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

６．アノテーション内容を集計する †

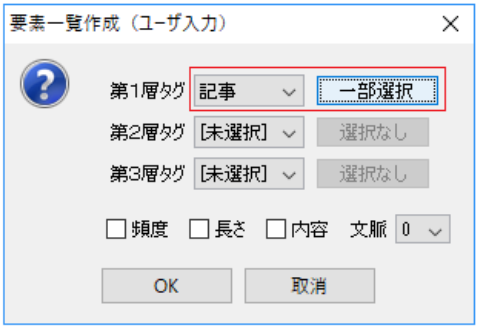
6.1 基本的な利用方法 †

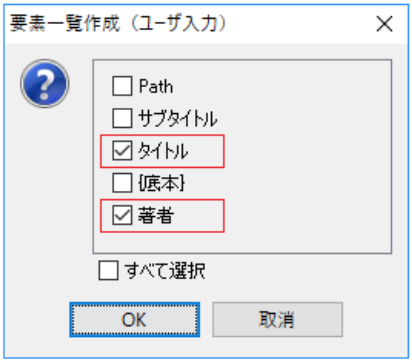
[ツール]⇒[一覧]⇒[ユーザ入力]で，言語資料に付与されているアノテーション内容の集計を行います。

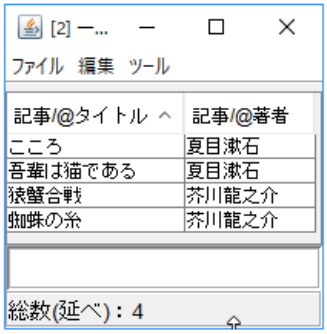
アノテーションは，タグによって記述されているため，タグを指定して集計することになります。例えば，『青空文庫』サンプルでは，一つの作品に対して，「記事」というタグが付与されています。タグはいくつかの属性を持つことができ，「記事」には作品名や著者名の属性があります。

『青空文庫』サンプルの「記事」タグを使って，作品一覧を作成する手順は，次のとおりです。

1. [ツール]⇒[一覧]⇒[ユーザ入力]で設定用のウィンドウを起動して，下図（左）のように「第１層タグ」のところに，「記事」を設定して下さい。
2. 選択メニューの右のボタンを押すと，下図（中央）のウィンドウが現れるので，「タイトル」「著者」にチェックを入れて下さい。
3. 二つのウィンドウの「OK」ボタンを押すと，下図（右）の記事一覧表が作成されます。

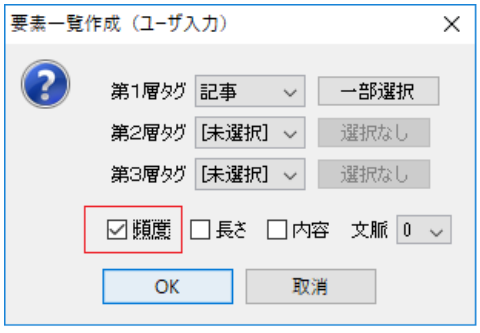


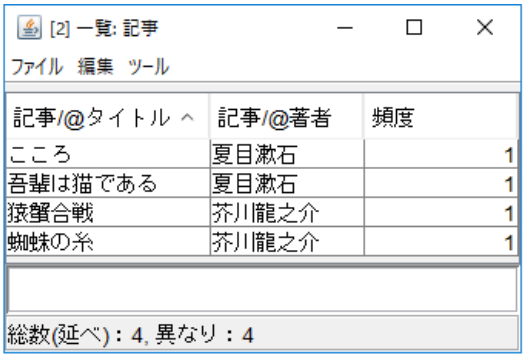




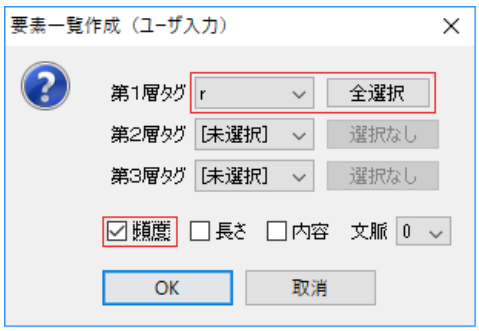
6.2 「頻度」オプション †

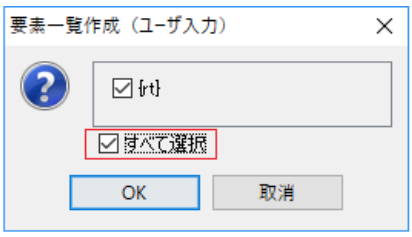
「頻度」オプションをチェックすると，一覧の各項目の出現頻度を計測することができます。下の図は，「記事」の頻度を表示したものです。『青空文庫』サンプルには，作品は重複して登録されていないため，当然，各作品の頻度は１になります。

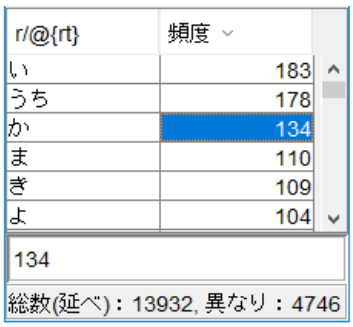




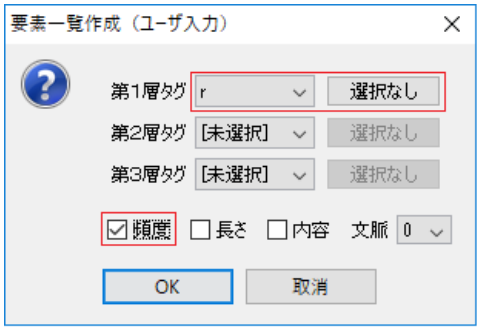
同様に，rタグ（ルビ）に対して，実行したのが次の図です。左図はrタグの属性rtを選択して表示したものです。rt属性には，ルビ本体が記述されているので，頻度付きのルビの一覧を作成することができます。

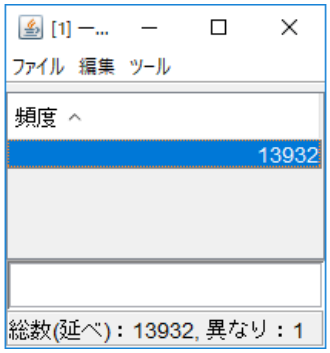






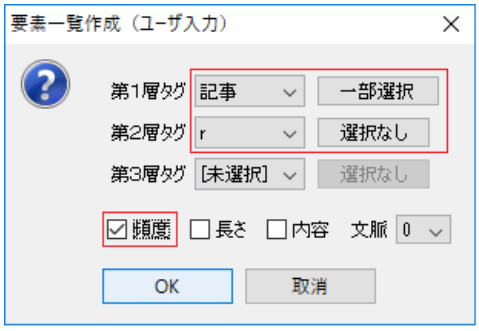
一方，右図は属性を選択しないで表示したものです。この場合，rタグの総数を計測することになります。

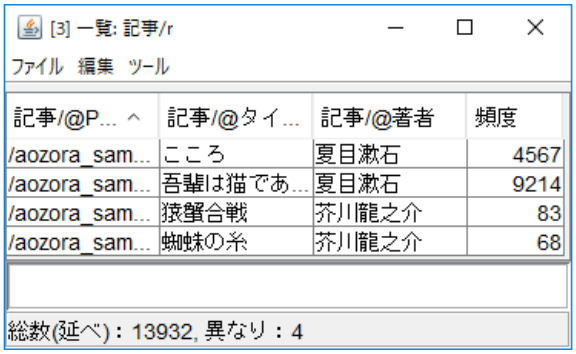




6. 3 「第x層タグ」の設定

タグは，「第１層タグ」「第２層タグ」「第３層タグ」に複数指定することにより，タグ間の包含関係を考慮した一覧の作成が可能です。下の図は，「第１層タグ」に「記事」タグ，「第２層タグ」にrタグを指定することにより，「記事」に含まれるルビの数を計測しています。



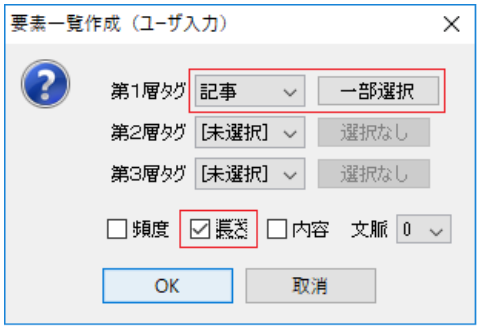


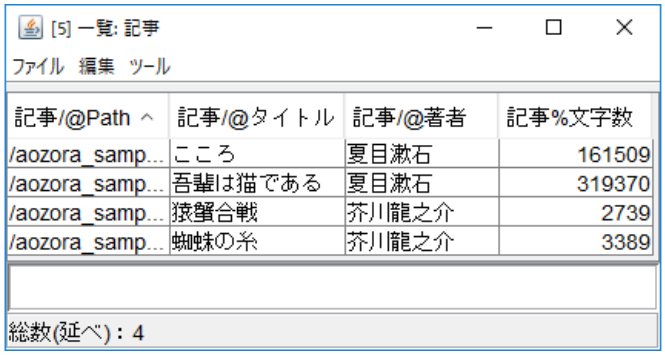
「頻度」は最下層のタグを対象に計測します。上の例の場合は，rタグの頻度を「記事」ごとに計測することになります。

6.4 「長さ」オプション

「長さ」オプションは，タグでマークアップされている文字列の長さを計測します。この際，マークアップされている文字列の中に含まれるタグや空白文字は，すべて長さ０として計測されます。

次の例は，「記事」タグでマークアップされている文字列（『青空文庫』サンプルの場合は一つの作品）に含まれる文字数を計測することになります。

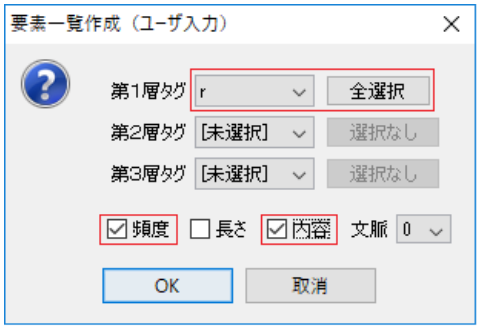


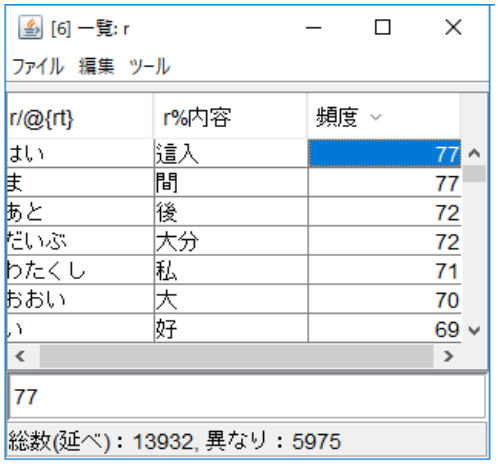


6.5 「内容」オプション †

「内容」オプションは，タグでマークアップされている文字列のための列を集計結果に追加します。

次の例は，rタグでマークアップされている文字列，つまり，ルビをつけられている文字列とルビをペアで集計しています。

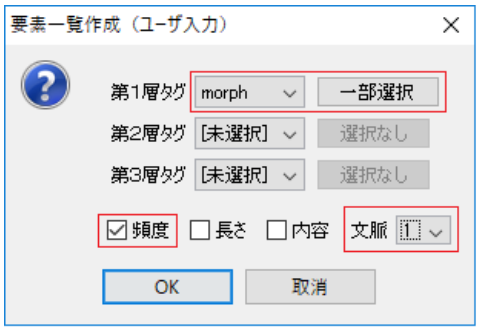


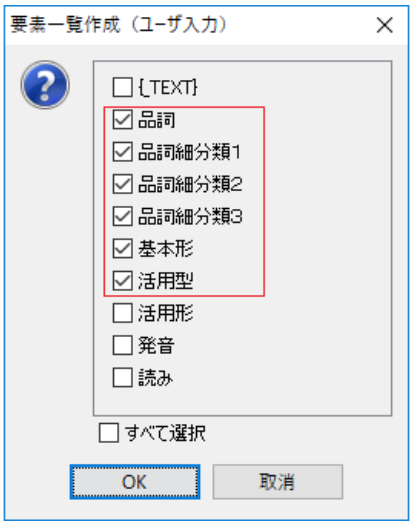


6.6 「文脈」オプション †

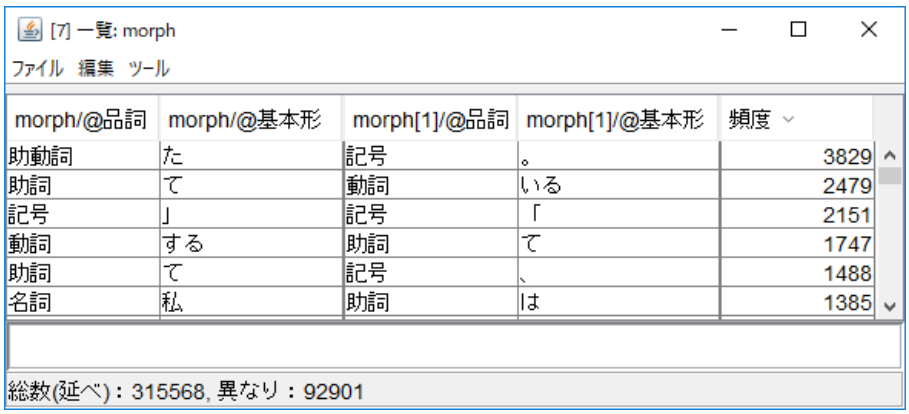
「文脈」オプションは，指定したタグのうち，最下層のタグに関して，後続するnタグ分の情報を集計結果に追加します。なお，nは「文脈」オプションで指定した値です。

例えば，『青空文庫』サンプル（形態素解析結果付き）のmorphタグを使って，単語bigramを作成してみます。morphタグは「単語」をマークアップするためのタグです。「文脈」オプションの値は１とします。また，あわせて，「頻度」オプションもチェックします。この場合，後続する１単語をペアにして計測することになるので，bigramが得られることになります。ただし，『ひまわり』の内部では，作品の最後の単語の次の単語は，次の作品の先頭の単語として，記述されているため，一部不要なデータbigramの定義に沿わないデータが含まれることに注意して下さい。





結果は，次のとおりです（一部の列のみ表示）。これを見ると，最も出現頻度の多いのは，「た」＋「。」であることがわかります。morph[0]，morph[1]がそれぞれ１番目，２番目の単語を表します。



6.7 外部アノテーション結果の表示 †

形態素解析結果など，外部アノテーションが施されている資料（言語資料の選択時に「外部DBあり」の資料）では，SHIFTキーを押しながら，検索結果をダブルクリックすると，当該の作品の外部アノテーション結果が一覧表示されます。

次の例は，『青空文庫』サンプル（形態素解析付き）で，「我輩」を検索し，その中の一つをSHIFT+ダブルクリックした結果です。１行１形態素で，「_TEXT」列が本文に相当します。