

文字単位の多対多自動アライメントを用いた 日本語歴史コーパスのルビアノテーションの自動修正 ver.1.2

岡 照晃[†]

[†] 国立国語研究所 コーパス開発センター

本論文では、国立国語研究所で整備している日本語歴史コーパスのルビアノテーション修正について述べる。現在のルビのアノテーションは国語研の規定する言語単位である短単位との整合がとれておらず、コーパス検索に不便である。そこで文字単位の多対多自動アライメント手法を使用し、ルビアノテーションをモノルビに修正する手法を提案する。この手法により、97.4%という高い正解率でルビアノテーションの修正が行えることが分かった。またこのアライメントのスコアを利用することで79%の適合率で近世の口語資料から当て字ルビの50%を検出できることが分かった。

Automatic Modification for Information of Ruby in Corpus of Historical Japanese using an Automatic many-to-many Character Alignment Teruaki Oka[†]

[†]Center for Corpus Development, National Institute for Japanese Language and Linguistics

In this paper, we describe modification of annotation of ruby tags in Corpus of Historical Japanese, that is creating in National Institute for Japanese Language and Linguistics. Since the tags are not corresponding with Short Unit Word unit, current ruby annotation is inconvenience for retrieving the corpus. Therefore, we introduce automatic many-to-many character alignment technique to modify the ruby annotation into mono-ruby annotation. By using this method, we achieved 98.9% of automatic modification accuracy. Furthermore, we confirm that we can discriminate “Ateji” rubes from rubes in historical documents written in spoken language of Edo era by utilizing the automatic alignment scores.

1 はじめに

現在、国立国語研究所では上代から近代にかけての日本語史資料を取めた日本語歴史コーパス(CHJ)¹の構築を行なっている[2]。CHJの大きな特徴として、短単位と呼ばれる齊一な言語単位で単語分割されていること、各短単位への詳細な形態論情報(発音、品詞、語彙素、語種...)が付与されていることがある[4]。短単位は認定基準が規程集で明確に定められており、アノテーション時の揺れも生じづらい。そのためコーパス検索に適しており、CHJでは通常の文

字列検索を超えて、さまざまな時代を横断した語の通時的变化を観測することが可能である。

CHJでは底本資料の本文だけでなく、本文に併記されたルビ情報のアノテーションも実施している。ルビ情報は全文検索システムひまわり[5]を使用することで検索可能であり、また2016年4月からはコーパス検索アプリケーション中納言²でもルビ情報の参照が可能となった。しかしあくまでルビは補助的な情報であり、一部を除き、本文テキストの短単位アノテーションとも整合がとられておらず、複数の短単位をまたぐ形でアノテーションされているルビも多い。例えば、図1は実際の中納言の検索結果であるが、「橋」という短単位に対し、「はし」とルビ振られているものもあるが、「みはし」や「ぼしあやふ」など、1つの短単位境界を越えたルビも見ら

本著作物の著作権は(社)情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。

¹http://pj.ninjal.ac.jp/corpus_center/chj/

²<https://chunagon.ninjal.ac.jp/>

れる。この状態のままではルビから短単位を検索したい場合、もしくはその逆を行う際の不便となる。また、ひまわりでのルビ文字列検索でも、 unnecessary 文字列まで検索結果に現れることや、必要な文字列が検索結果から漏れる恐れもある。

本研究では、現状の CHJ のルビアノテーションを修正し、本文テキストの短単位アノテーションと整合のとれるルビアノテーションの実現を目指す。コーパス中のルビの使用には、親文字もしくはルビが、カタカナ語やアルファベット文字列となっている場合もあるが、今回は親文字が漢字文字列で、それに対する音読み、訓読みの仮名文字ルビを対象とする。またルビアノテーション修正の基本方針として、各親文字1字にルビ情報を付与するモノルビを採用する。ただし「蝸牛 [かたつぶり]」のようにモノルビ化が困難な熟字訓に対してはグループルビを採用する³。また現状の CHJ 全体に対し、上述の修正をすべて人手で実施することのは人件費および時間が多大となる。そこで本研究では統計的機械学習による自動アライメント手法によって漢字1字ずつに対応するルビ文字列を自動判別する。その後、自動対応付けに失敗した箇所を人手で修正していく。自動アライメント手法には、文字単位の多対多アライメント手法 [3] を採用する。この手法は漢字の自動読み推定のために提案されたものであり、本研究の目的にも即した手法である。テストケースとして現在整備を行っているキリシタン資料エソポのハブラスに対し、この手法を使用した。その結果、97.4%という高い正解率でルビアノテーションの修正が行えることが分かった。ルビの修正に失敗した箇所も、親文字列側で「々」のような踊字を含む場合であったため、あらかじめ踊字を開いておくなどの前処理によって、さらに精度向上可能だとわかった。

また追加実験として、近世口語資料に頻出する当て字ルビの検出実験も行なった。近世の資料には「心驚 [びつくり]」「誘引 [さそはれ]」のように通用的でない特殊な振り仮名が多い。そのためこういった当て字に対し、自動アライメントは低いスコアを割り当てると予測できる。そこで現在整備中の近世の洒落本、人情本コーパスのルビアノテーションを対象にエソポのハブ

³以下、[] は直前文字列のルビを表す。

ラスと同様の方法で文字列アライメントを実施し、しきい値を使って各アライメントのスコアから当て字を検出する実験を行なった。その結果、79%の適合率で資料中の当て字ルビの50%を検出することができることが分かった。

2 JIS 規格の組版処理におけるルビの扱いと現状の CHJ のルビアノテーション

ルビは文字のそばに付けて文字の読み方、意味などを示す小さな文字である（以降、ルビ文字もしくはルビ文字列）。ルビ文字（列）を付与する際、その付与の対象となる文字を親文字という。ルビの組版処理は、JIS X 4051⁴において規定されており、親文字1字ごとに対応してルビ文字（列）を配置するモノルビと、2文字以上の親文字に対してルビ文字（列）を配置する熟語ルビおよびグループルビがある。熟語ルビは熟字訓でなければ、「凝 [ぎよう] (改行) 視 [し]」のように親文字列の途中で改行可能であるが、グループルビの場合、ルビ文字（列）の付与された親文字列を1つの塊として考え、親文字列中での改行はできない。

以上のような規則に対し現状の CHJ では、近代語のコーパスなど一部にはルビアノテーションに関する規程を設けているが、⁵ 基本的に、底本中で一定の空白を空けずに連続したルビ文字（列）を一まとめとし、それが併記されている本文文字列を親文字列とする、といった大雑把なアノテーションが実施されている。これは CHJ の大部分が小学館の新編日本古典文学全集を底本としており、小学館から直接提供された XML データのルビタグをそのまま使用しているためである。図2は、紙媒体の新編全集において実際に図1の「ばしあやふ」のルビが存在した箇所である。小学館の XML データは元々コーパス検索用に作成されたものではなく、紙のページの見た目を XML 形式でマークアップすることを目的としているため、上述のアノテーションで問題なかった。しかし検索を用途とする CHJ では、ルビのアノテーションにも何かしらの基準、少なくとも検索単位である短単位との整合が必要となる。

⁴<http://www.jisc.go.jp/app/pager?id=0&RKNP.vJISJISNO=X4051&\%23jps.JPSH0090D:JPS00020:/JPS/JPS00090.jsp>

⁵http://pj.ninjal.ac.jp/corpus_center/cmj/doc/07kondo.pdf

日本語歴史コーパス CHJ								
↑コーパス選択画面 ログアウト								
サンプル ID	前文脈	キー	後文脈	品詞	原文文字列	振り仮名	作品名	底本
20-大鏡 1100_01020	御 車 の 立ち やう など よ 。#尊者 の 御 車 をば 東 に 立て 、 牛 は 御	橋	の 平葱 柱 に つなぎ 、 こと上達 部 の 車 をば 、 河 より は 西 に 立て たる が めでたき を	名詞-普通名詞-一般	橋	みはし	大鏡	新編全集<34>
20-源氏 1010_00053	な ん と 思ふ も 、 惜 し か ら ぬ 身 なれ ど 、 例 の 心弱 さ は 、 一 つ	橋	危 がり て 帰 り 来 たり けん 者 の やう に 、 わ び しく お ぼ ゆ 。#こも ぎ 、 供 に 率	名詞-普通名詞-一般	橋	ばしあやふ	源氏物語	新編全集<25>
20-蜻蛉 0974_00003	かい 忍 び やか なれ ば 、 よ ろ づ に つ け て 涙 も ろ く お ぼ ゆ 。#そ の 泉 川 も 渡 り て 。#	橋	寺 と い ふ と こ ろ に と ま り ぬ 。#酉 の 時 ば かり に 降 り て 休 み た れ ば 、 旅 籠 と こ ろ	名詞-普通名詞-一般	橋	はしてら	蜻蛉日記	新編全集<13>
30-今昔 1100_12011	仏 道 を 行 ふ 。#其 の 時 に 、 其 の 郡 の 枇 花 の 郷 に 一 つ の	橋	有 り 。#其 の 橋 の 本 に 梨 の 木 を 伐 て 曳 き 置 て 、 年 来 を	名詞-普通名詞-一般	橋	はし	今昔物語集	新編全集<35>
30-今昔 1100_12011	の 時 に 、 其 の 郡 の 枇 花 の 郷 に 一 つ の 橋 有 り 。#其 の	橋	の 本 に 梨 の 木 を 伐 て 曳 き 置 て 、 年 来 を 経 たり 。#其 の	名詞-普通名詞-一般	橋	はし	今昔物語集	新編全集<35>

図 1: コーパス検索アプリケーション中納言の検索結果の例。語彙素「橋」で短単位検索した結果の一部（2016/11/1の検索結果）。各行が1短単位を表し、各短単位のルビ情報は「振り仮名」列に表示されている。また「前文脈」および「後文脈」列には検索目的の短単位（キー列）の前後の短単位列がそれぞれ表示されており、コンコーダンサ的な表示となっている。

そこで本研究では、コーパス中のルビアノテーションを熟字訓を除いて、すべてモノルビとし、熟字訓に対しては、熟語ごとのグループルビに修正する。このように可能な限り細かい粒度でのルビアノテーションを実現すれば、CHJ本文テキストの短単位アノテーションとの整合も容易となる。

3 文字単位の多対多自動アライメント手法に基づくルビアノテーションの自動修正

CHJのルビアノテーションをすべて人手で修正することは困難であるため、本研究では、文献[3]の文字単位の多対多自動アライメント手法を使用し、ルビアノテーションの自動修正を試みる。文献[3]の手法は、漢字の読み（発音）推定のために開発された文字列アライメント手法である。漢字列とその読み文字列を並べたパラレルコーパスを入力として、与えられた各漢字文字列とその読み文字列の最小の多対多アライメントを両者の共起関係に基づいた教師なし学習により

獲得する。本研究では、文献[3]の手法を実装したツール mpaligner⁶を使用し、読みの代わりにルビを入力することで、漢字文字列とルビ文字列の対応をとる。漢字とルビの共起に基づいた最小のアライメントを獲得するため、一定量以上の漢字列とそのルビ文字列のペアがあれば、特殊な共起関係である熟字訓を除き、各漢字にルビを対応させるモノルビが実現でき、同時に熟字訓はグループルビとして、アライメント結果に取得可能だと考えられる。

4 ルビアノテーション自動修正の性能評価実験

ルビアノテーションの自動修正の性能評価実験として、キリシタン資料エソポのハブラスを対象に、ルビアノテーションの自動修正を実施した。この資料は現在コーパス化作業の途中であり、大英図書館蔵（1593）を底本に、ローマ字本文のテキスト化の段階から作業を開始してい

⁶<https://osdn.jp/projects/mpaligner/>

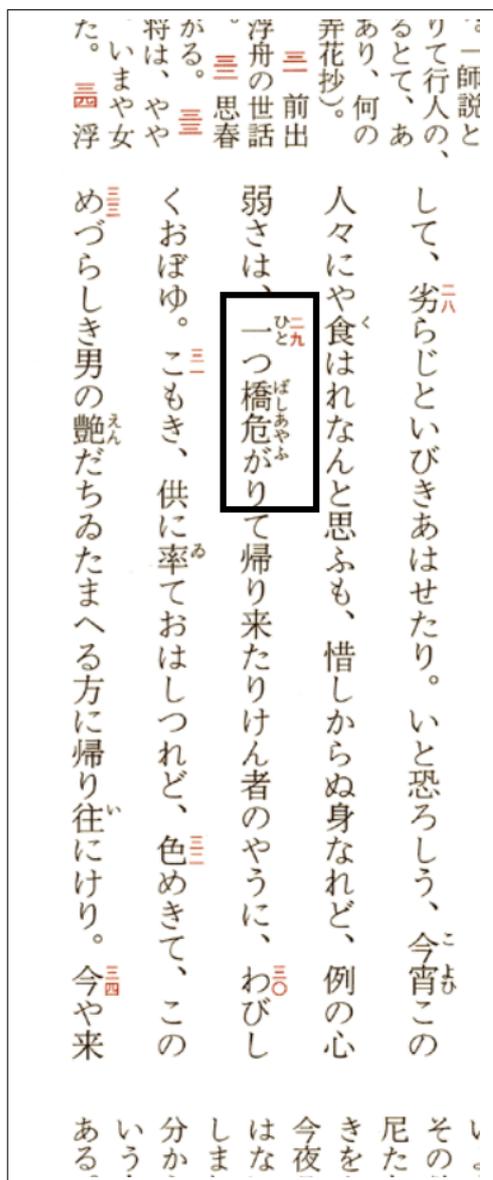


図 2: 新編日本古典文学全集 25 卷, 源氏物語 (6), 手習, p.329. より抜粋. 黒四角で囲った箇所が当該の「ばしあやふ」である.

る. ただし本実験で使用したのは元のローマ字テキストでなく, コーパスの漢字片仮名交じり本文テキストとして試験的に XML 化を試みている文献 [1] である⁷. 本実験はこの試験的 XML 化の一環として実施したため, 自動解析の性能

⁷著作権の関係で, コーパス公開時に実際に使用する漢字仮名交じり文での本文は未だ定まっていない. 今回のデータは XML 化方針の検討のため参考として使用しているものであり, 今回使用している漢字仮名交じり文を本文として公開するわけではない.

評価をすべて人手のチェックによって行うことができた.

文献 [1] の電子化は, (アノテーション作業の専門家でない) 業者依頼としたため, ルビアノテーションも新編全集と同様の大雑把な仕様となっている. 納品されたデータ (図 3) を単純な置換ルールのスクリプトで CHJ で採用している XML 形式 (図 4) に変換した. この際, 業者が実施したルビのアノテーションは XML 中では ruby タグに一括置換される. ruby タグで親文字列を括り, rubyText 属性でルビ文字列をアノテーションしている. エソポのハブラスに含まれる ruby タグは全 1,425 個であった.

mpaligner への入力として, エソポのハブラス中の親文字列と, ルビ文字列を並べたものに加え, 形態素解析用辞書 UniDic[4] の表層形と, 仮名形出現形⁸を並べたもの, および CHJ に含まれるすべてのルビタグを使用した. これにより, 全 1,374,214 の漢字文字列, 仮名文字列ペアが mpaligner への入力となった.

mpaligner が行なった自動アライメントの例を図 5 に示す. これを見ると, モノルビ化可能な漢字列に関しては, 正しく 1 字ずつにルビを当て, 「蝸牛 [かたつぶり]」のような熟字訓はそのままで残すことができている. また自動アライメントの結果をキリシタン資料のコーパス化担当者がすべて人手で確認したところ, エソポのハブラスにおいて改めて人手の修正が必要であった箇所は 37/1,425 であり, 正解率は 97.4% と高い精度でルビタグの自動修正が行えていることがわかった. また誤りのあった箇所を調べたところ, 「諸 [もろも]々[ろ]」のように踊字を親文字列に含む箇所での分割エラーと未分割が発生していることが分かった (エラー数: 5). そこで mpaligner への入力時, 踊字を直前の文字へ置換する前処理を実施したところ, 当該のエラーは 0 になり, 解析精度も 97.8% まで向上した. 残りのエラーを確認したところ, ほとんどが「一人 [ひとり]」や「昨日 [きのふ]」のような高頻度の文字列で, 担当者が主観的には分割してほしくない箇所の過分割であった. そのため, どういった文字列は過分割したくないかをあらかじめ規程集としてまとめておくことで, これらのエラーはなくすることができる. 実際, 「数字+人」および「昨日」の過分割 (エラー数: 16) を排

⁸UniDic の仮名形はカタカナ表記であるため, 平仮名表記に変換して使用した.

☆480
 1 言下 [ごんか] に人から見知られて、恥に及うで、退 [しりぞ] かう
 2 ず。
 3 馬と驢馬との事。
 4 ある人、驢馬と馬とに荷を駄 [おほ] せて行くが、驢
 5 馬の荷物があまり過ぎて、先へ行き着かう様 [やう] も
 6 なければ、驢馬から馬に佗言 [わびこと] をして言ふやうは、
 7 「そなたと我は一門で、そっとの高下をもって
 8 隔った。わが荷物があまり過ぎて、一足も
 9 引かうずる様がない。少しそなたの上に付けて、
 10 我を助けられいかし」と。馬は一向承引せいで、
 11 結句大きに嘲って、先へ行ったれば、驢馬は力
 12 に及ばいで、つひに倒れて死んだ。そこで、この
 13 馬追ひはせうことがなうて、驢馬に付けた荷物をも、
 14 ことごとく馬一匹に取り付けて、あまっさへ驢馬
 15 の皮をも剥いで、馬に駄 [おほ] せて行くところで、
 16 その時、馬は、わが愚痴 [ぐち] なることを顧み、
 17 「先に驢馬の佗びた時、そっと合力 [かぶりよく] したらば、これ
 18 ほどの重荷は持つまじいものを」と悔め
 19 ども、益がなかった。
 20 下心。
 21 「理のこうずるは非の一倍」というて、道理を承引
 22 せぬ者は、必ず非分の害に会はいで叶は
 23 ぬものぢや。
 ☆481
 1 二人 [ににん] 同道して行く事。

図 3: 納品された文献 [1] のテキストデータ. ☆はページ番号を表し、各行が紙のページ上での 1 行に相当している. また行頭の数字はページ中の行番号である. [] は直前の漢字列のルビ文字列を表している.

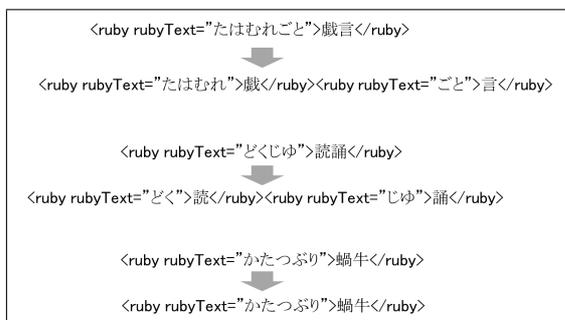


図 5: エソポのハブラスのルビタグ修正結果の例. 矢印の上が修正前, 下が修正後である.

除しただけで、精度は 98.9% まで向上した.

5 追加実験: アライメントスコアを利用した近世口語資料の当て字の検出

追加実験として、近世口語資料に頻出する当て字ルビの検出実験も行なった. 近世の資料には

「心驚 [びつくり]」「誘引 [さそは | れ]」⁹ のように通用的でない特殊な振り仮名 (当て字) が多い. これは近世の洒落本や人情本などの通俗小説類は大衆向けの出版物であり、難しい漢語が読めない読者でも、振り仮名と本文の平仮名の部分だけを目で追っていけば、内容を理解できるように工夫されていたためである. こういった当て字は、本文の漢字列の実際の読みと乖離しており、かつ著者ごとに独特の使用を行っていたため、自動アライメントを行なった場合に低いスコアが割り当てると予測できる. そこで現在整備中の近世の洒落本、人情本コーパスのルビアノテーションを対象にエソポのハブラスと同様の方法で文字列アライメントを実施し、しきい値を使って各アライメントのスコアから当て字を検出する実験を行なった.

対象としたのは、整備中の洒落本コーパスより花街鑑、人情本コーパスから比翼連理花洒志満台をそれぞれ 1 ファイルずつであり、どちら

⁹ 「|」は短単位境界を表す.

```

-<text series="エソポが作り物語の抜き書き。" textID="天草版伊曾保物語_044_馬と驢馬との事" title="馬と驢馬との事" year="1593" year_w="文禄2">
-<front>
-<titleBlock>
-<block type="title">
-<s>
<pb n="480"/>
<lb n="3"/>
馬と驢馬との事。
</s>
</block>
</titleBlock>
</front>
-<body>
-<article>
-<p>
-<s>
<lb n="4"/>
ある人、驢馬と馬とに荷を
<ruby rubyText="おほ">駄</ruby>
せて行くが、驢
<lb n="5"/>
馬の荷物があまり過ぎて、先へ行き着かう
<ruby rubyText="やう">様</ruby>
も
<lb n="6"/>
なければ、驢馬から馬に

```

図4: 図3からルールスクリプトで一括自動変換したXMLファイルの一部。エソポのハブラスは複数の小話から構成されており、各話を1つのXMLファイルとしている。上記は図3中に含まれている「馬と驢馬との事。」をのXML化したものである。ルビはrubyタグとして付与されており、rubyタグで括られたテキストが当該ルビの付与されている漢字仮名交じりの本文、rubyタグ中のrubyTextがルビ文字列をそれぞれ格納している。

もXMLでマークアップされておりルビもタグ付けされている。これらのXMLファイルから抽出したルビタグは数全部で5,153であった。コーパス整備の担当者がこれらのルビタグに対し、当て字と判断したルビにはtype属性として「当て字」という値を付与した。付与されたのは5,153中、206個であった。これをエソポのハブラスと同様の方法で自動アライメントし、スコアの昇順でソーティングを行なった。その結果、スコアに対し、しきい値-40を設定したとき-40以下のスコアを持つルビは130個、うち当て字は103個で約79%の精度で半分の当て字が検出できることが分かった。

6 おわりに

国語研がこれまでに蓄積してきたUniDicや通時コーパスといった言語資源を自動アライメントに活用することで、人手では高コストなルビタグの修正作業も高精度に自動化できることがわかった。今後の課題として、キリシタン資料だけでなく、他の時代の資料、特にルビの頻出する近世や近代の資料に対しても同様の手法を使用し、評価を行なっていく予定である。また近世口語資料の当て字は、自動解析、特に形態素

解析において精度向上の障害となっているため、追加実験として行なった当て字検出を今後発展させることで、近世口語資料の形態素解析性能向上にもつながると考えられる。

謝辞

本研究は、国立国語研究所共同研究「通時コーパスの構築と日本語史研究の新展開」の研究成果を報告したものである。

参考文献

- [1] 大塚光信, 来田隆編: エソポのハブラス 本文と総索引 本文篇, 清文堂出版(1999).
- [2] 近藤泰弘: 「日本語通時コーパスの設計」, NINJAL「通時コーパス」プロジェクト・Oxford VSARPS プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集, pp.1-10 (2012).
- [3] 久保慶伍, 川波弘道, 猿渡洋ほか: 日本語の未知語に対する発音付与のための多対多アライメント情報処理学会論文誌, Vol.54, No. 2, pp.452-462 (2013).
- [4] 伝康晴, 小木曾智信, 小椋秀樹ほか: 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』22号, pp.101-123 (2007).

- [5] 山口昌也：構造化テキストに対応した全文検索システム『ひまわり』，国立国語研究所報告 122, pp.49-82, 博文館新社 (2002).