국어대사전의 표제어에 나타나는 한자 정보

Hanja Information in the Entries of Korean Unabridged Dictionary

김철수
서남대학교 컴퓨터정보통신학과

Cheol-Su Kim(chskim@seonam.ac.kr)

요약

한글과 한자가 혼합되어 나타나는 분야의 언어정보처리를 수행하기 위해서는 한글 및 한자 정보를 동시에 지원할 수 있는 전자 사전이 요구된다. 본 논문에서는 국어대사전의 표제어에 나타나는 한자 통계 정보에 대하여 고찰하였다. 대상 정보는 KSC-5601 코드에 기초하여 엔트리에 한자가 포함된 엔트리 수, 엔트리에 나타나는 한자의 음과 훈의 출현빈도 수, 품사별 한자 출현빈도수, 엔트리당 평균 출현 한자 수 등이다.

440,594개의 표제어 중 303,951개에서 한 글자 이상의 한자가 나타나 68.99%의 표제어에서 한자가 출현하였다. 440,594개의 표제어에서 858,595글자의 한자를 포함하고 있어 표제어 당 평균 1.95개의 한자가 출현하였다. 표제어의 평균 음절길이 3.56이고 1.95개의 한자가 출현하므로 표제어를 구성하는 글자 중 54.78%가 한자임을 알 수 있다. 4,888개의 한자 코드 중 한번 이상 출현한 한자는 4,660개이며, 228개의 한자는 한 번도 출현하지 않았다. 4,000번 이상 출현하는 한자는 5개였다. 엔트리에 출현하는 858,595개의 한자에 대응하는 한글 음은 471개였다.

■ 중심어 : | 언어정보처리 | 전자사전 | 국어대사전 | 한자통계정보 |

Abstract

For language information processing that includes both Hangul and Hanja, an electronic dictionary supporting Hangul and Hanja simultaneously is necessary. This paper examined statistical information on Hanja entries of Korean Unabridged Dictionary such as the number of entries that include Hanja based on the KSC-5601 character set, the frequency of the pronunciation and meaning of each character of Hanja included in the entries, the frequency per part of speech of Hanja in entries and the average number of Hanja characters per entry.

At least one or more of Hanja characters appear in 303,951 entries out of 440,594, accounting for 68.99% of the total. 858,595 characters of Hanja are included in the 440,594 entries, which is 1.95 Hanja characters per entry. As the average syllable length of the entries is 3.56 and the average count of the Hanja characters per entry is 1.96, it can be said that 54.7% of all the characters of the entries are in Hanja. Among 4,888 Hanja character codes, 4,660 are used once or more, whereas 228 Hanja codes never appear in any entry. There were 5 characters which appear more than 4,000 times. A total of 858,595 Hanja characters used in all the entries correspond to 471 Hangeul codes.

■ keyword : | Language Information Processing | Electronic Dictionary | Korean Unabridged Dictionary | Hanja Statistical Information |

접수번호 : #091215-002
접수일자 : 2009년 12월 15일
심사완료일 : 2010년 01월 14일
교신저자 : 김철수, e-mail : 21chskim@hanmail.net

1. 서 론

언어 정보 처리 응용 분야는 형태소분석[1], 정보검색, 철자검색, 음성인식, 문자인식[2] 등 다양하다. 이러한 정보처리 시스템은 처리과정에서 전자사전을 매우 많이 참조하며, 전자사전은 언어정보처리 시스템의 필수적인 요소이다. 언어 정보처리는 응용환경에 따라 몇 단계의 처리과정을 거쳐야 하고, 각 단계마다 필요한 정보의 종류 및 형태는 다르지만 이 정보들은 전자사전에 저장된 정보로부터 얻어진다. 한글과 한자가 혼용되어 나타나는 응용환경에서는 한글에 대한 관련 정보뿐만 아니라 한자에 대한 관련 정보가 요구된다. 그러므로 한글에 관한 정보와 한자에 관한 정보를 동시에 지원할 수 있는 전자 사전이 요구된다.

우리 일상생활에서 사용하는 단어들을 살펴보면 고유 단어들도 있지만 외래어도 있으며 특히 한자 단어들이 많다. 우리나라는 오래전부터 중국과 많은 문화교류가 있었던 까닭에 일상생활에서 사용하는 많은 단어들이 한자를 포함하고 있다. 삶의 문화와 일상생활이 언어와 문자에 영향을 주기도 하지만 언어와 문자 정책이 일상생활에서 사용하는 어휘들에 많은 영향을 준다. 세종대왕 때 창제된 한글은 우리 고유의 글이면서도 사대주의 사상에 젖어 있던 당시 문화와 관습으로 많은 멸시와 천대를 받았으나 해방이후 한글 전용 정책으로 한글이 제자리를 찾아 왔다. 이런 까닭에 일상생활에서 많이 사용되어 왔던 한자들이 이제는 눈에 잘 뜨이지 않지만 한자 혼용론을 주장하는 의견들도 있다. 한글은 표음(表音) 문자이지만 한자는 한 글자마다 특정한 의미를 가지고 있는 표의(表意) 문자이다. 소리는 같지만 서로 다른 의미를 가진 단어(동음이의어)의 의미를 명확하게 전달하기에는 한자가 도움을 주므로 자연스러운 논의일지도 모른다. 그러나 영어의 경우 한글과 구조적으로 다른 특징을 가지고 있고, 동음이의어가 많이 존재하지만 전 세계의 많은 나라에서 사용되고 있다. 언어 및 문자 정책은 이러한 여러 언어들에 대하여 종합적인 비교검토와 연구, 민족 문화의 정체성 등을 고려하여 장기적이고 국가적인 안목에서 매우 심사숙고해야 할 문제이다.

본 연구에서는 한글과 한자가 혼용되어 나타나는 환경의 정보처리 과정에 도움이 될 수 있는 한자들에 관한 정보를 알아보기 위하여 국어 대사전의 표제어에 나타나는 한자들을 조사하였다. 한자들에 대한 통계 정보는 일상생활에서 사용하는 말뭉치에 나타나는 한자 단어들에 대한 통계 정보를 조사하거나 국어사전에 등록되어 있는 표제어에 나타나는 한자 단어들에 대한 통계 정보를 조사하는 방법이 있다. 국어사전에 등록되어 있는 단어들이 일상생활에서 일정한 비율로 활용되고 출현하는 것은 아니므로 말뭉치를 활용하는 방법이 좀 더 객관적이라고 할 수 있다. 그러나 한글단어와 한자가 동시에 병기되어 있고 다양한 분야를 포함하고 있는 말뭉치 선정도 쉽지 않다. 따라서 사용빈도수를 반영하지는 못하지만 모든 분야의 다양한 단어를 포함하고 있는 국어대사전[3]에 등록된 표제어의 외래어 표기 부분에 나타난 한자들에 대하여 고찰하였다.

2. 관련연구

2.1 문자정책의 변화 과정

1443년 세종 때 창제된 한글은 온 국민이 아닌 평민, 여자, 일부 계층에서만 사용되었고, 관리계층 및 중요문서들 대부분은 한자를 사용하였다. 이런 현상은 조선말까지 계속되었으며 고종 때부터 변화가 일기 시작한다.

한자 사용에 관한 문자 정책의 주요 변화 과정은 [표 1]과 같다[4][5][12].

표 1. 한자 정책 주요 변화 과정

년도	주요내용
1894	•외국국명 지명 인명 국문 표기 법령 공포 실시 •법률과 칙령 모두 국문을 근본으로 삼고, 附譯 혹은 국한문 혼용규정
1945	•한자 폐지와 가로쓰기 토의 결정(미군정청 학무국, 조선교육위원회 교과서 분과)
1948	•한글 전용에 관한 법률 제정 공포
1949	•교과서에 한자 혼용 의결
1950	•한자 혼용 결정
1951	•문교부 제정 '상용 일천 한자표' 공포
1954	•한글 전용 강조, 상용한자 1300자 제정, 초등학교 고학년 한자 혼용
1955	•문교부 한글 전용법 발포
1957	•'임시 제한 한자 일람표' 1,300자 심의 결정 •언론 기관을 겨냥한 한자 제한으로 만든 '임시 제한 한자 일람표' 이후 '상용한자 일람표'로 변경

	•한글 전용 적극 추진안과 한글 전용법 개정안 국무회의 상정
1958	•한글 전용 실천 요강실시
1959	•문교부, 내무부의 협조로 거리 간판의 한자를 강제로 추방함
1961	•한글 전용 법률안 개정(국가재건최고회의) •'정부 공문서 규정'
1962	•한글 전용 원칙 발표, 문교부에 '한글 전용 특별위원회' 설치
1963	•교과서에서 한자 노출시킴
1964	•한자 교육 부활(초등학교 600, 중학교 400, 고등학교 300)
1965	•한글 전용에 관한 개정 법률안 공포, 초중고 교과서에 임시 허용한자 1,300자 노출
1968	•대통령 한글 전용 선언, 한글 전용 5개년계획 수립
1969	•'교육과정 중 개정령' 공포, 한자교육근거 없애고 모든 교과서에서 한자를 없앰. 판서도 한자사용 금지
1970	•한글 전용 단행, 초중고 교과서에서 한자를 완전히 없앰
1971	•한자교육 부활 결정
1972	•교육법 시행령 개정 •문교부 선정 교육용 기초한자 1,781자 시안 발표 •'중학교 한자교육용 제한한자' 1,800자 확정공포(중학교 900, 고등학교 900)
1973	•각 대학에 한문교육과 신설
1974	•중고교 교과서에 한글한자 병용방침 결정(초등학교 제외)
1975	•중고교 교과서 기초한자범위 한자()로 병기 •중교교에서 한문을 선택과목으로 축소
1995	•초등 3~6학년 기초한자(600자 미지정) 학교재량교육 가능
1998	•정부공용문서 한자병용가능 도로표지등 한자병용 방침
2000	•주민등록증 성명에 한자병용 기초한자 일부 조정

문자 정책은 해방과 더불어 한글 전용으로 큰 변화를 가져왔다. 그러나 한자 사용에 관한 문자 정책은 일관성 없이 한글전용, 한자혼용, 한자병용 정책을 오가며 갈팡질팡한 모습을 쉽게 볼 수 있다. 한자 사용에 관한 찬반 의견 역시 문자 정책 못지않게 지금까지도 계속되고 있다[6][7]. 현재는 한글 전용 속에 한자를 병용하자는 의견들이 제시되고 있다. 한자 사용에 따른 장단점, 한글 전용에 따른 장단점들을 지적하고 있다[8][9].

2.2 국어대사전의 표제어 정보

국어대사전[3]은 50만이 넘는 표제어를 7,300여 쪽에 담고 있는 대사전으로 방대하고 다양한 정보를 담고 있어 백과사전의 기능을 겸하고 있다. 닿소리별 주표제어 수는 [표 2]와 같다[10].

표 2. 국어대사전의 닿소리별 표제어 수

닿소리	표제어수	백분율(%)	닿소리	표제어수	백분율(%)
ㄱ	65,524	14.87	ㅈ	50,747	11.52
ㄴ	17,510	3.97	ㅊ	19,484	4.42
ㄷ	31,133	7.07	ㅋ	5,267	1.20
ㄹ	9,589	2.18	ㅌ	9,687	2.20
ㅁ	26,710	6.06	ㅍ	12,574	2.85
ㅂ	39,478	8.96	ㅎ	28,173	6.39
ㅅ	55,706	12.64	합계	440,594	100.00
ㅇ	69,012	15.66			

가장 많은 표제어수를 가지는 닿소리는 ㅇ, ㄱ, ㅅ, ㅈ 순서로 69,012, 65,524, 55,706, 50,747개로 각각 15.66%, 14.87%, 12.64%, 11,52%를 차지하여 4개의 닿소리(ㄱ, ㅅ,ㅇ,ㅈ)로 시작하는 표제어가 전체 표제어의 절반 이상(54.69%)이다.

전자사전 구축에 따른 엔트리 수는 361,980개이다. 전자사전 구축에 사용되는 엔트리 361,980개에 출현하는 서로 다른 한글 음절수는 2,463개 이다. 전자사전에 출현하는 2,463개의 음절 중 단 한번 출현하는 음절은 372개였다. 361,980개의 엔트리에 사용된 전체 음절수는 1,289,659개, 단어의 평균 길이는 3.56이다.

2.3 한자의 특징

한자는 표의문자로 각 글자마다 하나의 형태, 하나의 음절, 하나의 개념을 나타내고 있다[11]. 하나의 글자를 이해하는데 있어서 형(形), 음(音), 의(義)의 세 요소가 동시에 필요하고 서로 밀접한 관계를 가지게 된다. 세 가지 중에서 한 가지라도 모르면 글자를 이해할 수 없게 되므로 한자 학습에 있어 중요한 기초가 되고 있다. 한자는 물체의 모양이나 형태를 본떠서 만들어진 글자로 각 글자마다 다른 뜻을 가지 있어 시각적 이해가 빠르다. 반면 획수가 너무 많고 구조가 복잡하며 자수가 많아서 모두 기억하기 어려운 단점을 가지고 있다. 한자의 발음에 있어서는 원칙적으로 한 글자는 하나의 음을 가지나 두 가지, 세 가지 음을 가질 수도 있다 즉 更(다시 갱, 고칠 경) 見(볼 견, 나타날 현)이나 樂(즐길 락, 풍류 악, 좋아할 요) 등이다. 또한 같은 음을 가지는 글자도 많다. 靑, 淸, 請, .. 등은 '청' 이란 하나의 음으로 많은 한자가 있다. 한자는 어순에 따라 의미와 직능이 달라지게 되는데 한자어에 있어서 그 위치에 의하여 의미의 변화를 가져오게 된다. 즉, 明月과 月明, 名人과 人名을 들 수 있다. 이러한 현상은 한자의 독특한 점이라고 볼 수 있다.

한자는 각 글자마다 고유한 뜻을 가지게 된다. 따라

서 어떠한 사물과 의미를 표현하는데 함축성이 있으며 문장을 만드는데 간결하여 문장의 강도를 높일 수 있다. 한자의 수가 많아 기억과 활용에 어려운 점이 있으나 한자는 그 구조와 응용의 종류에 따라 육서에 의하여 조자(造子)되었으므로 한자 구조 방식과 부수의 기본자의를 터득하면 오히려 처음 보는 한자라도 그 의미와 발음을 짐작할 수 있고 문자를 성격별로 분류할 수 있는 편리한 점도 지니고 있다.

3. 국어 대사전의 한자 정보

국어대사전의 표제어에 나타나는 한자 정보들을 알아보기 위하여 한자의 코드 체계를 설정하였다. 한글과 한자를 컴퓨터에서 표현하기 위한 코드 체계는 KSC-5601, UNICODE 등 다양하나 본 연구에서는 KSC-5601 코드 체계를 이용하였다.

KSC-5601 코드체계의 한자 코드 영역은 $CAA1_{(16)}$ ~ $FDFE_{(16)}$로, 상위 바이트는 $CA_{(16)}$ ~ $FD_{(16)}$이고, 하위바이트는 $A1_{(16)}$ ~ $FE_{(16)}$ 사이로 코드영역을 그림으로 표현하면 [그림 1]과 같다. 코드 영역에서 코드값 $CAA1_{(16)}$ ~ $FDFE_{(16)}$에 존재하는 한자는 전체 4,888 글자이다.

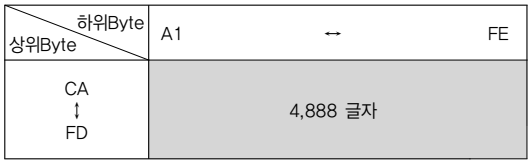

그림 1. KSC-5601 한자코드 영역

1) 한자 코드에 대응하는 한글 코드 수

KSC-5601 한자 코드 영역에 포함되어 있는 한자코드들이 몇 개의 음을 가지는지 조사하였다. 즉, 4,888개의 한자 코드 영역에 포함되어 있는 한자들이 가지는 각각의 음(音)과 의(義)에서 한자의 음들이 한글의 어떤 음들에 해당되는지 알아보기 위하여 한자 코드를 한글코드로 변환하였다. 한자 코드 영역에 포함된 한자를 한글 음으로 변환했을 때 대응되는 한자의 이의동음 코드 수는 [표 3]과 같다.

표 3. 한자 코드의 이의동음 개수

동음이의어 한자 수	음의 개수	해당 한글 음절
50 이상	6	기수사유정구
30 ~ 49	23	조경부비연전호도영주고요이소장진지선우상양서오
20 ~ 29	37	가방시저포순원초리자교반위적간강계모신여예용인치황제추무창현감개단배형회희
10 ~ 19	114	만미박분의준천노려령로성아완해거관담동병복재종항환공대보봉탁파규근나명석식염질차찬척취금류매빈악타태표하한화광량루민설역운윤작효후건과련료망묘문산숙채축함협각견당두면발벽애야임증패갈곤권독돈람번승심안어열참철첨첩총탄편폐필홍훈
5 ~ 9	125	괴난남등라락란린범세속약옹욱일점침판향혜낙누랑뢰마백삼섬송암와익절청휘검격곡균극급낭농력록롱말목변앙액육은집착통피합갑결겸국군궁궤귀내녹뇌능둔렬륜릉막맹붕손언엄온욕음잠촉충퇴투폭혼확궐납달답렴례림맥불살색생쇄습억옥왕외율융입잔좌직찰측탕평풍학행활휴흉
2 ~ 4	114	갱걸곽괄굉굴긍길념래룔립벌별빙삭삽섭술실알압앵엽왜응잉쟁족중책처촌탐토팽허헌혁혈훤흔흘흡겁게골괘년랍륙몽물밀새숭슬십울월읍접졸즙징최춘출췌칙칠탑택팔혹홀횡훼흘흠객날녕뉴늑니닉다덕돌랄략륙멱멸몰묵법쇠얼업웅존죽짐징탈특품핍할핵험획
1	54	갹곶귤글긴김끽냉녀논뇨눈눌늠댁득랭련론룡륭름몌본볼북솔쌍씨엔올왈을잡죄줄즉즐촬층친칩쾌탱터틈퍅폄헐혐훙흑흥힐
한자 4,888	473	

4.,888개의 한자 코드를 한글 코드로 변환했을 때 나타나는 한글 음절수는 473개로 전체 11,172개의 한글 음절 중 극히 일부(4.2%) 음절로 변환되었다. 이는 한자가 뜻글자에 기인한 것으로 한글로 표현했을 때 이의동음이 많기 때문이다. 한자코드를 한글코드로 변환했을 때 한자 코드 4,888개에 대응하는 한글음의 개수가 473개이므로 한자의 이의동음어 평균수는 10.33개이다. 가장 많은 음을 가지는 이의동음어 수의 순위는 '기'(64개), '수'(61개), '사'(60개) 순서였다. 한자코드를 한글코드로 변환했을 때 변환된 음의 개수가 하나인 이의이음어 수는 4,888개중 54개("곶귤글긴김끽 ...")로 1.2%에 불과하였다. 이는 한자가 음과 뜻을 이용하여 표현하는 한자 특성상 서로 다른 뜻을 가진 여러 개의 한자들이 소수의 음으로 표현됨을 알 수 있다.

2) 닿소리별 한자 출현 정도

표제어들이 한자를 포함하고 있는 정도를 알아보기 위하여 표제어에 한자들이 출현하는 단어들을 조사하였다. 한 글자 이상의 한자를 포함하는 표제어들의 닿소리별 출현 빈도는 [표 4]와 같다.

표 4. 표제어에 나타나는 한자 빈도

닿소리	표제어수	한자출현 표제어 수	한자출현 빈도	한자출현 비율(%)	단어당 평균한자수
ㄱ	65,524	46,371	135,428	70.77	2.07
ㄴ	17,510	9,699	25,792	55.39	1.47
ㄷ	31,133	16,882	47,619	54.23	1.53
ㄹ	9,589	6,937	17,736	72.34	1.85
ㅁ	26,710	15,602	43,080	58.41	1.61
ㅂ	39,478	25,182	71,191	63.79	1.80
ㅅ	55,706	41,091	117,285	73.76	2.11
ㅇ	69,012	49,411	137,555	71.60	1.99
ㅈ	50,747	39,488	118,050	77.81	2.33
ㅊ	19,484	16,445	44,119	84.40	2.26
ㅋ	5,267	1,315	2,440	24.97	0.46
ㅌ	9,687	6,366	17,813	65.72	1.84
ㅍ	12,574	7,924	21,213	63.02	1.69
ㅎ	28,173	21,238	59,240	75.38	2.10
합계	440,594	303,951	858,561	68.99	1.95

표제어 440,594개 중 303,951개의 표제어에서 한자가 출현하여 68.99%의 표제어에 한자가 포함되어 있다. 한자가 전혀 포함되지 않은 표제어는 136,643개로 31.01%였다. 표제어에 나타나는 한자 수는 858,561개로 전체 표제어에 대하여 평균적으로 1.95개의 한자가 출현하였다. 가장 많은 한자를 포함하는 표제어는 "남북조선정당사회단체대표자련석회의(南北朝鮮政黨社會團體代表者連席會議)" 17글자였다.

3) 한자코드들의 출현 빈도

표제어에는 한글뿐만 아니라 한자어, 영어와 같은 외래어도 포함되어 있다. 표제어에 나타나는 한자 글자들의 분포를 조사하였다. 4,888개의 한자 글자 중 4,659개의 글자가 858,536회 출현하여 한자들의 평균 출현 횟수는 전체 한자에 대해서는 175.6회, 출현한 한자만을 대상으로 했을 때는 184.3회 출현하였다. 한자글자들의 출현 빈도는 [표 5]와 같다.

가장 많이 나타나는 한자는 '法'으로 5,183회이고, 4,000회 이상 출현하는 한자가 5글자('法', '學', '性', '子', '大'), 1000~3999회 출현하는 한자가 176글자, 100~999회 출현하는 한자가 1,406글자이다. 단 1번 출현하는 한자는 153개이고, 한 번도 출현하지 않은 한자 수는 229개로 전체 한자의 4.7%였다.

한자가 출현한 표제어들만을 대상으로 한 한자들의 평균 출현 빈도는 2.82개이다. 이는 표제어들의 평균 길이가 3.56이므로 한자가 출현한 표제어일 경우 단어의 대부분인 79.2%가 한자로 표기됨을 의미한다.이러한 한자의 출현빈도는 습득해야할 권장 한자들의 범위를 결정하는데 단서를 제공한다.

표 5. 한자 글자들의 출현빈도

출현빈도	출현 음절 수 / 출현 음절
4,000이상	5 / 法學性子大
1000~ 3999	176 / 物地人生水動機化山氣國電金中體分行一文主三線無合器事會業自定石色理者心數力式音之度天的不間酸義上經工家計高道制書形神日海白用通外發面教長小出内時平方作成素空軍下同門本風論花戰症光語權民相重管產二土病官所木公科點正信部質王場原角變量世有字保四黃說熱實半頭草記期代反鐵手壓單政傳流藥車入曲對射調圖明言五運接節火魚直結全路靑馬食年交船板安解位多關別波感品名層
100~999	1406 / 口標社極新術臺造回加帶後農血命集放星轉種界月 ~
10~99	2147 / 傍巢例夷罷狗激雇愁禽焦輻婢殼燐迎莊栽眉枕泰廻 ~
2~9	772 / 翟湮砥嵬弁迪謂羔朧脛碣漱砦汪歿喔蔣箏剃駟儼鍾 ~
1	153 / 珂琫擄錡饍円淄縕鋌万潽荇瀏珦阮峽艸珷颰琛囹峀檍枏艅柒甍靚炫玹旹絇靺碉筫鐫杻苽岺阡堧鈺踈熢涑泂忖炅鋈鈑綷頊徇烟眕崙澔季亶邨祜昰炡浬旽坍澍蕭痺廾禑苒鉥恁秕蕓毗秘穖娃愿璥傃檍楫唇絪刺豈勍玭琠龐悓洹唵滉棄縝蕣杰劒璘啣磅怒湄澐峰絅郜弘噶斅鞨岩磻撚梏珽譬毛塡盼岵聆璟煊稶莫璨煇碍芯埼璹烋甬勣悰撅矔酩旼媏彧袓脩予圻憙渼槻
0	229 / 珏玜鉀畺舡盖迲俓倞暻烱堺蒿琨珙琯侊洸眏僬玖銶淃昀劤妗昑伋玘琪豈錤珞楠湳柰甯菉彖澾錟垪嶋焞燉嶝怜澧潞瀘菉瀧劉侖潾盉輞沔榆茗玅楙雯沕梶盃裴鱉昞昦棟軿湺珷宓輹盼斌浜乶岾捿敳琁鐥宬城笹篠琇璲雖潚琡栒橓珣柿侁拾瀁俺歟璵磯挺沇璉秊縝渶濚瀯煐聆鍈寠汭俉塢昨嫯瑥梡琓埇墉榕瑢玗玗釪勖栯櫰熉丂嫄寃暲蒿禕楢洧玧贇鈗瀜諡璘藺暲瀸吊畑点延晸柾湞瀞玎綎諄晁璪淙琮姝酎埈寯焌沚晋桭溱璔璡侹瓆潗鏶侘瀿宲仟玔苔愡萩瑃廁晫琸撑霸豊泌珌馝廈閒晛滎瀅迴時晧淏芦艧嬅碻晥晄榥湟鐄涍帿珝頊劤炘煕譆割懎熺詰
계	4888

4) 출현 한자에 대응하는 한글 음절 발생 빈도

440,594개의 표제어에 나타나는 한자 글자들을 한글 음절로 변환했을 때 나타나는 한글 음절 빈도를 조사하였다. 앞의 연구 "1) 한자 코드에 대응하는 한글 코드 수"에서는 한자 코드 4,888개를 한글 음절 코드로 변경하였을 때 대응하는 한글 음절 빈도로 표제어에 한자 글자들의 출현빈도는 고려되지 않았다. 따라서 표제어에서 실제로 발생하는 한자 글자들에 대응하는 한글 음절들의 발생 빈도를 조사하였다. 440,594개의 표제어에 나타나는 한자를 한글로 변환하였을 때 한글 음절의 발생 빈도는 [표 6]과 같다.

표 6. 표제어에 나타난 한자에 대응하는 한글음절

동음한자 출현범위	동음 개수	해당 한글 음절	전체출현 빈도
10,000~	7	사기수자전지정	94,385
5,000~ 10,000	34	성상대도장화산주조동제소선공부유계가구경인고신관원시연의문방식법이적	235,821
2,000~ 4,999	96	학무중비단일리양반교회보금분세진물호서국생천위석재간과해어용형체명종차영강초증각통병광음포심우삼미하감불행모복로파군합청판안치권백약력면설색역당류실목염표속업한절배발개오환거변직점작마만량아내	297,017
1,000~ 1,999	93	운육질열등독평예현황창태건결봉여두항곡외탄후극급민송격출토순매풍근망추요령승악향철축왕저본론편족노입언채은박농충남료집압야취피다좌난임옥접월련술투립폐궁침암담혈총록품년별견막준응혼균홍층	135,699
500~ 999	90	험귀액함검효타란능처례맥림락골규착책숙범십활습온협벽측덕완폭애척패칙엽라탁북찰필률죄쌍죽허밀팔특려루묘참객칠나번탕희맹갑손쇄익잡윤율래녀괴렬굴살랑잠훈념퇴득존쟁핵최탈김친촌욕낙촉흑	64,214
100~ 499	99	첩뇌찬춘낭녹택탐헌납달빙잔빈융섬말륙흥륜룡흡삭와답첨탑혜앙칭흉몽휴혁갈벌뢰징엄냉확람누웅휘억궤획릉할린묵솔랭략곤횡돈돌왜겸갱롱옹둔멸니섭몰논렴붕읍울길궐괘걸겁즉씨혹숭곽쇠잉솔쾌뇨촬긴삽즙슬팽랍긍흡흔	29,279
50~99	42	귤을앵핍륵훼알괄횰협혐계홀새굉녕늑헐췌댁륭욱흘닉훤얼멱끽날랄폄눌곶울즐왈칩짐눈몌뉴늠	2,076
1~49	10	름갹퍅터틈훙글줄볼엔	45
합계	471		858,536

표제어에 나타나는 한자들에 대응하는 한글 음절 가운데 10,000회 이상 출현하는 한글 음절은 7개(사, 기, 수, 자, 전, 지, 정)의 음이 94,385회이고 5,000~9,999회 출현하는 음은 34개, 50~99회 출현하는 음은 42개, 50회 미만 출현하는 음은 10개로 출현하는 전체 음은 471개였다. 그러므로 한자코드 4,888개를 한글 음절(코드)으로 변환하였을 때 대응하는 한글 음의 수는 473개 이지만, 실제로 대사전의 표제어에 나타나는 한자의 한글 음은 471개이다. 한자 코드를 한글 음으로 변환했을 때는 나타나지만 실제 사전의 표제에서는 출현하지 않은 음은 2개(힐, 탱)이다. 사전의 표제어에 출현하는 858,536개의 한자들은 471개의 한글 음으로 표현된다.

5) 한국어 표제어의 품사별 구성 비율

한자가 포함된 표제어들의 품사별 분포비율을 알아보기 위하여 먼저 등록된 전체 표제어들에 대한 한국어의 품사별 구성 비율을 조사하였다. 품사란 단어를 문법적인 특징의 공통성에 따라 나눈 부류이다. 문법적 특징이란 “형식”, “기능”, “의미”로 이 세 가지가 품사 분류의 기준이 된다. 우리말의 품사 분류는 학자에 따라 차이가 있기는 하나 학교문법에서는 9품사 체계로 나눈다. 9품사는 “명사, 대명사, 수사, 동사, 형용사, 관형사, 부사, 감탄사, 조사”이다. 한국어 표제어에 대한 품사별 구성 비율은 [표 7]과 같다.

표 7. 한글 표제어의 품사별 구성 비율

구분	단 어 수	백분율(%)
명 사	334,962	76.03
대명사	465	0.11
수 사	277	0.06
동 사	15,173	3.44
형용사	6,446	1.46
관형사	529	0.12
부 사	14,093	3.20
감탄사	812	0.18
조 사	357	0.08
품사소계	373,114	84.68
9품사 표제어수	371,904	84.41
비품사 표제어수	68,690	15.59
표제어 합계	440,594	100.00

9품사를 가지는 표제어 수는 373,114개로 전체의 84.68%이고, 9품사를 가지는 서로 다른 표제어 수는 371,904개 84.41%이다. 9품사 수의 합과 실제 9품사를 가진 실제 표제어 수는 1,210개의 차이가 난다. 이는 동일 표제어가 2개 이상의 품사를 가지는 요인에 기인한다. 1개의 품사를 가지는 표제어가 370,740개, 2개의 품사를 가지는 표제어가 1,121개, 3품사를 가지는 표제어가 40개, 4품사를 가지는 표제어는 3개였다.

9품사 범주에 포함되지 않는 표제어는 주로 복합명사와 북한어 등이 많았다. 품사별 구성 비율은 명사 76.03%, 동사 3.44%, 부사 3.20%, 형용사 1.46%로 나타났다. 9품사를 제외한 비품사 표제어들의 수는 68,690개로 전체 표제어의 15.59%이다.

6) 한자가 출현하는 표제어들의 품사별 비율

“2) 닿소리별 한자 출현정도”에서 440,594개의 전체

표제어 중 한 글자 이상의 한자를 포함하는 표제어수는 303,951개로 한자 출현 비율은 68.99%이다. 한 글자 이상의 한자가 포함된 표제어들의 품사별 분포를 조사하였다. 한자가 포함된 표제어들의 품사별 구성 비율은 [표 8]과 같다.

표 8. 한자 출현 표제어들의 품사별 비율

구분	개수	백분율(%)
명사	243,030	79.96
대명사	245	0.08
수사	90	0.03
동사	420	0.14
형용사	1,264	0.42
관형사	195	0.06
부사	796	0.26
감탄사	64	0.02
조사	0	0.00
합계	246,104	80.97
9품사 표제어수	245,941	80.91
비품사 표제어수	58,010	19.09
전 체 합	303,951	100.00

한자가 출현하는 303,951개의 표제어 중 9품사를 가지는 표제어수는 245,941개 전체 표제어의 55.82%이고, 한자가 출현하는 표제어 303,951개 중 비품사 표제어는 58,010개 나타났다. 한자가 출현하는 표제어의 대부분은 명사로 한자가 출현한 전체 표제어들의 79.96%를 차지하였다. 조사는 한자가 전혀 나타나지 않았다.

7) 품사별 한자 출현 비율

대사전을 구성하는 440,594개의 표제어에서 9품사 및 비품사 표제어들에서 한자가 출현하는 비율을 조사하였다. 440,594개의 표제어 중 303,951개의 표제어에서 한자가 출현하여 68.99%의 한자 출현율을 나타냈다. 품사별 한자 출현 비율은 [표 9]와 같다.

표 9. 품사별 한자 출현 비율

구분	표제어수(A)	한자출현 표제어 수(B)	한자포함 비율(%) B/A
명사	334,962	243,030	72.55
대명사	465	245	52.69
수사	277	90	32.49
동사	15,173	420	2.77
형용사	6,446	1,264	19.61
관형사	529	195	36.86
부사	14,093	796	5.65
감탄사	812	64	7.88
조사	357	0	0.00
9품사 소계	373,114	246,104	65.96
9품사표제어수	371,904	245,941	66.13
비품사 표제어수	68,690	58,010	84.45
표제어수 합	440,594	303,951	68.99

9품사에 속한 371,904개의 표제어 중 245,941개의 표제어에서 한자가 출현하여 9품사의 한자 출현 비율은 66.13%로 나타났다. 9품사를 제외한 69,690개의 비품사 표제어 중 58,010개의 표제어에서 한자가 출현하여 비품사표제어의 한자 출현율은 84.45%로 나타났다. 9품사에 대한 품사별 한자 출현 비율은 명사 72.55%, 대명사 52.69%이며, 나머지 품사는 40%미만이다. 한자가 출현하는 대부분은 명사이다.

8) 단어 전체가 한자인 표제어의 품사별 비율

"7) 품사별 한자 출현 비율"에서 한 글자 이상의 한자가 표제어에 나타나는 표제어 수는 303,951개 이였다. 이 절에서는 440,594개의 전체 표제어 중 표제어 문자열 전체가 한자인 표제어들을 조사하였다. 즉 하나의 표제어를 구성하는 문자열 전체가 한글과 한자로 나타나는 표제어들을 조사하였다. 9품사 및 9품사 이외 표제어 집단에서 표제어 문자열 전체가 한자인 표제어 수는 [표 10]과 같다.

표 10. 문자열 전체가 한자인 표제어의 품사별 비율

구분	표제어수(A)	한자출현 표제어 수(B)	한자포함 비율(%) B/A
명사	334,962	195,607	58.40
대명사	465	230	49.46
수사	277	85	30.69
동사	15,173	11	0.07
형용사	6,446	0	0.00
관형사	529	187	35.35
부사	14,093	492	3.49
감탄사	812	29	3.57
조사	357	0	0.00
9품사합	373,114	196,641	52.70
9품사 표제어수	371,904	196,641	52.87
비품사 표제어수	68,690	43,868	63.86
전체 표제어수	440,594	240,509	54.59

표제어 문자열 전체가 한자인 표제어 수는 전체 표제어 440,594개 중 240,509개로 54.59%로 나타났다. 9품사를 가지는 표제어는 371,904개 중 196,641개로 52.87%이며, 명사는 334,962개 중 195,607개로 58.40%이고 대명사 49.46%, 관형사 35.35%, 수사 30.69%가 표제어 문자열 전체가 한자인 표제어로 나타났다.

9) 음절 위치별 한자 출현 분포

이 절에서는 표제어를 구성하는 문자열들에 대하여 음절 위치별로 출현하는 한글 음절수와 한자음들의 출현 분포를 조사하였다. 표제어에 불완전 음절이 포함된 경우("ㄱ자쇠"-"__字__")에 완전 한글음절("자쇠")과 한자("字")는 포함하였다. 이런 까닭에 앞의 분석과 한자 출현 횟수에 다소간의 차이가 있다. 그러나 한글 표제어에 대하여 영문표기와 한자가 동시에 나타나는 경우("가고시마-Kagoshima[鹿兒島]")에는 한자의 출현 위치가 몇 번 째 위치인지 결정하기 모호한 면이 있어 이런 경우의 한자("鹿兒島")는 제외하였다. 완전한 한글 음절1,446,791개의 같은 음절 위치에 942,277개의 한자가 출현하여 음절 위치별 한자 출현 비율은 65.13%을 나타났다[표 11].

표 11. 음절위치별 한자 출현비율

음절위치	한글음절수	한자 출현수	한자출현비율(%)
1	438,369	272,314	62.12
2	431,545	315,926	73.21
3	290,283	190,051	65.47
4	169,598	103,770	61.19
5	67,822	36,343	53.59
6	29,477	14,508	49.22
7	11,626	5,547	47.71
8	4,790	2,210	46.14
9	1,913	885	46.26
10	783	388	49.55
11	338	186	55.03
12	143	94	65.73
13	57	32	56.14
14	22	9	40.91
15	11	7	63.64
16	8	4	50.00
17	5	3	60.00
18	1	0	0.00
합계	1,446,791	942,277	65.13

음절 위치별 한자가 가장 많이 나타나는 위치는 두 번째 음절로 한글 431454음절에서 315,926개의 한자가 출현하여 73.21%를 나타냈다. 첫 음절에서는 439,369개의 한글 음절에서 272,314개의 한자가 출현하여 62.12%를 나타냈다. 이는 두 번째 음절에서 더 많이 한자가 나타나고 있음을 보여주고 있다. 이와 같은 예는 북한어의 하나인 "일정한 지점을 중심으로 한 그 부근 일대" 뜻을 가지 단어 "가근방(__近方)"과 같이 첫음절에서는 한자가 나타나지 않지만 두 번째 위치에서는 한자가 출현하는 경우이다.

4. 결론

지금까지 국어대사전의 표제어에 나타나는 한자 정보에 대하여 알아보았다. 국어대사전의 주표제어 440,594개의 표제어 중 303,951개의 표제어에서 한 글자 이상의 한자가 출현하여 한자 출현 비율은 68.99%였다. 9품사를 가지는 표제어의 경우, 371,904개의 표제어 중 245,941개의 표제어에서 한자가 출현하여 9품사를 가지는 표제어들의 한자 출현 비율은 66.13%이다. 4,888개의 한자 코드를 한글음절로 변환했을 때 변환되는 한글음절수는 473음절이며 가장 많은 수의 한자가 변환되는 한글 음절은 '기'로 무려 64개의 한자가 변환되었다.

440,594개의 주표제어들을 대상으로 한글 전자사전을 구축했을 때 엔트리 수는 361,980개이고 361,980개의 엔트리에 사용된 한글 음절은 1,289,659개로 단어의 평균 음절길이는 3.56이다[10]. 440,594개의 주표제어에 출현하는 한자 수는 858,595개로 표제어당 명균 1.95개의 한자가 출현한다. 이는 표제어들의 평균 음절 3.56개 중 1.95개의 한자가 출현하므로 하나의 표제어에 출현하는 한자의 평균 비율은 54.78%이다. 이는 전체 표제어에 대한 한자 출현비율 68.99%와 14.21%의 차이는 보인다. 이는 일반 단어 중 한자 단어가가 차지하는 비율이 70%라는 일반적인 주장과 큰 차이를 보였다. 따라서 한자가 한 글자 이상 출현하는 단어의 비율은 68.99%이지만 음절 길이를 고려한다면 한자의 출현 비율은 54.78%로 볼 수 있다.

음절 위치별 한자 출현 비율은 첫 음절에서는 438,369개의 한글에서 272,314개의 한자가 출현하여 62.12%이

지만, 둘째 음절은 431,545개의 한글에서 315,926개의 한자가 출현하여 73.21%의 한자 출현 비율을 보였다. 주표제어에 출현하는 858,595개의 한자 음절은 4,659개의 한자 코드가 사용되므로 한자 글자는 평균적으로 184.3회 출현한다. 가장 많이 나타나는 한자는 '法'으로 5,183회이고 단 한 번 출현하는 한자 음절은 153개였고, 229개의 한자는 한 번도 출현하지 않았다.

표제어에 출현하는 858,595개의 한자들은 4,888개의 한자 코드 중 하나이고 4,888개의 한자 코드는 473개의 한글 음절 가운데 하나에 대응되므로 858,595개의 한자들은 473개의 한글 음절 중 하나에 대응되므로 한자에서 사용되는 음은 지극히 제한되어 사용되고 있음을 알 수 있다. 이는 한자자 뜻글자인 까닭에 음보다 훈이 발달된 요인이기도 하다. 즉, 한글은 음이 발단한 바면 한자는 훈(뜻)이 잘 발달한 까닭이다.

본 연구를 통해 얻어진 한자정보들은 교육용 한자 범위 선정의 참고 자료로 활용할 수 있고 한자정보처리 분야 등에 사용될 수 있다.

분석에 사용된 한자의 코드영역은 4,888개로 제한되어 이 범위를 벗어나는 한자들에 대한 분석은 이루어지지 않았다. 보다 객관적인 분석을 위해서는 모든 한자들로 영역을 넓혀서 이루어져야 할 것이다.

참 고 문 헌

[1] 유진희, 이종혁, 이근배, "형태소 분석과 언어 평가를 이용한 문자인식 후처리", 정보과학회 논문지(B), Vol.22, No.6, pp.880-891, 1995.

[2] 강승식, "*음절정보와 복수어 단어 정보를 이용한 한국어 형태소 분석*", 서울대학교 공학박사 학위 논문, 1993.

[3] 국립국어연구원, *표준국어대사전*, 두산동아출판사, 1999.

[4] 송재소, "한국의 한자교육", 새국어생활 Vol.9. No.2, pp.125-144, 1999.

[5] 이용주, "한자 정책 현안으로서의 한자 폐지", 국어 생활, 90 봄(20호), pp.11-31, 1990.

[6] 박양규, "국어정책", 국어학연감, Vol.2000, pp.21-39, 2000.

[7] 박천서, "한글 『專用』 정책과 그 功過", 어문 연구, Vol.27 No.2, 1999.

[8] 김영환, "한자 혼용론을 논박함 한글 전용론의 깊은 뜻", 배달말학회 논문지, No.41, pp.33-52, 2007.

[9] 심재기 "국한자 혼용의 타당성에 관한 연구", 관악어문연구, Vol.23, No.1, pp.5-39, 1998.

[10] 김철수, 김양범, "대용량 전자사전 구축을 위한 국어 대사전의 통계정보", 한국콘텐츠학회 논문지(B), Vol.7, No.6, pp.60-68, 2007.

[11] 최주열, "한자 교육 방법에 관한 고찰", 한글말 교육 논문지, Vol..5, pp.145-174. 1994.

[12] 오미선, "漢字環境의 實態와 學習·敎育", 일본연구, No.21, 2003.

저 자 소 개

김 철 수(Cheol-Su Kim) 정회원

- 1987년 2월 : 전북대학교전산통계학과(이학사)
- 1989년 2월 : 전북대학교전산통계학과(이학석사)
- 1998년 8월 : 전북대학교전산통계학과(이학박사)
- 1995년 3월 ~ 현재 : 서남대학교 컴퓨터정보통신학과 교수

<관심분야> : 자연어처리, 정보검색, 지식표현, U-Learning