全文検索システム『ひまわり』講習会

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

- 既存のテキストデータを全文検索システム『ひまわり』にインポートする方法を紹介
  - 『ひまわり』（ver.1.6.8）+ MeCab（ver.0.996）
  - 青空文庫
  - 日本語諸方言コーパス（COJADS）

- 全体的な流れ
  - 『ひまわり』の紹介と基本的な使い方
  - テキストデータのインポート　＋　形態素解析
    （青空文庫テキストファイル）
  - 『ひまわり』用データの構造と人手アノテーション
  - 自動アノテーション（日本語諸方言コーパスCSVデータ）

ツール・資料などの確認

- 『ひまわり』のインストール
- MeCabのインストール
- テキストエディタのインストール
- 『日本語話し言葉コーパス』サンプルデータのインストール
- 当日配布資料

  - なお, 日本語諸方言コーパス (COJADS) のデータについては, 当日配布資料にサンプルデータを同梱しました

『ひまわり』とは

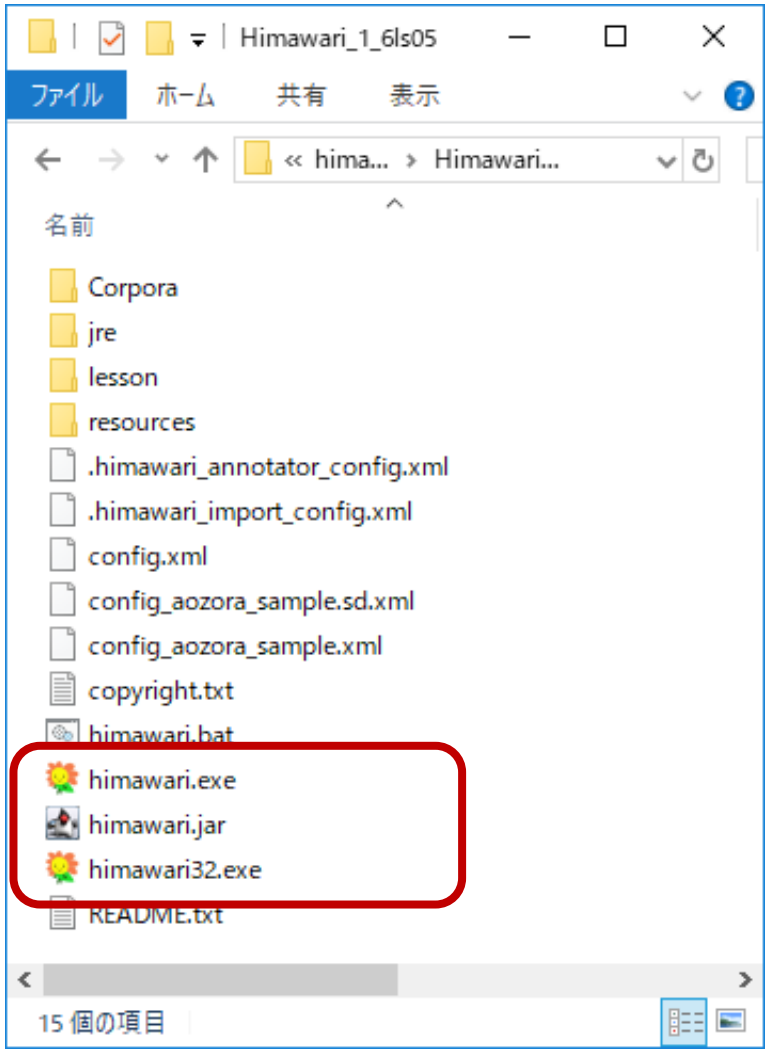
- 言語研究用の全文検索システム
  - 指定された文字列を網羅的に検索して，前後文脈付きで結果を表示します（コンコーダンサ）
  - 『太陽コーパス』（20世紀初頭の総合雑誌『太陽』）用の検索システムとして構築しました

- 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』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（64ビット版）
（Windows 専用）
himawari.exe

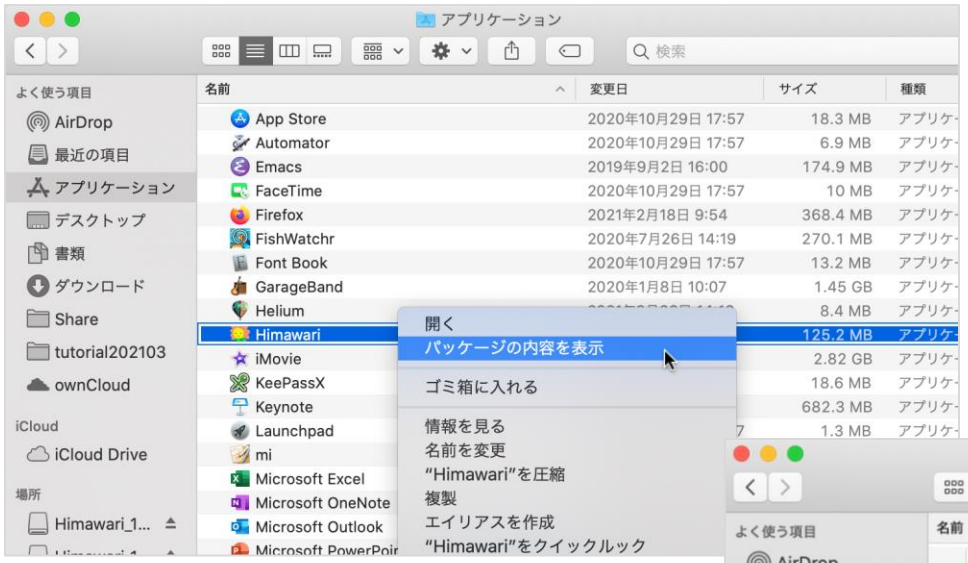
himawari32.exe

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

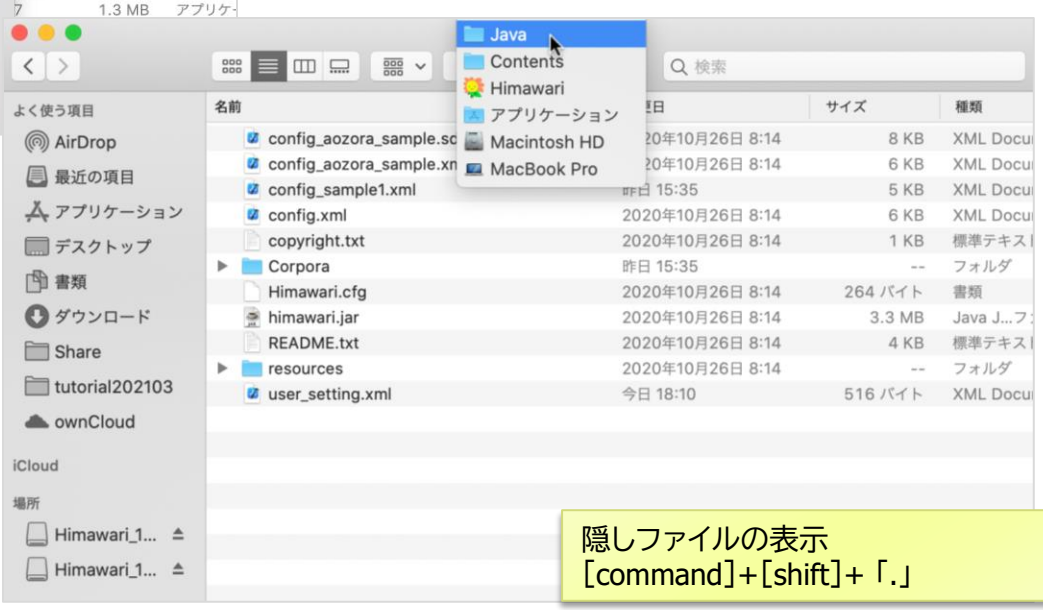
himawari.jar

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

『ひまわり』の起動と『ひまわり』フォルダの確認（macOSの場合）



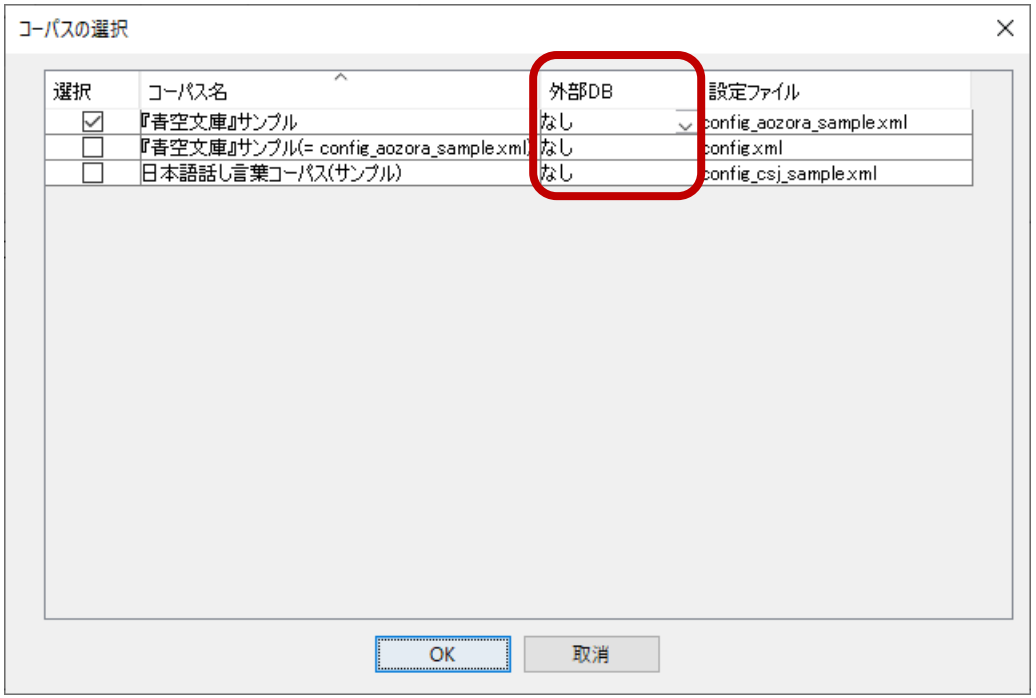
[アプリケーション]→ [Himawari]→ [Contents]→ [Java]



隠しファイルの表示
[command]+[shift]+「.」

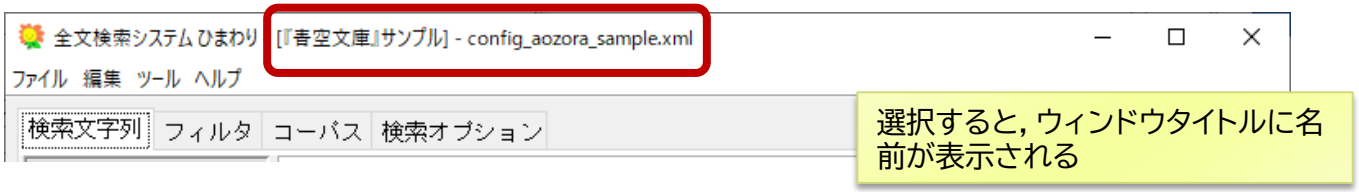
コーパスの選択

- [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

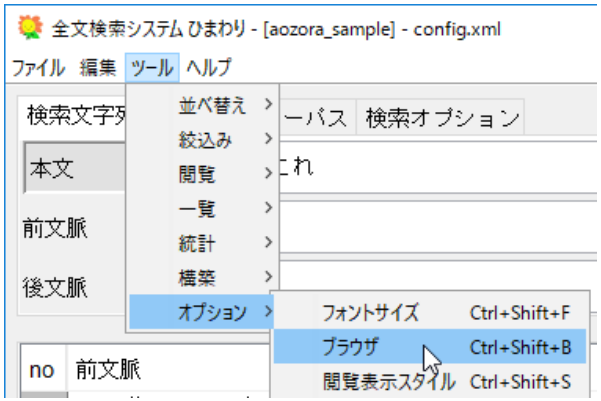

「検索文字列」欄では右クリックで履歴表示
全文検索システム ひまわり - [aozora_sample] - config.xml
ファイル 編集 ツール ヘルプ
検索文字列
検索の実行
検索文字列
本文
これ
前文脈
で終る
後文脈
で始まる
検索
字体変換
クリア
no 前文脈 キー 後文脈 Path タイトル 著者
検索結果
検索総数:597
検索総数
途中経過の表示

ブラウザでの閲覧

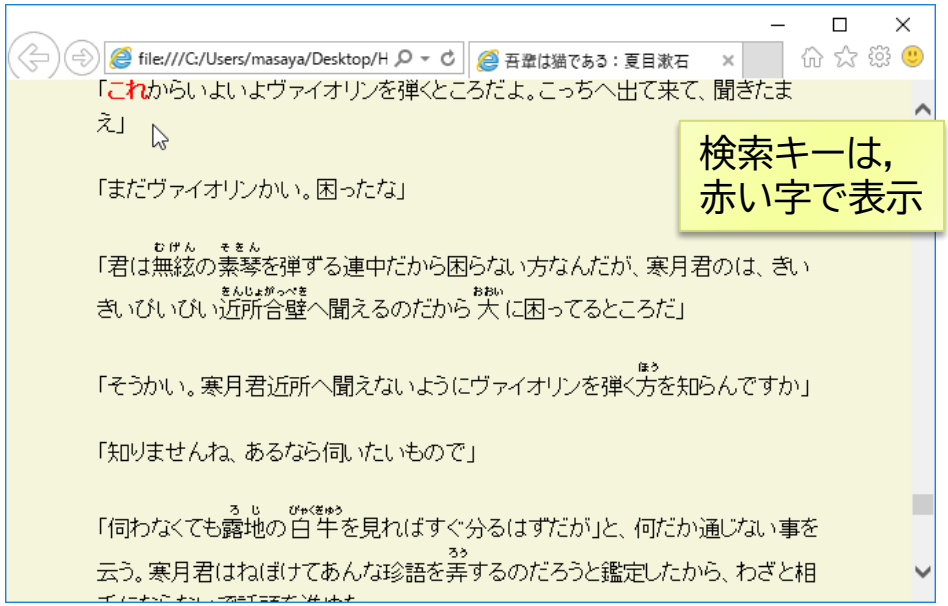
no	前文脈	キー	後文脈	Path
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...
3	弾くところです」「	これ	からいよいよヴァイオ	/aozora_s...
4	い話があるかい」「	これ	からいよいよヴァイオ	/aozora_s...
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aozora_s...
6	見当がつかない」「	これ	からいよいよ弾くとこ	/aozora_s...
7	めちゃなお困ります。	これ	からがいよいよ佳境に	/aozora_s...

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

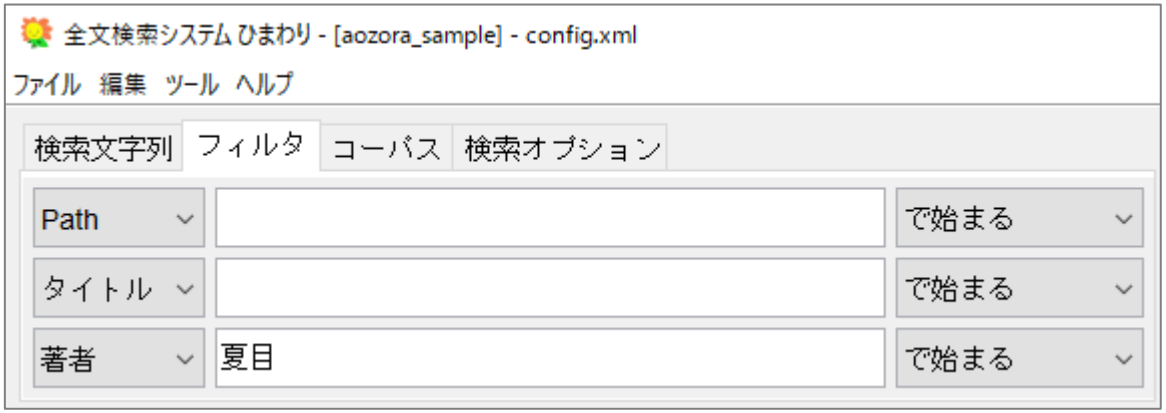
- 昇順
  - 列タイトルをクリック
- 降順
  - シフトキーを押しながら列タイトルをクリック

- 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:「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
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
	これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目
	「	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	「	これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	て、	これ	からいよいよ会津領へ	/aozora_s...	吾輩は猫...	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
	す。	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
⇒フィルタでもOK

検索結果の頻度集計

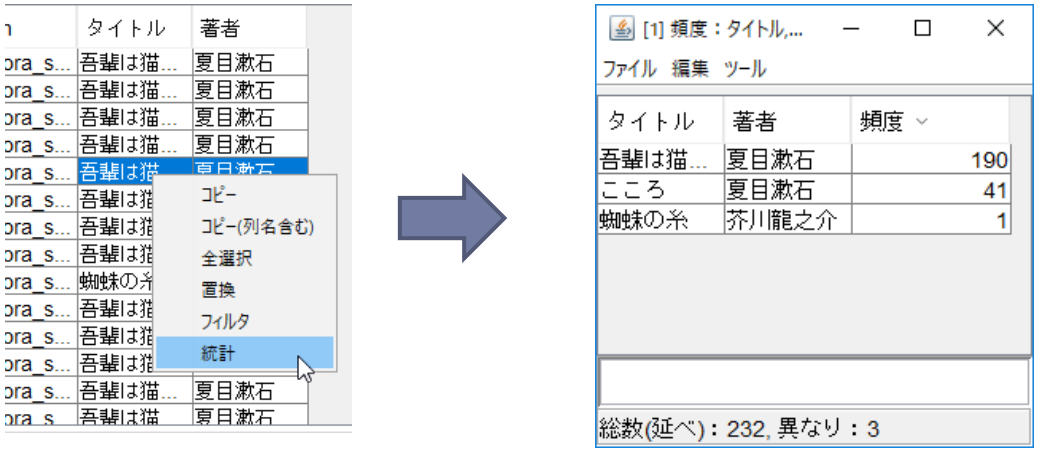
1. 集計したい列を選択

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

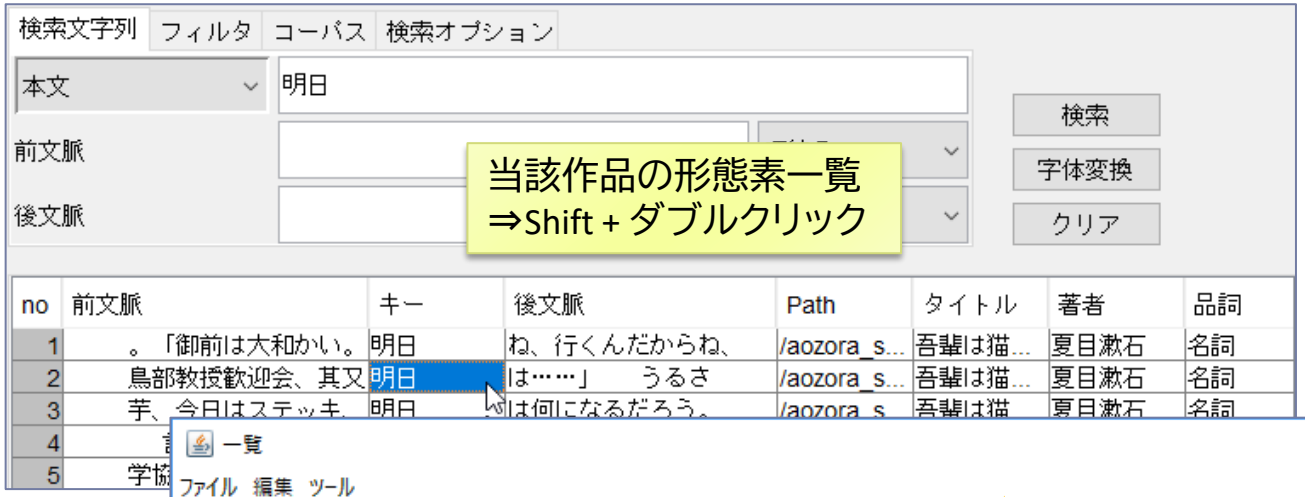
離れた列の選択
- WindowsはCtrlキー
- macOSはcommandキー

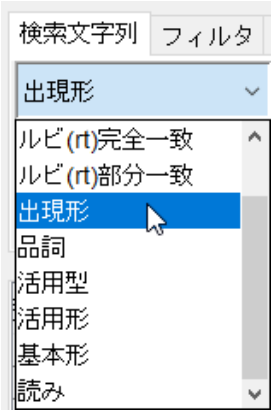
2. 右クリック⇒「統計」

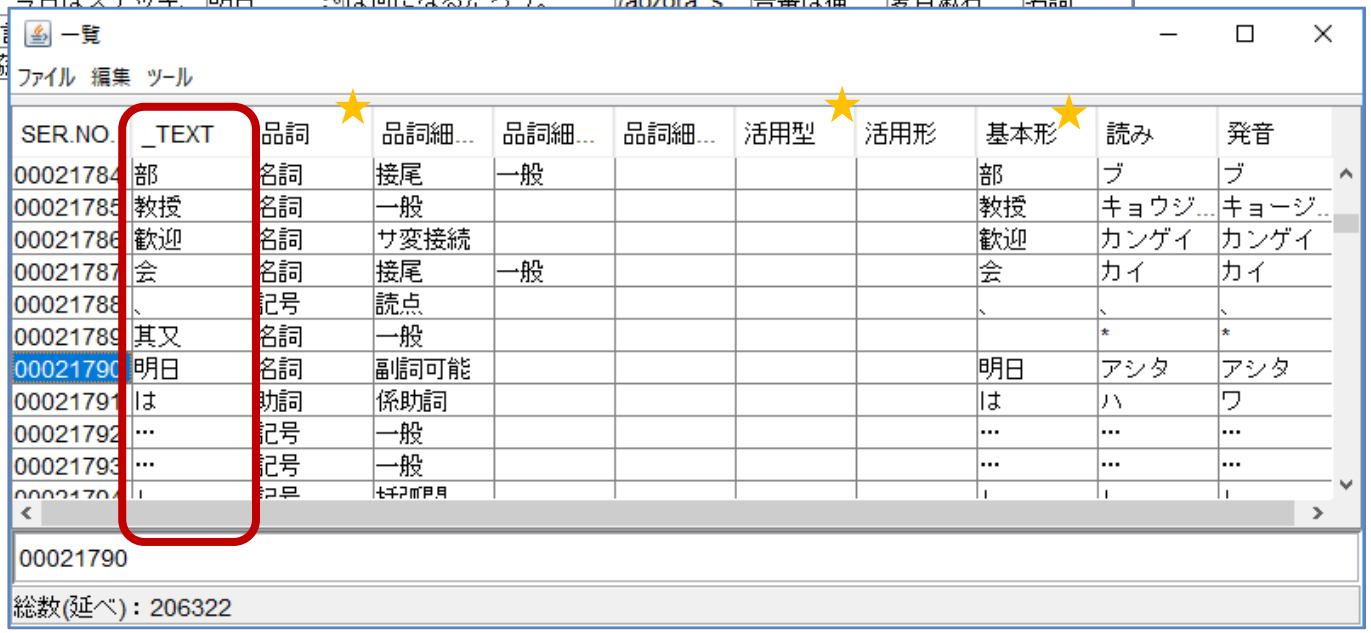


形態素解析結果の閲覧

この機能は，
外部DB「sd」の資料のみ実行可能







テキスト
進行方向

テキストデータのインポートと形態素解析結果のアノテーション

（一般利用者向け）

テキストファイルのインポート
ー青空文庫のテキストデータを例にー

配布資料/samples1/Akutagawa_ryunosuke/kumono_ito.txt

青空文庫の独自タグ
（3種類）

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）

蜘蛛の糸
芥川龍之介

【テキスト中に現れる記号について】

《》：ルビ
（例）蓮池《はすいけ》のふち

｜：ルビの付く文字列の始まりを特定する記号
（例）丁度｜地獄《じごく》の底に

［＃］：入力者注　主に外字の説明や、傍点の位置の指定
　　（数字は、JIS X 0213の面区点番号、または底本のページと行数）
（例）※［＃「特のへん＋𤇾＋聿」、第3水準1-87-71］

［＃8字下げ］一［＃「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。池の中に咲いている蓮《はす》の花は、みんな玉のようにまっ白で、そのまん中にある金色《きんいろ》の蕊《ずい》からは、何とも云えない好《よ》い匂《におい》が、絶間《たえま》なくあたりへ溢《あふ》れて居ります。極楽は丁度朝なのでございましょう。
　やがて御釈迦様はその池のふちに御佇《おたたず》みになって、水の面《おもて》を蔽《おお》っている蓮の葉の間から、ふと下の容子《ようす》を御覧になりました。この極楽の蓮池の下は、丁度｜地獄《じごく》の底に当って居りますから、水晶《すいしよう》のような水を透き徹して、三途《さんず》の河や針の山の景色が、丁度｜覗《のぞ》き眼鏡《めがね》を見るように、はっきりと見えるのでございます。

テキストファイルのインポート
ー青空文庫のテキストデータを例にー

蜘蛛の糸
芥川龍之介

【テキスト中に現れる記号について】

《》:ルビ
(例)蓮池《はすいけ》のふち

｜:ルビの付く文字列の始まりを特定する記号
(例)丁度｜地獄《じごく》の底に

［＃］:入力者注　主に外字の説明や、傍点の位置の指定
（または底本のページと行数）
準1-87-71］

［＃8字下げ］ー［＃「ー」は中見出し］

ルビ，注記は，本文とは区別され，全文検索の対象外となる

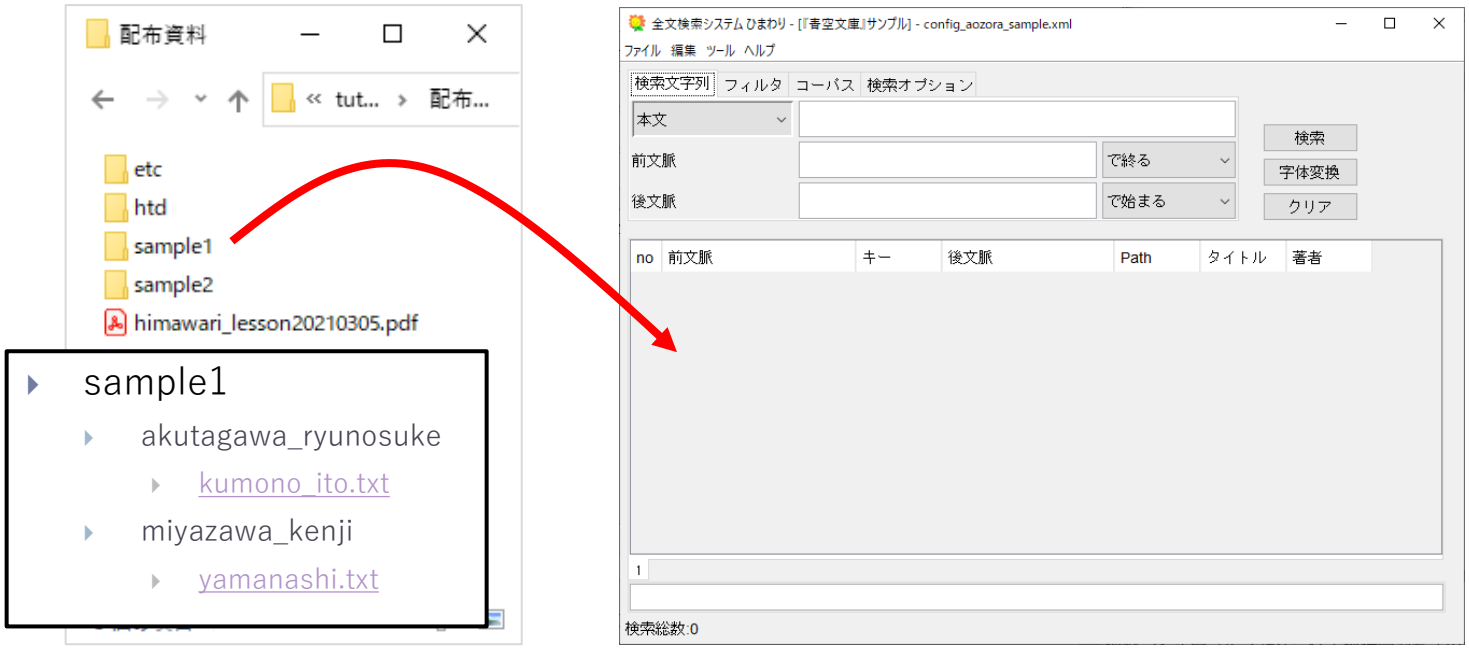
文字列の照合時は，タグは無視される
（例：「跳ねて」にも照合）

　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。池の中に咲いている蓮《はす》の花は、みんな玉のようにまっ白で、そのまん中にある金色《きんいろ》の蕊《ずい》からは、何とも云えない好《よ》い匂《におい》が、絶間《たえま》なくあたりへ溢《あふ》れて居ります。極楽は丁度朝なのでございましょう。
　やがて御釈迦様はその池のふちに御佇《おたたず》みになって、水の面《おもて》を蔽《おお》っている蓮の葉の間から、ふと下の容子《ようす》を御覧になりました。この極楽の蓮池の下は、丁度｜地獄《じごく》の底に当って居りますから、水晶《すいしよう》のような水を透き徹して、三途《さんず》の河や針の山の景色が、丁度｜覗《のぞ》き眼鏡《めがね》を見るように、はっきりと見えるのでございます。

芥川龍之介：蜘蛛の糸，青空文庫（テキストファイル版）

インポートの実行

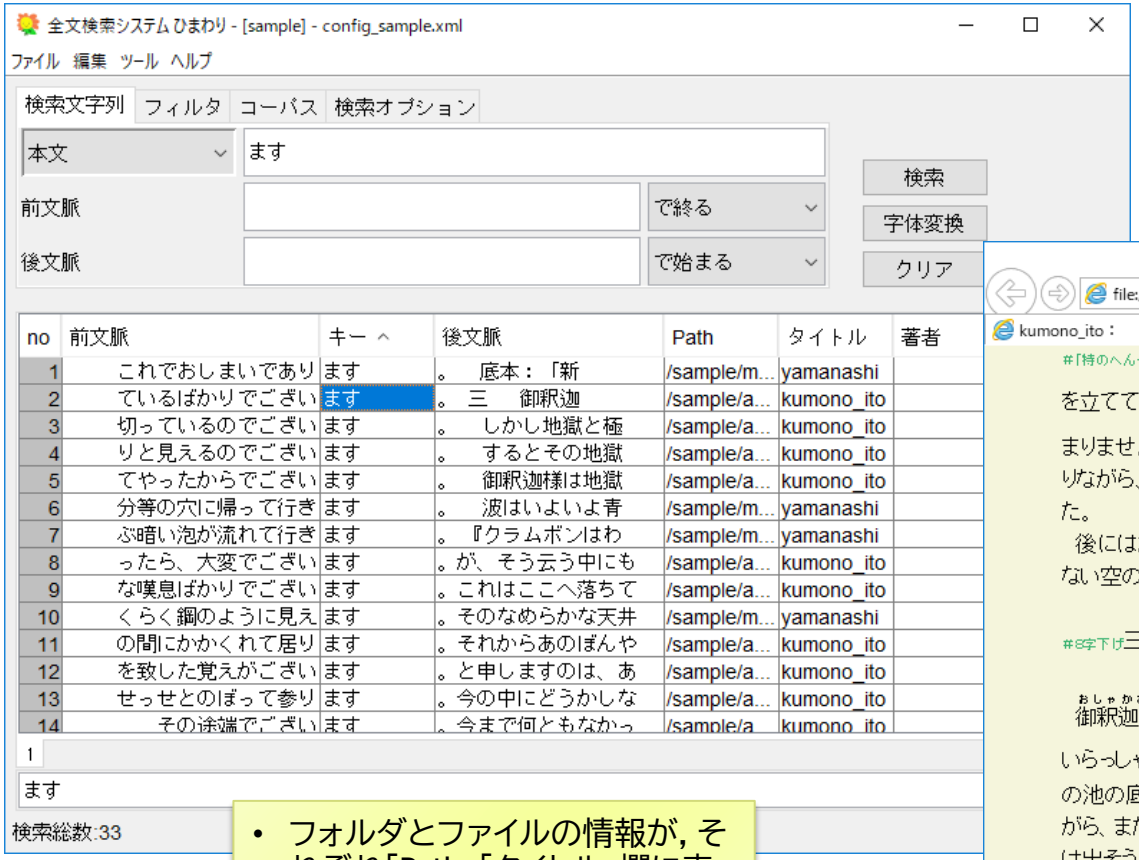
- sample1フォルダを，起動している『ひまわり』にドラッグ＆ドロップ



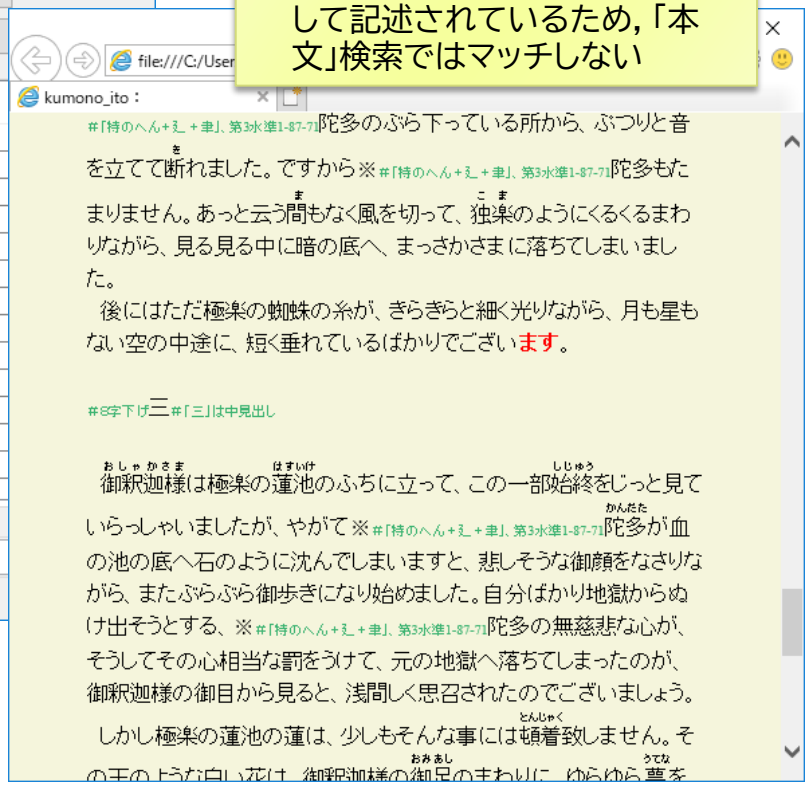
- フォルダの情報をインポート時に利用
  - フォルダ階層 ⇒ Path 欄
  - ファイル名 ⇒ タイトル欄
- ドロップしたフォルダ名がコーパス名になる

- インポート可能なファイル形式（ファイル末尾）
  .txt, .html, .xhtml, .xml
- 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

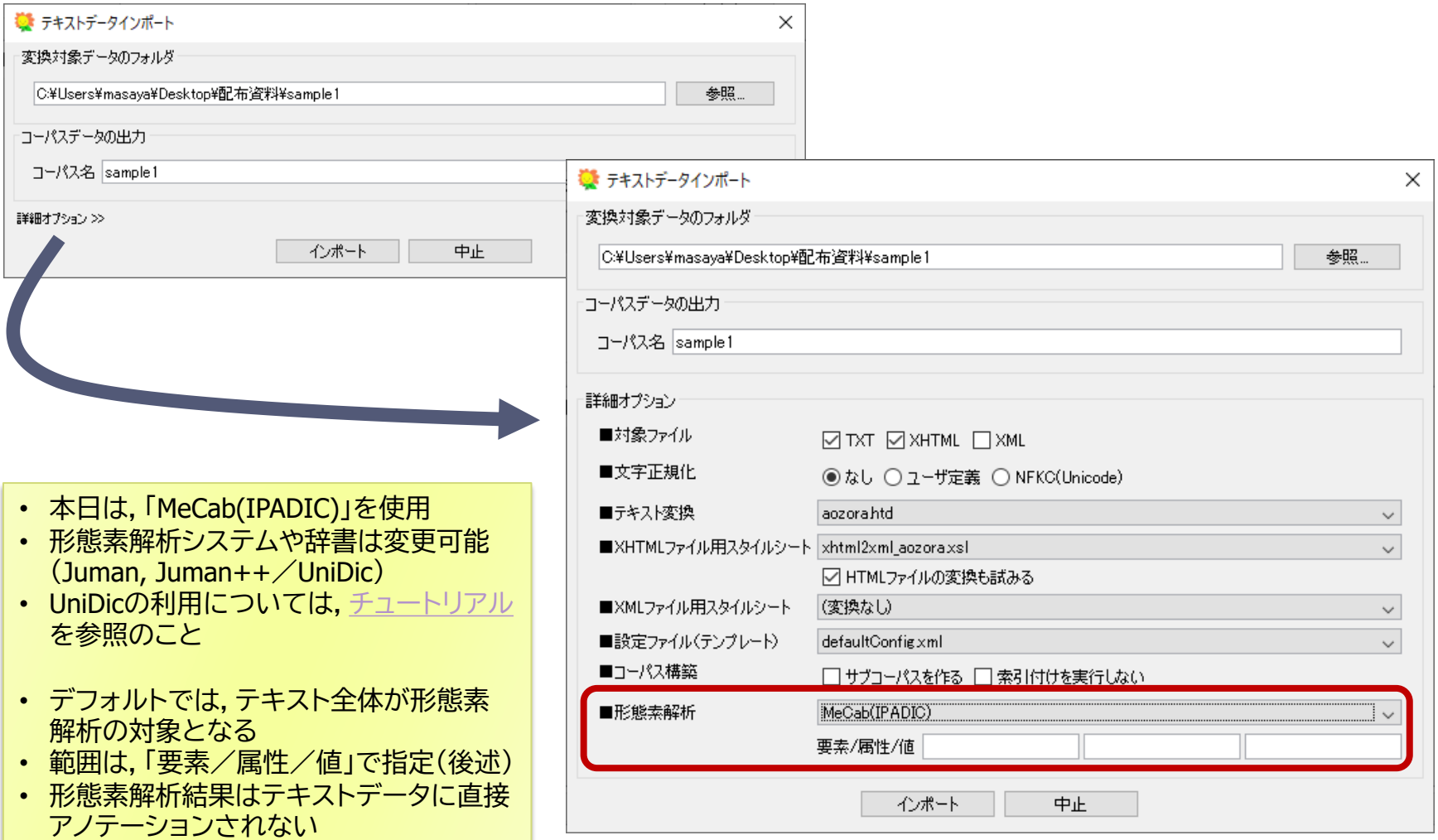


- フォルダとファイルの情報が，それぞれ「Path」「タイトル」欄に表示される
- 「著者」欄は空欄



- 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

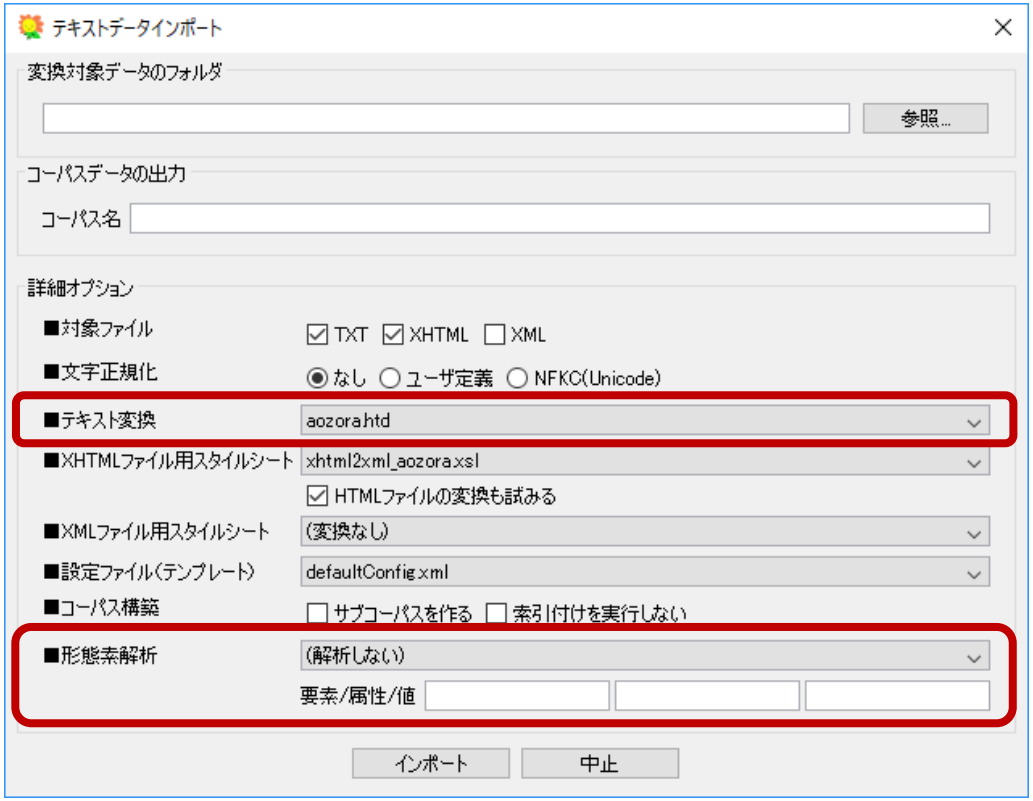
インポート時の形態素解析



- 本日は,「MeCab(IPADIC)」を使用
- 形態素解析システムや辞書は変更可能(Juman, Juman++／UniDic)
- UniDicの利用については, チュートリアルを参照のこと

- デフォルトでは, テキスト全体が形態素解析の対象となる
- 範囲は,「要素／属性／値」で指定(後述)
- 形態素解析結果はテキストデータに直接アノテーションされない

インポート時のオプション



- 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
- XHTMLファイル用スタイルシート
- XMLファイル用スタイルシート
  - XHTML, XML用の変換規則

『ひまわり』用データの構造と人手アノテーション

インポート処理の流れ

① XMLファイルへの変換

- XML（Extensible Markup Language）は，マークアップ言語
- ここでは，タグを自由に定義できるHTMLのようなものだと考えてください

② ファイルの統合

- 変換したXMLファイルを一つのXMLファイルに統合

③ 形態素解析（オプション）

XMLタグの基本

HTMLと異なり，自分でタグを定義できる

開始タグ

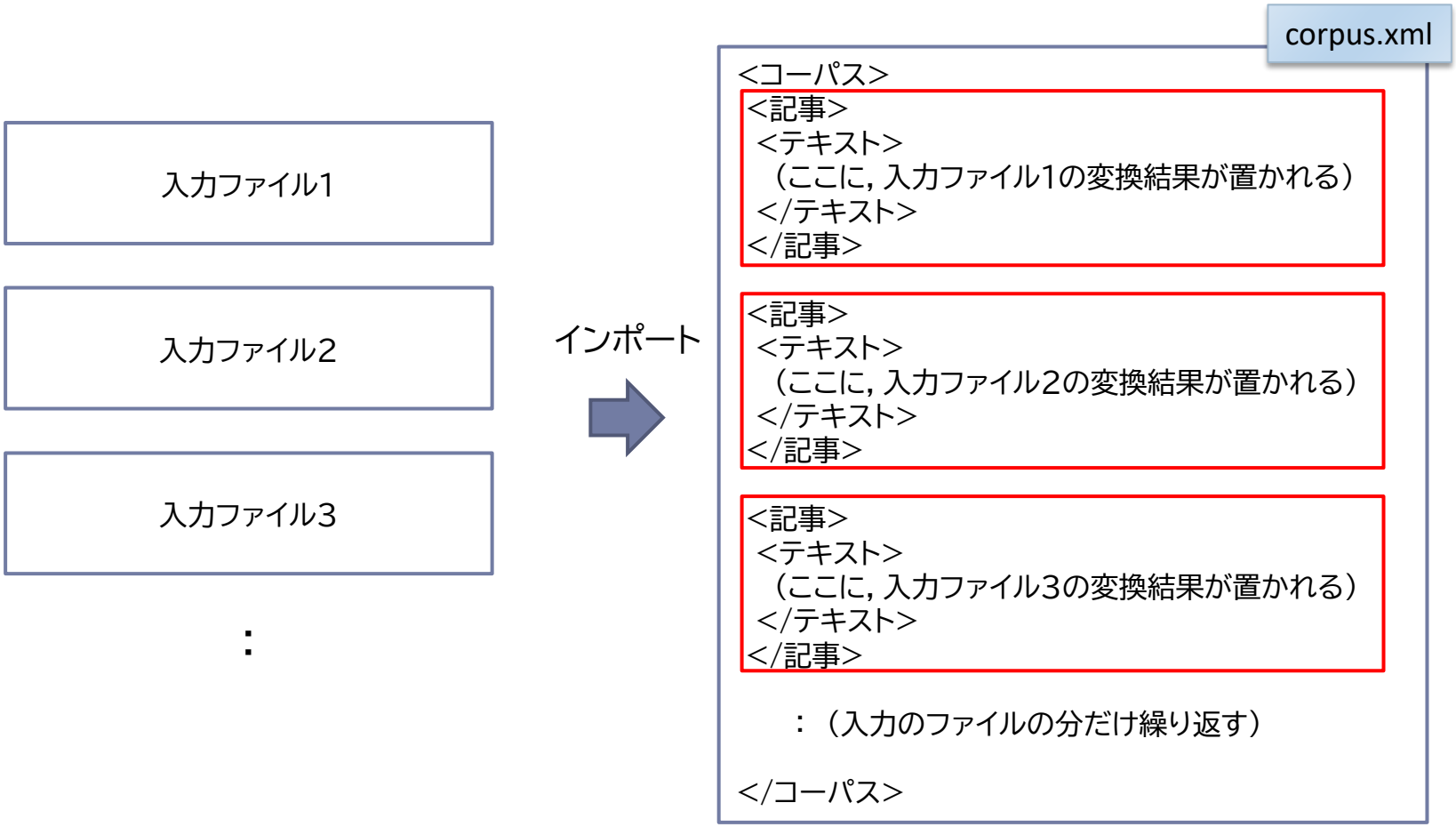
```
〈会話〉

〈発話 話者="太郎"〉こんにちは〈/発話〉
〈発話 話者="次郎"〉おはようございます〈/発話〉
〈ポーズ /〉
〈発話 話者="三郎"〉〈ruby yomi="タロウ"〉太郎〈/ruby〉さん，おはようございます〈/発話〉

〈/会話〉
```

終了タグ

- 一定範囲に意味づけ
  - 「会話」「発話」「ruby」タグ
- 特定の位置に意味付け（範囲がない場合）
  - 「ポーズ」タグ

『ひまわり』は検索時，タグを読み飛ばして，文字列照合する

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>

人手でタグ付けしてみる

- 教材の確認
  - 配布資料/sample2 フォルダ
  - 配布資料/etc/config_sample2.xml

- 配布資料/sample2 フォルダをインポートして下さい

file1.txt

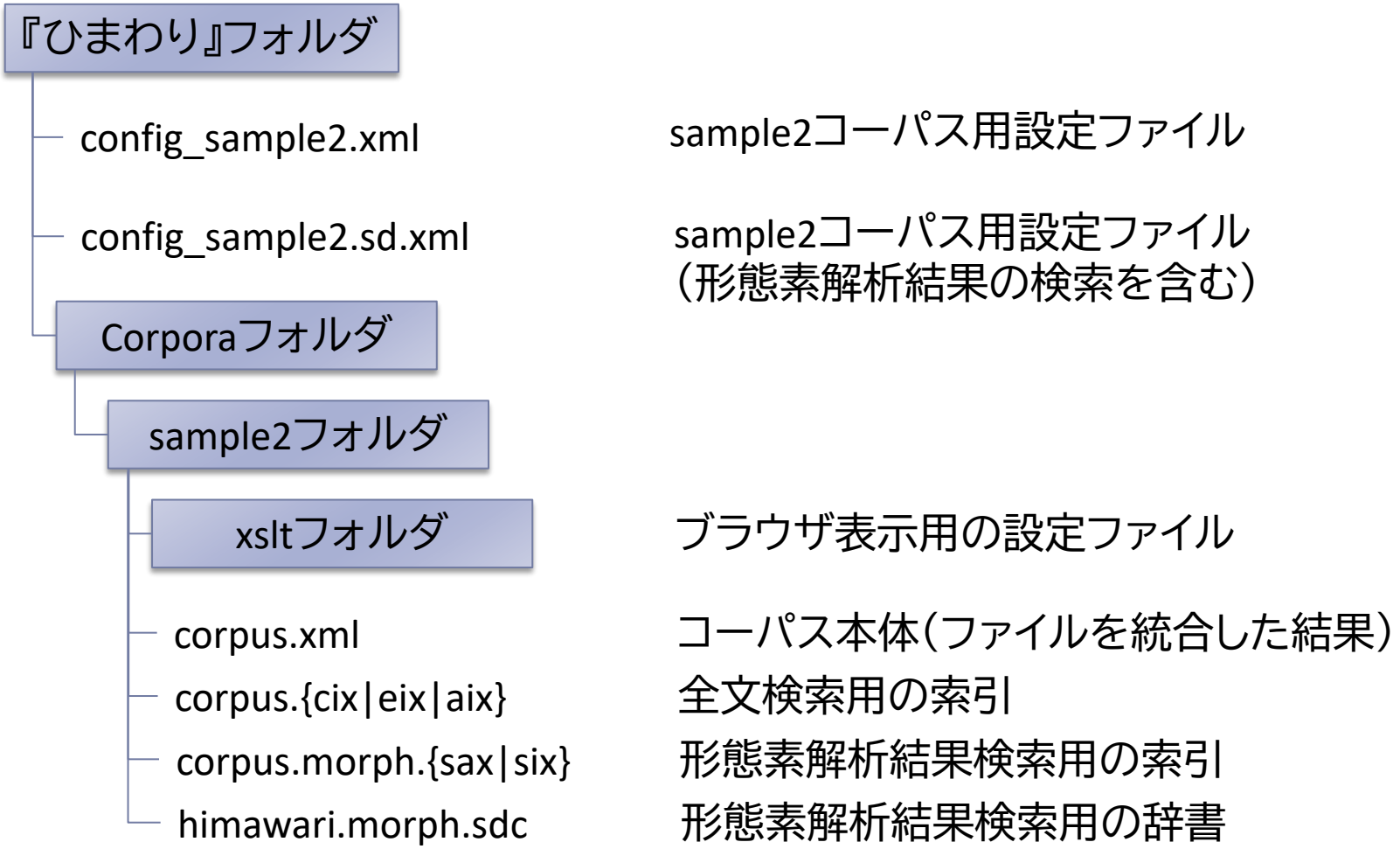
```
F1 まえがきです1
F1 まえがきです2

F1 本文です1
F1 本文です2


太郎 「太郎です。本文です。」
次郎 「次郎です。本文です。」
```

file2.txt

```
F2 まえがきです1
F2 まえがきです2

F2 本文です1
F2 本文です2
F2 本文です3・メタ文字(半角)&<>

F2 あとがきです1
F2 あとがきです2
```

- インポート後，必ず，『ひまわり』を終了してください

生成されるファイル(コーパス名sample2の場合)



エディタで見てみる

Corpora/sample2/corpus.xml

```
<?xml version="1.0" encoding="UTF-16"?>
<コーパス 名前="sample2" 備考="script:Himawari // source:C:/Users/masaya/Desktop/配布資料/sample2 // date:2021-03-02">
<記事 タイトル="file1" 著者="" path="/sample2/file1.txt" 備考="transDataType:テキスト">
<テキスト>
F1　まえがきです1<br />
F1　まえがきです2<br />
<br />
F1　本文です1<br />
F1　本文です2<br />
<br />
<br />
太郎「太郎です。本文です。」<br />
次郎「次郎です。本文です。」<br />
</テキスト>
</記事>

<記事 タイトル="file2" 著者="" path="/sample2/file2.txt" 備考="transDataType:テキスト">
<テキスト>
F2　まえがきです1<br />
F2　まえがきです2<br />
<br />
F2　本文です1<br />
F2　本文です2<br />
F2　本文です3・メタ文字(半角)＆＜＞<br />
<br />
F2　あとがきです1<br />
F2　あとがきです2<br />
</テキスト>
</記事>
</コーパス>
```

タグのメタ文字(&<>)は全角に変換されることに注意

文字コードは, UTF-16(Little Endian, BOM付)
改行コードは, LF

タグ付けしてみる
(「まえがき」「本文」「あとがき」「行末」「発話」タグ)

Corpora/sample2/corpus.xml

```
<?xml version="1.0" encoding="UTF-16"?>
<コーパス 名前="sample2" 備考="script:Himawari // source:C:/Temp/tutorial202103/配布資料/samples2 // date:2021-03-01">
<記事 タイトル="file1" 著者="" path="/samples1/file1.txt" 備考="transDataType:テキスト">
<テキスト>
<まえがき>
F1　まえがきです1<行末 番号="1" /><br />
F1　まえがきです2<行末 番号="2" /><br />
</まえがき>
<br />
<本文>
F1　本文です1<br />
F1　本文です2<br />
F1　本文です3<br />
<br />
<br />
<発話 話者="太郎">「太郎です。本文です。」</発話><br />
<発話 話者="次郎">「次郎です。本文です。」</発話><br />
</本文>
</テキスト>
</記事>
                    :(中略)
</コーパス>
```

- ✓ タグ付け前に『ひまわり』は終了してください
- ✓ mi工ディタの場合, [編集]→[自動インデントを一時的に無効にする]

- ✓ タグの<>/="は半角
- ✓ タグの範囲が交差するような記述は不可
  例:<本文><発話> … </本文></発話>

- タグ付けの体験が目的ですので, 全部のタグをつける必要はありません
- 編集が終わったら, 上書き保存してください

形式の検証（validation）

- 整形式（Well-formed ）のXMLファイルであることを検証
- 簡易的な方法ですが, Firefox, もしくは, IE に corpus.xml をドラッグ＆ドロップしてください

```
<?xml version="1.0" encoding="UTF-16"?>
- <コーパス 備考="script:Himawari // source:C:/Users/masaya/Desktop/配布資料/sample2 // date:2021-03-02" 名前="sample2">
  - <記事 備考="transDataType:テキスト" path="/sample2/file1.txt" 著者="" タイトル="file1">
    - <テキスト>
      - <まえがき>
          F1 まえがきです1
          <br/>
          F1 まえがきです2
          <br/>
        </まえがき>
        <br/>
      - <本文>
          F1 本文です1
          <br/>
          F1 本文です2
          <br/>
          <br/>
          <br/>
          <発話 話者="太郎">「太郎です。本文です。」</発話>
          <br/>
          <発話 話者="次郎">「次郎です。本文です。」</発話>
          <br/>
        </本文>
    </テキスト>
  </記事>
  - <記事 備考="transDataType:テキスト" path="/sample2/file2.txt" 著者="" タイトル="file2">
    - <テキスト>
      - <まえがき>
```

検証に成功した場合は, 木構造で表示される

設定ファイルの調整

- 独自に追加したタグは, 自分で設定を調整

- 配布資料/etc/config_sample2.xml を『ひまわり』フォルダにコピー
  - なお, インポートするたびに, 設定ファイルやcorpus.xmlは自動生成されるため, 人手で作成したファイルはバックアップを取っておいてください

設定ファイルの調整
（全文検索対象のタグ用の索引）

- 「まえがき」「本文」「あとがき」

config_sample2.xml

```
<index_cix>
  <li field_name="キー" label="全体" name="テキスト" middle_name="article" type="normal"/>
  <li field_name="キー" label="全体(正規表現)" name="テキスト" middle_name="article" type="null"/>
  <li field_name="キー" label="まえがきのみ" name="まえがき" middle_name="maegaki" type="normal"/>
  <li field_name="キー" label="本文のみ" name="本文" middle_name="honbun" type="normal"/>
  <li field_name="キー" label="あとがきのみ" name="あとがき" middle_name="atogaki" type="normal"/>
 </index_cix>
```

- label: 検索メニュー用のラベル
- name: 全文検索対象のタグ名
- middle_name: 他の設定値と重複しない文字列
- type: 全文検索の種類
  - normal: 通常の全文検索
  - null:正規表現を用いた全文検索（検索速度は遅い）

詳細は，「設定ファイルリファレンスマニュアル」参照

設定ファイルの調整（検索結果の列）

- 行末/@番号, 発話/@話者

config_sample2.xml

```
<!-- 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="著者" element="記事" name="著者" type="argument" width="80"/>
  <li attribute="話者" element="発話" name="話者" width="120"/>
  <li attribute="番号" element="行末" name="行番号" width="120"/>
</field_setting>
```

- attribute:　タグの属性名
- element:　タグ名
- name:　検索結果の列名

詳細は，「設定ファイルリファレンスマニュアル」参照

設定ファイルの調整
（検索結果表示用の索引）

- 発話, 行末

config_sample2.xml

```
<!-- 要素への索引 -->
 <index_eix>
  <li name="コーパス" middle_name="corpus" is_empty="false" top="true"/>
  <li name="記事" middle_name="article" isBrowsed="true" is_empty="false"/>
  <li name="発話" middle_name="utterance" is_empty="false"/>
  <li name="行末" middle_name="line" is_empty="true"/>
 </index_eix>
```

- name:　　タグ名
- middle_name:　他の設定と重複しない文字列
- is_empty:　　true→範囲があるタグ, false→範囲のないタグ

詳細は,「設定ファイルリファレンスマニュアル」参照

検索結果の表示に使うタグは, この設定をしてください

設定ファイルの調整（属性検索用の索引）

- 発話/@話者, 行末/@番号

config_sample2.xml

```
<!-- 要素属性への索引 -->
 <index_aix>
  <li name="r" argument="rt" isCompleteMatch="true" label="ルビ(rt)完全一致" middle_name="r" field_name="キー"/>
  <li name="r" argument="rt" isCompleteMatch="false" label="ルビ(rt)部分一致" middle_name="r2" field_name="キー"/>
  <li name="発話" argument="話者" isCompleteMatch="false" label="話者検索" middle_name="r" field_name="キー"/>
  <li name="行末" argument="番号" isCompleteMatch="true" label="行番号検索" middle_name="r2" field_name="キー"/>
 </index_aix>
```

- name: タグ名
- argument: 検索する属性名
- isCompleteMatch: true→完全一致検索, false→部分一致検索
- middle_name: 他の設定と重複しない文字列

詳細は,「設定ファイルリファレンスマニュアル」参照

索引付け

- 検索の高速化のために必要
- 索引付けが必要な場合
  - corpus.xmlを変更したとき
  - 形態素解析をするとき
  - 設定ファイルで索引の設定を変更したとき

- [ツール] ⇒ [構築] ⇒ [インデックス生成]

自動アノテーション
（自動形式変換）

インポート時の自動形式変換

- 背景
  - 人手でXMLタグを書くのは大変
  - 青空文庫のような独自タグや，カンマ区切りテキスト（CSVデータ）などをXMLに自動的に変換したい
- 『ひまわり』がサポートする変換方法
  - 正規表現を用いた文字列置換
    - 本日は，こちらをメインに扱います
    - 日本語諸方言コーパス（COJADS）
  - XSL（Extensible Stylesheet Language）による変換
    - XMLデータを『ひまわり』用のXMLデータに変換する
    - 『ひまわり』は，青空文庫のXHTML（＝XML）版用の変換規則を標準で同梱（XMLでアノテーションされている情報を自動で取り込める）

『ひまわり』にインポート可能なコーパス

- 日本語諸方言コーパス（COJADS）のCSVデータ
  - 正規表現を用いた文字列置換
- 多言語母語の日本語学習者横断コーパス（I－JAS）
  - インポートのみで変換は行っていない
- BNC（British National Corpus）コーパス
  - XSLによる変換
- TED字幕テキスト
  - XSLによる変換＋外部スクリプト

- リンク先のページでは，変換手順を解説しています
- 変換の仕組みは解説していませんが，興味のある方は試してみてください

テキストエディタで正規表現置換

- 処理のイメージをつかむため, テキストエディタで置換してみます

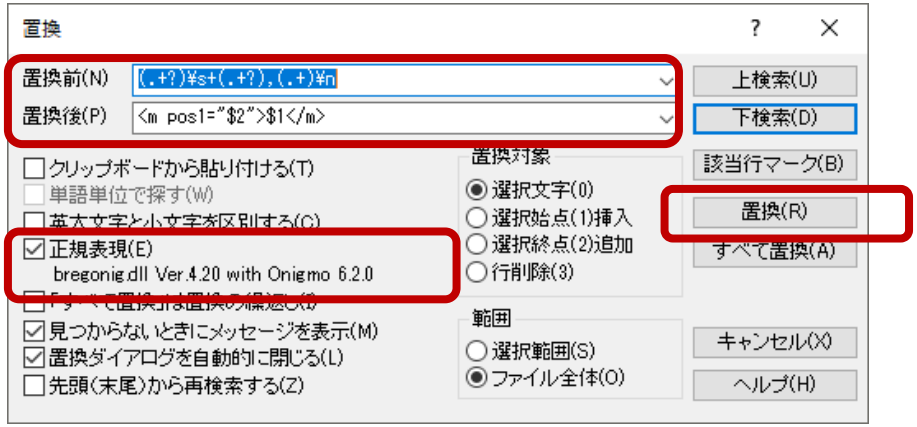
私　名詞,代名詞
の　助詞,連体化
名前　名詞,一般
は　助詞,係助詞
太郎　名詞,固有名詞
です　助動詞,*

配布資料/etc/test_regex.txt

変換前：(.+?)¥s+(.+?),(.+)¥n
変換後：<m pos1="$2">$1</m>

¥sは, 空白文字やタブ
¥nは, 改行
$1は, 一つ目の( ) にマッチする文字列

- サクラエディタ
[検索] → [置換]



- mi
[検索] → [検索／置換ダイアログを表示]

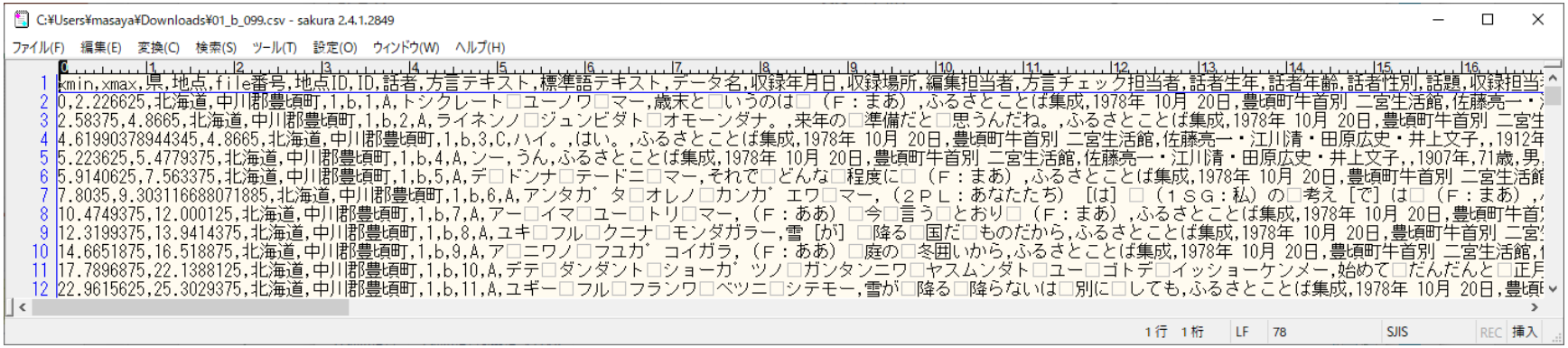
日本語諸方言コーパス(COJADS) CSVデータの利用

- Excelでの表示

01_b_099.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	xmin	xmax	県	地点	file番号	地点ID	ID	話者	方言テキスト	標準語テキスト	データ名	収録年月日	収録場所	編集担当者	方言チェッ	話者生年	話者年齢
2	0	2.226625	北海道	中川郡豊頃	1	b	1	A	トシクレート　ユーノワ　マー	歳末と　いうのは　（Ｆ：まあ）	ふるさとこ	1978年 10	豊頃町牛首	佐藤亮一・江川清・F		1907年	71歳
3	2.58375	4.8665	北海道	中川郡豊頃	1	b	2	A	ライネンノ　ジュンビダト　オモー	来年の　準備だと　思うんだね。	ふるさとこ	1978年 10	豊頃町牛首	佐藤亮一・江川清・F		1907年	71歳
4	4.619904	4.8665	北海道	中川郡豊頃	1	b	3	C	ハイ。	はい。	ふるさとこ	1978年 10	豊頃町牛首	佐藤亮一・江川清・F		1912年	66歳
5	5.223625	5.477938	北海道	中川郡豊頃	1	b	4	A	ンー	ン	ふるさとこ	1978年 10	豊頃町牛首	佐藤亮一・江川清・F		1907年	71歳
6	5.914063	7.563375	北海道	中川郡豊頃	1	b	5	A	デ　ドンナ　テードニ　マー	それで　どんな　程度に　（Ｆ：まあ	ふるさとこ	1978年 10	豊頃町牛首	佐藤亮一・江川清・F		1907年	71歳
7	7.8035	9.303117	北海道	中川郡豊頃	1	b	6	A	アンタガ゛タ　オレノ　カンカ゛エ	（２ＰＬ：あなたたち）［は］　（1	ふるさとこ	1978年 10	豊頃町牛首	佐藤亮一・江川清・F		1907年	71歳
8	10.47494	12.00013	北海道	中川郡豊頃	1	b	7	A	アー　イマ　ユー　トリ　マー	（Ｆ：ああ）　今　言う　とおり　（	ふるさとこ	1978年 10	豊頃町牛首	佐藤亮一・江川清・F		1907年	71歳
9	12.31994	13.94144	北海道	中川郡豊頃	1	b	8	A	ユキ　フル　クニナ　モンダガラー	雪［が］　降る　国だ　ものだから	ふるさとこ	1978年 10	豊頃町牛首	佐藤亮一・江川清・F		1907年	71歳

- サクラエディタでの表示



```
xmin,xmax,県,地点,file番号,地点ID,ID,話者,方言テキスト,標準語テキスト,データ名,収録年月日,収録場所,編集担当者,方言チェック担当者,話者生年,話者年齢,話者性別,話題,収録担当
0,2.226625,北海道,中川郡豊頃町,1,b,1,A,トシクレート□ユーノワ□マー,歳末と□いうのは□（Ｆ：まあ）,ふるさとことば集成,1978年 10月 20日,豊頃町牛首別 二宮生活館,佐藤亮一・
2.58375,4.8665,北海道,中川郡豊頃町,1,b,2,A,ライネンノ□ジュンビダト□オモーンダナ。,来年の□準備だと□思うんだね。,ふるさとことば集成,1978年 10月 20日,豊頃町牛首別 二宮生
4.61990378944345,4.8665,北海道,中川郡豊頃町,1,b,3,C,ハイ。,はい。,ふるさとことば集成,1978年 10月 20日,豊頃町牛首別 二宮生活館,佐藤亮一・江川清・田原広史・井上文子,,1912年
5.223625,5.4779375,北海道,中川郡豊頃町,1,b,4,A,ンー,うん,ふるさとことば集成,1978年 10月 20日,豊頃町牛首別 二宮生活館,佐藤亮一・江川清・田原広史・井上文子,,1907年,71歳,男,
5.9140625,7.563375,北海道,中川郡豊頃町,1,b,5,A,デ□ドンナ□テードニ□マー,それで□どんな□程度に□（Ｆ：まあ）,ふるさとことば集成,1978年 10月 20日,豊頃町牛首別 二宮生活館
7.8035,9.303116688071885,北海道,中川郡豊頃町,1,b,6,A,アンタガ゛タ□オレノ□カンカ゛エワ□マー,（２ＰＬ：あなたたち）［は］□（１ＳＧ：私）の□考え［で］は□（Ｆ：まあ）,
10.4749375,12.000125,北海道,中川郡豊頃町,1,b,7,A,アー□イマ□ユー□トリ□マー,（Ｆ：ああ）□今□言う□とおり□（Ｆ：まあ）,ふるさとことば集成,1978年 10月 20日,豊頃町牛首
12.3199375,13.9414375,北海道,中川郡豊頃町,1,b,8,A,ユキ□フル□クニナ□モンダガラー,雪［が］□降る□国だ□ものだから,ふるさとことば集成,1978年 10月 20日,豊頃町牛首別 二宮
14.6651875,16.518875,北海道,中川郡豊頃町,1,b,9,A,ア□ニワノ□フユカ゛コイガラ,（Ｆ：ああ）□庭の□冬囲いから,ふるさとことば集成,1978年 10月 20日,豊頃町牛首別 二宮生活館,
17.7896875,22.1388125,北海道,中川郡豊頃町,1,b,10,A,デテ□ダンダント□ショーカ゛ツノ□ガンタンニワ□ヤスムンダト□ユー□ゴトデ□イッショーケンメー,始めて□だんだんと□正月
22.9615625,25.3029375,北海道,中川郡豊頃町,1,b,11,A,ユギー□フル□フランワ□ベツニ□シテモー,雪が□降る□降らないは□別に□しても,ふるさとことば集成,1978年 10月 20日,豊頃
```

正規表現置換を用いた『ひまわり』用XMLへの変換

- 変換前と変換後のデータ

```
xmin,xmax,県,地点,file番号,地点ID,ID,話者,方言テキスト,標準語テキスト,データ名,収録年月日,...
0,2.226625,北海道,中川郡豊頃町,1,b,1,A,トシクレート　ユーノワ　マー,歳末と　いうのは　（F:まあ),ふるさとことば集成,1978年 10月 20日,豊頃町牛首別 二宮生活館,佐藤亮一・江川清・田原広史・井上文子,,1907年,71歳,男,年中行事, 昔の生活の様子,小野米一,自然談話,,,,,,,,

<utterance xmin="0" xman="2.226625" ken="北海道" chiten="中川郡豊頃町" ....>
  <standard>歳末と　いうのは　（F:まあ)</standard>
  <dialect>トシクレート　ユーノワ　マー</dialect>
</utterance>
```

01_b_099.csv

- 変換規則ファイル（配布資料/htd/cojads_sample.htd）

```
## ver.20210224
## 最初の行を削除
^xmin.*¥n
## 残りの行
(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?),(.*?)¥n<utterance xmin="$1" xmax="$2" ken="$3" chiten="$4" file_bango="$5" chiten_id="$6" id="$7" washa="$8" hogen_text="$9" hyojungo_text="$10" data_mei="$11" shuroku_nen="$12" shuroku_basho="$13" henshu_tantosha="$14" hogen_check="$15" washa_seinen="$16" washa_nenrei="$17" washa_seibetsu="$18" wadai="$19" shuroku_tantosha="$20" danwa_genre="$21" chu1_bango="$22" chu1_gokei="$23" chu1_chushaku="$24" chu2_bango="$25" chu2_gokei="$26" chu2_chushaku="$27" chu3_bango="$28" chu3_gokei="$29" chu3_chushaku="$30">		¥n<standard>$10</standard>		¥n<dialect>$9</dialect>		¥n</utterance>¥n
```

黄色い部分はタブ文字。この前後が変換前・変換後の置換規則となる

インポート時の正規表現置換の概要

- 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
  - 『ひまわり』フォルダ/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

実際に変換してみる

① 配布資料/htd/cojads_sample.htdを
『ひまわり』フォルダ/resources/htd にコピー

② CSV データ Ver.2021.01の配布ページから, 好みのファイル(複数可)をダウンロードし, 配布資料/htd/COJADS_TEST フォルダにコピーしてください。その際, ファイル名末尾(拡張子)の .csv を .txt にしてください。

② 配布資料/htd/COJADS_TEST フォルダを
『ひまわり』にドラッグ&ドロップして, インポート
- 「テキスト変換」オプションは, cojadas_sample.htdを選択
- ~~同梱データは, 01_b_099.csv から5行引用したもの~~

③ 配布資料/htd フォルダから, 次のファイルをコピー
- config_COJADS_TEST.xml ⇒ 『ひまわり』フォルダへ
- xslt フォルダ ⇒
『ひまわり』フォルダ/Corpora/COJADS_TEST フォルダへ

④ [ファイル] → [コーパス選択]で COJADS_TEST を選択
⑤ [ツール] → [構築] → [インデックス生成] を実行

ブラウザ表示用の設定

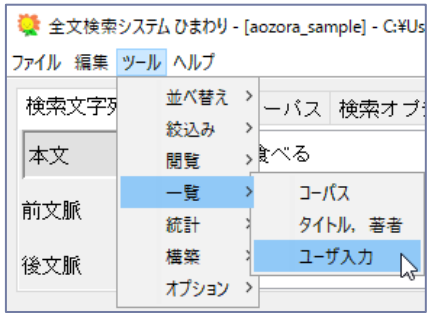
- 各資料の xsltフォルダ
  - XSLTとCSSの定義ファイル
- COJADS_TEST の場合
  - cojads_sample.xsl（XSL変換による XML→HTML変換規則）
  - cojads_sample.css（Cascade Style Sheet）

- XSL, CSS 共に標準規格なので，参考資料の入手は容易
- 各資料の xslt フォルダを参照してみてください

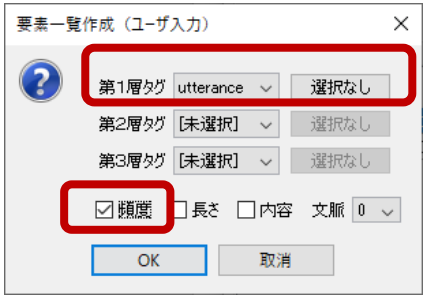
アノテーション結果の集計

- 一覧機能(ユーザ入力)で付与されたタグの情報を閲覧

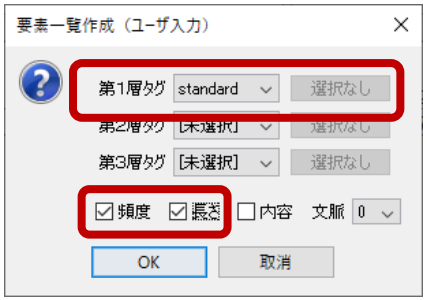
集計機能の詳細は,「利用者マニュアル」参照



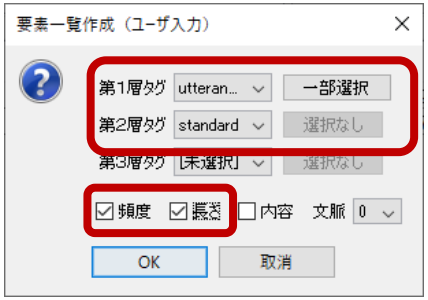
- 発話数



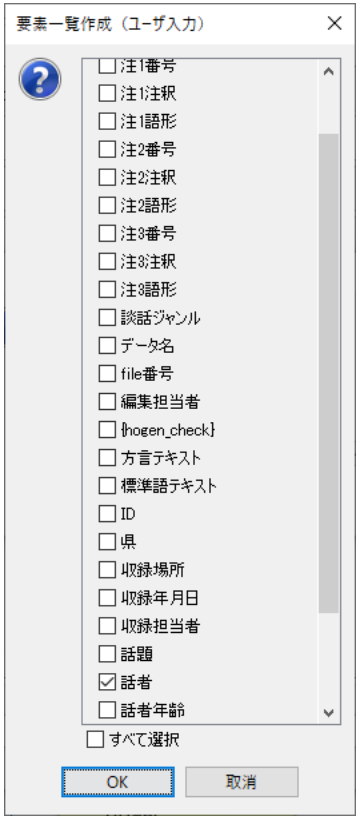
- 発話文字数(標準語)



- 発話文字数(話者別・標準語)



集計後,「standard」列のデータをどれか一つ選択して,[編集]→合算で「長さ」を合算する



おわりに

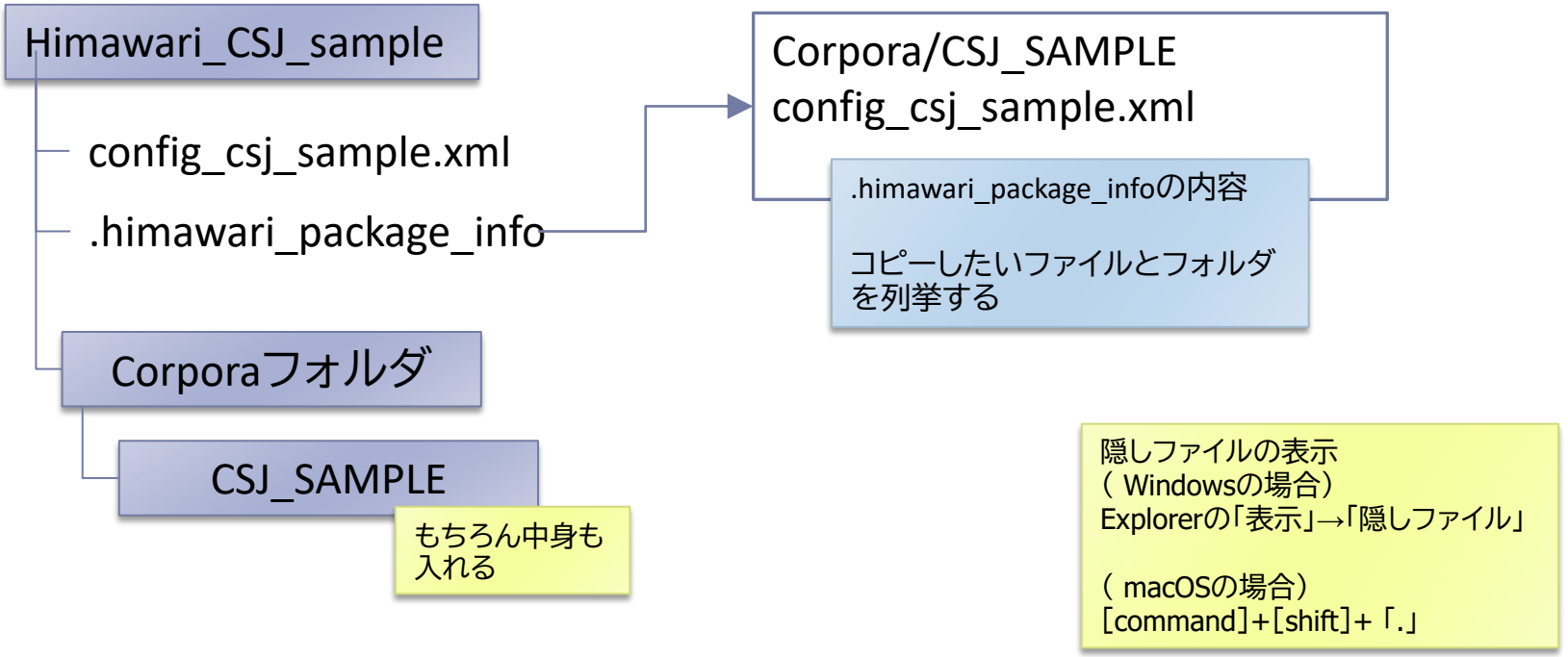
- 全文検索システム『ひまわり』チュートリアル（作成中心）
  - 『ひまわり』の紹介と基本的な使い方
  - テキストデータのインポートと形態素解析
  - 『ひまわり』用データの構造と人手アノテーション
  - 自動アノテーション（自動形式変換）

- 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』用各種パッケージや変換パッケージ
  - スクリプトによ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

参考資料

参考:『ひまわり』用パッケージの構造

- 日本語話し言葉コーパス(サンプル)の場合
  - 次の構造のフォルダを作成し, zipで圧縮



詳細は,「設定ファイルリファレンスマニュアル」参照

参考：変換規則でよく使う正規表現

- () 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

各種設定ファイル＆参考資料

- 『ひまわり』関連
  - 利用者マニュアル
  - 設定ファイルリファレンスマニュアル
  - ビデオチュートリアル
- XSL関連
  - サンプルで覚えるXSLTプログラミング
  - XSLTスタイルシートの基礎の基礎
  - スタイルシート入門（CSS）
- 正規表現関連
  - Java正規表現の使い方
  - Java Pattern クラス