日本語言語資源の包括的高度利用環境の整備

日本語言語資源の包括的高度利用環境の整備

C2

コーパス開発センター

2016-2021年度（6か年）のコーパス開発センタープロジェクト

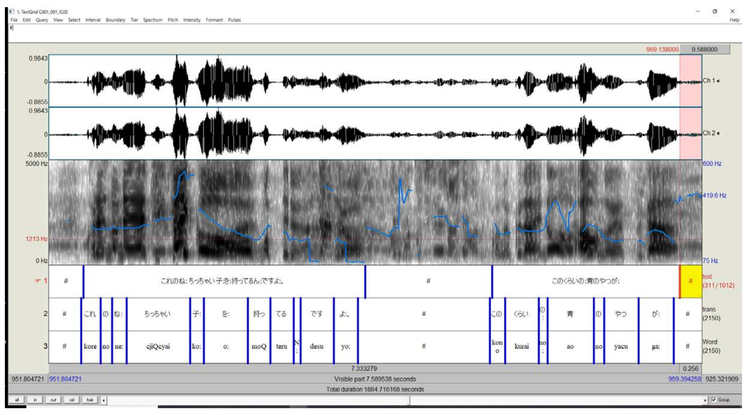
目的:
- 研究系のコーパス開発の支援のために必要な言語処理・音声処理・言語資源構築技術に関する共同研究を進める
- UniDic や分類語彙表などの語彙資源や、既存のコーパスに対するアノテーションの公開を行う
- 音声配信機能を含む統合検索環境の構築を行う

言語処理・音声処理・言語資源構築技術

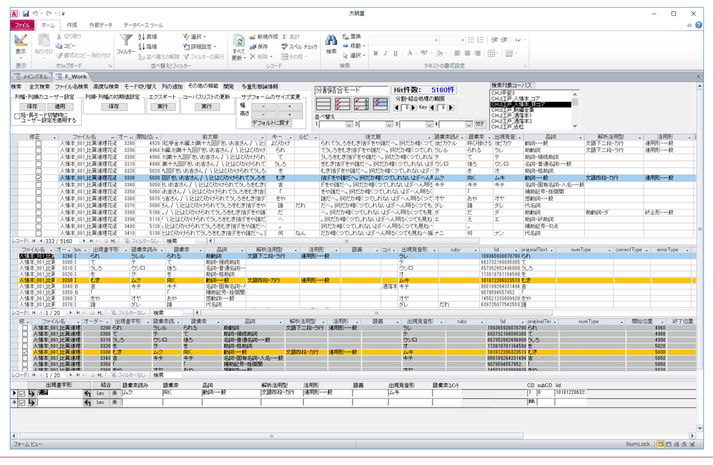
- 2テキスト間の対応付け技術

『日本語歴史コーパス』(CHJ) の古典⇔現代語訳
『日本語諸方言コーパス』(COJADS) の方言⇔共通語訳

- 音声と転記テキストとの対応付け技術



- 形態素解析用辞書『UniDic』とコーパスの協調整備環境「大納言」



語彙資源・アノテーションの整備

- 形態素解析用辞書『UniDic』の整備

現代書き言葉用 UniDic
現代話し言葉用 UniDic
古典用 UniDicS（時代別の形態素解析辞書）

- 『分類語彙表』の拡張

UniDic-分類語彙表番号対応表 (wlsp2unidic)
【開発中】反対語情報の整備
【開発中】代表義・プロトタイプ義情報の整備
【開発中】単語親密度の整備

- Universal Dependencies

70言語100以上のコーパスに対する言語横断統語情報付与
日本語 UD 言語資源の整備 (UD Japanese BCCWJ)

- 『分類番号表』関連データ

『現代日本語書き言葉コーパス』に対する分類番号付与
『日本語歴史コーパス』に対する分類番号付与
『日本語話し言葉コーパス』に対する分類番号付与
【開発中】機能語に対する用法アノテーション
反対語情報・単語親密度情報付与

統合検索環境の開発

複数コーパスの串刺し検索環境

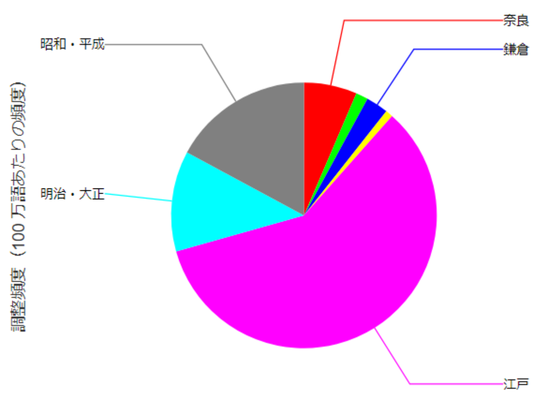
〔時代別〕

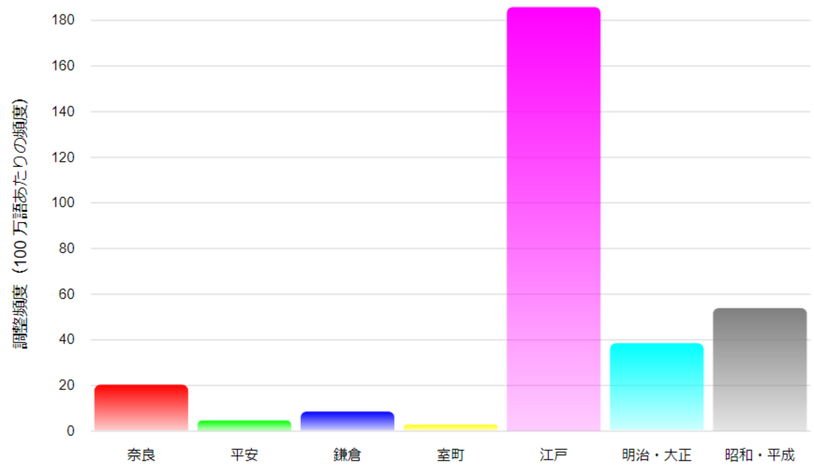
キー ---
書字形出現形 が 赤

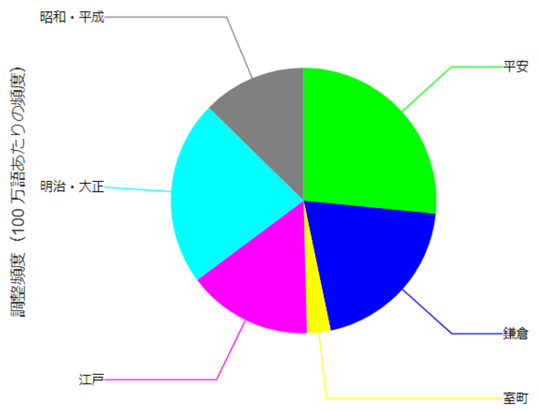
カテゴリ	コーパス	説明	検索結果の件数	検索対象語数	状態
奈良	CHJ	サブコーパス「奈良」のデータが対象	2	98,499	成功
平安	CHJ	サブコーパス「平安」のデータが対象	4	856,827	成功
鎌倉	CHJ	サブコーパス「鎌倉」のデータが対象	7	822,905	成功
室町	CHJ	サブコーパス「室町」のデータが対象	1	358,419	成功
江戸	CHJ	サブコーパス「江戸」のデータが対象	38	204,519	成功
明治・大正	CHJ	サブコーパス「明治・大正」のデータが対象	510	13,259,330	成功
昭和・平成	BCCWJ	全てのデータが対象	5,655	104,911,460	成功

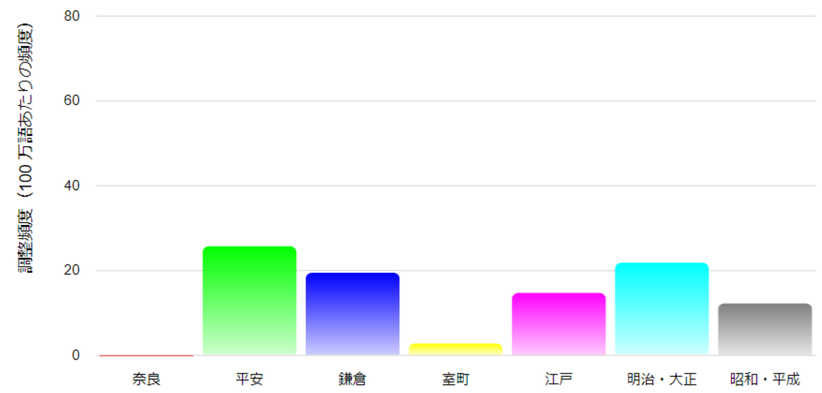
キー --- キーの条件を指定しない
書字形出現形 が 黄 条件を削除する 条件を追加する

カテゴリ	コーパス	説明	検索結果の件数	検索対象語数	状態
奈良	CHJ	サブコーパス「奈良」のデータが対象	0	98,499	成功
平安	CHJ	サブコーパス「平安」のデータが対象	22	856,827	成功
鎌倉	CHJ	サブコーパス「鎌倉」のデータが対象	16	822,905	成功
室町	CHJ	サブコーパス「室町」のデータが対象	1	358,419	成功
江戸	CHJ	サブコーパス「江戸」のデータが対象	3	204,519	成功
明治・大正	CHJ	サブコーパス「明治・大正」のデータが対象	289	13,259,330	成功
昭和・平成	BCCWJ	全てのデータが対象	1,281	104,911,460	成功









〔話し言葉・書き言葉／母語話者・学習者〕

キー ---
書字形出現形 が とても

カテゴリ	コーパス	説明	検索結果の件数	検索対象語数	状態
書き言葉	BCCWJ	全てのデータが対象	16,409	104,911,460	成功
	NWJC	一部のデータが対象	1,250	5,825,846	成功
話し言葉	CSJ	全てのデータが対象	3,158	7,576,046	成功
	名大会話コーパス	全てのデータが対象	60	1,131,971	成功
	職場コーパス	全てのデータが対象	7	186,906	成功
学習者の書き言葉	I-JAS	タスク「SW1、SW2」のデータが対象（日本語母語話者は除く）	163	115,322	成功
学習者の話し言葉	I-JAS	タスク「ST1、ST2、I、RP1、RP2、D」のデータが対象（日本語母語話者は除く）	3,430	1,925,558	成功