

OpenCHJ を活用した青空文庫データの公開 —国語教科書所収 6 作品を対象に—

木下 瞳、久保 柊子、呉 子凡（総合研究大学院大学）

謝 正科（一橋大学）

小木曾 智信（総合研究大学院大学／国立国語研究所）

Incorporating Aozora Bunko Data into OpenCHJ (the Open Corpus of Historical Japanese): Six Works from Japanese Kokugo Textbooks Used in Elementary through High School

Hitomi Kinoshita, Masako Kubo, Zifan Wu (The Graduate University for Advanced Studies, SOKENDAI)

Jian Kher SIAH (Hitotsubashi University)

Toshinobu Ogiso (The Graduate University for Advanced Studies, SOKENDAI / National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所（国語研）では、OpenCHJ と題して、国語研外部の研究者や日本語学研究に関心を持つ一般の人が、「中納言」で公開されている国語研のコーパスと同様に、インターネット上で資料を公開することができる環境の整備を進めている。

本稿では、青空文庫収録の 6 作品について単語情報を整備したデータを作成し、「中納言」で検索可能な形で公開した。データ整備の際には、Python を用いて青空文庫での XHTML 形式から形態論情報修正ツール「大納言」にインポート可能な XML に変換し、文字コードの変換も行った。データは GitHub でオープンデータとして公開した。

最後に、データの応用可能性の一例として、青空文庫所収の作品のうち国語教科書に収められている作品について、学年に見合ったレベルの語彙が使用されているかどうかや、各テキストの語彙的難度の検証が考えられることについて述べた。

1. はじめに

国立国語研究所（国語研）では、これまで国語研が中心として構築を行ってきた『日本語歴史コーパス』（CHJ）等に加え、多くの研究者が資料のデータを持ち寄る形でのコーパス構築を目指し、『OpenCHJ』の開発に取り組んでいる。本稿では、総合研究大学院大学の大学院生が講義内で行った、青空文庫¹収録の 6 作品のデータ整備の方法と OpenCHJ での公開について述べ、そのデータの応用可能性について述べる。

2. コーパスデータの構築

2.1 資料選定の背景

本研究で用いた小説テキストは、青空文庫に収録されている著作権保護期間が終了した作品を対象としている。青空文庫は、著作権が切れた作品や、著作者の許諾を得た作品を電子化し、学術研究や個人利用のために公開している。本研究では、学術研究を目的としてこれらのテキストを利用しており、著作権侵害の問題は生じない。本稿では、下記の 6 作品のデータ整備・公開について述べる。

¹ <https://www.aozora.gr.jp/>

表1 作品一覧

| 著者 | 作品名 | 成立年 | 青空文庫での底本 |
|-------|----------|------|-------------------|
| 芥川龍之介 | 羅生門 | 1915 | 芥川龍之介全集 1 (筑摩書房) |
| 芥川龍之介 | トロッコ | 1923 | 蜘蛛の糸・杜子春 (新潮社) |
| 太宰治 | 走れメロス | 1940 | 太宰治全集 3 (筑摩書房) |
| 中島敦 | 山月記 | 1942 | 李陵・山月記 (新潮社) |
| 宮沢賢治 | 注文の多い料理店 | 1924 | 注文の多い料理店 (新潮社) |
| 森鷗外 | 高瀬舟 | 1916 | 山椒大夫・高瀬舟他四篇 (講談社) |

これらは、大正から昭和初期に成立した小説で、現在の国語の検定教科書で採用されている作品である。OpenCHJ の最初の例として公開する資料は、日本語学・日本文学の研究対象として著名であるだけでなく、学校教育の場や一般の人にも親しみのある作品を取り上げるのが適切であると考え、これらの作品を選定した。

2.2 XML ファイルの作成

青空文庫に収録されている作品は、XHTML の形式でダウンロードが可能である。CHJ は XML の形でインポートする必要があるため、Python を用いて、XHTML ファイルを XML ファイルの形に変換した。テキストの範囲は、タイトル、作者名、本文のみとし、冒頭に付与された青空文庫テキストの凡例、末尾に付与された底本・入力者情報などは対象外とした。ファイル変換の際には、文字コードの変換を行ったほか、ルビタグをこれまでの CHJ の形式に合わせた。

| | |
|----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | 羅生門 |
| 2 | 芥川龍之介 |
| 3 | |
| 4 | ----- |
| 5 | 【テキスト中に現れる記号について】 |
| 6 | |
| 7 | 【】：ルビ |
| 8 | (例) 下人(げにん) |
| 9 | |
| 10 | ：ルビの付く文字列の始まりを特定する記号 |
| 11 | (例) 所々 丹塗《にぬり》 |
| 12 | |
| 13 | 【#】：入力者注 主に外字の説明や、傍点の位置の指定 |
| 14 | (数字は、JIS X 0213の面区点番号またはUnicode、底本のページと行数) |
| 15 | (例) ※【# 「てへん+丑」、第4水準2-12-93】 |
| 16 | ----- |
| 17 | |
| 18 | ある日の暮方の事である。一人の下人(げにん)が、羅生門(らしょうもん)の下で雨やみを待っていた。 |
| 19 | 広い門の下には、この男のほかに誰もいない。ただ、所々 丹塗《にぬり》の剥《は》けた、大きな門柱《まるばしら》に、蟻蜂《きりぎりす》が一匹とまっている。羅生門が、朱雀大路《すざくおおじ》にある以上は、この男のほかに、雨やみをする市女笠《いちめがさ》や採鳥帽子《もみえぼし》が、もう二三人はありそうなものである。それが、この男のほかに誰もいない。 |
| 20 | 何故かと云うと、この二三年、京都には、地震とか辻風《つじかぜ》とか火事とか饑饉とか云う災《わざわい》がつづいて起った。そこで洛中《らくちゅう》のさびれ方は一通りではない。旧記によると、仏像や仏具を打碎して、その丹《に》がついたり、金銀の箔《はく》がついたりした木を、路ばたにつみ重ねて、薪《たきぎ》の料《しろ》に売っていたと云う事である。洛中がその始末であるから、羅生門の修理などは、元より捨て捨て顧る者がなかった。するとその荒れ果てたのをよい事にして、狐狸《こり》が棲《す》む。盗人《ぬすびと》が棲む。とうとうしまいに、引取り手のない死人を、この門へ持ってきた。棄てて行くやうに出来た。そこで、日の目が見えなくなると、誰でも氣味を悪るが、この門の近所へは足がみせない事になってしまったのである。 |
| 21 | その代りまた鴉《からす》がどこからか、たくさん集って来た。昼間見ると、その鴉が何羽となく輪を描いて、高い鷗尾《しび》のまわりを啼きながら、飛びまわっている。ことに門の上の空が、夕焼けであかるい時には、それが胡麻《ごま》をまいたようにはっきり見えた。鴉は、勿論、門の上にある死人の肉を、啄《ついで》みに来るのである。一もつとも今日は、刻限《くげん》が遅いせい、一羽も見えない。ただ、所々、崩れかかった、そしてその崩れ目に長い草のはえた石段の上に、鴉の糞《ふん》が、点々と白くこびりついているのが見える。下人は七段ある石段の一番上の段に、洗いざらした紺の襦《あお》の尻を据えて、右の頬に出来た、大きな面皰《にきび》を氣にしながら、ほんやり、雨のふるのを眺めていた。 |
| 22 | 作者はさっき、「下人が雨やみを待っていた」と書いた。しかし、下人は雨がやんでも、格別どうしようも云う当てはない。ふだんなら、勿論、主人の家へ帰る可き筈である。所がその主人からは、四五日前に暇を出された。前にも書いたように、当時京都の町は一通りならず衰微《すいひ》していた。今この下人が、永年、使われていた主人から、暇を出されたのも、実はこの衰微の小さな余波にほかならない。だから「下人が雨やみを待っていた」と云うよりも「雨にふりこめられた下人が、行き所がなく、途方にくれた」と云う方が、適当である。その上、今日の空模様も少からず、この平安朝の下人の Sentimentalism に影響した。申《さる》の刻《く》下《さか》りからふり出した雨は、いまだに上るけしきがない。そこで、下人は、何をしても差当り明日《あす》の暮しをどうにかしようとして—云わばどうにもならない事を、どうにかしようとして、とりとめもない考えをたどりながら、さっきから朱雀大路にふる雨の音を、聞くともなく聞いていたのである。 |
| 23 | 雨は、羅生門をつつて、遠くから、ざあとう音をつめて来る。夕間は次第に空を低くして、見上げると、門の屋根が、斜につき出した雲《いらか》の先に、重たくうす暗い雲を支えている。 |
| 24 | どうにもならない事を、どうにかするために、手段を選んでいる連《いとま》はない。連んでいけば、築土《ついじ》の下か、道ばたの土の上で、餓死《うえじに》をするばかりである。そうして、この門の上へ持ってきた、犬のように棄てられてしまうばかりである。選ばないとすれば—下人の考えは、何度も同じ道を低徊《ていかい》した揚句《あげく》に、やっこの局所へ逢着《ほうちゃく》した。しかしこの「すれば」は、いつまでもたっても、結局「すれば」であった。下人は、手段を選ばないという事を肯定しながらも、この「すれば」のかたをつけるために、当然、その後に来る可き「盗人《ぬすびと》」になるよりほかに仕方がない」と云う事を、積極的に肯定するだけの、勇氣が出ずいたのである。 |
| 25 | 下人は、大きな嘘《くさめ》をして、それから、大籠《たいき》そうに立上った。夕冷えのする京都は、もう火桶《ひおけ》が欲しいほどの寒さである。風は門の柱と柱との間を、夕間と共に運慮なく、吹きぬける。丹塗《にぬり》の柱にまっていた蟻蜂《きりぎりす》も、もうどこかへ行ってしまった。 |

図1 変換前のテキストデータ

```

1 <?xml version="1.0" ?>
2 <text sampleID="芥川龍之介 羅生門">
3 <s>羅生門芥川龍之介ある日の暮方の事である。</s>
4 <s>一人の<rt="げにん">下人</rt>が、<rt="らしょうもん">羅生門</rt>の下で雨やみを待っていた。</s>
5 <s>広い門の下には、この男のほかに誰もいない。</s>
6 <s>ただ、所々<rt="にぬり">丹塗</rt>の<rt="は">は</rt>を<rt="まらばしら">円柱</rt>に、<rt="きりぎりす">蟋蟀</rt>が一匹とまっている。</s>
7 <s>羅生門が、<rt="すざくおおじ">朱雀大路</rt>にある以上は、この男のほかに、雨やみを<rt="いちめがさ">市女立</rt>や<rt="もみえほし">揉烏帽子</rt>が、もう三人はありそうなものである。</s>
8 <s>それが、この男のほかに誰もいない。</s>
9 <s>何故かと云うと、この二年、京都には、地震とか<rt="つじけさ">辻風</rt>とか火事とか饑饉とか云う<rt="わざわい">災</rt>がづついで起った。</s>
10 <s>そこで<rt="らくちゅう">洛中</rt>のさびれ方は一通りではない。</s>
11 <s>日記によると、仏像や仏具を打碎いて、その<rt="に">に</rt>丹</rt>がついたり、金銀の<rt="はく">はく</rt>箔</rt>がついたりした木を、路ばたにつみ重ねて、<rt="たきぎ">薪</rt>の<rt="しろ">しろ</rt>料</rt>に売っていたと云う事である。</s>
12 <s>洛中がその始末であるから、羅生門の修理などは、元より誰も捨てて顧る者がなかった。</s>
13 <s>するとその流れ果てたのをよい事にして、<rt="こり">狐狸</rt>が<rt="す">す</rt>棲</rt>む。</s>
14 <s><rt="ぬすびと">盗人</rt>が棲む。</s>
15 <s>どうとうしまいは、引取り手のない死人を、この門へ持って来て、棄てて行くと言ふ習慣さえ出来た。</s>
16 <s>そこで、目の目が見えなくなると、誰でも気味を悪るが、この門の近所へは足がみしな事になつてしまつたのである。</s>
17 <s>その代りまた<rt="からす">鴉</rt>がどこからか、たくさん集つて来た。</s>
18 <s>羅生門見ると、その鴉が何羽となく輪を描いて、高い<rt="しび">屋根</rt>のまわりを啼きながら、飛びまわっている。</s>
19 <s>ことに門の上の空が、夕焼けであかくなる時には、それが<rt="こま">胡麻</rt>をまいたようにはっきり見えた。</s>
20 <s>それは、勿論、門の上にある死人の肉を、<rt="ついは">威</rt>に來るのである。</s>
21 <s>—もつとも今日は、<rt="こげん">刻限</rt>が近いせいか、一羽も見えない。</s>
22 <s>ただ、所々、崩れかかった、そうしてその崩れ目に長い草のはえた石段の上に、鴉の<rt="ふん">糞</rt>が、点々と白くこびりついているのが見える。</s>
23 <s>下人は七段ある石段の一番上の段に、洗いざらした紺の<rt="あお">襦</rt>の尻を据えて、右の頬に出来た、大きな<rt="にきび">面皰</rt>を気にしながら、ぼんやり、雨のふるのを眺めていた。</s>
24 <s>作者はさつき、「下人が雨やみを待っていた」</s>
25 <s>と書いた。</s>
26 <s>しかし、下人は雨がやんでも、格別どうしよう云う当てはない。</s>
27 <s>ふだんなら、勿論、主人の家へ帰る可き筈である。</s>
28 <s>所がその主人からは、四五日前に暇を出された。</s>
29 <s>前にも書いたように、当時京都の町は一通りならず<rt="すいび">衰微</rt>していた。</s>
30 <s>今この下人が、永年、使われていた主人から、暇を出されたのも、実はこの衰微の小さな余波にほかならない。</s>
31 <s>だから「下人が雨やみを待っていた」</s>
32 <s>と云うよりも「雨にふりこめられた下人が、行き所がなく、途方にくれた」</s>
33 <s>と云う方が、適当である。</s>

```

図 2 変換後の XML テキストデータ

2.3 形態素解析

今回の対象作品はいずれも新字・新仮名遣いによって表記されているため、「現代書き言葉 UniDic」² (伝ほか 2007) と MeCab 0.996³ (Kudo et al. 2004) を用いて形態素解析を行い、コーパス用の短単位データを作成した。解析結果を国語研の「形態論情報データベース」(小木曾・中村 2014) にインポートできる形式とするため、Python を用いて文字表・短単位表・タグ表・ルビ表からなるデータを作成した。本文テキストのみを対象として s タグを文境界として MeCab による解析を行い、ふりがなの r タグも用いて上記の 4 つの表を生成し、それぞれの表をファイル頭からのオフセット位置によって関連付けている。

なお、この処理はオンライン上の「Web 茶まめ」⁴ (堤・小木曾 2023) で行うことができるように整備が行われている (堤・中村・小木曾 2023)。しかし、今回はタグセットの違いから独自のスクリプトによって解析を行った。

2.4 形態論情報の修正

上記の手順で作成したデータを、形態論情報修正ツール「大納言」(小木曾・中村 2014) にインポートし、手作業での形態論情報の修正を行った。参照した規定集は、『現代日本語書き言葉均衡コーパス形態論情報規程集上・下』である。

² https://clrd.ninjal.ac.jp/unidic/download.html#unidic_bccwj

³ <https://taku910.github.io/mecab/>

⁴ <https://chamame.ninjal.ac.jp/>

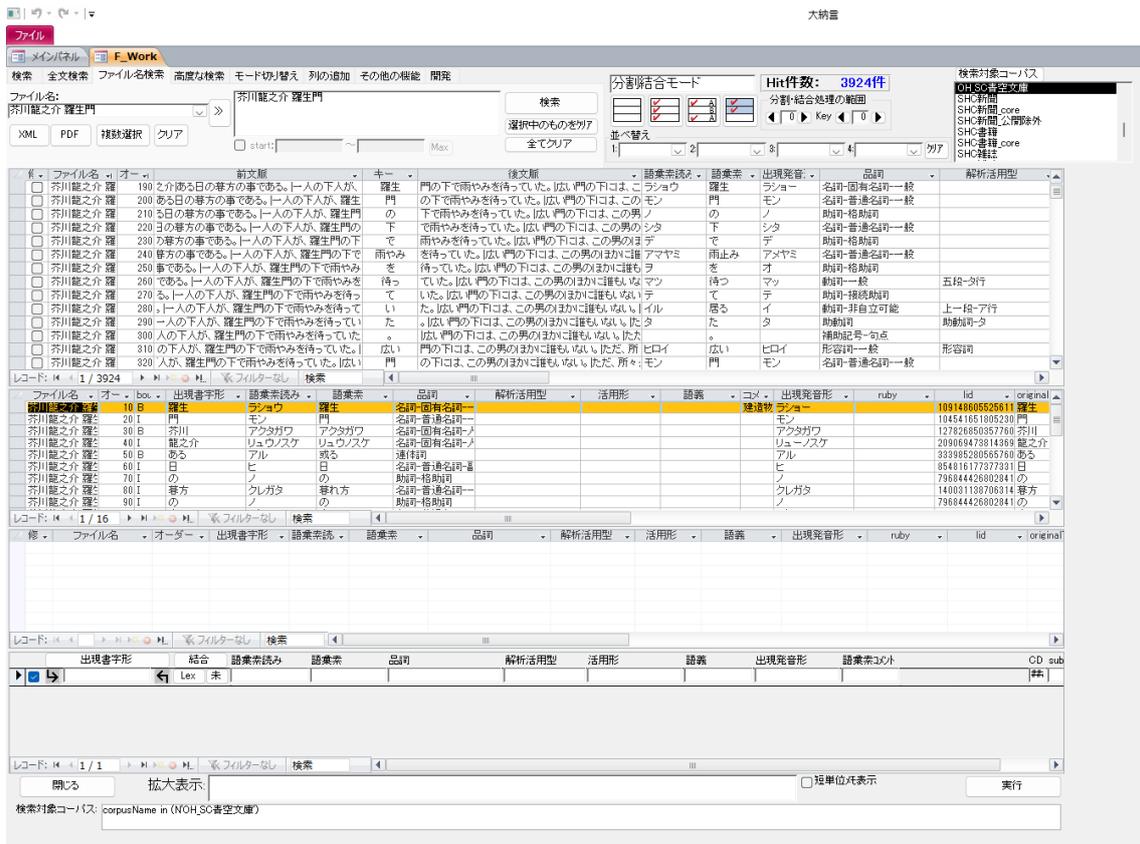


図3 「大納言」での修正画面

3. 中納言での表示

形態論情報を付与した青空文庫6作品は、コーパス検索アプリケーション「中納言」で検索・形態論情報、メタ情報等の表示が可能な状態となっている。

中納言では、「検索対象の選択」から、青空文庫作品全体や、個別の作品を選択することができる(図4)。

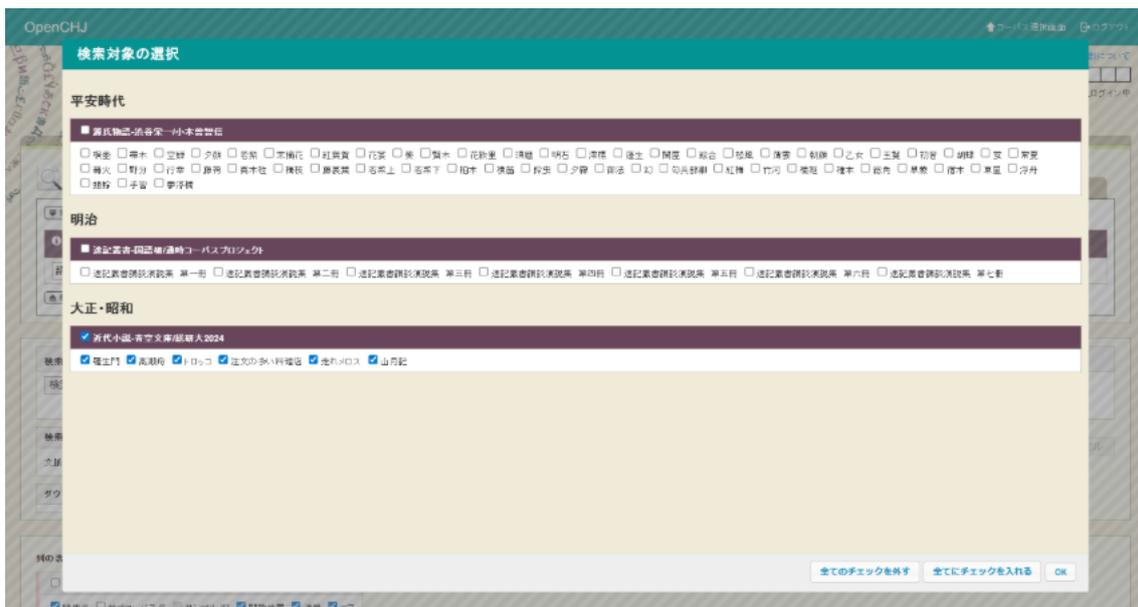


図4 「中納言」検索対象の選択

青空文庫テキストに付したタグや形態論情報は、図 5 のような形で検索結果として表示される。CHJ と主な仕様は共通しているが、一部異なる点について述べる。

8 件の検索結果が見つかりました。
検索対象語数: 26,311 記号・補助記号・空白を除いた検索対象語数: 22,458 検索対象サンプル数: 6

| 時代名 | サンプル ID | 開始位置 | 連番 | コア | 前文脈 | キ | 後文脈 | 語彙 | 語彙 | 語彙 | 品詞 | 活用型 | 活用形 | 原文文字列 | 振り仮名 | ジャンル | 作品名 | 成立年 | 巻名等 | 作者 | 生年 | 底本 | ページ番号 | テキストデータ | 形態論データ | |
|-----|---------------------|-------|-------|-----|-----------------------------------------------------------------------|---|--------------------------------------------------------------------------------------|----|----|----|-----------------|-----|-----|-------|------|------|---------|------|-----|-----------|------|-----------------|-------|---------|---------|------------|
| 8昭和 | 60N主れ 1940_11001 | 31940 | 21390 | 0.8 | に 行 わ れ た し、# 新 期 空 新 期 の 、 陣 々 の 宣 誓 が 満 ち た に ろ し、 黒 雲 が | 空 | を 強 い、 ぼ つ り ぼ つ り 雨 が 降 り 出 し、 や が て 車 輪 を 流 す よ う な 大 雨 と な っ た し、 | ソラ | 空 | ソラ | 名 詞 着 述 名 詞 一 般 | | | 空 | | 小説 | 走 れ ろ ス | 1940 | | 太 宰 治 | 1909 | 太 宰 治 金 集 3 | | | 青 空 文 庫 | 編 研 大 2024 |
| 7大正 | 60N羅生 1915_11001 | 5570 | 3800 | 0.8 | 届 いて、 黒 い 顔 毛 の ま わ り を 隠 さ な が ら 、 扱 ま わ っ て い る 。# こ こ に 戸 敷 の 上 の | 空 | が、 夕 陽 が 沈 み か か る 時 刻 に は、 そ れ が 顔 毛 を 隠 し、 ま わ り を 隠 さ な が ら 、 扱 ま わ っ て い る を 見 え | ソラ | 空 | ソラ | 名 詞 着 述 名 詞 一 般 | | | 空 | | 小説 | 羅 生 門 | 1915 | | 芥 川 龍 之 介 | 1892 | 芥 川 龍 之 介 全 集 1 | | | 青 空 文 庫 | 編 研 大 2024 |

図 5 「中納言」による検索結果の表示例

・ コア情報

CHJ では、コアデータであるか非コアデータであるかの 2 種のみであったため、コア情報は 1 か 0 で表示されていた。OpenCHJ では、形態論情報の修正状況や信頼度に合わせてコアの数値を分けることができるようになった。青空文庫では概ね信頼できる修正状況であるが、完全ではないと判断し、コアを 0.8 としている。

・ テキストデータ・形態論データ

OpenCHJ では、検索結果にテキストデータと形態論データの情報を示している。CHJ では底本情報と資料の参考リンクを示していた箇所である。これまでは、国語研が形態論情報を付与していたが、OpenCHJ では広く人々がデータを作成することとなる。元テキストの所蔵場所とともに、形態論情報付与者をこの箇所に示すことで、データ作成者の情報へのアクセスが容易となる。

4. データの公開

作成したデータは「中納言」での公開に加えてオープンデータとして GitHub 上で公開した⁵ (図 6)。形態論情報のライセンスは CC BY 4.0 とし、本文テキストの権利については青空文庫の記述⁶に従うものとした。

⁵ <https://github.com/togiso/OpenCHJ-Aozora>

⁶ <https://www.aozora.gr.jp/guide/kijyunn.html>

OpenCHJ-Aozora 形態論情報データ

青空文庫所収のテキストにUniDicによる短単位の形態論情報を付与し、誤りを修正したものです。

2025/03/03 作品一覧

このデータは[総合研究大学院大学日本語言語科学コース](#)の2024年度の授業「言語資源学演習1」の中で作られました。

| 作品名 | 作者 | 青空文庫・入力 / 校正 (図書カード) | 形態論情報整備・担当 / 副担当 | 形態論情報バージョン |
|----------|-------|-------------------------------------------|------------------------------|------------|
| 走れメロス | 太宰治 | 入力：金川一之 / 校正：高橋美奈子 | 小木曾智信 / 謝正科 | 0.8 |
| 羅生門 | 芥川龍之介 | 入力：野口英司・平山誠 / 校正：もりみつじゅんじ | 久保証子 / 木下瞳 | 0.8 |
| 高瀬舟 | 森鷗外 | 入力・校正：青空文庫 | 謝正科 / 呉子凡 | 0.8 |
| 注文の多い料理店 | 宮沢賢治 | 入力：土屋隆 / 校正：noriko saito | 木下瞳 / 久保証子 | 0.8 |
| 山月記 | 中島敦 | 入力：平松大樹 / 校正：林めぐみ | 呉子凡 / 小木曾智信 | 0.8 |
| トロッコ | 芥川龍之介 | 入力：蔣龍 / 校正：鈴木厚司 | 小木曾智信 / 久保証子 | 0.8 |

元のテキストデータの詳細は上記の青空文庫のリンク先（図書カード）を参照してください。

ファイル形式

- UTF-8 (BOMなし) LF改行, タブ区切り
- フィールド (左から)
 - ファイル名 (作者 作品名)
 - サブコーパス名
 - 開始文字位置 (ファイル頭からのオフセット値*10)
 - 終了文字位置 (同上)
 - 文境界 (B=文頭)

図 6 GitHub 上のデータ公開ページ (画面)

5. データの応用可能性

最後に、今回構築したデータの応用可能性について、日本語教育研究への応用という観点から述べる。今回対象とした青空文庫の作品は、小学校から高校までの国語教科書にも収められている著名な作品が多い。例えば、そのような作品を取り上げて、学年に見合ったレベルの語彙が使用されているかどうかを、各種語彙頻度表や教育基本語彙を参照しながら検討することが可能である。そして、学年に見合ったレベル以外の語彙が使用されていた場合、それらがどの程度使用されているか、また、それらの語にはどのような特徴があるのかなどを明らかにできる。

他にも、各テキストの語彙的難度を検証する一例として、青空文庫の各作品において、一定のテキストカバー率 (95%や 98%) に達するのに、各語彙表の頻度上位から何語必要かを調査することも可能である。そして基準から外れた語 (低頻度語) にはどのような語があるかを調査するために、オノマトペや意味的透明度の高い語 (白熊、横っ腹など) は理解できると仮定し、それらを除いた割合を見ることで、本当に難しい語は何なのかを明らかにできる。これらの研究課題については、今後取り組んでいく予定である。

6. 終わりに

本稿では、青空文庫のテキストを用いたオープンデータの作成・公開を通して、OpenCHJの使用例およびデータの応用可能性を示した。OpenCHJは、一般の方でも資料の公開が可能であり、データの活用法にもさまざまな可能性が考えられる。本発表で紹介したオープンデータを含め、今後多くのデータが公開されれば、研究や教育現場での幅広い応用が期待さ

れる。

謝 辞

本研究は国立国語研究所のプロジェクト「開かれた共同構築環境による通時コーパスの拡張」による成果である。

文 献

- 国立国語研究所 (2025). 『OpenCHJ』 <https://chunagon.ninjal.ac.jp/open-chj/> (2025年8月12日確認).
- 小椋秀樹・小磯花絵・富士池優美・宮内佐夜香・小西光・原裕 (2011) 『現代日本語書き言葉均衡コーパス』形態論情報規程集 第4版(上)(下).
- 小木曾智信 (2025). 「CHJ ver.2025.3 + OpenCHJ: 通時コーパス拡張進捗報告」 「通時コーパス」シンポジウム 2025.
- 小木曾智信・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的设计・実装・運用」自然言語処理, vol. 21, No. 2, pp. 301-332.
- 堤智昭・小木曾智信 (2023). 「複数の UniDic 辞書による形態素解析支援ツール『Web 茶まめ』の実装と運用」情報処理学会論文誌, Vol.64, No.3, pp.749-757.
- 堤智昭・中村壮範・小木曾智信 (2023). 「形態素解析ツール Web 茶まめによるコーパス作成支援」情報処理学会シンポジウム (じんもんこん 2023) 論文集, Vol.2023, No.31, pp.273-278.
- 伝康晴・小木曾智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007). 「コーパス日本語学のための言語資源: 形態素解析用電子化辞書の開発とその応用」日本語科学, Vol.22, pp.101-123. <http://doi.org/10.15084/00002185>
- Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. (2004) “Applying conditional random fields to Japanese morphological analysis”, In *EMNLP, Vol. 4*, pp. 230-237.