

国立国語研究所学術情報リポジトリ

既存の作文コーパスを『ひまわり』で活用する

メタデータ	言語: Japanese 出版者: 国立国語研究所 公開日: 2023-07-28 キーワード (Ja): キーワード (En): 作成者: 山口, 昌也 メールアドレス: 所属:
URL	https://doi.org/10.15084/0002000023



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

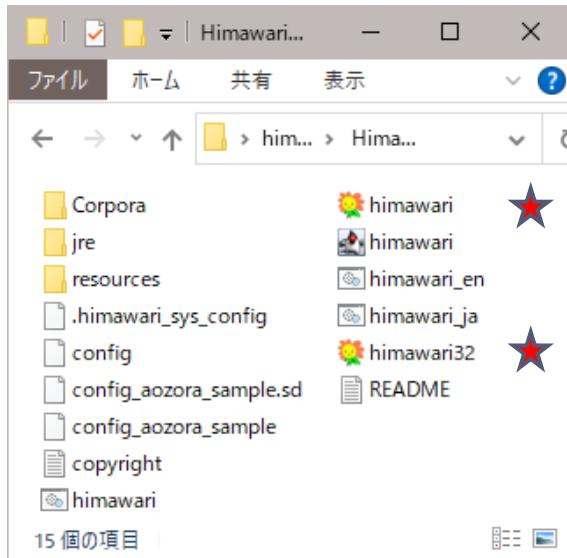
- ・ 全文検索システム『ひまわり』の基本的な使い方
- ・ 『ひまわり』による『小中高大生による日本語絵描写ストーリーライティングコーパス』(JASWRIC)の利用方法
- ・ JASWRICを使った簡単な分析

『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（コンコーダンサ）
 - ▶ 『[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

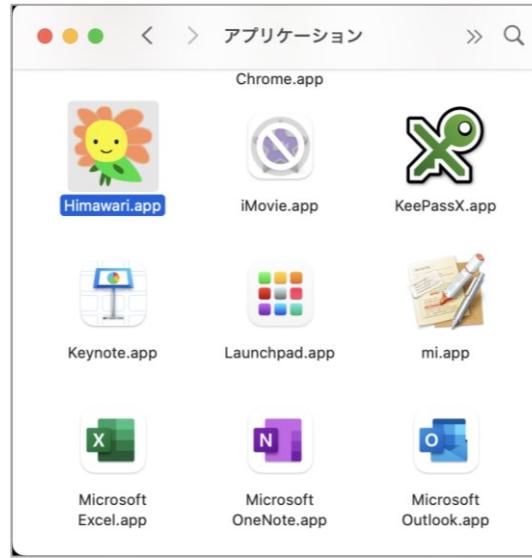
『ひまわり』の起動と『ひまわり』フォルダの確認

▶ 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

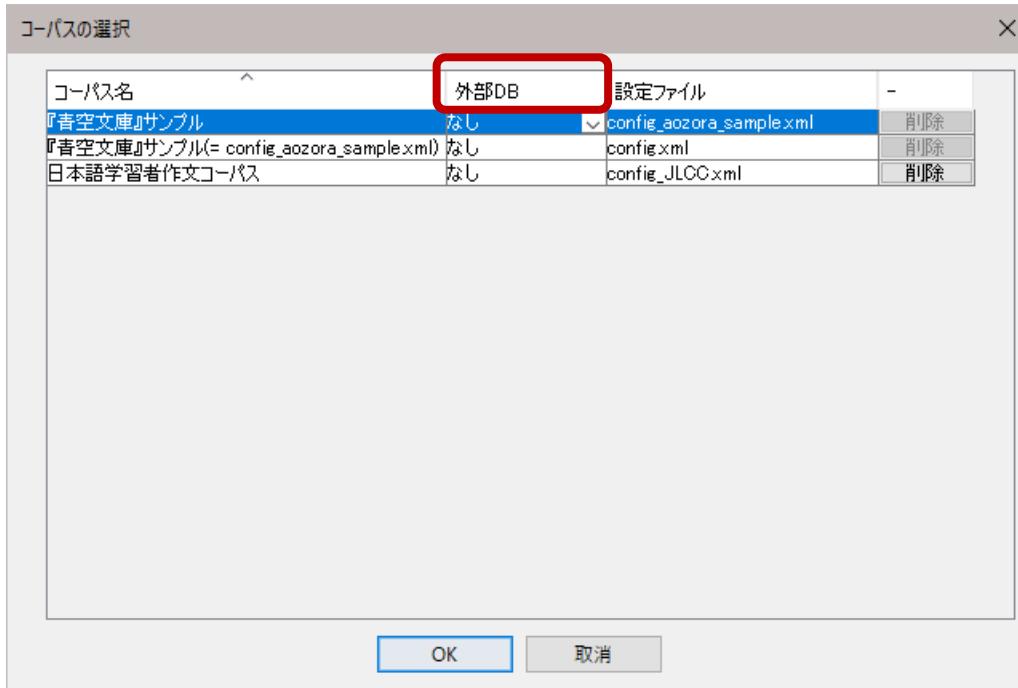
▶ macOSの場合



- ▶ 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るので「なし」



検索する

```
C:\WINDOWS\system32>  
C:\Users\masaya\Desktop\ProgramData\Oracle\em32\Wbem;C:\WINDOWS\System\Microsoft\Windows  
C:\Users\masaya\Desktop\t.jar  
osname: windows 10  
osname: windows 10  
osname: windows 10  
info:  
info: _sys _pred  
info: _sys _key  
info: _sys _full  
info: 記事 path  
info: 記事 タイ  
info: 記事 著者  
corpusname aozora_sample  
info(available memory  
history added: これ  
doSearch:start  
corpusname aozora_sample  
open_eix  
eix:記事  
n:597  
info(available memory  
time[milisec]: 65  
time[milisec]: 65
```

「検索文字列」欄では右クリックで履歴表示

検索の実行

検索結果

途中経過の表示

検索総数

検索文字列

本文

これ

検索

字体変換

クリア

検索総数: 597

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」「これ	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	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	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	これ	からいよいよ弾くとこ	/aozora_s...	吾輩は猫...	夏目漱石
7	めちゃなお困ります。これ	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	はすこぶる不憧だよ。これ	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	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	これ	からが結論だぜ。一	/aozora_s...	吾輩は猫...	夏目漱石
12	とーと息ついた。「これ	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	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	これ	からさき彼を相手にす	/aozora_s...	こころ	夏目漱石

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	これ	からいよいよヴァイオ	/aozora_s...
4	い話があるかい」「これ	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	これ	からいよいよ会津領へ	/aozora_s...
6	見当がつかない」「これ	これ	からいよいよ弾くとこ	/aozora_s...
7	めちゃなお困ります。これ	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



検索キーは、赤い字で表示

吾輩は猫である：夏目漱石

「これからいよいよ弾くところです」

「**これ**からいよいよヴァイオリンを弾くところだよ。こっちへ出て来て、聞きたまえ」

「まだヴァイオリンかい。困ったな」

「君は無絃の素琴を弾ずる連中だから困らない方なんだが、寒月君のは、きいきいぴいぴい近所合壁へ聞えるのだから 大に困ってるところだ」

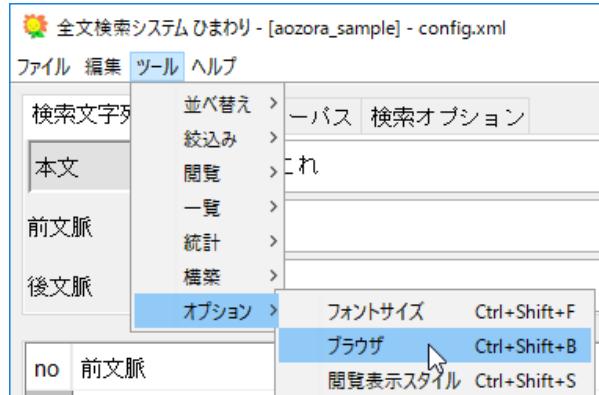
「そうかい。寒月君近所へ聞えないようにヴァイオリンを弾く方を知らんですか」

「知りませんね、あるなら伺いたいもので」

「伺わなくとも 露地の白牛を見ればすぐ分るはずだが」と、何だか通じない事を云う。寒月君はねばけてあんな珍語を弄するのだろうと鑑定したから、わざと相手にならないで話題を進めた。

「ようやくの事で一策を案出しました。あくる日は天長節だから、朝から

■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	/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」「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	めちゃなお困ります。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	はすこぶる不憮だよ。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。—	/aozora_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

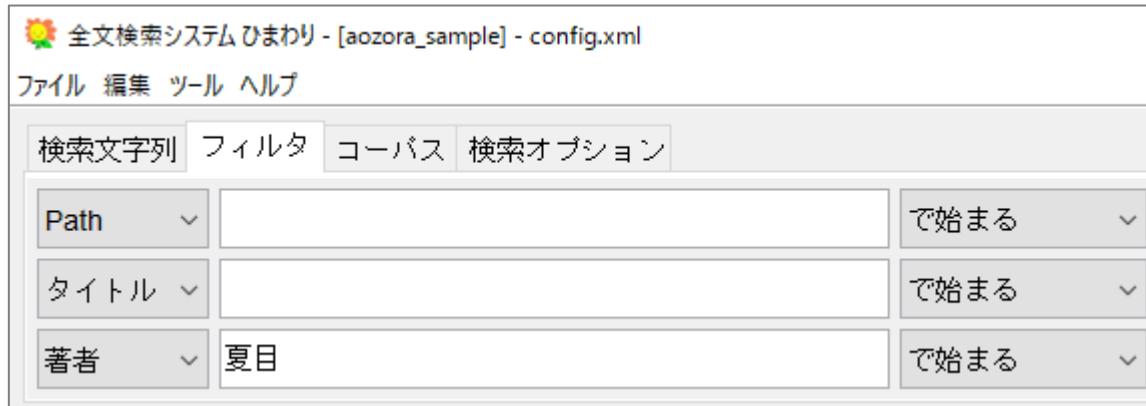
▶ 複数列を考慮したい場合

▶ 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



▶ 検索後に絞り込み

列名を右クリック

no	前文脈	キー	後文脈	Path	タイトル	著者	操作
1	セヒアトキイリーカレ	これ	いいたがるのを、始め	/aozora_s...	こころ	夏目	[文字列指定]
	これ	これ	かようかようしかじか	/aozora_s...	吾輩は猫...	夏目	[置換]
	「これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目	夏目漱石	
	「これ	からいよいよヴァイオ	/aozora_s...	吾輩は猫...	夏目	芥川龍之介	
	「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	/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		

絞り込みたい値を選択

⇒右クリック

⇒フィルタでもOK

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	これは本当の嘶だと、	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	くらいで充分淨土へ行	/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か	/aozora_s...	吾輩は猫...	夏目漱石
4	まで随分ひきましたが	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」



『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

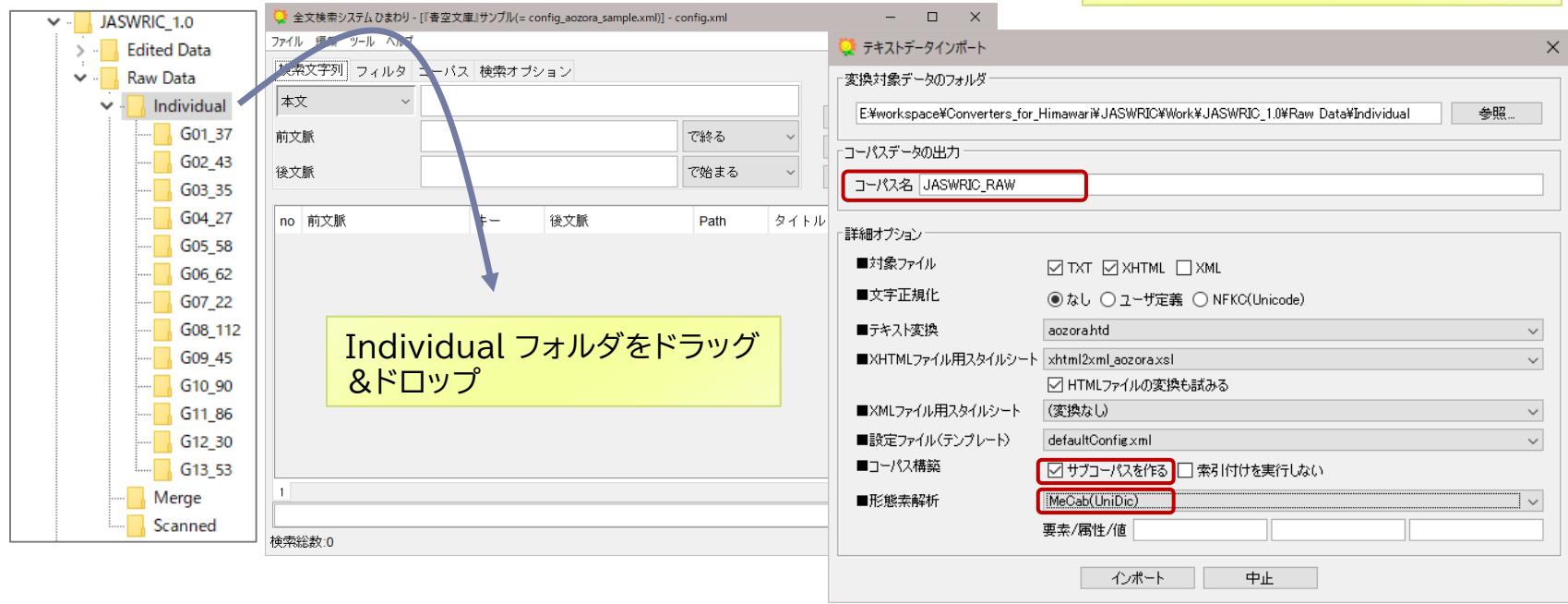
- 『ひまわり』の各種機能を利用できること
(例: 全文検索, 単語検索, 集計機能など)
- 検索システム中の元データを確認・検証できること
- 原資料に対するアノテーション・編集ができること
(例: 形態素解析システム JUMANでアノテーション)

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
- 「サーブコーパスを作る」
- MeCab(UniDic)



動作の確認：検索と作文全体の閲覧(1)

▶ 「行く」(出現形)で検索

The screenshot shows the search interface for the全文検索システム ひまわり. The search term '行く' is entered in the search field, and the search type is set to '出現形' (Occurrence). The results table lists 174 entries where '行く' appears as an occurrence. A specific result, G02_Pic_029, is highlighted and shown in a detailed view window.

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	暑にピクニックに	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	かの確任をしていたと	/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	しょに公園へデートに	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	どのルートをとおって	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、目を離した	/Individual...	G10_Pic...		動詞	非自立可能			五段-力行	終止形-一般	行く

検索結果数: 174

右側の詳細表示窓: G02_Pic_029 :
マリとケンはちずをみているあいだに、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、「せっかくつくったのに」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 フィルタ 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 行く

前文脈 で終る
後文脈 で始まる

検索 ボタン
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	星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

[20]一覧

ファイル フィルタ ツール

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
00000010	あいだ	名詞	普通名詞	副詞可能				間	あいだ	アイダ	和
00000011	に	助詞	格助詞					に	ニ	ニ	和
00000012	、	補助記号	読点					、	、		記号
00000013	犬	名詞	普通名詞	一般				犬	犬	イヌ	和
00000014	が	助詞	格助詞					が	が	ガ	和
00000015	入り	動詞	一般			五段-ラ行	連用形-一般	入る	入り	ハイリ	和
00000016	まし	助動詞				助動詞-マス	連用形-一般	ます	まし	マシ	和
00000017	た	助動詞				助動詞-タ	終止形-一般	た	タ	タ	和
00000018	。	補助記号	句点					。	。		記号
00000019	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
00000020	に	助詞	格助詞					に	ニ	ニ	和
00000021	行く	動詞	非自立可能			五段-力行	連体形-一般	行く	行く	イク	和
00000022	あい	接頭辞						相	あい	アイ	和
00000023	だり	助詞	副助詞								
00000024	犬	名詞	普通名詞	一般							
00000025	が	助詞	格助詞								
00000026	サンドイッチ	名詞	普通名詞	一般							

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

0000021
総数(延べ)：69

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

後文脈

で終る

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

後文脈

で終る

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

後文脈

で終る

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

出現形 基本形 歩く

前文脈

後文脈

で終る

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



B) 先頭が「日」の単語

正規表現の「^日」と同義
(先頭の文字が「日」)



C) 末尾が「日」の単語

正規表現の「日\$」と同義
(末尾の文字が「日」)



D) 単語「日」のみ

正規表現の「^日\$」と
同義



収録内容の確認

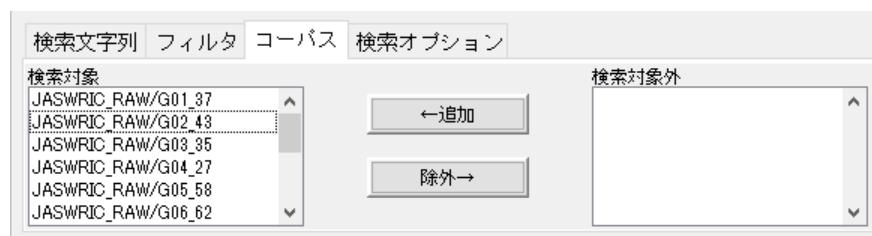
□ 作文一覧

([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])

The diagram illustrates the workflow for generating a word list based on user input. It shows three windows: 1. '要素一覧作成 (ユーザ入力)' dialog with '一部選択' and '頻度' checked. 2. '要素一覧作成 (ユーザ入力)' dialog with various checkboxes like 品詞, 品詞細分類1-3, 活用形, etc., checked. 3. ' [1] 一覧: morph' window showing a table of words with their morphological features and frequency counts.

要素一覧作成 (ユーザ入力)

要素一覧作成 (ユーザ入力)

[1] 一覧: morph

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
ど	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

チェックした属性を組として、
頻度が集計される

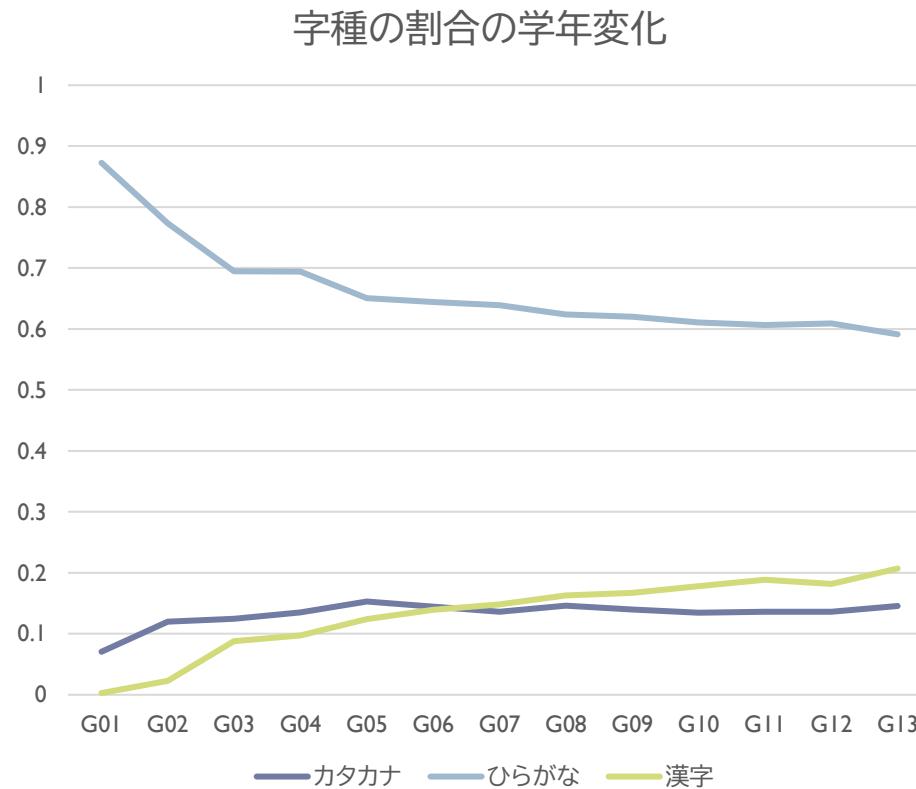
- 集計する属性にチェックを入れる
- 「活用形」ごとに集計しないので、
チェックはしない
- {TEXT}は、出現形

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)

The screenshot shows the 'Himawari' full-text search system interface. The search query in the search bar is '\p{InKatakana}'.

Search results table:

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	ミソファミレドミフ	ア	ソラファミレドうわ	/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	ミソファミレドミフ	/Individual...	G08_Pic...		名詞
6	ア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	/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、ケンはレディフ	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	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	があきません マリに	/Individual...	G02_Key...		名詞
13	をかけてもなかなかド	ア	が開かなかつたため、	/Individual...	G10_Key...		名詞

Search results count: 30169

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 フィルタ コーパス 検索オプション

本文(正規表現) `\p{InKatakana}`

検索

前文脈 で終る 文字数検索

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	ミソファミレドレミファ		ソラファミレド うわ	/Individual...	G08_Pic...	名語	
3	かみる ドドレミミファ		ファミレドソファミン	/Individual...	G08_Pic...	名語	
4	ミファファミレドソフア		ミソファミレソファミ	/Individual...	G08_P...	名語	
5	ファミソファミレソフア		ミソファミレドレミフ	/Individual...	G08_P...	名語	
6	アミレドソファミソフア		ミレスソファミソファミ	/Individual...	G08_P...	名語	
7	ミレドレミ フアソラフ	ア	ミレド うわ、大入っ	/Individual...	G08_P...	名語	
8	る ドドレミミファフア		ミレドソファミソフア	/Individual...	G08_P...	名語	
9	ファミレソファミソフア		ミレドレミ フアソラフ	/Individual...	G08_P...	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_Key...	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_P...	名語	
12	力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
G08_Pic_012

②学年情報(G01～G13)
以外を置換で削除

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) 取消

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

③再集計

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_054	66
G06_Pic_055	66
G04_Pic_041	65
G12_Pic_021	65
G13_Pic_016	65
G08_Pic_012	65
G10_Pic_010	65
G04_Pic_042	65
G06_Pic_056	65
G06_Pic_057	65
G06_Pic_058	65
G06_Pic_059	65
G06_Pic_060	65
G06_Pic_061	65
G06_Pic_062	65
G06_Pic_063	65
G06_Pic_064	65
G06_Pic_065	65
G06_Pic_066	65
G06_Pic_067	65
G06_Pic_068	65
G06_Pic_069	65
G06_Pic_070	65
G06_Pic_071	65
G06_Pic_072	65
G06_Pic_073	65
G06_Pic_074	65
G06_Pic_075	65
G06_Pic_076	65
G06_Pic_077	65
G06_Pic_078	65
G06_Pic_079	65
G06_Pic_080	65
G06_Pic_081	65
G06_Pic_082	65
G06_Pic_083	65
G06_Pic_084	65
G06_Pic_085	65
G06_Pic_086	65
G06_Pic_087	65
G06_Pic_088	65
G06_Pic_089	65
G06_Pic_090	65
G06_Pic_091	65
G06_Pic_092	65
G06_Pic_093	65
G06_Pic_094	65
G06_Pic_095	65
G06_Pic_096	65
G06_Pic_097	65
G06_Pic_098	65
G06_Pic_099	65
G06_Pic_100	65
G06_Pic_101	65
G06_Pic_102	65
G06_Pic_103	65
G06_Pic_104	65
G06_Pic_105	65
G06_Pic_106	65
G06_Pic_107	65
G06_Pic_108	65
G06_Pic_109	65
G06_Pic_110	65
G06_Pic_111	65
G06_Pic_112	65
G06_Pic_113	65
G06_Pic_114	65
G06_Pic_115	65
G06_Pic_116	65
G06_Pic_117	65
G06_Pic_118	65
G06_Pic_119	65
G06_Pic_120	65
G06_Pic_121	65
G06_Pic_122	65
G06_Pic_123	65
G06_Pic_124	65
G06_Pic_125	65
G06_Pic_126	65
G06_Pic_127	65
G06_Pic_128	65
G06_Pic_129	65
G06_Pic_130	65
G06_Pic_131	65
G06_Pic_132	65
G06_Pic_133	65
G06_Pic_134	65
G06_Pic_135	65
G06_Pic_136	65
G06_Pic_137	65
G06_Pic_138	65
G06_Pic_139	65
G06_Pic_140	65
G06_Pic_141	65
G06_Pic_142	65
G06_Pic_143	65
G06_Pic_144	65
G06_Pic_145	65
G06_Pic_146	65
G06_Pic_147	65
G06_Pic_148	65
G06_Pic_149	65
G06_Pic_150	65
G06_Pic_151	65
G06_Pic_152	65
G06_Pic_153	65
G06_Pic_154	65
G06_Pic_155	65
G06_Pic_156	65
G06_Pic_157	65
G06_Pic_158	65
G06_Pic_159	65
G06_Pic_160	65
G06_Pic_161	65
G06_Pic_162	65
G06_Pic_163	65
G06_Pic_164	65
G06_Pic_165	65
G06_Pic_166	65
G06_Pic_167	65
G06_Pic_168	65
G06_Pic_169	65
G06_Pic_170	65
G06_Pic_171	65
G06_Pic_172	65
G06_Pic_173	65
G06_Pic_174	65
G06_Pic_175	65
G06_Pic_176	65
G06_Pic_177	65
G06_Pic_178	65
G06_Pic_179	65
G06_Pic_180	65
G06_Pic_181	65
G06_Pic_182	65
G06_Pic_183	65
G06_Pic_184	65
G06_Pic_185	65
G06_Pic_186	65
G06_Pic_187	65
G06_Pic_188	65
G06_Pic_189	65
G06_Pic_190	65
G06_Pic_191	65
G06_Pic_192	65
G06_Pic_193	65
G06_Pic_194	65
G06_Pic_195	65
G06_Pic_196	65
G06_Pic_197	65
G06_Pic_198	65
G06_Pic_199	65
G06_Pic_200	65
G06_Pic_201	65
G06_Pic_202	65
G06_Pic_203	65
G06_Pic_204	65
G06_Pic_205	65
G06_Pic_206	65
G06_Pic_207	65
G06_Pic_208	65
G06_Pic_209	65
G06_Pic_210	65
G06_Pic_211	65
G06_Pic_212	65
G06_Pic_213	65
G06_Pic_214	65
G06_Pic_215	65
G06_Pic_216	65
G06_Pic_217	65
G06_Pic_218	65
G06_Pic_219	65
G06_Pic_220	65
G06_Pic_221	65
G06_Pic_222	65
G06_Pic_223	65
G06_Pic_224	65
G06_Pic_225	65
G06_Pic_226	65
G06_Pic_227	65
G06_Pic_228	65
G06_Pic_229	65
G06_Pic_230	65
G06_Pic_231	65
G06_Pic_232	65
G06_Pic_233	65
G06_Pic_234	65
G06_Pic_235	65
G06_Pic_236	65
G06_Pic_237	65
G06_Pic_238	65
G06_Pic_239	65
G06_Pic_240	65
G06_Pic_241	65
G06_Pic_242	65
G06_Pic_243	65
G06_Pic_244	65
G06_Pic_245	65
G06_Pic_246	65
G06_Pic_247	65
G06_Pic_248	65
G06_Pic_249	65
G06_Pic_250	65
G06_Pic_251	65
G06_Pic_252	65
G06_Pic_253	65
G06_Pic_254	65
G06_Pic_255	65
G06_Pic_256	65
G06_Pic_257	65
G06_Pic_258	65
G06_Pic_259	65
G06_Pic_260	65
G06_Pic_261	65
G06_Pic_262	65
G06_Pic_263	65
G06_Pic_264	65
G06_Pic_265	65
G06_Pic_266	65
G06_Pic_267	65
G06_Pic_268	65
G06_Pic_269	65
G06_Pic_270	65
G06_Pic_271	65
G06_Pic_272	65
G06_Pic_273	65
G06_Pic_274	65
G06_Pic_275	65
G06_Pic_276	65
G06_Pic_277	65
G06_Pic_278	65
G06_Pic_279	65
G06_Pic_280	65
G06_Pic_281	65
G06_Pic_282	65
G06_Pic_283	65
G06_Pic_284	65
G06_Pic_285	65
G06_Pic_286	65
G06_Pic_287	65
G06_Pic_288	65
G06_Pic_289	65
G06_Pic_290	65
G06_Pic_291	65
G06_Pic_292	65
G06_Pic_293	65
G06_Pic_294	65
G06_Pic_295	65
G06_Pic_296	65
G06_Pic_297	65
G06_Pic_298	65
G06_Pic_299	65
G06_Pic_300	65
G06_Pic_301	65
G06_Pic_302	65
G06_Pic_303	65
G06_Pic_304	65
G06_Pic_305	65
G06_Pic_306	65
G06_Pic_307	65
G06_Pic_308	65
G06_Pic_309	65
G06_Pic_310	65
G06_Pic_311	65
G06_Pic_312	65
G06_Pic_313	65
G06_Pic_314	65
G06_Pic_315	65
G06_Pic_316	65
G06_Pic_317	65
G06_Pic_318	65
G06_Pic_319	65
G06_Pic_320	65
G06_Pic_321	65
G06_Pic_322	65
G06_Pic_323	65
G06_Pic_324	65
G06_Pic_325	65
G06_Pic_326	65
G06_Pic_327	65
G06_Pic_328	65
G06_Pic_329	65
G06_Pic_330	65
G06_Pic_331	65
G06_Pic_332	65
G06_Pic_333	65
G06_Pic_334	65
G06_Pic_335	65
G06_Pic_336	65
G06_Pic_337	65
G06_Pic_338	65
G06_Pic_339	65
G06_Pic_340	65
G06_Pic_341	65
G06_Pic_342	65
G06_Pic_343	65
G06_Pic_344	65
G06_Pic_345	65
G06_Pic_346	65
G06_Pic_347	65
G06_Pic_348	65
G06_Pic_349	65
G06_Pic_350	65
G06_Pic_351	65
G06_Pic_352	65
G06_Pic_353	65
G06_Pic_354	65
G06_Pic_355	65
G06_Pic_356	65
G06_Pic_357	65
G06_Pic_358	65
G06_Pic_359	65
G06_Pic_360	65
G06_Pic_361	65
G06_Pic_362	65
G06_Pic_363	65
G06_Pic_364	65
G06_Pic_365	65
G06_Pic_366	65
G06_Pic_367	65
G06_Pic_368	65
G06_Pic_369	65
G06_Pic_370	65
G06_Pic_371	65
G06_Pic_372	65
G06_Pic_373	65
G06_Pic_374	65
G06_Pic_375	65
G06_Pic_376	65
G06_Pic_377	65
G06_Pic_378	65
G06_Pic_379	65
G06_Pic_380	65
G06_Pic_381	65
G06_Pic_382	65
G06_Pic_383	65
G06_Pic_384	65
G06_Pic_385	65
G06_Pic_386	65
G06_Pic_387	65
G06_Pic_388	65
G06_Pic_389	65
G06_Pic_390	65
G06_Pic_391	65
G06_Pic_392	65
G06_Pic_393	65
G06_Pic_394	65
G06_Pic_395	65
G06_Pic_396	65
G06_Pic_397	65
G06_Pic_398	65
G06_Pic_399	65
G06_Pic_400	65
G06_Pic_401	65
G06_Pic_402	65
G06_Pic_403	65
G06_Pic_404	65
G06_Pic_405	65
G06_Pic_406	65
G06_Pic_407	65
G06_Pic_408	65
G06_Pic_409	65
G06_Pic_410	65
G06_Pic_411	65
G06_Pic_412	65
G06_Pic_413	65
G06_Pic_414	65
G06_Pic_415	65
G06_Pic_416	65
G06_Pic_417	65
G06_Pic_418	65
G06_Pic_419	65
G06_Pic_420	65
G06_Pic_421	65
G06_Pic_422	65
G06_Pic_423	65
G06_Pic_424	65
G06_Pic_425	65
G06_Pic_426	65
G06_Pic_427	65
G06_Pic_428	65
G06_Pic_429	65
G06_Pic_430	65
G06_Pic_431	65
G06_Pic_432	65
G06_Pic_433	65
G06_Pic_434	65
G06_Pic_435	65
G06_Pic_436	65
G06_Pic_437	65
G06_Pic_438	65
G06_Pic_439	65
G06_Pic_440	65
G06_Pic_441	65
G06_Pic_442	65
G06_Pic_443	65
G06_Pic_444	65
G06_Pic_445	65
G06_Pic_446	65
G06_Pic_447	65
G06_Pic_448	65
G06_Pic_449	65
G06_Pic_450	65
G06_Pic_451	65
G06_Pic_452	65
G06_Pic_453	65
G06_Pic_454	65
G06_Pic_455	65
G06_Pic_456	65
G06_Pic_457	65
G06_Pic_458	65
G06_Pic_459	65
G06_Pic_460	65
G06_Pic_461	65
G06_Pic_462	65
G06_Pic_463	65
G06_Pic_464	65
G06_Pic_465	65
G06_Pic_466	65
G06_Pic_467	65
G06_Pic_468	65
G06_Pic_469	65
G06_Pic_470	65
G06_Pic_471	65
G06_Pic_472	65
G06_Pic_473	65
G06_Pic_474	65
G06_Pic_475	65
G06_Pic_476	65
G06_Pic_477	65
G06_Pic_478	65
G06_Pic_479	65
G06_Pic_480	65
G06_Pic_481	65
G06_Pic_482	65
G06_Pic_483	65
G06_Pic_484	65
G06_Pic_485	65
G06_Pic_486	65
G06_Pic_487	65
G06_Pic_488	65
G06_P	

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：

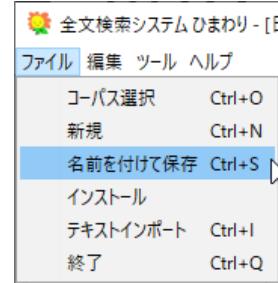
Ctrlキー+Shiftキー+C

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



- ▶ 「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

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	ないんだああああああ！	^	!!とケンは必死に	/Individual...	G13_Key...		補助記号
2	いんだああああああ！	!	!!とケンは必死に抵	/Individual...	G13_Key...		補助記号
3	…びっくりぎょうてん！		!!犬がせっかくつく	/Individual...	G04_Pic...		名詞
4	由を説明するしかない！		！「あのっ！」と言っ	/Individual...	G06_Key...		補助記号
5	た。「マリ！」「ケン！		！」今のことからよ	/Individual...	G03_Key...		補助記号
6	んに、マリが「お帰り！		！」と大きな声を出し	/Individual...	G06_Key...		補助記号
7	できました。「まさか！		！」と思いつながらバス	/Individual...	G05_Pic...		名詞
8	た。「よし、食べよう！		！」と言い、バスケッ	/Individual...	G06_Pic...		名詞
9	付きました。「そうだ！		！」ケンは物置きから	/Individual...	G05_Key...		補助記号
10	ら、「わあっ！」ボチ！		!お弁当全部食べちゃ	/Individual...	G06_Pic...		名詞
11	た…。ごはんにしよう！		！と、思ったら、「わ	/Individual...	G06_Pic...		名詞
12	なんだ。これで、よし！		！とそこに、「こんな	/Individual...	G06_Key...		補助記号
13	んだああああああ!!!		！とケンは必死に抵抗	/Individual...	G13_Key...		補助記号

検索結果

統計

① タイトル列で集計

正規表現「.」(半角のピリオド)
… 任意の1文字

② 学年情報(G01～G13)
以外を置換で削除
③ 再集計

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

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