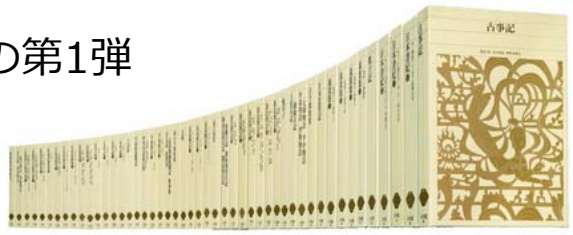
日本語歴史コーパス 平安時代編 の構築・公開

小木曽 智信　言語資源研究系・コーパス開発センター

・将来的に上代から近代までをカバーする「日本語歴史コーパス」の第1弾
・「通時コーパスの設計」プロジェクト（近藤泰弘リーダー）で構築
　「平安時代編」担当プロジェクト研究員：冨士池優美・鴻野知暁
・本文は小学館「新編日本古典文学全集」による
・全てのテキストに読み・品詞などの形態論情報を付与
　「中古和文UniDic」による自動解析後，誤りを人手修正済み



階層化された形態論情報により、必要なレベルでの検索・集計が可能。

・語彙素（辞書見出し）
　・語形（異語形）
　　・書字形（異表記）

「現代日本語書き言葉均衡コーパス」や「明六雑誌コーパス」と互換性あり。

収録作品・語数（太字は新規追加分）

作品名	短単位語数	長単位語数
古今和歌集	32540	**30550**
竹取物語	12750	**11880**
伊勢物語	16060	**15280**
大和物語	26940	**25010**
土佐日記	8210	**7810**
落窪物語	69170	**63040**
枕草子	80880	**75040**
源氏物語	520040	**462500**
和泉式部日記	12740	**11930**
紫式部日記	21040	**18600**
平中物語	**15100**	**14470**
堤中納言物語	**19660**	**18020**
更級日記	**17060**	**15950**
讃岐典侍日記	**18810**	**16600**
合計	*871000*	*786680*

・2012年12月：先行公開版（10作品の短単位データ）をWebインターフェイス「中納言」で公開済み
・2014年 3月：完成版（14作品の短単位・長単位データ）を「中納言」で公開予定

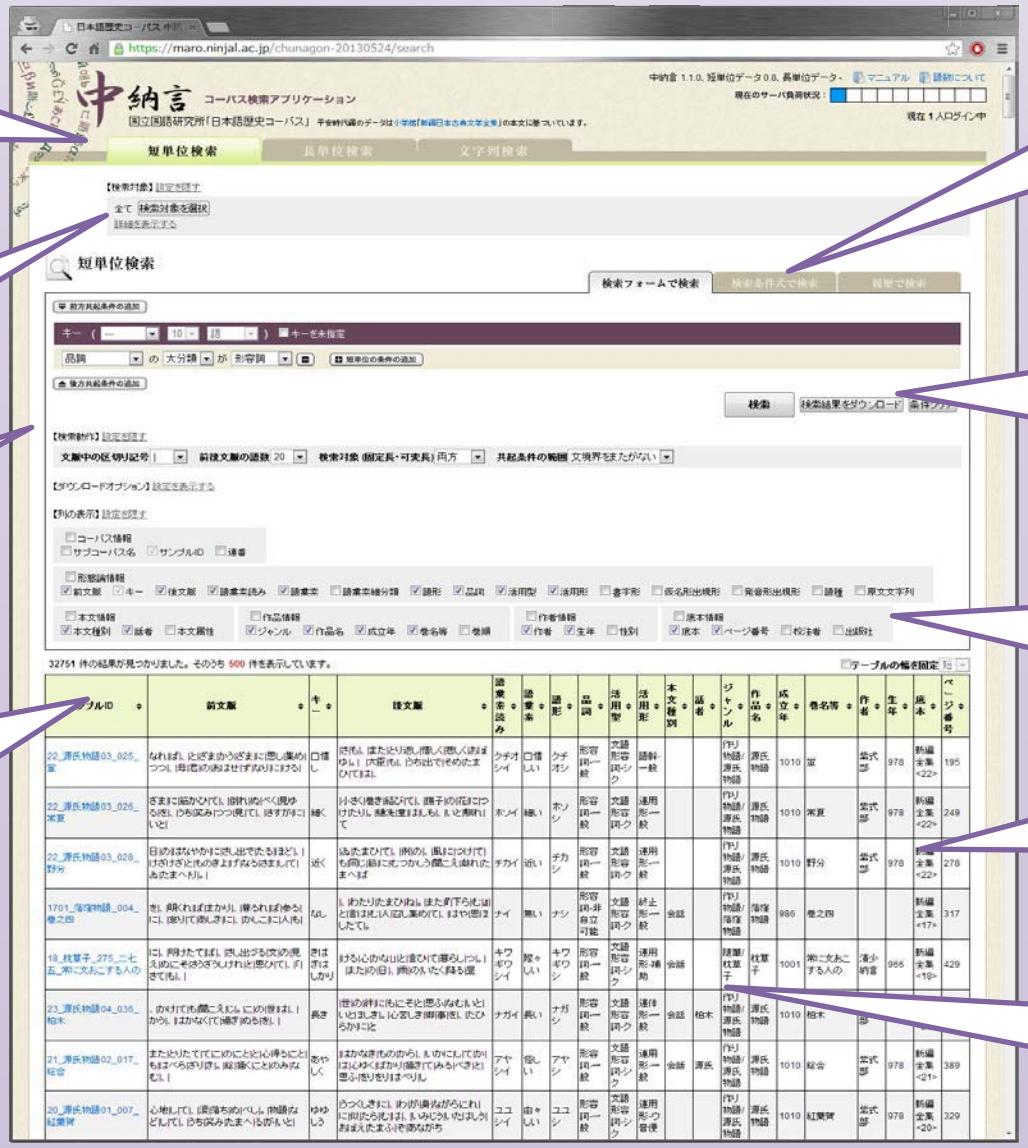
日本語歴史コーパス「中納言」（ Webインターフェイス ）

文字列検索と短単位検索・長単位検索に対応。
（長単位は3月以降）

作品別・ジャンル別などの検索対象指定が可能。

短単位の共起条件を前後計10単位まで指定できる。

検索結果を形態論情報と出典情報付きのKWIC形式で表示。
列見出しのクリックでソートも可能。



「検索フォームで検索」のほかに、
「検索条件式で検索」
「履歴で検索」が可能。

検索結果の全例（最大10万件）をテキスト形式でダウンロードできる。
（集計はダウンロード後Excelなどで）

検索対象や文脈の長さ、表示する項目（形態論情報・出典情報）を選択可能。

「底本情報」として“新編全集”の巻・ページ数等を表示。

「本文情報」として
会話文・地の文・和歌の区別、発話者、歌番号等を表示。

http://www.ninjal.ac.jp/corpus_center/chj/