2012 June 26

일본어 연구를 위한 "아오조라 문고" 데이터베이스 구축 및 활용
-전자 텍스트 처리 프로그램과 활용 모델 개발-

"AJ-Aozora-DB" 사용자 메뉴얼(한국어판) Ver.0.001

金曦泳(김유영-Yu Young, Kim)
yuiyu@korea.ac.kr・http://www.japanese.or.kr

1. 개요.

본 **"AJ-Aozora-DB(ver.0.201206-1)"**는 2012 년 6 월 16 일 한국일어일문학회 학술발표용으로 제작된 시험판 데이터베이스로 한정된 사용자에게 시험적으로 배포함.

2. "AJ-Aozora-DB"

1) **2012 년 04 월 18 일 현재의 아오조라 문고의 전체 텍스트 데이터 수록**.

1-1) 616 인의 저작 10,984 건, 약 9,286 만어 수록

1-2) 웹을 통해서도 전 목록 검색 가능

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

2) 「AJ-Aozora-Tool」ver1.001[1]에 의해 전자동으로 처리된 데이터 베이스.
AJ-Aozora-Tool 은 추후 논문과 함께 공개, 배포될 예정. 자세한 내용은 발표 레주메 참고.

3) 현재 웹을 통해 본문 텍스트에 대한 부분 검색이 가능. 추후 업데이트를 통해 전문 검색이 가능하도록 하고 속도를 개선할 예정임.
http://www.japanese.or.kr/japaneseutill/Corpus-Aozora/Corpus_TxtDB.aspx

4) 구체적인 데이터베이스의 구조는 다음의 그림 1 과 같다.

[1]「AJ-Aozora-Tool」: 프로그램 언어인「Visual Basic.Net 2010」을 사용하여 독자적으로 개발한 툴로, 데이터베이스는「MQ-SQL2008 이상」「MS-Access2010」에 대응하고 있으며, 금후「MySQL」에도 적용할 수 있도록 개선할 계획임. 버그 해결 및 기능 개선 후, 안정화 작업이 끝나는 대로 금후 프리 소프트웨어로서 공개할 예정. 2012 년 6 월 현재, 버전은 1.001.

[All_Aozora_XHTML] 아오조라문고 원시 데이터 디렉토리

- **[Data]** ⇒ **아오조라에 수록된 원시 XHTML파일**(10,984건) 수록 디렉토리
- **authors.xml** ⇒ 본 AJ-Aozora-DB에 수록된 저작의 독서카드. 단, XML형식.
- **log.txt** ⇒ 일괄 다운로드 기록 로그

[All_Aozora_Plain] 아오조라문고 플레인 텍스트 데이터 디렉토리

- **[All_Aozrora_Plain_NoRuby]** ⇒ 원시 XHTML파일을 가공한 **플레인 텍스트(루비 없음)** 수록(10,983건) 디렉토리
- **[All_Aozrora_Plain_Ruby]** ⇒ 원시 XHTML파일을 가공한 **플레인 텍스트(루비 있음)** 수록(10,983건) 디렉토리
- **[All_Aozrora_Plain_HTML]** ⇒ 원시 XHTML파일을 가공한 **플레인 텍스트(html 태그 있음)** 수록(10,983건) 디렉토리
- **AJ-Aozora_Log.txt** ⇒ 「AJ-Aozora-Tool」에 의한 일괄 가공중의 로그 데이터(누락 및 오류 내용 등)

[All_Aozora_Morp] 아오조라문고 플레인텍스트 형태소분석 결과 디렉토리

- **[All_Aozrora_Morp_Wakachi]** ⇒ **플레인 텍스트(루비 없음)**를 형태소 분석 가공한 **단순 띄어쓰기 텍스트**(10,983건) 수록 디렉토리
- **[All_Aozrora_Morp_Simple]** ⇒ 플레인 텍스트(루비 없음)를 형태소 분석 가공한 **단순 형태소 분석 텍스트**(10,983건) 수록 디렉토리
- **[All_Aozrora_Morp_ChasenStyle]** ⇒ 플레인 텍스트(루비 없음)를 형태소 분석 가공한 **차센(Chasen) 형식 텍스트**(10,983건) 수록 디렉토리

[All_Aozora_AccessDB] 아오조라 문고 MS-ACCESS 데이터베이스 디렉토리

- **AJ-Aozora-Tool.accdb** ⇒ 상기 아오조라문고의 **XHTML 파일** 및 이를 가공한 **플레인 텍스트(루비 없음)**와 **플레인 텍스트(루비 있음)**를 해당 저작의 **서지정보**(타이틀, 저본, 저자, 저자생몰 정보, 저자 생애, 입력 일자 등등)와 함께 DB화한 파일
- **AJ-Aozora_Log.txt** ⇒ 「AJ-Aozora-Tool」에 의한 일괄 가공중의 로그 데이터(누락 및 오류 내용 등)

AJ-Aozora-DB-Manual.docx 본 메뉴얼 문서(Root 디렉토리에 위치)

그림 1 AJ-Aozora-DB(ver.0.201206-1) 의 기본 구조

범례) [ ]안은 디렉토리(e.g. [Data]), 그 이외는 일반 확장자를 가진 파일(e.g. log.txt).

5) [All_Aozora_XHTML]의 [Data] 디렉토리에 수록된 아오조라문고의 원시 XHTML 파일 내용의 예시

아오조라에 수록된 원시 XHTML 파일 예시 - _348.html

```
<?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" >
<!-- XHTML? by AOMAME -->
<head>
<title>文七元結</title>
<meta HTTP-EQUIV="Content-Type" content="text/html; CHARSET='x-sjis'"/>
        <meta name="著者" content="三遊亭 円朝" />
        <meta name="著者 ID" content="a989" />
        <meta name="生没年" content="1839-1900" />
        <meta name="仮名遣い" content="" />
</head>
```

```
<body bgcolor="#FEFEFE">
<body>
<h1>文七元結</h1>
<h2>三遊亭圓朝</h2>
<h3>鈴木行三校訂編纂</h3>
<br/>
<br/>
<br/>
                    一<br/>
<br/>
　さてお短いもので、<ruby><rb>文七元結</rb><rt>ぶんしちもとゆい</rt></ruby>の由来という、ちとお古い処のお話を申上げますが、只今と徳川家時分とは余程様子の違いました事で、昔は遊び人というものがございましたが、只遊んで暮して居ります。よく遊んで喰って<ruby><rb>往</rb><rt>ゆ</rt></ruby>かれたものでございます。<ruby><rb>何</rb><rt>ど</rt></ruby>うして遊んでて暮しがついたものか

                                    ……
```

6) [All_Aozora_XHTML] 디렉토리에 수록된 XML 형식의 독서카드 예시.

XML 형식의 독서카드 예시 -渡辺 温

```
<著者>
        <ID>a20</ID>
        <氏名>渡辺 温</氏名>
        <異表記></異表記>
        <読み>わたなべ おん</読み>
        <ローマ字>Watanabe, On</ローマ字>
        <所属></所属>
        <分野></分野>
        <生年>1902-08-26</生年>
        <没年>1930-02-10</没年>
        <略歴>1902（明治 35）年 8 月 26 日、北海道生れ。1924（大正 13）年、慶応義塾大在学中に、プラトン社の映画筋書懸賞募集に「影」で一等入選。1927（昭和 2）年、博文館に入社し、横溝正史編集長のもとで雑誌「新青年」のモダニズム化を推進する。作品はほとんどが掌篇で、主に「新青年」に発表された。1930（昭和 5）年、谷崎潤一郎への原稿依頼の帰途、交通事故で死亡。なお、探偵作家の渡辺啓助は一歳年上の実兄、作品集『アンドロギュノスの裔』の挿絵を担当した渡辺東は姪にあたる。（森下祐行）「渡辺温」</略歴>
</著者>

                    ……
```

7) [All_Aozora_Plain] 디렉토리에 수록된 플레인 텍스트(루비 있음)의 예시

플레인 텍스트(루비 있음)의 예시 - _348.txt

```
文七元結
三遊亭 円朝
```

三遊亭圓朝
鈴木行三校訂編纂

一

さてお短いもので、文七元結(ぶんしちもとゆい)の由来という、ちとお古い処のお話を申上げますが、只今と徳川家時分とは余程様子の違いました事で、昔は遊び人というものがございましたが、只遊んで暮して居ります。よく遊んで喰って往(ゆ)かれたものでございます。何(ど)うして遊んでて暮しがついたものかというと、天下御禁制の事を致しました。只今ではお厳(やかま)しい事でございまして、中々隠れて致す事も出来んほどお厳しいかと思いますと、麗々と看板を掛けまして、何か火入れの賽(さい)がぶら下って、花牌(はなふだ)が並んで出ています、これを買って店頭(みせさき)で公然(おもてむき)に致しておりましても、楽(たのし)みを妨げる訳はないから、少しもお咎(とが)めはない事で、隠れて致し、金を賭(か)けて大きな事をなさり、金は沢山あるが退屈で仕方がない、負けても勝っても何うでも宜(よ)いと、退屈しのぎにあれをして遊んで暮そうという身分のお方には宜(よろ)しゅうございますが、其の日暮しの者で、自分が働きに出なければ、喰う事が出来ないような者がやりますと、自然商売が疎(おろそか)になります。慾徳ずくゆえ、倦(あ)きが来ませんから勝負を致し、今日で三日続けて商売に出ないなどということで、何うも障(さわ)りになりますから、厳(やかま)しゅう仰(おっ)しゃる訳で、併(しか)し賭博(ばくち)を致しましたり、酒を飲んで怠惰者(なまけもの)で仕方がないというような者は、何うかすると良い職人などにあるもので、仕事を精出して為(し)さえすれば、大して金が取れて立派に暮しの出来る人だが、惜(おし)い事には怠惰者だと云うは腕の好(よ)い人にございますもので、本所(ほんじょ)の達磨横町(だるまよこちょう)に左官の長兵衛(ちょうべえ)という人がございまして、二人前(ふたりまえ)の仕事を致し、早くって手際が好くって、塵際(ちりぎわ)などもすっきりして、落雁肌(らくがんはだ)にむらのないように塗る左官は少ないもので、戸前口(とまえぐち)をこの人が塗れば、必ず火の這入(はい)るような事はないというので、何(ど)んな職人が蔵を拵(こしら)えましても、戸前口だけは長兵衛さんに頼むというほど腕は良いが、誠に怠惰(なまけ)ものでございます。昔は、賭博に負けると裸体(はだか)で歩いたもので、只今はお厳(やかま)しいから裸体どころか股引も脱(と)る事が出来ませんけれども、其の頃は素裸体(すっぱだか)で、赤合羽(あかがっぱ)などを着て、「昨夜(ゆうべ)はからどうもすっぱり剥(むか)れた」と自慢に為(し)ているとは馬鹿気た事でございます。今長兵衛は着物まで取られてしまい、仕方なく十一になる女の子の半纏(はんてん)を借りて着たが、余程短く、下帯の結び目が出ていますが、平気な顔をして日暮にぼんやり我家(わがや)へ帰って参り、

……

8) [All_Aozora_Morp] 디렉토리에 수록된 "단순 형태소 분석 텍스트"의 예시

플레인 텍스트(루비 없음)을 단순 형태소 분석한 예시 - _348_m.txt

```
文	名詞-一般
七	名詞-数
元結	名詞-一般

三	名詞-数
円朝	名詞-固有名詞-人名-名


三遊亭圓朝	名詞-固有名詞-人名-一般

鈴木	名詞-固有名詞-人名-姓
行	名詞-固有名詞-人名-名
三	名詞-数
校訂	名詞-サ変接続
編纂	名詞-サ変接続
```

```
　　記号-空白
　　記号-空白
　　記号-空白
　　記号-空白
　　記号-空白
　　記号-空白
　　記号-空白
　　記号-空白
一　　名詞-数


　　記号-空白
さて　　接続詞
お　　接頭詞-名詞接続
短い　　形容詞-自立
もの　　名詞-非自立-一般
で　　助動詞
、　　記号-読点
文　　名詞-一般

……
```

3. "AJ-Aozora-DB(ver.0.201206-1)"의 이용상 주의사항

1) 본 데이터베이스는 본격적인 공개를 앞두고, 몇몇 사용자들에게 배포되는 시험판 DB 입니다. 따라서 재배포를 권장하지 않습니다. 발표의 논문화가 끝나면 추후 온전히 공개 할 예정입니다.

2) 현재 모든 데이터에 대한 전수검사를 실시하고 있습니다만, 시간 및 인력의 문제로 인해 몇몇 오류가 있는 텍스트가 있을 수 있음을 양해 바랍니다.

3) 본 데이터베이스 자동 구축 프로그램인 「AJ-Aozora-Tool」은, 추후 논문 공개 및 웹 검색의 정상화 후 홈페이지(http://www.japanese.or.kr)를 통해 공개 및 배포하도록 하겠습니다. 조금만 기다려 주세요.

4) 혹, 연구에 사용하시게 된다면 출처를 꼭 밝혀 주시기 바랍니다.

5) 상업적 목적에의 무단 사용을 금합니다.

6) 오류 보고 및 문의사항, 공동연구, 제언 등을 환영합니다.
연락처 : yuiyu@korea.ac.kr

2012 June 26