형태 정보에 기반한 전자사전에서의 3음절 명사 처리

이은전 최기선
KORTERM, AITRC, KAIST
{indian,kschoi}@world.kaist.ac.kr

Management of Three-Syllable Nouns in Electronic Dictionary based on Morphological Information

Eun-Jeon Lee Key-Sun Choi
KORTERM, AITRC, KAIST

요 약

언어학적 성과를 효과적으로 반영하고 운용할 수 있는 체계적인 전자 사전 구축을 위해선, 어휘들에 대한 총체적이고 체계적인 언어 정보 제공과 함께 효율적인 처리 방식이 무엇보다도 필요하다. 따라서 이번 전자 사전 구축 작업은 내용 면에서는 형태 정보를 중심으로 다양하고 상세한 어휘 특성들을 체계적으로 제시하였고, 기술 방식에 있어서는 모든 입력 정보를 코드화 시킴으로써 효율성을 추구했다. 또한 연구 과정에서 나타난 문제 유형에 대한 인식과 검토는 앞으로 사전 개발의 원칙 및 방향을 설정하는데 도움을 줄 수 있을 것으로 기대한다. 특히 단어 형성 정보에 있어서 접사 정보가 부착된 파생어 사전은 어휘 확장과 중의성 해결을 하는데 활용될 수 있을 것이다. 본고에서는 3음절 명사 사전 작업의 전반적인 과정, 분류 유형 , 어휘 정보, 기술 방법 및 앞으로 논의될 문제 유형들을 담고 있다.

1. 연구의 목적[1)]

국어 사전의 역할이 사람을 대상으로 의미적, 문법적 정보를 제공해 주는 것이라면, 전자 사전의 역할은 컴퓨터를 대상으로 의미적, 문법적 정보를 제공해 주는 일일 것이다. 따라서 전자 사전의 편찬 과정은 국어 사전의 편찬 과정 보다 좀 더 체계화되고 효율적인 정보 처리 방식이 요구되며, 정보의 내용에 있어서도 총체적인 언어 정보 즉, 형태적, 구문적, 의미적, 담화적 정보를 포괄한 범용적 사전이어야 한다.

따라서 이번 연구의 목적은 내용적인 면에서는 범용적 사전 구축을 위한 초석으로 형태소 해석 단계에서 우선적으로 필요한 형태적 정보를 총체적이고 체계적으로 제공하고자 하였다.

방법적인 면에서는 방대한 양의 데이터를 효과적으로 구축하기 위해선 모든 정보를 체계적으로 입력하는 방법이 필요하다. 모든 내용 및 문제 유형 등을 코드화하여 이후 추가 진행될 작업 내용이나 수정 될 사항들을 쉽고 용이하게 처리할 수 있도록 효율성을 추구하고자 하였다.

효용적인 면에서는 연구 과정에서 나타난 명사 유형을 관찰, 분석함으로써 앞으로 진행될 명사 사전의 원칙 및 방향을 설정하는데 도움을 주고자 하였다. 또한 연구 결과로 구축된 접사 사전 및 파생어 사전은 어휘 확장 및 형태소 분석에서 나타날 수 있는 '중의성(Ambiguity)' 문제를 해결하는데 있어서도 중요한 자료로 활용될 수 있을 것이다.

2. 선결 문제

어휘를 분석하고 정보를 설정하는 과정에서는 엄밀한 언어학적 기준이나 이론으로 해결할 수 없는 문제점들이 나타난다. 그 중 명사 사전에서 대표적으로 나타나는 문제점은 다음과 같다.

첫째, 표제어 선정과정에서 보통명사와 고유명사 및 전문용어의 구분이 모호하다. 예를 들어 '우울증'과 '염통핏줄신경증'의 경우, 같은 병명임에도 불구하고 전자는 경험적으로 보통명사로 처리해야 할 것 같지만 후자는 보통명사로 보기엔 전문용어의 성격이 짙게 나타난다. 그러나 현재 언어학적 이론으론 이들을 구분할 객관적이고 명확한 기준은 제시하기 어렵다.

둘째, 국어학적으로 통일되지 않은 품사들의 처리 문제다. 명사와 관련해서 대표적으로 나타나는 품사 처리 문제점은 접미사와 의존 명사, 접두사와 관형사, 접사와

1) 본 연구는 한국과학기술원 전문용어언어공학연구센터에서 수행한 과학기술부와 KISTEP의 핵심소프트웨어사업 중 "대용량 국어정보 심층처리 및 품질관리 기술개발" 과제의 일환으로 수행되었습니다.

경사의 기준 설정이 명확하지 않다는 것이다. 실제 사전이나 개인 논문마다 설정 기준이 다르기 때문에 어떠한 객관적, 절대적 기준을 따라 정보를 제공하기 어렵다.

본 연구에서는 이러한 문제점을 최소화하기 위한 방법으로 절대적 기준으로 어휘를 분석하지 않고 문제가 될 수 있는 유형들에 대해서는 그 문제 유형을 정보로 제공하는 방향으로 연구를 진행하였다. 보통 명사 선정에 있어 판단이 모호한 것들에 대해서는 같은 의미 그룹으로 묶어 정보를 제공하였고, 통일되지 않은 품사 처리는 각각의 품사 가능성을 모두 제공하여 범용적 사전으로서의 역할을 하도록 하였다.

3. 상위 정보 설정

본 연구는 KAIST 내에서 구축되어진 총 23420개의 3음절 명사 코퍼스를 대상으로 하여 실시되었다.

먼저 작업의 효율성을 높이기 위해 비서술성 명사, 동작명사, 상태명사로 세분화된 3음절 명사에 각각의 품사 태그 코드들 부착한 후 총 망라하는 작업을 실시하였다.

이렇게 종합된 3음절 명사를 체계적으로 분류하기 위해선 어떠한 정보들을 어떠한 방법으로 입력할 것인가가 결정되어야 한다. 따라서 형태 정보를 중심으로 다음과 같은 5개 범주의 상위 정보 분류 유형을 설정하게 되었다.

<1> 오류 추출 및 어휘 빈도에 따른 분류
<2> 어원 정보에 따른 분류
<3> 단어 형성에 따른 분류
<4> 기타 형태 분류
<5> 동형어 분류

이처럼 상위 5개 분류 유형을 설정한 다음에는 각각의 유형을 세분화시켜 하위 분류 내용을 설정하였다. 오류 추출 및 어휘 빈도에 따른 분류에서는 총 7개의 하위 분류 정보를, 어원 정보에 따른 분류에서는 총 5개의 하위 분류 정보를, 단어 형성에 따른 분류에서는 총 6개의 하위 분류 정보를, 기타 형태 분류에서는 총 35개의 하위 분류 정보를, 마지막으로 동형어 분류에서는 총 19개의 하위 분류 정보를 설정하였다. 이렇게 입력할 정보를 상위 5개 유형과 하위 72개의 유형으로 세분화한 후 설정된 정보를 실제 데이터작업에 입력하는 방법으로 작업을 진행하였다.

다음 <표1>은 분류된 정보가 실제 데이터에 입력된 예를 보여준다.

<표1> 실제 정보 입력의 예

3음절명사	빈도	어원	단어형성	문제유형	동형어
가로수ncn	1	2	S284		2B
가압류ncn	1	2	P3		
가이드ncn	1	3B	1	9	
가락동ncn	2				
강절도ncn	1	2		13	

4. 하위 정보 설정 및 내용

5개 상위 분류 설정에 따른 각각의 하위 분류 기준은 어떻게 설정되었으며, 어떠한 내용으로 구성되었고 각각의 정보 부착 코드는 어떻게 설정되었는지 자세히 살펴보겠다.

4.1 오류 추출 및 표제어 선정을 위한 어휘 빈도에 따른 분류

사전 작업에서 가장 기본적이고 기초적으로 이루어져야 할 작업이 바로 표제어 선정 작업일 것이다. 사전의 쓰임과 목적에 따라 가장 적절한 어휘를 선정하는 것은 그 어떤 작업보다도 신중하고 타당하게 이루어져야 한다.

따라서 이 항에서는 먼저 오류 데이터를 추출하고, 어휘의 빈도에 따른 표제어 선정을 목적으로 다음과 같은 하위 분류 유형을 설정하였다.

- 어휘 빈도가 높은 명사 <1> (예: 가압류)
- 같은 품사 태그 내에서 다의성이 발생하는 명사 <A> (예: 가스전)
- 다른 품사 태그 내에서 다의성이 발생하는 명사 <B> (예: 자연사)
- 어휘 빈도가 낮은 명사 및 고유 명사 <2>(예: 가솔화)
- 동작 명사로 처리되어야 할 것 <C> (예: 고갯짓)
- 상태 명사로 처리되어야 할 것 <D> (예: 무소득)
- 부사로 처리되어야 할 것 <E> (예: 갸우뚱)

표제어를 선정하기 위해선 먼저 어휘 빈도를 객관적으로 판단하는 일이 우선적으로 이루어져야 하지만 어휘 빈도를 객관적으로 판단하기란 결코 쉬운 문제가 아니다. 개인의 지식 정도, 처한 환경, 직업에 따라 개인이 인식하고 있는 어휘의 정보는 상당한 차가 있을 수 있기 때문에 연구자 개인이 표제어 선정 작업을 한다는 것은 커다란 위험을 감수해야 하는 일이다. 따라서 다음과 같은 기준을 설정하고 최대한 객관적인 표제어 선정을 하려고 노력하였다.

첫째는 연구자의 직관에 의해 선정하였다. 물론 객관적인 연구 방법이 될 수는 없지만 기본 어휘를 선정하는 데는 시간적으로 가장 효율적인 방법이 될 수 있다.

둘째는 KAIST 형태소 분석기를 이용하여 선정하였다. 모든 어휘를 검색하기에는 많은 시간이 필요했고, 분석기 역시 완벽한 결과를 내지 못하기 때문에 직관에 의해 선정하기 어렵거나 모호한 어휘에 대해서만 분석기를 통해 어휘의 빈도를 검사하고 선별하여 등재하였다.

셋째는 두 가지 기준에 의해서도 확실한 결과를 얻지 못한 경우는 한글학회사전 검색을 통해 정보를 얻어 실제 언어 생활에 많이 사용되는 어휘인지를 판단하였다.

넷째는 인명이나 지명 같이 확실한 고유명사는 등재를 하지 않는 것을 원칙으로 하였다. 그러나 기타 고유명사나 범위가 모호한 어휘에 대해서는 일단 등재해주고 기타 형태 유형이나 동형어 분류에 별도 표시를 해줌으로써 추후 논의를 거쳐 일괄적으로 처리할 수 있도록 하

였다.

이상과 같은 작업 기준에 의해 어휘 빈도를 측정해서 '가압류', '가속도' 같이 어휘 빈도가 높은 것과 '가솔화', '서울대'와 같이 고유명사나 어휘 빈도가 낮은 것을 구별하여 정보 코드를 입력하였다.

이렇게 어휘 빈도 검증을 거친 후에는 빈도가 높은 어휘 중 다의성이 발생하는 명사를 추려 별도의 정보 코드를 부착하였다. 다의성의 발생하는 명사의 경우는 다시 두 가지 상황으로 나누어 질 수 있는데, '가스전'처럼 비서술성 명사라는 같은 품사 태그 내에서 '가스층'이란 의미와 '가스를 사용한 싸움'의 의미가 나타나는 유형과 '자연사'처럼 '역사'의 의미를 나타낼 때는 비서술성 명사로 태깅되고, '죽음'의 의미를 나타낼 때는 동작명사로 태깅되는 유형이 있다. 이러한 유형의 경우는 3음절 명사를 세분화할 경우 각각의 품사에 이중으로 들어가 있어야 하며, 파생어의 경우는 각기 다른 접사 코드 정보가 부착되어야 하기 때문에 선별하여 정보 입력을 하였다.

이 항에서는 또한 기존의 작업에서 잘못 품사 태그를 부착한 어휘를 선별하여 새로 수정할 필요가 있음을 제시하였다.

4.2 어원 정보에 따른 분류

이 항에서는 표제어로 선정된 명사에 대해서 어원 정보를 제공하기 위해 다음과 같은 하위 분류 유형을 설정하였다.

- 고유어 표기 유형 <1> (예: 개울가)
- 한자어 표기 유형 <2> (예: 개정안)
- 외래어 표기 유형
  *일본어 표기 유형 <3A> (예: 기모노)
  *로마자 표기 유형 <3B> (예: 넌센스)
- 복합 유형 (로마자+고유어, 고유어+한자어 등)
  (예: 노벨상)

어원 정보는 전자 사전에 있어서 기본적으로 다루어져야 할 정보로 이번 작업에서는 고유어, 한자어, 외래어, 복합유형으로 크게 나누고, 외래어에 있어서만 일본어인 것과 로마자인 것으로 세부 분류하였다. 그러나 복합 유형의 경우는 '노벨상'처럼 '로마자+한자어'가 결합된 유형, '농구팀'처럼 '고유어+로마자'가 결합된 유형 등 많은 유형으로 다시 세분화 될 수 있으므로 추후 재검토 될 때에는 좀 더 세분화하여 정보를 입력할 필요가 있을 것이다.

4.3 단어 형성 정보에 따른 분류

한국어는, 특히 한자어 때문이기는 하나, 흔히 조어력이 강한 언어로 일컬어진다. 이를 달리 말하면 국어는 유연성(有緣性)이 큰 복합어(complex world)를 많이 지니고 있는 언어라고 할 수 있다(김정은 1995:7). 이러한 한국어의 성격은 사회, 문화의 변화에 따른 새로운 사물, 개념에 대한 명명이 필요할 때, 합성법과 파생법에 의해 생산적인 새 단어형성을 촉진시키는 장점을 가지고 있다. 때문에 전자사전에 있어서도 이러한 한국어의 단어 형성법을 잘 활용한다면 충분히 새롭게 생성되는 단어 및 표제어 이외의 단어를 구축하고 체계화하는데 중요한 역할을 하리라 생각이 된다.

단어형성법에 관한 기존의 연구는 많이 이루어져 있다. 그러나 각각의 연구마다 연구 방향 및 기준 설정이 다르기 때문에 어떤 연구 분류가 객관적으로 가장 타당한지 검증하기 어려운 점이 있다. 또한 우리가 구축하고자 하는 전자 사전은 기존의 문법의 틀에 가장 가깝게 접근하면서도 실제 데이터를 효과적으로 다룰 수 있는 방향으로 만들어져야한다는 요건을 다 충족시켜야 하기 때문에 실제 작업을 하는데 있어서는 다른 분류작업보다 까다롭고 어렵게 이루어졌다.

기존 문법에서의 단어 형성의 유형을 살펴보면 단일어(simpl world), 복합어(complex world)로 구분할 수 있고, 복합어는 다시 파생어(derivational world)와 합성어(compound world)를 포함한다. 단 용어의 경우 학자에 따라 사용에 차이를 보이고 있기 때문에 여기서는 한글맞춤통일안, 이희승(1955), 고등학교 문법 교과서(1991) 등에서 사용하는 용어를 따름을 밝힌다.

이렇게 기존 문법에서는 단어 형성의 유형을 단일어, 파생어, 합성어로 크게 구분 짓고 있지만, 이러한 구분 유형을 기준으로 실제 데이터를 처리하기에는 다음과 같은 문제와 장애에 부딪치게 된다.

첫째, 기존의 사전이나 논문의 경우 편찬 연구자의 관점에 따라 문법적 정의가 다르게 나타나고 있다. 현존 문법이 체계적으로 통일되고 있지 않기 때문에 어떠한 이론적 근거를 따라 작업을 진행해야 할지 객관적인 기준을 설정하기 어렵다.

둘째, 실제 언어 생활에서 사용하는 어휘 중에는 기존 문법에 근거한 단어 형성 분류로 설명할 수 없는 어휘들이 상당하다는 것이다. 보통 논문의 경우 다루어진 예들이 극히 한정된 것들이기 때문에 기존 문법의 기준에 따라 분류되지 않는 유형들이 많고 이를 어떻게 극복할 것인가가 어려움으로 남는다.

따라서 이 항에서는 이러한 문제점을 최소화하면서, 기존 문법의 틀에 가깝게 접근하고 데이터를 효율적으로 처리하기 위해 다음과 같은 단어 형성 분류를 실시하였다. 즉 단어를 기존 문법의 단어 형성 분류 유형인 단일어와 파생어 합성어로 구분하고 그 밖에 문제가 되는 단어에 대해서는 좀 더 세분화하여 정보 코드를 입력하였다. 이 항에서는 다음과 같은 하위분류 코드를 설정하였다.

- 단일어 유형 <1> (예: 농땡이)
- 파생어 유형 <P접사코드, S접사코드>
  (예: 노처녀, 노예제)
- 합성어 유형 <3> (예: 농구공)
- 1음절 관형사형 어기 + 명사 어기 <4>
  (예: 뜬구름, 군만두)
- 명사 + 비독립 성분의 형태 <5> (예: 대양저)
- 비독립 성분 + 명사의 형태 <6> (예: 중거리)

단일어는 어기가 하나의 형태소로 이루어진 단어로 단어 형성에 참여하는 기본 요소가 된다. 이 작업에서는 단일어의 경우 형태론적으로 더 이상 나눌 수 없는 비독립 성분으로 이루어진 단어로 한정하였다.

파생어란 형태론적 단위에 접두사나 접미사가 붙어 새로운 단어를 형성하는 복합적 구성이다. 따라서 단어의 파생여부를 확인하기 위해선 우선 접사와의 결합 여부를 살펴야 할 것이다. 그러나 기존 사전의 경우, 사전마다 임의적으로 접사를 수록하고 있을 뿐만 아니라 접사 선정 기준마저도 밝히고 있지 않지 있다. 연구 논문들 역시 접사의 설정 기준 등이 각기 다르며 형태론적 입장에서 체계적으로 접사를 구명한 논문을 찾아보기 힘들다.

따라서 이 작업에서는 먼저 기존에 나와 있는 사전과 개인 연구에서 인정하는 접사들을 종합하여 접사 목록을 구축하는 작업을 우선적으로 실시한 후, 파생어의 범위를 구축된 접사와 결합된 명사로 한정하였다. 물론 종합된 접사 중에는 학자마다 개인적으로 접사로 인정하지 않는 것도 포함되어 있지만 어법적 검증은 일단 추후에 하기로 하였다. 일단 모든 것이 코드화 되어 있기 때문에 재검토된 어휘는 추후에 쉽게 수정할 수 있기 때문이다.

접사 목록을 구축한 후에는 각각의 접사에 고유 코드 번호를 부착하고, 이를 실제 데이터에 정보를 입력하였다. 실례로 '노처녀'의 경우 '늙다'의 의미인 '老'가 접사 코드 63번으로 정해져 있으므로 '노처녀'에는 'P63'이란 정보가 입력된다. 이러한 방법으로 정보를 입력한 이유는 같은 접사와 결합된 파생어들을 쉽게 추출해냄으로써 접사의 생산성을 확인하거나 누락된 어휘들을 보완하는 데 도움을 줄 수 있기 때문이다.

합성어란 어기와 어기의 결합에 의해 구성된 단어로, 합성어 역시 설정 기준 및 성립 조건을 학자마다 다르게 설정하고 있다. 또한 합성어의 경우 배합 양식이 매우 다양하므로 이들의 결합 양상에 따라 구와의 구분이 모호할 때가 있다. 이런 상황에서 모든 합성어에 두루 적용되는 기준을 설정하기란 쉬운 문제가 아니다.

따라서 이 작업에서는 구조론적 기준으로 봤을 때 설령 내적 확장이 가능하여 구의 특징을 강하게 가지고 있는 단어에 대해서도 합성어로 인정하는 한편 의미론적 기준으로 봤을 때 새로운 의미변화 없이 단순한 결합으로 이루어진 단어에 대해서도 합성어로 인정하였다.

실례로 '작은집'의 경우는 내적 확장이 가능하여 구로 정의할 수도 있지만, 실제 언어 현실에서는 붙여쓸 수도 있다는 점을 생각해 일단 합성어로 인정하였다. 다만 구로도 쓰일 수 있다는 점을 감안해 이런 유형에 대해서는 별도의 정보 코드를 추가하였다. '봄가을'의 경우처럼 어떠한 의미변화 없이 구성성분의 단순한 결합으로 이루어진 단어에 대해서도 합성어로 인정하였다.

실제 작업 과정에서는 위와 같은 단일어, 파생어, 합성어의 분류만으로 모든 단어의 유형을 결정할 수 없는 상황이 있어 다음과 같은 단어 유형은 다른 정보 코드로 처리했다.

1음절 동사 어기 및 형용사 어기(관형사형)와 명사 어기와 결합된 유형이다. 위에서 언급한 바와 같이 어기와 어기가 결합된 유형은 설령 내적 확장이 가능하더라도 일단은 합성어로 처리하고 실제 구로도 쓰일 수 있는 점을 감안해 별도의 처리를 했음을 밝혔다. 그러나 1음절 동사 어기나 형용사 어기가 결합된 단어의 경우는 이미 굳어진 형태로 쓰이는 경우가 더 많기 때문에 별도의 유형으로 분류하였다. 이러한 유형의 예로는 군만두, 긴머리, 뜬구름 등이 있다.

<5>, <6>코드로 처리된 유형은 파생어와의 경계가 모호한 단어에 대한 처리 유형이다. 접사로 인정되지 않은 비독립 성분 + 어기가 결합된 유형이다. <5> 코드 유형은 접두 파생어 형태이지만 접두사로 인정되지 않은 비독립 성분과 결합된 유형이고, <6> 코드 유형은 접미 파생어 형태지만 접미사로 인정되지 않은 비독립 성분과 결합된 유형이다.

실례로 '단거리'의 경우, '短'은 '짧은'이란 의미로 '長'과 비슷한 역할을 하지만 '長'은 접사로 인정되어 '장거리'가 파생어로 처리되는 반면, '短'은 기존의 사전이나 연구 논문에서 접사로 인정하지 않기 때문에 파생어로 처리할 수 없는 형태이다. 이러한 '비독립 성분'의 경우 비록 기존 문법에서 접사로 인정하고 있지는 않지만 일괄성 있게 데이터를 처리하기 위해서는 필요한 성분이고 또한 국어 문법의 한계성을 극복할 수 있다는 점에서도 별도의 유형으로 처리하는 것이 필요하다고 생각한다.

4.4 기타 문제 유형 분류

이 항은 표제어 선정, 단어 형성 분류 과정 및 기타 사전 처리 과정에서 문제가 될 수 있는 형태 및 같은 특징을 나타내는 유형을 분류하고, 추후 일괄성 있게 처리할 목적으로 설정하였다. 총 35개의 세부 유형으로 분류하였는데 그 중에서는 작업 과정 중에서 이미 문제 해결을 한 유형도 일부 포함이 되어 있다.

이 항에서 분류한 하위 유형을 제시하면 다음과 같다.

- '적', '용'과 결합되어 관형어와 명사로 다 쓰이는 유형 <1>(예: 개인용)
- 부사와 명사로 다 쓰이는 유형 <2> (예: 가급적)
- 같은 의미가 이형태로 나타나는 유형 <3> (예: 가랑이, 가랭이)
- 동사 또는 형용사의 명사형 유형 <4>(예: 가려움)
- 파생어 중 접사의 생산성이 높은 유형 <5> (예: 가정간, 책상류)
- 속어나 방언 중 비교적 잘 쓰이는 유형 <6> (예: 가시내, 둥어리)
- 신조어 유형 <7> (예: 가요방)
- 특정 어휘와 주로 결합하는 유형 <8> (예: 간덩이(붓다))
- 비서술성 명사 중 '하다'와 결합할 수 있는 유형 <9>(예: 간막이, 요리사)
- 약자 혹은 준말 유형 <10>(예: 갈고리(갈고랑이))
- 단위성 명사와 결합된 유형 <11> (예: 갈래길)

- 명사형과 결합된 유형 <12> (예: 갈림길)
- A(B+C) 혹은 (A+B)C유형<13>(예: 출입구, 강절도)
- 형태변화 유형 <14> (예: 갯마을, 바닷길)
- 파생어 중 접사와 결합한 명사의 생산성이 낮은 유형 <15>(예: 공처가)
- A/B+C와 A+B/C 가 모두 가능한 유형 <16> (예: 대학교, 독문학)
- 2음절 동사, 형용사 어기(관형사형)+ 명사 어기 유형 <18> (예: 갓난애, 미친개)
- 동사의 어근 + 접사의 유형 <19> (예: 지우개)
- 초중고 유형 <20> (예: 육해공)
- 아라비아 숫자나 알파벳과 결합된 유형 <21> (예: 컬러TV)
- 접미사와 결합한 어근이 명사가 안되는 유형 <22> (예: 축음기, 순발력)
- 1음절 명사 어기+ 2음절 비독립 한자어 결합 유형 <23> (예: 폐활량)
- 형태상으로 부사와 명사가 다 되는 유형 <24> (예: 불화로, 전기로)
- 수사와 결합된 유형 <25> (예: 십자가)
- 그룹으로 나타날 수 있는 유형 <26> (예: 더블류)
- 합성어 중 의미의 변화가 일어난 유형 <27> (예: 오리발, 죽사발)
- 2음절 비독립 한자어+ 1음절 명사 어기 결합 유형 <28> (예: 도열병, 삼지창)
- 사이시옷이 들어간 형태로 쓰일 수 있는 유형 <29> (예: 소개장(소갯장))
- 단위성 명사로도 쓰이는 유형 <30> (예: 달구지)
- 앞에 수사를 동반해야 의미를 파악할 수 있는 유형 <31> (예: (한) 가닥쯤)
- 접사와 부사가 결합된 형태 <32> (예: 가일층)
- 접두파생어 중 독립적으로 잘 쓰이지 않고 뒤에 다른 명사나 접사가 오는 형태로 쓰이는 유형 <33> (예: 비독립(국))
- 2음절 비독립 고유어 + 1음절 명사 어기 결합 유형 <34> (예: 고명딸, 방아쇠)
- 1음절 명사 어기 + 2음절 비독립 고유어 결합 유형 <35> (예: 길잡이, 꽃꽂이)

4.5 동형어 분류

이 항의 분류 목적은 앞에서 언급된 표제어 선정 과정에서 나타나는 문제점 해결을 돕고 같은 특성을 지닌 단어를 묶어 봄으로써 일괄적 처리를 용이하게 하는 한편 누락되어 있는 데이터를 쉽게 찾아 구축하고자 함에 있다.

고유 명사의 경우 한국어 코퍼스 태깅 작업에서도 마찬가지 문제겠지만 그 경계를 확실히 구분 짓기 어려운 점이 있다. 인명과 지명을 고유명사로 인정하는데는 이의가 없지만 기타 다른 분야의 어휘는 고유명사와 일반 명사의 경계가 뚜렷이 나타나지 않는다.

실례로 동물·식물명의 경우 '개구리'나 '개나리' 같은 어휘는 사전에 등재를 해야 할 것 같지만 실제 어디까지 등재를 원칙으로 해야 하는가? 에 대해서는 개인적 기준을 설정하기 어렵고 여러 사람의 논의를 거쳐 이루어져야 할 문제다.

따라서 이 항에서는 문제가 될 수 있는 동형어들을 따로 분류해놓음으로써 문제가 되는 유형들을 인식하고, 추후 충분한 논의를 거쳐 새롭게 재정비할 수 있도록 하였다.

이 항에서 분류한 동형어 분류 유형은 다음과 같다.

- 동물명 <1>(예: 개구리)
  동물명 중 의미가 전성된 것 <1A>(예: 불여우)
- 같은 특성을 지닌 동물을 지칭하는 어휘 <1B> (예: 겨울새)
- 식물명 <2> (예: 개나리)
- 식물명 중 의미가 전성된 것 <2A>(예: 메밀꽃)
- 같은 특성을 지닌 식물을 지칭하는 어휘 <2B>(예: 보호수)
- 과일·채소명 <2C>(예: 토마토)
- 신체일부 <3>(예: 복사뼈)
- 광물명 <4>(예: 석회석)
- 철도·고속도로명 <5>(예: 장항선)
- 별자리 <6>(예: 천왕성)
- 공무원 직급명 <7>(예: 행정관)
- 종교명 (교파포함) <8>(예: 화엄종)
- 육십갑자+년, 명절, 기념일 <9>(예: 갑진년)
- 병명 <10>(예: 거절증)
- 종교명 외에 종교와 관련된 것(무속 신앙 포함) <11> (예: 교황청)
- 옛날 제도와 관련된 것(제도, 관직, 놀이 등) <12>(예: 농부가)
- 화학, 물리 관련 용어 <13>(예: 니코틴)
- 친인척 관련 용어 <14> (예: 당고모)

5. 1음절 접사 처리 과정 및 결과

5.1 접두사 처리

접두 파생어는 접두사와 단어와의 결합에 의해 만들어진다. 때문에 접두 파생어의 구성요소인 접두사를 정확히 이해해야 접두 파생어의 성격도 아울러 분명해질 것이다. 그러나 기존의 접두사 정의를 살펴보면 각 논문과 사전마다 접두사 설정이 다르게 되어 있고, 접두사의 정의도 일치하지 않고 있다.

이미 접두사 정의들에서 공통적으로 찾아볼 수 있는 점은 '접두사는 어근 앞에 붙는다, 독립성이 없다, 단음절어다, 문법적 기능 없이 의미 변화만 일으킨다'는 것이다. 그러나 대부분의 경우 접두사의 개념을 구체적인 기준에 의해 규정하지 않고 있다.

여기서는 이러한 문제점을 해결하기 위해 다음과 같은 방법으로 접사 목록을 종합하고 비교하였다.

5.1.1 조사 범위 및 방법

본 작업에서는 위에서 제기한 문제를 해결하기 위해 우선 기존에 연구되어진 접두사 관련 논문을 정리하는 작업을 하였다. 논문의 정리 결과 이미 출판된 다섯의 사전 즉 국어대사전(이희승,1988), 새우리말 큰사전(신기철외,1992), 국어대사전(김민수 외,1922), 우리말큰사전(한

글학회,1992), 우리말 분류사전(남영신,1993)에 수록된 접사와 김규선(1971), 성환갑(1972), 김종훈(1973), 서병국(1974), 이재성(1990)에서 합당한 접두사로 인정한 1음절 접두사 목록을 수집하였다. 이렇게 수집된 목록 중 현대 언어에서 잘 쓰이지 않아 생산성이 거의 없는 접두사는 가려내고 총 327개의 접두사를 종합하였다. 물론 접두사 '날'의 경우처럼 같은 형태라도 '날강도'나 '날생선'같이 접두사의 의미가 달라지는 것은 각각의 독립적인 접두사로 처리하였다.

접두사를 종합한 후에는 각 접두사마다 고유의 정보 코드를 붙여 실제 데이터작업 과정에서 파생어를 처리하는데 정보를 제공하였고, 가장 최근에 편찬된 국립국어연구원의 표준국어대사전과 전자통신연구원에서 인정한 접두사 목록을 토대로 비교·토론 과정을 거쳐 최종적으로 한국어 코퍼스 태깅 작업에 사용될 접두사 목록을 완성하였다.

5.1.2 문제 유형 및 기존 사전의 처리 결과

기존 사전들을 비교하는 과정에서 나타난 결과는 각 사전마다 임으로 접두사를 등재했기 때문에 동일한 어휘에 대해서도 어떤 사전에서는 접두사로 인정하는 한편 어떤 사전에서는 관형사로 처리하고 있다는 사실이다. 그만큼 아직 국어 문법 내에 통일성 없는 문법들이 실로 많이 존재하고 있다는 것을 입증하고 있다.

여기에서는 위에서 언급한 5개의 사전과 5명의 개인 연구 논문에서는 접두사로 인정된 바 있지만 표준국어대사전에서 다르게 품사 처리된 유형을 살펴봄으로써 앞으로 품사 처리를 하는데 참조하고자 한다.

<1> 관형사 인정 어휘
* 각(各, 각각의) - 각가정, 각고을, 각나라
* 당(當, 바로) - 당회사, 당열차
* 딴(고유어, 다른) - 딴소리, 딴생각
* 매(每, 매번) - 매시간 매주일
* 맨(고유어, 처음) - 맨먼저, 맨처음
* 뭇(고유어, 여러) - 뭇사람, 뭇백성
* 별(別, 다른) - 별사람, 별천지
* 본(本, 근원) - 본바탕, 본고장
* 새(고유어, 새로운) - 새신랑, 새나라
* 수(數, 여러) - 수십명, 수천명
* 순(純, 순수한) - 순이익, 순수입
* 양(兩, 양쪽의) - 양국가, 양도시
* 옛(고유어, 옛날의) - 옛국가, 옛도시
* 전(全, 전부) - 전세계, 전국민
* 전(前, 전의) - 전단계, 전담임
* 주(主, 주요) - 주원인, 주목적
* 첫(고유어, 처음의) - 첫사랑, 첫만남
* 타(他, 남) - 타지방, 타국가

<2> 명사로 인정 어휘
* 각(角) - 각도장
* 견(絹) - 견직물 견방직
* 관(官) - 관노비
* 꾀(고유어) - 꾀병
* 낱(고유어) - 낱글자, 낱그릇
* 면(綿) - 면직물
* 미(美) - 미소년, 미남자
* 반(半) - 반봉건, 반죽음
* 복(伏) - 복더위
* 산(山) - 산나리, 산난초 (식물명)
* 쌍(雙) - 쌍곡선, 쌍가마
* 안(고유어) - 안사돈, 안주인
* 알(고유어) - 알사탕
* 왜(倭) - 왜간장, 왜남비
* 우(右) - 우회전
* 의(義) - 의형제
* 좌(左) - 좌회전
* 질(고유어) - 질그릇
* 후(後) - 후삼국

<3> 부사 인정 어휘
* 갓(고유어) - 갓마흔, 갓스물

5.2 접미사 처리

접미 파생어 역시 접미사와 단어와의 결합에 의해 만들어진다. 접미사도 접두사와 같이 각 논문과 사전마다 설정 기준이 다르게 되어 있고 정의도 일치하지 않는다. 이 작업에서는 이미 KAIST 내에서 구축된 접미사 목록을 근거로 다른 사전 및 논문과의 비교를 거쳐 접미사 목록을 확정하였다.

5.2.1 조사범위 및 방법

접두사 목록의 경우는 기존에 사전과 논문에서 인정한 접사를 수집, 정리는 방법으로 만들어졌으나, 접미사의 경우는 이미 1차적으로 KAIST 내에서 구축한 접미사 목록을 근거로 보완, 수정하는 방법으로 만들어졌다. 이미 기존에 수집된 총 629개의 1음절 접미사 중에는 접미사라 칭하기 어려운 명사나 비독립적 한자어도 상당부분 포함되어 있다. 그러나 문법적 검증은 이후에 실시하기로 하고 각 접미사마다 고유의 정보 코드를 부착해 실제 데이터작업 과정에서 파생어를 처리하는데 정보를 제공하는 한편 접두사와 마찬가지로 가장 최근에 편찬된 국립국어연구원의 표준국어대사전과 전자통신연구원에서 인정한 접두사 목록을 토대로 비교·토론 과정을 거쳐 최종적으로 한국어 코퍼스 태깅 작업에 사용될 접미사 목록을 완성하였다.

5.2.2 문제 유형 및 기존 사전의 처리 결과

접미사 역시 각 사전마다 등재 여부가 다르게 나타나고 품사 역시 다르게 나타나고 있다.

특히 접미사의 경우는 접미사와 의존 명사의 경계가 모호한 것이 가장 많이 있다. 의존 명사라면 분명 띄어쓰기를 해야 하지만 실제 그 경계가 모호하거나 명사와 의존 명사의 결합력이 강해서 붙여쓰는 것이 자연스러운 경우가 많기 때문이다.

여기에서는 표준국어대사전에서는 의존 명사로 처리하고 있지만 실제 언어 생활에서는 명사와 결합함으로써 접미사와의 경계가 모호한 유형을 살펴보고자 한다.

<1> 의존 명사 인정 어휘

* 간(間, 사이) - 친척간, 부부간, 학교간
* 격(格, 역할) - 대부격, 주인격
* 결(고유어, 사이) - 아침결
* 내(內, 안) - 기안내, 회사내
* 년(고유어, 여자) - 도둑년
* 녘(고유어, 때) - 아침녘
* 대(代, 기간) - 20대, 30대
* 말(고유어, 끝) - 학년말, 학기말
* 분(고유어, 존칭) - 남자분, 여자분
* 시(고유어, 때) - 투자시, 판매시
* 양(고유어, 모양) - 어린양
* 조(調, 어조) - 웅변조, 경멸조
* 중(中, 진행) - 휴학중, 혈액중
* 초(初, 처음) - 올해초, 재임초
* 측(側, 쪽) - 학교측, 여당측
* 치(고유어, 몫) - 당년치, 하루치
* 편(고유어, 곳) - 행정편, 교통편
* 하(下, 아래) - 지배하, 영도하
* 회(回, 순서) - 최종회, 마지막회

6. 결론 및 향후 계획

전자 사전의 최종 목표 중의 하나는 자연 언어가 가지고 있는 모든 형태적, 구문적, 의미적, 또는 담화적 정보를 모두 포함한 범용적인 사전의 구축일 것이다.

그러나 언어의 모든 정보를 한 번에 빠짐없이 수록한다는 것은 결코 쉬운 일이 아닐뿐더러, 효율적인 면에서도 결코 옳은 방법은 아니라 생각된다.

따라서 이 번 연구는 우선 형태소 해석 단계에서 우선적으로 필요한 형태적 정보를 빠짐없이 포함시키려 노력하였고 방법에 있어서도 정보의 자질에 따라 상위 정보와 하위 정보로 나누어 체계적으로 분류·제시하였다.

특히 어휘 형성 정보를 기반으로 구축된 파생어 사전 및 접사 사전은 어휘 확장 및 형태소 단계에서 나타나는 중의성을 해결하는데 중요한 자료가 될 것으로 생각한다.

앞으로 과제는 형태 정보 뿐만 아니라 의미 정보, 구문 정보까지 확장한 사전 구축 작업이 필요하고, 특히 접사를 이용한 어휘 베이스 확장을 위해서는 명사의 체계적인 어휘 분류가 선행되어야 한다. 그러나 단순한 '의미적 분류'가 아닌 어휘의 쓰임이나 활용을 전제로 한 '어휘적 분류'가 되어야 할 것이다.

그 밖에 세부적으로는 연구 과정에서 나타난 언어 현상적인 문제들을 충분한 논의를 거쳐 어떻게 처리할 것인가를 결정지어야 한다. 또한 사전이 어느 정도 한국어 문법 정보 체계를 갖춘 후에는 기계 번역을 위해 기타 외국어와의 품사 체계를 비교하고 정보를 추가하는 작업도 이루어져야 할 필요성이 있다.

참고 문헌

[1]. 서정미(1994), 현대 한국어 접두 파생어, 경기대학교
[2]. 김정은(1995), 국어 단어형성법 연구, 박이정
[3]. 남지순, 어절 정보 사전'을 이용한 형태소 분석의 중의성 해결