近代語コーパスにおける資料選定の考え方

著者	田中 牧郎
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	13-26
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/10.15084/00002762

近代語コーパスにおける資料選定の考え方

田中　牧郎（国立国語研究所言語資源研究系）[1]

１．はじめに

国立国語研究所が 2011 年に公開した『現代日本語書き言葉均衡コーパス』は、代表性を担保する周到なサンプリングなされている点において（前川 2008）、これまでの日本語コーパスとは一線を画している。今後構築されるコーパスは、この代表性の担保にどのように対応するかが問われていくことになろう。一方、国立国語研究所では 2009 年から、上代から近世までの日本語の歴史をたどることのできる「通時コーパス」の設計に着手しているが、古典作品を対象とするコーパスでは、ランダムサンプリングを重視する代表性よりも、作品のアイデンティティを重視して資料の独自性を吟味する立場が重要になると見通されている（近藤 2012）。

この「通時コーパス」が対象とする近世までと、『現代日本語書き言葉均衡コーパス』が対象とする現代とをつなぐ位置にある近代における日本語を対象とした「近代語コーパス」の設計を考えるとき、資料選定において「代表性」や「独自性」はどのように考えていけばよいだろうか。この問いについて考えるには、近代語の資料のあり方を分析することを通して研究していくことが必要だろう。本稿では、近代語のコーパスを設計する際の資料選定の考え方を問題にする。

２．『太陽コーパス』から近代語コーパスへ

近代語のコーパスについて、国立国語研究所は既に『太陽コーパス』（国立国語研究所 2005a）を構築して公開している[2]。『太陽コーパス』は、言文一致を経て、口語体による書き言葉が安定し普及する時期（明治時代後期～大正時代）の書き言葉を代表できるコーパスとして作られたものであり、月刊の総合雑誌『太陽』（博文館）の、明治 28（1895）年、明治 34（1901）年、明治 42（1909）年、大正 6（1917）年、大正 14（1925）年の 60 冊分について、その全文（著作権処理ができなかった記事を除く）を対象にしたものである。年次が６年または８年刻みとなっている点はサンプリングコーパスと言えるが、対象になった年次の全体を含んでいる点では全文コーパスとも言える。

コーパスの重要な要件のひとつである代表性の担保については､対象とした総合雑誌『太陽』が、分量の多さ、ジャンルの広さ、執筆陣の多彩さ、読者層の厚さの四点で、当時の文献資料としては格別の価値を持っていることから、『太陽コーパス』にも「代表性」が備わっていると見ることもできる（田中 2005）。実際に例えば、図１は、『太陽コーパス』のジャンル（NDC）別の記事数とその比率を『現代日本語書き言葉均衡コーパス』（出版サブコーパスの書籍、図２）のサンプル数（丸山ほか 2011）と比較できる形で示したものであるが、社会科学が最も多く、文学がこれに次ぐところなど、『現代日本語書き言葉均衡コーパス』（出版サブコーパス書籍）と『太陽コーパス』は似ている面があることが分かるだろう。

しかし、大きく異なっている点として、『現代日本語書き言葉均衡コーパス』が、現代

[1] mtanaka@ninjal.ac.jp

[2] 同種のコーパスに、国立国語研究所『近代女性雑誌コーパス』があり、CD-ROM で公開している（その情報は、http://www.ninjal.ac.jp/corpus_center/）。これは、『太陽コーパス』とほぼ同時期の女性を読者とした３誌（『女学雑誌』『女学世界』『婦人倶楽部』）を対象とした約 120 万語の小規模なコーパスである。

書き言葉の種々の媒体を母集団に設定して、ランダムサンプリングが行われているのに対して、雑誌『太陽』は、そのような手続きを経て選ばれたものではないという点があげられる。むしろ、先に述べた、雑誌『太陽』が持つ、分量・ジャンル・執筆陣・読者層の四点の特徴がこの時期のコーパスの対象としてふさわしいと見た、「独自性」を重視した選定であったと言うこともできる。

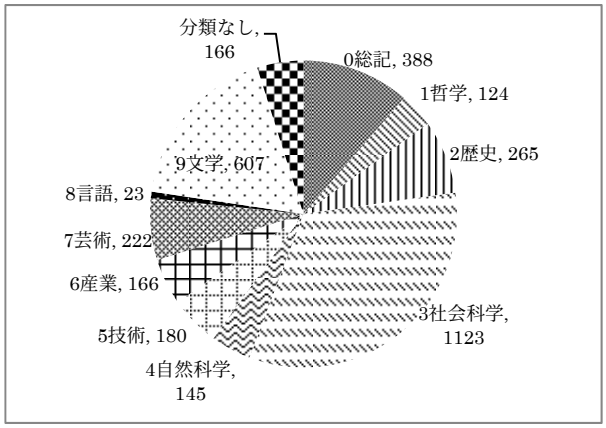

図１『太陽コーパス』のジャンル

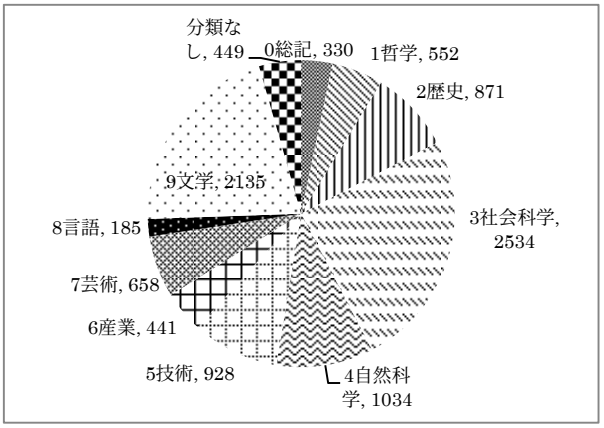

図２BCCWJ 出版サブコーパス書籍のジャンル

古典語と現代語をつなぐ位置にある近代語を対象としたコーパスに含める資料を決めていくには、「代表性」と「独自性」の両面を考慮することが望まれるのではないか。既にある近代語のコーパスとしての『太陽コーパス』を踏まえつつも、多様な近代語の資料の実態を整理した上で、コーパスの資料のあり方を考えていくことが必要である。

３．近代語の資料リストの作成

３．１　「国語辞典編集準備資料」

『太陽コーパス』は、国立国語研究所の史的国語辞典編集事業の系譜から生まれたものである。その史的国語辞典編集を行う準備研究のために設置された国語辞典編集準備室によって、用例採集の対象とすべき近代語資料をまとめた目録が、三つ作成されている。

(1)『用例採集のための主要文学作品目録』（国語辞典編集準備資料２、1980 年）
主要文学全集に収録された、明治元（1868）年～昭和 41（1966）年の 1506 作品をリスト化したもので、有識者 10 名が投票を行い得点化し、高得点の主要作品 139 点が「用語索引を作る作品」として選定されている。
(2)『用例採集のための主要雑誌目録』（国語辞典編集準備資料３、1983 年）
国立国会図書館の和雑誌目録の中から、昭和 25(1950)年以前に創刊され 20 年間以上発行されている雑誌2778件をリスト化したもので、有識者10名が投票を行い得点化し、高得点の主要誌 120 点が選定されている。
(3)『用例採集のためのベストセラー目録』（国語辞典編集準備資料４、1984 年）
ベストセラーに関する参考書に掲載された、明治元（1868）年～昭和 53（1978）年の書籍、1882 件をリスト化したもの。このリストについては得点化や主要作品の選定は、行われていない

実際の史的国語辞典編集のための用例採集事業[3]は紙媒体で開始されたが、すべての用語・用例を採集できるようにする「総索引方式」と、任意の用語・用例を選抜して採集する「スカウト式」の二段構えで着手された。総索引方式では国定国語教科書を対象とした

[3]「日本大語誌」と呼ばれるこの事業の記録は最近、飛田（2012）として公表された。

『国定読本用語総覧』（国立国語研究所 1985-1997 として完成公開）が作成され[4]、スカウト式では雑誌『太陽』の用例採集が進められた。ところが、この事業に本格的にコンピュータが導入されたことがきっかけとなって、『太陽』は途中からスカウト式を止めコーパス化の対象にされ、『太陽コーパス』が作成されたのである[5]。『太陽コーパス』の完成に先立って史的国語辞典編集のための用例採集作業は中断された形になっているが、実質的にはコーパス構築事業にその考え方は継承されており、平成 21 年度から通時コーパスと近代語コーパスの設計に関わるプロジェクトが同時に始まったことで、その側面はより色濃くなってきたと言える。近代語コーパスに含めるべき資料を検討する際に、上記の目録類は第一に参考にすべきものである。

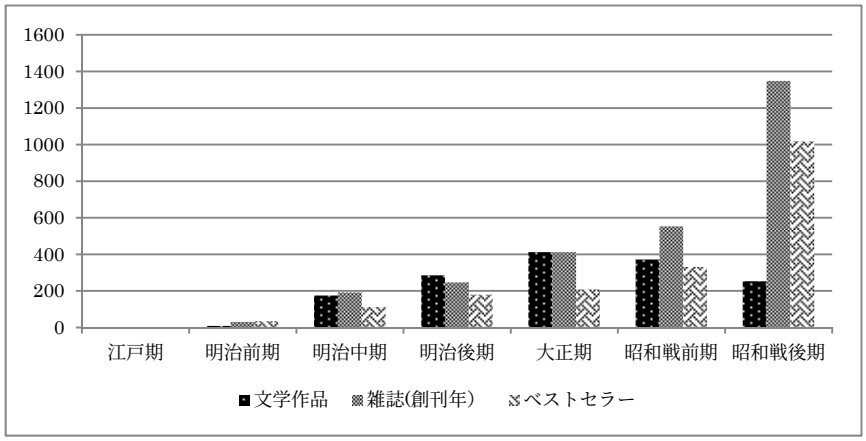

図３　国語辞典編集準備資料に掲載された資料数（時代別）

図３は、上記の三つの目録に掲載された資料の数を時代別にまとめたものである。時代区分は、明治から大正期をほぼ 15 年ごとに４つに区切り、昭和期を戦前と戦後に分けた。

明治前期：明治元～15(1868-1882)年
明治中期：明治 16～30(1883-1897)年
明治後期：明治 31～44(1898-1911)年
大正期：大正元～14(1912-1925)年
昭和戦前期：昭和元～20(1926-1945)年
昭和戦後期：昭和 21(1946)年～

明治・大正期と昭和期とで時間幅が異なっていて比較しにくい面はあるが、雑誌とベストセラーは時代を追って増加傾向にあり、文学作品は大正期まで増加し、昭和期に入って減少していると見ることができよう。こうした傾向はそれぞれの媒体が各時代にどの程度の量発行されたかという実態を反映している面もあるかもしれないが、直接的には目録作成の材料に何が使われたかということを反映しているのではないかと思われる。また、明治前期・中期が全般的に少ないのは、この目録作成が 20 世紀を主たる対象にしていたということも関係しよう。

雑誌とベストセラーは、『現代日本語書き言葉均衡コーパス』でも対象としており、文学作品は『現代日本語書き言葉均衡コーパス』では書籍の下位に NDC 分類に即して配置されている。『現代日本語書き言葉均衡コーパス』にはこのほか、新聞、教科書、白書、広報

[4]教科書については資料目録は作成されていない。国定読本の他には国定算数教科書の用語索引が作られたが、公開されてはいない（木村・加藤・田中 1999）。

[5]この間の経緯は、木村・加藤・田中（1999）参照。

誌、Yahoo!知恵袋、Yahoo!ブログ、法律、国会会議録などが含まれている。このうち、新聞、教科書、国会会議録などは、史的国語辞典編集のための資料目録作成は行われていないが、用例採集作業の対象として研究は行われており、対象資料の候補にはなっていた。一方、白書、広報誌という媒体は、昭和戦前期までは存在しておらず、Yahoo!知恵袋、yahoo!ブログのようなインターネット上の文章もまた同様である。しかし、政府や役所から国民や住民に告知する文書は戦前にもあり、知恵袋やブログを私的性格の強い文章と考えれば、手紙や日記など近代から存在していた媒体は多い。近代語コーパスの対象に含めるべき資料の候補は、さらに幅を広げて検討していくことが望まれよう。

３．２　叢書類

国語辞典編集準備資料の目録３冊は、近代語コーパスに含めるべき資料を考えるのにきわめて有益な資料であるが、不十分なところも多いため、他の材料を用いて増補していくことが必要である。特に、明治前期の資料の手薄さが目立つため、まずはこの時期の資料を豊富におさめる叢書類をもとに資料リストを増補していくことにした。用いた叢書は次の４つである。

(1) 明治文化全集　全 24 巻（1927～1932 年、日本評論社）
(2) 明治文化資料叢書　全 12 巻（1959～1963 年、風間書房）
(3) 日本近代思想大系　全 24 巻（1988～1992 年、岩波書店）
(4) 新日本古典文学大系 明治編　全 30 巻　（2001 年～刊行中、既刊 29 巻、岩波書店）

これらの叢書は、言語研究を目的として編纂されたものではないが、文化・思想・文学を中心に多様な分野の重要資料が選ばれていると考えられ、そこには、言語資料としても価値の高いものも含まれていると思われる。

表１　叢書類に収録される資料の数（時代別）

	江戸期	明治前期	明治中期	明治後期	計
明治文化全集	16	265	196	16	493
明治文化資料叢書	2	20	50	39	111
日本近代思想大系	70	959	504	7	1540
新古典大系明治編	1	26	99	14	140
計	89	1270	849	76	2284

表１は、四つの叢書に収録された資料の数を発行された時代別にまとめたものである（発行年代が大正期以後のものや不明のものは集計から除いてある）。明治前期・明治中期に集中しており、国語辞典編集準備資料の目録で不十分だった部分を補うことができよう。

この四つの叢書以外にも、資料リスト増補の材料として有用な叢書や図書目録は色々と考えられるが、まずは、上記の三つの目録と四つの叢書とから作成した資料リストの中身を分析することで、近代語史をとらえるための資料選定をどのように行っていくのがよいかを考えていきたい。

４．資料リストの分類と資料選定の考え方―明治前期・中期を例に―

４．１　文体の観点

４．１．１　文体の流れ

ここでは、明治前期・明治中期を例に取り上げたい。上記の、国語辞典編集準備資料と叢書類から作成した資料リストのうち、明治前期・明治中期の部分には、2000 点余りがお

さめられている。これについて、文体・ジャンル・媒体の三つの観点から分析を加えていこう。はじめに文体の観点から見る。

言文一致による口語体書き言葉の成立は、近代語史における最重要の出来事のひとつだが、その文体の流れを、森岡（1991）が示す図式をもとにまとめると、表２の通りである。明治初期には、文語体も口語体も多様な文体があったが、次第に統合されていき、明治 40 年代には言文一致体という口語体ひとつに統合されていく流れがあった。統合以前に多様に分かれていた文体は、研究者によって様々な分類や名付けがなされており、森岡説はそのひとつである。各文体は連続し交錯し、相互の識別が難しい場合も多い。要点は、近代の文体史は多様性から均質性へという明確な方向性をもっており、まずは文語体・口語体それぞれの内部で統合され、やがて口語体が全体に及んでいき、明治時代のうちにそれが完結するというところにある。文語体の内部、口語体の内部での文体の識別は、その指標が立てにくいが、文語体か口語体かの別については、文末辞を指標として明確に識別することが可能である[6]。

表２　近代語の文体統合の流れ（森岡 1991 に基づき作成）

<table>
<tr><td colspan="2"></td><td>明治初期</td><td>明治 10 年代</td><td>明治 20 年代</td><td>明治 30 年代</td><td>明治 40 年代</td></tr>
<tr><td rowspan="6">実用文系統</td><td rowspan="3">文語体</td><td>漢文訓読体</td><td rowspan="3">和漢折衷体</td><td colspan="2" rowspan="3">明治普通文</td><td rowspan="7">言文一致体</td></tr>
<tr><td>和漢折衷体</td></tr>
<tr><td>候文</td></tr>
<tr><td rowspan="3">口語体</td><td>問答体</td><td rowspan="3">演説体</td><td rowspan="3">演説体</td><td rowspan="3">初期言文一致体</td></tr>
<tr><td>講述体</td></tr>
<tr><td>談話体</td></tr>
<tr><td rowspan="2">文学系統</td><td>口語体</td><td>俗文体</td><td>講釈体</td><td>初期口語体</td><td>初期言文一致体</td></tr>
<tr><td>文語体</td><td>和漢折衷体</td><td colspan="2">雅俗折衷体</td><td>（雅俗折衷体）</td></tr>
</table>

４．１．２　文語体と口語体

表３　明治前期・明治中期の文体

	明治前期	明治中期
文語体	1187（93.1%）	773（91.1%）
口語体	31（2.4%）	47（5.5%）
文語体・口語体	3（0.2%）	0（0%）
その他	55（4.3%）	29（3.4%）
計	1276（100%）	849（100%）

表３は、明治前期・中期の 2000 点余りの資料について、文語体か口語体かを認定しその数と比率をまとめたものである[7]。文語体と口語体が混用されているものは、基調をなす文

[6]文末辞が「なり」「たり」「き」「けり」などで終わる文体は文語体、「だ」「である」「た」「です」「ます」などで終わる文体は口語体と識別できる。『太陽コーパス』の文体情報もこの基準で付与してある。

[7]明治前期には国語辞典編集準備資料と叢書類の両方を集計し、明治中期には叢書類のみを集計した。これは、国語辞典編集準備資料が示す資料のすべてを実際に見ることができなかったため、文体が未確認のものが残ったことによる。

体がどちらであるかによって区別した。「文語体・口語体」と記したのは、両者が同等であるもの、「その他」は漢文や英文あるいは文章でないもの（名簿など）である。明治前期では文語体がほとんどで、明治中期には口語体が数パーセント増加するものの、まだ大部分が文語体である。この時期、文語体が圧倒的に優勢であったことが確かめられる。

４．１．３　文語体

明治前期の文語体を、森岡（1991）は、漢文訓読体、和漢折衷体、候文の３種に分類するが、それぞれ、次のような文体のことを指す。上記の資料リストに含まれるものから１例ずつをあげてみよう。

○漢文訓読体

吾輩日常二三朋友ノ盍簪ニ於テ偶當時治亂盛衰ノ故政治得失ノ跡ナド凡テ世故ニ就テ談論爰ニ及ブ時ハ動モスレバカノ歐洲諸國ト比較スルコトノ多カル中ニ終ニハ彼ノ文明ヲ羨ミ我ガ不開化ヲ歎ジ果テ果テハ人民ノ愚如何トモスルナシト云フコトニ歸シテ亦欷歔長大息ニ堪ザル者アリ

（西周「洋字を以て国語を書するの論」、『明六雑誌』1、1874 年、明六雑誌原本による）

○和漢折衷体

輕重長短善惡是非等ノ字ハ相對シタル考ヨリ生ジタルモノナリ輕アラザレバ重アル可ラズ善アラザレバ惡アル可ラズ故ニ輕トハ重ヨリモ輕シ、善トハ惡ヨリモ善シト云フコトニテ此ト彼ト相對セザレバ輕重善惡ヲ論ズ可ラズ斯ノ如ク相對シテ重ト定リ善ト定リタルモノヲ議論ノ本位ト名ク諺ニ云ク腹ハ脊ニ替ヘ難シ又云ク小ノ虫ヲ殺シテ大ノ虫ヲ助クト

（福沢諭吉『文明論之概略』、1875 年、文明論之概略原本による）

○候文

浜田御預り所村々百姓共、衆訴落印と二つに相分り候に付、今度鶴田御役所より御役人様御上下拾六人、書添村へ御出張に相成、

（津山藩岡熊治郎による監察記録、1868 年、日本近代思想大系による）

候文は文末などに「候ふ」を伴うもので、文体類型として確立し、この類型に属する文章を特定していくことができるが、漢文訓読体と和漢折衷体との識別は難しい。漢文訓読体に和文や俗文の要素が交じった福沢諭吉の文章などが和漢折衷体の典型とされるが、個々の文章を漢文訓読体と和漢折衷体とに判別する明確な指標は立てることはできない。

４．１．４　口語体

森岡（1991）は、明治前期の口語体には、実用文系統に３種、文学系統に１種あったと見ているが、それぞれ、次のようなものを指すと思われる。やはり、上記の資料リストに含まれるものから例をあげよう。

○問答体の例

開化文明　サア／＼英吉君。是こそ僕が舊宅だ。

西海英吉　ホヽウ成程、茅葺の門長屋、廣庭の植ごみ、こなし部屋から牛部屋の景況、なんとなく古色を帶て、歴然たる舊家の豪農殿が兵衛が宅に來たやうだね。ソシテアノ異な歌を大勢が唱つて居るあれは何ンだね。

（横河秋濤『開化の入り口』、1873‐1874 年、明治文化全集による）

○講述体の例

世の諺にも｜不治是天福［しらぬがほとけ］と申す通りで、成程世の事國の事も自身に識らざる時は、更に心に掛［かゝ］らずして一向心配することはありますまい。だが、右の如く人間が箇［か］様［やう］に世間の物事を識らずして濟むものでありませう歟［か］。

（植木枝盛『民権自由論』、1879 年、明治文化全集による）

○談話体の例

なぐさみながら、よみあげまする。お経の文句はなにがなんだと、たずねてみたれば、作州五郡の庄屋がねんらい、あんまりおうきな盗みをしおった。そのしりだん／＼百姓がほりかけ、あちらもこちらも村々さわだち、中々ちよっこりちよっとにゃおさまりませんが、そのわけあらまし申してみふなら、ぬすんだそのかずおふひが中にも、とりわけ大きな事からあげます。

（本多応之助「鶴田騒動の阿呆陀羅経」、1868 年、日本近代思想大系による）

○俗文体の例

モシあなたヱ牛［ぎう］は至［し］極［ごく］高［かう］味［み］でござすネ此［この］肉［にく］がひらけちやアぼたんや紅葉［もみぢ］はくへやせんこんな清［せい］潔［けつ］なものをなぜいままで喰［く］はなかつたのでござウせう

（仮名垣魯文『安愚楽鍋』、1871 年、明治文学全集による）

明治前期の口語体資料は約 30 点あるが、それらが上の４種の文体のいずれであるか分類するのが難しい場合も多い。これらの種別は明確な類型としてではなく、口語体の多様な広がりの範囲を考える目安として考えるのが適切であろう。

４．１．５　資料選定における文体の扱い

以上見てきたように、明治前期に多様であった文体について、明確な類型と指標を立てて、個々の文章を分類していくことは困難である。一方、文語体と口語体の識別は文末辞を指標として明確に判別していくことが可能である。したがって、資料選定においては、文語体か口語体かの別については、これを選定の際の判断材料に用いることができるが、それぞれの中の細分類は、材料として採用しにくいと考えられる。むしろコーパスを作成した後に文体の詳細な研究が行われるべきだろう。

なお、明治前期・中期は、口語体の比率はきわめて低いが、それを理由として、当期のコーパスにおける口語体資料の構成比率をうんと低くするのは適切でないと考えられる。なぜなら、後代にすべての文体を統合していく口語体がどのように変容し発展したか、また普及し定着していったかを歴史的に把握するためには、まだ少数派だった初期段階のそれを積極的に採り、その変化の過程を研究できるようにしていくべきであるからである。このようなところは、言語史研究のためのコーパス設計における資料選定では、サンプリングによる代表性の尊重よりも、個々の資料の独自性の尊重が優先される部分だと言えるだろう。

４．２　ジャンルの観点

ジャンルの枠組みは、『現代日本語書き言葉均衡コーパス』の書籍や、『太陽コーパス』では、図書館における書籍の分類基準である NDC（日本十進分類法）が用いられている[8]。上述の資料リストに収録される資料についても図書館に収録されている書籍の場合は、

8『現代日本語書き言葉均衡コーパス』では国会図書館の書誌データに付されている NDC 番号を利用したが、『太陽コーパス』ではコーパス作成者が記事を読んで番号を付与した。

NDC 番号が取得できる場合がある。そこで、国立国会図書館の「近代デジタルライブラリー」を検索し、そこに収録されているものに NDC 番号を引き当て、明治前期・中期のジャンル分布を図４に表した。

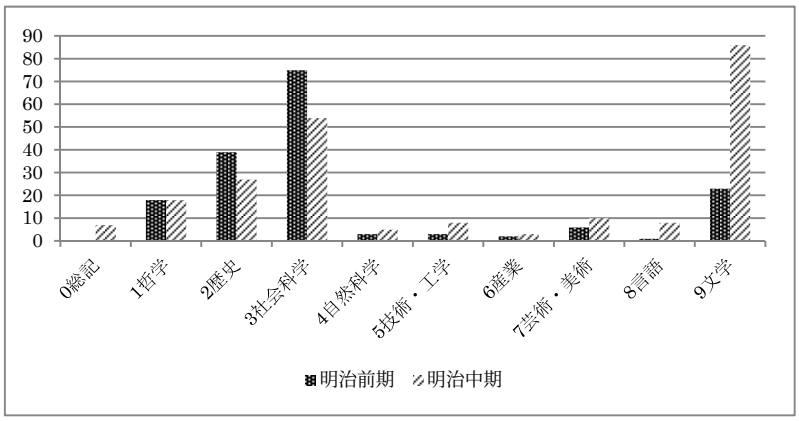

図４　明治前期・中期の資料のジャンル

明治前期は、社会科学が最も多く歴史がこれに次ぎ、さらに文学、哲学の順に多い。ところが、明治中期では文学が最も多くなっており、社会科学がこれに次ぎ、そして歴史、哲学という順となり、時代的な変容が大きい。これも、時代によるジャンルの多寡の違いが反映している面と、データ作成の典拠とした目録や叢書の性質を反映している面とがあろう。このような大きな変容があるところでは、単純に実際の構成比率にしたがってサンプルの比率を決めるだけでは適切でないように思われる。むしろまずは、資料リストの中身を見ながら、当期の当該ジャンルの資料として重要性の高いものであれば採ることを検討し、そうでなければ別に典拠とすべき叢書や目録がないか検討していくような研究段階が必要ではないだろうか。例えば、当期の自然科学や技術・工学の資料はきわめて少ないが、表４のような資料が含まれている。これらの資料を実際に見て、コーパス化の適否を考えていくことが望まれよう。このような点も代表性だけでなく個々の資料の性質への目配りが必要になるところである。

表４　明治前期の「4 自然科学」「5 技術・工学」の資料（部分）

資料	著者	NDC	文体	西暦	叢書	叢書巻
訓蒙 窮理図解	福沢諭吉	420	文語	1868	日本近代思想大系	科学と技術
物理了案	宇多健齋	420	文語	1880	明治文化全集	科学編
舎密局開講之説	三崎嘯輔	430	文語	1870	明治文化全集	科学編
天変地異	小幡篤次郎	440	文語	1868	明治文化全集	科学編
西洋時計便覧	柳河春三	535	文語	1870	明治文化全集	風俗編
男女普通家政小学	小林義則	590	文語	1880	日本近代思想大系	風俗 性
女房の心得	望月誠	590	文語	1878	日本近代思想大系	風俗 性
服製年中請負仕様書	鈴木篤右衛門	593	文語	1868	明治文化全集	風俗編
西洋料理通	仮名垣魯文	596	文語	1872	明治文化全集	風俗編
通俗男女自衛論	三宅虎太	598	文語	1878	日本近代思想大系	風俗 性

４．３　媒体の観点

資料リストを見ていくと、先に「ジャンル」として設定した NDC とは別の枠組みで分類した方がよいのではないかと思われるものが目につく。例えば、表５に示したものは、明治 8（1875）年に発行された新聞・雑誌の一群の一部である。

表５　明治 8（1875）年の新聞・雑誌（部分）

資料	著者	NDC	文体	西暦	叢書	叢書巻	出典
評論新聞	海老原穆	―	口語・文語	1875	明治文化全集	雑誌編	
仮名読新聞	―	―	口語	1875	日本近代思想大系	言論とメディア	
萬国叢話	―	―	文語	1875	明治文化全集	雑誌編	
国民気風論	西周	150	文語	1875	日本近代思想大系	天皇と華族	明六雑誌
華士族論	島地黙雷		文語	1875	日本近代思想大系	天皇と華族	共存雑誌
善良なる母を造る説	中村正直	370	文語	1875	日本近代思想大系	教育の体系	明六雑誌
真影の禁を論ず	高木登		文語	1875	日本近代思想大系	天皇と華族	朝野新聞

明治前期に次々に創刊される新聞や雑誌それ自体が叢書におさめられている場合（上の三つ）と、叢書に採られた資料の出典が新聞・雑誌である場合（下四つ）とがある。飛田（1973）は、新聞・雑誌は、近代に存在する多様な言語資料の性格をすべて合わせもっている「総合資料」という扱いをしており、雑誌『太陽』がそれ単体で代表性を持つと考えて『太陽コーパス』を設計したのも、そのような考え方に立ってのことであった。コーパス作成にあたっては、新聞・雑誌は、その総合性が生きるように、多様な資料をまとめて採集できる資料として扱うのが適切だろう。具体的には、総合性の高い新聞や雑誌をいくつか定め、その新聞や雑誌につついては、例えば、『太陽コーパス』で採ったような、等間隔の期間を置く方法などによってサンプリングを行うこが考えられる。どの雑誌・新聞を選ぶかは、資料の独自性を重視するものだが、その内部をサンプリングするのは、代表性を意識する選定方法と言うことができるだろう。

新聞・雑誌以外で目を引くのは、法令、文書、手紙・日記の類である。法令は、『現代日本語書き言葉均衡コーパス』の「特定目的サブコーパス」に「法律」として採られた枠組みに対応する。文書は、公的な文書については、同じく白書や広報誌と通じるところがあろう。手紙・日記のうち私的な性質を持っているものは、同じく Yahoo!知恵袋や Yahoo!ブログと共通する性格があろう。これらは、近代の重要資料として一群をなしているだけではなく、『現代日本語書き言葉均衡コーパス』への接続という点でも重要性の高いものである。こうした NDC によるジャンルとは別に立てることが必要だと思われる分類枠は、広い意味で「媒体」と呼ぶことができるだろう。

なお、上記の資料リストには少数しか入っていないが、近代語研究の重要資料には他に、教科書、演説や落語などの速記、日本語について記述した文典・辞書などが存在する。教科書は、『現代日本語書き言葉均衡コーパス』における教科書と対応する。速記は、同じく国会会議録や『日本語話し言葉コーパス』に対応づけられるものとしても重要であり、明治後期以後には演説や落語の録音資料も存在しており、近代語コーパスに話し言葉資料をどのように取り込むかという課題につながっていこう。また、文典・辞書などは、コーパスの直接の対象にはしにくい面もあるが、コーパスから記述できる近代語の文法や語彙の実態と対照すべき資料として重要性は高く、コーパス設計時において、その関連づけの方法を検討しておくことも有意義なことだろう。これら現段階の資料リストでは手薄な重要資料を補っていく作業も必要である。

４．４　その他の観点

上に記した、文体、ジャンル、媒体のほか、ある資料をコーパスに入れるかどうかを検討する際に考慮すべき点が、ほかにも想定される。まず、原本の参照可能性の高さという点である。文献資料に基づく日本語史研究においては、コーパスができれば原本を見なくてもよいということにはおそらくならず、コーパスのもとになった本文が原資料でどのような姿であったかを参照したいという要求が研究者には強く存在すると考えられる。そうした要求に応えられるように、コーパス作成と同時に原本の影印や画像などを作成し関連付けることも考えられるが、現実にはそこに開発コストをかけることは難しい面がある。そこで、複製本が出版されていたり、国立国会図書館などの電子図書館で画像が公開されていたりするものをコーパス化することが考えられる。同じような理由で、本文についての研究成果が反映した校訂本、注釈書、索引などが整備されている資料も、コーパス化する価値が高いであろう。

最後に指摘するのは、コーパスとして用いられる場合でなくとも、文献資料による言語研究一般において、価値が高いとされる資料は、コーパスの対象としても価値が高いという点である。例えば、振り仮名がついているものは語形が確定できる優位性があり、著者の自筆本に基づいているものは別人による改変の心配がないという優位性がある。

以上のような、コーパス化する資料そのものの優位性にかかわる情報も、資料リストに書き入れておき、選定の際の判断材料に使えるようにしておけるとよいだろう。

５．資料選定の実施に向けて

５．１　資料選定の基本的手順

以上述べてきたことを踏まえて、近代語コーパスを設計する際に、今後どのようにして資料を選定していけばよいかについて、現段階で想定される基本的な手順の見通しを記しておきたい。

(1) 時代、媒体、ジャンル、資料の四層を立て、この枠組みで分類しながら資料のリストを増補していく。利用する叢書や目録は、現在手薄となっている媒体やジャンルを中心に、範囲を広げていく。
(2) 第Ⅰ層には時代を立てる。時代区分は 5 年を一単位とし、明治・大正期は三つの単位をまとめた 15 年ごとの明治前期・明治中期・明治後期・大正期というまとまりを設定する。昭和戦前期は 20 年でひとまとまりとし、昭和戦後期も当面分割しない。
(3) 第Ⅱ層に媒体を立て、書籍（初出が雑誌・新聞等のものも含む）、新聞・雑誌、教科書、法令、文書、手紙・日記などに分類する。なお、文学作品とベストセラーの目録から収集した資料はまとめて「書籍」に入れる。
(4) 第Ⅲ層にジャンルを立て、書籍は NDC の第１階層を枠組みとし、NDC では細かすぎる場合は、部分的に統合する。書籍以外は各媒体の性質に応じて枠組みを検討するが、第Ⅲ層が不要な（直下の層が資料である）媒体もある。
(5) 第Ⅳ層は個々の資料とするが、資料リストには、各資料について、発行年、媒体、ジャンル、資料名のほか、著者名、文体、出典、複製本、注釈書、索引、所蔵図書館、表記法、底本の状態等、選定作業において有用と思われる情報をできるだけ書き加え、選定作業の判断材料とする。
(6) 四つの層による分類を見わたしながら、各資料の特質を吟味し、各層各枠の中で資料に優先順位を付けていく。
(7) 近代語コーパスの開発期間、開発予算、開発手順などが具体化してきたら、資料リストを活用して資料選定案を作成する。

上に記した作業手順は、一言で言えば、近代語資料全体のバランスと個々の資料の性質との両面を考慮した選定方法で、はじめに述べた「代表性」と「独自性」の両面を考慮し

たものである。このような作業仮説を立てて候補になる資料を実際に見ながら分類し、採否の基準やバランスの取り方を工夫していくことが重要だろう。近代語研究の最大の障壁は資料が多すぎることだと言われることもあるが（湯浅 2000）、資料論を重ねながらコーパスを設計することで、その障壁を乗り越えていく道筋も見えてくるのではないだろうか。そのような検討や工夫を議論する場を、多くの近代語研究者が参加できる形で設けていくことも大切だろう。

５．２　資料選定の実施例―明治前期を例に―

現段階では資料リストは作成途上であり、層による粗密があったり、資料の実物を見ていないために、リストに記入すべき情報が不足していたり、ジャンルや文体などの分類が不十分であったりするものも多い。資料リスト整備はさらに継続していく必要がある。ここでは、実際に資料選定を実施する場合に論点になりそうなことを、現段階の資料リストで、第Ⅰ層（時代）が明治前期（明治元～15 年）になっている、約 1300 件の資料をもとに、少し考えてみたい。

明治前期の資料の第Ⅱ層（媒体）の内訳は、書籍と新聞・雑誌がそれぞれ約 350 件、文書が 500 件弱、法令が 100 件弱で、ここまでがまとまった量があるものである。一方、手紙・日記、教科書、辞書・文典、速記、韻文等は、いずれも 10 件に満たない。これらの媒体については、書籍や文書等に分類されているものの中に、見方によってはこれらのいずれかに分類できるものがあったり、そもそも資料に関する情報収集が不十分なところがあったりするため、明治前期にあまり存在しなかった媒体だと言い切ることはできず、さらに精査していくことが求められる。文書がきわめて多くなっているのは、明治前期という社会体制が大きく変わる時期の資料を、文書から豊富に集めた叢書類の編集方針によるものである。文書における第Ⅲ層（ジャンル）をどのような枠組で分類していくかは課題であるが、例えば、叢書が立てる「宗教」「憲政」「風俗」「教育」といった内容から分類することや、典拠となっている「日本外交文書」「大久保利通文書」などのような編纂文書の種類ごとにまとめることなどが、想定できよう。

書籍の第Ⅲ層（ジャンル）は、NDC を用いるのが便利である。国会図書館等に所蔵があるなどして NDC 番号を引き当てることができた資料が 240 件ほどあり、0 番台「総記」から 9 番台「文学」までのすべてのジャンルにわたっている。そのうち、「文学」に分類されるものは、表６の 21 件である。表の中での資料の配列は刊行年順である。

第Ⅰ層の時代は、明治前期①（明治元～5 年）・明治前期②（明治 6～10 年）・明治前期③（明治 11～15 年）の三期に細分した。第Ⅱ層（媒体）、第Ⅲ層（ジャンル）はこれ以上の細分の必要はなさそうである。第Ⅳ層でどの資料を選ぶかの観点として、表６に示した「時代」「文体」「様式」「振り仮名」「日国用例数」「本の存在など」の情報を、この順で考慮したい。具体的には、時代は特定の期に偏らないようにすること、文体は口語を優先するが文語も採るようにすること、様式は多様になるようにすること、振り仮名は総ルビ・部分ルビの順に望ましいが無ルビでも排除しないようにすること、日国用例数（『日本国語大辞典第 2 版』でその資料から採られている用例の数）は多い方がよいこと、本の存在は国会図書館（近代デジタルライブラリー）や国語研に所蔵があるものが望ましいことなどを考慮するのがよいだろう。そうした考慮の結果、コーパスの対象として優先されると考えられるものから順位を付け、一番左側の列に記入した。具体的には、『安愚楽鍋』『通俗伊蘇普物語』『人間万事金世中』『西国立志編巻之貳　其粉色陶器交易』『怪化百物語』『西洋道中膝栗毛』『欧州奇事花柳春話』『近世紀聞』『鳥追阿松海上新話』『魯国奇聞烈女之疑獄』の順になり、他はさほど優先順位は高くないと考えられた。明治前期の書籍の文学では、これらがコーパス化に適切な資料ではないかと考えられ、この後は、他のジャンルや媒体などとのバランスから、さらに絞り込んでいくことになるだろう。

このような選定作業を、文学以外のジャンルや、書籍以外の媒体に対してもできるだけ行っていき、さらには、明治中期以後の時代にも行っていくことが考えられる。多くのジ

ャンル、媒体、時代について検討が進めば、それら全体を見わたした上でのバランスを取る作業も行うことができるようになるだろう。そのようにして、独自性と代表性の双方に目配りした選定作業を行っていくことが望まれよう。

表６　明治前期・書籍・文学の資料選定例

順位	Ⅰ時代	Ⅱ媒体	Ⅲジャンル	Ⅳ資料	著者	様式	刊行年	文体	振り仮名	日国用例数	本の存在など
6	明治前期①	書籍	913	西洋道中膝栗毛	仮名垣魯文	戯作	1870	口語	部分ルビ	1785 件	国会、複製あり
1	明治前期①	書籍	913	安愚楽鍋	仮名垣魯文	戯作	1871	口語	部分ルビ	1060 件	国語研蔵、索引あり
8	明治前期②	書籍	913	近世紀聞	染崎延房	記録	1875	文語	総ルビ	1954 件	国会、新古典大系
5	明治前期②	書籍	914	怪化百物語	高畠藍泉	戯作	1875	口語	総ルビ	54 件	国会、新古典大系
4	明治前期②	書籍	912	西国立志編巻之貳 其粉色陶器交易	佐藤富三郎	劇	1873	口語	部分ルビ	0 件	国会
	明治前期②	書籍	913	西国立志編巻之十 鞋補童教学	佐藤富三郎	劇	1873	口語	部分ルビ	0 件	国会
2	明治前期②	書籍	930	通俗　伊蘇普物語	渡部温	小説	1873	口語	総ルビ	46 件	国会
	明治前期②	書籍	930	開巻驚奇　爆夜物語	永峰秀樹	小説	1875	文語	部分ルビ	0 件	国会、翻訳
3	明治前期③	書籍	913	人間万事金世中	河竹黙阿弥	劇	1879	口語	総ルビ	49 件	国会、新古典大系
	明治前期③	書籍	912	日本美談	前田正名	劇	1880	口語	総ルビ	0 件	国会
	明治前期③	書籍	913	島衙月白浪	河竹黙阿弥	劇	1881	口語	総ルビ	70 件	国会、新古典大系
9	明治前期③	書籍	913	鳥追阿松海上新話	久保田彦作	小説	1878	文語	総ルビ	93 件	国会
	明治前期③	書籍	930	新説　八十日間世界 一周	川島忠之助	小説	1878	文語	無ルビ	3 件	国会、翻訳、新古典大系
7	明治前期③	書籍	930	欧州奇事花柳春話	丹羽純一郎	小説	1878	文語	無ルビ	2543 件	国会、翻訳
	明治前期③	書籍	914	高橋阿伝夜刃譚	仮名垣魯文	小説	1879	文語	総ルビ	27 件	国会、新古典大系
	明治前期③	書籍	930	哲烈禍福譚	宮島春松	小説	1879	文語	総ルビ	22 件	国会
	明治前期③	書籍	930	九十七時二十分間月 世界旅行	井上勤	小説	1880	文語	無ルビ	0 件	国会、翻訳
	明治前期③	書籍	913	民権演義　情海波瀾	戸田欽堂	小説	1880	文語	部分ルビ	0 件	国会、新古典大系
	明治前期③	書籍	930	春風情話	坪内逍遥	小説	1880	文語	総ルビ	0 件	翻訳、新古典大系
	明治前期③	書籍	930	欧州情譜　群芳綺話	大久保勘三郎	小説	1882	文語	部分ルビ	0 件	翻訳
10	明治前期③	書籍	930	魯国奇聞烈女之疑獄	柚田策太郎	小説	1882	文語	無ルビ	0 件	国会、翻訳

上に述べた書籍については、個々の資料の独自性を十分に考慮した選定を行うことがよいと思われたが、媒体やジャンルによっては、このような詳しい検討を行うことが現実的でないものもある。例えば、新聞・雑誌のような定期刊行物は、毎号の多様な記事のジャンルや文体などを逐一分析してから選定することは繁雑である。かといって、全号全文をコーパス化することもコストの面から難しい。この場合は、対象に定めた新聞や雑誌は、

4.3 に述べたように、総合資料としての性格を生かして、ランダムサンプリングを行って、コーパス化する号や記事を定めていくことが適切であろう。どの新聞・雑誌を対象とすべきかという点においては、各資料の性質を書籍の場合と同じように分類整理した上で、選定すべきである。本プロジェクトで作成するモデルコーパスには『明六雑誌』を選定したが[9]、そのような検討によって、明治前期の雑誌として、コーパス化の優先順位が最も高いと考えたことによる。ほかに、明治前期の新聞・雑誌では、『読売新聞』『東京日日新聞』『東洋学芸雑誌』『六合雑誌』などが、優先的にコーパス化されるべきだと考えられる。『明六雑誌』は、刊行された期間が短く、分量も多くないことから、モデルコーパスでは全文をコーパス化したが、刊行期間が長く分量の多い他の新聞・雑誌については、サンプリングを行ってコーパス化する号や記事を限ることが考えられる。

６．おわりに

以上本稿では、今後構築する「近代語コーパス」では、資料全体のバランスと個々の性質の双方をよく検討して、資料選定を行っていくことが重要であることを確認し、詳しい資料リストを作成して、このリストを分析しながら資料選定を実施する事例を示した。そのような資料選定を進めていくことは、「近代語コーパス」が、「通時コーパス」と『現代日本語書き言葉均衡コーパス』とをつなぐ役割を果たすためにも、欠かせないことだと考えられる。

本稿で用いた資料リストは現在作成途上であり、資料の増補や情報の整備を継続させていく必要がある。また、コーパス構築の事業規模に見通しがを立てた上での、実現性を重視した資料選定の段階に進むことも待たれよう。構築事業の立案のためには、短期・長期の両面での行程表の作成や、資料選定以外の設計にかかわる研究成果との統合などに着手することが求められよう。

文　献

小木曽智信(2005)「構造化テキストを直接利用するアプリケーション―『プリズム』と『たんぽぽ』―」（国立国語研究所 2005b 所収、pp.83-113）

木村睦子・加藤安彦・田中牧郎(1998)「国語辞典編集のための用例データベース」（『日本語科学』5、国書刊行会、pp.109-127）

国立国語研究所(1985-1997)『国定読本用語総覧』（三省堂）

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

国立国語研究所(2005b)『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―』（博文館新社）

近藤泰弘（2012）「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 VSARPJ 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』国立国語研究所、pp.1-10）

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」（国立国語研究所(2005b)、pp.1-48）

飛田良文(1973)「近代語研究の資料」（『文学・語学』66、三省堂、pp.45-60）

飛田良文(2012)『国立国語研究所「日本大語誌」構想の全記録』（港の人）

前川喜久雄(2008)「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」（『日本語の研究』4-1、pp.82-94）

丸山岳彦・柏野和佳子・田中牧郎（2011）「第 3 章 サンプリング」（『現代日本語書き言葉均衡コーパス 利用の手引　第1.0版』、国立国語研究所コーパス開発センター、pp.21-38）

[9]『明六雑誌コーパス』については、本報告書に収録した、近藤明日子・田中牧郎「『明六雑誌コーパス』の仕様」、及び近藤明日子「『明六雑誌コーパス』の概要」を参照。

森岡健二(1991)『近代語の成立 文体編』（明治書院）
湯浅茂雄(2000)「近代語研究の要点と課題」（『日本語学』19-11、明治書院、pp.138-148）

付 記

本論文は、「第１回コーパス日本語学ワークショップ」（2012 年 3 月 6 日、国立国語研究所）で発表した内容をもとに、加筆したものである。