

일본어 연구를 위한 「靑空文庫」데이터베이스 구축 프로그램, 「AJ-Aozora Tool」

金晞泳*

<目 次>

- | | |
|---------|-------------------------|
| 1. 들어가며 | 3. 「AJ-Aozora Tool」의 개발 |
| 2. 문제제기 | 4. 결론 |
-

Key Word : 靑空文庫(Aozora Bunko), Electronic Text, Japanese Corpus, Text Database, AJ-Aozora Tool, Morpheme Analysis

1. 들어가며

최근의 일본어 연구에는 텍스트에 대한 다양한 통계 자료를 제시하여 논증을 전개하는 연구가 적지 않다. 또한 그러한 연구 중 많은 수가 대량의 조사 대상을 보다 빠르고 보다 정확하게 그리고 보다 간단하게 분석하기 위해 전자 텍스트를 이용한 분석을 수행하거나, 조사결과 및 용례를 전자 텍스트화 하여 처리하는 경우가 대부분이다. 이와 같은 흐름을 반영하듯, 자연히 일본어 연구 분야에 있어서 전자 텍스트의 수요가 나날이 커지고 있으나, 실제 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 풀어야 할 과제가 적지 않다.

* 동덕여대 일본어과 조교수, 일본어학(어휘론, 코퍼스 언어학)

이와 같은 현황을 고려할 때, 이미 구축되어 있는 몇몇의 전자 텍스트 중에서도, 대량의 텍스트를 포함하고 있으면서도 더욱이 개개의 텍스트에 관한 서지정보를 포함한 부가정보도 함께 무상으로 공개하고 있는 인터넷 전자 텍스트 아카이브즈, 「靑空文庫(아오조라문고)」(<http://www.aozora.gr.jp>)는 일본어 연구에 있어서 매우 매력적인 일본어 전자 텍스트 자료라고 볼 수 있다. 이에 저자는 본고를 통해, 金鰐泳(2012)의 靑空文庫(아오조라문고) 데이터베이스 구축 모델의 실제적인 수단으로서 자동 프로그램, 「AJ-Aozora Tool」을 개발 및 업그레이드하여, 그 개요와 실제적인 사용방법을 구체적으로 소개하고자 한다. 이를 통해, 보다 많은 연구자들이 靑空文庫(아오조라문고)의 전자 텍스트를 보다 용이하게 접하고, 또한 활용할 수 있을 것으로 기대한다.

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 전자 텍스트 자료는, 다양한 형태(e.g. 플레인 텍스트, 코퍼스 등등, 眞島知秀・金鰐泳(2003) 참조)로 공개 및 교환하는 데에 용이하여, 데이터로서의 재이용성도 매우 높다. 그리고 전자 텍스트를 이용한 연구의 경우, 그 검증 및 재현도 보다 간단하고 정확히 이루어 질 수 있게 된다. 이와 같은 다양한 이점을 고려할 때, 앞으로의 일본어 연구를 위해서는 보다 적극적으로 텍스트의 전자화를 추진하여, 이제까지 전자화 되지 않은 많은 일본어 텍스트의 전자화를 추진해야만 할 것이다. 그러나 텍스트의 전자화는 여러 측면에서 많은 노력이 들 뿐만 아니라, 지금까지 구축된 코퍼스 등 많은 전자 텍스트조차도 저본 및 원본의 저작권 및 편집권 등과 관련된 문제로 인해 모처럼의 시간과 노력을 들여 구축했음에도 불구하고 공개에 제약이 있거나, 공개

되었다라고 해도 그 일부만을 공개할 수밖에 없는 경우가 적지 않은 것이 현실이다.

그런데, 본고에서 다루고자 하는 「靑空文庫」에는 이미 근세 말부터 근현대에 이르기까지의 시기를 중심으로 627명의 저작, 11,182건・약 9,383만어의 전자 텍스트(2012년 8월 14일 현재, 저자 조사)가 수록되어 있으며, 지금 이 시간에도 전자 텍스트는 증가하고 있다¹⁾. 그 뿐만 아니라 「靑空文庫」는, 저작권의 기한이 지난, 혹은 사용허가를 받은 저작만을 수록하고 있기 때문에, 그와 같은 저작권의 문제에서 자유로우며 「靑空文庫」가 밝히고 있는 바와 같이 텍스트에 대한 학문적인 측면에서의 재사용 또한 자유롭다. 즉, 「靑空文庫」는 다양한 시대의 수많은 저자의 텍스트가 대규모로 수록되어 있는 일본어 전자 텍스트의 보고일 뿐만 아니라, 저작권 문제에서 자유롭기 때문에 재이용성・개방성이라는 활용이라는 측면에서도 매우 매력적인 전자 텍스트 자료인 것이다. 따라서 저자는 「靑空文庫」의 전자 텍스트를 일본어 연구에 보다 폭 넓게 그리고 효과적으로 활용 할 필요가 있다고 본다.

하지만, 이와 같은 「靑空文庫」의 전자 텍스트를 손쉽게 읽을 수 있는 다양한 뷰어 등은 개발되어 왔지만, 일본어 연구에 있어서는 폭넓게 활용되지 못하였으며, 몇몇 선행연구에서 그 일부만의 전자 텍스트가 조사 대상으로서 활용되는 것에 그치는 경우가 많았다. 그리고 지금까지 「靑空文庫」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 「靑空文庫」에 수록되어 있는 전자 텍스트를 하나의 데이터베이스로 묶고, 이를 구축・관리・검색이 가능한 공개 코퍼스로 공개한다면, 앞으로 다양한 분야의 일본어 연구에 보다 효과적으로 활용될 수 있을 것으로 판단한다.

이에 본고에서는, 「靑空文庫」에 수록된 전자 텍스트를 체계적인 데이터베이스로서 구축하고 이를 효과적으로 관리하여 실제적인 일본어 연구에 응용할

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

수 있는 검색 툴을 개발한다고 하는 “공개 코퍼스 데이터베이스 개발 모델”(金
 晴泳(2012) 참고)의 일환으로서, 데이터베이스 구축 프로그램 「AJ-Aozora
 Tool」을 수정 및 보완하여, 선행연구인 金晴泳(2012)에서 자세히 언급할 수
 없었던 프로그램의 설치 및 이용에 관한 내용을 보다 자세히 언급하고자 한다.

3. 「AJ-Aozora Tool」의 개발

본 절에서는 「靑空文庫」 데이터베이스화 프로그램 「AJ-Aozora Tool」의 개
 발 방침과 기능에 대해서 소개하고자 한다. 그러나 그에 앞서, 金晴泳(2012)에
 서 제시된 “靑空文庫」공개 코퍼스 데이터베이스 개발 모델”의 방침을 언급하
 자면 다음과 같다.

(1) 「靑空文庫」공개 코퍼스 데이터베이스 개발 모델

- a. 텍스트 데이터와 부가정보를 함께 수록한 데이터베이스를 구축하고, 추후
 에 정보 필드를 확장할 수 있도록 확장성을 강구할 것.
- b. 「靑空文庫」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 강
 구할 것.
- d. 이용자가, 데이터베이스를 간단히 검색하고, 원본 및 부가 정보에 접근할
 수 있으며, 그리고 그 각각의 정보와 갱신 정보를 간단하게 확인할 수 있도
 록 공개할 것.
- e. 이용자가 독자적인 개개의 연구정보를 추가할 수 있는 수단을 강구할 것.

金晴泳(2012:208)

金晴泳(2012)에서는 상기 (1)의 방침에 따라, 「靑空文庫」의 코퍼스를 구축하
 였는데, 이를 구체적으로 언급하면 다음과 같다.

우선, 모든 텍스트 데이터를 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 단순히 플레인 텍스트를 추출하는 것에서 그치는 것이 아니라, 『靑空文庫』의 전자 텍스트에 포함되어 있는 부가정보를 정리하여 마찬가지로 동 데이터베이스에 입력하는 것을 통해, 일본어 연구에 있어서 필요한 정보를 전자 텍스트와 함께 일목요연하게 확인할 수 있도록 했다.

또한 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.



그림 1 『靑空文庫』공개 코퍼스 데이터베이스 개발 로드맵

『AJ-Aozora Tool』은, 『靑空文庫』를 기반으로 하는 공개 코퍼스 데이터베이스를 구축하는 데에 있어서 상기 (1)과 같은 조건을 충족시키면서 위와 같은 구체적인 공정을 시행하기 위한 툴(프로그램)로서 개발되었으며, 전체 로드맵에서는 그림 1의 1·2단계의 공정에 해당한다고 할 수 있다. 즉, 『靑空文庫』의 전자 텍스트를 효과적으로 수집 정제하여, 데이터베이스 형태의 코퍼스로 구축

해 내는 틀인 것이다. 참고로 현재 「Web 검색 시스템 개발」까지 공정이 완료된 상황이며 이와 같은 「靑空文庫」코퍼스의 「Web 검색 시스템」 및 이후 「검색 프로그램 개발」 및 「텍스트 태거 개발」에 관해서는 후속 연구를 통해 공개할 예정이다.

3.1. 「AJ-Aozora Tool」의 개발 방침

1) [데이터베이스화, 전 공정의 자동화]

「靑空文庫」의 전자화된 텍스트 데이터를 일본어 연구에 적용하기 위해 체계적인 코퍼스로 구축하기 위해서는 다음과 같은 작업이 요구된다. 우선, 텍스트를 다운로드하고 이를 정제 및 분류하여 데이터베이스화 하는 동시에, 전자 텍스트에 함께 포함되어 있는 서지정보 등 부가정보도 역시 정제 및 분류하여 데이터베이스화 해야 한다. 저자는 이와 같은 전 과정을 「AJ-Aozora Tool」을 사용하여 자동화할 수 있도록 강구했다.

2) [일본어 연구를 위한 순수한 본문 텍스트, 즉 플레인 텍스트의 가공이 가능할 것]

플레인 텍스트는 말 그대로 순수한 텍스트만을 가리키는 것으로, 품사정보 등 부가적인 언어정보를 포함하고 있지 않기 때문에, 문자열 검색 등 일본어 연구에 필요한 가장 기본적인 용례검색이 가능하다. 그런데, 「靑空文庫」가 제공하는 플레인 텍스트는 직접적인 언어연구의 대상이 되는 본문 텍스트 이외에도, 전자 텍스트화 하는 과정에서 입력자 및 교정자 등의 부가정보도 같은 전자 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 텍스트”를 정제하고 이를 분석해야만 한다.

3) [다양한 연구에 대응할 수 있는 다양한 형태의 전자 텍스트 데이터베이스를 구축할 것]

실제의 일본어 연구에 있어서 전자 텍스트는, 앞서 2)에서 언급한 플레인

텍스트를 시작으로, HTML, XHTML, XML 그리고 데이터베이스 파일 등등, 연구자의 필요나 텍스트 처리기술에 따라 일본어 연구에는 다양한 제 형식의 전자 텍스트가 요구된다. 따라서 본고에서는 『靑空文庫』 데이터베이스를 구축함에 있어서 프로그램의 개발을 통해 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 했다.

4) [구축된 데이터베이스로의 접근성을 강구할 것]

『靑空文庫』의 전자 텍스트를 데이터베이스화 하되, 개개인의 연구에 사용되는 것에 그치지 않고 장래에 타 연구자와의 자료 공유가 가능하도록, 데이터베이스 표준 포맷을 설정하고, 이에 맞추어 데이터베이스를 구축한다. 또한 프로그램을 통해 연구자 스스로 데이터를 정제할 수 있을 뿐만 아니라, 각각의 과정에서 얻을 수 있는 다양한 형식의 전자 텍스트를 연구자가 쉽게 이용할 수 있도록 공개할 수 있는 수단을 강구했다.

5) [서지정보의 보존 및 검색 기능이 가능할 것]

앞서 2)에서 언급한 『靑空文庫』플레인 텍스트에 포함된 부가정보는 전자 텍스트를 일본어 연구 자료로서 사용하는 데에 있어서 결코 불필요한 정보가 아니기 때문에, 코퍼스 데이터베이스에서 전자 텍스트 본문과 함께 확실히 그리고 즉각적으로 확인할 수 있는 형태로 보존해 둘 필요성이 있다. 따라서 단순히 플레인 텍스트만을 수집하여 데이터베이스를 구축하는 것뿐만 아니라, 일본어 연구에 활용 가능한 다양한 정보(eg. 서지정보, 전자화 정보)를 플레인 텍스트와 함께 체계적으로 정리하여 데이터베이스를 구축할 수 있도록 했다.

3.2. 『AJ-Aozora Tool』(ver.1.0.2.1)의 배포 및 인스톨

『AJ-Aozora Tool』은 다음의 (2)a와 같은 웹사이트에서 배포하고 있으며, (2)b의 사양을 충족하는 컴퓨터에서 다운로드하여 설치할 수 있다.

(2) 『AJ-Aozora Tool』- ver.1.0.2.1

a. 배포 웹사이트

-http://www.japanese.or.kr/japaneseutil/AJ-AoZoraTool/Download_AJ-AoZora_Tool.aspx

-<http://code.google.com/p/aj-aozora-tool>

b. 설치사양

-OS ver : MS Windows XP/Vista/7 이상 PC

-하드디스크 용량: 최소 200MB 이상

c. 배포 파일: AJ-Aozora Tool 1.0.2.1.zip

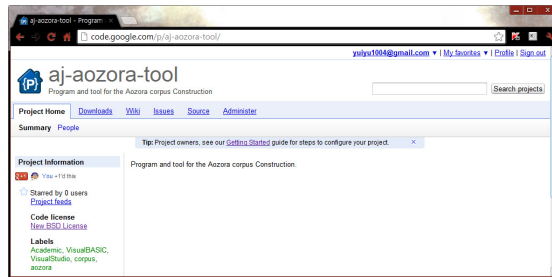
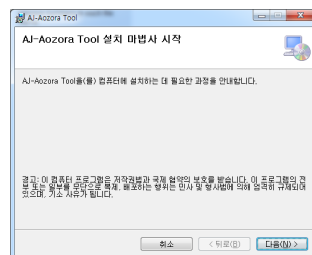


그림 2 「AJ-Aozora-Tool」배포 사이트

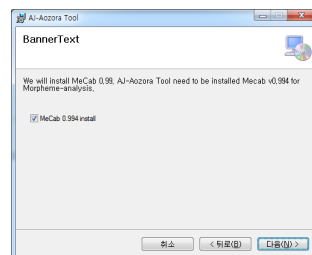
<http://code.google.com/p/aj-aozora-tool>

다운로드 한 (2)c와 같은 압축파일(AJ-Aozora Tool 1.0.2.1.zip)은 압축 해제 프로그램을 이용하여 압축을 푼 후, 더블클릭하여 다음의 표 1과 같은 과정을 통해 사용자의 PC에 인스톨하는 것이 가능하다.

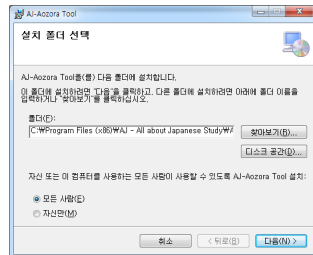
표 1 「AJ-Aozora-Tool」인스톨 과정 (일부 과정 생략)



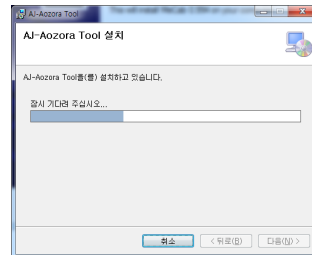
① 인스톨 시작



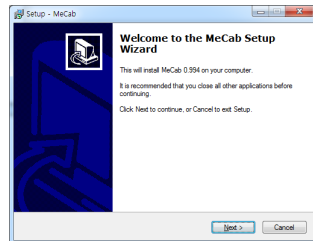
② MeCab install을 반드시 체크



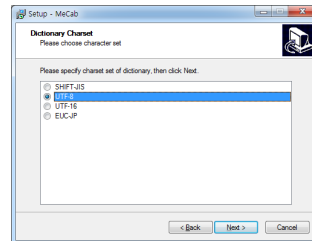
③ 설치 경로 입력



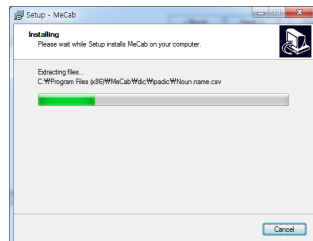
④ 설치 진행



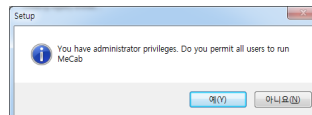
⑤ McCab 설치 자동 진행



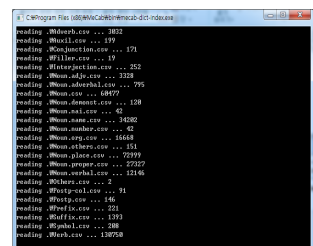
⑥ 한글 원도는 반드시 UTF-8 선택



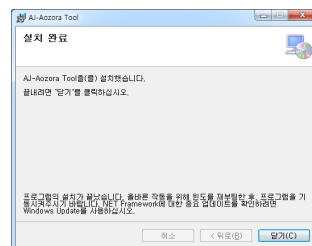
⑦ McCab 설치 진행



⑧ 반드시 Yes를 선택할 것



⑨ Mecab 사전 파일 인코딩



⑩ 설치 완료



⑪ 「AJ-Aozora Tool」바탕화면 아이콘

3.3. 「AJ-Aozora Tool」의 기능

1) [「青空文庫」수록 전자 텍스트 데이터의 일괄 다운로드 기능]

「AJ-Aozora Tool」의 기본적인 기능 중 하나로, 「青空文庫」에 수록되어 있는 모든 전자 텍스트를 일괄 다운로드할 수 있다. 다운받을 수 있는 파일은 XHTML 태그 텍스트 데이터로, 텍스트에는 태그 정보 및 부가 정보가 함께 포함되어 있다.

■ 결과물 : XHTML 태그 텍스트 데이터.

e.g. 236_19996.html : 『ア, 秋』太宰治(1975)

다만, 본 소프트웨어로 인한 전문 텍스트 다운로드로 인해 「青空文庫」의 서버에 과도한 부하를 끼칠 우려가 있기에, 「AJ-Aozora-Tool」의 일반 배포판에는 본 일괄 다운로드 기능은 비활성화 하여 배포하되, 그에 대한 대안으로 「AJ-Aozora-Tool」를 통해 다운 받을 수 있는 모든 텍스트 파일은 다음의 경로에서 다운받을 수 있도록 했다.

■ 「青空文庫」전문 전자 텍스트 다운로드 페이지(그림3 참조)

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

[2. Download AJ-Aozora-Corpus]→[2.1Full XHTML Text]→["AJ-Aozora XHTML 20120814.iso" : Ver.20120814]



그림 3 『靑空文庫』전문 전자 텍스트 다운로드 링크

2) [전자 텍스트 변환 및 처리]

2-1) [XHTML 태그 텍스트 데이터의 플레인 텍스트로 변환 기능]

위 1) 공정을 통해 특정 폴더에 일괄 다운로드 한 XHTML 태그 텍스트 파일들을 텍스트 안의 XHTML 태그뿐만 아니라, 서지정보 등의 부가정보도 분리하여 일본어 연구에 바로 사용할 수 있도록 플레인 텍스트로 일괄 변환할 수 있다. 자세한 프로그램 기동 순서는 다음의 그림 4와 같다.

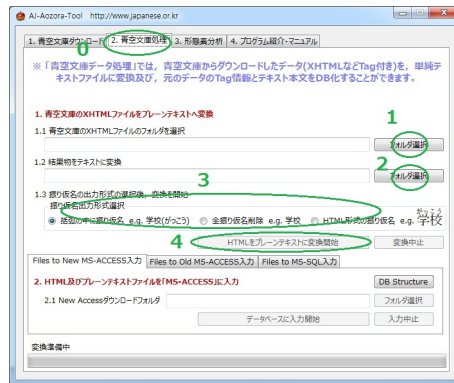


그림 4 HTML 태그 텍스트를 플레인 텍스트로 일괄변환

- ① 青空文庫処理 탭 선택
- ① XHTML 태그 텍스트 파일 폴더 선택
- ② 플레인 텍스트 파일을 저장할 폴더 선택
- ③ 결과물의 종류 선택
- ④ 변환개시

■ 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, 그림 4-3 참조).
e.g. 변환 전 : 236_19996.html ⇒ 변환결과 : 236_19996.txt

2-2) [플레인 텍스트 및 일반 텍스트 파일의 형태소 분석 기능]

위 2-1) 공정에서 얻은 플레인 텍스트(②후리가나振り仮名불포함 플레인
텍스트)를 그림 5와 같은 순서로 「AJ-Aozora-Tool」에 내장된 「MeCab」를 사
용하여 일괄 분석 가능하다.

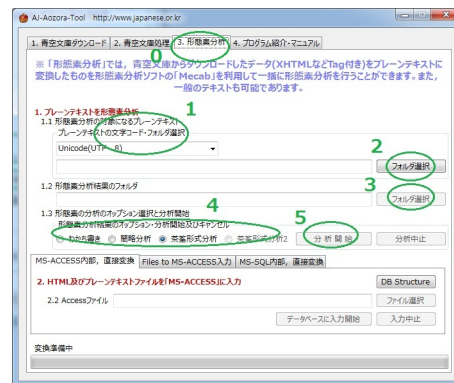


그림 5 플레인 텍스트 및 일반 텍스트 파일의 형태소 분석

- ① 形態素分析 탭 선택
- ① 플레인 텍스트 파일의 문자 코드선택(e.g. Unicode, Shift-JIS코드 등)
- ② 플레인 텍스트 파일이 들어 있는 폴더 선택
- ③ 형태소 분석 결과 파일을 저장할 폴더 선택
- ④ 형태소분석 결과의 형태 선택
- ⑤ 변환개시

■ 결과물 : 형태소 분석 텍스트 파일.

e.g. 변환 전 : 236_19996.txt ⇒ 변환결과 : 236_19996_m.txt

참고로, 본고의 『靑空文庫』의 전자 텍스트뿐만 아니라, 일반 일본어 텍스트 파일도 형태소 분석 및 띄어쓰기 분석 등이 가능하기 때문에, 타 분야의 일본어 학 연구에도 활용할 수 있다.

3) [데이터베이스 입력]

3-1) [플레인 텍스트의 데이터베이스화] - 「MS-Access」파일 및 「MS-SQL」DB

위 1) 공정을 통해 특정 폴더에 일괄 다운로드한 XHTML 태그 텍스트 파일 내부의 본문 플레인 텍스트, 그리고 그와 함께 내장되어 있는 각각의 부가정보를 그림 6과 같이 「MS-Access」 파일 혹은 「MS-SQL」 데이터베이스로 일괄 입력하여 하나의 코퍼스로 구축할 수 있다.

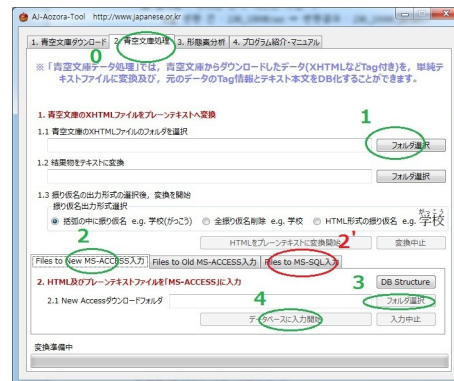


그림 6 靑空文庫 플레인 텍스트의 데이터베이스화

- ① 靑空文庫処理 탭 선택
- ① XHTML 태그 텍스트 파일 폴더 선택
- ② 데이터베이스 형식 선택(②: MS-ACCESS/②: MS-SQL)
- ③ 데이터베이스를 저장할 폴더 선택(②MS-ACCESS의 경우)
- ④ 변환개시

- 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 2·3 및 그림 8 참조

3-2) [형태소 분석 플레인 텍스트의 데이터베이스화]

- 『MS-Access』파일 및 『MS-SQL』DB

상기 2-2)를 통해 특정 폴더에 모은 형태소 분석 텍스트 파일들을 그림 7과 같이 『MS-Access』 파일 혹은 『MS-SQL』 서버의 데이터베이스에 일괄 입력 하여 구축 가능하다.

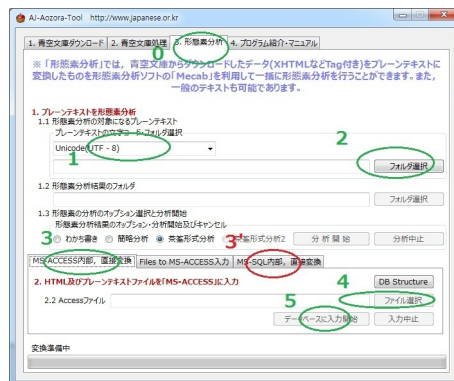


그림 7 青空文庫 플레인 텍스트의 데이터베이스화

- ① 形態素分析 탭 선택
- ① 플레인 텍스트 파일의 문자 코드선택(e.g. Unicode, Shift-JIS코드 등)
- ② 플레인 텍스트 파일이 들어 있는 폴더 선택
- ③ 데이터베이스 형식 선택(③: MS-ACCESS/③': MS-SQL)
- ④ 데이터베이스를 저장할 폴더 선택(③ MS-ACCESS의 경우)
- ⑤ 변환개시

- 결과물 : 『MS-Access』 데이터베이스 파일(『AJ-Aozora-Tool.accdb』 파일 중, 테이블 “Aozora_M”) 혹은 『MS-SQL』서버의 데이터베이스(테이블 “Aozora_M”). 표 3·4 및, 그림 8 참조

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

Table	Aozora		Aozora_P	
Table design	Id	bigint		
	FileName	nvarchar(500)		
	Title	ntext		
	Author	ntext		
	AuthorId	nvarchar(500)	Id	bigint
	BirthDeath	ntext	FileName	nvarchar(500)
	First	ntext	Text_Plane	ntext
	Base	ntext	Text_Plane_Kana	ntext
	Kana	ntext	Etc1	ntext
	About	ntext	Etc2	ntext
	BiblInfo	ntext	Etc3	ntext
	Text	ntext	Etc4	ntext
	Year	nchar(30)	Input_Date	nvarchar(50)
	Etc1	ntext		
	Etc2	ntext		
	Etc3	ntext		
	Etc4	ntext		
	Input_Date	nvarchar(50)		

표 4 「AJ-Aozora-Tool」에 의한 「靑空文庫」데이터베이스 구조 II
(「MS-SQL」의 경우)

Table	Aozora_M		Aozora_Author	
Table design			original_data	nvarchar(255)
			date	datetime
	Id	bigint	ID	nvarchar(255)
	FileName	nvarchar(500)	Name	nvarchar(255)
	Text_Morpheme	ntext	Name2	nvarchar(255)
	Etc1	ntext	[Read]	nvarchar(255)
	Etc2	ntext	Name_Eng	nvarchar(255)
	Etc3	ntext	Postion	nvarchar(255)
	Etc4	ntext	Field	nvarchar(255)
	Input_Dat	nvarchar(50)	Birth	nvarchar(255)
			Death	nvarchar(255)
			History	nvarchar(255)

ID	氏名	生年	没年	主な著作	備考
1.1060	比翼 藤元 利一	1888-1947	『戦城・春は風』	『戦城・春は風』	1969
2.1074	比翼 三子五	1891-1954	『比翼二十歳』	『比翼二十歳』	1969
3.1074	比翼 三子五	1891-1954	『日本現代文壇』	『日本現代文壇』	1969
4.1078	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
5.1081	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
6.1083	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
7.1103	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
8.1115	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
9.1115	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
10.1131	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
11.1255	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
12.1273	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
13.1283	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
14.1305	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969
15.1385	比翼 三子五	1891-1954	『比翼の足音』	『比翼の足音』	1969

그림 8 「MS-Access」로 구축한 「靑空文庫」데이터베이스

4) [그 외, 참고사항]

4-1) 데이터베이스 관리, 배포 및 검색

위와 같은 과정을 통해 구축된 데이터베이스는 「MS-Access」 프로그램을 사용하여 관리하거나, 「MS-SQL」 서버에 접속하여 관리할 수 있다. 본고를 통해 구축한 데이터베이스는 다음과 같이 공개했다.

- (3) a. 「MS-Access」 플레인 텍스트 데이터베이스 및 플레인 텍스트 파일 공개
[2. Download AJ-Aozora-Corpus]→[2.1 Full Plain Text]와 [2.4 Full XHTML + Plain Text of Aozora - MS-Access Ver.]
-http://www.japanese.or.kr/japaneseutil/Corpus-Aozora/About_Corpus_Aozora.aspx
- b. 「MS-Access」 형태소 분석 데이터베이스 및 형태소 분석 텍스트 파일 공개
[2. Download AJ-Aozora-Corpus]→[2.3 Full Morpheme Text of Aozora]와 [2.5 Full Morpheme Text - MS-Access Ver.]
-http://www.japanese.or.kr/japaneseutil/Corpus-Aozora/About_Corpus_Aozora.aspx

그리고 본고를 통해 「MS-SQL」 서버의 데이터베이스를 웹상에서 이를 관리 및 검색할 수 있도록 관리 및 검색 페이지를 구축했다.

- (4) a. 관리 : 관리자 아이디로 접속하여 데이터베이스를 조회하고 정보를 갱신하는 등의 제반 관리기능 개발. 단, 관리자 아이디 필요.
-<http://www.japanese.or.kr/japaneseutil/Corpus-Aozora/Corpus>

_Admin.aspx

b. 검색 : 『靑空文庫』데이터베이스 전문 검색 페이지.

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

4. 결론

본고에서는 인터넷 전자 텍스트 아카이브즈, 『靑空文庫(아오조라문고)』(<http://www.aozora.gr.jp>)를 일본어학의 연구 자료로서 보다 유용하게 이용하기 위해, 『靑空文庫』의 전자 텍스트에 대한 체계적인 데이터베이스화를 실현할 수 있는 자동 프로그램인, 『AJ-Aozora Tool』을 개발하고 본고를 통해 그 개요와 실제적인 방법을 구체적으로 소개했다. 선행연구에서 공개하지 않았으며, 금번 연구를 통해 수정 및 개선된 『AJ-Aozora Tool』의 장점과 기능을 정리하면 다음과 같다.

- 1) 『靑空文庫』에 수록된 모든 전자 텍스트의 일괄 다운로드하고 이 텍스트를 다양한 형태로 자동정제(e.g. 플레인 텍스트, XHTML 텍스트 등)
- 2) 『靑空文庫』의 전자 텍스트 및 일반 전자 텍스트의 형태소 분석
- 3) 다양한 형태로 정제된 『靑空文庫』의 전자 텍스트와 이를 형태소 분석한 텍스트를, 서지정보와 함께 일괄 코퍼스 데이터베이스로 구축(e.g. MS-ACCESS/MS-SQL)

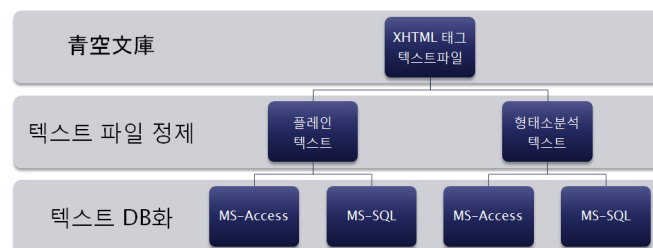


그림 9 『AJ-Aozora Tool』을 통해 얻을 수 있는 결과물

저자는 본 연구를 기반으로 구축된 「青空文庫」 코퍼스를 보다 효과적으로 이용할 수 있도록 하기 위해, 완료된 「Web 검색 시스템」의 공개, 그리고 한발 더 나아가 「검색 프로그램 개발」 「텍스트 태거 개발」과 같은 후속 연구를 금후의 과제로 삼고자 한다.

【參考文獻】

<참고논문>

- 青空文庫(2007) 『青空文庫10歳記念版「蔵書6300」』DVD1枚, Voyager.
- 石井正行(2010) 「招請講演：日本語コーパス言語學の新展開」 『일본언어문화』16, pp.5-21
- 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
- 金嘯泳(2012) 「일본어 연구를 위한 「青空文庫(아오조라문고)」 데이터베이스의 구축과 활용—전자 텍스트 처리 프로그램「AJ-Aozora Tool」을 활용한 데이터베이스 구축 모델 개발—」 『日本語学研究』35輯, 韓国日語日文学会, pp.35-53
- 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」 『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, pp.1-48
- 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, pp.103-110
- 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 靑空文庫』はる書房
 富田倫生(1999) 「<イネーブル・ライブラリー>としての靑空文庫」『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靑空文庫」『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・金嘯泳(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, pp.161-172
 손영석외(2012) 「구어 일본어 형태소 정보부여를 위한형태소분석 프로그램의 활용과
 문제점」『일본언어문화』22, pp.187-205
 김현주(2011) 「자기의식화 선행형 코퍼스, 콘코던스 활용학습에 대한 검토 및 평가 - 일
 본어 어휘, 표현 학습을 중심으로-」『일본언어문화』20, pp.55-78

<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

◇투 고 일 : 2013. 01. 30

◇심사개시 : 2013. 02. 14

◇게재확정 : 2013. 03. 28

〈要旨〉

Development of a Tool for Building Database
of “Aozora Bunko”, 「AJ-Aozora Tool」

“Aozora Bunko” is Japanese electronic text archives via the Internet, and it has mass and diverse electronic text by many authors. For that reason, “Aozora Bunko” is a rich repository of Japanese electronic resources.

The purpose of this paper is to develop and introduce a tool for building database of such an important “Aozora Bunko” in order to use more widely and efficiently for the study of Japanese. The functions and features of 「AJ-Aozora Tool」 is as in the following.

- 1) 「AJ-Aozora Tool」 allows to download all the Japanese electronic text of Aozora Bunko at once, and automatically handle or change all the text you downloaded (e.g. Plain text, XHTML text).
- 2) You can automatically analysis morpheme of all the electronic text of Aozora Bunko as well as other electronic text.
- 3) Moreover it allows you to build your own database of Aozora Bunko using the electronic text you downloaded or analysis morpheme before (e.g. MS-ACCESS/MS-SQL).