

全文検索システム『ひまわり』 講習会

「作文コーパスの活用」



本日の内容

- ▶ 第一部 既存の作文コーパスを『ひまわり』で活用する
(担当: 山口昌也・国語研)
 - ▶ 全文検索システム『ひまわり』の基本的な使い方
 - ▶ 『ひまわり』による『小中高大生による日本語絵描写ストーリーライティングコーパス』(JASWRIC)の利用方法
 - ▶ JASWRICを使った簡単な分析

『ひまわり』の紹介＋生テキストの活用

- ▶ 第二部 『日本語学習者作文コーパス』を『ひまわり』で検索する
(担当: 森篤嗣先生・京都外国語大)
 - ▶ 『日本語学習者作文コーパス』を『ひまわり』で検索する意義
 - ▶ 誤用の集計
 - ▶ 本来使用されるべき正用に基づく誤用の集計

作文コーパスの誤用分析

第一部

既存の作文コーパスを『ひまわり』で活用する

- 全文検索システム『ひまわり』の基本的な使い方
- 『ひまわり』による『小中高大生による日本語絵描写ストーリーライティングコーパス』(JASWRIC)の利用方法
- JASWRICを使った簡単な分析

『ひまわり』とは

▶ 言語研究用の全文検索システム

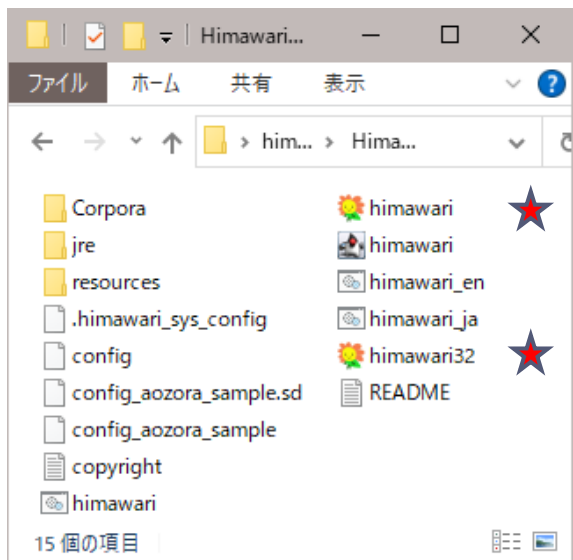
- ▶ 指定された文字列を網羅的に検索して, 前後文脈付きで結果を表示します (コンコーダンス)
- ▶ 『[太陽コーパス](#)』(20世紀初頭の総合雑誌『太陽』)用の検索システムとして構築しました
- ▶ 『[日本語日常会話コーパス](#)』, 『[日本語話し言葉コーパス](#)』, 『[分類語彙表](#)』, 『[青空文庫](#)』など多数の言語資料に対応

▶ 特徴

- ▶ XMLでタグづけされたコーパスに対する全文検索, 単語検索 (格納しているXMLデータは他のシステムでも利用可能)
- ▶ 検索結果, アノテーション結果の集計 (例: 総単語数)
- ▶ 資料の特徴に合わせた資料の閲覧 (例: 横書き / 縦書き表示)
- ▶ コーパス自作支援機能

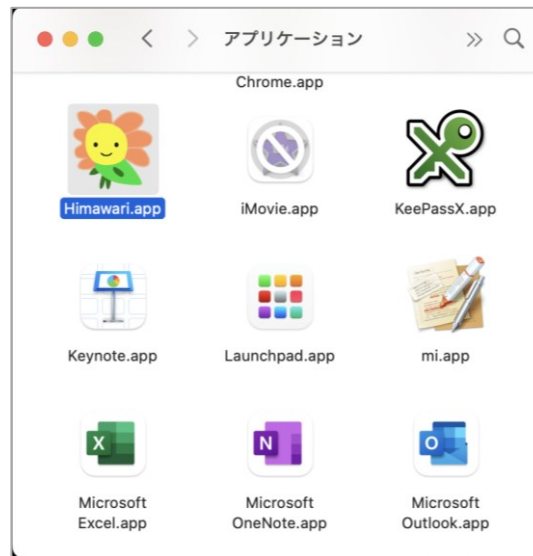
『ひまわり』の起動と『ひまわり』フォルダの確認

▶ Windowsの場合



- ▶ 通常は 🌻 のhimawariを使用
- ▶ 32bit版のWindowsの場合はWindows32を使用
- ▶ OSの設定によらず日本語を使いたい場合は, himawari_ja

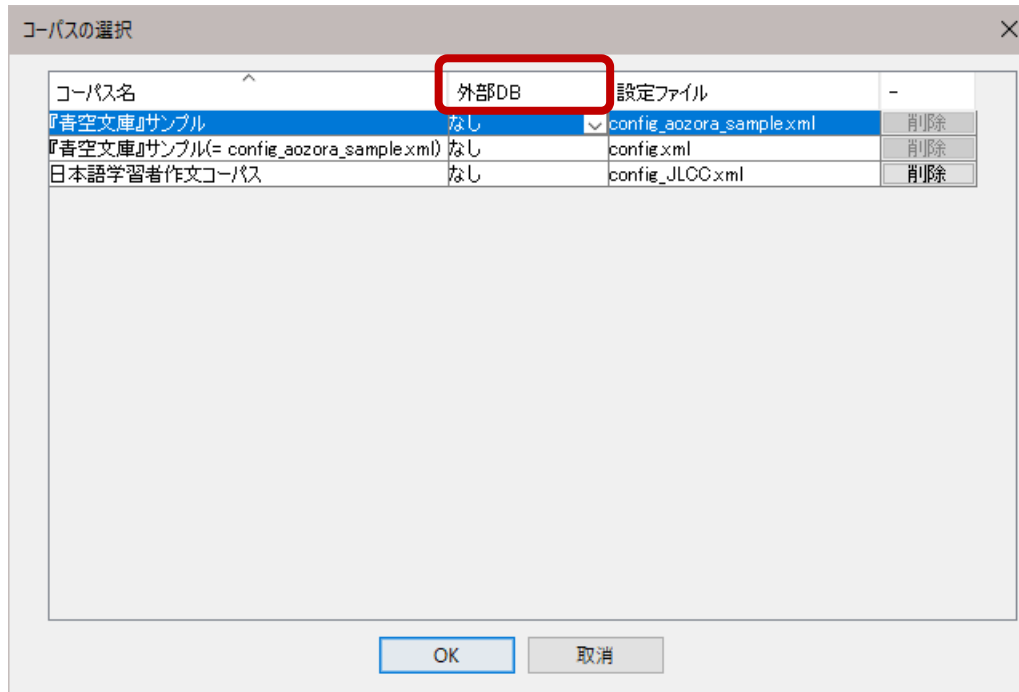
▶ macOSの場合



- ▶ Himawari.appを使用
- ▶ 右クリック→「パッケージの内容を表示」でフォルダを参照可能
- ▶ Contents → Resources フォルダにコーパスなどを格納

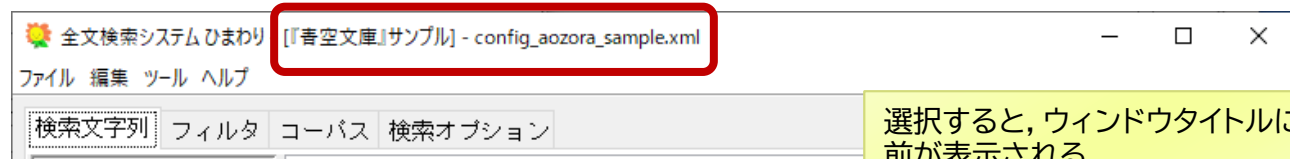
コーパスの選択

▶ [ファイル]⇒[コーパス選択]



▶ 「外部DB」

- ▶ 『青空文庫』サンプルなど、『ひまわり』のインポート機能で形態素解析を行った場合は、「あり」が選択可能
- ▶ 本日のJASWRICも「あり」で使用する
- ▶ 『日本語学習者作文コーパス』の形態素解析結果は原資料のデータを利用しているので「なし」



選択すると、ウィンドウタイトルに名前が表示される

検索する

「検索文字列」欄では
右クリックで履歴表示

全文検索システムひまわり - [aozora_sample] - config.xml

ファイル 編集 ツール ヘルプ

検索文字列 フィルタ コーパス 検索オプション

検索文字列

本文 前文脈 後文脈

検索 字体変換 クリア

no	前文脈	キー	後文脈	Path	タイトル	著者
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...	こころ	夏目漱石
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目漱石
3	弾くところですよ」	「これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目漱石
4	い話があるかい」	「これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目漱石
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aozora_s...	吾輩は猫...	夏目漱石
6	見当がつかない」	「これ	からいよいよ弾くとこ	/aozora_s...	吾輩は猫...	夏目漱石
7	めちゃんとお困ります。	これ	からがいよいよ佳境に	/aozora_s...	吾輩は猫...	夏目漱石
8	うと云うんです。さあ	これ	からがいよいよ失恋に	/aozora_s...	吾輩は猫...	夏目漱石
9	はすこぶる不慥だよ。	これ	からがいよいよ巧妙な	/aozora_s...	吾輩は猫...	夏目漱石
10	充分あらわれている。	これ	からが化物の記述だ。	/aozora_s...	吾輩は猫...	夏目漱石
11	か両君能く聞き給え、	これ	からが結論だぜ。一	/aozora_s...	吾輩は猫...	夏目漱石
12	と一と息ついた。「	これ	からが聞きどころです	/aozora_s...	吾輩は猫...	夏目漱石
13	んだ。「まだです。	これ	からが面白いところで	/aozora_s...	吾輩は猫...	夏目漱石
14	と信じました。同時に	これ	からさき彼を相手にす	/aozora_s...	こころ	夏目漱石

検索総数: 597

検索の実行

検索結果

途中経過の表示

検索総数

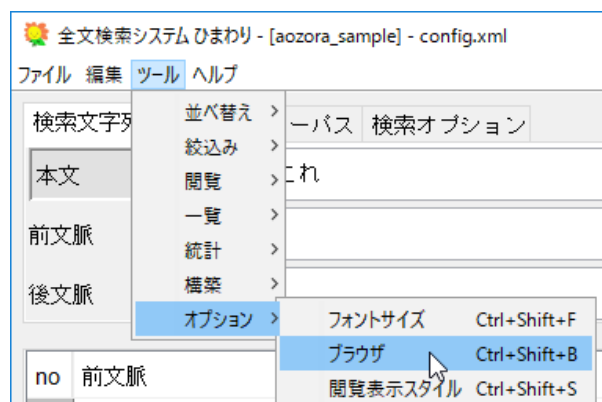
ブラウザでの閲覧

no	前文脈	キー	後文脈	Path
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aozora_s...
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aozora_s...
3	弾くところです」	これ	からいよいよヴァイオ	/aozora_s...
4	い話があるかい」	これ	からいよいよヴァイオ	/aozora_s...
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aozora_s...
6	見当がつかない」	これ	からいよいよ弾くところ	/aozora_s...
7	めちなお困ります。	これ	からがいよいよ佳境に	/aozora_s...

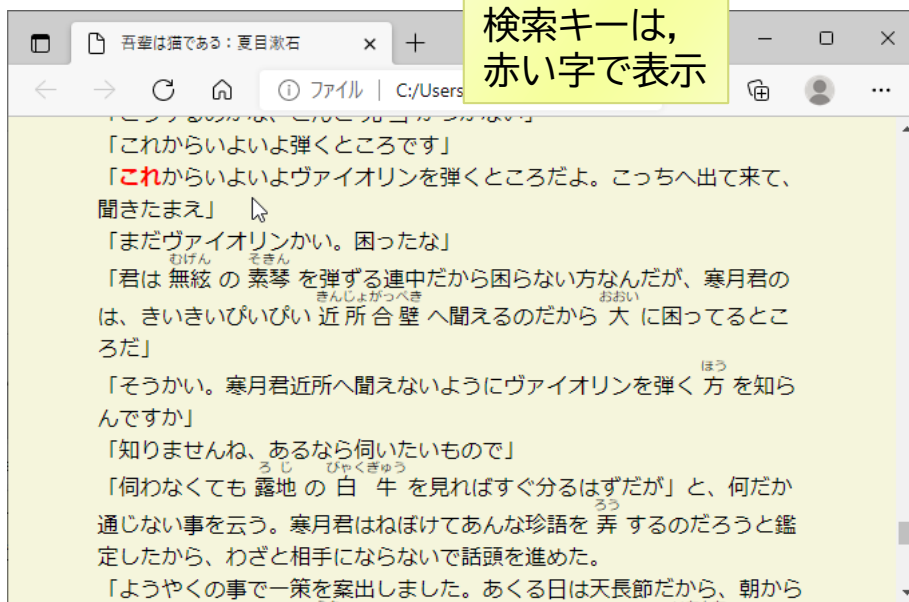
閲覧したい用例をダブルクリック



■ 閲覧用のブラウザの変更



[ツール]⇒[オプション]⇒[ブラウザ]



検索結果のソート

列名を左クリック

no	前文脈	キー	後文脈	Path	タイトル	著者
1	指して、しきりにかれ	これ	いいたがるのを、始め	/aозora_s...	こころ	夏目漱石
2	一軒屋を敲いて、これ	これ	かようかようしかじか	/aозora_s...	吾輩は猫...	夏目漱石
3	弾くところです」 「	これ	からいよいよヴァイオ	/aозora_s...	吾輩は猫...	夏目漱石
4	い話があるかい」 「	これ	からいよいよヴァイオ	/aозora_s...	吾輩は猫...	夏目漱石
5	、蛸壺峠へかかって、	これ	からいよいよ会津領へ	/aозora_s...	吾輩は猫...	夏目漱石
6	見当がつかない」 「	これ	からいよいよ弾くとこ	/aозora_s...	吾輩は猫...	夏目漱石
7	めちゃんお困ります。	これ	からがいよいよ佳境に	/aозora_s...	吾輩は猫...	夏目漱石
8	うと云うんです。さあ	これ	からがいよいよ失恋に	/aозora_s...	吾輩は猫...	夏目漱石
9	はすこぶる不慥だよ。	これ	からがいよいよ巧妙な	/aозora_s...	吾輩は猫...	夏目漱石
10	充分あらわれている。	これ	からが化物の記述だ。	/aозora_s...	吾輩は猫...	夏目漱石
11	か両君能く聞き給え、	これ	からが結論だぜ。一	/aозora_s...	吾輩は猫...	夏目漱石

▶ 昇順

列タイトルをクリック

▶ 降順

シフトキーを押しながら
列タイトルをクリック

▶ 複数列を考慮したい場合

▶ 優先順位の逆順でソートを実行

例:「タイトル」ごとに「後文脈」でソート
→ 「後文脈」「タイトル」の順

検索結果の絞り込み

▶ 検索時に指定

全文検索システム ひまわり - [aozora_sample] - config.xml
ファイル 編集 ツール ヘルプ

検索文字列 フィルタ コーパス 検索オプション

Path		で始まる
タイトル		で始まる
著者	夏目	で始まる

「著者」欄が「夏目」で始まる結果のみに絞り込まれる

▶ 検索後に絞り込み

no	前文脈	キー	後文脈	Path	タイトル	著者
1	...	これ	いいたがるのを、始め	/aozora_s...	こころ	夏目
	...	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目
	...	「これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	...	「これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目
	...	て、これ	からいよいよ会津領へ	/aozora_s...	吾輩は猫...	夏目漱石
	...	「これ	からいよいよ弾くところ	/aozora_s...	吾輩は猫...	夏目漱石
	...	これ	からがいよいよ佳境に	/aozora_s...	吾輩は猫...	夏目漱石

列名を右クリック

絞り込みたい値を選択
⇒右クリック
⇒フィルタでもOK

- [文字列指定]
- [置換]
- 夏目漱石
- 芥川龍之介

検索結果の頻度集計

1. 集計したい列を選択

no	前文脈	キー	後文脈	Path	タイトル	著者
1	これは本当の斬だと、	あの	うそつきの爺やが申し	/aozora_s...	吾輩は猫...	夏目漱石
2	ました、なに猫だから	あの	くらいで充分浄土へ行	/aozora_s...	吾輩は猫...	夏目漱石
3	が来ましたぜ。月並も	あの	くらいになるとなかな	/aozora_s...	吾輩は猫...	夏目漱石
4	まで随分ひきました	あの	くらい美しい音が出た	/aozora_s...	吾輩は猫...	夏目漱石
5	なら、立町は豚仙さ、	あの	くらい食い意地のきた	/aozora_s...	吾輩は猫...	夏目漱石
6	ますまい」と云う。「	あの	ちょっとくらい外出致	/aozora_s...	吾輩は猫...	夏目漱石
7	雪江さんが聞く。「	あの	ね。あとでおならは御	/aozora_s...	吾輩は猫...	夏目漱石
8	さんは謙遜した。「	あの	ね。坊たん、坊たん、	/aozora_s...	吾輩は猫...	夏目漱石

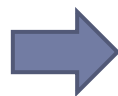
複数の列を
選択することも可

離れた列の選択

- WindowsはCtrlキー
- macOSはcommandキー

2. 右クリック⇒「統計」

1	タイトル	著者
ora_s...	吾輩は猫...	夏目漱石
ora_s...	吾輩は猫...	夏目漱石
ora_s...	吾輩は猫...	夏目漱石
ora_s...	吾輩は猫...	夏目漱石
ora_s...	吾輩は猫...	夏目漱石
ora_s...	吾輩は猫...	コピー
ora_s...	吾輩は猫...	コピー(列名含む)
ora_s...	吾輩は猫...	全選択
ora_s...	蜘蛛の糸	置換
ora_s...	吾輩は猫...	フィルタ
ora_s...	吾輩は猫...	統計
ora_s...	吾輩は猫...	夏目漱石
ora_s...	吾輩は猫...	夏目漱石



タイトル	著者	頻度
吾輩は猫...	夏目漱石	190
こころ	夏目漱石	41
蜘蛛の糸	芥川龍之介	1

総数(延べ): 232, 異なり: 3

『小中高大学生による日本語絵描写
ストーリーライティングコーパス』
(JASWRIC)のインポート

JASWRICの概要

▶ 収録データ

- ▶ 2種類の連続イラストに基づき, 日本の小・中・高・大学生, 合計700名が書いた作文を集めたもの
- ▶ 作文数は1400件, 総語数は約13万6千語

(<https://language.sakura.ne.jp/jaswric>)

▶ データ構成

▶ Edited Data

- ▶ 「形態素解析用に校閲を加えたデータ」

▶ Raw Data

- ▶ 手書き作文のスキャンデータ, 書き起こしテキスト

▶ JASWRIC_Participant Survey.xlsx

- ▶ 著者情報

▶ JASWRIC_Tagged.xlsx

- ▶ 形態素解析済みデータ(校閲済みのデータ)

本講習で利用
(書き起こしテキスト)

本家サイトの検索システム
で使用されている

『ひまわり』で利用する方法

- ▶ Raw Dataをインポート
 - ▶ 本講習
 - ▶ 『ひまわり』標準のテキストデータインポート機能を利用
- ▶ JASWRIC_Tagged.xlsxをインポート
 - ▶ 『ひまわり』のHPで[インポート方法を公開](#)
 - ▶ タブ区切りテキストを『ひまわり』用XMLデータに変換

利点

- 『ひまわり』の各種機能を利用できること
(例: 全文検索, 単語検索, 集計機能など)
- 検索システム中の元データを確認・検証できること
- 原資料に対するアノテーション・編集ができること
(例: 形態素解析システム JUMANでアノテーション)

インポートの実行

▶ インポート時の処理

- ▶ 学年ごとにサブコーパスとしてインポート
- ▶ テキスト自体には変更は加えない
- ▶ MeCab(UniDic)で形態素解析

- コーパス名: JASWRIC_RAW
- 「サブコーパスを作る」
- MeCab(UniDic)

Individual フォルダをドラッグ & ドロップ

動作の確認： 検索と作文全体の閲覧(1)

▶ 「行く」(出現形)で検索

全文検索システムひまわり - [JASWRIC_RAW] - config_JASWRIC_RAW.sd.xml

検索文字列 フィルタ

出現形

ルビ(rt)完全一致
ルビ(rt)部分一致
出現形
品詞
活用型
活用形
基本形
読み

検索文字列 フィルタ コーパス 検索オプション

出現形 行く

検索

前文脈 で終る

後文脈 で始まる

字体変換

クリア

no	前文脈	キー	後文脈	Path	タイトル	著者	品詞	品詞細分...	品詞細分...	品詞細分...	活用型	活用形	基本形
1	昼にピクニックに	行く	★定だったからです。	/Individual...	G06_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
2	ました。ピクニックに	行く	あいだり犬がサンドイ	/Individual...	G02_Pic ...		動詞	非自立可能			五段-力行	連体形-一般	行く
3	、どこへピクニックに	行く	か、地図を見て、さが	/Individual...	G03_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
4	二人で、どこへ	行く	かの確認をしていたと	/Individual...	G05_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
5	ぜなら、ピクニックに	行く	からです。マリとケン	/Individual...	G03_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
6	今日は、ピクニックに	行く	からです。リンゴや水	/Individual...	G03_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
7	しよに公園へデートに	行く	からです。作ったサン	/Individual...	G10_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
8	どのルートをとって	行く	かを2人で話しあうこ	/Individual...	G06_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
9	どこに	行く	かを二人で地図を見な	/Individual...	G04_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
10	て、地図を見てどこに	行く	かを確認していました	/Individual...	G08_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
11	らです。そしてどこに	行く	か地図を見ているとバ	/Individual...	G08_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
12	、どこにピクニックに	行く	か地図を見て二人で考	/Individual...	G07_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く
13	地図を見てどこに	行く	か相談中、目を離れた	/Individual...	G10_Pic ...		動詞	非自立可能			五段-力行	終止形-一般	行く

検索総数:174

G02_Pic_029 :

file:///C:/Users/masaya/AppData/Local/Temp/him: ☆

G02_Pic_029 :

マリとケンはずみみているあいだに、犬が入りました。ピクニックに**行く**あいだり犬がサンドイッチを、たべていました。ピクニックのところにつくと犬がでてきました。それでたべられてしまって、「せっかくつかったのに」とかなしくなりました。

動作の確認： 検索と作文全体の閲覧(2)

全文検索システムひまわり - [JASWRIC_RAW] - config_JASWRIC_RAW.sd.xml

ファイル 編集 ツール ヘルプ

検索文字列 フィルタ コーパス 検索オプション

出現形

前文脈 で終る

後文脈 で始まる

検索 字体変換 クリア

**Shift + ダブルクリック
⇒当該作文の形態素一覧**

no	前文脈	キー	後文脈	Path	タイトル	著者	品詞	品詞細分...	品詞細分...	品詞細分...	活用型	活用形	基本形
1	屋にピクニックに行く		★定だったからです。	/Individual...	G06_Pic...		動詞	非自立可能			五段-カ行	終止形-一般	行く
2	ました。ピクニックに行く		あいだり犬がサンドイ	/Individual...	G02_Pic...		動詞	非自立可能			五段-カ行	連体形-一般	行く
3	、どこへピクニックに行く		か、地図を見て、さが	/Individual...	G03_Pic...		動詞	非自立可能			五段-カ行	終止形-一般	行く
4	二人で、どこへ行く		かの確任をしていたと	/Individual...	G05_Pic...		動詞	非自立可能			五段-カ行	終止形-一般	行く

[20] 一覧

SER.NO.	TEXT	品詞	品詞細分類1	品詞細分類2	品詞細分類3	活用型	活用形	基本形	読み	発音	語種
0000010	あいだ	名詞	普通名詞	副詞可能				間	あいだ	アイダ	和
0000011	に	助詞	格助詞					に	に	ニ	和
0000012		補助記号	読点					、	、		記号
0000013	犬	名詞	普通名詞	一般				犬	犬	イヌ	和
0000014	が	助詞	格助詞					が	が	ガ	和
0000015	入り	動詞	一般			五段-ラ行	連用形-一般	入る	入り	ハイリ	和
0000016	まし	助動詞				助動詞-マス	連用形-一般	ます	まし	マシ	和
0000017	た	助動詞				助動詞-タ	終止形-一般	た	た	タ	和
0000018		補助記号	句点					。	。		記号
0000019	ピクニック	名詞	普通名詞	一般				ピクニック-p...	ピクニック	ピクニック	外
0000020	に	助詞	格助詞					に	に	ニ	和
0000021	行く	動詞	非自立可能			五段-カ行	連体形-一般	行く	行く	イク	和
0000022	あい	接頭辞						相	あい	アイ	和
0000023	だり	助詞	副助詞								
0000024	犬	名詞	普通名詞	一般							
0000025	が	助詞	格助詞								
0000026	サンドイッチ	名詞	普通名詞	一般							

検索総数

総数(延べ) : 69

**テキスト
進行方向**

■品詞での一覧表
①「品詞」列のどれかを選択
②右クリック→「統計」

語彙表の場合は、品詞・品詞細分類1~3・基本形・読みで集計

補足：単語での検索 (本講習会のJASWRICなど)

C) 先頭が「日」の単語

正規表現の「^」
(文字列の先頭)

検索文字列	フィルタ	コーパス	検索オプション
出現形	▼	^日	
前文脈			で終る ▼
後文脈			で始まる ▼

D) 末尾が「日」の単語

正規表現の「\$」
(文字列の末尾)

検索文字列	フィルタ	コーパス	検索オプション
出現形	▼	日\$	
前文脈			で終る ▼
後文脈			で始まる ▼

E) 単語「日」のみ

検索文字列	フィルタ	コーパス	検索オプション
出現形	▼	^日\$	
前文脈			で終る ▼
後文脈			で始まる ▼

F) 活用語の基本形

すべての語形を
一括して検索

検索文字列	フィルタ	コーパス	検索オプション
基本形	▼	歩く	
前文脈			で終る ▼
後文脈			で始まる ▼

補足：単語での検索 (「正規表現(前)」 「正規表現(後)」となる場合)

A) 「日」を含む単語

インターフェイスが
変わること
に注意

検索文字列	フィルタ	コーパス	検索オプション
出現形(短単位)		日	
正規表現(前)			正規表現
正規表現(後)			正規表現

B) 先頭が「日」の単語

正規表現の「^日」と同義
(先頭の文字が「日」)

検索文字列	フィルタ	コーパス	検索オプション
出現形(短単位)		日	
正規表現(前)		^	正規表現
正規表現(後)			正規表現

C) 末尾が「日」の単語

正規表現の「日\$」と同義
(末尾の文字が「日」)

検索文字列	フィルタ	コーパス	検索オプション
出現形(短単位)		日	
正規表現(前)			正規表現
正規表現(後)		\$	正規表現

D) 単語「日」のみ

正規表現の「^日\$」と同義

検索文字列	フィルタ	コーパス	検索オプション
出現形(短単位)		日	
正規表現(前)		^	正規表現
正規表現(後)		\$	正規表現

収録内容の確認

□ 作文一覧

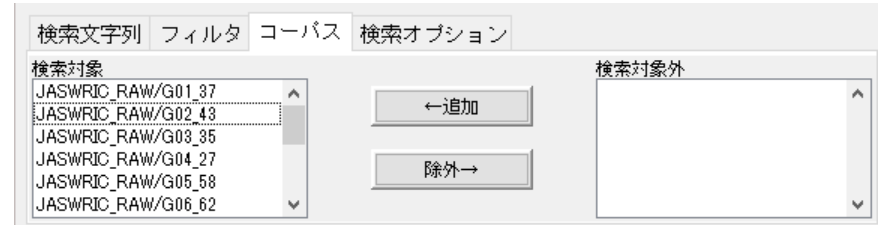
([ツール]→[一覧]→[タイトル・著者])



タイトル ^	サブタイトル	著者	Path
G01_Key_001			/Individual/G01_...
G01_Key_002			/Individual/G01_...
G01_Key_003			/Individual/G01_...
G01_Key_004			/Individual/G01_...
G01_Key_005			/Individual/G01_...
G01_Key_006			/Individual/G01_...
G01_Key_007			/Individual/G01_...
G01_Key_008			/Individual/G01_...
G01_Key_009			/Individual/G01_...
G01_Key_010			/Individual/G01_...
G01_Key_011			/Individual/G01_...
G01_Key_012			/Individual/G01_...

総数(延べ): 1400

□ サブコーパス



□ XMLデータ

- インポート結果のXMLファイルの場所
⇒Corpora/JASWRIC/サブコーパス名/copus.xml
- 個々の作文をXMLに変換した結果を連結したもの
- テキストエディタで閲覧することが可能
- 形態素解析結果は、外部データベースに保持し、XMLでは記述されない

収録内容の確認

□ 単語一覧(s) ([ツール]→[一覧]→[ユーザ入力])

要素一覧作成 (ユーザ入力)

第1層カテゴリー: morph (一部選択)

第2層カテゴリー: [未選択] (選択なし)

第3層カテゴリー: [未選択] (選択なし)

頻度 長さ 内容 文脈 0

OK 取消

チェックした属性を組として、
頻度が集計される

要素一覧作成 (ユーザ入力)

{TEXT}

品詞

品詞細分類1

品詞細分類2

品詞細分類3

基本形

活用型

活用形

発音

語種

読み

すべて選択

OK 取消

morph/@基本形	morph/@品詞	morph/@品...	morph/@品...	morph/@品...	morph/@活用型	頻度
て	助詞	接続助詞				8387
た	助動詞				助動詞-タ	8329
、	補助記号	読点				7081
ます	助動詞				助動詞-マス	5876
。	補助記号	句点				5699
に	助詞	格助詞				5530
を	助詞	格助詞				5462
は	助詞	係助詞				4105
が	助詞	格助詞				3960
居る	動詞	非自立可能			上一段-ア行	3277
と	助詞	格助詞				3202
為る	動詞	非自立可能			サ行変格	2861
の	助詞	格助詞				2778
て						

総数(延べ): 137463, 異なり: 2488

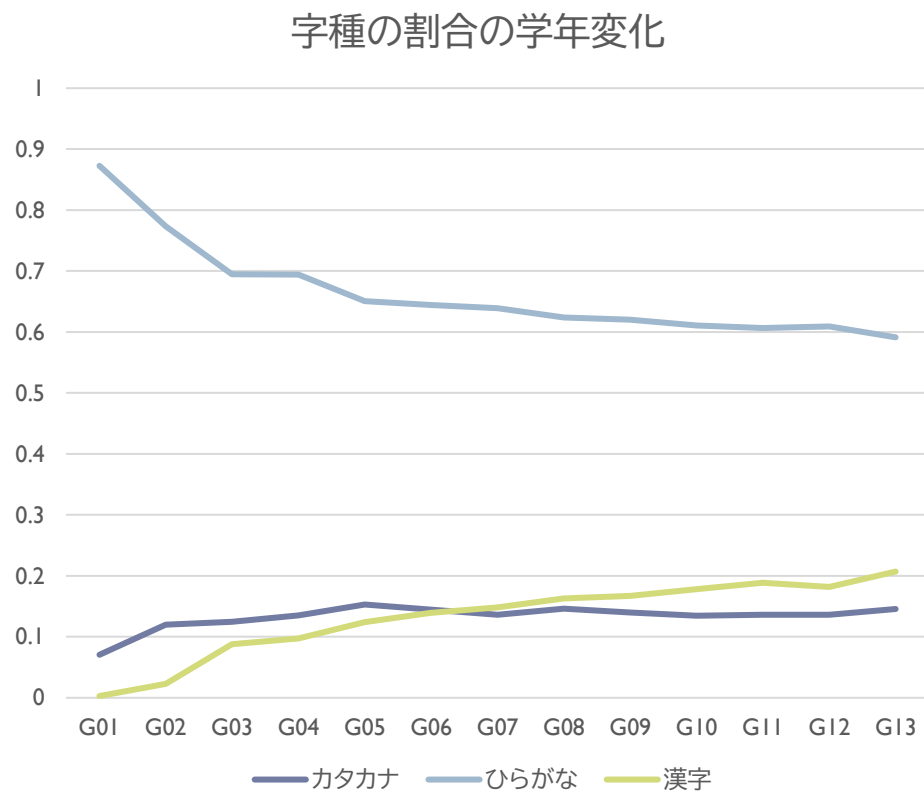
- 集計する属性にチェックを入れる
- 「活用形」ごとには集計しないので、チェックはしない
- {TEXT}は、出現形

JASWRICを用いた簡単な分析

— 学年ごとに字種の割合を調べてみる —

目標

▶ 次のようなグラフを作ります



手順

1. 字種Aを指定して全文検索(漢字, ひらがな, カタカナ)
2. 学年別(G1~G13)に頻度 f_{Gi} を集計
3. 学年別に総文字数 f_s を集計
4. 学年別に字種Aの割合を計算(f_{Gi} / f_s)
5. 1~4をすべての字種に対して行う

手順1 字種を指定して全文検索

■ 正規表現で字種を指定して全文検索(各字種ごとに個別に検索)

全文検索システムひまわり - [JASWRIC_RAW] - config_JASWRIC_RAW.sd.xml

検索文字列 フィルタ コーパス 検索オプション

本文(正規表現) \p{InKatakana}

前文脈 後文脈

no	前文脈	キー	後文脈	Path	タイトル	著者	品詞
1	を乗っ取ったんだゴル	ア	」男が言った、奴は女	/Individual...	G10_Key...		名詞
2	ミソファミレドレミフ	ア	ソファミレド うわ	/Individual...	G08_Pic ...		名詞
3	かみる ドドレミミフ	ア	ファミレドソファミソ	/Individual...	G08_Pic ...		補助記号
4	ミファファミレドソフ	ア	ミソファミレソファミ	/Individual...	G08_Pic ...		名詞
5	ファミソファミレソフ	ア	ミソファミレドレミフ	/Individual...	G08_Pic ...		名詞
6	ファミレドソファミソフ	ア	ミレソファミソファミ	/Individual...	G08_Pic ...		名詞
7	ミレドレミファソラフ	ア	ミレド うわ、犬入っ	/Individual...	G08_Pic ...		名詞
8	る ドドレミミファフ	ア	ミレドソファミソファミ	/Individual...	G08_Pic ...		名詞
9	ファミレソファミソフ	ア	ミレドレミファソラフ	/Individual...	G08_Pic ...		名詞
10	した、ケンハレディフ	ア	ーストだとバスケット	/Individual...	G10_Pic ...		名詞
11	、2階の開いているドア	ア	から入ろうとしました	/Individual...	G08_Key...		名詞
12	カギをわすれたのでドア	ア	があきません マリに	/Individual...	G02_Key...		名詞
13	をかけたもなかなかドア	ア	が開かなかったため、	/Individual...	G10_Key...		名詞

検索総数: 30169

¥p{InHiragana} ... ひらがな
¥p{InKatakana} ... カタカナ
¥p{InCJKUnifiedIdeographs} ... 漢字

日本語キーボードのmacの場合、「¥」は optionキー+「¥」を使用
なお、『ひまわり』の画面表示ではWindows, macとも逆スラッシュ「\」になる

手順2 学年ごとに頻度を計測

全文検索システムひまわり - [JASWRIC_RAW] - config_JASWRIC_RAW.sd.xml

検索文字列: 本文(正規表現) \p{InKatakana}

検索

①タイトル列で集計

no	前文脈	キー	後文脈	Path	タイトル	著者	品
1	を乗っ取ったんだゴル	ア	」男が言った、奴は女	/Individual...	G10_Key...		名語
2	ミンファミレドレミフ	ア	ソラファミレドうわ	/Individual...	G08_Pic...		名語
3	かみるドドレミミフ	ア	ファミレドソファミソ	/Individual...	G08_Pic...		名語
4	ミファファミレドソフ	ア	ミンファミレソファミ	/Individual...	G08_Pic...		名語
5	ファミソファミレソフ	ア	ミンファミレドレミフ	/Individual...	G08_Pic...		名語
6	ファミレドソファミソフ	ア	ミレソファミソファミ	/Individual...	G08_Pic...		名語
7	ミレドレミファソラフ	ア	ミレドうわ、犬入っ	/Individual...	G08_Pic...		名語
8	るドドレミミファフ	ア	ミレドソファミソファミ	/Individual...	G08_Pic...		名語
9	ファミレソファミソファミ	ア	ミレドレミファソラフ	/Individual...	G08_Pic...		名語
10	した、ケンはレディフ	ア	ーストだとバスケット	/Individual...	G10_Pic...		名語
11	、2階の開いているド	ア	から入ろうとしました	/Individual...	G08_Pic...		名語
12	カギをわすれたのでド	ア	があきません マリに	/Individual...	G02_Key...		名語
13	をかけてもなかなかド	ア	が開かなかったため	/Individual...	G10_Key...		名語

検索総数: 30169

②学年情報(G01~G13)以外を置換で削除

タイトル	頻度
G03_Pic_035	87
G09_Pic_042	84
G05_Pic_003	78
G05_Pic_044	75
G09_Pic_040	73
G05_Pic_056	70
G05_Pic_022	70
G13_Pic_015	69
G03_Pic_017	66
G06_Pic_057	66
G06_Pic_057	66
G04_Pic_057	66
G12_Pic_057	66
G13_Pic_057	66
G08_Pic_057	66
G10_Pic_057	66
G06_Pic_057	66

総数(延べ): 30169, 異なり...

③再集計

タイトル	頻度
G01	573
G02	1133
G03	1477
G04	1276
G05	2858
G06	2684
G07	1100
G08	4746
G09	2001
G10	4327
G11	3991
G12	1432
G13	2571

総数(延べ): 30169, 異なり...

確認

現在の「頻度」欄の値を考慮して、新しい頻度を計測しますか？

はい(Y) いいえ(N)

正規表現 「_*」
 ... 「_」+長さ0以上の文字列
 ⇒ 「_」以降を削除

置換 (正規表現)

検索する文字列: *_

置換後の文字列:

OK Cancel

補足： 結果のエクスポート

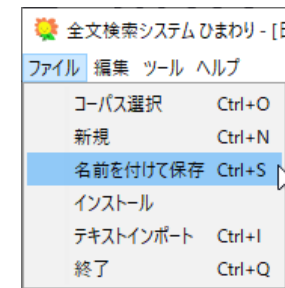
▶ 方法1： クリップボードを使用する方法

1. 結果を選択
 - ▶ 全選択したい場合は, Ctrl キー + A
2. 選択範囲をコピー
 - ▶ 通常のコピー:
Ctrl キー + C
 - ▶ 列名を含めたコピー:
Ctrl キー + Shift キー + C
3. Excel などにペースト



▶ 方法2： [ファイル] ⇒ [名前を付けて保存]

- ▶ タブ区切りのテキストとして保存



Excelは、日付や数値などを自動変換するので、
必要に応じて、コピー先のセルの書式を「文字列」にしておく

補足：「抽出」オプション

- ▶ 全数の検索が不可能な場合などに利用

The screenshot shows a dialog box titled '抽出' (Extract) with several tabs: '検索文字列' (Search string), 'フィルタ' (Filter), 'コーパス' (Corpus), and '検索オプション' (Search options). The '抽出' tab is active. Below the tabs, there are three radio button options: '全数' (All) which is selected, 'ランダム' (Random), and '頻度計測のみ' (Frequency measurement only). To the right of these options are three input fields: '抽出数上限' (Maximum number of extractions) with a text box, 'サンプル数' (Sample size) with a text box, and '表示方法' (Display method) with two radio buttons: '一覧' (List) which is selected, and '総計' (Total).

- ▶ 「頻度計測のみ(一覧)」
 - ▶ 指定した列(の組み合わせ)で頻度を計測
 - ▶ 手順
 1. 「全数」などで検索総数の少ない文字列を検索
 2. 検索結果(どの行でもよい)で、集計する列を選択
 3. 頻度計測のみ(一覧)を選択し、希望の条件で検索を実行(フィルタも使用可)

手順3 学年別に総文字数を集計

3. 学年ごとの総頻度

⇒ 正規表現「.」ですべての文字を検索し、学年ごとに集計
(指定する正規表現が異なるだけであとは同じ処理)

検索文字列 フィルタ コーパス 検索オプション

本文(正規表現) .

検索 字体変換 クリア

no	前文脈	キー	後文脈	Path	タイトル	著者	品詞
1	ないんだあああああ!		!!!とケンは無事に	/Individual...	G13_Key...		補助記!
2	いんだあああああ!		!!とケンは無事に抵	/Individual...	G13_Key...		補助記!
3	…びっくりぎょうてん!		!!犬がせっかくつく	/Individual...	G04_Pic...		
4	由を説明するしかない!		!「あっ!」と言っ	/Individual...	G06_Key...		
5	た。「マリ!」「ケン!		!「今のことからよ	/Individual...	G03_Key...		
6	んに、マリが「お帰り!		!と大きな声を出し	/Individual...	G06_Key...		
7	てきました。「まさか!		!と思いつながらパス	/Individual...	G05_Pic...		
8	た。「よし、食べよう!		!と言い、バスケッ	/Individual...	G06_Pic...		
9	付きました。「そうだ!		! ケンは物置きから	/Individual...	G05_Key...		
10	ら、「わあっ!」ポチ!		!お弁当全部食べちゃ	/Individual...	G06_Pic...		
11	た…。ごはんにしよう!		!と、思ったら、「わ	/Individual...	G06_Pic...		
12	たんだ。これで、よし!		!とそこに、「こんな	/Individual...	G06_Key...		補助記!
13	んだああああああ!!		!とケンは無事に抵抗	/Individual...	G13_Key...		補助記!



タイトル	頻度
G03_Key_015	756
G07_Key_001	573
G03_Pic_035	453
G03_Key_035	409
G10_Pic_079	374
G05_Pic_003	370
G04_Key_017	361
G10_Key_070	355
G10_Key_079	353
G09_Key_042	349
G10_Key_026	339
G13_Pic_015	338
G13_Key_038	331

総数(延べ): 220856, 異なり...



タイトル	頻度
G01	8162
G02	9477
G03	11850
G04	9442
G05	18713
G06	18566
G07	8085
G08	32461
G09	14333
G10	32226
G11	29335
G12	10535
G13	17671

総数(延べ): 220856, 異...

正規表現「.」(半角のピリオド)
… 任意の1文字

①タイトル列で集計

②学年情報(G01~G13)
以外を置換で削除
③再集計

手順4 字種の割合を計算

粗頻度

	カタカナ	ひらがな	漢字	総文字数
G01	573	7121	22	8162
G02	1133	7329	217	9477
G03	1477	8233	1038	11850
G04	1276	6554	917	9442
G05	2858	12170	2317	18713
G06	2684	11962	2586	18566
G07	1100	5166	1197	8085
G08	4746	20244	5291	32461
G09	2001	8891	2395	14333
G10	4327	19675	5737	32226
G11	3991	17796	5524	29335
G12	1432	6417	1913	10535
G13	2571	10447	3653	17671

② 選択して、「形式を選択して貼り付け」(除算)

総文字数に対する比率

	カタカナ	ひらがな	漢字	総文字数
G01	0.070203	0.872458	0.002695	8162
G02	0.119553	0.773346	0.022898	9477
G03	0.124641	0.694768	0.087595	11850
G04	0.135141	0.694133	0.097119	9442
G05	0.152728	0.65035	0.123818	18713
G06	0.144565	0.644296	0.139287	18566
G07	0.136054	0.638961	0.148052	8085
G08	0.146206	0.623641	0.162996	32461
G09	0.139608	0.620317	0.167097	14333
G10	0.13427	0.610532	0.178024	32226
G11	0.136049	0.606647	0.188307	29335
G12	0.135928	0.609112	0.181585	10535
G13	0.145493	0.591195	0.206723	17671

① 選択して、コピー



おわりに

- ▶ 既存の作文コーパスを『ひまわり』で活用する方法を紹介
 - ▶ 全文検索システム『ひまわり』の基本的な使い方
 - ▶ 『ひまわり』によるJASWRICの利用方法
 - ▶ JASWRICを使った簡単な分析

- ▶ さらに詳しく知るには
 - ▶ [『ひまわり』ホームページ](#)
 - ▶ [チュートリアルビデオ](#)
 - ▶ [利用者マニュアル](#)
 - ▶ [研究発表](#)
 - ▶ [過去の講習会](#)

補足：申込み時のコメントを受けて

▶ 発話データの集計

▶ 『ひまわり』で利用可能なパッケージ

- ▶ [名大会話コーパス](#), [昭和話し言葉コーパス](#), [国会会議録](#)
- ▶ [日本語日常会話コーパス](#), [日本語話し言葉コーパス](#)
- ▶ 自作したい場合は, [第13回コーパス利用講習会](#)の資料を参照

▶ タグの集計機能の利用

- ▶ [ツール] → [一覧] → [ユーザ入力] (本資料p.21)
- ▶ [第14回コーパス利用講習会](#)の資料 (p.20~)

▶ 文脈情報を利用した絞り込み

▶ 前後文脈欄の利用

- 第14回コーパス利用講習会の資料p.15

▶ 前後の単語の利用

- 検索結果の「基本形-2」「基本形-1」「基本形1」「基本形2」列で絞り込み