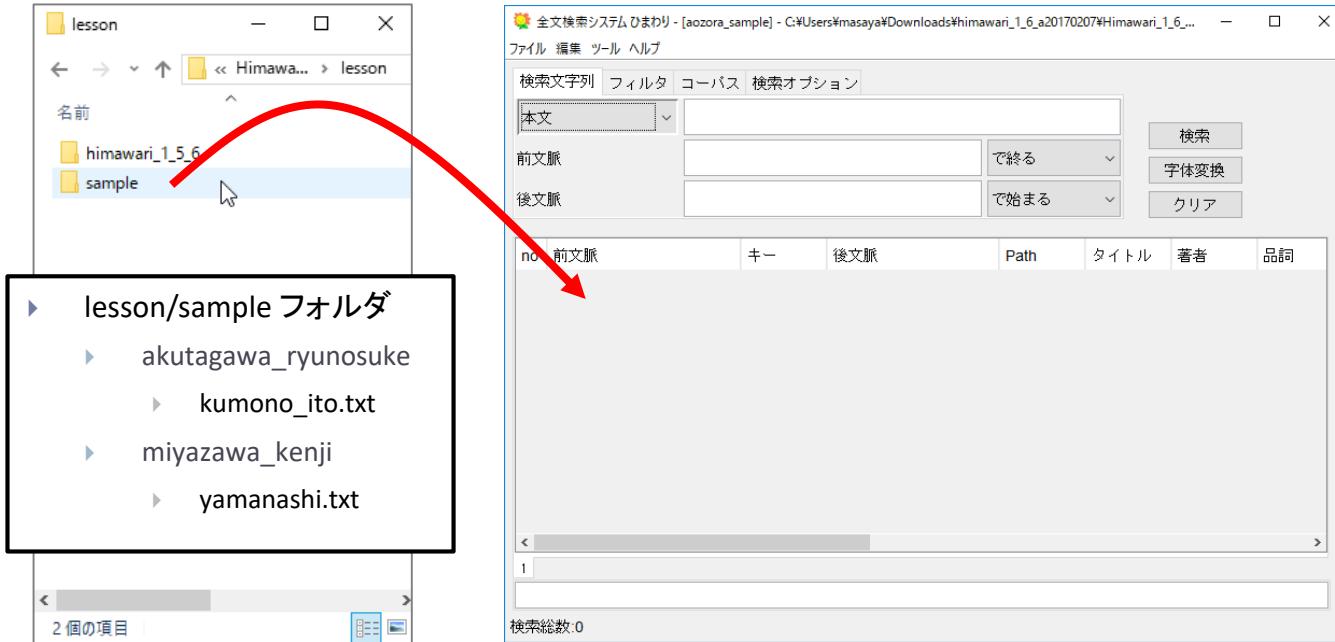
▶ ルビでの検索OK

▶ 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 ます

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	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	切っているのでございます		。しかし地獄と極	/sample/a...	kumono_ito	
4	りと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	ぶ暗い泡が流れています		。『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	くらく錆のよう見えます		。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ります		。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	その途端でござります		。今まで何ともなかっ	/sample/a...	kumono_ito	

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

#「持のへん+三+聿」、第3水準1-87-71 鮎多のふ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※「持のへん+三+聿」、第3水準1-87-71 鮎多もたまりません。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しました。

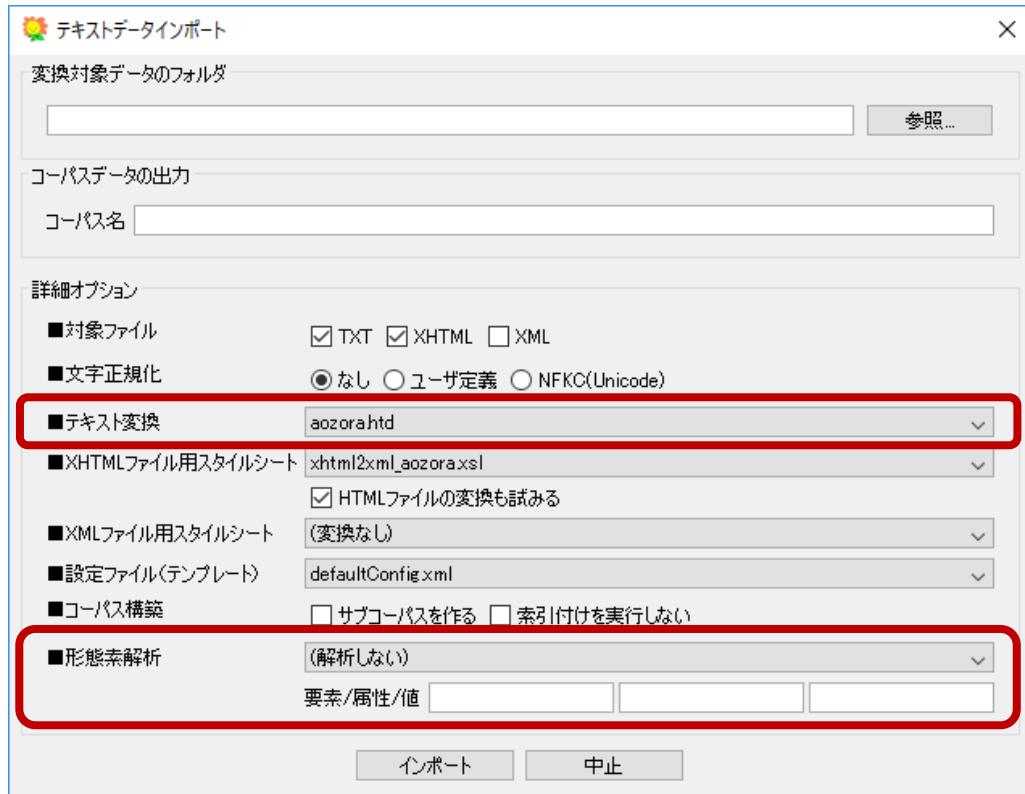
後にはただ極楽の蜘蛛の糸が、きらきらと細く光りながら、月も星もない空の中途に、短く垂れているばかりでございます。

#E字下げ 三 #「三」は中見出し

おしゃかさま はすいか しじゅう  
御釈迦様は極楽の蓮池のふちに立って、この一部敗戻終をじっと見て  
かんだいた  
いらっしゃいましたが、やがて※「持のへん+三+聿」、第3水準1-87-71 鮎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、※「持のへん+三+聿」、第3水準1-87-71 鮎多の無慈悲な心が、そ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、浅間し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。その王のとうた白い花は 御釈迦様の御足の主わりに ゆらゆら重を

# インポート時のオプション



- 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

## ▶ 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、<br />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

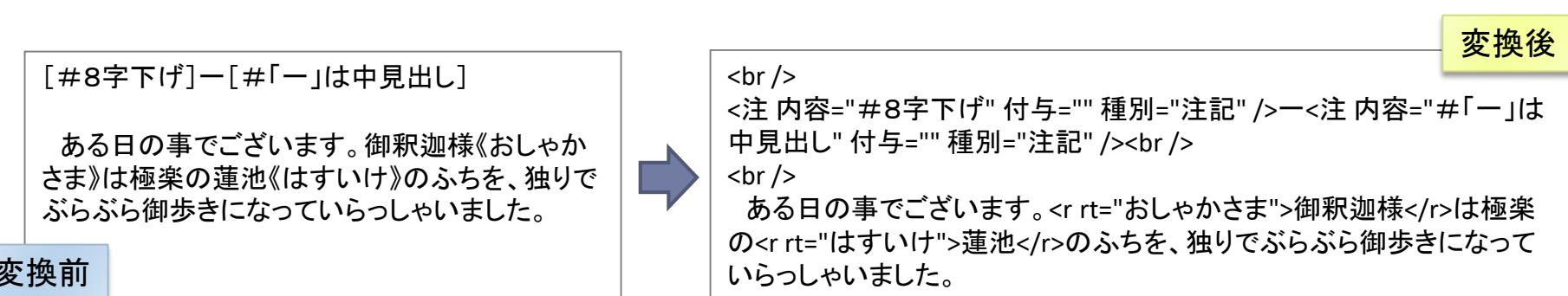
## ① テキスト変換

- ▶ 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)



# 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 : <br />

「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。<br />  
どこで生れたかとんと見当がつかぬ。……<br />

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(anzoza.htmの場合)

- 改行位置に<br />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: [#ここに注記を書く]
- 青空文庫のルビをrタグへ  
凡例1: 漢字列《ここにルビを書く》  
凡例2: | 文字列《ここにルビを書く》

[#8字下げ]  
[#「一」は中見出し]

御釈迦様《おしゃかさま》  
蓮池《はすいけ》  
人間中で一番 | 獐悪《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

変換後

```
<br />
<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" /><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。
```

# ファイルの統合

入力ファイル1

入力ファイル2

入力ファイル3

:

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)

## 『ひまわり』フォルダ

config\_sample.xml

sampleコーパス用設定ファイル

config\_sample.sd.xml

sampleコーパス用設定ファイル  
(形態素解析結果の検索を含む)

## Corpora フォルダ

### sample フォルダ

#### xslt フォルダ

ブラウザ表示用の設定ファイル

corpus.xml

コーパス本体(ファイルを統合した結果)

corpus.{cix|eix|aix}

全文検索用の索引

corpus.morph.{sax|six}

形態素解析結果検索用の索引

himawari.morph.sdc

形態素解析結果検索用の辞書

# 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]—[#「—」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》は極樂の蓮池《はすいけ》のふちを、独りでぶらぶら御歩きになつていらつしやいました。

:

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】

:

t2()  
開始タグ

[#8字下げ]—[#「—」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》は極樂の蓮池《はすいけ》のふちを、独りでぶらぶら御歩きになつていらつしやいました。

/t2

終了タグ

:

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)

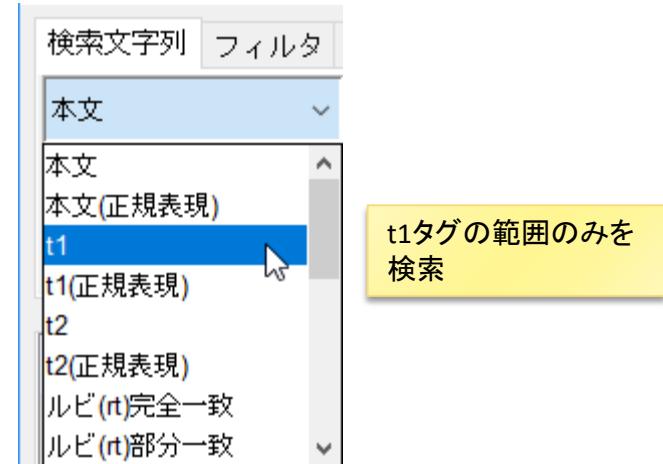
作品本体の  
範囲指定

# 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



## □ 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	/annotation	wagahai	夏目漱石	吾輩は猫...

## □ 変換結果のXML

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

タグは、すべて半角  
文字

<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つ（属性なしは不可）

## □ 検索対象



## □ 青空文庫タグのu1, u2での記述

ルビ

ニ | 死《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。

注記

この※[「特のへん+ゑ+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。



ニu1(ひき)死/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。



このu2(特のへん+ゑ+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)

# 空要素タグ e2

## ▶ 機能

- ▶ e1とほぼ同じ機能を持つ
- ▶ ただし、検索時は、マッチした文字列の先頭文字から見て、文書先頭方向の最も近いタグの属性値を表示

## □ 単語の区切り例

原資料

吾輩は猫である。

タグ付け後(e2)

e2/(名詞)吾輩e2/(助詞)はe2/(名詞)猫e2/(助動詞)でe2/(助動詞)あるe2/(記号)。

タグ付け後(e1)

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)

# 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

yamanashi.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。』

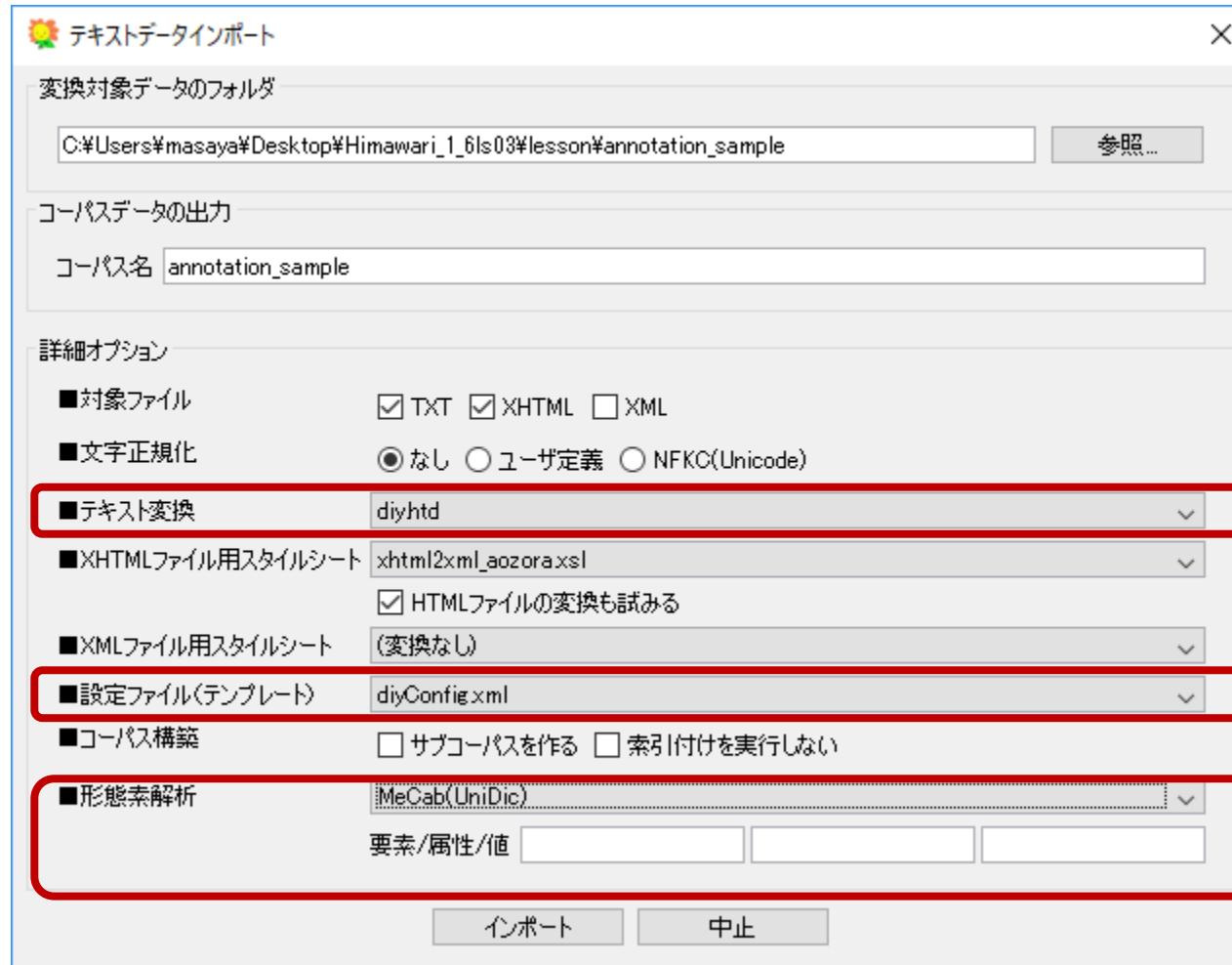
【お父さん】『どんなもんだ。』

【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

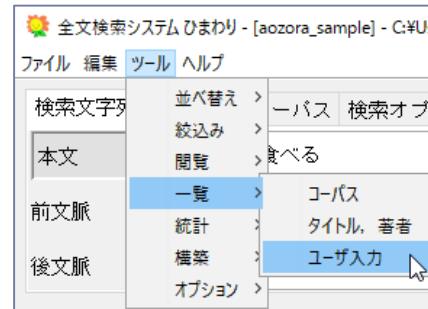
【】内は属性値とし、本文ではないものとする

# インポート時のオプション

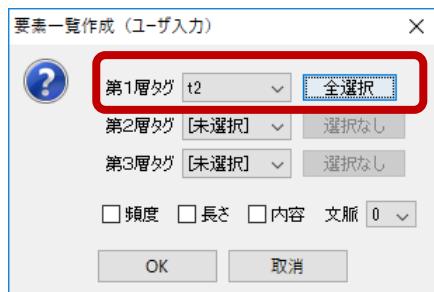


# アノテーション結果の集計

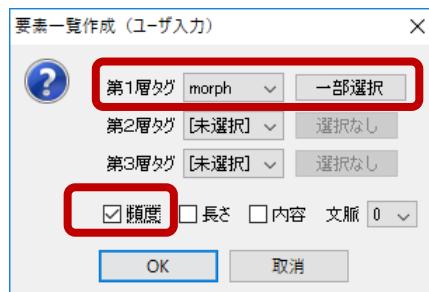
- ▶ 一覧機能(ユーザ入力)で付与情報を閲覧



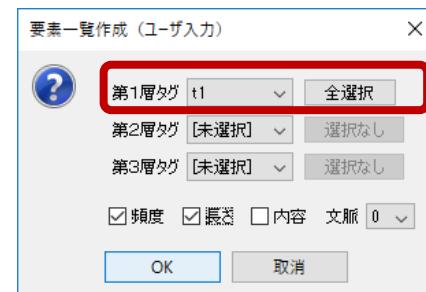
## ■ t2一覧



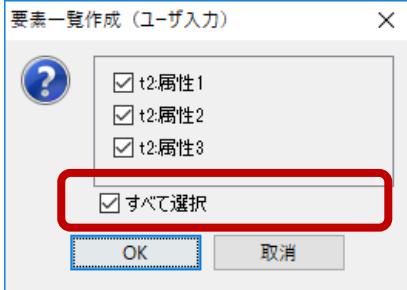
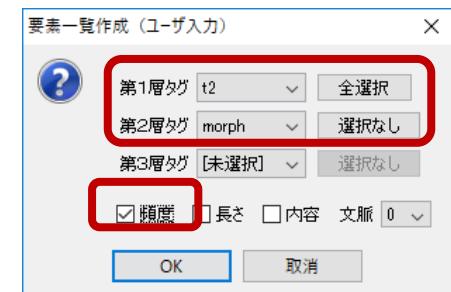
## ■ morph (形態素)一覧



## ■ 文字数(t1)



## ■ t2中の話者別形態素数



属性

- 品詞
- 品詞分類1～3
- 基本形
- 活用型

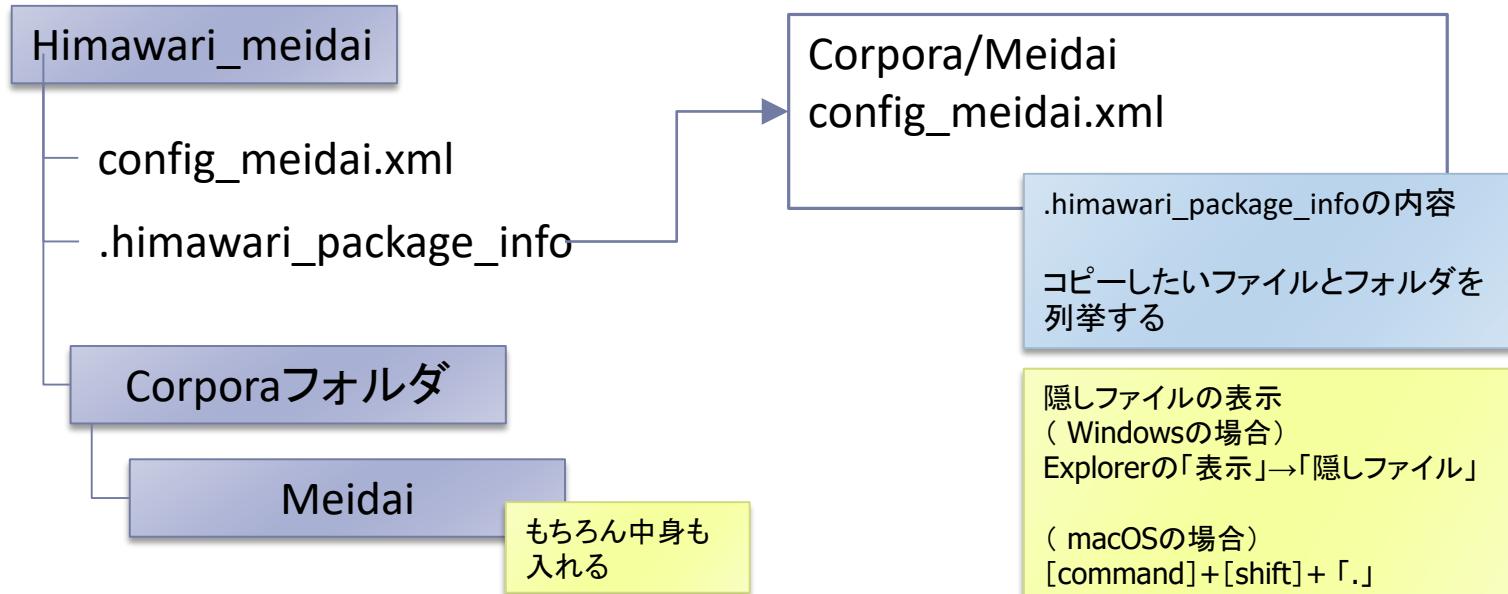
「長さ」: 最下位要素の長さ

タグの包含関係を用いる

# 『ひまわり』用パッケージの作成

# 『ひまわり』用パッケージの構造

- ▶ 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」

# パッケージを作成してみよう

---

1. パッケージを収めるフォルダ作成
2. 関連ファイルを1のフォルダにコピー
  - ▶ config ファイル
  - ▶ Corpora フォルダの中の関連フォルダ
3. .himawari\_package\_infoを作成
  - ▶ [Windowsの場合]  
Explorerで新規作成(ファイル名の末尾に「.」をつける)
  - ▶ [macOSの場合]  
miで[ファイル]→[別名で保存]
4. フォルダをzipで圧縮

# 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」

# 設定ファイル(config\_\*.xml)の調整

## ▶ 全文検索範囲の指定

```
<index_cix>
  <li field_name="キー" label="本文" middle_name="article" name="テキスト" type="normal"/>
  <li field_name="キー" label="本文(正規表現)" middle_name="article" name="テキスト" type="null"/>
  <li field_name="キー" label="t1" middle_name="t11" name="t1" type="normal"/>
  <li field_name="キー" label="t1(正規表現)" middle_name="t12" name="t1" type="null"/>
  <li field_name="キー" label="t2" middle_name="t21" name="t2" type="normal"/>
  <li field_name="キー" label="t2(正規表現)" middle_name="t22" name="t2" type="null"/>
</index_cix>
```

- ▶ index\_cix/li/@name                    対象とする要素名
- ▶ index\_cix/li/@type                    全文検索の種類
  - ▶ normal: 索引付きの通常の全文検索
  - ▶ null: 索引なしの正規表現(全文)検索

# 補足

# インポート時のテキスト変換(htdファイル)

---

- ▶ 正規表現による文字列置換を利用
  - ▶ 正規表現は、Java (クラス Pattern) に準ずる
- ▶ 変換規則
  - ▶ Himawari\_1\_6ls05/resources/htd に変換規則ファイルを配置
  - ▶ 変換規則の形式

変換前文字列(正規表現) タブ文字 変換後文字列
  - ▶ 規則の適用
    - ▶ 1入力ファイル全体(改行を含め)を一つの文字列と考える
    - ▶ 変換規則を上から順に適用する

# 変換規則の例

## ▶ aozora.htd

```
## 改行位置に、<br />を挿入
¥n          <br />¥n
## 注記
[(#.+?)]    <注 内容="$1" 付与="" 種別="注記" />
## ルビ(範囲指定あり)
[ | ](.+?)《(.+?)》      <r rt="$2">$1</r>
## ルビ(範囲指定なし)
(¥p{InCJKUnifiedIdeographs}+?) 《(.+?)》      <r rt="$2">$1</r>
```

※ [#小書き平仮名ん]  
⇒ <注 内容="#小書き平仮名ん" 付与="" 種別="注記" />

一番 | 獄惡《どうあく》な  
⇒ 一番<r rt="どうあく">獄惡</r>な

蓮池《はすいけ》  
⇒ <r rt="はすいけ">蓮池</r>

## ▶ diy.htd

```
##5 t1, t2 タグ(ブロックレベル要素、開始タグ)
##例 t1(蜘蛛の糸,芥川龍之介)
##      ⇒ <t1 arg1="蜘蛛の糸" arg2="芥川龍之介">
##
t([12])¥(([^\n]+)?,([^\n]+)?,([^\n]+)?)¥           <t$1 arg1="$2" arg2="$3" arg3="$4">
t([12])¥(([^\n]+)?,([^\n]+)?)¥                     <t$1 arg1="$2" arg2="$3">
t([12])¥(([^\n]+)?)¥                                <t$1 arg1="$2">
t([12])¥(¥)   <t$1>
```

t1, t2タグ(開始タグ)の変換  
属性の数ごとに定義している

```
##5 t1, t2 タグ(ブロックレベル要素、終了タグ)
##例 /t1
##      ⇒ </t1>
##
/t([12])    </t$1>
```

t1, t2タグ(終了タグ)の変換

# 参考: 正規表現の説明

---

- ▶ () は、マッチした文字列を記憶
- ▶ 「.」は任意の一文字
- ▶ 「+」は、前接する文字の1回以上の繰り返し
- ▶ 「?」はマッチングの処理を最短で行う
- ▶ \$1, \$2 は、マッチした文字列を展開する。番号は、マッチした位置を表す
- ▶ ¥p{InCJKUnifiedIdeographs} は、1文字の漢字を表す

# 全文表示用の設定

- ▶ 各資料の xslt フォルダ
  - ▶ XSLT と CSS の定義ファイル
- ▶ annotation\_sample の場合
  - ▶ kotobun\_written.xsl (XML → HTML 変換規則)
  - ▶ kotobun\_written.css (スタイルシート)



ブラウザとの表示と見  
比べてみてください

```
<!-- for diy.htm -->
<xsl:template match="t1">  
  <h3><xsl:text>t1:</xsl:text><xsl:value-of select="@arg1"/>/<xsl:value-of select="@arg2"/>/<xsl:value-of select="@arg3"/></h3>  
  <xsl:apply-templates/>  
</xsl:template>  
  
<xsl:template match="t2">  
  <h3><xsl:text>t2:</xsl:text><xsl:value-of select="@arg1"/>/<xsl:value-of select="@arg2"/>/<xsl:value-of select="@arg3"/></h3>  
  <xsl:apply-templates/>  
</xsl:template>  
  
<xsl:template match="e1">  
  <strong><xsl:text>[[</xsl:text><xsl:value-of select="@arg1"/>/<xsl:value-of select="@arg2"/>/<xsl:value-of  
select="@arg3"/>/<xsl:text>]]</xsl:text></strong>  
  <xsl:apply-templates/>  
</xsl:template>
```

t1用の変換規則

t2用の変換規則

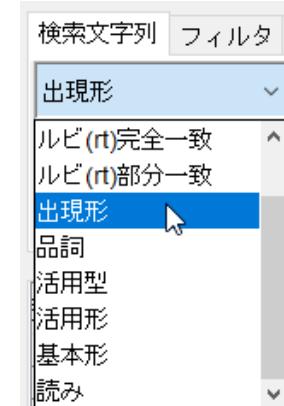
e1用の変換規則

# 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

The search interface has the following fields:

- 検索文字列: 日
- フィルタ: 出現形 (highlighted with a red box)
- 前文脈: で終る
- 後文脈: で始まる

no	前文脈	キー	後文脈	Path
1	は取れんはずである。	一両日	の後地等の土胆はさら	/aozc
2	でございましたのに、	一昨日	コピ-	/aozc
3	眼はその隙間の端に、	一昨日	コピー(列名含む)	/aozc
4	し親子兄弟の離れたる	今日	見付け出	/aozc
5	知れん、しかし太平の	今日	全選択	/aozc
6	はほっと一息ついて「	今日	フィルタ	/aozc
7	静岡から出て来てね、	今日	統計	/aozc

# 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)

検索文字列 フィルタ コーパス 検索オプション

出現形 ^日

前文脈

後文脈

で終る

で始まる

## 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 日\$

前文脈

後文脈

で終る

で始まる

## 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 ^日\$

前文脈

後文脈

で終る

で始まる

## 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

出現形 基本形 歩く

前文脈

後文脈

で終る

で始まる

# 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
  - ▶ 『ひまわり』用パッケージの作成
- ▶ 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
  - ▶ 『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」