形態論情報付き近代語コーパスのアノテーション
—『明六雑誌コーパス』を例として—

近藤明日子　小木曽智信　須永哲矢　田中牧郎
{ kondo, togiso, tsunaga, mtanaka } @ninjal.ac.jp

人間文化研究機構 国立国語研究所

１．はじめに

『現代日本語書き言葉均衡コーパス』[1]（以下、BCCWJ）の一般公開により、現代日本語のコーパス利用の環境が飛躍的に整備された今日、それに続けて、より古い時代の日本語資料の大規模コーパス構築に対する気運が高まっている。その代表的事例として、国立国語研究所の「通時コーパスの設計」プロジェクト[2]で古代から近世までの通時コーパスの構築に関する研究が進行していることがあげられる。また同研究所では、通時コーパスとBCCWJとの間をつなぐ位置にある近代語のコーパスのありかたについても研究を行ってきており、その中で、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』[3]を2012年に公開した。

近代語資料のコーパス化では、現代語とは異なる歴史的資料に特有の課題が存在し、アノテーション作業でもその課題に必然的に向き合わなければならない。現代語のコーパスではごく当たり前に施されるアノテーションが、近代語資料では技術的に実現困難であったり、実現するために現代語のコーパスでは必要のない前処理およびそれに付随するアノテーションが必要となったりする。

本稿では、『明六雑誌コーパス』構築の事例を通して、アノテーションにまつわる近代語資料のコーパス化特有の課題とその対処法について見ていく。

２．『明六雑誌コーパス』の概要

『明六雑誌コーパス』は、1874（明治7）年から1875（明治8）年にかけて全43号刊行された雑誌『明六雑誌』の全文コーパスである。『明六雑誌』は学術啓蒙を目的に結成された明六社の機関誌で、森有礼・津田真道・西周・西村茂樹・中村正直・加藤弘之・福沢諭吉・箕作麟祥ら16名の執筆する、西洋の近代思想を普及するために書かれた広範な論説がおさめられている。思想史上の重要資料であるとともに、明治前期の日本語の実態を知る上でも重要な資料と判断される。

『明六雑誌コーパス』は、約18万語からなる本文テキストにXMLを用いて文書構造・形態論・文字・表記等に関するアノテーションが施されている。使用するXMLタグセットは、近代語コーパスの先駆的存在である国立国語研究所（編）(2005)『太陽コーパス』のものを受け継ぎつつ、BCCWJや通時コーパス[4]との連続性を考慮して定めた。その主なタグを言語構造に沿って示すと、図1のようになる。

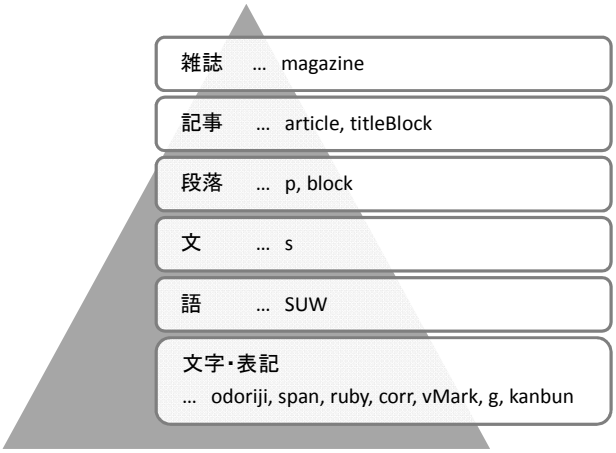

図１　『明六雑誌コーパス』の主なタグ

タグセットの全容については近藤・田中（2012）に詳しいが、本稿では特に近代語の資料であるがゆえに現代語の場合とは異なる対処法が必要となったアノテーションを中心に述べていく。

３．原本本文の校訂に関するアノテーション

コーパスの本文テキストの作成において、コーパスの利便性やアノテーションの精度を高めるために原本本文の校訂が行われる場合がある。例えば、本文の誤植の校訂などは現代語のコーパスでもしばしば行われる。それに加えて、『明六雑誌』のような近代語資料の本文を電子テキスト化する場合、現代語ではめったに見ることのない表記が出現するため、それに対処する種々の校訂およびそれに付随するアノテーションを行わなければならない。

３．１．漢字片仮名交じり文

現代語では文章を書くときに漢字平仮名交じりの書記体がとられることがほとんどである。しかし、『明六雑誌』では 155 記事中 154 記事が漢字片仮名交じりの書記体で書かれる。そこで、漢字片仮名交じりの場合、本文テキストは片仮名を平仮名に置き換えて作成した。原本での書記体に関する情報は、記事1編に対して付与された article タグの属性として明示的に表した。ただし、外来語のように片仮名表記のままテキスト化したほうがよいと判断した文字列については、span タグを付与して片仮名のままとした。コーパス全体で 509 個の span タグが付与された。

テキスト化の例[5]

```
<article title="駁旧相公議一題" author="西周" style="文語" script="漢字カタカナ">
（…中略…）故に<span type="カタカナ">ルウソウ</span>氏の説に據り政府を以て全く約束より成るとするも政府の事を與知するの權利は租稅を出すと相對するの權利に非ず（…中略…）
</article>
```

３．２．濁点無表記

現代語では濁音を表す仮名には必ず濁点を振ることが表記法として定着している。しかし『明六雑誌』では、濁音を表す場合であっても濁点を振らない濁点無表記の仮名が多く使用される。そこで、本文テキストは濁点付きの仮名に置き換えて作成し、vMark タグを付与して、原本では濁点無表記であることを明示的に表した。コーパス全体で 6645 個の vMark タグが付与された。

テキスト化の例

```
然<vMark>ど</vMark>も此弊に因て斯世の民幸福を蒙る𠆢を得<vMark>ず</vMark>衰弊の極救藥す<vMark>べ</vMark>から<vMark>ざ</vMark>るに至るは亦獨り政府の罪たるのみなら<vMark>ず</vMark>抑其國人民自己世道上の罪にて苟も賢智の徒たらんとする者は先ん<vMark>じ</vMark>て之を救ふ𠆢なくん<vMark>ば</vMark>亦世道上に於て其罪なしと謂ふ<vMark>べ</vMark>から<vMark>ず</vMark>
```

３．３．踊り字

直前の文字を繰り返すことを表す踊り字は、現代語では「々」1 種が用いられる程度であるが、『明六雑誌』ではその他にも「ゝ」「ゝ」「ゞ」「ヽ」「ヾ」「/ \」「/〃\」のような多種のものが出現する。そこで、本文テキストは踊り字で繰り返される文字列で入力し、odoriji タグを付与して、原本で使われている踊り字の種類を明示的に表した[6]。コーパス全体で 393 個の odoriji タグが付与された。

テキスト化の例

```
印書の發明より後凡そ百年になん<odoriji originalText="ゝ">なん</odoriji>として
```

```
己を捨て<odoriji originalText="ヽ">て</odoriji>人に従ふは大舜の美徳
```

```
府や縣でまち<odoriji originalText="/ \">まち</odoriji>なる法制を立て
```

３．４．漢文体

『明六雑誌』では、一部の語句が漢文体で書き表され、読み下すに際して、下の字から上の字に返って読む「返読」や助詞等を補って読む「補読」が必要な文字列が見られる。例えば、「やむをえざる」という語句を漢文体で「不得已」と書き表すような場合である。これについては、読み下した後の形でテキスト化し、kanbun タグを付与して、返読・補読の状況を明示的に表した。コーパス全体で 118 個の kanbun タグが付与された。

テキスト化の例

```
然ども是<kanbun type="返読前" originalText="不" id="00008"/><kanbun type="返読前" originalText="得" id="00009"/>已<kanbun type="補読">を</kanbun><kanbun type="返読後" id="00009">得</kanbun><kanbun type="返読後" id="00008">不</kanbun>の時なり
```

３．５．異体字

『明六雑誌コーパス』のテキスト入力では文字集合として JIS X 0213 を採用し、漢字の字体の包摂規準についても JIS X 0213 のものに準拠した。ただし、『明六雑誌』では JIS の包摂規準の適用できない字体差を持つ異体字が多数出現する。例えば、現代通用している字体「序」と明らかに同字であるが、JIS の包摂規準では包摂できない字体差のある異体字が『明六雑誌』には出現する（図 2）。

図 2　『明六雑誌』に出現する「序」の異体字

このような異体字をすべて JIS 外字として「=」で入力すればコーパスの利便性を大きく損ねることになる。そこで、『明六雑誌』用に新たに包摂規準を 28 種追加して字体包摂を行い、通用字体を用いてテキスト化し、g タグを付与して、原文では通用字体とは異なる字体であることを明示的に表した。図 1 の「序」の異体字は次のようにテキスト化した。

テキスト化の例

```
<g type”包摂”>序</g>
```

また、JIS 外字でかつ追加包摂規準の適用外の字体であっても、意味・用法の類似する他の漢字での代用が可能な場合は、その代用字を用いてテキスト化することで、「=」入力を極力避けるようにした。そして、代用字には g タグによって原本での字体に関する情報を付与した。例えば、『明六雑誌』には「減」のさんずいがにすいになっている字体「減」（Unicode コード：51CF）が出現するが、これは次のようにテキスト化した。

テキスト化の例

```
<g type="外字" ref="U+51CF">減</g>
```

その結果、包摂基準の追加により新たに延べ 1774 字が、別字代用により新たに延べ 295 字が JIS 内字を用いてテキスト化され、最終的に「=」のまま残されたのは延べ 31 字（コーパス全体の漢字の延べ字数 137866 字の 0.02%）となった（須永、2012）。

４．言語構造に関するアノテーション

以上のような本文の校訂およびそれに付随するアノテーションの作業を行った上で、言語構造にかかわるアノテーションを施すことになるが、ここでも近代語資料特有の課題に対処する必要があった。

４．１．文境界のアノテーション

現代語では、「。」によって文末を明示するという句読法が定着しており、それに基づいて文境界のアノテーションの自動化が実現している。しかし、句読法の確立していない時期の近代語資料では、文末を示す記号を全く使わない文章や、「、」を句点・読点の両義に用いる文章など、様々な形態があるため、文境界のアノテーションの自動化は今日に至るまで実現していない。そうした状況の中で、例えば『太陽コーパス』では便宜的に「、」「。」を手がかりとするアノテーションが試みられたが、しばしば過剰に区切られ文未満の単位になっているなど、実用性のあるアノテーションとは言えなかった。『明六雑誌』も句読法の確立していない時期の資料であり、文境界に関して自動的なアノテーションはできない。しかし、その必要性は高いと判断し、人手による文境界の認定を行い、各文に s タグを付与した。コーパス全体で 9172 個の s タグを付与した。

テキスト化の例

```
<s>　其他語格の若きは後日の成功を待つべし</s><s>右聊か愚考を陳じ諸先生の可否を請ふ</s><s>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</s>
```

４．２．形態論レベルのアノテーション

現代語のコーパスでは、形態素解析によって、単語の読みや品詞などの形態論レベルのアノテーションの自動化が実現している。一方、近代を含む古い時代の資料のコーパスでは、従来は高精度の形態素解析が困難であったため、形態論レベルのアノテーションはその必要性は意識されつつも事実上不可能であった。しかし、近代の文語文に対応した形態素解析辞書「近代文語 UniDic」[7]（2008-）や、平安時代の和文（仮名文）に対応した「中古和文 UniDic」[8]（2010-）の公開により、一部の歴史的資料について実用的な精度で形態素解析を行うことが可能になってきた。

『明六雑誌コーパス』では、この「近代文語 UniDic」による形態素解析結果に人手による修正を加え、SUW タグによってアノテーションを行った。一般公開された近代語のコーパスでは、形態論レベルのアノテーションが実施された初めての例となる。「近代文語 UniDic」は、①ゆれの少ない斉一な単位である「短単位」（Short-Unit Word）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（語彙素）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、といった特長を持つ辞書である。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」[9]や「中古和文 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。

テキスト化の例

```
<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" kanaToken="ヨウジ" orth="洋字" wType="漢" start="100" end="120" orderID="80">洋字</SUW>
<SUW orthToken="を" lForm="ヲ" lemma="を" pos="助詞-格助詞" form="ヲ" pronToken="オ" kanaToken="ヲ" orth="を" wType="和" start="120" end="130" orderID="90">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" kanaToken="モッ" orth="以つ" wType="和" start="130" end="140" orderID="100">以</SUW>
<SUW orthToken="て" lForm="テ" lemma="て" pos="助詞-接続助詞" form="テ" pronToken="テ" kanaToken="テ" orth="て" wType="和" start="140" end="150" orderID="110">て</SUW>
```

「近代文語 UniDic」による解析結果をそのまま利用するのではなく、人手による修正を加えたのは、明治後期以降のデータを中心に整備してきた「近代文語 UniDic」による解析では、明治前期の『明六雑誌』に特有の語が未知語となり、誤解析が多く生じたためである。そこで、新たに辞書登録を行いながら人手による修正作業を行ったところ、新規登録語は異なりで約 3700 語となった。これは、コーパス全体の異なり語数約 15500 語の 24%にあたる。ただし、新規登録語の大部分は低頻度語であり、新規登録語は延べで約 5600 語、コーパス全体の延べ語数約 180500 語の 3.2%にとどまる。(小木曽、2012)。

未知語の新規登録および誤解析の修正にあたっては、近代語資料のために独自の短単位の認定規程を定める必要も生じた。例えば、「異なる」は現代語の短単位規程（小椋・小磯・冨士池ほか、2011）では動詞「異なる」として 1 短単位と認定されるが、近代語ではその使用実態から名詞「異」と助動詞「なり」の 2 短単位と認定するよう規程の修正を行った。このように独自に設定した規程は、単位境界認定・同語異語判別・品詞認定・語形認定等と多岐にわたる（須永・近藤、2012)。

5．おわりに

以上、『明六雑誌コーパス』構築の事例を通して、アノテーションにまつわる近代語資料のコーパス化特有の課題とその対処法について述べた。今後の本格的な近代語コーパスの構築に向けて、近代語資料の課題の多くを見つけ出し、その対処法について雛形を示すことに成功したと言える。

ただし、本格的な近代語コーパス構築が開始されれば、また新たな課題に直面することは必至である。『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。その中で、アノテーションについても、BCCWJ や通時コーパスとの連続性を考慮しつつ改変していく必要がある。また、『明六雑誌コーパス』の最大の特長ともいえる形態論レベルのアノテーションについても、「短単位」による SUW タグだけでなく、より長い単位である「長単位」による LUW（Long-Unit Word）タグによるものも目指していく必要がある（BCCWJ では実現している)。さらに、『明六雑誌コーパス』はそれほどテキスト量が多くなかったため、全面的に人手に頼った作業をすることも可能であったが、大規模なコーパス構築に際しては、一部の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。岡（2012）によれば、濁点無表記の仮名に対するアノテーションの自動化については実用段階に近づいている。文境界のアノテーション等についても自動化の研究が進むことが期待される。

これらの新たに生じるであろう課題に対して一つ一つ対処法を提示し、今後の近代語コーパスの実現に着実につなげていきたい。

参考文献

市村太郎・河瀬彰宏・小木曽智信（2012）「近世口語テキストの構造化とその課題」『情報処理学会研究報告』Vol.2012-CH-96, No.1, pp.1-8

岡照晃（2012）「近代文語論説文を対象とした濁点の自動付与アプリケーション」『第 2 回コーパス日本語学ワークショップ予稿集』pp.305-314（http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no2_papers/JCLWorkshop2012_2_web.pdf よりダウンロード可）

小木曽智信（2012）「近代語テキストの形態素解析」『近代語コーパス設計のための文献言語研究　成果報告書』pp.83-92（http://www.ninjal.ac.jp/corpus_center/cmj/doc/05ogiso.pdf よりダウンロード可）

小椋秀樹・小磯花絵・冨士池優美・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)』、国立国語研究所

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

近藤明日子・田中牧郎（2012）「『明六雑誌コーパス』の仕様」『近代語コーパス設計のための文献言語研究　成果報告書』pp.118-143（http://www.ninjal.ac.jp/corpus_center/cmj/doc/07kondo.pdf よりダウンロード可）

須永哲矢（2012）「近代語文献を電子化するための異体字処理」『近代語コーパス設計のための文献言語研究　成果報告書』pp.65-82

須永哲矢・近藤明日子（2012）「近代語コーパスのための形態論情報付与規程の整備」『近代語コーパス設計のための文献言語研究　成果報告書』pp.93-117（http://www.ninjal.ac.jp/corpus_center/cmj/doc/06sunaga.pdf よりダウンロード可）

[1] http://www.ninjal.ac.jp/corpus_center/bccwj/
[2] http://www.ninjal.ac.jp/research/project/a/corpus/
[3] http://www.ninjal.ac.jp/corpus_center/cmj/meiroku/
[4] 特に、通時コーパスの一部として計画されている近世口語資料のコーパスのタグセット（市村・河瀬・小木曽、2012）との共通化を目指した。
[5] 以下、例として示すテキストでは、説明に不要なタグは省略する。
[6] ただし、「種々」「云ヒ」のように 1 短単位中で 1 文字を繰り返す「々」「ヒ」はそのままテキスト化した。
[7] http://www2.ninjal.ac.jp/lrc/index.php?UniDic
[8] http://www2.ninjal.ac.jp/lrc/index.php?UniDic
[9] http://sourceforge.jp/projects/unidic/