日本語學硏究 第35輯 (2012. 12. 20. pp.00~00)

일본어 연구를 위한「青空文庫(아오조라문고)」 데이터베이스의 구축과 활용

-전자 텍스트 처리 프로그램 "AJ-Aozora Tool"을 활용한 데이터베이스 구축 모델 개발-

金曙泳*

< 요 지 >

「青空文庫」는 웹을 통해, 대량의 일본어 전자 텍스트뿐만 아니라, 개개의 텍스트에 관한 서지정보를 포함한 부가정보를 함께 공개하고 있는 인터넷 전자 텍스트 아카이브즈로, 다양한 시대의 수많은 저자의 텍스트가 대규모로 수록되어 있는 일본어 전자 텍스트의 보고이다.

본고에서는 이와 같은「青空文庫」를 일본어 연구에 보다 폭 넓게 그리고 효과적으로 활용 할 필요가 있다는 판단 하에,「青空文庫」를 일본어학의 연구 자료로서 보다 유용하게 이용할 수 있는 수단으로서, 체계적인 데이터베이스화와 함께 이를 관리 및 검색하는 툴의 개발이라고 하는 구체적인 모델을 제시했다. 그리고 그와 같은 모델에 따라 실제적인「青空文庫」의 데이터베이스를 구축하고, 데이터베이스 및 텍스트 처리 툴을 일반에 공개했다.

본고를 통한 데이터베이스 구축의 대략의 공정은 다음과 같다.

1)「青空文庫」의 전자 텍스트 데이터를 일괄 다운로드 :「AJ-Aozora-Tool ver1.02」이용

2) 전자 텍스트 변환 및 처리 :「AJ-Aozora-Tool ver1.02」이용

2-1) 전 XHTML 태그 텍스트 데이터를 플레인 텍스트로 일괄 변환

2-2) 전 플레인 텍스트를 일괄 형태소 분석

3) 데이터베이스 입력 :「AJ-Aozora-Tool ver1.02」이용

3-1) 플레인 텍스트의 데이터베이스화 :「MS-Access」및「MS-SQL」

3-2) 형태소 분석 결과의 데이터베이스화 :「MS-Access」및「MS-SQL」

4) 데이터베이스 관리 :「MS-Access」파일 혹은「MS-SQL」서버 내 데이터베이스와 웹

5) 데이터베이스 검색 :「MS-Access」파일 혹은 웹 검색

「青空文庫」テキスト檢索 : http://www.japanese.or.kr/japaneseutill/Corpus-Aozora/Corpus_TxtDB.aspx

논문분야 : 코퍼스 일본어학

키 워 드 : 아오조라문고(青空文庫), 일본어 코퍼스, 데이터베이스, 형태소분석, AJ-Aozora Tool, MeCab

1. 들어가며

일본어 연구 분야에 있어서 대량의 텍스트를 보다 빠르고 보다 정확하게 그리고 보다 간단하게 분석하기

* 고려대학교 강사, 일본어학(어휘론, 통어론, 코퍼스 일본어학)

위해 일본어 전자 텍스트의 수요가 나날이 커지고 있는 지금, 이에 발맞추어 텍스트 자료의 전자화도 매년 크게 진척되고 있다. 그러나 텍스트의 전자화는 많은 인력과 비용 그리고 시간 등이 소요되는 결코 간단치 않은 작업이다. 그리고 텍스트를 전자화하고 이를 연구에 응용하고 공개하는 데에는, 텍스트를 전자화하는 물리적 노력 이외에도 텍스트에 대한 저작권과 편집권 및 전자 텍스트의 공개 제한 등 풀어야할 과제가 적지 않다. 따라서 일본어 연구를 위해 대량의 전자 텍스트의 분석이 필요한 경우, 공개되어 있는 전자 텍스트의 양이 한정되어 있기 때문에, 이미 구축된 한정된 코퍼스 자료에 의존해서 연구를 진행하거나, 비교적 소규모 전자 텍스트에 대한 분석이란 점을 감수하더라도 독자적으로 전자 텍스트를 수집 혹은 구축하여 연구를 진행할 수밖에 없다.

이와 같은 현상을 고려할 때, 이미 구축되어 있는 몇 몇의 전자 텍스트 중에서도, 대량의 텍스트를 포함하고 있으면서도 더욱이 개개의 텍스트에 관한 서지정보를 포함한 부가정보도 함께 무료로 제공하고 있는 인터넷 전자 텍스트 아카이브즈「青空文庫(아오조라문고)」(그림 1 참조, http://www.aozora.gr.jp)는 일본어 연구에 있어서 매우 매력적인 일본어 전자 텍스트 자료라고 사료된다. 이에 본고에서는「青空文庫」의 전자 텍스트 데이터를 일본어학의 연구 자료로서 보다 유용하게 이용할 수 있는 수단으로서,「青空文庫」의 전자 텍스트에 대한 체계적인 데이터베이스화와 함께 이를 관리 및 검색하는 툴의 개발이라고 하는 구체적인 모델을 제시하고자 한다.

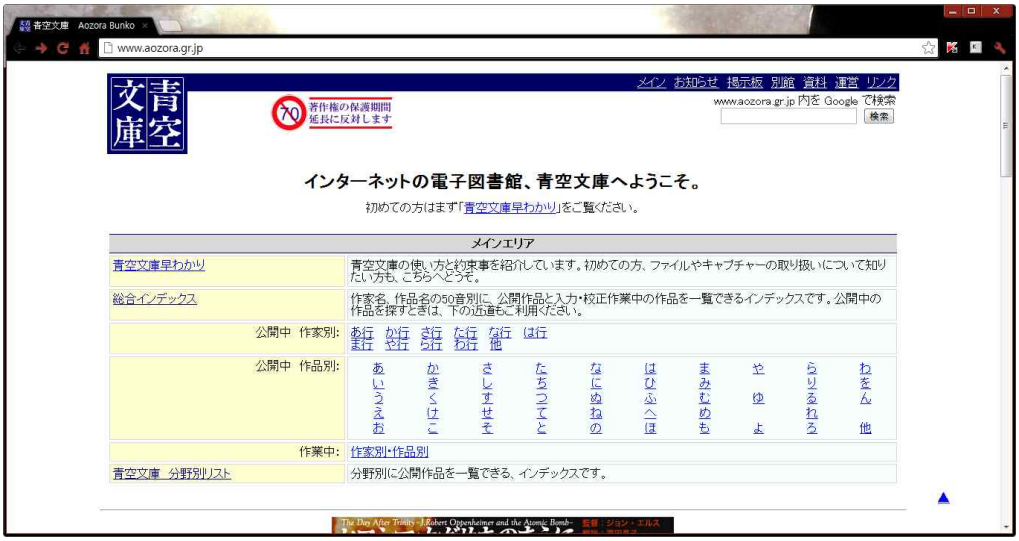

그림 1「青空文庫」의 웹페이지 http://www.aozora.gr.jp

2. 문제제기

전자 텍스트 혹은 코퍼스를 이용한 일본어 연구의 이점이라고 한다면, 무엇보다 대량의 전자 텍스트를 조사 대상으로 하여 단기간에 보다 간단하고 정확한 분석을 가능하게 해 준다는 점을 들 수 있는데, 이는 이제

와서 새삼스럽게 언급할 필요도 없을 것이다. 그뿐만 아니라, 개개의 연구를 통해 취사선택되어 수집된 용례나 통계 등의 전자텍스트 자료는, 다양한 형태(e.g. 플레인 텍스트, 코퍼스 등등, 眞島知秀·金暻泳(2003)참조)로 공개 및 교환하는 데에 용이하여, 재이용성도 매우 높다. 그리고 최근의 일본어 연구에는 전자 텍스트를 이용한 분석을 통한 통계를 제시하여 논증하는 연구가 적지 않은데, 이와 같은 연구의 연구 대상 및 결과물 등이 공개된다면, 연구의 검증 및 재현도 보다 간단하고 정확히 이루어 질 수 있게 된다. 앞서 언급한 전자 텍스트의 구축에는 다양한 측면에서 많은 노력이 필요로 하지만, 이와 같은 다양한 이점을 고려할 때, 앞으로의 일본어 연구를 위해서는 보다 적극적으로 텍스트의 전자화에 힘을 쏟아, 아직까지 전자화 되지 않은 많은 일본어 텍스트의 전자화를 추진해야만 할 것으로 보인다.

이에 저자는, 아직까지 전자화되지 않은 새로운 텍스트에 대한 전자화는 말할 것도 없거니와, 무엇보다 우선 이미 전자화된 양질의 일본어 텍스트를 체계적으로 정리하여 일본어 연구에 유용하게 활용할 필요가 있다고 생각한다. 예를 들어, 본고에서 다루고자 하는 「青空文庫」에는 이미 근세 말부터 근현대에 이르기까지 627명의 저작, 11,182건·약 9,383만어의 전자 텍스트(2012년 8월 14일 현재, 저자 조사)가 수록되어 있으며, 지금 이 시간에도 전자 텍스트의 양은 증가하고 있다[1]. 이와 같이 「青空文庫」는 다양한 시대의 수많은 저자의 텍스트가 대규모로 수록되어 있는 일본어 전자 텍스트의 보고이기 때문에, 저자는 이는 일본어 연구에 보다 폭 넓게 그리고 효과적으로 활용 할 필요가 있다고 본다.

한편, 지금까지 구축된 코퍼스 등 많은 전자 텍스트는 저본 및 원본의 저작권과 관련된 문제로, 모처럼의 시간과 노력을 들여 구축했음에도 불구하고 공개에 제약이 있거나, 공개 되었다라고 해도 그 일부만을 공개할 수밖에 없는 경우가 적지 않은 것이 현 실정이다. 이와 같이 저작권의 문제로 연구에 사용된 코퍼스 등 전자 텍스트에 접근하는 것에 제약이 따르게 되면, 연구 결과에 대한 검증·재현이 어려워질 뿐만 아니라, 전자 텍스트의 재이용 가능성이 떨어져, 일회성 전자 텍스트에 그쳐버릴 우려가 있다. 따라서 저자는, 전자 텍스트를 구축하여 이를 연구 대상으로 하는 연구자가 가능한 한 모든 노력을 기울여 그 전자 텍스트의 저작권에 관한 문제를 먼저 해결하고자 하는 노력을 기울일 필요가 있다고 생각한다. 그런데 「青空文庫」의 경우는, 저작권의 기한이 지난, 혹은 사용허가를 받은 저작만을 수록하고 있기 때문에, 그와 같은 저작권의 문제에서 자유롭다. 즉, 「青空文庫」는 수록된 전자 텍스트의 양적 측면뿐만 아니라, 재이용성·개방성이라는 활용이라는 측면에서도 매우 매력적인 전자 텍스트 자료인 것이다.

하지만, 이와 같은 「青空文庫」의 전자 텍스트는 지금까지 읽기 위한 대상으로서의 기능에만 그 활용이 치우쳐져, 전자 텍스트를 손쉽게 읽을 수 있는 다양한 뷰어[2] 등은 개발되어 왔지만, 일본어 연구에 있어서는 폭넓게 활용되지 못하였으며, 몇몇 선행연구에서 그 일부만의 전자텍스트가 조사 대상으로서 활용되는 것에 그치는 경우가 많았다. 그리고 지금까지 「青空文庫」의 방대한 전자 텍스트를 큰 하나의 데이터베이스, 즉 일본어 연구 자료로서 이용하고자 하는 관점에서의 접근하고자 하는 노력은 저자의 좁은 식견으로 판단하자면 田原広史·南場尚子(2001)의 시도를 제외하고 거의 찾아볼 수 없다. 이에 저자는 「青空文庫」에 수록되어 있는 전자 텍스트를 하나의 데이터베이스로 묶고, 이를 구축·관리·검색이 가능한 공개 코퍼스로 공개한다

1) 「青空文庫」에 수록된 전자 텍스트는 2009년 9월에 4,843건(野口榮司 : 2005), 2007년 7월에는 6,300건(青空文庫 : 2007)을 기록했으며, 2010년 5月에는 9,624건(青空文庫 : 2010)을 기록하는 등, 수록된 전자 텍스트는 매년 계속해서 증가하고 있음을 확인할 수 있다.

2) 「青空文庫」의 텍스트 파일을 읽는 기능에 특화된 프로그램으로서, 「smoopy」와 「扉～とびら～」와 같은 프리 소프트웨어를 들 수 있다. 그러나 「○○○○.ebk」와 같은 특수한 확장자를 가진 일부 파일은, (1)c와 같이 「익스팬드북 브라우저(Expand Book Browser·エキスパンドブックブラウザー)」라는 전용 툴이 필요하다.

면, 앞으로 다양한 분야의 일본어 연구에 보다 효과적으로 활용될 수 있을 것으로 판단한다. 본고에서는 小木曾智信·近藤明日子(2007)와 같은 전자 텍스트 처리방법에 관해 논한 선행논문을 참고하여, 「青空文庫」에 수록된 전자 텍스트를 체계적인 데이터베이스로서 구축하고, 이를 효과적으로 관리하면서 실제의 연구에 응용할 수 있는 검색 툴에 이르기까지의 일관된 시스템을 개발한다는 구체적인 "공개 코퍼스 데이터베이스 개발 모델"을 제시하고자 한다.

3.「青空文庫」

「青空文庫」는 1997년 이래, 저작권 문제가 해결된 일본어 텍스트를 전자화하여, 그 데이터를 작가·타이틀 및 분야별로 분류하여 웹상에 공개하고 있는 인터넷 도서관이다(富田倫生(1999 : 176)). 그리고 野口栄司(2005)가 언급한 것처럼, 「青空文庫」는 순수하게 텍스트 기록이나 자료를 수집하고 이를 체계적으로 보관·관리·공개하고 있다는 점에서 엄밀하게 말하자면 아카이브즈(Archives)에 가깝다고 하겠다. 앞서 언급한 바와 같이, 「青空文庫」는 인터넷 상에 공개되고 있기 때문에, 언제든지 인터넷을 통해 그 전자 텍스트 자료를 열람하는 것이 가능하다[3]. 이용자는 열람하고자 하는 전자 텍스트를 작가·작품 혹은 분야에서 선택하여 원하는 전자 텍스트를 다운로드하여 읽을 수 있는데, 이와 같이 「青空文庫」에서 다운로드 가능한 텍스트 파일은 크게 다음과 같은 세 가지 형식이 있다(그림 2, 예제(1) 참조).

ファイルのダウンロード

ファイル種別	圧縮	ファイル名(リンク)	文字集合／符号化方式	サイズ	初登録日	最終更新日
テキストファイル(ルビあり)	zip	1509_ruby_4945.zip	JIS X 0208／ShiftJIS	129395	2000-12-02	2010-11-02
エキスパンドブックファイル	なし	1509.ebk	JIS X 0208／ShiftJIS	413016	2000-12-02	2000-12-10
XHTMLファイル	なし	1509_40739.html	JIS X 0208／ShiftJIS	472532	2010-09-24	2010-11-02

그림 2「青空文庫」텍스트 파일 다운로드 :
島崎藤村의「家」http://www.aozora.gr.jp/cards/000158/card1509.html

(1) a. 텍스트 파일(후리가나振り仮名 유·무) : 태그 등이 없는 순수한 텍스트(이하, 플레인 텍스트)로, 압축된 상태로 다운로드하는 것이 가능하다. 텍스트를 확인하기 위해서는 우선, 다운로드한 파일의 압축을 푸는 작업이 필요하다. 하지만, 플레인 텍스트라고는 하지만, 해당 파일 안에는 "텍스트에 나타난 기호에 관하여(テキスト中に現れる記号について)", "저본(底本)", "초출(初出)", "입력자(入力者)", "교정자(校正者)", "작성일(作成日)", "그 외, 안내·주의사항(その他, 案内·注意事項)" 등의 다양한 부가정보가 함께 들어 있다. e.g. 1509_ruby_4945.zip : 그림 2 참조

b. HTML파일(혹은 XHTML) : HTML 혹은 XHTML이라고 하는 웹페이지 형식의 태그 파일로,

3) 다소간 시간은 지났으나, 인터넷뿐만 아니라 『青空文庫10歳記念版「蔵書6300」』와 같은 단행본의 부록으로 제공되는 DVD 등 전자매체를 통해 자료를 열람하는 것도 가능하다.

HTML혹은 XHTML의 규약에 따라 (1)a과 같은 플레인 텍스트에 태그가 부여되어 있다. 괄호가 쳐진 후리가나振り仮名 외에도, (1)a과 같은 부가정보도 함께 첨부되어 있다. e.g. 1509_40739.html : 그림 3 참조

c. 익스팬드북 파일(Expand Book file·エキスパンドブックファイル) : 익스펜디드북 브라우저(Expanded Book Browser·エキスパンドブックブラウザー)라고 하는 뷰어 전용 파일. 일본어판 혹은 영문판 뷰어의 다운로드는 다음의 경로와 같다.

http://www.voyager.co.jp/software/ebdl.html#download

e.g. 1509.ebk : 그림 2 참조

범례 1) 위 예시의 전자 텍스트 파일의 작품은 시마자키 토손島崎藤村의「家」로, 파일명의 번호는 작품번호. 2) 전자 텍스트 파일의 부호화 방식(encoding)은, Shift-JIS.

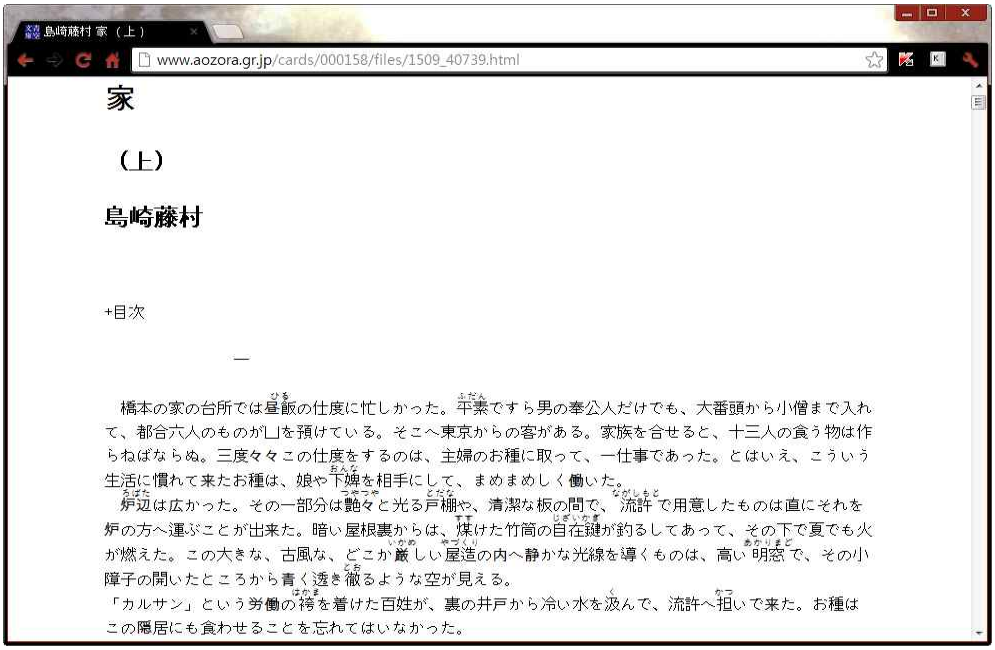

그림 3 島崎藤村의「家」의 전자 텍스트 파일, 1509_40739.html

「青空文庫」홈페이지는 위와 같이 각각의 전자 텍스트에 대한「図書カード(도서카드)」페이지를 설치하여, 전자텍스트 파일 뿐만 아니라, 각 텍스트에 관한 다양한 정보도 함께 제공하고 있다(작품명, 저자명, 가나 표기법 종별(仮名遣い種別), 작가 데이터(생몰, 인물설명), 저본(底本) 데이터, 공작원(工作員)[4] 데이터 등). e.g. 島崎藤村의「家」: http://www.aozora.gr.jp/cards/000158/card1509.html

4) 공작원(工作員) : 「青空文庫」는 다양한 자원봉사자들의 지원을 받고 있습니다. 우리들은 그 중에서 (전자 텍스트의) 입력이나 교정 작업을 담당해 주시는 분들을「공작원(工作員)」이라고 부르고 있습니다(출처 :「青空文庫」홈페이지. 본고의 저자에 의한 번역).

4.「青空文庫」의 데이터베이스화

4.1「青空文庫」의 문자열 검색과 전자텍스트의 형식

우선, 종래의 일본어 연구에 있어서「青空文庫」를 사용하여 문자열 검색을 실시하는 방법을 예로 들자면 다음의 (2)와 같다.

(2) a.「KWIC Finder[5]」,「秀丸(히데마루)」등, 정규표현식(정규식)을 사용하여「grep」「KWIC」검색이 가능한 소프트웨어를 사용한 문자열 검색
  1)「青空文庫」의 플레인 텍스트를 다운로드하여 한 개씩 압축을 해제하여 하나의 폴더에 넣어 정리
  2) 검색하고자 하는 어·구·문을 폴더내의 파일에 정규표현식을 사용하여 검색
 b. 전문검색 시스템「ひまわり」를 사용하여 문자열 검색
  1)「青空文庫」의 플레인 텍스트를 다운로드하여 한 개씩 압축을 해제하여 하나의 폴더에 넣어 정리
  2)「ひまわり」용 텍스트 데이터 변환 툴인「えだまめ[6]」를 사용하여 HTML 텍스트를「ひまわり」용의 XML형식의 파일로 변환
  3) 변환한 XML형식의 텍스트 데이터를「ひまわり」로 읽어 검색

그림 4「ひまわり」의 검색결과 화면

여기에서 플레인 텍스트는 말 그대로 순수한 텍스트만을 가리키는 것으로, 품사정보 등 부가적인 언어정보를 포함하고 있지 않기 때문에, 문자열 검색 등 최소한의 용례검색만이 가능하다. 물론, (2)b와 같은「ひまわり」에서는 HTML 파일을 사용하고 있으나, 부가적인 언어정보라는 측면에서 플레인 텍스트와 별다른 차이가 없다고 보아도 무방하다. 그러나 플레인 텍스트라고는 하지만,「青空文庫」의 플레인 텍스트에는 직접적인 언어연구의 대상이 되는 본문 텍스트 이외에도, 전자 텍스트화 하는 과정에서 입력자 및 교정자 등의 부가정보가 텍스트 안에 수록되어 있다. 따라서 텍스트 전문 검색 및 통계에 있어서 이와 같은 전자 텍스트를 분석하여 올바른 결과를 도출하기 위해서는 각각의 전자 텍스트의 부가정보를 제거한 "진정한" 플레인 텍스트를 정제하고 분석해야만 하는 필요성이 대두된다. 물론, 이와 같은 정보는 전자 텍스트를 일본어 연구 자

5)「KWIC Finder」 : http://www31.ocn.ne.jp/~h_ishida/KWIC.html
6)「えだまめ」 : 폴더에 분류한 파일(텍스트·XML·HTML)로부터「ひまわり」에서 이용 가능한 데이터(XML 파일)을 만들어내기 위한 소프트웨어.「http://www2.ninjal.ac.jp/lrc」의「『ひまわり』支援ツール『えだまめ』」메뉴에서 다운로드 가능.

료로서 사용하는 데에 있어서 결코 불필요한 정보가 아니기 때문에, 전자 텍스트 본문과 함께 확실히 그리고 즉각적으로 확인할 수 있는 형태로 보존해 둘 필요성이 있음은 두말할 나위가 없다. 예를 들어,「ひまわり」와 같이 본문뿐만 아니라, 작가, 작가의 생년월일, 출판사, 분야 등의 정보는 용례검색에 있어서 중요한 기초정보, 즉 서지정보를 검색 결과와 함께 제공해 주는 것은 일본어 연구에 너무나도 편리하고 필수적인 기능이라고 볼 수 있는 것이다.

「青空文庫」는 대규모의 텍스트 데이터를 수록하고 있기 때문에, 그것만으로 충분히 의미가 있지만, 언어연구에 유용한 자료로서 보다 광범위하게 활용되기 위해서는 그와 같은 전자 텍스트를 형태소 분석을 하거나, XML 등의 형식으로 언어 정보 태그를 부여하는 등의 가공을 거쳐야 할 필요가 있다. 또한 그와 동시에, 전자 텍스트를 효과적으로 검색하고 관리할 툴도 역시 요구된다. 이는 小木曾智信외(2007 : 147-148)에서 지적한 바와 같이, 일본어 연구에 있어서 XML과 같은 태그를 부여한 전자 텍스트를 이용함에 있어서 XML 문서에서 태그를 제거한 단순한 플레인 텍스트만을 이용하거나, 코퍼스와 함께 제공되는 검색 툴을 통해 출력되는 텍스트 데이터를 이용하는 것도 두말할 나위 없이 중요한 연구 방법이지만, 그와 같은 방식으로 연구를 진행할 경우, 자칫하면 원래의 텍스트 데이터를 잘못 이용할 수 있는 가능성[7]이 상존하기 때문이다. 따라서 본고의「「青空文庫」데이터 베이스 구축과 활용 모델」에서도 단순히 플레인 텍스트의 수집을 통해 데이터베이스를 구축하는 것뿐만 아니라, 일본어 연구에 활용 가능한 다양한 정보(e.g. 서지정보)를 플레인 텍스트와 함께 체계적으로 정리하여 데이터베이스를 구축했다.

그러나 실제의 연구에 있어서 전자 텍스트의 수요는, 플레인 텍스트를 시작으로, HTML, XHTML, XML 그리고 코퍼스의 데이터베이스 파일 등등, 연구자의 필요나 텍스트 처리기술에 따라 일본어 연구에는 다양한 재형식의 저자 텍스트가 요구된다. 따라서 본고에서는 데이터베이스를 구축함에 있어서 자동화를 전제로 하되, 전자 텍스트를 처리하는 단계·과정을 세분화하여 다양한 전자 텍스트 형식을 연구자가 얻을 수 있도록 하는 것을 통해, 다양한 연구에 대응할 수 있도록 하는 것을 목표로 했다. 그 공정을 간단히 정리하자면, 우선「青空文庫」에 수록된 XHTML형식의 전자 텍스트를 다운로드 하고, 플레인 텍스트 혹은 해당 플레인 텍스트를 형태소 분석한 전자 텍스트 등에 각각의 서지정보 등 부가정보를 추가하여 데이터베이스를 구축했다. 그리고 그와 같은 다양한 형식의 전자 텍스트의 검색시스템으로서 플레인 텍스트와 형태소 분석 텍스트를 넷 상에 업로드 하고 검색 가능한 형태로 공개했다. 이를 통해, 전자 텍스트의 수집과 가공 그리고 검색 및 공개라고 하는 일련의 데이터베이스화 모델을 개발하고자 했다. 그 구체적인 방법은 다음 절에서 자세히 다루고자 한다.

4.2「青空文庫」의 데이터베이스화

본고의 데이터베이스화를 논하기에 앞서 우선, 선행연구에 있어서의「青空文庫」의 데이터베이스화에 관해 간단히 언급한 후, 본론으로 들어가고자 한다. 田原広史·南場尚子(2001 : 1-4)는, 이하 (3)과 같이, 전자 텍스트 데이터를 정제하고, 해당 전자 텍스트와 함께 유효한 정보를 각각의「MS-Excel」의 데이터 필드에

---

7) a. 태그를 제거한 전자 텍스트만을 이용할 경우, 텍스트 데이터의 오용으로 이어질 가능성이 있다.
b. 원래의 텍스트 데이터가 가진 정보 중에서 극히 일부만을 이용할 수밖에 없다.
c. 조사결과가 원래의 텍스트 데이터에 반영되지 않기 때문에, 조사 당시에 한정된 자료로서만 활용되고 버려지게 되고, 조사 결과를 다른 조사 결과와 함께 종합적으로 분석하는 것이 불가능하게 된다.
小木曾智信·近藤明日子(2007 : 147-148)

입력하여, 이를 검색 가능할 수 있는 형태로 묶어 내는 수법을 통해, 「青空文庫」의 데이터베이스화를 시도했다.

(3) a. 「青空文庫」의 전자 텍스트 다운로드 : 수작업으로 「青空文庫」 홈페이지에서 임의의 작품을 개별적으로 다운로드
b. 압축 해제 : zip형식의 압축파일을 텍스트 파일로 압축해제
c. 파일명 변환 : 직관적이지 않은 파일명(e.g. kakekomi-rubi.txt)을 "0018駆け込み訴え.txt"와 같이 관리하기 쉬운 이름으로 변경
d. 「MIFES[8]」를 사용하여 텍스트를 문단위로 구분 : 메가소프트사의 소프트웨어인 「MIFES」를 사용하여 텍스트를 문단위로 구분
e·f. 텍스트 파일을 「MS-Excel」에 입력 : 한 문장이 한 줄로 구분된 텍스트 데이터를 「MS-Excel」에서 읽고, 추가로 작품명·작가·작품 내 문 번호 등의 서지정보와 그 외에 필요한 데이터를 추가
g. 「MS-Excel」의 각 시트를, 새로 만든 시트에 이동시켜 병합 : 「MS-Excel」에서는 행의 개수에 제한이 있으나, 대규모의 데이터를 한 번에 검색할 수 있기 때문에 가능한 한 하나의 파일로 묶어 두되, 한 작가별로 하나의 시트로 병합하는 것을 원칙으로 함
h. 「파일메이커[9]」에 「MS-Excel」 파일의 각 시트를 입력 : 몇 몇의 파일로 병합한 「MS-Excel」 파일을 데이터베스화하기 위해, 파일메이커에 「MS-Excel」 파일을 입력. 단, 파일메이커에는 행수의 제한이 없기 때문에 둘 이상의 시트로 분할해 두었던 「MS-Excel」 파일을 작가별로 하나의 파일로 병합하는 것이 가능. 예를 들어 芥川龍之介의 경우 180건의 작품 약 52,000행, 太宰治는 112건의 작품 약 65,000행 등을 각각 하나의 파일로 병합
i. 파일메이커용 레이아웃 작성 : 입력할 때 자동으로 작성되는 단순한 레이아웃의 경우, 이후의 사용에 불편이 따르기 때문에, 「검색용 레이아웃」 혹은 「인자용(印字用) 레이아웃」의 두 종을 작성. 단, 검색 결과가 대량일 경우, 「MS-Excel」 파일로 돌아가 검색하는 편이 효율적
j. 데이터베이스 완성 : 부가정보를 입력하기 위한 몇몇의 필드를 작성

범례) 요약 및 번역은 본고의 저자. 田原広史·南場尚子(2001 : 1-4)

그렇지만, 위와 같은 공정은, 그 과정이 복잡하고, 거의 대부분의 공정이 수작업이기에 효율이 나쁘고, 따라서 자연스럽게 데이터베이스의 규모가 작아질 수 밖에 없다. 또한 데이터베이스에 구축 및 접근 시, 라이선스가 필요한 상용 소프트웨어만을 사용하고 있는 점을 문제점으로 지적할 수 있다. 하지만, 전자 텍스트 데이터를 부가정보와 함께 검색 가능한 형태로 묶고자 했다는 점, 그리고 사용자가 자유롭게 데이터베이스의 원문을 수정하는 것이 가능하고 필요한 정보 필드를 추가하는 것이 가능하도록 방안을 강구했다는 점에서는 높게 평가할 만 하다고 하겠다. 이는 방법은 다소 다르지만, 小木曾智信·近藤明日子(2007)에서 언급하고 있는 "연구자 자신이 XML문서에 태그를 부여하고, XML관련 기술을 적극적으로 이용하여 태그 정보를 추출해 낸다"는 개념과 일맥상통하는 것이라고 볼 수 있다.

8) 「MIFES(마이페스)」 : 프로그램 개발, 로그/CSV편집, Web제작, 원고작성 등의 기능을 가진 텍스트 에디터. 라이선스가 필요한 상용 소프트웨어. 참고 : http://www.megasoft.co.jp/mifes/about.html

9) 파일메이커사(Filemaker Inc.)가 개발 공급하는 관계형 데이터베이스 소프트웨어. 1985년 처음 출시된 이래 현재 버전 11.0까지 출시되었으며, 윈도와 매킨토시 OS-X용 버전이 존재. 아이폰과 아이패드에서도 이용할 수 있는 간이형 버전도 출시되었다.

앞서 언급한 바와 같이「青空文庫」는 세 가지 형식의 텍스트 데이터를 제공하고 있으나((1) 참조), 일반적으로 일본어 연구를 목적으로 한다면, HTML 태그가 부여된 텍스트나 전용 뷰어(익스팬드북)의 태그가 부여된 텍스트 보다는, 플레인 텍스트 쪽이 보다 많이 사용되고 있는 것이 지금의 현실이라고 할 수 있다. 따라서 문자열 검색이나 형태소 분석을 수행할 때에는 우선, (1)a·b에서 언급한 것과 같은 태그 및 부가정보를 분리하는 작업을 필요로 하게 된다. 물론,「えだまめ」와 같은 툴을 사용하면, (2)b와 같이 다소간의 준비공정이 필요하지만, 한번에「青空文庫」의 전자 텍스트를 변환하여 검색 소프트웨어인「ひまわり」에 탑재할 수 있는데, 전자 텍스트에 관한 검색 속도가 빨라질 뿐만 아니라, 그 정확도 또한 높아지게 된다. 또한 무엇보다 직관적인 인터페이스로 알기 쉽기 때문에, 문자열 검색뿐만 아니라,「青空文庫」에 대한 언어조사에 편리할 것이라고 생각된다. 그러나 (2)b와 같은 과정을 통해 데이터베이스를 구축하게 되면, 검색대상이 되는 텍스트 데이터가「ひまわり」의 고유의 독자적인 파일 형식으로 변환되기 때문에 원래의 텍스트 파일의 내용에 접근하기 어렵게 된다. 따라서「ひまわり」에 의해 XML형식으로 변화된 텍스트는, 그 플레인 텍스트의 내용에 접근하기 어렵기 때문에, 결과적으로 형태소 분석이나, 태그 코퍼스 등에 재이용하는 것이 용이하지 않은 폐쇄적인 면이 존재한다. 또한 이 시스템에서는 새로운 데이터가 추가될 때마다, 일부 데이터만을 추가하는 것도 용이하지 않다.

이에 본고에서는「青空文庫」의 텍스트 데이터를 우선 순수한 본문 텍스트와 텍스트에 관한 정보(서지 정보 등)으로 분리한 뒤, 체계적으로 정리하여 데이터베이스를 구축했다. 또한 이와 같은 플레인 텍스트를 형태소 분석하여 동 데이터베이스에 추가하였다. 그리고 전자 텍스트 데이터를 정제하고, 순수한 본문 텍스트를 추출하여 체계적으로 정리하는 데에 있어서는 지금까지 "읽기"에 초점을 두었던「青空文庫」의 텍스트를, 언어연구에 적합한 텍스트 데이터로 변환한다는 차원에서 큰 의미가 있다. 또한 단순히 플레인 텍스트를 추출하는 것에서 그치는 것이 아니라,「青空文庫」의 전자 텍스트에 포함되어 있는 부가정보를 정리하여 마찬가지로 동 데이터베이스에 입력하는 것을 통해,「ひまわり」의 검색결과와 같이, 일본어 연구에 있어서 필요한 정보를 전자 텍스트와 함께 일목요연하게 확인 하는 것이 가능하다. 이에 더해 플레인 텍스트를 형태소 분석한 텍스트를 동일선상에서 데이터베이스화하는 것을 통해, 본 데이터베이스가 전자 텍스트에 대한 단순 문자열 검색에 활용되는 것에 그치는 것이 아니라, 앞으로 보다 고도의 언어 연구에 응용될 수 있는 기초자료로서의 가치를 가질 수 있도록 강구했다.

본고에서는 이상과 같은 점을 고려하여,「青空文庫」를 데이터베이스화 했으나, 그 과정에서 다음과 같은 점에 중점을 두어 작업을 진행했다.

(4) a. 텍스트 데이터와 부가정보를 함께 수록한 데이터베이스를 구축하고, 추후에 정보 필드를 확장할 수 있도록 확장성을 강구할 것.
b.「青空文庫」의 텍스트 데이터를 보다 간단하고 정확하게 추출해 낼 수 있는 수단(프로그램과 같은 툴)을 강구할 것.
c. 관리자가 데이터베이스를 간단히 유지·보수·갱신할 수 있는 수단을 강구할 것.
d. 이용자가, 데이터베이스를 간단히 검색하고, 원본 및 부가 정보에 접근할 수 있으며, 그리고 그 각각의 정보 와 갱신 정보를 간단하게 확인할 수 있도록 공개할 것.
e. 이용자가 독자적인 개개의 연구정보를 추가할 수 있는 수단을 강구할 것.

4.3「AJ-Aozora-Tool」에 의한「青空文庫」의 데이터베이스화

본고에서는 앞선 절의 (4)과 같은 점을 고려하여 「青空文庫」의 데이터베이스를 구축했으며, 그 공정 및 순서는 다음과 같다. 참고로 본고에서는 데이터베이스 구축의 자동화를 위해 전용 소프트웨어 「AJ-Aozora-Tool[10]」를 개발하여, 웹페이지(그림5 참조)를 통해 공개하고 있다. 프로그램에 관한 자세한 설명은 별도의 논문을 통해 논할 예정이며 프로그램의 배포 및 간단한 매뉴얼은 홈페이지를 통해 공개하고 있다.

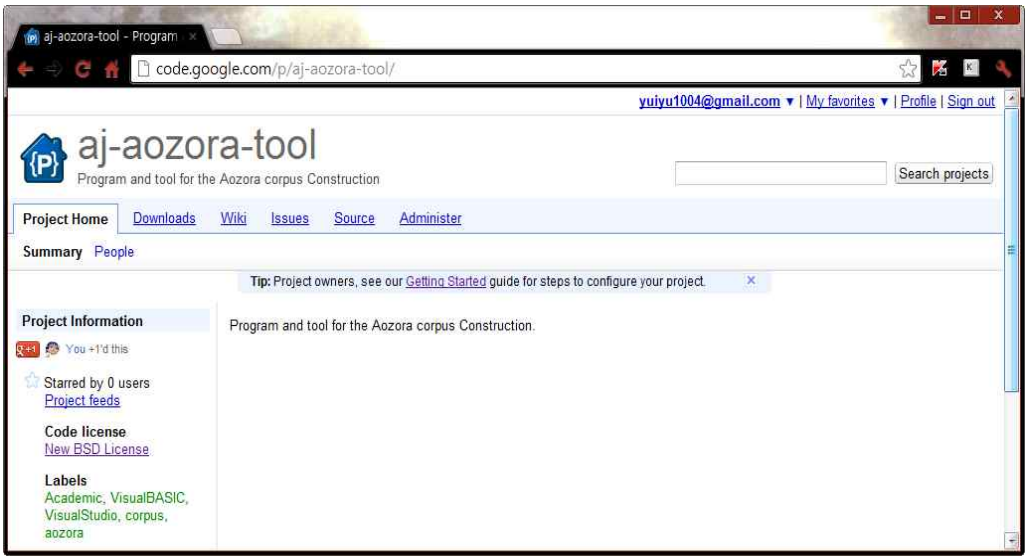

그림 5「AJ-Aozora-Tool」배포 사이트 http://code.google.com/p/aj-aozora-tool

(5) a.「青空文庫」의 전자 텍스트 데이터를 일괄 다운로드 : 「AJ-Aozora-Tool ver1.02」를 사용하여 모든 텍스트를 일괄 다운로드. ⇒ 결과물 : XHTML 태그 텍스트 데이터. e.g. 236_19996.html : 『ア, 秋』 太宰治(1975)

cf. 단,「AJ-Aozora-Tool」의 일반 배포판에는 본 소프트웨어로 인한 전문 텍스트 다운로드로 인해「青空文庫」의 서버에 과도한 부하를 끼칠 우려가 있기에,「青空文庫」전자 텍스트 일괄 다운로드 기능은 비활성화 하여 배포함. 그러나 연구자를 위해 모든 텍스트 파일은 다음의 경로에서 다운받을 수 있도록 했다.

「2. Download AJ-Aozora-Corpus」메뉴의 「2.1 Full XHTML Text」

- "AJ-Aozora XHTML 20120814.iso" : Ver.20120814

: http://www.japanese.or.kr/japaneseutill/Corpus-Aozora/About_Corpus_Aozora.aspx

10)「AJ-Aozora-Tool」 : 본고의 저자가 프로그램 언어인 「Visual Basic.Net 2010」을 사용하여 독자 개발한 MS 윈도 기반(윈도 XP 이상 지원) 툴로, 기본적으로 「MS-Access2010」와 「MQ-SQL2008」 이상의 데이터베이스에 대응한다. 금후, 「MySQL」 데이터베이스까지 대응할 수 있도록 개선할 예정. 2012年9月 현재 버전은 1.001이며, 다음의 배포 웹페이지에서 공개 배포 중.

- 배포 사이트 : http://www.japanese.or.kr/japaneseutill/AJ-AoZora Tool/About AJ-AoZora_Tool.aspx

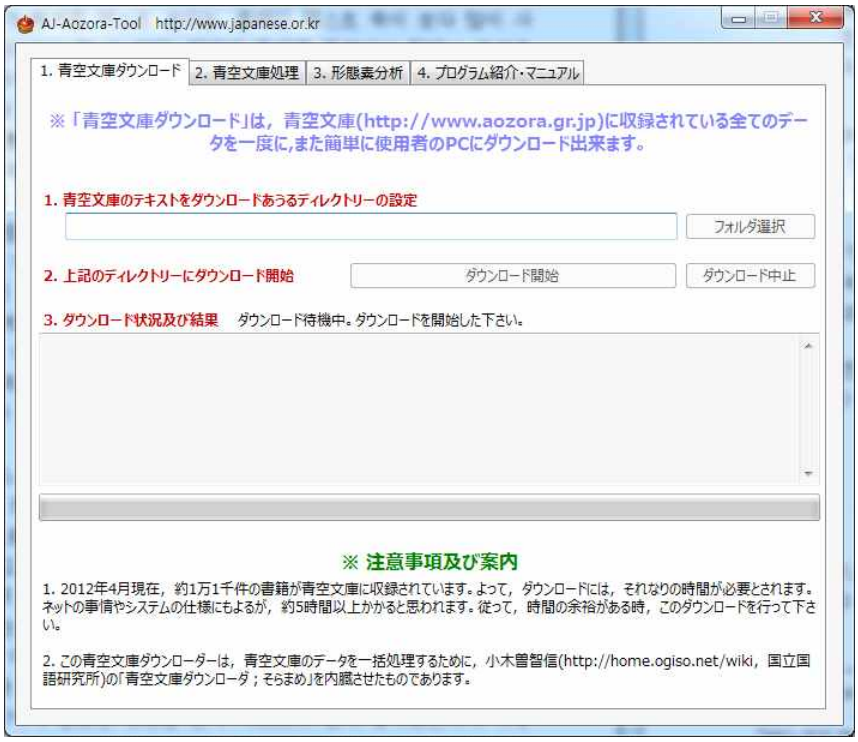

그림 6「AJ-Aozora-Tool」을 이용하여 다운로드

표 1 일괄 다운로드한 파일 중, 236_19996.html :『ア，秋』太宰治(1975)

```
<?xml version="1.0" encoding="Shift_JIS"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
    "http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="ja" >
<head>
        <meta http-equiv="Content-Type" content="text/html;charset=Shift_JIS" />
        <meta http-equiv="content-style-type" content="text/css" />
        <link rel="stylesheet" type="text/css" href="../../default.css" />
        <title>太宰治 ア、秋</title>
                                ..... 중략 .....
<br />
<br />
本職の詩人ともなれば、いつどんな注文があるか、わからないから、常に詩材の準備をして置くのである。
<br />
                                ..... 이하 생략 .....
```

그림 7「AJ-Aozora-Tool」을 이용하여 HTML 태그 텍스트를 플레인 텍스트로 일괄변환

b. 전자 텍스트 변환 및 처리

1) XHTML 태그 텍스트 데이터를 플레인 텍스트로 변환 : (5)a에서 특정 폴더에 다운로드 한 XHTML 태그 텍스트 파일들을 「AJ-Aozora-Tool」을 사용하여 플레인 텍스트로 일괄 변환 ⇒ 결과물 : 플레인 텍스트(①후리가나振り仮名포함 플레인 텍스트, ②후리가나振り仮名불포함 플레인 텍스트, ③서지정보 등 부가정보만을 제거한 XHTML 태그 텍스트와 같은 세 가지 옵션 중에서 선택 가능, 그림7 참조). e.g. 변환전 : 236_19996.html ⇒ 변환결과 : 236_19996.txt

2) 플레인 텍스트를 형태소 분석 : (5)b-1의 공정에서 얻어진 플레인 텍스트(②후리가나振り仮名불포함 플레인 텍스트)를 「AJ-Aozora-Tool」에 내장된 「MeCab[11]」을 사용하여 일괄 분석 ⇒ 결과물 : 형태소 분석 텍스트 파일. e.g. 변환전 : 236_19996.txt ⇒ 변환결과 : 236_19996_m.txt

표 2 일괄 변환한 파일 중, 후리가나振り仮名불포함 플레인 텍스트,
236_19996.txt :『ア, 秋』太宰治(1975)

ア、秋
太宰 治

…… 중략 ……

本職の詩人ともなれば、いつどんな注文があるか、わからないから、常に詩材の準備をして置くのである。「秋について」という注文が来れば、よし来た、と「ア」の部の引き出しを開いて、愛、青、赤、アキ、いろいろのノオトがあって、そのうちの、あきの部のノオトを選び出し、落ちついてそのノオトを調べるのである。

…… 이하 생략 ……

11) 「MeCab(和布蕪·메카부)」 : MeCab는 교토대학 정보학연구과와 일본전신전화주식회사 커뮤니케이션과학기반연구소의 공동연구유닛 프로젝트를 통해 개발된 오픈소스 형태소분석 엔진이다. MeCab는 언어, 사전, 코퍼스에 의존하지 않는 범용적인 설계를 기본 방침으로 하고 있다. 파라미터의 추정에 컨디셔널 랜덤 필드(Conditional Random Fields·CRF) 사용하고 잇으며, ChaSen이 채용하고 있는 은익 마르코프모델과 비교해 볼 때 보다 나은 성능을 보여준다. 또한 평균적으로 ChaSen, Juman, KAKASI보다 고속으로 동작한다. - 「MeCab: Yet Another Part-of-Speech and Morphological Analyzer」(http://mecab.sourceforge.net). 범례) 본고의 저자 번역.

c. 데이터베이스 입력

1) 플레인 텍스트의 데이터베이스화 : (5)a에서 특정 폴더에 다운로드한 XHTML 태그 텍스트 파일의 플레인 텍스트와 각각의 부가정보를「AJ-Aozora-Tool」를 사용하여「MS-Access」혹은「MS-SQL」데이터베이스로 일괄 입력하여 구축 ⇒ 결과물 :「MS-Access」데이터베이스 파일(「AJ-Aozora-Tool.accdb」파일 중, 테이블 "Aozora"·"Aozora_P") 혹은「MS-SQL」서버의 데이터베이스 중 테이블 "Aozora"·"Aozora_P". 표 3 및 그림 9 참조.

cf.「MS-Access」플레인 텍스트 데이터베이스 및 플레인 텍스트 파일 공개

「2. Download AJ-Aozora-Corpus」메뉴의「2.1 Full Plain Text」와「2.4 Full XHTML + Plain Text of Aozora - MS-Access Ver.」

→ http://www.japanese.or.kr/japaneseutill/Corpus-Aozora/About_Corpus_Aozora.aspx

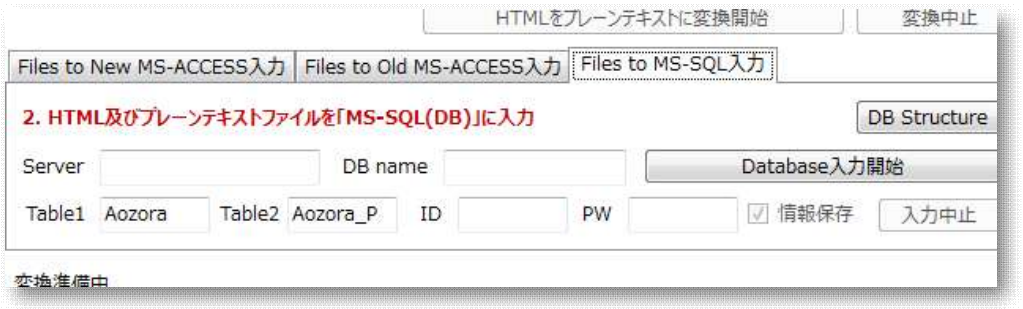

그림 8「AJ-Aozora-Tool」을 이용하여「MS-SQL」일괄 DB 구축

2) 형태소 분석 결과의 데이터베이스화 : (5)b-2를 통해 특정 폴더에 모은 형태소 분석 텍스트 파일들을「AJ-Aozora-Tool」을 사용하여,「MS-Access」혹은「MS-SQL」서버의 데이터베이스에 일괄입력 ⇒ 결과물 :「MS-Access」데이터베이스 파일(「AJ-Aozora-Tool.accdb」파일 중, 테이블 "Aozora_M") 혹은「MS-SQL」서버의 데이터베이스(테이블 "Aozora_M"). 표3 참조, 그림 9 참조.

cf.「MS-Access」형태소 분석 데이터베이스 및 형태소 분석 텍스트 파일 공개

「2. Download AJ-Aozora-Corpus」메뉴의「2.3 Full Morpheme Text of Aozora」와「2.5 Full Morpheme Text - MS-Access Ver.」

→ http://www.japanese.or.kr/japaneseutill/Corpus-Aozora/About_Corpus_Aozora.aspx

표 3「AJ-Aozora-Tool」에 의한「青空文庫」데이터베이스 구조(「MS-SQL」의 경우)

Table	Aozora	Aozora_P	Aozora_M	Aozora_Author
Table design	Id bigint FileName nvarchar(500) Title ntext Author ntext AuthorId nvarchar(500) BirthDeath ntext First ntext Base ntext Kana ntext About ntext BibInfo ntext Text ntext Year nchar(30) Etc1 ntext Etc2 ntext Etc3 ntext Etc4 ntext Input_Date nvarchar(50)	Id bigint FileName nvarchar(500) Text_Planen text Text_Plane_Kana ntext Etc1 ntext Etc2 ntext Etc3 ntext Etc4 ntext Input_Date nvarchar(50)	Id bigint FileName nvarchar(500) Text_Morpheme ntext Etc1 ntext Etc2 ntext Etc3 ntext Etc4 ntext Input_Date nvarchar(50)	original_data nvarchar(255) date datetime ID nvarchar(255) Name nvarchar(255) Name2 nvarchar(255) [Read] nvarchar(255) Name_Eng nvarchar(255) Postion nvarchar(255) Field nvarchar(255) Birth nvarchar(255) Death nvarchar(255) History nvarchar(255)

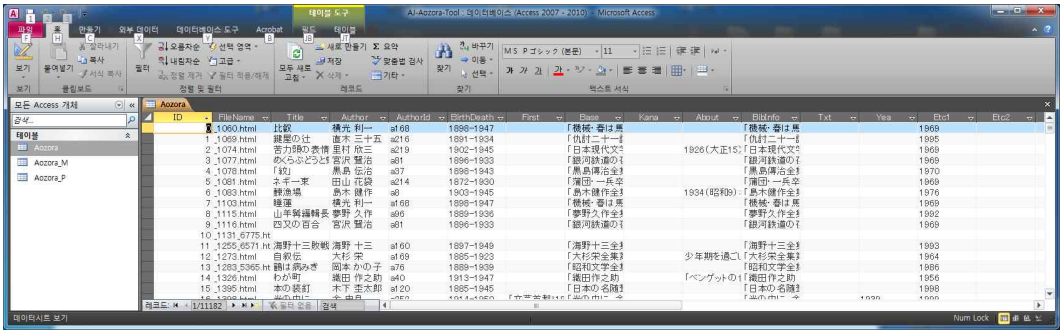

그림 9「MS-Access」로 구축한「青空文庫」데이터베이스

d. 데이터베이스 관리 : (5)c 공정을 통해 사용자가 각자 구축한 데이터베이스의 종류에 따라,「MS-Access」프로그램을 사용하여 관리하거나(그림9 참조),「MS-SQL」서버에 접속하여 관리. 참고로 본고의 저자는,「MS-Access」데이터베이스를 다음의 웹을 통해 공개하고 있으며,「MS-SQL」서버에도 동일한 데이터베이스를 구축하고 웹상에서 이를 관리할 수 있도록 관리 페이지를 구축하고 있다.

1) 관리 : 관리자 아이디로 접속하여 데이터베이스를 조회하고 정보를 갱신하는 등의 제반 관리기능 개발. 단, 관리자 아이디 필요(그림 10·11참조).

http://www.japanese.or.kr/japaneseutill/Corpus-Aozora/Corpus_Admin.aspx

그림 10「青空文庫」데이터베이스 관리자 페이지 – ① 메인 페이지

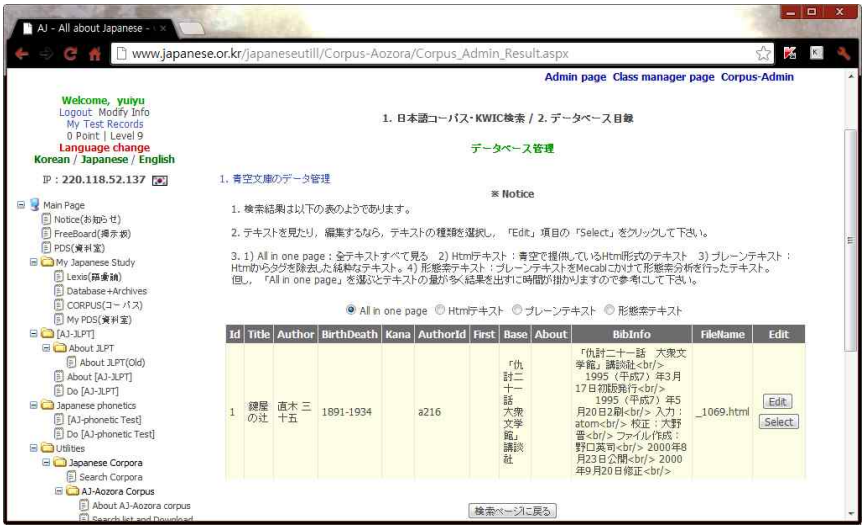

그림 11「青空文庫」데이터베이스 관리자 페이지 – ② 수정 및 추가 등 관리 페이지

e. 데이터베이스 검색 : (5)c 공정을 통해 사용자가 각자 구축한 데이터베이스의 종류에 따라,「MS-Access」프로그램을 사용하여 검색하거나,「MS-SQL」서버에 접속하여 검색.
참고로 저자는「MS-SQL」서버를 기반으로 하는「青空文庫」본문 텍스트 검색 페이지를 개발하고 웹상에 공개. 단, 현재(2012년 09월 18일 현재) 서버의 속도 문제로, 검색결과는 한 작품 당 한 건의 결과만을 제공. 추후, 속도 개선을 통해 전부 제공할 수 있도록 수정할 예정.
1) 검색 :「青空文庫」데이터베이스 전문 검색 페이지. 그림 12참조.
http://www.japanese.or.kr/japaneseutill/Corpus-Aozora/Corpus_TxtDB.aspx

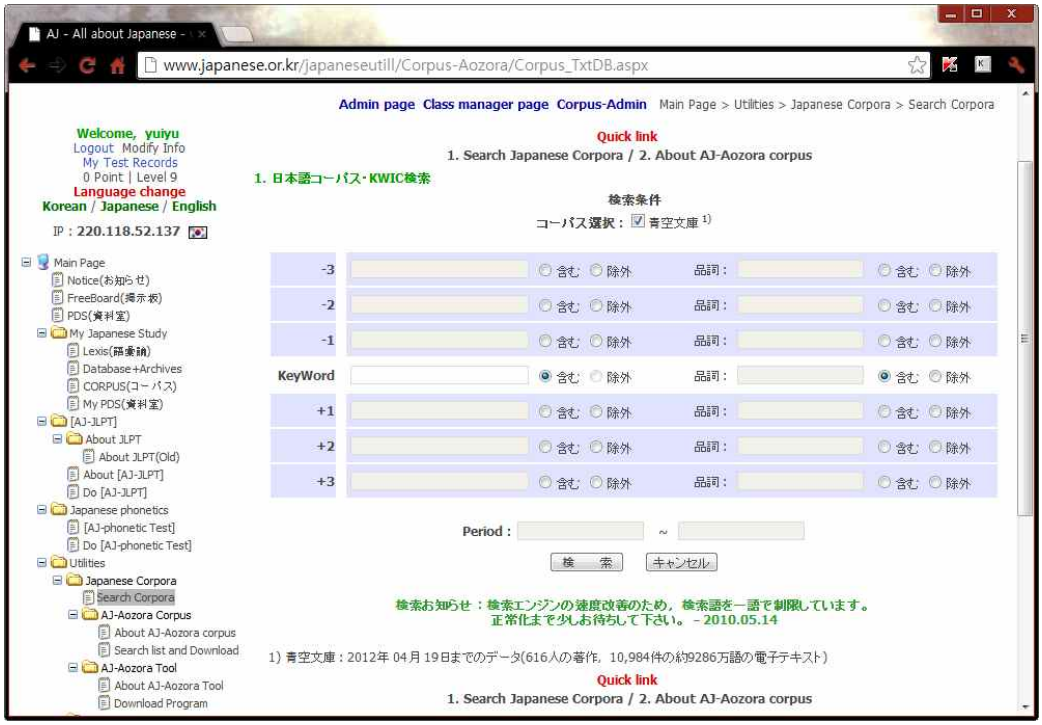

그림 12 「青空文庫」 데이터베이스 전문 검색 페이지

5. 마치면서

이상과 같이 저자가 독자 개발한 「AJ-Aozora-Tool」 을 이용하여 구축한 「青空文庫」 데이터베이스 모델은, 「青空文庫」 의 전자 텍스트를 다운로드하는 것부터 데이터베이스의 구축 및 형태소 분석에 이르기까지 전 과정을 자동으로 수행할 수 있다는 점에서 획기적인 시스템이라고 할 수 있다. 또한 「AJ-Aozora-Tool」 의 형태소 분석에는 「UniDic[12)]」 을 내장한 「MeCab」 라고 하는 오픈소스 프로그램을 탑재하는 것을 통해, 형태소 분석 알고리즘이나 과정을 투명하게 했다는 점도 장점이라고 할 수 있다. 그리고 본고의 대량 텍스트에 대한 형태소 분석 결과는 「MeCab」 의 학습에도 응용할 수 있기 때문에, 금후 보다 정확성이 높은 형태소

12) 「UniDic(유니딕·현대어판)」 : 일본어 텍스트를 단어별로 분할하여, 형태론 정보를 부여하기 위한 전자화사전으로, 형태소분석기 「茶筌(차센·ChaSen)」, 「MeCab(和布蕪·메카부)」 의 사전으로서 이용가능하다. 「UniDic」 의 장점은 다음과 같다.
a. 국립국어연구소가 규정한 「단단위(短單位)」 라고 하는 일정하고 명확한 단위를 그 기반으로 설계.
b. 어휘소(語彙素)·어형(語形)·서자형(書字形)·발음형(發音形)의 단층구조를 갖고 있으며, 표기의 베리에이션이나 어형의 변이에 관계없이 동일한 하나의 표제어를 부여하는 것이 가능.
c. 엑센트나 음편화의 정보를 부여하는 것도 가능하여, 음성처리의 연구에도 이용 가능.
– 「形態素解析辭書UniDic」 에서 발췌. 본고의 저자 번역. http://www.tokuteicorpus.jp/dist

분석 결과를 기대할 수 있다는 측면에서 일본어 연구에 기여하는 바가 크다고 하겠다. 참고로,「青空文庫」에는 현대 일본어 텍스트뿐만 아니라, 근대 일본어의 텍스트자료도 다수 수록하고 있다. 따라서「AJ-Aozora-Tool」에 내장되어있는「MeCab」의 내장 전자사전인「UniDic」을, 금후 선택(e.g.「近代文語UniDic[13)]」)가능할 수 있게 하는 것을 통해, 보다 폭 넓은 시대에 걸친 텍스트를 분석할 수 있게 될 것이다. 또한 이와 같은 시스템은「青空文庫」뿐만 아니라, 이제부터 전자화 하고자 하는 텍스트에도 재 이용가능하다는 점에서도 금후의 데이터베이스의 확장도 기대할 수 있어, 앞으로의 일본어 코퍼스 연구에 기여하는바 또한 크다고 하겠다.

이어서 데이터베이스의 검색과 관리에 관한 내용으로, 본고의 저자는 본 데이터베이스와 검색·관리를 모두「MS-SQL」데이터베이스를 기반으로 한 "웹"상에 공개했다((5)d·e참조). 이는 일본어 연구에 특화된「青空文庫」의 텍스트를 누구라도 보다 간단하고 빠르고 정확하게 검색할 수 있도록 하고자 함이다. 또한 그 검색결과·원문·형태소 분석 등의 모든 정보를 웹에서 간단하게 확인할 수 있도록 하여 본「青空文庫」데이터베이스의 신뢰성을 높였다. 마지막으로, 관리 툴 역시 웹에서 처리할 수 있도록 한 것은, 다수의 관리자에 의한 원격 공동관리가 가능하도록 하고자 함이다(관리 ID부여 및 관리 그룹 지정도 역시 가능). 물론, 이와 같은 웹상에 있어서의 데이터베이스와 그 검색 시스템은 개별 연구자에 의해 특화된 연구를 위해 원본 데이터를 수정하거나, 추가 정보를 부여하는 것이 쉽지 않다는 문제점이 존재한다. 그와 같은 경우, (5)c-1와 같이, 데이터베이스를 구축하는 단계에서「AJ-Aozora-Tool」을 사용하여「青空文庫」를「MS-Access」파일 형식으로 구축하는 것을 통해 해결할 수 있다.「MS-Access」파일로 구축한 데이터베이스의 활용은 각 연구자의 재량 하에 자유롭게 연구를 진행할 수 있기 때문이다. 저자는 본고와 같은 연구를 통해, 앞으로「青空文庫」에 수록되어 있는 전자 텍스트가 일본어 연구에 보다 활발히 이용될 수 있기를 희망하며, 이를 위해 후속 연구로서「AJ-Aozora-Tool」를 수정 및 업그레이드하고, 이를 통해 얻어진 데이터를 적극적으로 공개함과 동시에 사용 매뉴얼에 대한 저작을 공개하고자 한다. 그리고 무엇보다 본고를 통해 정제한 전자 텍스트 파일을 활용하여, 근·현대 일본어에 관한 연구도 함께 진행하고자 한다.

【참고문헌】

青空文庫(2007)『青空文庫10歳記念版「蔵書6300」』DVD1枚, Voyager.

小木曽智信·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.

小木曽智信·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.

小木曽智信·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

13)「근대문어UniDic」：UniDic을 기반으로 근대 문어문을 해석할 수 있도록 제작한 형태소분석 사전으로, 주로 근대의 논설문(메이지보통문(明治普通文))을 그 대상으로 한다. 문학작품이나 타 시대의 텍스트는 반드시 좋은 결과를 얻을 수 없다. MeCab버전과 ChaSen버전으로 공개하고 있으나(Windows용 패키지는 양쪽 사전을 함께 포함), 해석 정확도가 높은 MeCab 버전을 사용하는 것을 권장함.
범례)「形態素解析辭書：近代文語UniDic」에서 발췌. 본고의 저자 번역.
「http://www2.ninjal.ac.jp/lrc」の「形態素解析辭書：近代文語UniDic」메뉴

14回年次大会発表論文集』：pp.225－228.
小木曽智信·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：pp103～110.
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別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：pp176－181.
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)』立命館大学アートリサーチセンター, Vol. 2：pp49～56.
眞島知秀·金暻泳(2003)「일본어 주석 코퍼스(tagged corpus)의 구축 방법에 대하여」『日本學報』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：pp161－172.

【데이터베이스·프로그램·웹페이지】

『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, 博文官新社, 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
『讀賣新聞』(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.fll.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.statice.jp/soft_smoopy.php

< 要 旨 >

日本語研究のための「青空文庫」データベースの構築と活用

－電子テキスト処理プログラム "AJ－Aozora Tool"を活用したデータベース構築モデル開発－

「青空文庫」はウェブを通じて、大量の日本語の電子テキストだけではなく、個々のテキストに関する書誌情報を含む付加情報を共に公開しているインターネット電子テキストアーカイブズであって、様々な時代の数多くの著者のテキストが大規模で収録されている日本語の電子テキストの宝庫である。

本稿ではこのような「青空文庫」を日本語の研究により幅広く、そして効果的に活用する必要があるという判断の上、「青空文庫」を日本語学の研究資料としてより有用に利用する手段として、体系的なデータベース化と共にその管理及び検索ツールの開発という具体的なモデルを提示した。そして、そのようなモデルに基づいて実際的な「青空文庫」のデータベースを構築し、データベース及びテキストの処理ツールを一般に公開した。

本稿におけるデータベースの概略的な構築工程は以下のようである。

1)「青空文庫」の電子テキストデータを一括ダウンロード：「AJ－Aozora－Tool ver1.02」利用

2) 電子テキスト変換及び処理：「AJ－Aozora－Tool ver1.02」利用

2-1) すべてのXHTMLタグテキストデータをプレーンテキストに一括変換

2-2) すべてのプレーンテキストを一括して形態素分析

3) データベース入力：「AJ－Aozora－Tool ver1.02」利用

3-1) プレーンテキストの電子テキスト化：「MS－Access」及び「MS－SQL」

3-2) 形態素分析結果のデータベース化：「MS－Access」及び「MS－SQL」

4) データベース管理：「MS－Access」ファイル或いは「MS－SQL」サーバー内のデータベースとウェブ

5) データベース検索：「MS－Access」ファイル或いはウェブ検索

http://www.japanese.or.kr/japaneseutill/Corpus－Aozora/Corpus_TxtDB.aspx

論 文 分 野 ：コーパス日本語學

キーワード：青空文庫、コーパス、日本語データベース、形態素分析、AJ－Aozora Tool、MeCab

▣ 김유영(金嚅泳)

고려대학교 시간강사

yuiyu@korea.ac.kr

■ 投 稿 日 ： 2012년 0월 00일

■ 審査開始 ： 2012년 0월 00일

■ 審査完了 ： 2012년 0월 00일

■ 掲載確定 ： 2012년 0월 00일