울산어문논집 제15집 (2001.11.30.) pp.65~94.
울산대학교 인문대학 국어국문학부

한국어 학습자 사전 개발을 위한 어휘 계량적 접근

한 영 균*

1. 서론

1.1. 이 글은 한국어 학습자 사전(Learners' Dictionary of Modern Korean)의 개발에 필요한 기초 자료의 수집·정리와 관련하여, 말뭉치의 계량적 분석을 통해 얻어낼 수 있는 어휘 관련 정보에는 어떤 것들이 있는가를 살펴보고, 그와 관련된 문제들을 정리하는 것을 목적으로 한다.

1.2. 학습자 사전은 제2언어 혹은 외국어를 학습하는 이를 위한 일종의 선별형 사전이다. 따라서 학습자 사전은 학습자의 학습 단계에 맞추어[1] 학습 대상 언어의 어휘 단위 중에서 그 언어를 이해하고 사용하는 데에 필요하다고 판단되는 어휘 단위들을 등재하게 된다. 결국 어떤 것들을 사전의 표제항[2]으로 등재할 것인가를 결정하는 일이 학습자 사전 개발의

* 울산대학교 국어국문학부 교수

1) 학습자 사전이 학습 단계에 맞추어야 한다는 것과 학습 단계에 맞춘 학습자 사전이란 어떤 것이며 외국어 학습의 어느 단계까지 학습자 사전이 용인될 것인가 하는 문제는 별개의 것이다. 이에 대해서 여기서는 길게 논의하지 않는다.

2) 이 글에서는 거시구조의 구성 요소를 지칭할 때에는 '표제어' '부표제어'라는 용어 대신 '표제항, 부표제항'이라는 용어를 쓰고, 어휘부의 구성 요소 중에서 사전의 표

첫 단계가 되는 것이다.

사전의 표제항으로 등재할 어휘 단위를 고르는 데에는 빈도가 가장 신뢰할 수 있는 기준이 된다. 빈도만을 기준으로 거시구조를 이루는 어휘 단위를 고르는 것이 합당한가 하는 데에는 논의의 여지가 있지만, 아직까지 그보다 나은 객관적 기준이 없는 것이다.[3] 그러나 표제항으로 등재할 어휘 단위의 선별이라는 목적을 염두에 두고 말뭉치에서 빈도를 산출하고 그 결과를 이용해서 기초 자료를 개발한다고 하더라도 무엇을 대상으로 할 것인가에 따라서 방법이 달라질 수 밖에 없다. 이 글은 이런 관점에서 학습자 사전의 거시구조를 이루는 어휘 단위를 선정하는 데에 말뭉치 분석의 결과를 활용할 수 있는 방법을 정리해서 보다 바람직한 학습자 사전을 개발할 수 있게 하려는 데에 목적이 있다.

1.3. 2장에서는 말뭉치 분석을 통해서 얻을 수 있는 어휘 관련 빈도를 형태 빈도, 갈래뜻 빈도, 개념 빈도, 단어족 빈도의 네 가지로 나누어 정리한다. 이 글에서는 지면 관계상 형태 빈도 중에서 어절 빈도의 추출과 관련된 문제를 주로 다루게 될 것이고, 나머지는 간단히 개념 중심으로 정리하는 데에 그친다. 3장에서는 2장에서 정리한 각종 어휘 관련 빈도 정보를 학습자 사전의 거시구조 및 미시구조 구성에 어떻게 적용할 수 있는가를 살펴본다. 이 과정에서는 특히 제2언어 학습의 관점에서 말뭉치에 반영되어 있는 한국어 어휘부의 조직을 분석하고 어휘 단위 사이의

제항이 되는 것을 가리킬 때에는 '단어, 어휘소'라는 용어 대신 '어휘 단위'라는 용어를 쓰기로 한다. 사전의 표제항이 되는 것이 기본적으로 단어이기는 하지만 단어만 사전의 표제항이 되는 것은 아니고, 또 둘 이상의 어휘소가 결합한 형태가 사전의 표제항이 될 수도 있고 단일 어휘소가 아닌 것도 사전의 표제항이 될 수 있어서 어휘소(lexeme)라는 용어를 쓰는 것도 적절치 않다고 판단되기 때문이다.

3) 그러나 이른바 '기초 어휘·기본어휘'류를 선별하는 경우에는 조금 다르다. 기본어휘에 들어가는 어휘 수는 국어 어휘의 총량에 비해 아주 적기 때문에 빈도 이외에도 텍스트 안에서의 분포, 단어들 사이의 관계 등이 고려되어야 하고, 게다가 제2언어로서 한국어를 학습하는 이들을 위한 교육용 기본어휘의 선별에서는 학습자의 모어를 고려해야 하는 등 여러 측면에서의 접근 방법론이 필요한 것이다. 기본어휘·기초어휘의 개념과 기본어휘 선정방법론에 대해서는 서상규·남윤진·진기호(1998 : 7-17)에 비교적 자세히 정리되어 있다.

관계를 밝혀 그것을 학습자 사전의 개발에 응용하기 위해서는 어떤 방법과 기준에 의한 말뭉치 분석이 이루어져야 하는가를 단어형성론 및 어휘의미론의 관점에서 정리할 것이다.[4)]

2. 어휘 관련 빈도 정보의 유형

2.1. 형태 빈도

형태 빈도란 말 그대로 말뭉치에 출현하는 언어 단위의 형태별 빈도를 가리키는데, 언어 단위의 연쇄를 대상으로 빈도를 구하는가 그렇지 않은가에 따라 단순 빈도와 연접 빈도의 두 유형으로 나눌 수 있다.[5)] 단순 빈도란 하나의 언어 단위가 출현하는 빈도를 말하며, 연접빈도는 둘 이상의 언어 단위가 연쇄를 이루는 빈도를 가리킨다. 단순 빈도로는 어절 빈도, 단어 빈도, 문법형태소 빈도가 대표적인 것이고,[6)] 연접 빈도는 단순

---

4) 필자는 학습자 사전은 단순한 참조용 문헌(reference book)이 아니며, 제2언어 학습 교재(textbook)으로서의 기능도 겸비해야 한다고 생각한다. 특히 기본적인 것들을 등재하게 되는 초·중급자를 위한 한국어 학습자 사전에서는 이런 기능이 충실히 반영되어야 한다고 믿는다. 이 글에서 학습자 사전의 개발을 다루면서 제2언어 습득에 있어서의 어휘 학습의 문제를 언급하는 것은 이 때문이다.

5) 단어의 연접 빈도를 구하는 일과 연쇄를 이루는 어휘 단위에 대한 빈도를 구하는 일은 별개의 작업이다. 연접 빈도를 구하는 일이 개별 언어 단위를 대상으로 각 단위가 어떤 언어 단위와 어느 정도 자주 연쇄를 이루는가를 밝히는 것이 목적이라면 다어기어휘소(Multiword Lexical Unit)라고 지칭되는 연쇄를 이루는 어휘 단위에 대한 빈도를 구하는 일은 연쇄를 이루는 어휘 단위 자체를 대상으로 각각의 어휘가 어느 정도 자주 쓰였는가를 확인하는 작업이기 때문이다. 물론 전자는 후자의 작업을 위한 기초자료로 사용될 수 있다. 다어기어휘소는 하나의 어휘소이면서 여러 개의 어절로 구성되기 때문에 그것의 빈도를 구하기 위해서는 국어의 다어기어휘소에는 어떤 것들이 있는가를 알아 목록을 확정하는 일이 선행되어야 하기 때문이다. 연어의 빈도를 구하는 문제도 이와 비슷하다. 여기서는 다어기 어휘소나 연어와 같은 단어 연쇄의 빈도를 구하는 문제에 대해서는 구체적으로 다루지 않을 것이다.

빈도 추출 대상이 되는 대부분의 언어 단위에 대해서 추출할 수 있는데, 여기서는 단순 빈도 중 어절 빈도를 중심으로 논의를 진행하기로 한다. 단어나 개별 문법형태소의 빈도 정보가 지니고 있는 중요성에 대해서는 이미 잘 알려져 있는 반면, (학습자) 사전의 개발에 어절 빈도를 어찌 활용할 것인가 하는 문제에 대해서는 별로 논의가 이루어진 바 없다고 생각되기 때문이다.

어절 빈도는 간단히 말해서 띄어쓰기에 의해 구분된 언어 단위의 빈도를 말하는데,[7] 지금의 맞춤법에서 단어와 단어는 띄어쓰고 조사와 어미는 붙여 쓰도록 규정하고 있으면서 또한 형태음소적 교체를 표기에 그대로 반영하는 경우가 많기 때문에,[8] 어절 빈도에는 이른바 형태적 중의성과 품사 중의성을 띄는 것들이 구별되지 않은 채로 하나의 단위로 처리된다. 이러한 까닭에 어절 빈도 정보는 그 쓰임이 극히 한정적이었다.[9] 어절 빈도의 활용을 위해서는 중의적 어절을 대상으로 한 2차 처리가 필요하다는 것이 문제가 되었기 때문이다. 그러나 이제는 날말뭉치(raw corpus)와 문법 정보 주석 말뭉치가 함께 공개된 자료가 많기 때문에 그러한 문제는 어느 정도 극복되었다고 할 수 있다. 오히려 두 가지 말뭉치

---

6) 이 이외에 품사별 빈도와 접사 및 첨사류(clitics)의 빈도와 분포에 대한 계량적 정보도 필요하지만 앞에서 언급한 세 가지가 기본적인 것으로 생각된다. 자소 빈도, 음절 빈도 등 2차 분절을 이루는 요소의 빈도는 학습자 사전 개발과는 크게 관련이 없다.

7) 그런데 사람은 띄어쓰기만이 아니라 한글 자모 이외의 문장 부호를 어절 구분의 기준으로 인식한다. 어절 빈도의 추출도 생각처럼 간단치는 않은 것이다. 이에 대해서는 부록 참조.

8) 국어 맞춤법의 '어법에 맞게 쓴다'는 규정을 이형태를 표기에 반영하지 않는다는 의미로 해석하는 경우도 있으나, 이 규정은 형태소 경계에서의 형태음소적 교체를 어찌 다룰 것인가에 대한 일반 원칙일 뿐이다. 실제 불규칙 활용 어간과 어미의 이형태는 표기에 그대로 반영되기 때문에 형태음소적 교체의 많은 부분이 표기에 반영된다.

9) 어절 빈도를 적극적으로 이용한 예로 어절 사전을 이용한 형태소 분석기와 반자동 문법 정보 주석 시스템의 개발을 들 수 있다. 주로 자연언어처리시스템을 개발하는 경우에 응용한 것이다. 자연언어처리시스템의 경우 한국어에 대해 전혀 직관이 없다는 점에서 제2언어로 한국어를 학습하는 이와 마찬가지라는 점 시사적이다.

를 함께 활용함으로써 학습자 사전의 개발에서의 말뭉치의 효용성을 높일 수 있는 것으로 생각된다.

우선 한국어 텍스트 안에서 어느 정도의 어절이 얼마나 중의적인 양상을 보이는가를 말뭉치를 분석한 자료를 통해 보이기로 한다.

<표 1 > 고빈도 어절의 텍스트 점유율과 중의도

빈도순 어절수	중의적 어절수	누적 중의적 어절수	빈도순 어절수	중의적 어절수	누적 중의적 어절수
0-100	38	38	-600	27	218
-200	45	83	-700	30	248
-300	48	131	-800	30	278
-400	31	162	-900	27	305
-500	29	191	-1001	24	329

<표 1>은 150만 어절 말뭉치에서 1,001 개의 상위 고빈도 어절[10]이 얼마나 중의적인 쓰임새를 보이고 있는가를 정리한 것인데,[11] 실제 1,001개 어절 중에서 중의적 용법을 보이는 어절은 50%가 되지 않는 것을 보여준다.[12] 빈도가 높은 어절일수록 중의적인 용법을 보일 가능성이 크다는

10) 150만 어절 말뭉치에서 151회 이상 사용된 것들이다.

11) 대규모 말뭉치를 대상으로 한 어절 빈도의 추출은 어절빈도 추출프로그램에 의존할 수 밖에 없는데, 프로그램에 따라 기능에 차이가 있어서 동일한 말뭉치를 대상으로 하는 경우에도 다른 결과를 얻게 되는 경우가 있다. 여기서 제시한 것은 21세기 세종계획의 결과물로 보급된 <글잡이Ⅱ(직접)>을 사용한 결과인데, 전주대 소강춘 교수팀에서 개발·공개한 <깜짝새>의 분석 결과와는 상당한 차이가 있다. 어절 빈도 분석 프로그램을 어떤 것을 쓰느냐에 따라서 분석 결과가 달라진다는 뜻인데, 이 글의 논의와는 직접 관계가 없지만 이 문제에 대해 확실히 인식할 필요가 있다고 생각되어 부록에서 별도로 다루었다. 어절 빈도 추출과 관련된 몇 가지 문제에 대해서는 부록을 참조하기 바란다.

12) 중의성 판별은 『연세한국어사전』에서의 동음이의어 처리에 따랐다.

점을 고려하면 전체 텍스트를 구성하는 어절 중에서 중의적 용법을 보이는 것의 비율이 그리 크지 않을 것임을 짐작할 수 있다.[13] 더구나 고빈도 어절의 텍스트 점유율을 고려하면,[14] 각 어절의 빈도 순위와 중의성이라는 정보는 학습자 사전의 편찬에서 활용 가능성이 높다.

굳이 유타니(1999, 2001)의 지적을 빌지 않더라도, 한국어 학습자들이 중의적 어절을 분석하고 어간을 찾아내는 데에 상당히 어려움을 겪는다는 것은 잘 알려진 사실이다. 기본적으로 한국어 어절이 어떻게 분절되는가를 알기 어렵기 때문이다. 특히 한 어절이 여러 가지로 해석될 수 있는 경우에 그것을 올바로 분석하는 일은 대단히 어려운 일이다. 어휘 학습에서도 용언의 어간과 활용형 사이의 관계를 익히는 일은 매우 중요한데,[15] 초•중급 학습자를 위한 사전이 이러한 문제에 도움이 될 수 있으면 좋겠지만, 구체적으로 어느 정도의 어절이 형태적 중의성을 가지는 것이며 그 중에서 어떤 것들이 한국어 학습에 어려움을 일으키는 것인가에 대한 정보가 제대로 정리된 적이 없기 때문에 체계적으로 수용하기 어렵다는 문제가 있다.[16] 어절 빈도의

---

13) 전체 어절에 대한 중의성 검토는 불필요하다고 판단했기 때문에 하지 않았다. 실제 150만 어절 말뭉치의 어절 빈도가 30만 가지가 넘기 때문에 이를 하나 하나 검토하려면 한 사람이 하루에 5,000어절의 용례를 처리한다 해도 약 10개월이 걸리는 일이다.

14) 150만 어절 말뭉치에 쓰인 어절의 종수 및 텍스트 점유율에 대해서는 부록의 <표 1> 및 <표 2> 참조.

15) 일본의 일부 지역에서 한국어의 동사 어간을 3유형의 어기로 나누어 교육하는 것도, 한국어의 학습에는 용언의 어간의 분석 능력이 중요한데 일본인에게는 일본어 문법 체계 위에서 한국어 용언의 활용체계를 학습하는 것이 효율적이라는 관점에서 재구성한 것이라고 할 수 있다. 옳고 그름을 떠나서, 국어 굴절형의 분석이 외국인에게는 큰 부담임을 보여주는 단적인 예라고 할 것이다. 한국어와 매우 유사한 활용체계를 가지고 있는 일본인에게조차 국어 용언의 활용형을 분석하는 일이 상당히 어려운 일임을 말해주는 것이기 때문이다.

16) 사전 표제항으로의 체계적 수용이라는 문제는 한정된 지면에 어떻게 하면 꼭 필요한 정보를 담을 수 있느냐 하는 문제라고 할 수 있다. 이른바 경제성의 문제인데, 이와 관련해서 모든 활용형을 표제항으로 수용하기는 어렵지 않느냐는 것이 2001년 11월 24일 연세대에서 열린 제2회 한국어교육 국제워크숍-『한국어 교육과 학습사전』의 유타니 선생의 발표에 대한 토론에서 지적되기도 했다. 이는 말뭉치 안에서의 어절 빈도가 제대로 정리되면 해결될 문제로 생각된다.

활용은 이러한 문제를 해결하는 데에 도움을 준다.

또 어절 빈도 정보는 단순히 표제항 선별에만 활용할 수 있는 것이 아니다. 앞에서 언급한 바와 같이 주석 말뭉치와 날말뭉치를 함께 활용하면, 용례를 선택하는 경우에도 여러 활용형 중에서 개별 어휘 단위의 전형적 용법을 보이는 것을 확보할 수 있고, 어휘 항목별 활용형의 빈도가 확보되는 것이므로 한정된 활용형으로만 쓰이는 어휘 항목에 대한 정보의 수집도 함께 이루어지며, 연어 관련 정보도 함께 정리할 수 있는 것이다. 이 글의 논의는 거시구조에서의 말뭉치 활용 문제를 다루는 것이 주목적이므로 자세히 언급하지 않는다.

2.2. 갈래뜻 빈도

갈래뜻 빈도는 개별 어휘 항목이 말뭉치 안에서 실제 어떤 구체적 의미 로 쓰였는가를 계량적으로 분석한 결과라고 할 수 있다. 졸고(1998)에서 기본적인 개념과 방법을 설명한 바 있고, 서상규 • 강현화(2000)에서 실제 작업의 결과를 볼 수 있으므로, 예만 들어 두기로 한다.

<예시 1> '하늘'의 갈래뜻 빈도
- 서상규 • 강현화(2000)에서 인용

하늘_NN	544	-①	453	83.27%
		-②	30	5.51%
		-③	46	8.46%
		-x01	3	0.55%
		-x02-1	1	0.18%
		-x03	1	0.18%
		-x06-2	1	0.18%
		-x07	4	0.74%
		-x08	1	0.18%
		-x09	1	0.18%
		-x10	1	0.18%

<예시 1>은 약 100만 어절의 말뭉치 안에서 '하늘'이라는 단어가 어떤 용법으로 쓰였는가를 보여 준다. 이 때의 갈래뜻 분석의 기준이 된 것은 『연세 한국어 사전』의 뜻풀이인데, ①은 '하늘이 맑다'에서와 같은 물리적 대상으로서의 하늘, ②는 '하늘로 가다'에서와 같은 '현세(現世)'에 대립되는 개념으로서의 하늘, ③은 '하늘 무서운 줄 알아라'와 같

은 신적 존재를 가리키는 하늘로 쓰인 것을 나타내며, 나머지 -x의 예들은 여러 가지 관용구에서 쓰인 '하늘'의 빈도이다.[17] 자료를 통해서 우리는 100만 어절 말뭉치에서 '하늘'이라는 단어가 모두 544회 쓰였는데, 그 중에서 453개의 예가 물리적 대상으로서의 '하늘'을 가리키는 것이었고, ②, ③의 의미로 쓰인 것은 각각 30, 46개의 예가 있음을 알 수 있다. 이러한 분석 결과를 바탕으로 우리는 사전에서의 '하늘' 항목의 기술을 좀 더 구체화할 수 있게 된다.

2.3. 의미장과 단어군 : 의미장 빈도와 관련해서

한국어를 모어로 하는 사람은 '사과, 배, 감, 수박, 참외'라는 단어들이 같은 부류에 속한다고 생각한다. 또 이들 중에서 '수박, 참외'는 다시 '여름 과일'이라는 좀더 작은 부류를 이룬다는 것도 안다. '호두, 잣' 따위의 단어들은 앞의 것들과는 다른 부류를 이루지만, 나무에 달리는 '열매'로 '먹을 것'이라는 점에서는 같은 부류라고 인식한다. 이때의 같은 부류라는 개념은 존재론적 범주를 기반으로 한 단어군과는 조금 다르다. 단어들 사이의 어휘의미론적 관계에 의한 묶음인 것이다. '여름 과일' '과일' '열매'와 같은 것들은 앞서 나열한 단어들의 상위 개념어라고 할 수 있고 이들끼리도 계층적인 의미관계를 보인다. 이렇게 동일한 상위 개념어에 속하

17) 관용구에 쓰인 '하늘'의 빈도를 '하늘'의 갈래뜻 빈도에 넣어도 좋은가는 생각해 볼 필요가 있다. 갈래뜻 빈도란 개개 단어의 의미 용법 상의 빈도를 가리키는 것인데, 다른 요소와 결합하여 '하늘'의 어휘적 의미와 무관한 용법을 가지게 된 것을 '하늘'의 하위 의미와 함께 처리해도 좋은가 하는 문제가 남기 때문이다. '하늘이 두 쪽이 나도 약속은 지킨다'는 표현에서의 '하늘이 두 쪽 나다'라는 관용구를 '하늘'의 미시구조 속에 넣는 것은 사용자가 그것을 찾기 쉽게 해주려는 편의상의 처리일 뿐 실제로 하늘의 하위 의미 속에 그 관용구의 의미가 들어있다는 뜻은 아닌 것이다. 따라서 관용적 용법에 사용된 '하늘'의 용례는 일일이 열거하여 의미빈도사전의 미시구조에 넣기보다는 기타 용법으로 처리하여 묶어주는 것이 의미빈도사전 본래의 목적에 적합할 것으로 생각된다. 비슷한 예로 속담 속에 들어있는 용례를 들 수 있는데, '마당 터지는데 솔뿌리 걱정한다'는 속담에서의 마당이라는 단어의 용법을 '마당'의 의미 용법 속에 넣을 수는 없는 일이다.

는 단어들이 엮어내는 의미 관계를 보통 낱말밭(word field) 혹은 의미장(semantic field)이라고 한다.

단어군(word cluster : 單語群)이라는 개념은 의미장과 유사하지만 조금 다르다.

'창문, 문, 책상, 의자, 칠판, 교탁, 교단, 벽'이라는 여덟 개의 단어를 보면 무엇이 연상되는가?

대개는 자연스럽게 '교실'을 떠올리게 될 것이다. 그러나 교실을 한번도 본 적이 없는 사람이나 교탁이나 교단이 없는 교실만을 아는 사람이 이들 단어를 통해서 '교실'을 떠올리는 일은 쉽지 않다. 이들 단어의 의미적 속성을 통해서 추량해 낼 수 있는 것이 아니기 때문이다. 즉 '교실'은 이들 단어의 상위 개념어가 아닌 것이다. 여기서 중요한 것은 이들 단어가 어휘의미론적으로는 동일한 상위 개념어의 하위 구성요소가 아님에도 불구하고 '교실'이라는 인지 대상을 통해 하나로 묶인다는 점이다. 이렇게 하나의 인지 대상 속에 묶이는 -존재론적으로(ontologically) 같은 범주를 이루는- 일련의 단어들의 집합을 단어군이라고 부른다.

낱말밭을 이루는 단어들 사이의 어휘의미론적 관계는 범어적 속성을 지니며, 대역 사전에서의 어휘의 대응은 이러한 어휘의미론적 속성을 바탕으로 한 것이라고 할 수 있다. 그러나 실제 번역에 있어서는 어휘 의미의 일대일의 대응관계를 그대로 적용하는 경우 시•공의 차이에서 빚어지는 인식론적 의미를 살리지 못한다는 문제가 일어난다. 이른바 번역에 있어서의 의미 해석의 문제를 일으키는 바(김용옥 1990), 이는 바로 범어적 속성으로서의 어휘의미론적 의미장 안에서의 어휘 단위의 의미와 개별 언어에 있어서의 존재론적 단어군 안에서의 각 어휘 단위의 의미 사이의 차이에서 빚어지는 것이라고 볼 수 있는 것이다.[18]

---

18) 실제 어떤 어휘 단위가 가지고 있는 어휘의미론적 관계와 존재론적 관계를 구분하는 일은 쉽지 않다. 개별 어휘 단위의 존재론적 의미 속성이 그 사회의 문화적•정서적 함의에 의해 결정되는 상대적인 것이기는 하지만 다른 한편으로는 언어 기호가 가지고 있는 내재적 의미가 어휘 단위를 통해 동시적으로 표출되기 때문이다. 그러나 둘 사이의 차이는 분명하다. 예를 들어 한국인은 '비동양인=미국인'이라

의미장 빈도란 일반적으로 말하는 의미 주석을 기반으로 한 빈도라고 할 수 있는데, 이때의 의미 주석이 앞에서 이야기한 의미장을 토대로 한 분류체계를 전제로 한 주석체계에 의하는 것인가 아니면 단어군을 전제로 한 주석체계를 바탕으로 하는가가 문제로 제기될 수 있다. 의미 빈도를 구하기 위한 의미 주석에서 존재론적 의미와 어휘의미적 의미를 구분하지 않고 주석체계를 설정하게 되면 의미장 빈도와 단어군 빈도가 뒤섞이게 된다는 문제가 있는 것이다. 이를 분명하게 하기 위해 여기서는 의미장 빈도라는 용어를 사용한 바, 의미장 빈도란 각각의 어휘 단위에 대해 그것이 속하는 의미장 상의 상위 개념을 통해서 주석을 붙이고 그를 이용해서 개념별로 빈도를 산출한 것이다. 이는 무엇보다도 표현사전의 거시구조를 결정하는 데에 활용할 수 있다. 즉 같은 개념에 속하는 여러 단어 중에서 가장 핵심적인 의미를 가지고 있는 단어가 어떤 것인지를 파악하고, 유사한 의미를 가지는 단어들을 모아서 서술하려고 할 때에는 이러한 정보가 있어야 효율적인 사전 기술이 가능한 것이다.[19]

2.4. 단어족과 단어족 빈도

우선 『표준국어대사전』의 "문화" 항목의 내용을 보이는 것으로 논의를 시작하기로 하자.

고 생각하는 경향이 있다(서양인=미국인이 아니다). 그렇다고 해서 '미국인'의 상위개념어를 '비동양인'이라고 할 수는 없는 일인 것이다. 그러나 일본인만 해도 그렇지 않다. 일본인들은 서양 사람조차도 유럽인과 미국인을 구분한다. 게다가 일본인과 한국인이 지니고 있는 '미국인'에 대한 정서적 반응은 같지 않다(물론 한국인 사이에도 그 정서가 다를 수 있다). 이러한 차이를 어휘의미론적인 것으로 볼 것인가 아니면 존재론적인 것으로 볼 것인가 하는 데에는 견해의 차이가 있을 수 있는데, 학습자 사전의 개발에서는 이러한 문제를 고려한 사전 정보 구성이 필요하고 생각된다.

19) 국어의 의미 주석 및 개념 분류에 대해서는 졸고(1999)에서 간단히 정리한 바 있다. 그러나 거기서도 여기서 언급한 어휘 단위 사이의 존재론적 관계와 어휘의미론적 관계의 구분 문제에 대해서는 미처 다루지 못했는데, 이는 앞으로의 연구 과제로 삼으려 한다.

<예시 2> 『표준국어대사전』 '문화' 항목

문화[1](文化) 명 ① 자연 상태에서 벗어나 일정한 목적 또는 생활 이상을 실현하고자 사회 구성원에 의하여 습득, 공유, 전달되는 행동 양식이나 생활양식의 과정 및 그 과정에서 이룩하여 낸 물질적 정신적 소득을 통틀어 이르는 말. 의식주를 비롯하여 언어, 풍습, 종교, 학문, 예술, 제도 따위를 모두 포함한다. ¶구석기 문화/고급 문화/귀족 문화/근대 문화/유목 문화/전통 문화/문화를 교류하다/문화를 창조하다/새로운 문화에 접하다/찬란한 문화의 꽃을 피우다. ② 권력이나 형벌보다는 문덕(文德)으로 백성을 가르쳐 인도하는 일. ③ 학문을 통하여 사람들의 인지(人智)가 깨어 밝게 되는 것. 참 문명[3].

문화-적(一的) 관명 ① 문화와 관련된. 또는 그런 것. ¶문화적 특성//문화적인 터전/문화적인 충격. ② 높은 문화 수준에 있는. 또는 그런 것.

<예시 2>의 부표제항 '문화적'과 용례에 들어 있는 '구석기문화*, 고급문화, 귀족문화*, 근대문화*, 유목문화*[20], 전통문화'라는 어휘 단위는 모두 '문화'라는 어기를 공유한다. 어휘 구성이라는 면에서는 '문화'라는 어기를 바탕으로 만들어진 파생어이거나 합성어이며, 어휘 의미론적으로는 '문화'라는 어기의 의미를 기본 의미로 해서 통합되는 접사 및 선행하는 구성 요소 '구석기, 고급, 귀족, 근대, 유목, 전통'이라는 어휘 항목의 의미가 첨가되어 새로운 의미를 지니게 된 어휘 단위인 것이다.[21] 이들 같은

20) *를 한 것들은 『표준국어대사전』에 표제항으로 등재되지 않은 것들이다. 용례에 단어의 형태로 들어 있는 것인데 어떤 것은 표제항에 등재되고 어떤 것은 등재되지 않은 것들이 있을 경우에 등재 기준이 분명해야 할 것이다. 단어족에 대한 논의는 이 문제를 처리하는 기준에 대한 것도 포함한다.

21) 단순히 의미를 첨가한 것은 아니다. 어휘 항목별로 의미의 융합도가 다르기 때문에 일률적으로 이야기하기는 어렵지만, 단순히 구성 요소의 의미 조합으로 합성명사의 의미가 결정되는 것이 아님은 알고 있는 바와 같다.

어기를 바탕으로 만들어지는 파생어나 합성어들은 형태적으로 유사할 뿐만 아니라 의미적으로도 유연성(有緣性)을 지닌다. 이렇게 같은 어기를 가지고 있으면서 의미적으로 유연성을 지닌 일련의 단어들의 집합을 단어족(word family : 單語族)이라 한다.[22]

그렇다면 '국유문화재, 매장문화재, 문화재청, 불량문화재, 인간문화재' 등은 어떤가? 이들도 '문화'를 기본어기로 하는 어휘 단위인가? 이들은 이른바 2•3차 조어 과정을 거친 것들로 '문화'라는 어휘 단위가 들어 있기는 하나, 그것이 이들의 어기는 아니다. 이들의 어기는 '문화재'이며,[23] 각 어휘 항목의 의미는 '문화'의 의미보다는 '문화재'의 의미를 기본으로 한다. 『표준국어대사전』의 풀이를 보면 이것이 좀더 분명히 드러난다.

<예시 3> 『표준국어대사전』의 '문화재' 항목

문화-재(文化財) 명 ① 문화 활동에 의하여 창조된 가치가 뛰어난 사물. ② 문화재 보호법이 보호의 대상으로 정한 유형 문화재, 무형 문화재, 민속 문화재, 천연기념물, 사적, 명승지 따위를 이르는 말. 비① 문화물.

<예시 3>의 '문화재'에 대한 뜻풀이는 둘인데, 우리가 앞에서 예를 든 일련의 어휘 항목들은 '유형 문화재, 무형 문화재, 민속 문화재' 등을 예시한 뜻풀이 ②와 관련된 것임을 알 수 있다. '문화'라는 동일한 어휘 항목을 포함하고 있더라도 '유형 문화재, 무형 문화재, 민속 문화재'와 '구석기문화, 고급문화, 귀족문화, 근대문화, 유목문화, 전통문화'는 별개의 단어족을 이루는 것이다.

단어족 빈도(word family frequency)란 기본적으로 동일 어기를 바탕으로 만들어진 파생어•합성어의 빈도의 총합을 가리킨다.[24] 그러나 단순

22) 영어 어휘론에서의 단어족의 개념은 여기서의 개념 정의와 조금 차이가 있다. 굴절형이 단어족의 구성 요소의 하나가 되며, 합성어는 단어족에 넣지 않는다. 굴절 inflection과 파생derivation에 의한 것들만을 단어족의 구성 요소로 파악하는 것이다(Bauer & Nation 1993). 이와 관련된 문제는 주 24) 참조.

23) 이는 사전의 표제항 처리에도 반영된다. 하이픈이나 ∩(전문용어의 어휘 구성표시)는 모두 '문화재'라는 요소와 다른 요소 사이에 들어간다.

히 형태론적인 기준만으로 판별되는 것은 아니다. 앞에서 예를 든 것처럼 '문화재'는 '문화'에서 만들어진 단어이지만 그것이 별개의 기본의미를 지니게 되면서 또다른 어휘 단위를 생성하는 경우, 두 어기에서 만들어진 파생어나 합성어는 각각 별개의 단어족을 이룬다. 따라서 단어족 빈도를 구할 때에도 이들이 별도로 처리되어야 할 것임은 두 말할 나위도 없다.

3. 학습자 사전의 개발과 어휘 관련 빈도 정보의 활용

3.1. 단일어 학습자 사전의 유형과 거시구조 구성

학습자 사전의 유형은 분류 기준에 따라서 여러 가지로 나눌 수 있지만, 여기서는 한국어를 제2언어로 배우고자 하는 사람이 사용할 사전을

---

24) 앞에서 언급한 바와 같이 Bauer & Nation(1993)에서는 합성어를 단어족에 넣지 않았다. 기본적으로 그 논문의 목적이 영어 접사의 순위를 매기는 데에 있었기 때문이기도 하지만 굴절과 파생에 의한 어형들을 단어족으로 다루는 영어 어휘론의 전통적인 견해를 따른 것이라고 할 수 있다. 그러나 단어족이라는 개념을 국어에 대해 구체적으로 적용하는 데에 있어서는 몇 가지 문제가 있다. 우선 단어족의 한계를 정하는 데에 어휘의미론적 요소가 전혀 무관한 것이 아니라는 점을 지적해 두어야 할 것이다. 예를 들어 social과 socialism은 같은 단어족을 구성하는 것이 아니라고 보는 견해도 있는 것이다(Cruse 1986). 단어 형성이라는 측면에서 합성어를 단어족에 넣지 않는 것 역시 검토가 필요하다. 단어족이라는 개념이 어차피 단어 형성에 있어서의 생산성과 관련된 개념이라면 굳이 합성어를 단어족 안에 포함시키지 못할 까닭이 없는 것이다. 더구나 접두사와 관형사의 구분이 어렵고 합성명사나 파생명사는 만들어 낼 수 있지만 단독으로는 사용되지 않는 특수한 부류의 어기가 존재하는 국어를 대상으로 하는 경우 '어기(word-base)'의 빈도 및 생산성 그리고 어휘부 안에서의 어기 사이의 위계를 판별하는 기준으로 활용하기 위해서는 반드시 합성어를 단어족 논의에 포함시켜야 한다는 것이 필자의 생각이다. 물론 합성어를 대상으로 한 단어족 판별은 신중할 필요가 있다. 예를 들어 '유목 문화'나 '귀족 문화'와 같은 것들이 '문화'의 단어족을 이룬다는 데에는 별 이견이 없는 경우에도, '문화 정책'이나 '문화 공간'을 '문화'라는 어기를 중심으로 한 단어족에 포함시키는 데에는 이견이 있을 수 있는 것이다.

논의의 대상으로 삼는 바, 사전에 사용되는 언어에 따라서 단일어 사전(monolingual dicrtionary)과 두언어 사전(bilingual dictionary), 사전의 기능에 따라서 이해를 위한 사전(dictionary for comprehension)과 표현을 위한 사전(dictionary for expression)의 두 가지로 나누어 생각하면 충분할 것이다.[25] 그런데 이 글은 학습자 사전의 거시 구조 구성에 필요한 정보를 말뭉치에서 추출하는 방법을 다루는 것이므로, 두언어 사전 중에서 이해를 위한 사전과 단일어 학습자 사전에서의 거시구조 구성을 위한 언어정보 추출에 관련된 문제가 논의의 대상이 된다. 표현을 위한 두언어 사전의 거시구조를 구성하는 어휘 단위는 학습자의 모어이므로 여기서는 논의의 대상이 되지 않는 것이다.[26] 또한 두언어 사전 중 이해를 위한 사전의 거시구조 구성과 단일어 사전에서의 거시구조 구성은 기본적으로 동일한 원칙에 의해 이루어져도 크게 문제가 없다. 따라서 여기서는 단일어 사전에서의 거시구조 구성 문제를 중심으로 논의를 진행하기로 한다. 경우에 따라서는 단일어 사전의 번역에 의한 학습자 사전 개발이 용인되는 것도 그러한 까닭이다.

사전의 거시구조를 결정하는 과정은 사전의 편찬 목적과 방침에 따라 표제항을 선별하고 배열하는 것이라고 요약할 수 있다. 여기서 우선 결정되어야 할 것은 단일어 학습자 사전으로서의 이해를 위한 사전과 표현을

---

25) 단일어 학습자 사전은 기본적으로 이해를 위한 사전과 표현을 위한 사전으로서의 두 기능을 다 가진다고 할 수 있다. 이해를 위한 사전일지라도 표제항을 통해서 올바른 표기를 확인한다든가 발음을 익히는 등의 표현과 관련된 언어 지식의 습득을 목적으로 한 활용이 가능하고, 또 예문이나 재미있는 용례를 배워서 표현에 응용하는 등 사용자의 의도에 따라서 원래의 목적과는 다르게 이용할 수 있기 때문이다. 그것이 학습자 사전의 학습 교재로서의 기능이기도 한데, 여기서의 구분은 사전의 사용자가 이해와 표현이라는 구체적 목표를 위해 그 사전을 사용할 수 있는가 하는 점과, 사전 편찬자가 어떤 사용자를 염두에 두고 개발한 사전인가 하는 원래의 목적을 토대로 이야기하는 것이다.

26) 혹자는 비모어화자를 위한 단일어 사전이 이해를 위한 사전과 표현을 위한 사전으로 나뉠 수 있는가 하는 의문을 제기할 수도 있다. 그러나 영어 학습자 사전 중에서 *Longman Activator*와 *Cambridge International Dictionary of English*를 참조하면 그러한 의문은 자연스럽게 풀릴 것이다.

위한 사전이 표제항의 선별과 배열이라는 점에서 차이가 있느냐 그렇지 않으냐 하는 문제다. 우선 표제항 선별의 문제부터 살펴 보기로 하자.

표제항 선별의 문제란 이해를 위한 학습자 사전과 표현을 위한 학습자 사전의 거시구조를 이루는 어휘 단위가 별 차이가 없는가 아니면 달라야 하는가, 차이가 있다면 그것은 무엇인가 하는 점을 밝히는 것이다. 결론부터 이야기하자면, 이해를 위한 사전과 표현을 위한 사전은 거시구조를 이루는 어휘 단위가 분명히 다르다. 얼핏 생각하면 이해를 위한 사전에 꼭 등재되어야 할 어휘 단위라면 표현을 위한 사전에도 필요할 것이라고 생각하기 쉽기 때문에, 표제항의 선별이라는 문제는 그것이 이해를 위한 사전이건 표현을 위한 사전이건 큰 차이가 없는 것이 아닌가 생각할 수도 있다. 그러나 그렇지 않다.

사전의 표제항이 가지는 중요한 기능 중의 하나는 지시성이다. 사용자가 자신이 필요로 하는 정보를 찾을 수 있도록 해주는 지표(index)로서의 역할을 하는 것이다. 이런 관점에서 보면, 이해를 위한 사전과 표현을 위한 사전이 지니고 있는 가장 큰 차이 중의 하나가 사용자의 기지성(旣知性)이다. 즉 사전의 사용자가 자신이 필요한 정보를 찾아가는 데에 지표로 사용하는 언어 기호에 대해 무엇을 알고 있느냐 하는 점인 것이다.

이해를 위한 사전은 기본적으로 어떤 언어 표현을 접했을 때 사용자가 알지 못하는 정보를 확인하는 데에 사용한다.[27] 사용자는 언어 기호의 외적 형태만 알 뿐 의미·용법을 알지 못하기 때문에 사전을 이용하는 것이다. 따라서 이해를 위한 사전의 경우, 거시구조를 이루는 표제항은 사용자가 현실적으로 접하게 되는 언어 형식을 고려해서 선별할 필요가 있다. 즉 사용자가 어떤 어휘 단위를 자주 접하는가와, 그들 어휘 단위의 의미·용법을 사전에서 검색하려 할 때에 어려움을 겪는 '지표'로서의 기능을 고려해서 선별할 필요가 있는 것이다.

---

27) 앞에서 간단히 언급한 바 있지만, 이해를 위한 사전이라고 해도 맞춤법이나 한자를 확인하기 위해서 사용하는 경우에는 표현적인 기능을 발휘하게 된다. 이 경우는 사용자는 이미 자신이 찾고자 하는 것을 분명히 알고 있다기보다는 막연하게 알고 있는 것을 확인하는 과정을 거치게 된다.

그러나 표현을 위한 사전의 경우는 다르다. 사용자는 어느 정도 어휘력을 지니고 있으며, 자신이 표현하려는 바가 자신이 알고 있는 것과 일치하지 않는다고 생각할 때에 사전을 찾는다. 즉 형태와 의미를 모두 아는 어휘 단위에 대해 만족하지 못해서 좀더 적절한 표현을 찾고자 할 때에 사전을 사용하게 되는 것이다. 따라서 표현을 위한 사전의 거시구조를 이루는 표제항은 기본적인 어휘 단위의 의미와 용법에 대해서 알고 있는 사용자가 보다 세분화된 의미 용법에 따라 적절한 어휘 단위를 선택하려는 노력을 도울 수 있도록 선별되어야 한다.

표제항의 배열이라는 면에서도 표현을 위한 사전과 이해를 위한 사전이 달라질 수 밖에 없다는 것은 분명하다.

모어 화자조차 자신이 원하는 표현에 적절한 어휘를 선택하는 데에 어려움을 겪는 경우가 많고, 이러한 이들을 위해 표현을 위한 별개 형식의 사전이 요구된다는 것이 잘 알려져 있는 바와 같다. 모어 화자를 위한 표현을 위한 사전의 대표적인 형태가 유의어 사전과 시소러스(thesaurus)이다.[28] 유의어 사전은 동일한 의미영역에 속하는 단어들 - 유사한 의미를 지니고 있지만 의미상의 차이를 가진 단어들-을 한데 모으고, 용법과 의미를 구체적으로 설명해 줌으로써 사용자가 자신이 표현하고자 하는 내용에 가장 적절한 단어를 선택할 수 있도록 한 사전이다. 표제항의 배열 방식이 이해를 위한 사전과는 전혀 달라지는 것이다. 따라서 이해 어휘와

---

28) 전통적인 개념의 시소러스(thesaurus)는 유의어 사전(Dictionary of Synonyms)과는 구별되는 것이었다. 배열도 자모순에 의한 것이 아니고, 표제항의 의미를 풀이하지도 않았던 것이다. 로젯Roset의 시소러스가 그 대표적인 예라고 할 것이다. 그러나 시소러스와 같은 비자모순 배열 방식의 사전은 검색에 불편을 느낄 수 밖에 없고, 따라서 사용자에게 환영을 받지 못한다. 이를 극복하기 위해 고안된 것이 일반사전과의 혼합형 시소러스와 자모순 시소러스라고 할 수 있다. 전통적인 형태를 고수하던 로젯Roset의 시소러스조차 1990년대에 들어 21세기형 시소러스(21st Century Thesaurus)라고 해서 자모순으로 그 형태를 바꾼 것을 동시에 출판하고 있는 사실이 그것을 잘 보여준다. 결국 표현을 위한 사전의 배열 방식도 자모순 배열을 지향하게 된 것이고, 여기에 개개 표제항의 의미에 대한 구체적 기술이 더해지면 시소러스와 유의어 사전은 구분이 어렵게 된다(Hartmann & James 1998 : 142-143).

사용 어휘 사이에 큰 간격이 있는 제2언어 학습자에게 표현을 위한 사전이 필요하리라는 것은 두말할 나위도 없는 것이다. 무엇보다도 학습자가 자신이 알고 있는 단어를 통해서 그것과 유사한 의미를 가지는 어휘 단위들의 집합에 접근할 수 있어야 하고, 또 자신이 원하는 표현을 쉽게 찾을 수 있는 구조를 지녀야 할 것이기 때문이다.[29]

그러나 이해를 위한 사전과 표현을 위한 사전의 표제항 배열은 가능하면 같은 방식에 의하는 것이 바람직하다. 기본적으로 이해를 통해서 언어를 습득하고, 그를 바탕으로 표현 능력을 늘려가는 것이 제2언어의 습득 과정이라면, 학습자는 이해를 위한 사전을 먼저 그리고 자주 사용하게 될 것이며, 따라서 표현을 위한 사전의 표제항 배열이 이해를 위한 사전의 그것과 달라진다면 사용자에게는 사용하기 불편한 사전이 될 것이기 때문이다. *Longman Activator*나 *CIDE*(*Cambridge International Dictionary of English*)가 취한 방식은 그러한 문제를 고려한 것이라고 할 수 있다.[30]

3.2. 거시구조 구성과 빈도 정보의 활용

3.1.에서 논의한 바와 같이 동일한 언어를 다루는 학습자 사전이라 해도 이해를 위한 사전과 표현을 위한 사전은 그 거시구조 및 미시구조의 구성과 배열이 달라질 수 밖에 없다. 따라서 말뭉치의 분석을 통해 얻어낸 빈도 정보의 활용에 있어서도 서로 다른 정보가 활용되는 것은 당연

---

29) 뿐만 아니라 미시구조의 측면에서 유의어군 중 특정 어휘 단위를 중심으로 유의어군을 다룰 것인가 아니면 유의어들을 대등하게 처리하되 비교가 가능하도록 지표를 만들어 사용할 것인가에 따라서 표현을 위한 학습자 사전의 거시구조는 달라질 것이다.

30) 그러나 거시구조의 배열방식이 같아진다고 해서 미시구조까지 같아지는 것은 아니다. 이해를 위한 사전에서는 표제항의 의미를 위주로 미시구조가 구성되지만, 표현을 위한 사전은 언어 표현 사이의 차이를 사용자가 쉽게 비교할 수 있도록 하기 위해서 대부분 유의어의 비교라는 방법을 택하게 되고, 따라서 표제항의 유의어들을 여하히 배열하고 그 차이를 보여줄 것인가 하는 점을 중심으로 미시구조가 구성되기 때문이다.

하다. 여기서는 우선 이해를 위한 사전의 거시구조 구성에 활용될 수 있는 것들을 검토해 보기로 한다.

이해를 위한 학습자 사전의 표제항 선별 및 배열에는 기본적으로 형태빈도가 큰 몫을 맡는다고 할 수 있다. 사용 빈도가 높은 어휘 단위일수록 학습자가 자주 접하는 것일 터이고, 따라서 이해를 위한 학습자 사전의 표제항으로 등재할 필요성이 높아지는 것이다. 형태빈도에는 2.1.에서 정리한 것처럼 어절 빈도•단어 빈도•문법 형태소 빈도가 기본적인 것이지만, 단어빈도나 문법형태소 빈도는 이미 사전 개발에 활용되고 있어서 그 중요성을 재삼 강조하지 않아도 될 것이라고 생각되므로 어절 빈도를 예로 들기로 한다.[31]

2.1.에서 이야기한 것처럼 한국어를 모어로 하지 않는 학습자는 국어의 굴절형이 어떻게 분절되는가를 아는 데에 크게 어려움을 겪는다. 유타니(1999, 2001)에서 지적하고 있는 용언의 활용형 및 동형이어(同形異語)의 문제가 그 대표적인 예라고 할 것인데, '줄'이라는 어절을 예로 들어 보기로 한다.

'줄'이라는 150만 어절 말뭉치에서 134번째로 자주 쓰이는 어절로 모두 701번 사용되어 어절의 텍스트 점유율이 0.0466% 누적점유율 15.4511%이다.[32] 이 어절은 150만 어절 말뭉치에서 쓰인 예에서 모두 6개의 어간으로 해석되는데, 701개의 용례가 각각 어떤 어간의 굴절형이며 무슨 의미로 쓰이는가를 분석한 본 결과는 다음과 같다.[33]

---

31) 그렇다고 해서 단어나 문법형태소의 빈도 정보를 활용한 사전 정보 구성 방식에 문제가 전혀 없다는 뜻은 아니다. 이는 후일의 과제로 남긴다.

32) 여기서 제시한 것은 21세기 세종 계획의 결과물 중 하나로 공개된 <글잡이Ⅱ(직접)>로 어절 빈도를 분석한 결과이다. 다만 누적 점유율은 <글잡이Ⅱ(직접)>에서는 구현되지 않아 별도로 계산하였다.

33) 분석을 위한 KWIC의 작성에는 hgrep97을 이용하였다. <글잡이Ⅱ(직접)>의 용례 검색 결과는 정렬이라든가 배열을 사용자가 결정할 수 없고 불필요한 내용이 지나치게 많이 섞이기 때문이다. 또한 <깜짝새>의 KWIC 검색기능을 이용해서 어절 단위 "줄"을 검색하면 681개 밖에 없는 것으로 나타남도 지적해 두어야겠다. <깜짝새>의 어절 빈도 산출 기능을 사용할 때와 문맥 검색 기능을 사용할 때에 처리 방식에 차이가 있음을 보여주는 것이기 때문이다.

줄1 (110) : 주(타동사)+ㄹ
ex. 그 꽃 누구에게 줄 거니?
줄2 (219) : 주(보조동사)+ㄹ
ex. 집안어른이나 가까이서 돌봐 줄 남자 친척 하나 없이
줄3 (2) : 줄(자동사)+(ㄹ)
ex. 재수생의 비율이 예년에 비해 크게 줄 것으로 전망했다.
줄4 (17) : 줄(명사)
ex. 눈 감고 줄 없는 거문고를 타는 마음이
줄5 (353) : 줄(의존명사)
ex. 한창 시간 가는 줄 모르게 호조황(好釣況)을 즐겼다.
줄6 (1) : 줄(명사)
ex. 쇠붙이를 깎거나 다듬는 데 쓰이는 줄

분석 결과를 간단히 정리하면, '줄'이라는 어절의 쓰임새는 의존명사로 쓰이는 비율이 가장 높으며, 그 다음이 보조동사 '주다'의 활용형이고, 본동사 '주다'의 활용형인 경우가 세 번째이다.

당연한 일이지만 일반적인 국어사전에서는 이러한 사실은 고려하지 않는다. 예를 들어 『표준국어대사전』에는 12개의 '줄'이, 『연세한국어사전』에는 모두 4개의 '줄'이 표제항으로 실려 있는데 의존명사 '줄'이 네 번째로 실릴 뿐,[34] '줄'이 동사 '줄-'과 '주-'의 활용형이라는 정보는 담겨 있지 않다.[35] 모어화자라면 대부분 어절을 분석해서 어간형을 찾아내는 데에

34) 동음이의어 배열도 학습자 사전과 일반 국어사전은 달라져야 할 것이다. 지금의 국어사전에서의 동음이의어 배열은 단어의 역사와 품사, 한자 표기 등을 고려한 것인데, 이는 한국어의 학습과는 무관한 것이기 때문이다. 그러나 현재로서는 빈도 이외에 어떤 순서로 동음이의어를 배열하는 것이 바람직한가에 대한 구체적인 방안은 찾지 못했다. 그러나 단순히 빈도에 의존하는 방식은 경우에 따라서 배열의 기준이 달라진다는 문제를 안고 있어서 그대로 채용할 수 없다.

35) 별개의 문제이지만, 국어사전에서 용언의 표제형을 '-다'와 결합한 형태로 등재하는 관행도 학습자 사전에서는 사용자에게 불편을 주는 요소일 수 있다. 형용사와는 달리 동사는 '기본형'이 문어에서나 구어에서 실제 쓰이는 경우는 거의 없기

크게 어려움을 느끼지 않기 때문이다.

그러나 앞에서 이야기한 바와 같이 한국어의 어절이 어떻게 분절되는가를 잘 알지 못하는 학습자들에게는 기존 국어사전과 같은 표제항 구성 방식은 적절하지 못하다고 할 수 있다. 학습자 사전에 대해서 단순한 참조용 문헌이 아닌 한국어 학습의 교재로서의 역할을 기대한다면 우선 표기된 형태를 바탕으로 학습자가 알아야 할 정보를 제공하는 것이 바람직하다는 것이 그 이유의 첫째고, 둘째는 참조용 문헌으로서도 표제항이 가지고 있는 지표(index)로서의 기능을 제대로 수행하지 못하는 것이라고 할 수 있기 때문이다. 이러한 문제를 해결하기 하는 데에 어절 빈도 정보를 적극적으로 활용하는 것이 바람직한 것이다.

표현을 위한 학습자 사전의 표제항 선별 및 배열에는 형태 빈도도 중요하지만, 앞에서 이야기한 단어족 및 의미장 빈도가 큰 몫을 할 수 있다. 특히 의미장을 바탕으로 한 개념빈도의 조사와 그것을 바탕으로 한 대표어의 선정 그리고 하위어의 배열은 표현을 위한 사전의 효율성을 높이는 데에 기여할 수 있다.

4. 남은 문제들 : 결론을 대신하여

이 글은 학습자 사전의 개발에 필요한 기초적인 어휘 관련 정보를 말뭉치의 계량적 분석을 통해 얻어내는 것과 관련된 문제를 정리하는 것을 목적으로 한 것이었다. 그러나 지면 관계상 미시구조의 구성과 관련된 문제나 실제 정보의 획득 방법은 구체적으로 서술하지 못하고, 학습자 사전의 유형에 따른 거시구조 구성과 관련된 문제를 중심으로 개념을 다루는 정도에 그치게 되었다. 이제 앞에서의 논의에서 미처 다루지 못했거나 논

때문에 형용사와 동사를 구분하지 못하는 학습자로서는 소위 기본형을 재구성하는 데에 어려움을 겪는 것이다. 사전에 등재할 용언의 대표형이 어떤 것이 되어야 할 것인지는 재검토할 필요가 있는 것이다.

의가 부족했다고 생각되는 것들을 정리함으로써 결론에 대신하고자 한다.

2장에서의 논의는 말뭉치에서 얻어낼 수 있는 빈도 정보의 유형을 정리하였다. 어절 빈도의 활용과 관련해서는 중의성을 가지는 어절들의 중의도[36]를 검토하지 못했다는 것이 문제로 남는다. 이는 빠른 시일 안에 다시 정리해서 그 의미와 함께 다루어보기로 한다. 또한 중의적이지 않더라도 어절의 분절을 통해 어간형을 확인하기 쉽지 않은 것들의 유형과 각 유형별 종수를 확인하는 문제도 앞으로 해결해야 할 문제 중의 하나다. 단어족 및 단어족 빈도의 활용과 관련해서는 단어족 빈도 추출의 대상이 되는 어기의 유형과 각각의 유형을 선정하는 방법 등이 서술되어야 할 것이고, 이와 관련해서 한자 형태소 및 이른바 단어형성적 어기의 처리에 대한 논의가 보충되어야 할 것으로 보인다. 이 역시 앞으로의 과제로 남긴다. 의미 빈도와 관련해서는 어휘의미론적 의미장의 하위범주와 존재론적 단어군의 하위 범주가 어떻게 구체적으로 다를 것인지와 그것들을 어떻게 정량화해서 사전 개발에 활용할 것인가가 가장 큰 문제라고 할 수 있다. 앞에서도 언급한 바 있지만, 외국어의 학습이 기본적으로는 어휘의미론적 의미 대응을 바탕으로 하고 있지만 존재론적 의미에 대한 이해 없이는 진정한 의미의 외국어 습득이 어려운 일이기 때문에. 어떻게 이러한 구분을 사전에 반영할 수 있을 것인가가 구체적으로 검토되어야 하는 것이다.

3장에서는 학습자 사전의 유형에 따른 거시구조의 차이 및 거시구조 구성에 있어서이 어절 빈도의 활용과 관련된 문제가 논의의 중심이었는데, 실제 사전 구성의 차이를 구체적으로 보여주지 못했다는 것이 가장 큰 문제일 것이다. 그러나 이것은 표현을 위한 학습자 사전의 거시구조의 형태를 결정하는 문제와 함께 다루어야 하는 문제이기 때문에 그리 간단한 것은 아니다. 이것 역시 앞으로의 과제 중 하나로 삼는다.

---

36) 중의도(重意度)란 일정한 범위에 속하는 어절들이 몇 가지의 중의적 의미를 가지는가를 중의성을 가지는 어절의 수로 나눈 것으로, 고빈도 어절과 저빈도 어절 사이의 중의도 비교를 통해서 어느 정도의 텍스트 점유율을 가지는 어절들이 높은 중의성을 가지는가를 밝히는 척도가 된다.

참고 문헌

김용옥(1990), "번역에 있어서의 시간과 공간", 『동양학 어떻게 할 것인가』, 통나무.

배주채(2000), 『초급 한국어 사전 개발을 위한 연구』, 연구결과보고서, 문화관광부.

배주채(2001), "외국인을 위한 한국어 사전의 방향", 『성심어문논집』 23, 성심어문학회.

서상규 • 남윤진 • 진기호(1998), 『외국어로서의 한국어 교육을 위한 기초 어휘 선정 I -기초어휘빈도조사결과』 사업결과보고서, 한국어 세계화 추진위원회, 문화관광부.

서상규 • 최호철 • 강현화(1999), 『한국어 교육 기초어휘 의미빈도사전의 개발』, 사업결과보고서, 한국어세계화추진위원회 • 문화관광부.

서상규 • 강현화(2000), 『한국어 교육 기초어휘 의미빈도사전의 개발』, 사업결과보고서, 한국어세계화추진위원회 • 문화관광부.

서상규 • 한영균(2000), 『국어정보학입문』, 태학사.

유타니 油谷行利(1999), "한국어 학습자를 위한 한일사전에 대하여", 『사전편찬학연구』 9, 연세대학교 언어정보개발연구원.

유타니 油谷行利(2001), "조선어 사전의 편찬 - 학습자를 위한 한일사전", 『한국어 교육과 학습사전』- 제2회 한국어 교육 국제워크숍 발표자료집, 연세대학교 언어정보개발연구원.

한영균(1998), ≪한국어 기본어휘의미빈도용례사전≫의 개발을 위한 기초적 연구, 연구결과보고서, 양영학술재단.

한영균(1999), 『전자말뭉치를 이용한 사전편찬론』, 연구결과보고서, 문화관광부.

한영균 • 서상규(2000), 『한국어 말뭉치의 활용 I』, 연구결과보고서, 문화관광부.

한영균(2001), "파생접사 및 파생어의 사전적 처리와 말뭉치 활용 - '반(反)-'과 '-질'의 경우", 『이광호선생회갑기념논문집』, 태학사.

한영균(준비중), "한국어 학습자 사전 개발을 위한 몇 가지 전제".

한영균(준비중), "한국어의 기본어휘 - 선별과 응용을 중심으로".

한영균(준비중), "어휘 학습 자료의 개발을 위한 계량적 연구의 한 방향 - 기본어기 단어족 빈도 사전의 개발"

Asher, R. E. & J. M. Y. Simpson(19?? eds), *The Encyclopedia of Language and Lingustics*, Pregamon Press.

Bauer, L & P. Nation(1993), "Word Families", *International Journal of*

Lexicography, Vol. 6(4):253-297.

Carter, Ronald & Michael McCarthy(1988), *Vocabulary and Language Teaching*, Longman.

Hartmann, R.R.K & G. James(1998) *Dictionary of Lexicography*, Routledge.

Herbst, Thomas. & Kernst Popp (eds.) (1999), *The Perfect Learners' Dictionary*, Niemeyer.

Marzano, Robert J. & Jana S. Marzano(1988), *A Cluster approach to Elementary Vocabulary Instruction*, IRA.

Nation, I.S.P.(2001), *Learning vocabulary in Another Language*, Cambridge.

Stark, Martin(1999), *Encyclopedic Learners' Dictionary*, Niemeyer.

<부록> 어절 빈도 분석 프로그램의 기능 비교 분석

- <글잡이Ⅱ(직접)>와 <깜짝새>

말뭉치를 대상으로 한 어절 빈도의 추출은 말뭉치를 활용한 어휘계량적 연구에서 가장 기초적인 것에 속한다. 그러나 이것조차도 대규모 말뭉치를 대상으로 하는 경우 간단치 않다. 어절에 대한 인식 방식에 따라서 어절 빈도 추출 프로그램의 기능 구성이 달라지고 그에 따라 분석 결과도 상당히 달라지기 때문이다. 논문의 본론과 직접 관련은 없지만, 어절 빈도의 활용과 밀접한 관계를 가지는 것이므로 정리해 두기로 한다.

어절 빈도는 기본적으로 띄어쓰기를 그 처리 단위로 한다. 그런데 우리는 띄어쓰기만이 아니라 문장 부호[1]를 어절 구분의 기준으로 인식한다. "**있다(그림 있다(헌법재판소**"와 같은 형식은 띄어쓰기를 바탕으로 한다면 한 어절로 다루어야 할 것이지만, 사람은 이것을 한 어절이라고 생각하지는 않는다는 말이다. "**있다(그림**"은 "**있다 (그림**"의 조합으로 "**있다(헌법재판소**"는 "**있다 (헌법재판소**"라는 세 요소의 조합으로 인식하는 것이다. 반면 띄어쓰기에만 의존하는 어절 빈도 추출 프로그램은 "**있다(그림 있다(헌법재판소**"를 각각 하나의 어절로 인식한다. 직관과 다른 결과를 낳는 것이다. 이러한 문제를 해결하려면 문장 부호가 개재된 어절은 그것을 문장 부호를 기준으로 분리하면 될 것이 아닌가 생각하기 쉬운데, 그것이 그렇게 단순하지 않다. 그렇게 처리할 경우 '**있다)에 있다) 와도**'와 같은 형태를 '**있다) 에 와도**'로 분석하게 되어 실제 존재하지 않는 가공의 어절 **에 와도**를 어절 빈도에 포함하게 된다는 또다른 문제

1) 사실은 문장 부호만이 아니라, 시각적 효과를 위해 인쇄에 쓰이는 특수 기호와 문자 모두를 그렇게 인식한다. 편의상 이렇게 표현한 것이다.

를 야기하는 것이다.[2)]

본 연구를 위해 사용한 프로그램의 예를 들어 문제를 좀더 분명히 드러내 보이기로 한다. 대규모 말뭉치를 대상으로 한 계량적 분석에서 수작업에 의존할 수는 없다는 것은 두말할 나위도 없는데, 어절 빈도 추출 프로그램별로 기능에 조금씩 차이가 있어서 다른 수치를 보여 주는 바, 그 차이가 구체적으로 어느 정도의 수치로 나타나는가를 분명히 인식할 필요가 있기 때문이다.[3)]

여기서의 논의를 위한 말뭉치로는 고려대학교 민족문화연구소 전자텍스트연구소에서 구축한 150만 어절 형태 분석 말뭉치를 활용하였고, 자료처리 프로그램으로는 21세기 세종계획의 결과물로 보급된 <글잡이Ⅱ(직접)>과 전주대학교 소강춘 교수가 공개한 <깜짝새>를 이용하였다.[4)] 불과 150만 어절 규모의 말뭉치를 대상으로 한 것이었는데도 어절 빈도의 차이는 결코 적은 것이 아니었다. <글잡이Ⅱ(직접)>를 이용해 분석하면 150만 어절 중의 이형 어절의 가짓수가 347,949 개가 검출되는데, <깜짝새>에서는 308,586 개가 검출되는 것이다. 두 프로그램의 결과가 39,363 개나 차이를 보여, 10%를 넘는 오차를 보이는 것이다. <표 1>은 그것을 확인할 수 있도록 정리한 것인데, 빈도 순위 순으로 정리한 상위의 어절이 차지하는 텍스트 점유율, 누적점유율, 그리고 각 구간에서의 1000어절당 평균 텍스트 점유율을 보인 것이다.[5)]

---

2) 실제로는 이 '와'와 '에도'는 같은 문장 안의 앞 부분에 있는 (앞의 요소에 첨가되는 것이므로, 그것을 찾아 결합형으로 처리해 줄 수 있어야 할 것이다. 그러나 자료 처리 시스템 개발자들은 그러한 문제까지는 고려하지 않으려 한다.

3) 이러한 문제의 인식을 통해서 어절 빈도 추출 프로그램의 개선이 이루어졌으면 하는 바람도 포함된다.

4) 이 말뭉치와 두 프로그램을 활용할 수 없었으면 이 글을 쓰는 것은 애당초 불가능했을 것이다. 이 자리를 빌어 자료와 프로그램을 개발하고 공개한 분들에게 감사하는 마음을 전한다.

5) 빈도 순위 50,000 까지만을 보인 것은 두 가지 이유에서다. 첫째는 그 이상을 비교하는 것은 의미가 없다고 판단되었기 때문이다. 빈도 순위 50,000 번째를 넘어가면 각 어절의 빈도는 4를 넘지 못한다. <글잡이Ⅱ>의 분석결과는 빈도수 4 이상인 어절의 총수가 48,994개이며 빈도수 3인 어절은 18,344개다. <깜짝새>의 분석 결과

<표 2> 어절 빈도 순위별 텍스트 점유율

빈도 순위	깜짝새			글잡이Ⅱ(직접)		
	텍스트 점유율	누적 점유율	1000어절 평균 점유율	텍스트 점유율	누적 점유율	1000어절 평균 점유율
0 - 1000	32.6661	32.6661	32.6661	30.8058	30.8058	30.8058
- 2000	7.1521	39.8182	19.9091	6.7314	37.5372	18.7686
- 3000	4.5081	44.3263	14.7754	4.3344	41.8716	13.9572
- 4000	3.3301	47.6564	11.9141	3.2191	45.0907	11.2727
- 5000	2.6624	50.3188	10.0638	2.5715	47.6622	9.5324
- 10,000	8.4542	58.7830	5.8783	8.2486	55.9108	5.5911
- 30,000	12.4565	72.1395	2.4047	13.1384	69.0492	2.3016
- 50,000	6.0007	78.1402	1.5628	5.9359	75.0031	1.5001

결국 어떤 기능상의 차이 때문에 이런 문제가 발생하는가를 확인할 필요가 있었는데, 두 프로그램에서 추출된 어절들을 비교 검토한 결과 잠정적으로 다음과 같이 추론할 수 있었다.

<글잡이Ⅱ(직접)>는 어절 내부 혹은 어절의 앞뒤에 문장 부호가 사용된 경우 각각 다른 어절로 처리한다. 그러나 <깜짝새>는 이들 어절을 문장 부호를 기준으로 다시 분리하고 분리된 각각의 어절을 처리한다. 예를 들어 **"있다! 있다" 있다"고 있다"는 있다"며 있다"면서 있다' 있다'고 있다'는 있다(그럼 있다(헌법 있다(헌법재판소법 있다) 있다). 있다)에 있다, 있다. 있다?"**와[6] 같은 어절들이 있을 때, <글잡이Ⅱ(직접)>은 이

빈도수 4 이상인 어절의 총 수는 49,968개이고 빈도수 3인 어절은 18,273개, 빈도수 2인 어절은 41,942개이다. 둘째는 두 프로그램으로 분석한 전체 어절 수가 다르기 때문에 빈도순으로 전체를 보일 수 없었기 때문이다. 전체 어절수 비교는 점유율을 중심으로 정리한 <표 2>에 제시하였다.

6) 굵은 글씨로 된 문장 부호는 인용 대상을 나타내며, 보통 글씨의 " "은 인용 부호이다.

들을 모두 각각 다른 어절로 처리하지만, <깜짝새>에서는 문장 부호가 개재된 **있다!, 있다"는 있다"며** 등을 각각 **있다**와 **!, 는, 며**로 분리하여 합계를 낸다. 뿐만아니라 "**있다(그림 있다(헌법**"과 같은 어절의 분석 결과는 각각 "**있다 (그림 헌법**"으로 분석된다. 그 결과 <깜짝새>에서의 분석 결과는 "**있다**"의 빈도가 9,158 개인데, <글잡이Ⅱ(직접)>에서는 "**있다.**"만 8,570 개인 것으로 검출되는 것이다(다른 것은 일일이 예로 들지 않는다).

이러한 양자의 기능은 서로 장단점이 있다. <글잡이Ⅱ(직접)>는 있는 그대로의 어절의 모습을 보여주는 것이라는 점이 장점인 반면 **있다. 있다! 있다"**와 같이 부호가 첨가되어있는 동일 어절을 별개의 어절로 간주하기 때문에 다시 한번 계산을 할 필요가 있다거나, **있다(그림 있다(헌법재판소**와 같이 도저히 한 어절로 인정하기 어려운 형태를 한 어절로 처리한다는 것이 단점이라면, <깜짝새>의 경우에는 문장 부호를 모두 자동적으로 분리해냄으로써 '있다'의 빈도 안에 **있다. 있다! 있다"** 등의 빈도가 함께 포함되어 있어서 별도의 처리가 필요없는 점이 장점이지만, 문장 부호 다음의 비자립 단위를 모두 자립적인 어절로 계산해 내기 때문에 실제 존재하지 않는 가공의 어절이 어절 빈도 안에 포함되는 점이 단점이고, 프로그램을 통해 추출된 어절을 KWIC을 작성하는 경우 어절 빈도 산출된 모든 어절의 용례가 추출되지 않는 것도 문제다. 문자열 구성이 다른 것을 같은 어절로 계산한 것들은 KWIC의 작성에서는 추출되지 않기 때문이다.

두 프로그램이 지니고 있는 이러한 기능상의 차이로 말미암아 <글잡이Ⅱ(직접)>와 <깜짝새>는 말뭉치 안에 들어 있는 어절 종수 분석에서 상당한 차이를 보이게 된 것이다. 이러한 차이를 구체적인 수치로 비교한 것이 <표 2>이다.

<표 2>는 텍스트 점유율을 높이는 데에 몇 개의 어절이 추가되는가를 비교할 수 있도록 만든 것인데, 상위 고빈도 어휘의 수가 <깜짝새>의 분석 결과가 <글잡이Ⅱ>의 분석 결과에 비해 상대적으로 적은 것을 확인할 수 있다. 흥미로운 것은 점유율 10%을 이루는 상위 고빈도 어휘의 수

<표 3 > 텍스트 점유율에 따른 어절 수의 증가

텍스트점유율	깜짝새		글잡이Ⅱ(직접)	
	추가어절수	누적어절수	추가어절수	누적어절수
5%	12	12	12	12
10%	36	48	36	48
15%	71	119	77	125
20%	119	238	141	266
25%	200	438	246	512
30%	319	757	402	914
35%	505	1,262	639	1553
40%	772	2,034	972	2525
45%	1,148	3,182	1,444	3,969
50%	1,686	4,869	2,131	6,100
100%	303,690	308,586	341,847	347,947

가 동일하다는 것인데, 실제 어휘 구성을 비교하면 양자의 분석 결과에 차이가 있다. <예시 1>에 그것을 정리해 두었다.

<깜짝새>의 분석 결과에는 들어 있는데, <글잡이Ⅱ>에는 들어 있지 않은 것은 "의, 을, 는"이라는 세 형태이고, 그 반대의 것은 "/", ""라는 두 개의 문장 부호와 "가장"이다. 앞에서 지적한 바와 같이 <깜짝새>의 분석 결과에 들어 있는 것은 가공의 어절이다. 실제 띄어쓰기를 한 형태로 나타나는 "의, 을, 는"이라는 형태는 어디서도 나타나지 않는 것이다. 반면 "가장"의 경우에는 <깜짝새>의 분석 결과에서는 빈도수로 51번째의 순위를 차지해서 점유율 10%-15%를 구성하는 어휘 중에 들어 있는 것이다. 비교 범위를 확대하면 차이가 더욱 잘 드러난다.

<예시 2>는 상위 빈도 1000어절을 비교해서 각각의 분석 결과에만 들어 있는 어절을 보인 것이다. <깜짝새>의 분석 결과에만 들어 있는 어절 중에서 진한 글씨로 보인 것들은 많은 경우 문말에서 문장 부호와 함께

<예시 1> <깜짝새>와 <글잡이Ⅱ>의 고빈도 어절 분석 결과(1)

	점유율 10% 상위 고빈도 어휘 구성
깜짝새	그 수 이 있다 있는 것이다 한 것은 그러나 것이 대한 나는 한다 하는 있었다 할 같은 하고 또 우리 때 등 것 그리고 것을 의 그는 없는 했다 더 다른 없다 것으로 때문에 두 다시 그런 아니라 어떤 을 내 내가 이런 이러한 위해 잘 그의 는
글잡이Ⅱ	그 수 있다. 있는 이 것이다. 한 " 것은 것이 그러나 대한 나는 있었다. 하는 한다. 할 같은 또 우리 등 그리고 그는 것을 없는 했다. 때 더 다른 하고 것으로 때문에 없다. 두 다시 그런 어떤 것 이러한 이런 내 / 내가 위해 그의 잘 아니라 **가장**

<예시 2> <깜짝새>와 <글잡이Ⅱ>의 고빈도 어절 분석 결과(2)

	상위 빈도 1,000 어절 중 공통되지 않는 어절
깜짝새	개혁 **것인가 것일까** 과 광주 교수 군 **그래 그럼 그렇다** 길 는 대표 도 도시 돈 **됐다** 둘째 라고 라는 로 를 만 며 **무엇인가** 뭐 반면 사건 사설 생산 소리 시 시대 아 **아닌가** 아들 아버지 어머니 엄마 에 에서 여자 와 운동 원 으로 은 을 의 의원 의하면 이다 이라는 이란 인 **있는가** 있었고 있었는데 **있을까** 자장 좋다 주 하는데 **하자**
글잡이Ⅱ	공동 관련된 국가의 그와 기간 기능을 기본 기존의 나누어 나머지 날이 대통령이 데리고 동네 5월 땅을 때로는 떨어져 떨어진 마치고 말에 먹을 멀리 가까이 모를 모여 몹시 무엇이 문득 문화를 미적 발전을 배가 불을 비가 사회를 사회에서 상호 생긴 손에 시를 씨의 아이들은 아이들이 아파트 알게 어쩔 없게 역사의 열 열린 오늘의 이루고 이르는 작업을 잡고 장윤 저녁 적은 점차 찾는 최초의 특별 표정을 하나를 학교에 현재의 활동을

쓰이는 것이고, 또 '과,군, 는, 도, 라고, 라는, 로, 를, 만, 며, 에, 에서, 와, 으로, 은, 을, 의, 이다, 이라는, 이란, 인' 등은 문법형태소임을 짐작할 수 있는 것들이다. <글잡이Ⅱ(직접)>의 분석 결과가 문장 부호가 포함된 실제 텍스트에 나타난 모든 띄어쓴 형태를 어절로 처리하여 보여준 것이라면, <깜짝새>의 분석 결과는 가공의 어절들(주로 문장 부호가 선행하는 문법형태소가 그 대부분이다)이 통계에 포함되어 있어서 말뭉치에 들어 있는 실질적인 어절 수[7]는 <깜짝새>의 분석 결과보다 적다고 할 수 있다.[8] 이러한 분석 기능의 차이는 결국 전체 어절수 뿐만 아니라, 각 어절별 빈도, 텍스트 점유율 등에서 전반적인 차이를 가져온다.

---

7) 실질적 어절 수란, 문장 부호나 기호가 포함된 어절을 사람이 인식하는 대로 계산했을 때의 어절 수를 가리킨다.

8) 여기서 구체적으로 검토하지 않았지만, 숫자와 영문자가 포함된 어절의 처리 방식에 따라서도 어절 총수는 크게 달라지는 것으로 보인다. 이 문제에 대해서도 검토할 필요가 있는 것이지만, 이 글의 본래의 목적에서 크게 벗어나는 것으로 보여 상론하지 않는다.