全文検索システム『ひまわり』/利用者マニュアル/1　6/７．言語資料をインポートする

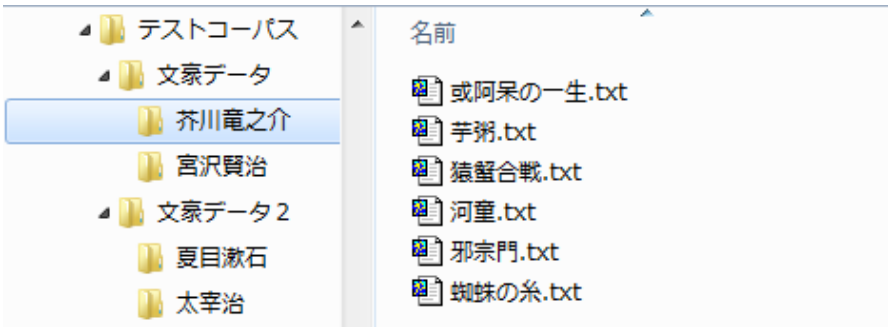
７．言語資料をインポートする †

7.1 一般的な手順 †

『ひまわり』は，テキストファイル，HTML， XHTML，XML などさまざまな形式のテキストをインポートして，検索することができます。以下，順序をおって，一般的な操作手順を説明します。

7.1.1 言語資料の準備 †

まず，インポートする言語資料を一つのフォルダにまとめます。フォルダの中にフォルダを作って，細かく分類しても，かまいません。例えば，次のように，作家ごとにフォルダを作ったり，作家をグループにまとめたりします。後述のとおり，フォルダ構造やファイル名もコーパスに取り込まれ，検索にも利用できます。タグ付けされていない生テキストでは，上図のように，著者名や書名などの書誌情報を記述するのに利用するとよいでしょう。



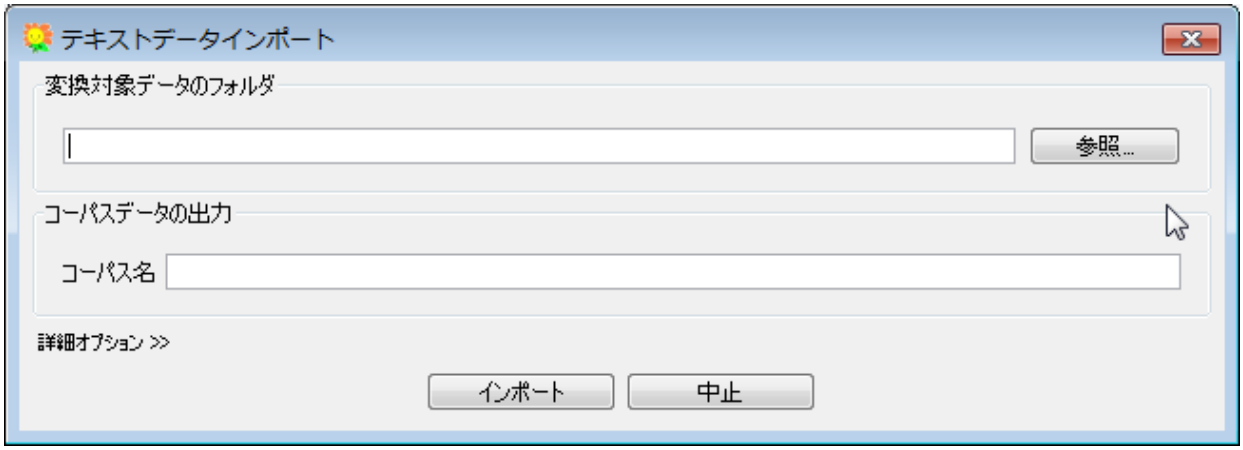
言語資料を集める際は，次のことに留意してください。

- ファイル名の末尾は，ファイル形式に応じて，次のようにつけてください。複数のファイル形式が混在していてもかまいません。
  - .txt ... タグ付けされてない生のテキストファイル
  - .html ... または .htm: HTML，XHTML ファイル
  - .xml ... XML ファイル
- 言語資料の文字コードは，自動判別します。

7.1.2 インポートの実行 †

『ひまわり』のメニューから[ファイル]⇒[テキストインポート]を実行します。なお，Windows を利用している方は，エラーが出ていないか確かめるために，(himawari.exe ではなく) himawari_debug.exe を使うとよいでしょう。

次のウィンドウが現れたら，「参照」ボタンを押して，言語資料をまとめたフォルダを指定します。コーパス名は，指定したフォルダ名となります。例えば，7.1.1 の図の言語資料の場合，「テストコーパス」がコーパス名となります。なお，以上の操作は，指定するフォルダを『ひまわり』にドラッグ＆ドロップすることによっても実行できます。



そのままでよければ，「インポート」ボタンを押してください。デフォルトでは，変換対象フォルダ中のファイルのうち，テキストファイル，および，HTMLファイルが処理対象になります。想定する形式は，『青空文庫』の形式です。

インポート処理が終わると，次のようなウィンドウがでれば，インポート完了です。なお，より詳しい設定を行う場合は，「詳細オプション」を選択してください。詳しくは，7.2 節以降でファイルの種類ごとに説明します。

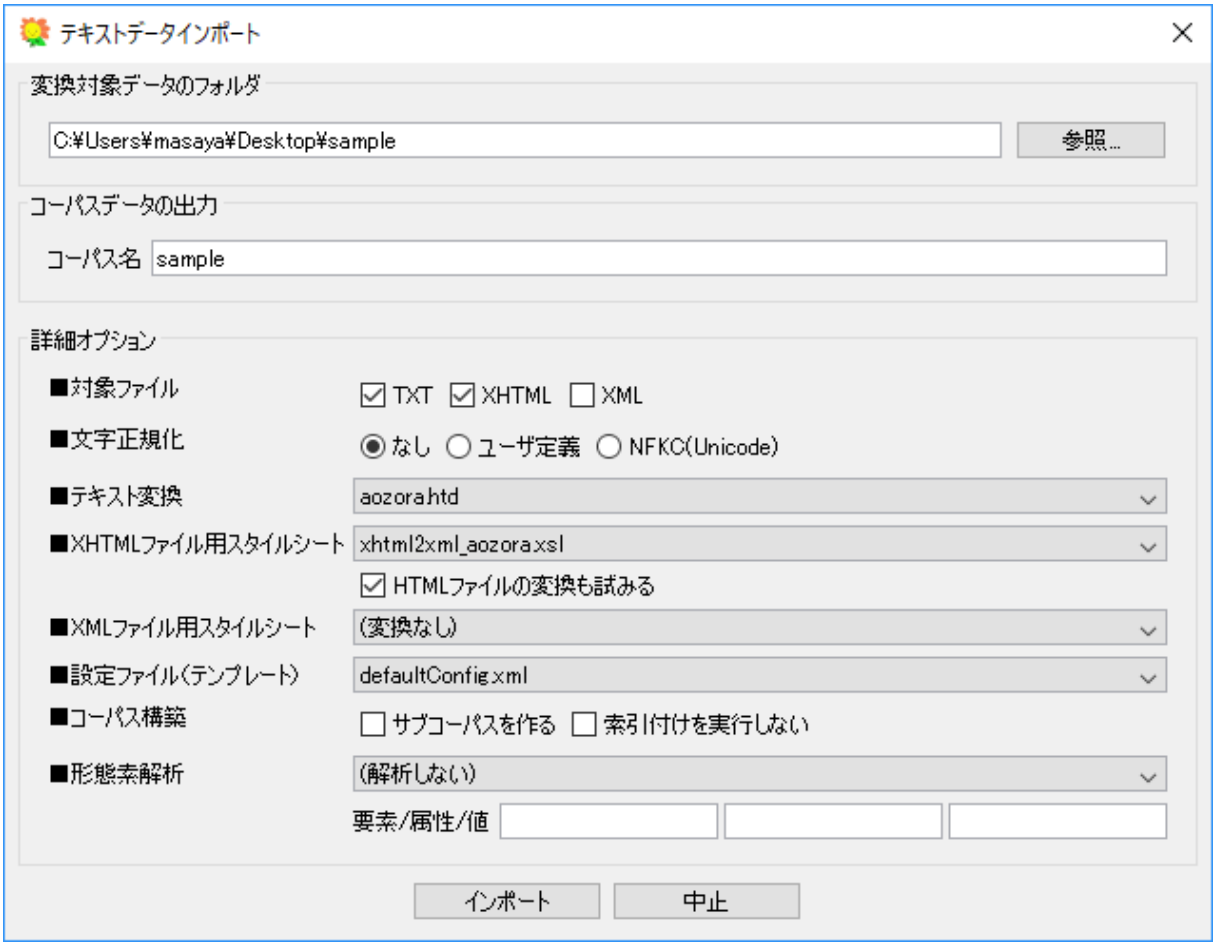
7.1.3 言語資料の利用

まずは，処理途中でエラーが出ていないか確認します。どのようなファイルが取り込まれたかは，[ツール]→編集で確認してください。問題がない場合は，実際に検索してみましょう。

インポート直後から，検索できる状態になります。検索方法と検索結果の見方は，『ひまわり』に同梱している『青空文庫』サンプルの使い方を参照してください（『青空文庫』サンプルは，『ひまわり』のテキストインポート機能を使って作られています）。

7.2 インポート時の詳細オプション

インポート時に詳細な設定は，は，詳細オプション(7.1.2 節参照)で行います。詳細オプションをクリックすると，次のようなウィンドウが現れます。



変換対象ファイル
変換対象のファイルの種類を設定します。

文字正規化
変換時の文字正規化処理の種類を設定します。

- **なし:** 正規化処理は基本的に行いません。ただし，変換後のファイルはXMLなので，XMLのマークアップで使用される文字(<>&の3文字)は強制的にいわゆる全角文字に変換されます。
- **ユーザ定義:** ユーザが定義した変換規則に基づいて，文字を正規化します。変換規則は，『ひまわり』の設定ファイルの import/char_convertion_table 要素で定義します。
- **NFKC (Unicode):** Unicode で定義されている正規化方式 NFKC(Normalization Form Compatibility Composition)に基づいて，正規化する。詳細は，Unicode Standard Annex #15，Wikipediaなどを参照のこと。

テキスト変換
テキストファイル中の文字列を変換するための規則を指定します。２種類の変換規則が用意されています。詳細は，7.3.1節を参照して下さい。

- aozora.htd ... 『青空文庫』テキスト版をインポートするための規則
- diy.htd ... 自作用テキストをインポートするための規則（aozora.htdの規則を包含）

XHTMLファイル用スタイルシート
XHTMLファイルを変換するためのスタイルシートを指定します。また，オプションにより，対象ファイルがHTML ファイルだった場合，XHTML への変換の可否を指定します。

XMLファイル用スタイルシート
XMLファイルを変換するためのスタイルシートを指定します。

設定ファイル（テンプレート）
インポート結果を利用するための設定ファイルの雛形です。

- defaultConfig.xml ... aozora.htd 向けの設定
- diyConfig.xml ... diy.htd 向けの設定

コーパス構築
コーパス構築時のオプションです。

- サブコーパスを作る：インポートするフォルダの直下のフォルダをサブコーパスとして利用します。
- 索引付けを実行しない：インポート時に索引付けをしません。インポート後，手動で，［ツール］⇒［構築］⇒［インデックス作成］を実行して下さい。

形態素解析
形態素解析時のオプションです。

- 形態素解析器を指定すると，インポート時に形態素解析を行います。解析結果を利用するには，コーパス選択時に「外部DB」欄で，「あり(sd)」が選択してください。
- 形態素解析器がインストールされていないと，エラーになります。7.4節を参照して，セットアップを済ませておいて下さい。
- 「要素/属性/値」は，形態素解析対象のXML要素（インポート後のXMLファイル）を指定します。何も指定しなければ，すべてのテキスト要素が形態素解析対象になります。

7.3 インポート時の処理

7.3.1 TXT ファイルのインポート

TXT ファイルをインポートする際の詳細設定について説明します。ここで言う「TXT ファイル」とは，ファイル名の末尾が ".txt" のファイルで，HTML，XML でアノテーションされていないファイルのことです。

インポートするファイルの中に，TXT ファイルが含まれる場合は，変換オプションの「対象ファイル」で，「TXT」を選んで下さい。この項目が選択されていない場合は，指定したフォルダの中に生テキストファイルが含まれていても，インポートされません。

TXT ファイルのインポートに関連するオプションは，「文字正規化」「テキスト変換」オプション(7.2.1参照)です。インポート時は，「テキスト変換」の結果に「文字正規化」の処理が適用されます。

テキスト変換オプションで指定する変換規則は，aozora.htd がデフォルトで同梱されています。このファイルには，『青空文庫』（テキスト版）に含まれる独自形式のアノテーションに対応するための変換規則が記述されています。具体的には，次の三つのアノテーションです（「坊ちゃん」から引用）。

《》：ルビ
（例）坊《ぼ》っちゃん

｜：ルビの付く文字列の始まりを特定する記号
（例）夕方｜折戸《おりど》の ... ルビの範囲が「折戸」までであることを示します

［＃］：入力者注　主に外字の説明や，傍点の位置の指定
（例）おくれんかな［＃「おくれんかな」に傍点］

aozora.htd は，（『ひまわり』フォルダ）/resource/htd/ に配置されています。ファイルの仕様は，設定ファイルリファレンスマニュアルの import / text_transformation_definition 要素を参照してください。

7.3.2 HTML，XHTML ファイルのインポート

HTML，XHTML ファイルをインポートする場合は，変換オプションの「対象ファイル」で，「XHTML」を選んで下さい。このオプションが選択されると，ファイル名の末尾が .html もしくは .htm のファイルがインポート対象となります。

インポート時のオプションには，「文字正規化」「XHTMLファイル用スタイルシート」があります。インポート時は，XHTML用スタイルシートによる変換処理のあと，「文字正規化」の処理が適用されます。

「XHTMLファイル用スタイルシート」は，デフォルトで次の二つのスタイルシートが用意されています。スタイルシートを指定しなければ，そのままインポートします。なお，デフォルトのスタイルシートは，（『ひまわり』フォルダ）/resource/xsl/xhtml/ に配置されています。

xhtml2xml.xsl
XHTML 汎用のスタイルシートです。

xhtml2xml_aozora.xsl
青空文庫専用スタイルシートです。『青空文庫』（XHTML版）のアノテーションをできるだけ取り込みます。ルビ，注記などのほか，タイトルや著者の情報も取り込みます。

HTML ファイルの場合，そのままではスタイルシートは適用できませんが，「HTMLファイルの変換も試みる」オプションをチェックすると，XHTML ファイルへの変換を試みた後に，スタイルシートを適用します。ただし，常に XHTML ファイルに変換できるとは限りません。

7.3.3 XML ファイルのインポート

HTML，XHTML ファイルをインポートする場合は，変換オプションの「対象ファイル」で，「XML」を選んで下さい。このオプションが選択されると，ファイル名の末尾が .xml のファイルがインポート対象となります。

インポート時のオプションとして，XHTML ファイル用のスタイルシートを指定できます。インポート時は，XML用スタイルシートによる変換処理のあと，「文字正規化」の処理が適用されます。

スタイルシートを指定しなければ，そのまま変換せずにインポートします。特定のスタイルシートは同梱されていませんが，（『ひまわり』フォルダ）/resource/xsl/xml フォルダにスタイルシートを入れると，メニューから利用できるようになります。

7.4 形態素解析システムのセットアップ

形態素解析を実行する場合は，PCに事前にセットアップしておく必要があります。対応している形態素解析器は，次のとおりです。

- MeCab (IPADIC)
  - デフォルトのインストールを行ってください。
- MeCab (UniDic)
  - MeCabのインストールをした後，UniDic配布サイトからダウンロードしたファイルを展開し，『ひまわり』フォルダのresources以下の unidic というフォルダに置いて下さい。内部的には，MeCabの -d オプションで使用する辞書を参照しています。
- Juman
  - デフォルトのインストールを行ってください。
- Juman++
  - 今のところ，Linux版のみの対応です。

形態素解析関連の設定は，『ひまわり』フォルダの .himawari_annotator_config.xml で行います。詳細は，リファレンスマニュアルannotator要素を参照して下さい。

Prev
全文検索システム『ひまわり』/利用者マニュアル/1_6/６．アノテーション内容を集計する
Home
Up
Next
全文検索システム『ひまわり』/利用者マニュアル/1_6/８．各種機能