



「日常会話コーパス」モニター版 『ひまわり』講習会

山口昌也(国立国語研究所)



本日の内容

- ▶ 全文検索システム『ひまわり』を使った『日本語日常会話コーパス』の利用方法を紹介
 - ▶ 『日本語日常会話コーパス』(モニタ公開版の講習会用パッケージ)
 - ▶ 『ひまわり』(ver.1.6.4)
 - ▶ FishWatchr (ver.0.9.15b)

- ▶ 全体的な流れ
 - ▶ 講習会用パッケージの説明
 - ▶ 『ひまわり』の紹介と基本的な使い方
 - ▶ 検索機能の詳細
 - ▶ コーパスの基礎情報の集計とコーパスの構造
 - ▶ 動画の閲覧・アノテーション

『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

▶ 特徴

- ▶ XMLでタグづけされたコーパスを全文検索できます
- ▶ Windows, Mac OS, Linux など、多くのOS上で動作します
- ▶ 無料です

Ver.1.6 ⇒ 統計的な分析に必要なデータの収集支援機能を強化
(例:総文字数, 総単語数)

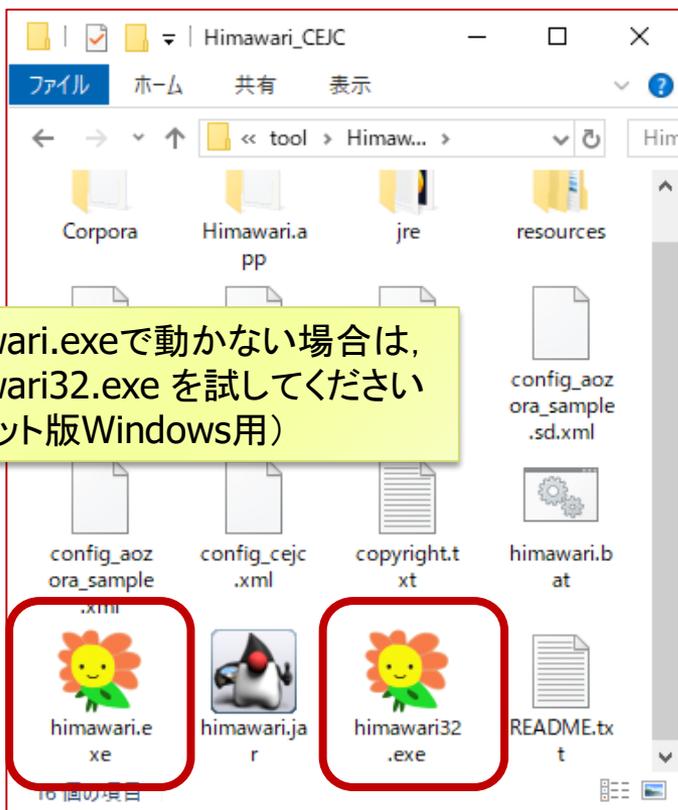
『ひまわり』の基本的な使い方 (CEJC編)



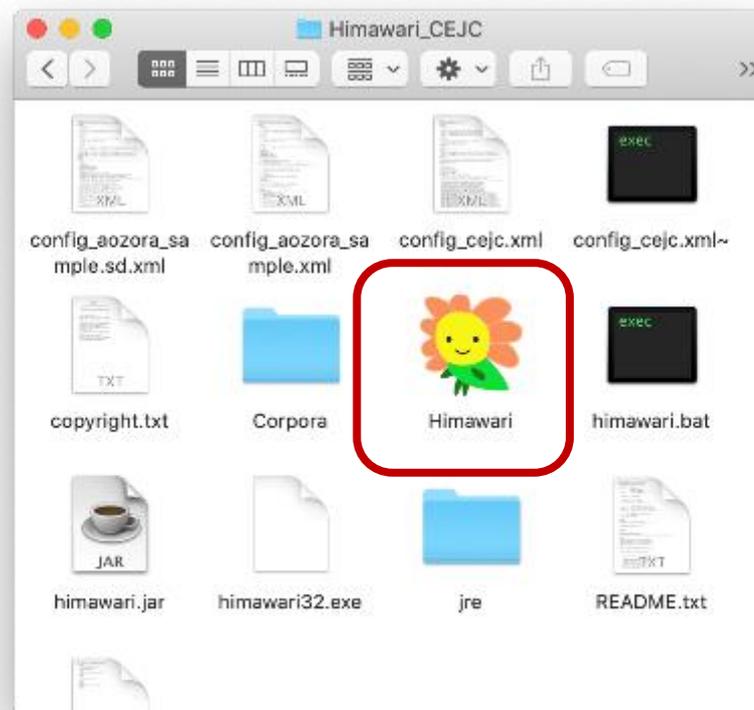
『ひまわり』を起動する

▶ tool ⇒ Himawari_CEJC フォルダ

☐ Windowsの場合



☐ macOSの場合



トラブルシューティング(2)

▶ Windowsで『ひまわり』の文字が小さすぎる

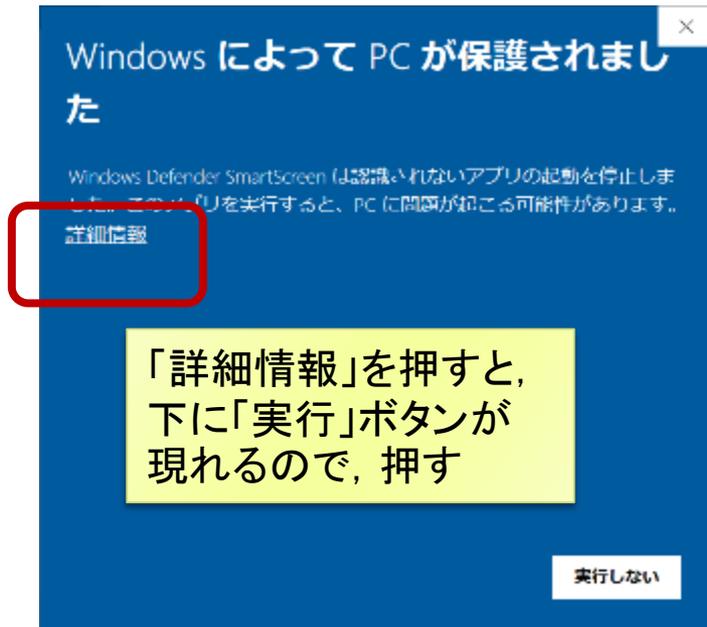
himawari.exe (himawari32.exe) を
右クリックして、プロパティを選択

クリック！！

「システム」
を選択

トラブルシューティング(3)

- ▶ 起動時にエラーが出る
(Windowsの場合)



- ▶ 起動時にエラーが出る
(macOSの場合)

- ▶ ダウンロードした講習会用パッケージは、正常に動作しません。
- ▶ 別途ファイルを配布しますので、ご相談ください。

検索する

「検索文字列」欄では
右クリックで履歴表示

```
C:\Users\masaya\Desktop
eix:s
eix:cejc
eix:u
n:71
info(available memory
time[milisc]: 64
time[milisc]: 79
history added: じゃん
doSearch:start
corpusname CEJC
open eix
eix:s
eix:cejc
eix:u
n:71
info(available memory
time[milisc]: 46
time[milisc]: 62
history added: じゃん
doSearch:start
corpusname CEJC
open eix
eix:s
eix:cejc
eix:u
n:71
info(available memory
time[milisc]: 31
time[milisc]: 31
```

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

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

全文 じゃん

前文脈 後文脈

検索

検索の実行

| no | 前文脈 | キー | 後文脈 | 会話ID | 話者ID | 話者ラベル | 性別 |
|----|------------|-----|------------|----------|----------|------------|----|
| 1 | ナはあの小ささがいい | じゃん | ◇ あの細長さでさ | T010_004 | T010 | IC01 徹 | 男性 |
| 2 | だよ キャッチする人 | じゃん | ◇ お前超怖えから | T010_004 | T010_004 | IC02 大場 | 男性 |
| 3 | 器吹い 腹式やってん | じゃん | ◇ やってるにも聞 | T010_004 | T010 | IC01 徹 | 男性 |
| 4 | てもう就活始まって | じゃん | ああ だから今年度 | T004_001 | T004_002 | IC03 さと... | 男性 |
| 5 | ついてね あれもあん | じゃん | あのタッチするやつ | T010_004 | T010 | IC01 徹 | 男性 |
| 6 | だからカットしに来る | じゃん | あー はいはいはい | T010_004 | T010 | IC01 徹 | 男性 |
| 7 | ん うん ばちーなる | じゃん | あー 俺痛いじゃん | T010_004 | T010 | IC01 徹 | 男性 |
| 8 | かみたいでさ全部やる | じゃん | うん うん あー | T010_004 | T010_003 | IC03 龍之... | 男性 |
| 9 | 回ナセカンドに投げん | じゃん | うん うん うん | T010_004 | T010 | IC01 徹 | 男性 |
| 10 | だからキャッチャー | じゃん | うん うん うん | T010_004 | T010 | IC01 徹 | 男性 |
| 11 | たら絶対もう合わない | じゃん | うん うん だから | T010_004 | T010 | IC01 徹 | 男性 |
| 12 | いや みんなねゆう | じゃん | うん うん んで俺 | T010_004 | T010 | IC01 徹 | 男性 |
| 13 | ヒットも一本のヒット | じゃん | うん うん うん じ | T010_004 | T010_004 | IC02 大場 | 男性 |

検索総数: 71

途中経過の表示

検索総数

検索結果

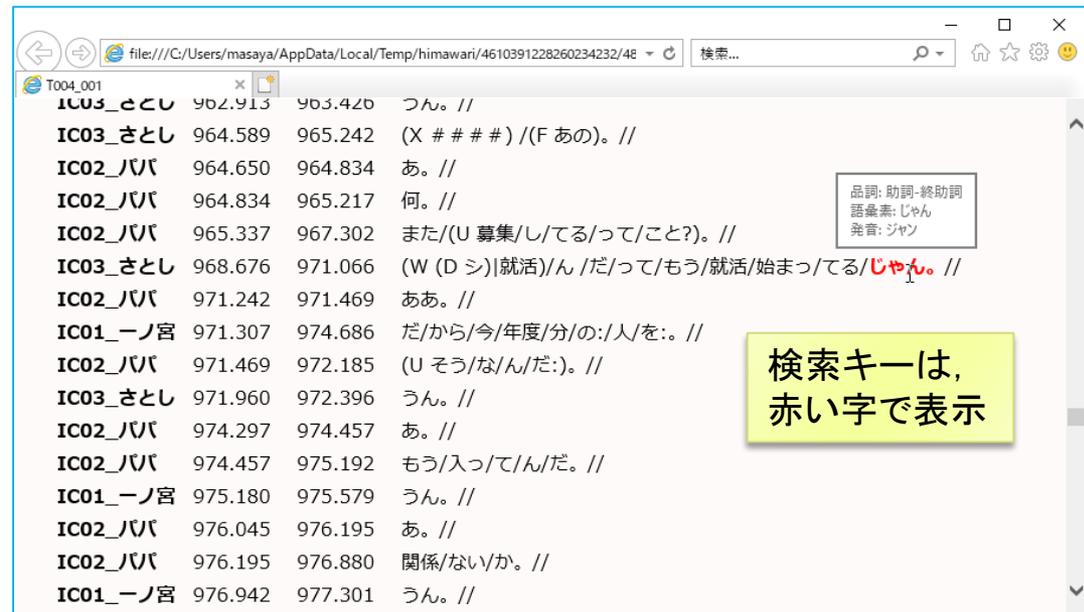
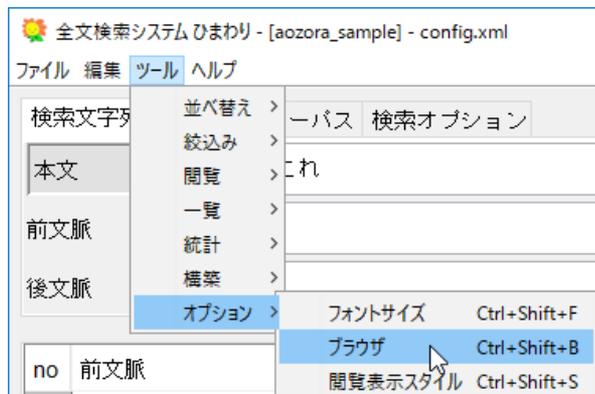
転記テキストの閲覧

| no | 前文脈 | キー | 後文脈 | 会話ID | 話者ID | 話者ラベル | 性別 |
|----|------------|-----|-----------|----------|----------|---------|----|
| 1 | ナはあの小ささがいい | じゃん | ◇あの細長さでさ | T010_004 | T010 | IC01_徹 | 男性 |
| 2 | だよ キャッチする人 | じゃん | ◇お前超怖えから | T010_004 | T010_004 | IC02_大場 | 男性 |
| 3 | 器吹い 腹式やってん | じゃん | ◇やってるにも関 | T010_004 | T | | |
| 4 | てもう就活始まって | じゃん | ああ だから今年度 | T004_001 | T | | |
| 5 | ついてね あれもあん | じゃん | あのタッチするやつ | T010_004 | T | | |
| 6 | だからカットしに来る | じゃん | あー はいはいはい | T010_004 | T010 | IC01_徹 | 男性 |
| 7 | ん うん ばちーなる | じゃん | あー 俺痛いじゃん | T010_004 | T010 | IC01_徹 | 男性 |

閲覧したい用例の「キー」列などをダブルクリック



閲覧用のブラウザの変更



[ツール]⇒[オプション]⇒[ブラウザ]

話者，会話情報の閲覧

| no | 前文脈 | キー | 後文脈 | 会話ID | 話者ID | 話者ラベル | 性別 |
|----|------------|-----|-----------|----------|----------|------------|----|
| 1 | ナはあの小ささがいい | じゃん | ◇ あの細長さでさ | T010_004 | T010 | IC01_徹 | 男性 |
| 2 | だよ キャッチする人 | じゃん | ◇ お前超怖えから | T010_004 | T010_004 | IC02_大場 | 男性 |
| 3 | 器吹い 腹式やってん | じゃん | ◇ やってるにも関 | T010_004 | T010 | IC01_徹 | 男性 |
| 4 | でもう就活始まって | じゃん | ああ だから今年度 | T004_001 | T004_002 | IC03_さと... | 男性 |
| 5 | ついてね あれもあん | じゃん | あのタッチするやつ | T010_004 | T010 | IC01_徹 | 男性 |
| 6 | だからカットしに来 | じゃん | あー はいはいはい | T010_004 | T010 | IC01_徹 | 男性 |
| 7 | ん うん ばちーなる | じゃん | あー 俺痛いじゃん | T010_004 | T010 | IC01_徹 | 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



会話DB ×

i 会話ID: T004_001
 会話時間: 27
 会話概要: 自宅で夫・息子と野球中継を見ながら夕食
 話者数: 3
 形式: 雑談
 場所: 自宅
 活動: 食事
 話者間の関係性: 家族
 備考: 夫・息子は巨人ファンで野球中継をよく見る。
 時間帯が合えば家族揃って家や外で食事をする。
 お酒を飲みながら食事することが多い。

OK

話者DB ×

i 話者ID: T004_002
 年齢: 20-24歳
 性別: 男性
 職業: 会社員・役員・公務員・専門職
 出身地: 東京都
 居住地: 東京都
 協力者からみた関係性: 家族親族_息子
 備考: 就職1年目で一人暮らしを始めたところ。
 姉と2人兄弟。
 実家によく立ち寄る。

OK

会話一覧: [ツール] ⇒ [一覧] ⇒ [会話DB]
 話者一覧: [ツール] ⇒ [一覧] ⇒ [話者DB]

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 | 会話ID | 話者ID | 話者ラベル | 性別 |
|----|------------|------|-----------|----------|----------|------------|----|
| 1 | ナはあの小ささがいい | じゃん | ◇ あの細長さでさ | T010_004 | T010 | IC01_徹 | 男性 |
| 2 | だよ キャッチする人 | じゃん | ◇ お前超怖えから | T010_004 | T010_004 | IC02_大場 | 男性 |
| 3 | 器吹い 腹式やってん | じゃん | ◇ やってるにも関 | T010_004 | T010 | IC01_徹 | 男性 |
| 4 | てもう就活始まって | じゃん | ああ だから今年度 | T004_001 | T004_002 | IC03_さと... | 男性 |
| 5 | ついてね あれもあん | じゃん | あのタッチするやつ | T010_004 | T010 | IC01_徹 | 男性 |
| 6 | だからカットしに来る | じゃん | あー はいはいはい | T010_004 | T010 | IC01_徹 | 男性 |
| 7 | ん うん ばちーなる | じゃん | あー 俺痛いじゃん | T010_004 | T010 | IC01_徹 | 男性 |
| 8 | かみたいでさ全部やる | じゃん | うん うん あー | T010_004 | T010_003 | IC03_龍之... | 男性 |
| 9 | 回ナセカンドに投げん | じゃん | うん うん うん | T010_004 | T010 | IC01_徹 | 男性 |
| 10 | だからキャッチャー | じゃん | うん うん うん | T010_004 | T010 | IC01_徹 | 男性 |
| 11 | たら絶対もう合わない | じゃん | うん うん だから | T010_004 | T010 | IC01_徹 | 男性 |
| 12 | いやみんなねゆう | じゃん | うん うん んで俺 | T010_004 | T010 | IC01_徹 | 男性 |

▶ 昇順・降順

- ▶ 列タイトルをクリックで、昇順・降順が切り替わる
- ▶ シフトを押しながらクリックすると、降順

▶ 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:「性別」ごとに「後文脈」でソート
→ 「後文脈」「性別」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 - [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

検索文字列 フィルタ コーパス 検索オプション

性別 ▼ 男性 で始まる ▼
話者ID ▼ で始まる ▼
話者ラベル ▼ で始まる ▼

「性別」列が「男性」
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 | 会話ID | 話者ID | 話者ラベル | 性別 | 年齢 |
|----|------------|-----|-----------|----------|----------|------------|----|--------|
| 1 | ナはあの小ささがいい | じゃん | ◇ あの細長さでさ | T010_004 | T010 | IC01_徹 | 男性 | |
| 2 | だよ キャッチする人 | じゃん | ◇ お前超怖えから | T010_004 | T010_004 | IC02_大場 | 男性 | |
| | 聞かして聞かして | じゃん | ◇ やってるにも関 | T010_004 | T010 | IC01_徹 | 男性 | |
| | ってるじゃん | じゃん | ああ だから今年度 | T004_001 | T004_002 | IC03_さと... | 男性 | 20-24歳 |
| | もあん | じゃん | あのタッチするやつ | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |
| | に来る | じゃん | あー はいはいはい | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |
| | ーなる | じゃん | あー 俺痛いじゃん | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |

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

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 | 会話ID | 話者ID | 話者ラベル | 性別 | 年齢 |
|----|------------|-----|-----------|----------|----------|------------|----|--------|
| 1 | ナはあの小ささがいい | じゃん | ◇あの細長さでさ | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |
| 2 | だよ キャッチする人 | じゃん | ◇お前超怖えから | T010_004 | T010_004 | IC02_大場 | 男性 | 20-24歳 |
| 3 | 器吹い 腹式やってん | じゃん | ◇やってるにも関 | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |
| 4 | てもう就活始まって | じゃん | ああ だから今年度 | T004_001 | T004_002 | IC03_さと... | 男性 | 20-24歳 |
| 5 | ついてね あれもあん | じゃん | あのタッチするやつ | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |
| 6 | だからカットしに来 | じゃん | あーはいはいはい | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |
| 7 | ん うん ばちーなる | じゃん | あー 俺痛いじゃん | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |
| 8 | かみたいでさ全部やる | じゃん | うん うん あー | T010_004 | T010_003 | IC03_龍之... | 男性 | 20-24歳 |
| 9 | 回ナセカンドに投げん | じゃん | うん うん うん | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |
| 10 | だからキャッチャー | じゃん | うん うん うん | T010_004 | T010 | IC01_徹 | 男性 | 20-24歳 |

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」

| 話者ID | 話者ラベル | 性別 | 年齢 |
|----------|------------|----|--------|
| T010 | IC01_徹 | 男性 | 20-24歳 |
| T010_004 | IC02_大場 | 男性 | |
| T010 | IC01_徹 | 男性 | |
| T004_002 | IC03_さと... | 男性 | |
| T010 | IC01_徹 | 男性 | |
| T010 | IC01_徹 | 男性 | |
| T010 | IC01_徹 | 男性 | |
| T010_003 | IC03_龍之... | 男性 | |
| T010 | IC01_徹 | 男性 | 20-24歳 |
| T010 | IC01_徹 | 男性 | 20-24歳 |
| T010 | IC01_徹 | 男性 | 20-24歳 |
| T010 | IC01_徹 | 男性 | 20-24歳 |
| T010_004 | IC02_大場 | 男性 | 20-24歳 |



| 性別 | 年齢 | 頻度 |
|----|--------|----|
| 男性 | 20-24歳 | 51 |
| 女性 | 35-39歳 | 9 |
| 女性 | 60-64歳 | 9 |
| 男性 | 5-9歳 | 1 |
| 女性 | 30-34歳 | 1 |

総数(延べ)：71, 異なり：5

結果のエクスポート

▶ クリップボードを使用する方法

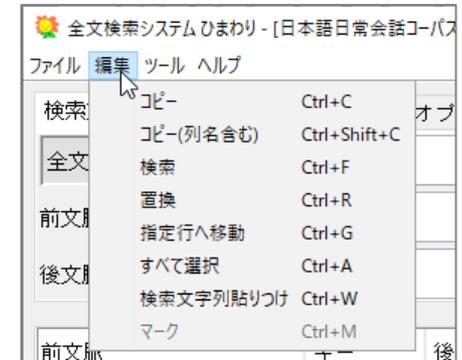
1. 結果を選択

- ▶ 全選択したい場合は, Ctrl キー + A

2. 選択範囲をコピー

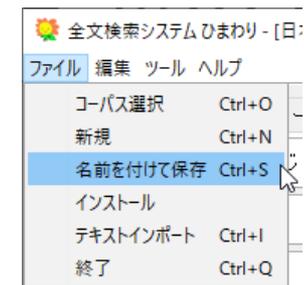
- ▶ 通常のコピー: Ctrl キー + C
- ▶ 列名を含めたコピー: Ctrl キー + Shift キー + C

3. Excel などにペースト



▶ [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



動画の閲覧

講習会用パッケージには、会話ID「T001_009」の動画のみ収録

| 会話ID ^ | 話者ID | 話者ラベル | 性別 | 年齢 |
|----------|----------|---------|----|----------|
| T001_009 | T001_001 | IC03_優香 | 女性 | 30-34歳 ^ |
| T003_001 | T003 | IC01_由美 | 女性 | 35-39歳 |
| T003_001 | T003_002 | IC03_大和 | 男性 | 5-9歳 |
| T003_001 | T003 | IC01_由美 | 女性 | 35-39歳 |
| T003_001 | T003 | IC01_由美 | 女性 | 35-39歳 |
| T003_001 | T003 | IC01_由美 | 女性 | 35-39歳 |
| T003_001 | T003 | IC01_由美 | 女性 | 35-39歳 |

閲覧したい用例の
「性別」か「年齢」列のセルを
ダブルクリック

観察支援システムFishWatchr

「注記連動」をクリックすると、
転記テキストが動画と連動してスクロール

| 番号 | 時間 | 注釈者 | 話者 | ラベル | テキスト | 転記テキスト | 補助情報 |
|----|----------|--------|---------|--------------|-------------------------|--------|------|
| 72 | 00:01:43 | system | IC04_広瀬 | T001_009_... | そうなんすか。 | | |
| 73 | 00:01:43 | system | IC02_義母 | T001_009_... | うん。 | | |
| 74 | 00:01:43 | system | IC02_義母 | T001_009_... | なんか肺ガンとかって出た。 | | |
| 75 | 00:01:45 | system | IC02_義母 | T001_009_... | うん。 | | |
| 76 | 00:01:46 | system | IC02_義母 | T001_009_... | でもまあおかしいなみたいな感じだったんだけど。 | | |
| 77 | 00:01:48 | system | IC03_優香 | T001_009_... | でもこの間の温泉のやつはまってたじゃん。 | | |
| 78 | 00:01:48 | system | IC02_義母 | T001_009_... | うん。 | | |
| 79 | 00:01:50 | system | IC02_義母 | T001_009_... | 出てたね。 | | |
| 80 | 00:01:51 | system | IC03_優香 | T001_009_... | ちょっと。 | | |
| 81 | 00:01:51 | system | IC02_義母 | T001_009_... | そうそう。 | | |

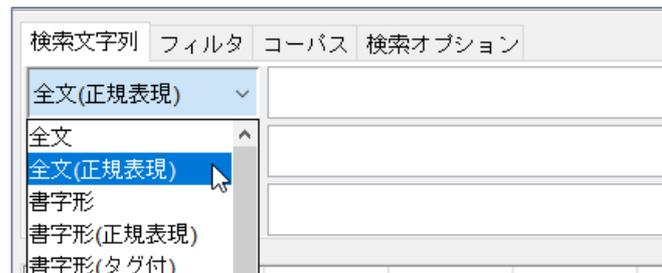
注意

- この後の講習で使うので、再生を停止し、ウィンドウを「最小化」しておいてください
- 個別に起動するので、通常は、閲覧し終わったら終了したほうがよい

検索機能の詳細

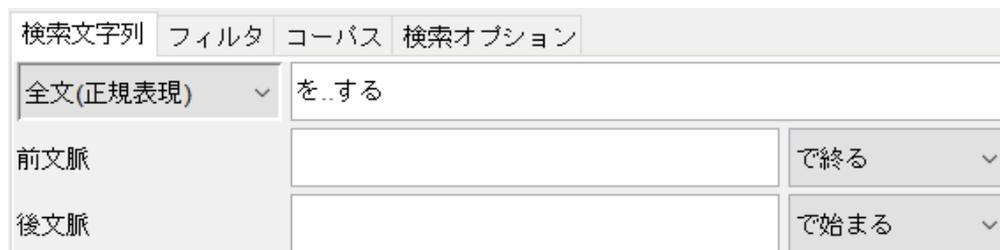
全文(正規表現)

- ▶ 本文検索に正規表現(Java)が利用可能
- ▶ 検索速度は「全文」検索より低速
- ▶ マッチングの範囲は、1発話単位(転記テキスト1行)



A) 「です。」

. (ピリオド)や
カッコは半角!



B) 「じゃ(ない|ねえ)」

C) 雑多な例

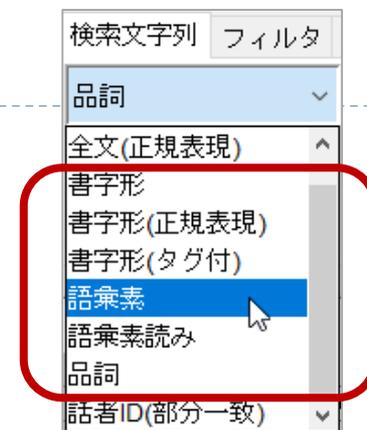
- ▶ 私[がは]
- ▶ です.*\$
- ▶ (..)¥1

| | |
|--------|-----------------------|
| . | ... 任意の1文字 |
| (A B) | ... AもしくはB |
| [がをにへ] | ... 「が」「を」「に」「へ」のいずれか |
| ^ | ... 行頭(この場合、発話単位頭) |
| \$ | ... 行末(この場合、発話単位末) |
| * | ... 直前要素の0個以上の繰り返し |
| + | ... 直前要素の1個以上の繰り返し |
| () | ... マッチした範囲を記録 |
| ¥1 | ... 1個目の記録した要素 |

macOSの場合、「¥」は逆スラッシュ (optionキー+「¥」) を使用

単語(短単位)での検索

- ▶ 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」)
- ▶ 「書字形(正規表現)」だけ、検索文字列の指定方法が異なる
- ▶ 詳細は、[ヘルプ]⇒[『ひまわり』マニュアル]参照



❑ 「国」を含む単語

書字形
書字形(タグ付)
語彙素
語彙素読み
品詞

| 検索文字列 | フィルタ | コーパス | 検索 |
|---------|------|------|----|
| 書字形 | ▼ | 国 | |
| 正規表現(前) | | | |
| 正規表現(後) | | | |

❑ 「国」で始まる単語

| 検索文字列 | フィルタ | コーパス | 検索 |
|---------|------|------|----|
| 書字形 | ▼ | 国 | |
| 正規表現(前) | | ^ | |
| 正規表現(後) | | | |

❑ 単語「国」のみ

| 検索文字列 | フィルタ | コーパス | 検索 |
|---------|------|------|----|
| 書字形 | ▼ | 国 | |
| 正規表現(前) | | ^ | |
| 正規表現(後) | | \$ | |

書字形(正規表現)

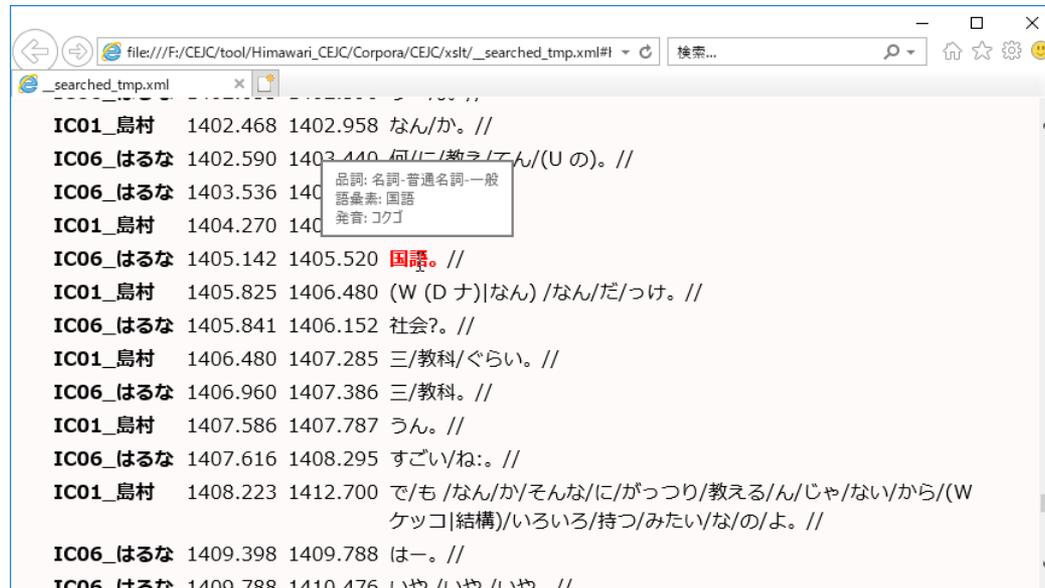
| 検索文字列 | フィルタ | コーパス | 検索 |
|-----------|------|------|----|
| 書字形(正規表現) | ▼ | 国 | |
| 前文脈 | | | |
| 後文脈 | | | |

| 検索文字列 | フィルタ | コーパス | 検索 |
|-----------|------|------|----|
| 書字形(正規表現) | ▼ | ^国 | |
| 前文脈 | | | |
| 後文脈 | | | |

| 検索文字列 | フィルタ | コーパス | 検索 |
|-----------|------|------|----|
| 書字形(正規表現) | ▼ | ^国\$ | |
| 前文脈 | | | |
| 後文脈 | | | |

検索と転記テキストの関係

- ▶ 「全文」「書字形」などでの検索
 - ▶ タグを除外した上で検索
- ▶ 「書字形(タグ付)」での検索
 - ▶ タグを除外しないで検索



| タグあり | タグなし |
|------------|------|
| 国語。 | 国語 |
| 社会?。 | 社会 |
| (Uの)。 | の |
| (Wケッコ 結構) | 結構 |
| (W(Dナ) なん) | なん |

ブラウザ表示されるのは、タグ付きの本文
(ただし、/は単語区切り。転記テキストのタグではなく、表示上の工夫)

「全文」検索では、タグや単語の区切りは無視して検索
(例: 「結構いろいろ」)

練習問題1

- ① 代名詞の「私」を検索
(「語彙素」で検索してみましょう)

- ② ①の検索結果の発音のバリエーションを集計
(検索結果の「発音」列や「キー」列で集計してみましょう)

コーパスの構造 と 基礎情報の集計

基礎情報の集計

▶ 基礎情報

- ▶ 「コーパス中の単語数」
- ▶ 「会話データ中の発話数」



量的な分析を行う際に必要

▶ CEJCの構造

- ▶ 会話データは発話列から構成
- ▶ 発話は単語列から構成

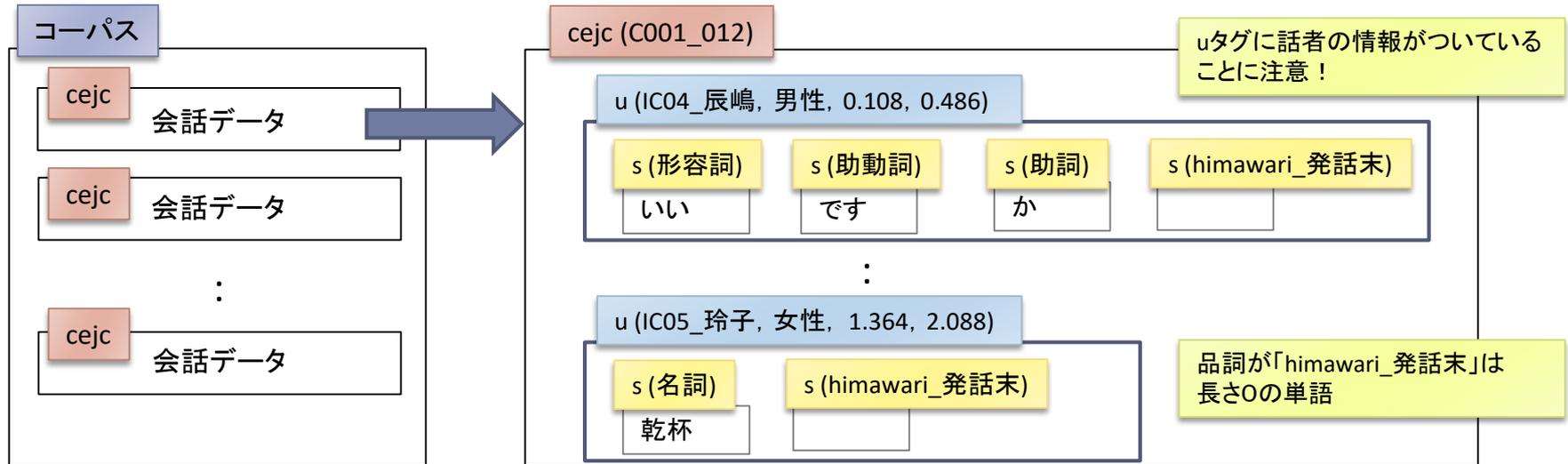
『ひまわり』は、構造を考慮した構成要素の集計が可能

| 話者ラベル | 開始時刻 | 終了時刻 | 転記テキスト |
|---------|-------|-------|------------------------|
| IC04_辰嶋 | 0.108 | 0.486 | いいですか?。// |
| IC05_玲子 | 0.519 | 0.715 | うん。// |
| IC04_辰嶋 | 0.668 | 0.847 | はい。// |
| IC04_辰嶋 | 0.956 | 1.364 | じゃあ。// |
| IC05_玲子 | 0.956 | 1.221 | はい。// |
| IC05_玲子 | 1.364 | 2.088 | 乾(Kバ:イ)杯。// |
| IC05_玲子 | 6.524 | 8.097 | うはうは/みたい/な/感じ/だった/よ。// |
| IC04_辰嶋 | 6.700 | 7.779 | うーん。// |

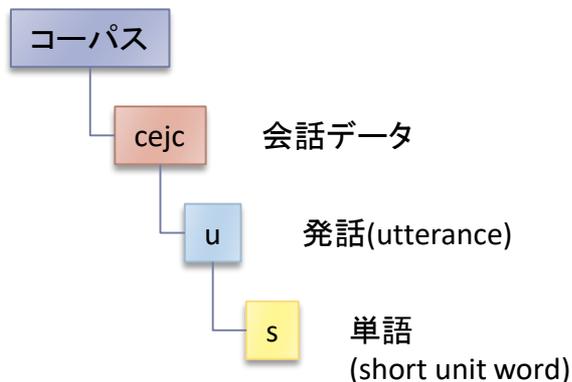
コーパス本体を見たい場合
Corpora/CEJC/corpus.xml

※「秀丸」などのテキストエディタを利用のこと

『ひまわり』用データの全体構造



□ タグの階層構造

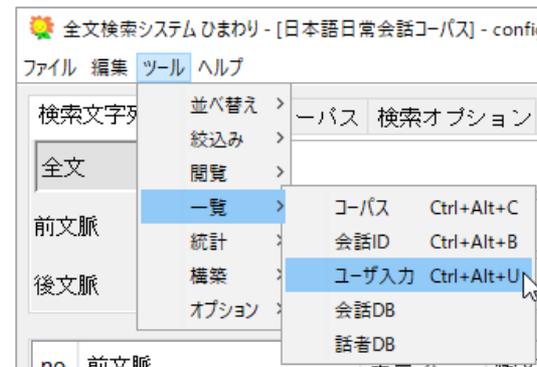


| 話者ラベル | 開始時刻 | 終了時刻 | 転記テキスト |
|---------|-------|-------|--------------------------|
| IC04_辰嶋 | 0.108 | 0.486 | いい/ですか? // |
| IC05_玲子 | 0.519 | 0.715 | うん。 // |
| IC04_辰嶋 | 0.668 | 0.847 | はい。 // |
| IC04_辰嶋 | 0.956 | 1.364 | じゃあ。 // |
| IC05_玲子 | 0.956 | 1.221 | はい。 // |
| IC05_玲子 | 1.364 | 2.088 | 乾(Kパ:イ 杯)。 // |
| IC05_玲子 | 6.524 | 8.097 | うはうは/みたい/な/感じ/だっ/た/よ。 // |
| IC04_辰嶋 | 6.700 | 7.779 | うーん。 // |

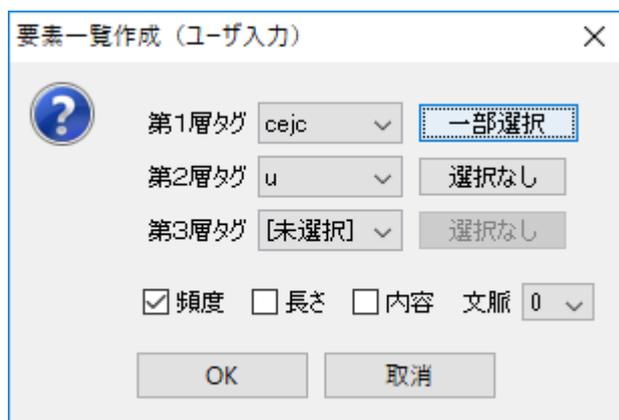
タグの集計

▶ [ツール] ⇒ [一覧] ⇒ ユーザ入力

- ▶ タグの階層構造を利用しつつ、タグの数や属性を集計する
- ▶ 例1: 発話数
- ▶ 例2: 語彙表



□ 各会話に含まれる発話数



- ▶ 頻度: 指定したタグの頻度
- ▶ 長さ: マークアップされている文字列の長さ (空白やXMLタグは除く)
- ▶ 内容: マークアップされている文字列
- ▶ 文脈: 後続する同種の要素の属性をn個表示 (単語の場合n+1 gramになる)

タグの集計の例1

□ 会話一覧

要素一覧作成 (ユーザ入力)

第1層タグ **cejc** 一部選択

第2層タグ [未選択] 選択なし

第3層タグ [未選択] 選択なし

頻度 長さ 内容 文脈 0

OK 取消

要素一覧作成 (ユーザ入力)

会話ID

{path}

話者間の関係性

すべて選択

OK 取消

[1] ...

ファイル 編集 ツール

| cejc/@会話ID ^ | 頻度 |
|--------------|----|
| T001_009 | 1 |
| T003_001 | 1 |
| T004_001 | 1 |
| T004_013 | 1 |
| T010_004 | 1 |

総数(延べ): 5, 異なり: 5

□ 発言者一覧

要素一覧作成 (ユーザ入力)

第1層タグ **u** 一部選択

第2層タグ [未選択] 選択なし

第3層タグ [未選択] 選択なし

頻度 長さ 内容 文脈 0

OK 取消

要素一覧作成 (ユーザ入力)

年齢

出身地

コメント

終了時刻

{n}

職業

{residence}

性別

話者ラベル

話者ID

開始時刻

すべて選択

OK 取消

[2] 一覧: u

ファイル 編集 ツール

| u/@性別 ^ | u/@話者ID | 頻度 |
|---------|----------|-------|
| 女性 | T003 | 147 ^ |
| 女性 | T004_031 | 205 |
| 女性 | T004 | 971 |
| 女性 | T004_032 | 495 |
| 女性 | T004_033 | 444 |
| 女性 | T004_030 | 302 |
| 女性 | T001_001 | 136 |
| 女性 | T001_002 | 422 |
| 女性 | T003_002 | 181 |
| 男性 | T003_001 | 78 v |

総数(延べ): 6027, 異なり: 16

タグの集計の例2

□ 語彙表

要素一覧作成 (ユーザ入力)

第1層タグ 一部選択

第2層タグ [未選択] 選択なし

第3層タグ [未選択] 選択なし

頻度 長さ 内容 文脈 0

OK 取消

要素一覧作成 (ユーザ入力)

仮名

活用型

語彙素読み

活用形

発音形出現形

語彙素

{n}

品詞

発音

タグ付き書字形

書字形

すべて選択

OK 取消

□ 各会話に含まれる発話数

要素一覧作成 (ユーザ入力)

第1層タグ 一部選択

第2層タグ 選択なし

第3層タグ [未選択] 選択なし

頻度 長さ 内容 文脈 0

OK 取消

要素一覧作成 (ユーザ入力)

盗話ID

{path}

話者間の関係性

すべて選択

OK 取消

タグの集計時のフィルタ

▶ タグ集計時にもフィルタをかけることが可能

▶ 例：品詞名に「himawari」を含む単語は集計から除外

| | | | |
|-------|----------|-------|---------|
| 検索文字列 | フィルタ | コーパス | 検索オプション |
| 品詞 | himawari | を含まない | |
| 話者ID | | で始まる | |
| 話者ラベル | | で始まる | |

- フィルタの対象は、集計する項目に必ず含めること
- 左例の場合、「要素一覧作成」ダイアログで、sの「品詞」属性を含める

| s/@品詞 | 頻度 |
|--------------|--------|
| himawari 発話末 | 147867 |
| 助動詞 | 76091 |
| 感動詞-一般 | 57729 |
| 名詞-普通名詞-... | 53020 |
| 助詞-格助詞 | 51714 |
| 助詞-終助詞 | 46466 |
| 副詞 | 37297 |
| 動詞-一般 | 34715 |
| 動詞-非自立可能 | 29510 |
| 助詞-接続助詞 | 27311 |

総数(延べ)：770169, 異なり：49

フィルタをかけると...



| s/@品詞 | 頻度 |
|-------------|-------|
| 助動詞 | 76091 |
| 感動詞-一般 | 57729 |
| 名詞-普通名詞-... | 53020 |
| 助詞-格助詞 | 51714 |
| 助詞-終助詞 | 46466 |
| 副詞 | 37297 |
| 動詞-一般 | 34715 |
| 動詞-非自立可能 | 29510 |
| 助詞-接続助詞 | 27311 |
| 代名詞 | 23579 |

総数(延べ)：622302, 異なり：48

練習問題2

- ① 会話データごとの単語数を求める
(cejcとsタグを使いましょう)
- ② 会話データごとに、品詞が「himawari_発話末」の語数を求める
(フィルタで「品詞」が「himawari」で始まるもの限定しましょう)

動画の閲覧・アノテーション



FishWatchr の紹介

▶ 特徴

- ▶ 『ひまわり』で検索した場所の映像を見るのに便利
- ▶ 転記テキストと映像を同期させて見ることができる
- ▶ 簡単に注釈付け(アノテーション)できる

▶ 簡易的な利用を想定

⇒ 専門的なアノテーションや分析を行う場合は, ELAN, Praatを利用のこと

- ▶ 元々, 大学におけるディスカッション練習を, 学生がリアルタイムに観察するためのシステムとして開発しました

動画の再生位置の指定

▶ 三つの方法で指定可能

プロット画面に何も表示されない場合、「リセット」してください

①「番号」列をダブルクリック

②ダブルクリック
(横軸は時間)

③クリック

現在の再生位置

この色の範囲は、
現在位置の前後
10秒

| 番号 | 時間 | 注釈者 | 話者 | ラベル | セット | 転記テキスト | 補助情報 |
|----|----------|--------|---------|-----|--------------|--------------------------|------|
| 72 | 00:01:43 | system | IC04_広瀬 | | T001_009_... | そうなんすか。 | |
| 73 | 00:01:43 | system | IC02_義母 | | T001_009_... | うん。 | |
| 74 | 00:01:43 | system | IC02_義母 | | T001_009_... | なんか肺ガンとかって出た。 | |
| 75 | 00:01:45 | system | IC02_義母 | | T001_009_... | うん。 | |
| 76 | 00:01:46 | system | IC02_義母 | | T001_009_... | でもどうもおかしいなみたいな感じだったんだけど。 | |
| 77 | 00:01:48 | system | IC03_優香 | | T001_009_... | でもこの間の温泉のやつお泊ってたじゃん。 | |
| 78 | 00:01:48 | system | IC02_義母 | | T001_009_... | うん。 | |
| 79 | 00:01:50 | system | IC02_義母 | | T001_009_... | 出たたね。 | |
| 80 | 00:01:51 | system | IC03_優香 | | T001_009_... | ちょっと。 | |
| 81 | 00:01:51 | system | IC02_義母 | | T001_009_... | そうそう。 | |

動画の再生(微調整)

- 再生速度の変更
- 前後のコメント(発話)へのジャンプ

[resources/FishWatchr/xml/T001_009.fw.xml] - FishWatchr

ファイル コントロール 注釈 分析 オプション ヘルプ

再生・一時停止 (■) Ctrl-P
スキップ (◀) Ctrl-K
スキップ (▶) Ctrl-L
停止 (■) Ctrl+Shift-H
録音録画 (●) Ctrl+Shift-K
前のコメント Ctrl+Shift-L
次のコメント Ctrl+Shift-L
再生位置の注釈表示 Ctrl-J
再生速度+ Ctrl-ピリオド
再生速度- Ctrl-カンマ
再生速度リセット Ctrl-スラッシュ

再生開始時の時間補正
2秒前から再生 (デフォルト)
[オプション]⇒[再生位置の補正]

再生のコントロールは
すべてキーボードからも可能

| 番号 | 時間 | 発話者 | 話者 | ラベル | セット | 転記テキスト | 補助情報 |
|-----|----------|---------|---------|--------------|-----|--|------|
| | | IC04_広瀬 | IC04_広瀬 | T001_009_... | | でもなん(W(Dブ)分類)分類でゆうと一緒にです。 | |
| | | IC02_義母 | IC02_義母 | T001_009_... | | (Wモ(もう)とにかく。 | |
| | | IC02_義母 | IC02_義母 | T001_009_... | | 分類?。 | |
| | | IC02_義母 | IC02_義母 | T001_009_... | | (Fあの)ねなんだっけーチャンのアナウンサーなんだっけ。 | |
| 132 | 00:02:39 | system | IC03_優香 | T001_009_... | | 有働さん。 | |
| 133 | 00:02:40 | system | IC02_義母 | T001_009_... | | 有働さん 年がら年中ゆわれる。 | |
| 134 | 00:02:42 | system | IC04_広瀬 | T001_009_... | | (T有働さんて誰だっけ)。 | |
| 135 | 00:02:43 | system | IC03_優香 | T001_009_... | | いのっちと出てる人。@いのっち=ジャニーズグループV6メンバー・井ノ原快彦のこと | |
| 136 | 00:02:45 | system | IC04_広瀬 | T001_009_... | | あー。 | |
| 137 | 00:02:45 | system | IC02_義母 | T001_009_... | | うーん。 | |

ラベル 1 [F1] ラベル 2 [F2]

発話のプロット画面

『日本語日常会話コーパス』の場合は、「表示」を「話者」に変更するのが便利

ドラッグすると
拡大表示
(戻すときは「リセット」)

発話の
ヒストグラム

| 番号 | 時間 | 注釈者 | 話者 | ラベル | セット | 転記テキスト | 補助情報 |
|-----|----------|--------|---------|-----|--------------|---|------|
| 542 | 00:13:13 | system | IC02_義母 | | T001_009_... | (Dイ) なかなか行かれないよね。 | |
| 543 | 00:13:15 | system | IC04_広瀬 | | T001_009_... | (T(Uそう)すね)。 | |
| 544 | 00:13:15 | system | IC02_義母 | | T001_009_... | うん。 | |
| 545 | 00:13:16 | system | IC04_広瀬 | | T001_009_... | 東京東京の地震発生率ってなん ここ何年で 何年かで七十パーセントでしたっけ。 | |
| 546 | 00:13:23 | system | IC02_義母 | | T001_009_... | 知らないけど。 | |
| 547 | 00:13:24 | system | IC02_義母 | | T001_009_... | とにかくわたし(私たち)小さいころから もう地震は来ってゆうので(Fあの)小(W... | |
| 548 | 00:13:35 | system | IC02_義母 | | T001_009_... | でうん 年に二三次? ちゃんと防災訓練があった。 | |
| 549 | 00:13:40 | system | IC02_義母 | | T001_009_... | うん。 | |
| 550 | 00:13:40 | system | IC02_義母 | | T001_009_... | もうほんとにだから 五十(T五年ぐらい前)四五年前からあったから。 | |
| 551 | 00:13:45 | system | IC04_広瀬 | | T001_009_... | (Tはー)。 | |

発話の絞り込み(1)

▶ 列名の部分を右クリックすると絞り込み

[resources/FishWatchr/xml/T001_009.fw.xml] - FishWatchr
ファイル コントロール 注釈 分析 オプション ヘルプ

全体 詳細
表示 ラベル フィルタ連動 リセット < >

ラベル1
ラベル2

00:10:00 00:19:19

| 番号 | 時間 | 注釈者 | 話者 | ラベル | セット | 転記テキスト | 補助情報 |
|----|----------|--------|---------|------------|-------------|--------------------------|------|
| 72 | 00:01:43 | system | IC04_広瀬 | [検索文字列の指定] | 01_009... | そうなんすか。 | |
| 73 | 00:01:43 | system | IC02_義母 | IC02_義母 | 01_009... | うん。 | |
| 74 | 00:01:43 | system | IC02_義母 | IC02_義母 | 01_009... | なんか肺ガンとかって出た。 | |
| 75 | 00:01:45 | system | IC02_義母 | IC03_優香 | 01_009... | うん。 | |
| 76 | 00:01:46 | system | IC02_義母 | IC04_広瀬 | 01_009... | でもどうもおかしいなみたいな感じだったんだけど。 | |
| 77 | 00:01:48 | system | IC03_優香 | | T001_009... | でもこの間の温泉のやつは治ってたじゃん。 | |
| 78 | 00:01:48 | system | IC02_義母 | | | | |
| 79 | 00:01:50 | system | IC02_義母 | | | | |
| 80 | 00:01:51 | system | IC03_優香 | | | | |
| 81 | 00:01:51 | system | IC02_義母 | | T001_009... | そうそう。 | |

ラベル 1 2

注記連動

選択メニュー

絞り込みを解除するには、選択メニューから「全フィルタ解除」を実行

発話の絞り込み(2)

▶ 文字列を指定しての絞り込み

[resources/FishWatchr/xml/T001_009.fw.xml] - FishWatchr
ファイル コントロール 注釈 分析 オプション ヘルプ

全体 詳細
表示 話者 フィルタ連動 リセット < >

IC02_義
IC03_優
IC04_広

検索文字列の例

- 例1: 「うん」
- 例2: 「L」 ... 笑いを含む発話
- 例3: 「¥?」 ... 疑問上昇調の発話

※macの場合¥を半角の\に
※¥を入れるのは、?が正規表現のメタ文字のため

00:13:27 00:19:19

| 番号 | 時間 | 注釈者 | 話者 | ラベル | セット | 転記テキスト | 補助情報 |
|-----|----------|--------|---------|-----|-------------|------------------|------|
| 542 | 00:13:13 | system | IC02_義母 | | T001_009... | (Dイ) なかなか行かないよね。 | |
| 543 | 00:13:15 | system | IC04_広瀬 | | | | |
| 544 | 00:13:15 | system | IC02_義母 | | | | |
| 545 | 00:13:16 | system | IC04_広瀬 | | | | |
| 546 | 00:13:23 | system | IC02_義母 | | | | |
| 547 | 00:13:24 | system | IC02_義母 | | | | |
| 548 | 00:13:35 | system | IC02_義母 | | | | |
| 549 | 00:13:40 | system | IC02_義母 | | | | |
| 550 | 00:13:40 | system | IC02_義母 | | | | |
| 551 | 00:13:45 | system | IC04_広瀬 | | | | |

絞り込み条件の指定

[検索文字列の指定]

(Dア) アスミの木のねんどうけ(アア) #####
(Dイ) なかなか行かないよね。
(Dイ) 効く人で効くタイミングの時もあれば 効かないタイミングの時もあるし。
(Dオ) え。
(Dオー) お手伝いもね: うんわたしはいい。
(Dハ) (Mハ) リカとか(Mハ) ルタとかなんとかってゆう そうゆうあれの(U こと)。
(Dハ)。

OK 取消

アノテーション(追加と削除)

データの保存

- 変更すると, 終了時に自動保存
- ただし, 『ひまわり』やdataフォルダのデータには影響なし

The screenshot shows the FishWatchr application window. On the left, there are tabs for '全体' (Overall) and '詳細' (Details). Below them are controls for '表示' (Display) and 'フィルタ連動' (Filter联动). The main area is split: the top left shows a list of annotations with columns for '注釈者' (Annotator), '話者' (Speaker), 'ラベル' (Label), and 'セット' (Set). The top right shows a video feed. The bottom part of the window features a timeline and a table of annotations. A red box highlights the '補助情報' (Auxiliary Information) column in the table. Another red box highlights two buttons labeled 'ラベル 1' and 'ラベル 2' at the bottom of the interface.

| 番号 | 時間 | 注釈者 | 話者 | ラベル | セット | 転記テキスト | 補助情報 |
|-----|----------|--------------|---------|------|---------------|--|------|
| 543 | 00:13:15 | system | IC04_広瀬 | | T001_009... | (T(Uそう)すね:) | |
| 544 | 00:13:15 | system | IC02_義母 | | T001_009... | うん。 | |
| 545 | 00:13:16 | system | IC04_広瀬 | | T001_009... | 東京東京の地震発生率って:なんここ何年で何年かで七十パーセントでしたっけ。 | |
| 546 | 00:13:23 | system | IC02_義母 | | T001_009... | 知らないけど。 | |
| 547 | 00:13:24 | system | IC02_義母 | | T001_009... | とにかくわたし(私たち)小さいころから もう地震は来るってゆうので(Fあの)小(W... | |
| 800 | 00:13:27 | masaya_in... | 不特定 | ラベル1 | T001_009.f... | | |
| 548 | 00:13:35 | system | IC02_義母 | | T001_009... | でうん年に二三回?ちゃんと防災訓練があった。 | |
| 549 | 00:13:40 | system | IC02_義母 | | T001_009... | うん。 | |
| 550 | 00:13:40 | system | IC02_義母 | | T001_009... | もうほんとにだから 五十:(T五年ぐらい前)四五年前からあったから。 | |
| 551 | 00:13:45 | system | IC04_広瀬 | | T001_009... | (Tほー) | |

「注釈者」「話者」「ラベル」「セット」
「補助情報」列は変更可

- コメントは「補助情報」列に記入
- 直接入力できるが, タブルクリックすると, 入力用のフォームから入力可

- 再生位置に1行追加
(一つのアノテーション)
- 削除は, 削除したい行で,
右クリック⇒[行の削除]

アノテーションボタン

アノテーション(ラベルの変更)

▶ ラベルは自由に定義可

「マーク色」はアノテーション・プロット時の色

| 注記名 | マーク色 |
|--------|------|
| うなずき | 赤 |
| 疑問 | 黄 |
| 躊躇 | 緑 |
| ジェスチャー | 青 |
| | |
| | |
| | |
| | |
| | |

| 番号 | 時間 | 注釈者 | 話者 | ラベル | セット | 補助情報 |
|-----|----------|--------------|---------|------|---------------|---|
| 543 | 00:13:15 | system | IC04_広瀬 | | T001_009... | |
| 544 | 00:13:15 | system | IC02_義母 | | T001_009... | |
| 545 | 00:13:16 | system | IC04_広瀬 | | T001_009... | 東の東の地震がエキスプ...の...で七十パーセントでしたっけ。 |
| 546 | 00:13:23 | system | IC02_義母 | | T001_009... | 知らないけど。 |
| 547 | 00:13:24 | system | IC02_義母 | | T001_009... | とにかくわたし(私たち)小さいころからもう地震は来るってゆうので(Fあの)小(W... |
| 800 | 00:13:27 | masaya_in... | 不特定 | うなずき | T001_009.f... | |
| 548 | 00:13:35 | system | IC02_義母 | | T001_009... | でうん年に二三回?ちゃんと防災訓練があった。 |
| 549 | 00:13:40 | system | IC02_義母 | | T001_009... | うん。 |
| 550 | 00:13:40 | system | IC02_義母 | | T001_009... | もなほんと前から五中(五年くらい前)四十年前からあったから。 |
| 551 | 00:13:45 | system | IC04_広瀬 | | T001_009... | (T1一)。 |

うなずき [1] 疑問 [2] 躊躇 [3] ジェスチャー [4]

ただし、ラベルの設定は個々のファイルごと

アノテーション結果の利用

- ▶ 講習会用のパッケージからFishWatchrを単独で起動するのは難しいので、独立してインストールして、利用するのがお薦め
 - ▶ [ヘルプ] ⇒ [マニュアル]
- ▶ アノテーション結果ファイルの所在(※移動しないこと)
 - ▶ tool/Himawari_CEJC/resources/FishWatchr/xml
 - ▶ FishWatchrにドラッグ & ドロップすると、直接閲覧できる
- ▶ バックアップファイル
 - ▶ tool/Himawari_CEJC/resources/FishWatchr/xml/BAK
 - ▶ 保存前のファイルはすべてバックアップされる
 - ▶ 元に戻したい場合は、一番古いファイルを使えばよい

おわりに

- ▶ 全文検索システム『ひまわり』を使った『日本語日常会話コーパス』の利用方法を紹介
 - ▶ 検索機能(全文検索・単語検索)
 - ▶ コーパスの構造を利用した基礎データの集計方法
 - ▶ FishWatchrによる動画の再生・簡単なアノテーション
- ▶ さらに詳しく知るには
 - ▶ [ヘルプ]⇒[マニュアル]
 - ▶ 資料末の参考資料を参照

参考資料

- ▶ [全文検索システム『ひまわり』](http://www2.ninjal.ac.jp/lrc/index.php?himawari)
(<http://www2.ninjal.ac.jp/lrc/index.php?himawari>)
- ▶ [ヘルプ]⇒[『ひまわり』マニュアル]
- ▶ 観察支援システムFishWatchr
(<http://www2.ninjal.ac.jp/lrc/index.php?fw>)

- ▶ 正規表現
 - ▶ [Java Pattern クラス](https://docs.oracle.com/javase/jp/8/docs/api/java/util/regex/Pattern.html) (『ひまわり』で利用できる正規表現の仕様)
(<https://docs.oracle.com/javase/jp/8/docs/api/java/util/regex/Pattern.html>)
 - ▶ [「Java正規表現の使い方」](http://www.javadrive.jp/regex/)
(<http://www.javadrive.jp/regex/>)

練習問題1の解答例

- ① 代名詞の「私」を検索
(「語彙素」で検索してみましょう)

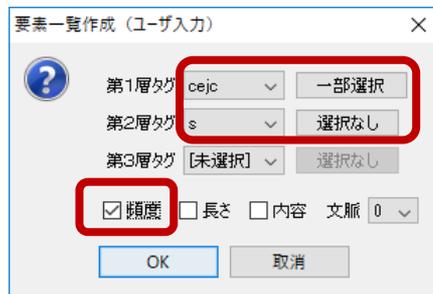
| | | | |
|---------|------|------|---------|
| 検索文字列 | フィルタ | コーパス | 検索オプション |
| 語彙素 | 私 | | |
| 正規表現(前) | ^ | 正規表現 | ▼ |
| 正規表現(後) | \$ | 正規表現 | ▼ |

- ② 発音のバリエーションを集計
(検索結果の「発音」列や「キー」列で集計してみましょう)

| 素読み | 発音形出... | 発音 | 書字形 | タグ付き... | 語彙素-2 | 語 |
|-----|---------|------|-----------|---------|-------|---|
| クシ | アタクシ | アタクシ | あたくし | あたくし | から | |
| クシ | アタクシ | アタクシ | コピー | あたくし | | も |
| シ | アタシ | アタシ | コピー(列名含む) | あたし。 | | |
| シ | アタシ | アタシ | 全選択 | あたし。 | の | |
| シ | アタシ | アタシ | 置換 | あたし。 | うん | |
| シ | アタシ | アタシ | 置換 | あたし。 | よ | |
| シ | アタシ | アタシ | フィルタ | あたし。 | もの | |
| シ | アタシ | アタシ | 統計 | あたし。 | よ | |
| シ | アタシ | アタシ | あたし | あたし？ | | 此 |

練習問題2の解答例

① 会話データごとの単語数を求める

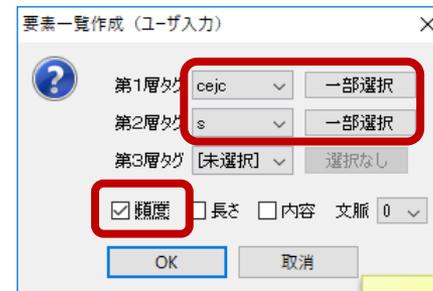


cejcの属性「会話ID」チェック

| cejc/@会... | 頻度 |
|------------|-------|
| C001_001 | 11929 |
| C001_002 | 4659 |
| C001_005 | 2513 |
| C001_007 | 6873 |
| C001_012 | 8430 |
| C002_003 | 925 |
| C002_004 | 3396 |
| C002_006a | 3669 |

総数(延べ): 770169, 異なり: 1...

② 会話データごとに、品詞が「himawari_発話末」の語数を求める



- cejcの「会話ID」属性チェック
- sの「品詞」属性チェック