「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

- 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

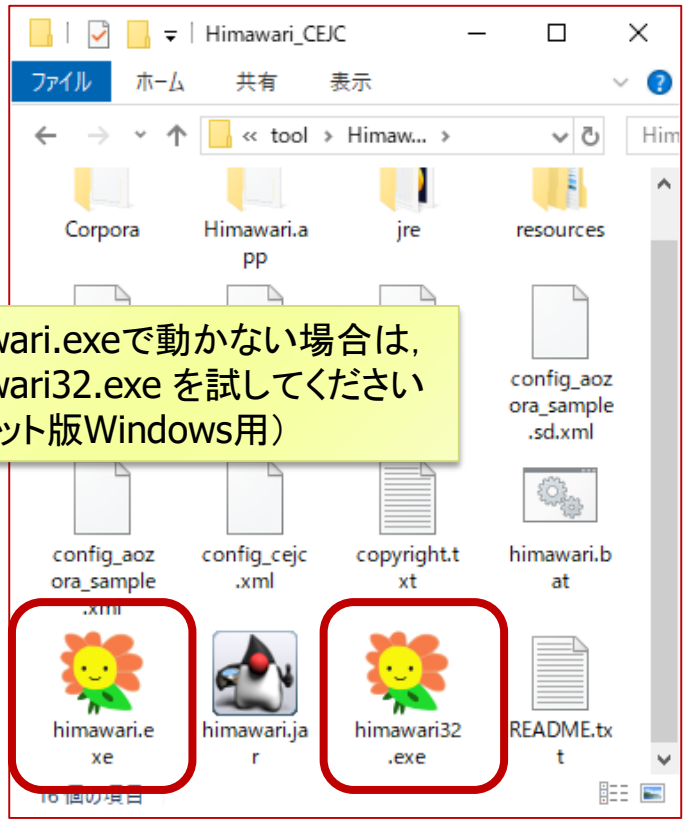
Ver.1.6 ⇒　統計的な分析に必要なデータの収集支援機能を強化
（例：総文字数，総単語数）

『ひまわり』の基本的な使い方
（CEJC編）

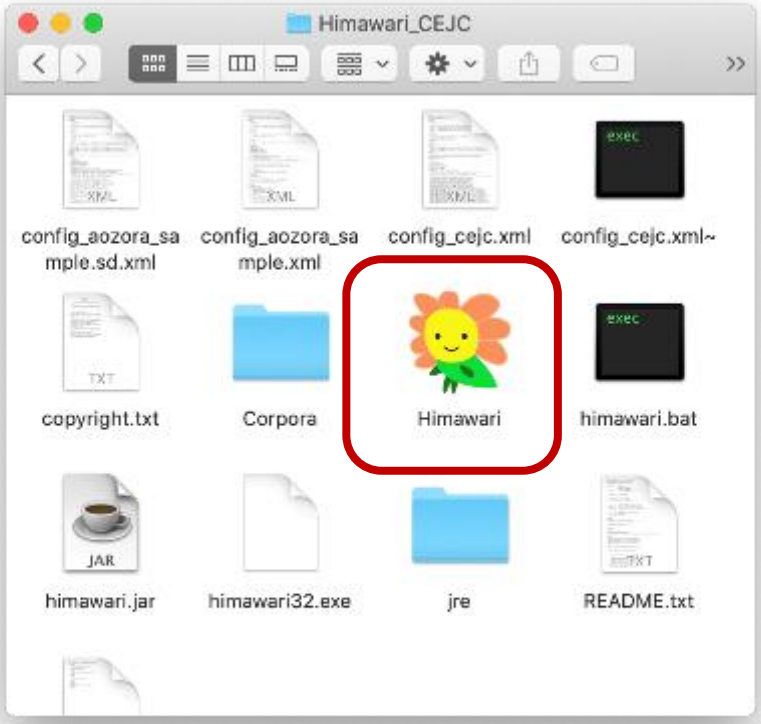
『ひまわり』を起動する

- tool ⇒ Himawari_CEJC フォルダ

❑ Windowsの場合

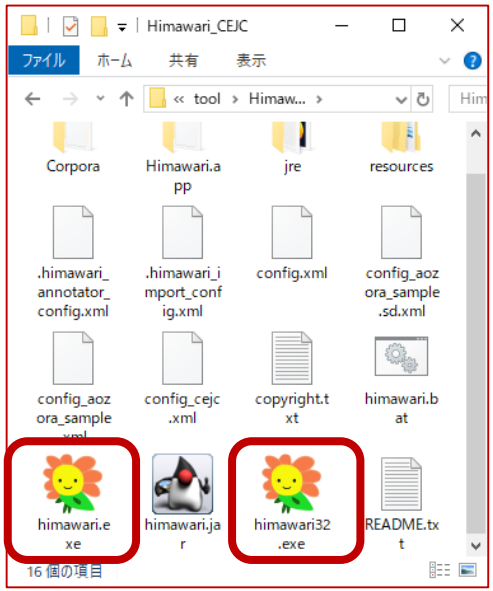


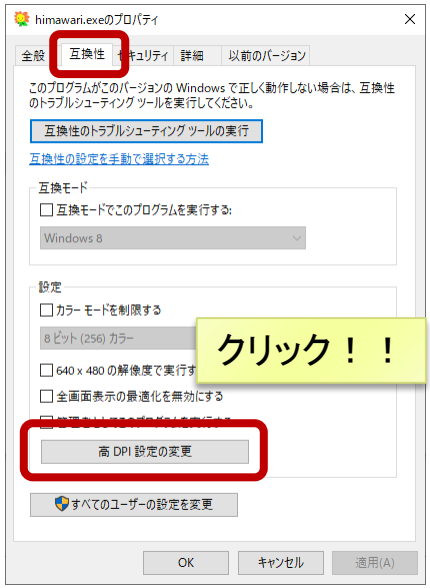
❑ macOSの場合



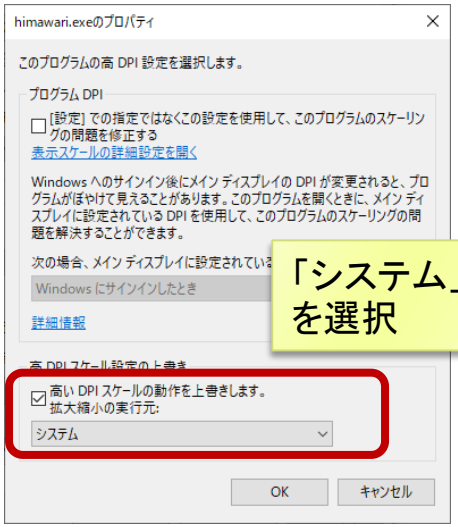
トラブルシューティング（2）

- Windowsで『ひまわり』の文字が小さすぎる





クリック！！

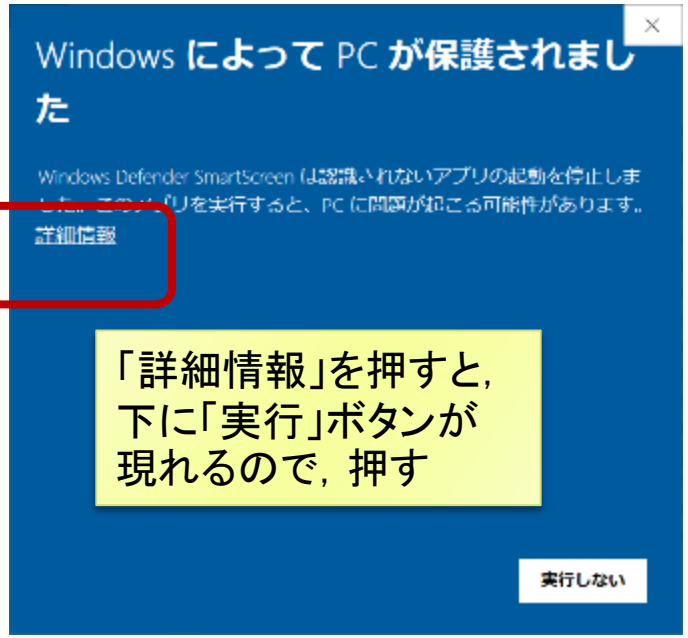


「システム」を選択

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

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

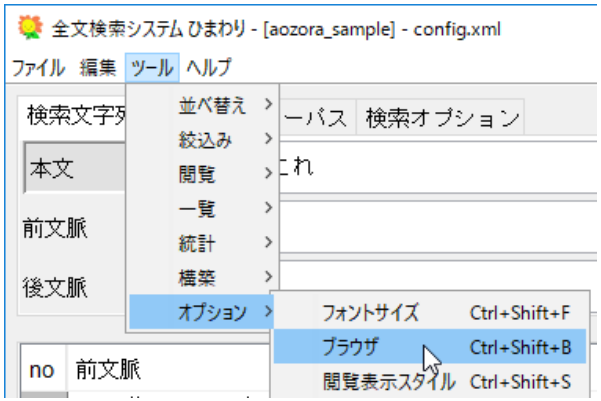

「検索文字列」欄では
右クリックで履歴表示
全文検索システム ひまわり - [日本語日常会話コーパス] - config.xml
ファイル 編集 ツール ヘルプ
検索文字列
フィルタ
コーパス
検索オプション
検索文字列
検索の実行
全文
じゃん
前文脈
で終る
後文脈
で始まる
検索
字体変換
クリア
no
前文脈
キー
後文脈
会話ID
話者ID
話者ラベル
性別
1 ナはあの小ささがいい じゃん ◇ あの細長さでさ T010_004 T010 IC01_徹 男性
2 だよ キャッチする人 じゃん ◇ お前超怖えから T010_004 T010_004 IC02_大場 男性
3 器吹い 腹式やってん 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る 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
検索結果
検索総数:71
検索総数
途中経過の表示

転記テキストの閲覧

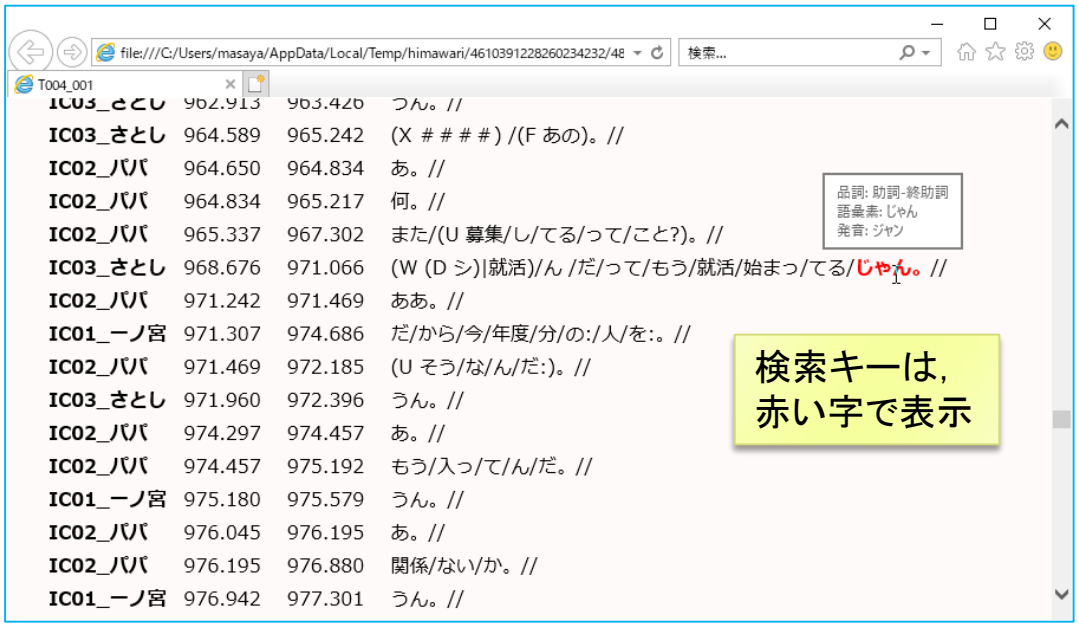
no	前文脈	キー	後文脈	会話ID	話者ID	話者ラベル	性別
1	ナはあの小ささがいい	じゃん	◇ あの細長さでさ	T010_004	T010	IC01_徹	男性
2	だよ キャッチする人	じゃん	◇ お前超怖えから	T010_004	T010_004	IC02_大場	男性
3	器吹い 腹式やってん	じゃん	◇ やってるにも関	T010_004	T		
4	てもう就活始まってる	じゃん	ああ だから今年度	T004_001	T		
5	ついてね あれもあん	じゃん	あのタッチするやつ	T010_004	T		
6	だからカットしに来る	じゃん	あー はいはいはい	T010_004	T010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



検索キーは,
赤い字で表示

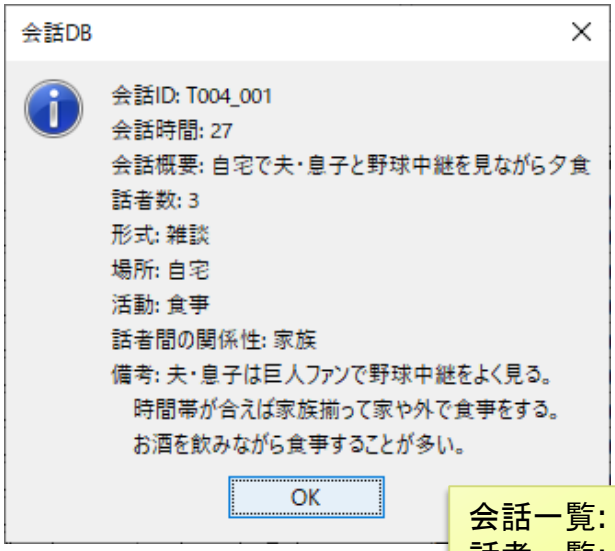
話者，会話情報の閲覧

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	てもう就活始まってる	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	T0		
6	だからカットしに来る	じゃん	あー はいはいはい	T010_004	T0		
7	ん うん ばちーなる	じゃん	あー 俺痛いじゃん	T010_004	T0		

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る	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

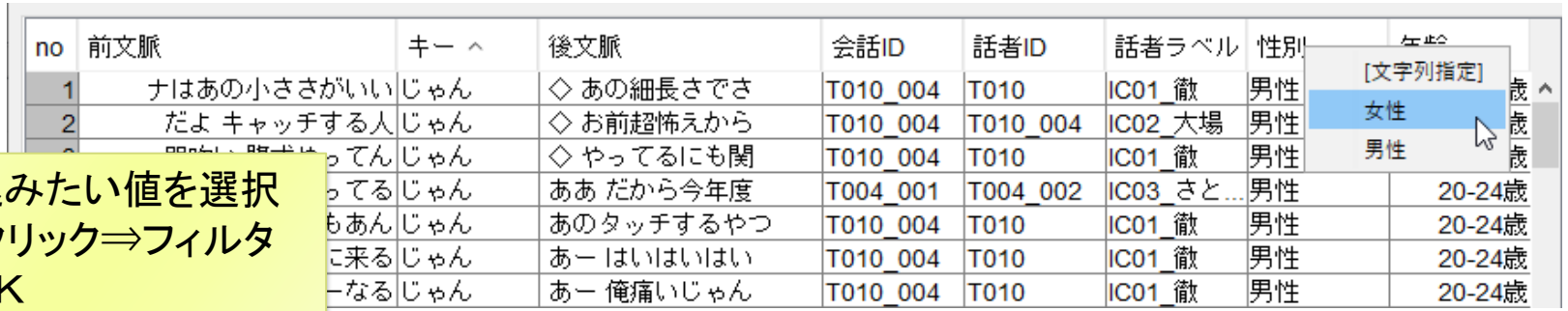
検索結果の絞り込み

- 検索時に指定

「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

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

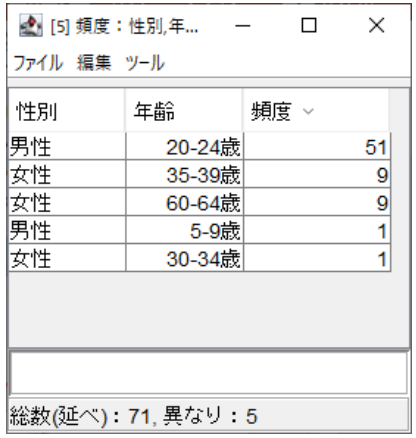
1. 集計したい列を選択

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る	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る	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

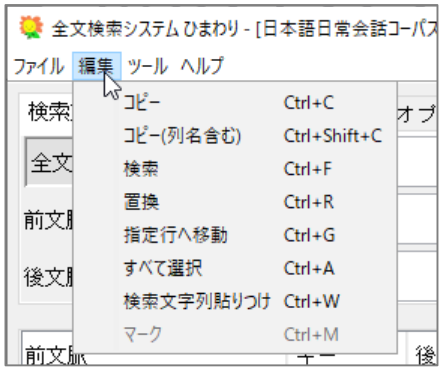
離れた列の選択
- WindowsはCtrlキー
- macOSはcommandキー

2. 右クリック⇒「統計」

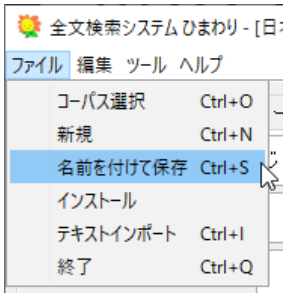


結果のエクスポート

- 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

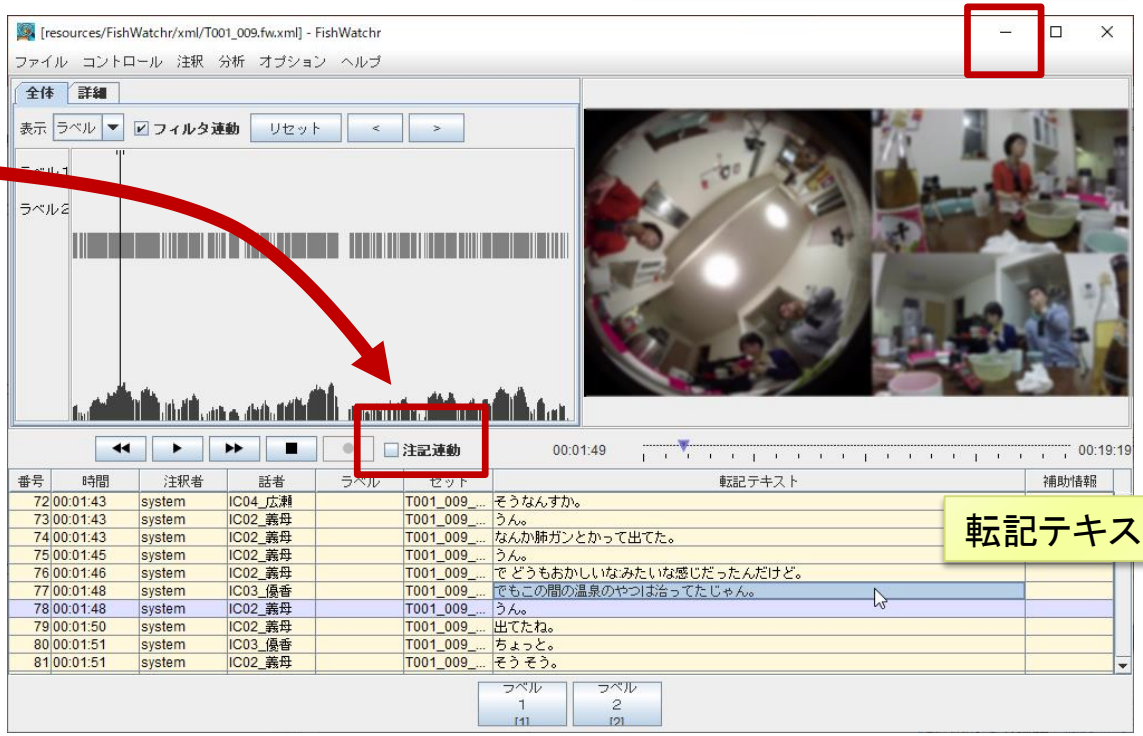
講習会用パッケージには，会話ID「T001_009」の動画のみ収録

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



転記テキスト

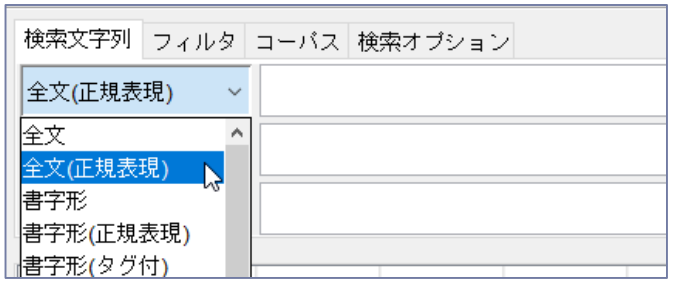
注意

- 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）

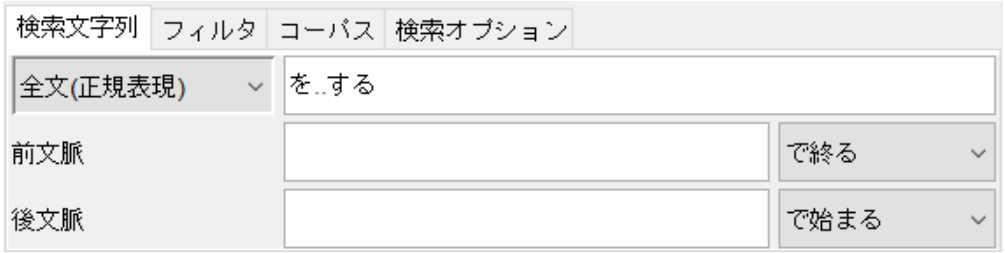
- 本文検索に正規表現（Java）が利用可能
- 検索速度は「全文」検索より低速
- マッチングの範囲は，１発話単位
（転記テキスト1行）



A) 「です.」

. （ピリオド）や
カッコは半角！

B) 「じゃ(ない|ねえ)」

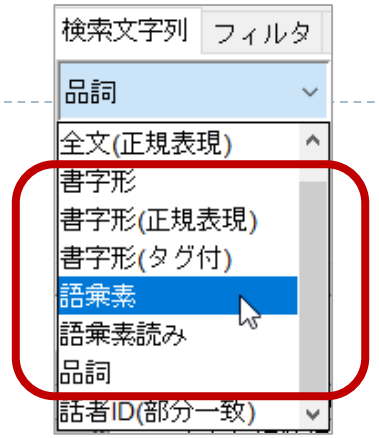


C) 雑多な例
- 私[がは]
- です.*$
- (..)¥1

.	… 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



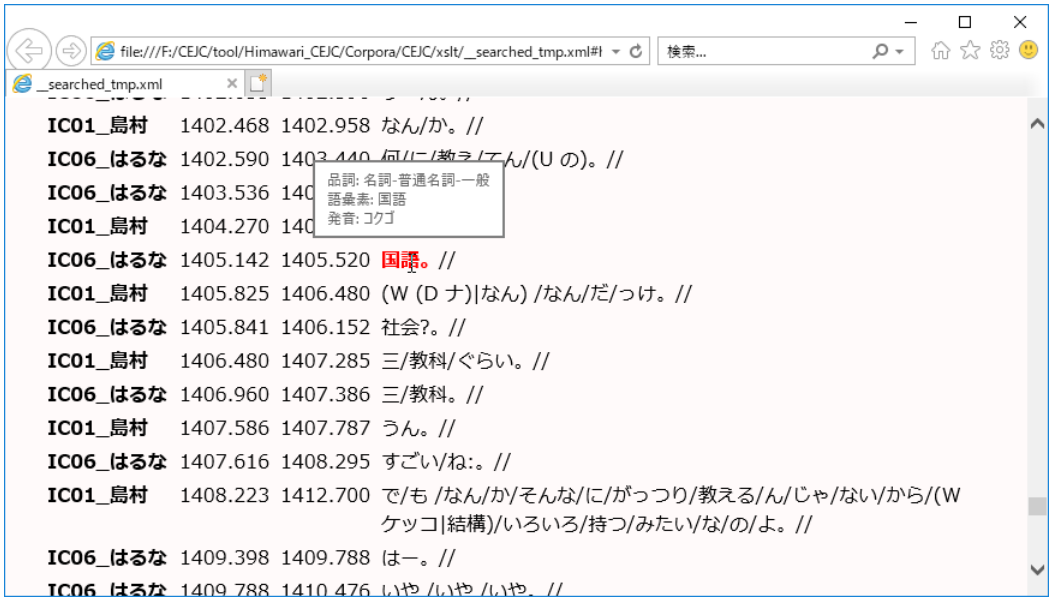
- マッチングの範囲は単語(短単位)
- 単位をまたいだ検索はできない
- 前後2単語の語彙素も表示(例:「語彙素1」「語彙素-1」)
- 「書字形(正規表現)」だけ，検索文字列の指定方法が異なる
- 詳細は，[ヘルプ]⇒[『ひまわり』マニュアル]参照

	❑ 「国」を含む単語	❑ 「国」で始まる単語	❑ 単語「国」のみ
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)	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)	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

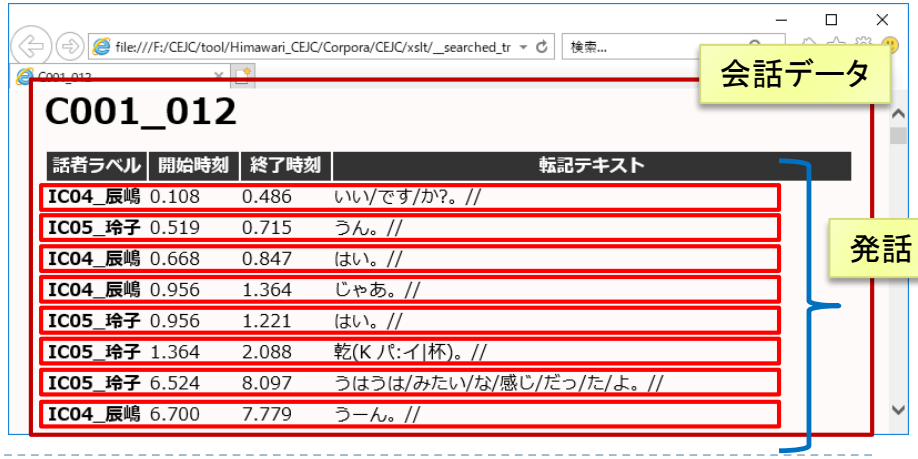
基礎情報の集計

- 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

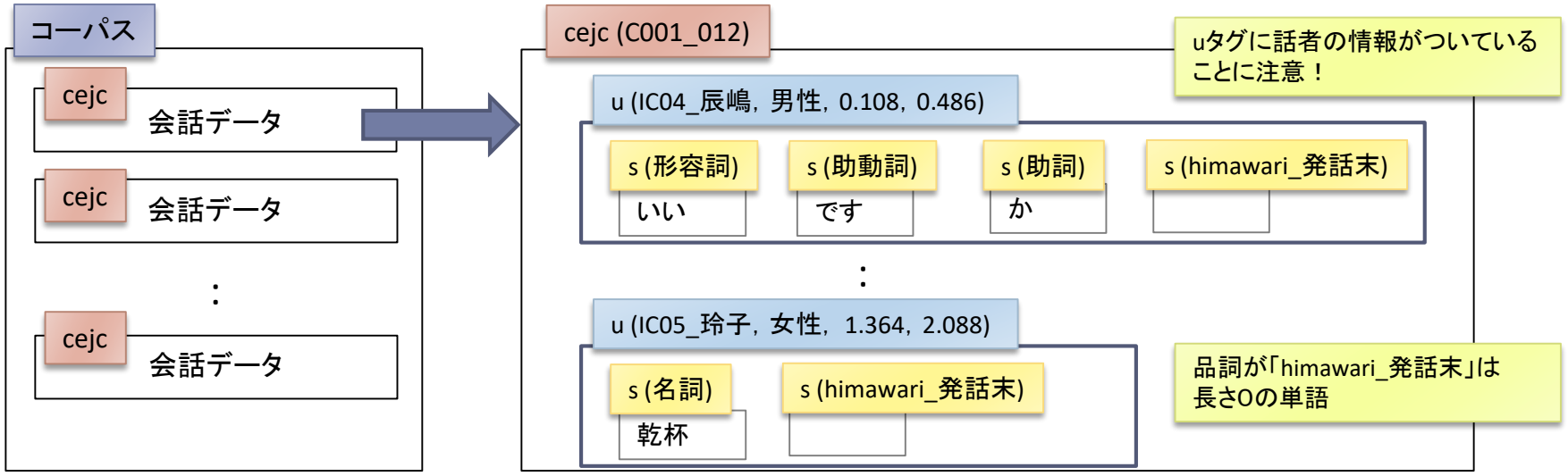
『ひまわり』は，構造を考慮した構成要素の集計が可能



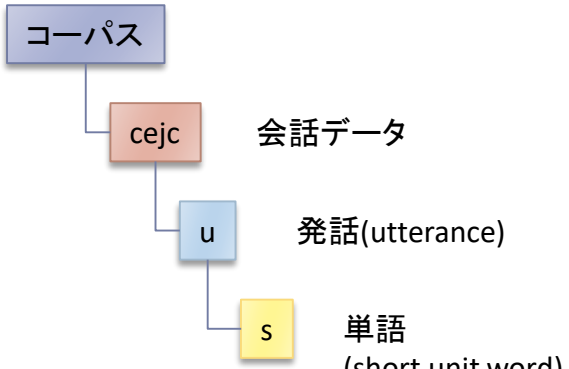
『ひまわり』用データの全体構造

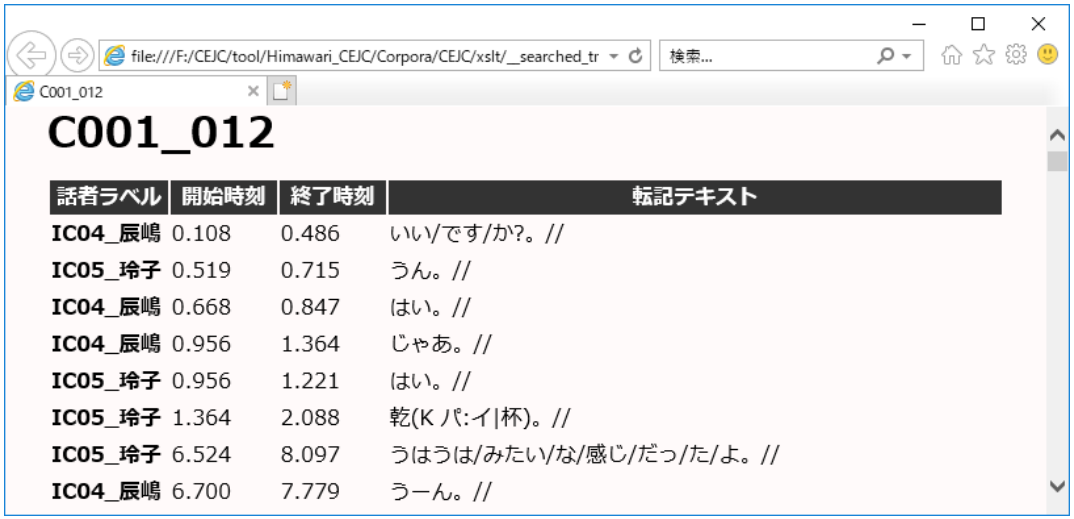
コーパス本体を見たい場合
Corpora/CEJC/corpus.xml

※「秀丸」などのテキストエディタを利用のこと



- タグの階層構造

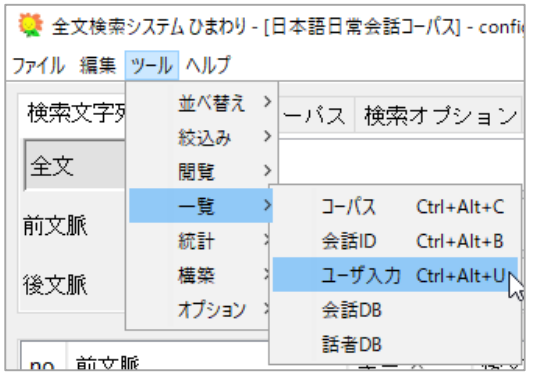




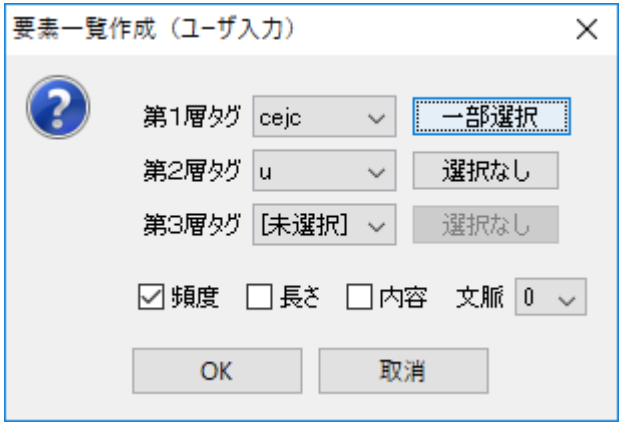
話者ラベル	開始時刻	終了時刻	転記テキスト
IC04_辰嶋	0.108	0.486	いい/です/か?。//
IC05_玲子	0.519	0.715	うん。//
IC04_辰嶋	0.668	0.847	はい。//
IC04_辰嶋	0.956	1.364	じゃあ。//
IC05_玲子	0.956	1.221	はい。//
IC05_玲子	1.364	2.088	乾(K パ:イ|杯)。//
IC05_玲子	6.524	8.097	うはうは/みたい/な/感じ/だっ/た/よ。//
IC04_辰嶋	6.700	7.779	うーん。//

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



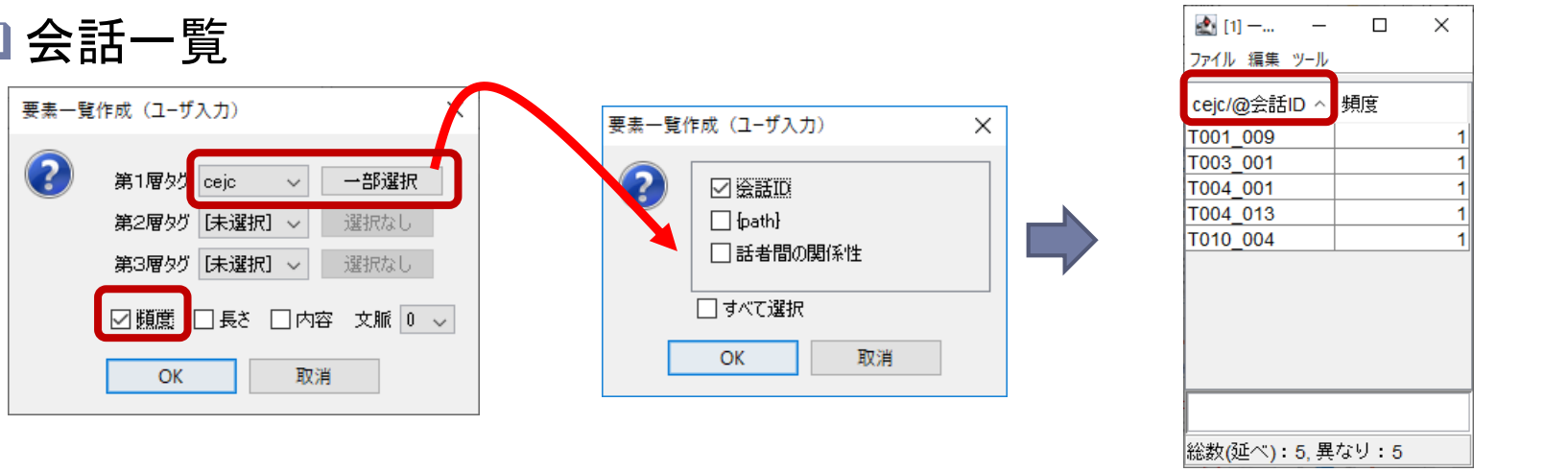
- 各会話に含まれる発話数



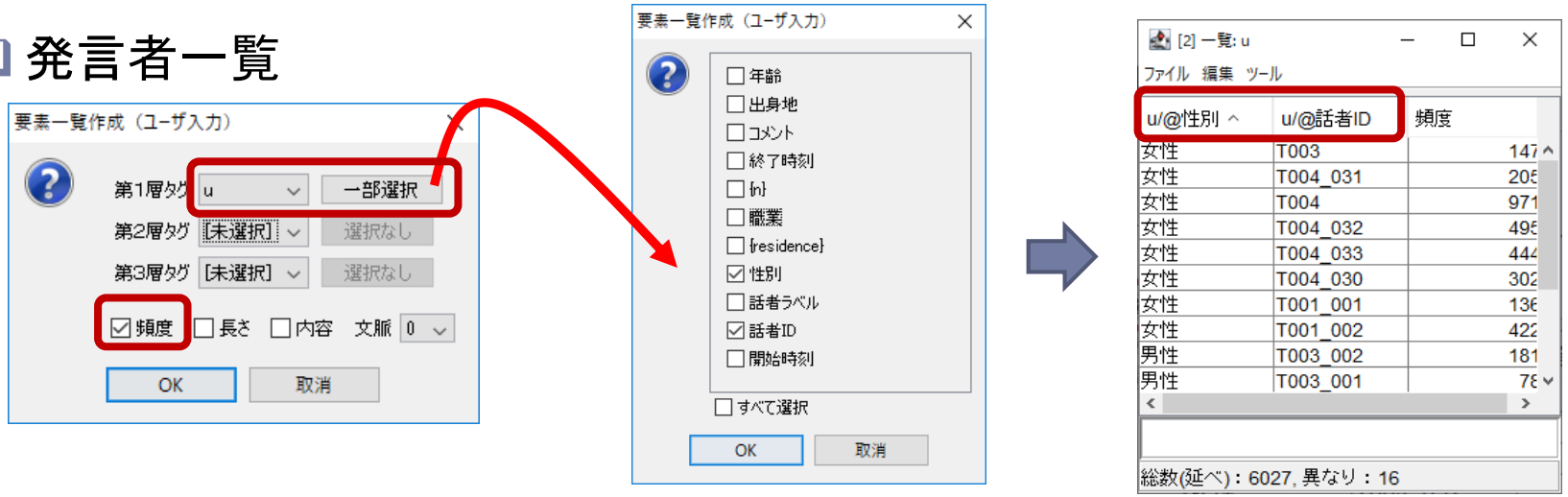
- 頻度： 指定したタグの頻度
- 長さ： マークアップされている文字列の長さ（空白やXMLタグは除く）
- 内容： マークアップされている文字列
- 文脈： 後続する同種の要素の属性をn個表示（単語の場合n+1 gramになる）

タグの集計の例1

❑ 会話一覧

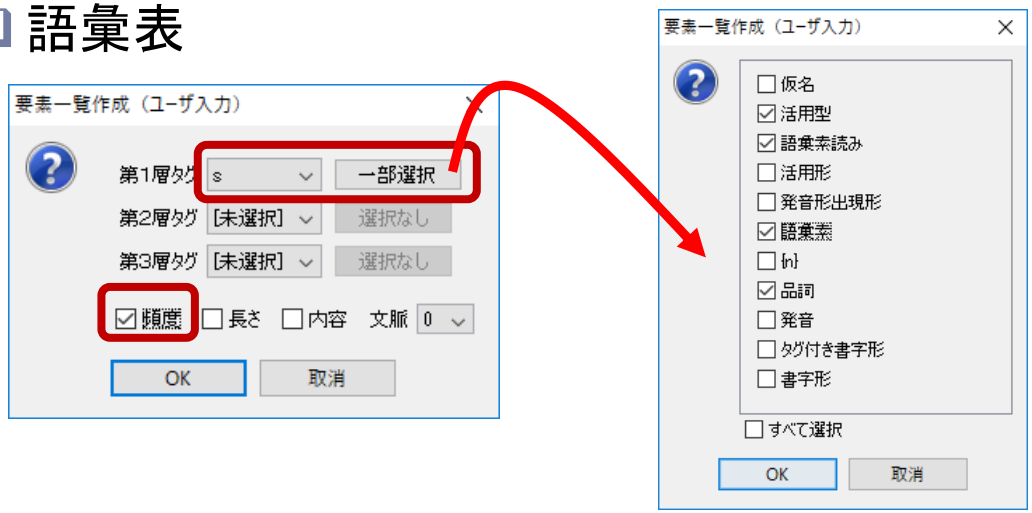


❑ 発言者一覧

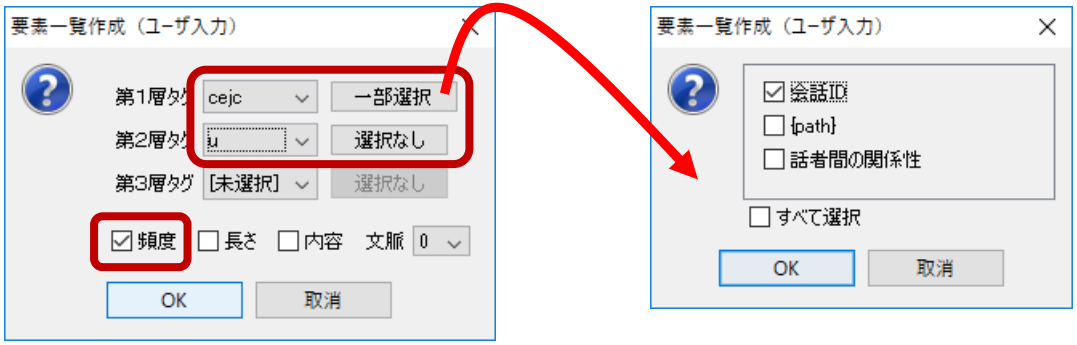


タグの集計の例2

- 語彙表

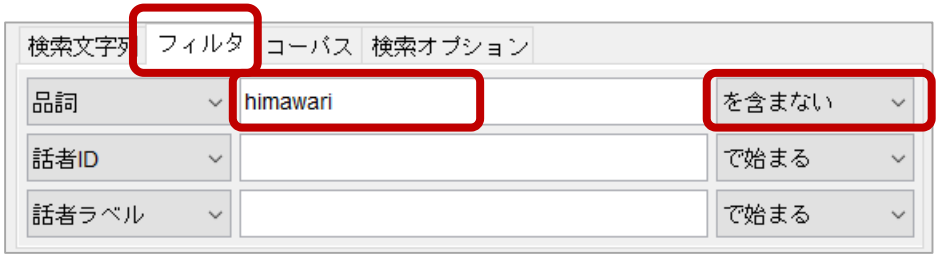


- 各会話に含まれる発話数

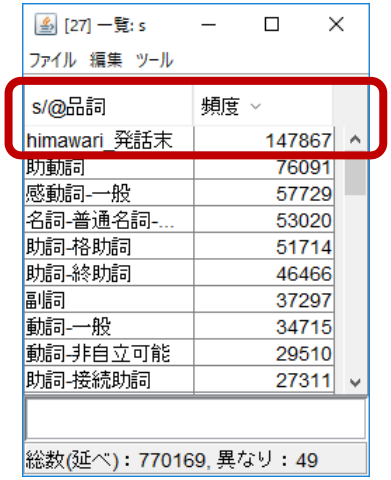


タグの集計時のフィルタ

- 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



- 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



フィルタをかけると…

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」で始まるものに限定しましょう）

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

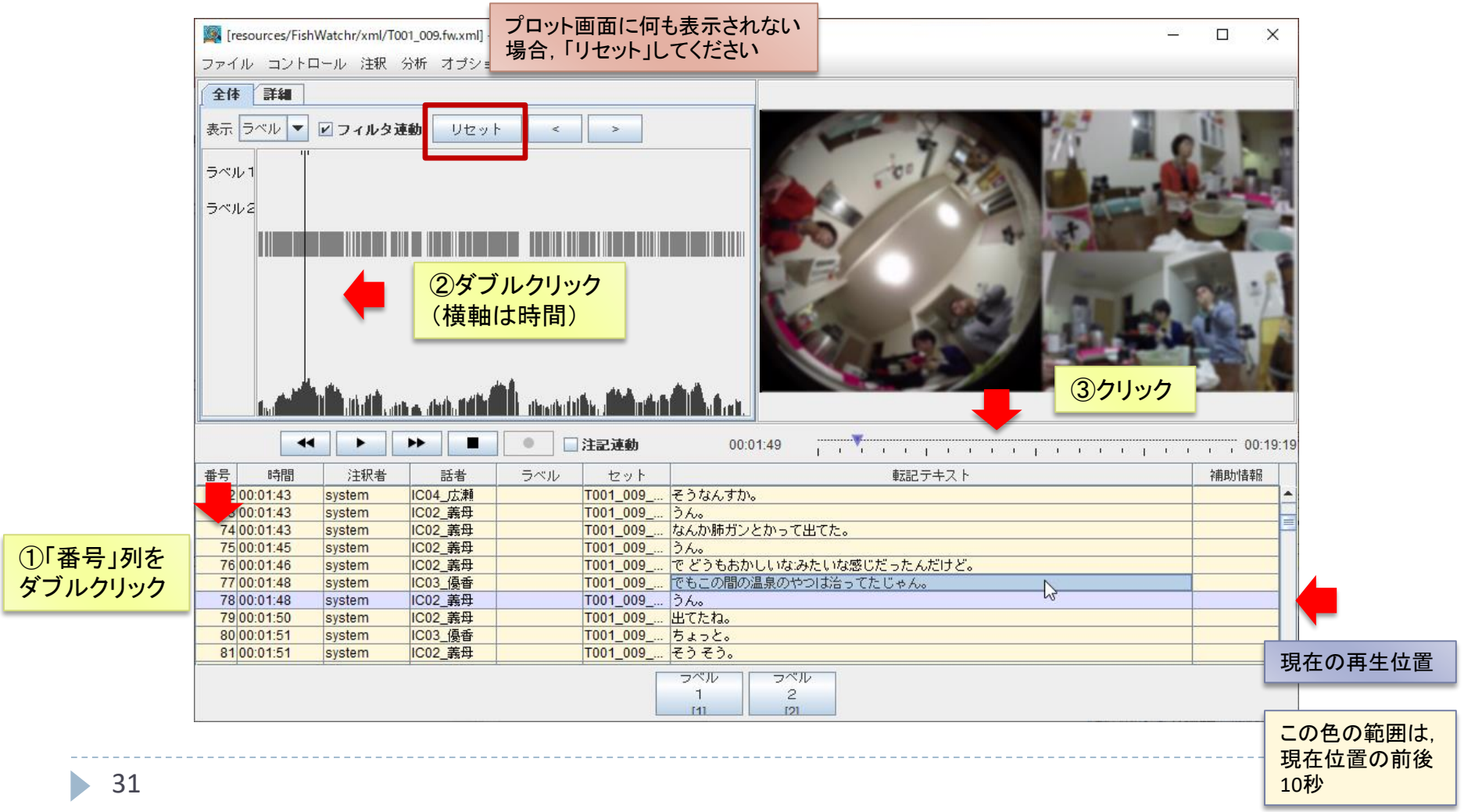
- 簡易的な利用を想定

  ⇒　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

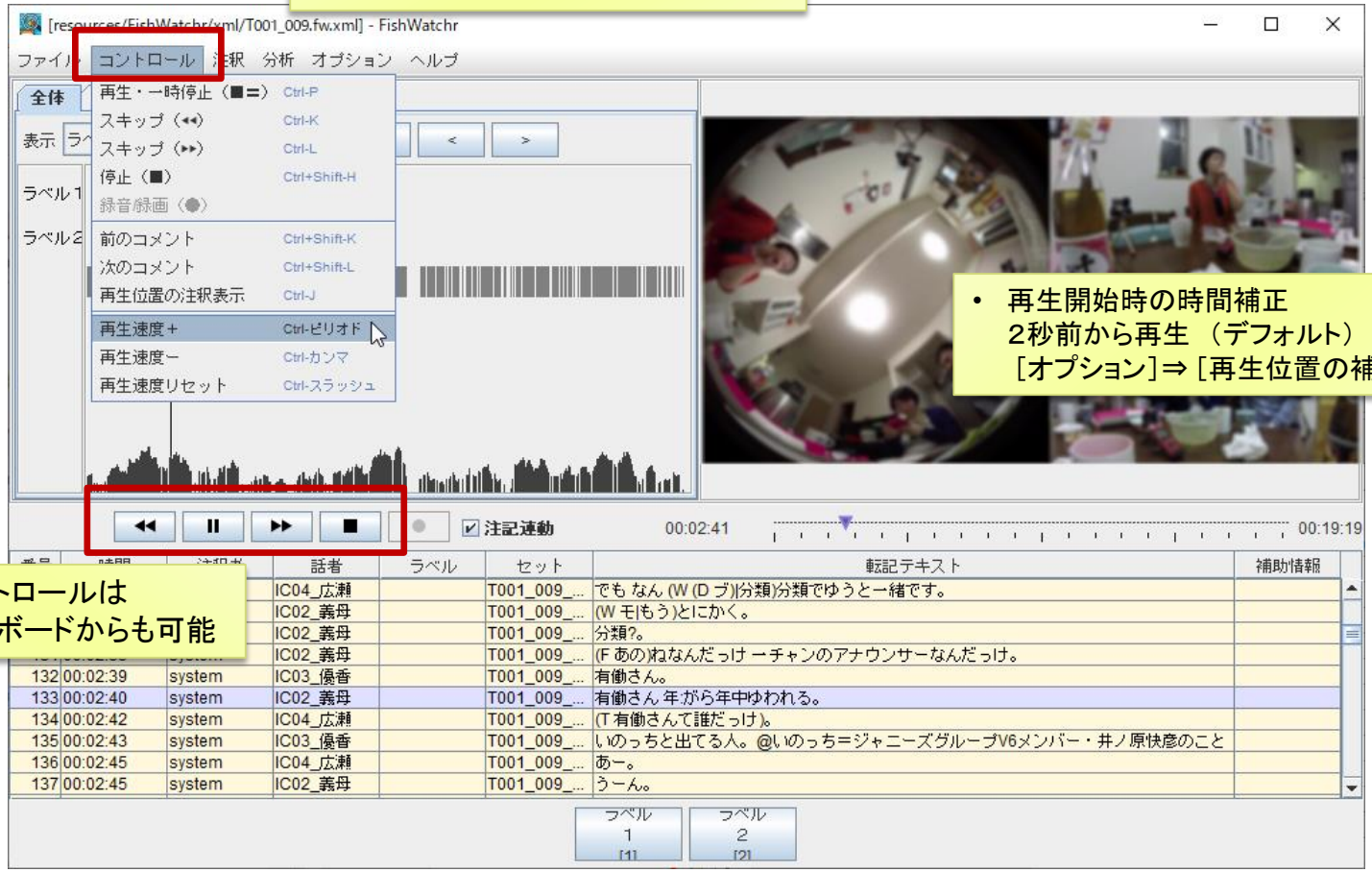
動画の再生位置の指定

- 三つの方法で指定可能



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

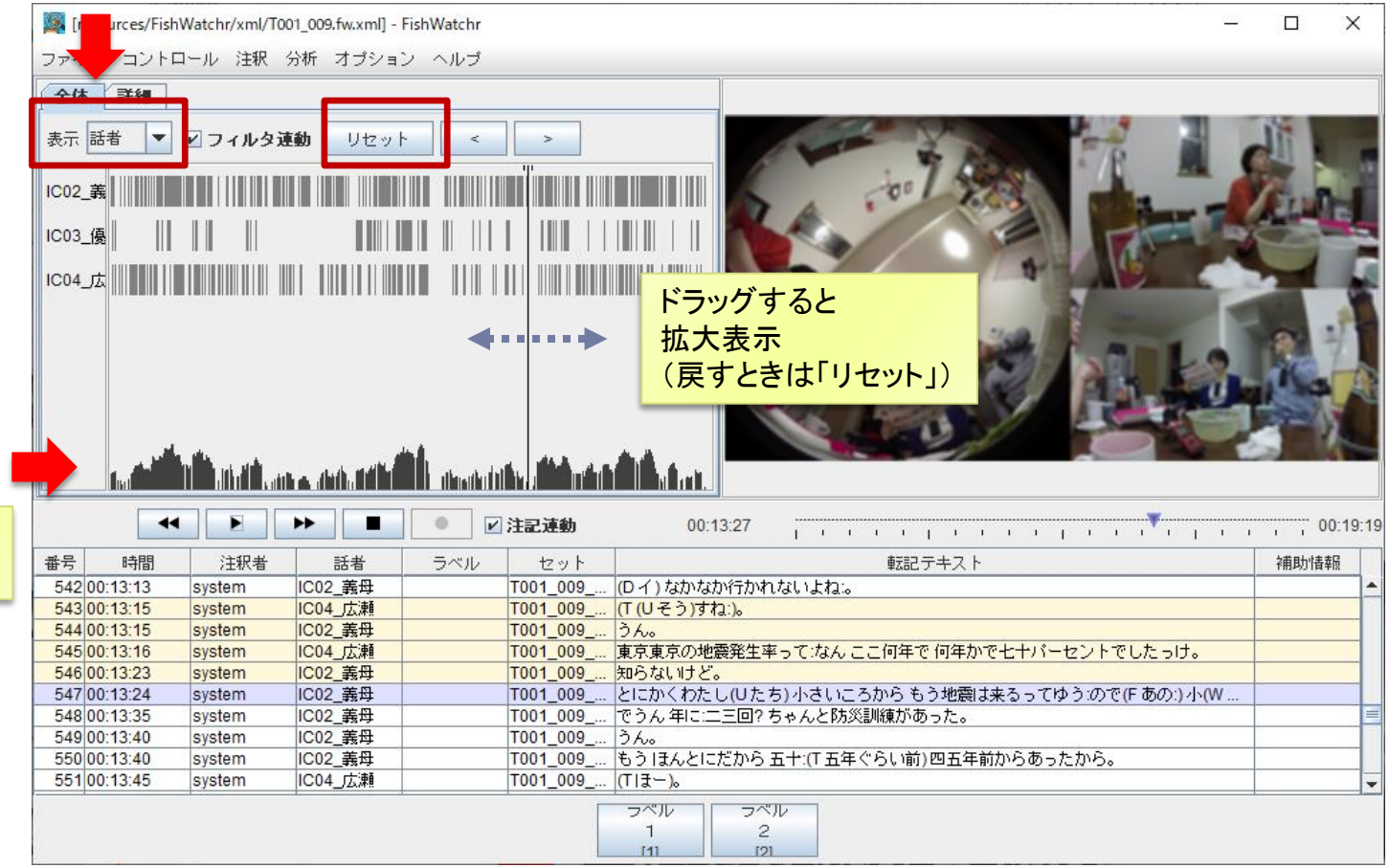


- 再生開始時の時間補正
  2秒前から再生 （デフォルト）
  ［オプション］⇒［再生位置の補正］

再生のコントロールは
すべてキーボードからも可能

発話のプロット画面

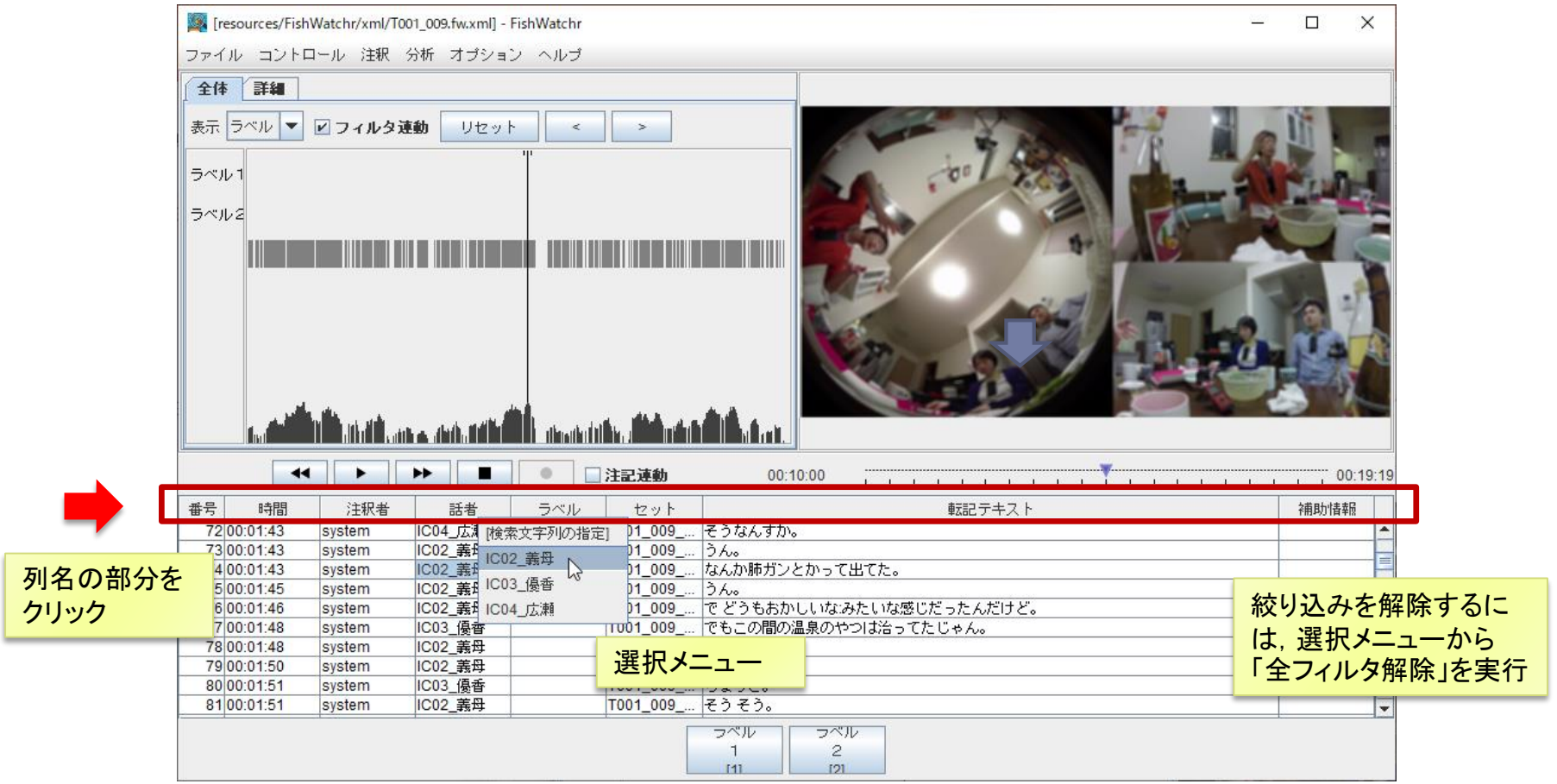
『日本語日常会話コーパス』の場合は，
「表示」を「話者」に変更するのが便利



発話の
ヒストグラム

発話の絞り込み（1）

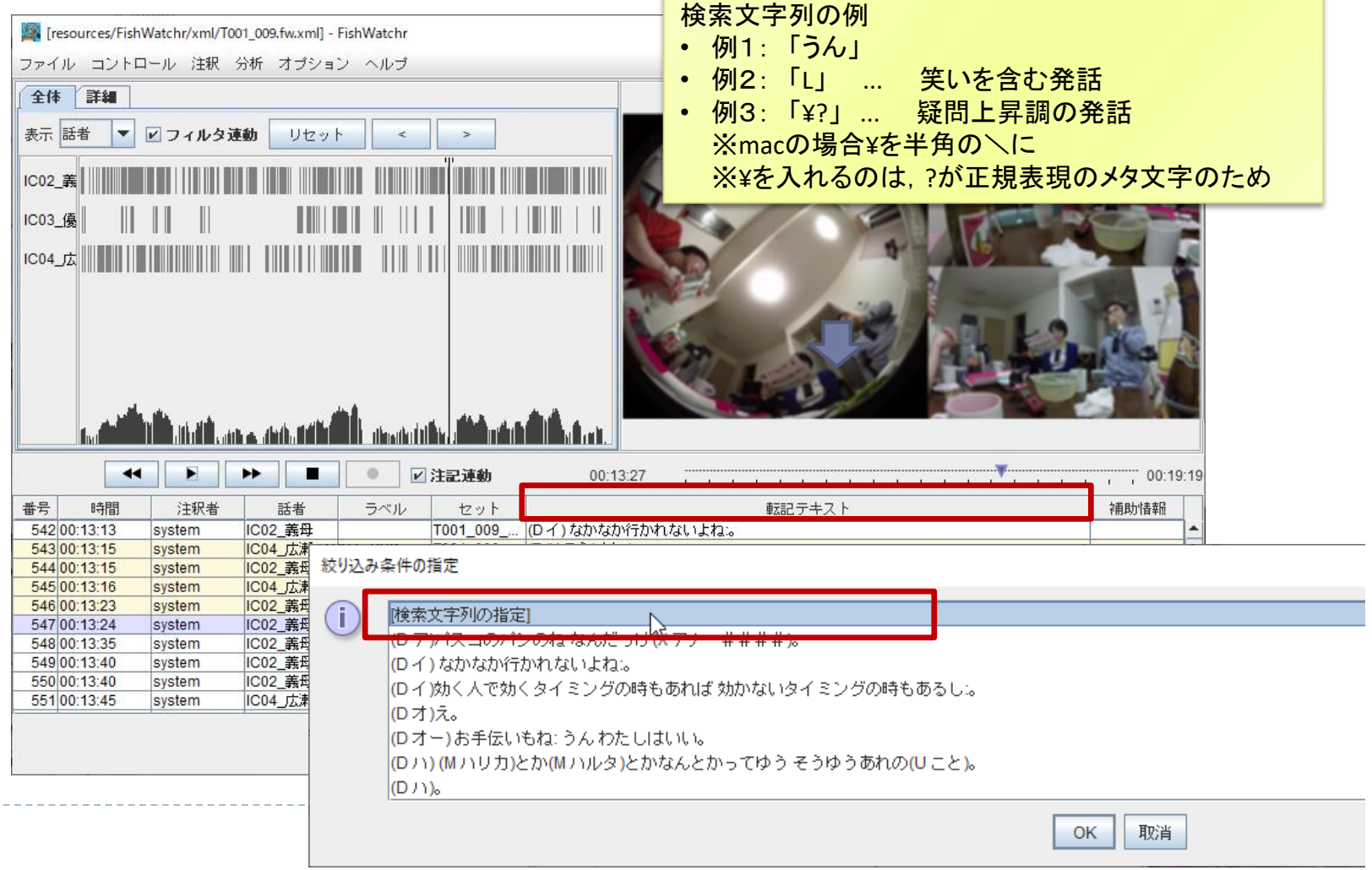
- 列名の部分を右クリックすると絞り込み



発話の絞り込み（2）

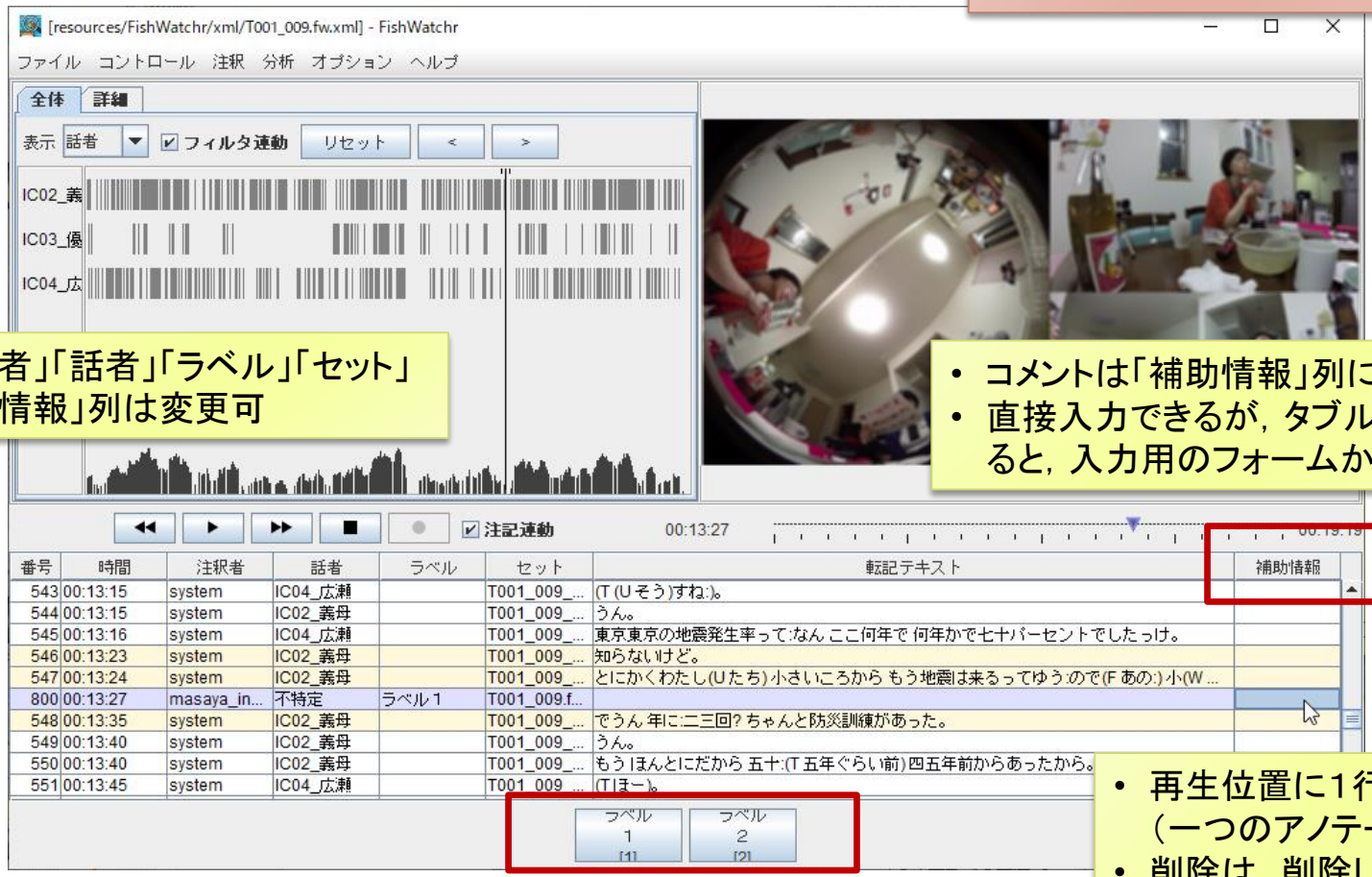
- 文字列を指定しての絞り込み

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



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



「注釈者」「話者」「ラベル」「セット」「補助情報」列は変更可

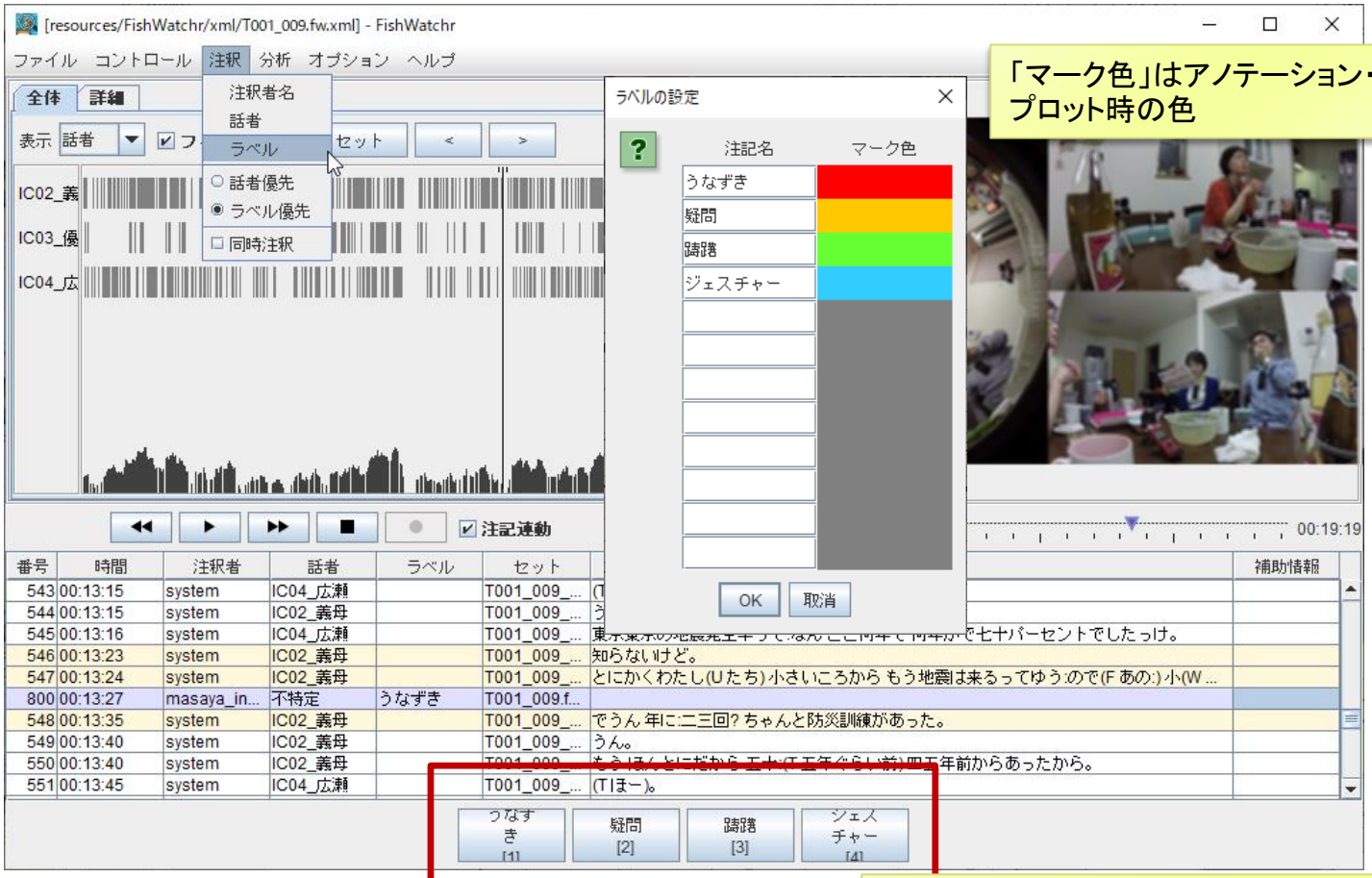
- コメントは「補助情報」列に記入
- 直接入力できるが，タブルクリックすると，入力用のフォームから入力可

アノテーションボタン

- 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(ラベルの変更)

- ラベルは自由に定義可



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

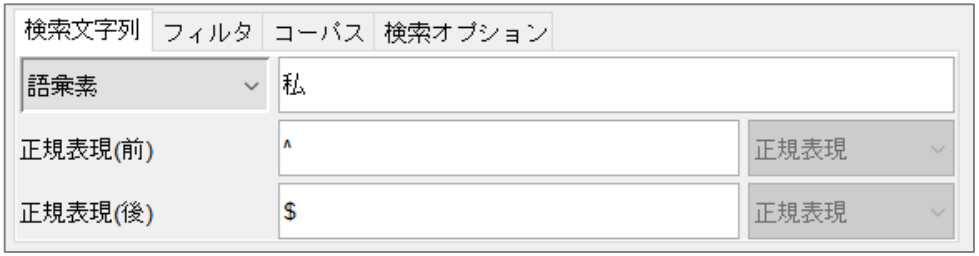
- 全文検索システム『ひまわり』
  (http://www2.ninjal.ac.jp/lrc/index.php?himawari)
- ［ヘルプ］⇒［『ひまわり』マニュアル］
- 観察支援システムFishWatchr
  (http://www2.ninjal.ac.jp/lrc/index.php?fw)

- 正規表現
  - Java Pattern クラス（『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/)

練習問題1の解答例

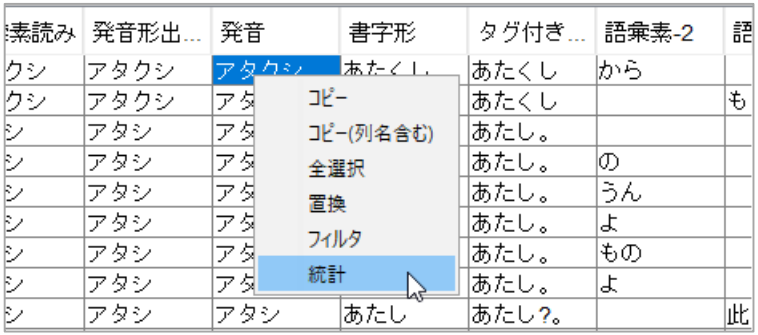
① 代名詞の「私」を検索

（「語彙素」で検索してみましょう）



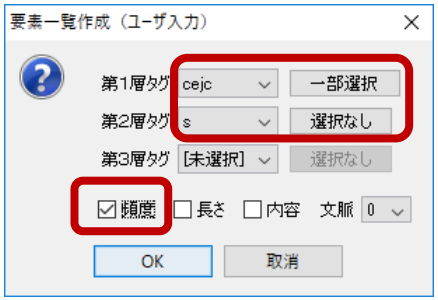
② 発音のバリエーションを集計

（検索結果の「発音」列や「キー」列で集計してみましょう）

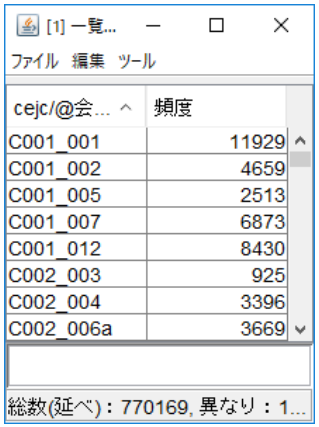


練習問題2の解答例

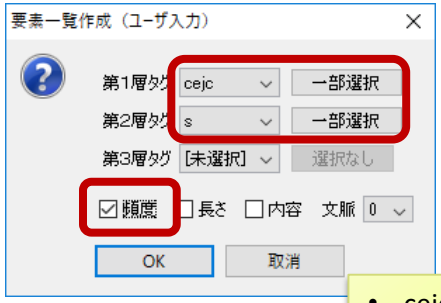
① 会話データごとの単語数を求める



cejcの属性「会話ID」チェック



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