『日本語歴史コーパス 明治・大正編III明治初期口語資料』の拡充
―『春秋雑誌会話篇』の追加とデータ設計の改訂―

近藤 明日子　KONDO Asuko　（国立国語研究所）
常盤 智子　TOKIWA Tomoko　（白百合女子大学）
小木曽 智信　OGISO Toshinobu　（国立国語研究所）

1. はじめに

国立国語研究所では、明治 0 年代～10 年代の口語体資料を収録するコーパスとして『日本語歴史コーパス 明治・大正編Ⅲ明治初期口語資料』を構築・公開している。2021 年 3 月、本コーパスにアーネスト・サトウ著の日本語学習書 *Kuaiwa Hen* の 1 パートである『春秋雑誌会話篇』を追加収録し、さらにコーパスの設計を一部改訂し、新バージョン（短単位データ 0.9）として公開した。本発表では、追加資料の『春秋雑誌会話篇』の資料性やコーパスデータの設計の改訂点を中心にコーパスのバージョンアップについて報告する。

2. コーパスに収録した資料

次の表 1 に『日本語歴史コーパス 明治・大正編Ⅲ明治初期口語資料』(短単位データ 0.9) に収録した資料とその延べ語数を示す。

表1　コーパス収録資料とその延べ語数

資料名	著者	刊行年	資料ジャンル	底本所蔵場所	延べ語数*
交易問答	加藤弘之	明治2（1869）	非文芸	国立国語研究所	10,176
安愚楽鍋　初編/二編/三編	仮名垣魯文	初編/二編：明治4（1871） 三編：明治5（1872）	文芸	国立国会図書館	20,907
開化のはなし	曲肱軒主人	明治5（1872）か	非文芸	国立国語研究所	4,765
春秋雑誌会話篇	アーネスト・サトウ	明治6（1873）か	非文芸	早稲田大学図書館	17,775
文明開化　初編/二編	加藤祐一	初編：明治6（1873） 二編：明治7（1874）	非文芸	国立国語研究所	29,156
百一新論	西周	明治7（1874）	非文芸	国立国会図書館	24,029
よりあひばなし　初編	榊原伊裕	明治7（1874）	非文芸	国立国語研究所	14,573
開化問答　初編/二編	小川爲治	初編：明治7（1874） 二編：明治8（1875）	非文芸	国立国語研究所	47,111
明治の光	石川富太郎	明治8（1875）	非文芸	国立国会図書館	10,612
文明田舎問答　初編	松田敏足	明治11（1878）	非文芸	国立国会図書館	12,638
民権自由論	植木枝盛	明治12（1879）	非文芸	国立国会図書館	9,888
計					201,630

*記号類・未知語類を除いた主本文の延べ語数。

表 1 に示したように、本コーパスは明治 2（1869）～明治 12（1879）年に刊行された 11

資料、延べ約 20 万 2 千語（短単位）を収録する。11 資料中、『安愚楽鍋』『春秋雑誌会話篇』は当時の話し言葉の実態を知る資料として、その他の 9 資料は明治 20 年代以降に論説文等の実用文において進展する口語体書き言葉の源流を示す資料として重要なものである。

『春秋雑誌会話篇』は新バージョンで新たに追加収録した資料で、アーネスト・サトウ著の日本語学習書 *Kuaiwa Hen*（『会話篇』）の Part III にあたるものである。*Kuaiwa Hen* はローマ字表記の日本語の対話文とその英訳を収めた Part I、その注解を内容とする Part II、そして Part I の日本語の対話文の和文テキストである Part III の 3 パートからなる[1]。Part III には「春秋雑誌会話篇」という内題が示されており、本コーパスではそれを資料名とした。『春秋雑誌会話篇』に収められた日本語の対話文は、外国人が日本語の日常会話を学ぶことを意図したもので、幕末から明治初年の定型的な口語表現や、様々な場面の中での多様な発話者による対話が含まれ、当時の話し言葉を知るための重要な資料となっている。表記の面からは、前半は仮名文、後半は漢字仮名交じり文となっており、当時の仮名遣いの問題や文字教育という点からも注目されるべき資料である。『春秋雑誌会話篇』には刊記がなく刊年は不明であるが、Part I・II の刊行された明治 6（1873）年と同時期と推測される。コーパスの底本には早稲田大学図書館所蔵本（請求記号：文庫 08 C0763 1-2）を用いた。

3. コーパスデータの設計の改訂点

コーパスデータの設計[2]は旧バージョンに準じ、電子化した本文テキストに、形態論情報をはじめとする言語研究に有用なアノテーションを付与した。コーパスの利用は検索ツール「中納言」（https://chunagon.ninjal.ac.jp/）を通じて行い、短単位検索・文字列検索を可能とした。検索結果には本文テキストにアノテーションが同時に示され、「早稲田大学図書館古典籍総合データベース」（https://www.wul.waseda.ac.jp/kotenseki/）で公開されている底本[3]の該当ページの画像へのリンクも表示される。

新バージョンでは以下の①②の点について設計の改訂を行った。

① 漢字の電子化

本コーパスの電子化テキストに使用した文字の範囲は、JIS X 0213: 2004（JIS の文字コードの規格）の文字集合に準拠している（一部使用しない文字がある）。漢字の電子化については、収録資料の書体が楷書・明朝体のように字画を崩さない場合と、行書・草書のように字画を崩す書体の場合で、電子化の方法が異なる。

字画を崩さない書体の資料の場合、文字集合に含まれない漢字については JIS の包摂規準や独自に追加した包摂規準を適用し包摂したり、意味・読み・漢字部品の一致・類似する文字集合内の代用字に置き換えたりすることで、できる限り文字集合に含まれる漢字で電子化を行った。また、字画を崩す書体の資料の場合、漢字の電子化の候補として異体字関係にある複数の字体が想定される際、コーパスで使用する文字集合内漢字について異体字グループを設定し、グループ中で現代もっとも一般的に使用される 1 字体（常用漢字表字体・表外漢字字体表印刷標準字体等）によって電子化した[4]。また、文字集合に含まれない漢字は文

---

[1] *Kuaiwa Hen* の書誌については松村（1998、pp.351-419）に詳しい。

[2] コーパスデータの設計の詳細については近藤（2021）を参照のこと。

[3] https://www.wul.waseda.ac.jp/kotenseki/html/bunko08/bunko08_c0763/

[4] 例えば、原本の「  」の電子化候補として「覚」「覺」の2字体が考えられるが、常用漢字表字体の「覚」に統一して電子化した。

字集合内に代用字がある場合はそれに置き換えて入力した。なお、新バージョンで追加収録した『春秋雑誌会話篇』は字画を崩す書体の資料であるが、漢字の字体は仏訳本『春秋雑誌会話篇』[5]の活字字体を参照し、それに字画を崩さない資料の電子化方法を適用して電子化を行った。

以上の方法でも電子化のできない漢字については、旧バージョンでは「〓」記号で電子化していたが、新バージョンではそのうち Unicode に収録されている漢字（「慺」「獃」「詖」等）はそれにより電子化した。さらに、旧バージョンでは JIS X 0213: 2004 の文字集合中の Unicode における CJK 統合漢字拡張 B をコーパスの文字集合には含めていなかったが、新バージョンでは文字集合に含める改訂も行った。これにより旧バージョンでは代用字で入力していた漢字の一部を新バージョンでは原本の字体で電子化した（例：[旧]春→[新]卷、[旧]蹈→[新]蹈）。

以上の漢字の電子化ルールの改訂により、「〓」記号の入力を減らし、原本の漢字にそった字体での入力をいっそう進め、検索等での利便性を高めた。

② 特殊な表記への形態論情報付与

本コーパスでは形態論情報は振り仮名による読みに基づく語(短単位)に対して付与した。しかし、例えば「無據（右振り仮名：よんどころなく)」「一（右振り仮名：ひとつ)」「開化文明（右振り仮名：よのひらける)」のように、短単位の境界や順序と本行の漢字との間に対応関係が見られない場合、旧バージョンでは、返読・補読によって本行のテキストを校訂した上で形態論情報を付与したり、臨時的に複数短単位を一つにまとめて形態論情報を付与したりする対処方法をとり、それも困難な場合はやむなく未知語として扱っていた。これを新バージョンでは、表 2 に示すように形態論情報の多重化の技術（村山・小木曽・中村 2017）を新たに用い、形態論情報を付与することを実現した。

表 2　形態論情報の多重化の例

本文テキスト	書字形*		付与する形態論情報（一部）*
よんどころなく 無　據	主本文	よんどころ／なく	語彙素「拠り所」語彙素読み「ヨリドコロ」語形「ヨンドコロ」品詞「名詞-普通名詞-一般」／語彙素「無い」語彙素読み「ナイ」品詞「形容詞-非自立可能」活用型「形容詞」活用形「連用形-一般」
	副本文	無據	品詞「対象語無し」
ひとつ 一	主本文	ひと／つ	語彙素「一」語彙素読み「ヒト」品詞「名詞-数詞」／語彙素「つ」品詞「接尾辞-名詞的-助数詞」
	副本文	一	品詞「対象語無し」
よのひらける 開 化 文 明	主本文	よ／の／ひらけ／る	語彙素「世」語彙素読み「ヨ」品詞「名詞-普通名詞-一般」／語彙素「の」品詞「助詞-格助詞」／語彙素「開く」語彙素読み「ヒラク」品詞「動詞-一般」活用型「文語四段-カ行」活用形「命令形-一般」／語彙素「り」品詞「助動詞」活用型「文語助動詞-リ」活用形「連体形-一般」
	副本文	開化文明	品詞「対象語無し」

*「／」は短単位の境界を表す。

[5] 松村（1998、pp.521-563）の影印版による。

表 2 に示したように、振り仮名の表記を「主本文」、本行の表記を「副本文」として設定し、主本文のほうに形態論情報を付与することで、振り仮名による読みに出現する各短単位に対し形態論情報を付与した。なお、副本文への形態論情報は付与しない方針とし、すべて品詞「対象語無し」として未知語扱いとした。以上のような形態論情報の多重化を行った箇所はコーパス全体で 284 箇所、延べ 626 短単位にのぼる。

以上の形態論情報付与方法の改訂により、短単位の検索や語彙の計量的分析がより精確に行えるようになった。

4. おわりに

本発表では、『日本語歴史コーパス 明治・大正編Ⅲ明治初期口語資料』のバージョンアップについて報告した。このバージョンアップにより、当時の話し言葉や口語体書き言葉の実態を知る上で、一層有用なコーパスとすることができた。

残る課題として、今回のバージョンアップでは十分になしとげられなかった形態論情報の精度（適合率 98%）の向上とともに、形態論情報の多重化箇所の増補および副本文への形態論情報の付与があげられる。形態論情報を付与する副本文の候補の一つは、本行の表記に対し振り仮名とは別の形態論情報の付与が考えられる箇所である。例えば表 2 の 3 例目の場合、副本文である本行の「開化文明」に「開化（カイカ）」「文明（ブンメイ）」の 2 短単位の形態論情報を付与することが考えられる。このようなテキストは新バージョンで形態論情報の多重化を行った箇所だけではなく、「登楼（右振り仮名：あがつ）た」「童蒙（右振り仮名：こども）」等、多数存在する。候補のもう一つは、「激発（右振り仮名：げきはつ、左振り仮名：ヤケニナル）の徒」「恐怖（右振り仮名：けうふ、左振り仮名：ヲジヲソル）し」のような注釈的に使用される本行左側に振られた振り仮名で、本コーパス収録の資料に多く存在する。こうした候補群から形態論情報を付与する副本文を網羅的に選定するためには、その選定方針を策定し選定作業を行う必要があるが、今回のバージョンアップではそこまで至らず、副本文への形態論情報の付与は一律に見送った。この課題に対処し、資料に出現する語の情報を遺漏なく付与したコーパスに進化させることを今後検討していきたい。

参考文献

国立国語研究所（近藤明日子・市村太郎・常盤智子ほか）編（2021）『日本語歴史コーパス 明治・大正編Ⅲ明治初期口語資料』（短単位データ 0.9）
https://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html#shokikogo

近藤明日子（2021）「『日本語歴史コーパス 明治・大正編Ⅲ明治初期口語資料』（短単位データ 0.9）概説書」https://pj.ninjal.ac.jp/corpus_center/chj/meiji_taisho.html#shokikogo にて公開

松村明（1998）『増補 江戸語東京語の研究』東京堂出版

村山実和子・小木曽智信・中村壮範（2017）「形態論情報の多重化による洒落本コーパスの質的拡張」『研究報告人文科学とコンピュータ（CH）』2017-CH-114 (8)、pp.1-8

謝辞

本発表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」の成果である。また、『春秋雑誌会話篇』のコーパスデータの構築の一部は JSPS 科研費 JP17K02786「英学資料のテキストデータ化に関する研究」の成果である。