

『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

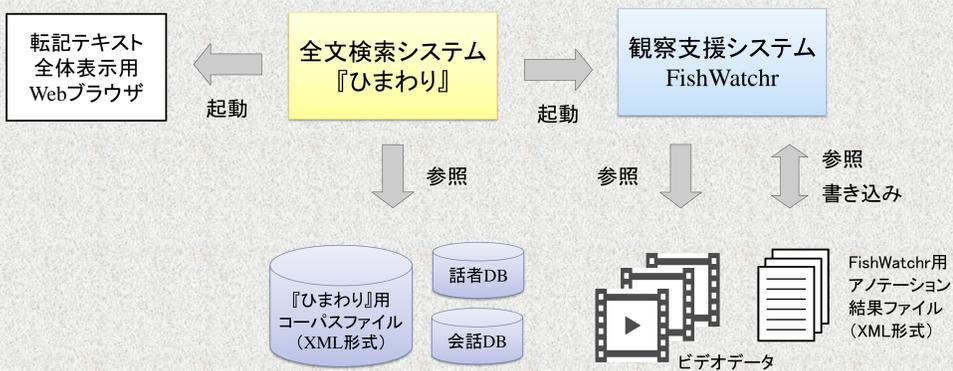
本活用環境の設計方針

- 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

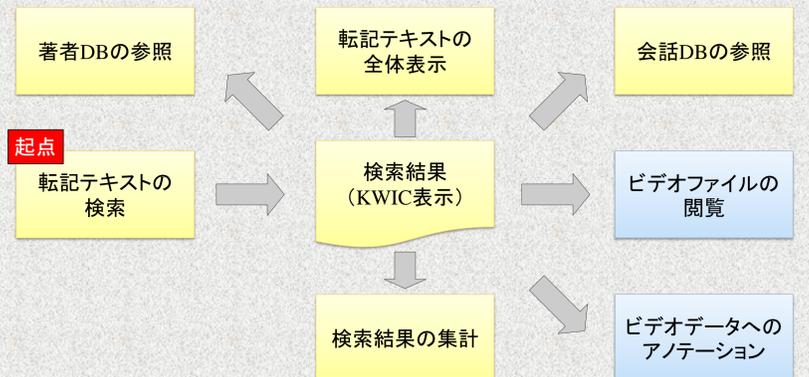


活用環境の構造

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ID	頻度
C001_001	31
C001_002	16
C001_005	17
C001_007	29
C001_012	33
C002_003	3
C002_004	17
C002_00...	9
C002_008	14
C002_01...	37
C002_01...	29
C002_016	43
K001_00...	54
K001_00...	91
K001_008	12
K001_011	9

会話データごとの総単語数

cejc/@名前	頻度
C001_001	8990
C001_002	3814
C001_005	2050
C001_007	5641
C001_012	7214
C002_003	751
C002_004	2809
C002_006a	2890
C002_008	3011
C002_013a	3174
C002_014b	2442
C002_016	8070
K001_003a	4914
K001_003b	7798
K001_008	2728
K001_011	3468

二つの表を合併

検索結果の集計

会話ID	頻度:cejc/@名前	頻度
C001_001	8990	31
C001_002	3814	16
C001_005	2050	17
C001_007	5641	29
C001_012	7214	33
C002_003	751	3
C002_004	2809	17
C002_008	2890	9
C002_013	3011	14
C002_016	3174	37
C002_014b	2442	29
C002_016	8070	43
K001_003a	4914	54
K001_003b	7798	91
K001_008	2728	12
K001_011	3468	9

観察支援システムFishWatchr

