국어사전 표제어의 한자 빈도

남윤진
日本 神田外語大學 韓國語學科 傳任講師

1. 서　　론

1.1. 연구의 목적

한자는 수천 년 동안 우리 문자 생활의 중심이 되어온 문자로서 정신문화의 형성과 계승의 주된 수단이 되어 왔다. 따라서 사학, 철학이나 한문학 등 인문학의 제반 분야를 연구하기 위해서는 한자에 대한 정확하고 풍부한 지식이 필수적인 것으로 인식되고 있다. 또한 최근의 문자 생활이 한글 전용으로 나아가는 추세이기는 하지만, 한자는 아직도 우리 국어생활에서 커다란 비중을 차지하고 있다. 이는 사전에 등재된 국어 어휘의 절대 다수를 한자어가 차지하고 있다는 점, 그리고 시시때때로 변화하는 사회적 여건에 따라 끊임없이 만들어지는 신조어 및 약어의 많은 숫자가 한자어라는 점 등을 통해 드러난다. 이것은 모두 한자가 가지는 풍부한 재원과 막강한 조어력에 기인하는 것으로 이에 대한 지식이 없이는 정확하고 효율적이며 풍부한 국어생활을 영위하기가 어려운 것이다.

본 연구에서는 이러한 인식을 바탕으로, 국어 사전에 등재된 한자어의 한자 빈도를 조사함으로써 일반적인 국어생활에 사용되는 한자의 중요도를 판정하는 객관적인 자료를 제시하는 것을 목적으로 삼는다. 이러한 작업은 다음과 같은

측면에서 그 의의를 찾아볼 수 있다.

첫째, 한자와 관련된 국어 정책 및 국어 교육, 나아가 한문 교육에 유용한 지침을 마련하게 된다. 한자 빈도 조사가 이루어진 중국의 경우를 보면 빈도 순위 1,000에 해당하는 글자가 '孫'이며 이는 누적 빈도 95.36%에 해당한다는 결과를 얻고 있어서 중국어 문장을 95% 정도 이해할 수 있기 위해서 알아야 할 한자의 숫자는 1,000개 정도라는 추정이 가능하다. 이러한 사실은 한자 빈도 조사 결과가 중국어의 기본 어휘 선정에 중요한 기준이 됨을 의미한다.

한자의 빈도 조사가 중국어에서 갖는 의의와 국어에서의 그것은 동일하게 다루어질 수 없는 것이겠지만, 국어에서 한자어가 차지하는 중요도를 감안해 볼 때, 기본 한자, 즉 상용 한자의 선정은 국어 정책이나 국어 교육에서 결코 가벼이 다루어질 수 없는 사안이며, 이의 결정 혹은 평가에 본 연구의 결과가 유용한 기준으로 이용될 수 있을 것으로 생각된다.

둘째, 국어 어휘론 특히 한자어 연구의 기초적인 자료를 제공한다. 본 연구를 통해 얻게 되는 한자의 목록 및 빈도는 사전 표제어를 구성하는 한자를 바탕으로 한 것이며 각 한자의 단순 빈도와 아울러 각 한자어(사전 표제어) 내에서의 위치별 빈도나 한자어 구성상의 특성에 따른 빈도도 제시될 것이다. 이러한 조사 결과는 한자어 단어 형성론이나 의미론 연구 등에 유용한 자료가 될 것으로 생각된다.

셋째, 한자의 전산 처리에 필요한 일차적인 언어 정보를 구비하게 된다. 이는 국제적인 정보 교환에 적극적으로 참여할 수 있는 기본적인 여건을 조성하게 됨을 의미한다. 즉, 국제 표준 문자 코드 (uni-code)의 국제 공용 한자 20,902자를 중요도별로 분류할 수 있게 되어 이의 활용 방침을 표준화할 수 있게 된다.

1.2. 연구의 내용

본 연구에서는 국어 사전 표제어에 사용된 한자의 빈도를 조사하기 위하여 빈도 조사의 대상이 되는 사전으로 현재 국립국어연구원에서 편찬하고 있는 『표준 국어 대사전』(이하 대사전)을 정하였다. 이를 대상으로 한자의 빈도를 구하는 작업은 다음과 같은 두 가지 방향에서 이루어지게 된다.

① 한자의 단순 빈도 조사

한자어를 구성하는 각 한자의 빈도가 조사되어 한자어 구성에 사용되는 한자 전체의 목록과 빈도수, 빈도 순위, 사용률(누적 빈도 비율) 등이 얻어지게 된다. 이 목록은 한자음의 가나다순에 따라 배열된 것과 빈도순에 따라 배열된 것의 두 가지 형태로 제시되며 빈도순 배열에는 누적 빈도수 및 누적 빈도 비율이 덧붙여질 것이다.

② 한자어의 어휘 특성에 따른 한자의 빈도 조사

중국의 문자인 한자가 어휘를 구성하여 국어의 질서에 편입되는 과정에서 그 용법이나 사용 범위가 변화를 겪게 되며, 또 그러한 과정을 통해 형성된 한자어가 고유어와는 다른 어휘 특성을 지니게 됨은 그동안의 국어 연구를 통하여 지적된 바 있다. 본 연구에서는 이러한 국어 연구의 성과를 검증하고 확대한다는 의미에서 다음과 같은 작업을 수행한다.

첫째, 한자어의 음절수에 따른 한자의 빈도를 조사하여 그 분포를 밝히는 작업이다. 일반적으로 한자어는 '山, 江, 冊' 등과 같은 특수한 경우를 제외하면 2개 이상의 한자로 구성됨이 지적되고 있다. 따라서 본 연구에서는 1음절 한자어를 이루는 한자는 어떤 것이 있는지를 조사하고, 나아가서 2음절 이상의 한자어에 대해서는 각 단어 내에서의 위치별로 분포를 조사하여 제시하고자 한다.

둘째, 국어에는 하나의 어휘로 볼 것인지 어휘 이상 혹은 어휘 이하의 단위로 볼 것인지가 애매한 구성들이 다수 존재한다. 특히 한자어와 밀접한 관련이 있는 것으로는 고사성어나 전문용어, 준말 등을 들 수 있다. 여기서는 이러한 구성들을 그 특징에 따라 분류하고 특히 전문용어와 일반어, 단어와 단어 이외의 구성을 구별하여 각 범주별로 한자의 빈도를 조사하고 그 분포상을 제시한다.

2. 대사전 표제어의 한자 빈도 (I) : 단순 빈도

2.1. 대사전 표제어의 특성

앞에서도 언급한 바 있듯이 본 연구는 국어연구원에서 편찬중인 『표준 국어대사전』을 대상으로 하고 있다. 이 사전은 국어에 사용되는 모든 어휘를 망라할 것을 목표로 하는 것으로서 표제어 구성에 있어 전문용어와 고유명사, 고어와

북한어 등이 포함되어 있다. 그 결과 언어학적인 관점에서 볼 때 단어가 아닌 구성들, 즉 구나 숙어, 속담, 고사성어 등 다양한 언어 단위들이 등재되고 있다.

어떤 대상에서 통계적으로 의미 있는 빈도 정보를 얻어내기 위해서는 그 대상의 구성과 특성을 잘 파악하고 대상의 성격에 맞는 인자들을 정하여 빈도를 구하여야 한다. 본 연구는 이러한 원칙에 따라 앞에서 언급한 대사전의 성격을 고려하여 다음과 같이 여러 각도에서 한자의 빈도를 추출하였다.

첫째, 대사전에는 '단어' 뿐만 아니라 구나 절 등이 표제어로 실린 경우가 있다. 따라서 본 연구에서는 이들 언어 단위의 성격에 따라 별도의 빈도를 구한다. 즉 전체 표제어에 나타난 한자의 빈도와 단어 표제어에 나타난 한자의 빈도를 별도로 구한다. 특히 단어를 구성하는 한자의 빈도가 국어 어휘에 사용되는 한자의 양상을 보여주는 기본적인 자료가 될 것이라고 보고, 이를 다각도로 분석하여 빈도를 제시한다.

둘째, 대사전에는 일반 어휘뿐만 아니라 전문용어도 포함하고 있다. 그런데 일반 어휘와 전문용어는 그 구성방식이나 사용 범위 등에서 차이를 보이기 때문에 한자의 사용 양상도 각각 다르게 나타날 것으로 추정된다. 따라서 이들을 구별하여 빈도를 제시한다.

셋째, 대사전은 북한 사전의 표제어도 다수 포함하고 있다. 그런데 북한은 독자적인 '말다듬기' 작업을 추진함으로써 남한과의 언어 이질화가 상당히 진행된 상태이며 이는 특히 조어법 부문에서 두드러지는 것으로 생각된다. 이러한 점에 착안하여 본 연구에서는 북한어에 사용된 한자의 빈도를 별도로 제시한다. 한편, 대사전에서는 북한어의 표제어에는 띄어쓰기 정보나 구 표지 등을 보이지 않고 있다. 이는 북한어에 한해서는 북한 사전의 표기방식을 따른다는 원칙에서 비롯한 것이다. 이런 이유로 하여, 단어와 단어 이상의 단위를 구별하여 빈도를 제시하는 과정에서는 북한어 표제어를 제외하였다. 북한어 표제어에 대하여 그것이 단어인지 구인지 연구자 개인이 판단을 내리는 것은 무의미하다고 판단되었기 때문이다.

2.2. 전체 표제어에 사용된 한자의 빈도

앞 절에서 언급하였듯이 대사전은 다양한 성격을 지니는 표제어들로 구성되어 있어서 의미 있는 통계 정보를 얻기 위해서는 사전 표제어 전체의 빈도를 구할 뿐만 아니라 사전 표제어를 어휘적 특성에 따라 분류하고 각 어휘 범주별로 한자의 빈도를 구해야 할 것이다. 이러한 맥락에서 본 절에서는 먼저 사전 표제어 전체의 빈도를 제시하고자 한다.

대사전의 항목 가운데 표제어에 한자를 포함한 항목은 모두 316,721 항목이었다. 이들 표제어를 구성하는데 사용된 한자는 모두 7,310자였으며 이들 7,310자의 총 사용 횟수는 910,849회였다. 이 수치에는 '상하이'라는 표제어에 대하여 [上海]라는 한자 어원이 제시되는 경우처럼 중국이나 일본의 地名, 人名 등을 原語의 발음대로 읽고, 어원표시를 위해 사용된 한자는 제외된 것이다. 반면, '나무아미타불'이라는 표제어를 '南無阿彌陀佛,'로 표기할 때 사용되는 '南'과 같이 외래어 표기를 위해 원래의 음과 달리 사용된 것이나, '사탕'이라는 표제어의 어원 '砂糖'과 같이 한자어를 구성하는 과정에서 한자가 音變化를 겪어 원래의 음으로부터 멀어진 경우는 포함되었다.

2.2.1. 가나다순 목록

대사전 표제어 한자의 가나다순 목록 가운데 ㄱ항의 일부만을 아래에 보이기로 한다. 이 목록에서 '가0, 간1' 등으로 표기된 글자들은 문서편집기 '흔글'로 구현되지 않는 글자들이다.

[표1] 대사전 표제어 한자의 가나다순 목록 - 일부

茄, 價, 呵, 猳, 蚵, 可, 珈, 咖, 加, 歌, 痂, 笳, 嫁, 椵, 架, 袈, 軻, 斝, 街, 坷, 跏, 枷, 哥, 葭, 家
柯, 訶, 稼, 佳, 駕, 嘉, 舸, 伽, 苛, 假, 暇, 迦, 榎, 珂, 檟, 駕, 賈, 가0, 殼, 各, 覺, 擱, 桷, 恪, 閣,
咯, 却, 搉, 刻, 埆, 榷, 脚, 慤, 角, 杆, 鐗, 趕, 玕, 姦, 榦, 癇, 間, 慳, 艮, 간1, 齦, 諫, 看, 柬, 幹,
奸, 䵟, 盰, 澗, 旰, 茛, 衎, 干, 稈, 竿, 刊, 簡, 奸, 侃, 肝, 懇, 墾, 桿, 馯, 揀, 艱, 癎, 간2, 葛, 曷,
褐, 喝, 鶡, 楬, 暍, 渴, 羯, 乫, 噶, 鞨, 竭, 蝎, 碣, 敢, 嵌, 柑, 淦, 憨, 泔, 轗, 勘, 莰, 嵁, 疳, 龕
撼, 椷, 橄, 坎, 欿, 堪, 鑑, 輡, 埳, 鹻, 減, 坩, 减, 鑒, 邯, 紺, 甘, 戡, 感, 酣, 憾, 監, 瞰, 甲, 岬,
胛, 匣, 閘, 玒, 糠, 康, 罡, 羌, 蜣, 鋼, 崗, 疆, 講, 岡, 殭, 江, 杠, 鱇, 堈, 橿, 慷, 降, 姜, 强, 綱,
彊, 絳, 鏹, 剛, 羗, 薑, 扛, 繈, 掆, 椌, 腔, 襁, 僵, 강1, 礚, 鎧, 疥, 丐, 塏, 凱, 槪, 介, 個, 槩, 改,
皆, 箇, 開, 愾, 匃, 闓, 揩, 愷, 价, 慨, 漑, 蚧, 芥, 蓋, 개0

2.2.2. 빈도순 목록

대사전 표제어에 사용된 한자를 빈도순으로 정리한 목록의 일부를 보이면 [표2]와 같다. 이를 통해, 대사전 표제어에 가장 많이 사용된 한자는 5,319회 사용된 '法'이며, '學, 性, 大, 子, 物, 地, 人, 動, 生…'의 순으로 높은 사용률을 보임을 알 수 있다.

이 목록에서 原音빈도란, 각 한자가 원래의 음으로 사용된 빈도를 나타내며, 異音빈도는 앞에서 설명한 바와 같이 차용이나 한자어 구성 과정에서 변화된 음으로 사용된 빈도를 나타낸다. 전체빈도는 原音빈도와 異音빈도의 합으로서 각 한자가 사용된 전체 횟수를 나타낸다.

[표2] 대사전 표제어 한자의 빈도순 목록 - 상위 10%

순 위	한자	原音빈도	異音빈도	전체빈도	누적빈도	누적비율
1	法	5,317	2	5,319	5,319	0.58
2	學	5,014		5,014	10,333	1.13
3	性	4,786		4,786	15,119	1.66
4	大	4,539	3	4,542	19,661	2.16
5	子	4,319	2	4,321	23,982	2.63
6	物	4,119		4,119	28,101	3.09
7	地	4,037		4,037	32,138	3.53
8	人	3,956	3	3,959	36,097	3.96
9	動	3,925	1	3,926	40,023	4.39
10	生	3,858	2	3,860	43,883	4.82
11	水	3,838		3,838	47,721	5.24
12	金	3,731	1	3,732	51,453	5.65
13	機	3,729		3,729	55,182	6.06
14	化	3,662		3,662	58,844	6.46
15	山	3,530		3,530	62,374	6.85
16	國	3,484		3,484	65,858	7.23
17	氣	3,293		3,293	69,151	7.59
18	的	3,257	3	3,260	72,411	7.95
19	電	3,193		3,193	75,604	8.30
20	中	3,174	1	3,175	78,779	8.65
21	體	3,089		3,089	81,868	8.99
22	行	2,888		2,888	84,756	9.30
23	主	2,881		2,881	87,637	9.62
24	分	2,866		2,871	90,508	9.94
25	文	2,829	5	2,829	93,337	10.25

이 목록은 빈도가 높은 한자부터 순차적으로 배열한 것으로, 이를 통해 각 한자의 빈도 순위를 확인할 수 있다. 이러한 빈도순 배열을 통하여 누적 빈도와 누적비율을 구할 수 있다. 누적 빈도는 최상위 빈도의 한자부터 해당 한자까지의 빈도의 총합으로 빈도수 최하위의 한자에 이르면 누적빈도는 전체 한자의 총 사용 횟수(910,849회)와 일치하게 된다. 누적비율은 해당 한자의 누적빈도를 총 사용 횟수로 나누어 100을 곱한 수치로서, 바꾸어 말하면 최상위 빈도 한자로부터 해당 한자까지의 사용 횟수가 전체 사용 횟수의 몇 퍼센트에 해당하는가를 보여주는 항목이다.

누적비율은 빈도조사 결과를 응용하고자 할 때 가장 기초적인 통계자료로 이용되는 항목으로, 이를 통해서 우리는 '사전 표제어에 사용된 한자의 90퍼센트에 해당하는 한자의 수는 1,589자이며 95%에 해당하는 한자는 2,256자'라는 사실을 알게 되는데 이러한 사실을 바탕으로 '대사전의 표제어를 구성하는 한자를 모두 알기 위해서는 7,310자의 한자를 학습하여야 하지만, 90%의 학습성취도를 목표로 할 때는 1,589자를, 95%를 목표로 할 때는 2,256자의 한자를 학습하면 가능하다'는 식의 결정을 내릴 수 있다.

이러한 결과는 국어 교육 내지 한자 교육의 내용과 量的 지표를 정하는데 객관적인 근거로 활용될 수 있다. 현재의 한자교육에서는 1,800자의 '漢文敎育用 基礎漢字'를 정하고 있지만 기초 한자를 1,800자로 정한 근거나 개별 한자를 기초 한자로 선정한 객관적인 기준은 제시되지 않고 있다. 이러한 부분에서 본 연구에서 제시되는 빈도 목록과 같은 자료의 통계정보를 활용할 수 있게 된다면 상당 부분의 문제점이 해소될 것으로 생각된다. 사전 표제어 한자 목록에서 빈도 순위가 상위 1,800에 해당하는 한자는 78회 이상의 빈도를 가지며 누적 비율은 92.03%까지인 글자들임을 알 수 있다. 즉 교육부 제정 기초 한자의 수량인 1,800자는 대사전에 실린 표제어 한자의 92% 가량을 포괄하는 것이 된다. 이는 1,800자의 한자를 익히면 사전 표제어의 한자를 92%정도 이해할 수 있음을 의미한다. 그런데 교육부 제정의 기초 한자와 대사전 표제어 한자의 빈도순위 1,800자는 적지 않은 相違를 보인다. 즉 기초 한자에 포함된 한자가 사전 표제어에는 쓰이지 않은 경우(劒 등), 사전 표제어에서는 빈도순위 1,800위 이하에 해당하는 경우(暇, 簾, 陋, 栗, 肩 등)가 다수 있으며, 사전 표제어의 한자 빈도 순위 1,800위

이상인 한자가 기초 한자에 포함되지 않은 경우(伽, 姜, 闕, 掘, 圈 등)도 있는 것이다. 따라서 기초 한자 1,800자의 학습 성취와 사전 표제어의 이해도는 일치하지 않는다.

위와 같은 작업을 통해 알 수 있듯이, 교육용 한자 선정이 보다 객관적이고 타당성 있는 선정 근거를 제시할 수 있기 위해서는 본 연구에서 제시하는 것과 같은 빈도 목록이 필수적이라 할 수 있을 것이다.

3. 대사전 표제어의 한자 빈도 (Ⅱ) : 어휘 특성별 빈도

3.1. 전문용어와 일반어의 한자 빈도

대사전의 한자어 표제항 가운데 전문용어는 170,369 항목, 일반어는 146,352 항목이었다. 이들 각각의 한자 빈도를 조사한 결과 이들 어휘의 성격에 따라 한자의 분포가 다른 양상을 보여서, 전문용어에 사용된 한자는 5,774종임에 비하여 일반어에 사용되는 한자는 6,170 종이었다. 이는 전문용어보다 일반어에 사용되는 한자의 종류가 훨씬 다양함을 의미하는 것으로, 특정 분야에 한정된 내용을 전달하는 어휘라는 전문용어의 속성이 반영된 것으로 해석된다. 또 누적빈도 90%인 한자의 종류는 일반어가 1,800 종, 전문용어가 1,290 종임이 밝혀졌다. 빈도와 다양도의 상관관계는 어휘 사용의 집중도를 의미하는 것이라고 해석할 때, 전문용어에 사용된 한자의 집중도는 일반어에 사용된 한자보다 더 높다고 할 수 있을 것이다. 한편, 상위 빈도를 보이는 한자의 종류에서도 두 부류는 차이를 보인다. 전문용어를 구성하는 한자에서 '法, 學, 性' 등이 빈도 1, 2, 3위를 차지한 데 반하여 일반어에서는 '人, 之, 大' 등이 그에 해당하는 것이다. 전문용어와 일반어의 상위빈도 한자의 일부를 표로 보이면 다음과 같다.

[표3] 전문용어의 한자 빈도-상위 10%

순위	한자	빈도	누적빈도	누적비율	순위	한자	빈도	누적빈도	누적비율
1	法	4,687	4,687	0.85	11	金	2,747	35,036	6.36
2	學	4,006	8,693	1.58	12	地	2,668	37,704	6.85
3	性	3,441	12,134	2.20	13	山	2,638	40,342	7.33
4	動	3,055	15,189	2.76	14	水	2,552	42,894	7.79
5	子	2,990	18,179	3.30	15	國	2,497	45,391	8.24
6	物	2,888	21,067	3.83	16	體	2,390	47,781	8.68
7	機	2,851	23,918	4.34	17	氣	2,387	50,168	9.11
8	化	2,817	26,735	4.85	18	生	2,231	52,399	9.51
9	大	2,785	29,520	5.36	19	分	2,137	54,536	9.90
10	電	2,769	32,289	5.86	20	線	2,123	56,659	10.29

[표4] 일반어의 한자 빈도 - 상위 10%

순위	한자	빈도	누적빈도	누적비율	순위	한자	빈도	누적빈도	누적비율
1	人	2,084	2,084	0.58	15	物	1,231	22,595	6.27
2	之	1,759	3,843	1.07	16	行	1,152	23,747	6.59
3	大	1,757	5,600	1.55	17	中	1,116	24,863	6.90
4	不	1,713	7,313	2.03	18	心	1,101	25,964	7.21
5	生	1,629	8,942	2.48	19	事	1,097	27,061	7.51
6	者	1,614	10,556	2.93	20	主	1,087	28,148	7.81
7	一	1,453	12,009	3.33	21	色	1,025	29,173	8.10
8	無	1,416	13,425	3.73	22	學	1,008	30,181	8.38
9	地	1,369	14,794	4.11	23	文	1,002	31,183	8.66
10	的	1,364	16,158	4.49	24	自	993	32,176	8.93
11	性	1,345	17,503	4.86	25	國	987	33,163	9.21
12	子	1,331	18,834	5.23	26	金	985	34,148	9.48
13	水	1,286	20,120	5.59	27	天	965	35,113	9.75
14	家	1,244	21,364	5.93	28	日	932	36,045	10.01

3.2. 북한어 표제어에 사용된 한자의 빈도

한자를 어원으로 가진 북한어는 44,240 항목이 등재되어 있는데, 여기에 사용된 한자는 3,833자이며 이들의 총 사용 빈도는 136,643회이다. 이들의 상위빈도 한자는 '機, 動, 性, 地, 物' 등으로 전문용어의 상위 빈도 한자들과 상당 부분 일치를 보이고 있다. 이와 같은 사실은 대사전에 등재된 북한 한자어가 주로 전문용어이거나 일반어라 하더라도 남한에서 사용되지 않거나 쓰임이 다른 것들에 한정된다는 점을 고려해 볼 때 당연한 결과라고 할 것이다. 북한어 표제어에 사용된 한자를 상위 빈도에 해당하는 한자에 한하여 제시하여 보면 다음과 같다.

[표3] 북한어 표제어 한자 빈도 - 10%

순위	한자	빈도	누적빈도	누적비율	순위	한자	빈도	누적빈도	누적비율
1	機	1,524	1,524	1.12	17	工	673	14,131	10.34
2	動	1,121	2,645	1.94	18	化	662	14,793	10.83
3	性	985	3,630	2.66	19	器	628	15,421	11.28
4	地	850	4,480	3.28	20	分	587	16,008	11.71
5	物	849	5,329	3.90	21	合	574	16,582	12.13
6	電	792	6,121	4.48	22	大	569	17,151	12.55
7	式	789	6,910	5.06	23	力	556	17,707	12.96
8	線	777	7,687	5.63	24	中	537	18,244	13.35
9	學	769	8,456	6.19	25	理	523	18,767	13.73
10	體	753	9,209	6.74	26	形	498	19,265	14.10
11	法	742	9,951	7.28	27	色	480	19,745	14.45
12	子	723	10,674	7.81	28	度	477	20,222	14.80
13	水	720	11,394	8.34	29	車	473	20,695	15.14
14	生	692	12,086	8.84	30	人	462	21,157	15.48
15	的	691	12,777	9.35	31	數	452	21,609	15.81
16	氣	681	13,458	9.85	32	間	448	22,057	16.14

순위	한자	빈도	누적빈도	누적비율	순위	한자	빈도	누적빈도	누적비율
33	行	446	22,503	16.47	40	文	389	25,441	18.62
34	音	442	22,945	16.79	41	山	378	25,819	18.89
35	主	432	23,377	17.11	42	勞	374	26,193	19.17
36	流	428	23,805	17.42	43	壓	373	26,566	19.44
37	業	427	24,232	17.73	44	作	371	26,937	19.71
38	金	414	24,646	18.04	45	反	369	27,306	19.98
39	自	406	25,052	18.33	46	連	358	27,664	20.24

3.3. 고유명사의 한자 빈도

대사전에는 12,472종의 고유명사가 등재되어 있으며 그 가운데 단어인 것은 10,470개이다. 이들은 인명, 지명이 대부분이며 적지만 책명, 고적명 등이 포함되어 있다. 본 연구에서는 고유명사에 사용된 한자의 빈도와 아울러 인명과 지명에 사용된 한자를 조사하였다. 한자를 이용하여 인명이나 지명을 짓는 것은 가장 일반화된 형태의 한자어 형성 과정임에도 불구하고 이에 대한 실태조사나 그 기제에 대한 국어학적 연구는 활발히 이루어지지 못했다. 사전에 등재된 인명이나 지명은 고유명사이지만 지명도가 상대적으로 높은 것인 만큼 타당성과 객관성이 확보된 고유명사 자료라고 할 수 있다. 대사전 수록 고유명사에 사용된 한자의 종류는 모두 2,804 종이며 인명에 사용된 한자는 1,942종, 지명에 사용된 것은 1,423종이었다. 상위 빈도를 차지하는 한자는 인명의 경우 '金, 李, 王, 鄭, 宗' 등 姓이나 왕명과 관련된 것이었으며 지명에서는 '山, 島, 江, 峯, 郡' 등 지형이나 행정구역과 관련된 한자들이었다. 다음 [표6~8]로 이들의 목록 일부를 제시한다.

[표6] 고유명사 한자 빈도 목록 - 상위 20%

순위	한자	빈도	누적빈도	누적비율	순위	한자	빈도	누적빈도	누적비율
1	山	1,101	1,101	3.05	11	國	255	5,346	14.81
2	島	875	1,976	5.47	12	海	246	5,592	15.49
3	金	575	2,551	7.07	13	東	241	5,833	16.16
4	李	455	3,006	8.33	14	郡	235	6,068	16.81
5	大	434	3,440	9.53	15	州	235	6,303	17.46
6	王	374	3,814	10.56	16	文	218	6,521	18.06
7	江	362	4,176	11.57	17	湖	200	6,721	18.62
8	世	359	4,535	12.56	18	德	175	6,896	19.10
9	峯	292	4,827	13.37	19	三	168	7,064	19.57
10	南	264	5,091	14.10	20	錄	167	7,231	20.03

[표7] 인명의 한자 빈도 목록 - 상위 20%

순위	한자	빈도	누적빈도	누적비율	순위	한자	빈도	누적빈도	누적비율
1	金	494	494	3.44	10	趙	95	2,331	16.22
2	李	447	941	6.55	11	朴	91	2,422	16.85
3	王	343	1,284	8.93	12	子	91	2,513	17.49
4	世	323	1,607	11.18	13	張	90	2,603	18.11
5	鄭	140	1,747	12.16	14	權	88	2,691	18.72
6	大	137	1,884	13.11	15	安	87	2,778	19.33
7	一	128	2,012	14.00	16	柳	86	2,864	19.93
8	宗	128	2,140	14.89	17	師	78	2,942	20.47
9	文	96	2,236	15.56					

[표8] 지명의 한자 빈도 목록 - 상위 40%

순위	한자	빈도	누적빈도	누적비율	순위	한자	빈도	누적빈도	누적비율
1	山	1,019	1,019	7.72	14	嶺	127	4,254	32.25
2	島	871	1,890	14.33	15	川	118	4,372	33.14
3	江	347	2,237	16.96	16	東	112	4,484	33.99
4	峯	269	2,506	19.00	17	國	110	4,594	34.82
5	郡	230	2,736	20.74	18	德	101	4,695	35.59
6	州	229	2,965	22.48	19	灣	97	4,792	36.33
7	大	197	3,162	23.97	20	白	96	4,888	37.05
8	海	191	3,353	25.42	21	西	95	4,983	37.77
9	湖	189	3,542	26.85	22	城	94	5,077	38.49
10	南	158	3,700	28.05	23	平	94	5,171	39.20
11	脈	149	3,849	29.18	24	北	90	5,261	39.88
12	地	141	3,990	30.25	25	水	78	5,339	40.47
13	諸	137	4,127	31.28					

3.4. 한자 단어 표제어에 사용된 한자의 빈도

앞에서도 언급하였듯이 대사전은 '학교, 가정, 구체제, 초등학교' 등과 같은 단어는 물론, '부관 참시, 아르키메데스의 법칙' 등의 구, '잃어버린 시간을 찾아서' 등의 절과 같이 다양한 언어 단위를 표제어로 올리고 있다. 따라서 이러한 언어 단위에 대한 고려 없이 일괄적으로 추출된 빈도정보만으로는 국어의 어휘 구성에 있어서 한자가 차지하는 비중 및 기능을 왜곡할 우려가 있다. 따라서 본 절에서는 이러한 어휘 단위의 특성에 따라 표제어를 분류하고 그 가운데 단어로 판단되는 표제어에 대해서 다음과 같은 빈도 조사 작업을 진행하였다[1].

1) 앞에서도 언급하였듯이 이 작업에서는 북한어를 배제하였다. 북한어 항목에 관한 한 대사전에서는 언어 단위에 대한 정보를 싣지 않고 있어서(물론 이는 북한어의 모습을 그대로 제시한다는 취지에서 비롯한 것이다), 이 정보를 새로이 확보하자면 많은 작업이 필요한데 이는 본 연구의 범위를 벗어나는 것일 뿐만 아니라 왜곡된 정보를

3.4.1. 한자 단어 내에서의 위치별 한자 빈도

학교 문법의 기준에 따라 단어로 처리되는 한자어 표제어의 수는 213,000종이었다. 여기에 사용된 한자는 7,197자, 총 빈도는 524,163회였다. 상위 빈도의 한자로는 '大, 人, 法, 山, 子, 金' 등이 있으며 누적빈도 90%를 이루는 한자는 1,888 종, 95%를 이루는 한자는 2,641 종이었다. 상위 빈도를 보이는 한자의 목록을 제시하면 [표9]와 같다.

본 연구에서는 이들의 총 빈도를 구하는 데서 나아가, 한 단어 내에서 각 한자가 놓이는 위치에 따른 빈도를 구하였다. 구체적 내용을 살펴보면, '出, 身, 交, 血, 油, 政, 防…' 등의 한자는 단독으로는 쓰이지 못하며, '枸, 橢…' 등의 한자는 어두에만 사용되었다. 또 누적 비율 95%의 한자 가운데 '假, 姜, 旣, 劉, 李, 竝, 普, 複, 吳, 再, 鄭, 粗, 趙, 超, 最, 崔, 追' 등은 어두에서 사용되는 비율이 각 한자의 사용 빈도의 90% 이상을 차지하며, '匣, 畓, 島, 餠, 者, 的, 劑, 症, 峙, 混, 膾' 등은 어말에 쓰이는 비율이 90% 이상이고 '而'는 어중에 사용되는 비율이 90%이상이었다.

[표9] 단어 표제어의 한자 목록 - 상위10%

한자	1음절어	어두	어말	어중	1음절비	어두비	어말비	어중비	합계	누적비
大	3	1,918	151	621	0.11	71.22	5.61	23.06	2,693	0.51
人	3	535	1,250	595	0.13	22.45	52.45	24.97	2,383	0.97
法	3	358	1,756	212	0.13	15.37	75.40	9.10	2,329	1.41
山	2	780	978	533	0.09	34.02	42.65	23.24	2,293	1.85
子	4	148	1,253	815	0.18	6.67	56.44	36.71	2,220	2.27
金	7	1,281	539	305	0.33	60.08	25.28	14.31	2,132	2.68
水	3	783	577	762	0.14	36.85	27.15	35.86	2,125	3.08
性	2	114	1,491	340	0.10	5.86	76.58	17.46	1,947	3.46
之	1	5	120	1,810	0.05	0.26	6.20	93.49	1,936	3.83
生	5	711	528	675	0.26	37.05	27.51	35.17	1,919	4.19

제공할 우려가 있기 때문이다.

한자	1음절어	어두	어말	어중	1음절비	어두비	어말비	어중비	합계	누적비
不	3	1,165	7	736	0.16	60.96	0.37	38.51	1,911	4.56
一	2	1,151	79	544	0.11	64.81	4.45	30.63	1,776	4.89
化	4	174	682	916	0.23	9.80	38.40	51.58	1,776	5.23
科	3	95	1,566	94	0.17	5.40	89.08	5.35	1,758	5.57
無	2	1,250	32	440	0.12	72.51	1.86	25.52	1,724	5.90
中	3	905	334	432	0.18	54.06	19.95	25.81	1,674	6.22
地	2	458	770	429	0.12	27.61	46.41	25.86	1,659	6.53
者	2	2	1,528	125	0.12	0.12	92.21	7.54	1,657	6.85
主	3	392	318	934	0.18	23.80	19.31	56.71	1,647	7.16
學	1	218	871	549	0.06	13.30	53.14	33.50	1,639	7.48
三	1	1,230	30	325	0.06	77.55	1.89	20.49	1,586	7.78
石	3	459	724	327	0.20	30.34	47.85	21.61	1,513	8.07
天	2	803	325	379	0.13	53.21	21.54	25.12	1,509	8.35
物	3	144	952	387	0.20	9.69	64.06	26.04	1,486	8.64
文	6	465	547	448	0.41	31.72	37.31	30.56	1,466	8.92
國	3	467	554	361	0.22	33.72	40.00	26.06	1,385	9.18
心	2	312	658	407	0.15	22.63	47.72	29.51	1,379	9.44
行	6	335	505	485	0.45	25.17	37.94	36.44	1,331	9.70
色	1	203	748	359	0.08	15.48	57.06	27.38	1,311	9.95
事	2	172	689	416	0.16	13.45	53.87	32.53	1,279	10.19

그간의 한자어 연구에서 어떤 한자어가 단일어인가, 파생어인가, 복합어인가를 판별하는 기준을 찾기가 어려움이 지적되었으며 이런 문제에 대한 해결은 한자어를 구성하는 각 한자가 국어의 어휘형성에 참여하는 양상을 밝힘으로써만 이루어질 수 있을 것이라는 인식이 커져 왔다. 이러한 상황을 고려할 때, 본 연구에서 제시되는 한자의 단어내 위치 정보는 국어 어휘론 연구의 기초 자료로서 매우 유용하게 쓰일 수 있을 것이다.

3.4.2. 한자 이외의 요소와 결합하는 한자의 빈도

국어 단어의 형성소로서 한자는 다른 한자와 결합하여 한자어를 이룰 뿐만 아니라 고유어나 외국어와 결합하여 새로운 단어를 만들어낸다. 외국어와 한자가 결합하는 경우는 대개 신어의 생성과 밀접한 관련이 있으며, 고유어와 한자의 결합은 해당 한자의 국어화 정도를 반영하는 것이라는 추정이 가능한바, 이들의 결합 빈도 및 단어 내에서의 위치정보를 확보하게 된다면 이러한 연구에 상당한 도움을 줄 수 있을 것으로 기대된다. 이러한 점에 착안하여 본 연구에서는 고유어와 결합하는 한자와 외국어와 결합하는 한자의 결합 빈도를 조사하였다. 그 결과 고유어와 결합하여 단어를 형성하는 한자는 2,861종, 외국어와 결합하여 단어를 형성하는 한자는 867종이었으며 그 가운데 상위 빈도를 차지하는 것은 각각 '科, 子, 草, 大, 半', '酸, 化, 語, 族, 江' 등이었다. 상위 빈도를 갖는 한자들의 목록을 제시하면 다음 [표10~11]과 같다.

[표10] 고유어와 결합하는 한자의 빈도 목록 - 일부

순위	한자	어두	어말	어중	총빈도	순위	한자	어두	어말	어중	총빈도
1	山	7	4	14	25	14	半	6	3	9	18
2	色	5	4	16	25	15	方	4	2	12	18
3	紅	7	2	15	24	16	鬚	5	2	11	18
4	點	6	4	13	23	17	字	3	4	11	18
5	靑	6	4	11	21	18	長	6	3	9	18
6	金	6	4	10	20	19	天	6	2	10	18
7	子	1	6	13	20	20	蟲	2	5	11	18
8	王	7	2	10	19	21	花	3	5	10	18
9	將	4	4	11	19	22	黃	6	2	10	18
10	香	5	4	10	19	23	間	5	5	7	17
11	角	4	5	9	18	24	女	3	5	9	17
12	果	3	4	11	18	25	法	6	7	4	17
13	大	6	2	10	18						

[표11] 외국어와 결합하는 한자의 빈도 목록 - 일부

순위	한자	어두	어말	어중	총빈도	순위	한자	어두	어말	어중	총빈도
1	酸	32	194	340	566	15	素		16	35	51
2	化		45	277	322	16	族		49	2	51
3	鹽	74	30	17	121	17	基		48		48
4	法		104	1	105	18	世		45	1	46
5	黃	40	7	25	72	19	亞	26	1	19	46
6	水	17	12	42	71	20	機		42	2	44
7	主	2		66	68	21	石	6	36	2	44
8	油	4	55	7	66	22	窒	30		11	41
9	義		61	5	66	23	二	12		28	40
10	派		63		63	24	劑		40		40
11	病		58	1	59	25	子		30	8	38
12	管	1	46	6	53	26	三	21		16	37
13	種		52		52	27	人	4	26	6	36
14	線		45	6	51	28	紙		36		36

4. 결 론

지금까지 국립국어연구원에서 편찬 중인 『표준 국어 대사전』을 대상으로 사전 표제어의 한자 빈도를 살펴 보았다. 그 결과 대사전 표제어에는 총 7,310자의 한자가 사용되었음이 밝혀졌으며, 누적 사용율 90%에 해당하는 한자의 수는 1,589자, 95%에 해당하는 한자는 2,256자'라는 사실이 밝혀졌다. 대사전의 대표성과 포괄성을 고려할 때 일반적인 국어생활에 사용될 수 있는 한자어의 한자를 모두 알기 위해서는 7,310자의 한자를 학습하여야 하고, 90% 이상의 학습성취를 목표로 할 때는 약 1,600자를, 95%이상을 목표로 할 때는 약 2,300자의 한자를 학습하여야 한다는 결론을 얻을 수 있었다. 이는 전문용어와 일반용어, 구와 단어, 북한어, 고유명사 등을 모두 포함하여 얻은 빈도로서, 그 범위를 일반어 단어로 좁힌다면 동일한 학습성취에 필요한 한자의 수는 더 줄어들 수 있을 것이다.

이 외에 한자어의 어휘적 특성에 따라 전문용어와 일반용어, 북한어, 고유명사, 단어 등으로 나누어 각각의 빈도를 구하였으며, 그 결과 이러한 어휘의 특성에 따라 한자의 빈도 분포가 다른 양상을 보이고 있음을 알 수 있었다. 특히 사전에서 복합어나 파생어를 포함하여 단어로 처리된 한자어의 경우에는 각 한자의 위치에 따른 분포를 구하고, 한자+한자 이외의 구성에 나타나는 빈도를 구하였다.

이러한 빈도 조사 결과의 활용 방안을 살펴봄으로써 논의를 끝맺고자 한다.

① 교육용 기초 한자의 선정

국어 어휘 구성에서 한자어가 차지하는 비중을 고려해 볼 때 한자에 대한 지식은 필수적인 것이라 함은 이론의 여지가 없다 할 것이다. 지금까지 국어 교육에 있어서 한자 교육 여부가 찬반 논쟁의 중심이 되어 온 것은 국어의 현실을 올바르게 파악할 수 있는 구체적인 자료가 제시되지 못한 데 주된 요인이 있는 것으로 생각되며, 앞으로의 논의의 초점은 어느 정도의 한자를 가르쳐야 할 것인가 하는 데 있어야 할 것이다. 이런 문제와 관련하여 한자어 구성에 사용되는 한자의 빈도가 제시된다면 교육용 한자를 선정하는 객관적이고 타당성 있는 기준이 될 것이다.

② 국제 공용 한자 코드의 활용 지침

한자 문화권의 각국의 정보 교환의 표준화를 위하여 제정되는 국제 공용 한자 코드를 우리 실정에 맞게 실용화하기 위하여는 한자 사용 실태에 맞도록 그 배열 및 입력 방식이 정해져야 한다. 이를 위해서 본 연구 결과로 제시되는 각 한자의 사용 빈도 및 중요도가 필수적인 정보로 제공될 수 있다.

③ 어문 규범 결정의 지침

각 한자가 한자어를 구성하는 데 있어서 사용되는 위치나 어휘적 특성들이 밝혀짐으로써 복합어 및 파생어 형성과 관련된 맞춤법의 세부 사항이나 국어 순화 등의 지침으로 사용할 수 있다.

④ 한국어 자연 언어 처리 시스템용 전자 사전

현재의 한국어 자연 언어 처리 시스템을 개발하는 데 있어서 어휘 처리와 관련하여 문제가 되는 것은 동음이의어 및 사전 미등재어의 처리라 할 수 있다. 본 연구의 결과를 자연 언어 처리용 전자 사전에 수용하면 국어의 동음이의어나 사전 미등재어의 많은 부분을 차지하는 한자어 구성에 있어서 각 한자의 위

치 및 품사별, 어휘 특성별 분포에 따른 빈도 정보를 제시할 수 있게 된다. 이와 같은 정보를 전자 사전이 제공할 수 있게 되면 분석 시스템의 개발에서 이를 기반으로 한 동음이의어 해석이나 미등재어에 대한 추정 알고리즘 등을 제공할 수 있게 되어 전반적인 시스템의 향상이 이루어질 수 있다.

참 고 문 헌

高永根(1989).『國語形態論硏究』. 서울 : 서울대학교 출판부.

고영근 · 남기심(1985).『표준국어문법론』. 서울 : 탑출판사.

김광해(1993).『국어 어휘론 개설』. 서울 : 집문당.

남윤진(1994).「형태소분석기에서의 접미파생어 처리를 위한 연구」. 울산대학교 대학원 석사학위 논문.

宋喆儀(1992).『국어의 파생어형성 연구』. 國語學叢書18. 서울 : 태학사.

沈在箕(1982).『國語語彙論』. 서울 : 집문당.

정광 외(1995).『한국어 데이터 베이스의 설계 및 응용을 위한 기초 연구』. 서울 : 민음사.