기초어휘 선정 방법론

임칠성
전남대학교 국어교육학과

1. 서론

어휘를 계량하는 일은 어휘의 내적 구조를 통계학적으로 구명하고자 하는 목적 이외에 기본어휘의 설정, 기초어휘의 설정, 언어 정책의 기초 자료 확보와 같은 실용적인 목적 때문에 이루어진다. 이 때문에 우리나라에서는 1956년 문교부 주관으로 교과서, 신문, 잡지 등 다양한 종류의 기록문들을 대상으로 개별 어휘 56,485어, 운용 어휘[1] 2,218,792어를 대상으로 대대적인

1) 운용 어휘(運用 語彙 running word)란 연어휘(延語彙)를 대신하는 용어다. 이것은 어휘를 계량할 경우 계량 대상이 되는 언어 집단 내에 사용된 모든 어휘를 가리키는 말로서, 대체로 대상 언어 집단의 크기를 가리킨다. 운용어휘는 언어 집단 내에 사용된 어휘들 가운데 의미적 변별성을 가진 어휘들을 가리키는 개별어휘와 대립되는 개념이다.

이를 특별히 운용 어휘라 부르는 것은 일단 '연어휘'라는 용어가 일본 한자를 그래도 옮겨 온 것으로서, 우리의 인식 체계에 쉽사리 자리잡기 힘들기 때문이다. 운용어휘라 부르는 것은 이것이 가리키는 어휘들이란 실제로 운용이 된 어휘들이고, 또 이들 어휘가 어느 정도의 비율로 운용되었는가를 나타내주는 지표를 '운용 비율'이라고 부르기 때문이다. 개별 어휘에 대한 일본 한자말은 異語다. 異語라는 용어는 개별 어휘로 자리매김을 하였다고 본다.

어휘 계량 사업이 있은 후 크고 작은 어휘 계량 연구들이 끊임없이 있어 왔다.[2] 특히 최근에는 수백만 단위의 어휘 코퍼스를 구축하고 이를 기초로 어휘 빈도수에 기초한 사전을 편찬하는 등 여러 가지 사업이 진행되고 있다. 북한에서도 1993년에 개별 어휘 39,389어, 운용 어휘 1,047,375어를 대상으로 한 그간의 어휘 조사를 「조선어빈도수사전」으로 정리하여 출판하였다.

이충우(1994)에서는 계량으로 추출할 수 있는 어휘 집단을 기초어휘, 기본어휘, 학습용 기본어휘, 교육용 어휘라고 분류하고 있다. 여기에서는 기초어휘란 '사용 빈도보다는 한정된 소수의 어휘 자료[어휘소]에 의해서 가장 기본이 되는 일상 생활 각 영역에서의 필요가 충족될 수 있도록 계획적으로 선정된 것'(33쪽)이라고 정의하고 있다. 기초어휘는 '사회적 격변 영향이 적고, 차용어의 침투도 적으며, 장시간 지나도 잔존 가능성이 크다. 따라서 동계 언어 연구에 이용되며 언어 연대학·어휘통계학적 연대 산정 등의 연구 분야에서 자료로 쓰인다(33-34쪽)'라고 지적하고 있다.

기본어휘에 대해서는 '일상 생활에서 가장 일반적으로 사용하고 사용 빈도가 높은 어휘 가운데는 모든 사람에게 공통되는 것이 상당수 있다. 이 공통 어휘 중, 그 사회 구성원으로서 정상적인 기본 생활을 하는 데 필요하다고 간주되는 것을 말한다.(34쪽)'고 지적하고 있다.

다른 측면에서 김광해(1993 : 55)에서는 기본어휘라는 개념을 두 가지로 나누어 살피고 있다. 첫째는 넓은 개념으로서, 계량 대상 언어 집단을 각 성격에 따라 몇 개의 무리로 나눌 때 각 무리에 공통으로 출현하는 어휘들의 집합을 의미한다. 예컨대, 잡지라면 실용 기사, 문예 작품, 취미 등 게재되는 내용별로 일종의 층을 형성하거나, 또는 작품별로 층을 가상하는 일이 가능한데 이러한 경우 여러 층에 걸쳐서 공통적으로 출현하는 어휘의 집합이 기본어휘라는 개념이다.

둘째는 좁은 개념으로서, 언어 사용의 국면이 다양한 여러 영역으로 분

2) 우리나라의 어휘 계량 연구의 역사에 대해서는 임칠성·水野俊平·北山一雄(1997)에 정리되어 있다.

리될 수 있다는 것을 전제로 하여 특정 영역의 전개를 위하여 가장 기본이 되는 어휘의 집합을 가리키는 개념이다. 이러한 경우의 기본어휘란 특정한 목적, 특정한 분야를 위한 '○○ 기본어휘'라는 식의 표현이 가능하다. 예컨대, '생활 기본어휘', '학습 기본어휘'처럼 사용될 수 있거나, 나아가서는 '국민학교 교육을 위한 기본어휘', '중학교 수학 교육을 위한 기본어휘' 등처럼 '분야별 기본어휘'라는 개념으로 사용될 수 있는 것이다. 우리나라에서는 이 기본어휘라는 용어가 주로 학습용 기본어휘 목록을 조사하기 위한 작업에서 사용되었다.

이렇게 보면 기초어휘는 어휘의 체계의 바탕[素]이 되는 어휘들을 말한다고 할 수 있다. 기본어휘란 특정 집단이나 성격과 관련한 어휘들을 의미한다. 따라서 기초어휘는 어휘들의 빈도와 함께 어휘들의 체계 분석 및 그 체계를 이루는 바탕이 연구의 대상이 된다. 그렇지만 기본어휘는 대체로 빈도수의 계량에 의존하게 된다.

이 글에서는 기본적으로 기본어휘의 선정 방법에 대하여 살피고자 한다. 그렇지만 여기에 소개된 방법은 그대로 기초어휘 선정의 기본 작업이 될 것이다. 이 글에서는 특히 필자의 계량 경험을 바탕으로 언급할 수 있는 실제적인 계량의 방법을 소개하고자 한다.

이 글은 크게 두 가지 문제를 다룬다. 하나는 어휘를 계량하기 위한 단위화 문제이고, 다른 하나는 실제로 계량하기 위한 방법의 문제이다. 후자는 다시 자료가 방대할 경우와 자료가 적을 경우(대체로 10만 어휘 이하)로 나누어 각각의 경우 계량 결과 추출 방법을 다루고자 한다.[3]

3) 어휘 계량에서 우선 문제가 되는 것은 계량 대상의 선정이다. 모든 자료를 계량할 것인가 아니면 일부를 표본 추출하여 계량할 것인가에 관한 것이다. 대체로 자료가 방대할 때는 표본 추출하여 계량한 결과가 매우 신빙성이 높다. 이런 경우 어떤 대상에 어느 정도의 비율로 표본을 추출할 것인가 등의 문제가 생기게 된다. 그렇지만 본고에서는 표본 추출에 관한 내용은 다루지 못한다.

2. 어휘 자료의 계량 단위화[4)]

어떤 대상이든지 그 대상을 통계학적으로 처리하기 위해서는 기초 자료를 어떻게 처리하느냐가 가장 기본적인 문제가 된다. 기초 자료를 처리하는 방법에 따라 통계의 결과는 달라지기 때문이다. 예를 들어, 지금까지 어휘 계량을 보면 조사를 대상에 포함하기도 하고 더러는 빼기도 했는데 이는 결과에 엄청난 차이를 가져온다. 조사는 고빈도어에 속하기 때문이다. 따라서 어휘를 통계학적으로 처리하기 위해서는 무엇보다 어휘 계량의 단위를 어떻게 설정하느냐가 가장 기본적인 작업이 된다. 그러나 우리나라 어휘 계량의 경우 대개의 연구들은 어휘 계량의 단위를 설정한 방식에 대한 명확한 설명 없이 조사 결과를 내놓은 상태로서 아직껏 이런 문제에 대해 이렇다 할 합의가 이루어지지 않은 상태이다.

이런 실정에서 국어의 어휘 계량 단위 설정은 국가적인 계량 결과의 통합, 계량 결과간의 비교, 계량를 위한 문장의 계량 단위화 프로그램의 기초라는 측면에서 의의를 가지게 된다. 특히 세 번째 문제는 국어 계량의 발목을 잡고 있는 부분이다. 지금까지 여러 가지 계량 프로그램이 있지만 만족할 정도로 계량 단위화 할 수 있는 프로그램은 없기 때문이다.

이 글에서는 수 차례에 걸친 어휘 계량의 경험을 바탕으로, 우리 어휘 계량의 기본 단위를 설정해야 하는가에 대해 정리하고자 한다.[5)] 특히 모든 사항을 다 다루지 않고 필자가 계량을 하면서 처리하기 힘들었었던 사항들을 중심으로 다루고자 한다.

국어 어휘 계량 단위 설정의 원칙은 다음 두 가지이다.

---

4) 이에 대한 구체적인 내용은 임칠성・水野俊平・北山一雄(1997)을 참조..

5) 계량 단위 설정은 계량의 목적이 무엇이냐에 따라 달라질 수밖에 없다. 조사(助詞)의 쓰임을 연구하기 위해서는 당연히 조사를 별도의 계량 단위로 삼아야 한다. 이런 경우는 '-에서부터, -로까지'와 같은 복합 조사의 처리나, '-하고, -라며, -요'와 같은 문장 조사들의 처리가 문제가 될 것이다. 낱말 파생의 양상을 조사하기 위해서는 접두사와 접미사의 처리, 복합어나 파생어의 처리와 같은 것들이 문제가 될 수 있다.

첫째, 띄어쓰기를 단위로 한다.
둘째, 기본형을 단위로 한다.

2.1. 띄어쓰기의 원칙

띄어쓰기를 단위의 원칙으로 삼는 것은 띄어쓰기가 인식의 기본 단위를 의미한다고 볼 수 있기 때문이다. 한글맞춤법 총칙의 제2항에서 '문장의 각 단어는 띄어 씀을 원칙으로 한다'고 규정하고 있으니 띄어쓰기 원칙이라는 것은 각 단어를 계량의 단위로 삼는다는 것을 의미한다. 여기서 '단어'의 규정이 무엇이냐는 문제가 될 수 있지만 단어가 인식의 기본 단위를 가리키는 것으로 보아도 무리가 없을 것이다.

1) 보조 용언

현행 맞춤법에 의하면 '꺼져 간다, 올 듯하다'와 같은 경우는 '꺼져간다, 올듯하다'와 같이 붙여 쓸 수 있다. 그러나 계량의 단위로 설정할 때는 보조 용언은 띄어 씀을 원칙으로 한 맞춤법 제47항의 규정에 따라 띄어 쓴 단위를 계량의 단위로 삼는다.

다만 본 용언과 보조 용언이 합한 단위가 이미 한 낱말로 굳어져 하나의 단위로 인식되는 경우, 즉 합성 동사로서 띄어 쓰는 것이 원칙이 될 수 없는 경우에는 붙여 쓴 경우를 그대로 한 단위로 인식한다.

2) 고유명사와 전문 용어

한글 맞춤법 제48항에서는 '성과 이름, 성과 호 등은 붙여 쓰고, 이에 덧붙는 호칭어, 관직명 등은 띄어 쓴다'고 되어 있다. 따라서 성과 이름, 성과 호 등은 한 단위로 인식하되 그 뒤에 덧붙는 호칭어, 관직명은 별개의 단위로 인식한다. 예 박동식 박사－박동식+박사. 그런데 제48항 다만 규정은 성과 이름, 성과 호를 분명히 구분할 필요가 있을 경우에는 띄어 쓸 수 있다고 규정하고 있다. 이 경우에는 인식의 단위가 다르기 때문에 띄어 썼다기

보다는 성을 이름과 혼동하지 않도록 하기 위해 띄어 쓴 것이므로 성과 이름을 붙여 쓴 단위를 한 단위로 인식한다. 예 남궁 억－남궁억, 황보 지봉－황보지봉.

이름이나 성 뒤에 붙는 호칭어가 '선생, 박사, 계장' 등의 경우에는 호칭어를 별개의 단위로 인식한다. 그러나 이름이나 성 뒤에 '씨, 님, 군, 양'과 같은 유형이 붙을 경우에는 띄어쓰기 원칙에 따라 두 가지 경우를 달리 한다. 이름 뒤에 이들이 붙은 경우에는 이들을 별개의 단위로 인식한다. 그렇지만 '김씨'가 김씨 가문을 의미하는 경우를 제외하고는 모두 별개의 단위로 인식한다.[6]

성 뒤에 호가 붙은 경우에도 성 뒤에 이름이 붙은 경우와 마찬가지로 취급한다. 그러나 이름 앞에 호가 붙은 경우에는 호와 이름을 별개의 단위로 인식한다. 예 정송강(鄭松江)－정송강, 송강 정철－송강+정철

맞춤법 49항에서는 '성명 이외의 고유 명사는 단어별로 띄어 씀을 원칙으로 하되, 단위별로 띄어 쓸 수 있다'고 하고 있다. 이들은 허용 규칙에 따르지 않고 원칙에 따라 단위별로 한 단위로 인식하지 않고 단어별로 한 어휘로 인식한다. '한국대학교'의 경우 '한국'과 '대학교'는 서로 분리 가능한 단위들로서, '한국'은 고유 이름의 성격을 지니지만 뒤에 오는 '대학교'는 일반 명사의 성격을 지닌다. '창억 떡집'과 같이 일반 명사('떡집') 앞에 붙는 이름('창억')이 일반화된 의미가 아닐 경우에도 일반 명사와 그 앞에 붙는 이름을 별개의 단위로 인식한다.

고유 명사로 일컬어지는 대상물이 아니라 그 대상물의 존재 관계를 나타내는 '부설, 부속, 직속, 산하' 따위는 별개의 단위이지만 '부속학교, 부속

6) 졸고(1997)에서는 '김씨'의 '씨'를 접미사로 보아 '김씨'를 하나로 처리했으나, 국립국어연구원의 국어대사전에 '씨'를 의존명사로 처리하였기 때문에 이를 떼어 별개의 단위로 인식하게 되었다. '씨'의 경우처럼 사전에서 어떤 식으로 풀이하여 놓았는지는 단위 설정에 중요한 기초가 된다. 국립국어연구원의 국어대사전이 있기 전까지는 국가를 대신할 수 있는 사전이 없었기 때문에 다른 큰사전들을 이용했지만 이제 국립국어연구원의 국어대사전이 출판되었기 때문에 이 사전에 따라 몇 가지를 수정할 수밖에 없게 되었다.

국민학교' 따위는 대상물 자체를 나타내므로 붙여 쓸 수 있다. 그러나 이것들도 '부속'과 '학교'가 서로 다른 단위를 나타내며 띄어 쓸 수 있다는 점에서 '한국 대학교'를 '한국'과 '대학교'의 두 가지로 나누어 인식하는 것과 같이 '부속'과 '학교' 두 단위로 인식한다.

이는 맞춤법 제50항에서는 전문 용어에 대해서도 마찬가지이다. 예 만국 음성 기호－만국+음성+기호, 모음 조화－모음+조화.

이런 인식은 일상적인 개념으로도 그 뜻을 충분히 전달할 수 있는 경우에 이들을 굳이 전문 용어로 구분하여 전문어의 범주에 포함시킨 예들로 인한 혼란이나, 어느 정도까지를 이른바 '전문' 용어로 보아야 하는지에 대한 문제를 해결해 준다. 예를 들어, 국어연구소에서 발행한 한글맞춤법 해설서에 전문 용어의 예로 나와 있는 '도면그리기, 여름채소가꾸기, 감자찌기, 기구만들기'와 같은 경우는 전문 용어로 보기보다 그냥 일상어로 보아 '도면 그리기, 여름채소 가꾸기, 감자 찌기, 기구 만들기'로 처리하는 편이 좋다는 의견이 제시될 수 있는 것들이다.[7]

3) 관용어

'남의집살이, 처남의댁'과 같이 한 단위로 붙여 쓰는 관용어들은 한 단위로 인식한다. 한 단위로 붙여 쓰는 것들은 본래 구나 문장의 형태이지만 이것이 한 단위로 굳어져 인식되는 것들이기 때문이다. 그러나 '낫 놓고 기역자'와 같이 한 단위로 붙여 쓰는 것이 아니면 한 단위로 인식하지 않는다.

'여보세요, 맛있다'와 같이 구나 문장 단위 형태이지만 굳어져 관용적으로 한 단위로 쓰이는 표현들은 그대로 인정하여 한 단위로 본다. 이것들은 사전에 그대로 표제어로 등록되어 있으며, 띄어 쓸 경우 다른 의미를 가지

7) 다만 고유 명사의 사용 예를 살펴보고 싶으면 이들 고유 명사들을 두 가지로 분류하여 각각 그 결과를 도출할 수 있다. 다시 말하여, 위에서 말한 원칙을 따르는 한 가지와 고유명사를 붙여 써서 한 단위로 계량 단위화하는 방법 한 가지이다. 물론 이 두 가지 중 한 가지는 계량 결과에서 제외하여야 하며, 어떤 방법을 사용하였는지 연구 결과서에 분명하게 밝혀야 한다.

게 된다고 인정되어 띄어 쓸 수 없도록 된 말들이기 때문이다. '여보세요'는 기본형을 굳이 설정하자면 '*여보다'가 된다. 그러나 실제로 '여보다'라는 기본형은 존재하지 않는다. 따라서 '여보세요'를 그 자체로 한 단위로 인식한다. 간투사의 기능을 하는 '있잖아'도 마찬가지다. 그렇지만 '여보게, 여봐'와 같은 표현은 '여보세요'에 통합하여 합산한다. '*여보다'라는 기본형이 존재하지 않지만 '여보게', '여봐'는 '여보세요'와 높임상의 차이만 보이는데 이 차이가 어미의 활용에 의한 것이기 때문이다.

2.2. 기본형의 원칙

기본형이란 대표성을 가질 수 있는 바탕이 되는 형으로서 대체로 형태상 기본이 되는 꼴을 의미하는 것이지 가장 많이 쓰이는 꼴을 의미하는 것은 아니다. 기본형이란 용언의 경우 활용의 기본형을 의미하고, 준말과 본디말의 경우 본디말을 의미한다.

1) 용언

용언은 여러 가지 활용형을 갖는다. 이들은 각기 그 활용형의 기본형을 계량의 단위로 삼는다. 용언의 기본형은 거의 '어간 + -다'의 꼴을 취하지만 '먹어 달라고'의 '달라'와 같은 경우는 '달다'가 아니라 '달라'가 기본 꼴이기 때문에 그대로 '달라'를 계량의 단위로 삼는다.

활용의 기본형을 계량의 단위로 삼을 경우 가장 문제가 되는 것은 '그러나, 그러니까, 그러다가, 그런, 그래서, 그리고' 등과 같은 '그러(리)하다'류의 경우이다. '이런, 이러다가, 이렇게, 이래서' 등의 '이러(리)하다'류, '저런, 저렇게, 저러다가, 저래서' 등의 '저러(리)하다'류도 마찬가지다. 이런 유형들은 대체로 활용형 그대로를 계량의 단위로 삼고 있다. 이들을 사전에서 살펴보면 일부를 제외하고는 대부분 다른 용언들의 경우와 달리 활용형들이 표제어로 등재되어 있다. 따라서 이들 중 어느 정도까지를 활용형이 대표형으로 굳어진 꼴로 인정하느냐가 문제된다. 그 정도에 있어 각 사전에서

도 차이를 보이고 있다. 국립국어대사전에서 기본형으로 올려진 표제어는 기본형을 그대로 단위로 삼고, 만약 어떤 말의 준말로 올려진 표제어가 있으면 준말로 올려진 것은 본디말로 고쳐 그것을 다시 기본형화하여 단위화한다.

'많다'와 '많이', '깨끗하다'와 '깨끗이' 등도 이와 유사한 문제를 보인다. 많은 경우 용언 어간에 부사형 어미 '이, 히'가 붙어 된 말이 대표형으로 굳어져 있기 때문이다. 이 경우에도 정도의 문제가 발생하는데 결국 그 해답은 사전에서 표제어로 취급하고 있느냐 아니냐에서 찾아야 할 것이다.

'먹다'와 '먹이다'와 같은 사동 관계의 형태와 '잡다'와 '잡히다'와 같은 피동 관계의 형태는 사동형이나 피동형이 그대로 사전의 표제어로 처리되는 대표형이므로 개별 단위로 인식한다.

'-어지다'의 꼴은 '-지다'가 본래 보조 용언의 모습을 취했을 것이지만 지금은 '-어지다'로 융합되어 하나의 어미로 처리되기 때문에 '지다'를 개별 단위로 인식하지 않는다.

2) 의미 없는 소리

실제 조사를 하다 보면 대표형을 정할 수 없는 경우들이 있다. 한글 자모음을 비롯한 기호들과 순서를 나타내는 아라비아 숫자 등과 소리나는 대로 적은 표현들(예 '밭에'를 소리나는 대로 적어 '바테'라고 한 경우)처럼 의미 없이 쓰인 표현들이 그것이다. 이것들은 모두 조사의 대상에서 제외했던 것이 관례다.

그러나 1) 2) 3), ㄱ, ㄴ, ㄷ이나 a, b, c와 같이 순서를 나타내는 아라비아 숫자, 한글 자모, 영어 알파벳은 그것이 단순히 순서를 나타내기 위한 표시 이외에 내용상의 의미가 전혀 없으므로 조사 대상에서 제외한다. 그러나 기호가 내용의 일부로 들어간 경우는 이것을 계량의 단위로 삼는 것이 전체적인 국어 사용의 실태를 파악하는 데 효과적이라고 본다. 예를 들어 컴퓨터 통신에 관한 이야기라면 각 개인을 가리키는 ID들이 다수 등장할 것인데 그 ID들은 실제로는 사전에 등록되지도 않았고 아무런 의미를 가지지

못하는 것으로 분류되기 때문이다. 또 한글 자모에 대해 설명하기 위해 ㄱ, ㄴ, ㄷ과 같은 한글 자모를 열거한 경우 이들은 내용의 일부로서 국어 사용의 일단이라고 보아야 하기 때문이다.

3) 수사

대표형을 계량의 단위로 삼다 보면 수사가 문제된다. 예를 들어 '23'과 같은 표현이나 이것을 한글로 표현한 '이십삼', 혹은 '25,000'이나 '2만 5천'과 같은 경우가 문제될 수 있다. 수사에 관해서는 10 이하의 숫자만 선택하고 그 이상은 버리는 경우도 있다. 10 이상의 숫자는 일상적으로 많이 쓰이지 않는 것들로서 단지 숫자로서만 의미가 있기 때문일 것이다.

그러나 실제 계량을 하다 보면 이들 숫자가 상당히 많이 쓰임을 알 수 있다. 일상적인 대화에서도 이들 숫자는 많이 쓰인다. 특히 지금은 대상을 수량화함으로써 기호화하고(예 전화 번호, 시내버스 번호, 주민등록번호, 회원 번호 등) 있기 때문에 이들 숫자를 무시할 수 없다.

263-1538과 같은 전화번호는 그것이 국어 사용의 내용이기 때문에 어떻게든 계량에 포함되어야 한다. 이들은 하나의 운용 단위이므로 개별 어휘들이다. 다만 이들 숫자는 단지 하나 하나의 숫자들의 모음일 따름이기 때문에 263-1538의 경우 2, 6, 3, 1, 5, 3, 8이 모두 동등한 의미를 가지지만, 그것이 전화번호 한 단위로 인식되기 때문에 전화번호 전체를 하나의 단위로 인식해야 할 것이다. 이것은 주민등록번호나 회원 번호의 경우에도 마찬가지다. 그러나 이들은 '263'이라는 양적 의미를 가지는 숫자와는 구분되어 인식되어야 할 것이다.

'40자, 20개국, 30여 년'과 같은 경우는 '40, 20, 30여'와 '자, 개국, 년'을 개별 단위로 인식한다. 이 가운데 '여분'의 의미를 갖는 '여'는 '30'에 붙는 것으로서 뒤에 오는 의존명사 '년'과 성격이 다르다. 따라서 '여'는 '30'에 붙는 단위로 인식해야 한다. 이 경우 '30여'는 '30'과는 다른 단위다. 그러나 '제1과'와 같이 아라비아 숫자 앞에 '제'가 붙은 경우는 '제'가 접사이므로 '제1과' 전체를 하나의 단위로 인식한다.

‘40’과 ‘마흔’, ‘이’와 ‘둘’ 등은 구분하여 인식한다.

맞춤법 제44항에 의하면 수를 적을 적에는 ‘만(萬)’ 단위로 띄어 쓴다고 규정하여 놓고 있기 때문에 ‘345,673’은 ‘34만 5673’와 같이 띄어 써야 한다. 그러나 이것은 ‘황보 지붕’의 경우처럼 숫자 인식의 편의를 위한 장치이지 숫자 자체가 나누어 두 단위로 인식되는 것은 아니기 때문에 ‘345673’을 하나의 단위로 인식한다. ‘34만 5673’과 같이 띄어 씌어진 경우라도 이것은 하나의 단위로 인식한다.

4) 파생어

기본형을 계량의 단위로 삼을 때 파생어의 처리가 문제된다. ‘짓푸르다’는 ‘푸르다’를 기본형으로 볼 수 있기 때문이다. 그러나 대부분의 경우 접두사가 붙어 파생된 말들은 또 다른 의미 단위를 형성하며 이것들은 실제로 사전에 표제어로 등록되어 있다. 따라서 접두사가 붙어 파생된 말들은 모두 계량의 한 단위로 보아야 한다.

대부분의 접미사는 접두사의 경우와 마찬가지로 또 다른 의미 단위를 형성하기 때문에 ‘이쪽’과 ‘이’를, ‘철수’와 ‘철수네’를 각각 별개의 단위로 계량하여야 한다. 이런 기준에 비추어 현행 계량 단위에서 문제가 될 수 있는 것들은 ‘-님, -들, -적’과 같이 그것이 붙는 경우와 그렇지 못한 경우가 특별한 의미 차이를 보이지 않는다고 생각할 수 있는 경우들이다. ‘선생님’과 ‘선생’, ‘우리들’과 ‘우리’, ‘민주’와 ‘민주적’을 전혀 다른 별개의 단위로 인식하기에는 무리가 있다고 볼 수도 있다는 것이다. 사전에서도 ‘선생, 우리, 민주’는 표제어로 다루어지고 있지만 ‘선생님, 우리들, 민주적’은 표제어로 다루어지지 않고 있다. 그럼에도 불구하고 이들의 쓰임이 다르며, 다른 경우와 형평을 유지하기 위하여 이들을 각각 별개의 단위로 인식해야 한다.

이들을 별개의 단위로 인식하지 않을 경우 전체 접미사 처리의 형평성에 관한 문제가 제기될 수 있다. 예를 들어 ‘우리 인생살이’와 ‘우리네 인생살이’의 ‘우리’와 ‘우리네’는 ‘우리’와 ‘우리들’의 경우처럼 다른 의미 차이, 즉 지칭 대상의 차이나 지칭 대상에 대한 태도의 차이를 보이지 않는데도 불

구하고 '우리'와 '우리네'는 별개의 단위로 처리하는 한편 '우리'와 '우리들'은 '우리'라는 한 단위로 처리해야 하기 때문이다. 차라리 비록 그 의미 차이가 없을지라도 사용상의 선택으로 보아 '우리'와 '우리들'은 각각의 단위로 처리하는 것이 접사가 붙은 말은 별개의 단위로 처리한다는 원칙에도 일관성을 유지할 수 있게 해 준다. '님'의 경우는 더욱 그렇다. '선생'이라고 부르는 경우와 '선생님'이라고 부르는 경우는 대상에 대한 태도에서 엄연한 차이를 수반하게 되기 때문이다. 또, '아들, 딸, 아버지, 할아버지'라는 호칭이나 지칭과 '아드님, 따님, 아버님, 할아버님'이라는 호칭이나 지칭은 그 쓰임에서 상당한 차이를 수반하고 있다. 따라서 이들 모두는 각각 별개의 단위로 처리하는 것이 좋다. 더구나 이들이 갖는 차이를 인정하지 않게 되면 '밥'과 '진지', '있다'와 '계시다' 등도 동일하게 취급해야 하는 지경에 이르게 된다. '-적'의 문제도 이러한 원칙에 따라 '-적'이 붙은 어휘는 그것이 붙지 않은 어휘와 별개의 단위로 다룬다.

부사에 강조 접미사나 조사가 붙은 경우에도 각각 별개의 단위로 인식하여 계량의 단위로 삼는다. 예 너무나, 정말로, 자꾸만, 이따가.

5) 큰말과 작은말

'졸졸'과 '줄줄', '촐랑촐랑'과 '출렁출렁'은 같은 말이면서 어감의 차이만 나는 말들로 취급된다. 그러나 이들은 단순한 어감의 차이 외에도 쓰이는 유형이 다르다. 예를 들어 시냇물은 졸졸 흘러내리지만 콩자루에서 콩은 구멍을 통해 줄줄 흘러내린다. 대야의 물은 촐랑거리지만 바닷물은 출렁거린다. 아가는 아장아장 걷지만 어른은 어정어정 걷는다. 따라서 이들은 별개의 단위로 인식한다.

마찬가지로 의성어나 의태어의 경우 강조를 나타내기 위해 반복해 쓴 첩어는 별개의 단위로 인식한다. 예 하나하나, 움찔움찔, 흔들흔들

그러나 단음절로 된 의성어를 반복해 쓴 경우는 특별히 강조의 의미를 가지는 것이 아니기 때문에 단음절어이든('앙'), 3음절어이든('앙앙앙'), 혹은 그 이상이든('앙앙앙앙앙') 모두 이음절어로 합하여 산정한다. 예 하하하하—하

하, 흑흑흑흑－흑흑, 졸졸졸－졸졸

6) 준말

준말은 본디말을 계량의 단위로 삼는 것을 원칙으로 한다. 그러나 '좀'의 경우와 같이 준말이 이미 굳어져 본디말과 전혀 다른 의미를 가지거나 '막, 않는다'의 경우와 같이 본디말의 형태가 거의 쓰이지 않고 준말의 형태만 쓰이는 경우는 준말을 그대로 계량의 단위로 삼는다.[8] 이런 경우는 '점잖다, 편찮다'와 같은 경우도 마찬가지다.

요즘 들어 준말의 사용이 늘어나면서 '고삼, 노개위, 일고, 한투'와 같은 표현들이 등장했다. 이들은 '고등학교 삼학년, 노동법 개정 위원회, 제일고등학교, 한국투자신탁'의 준말들이다. 이들은 대부분 준말 이외에 다른 의미가 없으므로 본디말을 계량의 단위로 삼아야 한다고 볼 수도 있다. 그러나 앞서 지적했듯이 '한국 투자 신탁'의 준말인 '한투'를 본디말로 고쳐 인식할 경우 이것은 한 마디 말을 세 마디로 인식해야 한다는 문제점을 낳는다. 또 '고삼'과 같은 경우는 본디말을 계량의 단위로 삼을 경우 '고등학교 삼학년'이 되는데 '고삼'이 단지 '고등학교 삼학년'의 의미를 넘어 특별한 의미를 가진 일종의 관용어로 쓰인다. 이것은 '미나공'(미안해, 나 공주야)과 같은 경우도 마찬가지다. 따라서 이들은 본디말과는 별도로 한 단위로 인식하여 계량한다.

결국, 음운론적으로 명백하게 줄어진 말들 이외에 관용적으로 굳어진 형태만이 쓰이거나, 굳어진 형태가 본디말과는 다른 어감을 가질 수 있다거나 하는 모든 경우의 준말은 본디말과 별도로 한 단위로 산정한다.

7) 이름

고유 명사 가운데 '이철수, 두암동, 태양주식회사, 삼양라면'과 같은 사

8) 1) ㄱ. 좀 먹어라. ㄴ. 조금 먹어라
2) ㄱ. 막 도착하니까 ㄴ. ?마구 도착하니까('바로 도착하니까'의 의미일 경우)
3) ㄱ. 먹지 않는다. ㄴ. ?먹지 아니한다.

람의 이름, 지명, 단체명, 상품명과 같은 경우는 사전에 올라가지 않는다. 따라서 이들 고유명사들을 계량의 단위에서 제외한 경우들도 있다.

그러나 이들이 국어 사용에서 차지하는 비중은 상당히 크다. 특히 구어인 경우에는 그 비중이 무시할 수 없을 정도다. 따라서 이들은 그대로 한 단위로 계량한다. 다만 '삼양주식회사'와 같이 몇 개의 단위가 결합하여 한 낱말을 구성한 경우는 이들을 띄어쓰기 단위에 따라 각각 별개의 단위로 보아 '삼양, 주식, 회사'의 세 개로 나누어 인식한다.

8) 품사 통용

품사가 달라지면 기본 단위가 달라짐을 원칙으로 한다. 예들 들어, '밤낮'이 명사로 쓰인 경우와 부사로 쓰인 경우, '그'가 관형사로 쓰인 경우와 대명사로 쓰인 경우는 각기 하나의 단위로 인식한다.

그러나 용언에 전성 접미사인 '-음, -기'가 붙어 명사로 전성된 경우에는 의미론적으로 볼 때 용언의 활용과 별다른 차이를 보이지 않기 때문에, '꿈, 잠'과 같이 그것이 굳어져 별개의 단위로 인식되는 경우에만 별개의 단위로 인식하고 그렇지 못한 경우는 활용의 한 형태로 본다.

체언에 -하다, -되다, -스럽다, -직하다 등이 붙은 말은 하나의 단위로 인식한다. '-하다, -되다'는 거의 모든 체언에 붙을 수 있는데 명백한 용언으로서 체언과는 전혀 다른 의미를 가지기 때문이다. 예 공부하다, 참되다, 사랑스럽다, 널찍하다.

용언에 부사화 접미사가 붙어 있는 형식은 기본형을 기본 단위로 삼지만 부사화 접미사가 붙은 표현이 굳어져 하나의 독립적인 표현으로 인식될 때는 그것을 인정한다. 예 깨끗이, 빨리, 같이.

9) '내, 네, 제'

'내, 네, 제'가 주격으로 쓰이는 경우는 본시 통시적으로는 그 구성이 다르지만 공시적으로는 '나, 너, 저'와 음운론적 변이 관계에 불과하다고 볼 수 있다. 따라서 '네가'의 '네'와 '너는'의 '너'는 두 가지로 나누어 인식하는

것은 불합리하다. 주격으로 쓰인 '내, 네, 제'는 각각 '나, 너, 저'로 고쳐 인식해야 한다.

'내, 네, 제'가 소유격으로 쓰인 경우에도 이것들이 본래 '나의, 너의, 저의'에서 비롯된 것들일 뿐만 아니라 실제로 '내'와 '나의'는 아무런 의미적 변별성을 가지지 못한다. 따라서 이들도 조사 '의'를 제외한 '나, 너, 저'를 인식의 단위로 설정한다. 이 경우 '내, 네, 제'는 '나의, 너의, 저의'의 축약형으로 본다는 것이다.

3. 어휘의 계량 절차와 방법

3.1. 어휘의 규모가 방대할 때

우리나라의 경우 기본어휘를 정함에 있어 모든 연구가 지금까지 빈도수에만 의존해 왔다. 그러나 기본어휘는 빈도가 높아야 함은 물론 그 사용 폭도 넓어야 하며, 언중들의 인식도 고려되어야 한다. 그리고 기본어휘는 단순히 그 순위만을 정할 것이 아니라 기본의 정도를 수치화하여야 한다. 이런 점에서 일본 국립국어연구소(1962)의 현대 잡지 90종의 어휘 조사는 의의가 있다. 이 조사에서는 빈도수 외에 어휘의 편차, 언중의 기본 인식 정도를 포함시켰다. 다시 말하여, 빈도수를 바탕으로 하되, 여기에 편차와 언중의 기본 인식 정도를 수식화하여 이를 바탕으로 어휘의 기본도, 즉 기본의 정도를 수치화하였다.[9)]

일본의 국립국어연구소(1962)에 소개된 조사 방법을 필자가 텔레비전으로 방영된 드라마 '일출'의 대본을 대상으로 조사해 본 절차에 따라 하나씩 설명하겠다.

---

9) 실제로 수백만 어휘를 코퍼스로 구축한 연세한국어사전에 기본적인 어휘들이 많이 빠져있다는 지적을 받게 되는 것은 이 사전이 단순히 빈도수에만 의존했기 때문이다. 빈도수에만 의존하는 어휘 계량의 결과는 언중의 인식과 다른 경우가 많다.

1) 대본의 입력

계량할 자료를 입력한다. 「일출」의 대본 중 대사 부분만 제1회부터 제100회까지 입력한다. 이를 편의상 1차 자료라고 하겠다.

2) 계량 단위화

입력된 1차 자료를 대상 어휘 선정 기준에 따라 계량 대상 어휘로 고친다. 계량 단위화한 자료를 편의상 2차 자료라 하겠다. 계량 대상 어휘로 고칠 때 주의할 것은 태그를 다는 것이다. 필자는 품사별, 의미별로 구분할 필요가 있을 경우 태그를 달았다. 의미별 구분은 단위 어휘에 의미를 나타내는 한자어를 병기하는 방식을 택했다. 어떤 경우이든지 태그를 다는 방법은 일관성이 있어야 한다. 계량 프로그램에서는 태그까지를 포함하여 한 단위로 인식하기 때문이다.

계량 단위화는 항시 사전을 곁에 두고 작업을 하여야 한다. 의문이 갈 때마다 사전을 찾아 표제어 등재 여부, 기본형 여부, 복합어인지 아니면 복합 용언의 조합인지를 판단하여야 하기 때문이다. 국립국어연구원에서 출간된 국어대사전을 CD로 설치하여 작업하는 것이 편리하다.

계량에 미숙한 조사자라면 일정량을 계량 단위화하여 이를 계량 프로그램으로 돌려 결과를 살펴보면 잘못된 부분을 쉽게 찾아 고칠 수 있다. 계량 단위화하는 작업이 계량 연구에서 가장 많은 시간을 필요로 하고, 끈기를 필요로 한다. 따라서 처음 몇 차례의 시행착오를 통해 잘못된 부분을 잡아가는 것이 필요하다. 일단 2차 자료화하고 나서 나중에 계량 단위화에서의 잘못이 발견되면 난감한 경우가 많다. 예를 들어, '빨리'를 '빠르다'라고 처리하였는데 나중에 살펴보니 '빨리'가 부사로서 사전에 등재되어 있는 경우가 그런 예이다. 이럴 경우 1차 자료를 다시 살펴가며 다시 작업해야 하는 불편함이 있다. 또 일단 계량을 하다 보면 계량에 포함되지 않아야 할 어휘들이 포함되거나 아니면 어휘들이 쪼개져서 계량이 되기도 한다. 이는 조사자가 입력을 잘못해서 생긴 문제이거나 아니면 프로그램 자체의 문제일 경우도 있다. 시행착오를 통해 이러한 잘못을 바로잡아야 한다.

3) 2차 자료의 집단화

2차 자료를 다섯 집단으로 나눈다. 집단을 나누는 기준은 다섯 개의 집단의 운용 어휘 수가 비슷해야 한다는 것이다. 집단의 어휘 크기는 계량에서 매우 중요한 의미를 가진다. 각 집단의 크기가 달라지면 각 어휘가 그 속에서 차지하는 비율에 차이가 생기게 된다. 어떤 식으로 나누든지 각 집단의 전체 어휘 수가 비슷해야 한다. 필자는 100회 대본을 1회-20회, 21회-40회, 41회-60회, 61회-80회, 81회-100회의 다섯 무리로 나누었다.

4) 어휘의 계량

각 집단의 어휘를 계량 프로그램을 이용하여 계량하고, 그 결과를 데이터베이스화한다. 이렇게 하면 예를 들어 '가다'라는 어휘의 경우 무리별로 '247, 342, 235, 321, 299'와 같은 빈도수가 나오게 된다. 어휘를 데이터베이스화하여야 자신이 원하는 통계 값을 쉽게 추출할 수 있다. 예를 들어, 어휘 속에 '-하다'가 들어 있는 경우만 추출한다든지, 아니면 명사이면서 빈도수가 100회 이상인 경우만 추출한다든지 등 데이터베이스 프로그램의 연산식을 사용하여 많은 작업을 할 수 있다.

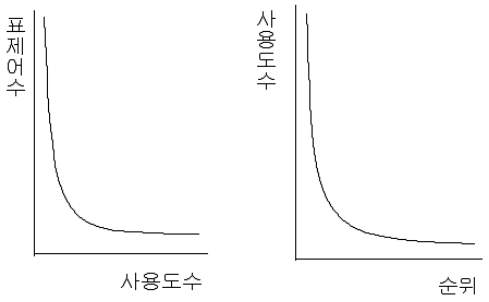
5) 어휘의 평균 사용률(p) 계산

무리별로 빈도수(운용 어휘 수)의 합을 해당 무리의 어휘의 총 운용 어휘 수로 나누어 각 어휘의 무리별 사용률을 구한다. 사용 빈도를 굳이 사용률로 고쳐 사용하는 데는 이유가 있다. 아무리 집단의 어휘 크기를 비슷하게 한다고 하더라도 이들이 어느 정도 차이를 가질 수밖에 없다. 따라서 20,000개 운용 어휘 수에서 한 어휘가 300번 출현했다(빈도수 300)는 경우와 20,500개 운용 어휘 수에서 그 어휘가 300번 출현했다(빈도수 300)는 것의 의미는 다르다. 즉 빈도수는 같더라도 그 빈도수가 해당 집단에서 차지하는 비중은 다르다는 것이다.

어휘의 크기에 따라 어휘의 양적 구조가 달라지기 때문에 계량의 결과가 달라진다는 사실은 매우 유의미함에도 불구하고 어휘 계량에 미숙한 조

사자들이 간과하기 쉬운 부분이다. 어휘를 계량해 보면 조사 대상 어휘의 크기가 크면 클수록 저빈도 어휘의 수가 매우 방대해지고 또 극히 소수의 고빈도 어휘가 차지하는 비중이 높아진다. 다시 말하여, 아주 많이 쓰이는 어휘의 수는 상대적으로 상당히 적어지고 한두 번밖에 쓰이지 않는 어휘의 수가 상대적으로 매우 많아진다는 것이다. 사용도수와 표제어 수 및 순위의 관계를 개념적으로 그래프로 나타내 보이면 다음과 같이 L형 분포(L sharped distribution)를 형성한다.[10]

표제어수, 사용도수, 순위의 관계



10) 이러한 사용도수(또는 사용률)의 분포는 어느 나라에서 어떤 자료를 대상으로 하든지 모두 같은 양상을 보이기 때문에, 이러한 현상에 어떤 함수가 존재하리라고 예상할 수 있다. 실제로 그런 함수가 존재한다면, 기본어휘의 양을 합리적으로 선정하는 데 큰 도움이 된다. 기본어휘 선정에 있어서 기본어휘를 어느 정도까지 설정해야 하는가에 관한 객관적인 기준이 되기 때문이다. 단어 사용률의 보편적인 분포 상황을 밝히고, 이 보편적인 분포 상황에 내재된 함수를 밝혀내면, 실제로 단어 사용도수 분포를 일일이 조사하지 않아도 이론적으로 사용도수 분포를 그려서 그것을 바탕으로 기본어휘의 양을 결정할 수 있다. 따라서 사용률 분포 상황을 잘 나타내주는 함수식을 밝히려는 연구가 20세기 초기부터 있었다. 이 연구는 언어학자·통계학자·심리학자들의 많은 관심을 모았으며, 계량어휘론에서 별도의 연구 영역으로 자리잡을 만큼 많은 연구가 이루어졌지만 불행하게도 아직까지 이렇다 할 해결에 이르지 못한 문제이기도 하다. Zipf의 법칙이나 Waring분포를 이용한 시도가 그 대표적인 예이다.

따라서 각 어휘가 차지하는 비중을 일정하게 해야 하기 때문에 사용 빈도 대신에 각 어휘의 빈도수를 집단의 운용 어휘 수로 나눈 사용률을 구하여, 이를 빈도수 대신에 사용하게 된다. 사실 개별 어휘의 빈도수는 조사 대상이 되는 전체 어휘의 크기 속에서만 의미가 있다. 그 외에는 아무런 의미가 없다. 조사 대상 어휘의 크기에 따라 빈도수의 값은 달라질 수밖에 없다. 중요한 것은 그 어휘가 전체 어휘 크기 속에서 어느 정도의 비중으로 사용되느냐, 거칠게 표현하면 어느 정도의 기본어휘로 인식될 수 있느냐 하는 것이다.

대체로 필자의 총 운용 어휘 수는 105,252개이었다. 다음과 같이 사용률에 1000을 곱하면 대체로 사용 빈도와 비슷한 수를 구할 수 있었다.

개별 어휘의 사용률(p)=(개별 어휘의 빈도수÷집단의 운용 어휘 수) ×1,000

다섯 집단에서 개별 어휘의 빈도수에 대한 사용률이 구해지면 이들의 평균을 구하여 평균 사용률을 구하게 된다. 결국 해당 어휘가 전체 조사 대상 어휘 속에서 몇 번 사용되었는가 하는 것이 문제가 되는 것이 아니라 그 어휘 속에서 어느 정도의 비율로(비중으로) 사용되었는가 하는 것이 중요하다.

6) 개별 어휘의 표준편차(sc) 계산

다섯 가지 사용률의 표준편차를 구한다. 개별 어휘가 다섯 집단에서 어느 정도의 비중으로 사용되었는지를 알기 위한 것이다. 표준편차 부분은 어휘 계량에서 매우 중요한 의미를 가진다. 구체적인 내용은 소규모 어휘의 계량에서 설명하겠다.

다음 이 표준편차를 이론적 상한치인 최고 표준편차로 나누어 이것을 각 어휘의 표준편차로 사용한다(이하 표준편차란 이렇게 구해진 표준편차의 값을 가리킨다). 이론적 상한치인 최고 표준편차란 운용 어휘 수가 가장

적은 무리 한 곳에 해당 어휘가 모두 출현하고 나머지에는 전혀 출현하지 않은 경우의 표준편차로 '가다'의 경우 1444, 0, 0, 0, 0일 경우의 표준편차를 의미한다. 다시 말해, 개별 어휘 사용률의 표준편차가 그 어휘가 한 곳에서만 전부 사용되었을 때의 표준편차값에서 어느 정도의 거리 비율을 가지는지 판단하는 것이다.

개별 어휘의 표준편차(sc)=개별 어휘의 사용률의 표준편차÷이론적인 최고 표준편차

7) 언중의 인지도(z′) 조사

어휘를 계량하고 보면 우리의 직관에 의한 기본도와 출현 빈도에 의한 기본도가 아주 다르다는 것을 알게 된다. 따라서 언중의 직관은 상당한 의의를 가진다. 다시 말하여, 언중이 어느 어휘가 더 기본적인 어휘라고 인식하는가 하는 언중의 인지도를 수식에 포함하여 기본도를 계산할 필요가 있다는 것이다. 언중의 직관적인 인지도를 기본도에 포함한 점이 일본의 어휘 계량의 장점이다.

언중의 인지도 조사는 다음과 같은 방법으로 실행한다.

- 조사 대상 어휘를 표준편차별, 사용률별로 각 5단계씩 나누어 25개 무리로 나눈다.
- 25개 무리 각각에서 임의로 선정한 5개의 어휘로 한 단위씩 25개의 어휘군을 이룬 뒤 이것을 2개가 한 쌍이 되도록 300개의 짝을 이루어 언중이 느끼는 인지도를 측정한다.
- 언중에게 설문을 하여 2개 쌍 중 인지도가 높다고 생각하는 단어에 표를 하게 한다. 2개 쌍 중 선택된 쌍 하나에 1점씩을 부과하여 25개 무리의 기본도 의식치(z′)를 구한다.

8) 상수(常數)의 계산

이제 25개 무리를 한 단위씩으로 하여 각 무리의 평균 사용률(x), 평균 표준편차(y)를 구한다. 그리고 이렇게 해서 나온 x, y와 (7)에서 구한 기본도 의식치 z의 값을 다시 각각 log x + 5, log y + 3, z × 0.01의 식을 통해 x′, y′, z′의 값을 구하여 이를 다음의 방정식에 사용하여, 상수 a, b, c의 값을 구한다.

25a + [x′]b + [y′]c = [z′]
([x′]는 25개 x의 합)
[x′]a + [x′x′]b + [x′y′]c = [x′z′]
([x′x′], [x′y′]는 각각 x′과 x′의 곱의 25개의 합과 x′와 y′의 곱의 25개의 합)
[y′]a + [x′y′]b + [y′y′]c = [y′z′]
([y′y′], [y′z′]는 각각 y′과 y′의 곱의 25개의 합과 y′와 z′의 곱의 25개의 합)

9) 개별 어휘의 기본도(z) 계산

구해진 a, b, c의 값을 다음 식에 적용하여 각 어휘의 기본도를 구한다.

$$z = a + bx'' - cy''$$

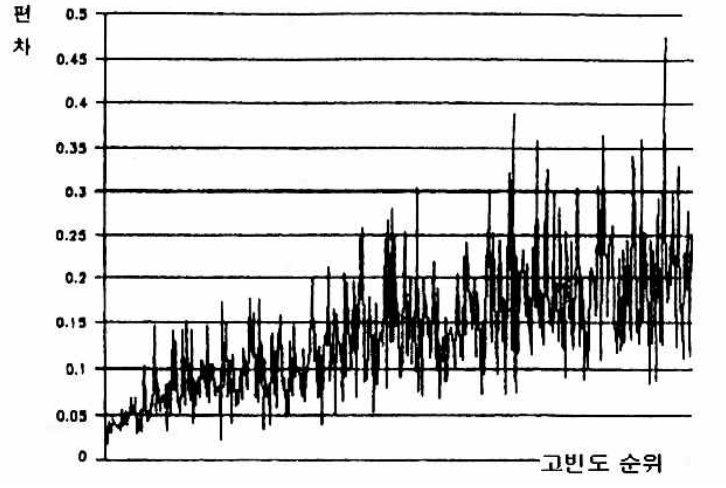
(x″ = log p + 5 : p는 각 어휘의 사용률,
y″ = log sc + 3 : sc는 각 어휘의 표준편차,
z는 각 어휘의 기본도)

3.2. 어휘의 규모가 적을 때 - 빈도수와 표준편차를 이용한 계산

개인이 어휘를 계량할 때는 어휘의 규모가 방대하면 여러 가지 어려움이 있다. 대체로 어휘 규모가 10만개 이하일 경우가 많다. 이런 경우 언중의

인지도를 측정하는 것은 사실 별다른 의미가 없다. 실제 작업을 해 보면 25개의 무리로 나누는 것 자체가 무리가 되기 때문이다.

그럼에도 불구하고 어휘 규모가 적을지라도 여전히 표준편차는 큰 의미를 가진다. 실제로 필자가 105,252개의 어휘를 계량하고서 일정한 빈도수 이상을 대상으로 빈도 순위와 표준편차와의 관계를 살펴보니 다음 그래프와 같았다.



이 그래프는 다음과 같은 사실을 설명해 준다. 즉, 빈도수의 차이에도 불구하고 표준편차의 차이가 심하여 결국 기본도가 바뀌어지는 경우이다.

어휘	빈도수	표준편차	기본도 순위
엄마	830	0.146	40
아니다	818	0.039	12
가다	1,043	0.149	4
그러하다	1,887	0.423	5

'가다'와 '그러하다'는 '엄마'와 '아니다'와 마찬가지로 각각 언중의 인지도에서 같은 수치를 받았으므로 언중의 인지도가 기본도 순위에 영향을 미치지는 않았다. 다시 말해, 기본도 순위를 결정한 것은 빈도수와 표준편차이다. 그런데 '아니다'는 '엄마'보다 빈도수가 낮음에도 불구하고 표준편차가 낮기 때문에, 즉 '엄마'보다는 상대적으로 골고루 분포되어 있었기 때문에 기본도 순위가 '엄마'보다 무려 38위나 높다. '가다'는 '그러하다'보다 빈도수에 있어서는 무려 843번이나 낮지만 표준편차가 상대적으로 낮기 때문에 기본도 순위가 높다. 역으로, '그러하다'는 1,887번이나 출현했음에도 불구하고 표준편차에서는 0.423이나 되는 것을 알 수 있다. 이러한 사실은 빈도수에만 의존해서는 기본도를 알 수 없다는 것이며, 기본어휘의 선정에 반드시 표준편차를 반영해야 한다는 것을 극명하게 보여주고 있다.

그러나 불행히도 통계를 통해서는 계량 연구에서 원하는 빈도수와 표준편차를 적용한 기본어휘 선정의 방법을 택할 수 없었다. 따라서 필자는 이를 이 두 가지 요소를 반영하여 기본도를 계산할 수 있는 방법을 나름대로 고안하였고, 그 수식을 바탕으로 한국어 초급 교육용 어휘를 선정한 바 있다. 여기에 소개하는 방법은 졸고(2003)에서 필자가 사용한 방법을 다시 수정 보완한 방법이다. 필자는 이 방법이 상당히 간단하지만 비교적 적절하게 사용될 수 있는 방법이라고 생각한다.

1) 졸고(2003)에서는 표준편차와 빈도수를 활용하여 기본의 정도인 기본도를 계산하기 위하여 몇 가지 방법을 사용하였다. 우선 단순히 표준편차와 분산을 이용한 수식은 이 논문에서 추구하고자 하는 결과를 가져다주지 못하였다.

기존의 통계 처리 수식 가운데 가장 매력적인 것은 '변동계수'였다. 변동계수는 표준화를 하는 데 사용하는 수식으로서, 각 항목의 빈도의 합을 표준편차로 나눈 값이다. 이것은 동일한 빈도의 합일 경우 편차가 크면 상대적으로 적은 값을 부여받도록 한 것이었다. 그러나 빈도의 값이 커지면

편차의 값도 상대적으로 커져서 원하는 값을 얻을 수 없었다. 여기서 원하는 값이란 빈도의 합과 표준편차를 직관적으로 분석한 값과 수식 결과의 값이 대체로 일치하는 값을 말한다.

궁리 끝에 편차의 값을 조정하기 위하여 편차의 제곱근(s′)을 사용하였으나 편차가 차지하는 비중이 너무 커서 원하는 값을 얻을 수 없었다. 따라서 s′를 다시 제곱근한 값을 사용하여 변동계수의 수식을 이용하였다. 이를 통해 상당히 원하는 값이 근접할 수는 있었지만 몇 개의 어휘에서 결정적으로 문제가 있었다. 따라서 이 식을 포기할 수밖에 없었다.

2) 원하는 값을 도출해 줄 수 있는 수식을 발견할 수 없었다.[11] 따라서 새로운 수식을 만들어 사용하였다. 기본적인 개념은 각 항목의 빈도의 합과 분산의 정도를 함께 고려한 값을 구하는 것이었다. 각 항목 빈도의 합은 앞서 밝힌 평균 사용률을 의미한다. 그러나 여기에서 소개하는 방법은 그 방법을 수정 보완한 방법이다.

분산의 정도는 각 항목의 사용률이 그 항목의 평균 사용률에서 벌어진 거리의 비율로 계산한다. 각 집단별로 한 항목의 사용률이 전체 집단의 평균 사용률에서 어느 정도나 떨어져 있는가를 계산한 다음, 그것을 '1'에서 뺀 값을 그 항목의 사용률에 곱하여 해당 집단의 어휘 항목이 가지는 가치로 산정하는 것이다. 다시 말해, 평균 사용률에서 벌어진 거리의 비율에 해당하는 값을 본래 어휘의 사용률에서 제거하는 것이다.

이것은 어떤 어휘 항목의 나누어진 각 집단에서 완전하게 고르게 사용된 경우를 '1'로 계산하고, 그 평균 사용 비율에서 각 집단마다 떨어진 비율들을 '1' 이하로 계산하여, 떨어진 비율만큼 본래의 사용률을 감소시켜 기본

11) 원하는 값이란 집단별 편차가 고른 어휘 항목의 기본도가 편차가 고르지 않은 어휘 항목의 기본도보다 비례적으로 높아지는 값을 말한다. 수학, 통계학, 교육통계학을 전공한 교수들의 자문을 얻어 보았지만, 기존의 수식에서는 원하는 수식을 얻을 수 없었다.

도의 값을 구하는 것이다. 따라서 편차가 고를수록 본래의 평균 사용률에 가까워지고, 편차가 고르지 않을수록 본래의 평균 사용률에서 점점 멀어지도록 하는 것이다.

이것을 구하는 식은 다음과 같다.

RD(Ak)= ABS(AVERk - Ak)/AVERk
RD(Ak) : A집단에서 k의 거리의 비율
ABS : 절대값
AVERk : 전체 집단에서 k의 평균 사용률
Ak : A집단에서 k의 사용률

이런 과정을 통해 나온 RD(Ak)의 값을 '1'에서 뺀 값이 A집단에서 분산을 고려한 k의 사용률이 된다.

Ak′의 사용률 = 1 - RD(Ak)

이런 과정을 통해 나온 각 집단의 k의 사용률, 즉 Ak′, Bk′, Ck′…를 합하여 전체 k의 사용률을 구하고, 이것이 기본도 값이 된다.

3) 구체적인 예를 들어 이 과정을 다시 살펴보자. 전체 어휘 집단을 A, B, C 세 집단으로 나누어 2차 어휘를 계량하여 다음과 같은 결과를 얻었다고 하자.

어휘 항목	A	B	C	계	평균
가다	145	200	210	555	185
먹다	160	170	165	495	165
…	…	…	…	…	…

사용률로 예시를 제시하면 읽기 복잡하므로 편의상 빈도수로 사용률을 대신한다.[12] '가다'의 거리 비율을 구하는 과정은 다음과 같다.

A집단 |(185 - 145)/185| = 0.21
B집단 |(185 - 200)/185| = 0.08
C집단 |(185 - 210)/185| = 0.13
A(가다) = 145 × (1 - 0.21) = 114.55
B(가다) = 200 × (1 - 0.08) = 184.00
C(가다) = 210 × (1 - 0.13) = 182.70
'가다'의 사용률 111.55 + 184.00 + 182.70 = 478.25

이렇게 되면 '가다'는 본래 전체 빈도수가 555회임에도 불구하고, 고르지 못한 편차 때문에 실제로는 478.25회의 빈도 값을 가지게 되는 것이다. 편차가 고른 경우에는 본래의 빈도 값에 근사한 값을 갖게 된다. '먹다'의 경우를 예로 들어보자.

A집단 |(165 - 160)/165| = 0.03
B집단 |(165 - 170)/165| = 0.03
C집단 |(165 - 165)/165| = 0
A(먹다) = 160 × (1 - 0.03) = 155.20
B(먹다) = 170 × (1 - 0.03) = 164.90
C(먹다) = 165 × (1 - 0) = 165
'먹다'의 사용률 155.20 + 164.90 + 165 = 485.10

'먹다'는 비교적 편차가 고르기 때문에 본래의 빈도 값 495에 근사한

12) 그렇지만 실제 작업에서는 빈도수 대신 사용률을 사용해야 한다. 다만 나누어진 집단의 어휘 총수가 거의 비슷한 경우, 즉 세 집단의 크기가 어휘의 구조에 영향을 거의 미치지 않을 경우에는 빈도수를 그대로 사용할 수 있다.

485.10의 값을 얻게 된다.

이제 '가다'와 '먹다'의 경우를 비교해 보면 상당히 흥미로운 결과를 얻게 된다. '가다'가 '먹다'보다 출현 빈도에서 60회(555-495)나 높지만 '가다'는 편차가 매우 고르지 못하고 대신 '먹다'는 아주 고른 편차를 보이기 때문에 기본도에서 보자면 '먹다'가 '가다'보다 높은 순위를 받게 된다.

어휘 항목	A	B	C	계	평균	기본도
가다	145	200	210	555	185	478.25
먹다	160	170	165	495	165	485.10
…	…	…	…	…	…	…

이 식에서 보듯이 편차를 고려하면 한쪽에서 많이 출현하고 다른 한쪽에서는 적게 출현하는 어휘의 값이 모든 곳에서 고르게 출현하는 어휘의 값보다 낮아지는 경우가 많아진다. 어휘 계량을 하다 보면 어휘가 편중되어 출현하는 경우가 많을 수밖에 없다. 특히 어휘 규모가 적을수록 이러한 편중 현상은 심하다. 그렇지만 빈도수만 고려해가지고는 이러한 편중 현상을 처리할 수 있는 방법이 없다. 따라서 이런 경우 위의 식을 사용함으로써 어휘 편중을 적절하게 처리하여 올바른 빈도 순위를 얻을 수 있으리라 확신한다.

4. 결론

어휘 계량은 끈기를 요구하는 지난한 작업이다. 그럼에도 불구하고 국가 언어 정책의 수립이나 사전의 출판, 교육용 어휘 선정 등 매우 중요한 작업이 아닐 수 없다. 많은 연구자들이 여러 가지 방법으로 어휘를 계량하여 그 결과를 발표하고 있다. 조사 과정에서 여러 가지 잘못들이 발견될 수 있다. 그렇지만 이러한 많은 시행착오들이 오히려 연구를 촉진시킬 수 있는 촉매제가 되리라 확신한다.

참 고 문 헌

김광해(1993), 국어어휘론개설, 서울: 집문당.
이충우(1994), 한국어 교육용어휘 연구, 서울: 국학자료원.
임칠성・水野俊平・北山一雄(1997), 한국어 계량 연구, 광주: 전남대출판부.
임칠성(2002), 초급 한국어 교육용 어휘 선정 연구, 국어교육학연구 14집.
國語硏究所(1962), 現代 雜誌 九十種の 用語用字, 일본: 秀英出版.