『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）

背景

- 「日常会話コーパス」
  - 日常場面で生じた自発的な会話を収録したコーパス
  - ビデオデータ（映像，音声）に対して，さまざまな言語学的情報（転記テキスト，単語，話者情報，会話情報など）がアノテーションされる

⇒ 複数の情報を統合的に利用する環境の必要性

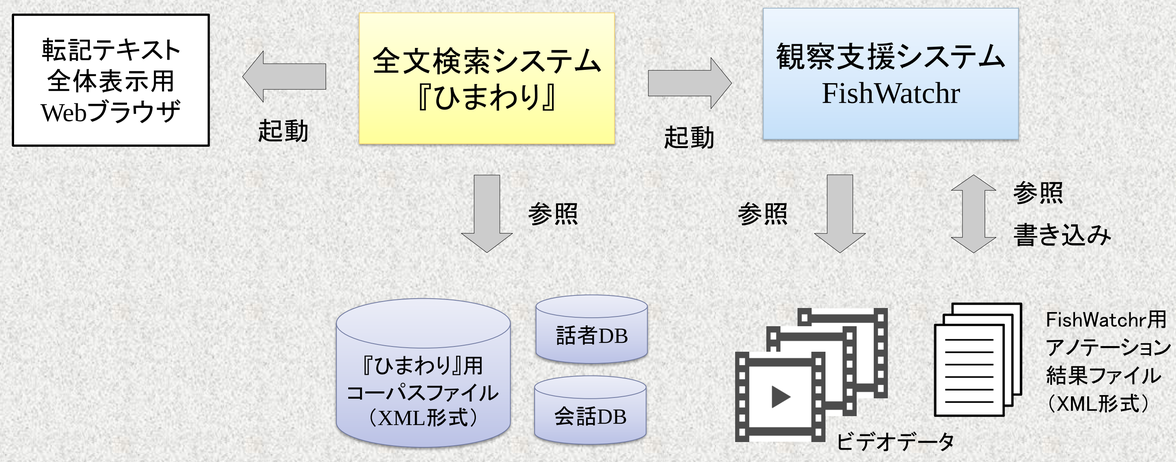
例1： 転記テキストの検索箇所のシーンを視聴
例2： 転記テキストの検索箇所の話者の詳細情報を参照

- 本活用環境の設計方針
  - 想定する利用方法
    - 転記テキストの全文検索結果を起点に，さまざまな情報を参照
    - ビデオに簡易的なアノテーション
  - 複雑な操作方法を覚えることなく，すぐに使える
    - 「日常会話コーパス」の配布ディスクに同梱

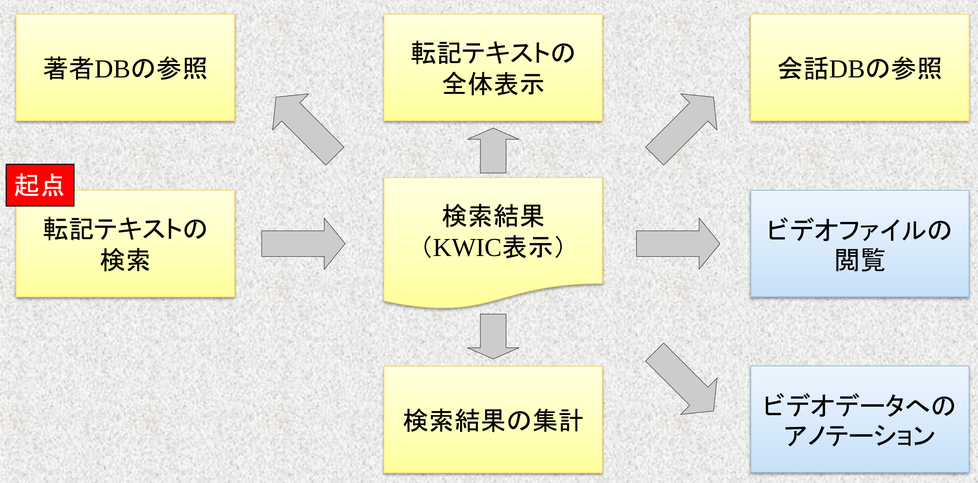
転記テキストの全文検索	全文検索システム『ひまわり』
ビデオデータの閲覧 アノテーション	観察支援システム FishWatchr

活用環境の構造

- 全体構造

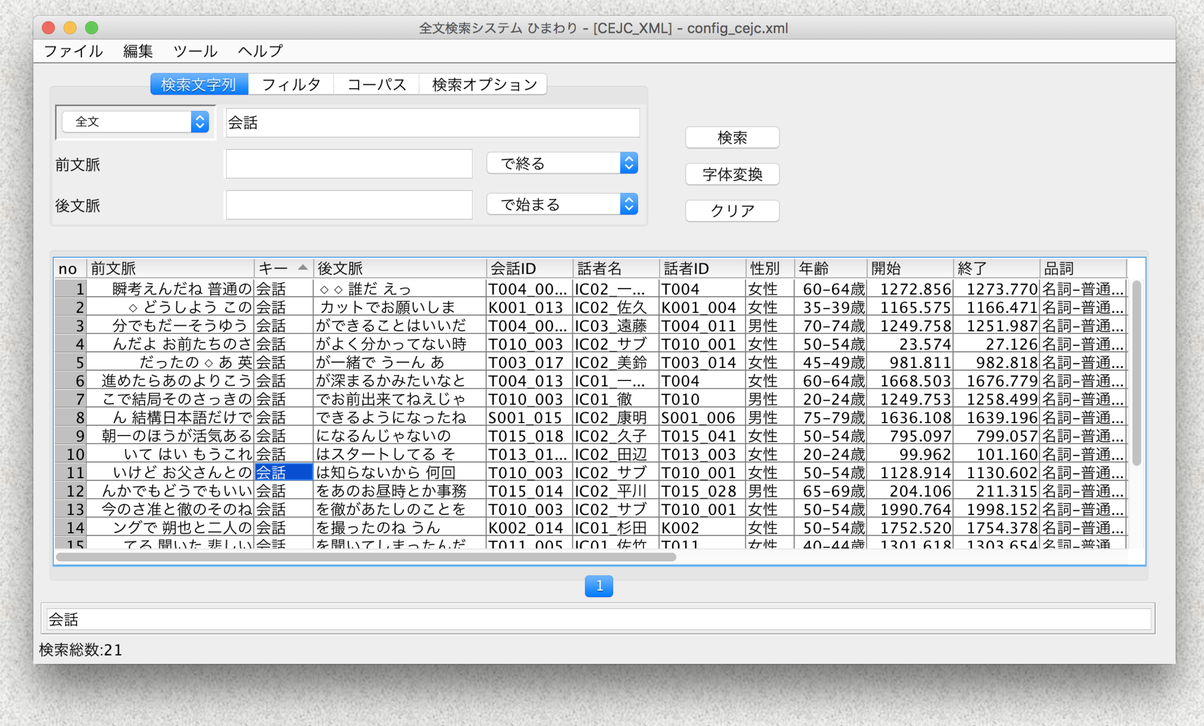


- 活用環境の機能

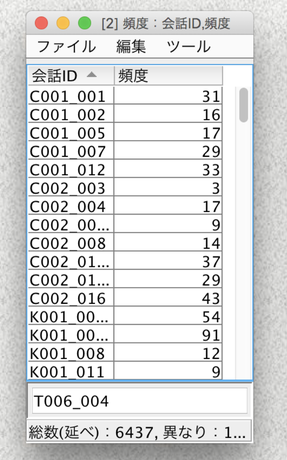


構築結果

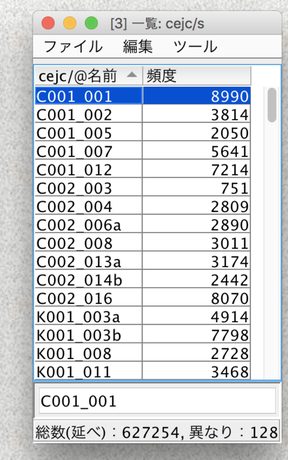
- 全文検索システム『ひまわり』



- 検索文字列の出現頻度

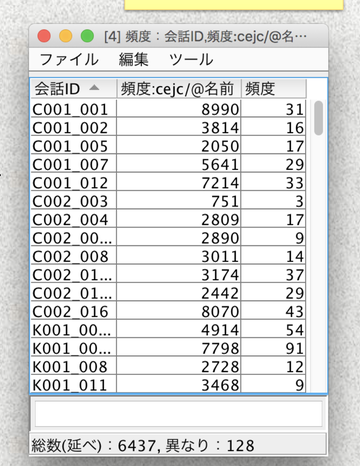


- 会話データごとの総単語数



二つの表を合併

検索結果の集計



- 観察支援システムFishWatchr

- アノテーションボタンを押したタイミングで話者，ラベル，コメントをアノテーション可能
- ラベルは自由に定義可能