

日本語研究のための「青空文庫」データベース構築と活用

—電子テキスト処理プログラムとその活用モデル開発—

金囁泳(高麗大学講師)

yuiyu@korea.ac.kr · <http://www.japanese.or.kr>

1. はじめに

日本語研究において、大量のテキストをより早く、より正確に、より簡単に分析するために、電子テキストの需要は大きくなる一歩であって、それに応じてテキスト資料の電子化も年々大きく進んでいる。しかし、テキストの電子化には多くの人力、費用、時間などが必要とされる決して容易ではない作業である。よって大量のテキストの分析が必要とされる場合、すでに構築されている限られたコーパスに頼って研究を進めるか、比較的小規模であっても独自の電子テキストを収集して研究を進めるしかない。このような現状を考慮すると、すでに構築されているいくつかの電子テキストの中でも、大量のテキストを有しているながら、さらに個々のテキストに関する付加情報も含めて全てが無料で提供されているインターネット電子テキストアーカイブス・「青空文庫」は、日本語研究において大変魅力的なテキスト資料であると考えられる。そこで本発表では、「青空文庫」のテキストデータを日本語史の研究資料としてより有効に利用できる手法として、データベース化と管理・検索ツールの開発という具体的なモデルを提示したい。

2. 問題提起

電子テキスト或いはコーパスを用いた日本語研究の利点といえば、大量のテキストを対象とし、短時間でより簡単に正確な分析を可能にしてくれる点だということは今更言うまでもない。さらに、そのような個々の研究において取捨選択されて収集された用例や統計などの電子テキスト資料は、様々な形(e.g. プレーンテキスト、コーパスなど、詳しくは真島知秀・金囁泳(2003)参照)で公開・交換することが容易であって、再利用性も高い。また、最近の日本語研究には電子テキストを利用した言語統計を提示して論証する研究が少なくないが、そのような研究の対象や結果物などが公開されれば、研究の検証及び再現もより簡単、かつ正確に行うことが可能になる。前述したように電子テキストの構築には様々な面で多くの労力が必要とされるが、上記のような利点を考慮すると、これからは日本語研究のためにより積極的にテキストの電子化に取り組んで、未だに電子化されていない多くの日本語テキストにおける電子化を推進するべきであると考えられる。

そこで発表者は、未だに電子化されていない新しいテキストにおける電子化を進めることは言うまでもないが、すでに電子化されている良質のテキストをまず整えて有効に活用する必要もあると考える。例えば、今回の発表で取り上げる「青空文庫」には、すでに近世末から近現代に至るまで616人の著作、10,984件・約9286万語の電子テキスト(2012年4月19日現在、発表者の調査による)が収録されていて、現在にもその量は増加している¹⁾。そのような「青空文庫」は、様々な時代の著者による数多くの著作が収録されている電子テキストの宝庫であって、日本語研究にもっと幅広く効果的に活用されるべきであると考えられる。

一方、今まで構築されたコーパスなど多くの電子テキストは、底本の著作権に関する問題で、せっかくの時間と労力をかけて構築したにも関わらず公開に制約が生じたり、公開されても一部のみを公開せざるを得ない場合が少なくないのが現状である。そのように著作権の問題で研究に使用されたコーパスなど電子テキストに接近することに制約が生じると、研究結果に対する検証・再現が難しくなる上、その電子テキストは再利用の可能性が薄れ、一回性に止まってしまう恐れがある。よって発表者は、電子テキストを構築してそ

1) 「青空文庫」に収録されている電子テキストは、2009年9月に4,843件(野口栄司：2005)で、2007年7月に6,300件(青空文庫：2007)で、2010年5月に9,624件(青空文庫：2010)であって、年々増加している。

それを研究対象とする研究者が出来る限りその電子テキストの著作権に関する問題を先決するように努力する必要があると考える。ところが「青空文庫」の場合、著作権の期限が過ぎた、或いは使用許可を得た著作のみが収録されているため、そのような著作権の問題から自由である。つまり、「青空文庫」はその収録されている電子テキストの言語量という面だけではなく、再利用性・開放性という面でも大変魅力的な電子テキスト資料であることが分かる。

ところで、そのような「青空文庫」は今まで、読む対象としての電子テキストの機能だけに焦点があてられて多数のビューアなどは開発されてきたものの、先行研究ではその一部の電子テキストが使われることに止まっている場合が多い。しかし、「青空文庫」の膨大な電子テキストを大きい一つのデータベース、つまり日本語研究資料として利用しようとする観点からの接近や工夫は、管見の限り田原広史・南場尚子(2001)の試しを除くとあまり見られない。そこで発表者は、「青空文庫」に収録されている電子テキストを一つのデータベースとして一括し、構築・管理・検索が可能な公開コーパスにすることには、様々な分野の日本語研究により効果的に活用される可能性があると考えたのである。今回の発表では、小木曾智信・近藤明日子(2007)のような電子テキスト処理手法に関して論じた先行研究を踏まえて、「青空文庫」に収録されている電子テキストを体系的なデータベースとして構築し、それを効果的に管理しながら、実際の研究に応用できる検索ツールまでの一貫したシステムを開発するという、具体的な「公開コーパスデータベース開発モデル」を提示したい。

3. 「青空文庫」

「青空文庫」は1970年以来、著作権の問題が解決されたテキストを電子化して、そのデータを作家・タイトル及び分野によって分類してウェブ上に公開しているインターネット上の図書館である(富田倫生(1999:176))。また、野口栄司(2005)が述べたように、「青空文庫」は純粹にテキストの記録や資料を収集して、それを体系的に保管・管理・公開している点から厳密にいうと、アーカイブス(Archives)に近いといえよう。このような「青空文庫」からダウンロードできる電子テキストファイルには主に以下のような三つの形式がある。

- (1) a. テキストファイル(ルビあり・ルビなし)：タグなどが無い純粹なテキスト(以下、プレーンテキスト)で、圧縮された状態でダウンロードできる。テキストを確認するにはまず、ダウンロードしたファイルを解凍する作業が必要とされる。しかし、プレーンテキストとはいえ、テキストの中には、「テキスト中に現れる記号について」、「底本」、「初出」、「入力者」、「校正者」、「作成日」、「その他、案内・注意事項」などの付加情報が含まれている。e.g.) 1509_ruby.zip
 - b. HTMLファイル(或いはXHTML)：HTMLというウェブページ形式のファイルで、HTMLの規約に沿ったタグがつけられている。括弧で囲まれているルビの他、(1)aのような付加情報も含まれている。同じく本文テキストと共に、付加情報も含まれている。e.g.) 1509.html
 - c. エキスパンブックファイル：エキスパンドブックブラウザ(Expanded Book Browser)という電子テキストを読むためのプログラム用の専用ファイルである(日本語版および英語版のダウンロード：<http://www.voyager.co.jp/software/ebdl.html#download>)。e.g) 1509.ebk
- 凡例1) 例の作品は島崎藤村の「家」で、ファイル名の番号は作品番号である。 2) 符号化方式(encoding)は、ShiftJIS。

4. 「青空文庫」のデータベース化

4.1 「青空文庫」における文字列検索と電子テキストの形式

まず、従来の日本語研究における「青空文庫」を使った文字列検索の実例は、以下のようである。

- (2) a. 「秀丸」「KWIC Finder」など、正規表現(正規式)による「grep」「KWIC」検索が可

能なソフトウェアを用いた文字列検索

- ① 「青空文庫」のプレーンテキストをダウンロードし、一個ずつ解凍して一つのフォルダに入れる。
- ② 検索したい語・句・文を、フォルダ内のファイルに正規表現を使って検索。

b. 全文検索システム「ひまわり」を使って文字列検索

- ① 「青空文庫」のテキストファイルをダウンロード。但し「そらまめ」を使えば、「青空文庫」における全HTML(XHTML)テキストを一括ダウンロードすることが可能。
- ② ひまわり用データ作成ツール「えだまめ」を使って、HTMLテキストを「ひまわり」用のXML形式ファイルに変換。
- ③ 変換したデータを「ひまわり」に取り込み、検索。

そこで、プレーンテキストと(Plain Text)は、その言葉通りに純粋なテキストだけであって、品詞情報など言語情報は付されていないため、文字列検索などの最低限の用例検索しかできない。もちろん(2)bのような「ひまわり」ではHTMLファイルが使われているが、HTMLファイルにおけるタグは主に編集に関わる様式情報であって、言語情報においてはプレーンテキストとほとんど変わらない。しかし、プレーンテキストとはいえ、「青空文庫」のプレーンテキストには、言語研究の対象になる本文以外に、入力・校正者や注意事項などの付加情報がテキスト内に埋め込まれている。このような情報は、テキストを研究資料として利用するに際して決して不要な情報ではないため、本文とともに確実にまたすぐに確認できる形で保存しておくべきである。例えば「ひまわり」のように、本文のみならず、作家(Author)、生年(Birthday)、出版社(Publisher)、分野(Genre)などの情報は、用例検索において重要な情報になる。

また、「青空文庫」は大規模なテキストデータをもつため、それだけでも十分意味があるが、より多くの言語研究に有用な資料として活用するためには、そのテキストを形態素分析したり、XMLなどの形式で言語情報タグを付したりする加工を加える必要があると考えられる。またその同時に、効果的にそのテキストを検索・管理するツールも要求される。小木曾智信他(2007: 147-148)が指摘したように、日本語研究においてXMLのようなタグ付き電子テキストを利用するに際して、XML文書からタグを取り去って単なるプレーンテキストのみを利用したり、コーパス付属の検索ツールが出力するテキストデータと表形式に利用したりすることも十分研究に役立つ。しかし、そのような場合は元のデータの誤った利用に繋がる恐れがあるなど、いくつかの問題²⁾が存在するからである。従って発表者は、日本語研究ではもっと究極的にタグが付与された、例えばXMLタグテキストのような電子テキストを活用することを目指すべきであると考え。よって今回の「「青空文庫」データベース構築と活用モデル」でも、単にプレーンテキストの収集を通じてデータベースを構築することではなく、日本語研究に活用できる様々な情報をプレーンテキストと共に体系的に整理してデータベース化したものである。

しかし、実際の研究における電子テキストの需要は、プレーンテキストから、HTML、XHTML、XML、またコーパスのDBファイル等々、研究者のニーズやテキスト処理技術によって様々な形式のテキストが必要とされる。よって、本発表ではデータベースを構築するに際して自動化を前提にした上、段階・課程別に細分化して多様なテキスト形式を得られるようにすることによって、様々な研究に対応させることを目標とした。その課程を簡単にまとめると、まず「青空文庫」に収録されているXHTML形式の電子テキストをダウンロードし、プレーンテキスト、またその形態素分析テキストなどをそれぞれの書誌情報などの付加情報を加えてデータベース化した。また、その検索システムとして、プレーンテキストと形態素分析テキストをネット上に載せ、検索出来る形態に公開した。そうすることによって、電子テキストの収集、加工、検索と公開といっ

2) a. タグを取り去ったテキストだけを利用する場合、データの誤った利用に繋がる可能性がある。

b. 元のデータが持つ情報のごく一部しか利用できない。

c. 調査結果が元のデータに反映されないため、その場限りの使い捨てになり、別の調査結果との複合的な分析ができない。

小木曾智信・近藤明日子(2007: 147-148)

た一回りのデータベース化モデルを開発した。その詳しい手法は次のようである。

4.2 「青空文庫」の電子テキストとデータベース化

そこでまず、先行研究における「青空文庫」のデータベース化に関して簡単に触れてから論を進めたい。田原広史・南場尚子(2001:1-4)はテキストデータを精製し、テキストと共に有効な情報をそれぞれの「MS-Excel」のデータフィールドに入れて、検索できるようにする手法で「青空文庫」のデータベース化を試みた。しかし、その手順が複雑で、殆どが手作業であって効率が悪く、自然にデータベースの量が少ない。また、データベースに接近するに際してライセンスが要る専用のソフトウェアのみを使っている点などが指摘できると思う。ところが、テキストデータを付加情報とともに一緒に検索できるようにしようとした点と、使用者が自由にデータベースの原文を修正することが可能で、必要な情報フィールドを追加できる点などは、評価すべきである。これは、方法は少し異なるが、小木曾智信・近藤明日子(2007)で述べられている《研究者自身がXML文書にタグとして埋め込み、XML関連記述を積極的に利用してタグの情報を引き出す》という概念と一脈通ずるものである。

前述したように、「青空文庫」は三つの形式のテキストを提供しているが((1)参照)、日本語研究のためであれば一般的にHTMLタグ付きテキストと専用ビューア用のタグ付きテキストよりは、プレーンテキストの方が多く使われているのが現状である。そこで、文字列検索や形態素分析を行う時にはまず上記の(2)aのようなタグを分離する作業が必要とされる。もちろん、(2)bのように少し準備工程が必要とされるが「ひまわり」のようなツールを使用して一度「青空文庫」におけるテキストを変換して取り込んでおくと、そのテキストにおける検索速度が速いだけでなく、その精度も高い。また、何より直観的なインタペースで分かりやすいので、文字列検索だけであれば「青空文庫」における言語調査に便利だと考えられる。しかし、検索対象になるテキストデータは「ひまわり」の独自のファイル形式に変換されて、元のファイルに接近しがたい。その故、「ひまわり」によってXML形式に変換されたテキストは、そのプレーンテキストに接近しがたくて結果的に形態素分析やタグ付コーパスなどに再利用するに容易ではない閉鎖的な面が存在する。また、新しいデータが追加するに際して、一部のデータのみを追加することも容易ではない。

そこで、本発表では「青空文庫」におけるテキストデータを、まず純粋な本文テキストとテキストに関する情報(書誌情報など)を分けた上、体系的に整理してデータベース化した。また、このようなプレーンテキストを形態素分析して、同じデータベースに加えることにした。まず、テキストデータを精製し、純粋な本文テキストを抽出して体系的に整理するということには、「読む」ことにフォーカスが当てられている「青空文庫」のテキストを、言語研究に適するテキストデータに変換するという意味がある。また、単にプレーンテキストを取り出すだけでなく、テキストにおける付加情報を整理して同じく同一データベースに入力することによって、「ひまわり」の検索結果のように、日本語研究における必要な情報もテキストとともに一目瞭然に確認することができる。それに加えて、プレーンテキストを形態素分析したテキストを同一線上でデータベース化することによって、単純に文字列検索で終わるのではなく、より高度な言語研究に応用できる基本資料になると考えられる。

今回の発表では、以上のような点を踏まえて「青空文庫」をデータベース化するが、その際にして以下の点に重点をおいた。

- (3) a. テキストデータと付加情報と共にをデータベース化し、情報フィールドを拡張できるようにすること。
- b. 「青空文庫」のテキストデータをより簡単に、かつ正確に取り出す手段(ツールの開発)を講じること。
- c. 管理者にとって、データベースを簡単に維持・補修・更新できる手段を講じること。
- d. 利用者にとって、データベースを簡単に検索し、原本或いは付加データやその更新情報などを簡単に確認できるように公開すること。

- e. 利用者が独自の研究情報を加えるような手段を講じること。

4.4 「AJ-Aozora-Tool」による「青空文庫」のデータベース化

(3)における点を考慮した今回のデータベース化の手順は以下のようになる。

- (4) a. 「青空文庫」テキストデータをダウンロード：「そらまめ」を内蔵した「AJ-Aozora-Tool ver1.001³⁾」を使用して全テキスト一括ダウンロード ⇒ 結果物：XHTMLタグテキスト。e.g. 236_19996.html 『ア, 秋』太宰治(1975)

b. テキスト変換・処理

- 1) XHTMLをプレーンテキストへ変換：(5)aで特定のフォルダにダウンロードしたXHTMLタグテキストファイルを「AJ-Aozora-Tool」を使用してプレーンテキストへ一括変換 ⇒ 結果物：プレーンテキスト(仮名付き・仮名なし・書誌情報のみを除いたXHTMLテキストの中で選択可能)。

e.g. 236_19996.txt

- 2) プレーンテキストを形態素分析：(5)b-1で得られたプレーンテキストを「AJ-Aozora-Tool」に内蔵した「Mecab⁴⁾」を使って一括分析 ⇒ 結果物：形態素分析テキストファイル。e.g. 236_19996_mtxt

c. データベース入力

- 1) プレーンテキストのデータベース化：(5)aで特定のフォルダにダウンロードしたXHTMLタグテキストファイルのプレーンテキストと付加情報を「AJ-Aozora-Tool」を使用して「MS-Access」或いは「MS-SQL」へデータベースへ一括入力 ⇒ 結果物：「Access」データベースファイル(「AJ-Aozora-Tool.accdb」ファイルの中、テーブル“Aozora”・“Aozora_P”)或いは「MS-SQL」サーバーのデータベースの中、テーブル“Aozora”・“Aozora_P”。

- 2) 形態素分析結果のデータベース化：(5)b-1で特定のフォルダに分析された形態素分析結果テキストファイルを「AJ-Aozora-Tool」を使用して「Access」或いは「MS-SQL」にデータベースへ一括入力 ⇒ 結果物：「Access」データベースファイル(「AJ-Aozora-Tool.accdb」ファイルの中、テーブル“Aozora_M”)或いは「MS-SQL」サーバーのデータベースの中、テーブル“Aozora_M”。

- e. 管理及び検索ツール：データベースによって、研究者各自に「Access」プログラムを資料するか、「MS-SQL」サーバーに接続して管理。発表者の場合、データベースの管理及び検索をウェブ上で行われるようにウェブページ基盤のツールを開発。

①管理：http://www.japanese.or.kr/japaneseutil/Corpus-Aozora/Corpus_Admin.aspx
http://www.japanese.or.kr/JapaneseUtil/Corpus-Aozora/Corpus_Aozora.aspx

②検索：http://www.japanese.or.kr/japaneseutil/Corpus-Aozora/Corpus_TxtDB.aspx

以上のような今回の「AJ-Aozora-Tool」を使用した「青空文庫」のデータベース化システムは、「青空文庫」の電子テキストをダウンロードすることからデータベースの構築、また形態素分析までの全課程を自動に行える点で画期的に便利であると発表者は考える。また、形態素分析を「UniDic」を内蔵した「Mecab」というオープンソースプログラムを搭載することによって、分析におけるアルゴリズムや過程が

3) 「AJ-Aozora-Tool」：発表者がプログラム言語「Visual Basic.Net 2010」で独自開発したツールで、データベースは基本「MQ-SQL2008以上」「MS-Access2010」に対応しているが、今後「MySQL」にも対応できるように改善する計画である。バグや機能を改善し、安定化した上、今後フリーソフトとして配布する予定。2012年6月の現在バージョンは、1.001。

4) 「Mecab(和布蕪)」：MeCabは 京都大学情報学研究所-日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発されたオープンソース 形態素解析エンジンです。言語、辞書、コーパスに依存しない汎用的な設計を基本方針としています。パラメータの推定に Conditional Random Fields (CRF) を用いており、ChaSenが採用している 隠れマルコフモデルに比べ性能が向上しています。また、平均的に ChaSen, Juman, KAKASIより高速に動作します。 - 「MeCab: Yet Another Part-of-Speech and Morphological Analyzer」(<http://mecab.sourceforge.net>)

透明になる。それに、今回の大量のテキストにおける形態素分析結果は、「Mecab」を学習にも応用できるので、今後より精度が高い形態素分析の結果が期待できる。ちなみに「青空文庫」には、現代日本語テキストが多数あるが、著作権の問題で近代におけるテキスト資料も多く存在している。従って、「AJ-Aozora-Tool」に内蔵されている「Mecab」の内臓電子事典である「UniDic」を、今後選択(e.g. 「近代文語UniDic」)できるようにすることによって、より幅広い時代におけるテキストデータを分析できるようになる。また、このようなシステムは「青空文庫」だけではなく、これから電子化するテキストにもそのまま再利用可能な点から考えてみると、今後のデータベースの拡張も期待できる。

続いて、データベースの検索と管理であるが、発表者は今回のデータベースや検索・管理はすべて、「MS-SQL」データベースを基盤とした「ウェブページ」上で行えるようにした((5)e参照)。それは、日本語研究に特化した「青空文庫」におけるテキストを、誰でもより簡単に早く検索できるようにするためである。また、その検索結果・原文・形態素分析とその結果におけるすべての情報をウェブで簡単に確認できるようにして、データにおける信頼性を高めた。最後に、管理ツールもまたウェブ化したのは、多数の管理者による遠隔での共同管理が可能にするためである(ID付与や管理グループ指定も可能)。もちろん、このようなウェブ上におけるデータベースとその検索システムは、個別の研究者によって特化された研究のために元のデータを修正したり、さらなる情報を付与することが容易ではない問題点がある。そのような場合、(5)c-1のように、データベースの構築段階で「AJ-Aozora-Tool」を使用して「青空文庫」を「MS-Access」ファイル形式にデータベース化することによって解決できる。「Access」ファイルで生成したDBの活用は各研究者の自由であるからである。

5. 参考文献

◆青空文庫(2007)『青空文庫10歳記念版「蔵書6300」』DVD1枚, Voyager.◆小木曾智信・近藤明日子(2007)「日本語研究のためのXMLタグ付けプログラム—その開発と活用例—」『日本語科学』22号: pp.147-159.◆小木曾智信・伝康晴・渡部涼子・近藤明日子(2009)「現代語コーパスの利用による近代語形態素解析の精度向上」『言語処理学会第15回年次大会発表論文集』: pp.801-804.◆小木曾智信・小椋秀樹・近藤明日子(2008a)「近代文語文を対象とした形態素解析辞書の開発」『言語処理学会第14回年次大会発表論文集』: pp.225-228.◆小木曾智信・小椋秀樹・近藤明日子(2008b)「近代文語文を対象とした形態素解析辞書—近代文語UniDic」『日本語学会2008年度春季大会予稿集』: pp.211-218.◆国立国語研究所(2005)『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所収.◆田原広史・南場尚子(2001)「『青空文庫』のデータベース化と研究への利用」『大阪樟蔭女子大学日本語研究センター報告』大阪樟蔭女子大学日本語研究センター, 第9号: p103~110.◆寺村秀夫(1990)「日本語学習者の日本語誤用例集」(科学研究費 特別推進研究「日本語の普遍性と個性性に関する理論的及び実証的研究」代表者井上和子, 分担研究「外国人学習者の日本語誤用例集, 整理及び分析」資料).◆野口栄司編(2005)『インターネット図書館 青空文庫』はる書房.◆富田倫生(1999)「〈イネーブル・ライブラリー〉としての青空文庫」『現代の図書館』日本図書館協会, 9月号 Vol.37 No.3: p176-181.◆富田倫生(2002)「永久機関の夢を見る青空文庫」『アート・リサーチ (Art research)』立命館大学アトリサーチセンター, Vol. 2: p49~56.◆真島知秀・金囁泳(2003)「일본어 주석 코퍼스 (tagged corpus)의 구축 방법에 대하여」『日本学報』韓国日本学会, Vol.57 No.1 pp.93-107.◆深田敦(2007)「日本語用例・コロケーション情報抽出システム『茶漉』」『日本語科学』特集 コーパス日本語学の射程, 国書刊行会: 161-172pp

【データベース・プログラム・ウェブページ】

◆『日本語話し言葉コーパス(モニター版2002)』(2002)「開放的融合研究『話し言葉工学』による」国立国語研究所.◆『朝日新聞戦前紙面データベース』(2001~2002)東京朝日新聞社, CD-ROM.◆『太陽コーパス』—雑誌『太陽』(1895-1928)日本語データベース(2005)国立国語研究所, 博文館新社, CD-ROM.◆『婦人畫報』臨川書店編集部編(2004~2005)臨川書店編集部, 臨川書店, DVD-ROM.◆『婦人公論』(2006) 臨川書店編集部, 臨川書店, DVD-ROM.◆『読売新聞』(1999~2002)読売新聞社メディア企画局データベース部, CD-ROM.◆『現代日本語書き言葉均衡コーパス』「BCCWJ領域内公開データ(2008年度版のモニター公開データ)」(2008), 国立国語研究所, DVD-ROM.◆『用例採集のための主要雑誌目録』(1983)国立国語研究所, 国立国語研究所国語辞典編集準備室.◆「国立国語研究所のデータベース目録」<http://www.ninjal.ac.jp/database>.◆「茶漉」<http://tell.fl.purdue.edu/chakoshi-wiki>.◆「ひまわり」<http://www.kokken.go.jp/lrc/index.php> (全文検索システム『ひまわり』).◆「扉~とびら~」http://karasu.xrea.bz/soft_tobira.shtml.◆「smoopy」http://site-clue.static.jp/soft_smoopy.php