

Ankou CSJ XML/Tamino Client

Version. 1.0

Copyright © 2004 Teru KAMOGASHIRA

Permission is granted to make and distribute verbatim copies of this manual provided the copyright notice and this permission notice are preserved on all copies.

Permission is granted to copy and distribute modified versions of this manual under the conditions for verbatim copying, provided that the entire resulting derived work is distributed under the terms of a permission notice identical to this one.

Permission is granted to copy and distribute translations of this manual into another language, under the above conditions for modified versions, except that this permission notice may be stated in a translation approved by Free Software Foundation, Inc.

Copyright © 2000-2004 The Apache Software Foundation. All rights reserved.

This product includes software developed by the Apache Software Foundation (<http://www.apache.org/>).

目次

1	はじめに	1
2	使い方	2
2.1	検索する対象の指定	3
2.2	対象をリストに追加	3
2.2.1	Tamino に対する検索を指定する時	4
2.2.2	file に対する検索を指定する時	4
2.2.3	複合的な検索を行う場合	6
2.3	検索条件の指定	6
2.4	表示する結果の指定	8
2.5	検索実行	10
2.6	検索条件の保存 / 読み込み	11
3	うまくいかないときは	11
3.1	Ankou 起動時にメッセージが表示される	11

1 はじめに

Ankou は国立国語研究所で作られた日本語話しことばコーパス (CSJ) の XML データベースに対し様々な条件のもとで検索をかけられるツールです。

このツールの動作には以下の環境が必要です。満たしていない場合検索中に Ankou が異常終了したりする可能性があります。

- ハードウェア

プロセッサ Pentium II 450MHz 以上

メモリ 128 Mbyte 以上

標準の 1 XML ファイルに対する検索で最低限必要なメモリです。複数検索を行なったり複雑な検索を行なった場合は更にメモリが必要となります。検索中に頻繁に応答しなくなるようならメモリが足りないかチェックして下さい。

- ソフトウェア

Sun Java 2 Platform, Standard Edition (J2SE) Version 1.4.2 以上 (1.5.0 での動作確認済み) の動作する環境

音声を聞く場合 OS のサポートに加え Java もサポートしている必要があります。

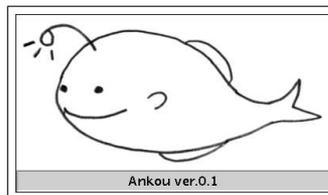
現在存在する問題として、blackdown-java (少なくともバージョン 1.4.1-01) では音声ファイルのシークがうまくいかず結果としてうまく再生されないというものがあります。この場合は、Sun 純正の JavaVM を使用して下さい。

- Tamino に対する検索を使用する場合上記に加えて Tamino XML Server version 4.1.4.1 以上の稼働しているサーバにアクセスできる環境が必要となります。

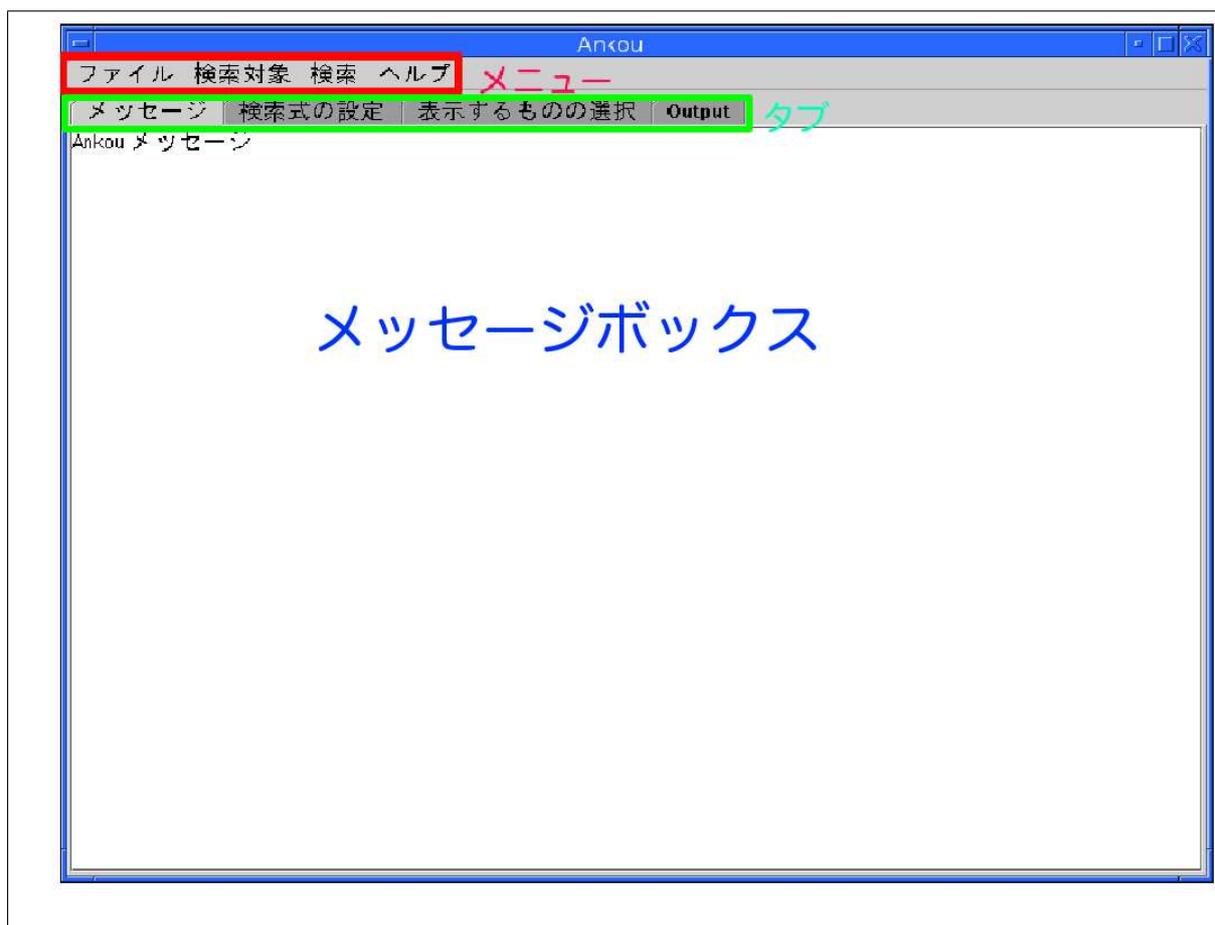
2 使い方

検索する対象の指定、検索条件の指定、表示する結果の指定、検索実行の順に行ないます。
まず Ankou を起動して下さい。

右に示すような Window が表示されます。この間に、プログラムは検索に必要な XML データのツリー構造を読み込みと内部データベースの初期化を行ないますので、若干時間がかかる事がありますが、そのままこの Window が消えるまでお待ち下さい。



すると以下のような Window が現れます。



主に以下の3つを使って検索を行ないます。

- メニュー ファイルに検索結果を保存などといった主な動作をこのプログラムに実行させる時や検索対象を設定する時に使います。
- タブ 検索条件の設定や検索時に表示する対象を設定する時に使います。
- メッセージボックス このプログラムの内部的な情報を表示します。エラーが起こった時は一番下の行に表示された内容からエラーの内容を判断できることがあります。

では実際に検索をしてみましょう。

2.1 検索する対象の指定

まずメニューの「検索対象」から「検索対象の設定」を選んで下さい。すると右のような Window が開きます。

一番上のボックスはリストボックスとって、検索対象となる XML ファイルまたは Tamino XML Server を追加していく所です。このリストボックスでは対象を Click して選択するというを行ないません。

検索対象の種類のリ側のプルダウンメニューは、上のリストボックスで選択したものが XML ファイル対象なのか Tamino XML Server なのかを選択 / 表示するものです。XML ファイルが検索対象の場合 “doc” が、Tamino XML Server の場合は “Tamino” がそれぞれ選択されています。

その下のファイル名 / Tamino の URI のテキストボックスは、ファイル名または Tamino データベース URI を指定します。ファイル名を入力する場合はテキストボックスのすぐ右下のファイル名を参照でマウスにより選択できますが Tamino データベース URI はキーボードから入力する必要があります。Tamino データベース URI については後述します。

先ほど述べたリストボックスに検索対象を追加したりリストボックスから対象を削除する場合は、一番下の段の検索対象を追加 / 削除ボタンを使います。

設定し終わったら閉じるボタンで Window を閉じます。

では、実際に検索対象を設定してみましょう。



2.2 対象をリストに追加

まず一番下の段の検索対象を追加を押して下さい。すると以下のように上のリストに一つ doc(“”) が追加されます。



ここで検索する対象が Tamino かファイルかによって方法が違います。

2.2.1 Tamino に対する検索を指定する時

次に Tamino XML Server に対して検索を実行してみましょう。この場合、Tamino には複数の文書が IPU をルートとして格納されていることに注意して下さい。

まず、Tamino の動いているサーバの IP アドレスまたは名前、Tamino/Apache の動いているポート、Tamino でのデータベース名、コレクション名を取得して下さい。

例：

- サーバの IP アドレス 192.168.1.60
- Tamino/Apache の動いているポート 80
- Tamino でのデータベース名 base1219
- コレクション名 IPUCollection

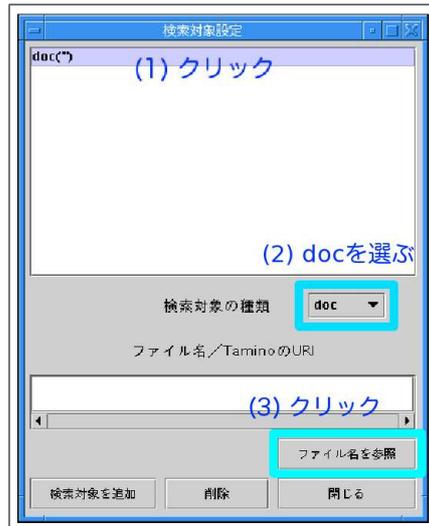
そうしたら、さきほど追加したリストをクリックし、下の検索対象の種類で「Tamino」を選んで下さい。すると、表示が「☐」になります。そうしたら、「ファイル名 / Tamino の URI」で、
http://サーバの IP アドレス:80/tamino/データベース名?コレクション名
を入力して下さい。上で挙げた例でいえば、
http://192.168.1.60:80/tamino/base1219?IPUCollection
の入力します。



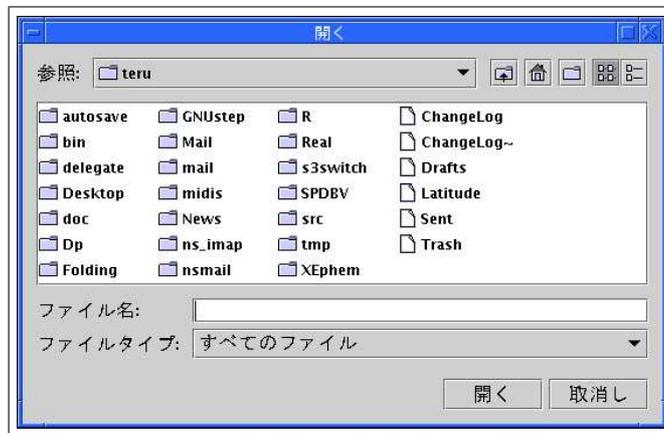
これで Tamino を検索対象として設定できました。閉じるボタンを押して Window を閉じて下さい。

2.2.2 file に対する検索を指定する時

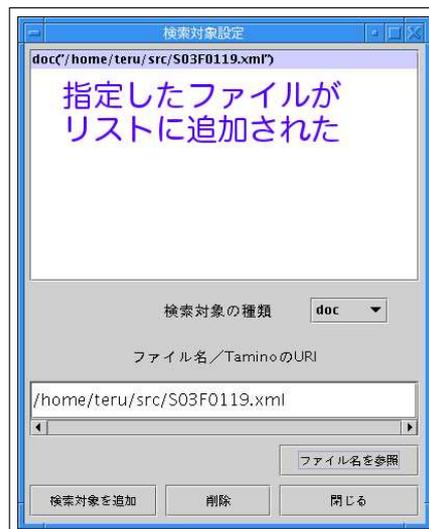
リストのなかの、先に追加した物をクリックして下さい。そして検索対象の種類で doc が選ばれているのを確認した上でファイル名を参照をクリックして下さい



すると、以下のような Window が開きます。



ここで検索対象となるファイルを指定して下さい。していしたら「開く」ボタンを押します。すると、さっきのリストでは指定したものに置き換わっているのがわかると思います。



これで XML ファイルを検索対象として設定できました。閉じるボタンを押して Window を閉じて下さい。

2.2.3 複合的な検索を行う場合

以上で述べた方法を繰り返してリストに複数の対象を登録することで複数のファイルを対象とした検索を行うことができます。ただし、サーバ/ファイル両方同時には検索できません。また、サーバは複数してしても検索対象となるのははじめのものです。しかし、対象ファイルが増えるほど検索にかかる時間や必要なメモリが多くなり検索が失敗する可能性もあるので、そのことを念頭においた上で利用して下さい。

2.3 検索条件の指定

では今度は検索条件を指定してみましょう。もし XML に関する知識が全くない場合は後のコーパス附属のマニュアルをお読み下さい。

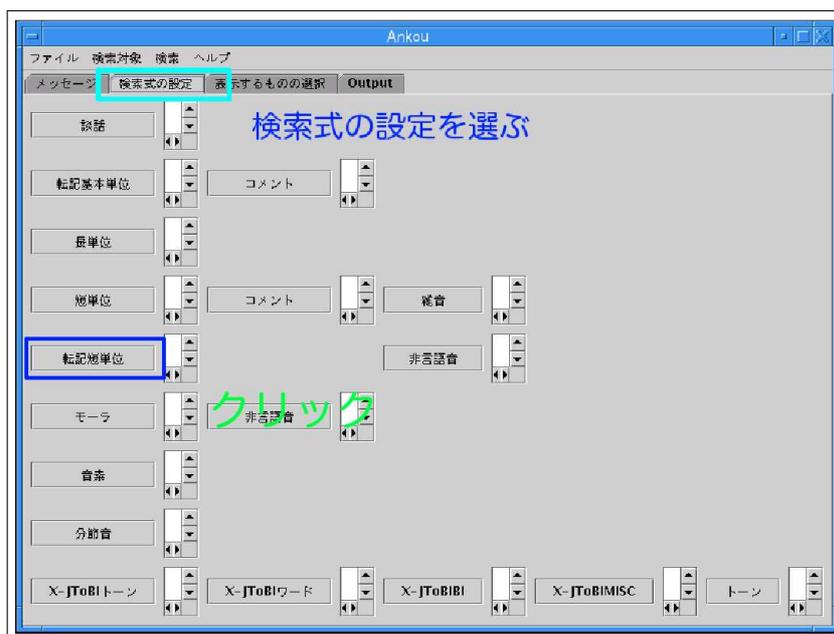
まず、言い誤り時の促音化「ッ」の検索を先ほど指定したファイルに対してかけてみましょう。ここで、コーパスに附属するマニュアルのなかの「XML 文書について」というマニュアルを見て下さい。

ここで探したいのは、まずモーラが「ッ」であるものですから、モーラ記号 (MoraEntity) が「ッ」であるものを探します。さらに、言い誤りをしている必要がありますので、転記短単位 (TransSUW) において発音エラー (TagIncorrectNorm) が指定されている (つまり言い誤りをしたもの) 必要もあります。

以上をまとめると、TransSUW の属性 TagIncorrectNorm が空文字列ではなく、かつ、Mora の属性 MoraEntity が「ッ」であるものを検索すれば良い事になります。では実際に指定してみましょう

まず転記短単位 (TransSUW) を設定します。

はじめにタブで検索式の設定を選んで下さい。

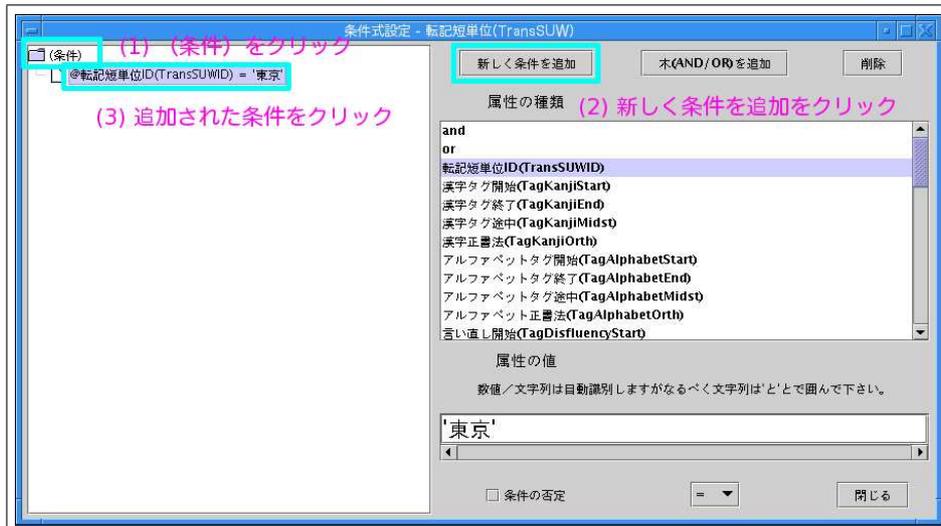


次に転記短単位と書いてあるボタンを押して下さい。すると、以下のような Window が現れます。



条件一覧には、いま Click したタグについて設定した条件のすべてが表示されます。属性一覧には、今 Click したタグについて設定できる属性一覧がリスト表示されます。テキスト入力では、「ッ」といった条件のテキストを入力します。条件設定では、～でないなどといった否定を指定したり、～以上といった条件を指定できます。

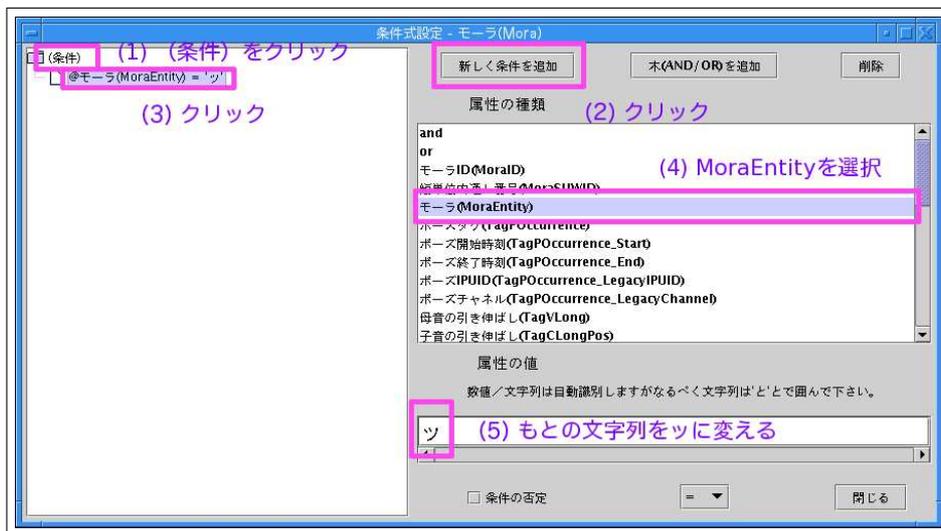
では、転記短単位の属性 TagIncorrectNorm が空文字列ではないと指定してみましょう。まず、条件一覧のなかの「(条件)」をクリックします。次に「新しく条件を追加」をクリックして左の条件一覧に条件を追加します。そして、条件一覧に新しく追加された条件をクリックします。



そうしたら、今度は属性一覧から言い間違い正式発音 (TagincorrectNorm) を選んで Click し、テキストエリアで文字をすべて消去し、条件設定の右側のプルダウンメニューから「!=」を選びます。これで転記短単位の設定はできました。もう一度左の条件一覧を見て下さい。@言い間違い正式発音 (TagIncorrectNorm) != ” と入力されています。これが条件文です。設定が終わったので閉じるボタンを押してこの Window を閉じて下さい。



同様にして Mora の属性 MoraEntity を「ッ」である条件を追加しましょう。まずメインの Window でモーラを Click します。そして出た Window で先ほどと同じように条件を設定します。以下のようにできましたか。



できたらまたこの Window は閉じて下さい。

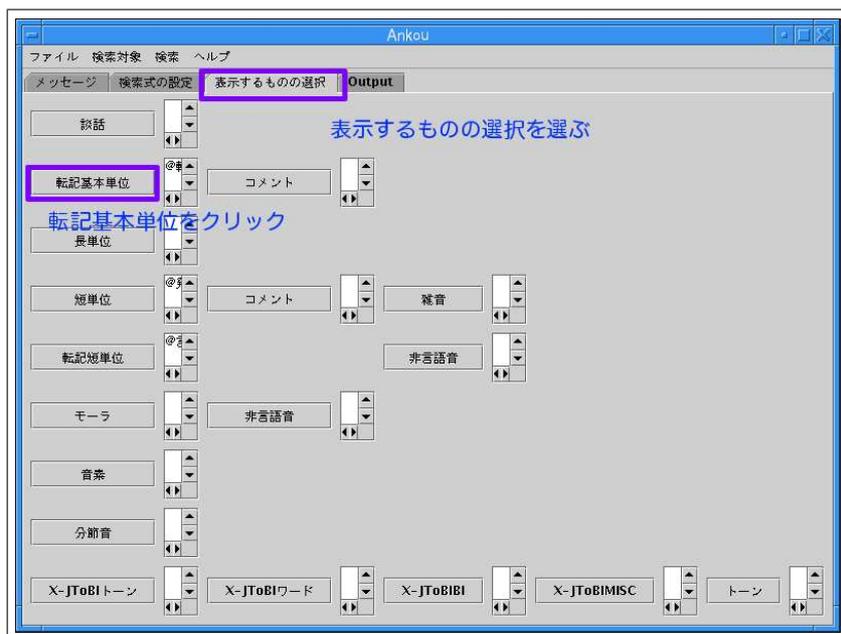
2.4 表示する結果の指定

さて、半分まで来ました。今度は表示するものを設定します。いくら検索してもこのままではなにも表示されません。

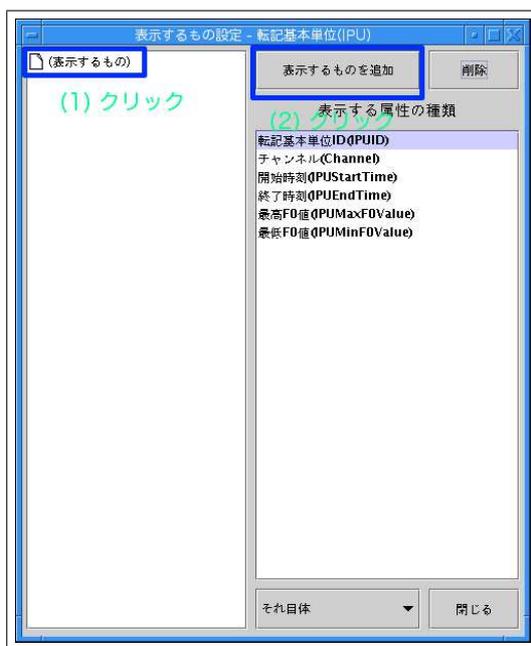
ここでは、転記基本単位 ID、発音形、発音エラー正式発音、を表示することにします。

まず転記基本単位 ID を設定しましょう。

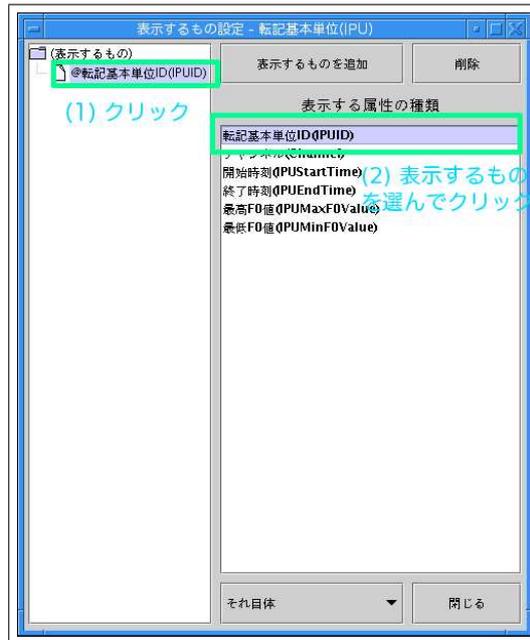
先ほどと同じように今度はタブで表示するものの選択を選びます。そして、転記基本単位を Click します。



すると今度は以下のような Window が表示されます。



ここで、まず左の (表示するもの) を Click します。次に表示するものを追加を Click します。すると、左のツリーに新しく追加されるのでそれを Click してから表示する属性を右のリストから選びます。これで設定できました。



発音形、発音エラー正式発音についても同様に設定できます。表示したい属性は、先ほどの XML マニュアルを参照して探してください。

発音形は短単位に、発音エラー正式発音は転記短単位にそれぞれ含まれます。

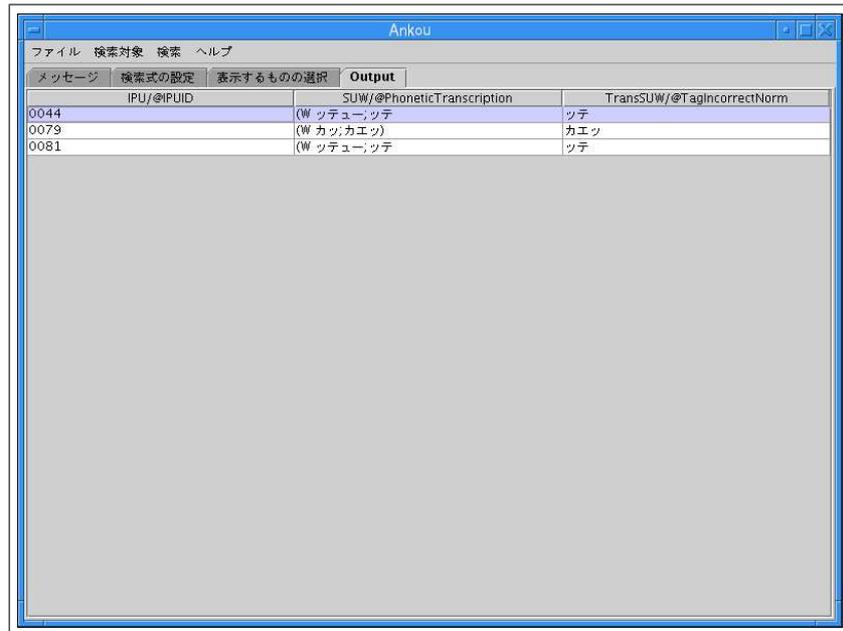
2.5 検索実行

あとは検索を実行するだけです。メニューから検索を実行を選んで下さい。



のような Window が表示され、検索が始まります。

何十秒間かたつとこの Window が消えます。そうしたらタブの Output を選んで下さい。うまくいってればテーブルが表示されています。



これが検索していたものです。ファイルに保存する場合はメニューのファイルからファイルに保存 結果を CSV として保存を選んで保存できます。勝手に上書きするので注意して下さい。

テーブルをダブルクリックすると該当する箇所の転記基本単位の音声部分が再生されます。ファイルがなかったり OS が対応していない場合は鳴りません。

2.6 検索条件の保存 / 読み込み

ファイルから選べる設定内容の保存 / 読み込みで行えます。ただし、ファイル名に関しては、それぞれのコンピュータの構成によって違うので、ほかのところでもうまく動かない事があります。その場合は、検索対象の設定で正しいファイル名を再入力して下さい。

3 うまくいかないときは

3.1 Ankou 起動時にメッセージが表示される

Ankou を起動した時に以下のようなメッセージが出る事があります。



これは、メッセージの通り、Java が古いので、Java を新しくインストールします。http://java.sun.com/ からリンクをたどっていき、新しい Java を入手します。頻繁にこのサイトは変わりますが、一般的には右側の Popular Downloads から J2SE という項目を探し、そのなかで Beta とついておらずかつバージョンが一番新しいものを Click します。ここで Download J2SE JRE を Click し、ライセンスに同意してから現れるダウンロードリストで自分の OS に適したものをダウンロードして実行します。詳しいことは http://java.sun.com/ でドキュメントを探して読むか、他のドキュメントを参照して下さい。