자동 형태소 분석 기술을 이용한 한국어 읽기 보조 도구의 개발

- 일본어 모어화자를 위한 기능을 중심으로 -

스가이 요시노리(須賀井 義教)
(긴키대학)

Yoshinori Sugai. 2013. The Development of a Reading Support Tool for the Korean Language Using Automatic Morphological Analysis Technology. *Journal of Korean Language Education* 24-3: 139-159. The purpose of this paper is to give an outline of a web-based reading support tool which provides valuable information for learners of the Korean language, by using the automatic morphological analysis technology. This tool processes Korean texts input by a user and provides information such as: 1) word frequency lists according to the part of speech, 2) learning level of words, 3) Hanja notation. These kinds of information will be helpful for learners to encourage both bottom-up and top-down processes in reading comprehension.

In this tool, an open source software "MeCab" is used as the morphological analysis engine, with a dictionary which has approximately 60,000 entries.

The reading support tool developed in this study is available on the internet and anyone can use it via a web browser. It is expected that learner autonomy can be fostered by providing the tool through the internet. However, there are some problems to be solved in the future as follows: a) shortage of dictionary entries, b) additional functions necessary for learners. (Kinki University)

주제어: 읽기 교육(reading education), 웹 기반 한국어 교재(web-based Korean learning materials), 상향식 처리 과정(bottom-up model), 하향식 처리 과정(top-down model), 형태소 분석 (morphological analysis), 한자 표기(Hanja notation)

1. 들어가기

본고[1]는 현대 한국어의 자동 형태소 분석 기술을 이용하여 필자가 개발한 읽기 보조 도구 “한국어 독해 보조 툴”에 대해 그 개발 과정과 기능을 소개하는 것을 목적으로 한다. “한국어 독해 보조 툴”의 소개와 함께 형태소 분석을 비롯한 자연언어처리 기술이 한국어 교육에서 어떻게 이용될 수 있는지에 대해서도 논의하고자 한다.

현재 많은 “웹 교재”(인터넷에서 공개, 운용되고 있는 웹 기반 한국어 교재)가 인터넷을 통해 제공되고 있는데, 그 중에 몇 가지를 소개하면 국립국제교육원이 제공하는 “KOSNET”,[2] 서강대학교 한국어교육원이 제공하는 “서강대학교 한국어 강좌”,[3] 그리고 서울대학교 언어교육원이 제공하는 “온라인 한국어”[4] 등이 있다. 그리고 한국 이외의 지역에서 제공되고 있는 웹 교재로 일본 후쿠오카[福岡]대학에서 제작한 “한글마당”[5] 등이 있다. 최근에는 사이버 대학이 제작한 웹 교재들도 등장하여 다양한 웹 교재가 제공되고 있다.[6] 이들 한국어 웹 교재는 대부분 단계적으로 학습할 수 있도록 설계되어 있고, 또 하이퍼링크나 다양한 미디어를 이용하는 등 책 형식으로 만들어진 교재와는 많은 면에서 차이를 보인다. 다만 이들 교재는 그 내용이 고정되어 있기에 그것을 이용하는 교사나 학습자가 마음대로 내용을 추

---

1) 본고는 日本學術振興會 科學研究費補助金 基盤研究(B)(2010-2012年度, 課題番號: 22320115), 若手研究(B)(2013年度-, 課題番號: 25770164)에 의한 연구 성과의 일부이며 제54회 朝鮮語教育研究會(2012년 6월 17일, キャンパスプラザ京都)에서 발표한 내용을 수정, 보완한 것이다. 원고 작성 과정에서 많은 도움을 주신 목종균 선생님(긴키대학), 그리고 귀중한 지적을 해 주신 익명의 심사위원 선생님들께 감사를 드린다.

2) http://www.kosnet.go.kr/

3) http://korean.sogang.ac.kr/

4) https://lei.snu.ac.kr/site/kr/klec/click-korean/index.jsp

5) http://ccc.cis.fukuoka-u.ac.jp/~user03/index.html

6) 한국어 웹 교재에 대해서는 방성원(2008) 등을 참조할 것.

가하거나 수정할 수가 없다는 단점도 존재한다.

이러한 웹 교재 이외에도 다양한 보조 교재나 도구들이 인터넷을 통해 제공되고 있다. 그 중에서 서울대학교 IDS 연구실이 제공하는 "꼬꼬마 세종 말뭉치 활용 시스템"[7]은 그 대표적인 예로, 21세기 세종계획에서 구축된 말뭉치 데이터를 바탕으로 용례 검색을 하거나 통계 정보를 얻을 수 있다. 그리고 함께 제공하고 있는 '한국어 쓰기 연습'은 한국어 교육이라는 측면에서 볼 때 주목할 만하다. '한국어 쓰기 연습'은 이용자가 입력한 문장을 분석하여 '양식'을 추출하거나 그 '양식'이 쓰이는 용례를 말뭉치에서 찾아주는 등, "주로 학문적 목적으로 한국어를 공부하는 외국인 및 한국어 교육자에게 도움을 주기 위해 개발된 기능"('한국어 쓰기 학습' 기능 사용 설명서에서 인용)을 가지고 있다. 또한 구문 분석 기능도 있으며 그 분석 결과를 수형도로 표시해 주기도 하는데 이 시스템이 "학문적 목적"을 염두에 두고 있기 때문에 제공되는 정보들은 사실 초/중급 학습자가 직접 활용하기에는 어려움이 수반할 수 있다.

보조 교재라는 관점에서 유타니[油谷幸利]가 개발한 웹 사전(유타니[油谷幸利] 2008)이나 오나[大名力]가 개발한 듣기 자동 채점 시스템(오나[大名力] 2007) 등의 도구들도 주목할 만하다. 이러한 보조 도구들은 학습자의 외국어 학습을 돕기 위해 개발되었으나 교사가 교재 작성 시에 유용하게 이용할 수 있는 보조 도구라고도 할 수 있다. 이러한 보조 도구들은 완성된 교재와는 달리 학습자에게 일정한 틀을 제공하여 그로부터 이용자가 학습에 필요한 정보를 얻을 수 있도록 제작되었다.

다음으로는 일본어 교육에서의 성과물에 대해 잠시 살펴보도록 한다. 가와무라[川村よし子](2009)가 개발한 "리딩 추타"(リーディングチュウ太)[8]도 이러한 보조 도구의 특성을 가지고 있다. "리딩 추타"는 이용자가 일본어 문장을 입력하면 그 문장에 대한 형태소 분석 결과를 표시해 준다. 그리고 분

7) http://kkma.snu.ac.kr/
8) http://language.tiu.ac.jp/

석 결과를 바탕으로 사전을 검색해 주고 어휘의 학습 수준도 표시해 준다. 학습자가 읽고 싶은 문장을 입력하면 사전을 찾아서 단어 뜻을 표시해 주는 것이다. 또 가와무라[川村よし子](2009)에 따르면 교사가 읽기 자료를 작성할 때에도 이 도구를 활용할 수 있다고 한다.

가와무라, 오나, 유타니 등의 보조 도구, 그리고 앞에서 언급한 "꼬꼬마 세종 말뭉치 활용 시스템"은 완성된 교재와 같이 단계별로 학습 내용을 제공하는 것이 아니라 학습자나 교사가 필요한 내용을 입력하여 그 입력 내용에 대해 유용한 정보를 얻을 수 있다는 것이 큰 공통점이라 할 수 있다. 일정한 학습 항목을 단계적으로 학습해 나가는 교재도 중요하지만 학습자가 원하는 정보를 필요에 따라 제공해 주는 보조 도구 또한 외국어 학습에 있어 중요하며 앞으로 그 기대는 더욱 커질 것으로 예상된다.

위에서 소개한 보조 도구들은 모두 인터넷에 공개된 것으로 학습자가 때나 장소를 가리지 않고 이용할 수 있다는 장점이 있다. 그런 점에서 학습자의 자율적인 학습을 촉진시켜 주는 교육 방법의 하나로 주목받고 있다. 본고에서 소개하고자 하는 "한국어 독해 보조 툴" 또한 이러한 관점에서 개발되었음을 밝히며 가와무라[川村よし子](2009) 등을 참고로 읽기 보조 도구로서 어떻게 쓰일 수 있는지 그 가능성을 검토하도록 하겠다.

2. 읽기 교육과 컴퓨터 기술

여기서는 학습자의 읽기 활동에서 컴퓨터 기술을 어떻게 활용할 수 있을지 생각해 보고자 한다.

2.1 읽기 교육에서의 컴퓨터 기술 이용

읽기 교육뿐만 아니라 한국어 교육 전반에 걸쳐 컴퓨터나 인터넷 기술을 어떻게 이용할 수 있는지에 대해 많은 논의가 있었는데, 그 예료서 멀티미

디어 자료를 이용한 학습, 쌍방향적인 커뮤니케이션이 가능한 메일이나 채팅을 통한 학습 등을 들 수 있다(김중섭 2001, 방성원 2008 등).

또한 컴퓨터를 이용함으로써 학습자 개인이 개별적인 학습이 가능하게 된다. 김중섭(2001:205), 요시다[吉田晴世]외 편저(2008:19) 등에서 지적되었듯이 컴퓨터를 이용하게 되면 학습자가 학습 속도 등을 조절할 수 있게 된다. 따라서 학습자는 스스로 자신에 맞는 수준과 속도로 학습을 진행할 수 있다. 결과적으로 학습자가 스스로 학습 과정을 관리하고 조절할 수 있는 자율적인 학습이 이루어질 것으로 기대된다.

그리고 이러한 자율적인 학습을 고려할 때, 읽는 글의 내용은 학습자의 학습 의욕에 큰 영향을 미칠 것이다.[9)]

이러한 점을 감안하여 필자는 학습자가 읽고 싶은 내용을 자유롭게 입력할 수 있고, 그 내용을 분석하여 다양한 정보를 표시해 주는 읽기 보조 도구를 설계하였다. 특히 학습자가 교실 밖에서 수행하는 읽기 활동 지원을 목적으로 인터넷을 통해 제공한다. 자율적인 학습을 위해서는 때나 장소를 가리지 않고 이용할 수 있다는 것이 필수적인 여건이라 할 수 있다. 인터넷에 접속할 수 있는 PC나 스마트폰만 있으면 이용이 가능하도록 설계하였다.

2.2 '읽기'의 세 가지 모델

다음으로 실제 읽기 활동에서 보조 도구를 어떻게 이용할 수 있는지 검토해 보도록 한다. 검토에 앞서 먼저 '읽기'라는 활동에 대해 잠시 살펴보고자 한다.

읽기 과정에는 다음 세 가지 모델이 있다고 한다(국제교류기금[國際交流基金] 2006:5, 한재영 외 2005:223-224 등):

9) 김중섭(2001:197-198)에서 읽기 수업에 대한 중국인 학습자의 설문 조사 결과가 소개되어 있어 참고가 된다. 설문 조사의 결과, 학습자들은 읽기 능력 향상에 도움이 되는 읽기 수업으로 "다양한 텍스트 사용"이라는 의견을 제시하였고, 또 신문, 수필, 소설 등을 선호하는 것으로 대답하였다고 한다.

(1) a. 상향식 처리 과정
b. 하향식 처리 과정
c. 상호적인 처리 과정

먼저 (1a)의 '상향식 처리 과정'이란 텍스트에 나타난 단어나 문법 항목과 같은 언어적인 정보를 이용하여 단어로부터 절, 문장, 글 전체로 이해해 나가는 읽기 모형이다. 이 읽기 모형은 "쓰인 의미를 회복하려는 작업으로서 누가 해도 같은 결과를 얻을 수 있을 것"(오카자키[岡崎眸] · 오카자키[岡崎敏雄] 2001:29-30)이라는 점에서 학습자의 활동이 수동적일 수밖에 없을 것으로 추정된다.

한편 (1b)의 '하향식 처리 과정'은 텍스트 내의 특정 언어 정보를 힌트로 학습자가 예측, 추측을 한 뒤, 자신의 추측이 맞는지 확인하면서 텍스트를 읽어 나가는 읽기 모델이다. 예측할 때에는 학습자가 가지고 있는 배경 지식 등이 동원된다. (1a)의 '상향식 처리 과정'과는 달리 학습자가 텍스트에 더 주체적으로 관여하는 읽기 방식으로, '상향식 처리 과정'에 대한 비판과 함께 주목을 받고 있는 읽기 모델이다.

마지막으로 (1c)의 '상호적인 처리 과정'이란 "하향적인 모델과 마찬가지로 학습자 위주로 정보를 처리하는 것"으로 "텍스트에 대한 추측을 언어 정보에 근거하여 확인하고 다시 추측하고 다시 언어 정보를 확인하는 과정을 반복"하는 처리 과정이다(한재영 외 2005:224). 실제로 초급 학습자들은 어휘나 문법이 부족해서 '하향식 처리 과정'이 적절하지 않은 경우가 있다. 또 하향적인 정보 처리 과정에서도 결국에는 상향적인 정보 처리가 행해지기 때문에 읽기 과정에서 상향적인 처리와 하향적인 처리가 순환적으로 이루어진다. 한국어 교육에서도 이 모델에 근거한 읽기 수업이 제안되기도 하였다.[10]

위의 세 가지 모델 중 본고에서 소개하는 보조 도구는 주로 상향식 처리 과정에서 도움을 줄 수 있을 것이다. 입력된 문장을 분석하여 각 단어의 품

10) 권미정(1999), 김현진(2005) 등을 들 수 있다.

사나 어휘 학습 수준을 표시해 줌으로써 학습자가 문장의 의미를 파악하거나 난이도를 짐작하는 데 도움이 될 것으로 예상된다. 또한 인터넷 상의 웹 사전으로 연결해 주면 단어의 의미 찾기가 더 쉬워질 것이다.

또 형태소 분석 처리를 함으로써 글에 나타난 단어의 리스트를 만들고 그 출현 빈도를 계산할 수도 있다. 이러한 처리를 통해 글의 키워드 추출이 가능하게 되는데, 이와 같은 기능은 학습자가 글의 내용을 추측하는 데 도움을 줄 수 있을 것으로 생각된다. 즉 하향식 처리 과정에서 유용한 정보를 제공하는 기능이라 하겠다.

이러한 점을 고려하여 보조 도구의 기능을 설계하였다. 각 기능에 대해서는 4장에서 자세히 설명될 것이다. 다음 3장에서는 본고에서 이용한 형태소 분석 프로그램에 대해 간단히 소개하고자 한다.

3. MeCab(메카브)란

본고에서 소개하는 “한국어 독해 보조 툴”에서는 형태소 분석 엔진으로서 오픈소스 소프트웨어인 “MeCab”(めかぶ, 메카브)를 이용하였다. MeCab는 京都大學情報學研究科-日本電信電話株式會社커뮤니케이션科學基礎研究所 共同研究유닛 프로젝트를 통해 개발된 것으로 최신 버전은 0.996이다. 홈페이지[11]에서 프로그램과 일본어 분석을 위한 사전을 얻을 수 있다. MeCab의 특징 중 본고와 관련되는 것은 다음과 같다.

(2) a. 특정 문법에 의존하지 않기 때문에 형태소 분석을 위한 사전만 만들면 어떤 언어도 분석할 수 있다
b. Java나 Perl 등의 프로그래밍 언어를 통해 이용할 수 있다
c. 문자 코드로 UTF-8을 지원한다

읽기 보조 도구 개발과 관련해 (2a)가 특히 중요한 특징이라 할 수 있다.

11) http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

종래의 한국어 형태소 분석기들, 예를 들어 21세기 세종계획에서 개발된 “지능형 형태소 분석기”와 같은 경우에는 형태소 분석을 위한 사전 데이터(이하 ‘분석 사전’)가 분석 처리를 수행하는 프로그램(이하 ‘분석 프로그램’)에 내포되어 있어 이용자가 새 단어나 정보를 추가, 수정할 수가 없다. 그러나 MeCab는 분석 프로그램과 분석 사전이 따로 분리되어 있어 분석 사전을 새로 만들면 일본어 이외의 언어를 분석할 수 있다.[12] 또 분석 사전에 이용자가 항목을 추가하여 사전을 다시 구축할 수도 있다. 분석 사전에서는 문자 코드로서 UTF-8을 지원하기 때문에 한글이나 한자 등 다양한 문자 집합(character set)을 포함시킬 수 있다(2c).

또 (2b)와 같이 Java나 Perl 등의 프로그래밍 언어로부터 MeCab를 이용할 수 있기 때문에 이용자가 자기 컴퓨터에 MeCab를 설치할 필요 없이 웹 브라우저를 통해 분석 결과만 얻을 수 있는 시스템을 구축할 수 있다.

위와 같은 여러 사항을 고려한 결과, 형태소 분석 엔진으로 MeCab를 이용하여 읽기 보조 도구를 개발하는 것이 여러모로 유용하다고 판단하였다. 그리고 Perl/CGI를 이용해서 형태소 분석 결과를 활용할 수 있는 읽기 보조 도구 개발에 대해서는 4장에서 소개하도록 하겠다.

4. 보조 도구 개발의 실제

4.1 분석 사전의 규모

본고에서 이용한 분석 사전은 스가이[須賀井義教](2013a, 2013b)에서 개발

12) 일본어 이외의 언어를 분석하는 데 MeCab를 이용한 예로서는 고전 중국어를 대상으로 한 모리오카[守岡知彦](2008), 15세기 한국어를 대상으로 한 무라타[村田寛](2010), 스가이[須賀井義教]·무라타[村田寛](2011), 그리고 현대 한국어를 대상으로 한 스가이[須賀井義教](2013b)가 있다. 또한 MeCab는 범용적인 텍스트 변환 도구로 이용할 수도 있는데, 필자는 그 기능을 응용하여 히라가나를 한글로 轉寫해 주는 도구를 개발하여 공개한 바 있다.
http://porocise.sakura.ne.jp/korean/hira2han/

한 사전이며 현재 항목 수는 총 60,177개이다.[13] 사전 구축에 있어서는 "한국어 학습용 어휘 선정 결과"(조남호 2003)[14]를 기본 데이터로 삼고 "현대국어 사용 빈도 조사"(조남호 2002)의 일부 항목을 추가하였다.[15] 또 인터넷상의 자료를 이용하여 서울시 행정 구역명이나 지하철 역명, 나라와 수도 등의 지명, 한국인의 성(姓) 등을 임의로 추가하였다.

MeCab용 분석 사전을 구축할 때 학습 데이터가 필요한데, 본고에서는 문어를 중심으로 1000개 문장[16]을 이용해 분석 사전을 구축하였다.[17] 분석 사전에는 표제항과 함께 다음 정보들, 즉 소성(素性)이 등록되어 있다.

(3) 품사1, 품사2, 품사3, 활용형, 접속 정보, 사전 항목, 표층형, 한자, 비고, 학습수준

(3)의 '품사1'은 동사, 형용사 등 큰 분류를 가리키며, '품사2'와 '품사3'은 각각 그 하위분류를 가리킨다. 품사 분류는 용언을 제외하면 대부분 "표준국어대사전"을 따랐다. 용언으로서 동사, 형용사, 지정사(指定詞: '-이다', '아니다'), 존재사(存在詞: '있다', '없다' 등)를 인정한다.

'활용형'은 용언의 활용 어기(語基)[18]를 구분하여 표시하도록 하였으며 '접

13) 분석 사전에 관한 자세한 정보는 필자 홈페이지도 참조하기 바란다. 분석 사전은 향후 공개할 예정이다. http://porocise.sakura.ne.jp/wiki/korean/mecab.ko

14) 전체 항목 수는 5,965개인데, 분류불명 항목 등 일부를 제외하였다.

15) 모두 국립국어원 홈페이지 "자료실"에서 Excel 형식의 데이터를 다운 받아 이용하였음을 밝혀 둔다.

16) 학습 데이터로서는 21세기 세종계획의 성과물 중 "문어"의 "원시 말뭉치"에서 임의로 10개 파일을 뽑아 이용하였다. 각 파일 앞 부분에서 머리말이나 제목을 제외한 100개 문장을 분석한 것이다.

17) 분석 사전 구축에 관해서는 MeCab 홈페이지나 스가이[須賀井義教]·무라타[村田寛](2011), 스가이[須賀井義教](2013b) 등을 참조할 것.

18) 본고에서는 용언 활용을 기술하는 데 "어기"(語基) 개념을 이용하였다. MeCab는 입력된 문자열(文字列)을 그대로 처리하기 때문에 축약, 생략된 문자를 복원하지 못한다. 예를 들어 '기다려요'를 '기다리-'와 '-어요'로 분석하기 어렵다. 그러나 어기 개념을 이용하면 '기다려-'(第Ⅲ語基)와 '-요'(어미)로 분석할 수 있기

속 정보'는 용언 어미가 어떤 어기에 접속하는 것인지를 표시하는 것이다.

'사전 항목'은 사전 표제항으로 등록된 형태[19]를 입력한 것이다. 그리고 '표층형'은 '사전 항목'과 달리 실제로 나타난 형태를 입력한 것인데, 예를 들어 아래 (4)의 첫째 줄 '지금'의 경우는 '사전 항목'이 '지금03', '표층형'이 '지금'과 같이 달리 표기된다.

'한자'는 한자어 등을 한자로 표시한 것으로 어휘의 일부만 한자로 표기할 수 있는 경우, 예를 들어 '결코'는 '決코'와 같이 입력하였다. '비고'는 기본 데이터에 기술되어 있던 부가적인 정보를 그대로 기술하였다.

마지막으로 '학습 수준'은 "한국어 학습용 어휘 선정 결과 보고서"에 표시된 세 단계의 학습 수준을 입력하여 표시할 수 있도록 하였다.

이렇게 구축된 분석 사전으로 "지금 바로 이것들을 생활 속에서 실천해 보는 것은 어떨까?"(서강대학교 한국어교육원 2009:64)라는 문장을 분석하면 다음과 같은 출력 결과를 얻을 수 있다.[20]

(4)
```
ㅈㅣㄱㅡㅁ          Adverb,一般,名詞可能,*,*,지금03,지금,只今,*,A
ㅂㅏㄹㅗ            Adverb,一般,*,*,*,바로02,바로,*,*,A
ㅇㅣㄱㅓㅅ          Noun,代名詞,*,*,*,이것,이것,*,*,A
ㄷㅡㄹ              Suffix,名詞派生,*,*,*,들09,들,*,우리들,*
ㅇㅡㄹ              Ending,助詞,對格,*,*,을,을,*,*,*
ㅅㅐㅇㅎㅘㄹ        Noun,普通,*,*,*,생활,생활,生活,*,A
ㅅㅗㄱ              Noun,普通,*,*,*,속01,속,*,*,A
ㅇㅔㅅㅓ            Ending,助詞,處格,*,*,에서,에서,*,*,*
ㅅㅣㄹㅊㅓㄴㅎㅐ    Verb,自立,*,語基3,*,실천하다01,실천해,實踐ㅎㅐ,*,C
ㅂㅗ                Verb,非自立,*,語基1,3接續,보다01,보,*,*,A
```

때문에 MeCab의 처리 방식에 맞출 수 있다. 그리고 필요에 따라 MeCab의 출력 결과를 다시 처리하여 '기다리-'와 '-어요'로 바꾸든지 원하는 형식으로 변환할 수도 있다. 어기 개념에 대해서는 간노[菅野裕臣](1997) 등을 참조.

19) 기본적으로 "한국어 학습용 어휘 선정 결과 보고서", "현대 국어 사용 빈도 조사"를 따른 것이지만 일부 "표준국어대사전"에 따라 동음이의어 번호를 수정하였다.

20) 예에서 '*'로 표시된 부분은 해당되는 정보가 공백임을 나타낸다.

```
ㄴㅡㄴ          Ending,語尾,連体形,*,1接續,는,는,*,*,*
ㄱㅓㅅ          Noun,不完全名詞,*,*,*,것01,것,*,*,A
ㅇㅡㄴ          Ending,助詞,題目,*,*,은,은,*,*,*
ㅇㅕㄸㅓ        Adjective,自立,ㅎ変則,語基2,*,어떻다,어떠,*,*,A
ㄹㄲㅏ          Ending,語尾,終止形,*,2接續,ㄹ까,ㄹ까,*,*,*
?               Symbol,疑問符,*,*,*,?,?,*,*,*
EOS
```

(4)와 같이 한 형태에 한 줄씩 분석 결과가 출력되는데 분석된 항목과 탭(tab) 문자에 이어 그 항목의 소성(素性)이 출력된다. 표제항들은 모두 풀어쓰기 형식으로 기술하였는데 이는 '-ㄴ다'처럼 자모를 단위로 분석해야 추출이 가능한 항목들이 있기 때문이다.

현재 분석 사전에는 (4)와 같은 정보들이 포함되어 있는데 필요에 따라 다른 정보들, 예를 들어 "한국어 능력 시험"(TOPIK)이나 "<한글> 능력 검정 시험"[21] 등의 어휘 수준과 같은 다양한 정보들을 추가하여 사전을 업그레이드할 수도 있다.

본고에서 사용한 분석 사전의 분석률[22]은 아래의 <표 1>과 같다. <표 1>의 결과는 한국어 교재 "서강한국어" 5A(6과 읽기I)와 5B(3과 읽기I)의 '읽어 봅시다' 본문 95개 문장을 분석한 것(이하 "한국어 교재"라 함)과 21세기 세종계획에서 구축한 문어 원시 말뭉치 중 "소설 창작 강의"('BRHO0402.txt'. 이하 "세종 말뭉치"라 함)의 앞 부분 50개 문장을 분석한 것이다.

---

21) 일본에서 시행되고 있는 한국어 능력 평가이다. 봄과 가을에 실시되는 시험으로 2013년 봄에 제40회 시험이 시행되었다. 시험에 관한 정보는 한글 능력 검정 협회 홈페이지(http://www.hangul.or.jp/)에서 얻을 수 있다.

22) 여기서 '분석률'은 '재현율'과 '정확률'의 평균치를 가리키며 모두 MeCab와 함께 배포되는 mecab-system-eval 스크립트를 이용하여 산출하였다. '재현율'은 분석 대상의 전체 형태소 중에서 분석된 형태소가 차지하는 비율이며 '정확률'은 분석된 형태소 중에서 정확하게 분석된 형태소가 차지하는 비율이다.

<표 1> 한국어 교재와 세종 말뭉치의 분석률

	형태소 경계	품사1	전체
한국어 교재	99.0514	98.5022	96.8547
세종 말뭉치	95.9538	95.5925	93.1358

“한국어 교재”에 대해서는 사전 미등록 항목이 1개만 포함되어 있는데 형태소 경계에 관해서는 99% 정도의 높은 분석률을 보이고 있다. 한편 미등록 항목 18개를 포함한 세종 말뭉치의 경우, 형태 경계의 분석률이 95% 정도, ‘사전 항목’ 즉 동음이의어 판별까지 포함한 전체 분석률은 93% 정도로 약간 떨어진다. <표 1>의 결과로부터 위에서 언급한 보조 도구 개발은 물론 분석률을 높이기 위한 사전 개발도 진행해야 할 과제라 할 수 있다.

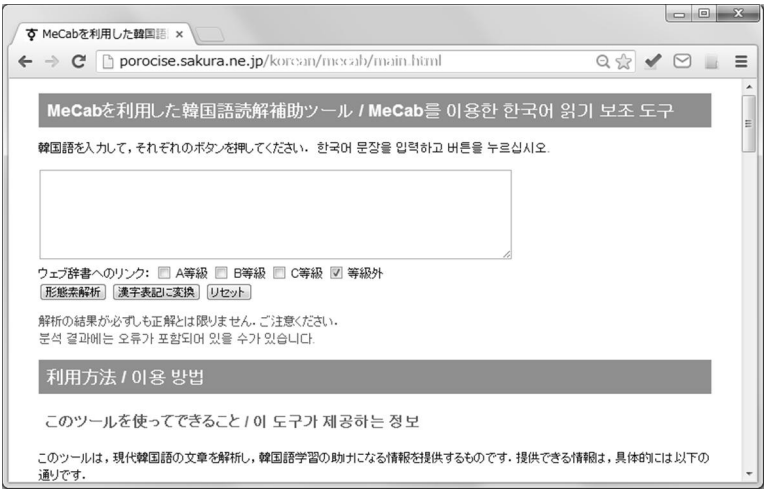
본고에서 소개하는 보조 도구는 프로그래밍 언어 Perl을 이용하여 인터넷을 통해 학습자가 쉽게 접속할 수 있도록 개발하였다. 웹 서버에서 움직이는 MeCab로 분석 처리를 하고 그 출력 결과를 Perl로 가공하여 이용자의 웹 브라우저에 표시한다.[23]

4.2 읽기 보조 도구의 기능

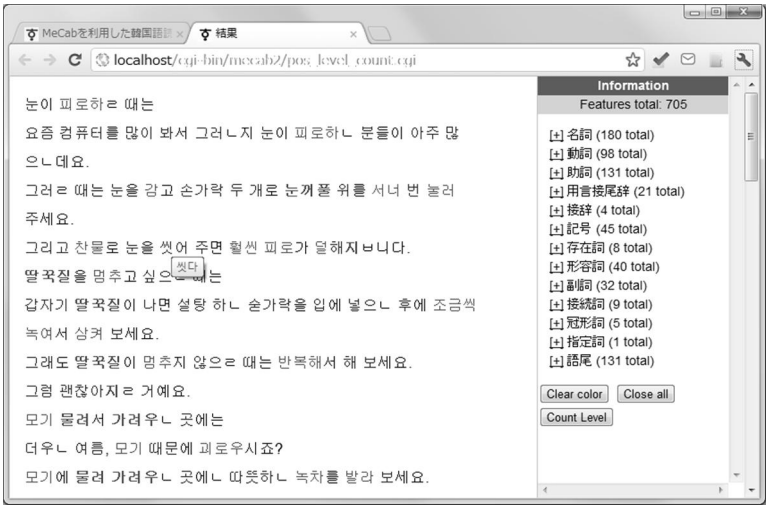
필자가 개발한 보조 도구에는 입력된 글에 대해 (a) 형태소 분석 결과를 표시, (b) 한자어를 한자 표기로 변환, (c) 어휘 학습 수준을 표시, 등의 기능이 있다.

보조 도구의 초기 화면은 <그림 1>과 같다. 화면 중앙에 있는 입력란에 한국어 문장을 입력하고 “形態素解析”(형태소분석) 혹은 “漢字表記に變換”(한자 표기로 변환) 버튼을 누르면 새로 창이 열려 결과가 표시된다(<그림 2> 참조).

23) 필자 홈페이지(http://porocise.sakura.ne.jp/korean/mecab/main.html)에서 공개하고 있다. 웹 서버에는 MeCab가 이미 설치되어 있었고 그 버전은 0.97이다. 또 Perl을 통해 MeCab를 이용하기 위한 Perl 모듈 ‘Text:MeCab-0.20013’을 필자가 따로 설치하였다.



<그림 1> 보조 도구 초기 화면

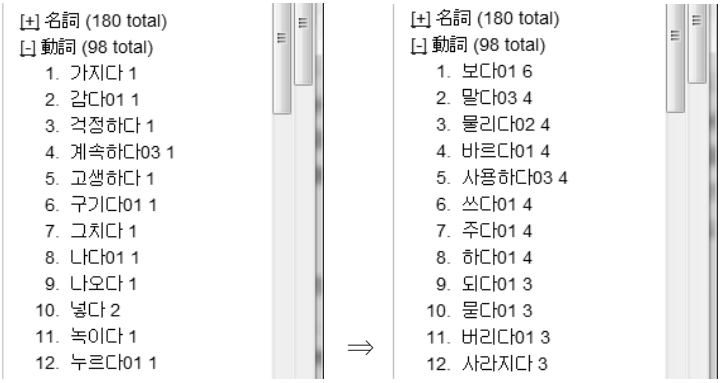


<그림 2> 결과 출력 화면

화면의 오른쪽에 있는 “Information”란(이하 “정보창”이라 함)에는 보조적 정보, 예를 들어 품사별 어휘 빈도수나 한자어 빈도수 등이 표시된다.

4.2.1 형태소 분석 결과와 어휘 학습 수준 표시

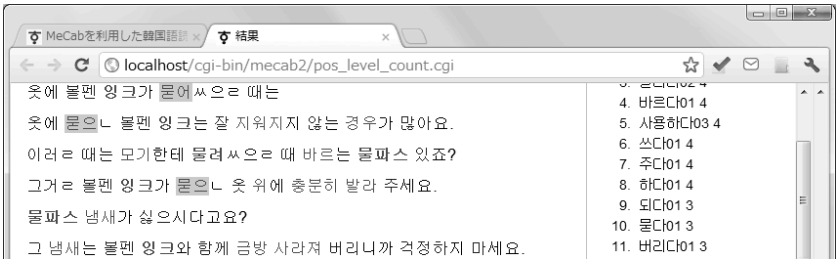
이 기능은 MeCab로 형태소 분석을 하여 분석 사전에 등록된 어휘의 학습 수준에 따라 색으로 구분하여 결과를 표시해 주는 것이다(위의 <그림 2> 참조). 정보창에는 품사별로 어휘 빈도수가 표시되며 또 품사명 옆에 있는 "[+]"를 클릭하면 어휘별 빈도수가 표시된다(<그림 3> 참조). 그리고 품사명을 클릭하면 빈도의 내림순으로 정렬할 수가 있다. 그 결과 글의 핵심어 찾기가 쉬워질 것으로 예상된다.



<그림 3> 품사별 어휘 리스트 정렬 후의 리스트

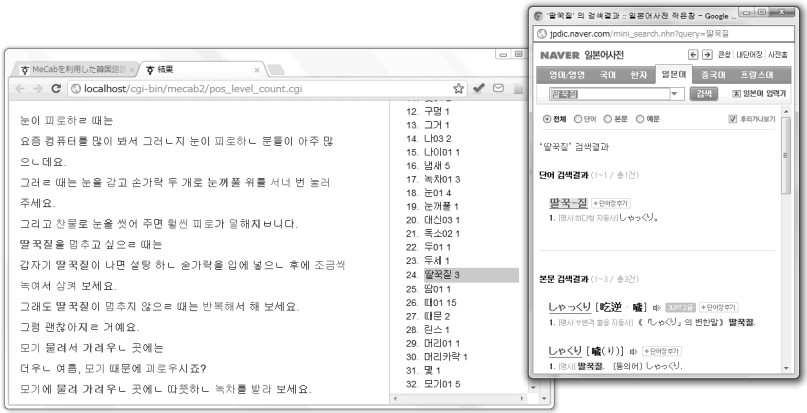
어휘의 학습 수준 표시는 "한국어 학습용 어휘 선정 결과"를 따랐다. 예를 들면 학습 수준이 A등급인 어휘는 파란색, B등급은 녹색, C등급은 빨간색으로 표시되며 등급 외의 어휘(이하 "D등급"이라 함)는 빨간색의 굵은 글씨로 표시된다.

또 정보창에 나타난 어휘 중 임의의 어휘를 클릭하면 왼쪽의 본문에서 해당 어휘가 강조 표시된다(<그림 4>는 '묻다01'을 클릭한 후의 상태).



<그림 4> 정보창에서 '묻다01'을 클릭한 후의 화면

반대로 왼쪽 본문에서 임의의 단어를 클릭하면 따로 창이 열려서 웹 사전으로 해당 항목이 표시된다(<그림 5>는 '딸꾹질'을 클릭한 후의 상태).

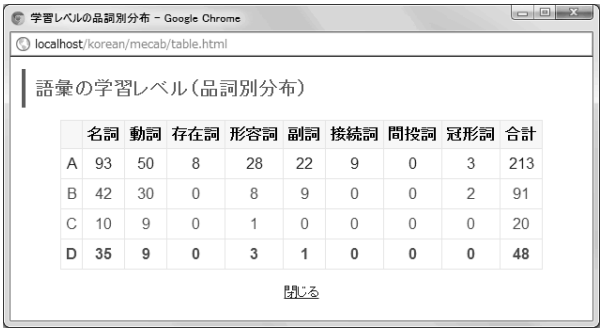


<그림 5> 본문 부분에서 '딸꾹질'을 클릭한 후의 화면

현재는 '네이버 일본어 사전'24)을 참조하도록 설정하였는데, 클릭한 어휘에 대해 그 품사를 알 수 있도록 오른쪽에 있는 정보창에 강조 표시를 해 준다.

그리고 정보창 아래 부분에 있는 'Count Level' 버튼을 누르면 새로 창이 열리며 <그림 6>과 같이 입력된 글 전체의 품사별, 학습 수준 별로 빈도수를 표시해 준다.

24) http://jpdic.naver.com/



	名詞	動詞	存在詞	形容詞	副詞	接続詞	問投詞	冠形詞	合計
A	93	50	8	28	22	9	0	3	213
B	42	30	0	8	9	0	0	2	91
C	10	9	0	1	0	0	0	0	20
D	**35**	**9**	**0**	**3**	**1**	**0**	**0**	**0**	**48**

<그림 6> 품사별/학습 수준별 빈도수

<그림 6>은 “서강한국어” 5B 제3과의 읽기 본문을 분석한 결과이다. 대체로 A등급이나 B등급의 어휘가 사용되었는데 “D등급”, 즉 등급 외의 어휘도 적지 않게 포함되어 있음을 알 수 있다. 이러한 빈도수 표시는 글의 난이도를 대충 알 수 있다는 점에서 학습자에게 큰 도움이 될 것으로 예상된다. 그뿐만 아니라 읽기 교재 작성 시 이와 같은 기능을 이용하면 글의 난이도를 체크할 수 있어 교사에게도 도움이 되리라 믿는다.

위의 통계에는 조사나 어미는 포함되지 않는다. 그리고 이 기능은 글 전체의 난이도를 표시하는 것이 아니라 글에 사용된 어휘의 학습 수준을 표시하는 것이기에 앞으로 글의 난이도를 표시할 수 있도록 개량할 예정이다.

4.2.2 한자 표기로 변환

한국어 어휘 중에 한자어가 많이 포함되어 있다는 것은 일본어 모어화자에게는 한국어 학습에 있어 큰 장점으로 작용된다. 그러나 실제로는 대부분이 한글로만 표기되어 있어 큰 도움을 받지 못하고 있는 것이 실정이다. 물론 한국어와 일본어의 한자어가 완전히 일치되는 것은 아니지만 한자어가 한자로 표기된다면 일본어 모어화자에게는 큰 도움이 될 것으로 예상된다. 이런 점을 감안하여 이 보조 도구는 입력된 문장 중에 한자 표기가 가능한 항목이

있으면 그것을 모두 한자로 변환하여 표시해 주는 기능을 제공하였다. 이러한 기능은 종래의 형태소 분석기에는 거의 볼 수 없었던 것으로, 분석 사전을 임의로 구축해 이용할 수 있는 MeCab로만 제공 가능한 것이다. 이런 점에서 이 보조 도구만이 제공할 수 있는 특징적인 기능이라 할 수 있다.

보조 도구 초기 화면(앞의 <그림 1> 참조)에서 한국어 문장을 입력하여 "한자 표기로 변환" 버튼을 누르면 새 창에 그 결과가 표시된다.

<그림 7> 한자 표기 변환 결과

<그림 7>과 같이 오른쪽 정보창에는 한자어를 추출하여 각각의 빈도수를 제시해 준다. 그러나 한자 표기에 오류도 있다. 예를 들어 <그림 7>의 넷째 줄의 "日帝 時代"가 "一齊 時代"로 표기되었다. 이것은 형태소 분석 결과에 오류가 있기 때문에 한자 표기에도 오류가 생긴 것이다.

여기서 소개한 읽기 보조 도구는 문법성을 판단하는 것이 아니므로 입력된 문장이 문법적인지 아닌지를 확인할 수는 없다. 그러나 한국어 교사의

입장에서 보면 이 읽기 보조 도구를 활용하여 자신이 만든 읽기 교재의 문장 난이도를 알아볼 수 있다. 그뿐만 아니라 분석 결과를 참고로 어려운 어휘를 더 쉬운 것으로 바꾸거나 삭제한다든지 학습자 수준에 맞는 교재를 개발할 수도 있다. 그리고 학습자 입장에서는 이 도구를 이용함으로써 사전을 찾는 시간을 줄일 수 있기에 그만큼 내용 파악에 더 시간을 투자할 수 있을 것으로 판단된다.

5. 나가기

본고에서는 자동 형태소 분석 기술을 이용한 한국어 읽기 보조 도구 “한국어 독해 보조 툴”에 대해 소개하였다. 본고에서 다룬 읽기 보조 도구는 형태소 분석 결과는 물론 한자 표기로 변환하거나 학습 수준까지 표시할 수 있다. 학습 수준에 따라 인터넷 상의 웹 사전을 참조할 수 있도록 링크 기능도 부여하였다. 한자 표기로 변환하는 기능은 종래의 형태소 분석기나 보조 도구에 없었던 것으로 본고에서 다룬 보조 도구가 제공하는 독자적인 기능이다.

마지막으로 향후 검토해야 할 과제는 간단하게 언급하도록 하겠다.

(5) a. 형태소 분석의 분석률 향상
b. 기능 추가

(5a)에 관해서는 4.2.2절에서도 언급하였듯이 약 60,000 개 정도의 항목은 분석 사전으로서는 미흡하다고 할 수밖에 없다. 입력 내용에 미등록 항목이 포함되어 있을 경우, 당연히 형태소 분석의 분석률이 낮아진다. 따라서 분석률을 향상시키기 위해 충분한 항목을 분석 사전에 추가해야 할 것이다. 또한 사전 구축 시에 필요한 학습 데이터도 늘려야 할 것이다.

(5b)에 대해서도 더 많은 기능을 추가할 필요가 있다고 판단된다.[25] 이 점

25) 이 점에 관해서 익명의 심사자가 지적한 바와 같이 보조 도구를 이용하는 데 다

에 대해서는 학습자가 읽기 활동을 수행할 때 어떠한 도움이 필요한지, 그리고 학습자에게 어떠한 정보가 학습에 도움이 되는지 등, 다각적인 학습자 요구 분석을 통해 추가 기능을 검토해 가고자 한다.

본고에서 소개한 보조 도구는 그 사용 용도가 읽기 활동 지원에 국한되어 있으나 형태소 분석과 같은 자연언어처리 기술을 한국어 교육에 실질적으로 적용한 하나의 예로서 큰 의의가 있다. 구문 분석이나 음성 인식을 비롯한 기타 자연언어처리 기술을 이용하여 앞으로 더 다양한 웹 교재, 보조 도구가 개발되기를 기대한다.

참고문헌

강승식(2002;2003), 한국어 형태소 분석과 정보 검색(수정판), 홍릉과학출판사.
권미정(1999), “외국어로서의 한국어 읽기 교육: 독해 전략을 통한 효율적인 읽기 방안”, 한국어 교육, 10권 1호, 국제한국어교육학회, 1-28쪽.
김중섭(2004), 한국어 교육의 의해, 한국문화사.
김현진(2005), “읽기 교육의 교수 학습: 스키마 활성화를 통한 효과적인 읽기 활동 방안”, 한국어교육론 3, 한국문화사, 127-144쪽.
방성원(2008), “웹 기반 한국어 교재”, 한국어 교재 연구, 도서출판 하우, 59-80쪽.
서강대학교 한국어교육원(2009), 서강한국어 Student’s Book 5B 읽기 · 말하기, 서강대학교 국제문화교육원 출판.
안은희(2006), “다언어 학습 시스템을 통한 웹 기반 한국어 교육 프로그램”, 한국어 교육, 17권 2호, 국제한국어교육학회, 157-181쪽.
유혜원(2004), 한국어 정보 처리의 이론과 실제, 제이앤씨.
이동주 외(2010), “꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구”, 정보과학회논문지, 16권 11호, 한국정보과학회, 1046-1050쪽.
조남호(2002), 현대 국어 사용 빈도 조사: 한국어 학습용 어휘 선정을 위한 기초 조

소의 불편함이 있거나 기능상 미흡한 부분이 아직 많다. 필자 개인이 개발하고 있고 아직 완성된 상태가 아니기 때문이다. 그러나 이미 완성된 상태로 제공되는 제품이나 서비스와는 달리 개인이 개발하고 있기 때문에 이용자의 요구를 바로 반영시킬 수 있다는 장점도 있다. 물론 기술적인 한계는 있겠으나 앞으로 이용자의 의견을 모아서 도구를 개선해 나갈 예정이다.

사(국립국어연구원 2002-1-17), 국립국어연구원.
조남호(2003), 한국어 학습용 어휘 선정 결과 보고서(국립국어연구원 2003-1-4), 국립국어연구원.
진기호(2005), “읽기 교육의 과제와 발전 방향”, 한국어교육론 3, 한국문화사, 145-156쪽.
한재영 외(2005), 한국어 교수법, 태학사.
황화상(2006), 한국어와 정보, 박이정.
가와무라[川村よし子](2009), チュウ太の虎の巻: 日本語教育のためのインターネット活用術, 東京: くろしお出版.
간노[菅野裕臣](1997), “朝鮮語の語基について”, 日本語と外國語との對照研究IV 日本語と朝鮮語 下巻 研究論文編, 東京: くろしお出版, 1-21쪽.
국제교류기금[國際交流基金](2006), 讀むことを教える(國際交流基金 日本語教授法シリーズ 第7巻), 東京: ひつじ書房.
모리오카[守岡知彦](2008), “MeCabを用いた古典中國語の形態素解析の試み”, 情報處理學會研究報告[人文科學とコンピュータ], 2008-CH-73, 東京: 情報處理學會, 17-22쪽.
무라타[村田寬](2010), “15世紀朝鮮語の形態素解析の試み: MeCabを利用して”, 福岡大學研究部論集A: 人文科學編, Vol.10 No.3, 福岡: 福岡大學, 17-28쪽.
스가이[須賀井義教](2013a), ‘MeCab用韓國語形態素解析辭書の構築’, 言語處理學會第19回年次大會 發表論文集 781-784쪽.
스가이[須賀井義教](2013b), “MeCabを用いた現代韓國語の形態素解析”, 朝鮮語研究, 5, 東京: ひつじ書房, 283-312쪽.
스가이[須賀井義教]·무라타[村田寬](2011), “15世紀朝鮮語の形態素解析について”, 教養·外國語教育センター紀要, 第1巻 第2号, 東大阪: 近畿大學教養·外國語教育センター, 41-56쪽.
오나[大名力](2007), “ウェブを利用した韓國語ディクテーション自動採点システム”, 外國語教育メディア學會中部支部研究紀要, 18号, 愛知: 外國語教育メディア學會中部支部, 11-20쪽.
오카자키[岡崎眸](1996), “讀み方の指導: ボトムアップ的讀みから相互交流的讀みへ”, お茶の水女子大學人文科學紀要, 第49卷, 東京: お茶の水女子大學, 205-218쪽.
오카자키[岡崎眸]·오카자키[岡崎敏雄](2001), 日本語教育における學習の分析とデザイン: 言語習得過程の視点から見た日本語教育, 東京: 凡人社.
요시다[吉田晴世] 외 편저(2008), ICTを活用した外國語教育, 東京: 東京電氣大學出版局.
유타니[油谷幸利](2008), “朝鮮語Web辭典の設計”, 朝鮮學報, 第206輯, 天理: 朝鮮學會, (1)-(37)쪽.

이케다[池田伸子](2003), CALL導入と開發と實踐: 日本語教育でのコンピュータの活用, 東京: くろしお出版.

스가이 요시노리(須賀井 義教)
日本 大阪府東大阪市小若江3-4-1
近畿大學 綜合社會學部
577-8502
전화번호: +81-6-4307-4186
전자우편: sugaiy@kindai.ac.jp

접수일자: 2013. 6. 15
심사일자: 2013. 7. 10
게재확정: 2013. 8. 23