국어 어휘의 역사 검색 프로그램 개발

두길수*°, 황호전**, 김법균**, 안동언**, 정성종**, 이신원***
*서남대학교 전기전자공학부, **전북대학교 전자정보공학부, ***전북과학대학 컴퓨터정보계열
dgs@tiger.seonam.ac.kr, {hjhwang, kyun}@duan.chonbuk.ac.kr,
{duan, sjchung}@moak.chonbuk.ac.kr, swlee9237@hanmail.net

A Development of Retrieval Program for Korean Vocabulary History

*Gil-Su Doo°, **Ho-Jeon Hwang, **Beob-Kyun Kim, **Dong-Un An, **Sung-Jong Chung, ***Shin-Won Lee
*Faculty of Electric and Electronic Engineering, Seonam University,
**Department of Computer Engineering, ChonBuk National University,
***Department of Computer Information, JeonBuk Science College

요약

"국어 어휘의 역사 검색 프로그램 개발"은 국민들에게 국어 어휘에 대한 역사 정보를 제공하여 국어에 대한 지식과 이해의 폭을 넓힘으로써 국어 생활을 더욱 윤택하게 하고, 국어의 정확한 사용을 통하여 국민들의 의사소통의 정확성과 신속성을 제고함을 그 목적으로 한다. 아울러 이러한 국어 어휘 역사에 대한 정보를 컴퓨터 프로그램을 통하여 국민들에게 제공함으로써, 국민들의 국어 정보화에 기여함은 물론, 잊혀져만 가는 국어에 대한 역사적 사실을 밝혀 줌으로써, 국어의 보존에도 크게 기여하게 될 것이다. 따라서 이 연구는 거시적으로는 한민족 언어의 정보화를 확충시켜 나아가고, 미시적으로는 국민들의 국어에 대한 자긍심을 갖게 함으로써, 국어 발전의 기틀을 마련하는 데에 그 목적이 있다. 본 논문에서는 국어 어휘의 역사 검색 프로그램에 대한 개발 개요와 어휘의 시대별 분류 방식에 대해서 논한다.

1. 서론

"국어 어휘의 역사 검색 시스템 (1) 개발"은 국민들에게 국어 어휘에 대한 역사 정보를 제공하여 국어에 대한 지식과 이해의 폭을 넓힘으로써 국어 생활을 더욱 윤택하게 하고, 국어의 정확한 사용을 통하여 국민들의 의사소통의 정확성과 신속성을 제고함을 그 목적으로 한다. 아울러 이러한 국어 어휘 역사에 대한 정보를 컴퓨터 프로그램을 통하여 국민들에게 제공함으로써, 국민들의 국어 정보화에 기여함은 물론, 잊혀져만 가는 국어에 대한 역사적 사실을 밝혀 줌으로써, 국어의 보존에도 크게 기여하게 될 것이다. 따라서 이 연구는 거시적으로는 한민족 언어의 정보화를 확충시켜 나아가고, 미시적으로는 국민들의 국어에 대한 자긍심을 갖게 함으로써, 국어 발전의 기틀을 마련하는 데에 그 목적이 있다[1][2][3].

국민들이 관심을 가장 많이 가지고 있다고 추정되는 어휘 1,000 개의 어원 정보 및 어휘의 역사를 제시하여 줌으로써, 국민들의 국어 생활을 윤택하게 하고, 아울러 이것을 인터넷 상에서 공개하여 국민들의 국어에 대한 관심도를 높이도록 한다. 2002년도에는 약 1,000개의 표제항을 조사하되, 이를 15세기부터 20세기 및 방언형에 이르기까지의 형태, 음운, 의미 변화들을 쉽게 기술하도록 한다. 따라서 그 변이 형태는 약 5,000개에 이를 것으로 추정된다. 지금까지 국어의 어휘에 대한 어원 정보나, 그 역사를 기술해 놓은 문헌을 조사한다. 이를 통해 약 1,000개의 어휘를 선정한다. 이 1,000개의 어휘는 각각의 어휘에 대하여 형태, 음운, 의미 변화들을 쉽게 기술하도록 한다[1][2].

본 논문에서는 이와 같은 국어 어휘의 역사 검색 시스템을 위한 어휘 자료 데이터베이스 구축 방식과, 어휘의 검색, 검색된 어휘의 표시 방식 등을 설명한다. 또한, 시대별로 혼재되어 있는 어휘들을 출현 시대에 따라 정렬하는 정렬 방식에 대해서 논한다.

2. 국어 어휘의 역사 검색 프로그램의 데이터베이스 구축

2.1 국어어휘 자료의 수집과 정리

국어 어휘의 역사 검색 프로그램은 현대국어의 어휘 중에서 약 1,000개의 일상 어휘를 선택하여 그 어휘들 각각의 어원정보와 표기법 및 형태의 변화, 음운변화, 의미변화 과정을 문헌자료에 대한 철저한 실증적 자료를 제시하여 주어서, 오늘날 많은 사람들이 국어 어휘에 대하여 잘못 인식하고 있는 내용들을 제시하여 줌으로써, 국민들의 올바른 언어생활을 유도하는 데에도 그 목적을 둔다.

이 프로그램에서 제시된 자료는 21세기 세종계획의 1단계 사업 결과물에 공개된 파일을 대상으로 하되, 훈민정음 창제 이후부터 1949년 이전까지의 문헌을 중심으로 하였다.

(1) 자료명 : 제목이 당시의 표기 관습을 반영한 예들은 문헌에 반영된 표기를 그대로 옮겼다. '명주보월빙'과 같이 현대국어의 표기 방식에 따르지 않고 '명듀보월빙'처럼 한 것이 다.
(2) 이본 약호 : 같은 제목을 가진 여러 이본이 제공된 경우에는 그 이본명을 될 수 있는 대로 반영하여 넣었다.
(3) 자료의 장차 앞뒷면은 대부분 'a, b'로 표시하였다.
(4) 20세기의 자료는 1949년 이전의 자료만을 대상으로 하였다.
(5) 이용한 문헌자료의 간행연도와 약호, 문헌명을 보이면 다

음과 같다.

18?? 가곡원 : 가곡원류(歌曲原流)(국악원본)
1792 가례석 : 가례석의
1632 가례해 : 가례언해
18?? 가체신 : 가체신금사목
1525 간벽온 : 간이벽온방
1500 개법화 : 개간법화경 규장각 소장
1748 개첩해 : 개수첩해신어

(6) 이 중에서 물음표(?)의 표시를 한 것은 간행연도나 필사연도가 분명하지 않은 것을 의미한다. 그래서 '16??'은 17세기에, 그리고 '17??'은 18세기에, '18??'은 19세기에 간행된 문헌임을 뜻한다.

2.2 국어 어휘 데이터 베이스

국어 어휘의 역사 검색 프로그램은 1,016개의 표준어휘인 표제어와 4,146개의 시대별 어휘인 검색어를 가지고 있다. 표준 어휘는 어휘에 대한 기본적인 자료들과 시대별 어휘에 대한 세기별 형태와 예문을 가지고 있다. 시대별 어휘 자료 데이터 베이스에는 각 어휘가 문헌에 출현한 시대 정보를 담고 있다. [표 1]과 [표 2]는 표준어휘 자료 데이터베이스와 시대별 어휘 자료 데이터베이스의 구조를 각각 보여주고 있다.

[표 1] 표준 어휘 자료 데이터베이스

테이블 항목	항목 설명
ID	일련 번호
SNum	작성자의 일련번호
Writer	작성자
Date	작성일
SWord	검색어 ','로 구분
Standard	표준어
Part	품사
Mean	현대 뜻 풀이
Chinese	관련 한자어
Form15	15세기 형태
Exam15	15세기 예문
Form16	16세기 형태
Exam16	16세기 예문
Form17	17세기 형태
Exam17	17세기 예문
Form18	18세기 형태
Exam18	18세기 예문
Form19	19세기 형태
Exam19	19세기 예문
Form20	20세기 형태
Exam20	20세기 예문
Explain	종합 설명

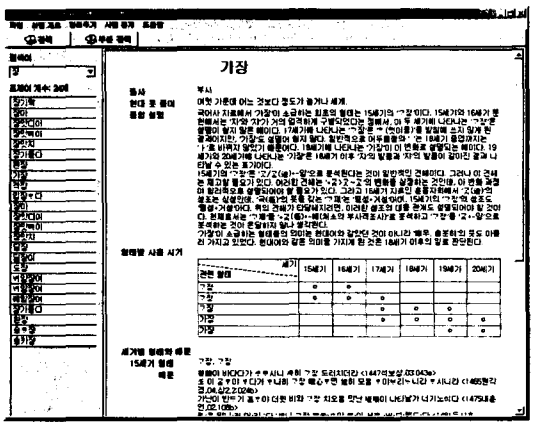
[표 2] 시대별 어휘 자료 데이터베이스

테이블 항목	항목 설명
ID	일련 번호
Word	어원 형태
StandardID	표준어 ID
Appear15	15세기 출현 플래그
Appear16	16세기 출현 플래그
Appear17	17세기 출현 플래그
Appear18	18세기 출현 플래그
Appear19	19세기 출현 플래그
Appear20	20세기 출현 플래그

3. 국어 어휘의 역사 검색 프로그램

3.1 국어 어휘의 역사 검색 프로그램의 실행

한국 방언 검색 프로그램은 Microsoft Visual C++로 작성되었으며, 데이터베이스는 Microsoft Access를 사용하였다.



<그림 1> 국어 어휘의 역사 검색 프로그램 실행 화면

검색어를 입력하면 데이터베이스의 단어와 부분 일치되는 모든 단어를 검색하여 준다. 이 검색어가 표준어인지 옛말인지를 구분하지 않는다. 옛말일 경우에는 해당하는 표준어를 찾아준다. 검색된 표준어는 앞에서부터 일치하는 항목들이 먼저 나오고 나중에 중간에 일치하는 내용들이 나온다. 이런 방식을 통해 사용자가 원하는 가장 적절한 결과가 앞쪽에 나올 수 있도록 한다.

국어 어휘의 역사 검색 프로그램은 옛 글자(고어)를 완전하게 표시해 주고 있다. 어휘 역사 자료들의 경우 대부분의 자료들이 옛 글자를 포함하고 있어 옛 글자 표시 지원이 필수적이다. 따라서 출력된 결과뿐만이 아니라 검색된 결과들도 모두 옛 글자를 지원할 수 있어야 한다. 검색 창에서 옛글자에 대한 직접적인 입력은 지원되지 않는다. 옛 글자 입력은 옛 글자를 입력할 수 있는 다른 프로그램을 이용하여 수행하고,

입력된 내용을 검색 창에 붙이기 기능을 사용하면 된다.

부분검색 기능을 이용하면 검색범위를 지정하여 전체 표제어의 일부를 검색할 수 있다. 선택사항으로 주어지는 'ㄱ'에서 'ㅎ'까지 중의 하나를 선택할 수도 있고 검색 범위를 입력할 수도 있다.

3.2 형태별 사용시기 정렬

원하는 어휘를 선택하면 그 어휘에 대한 자세한 정보를 보여준다. <그림 1>은 '가장'를 선택했을 때, 표준어 '가장'에 대한 정보를 보여주고 있다. 또한 출력 화면에는 <그림 2>와 같이 각 어휘의 형태별 사용시기를 표로 보여주고 있으며, '국어 어휘의 역사 검색 프로그램'은 입력된 순서와 관계없이 어휘의 출현 연대에 따라 자동으로 정렬해 주는 기능을 갖고 있다.

관련 형태 \ 세기	15세기	16세기	17세기	18세기	19세기	20세기
ㄱ장	ㅇ	ㅇ				
ㄱ장	ㅇ	ㅇ	ㅇ			
ㄱ장			ㅇ	ㅇ	ㅇ	
가장			ㅇ	ㅇ	ㅇ	ㅇ
가장					ㅇ	ㅇ

<그림 2> 형태별 사용 시기

형태별 사용시기에 따른 검색어 정렬을 위해서는 다음과 같은 규칙에 따라야 한다.

(1) 이전 시기에 나온 검색어가 먼저 표시되어야 한다.
(2) 출현시기가 동일할 경우 나중까지 출현한 검색어가 나중에 표시되어야 한다.

첫 번째 조건을 만족시키기 위하여 출현시기를 비트별로 표시하였다. [표 3]과 같이 15세기를 상위비트로 잡고 20세기를 최하위 비트로 설정하여 출현시기에 1 값을 세팅하는 경우 각각의 비트값들은 다음과 같다.

[표 3] 출현 세기별 비트 할당

비트	5	4	3	2	1	0
표현값	15세기	16세기	17세기	18세기	19세기	20세기

① ㄱ장: 110000 ② ㄱ장: 111000
③ ㄱ장: 001110 ④ 가장: 001111
⑤ 가장: 000011

위 값들을 내림차순으로 정렬하는 경우 ②①④③⑤ 의 순서가 되어 (2)의 규칙에 위배된다. 같은 시기에 시작되었어도 끝나는 시기가 먼저이면 앞쪽에 표시되어야 한다. (2)의 조건을 만족시키기 위해 [표 4]와 같이 종료 세기 비트를 설정하였다.

[표 4] 출현 세기별 비트 할당

비트	11	10	9	8	7	6	5	4	3	2	1	0
표현값	15세기	15종료	16세기	16종료	17세기	17종료	18세기	18종료	19세기	19종료	20세기	20종료

[표 4]에 의해 비트값들을 재할당하면 다음과 같다.

① ㄱ장: 101100000000 ② ㄱ장: 101011000000
③ ㄱ장: 000010101100 ④ 가장: 000010101011
⑤ 가장: 000000001011

앞에서 보는 바와 같이 각각의 비트값이 ①②③④⑤ 의 순서로 정렬됨을 알 수 있다.

4. 결 론

본 논문에서는 21세기 세종계획의 한민족 언어 정보화 분과의 결과물인 국어 어휘의 역사 검색 시스템에 대해서 소개하고 검색 및 형태별 사용시기 정렬 방식에 대해서 논하였다. 특히 형태별 사용시기 정렬 기능은 세기별로 출현하는 어휘를 보기 쉽게 정렬하는 기능을 가지고 있으며, 이를 위해 세기별 출현 정보를 비트에 할당하는 방식을 사용했다. 같은 시기에 출현해서 먼저 사라지는 어휘를 앞쪽에 위치시키기 위해 세기별 종료 비트를 두어 문제를 해결하였다. 이 방식은 주어진 문제를 해결하기 위한 단순하면서도 명료한 방식이라 생각된다.

참고문헌

[1] 이태영 외(2002), 21세기 세종계획 한민족 언어 정보화 연구보고서, 문화관광부/국립국어연구원
[2] 이태영 외(2001), 21세기 세종계획 한민족 언어 정보화 연구보고서, 문화관광부/국립국어연구원
[3] 세종계획 홈페이지, http://www.sejong.or.kr/
[4] 이동광, 안동언, 정성종, 김호영, 두길수(2002), "한국 방언 검색 프로그램 개발", 대한전자공학회 하계 학술발표 논문집
[5] 김금영, 조시성, 안동언, 정성종, 두길수(2002), "남북한 언어 비교 사전 검색 프로그램 개발", 대한전자공학회 하계 학술발표 논문집
[6] 오형진, 정성종, 안동언, 이신원, 두길수(2002), "국어 어문규정 검색 프로그램 개발", 대한전자공학회 하계 학술발표 논문집