多言語対応コンコーダンサー『HASHI』機能説明書

ver 0.8.10 対応

2012/04/16

『HASHI』は多言語を扱えるコンコーダンサーです。コンコーダンサーとは、コーパスを分析するためのソフト、つまり言語を分析するソフトです。使用するテキストファイルは何もタグ情報の付与されていないプレーンテキストです。ファイルを指定すると自動で様々なタグ情報が付与され、それらのタグ情報を分析に使います。多言語を分析できますが、主に日本語の分析を主軸に開発されています。一部、英語分析に有利な機能も追加されています。コーパスの利点は、瞬時の検索と、その結果に基づく統計処理と言えますが、まず検索が目的通り詳細に行えなければ、その結果を扱う統計が正しく行われません。本ソフトではこの考えを元に、検索を高度に突き詰めることを目的に開発しています。テキストにできうるあらゆるタグを付与し、それらを全て個別や組み合わせて自由に扱え、言語のあらゆる面を自由に見ることができるソフトです。

本ソフトは極めて簡単に扱うことができます。コーパスやコンコーダンサーの初心者でも、現れる誘導通りに操作をすればすぐに最初の結果を見ることができます。１つ１つの操作も簡単で、基本的に１つのボタンを押せば１つの処理が行われるように作ってあります。また、コンコーダンサーの使用に慣れた方にとっても煩わしい工程はなく、何かをするたびにいちいち初心者に合わせた丁寧な説明が現れ手間をかけることもありません。１ボタン＝１動作となることは、使用者自身が作業の全ての工程を確認しながら操作を進められるため、極めて自由度の高い操作方法と言えます。いくつもの設定をした上で実行をすることで高度な結果がいきなり現れるのではなく、そこにいたる過程を１つ１つ確かめながら操作し目的の形式や結果に導くように使います。それにより、工程の途中で選択の違いが有ればその段階で気づきすぐに修正できる、また面白い結果が現れれば更にそこから１つ先へ操作を進めるなど、その都度その都度、頭に浮かんだことをすぐに実行して確認し使用を進めることができます。高度なソフトが素晴らしい結果を見せてくれるのではなく、あたかも手作業で直接データを扱い、自分自身でデータの様々な面を引きだしてくる、そういうソフトになることを目的に開発しています。

※HASHI 0.8.10 は現在未完成の部分を含み、開発中ですので開発の進捗に従って本マニュアルも変更されます。

目次

使用準備 8
起動 9
処理の種類 10
分析言語 10
基本的な画面構成 11
ファイル選択 13
全文表示(Sentence) 15
　表示変更 16
　複数行表示 17
　検索 18
　検索結果のみ 19
　集計のみ 19
　最少語数、最大語数 20
検索語と使われている文脈(KWIC) 21
　表示項目の変更 21
　ソート 22
　複数条件でのソート 22
　Sort Type 23
　頻度でのソート 23
　位置毎分割 24
　ソート時の表示項目 24
　ソート後の表示項目変更 25
　補助表示 25
　降順 26
　本文リンク 26
共起語の頻度とスコア(Collocates) 27
　表示項目の変更 27
　スコア表示 28
　ソート 28
　表示最低数の指定 29
　降順 30
　50 音順の降順 30
位置ごとの共起語の頻度(Picture) 31
　表示項目の変更 31

算出する値の変更 32
表示最低数の指定 32
ソート 33
降順 33
Words for Display 34
頻度数での KWIC(POPAK) 35
語を表示 35
表示項目の変更 36
ソート 36
合計に加える最低数の指定 37
Number of Calculate 37
集計値 38
算出する値の変更 38
降順 39
テキスト全体の語の頻度(Freq) 40
表示項目の変更 40
ソート 41
検索 41
降順 42
同じ並びの語の数(N-gram) 43
Ngram のサイズ 43
記号排除 44
表示項目の変更 44
ソート 45
表示最低数の指定 45
作成する Ngram の単位 46
降順 47
検索 47
穴空きの Ngram 48
特徴的な語(Keyness) 49
参照ファイル 49
対数尤度比とカイ二乗 50
表示項目の変更 51
ソート 51
降順 52
検索 53

文字列の検索(Grep) 54
ファイル選択 54
分析ファイルの設定 55
検索 55
ソート 56
複数条件でのソート 56
Sort Settings 57
本文リンク 57
Sub Key 58
KWIC 形式 58
検索文字列のみの表示 59
検索文字列を含まない結果のみの表示 59
行番号表示 60
検索文字列を単語として検索する 60
通常の処理にまつわること 62
「検索」「再描写」の違い 62
分析言語 63
大文字小文字同時検索 64
検索語の文字列の小文字化 64
整形単位 66
整形単位の選択 66
語単位の整形ルール 67
語単位の整形例 68
構成形態素 69
検索 69
語末ソート 70
構成形態素の検索 70
連続する構成形態素の指定 71
複合検索 71
タグ項目 72
日本語の場合 72
英語の場合（TreeTagger 有り） 74
検索 75
通常検索 75
検索語句の詳細指定 79
各項目での検索の実例 81

複数条件の指定 86
and 条件、or 条件 86
入力方法 87
簡易選択 87
正規表現自動生成 88
Grep での特殊ボタン 93
周囲の語句の指定 95
検索のメカニズム 97
柔軟検索 97
詳細検索 99
周囲の語句 99
再描写 100
停止 100
オリジナルコーパス作成 101
マーカーを引く(Mark) 102
色選択 102
マーカーの付け方 102
編集を１つやり直す 103
編集全体をやり直す 103
色に名前を付ける 104
検索 105
表示項目の変更 105
編集結果の保存 105
付与したマーカーの利用 106
Sentence でのマーカーの利用 106
KWIC でのマーカーの利用 108
Collocates、Picture、POPAK でのマーカーの利用 110
Freq でのマーカーの利用 111
テキストデータの編集(Edit) 112
語タグ 113
タグ名と要素リストの作成 113
タグ付与 114
語タグの追加 115
語タグの削除 116
語タグの入れ替え 116
行タグ 117

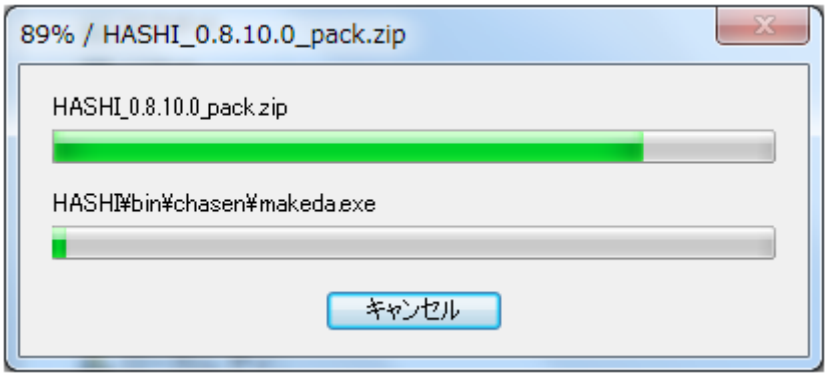
行タグの名前変更と付与.....117
行タグの追加、削除、入れ替え.....118
属性タグ.....119
属性タグ要素の設定.....119
属性タグの付与.....120
属性タグ、項目の削除.....120
ファイルタグ.....122
ファイルタグの追加と設定.....122
音声ファイルの指定.....122
音声コーパス化.....123
テキスト本体の編集.....125
語の編集.....125
形態素解析確認.....127
行の編集.....129
表示項目の変更.....130
検索.....130
表示行数の制限.....131
編集のやり直し.....131
編集結果の保存.....132
通常の処理でのオリジナルタグの使用.....133
追加ボタン.....133
オリジナルタグの使用.....134
オリジナルタグの使用可能処理.....137
音声再生.....137
Edit の設定変更.....137
フォルダでの一括ファイル選択.....139
テキストの整形段階での行情報の付与.....141
ファイルタグ、属性タグ、行タグの指定書式.....141
ファイルタグ名、属性タグ名、行タグ名の指定.....144
ファイルタグ名、属性タグ名、行タグ名の指定書式.....145
テキストデータ編集での行情報のタグ名の変更.....146
語単位の整形ルールの変更.....147
整形済みファイルの処理(Files).....149
テキストファイルへの書き出し.....149
ファイルの削除.....151
ファイルの複製.....152

ファイルの分割 152
別編集の同一ファイルの統合 153
複数のファイルの連結 155
整形データの移動、配布 156
Input menu 157
文の区切り文字記号 157
左の取得幅、右の取得幅 158
分割表示の際の 1 語の幅 159
補助表示の語数 160
語数を合わせる 160
表示フォント 161
Ngram での Input menu 161
Grepd の Input menu 161
ファイルへの保存 162
.slk 162
.tsv 163
.txt 163
画面の直接コピー 163
Output menu 164
オプション 165
パソコンの文字コード 165
Window 内の文字サイズ 165
Window の文字サイズ 166
メインカラー 167
各種統計の定義 168
ファイルの総語数と使用するデータの範囲 169
形態素解析ソフトの設置 170
茶筌用辞書の置き換え 170
形態素解析ソフトの設置 170
著作権 171
『HASHI』について 171
HASHI を使って作成したコーパスの公開について 171
『茶筌』について 172
その他のソフトについて 173

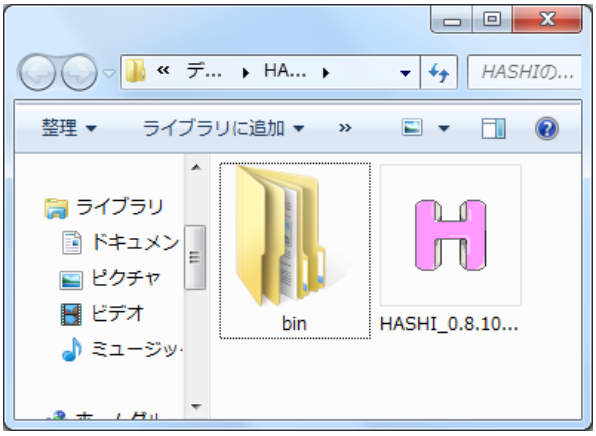
使用準備

ホームページから最新のバージョンをダウンロードすると、「HASHI_(バージョン)_pack.zip」というファイルが保存されます。

圧縮ファイルですので、解凍してお使いください。ダブルクリックで圧縮されたまま内部のフォルダが見られる設定のパソコンもありますが、圧縮されたままだと動作しません。必ず解凍してからお使いください。



解凍されると「HASHI」というフォルダが現れます。



「HASHI」フォルダの中に更に「bin」というフォルダと「HASHI_(バージョン).exe」が入っています。「bin」の中には ChaSen などの HASHI の動作に必要なソフトが入りますので、HASHI 本体だけ別の場所に移動するなどせず配置関係はそのままでお使いください。他に、同じホームページでは補助ソフトや旧バージョンの HASHI の公開もありますが、それらはソフト単体での公開になりますので、解凍後、このフォルダに入れて使用ください。

起動

「HASHI_(バージョン).exe」がソフトの本体ですのでダブルクリックで起動します。

起動後に出てきたウィンドウに並ぶボタンが、このソフトでできる処理の一覧です。

処理名が書かれたそれぞれのボタンで各処理のウィンドウが現れます。

各ボタンで、それぞれの処理のウィンドウが開きます。

処理の種類

処理名	単位	対象範囲	説明
全文表示 (Sentence)	単語	テキスト全文	テキスト全文を読みながら検索語も調べられる
検索語と使われている文脈 (KWIC)	単語	検索語と 周囲の語	検索語のテキスト中の振る舞いを実例で見る
共起語の頻度とスコア (Collocates)	単語	検索語と 周囲の語	検索語と共起語の傾向を数字で見る
位置ごとの共起語の頻度 (Picture)	単語	検索語と 周囲の語	位置ごとでの検索語と共起語の数を見る
頻度数での KWIC (POPAK)	単語	検索語と 周囲の語	検索語と共起語の非連続の並びを実例で見る
テキスト全体の語の頻度 (Freq)	単語	テキスト全文	テキスト全体での語の頻度を見る
同じ並びの語の数 (N-gram)	単語	テキスト全文	実際に使われた語の並びのどれが多いかを見る
特徴的な語 (Keyness)	単語	２ファイルの テキスト全文	２つのファイルを比べ、テキストに特徴的な語を抽出する
マーカーを引く (Mark)	単語	テキスト全文	テキストに使用者独自の簡易タグを付ける
テキストデータの編集 (Edit)	単語	テキスト全文	テキストに使用者独自の様々な本格的なタグを付ける
整形済みファイルの処理 (Files)		ファイル	ファイル全体の削除や分離や結合などの処理を行う
文字列の検索 (Grep)	文字列	検索文字列と 周囲の文字列	整形を行わないプレーンテキストに、文字列で検索をする

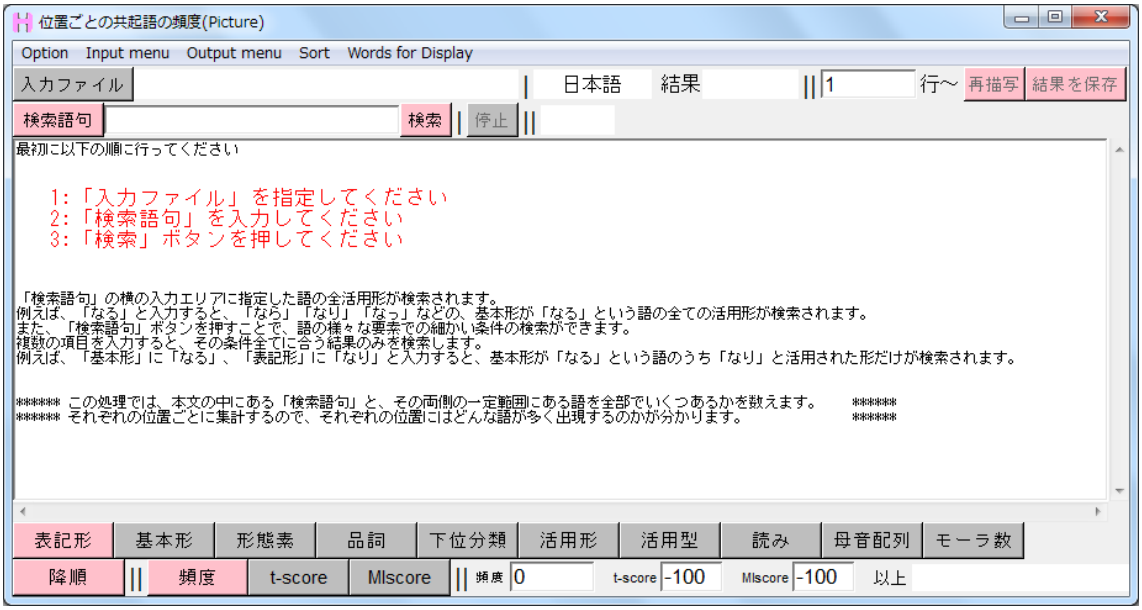
分析言語

「日本語」「英語」他
本稿では日本語を分析言語として説明を行います。
分析対象を他の言語に変更する方法は後で説明します。

基本的な画面構成

メインウィンドウの各ボタンを選択するとそれぞれの処理ウィンドウが現れます。ウィンドウはいくつかの区画に分かれています。

処理ごとにウィンドウ内のボタンの種類が違いますが、各処理でだいたい共通しているボタンの意味合いの説明をします。ここでは「位置ごとの共起語の頻度(Picture)」のウィンドウを例に、まとまりごとに大きく分けて説明します。



ウィンドウ最上部のツールバーは、主に何かの指定をし、その後「再描写」を押して変更を反映させるなど、組み合わせなどで使うものです。

ウィンドウ上部のパネルのボタンは、読み込みデータを指定したり、検索語句を指定し、検索をするなど、ソフトの最も基本機能を扱うものです。

ウィンドウ下部のパネルのボタンは、表示形式を変えるもので押すとすぐに結果に反映されます。

- 「Option」から始まる横一列は、データを検索する前と後の両方で使います。原則的に複数の項目から一つを選ぶなどを行いますが、それらの選択項目を常に表示させておくと画面が見づらくなりますので、普段はここに収まっています。
- 「入力ファイル」から始まる横一列は、本文のあるテキストファイルと、その種類などを指定するものです。また、データのあった数などの情報が表示されます。
- 「検索語句」から始まる横一列は、調べたい語を指定し、検索などを行うものです。
- スクロールバーの付いている大きな枠の中は、検索したデータが並べられて表示される部分です。

●「表記形」から始まる横一列は、データの検索後に行う項目で、データが表示される際、それぞれの語をどの形で表示させるかを選択するものです。この中から一つを選びます。

●「降順」から始まる横一列は、データの検索後に行う項目で、データの表示の並べ方を決めるものです。それぞれが独立していて、バラバラに ON OFF で選ぶものと、あるセットの中から１つを選ぶものとに分かれます。

実際の検索では、主に以下の手順になります。

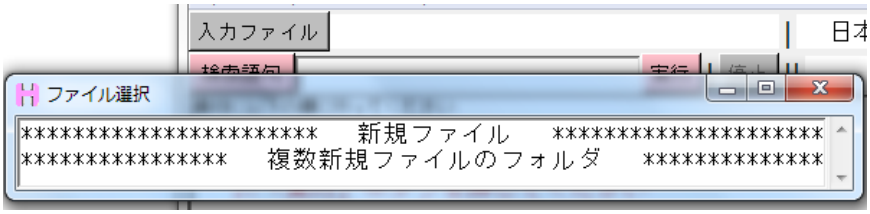
①「入力ファイル」のボタンで調べたい本文の入っているテキストファイルを選択。

②「検索語句」に検索したい語を入力。

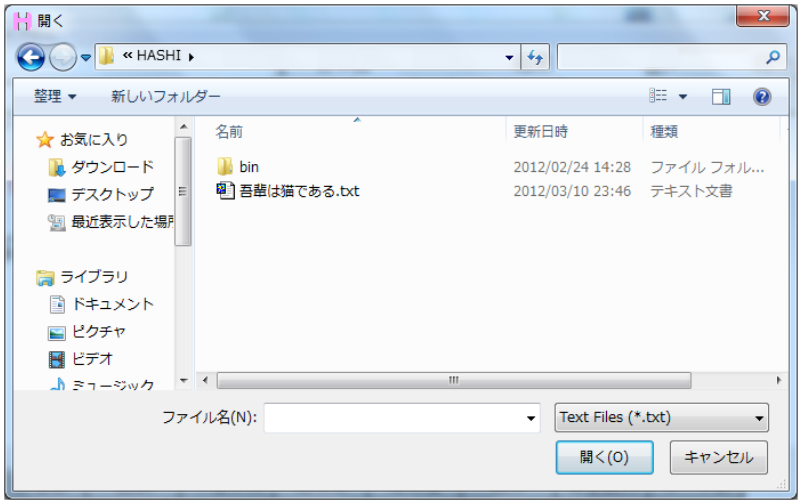
③「検索」ボタンをクリックして検索を開始。

ファイル選択

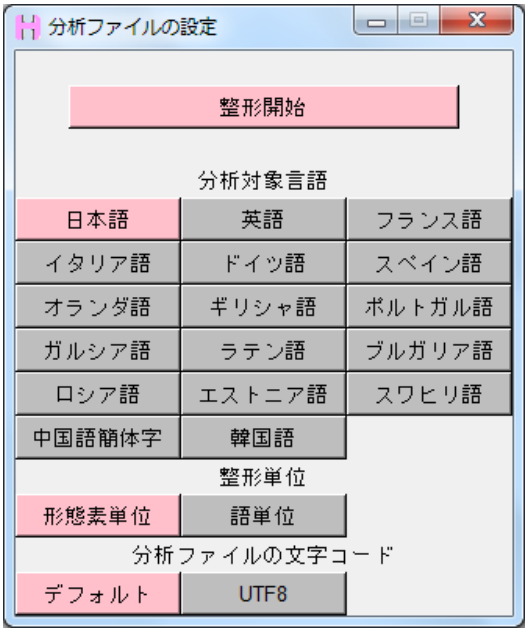
本ソフトで扱うテキストファイルの指定方法を説明します。
ウィンドウ上部、一番左の「入力ファイル」ボタンでテキストファイルの指定をします。



ファイル選択リストが現れたら、「新規ファイル」を選択します。



すると、ファイル選択ウィンドウが現れるので、使用したいテキストファイルを選択し、「開く」をクリックします。



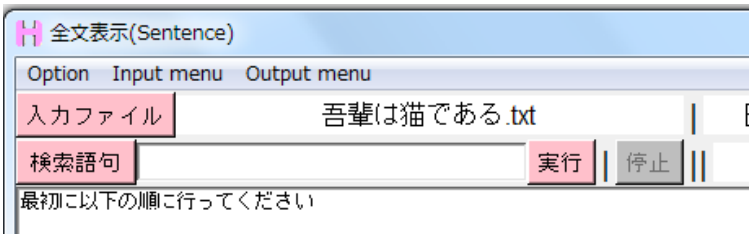
続いて、分析ファイルの設定ウィンドウが現れます。

分析するテキストファイルの言語と、分析したい単位、分析するファイル内の文字コードを選択し、「整形開始」ボタンをクリックします。

テキストの整形が開始されます。

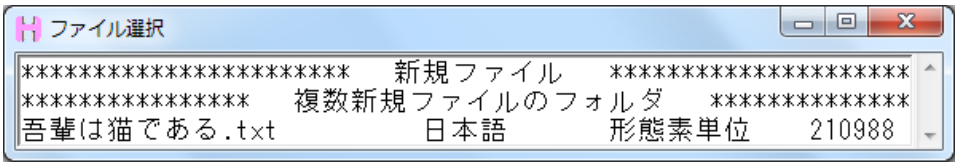
「一次整形」「二次整形」「語数調査」「最終整形」の順で整形が行われます。
この処理は時間のかかるもので、小説１冊で１分程度かかることがあります。テキストの分量が多いと時間も長くかかります。

整形が終わると「入力ファイル」の欄に指定したファイル名が入り扱えるようになります。



その後、「実行」などでそれぞれの処理を行えるようになります。

一度整形したファイルはファイル選択リストに載るので、次回からはそれを選択します。



ファイル名、ファイルの言語、分析単位、総語数の順で情報が表示されます。
整形済みファイルを選択した場合は、そのまますぐに扱えます。

全文表示(Sentence)

テキスト本文がそのまま読める形式で、本文の左に行番号と行中の語数が表示されます。語数は記号を抜いた数です。語句の検索も行え、表示された本文の中の検索語句の箇所が赤く表示されます。検索語句の数や、テキスト内での散らばりも表示されます。

	行	語	1行平均
1	234	29944	127.97
2	234	32539	139.06
3	234	14465	61.82
4	234	45518	194.52
5	234	15134	64.68
6	234	14971	63.98
7	234	9010	38.50
8	234	6534	27.92
9	234	7962	34.03
10	241	12214	50.68
total	2347	188291	80.23

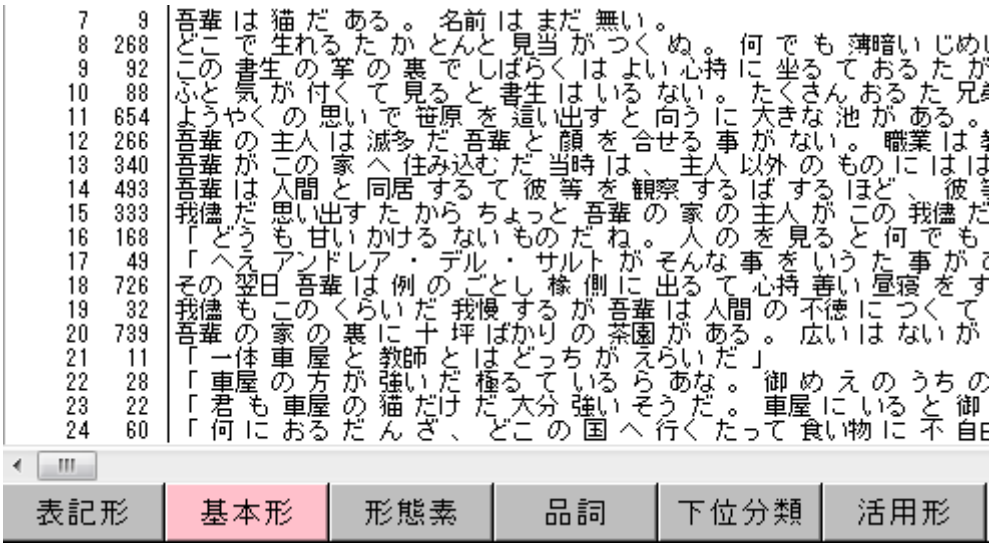
行	語数	
1	5	吾輩 は 猫 で ある
2	2	夏目 漱石
3	0	
4	0	
5	1	一
6	0	
7	9	吾輩 は 猫 で ある 。 名前 は まだ 無い 。
8	268	どこ で 生れ た か とんと 見当 が つか ぬ 。 何 で も 薄暗い じめじめ し た 所 で ニャーニャー 泣い て い た 事 だけ は 記憶 し て い
9	92	この 書生 の 掌 の 裏 で しばらく は よい 心持 に 坐っ て おっ た が 、 しばらく する と 非常 な 速力 で 運転 し 始め た 。 書生 が
10	88	ふと 気 が 付い て 見る と 書生 は い ない 。 たくさん おっ た 兄弟 が 一 疋 も 見え ぬ 。 肝心 の 母親 さえ 姿 を 隠し て しまっ た
11	654	ようやく の 思い で 笹原 を 這い出す と 向う に 大きな 池 が ある 。 吾輩 は
12	266	吾輩 の 主人 は 滅多 に 吾輩 と 顔 を 合せる 事 が ない 。 職業 は 教師 だ そ
13	340	吾輩 が この 家 へ 住み込ん だ 当時 は 、 主人 以外 の もの に は はなはだ 不
14	493	吾輩 は 人間 と 同居 し て 彼 等 を 観察 すれ ば する ほど 、 彼 等 は 我儘
15	333	我儘 で 思い出し た から ちょっと 吾輩 の 家 の 主人 が この 我儘 で 失敗 し
16	168	「 どう も 甘く かけ ない もの だ ね 。 人 の を 見る と 何 で も ない よう
17	49	「 へえ 、 アンドレア ・ デル ・ サルト が そんな 事 を いっ た 事 が ある か い
18	726	その 翌日 吾輩 は 例 の ごとく 縁側 に 出 て 心持 よく 昼寝 を し て い たら

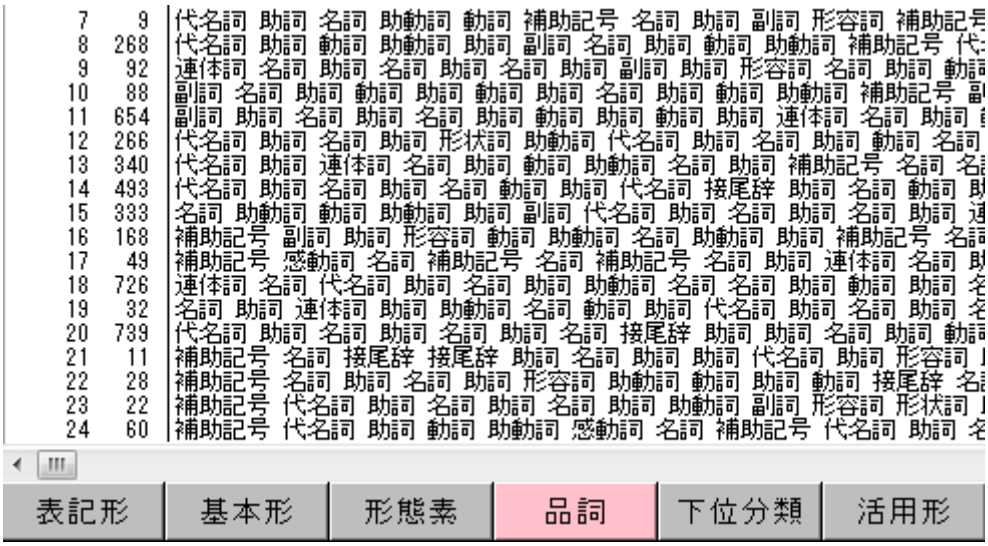
この処理では、本文がそのまま文章として読める形式で表示されます。

「実行」ボタンで本文が表示されます。
左に行番号と行中の語数が表示され、右側に本文が表示されます。1 行は、元のテキストでの改行で区切られた範囲です。行中の語数は記号を抜いた数です。
本文での実際の文章の順に表示されていき、表示が最後まで完了したら表示画面の一番上にテキストを 10 分割した内の各、語数と平均語数、1 行平均での語数が表示され、一番下に全行数と全語数、1 行平均での語数が表示されます。

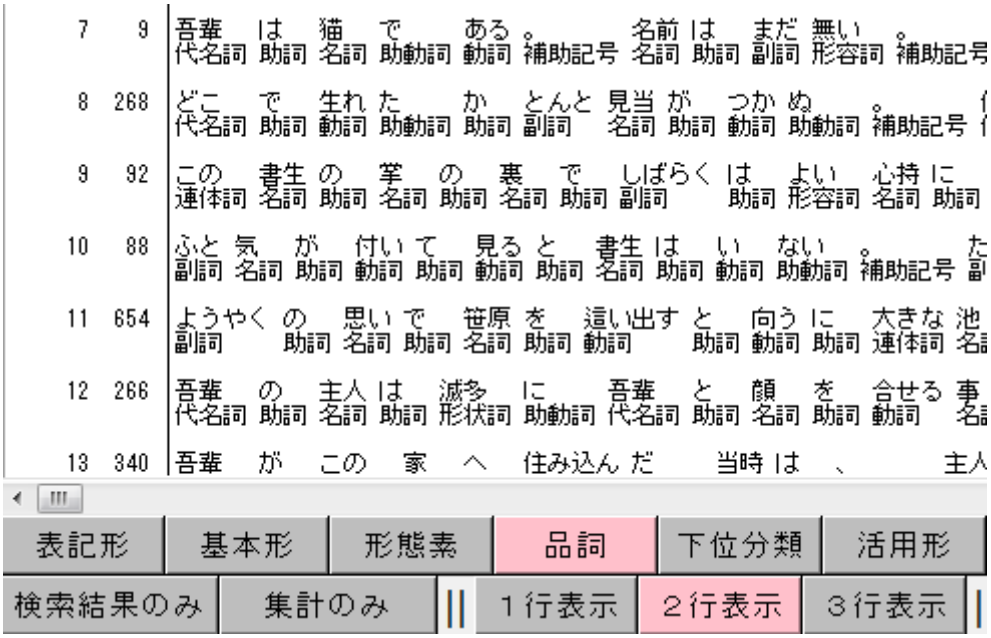
表示変更

「基本形」「品詞」「活用形」などウィンドウ下部のボタンで表示される項目が変わります。

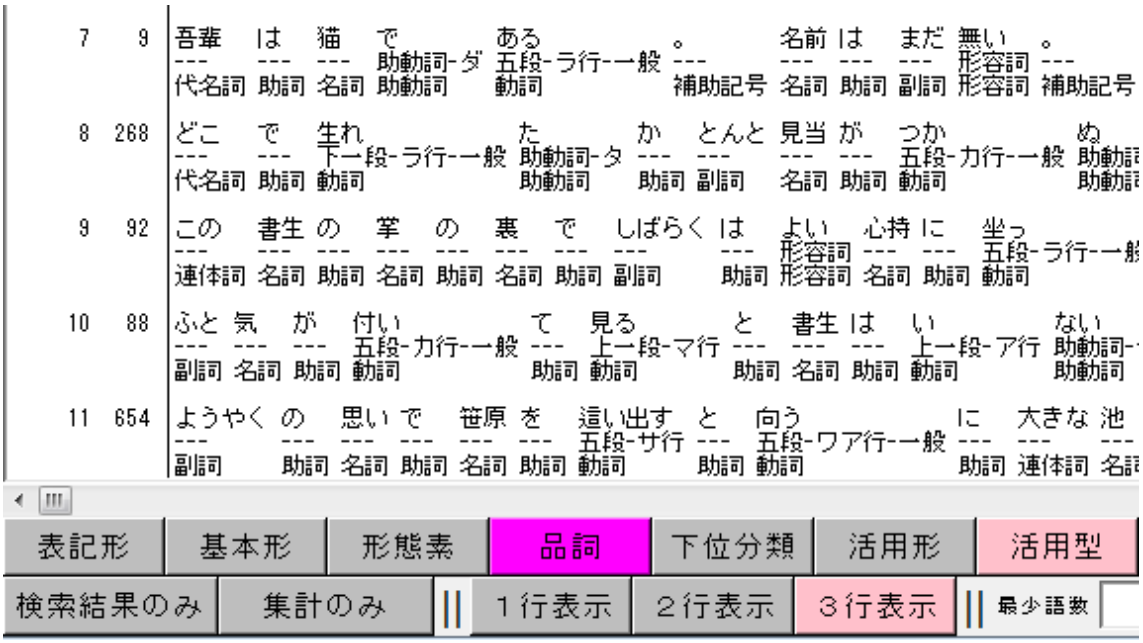




複数行表示



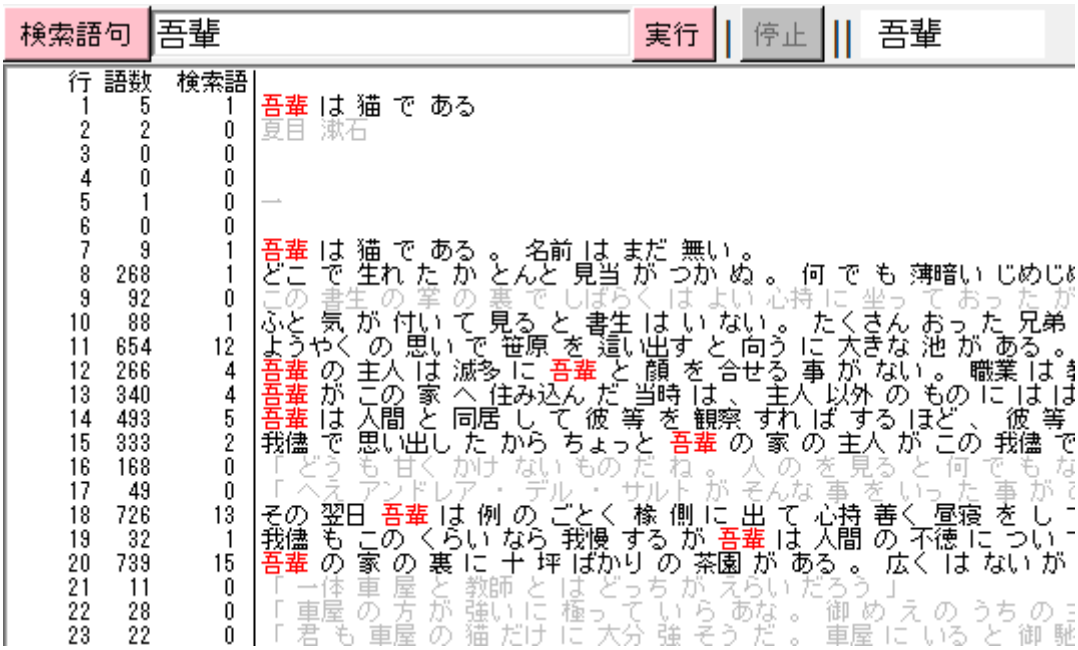
「２行表示」を選択すると、上の行に表記形、下の行に選択した項目の形式で本文が２行で表示されます。



同様に「３行表示」で、本文が３行の形式で表示されます。上の行が表記形、真ん中の行が、現在選択している項目、下の行が、直前に選択していた項目で表示されます。直前の選択項目は、ボタンが通常とは違う色で表示されます。

検索

検索語句を指定してから「実行」を押すと、本文中にある検索語句が検出されます。



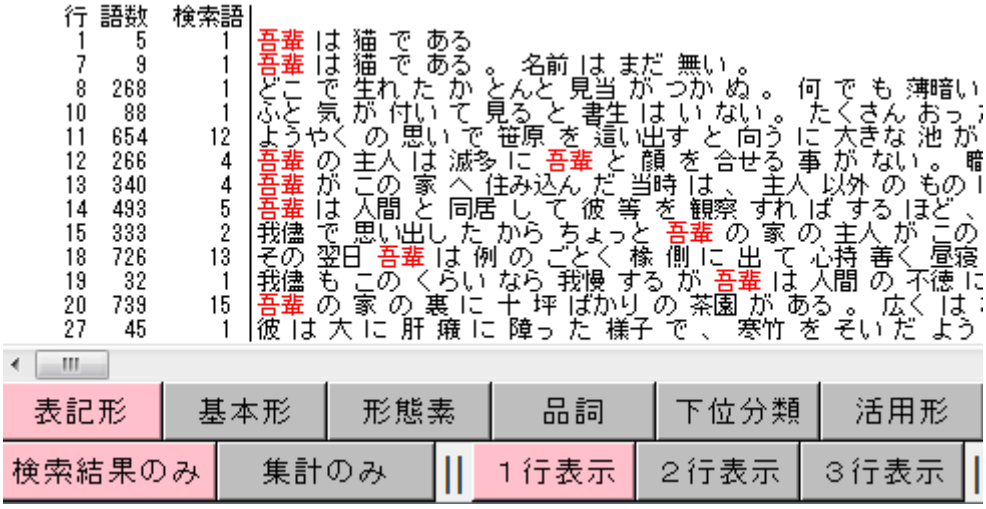
検索語が含まれない行がグレーになり、検索語の含まれる行が浮き出ます。本文表示部分にある検索語句は赤く表示されます。

各行の語数の右側に検索語の出現数の表示が加わります。

同様に、本文表示の完了後に画面の一番上に表示される本文全体の行数、語数を 10 分割した数字の右側にも検索語の出現数が表示されます。

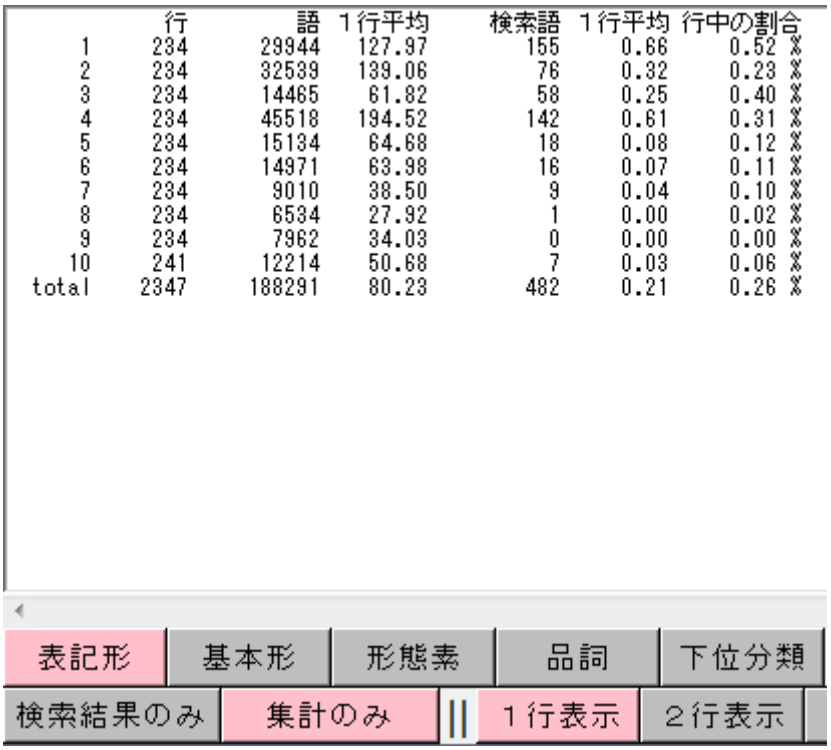
	行	語	1行平均	検索語	1行平均	行中の割合
1	234	29944	127.97	155	0.66	0.52 %
2	234	32539	139.06	76	0.32	0.23 %
3	234	14465	61.82	58	0.25	0.40 %
4	234	45518	194.52	142	0.61	0.31 %
5	234	15134	64.68	18	0.08	0.12 %
6	234	14971	63.98	16	0.07	0.11 %
7	234	9010	38.50	9	0.04	0.10 %
8	234	6534	27.92	1	0.00	0.02 %
9	234	7962	34.03	0	0.00	0.00 %
10	241	12214	50.68	7	0.03	0.06 %
total	2347	188291	80.23	482	0.21	0.26 %

検索結果のみ



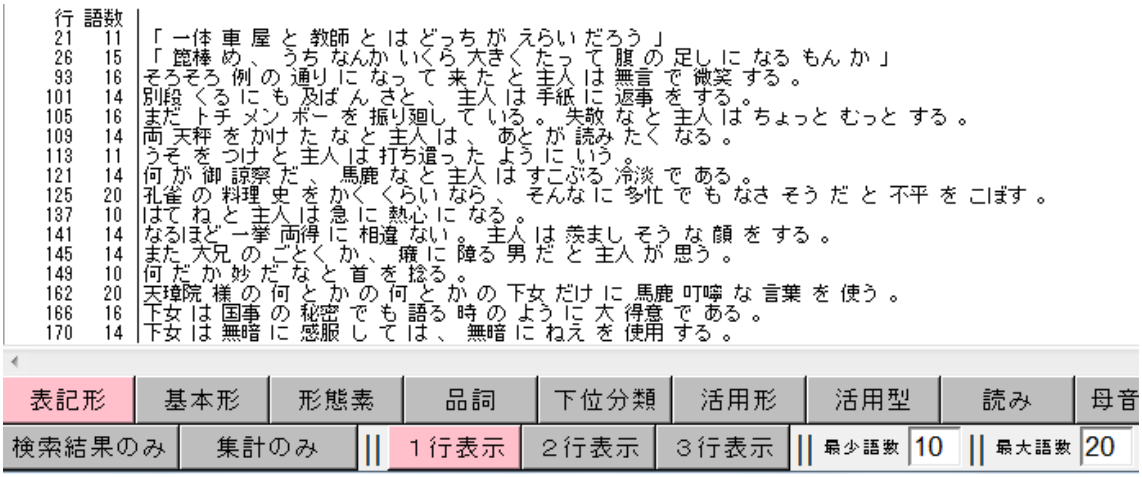
ウィンドウ下部の「検索結果のみ」ボタンで、検索語を含む行のみが表示されるようになります。

集計のみ



ウィンドウ下部の「集計のみ」ボタンで、本文表示がされずに、本文の語数や検索語の数などの集計結果のみの表示なります。本文の分量が多く表示に時間がかかる場合に使用します。

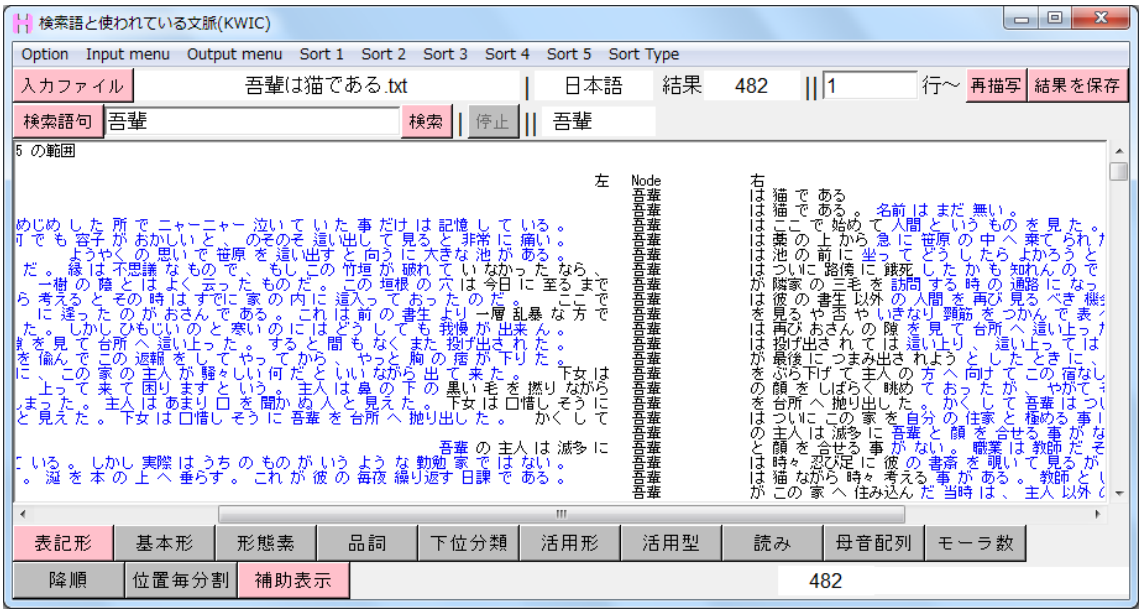
最少語数、最大語数



ウィンドウ下部、右側の「最小語数」「最大語数」の数値を指定すると、指定した範囲の語数の行のみが表示されます。最小と最大を片方のみ指定することもできます。両方の数値を一致させれば、完全に決まった語数の行のみを表示させることができます。

検索語と使われている文脈(KWIC)

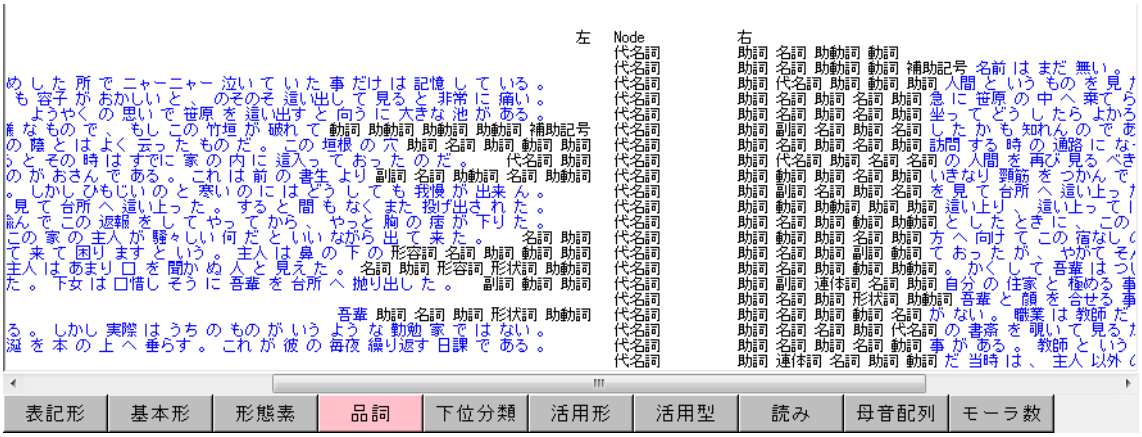
この処理は、指定した語を検索し、見つかった検索語を中心に、左右にそれぞれの左右文脈を配置して表示するものです。



検索語に焦点を当てながら、実際の文脈も確認できる形式です。左右の表示内容を絞ったりソートをかけることで、実際の本文の確認と、頻度などの数値としての利用の両方を兼ね備えた処理になります。検索語句を指定してから「検索」ボタンで検索を開始します。

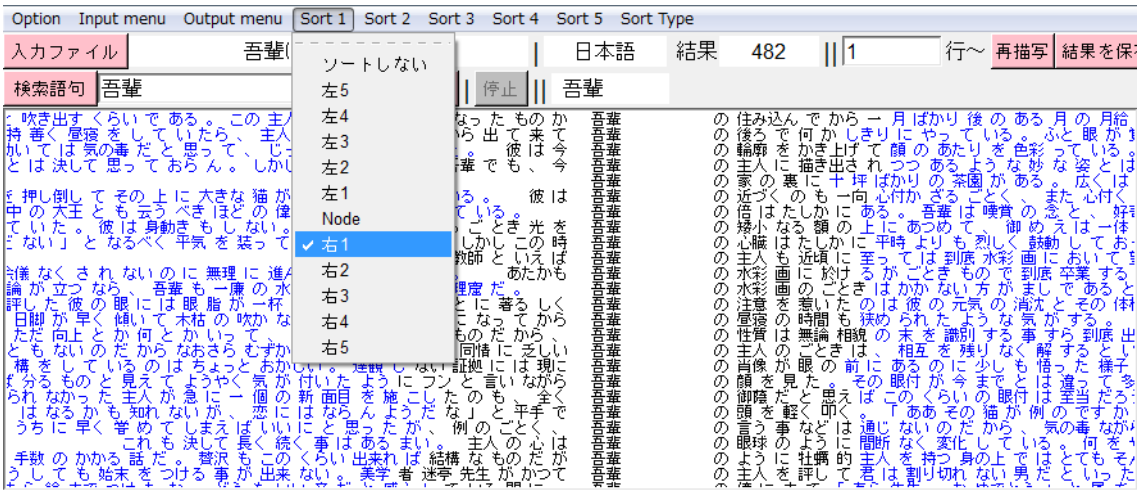
表示項目の変更

他の処理と同様、ウィンドウ下部のボタンで表示項目を切り換えることができます。



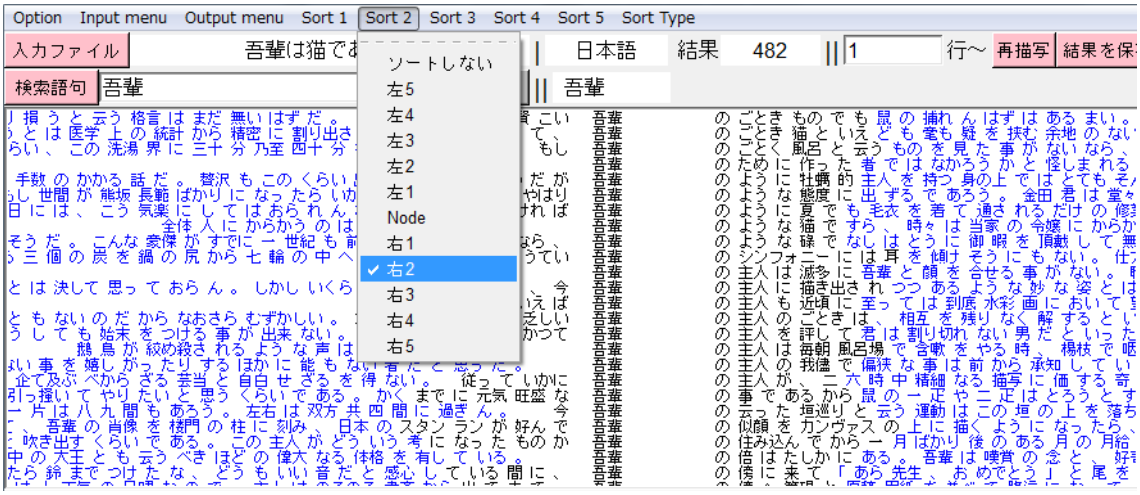
ソート

ウィンドウ上部の Sort1～Sort5 を指定すると、指定した位置で並び変えができます。



Sort1 を指定し、「再描写」ボタンを押すと、指定した位置の語の 50 音順を基準に、全ての行が並び変わります。「位置」は検索語を基準として左右いくつ分離れた場所かを意味します。検索語の位置を「Node」とし、それより 1 つ右側を「右 1」、2 つ右側を「右 2」、1 つ左を「左 1」とします。左右の取得幅の範囲でしていできます。

複数条件でのソート

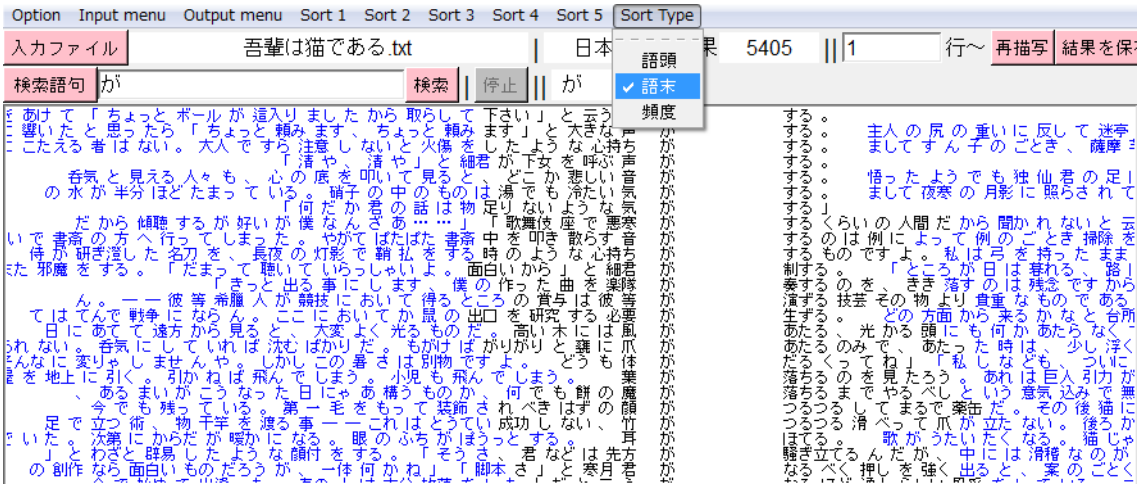


Sort1 を指定した上で、Sort2 を指定し、「再描写」を押すと、複数の条件でソートができるようになります。まず Sort1 の位置で並び変え、そこにある語が全く同じだった場合に、その中で次の条件として Sort2 で指定した位置の語を基準に並び変えが行われます。

ソートは第五条件まで指定できます。必ず、Sort1、Sort2、Sort3・・・の順で指定します。Sort1、Sort3 のように間を飛ばして指定した条件は反映されません。

Sort Type

位置以外のソートする基準として、タイプも選べます。ウィンドウ上部の Sort Type で選択します。ソートタイプは「語頭」「語末」「頻度」です。通常では「語頭」が選ばれています。これを「語末」に変えると、条件に指定した位置の語の語末から語頭に向かって 50 音順に並び変わります。活用形ごとにまとめて並べたいときなどに使います。

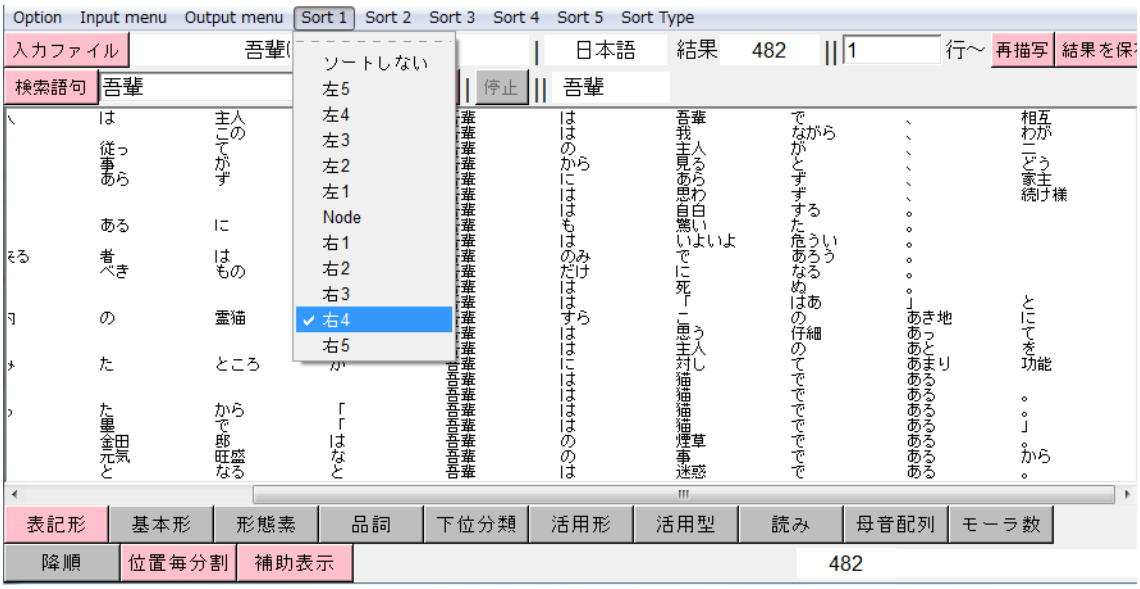


頻度でのソート

ソートタイプを頻度に変えると、50 音順ではなく、多い順に並び変わります。どの形が一番多く使われているかなどの確認に使います。

位置毎分割

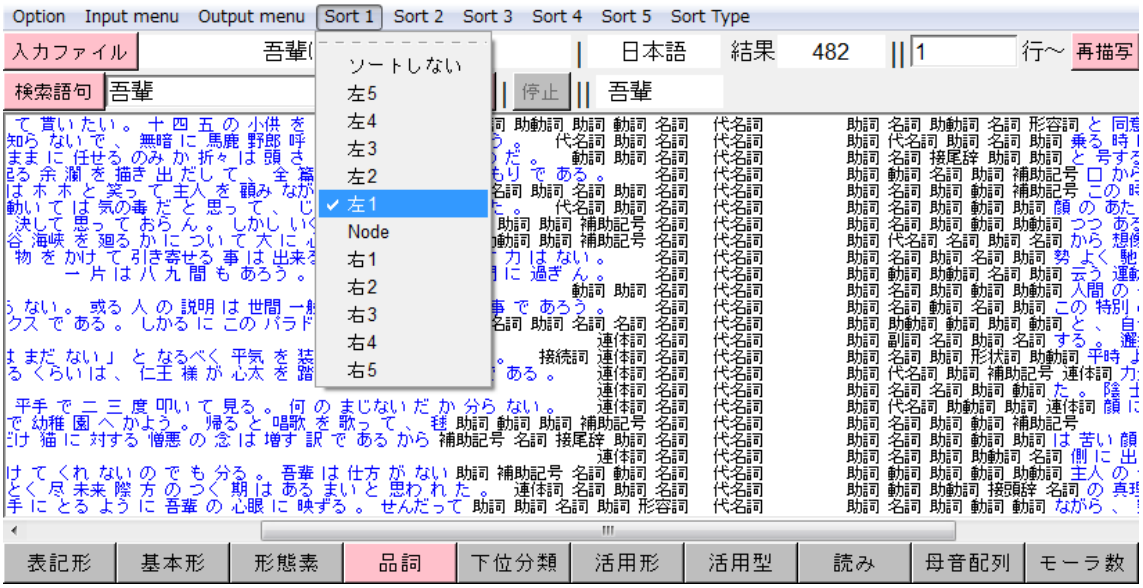
ソート位置が Node よりも遠かった場合、並び変えた語がうまく縦に揃わずによく分からないことがあります。その際に、ウィンドウ下部の「位置毎分割」ボタンを使います。



これを使うと、それぞれの位置で縦に語の幅が揃って表示されるため、ソートした位置の語が見やすくなります。

ソート時の表示項目

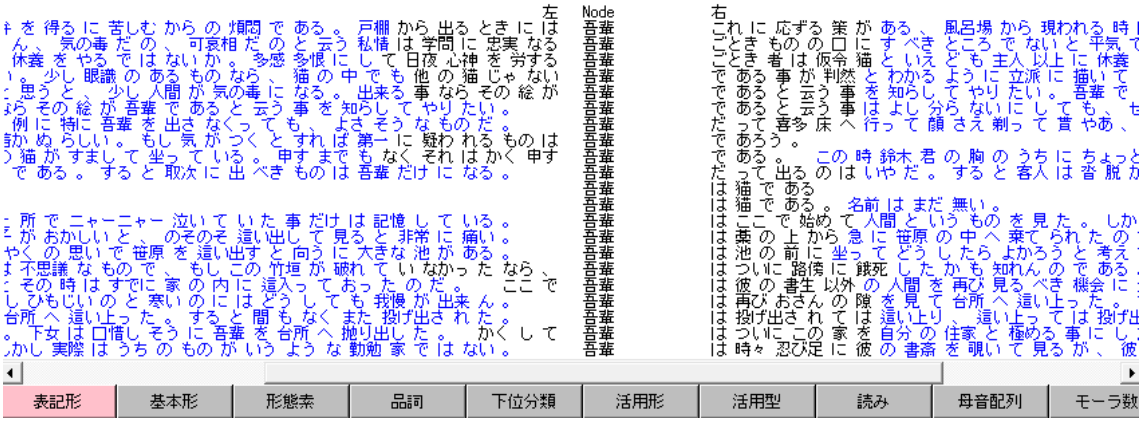
ソートをする際に表示している項目での順に並べ替えが行われます。



ソート位置を指定して再描写でソートが実行されるときに表示している項目で並びます。

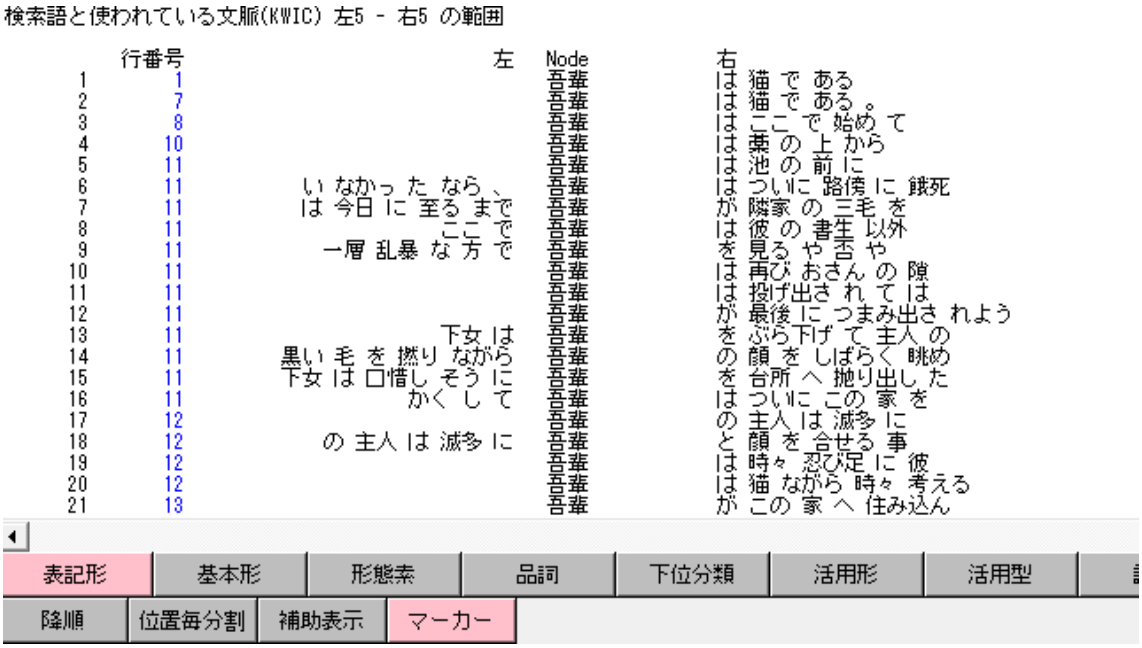
ソート後の表示項目変更

ソートした後に表示する項目を変えた場合、ソート時の行の順番は変わらないので、表示項目だけが変わり、表示される行の順番はそのまま維持されます。



補助表示

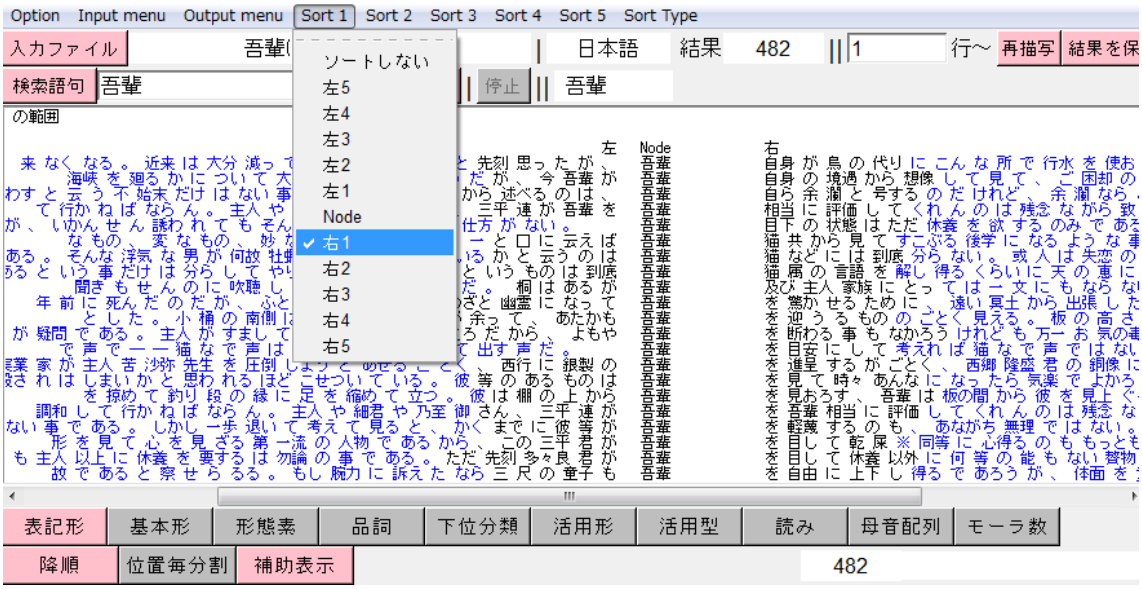
検索結果の周囲の青い文字での補助表示の表示、非表示の切り替えができます。



検索結果を表す黒い文字で表示された語はその後の統計などに使うため、左右の範囲が限定されています。しかし、その範囲の語だけの表示だと文脈を読むには短すぎて足りない場合があるため、通常ではそれに加えて左右に更に長い範囲で補助表示が加わります。補助表示は青い文字で表示されます。この文字列は、検索結果には加わらないので、その後の統計などでも扱われません。表示項目の切り替えても表示は変更されず、文脈がそのまま表示されます。ウィンドウ下部の「補助表示」をオフにすると消すことができます。

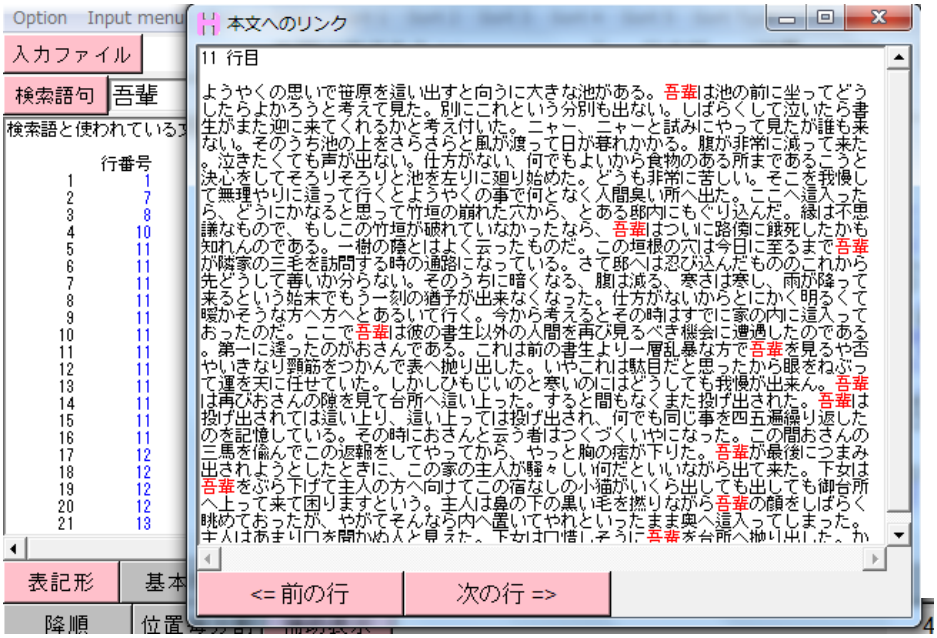
降順

表示順を逆にすることができます。



ウィンドウ下部の「降順」ボタンで、表示させる順を逆にできます。ソート指定をしていなければ単純に出現順の逆になり、ソートを指定している場合は、位置の中での条件が逆順になります。「語頭」や「語末」での 50 音順指定の場合は文字列順の逆順で、「頻度」の場合は少ないものが上に来ます。本来、数値で降順の場合は数字の大きい順になりますが、この KWIC の場合のみ少ない順が降順になります。

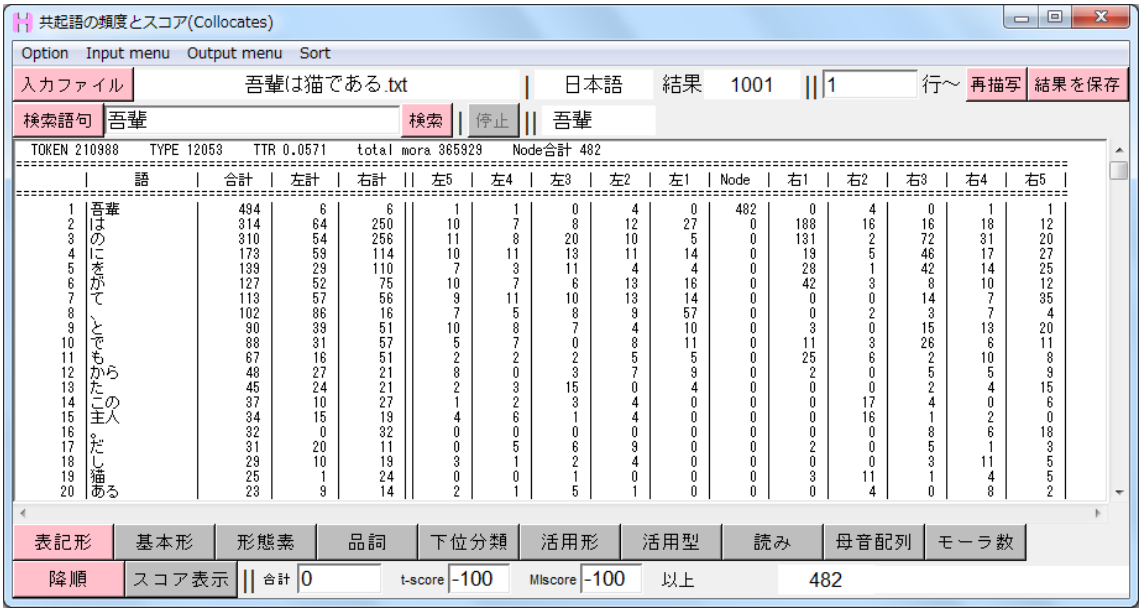
本文リンク



表示の右側の青い行番号をクリックで行の全文が読めます。
更に、「<=前の行」「次の行=>」で、前後の行の確認もできます。

共起語の頻度とスコア(Collocates)

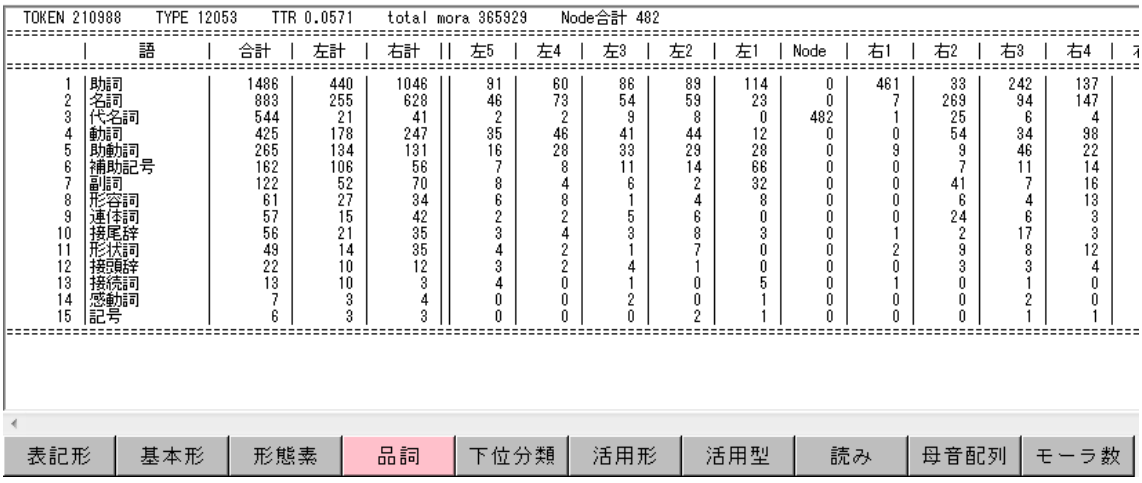
この処理は、検索結果の周囲の語の数を元に統計をする処理になります。



検索語と左右の指定幅の語、つまり KWIC の際に黒い文字で表示された語を全てまとめて集計した結果になります。

一番左が順位で、１つ右に語のリストが表示されます。その右から集計結果になり、集計結果の一番左から順に、範囲内の各語の合計数、Node より左側の合計数、Node より右側の合計数、それぞれの位置の中での合計数の順になります。位置の中での合計数では、例えば、第２位の「は」は、左１では 27 回、右３では 16 回出現していることが分かります。

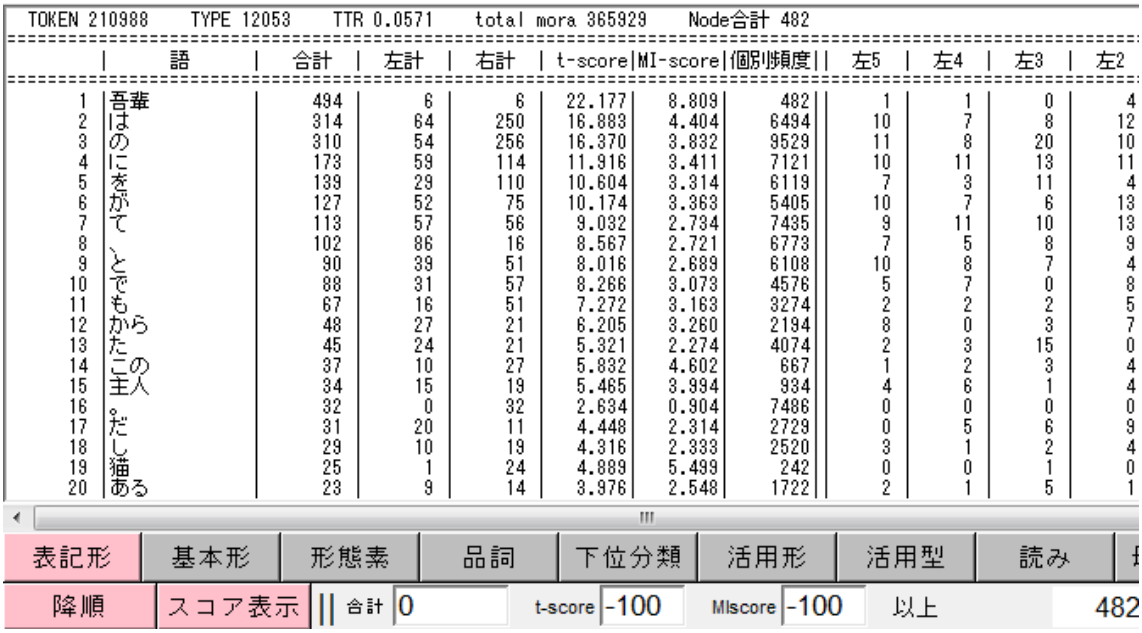
表示項目の変更



他の処理と同様に表示項目の変更ができます。数値を扱う処理では表示項目が変わるとその項目ごとの集計値に再計算され表示されます。

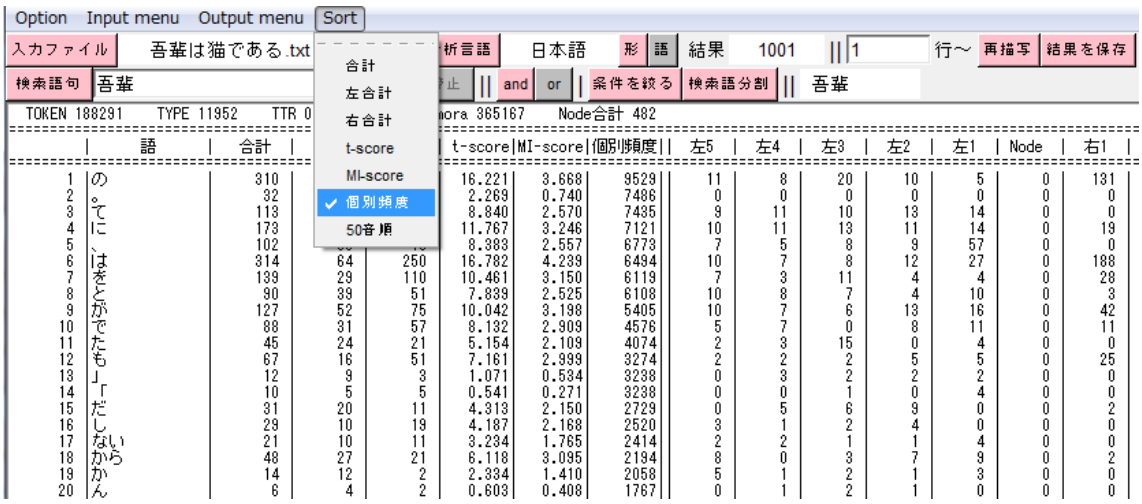
スコア表示

合計数などの数値に加えて、各種統計結果を表示に加えることができます。

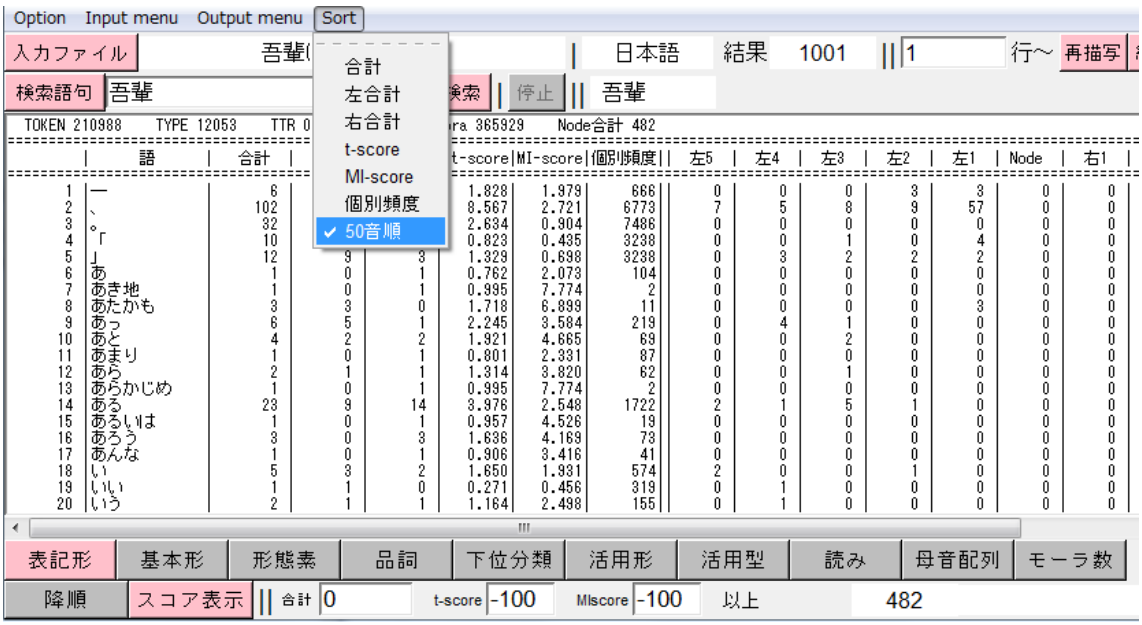


ウィンドウ下部の「スコア表示」ボタンで、統計値の表示ができます。加わる数値は、t-score、MIscore、個別頻度です。t-score は、検索語の周囲の各語と検索語の繋がりの強さを表す数値で、数値が高いほど結びつきが強いと言われます。範囲内の合計値、いわゆる共起語頻度が高いと高く出る傾向があります。MIscore も同様に検索語とその周囲の語の結びつきの強さを表す数値ですが、範囲内の合計値が低いと高く出る傾向があります。個別頻度は、各語のテキスト全体での頻度で、検索語との関係ではなくその語がテキストの中で何回出現しているかの数値です。

ソート



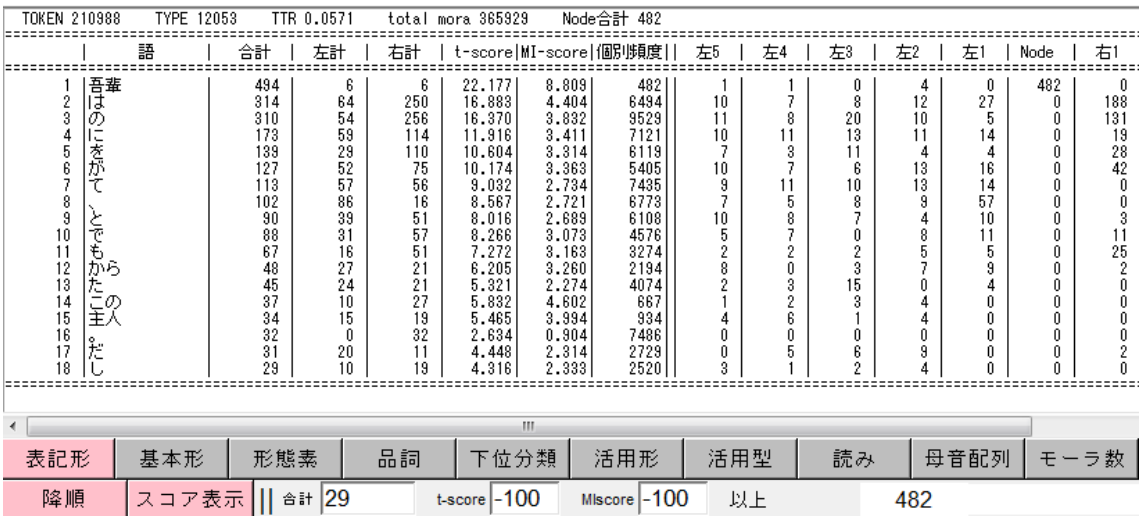
ウィンドウ上部のツールバーの Sort で、表示順を変えることができます。ソートの種類は「合計」「左合計」「右合計」「t-score」「MIsocre」「個別頻度」「50 音順」です。ソート条件を指定してから「再描写」ボタンでソートを行います。



「合計」から「個別頻度」までは各数値の順、「50 音順」は語の並び順でソートされます。

表示最低数の指定

いくつか数値で、表示させる最低数の指定ができます。



ウィンドウ下部の「以上」の前にある「合計」「t-score」「MIscoer」の数値を指定することで、指定した未満の数値の語は表示から外れます。数値は複数組み合わせて指定できます。

降順

降順指定のオンオフで、表示を逆順にできます。

ウィンドウ下部の「降順」ボタンで表示順を逆に変えられます。通常は降順が指定されていますが、ボタンを押すことで、指定がはずれ昇順になります。ソートを頻度などの数値で指定いている場合は、降順だと数値の大きい順、昇順だと数値の小さい順になります。

50 音順の降順

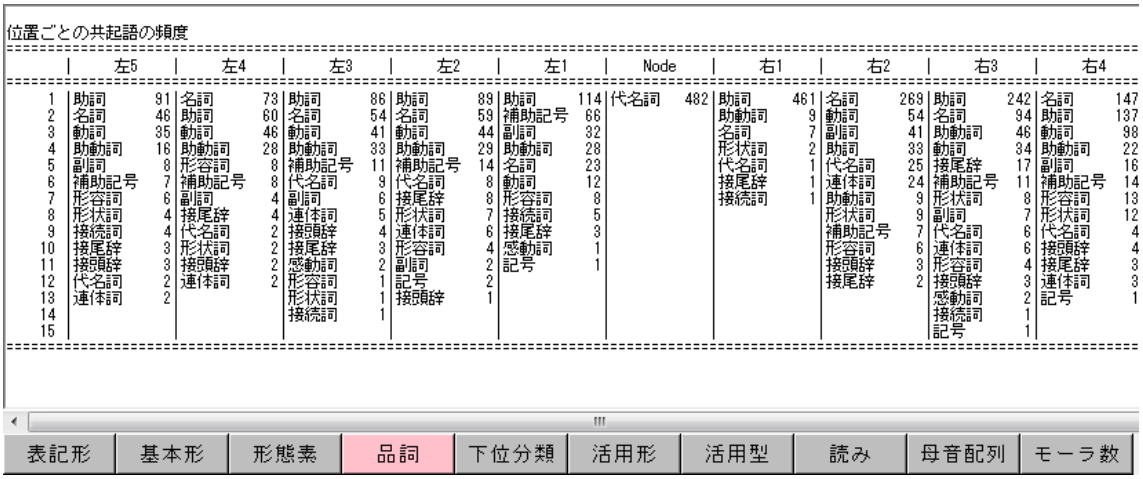
ソートを 50 音順にしているときは、降順だと「あ～ん」の順、昇順だと「ん～あ」」の順で表示されます。

位置ごとの共起語の頻度(Picture)

この処理は、それぞれの位置で、どの語が多いかを表示するものです。

Collocates の場合は、範囲内の語を全てまとめて集計し、位置ごとの頻度は各語がそれぞれの位置でいくつあるかを表示したもので、語ごとの数値でしたが、この処理ではそれぞれの位置が基準で、その位置に何の語がどれだけあるかを表示するものです。この例では右 1 では、「は」が 188 回、「の」が 131 回、「が」が 42 回出現しています。

表示項目の変更



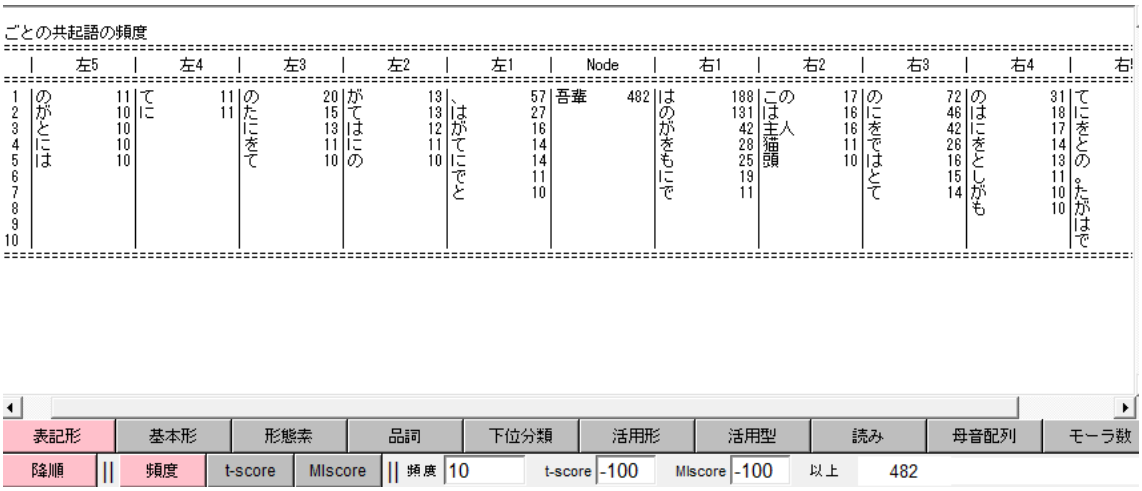
他の処理と同様に表示項目の変更ができます。これも数値を扱う処理ですので、表示項目が変わるとその項目ごとの集計値に再計算され表示されます。

算出する値の変更

算出する数値の変更もできます。

ウィンドウ下部の「頻度」「t-score」「MIscore」ボタンのどれか１つを選択し、算出する数値を選びます。通常では頻度が選ばれています。これは各位置での単純な出現頻度です。t-score と MIscore は、各位置でのスコアになりますので、Collocates で扱う、左右の範囲全体での数値とは異なります。

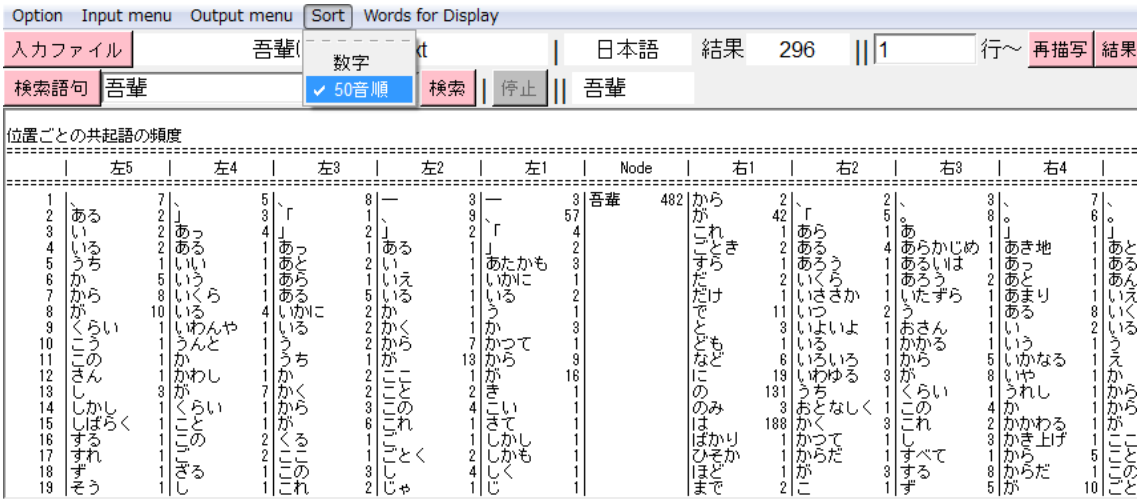
表示最低数の指定



ウィンドウ下部の「以上」の前にある「頻度」「t-score」「MIscoer」の数値を指定することで、指定した未満の数値の語は表示から外れます。実際に表示させている数値だけでなく、現在選択していない値も選択できます。数値は複数組み合わせて指定できます。

ソート

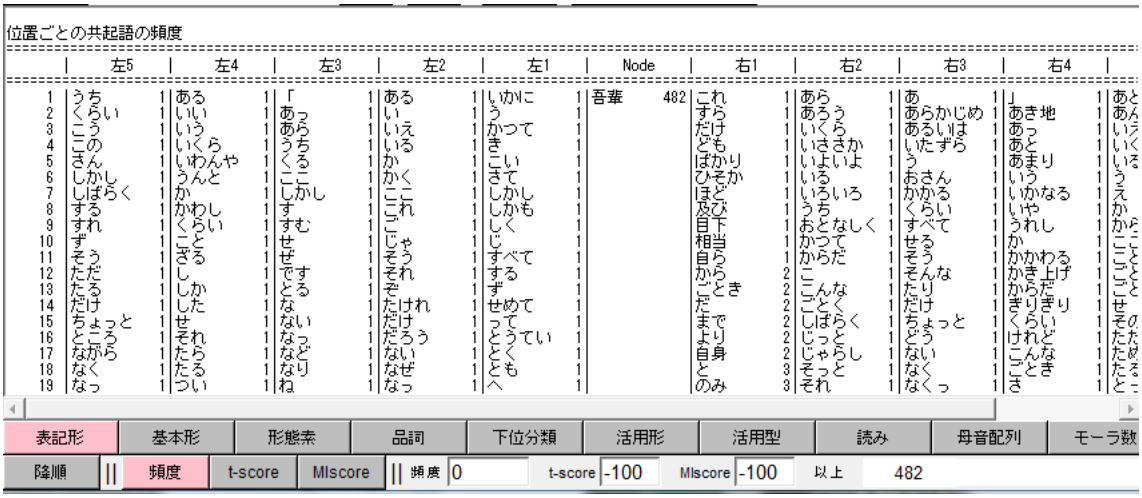
表示させる基準を変更できます。



ウィンドウ上部のツールバーの Sort で、表示順を変えることができます。ソートの種類は「数字」「50 音順」です。ソート条件を指定してから「再描写」ボタンでソートを行います。ソート条件を「数字」と指定している時は、算出する値が「頻度」「t-score」「MIscoer」のどれであってもその数値で並び変えをします。

降順

他の処理と同様に、降順の指定ができます。



ウィンドウ下部の「降順」ボタンのオンオフで、表示を昇順、降順に切り替えられます。数字でソートしている時は降順だと数字の大きい順、昇順だと数字の小さい順です。50 音順でソートしているときは、降順だと「あ～ん」の順、昇順だと「ん～あ」」の順で表示されます。

Words for Display

それぞれの位置で表示させる語の条件を絞ることができます。

ウィンドウ上部、ツールバーの「Words for Display」で現れるウィンドウで条件を指定します。条件は位置をまず決め、そこの位置での条件を絞る項目を選択し、その中身を文字列で指定します。この例では右 2 の位置、「品詞」の項目が、「名詞」のものとしています。

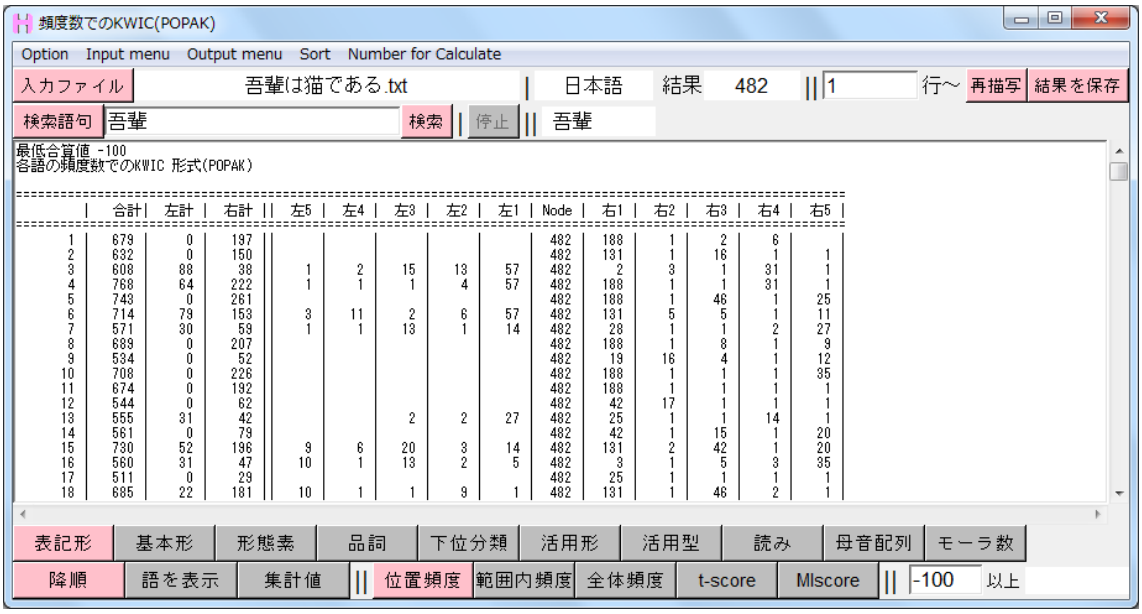
位置ごとの共起語の頻度

	左5		左4		左3		左2		左1		Node		右1		右2		右3		右4		右5	
1	の	11	て	11	の	20	が	13	、	57	吾輩	482	は	188	主人	16	の	72	の	31	て	35
2	のが	10	にと	11	たに	15	がて	13	、は	27			のが	131	猫	11	にを	46	はに	18	にを	27
3	とには	10	にとの	8	にを	13	はがて	12	がて	16			をも	42	頭	10	をでは	42	にをと	17	をとの	25
4	てから	10	のがでは	8	て	11	はにの	11	にでと	14			にでなど	28	顔	6	とて	26	をとしが	14	とのが	20
5	、をかで	10	主人	7	、は	10	にの	10	にでとから	14			との	25	例	4	金田	16	もある	13	。たがはで	20
6	主人	9	だあいる	7	とがだ	8	、だでから	9	から	11			猫から	19	金田	4	今	15	。たがはで	11	からもこの	18
7	なら	8	ーたもの	7	ら	8	ら	9	のも	10			ごだまよ	11	仕方	3	少々	14	からもこし	10	事猫	15
8	ある	7	をここず	6	のれ	7	もこし	8	ても	9			自己	11	尻尾	3	がす君から	8	、	10	、ない	12
9	いる	7	来	5	もの	7	もの	7	たい	9			すら	6	尻尾	3	。	8	。でない	8	へ	12
10	た	7	あと	5	を	6	を	7	を	9			先刻	6	眼	3	。	8	からとく	8		11
11		6		5	主人	6	主人	6	到底	8				5	耳	3	。	7	かごたなっ	7		9
12		6		5	吾輩	6	一	6	時	8				5	触	3	。	7	へ	7		8
13		6		5		6		6	一	8				5	鼻づら	3	。	7		7		8
14		6		5		6		6		8				5	人間	3	。	7		7		8
15		6		5		6		6		8				5	今朝	3	。	7		7		8
16		6		5		6		6		8				5	倩	3	。	7		7		8
17		6		5		6		6		8				5	先刻	3	。	7		7		8
18		6		5		6		6		8				5		3	。	7		7		8
19		6		5		6		6		8				5		3	。	7		7		8

その後、「再描写」を押すと、右 2 の位置は名詞の語のみが表示されるようになります。実際に表示させている項目だけでなく、全ての項目での選択ができます。

頻度数での KWIC(POPAK)

この処理は、KWIC でのそれぞれの語をそれぞれの位置で出てきた回数などに置き換えたもので、語を KWIC 形式の並びで Picture の位置ごとの頻度数で表示させたものです。

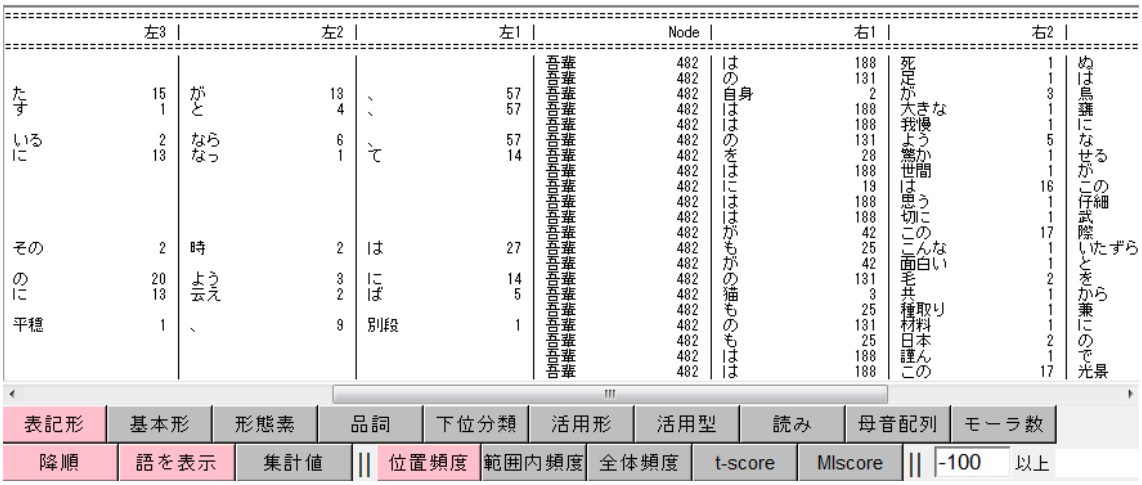


KWIC では、各語をその語自体や選択した項目で表示しますが、この処理では、語を数値に置き換えて表示します。数値は、基本では位置ごとの頻度数です。つまり、語を KWIC の並び、Picture の頻度で表示する処理です。

表の中で「左 5」～「右 5」の位置にある数値は KWIC での語を数値に置き換えたものです。KWIC での位置毎分割された語の表記が数値に置き換わったイメージです。

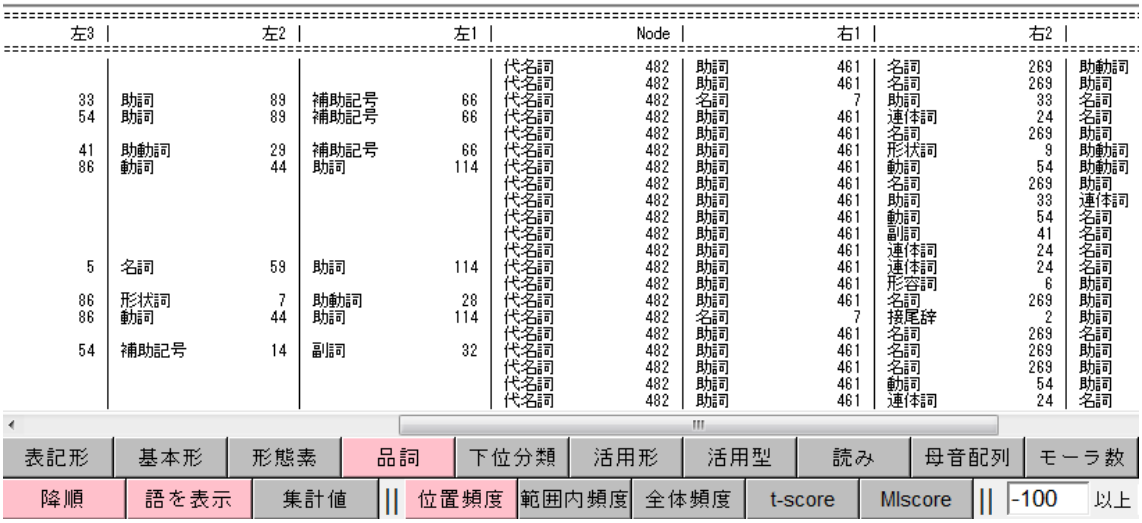
表の左側にある「合計」「左計」「右計」は、各行にある語の数値を足したものです。それぞれの行の中の語の数値を全部合計した数値が各行の合計や左計などになります。

語を表示



各数値がそれぞれ何の語のものなのかを表示します。ウィンドウ下部の「語を表示」ボタンでそれぞれの位置での語が表示されます。ここで表示される語のみを抜き出すと KWIC での表示と同じになります。

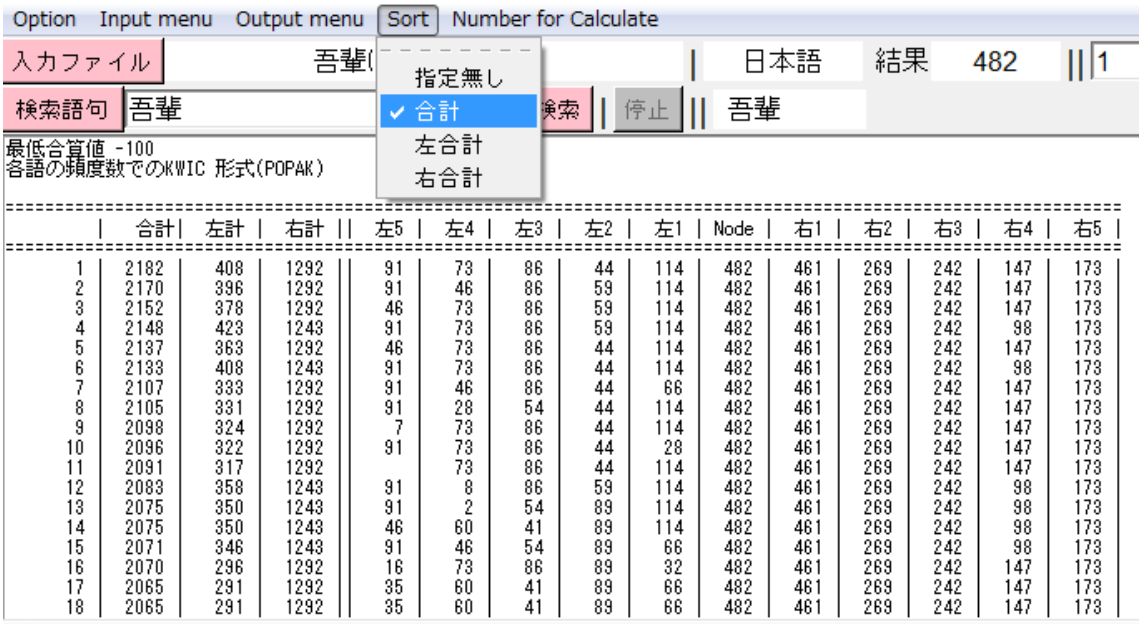
表示項目の変更



他の処理と同様に表示項目の変更ができます。これも数値を扱う処理ですので、表示項目が変わるとその項目ごとの集計値に再計算され表示されます。

ソート

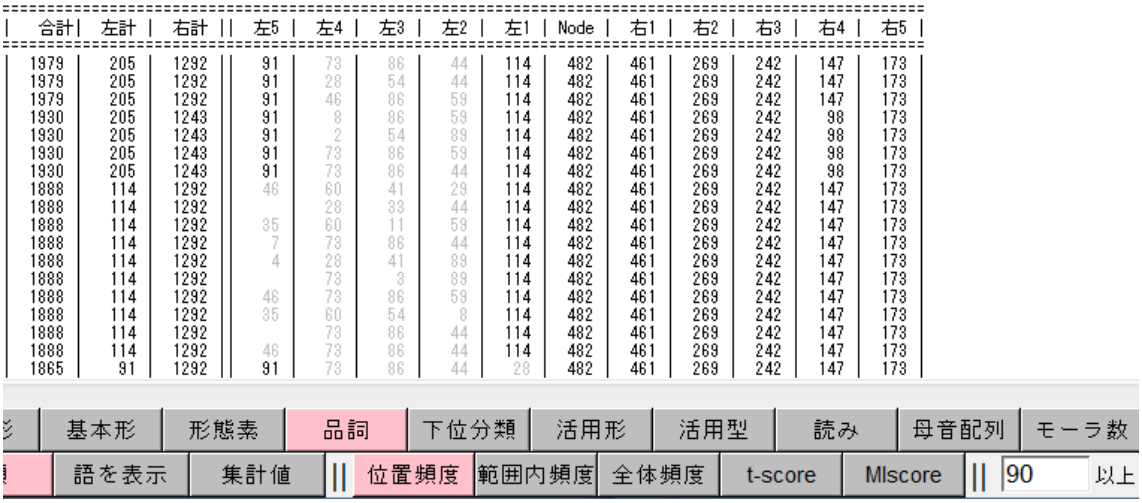
表示させる基準を変更できます。



ウィンドウ上部のツールバーの Sort で、表示順を変えます。ソートの種類は「指定無し」「合計」「左合計」「右合計」です。ソート条件を指定し「再描写」ボタンでソートします。

合計に加える最低数の指定

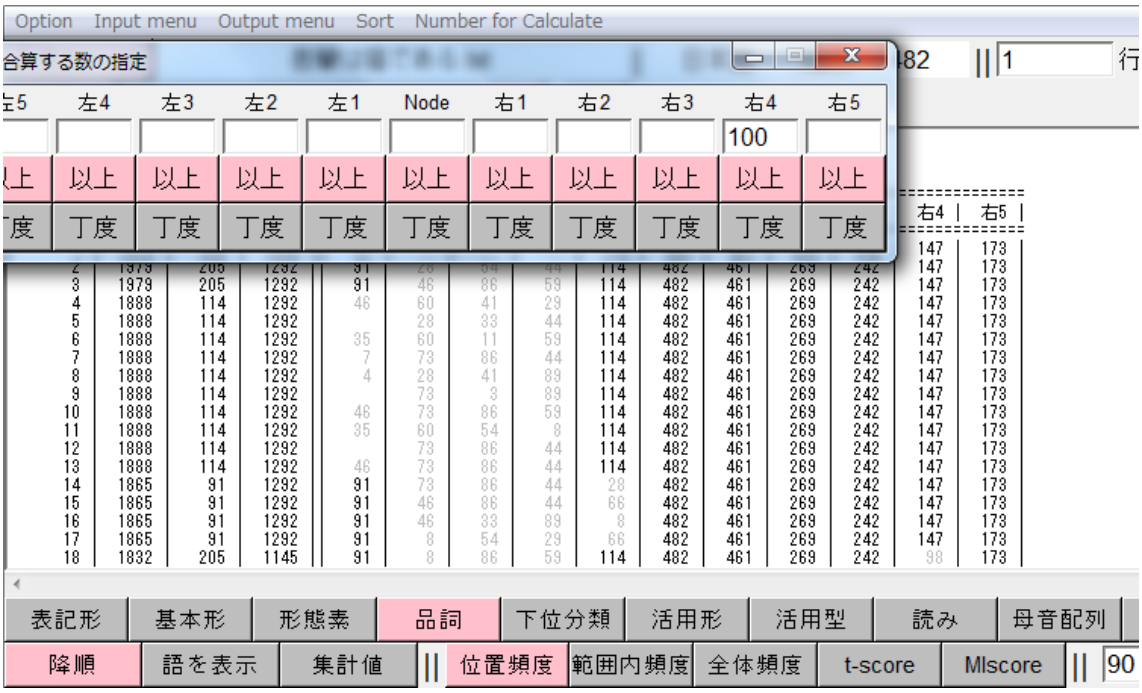
合計する数値の最低数を指定できます。



ウィンドウ下部の右の「以上」の前の数を指定し「再描写」で、それ以上の数値だけでの合計が行われます。これで数値の低い語、つまり使頻度の少ない語は合計値からはずれ、頻度の高い語だけが並ぶことになります。特に上位では同じ語が使われた結果が連続して並び、これを上から順に見ることで同じ語の並びがいくつ出現したのか実数や割合が分かります。

Number of Calculate

合計に加える最低数の指定で、特定の位置だけ条件を厳しくしたり緩くしたりできます。

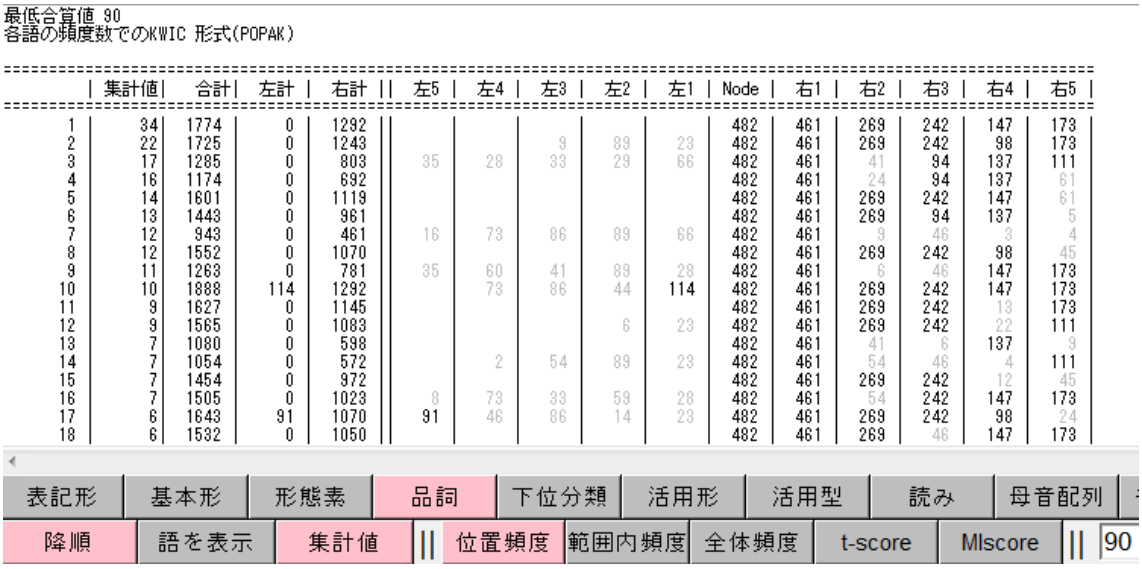


ウィンドウ上部、ツールバーの「Number of Calculate」で現れるウィンドウで条件を指定

します。条件はまず位置を決め、条件とする数値を指定し、合計するのはその数値以上か丁度かを選択し「再描写」を押します。これで指定した位置と数値の条件に沿うもののみが合計値に加わります。この例では右 4 が 100 以上としています。全体条件よりも個別の位置の条件が優先されますので、全体位置の条件が 90 以上の場合でも、個別の位置で 100 以上としたらその位置だけは 100 以上、30 以上とされたら 30 以上が合計値に加わります。

集計値

合計値が全く同じ行をまとめて数えることができます。

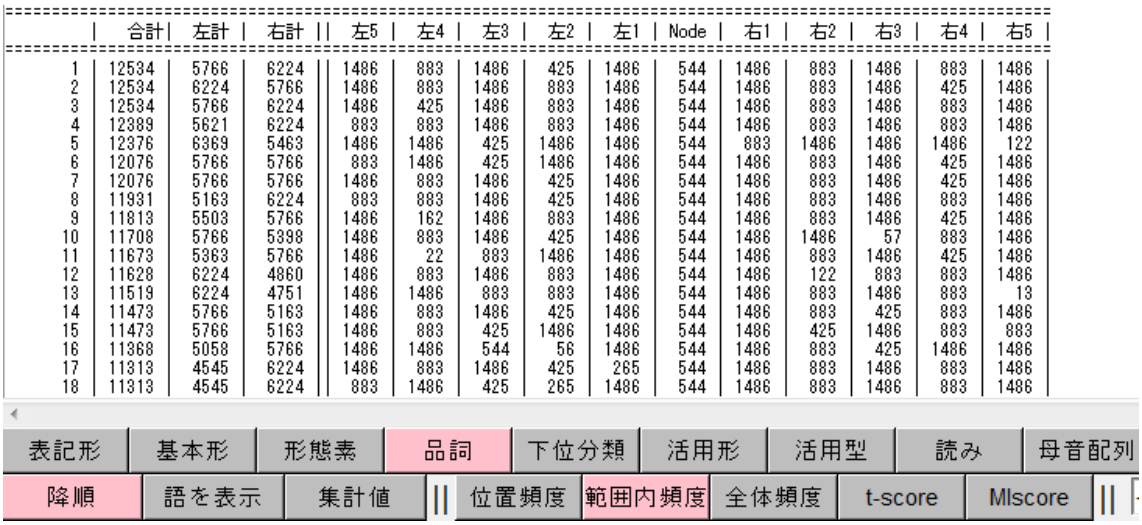


	集計値	合計	左計	右計	左5	左4	左3	左2	左1	Node	右1	右2	右3	右4	右5
1	34	1774	0	1292						482	461	269	242	147	173
2	22	1725	0	1243			9	89	23	482	461	269	242	98	173
3	17	1285	0	803	35	28	33	29	66	482	461	41	94	137	111
4	16	1174	0	692						482	461	24	94	137	61
5	14	1601	0	1119						482	461	269	242	147	61
6	13	1443	0	961						482	461	269	94	137	5
7	12	943	0	461	16	73	86	89	66	482	461	9	46	3	4
8	12	1552	0	1070						482	461	269	242	98	45
9	11	1263	0	781	35	60	41	89	28	482	461	6	45	147	173
10	10	1888	114	1292		73	86	44	114	482	461	269	242	147	173
11	9	1627	0	1145						482	461	269	242	13	173
12	9	1565	0	1083				6	23	482	461	269	242	22	111
13	7	1080	0	598						482	461	41	6	137	9
14	7	1054	0	572		2	54	89	23	482	461	54	46	4	111
15	7	1454	0	972						482	461	269	242	12	45
16	7	1505	0	1023	8	73	33	59	28	482	461	54	242	147	173
17	6	1643	91	1070	91	46	86	14	23	482	461	269	242	98	24
18	6	1532	0	1050						482	461	269	46	147	173

ウィンドウ下部の「集計値」ボタンで、合計値の値が全く同じ行の数を数えて表の左に表示します。これで、どの語の並びがいくつ実際に出現したのかの実数が分かります。

算出する値の変更

算出する値の変更もできます。

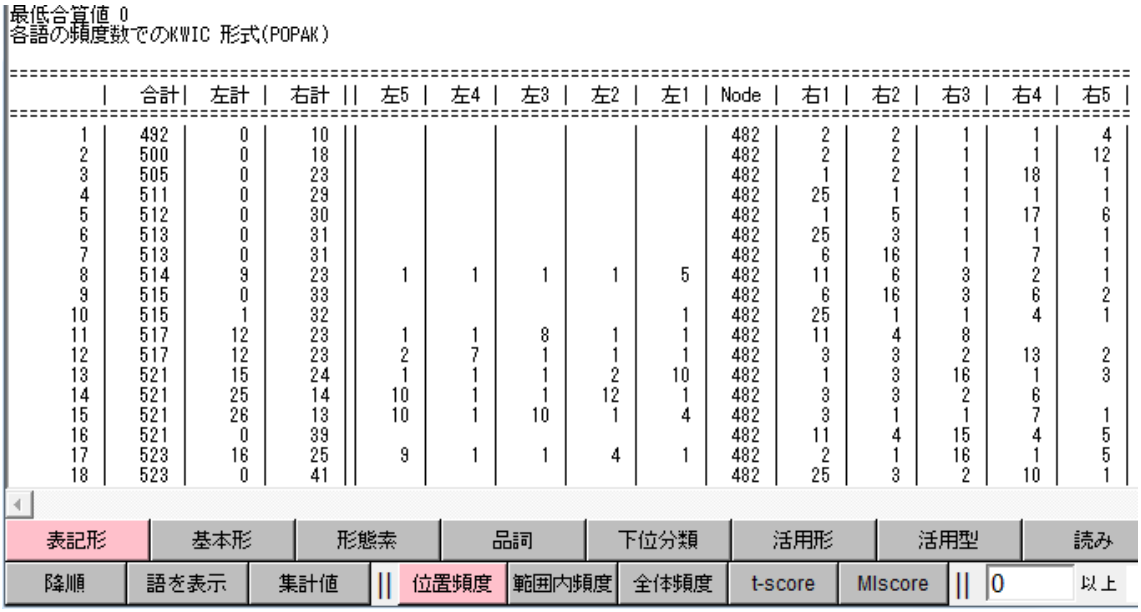


	合計	左計	右計	左5	左4	左3	左2	左1	Node	右1	右2	右3	右4	右5
1	12534	5766	6224	1486	883	1486	425	1486	544	1486	883	1486	883	1486
2	12534	6224	5766	1486	883	1486	883	1486	544	1486	883	1486	425	1486
3	12534	5766	6224	1486	425	1486	883	1486	544	1486	883	1486	883	1486
4	12389	5621	6224	883	883	1486	883	1486	544	1486	883	1486	883	1486
5	12376	6369	5463	1486	1486	425	1486	1486	544	883	1486	1486	1486	122
6	12076	5766	5766	883	1486	425	1486	1486	544	1486	883	1486	425	1486
7	12076	5766	5766	1486	883	1486	425	1486	544	1486	883	1486	425	1486
8	11931	5163	6224	883	883	1486	425	1486	544	1486	883	1486	883	1486
9	11813	5503	5766	1486	162	1486	883	1486	544	1486	883	1486	425	1486
10	11708	5766	5398	1486	883	1486	425	1486	544	1486	1486	57	883	1486
11	11673	5363	5766	1486	22	883	1486	1486	544	1486	883	1486	425	1486
12	11628	6224	4860	1486	883	1486	883	1486	544	1486	122	883	883	1486
13	11519	6224	4751	1486	1486	883	883	1486	544	1486	883	1486	883	13
14	11473	5766	5163	1486	883	1486	425	1486	544	1486	883	425	883	1486
15	11473	5766	5163	1486	883	425	1486	1486	544	1486	425	1486	883	883
16	11368	5058	5766	1486	1486	544	56	1486	544	1486	883	425	1486	1486
17	11313	4545	6224	1486	883	1486	425	265	544	1486	883	1486	883	1486
18	11313	4545	6224	883	1486	425	265	1486	544	1486	883	1486	883	1486

ウィンドウ下部の「位置頻度」「範囲内頻度」「全体頻度」「t-score」「MIscore」ボタンのどれか１つを選択し、算出する数値を選びます。通常では位置頻度が選ばれています。これは各位置での単純な出現頻度です。範囲内頻度は、検索時の左右の取得幅、通常では左 5 語、右 5 語の範囲の中での各語の出現頻度になります。全体頻度は、各語がテキスト全文で使われた頻度です。t-score と MIscore は、各位置でのスコアです。Picture の時と同じ数値になります。

降順

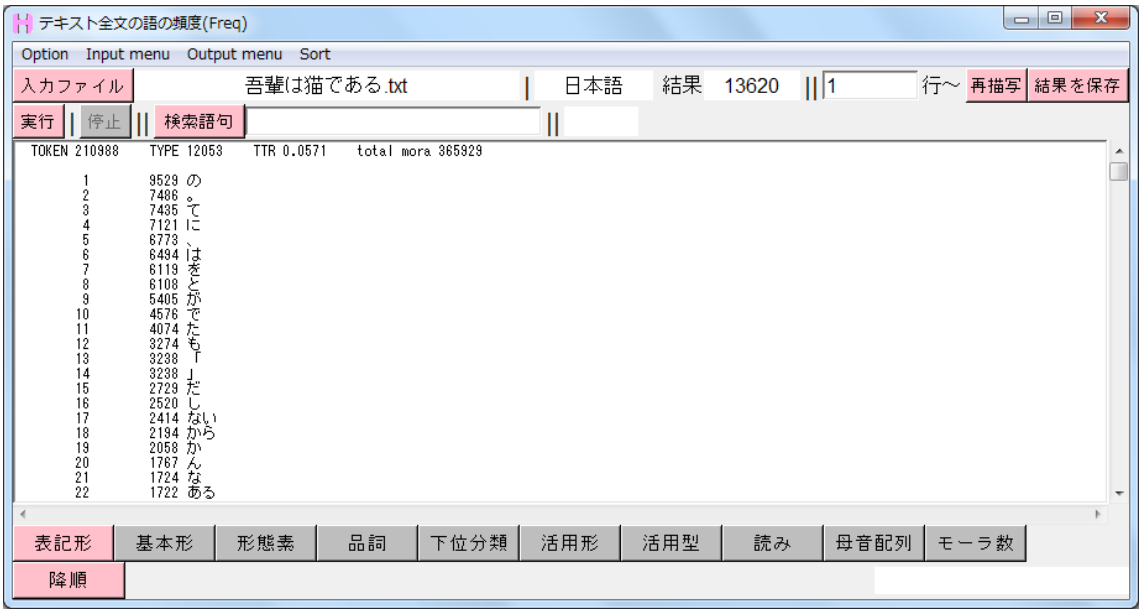
他の処理と同様に、降順の指定ができます。



ウィンドウ下部の「降順」ボタンのオンオフで、表示を昇順、降順に切り替えられます。ソート条件で指定した値の順になります。降順だと数字の大きい順、昇順だと数字の小さい順です。ソート条件が「指定無し」の場合は、出現順の逆になります。

テキスト全体の語の頻度(Freq)

　ここからは検索語句と関係なくテキスト全文の語を対象に統計を行う処理になります。この処理はテキスト全文の中の語の数を数えるもので、各語の使用数を数えて表示します。



「実行」を押すと、そのまま集計結果が表示されます。ごく単純な単語の出現頻度表です。

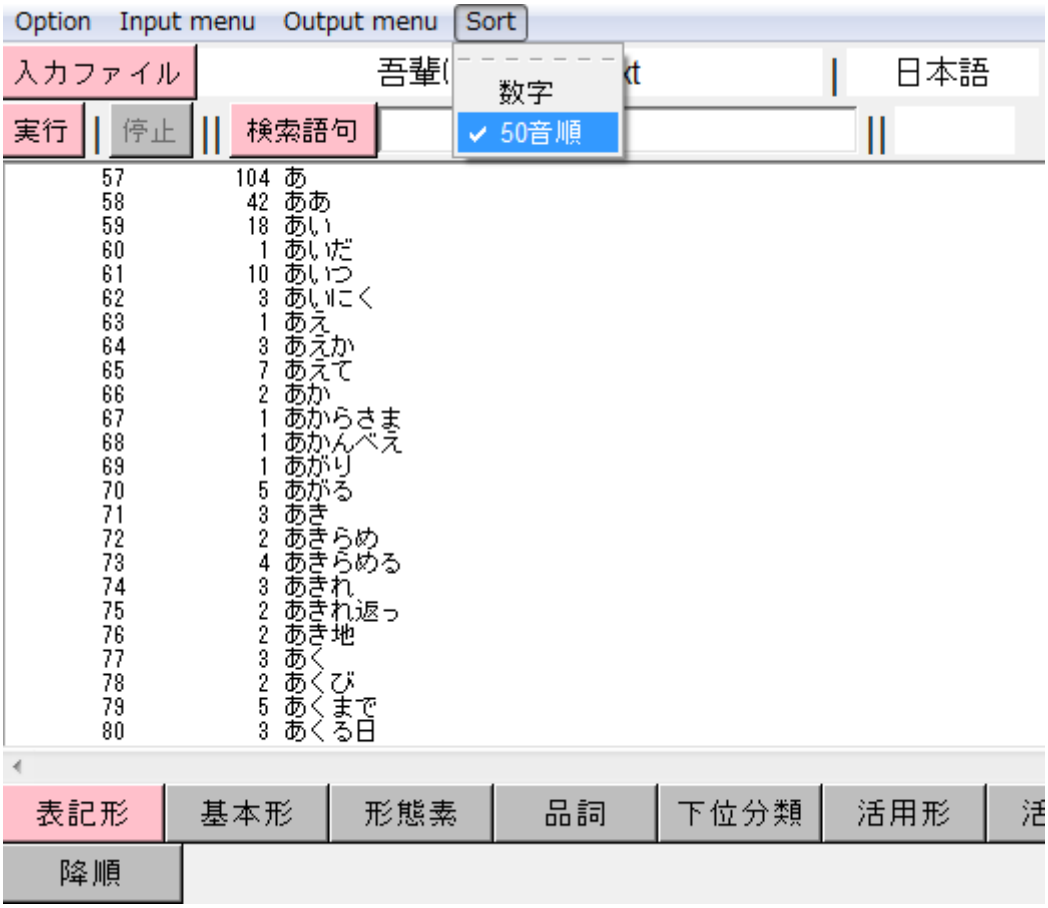
表示項目の変更

他の処理と同様に表示項目の変更ができます。これも数値を扱う処理ですので、表示項目

が変わるとその項目ごとの集計値に再計算され表示されます。

ソート

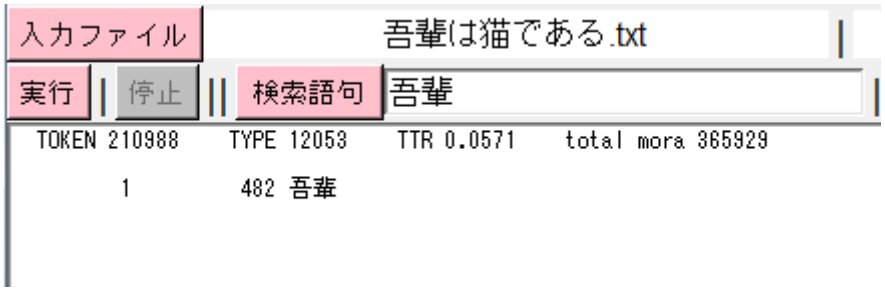
表示させる基準を変更できます。



ウィンドウ上部のツールバーの Sort で、表示順を変えることができます。ソートの種類は「数字」「50 音順」です。ソート条件を指定してから「再描写」ボタンでソートを行います。

検索

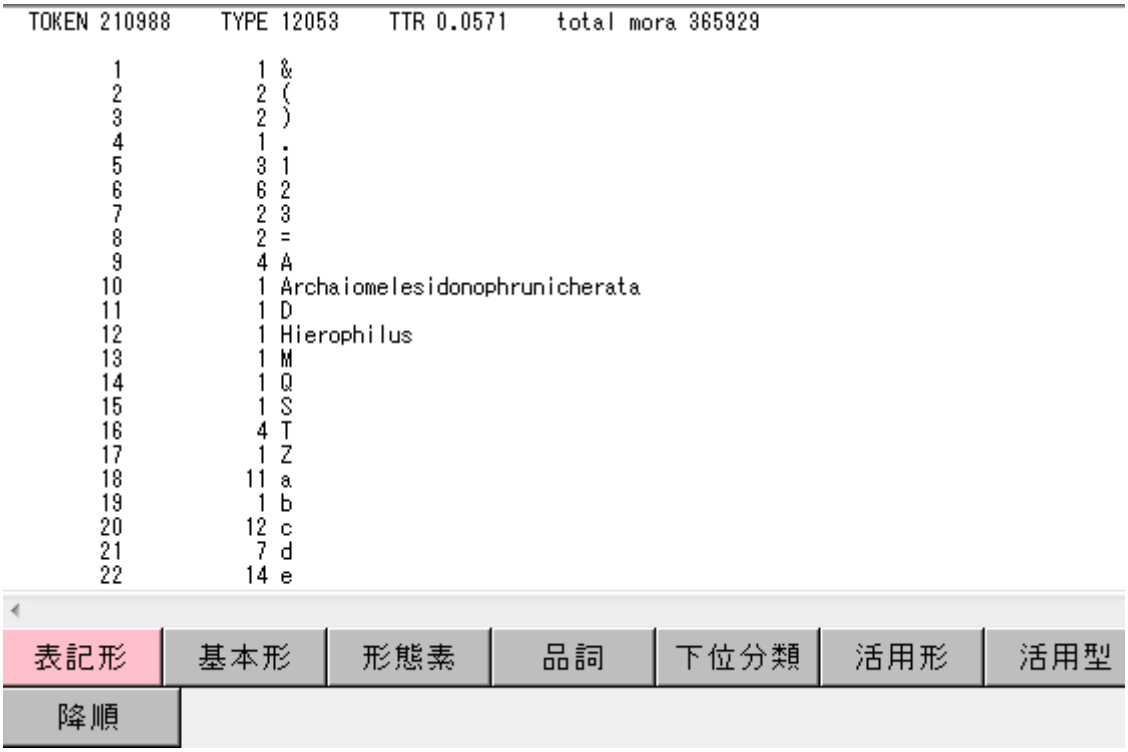
個別の語を指定して表示させることができます。



検索語句を指定してから「実行」を押すと、指定した語句のみが表示されます。決まった語の頻度を探したいときに使います。

降順

他の処理と同様に、降順の指定ができます。



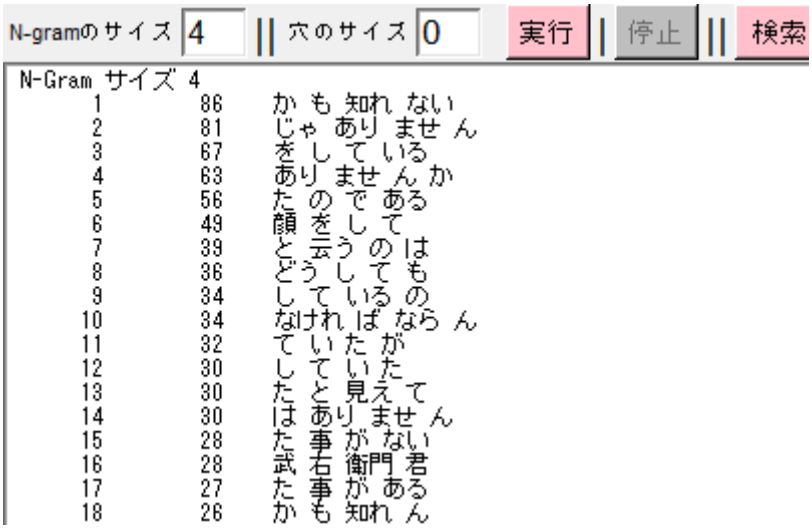
ウィンドウ下部の「降順」ボタンのオンオフで、表示を昇順、降順に切り替えられます。数字でソートしている時は降順だと数字の大きい順、昇順だと数字の小さい順です。50 音順でソートしているときは、降順だと「あ～ん」の順、昇順だと「ん～あ」」の順で表示されます。

同じ並びの語の数(N-gram)

この処理では、本文の中に出てきた語のどの並びがいくつあるかを表示します。

例えば「私はその人を常に先生と呼んでいた。」という文を形態素に分けると「私 は その 人 を 常 に 先生 と 呼ん で い た 。」で、これを N-gram サイズごとに順に保管していきます。サイズを２の場合、「私 は」「は その」「その 人」「人 を」「を 常」「常 に」「に 先生」「先生 と」「と 呼ん」「呼ん で」「で い」「い た」「た 。」となり、このうち文の区切りに指定してある「。」を含まない結果のみを保存し、同じ形の語の並びの数を数えます。これによりどの語の並びがいくつあるか分かります。「実行」を押すと集計が開始されます。

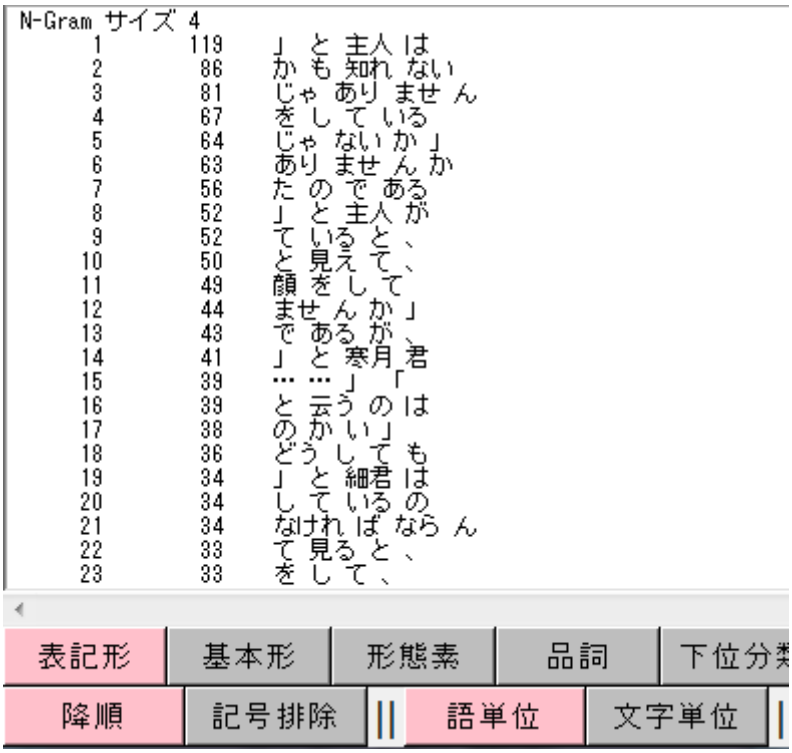
Ngram のサイズ



N-gram のサイズの変更もできます。ウィンドウ上部の Ngram のサイズで指定します。

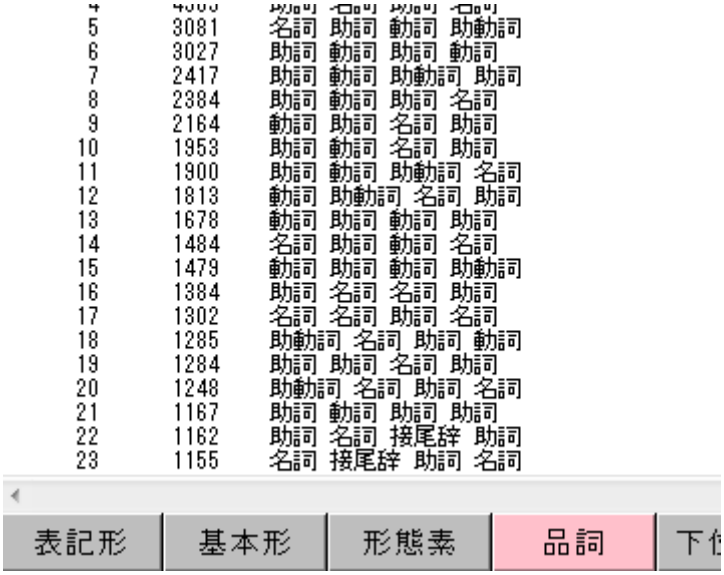
記号排除

作成される Ngram の中に記号が含まれる結果を省くことができます。



ウィンドウ下部の「記号排除」ボタンで、Ngram の集計に記号を含む結果を入れるか選択できます。通常は記号排除はオンになっています。記号が含まれる結果を集計から省くだけなので、元データの記号を消して語だけにして Ngram を生成するわけではありません。

表示項目の変更

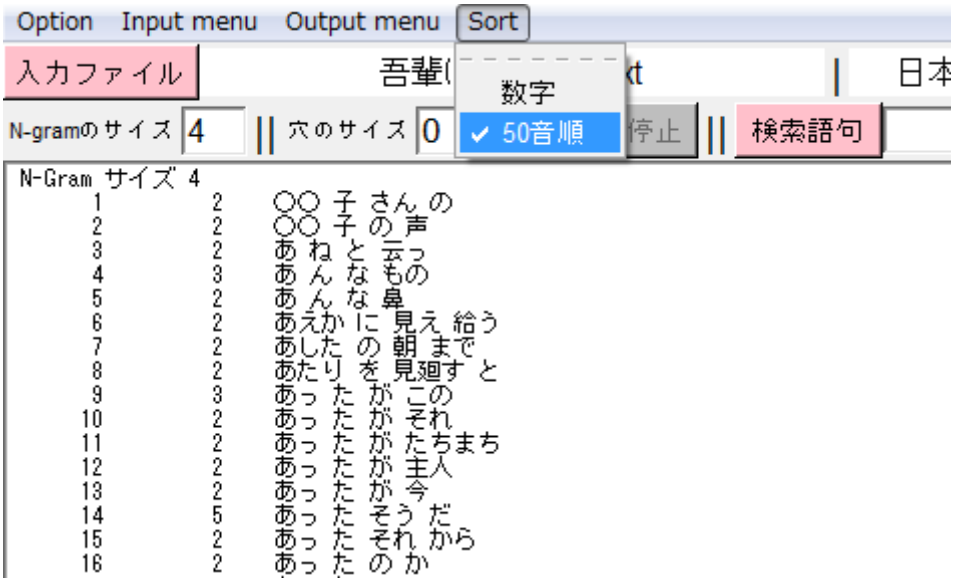


他の処理と同様に表示項目の変更ができます。これも数値を扱う処理ですので、表示項目

が変わるとその項目ごとの集計値に再計算され表示されます。選択した項目の要素の並びでの集計になります。

ソート

表示させる基準を変更できます。



ウィンドウ上部のツールバーの Sort で、表示順を変えることができます。ソートの種類は「数字」「50 音順」です。ソート条件を指定してから「再描写」ボタンでソートを行います。

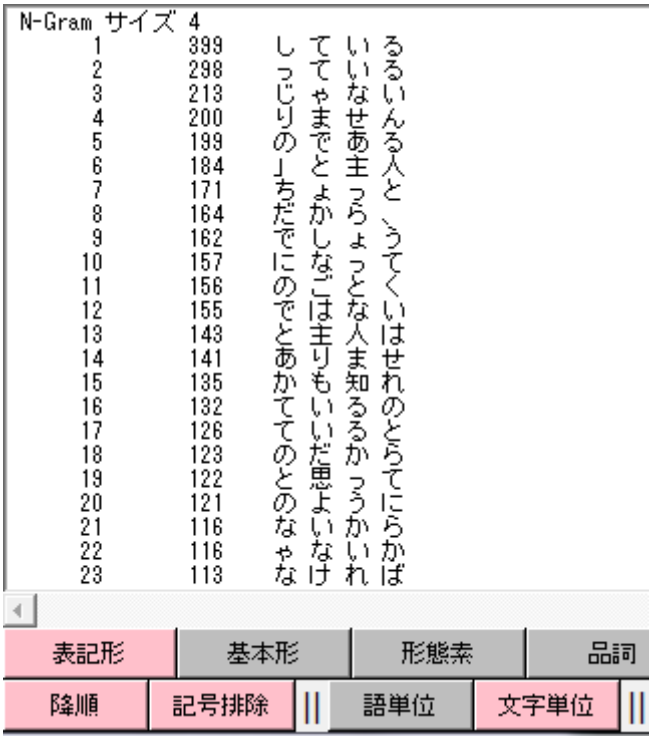
表示最低数の指定

1 63 あり ませ ん か
2 86 か も 知れ ない
3 30 し て い た
4 34 し て いる の
5 81 じゃ あり ませ ん
6 30 た と 見え て
7 56 た の で ある
8 32 て い た が
9 39 と 云う の は
10 36 どう し て も
11 34 なけれ ば なら ん
12 30 は あり ませ ん
13 67 を し て いる
14 49 顔 を し て

ウィンドウ下部の「以上」の前の数を指定し「再描写」で、それ以上の数値の結果だけが表示されます。これを多めに指定することで処理が高速になり、またテキストサイズが大きい場合パソコンのメモリ不足で Ngram の算出ができないことがありますが、負荷を軽減し算出が可能になることがあります。

作成する Ngram の単位

Ngram を作成する単位の変更ができます。



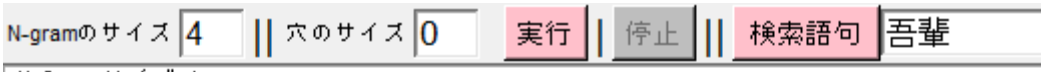
ウィンドウ下部の「語単位」「文字単位」を切り換えることで作成する Ngram の単位を変えられます。通常では語単位になっています。これを文字単位にすると、Ngram の集計が１文字ずつの並びで算出されます。

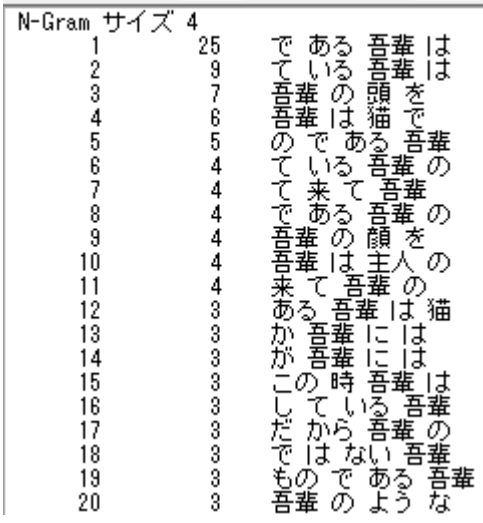
降順

他の処理と同様に、降順の指定ができます。

ウィンドウ下部の「降順」ボタンのオンオフで、表示を昇順、降順に切り替えられます。数字でソートしている時は降順だと数字の大きい順、昇順だと数字の小さい順です。50 音順でソートしているときは、降順だと「あ～ん」の順、昇順だと「ん～あ」」の順で表示されます。

検索

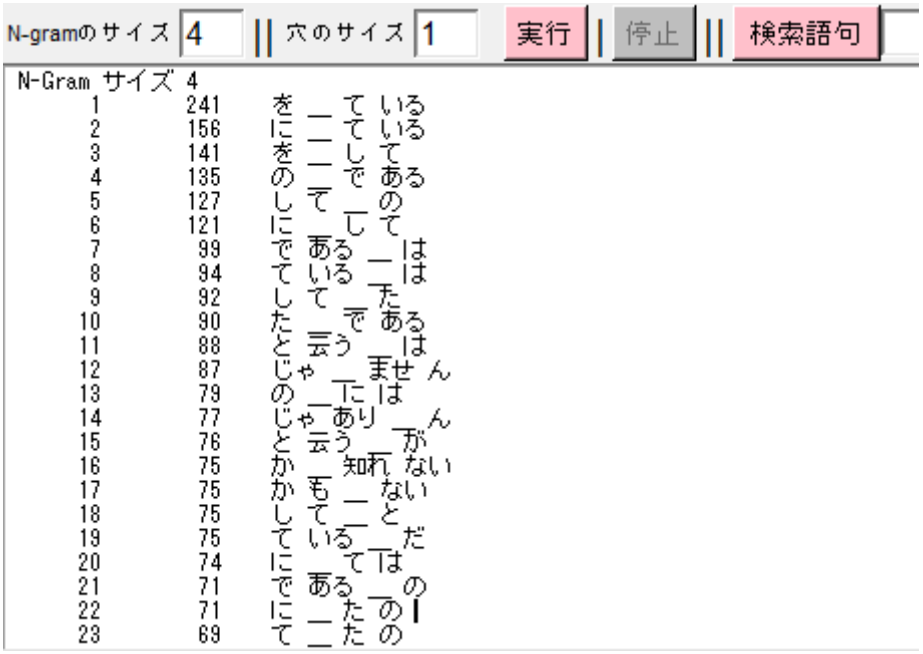




検索語句を指定してから「実行」で、指定した語のある結果のみを表示することができます。特定の語の使用結果のみを見るのに使います。

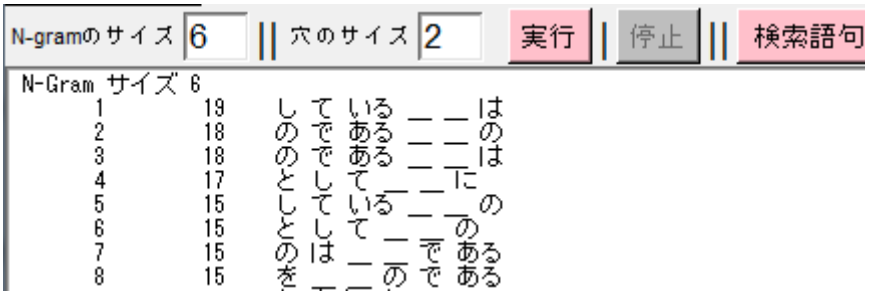
穴空きの Ngram

Ngram が完全に連続していなくても集計を行うことができます。



ウィンドウ上部の「穴のサイズ」に数字を指定するとその数字の分だけ Ngram に穴を空けることができます。Ngram は通常、完全に同じ語の並びの数だけを集計しますが、これだと、例えば４つの語のうち、３つは非常に出現頻度が高くてもそのうちの１つが頻度の低い結果の場合にその結果は少ないものとして埋もれてしまいます。そこで、Ngram の語の並びのうち、何が来ても良い箇所を１つ設けます。すると、そこに来る語は何でもいいが、その周囲の語は決まった同じ語の並びの結果が浮き出てきます。これにより、語の使われるフレームが見つかります。穴は端の語には来ないようになっていますので、両端を抜かした間の語で空くようになっています。

これを本ソフトでは「N－mgram（Nマイナスmグラム）」と呼びます。



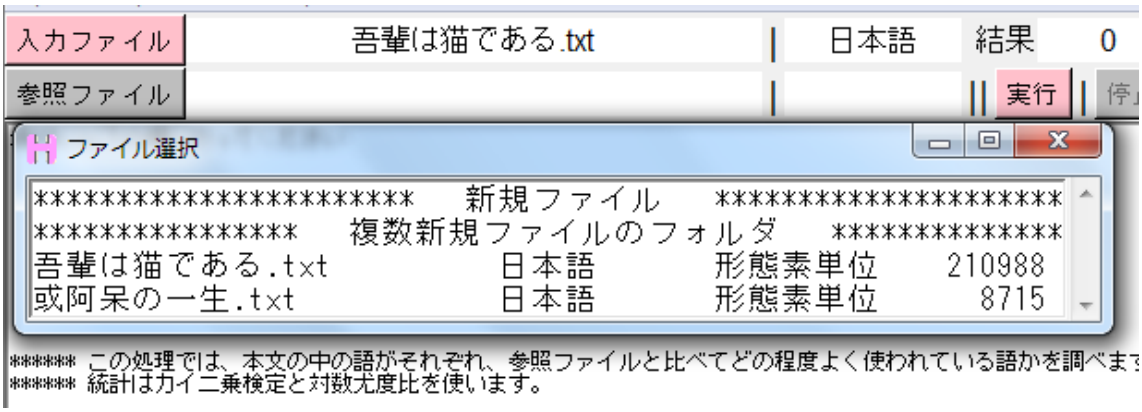
穴のサイズを２以上に指定した際も空く箇所は１箇所で、そこに２語連続の穴が開きます。

特徴的な語(Keyness)

この処理では２つのファイルを比較しメインファイルに特徴的に現れる語を表示します。

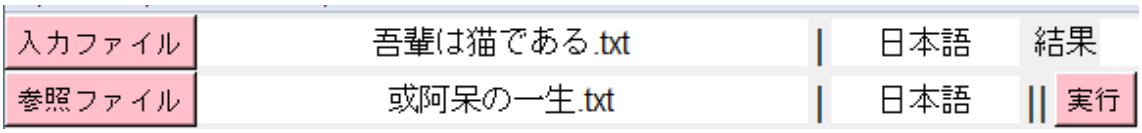
他の処理で使用するファイルと同じ、「入力ファイル」で指定したファイルがメインファイルとなり、もうひとつ指定する「参照ファイル」を比較対象のファイルとして使用します。

参照ファイル



参照ファイルの指定は、ウィンドウ上部の「入力ファイル」の並びにある「参照ファイル」のボタンで指定します。既に複数のテキストファイルが整形されていれば整形されたファイルのリストから選択できます。参照ファイルに指定するテキストファイルがまだ整形されていなければ、新規ファイルから選択します。その際は「入力ファイル」の時と同様に分析ファイルの設定をし、整形が終わるまで待ちます。

２つのファイルを選択してから「実行」で、結果が表示されます。



対数尤度比とカイ二乗

メインファイルに特徴的な語を算出する指標は対数尤度比とカイ二乗値の２つがあります。

	語	尤度比	カイ二乗	メイン	%	参照	%
				210988		8715	
1	で	201.44	133.14	4576	2.169	31	0.356
2	がとん	161.90	117.33	5405	2.562	62	0.711
3	とん	133.79	102.58	6108	2.895	92	1.056
4	いる	107.13	62.69	1767	0.837	5	0.057
5	事	101.89	51.09	1255	0.595	0	0.000
6	から	98.72	49.47	1216	0.576	0	0.000
7	だある	95.39	62.45	2194	1.040	15	0.172
8	な	95.33	66.13	2729	1.293	26	0.298
9	云う	81.32	51.75	1722	0.816	10	0.115
10	主人	77.69	50.11	1724	0.817	11	0.126
11	ない	70.24	34.87	866	0.410	0	0.000
12	御	66.54	35.71	934	0.443	1	0.011
13	です	59.15	43.84	2414	1.144	33	0.379
14	でよう	57.81	28.51	713	0.338	0	0.000
15	よて	48.58	29.97	975	0.462	5	0.057
16	そう	46.93	25.38	685	0.325	1	0.011
17	なる	46.88	40.93	7435	3.524	195	2.238
18	鼻	45.96	22.45	567	0.269	0	0.000
19	たら	45.23	22.08	558	0.264	0	0.000

表の左側が対数尤度比で、右がカイ二乗値です。

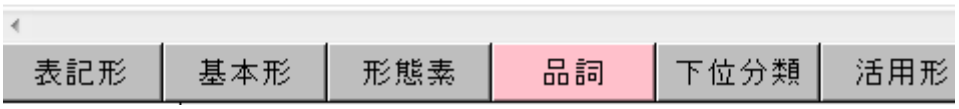
54	とく	15.79	7.05	195	0.092	0	0.000
55	てる	14.75	8.10	265	0.126	1	0.011
56	ちょっと	14.66	6.48	181	0.086	0	0.000
57	もので	14.58	6.44	180	0.085	0	0.000
58	まば	13.85	6.07	171	0.081	0	0.000
59	でしょう	13.71	10.67	1012	0.480	20	0.229
60	する	13.42	7.98	304	0.144	2	0.023
61	訳	13.37	9.57	608	0.288	9	0.103
62	ね	13.04	5.65	161	0.076	0	0.000
63	気	12.95	10.20	1061	0.503	22	0.252
64	いう	12.88	5.57	159	0.075	0	0.000
65	来	12.73	9.34	671	0.318	11	0.126
66	た	12.56	6.88	235	0.111	1	0.011
67	だ	12.55	5.41	155	0.073	0	0.000
68	私	12.36	8.04	384	0.182	4	0.046
69	聞い	12.31	5.28	152	0.072	0	0.000
70	思う	12.23	5.24	151	0.072	0	0.000
71	行っ	12.15	5.20	150	0.071	0	0.000
72	おい	12.07	5.16	149	0.071	0	0.000
73	よく	11.66	4.96	144	0.068	0	0.000
74	大	10.85	4.55	134	0.064	0	0.000
75	東風	10.45	4.34	129	0.061	0	0.000
76	人	10.40	5.68	205	0.097	1	0.011
77	東	9.55	3.89	118	0.056	0	0.000

表の中の、赤い数値は 0.1%水準、ピンクの数値は 1%水準、黄色い数値は 5%水準で有意差のある語です。２つの指標のそれぞれで水準ごとに色が付けられます。これらの数値は、

メインファイルにより特徴的に現れた語を表す指標になります。
その右がメインファイルでの各語の個別頻度、各語がメインファイルの総語数に占める割合を%で表した数値です。またその右が参照ファイルでの同じ語の個別頻度と割合です。

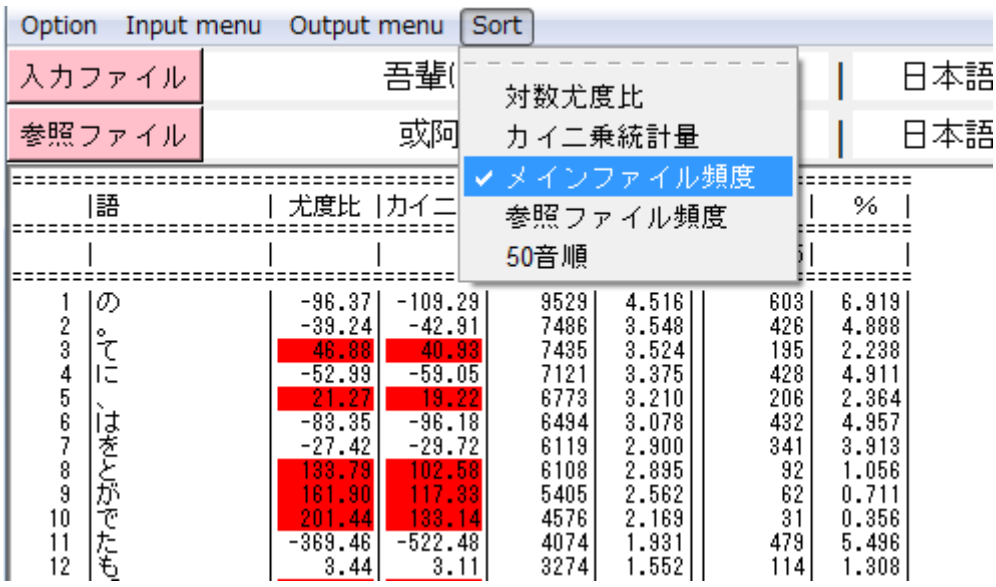
表示項目の変更

	語	尤度比	カイ二乗	メイン	%	参照	%
				210988		8715	
1	副詞	72.29	59.94	6226	2.951	133	1.526
2	動詞	40.35	38.35	29917	14.179	1030	11.819
3	形状詞	34.28	27.70	2661	1.261	54	0.620
4	助詞	20.83	20.44	63475	30.085	2424	27.814
5	接頭辞	20.55	16.45	1652	0.783	34	0.390
6	感動詞	14.40	11.54	1297	0.615	28	0.321
7	助動詞	5.02	4.84	20744	9.832	794	9.111
8	形容詞	3.68	3.38	4393	2.082	156	1.790
9	補助記号	-0.10	-0.09	22216	10.530	927	10.637
10	連体詞	-10.11	-10.80	2109	1.000	119	1.365
11	接尾辞	-12.34	-12.99	4881	2.313	254	2.915
12	名詞	-21.84	-22.17	45765	21.691	2076	23.821
13	接続詞	-46.83	-63.91	624	0.296	69	0.792
14	記号	-128.40	-217.24	481	0.228	92	1.056
15	代名詞	-394.48	-553.79	4547	2.155	525	6.024



他の処理と同様に表示項目の変更ができます。これも数値を扱う処理ですので、表示項目が変わるとその項目ごとの集計値に再計算され表示されます。

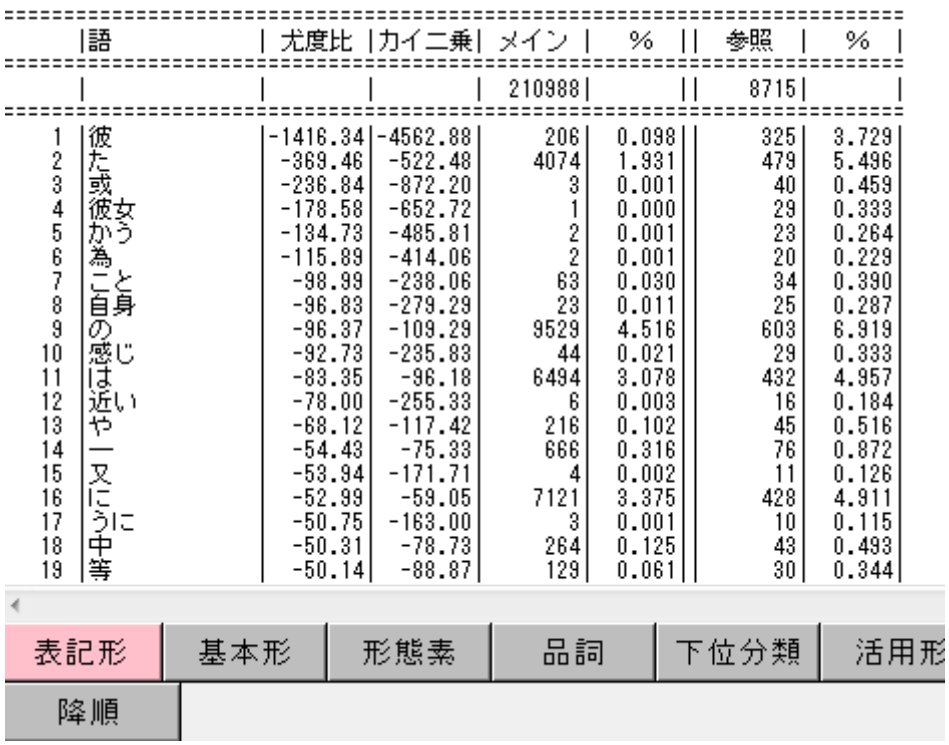
ソート



ウィンドウ上部のツールバーの Sort で、表示順を変えることができます。ソートの種類は「対数尤度比」「カイ二乗統計量」「メインファイル頻度」「参照ファイル頻度」「50 音順」です。ソート条件を指定してから「再描写」ボタンでソートを行います。

降順

他の処理と同様に、降順の指定ができます。

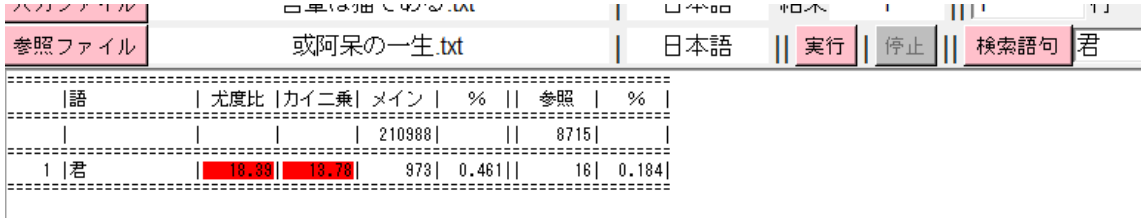


	語	尤度比	カイ二乗	メイン	%	参照	%
				210988		8715	
1	彼	-1416.34	-4562.88	206	0.098	325	3.729
2	た	-369.46	-522.48	4074	1.931	479	5.496
3	或	-236.84	-872.20	3	0.001	40	0.459
4	彼女	-178.58	-652.72	1	0.000	29	0.333
5	かう	-134.73	-485.81	2	0.001	23	0.264
6	為	-115.89	-414.06	2	0.001	20	0.229
7	こと	-98.99	-238.06	63	0.030	34	0.390
8	自身	-96.83	-279.29	23	0.011	25	0.287
9	の	-96.37	-109.29	9529	4.516	603	6.919
10	感じ	-92.73	-235.83	44	0.021	29	0.333
11	は	-83.35	-96.18	6494	3.078	432	4.957
12	近い	-78.00	-255.33	6	0.003	16	0.184
13	や	-68.12	-117.42	216	0.102	45	0.516
14	一	-54.43	-75.33	666	0.316	76	0.872
15	又	-53.94	-171.71	4	0.002	11	0.126
16	に	-52.99	-59.05	7121	3.375	428	4.911
17	うに	-50.75	-163.00	3	0.001	10	0.115
18	中	-50.31	-78.73	264	0.125	43	0.493
19	等	-50.14	-88.87	129	0.061	30	0.344

ウィンドウ下部の「降順」ボタンのオンオフで、表示を昇順、降順に切り替えられます。数字でソートしている時は降順だと数字の大きい順、昇順だと数字の小さい順です。50 音順でソートしているときは、降順だと「あ～ん」の順、昇順だと「ん～あ」」の順で表示されます。２つの指標でソートしているときの表の下にある語や数値が上に来ますが、これらはマイナスの数値になっています。対数尤度比とカイ二乗値は特徴度を表す数値に過ぎないため、「特徴的に多い」も「特徴的に少ない」も同じように高い数値になります。そこで本ソフトでは、特徴的に少ない場合は数値に－1 をかけて、マイナスの数値にしています。特徴的に少ないの基準は、%の数字が参照ファイルの方が高いものです。これで、「メインファイルに特徴的に少ない」≒「参照ファイルに特徴的に多い」語が分かりますが、あくまでもメインファイルに使われた語だけを対象としたものなので、参照ファイルのみで出現した語はリストアップされません。参照ファイルに特徴的な語を見るためには、メインファイルと参照ファイルの指定を逆にして再度算出します。

検索

個別の語を指定して表示させることができます。



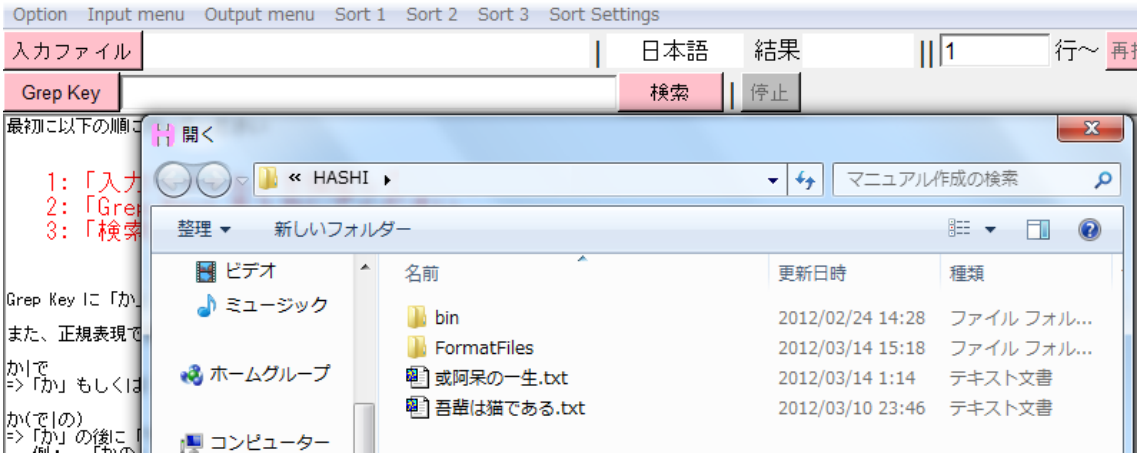
検索語句を指定してから「実行」を押すと、指定した語句のみが表示されます。決まった語の特徴度を探したいときに使います。

文字列の検索(Grep)

　この処理は、整形しないプレーンテキストを対象に、タグを全く利用せずに文字列のみで検索を行うものです。非常に高速に検索が行えます。

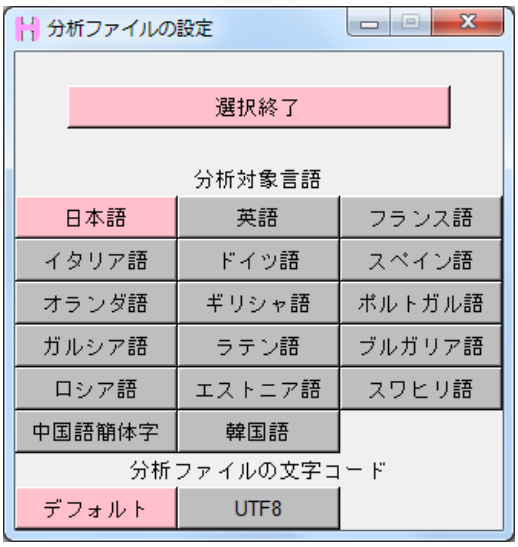
他の処理とは違い、テキスト中に付く様々なタグ情報を一切使わずに文字列のみで検索をする処理になります。使用するファイルも整形しないプレーンテキストのみです。タグ情報を使わないため、多層的な複雑な検索はできませんが、その代わり分かち書きされて分けられた語ではできない、連続した長い文字列を自在に条件付けて検索することができます。整形を行わないためファイル選択後にすぐに検索を開始でき、文字列と言う単一の条件のため検索も高速に行えます。検索結果を使い他の処理で統計を扱うことはできません。

ファイル選択



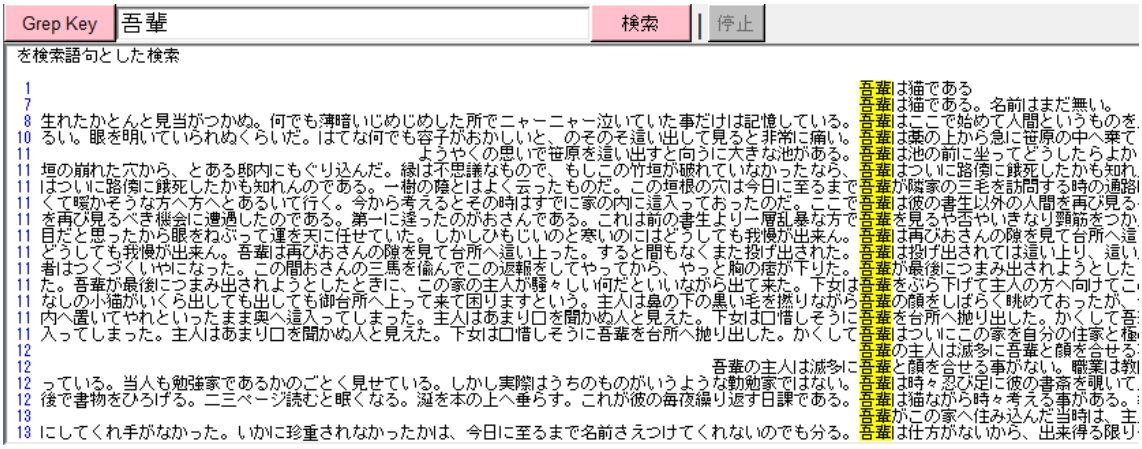
ファイルはウィンドウ上部の「入力ファイル」で選択します。整形しないプレーンテキストのみ扱うため、常に新規のテキストファイルを選択します。

分析ファイルの設定



他の処理での新規ファイルの選択時と同様に分析するテキストファイルの言語と、分析するファイル内の文字コードを選択します。
文字列としてのみ扱い、整形はしませんので、分析したい単位の項目はありません。
選択後は「選択終了」ボタンをクリックします。

検索



検索は、ファイル指定後、Grep Key に検索する文字列を指定し「検索」ボタンで行います。
結果は KWIC 形式で表示され、真ん中に検索語が、左右に本文中での文脈が並びます。検索語は黄色いバックで表示されます。

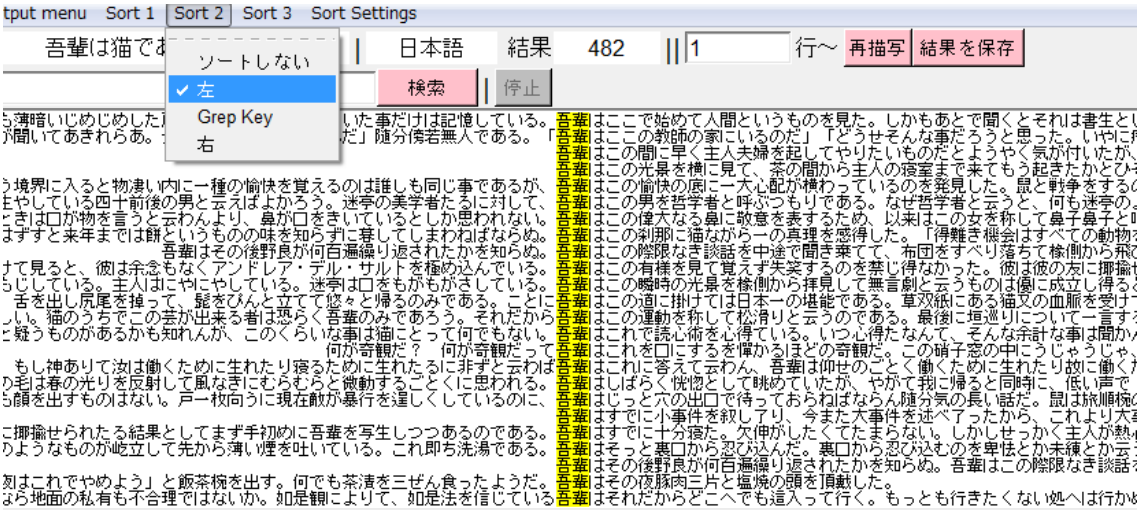
ソート

KWIC と同様に結果の並べ替えができます。

ウィンドウ上部の Sort1～Sort3 を指定すると、指定した位置で並び変えができます。Sort1 を指定し、「再描写」ボタンを押すと、指定した位置の語の 50 音順を基準に、全ての行が並び変わります。「位置」は検索語とその左右です。検索語の位置が「Grep Key」となり、その右側か左側かです。右側を指定した場合は、Grep Key の直後の文字から規定の幅の文字列を順に比べて並べ替えます。

複数条件でのソート

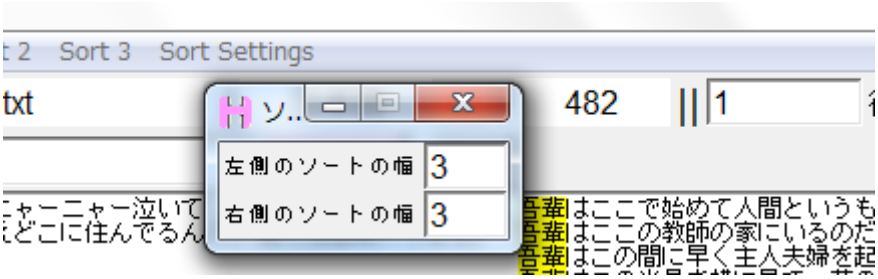
KWIC と同様に、ソート条件を複数組み合わせて指定できます。



第２条件は、第１条件の位置の文字列が全く同じだった場合にその中で並べ替えをする条件に使われます。左側を選択した場合は、Grep Key の直前の文字から規定の幅の文字列を文字列末から順に比べて並べ替えます。

Sort Settings

ソート時の条件に使う幅の指定ができます。



ウィンドウ上部の「Sort Settings」で現れるウィンドウでソート時の基準とする文字列の規定幅を指定できます。通常では 3 文字分となています。

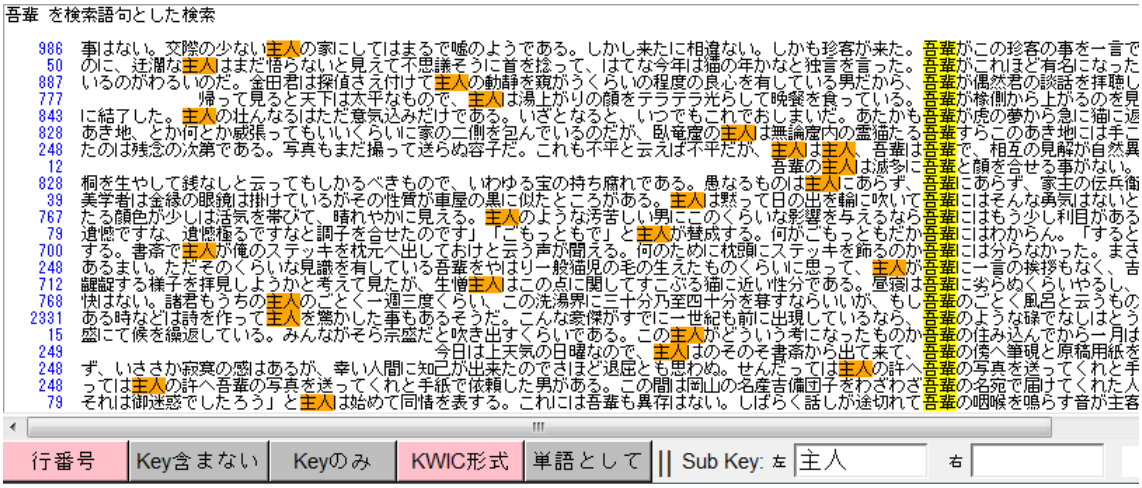
本文リンク

KWIC と同様に、検索結果の本文を確認することができます。

表示の右側の青い行番号をクリックで行の全文が読めます。更に、「<＝前の行」「次の行=>」で、前後の行の確認もできます。検索文字列の箇所は黄色いバックで表示されます。

Sub Key

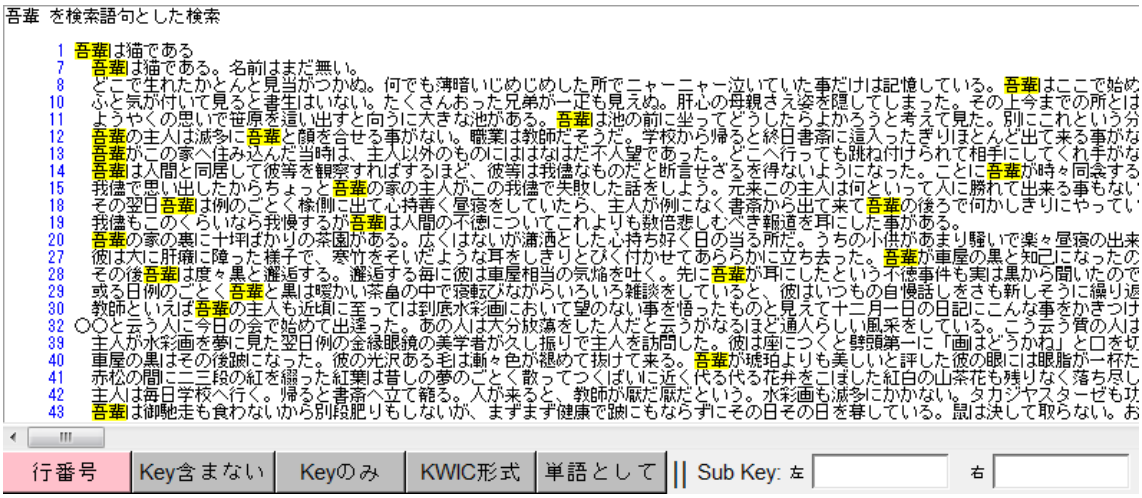
検索結果を更に左右の文字列を指定して絞ることができます。



ウィンドウ下部の Sub Key を指定し、「再描写」で、表示を左右に指定した文字列がある結果のみに絞ることができます。Sub Key は左と右で別々に指定します。片方のみの指定もできます。両方指定をした場合は、２つの条件が両方とも揃った結果のみが表示されます。Sub Key に指定された文字列はオレンジ色のバックで表示されます。

KWIC 形式

検索結果を、通常のテキストのように書かれた形式まま１行丸々表示することができます。



ウィンドウ下部の「KWIC 形式」のオンオフで、検索文字列が中央に揃う KWIC 形式と通常のテキストのように１行がそのまま表示される形式を切り換えることができます。通常では KWIC 形式はオンになっています。

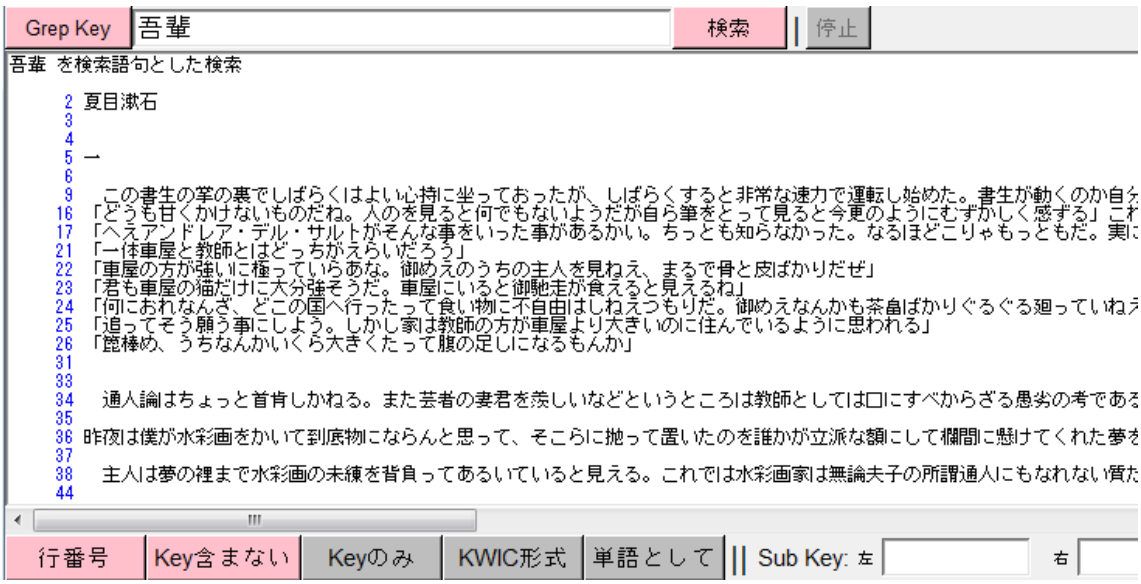
ソート、Sub Key の指定は KWIC 形式がオンになっていないと使えない機能です。

検索文字列のみの表示

検索結果の中で検索文字列のみを表示することができます。

ウィンドウ下部の「Key のみ」で、検索文字列のみの表示にできます。KWIC 形式をオンにしていると、縦 1 列に検索文字列が並びます。

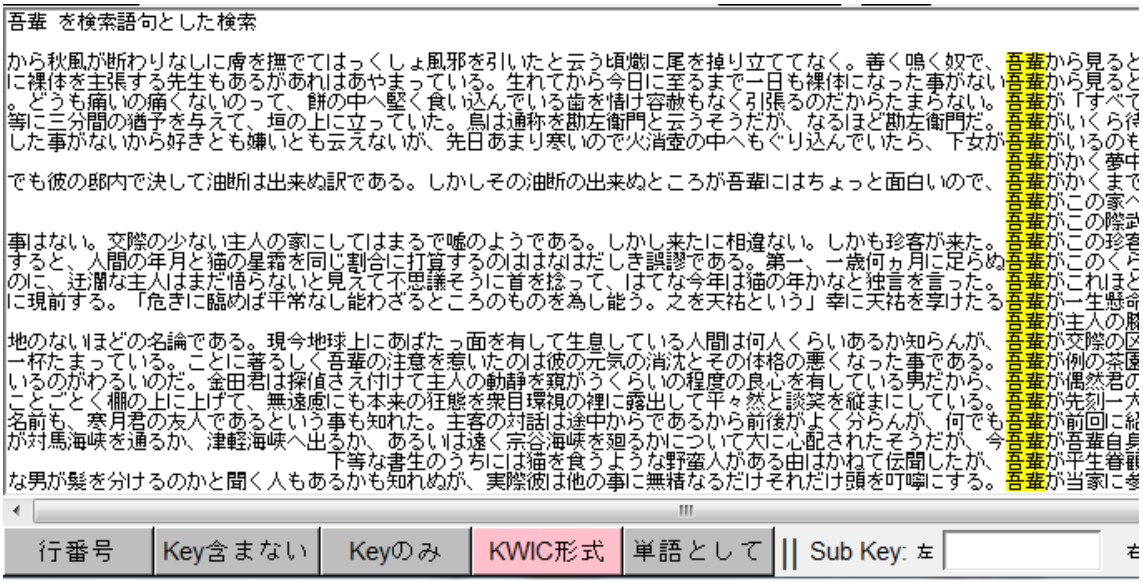
検索文字列を含まない結果のみの表示



ウィンドウ下部の「Key 含まない」で、指定した検索文字列が含まれない行のみを表示することができます。これと「Key のみ」、「KIWC 形式」は共存できない機能になります。

行番号表示

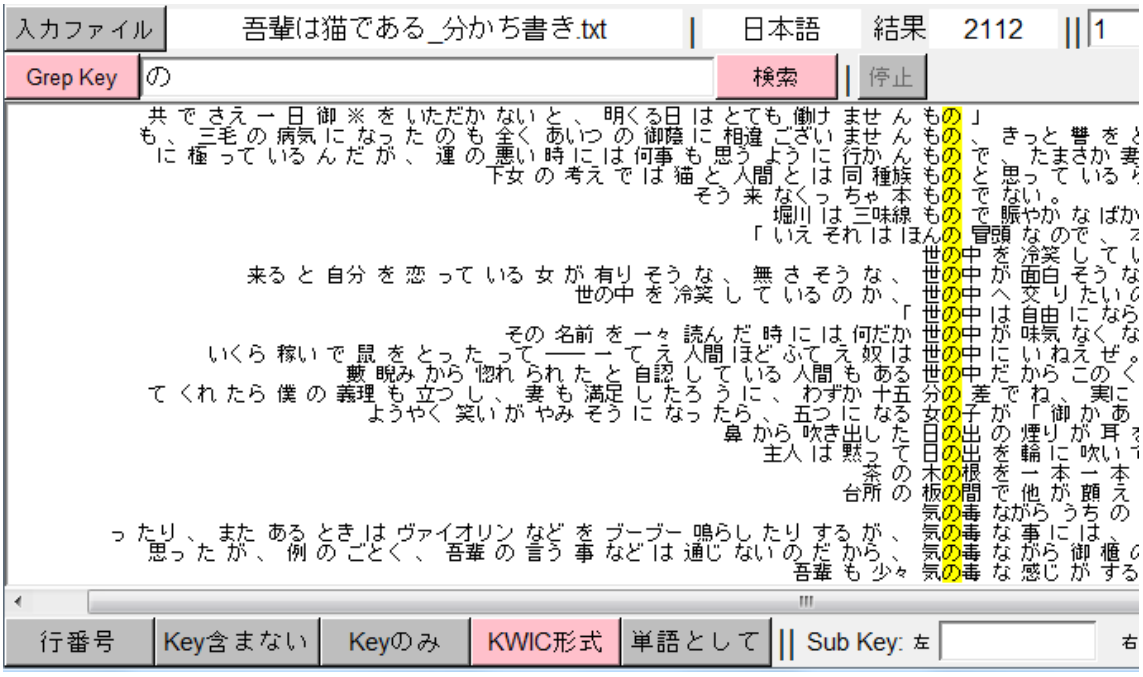
画面左端の行番号表示の有無を選択できます。



ウィンドウ下部の「行番号」のオンオフで、画面の左端に表示される行番号の表示、非表示を決定できます。通常はオンになっています。これをオフにすると行番号が表示されないので、本文へのリンクができなくなります。

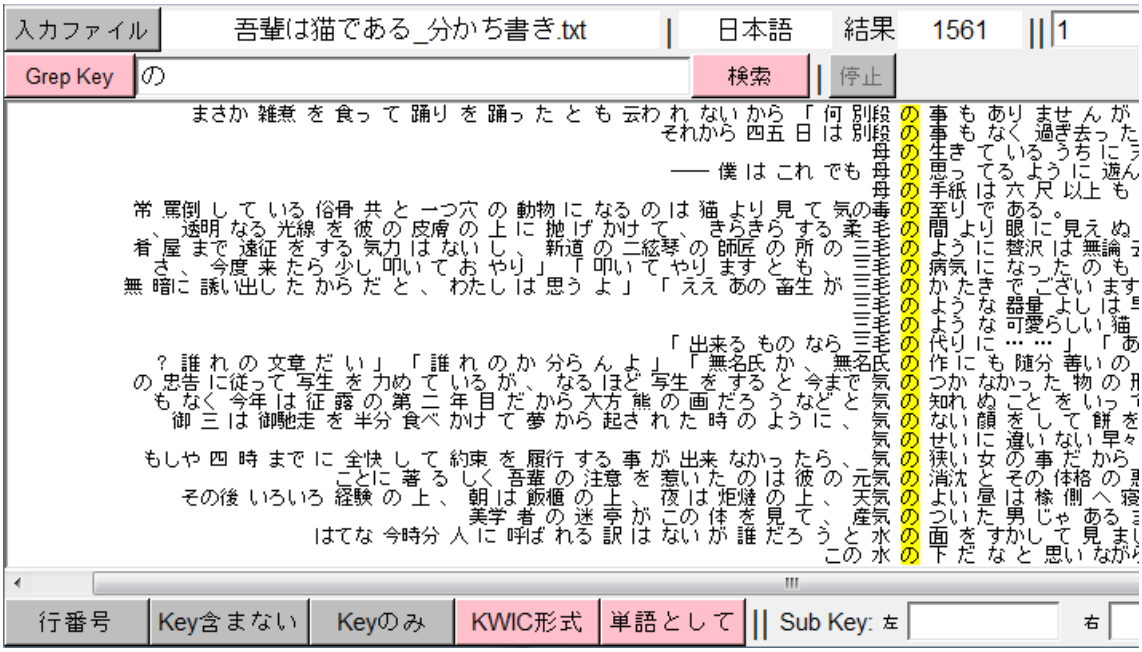
検索文字列を単語として検索する

分かち書きのされた言語やテキストを検索する際に検索文字列を単語として指定できます。



この処理は、単語の区切りを関係なく全てを単純に連続した文字列として検索を行うため、語の区切りを認識せず、検索文字列が語の中の一部かどうかに関わらず検索を行います。

そこで、ウィンドウ下部の「単語として」を押すと、以降の検索では検索文字列を単語として検索します。つまり、指定した文字列の両脇に語の区切りがあるもののみを検索するようになります。



この機能を使い通常の単語を検索しようとする際に、分かち書きされていないテキストを指定していると何も検索できなくなります。

通常の処理にまつわること

「検索」「再描写」の違い

「検索」は、検索語句やソート条件などを全て読み込みなおして新しく分析テキストを読み込み検索するものです。
「再描写」は、一度検索した後にソートや「周囲の語句」の指定などを追加した際に新たに検索からはし直さなくてもいいが、変更した表示条件を反映させる際に使います。基本的にウィンドウ上部のツールバーの各条件を変更した後に使います。

分析言語

分析言語は「日本語」「英語」「ドイツ語」「フランス語」「イタリア語」「ポルトガル語」「ガルシア語」「ラテン語」「ブルガリア語」「ロシア語」「エストニア語」「スワヒリ語」「中国語」「韓国語」を扱えます。

※日本語、英語、韓国語以外の言語は整形の対応がまだできていません。付与される項目の、品詞タグの作成や簡易入力リストも英語のものを流用しているので、言語ごとの解析を反映したものになっていません。

HASHI では分析言語にさまざまな文法タグを付けるために形態素解析ソフトが必要です。日本語の場合は同梱されている茶筌で形態素分けをし、文法などのタグが付けられます。英語を始めとする他の言語はで初期状態では形態素解析がされず、文法などのタグは付きませんが、文法などのタグ無しの状態で分析することができます。

その際タグは「品詞タグ」のみが付きます。品詞の中では「単語」と「記号」の区別だけがされます。付与されるタグの制限がある以外はすべての機能が使えます。

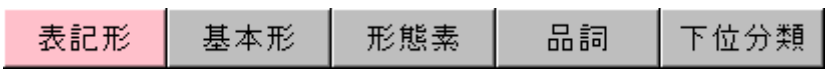
茶筌の場合も、使用する内部辞書を置き換えることで解析結果を変えることができます。茶筌用辞書の置き換えや、日本語以外の言語の形態素解析ソフトの設置方法については後で詳しく説明します。

各言語を所定の形態素解析ソフトがある状態で整形すると、使用できるタグにそれぞれ違いが出ます。英語、中国語、韓国語を例に提示します。

TreeTagger 有りの英語

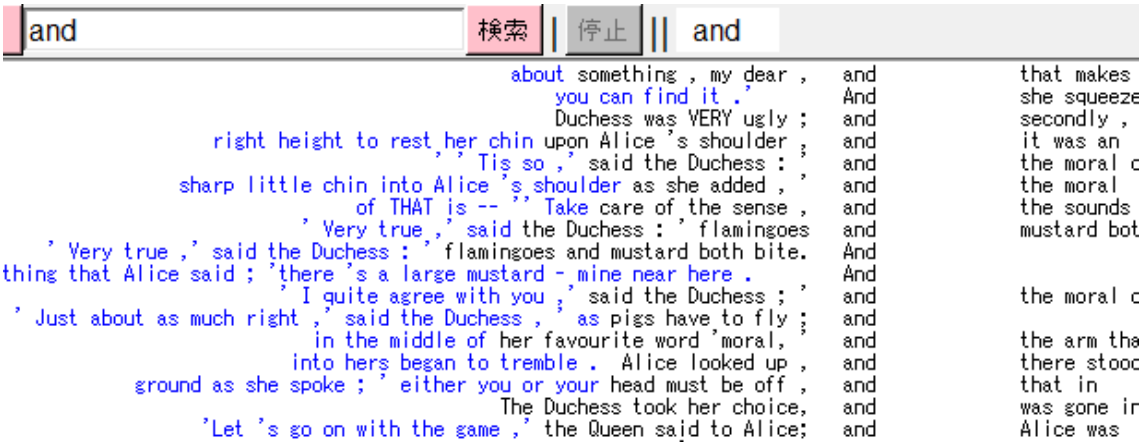
TreeTagger 有りの中国語

MACH 有の韓国語



大文字小文字同時検索

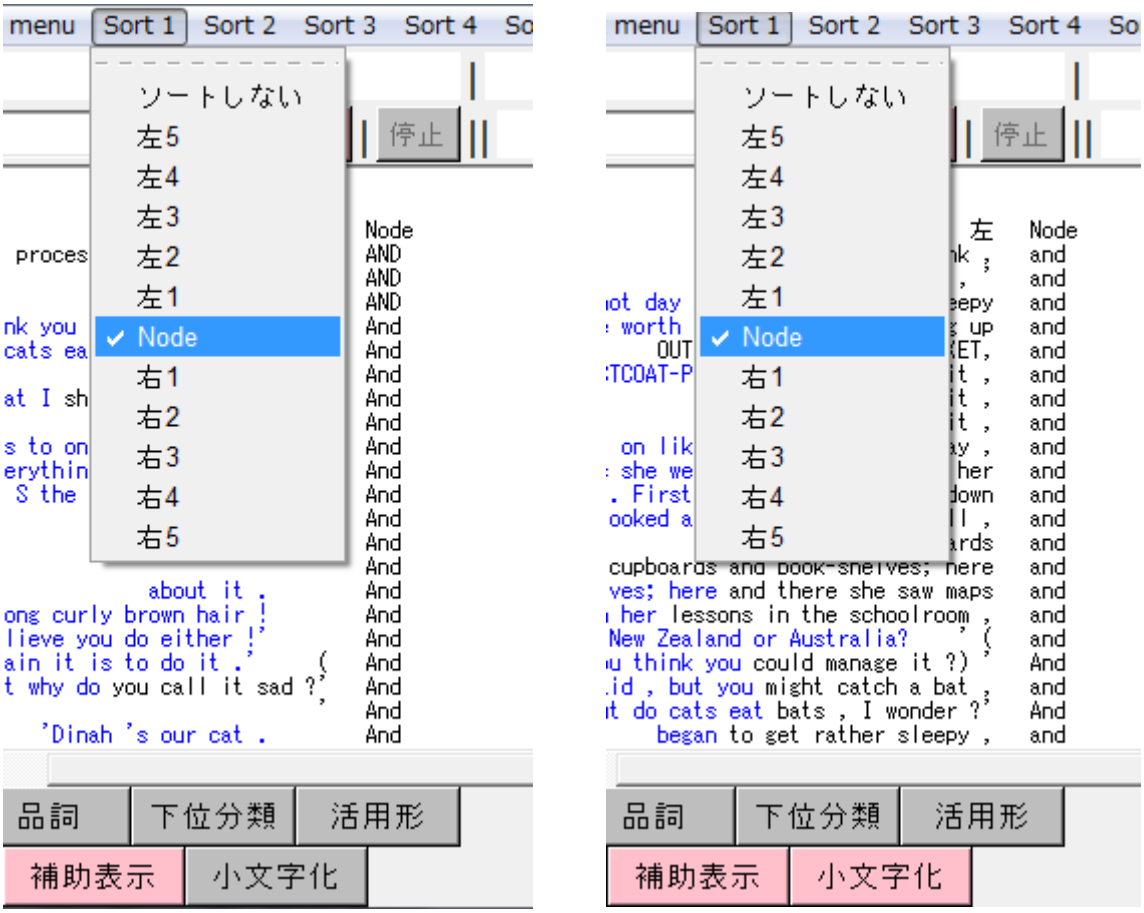
英語などアルファベットを使う言語の場合大文字と小文字は同じ文字として検索されます。



検索文字列の指定欄に入力した文字は常に大文字でも小文字でも同じ文字列を検索しているという扱いになって検索されます。

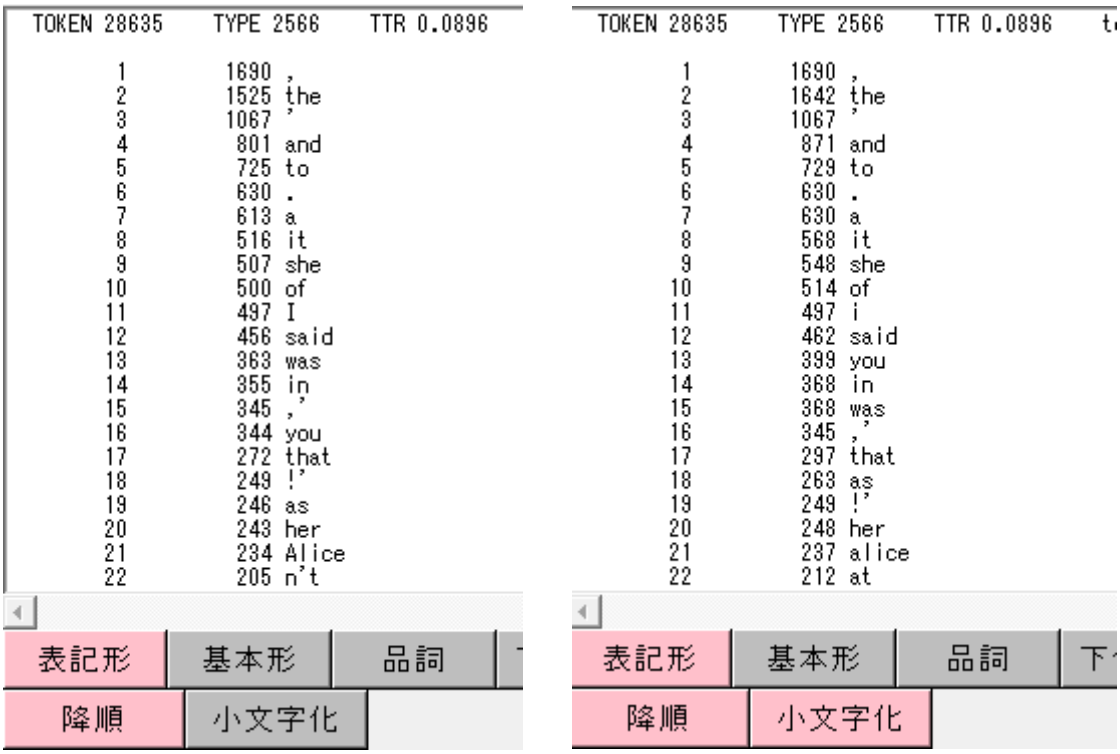
検索語の文字列の小文字化

検索語に表示された文字列を全て小文字として扱うこともできます。



KWIC でのソートの場合、通常では大文字と小文字の語は別々に並びます。しかし、大文字での語でも小文字での語でも同じように扱いたい場合もあります。その際にウィンドウ下部の「小文字化」ボタンで、大文字小文字の差を無視して並べ替えることができます。KWIC の場合は表示上は大文字小文字の区別がされていますが、並べ替えの判断時には全てが小文字であるとして扱われます。

同様に、数値を扱う処理の際も大文字と小文字の語を統一して扱うことができます。

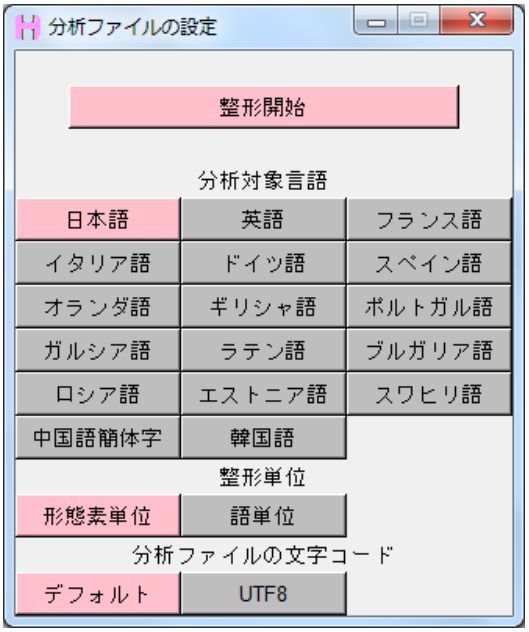


全ての数値を扱う処理でいえることですが、同一単語でも大文字と小文字の語を別集計すると不都合なときがあります。その際も小文字化をすることで全ての単語を小文字に変えて集計します。これで使用された単語の確実な数を知ることができます。

整形単位

本ソフトで使用する語の区切りは２種類あります。１つは「形態素単位」で１つは「語単位」です。日本語の場合、形態素単位は形態素解析ソフトで分かち書きされた単位、語単位はそれを元にルールによりいくつかの形態素を結合し作成したオリジナル単位です。

整形単位の選択



テキスト選択後の分析ファイルの設定時に整形単位の選択を行います。
日本語と韓国語以外では語単位のみしか選択できません。

語単位では形態素を結合していくつかの語がつくられますが、大きくは、

サ変動詞が復元される	「勉強する」「オープンする」
ナ形容詞が復元される	「勝手な」「古風な」
「～と」の副詞が復元される	「公然と」「堂々と」
活用語尾の助動詞が連結する	「助けてくれるだろう」「可哀そうでした」
接頭辞や接尾辞を本体と連結する	「おばあさん」「私達」

などになります。

語単位での整形は形態素単位での整形過程に加えてテキスト中の全ての語に複雑なルールを適応させていきますので非常に時間がかかります。形態素単位での整形の５倍程度かかることもあります。

語単位の整形ルール

UniDic の場合の結合ルールを提示します。
ルールは上から順に適応されていきます。

連続する半角アルファベット	→ 連結
固有名詞-人名-姓の後の固有名詞-人名-名	→ 連結し固有名詞-人名-フルネーム'
連続する数詞	→ 連結
数詞の後の助数詞	→ 連結し数詞
接頭辞と後の名詞,動詞,形容詞,形状詞	→ 連結し後ろの品詞
名詞,動詞,形容詞,形状詞,接頭辞と後の接尾辞	→ 連結し前の品詞-後の品詞転成
代名詞と後の接尾辞	→ 連結し代名詞
形容詞	→ イ形容詞
サ変動詞語幹と後の「する,できる,なさる,いたす」	→ 連結
助動詞語幹と後の「だ」	→ 連結し助動詞
形状詞の助動詞語幹	→ 助動詞
形状詞の一般と転成	→ ナ形容詞
ナ形容詞と後の助動詞「だ」	→ 連結しナ形容詞
名詞の形状詞可能と後の助動詞「だ」	→ ナ形容詞
動詞の連用形と後の接続助詞「て,で」	→ 連結し動詞
「て,で」で終わる動詞の連用形と後の非自立可能の動詞	→ 連結し動詞
動詞の連用形と後の副助詞「たり」	→ 連結し動詞
動詞の仮定形と後の接続助詞「ば」	→ 連結し動詞
動詞,イ形容詞,ナ形容詞,助動詞と後の助動詞	→ 連結し前の品詞
形状詞-タリと後の「と」	→ 連結し副詞
形状詞-タリと後の「たる」	→ 連結し連体詞
連続する終助詞	→ 連結
動詞の連用形と後の非自立可能の動詞かイ形容詞	→ 連結し複合動詞

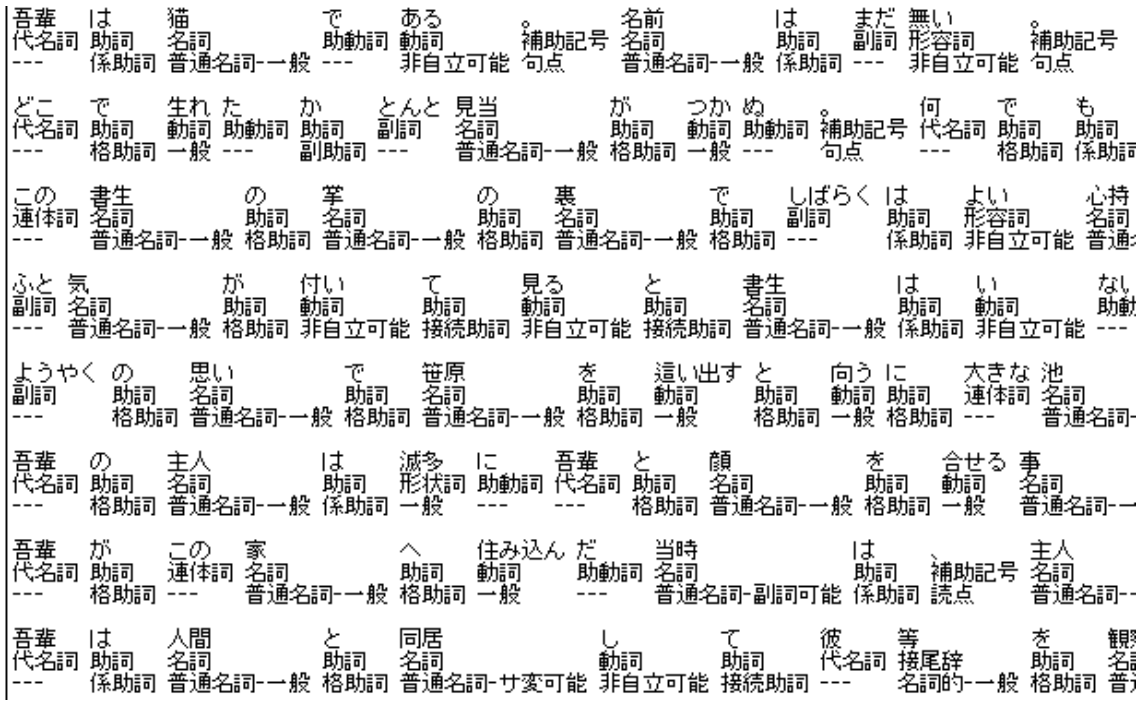
本ソフトの配布版に同梱している IpaDic でもこれに近い整形が行われています。

このルールは、HASHI のフォルダ内の「bin」→「Format_Rules」フォルダの中にルールモジュールファイルを入れて、中身を書きかえれば変更できます。ルールファイルは、UniDic 版は「O_Uni_Format_Rules.pm」、IPADic 版は「O_Ipa_Format_Rules.pm」です。

ルールファイルは本ソフトのホームページで公開しますので、各自で設置してください。

語単位の整形例

まず、形態素単位の場合の整形例を示します。



吾輩	は	猫	で	ある	。	名前	は	まだ	無い	。
代名詞	助詞	名詞	助動詞	動詞	補助記号	名詞	助詞	副詞	形容詞	補助記号
---	係助詞	普通名詞-一般	---	非自立可能	句点	普通名詞-一般	係助詞	---	非自立可能	句点

どこ	で	生れ	た	か	とんと	見当	が	つか	ぬ	。	何	で	も
代名詞	助詞	動詞	助動詞	助詞	副詞	名詞	助詞	動詞	助動詞	補助記号	代名詞	助詞	助詞
---	格助詞	一般	---	副助詞	---	普通名詞-一般	格助詞	一般	---	句点	---	格助詞	係助詞

この	書生	の	筆	の	裏	で	しばらく	は	よい	心持
連体詞	名詞	助詞	名詞	助詞	名詞	助詞	副詞	助詞	形容詞	名詞
---	普通名詞-一般	格助詞	普通名詞-一般	格助詞	普通名詞-一般	格助詞	---	係助詞	非自立可能	普通

ふと	気	が	付い	て	見る	と	書生	は	い	ない
副詞	名詞	助詞	動詞	助詞	動詞	助詞	名詞	助詞	動詞	助動
---	普通名詞-一般	格助詞	非自立可能	接続助詞	非自立可能	接続助詞	普通名詞-一般	係助詞	非自立可能	---

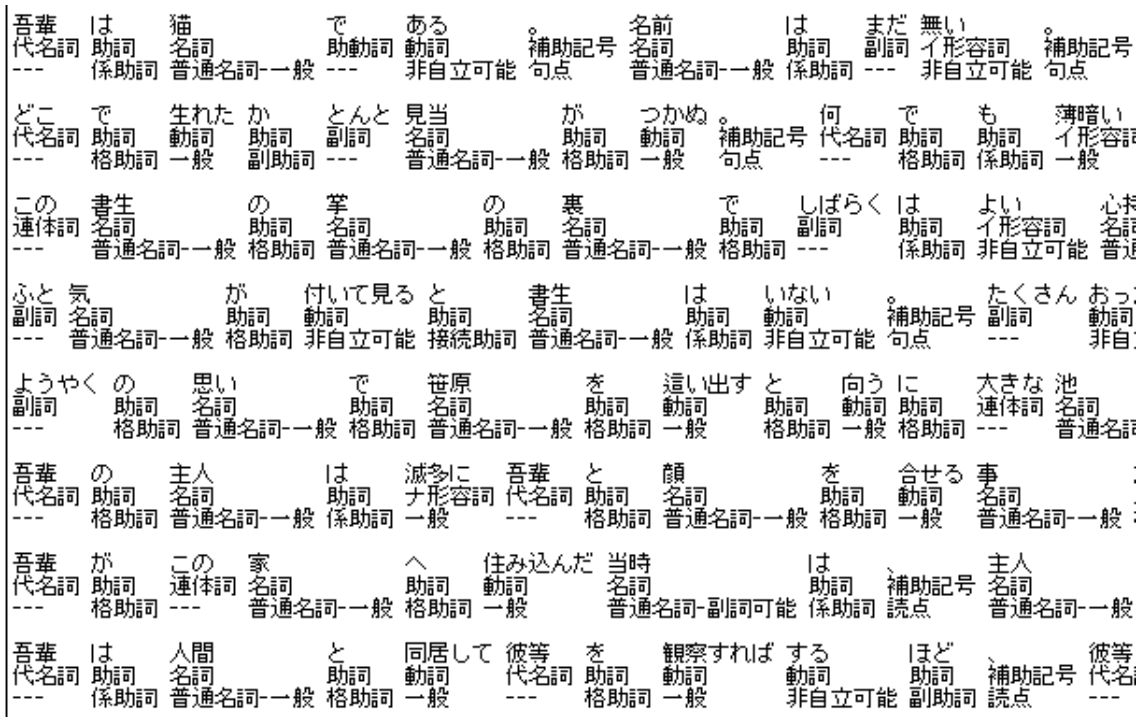
ようやく	の	思い	で	笹原	を	這い出す	と	向う	に	大きな	池
副詞	助詞	名詞	助詞	名詞	助詞	動詞	助詞	動詞	助詞	連体詞	名詞
---	格助詞	普通名詞-一般	格助詞	普通名詞-一般	格助詞	一般	格助詞	一般	格助詞	---	普通名詞-

吾輩	の	主人	は	滅多	に	吾輩	と	顔	を	合せる	事
代名詞	助詞	名詞	助詞	形状詞	助動詞	代名詞	助詞	名詞	助詞	動詞	名詞
---	格助詞	普通名詞-一般	係助詞	一般	---	---	格助詞	普通名詞-一般	格助詞	一般	普通名詞-

吾輩	が	この	家	へ	住み込ん	だ	当時	は	、	主人
代名詞	助詞	連体詞	名詞	助詞	動詞	助動詞	名詞	助詞	補助記号	名詞
---	格助詞	---	普通名詞-一般	格助詞	一般	---	普通名詞-副詞可能	係助詞	読点	普通名詞-

吾輩	は	人間	と	同居	し	て	彼	等	を	観
代名詞	助詞	名詞	助詞	名詞	動詞	助詞	代名詞	接尾辞	助詞	名
---	係助詞	普通名詞-一般	格助詞	普通名詞-サ変可能	非自立可能	接続助詞	---	名詞的-一般	格助詞	普

次に、語単位の整形例を示します。



吾輩	は	猫	で	ある	。	名前	は	まだ	無い	。
代名詞	助詞	名詞	助動詞	動詞	補助記号	名詞	助詞	副詞	イ形容詞	補助記号
---	係助詞	普通名詞-一般	---	非自立可能	句点	普通名詞-一般	係助詞	---	非自立可能	句点

どこ	で	生れた	か	とんと	見当	が	つかぬ	。	何	で	も	薄暗い
代名詞	助詞	動詞	助詞	副詞	名詞	助詞	動詞	補助記号	代名詞	助詞	助詞	イ形容詞
---	格助詞	一般	副助詞	---	普通名詞-一般	格助詞	一般	句点	---	格助詞	係助詞	一般

この	書生	の	筆	の	裏	で	しばらく	は	よい	心
連体詞	名詞	助詞	名詞	助詞	名詞	助詞	副詞	助詞	イ形容詞	名
---	普通名詞-一般	格助詞	普通名詞-一般	格助詞	普通名詞-一般	格助詞	---	係助詞	非自立可能	普

ふと	気	が	付いて見る	と	書生	は	いない	。	たくさん	おっ
副詞	名詞	助詞	動詞	助詞	名詞	助詞	動詞	補助記号	副詞	動
---	普通名詞-一般	格助詞	非自立可能	接続助詞	普通名詞-一般	係助詞	非自立可能	句点	---	非自

ようやく	の	思い	で	笹原	を	這い出す	と	向う	に	大きな	池
副詞	助詞	名詞	助詞	名詞	助詞	動詞	助詞	動詞	助詞	連体詞	名詞
---	格助詞	普通名詞-一般	格助詞	普通名詞-一般	格助詞	一般	格助詞	一般	格助詞	---	普通名詞

吾輩	の	主人	は	滅多に	吾輩	と	顔	を	合せる	事	
代名詞	助詞	名詞	助詞	ナ形容詞	代名詞	助詞	名詞	助詞	動詞	名詞	
---	格助詞	普通名詞-一般	係助詞	一般	---	格助詞	普通名詞-一般	格助詞	一般	普通名詞-一般	

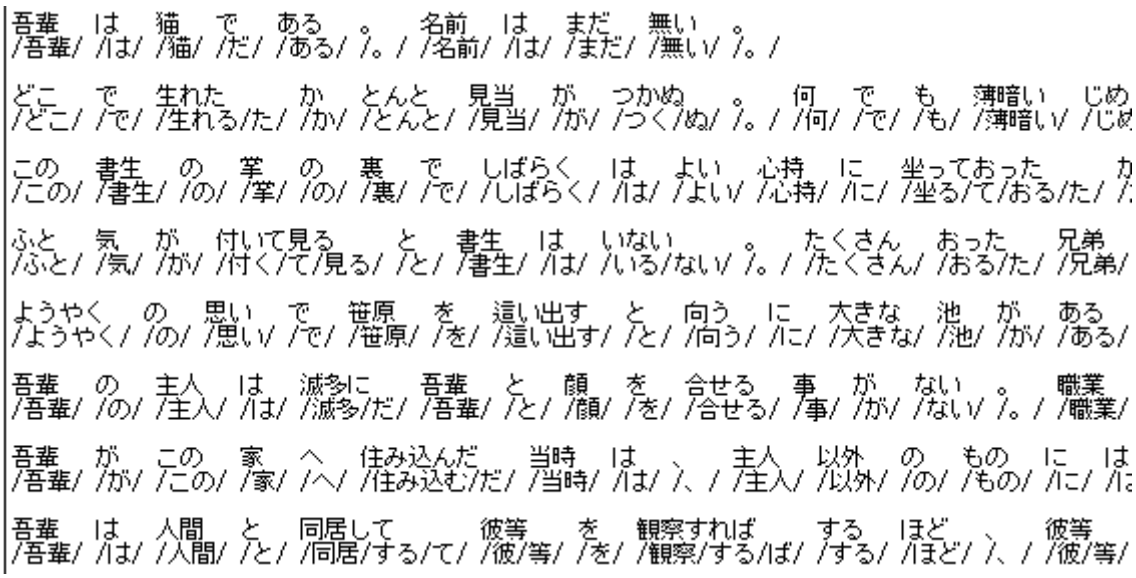
吾輩	が	この	家	へ	住み込んだ	当時	は	、	主人
代名詞	助詞	連体詞	名詞	助詞	動詞	名詞	助詞	補助記号	名詞
---	格助詞	---	普通名詞-一般	格助詞	一般	普通名詞-副詞可能	係助詞	読点	普通名詞-一般

吾輩	は	人間	と	同居して	彼等	を	観察すれば	する	ほど	、	彼等
代名詞	助詞	名詞	助詞	動詞	代名詞	助詞	動詞	動詞	助詞	補助記号	代名
---	係助詞	普通名詞-一般	格助詞	一般	---	格助詞	一般	非自立可能	副助詞	読点	---

ナ形容詞「滅多に」、サ変動詞「観察すれば」、動詞のテ形「付いて見ると」などが復元されていることが分かります。

構成形態素

結合された語が、元々何の形態素が繋がったものか、「構成形態素」の項目で確認できます。

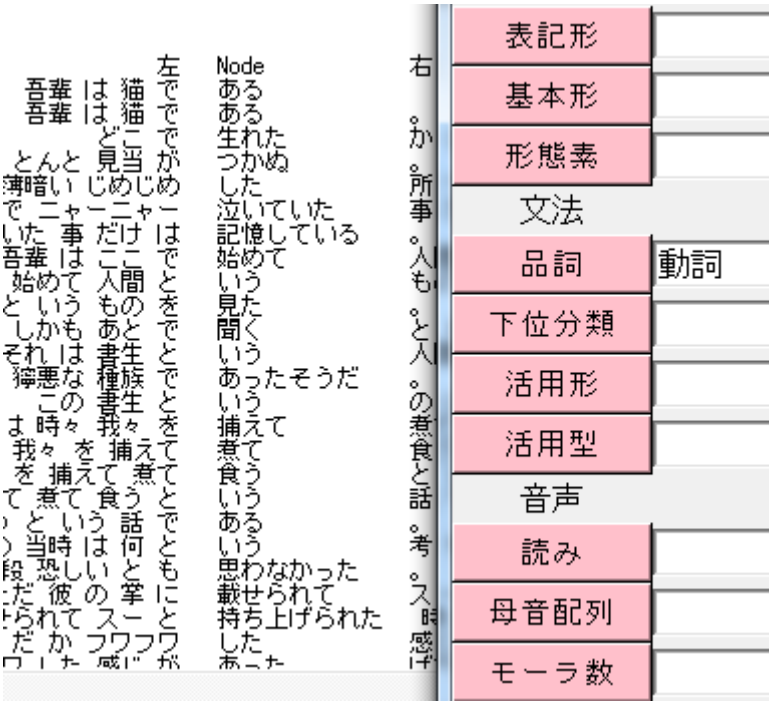


構成形態素は、ソフト上では「形態素」という名前のボタンで扱います。

例えば、「観察すれば」は「/観察/する/ば/」と３つの形態素の連結で作られていることが分かります。構成形態素は元々の形態素の基本形になります。

検索

検索は形態素単位の際と同様に行えます。



動詞に連なる接続助詞「て」や助動詞なども結合して１つの動詞扱いになりますので、通常で想像される「語」に近い単位と言えます。

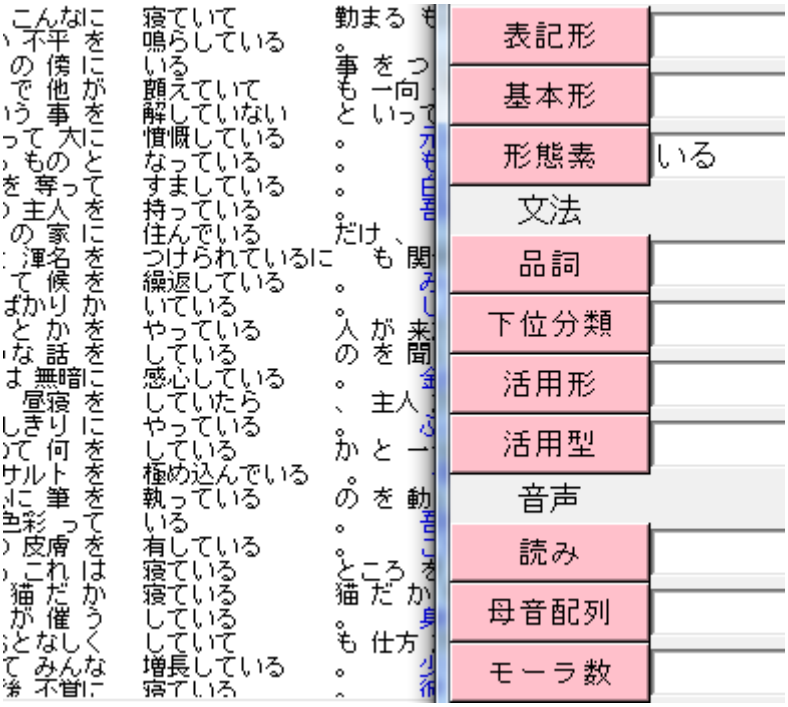
語末ソート

１語の中に様々な助動詞などを内在しますので、その助動詞が語の右側に現れます。

例えば、動詞をすべて検索した後に語末でソートすれば、活用形や内在する補助動詞ごとに塊で現れます。

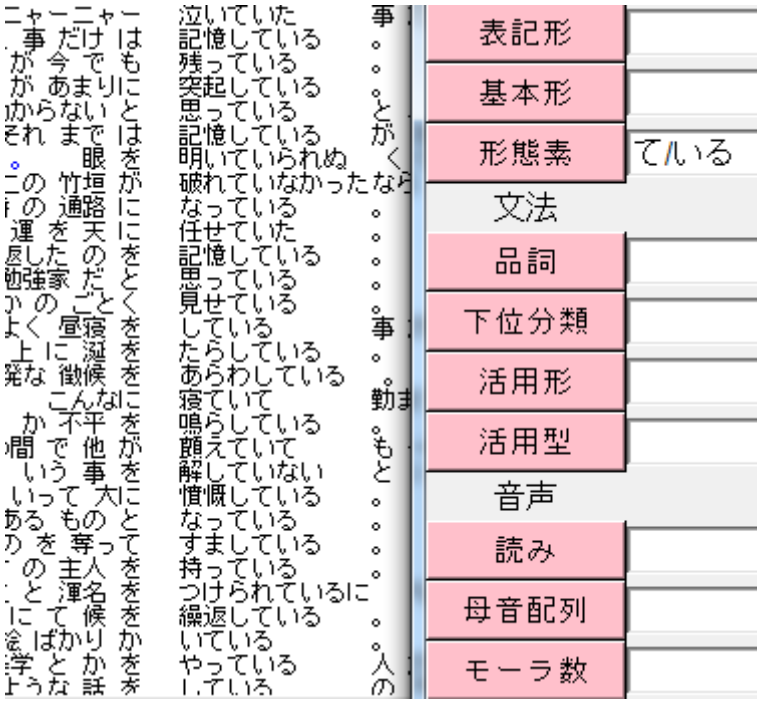
これで、動詞全体でどの活用形や補助動詞が多く使われているのか確認ができます。

構成形態素の検索



構成形態素で検索をした場合、指定した形態素が１つでのその語の中にあれば検出されます。つまり、本動詞の場合も補助動詞の場合も同時に検索されます。

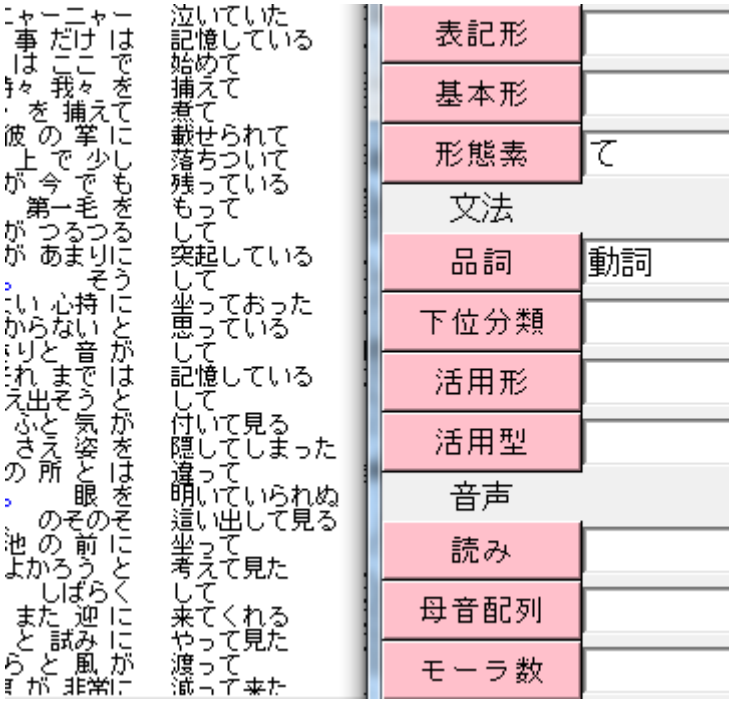
連続する構成形態素の指定



構成形態素は全て「/」で区切られて記録されています。検索にもこれを応用できます。

「て/いる」などの形で、形態素の区切りを付け２つ以上の形態素を指定すればより詳細に語形の指定ができます。

複合検索



構成形態素も品詞などの他の要素と組み合わせることでより明確な条件とできます。例えば形態素「て」、品詞「動詞」なら、すべての動詞のテ形が検索できます。

※ここでのテ形の例は実際には形態素は「(て | で)」とするべきです。

タグ項目

以下に、本ソフトで使用できる、整形時に自動で付与されるタグを示します。
日本語と英語を例にします。

日本語の場合

表記形
元テキストでの記述そのままの形

```
どこ で 生れ た か とんと 見当 が つか ぬ 。
どこ で 生れ た か とんと 見当 が つか ぬ 。
```

基本形
表記形の活用がされていない形

```
どこ で 生れ  た か とんと 見当 が つか ぬ 。
どこ で 生れる た か とんと 見当 が つく ぬ 。
```

構成形態素
１語の中に含まれている各形態素（「語単位」の時のみ意味を持つ）

```
どこ   で   生れ     た   か   とんと   見当   が   つか   ぬ   。
/どこ/ /で/ /生れる/ /た/ /か/ /とんと/ /見当/ /が/ /つく/ /ぬ/ /。/
```

「語単位」の際の構成形態素

```
どこ   で   生れた     か   とんと   見当   が   つかぬ     。
/どこ/ /で/ /生れる/た/ /か/ /とんと/ /見当/ /が/ /つく/ぬ/ /。/
```

品詞
茶筌での品詞第一分類

```
どこ   で   生れ た     か   とんと 見当 が   つか ぬ     。
代名詞 助詞 動詞 助動詞 助詞 副詞   名詞 助詞 動詞 助動詞 補助記号
```

下位分類
茶筌での品詞第二分類以下

```
どこ で     生れ た  か     とんと 見当         が     つか ぬ  。
---  格助詞 一般 --- 副助詞 ---    普通名詞-一般 格助詞 一般 --- 句点 -
```

活用形

茶筌での活用形

```
どこ で 生れ た か とんと 見当 が つか ぬ 。
--- --- 連用形-一般 終止形-一般 --- --- --- --- 未然形-一般 終止形-一般 ---
```

活用型

茶筌での活用型

```
どこ で 生れ た か とんと 見当 が つか ぬ 。
--- --- 下一段-ラ行-一般 助動詞-タ --- --- --- --- 五段-カ行-一般 助動詞-ヌ ---
```

読み

茶筌での読み

```
どこ で 生れ た か とんと 見当 が つか ぬ 。
ドコ デ ウマレ タ カ トント ケントウ ガ ッカ ヌ 。
```

母音配列

茶筌での読みから、母音のみを抜き出したもの

```
どこ で 生れ た か とんと 見当 が つか ぬ 。
OO  E  UAE  A  A  ONO   ENO- A  UA  U  XXX
```

モーラ数

母音配列の母音数

```
どこ で 生れ た か とんと 見当 が つか ぬ 。
2   1  3   1  1  3     4   1  2   1  0
```

英語の場合（TreeTagger 有り）

表記形

```
Alice was beginning to get very tired of sitting by her sister on the bank ,
Alice was beginning to get very tired of sitting by her sister on the bank ,
```

基本形

```
Alice was beginning to get very tired of sitting by her sister on the bank ,
Alice be  begin     to get very tired of sit     by her sister on the bank ,
```

品詞

TreeTagger でのタグを品詞としての共通部分をまとめたもの、本ソフト独自判別

```
Alice was  beginning to   get  very tired of       sitting by       her  sister on       the bank  ,
SUBST VERB VERB      PREP VERB ADV  ADJ   PREP/SBCJ VERB    PREP/SBCJ PRON SUBST  PREP/SBCJ ADJ SUBST PUN
```

下位分類

TreeTagger で付与されるタグそのまま

```
Alice was beginning to get very tired of sitting by her sister on the bank ,
NP    VBD VVG       TO VV  RB   JJ    IN VVG     IN PP$ NN     IN DT  NN   ,
```

活用形

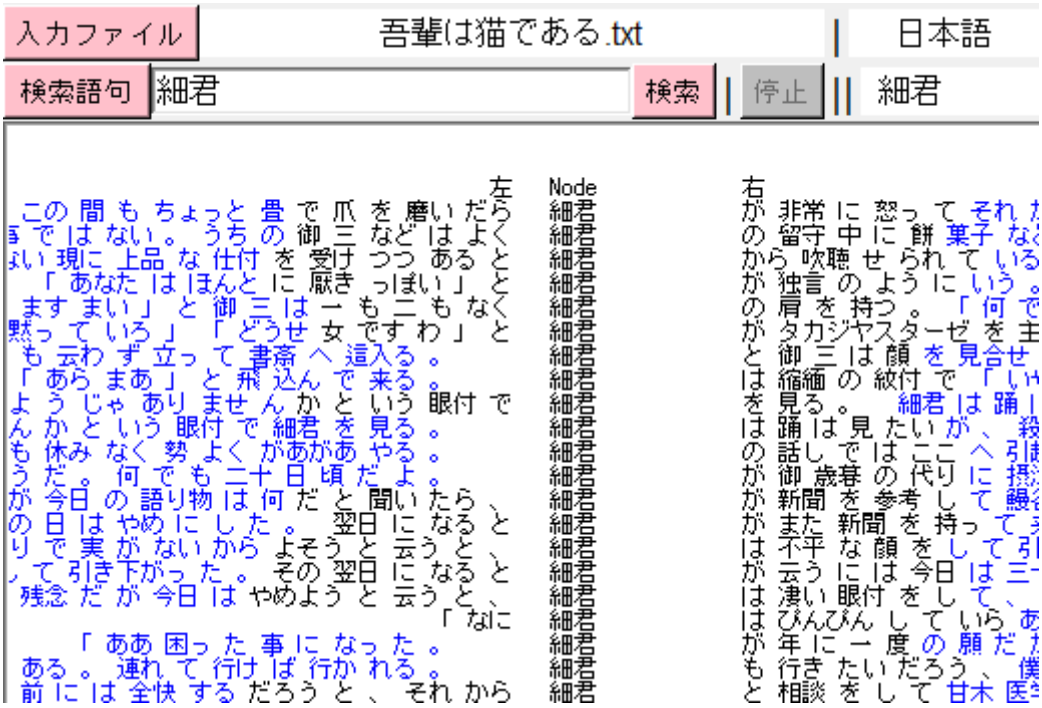
TreeTagger のタグを活用形としての共通部分でまとめたもの、本ソフト独自判別

```
Alice was  beginning to  get  very     tired    of  sitting by  her sister on  the bank ,
---   past -ing      --- base unmarked unmarked --- -ing    --- --- ---    --- --- ---  ,
```

検索

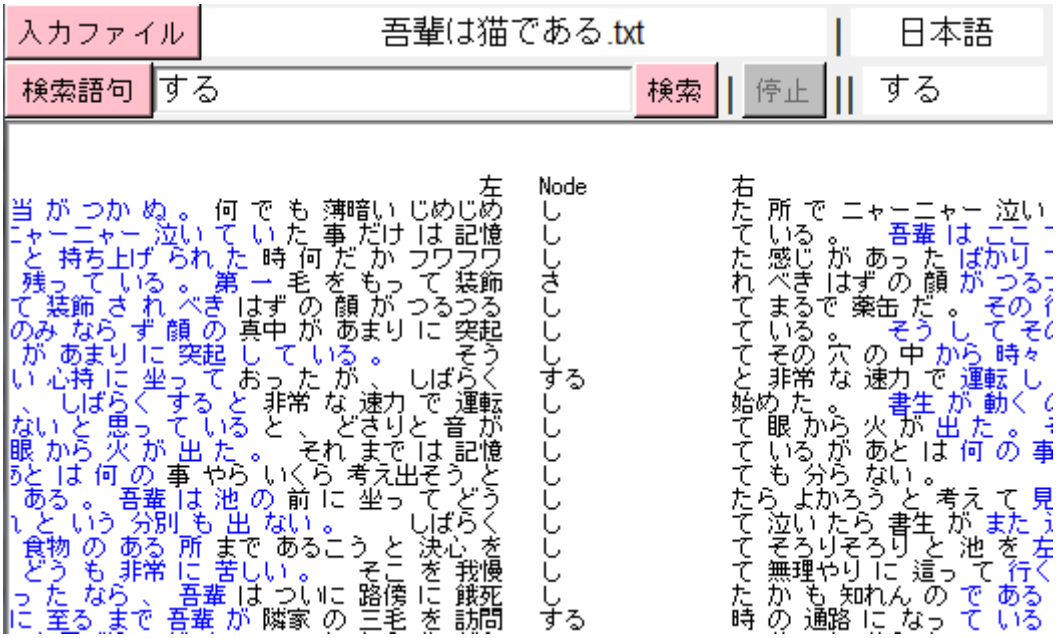
通常検索

通常の検索は、各処理のメインのウィンドウの「検索語句」の右の欄に検索する文字列を指定してから「検索」ボタンで行います。



複数の語の指定をする際は「 | 」で区切ります。

この欄に入力した文字列は単語の基本形として扱われます。つまり、活用する語が指定された場合にはその活用形が全て検索されます。

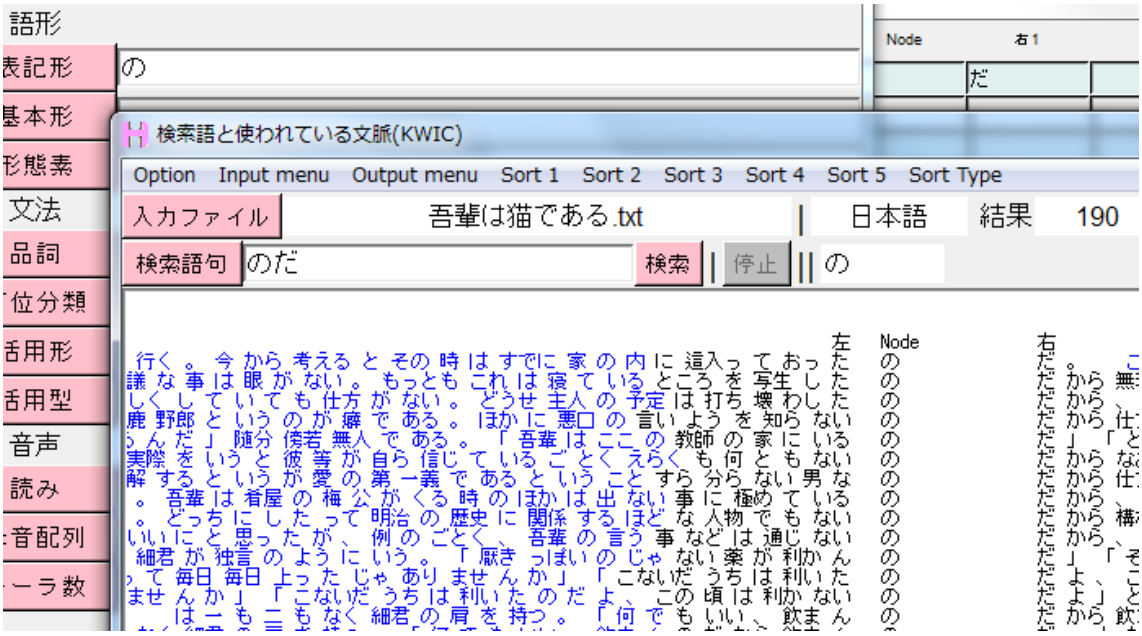


ただし、指定された語が基本形として１例も使用されていなかった場合、次にそれを表記形として使用の有無を確認し、有れば検索が開始されます。

指定した検索語が基本形としても存在するが、基本形としてではなく表記形として扱ってほしい時があります。その際は、検索語のすぐ後ろに半角のスペースを入れて検索します。

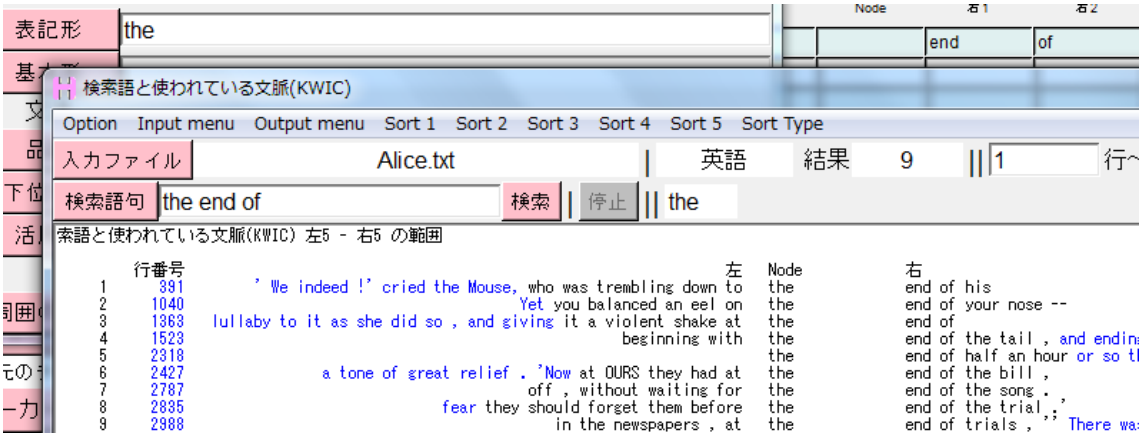
するとその語は表記形として扱われます。

本ソフトでは、Grep 以外の処理では検索は語や形態素の単位でのみ行います。したがって、語をまたいでの検索は特別な指定が必要になります。しかし、これを簡易に行うことができます。検索語句の右の欄に２形態素以上の文字列を入力して「検索」をすると、自動で検索語が複数の形態素に分割されて検索されます。

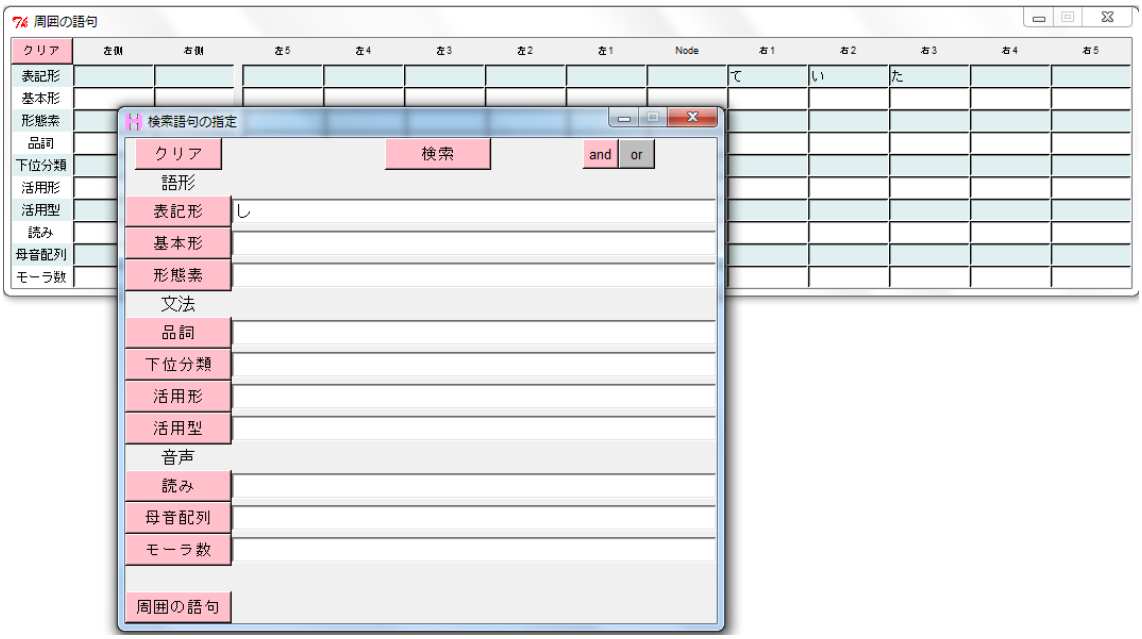


ただし、これは検索語句の文字列を形態素解析しているのではなく、検索テキスト中に一致する形態素の並びが有るかを確認するという仕組みで行われているため、指定した文字列の並びが元々テキスト中に無いと検索語句の分割はされません。また指定した語が複数形態素にまたがる場合、各形態素の最初から最後までが完全に一致していないと分割されません。つまり「歩い て いく」を検索する場合は「歩いていく」であれば分割されて検索できますが、「歩いてい」では分割されずに検索されません。

英語などの元々分かち書きされている語の場合検索語句の右の入力欄に半角スペースを空けて単語を入力すれば、最初から自動で分割されます。その結果がテキスト中に無かった場合でも必ず分割は行われます。



検索語句の自動分割がされる際に２つのウィンドウが出現し分割した形態素が収まります。



次以降でこれについての説明をします。

検索語句の詳細指定

本ソフトでは言語ごとに違いますが最大 10 項目のタグが自動で付与されます。日本語であれば「表記形」「基本形」「構成形態素」「読み」「品詞」「品詞下位分類」「活用形」「活用型」「読み」「母音配列」「モーラ数」です。表示や集計でこれらの項目を切り換えることができましたが、検索の際にもこれらの項目の全てを使うことができます。

通常の検索の際にメインの検索語句入力ボックスに指定した語は基本形や表記形、自動分割など、状況に応じて条件が変わり索されますが、基本形としては扱われたくない場合や、分割される語の区切りを自分で決めたい場合、語形ではなく品詞などの他の項目も検索に利用したい場合などがあります。その際に検索語句の詳細指定を行います。

ウィンドウの上部の「検索語句」の文字自体がボタンになっていますので、これをクリックするとウィンドウが現れます。

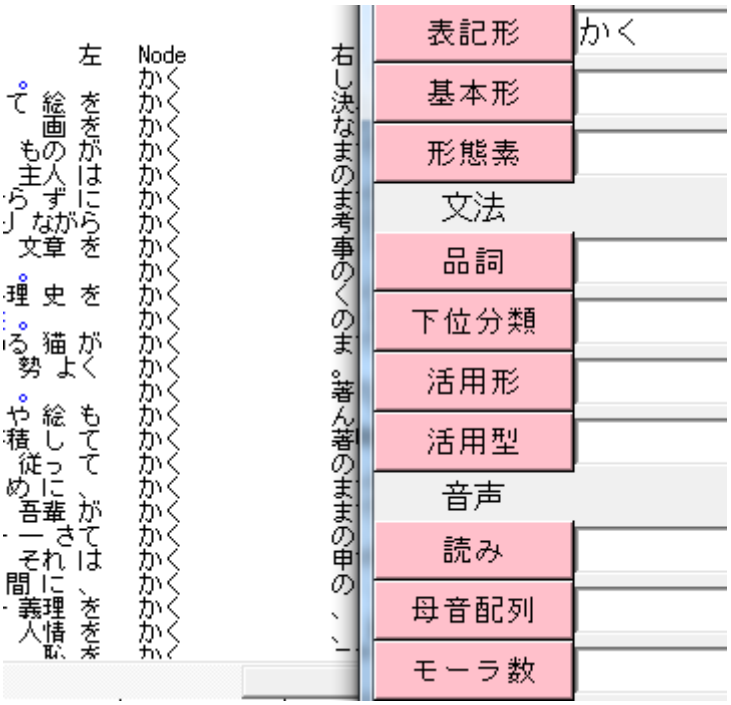
ここに、現在指定している分析言語で使用できる項目が並んでいます。各項目名の右の欄に検索文字列を指定します。1 つから全部の項目を単独または組み合わせて指定できます。検索語句の指定後、このウィンドウの「検索」を押して検索を開始します。このウィンドウに付いている「検索」ボタンはこのウィンドウで指定された文字列しか対象としません。逆にメインのウィンドウの「検索」ボタンではこのウィンドウの指定は無効になります。

具体的な検索方法として、例えば表記形に「し」と入力して検索すると、表記形が「し」の語が全て検索されます。サ変動詞「する」の未然形「し」も、接続助詞「し」も区別せずに検索されます。
また、基本形に「する」と入力して検索をすると、基本形「する」の語が全て選ばれますので、その活用形である「さ」「し」「すれ」「せ」などが全て検索されます。
ここで、この２つの条件を組み合わせて「表記形」に「し」、「基本形」に「する」と入力すると、検索の条件が２つになり、その両方が適う結果、つまり基本形が「する」の語のうちの活用形「し」のみが検索結果として選ばれます。
その他の文法的な項目は語形ではなく、文法の要素名を入力します。例えば、「品詞」に「名詞」と指定すると名詞の語のみが全て検索され、「活用形」に「連用形-一般」と指定すれば連用形-一般の語のみが検索されます。

各項目での検索の実例

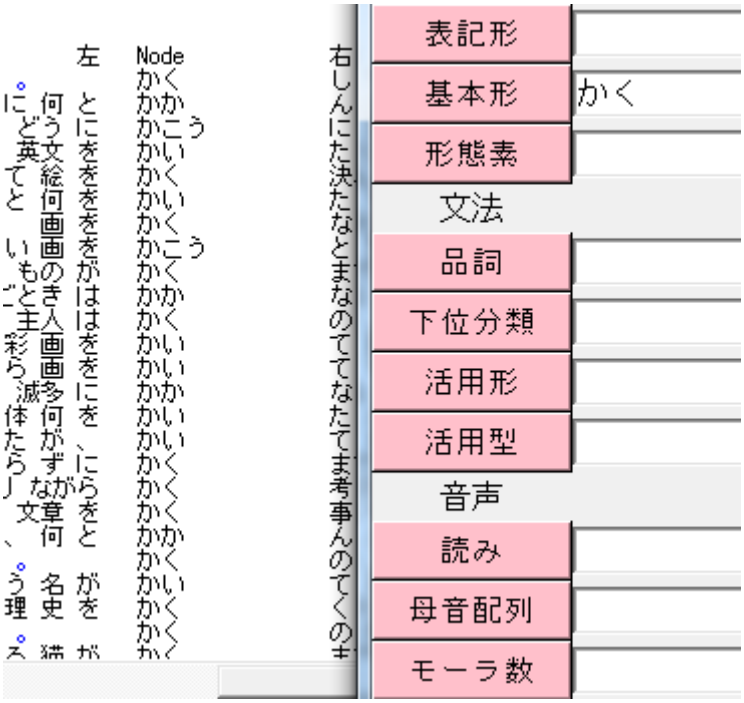
以下に、項目ごとでの検索の実際の例を示します。

表記形



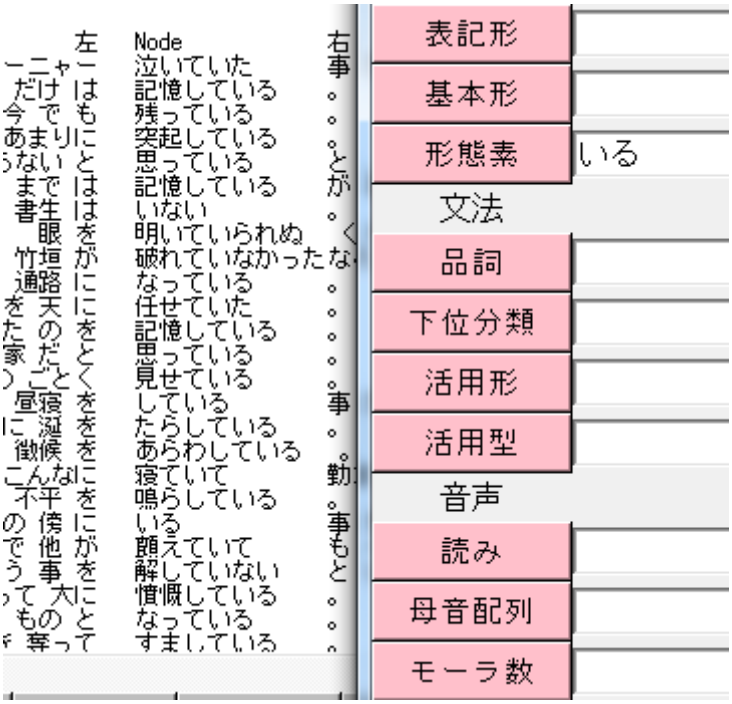
表記形は、テキスト中の文字列そのままなので、指定した文字列と完全に一致する語を検索します。

基本形



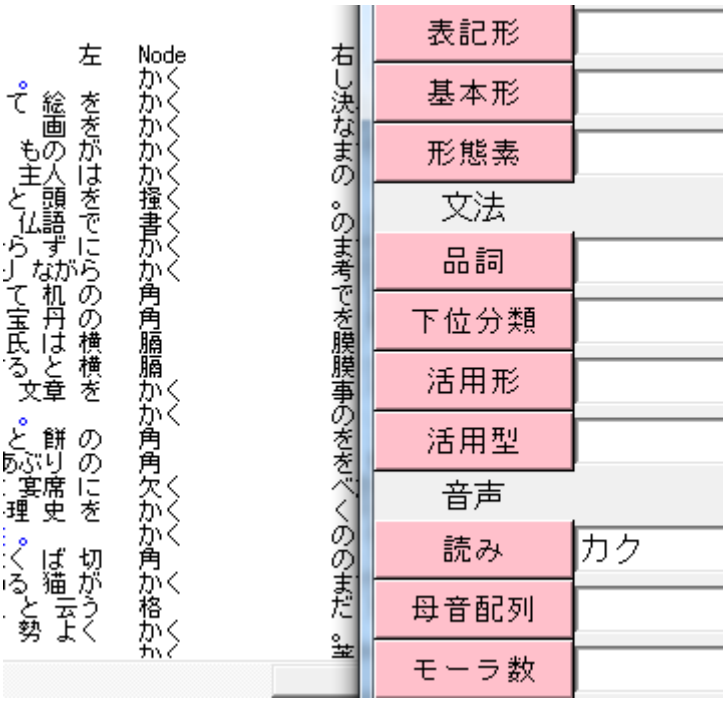
基本形は、語の活用がされていない形なので、指定した語の全活用形が検索されます。

構成形態素



形態素は、構成形態素の意味で、１語の中に存在する複数の形態素を個別に指定し検索します。整形単位が「語単位」の時のみ意味を持ちます。画面も語単位のときの例です。これを使えば、１語の中に内在する補助動詞や接辞のみを指定し、語全体を検索できます。

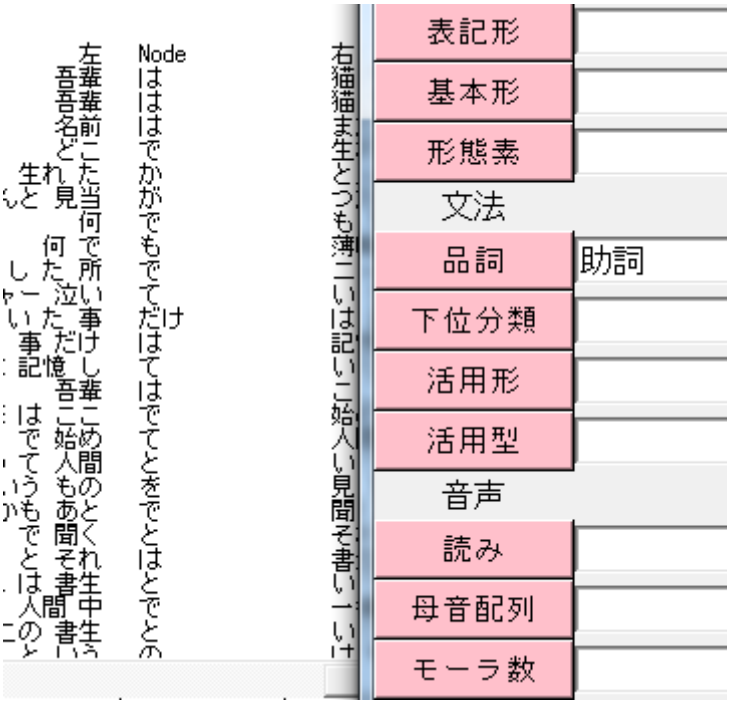
読み



読みは、それぞれの語の読みをカタカナで指定します。読みが同じであれば、元の語の違いや活用形の違いなどを一切問わず検索されます。

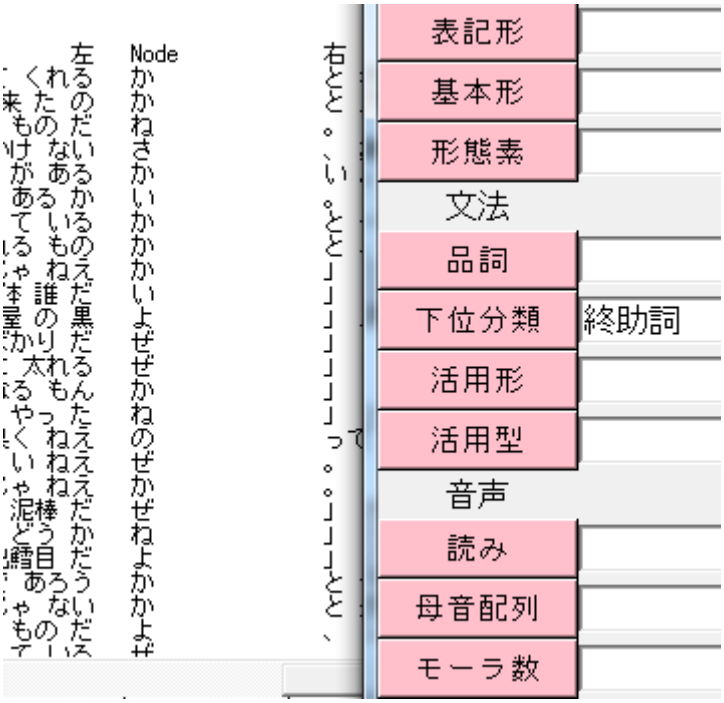
以上の項目は、項目名のボタンを押すと文字列を複雑なルールとして指定できる正規表現を簡単に生成するための機能が使えます。後で詳しく説明します。

品詞



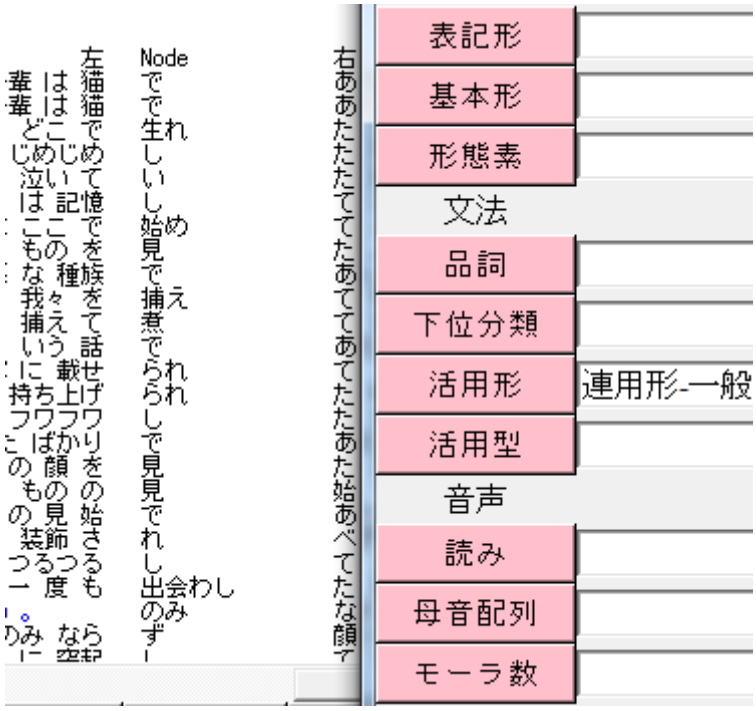
品詞は、語の品詞のうち一番の上位分類です。例えば「名詞 普通名詞」であれば「名詞」の部分です。これを指定すれば「普通名詞」「固有名詞」などに関わらず名詞が検索されます。

下位分類



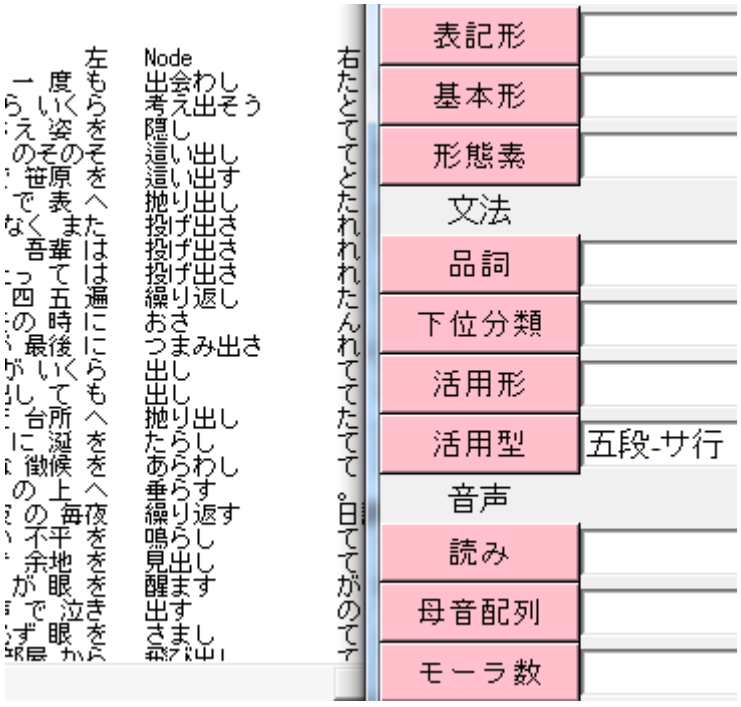
下位分類は、品詞の第２分類以降のことで、「助詞 接続助詞」の「接続助詞」の方になります。ただし、「名詞 普通名詞」は、実際は「名詞 普通名詞 一般」などのように、第３分類、第４分類など更に下位分類に分かれるものもあります。この第２以降の全ての下位分類を一括で扱うのが下位分類になります。

活用形



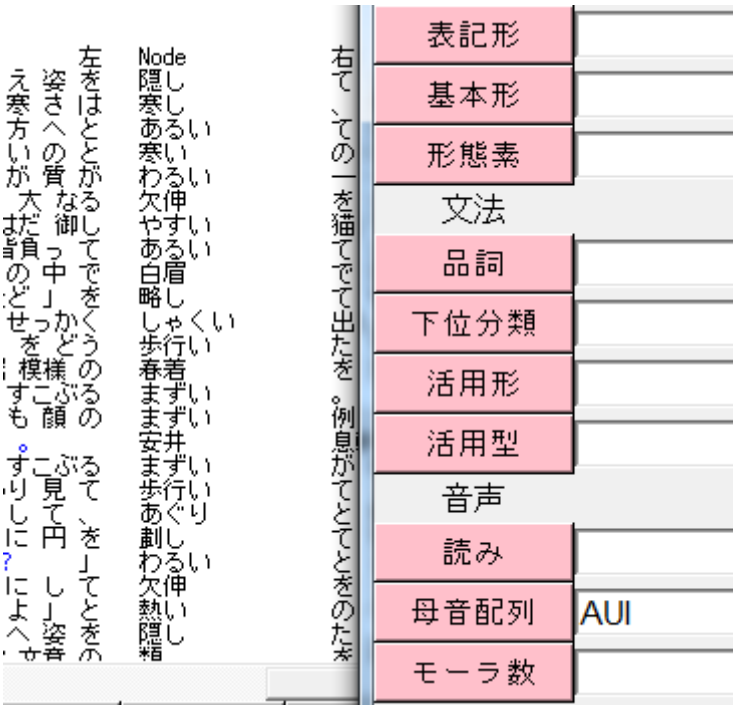
活用形は、「未然形」や「連用形」など、語の活用された形の名前です。これも「未然形-一般」や「連用形-イ音便」などのように下位分類がありますが、１項目の中で上位分類、下位分類を含んで扱います。

活用型



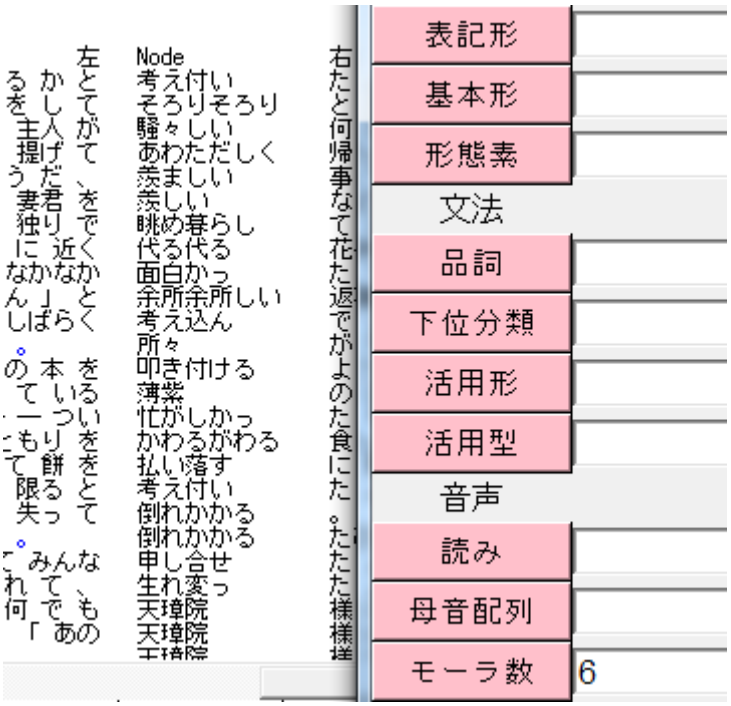
活用型は、「五段」や「サ行変格」など、語自体の活用されるルールの名前です。これも「五段-ワア行ー一般」「下一段-ア行」などのように下位分類がありますが、１項目の中で上位分類、下位分類を含んで扱います。

母音配列



母音配列は、語の発音のうち、母音だけを抜き出したものです。母音
「AIUEO」
と、特殊拍の
「-(長音)」
「Q(促音)」
「N(撥音)」
の組み合わせで指定します。

モーラ数

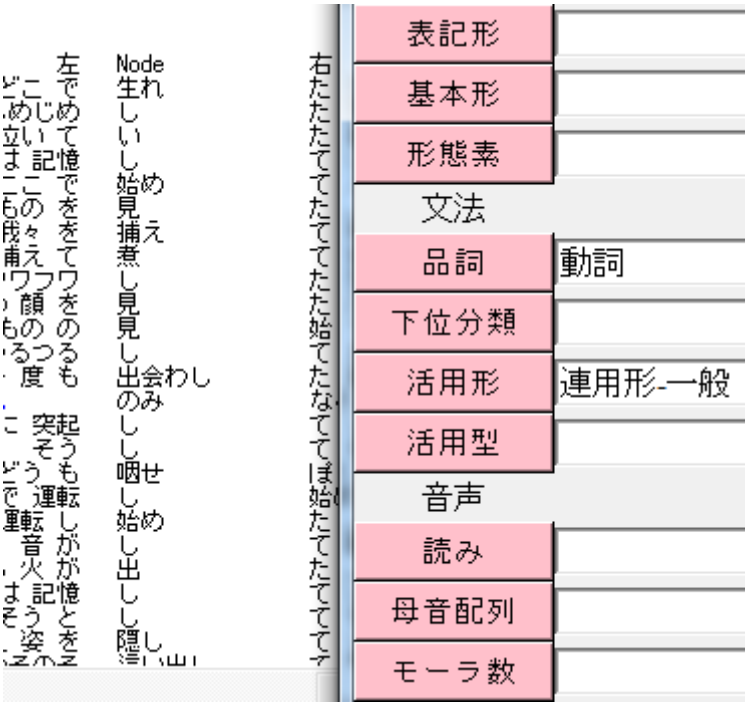


モーラ数は、各語のモーラの数になりますが、具体的には母音配列の母音と特殊拍の文字数になります。

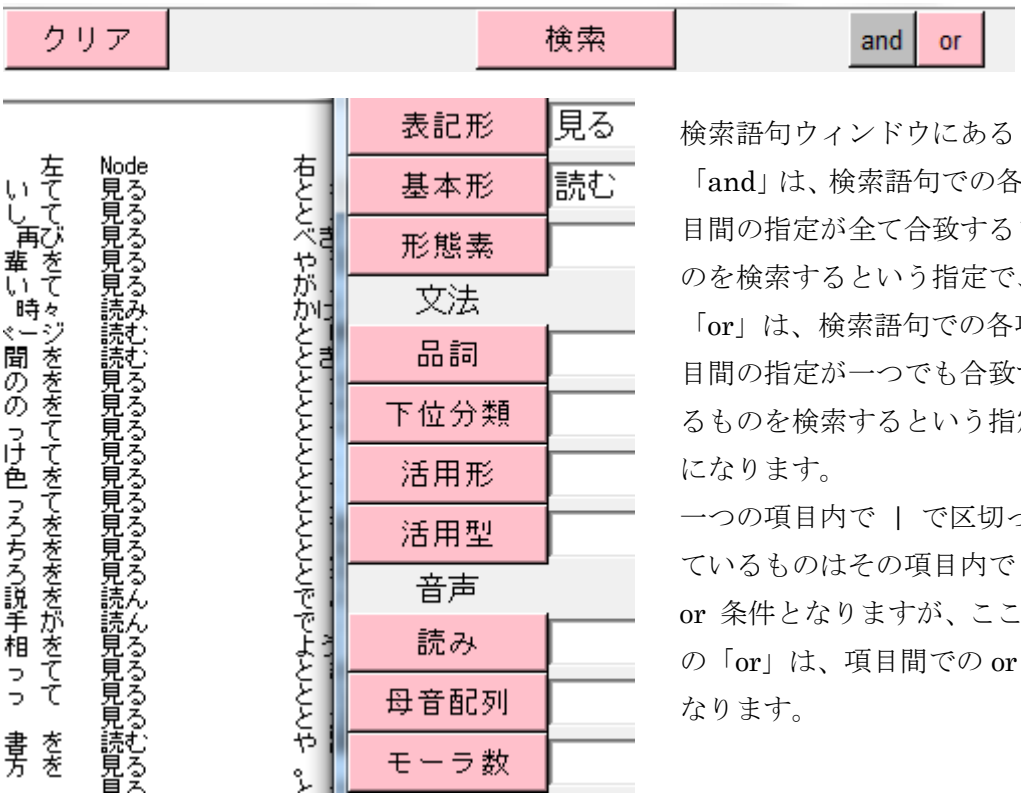
項目名のボタンを押すと現れるウィンドウのリストから選択するだけで指定できる簡易選択ができます。後で詳しく説明します。

複数条件の指定

複数条件で指定すれば指定した項目が全て合致する語のみが検索されます。



and 条件、or 条件



検索語句ウィンドウにある「and」は、検索語句での各項目間の指定が全て合致するものを検索するという指定で、「or」は、検索語句での各項目間の指定が一つでも合致するものを検索するという指定になります。

一つの項目内で ｜ で区切っているものはその項目内で or 条件となりますが、ここでの「or」は、項目間での or になります。

入力方法

簡易選択

「品詞」や「活用形」など、言語によって決まった数の要素しかない項目は、入力ボックスに直接文字列を入力する以外にリストから簡易的に選択することができます。それぞれの項目名のボタンを押すとその項目で扱える要素のリストが出ます。

・リストにある要素をダブルクリックするとその要素が項目指定に単独で選択されます。
・シングルクリックすると、 | で区切られて複数選択できます。
・Ctrl + クリックすると、「指定した要素以外」という選択方法になります。

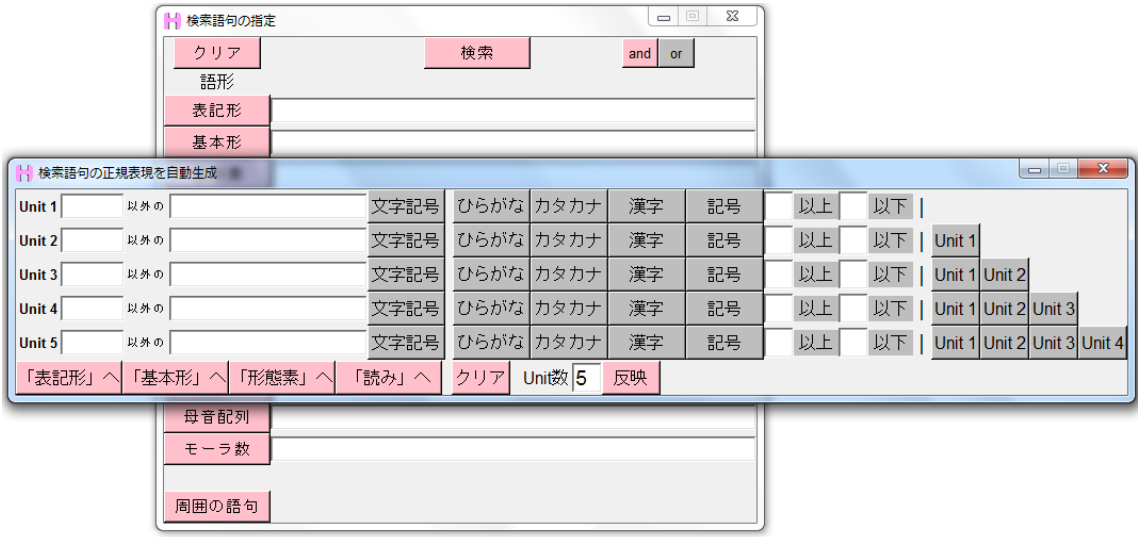
それぞれの項目のリストの一番左から始まっている文字列は各要素の上位分類で、スペースが空いて右にずれて始まっているのがその下位分類になります。上位分類を選べばその下位分類がすべて選択されたと同じことになります。

品詞だけは「最上位分類」と「それ以降」に分かれて保持されていて、「品詞」が最上位分類、「下位分類」が第二分類以降となります。「品詞」と「下位分類」は簡易選択リストが共有され連動していますので、どちらを押しても同じリストが出ます。最上位分類を選択したときだけ「品詞」のエリアに語が入力され、それ以降を選択した場合は「品詞」と「下位分類」それぞれのエリアに語が入力されます。

シングルクリックの場合、基本的には | で区切られて複数選択ですが、「母音配列」と「モーラ数」は選択項目がそのまま連なって１つのものとして指定されます。

正規表現自動生成

「表記形」や「読み」など、要素の数に上限が無い項目は、項目名のボタンを押すと正規表現の簡易的な自動生成ウィンドウが開きます。

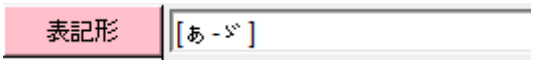


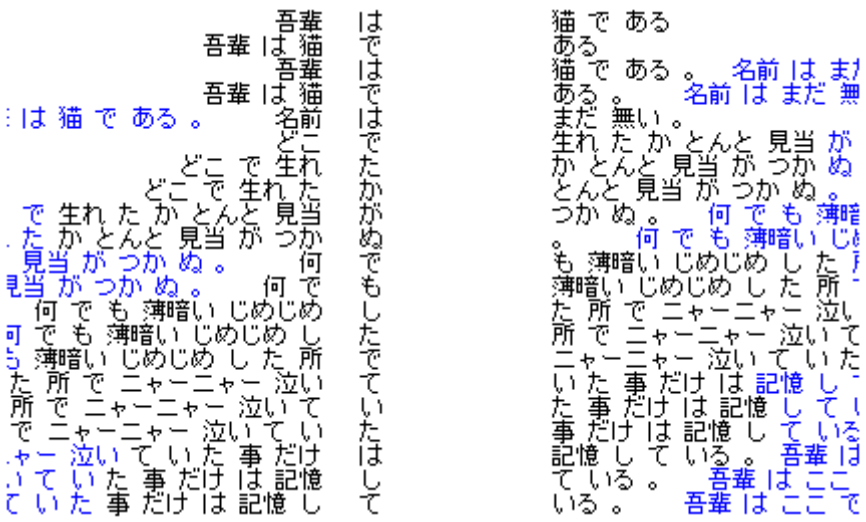
正規表現の生成ウィンドウは横列がそれぞれのユニットになっていて、Unit 1 ～ 5 に分かれています。各ユニットは独立して条件を指定できます。Unit1 が最終的に生成される正規表現の一番左側になり、次から順に右側の文字列になります。
それぞれのボタンを押すと自動で指定される正規表現文字列は以下のようになります。

「文字記号」	¥S
「ひらがな」	[ぁ-ゞ]
「カタカナ」	[ァ-ヾ]
「漢字」	[一-龥]
「記号」	[、-〓]
「以上」	{N,}
「以下」	{,N}
「Unit1」～「Unit5」	¥1～¥5

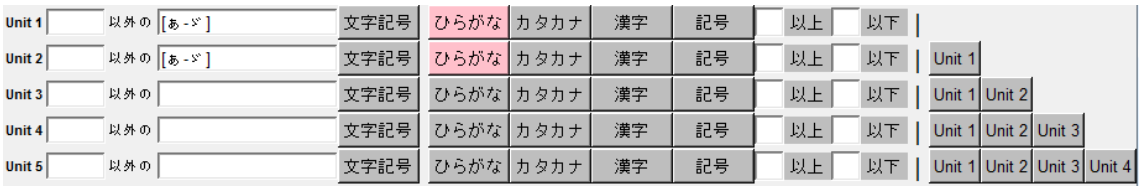
正規表現の作成ルールの指定ができたら、「表記形へ」など、指定したい項目名のボタンを押すと最終的な正規表現となって指定エリアに入力されます。
以下に押されるボタンなどの指定と、生成された正規表現の文字列と、それによる検索結果のいくつかを提示します。

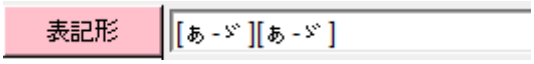
ひらがな１文字

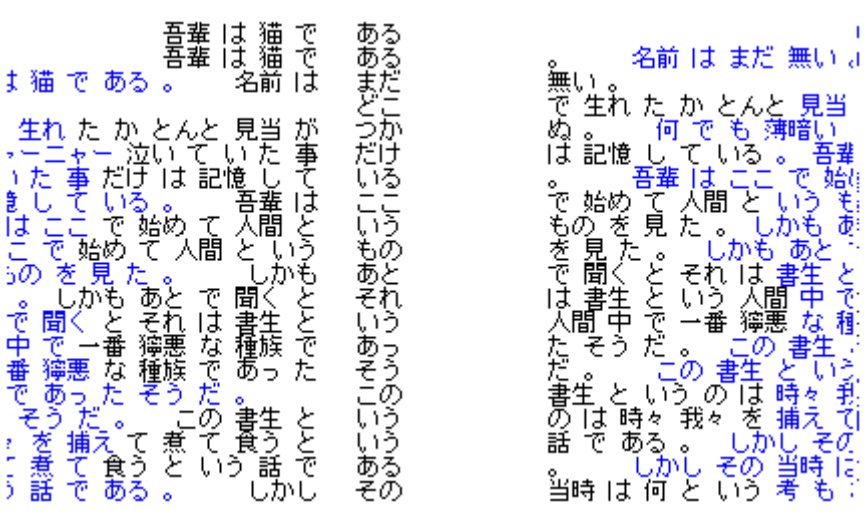




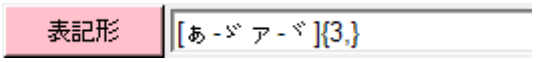
ひらがな１文字、ひらがな１文字

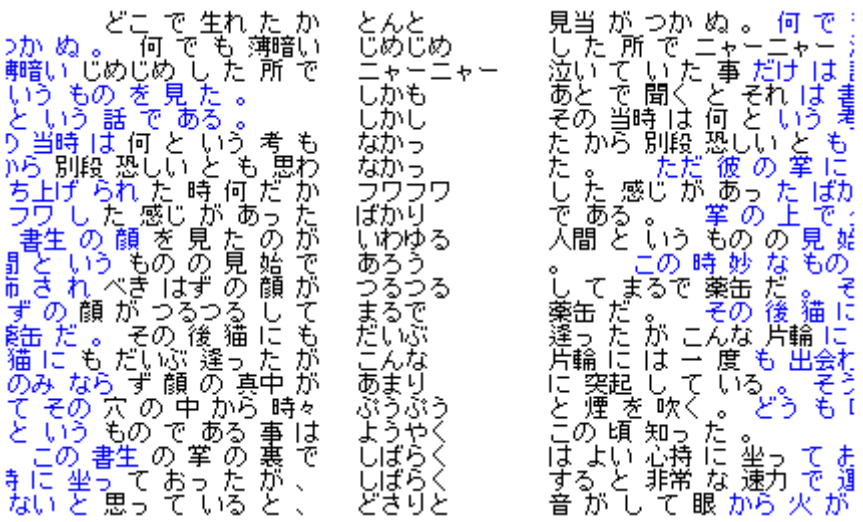




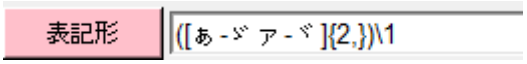


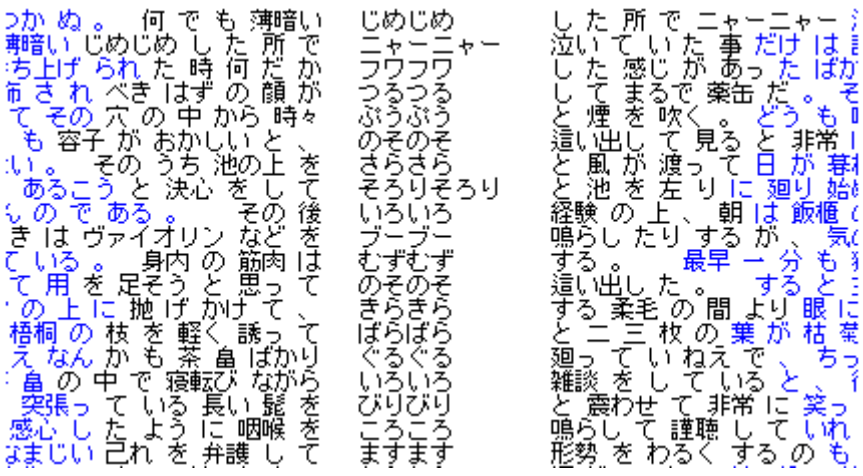
ひらがなかカタカナ３文字以上



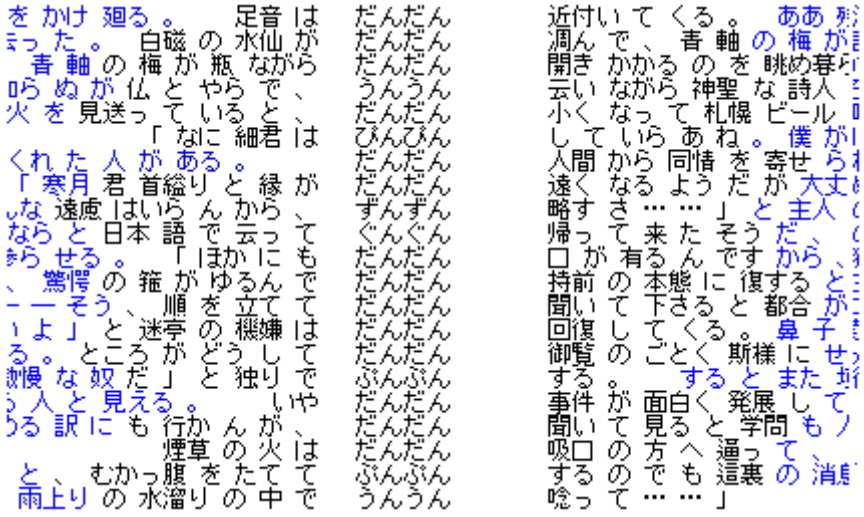


ひらがなかカタカナ２文字以上、その２文字以上のものと全く同じ文字列が１回

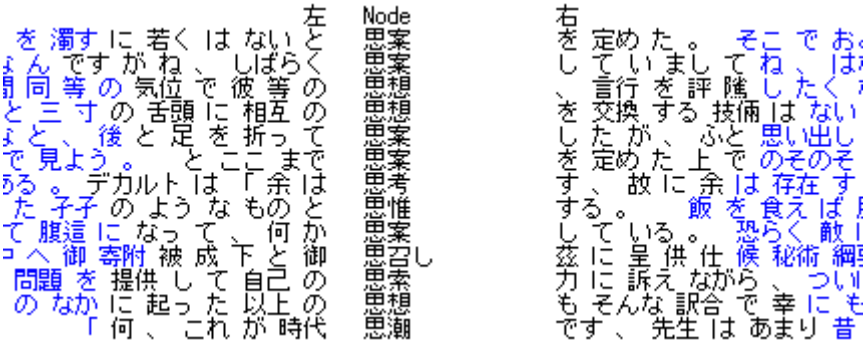




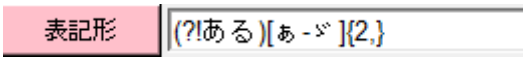
ひらがな１文字以上、「ん」最初のひらがな１文字以上と同じ文字列、「ん」と同じ文字列

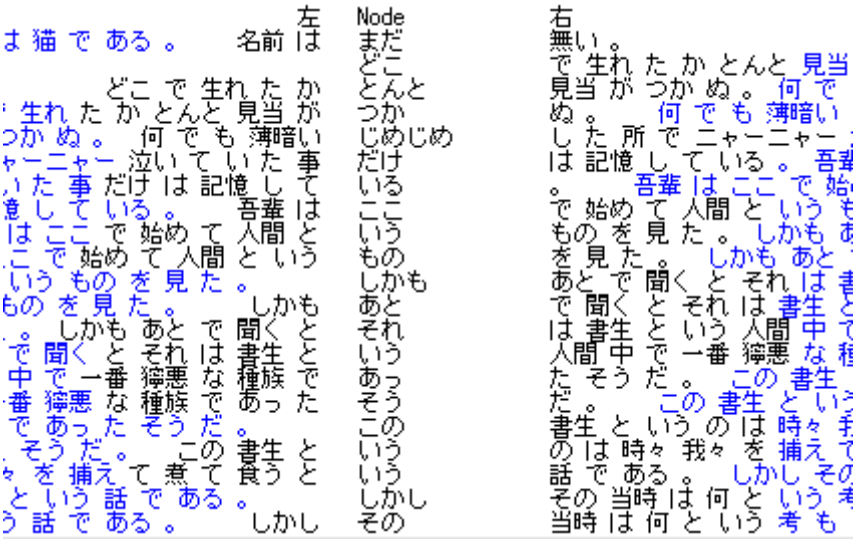


「思」、漢字１文字以上、ひらがな０文字以上



「ある」以外のひらがな２文字以上





Unit 数の変更

Unit 数が５つでは足りない場合はウィンドウ下部の「Unit 数」の数字を変更し、「反映」ボタンを押すことで増やしたり減らすことができます。

Grep での特殊ボタン

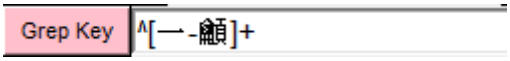
Grep では、形態素分けされていない文字列を検索するので、正規表現が更に効果的に扱えます。そのため Grep のみ使える自動生成用のボタンが用意されています。

特別なボタンは以下の通りになります。

「行頭」　　指定した文字列がテキスト中の行頭にあるときのみ
「行末」　　指定した文字列がテキスト中の行末にあるときのみ
「前に来る」　規定の入力欄の文字列が、生成する正規表現の前に来る文字列のみ
「後ろへ来る」　規定の入力欄の文字列が、生成する正規表現の後に来る文字列のみ
「前に来ない」　規定の入力欄の文字列が、生成する正規表現の前に来ない文字列のみ
「後ろへ来ない」規定の入力欄の文字列が、生成する正規表現の後に来ない文字列のみ

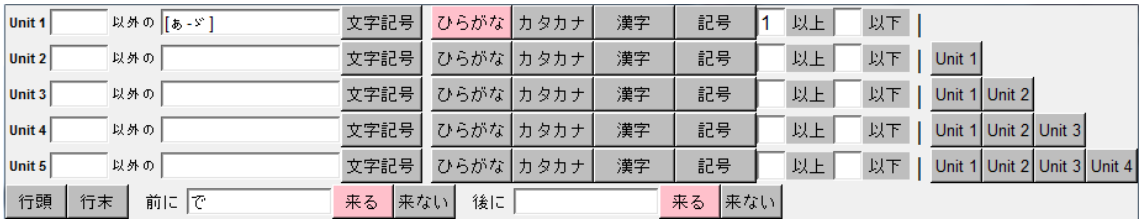
以下、使用の具体例を示します。

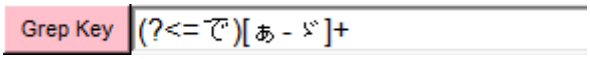
行頭にある漢字１文字以上

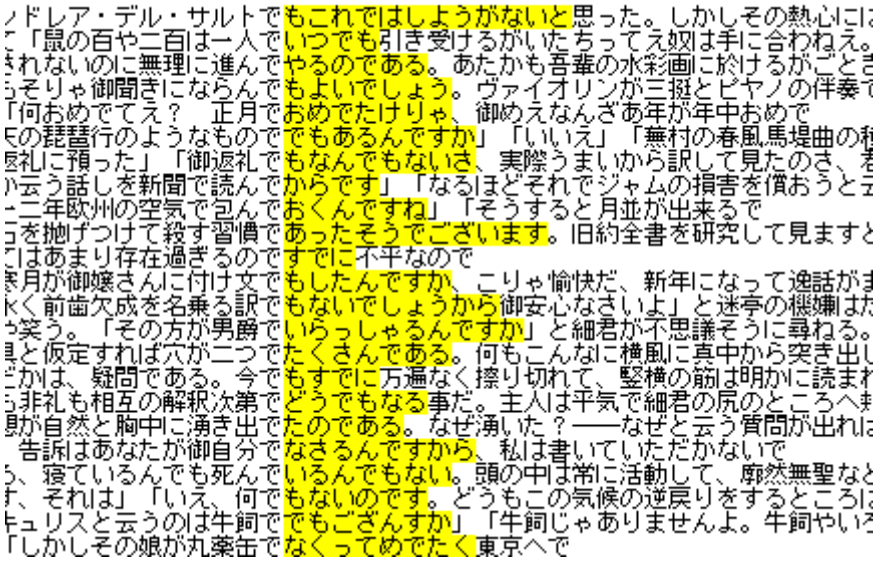



```
  1 吾輩は猫である
  2 夏目漱石
  5 一
 36 昨夜は僕が水彩画をかいて到底物にならんと思って、そこらに抛って置いたのを誰か
 45 二
 57 寒月と、根津、上野、池の端、神田辺を散歩。池の端の待合の前で芸者が裾模様の春
 61 宝丹の角を曲るとまた一人芸者が来た。これは背のすらりとした撫肩の恰好よく出来
 65 神田の某亭で晩餐を食う。久し振りで正宗を二三杯飲んだら、今朝は胃の具合が大変
246 三
280 四
491 五
709 六
736 世の人に似ずあえかに見え給う
742 倦んじて革ずる香裏に君の
743 霊か相思の煙のたなびき
764 七
826 八
1012 九
1034 拝啓愈御多祥奉賀候回顧すれば日露の戦役は連戦連勝の勢に乗じて平和克復を告げ吾
1038 時下秋冷の候に候処貴家益々御隆盛の段奉賀上候陳れば本校儀も御承知の通り一昨々
1040 大日本女子裁縫最高等大学院
1041 校長　　縫田針作　九拝
```

「で」の直後に来るひらがな１文字以上



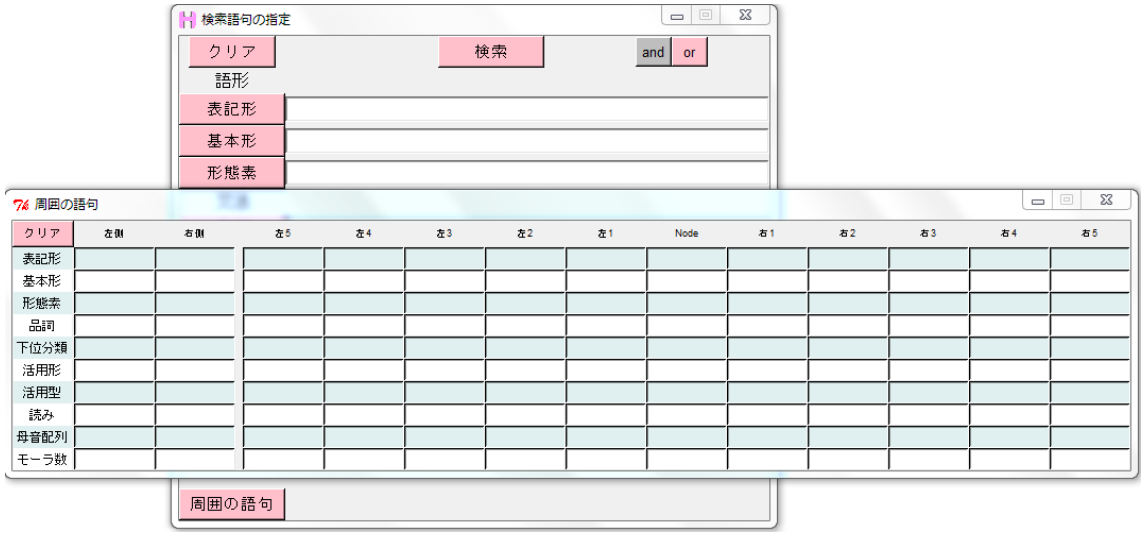




これ以外にも、本ソフトでは全ての入力ボックスは、自分で打ち込んだ正規表現を認識しますのでごく簡単な指定や更に複雑な指定の場合、直接正規表現を打ち込んで検索できます。

周囲の語句の指定

検索結果を更に詳細に条件づけるために、検索結果の周囲の語句の指定も行えます。



検索語句のウィンドウの「周囲の語句」ボタンで出現するウィンドウで指定します。周囲の語とは、検索語自体の位置と、その左右の取得幅の語数分の位置です。通常では左 5 から右 5 の範囲になります。この範囲の左右の語で、分析言語ごとに扱えるタグの分だけ項目を指定できます。

基本の検索語句に加えて、周囲の語句として項目での指定した文字列がある結果のみに絞られます。

「品詞」などの文法項目での指定もできます。

検索語句と別の項目でも構いません。指定は通常の検索語句と同じように行います。

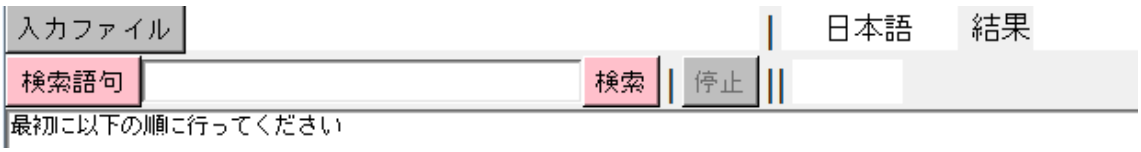
ただし、正規表現の自動生成や簡易入力は行えません。

検索のメカニズム

検索語句は、メインのウィンドウでの検索と、検索語句ウィンドウでの検索とで、入力する場所によって２つの検索があります。それぞれ「柔軟検索」と「詳細検索」とします。

柔軟検索

まず、それぞれのウィンドウに検索語句の入力エリアと「検索」ボタンがあります。これで指定する検索は「柔軟検索」です。



柔軟検索の検索ボタンは、同じく柔軟検索の入力エリアに指定された語句を検索条件として検索を行います。その際に詳細検索の入力エリアに入っている全ての語句が消されます。

柔軟検索の際の内部処理の手順

① 詳細検索の検索語句と周囲の語の指定を全て消去
② 柔軟検索の指定語句が半角スペースで区切られていれば自動で検索語を分割。分割された最初の語を詳細検索の表記形へ、それ以下を周囲の語の表記形の左側へ順に指定。その際に検索語の指定が１つの文字列のみで直後に半角スペースが有れば、検索語を単独で表記形として扱う。⑦検索開始へ
半角スペースでの区切りが無ければ③へ
③ 柔軟検索の指定語句を、詳細検索の基本形へそのままの形でコピー
④ 単語リストにその語が有るか確認。有れば⑦検索開始へ、無ければ⑤へ
⑤ 検索語句の自動分割が可能か確認。テキスト中に、検索語句がいくつかに分解されたとして、それと同じ並びでの語があるか調べる。一致する語の並びが有ればその語の区切りと同じ個所で検索語句を分割し、分割された最初の語を詳細検索の表記形へ、それ以下を周囲の語の表記形の左側へ順に指定。有れば⑦検索開始へ
⑥ 検索語分割処理の際に、柔軟検索の指定語句が１語で表記形として存在することが分かれば詳細検索の表記形に指定。有れば⑦検索開始へ
⑦ 検索開始

の順で処理されます。

処理メニューによる動作の違い

Sentence、KWIC、Collocates、Picture、POPAK という、語を検索してその結果を元に行う処理４つと、全文表示の Sentence では、柔軟検索の語句の指定が無い時に詳細検索のどれかにでも指定が有った場合は、詳細検索に切り替わります。また、これらの処理では語の並びが意味を持つので検索語句の自動分割が行われます。

Freq、N-gram、Keyness、Mark、Edit という、語の検索が行われなくても基本処理が行われるものでは処理過程が多少変わります。

まず、柔軟検索の基本的な流れは同じですので、柔軟検索語句として指定された語は通常では基本形として扱われます。

ただし、特殊な条件として、これらの処理では、文脈ではなく１語１語のみを対象にするため、連続した語は扱えなく、“検索語句を分割”するという概念がありません。したがって、検索語句が半角スペースで区切られている際は、一番左の文字列のみが表記形として検索語句になります。同様に検索語句の直後に半角のスペースが有る場合、検索語句を表記形として扱います。

複数形態素の文字列でも自動分割は行われません。

また、柔軟検索語句の指定が無い場合、この条件を付けなくても、全ての内容を対象にするという意味がありますので、柔軟検索語句の指定が無い場合は、詳細検索と周囲の語の全ての指定が消され、完全な初期状態として処理が行われます。

処理ごとの柔軟検索の内部手順をまとめると以下のようになります。番号は上から順に実行され、条件に一致しなければ順次下に流れます。

◎Sentence、KWIC、Collocates、Picture、POPAK

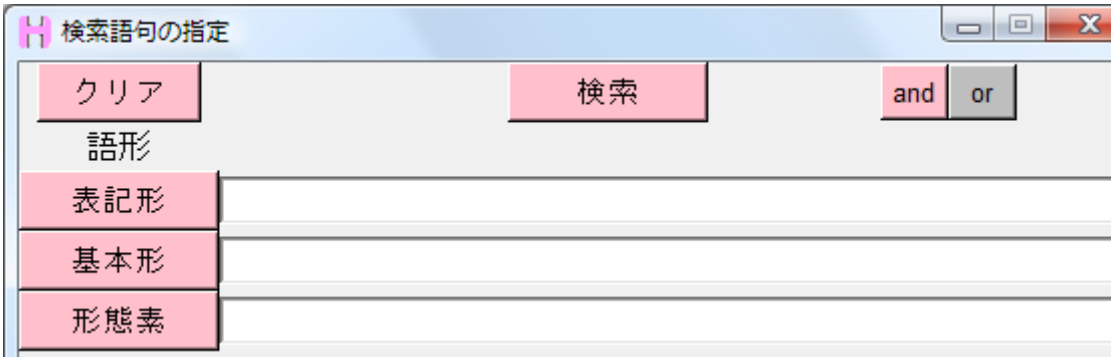
①半角スペースで終わっていれば表記形として検索
②半角スペースで区切られていれば分割して検索
③基本形として検索
④分割して検索
⑤表記形として検索

◎Freq、N-gram、Keyness、Mark、Edit

①指定が無ければ全体の実行
②半角スペースで終わっていれば表記形として検索
③半角スペースで区切られていれば一番左のみを表記形として検索
③基本形として検索
⑤表記形として検索

詳細検索

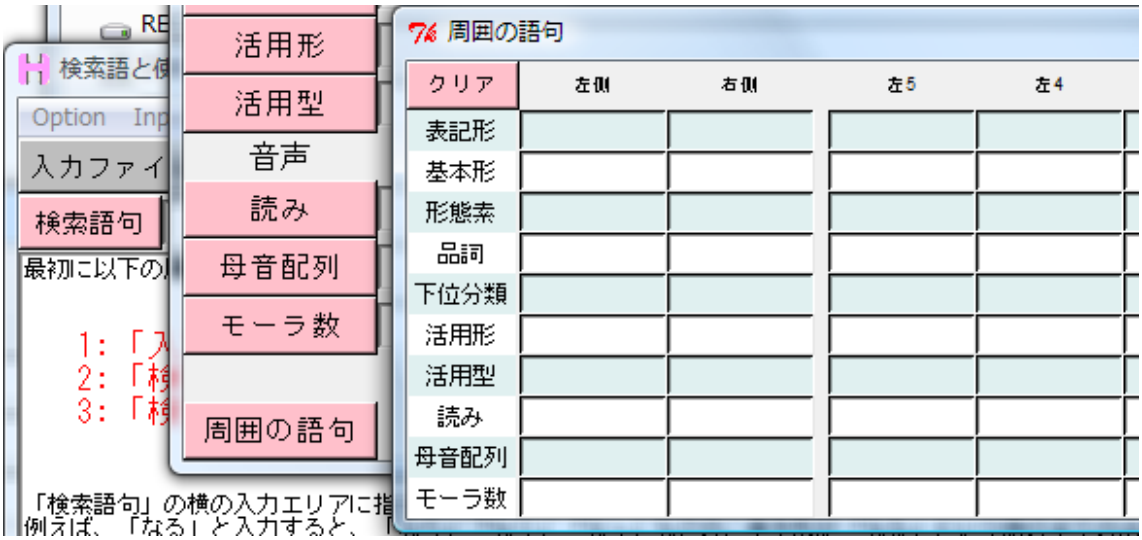
メインのウィンドウにある「検索語句」ボタンで現れる小さなウィンドウで、全ての項目を個別に指定して検索ができます。これは詳細検索です。



表記形からモーラ数まで、各タグ項目を個別に指定して検索を行える処理です。柔軟検索と違い、ここでの検索ボタンでの検索では、指定された各項目をそのまま使い、状況に応じて処理方法が変更されたりせずに指定内容通りに検索が行われます。その際に柔軟検索の指定語句は消去されます。ここで指定して行う検索は、柔軟検索のように自動では処理される先が変わらないので、より詳しく、細かく指定する際に使います。

周囲の語句

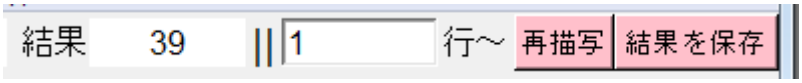
詳細検索ウィンドウから、更に検索語句の周囲の語句の指定ができます。



ここでは、各項目全てを検索語を中心として、その周囲の一定範囲分一気に指定できるものです。既に何らかの検索結果があって、更に条件を絞るために行う際は、周囲の語句を指定した後に再描写ボタンを押します。検索語句自体を新しく指定し、同時に周囲の語句を指定した場合は、検索ボタンで検索そのものをします。メインのウィンドウにある柔軟検索での検索ボタンでは周囲の語句の条件は反映されません。

再描写

検索後に何らかの条件を追加した際に一から検索を行うのではなく、その条件だけを追加した結果を表示させるのに再描写をおこないます。



主に、sort 条件や周囲の語句の指定の変更をした後に再描写ボタンで行います。特に、周囲の語句の指定を追加した後などに、「検索」を押すとまた一から検索が行われてしまうので、扱うデータが大きい場合に時間がかかってしまいます。検索語句自体以外の条件の変更のみのときは再描写を行うと短時間で処理がされます。

停止

処理を途中で止めたい場合に「停止」ボタンを押します。

この停止ボタンは、処理の状況によって使用可能、不可能が変わったり、停止後の動作が変わります。例えば、KWIC の時には内部での処理は以下の手順で行われます。

① 検索語句の有無や分割の可能性を調べる
② 検索の実行
③ 検索結果データのソートや整形
④ 画面表示

このうち、①と③の時には停止ボタンは使用できません。この処理中に停止を行うと内部データが著しく崩れてしまうためです。しかし、この２つの工程は共にそれほど長い時間がかからないため問題はありません。対して、②と④のときに停止ボタンが使えます。②の検索の実行時が一番時間がかかります。検索結果数は 1000 単位で画面の右下に加算されていきます。全ての例を見る必要はないが十分に必要な数の検索例が取得できた場合に停止ボタンを押します。すると検索がそこで止まり、そこまでの検索数で以下の処理へ移行します。④の画面表示中に停止をすると、画面表示自体がそこで止まります。再描写をすれば、取得してある検索例分だけ再度一から描写を行います。他の処理でも同様の仕組みになります。

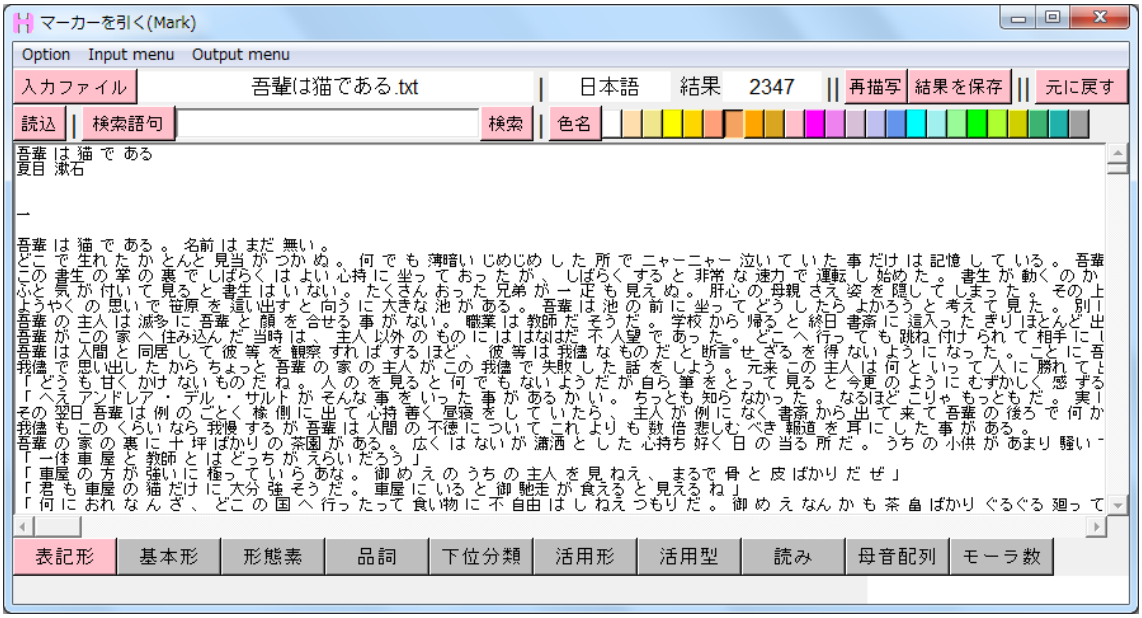
オリジナルコーパス作成

本ソフトでは自動で最大 10 の項目タグが付与されますが、これに加え、使用者独自のタグを付与してオリジナルコーパスの作成ができます。また、整形されたファイル同士を連結や結合などをして、更に大きいコーパスにすることもできます。以降ではこれらの方法を提示します。

簡易タグの付与によるデータへのマーク付け
本格的なタグの付与による完全オリジナルコーパス化
行単位のタグ付与による複雑なデータ構造の作成
音声再生による会話コーパス化
コーパス本体の修正
大規模コーパス化
語単位の作成ルールの変更による独自整形

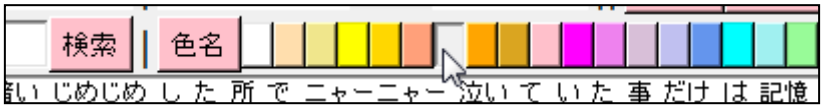
マーカーを引く(Mark)

　この処理では、本文の中に利用者自身が簡易的なタグを付け、他の処理でそのタグを利用できるようにします。



タグは色で表されます。本ソフトではこれを「マーカー」と呼びます。
テキストを指定後「読込」ボタンを押すと、本文が画面上に表示されます。

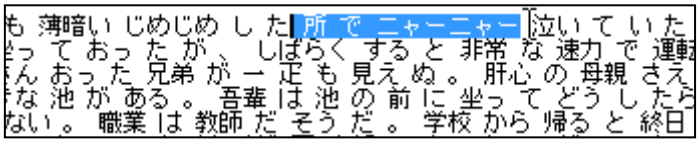
色選択



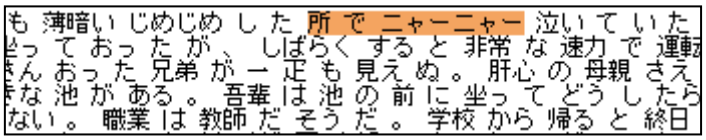
その後、ウィンドウ上部の色が配置されているパレットの好きな色をクリックすると、その色が選択されます。

マーカーの付け方

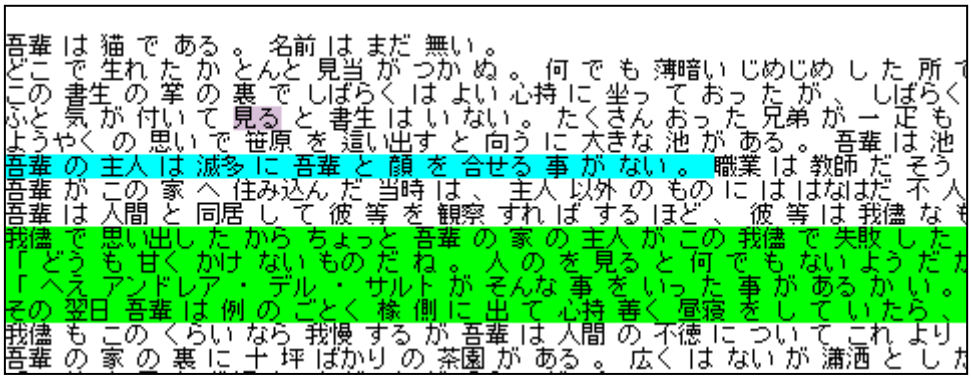
色を選択後、本文中の語をマウスで範囲選択します。



範囲指定の後、マウスのクリックを離すとその範囲の語に選択された色が付きます。

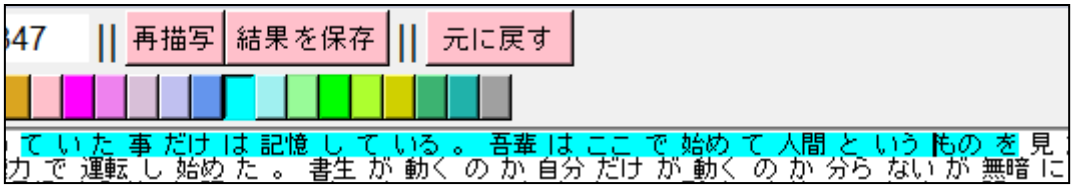


マーカー引く単位は自由に決められます。

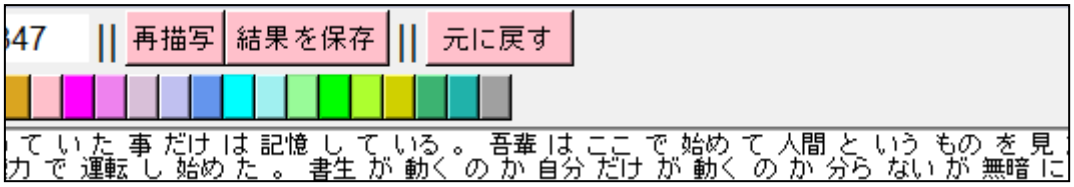


文字ごと、単語ごと、範囲ごと、段落などの大きな範囲ごとなどマウスでの範囲指定の幅だけ好きに決められます。

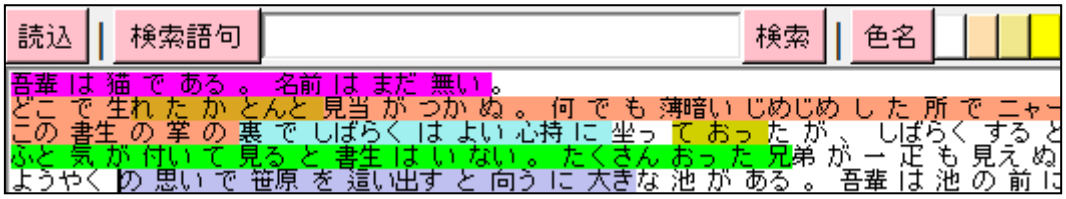
編集を１つやり直す



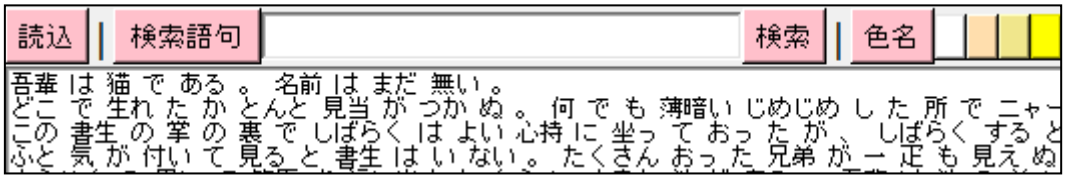
範囲指定を間違えてしまった場合、ウィンドウ上部右端の「元に戻す」ボタンで１工程だけ前に戻ることができます。



編集全体をやり直す

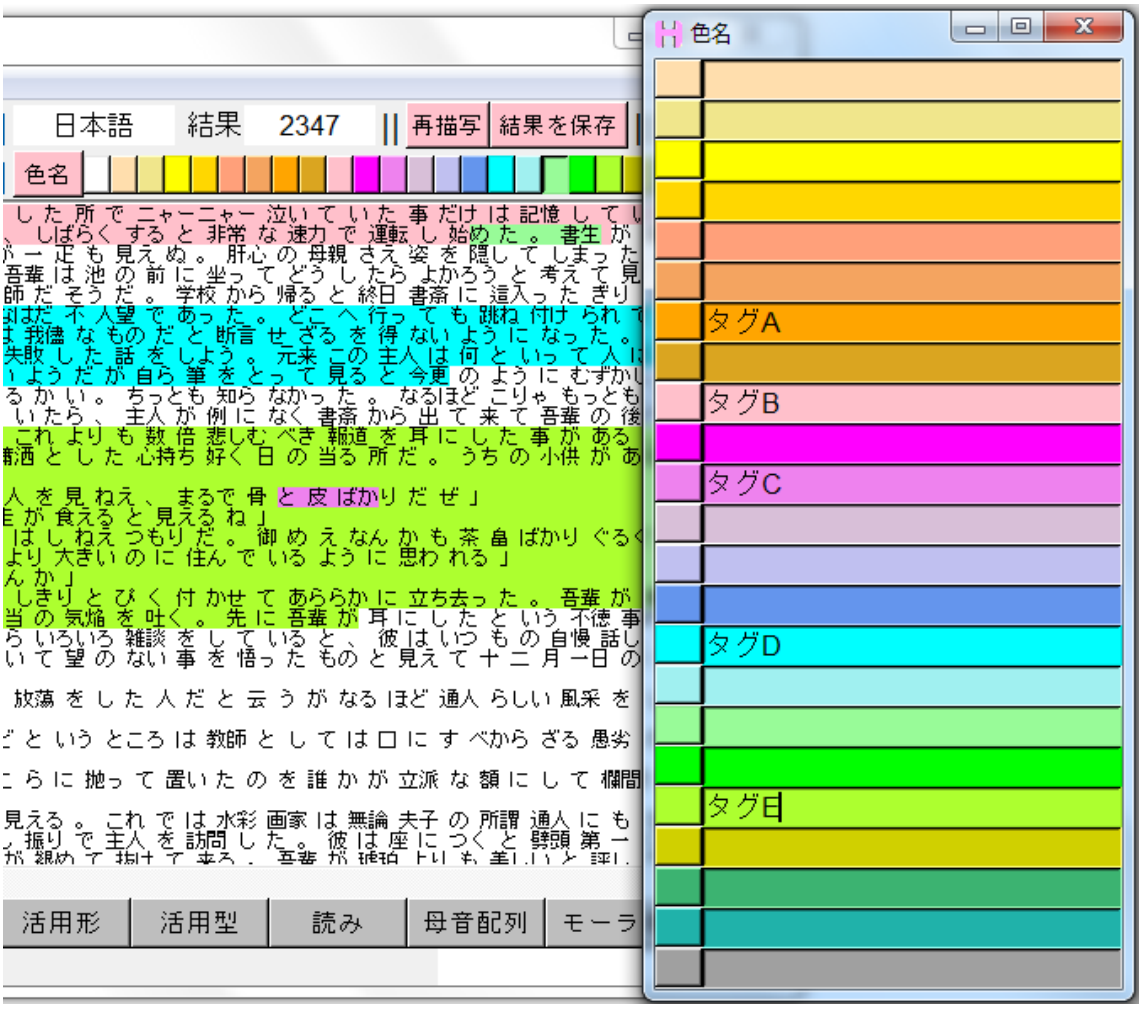


編集した内容を全てやり直す場合は、「読込」ボタンを押すと、ファイル選択直後に戻すことができます。



色に名前を付ける

タグである色に名前を付けることができます。



この処理ではタグは色で表されるので、その意味合いは使用者が独自に設定します。しかし、後からそのタグごとの意味合いを確認するためや別の使用者がそのデータを見たときにタグの意味合いを共有するのに、文字としてタグに色を付けておくと便利です。

タグの色名は、ウィンドウ上部、色の一覧のボタンの左側にある「色名」ボタンで現れるウィンドウで付けられます。

色ごとに横長のバーに文字の入力ができるようになってますので、名前を付けたい色のバーに名前を入力します。

ここで入力した名前はマーカータグの保存後に他の処理でも使用されます。

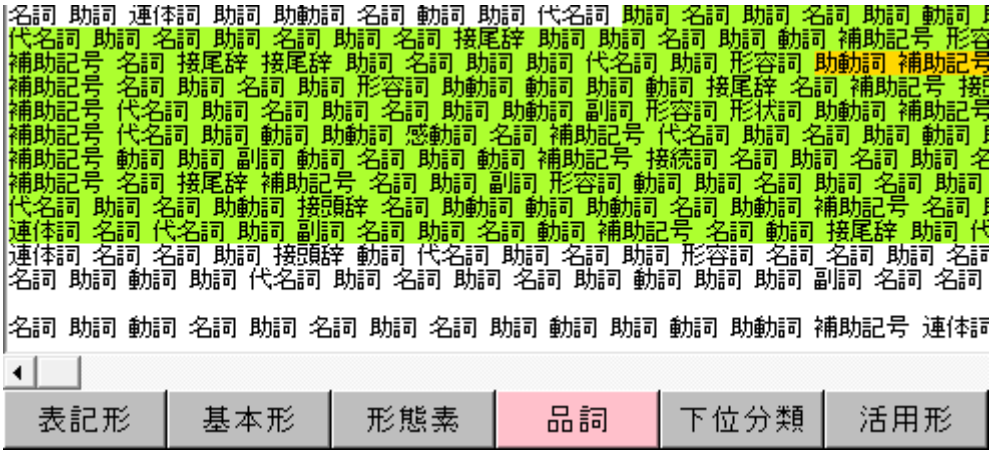
色のバーの右にあるボタンは、テキストへ付ける色の選択ができます。メインのウィンドウの各色のボタンと連動しています。

検索

マーカーを引く際に所定の語を探すのに検索が使えます。

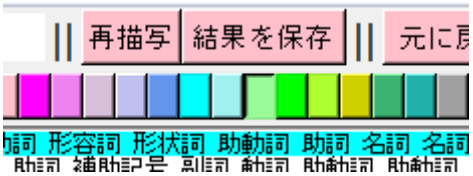
検索された語は赤く表示されます。通常の処理での検索と全く同等に行えます。

表示項目の変更



他の処理と同様に表示項目の変更ができます。表示項目を変更しても色は付けられます。

編集結果の保存

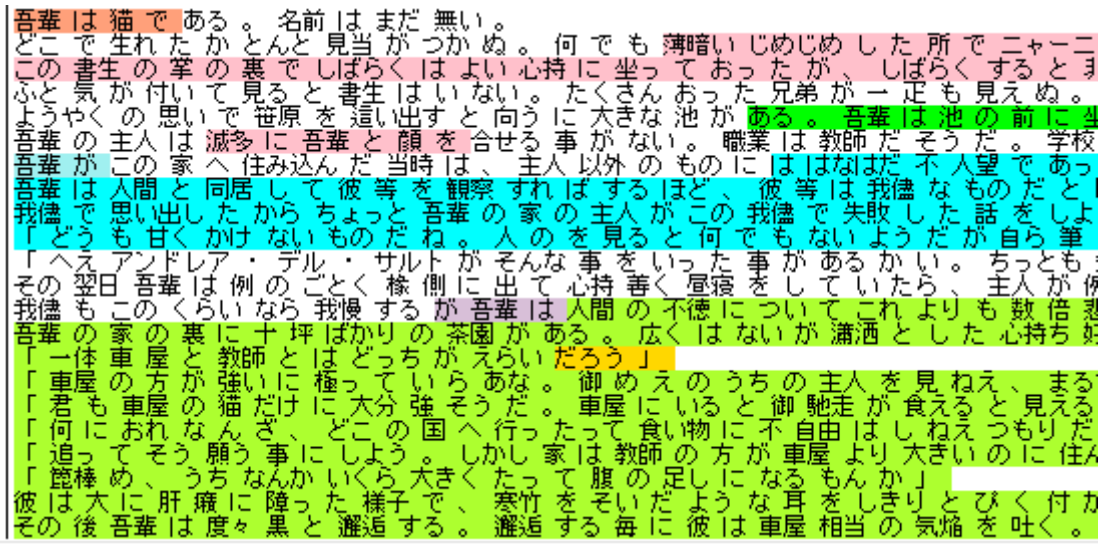


編集の完了後ウィンドウ上部の「結果を保存」ボタンで編集内容を保存します。編集しているテキストのサイズが大きい場合保存に時間がかかりますが、途中でウィンドウを消してしまうと保存がうまくされません。「結果を保存」ボタンがへこんでいる間が保存中です。

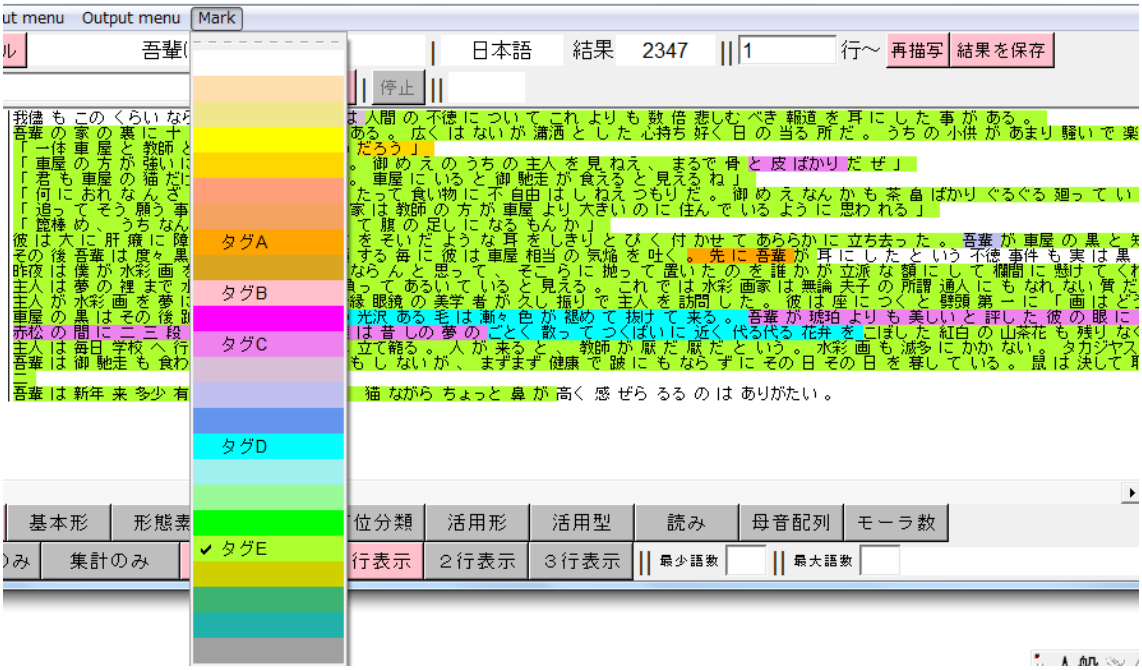
付与したマーカーの利用

編集したマーカーの保存後に通常の処理で付与したマーカーが使えます。選択したテキストファイルにマーカータグが付けられている場合、いくつかのボタンが追加されます。

Sentence でのマーカーの利用

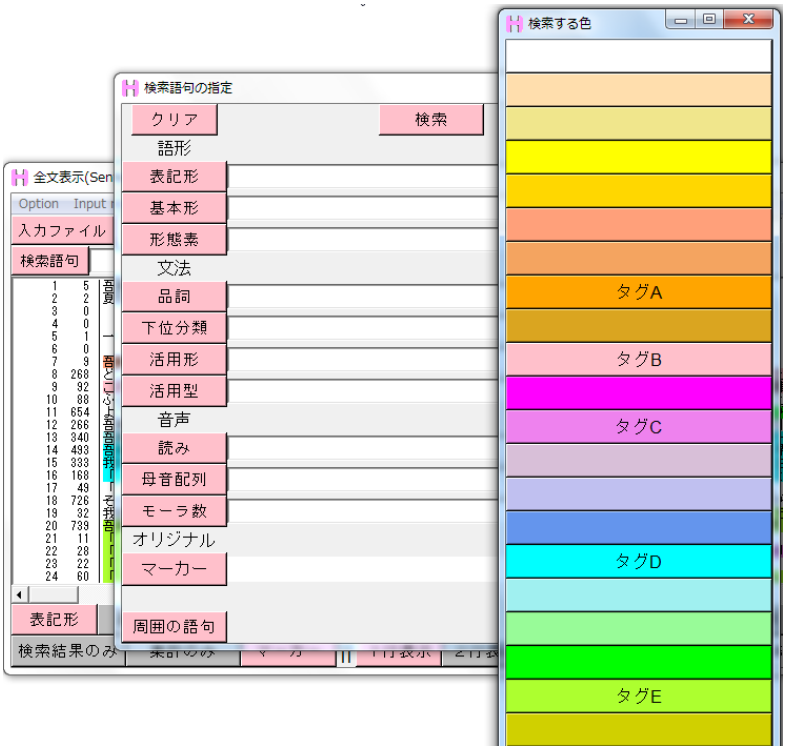


Sentence での表示では、そのまま表示すれば編集した箇所に色が付けられています。



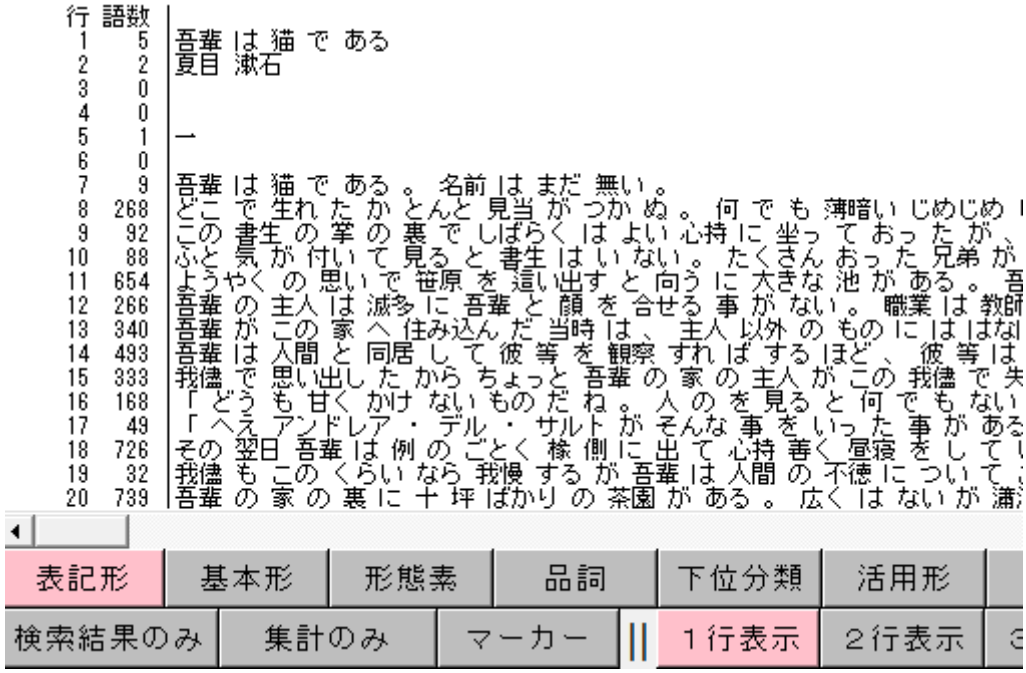
ウィンドウ上部のツールバーに追加された「Mark」ボタンで色を選択し「再描写」で、選択した色のマーカータグが付けられた語のある行のみが表示されます。

色のリストには、編集中に付けた各色の名前が反映しています。



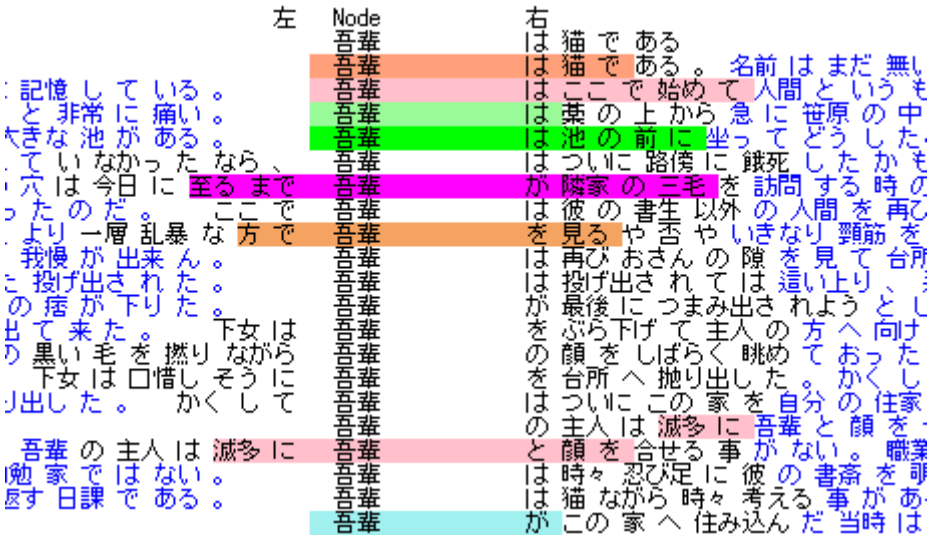
検索語句ウィドウにも「マーカー」ボタンが追加されます。

マーカーも通常の語の検索のように検索語句の条件に追加できますが、マーカーだけを選択しても検索は行われません。何か他の項目と組み合わせて指定した場合のみ検索に利用できます。

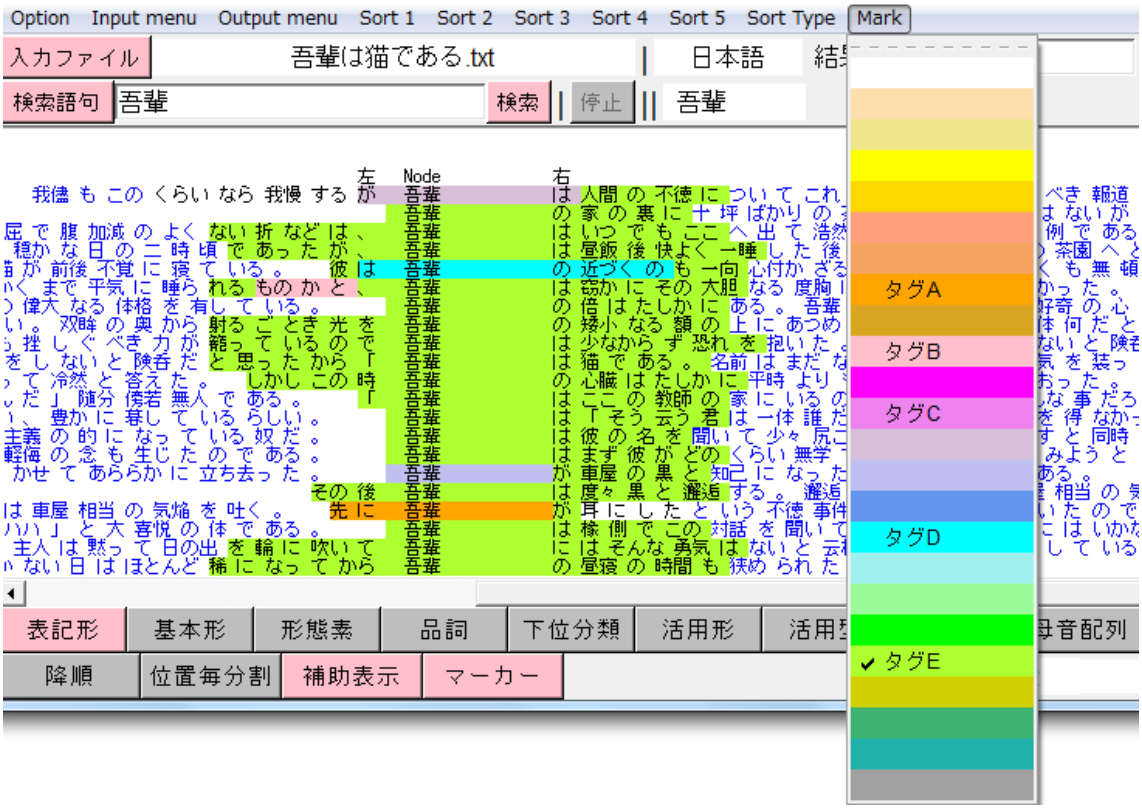


ウィンドウ下部に追加された「マーカー」ボタンで、画面上の色の表示、非表示を選択できます。

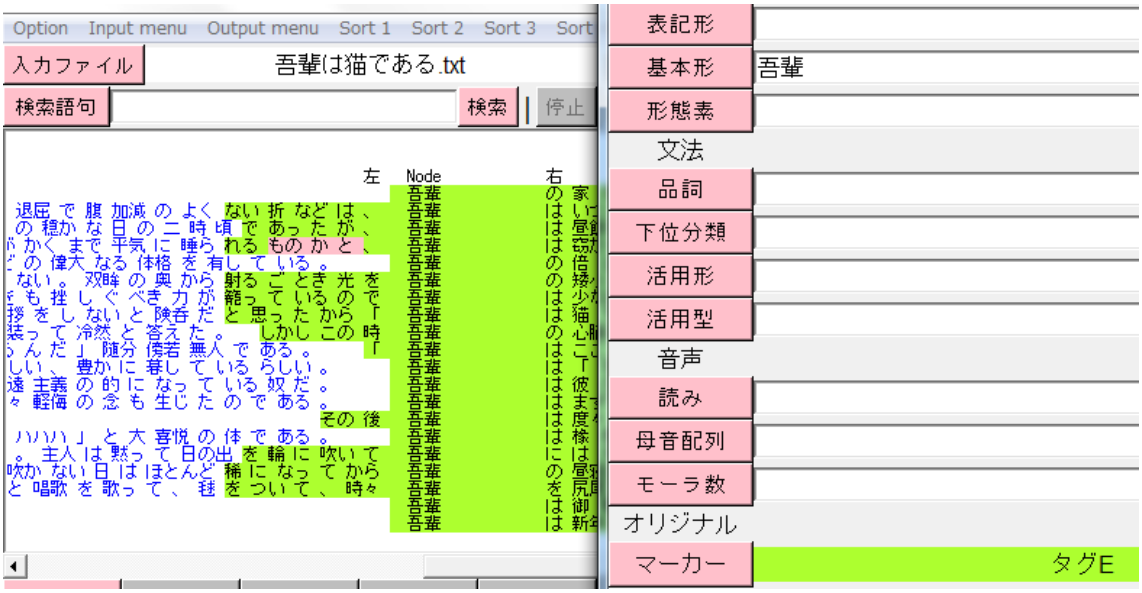
KWIC でのマーカーの利用



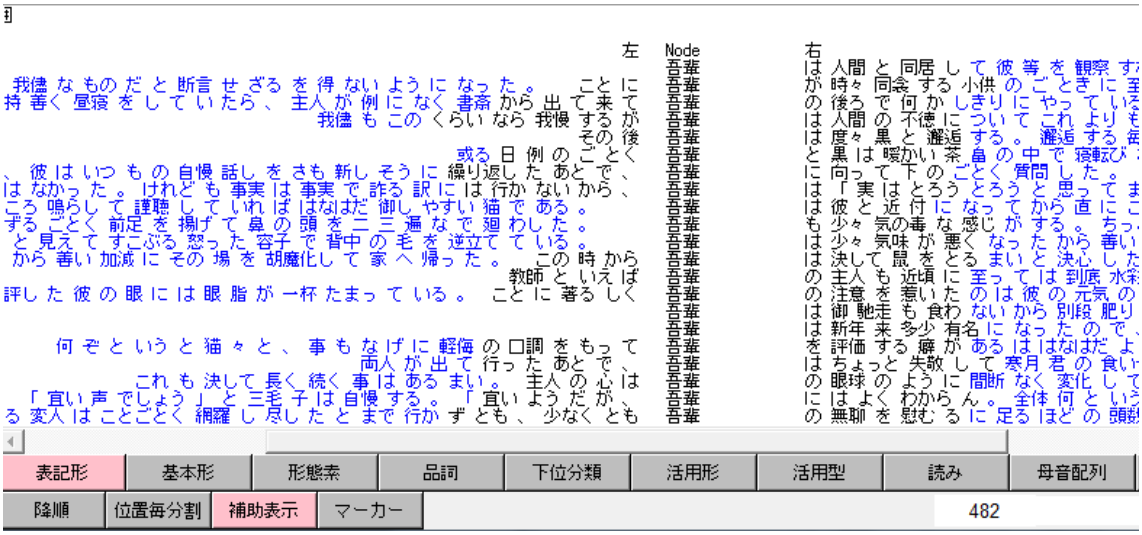
通常での KWIC の検索結果にも色が反映されます。



ウィンドウ上部のツールバーに追加された「Mark」ボタンで色を選択し「再描写」で、選択した色のマーカータグが付けられた語が左右の取得範囲内に１つでも有る結果のみが表示されます。検索語そのものにそのマーカータグが付いていなくても構いません。

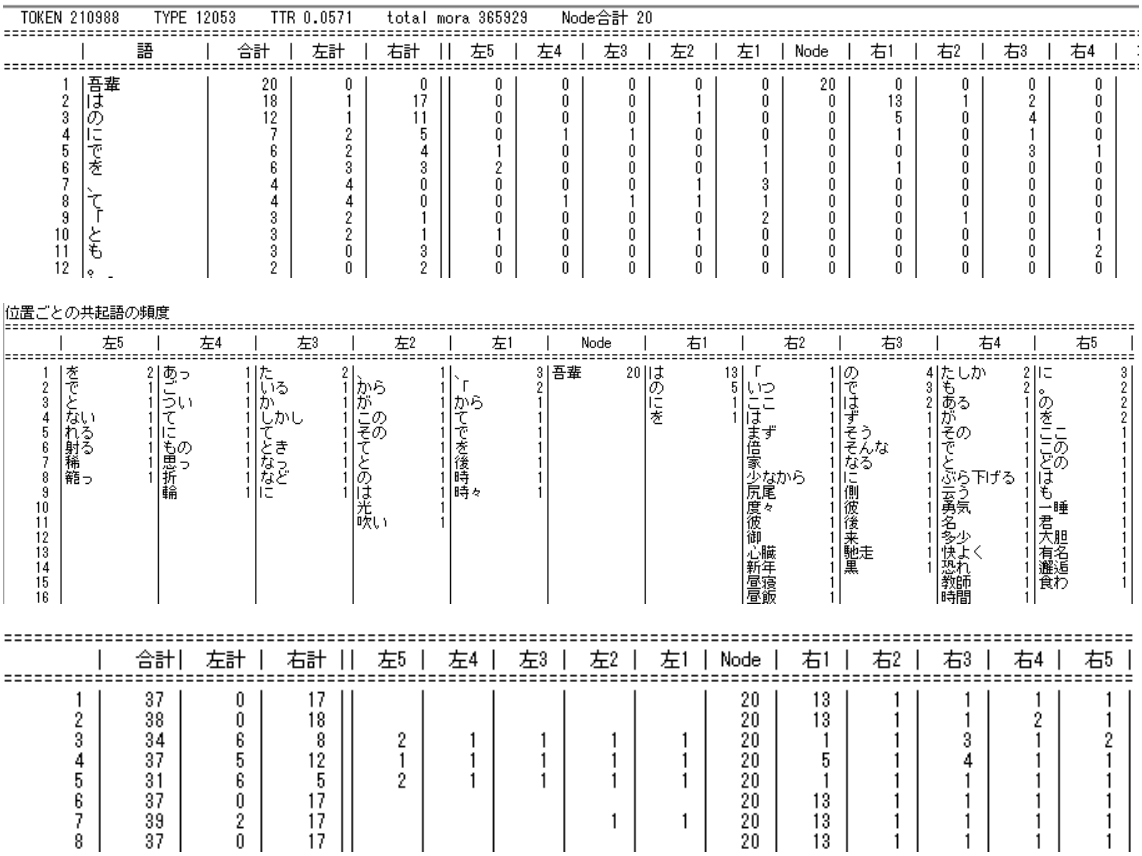


検索語句ウィンドウでマーカータグを他の項目と組み合わせると検索条件に加えることができます。



ウィンドウ下部の「マーカー」ボタンで KWIC 結果の色の表示、非表示を切り替えられます。

Collocates、Picture、POPAK でのマーカーの利用

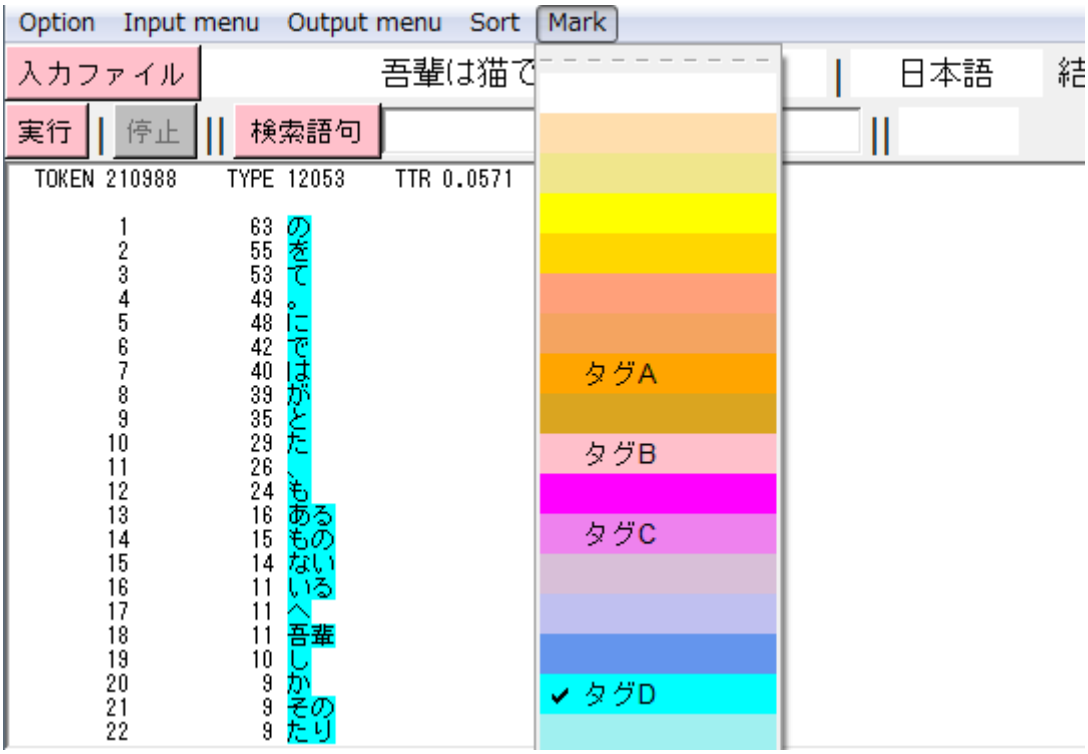


この３つの処理は KWIC での検索結果の左右の取得幅の語を共有しますので、検索結果を使った統計値にそのままマーカーを指定した条件の結果が反映しています。ただし、検索語句以外のウィンドウ上では特にボタンなどは現れません。

Freq でのマーカーの利用

Freq の場合、ウィンドウ下部のマーカーボタンをオンにすると、マーカータグが結果に反映されます。その際、マーカータグごとに語の数が集計されますので、同じ語でも違う色が付いていれば別の語という扱いで集計されます。

つまり、色と語の２条件の組み合わせで集計されます。



ウィンドウ上部のツールバーの「Mark」ボタンで色を選択し「再描写」で、選択した色のマーカータグが付けられた語のみが表示されます。

テキストデータの編集(Edit)

この処理では、テキストに自動で付与されるタグと全く同等のタグを使用者が独自に付与できます。また行全体にまたがるタグ、複数行を一括で管理するタグなど様々な単位でのタグの付与も行えます。更にテキスト自体の形態素解析ミスなどの修正も行えます。

※この処理は、テキストデータの中身を書き変えることになりますので、他で検索などの処理をしているとデータの関連性が崩れる可能性がありとても危険なので、この処理のウィンドウが表示されている間は他の全ての処理のウィドウは起動しません。この処理のウィンドウを消去すると他の処理が使えるようになります。

※この処理は大量のメモリを消費しますので、新聞１年分などのサイズの極端に大きいファイルでは扱えない可能性があります。

タグは大きく４種類あります。

語タグ	１語ごとに付くタグ
行タグ	１行ごとに付くタグ
属性タグ	いくつかの項目をまとめて１行ごとに付くタグ
ファイルタグ	複数のまとまった行に付くタグ

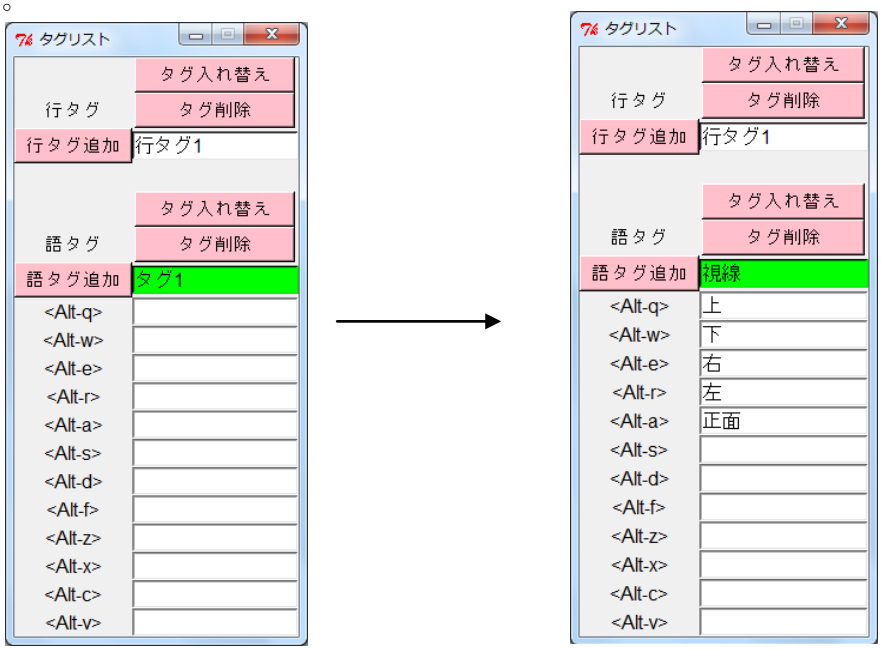
「入力ファイル」ボタンで他の処理と同様にテキストファイルを指定します。その後「読込」ボタンで画面に本文が表示され、タグ編集画面の準備が完了します。
テキストエリアの右側が語タグの編集エリアで、左側が行タグの編集エリアです。

語タグ

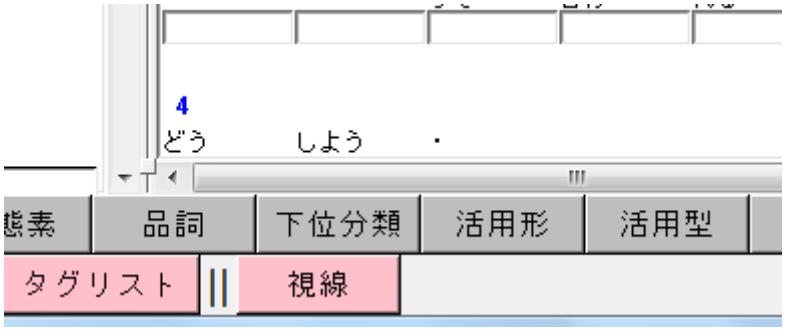
画面右側の本文が表示されている、語タグの編集エリアでは、各語の下にそれぞれ入力ボックスがあります。ここに新しいタグの内容を付与して行きます。ワンクリックでの簡易な入力方法もあります。

タグ名と要素リストの作成

タグに名前を付け、新しい項目タグを作成します。そして、その項目の要素のリストを作成します。

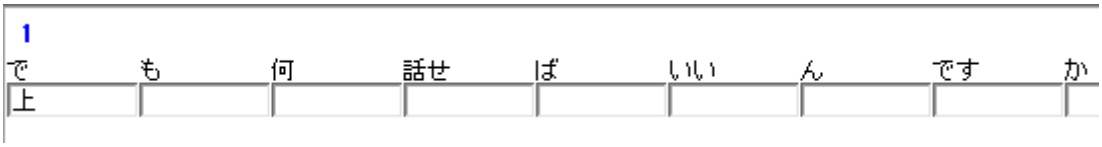


ウィンドウ下部の「タグリスト」をクリックするとタグリストのウィンドウが出ます。
そのウィンドウの「語タグ」の欄の緑の入力ボックスがタグ名になります。「タグ 1」というのが語のタグのデフォルト名ですのでこれを変更します。また、その項目にはどんな要素が来るのかを下に並ぶ白い入力ボックスに順に入力します。

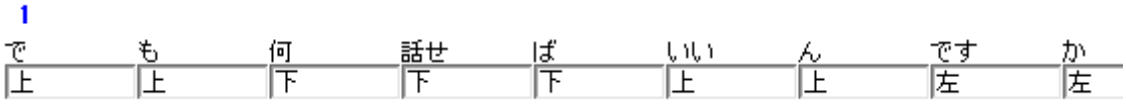


タグの名前を変えると、メインのウィンドウ下部のボタンの名前もそれに変わります。
「タグリスト」の語のタグの各要素の左側に「<Alt+q>」「<Alt+w>」「<Alt+e>」などが書いてあります。これはワンアクションで簡単にタグを付与するためのショートカット名です。

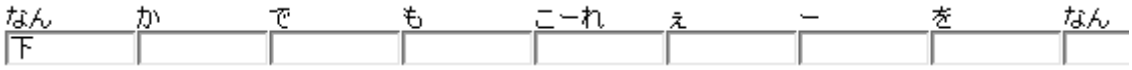
タグ付与



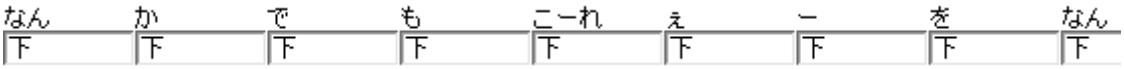
本文が表示されている右側のエリアの各語の下の入力ボックスのどれかをクリックし、先ほどのショートカットの１つを押すと、対応する要素が自動で付与されます。「<Alt+q>」であれば、「Alt キー」と「q」のキーを同時押しします。他の語も同様にできます。別の要素を付けたいときは対応するショートカット名の通りに押せば対応する要素がタグとして付与されます。



複数の連続した語に一気に同じタグを付けたい場合は、まず「Alt+アルファベットキー」で、最初の位置の語にタグを付けます。

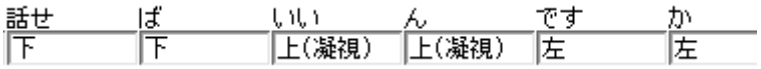


その後、同じタグを付ける終了位置の語を「Ctrl+左クリック」すると、その範囲の語に一気に同じタグが付きます。



これは、行をまたいでもできますので、複数行にまたがる連続する語に同じタグを付けることもできます。

タグは簡易に付与することができますが、例えば「視線」でも「目が合う」「凝視」など、特別な場合により詳しく付けたい場合があります。ただそのような要素はそれほど出現せず、リストが長くなりすぎたり細分化しすぎてのリスト化が難しいこともあります。そのような情報を付与したいときは、タグの入力ボックスに直接入力できるようになっています。



キーボードで直接、自動で入力された要素に付け足すなどをします。

語タグの入力は、リストからの簡易入力と、キーボードからの直接入力の２通りがあります。

語タグの追加

タグリストボタンで、「語タグ追加」ボタンを押すと、語のタグが増えます。新しいタグに対応するためにウィンドウがいったん消えて再度現れます。タグウィリスとンドウも消えますので、再度「タグリスト」ボタンで出現させます。

ウィンドウ下部の右に新しく「タグ 2」ができています。先程のタグの時と同様にタグ名と要素を指定していきます。

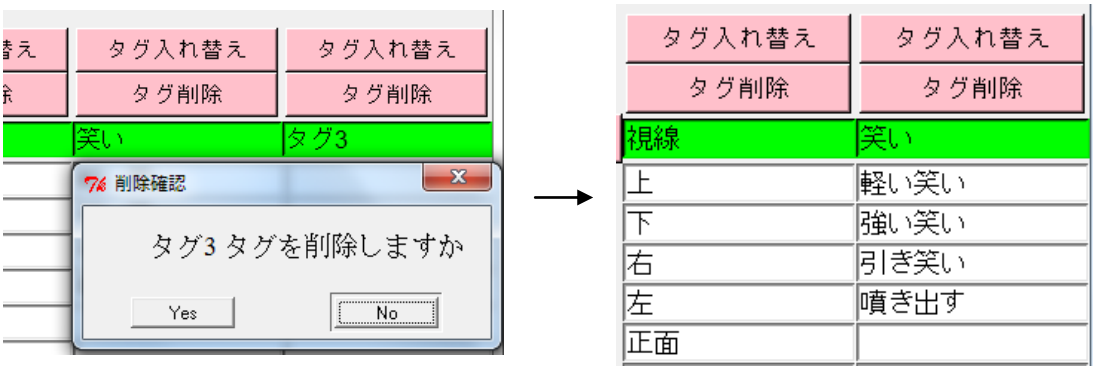
「タグ 2」も同様に名前が変わります。新しく付与したタグの項目表示ボタンを押すと、編

集エリア内の今まで付いていたタグの情報が全て消えます。これは、編集するオリジナルタグが新しいタグのものに変わったためで、先ほど付けた語タグの情報は残っていますが、現在選択されているタグではなくなったため一時的に表示から消えているのです。

新しい語タグも同様の操作で付与ができます。

語タグの削除

新しく作ったタグを削除したい場合は、タグリストウィンドウのタグ名の上の「タグ削除」ボタンで消します。



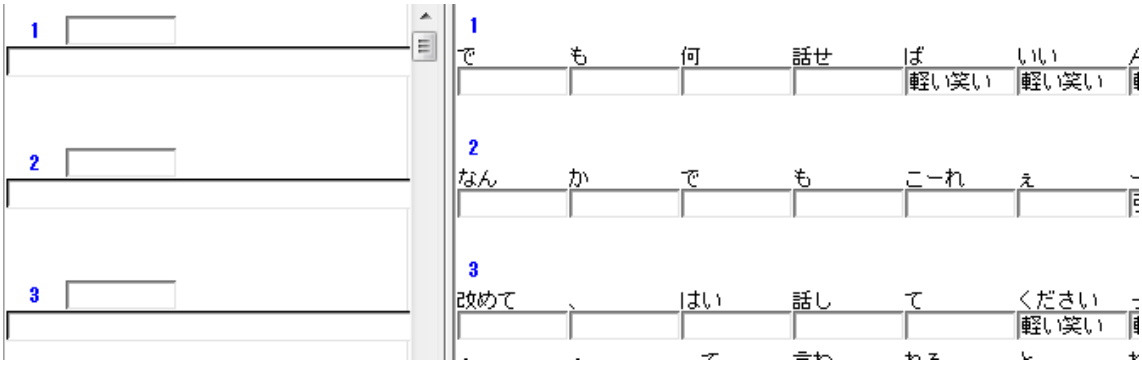
テキストへ既に付与されたこのタグの要素も全てこれで一括で消えます。

語タグの入れ替え

語タグの順番を入れ替えるときは、タグウィドウの入れ替えたい語タグの上の「タグ入れ替え」を押し、出てきたリストから入れ替える先のタグを選択し、確認に「Yes」をします。

行タグ

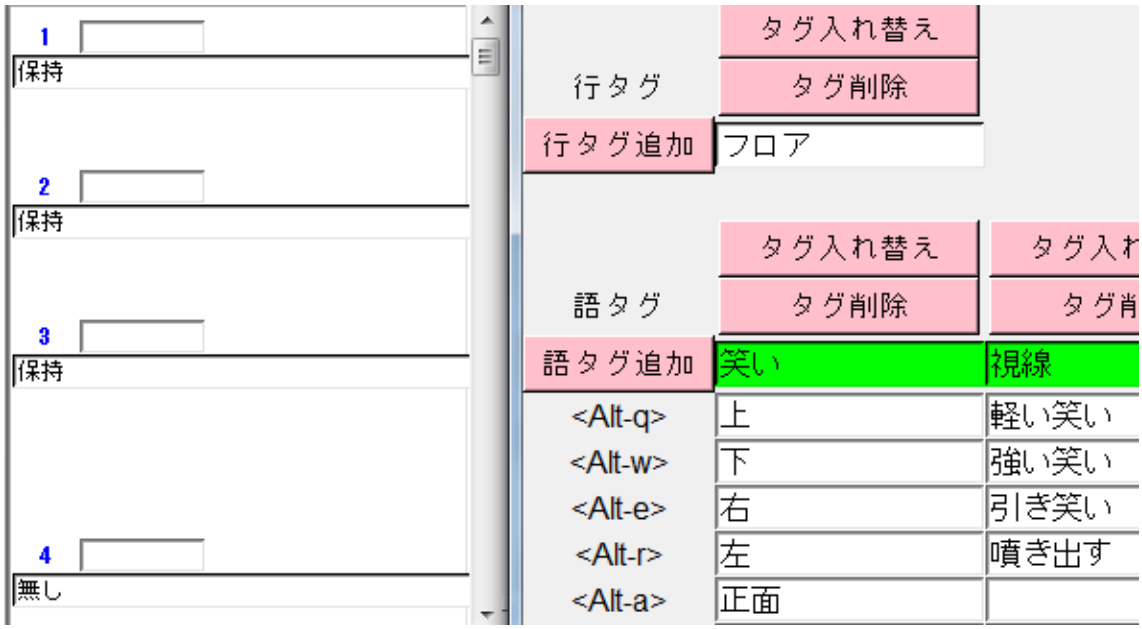
行タグは、画面の左側の編集エリアで付与します。行タグは、右側の語の編集エリアでのその行の内容とほぼ同じ位置に表示されています。



各行の頭にある数字が行番号名で、左右の対応する番号が同じ行になります。左右２つの編集エリアのスクロールバーは連動していて、２つ編集エリアが同時にスクロールします。スクロールバーを使わずに、編集エリア内でマウスの移動でスクロールさせると片方のウィドウのみがスクロールし左右の行がずれます。その場合もスクロールバーを使えばすぐに行位置が揃います。

行タグの名前変更と付与

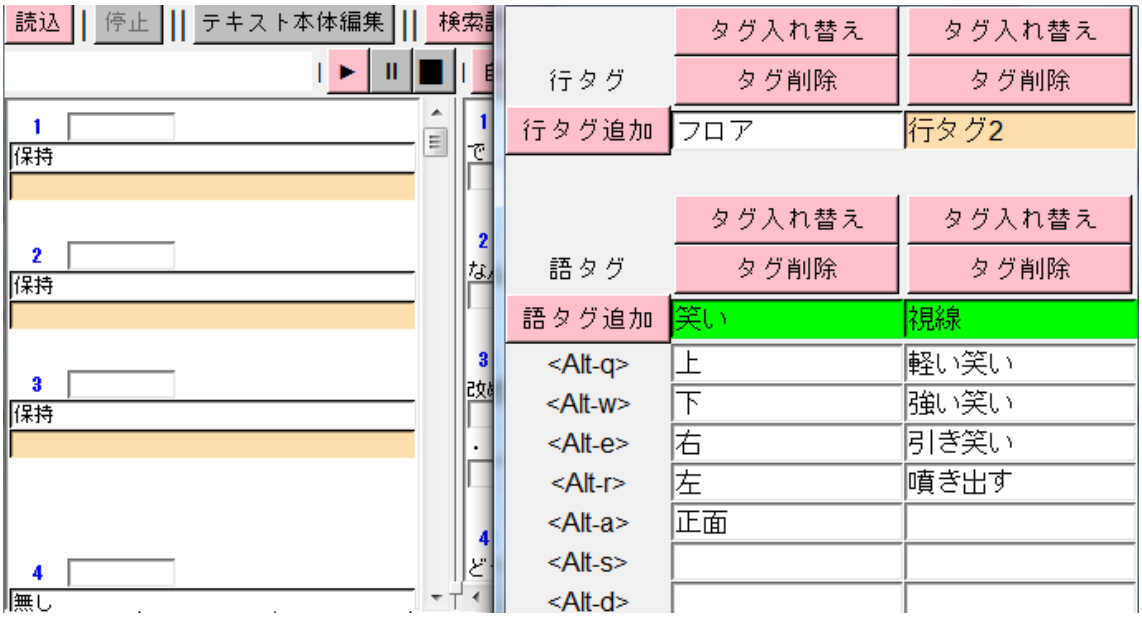
タグ名は、語タグと同様に編集できます。



語タグと同様に、ウィンドウ下部の「タグリスト」ボタンで行タグ名を編集します。

行タグは簡易編集が無いので行タグの入力ボックスに直接入力します。行タグの入力ボックスは、行番号の下にある横に長いボックスです。

行タグの追加、削除、入れ替え

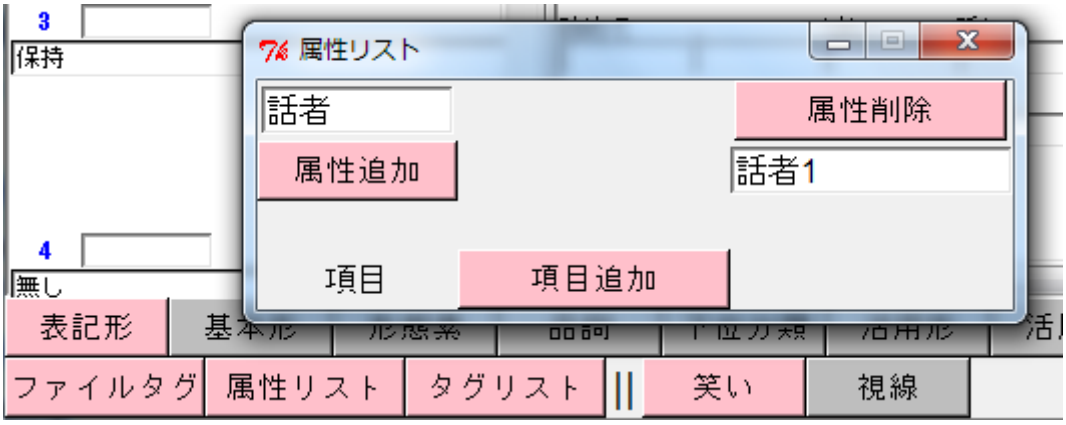


行タグも語タグと同様に、「行タグ追加」「タグ削除」「タグ入れ替え」のボタンで数や順番を変えられます。

行タグは、増えるごとに編集ボックスに別の色が付きます。タグリストでの行タグの色と対応しています。行タグの数が増えると何番目がどの項目だったか分かりにくくなるため、色で区別します。この色は、行タグ自体に割り振られるのではなく、１つめからの順番で自動的に割り振られるものなので、行タグの順番を入れ替えた場合でも付く色の順番は同じになります。

属性タグ

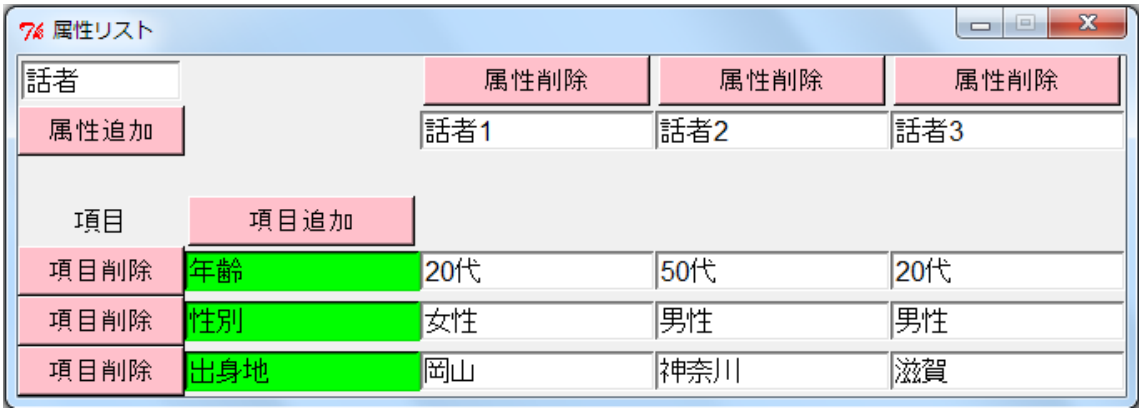
属性タグは、行ごとに付くタグですが、行タグとの違いは、１つのタグに様々な要素がまとまって入っている点です。



ウィンドウ下部の「属性リスト」で編集します。

属性タグとは、「話者 01」「話者 02」のように、１つの名前がその中に様々な情報を保持しているものです。たとえば一人の人が「年齢」「性別」「出身地」などの様々な情報を保持していたとして、これを全て別々の行タグとして作成し、この発話者の発話行全てに毎回「年齢」「性別」「出身地」のように付与すると非常に効率が悪くなります。そこで「話者 01」は「年齢」-「30 代」、「性別」-「女性」、「出身地」-「秋田」のように一度だけ設定し、あとは各行に「話者 01」とだけ指定すれば、同時に内在する全ての要素が指定されたと同じように扱えるようにします。

属性タグ要素の設定



属性リストウィンドウ上部左の「話者」となっているのが属性の名前です。その右側でメンバーを作ります。また、各メンバーの下にそれぞれの項目を設定します。項目数も変えられます。また属性の名前も変えられます。

属性の追加は「属性追加」ボタン、内部の項目の追加は「項目追加」ボタンです。

属性タグの付与

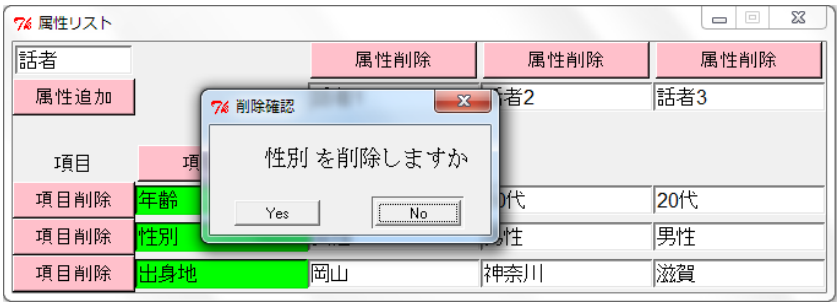
属性タグの入力ボックスは画面左側、行タグの編集エリアの行番号の右側にある少し短い入力ボックスです。

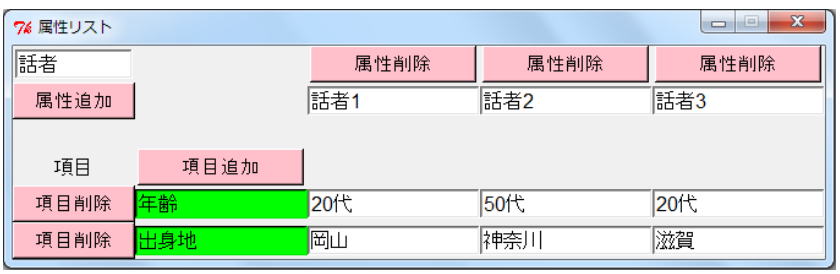
属性タグは簡易付与できます。行番号の上で「Alt」+「1」~「9」,「a」~「z」キーがショートカットになっています。「Alt+1」で属性の 1 番目、「Alt+2」で属性の 2 番目が入ります。「9」まで行ったら、「a」が 10 番目の属性になります。

語タグと同様に、入力ボックスをクリックしてからショートカットキーで簡易付与します。直接入力もできます。属性リストに作成していない属性名を付与することもできますが、結果を保存後に他の処理での扱いに制限が出ます。

属性タグ、項目の削除

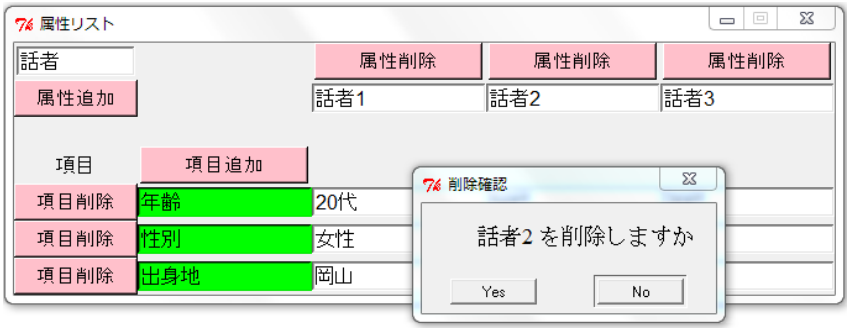
属性タグ、内部項目の削除は、語タグ、行タグとほぼ同様です。項目削除は消したい項目の左の「項目削除」ボタンで行います。

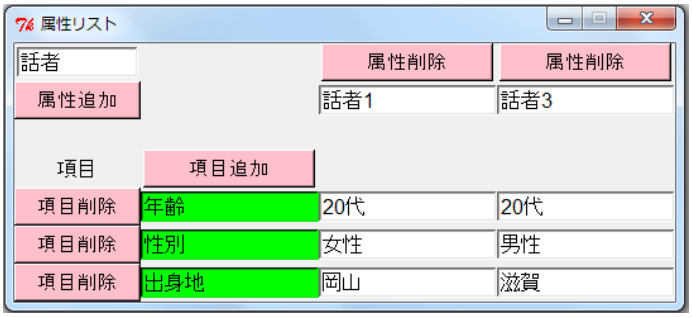




項目削除では、それぞれの属性の中でその項目だけが消されます。

属性タグの削除は消したい属性の上にある「属性削除」で行います。





１つの属性を消すと、内在している全ての項目の要素も消えます。
属性タグを消しても、テキストに付与した属性タグの名前はそのまま残ります。必要に応じて付与し直す必要があります。

属性タグ、内部項目の入れ替えはできません。

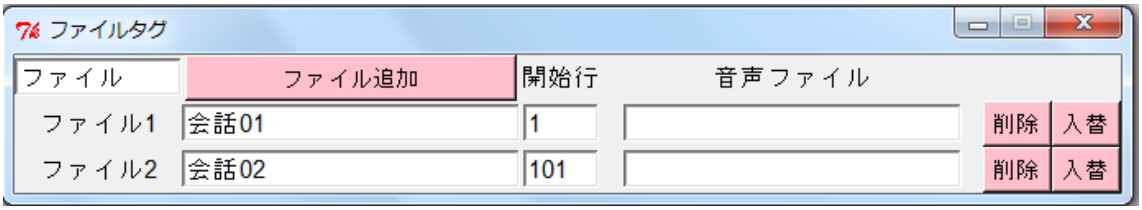
ファイルタグ

複数の行をまとめて管理するファイルタグの付与を行います。

ウィンドウ下部の「ファイルタグ」ボタンで編集を行います。

ファイルタグは、複数のテキストを結合してコーパス化した時などに使います。通常はテキスト選択時にフォルダを選択して複数テキストを一括で選択した際に、元々のファイルの内容ごとに区切られて自動的に付与されるタグです。これを、タグ編集時にも任意に決めることができます。

ファイルタグの追加と設定

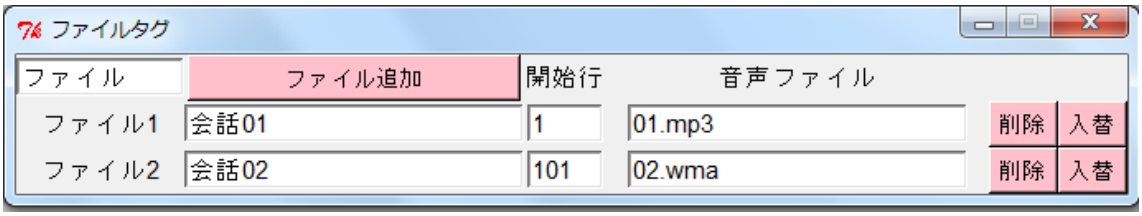


ファイルタグウィンドウの左上がファイルタグそのものの名前です。その下にある各行が１つ１つのファイルタグの名前や中身です。開始行は、各ファイルタグの始まる行で、前のタグの開始行よりも後の行番号になる必要があります。

ファイルタグの追加は「ファイル追加」ボタンで行います。

各行の右にある「削除」「入替」ボタンで、各ファイルタグの順番を入れ替えられます。

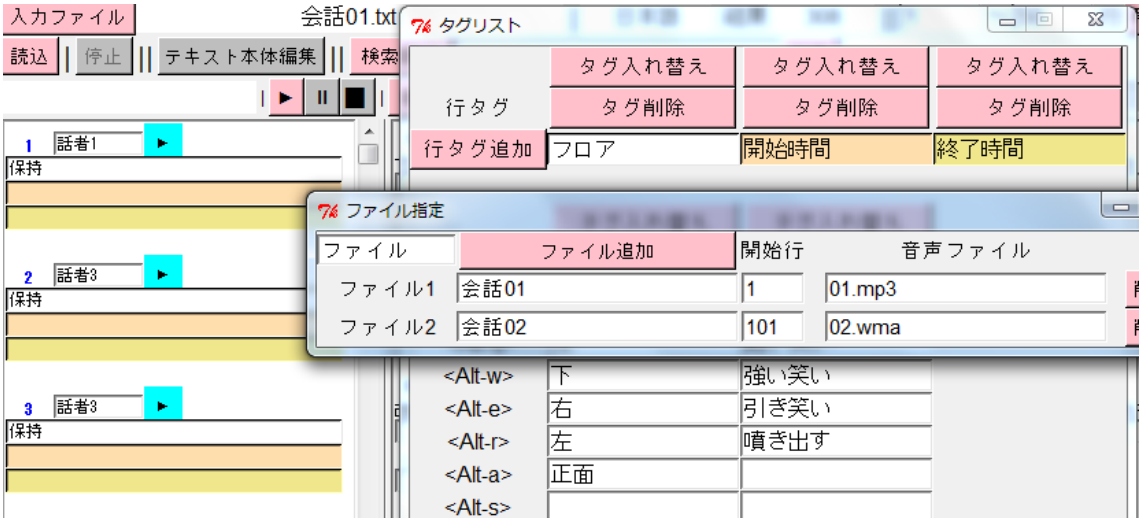
音声ファイルの指定



１つのファイルタグの範囲に１つの音声ファイルを指定できます。この音声ファイルはコーパスの中身と連動させ、いくつかの通常の処理の際に該当部分の音声を聞くことができるようになります。これによって音声コーパスの作成ができるようになります。

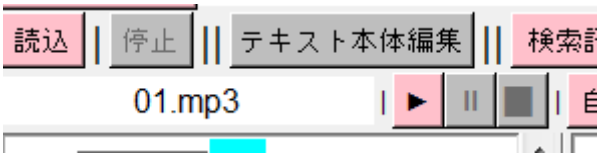
音声コーパス化

通常の処理を行う際に、一定の準備をすれば該当する行の音声の再生ができるようになります。音声コーパス化に必要な準備は２つあります。１つめは行タグを２つ用意し、片方を「開始時間」、もう片方を「終了時間」と名付けることです。２つめはファイルタグの編集の際に音声ファイルを指定することです。これで音声コーパス化の準備はできましたが、各行の開始時間と終了時間に所定の時間を付与する必要があります。



音声ファイルは、Windows Media Player で再生できる形式のファイルから指定できます。ファイルの名前には２バイト文字も使えます。ファイル名は拡張子まで書き込みます。音声ファイルは「Formatfiles」の中の「SoundFiles」フォルダに入れます。音声ファイルと開始時間、終了時間の２つの条件が揃ったら、「再描写」をします。

行タグの編集エリアの各行の属性タグの入力ボックスの右に、青い四角に囲まれて「▶」の記号が表示されます。これをクリックします。

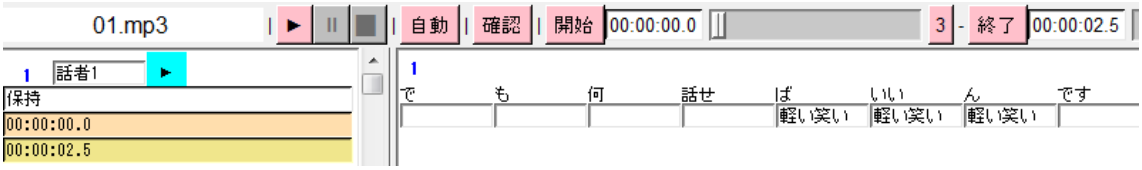


ウィンドウ上部の左端にその行が属するファイルタグで指定された音声ファイルの名前が表示されます。

「▶」をクリックするとその行が時間編集の処理にセットさます。この状態で「自動」ボタンを押します。

すると、その行の開始時間と終了時間と予測され仮の時間が設定されます。開始時間は、前の行の終わり時間がそのまま入り、終了時間は、その行の語の合計モーラ数（英語の場合は文字数）に規定の秒数をかけた時間＋開始時間となります。時間は 0.1 秒単位です。

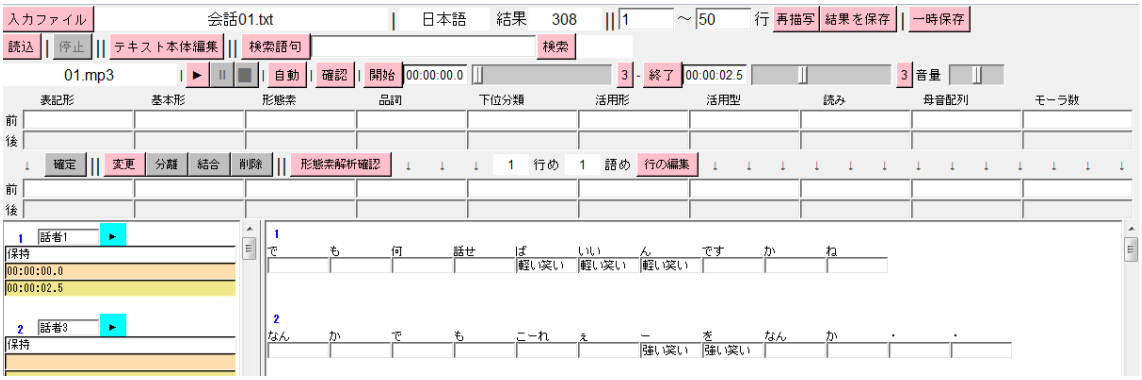
自動で付与された時間が本当に合っているかどうか、「確認」ボタンを押すと該当時間の音声が再生されます。ずれている場合、「開始」や「終了」の右にあるスライダーを左右にずらすと仮の時間がそれぞれ変わります。
編集後は「開始」と「終了」を押すとそれぞれの時間が行のタグに入り、確定されます。



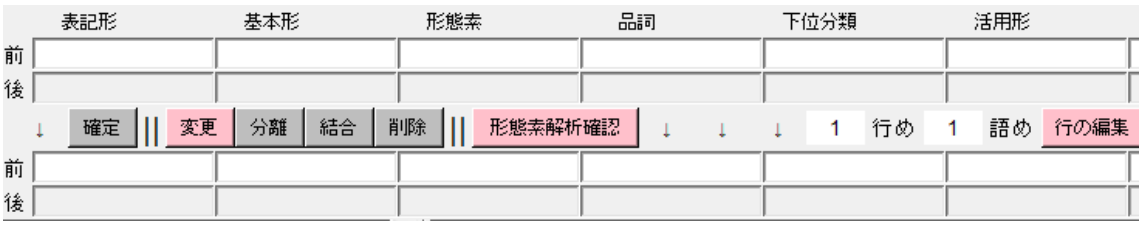
音声再生方法は全部で３つあり、
「確認」は設定時間の最初から最後までがぴったり再生されます。
「3」というボタンは、それぞれ、開始位置から３秒間、終了位置まで３秒間のみが再生されます。これは微妙な調整をする際、全てが再生されると長すぎる場合に使います。
「▶」ボタンは、開始位置から再生が始まりますが、終了は好きなところで自分で決められるものです。「■」ボタンを押すとそこで再生が止まり、その時間が仮の終了時間に入ります。耳で聞きながら終了位置を決定できるのでより確実です。その後、微調整をスライダーでします。「❙❙」ボタンは一時停止で、終了時間への付与はされませんが、音声の再生だけを止められます。また「❙❙」を押せばそこから再開されますし、そのまま「■」を押せばその時間が終了時間に入ります。この状態では「▶」は利きませんが、行タグ編集エリアの「 ▶ 」を押せばまた初期化されます。

テキスト本体の編集

テキストの形態素解析結果の修正が行えます。

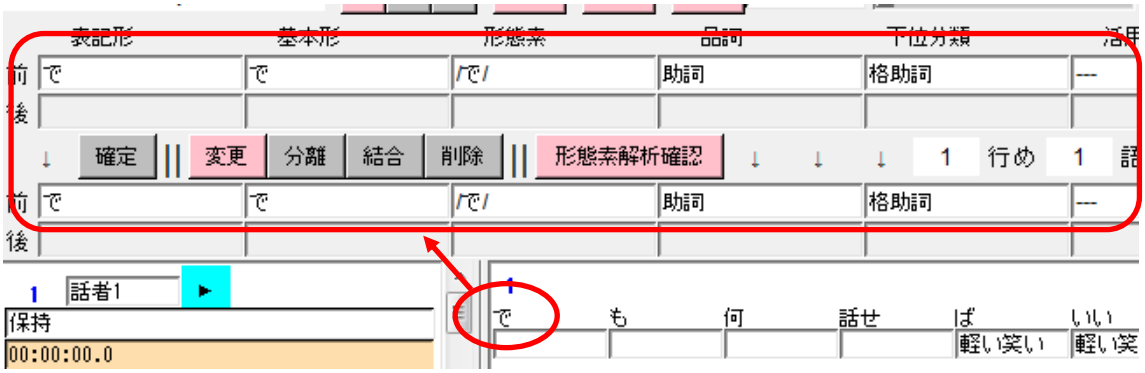


ウィンドウ上部の「テキスト本体編集」ボタンを押すと、ウィンドウ上部の編集ボタンのエリアが増えます。



新しく増えたエリアは大きく３つのことができます。１つは、語の区切り方や各項目タグなどの、１語を詳細に編集することと、行単位でテキストを編集したり、新しい内容を本文に加えるなどの、行単位で編集することと、形態素解析ソフトの解析結果を確認することです。

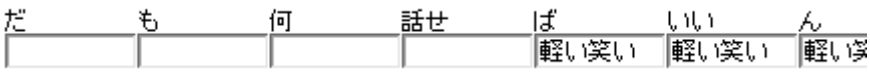
語の編集



語タグの編集エリアの中で編集したい語をクリックすると、語の編集エリアにその語がセットされます。

各項目が羅列されている上側が編集前の内容で、下側が編集後の内容です。下側の変更をします。

修正すべき箇所を修正したら、「確定」ボタンを押します。



語タグの編集エリアの該当する語が変更されます。

語の編集は４種類あります。編集する語を選択後、編集方法のボタンで選択します。

「変更」は、１語の中だけでの変更で、各項目を変更します。

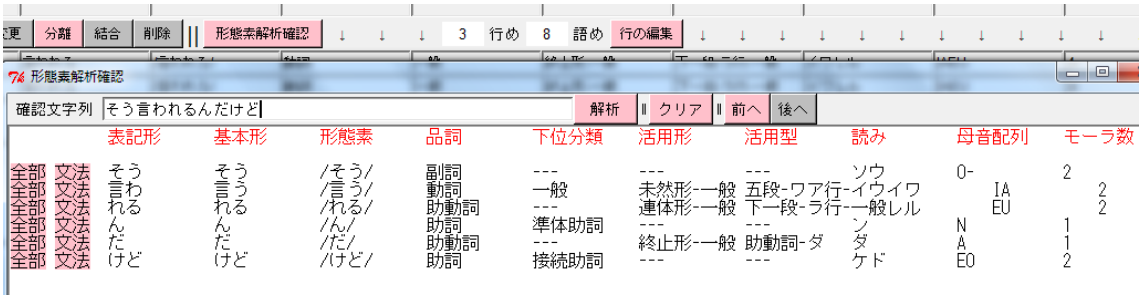
「分離」は１語を２語に分離するものです。編集後の「前」と「後」にそれぞれ分離後の前後の語の内容を指定します。分離する個所などは自動で判別できないため、完全に１つ１つ指定します。

「結合」は、２語を１語に結合するものです。編集前の「前」と「後」の２語を結合し、編集後の「前」の語として指定します。結合内容はほぼ予測できるため、自動で結合結果が出ますので、細かい修正をします。

「削除」は、１語を単純に消します。修正後の語の指定はできません。

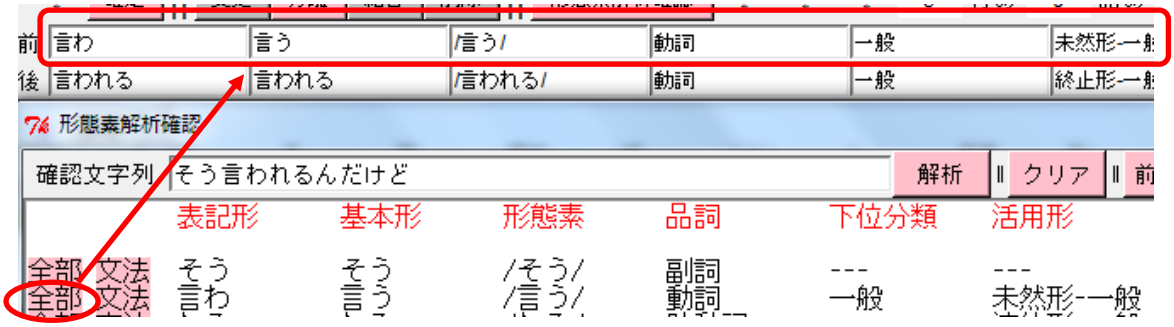
形態素解析確認

語の編集で、特に分離の場合など、編集後の文法タグをどう付ければいいか、形態素解析ソフトのタグについて詳しくなければなりません。本ソフトでは、これを簡易に利用するために形態素解析確認機能が有ります。語の編集エリアの中にある「形態素解析確認」ボタンで形態素解析確認ウィンドウが出現します。

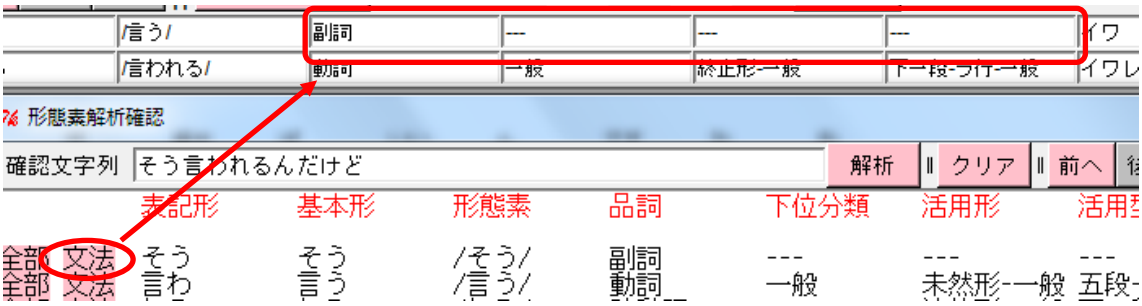


この上の方にある入力ボックスに解析をテストしたい文字列を入れて「解析」ボタンを押します。すると、下の広いテキストエリアに解析結果が表示されます。解析結果は縦に１行１語で詳しく表示されます。またそれぞれの語の結果の右に「全部」「文法」というピンクのバックになっている文字が有ります。これをクリックすると語の編集先に簡単に入力ができるようになります。

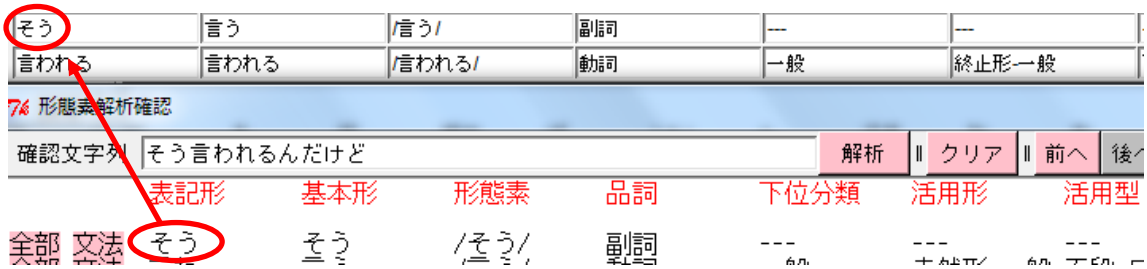
参考にしたい語の解析結果の「全部」をクリックすると、語の編集エリアの編集後の内容にそのまま入ります。



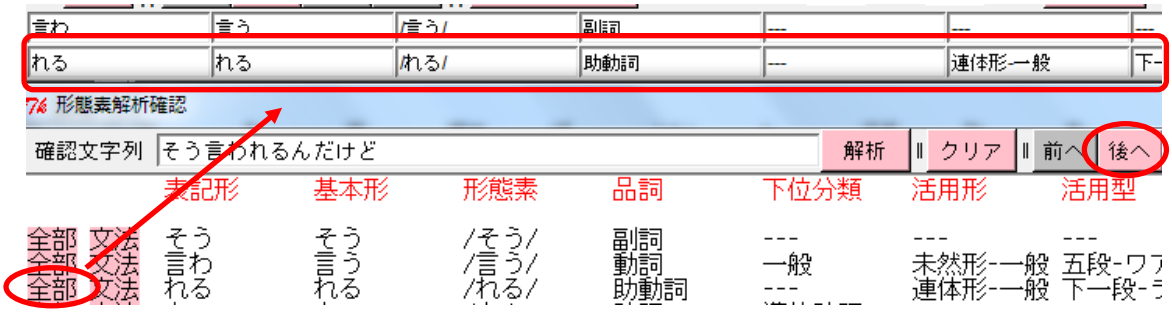
「文法」をクリックすると文法項目の「品詞」「下位分類」「活用形」「活用型」が入ります。



単独の要素をクリックするとその項目のみが使用されます。

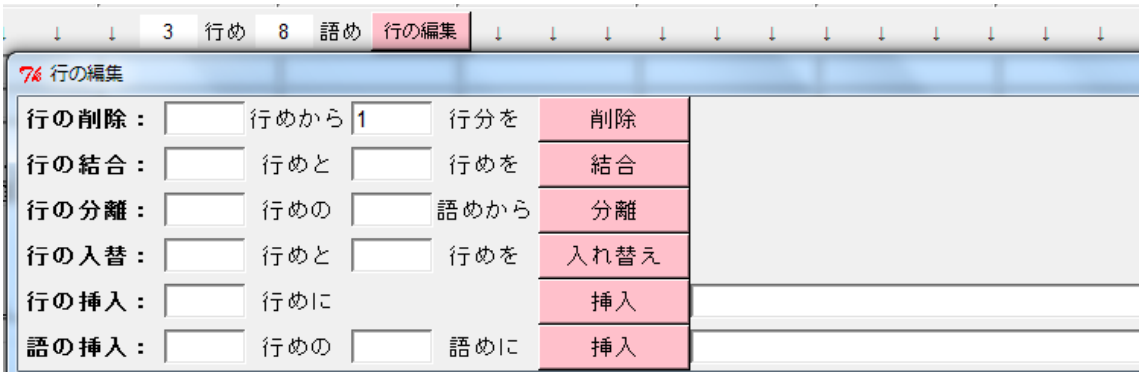


編集後の「後」の語に入力したい場合は、形態素解析確認ウィンドウの「後へ」ボタンを押してから簡易入力する項目をクリックします



行の編集

行の編集は６種類あります。



語の編集エリアの中の「行の編集」ボタンで行の編集ウィンドウが現れます。

行の編集は以下の種類があります。

行の削除

　指定した行から指定した行分の行が削除されます。

行の結合

　指定した２つの行を結合します。

行の分離

　指定した行を、指定した番号の語の場所から２つに分離します。

行の入替

　指定した２つの行の位置を入れ替えます。

行の挿入

　指定した行の指定した位置に新しい行を挿入します。

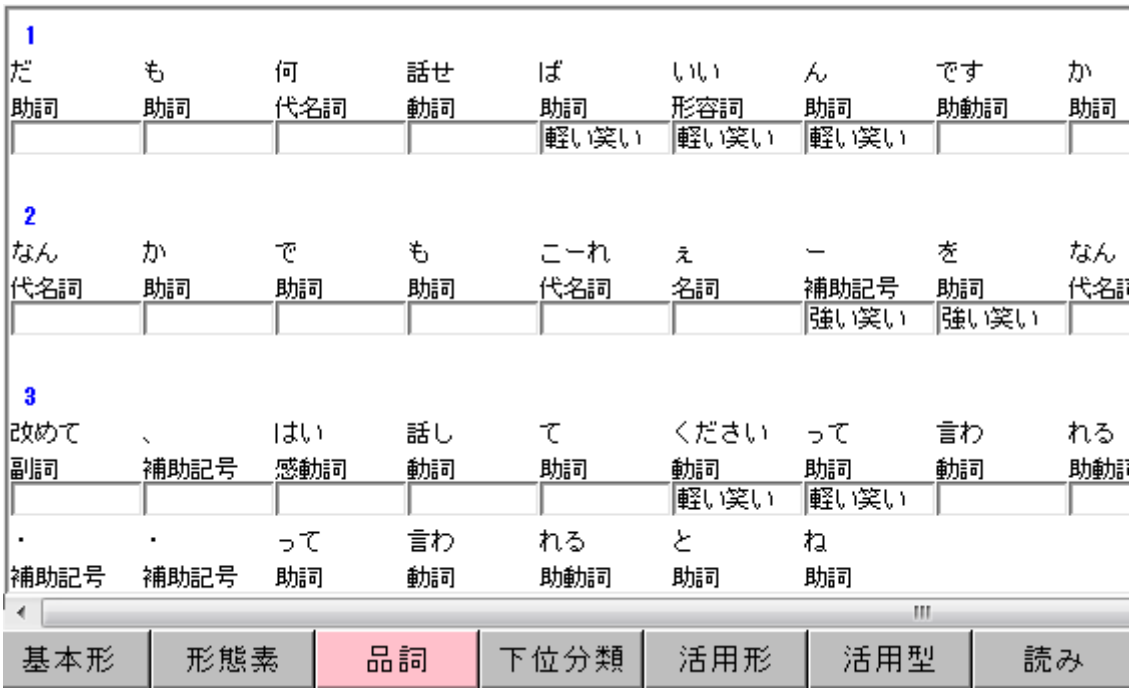
　「挿入」ボタンの右の入力ボックスに入力された文字列が新しい行の内容になります。

語の挿入

　指定した行の指定した位置に新しい語を挿入します。

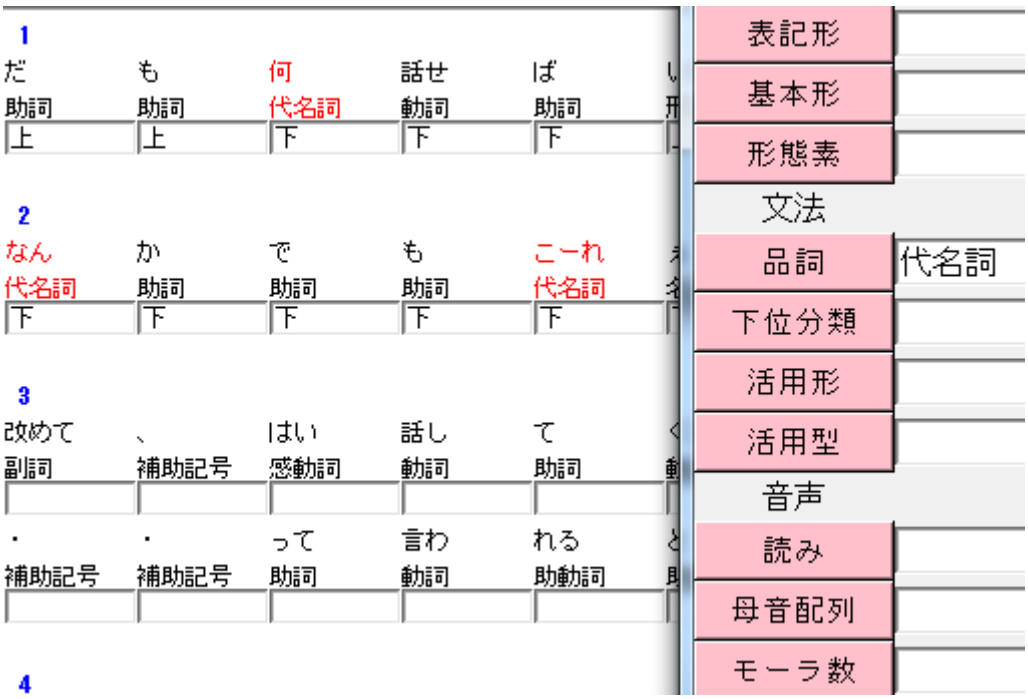
　「挿入」ボタンの右の入力ボックスに入力された文字列が挿入される内容になります。

表示項目の変更



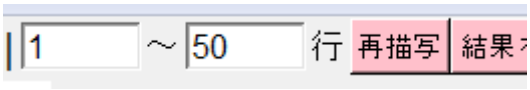
他の処理と同じように表示項目を変えられますが、表記形以外の項目で表示した際は自動で２行表示になり、上が表記形、下が選択した項目になります。

検索



他の処理と同じように検索もできます。検索された語が赤く表示されます。

表示行数の制限



この処理は非常にパソコンのメモリに負荷をかけ、処理によっては動作時間も遅くなります。特に画面に多くの語が表示されると顕著に遅くなるため、編集画面に一度に表示する行数を制限できます。ウィンドウ上部、右にある[　]～[　]行の２つの入力ボックスに表示開始行と表示終了行を入力します。初期状態では１行めから 50 行めまでになっています。50 行以上あるテキストの場合、これを変更しないと画面上に表示されている内容のみを編集したとしてもそれ以外の場所の編集していないままになります。

この数字は直接入力して指定します。表示開始行と表示最終行を別々に指定しますので、この続きを編集する場合は 51 行めからとし、終わりはこの分量のままであれば 100 行めまでとします。

１行の文字数の長いテキストの場合は行数を少なく、１行がそれぞれ短いテキストの場合は長くすると作業効率が上がります。また、パソコンの性能によって処理速度は大きく変わりますので、環境に合わせて変更してください。

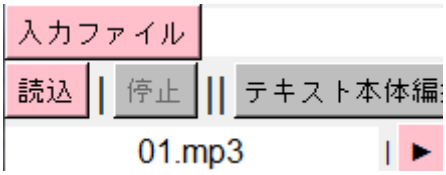
画面描画と再描写、編集後にウィンドウを消す際に特に処理の重さが顕著に出ます。

表示項目を変更する際は画面を再描写しますので、処理が重くなります。

タグを付与する際は処理時間はほぼかかりませんので、行数が多くても問題は出にくいです。また検索も再描写しませんので問題なくできます。

編集するタグを変更する際は画面の再描写はしませんが、全ての語のデータの対応を変えますので多少時間がかかります。ただこれは画面描写の分量に関わらず同じ処理になりますので、表示行数の分量には関係しません。

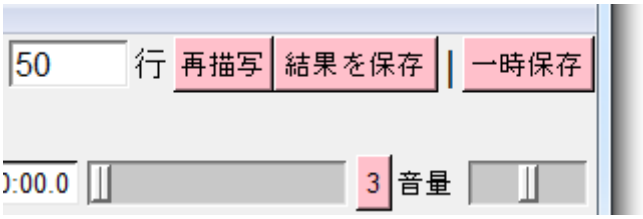
編集のやり直し



ウィンドウ上部、「読込」ボタンを押すと、その回で編集した内容を全て破棄し、ファイル選択直後の段階まで戻ります。１工程のみを戻す処理はありません。

編集結果の保存

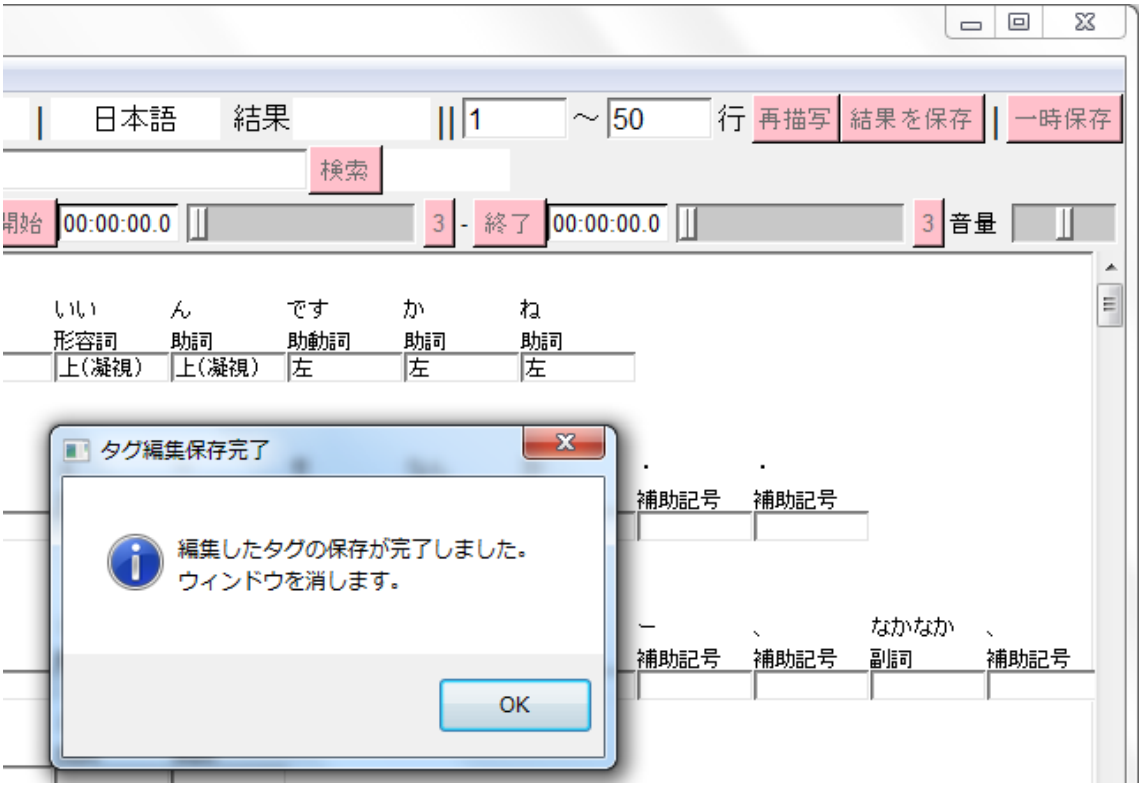
編集が完了していないが、いったん止めたいときは「一時保存」ボタンを押します。仮編集中のデータが保存されます。



マーカーの保存同様に、編集しているテキストのサイズが大きい場合保存に時間がかかりますが、途中でウィンドウを消してしまうと保存がうまくされません。「結果を保存」ボタンがへこんでいる間が保存中です。

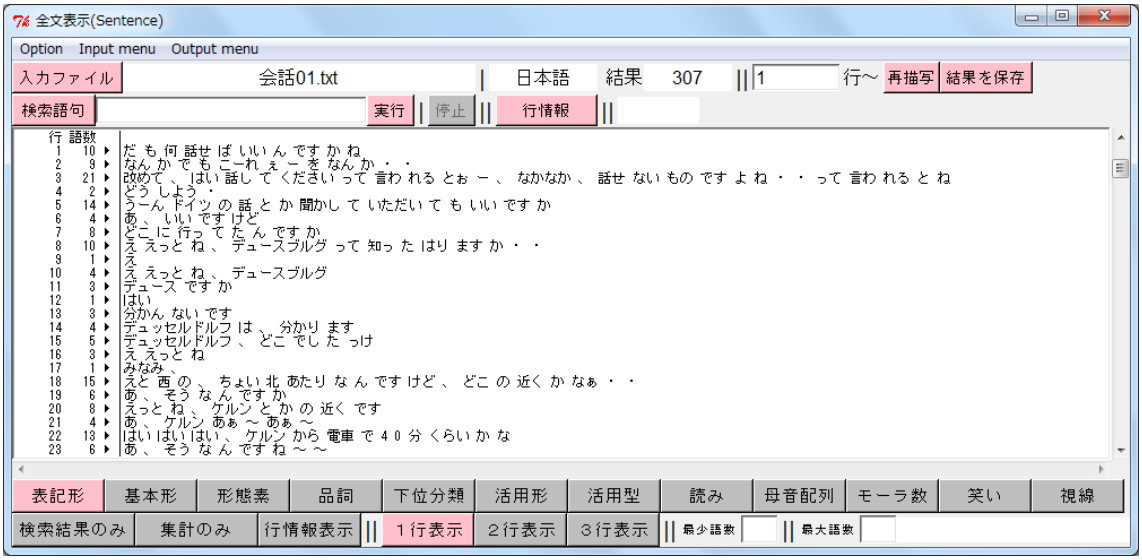
一時保存をすると、HASHI を終了しても再度同じ内容が読み込まれますので、続きを編集できます。

完全に編集が終わった場合は、ウィンドウ上部の「結果を保存」ボタンをクリックします。



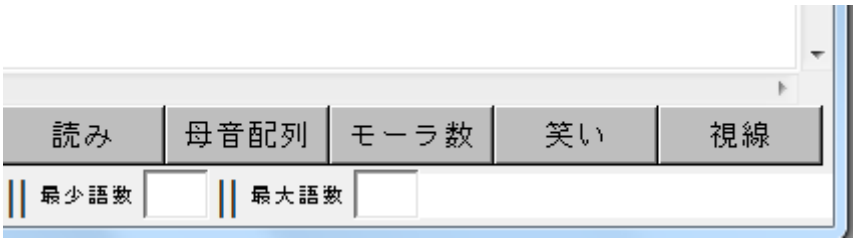
編集内容が保存され、HASHI 用の整形ファイルがいったん解凍され新しいタグデータが追加され再度整形されます。多少時間がかかりますが、保存が完了すると「タグ編集保存完了」というウィンドウが出てきて、編集ウィンドウを消しますと表示されますので「OK」します。すると編集ウィンドウが自動で消えます。

通常の処理でのオリジナルタグの使用



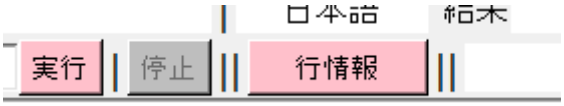
追加ボタン

オリジナルタグを付けたファイルを選択するといくつかのボタンがウィンドウに加わります。



ウィンドウ下部の項目一覧にオリジナルタグの項目が加わります。

これは、オリジナルの語タグの付与を行った場合のみ加わり、追加した語タグの数だけボタンが加わります。

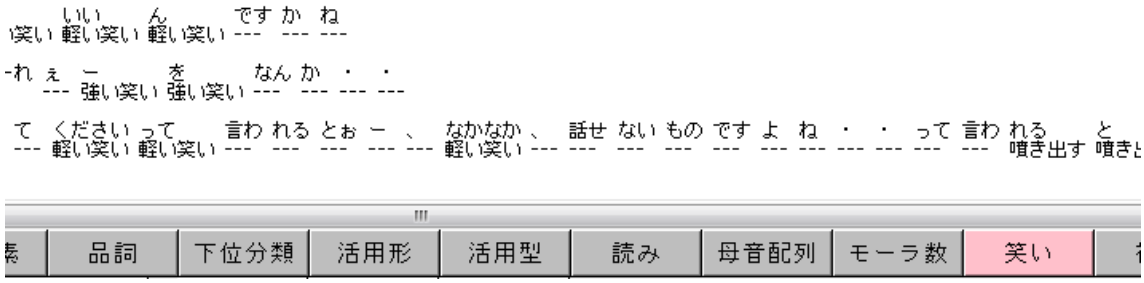


ウィンドウ上部に「行情報」ボタンが加わります。

ウィンドウ下部に「行情報表示」ボタンが加わります。

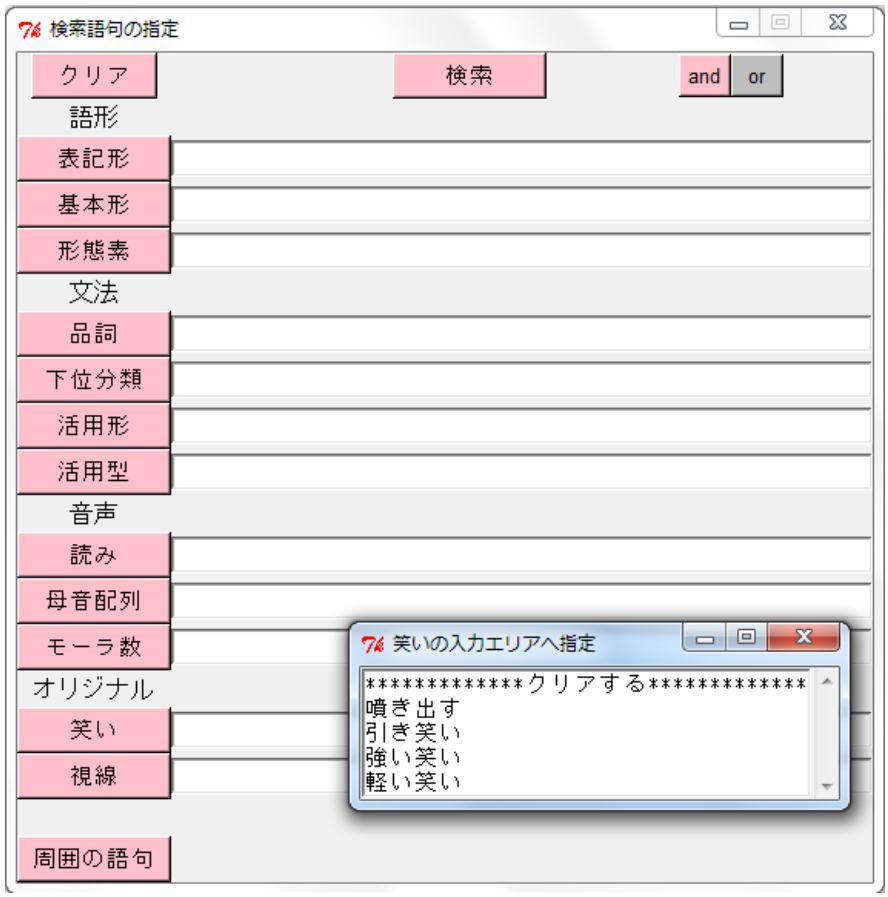
これは、行タグ、属性タグ、ファイルタグをどれか１つでも付与したら追加されます。「行情報」ボタンと「行情報表示」ボタンは必ず同時に現れます。どのタグをいくつ追加してもそれぞれのボタンは１つだけ加わります。

オリジナルタグの使用



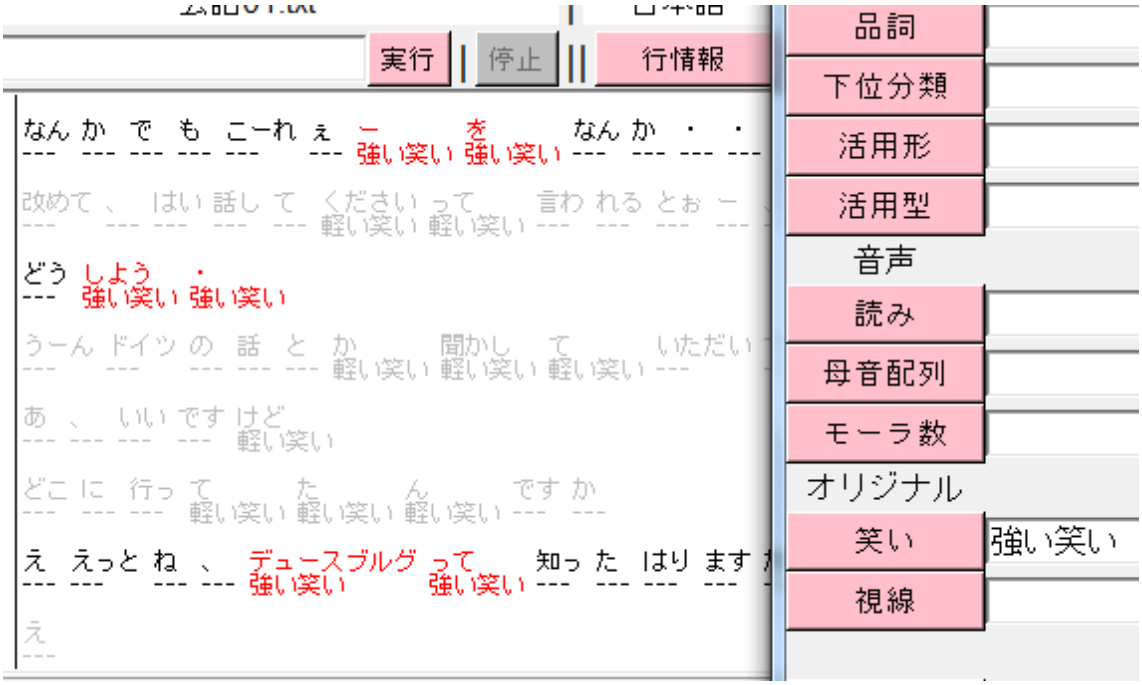
オリジナルの語タグも通常の語タグと全く同じように扱えます。表示の切り替えで独自に付与した語タグが表示されます。タグを付与しなかった箇所は「---」になります。

検索語句の指定でもオリジナルタグの項目が追加されます。

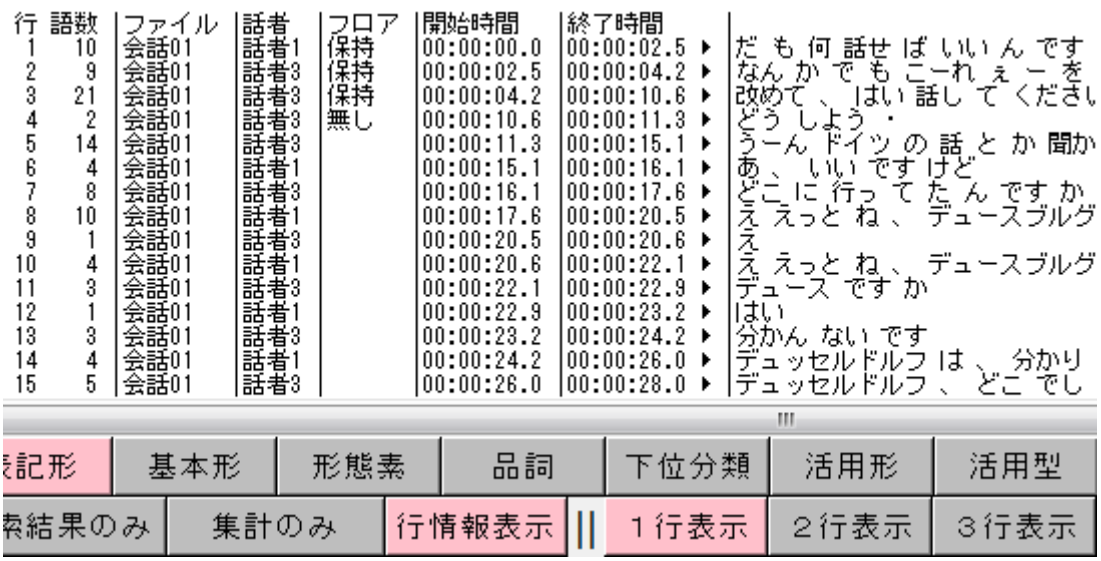


検索語句ウィンドウにもオリジナルタグの項目が追加されています。項目名のボタンを押すと簡易入力リストも現れます。タグ編集でその項目に付けた全ての要素が 50 音順でリスト化されていますので、そこから選ぶことができます。

検索も他のタグと全く同じようにできます。

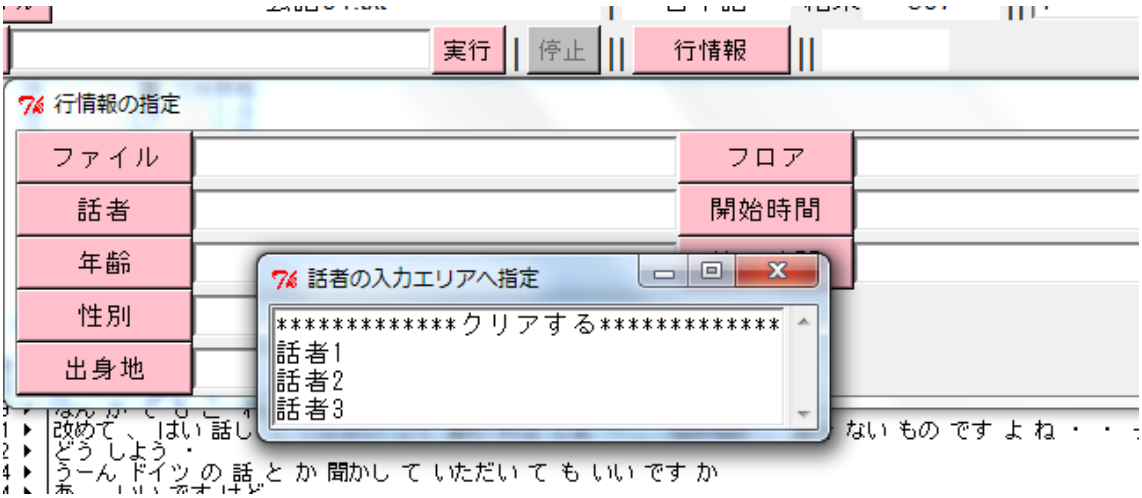


各行に付けられている行タグ、属性タグ、ファイルタグを表示させるには、ウィンドウ下部の「行情報表示」ボタンを使います。



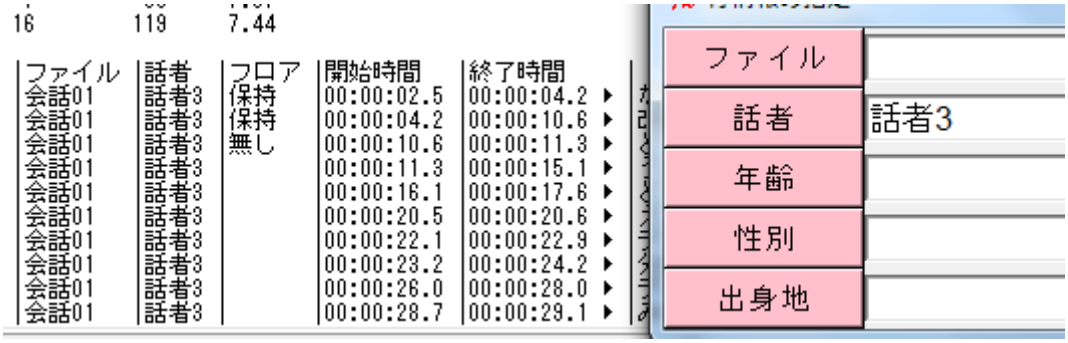
行ごとに付与されているタグ情報が本文表示の左側に表示されます。付与していない項目は空欄になります。

行タグで検索条件を絞る場合は、ウィンドウ上部に現れる「行情報」ボタンで指定します。



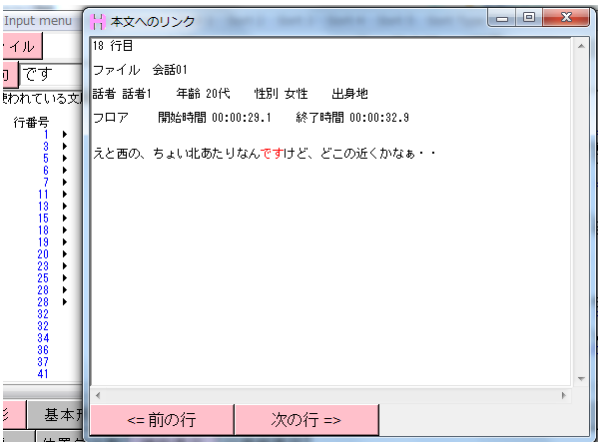
行情報の指定ウィンドウが現れるので、各項目を指定します。語タグの簡易入力と全く同じように選択できます。

行タグ、属性タグ、ファイルタグは１つもしくは組み合わせて指定できます。属性タグの内部項目も個別に指定でき、これを指定するとその要素を含んだ属性のみが抽出されます。



どれかを指定すると、指定した項目の要素のある行のみが残ります。

本文リンクでの行タグ、属性タグ、ファイルタグ内容の表示



行タグ、属性タグ、ファイルタグのうち、設定されている項目は、KWICでの本文リンクの際にも表示されます。

オリジナルタグの使用可能処理

オリジナル語タグの表示、検索は全ての通常処理で可能です。

行タグ、属性タグ、ファイルタグで条件を絞るのは Sentence, KWIC, Collocates, Picture, POPAK, Ngram のみで可能です。ただし、Freq, Keyness では、ファイルタグのみ使用可能です。

行情報表示は、Sentence, KWIC のみで利用可能です。

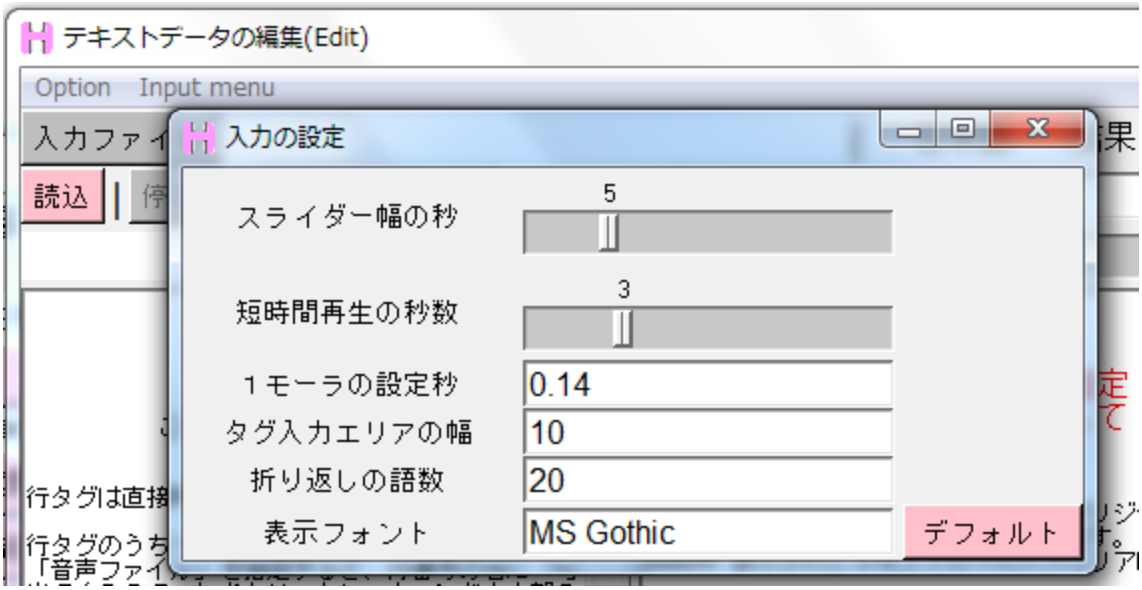
音声再生

条件が整っていれば行の音声が再生されます。

ア	開始時間	終了時間	
	00:00:00.0	00:00:02.5 ▶	だ も 何 話せ ば いい ん です か ね
	00:00:02.5	00:00:04.2 ▶	なん か で も こーれ ぇ ー を なん か ・ ・
	00:00:04.2	00:00:10.6 ▶	改めて 、 はい 話し て ください って 言わ れる とぉ ー 、
	00:00:10.6	00:00:11.3 ▶	どう しよう ・
	00:00:11.3	00:00:15.1 ▶	うーん ドイツ の 話 と か 聞かし て いただい て も いい で
	00:00:15.1	00:00:16.1 ▶	あ 、 いい です けど
	00:00:16.1	00:00:17.6 ▶	どこ に 行っ て た ん です か
	00:00:17.6	00:00:20.5 ▶	え えっと ね 、 デュースブルグ って 知っ た はり ます か ・
	00:00:20.5	00:00:20.6 ▶	え
	00:00:20.6	00:00:22.1 ▶	え えっと ね 、 デュースブルグ

行ごとの開始時間、終了時間が付与されていて、該当範囲のファイルタグに音声ファイルが指定されていて、その音声ファイルが、「Formatfiles」フォルダの中の「SoundFiles」フォルダに入っていれば、結果表示の左側に「▶」が表示され、これをクリックすると開始時間から終了時間の幅の音声が再生されます。

Edit の設定変更

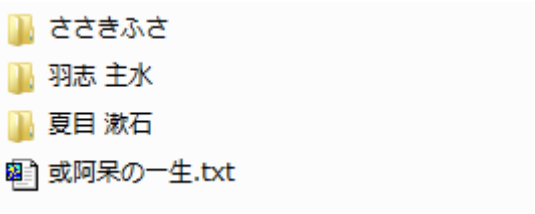


タグ編集処理で、いくつかの項目を変更できます。

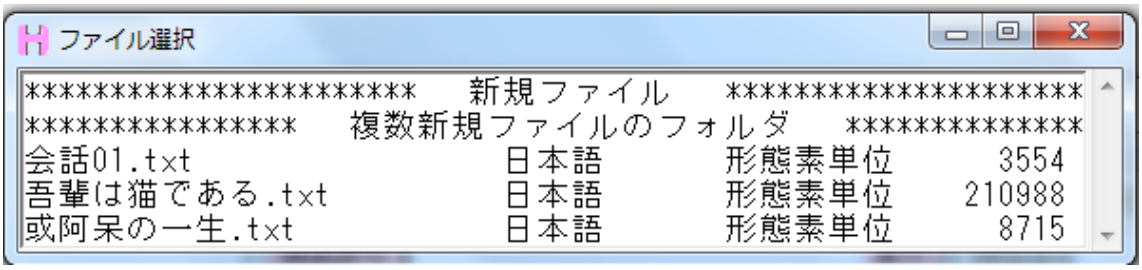
スライダー幅の秒	音声編集の際スライダーで選択できる時間の幅です。仮の開始時間と終了時間の前後の指定秒までスライダーで選択できます。
短時間再生の秒数	開始と終了位置付近の音声の簡易再生の秒数です。
１モーラの設定秒	自動で音声を付与する際の計算に使う単位秒です。
タグ入力エリアの幅	語のタグの入力エリアのそれぞれの大きさを決めます。
折り返しの語数	テキスト表示画面は規定の語数で表示が折り返されるので、その語数を指定します。

フォルダでの一括ファイル選択

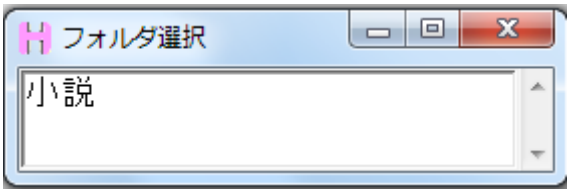
複数のファイルを一括で扱いたいときは、フォルダごと選択をします。
まず、HASHI のフォルダの直下に任意の名前のフォルダを作成し、その中に一括指定したいファイルを全て入れます。フォルダの中に更にフォルダを作り階層式にしても構いません。フォルダの階層は何層でも構いません。



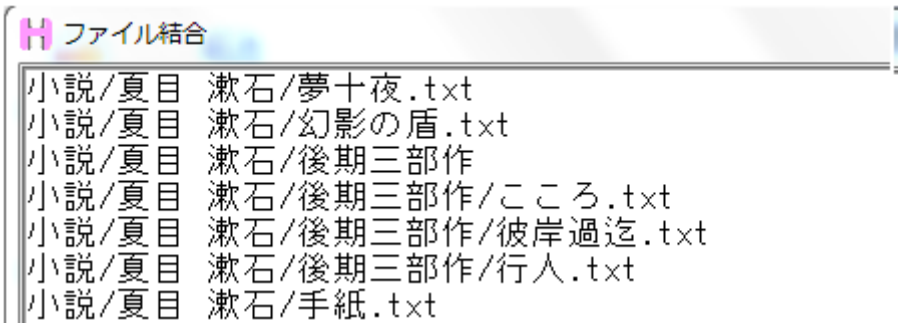
フォルダを用意したら、通常のテキストファイルを選択する際と同様に「入力ファイル」ボタンでファイル選択ウィンドウを開きます。次に、リストのうち、「複数新規ファイルのフォルダ」を詮索します。



すると、HASHI フォルダの中にある選択可能なフォルダの一覧が出ます。



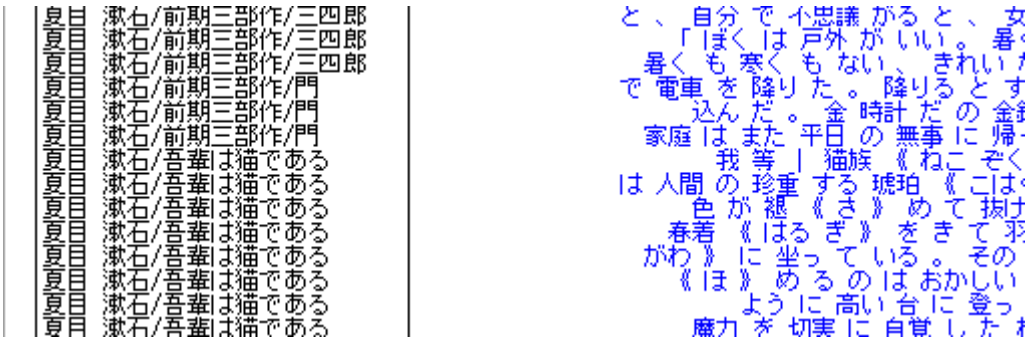
このうち１つのフォルダを選択します。その後、通常通りに分析ファイルの設定を行うとまずファイルが全て結合されます。



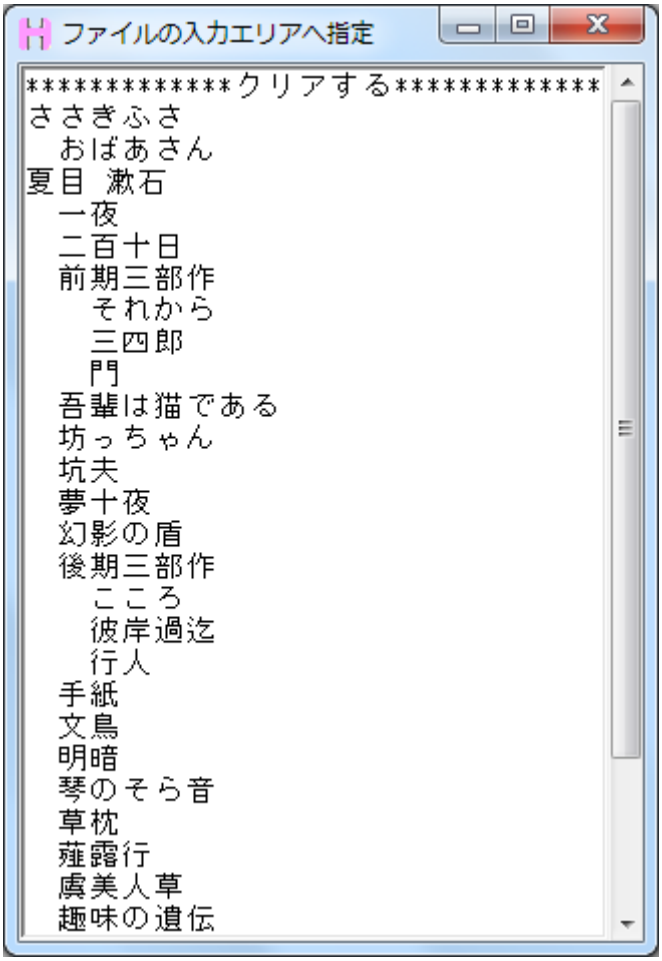
ファイルの結合が完了すると整形が開始されます。テキストの分量が多くなるので、整形には時間がかかります。

結合された内容はフォルダ名と同じ名前のテキストファイルとして HASHI フォルダの中に置かれます。

複数ファイルを一括で選択した場合は、最初からファイルタグが付与されています。



元のファイルが格納されたフォルダ構造がそのまま記録されています。



ファイルタグの指定のリストでは、元ファイルのフォルダ構造ごとに上位分類、下位分類の関係で一括指定できます。つまり、元のフォルダ名を選択すれば、そのフォルダに入っていたファイルが一括で全て指定できます。

テキストの整形段階での行情報の付与

テキストファイルにあらかじめ一定のタグを付与しておけば、整形後すぐに行タグ、属性タグ、ファイルタグの行情報を扱うことができます。

※「く」からはじまる行はタグ情報と認識されますので、通常の本文では使用できません。

ファイルタグ、属性タグ、行タグの指定書式

ファイルタグ

```
〈#ファイルタグ#〉
```

ファイルタグは、「〈# #〉」で囲まれた中に記入します。

属性タグ

```
〈%属性タグ%〉
```

属性タグは、「〈% %〉」で囲まれた中に記入します。属性タグの内部項目の指定はできません。tagname ファイルか、整形後に「テキストデータの編集(Edit)」で指定します。

行タグ

```
〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
```

行タグは全体を「く 〉」で囲みます。その中に各行タグを指定しますが、「: :@, 」で囲まれた中に記入します。「, 」の後には必ず半角のスペースを入れます。

タグの記入位置

タグは本文内容とは別に独立した行に記入します。

ファイルタグはどのタグとも同じ行にならないように単独の行に記入します。

```
〈#ファイルタグ#〉
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
本文の文章・・・・・・
```

ファイルタグが記入された以降は次のファイルタグが出るまで同じファイルタグの範囲とされます。

属性タグは行タグと同じ行に記入し、必ず一番左端の行頭に記入します。

```
〈#ファイルタグ#〉
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
本文の文章・・・・・・
```

属性タグは半角のスペースを含めることができません。必ず連続した文字列にします。
行タグは属性タグと同じ行に記入し、属性タグに続いて右側に記入します。

```
〈#ファイルタグ#〉
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
本文の文章・・・・・・
```

行タグの数は自由で、１つでも複数でも可能です、必要な数だけ記入します。行タグの各項目は、左からの位置で決まりますので、各タグの行で一致させる必要があります。例えば、新聞データで「刊種別、面名、ページ」の行タグを付与する場合は、全ての行タグの行を「〈:刊種別:@, ： 面名:@, ： ページ:@, 〉」同じ順番で記入します。
属性タグでも行タグも、もしどれかの行だけ項目が不明なものがあった場合、タグ形式は維持したままで中を記入しないようにします。例えば、先程の例で１つの行だけ面名が不明な場合は、「〈:刊種別:@, ::@, ： ページ:@, 〉」とします。空白も空けずに単に何も記入しません。
属性タグ、行タグは、記入された次以降の行にある本文のタグになります。次の属性タグ、行タグが出るまで、そのタグ情報の範囲となります。

ファイルタグ、行タグ、属性タグとも、必要がなければ１つも記入しなくても構いません。
ファイルタグのみや、行タグ１つのみの記入でも構いません。

```
〈#ファイルタグ#〉
本文の文章・・・・・・
```

```
〈:行タグ１:@, 〉
本文の文章・・・・・・
```

以降にタグ付与の例をいくつか示します。

完全に毎行タグを付与する方式

```
〈#ファイルタグ#〉
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
本文の文章・・・・・・
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
本文の文章・・・・・・
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
本文の文章・・・・・・
〈#ファイルタグ#〉
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
```

いくつかの行で属性タグ、行タグを共有する方式

```
〈#ファイルタグ#〉
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
本文の文章・・・・・・
本文の文章・・・・・・
本文の文章・・・・・・
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
本文の文章・・・・・・
〈#ファイルタグ#〉
〈%属性タグ%〉〈:行タグ１:@, :行タグ２:@, :行タグ３:@, 〉
```

ファイルタグのみの方式

```
〈#ファイルタグ#〉
本文の文章・・・・・・
本文の文章・・・・・・
本文の文章・・・・・・
〈#ファイルタグ#〉
本文の文章・・・・・・
本文の文章・・・・・・
```

属性タグのみの方式

```
〈%属性タグ%〉
本文の文章・・・・・・
〈%属性タグ%〉
本文の文章・・・・・・
〈%属性タグ%〉
本文の文章・・・・・・
```

行タグのみの方式

```
〈:行タグ１:@, :行タグ２:@, 〉
本文の文章・・・・・・
〈:行タグ１:@, :行タグ２:@, 〉
本文の文章・・・・・・
〈:行タグ１:@, :行タグ２:@, 〉
本文の文章・・・・・・
```

実際の整形例として、以下のようにテキストにタグを付与しておくとします。

```
<#吾輩は猫である#>↓
<%夏目漱石%><:明治後期:@, :ホトトギス:@, :連載:@, >↓
　吾輩は猫である。名前はまだ無い。↓
　どこで生れたかとんと見当がつかぬ。何でも薄暗いじめじめした所でニャーニャー泣いてい
　この書生の掌の裏でしばらくはよい心持に坐っておったが、しばらくすると非常な速力で運
　ふと気が付いて見ると書生はいない。たくさんおった兄弟が一疋も見えぬ。肝心の母親さえ
<#或阿呆の一生#>↓
<%芥川龍之介%><:昭和初期:@, :改造:@, :短編:@, >↓
　僕はこの原稿を発表する可否は勿論、発表する時や機関も君に一任したいと思つてゐる。↓
　君はこの原稿の中に出て来る大抵の人物を知つてゐるだらう。しかし僕は発表するとして
　僕は今最も不幸な幸福の中に暮らしてゐる。しかし不思議にも後悔してゐない。唯僕の如き
　最後に僕のこの原稿を特に君に托するのは君の恐らくは誰よりも僕を知つてゐると思ふか
<#にごりえ#>↓
<%樋口一葉%><:明治中期:@, :文芸倶楽部:@, :短編:@, >↓
　おい木村さん信さん寄つてお出よ、お寄りといつたら寄つても宜いではないか、又素通り
　店は二間間口の二階作り、軒には御神燈さげて盛り鹽景氣よく、空壜か何か知らず、銘酒と
　お高は往來の人のなきを見て、力ちやんお前の事だから何があつたからとて氣にしても居ま
　やがて雁首を奇麗に拭いて一服すつてポンとはたき、又すいつけてお高に渡しながら氣を
<#山椒大夫#>↓
<%森鴎外%><:大正:@, :中央公論:@, :短編:@, >↓
　越後の春日を経て今津へ出る道を、珍らしい旅人の一群れが歩いている。母は三十歳を踰え
　道は百姓家の断えたり続いたりする間を通っている。砂や小石は多いが、秋日和によく乾い
　藁葺きの家が何軒も立ち並んだ一構えが柞の林に囲まれて、それに夕日がかっとさしてい
「まあ あの美しい紅葉をごらん」と、先に立っていた母が指さして子供に言った。↓
<#文七元結#>↓
<%三遊亭円朝%><:明治中期:@, :やまと新聞:@, :落語:@, >↓
　さてお短いもので、文七元結の由来という、ちとお古い処のお話を申上げますが、只今と
　長「おう今｜帰ったよ、お兼……おい何うしたんだ、真暗に為て置いて、燈火でも点けね
　兼「あゝ此処にいるよ」↓
　長「真暗だから見えねえや、鼻ア撮まれるのも知れねえ暗え処にぶっ坐ッてねえで、燈火
　兼「お燈明どこじゃアないよ、私は今帰ったばっかりだよ、深川の一の鳥居まで往って来
```

これを整形すると以下のようにタグが保存されます。

行	語数	ファイル	話者	行タグ1	行タグ2	行タグ3	
1	9	吾輩は猫である	夏目漱石	明治後期	ホトトギス	連載	吾輩 は 猫 で ある 。 名前 は まだ 無い 。
2	268	吾輩は猫である	夏目漱石	明治後期	ホトトギス	連載	どこ で 生れ た か とんと 見当 が つか ぬ 。 何 で も
3	92	吾輩は猫である	夏目漱石	明治後期	ホトトギス	連載	この 書生 の 掌 の 裏 で しばらく は よい 心持 に 坐っ
4	88	吾輩は猫である	夏目漱石	明治後期	ホトトギス	連載	ふと 気 が 付い て 見る と 書生 は いない 。 たくさん
5	25	或阿呆の一生	芥川龍之介	昭和初期	改造	短編	僕 は この 原稿 を 発表 する 可否 は 勿論 、 発表 する
6	40	或阿呆の一生	芥川龍之介	昭和初期	改造	短編	君 は この 原稿 の 中 に 出 て 来る 大抵 の 人物 を 知
7	74	或阿呆の一生	芥川龍之介	昭和初期	改造	短編	僕 は 今 最も 不幸 な 幸福 の 中 に 暮らし て ゐる 。
8	52	或阿呆の一生	芥川龍之介	昭和初期	改造	短編	最後 に 僕 の この 原稿 を 特 に 君 に 托する の は 君
9	703	にごりえ	樋口一葉	明治中期	文芸倶楽部	短編	おい 木村 さん 信 さん 寄つ て お 出よ 、 お 寄り と
10	352	にごりえ	樋口一葉	明治中期	文芸倶楽部	短編	店 は 二 間 間口 の 二 階 作り 、 軒 に は 御 神 燈 さ
11	318	にごりえ	樋口一葉	明治中期	文芸倶楽部	短編	お 高 は 往 來 の 人 の なき を 見 て 、 力 ちやん と
12	186	にごりえ	樋口一葉	明治中期	文芸倶楽部	短編	やがて 雁首 を 奇麗 に 拭い て 一服 すつ て ポン と は
13	183	山椒大夫	森鴎外	大正	中央公論	短編	越後 の 春日 を 経 て 今津 へ 出る 道 を 、 珍らしい
14	56	山椒大夫	森鴎外	大正	中央公論	短編	道 は 百姓 家 の 断 え たり 続い たり する 間 を 通っ
15	31	山椒大夫	森鴎外	大正	中央公論	短編	藁葺き の 家 が 何 軒 も 立ち並ん だ 一 構え が 柞 の
16	21	山椒大夫	森鴎外	大正	中央公論	短編	「 まあ あの 美しい 紅葉 を ごらん 」 と 、 先 に 立っ
17	592	文七元結	三遊亭円朝	明治中期	やまと新聞	落語	さて お 短い もの で 、 文 七 元結 の 由来 という 、
18	48	文七元結	三遊亭円朝	明治中期	やまと新聞	落語	長「おう 今 ｜ 帰っ た よ 、 お 兼 …… おい 何う
19	6	文七元結	三遊亭円朝	明治中期	やまと新聞	落語	兼「あゝ 此処 に いる よ」
20	36	文七元結	三遊亭円朝	明治中期	やまと新聞	落語	長「真暗 だ から 見え ねえ や 、 鼻 ア 撮ま れる の
21	54	文七元結	三遊亭円朝	明治中期	やまと新聞	落語	兼「お 燈明 どこ じゃ ア ない よ 、 私 は 今 帰っ た
22	8	文七元結	三遊亭円朝	明治中期	やまと新聞	落語	長「エ 、 お 久 が 、 何処 へ 往っ た ん だ 」
23	64	文七元結	三遊亭円朝	明治中期	やまと新聞	落語	兼「何処 へ 往っ た か 解ら ない から 方々 探し て
24	81	文七元結	三遊亭円朝	明治中期	やまと新聞	落語	長「ナ … ナニ 知れ ねえ 、 年頃 の 娘 だ 、 え 、

ファイルタグ名、属性タグ名、行タグ名の指定

各行情報のタグ名は、あらかじめ指定するこができます。通常、ファイルタグは「ファイルタグ」、属性タグは「話者」、行タグは「行タグ 1」「行タグ 2」「行タグ 3」. . . というタグ名が与えられますが、それをテキスト整形前に指定することができます。整形するテキストファイルと同じ場所に「tagname_ファイル名」というテキストファイルで指定します。例えば「小説.txt」というテキストファイルを整形する際は「tagname_小説.txt」とします。

ファイルタグ名、属性タグ名、行タグ名の指定書式

ファイルタグ名

```
<#ファイルタグ名#>
```

ファイルタグ名は、「<# #>」で囲まれた中に記入します。

属性タグ名

```
<%属性タグ名%><項目 1 名><項目 2 名><項目 3 名>
```

属性タグ名は、「<% %>」で囲まれた中に記入します。
属性内の各項目名は「< >」で囲まれた中に記入します。

属性タグの内部項目

```
#<%属性１%><項目 1><項目 2><項目 3>
#<%属性１%><項目 1><項目 2><項目 3>
#<%属性１%><項目 1><項目 2><項目 3>
```

各属性タグの内部項目は整形するファイル内では指定できませんので、tagname ファイルで指定します。「#」で始まる行に記述します。その後は属性タグ名と内部項目名を指定するのと同様の書式で指定します。

行タグ名

```
<:行タグ１名:@, :行タグ２名:@, :行タグ３名:@, >
```

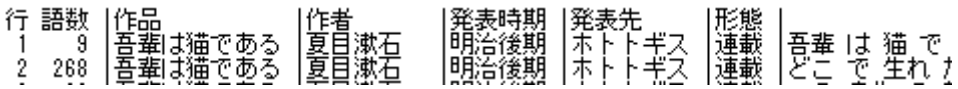
行タグ名は全体を「< >」で囲みます。その中に各行タグ名を指定しますが、「: :@, 」で囲まれた中に記入します。「, 」の後には必ず半角のスペースを入れます。

具体的には、前述の整形例のテキストの場合以下のような tagname ファイルを設定します。

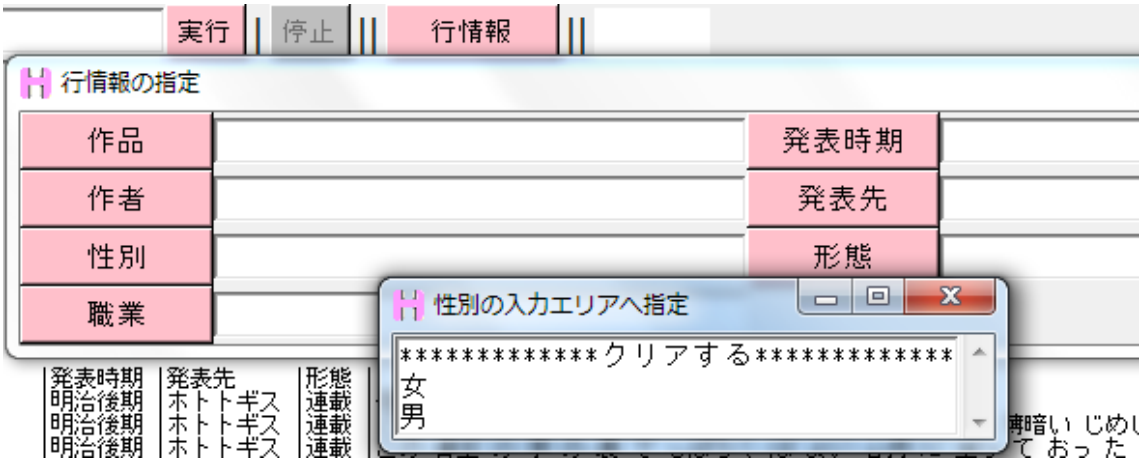
```
<#作品#>
<%作者%><性別><職業>
<:発表時期:@, :発表先:@, :形態:@, >

#<%夏目漱石%><男><教師>
#<%芥川龍之介%><男><小説家>
#<%樋口一葉%><女><小説家>
#<%森鴎外%><男><医者>
#<%三遊亭円朝%><男><落語家>
```

この内容のファイルを作成しておくと、テキストファイル整形時に読み込まれ、各行情報のタグ名が置き換わります。



行情報の指定時にも反映され、属性タグの内部項目も指定どおりに割り振られます。



テキストデータ編集での行情報のタグ名の変更

各タグ名は「テキストデータの編集(Edit)」でも変更できます。

tagname ファイルを用意しなかった場合や、整形後にタグ名の変更する必要ができた場合などは、「テキストデータの編集(Edit)」によって修正、変更をします。

小説数千冊、新聞 1 年分など、膨大なデータを扱う場合、テキストデータの編集はパソコンのメモリの関係で行えない可能性が高いため、そういう際は整形前に tagname ファイルを用意してタグ名の指定を行います。

語単位の整形ルールの変更

語単位は形態素解析ソフトでの解析結果である「形態素単位」を一定のルールに基づいて結合して少し大きな単位を再現するものです。これを変更することでテキスト整形の段階で、語の区切りや語に付く文法タグを一括で自由に変更できます。語単位の整形は、67ページで提示されたルールで行われますが、ルールモジュールファイルを書きかえることで変更できます。大幅な整形ではなくテキストを少しだけ変更したいときは、ルールのほとんどを消して１つ２つだけ加えるなどで、好みの整形結果にできます。

書き換えたルールファイルは、HASHI のフォルダ内の「bin」→「Format_Rules」フォルダの中に入れます。ルールファイルは、UniDic 版は「O_Uni_Format_Rules.pm」、IPADic 版は「O_Ipa_Format_Rules.pm」です。

以下に整形ルールの書式を示します。

書式

[{その語の適合ルール}, {次の語の適合ルール}], {整形ルール}]

※語を結合せずに１語の中でのみの変形なら、**{次の語の適合ルール}**を記入しない
※結合後の語の各項目が全てデフォルトなら、**{整形ルール}**を記入しない

適合ルール

{' 適合ルール１', ' 適合ルール２', ...}

各適合ルールの内部書式

' 項目' =>' 文字列'

整形ルール

{' 整形ルール１', ' 整形ルール２', ...}

各適合ルールの内部書式

' 項目名' =>' 変更後文字列'

以下に、整形例を示します。

品詞「形容詞」を、品詞「イ形容詞」に変更

```
[ [{' 品詞' =>' 形容詞' } ], {' 品詞' =>' イ形容詞' }],
```

表記形「と」と、後続の表記形「か」を結合

```
[{' 表記形' =>' と' }, {' 表記形' =>' か' }] ],
```

下位分類「固有名詞-人名-姓」と、後続の下位分類「固有名詞-人名-姓」を結合し、下位分類「固有名詞-人名-フルネーム」にする

```
[ [{' 下位分類' =>' 固有名詞-人名-姓' }, {' 下位分類' =>' 固有名詞-人名-名' }], {' 下位分類' =>' 固有名詞-人名-フルネーム' } ],
```

品詞「形状詞」下位分類「タリ」と、基本形「と」品詞「助詞」を結合し、品詞「副詞」下位分類「---」にする

```
[ [{' 品詞' =>' 形状詞', ' 下位分類' =>' タリ' }, {' 基本形' =>' と', ' 品詞' =>' 助詞' }], {' 品詞' =>' 副詞', ' 下位分類' =>' ---' } ],
```

下位分類「名詞的-サ変可能」と、後続の基本形「する」か「できる」か「なさる」か「いたす」を結合し、品詞「動詞」下位分類「一般」にする

```
[ [{' 下位分類' =>' 名詞的-サ変可能' }, {' 基本形' =>' (する|できる|なさる|いたす)' } ], {' 品詞' =>' 動詞', ' 下位分類' =>' 一般' } ],
```

品詞「動詞」と後続の品詞「助動詞」を結合し、前側の語の基本形、前側の語の活用形にする

```
[{' 品詞' =>' 動詞' }, {' 品詞' =>' 助動詞' }], {' 基本形' =>[0, ' 基本形' ], ' 活用型' =>[0, ' 活用型' ]} ],
```

品詞「動詞」と後続の品詞「助動詞」を結合し、後側の語の基本形、後側の語の活用形にする

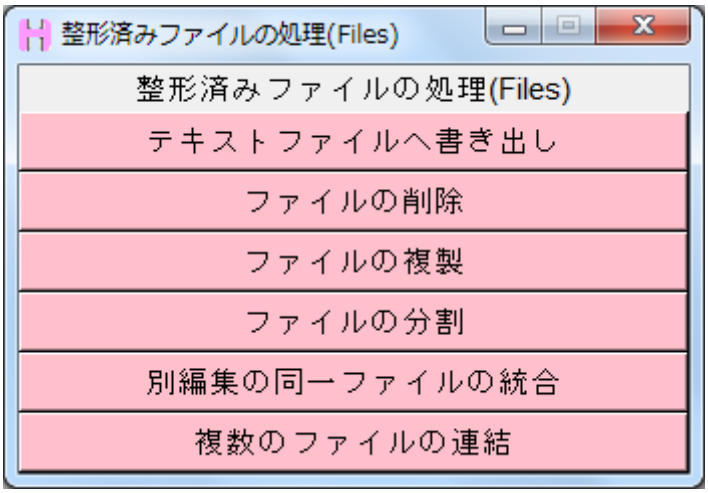
```
[{' 品詞' =>' 動詞' }, {' 品詞' =>' 助動詞' }], {' 基本形' =>[1, ' 基本形' ], ' 活用型' =>[1, ' 活用型' ]} ],
```

品詞「形状詞」下位分類「タリ」と、後続する表記形「たる」基本形「たり」品詞「助動詞」を結合し、品詞「連体詞」下位分類「---」活用形「---」活用型「---」にする

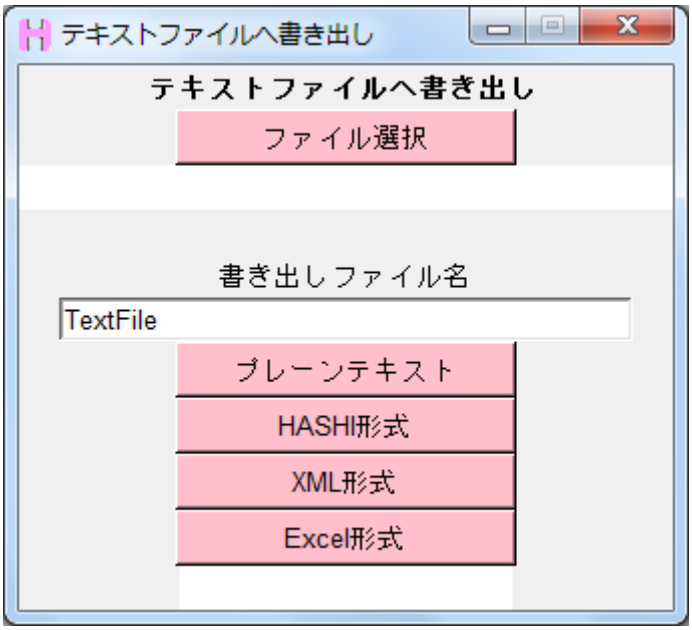
```
[ [{' 品詞' =>' 形状詞', ' 下位分類' =>' タリ' }, {' 表記形' =>' たる', ' 基本形' =>' たり', ' 品詞' =>' 助動詞' }], {' 品詞' =>' 連体詞', ' 下位分類' =>' ---', ' 活用形' =>' ---', ' 活用型' =>' ---', } ],
```

整形済みファイルの処理(Files)

すでに整形されているファイルを対象に、ファイル単位での処理が行えます。

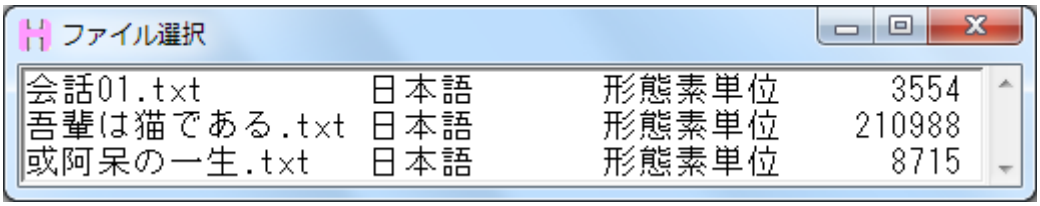


テキストファイルへの書き出し



整形したりオリジナルコーパスとして編集したファイルをテキストファイルに戻して書き出すことができます。ファイル形式はプレーンテキスト、HSHI 形式、XML 形式、Excel 形式から選べます。

「ファイルを選択」で書き出したいファイルを選択します。



その後に出力形式ボタンを押すとファイルの書き出しが始まります。

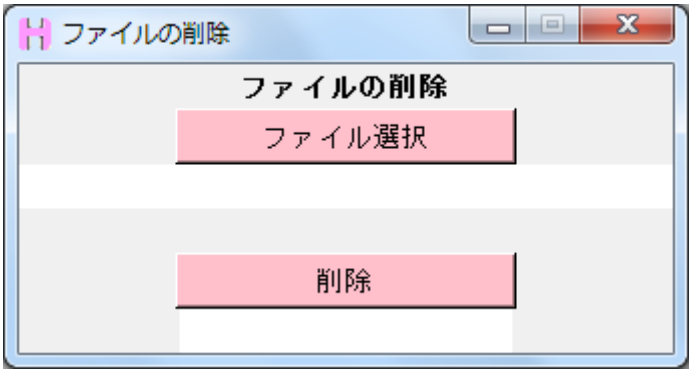
XML 形式は、使用されている語タグ、行タグ、属性タグが全て XML タグ化して出力されます。
ファイルタグごとに別々の XML ファイルとなり、選択ファイル名のフォルダに一括で保存されます。

```
<?xml version="1.0" encoding="UTF-8"?>
<hashiCorpus xml:id="会話01">
 - <u who="話者1">
    - <s n="1">
        <w mora="1" vowl="E" yomi="デ" type="---" conj="---" subc="格助詞" posp="助詞" morp="/で/" lema="で">だ</w>
        <w mora="1" vowl="O" yomi="モ" type="---" conj="---" subc="係助詞" posp="助詞" morp="/も/" lema="も">も</w>
        <w mora="2" vowl="AN" yomi="ナン" type="---" conj="---" subc="---" posp="代名詞" morp="/何/" lema="何">何</w>
        <w mora="3" vowl="AAE" yomi="ハナセ" type="五段-サ行" conj="仮定形-一般" subc="一般" posp="動詞" morp="/話す/" lema="話す">話せ</w>
        <w mora="1" vowl="A" yomi="バ" type="---" conj="---" subc="接続助詞" posp="助詞" morp="/ば/" lema="ば">ば</w>
        <w mora="2" vowl="I-" yomi="イイ" type="形容詞" conj="連体形-一般" subc="非自立可能" posp="形容詞" morp="/いい/" lema="いい">いい</w>
        <w mora="1" vowl="N" yomi="ン" type="---" conj="---" subc="準体助詞" posp="助詞" morp="/ん/" lema="ん">ん</w>
        <w mora="2" vowl="EU" yomi="デス" type="助動詞-デス" conj="終止形-一般" subc="---" posp="助動詞" morp="/です/" lema="です">です</w>
        <w mora="1" vowl="A" yomi="カ" type="---" conj="---" subc="終助詞" posp="助詞" morp="/か/" lema="か">か</w>
        <w mora="1" vowl="E" yomi="ネ" type="---" conj="---" subc="終助詞" posp="助詞" morp="/ね/" lema="ね">ね</w>
      </s>
  </u>
 - <u who="話者3">
    - <s n="2">
        <w mora="2" vowl="AN" yomi="ナン" type="---" conj="---" subc="---" posp="代名詞" morp="/なん/" lema="なん">なん</w>
        <w mora="1" vowl="A" yomi="カ" type="---" conj="---" subc="副助詞" posp="助詞" morp="/か/" lema="か">か</w>
        <w mora="1" vowl="E" yomi="デ" type="---" conj="---" subc="格助詞" posp="助詞" morp="/で/" lema="で">で</w>
        <w mora="1" vowl="O" yomi="モ" type="---" conj="---" subc="係助詞" posp="助詞" morp="/も/" lema="も">も</w>
        <w mora="3" vowl="O-E" yomi="コーレ" type="---" conj="---" subc="---" posp="代名詞" morp="/これ/" lema="これ">これ</w>
        <w mora="0" vowl="XXX" yomi="ぇ" type="---" conj="---" subc="普通名詞-一般" posp="名詞" morp="/ぇ/" lema="ぇ">ぇ</w>
        <w mora="0" vowl="XXX" yomi="ー" type="---" conj="---" subc="一般" posp="補助記号" morp="/ー/" lema="ー">ー</w>
        <w mora="1" vowl="O" yomi="ヲ" type="---" conj="---" subc="格助詞" posp="助詞" morp="/を/" lema="を">を</w>
        <w mora="2" vowl="AN" yomi="ナン" type="---" conj="---" subc="---" posp="代名詞" morp="/なん/" lema="なん">なん</w>
        <w mora="1" vowl="A" yomi="カ" type="---" conj="---" subc="副助詞" posp="助詞" morp="/か/" lema="か">か</w>
        <w mora="0" vowl="XXX" yomi="・" type="---" conj="---" subc="一般" posp="補助記号" morp="/・/" lema="・">・</w>
        <w mora="0" vowl="XXX" yomi="・" type="---" conj="---" subc="一般" posp="補助記号" morp="/・/" lema="・">・</w>
```

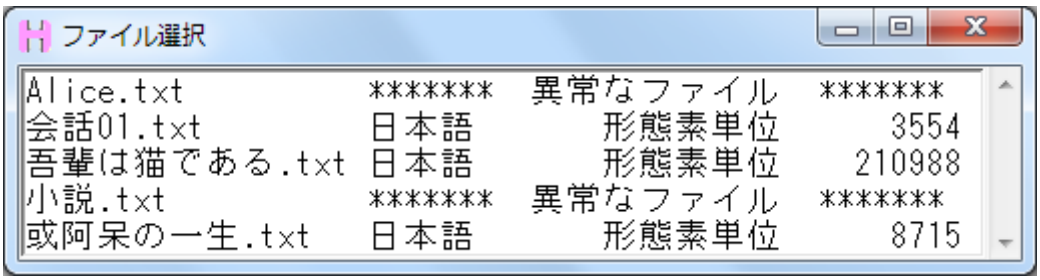
	1	2	3	4	5	6	7	8	9	10
4	3	改めて、はい話してくださいって言われるとぉー、なかなか、話せない	会話01	話者3	20代	男性	滋賀	保持	00:04.2	00:10.6
5	4	どうしよう・	会話01	話者3	20代	男性	滋賀	無し	00:10.6	00:11.3
6	5	うーんドイツの話とか聞かしていただいてもいいですか	会話01	話者3	20代	男性	滋賀		00:11.3	00:15.1
7	6	あ、いいですけど	会話01	話者1	20代	女性	岡山		00:15.1	00:16.1
8	7	どこに行ってたんですか	会話01	話者3	20代	男性	滋賀		00:16.1	00:17.6
9	8	ええっとね、デュースブルグって知ったはりますか・・	会話01	話者1	20代	女性	岡山		00:17.6	00:20.5
10	9	え	会話01	話者3	20代	男性	滋賀		00:20.5	00:20.6
11	10	ええっとね、デュースブルグ	会話01	話者1	20代	女性	岡山		00:20.6	00:22.1
12	11	デュースですか	会話01	話者3	20代	男性	滋賀		00:22.1	00:22.9
13	12	はい	会話01	話者1	20代	女性	岡山		00:22.9	00:23.2
14	13	分かんないです	会話01	話者3	20代	男性	滋賀		00:23.2	00:24.2
15	14	デュッセルドルフは、分かります	会話01	話者1	20代	女性	岡山		00:24.2	00:26.0
16	15	デュッセルドルフ、どこでしたっけ	会話01	話者3	20代	男性	滋賀		00:26.0	00:28.0
17	16	ええっとね	会話01	話者1	20代	女性	岡山		00:28.0	00:28.7
18	17	みなみ、	会話01	話者3	20代	男性	滋賀		00:28.7	00:29.1
19	18	えと西の、ちょい北あたりなんですけど、どこの近くかなぁ・・	会話01	話者1	20代	女性	岡山		00:29.1	00:32.9
20	19	あ、そうなんですか	会話01	話者3	20代	男性	滋賀		00:32.9	00:34.0
21	20	えっとね、ケルンとかの近くです	会話01	話者1	20代	女性	岡山		00:34.0	00:36.1
22	21	あ、ケルンあぁ～あぁ～	会話01	話者3	20代	男性	滋賀		00:36.1	00:37.2
23	22	はいはいはい、ケルンから電車で40分くらいかな	会話01	話者1	20代	女性	岡山		00:37.2	00:40.8
24	23	あ、そうなんですね～～	会話01	話者3	20代	男性	滋賀		00:40.8	00:41.9

Excel 形式は、全てのファイルタグ、属性タグと内部項目、行タグを本文の右側にセルに分かれる形式で、タブ区切りで出力されます。語タグは使用されません。

ファイルの削除



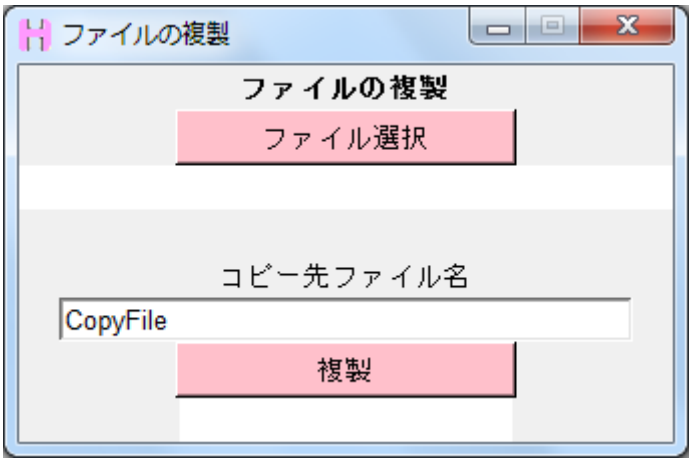
整形したテキストファイルが不要になった場合に削除することができます。



「ファイル選択」ボタンで、整形済みファイルのリストが表示されますので選択します。ここには整形が失敗したり、編集結果の保存を途中で切ってしまった場合など、異常な内容になったファイルの名前も表示されます。これも削除することができます。通常のファイル選択や他の処理では異常なファイルは表示されません。ファイルを選択したら「削除」ボタンを押します。

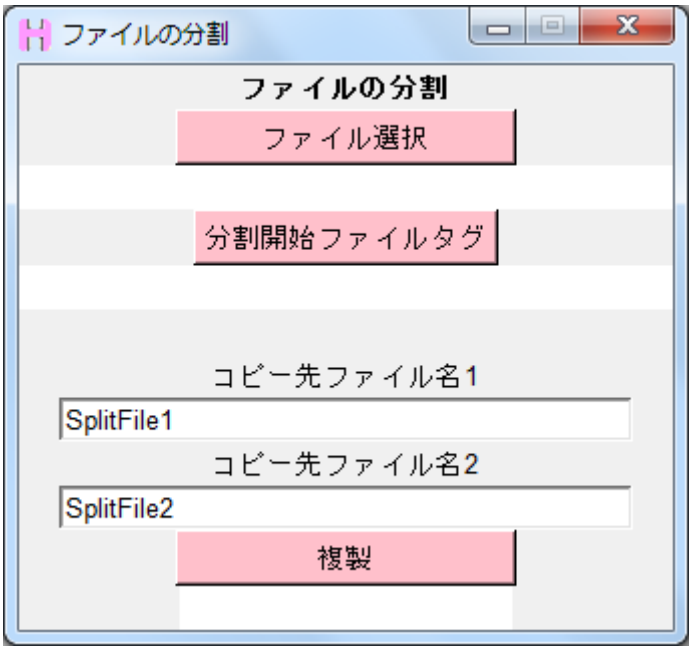
削除の最終確認をしてきますので、「Yes」をするとファイルは削除されます。

ファイルの複製

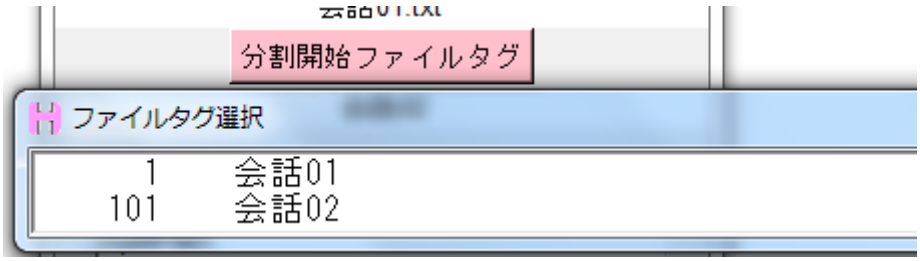


ファイルの複製が行えます。内部に保持されたタグ情報は全てそのままコピーされます。

ファイルの分割



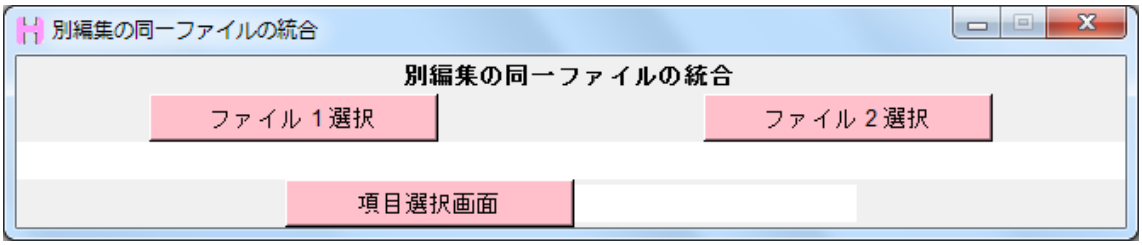
ファイルの分割を行います。分割の位置はファイルタグごとに行えます。ファイルを選択後に「分割開始ファイルタグ」ボタンを押すと、そのテキスト内のファイルタグの一覧が表示されます。最初のファイルタグは選択できませんので２つ目以降を選択します。



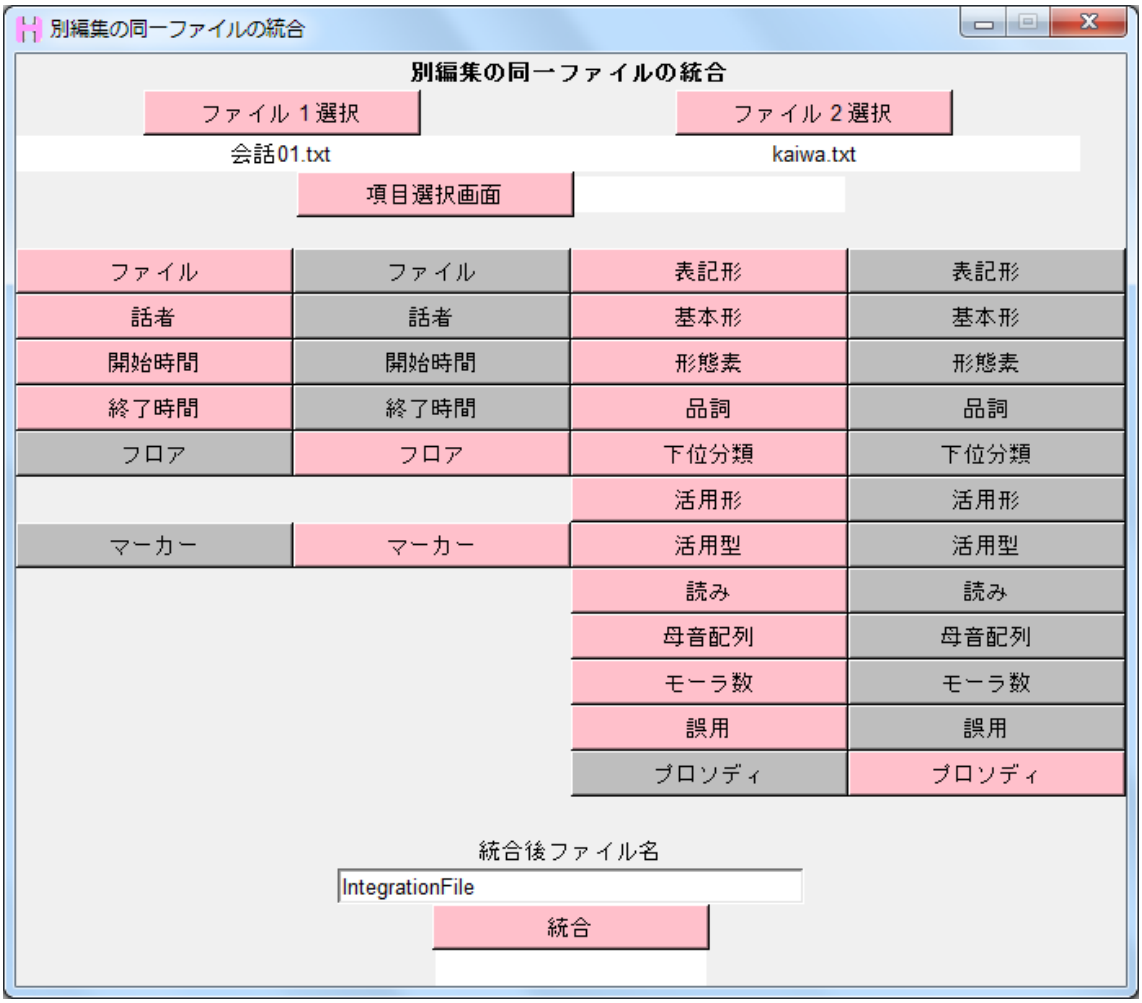
分割をしたい行から新しいファイルタグが始まっていない場合は、Edit でのその行にファイルタグを設定します。

別編集の同一ファイルの統合

１つのファイルに対して、複数のスタッフがそれぞれの担当のタグの付与をしている場合、後で各自が編集した各タグの統合が必要になります。そのための処理が行えます。

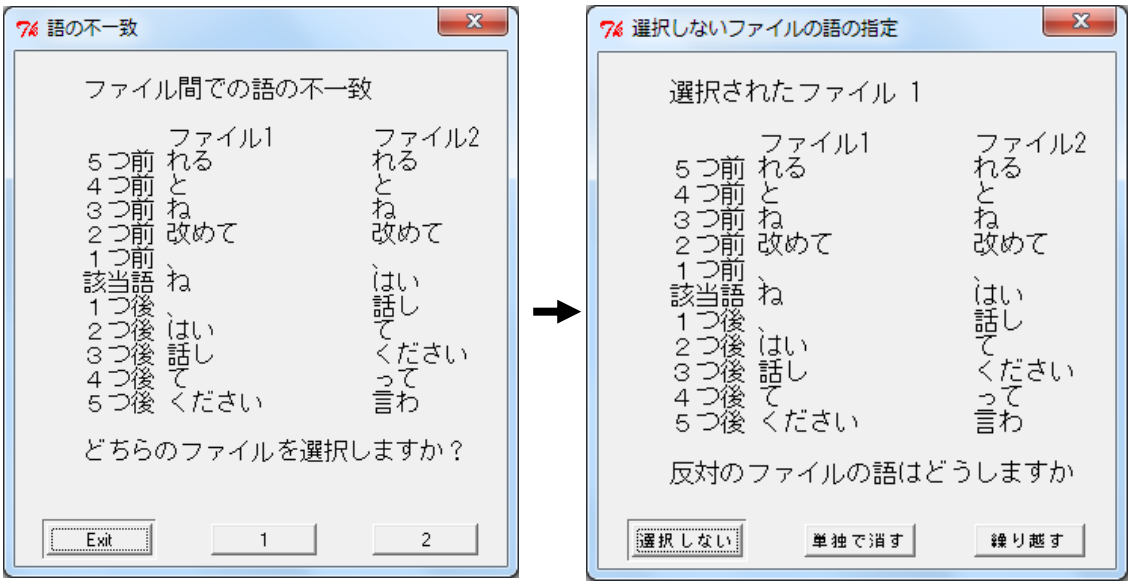


「ファイル１」と「ファイル２」で編集別のファイルをそれぞれ選択し、「項目選択画面」ボタンを押します。



２つのファイルで使われている全項目のボタンが出ます。それぞれ、左のボタンがファイル１、右のボタンがファイル２です。各項目をどちらのファイルから使用するか全て選択し、「統合」ボタンで統合が開始されます。

テキスト内容を修正したりなどで２つのファイルで使われている語が一致しない個所は、小さなウィンドウが出て、どちらのファイルの語を使用するか個別に選択します。



たとえば、２つのファイルで「改めて、ね、はい話して」と「改めて、はい話して」のように違う箇所が有ったとします。まず「ね」の箇所が違うので、統合処理中に「ね」の箇所に来たら、「ね」と、もう片方のファイルで同じ語の位置の「はい」とのどちらを使用するか選択します。選択は、ファイルの番号で指定します。

次に、選択しなかった方の語である「話し」をどうするか聞いてきます。

「選択しない」は、単純に「話し」を選択せずに、その位置の語は「ね」を使用し、また次の語の一致の確認に流れ「ね」の次の「、」と、「話し」の次の「て」の比較になります。

「単独で消す」は、「ね」の使用をいったん保留しておいて、「話し」だけをその行から消します。次には、また「ね」と、「話し」の次の「て」の比較になります。

「繰り越す」は、その位置は「ね」を使用しますが、相手のファイルの「話し」は消さずに、次の位置の比較に繰り越します。次の位置では「ね」の次の「、」と、「話し」の比較になります。

このように語数の違うファイルの統合もできますが、行数の違う、または行の入れ替えが行われたファイル同士の対応はできません。その際は、「テキストデータの編集(Edit)」の「テキスト本体編集」から「行の編集」で２つのファイルの行数を合わせてから統合をします。

複数のファイルの連結

全く別のファイルを順番に連結して大きなコーパスとすることができます。ファイル１から順に選択し、「連結」ボタンを押すと１つになって保存されます。保存されたファイルでは、ファイルタグは元のファイルのファイルタグが引き継がれます。ファイルタグが付いていないファイルを連結した際は元のファイルのファイル名がその範囲のファイルタグとされます。

整形データの移動、配布

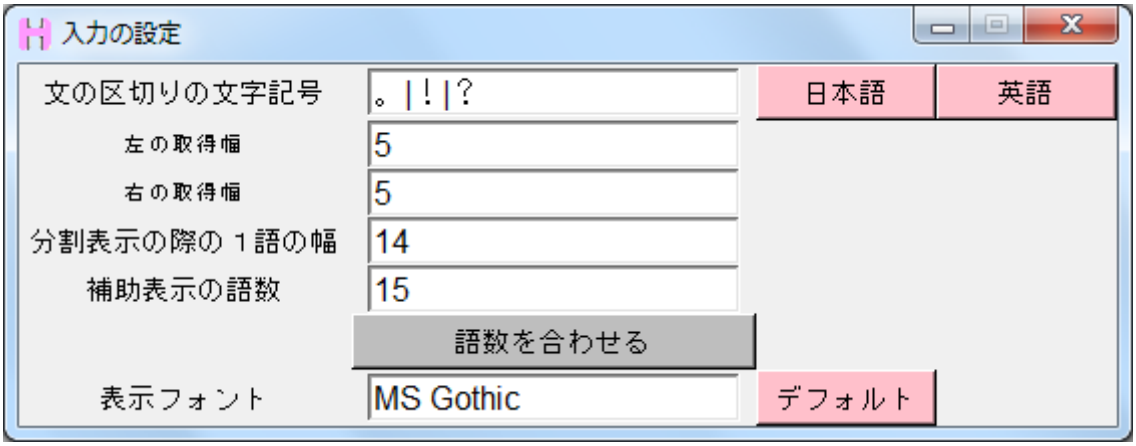
HASHI で使用される、整形や編集されたテキストデータは「FormatFiles」というフォルダに全て入っています。このフォルダの中にはさらにいくつかのフォルダが有り、「SoundFiles」が音声ファイルを入れるフォルダ、「Test」が形態素解析ソフトの解析結果を試す一時ファイルが入るフォルダですが、それ以外のフォルダにはテキストファイルの整形結果が入っています。この各テキストファイルの整形結果のフォルダの中身がそれぞれのコーパスデータと言えます。コーパスデータのフォルダの名前は、最初に読み込んだテキストファイルの名前になっています。「整形済みファイルの処理(Files)」で、複製や統合などをしたフォルダはその作業時に指定した名前になっています。

１つの HASHI の中でデータを複製する場合は、「整形済みファイルの処理(Files)」で「ファイルの複製」で行います。１つのファイルの各項目を複数人で同時に編集したい場合など複数人でデータを共有するために複製する場合は、各コーパスデータのフォルダをそのまま他の HASHI の「FormatFiles」へコピーします。通常のパソコン操作でそのままフォルダごと移動すれば他の HASHI でそのデータがそのまま使えるようになります。ただし、バージョンの違う HASHI へ移動しても読み込まれないことがあります。ver0.8.8 台と ver0.8.9 台と ver0.8.10 台ではそれぞれ内部のデータ形式が全く違うため、データの共有はできません。

データをオリジナルコーパスとして編集し、配布する場合も「FormatFiles」の中の編集したデータの入っているフォルダをそのまま配布すれば他の人とデータの共有ができます。その場合、編集に使用した HASHI のバージョンを明記してください。

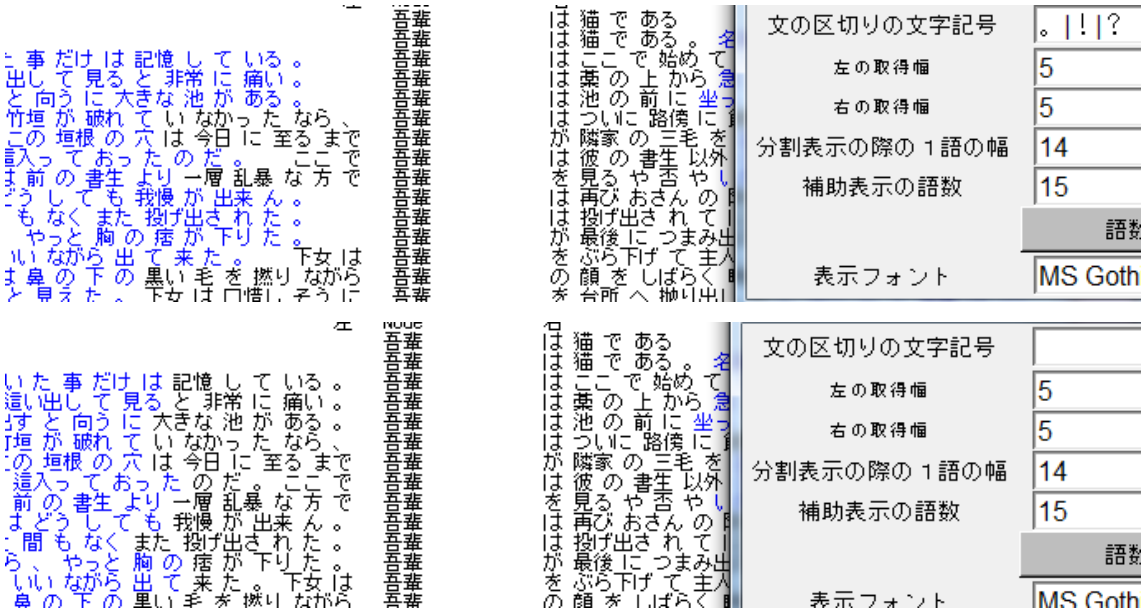
Input menu

各処理のウィンドウ上部のツールバーにある「Input menu」をクリックすると、検索の際のデータ取得範囲やデータの区切りの扱いなどを変更できます。
このメニューは Ngram,　Edit, grep 以外のすべて出てきますが、基本的に KWIC, Collocates, Picture, POPAK 以外ではほぼ意味を持ちません。



文の区切り文字記号

本ソフトでは通常文の区切りを超えてデータを取得しません。これは、ある語と近隣の語の共起関係を考える際に語の並び上たまたま近かっただけで、別の文にある語との関係性が強いと言えるかに疑問があるためです。このために文の区切りを表す記号が指定されています。これを変更できます。

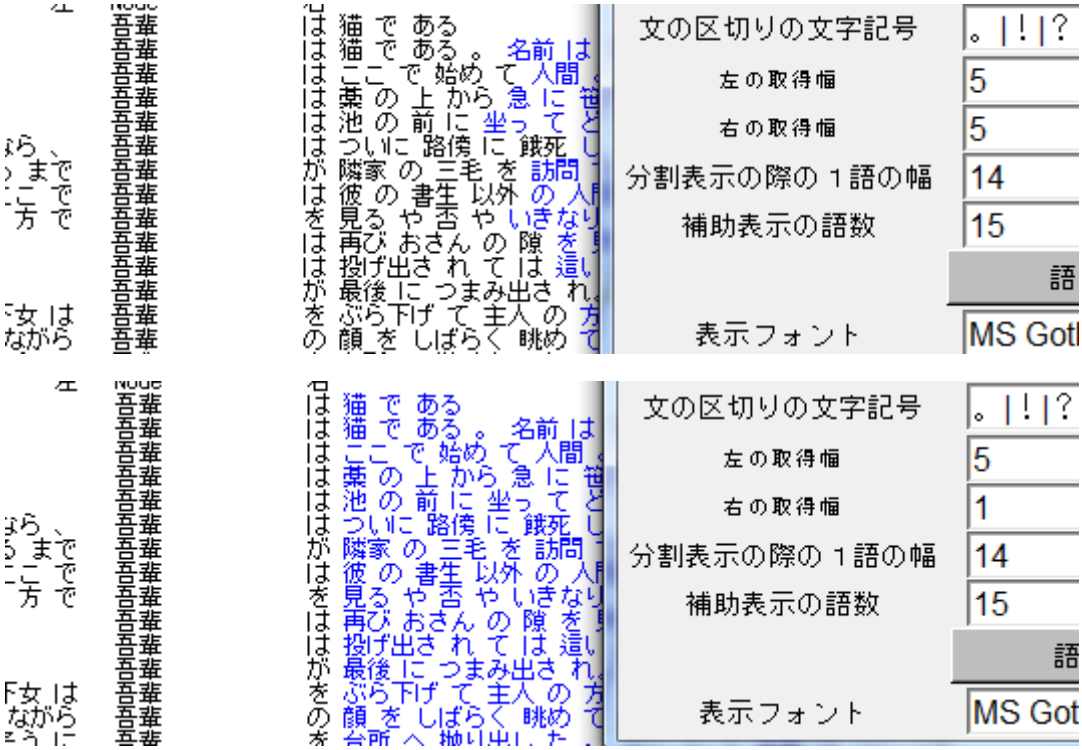


区切りの文字記号を別のものに変えると変えた文字が現れた箇所が文の区切りと認識され

ます。また区切り記号を何も指定しなければ文の区切りを設けませんので、左右の幅いっぱいまでデータが取得されます。
ただし、本ソフトでは行単位でしかデータを取得しません。つまり、文末記号が無かったとしても元々テキスト自体がそこで改行されていたらそれを超えては左右の幅としてデータを取得しません。
入力ボックスの右の「日本語」を押すと、日本語での標準の文の区切りである「。|！|？」が指定されます。「英語」を押すと、英語の標準の文の区切りである「.|!|?」が指定されます。ここでの指定は正規表現が使えますので、２つ以上を指定する場合は | で区切ります。

左の取得幅、右の取得幅

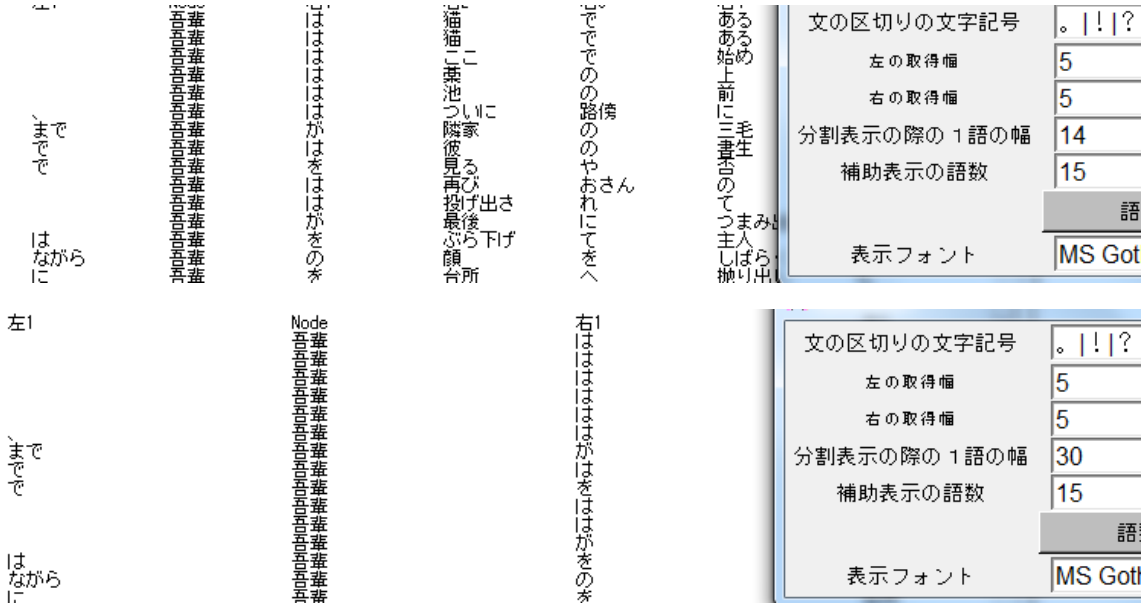
データとして取得する左右の幅を個別に指定できます。



取得結果が変わると統計処理での結果も変わります。

分割表示の際の１語の幅

分割表示の際に位置ごとの幅を変えられます。



長い文字幅の語が多く、１つの位置に収まりきれずに表示が大きく崩れるような際に使います。数字を変更後に「再描写」をします。

統計を扱うような他の処理でも長い語が多く表が崩れるような際に使います。

位置ごとの共起語の頻度

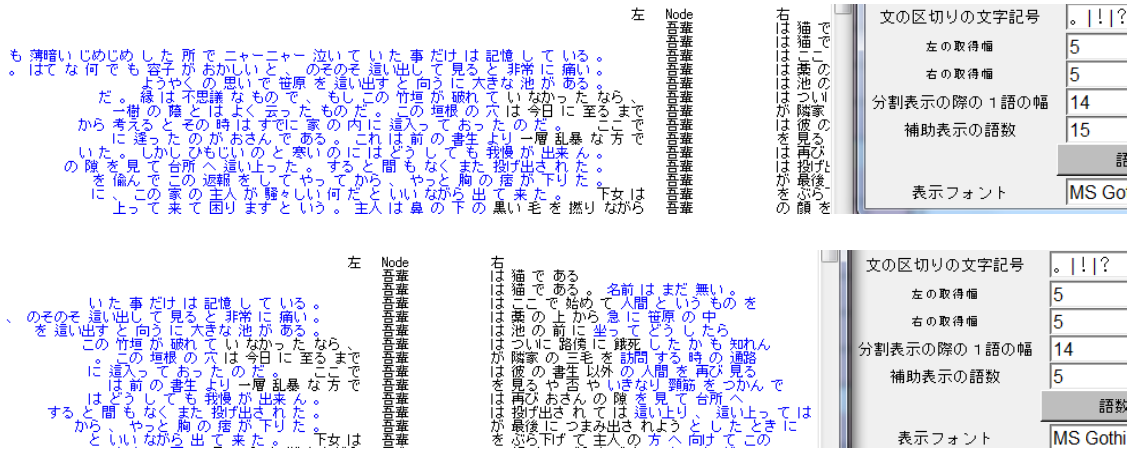
	左5	左4	左3	左2	左1	Node	右1	右2
1	--- 170	--- 157	--- 182	--- 196	--- 244	--- 482	--- 473	--- 413
2	五段-ラ行-一般 8	助動詞-ダ 17	五段-ラ行-一般 15	助動詞-ダ 22	助動詞-ダ 18			助動詞
3	助動詞-ダ 8	五段-ラ行-一般 14	助動詞-タ 15	五段-ラ行-一般 10	形容詞 9		文語助動詞-ゴトシ 2	五
4	形容詞 6	形容詞 8	助動詞-ダ 8	五段-ワア行-イウ 7	五段-ラ行-一般 4			五段-
5	サ行変格 5	上一段-ア行 4	サ行変格 5	サ行変格 5	助動詞-タ 4			助動詞-ダ 6
6	上一段-ア行 5	五段-ワア行-一般 4	カ行変格 4	形容詞 4	助動詞-ヌ 3			形容詞 5
7	五段-カ行-一般 3	助動詞-タ 4	五段-ワア行-イウ 4	下一段-ア行 3		サ行変格 2		
8	下一段-ア行 2	上一段-マ行 3	助動詞-ヌ 4	五段-カ行-一般 3	上一段-ア行 2			五段-ワア行-
9	下一段-ダ行 2	下一段-ダ行 3	五段-ワア行-一般 3	五段-サ行 3	五段-カ行-一般 2			
10	五段-サ行 2	五段-サ行 3	上一段-ア行 2	助動詞-ヌ 3	文語助動詞-タリ-断定 2			文語助動詞-ニ

位置ごとの共起語の頻度

	左5		左4		左3	
1	---	170	---	157	---	182
2	五段-ラ行-一般	8	助動詞-ダ	17	五段-ラ行-一般	15
3	助動詞-ダ	8	五段-ラ行-一般	14	助動詞-タ	15
4	形容詞	6	形容詞	8	助動詞-ダ	8
5	サ行変格	5	上一段-ア行	4	サ行変格	5
6	上一段-ア行	5	五段-ワア行-一般	4	カ行変格	4
7	五段-カ行-一般	3	助動詞-タ	4	五段-ワア行-イウ	4
8	下一段-ア行	2	上一段-マ行	3	助動詞-ヌ	3
9	下一段-ダ行	2	下一段-ダ行	3	五段-ワア行-一般	3
10	五段-サ行	2	五段-サ行	3	上一段-ア行	3

補助表示の語数

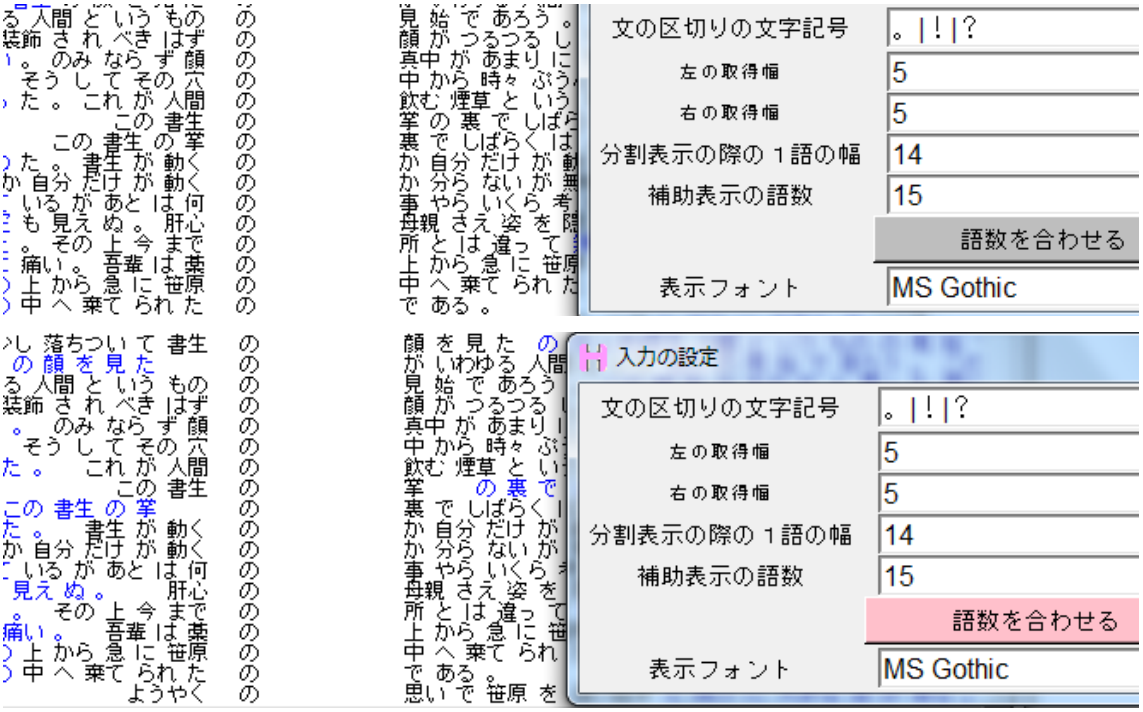
補助表示の際の文字幅を決めます。KWIC のみ意味を持ちます。



補助表示はあくまでも目視での確認用なので、この数値の変更はデータの取得結果に影響を及ぼしません。

語数を合わせる

取得する語の数の重複を避けることができます。

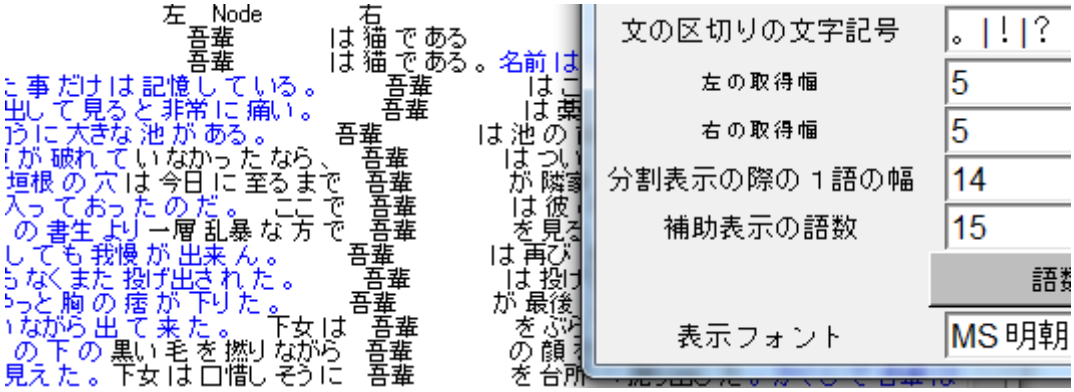


例えば「の」を検索して共起語を調べる際に「この 書生 の 掌 の 裏 で しばらく」という文があるとします。この中に「の」は２回でてきます。検索語の左右３語を共起語として保存する際に「この 書生 <u>の</u> 掌 の 裏」と「書生 の 掌 <u>の</u> 裏 で しばらく」とそれぞ

れ取得されます。検索語の左右一定の幅に有る語がその語の共起語と言えますが、「書生」「掌」「裏」は２つ検索語で別々に習得されています。実際には１度しか使われていない語ですが、共起語としては２と数えられています。これを防ぐために、検索語の左右に別の検索語が有った場合、前の検索語が取得した語は共起語として取得しないように選択できます。つまり「この 書生 の 掌」と「の 裏 で しばらく」として取得されるようになります。

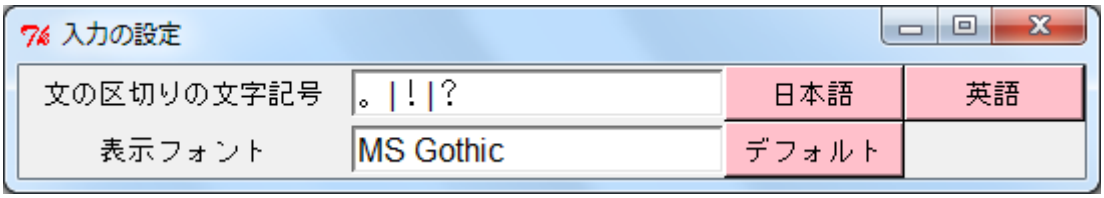
表示フォント

画面表示用のフォントを変更できます。



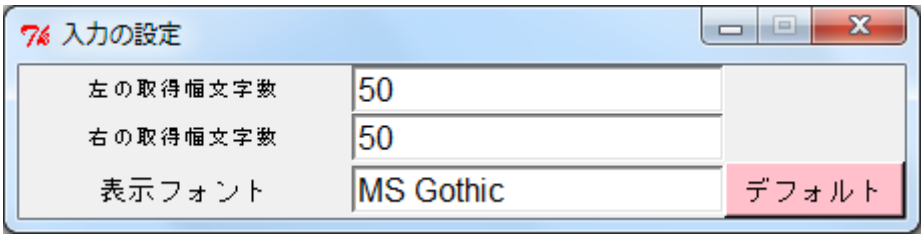
言語によっては KIWC の表示などが崩れることがあります。フォントの変更で対応できる可能性があるため、フォントを指定できるようになっています。リストはでませんので、直接入力します。基本的にフォントを変えると画面描写は崩れます。

Ngram での Input menu



Ngram では文の区切りを超えて作成されないようになっていますので、他の処理と同様に文の区切りを指定できます。

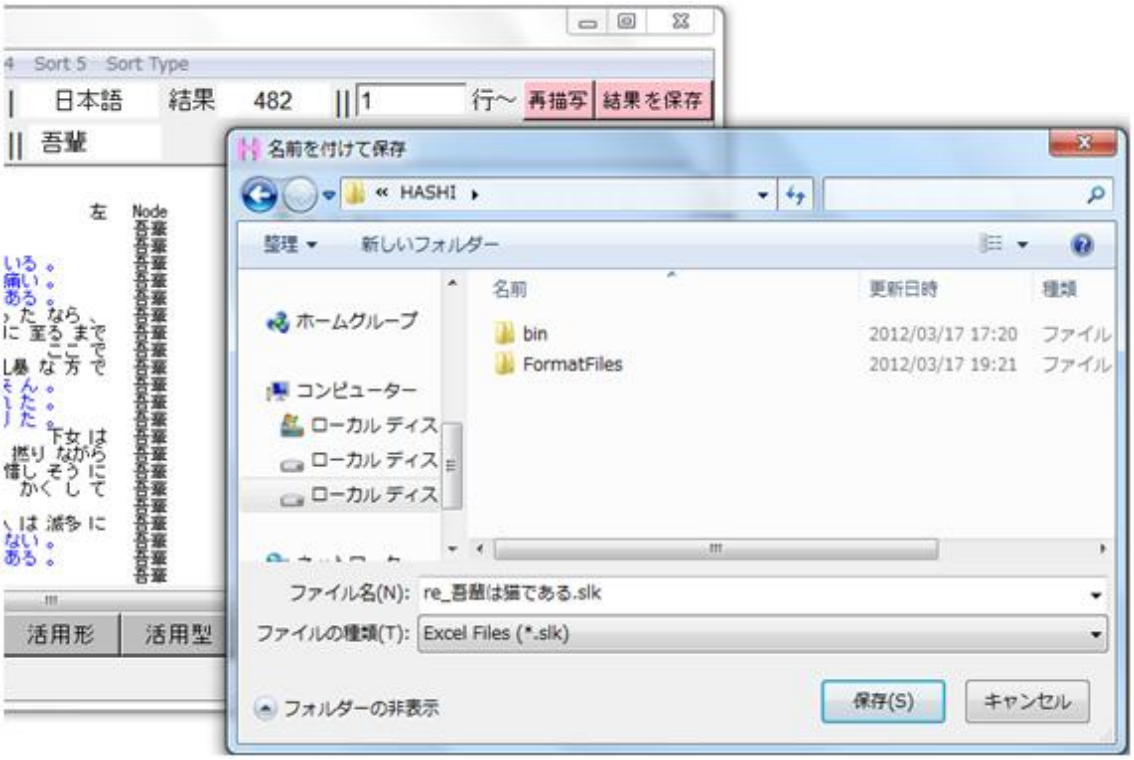
Grepd の Input menu



Grep では KWIC 形式の際の左右の取得文字数を変更できます。

ファイルへの保存

各処理の結果をファイルへ保存できます。



ウィンドウ上部の右端の「結果を保存」ボタンで各処理の結果をファイルへ書き出して保存することができます。保存の際には保存先や保存するファイル名を指定できます。ここで「ファイルの種類(T):」で拡張子を選択できますが、選択する拡張子によってファイルに書き込まれる形式が変わります。

.slk

MS-Excel 対応の sylk 形式で保存され、そのまま保存されたファイルをダブルクリックするだけで自動的に MS-Excel で読み込まれます。

	1	2	3	4	
1	検索語と使われている文脈(KWIC) 左5 - 右5 の範囲				
2		行番号	左	Node	右
3	1	1		吾輩	は猫である
4	2	4		吾輩	は猫である。名前 は まだ 無い 。
5	3	5	ーニャー 泣い て い た 事 だけ は 記憶 し て いる 。	吾輩	はここで始めて人間 という もの
6	4	7	い と 、 のそのそ 這い出し て 見る と 非常 に 痛い 。	吾輩	は薬の上から急 に 笹原 の 中 へ
7	5	8	で 笹原 を 這い出す と 向う に 大きな 池 が ある 。	吾輩	は池の前に坐っ て どう し たら よ
8	6	8	もの で 、 もし この 竹垣 が 破れ ていなかったなら、	吾輩	はついに路傍に餓死し た か も 知
9	7	8	云っ た もの だ 。この 垣根 の 穴は今日に至るまで	吾輩	が隣家の三毛を訪問 する 時 の
10	8	8	すでに 家 の 内 に 這入っ て おっ た の だ 。ここで	吾輩	は彼の書生以外の 人間 を 再び
11	9	8	で ある 。これ は 前 の 書生 より一層乱暴な方で	吾輩	を見るや否やいきなり 頸筋 を つ

統計を扱う処理の場合、表がそのまま各セルに収まります。

	1	2	3	4	5	6	7	8	9	10	11	12	
1	TOKEN 210988	TYPE 12053	TTR 0.0571	total mora 365929	Node合計 482								
2		語	合計	左計	右計	左5	左4	左3	左2	左1	Node	右1	右
3	1	吾輩	494	6	6	1	1	0	4	0	482	0	
4	2	は	314	64	250	10	7	8	12	27	0	188	
5	3	の	310	54	256	11	8	20	10	5	0	131	
6	4	に	173	59	114	10	11	13	11	14	0	19	
7	5	を	139	29	110	7	3	11	4	4	0	28	
8	6	が	127	52	75	10	7	6	13	16	0	42	
9	7	て	113	57	56	9	11	10	13	14	0	0	

.tsv

内部の各パーツがタブで区切られた形式で保存されます。これも MS-Excel が対応できる形式ですが、読み込む際は予め MS-Excel を起動しておき、ファイル指定をし、区切りをタブと指定して開く必要があります。

MS-Excel で開いた際の画面構成は基本的に.slk 形式と同じになります。

.txt

画面に表示されているのとほぼ全く同じ書式で、テキストファイル形式で保存されます。

```
;と使われている文脈(KWIC) 左5 - 右5 の範囲 ↓

    行番号                                                                左   Node          右
 1       1                                                                    吾輩          は 猫 で
 2       4                                                                    吾輩          は 猫 で
 3       5          も 薄暗い じめじめ した 所 で ニャーニャー 泣い て い た 事 だけ は 記憶 し て いる 。    吾輩          は ここ て
 4       7          。 はて な 何 で も 容子 が おかしい と 、 のそのそ 這い出し て 見る と 非常 に 痛い 。    吾輩          は 藁 の
 5       8                    ようやく の 思い で 笹原 を 這い出す と 向う に 大きな 池 が ある 。          吾輩          は 池 の
 6       8                 だ 。 縁 は 不思議 な もの で 、 もし この 竹垣 が 破れ て い なかっ た なら 、    吾輩          は ついに
 7       8                   一樹 の 蔭 と は よく 云っ た もの だ 。 この 垣根 の 穴 は 今日 に 至る まで    吾輩          が 隣家 の
 8       8               から 考える と その 時 は すでに 家 の 内 に 這入って おっ た の だ 。 ここ で    吾輩          は 彼 の
 9       8                 に 逢っ た の が おさん で ある 。 これ は 前 の 書生 より 一層 乱暴 な 方 で    吾輩          を 見る
10       8               い た 。 しかし ひもじい の と 寒い の に は どう し て も 我慢 が 出来 ん 。      吾輩          は 再び
11       8             の 隙 を 見 て 台所 へ 這い上っ た 。 する と 間 も なく また 投げ出さ れ た 。      吾輩          は 投げ出
12       8                を 偸ん で この 返報 を し て やっ て から 、 やっ と 胸 の 痞 が 下り た 。      吾輩          が 最後
13       8                に 、 この 家 の 主人 が 騒々しい 何 だ と いい ながら 出 て 来 た 。 下女 は    吾輩          を ぶら下
14       8                   上って 来 て 困り ます という 。 主人 は 鼻 の 下 の 黒い 毛 を 撚り ながら    吾輩          の 顔 を
15       8               しまっ た 。 主人 は あまり 口 を 聞か ぬ 人 と 見え た 。 下女 は 口惜し そう に    吾輩          を 台所
```

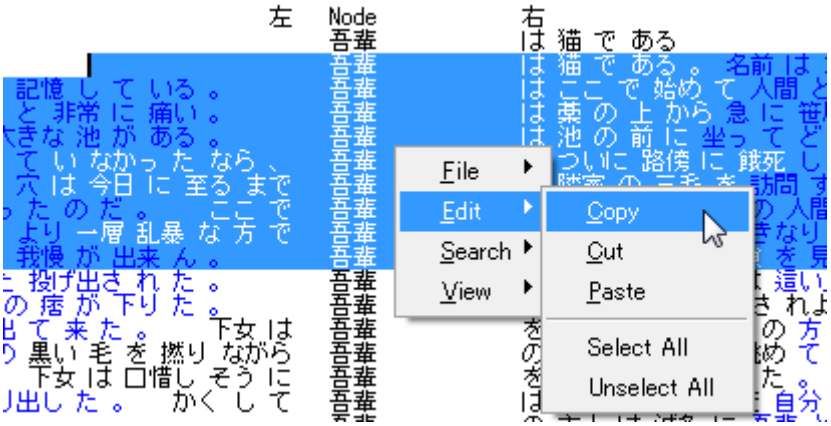
このファイルをコピーすればそのまま MS-Word などに貼り付けることもできます。

統計などの処理結果でも表の区切りが | や --- などで再現されそのまま記入されます。

```
 1|  TOKEN 210988    TYPE 12053    TTR 0.0571    total mora 365929    Node合計 482↓
 2|=============================================================================================================
 3|        |      語       |   合計   |  左計   |  右計   ||  左5  |  左4  |  左3  |  左2  |  左1  | Node  |  右
 4|=============================================================================================================
 5|       1|吾輩           |     494  |      6  |      6  ||     1  |     1  |     0  |     4  |     0  |   482  |
 6|       2|は             |     314  |     64  |    250  ||    10  |     7  |     8  |    12  |    27  |     0  |   1
 7|       3|の             |     310  |     54  |    256  ||    11  |     8  |    20  |    10  |     5  |     0  |   1
 8|       4|に             |     173  |     59  |    114  ||    10  |    11  |    13  |    11  |    14  |     0  |
 9|       5|を             |     139  |     29  |    110  ||     7  |     3  |    11  |     4  |     4  |     0  |
10|       6|が             |     127  |     52  |     75  ||    10  |     7  |     6  |    13  |    16  |     0  |
11|       7|て             |     113  |     57  |     56  ||     9  |    11  |    10  |    13  |    14  |     0  |
12|       8|、             |     102  |     86  |     16  ||     7  |     5  |     8  |     9  |    57  |     0  |
```

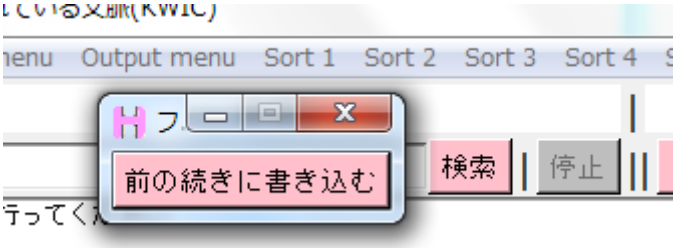
画面の直接コピー

また他の保存方法として画面の直接コピーが有ります。本ソフトでは処理結果が全てただの文字と記号で画面上に表現されます。そのため、画面上の好きな範囲を指定し、「Ctrl + c」や右クリックから「Edit」→「Copy」でその範囲のみをコピーできます。これをそのままテキストファイルや MS-Word 文書に貼り付けることで簡単に結果の保存や貼り付けができます。



Output menu

ウィンドウ上部のツールバーにある「Output menu」をクリックすると、ファイルへの書き込み方法を指定できます。



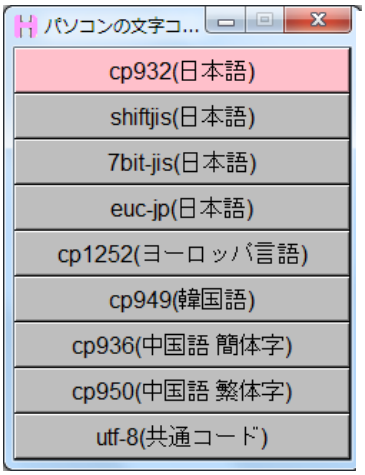
「前の続きに書き込む」がオンになっていると、結果をファイルへ書き込む際に、すでに他の結果を書き込んである場合、前回書き込んだ位置に続けて書き込みます。オフになっていると、前回の結果を消して新しい結果を上書きします。

ただし、これがオンでもオフでも書き込み先選択のウィンドウでの決定時には上書きの許可を求めてきますが、実際の書き込み動作は指定どおりに行われます。

オプション

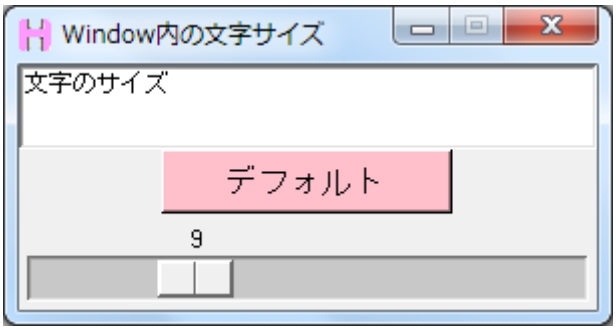
各処理のウィンドウの上部、一番左上にある「Option」ボタンで、ソフトのいくつかの設定の変更ができます。

パソコンの文字コード



パソコン自体の文字コードを変更できます。通常では日本語版 Windows 用になっていますがこれを変更することで別言語の Windows でも動くようになります。
分析するファイルの中身ではなく、読み込むファイル名自体のや途中のパスの文字コードへの対応です。
別言語での動作確認は取っていませんので、結果は不確かです。変更後も読み込みがうまくいかない場合は、分析するファイルや途中のパスの文字を半角英数のみにすることで読み込める可能性が高まります。

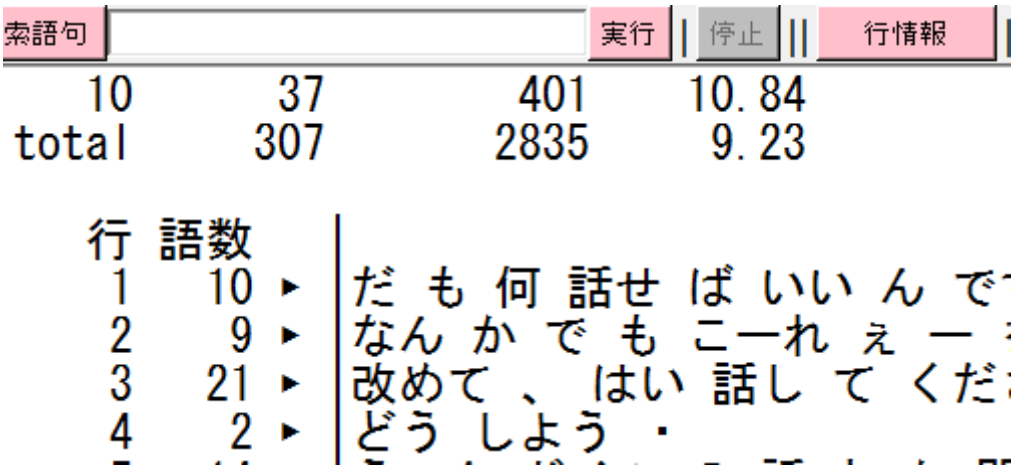
Window 内の文字サイズ



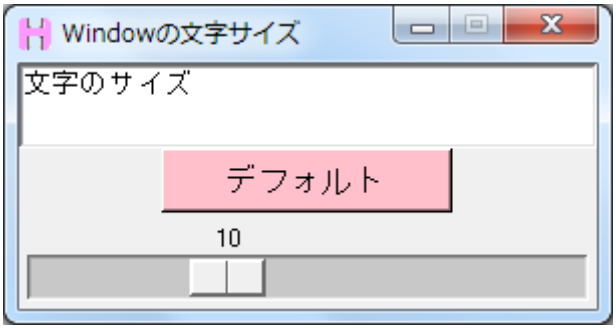
各処理の結果を表示するウィンドウ内の文字のサイズを変更します。結果の表示文字が小さすぎて見づらい場合などに使います。

スライダーを移動させてサイズを変更します。元に戻すときは「デフォルト」と押します。

サイズの変更ができたら文字サイズ変更のウィンドウを消します。
メインのウィンドウの「再描写」で、新しい文字サイズで画面内の表示が更新されます。



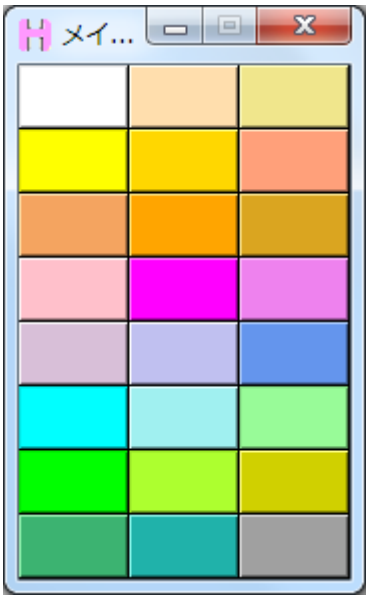
Window の文字サイズ



処理結果の表示ではなく、ウィンドウそのものの文字サイズを変更します。

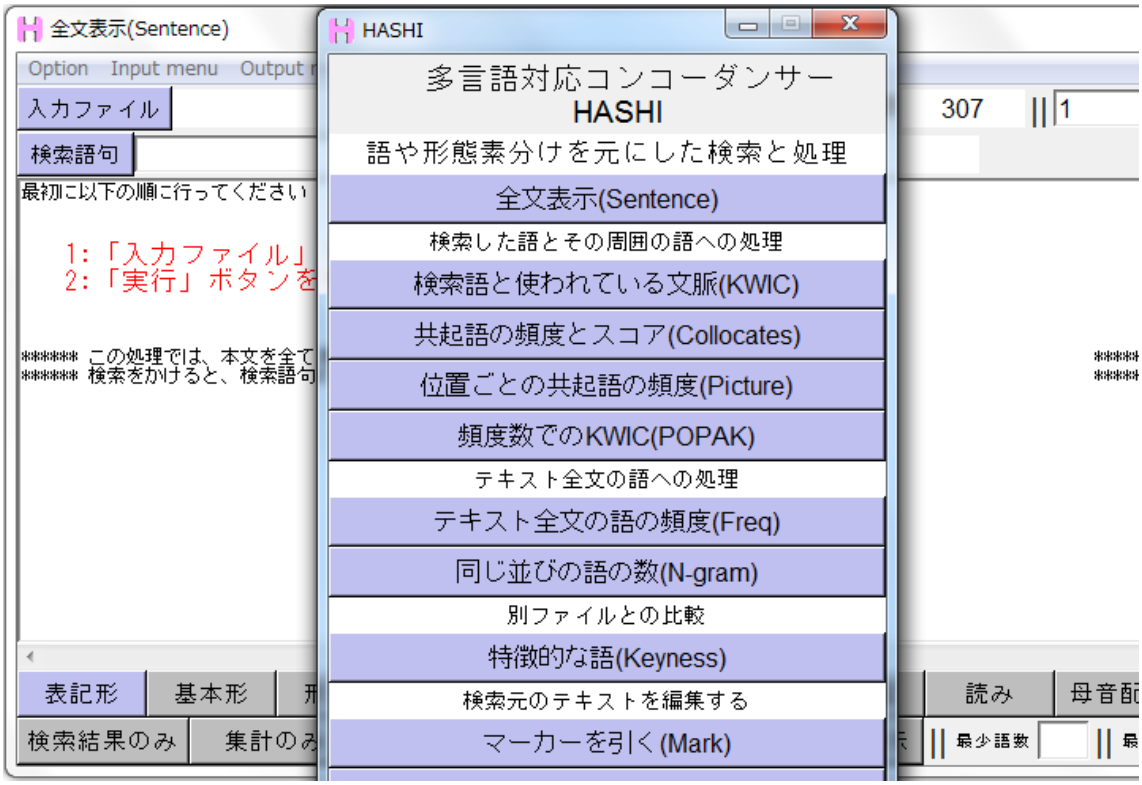
スライダーを移動させてサイズを変更します。元に戻すときは「デフォルト」と押します。
サイズの変更ができたら文字サイズ変更のウィンドウを消します。メインのウィンドウ自体を消して、再度その処理を選択すると新しい文字サイズでウィンドウが生成されます。

メインカラー



ソフト自体の色を変えることができます。厳密には、
有効になっているボタンの色を変更できます。
様々な色のボタンから１つ色を選択します。

すべてのボタンの色が変わります。



本ソフトでは２つ以上のファイルを同時に分析することはできませんが、それを行いたい場合は２つ同時に本ソフトを立ち上げることで疑似的に行うことができます。その際にどちらのデータの分析結果か区別するためにソフトの色を変えておきます。

各種統計の定義

TOKEN　　：記号を抜いた全ての語数
TYPE　　：記号を抜いた全ての基本形の種類数
TTR　　：TYPE 割る TOKEN
TOTAL MORA　：全語のモーラ数を足したもの
共起頻度：各語ごとに、検索語句の周囲の指定幅内に出現した頻度
個別頻度：各語ごとに、テキスト全文の中に出現した頻度
位置頻度：各語ごとに、検索語句の周囲の指定幅内の各位置別に出現した頻度

計算式

t-score (Collocates, POPAK)
（ 共起頻度 －(検索語数 × 個別頻度 ÷ TOKEN)） ÷ √共起頻度

MIscore (Collocates, POPAK)
log2 ((共起頻度 × TOKEN) ÷ (検索語数 × 個別頻度))

t-score (Picutre)
(位置頻度 －(検索語数 × 個別頻度 ÷ TOKEN)） ÷ √共起頻度

MI-score (Picutre)
log2 ((位置頻度 × TOKEN) ÷ (検索語数 × 個別頻度))

カイ二乗検定
(((期待値 － 個別頻度) の絶対値 - 0.5) ^2 ÷ 期待値) の総和
ただし、(期待値 － 個別頻度) が 0.5 未満の場合は、((期待値 － 個別頻度) の絶対値 - 0.5) を 0 とする。
期待値は (メインコーパスの個別頻度 + 参照コーパスの個別頻度) × (メイン or 参照) コーパスの TOKEN ÷ (メインコーパスの TOKEN + 参照コーパスの TOKEN)とする

対数尤度比検定
2 × ((個別頻度 × (log(個別頻度) － log(期待値)) の総和)
ただし、対数を取るため、各、個別頻度、期待値が 0 であった場合、(個別頻度 × (log(個別頻度) － log(期待値)自体を 0 とする。

基準統計量　　0.1%：10.83、1%：6.63、5%：3.84

ファイルの総語数と使用するデータの範囲

各処理で「行情報」ボタンによって、行単位のタグでの使用制限を行えます。その際に「ファイル」「属性」「行」で使用範囲を絞ることができますが、Collocates や Keyness などの統計値で使う TOKEN は、その使用範囲によって変動します。しかし、「ファイル」「属性」「行」など様々な条件が変動するたびに総語数である TOKEN が変動すると統計値が一定しない可能性があります。そこで、本ソフトでは、TOKEN は選択して使用される「ファイル」タグに一致するデータの範囲の総語数とします。ファイルタグを選択していなければ、データ全体から、指定していれば、そのファイルタグに一致する範囲の総語数とします。

Freq、N-gram、Keyness では行情報はファイルタグしか選択できません。これは上記の理由によるもので、選択したファイルタグの範囲での頻度などを扱います。

Keyness では、メインファイルと参照ファルの両方で別々にファイルタグの選択ができます。これで、同一のデータを選択したとしても別々のファイルタグを指定すれば、同じデータの中の別の箇所の比較ができます。

形態素解析ソフトの設置

茶筌用辞書の置き換え

初期状態では、同梱の茶筌の内部辞書は「ipadic-2.7.0」ですが、これを UniDic に置き換えることができます。UniDic は伝康晴・山田篤・小椋秀樹・小磯花絵・小木曽智信、各氏と国立国語研究所などの共同開発により作られている辞書です。UniDic の使用には登録が必要であるため、所定のサイトより登録し、ダウンロードしてください。
ダウンロード後、「HASHI」/「bin」/「chasen」の中にある「dic」というフォルダと置き換えてください。元々の「dic」には ipadic が入っていますので、安全のためにこれを「ipadic」などと名前を変更しておくと後から元に戻すことができます。HASHI では、常に、ここにある「dic」という名前のフォルダを茶筌用の辞書として使用します。

形態素解析ソフトの設置

日本語と韓国語以外の言語は「TreeTagger」によって形態素解析されます。これも所定のサイトより各自でダウンロードしてください。
設置場所は、「HASHI」/「bin」の中に「TreeTagger」というフォルダで置いてください。更にその中に「bin」と「lib」というフォルダを置き、「bin」の中に「tree-tagger.exe」を、「lib」の中に「english.par」、「english-abbreviations」などの「.par」「-abbreviations」形式のファイルを置いてください。中国語は「zh.par」「lcmc-bigrams2.dat」「lcmc-uni2.dat」の３つを「lib」に入れます。
TreeTagger で解析する言語の場合、パソコンに Perl がインストールされていないと tokenize が行われませんので、直にタグ付けが行われます。

韓国語は「MACH」を使います。これも所定のサイトからダウンロードし、「HASHI」/「bin」の中に「MACH」/「mach.exe」という配置で置きます。

これで、日本語以外の言語でも形態素解析の結果を利用することできるようになります。

HASHI 立ち上げ時に各形態素解析ソフトや該当言語の辞書などの設置が確認できれば、分析言語でその言語を指定したあと、扱える項目が増えます。

著作権

『HASHI』について

『HASHI』の著作権は、田中良が保持するものですが、広く各分野の研究に活用してもらえるよう、無料で配布するものです。ソースコードは Perl で書かれていますが、Active State 社による Perl Dev Kit の Perlapp コマンドによって exe 化をしているため、Perl の実行環境のない Windows 環境でも動作します。改編や再配布の許可に関しては現在検討中ですが、現状は不可とします。ただし、個人的な譲渡や授業においてクラス内で配布するなどは可能とします。Web 上などに設置して不特定多数に配布することは禁じます。

また、本ソフト「HASHI」を使用した各種の結果は一切保障されないものであり、それによって生じる、直接、間接的あらゆる不利益に関して田中良は一切の責任、保障を負わないものとします。

ただ、不具合や問題点などご報告いただければ、できる限り修正をしたいと考えています。今後、HASHI を各自の研究に利用していただくことは全くの自由ですが、使用結果を用いて論文や学会などで発表される場合、分析に本ソフトを使用した旨を記述いただきたいと思います。またその際にご一報いただけると大変嬉しく思います。

連絡先 gr021071@ed.ritsumei.ac.jp
agc59660@kwansei.ac.jp
田中 良

HASHI を使って作成したコーパスの公開について

HASHI ではオリジナルコーパスの作成が可能です。作成されたオリジナルデータに関しては作成者に著作権があるものですので、データの配布は自由です。その際データを公開するに当たって、作成されたデータの利用という目的で、コーパスを作成したバージョンの HASHI 本体も同梱して配布されたいという場合はご一報ください。その場合のみ再配布を許可するかご返答します。

HASHI 本体の同梱再配布の許可の如何に関わらず、コーパスデータの配布の際にはデータ作成に使用した HASHI のバージョンを明記してください。バージョンが変わるとデータの読み込みができない場合があります。

HASHI 形式でコーパスを作成され、一般配布される場合はご一報くだされば、HASHI の配布ホームページにてそのコーパスの紹介をさせていただきます。

『茶筌』について

『茶筌』および『ipadic』の著作権は奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座が保有しています。

以下、「形態素解析システム『茶筌』version 2.4.0 使用説明書」より抜粋

茶筌システムは，広く自然言語処理研究に資するため無償のソフトウェアとして開発されたものである．茶筌の著作権は，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室) が保持する．本ソフトウェアの使用，改変，再配布については，特に制限を課すことはしないが，再配布については，次の事項を条件とする．

_ 再配布されるソフトウェアに，著作権に関する本節の記述と使用説明書の表紙裏のページの著作権に関する但し書きを必ず含むこと．

なお，本ソフトウェアの著作権者である奈良先端科学技術大学院大学は，原形あるいは改変された形で配布された本ソフトウェアに関連して生じる一切の損失に対して保証の責を負わないこととする．また，上に述べた著作権は茶筌システム本体についてのものであり，ipadic をはじめとする他の辞書については，各辞書についての著作権条項があるためそちらを参照すること．

以下、「ipadic version 2.7.0 ユーザーズマニュアル」より抜粋

Copyright @copyright 2000, 2001, 2002, 2003 Nara Institute of Science and Technology. All Rights Reserved.

Use, reproduction, and distribution of this software is permitted. Any copy of this software, whether in its original form or modified, must include both the above copyright notice and the following paragraphs. Nara Institute of Science and Technology (NAIST), the copyright holders, disclaims all warranties with regard to this software, including all implied warranties of merchantability and fitness, in no event shall NAIST be liable for any special, indirect or consequential damages or any damages whatsoever resulting from loss of use, data or profits, whether in an action of contract, negligence or other tortuous action, arising out of or in connection with the use or performance of this software.

A large portion of the dictionary entries originate from ICOT Free Software. The following conditions for ICOT Free Software applies to the current dictionary as well.

Each User may also freely distribute the Program, whether in its original form or modified, to any third party or parties, PROVIDED that the provisions of Section 3 (”NO WARRANTY”) will ALWAYS appear on, or be attached to, the Program, which is distributed substantially in the same form as set out herein and that such intended distribution, if actually made, will neither violate or otherwise contravene any of the laws and regulations of the countries having jurisdiction over the User or the intended distribution itself.

NO WARRANTY

The program was produced on an experimental basis in the course of the research and development

conducted during the project and is provided to users as so produced on an experimental basis.
Accordingly, the program is provided without any warranty whatsoever, whether express, implied, statutory or otherwise.
The term ”warranty” used herein includes, but is not limited to, any warranty of the quality, performance, merchantability and fitness for a particular purpose of the program and the nonexistence of any infringement or violation of any right of any third party.
Each user of the program will agree and understand, and be deemed to have agreed and understood, that there is no warranty whatsoever for the program and, accordingly, the entire risk arising from or otherwise connected with the program is assumed by the user.
Therefore, neither ICOT, the copyright holder, or any other organization that participated in or was otherwise related to the development of the program and their respective officials, directors, officers and other employees shall be held liable for any and all damages, including, without limitation, general, special, incidental and consequential damages, arising out of or otherwise in connection with the use or inability to use the program or any product, material or result produced or otherwise obtained by using the program,regardless of whether they have been advised of, or otherwise had knowledge of, the possibility of such damages at any time during the project or thereafter. Each user will be deemed to have agreed to the foregoing by his or her commencement of use of the program. The term ”use” as used herein includes, but is not limited to, the use, modification, copying and distribution of the program and the production of secondary products from the program.
In the case where the program, whether in its original form or modified, was distributed or delivered to or received by a user from any person, organization or entity other than ICOT, unless it makes or grants independently of ICOT any specific warranty to the user in writing, such person, organization or entity, will also be exempted from and not be held liable to the user for any such damages as noted above as far as the program is concerned.

その他のソフトについて

『UniDic』、『TreeTagger』、『TreeTagger 用パラメータファイル等』、『MACH』に関してもそれぞれの著作者の権利物ですが、HASHI で同梱して配布していないものなのでここでは明記しません。各自で著作権および使用権についてご確認ください。