아동문헌의 구문분석을 위한 모아쓰기식 어휘사전 구축에 대한 연구

A study on construction of lexicon based on assorted writing style for syntax analysis of children literature

안지은, 이태영, 남궁황 (전북대학교 문헌정보학과)

Ji Eun An, Tae Young Lee, Hwang Namgoong
Dept. of Library and Information Science, Chonbuk National University

모아쓰기식 어휘사전은 풀어쓰기에 비해 용언어간의 크기가 늘어나고 용언어미도 많이 증대된다. 본 논문에서는 초등학생을 위한 홈페이지에서 사용되는 어휘가 상대적으로 적기 때문에 용언어간과 어미, 명사와 조사를 조화시켜 간단한 모아쓰기식 기계사전을 제시하였다.

1. 서 론

최근 수년간 전 국민적으로 컴퓨터 교육이 확대되면서, 1990학년도부터 개정 실시되고 있는 초등학교 교육과정에서도 컴퓨터를 지도할 수 있게 하기 위하여 상급학년 실과 교과서에 컴퓨터 단원을 넣어 교육을 실시하고 있다. 이를 통해 초등학생들의 컴퓨터 활용 능력이 향상되고 인터넷 사용이 증가하고 있으며, 초등학생들의 온라인 정보에 대한 요구 또한 다양해졌다. 이에 따라 초등교사들은 컴퓨터 활용 능력을 습득하여 홈페이지를 개설하고 인터넷을 통해 학생들에게 많은 학습 자료 및 정보를 제공하고 있다. 본 연구는 초등학교 교사들이 개인 홈페이지를 개설하고 필요에 따라 데이터베이스를 구축하며 더 나아가 데이터베이스에 있는 정보의 내용(본문)을 분석할 수 있게 문장분석에 필수적인 어휘사전을 구축하는데 있어서 용언의 어간, 어미와 명사구에 대한 처리방식을 모아쓰기 형태로 살펴보고자 한다.

2. 모아쓰기의 필요성 및 문제점

2.1. 필요성

구문분석을 위한 컴퓨터용 어휘사전을 구축할 때 풀어쓰기 형태로 어휘사전의 표제어를 등록·배열하면, 한 용언에 대해서 용언어간과 변화되는 용언어미들 만이 등록되기 때문에 어휘사전 구축에 별로 어려움이 생기지 않는다. 예를 들면 'ㄱㅏ ㄷㅏ'란 용언은 어간 'ㄱㅏ'와 어미 'ㄷㅏ'로 나누어져 등록되고 'ㄱㅏ ㄷㅏ'에서 변화한 단어인 'ㄱㅏ ㄴ'은 어간이 역시'ㄱㅏ'와 같기 때문에 어미만 'ㄴ'으로 등록한다. 그러나 모아쓰기에서는 '가다'는 어간 '가'와 어미 '다'가 등록이 됨은 물론

"간, 갈, 갔" 등과 같이 변화한 단어들도 용언어간에 등록되어야 하는 문제가 발생한다.

이러한 이유로 하여 교사들의 싸이트에서 구문분석이 필요할 때 풀어쓰기 형태의 단어를 적용시키면 편리하게 어휘사전을 구축할 수 있고 또 분석에 사용할 수 있다. 그런데 문제는 현재의 컴퓨터의 한글 코드들은 모아쓰기 형태의 완성형 코드를 사용하고 있기 때문에 별도의 풀어쓰기 변환 프로그램을 장착하지 않는 한 풀어쓰기 형태의 어휘사전은 만들어질 수가 없는 것이다. 따라서 현재 개인의 홈페이지에서 간단한 구문분석을 위한 아동용 어휘사전을 모아쓰기식으로 만들 필요성이 있다.

2.2 문제점

풀어쓰기식 대신에 모아쓰기식 어휘사전을 구축하게 되면 몇가지의 문제들이 발생한다. 첫째, 전절에서 말했듯이 '가다'류와 같은 용언들의 어간의 종수가 늘어나 사전의 몸체를 부풀리는 현상이다('가다'는 "가, 간, 갈, 갔, 감"과 같이 5배로 늘어난다). 둘째, '생각하다'와 같은 '하다'류 동사는 음절 '하'의 위치에서 "하, 한, 할, 해, 했, 함"과 같이 6배로 늘어나면서, 보통 "생각되다, 생각시키다"로 이어지며 "되다, 시키다"가 "하다" 만큼 어미변화를 파생시킨다는 것이다. 셋째, "구어지다, 구어내다, 구어보다, 찾아보다, 읽어보다"와 같은 용언들은 "-지다, -내다, -보다"에서 앞의 '가다'류의 용언과 같은 현상이 벌어진다. 넷째, "밀다, 밀리다, 밀치다"와 같은 단어들은 셋째 문제와 결부하여 "지다, 내다, 보다" 등이 어미로 취급될 때 어간, '밀'과 어미 "다, 리다, 치다"로 분리되어 용언어간 '밀'에 세 가지의 다른 뜻을 가진 단어가 결부될 수 있다. 다섯째, 명사 '가을'은 용언어간 '가'와 어미 '을'로 인식될 수 있다.

3. 명사/용언 사전 구축

3.1 모아쓰기식의 규약

앞장에서 지적되어진 문제점들에 대해 해결할 수 있는 원칙을 정하는데 풀어쓰기식에서 차지하는 사전의 공간크기에 될 수 있는대로 최대한 근접하려는 노력을 하였다.

(1) '가다'류는 용언어간으로 사전에 "가, 간, 갈, 감, 갔"을 올려서 다섯가지 변형어간으로 탐색하게 만듬과 동시에 '가다'가 그 단어들이 원형이라는 점을 레코드에 기록하여 한 뿌리의 단어란 점을 명시한다. 그리고 아동용 사전으로 어휘가 많지 않은 점과 '간, 감'이 명사도 된다는 것을 고려하여 명사와 용언어간을 혼합 배열한다.(표1 참조)

<표1> 명사와 용언어간 혼합 배열

단어	원형	품사	역할	결합	유사
가	가다	명용		@	가르
간	가다	명용		#	
갈	가다	용언		@	갈,가르
곰		명사		#	
구어		용언		@	
말	말하다	명용		#	
밀	밀다	명용		@	밀리,밀치

(2) '하다'류 용언은 '하다' 이전의 음절을 명사들과의 사전에 혼합 배열하고 '하다, 시키다, 되다'의 변화들을 어미사전에 등재하여 식별케 한다.(표2 참조)

예를 들어 '말시킨' 단어가 입력되었을 때에는 명사/용언어간 사전에서 '말'을, 또 용언어미 사전에

서 '시킨'을 찾아 '말하다'란 용언 원형으로 인식되는 것이다. 품사'에서 '명용'이라고 명명한 것은 그 단어가 명사도 되고 용언도 된다는 뜻이다. 용언으로 쓰였을 때는 '원형'의 용언원형 단어를 참조하여 여러 가지 값을 구하면 된다.

<표2> 용언어미 배열

단어	역할	시제	존칭	결합	보기
낸다		현재			내다
라					다
은	수식			%	다
을		미래		%	다
시킨	수식				
지니					지다
쳤으니까		과거			치다
혔고		과거			히다

(3) '구어지다'류도 '구어'를 명사/용언어간 사전에, '지다'를 어미사전에 포함시킨다.(표1과 표2 참조)

<표1>에서 '결합'란에 '@'가 있는 단어들은 명사/용언어간 사전에 있는 '단어' 항목의 용언어간과 용언어미 사전의 '보기' 항목의 어미 원형이 결합되어 용언원형이 이루어지는 단어들이다. 즉, '구어지니'란 단어는 <표1>에서 '구어'를 가져오고 <표2>에서 '지다'를 가져와 결합하여 '구어지다'의 용언원형을 갖고 차후의 여러 가지 값 계산에 참여하게 된다.

(4) "밀다, 밀치다, 밀리다"류는 '밀'만 명사/용언어간 사전에 '@'와 함께 포함시키고 '다, 치다, 리다'의 어미변화들은 용언어미 사전에 원형과 함께 등록시켜(표2 참조) 나중에 '밀'과 '다', '치다', '리다'가 결합하여 원형을 찾아가는 형식으로 설계한다. 예를 들면 '밀쳤으니까'란 단어는 명사/용언어간 사전에서 '밀'을 찾고, 다시 용언어미 사전에서 '쳤으니까'를 찾아서 '밀치다'란 용언 원형으로 자리잡게 된다.

(5) 명사/용언어간 사전에서 품사로 '명사'나 '명용'으로 등재된 단어들은 품사분석시에 우연하게 용언으로 분류될 수가 있다. 그것은 '곰은'이나 '곰을'과 같은 명사구가 '곰'이란 용언어간으로 오인이 되고 '은, 을'이란 용언어미로 식별이 되어 용언으로 둔갑 할 수 있다. 이러한 문제를 처리하기 위해 명사/용언어간 사전의 단어들에 오인의 소지가 있는 것들은 '#' 표시를 하고 또 한편으로 용언어미 사전에서 조사로도 쓰이는 어미형에 '%'표시를 하여 두 단어가 만나면 용언으로 구분될 수 없게 한다.

(6) 보조용언인 "이다, 우다"는 '이다'는 조사의 '이다'와 '우다'는 'ㅂ' 불규칙 변화 동사이기 때문에 위의 (3)과 같이 처리될 수가 없다. 따라서 "모이다, 먹이다, 죽이다 등"은 용언어간을 "모이, 먹이, 죽이"로 등록하고 "내세우다, 배우다 등"은 어간을 "내세우, 배우"로 한다.

3.2 운영상의 규약

위와 같은 형식으로 모아쓰기식 시스템을 운영하면 먼저 "-지다, -내다, -히다 등"과 같은 보조용언의 종류를 명확히 해야 한다. 왜냐 하면 한국어의 두 음절로 된 동사들은 명사 뒤에 붙어 그 단어를 용언화하기 때문이다. "꿈꾸다, 밥먹다, 책읽다, 재미있다 등"과 같이 수많은 단어들이 용언으로 등록되기를 원하면 곤란하기 때문이다. 1989년 맞춤법 통일안에 보면 "-" 등은 띄어쓰기를 권장하나 "-"와 같이 붙여써도 무방하다고 나와 있다. 그러므로 시스템에 입력시에 위와 같은 사례들은 "책 읽다, 밥 먹다"와 같이 띄어써야 한다.

그밖에 편의에 따른 어휘사전 시스템의 가동내역은 다음과 같다.

(1) 명사/용언어간 사전을 1음절, 2음절, 3음절 등으로 나누어 1음절에서 3음절까지는 조사와 어미를 분석하기 전에 먼저 입력된 단어를 사전의 어휘와 비교하는 방법을 취 한다. 이는 한국어의 쓰임새에 있어 "현실적 해결책 제출"과 같이 명사가 조사 없이 연이
어 기술되는 사실이 많으며 대부분의 명사는 1음절에서 3음절 사이에 있다.

따라서 <표1>에 기재되어 있는 '구어'는 <표3>과 같은 2음절 어휘사전에 속하게 된다.

<표3> 2음절 명사/용언어간 사전

단어	원형	품사	역할	결합	유사
구어		용언		@	
마음		명사		#	

(2) 고유명사는 고유명사의 전체 음절로 찾게 만든다. 고유명사임을 알게 하기 위해서이다.

(3) 어미의 종류는 초등학교 책에 나오는 것으로 한정하는데 보조용언의 종류는 <표4>와 같이 제한하였다. 여기서 '**'는 10종이상, '*' 는 5종 이상 나타난 보조용언들 이다.

<표4> 보조용언의 종류

**~가다, *~기다, ~꾸다, *~나다, *~내다, ~놓다, ~눕다, ~니다, ~달다, ~두다, ~들다, ~들이다, ~럽다, ~롭다, **~르다, **~리다, ~먹다, ~보다, ~붙다, ~서다, ~시다, ~쓰다, ~앉다, *~오다, **~우다,~잡다, ~주다, **~지다, ~추다, **~치다, ~키다, ~타다, ~피다, *~히다, ~받다, ~깝다

(4) 조사의 종류도 초등학교 책에서 식별되는 것으로 한정하는데 그 종류가 100여가 였다.

4. 품사 탐색 순서

한국어의 단어는 명사와 용언(동사, 형용사) 외에 부사, 관형사, 수사, 대명사, 감탄사
들이 있다. 단어가 입력되어 이들 어휘사전을 방문하여 비교되는 순서는 관형사, 부사/감탄사, 용언, 명사, 대명사/수사 순으로 정한다.

5. 결 론

정보검색시스템 중의 질문응답시스템과 유사하게 초등학교의 교사와 아동들이 자연언어로 정보를 주고 받을 수 있는 시스템을 구축할 수 있도록 모아쓰기식 어휘사전에 대한 방안을 조사하였다. '가다'와 같은 용언은 어간에 '가'말고도 "간, 갈, 감, 갔"이 더 등재가 되었으나 '하다'류 동사는 '하다' 이전의 어간을 명사/용언어간 파일에 등재하여 이중으로 등록되는 것을 막았다. '구어내다'의 '내다'와 '미치다'와 '치다'와 같은 보조용언이나 보조용언에 상당하는 것을 어미로 간주하여 어미사전에 등록하였다. 이는 어간이 많은 수로 불어나는 것을 막고자 함이였다.

참고문헌

김석주. 1991. 한글 BASIC의 개발을 위한 한글 명령어 선정에 관한 연구, 동국대학교 교육대학원 석사학위논문

김설희. 1986. 한글모아쓰기 기법에 관한 연구, 동국대학교 경영대학원 석사학위논문.