한국어 처리를 위한 품사 체계 연구

°안 미정, 김 재한, 옥 철영
울산대학교 전자계산학과

A Study on a Part of Speech for Korean Natural Language Processing

°Mi-Jung Ahn, Jae-Han Kim, Cheol-Young Okcy
Dept. of Computer Science, Ulsan University

요 약

지금까지의 한국어 자연언어 처리에 기반이 되는 사전 품사 체계에 대한 연구는 형태소 분석, 구문 구조 분석, 그리고 의미 분석 등의 다양한 분야에서 이루어져 왔다. 한국어 자연언어 처리 각 분야는 자체의 고유한 독립성을 가지는데, 이러한 특성은 사전 품사 체계의 다양화를 초래하였으며, 연계성있는 자연언어 처리를 위한 통합 환경 조성을 저해시켜 왔다. 본 논문에서는 한국어 자연언어 처리 전반에 걸친 통합 환경 조성을 위한 범용적인 사전 품사체계의 필요성에 따라 한국어 자연언어 분석의 각 분야에 적합한 사전 품사체계에 대하여 살펴 본 후, 한국어 자연 언어 처리 전반에 사용될 범용적이고 통합적인 기본 사전 품사체계 구축을 위한 방안을 제시한다.

I. 서 론

근래의 한국어 분석에 대한 연구는 사전을 기반으로 이루어지고 있다[2,5]. 한국어 분석에서 사전으로 인한 오류가 시스템의 오류로 직접 이어지므로[4], 사전은 전체 시스템의 효율성을 좌우하는 중요한 부분이라 할 수 있다. 한국어 사전은 문장 성분을 결정하는데 중요한 구실을 하는 기능어(조사 및 어미)가 제대로 등록되어 있지 않고, 실제 많이 쓰이는 단어가 미등록되어 있다거나 필요없는 고어가 그대로 등록되어 있어서, 그리고 특히 보조 용언, 접미사 그리고 복합 명사와 같은 표제어를 다루는 부분은 일관성을 찾아 보기가 힘들기 때문에 컴퓨터로 한국어를 처리하는 데 그대로 사용하기에는 크게 미흡하다[9].

한국어 처리에 기반이 되는 해석 과정에 대한 연구는 한국어 형태소 분석[2,4,7,9], 구문 분석[1,11], 의미 표현[10,12] 등에 걸쳐 이루어져 왔으며, 이를 통해 어느 정도의 기술이 축적되었다고 볼 수 있다. 한국어 처리에 기반이 되는 사전 품사 체계에 대한 연구 또한 형태소 분석, 구문 분석, 의미분석 등 각 단계에서 행해져 왔으나[2,5,6,7,8], 특정 품사에 대한 분류만이 행해졌을 뿐 각 단계간의 연계성을 고려한 분류는 행해지지 못했다. 또한 해석에 대한 연구들은 기존의 각 분야에 대한 독립적인 연구 결과, 단지 특정한 단계에 치중되었거나, 세 단계를 모두 다루었더라도 각 단계 간의 모호성 해결을 위한 적극적인 접근은 하지 못했다. 이는 사전 품사 체계의 다양화를 초래하였으며, 연계성있는 자연 언어 처리를 위한 통합 환경 조성을 저해시켜 왔다.

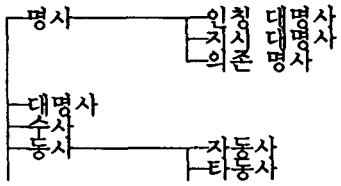
본 논문에서는 한국어 처리 전반에 걸친 통합 환경 조성을 위한 범용적인 사전 품사 체계의 필요성에 따라 한국어 분석의 각 분야를 위한 사전 품사 체계에 대하여 살펴 본 후, 한국어 처리 전반에 사용될 범용적이고 통합적인 기본 사전 품사 체계 구축을 위한 방안을 제시한다.

본 논문의 2장에서는 한국어 형태소 분석을 위한 품사 분류에 대하여 논하고, 3장에서는 국어학에서의 품사 체계를 그대로 수용하면서 구문 분석 과정에서 필요로 하는 정보를 제공하기 위해 세분화된 품사 체계에 대하여 논한다. 4장에서 구문 분석시 발생하는 중의성을 배제시키기 위해 필요로 하는 용언 통사규칙 정보와 용언의 의미 정보에 대하여 간단히 논한 후, 5장에서는 본 논문에서 한국어 해석의 각 단계에서 필요로 하는 사전 품사 체계를 통합하여 계층적으로 구성한 한국어 처리 전반에 사용될 수 있는 기본 사전 품사 체계를 제시한다.

II. 형태소 분석을 위한 품사 분류

형태소 분석은 입력 어절을 이루고 있는 형태소의 분리 및 복원 과정으로 특별히 세분화된 품사 체계를 필요로 하지 않는다. 따라서 한국어 형태소 분석에 필요한 사전은 국어 사전으로부터 필요한 정보를 추출하여 구성한다. 이를 위하여 국어 사전에 등록된 표제어, 품사, 불규칙 정보, 용언화접미사 중 '-하다', '-되다', '-스럽다', '-거리다', '-대다'와 결합 가능한 명사를 추출하여 사용한다.

용언화 접미사는 명사나 의성어 또는 의태어에 붙어 용언으로 전성시키는 접미사로 특히 빈도수가 높은 것은 '-거리다', '-대다', -하다'로 각각 전체 용언의 14%, 14%, 19% 정도(전체 용언의 47%)를 차지한다. 본 논문에서는 어휘 사전 크기와 사전 검색 빈도를 감소시키기 위해 명사에 접미사 -하다'가 붙어서 용언이 되는 경우(명사 18,394 중 명사 83,916개가 해당됨)를 **하다용언류**로, '-스럽다'와 '-되다'와 결합하여 자동사가 되는 경우와 형용사가 되는 경우를 각각 **되다자동사, 스럽다형용사**로 품사를 분류한다. 의성어나 의태어에 붙어 용언이 되게 하는 '-거리다'와 '-대다'는 그 특성이 동일함을 이용하여 **거리다대다용언류**로 분류한다. 형태소 분석에 필요한 사전을 효율적으로 검색할 수 있도록 구성하기 위하여 접미사가 붙어서 용언이 되는 모두를 사전에 표제어로 수록하지 않고 각각의 명사 또는 부사에 품사 코드정보를 추가하여 표현한다[2,4]. 그림 2-1은 본 논문에서 정의한 형태소 분석을 위한 품사 분류이다.



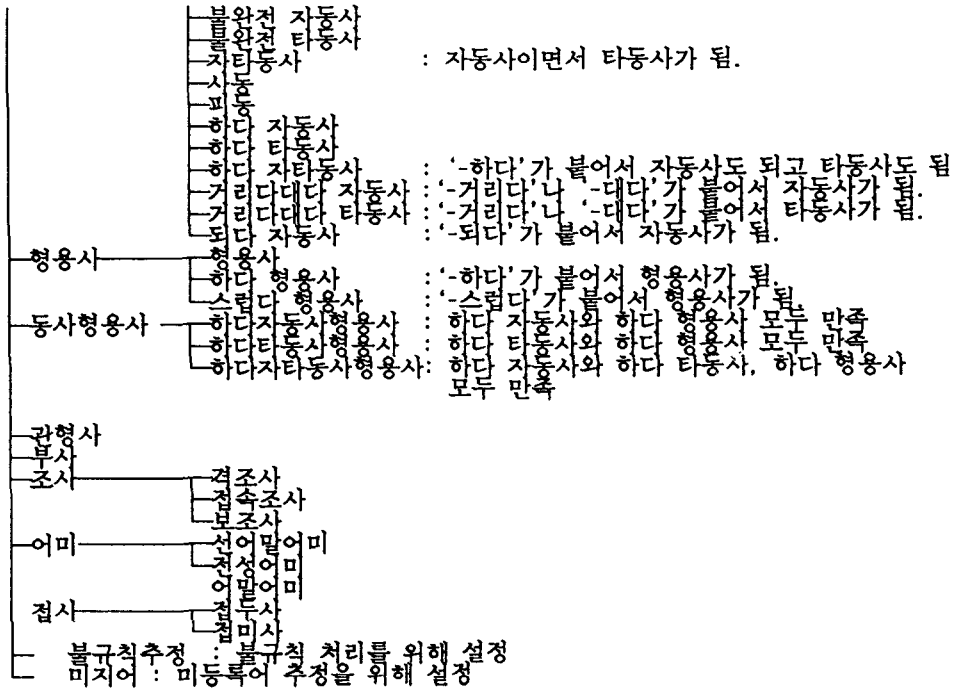


그림 2-1 형태소 분석을 위한 품사 분류

Ⅲ. 구문 구조 분석기에 적합한 사전 품사 체계

한국어가 부분 자유어순을 가지며 교착어라는 특성은 구문 해석 과정에서 다양한 중의성을 발생시켜 구문 해석을 비결정적으로 만든다[11]. 따라서 중의성을 최대한 줄이기 위해서는 중의성을 발생시키는 어절에 대한 정확한 정보를 제공하는 것이 중요하다. 구문 분석 단계에서의 중의성은 주로 생략된 문장성분, 미비한 구정보, 복합 명사구, 복합 조사 또는 보조사로 인한 격정보 애매성으로 인해 일어난다[10]. 이를 해결하기 위해서는 형태소 분석시 각 표제어가 가지는 품사 정보보다 더 상세한 구문 정보 및 의미 정보를 제공할 수 있도록 품사 체계를 더 세분화시켜야 한다. 따라서 본 논문에서는 기존 국문법상의 품사체계를 기본 바탕으로 하고, 전산 처리에 적합하도록 일부를 수정하여 구문 구조 분석에 필요한 품사 체계를 하위범주화시켰다.

3.1 조사 분류

한국어에서는 조사에 의해서 정확한 문법관계의 파악이 곤란한 모호성이 발생된다. 한국어에서 조사는 격조사와 보조사, 접속 조사로 분류된다. 격조사는 주로 체언이나 용언의 명사형이 일정한 자격을 가지도록 하여 문법적 관계를 나타내는 역할을 하나, 보조사와 접속조사는 여러 성분에 두루쓰여 문법적 관계보다는 특별한 의미를 더해주거나 두 단어를 같은 자격으로 이어준다[10,13]. 보조사는 격조사를 대신하거나 격조사나 어미에 덧붙을 수 있으므로 컴퓨터로 한국어를 분석할 때에 구문 구조 파악을 어렵게 만든다. '격조사+격조사',

'보조사+격조사', '격조사+보조사', '보조사+보조사'의 조사끼리 두 결합 이상의 결합 형태는 규칙으로 처리할 경우, 오류도 많고 규칙 또한 찾기가 힘들므로 통합형 조사로 취급하여 각각 하나의 조사로 분류하여 사전에 수록한다[5,14]. 통합형 조사를 하나의 격조사로 처리하기 위해 본 논문에서는 [14]에서의 가정을 일부 사용하여 표3-1과 같은 규칙을 새롭게 정의하였다.

(a) '격조사 + 격조사'로 이루어진 통합형 조사는 뒤에 오는 격조사로 조사를 부여한다.
예) '에서가'는 '가'가 주격조사이므로 주격조사로 등록
(b) '격조사 + 보조사' 또는 '보조사 + 격조사'로 이루어진 통합형 조사는 격조사로 조사를 부여한다.
예) '에서만은'은 '에서'가 부사격조사이므로 부사격조사로 등록
(c) 단독 보조사 또는 '보조사+보조사', '조사'와 '보조사'로 이루어진 통합형 조사는 격조사를 나타내지 않고, 용언의 의미 정보와 구문 정보를 이용하여 적합한 격조사를 정한다.

(d) 격조사와 보조사간의 두 결합 이상의 통합형 조사는 격조사로 조사를 부여한다.
예) '만으로도'는 '으로'가 부사격 조사이므로 부사격으로 등록

표 3-1 통합형 조사에 격조사를 부여하기 위한 규칙

(a)는 [14]에서는 '격조사+격조사' 형태의 통합형 조사가 10여개로 잘 사용되지 않는다고 하여 처리하지 않았으나, 본 논문에서는 실제 한국어 문장에서 자주 사용되므로 새롭게 정의하였다. (b),(c)는 [14]에서의 가정을 그대로 수용한 것이다. 이러한 규칙을 이용하면 (1) 관용절에서 보조사를 가지는 어절이 용언 앞 또는 뒤, 그리고 앞과 뒤 양쪽에서 나타났을 경우, (2) 보조사와 조사가 생략된 어절에서 발생하게 되는 중의성을 해결할 수 있다. 표 3-2는 본 논문에서 정의한 조사 분류 및 규칙을 적용하여 구성한 조사의 분류를 나타낸 것이다.

조사분류	격 정 보	대표 조사
주격조사	주어	이/가, 께서, 에서, 서, 에서는, 에서가, 까지가
목적격조사	목적어	을/를, 까지를, 만을, 부터를,
관형격조사	관형어	의, 에서의, 부터의, 까지의,
부사격조사	부사어	에, 에게, 한테, (으)로, 으로까지, 만으로는,
접속 조사	*	와/과, (이)고, (이)며, 하고, 하며, 와는, 와도, 까지와
호격 조사	독립어	아/야, (이)여, (이)시여

* 접속 조사는 격정보를 부여하지 않는다.

표 3-2 조사의 분류

본 논문이 가지고 있는 사전에 등록된 조사는 모두 260개로 금성판 국어 사전을 기반으로 하여 구축한 사전에서 추출한 조사 145개 이외에, 사전에 등록되지 않은 복합 조사는 국민학교 읽기 교과서(2학년 ~ 6학년)와 중학교, 고등학교 국어 교과서에서 빈도수가 높은 복합 조사 115를 기본 사전에 등록시켜서 구축한 것이다. 이들 중에서 단독 조사(조사, 보조사, 격조사)가 60 개, 두 결합, 세 결합, 네 결합을 포함한 복합 조사는 각각 143, 45, 12 개로 조사에서 복합 조사가 상당수를 차지함을 알 수 있었다. 그러나 국어 사전에는 실제 많이 사용되는 복합 조사 미등록되어 있기 때문에 범용성있는 처리를 위해서는 이를 계속 보강해야 한다.

3.2 용언의 분류

3.2.1 보조 용언의 품사분류

구문 분석 과정에서 보조 용언과 본용언이 붙여 쓰기된 경우, 그 용언이 합성 용언인지 또는 어느 것이 본용언인지, 보조 동사인지 보조 형용사인지에 대한 구별이 필요하다. 국어학에서는, 최현배(1837)에서 보조 동사의 설정과 그 판별 기준을 정의한 이후로 많은 학자들에 의해 연구가 이루어져 왔으나 논자에 따라 견해가 엇갈린다.[15] 따라서 컴퓨터를 이용하여 구문 구조를 분석하기 위해서는 전산학적 측면에서 국어학에서의 객관적인 기준을 근거로 하여 보조 용언을 분류할 필요가 있다. 보조 용언은 분류하기가 어려울뿐 아니라, 용언이 요구하는 구문 정보에 대해서도 모호성을 가지게 한다[16]. 따라서 본 논문에서는 구문 분석의 효율을 위해서 표 3-3과 같은 용언 처리를 위한 규칙을 정의하여 이를 해결하고자 한다.

(1) 본 용언과 보조 용언이 독립된 어절로 입력된 경우 본용언 바로 뒤에 붙는 연결 어미에 따라 보조 용언의 종류 및 양태적 의미를 구별한다. (단, 용언이 가지는 구문 정보와 의미 정보를 먼저 고려한다.) (2) '본용언 + 연결 어미 + 보조 용언'형태의 한 어절로 입력된 경우에는 보조 용언이 가지는 문법 범주를 본용언이 가지게 하고 본 용언의 구문 정보와 의미 정보를 참조하여 구문 분석한다.

표 3-3 용언처리를 위한 규칙

(a) 철수가 책을 들고 갔다.
(b) 나는 철수를 학교에 가게 했다.
(c) 그것만 이루어 놓고 가거라.

본 논문에서 정의한 규칙을 이용하면 형태소 분석 단계에서 생기는 중의성 문제도 처리할 수 있으며, (a)와 같이 연결 어미가 본동사 사이에 위치할 때, (1)에 의해 '가다'라는 보조 동사는 '-아/-어'형의 연결 어미를 필요로 한다는 정보에 따라 이를 본용언과 보조 용언으로 처리하여 원래 문장 구조와 다른 그릇된 분석을 사전에 막아 준다. (b)에서는 '하다' 보조 동사가 '-게'형의 연결 어미를 필요로 한다는 정보와 문법 규칙 (2)를 적용하여 보조 동사가 사동의 뜻을 내포한다는 의미정보를 이용하여 문장을 올바르게 분석한다. 또한 (c)와 같이 '본용언+보조 동사+본용언'도 문법 정보와 각각의 보조 용언이 갖는 정보를 이용하여 분석할 수 있다. 표 3-3에 정의된 규칙을 이용하면 본용언의 어미가 '읽기도 한다'처럼 '전성 어미에 보조사(도, 만, 는 등)가 붙어서 뒤에 오는 용언이 보조 용언임을 나타내는 경우도 처리 가능하다. 표 3-4는 본 논문에서 국어 사전에 등록된 모든 보조 용언 분류를 기반으로 보조 용언을 정의한 것이다.

보조용언의 분류	종 류
보조동사 보조형용사 불완전보조동사 하다보조동사 하다보조형용사	가다, 굴다, 나다, 내다, 놓다, 등 법하다, 아니하다, 않다, 있다, 등 달다, 다오, 달라뿐

표 3-4 보조 용언의 분류

본 논문에서는 동사 또는 형용사의 어미 '기'에 보조사가 붙는 말 아래에 '하다'가 쓰이는 경우는 다른 보조 용언과는 달리 일반 동사와 성격이 거의 같다는 점에서 국어학에서와 같이 보조 동사의 범주에서 제외 시켰다. 또한 '하다'가 붙어서 보조 용언이 되는 유형들은 각각 **하다보조동사, 하다보조형용사**로 분류하였다.

국어 사전에 등록되어 있는 모든 보조 용언에 규칙 (1)을 적용하여 얻은 정보와 [3]에서 보조 동사에 대한 문법범주를 기반으로 표 3-5과 같이 세부 코드를 부여하여 본 용언과 보조 용언이 하나의 통합된 용언으로 입력되었을 경우에 본 용언을 구별하는 데 사용한다.

	대표 형태소	품 사	종 류
'아' 형	-아/-어/-여	보조동사, 보조형용사	가다, 가지다, 계시다, 나다, 내다
'게' 형	-게/-도록	보조동사, 보조형용사	하다, 만들다, 생기다
'지' 형	-지	보조동사, 보조형용사	아니하다, 못하다, 말다, 않다
'고' 형	-고	보조동사, 보조형용사	싶어하다, 있다,
'보다' 형	-는가/-은가/나/	보조형용사	보다, 싶프다,

표 3-5 보조용언의 문법범주

3.2.2 구정보 분류

한국어 문장 분석시의 통사 모호성은 하나의 어절이 여러 형태소로 분석되고, 각 형태소가 문법에 맞는 구문 구조를 형성할 수 있을 때 발생된다. 그러나 대부분의 분석 결과는 의미적으로 타당하지 않은 구문 구조를 생성한다. 한국어 문장의 구문 구조를 분석하기 위한 방법으로 특정 언어 이론이나 분석 기법을 적용한 연구들이 있어왔으며[10,12], 구문 구조 분석과정에서 의미적으로 타당하지 않은 결과를 제거하기 위해 동사가 가지는 하위범주화 성분에 제약을 가하는 선택제약에 관한 연구도 있어왔다[10,12]. 그러나 이러한 연구들은 하나의 어절이 여러 형태소로 분석되는 경우에 발생하는 통사적 모호성은 근본적으로 해결하지 못하며, 비문법적인 관용 표현이나 문장에서의 일부 성분이 생략된 대화체의 분석에는 어려움이 있다. 또한 개별 어휘의 하위범주화 정보를 수록하는 어휘 사전을 구축하여 관리하는 데에도 문제점이 있다. 본 논문에서는 관용구로 인하여 생기는 모호성을 해결하기 위해서 관용구를 크게 세 가지로 분류하여 구문 분석에 사용한다.

(1) '의존형태소+용언류'(관용어): -에 대하다, 로 더불다, -ㄹ 까 보다, -되지 못하다, -은가 보다 등으로 관용어로 코드를 부여한다.

(2) '체언(조사)+용언류'(관용구): 거드름(을)피우다, 거울(로)삼다, 귀(를)기울이다, 벼락(이) 치다 등으로 관용구로 코드를 구성한다.

(3) '용언+용언'(합성 용언): 합성용언은 보조 용언이 붙어서 된 것은 제외하고 나머지는 사전에 등록한다.

(2), (3) 같은 형태는 원래는 조사가 있는 형태로 쓰이다가 이미 한말로 굳어졌다고 인정되는 말로 주로 사전에 표제어로 등록된 것도 있으나 한국어 사전에서 이에 대한 예외가 많기 때문에 본 논문에서는 사전에서 관용구 정보를 뽑아서 (1)은 관용어로, (2)는 관용구로 따로 표제어로 등록하고 각각에 적합한 품사를 부여하고 어떠한 조사가 어떤 음절뒤에 있었던 것인가에 대한 정보를 코드화하여 '거드름피우다'가 '거드름을 피우다'의 형태로 들어온 경우에도 처리되도록 하였다. 이와 같이 관용구 사전을 따로 구성하여 구 단위 정보를 이용함으로써 복잡한 구문 규칙없이도 정확한 구문 구조를 분석할 수 있을 뿐만 아니라, 의미 해석

을 하지 않고도 상당수의 의미 모호성을 구문구조 분석 과정에서 제거할 수 있다. 또한 기존의 구문 분석기에서 처리하기 어려웠던 대화체 문장이나 관용 표현들도 특별한 구문 규칙을 설정하지 않고서도 처리가 가능하다.

3.3 명사 분류

본 논문에서는 금성판 국어 사전에서 분류해 놓은 명사 분류를 기반으로 하여 명사를 크게 보통 명사와 의존 명사로 분류하였다. 의존 명사는 다른 명사에 비해 조사의 생략이나 격조사의 결합에 제약이 많기 때문에 구문 분석시에 많은 중의성이 발생하게 된다. 이를 해결하기 위해서 의존 명사의 기능 측면에서 분류하여 의존 명사가 단독으로 쓰여 격정보를 알 수 없음으로 해서 생기는 오류를 막을 수 있도록 하였다.

(a) 노력한 만큼 보답을 받는다.
(b) 죽은 줄 알았다.
(c) 죽은 줄로만 알았다.
(d) 나는 빨리 달릴 수 있다.

(a),(d)에서와 같이 체언류 뒤에 오는 조사가 생략됨으로 해서 격정보를 알 수 없을 경우, '만큼'은 문장에서 부사어로만 쓰이는 부사성 의존 명사이고, '수'는 문장에서 주어로만 쓰이는 주어성 의존 명사라는 정보가 주어지면 격정보를 알아낼 수 있다. 또한 (b),(c)에서처럼 '줄'이 목적성 의존 명사로서 문장 분석 과정에서 '알았다'('알다'(타동사)+'았'+'다')가 가지는 구문 정보에 만족할 수 있도록 격정보를 줄 수 있기 때문에 파싱 과정에서 불필요한 분석 구조를 제거할 수 있다. 표 3-6은 본 논문에서 의존 명사를 분류한 것을 나타낸 것이다. 이러한 분류를 이용하여 용언의 의미 정보를 표현하면 조사의 생략, 보조사와의 결합으로 인한 격정보를 정할 수 없는 경우, 잘못된 문장 판별 등으로 인해 생기는 모호성 등을 효율적으로 처리할 수 있다.

코 드	기 능	해당 의존 명사
보편성	여러 성분으로 가능	분, 이, 것, 데, 바, 따위 ....
주어성	주어로만 사용됨	지, 수, 리, 나위....
서술성	서술어로만 사용됨	따름, 뿐, 터, ...
부사성	부사어로만 사용됨	대로, 양, 듯, 체, 둥, 만큼, 뻔, 채, 만, 줄...
목적성	목적어로만 사용됨	줄, 지
단위성	수량이나 단위를 나타냄	개, 채, 마리, 자, 푼, 평

표 3-6 의존 명사 분류

3.4 부사의 분류

부사는 문법적인 관계를 표시하는 말이 특별히 나타나지 않으며, 다른 품사에 비하여 문장에서의 위치가 자유롭기 때문에 특정한 구문 정보를 도출해 내기가 어렵다. 따라서 부사와 동사, 부사와 체언, 부사와 부사, 부사와 관형사, 부사와 구, 또는 부사와 절 그리고 부사와 한 문장 사이의 결합 관계를 설정하는 것은 국어학에서뿐만 아니라 전산학적 관점에서 볼 때에도 큰 의미를 지닌다고 할 수 있다. 부사는 일반적으로 수식할 수 있는 품사가 한정되어 있으므로 용언류, 명사등에 대한 하위 분류와 함께 부사에 대한 하위 분류가 이루어져야 하며, 분포상의 기준 및 의미상의 기준이 수립되어야 한다. 그러나 이제까지의 연구에서는 주로 명사나 용언에 대한 품사 분류는 많이 있어 왔지만 부사를 세분화하여 제시하지는

못했다. 또한 국어 사전에 시간을 나타내는 부사의 대부분이 명사로만 등록되어 있고 의성, 의태 부사가 등록되지 않아 파싱 트리만 많이 생성한다. 본 논문에서는 이러한 문제를 해결하기 위하여 구문 분석에 적합하도록 부사를 세분화하여 표 3-7로 제시한다. 접속형은 호응 정보를 필요로 하기 때문에 각각의 부사와 호응이 되는 어미류에 대한 정보도 제공되어야 한다.

분 류	기 능	해당 예
일반형	모든 용언과 어울릴 수 있다.	모두, 다, 갑자기, 많이, 웅기종기 대략, 대관절, 몽땅 등
시간형	시간을 나타내는 명사와 시간 부사로 보든 용언과 어울릴 수 있다.	그때, 접때, 가끔, 이따금, 그젠 어제, 오늘, 일찍, 이제, 하마 등
처소형	모든 용언과 어울린다.	이리, 그리, 어디, 요리, 고리, 여기, 거기 등
동사제한형	동사만 수식한다.	상당히, 무척, 굉장히, 퍽 등
형용사제한형	형용사만 수식한다.	휘어청, 똑, 마치, 작히, 정히, 일반 정도부사
명사제한형	명사적 서술어 수식	매우, 꽤, 퍽 등
부사제한형	일부 부사 수식	잘, 퍽, 빨리, 자세히 등
부정형	부정부사로 용언을 부정.	안, 아니, 못 등
활용형	형용사 + -이로 서술어의 역할을 한다.	없이, 있이, 같이, 달리, 듯이 등
호응형	단어 문장전체를 수식.	진실로, 실로, 마땅히, 모름지기 과연, 단연코, 꼭, 반드시, 기필코 흡사, 마치, 틀림없이 결코, 조금도, 전혀 왜, 어찌, 설마, 하물며 아마, 혹시 만약, 혹시, 가령, 비록, 아무리 제발, 부디, 아무쪼록 정녕, 정작
접속형	단어와 단어, 문장과 문장 이어주면서 뒤의 말을 수식	또, 또는, 곧, 그러므로, 그리고, 그러나, 하지만, 더욱, 그런데

표 3-7 부사 분류

Ⅳ. 구문 분석을 위한 용언의 통사규칙 정보와 의미 정보

한국어는 문장의 중심어인 술어가 문장의 맨 뒤에 오며, 피수식어가 모든 수식어의 뒤에 나타나는 중심어 후행언어이다. 한국어에서 중심어 후행에 위배되는 것도 있는데, 그 대표적인 예가 체언이 조사의 앞에 위치하는 경우와 본용언이 보조 용언의 앞에 위치하는 경우이다. 이들의 처리를 위해서는 구문 정보가 특별히 필요하지 않기 때문에 구문 분석 이전의 단계에서 처리한다. [17]

술어가 가지는 정보를 이용한 문장 분석은 술어가 지니고 있는 의미 특성과 통사 특성을 이용하여 우좌 중심어 우선의 파싱을 해 나가는 것이 기존의 하향식이나 상향식 파싱 과정에서 여러 번 반복 되었던 backtracking을 감소시키고 조사 생략 및 보조사로 인한 격성분

의 모호성 해결, 한국어 어순의 자유로움으로 인한 분석의 어려움 및 내포문의 분석을 보다 효율적으로 처리할 수 있다[9].

용언이 가지는 구문 정보는 각 용언의 통사적 규칙을 살펴보면 쉽게 얻을 수 있다. 본 논문에서는 동사가 지니는 통사적 규칙을 이용하여 구문 정보를 형성할 수 있음을 보인다.

동사는 크게 나누어 보면 자동사와 타동사로 나눌 수 있다. 동사는 형용사와는 달리 그 유형을 분류하기가 힘들다. 또한 이중 주어 유형이나 이중 목적어 구문의 경우 양상이 매우 복잡하고 자동사이면서 목적어를 필요로 하는 경우도 있기때문에 세분화가 다른 품사에 비하여 다양하다고 볼 수 있다. [6]에서는 동사를

(1) 주어 + 동사
(2) 주어+보어+동사
(3) 주어+목적어+동사
(4) 주어+주어+동사
(5) 주어+간접목적어+직접목적어+동사
(6) 주어+목적어+목적보어+ 동사
(7) 주어+목적어+목적어+동사

의 7가지로 자동사류와 타동사류의 구별없이 영어구문에 맞도록 분류하였었다. 본 논문에서는 국어 사전에서 예문을 조사하여 자동사와 타동사가 갖는 통사규칙을 세분화하여 구문 정보로 사용한다.

자동사가 갖는 통사규칙

(1) 주어 + 동사
(2) 주어 + 부사어(-에/-에게/-께) + 동사 : 예) 순이가 차에 올라타다.
(3) 주어 + 부사어(-와/-과/-함께) + 동사 : 예) 철수가 순이와 다투다.
(4) 주어 + 부사어(-(으)로/-로서/-로써) + 동사 : 예) 철수가 시의원으로 나가다
(5) 주어 + 부사어(-보다) + 동사 : 예) 벌레가 전보다 멸하다.
(6) 주어 + 주어 + 동사 : 예) 얼음이 물이 되다.

타동사가 갖는 통사규칙

(1) 주어+목적어+동사 : 가리다, 골다, 때리다, 부수다.
(2) 주어+부사어(-에서/-으로부터)+목적어+동사 : 뽑다, 긷다, 지새다, 갈아타다.
(3) 주어+부사어(-에게서/-한테서)+목적어+동사 : 빼앗다, 보다, 얻다, 구하다.
(4) 주어+부사어(-에게/-게)+목적어+동사 : 끼치다, 선물하다, 보내다, .
(5) 주어+부사어(-에)+동사 : 가져오다, 먹다, 두다, 흘리다, 감다.
(6) 주어+목적어+부사어(-(으)로/-로서)+동사 : 가공하다, 가꾸다, 삼다,
(7) 주어+부사어(-와/-과/-하고/-랑+목적어 +동사 : 나누다, 섞다, 비교하다.
(8) 주어+목적어+부사어(-(이)라/-라고)+동사 : 이르다, 부르다, 일컫다, 칭하다.

용언의 의미 자질과 통사규칙간의 관계에서 보면 서술어의 의미에 따라서 적합한 명사의 종류가 한정될 수 있음을 알 수 있다. 용언화 접미사의 경우 앞에 붙을 수 있는 명사가 한정되어 있으나 일정한 규칙을 적용하기 위한 명사 분류가 제대로 이루어지지 않으므로 인해 예를 들어, "아름답다"와 같은 경우에는 '아름답다'(형용사)와 '아름(명)+-답다(접미사)'로 분석해 낸다. 따라서 술어의 의미 정보를 위해서는 신빙성있는 명사에 대한 의미 분류가 선행되어야 한다. 기존의 연구에서도 명사에 대한 분류를 시도해 왔으나 국한된 범위에서의 한국어 처리에만 사용되었다. 아직 이에 대한 연구가 진행 중에 있으며, 빈도수가 높은 명사에 대하여 분류를 시도하고 있다. 따라서 신빙성 있는 분류를 위해서는 다양한 분야에 걸

친 한국어 문장 수집 및 분류가 시급하다. 본 논문에서는 술어가 가지고 있는 정보를 각 용언이 가지는 구문 자질과 의미 자질을 부여하여 표 4-1와 같이 구성한다.

[표제어(동사,형용사):구문정보코드(필수격1, 의미자질1 ...):(필수격2, 의미자질2...)]

주다:타동사(주격,인성 명사):(목적격, 무정명사 비인성명사):(여격, 인성명사)]

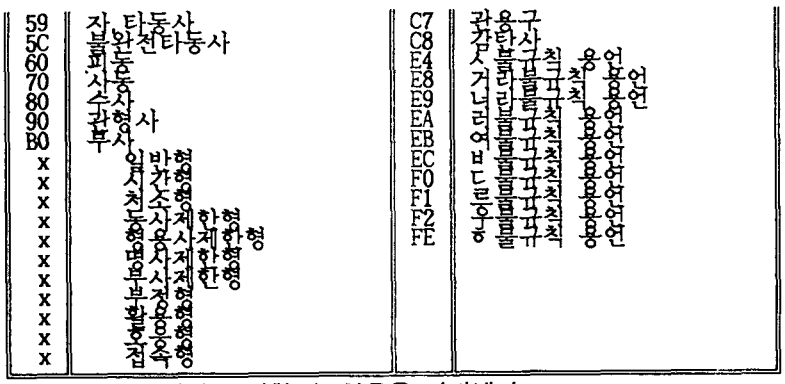
표 4-1 : 용언이 가지는 의미 정보 및 구문 정보

V. 한국어 처리 전반에 적합한 범용적인 사전 품사 체계

1장에서 3장에 걸쳐 한국어 처리에 있어서 형태소 분석기와 구문 구조 분석기를 위한 사전 품사 체계에 대하여 논하였다. 본 장에서는 각 단계에서 제시한 품사 체계를 통합하여 한국어 처리 전반에 적합한 범용성있는 사전 품사 코드 체계를 제시한다.

본 논문에서는 한국어 전체 사전에 표기된 품사를 자연언어 처리에 사용하기 위해 형태소 분석기나 문장 분석기에서는 아스키 형태의 코드보다는 품사 계층적 구별이 용이한 2진 코드로 변환하여 사용했다. 하나의 품사는 한 바이트에 표현되며, 비트별로 계층적 의미를 가진다. 예를 들어 모든 명사류는 비트 00으로, 모든 용언류는 비트 01로 시작한다. 이러한 2진 형태의 품사 체계 구성은 품사 비교 및 검증이 용이하다. 표 5-1은 본 논문에서 제시하는 한국어 처리 전반에 적합한 사전 품사 체계를 나타낸 것이다.

2 진 코드	코드 명칭	2 진 코드	코드 명칭
01	인칭대명사	A0	거리다,대다자동사
02	지시대명사	A4	거리다,대다타동사
03	의존 명사	A8	거리다,대다형용사
x	보편성	C0	조사
x	주어성	x	주격조사
x	서술성	x	목적격조사
x	부사성	x	관형격조사
x	목적성	x	부사격조사
x	단위성	x	접속조사
04	명사	x	호격조사
21	하다자동사	C1	준말
22	하다타동사	C2	접두사
23	하다자,타동사	C3	접미사
24	하다형용사	C4	선어말어미
25	하다자동사형용사	x	존칭선어말어미
26	하다타동사형용사	x	과거형선어말어미
27	하자자,타동사형용사	x	
28	하다보조동사	C5	어미
2C	하다보조형용사	x	평서형종결어미
30	되다자동사	x	의문형종결어미
31	스럽다형용사	x	명령형종결어미
40	형용사	x	청유형종결어미
42	보조형용사	x	감탄형종결어미
50	자동사	x	대등적연결어미
52	보조동사	x	종속적연결어미
54	불완전자동사	x	보조적연결어미
56	불완전보조동사	x	관형격어말어미
58	타동사	C6	관용어



x는 code를 임의로 정할 수 있음을 나타낸다.

표 5-1 한국어 처리 전반에 적합한 사전 품사 체계

5. 결론

지금까지의 한국어 해석을 위한 기본 사전 품사 체계에 대한 연구는 각 단계에 필요한 사전 품사 체계에 대한 연구로 국한되어 이루어져 왔다. 각 단계에 대한 연구에서는 특정 품사에 대한 분류만을 했을 뿐, 각 단계간의 연계성을 고려한 분류는 행해지지 않았다.[2,5,6,7,8] 따라서 본 논문에서는 한국어 해석 과정의 각 단계에서 필요로 하는 정보를 모두 제공해 줄 수 있는 기본 사전 품사 체계에 대하여 알아보았다. 형태소 분석 단계는 형태소의 분리 및 복원 과정으로 특별히 세분화된 품사 체계를 필요로 하지 않으므로 국어학에서의 품사 체계를 기반으로 처리 효율을 높이는 관점에서 품사를 분류하였다. 구문 분석 단계에서는 형태소 분석 단계의 품사 체계를 그대로 수용하면서 문장 구조를 정확히 파악하기 위해 필요로한 정보를 고려하여 각 품사에 대한 세분화를 하였다. 그리고 구문 분석의 효율을 위해 통합형 조사에 격조사를 부여하기 위한 규칙과 용언 분류를 위한 규칙을 새롭게 정의하였다.

본 논문에서는 다양한 분야에 대한 적응성, 처리 효율성, 점진 가능성을 기반으로 하여 국어학에서의 품사 체계를 그대로 수용하면서 전산학적 관점에서 나름대로 새롭게 정의한 한국어 처리 전반에 적합한 기본 사전 품사 체계를 표 5-1과 같이 정의하였다. 이러한 품사 체계와 용언이 가지는 구문 정보, 의문 정보를 이용해 중의성이 없는 한국어 해석을 하기 위해 본 논문에서는 서술어가 가지는 구문 정보와 의미 정보를 이용하여 우좌방향 중심어 우선의 파싱 기법을 도입했다. 오른쪽에서 왼쪽으로 의미적인 분석과 통사적인 분석을 동시에 수행해 나가는 파싱 기법은 기존의 하향식이나 상향식 파싱 과정에서 여러 번 반복 되었던 backtracking을 감소시키고 조사 생략 및 보조사로 인한 격성분의 모호성 해결, 한국어 어순의 자유로움으로 인한 분석의 어려움 및 내포문의 분석을 보다 효율적으로 처리할 수 있다.

실제 자연어 처리를 위한 분석에 필요한 알고리즘 개선에 의한 성능향상이 한계에 이르렀다고 볼 때, 시스템 성능향상을 위해서는 한국어 해석에 사용되는 사전을 대폭 정리하고 보강해야 할 것이다. 또한 사전 품사 체계를 한국어 처리 전반에 적합하도록 구축하고 이를 검증해 나가는 것은 계속되어져야 할 것이다.

참고문헌

[1] 강 원석, 박 재득, 최 기선, 김 길창, "형태/구문/의미 정보의 interaction을 이용한 한국어 파서 프로토타잎의 설계 및 실험", 한국정보과학회 봄 학술발표논문집 16권 1호, PP 237-240, 1989

[2] 여 상화, 김 용호, 이 학주, 이 정현, "다단계 필터링 능력을 갖는 형태소 분석기의 설계 및 구현", 한국정보과학회 가을 학술발표논문집 18권 2호, PP 797-800, 1991

[3] 김 영식, 권 철중, 박 재득, 최 기선, 김 길창, "QMC: 사전 항목의 자질-값 도출을 위한 방법론", 한국정보과학회 봄 학술발표논문집 16권 1호, PP 245-248, 1989

[4] 강 승식, 김 영택, "사전 정보에 기반한 효율적인 한국어 형태소 분석기의 설계 및 구현", 한국정보과학회 봄 학술발표논문집 18권 1호, PP 529-532, 1991

[5] 조 성원, 송 만석, "사전에 기반한 한국어 문장 해석 시스템 원형의 연구, 한국정보과학회 가을 학술발표논문집 18권 2호, PP 809-812, 1991

[6] 윤 준태, 송 만석, "문장 분석기 및 전자 사전 구성에 대한 연구", 제 4회 한글 및 한국어 정보처리 학술발표논문집, PP 151-158, 1992

[7] 이 미선, 박 성숙, 한 성국, 최 운천, 지 민제, 이 용주, "한국어 형태소 분석기의 정형화", 한국정보과학회 봄 학술발표논문집 20권 1호, PP 777-780, 1993

[8] 김 은자, 이 종혁, "일-한 기계번역 시스템의 구현:휴리스틱을 이용한 일본어 형태소 해석 기법", 한국정보과학회 봄 학술발표논문집 20권 1호, PP 797-800, 1993

[9] 박 경환, 김 경서, 송 만석, "말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현", 한국정보과학회 봄 학술발표논문집 20권 1호, PP 801-804, 1993

[10] 윤 덕호, 김 영택, "미지문법관계 속성을 이용한 LFG에서의 한국어 문장분석 연구", 한국정보과학회 9월 16권 5호, PP 434-443, 1989

[11] 조 혁규, 박 용욱, 권 혁철, 윤 애선, "자연언어 구문분석의 비결정성 처리에 관한 연구", 한국정보과학회 가을 학술발표논문집 16권 2호, PP 575-578, 1989

[12] 양 재형, 김 영택, "KLASH: HPSG에 기반한 한국어 분석기", 한국정보과학회 가을 학술발표논문집 16권 2호, PP 587-590, 1989

[13] 조 규빈, "하이라이트 고교문법", 지학사, 1987

[14] 조 기용, 송 만석, "한국어 필수적 문장 성분을 찾는 말뭉치 분석 도구", 한국정보과학회 봄 학술발표논문집 19권 1호, PP 651-654, 1992

[15] 최 윤호, 박 혜준, 송 만석, "통사적 특징을 이용한 용례 색인기의 구현", 한국정보과학회 봄 학술발표논문집 20권 1호, PP 801-804, 1993

[16] 서영훈, 김 영택, "한국어 통사 분석기의 설계에 관한 연구", 한국정보과학회 가을 학술발표논문집 14권 2호, PP 581-584, 1987

[17] 서 영훈, 김 영택, "활성 차트를 이용한 중심어 후행 언어의 파싱", 한국정보과학회 1월 17권 1호, PP 84-90, 1990