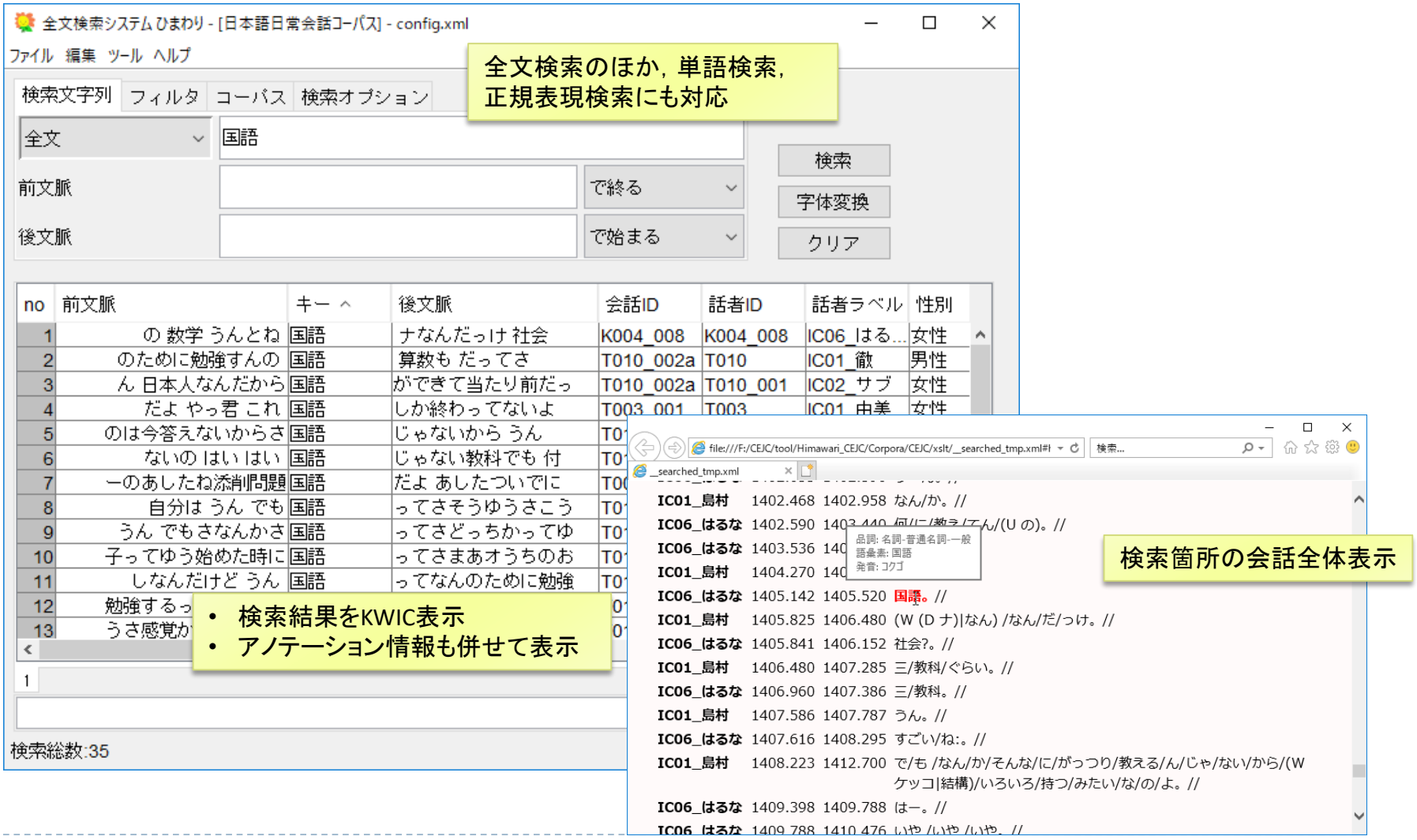
『ひまわり』用CEJC

- CEJCを全文検索システム『ひまわり』にインポートしたもの
  - 転記テキスト（発話単位版，短単位情報付き）
  - 話者データベース
  - 会話データベース
  - 動画

- できること
  - 全文検索，単語検索
  - メタデータの参照
  - コーパス基礎データの集計
  - 動画の閲覧・アノテーション

全文検索


全文検索システム ひまわり - [日本語日常会話コーパス] - config.xml
ファイル 編集 ツール ヘルプ
全文検索のほか，単語検索，
正規表現検索にも対応
検索文字列 フィルタ コーパス 検索オプション
全文
国語
前文脈
で終る
後文脈
で始まる
検索
字体変換
クリア
no 前文脈 キー 後文脈 会話ID 話者ID 話者ラベル 性別
1 の 数学 うんとね 国語 ナなんだっけ 社会 K004_008 K004_008 IC06_はる... 女性
2 のために勉強すんの 国語 算数も だってさ T010_002a T010 IC01_徹 男性
3 ん 日本人なんだから 国語 ができて当たり前だっ T010_002a T010_001 IC02_サブ 女性
4 だよ やっ君 これ 国語 しか終わってないよ T003_001 T003 IC01_由美 女性
5 のは今答えないからさ 国語 じゃないから うん
6 ないの はい はい 国語 じゃない教科でも 付
7 ーのあしたね添削問題 国語 だよ あしたついでに
8 自分は うん でも 国語 ってさそうゆうさこう
9 うん でもさなんかさ 国語 ってさどっちかってゆ
10 子ってゆう始めた時に 国語 ってさまあオうちのお
11 しなんだけど うん 国語 ってなんのために勉強
12 勉強するっ
13 うさ感覚か
• 検索結果をKWIC表示
• アノテーション情報も併せて表示
検索総数:35
file:///F:/CEJC/tool/Himawari_CEJC/Corpora/CEJC/xslt/_searched_tmp.xml#I
_searched_tmp.xml
IC01_島村 1402.468 1402.958 なん/か。//
IC06_はるな 1402.590 1403.440 何/に/教える/ん/(U の)。//
IC06_はるな 1403.536 1404.270
IC01_島村 1404.270 1405.142
IC06_はるな 1405.142 1405.520 国語。//
IC01_島村 1405.825 1406.480 (W (D ナ)|なん) /なん/だ/っけ。//
IC06_はるな 1405.841 1406.152 社会?。//
IC01_島村 1406.480 1407.285 三/教科/ぐらい。//
IC06_はるな 1406.960 1407.386 三/教科。//
IC01_島村 1407.586 1407.787 うん。//
IC06_はるな 1407.616 1408.295 すごい/ね:。//
IC01_島村 1408.223 1412.700 で/も /なん/か/そんな/に/がっつり/教える/ん/じゃ/ない/から/(W ケッコ|結構)/いろいろ/持つ/みたい/な/の/よ。//
IC06_はるな 1409.398 1409.788 はー。//
品詞: 名詞-普通名詞-一般
語彙素: 国語
発音: コクゴ
検索箇所の会話全体表示

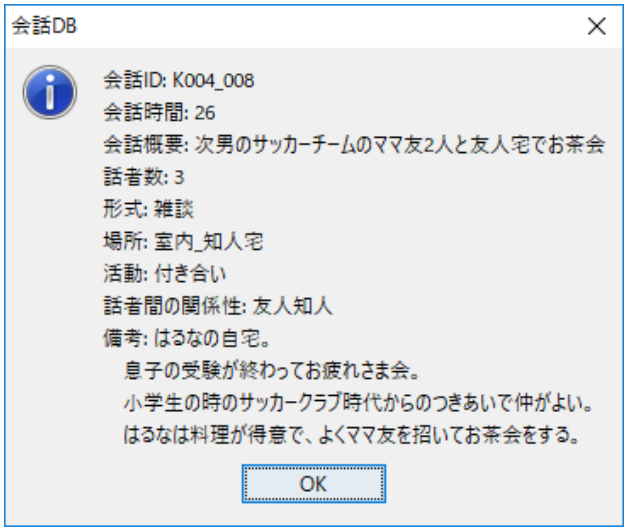
メタ情報（話者，会話）の参照

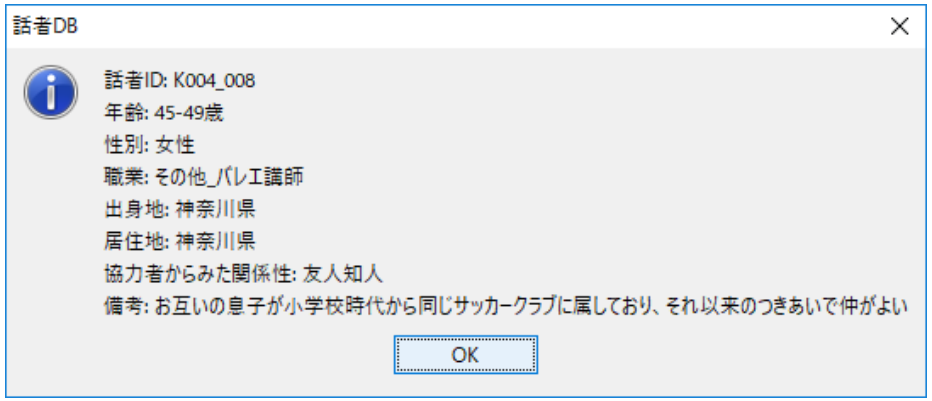
…	前文脈	キー	後文脈	会話ID	話者ID	話者ラ…	性別	年齢
1	の 数学 うんとね	国語	ナなんだっけ 社会	K004_008	K004_008	IC06_は…	女性	45
2	のために勉強すんの	国語	算数も だってさ	T010_002a	T010	IC01_徹	男性	20
3	ん 日本人なんだから	国語	ができて当たり前…	T010_002a	T010_001	IC02_サ…	女性	50
4	だよ やっ君 これ	国語	しか終わってないよ	T003_001	T00			
5	のは今答えないからさ	国語	じゃないから うん	T010_002a	T01			
6	ないの はい はい	国語	じゃない教科でも 付	T010_007	T01			
7	ーのあしたね添削問題	国語	だよ あしたついでに	T003_001	T003_002	IC03_大…	男性	

閲覧したい用例の「会話ID」「話者ID」列をダブルクリック

話者ID

会話ID

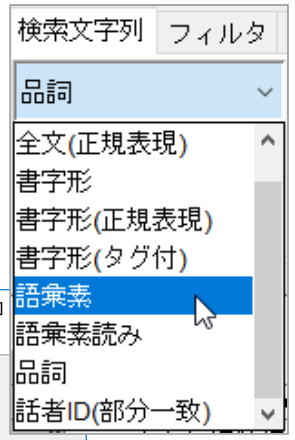




話者一覧，会話情報一覧を表示することも可能

単語検索

- マッチングの範囲は単語（短単位）
- 単位をまたいだ検索はできない
- 前後2単語の語彙素も表示（例：「語彙素1」「語彙素-1」）

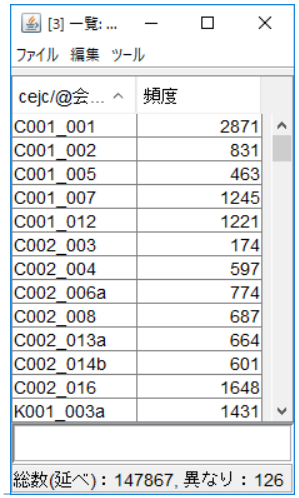


基礎データの集計

- アノテーションされている情報をもとにコーパスの基礎データを集計
  - 会話データごとの発話数，単語数
  - 語彙頻度表
  - 文字・単語n-gram

詳しくは，マニュアルページ参照
[ヘルプ]⇒[『ひまわり』マニュアル]

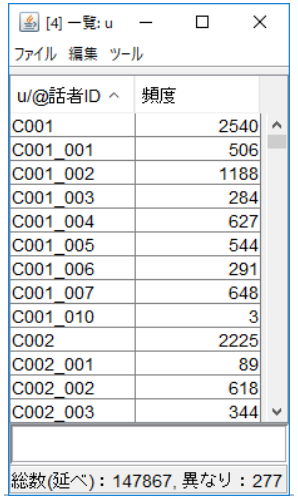
- 会話データごとの発話数



cejc/@会...	頻度
C001_001	2871
C001_002	831
C001_005	463
C001_007	1245
C001_012	1221
C002_003	174
C002_004	597
C002_006a	774
C002_008	687
C002_013a	664
C002_014b	601
C002_016	1648
K001_003a	1431

総数(延べ)：147867, 異なり：126

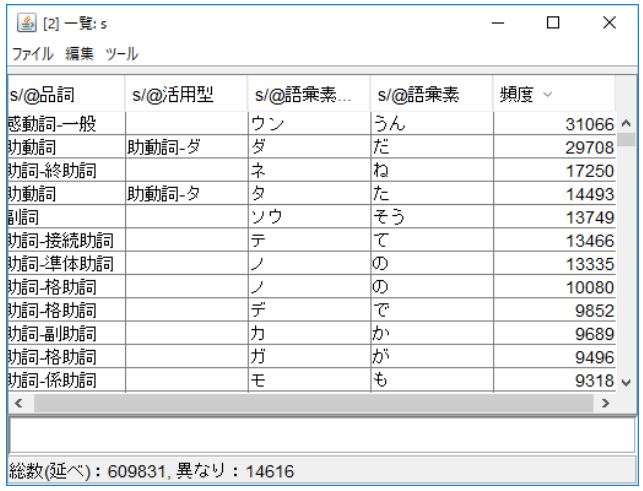
- 話者ごとの発話数



u/@話者ID	頻度
C001	2540
C001_001	506
C001_002	1188
C001_003	284
C001_004	627
C001_005	544
C001_006	291
C001_007	648
C001_010	3
C002	2225
C002_001	89
C002_002	618
C002_003	344

総数(延べ)：147867, 異なり：277

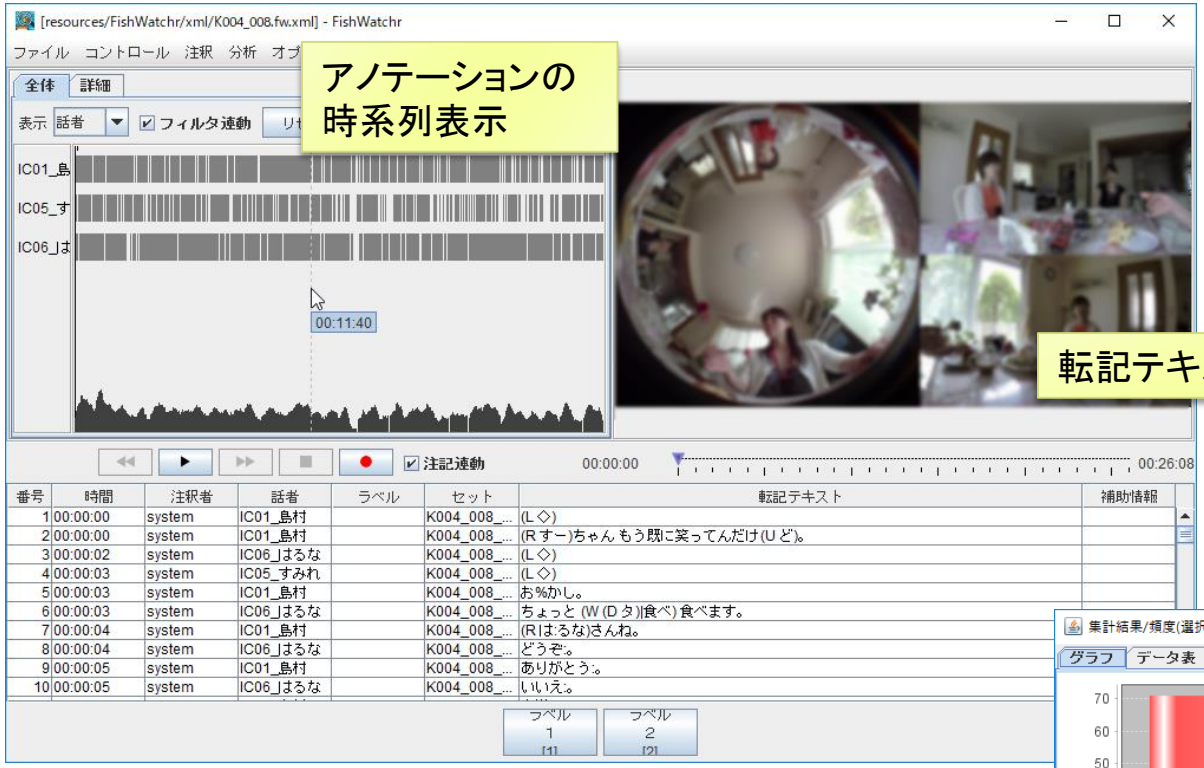
- 語彙頻度表



s/@品詞	s/@活用型	s/@語彙素...	s/@語彙素	頻度
感動詞-一般		ウン	うん	31066
助動詞	助動詞-ダ	ダ	だ	29708
助詞-終助詞		ネ	ね	17250
助動詞	助動詞-タ	タ	た	14493
副詞		ソウ	そう	13749
助詞-接続助詞		テ	て	13466
助詞-準体助詞		ノ	の	13335
助詞-格助詞		ノ	の	10080
助詞-格助詞		デ	で	9852
助詞-副助詞		カ	か	9689
助詞-格助詞		ガ	が	9496
助詞-係助詞		モ	も	9318

総数(延べ)：609831, 異なり：14616

動画の閲覧・アノテーション

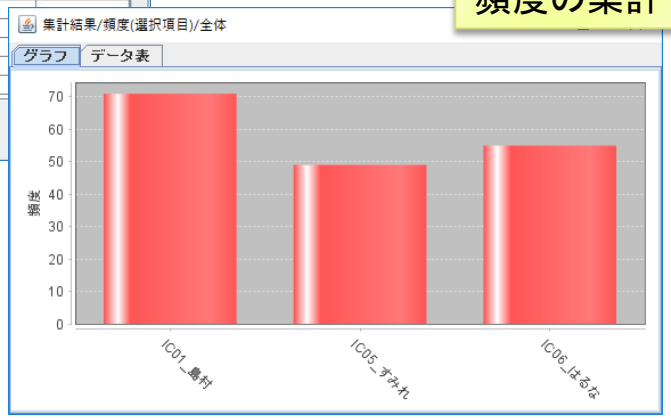


アノテーションの
時系列表示

転記テキストと同期しつつ，再生

ボタン操作で再生位置に簡単な
アノテーションができる

頻度の集計



試してみるには

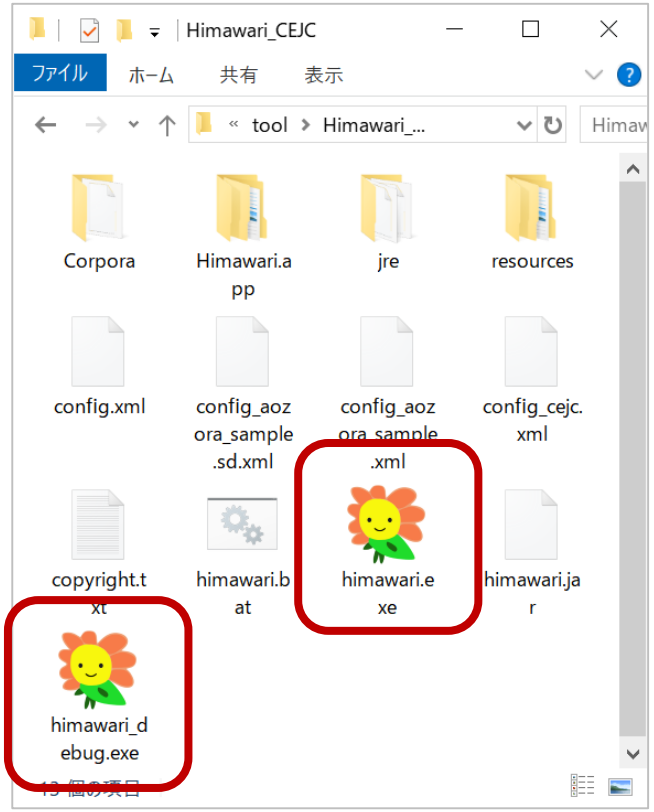
- WindowsとmacOSに対応しています

- 配布ハードディスクをお持ちの方
  - PCに接続するだけです

- 配布ハードディスクをPCにコピーして利用する場合
  - ハードディスクの内容をすべてPCにコピーしてください
  - ただし，コピーする際は，フォルダの構造を変更しないでください

『ひまわり』の起動

- tool ⇒ Himawari_CEJC フォルダ
  - Windowsの場合



  - macOSの場合