『大阪樟蔭女子大学日本語研究センター報告』第9号（103～110頁）

「青空文庫」のデータベース化と研究への利用

大阪樟蔭女子大学　田原 広史
大阪樟蔭女子大学非常勤講師　南場尚子

1.はじめに

本論では、良質で量的にも研究に耐えうるウエッブサイトである「青空文庫」のデータを利用し、語彙検索、用例検索用のデータベースを作成する手順とその意義について述べる。

これは、他人が作成したデータを、研究に利用できる形に加工するというものだが、以前、第 14 回の DB-West において「邦画の題名における字種の移り変わり -日本映画データベースを用いて-」（田原広史・渡邊陽子）と題して発表したものと同様の手順踏んでいる。「日本映画データベース」（http://www.jmdb.ne.jp）では、約 100 年にわたる邦画の題名 33,758 タイトルをデータベース化し、タイトルの長さ、字種の移り変わりを分析の対象とした。

ここでは、「青空文庫」について、このデータがどのようなものなのか、研究に利用するためには、どのような作業をおこなえばよいのか、具体的な研究に利用する場合の資料としての位置づけ、利点あるいは問題点などについて述べる。

2.「青空文庫」というサイト　http://www.aozora.gr.jp

「青空文庫」は、富田倫生氏が呼びかけ人となり運営されているサイトで、著者の死後 50 年後を経て著作権の切れた作品を対象に、フルテキストで掲載している。掲載ファイルの形式は、テキストファイル（ルビなし、ルビありの 2 種類）、HTML 形式、エキスパンドブック形式の 3 形式があり、いずれの形式もホームページ上から自由にダウンロードできる。1997 年 8 月に 5 作品でスタートしたが、2001 年 3 月時点で 1300 作品を超えており、現在も作品を追加ししつつある。

3.データベース化の手順

青空文庫からダウンロードしたフルテキスト文書のままでは、語彙検索がしにくいので、今回はすべての作品について行単位で区切り、行単位のデータベースとした。具体的なデータベース化の手順（1 ～ 10）を以下に述べる。

1.ダウンロード：ダウンロードはホームページ上から自分のパソコン内にデータを保存する作業である。ここでは青空文庫から任意の作品について、ルビ情報付きのテキストファイルのものをダウンロードした。

2.解凍：ホームページ上からダウンロードできるファイルは、サーバーの容量を節約するため、また、ダウンロードの時間短縮のために圧縮ファイルになっていることが多い。青空文庫上のテキスト形式ファイルも zip 形式で圧縮されている。lhasa（解凍用のソフト）がインストールされていればダブルクリックで解凍することができる。

3.ファイル名付け替え：青空文庫のテキスト形式のファイル名はローマ字で付けられている。そのままのファイル名では読みにくい上に、多くの作品を管理しにくいため、次のようにファイル名を変更した。例えば、「kakekomi-rubi.txt」（-rubi はルビ情報付きテキストファイルであることを表す）は、「0018 駆け込み訴え.txt」というようにした。数字は作家番号、作品番号等を自分で決めて付けておくとよいだろう。

4.MIFES を使って文単位に分ける：一般のワープロソフトでは、句点を手がかりに改行して別のレコード（行）にしていくことはできないので、ここではメガソフト社の MIFES を使い、次のような置換をおこなった。すなわち、句点を、句点＋改行に置換した。MIFES にあるグローバル置換機能と呼ばれる機能を使うと、多くのテキストファイルが入っているフォルダを指定するだけで、ファイルを一々開くことなしに中身をすべて置換することができる。

5.テキストファイルをエクセルに読み込む：一文一行（あるいは一段落）になったデータに、作品名・作家名・作品内の文番号を付加するための作業をおこなうため、いったんエクセルに読み込む。テキストファイルをエクセルに読み込むには、エクセルからテキストファイルを指定して開けばよい。ここでは、1 作品 1 シートに読み込んでいき、作家別にファイル化した。ただし、作品数が多い作家については、幾つかのファイルに分けた。エクセルでテキストファイルを開くと、開いた時点でテキストファイルのファイル名がシート名になるので、3 で付け替えたファイル名がそのまま生きてくる。

6.各シートの各行に必要データを入力：本文テキストの左側に 3 列ほど挿入し、「作品番号＋作品名」（シート名をコピーしてセルに貼り付ければ、打ち直さなくても良い）、「作家名」、「作品中の行番号」を、オートフィルを使って入力する。これで、任意の行を検索して取り出しても、誰の何という作品の何番目の文かが特定できる。

7.エクセルの各シートを新たなシート上で一つにまとめる：データベースの利点は大量のデータを一度に検索できることである。したがって、可能な限り多くの行をひとかたまりにしておく方が望ましい。ここでは、一作家の全作品を目安に一つのシートにまとめた。エクセルのシートには行数の制限があるので、これを超える場合は、とりあえず複数のシートに分けておく必要がある（エクセル 2000 の最大行数は 16 の 4 乗の 65,536 行である）。

8.ファイルメーカーにエクセルのシートを読み込む：いよいよデータベース化のため、ファイルメーカーにエクセルのデータを読み込んでいく。テキストファイル→エクセル同様、エクセル→ファイルメーカーについても、ファイルメーカーからエクセルのファイルを直接開くことができる（「開く」でエクセルのファイルを指定すると、どのシートを読み込むかを聞いてくる)。ファイルメーカー Pro5 から、エクセルの先頭行をフィールド名として読み込むことができるようになったので、読み込んだ後にフィールド名を付け直す手間もかからなくなった。ファイルメーカーでは行数の制限はないので、もし、エクセルで 2 つ以上のシートに分割せざる得ないものについても、何回かに分けて読み込んでいけば、一作家の全作品を一つのファイルにまとめることが容易にできる。今回のデータで言えば、芥川龍之介 180 作品が約 52,000 行、太宰治 112 作品が約 65,000 行となった。

9.ファイルメーカー用のレイアウトを作成する：読み込み時に自動的に作成されるレイアウトは単純なもので使いにくいので、「検索用レイアウト」及び「印字用レイアウト」の 2 種類程度を作っておく方がよい。なお、検索用については、カード型と一覧表型の両方ある方が便利である。なお、検索結果が大量の場合は、再びエクセルに移して印字する方が、効率よく印字できる。

10.データベースの完成：付加情報を入力するためのフィールドを幾つか作成しておき、検索をおこないながら逐次情報を蓄積していく。データベースの価値はこの付加情報である。したがって、データベースの構造については、使用者が自分の目的に応じて情報を柔軟に付加することができる仕組みになっていることが望ましい。付加すべき情報とは、機械的に検索できない情報、すなわち人の目による判断が必要とな情報のことである。ここで扱っている研究で言えば、特定の複合動詞を検索したときに、意味によってさらに絞り込むような場合が該当する。いったん、この情報を付加しておけば、次からこの情報を利用して機械的に、一気に検索することができる。また動詞の拍数や活用の種類等の情報を別途入力しておけば、それらの情報と組み合わせて検索することにより、新たな知見を生み出すことが可能となるだろう。こうして進化をとげた良質のデータベースは、一個人の研究にとって有用であるだけでなく、他の研究者あるいは後進の研究へと受け継がれていくことになる。

使用したソフトと、作業との関係を簡単にまとめておく。エディタソフト(ここではMIFES)は大量データの機械的加工用に、エクセルは作品ごとの整理をおこなう中間作業用、ファイルメーカーは完成したデータを蓄積し、検索をおこなったり、検索後にさまざまな情報を付け加えたり、結果を印刷したりするためのツールとして位置づけている。すなわち、以下の通りである。

フルテキストデータ	→	行単位データ＋基本情報の付加	→	データベース化
MIFES		エクセル		ファイルメーカー

4.データベース化する意義

フルテキストデータを文単位でデータベース化する意義について述べる。

大きな意義としては、フルテキストデータに比べると、語彙検索が手軽にできることである。フルテキストでは検索部分を抜き出したり、検索結果のみを寄せ集めて表示したりできない。したがって、語彙検索をするにあたっては、データを何らかの形で分割する必要がある。文節単位や単語単位で分けることも考えられるが、文節、単語の判定を一々おこなう必要があり、作業量が膨大になるので、ここでの目的（用例検索用）からすると、現実的ではない。データを細かく分けてあることは、一見便利そうではあるが、分ける段階で編者の意図や解釈が入り込んでくるということでもあり、却って汎用性がなくなることにつながる場合も多い。

また、多数の作品、場合によっては複数の作家を同時に検索対象とできるので、作品ごとに何度も同じ検索手順を繰り返す必要がなくなる。この点は作業の効率化、手順の単純化という観点からも大きな魅力である。単に手軽で楽になるというだけではない。何度も同じ手順を繰り返すということは、その際に間違いが生じる率が高くなるということである。他の研究者が検証する場合にも、手順は可能な限り単純な方がよい。

さらに、作成したデータベースを公開した場合、他の研究者がそのまま利用でき、データ資源の共有という点で優れている。データベースの共有は、単にデータの共有にとどまらないものである。それは、これまでに述べてきたように、データベースは知の集積でもあるからである。極論を言えば、作成者の到達点から研究を始めることができるのである。この点に関して、ある種の感情的な抵抗感があることは事実であり、この感情が、データベースの公開・共有の障害となっている面がある。これを回避し、データベース研究の発展を促すためには、データベース研究の研究的側面を正当に評価すること、および利用にあたっての一般的なルールを明確化することが不可欠であろう。

5.研究への利用について

ここでは、上記の方法で作成したデータベースを使っておこなった研究例について紹介する。筆者（南場）は、これまで、「―きる」「―ぬく」「―とおす」や「―きる」「―つくす」「― はたす」のように、複合動詞の後項にあって形式的な意味を前項に添える働きをする動詞の中で、類義関係にあるものについて考察してきた。その際、主に文学作品や新聞から用例を集めてきた。

文学作品に限っていえば、これまで語彙検索の方法としては、次の二つの方法が有効であった。

① 索引を利用する

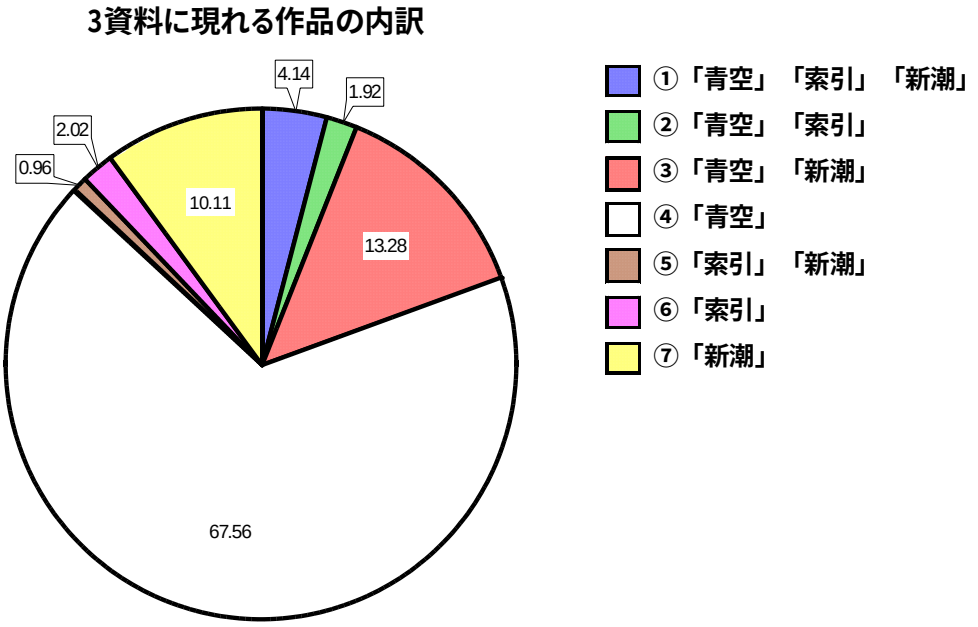
② CD-ROM を利用する

索引を利用する場合は、『作家別用語索引』（近代作家用語研究会　教育技術研究所編　教育社）で、夏目漱石・芥川龍之介・森鴎外・志賀直哉・太宰治の作品の一部について検索ができる。また、『国語国文学資料索引総覧』（国立国語研究所図書館編　笠間索引叢刊 109）では、作品ごとに索引の有無が調べられる。

CD-ROM を利用する場合は、『新潮文庫の 100 冊』『新潮文庫 明治の文豪』『新潮文庫 大正の文豪』がある。

以上の方法に加えて最近ではホームページを利用するという方法も出てきた。岡島昭浩氏の「日本文学等テキストファイル」は、古典文学から近代文学までの幅の広いテキストファイルで、一部「青空文庫」からもダウンロードされている。

さて「青空文庫」(図では「青空」)、『作家別用語索引』(図では「索引」)、上記の CD-ROM 3 種（図では「新潮」。組み合わせが煩雑になるので三資料をまとめた）に収められている作品の重複状況を調べてみると、以下のようなことがわかった。論文末に各パターンの作品例について掲載したので参照いただきたい。



	①	②	③	④	⑤	⑥	⑦	計
「青　空」	○	○	○	○	×	×	×	
「索　引」	○	○	×	×	○	○	×	
「新　潮」	○	×	○	×	○	×	○	
作品数	43	20	138	702	10	21	105	1039
全体に占める率	4.14	1.92	13.28	67.56	0.96	2.02	10.11	

○は掲載あり、×は掲載なし

「青空文庫」は、その作品数において他の資料を凌駕する。通時的な考察をする場合、「青空文庫」でこの時期の語彙が俯瞰でき、特定の語の消長の調査にも使えるのではないか。

また、この時期の特定の作家の語彙については、このサイトの検索でかなりカバーできるであろう。今、芥川龍之介、森鴎外、太宰治の作品数を比較してみると次の表のようになる。

	芥川龍之介	森鴎外	太宰治
青空文庫	180 作品	30 作品	112 作品
作家別用語索引	24	15	23
新潮文庫 100 冊	8	11	1
新潮文庫 明治の文豪	―	23	―
新潮文庫 大正の文豪	59 ※	―	―

※ほかに小品、随筆として 7 題

「青空文庫」には作品が随時追加掲載されているので、こまめにチェックして、一連の作業をし、データを追加していくことができる。これは、他の資料と大きく異なる点である。ただ、複合動詞の後項についての考察という点から言えば、たとえば、「―きる」を検索すると、「できる」「飽きる」等まで入ってくるのでいくつかの段階を経なければ必要なものが集められないという不便さがある。

第 17 回 DB-West において本内容を発表したときに、「青空文庫」のデータの信憑性はどうかというご質問を頂いた。「青空文庫」では、入力者とは別の校正者がいるが、それでも後日入力ミスが見つかった場合は、サイトの中でその情報を掲示している。利用する側としては、万全を期する意味でも、入力者が底本としなかった他の資料における同一作品との照合は必要であろう。この発表は作業が完全に終了したとは言えない段階のものであり、また、入力作品が増えていくというこのサイトの性格もあるため、今後も作業、考察等を進めていきたいと考えている。

付記

この報告は、2000 年 12 月 3 日に、大阪樟蔭女子大学でおこなわれた第 17 回 西日本国語国文学データベース研究会（大阪樟蔭女子大学）で口頭発表した、南場尚子・田原広史「「青空文庫」のデータベース化と研究への利用」（当日の発表資料）に加筆修正を加え、まとめ直したものである。当日の発表に対して、ご意見をいただいた方々にお礼申し上げる。1 ～ 4 については田原が、5 については南場がそれぞれ執筆をおこなった。

各パターンの作品の例（1）

パターン	「青空」	「索引」	「新潮」	著者	作品名
①	○	○	○	芥川龍之介	トロツコ
①	○	○	○	芥川龍之介	或阿呆の一生
①	○	○	○	芥川龍之介	芋粥
①	○	○	○	芥川龍之介	河童
①	○	○	○	芥川龍之介	戯作三昧（新字・新仮名）
①	○	○	○	芥川龍之介	玄鶴山房
①	○	○	○	夏目漱石	こころ
①	○	○	○	夏目漱石	虞美人草
①	○	○	○	森鷗外	かのように
①	○	○	○	森鷗外	阿部一族
②	○	○	×	芥川龍之介	海のほとり
②	○	○	×	太宰治	ヴィヨンの妻
②	○	○	×	太宰治	めくら草紙
②	○	○	×	太宰治	ロマネスク
②	○	○	×	太宰治	陰火
②	○	○	×	太宰治	右大臣実朝
②	○	○	×	太宰治	猿面冠者(新字・新仮名)
②	○	○	×	太宰治	玩具
②	○	○	×	太宰治	逆行(新字・新仮名)
②	○	○	×	森鷗外	安井夫人
③	○	×	○	芥川龍之介	アグニの神
③	○	×	○	芥川龍之介	あばばばば
③	○	×	○	芥川龍之介	おぎん
③	○	×	○	芥川龍之介	おしの
③	○	×	○	芥川龍之介	お富の貞操
③	○	×	○	芥川龍之介	きりしとほろ上人伝
③	○	×	○	芥川龍之介	さまよえる猶太人
③	○	×	○	夏目漱石	カーライル博物館
③	○	×	○	夏目漱石	ケーベル先生
③	○	×	○	梶井基次郎	Kの昇天
③	○	×	○	梶井基次郎	ある崖上の感情
③	○	×	○	梶井基次郎	ある心の風景
③	○	×	○	国木田独歩	たき火
③	○	×	○	森鷗外	うたかたの記
③	○	×	○	森鷗外	カズイスチカ
③	○	×	○	森鷗外	ぢいさんばあさん
③	○	×	○	徳田秋声	あらくれ
③	○	×	○	二葉亭四迷訳	あいびき
③	○	×	○	樋口一葉	うつせみ
③	○	×	○	樋口一葉	たけくらべ
④	○	×	×	芥川龍之介	「鏡花全集」目録開口
④	○	×	×	芥川龍之介	LOS CAPRICHOS
④	○	×	×	芥川龍之介	MENSURA ZOILI
④	○	×	×	芥川龍之介	あの頃の自分の事
④	○	×	×	芥川龍之介	お時儀
④	○	×	×	芥川龍之介	お律と子等と
④	○	×	×	芥川龍之介	カルメン
④	○	×	×	かぐつちみどり（夢野久作）	オシャベリ姫
④	○	×	×	太宰治	HUMAN LOST
④	○	×	×	太宰治	ア,秋
④	○	×	×	太宰治	あさましきもの
④	○	×	×	太宰治	おさん
④	○	×	×	太宰治	おしゃれ童子
④	○	×	×	太宰治	お伽草子
④	○	×	×	海若藍平（夢野久作）	お菓子の大舞踏会
④	○	×	×	海若藍平（夢野久作）	キキリツツリ
④	○	×	×	海若藍平（夢野久作）	クチマネ
④	○	×	×	菊池寛	M侯爵と写真師
④	○	×	×	九鬼周造	「いき」の構造
④	○	×	×	国木田独歩	あの時分
④	○	×	×	小中千昭	Alice6
④	○	×	×	小葉武史	sophia
④	○	×	×	清水哲男	RETURN

各パターンの作品の例 (2)

パターン	「青空」	「索引」	「新潮」	著者	作品名
④	○	×	×	清水哲男	きみとは往けない。
④	○	×	×	清水鱗造	1992～1993年現代詩時評コラム
④	○	×	×	素木しづ	かなしみの日より
④	○	×	×	束原和多志	いないないヴァーチャる
④	○	×	×	長尾高弘	イギリス観光旅行
④	○	×	×	津野潤	〈子供〉をめぐる虚言
④	○	×	×	田中英光	オリンポスの果実
④	○	×	×	藤下真潮	イブ― 覚醒儀式―
④	○	×	×	藤下真潮	インターミッション
④	○	×	×	藤下真潮	エア―黄泉戸喫―
④	○	×	×	藤本和子	オリエントの舌
④	○	×	×	南部修太郎	S中尉の話
④	○	×	×	北條民雄	いのちの初夜
④	○	×	×	牧野剛	21世紀への遺言
④	○	×	×	有島武郎	カインの末裔
⑤	×	○	○	夏目漱石	それから
⑤	×	○	○	志賀直哉	山科の記憶
⑤	×	○	○	志賀直哉	十一月三日午後の事
⑤	×	○	○	志賀直哉	小僧の神様
⑤	×	○	○	志賀直哉	城の崎にて
⑤	×	○	○	志賀直哉	真鶴
⑤	×	○	○	志賀直哉	赤西蠣太
⑤	×	○	○	森鴎外	雁
⑤	×	○	○	森鴎外	青年
⑤	×	○	○	森鴎外	普請中
⑥	×	○	×	太宰治	ダス・ゲマイネ
⑥	×	○	×	太宰治	猿ヶ島
⑥	×	○	×	太宰治	魚服記
⑥	×	○	×	太宰治	思い出
⑥	×	○	×	太宰治	斜陽
⑥	×	○	×	志賀直哉	或る朝
⑥	×	○	×	志賀直哉	暗夜行路
⑥	×	○	×	志賀直哉	灰色の月
⑥	×	○	×	志賀直哉	正義派
⑥	×	○	×	志賀直哉	清兵衛と瓢箪
⑦	×	×	○	夏目漱石	幻影の盾
⑦	×	×	○	葛西善蔵	葛西善蔵集
⑦	×	×	○	菊池寛	ある恋の話
⑦	×	×	○	菊池寛	海の勇者
⑦	×	×	○	菊池寛	義民甚兵衛
⑦	×	×	○	菊池寛	極楽
⑦	×	×	○	菊池寛	形
⑦	×	×	○	久米正雄	学生時代
⑦	×	×	○	幸田露伴	王義之
⑦	×	×	○	幸田露伴	画題とその詩仙
⑦	×	×	○	高浜虚子	虚子自選句集
⑦	×	×	○	国木田独歩	おとづれ
⑦	×	×	○	国木田独歩	まぼろし
⑦	×	×	○	国木田独歩	わかれ
⑦	×	×	○	国木田独歩	遺言
⑦	×	×	○	国木田独歩	岡本の手紙
⑦	×	×	○	国木田独歩	河霧
⑦	×	×	○	国木田独歩	郊外
⑦	×	×	○	志賀直哉	雨蛙
⑦	×	×	○	志賀直哉	好人物の夫婦
⑦	×	×	○	森鴎外	興津弥五右衛門の遺書
⑦	×	×	○	森鴎外	鶏
⑦	×	×	○	泉鏡花	歌行燈
⑦	×	×	○	田山花袋	蒲団
⑦	×	×	○	島崎藤村	ある女の生涯
⑦	×	×	○	島崎藤村	家