

国立国語研究所学術情報リポジトリ

精緻な表記情報を付与した近世版本コーパスの構築 とその展開

メタデータ	言語: Japanese 出版者: 公開日: 2021-06-25 キーワード (Ja): キーワード (En): 作成者: 間淵, 洋子 メールアドレス: 所属:
URL	https://doi.org/10.15084/00003411



精緻な表記情報を付与した近世版本コーパスの構築とその展開

間瀬 洋子（言語変化研究領域 特任助教）

1. 私の仕事をざっくり言うと

・『日本語歴史コーパス』を作っています

- ▶ 奈良時代から明治・大正まで幅広い時代の多種多様な資料を収録 → 現代日本語へと連なる「**口语資料**」が充実
- ▶ すべての時代に共通の言語単位を用い、詳細な単語情報を付与 → 日本語の時代間比較、通時的変化研究が可能

・近世版本コーパスの構築を担当しています

- ▶ 『日本語歴史コーパス』構成資料の一つ。発話を主体とし、**江戸時代後期の話し言葉**を反映した**人情本**を収録
- ▶ **版本を底本**とし、忠実な翻刻を経てコーパス本文を作成。言語・表記・書物の重層構造を記述したコーパスへの発展を目指す

2. 近世版本コーパスの特徴

・精緻な表記情報を持つ

- 連綿の切れ目位置の情報
- 変体仮名の字体情報
 - 仮名の元になった字母の情報
 - Unicode (国際符号化文字集合) コードポイント

Diagram illustrating the relationship between characters and their original letters (kanji). It shows examples for '者' (者), '奈' (奈), 'え' (え), and '多' (多). Each character is linked to its original letter with red arrows. Below the diagram are three buttons: '字体' (Font), '連綿' (Connected), and '単語' (Word).

Callout bubble: 言語単位と書記単位の相関 仮名文字違いなど の研究に使います！

・図版情報

- 画像メタ情報
- タイトル
- キャプション
- 図版内容情報 など
 - 〔名前、服装、髪型、持ち物など〕

Diagram illustrating the relationship between characters and their original letters. It shows examples for 'キャラクション' (Character) and 'タイトル' (Title). Below the diagram is a callout bubble: 登場人物の造形や場面 の理解に必要な情報が含まれるので大事です！

5. まとめと展望

・まとめ

- ▶ 単語と表記や書籍の形態の関係性を解明するための、近世版本コーパスを構築中。作るのは結構大変&課題も多い。
- ▶ 応用コンテンツは、変体仮名を含めた文字認識技術の向上や一般向けのくずし字学習、前近代資料や古典に親しむための窓口として利用価値を持つ。

・今後の展望

- ▶ 『日本語歴史コーパス 江戸時代編Ⅱ人情本』は今年度末（2019年3月）に、8作品96巻（約50万短単位）を公開予定。
- ▶ 表記・図版情報付きコーパスは、このうちの2作品6巻分について上記と合わせて再構築し、応用コンテンツと共に2019年度中に公開予定。

3. 情報付け（アノテーション）の実践

・連綿（文字の切れ続ぎ）の判定

… 2名の判定者の結果を突合させて一致を見る

- 判定に搖れが出るところを調整



・変体仮名字体の判定

… 1名の判定者 + 結果の総点検

- 判定の難しい文字

音節	現行 仮名	変体 仮名	弁別の観点
「し」と「し」(U+1B045)	し	し	右方向への運び
「な」と「あ」(U+1B082)	な	あ	3画目の点
「や」と「や」(U+1B0DD)	や	や	「ゆ」に近い形
「ゆ」と「ゆ」(U+1B0E5)	ゆ	ゆ	中央のとめ
「る」と「つ」(U+1B0FB)	る	つ	上の横画
「を」と「ど」(U+1B11C)	を	ど	2画目の角度
「て」と「そ」(U+1B073)	て	そ	
「も」と「も」(U+1B0DA)	も	も	

- 異なる字体を同一視する必要がある場合も（包摂）

「毛」を字母とする変体仮名「モ」(U+1B0DE) ...



「本」を字母とする変体仮名「ボ」(U+1B0C0) ...



4. コーパスの応用

・変体仮名字形データベース

- ▶ 字母とUnicodeで層別した変体仮名字形のデータベース
- ▶ 字形画像と本文画像を相互にリンク

・前近代書物に親しむWebコンテンツ

- ▶ LINEトーク風表示による会話記述で楽しく
- ▶ ライトノベル風現代語訳付きで分かりやすく
- ▶ 原文画像を有効活用してビジュアル的に

人情本 Light

The screenshot shows a database interface for hiragana variants (な, ふ, 那) with their corresponding Unicode codes (U+306A, U+1B081, U+1B084) and various font samples. Below this is a conversational interface with characters like Dan and Nao, using LINE-style bubbles to show dialogue and modern Japanese translations of the original text.