全文検索システム『ひまわり』を用いた既存資料の活用

山口昌也（国立国語研究所）

第7回コーパス日本語学ワークショップにおけるチュートリアル資料（2015-03-10）

本日の内容

- 準備状況の確認
- 全文検索システム『ひまわり』の簡単な紹介
- 既存資料のインポート
  - 生のテキストをそのままインポート
  - 生のテキストの構造を生かしたインポート
- インポートした資料の活用

インポート例

- 『CD-毎日新聞データ集』
- 米国議会図書館蔵
『源氏物語』
- 『青空文庫』パッケージ
- 日本語話し言葉コーパス

米国議会図書館蔵『源氏物語』
桐壺

きりつほ

（1オ）
いつれの御時にか女御更衣あまたさふらひ給けるなかに
いとやむことなききはにはあらぬかすくれてときめき
給ふありけりもとより我はと思ひあかりたまへる御かた／＼
めさましき物におとしめそねみ給ふおなし程それより
けらうの更衣たちはましてやすからすあさ夕のみや
つかへにつけても人の心をのみうこかしうらみをおふつもり
にやありけんいとあつしくなりゆき物心ほそけにさとかちに
なるをいよ／＼あかすあはれなる物におほして人のそしりをも

（1ウ）
えはゝからせたまはす世のためしにもなりぬへき御もてなし也

準備状況の確認

- チュートリアルキットのインストール
  - tutorialkit_20150310.zip
- 動作の確認
  - 『ひまわり』(ver.1.5)
  - TeraPad
  - MeCab

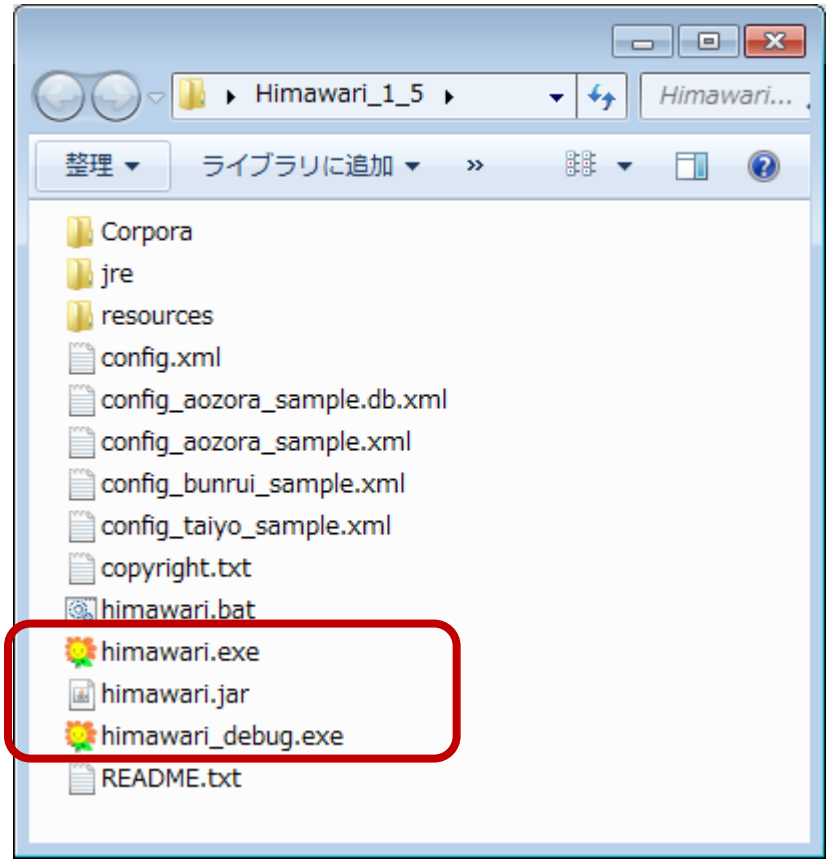
全文検索システム『ひまわり』の簡単な紹介

『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
  - 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

『ひまわり』を起動する



普段使うとき
（Windows 専用）
himawari.exe

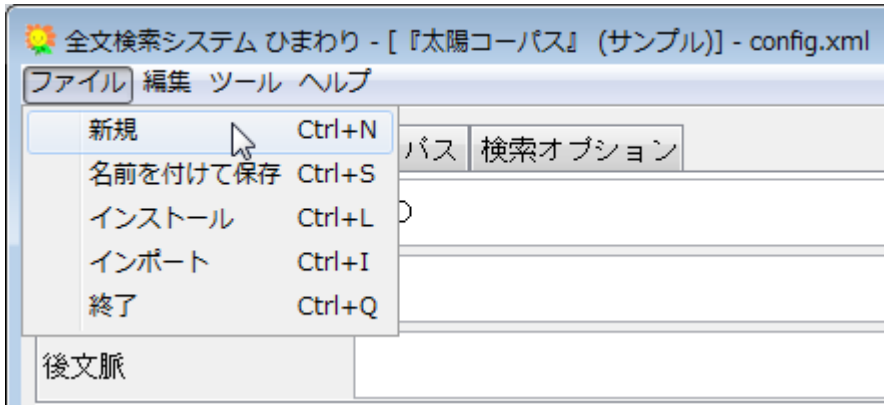
コーパスを作るとき
巨大なデータを検索するとき
（Windows 専用）
himawari_debug.exe

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

検索する


全文検索システム ひまわり - [『太陽コーパス』 (サンプル)] - 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 フィルタ コーパス 検索オプション
本文
あの
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 雑誌名 年 号 題名 著者
1 か涙を浮めた。『 あの 人ッて。』と、勝 太陽 1901 01 櫨紅葉 広津柳
2 が又立戻つて、『 あの う、お花さんもお松 太陽 1901 01 櫨紅葉 広津柳
3 …あのう、何時から あの う、何か上げたいと 太陽 1901 01 櫨紅葉 広津柳
4 んもお松さんもね…… あの う、何時からあのう 太陽 1901 01 櫨紅葉 広津柳
5 『何處でッてね、 あの 何なの、早稻田田甫 太陽 1901 01 櫨紅葉 広津柳
6 ぢや無えんですが、 あの 何なんで……へい、 太陽 1901 01 櫨紅葉 広津柳
7 吉の語を遮り、『 あの 何かい。世間でも何 太陽 1901 01 櫨紅葉 広津柳
8 アにね……。』『 あの 何かい。』と、老 太陽 1901 01 櫨紅葉 広津柳
9 ね。』『えゝ、 あの 、鳥渡……。』と 太陽 1901 01 櫨紅葉 広津柳
10 うで、『えゝ、 あの 、何でげすよ。』 太陽 1901 01 櫨紅葉 広津柳
11 ね。』『いえ、 あの 、何ですよ。今少 太陽 1901 01 櫨紅葉 広津柳
12 事をも知つて居様。 あの お花の饒舌が饒舌たら 太陽 1901 01 櫨紅葉 広津柳
13 『私だつてね、 あの 勝公の顔や様子を見る 太陽 1901 01 櫨紅葉 広津柳
14 …え、お神さん、 あの 何でげさア、詰らね 太陽 1901 01 櫨紅葉 広津柳
15 なの。私は何を、 あの う……今朝ッから頭が 太陽 1901 01 櫨紅葉 広津柳
検索総数:21
検索結果
C:¥Users¥masaya¥Deskto
osname: windows 7
info:
info: _sys _prece
info: _sys _key
info: _sys _follo
info: 雑誌 雑誌名
info: 雑誌 年
info: 雑誌 号
info: 記事 題名
info: 記事 著者
info: 記事 欄名
info: 記事 ジャン
info: 記事 文体
info: 引用 話者
info: 引用 種別
info: 1 位置
history added: あの
doSearch:start
corpusname 『太陽コー
open eix
info(available memory)
time[milisec]: 47
time[milisec]: 47
途中経過の表示
検索総数

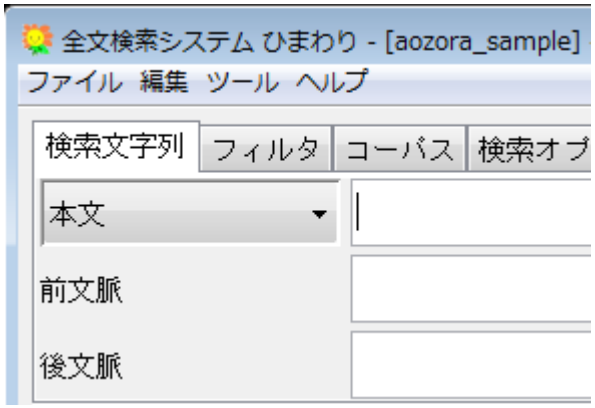
検索対象のコーパスを切り替える

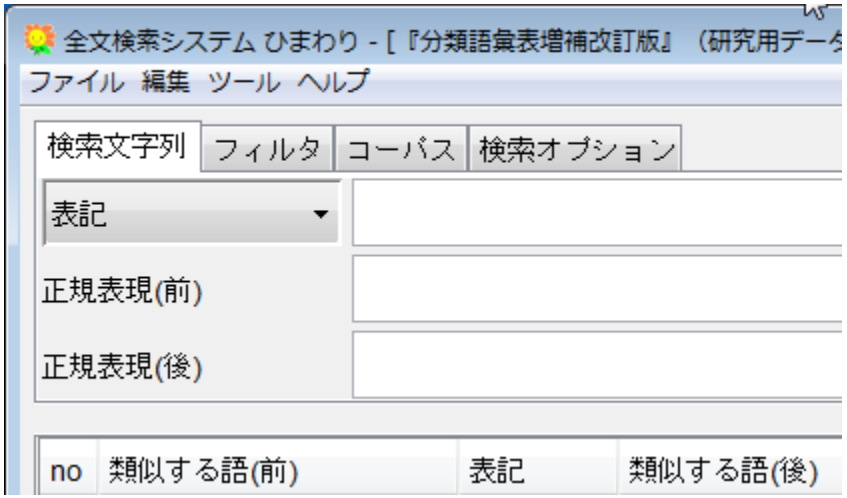


config.xml
（config_taiyo_sample.xml と同じ）

config_bunrui_sample.xml を選択
『分類語彙表』サンプル

config_aozora_sample.xml を選択
『青空文庫』サンプル





検索結果のソート

列名を左クリック

前文脈	キー	後文脈	タイトル	著者
ながら、「おめでとう	ございます	」といったまま席を立	こころ	夏目漱石
なるか、ならないかで	ございます	」と奥さんは気の毒そ	こころ	夏目漱石
うのだから、宜しゅう	ございます	、どうせ英語なんかは	吾輩は猫...	夏目漱石
くらいで、大変御暑う	ございます	。一でも御変りもご	吾輩は猫...	夏目漱石
それは本当のところで	ございます	。もう少し召し上って	吾輩は猫...	夏目漱石
には碟を用いたようで	ございます	。但し生きているうち	吾輩は猫...	夏目漱石
番附を染め出したので	ございます	。妾しには地味過ぎて	吾輩は猫...	夏目漱石
絞殺するという条りで	ございます	。希臘語で本文を朗読	吾輩は猫...	夏目漱石
す習慣であったそうで	ございます	。旧約全書を研究して	吾輩は猫...	夏目漱石
け易えてもよろしゅう	ございます	。金田家の結婚式には	吾輩は猫...	夏目漱石
のあるところに愛嬌が	ございます	。鼻高きが故に貴から	吾輩は猫...	夏目漱石
事でしょう。宜しゅう	ございます	」「それから、あの	吾輩は猫...	夏目漱石
おっしゃった、あれで	ございます	」「あらいやだ。善く	吾輩は猫...	夏目漱石
現象を呈出したもので	ございます	」「伴りのない愚見だ	吾輩は猫...	夏目漱石

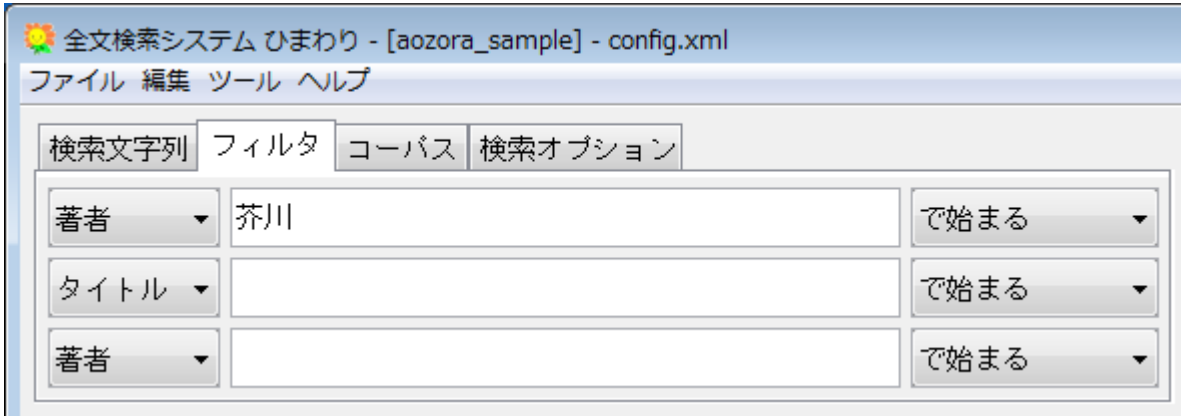
- 昇順
  - 列タイトルをクリック
- 降順
  - シフトキーを押しながら
  - 列タイトルをクリック

- 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

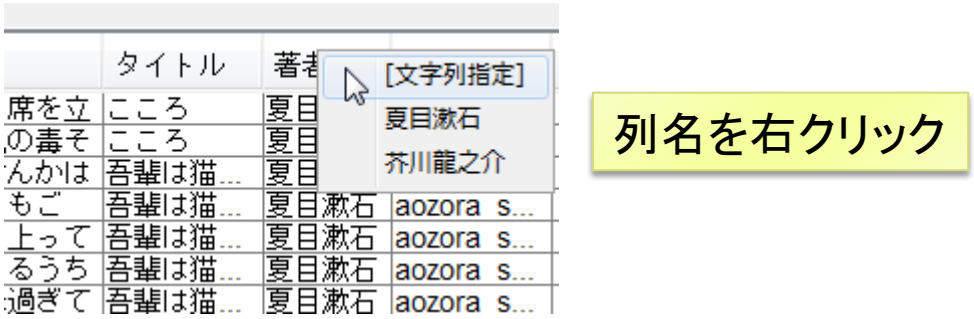
例:「著者」ごとに「後文脈」でソート
→ 「後文脈」「著者」の順

結果の絞り込み

- 検索時に指定



- 検索後に絞り込み



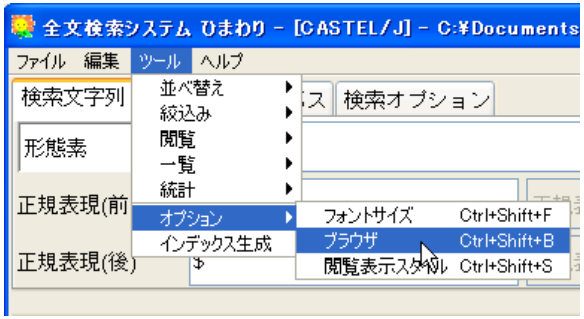
列名を右クリック

ブラウザでの閲覧

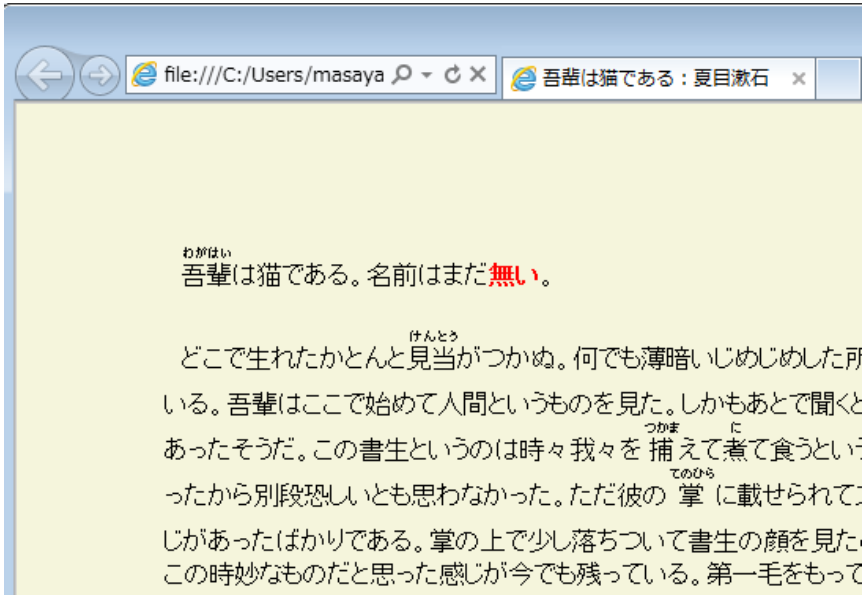
no	前文脈	キー	後文脈	タイトル
1	猫である。名前はまだ	無い	。　どこで生れたか	吾輩は猫...
2	のにそれほどの勇気も	無い	。いよいよ牡蠣の根性	吾輩は猫...
3	な事には記憶が人一倍	無い	。美学原論を著わそう	吾輩は猫...
4	。まだ頂戴するものは	無い	かなと、あたりを見廻	吾輩は猫...
5	る戒名ほど俗なものは	無い	からな」と天然居士は	吾輩は猫...
6	自分くらいえらい者は	無い	つもりでいるんだよ」	吾輩は猫...
7	がただ生きてるんじゃ	無い	です。頭にちょん髷を	吾輩は猫...

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]



既存資料のインポート
（簡単な例）

簡単な例

- 生の（タグなし）テキストファイルをインポートする
- インポートするファイル
  - 配布資料の「簡単サンプル」フォルダ中の2ファイル
  - タグなしテキストを自分で作成する場合は，ファイル名の末尾を「.txt」としてください

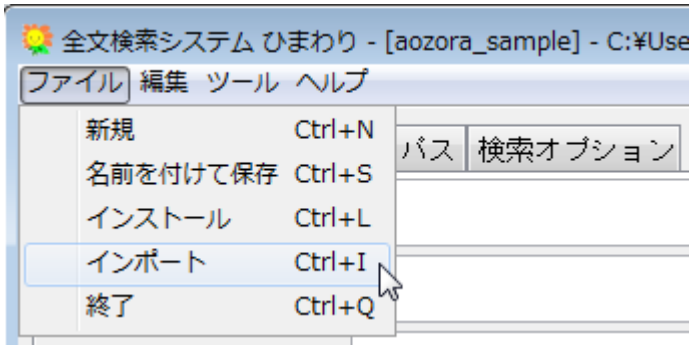
■ テキスト1.txt

```
これはテスト文1Aです。
これはテスト文1Bです。
これはテスト文1Cです。
これはテスト文1Dです。
これはテスト文1Eです。
```

■ テキスト2.txt

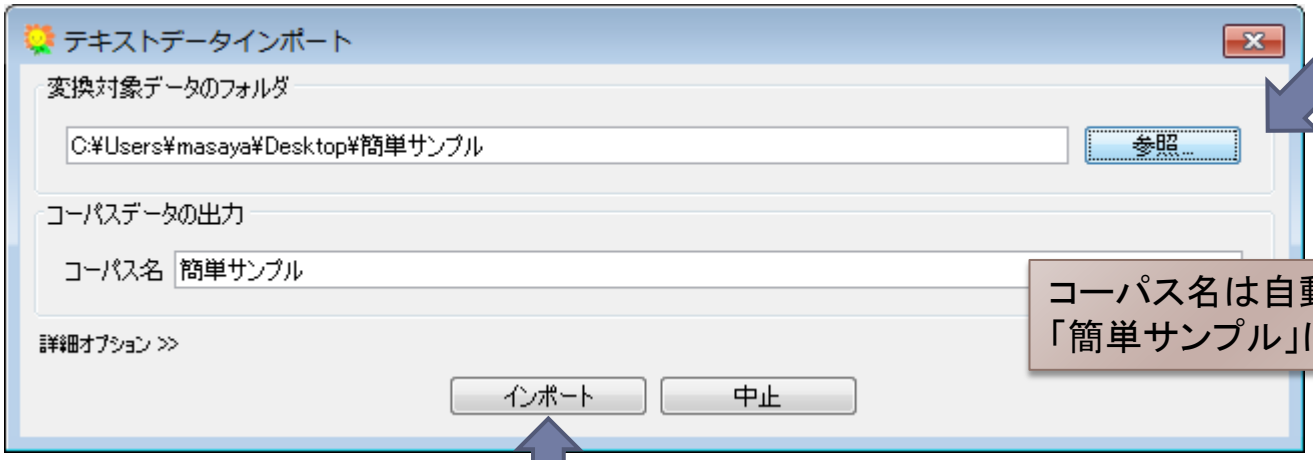
```
これはテスト文2Aです。
これはテスト文2Bです。
これはテスト文2Cです。
これはテスト文2Dです。
これはテスト文2Eです。
```

『ひまわり』用に変換する



「インポート」機能を実行

資料のフォルダを指定する
(「簡単サンプル」)

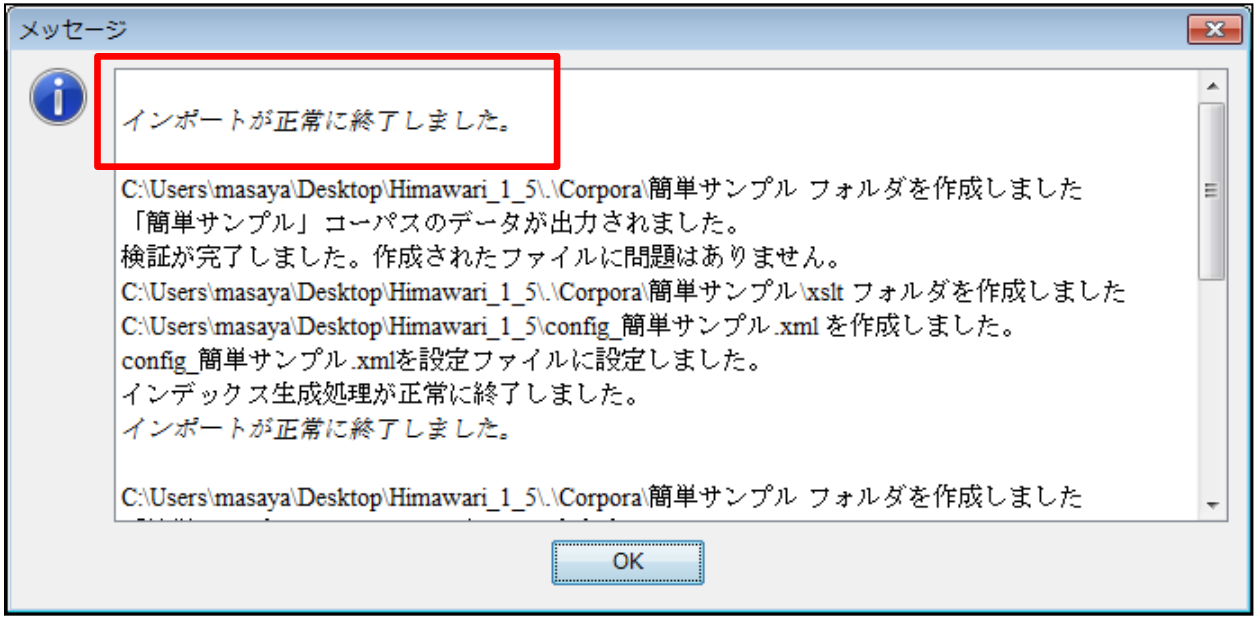


コーパス名は自動的に
「簡単サンプル」になる

「インポート」ボタンを押すと変換が始まる

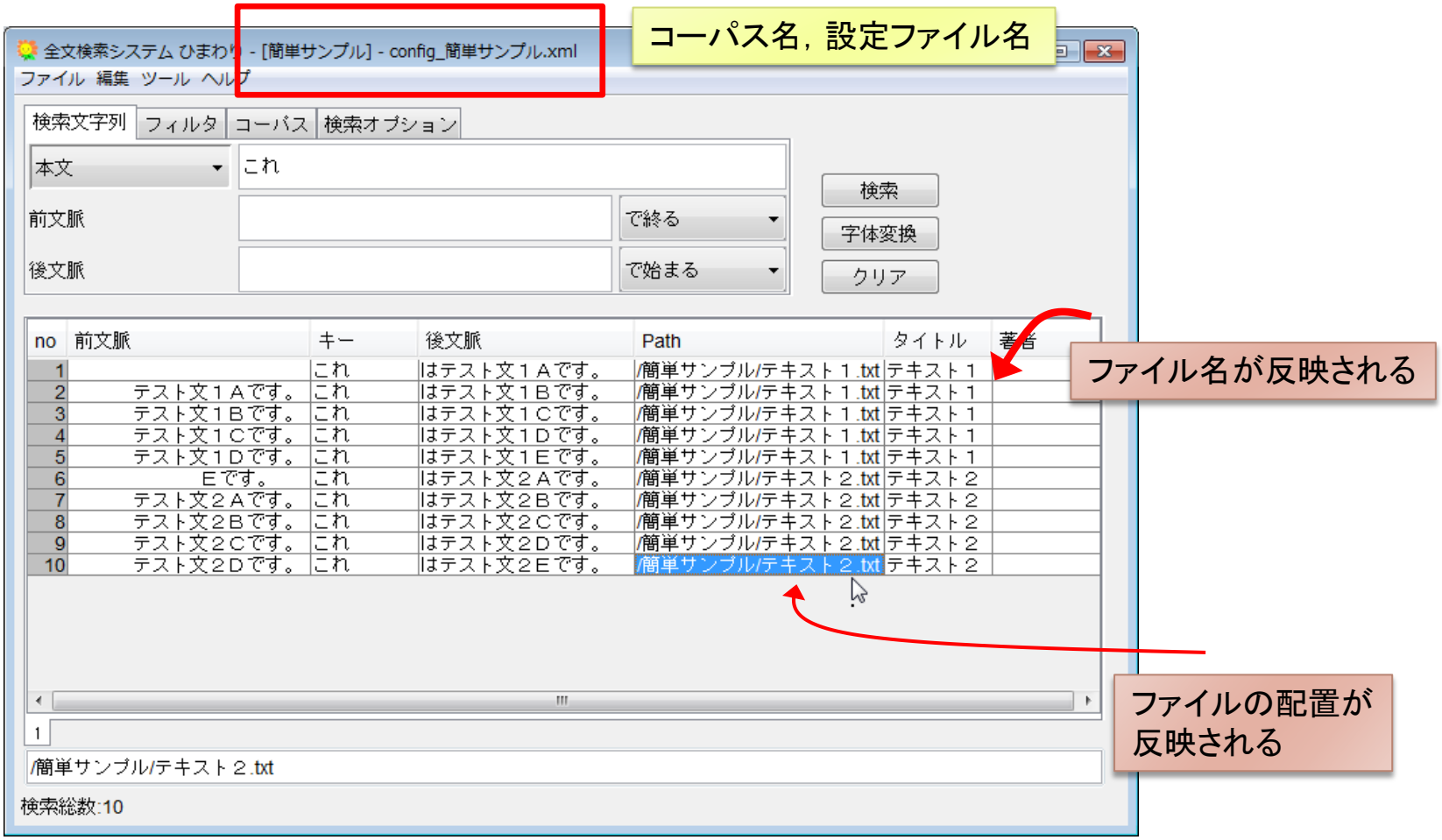
変換結果の確認

- 「インポートが正常に終了しました」となることを確認



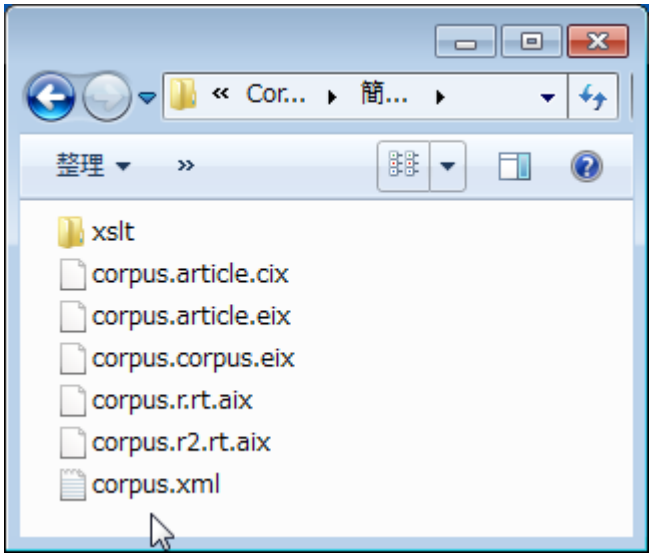
- 設定ファイル「config_簡単サンプル.xml」,
  「 config_簡単サンプル.db.xml 」が生成される
- すぐ使える状態になる（config_簡単サンプル.xml ）

検索してみる

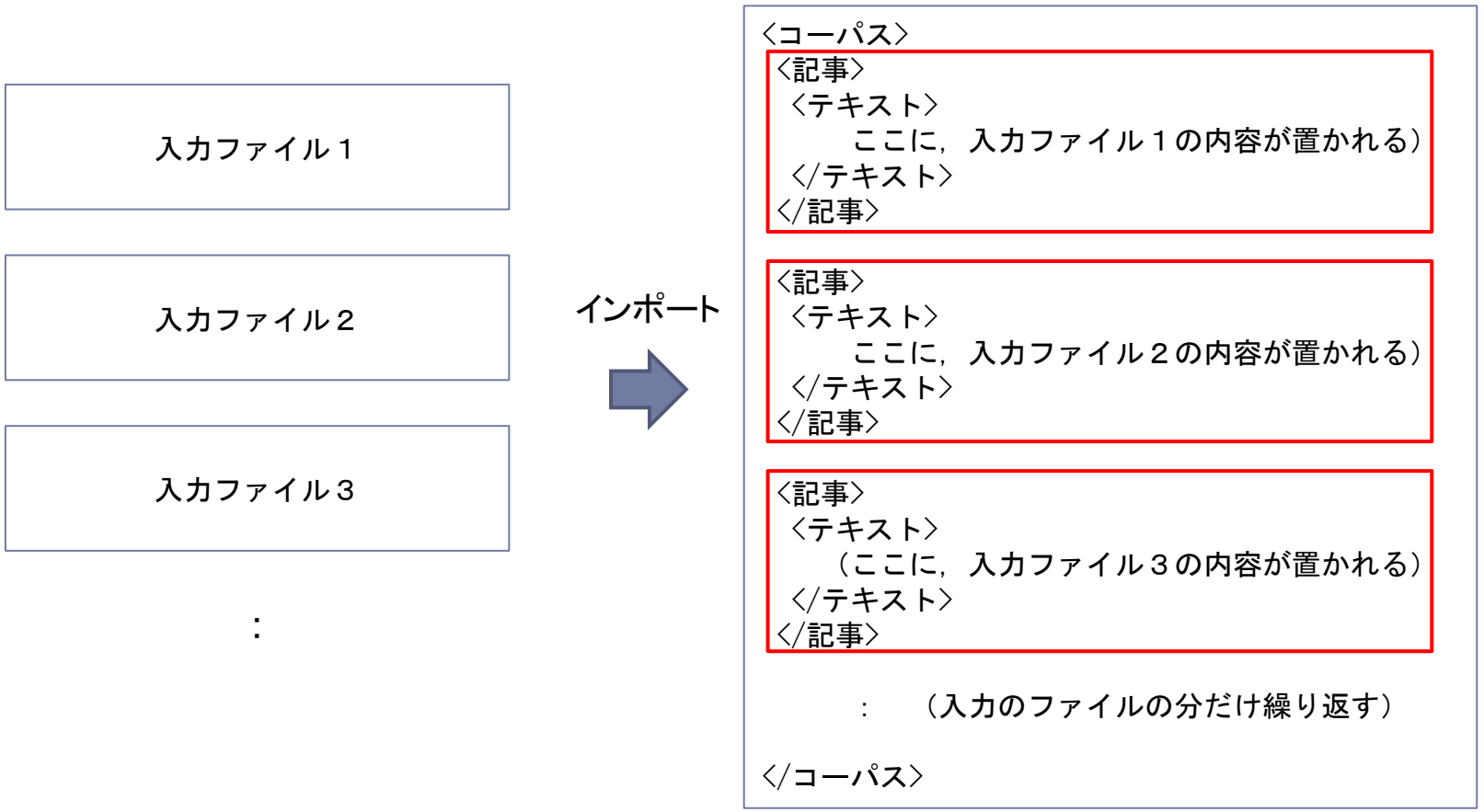

コーパス名，設定ファイル名
全文検索システム ひまわり - [簡単サンプル] - config_簡単サンプル.xml
ファイル 編集 ツール ヘルプ
検索文字列 フィルタ コーパス 検索オプション
本文
これ
前文脈
で終る
後文脈
で始まる
検索
字体変換
クリア
no 前文脈 キー 後文脈 Path タイトル 著者
1 これ はテスト文１Ａです。 /簡単サンプル/テキスト１.txt テキスト１
2 テスト文１Ａです。 これ はテスト文１Ｂです。 /簡単サンプル/テキスト１.txt テキスト１
3 テスト文１Ｂです。 これ はテスト文１Ｃです。 /簡単サンプル/テキスト１.txt テキスト１
4 テスト文１Ｃです。 これ はテスト文１Ｄです。 /簡単サンプル/テキスト１.txt テキスト１
5 テスト文１Ｄです。 これ はテスト文１Ｅです。 /簡単サンプル/テキスト１.txt テキスト１
6 Ｅです。 これ はテスト文２Ａです。 /簡単サンプル/テキスト２.txt テキスト２
7 テスト文２Ａです。 これ はテスト文２Ｂです。 /簡単サンプル/テキスト２.txt テキスト２
8 テスト文２Ｂです。 これ はテスト文２Ｃです。 /簡単サンプル/テキスト２.txt テキスト２
9 テスト文２Ｃです。 これ はテスト文２Ｄです。 /簡単サンプル/テキスト２.txt テキスト２
10 テスト文２Ｄです。 これ はテスト文２Ｅです。 /簡単サンプル/テキスト２.txt テキスト２
1
/簡単サンプル/テキスト２.txt
検索総数:10
ファイル名が反映される
ファイルの配置が反映される

インポート時に生成されるファイル

- 「Himawari_1_5¥Corpora¥簡単サンプル」フォルダ
  - 索引ファイル（検索の高速化）
    - corpus.～.cix
    - corpus.～.eix
    - corpus.～.aix
  - 『ひまわり』形式のXMLファイル
    - corpus.xml
  - 『ひまわり』用の外部データベース
    （形態素解析結果を取り込んだときなどに作成）
    - himawari.h2.db



インポートされたテキストデータの構造（1）



インポートされたテキストデータの構造（2）

- テキスト1.txt

```
これはテスト文１Ａです。
これはテスト文１Ｂです。
これはテスト文１Ｃです。
これはテスト文１Ｄです。
これはテスト文１Ｅです。
```

- テキスト2.txt

```
これはテスト文２Ａです。
これはテスト文２Ｂです。
これはテスト文２Ｃです。
これはテスト文２Ｄです。
これはテスト文２Ｅです。
```

インポート

- corpus.xml（一部，省略）

```
<?xml version="1.0" encoding="UTF-16"?>
<コーパス 名前="簡単サンプル">
<記事 タイトル="テキスト１" 著者=""
        path="/簡単サンプル/テキスト１.txt">
<テキスト>
これはテスト文１Ａです。<br />
これはテスト文１Ｂです。<br />
これはテスト文１Ｃです。<br />
これはテスト文１Ｄです。<br />
これはテスト文１Ｅです。<br />
</テキスト>
</記事>

<記事 タイトル="テキスト２" 著者=""
        path="/簡単サンプル/テキスト２.txt">
<テキスト>
これはテスト文２Ａです。<br />
これはテスト文２Ｂです。<br />
これはテスト文２Ｃです。<br />
これはテスト文２Ｄです。<br />
これはテスト文２Ｅです。<br />
</テキスト>
</記事>
</コーパス>
```

XMLの簡単な説明

タグの基本

- 一定範囲に文書に意味づけ
  - 「開始タグ」: <テキスト> / 「終了タグ」: </テキスト>
  - <記事 ....> / </記事>
- 範囲がない場合
  - 「空タグ」:

「記事」タグは，「タイトル」，「著者」，「path」属性を持つ

開始タグ →

「記事」要素の要素内容

終了タグ →

```
〈記事 タイトル="テキスト１" 著者="" path="/簡単サンプル/テキスト１.txt"〉
〈テキスト〉
これはテスト文１Ａです。〈br /〉
これはテスト文１Ｂです。〈br /〉
これはテスト文１Ｃです。〈br /〉
これはテスト文１Ｄです。〈br /〉
これはテスト文１Ｅです。〈br /〉
〈/テキスト〉
〈/記事〉
```

XMLファイルを作成するときのルール

- 最上位の要素は一つ
- タグの範囲は交差しない
- メタ文字（< > & など）は使わない
- ファイルの先頭でXML宣言を行う（任意）

整形式（well-formedな）XML 文書

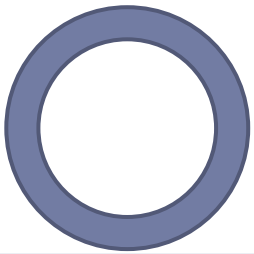
（通常はこれに加えて，文書構造を検証する）

XMLファイルを作成するときのルール

- 最上位の要素は一つ／ファイルの先頭でXML宣言

```
<?xml version="1.0" encoding="UTF-16"?>
<コーパス 名前="簡単サンプル">
<記事 タイトル="テキスト１" 著者=""
        path="/簡単サンプル/テキスト１.txt">
<テキスト>
これはテスト文１Ａです。<br />
これはテスト文１Ｂです。<br />
これはテスト文１Ｃです。<br />
これはテスト文１Ｄです。<br />
これはテスト文１Ｅです。<br />
</テキスト>
</記事>

<記事 タイトル="テキスト２" 著者=""
        path="/簡単サンプル/テキスト２.txt">
<テキスト>
これはテスト文２Ａです。<br />
これはテスト文２Ｂです。<br />
これはテスト文２Ｃです。<br />
これはテスト文２Ｄです。<br />
これはテスト文２Ｅです。<br />
</テキスト>
</記事>
</コーパス>
```

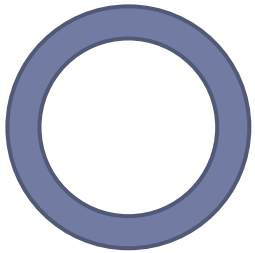


```
<?xml version="1.0" encoding="UTF-16"?>
<コーパス 名前="簡単サンプル１">
        :
        :
</コーパス>

<コーパス 名前="簡単サンプル２">
        :
        :
</コーパス>
```

XMLファイルを作成するときのルール

- タグの範囲は交差しない
  - <著者> <姓>芥川</姓> <名>龍之介</名> </著者>



  - <著者> <姓>芥川</姓> <名>龍之介</著者></名>

交差

XMLファイルを作成するときのルール

- メタ文字（半角）は，そのままでは使えない
- 一般的なXML文書では，次の記号で代替する

  - ＜　⇒　<
  - ＞　⇒　>
  - ＆　⇒　&

既存資料のインポート
（テキスト構造の利用）

生テキストから得られる情報

- 実習用サンプルデータ/青空文庫_txt/芥川龍之介/羅生門.txt

羅生門
芥川龍之介

【テキスト中に現れる記号について】

《》:ルビ
(例)下人《げにん》

｜:ルビの付く文字列の始まりを特定する記号
(例)所々｜丹塗《にぬり》の剥《は》げた

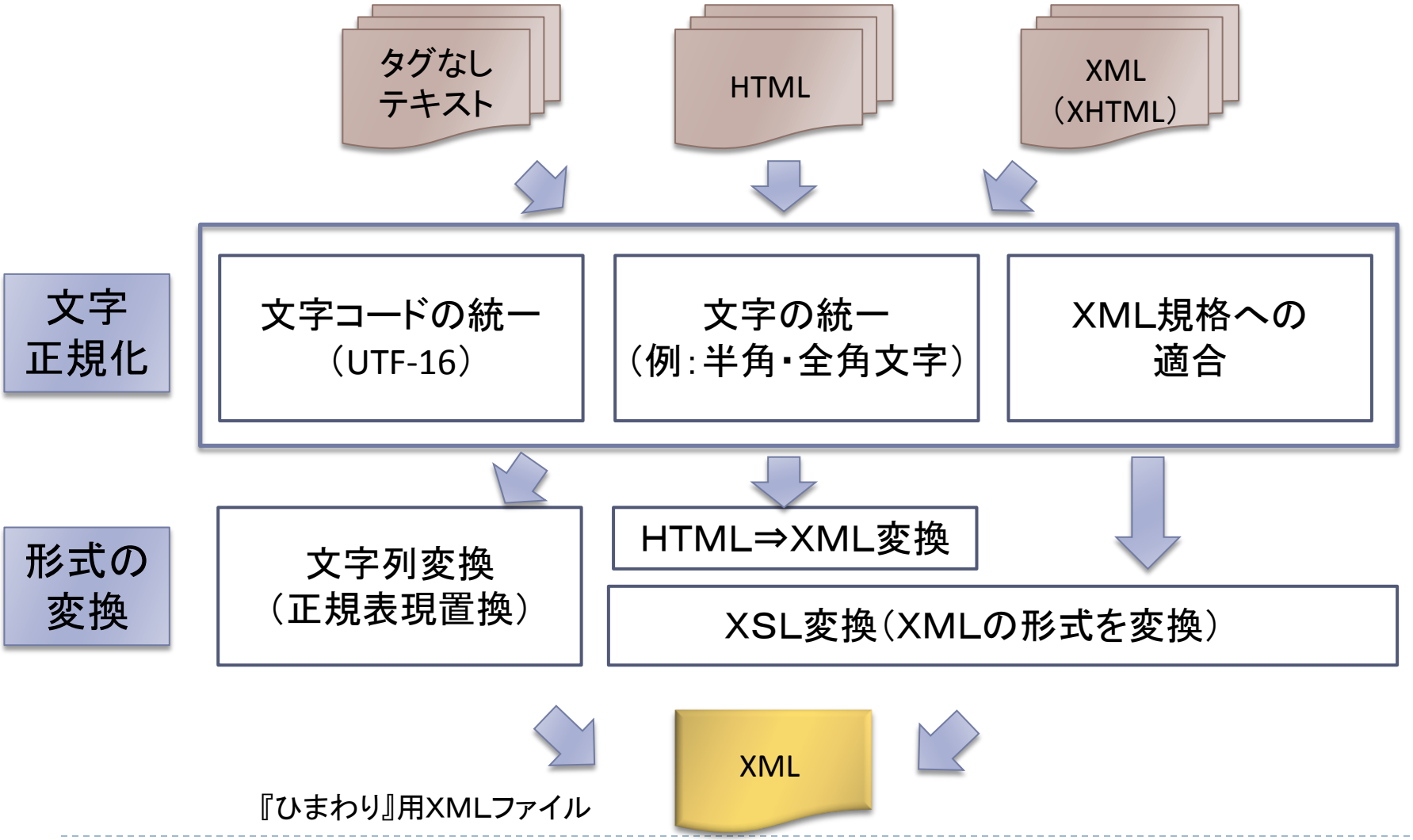
[＃]:入力者注　主に外字の説明や、傍点の位置の指定
　　(数字は、JIS X 0213の面区点番号、または底本のページと行数)
(例)※[＃「てへん＋丑」、第4水準2-12-93]

　ある日の暮方の事である。一人の下人《げにん》が、羅生門《らしょうもん》の下で雨やみを待っていた。
　広い門の下には、この男のほかに誰もいない。ただ、所々｜丹塗《にぬり》の剥《は》げた、大きな円柱《まるばしら》に、蟋蟀《きりぎりす》が一匹とまっている。羅生門が、朱雀大路《すざくおおじ》にある以上は、この男のほかにも、雨やみをする市女笠《いちめがさ》や揉烏帽子《もみえぼし》が、もう二三人はありそうなものである。それが、この男のほかには誰もいない

注記の説明　⇒　---- で囲われている
検索対象からは外したい

ルビの情報　⇒　独自の表記法で記述されている
ただし，このままだと，検索のときに問題となる(例:「下人が」)

インポートの流れ


タグなし
テキスト
HTML
XML
（XHTML）
文字
正規化
文字コードの統一
（UTF-16）
文字の統一
（例：半角・全角文字）
XML規格への
適合
形式の
変換
文字列変換
（正規表現置換）
HTML⇒XML変換
XSL変換（XMLの形式を変換）
XML
『ひまわり』用XMLファイル

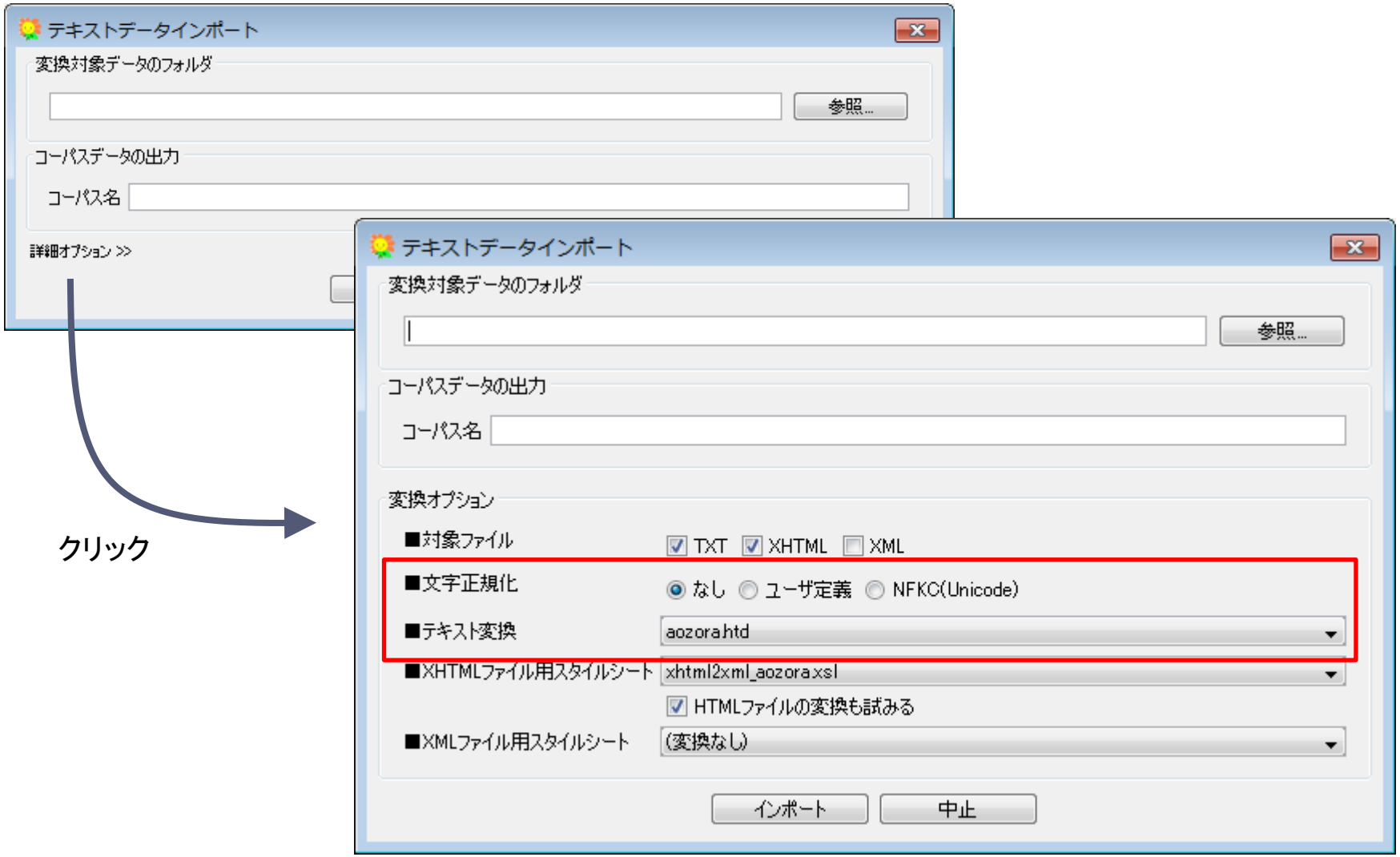
文字の正規化

- 文字コード： UTF-16 に統一（自動）

- 文字の統一
  - ユーザ定義
    - 文字レベルの変換規則を定義できる
    - 設定ファイル（config*.xml）の char_convertion_table 要素
  - NFKC (Normalization Form Compatibility Composition)
    - Unicode で規定されている正規化方法
    - おおまかな規則（参考：Wikipedia，Unicode正規化とは）
      - 半角カナ ⇒ 全角になる
      - 英数字，一部の記号 ⇒ 半角になる
      - ℡⇒ TEL，Ⅳ⇒ IV，②⇒ 2
        （参考：Wikipedia，Unicode正規化とは）

- XMLのメタ文字（<>&）は，全角文字に置換（自動）

インポート時の設定(『ひまわり』)



文字列変換

- 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_5/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
## 例：　※［＃小書き平仮名ん］　⇒　〈注 内容=″＃小書き平仮名ん″ 付与=″″ 種別=″注記″ /〉
※[(＃.+?)]          <注 内容="$1" 付与="" 種別="注記" />


## ルビ(範囲指定あり)
## 例：　所々｜丹塗《にぬり》　⇒　所々〈r rt=″にぬり″〉丹塗〈/r〉
｜(.+?)《(.+?)》        <r rt="$2">$1</r>


## ルビ(範囲指定なし)
## 例：　下人《げにん》　⇒　〈r rt=″げにん″〉下人〈/r〉
(¥p{InCJKUnifiedIdeographs}+?)《(.+?)》              <r rt="$2">$1</r>
```

参考：正規表現の説明

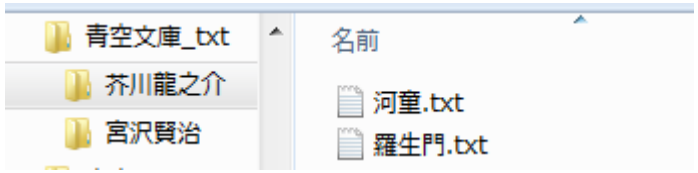
- () は，マッチした文字列を記憶
- 「.」は任意の一文字
- 「+」は，前接する文字の１回以上の繰り返し
- 「?」はマッチングの処理を最短で行う
- $1, $2 は，マッチした文字列を展開する。番号は，マッチした位置を表す
- ¥p{InCJKUnifiedIdeographs} は，１文字の漢字を表す

インポートする資料

- 『青空文庫』から4作品
  - 芥川龍之介： 羅生門，河童
  - 宮沢賢治： 風の又三郎，銀河鉄道の夜

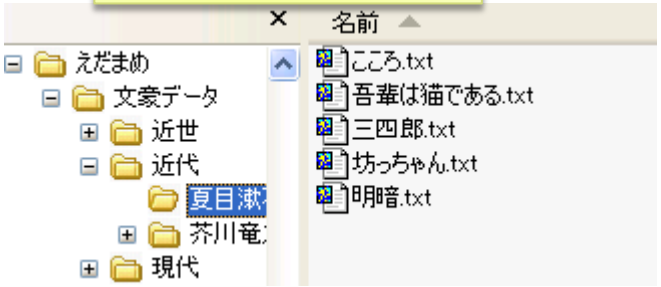
- ファイルの配置

著者情報をフォルダに付与



実習用データのファイル配置

もっと細かくしてもよい



「青空文庫_txt」フォルダをインポートしてみてください

変換規則の例（追加）

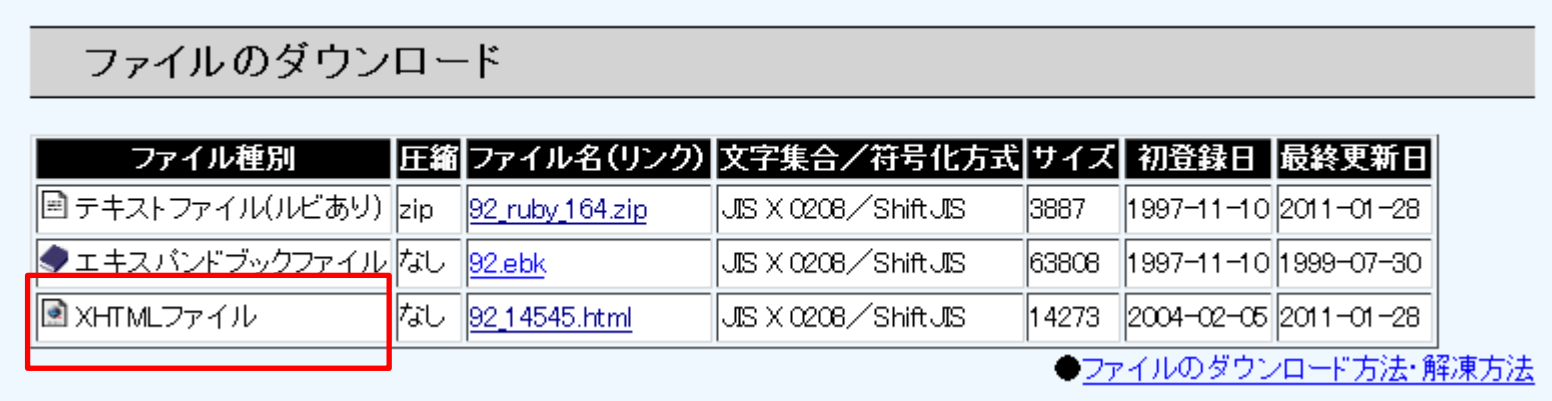
```
## 注記のタグ化
 [(#.+?)]　　<注 内容="$1" 付与="" 種別="注記" />
## 注記凡例の削除
(?s)----+.+?----+.+?¥n
```

表示されていないが，タブ文字があることに注意

- 規則の内容
  - 「※」がない注記にも対応
  - 資料冒頭の注記（--- で囲まれた範囲）の凡例を削除
  - (?s) を指定すると，「.」が改行にもマッチするようになる（正規表現の規則）
- Himawari_1_5/resources/htd/aozora.htd の末尾に追加

参考:『青空文庫』の作品の利用方法

- 今回は，説明の都合上，「テキストファイル」を利用しています
- ただし，通常は，「XHTMLファイル」を使ってください
  - 著者，タイトルの情報は，ファイル内のタグから自動的に抽出されます
  - 凡例や著作権表示などは，検索対象から自動的に除外されます



ファイルのダウンロード

ファイル種別	圧縮	ファイル名(リンク)	文字集合／符号化方式	サイズ	初登録日	最終更新日
テキストファイル(ルビあり)	zip	92_ruby_164.zip	JIS X 0208／ShiftJIS	3887	1997-11-10	2011-01-28
エキスパンドブックファイル	なし	92.ebk	JIS X 0208／ShiftJIS	63808	1997-11-10	1999-07-30
XHTMLファイル	なし	92_14545.html	JIS X 0208／ShiftJIS	14273	2004-02-05	2011-01-28

●ファイルのダウンロード方法・解凍方法

『ひまわり』の検索機能

全文検索機能(要素内容)

- 指定した要素の要素内容を全文検索
  - 例:「テキスト」要素の要素内容(赤い字の部分)

- 照合時にタグは無視される
  - 「下人が」とマッチングする

<記事 タイトル="羅生門" 著者="" path="/青空文庫_txt/芥川龍之介/羅生門.txt">

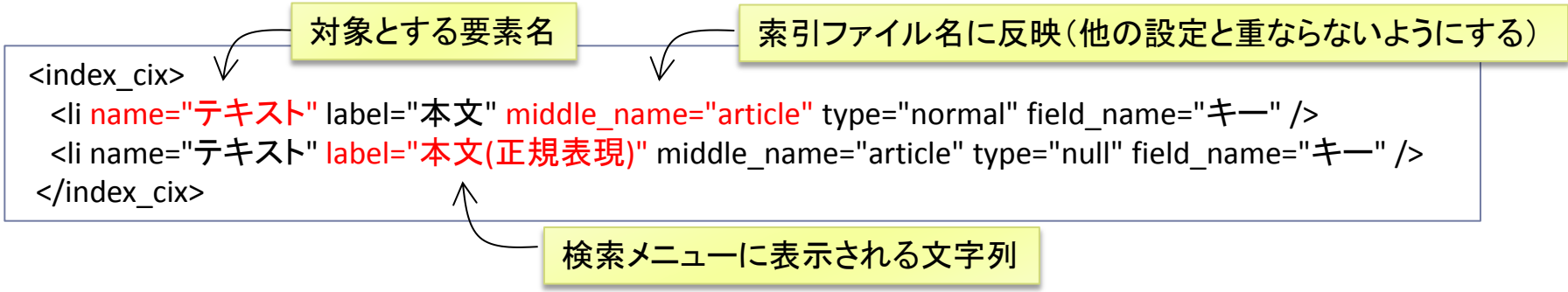
<テキスト>

ある日の暮方の事である。一人の<r rt="げにん">下人</r>が、<r rt="らしょうもん">羅生門</r>の下で雨やみを待っていた。

広い門の下には、この男のほかに誰もいない。ただ、所々<r rt="にぬり">丹塗</r>の<r rt="は">剥</r>げた、大きな<r rt="まるばしら">円柱</r>に、<r rt="きりぎりす">蟋蟀</r>が一匹とまっている。羅生門が、<r rt="すざくおおじ">朱雀大路</r>にある以上は、この男のほかにも、雨やみをする<r rt="いちめがさ">市女笠</r>や<r rt="もみえぼし">揉烏帽子</r>が、もう二三人はありそうなものである。それが、この男のほかには誰もいない。

全文検索対象の設定（config_青空文庫_txt.xml）

- 索引の設定



- 注意
  - 手動で設定ファイルや corpus.xml を書き換えた場合は，[ツール]⇒[インデックス生成]を実行してください

要素属性の取得

- 指定した要素の属性を取得する
  - 例1:「記事」要素の「タイトル」属性
  - 例2:r 要素の rt 属性

〈記事 タイトル="羅生門" 著者="" path="/青空文庫_txt/芥川龍之介/羅生門.txt"〉

〈テキスト〉

ある日の暮方の事である。一人の〈r rt="げにん"〉下人〈/r〉が、〈r rt="らしょうもん"〉羅生門〈/r〉の下で雨やみを待っていた。〈br /〉

no	前文脈	キー	後文脈	Path	タイトル	著者
1	衰微していた。今この	下人が	、永年、使われていた	/青空文庫...	羅生門	
2	方の事である。一人の	下人が	、羅生門の下で雨やみ	/青空文庫...	羅生門	
3	「雨にふりこめられた	下人が	、行き所がなくて、途	/青空文庫...	羅生門	
4	かならない。だから「	下人が	雨やみを待っていた」	/青空文庫...	羅生門	
5	作者はさっき、「	下人が	雨やみを待っていた」	/青空文庫...	羅生門	

取得する要素属性の設定

- 索引の設定

対象とする要素名

```
<index_eix>
 <li name="コーパス" middle_name="corpus" is_empty="false" top="true" />
 <li name="記事" middle_name="article" is_empty="false" isBrowsed="true" />
 <li name="r" middle_name="r" is_empty="false" />
</index_eix>
```

- 検索結果の表示の設定

「記事」要素のpath属性を表示

```
<field_setting>
        :
 <li name="Path" type="argument" element="記事" attribute="path" width="80" />
 <li name="タイトル" type="argument" element="記事" attribute="タイトル" width="80" />
 <li name="著者" type="argument" element="記事" attribute="著者" width="80" />
 <li name="ルビ" type="argument" element="r" attribute="rt" width="80" />
```

検索結果の列名

属性の検索

- 指定された要素属性を全文検索
  - 例の場合は，r 要素の rt 属性
  - 属性の範囲内だけで文字列照合

<記事 タイトル="羅生門" 著者="" path="/青空文庫_txt/芥川龍之介/羅生門.txt">

<テキスト>

　ある日の暮方の事である。一人の<r rt="げにん">下人</r>が、<r rt="らしょうもん">羅生門</r>の下で雨やみを待っていた。

　広い門の下には、この男のほかに誰もいない。ただ、所々<r rt="にぬり">丹塗</r>の<r rt="は">剥</r>げた、大きな<r rt="まるばしら">円柱</r>に、<r rt="きりぎりす">蟋蟀</r>が一匹とまっている。羅生門が、<r rt="すざくおおじ">朱雀大路</r>にある以上は、この男のほかにも、雨やみをする<r rt="いちめがさ">市女笠</r>や<r rt="もみえぼし">揉烏帽子</r>が、もう二三人はありそうなものである。それが、この男のほかには誰もいない。

属性検索対象の設定

- 索引の設定

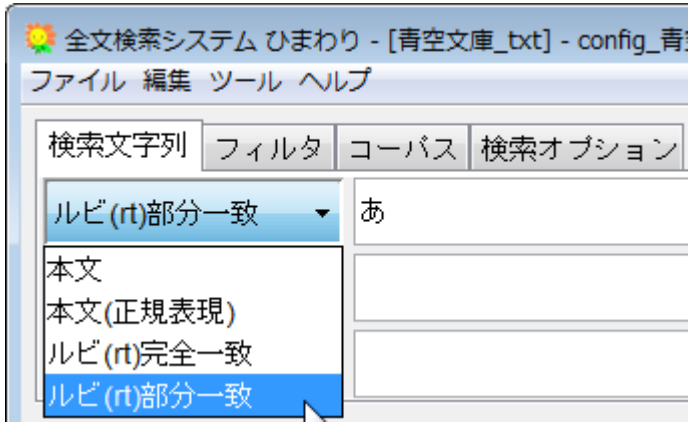
```
<index_aix>
                :
  <li label="ルビ(rt)完全一致" name="r" middle_name="r" argument="rt"
                                                    isCompleteMatch="true" field_name="キー" />
  <li label="ルビ(rt)部分一致" name="r" middle_name="r2" argument="rt"
                                                    isCompleteMatch="false" field_name="キー" />
</index_aix>
```

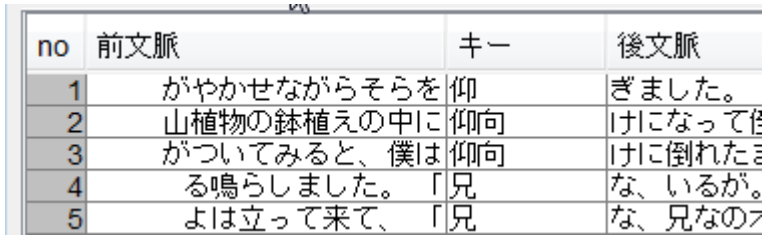
対象とする要素名

対象とする属性名

検索メニューに表示される文字列

true: 完全一致，　false: 部分一致



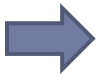


no	前文脈	キー	後文脈
1	がやかせながらそらを	仰	ぎました。
2	山植物の鉢植えの中に	仰向	けになって
3	がついてみると、僕は	仰向	けに倒れた
4	る鳴らしました。「	兄	な、いるが。
5	よは立って来て、「	兄	な、兄なの

インポートした資料の活用

語の区切りと品詞の情報を付与する(1)

- 形態素解析システム
  - MeCab（工藤拓氏）
  - JUMAN （京都大学黒橋・河原研究室）



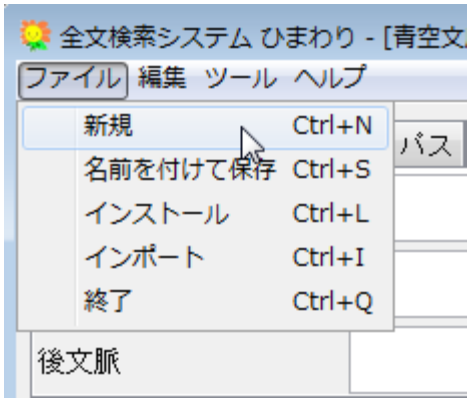
形態素解析システムの解析結果をインポートした資料に付与

- 実行例 （入力文:「文を単語に区切ることができます。」）

```
文　　名詞,一般,*,*,*,*,文,ブン,ブン
を　　助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
単語　名詞,一般,*,*,*,*,単語,タンゴ,タンゴ
に　　助詞,格助詞,一般,*,*,*,に,ニ,ニ
区切る 動詞,自立,*,*,五段・ラ行,基本形,区切る,クギル,クギル
こと　名詞,非自立,一般,*,*,*,こと,コト,コト
が　　助詞,格助詞,一般,*,*,*,が,ガ,ガ
でき　動詞,自立,*,*,一段,連用形,できる,デキ,デキ
ます　助動詞,*,*,*,特殊・マス,基本形,ます,マス,マス
。　　記号,句点,*,*,*,*,。,。,。
EOS
```

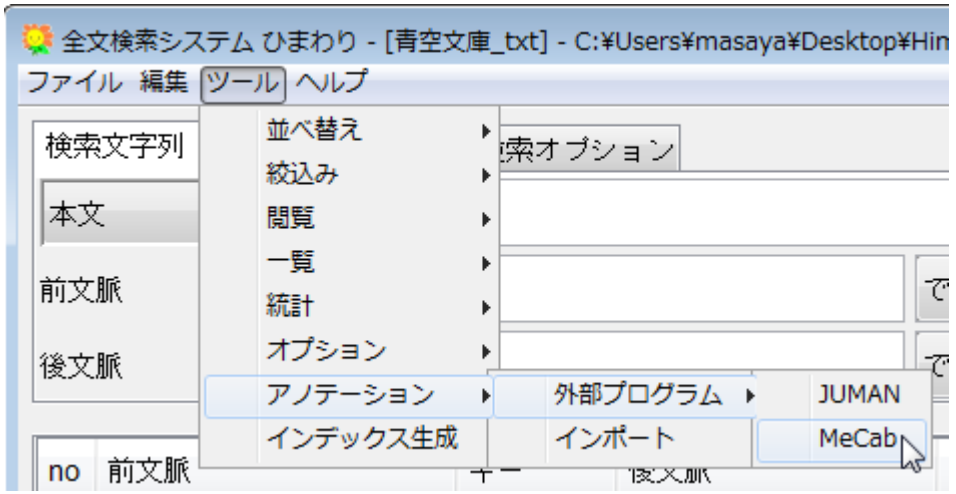
区切り位置，品詞体系は，システム，使っている辞書によって異なります。

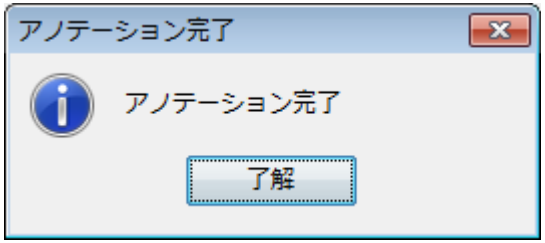
語の区切りと品詞の情報を付与する(2)

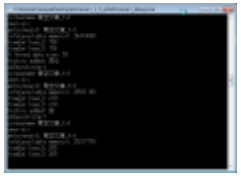


config_青空文庫_txt.db.xml
を読み込む

MeCab の実行



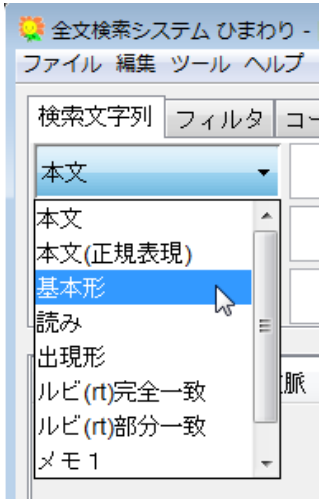




監視しながら，待ちます

形態素解析結果の検索

■ 検索対象の選択



- 基本形： 終止形（活用語の場合）
- 読み： 出現形の読み（カタカナ）
- 出現形：テキスト中での語形（キー欄の値）

■ 検索結果の表示

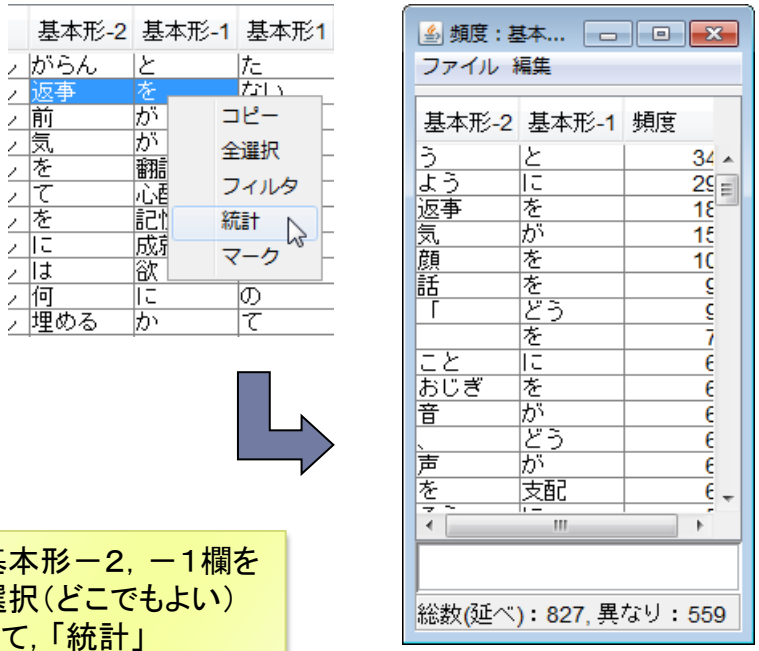
取り込まれた形態素解析結果

基本形－２ … ２語前
基本形２ … ２語後

キー	後文脈	Path	タイトル	著者	基本形	品詞	活用形	基本形-2	基本形-1	基本形1	基本形2
し	てぼんやりそっちを見	/青空…	銀河鉄道…		する	動詞	サ変・スル	気	が	て	ぼんやり
する	と、だいたいこういう	/青空…	河童		する	動詞	サ変・スル	を	翻訳	と	、
し	ていらっしゃるんです	/青空…	銀河鉄道…		する	動詞	サ変・スル	て	心配	て	いらっ…
する	のみ。　問　その詩	/青空…	河童		する	動詞	サ変・スル	を	記憶	のみ	。
し	た時の、安らかな得意	/青空…	羅生門		する	動詞	サ変・スル	に	成就	た	時
せ	ざるあたわず。しかれ	/青空…	河童		する	動詞	サ変・スル	は	欲	ぬ	あたう
する	のですか？」「もち	/青空…	河童		する	動詞	サ変・スル	何	に	の	です
し	ているらしく、立った	/青空…	銀河鉄道…		する	動詞	サ変・スル	埋める	か	て	いる
し	からえでもおら知らな	/青空…	風の又三郎		する	動詞	サ変・スル	。	「	から	え
し	てならんだのです。	/青空…	風の又三郎		する	動詞	サ変・スル	ならえる	を	て	ならぶ

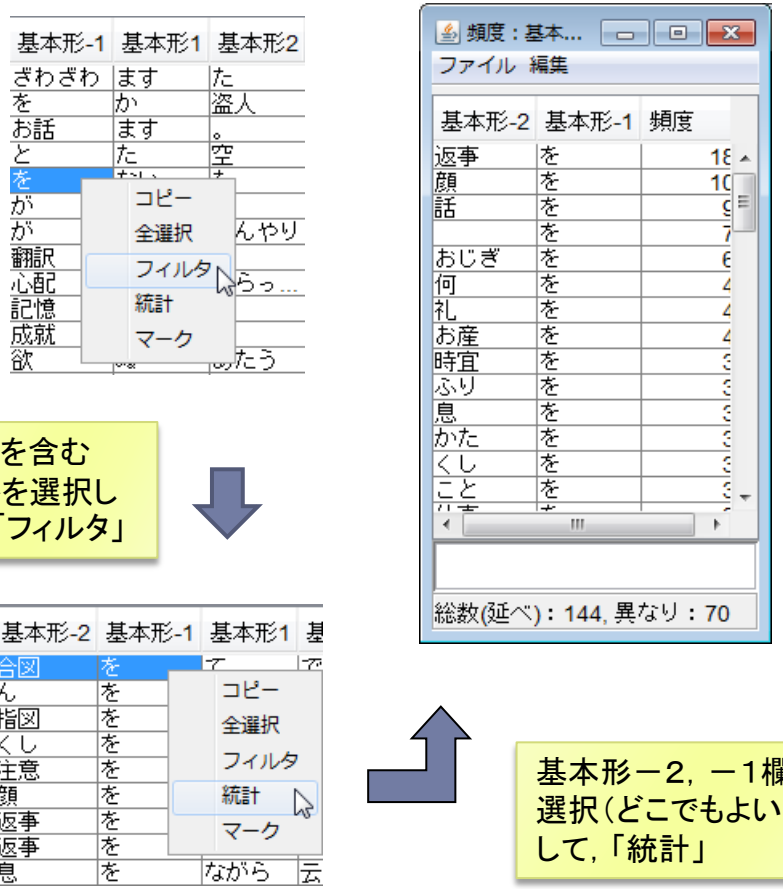
検索結果の集計

- 選択した列での集計
  （「する」の前接形態素の分布）
  - 基本形で「する」を検索



基本形－2，－1欄を選択（どこでもよい）して，「統計」

- フィルタとの連携
  （「する」のヲ格要素の分布）



「を」を含むセルを選択して，「フィルタ」

基本形－2，－1欄を選択（どこでもよい）して，「統計」

参考資料

- 全文検索システム『ひまわり』
  - 設定ファイル作成の手引き
  - 設定リファレンスマニュアル
  - 簡単な検索用データの作成方法2

- 本チュートリアルで使用したソフトウェア・資料
  - MeCab
  - Terapad
  - 青空文庫