INDEX Converter の使い方

STEP1：テキストデータの作成

まず語彙索引を作りたいテキストを準備します。エディタ（秀丸や Emditor）などで、UTF-8 のテキストファイルを作ります。準備したテキストは自分で単語の切れ目記号やまたがり行記号などを入れていってください。サンプルは簡体字中国語で作成していますが、言語に関係なく利用できます。

【単語の切れ目記号】日本語、中国語などの言語の場合、単語の切れ目を明示的に入れる必要があります。単語の切れ目の印は、「半角スペース」をお使いください。

【またがり行記号】　1 単語が行をまたぐ時には、それを 1 単語と認識させる記号を使います。半角の「=」をお使いください。

以下に、中国語の場合の Sample を示します。

```
我 是 關西大學 的 學生 我 今=
年 二十一 歲 我 學 漢語 專業 我
住在 大阪 我 有 父親 母親 弟弟
我 弟弟 今年 十七 歲 明年 要 考 大=
學 我 父親 今年 五十四 歲 母親 五=
十二歲
```

上の例では、1 行目の最後の「今」と次の行の「年」は行をまたいで 1 単語なので、「今」の後ろに「=」を付けてやるわけです。これで、1 単語として処理されることになります。

また、巻数、ページ数、行数を結果に表示させたいときには、先のテキストに次のように標識を付けます。

<V 1>　…V と数字の間には半角スペースが必要です。この V の後ろは数字ではなくても <V ○○編><V 1 章>のように漢字やひらがなを混ぜてもかまいません。

<P 1>　…ページ数（漢籍の場合は、1 葉の表は 1a。裏の場合は 1b のように示すこともできます。）

<L 1>　…本文の最初の行を 1 行と示します。

※行数は、本文の最初の行が 1 行目であれば、明示する必要はありません。もし、5 行目などから始まる場合には、<L 5> というように明示して下さい。

※なお、ページ数、行数は、ワープロなどのページ数、行数とは無関係であることに注意

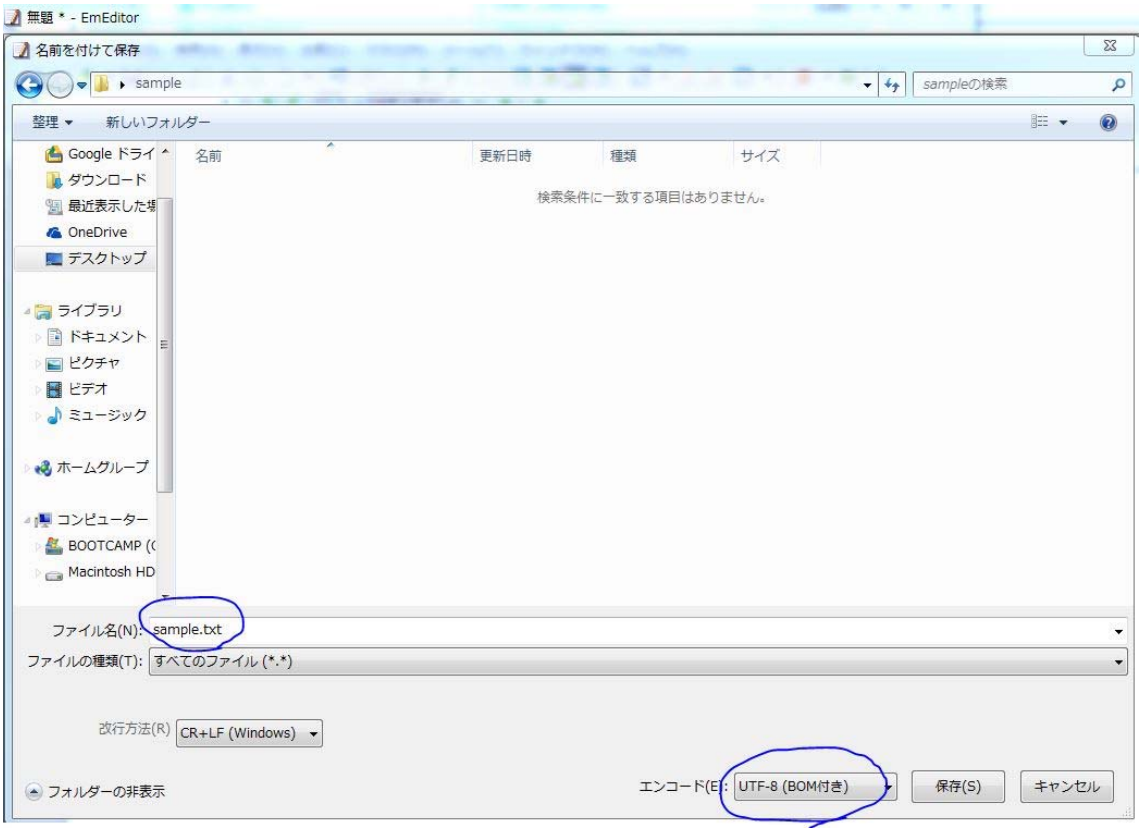
して下さい。

※上記の記号はいずれも省くことができます。

<V 1>
<P 1>
<L 1>
我 是 關西大學 的 學生 我 今=
年 二十一 歲 我 學 漢語 專業 我
住在 大阪 我 有 父親 母親 弟弟
我 弟弟 今年 十七 歲 明年 要 考 大=
學 我 父親 今年 五十四 歲 母親 五=
十二歲

<P 2>
<L 1>
…

完成したらエンコードを UTF-8 に指定して txt 形式で保存してください。



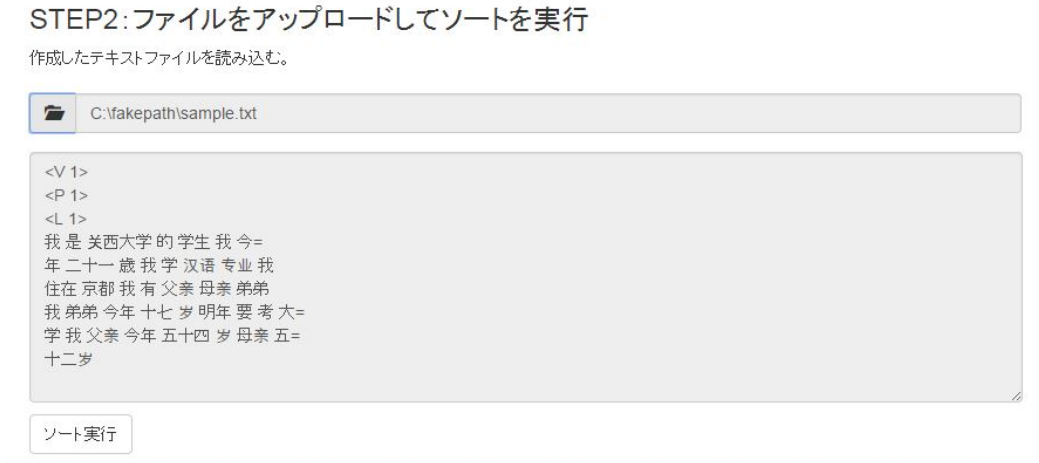
STEP2：ファイルをアップロードしてソートを実行

1.「INDEX CONVERTER」が公開されているページを表示します。

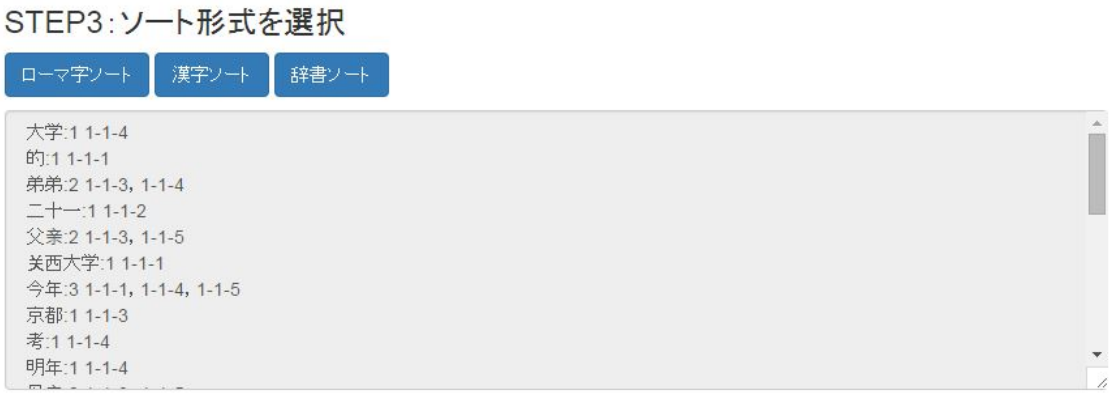
http://www.chlang.org/contents/index-converter/

2.完成したファイルを「フォルダマーク」のアイコンから選択し読み込みます。

3.読み込まれるとファイルの内容が下のボックスに表示されます。



4.「ソートの実行」をクリックする。
ソートを実行すると「STEP3：ソート形式を選択結果」という画面が表示され、ソート結果が表示されます。



※このプログラムは Javascript で構築しているためめ処理速度はお使いの PC の能力に依拠します。そのため環境によっては実行に非常に時間がかかることがあります。（※たまにタイムアウトで実行されない場合もあるようです。そのためブラウザは Firefox を推奨しています。）

STEP3：ソート形式を選択

ソート結果は、「漢字ソート」「ローマ字ソート」「辞書ソート」の 3 種類で出力されます。それぞれのソート結果は青いボタンで切り替えることができます。

ローマ字ソート	漢字ソート	辞書ソート
大学:1 1-1-4 的:1 1-1-1 弟弟:2 1-1-3，1-1-4 二十一:1 1-1-2 父亲:2 1-1-3，1-1-5 关西大学:1 1-1-1 今年:3 1-1-1，1-1-4，1-1-5 京都:1 1-1-3 考:1 1-1-4 …… Total words:25	专业:1 1-1-2 二十一:1 1-1-2 五十二岁:1 1-1-5 五十四:1 1-1-5 京都:1 1-1-3 今年:3 1-1-1，1-1-4，1-1-5 住在:1 1-1-3 关西大学:1 1-1-1 十七:1 1-1-4 …… Total words:25	大学:1 1-1-4 的:1 1-1-1 弟弟:2 1-1-3，1-1-4 二十一:1 1-1-2 父亲:2 1-1-3，1-1-5 关西大学:1 1-1-1 今年:3 1-1-1，1-1-4，1-1-5 京都:1 1-1-3 考:1 1-1-4 …… Total words:25

1.ソート結果の見方

　単語　出現数　巻数（V）- ページ数（P）- 行数（L）　の順番で表示されます

次の例であれば　大阪という単語が 1 回、第 1 巻の 1 ページ 3 行目に出現するという意味で、今年は 3 回、第 1 巻 1 頁の 1 行目と 4 行目と 5 行目に出現するという意味になります。

　大阪:1 1-1-3

　今年:3 1-1-1，1-1-4，1-1-5

Total words は異なり語数（同一の単語が何度用いられていてもこれを一語とし，全体で異なる単語がいくつあるかをかぞえた数。）を示しています。

2.ソートの違い

ローマ字ソート…1 文字目の漢字のピンイン順に表示されます。同じ漢字の場合はさらに 2 文字目…3 文字目…という順番に並びます。ただしプログラム付属の漢字とピンインの対象辞書（25535 文字）にデータが無い漢字を含む場合は各セクションの一番最初に表示されるようになっていますのでご注意ください。

漢字ソート…文字コードの順番に並びます。

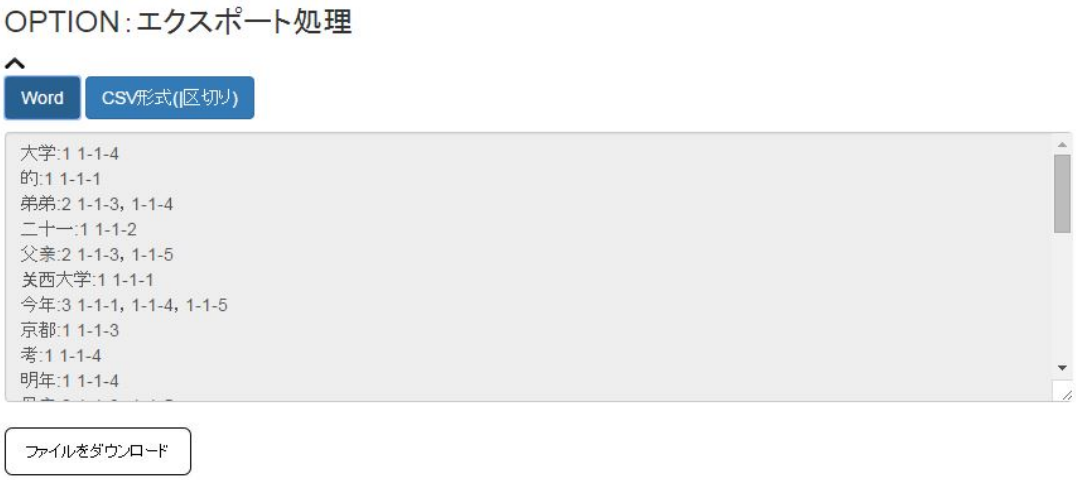
辞書ソート…上記の漢字ソートにピンイン順を組み合わせた並び順になります。1 文字目の漢字が同じものが並び、その漢字の中でピンイン表記順になっています。基本的には一般的な中日辞書と同じ並びと考えていただければイメージしやすいと思います。

3.多読語の処理

基本的にはテキストファイル作成段階で重 chong2　重 zhong4 のように目印を付けることをおすすめしますが、「STEP3：ソート形式を選択」の段階で多読語を選択することもできます。その場合右のほうに選択用ボックスが表示されるので、1 ワードずつ確認していってください。選択した読み方がソートにそのまま反映されます。

OPTION：エクスポート処理

直接出力結果からデータをコピーすることもできますが、Word と CSV 形式の出力が可能となっています。OPTION をクリックするとエクスポート用のボックスが開き、Word と CSV の出力が可能です。形式を選択して「ファイルをダウンロード」ボタンをクリックしてください。なお拡張子が無い状態でダウンロードされるので、word 形式であれば.doc を CSV 形式であれば.csv と付け加えてから開いてください。



Word 形式：Word で開いた際に見やすいように整形しています。
CSV 形式：区切り線は「｜」を利用しています。他のデータベース連携用に作成したため、少し順番や余分なセルが入った状態で出力されます。利用はあまりおすすめしません。

[謝辞]
このプログラムの作成に当たっては内田慶市（関西大学)、弥永信美氏、齋藤希史氏（東京大学）が協同で作成した Mac 専用の索引作成ツールをヒントに、氷野善寛（関西大学）と北田祐平氏（関西大学大学院理工学研究科・院生）とが共同で、ブラウザで利用できるようウェブプログラムとして設計しなおしたものである。使用した感想、バグ・レポートは氷野までお寄せ下さい。2015.03.11