『明六雑誌コーパス』の開発　―近代語コーパスのモデルとして―

近藤 明日子　（国立国語研究所コーパス開発センター）†
小木曽 智信　（国立国語研究所言語資源研究系）††
須永 哲矢　（国立国語研究所コーパス開発センター）†††
田中 牧郎　（国立国語研究所言語資源研究系）††††

Development of *Meiroku Zasshi Corpus*: A Model of Modern Japanese Corpora

KONDO Asuko　(National Institute for Japanese Language and Linguistics)
OGISO Toshinobu　(National Institute for Japanese Language and Linguistics)
SUNAGA Tetsuya　(National Institute for Japanese Language and Linguistics)
TANAKA Makiro　(National Institute for Japanese Language and Linguistics)

１．はじめに

国立国語研究所共同研究プロジェクト（独創・発展型）「近代語コーパス設計のための文献言語研究」（プロジェクトリーダー：田中牧郎、プロジェクト略称「近代語コーパス」）では、同研究所共同研究プロジェクト（基幹型）「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘）が対象とする古代から近世までの通時コーパスと、2011 年に公開された現代語の大規模均衡コーパスである『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）との間をつなぐ位置にある、近代語のコーパスのあり方について研究を行ってきた。そして、その成果を活かし、今後の本格的な近代語コーパス構築に向けてのモデルとして『明六雑誌コーパス』を開発中であり、2012 年秋の公開を予定している。本発表では、この『明六雑誌コーパス』の概要を紹介する[1]。

２．『明六雑誌コーパス』以前の近代語のコーパス

これまでに公開された大規模な近代語のコーパスとして国立国語研究所（編）(2005)『太陽コーパス』がある。『太陽コーパス』は、1895（明治 28）年から 1928（昭和 3）年にかけて博文館から刊行された総合雑誌『太陽』の 5 カ年分 60 冊の全文を電子テキスト化し、そこにマークアップ言語である XML を用いて文書構造・文字・表記等に関する情報を付与するという、当時、近代語のコーパスとして画期的な構造を持つものであった。この『太陽コーパス』により、近代語のコーパスのあり方の方向性が示されたが、その公開から 7 年以上が経過し、その本文の電子テキスト化の方法や XML タグによって付与する情報の種類やその付与方法については、今日の研究成果や技術水準から見ると不十分な部分も多く、これをそのまま今後の本格的な近代語コーパスのモデルとするには不足の感は否めない。

† kondo@ninjal.ac.jp
†† togiso@ninjal.ac.jp
††† tsunaga@ninjal.ac.jp
†††† mtanaka@ninjal.ac.jp

[1] 『明六雑誌コーパス』は現在開発中であり、本発表で紹介する事項等は公開時までに変更される可能性がある。

そこで、「近代語コーパス」プロジェクトでは、国立国語研究所において作成済みの近代語資料の電子テキストの一つである『明六雑誌』を対象とし、『太陽コーパス』の構造を引き継ぎつつ、本プロジェクトでの最新の研究成果を盛り込んだ形でコーパス化を行い、これからの近代語コーパスのモデルとして示すことにした。それが、本発表で紹介する『明六雑誌コーパス』である。

３．『明六雑誌コーパス』の概要

『明六雑誌コーパス』は、1884（明治 7）年から 1885（明治 8）年にかけて刊行された『明六雑誌』全 43 号の全文[2]を対象とするコーパスである。『明六雑誌』は当時の洋学者によって結成された学術団体である明六社の機関誌で、社中での討論や演説を掲載した、日本の学術雑誌の先駆けの一つとされるものである。近代語の実態を知る上で重要度の高い資料であり、コーパス化の望まれるものの一つと言える。

『明六雑誌コーパス』は、『太陽コーパス』の構造を引き継ぎ、電子テキスト化した本文データに、XML を用いて文書構造・形態論・文字・表記等に関する情報を付与する構造となっている。『明六雑誌コーパス』で用いられる主な XML タグをあげたものが表 1 である。

表 1　『明六雑誌コーパス』の XML タグ（一部）

タグ名	説明
雑誌	雑誌 1 号分を表す。
記事	記事を表す。
div	雑誌タイトル等を表す。
paragraph	段落を表す。
block	記事タイトル・記事著者・小見出し等を表す。
figureBlock	図表を表す。
割書	割書された文字列を表す。
引用	当該記事以外の文献や会話・心話等の引用を表す。
sentence	文を表す。
SUW	語（短単位）を表す。
ruby	振り仮名を表す。
外字	文字集合 JIS X 0213 外の文字を表す。
包摂	拡張包摂基準により包摂した文字を表す。
注	誤植・濁点脱落を校訂したことを表す。
小書	小書きされた仮名文字を表す。
敬意欠字	皇室への敬意を示す欠字を表す。
踊字	踊字を表す。
pb	原本の改ページ位置を表す。
lb	原本の強制改行位置を表す。

４．『明六雑誌コーパス』の特徴

次に、『明六雑誌コーパス』の特徴について、特に『太陽コーパス』からの改良という

[2] 表紙、目次、識語、奥付、図表中の文字列は除く。

観点から説明する。

４．１．文字集合 JIS X 0213 の採用

『明六雑誌コーパス』では、本文の電子テキスト化において JIS X 0213:2004 に準拠した文字集合を採用した。『太陽コーパス』の開発中に JIS X 0213:2000 が制定されたが、当時はまだ実装が限られており普及が見通せなかったため、『太陽コーパス』では文字集合としての採用は見送られ、代わりに JIS X 0208:1997 が採用された（田中、2005）。しかし、その後の JIS X 0213 の普及により、BCCWJ のように JIS X 0213 に準拠した文字集合を用いるコーパスも登場した。これを受けて『明六雑誌コーパス』でも JIS X 0213:2004 を採用することにしたものである。ただし、JIS X 0213:2004 そのものを文字集合とするのではなく、JIS 包摂規準の例外にあたる漢字[3]や動作環境によっては適切に表示されない問題のある漢字[4]については使用しないこととした。

これにより、非漢字では例えば合字「㌧」（コト）や繰り返し符号「〳〵」「〳〵゛」「〲」、漢字では第 3・第 4 水準漢字の電子化が可能となった。須永・堤・高田（2011）では、『明六雑誌』の本文中の漢字延べ 137,897 字の内訳について表 2 のように報告しており、ここから、JIS X 0213 の採用により、延べ 201 字の第 3・第 4 水準漢字の電子化が新たに可能となったことがわかる。

表 2　JIS X 0213 文字集合と『明六雑誌コーパス』漢字

文字区分	延べ字数
JIS X 0213	135,797
第 1 水準漢字	117,643
第 2 水準漢字	17,953
第 3 水準漢字	118
第 4 水準漢字	83
外字	2,100
計	137,897

なお、JIS X 0213 を用いてもなお外字となる漢字延べ 2,100 字については、言語研究での実用性を鑑み、包摂規準の拡張や別字での代用を行うことでできるだけ電子化する方針をとり、最終的に外字として〓表示になるものは延べ 31 字までに減らした（須永・堤・高田、2011）。

４．２．形態論情報の付与

BCCWJ で現代語の形態素解析辞書「UniDic」を用いて形態論情報の付与がなされたように、現代語のコーパスでは、形態素解析の技術を用い、文字列を語に分割し、語彙素（見出し語に相当）や品詞といった形態論情報を言語研究への利用に適した形で高精度に付与することが可能となっている。近代語の資料では、現代語以上に語形や表記のバリエーションが豊富であることを考えると、近代語のコーパスでは、現代語のコーパス以上に形態

[3] 康熙別掲字（104 字）と UCS 互換字（10 字）。

[4] CJK 統合漢字拡張Ｂに符号位置が割り当てられる文字（302 字）。

論情報の付与の必要度が高いと言える。しかし、近代語を含めた古い時代の日本語資料について、実用的な精度で形態素解析することは『太陽コーパス』開発時には実現されておらず、『太陽コーパス』では形態論情報の付与はなされなかった。それが近年になり近代の文語論説文を対象とする形態素解析辞書「近代文語 UniDic」（小木曽、2009）が開発され、近代語のコーパスに形態論情報を付与する環境が整備されつつある。『明六雑誌コーパス』では、その最新の技術を活かし、「近代文語 UniDic」を用いて本文を形態素解析した後、人手修正を加えたものを SUW（Short-Unit Word）要素として付与した。解析に用いた「近代文語 UniDic」は、①ゆれの少ない斉一な単位である「短単位」を解析単位とする、②表記の揺れや語形の変異にかかわらない見出し語が付与される、③和語・漢語・外来語・混種語といった語種情報が付与される、といった日本語研究に適した特長を持つ。また、「近代文語 UniDic」の構造は開発の基盤となった現代語用「UniDic」と共通するため、それぞれの UniDic で解析したコーパスを利用した通時的研究も可能となる。

『明六雑誌コーパス』の SUW 要素の例

```
<SUW orthToken="洋字" lForm="ヨウジ" lemma="洋字" pos="名詞-普通名詞-一般" form="ヨウジ" pronToken="ヨージ" wType="漢" start="100" end="120" orderID="80" section="v">洋字</SUW>
<SUW orthToken="を" lForm="ヲ" lemma="を" pos="助詞-格助詞" form="ヲ" pronToken="オ" wType="和" start="120" end="130" orderID="90" section="v">を</SUW>
<SUW orthToken="以" lForm="モツ" lemma="持つ" pos="動詞-一般" form="モツ" cType="文語四段-タ行" cForm="連用形-促音便" pronToken="モッ" wType="和" start="130" end="140" orderID="100" section="v">以</SUW>
<SUW orthToken="て" lForm="テ" lemma="て" pos="助詞-接続助詞" form="テ" pronToken="テ" wType="和" start="140" end="150" orderID="110" section="v">て</SUW>
```

SUW 要素を利用して『明六雑誌コーパス』の語数を概観すると、延べ語数約 18 万語（記号類除く）、異なり語数（語彙素レベル）約 1 万 2 千語となる。

４．３．文情報の付与

近代語のコーパスの開発では、形態論情報以外にも言語の階層的構造に適切に対応した情報の付与が望まれる。例えば、文という単位について情報の付与は、近代語のコーパスにおいても優先度の高い事項と考えられる。現代語の文字資料であれば、文の末尾は、多くの場合「。」「！」等の記号や論理改行によって明示され、それに基づきコンピュータが自動的に文認定をすることも可能であり、BCCWJ でもそのように文認定が行われている（山口・高田・北村・他、2011、pp.136-138）。しかし、まだ句読法の確立していない時期の近代語の資料では、文末を示す記号を全く用いない文章や、「、」を句点・読点の両義に用いる文章など、様々な形態があり、コンピュータによる自動的な文認定が今日に至るまで実現していない。そのため、『太陽コーパス』では文という単位に関する情報の付与を事実上諦め、代わりに「、」と「。」を手がかりとして自動的に認定された単位を s 要素として情報付けを行った。しかし、結果的に s 要素は言語研究で利用しづらい情報となってしまっている。

『太陽コーパス』の s 要素の例

```
<s>詩仙堂　仝村に在り、</s><s>石川丈山の山荘なり、</s><s>遺物少からず、</s><s>
```

堂の四壁に、</s><s>漢、</s><s>晋、</s><s>唐、</s><s>宋の詩人三十六輩の像を畫き、</s><s>（狩野尚信の筆）自から其人の詩を書きて掲げたり、</s>

『明六雑誌コーパス』ではこの問題を克服するため、人手による文認定を行い、sentence 要素として情報付けを行った。これにより、文という単位に基づいたコーパス利用が可能となっている。

『明六雑誌コーパス』の sentence 要素の例

<sentence>　其他語格の若きは後日の成功を待つべし</sentence><sentence>右聊か愚考を陳じ諸先生の可否を請ふ</sentence><sentence>敢て採用を望むにあらずと雖ども諸先生幸に電覽を賜はゞ幸甚</sentence>

４．４．コーパスの公開形式

『明六雑誌コーパス』の公開形式は次の 3 種類を予定している。

① XML ファイル
② 形態論情報タブ区切りデータ
③ 「ひまわり」版

①はコーパスの全情報（本文データと XML タグによる付加情報）を納めたデータである。②は、①から特に形態論情報（SUW 要素）を抽出し表形式に整形したデータで、語彙研究等への活用を想定し作成した。③は XML 文書全文検索システム「ひまわり」に搭載した形式で、GUI による簡便な検索が可能であり、コーパス利用の幅が広がることを期待して作成した。「ひまわり」は当初『太陽コーパス』のための検索システムとして開発されたものであるが、『明六雑誌コーパス』への適用にあたっては、本文データに対する文字列検索だけでなく、語彙素・語彙素読み・品詞・語種といった形態論情報に対する検索にも対応した形に整備した（図 1、稿末）。形式②③での公開は、『明六雑誌コーパス』に付与した形態論情報を活かした新規の試みとなる。

５．おわりに

以上、『明六雑誌コーパス』の概要および特徴について紹介した。従来の近代語のコーパスでは不十分であった点を改良し、今後の近代語コーパス構築に向けての一つのモデルを示すことができたと考える。

ただし、当然のことながら、本格的な近代語コーパス構築にあたっては、まだ克服すべき課題も多い。『明六雑誌コーパス』では学術雑誌という一媒体・一ジャンルのコーパス化のモデルを示し得たに過ぎず、新聞・小説等の他媒体・他ジャンルの資料のコーパス化の方法については、別途検討が必要である。タグ付けする情報の種類や付与方法については、BCCWJ や今後開発が予定されている古代から近世までの通時コーパスとの連続性を考慮しつつ、更に改変していく必要があろう。また、『明六雑誌コーパス』最大の特徴である形態論情報についても、「短単位」による SUW 要素だけでなく、より長い単位である「長単位（Long-Unit Word）」による LUW 要素の付与（BCCWJ では実現している）も目指していく必要がある。さらに、『明六雑誌コーパス』はそれほどデータ量が多くなかったため、文の認定等、全面的に人手に頼る作業も可能であったが、大規模なコーパス構築に際しては、そうした作業の自動化について研究を進める必要が生じるかもしれない。これらの残された課題を一つ一つ克服し、今後の近代語コーパスの開発の実現に着実につなげていきたい。

文　献

小木曽智信（2009）『科学研究費補助金研究成果報告書　近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』（http://dl.dropbox.com/u/73297026/report/unidic-MLJ_report2009.pdf よりダウンロード可能）

国立国語研究所（編）（2005）『太陽コーパス―雑誌『太陽』日本語データベース』博文館新社

須永哲矢、堤智昭、高田智和（2011）「明治前期雑誌の異体漢字と文字コード―『明六雑誌』を事例として―」人文科学とコンピュータシンポジウム論文集、2011:8、pp.381-388

田中牧郎（2005）「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―、pp.1-48、博文館新社

山口昌也、高田智和、北村雅則、間淵洋子、大島一、小林正行、西部みちる（2011）『特定領域研究「日本語コーパス」平成 22 年度研究成果報告書　『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2』(JC-D-10-04)

関連 URL

「近代語コーパス」プロジェクト	http://www.ninjal.ac.jp/research/project/b/kindaigo/
近代文語 UniDic	http://www2.ninjal.ac.jp/lrc/index.php?UniDic
ひまわり	http://www2.ninjal.ac.jp/lrc/index.php
UniDic	http://download.unidic.org

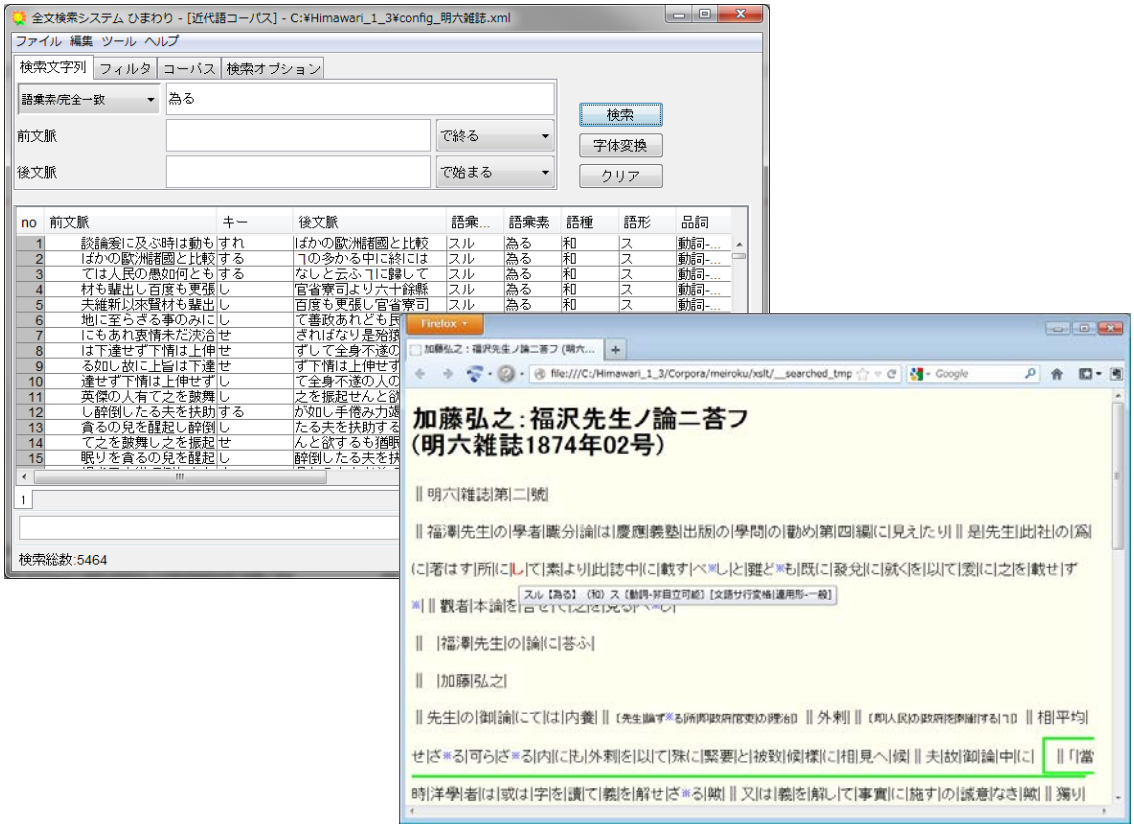

図１　「ひまわり」による『明六雑誌コーパス』の検索画面（左）および文脈確認画面（右）