# 국어 어휘의 분야별 분포 양상

조 남 호*

## Ⅰ. 시작하는 글

국어의 어휘는 한정된 수로 이루어져 있지 않다. 사람마다 알고 있는 단어가 다르고, 있던 단어가 없어지는 한편 새로운 단어가 계속 만들어지기 때문에 국어 어휘의 총량을 아는 사람은 없다. 어휘의 총량을 모르고 일부 단어에 대한 세밀한 검토를 통해서 국어 어휘의 특성을 밝히는 것은 한계가 있다. 컴퓨터를 이용한 대규모 자료 처리가 가능해지면서 어휘의 계량에 관심이 늘게 된 것은 어휘 계량을 통해 일부 단어에 대한 검토만으로는 드러나기 힘든 어휘의 특성을 밝히는 것이 가능해졌기 때문이다. 본고에서도 어휘의 계량을 통해 국어 어휘에 관한 하나의 면모를 밝혀 보고자 한다.

본고에서 관심을 가지고자 하는 것은 현대 국어에서 분야별로 사용되는 어휘가 어떤 공통점과 차이점이 있는가 하는 것이다. 말을 할 때와 글을 쓸 때 사용하는 어휘에서 차이가 있다. 또한 같은 글이라 해도 논리적인 성격의 글을 쓸 때와 감성적인 성격의 글을 쓸 때 어휘 사용에서 차이가 있다. 그렇다고 차이만 있는 것은 아니고 공통으로 나타나는 경향도 있을 수 있다. 본고는 몇 개의 분야를 구분하여 그 분야에서 사용되는 어휘를 계량하여 비교함으로써 각각의 분야에서 드러나는 어휘 사용 양상을 밝히고자 한다. 어휘 계량에 이용한 자료는 필자가 2000년부터 담당 연구원으로 책임을 맡아 3년간 150만 어절 분량의 말뭉치에서 빈도 조사를 한 자료이다.[1] 이

* 국립국어연구원 학예연구관

1) 이 작업은 한국어 학습용 어휘 선정을 위한 기초 조사로 이루어진 것으로 현대 국

자료는 9개 분야로 나누어 각각의 분야에서의 사용 빈도 및 전체 150만 어절에서의 사용 빈도를 조사한 결과를 담고 있다.

그런데 이 자료는 분야별로 조사한 말뭉치의 양이 균등하지 않다. 조사의 목적이 현대 국어에서 사용되는 어휘의 빈도를 파악하는 데 있었지 분야별 경향을 파악하기 위한 데 있었지 않았기 때문이다. 이에 따라 분야별로 비중에 따라 조사 대상 말뭉치의 양에 차등을 두었다. 말뭉치의 양이 균등하지 않은 상태대로의 자료를 분석하는 것도 충분한 의의가 있기는 하지만 논의를 진행하는 중에 때로 분야별로 말뭉치의 양을 균등하게 하여 조사한 빈도 자료를 확인할 필요가 있다. 이를 위해 분야별로 40,000단어에서 조사한 빈도 자료를 보조 자료로 이용하도록 한다.

## II. 자료 소개

본고의 논의를 위해 이용한 자료는 졸저(2002)에 수록된 빈도 조사 결과이다. 이 자료는 이미 앞서 언급했듯이 9개의 분야로 나누어 각 분야별로 빈도를 냈기 때문에 본고에서도 9개로 분야를 나누어 어휘의 분포 양상을 살피고자 한다.

신뢰할 만한 빈도 조사가 이루어지기 위해서는 분야의 구분 및 분야별 조사 양 배정 비율의 결정이 중요하다. 이에 대한 국내외의 선행 연구도 여럿 있다. 그렇지만 아직 충분하게 검증이 되어 객관적으로 믿고 따를 만한 결론이 내려지지는 않은 상태이다. 그래서 선행 연구를 참조하기는 했지만 최종적으로는 필자의 주관적인 판단에 의존하여 분야를 구분할 수밖에 없었다. 졸저(2002)를 작성하면서 필자는 몇 차례 수정 끝에 최종적으로 9개 분야로 구분하였다. 9개의 분야는 교재, 교과서, 교양, 문학, 신문, 잡지, 대본, 구어, 기타이다. '교재'에 포함된 것은 외국인에게 한국어를 가르치기 위해 만든 한국어 교재들에 나오는 본문이다. 졸저(2002)가 한국어 학습용 어휘를 선정하기 위한 목적으로 진행된 것이기 때문에 한국어 교재를 빈도 조사

---

어 사용 빈도를 일반 단어, 조사, 어미, 고유명사로 나누어 조사하였다. 빈도 조사 결과를 보고서로 낸 것이 졸저(2002)이다.

대상 자료로 포함한 것이고 분야도 따로 설정한 것이다. '교과서'에는 주로 초등학교에서 사용하는 교과서를 포함하였으며 일부 중학교 교과서도 포함하였다. '문학'은 대부분 소설로 이루어져 있으며 동화를 일부 포함하였다. '신문, 잡지, 대본, 구어'에 대한 설명은 생략한다. 필자가 분야를 구분하면서 가장 고심한 것은 '교양'과 '기타'이었다. 최종적으로 '교양'에는 인문, 사회, 자연과학 분야에 관한 글들로 교양서의 성격이 있다고 판단한 것을 포함하였다. '기타'에는 문학적인 성격의 글이기는 하지만 정식 문학 작품으로 보기 어려운 것들을 배당하였다.

150만 어절의 규모이기는 하지만 한 종의 문헌에서 추출하는 분량은 그 문헌 전체 분량의 1/3이 넘지 않는 것을 원칙으로 하였다. 그렇지만 어절 규모가 적은 경우에는 전체를 다 조사하기도 하였다. 이렇게 해서 조사된 자료의 규모를 제시하면 아래와 같다.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 단어 총수(A) | 73,885 | 103,562 | 372,112 | 273,977 | 289,198 | 207,129 | 40,929 | 46,221 | 77,450 | 1,484,463 |
| 비율ㄱ(%) | 4.98 | 6.98 | 25.07 | 18.46 | 19.48 | 13.95 | 2.76 | 3.11 | 5.22 | |
| 단어 종수(B) | 9,512 | 10,464 | 28,671 | 23,666 | 25,429 | 22,769 | 6,192 | 4,636 | 11,059 | 58,437 |
| 비율ㄴ(%) | 16.28 | 17.91 | 49.06 | 40.50 | 43.52 | 38.96 | 10.60 | 7.93 | 18.92 | |
| 평균(A/B) | 7.77 | 9.90 | 12.98 | 11.58 | 11.37 | 9.10 | 6.61 | 9.97 | 7.00 | 25.40 |

〈표 1〉 분야별 단어 분포 현황

〈표 1〉에서 '단어 총수'라고 한 것은 150만 어절 분량의 자료 중에서 나타난 단어들의 합이다. 본고에서는 졸저(2002)에서 규정한 단어만을 논의의 대상으로 삼기 때문에 단어로 조사된 것들의 전체 합계는 150만에서 약간 부족한 1,484,463개로 표에 나와 있다.[2] 빈도 조사를 하면서 분야별로 조사

2) 어디까지 단어로 보았는지는 V. 품사별 분포 양상에 나온다. '어절'이라고 한 것은 띄어쓰기를 기준으로 한 것인데 조사 대상이 되었던 말뭉치의 띄어쓰기가 일관성이 없었기 때문에 하나의 어절에서 두 개 이상의 단어가 나오기도 하고 여러 어절이 하나의 단어로 통합되기도 하였다. 또 단어로 보지 않은 항목이 포함된 어절은 어절은 하나이어도 단어는 0개로 처리되기도 하였다. 단어에서 제외한 고유명사가 들어간 어절이 그 예가 된다. 이런 점을 감안하면 어절의 개수와 단어의 개

비율에 차등을 두었기 때문에 단어 총수가 분야마다 다르다. '교양'에서 제일 많이 조사하였으며 다음으로 '신문, 문학' 등의 순서로 많이 조사하였다. '비율ㄱ'에 제시한 수치는 각 분야에서 조사된 단어 총수를 총합인 1,484,463개로 나누어 얻은 값이다.

'단어 종수'는 중복해서 나타난 단어들은 무조건 1로 계산하여 단어들의 개수를 합산한 것이다.[3] '전체'를 예로 들어 설명하면 1,484,463개를 조사한 결과 58,437개의 단어가 확인된 것이다. 여러 분야에 걸쳐 나타나는 단어는 그 분야에서 각각 계산됐기 때문에 9개 분야의 단어 종수를 합하면 '전체'보다 많다. '비율ㄴ'에 제시한 수치는 '비율ㄱ'과 마찬가지로 단어 종수를 총합인 58,437개로 나누어 얻은 값이다.

마지막 줄에 제시한 '평균'은 단어 총수를 단어 종수로 나누어 산출한 것이다. 한 단어의 평균 출현 횟수가 되는 셈이다. 예를 들어 '교재'의 경우 한 단어당 평균 7.77회 사용되었다고 할 수 있다. '평균'에 제시된 수치가 높을수록 단어를 반복하여 사용한 비율이 높다고 해야 한다. 그렇지만 이 표만으로 평균이 높게 나온 분야가 적게 나온 분야보다 단어의 반복 사용율이 높다고 단정할 수는 없다. 단어 총수에서 차이가 있기 때문이다. 단어 총수가 늘어날수록 반복적으로 사용되는 단어가 많게 되면 평균도 올라갈 가능성이 있다. 아래에서 이 문제를 다시 다루게 된다.

보조 자료는 1,484,463개 중에서 분야마다 똑같은 양을 추출하여 만들었다. 즉, 9개의 분야에서 각각 40,000개씩 뽑아 만들었다. 40,000개로 한 이유는 '대본' 때문이다. 〈표 1〉에서 보면 단어 총수가 가장 적은 것이 '대본'으로 40,929개이다. 그래서 929개를 제외한 40,000개를 추출하였고 이것이 기준이 되었기 때문에 다른 분야에서도 40,000개씩 추출한 것이다. 40,000개라는 것은 어절 수가 아니다. 단어로 규정한 것 중에서 뽑은 것이다. 분야별로 포함되는 말뭉치 파일의 개수도 되도록 똑같도록 하였다. '대본'이 12

---

수는 실제로는 많은 차이가 있다.

3) 본고에서 잠정적으로 사용하는 '단어 총수, 단어 종수'는 김광해(1993:73)에서는 '연어휘, 개별어휘'라고 한 바 있으며, '토큰, 타입'이라는 용어가 사용되기도 한다.

개 파일로 구성되어 있기 때문에 다른 분야에서도 12개 파일에서 40,000개를 뽑았다. 다만, '구어'와 '기타'는 파일의 수가 12개가 안 되어서 각각 8개 파일과 6개 파일에서 40,000개를 뽑았다.[4] 이렇게 하여 총 360,000개 단어로 보조 자료를 만들었다.

조사된 자료의 규모를 제시하면 아래와 같다.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 단어 총수(A) | 40,000 | 40,000 | 40,000 | 40,000 | 40,000 | 40,000 | 40,000 | 40,000 | 40,000 | 360,000 |
| 단어 종수(B) | 6,718 | 5,991 | 7,863 | 8,006 | 8,690 | 9,249 | 6,083 | 4,363 | 7,254 | 28,317 |
| 비율(%) | 23.72 | 21.16 | 27.77 | 28.27 | 30.69 | 32.66 | 21.48 | 15.41 | 25.62 | |
| 평균(A/B) | 5.95 | 6.68 | 5.09 | 5.00 | 4.60 | 4.32 | 6.58 | 9.17 | 5.51 | 12.71 |

〈표 2〉 보조 자료의 분야별 단어 분포 현황

'비율'에 제시한 수치는 〈표 1〉에서와 마찬가지로 단어 종수를 총합인 28,317개로 나누어 얻은 값이다. 〈표 1〉과 비교하면 수치가 크게 차이를 보인다. '평균'도 앞서 〈표 1〉에서와 마찬가지로 단어 총수를 단어 종수로 나누어 산출하였다. 〈표 1〉에서 평균이 높았던 '교양, 문학, 신문'의 평균이 오히려 다른 분야보다 낮게 나오는 것을 볼 수 있다. 이 세 분야는 다른 분야보다 단어 총수가 많았던 분야인 만큼 보조 자료를 만들기 위해 조사 양이 줄어든 비율이 상대적으로 높은 분야이다. 평균적으로 단어의 사용 횟수가 높았던 것이 아니라 단어 총수가 다른 분야보다 많았기 때문에 단어의 반복 사용이 그만큼 많아져 높았던 것임을 알 수 있다. 〈표 2〉를 기준으로 하면 '구어'에서 단어의 반복 사용이 제일 빈번한 것으로 나온다. 다른 분야와 차별되는 구어의 특성인지 아니면 본 조사에 이용된 자료의 특성에 기인한 것인지는 좀더 많은 자료를 검토한 후에 결론을 내려야 할 것으로 보인다.

4) 파일 목록에 대한 자세한 정보는 졸저(2002)에 나온다. 12개의 파일은 졸저(2002)에서 배열된 순서대로 먼저 앞 파일에서부터 뽑는 방식으로 하였다. 12개의 파일에서는 되도록 똑같은 개수의 단어를 추출하도록 하였으나 단어 개수가 적은 파일은 전체를 다 포함하였다.

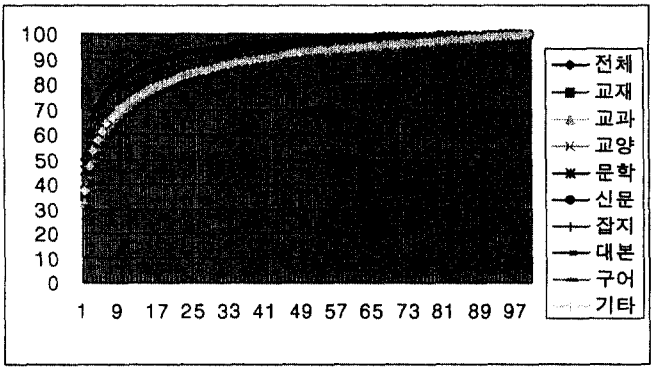
## Ⅲ. 어휘의 누적 사용 비율

여기서는 각 분야에서 빈도가 높은 순서대로 단어를 배열하고 단어 종수의 비율이 늘어날수록 단어 총수의 누적 사용 비율이 어떻게 변하는지를 살피도록 한다. 분야별로 누적 사용 빈도를 산출하여 이를 비율로 환산하여 정리한 것이 아래의 표이다. 대략 10%씩 증가할 때의 변화 양상을 볼 수 있도록 하였다.

| 비율 | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1% | 35.1316 | 32.6306 | 42.5764 | 45.3760 | 36.2696 | 36.3150 | 30.7899 | 38.7421 | 37.3105 | 49.7859 |
| 10% | 71.7439 | 69.6703 | 77.4272 | 77.2572 | 72.8359 | 71.8146 | 67.3629 | 76.5020 | 71.5403 | 82.8274 |
| 20% | 82.0572 | 81.4970 | 86.4613 | 85.8583 | 83.5482 | 82.4761 | 78.6092 | 85.2166 | 81.1865 | 90.3701 |
| 30% | 87.3411 | 87.6199 | 90.9777 | 90.2820 | 89.0338 | 87.9041 | 84.8053 | 89.6865 | 86.4506 | 93.9584 |
| 40% | 90.7098 | 91.3298 | 93.6712 | 93.1355 | 92.3726 | 91.4381 | 88.9100 | 92.4925 | 89.9238 | 96.0413 |
| 50% | 93.2841 | 93.8693 | 95.4709 | 95.0050 | 94.6192 | 93.7512 | 91.9348 | 94.4960 | 92.7798 | 97.3645 |
| 60% | 94.8514 | 95.8913 | 96.9181 | 96.5449 | 96.3779 | 95.6032 | 93.9504 | 95.9888 | 94.2892 | 98.2136 |
| 70% | 96.1385 | 96.9689 | 97.6885 | 97.4089 | 97.3623 | 96.7025 | 95.4628 | 96.9927 | 95.7172 | 98.8190 |
| 80% | 97.4257 | 97.9799 | 98.4590 | 98.2724 | 98.2416 | 97.8018 | 96.9752 | 97.9944 | 97.1452 | 99.2127 |
| 90% | 98.7128 | 98.9899 | 99.2295 | 99.1364 | 99.1210 | 98.9011 | 98.4876 | 98.9982 | 98.5732 | 99.6063 |
| 99% | 99.8714 | 99.8995 | 99.9231 | 99.9138 | 99.9121 | 99.8904 | 99.8509 | 99.9004 | 99.8579 | 99.9606 |

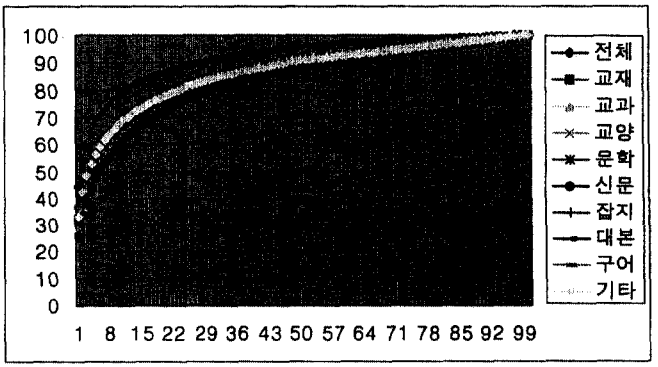
〈표 3〉 누적 사용 빈도의 증가 비율

표에서 맨 왼쪽에 %로 표시된 것은 단어 종수에서의 비율이다. '전체'를 예로 들면 단어 종수가 58,437개이므로 1%이면 빈도 순위로 1위부터 584위까지의 단어가 포함된 것이다. 분야마다 단어 종수가 다르므로 1%에 드는 단어의 개수도 당연히 다르다. 〈표 3〉을 보면 분야의 구분 없이 빈도에서 상위를 차지하는 단어에서 단어 총수의 비율이 가파르게 상승함을 알 수 있다. 1%가 차지하는 비율도 높을 뿐더러 상위 10%에 이르면 '전체'의 경우는 단어 총수의 82% 가까운 비율로 나타난다. 즉, 불과 10%에 속하는 단어가 전체 사용 횟수의 82%를 차지하는 것이다. 그렇지만 그 이후로는 비율의 상승 폭이 급격히 둔화된다. 이를 보기 쉽게 차트로 표시하면 아래와 같다.




〈차트 1〉 누적 빈도 증가 모습

처음에는 가파르게 상승하다가 뒤로 갈수록 완만해지는 것을 볼 수 있다. 이러한 모양의 분포 곡선은 새삼스러운 것은 아니며 이미 기존의 조사에서도 드러나서 널리 알려져 있는 사실이다. 분포 곡선의 모양이 분야마다 크게 차이를 보이지 않고 비슷한 모양을 보이는 점도 흥미롭다. 분야와 상관없이 단어 종수의 증가에 따른 단어 총수의 증가 비율은 비슷함을 이 차트에서 쉽게 파악할 수 있다. 아래 차트 2는 보조 자료의 것으로 자료의 양이 적더라도 누적 빈도의 증가 양상은 비슷함을 알 수 있다.




〈차트 2〉 보조 자료 누적 빈도 증가 모습

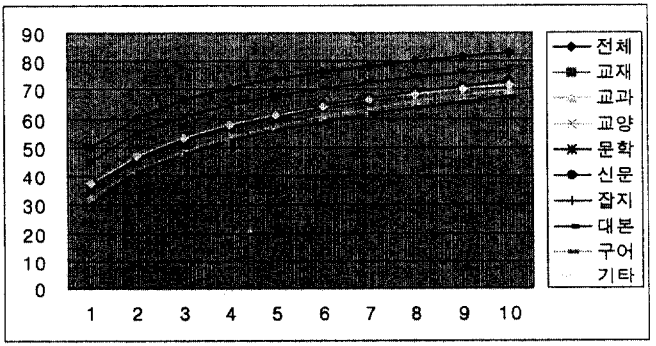
단어 종수의 비율이 높아질수록 단어 총수의 상승 폭이 급격하게 둔화되는 것은 저빈도로 출현하는 단어의 수가 많은 것과 연관된다.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 단어 종수 | 9,512 | 10,464 | 28,671 | 23,666 | 25,429 | 22,769 | 6,192 | 4,636 | 11,059 | 58,437 |
| 1회 출현 단어 | 4,550 | 4,115 | 11,817 | 9,981 | 9,867 | 9,825 | 2,891 | 2,092 | 5,466 | 20,231 |
| 비율(%) | 47.8 | 39.3 | 41.2 | 42.1 | 38.8 | 43.1 | 46.6 | 45.1 | 49.4 | 34.6 |

〈표 4〉 1회 출현 단어의 개수와 비율

〈표 4〉는 각 분야에서 나타나는 단어 중에서 1회만 출현하는 단어의 비율을 조사한 것이다. 분야마다 차이가 있기는 하지만 전체 단어 중에서 1회만 출현하는 단어의 비율이 34%~49% 사이에서 나타난다. 이는 〈표 3〉에서 66%~51%의 위치에 해당하는 것으로 상승 폭이 매우 둔화되는 곳이다. 보조 자료는 따로 표를 제시하지 않지만 비율이 더 높아 40%~53% 사이에서 나타난다.

상위 빈도에서 급격하게 변화를 보이므로 상위 10%에 대해서만 1% 간격으로 누적 사용 빈도가 어떻게 변화하는지를 차트로 나타내면 아래와 같다.




〈차트 3〉 누적 빈도 증가 모습(상위 10%)

이 차트에서 보면 상위 10%에 도달할 때도 분야와 상관없이 상승 폭이

비슷함을 알 수 있다. 단지 '구어'에서 상승 폭이 좀더 가파르게 나타날 뿐이다.

## IV. 기원별 분포 양상

우리 국어의 중요한 특징 중의 하나로 꼽히는 것은 한자어가 많다는 점이다. 『표준국어대사전』을 대상으로 조사한 이운영(2002:50)에 따르면 사전에 수록된 표제어의 기원별 분포는 아래의 표와 같다.

| 구분 | 고유어 | 고+한 | 고+외 | 고+한+외 | 한자어 | 한+외 | 외래어 | 계 |
|---|---|---|---|---|---|---|---|---|
| 횟수 | 111,156 | 36,618 | 1,323 | 720 | 252,278 | 14,480 | 24,019 | 440,594 |
| 비율(%) | 25.23 | 8.31 | 0.3 | 0.16 | 57.26 | 3.29 | 5.45 | 100 |

〈표 5〉『표준국어대사전』 표제어의 기원별 분포

〈표 5〉는 주표제어만 조사한 것으로 부표제어 68,482개는 조사에서 빠졌다. 따라서 모든 단어에 대한 것은 아니지만 대체적인 경향을 볼 수는 있다. 이 자료에 따르면 사전에 실린 단어 중에서 순수한 한자어만도 57.26%에 이른다. 그러면 실제 빈도 조사 자료에서 나타나는 양상은 어떨까?

먼저 단어 종수를 기준으로 하여 기원별 분포를 살피도록 한다.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 고유어 | 37.67 | 35.95 | 22.25 | 40.16 | 17.3 | 28.26 | 56.88 | 38.46 | 41.44 | 24.05 |
| 고+한 | 17.3 | 17.66 | 19.88 | 18.12 | 19.01 | 18.31 | 13.52 | 14.37 | 16.6 | 18.06 |
| 고+외 | 0.13 | 0.14 | 0.15 | 0.22 | 0.18 | 0.36 | 0.39 | 0.41 | 0.07 | 0.37 |
| 고+한+외 | 0 | 0 | 0.05 | 0.01 | 0.01 | 0.02 | 0.02 | 0 | 0 | 0.04 |
| 한자어 | 41.69 | 43.13 | 54.08 | 37.93 | 57.74 | 47.15 | 24.52 | 41.26 | 39.19 | 51.8 |
| 한+외 | 0.32 | 0.39 | 0.53 | 0.38 | 0.84 | 0.61 | 0.32 | 0.3 | 0.16 | 0.94 |
| 외래어 | 2.89 | 2.72 | 3.06 | 3.17 | 4.94 | 5.28 | 4.36 | 5.2 | 2.53 | 4.74 |

〈표 6〉 단어 종수에서의 기원별 분포

'전체'에서의 비율을 기준으로 하면 많이 나타나는 순서가 한자어, 고유어, 고+한, 외래어, 한+외, 고+외, 고+한+외로 〈표 5〉와 〈표 6〉이 동일하다. 즉, 사전에서의 순서대로 실제 자료에서도 나오는 것이다. 그렇지만 기원별로 차지하는 비율은 차이가 있다. 특히 고+한 구성의 단어 비율이 10% 가까이 차이가 난다. 이 차이는 앞서 〈표 5〉에서 부표제어가 합산에서 빠진 것과 관련될 듯하다. 부표제어에서 많은 비중을 차지하는 'ㅇ하다, ㅇ되다'에서 고유어+한자어의 결합이 많이 나타나는데 표 5에서는 이런 종류의 단어를 대부분 제외하고 계산하였고 〈표 6〉에서는 포함하고 계산하였다.

그런데 분야를 나누어 살펴보면 기원별 분포는 분야마다 크게 차이를 보이고 있다. '교양, 신문, 잡지'에서는 한자어의 비율이 고유어의 비율보다 매우 높게 나타난다. '대본'에서는 반대로 고유어의 비율이 매우 높게 나타나며, '문학, 기타'에서도 고유어의 비율이 더 높다. 나머지 '교재, 교과, 구어'에서는 한자어의 비율이 상대적으로 고유어보다 높게 나타난다.

이제 단어 총수에서의 분포 양상을 보도록 하자.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 고유어 | 67.12 | 58.25 | 44.64 | 71.43 | 34.35 | 54.04 | 78.49 | 73.56 | 69.93 | 54.11 |
| 고+한 | 6.73 | 9.54 | 11.21 | 7.39 | 10.52 | 8.85 | 5.46 | 5.33 | 6.19 | 9.1 |
| 고+외 | 0.03 | 0.03 | 0.02 | 0.04 | 0.03 | 0.07 | 0.09 | 0.05 | 0.01 | 0.04 |
| 고+한+외 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 0 | 0 |
| 한자어 | 24.78 | 30.99 | 42.34 | 20.01 | 52.02 | 33.57 | 13.89 | 19.29 | 22.29 | 34.67 |
| 한+외 | 0.09 | 0.1 | 0.13 | 0.06 | 0.19 | 0.14 | 0.16 | 0.03 | 0.05 | 0.12 |
| 외래어 | 1.25 | 1.08 | 1.64 | 1.07 | 2.89 | 3.31 | 1.91 | 1.74 | 1.52 | 1.96 |

〈표 7〉 단어 총수에서의 기원별 분포

〈표 7〉과 〈표 6〉을 비교하면 단어 총수에서는 단어 종수와 사뭇 다른 양상을 보임을 알 수 있다. 대체적으로 한자어보다 고유어의 비율이 높게 나타난다.[5] '신문'만 예외적으로 한자어의 비율이 고유어보다 높다. 단어 총수에서 고유어 비율이 높아진 것은 고유어는 반복해서 사용되는 비율이 높

5) 기원별로 어휘 출현 빈도를 조사하면 단어 총수에서 고유어가 많이 나타난다는 지적은 이미 있었다. 임칠성 외(1997:18) 참조.

은 데 비해 한자어는 반복해서 사용되는 비율이 낮은 데 원인이 있을 것이다. VI. 상위 빈도에서의 분포에서 자세히 밝히겠지만 상위 빈도에서는 고유어의 개수가 한자어의 개수를 크게 앞선다.

보조 자료에서도 단어 종수이든 단어 총수이든 기원별 분포에서 지금까지 논의한 바와 경향이 크게 다르지 않다. 구체적인 수치에서는 차이가 있으나 이로 인해 순위가 바뀌거나 하지 않는다. 다만, 단어 총수에서 '교양'이 40.89:45.69로 한자어가 더 많은 것으로 나와 순위가 바뀐다. 한자어가 더 많은 비중을 차지하는 분야가 '신문'에서 '교양, 신문' 둘이 되는 것이다.

## V. 품사별 분포 양상

졸저(2002)에서는 어휘 빈도를 조사할 때 21세기 세종계획에서 정한 분석 표지를 그대로 이용하였다. 다만 '분석 불능' 범주를 설정하여 그에 대한 하위 범주로 '통계 가치가 있는 범주'와 '통계 가치가 없는 범주'를 추가하였다.[6] 이 중에서 단어에 속하는 것으로 '일반명사(NNG), 의존명사(NNB), 대명사(NP), 수사(NR), 동사(VV), 형용사(VA), 보조용언(VX), 부정지정사(VCN), 관형사(MM), 일반부사(MAG), 접속부사(MAJ), 감탄사(IC), 통계 가치가 있는 범주(NV)[7]'를 인정하였다. 학교 문법에서는 조사를 단어로 인정하고 있으나 졸저(2002)에서는 단어로 인정하지 않았으며 고유명사도 제외하였다. 여기서는 졸저(2002)의 기준을 그대로 따르면서 품사별 분포 양상을 살피도록 한다.[8]

먼저 단어 총수를 기준으로 했을 때의 분야별 양상부터 살피도록 한다. 지면을 절약하기 위해 개수는 밝히지 않고 비율로만 제시하도록 한다.

---

6) 자세한 사항은 졸저(2002)에 밝혔다.

7) 통계 가치가 있는 범주로 처리한 항목들은 '그래, 어때, 저래, 개, 암말, 인마'와 같이 줄어든 말이어서 『표준국어대사전』에서 품사를 배정하지 않은 것들이다. 분석하기 곤란하다고 판단하여 따로 분석 표지를 배당한 것이다.

8) 통계 가치가 있는 범주는 품사로 인정할 수는 없으나 여기서는 품사처럼 취급하도록 한다. 앞으로는 '통계 가치가 있는 범주'라는 용어 대신 글자 수를 줄이기 위하여 '분석불능'이라는 용어를 사용하도록 한다.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 일반명사 | 38.29 | 46.23 | 48.10 | 36.50 | 54.39 | 46.41 | 32.18 | 26.40 | 39.36 | 44.76 |
| 동사 | 22.64 | 24.49 | 20.05 | 24.15 | 18.71 | 21.79 | 29.62 | 20.30 | 23.55 | 21.68 |
| 형용사 | 9.32 | 7.91 | 7.44 | 7.53 | 4.96 | 7.29 | 6.16 | 8.64 | 7.62 | 7.09 |
| 의존명사 | 6.39 | 4.92 | 6.35 | 5.99 | 9.14 | 7.31 | 4.95 | 7.29 | 5.98 | 6.83 |
| 일반부사 | 7.33 | 4.49 | 4.76 | 7.56 | 3.63 | 5.35 | 9.42 | 10.96 | 7.47 | 5.71 |
| 보조용언 | 4.32 | 5.01 | 3.76 | 5.48 | 3.77 | 4.60 | 4.35 | 3.39 | 4.52 | 4.36 |
| 관형사 | 3.55 | 2.82 | 4.45 | 3.79 | 2.56 | 2.90 | 2.97 | 5.98 | 3.89 | 3.56 |
| 대명사 | 4.09 | 2.47 | 2.62 | 6.25 | 1.47 | 2.55 | 5.28 | 6.85 | 4.88 | 3.44 |
| 접속부사 | 1.79 | 1.20 | 1.67 | 1.18 | 1.00 | 1.05 | 0.84 | 3.00 | 1.39 | 1.34 |
| 감탄사 | 1.40 | 0.19 | 0.09 | 0.57 | 0.01 | 0.12 | 2.88 | 4.70 | 0.52 | 0.48 |
| 부정지정사 | 0.42 | 0.17 | 0.50 | 0.51 | 0.27 | 0.40 | 0.49 | 0.82 | 0.41 | 0.42 |
| 수사 | 0.24 | 0.08 | 0.18 | 0.31 | 0.08 | 0.15 | 0.28 | 0.45 | 0.24 | 0.19 |
| 분석불능 | 0.22 | 0.02 | 0.04 | 0.17 | 0.01 | 0.06 | 0.58 | 1.22 | 0.17 | 0.13 |

〈표 8〉 단어 총수의 품사별 분포 비율

'전체'를 기준으로 비율이 높게 나타나는 품사부터 배열하였다.

균등한 조사 양에서 빈도를 추출한 자료인 보조 자료에서의 품사별 비율은 아래와 같다.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 일반명사 | 37.91 | 42.59 | 49.43 | 37.32 | 55.17 | 46.72 | 32.02 | 26.82 | 39.37 | 40.81 |
| 동사 | 23.06 | 26.43 | 20.06 | 23.96 | 17.99 | 21.24 | 29.79 | 20.28 | 23.87 | 22.96 |
| 형용사 | 9.40 | 8.25 | 6.97 | 7.81 | 5.03 | 7.47 | 6.15 | 8.55 | 7.72 | 7.48 |
| 의존명사 | 6.31 | 4.38 | 6.35 | 6.02 | 8.88 | 7.52 | 5.00 | 7.35 | 5.90 | 6.41 |
| 일반부사 | 7.26 | 4.69 | 3.95 | 7.58 | 3.79 | 5.43 | 9.34 | 10.65 | 7.57 | 6.69 |
| 보조용언 | 4.65 | 5.75 | 3.81 | 5.35 | 3.71 | 4.59 | 4.38 | 3.40 | 4.52 | 4.46 |
| 관형사 | 3.63 | 3.11 | 4.69 | 3.88 | 2.61 | 2.81 | 2.99 | 5.89 | 3.81 | 3.71 |
| 대명사 | 4.19 | 3.09 | 2.33 | 5.27 | 1.53 | 2.52 | 5.3 | 6.69 | 4.76 | 3.96 |
| 접속부사 | 1.70 | 1.10 | 1.76 | 1.22 | 0.90 | 1.05 | 0.85 | 3.01 | 1.44 | 1.45 |
| 감탄사 | 0.98 | 0.30 | 0.01 | 0.59 | 0.01 | 0.12 | 2.84 | 4.86 | 0.34 | 1.12 |
| 부정지정사 | 0.49 | 0.18 | 0.46 | 0.52 | 0.28 | 0.35 | 0.50 | 0.82 | 0.38 | 0.44 |
| 수사 | 0.20 | 0.10 | 0.18 | 0.35 | 0.10 | 0.15 | 0.28 | 0.47 | 0.20 | 0.22 |
| 분석불능 | 0.24 | 0.03 | 0.03 | 0.14 | 0.02 | 0.05 | 0.58 | 1.22 | 0.15 | 0.27 |

〈표 9〉 보조 자료 단어 총수의 품사별 분포 비율

품사의 배열 순서는 〈표 8〉의 배열 순서를 따랐다. 〈표 8〉과 〈표 9〉를 비교하면 둘 사이의 분포가 크게 차이가 나지 않음을 볼 수 있다. 1% 이상의 차이를 보이는 것은 '전체'의 일반명사와 동사, '교과'의 일반명사와 동사, '교양'의 일반명사뿐이다. 그나마 최대 4%의 차이도 안 난다. '대본'과 '구어'의 경우에는 두 자료 사이의 조사 양이 크게 차이가 나지 않아 논외로 해도 조사 양이 상당히 다른 '교양', '문학', '신문', '잡지' 등에서도 비슷한 분포를 보이는 것을 보면 위의 표에 나오는 분포가 해당 분야의 품사별 분포를 보여주는 자료가 될 가능성이 없지 않다.

분야별로 품사 분포의 차이가 어떻게 나는지를 파악하기 위하여 〈표 8〉을 기준으로 각 분야에서 높게 나타나는 품사 순서대로 배열해 보자.

| 전체 | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 |
|---|---|---|---|---|---|---|---|---|---|
| 일반명사(①) | ① | ① | ① | ① | ① | ① | ① | ① | ① |
| 동사(②) | ② | ② | ② | ② | ② | ② | ② | ② | ② |
| 형용사(③) | ③ | ③ | ③ | ⑤ | ④ | ④ | ⑤ | ⑤ | ③ |
| 의존명사(④) | ⑤ | ⑥ | ④ | ③ | ③ | ③ | ③ | ③ | ⑤ |
| 일반부사(⑤) | ④ | ④ | ⑤ | ⑧ | ⑥ | ⑤ | ⑧ | ④ | ④ |
| 보조용언(⑥) | ⑥ | ⑤ | ⑦ | ④ | ⑤ | ⑥ | ④ | ⑧ | ⑧ |
| 관형사(⑦) | ⑧ | ⑦ | ⑥ | ⑥ | ⑦ | ⑦ | ⑥ | ⑦ | ⑥ |
| 대명사(⑧) | ⑦ | ⑧ | ⑧ | ⑦ | ⑧ | ⑧ | ⑦ | ⑩ | ⑦ |
| 접속부사(⑨) | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑨ | ⑩ | ⑥ | ⑨ |
| 감탄사(⑩) | ⑩ | ⑩ | ⑪ | ⑩ | ⑪ | ⑪ | ⑨ | ⑨ | ⑩ |
| 부정지정사(⑪) | ⑪ | ⑪ | ⑫ | ⑪ | ⑫ | ⑫ | ⑬ | ⑬ | ⑪ |
| 수사(⑫) | ⑫ | ⑫ | ⑩ | ⑫ | ⑬ | ⑩ | ⑪ | ⑪ | ⑫ |
| 분석불능(⑬) | ⑬ | ⑬ | ⑬ | ⑬ | ⑩ | ⑬ | ⑫ | ⑫ | ⑬ |

〈표 10〉 단어 총수에서의 분야별 품사 순위

'전체'를 기준으로 하여 '전체'에서의 순위가 다른 분야에서는 어떻게 변동이 되었는지를 표시하였다. '전체'에서 원문자로 순위를 표시하였고 이 원문

자로 표시된 품사가 각 분야에서 차지하는 순위에 해당 원문자를 제시하였다. 색으로 표시된 부분이 '전체'와 비교했을 때 순위에서 차이가 나는 것들이다. '교과'가 보조용언이 순위의 변동을 준 것을 빼면 나머지 순위는 '전체'와 동일하여 가장 비슷한 모습을 보이며 '대본'과 '구어'는 형용사 이후의 순위가 모두 '전체'와 차이를 보여 가장 다른 모습을 보인다. 〈표 10〉에서 나타난 바로는 일반명사와 동사는 어느 분야에서든 1위와 2위를 차지하여 부동의 지위에 있음을 알 수 있다.

이제 단어 종수에서의 차이를 보도록 하자.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 일반명사 | 58.55 | 61.17 | 64.35 | 57.64 | 68.37 | 64.62 | 46.35 | 51.32 | 57.70 | 68.20 |
| 동사 | 22.07 | 23.10 | 21.38 | 22.63 | 20.68 | 20.99 | 25.95 | 23.25 | 22.21 | 18.68 |
| 형용사 | 7.43 | 6.41 | 5.49 | 7.85 | 4.38 | 6.33 | 8.38 | 6.97 | 7.90 | 4.66 |
| 일반부사 | 6.07 | 4.90 | 3.60 | 7.14 | 2.93 | 4.59 | 11.16 | 7.59 | 7.07 | 4.02 |
| 관형사 | 1.75 | 1.30 | 3.46 | 1.69 | 2.03 | 1.38 | 1.45 | 3.11 | 1.24 | 2.53 |
| 의존명사 | 1.53 | 1.34 | 0.77 | 0.98 | 1.01 | 0.92 | 1.57 | 1.96 | 1.44 | 0.66 |
| 감탄사 | 0.81 | 0.48 | 0.29 | 0.71 | 0.05 | 0.29 | 1.92 | 1.57 | 0.70 | 0.43 |
| 수사 | 0.33 | 0.21 | 0.13 | 0.40 | 0.09 | 0.16 | 0.44 | 1.16 | 0.27 | 0.27 |
| 대명사 | 0.55 | 0.43 | 0.20 | 0.38 | 0.15 | 0.26 | 1.13 | 1.01 | 0.55 | 0.22 |
| 분석불능 | 0.26 | 0.09 | 0.07 | 0.18 | 0.05 | 0.11 | 0.57 | 0.75 | 0.24 | 0.11 |
| 접속부사 | 0.29 | 0.26 | 0.12 | 0.19 | 0.12 | 0.15 | 0.37 | 0.67 | 0.31 | 0.10 |
| 보조용언 | 0.36 | 0.31 | 0.14 | 0.20 | 0.15 | 0.18 | 0.66 | 0.60 | 0.36 | 0.10 |
| 부정지정사 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.02 | 0.02 | 0.01 |

〈표 11〉 단어 종수의 품사별 분포 비율

역시 '전체'를 기준으로 높은 순위부터 배열하였다. 〈표 11〉과 〈표 8〉을 비교해 보면 순위에서 차이가 있음을 알 수 있다. '전체'를 기준으로 해서 보면 일반명사, 동사, 형용사를 제외한 나머지의 순위는 다르다. 예를 들어 보조용언의 경우 〈표 8〉에서는 6위에 있었으나 지금은 12위로 밀려 있다. 이것은 하나의 단어가 반복해서 나타나는 비율이 높은 데 원인이 있다. 단

어 종수는 적지만 개개의 단어가 자주 사용되어 단어 총수는 높았던 것이다. 반대로 끌찌였던 분석불능이 9위로 올라선 것은 그만큼 단어가 반복 사용된 비율이 낮다는 것을 의미한다.

보조 자료에서는 어떻게 나타나는지를 정리한 것이 〈표 12〉이다.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 일반명사 | 56.39 | 54.65 | 57.97 | 50.04 | 64.82 | 59.92 | 46.13 | 51.34 | 55.47 | 62.92 |
| 동사 | 22.37 | 25.22 | 25.79 | 25.08 | 22.05 | 23.18 | 26.14 | 22.60 | 23.04 | 20.97 |
| 형용사 | 8.37 | 8.20 | 5.96 | 9.47 | 4.74 | 7.13 | 8.40 | 7.04 | 8.11 | 6.07 |
| 일반부사 | 6.61 | 6.41 | 4.21 | 9.17 | 3.92 | 5.47 | 11.10 | 7.72 | 7.64 | 5.31 |
| 관형사 | 1.52 | 1.37 | 3.57 | 1.54 | 2.07 | 1.32 | 1.48 | 3.23 | 1.41 | 2.03 |
| 의존명사 | 1.74 | 1.52 | 1.25 | 1.74 | 1.38 | 1.50 | 1.59 | 2.09 | 1.75 | 0.85 |
| 감탄사 | 0.91 | 0.65 | 0.04 | 0.72 | 0.02 | 0.21 | 1.94 | 1.65 | 0.65 | 0.58 |
| 수사 | 0.28 | 0.25 | 0.17 | 0.47 | 0.12 | 0.12 | 0.39 | 1.24 | 0.25 | 0.36 |
| 대명사 | 0.65 | 0.70 | 0.33 | 0.69 | 0.29 | 0.42 | 1.17 | 1.01 | 0.66 | 0.35 |
| 분석불능 | 0.28 | 0.12 | 0.04 | 0.21 | 0.06 | 0.10 | 0.56 | 0.76 | 0.19 | 0.19 |
| 접속부사 | 0.39 | 0.42 | 0.34 | 0.36 | 0.23 | 0.29 | 0.38 | 0.66 | 0.41 | 0.18 |
| 보조용언 | 0.48 | 0.48 | 0.32 | 0.50 | 0.29 | 0.34 | 0.67 | 0.64 | 0.40 | 0.18 |
| 부정지정사 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.05 | 0.02 | 0.03 | 0.01 |

〈표 12〉 보조 자료 단어 종수의 품사별 분포 비율

〈표 12〉와 〈표 11〉을 비교하면 일반명사의 비율 증가가 두드러짐을 볼 수 있다. '구어'를 제외하고 나머지 분야에서 일반명사의 비율이 모두 증가하였으며 그만큼 다른 품사가 차지하는 비중은 줄어들었다. 제일 많이 늘어난 것은 '문학'으로 일반명사의 비중이 7.6% 늘었다.

단어 종수에서는 분야별로 품사 분포의 차이가 어떻게 나는지를 파악하기 위하여 단어 총수에서와 마찬가지로 〈표 11〉을 기준으로 각 분야에서 높게 나타나는 품사 순서대로 배열해 보자. 정리 방식은 〈표 10〉과 동일하다.

| 전체 | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 |
|---|---|---|---|---|---|---|---|---|---|
| 일반명사(①) | ① | ① | ① | ① | ① | ① | ① | ① | ① |
| 동사(②) | ② | ② | ② | ② | ② | ② | ② | ② | ② |
| 형용사(③) | ③ | ③ | ③ | ③ | ③ | ③ | ③ | ④ | ③ |
| 일반부사(④) | ④ | ④ | ④ | ④ | ④ | ④ | ④ | ③ | ④ |
| 관형사(⑤) | ⑤ | ⑥ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑤ | ⑥ |
| 의존명사(⑥) | ⑥ | ⑤ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑥ | ⑤ |
| 감탄사(⑦) | ⑦ | ⑦ | ⑦ | ⑦ | ⑫ | ⑦ | ⑦ | ⑦ | ⑦ |
| 수사(⑧) | ⑨ | ⑨ | ⑨ | ⑧ | ⑨ | ⑨ | ⑨ | ⑧ | ⑨ |
| 대명사(⑨) | ⑫ | ⑫ | ⑫ | ⑨ | ⑪ | ⑫ | ⑫ | ⑨ | ⑫ |
| 분석불능(⑩) | ⑧ | ⑪ | ⑧ | ⑫ | ⑧ | ⑧ | ⑧ | ⑩ | ⑪ |
| 접속부사(⑪) | ⑪ | ⑧ | ⑪ | ⑪ | ⑦ | ⑪ | ⑪ | ⑪ | ⑧ |
| 보조용언(⑫) | ⑩ | ⑩ | ⑩ | ⑩ | ⑩ | ⑩ | ⑩ | ⑫ | ⑩ |
| 부정지정사(⑬) | ⑬ | ⑬ | ⑬ | ⑬ | ⑬ | ⑬ | ⑬ | ⑬ | ⑬ |

〈표 13〉 단어 종수에서의 분야별 품사 순위

〈표 13〉과 〈표 10〉을 비교하면 단어 종수에서는 단어 총수에서만큼 순위 변동이 두드러지지 않는다. 7위인 감탄사까지의 순위가 크게 변동을 하지 않는다. 한 가지 주목할 만한 점은 '신문'에서 감탄사가 4칸이나 순위가 밀려 표에 나와 있는 것 중에서 제일 변동이 크다는 점이다. 그만큼 '신문'에서는 감탄사가 별로 사용되지 않는다고 해석할 수 있을 것이다.

마지막으로 평균 출현 비율을 보자. 단어 총수를 단어 종수로 나누었을 때의 비율이다.

| | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 | 전체 |
|---|---|---|---|---|---|---|---|---|---|---|
| 부정지정사 | 307.00 | 180.00 | 1860.00 | 1401.00 | 788.00 | 825.00 | 67.00 | 378.00 | 159.00 | 2086.00 |
| 보조용언 | 93.82 | 162.09 | 341.63 | 312.54 | 294.43 | 226.90 | 43.41 | 56.04 | 87.48 | 1059.97 |
| 대명사 | 58.15 | 56.96 | 174.18 | 192.51 | 111.71 | 88.15 | 30.87 | 67.38 | 61.90 | 399.31 |
| 접속부사 | 47.36 | 45.96 | 182.24 | 70.52 | 96.43 | 63.97 | 14.96 | 44.68 | 31.65 | 325.90 |
| 의존명사 | 32.33 | 36.38 | 106.42 | 70.73 | 102.84 | 72.12 | 20.90 | 37.04 | 29.14 | 264.21 |
| 형용사 | 9.74 | 12.21 | 17.59 | 11.11 | 12.88 | 10.48 | 4.86 | 12.37 | 6.76 | 38.68 |
| 일반부사 | 9.38 | 9.06 | 17.17 | 12.25 | 14.11 | 10.61 | 5.58 | 14.39 | 7.40 | 36.09 |
| 관형사 | 15.80 | 21.45 | 16.69 | 25.97 | 14.33 | 19.09 | 13.51 | 19.19 | 21.98 | 35.83 |
| 동사 | 7.97 | 10.49 | 12.17 | 12.36 | 10.29 | 9.45 | 7.54 | 8.70 | 7.43 | 29.48 |
| 분석불능 | 6.60 | 2.56 | 6.48 | 11.33 | 3.50 | 5.24 | 6.74 | 16.14 | 5.00 | 28.42 |
| 감탄사 | 13.39 | 4.02 | 3.89 | 9.21 | 1.77 | 3.71 | 9.92 | 29.74 | 5.26 | 28.19 |
| 수사 | 5.71 | 3.64 | 18.35 | 8.89 | 9.61 | 8.38 | 4.26 | 3.81 | 6.30 | 17.75 |
| 일반명사 | 5.08 | 7.48 | 9.70 | 7.33 | 9.05 | 6.53 | 4.59 | 5.13 | 4.78 | 16.67 |
| 전체 평균 | 7.77 | 9.90 | 12.98 | 11.57 | 11.37 | 9.10 | 6.61 | 9.97 | 7.00 | 25.40 |

〈표 14〉 품사별 평균 출현 비율

부정지정사가 평균 출현 비율이 가장 높게 나타나고 다음으로 보조용언이 높게 나타난다. 부정지정사는 모든 분야에서 1위이지만 보조용언은 '구어'에서만은 대명사에 밀렸다. 〈표 11〉에 제시된 단어 종수에서는 부정지정사가 맨 마지막, 보조용언이 그 바로 위 순위에 있었다. 반대로 최상위를 차지하던 일반명사는 '전체'를 기준으로 했을 때 평균 출현 비율이 제일 낮아 여기서는 정반대로 맨 마지막에 있다. 그렇지만 분야별로 보면 일반명사가 마지막 순위에 있는 것은 '교재, 문학, 기타'만이다.

## VI. 상위 빈도에서의 분포

빈도 조사에서 상위 순위를 차지하는 고빈도어는 주목할 만하다. 상위 빈도에서 나타나는 양상에 대해 좀더 세밀히 검토해 보자.

먼저 분야별로 최상위 빈도에 속하는 20위까지의 단어만을 추출하여 높은 출현 횟수를 보이는 단어들이 분야별로는 어떤 양상으로 나타나는지를

보자. 지면상의 문제 때문에 20위까지만 검토하기로 한다.[9)]

| 순위 | 전체 | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 것01/의 | 하다01/동 | 하다01/동 | 것01/의 | 나03/대 | 것01/의 | 것01/의 | 하다01/동 | 거01/의 | 하다01/동 |
| 2 | 하다01/동 | 있다01/형 | 있다01/형 | 있다01/형 | 것01/의 | 있다01/보 | 있다01/보 | 거01/의 | 하다01/동 | 것01/의 |
| 3 | 있다01/보 | 것01/의 | 것01/의 | 하다01/동 | 있다01/보 | 동05/의 | 하다01/동 | 나03/대 | 있다01/형 | 나03/대 |
| 4 | 있다01/형 | 되다01/동 | 되다01/동 | 있다01/보 | 하다01/동 | 있다01/형 | 있다01/형 | 보다01/동 | 그런01/관 | 있다01/형 |
| 5 | 되다01/동 | 사람/명 | 사람/명 | 수02/의 | 그01/대 | 하다01/동 | 하다01/보 | 있다01/보 | 되다01/동 | 그01/관 |
| 6 | 수02/의 | 가다01/동 | 가다01/동 | 되다01/동 | 그01/관 | 하다01/보 | 되다01/동 | 가다01/동 | 나03/대 | 되다01/동 |
| 7 | 하다01/보 | 하다01/보 | 하다01/보 | 하다01/보 | 않다/보 | 수02/의 | 수02/의 | 되다01/동 | 같다/형 | 우리03/대 |
| 8 | 나03/대 | 있다01/보 | 있다01/보 | 그01/관 | 있다01/형 | 일07/의 | 않다/보 | 오다01/동 | 사람/명 | 하다01/보 |
| 9 | 그01/관 | 없다01/형 | 없다01/형 | 없다01/형 | 없다01/형 | 되다01/동 | 나03/대 | 너01/대 | 그거/대 | 없다01/형 |
| 10 | 없다01/형 | 나03/대 | 나03/대 | 이05/관 | 되다01/동 | 않다/보 | 없다01/형 | 있다01/형 | 안02/부 | 가다01/동 |
| 11 | 않다/보 | 수02/의 | 수02/의 | 않다/보 | 거01/의 | 이05/대 | 그01/관 | 안02/부 | 그01/관 | 있다01/보 |
| 12 | 사람/명 | 않다/보 | 않다/보 | 대하다02/동 | 하다01/보 | 없다01/형 | 년02/의 | 없다01/형 | 좋02/부 | 사람/명 |
| 13 | 우리03/대 | 그01/관 | 그01/관 | 아니다/지 | 수02/의 | 대하다02/동 | 그01/대 | 주다01/보 | 우리03/대 | 어머니01/명 |
| 14 | 이05/관 | 오다01/동 | 오다01/동 | 사회07/명 | 사람/명 | 년02/의 | 사람/명 | 보다01/보 | 아02/감 | 집01/명 |
| 15 | 그01/대 | 저03/대 | 저03/대 | 우리03/대 | 말01/명 | 이05/관 | 때01/명 | 뭐/대 | 것01/의 | 때01/명 |
| 16 | 아니다/지 | 거01/의 | 거01/의 | 보다01/동 | 그녀/대 | 씨07/의 | 주다01/보 | 그01/관 | 아니다/지 | 수02/의 |
| 17 | 보다01/동 | 보다01/동 | 보다01/동 | 이05/대 | 아니다/지 | 위하다01/동 | 동05/의 | 듯01/의 | 때01/명 | 오다01/동 |
| 18 | 동05/의 | 일01/명 | 일01/명 | 사람/명 | 가다01/동 | 명03/의 | 아니다/지 | 사람/명 | 뭐/감 | 않다/보 |
| 19 | 때01/명 | 말01/명 | 말01/명 | 그01/대 | 같다/형 | 우리03/대 | 이05/관 | 집01/명 | 얘기/명 | 보다01/동 |
| 20 | 거01/의 | 때01/명 | 때01/명 | 동05/의 | 말하다/동 | 원01/의 | 한01/관 | 나오다/동 | 보다01/동 | 날01/명 |

〈표 15〉 분야별 상위 빈도 단어(20위까지)

'전체'에서 빈도가 높은 단어가 각 분야에서도 대체로 상위에 속하는 단어들로 나타난다. 그만큼 고루 분포하는 단어들이라는 뜻이다. '전체'에서 1위에서 5위에 속하는 단어들로 국한하면 거의 모든 분야에서 상위 20 안에 나타난다. 좀더 구체적으로 보면 '하다01/동, 있다01/형, 되다01/동'은 모든 분야에서 상위 20위 안에 들었다. 특히 '하다01/동'은 모든 분야에서 상위 5위 안에 들었다. 전체 빈도로는 1위인 '것01/의'는 '대본'에서 57위가 되어 상위 20위 안에 들지 못했다. '있다01/보'는 '구어'에서 37위가 되어 역시 한 분야에서 20위 안에 들지 못했다.

분야별로 빈도순 1위를 차지한 단어는 대부분 '하다01/동'이거나 '것01/의'인데 '문학'과 '구어'에서만은 다르다. '문학'에서는 '나03/대'가 가장 많이 사

9) 칸이 비좁아 품사는 약호로 제시하였다. '01, 02' 등 숫자가 붙은 것은 『표준국어대사전』의 어깨번호를 표시한 것이다.

용되었으며 '구어'에서는 '거01/의'가 그 자리를 차지하였다. '전체'에서의 순위로는 '나03/대'는 8위, '거01/의'는 20위이다. 구어투의 말투인 '거01/의'는 '대본'에서도 2위로 나타난다. '거01/의'는 이 두 분야를 제외하면 '교재'에서 16위, '문학'에서 11위를 차지한 외에 다른 분야에서는 20위 안에 들지 못했다. '나03/대'는 '교양'과 '신문'을 제외하고 다른 분야에서는 모두 20위 안에 나타난다. '교양'에서는 45위, 신문에서는 218위에 그쳤다. '구어'에서 한 가지 더 주목할 점은 '어02/감'가 14위, '뭐/감'이 18위를 차지하여 감탄사 2개가 상위 빈도에 포함되었다는 점이다. 감탄사가 20위 안에 든 것은 여기서가 유일하다.

이제 상위 빈도에서 기원별 분포가 어떻게 나타나는지를 보자. 상위 빈도 100위까지만 대상으로 한다.

| | 전체 | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 |
|---|---|---|---|---|---|---|---|---|---|---|
| 고유어 | 81 | 87 | 77 | 64 | 88 | 53 | 78 | 93 | 88 | 87 |
| 한자어 | 15 | 12 | 18 | 33 | 7 | 41 | 19 | 7 | 10 | 10 |
| 고+한 | 4 | 1 | 5 | 3 | 5 | 5 | 3 | 0 | 2 | 1 |
| 외래어 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 |

〈표 16〉 상위 빈도 단어에서의 기원별 분포

앞에서 검토했던 기원별 분포에서 7가지의 유형이 나타났었는데 〈표 16〉을 보면 상위 빈도에서는 이 중에서 고+외, 한+외, 고+한+외 등 외래어가 포함된 유형 3가지가 전혀 나타나지 않음을 볼 수 있다. 순수 외래어도 모든 분야를 통해 단지 '신문'에서 하나, '기타'에서 둘 등 세 단어만 상위 빈도에 나타났다. '신문'에서 100위 안에 든 외래어는 23위로 나오는 '퍼센트'이다. '기타'에서는 33위에 '아나운서', 70위에 '뉴스'가 나타났다. '기타'에서 '아나운서'와 '뉴스'가 높게 나타나는 것은 조사 대상 자료 중에 방송 관련 자료가 포함되었기 때문이다. '기타'에서 '뉴스'가 132회 나오는 중에 131회가 한 자료에서 나왔다. '아나운서'는 237회 모두 한 자료에서 나왔다.

〈표 16〉에서 두드러지게 나타나는 양상은 모든 분야를 통틀어 상위 빈도

에서 고유어가 한자어보다 많다는 점이다. '대본'에서는 93:7로 그 격차가 가장 심하다. '신문'과 '교양'에서만 상대적으로 고유어와 한자어의 격차가 줄었을 뿐이다. 이 두 분야는 앞서 전체 단어를 대상으로 기원별 분포를 살필 때 한자어의 비율이 고유어보다 2배 이상 높던 분야이다.

마지막으로 상위 빈도에서 품사별로는 어떻게 나타나는지를 보도록 하자. 역시 상위 빈도 100위까지만 대상으로 한다.

| | 전체 | 교재 | 교과 | 교양 | 문학 | 신문 | 잡지 | 대본 | 구어 | 기타 |
|---|---|---|---|---|---|---|---|---|---|---|
| 동사 | 24 | 23 | 26 | 16 | 26 | 16 | 22 | 25 | 17 | 24 |
| 일반명사 | 19 | 22 | 36 | 35 | 26 | 39 | 25 | 21 | 15 | 28 |
| 의존명사 | 13 | 7 | 6 | 7 | 6 | 16 | 15 | 6 | 5 | 6 |
| 관형사 | 8 | 8 | 7 | 8 | 6 | 5 | 7 | 7 | 6 | 7 |
| 대명사 | 8 | 9 | 4 | 7 | 12 | 3 | 6 | 8 | 9 | 7 |
| 보조용언 | 8 | 7 | 5 | 7 | 8 | 7 | 9 | 9 | 6 | 9 |
| 일반부사 | 8 | 7 | 4 | 4 | 5 | 5 | 5 | 9 | 14 | 5 |
| 형용사 | 8 | 9 | 8 | 10 | 6 | 5 | 7 | 6 | 8 | 8 |
| 접속부사 | 3 | 6 | 3 | 5 | 3 | 3 | 3 | 2 | 8 | 5 |
| 부정지정사 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 감탄사 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 6 | 9 | 0 |
| 분석불능 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| 수사 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

〈표 17〉 상위 빈도 단어에서의 품사별 분포

기원별 분포와 달리 품사별 분포에서는 본고에서 구분하는 품사들이 최소 하나씩은 100위 안에 들었다. 그렇지만 감탄사, 분석불능, 수사의 분포는 일부 분야에서만 나타난다. 감탄사는 '교재'와 '대본', '구어'에서만 나타난다. '교재'에서 나타나는 감탄사는 32위를 차지한 '네03'이다. 대답할 때 사용하는 말인 '네03'이 높게 나타난 것은 '교재'의 성격을 감안하면 수긍이 가는 분포이다. '대본'과 '구어'에서 나타나는 것은 다음과 같다. ( ) 안의 숫자는 순위이다.

대본 : 그래01(48), 아04(68), 아02(73), 네03(74), 예06(85), 응01(89)

구어 : 어02(14), 뭐(18), 아02(33), 그02(46), 예06(70), 음01(72), 아니02(82), 야04(91), 그래01(94)

'구어'에서 분석불능으로 100위 안에 든 두 항목은 62위의 '걔', 79위의 '그랬'이다. 수사로 유일하게 '문학'에서만 나타나는 단어는 '하나'이다.

〈표 17〉에서 모든 품사가 나타나기는 했지만 순위는 앞에서 검토했던 품사별 분포와 차이가 있다. 동사가 일반명사를 제치고 1위로 올라섰다는 점을 가장 두드러진 특징으로 꼽아야 할 것이다. '전체'를 보면 동사가 1위로 올라섰으며 '교재, 대본, 구어'에서도 동사가 일반명사보다 많다. '문학'에서는 명사와 동사의 개수가 똑같다. 전체 단어를 대상으로 했을 때 단어 총수에서도 단어 종수에서도 부동의 1위를 차지하던 품사가 일반명사였다. 그런데 상위 빈도에서는 그렇지 못한 것이다.[10]

## VII. 맺는 글

지금까지 졸저(2002)에서 분야별로 빈도 조사를 한 자료를 이용하여 어휘가 분야별로 어떻게 분포하는지를 한국어 교재, 교과서, 교양, 문학, 신문, 잡지, 방송 대본, 구어, 기타의 9개 분야로 나누어 살폈다. 논의 중에 나왔던 주요 사항을 정리하면 다음과 같다.

1. 단어 종수의 증가에 따른 단어 총수의 누적 사용 빈도는 모든 분야에서 비슷하게 증가하였다.

2. 기원별 분포에서는 단어 종수를 기준으로 했을 때는 대체로 한자어의 비율이 높게 나타났고 단어 총수를 기준으로 했을 때는 대체로 고유어의 비율이 높게 나타났다. 단어 종수에서는 '대본, 문학, 기타'에서만 고유어의 비율이 높았고 나머지는 한자어의 비율이 더 높았다. 반대로 단어 총수에서는 '신문'을 제외한 다른 분야에서는 모두 고유어의 비율이 한자어의 비율보다 높게 나왔다.

---

10) 보조 자료를 이용하여 동일한 조건으로 상위 빈도에서의 분포를 검토한 결과는 지금까지의 논의와 경향에서 큰 차이가 없으므로 자세한 설명은 생략한다.

3. 품사별 분포에서는 단어 총수를 기준으로 했을 때 총합이 148만여 단어가 되는 주 자료와 36만 단어가 되는 보조 자료에서의 분야별 품사 분포 비율이 큰 차이가 없는 것으로 나타나 자료 양과 상관없이 분야별 품사 분포 비율은 일정한 편이었다. 단어 종수에서는 보조 자료에 비해 주 자료에서 일반명사의 비율 증가가 두드러졌다. 품사별 분포 순위를 보면 일반명사와 동사는 단어 총수와 단어 종수 둘 다에서 모든 분야에서 1위와 2위를 차지하였다.

4. 분야별로 빈도순 1위를 차지한 단어는 대부분 '하다01/동사'이거나 '것01/의존명사'인데 '문학'에서는 '나01/대명사'가, '구어'에서는 '거01/의존명사'가 1위를 차지하였다.

5. 빈도순 20위까지 살폈을 때 '구어'에서만 유일하게 감탄사 2개가 20위 안에 들었다.

6. 빈도순 100위까지만을 대상으로 했을 때 고유어가 한자어보다 압도적으로 많이 나타났다. 외래어의 경우는 모든 분야를 통틀어 단지 3개의 단어만 100위 안에 들었다.

7. 빈도순 100위까지만을 대상으로 했을 때 전체 단어를 대상으로 한 조사에서는 부동의 1위를 차지하던 일반명사가 '교재, 대본, 구어'에서 동사에 밀려 2위로 내려앉았다.

본고는 148만여 단어 분량의 자료와 36만 단어 분량의 자료를 이용하여 어휘의 분야별 분포에 대해 검토하였다. 그렇지만 이 자료의 양이 국어 어휘의 분포 양상을 파악하는 데 충분한 정도의 양이었는가에 대해서는 의문이 제기될 수 있다. 더 많은 자료를 확보하여 논의를 하는 것이 바람직하기는 하지만 본고와 같은 정도의 논의를 하기 위한 자료를 만드는 일 자체가 많은 시간을 요하는 일이기 때문에 자료의 양을 더 늘리지는 못했다. 자료의 양이 미흡하여 결과가 일부 실제 언어 현실과 다르게 나왔을 소지도 전혀 없지는 않다. 앞서 상위 빈도에 대한 기원별 검토에서 '뉴스, 아나운서'가 100위 안에 든 것이 그 사례가 되지 않을까 의심하고 있다.

어휘 계량을 통한 연구가 많이 부족하기 때문에 이 정도의 자료만으로

분포 양상을 검토한 것도 깊이있는 논의가 나오는 데 좋은 밑거름이 된다고 생각하지만 본고에서 확인된 사실은 앞으로 더 많은 양의 자료에서 검증할 필요가 있다. 국립국어연구원에서 새로운 조사에 착수하였기 때문에 몇 년 후 조사가 완료되면 더 많은 자료에서 본고의 논의를 검증할 기회가 있을 것으로 생각한다.

# 참고문헌

강범모(1999), 『한국어의 텍스트 장르와 언어 특성』, 고려대 출판부.

_____ 외(2000), 『한국어의 텍스트 장르, 문체, 유형-컴퓨터와 통계적 기법의 이용』, 태학사.

김광해(1993), 『국어 어휘론 개설』, 집문당.

김홍규 · 강범모(2000), 『한국어 형태소 및 어휘 사용 빈도의 분석 1』, 고려대 민족문화연구원.

남윤진(1999), 「균형 말뭉치 구축을 위한 실험적 연구(1) - 표본 크기 및 텍스트 범주의 문제를 중심으로」, 서상규 편 『언어 정보의 탐구』(도서출판 월인).

_____(2002), 「국어 연구와 빈도 정보」, 홍윤표 외 『한국어와 정보화』(태학사).

문영호(2001), 『어휘 통계학』, 박이정.

문화관광부(1999), 『21세기 세종계획 국어 기초자료 구축 분과 보고서』.

서상규(1998), 「연세 말뭉치 1-9를 대상으로 한 현대 한국어의 어휘 빈도- 빈도 7 이상」, 연세대 언어정보개발연구원 내부 보고서.

이상억(2001), 『계량국어학 연구』, 서울대 출판부.

이운영(2002), 『표준국어대사전』 연구 분석, 국립국어연구원 보고서.

이익환(2002), 『기본 어휘 선정 및 사용 실태 조사를 위한 기초 연구』, 국립국어연구원.

임칠성 외(1997), 『한국어 계량 연구』, 전남대 출판부.

조남호(2002), 『현대 국어 사용 빈도 조사 - 한국어 학습용 어휘 선정을 위한 기초 조사』, 국립국어연구원 보고서.