「青空文庫」を言語コーパスとして使おう

—メタデータ構築による歴史的・社会言語学的研究への応用の試み—

千葉庄寿[†]　夷石寿賀子[††]　陳君慧[††]

[†]麗澤大学外国語学部　　[††]麗澤大学大学院 言語教育研究科

schiba@reitaku-u.ac.jp, {siseki, cchen30}@cs.reitaku-u.ac.jp

1 はじめに

青空文庫[1]は 1997 年 7 月数名の有志でスタートしたオンラインの無料アーカイブである。青空文庫では以下のように述べられている。

> 青空文庫は，利用に対価を求めない，インターネット電子図書館です。著作権の消滅した作品と,「自由に読んでもらってかまわない」とされたもの[2]を，テキストと HTML 形式でそろえています。より多くの人に作品を味わってもらい，より自由にファイルを使ってもらうことは[3]，この場を整えている私たちの願いです。どうか青空文庫を，活用してください。(「青空文庫早わかり」[4])

現在の収録作品数は 5083 作品（2006 年 2 月 1 日現在), ほぼ毎日新たな作品が一作品以上公開され拡張し続けている。但し，野口（2005: 3-4）にもあるように作品数は本の単位「冊」ではなく一作品毎のカウントであり，一作品の長短はさまざまである。作品のジャンルは小説，論文，評論，随筆，エッセイなどの散文から詩，翻訳作品までと多岐に渡る。執筆年は青空文庫上で確認が可能な初出年の記載では 1890 年から始まり，死後 50 年で著作権の切れる 1956 年（2006 年現在）までに死去した作家 300 人あまりの作品が収録されている[5]。

本研究では作家・作品データベースの構築にあたり, 青空文庫の提供している作品一覧の CSV ファイルと図書カードを利用している。図書カードには作家のカード，作品毎のカードの二種類がある。作家のカードは作家の情報と作品のリスト，作品のカードは作品の情報と作家の情報が記載されている。また CSV ファイルには主な情報がひとつのファイルに収められている。

CSV ファイル及び図書カードの情報には，作家のジェンダーの区別，そして作品の小説・詩などといったジャンル分けの情報はない。また，例えば作品カードにおいて初出年や作家の生年の記載の有無というようなカード毎の情報にばらつきがみられる。

2 作家・作品データベースの構築

データベース構築のために，選出したパラメータは，具体的に以下のようなものである。作家のカードからは作家 ID，作家名（姓，名），作家名読み（かなとローマ字), 生年, 没年, 作家カードの URL を，作品のカードからは, 作品 ID, 作品名, 作品名読み，原題（翻訳作品のみ），初出情報，初出年，仮名遣い種別，著者名，翻訳者名，底本情報（詳細は省略），作品ファイルの URL，作品カードの URL，ファイルの登録日，更新日，ファイルサイズを選出した。それらについては, 作品一覧の CSV ファイルおよび図書カードから抽出が可能であるが，作家，作品毎に欠けている情報については，本研究で調査して補った。更に, 図書カードに掲載されている情報以外に，作家のジェンダー，DVD-ROM のファイルの位置をパラメータとして採用した。作品ファイルの URL は原則としてルビつきファイルを参照した。

そして，これらを基に作家・作品データベースを作成するにあたって，作家テーブル，作品テーブル

---

[1] http://www.aozora.gr.jp/

[2] 野口(2005:38)によると 2005 年 11 月現在では著作権存続中の「自由に読んでもらってかまわないとされた作品」は新たな作品としては追加されていない。

[3] ファイルの複製と再配布の際の取り扱いは著作権が切れたものと著作権存続中のものでは基準が違う。詳しくは,「青空文庫収録ファイルの取り扱い規準」(http://www.aozora.gr.jp/guide/kijyunn.html)を参照。

[4] http://www.aozora.gr.jp/guide/nyuumon.html

[5] 著作権がきれていない作家及び翻訳者などを含めると 400 人あまりになる。

という主要な二つのテーブルと，著者テーブル，翻訳者テーブル，ペンネームテーブルという関連付けを行うための 3 つのテーブルを設定した（**図 1**）。まず，作家テーブルは作家カードから抽出した情報に作家のジェンダーを加えたものであり，作品テーブルは，作品カードから抽出した情報に DVD-ROM のファイルの位置を加えたものから構成される。ただし，作品カードにある情報のうち，著者名及び翻訳者名については，著者テーブル，翻訳者テーブルの構成要素としてある。著者テーブルは，作品と作家との，翻訳者テーブルは作品と翻訳者との関連付けをおこなうものである。

なお，青空文庫における作家 ID は，同じ作家であっても，名前が異なる場合には異なる ID が振り当てられている。例えば，森林太郎で書かれた作品を森鴎外でも検索可能にするため，更にペンネームテーブル（作家関連付けテーブル）を作成し，その関連付けを行っている。

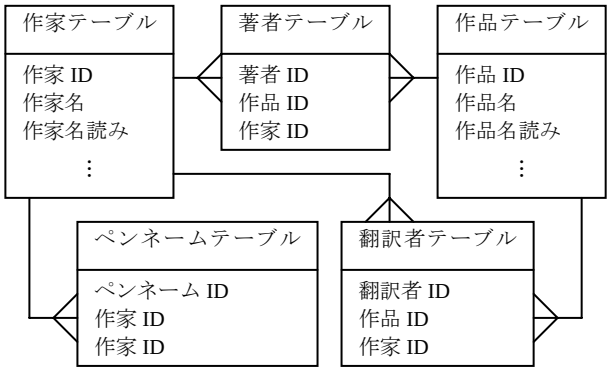

図 1：作家・作品データベースの構造

3 青空文庫の歴史的・社会言語学的利用の要件

§ 1 でみたように，青空文庫のデータは量的にも多様性の点からも様々なタイプの近・現代日本語研究の一次資料として用いることのできる潜在性をもっている。一方で，青空文庫は電子化された作品データが単純に蓄積されたものであり，そのままでは量的な分析に耐える言語コーパスとして利用することは難しい。例えば『太陽コーパス』は対象としての多様性を雑誌というメディアに求める（国立国語研究所 2005: iv）。本節では，青空文庫を用いておこなう調査を実際に想定しながら，青空文庫を言語コーパスとして利用する際の要件を考察する。

歴史（言語学）的研究の具体的な例としては，戦前戦後の文体差を比較，またはある語の通時的変化をみること等が想定される。そのためには，初出年の一覧表示，年代のラベル付け（明治・大正・昭和（戦前）・昭和（戦後）といった 4 分類や数年毎など），そして時代順のソートが出来なければならない。また文献学上，初出年と入力に使用した版と異なるか否かの確認のためには底本情報も必要である。

次に，社会言語学的な研究では，例えば作家名に「夏目漱石」を指定することにより特定の作家における文体研究ができることは言を待たない。また，ジェンダーで「女」を指定することにより，女性に特徴的な言語表現に対する調査が可能となる。そして初出年から生年を引くことにより算出される著作年齢の指定により，例えば「50 歳代」等特定の年齢層の言語表現を調査することも可能である。

更に「非翻訳」「翻訳」を指定して得られたデータに基づき，翻訳文体に特徴的な言語表現を分析することもできる。なお，ジャンルについては青空文庫から関連する情報が提供されていないので，ジャンルによる絞込みを可能にするためには，新たに情報を追加しなければならない。

また，仮名遣い種別については青空文庫から情報の提供はあるが，旧字，旧漢字から現代表記への変換が認められているため[6]，表記に関する文献学的な研究を行う場合には注意が必要である。

収録されている作品はサイズの点でもさまざまである。青空文庫のデータに言語コーパスとして量的な考察を加える場合，データから作品による偏りをできるだけ排除することが望ましい。具体的には，データ中に作品・作家が占める比率を量的に確認したり，一作品から一定のデータ量をサンプルとして検索データのサブセットを取得できることが考えられる。歴史的・社会言語学的なパラメータによる作品絞込みの結果に基づき対象となるテキストを簡易サンプリングする機能をもたせることで，「バランスド・コーパス」(Atkins *et. al.* 1992)を意識したコーパスサンプルの抽出効果が得られるだろう。

[6] 旧字，旧漢字から現代表記への変換については，「旧字、旧仮名で書かれた作品を、現代表記にあらためる際の作業指針」(http://www.aozora.gr.jp/KOSAKU/genn daihyouki.html)，「本という財産とどう向き合うか」の §3.1「旧漢字、旧かなづかいの書き換え」(http://www.aozora.gr.jp/KOSAKU/MESSAGE.html#ANK10)を参照。

更には，歴史（言語学）的・社会言語学的コーパスとバランスド・コーパスとしての特性を複合的に利用し，例えばある作家における文体の通時的移行，同時代の 30 歳代におけるジェンダー差，女性の文体の時代間における差異等の調査に必要なデータを青空文庫から絞込むことも可能になる。

４ 作品抽出機能つき CGI 検索システムの試作

以上，構築したデータベースを利用する検索システムの具体例として作品抽出機能を備えた CGI 検索システムを試作した。この CGI 検索システムは Lincoln D. Stein 氏の作による CGI モジュールを利用し，Perl5.8 で動作する。データベースサーバ（MySQL）への接続は DBI モジュールを利用している。

このシステムでは，作成したデータベースに収録されたテキストの検索が出来るほか，作品のテキスト検索には現在出版されている野口（2005）付属 DVD-ROM の青空文庫 4843 作品テキストファイルからの検索も可能である。またテキスト検索のキーワードには正規表現の利用が可能である[7]。

なお，テキストは原則としてルビありテキストを参照することとしているが，ルビを除いたテキストを検索対象とすることも可能である。

また書誌情報検索画面から抽出された書誌情報一覧から，テキストの検索，テキストの一覧情報のダウンロード，テキストのさらなる絞込み，テキスト抽出（簡易サンプリング）ができる。書誌情報は，さまざまなパラメータを使ってソートすることができ，必要なリソースを直接選択して検索することも可能である。

５ XML による書誌情報のメタデータ化

「メタデータ」は，インターネット上で公開されているリソースを効率よく検索するために「機械可読可能なかたちで記された，ウェブリソースなどに関する情報」(Berners-Lee 1997)[8]として考案され，あらかじめ生成した情報を使って検索やフィルタリングを的確に行うことが期待されている。さらに，リソースを記述する語彙（オントロジー）を定義することによりコンピュータの自動類推を可能にするセマンティック・ウェブの実現にむけ，言語資料のアーカイブのためのオントロジーの規格化がすすめられている（Wittenburg *et. al*. 2002)。

本研究では，作家・作品データベースから絞り込んだ作品データをメタデータとして XML 形式で抽出するシステムのプロトタイプを構築し，CGI 検索システムに実装した。青空文庫を言語コーパスとして利用する場合，このような書誌情報のメタデータを検索結果と切り離して生成，配信するメリットは少なくとも 2 つ考えられる：

- 検索対象となる抽出作品を書誌情報としてまとめて保存できること
- 検索対象となる作品の場所を含めた書誌情報を汎用的なファイル形式で取得することで，異なるツールを使ってデータを処理できること

常に更新・拡張されるという青空文庫の性質を考えると，検索結果の検証可能性は言語コーパスとしての青空文庫の信頼性を大きく左右する。その意味で，検索対象となったデータの書誌情報をそのまま保存できる前者の利点は大きい。（今後，書誌情報としてのメタデータに加え，検索条件自体を XML 形式で記録できるようにすることで，言語コーパスとしての利便性をさらに高めることができよう。）

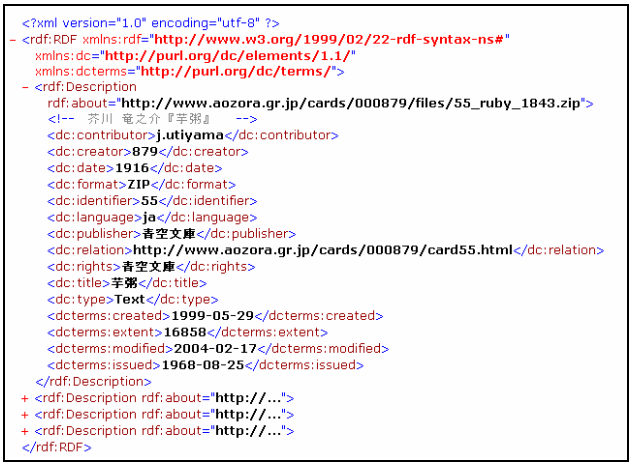

図２：ダブリン・コアに基づく書誌の出力例

また後者について，メタデータ出力形式のプロトタイプとして，本研究の現在のシステムでは(1) RDF によるダブリン・コアの記述（Apps 2005; Powell 2003,

[7] MySQL (安定版の最新バージョンは 5.0)は正規表現を使ってフィールドのデータを検索することが可能であるが，現時点では日本語を含むデータの検索には利用できない。

[8] 訳は神崎(2005:8)による。

図２）[9]，(2) TEI Lite (TEI U5)のヘッダ部分（teiHeader 要素）を取り出したリスト[10]に加え，さらに(3)青空文庫の図書カードで使われている書誌情報のラベルを独自のスキーマで出力する形式，の 3 種類を実現させている。

６ おわりに

本研究では，オープンな電子図書館としての青空文庫を言語資源として利用するための作家・作品データベースの整備を通じ，歴史（言語学）的・社会言語学的研究のために青空文庫を利用する方向性を検討し，CGI 検索システムを試作した。

また，書誌情報を XML 形式のメタデータとして提供することにより，検索結果の記録や検証が容易になること，また，様々な XML 対応のツールで作品データを処理できることを示した。本データベースを土台として，学術利用にも活用できる本格的な「電子図書館」としての役割を青空文庫が担うためには，今後図書館情報学的な観点からのメタデータの精密化や，より広い利用場面を想定したパラメータの追加や細分化が必要になると思われる。

青空文庫が提供する作品ファイルは自由な利用が可能である。一方，本データベースは青空文庫が提供する図書カードと作品一覧の CSV ファイルの情報を含んでいるが，それらの書誌情報には当然ながら編纂者の著作権が存在する。本データベース，検索システムを含め，メタデータという二次的な著作物を一般公開するのに必要な著作権処理の手続きについては今後慎重に議論する必要がある。また，本データベースを一般利用に供する場合には，青空文庫を直接オンラインで検索することでサーバ側にかかる負担についても考慮する必要があろう。

青空文庫のデータは頻繁に更新・追加され，そのデータ量及びそのカバーする時代区分はさらに充実していくと思われる。本研究が構築したデータベースを今後最大限に活用するためには，人手による修正を最小限に抑え，データベースを自動更新できることが望ましい。図書カードの規格の整理，およびその一貫した管理といった面で青空文庫と提携していくことも考えられるだろう。本研究が今後模索する道は，動的な性質をもつオンラインデータを学術的に信頼できる資料として利用するためのモデルケースになっていくと思われる。

[9] ダブリン・コアの精密化要素 Elements Refinements (名前空間 http://purl.org/dc/terms) を用いて date 要素を精密化し，底本の出版年(issued)，作品データの初登録日(created)と最終更新日(modified)，またファイルのバイト数(extent)を記述している。

[10] TEI ヘッダ要素を部分木とする整形式 XML として出力する。

謝辞

本研究は麗澤大学言語研究センター言語情報学プロジェクト(2004-)「言語研究のための多言語データベースの構築」[11]の研究成果の一部である。

参考文献

Apps, Ann (2005). Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata. Dublin Core Metadata Initiative Recommendation. [URL: http://dublincore.org/documents/dc-citation-guidelines/]

Atkins, Sue, Jeremy Clear & Nicholas Ostler (1992). Corpus design criteria. Literary and Linguistic Computing. 7: 1—16.

Berners-Lee, Tim (1997). Metadata architecture. World Wide Web Consortium Design Issues. [URL: http://www.w3.org/DesignIssues/Metadata.html]

Burnard, Lou & C. Michael Sperberg-McQueen (1999-2002). TEI Lite: An Introduction to Text Encoding for Interchange. TEI Consortium. [URL: http://www.tei-c.org/Lite/]

Powell, Andy & Pete Johnston (2003). Guidelines for implementing Dublin Core in XML. Dublin Core Metadata Initiative Recommendation. [URL: http://dublincore.org/documents/dc-xml-guidelines/]

Wittenburg Peter & Daan Broeder (2002). Metadata Overview and the Semantic Web. Paper read at the IMDI 2002 Workshop, 14. and 15. November, 2002, Nijmegen, the Netherlands. [downloadable from: http://www.mpi.nl/IMDI/documents/documents.html#IMDI_2002]

神崎正英 (2005). セマンティック・ウェブのための RDF/OWL 入門. 森北出版.

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.

野口英司 (2005). インターネット図書館青空文庫. はる書房.

[11] http://www.fl.reitaku-u.ac.jp/LINC/projects/langTech/