# Development of an Environment to Make Use of "Corpus of Everyday Japanese Conversation"

# 『日本語日常会話コーパス』活用環境の構築

山口昌也 (国立国語研究所音声言語領域) †

# Development of an Environment to Make Use of "Corpus of Everyday Japanese Conversation"

Masaya YAMAGUCHI (Spoken Language Division, NINJAL) †

**要旨**

本発表では，『日本語日常会話コーパス』を活用するための環境構築について述べる。『日本語日常会話コーパス』は動画・音声，転記テキストを含み，転記テキストには形態素解析結果などの言語学的な情報がアノテーションされている。本発表で提案する活用環境は，全文検索システム『ひまわり』と観察支援システム FishWatchr を統合することにより実現した。本環境を用いることにより，次のことが可能になる。(1)『ひまわり』で転記テキストを全文・単語検索し，当該位置の映像を FishWatchr で閲覧すること，(2)FishWatchr 上で動画再生位置に簡易なアノテーション（二つのユーザ定義ラベル，自由テキストを記述可能）を付与すること，(3)FishWatchr 上で転記テキストを表形式で表示し，選択した転記テキスト位置の動画を再生すること。また，動画の再生と同期させて転記テキストをスクロール表示すること。

## 1 はじめに

本稿では，現在，国立国語研究所で構築中の『日本語日常会話コーパス』（以後，CEJC）(小磯花絵ほか, 2017) を活用する環境の構築について述べる。CEJC は日常場面で自発的に生じた多様な日常会話を収録したコーパスで，ビデオデータ，転記テキストを含み，転記テキストには単語情報，会話情報，話者情報などの言語学的な情報がアノテーションされている。

このように，CEJC は一次資料のビデオデータに対して，さまざまなデータがアノテーションされている。これらのデータを扱うには，既存のツールを利用することができる。例えば，映像の閲覧や分析には，メディアプレーヤー (例：VLC[1])，ビデオアノテーションシステム (例：ELAN[2] (Brugman and Russel, 2004))，音声分析であれば，Praat[3](Boersma and Weenink, 2001) などの音声分析ソフトウェア，転記テキストに対する全文検索や単語検索に対しては，KHCoder[4](樋口耕一, 2003) や『ひまわり』[5](山口昌也・田中牧郎, 2005) などである。

その一方で，CEJC に含まれるデータを効率的に利用するには，複数の種類のデータを統合して利用する環境が必要である。例えば，転記テキストを検索したとき，検索結果の当該シーンや，話者・会話データなどの発話状況を迅速に参照できれば，効率的な分析が可能になる。

そこで，本稿では，複数のツールを組み合わせて，CEJC を有効に活用できる利用環境を構築する。活用環境の設計にあたっては，上で挙げた例のように，転記テキストを検索し，一次資料であるビデオ

† http://www2.ninjal.ac.jp/masaya

[1] https://www.videolan.org/vlc/

[2] https://tla.mpi.nl/tools/tla-tools/elan/

[3] http://www.fon.hum.uva.nl/praat/

[4] http://khc.sourceforge.net/

[5] http://www2.ninjal.ac.jp/lrc/index.php?himawari

データを参照したり，ビデオデータに簡易なアノテーションを行う利用形態を想定する。また，本環境は，CEJC 公開時に同梱することを想定しているため，容易に利用できることを目指す。

以上の背景から，本環境は，全文検索システム『ひまわり』と観察支援システム FishWatchr とを組み合わせて，構築する。『ひまわり』は，XML でアノテーションされたテキストに対するコンコーダンサで，全文検索・単語検索，検索文字列の KWIC 表示，アノテーション内容の表示・集計をすることができる。2004 年から一般公開され，これまでに『太陽コーパス』[6]『日本語話し言葉コーパス』[7]などの検索システムとして同梱されてきた実績がある。

一方，FishWatchr はディスカッション練習やプレゼンテーション練習などの協同型教育活動の観察を支援するためのシステムである。学習者が利用することを前提としたシステムであり，ELAN のように複雑なアノテーションを行うことはできないが，特定のシーンに対して，アノテーション専用ボタンで容易にアノテーションすることが可能である。

この後の本稿の構成は，次のようになっている。まず，次節では，本環境の設計を行うために，基本的な利用形態を定めた上で，必要とされる機能を示す。3 節では，本環境を構築するための方法として，CEJC のデータを『ひまわり』と FishWatchr にインポートする方法を示す。さらに，4 節で実現した環境での実行例を示し，5 節でまとめを述べる。

## 2　活用環境の設計

### 2.1　CEJC のデータ構成

前述のとおり，CEJC には複数の種類のデータが含まれている。ここでは，本環境の設計を行う前に，CEJC のデータ構成（図 1）について説明しておく。なお，ここで述べるデータ構成は，本環境に関連する部分のみであり，全データ構成については，小磯花絵ほか (2017) などを参照されたい。

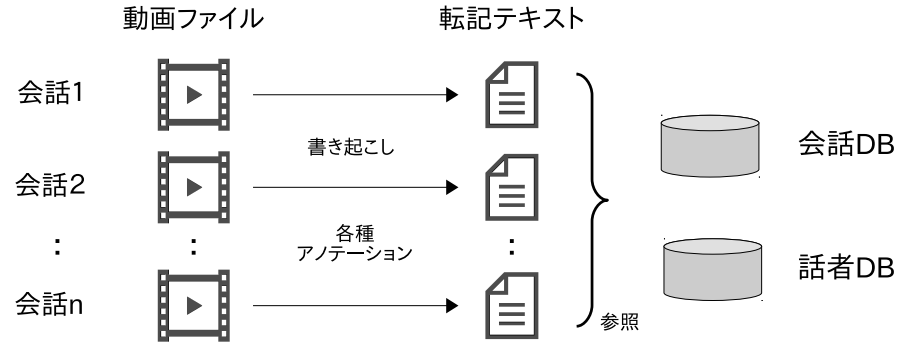


図 1　CEJC のデータ構成

このように，各会話はビデオファイルとして格納される。会話内の発話は書き起こされ，転記ファイルに記述される。発話には，話者名，ビデオファイルにおける時間情報，単語の情報（短単位），言いさしや言い誤りなどの情報がアノテーションされる。会話や発話者の詳細情報は，データベースとしてまとめられており，適宜参照できるようになっている。

[6] http://pj.ninjal.ac.jp/corpus_center/cmj/taiyou/
[7] http://pj.ninjal.ac.jp/corpus_center/csj/

### 2.2 基本的な利用形態と機能

本稿では，図 2 のような利用形態を想定する。この図のとおり，転記テキストに対する全文検索，もしくは，単語検索を行い，その結果を KWIC 表示するのが最も基本的な利用方法である (図 2 ①②)。この際，個々の検索結果には，KWIC キーに関連する発話者，会話などの情報も併記する。

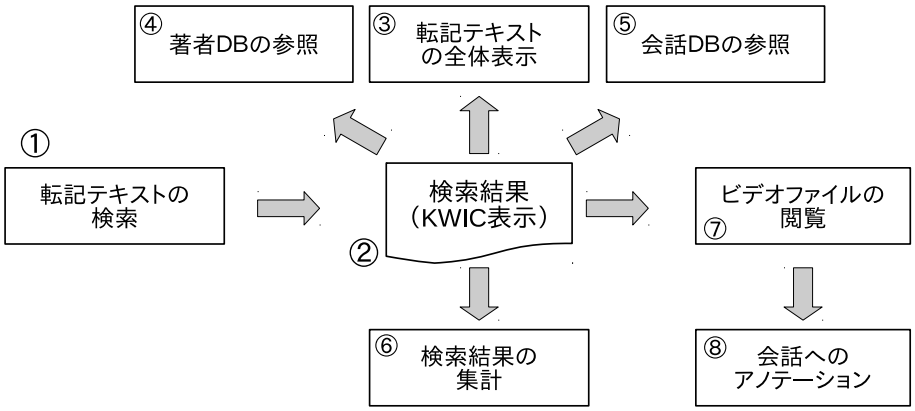


図 2 想定する基本的な利用形態

この検索結果を起点として，さまざまな処理を行う。処理は，(1) 転記テキスト関連情報の取得 (図 2 ③④⑤)，(2) 検索結果の集計 (図 2 ⑥)，(3) ビデオファイル閲覧 (図 2 ⑦⑧)，の三つに分けられる。

**■転記テキスト関連情報の取得** 検索結果の KWIC 表示は，表示範囲が限定されるため，転記テキスト全体を閲覧できるようにする。また，詳細な話者情報や会話情報を個々の検索結果に併記するのは表示量が多くなりすぎるので，必要に応じて，閲覧したい検索結果を選択し，話者 DB，会話 DB を参照できるようにする。

**■検索結果の集計** 大量の検索結果が得られた場合，その結果を集約したり，統計的な分析を支援することができるようにする。例えば，検索結果から発話者別の頻度を求めたり，検索文字列の調整頻度を計算するために，会話ごとの単語数を収集するといった処理である。

**■ビデオファイルの閲覧** 検索結果の当該シーンのビデオを転記テキストと並行して，閲覧できるようにする。また，閲覧しているビデオファイル中のシーンを指定して，ラベルやコメントを付与するなどの簡易的なアノテーションができるようにする。

## 3 活用環境の実現

### 3.1 全体構成

本稿で提案する CEJC 活用環境の全体構成を図 3 に示す。

活用環境は，大きく分けて，『ひまわり』と FishWatchr の二つのシステムから構成される。図 2 の利用形態と対応させると，『ひまわり』は①〜⑥を担当し，FishWatchr は⑦⑧を担当する。ただし，転記テキストの全体表示（③）には，Web ブラウザを用いる。

CEJC に含まれるデータのうち，『ひまわり』側で扱うのは，転記テキスト，会話データベース，話者データベースである。転記テキストは，『ひまわり』用のコーパスファイルに変換する。その際，転記テキストにアノテーションされている単語情報は，XML タグとして記述される。詳細は，この後で述

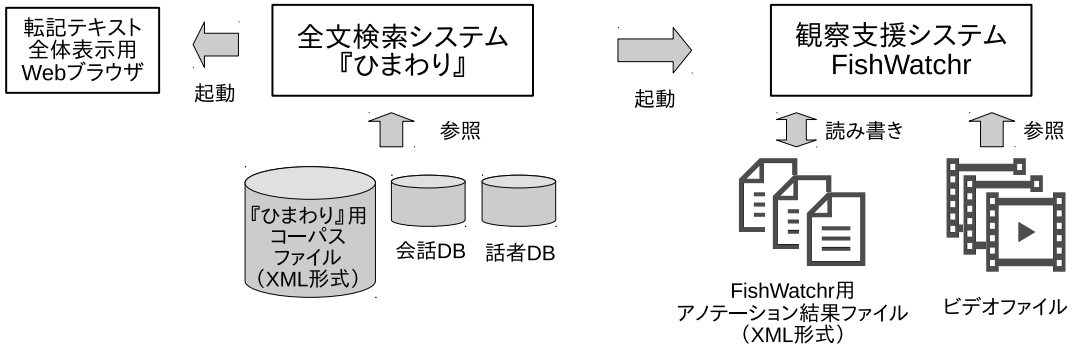


図 3　活用環境の全体構成

べる。会話データベース，話者データベースは，『ひまわり』内部のデータベース（書誌情報などを格納するために設計されたもの）のデータベースに格納される。

一方，FishWatchr 側で扱うのは，ビデオファイルと転記テキストである。転記テキストは，会話ごとに FishWatchr 用のアノテーション結果ファイルに変換され，ビデオとの連動表示（4.2 節参照）を行うために利用される。このデータファイルには，FishWatchr で新規に行ったアノテーションも追記される形で，保存される。

この後の節では，CEJC のデータが『ひまわり』，FishWatchr にどのようにインポートされるかを詳しく説明する。なお，図 2 で示した各種の処理については，両システムの機能を組み合わせて利用しているため，動作例を 4 節で示すにとどめる。

### 3.2　『ひまわり』への CEJC データのインポート

図 4 は，CEJC の CSV 形式の転記テキストの例である。1 行 1 発話で記述され，発話の開始・終了時間，話者名が付与されている。発話部分には，独自の形式のタグが付与されている。例えば，2 行目の（F あの:）は F タグでフィラーであることを示している。4 行目の（W ナ|なん）は W タグで言い間違いを修正している (|の前が修正前，後が修正後である)。なお，転記テキストには，単語情報もアノテーションされているが，表記が複雑になるため，ここでは除外している。

| startTime | endTime | speaker | text |
|---|---|---|---|
| 381.949 | 383.086 | IC02_美沙 | なんかね（F あの） |
| 383.175 | 385.061 | IC02_美沙 | えーとねー ちょっと待ってね:。 |
| 384.607 | 386.226 | IC01_玲子 | （W ナ\|なん）かもうすごい未知の世界。 |
| 386.209 | 386.825 | IC02_美沙 | でしょ:。 |

図 4　転記テキストの例 (C001_002 から一部引用)

『ひまわり』にインポートする際は，これらを XML 形式で記述する。図 5 は，図 4 の転記テキストを XML 形式に変換した結果である。なお，紙面上の見やすさの関係上，転記テキストの 4 行目の発話のみ示した。また，適宜改行を挿入するとともに，タグの属性は説明に必要なもののみ記述している。

発話は u タグでマークアップされ，1 会話分の発話が cecj タグでマークアップされる。どちらのタグも属性を持ち，cejc タグの name 属性値は会話の ID の役割を果たす。u タグの startTime，endTime 属性は，ビデオファイルにおける発話の開始時間・終了時間を表し，発話をビデオファイルと関連付ける。speaker，speakerID 属性は，発話者名，発話者 ID である。speakerID 属性値は，発話 DB を参照

する際のキーとなる。単語情報 (短単位) は，s タグでマークアップされる。s タグの p，l，t 属性はぞれぞれ品詞，「語彙素」，「タグなし出現形」を保持する。

```
<cejc name="C001_002">
                    :
<u startTime="384.607" endTime="386.226" speaker="IC02_玲子" speakerID="C001">
 <s p="代名詞" l="何" t="なん">(W ナ|なん)</s>
 <s p="助詞-副助詞" l="か" t="か">か</s>
 <s p="副詞" l="もう" t="もう">もう</s>
 <s p="形容詞-一般" l="凄い" t="すごい">すごい</s>
                    :
 <s p="名詞-普通名詞" l="世界" t="世界">世界。</s>
</u>
                    :
</cejc name="C001_002">
```

図 5　『ひまわり』へのインポート例

『ひまわり』は，XML タグを無視して全文検索する。したがって，全文検索時は，CEJC の独自タグを考慮しつつ，検索文字列を指定する必要がある。独自タグを除外して検索したい場合は，単語検索で「タグなし出現形」を検索すればよい。前後 2 単語だが，前後の文脈を指定できる。

### 3.3　FishWatchr への CEJC データのインポート

FishWatchr では，『ひまわり』と異なり，一つの転記ファイル（会話）が一つの FishWatchr 用のアノテーション結果ファイルとしてインポートされる。形式は，XML である。FishWatchr のアノテーションは，時間的範囲を持たない，特定の 1 シーンに対して行われる。そのため，一つの発話はその開始時間を基点とするアノテーションとして記述される。

図 6 は，図 4 の転記ファイルを FishWatchr 用のアノテーション結果ファイルに変換した結果である。comment_list タグは，1 会話分のアノテーション結果を表し，media_file 属性にビデオファイル名を格納している。発話は comment タグでマークアップする。commnet タグの commenter，comment_time 属性は，それぞれ発話者，ビデオファイルにおける発話開始時間を保持する。さらに，comment_type 属性にはユーザ定義のラベル，aux 属性には自由記述のコメントが格納される。

```
<comment_list media_file="C001_002_MIX.mp4">
 <comment commenter="IC02_美沙" comment_type="" aux="" comment_time="381949">
    なんかね (F あの)</comment>
 <comment commenter="IC02_美沙" comment_type="" aux="" comment_time="383175">
    えーとねー ちょっと待ってね:。</comment>
 <comment commenter="IC01_玲子" comment_type="" aux="" comment_time="384607">
    (W ナ|なん) かもうすごい未知の世界。</comment>
 <comment commenter="IC02_美沙" comment_type="" aux="" comment_time="386209">
    でしょ:。</comment>
</comment_list>
```

図 6　FishWatchr へのインポート例

インポート後の初期状態では，転記テキストからインポートした発話のみがアノテーションされている状態だが，後述するように，ユーザはビデオを参照しながら，任意のシーンに対してアノテーションを追加することができる。

## 4　実現結果

### 4.1　『ひまわり』

まず，『ひまわり』で CEJC を検索した結果を図 7 に示す。検索結果には，画面左から，検索文字列「会話」に対する KWIC，会話 ID，話者関連情報，発話情報，単語情報が含まれる。

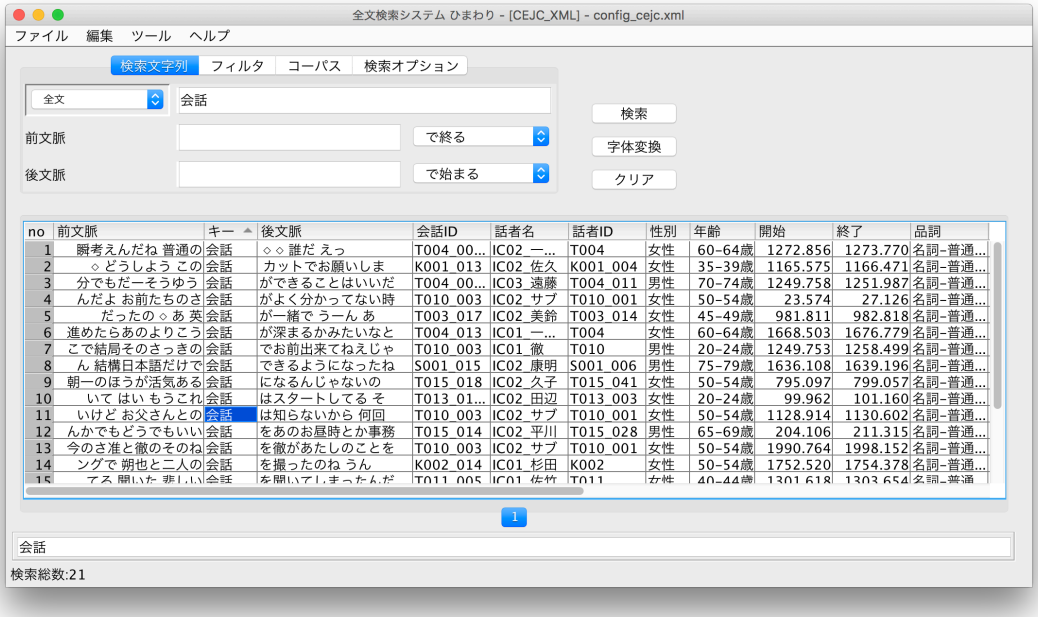


図 7　『ひまわり』による CEJC の検索例

このうち，KWIC 部分のセルをダブルクリックすると，図 8 のように，当該の転記テキスト全体が Web ブラウザで表示される。また，「会話 ID」「話者 ID」列のセルをダブルクリックすると，それぞれ会話 DB，話者 DB から検索された情報が表示される（図 9 は話者情報）。なお，それぞれのデータベースの内容は一覧することも可能である。これら以外の列をダブルクリックした場合は，FishWatchr が起動し，当該シーンがビデオ再生される。

検索結果の分析には，『ひまわり』の分析支援機能を利用する。ここでは，単語「です」の会話ごとの調整頻度を求めるのに必要なデータを収集してみる。図 10 左が「です」を単語検索し，会話ごとの出現頻度を集計した結果である。図 10 中央は s タグを会話別に集計し，会話別の総単語数を求めた結果である。さらに，図 10 右は，『ひまわり』の集計結果連結機能を用いて，二つの結果を連結した結果である。この結果を Excel などの表計算ソフトウェアにコピー＆ペーストすれば，例えば，会話ごとの調整頻度を計算することができる。

### 4.2　FishWatchr

ここでは，『ひまわり』から起動した FishWatchr でビデオファイルを閲覧し，特定のシーンにコメントをつけてみる。


図 8　転記テキストの全体表示


図 9　話者情報の表示

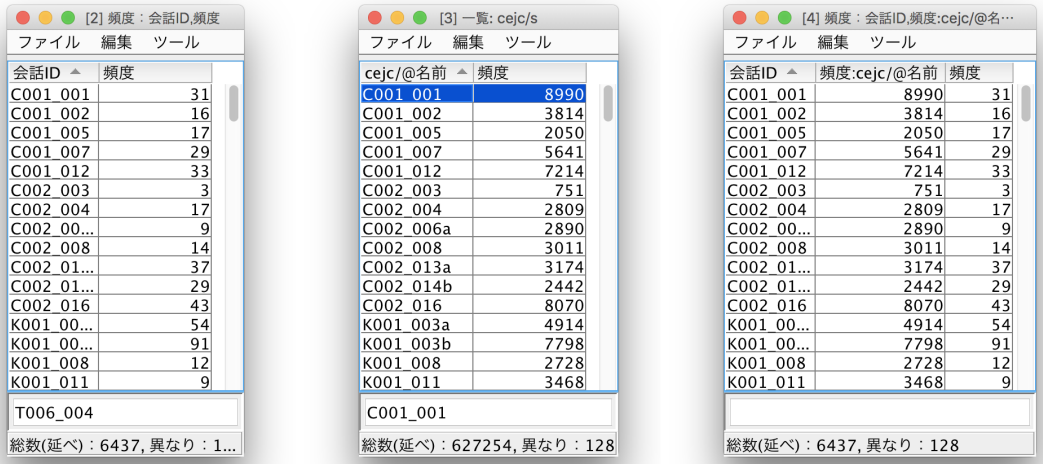


図 10　調整頻度を計測するためのデータ収集（「です」の出現頻度，会話ごとの総単語数，両者の結合結果）

図 11 は，FishWatchr にアノテーション結果ファイルを読み込んだ例である。ウィンドウ右上が会話のビデオである。下部の表はアノテーション表であり，1 アノテーション（つまり 1 発話）1 行で表示される。ウィンドウ左上はアノテーション表を時系列にプロットした図である。

アノテーション表の各アノテーションには発話の開始時間，発話者名，会話 ID，発話の転記テキストが含まれる。このうち，発話の開始時間，転記テキストは変更できないように設定されている。そのため，発話に対してコメントしたい場合は，最右列の「補助情報」欄を用いる。

アノテーション表の表示はビデオの再生と連動してスクロール表示させることもできる。また，閲覧したい行をダブルクリックすると，当該のシーンが再生される。

画面最下部の二つのボタン (「ラベル 1」「ラベル 2」) はアノテーション専用のボタンである。押下すると，ビデオの再生位置にアノテーションが付与される。ボタンのラベルはユーザが 8 個まで定義可能である。

## 5　おわりに

本稿では，CEJC を有効に活用するための環境の構築方法として，全文検索システム『ひまわり』と観察支援システム FishWatchr を組み合わせて用いる方法を提案し，その実現結果を示した。今回構築したシステムは，本年度に予定されている CEJC のモニタ公開版にも同梱される予定である。

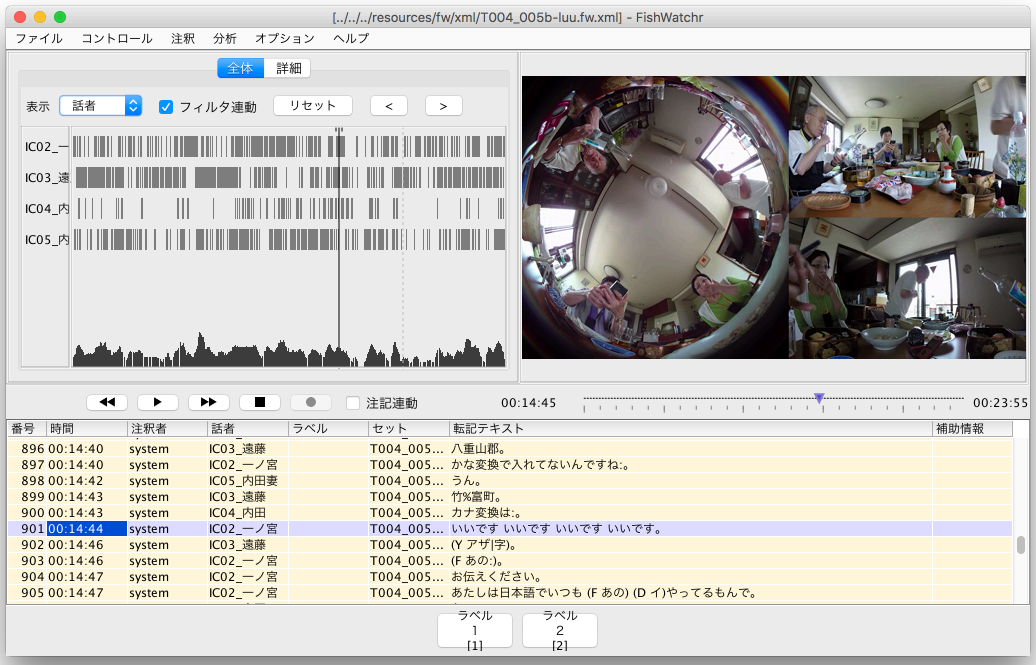


図 11　FishWatchr の動作例

## 謝　辞

本研究は，国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」，および，科研費基盤研究 (B)『「昭和話し言葉コーパス」の構築による話し言葉の経年変化に関する実証的研究』の一環で行われたものである。本環境を設計するにあたり，国立国語研究所の川端良子氏から貴重なご意見をいただいた。また，国立国語研究所の西川賢哉氏，小磯花絵氏には，データの作成・利用方法に関して情報を提供していただいた。深く感謝いたします。

## 文　献

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉 (2017).「『日本語日常会話コーパス』構築」　言語処理学会第 24 回年次大会発表論文集, pp. 775–778.

H. Brugman, and A. Russel (2004). “Annotating Multimedia/ Multi-modal resources with ELAN.” *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.*

Paul Boersma, and David Weenink (2001). “PRAAT, a system for doing phonetics by computer.”, 5, pp. 341–345.

樋口耕一 (2003).「コンピュータ・コーディングの実践—漱石『こころ』を用いたチュートリアル—」, 24, pp. 193–214.

山口昌也・田中牧郎 (2005).「構造化された言語資料に対する全文検索システムの設計と実現」　自然言語処理, 12:4, pp. 55–77.