『日本語歴史コーパス』の誤り修正プラットフォームの開発

小木曽 智信（人間文化研究機構 国立国語研究所）
八木 豊（株式会社 ピコラボ）

概要：国立国語研究所で開発・公開されている『日本語歴史コーパス』は多くの研究者に利用されているが，多くのテキストが収録されているため，その形態論情報には誤りが含まれている．この誤りをコーパスの利用者が報告し，研究者コミュニティ全体でコーパスの精度向上を図るためのプラットフォームとなるシステムを開発した．

キーワード：コーパス，形態論情報，共同研究プラットフォーム，オープンサイエンス

Development of a platform for error correction of the Corpus of Historical Japanese

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics, NIHU)
Yutaka Yagi (Picolab Co., Ltd.)

Abstract: The "Corpus of Historical Japanese" developed and published by the National Institute of Japanese Language and Linguistics is used by many researchers, but since it contains many texts, its morphological information contains errors. We have developed a system that allows users of the corpus to report these errors and provides a platform for the entire research community to improve the accuracy of the corpus.

Keywords: Corpus, Morphological Information, Collaborative Research Platform, Open Science

１．はじめに

『日本語歴史コーパス』（CHJ）[1]は，デジタル時代における日本語史研究の基礎資料として，国立国語研究所で開発・公開されているコーパスである．奈良時代以前から明治・大正時代までの多様なテキストを収録している．その全てのテキストに読み・品詞などの単語情報（形態論情報）が付与されていることが特長で，これにより高度な検索や集計が可能になっている．

このコーパスはオンラインのコーパス検索アプリケーション「中納言」を通して公開されているが，現在登録ユーザー数が約 2 万人，年間の検索クエリ数が約 45 万件，このコーパスを利用した研究発表（論文・全国大会予稿集）数が年間に約 100 本と，日本語の歴史研究の分野では研究に欠かせないインフラとして機能しつつある．

一方で，膨大な量のテキストの全てに形態論情報を付与しているため，その精度には限界がある．この精度を高めるために，コーパスのユーザーが誤りを報告し，その報告を相互にチェックする仕組みを導入することができれば，クラウドソーシングによってコーパスの精度向上を図ることができることになる．本論文は，このような誤り修正の機能を「中納言」に付与するためのシステムの開発を報告するものである．

２．コーパス中の誤り

『日本語歴史コーパス』は，コーパスに収録した資料（サンプル）ごとに ID を付与し，原則としてその全文テキストを収録し，その全文に形態論情報を付与したものである．個々のサンプルには，出典・著者等のメタ情報が付与されているほか，テキストの一部分には，その箇所が地の文なのか会話文なのか和歌なのかといった「文体種別」と呼ぶ情報や，それに対応した話者・歌人等の情報が付与されている．

コーパスに含まれるこれらの情報の誤りとして，小さな単位から順に見ると，おおよそ次のような段階が考えられる．

1. 文字（テキスト入力）の誤り
2. 形態論情報（形態素解析）の誤り
3. 範囲情報（文体種別等）の誤り
4. サンプルのメタ情報の誤り

いずれもコーパスにとって重要な情報であり，誤りは修正されるべきものであるが，本研究では次に示す理由から，このうちの 2.形態論情報に着目し，その誤り修正について扱うこととした．

『日本語歴史コーパス』には，各種の資料・作品を収録しているが，主として中世以前の資料は小学館『新編日本古典文学全集』をもととしてい

る．このデータは，株式会社小学館と国立国語研究所との契約の下で利用協定に基づいて提供を受けたもので，出版やウェブサービスで利用されている極めて信頼性の高い本文データを，ほぼそのまま用いている．そのため，これらの資料では，1.文字の誤りや，3.範囲情報の誤りは極めて少ない．また，4.サンプルのメタ情報については，1つの資料について限られた項目の内容を確認するだけで事足りるため，十分に修正が行き届く範囲の量である．

一方で，2.形態論情報については課題が残る．テキストを，各時代の資料にあわせた形態素解析用の辞書を用いて解析しているが，自動解析の精度は 95%程度に過ぎない．コーパスに収録したデータでは，自動解析結果に人手による修正を加えているものの，UniDic による非常に豊富な情報の全てについて修正を及ぼすことは難しい．

『日本語歴史コーパス』の形態論情報は，人手による十分な修正を加えたコアデータ（約 406 万語）と，自動解析に一定程度のしゅうせいをくわえただけの非コアデータ（約 1,354 万語）に分けられるが，コアデータにも一定程度の誤りは含まれている．

非コアデータでは，精度はおおよそ 98%程度に留まる．コーパスの形態論情報はテキストの全文に付与されているため，確認が必要なレコード数はコーパスの総語数であり，しかも 1 レコードが語彙素・語彙素読み・語種・品詞・活用型・活用形・書字形・発音形等の多数の情報を含むことから，非常に多くの誤りの可能性を含んでいる．

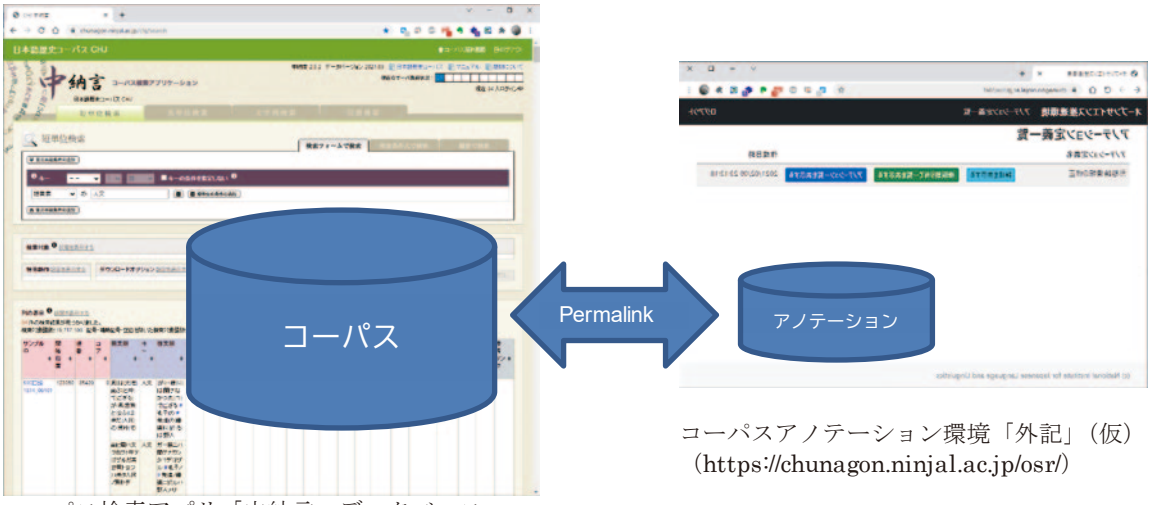
一般のテキストデータベースと異なり，このコーパスは言語研究に用いられるものであるため，形態論情報の利用頻度は極めて高い．そこで，本研究では，形態論情報の修正を主たるターゲットとした．

なお，『新編日本古典文学全集』以外をソースとする資料では，文字レベルの誤りや範囲情報の誤りも残されていると考えられる．こうした情報についても同一のプラットフォームでアノテーション共有の機能によって報告を行うことができる機能を付与する予定である．

3．システムの構成

先述したとおり『日本語歴史コーパス』の中世以前のデータの多くは，小学館「新編日本古典文学全集」に依拠しており，契約によりテキストの一般向けの提供は，前述したオンラインの検索アプリケーション「中納言」を通したものに限られている．そのため，全文データを配布することはできず，ユーザーとコーパスのデータとの接点は基本的にオンライン上の「中納言」を介したものに限られる．誤りの修正のみならず，コーパスを活用する上で，本文に情報（アノテーション）を付与したい場面は少なくないが，こうした情報付与も，基本的にユーザーとの接点となる「中納言」を利用して行う必要がある．

そのため，本稿の執筆者等は，「中納言」を介してユーザーがアノテーション情報を共有するシステム（コーパスへのアノテーションを核としたオープンサイエンス推進環境）を構想し，構築を進めてきた（科研費 20K20411）．このシステムは，コーパスの任意の箇所に任意の情報を付与できるように高い汎用性をもつものとして設計を進めてきたものであったが，本稿で提案する形態論情報修正の報告システムは，このシステムを一部拡張し，形態論情報修正に特化する形で実装し



コーパス検索アプリ「中納言」データベース
（https://chunagon.ninjal.ac.jp/chj）

コーパスアノテーション環境「外記」（仮）
（https://chunagon.ninjal.ac.jp/osr/）

※URL を共有するが別サーバー

図 1　システムの構成

図 2　「外記」アノテーション画面

たものである．

このコーパスアノテーション環境（「外記」と仮称）は，コーパスを格納する「中納言」のデータベーストは切り離し，サーバーも切り分けた，疎結合のシステムとして構築した．図 1 に示すように，両者は別のデータベース，別のインターフェイスをもち，「中納言」が提供する用例への permalink によって連携する．「中納言」のコーパスデータの直接の更新は行わず，「外記」の側はコーパスへのアノテーションデータを蓄積する形で情報を保持する．

４．コーパス中の位置情報と誤り修正

データベースの連携に用いる用例への permalink とは，「中納言」に格納されたデータベースの用例を一意に指し示すことのできる位置情報をもとにしたものである．これは，小木曽（2019）[2]で提案した形式であり，「中納言」の検索結果出力のうちの「サンプル ID」と「開始位置」の組み合わせにより，用例の箇所を一意に示す方法によっている．サンプル ID は収録資料に与えられた固有の ID であり，開始位置は，当該サンプル先頭からの文字のオフセット値を 10 倍した数字となっている．

佐々木（2021）[3]は『日本語歴史コーパス』の誤りを指摘し情報を報告する方法として，この位置情報を用いるについて提案を行っているが，この中で指摘のある「破（やぶ）る」を「破（や）る」とした誤りの例は，

20-源氏 1010_00047,156720

で指し示すことができる．サンプル ID「20-源氏 1010_00047」が『源氏物語』中の「総角」の巻であることを示し，開始位置「156720」はその 15672 文字目であることを示している．permalink は下記のようになる．

https://chunagon.ninjal.ac.jp/chj/permalink?unit=short&position=20-源氏 1010_00047,156720

「外記」はこの permalink を引数として与えることで，当該位置に対するアノテーションとして誤り修正情報を記述できる．後述するとおり，同じ長さの語でない場合には，1 つの語を複数に分割したり，複数の語を 1 つに統合したりして形態論情報を付与することも可能である．この場合，本文の文字列（出現書字形）をアノテーションデータに保持することで，本文中の開始位置・終了位置を保持している．開始位置はサンプル先頭からの文字数にもとづくものであるため，アノテーション対象の文字数が確定すれば，コーパス上での文字位置をユニークに特定できるからである．

５．アノテーションデータの形式

「外記」では，Web インターフェイス上で，図 2 に示す画面によって，コーパスの当該箇所の文脈と現在付与されている形態論情報を確認しながら，その部分に対してアノテーションを付与することができる．アノテーション自体は汎用的なものであるが，このとき，「形態論情報の修正」のアノテーションを選択して，現在の誤った内容に対して正しい内容を付与すれば，誤り修正としてのデータを報告することになる．

このときのアノテーション（誤り修正）のデータは，図 3 に示したように permalink に対する JSON 形式のデータとして，ユーザー名と更新日

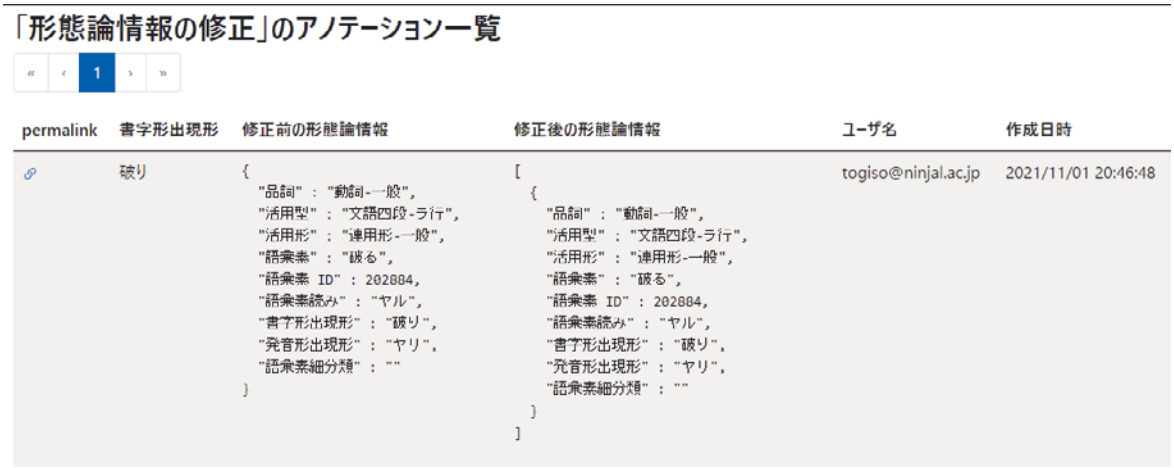


図 3　JSON 形式のアノテーション情報

時とともに記録される.

実際のコーパス中の誤りに対する修正データとして, いくつかのタイプ別の修正例を下記に示す.

1. 同じ長さの単語の修正

エスキモ人は旅行する時には, 雪小屋に泊り, 夏は	あざらし海豹	の皮でつくつたテントに寝ね, 冬は半ば地下室になつてゐる岩穴の

https://chunagon.ninjal.ac.jp/chj/permalink?unit=short&position=60M%E5%A4%AA%E9%99%BD1925_02073,250

修正前			修正後	
語彙素	語彙素読み	→	語彙素	語彙素読み
海豹	カイヒョウ		海豹	アザラシ

修正前：


```
{
  "品詞": "名詞-普通名詞-一般",
  "活用型": "*",
  "活用形": "*",
  "語彙素": "海豹",
  "語彙素 ID": 55635,
  "語彙素読み": "カイヒョウ",
  "書字形出現形": "海豹",
  "発音形出現形": "カイヒョー",
  "語彙素細分類": ""
}
```


修正後：


```
[
  {
    "品詞": "名詞-普通名詞-一般",
    "活用型": "",
    "活用形": "",
    "語彙素": "海豹",
    "語彙素 ID": 602,
    "語彙素読み": "アザラシ",
    "書字形出現形": "海豹",
    "発音形出現形": "アザラシ",
    "語彙素細分類": ""
  }
]
```


このように 1 対 1 で本文文字列が対応する場合は特段の問題なはい. ID と修正箇所以外にも, 辞書見出しとしてキーとなる情報を一律に持たせている.「出現書字形」が本文の文字列に相当する.

2. 2 語（以上）を結合して修正

議員の演説が濟むと議長の	カリー	ニンが其の採決をする, 擧手が凡そ三分の一位かと思つてゐると,

https://chunagon.ninjal.ac.jp/chj/permalink?unit=short&position=60M%E5%A4%AA%E9%99%BD1925_05028,23700

修正前			修正後	
語彙素	書字形出現形	→	語彙素	書字形出現形
カレー	カレー		カリーニン	カリーニン
ニン	ニン			

修正前：


```
{
  "品詞": "名詞-普通名詞-一般",
  "活用型": "*",
  "活用形": "*",
  "語彙素": "カレー",
  "語彙素 ID": 7342,
  "語彙素読み": "カレー",
  "書字形出現形": "カリー",
  "発音形出現形": "カリー",
  "語彙素細分類": "curry"
}
```


修正後：


```
[
```

```
  {
    "品詞": "名詞-固有名詞-人名-一般",
    "活用型": "",
    "活用形": "",
    "語彙素": "カリーニン",
    "語彙素 ID": 262785,
    "語彙素読み": "カリーニン",
    "書字形出現形": "カリーニン",
    "発音形出現形": "カリーニン",
    "語彙素細分類": "Kalinin"
  }
]
```


この場合は 2 対 1 で本文文字列が対応するため，修正前と修正後で「出現書字形」＝本文文字列の長さが異なっている．修正後に消えることになる「ニン」の情報は保持せず，修正後の出現書字形によって上書き部分を示すことになる．

3．1 語を分割して修正

程もあらせず表のかた「ハイ	今日は	．どふかお天気になればよふございますねへ．

https://chunagon.ninjal.ac.jp/chj/permalink?unit=short&position=53-%E4%BA%BA%E6%83%851832_06002,60700

修正前	
語彙素	語彙素読み
今日は	コンニチハ

→

修正後	
語彙素	語彙素読み
今日	コンニチ
は	ハ

修正前：


```
{
  "品詞": "感動詞-一般",
  "活用型": "'",
  "活用形": "'",
  "語彙素": "今日は",
  "語彙素 ID": 138112,
  "語彙素読み": "コンニチハ",
  "書字形出現形": "今日は",
  "発音形出現形": "コンニチワ",
  "語彙素細分類": ""
}
```


修正後：


```
[
  {
    "品詞": "名詞-普通名詞-副詞可能",
    "活用型": "",
    "活用形": "",
    "語彙素": "今日",
    "語彙素 ID": 13244,
    "語彙素読み": "コンニチ",
    "書字形出現形": "今日",
    "発音形出現形": "コンニチ",
    "語彙素細分類": ""
  },
  {
    "品詞": "助詞-係助詞",
    "活用型": "",
    "活用形": "",
    "語彙素": "は",
    "語彙素 ID": 29321,
    "語彙素読み": "ハ",
    "書字形出現形": "は",
    "発音形出現形": "ワ",
    "語彙素細分類": ""
  }
]
```


この場合は 1 対 2 で本文文字列が対応する．修正後の形態論情報は配列になっており，複数個の形態論情報に対応している．

このように，アノテーションデータは JSON 形式のデータとして保持され，誤り修正の場合も，コーパスのデータを直接修正するものではない．定期的にコーパス管理者側でデータを確認し，「中納言」側のデータをアップデートすることを計画している．

なお，誤り修正データにかかわらずこのような形で付与されるアノテーションデータはコーパスからは独立しており，本文の権利にも抵触しないため，自由に再配布ができる．そのため，アノテーションデータセットを GitHub に出力し，オープンデータとして共有可能にすることを計画している．

６．辞書引き機能の実装

上記のような形態論情報の修正にあたっては，個々の情報を人手で入力することは現実的ではない．形態素解析に用いた辞書の情報を参照しながら修正することがきわめて有効であるため，インターフェイスに辞書引き機能を付与することとした．この機能の追加が，汎用のアノテーション環境「外記」にとって，誤り修正機能に特化した拡張部分に相当する．

図 2 の画面上でアノテーション（修正）箇所を選択し，辞書呼び出しのボタンをクリックすると，当該の出現書字形に相当する見出し語（活用語であれば，出現形が一致する活用形）を一覧で表示することができる（図 4）．ここから選択することで，形態論情報を一々手作業で入力することなく，正しいものを選択することで入力することができるようにした．

図 4　辞書引き機能の利用例（「破り」の候補）

また，語の分割・結合を伴う辞書引きのためのユーザーインターフェイス（図 4）は，国立国語研究所内のコーパス構築用のシステム「大納言」[4]で用いられていたものに準拠した．これにより，コーパス構築に従事した経験者であれば容易に使いこなすことができるものとなっている．

7．誤り修正報告の評価機能

現在は実装途中であるが，先述の機能によってユーザーが行ったアノテーションに対して他のユーザーが評価を加えられる機能を開発中である．これによって，誤り修正の提案が正しいのかどうかをユーザー同士で評価し合うことが可能になる．

『日本語歴史コーパス』の総語数は，現時点で約 1760 万語あり，そのうち十分な人手修正を行っていない非コアデータが約 77%を占める．非コアデータの解析精度は 98%を下回ると見込まれるから，仮にコアデータに誤りが含まれないとしても，ざっと 27 万箇所の誤りが含まれていると考えられる．これだけのサイズのコーパスのメンテナンスを続けることは国立国語研究所にとっても極めて大きな負担である．ユーザーの誤り報告が行われるようになったとしても，相当多数の誤り報告が送られると見込まれ，その確認作業だけであっても，研究所側だけで行うことは不可能であると考えられる．

今回構築するシステムに相互評価機能を付与することによって，ユーザーが相互に誤り修正とそのチェックを行えるようにすることで，その結果をコーパスの定期的にアップデートに反映して精度向上に活かすことが可能になる．幸い，『日本語歴史コーパス』は研究者や学生に活発に利用されている．誤りの全てを修正することは現実的ではないが，重要な資料の一部だけでもこのようなユーザー間の相互扶助で精度向上を図ることができれば，このコーパスを学界が共有する財産として育てていくことが可能になると思われる．

7．おわりに

本発表では，『日本語歴史コーパス』の検索システム「中納言」に形態論情報の誤り報告機能を付加することで，コーパスの精度向上を図る試みについて報告した．現時点では公開前の試験段階であり，実運用上では多くの課題が発生することと思われるが，ユーザーの意見を取り入れつつ改善を重ねていくことを計画している．

今後，このシステムを運用することで，ユーザーが知見を持ち寄り，相互に確認を行うことを通して，学界でコーパスをともに育てていくプラットフォームとなることが期待される．これは，研究データを共有しお互いに活用していく日本語研究のオープンサイエンスの基盤として，さらなる発展の可能性を持つものであると考える．

参考文献

[1] 国立国語研究所(2021)『日本語歴史コーパス』（バージョン 2021.3） https://ccd.ninjal.ac.jp/chj/, (参照 2021-11-01)
[2] 小木曽智信（2019)『日本語歴史コーパス』への追加情報の付与と共有―中古和文の「る」「らる」を例に―,日本語学会 2019 年度春季大会予稿集
[3] 佐々木勇（2021）「「日本語歴史コーパス」修正点報告の提案」『日本語の研究』17-2.
[4] 小木曽智信, 中村壮範（2014）「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」.自然言語処理 21(2)

謝辞

本研究は，国立国語研究所共同研究プロジェクト「通時コーパスの構築と日本語史研究の新展開」の成果の一部であり，また JSPS 科研費 20K20411 の助成を受けたものです．