『표준국어대사전』 수록 정보의 통계적 분석

정호성
국립국어연구원 학예연구사

1. 들어가기

이 글은 『표준국어대사전』에 수록된 다양한 정보를 분석·정리하고 그 결과를 수치로 제시하여 사전 전체의 윤곽을 드러내려는 목적에서 쓰는 글이다.

기존의 사전에는 주·부표제어를 합쳐 3~40만여 항목의 표제어가 실려 있는 것으로 알고 있다. 이렇듯 많은 양의 자료이지만 사전의 구체 항목에 대한 수치화된 정보, 즉 명사의 개수, 한자어의 개수, 표제어의 음절 빈도수 등과 같은 정보는 아직 어떤 사전을 대상으로 해서도 제시된 바가 없다. 이에 『표준국어대사전』 전체를 대상으로 표제어와 하위 항목을 유형별로 분류하고 정리하여 그 계량적(計量的)인 모습을 보이려 한다. 이 작업을 통하여 『표준국어대사전』에 수록된 정보를 수치로 파악할 수 있을 것이며, 나아가서 국어 전반에 걸친 다양한 정보를 얻을 수 있을 것으로 기대한다.

이번 작업에서는 우선 표제항과 그 하위 분류 항목인 원어, 발음, 부표제어, 관용구, 속담 등과, 각 항목에서 제시된 각종 기호를 대상으로 하여 분석을 실시하였다. 더욱 구체적인 정보를 얻기 위해서는 표제어에 따른 뜻풀이 유형과 의미 기술 등을 분석하여야 할 것이나, 전산 자료의 변환이 완벽하게

이루어지지 못한 관계로 이 작업은 다음 기회로 미룬다.

사전 원고 검색에 쓰인 프로그램은 국립국어연구원에서 연구·개발한 사전 검색 프로그램인 'Hdb97.exe'와 문자열 검색 프로그램인 'Hgrep97.exe'를 이용하였다.

2. 『표준국어대사전』의 구성

『표준국어대사전』에 수록된 정보를 분석하려면 먼저 사전에 어떤 정보가 어떤 형식으로 제시되어 있는지를 살펴보아야 한다. 그래야만 어떤 정보를 검색할 것인지가 명확히 드러나기 때문이다.

사전의 각 표제항은 다음과 같은 구성으로 이루어져 있다(각 항목의 구성을 보이고 각 항목 안에서 사용된 기호[1]를 보인다).

2.1. 표제어 : 주표제어와 부표제어로 구분된다. 표제어는 한글 표기 이외에 붙임표(-)[2], 띄어쓰기 기호(^)[3], 어깨 번호[4], 의존 형태소 기호(-)[5], 고유명사 구분 기호(·)[6]로 구성된다.

2.2. 원어 : 표제어가 한자어인 경우에는 한자로, 외래어인 경우에는 로마자로 그 원어를 밝힌다. 한자와 로마자 이외에 고유어 기호(-)[7], 언어명 약호[8], 한자음이 원음과 다름을 나타내는 기호(▾)[9], 한자어의 병용자 구분 기호(／)[10], 해당 언어에 존재하지 않는 원어 기호(▾)[11], 원래의 형태에서 변한 외래어 기호(←)[12] 등으로 구성된다.

---

1) 각종 기호에 대한 구체적인 설명은 『표준국어대사전』의 〈일러두기〉 참조.
2) 예) '큰-아버지', '첫-눈'
3) 예) '노래기^부적', '성격^묘사'
4) 예) '거문[1](去文)', '거문[2](巨門)', '거문[3](拒門)'
5) 예) '-었-', '-이'
6) 예) '독일·오스트리아^전쟁', '보일·샤를의 법칙'
7) 예) '거만스럽다(倨慢---)', '거멱(擧-)', '나비넥타이(--necktie)'
8) 예) '가격^카르텔(價格㊎Kartell)', '가라테(㊐karate〔唐手〕)'
9) 예) '시월(十▾月)', '사탕(砂糖▾)'
10) 예) '기념(紀念/記念)', '견강(堅剛/堅强)'
11) 예) '콩글리시(▾konglish)', '고스톱(▾go stop)'
12) 예) '케미 슈즈(←chemical shoes)', '헬기(←helicopter機)'

2.3. 발음 : 표제어의 표기와 발음이 일치하지 않는 음절에 대하여 발음을 제시한다. 그 밖에도 발음 변화가 없는 음절 표시 기호(-)[13], 장음 부호(:)[14], 허용 발음 표시 기호(/)[15] 등으로 구성된다.

2.4. 뜻풀이 : 뜻풀이는 표제어의 의미를 설명하는 항목이다. 품사 기호(명, 동 등)와 하위 분류 번호(Ⅰ, 1, ① 등)[16]가 표시되고 그 뒤로 전문어, 북한어, 방언, 옛말임을 나타내는 범주 표시 기호가 위치한다. 그 뒤로 문형 정보(【…을】, 【(…과)】 등)와 문법 정보(((‘…과’가 나타나지 않을 때는 여럿임을 뜻하는 말이 주어로 온다)) 등)가 제시되고 구체적인 뜻풀이가 나오게 된다. 각각의 세부 뜻풀이 뒤에는 용례가 제시된다. 용례는 다시 작성례와 인용례로 나뉜다. 관련 어휘 정보와 참고 어휘 정보가 그 뒤를 따르며, 뜻풀이 맨 끝에 어원 혹은 최초 출현형이 제시된다.

2.5. 부표제어와 뜻풀이 : ‘-하다, -되다, -거리다, -대다, -이다, -이/히, -적’ 등이 결합한 파생어는 해당 어기의 부표제어로 처리하였다. 뜻풀이는 부표제어 뒤에 제시하였다.

2.6. 관용구・속담 : 관용구와 속담을 제시하고 뜻풀이를 제시하였다. 교체가 가능한 성분임을 나타내는 기호(〔 〕)와 여러 개의 교체 가능 성분을 구분해 주는 기호(／)[17] 등으로 구성된다.

3. 사전 각 항목의 통계적 분석

3.1. 표제어 항목

『표준국어대사전』의 표제어 수는 모두 44만여 항목이고, 여기에 부표제

13) 예) ‘가계(家計)〔-계/-게〕’, ‘필획(筆劃)〔-획/-휙〕’
14) 예) ‘구명(救命)〔구:-〕’, ‘하급생(下級生)〔하:-쌩〕’
15) 예) ‘가계(家計)〔-계/-게〕’, ‘나뭇가지〔-무까-/-묻까-〕’
16) Ⅰ Ⅱ : 같은 표제어의 품사가 달라진 경우에 쓰인다.
1 2 : 같은 표제어의 문형 정보가 달라진 경우에 쓰인다.
①② : 같은 표제어의 뜻풀이가 달라진 경우에 쓰인다.
17) 예) ‘가슴이 〔가슴에〕 찔리다’, ‘같은 값이면 다홍치마 〔검정 송아지 / 과부 집 머슴살이 / 처녀〕’

어 6만 8천여 항목을 합하면 전체 50만 8천여 항목에 이른다. 먼저 주표제어와 부표제어 모두를 품사별로 분류해 보면 〈표 1〉과 같다.[18]

〈표 1〉 표제어의 품사별 분류[19]

표제어 / 품사	주표제어		부표제어		주표제어+부표제어	
	항목수	백분율	항목수	백분율	항목수	백분율
	440,262	100.00	68,509	100.00	508,771	100.00
명사	333,226	75.68	1,156	1.68	333,226	65.49
의존 명사	1,049	0.23	0	0.00	1,049	0.20
대명사	462	0.10	0	0.00	462	0.09
수사	275	0.06	0	0.00	275	0.05
동사	15,135	3.43	53,235	77.70	68,370	13.43
형용사	6,438	1.46	10,915	15.93	17,353	3.41
부사	14,076	3.19	3,883	5.66	17,959	3.52
관형사	525	0.11	1,156	1.68	1,681	0.33
감탄사	811	0.18	0	0.00	811	0.15
조사	356	0.08	0	0.00	356	0.06
어미	2,523	0.57	0	0.00	2,523	0.49
접두사	204	0.04	0	0.00	204	0.04
접미사	450	0.10	0	0.00	450	0.08
어근	7,353	1.67	0	0.00	7,353	1.44
무품사	58,884	13.37	0	0.00	58,884	11.57
품사 통용	1,526	0.34	1,853	2.70	2,223	0.43

18) 표에서 보듯이 표제어의 총 개수와 각 품사의 합은 일치하지 않는다. 이는 품사가 통용되는 표제어('가까이(명~부)', '자라다(동~형)', '그따위(대~관)' 등)가 각각의 품사에 한 번씩 두 번 계산되기 때문이다.

19) 표에 제시된 항목수는 실제 사전의 항목수와 약간의 차이가 있을 수 있다. 그것은 사전 모습 그대로의 원고를 전산 자료 파일로 변환하는 과정에서 오류가 일부 발생하였기 때문이다. 오류의 대부분은 전산 코드의 불일치로 인하여 글자가 실현되지 않거나 엉뚱한 기호로 변환된 경우, 혹은 일정한 문자열을 만나면 그 뒤의 정보가 삭제되는 경우 등이 있다. 그러나 오류를 일으킨 항목의 개수는 사전 전체에서 보면 그리 큰 비중을 차지하는 것이 아니므로, 이러한 오류에서 발생하는 차이는 무시하고 작업을 진행하였다. 더욱 정확한 통계 분석 자료를 제시하기 위하여 오류를 일으킨 항목을 수정한 후 다시 한번 작업을 진행할 예정이다.

〈표 1〉은 주표제어와 부표제어 각각의 품사별 분류를 보이고 또한 이들을 모두 합친 수치를 보인 것이다. 품사별로는 명사가 전체의 2/3 가량을 차지하고 있고, 그 다음으로는 동사, 부사, 형용사 순으로 그 순서가 매겨짐을 알 수 있다. 부표제어에서 제시된 명사의 1,156항목은 '-적(的)'으로 파생된 단어들로 모두 '「관」「명」'과 같이 관형사와 통용되는 것으로 처리된 것이다.

〈표 1〉에서 무품사 58,884항목 중에는, 띄어 쓰는 것이 원칙이나 붙여 쓸 수 있는 전문 용어나 고유 명사('^' 표시) 53,477항목이 포함된 것이다. 나머지는 구 구성이 줄어든 말(내(〈나+의), 걔(〈그 아이))이 차지한다.

다음으로, 표제어를 자모별로 나누어 그 개수를 세어보면 다음과 같다.

〈표 2〉 주·부표제어의 자모별 개수

자모	주표제어	부표제어	계	자모	주표제어	부표제어	계
ㄱ	65,516	10,207	75,723(14.88%)	ㅇ	69,007	10,131	79,138(15.55%)
ㄴ	17,494	1,999	19,493(3.83%)	ㅈ	50,735	9,383	60,118(11.81%)
ㄷ	31,130	4,052	35,182(6.91%)	ㅊ	19,480	4,353	23,833(4.68%)
ㄹ	9,588	857	10,445(2.05%)	ㅋ	5,254	428	5,682(1.11%)
ㅁ	26,448	3,264	29,712(5.84%)	ㅌ	9,687	1,850	11,537(2.27%)
ㅂ	39,475	6,917	46,392(5.19%)	ㅍ	12,572	2,239	14,811(1.97%)
ㅅ	55,705	7,415	63,120(12.40%)	ㅎ	28,170	5,348	33,518(6.59%)

표제어의 자모별 개수로는 'ㅇ'으로 시작하는 표제어가 가장 많아 주·부표제어를 모두 합치면 모두 7만 9천여 항목에 이른다. 자모별 빈도는 'ㅇ〉ㄱ〉ㅅ〉ㅈ〉ㅂ〉ㄷ〉ㅎ〉ㅁ〉ㅊ〉ㄴ〉ㅍ〉ㅌ〉ㄹ〉ㅋ'의 순서로 나타난다.[20]

20) 표에서는 'ㅋ'이 가장 적은 것으로 나타난다. 하지만 'ㄹ' 표제어의 대부분이 두음법칙을 적용하지 않는 북한어임을 감안한다면 남한어 중에서는 'ㄹ'로 시작하는 단어의 수가 가장 적다고 할 수 있을 것이다. 참고로, 'ㄹ' 주표제어 중에서 북한어는 6,182항목이고, 'ㄹ' 부표제어 중에서 북한어는 848항목이다.

다음으로 부표제어의 구성을 자세히 살펴보도록 하자. 부표제어는 〈표 3〉에서 보는 것과 같이 '-하다, -되다'(동사, 형용사), '-거리다, -대다, -이다'(동사), '-이/히'(부사), '-적(的)'(관형사~명사) 들이 결합한 파생어로 구성된다. 부표제어는 모두 68,509항목이고, 그 가운데 '-하다' 항목이 5만 1천여 항목으로 부표제어의 75%를 차지하고 있음을 볼 수 있다. '동~형'의 697항목은 '-하다' 부표제어에서 '동사'와 '형용사'로 품사가 통용되는 항목의 개수를 나타낸 것이다. '-되다' 부표제어는 4,800여 항목으로 이들 중 형용사는 단지 18항목이고 나머지는 모두 동사로 나타난다. '-거리다'와 '-대다'는 대부분 동일한 어기에 결합하는 것으로 나타난다. 그 개수에 차이가 나는 것은 북한어 가운데 일부 '-거리다' 표제어의 어근이 '-대다'와 결합하지 못하는 경우가 몇 개 있기 때문이다(뒤범벅거리다, 부쩍거리다…).

〈표 3〉 부표제어의 구성

부표제어	항목수	동사	형용사	부사	관~명	동~형
-하다	51,537	41,356	10,861	-	-	697
-되다	4,877	4,859	18	-	-	-
-거리다	3,200	3,200	-	-	-	-
-대다	3,193	3,193	-	-	-	-
-이다	509	509	-	-	-	-
-이/히	3,792	-	-	3,792	-	-
-적(的)	1,156	-	-	-	1,156	-
옛말	245	118	36	91	-	-
계	68,509	53,235	10,915	3,883	1,156	697

'-이/히' 부표제어는 '-하다, -되다' 부표제어가 다시 부사로 파생되는 경우(깨끗하다~깨끗이, 헛되다~헛되이), 혹은 '-스럽다, -롭다' 파생어가 다시 부사로 파생되는 경우(정성스럽다~정성스레, 자유롭다~자유로이), 이들 부사형들을 부표제어로 등재한 것이다.

다음으로 표제어에 제시되어 있는 정보들을 구체적으로 분석해 보기로 한다. 먼저 표제어가 복합어[21]임을 나타내는 붙임표(-)가 쓰인 항

21) 여기서 '복합어'는 '합성어'와 '파생어'를 모두 이르는 말이다.

목들을 품사별로 분류하면 〈표 4〉와 같다.[22] (표의 백분율에서 'ㄱ'은 해당 항목의 전체 표제어에 대한 비율이고, 'ㄴ'은 해당 부류 총 항목에 대한 비율이다.)

〈표 4〉 붙임표가 있는 표제어

분류 / 품사	주표제어		붙임표		
	항목수	백분율	항목수	백분율	
				ㄱ	ㄴ
전체 항목수	440,262	100.00	158,819	36.07	100.00
명사	333,226	75.68	139,786	31.75	88.01
의존 명사	1,049	0.23	146	0.03	0.09
대명사	462	0.10	111	0.02	0.06
수사	275	0.06	101	0.02	0.06
동사	15,135	3.43	8,188	1.85	5.15
형용사	6,438	1.46	3,740	0.84	2.35
부사	14,076	3.19	6,060	1.37	3.81
관형사	525	0.11	196	0.04	0.12
감탄사	811	0.18	236	0.05	0.14
조사	356	0.08	38	0.00	0.02
어미	2,523	0.57	0	0.00	0.00
접두사	204	0.04	0	0.00	0.00
접미사	450	0.10	0	0.00	0.00
어근	7,353	1.67	0	0.00	0.00
무품사	58,884	13.37	454	0.10	0.28
품사 통용	1,526	0.34	237	0.05	0.14

〈표 4〉에서 볼 수 있듯이 어미나 접사 등에는 붙임표 기호가 나타나지 않는다. 어미나 접사에 제시된 '-'는 자립적으로 쓰이지 않고 반드시 다른 말과 결합해야 하는 표제어임을 밝히는 기호이므로 붙임표와는 다른 것이다.

어미나 접사와는 달리, 조사에는 붙임표가 제시되었다. 이는 '에서-부터', '이야-말로'와 같이 조사가 통합되어 쓰이는 형태도 표제어로 실었기 때문이다.

다음으로 어깨 번호를 붙인 표제어에 대하여 살펴보자. 어깨 번호를 붙인 표제어는 둘 이상의 동음이의어가 있다는 의미이다. 『표준국어대사전』에서는 원어나 품사가 다르더라도 붙임표나 의존 형태소 기호, 띄어쓰기 기호 등을 제외한 한글 표기가 동일하다면 주표제어와 부표제어 구분 없이 모두 동음이의어로 처리하였다.

예) 가다[1] 「1」동「1」 ① 한 곳에서 다른 곳으로 장소를 이동하다. …
가다[2] 부『북』 '이따금'의 북한어.

22) '눌리다, 빨리'처럼 구성 성분이 음절로 나누어지지 않을 때는 붙임표를 제시하지 않았으므로, 실제 복합어는 위 표에 제시된 것보다 많다.

가다[3] 명 '어깨[1]'의 잘못.

가다[4] 〔Gadda, Carlo Emilio〕 명『인』이탈리아의 작가.

그러므로 표제어의 어깨 번호 정보에 대하여 주표제어와 부표제어로 나누거나 품사별로 구분하여 살펴보는 것은 별 의미가 없다.

어깨 번호가 쓰인 표제어는 모두 125,517항목이다(주표제어 106,778항목, 부표제어 18,739항목). 이들 중에서 어깨 번호가 '1'인 것은 모두 41,499항목이다(주표제어 33,547항목, 부표제어 7,952항목). 이는 125,517항목의 동음이의어에서 중복되어 나타난 것을 헤아리지 않은 개별 표제어가 41,499항목이라는 의미이다. 한편, 동음이의어의 개수가 가장 많은 것은 '장'으로 46항목이 제시되어 있다.

3.2. 원어 항목

원어는 표제어에 따라 한자어, 외래어, 그리고 이들과 고유어가 결합한 혼종어로 구성된다. 한자어 원어는 한자로, 그 외의 외래어 원어는 모두 로마자로 제시되어 있다. 먼저 원어가 쓰인 표제어 전체에 대하여 살펴보도록 한다.

〈표 5〉에서 알 수 있듯이, 44만여 항목의 표제어에서 한자어나 외래어가 한 자라도 포함된 표제어는 대략 32만 9천 항목으로, 이는 주표제어의 75% 가량을 차지한다.

〈표 5〉 원어가 제시된 표제어

품사 \ 항목	주표제어		원어 제시 표제어		
	항목수	백분율	항목수	백분율	
				ㄱ	ㄴ
전체	440,262	100.00	328,963	74.71	100.00
명사	333,226	75.68	263,338	59.81	80.05
의존 명사	1,049	0.23	675	0.15	0.20
대명사	462	0.10	243	0.05	0.07
수사	275	0.06	89	0.02	0.02
동사	15,135	3.43	438	0.09	0.13
형용사	6,438	1.46	1,293	0.29	0.39
부사	14,076	3.19	808	0.18	0.24
관형사	525	0.11	193	0.04	0.05
감탄사	811	0.18	73	0.01	0.02
조사	356	0.08	0	0.00	0.00
어미	2,523	0.57	0	0.00	0.00
접두사	204	0.04	90	0.02	0.02
접미사	450	0.10	220	0.04	0.06
어근	7,353	1.67	4,355	0.98	1.32
무품사	58,884	13.37	57,703	13.10	17.54
품사 통용	1,526	0.34	569	0.12	0.17

원어가 제시된 표제어 가운데에서 가장 많은 것은 역시 명사로 26만 3천여 항목에 이른다. 그 다음으로는 품사 표시가 없는 표제어가 5만 7천여 항목에 이른다. 이들은 대부분 구 구성의 전문어로 띄어쓰기 기호(^)를 가지는 표제어들이다.

여타의 품사들과는 달리, 조사와 어미는 고유어만으로 구성되어 있음을 볼 수 있다. 그러나 접사에는 원어가 상당수 제시된 것으로 보아 어미·조사와 접사는 그 성격이 다름을 알 수 있다.

다음으로 원어가 제시된 표제어를 원어의 종류별로 분류해 보자. 〈표 6〉에서 볼 수 있듯이, 순수 고유어는 11만여 항목으로 전체 44만여 주표제어의 25%를 차지하고 있다.

원어가 제시된 항목은 주표제어의 74.72%인 약 32만 9천 항목이고, 그 가운데 한자어가 포함된 항목은 30만여 항목이다. 이들 중 원어가 한자로만 구성된 표제어를 찾는다면 모두 25만여 항목으로, 이는 주표제어 전체의 57%를 차지하는 수치다.

원어에 외래어가 한 자라도 제시된 표제어는 4만여 항목으로, 주표제어의 9.27%에 이른다. 이들 중 순수한 외래어로만 구성된 표제어는 2만 3천여 항목으로 주표제어의 5.27%를 차지한다.

〈표 6〉 주표제어의 원어별 분류

항목 / 품사	항목수	고유어	한자어	외래어	한+고	외+고	한+외	한+외+고
전체 (백분율)	440,262 (100%)	111,299 (25.28%)	251,478 (57.12%)	23,196 (5.26%)	36,461 (8.28%)	1,331 (0.30%)	15,548 (3.53%)	751 (0.17%)
명사	333,226	69,888	205,229	19,443	31,221	845	6,142	312
의존 명사	1,049	411	207	350	9	6	65	1
대명사	462	219	236	0	7	0	0	0
수사	275	186	89	0	0	0	0	0
동사	15,135	14,701	0	0	433	1	0	0
형용사	6,438	5,145	0	0	1,266	27	0	0
부사	14,076	13,268	528	0	280	0	0	0
관형사	525	332	191	0	2	0	00	0
감탄사	811	738	30	11	32	0	0	0
조사	356	356	0	0	0	0	0	0
어미	2,523	2,523	0	0	0	0	0	0
접두사	204	114	90	0	0	0	0	0
접미사	450	230	220	0	0	0	0	0
어근	7,353	3,003	4324	0	26	0	0	0
무품사	58,884	1,181	40,782	3,426	3,271	452	9,342	439
품사 통용	1,526	957	490	34	43	0	2	0

이 외에 원어 항목과 관련하여 산출할 수 있는 정보는 다음과 같다.

원어 항목에서 영어 이외의 원어에는 원어명을 보이고 있는데, 모두 32개의 언어명이 4,139항목에 제시되어 있다. 또한 두 가지 이상의 원어 표기를 병기한 표제어는 한자어 2,023항목과 외래어 52항목이고, 형태의 변화가 있는 원어가 978항목, 한자어의 음이 원음과 다른 경우가 708항목, 원어에는 없는 형태, 즉 해당 언어에 존재하지 않는 원어가 170항목 등으로 나타난다.

〈표 7〉의 맨 밑줄에 제시된 '외래어〔한자〕' 구성은, 일본어와 중국어를 원음대로 차용한 경우에 로마자로 원어를 제시하고 '〔 〕'안에 해당 한자를 제시한 것이다. 그러므로 이 항목은 외래어만으로 구성된 표제어에 속할 수 있을 것이다. 중국어 810여 항목, 일본어 376여 항목에 걸쳐 제시되어 있는 정보이다.

〈표 7〉 원어 표제어 기타 세부 사항

구분	표제항	예
언어명 표시	4,139	가제(㉤Gaze), 솔(㉥sol), 수드라(㉲sudra)
한자어 병기 (/)	2,023	기념(記念/紀念), 탄복(歎服/嘆服)
외래어 병기 (/)	52	미스터(mister/Mr.)
형태 변화 (←)	978	메리야스(← ㉦medias), 크레파스(㉨← kurepasu)
한자음 변음 (▽)	708	돈쭝(-重▽), 모과(木▽瓜), 보시(布▽施)
원어에 없는 형태 (▼)	170	오버센스(▼over sense), 카센터(▼car center)
원어 중간점 (·)	166	앵글로·색슨(Anglo·Saxon)
외래어〔한자〕	1,186	베이징(Beijing〔北京〕), 다쿠앙(㉨takuan〔澤庵〕)

다음으로 부표제어를 원어별로 분류해 보자. '-적(的)'이 결합한 부표제어를 제외한다면, 부표제어는 '-하다, -되다, -거리다, -이/히' 등과 같은 접미사가 결합한 형태이므로 한자어나 외래어만으로 구성된 부표제어는 있을 수 없다. 그러므로 〈표 8〉의 각 항목은 '고유어 + -하다', '한자어 + -되다'와 같은 형식으로 이해해야 한다.

〈표 8〉 부표제어의 원어별 분류

부표제어	항목수	고유어+	한자어+	외래어+	한+고+	외+고+	한+외+
전체	68,509 (100%)	20,672 (30.17%)	46,438 (67.78%)	165 (0.24%)	1,624 (2.37%)	2 (0.00%)	41 (0.05%)
-하다	51,537						
동사	41,356	8,113	32,114	142	954	2	31
형용사	10,861	4,018	6,805	7	31	0	0
-되다	4,877						
동사	4,859	47	4,761	14	27	0	10
형용사	16	6	10	0	0	0	0
-거리다	3,200	3,193	7	0	0	0	0
-대다	3,193	3,186	7	0	0	0	0
-이다	509	507	2	0	0	0	0
-이/히	3,792	1,602	1,578	0	612	0	0
-적	1,156	0	1,154	2	0	0	0

'-하다, 되다'의 경우는 고유어보다 한자어와 결합한 개수가 훨씬 많음을 볼 수 있다. 그러나 '-거리다, -대다, -이다'와 같은 접미사는 대부분 고유어와 결합하고 있음을 알 수 있다. 한자어에 '-거리다'가 결합한 예는 '쟁쟁거리다(琤琤-), 주저거리다(躊躇-), 희희낙락거리다(喜喜樂樂-)' 등의 예가 있고, 한자어에 '-이다'가 결합한 예는 '충동이다(衝動-), 흥성이다(興盛-)'의 두 예가 있다. '-적'도 거의 한자어와 결합하지만 외래어와도 결합할 수 있음을 보여준다.(미스터리적, 카리스마적)

3.3. 발음 항목

표제어의 표기와 실제 발음이 차이가 나는 경우 발음 표시를 하였다. 발음 정보는 외래어, 북한어, 방언, 옛말, 비표준어를 제외한 현대 표준어에만 제시하였다. 또한 문법 형태와 어근, 그리고 품사 정보가 없는 전문어 구에서도 발음 정보를 제시하지 않았다. 그러므로 발음 정보가 표시된 항목은 현대 표준어 가운데 일부에 국한된다고 할 수 있다.

발음 정보가 제시된 주표제어는 전부 130,406항목이다. 이 중에서 장음 기호(:)가 있는 항목은 모두 67,314항목이고, 이들 중에서 원어가 한자어인 것이 59,506항목이다.

〈표 9〉 발음 정보를 가지는 표제어

구분	표제항
발음이 제시된 주표제어	130,406
장음 제시 (:)	67,314
장음이 고유어인 경우	7,808
장음이 한자어인 경우	59,506
복수 발음 (/)	16,545

표준어 규정에 따라 발음이 둘 이상으로 될 경우에는 '/' 기호를 이용하여 왼쪽에는 원칙적인 발음을, 오른쪽에는 허용되는 발음을 제시하였다.('강의〔강:의/강:이〕', '되다〔되:-/뒈:-〕') 이렇게 복수 발음을 제시한 항목은 16,545항목에 이른다.

3.4. 뜻풀이 항목

품사별 구분은 이미 위에서 제시되었으므로 여기서는 활용 정보, 문형 정보, 문법 정보 등에 대하여 알아보기로 한다.

〈표 10〉 활용·문법 정보

	활용 정보	문형 정보
명사	32,391	-
수사	47	-
대명사	32	-
동사	6,478	29,780
형용사	2,192	1,313
계	41,140	31,093

활용 정보는 체언과 조사의 결합형 및 용언의 활용형과 발음 변화를 제시한 것으로 북한어, 옛말, 방언, 비표준어 등을 제외한 표준어에만 모두 41,138항목이 제시되어 있다. 부표제어에는 활용 정보를 제시하지 않았다. 품사별로는 명사에 32,391항목, 수사에 47항목, 대명사에 32항목, 동사에 6,478항목, 형용사에 2,192항목이 제시되어 있다.

문형 정보는 주어를 제외한 용언의 필수적 성분을 제시한 것으로 동사 29,780항목(주표제어 4,727항목, 부표제어 25,053) 형용사 1,313항목(주표제어 297항목, 부표제어 1,016항목) 등 모두 3만 1천여항목에 제시되어 있다.

문법 정보는 문형 정보를 제시할 수 없는 경우, 굳어진 형식으로 쓰이는

경우, 문장 성분이 가지는 통사·의미론적 제약을 보일 경우, 음운이나 형태 결합상의 제약이나 통사 환경을 보일 경우, 의미의 선택 제한을 보일 경우, 활용상의 제약을 보일 경우, 제한된 환경에서 쓰일 경우 등을 나타내는 데에 사용된다.

〈표 11〉 문법 정보

	문법 정보
명사	1,020
수사	21
대명사	17
동사	1763
형용사	309
부사	134
관형사	66
조사	198
감탄사	9
어미	2,018
접사	597
계	6,152

문법 정보는 문형 정보와 함께 쓰여 문형 정보를 이해하기 위해 필요한 정보로 쓰이는 경우도 있지만, 문형 정보 없이 해당 표제어의 문장에서의 쓰임을 보이는 경우에도 사용된다. 문법 정보는 주표제어와 부표제어를 통틀어 6,152항목에 제시되어 있다. 〈표 11〉에서 동사 1,763항목은 주표제어 542항목과 부표제어 1,221항목을 합한 것이고, 형용사 309항목은 주표제어 132항목과 부표제어 177항목을 합한 것이다. 문법 정보가 제시된 부표제어 가운데 부사는 '남짓이' 단 하나로 "((수량을 나타내는 말 뒤에 쓰여))"와 같은 문법 정보가 제시되어 있다.

다음으로 용례에 대하여 살펴보자. 용례는 다시 작성례와 인용례로 나뉘는데, 작성례는 해당 표제어의 가장 전형적인 쓰임을 구와 문장으로 보인 것으로 사전 편찬자들이 작성한 것이고, 인용례는 문헌 자료를 중심으로 구성되어 있는 7천만 어절 가량의 문학 작품 데이터베이스에서 해당 표제어의 쓰임을 잘 보이고 있는 예를 선별하여 직접 인용한 예이다. 그러므로 모든 인용례에는 그 작자와 작품 이름이 제시되어 있다.

용례는 모두 99,706항목에 걸쳐 제시되어 있다(주표제어 74,092항목, 부표제어 25,614항목). 이 가운데 인용례가 제시된 표제어는 49,651항목이다. 개별 용례는 모두 229,196 개가 제

〈표 12〉 주·부표제어 용례 개수(옛말 제외)

용례	주표제어	부표제어	계
전체 표제어	440,262	68,443	508,905(100 %)
용례 표제어	74,092	25,614	99,706(20%)
인용례 표제어	39,099	10,552	49,651(10%)
용례 총 개수	172,232	56,964	229,196
인용례 총 개수	50,626	13,700	64,326

시되었으며, 인용례는 64,326 개로 전체 용례의 28%에 이른다.

의미의 차이가 거의 없어 동의 관계에 있는 표제어들은 어느 하나를 기본으로 삼아 뜻풀이를 하고 나머지에서는 '=' 기호를 사용하여 기본 표제어로 뜻풀이를 돌렸다. 또한 기본 표제어에서는 뜻풀이 다음에 '='를 사용하여 기본 표제어로 뜻풀이를 돌린 해당 동의어를 모두 보여 주었다. 이렇게 뜻풀이에 동의어가 제시된 기본 표제어는 46,651항목이고, 뜻풀이를 기본 표제어로 돌린 표제어는 66,635항목이다.

다음으로는 관련 어휘와 참고 어휘를 제시한 항목을 살펴보자.

관련 어휘는 본말, 준말, 비슷한말, 반대말, 높임말, 낮춤말로 나뉘는데, 본말은 1,012항목에, 준말은 2,412항목에, 비슷한 말은 18,411항목에, 반대말은 5,285항목에, 높임말은 64항목에, 낮춤말은 29항목에 각각 제시되어 있다.

또한 어떤 표제어의 의미를 이해하는 데 도움이 될 수 있는 어휘를 관련 어휘 뒤에 참고 어휘로 보여 주고 있는데, 이렇게 참고 어휘를 밝힌 표제어는 모두 31,185항목에 이른다.

다음으로 뜻풀이 항목의 제일 마지막에 위치하는 어원 정보에 대하여 알아보자. 로마자 이니셜로 이루어진 표제어에 원말을 제시한 것은 973항목, 표제어의 어원적 분석을 제시한 것은 2,197항목, 현재는 고유어처럼 보이나 원래 한자어이거나 몽골어, 중국어 따위에서 차용한 단어는 150항목, 제3언어를 통해 들어온 간접 차용어는 104항목들이 제시되어 있다. 또한 해당 표제어가 15~17세기에 처음으로 문헌에 나타난 예를 찾아 어원 정보에 제시하고 그 출현형과 문헌명을 밝힌 표제어는 모두 3,566항목에 이른다.

3.5. 관용구·속담 항목

관용구와 속담은 그 시작 단어를 주표제어로 삼아 해당 주표제어 밑에서 한 번만 제시하였다. 관용구는 2,272항목의 표제어에 모두 4,623개가 제시되어 있다. 이 가운데에는 북한에서 쓰이는 관용구 885개도 포함되어 있다. 속담은 3,358항목의 표제어에 모두 9,475개가 제시되어 있는데 여기에도 역

시 북한에서 쓰이는 속담 2,622 개가 포함되어 있다.

3.6. 표제어의 범주별 분류

이상에서는 전체 표제어를 대상으로 각각의 하위 항목을 살펴보았다. 하지만 이에는 일반어 뿐만 아니라 전문어, 북한어, 방언, 옛말이 포함되어 있으므로 순수한 국어의 일면을 나타내는 것이라고는 말하기 어려운 면이 있다. 이에 표제어를 일반어, 전문어, 북한어, 방언, 옛말로 어휘 범주를 나누어 각각의 세부 사항을 살펴보기로 한다.[23]

먼저 주표제어의 범주를 나누어 살펴보면 〈표 13〉과 같다.

〈표 13〉 주표제어의 범주에 따른 분류

	계	일반어	전문어	북한어	방언	옛말
계	440,262	170,792	221,733	64,313	20,482	12,194
명사	333,226	141,609	163,124	55,467	15,231	6,062
의·명	1,049	531	335	65	59	63
대명사	462	330	9	16	84	23
수사	275	175	0	7	34	59
동사	15,135	7,521	157	3,083	2,184	2,824
형용사	6,438	3,049	6	1,629	1,217	766
부사	14,076	8,700	0	3,988	1,209	885
관형사	525	370	11	18	63	66
감탄사	811	529	87	85	122	20
조사	356	172	0	4	95	87
어미	2,523	808	0	22	341	1,360
접두사	204	168	0	23	23	18
접미사	450	336	0	0	0	86
어근	7,353	7,353	0	0	0	0
무품사	58,884	595	58,266	0	0	0
통용	1,526	725	0	0	0	-120

23) 주표제어의 총 항목수는 44만여 항목이지만 일반어, 전문어, 북한어 등을 모두 합한 수치는 훨씬 많은 49만여 항목임을 볼 수 있다. 이는 품사의 통용과 같이, 일반어이면서 전문어가 된다든지(가슴, 배), 일반어~북한어, 일반어~방언, 전문어~북한어 등과 같은 범주의 통용을 보이는 표제어가 많이 나타나기 때문이다.

표제어를 범주별로 나눠 본 결과 전문어가 일반어보다 훨씬 많음을 보게 된다. 이는 물론 사전 편찬의 태도에 따른 것으로, 순수 언어 사전적 정보뿐만 아니라 백과 사전적 정보도 제공하고 있는『표준국어대사전』의 성격이 그대로 드러난다고 할 수 있겠다.

전문어는 다음과 같이 모두 53개 분야로 나뉘어 제시되어 있다.

〈표 14〉 전문어 분야별 개수

가톨릭	1,362	군사	4,796	민속	4,655	심리	1,524	의학	10,264	철학	1,988
건설	6,105	기계	2,347	법률	8,696	약학	1,476	인명	10,304	출판	1,358
경제	8,328	기독	1,121	불교	9,271	언론	551	전기	2,951	컴퓨터	1,454
고적	2,392	논리	669	사회	2,300	언어	4,110	정치	2,082	통신	1,061
고유명사	498	농업	4,369	생물	4,104	역사	19,401	종교	970	한의학	4,837
공업	3,606	동물	11,559	수학	3,991	연영	1,670	지리	5,905	항공	871
광업	3,242	문학	3,751	수산	898	예술	1,283	지명	6,989	해양	1,117
교육	1,453	물리	7,710	수공	2,192	운동	4,653	책명	2,050	화학	8,605
교통	1,530	미술	1,458	식물	12,942	음악	6,842	천문	2,072		

전문어 가운데 가장 표제어 수가 많은 분야는 역사 분야로 1만 9천여 항목이 제시되었고, 가장 적은 분야는 고유 명사로 498항목이 제시되었다.

전문어와 일반어 다음으로 개수가 많은 것은 북한어이다. 북한어는『조선말 대사전』(1992)에 수록된 단어 가운데 남한에서 쓰임이 확인되지 않은 단어와 어문 규정의 차이로 달리 표기하는 단어를 편찬 원칙에 따라 선정하여 수록한 것이다. 남한에서 쓰는 단어라도 북한에서만 쓰는 용법이 있다면 북한어 뜻풀이를 덧붙였다. 그러므로 하나의 표제어에 남한어와 북한어가 공존할 수도 있다.

부표제어를 범주별로 자세히 분류하면 〈표 15〉의 결과를 얻을 수 있다.

〈표 15〉 부표제어의 범주에 따른 분류

	계	일반어	전문어	북한어	방언	옛말
총 항목수	68,443	57,104	6,299	7,796	2	244
동사	하다 41,356 되다 4,859 거리다 3,200 대다 3,193 이다 509 옛말 118	하다 33,541 되다 4,118 거리다 2,499 대다 2,487 이다 476	하다 5,440 되다 721	하다 4,288 되다 360 거리다 847 대다 846 이다 49	1	118
형용사	하다 10,861 되다 18 옛말 36	하다 10,052 되다 15	하다 64	하다 1,022 되다 3	1	36
부사	3,883	3,368	5	411	-	90
관~명	1,156	1,073	69	38	-	-
통용	-697	562	-	159	-	-

부표제어에서는 일반어의 개수가 다른 범주에 비하여 월등히 많음을 볼 수 있다. 일반어 부표제어 가운데에서는 '-하다'가 가장 많아 전체 부표제어의 49%에 이르고 있다.

다음은 주·부표제어를 음절수별로 분류한 것이다(옛말과 방언은 제외). 동사와 형용사의 경우에는 어미 '-다'를 제외한 어간만을 대상으로 하였으므로, '먹다, 보다' 등은 1음절, '가리다, 높이다' 등은 2음절, '깨끗하다, 성공하다' 등은 3음절로 처리하였다(표에서 괄호 속의 숫자는 해당 음절에 포함된 동사와 형용사의 개수이다).

〈표 16〉에서 보듯이 표제어를 음절수별로 분류해 본 결과 3음절 표제어가 가장 많음을 볼 수 있다. 음절수가 가장 많은 표제어는 '프로테스탄티즘의 윤리와 자본주의의 정신'의 18음절을 가진 책명이다.

〈표 16〉 주·부표제어 음절수별 분류(방언·옛말 제외)

음절수	표제항 수		백분율	예
1음절	6,318	(714)	1.28%	강
2음절	140,836	(4,142)	28.60%	가슴
3음절	164,619	(49,272)	33.43%	가곡집
4음절	105,944	(11,634)	21.51%	가가호호
5음절	44,996	(10,427)	9.13%	가감저항기
6음절	17,095	(327)	3.47%	자동호출장치
7음절	7,917	(1,189)	1.60%	버드나무하늘소
8음절	2,823	(3)	0.57%	개구리과동편모충
9음절	1,102	(12)	0.22%	강원도자진방아타령
10음절	429		0.08%	가는다리애기좀진드기
11음절	190	(2)	0.03%	가로자기마당전류발전기
12음절	85		0.01%	고용이자및화폐의일반이론
13음절	35		0.00%	가로자기마당전기기계증폭기
14음절	11		0.00%	국제연합난민고등판무관사무소
15음절	3		0.00%	라이프치히게반트하우스관현악단
16음절	3		0.00%	감지금니대방광불화엄경보현행원품
17음절	4		0.00%	국제연합팔레스타인난민구제사업기관
18음절	1		0.00%	프로테스탄티즘의윤리와자본주의의정신
계	492,411		100.00%	

4. 맺음말

이 글은『표준국어대사전』에 수록된 정보를 통계적으로 분석해 보려는 의도에서 집필되었다. 이 글에서는 주로 기계적으로 처리될 수 있는 표제어의 항목별·기호별 분류 등을 중심으로 하여 각종 정보를 분석해 보았다.

아쉬운 점은 심도 있는 분석이 되기 위해서는 뜻풀이 항목의 각종 의미 정보도 함께 분석의 대상으로 다루었어야 하지만 그렇지 못했다는 점이다. 앞으로의 작업에서는 각종 빈도수 조사(표제어에 쓰인 음절의 빈도, 원어에 쓰인 한자어의 빈도, 뜻풀이에 쓰인 어휘의 빈도 등)와 문형 정보에 따른 동사·형용사의 구분, 색채어나 의성·의태어와 같은 어휘 부류에 대한 세부적인 분석, 그리고 표제어의 뜻풀이 방식에 따른 분류 등과 같은 구체적인 작업이 진행되어야 할 것이다.