어휘 조사의 전산 처리

박 민 규
(국어연구소 연구원)

Ⅰ. 머 리 말

어휘 조사의 전산 처리는 단어 구분을 어느 단계에서 하는 것이 효율적인가 하는 점이 관건이 된다. 단어를 구분하는 과정에 따라 두 가지 가능성을 생각할 수 있는데 첫째, 입력 과정에서 단어를 구분하는 방법과 둘째, 일단 입력하여 동일한 형태를 모은 후에 단어를 구분하는 방법이 그것이다.

첫째 방법은 조사하는 자료의 실사와 허사를 하나하나 구분한 후에 입력하여 실사를 기준으로 재배열해서 단어별로 정리하는 방법이다. 이러한 방법은 카드 작업과 그 성격이 거의 같다고 할 수 있다. 양적으로는 동일한 자료가 여러 번 반복되어 나오더라도 그 수만큼 각기 실사와 허사로 구분해 주어야 한다. 질적인 면 즉 단어 구분의 정확성에 있어서도 자료가 나열된 순서대로 하나씩 구분해야 하므로 동형어(homograph)를 구별하거나 실사와 허사를 구분하는 데 부정확하거나 일관성이 없을 수도 있다. 이러한 점들은 카드 작업의 경우에도 마찬가지로 문제가 된다. 그러나 전산 처리의 경우는 카드 작업과 동일한 방법으로 단어를 구분하더라도 일단 단어로 구분된 자료를 신속히 정확하게 정리하는 데는 월등히 효율적이다. 뿐만 아니라 카드 작업의 경우에 단어별로 정리한 후에는 확인 점검이 쉽지 않은 데 비해 전산 처리는 KWIC색인으로 용례를 정리하여 확인 점검을 쉽게 할 수 있다.[1]

둘째 방법은 일괄 처리 방법이다. 조사 자료를 원문대로 입력하고 조사

1) 福井玲의 '月印千江之曲 上 KWIC索引'이 이러한 경우라고 할 수 있다. 실사뿐만 아니라 허사의 용례까지 정리되어 있는데, 이는 실사와 허사를 하나하나 구분한 후에 입력하여 KWIC색인을 출력해 낸 것이다.

하고자 하는 단위로 끊어서 자모순으로 재배열(sort)한다. 이렇게 하면 동일한 형태의 자료들을 모두 모아서 단어를 구분할 수 있게 되므로, 일관성 있게 처리할 수 있을 뿐만 아니라 동일한 자료에 대한 반복 작업을 줄일 수 있게 된다.

이 글에서는 둘째 방법으로 추진한 국어연구소의 어휘 조사 과정을 살펴보고 첫째 방법을 보완한 한 가지 방안을 제시해 보기로 하겠다.

Ⅱ. 교과서 어휘 조사

국어연구소에서는 1984년에 산업연구원(KIET)의 컴퓨터를 이용하여 국민학교 전과목 교과서의 어휘를 조사하였고, 1985년에는 한국과학기술원(KAIST) 부설 시스템공학센터의 컴퓨터로 중학교 국어 교과서의 어휘 조사를 추진하였다. 두 차례 모두 일괄 처리 방식으로 조사하였으나 구체적인 전산 처리 과정은 차이가 있으므로 조사 과정을 비교하여 살펴보기로 하겠다. 앞으로 편의상 국민학교 전과목 교과서 어휘 조사를 '1차 조사'로, 중학교 국어 교과서 어휘 조사를 '2차 조사'로 줄여서 언급하겠다.

1. 조사 자료

국민학교 교과서는 13과목으로 70권인데 모두 1984년 판으로 조사하였다. 전산 처리시 부여된 과목별 코드는 다음과 같다.

우리들은 1학년(01), 바른 생활(02), 슬기로운 생활(03), 즐거운 생활(04), 국어(05), 사회(06), 도덕(07), 산수(08), 자연(09), 미술(10), 음악(11), 체육(12), 실과(13)

중학교 국어 교과서는 6권인데 역시 1984년 판으로 조사하였다.

2. 조사 단위

교과서의 어휘 조사는 교과서에서 띄어 쓴 어절을 기준으로 하였다.[2] 이러한 기준의 어휘 조사에는 몇 가지 문제점이 있다. 예를 들어 '국민 학교'는 교과서에 두 어절로 되어 있기 때문에 '국민'과 '학교'로 조사되고 '국민 학교'는 조사되지 않게 된다. 교과서에 두 어절로 띄어져 있다고 하더라도 어절 사이에 일정한 부호를 넣어서 입력하면 '국민@학교'도 쉽

2) 교과서의 띄어쓰기에 일관성이 없거나 문제가 있는 부분은 별도로 정리하였다. 국어연구소(1987), 국민 학교 교육용 어휘 —4, 5, 6 학년용—, pp. 406~426.

사리 조사할 수 있게 된다. 그러나 합성어나 파생어의 처리에 여러 문제가 있어서 어휘 조사를 일관성 있게 추진하는 데는 어려움이 있다. 이 문제는 국어연구소에서 별도의 연구 과제로 현재 추진 중이다.

둘 이상의 어절로 된 고유 명사나 전문 용어의 처리도 역시 문제가 된다. 1989년 3월부터 시행된 한글 맞춤법의 띄어쓰기 규정에 띄어 쓰는 것을 원칙으로 하고 붙여 써도 좋도록 허용된 규정이 여럿 있다. 앞으로의 어휘 조사는 이처럼 동일한 경우에 붙여 쓰기도 하고 띄어 쓰기도 한 자료들의 처리에 유의해야 할 것이다.[3]

교과서는 규범적인 성격의 자료이므로 원문대로 조사하였다. 그러나 신문 등과 같이 성격이 다른 자료를 조사할 때는 띄어쓰기를 비롯한 여러 가지 문제점을 적절히 처리하는 과정이 필요할 것이다.

3. 전산 처리 과정

1차 조사와 2차 조사의 전체적인 전산 처리 과정은 비슷한데 간략히 정리하면 다음과 같다.

i) 교과서 자료를 입력한다.
ii) 문장 단위로 끊어서 문장 화일을 만든다.
iii) 다시 어절 단위로 끊어서 어절 화일을 만든다.
iv) 각 어절의 실사와 허사가 구분되도록 처리하고, 어절의 문맥을 볼 수 있도록 문장 화일을 병합(merge)하여 출력한다.
v) 실사와 허사 구분이 잘못된 부분을 고치고 동형어를 구별한다.
vi) 실사를 기준으로 하여 자모순으로 재배열해서 빈도, 색인 등의 결과물을 정리해 낸다.

두 차례의 교과서 조사를 비교하면서 전산 처리 과정을 좀더 자세히 살펴보기로 하겠다.

(1) 입 력

조사 자료의 입력은 가능한 한 원문대로 하였다. 특히 많은 분량의 자료를 조사할 때는 여러 측면의 조사가 가능하도록 이 점에 유의하는 것이 바람직할 것이다. 1차 조사에서는 부호까지도 조사가 가능하도록 그대로 입력하였다. 이는 산업연구원(KIET)의 완성형 코드를 사용하였기 때문에

3) 일본의 국립국어연구소에서 추진한 교과서 어휘 조사의 경우에 長單位(W單位)와 短單位(M單位)의 두 가지 기준으로 조사되었다. 國立國語研究所(1983), 高校教科書の語彙調査, pp. 4~15.

다양한 부호의 구분이 가능하였다. 그러나 2차 조사에서는 IBM 1바이트 코드를 사용하였기 때문에 부호 사용에 제약을 받게 되었다.

교과서 자료의 입력 과정에서 원문에 두 가지 표시를 하였다. 문장이 아닌 자료의 시작과 끝을 '「'과 '」'로 표시하였고 문장의 계속을 나타내는 부호를 사용하였다. 이에 대한 자세한 설명은 뒤에서 하도록 하겠다.

(2) 문장 구분

어휘 조사에서 문장을 구분하는 과정을 반드시 거쳐야 하는 것은 아니다. 어휘의 빈도 조사만을 목적으로 한다면 바로 어절 화일을 작성하는 것이 좋을 것이다. 그러나 문장 용례나 문장에 대한 통계 처리가 필요하다면 문장 화일을 작성한 후에 어절 화일을 작성하는 것이 바람직할 것이다.

입력할 때 문장 끝에 일정한 부호를 표시하여 이 부호를 기준으로 문장을 구분하는 방법을 생각해 볼 수 있다. 이 방법은 자료를 입력할 때에 대부분 마침표[4] ——온점이나 고리점, 물음표, 느낌표——와 문장의 끝을 표시하는 부호가 중복된다는 점에서 많은 분량의 자료를 처리할 때에는 효율적인 방법이라고 할 수 없다.

문장을 구분하는 다른 방법은 마침표를 문장 구분의 표시로 간주하여 전산 처리하는 것이다. 이러한 방법으로 문장을 구분하려면 두 가지 경우를 보완해야 한다.

첫째는 문장인 자료와 문장이 아닌 자료 즉 단어나 구가 뒤섞여 있는 경우다. 문장이 아닌 자료가 시작되는 부분과 끝나는 부분을 '「'과 '」'로 표시해 문장과 구별하였다. 이렇게 구별하지 않으면 단어나 구로 된 자료가 문장에 연결되어 처리되고 만다.

둘째는 직접 인용문을 안은 문장이다. 예문 ①과 같은 경우가 있으므로, 인용문이 끝나고 따옴표 뒤에 바로 공백이 있으면 문장이 끝난 것으로 본다. '마침표+따옴표' 바로 뒤가 공백이 아니면 문장이 계속되는 것으로 보아서, 그 뒷부분에 마침표가 나와야만 문장이 끝난 것으로 본다. 이와 같은 방법으로 문장을 구분하면 ②와 같이 '라고'에 의해 직접 인용되는 경우에는 문제가 없지만 ③과 같이 인용문 다음에 '하고'가 오는 경우에는 띄어 쓰기 때문에 문제가 된다. ④와 같이 앞뒤에 따옴표를 붙인 인용문들이 나열되는 경우에도 그 사이에 공백이 있으므로 문제가 된다. 결국 ③과 ④에서 따옴표 뒤의 공백이 문제가 되는데 이 공백에 문장의 계속을

4) 문장 부호의 이름은 개정된 한글 맞춤법에 따른다.

표시하는 부호를 넣어 보완하였다. ⑤와 같은 경우에도 이러한 방법이 그대로 적용된다.

① '이웃 사촌이 친형제보다 낫다.'
② 어떤 청년이 "이 근처에 혹시 절이 있읍니까?"라고 물었다.
③ 스님께서 "너도 어제 큰절 구경을 했느냐?"—하고 물으신다.
④ '경제 개발 5개년 계획'—'우리 나라 경제는 크게 발전했다.'—'그렇게 하는 것이 경제적이다.'—등에서와 같이 경제라는 말은 나라 살림에서부터 개인의 살림살이에 이르기까지 두루 쓰는 말이다.
⑤ "토끼가 단식 투쟁을 하는 모양이다. '나에게 자유를 달라. 나를 엄마 곁으로 보내 달라.'—하고 말이다."라고 형이 웃으며 말하였다.

직접 인용문을 안은 문장은 길므로, 인용문에 하위 문장 번호를 붙여 분리하는 방법을 생각해 볼 수 있다. ⑤를 이와 같은 방법으로 분리하면 다음과 같이 된다.

001—000—000 "　"라고 형이 웃으며 말하였다.
001—001—001 토끼가 단식 투쟁을 하는 모양이다.
001—002—000 '　'—하고 말이다.
001—002—001 나에게 자유를 달라.
001—002—002 나를 엄마 곁으로 보내 달라.

이처럼 직접 인용문을 분리하여 문장 화일을 작성하면 문장 레코드의 길이가 짧아져서 KWIC 색인을 할 때 효율적이 될 것이다. 뿐만 아니라 문장의 길이를 계산할 때 문장 전체의 길이와 인용문의 길이를 구분하여 확인할 수 있게 된다.

조사하는 자료에 쓰인 문장 부호를 기준으로 하여 문장을 구분할 때 주의해야 할 점이 있다. 마침표 중 온점은 문장의 끝에 쓰일 뿐만 아니라, ⑥처럼 아라비아 숫자만으로 연월일을 표시하거나 준말을 나타내는 데도 쓰인다. 그리고 ⑦과 같이 표시 문자 다음에도 쓰인다.

⑥ 서. 1989. 3. 1.(서기 1989년 3월 1일)
⑦ 2. 본 론
　ㄱ. 지 명

이처럼 온점이 문장의 끝을 표시하지 않는 경우는 코드를 바꾸거나 다른 부호를 사용해야 한다.

(3) 어절 구분과 부호 제거

공백(space)을 기준으로 어절을 구분하여 자모순으로 재배열하면 동형의 어절들이 모이게 된다. 어절을 구분하는 과정에서 부호를 제거해야 하는데 주의해야 할 점들은 다음과 같다.

i) 가운뎃점과 빗금은 제거하지 않는다.
ii) 쉼표는 제거하면서 그 자리도 삭제한다.
iii) 바로 앞이 공백이 아니고 마침표가 아닌 따옴표만 남겨두고 다른 따옴표는 그 자리와 함께 제거한다.
iv) 나머지 부호는 제거하면서 그 자리는 공백으로 비워둔다.

가운뎃점과 빗금은 '6·25', '경남·북', '1/4분기', '2/5(분수)'에서와 같이 그대로 하나의 어절로 처리되어야 할 경우뿐만 아니라 '방언의 조사 연구', '가물거리다/가물대다'와 같이 둘 이상의 어절로 구분되어야 하는 경우에도 쓰인다. 후자의 경우는 실사를 구분하는 과정에서 확인하여 어절을 분할해야 한다.[5]

쉼표는 수의 자릿점을 나타낼 때도 쓰기 때문에 제거하면서 그 자리를 비워 두면, 예를 들어 '2,714,653'과 같은 수는 '2', '714', '653'의 세 어절로 분리되어 잘못 처리된다.

따옴표는 대부분의 경우에 쉼표와 동일하게 처리해도 문제가 없다. 그러나 예문 ⑧에서 "'나는'은"과 "'날다'의"와 같은 어절의 처리가 문제된다. 쉼표와 같은 방법으로 처리하면 '나는은', '날다의'로 되어 허사 부분의 처리가 곤란하게 된다. iii)과 같이 처리하면 어절은 "나는'은", "날다'의"로 되고 허사 부분은 "는'은", "다'의"로 구분될 것이다.[6]

⑧ '나는'은 대명사에 조사가 연결된 것으로도 볼 수 있고 동사 '날다'의 활용형으로도 볼 수 있다.

어절 구분까지 되면 각 어절에 필요한 색인 부분이 거의 만들어지게 되는데 다음과 같다.

자료	구분	01
과목		05

5) 이 경우에 하나의 레코드를 나누어 각각의 색인에 하위 번호를 부여해야 한다. 이처럼 번거러운 과정을 거치지 않으려면 용법에 따라 코드를 구별해서 입력해야 할 것이다. 이외에 부호 사용 규정을 재검토하는 방법도 있을 것이다.
6) "는'은"은 필요에 따라 다시 '는'과 '은'으로 쉽게 구분할 수 있을 것이다.

학년	3
학기	1
페이지	174
문장	006
어절	012

우측의 코드는 1차 조사 때의 예인데 국민학교 교과서 국어 3학년 1학기 174페이지에 여섯번째로 나오는 문장의 열두번째 어절을 나타낸다. 이처럼 문장 색인을 거쳐서 어절 색인을 작성하면 그 페이지 마지막 문장의 일부가 다음 페이지로 이어지는 경우에, 다음 페이지로 넘어간 어절들의 페이지 코드는 그 문장이 시작된 페이지로 처리되기 때문에 실제와 달라지게 된다. 2차 조사에서는 페이지 다음에 행 번호를 추가로 부여하여 이러한 문제점을 보완하였다. 이밖에도 2차 조사에서는 본문 여부를 구분하는 코드도 추가하였다. 그리고 입력이나 처리 과정에서 빠뜨린 어절을 끼워 넣을 수 있도록 코드 뒤에 여분의 자리를 비워두었다.

(4) 실사와 허사 구분

어절 자료를 단어별로 정리하기 위해서는 실사를 기준으로 다시 재배열해야 하므로 어절에서 실사와 허사를 구분하여 다음과 같은 형태가 되도록 해야 한다.

색인	어절	실사	허사

실사 필드와 허사 필드를 작성하는 방법은 여러 가지가 있을 수 있다. 실사와 허사를 새로이 입력하는 방법에서부터 실사 허사 구분이 완전히 자동화되도록 전산 처리하는 방법까지 생각할 수 있다. 전자의 경우는 사실상 카드 작업을 한 후에 입력하여 재배열과 결과물 출력만 전산 처리하는 것과 거의 같다고 하겠다. 후자의 경우 현재로서는 완벽한 자동화가 불가능하다고 보아도 좋을 것이다. 결국 실사와 허사의 분리는 수작업이나 완전한 자동화가 아닌 일부 전산 처리로 방향을 잡아야 하는데, 어느 정도까지 자동화하는가에 따라 전산 처리와 관련된 부담이 크게 달라지게 된다. 이러한 문제는 조사 여건과 밀접히 관련될 것이다.

실사와 허사를 구분하는 간단한 방법으로 컴퓨터에 허사 목록을 입력하여 각 어절에서 허사를 분리하도록 할 수 있다. 1차 조사에서 이러한 방법을 사용하였다. 그러나 2차 조사에서는 정확도를 좀더 높이기 위해서

1차 조사에서 정리된 결과를 용언의 활용 정보와 함께 이용하였다.

실사와 허사를 분리하는 과정에서 생기는 어려움으로 불규칙 용언의 활용형과 축약된 어절은 실사 부분의 형태가 하나로 고정되어 있지 않아서 자모순으로 재배열한 후에도 흩어져 있게 된다.

(가) 불규칙 활용형

불규칙 활용형들을 한곳에 모으기 위해서는 재배열의 기준이 되는 실사 필드의 형태를 일치시켜야 한다. 1차 조사에서는 불규칙 활용형들간에 앞에서부터 일치하는 부분까지를 실사로 처리하였다.

⑨ ㄱ. 덥고
ㄴ. 더워서

⑨에서 '더'까지를 실사로 구분하여 이것을 기준으로 재배열하면 '덥고'와 '더워서'가 한곳에 모이게 된다. 그러나 '더'를 실사로 처리할 경우에는 부사 '더'와 동형어가 되어 구별해 주어야 할 뿐만 아니라 자모순으로 재배열하면 '더'의 위치에 정리되어 정상적인 기본형 '덥다'의 위치와는 달리 처리된다. 즉 새로운 동형어가 생기고 자모순 정리에 문제가 생긴다.

1차 조사의 두 가지 문제점을 보완하기 위해서 2차 조사에서는 실사 필드에 정상적인 어간이 작성되도록 하였다.[7] 이는 1차 조사에서 정리된 용언 목록에 규칙 활용 여부, 불규칙 활용의 종류 등의 활용에 대한 정보를 표시하여 입력하고, 활용의 종류에 따라 작성될 활용표를 이용하여 전산 처리한 결과였다. 이러한 전산 처리가 2차 조사에서 크게 향상된 부분이다.

2차 조사에서 사용한, 단어 구분 전산 처리의 흐름도(flow chart)는 142 페이지부터 143 페이지까지의 그림과 같다.[8] 이에 따른 전산 처리의 부담은 소프트웨어 구축이나 컴퓨터 사용 시간 등에서 1차 조사에 비해 크게 증가하였다.

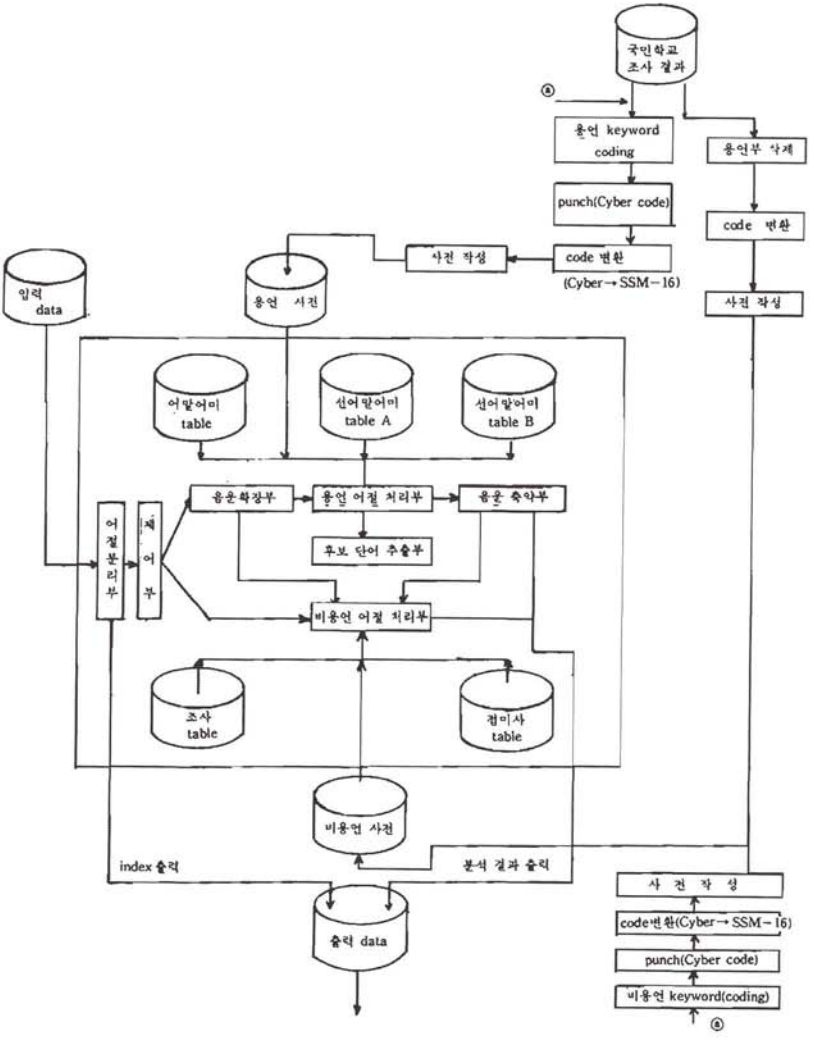
(나) 축약형

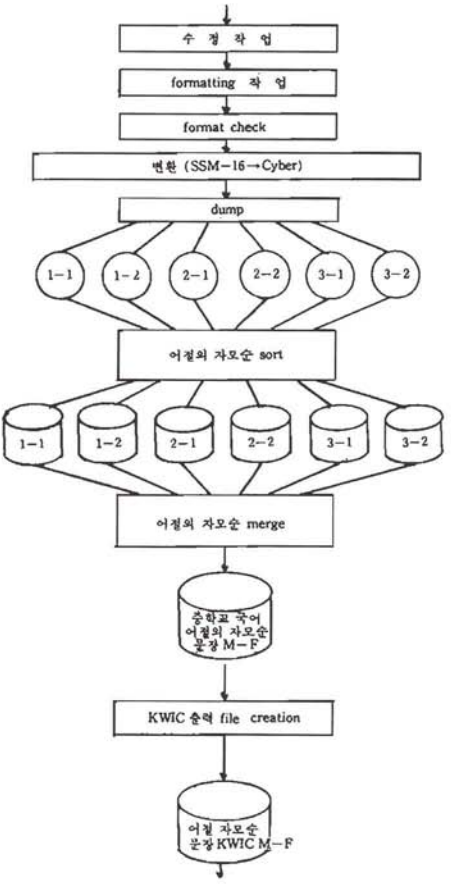
실사와 허사가 결합될 때 축약된 어절의 경우에는 한 음절에 실사와 허사가 공존하게 된다. 이런 문제를 해결하기 위해 1차 조사에서는 각 어절의 모음을 ㅏ, ㅓ, ㅗ, ㅜ, ㅡ, ㅣ 등의 기본 모음으로 풀어서 작성한 후에 실사와 허사로 구분하였다.

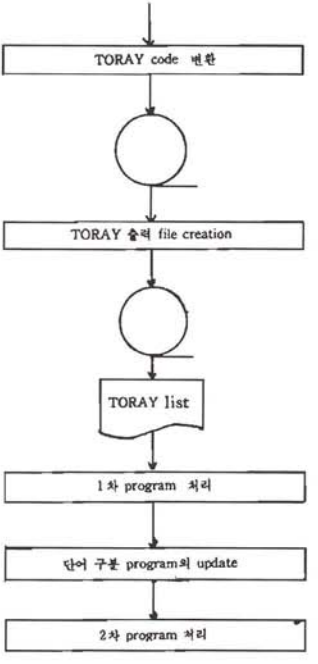
---

7) 이러한 전산 처리는 한국과학기술원(KAIST) 시스템공학센터가 日韓 기계번역으로 축적한 기술을 활용하여 추진되었다.

8) 한국과학기술원 시스템공학센터(1985), 어휘 조사 전산 처리에 관한 연구, pp. 67~91.


국민학교
조사 결과
용언 keyword
coding
punch(Cyber code)
code 변환
(Cyber→SSM-16)
사전 작성
용언부 삭제
code 변환
사전 작성
입력
data
용언 사전
어말어미
table
선어말어미
table A
선어말어미
table B
음운확장부
용언 어절 처리부
음운 축약부
후보 단어 추출부
어절분리부
제어부
비용언 어절 처리부
조사
table
접미사
table
비용언 사전
index 출력
분석 결과 출력
출력 data
사 건 작 성
code변환(Cyber→SSM-16)
punch(Cyber code)
비용언 keyword(coding)


수 정 작 업
formatting 작 업
format check
변환 (SSM-16→Cyber)
dump
1-1
1-2
2-1
2-2
3-1
3-2
어절의 자모순 sort
1-1
1-2
2-1
2-2
3-1
3-2
어절의 자모순 merge
중학교 국어 어절의 자모순 문장 M-F
KWIC 출력 file creation
어절 자모순 문장 KWIC M-F


TORAY code 변환
TORAY 출력 file creation
TORAY list
1차 program 처리
단어 구분 program의 update
2차 program 처리

⑩ ㄱ. 가져 (가지-ㅕ)
ㄴ. 가지고 (가지-고)
⑪ ㄱ. 했다 (하-ㅆ다)
ㄴ. 하고 (하-고)

이 풀어 쓴 실사를 기준으로 재배열하면 ⑩과 같은 경우에 실사 '가지'를 기준으로, ⑪은 '하'를 기준으로 모이게 되고 허사 부분은 'ㅕ' 'ㅆ다'와 같은 형태가 된다.[9]

2차 조사에서는 모음이 확장되는 과정을 거쳐 실사와 허사를 구분한 후에 관계없는 어절의 모음이 축약되어 다시 원래의 형태로 바뀌도록 하였다. ⑩ㄱ은 '가지-어'로, ⑪ㄱ은 '하-였다'로 구분되어 허사의 형태가 완전하도록 처리되었다.

(5) 동형어 구별

표기상으로 동일한 어절이 되는 것은 단어 자체의 표기가 동일한 경우와 본래 표기상으로 다른 단어이나 일부 조사나 어미가 붙어서 동일하게 된 경우가 있다. 전자는 동형어(homograph)로 어떤 조사나 어미가 붙더라도 표기 형태가 동일하다. ⑫처럼 장단 등 발음상의 차이가 있더라도 표기상으로 동일하면 이에 포함된다. 후자는 ⑬, ⑭, ⑮ 등과 같이 일부 조사나 어미가 붙을 때 표기상으로 동일하게 되는 경우로 이를 부분동형어라고 하겠다.

⑫ ㄱ. 밤이 밤(夜)
ㄴ. 밤(栗)
⑬ ㄱ. 주는 줄다(縮)
ㄴ. 주다(與)
⑭ ㄱ. 남을 남(他)
ㄴ. 남다(餘)
⑮ ㄱ. 가게 가게(廛房)
ㄴ. 가다(行)

⑬은 불규칙 활용 때문에 일부 활용형의 표기가 같아진 경우이고 ⑭는 체언의 표기 형태가 용언의 어간과 같은 경우이며 ⑮는 체언의 표기 형태가 용언의 활용형과 일치하는 경우다.

9) '돼'의 경우는 'ㄷㅗㅏㅣ'가 되어 이러한 방법으로 정상적인 어간을 구분해 낼 수 없다.

동형어에서 실사·허사의 구분과 관련된 문제는 다음과 같은 방법으로 대부분 전산 처리가 가능하다. ⑫와 ⑭는 실사나 허사 부분의 표기가 동일하므로 실사와 허사를 분리하는 데는 어려움이 없다. ⑬의 경우에 활용형 '주는'의 어간으로 '줄-'과 '주-'가 가능하므로 2차 조사에서는 둘 다 출력하여 선택할 수 있도록 처리하였다. ⑮와 같은 유형의 처리는 1차 조사에서 허사 목록에 '-게'를 포함시켜 모든 어절에서 '-게'를 분리했으므로, 용언이 아니면서 '-게'로 끝난 경우에는 모두 수정해야 했다. 2차 조사에서는 내장된 사전에 등재된 단어는 허사 목록의 형태로 끝났더라도 분리되지 않도록 하여 ⑮ㄱ이 정상적으로 처리되도록 하였다. 그러나 ⑮ㄱ과 동형의 어절인 ⑮ㄴ은 동사의 활용형이어서 '-게'를 허사로 분리시켜야 하므로 확인하여 수정해야 한다. 그리고 내장된 사전에 기억되지 않은 단어가 허사 목록과 같은 형태로 끝나면 분리되므로 이것도 보완해야 한다.

부분동형어 중 ⑬이나 ⑮처럼 실사의 표기를 달리하는 부분동형어는 실사와 허사를 구분하는 것만으로 동형어 구별이 된다. 그러나 동형어나 ⑭와 같이 실사의 표기가 동일한 부분동형어는 재배열의 기준이 되는 실사에 적절한 구별 표시를 해야 한다. 이러한 표시 방법으로 사전처럼 번호를 부여하거나 한자, 간단한 뜻풀이, 문맥 등으로 표시할 수 있는데 1차 조사와 2차 조사 모두 번호로 구별하였다.

동형어의 구별은 문맥에 의존해야 한다. 문맥과 관련된 정보를 정리하여 동형어의 구별을 자동화하는 것은 추진하기 어려우므로, 자모순으로 배열된 어절의 문맥을 출력하여 동형어 구별에 활용하는 방법을 사용하였다. 1차 조사에서는 KWOC(Key Word out Context) 색인을 출력하였고 2차 조사에서는 KWIC(Key Word in Context) 색인을 출력하였다.

(6) 빈도와 색인 출력

각 어절의 실사와 허사를 구분하고 동형어를 구별하면 레코드 형식이 다음과 같이 된다.

색인	어절	실사	번호	허사

이것을 실사와 동형어 구별 번호를 기준으로 재배열하면 단어별로 모아진다. 이처럼 정리된 화일을 이용하여 빈도를 비롯한 각종 통계와 색인을 정리해 냈다.

1차 조사에서는 국민학교 전과목 교과서를 전산 처리하여 전과목 자모

순, 전과목 빈도순, 과목별 자모순, 과목별 빈도순 등의 어휘 목록을 빈도, 색인과 함께 출력하였다. 2차 조사에서는 중학교 국어 교과서를 전산 처리하여 자모순과 빈도순으로 정리하고, 국민학교와 중학교 교과서의 어휘를 학년·학기별로 정리하여 신출어를 작성하였다.

Ⅲ. 어휘 조사 단순화 방안

앞에서 살펴본 바와 같은 대형 컴퓨터를 이용한 일괄 처리 방법은 단어 구분의 정확도에서 상대적으로 나으나 여러 단계의 처리 과정에 필요한 소프트웨어 구축과 컴퓨터 사용에 따른 부담이 적지 않다. 이는 단어 구분의 정확도를 높이기 위해 어절 범위 안에서 필요한 품사, 활용 등의 문법 정보를 전산화하여 활용하였기 때문이다. 단어 구분의 정확도를 좀더 높이기 위해, 예를 들어 동형어의 구분을 자동화하려면 감당하기 어려운 부담을 안아야 할 것이다. 게다가 정확도를 향상시킨 후에도 완벽한 자동 처리는 거의 불가능하기 때문에 전체 분량을 모두 확인해야 한다. 결국 전산 처리와 수작업을 병행할 수밖에 없게 되는데, 여건에 따라 두 가지의 비율을 적절히 조절하는 것이 바람직 할 것이다.

여러 가지 여건 때문에 가능한 한 적은 부담으로 간단히 어휘 조사를 해야 한다면 대형 컴퓨터를 이용한 일괄 처리 방법과 같이 복잡한 여러 과정을 거치기 보다는 전산 처리 과정을 단순화시켜 보는 것도 한 방법이라고 하겠다. 그러한 방안으로 메뉴 방식의 한자 입력과 유사한 어휘 조사 방법을 생각해 볼 수 있다.

이미 완료된 어휘 조사 결과를 정리하여 실사 목록, 동형어 구분을 위한 정보, 빈도 등의 내용을 입력하고 검색이 가능하도록 해서 내장 사전을 만든다.

조사 자료를 입력할 때 먼저 어절을 입력하고 나서 실사 부분을 다시 입력한다. 실사 입력과 동시에 실사가 사전에 있는지 검색되어, 있으면 하단에 동형어까지 간략한 주석과 함께 일련 번호가 부여되어 나열되고, 조사자가 이것을 보고 번호를 선택하면 되도록 한다. 검색한 결과 내장된 사전에 없으면 신출어이므로 사전에 추가되도록 한다. 그리고 허사 부분을 별도의 필드에 입력하면 허사에 대한 조사도 별도로 가능하게 될 것이다. 색인은 조사 자료의 코드와 페이지만 입력해 주고 어절 번호는 순서대로 부여되도록 한다. 이상과 같은 방법으로 입력하여 작성된 각 어절 레코드

의 형식은 일괄 처리 방법의 최종 화일의 레코드 형식과 같이 될 것이다. 결국 일괄 처리 방법에서 전산 처리와 수작업으로 여러 단계를 거쳐 최종적으로 만들어 내는 자료와 동일하게 된다. 수작업과 전산 처리로 여러 단계를 거치게 되면 각 단계마다 오류가 생기게 되어서 이를 여러 차례 점검하고 수정해야 한다. 교과서의 어휘 조사를 추진하면서 입력에서부터 최종 결과를 출력까지 단계마다 확인 점검과 오류의 수정이 큰 부담이 되었다. 그러나 여러 단계를 거쳐서 정리해 내는 결과를 한 번의 처리 과정으로 정리해 낼 수 있게 되면 상대적으로 잘못 처리되는 비율을 크게 줄일 수 있을 것이다. 뿐만 아니라 내장된 사전의 검색으로 이미 조사된 결과를 확인하면서 조사하게 되므로 조사 기준의 일관성을 유지하는 데에도 효과적일 것이다.

이 방법에서도 단어 구분이 정확히 되었는지 확인 점검하는 과정이 필요하다. 해당 레코드의 실사와 허사를 중심으로 하여 전후 레코드의 어절들을 일정한 길이만큼 보여주는 KWIC 출력을 내면, 확인 점검에 사용할 수 있을 뿐만 아니라 수정이 끝난 다음에는 용례 색인으로 활용할 수 있을 것이다.

Ⅳ. 마 무 리

두 차례에 걸쳐 추진된 교과서의 어휘를 조사하는 과정을 살펴보고 조사 과정을 단순화하는 방안 한 가지를 제시하여 보았다. 결국 단어 구분을 어느 과정에서 할 것인지 그리고 어떠한 방법으로 어느 정도까지 자동화할 것인지에 따라서 어휘 조사의 전산 처리 방법이 크게 달라지게 된다.

어휘 조사의 전산 처리는 그 목적과 여건에 따라 여러 가지 방법이 있을 수 있다. 문제는 추진이 가능한 전산 처리 여건에서 필요한 결과물을 가장 효율적으로 출력해 낼 수 있는 방안이다. 조사 과정을 축소한 어휘 조사 방안은 전산 처리 여건의 제약—근본적으로는 예산상의 제약—을 극복해 보기 위한 하나의 방편에 불과할 것이다. 어휘 조사가 단순히 어휘 빈도와 색인을 정리하는 데 그치지 않고 용례 색인과 검색을 비롯하여 문장의 길이, 부호 등의 다양한 처리를 목적으로 충분한 지원하에서 추진된다면 2차 조사 방법을 좀더 보완하여 조사하는 것이 바람직할 것이다. □