# 電子化が望まれる近代語資料探索 ： 日本語史を研究する大学院生の報告から


| | |
|---|---|
| 著者 | 岡島 昭浩，森 勇太，金 〓泳，竹村 明日香，坂井 美日 |
| 雑誌名 | 近代語コーパス設計のための文献言語研究 成果報告書 |
| ページ | 27-35 |
| 発行年 | 2012-10-31 |
| シリーズ | 国立国語研究所共同研究報告 ； 12-03 |
| URL | http://doi.org/10.15084/00002763 |

# 電子化が望まれる近代語資料探索
## ―日本語史を研究する大学院生の報告から―

岡島　昭浩（大阪大学大学院文学研究科）[1]
森　　勇太（日本学術振興会特別研究員）
金　　曦泳（高麗大学校言語情報研究所）
竹村明日香（大阪大学大学院博士後期課程学生）
坂井　美日（大阪大学大学院博士後期課程学生）

### １．趣旨

本稿の筆頭著者である岡島は、2010 年度に大阪大学大学院文学研究科における「国語史演習」[2]において、「近代語資料探索」をテーマとした。岡島が、本プロジェクトに参加することになったことを受け、これから近代語コーパスを利用することになるであろうと思われる若い研究者たちが､どのような資料がコーパスに含まれることになることを望むか、ということを知りたい、という目的があった。もちろん、日本語史研究に常に付随する資料論を意識した演習でもある。対象の学生は大阪大学大学院文学研究科に属する、博士前期課程・博士後期課程の学生である。

シラバスは、おおよそ以下のようであった。

講義題目　近代語資料探索
授業の目的　幕末・明治初期から昭和初期あたりまでの言語史資料は、英学資料などを除くと、文学作品以外への開拓はまだまだ進んでいない。また、文学作品でも、文学史上有名な一部の作品が使われているのが現状である。そこで、言語史研究に有用な資料を探し、その有効利用を探る。
講義内容　電子化されたら有用であろうと思われるものを探ることを、今回は主眼としたい。
授業計画　発表者が資料を探索し、「電子化されると大変便利であるから、電子化されるべきである」ということを主張する発表をしてもらう。どの方面(分野)に役に立つ資料か、また、底本とすべき資料、その分量(抄出ならばその度合）、必要な精度、タグの必要度なども含めること。

---

1 okajima@let.osaka-u.ac.jp　岡島以外の執筆者は、みな、2010 年度に於て大阪大学大学院文学研究科文化表現論専攻国語学専門分野の学生であった。

2 正式には、博士前期課程の学生が受講するものが「国語史演習」、博士後期課程の学生が受講するものが「国語史特殊演習」であるが、同時に行われるものである。

参考文献　『太陽コーパス』(博文館新社)

上記シラバスの他に、日本文学・国語学を専門分野とする学生を集めてのガイダンスの際や、最初の授業の際に求めたものは、次のようなことである。

膨大で実現性に乏しいものは外す（『明治文学全集』『明治文化全集』を全部、など）
資料の文字数等、言語量を大まかでよいので明示すること
どのようなジャンルの研究に有効か。電子化することによる効果[3]。（プレゼンテーション）
著作権の確認(公開可能な電子資料を目指すために)
OCR との相性チェック[4]

上記のような説明を経て、実際に受講したのは、
国語学専門分野[5]博士前期課程 4 名　博士後期課程 5 名
日本文学専門分野[6]博士前期課程 1 名
の計 10 名であった[7]。

最初の授業で岡島の例示したものは、『旧事諮問録』『史談会速記録』[8]等の歴史系の速記資料であった。「膨大で実現性に乏しいものは外す」という趣旨からは外れるが、抽出などで言語量を減らす方向で考えた。

## ２．提案されたもの

提案された主な資料群としては以下のようなものがあった。

---

3 電子資料のない状態で、手作業で用例数等を求め、ケーススタディを示した学生が何人もあった。

4 新字体の資料で入力の上、原本等に戻って訂正するなどの場合も考慮に入れた。現在は、版面権は認められていないので、校訂者の権利のみを気にすればよい。

5 同じ文学研究科内に、別に日本語学専門分野があることもあり、国語学専門分野では、歴史的研究・文献研究を中心にしている。

6 国語学専門分野と日本文学専門分野は、研究室を共有するに留らず、院生発表会など、多くの場面で研究を共にしている。課程在籍の間に、国語学専門分野の学生は日本文学の演習を一つ以上、日本文学専門分野の学生は国語学の演習を一つ以上取ることを求めている。

7 本稿の共著者の他に、国語学専門分野のものは、伊藤由貴・清田朗裕・目黒陽子・鈴木久恵・山本一巴。

8 原書房から復刻版が出ており、『幕末明治/研究雑誌目次集覧』古書通信社（1968）に目次がある。明治 25 年から昭和 13 年まで刊行。江戸時代から明治の初め頃までの歴史的証言を集めようとして速記したものの刊行である。史料編纂所に、出版されなかった速記録が蔵されているとのことである（『幕末明治/研究雑誌目次集覧』）

○共通語ないし東京語系

『東洋学芸雑誌』(雑誌)　　〔後述〕

『丁酉倫理会倫理講演集』(雑誌)

丁酉倫理会"が 1900(明治 33)年に発刊した雑誌(終刊は 1946(昭和 21)年。ただし継続誌有り)で、主に"丁酉倫理会"で行われた講演が収められている。口語体の講演が多く収録され、一定の言語量をもった、多数の話者によるコーパスの作成が期待される。

1)　講演録+"雑録"+"時潮"+"出版界(新刊)"+"応問[読者の質問欄]

2)　講演録は初期はすべて口語体。後世では文語体の論文も加わる。

3)　"雑録"以下は文語体・口語体の両方がみられる。

4)　表記はすべて漢字ひらがな交じり文。

縦 35 字×横 15 行　1 号あたりおよそ 70〜100 頁前後

『旧幕府』(雑誌)

明治 30〜34 年。戸川残花編。勝左衛門太郎「夢酔独言」などあり。

「帝国議会会議録」

国会図書館で現在電子テキスト化されているもの（昭和 20 年以降）よりも前のものについてのテキスト化の提案。全体では大きすぎるので、抽出して行うことを提案。

篠田鉱造『幕末百話』『明治百話』『幕末明治女百話』など

篠田鉱造（1871〜1965。報知新聞記者）による、聞書スタイルのもの。

河竹黙阿弥『狂言百種』　〔後述〕

SP レコード文句集

大空社より『大正期 SP 盤レコード/芸能・歌詞・ことば全記録』（倉田善弘・岡田則夫監修。1996〜1997）として復刊されている。金水(2001)参照。

東江学人『文明開化/内外事情』

福沢諭吉『西洋事情』との比較などで、明治初期の言語資料として。近代デジタルライブラリ所収。

松林伯円の講談「安政三組盃」

明治 18 年刊。講談速記本の濫觴とされるもの。前年刊で落語速記本の初めとされる三遊亭円朝の「怪談牡丹燈籠」と比して注目されることが少ない。リライトされた形でしか再版はなされていないようで、原刊本からの電子化が必要となる。近代デジタルライブラリには欠巻あり。

巌谷小波の言文一致もの（「初紅葉」など）

「こがね丸」の文体を換えた執筆などでは言及される巌谷小波ではあるが、その他の言文一致作品も重要である、という指摘。

○上方語系

柴田鳩翁『鳩翁道話』等。

江戸期のもの。『鳩翁遺稿』（昭和 4、柴田寅三郎[9]）による。平凡社東洋文庫の柴田実『鳩翁道話』(1970)は表記の改訂有り。『日本思想大系・石門心学』には、初編のみ。「『鳩翁道話』の文字数は、正編、続編、続々編はそれぞれ約 47,000 字、拾遺に含まれる未刊行の筆記は約 18,000 字である」との情報があった。

一荷堂半水（『諺 臍の宿替え』・『穴さがし心の内そと』）

特に『穴さがし 心のうちそと』は近世上方語資料として知られている。前田(1974)参照。『諺 臍の宿替え』は太平書屋より影印・翻刻あり(1992、武藤禎夫)。また南和男『江戸のことわざ遊び』（平凡社新書、2010)で、一部影印・現代語訳もあり）。

大阪文芸誌『なにはがた』

明治 24 年。西村天囚など大阪朝日新聞関係。また、堺利彦なども参加し、言文一致体のものを書いている。

『上方はなし』　〔後述〕

今村信雄速記『名作落語全集』

騒人社書局,1929-1930 年。紙型の流用で他の出版社からも後に刊行されている。今村信雄は 1959 年歿で、著作権保護期間終了。演者の著作権を確認する必要有り。東西の落語が収録されているが、上方落語を中心の電子化を考える。

○その他

『日本外交文書』『条約改正関係・大日本外交文書』　〔後述〕

永井荷風の小説

日本新聞歴史

番外　青空文庫のデータベース化

他にもあったが、有意義と考えられるものを示した。以下には、その数例を摘記する。

## ３．例

### ３．１　河竹黙阿弥『狂言百種』　　坂井美日

『狂言百種』は、明治二十五年四月から明治二十六年二月にかけて出版された、河竹黙阿弥脚本集である。これは、大正期に出版された『黙阿弥全集』などとは異なり、河竹黙阿弥自身が編纂したという点が特徴である。『狂言百種』におさめられた作品は、全て黙阿弥の「世話物」である。黙阿弥の自筆台帳は関東大震災で多くが消失している。

さらに多くの黙阿弥作品を扱おうとするならば、大正期に翻刻された『河竹黙阿弥脚本集』(河竹糸補修・河竹繁俊校訂、全 28 巻、1924-1926)や、その増補修正版の『河竹黙阿弥全集』(河竹糸補修・河竹繁俊校訂、全 28 巻、1924-1926、春陽堂)などがある。この中にはさらに多くの散切物が入っており、『東京日日新聞』『女書生繁』『人間萬事金世中』

9 1942 年没で、著作権保護期間終了。

などがある。

しかし、これらの資料は次の点で問題がある。

1.『河竹黙阿弥脚本集』『河竹黙阿弥全集』はともに、2017 年まで著作権期間が残っている(河竹糸(=黙阿弥の娘)の養子である河竹繁俊(1889-1967)が校訂に加わっているため。)

2.校訂が加わっているため、どの時期の言語資料として扱うべきか、検討の必要がある。

しかし二点目について、『狂言百種』、と『黙阿弥全集』には次のような有意差のある補修があるようにも思われるため、二者の異同比較は有効かもしれない。

○無生物主語、結果相、自動詞の「てある」→「ている」

(1)わずか五錢の所十五錢残ってある(『狂言百種』、木間星、12)

(2)わずか五錢の所十五銭残っている(『黙阿弥全集』木間星、672)

(無生物主語の「～ている(自動詞)」が無生物主語の「～てある(自動詞)」を圧倒して現代語と同様の体系になるのがいつなのかはまだよくわかっていないが、この異同例などからはその整理状況が伺える。)

A.散切物(明治期世話狂言)だけを抜粋した場合、文字数は約 65 万字、

B.全ての世話狂言を含む場合は 145 万字となる。

一作品あたり十三万字として、

A) 13×5 作品=65 万字

B) 13×11 作品=145 万字

OCR→ほぼ読み取れず。手打ち作業がはやいと思われる。コスト計算は次のようになる。

A)

・コピー代: 一作品あたり約 60 枚×5 作品

・手打ち作業: 30 時間×10 名

・確認作業: 20 時間×10 名

B)

・コピー代: 一作品あたり約 60 枚×11 作品

・手打ち作業: 30 時間×20 名

・確認作業: 20 時間×15 名

### ３．２ 『東洋学芸雑誌』 森勇太

明治 14～昭和 5 年の雑誌。567 冊。

1)　表記法としては、かなの選択(漢字片仮名交じり文か、平仮名交じり文か)、記号法・用字法の面でさまざまな表記法が混在している様相がある。

2)演説口調に近い口語文の特徴は認められるが(談話標識、「です」などの敬語表現)、演説の文章をそのまま筆録したものかどうかは確証がない。ただし、発話者(筆者)を特定するのは容易であり、個人の体系として記述を行うことには適している。

3)理化学系統の論文が収録されており、それらの分野の語彙の出現をたどったり、また、明治に入って導入された新しい概念を訳出するための語彙を探し出すこと、および訳語の変遷等に資することが期待される。

4)待遇表現の面では、「であります」「です」「でございます」などが混在している。用法の面でも、形容詞述語や動詞述語に「です」が充てられている例もあり、丁寧表現の整然とした使用がなされていない。

5)総じて、近代に入っての、表記規範や言文一致体など文体形成に至るまでの書きことば形成の過程を追う資料となりうる。特に学者などの知識層の影響を明らかにすることが期待される。

太陽コーパスからの比較の観点からいえば、以下の点で有用な資料となりうる。

1)太陽コーパス発刊以前の状況を調査する。

2)太陽コーパスにあまり掲載のない理化学系統の語彙の調査を行う。

3)太陽コーパスの並行資料として、太陽コーパスを相対的な視点から捉える。

字数は漸増傾向にあるが、約 500,000 字平均として、約 4,000,000 字程度のコーパスとなることが期待される。

OCR

1881 年と 1917 年のものを OCR にかけてみると(e.Typist　12.0 使用)、ルビが多い 1881 年のものは崩れがあるが、全く認識しない、というわけではない。1917 年のものはきれいに読めている。

著作権

『東洋学芸雑誌』そのものの著作権(東洋学芸社)は最終号発行から 79 年以上経過しており問題とならない。ただし、著作権が明確に雑誌社にあると明記された箇所がないため、論文著者の著作権が問題となる可能性がある。

コスト試算

太陽コーパスと比較検討できる材料とするためには、(タグまであれば一番良いが)少なくともルビまでを明示化した形で電子化することが必要になるだろう。しかし、以下の試算では非常にコストが高く、この点で問題となる。

『東洋学芸雑誌』電子化にかかるコスト

内訳　　個数
原本コピー代金　　3360 枚
OCR 読み込み、修正処理　　1120 時間
▼内訳
[原本コピー代金]コピー枚数の内訳:
1 号平均 70 ページ(論文部分)を見開きコピー→35 枚分×12(ヵ月)×8(年)=3360
[OCR 読み込み、修正処理]時間数の内訳:6 ページ/1 時間→1 時間に 3 枚処理。

その後、DVD-ROM 版が出た（大空社、2011）。

### ３．３　『上方はなし』五代目笑福亭松鶴　　竹村明日香

「上方はなし」(昭和 11〜15 発刊)は会話体中心の落語速記資料で、原稿用紙 1611 枚分の豊富な口語性を有する。資料性については矢島(2006)(2007a)(2007b)を参照。

電子化に当たってのコスト試算
コピー　　約 280 枚
OCR 読み込み・修正処理　　30 時間×8 名
原本との対照チェック　　20 時間×8 名
→すでに活字化されているので､最初から打ち込む資料よりは格段に低コストで行える。
著作権問題
原本　昭和 11-15(1936-1940)ならば、今でも自由な電子化が可能か。
図書、三田編(1971-1972)を使用すると、三田純一(別名・三田純市)氏死後の 1994+50=2044 年以降しか不可か。

一部に特殊な語彙・語法が見えるが(武士の言葉など〉、全体的に近代上方語を代表する語彙・語法を含んでおり、近代上方語資料として有用であると考えられる。

復刻版は OCR にかけてもほぼ問題なく読み込むので、電子化に大きなコストがかからず、短期間で行えるという長所がある。

### ３．４　『日本外交文書』　金曦泳

日本語では、「遺憾」「遺憾の意を表する」「積極的に」「前向きに検討する」「可及的速やかに」のような定型表現が存在する。（中略）これには漢文脈の名残の語形を含め、公文書の文書としての構成を背景にした言い回しも存在する。

『日本外交文書』は日本の外務省によって明治期から昭和に至るまで(進行中)の外交文書を、編年体を基本として編纂・公表された公文書集である[10]。量的な面から見ても、例

10 外務省の web ページ上で公開されている。

えば明治期や大正期の公文書における文体や語彙などの変遷を知るに有用な資料であると考えられる。

例　漢語接尾辞「-的」の用例

・「1 巻 1 冊(明治元年)」　全 942 ページ:0 例

・「22 巻(明治 22 年)」　事項 11 の 14 ページ:3 例

・「45 巻 1 冊(明治 45 年)」　事項 1 の 43 ページ:26 例

『日本外交文書』は、基本的に各項目ごとに時間順に番号が付されている。また、一定の形式に沿っているのであって、その形式情報をタグを用いて付与しておくと、より素早くまた的確に必要とする情報に接近できるようになる利点がある。また、日本語だけではなく、外国語(英語・ドイツ語など)とその和訳が同時に掲載されている場合が少なくないため、対訳文における情報をタグを使って付与しておくと、対訳コーパスとしての活用も期待される。このような点から、単に電子化することにとどまらずタグを付しておくと様々な利点があると考えられる。

ただし、

> 「日本外交文書」および日本外交文書デジタルアーカイブの著作権は外務省に帰属します。本コンテンツを著作権の保護期間内に著作権法上認められる範囲(私的使用のための複製など)を超えて使用する場合は、当省の許諾を得る必要があります。

と明記されている著作権などの問題がある。外交文書自体は、団体名義の著作物であり、公表後 50 年で著作権は切れるはずであるが、このデジタルアーカイブに新たに著作権が発生しているとすれば、許諾を得る必要がある[11]。

語数及びコスト試算

・明治期における『日本外交文書』

本巻 63 冊,その別冊 9 冊,別巻 17 冊,追補 2 冊,合計 91 冊,延べ約 73,000 ページ

・語数:約 288,373　÷　946 ページ　＝　約 305 語/頁

合計語数:73,000 ページ×305 語＝約 22,265,000 語

cf.　『現代日本語書き言葉均衡コーパス』「BCCWJ 領域内公開データ(2009 年度版のモニター公開データ)」:4,490 万語

## ４．まとめ

上記の注にも記したように、大阪大学大学院文学研究科の国語学専門分野では、日本語の歴史的研究を中心にしており、提案された資料については、自分の言語史研究の中で、調べるのに苦労した時代・位相などについて、それを埋めるものを探そうとしたもののように思われる。

---

http://www.mofa.go.jp/mofaj/annai/honsho/shiryo/archives/

11 その後、公開されている電子化の方式が変更されたようで、演習時のように容易に画像データを取得できるものではなくなったようである。

文法史の面で特に必要となる口語性の高い資料を求める一方、語彙的な面などから文語（ただし文芸性の強いものではなく、公用文的なもの)を求める声もあった。

口語については、近世中期頃までの上方語系資料と現代関西方言の間を埋めるような上方語資料を求めており、また、共通語・東京語系資料についても、言文一致の定着よりも前のものが主として求められている感があった。それ以後のものは、既存のもの（青空文庫や『CD-ROM 新潮の百冊』「国会会議録」など）で或る程度覆えるという感触によるものと思われる。

具体的な提案もあり､予算さえあればすぐにでも取りかかりたいと思える企画もあった。特に上方語資料などは、大阪の大学として作れないものかと思っている。

また、例として挙げたもの以外にも、小規模のテキストデータとしての存在意義は充分なものや、日本語研究者以外の手によるテキスト化を待って、日本語史資料として使うことが望まれるものもあった。

本稿は、演習時の資料および、期末のレポートによって編集したが、言語的な調査の多くについては、今後、発表されることもあるだろうということや、全体のバランスを考えて、本稿においては、岡島の責任で削除したものが多い。その他の抽出・削除・並べ替えなどについても、責任は岡島にある。

## 文 献

金水敏(2001)「《資料紹介》明治・大正時代 SP レコード文句集について」『語文』75,76 pp.80-88

前田勇(1974)「穴さがし心の内そと」『近代語研究』4、武蔵野書院 pp.429-484

三田純一編(1971-1972)　五代目笑福亭松鶴『上方はなし』(上・下・解説)、三一書房

矢島正浩(2006)「落語録音資料と速記本――五代目笑福亭松鶴の仮定表現の用法から――」『愛知教育大学国語国文学報』64 pp.132-116

――――(2007a)「五代目笑福亭松鶴落語における原因・理由表現の用法」『愛知教育大学大学院国語研究』15 pp.70-56

――――(2007b)「近代関西言語における条件表現の変遷原理に関する研究」平成 17-18 年度科学研究費補助金(基盤研究 C)研究成果報告書　課題番号 17520298