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

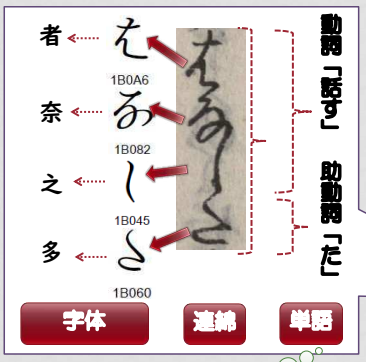
間淵 洋子（言語変化研究領域　特任助教）

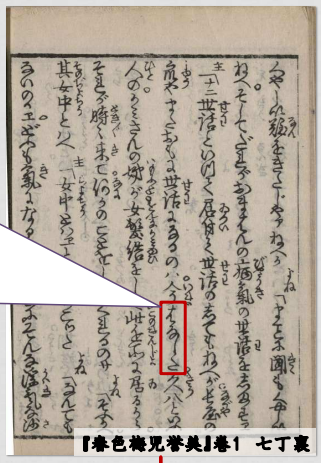
1．私の仕事をざっくり言うと

- **『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「**口語資料**」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





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！

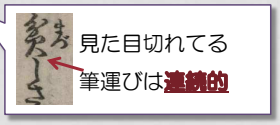
- **図版情報**
  - 画像メタ情報
  - タイトル
  - キャプション
  - 図版内容情報　など

名前，服装，髪型，持ち物など

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！

3．情報付け（アノテーション）の実践

- **連綿（文字の切れ続き）の判定**
  … 2名の判定者の結果を突合せて一致を見る
  - 判定に揺れが出るところを調整



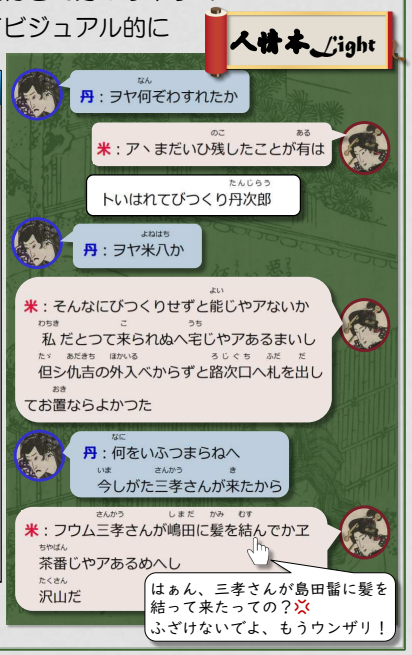
- **変体仮名字体の判定**
  … 1名の判定者 ＋ 結果の総点検
  - 判定の難しい文字
    - 「し」と「𛀁」(U+1B045)
    - 「な」と「𛀂」(U+1B082)
    - 「や」と「𛀍」(U+1B0DD)
    - 「ゆ」と「𛀥」(U+1B0E5)
    - 「る」と「𛀻」(U+1B0FB)
    - 「を」と「𛁜」(U+1B11C)
    - 「て」と「𛁳」(U+1B073)
    - 「も」と「𛀦」(U+1B0DA)

音価	現行仮名	変体仮名	弁別の観点
し			右方向への運び
な			３画目の点
や			「ゆ」に近い形
ゆ			中央のとめ
る			上の横画
を			２画目の角度

  - 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「𛀦」(U+1B0DE) … 
    「本」を字母とする変体仮名「𛀰」(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



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．

本研究は，人間文化研究機構広領域連携型基幹研究プロジェクト「異分野融合による「総合書物学」の構築」
国語研ユニット「表記情報と書誌形態情報を加えた日本語歴史コーパスの精緻化」による成果の一部です