『日本語歴史コーパス』

- 『日本語歴史コーパス』は、デジタル時代の日本語史研究の基盤とすることを目標に国立国語研究所で構築・公開中の言語データベースです。日本語史研究だけでなくさまざまな関連分野で利用可能です。
- ユーザー登録が必要ですがWeb上のコーパス検索アプリケーション「中納言」上で無料で利用できます。
- 全てのテキストに読み・品詞などの単語情報が付与されているため高度な検索や集計などが行えます。
- 底本や原文画像などにリンクしており当該箇所の現代語訳や原本画像を確認することができます。

構築計画と進捗

奈良時代～8c末	☑万葉集　□宣命	
平安時代～12c末	☑仮名文学	□和歌
鎌倉時代～14c中	☑説話・随筆　☑日記・紀行　□軍記	
室町時代～17c初	☑狂言　☑キリシタン資料	
江戸時代～1868	☑洒落本　□人情本　□近松	
明治・大正～1945	☑雑誌　□教科書　□文学作品　□新聞	

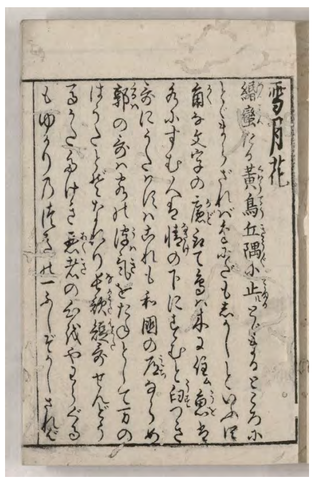
※2016～2022年度の6年間で構築・公開する予定のもの
□は着手済み、☑は公開済み

構築済みコーパス語数

		短単位	長単位
奈良時代編Ⅰ万葉集		10万	9.5万
平安時代編（「源氏物語」等16作品）		101万	91万
鎌倉時代編	Ⅰ説話・随筆	84万（うち非コア※41万）	79万（うち非コア39万）
	Ⅱ日記・紀行	11万	10万
室町時代編	Ⅰ狂言	28万	21万
	Ⅱキリシタン資料	14万	13万
江戸時代編Ⅰ洒落本		22万	―
明治・大正編Ⅰ雑誌		1400万（うち非コア1344万）	―

※非コアとは自動解析のみで単語情報の人手修正を経ていないデータ

コーパスの構築手順



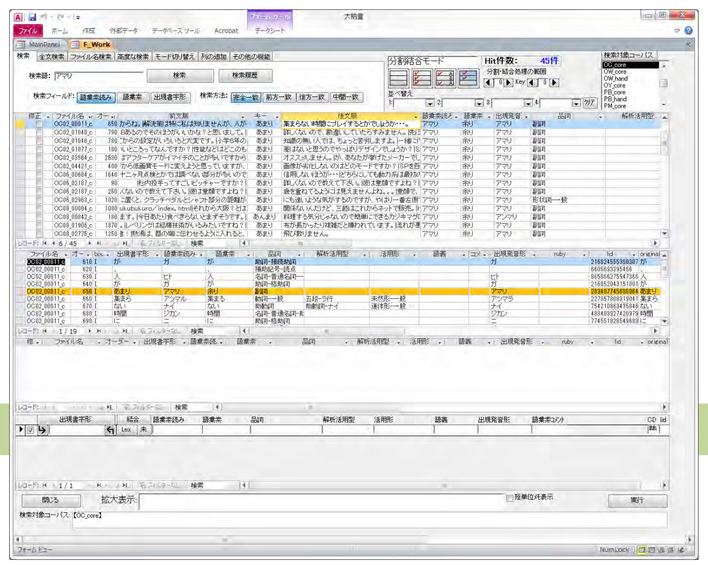
①（翻字・テキスト化）
原本からテキストデータを作成

②文書構造タグ付け
テキストデータに文書構造やフリガナ、話者情報などをタグ付け

```
1 ktext textID="虎明本狂言 035 大名 鷹盗人" series="虎明本狂言・大名狂言之類V上#35"
  title="鷹盗人" year="1642" year_w="寛永19">↓
2 <front>↓
3 <titleBlock><block type="title"><s><pb n="177"/></lb/><info originalPage="256"/>
  鷹盗人</s>↓
4 </block></titleBlock></front>↓
5 <body>↓
6 <article>↓
7 <p>↓
8 <stage><s><lb/>「名乗<char script="カタカナ">は</char><hi rend="傍線">入間</hi>
  川の<vMark>ご</vMark>とく下人よび出し</s>↓
9 </stage><speech><s><speaker value="大名"/>「追付て国本へ下らふが、さりなが</lb/>
  ら、<ruby resp="校注者" rubyText="この">此</ruby><ruby resp="校注者" rubyText="
  ぢゆう">中</ruby>某が事に色々<ruby resp="校注者" rubyText="肝">きも</ruby>を<rub
  y resp="校注者" rubyText="煎">い</ruby>らせられて、御<corr originalText="地" typ
  e="erratum" resp="annotator">馳</corr>走なされたる衆を、<ruby resp="校注者" ruby
  Text="振">ふる</ruby></lb/><ruby resp="校注者" rubyText="舞">まふ</ruby>て下りた
  ひと思ふが、何とあらふぞ</s>↓
0 </speech><speech><s><speaker value="太郎冠者"/>「誠におの／＼の<ruby resp="校注
```

③形態素解析
独自に開発した電子化辞書（UniDic）を用いて自動で単語の情報をタグ付け

④「大納言」DB上で修正
データベース上で電子辞書の見出しと関連づけながらコーパスのデータを修正



「中納言」による公開

- Web上で形態論情報を使用した高度な検索を実現

Web上の資料へのリンク

「中納言」検索結果から底本や原文の当該箇所にリンク

- **現代語訳等へ**
  - JapanKnowledge 小学館「新編日本古典文学全集」
- **原文画像へ**
  - 今昔物語集（京大・実践女子大）
  - 洒落本（早大・東大・阪大・国文研ほか）
  - 明六雑誌・国民之友（国語研）
  - 太陽（JKBooks）　etc.

https://pj.ninjal.ac.jp/corpus_center/chj