近代語史をとらえるための文献選定とコーパス

田中　牧郎（国立国語研究所言語資源研究系）†

Material Selection and Corpus Compilation for Historical Study of the Modern Japanese

Makiro Tanaka (Dept. Corpus Studies, NINJAL)

１．はじめに

本稿は、平成 21 年度から行っている国立国語研究所の共同研究プロジェクト「近代語コーパス設計のための文献言語研究」で研究している内容のうち、近代語のコーパスの対象にする文献資料をどのように選ぶかという課題について考えるものである。

２．近代語のコーパスへ

２．１　近代語コーパスの位置

『現代日本語書き言葉均衡コーパス』が完成したが、日本語研究のためのコーパスはこれで十分というわけではもちろんなく、国立国語研究所においてはすでに、「超大規模コーパス」「通時コーパス」など新たなコーパスの設計が始まっている。

『現代日本語書き言葉均衡コーパス』（1 億語以上）は、明治時代から現代に至る近現代日本語の全体を把握するためのコーパス群（KOTONOHA）の最重要構成要素と位置付けられて開発が始まった（前川 2008）。KOTONOHA を構成する国立国語研究所のコーパスとしては、『日本語話し言葉コーパス』（約 750 万語、2004 年公開）、『太陽コーパス』（約 750 万語、2005 年公開）が先行している。現代共通語話者の独話を対象とした『日本語話し言葉コーパス』は、東京工業大学・情報通信研究機構と協力して構築されたもので、工学的応用の側面も色濃く反映した設計になっており、この性質は『現代日本語書き言葉均衡コーパス』にも継承され、さらに「超大規模コーパス」へと受け継がれていくことが見込まれる。明治後期から大正期の総合雑誌『太陽』を対象とした『太陽コーパス』は、国立国語研究所に従来あった近代語研究や史的国語辞典編集といった文献言語研究の系譜の中から生まれたものであり、その側面はやはり『現代日本語書き言葉均衡コーパス』に受け継がれており、さらに「通時コーパス」へと連なっていくものと考えられる。

「通時コーパス」の設計については、平成 21 年度から新規に始まった「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘客員教授）において扱われているが、そこでは奈良時代から江戸時代までが対象とされており、明治時代以後『現代日本語書き言葉均衡コーパス』までをつなぐコーパスとしても、近代語コーパスの重要性は高い。

２．２　『太陽コーパス』

近代語のコーパスとして公開済みの『太陽コーパス』（国立国語研究所 2005a）について概観しておこう[1]。『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

† mtanaka@ninjal.ac.jp

[1] 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（http://www2.ninjal.ac.jp/lrc/）。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。

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、読者層の厚さなどの点で、当時の文献資料としては格別の価値を持っていることに、根拠を置いている（田中 2005）。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、明治後期から大正期の書き言葉の広がりを母集団として把握した上で『太陽』の代表性を検証して設計したものではなく、一資料のみを対象としたコーパスにとどまる限界は否めない。

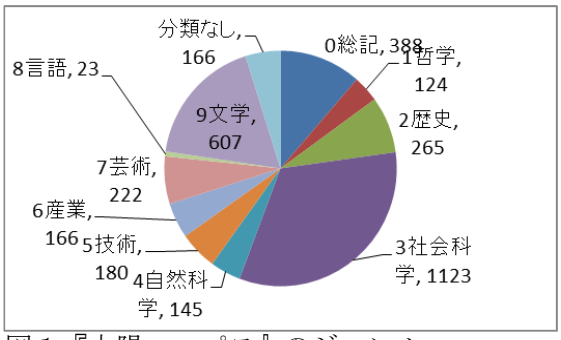

図１『太陽コーパス』のジャンル

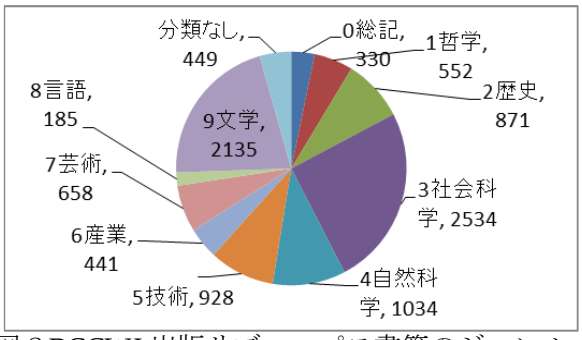

図２BCCWJ 出版サブコーパス書籍のジャンル

コーパスの要件としてやはり重要な、言語研究に役立つ付加情報については、総合雑誌という特性を生かせるように、記事や引用などの範囲をマークアップし、そこにジャンル・文体・著者（話者）などの情報を属性として埋め込んだ構造化テキストを実現し、校訂注記や異体字情報などもタグを付与して豊富に表現した（田中 2005）。あわせて、それらの情報を自在に利用できるように、『ひまわり』（山口 2005）、『プリズム』『たんぽぽ』（小木曽 2005）などの検索ツール群を開発した。このように、『太陽コーパス』は、文献資料を対象としたコーパスとしては画期的なものであったが、『日本語話し言葉コーパス』や『現代日本語書き言葉均衡コーパス』に付与されている形態論情報が全く付与されていないなど、不十分なところも残されていた。

『太陽コーパス』は公開後約７年が経過したが、ほぼ同時期に公開した『日本語話し言葉コーパス』に比較すると、これを利用した研究成果は、必ずしも多くない。これは、上述した代表性や付加情報の不十分さに起因するほか、明治後期から大正期の約 30 年間だけを切り取っていて近代語史の中で浮いた存在になっていることにもよっている。前後の時代をも対象に加えて『太陽コーパス』を相対化し、近代語史の全体がとらえられる近代語のコーパスを設計し構築を始めることが求められている。

３．近代語の文献リストの作成

３．１　「国語辞典編集準備資料」

先に『太陽コーパス』は国立国語研究所の史的国語辞典編集の系譜から生まれたと述べたが、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『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[2]、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(この間の経緯は、木村・加藤・田中 1999 参照)。『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文献を検討する際に、上記の目録類は第一に参考にすべきものである。

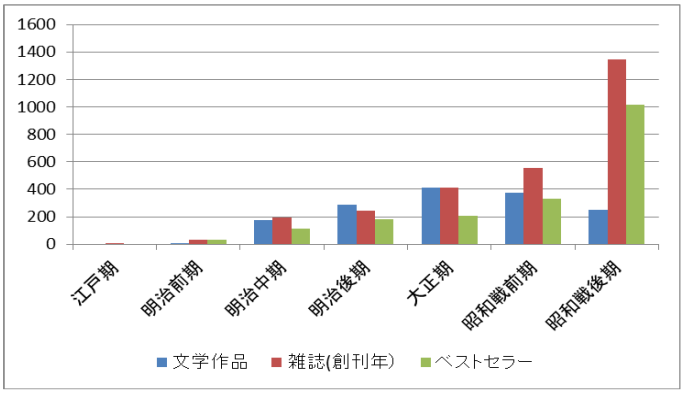

図３　国語辞典編集準備資料に掲載された文献数(時代別)

図３は、上記の三つの目録に掲載された文献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 1～15(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 1～14(1912-1925)年
昭和戦前期：昭和 1～20(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。

[2]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)。

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文献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、対象文献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文献の候補は、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文献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、明治前期の文献の手薄さが目立つため、この時期の文献を豊富におさめる叢書類をもとに文献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・文学の分野の重要文献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が含まれていると思われる。

表１　叢書類に収録される文献の数（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、４つの叢書に収録された文献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４つの叢書以外にも、文献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３つの目録と４つの叢書とから作成した文献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文献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。ここでは、明治前期・明治中期を例に取り上げたい。

４．文献リストの分類と文献選定の考え方―明治前期・中期を例に―

４．１　文体の観点

4.1.1　文体の流れ

上記の文献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３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[3]。

表２　近代語の文体統合の流れ（森岡 1991 に基づき作成）

<table>
<tr><th colspan="2"></th><th>明治初期</th><th>明治 10 年代</th><th>明治 20 年代</th><th>明治 30 年代</th><th>明治 40 年代</th></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="8">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 点余りの文献について、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[4]。文語体と口語体が混用されているものは、基調をなす文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

[3]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。

[4]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文献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。

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。上記の文献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「候う」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、上記の文献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うせう

（仮名垣魯文『安愚楽鍋』、1871 年、明治文学全集による）

明治前期の口語体文献は 31 点あるが、それらが上の 4 種の文体のいずれであるかに分類するのは難しい場合も多く、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　文献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、明確な類型を立てて指標にしたがっ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、文献選定においては、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。

また、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文献の構成比率をぐっ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新しい文体がどのように広がり定着していく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、その発展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。

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、NDC（日本十進分類法）が用いられている[5]。上述の文献リストに収録される文献についても図書館に収録されている書籍の場合は、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。

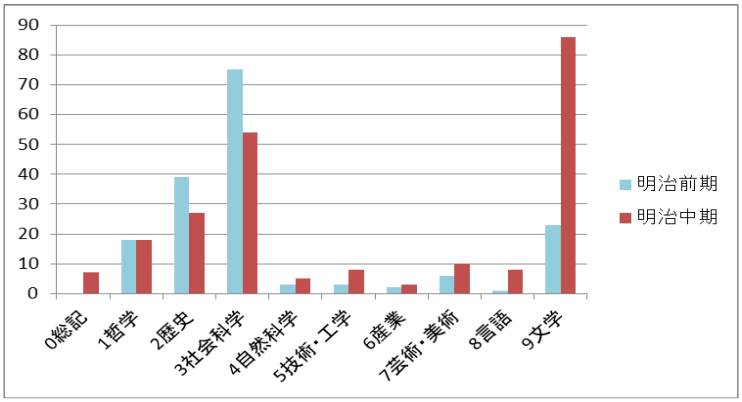

図４　明治前期・中期の文献のジャンル

[5]『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。

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、資料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ことは適切でないように思われる。むしろまずは、文献リストの中身を見ながら、当期の当該ジャンルの文献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あろう。例えば、当期の自然科学や技術・工学の文献はきわめて少ないが、表４のような文献が含まれている。これらの文献を実際に見て、コーパス化の適否を考えていくことが望まれよう。

表４　明治前期の「4 自然科学」「5 技術・工学」の文献（部分）

文献	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

文献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）

文献	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、叢書に採られた文献の出典が新聞・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、多様な文献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、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。『太陽コーパス』と同様の方法である。

新聞・雑誌以外で目を引くのは、教科書、法律、文書の類である。教科書や法律は、『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」に採られた枠組みである。文書は、公文書については、同じく白書や広報誌と通じるところがあろう。また、日記や手紙などの一群もあるが、これら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、近代の重要文献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、上記の文献リストでは目立たなかったが、近代語研究の重要資料には他に、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。速記は、『現代日本語書き言葉均衡コーパス』における国会会議録や『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いので、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。

４．４　その他の観点

上に記した、文体、ジャンル、媒体のほか、ある文献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提供す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や注釈書類などが整備されている文献も、コーパス化する価値が高いであろう。

次に指摘するのは、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文献は、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、コーパス化する文献そのものの優位性にかかわる情報も、文献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と思われる。

５．文献選定の実施に向けて

最後に、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文献を選定していけばよいかについて、現段階での見通しを記しておきたい。

(1) 発行年代、媒体、ジャンル、文献の四層を立て、この枠組みで分類しながら文献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、法律、文書、日記・手紙、速記（会議録を含む）、文典・辞書などを立て、近代語資料として重要なもの、『現代日本語書き言葉均衡コーパス』の媒体と対応付けられるものを生かした枠組みとする。なお、文学作品とベストセラーの目録から収集した文献はまとめて「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（直下の層が文献である）媒体もある。
(5) 第Ⅳ層は個々の文献とするが、文献リストには、各文献について、発行年、媒体、ジャンル、文献名のほか、著者名、文体、出典、複製本、注釈書、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加え、選定作業の判断材料とする。目録における優先候補、叢書における扱いなどもできるだけ書き入れる。
(6) 四つの層による分類を見わたしながら、バランスを考慮して文献選定の考え方を議論し、各層各枠の中で文献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、文献リストを活用して文献選定案を作成する。

ここに記したことの中には、プロジェクトで十分に議論を行っていない案も交じっているが、このような作業仮説を立てて候補になる文献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。

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）
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）
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）
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）
山口昌也(2005)「構造化テキストに対応した全文検索システム『ひまわり』」（国立国語研究所2005b所収、pp.49-82）