한국어 품사 분류에 대한 제안

서 민 정
부산대학교 국어국문학과

Study on the parts-of-speech in Korean

Seo Min-Jeong
Department of Korean Language and Literature, Pusan National University

요 약

인터넷의 발달 등으로 많은 정보들이 문서화되기도 하고 그런 정보들이 공유되고 있는 지금, 언어학이나 전산학의 요구를 함께 충족시킬 수 있는 문법 모델 개발의 필요성이 극대화되고 있다. 이 글은 한국어 품사 분류에 대해서 국어학과 전산학에서의 처리 방법과 결과를 검토하고 정리하여 우리말의 특성을 잘 설명하면서도 국어를 전산 처리하는데도 도움을 줄 수 있는 품사분류를 제안하는데 그 목적이 있다. 한국어의 특성을 고려하여 음운, 형태, 통어, 의미 정보를 함께 처리할 수 있는 어휘부 중심의 문법인 HPSG의 모형을 도입하여 한국어 품사 분류를 정보 전달에 기반을 두어 자질 체계와 통합 연산을 핵심으로 기술하려고 한다. 문법기술은 주로 자질 구조를 속성과 값의 행렬인 AVM(attribute-value matrices)으로 제시할 것이다.

1. 들어가기

1.1. 자연언어는 그 양상이 무척 다양해 확실한 법칙성을 찾기가 힘들다는 특성이 있다. 이러한 특성을 가진 자연언어를 대상으로 연구하는 분야에 '언어학'과 '전산학'이 있다. 그런데 이들은 자연언어를 대상으로 한다는 점은 같으나, '자연언어'를 바라보는 관점은 다르다. 언어학에서는 언어 연구의 목적을 인간의 지적 활동 중에서 언어의 생성과 이해 과정에 대한 탐구 즉, 언어의 본질을 규명하는데 둔다. 그래서 언어 능력과 언어 행위의 과정을 기술, 형식화하는데 목적이 있다. 한편, 전산학에서는 언어는 정보를 담고 있는 도구이며, 언어행위의 과정을 이해하기 위한 자료로 보기 때문에 자료 처리가 목적이 된다.[1)]

이와 같이 언어학과 전산학에서 '언어'를 바라보는 입장이 다르기 때문에 지금까지 언어학과 전산학의 연구성과들이 서로의 연구에 크게 이용되지 못했던 것이 사실이다.

인터넷의 발달 등으로 많은 정보들이 문서화되기도 하고 그런 정보들이 공유되고 있는 지금, 언어학이나 전산학의 요구를 함께 충족시킬 수 있는 문법 모델 개발의 필요성이 극대화되고 있다. 방대한 양의 언어를 처리하기 위해 전산학에서 개발한 검색 프로그램 등을 이용한다거나, 구문 분석기나 의미 분석기 개발에 언어학의 연구성과들이 거의 직접적으로 이용된다면 두 학문의 발전을 위해서도 분명히 필요한 일일 것이다.

1.2. 어떤 학문에서든지 학문에서 다루어야 할 대상 모두를 하나하나 다루기 어려우므로 보통 어떤 기준에 의해 범주로 갈라서 그 범주의 특성이나 기능 등을 다룸으로써 대상들을 설명하게 된다. 언어학도 마찬가지인데, 언어학에서 다루어야 할 대상인 형태소나 단어, 문장 등을 범주로 나누어서 연구하게 되는 것이다.

이러한 범주 구분 가운데서 단어 범주를 구분한 '품사 분류'는 국어학과 전산학에서 각각 연구해 왔다. 언어 연구의 기본이 범주 구분이라고 한다면, 국어학에서는 당연히 국어연구의 출발에서 이 품사 분류를 했고, 전산학에서는 자연언어처리를 하기 위해서는 전자사전이 거의 필수적으로 요구되는데, 그 전

1) 서상규 · 한영균(1999:13~16)에서는 이들을 각각 '전산언어학'과 '자연언어처리'로 나누어 설명하고 있다.

자 사전을 구축하기 위해서 품사분류가 반드시 선행되어야 했던 것이다.

이 글은 이러한 품사 분류에 대한 국어학과 전산학에서의 처리 방법과 결과를 검토하고 정리하여 한국어의 특성을 잘 설명하면서도 국어를 전산 처리하는데도 도움을 줄 수 있는 품사분류를 제안하는데 그 목적이 있다.

2. 앞선 연구의 검토

2.1. 품사 분류는 국어학에서도 학자들의 연구 관점에 따라 조금씩 다른 양상을 보인다. 주로 전통문법에서 논의되었는데, 체언이나 용언, 부사, 관형사, 감탄사 등의 분류[2]는 학자들의 입장에 따라 세부 분류를 인정하느냐, 묶어서 처리하느냐 정도로 비교적 그 차이가 간단하다고 할 수 있지만, 조사나 어미의 경우는 다소 복잡한 양상을 보였다. 간단하게 (1)을 통해서 살펴보면, 다음과 같이 세 가지로 나누어 살필 수 있다.[3]

(1) 하늘만 보면, 눈물이 흐른다.

첫째, 분석적 체계이다. (1)에서 '만, 이'와 같은 조사나 '면, ㄴ, 다'와 같은 어미를 독립된 품사로 인정하는 경우로, 주시경(1910), 김두봉(1916), 김윤경(1948) 등이 여기에 해당한다.

둘째, 절충적 체계이다. (1)에서 '만, 이'와 같은 조사는 독립된 품사로 인정하지만, '면, ㄴ, 다'와 같은 어미는 독립된 품사로 인정하지 않는다. 최현배(1937), 허웅(1983)의 논의가 여기에 해당한다.

셋째, 종합적 체계이다. (1)에서 '만, 이'와 같은 조사나 '면, ㄴ, 다'와 같은 어미를 독립된 품사로 인정하지 않는 경우로, 정렬모(1946), 김민수(1957), 이숭녕(1956) 등의 논의가 여기에 해당한다.

2.2. 전산학에서는 자연언어처리의 시작 단계가 형태소 분석이기 때문에 대부분의 논의에서 품사 분류를 다루고 있다. 그러나 그것이 언어의 본질을 규명하거나 언어의 특성을 살피는 것이 목적이라기보다는 자료처리를 주된 목적으로 하기 때문에 개발 프로그램의 성격에 따라서 품사분류가 다른 모습을 가지고 있었다. 전산학에서는 어떤 응용 프로그램이냐에 따라서 품사분류가 조금씩 다름을 알 수 있었다. 전산학에서는 가능하면 더 세분화하는 것이 형태소 정보를 잘 제공할 수 있고, 형태소 해석 관점에서 애매성을 줄이고 정확한 결과를 제공할 수 있다고 보고 그렇게 해 왔다. 그러나 이 글은 그러한 세분화가 가져오는 문제점도 무시하지 못할 것이다.

전산학에서 품사분류는 시스템의 목적에 따라 다른데, 여기서는 [10]의 분류만 간단히 제시하면, 최상위 9개 부류로 기호, 외국어, 체언, 용언, 수식언, 관계언, 독립언, 어미, 접사가 있고, 이들을 다시 세분류해서 54개의 품사 태그를 설정하고 있었다.

3. 한국어의 특성과 문법 모델

3.1. 어휘부와 HPSG

잘 알다시피, 한국어는 어순이나 격표지 등에 있어서 영어와는 상당히 다른 언어이다. 따라서 영어와 같은 인구어의 분석에 많이 이용된 문법 이론을 우리 문법 분석에 그대로 이용해서는 무리가 따르는 경우가 많다. 이 글은 가능하다면 한국어의 특성을 있는 그대로 분석하려고 할 것이다. 한국어는 어순도 고정되어 있지 않고, 생략도 빈번하게 발생하며, 특히 문장 성분의 기능을 밝혀주는 표지가 나타나기도 하고 나타나지 않기도 하는 등 규칙화하기 어려운 점이 많이 있다. 따라서 이러한 언어적 특성을 가지기 때문에 전산으로 실현시키기도 무척 어려울 것 같다. 이러한 한국어의 특성을 고려해서, 가능하다면 규칙의 수는 줄이면서 필요한 정보는 다 제시할 수 있다고 생각되는 '어휘부'의 중요성이 커지게 되었다.

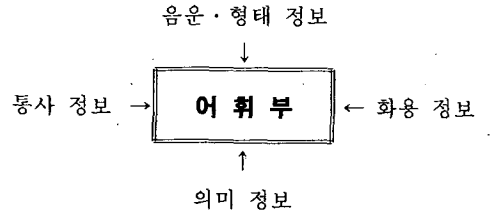
이러한 어휘부를 중시하는 문법 가운데서 HPSG[4]가 있는데, 이 글은 HPSG의 문법모형을 도입하여 한국어 기본 문법을 정보 전달에 기반을 두어 자질 체계와 통합 연산을 핵심으로 하는 문법모형을 기술하려고 한다. 문법기술은 주로 자질 구조를 속성과 값의 행렬인 AVM(attribute-value matrices)으로 제시할 것이다.

(3) 어휘부

2) 국어학에서 논의된 품사 분류에 대한 전반적인 내용은 [5]를 참고하시오.

3) 전통문법의 품사분류에 대한 자세한 내용은 [8]을 참고하시오.

4) HPSG의 핵심적 개념인 자질구조나, 통합 등에 대해서는 3.2에서 다시 다루게 될 것이다.



3.2. 자질구조와 통합

자질구조는 LFG, HPSG 등의 통합(*unification*) 문법에서 많이 사용하는 설명방법인데, 자질과 값으로 구성된다.

(4) [자질 값]

예를 들어 (5)와 같이 설명할 것을 (6)과 같이 자질 구조를 이용하여 설명하면, 좀더 도식적으로 볼 수 있다.

(5) 영이는 나이가 20세이고, 부산대학교 국어국문학과 2학년 여학생이다.

(6)

$$\begin{bmatrix} \text{이름} & \text{영이} \\ \text{나이} & 20 \\ \text{학교} & \text{부산대학교} \\ \text{학과} & \text{국어국문학과} \\ \text{학년} & 2 \\ \text{성별} & \text{여} \end{bmatrix}$$

그리고 이 자질구조는 공유된 구조를 가리키거나, 정보의 통합을 나타내기도 적당하다.[5]

따라서 여기서 기술할 한국어 문법은 언어적 정보의 표현 도구로 자질 구조를 이용할 것이다. 단어, 구, 문장 등의 언어형식은 (7)과 같은 자질 구조를 가진다.

(7)

$$\begin{bmatrix} \text{음운} & \text{값} \\ \text{통사의미} & \text{값} \\ \text{구성소} & \text{값} \end{bmatrix}$$

(7)의 자질 가운데 '구성소'는 언어 형식이 구 이상의 단위일 때 그 언어 형식을 구성하는 성분을 표시하는 것이다.

(7)에서 통사의미의 값은 다시 (8)과 같은 자질 구조를 가진다.

(8)

$$\begin{bmatrix} \text{통사의미} & \begin{bmatrix} \text{LOC} & \begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} \\ \text{SUBCAT} \\ \text{LEX} \end{bmatrix} \\ \text{CTT} & \begin{bmatrix} \text{IND} \\ \text{RESTR} \end{bmatrix} \\ \text{CXT} & [\text{BKG}\ \ ] \end{bmatrix} \\ \text{NONLOCAL} & \begin{bmatrix} \text{INHERITED}[\text{SLASH}] \\ \text{TO-BIND}\ [\text{SLASH}] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

이 글에서 다룰 품사분류에 대한 것은 [LOC | CAT | HEAD] 자질과 직접적으로 관련되고, 그 외 'SUBCAT'이나 'RESTR' 자질은 더 세분할 때 관련된다. 따라서 이 글에서는 위의 자질 가운데서 [통사의미 | LOC] 자질에 한정해서 살필 것이다.

4. 품사 분류의 검토와 제안

4.1. 조사와 어미의 처리

조사를 독립된 품사로 인정했을 때의 문제점은, 조사의 자질 자체가 독립된 품사로 인정하기 힘든 의존성을 가지고 있다는 국어학적 문제도 있고, 중심어 - 뒤 구성이라는 한국어의 어순 특성에 비추어 자연언어처리를 했을 때 조사가 중심어가 되어 조사의 자질만 남으므로, 문장 분석 후에 조사와 결합한 성분의 어휘적 의미나 통사적 자질은 사라지고 조사의 정보만 남게 된다는 것이다.

사실 조사의 문법적 지위는 국어학 안에서도 많은 논의가 있어 왔는데, 그 중에서 [9] 의 논의에 의하면, 한국어의 조사와 어미[6]는 통어론에서는 구, 마디, 월 따위의 통어적 짜임새와 결합하면서, 형태론에서는 그 앞의 말과 결합하여 최소 자립 형식을 형성하는 특성을 지녔다. 즉 조사는 형태적으로는 앞에 결합한 성분 - 주로 명사 -에 의존적이면서 통어적으로는 용언과 관련된다.

5) HPSG에서 설명하는 자질구조, 구조 공유, 통합에 대한 자세한 것은 [11]의 2장과 [12]의 1장을 참고하시오.

6) [9]에서는 조사와 어미를 묶어 '토'라고 했는데, 여기서는 아직 '토'에 대한 개념을 설명하지 않았기 때문에 '조사와 어미'로 풀어서 설명한 것이다.

(9) ㄱ. 나는 오월이 좋다.
ㄴ. 나는 오월을 좋아한다.

(10)에서 밑줄 그은 조사 '이'와 '을'의 경우를 보자. 이들은 형태적으로는 '오월'과 결합하고 있으므로 '오월+이', '오월+을'로 각각 분석할 수 있지만, 통어적으로는 (10ㄱ)의 '이'는 '좋다'와 (10ㄴ)은 '좋아하다'와 관련되기 때문에 형태적 분석과 통어적 분석이 일치하지 않으므로 다른 품사와는 성격이 다르다. 이런 점은 '어미'의 경우도 마찬가지다. 즉 그것이 형태적으로는 용언과 결합하고 있으나, 통어적으로는 문장 전체와 결합하고 있기 때문에 그것의 문법적 기능이 어떤 면에서는 조사와 많이 닮아 있다. 이러한 특성 때문에 최규수(2001)에서는 우리말의 조사와 어미는 통어론에서의 문법적 지위와 형태론에서의 문법적 지위가 달라, 통어론과 형태론 사이에서 불일치 현상을 불러일으키므로, 이들을 '접어'로 보고, 통어론과 형태론을 자율적으로 처리하여, 어휘부에서 그 정보들을 통합하는 방식으로 분석할 수 있음을 제시하였다.

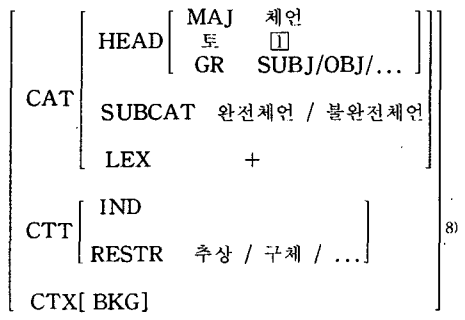
한편 조사는 보통 체언과만 결합하는 것으로 알려져 있는데, 많은 언어 자료 가운데서는 조사가 꼭 체언과만 결합하지 않는 예가 보이기 때문에 각 품사의 특성을 살필 때 이러한 점도 함께 고려해야 할 것이다.

그래서 이 글에서는 위와 같은 국어학적 문제와 전산학적 문제를 함께 고려하여, 조사와 어미를 기본적으로 독립된 품사로 인정하지 않고, 이 둘을 묶어서 '토'라고 하여 각 품사의 HEAD 자질의 값 중의 한 자질로 설정하여 다룰 것이다.

4.2. 명사, 대명사, 수사의 분류

이들은 체언의 하위부류인데, 이들은 통사적으로는 차이를 보이지 않는다. 단지 의미적으로 차이를 보일 뿐이므로 중심어 자질에서 굳이 구분할 필요가 없을 것이다. 즉 명사를 의미에 따라 추상이나 보통 등으로 구분하는 것과 명사와 대명사를 구분하는 것이 비슷하게 보이기 때문이다. 이들의 구분은 (10)에서 [CTT | RESTR][7]과 [CXT | BKG] 등의 값으로 구분할 수 있을 것이기 때문에 [CAT | HEAD | MAJ]에서는 구분하지 않고 '체언'으로 묶어 처리하는 것이 나을 것이다.

(10)



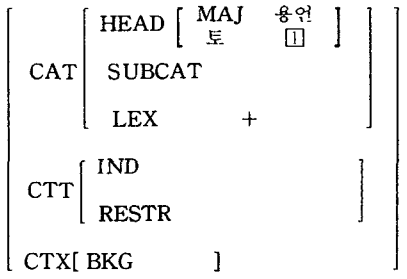
[8]

한편, 이들 체언들에는 '토'가 붙는데, 이것은 HEAD 자질에서 분석될 수 있다. 그리고 이 토의 결합으로 체언의 문법관계가 설정되므로 체언의 어휘적 의미가 사라지거나 토의 문법적 기능을 나타내지 않는 경우는 없다.

4.3. 동사, 형용사, 지정사의 분류

한국어에서 - 최소한 형태적, 통어적 특성으로는 - 동사와 형용사와 지정사의 구분은 크게 의미를 가지지 못한다. 그것은 타동사와 자동사의 차이나 자동사와 형용사의 차이가 거의 비슷하기 때문이다. 이들의 구분은 (11)에서 [CAT | SUBCAT]의 값으로 하면 적절할 것이므로 용언으로 묶어 [CAT | HEAD | MAJ]의 값은 같이 '용언'으로 해도 무리가 없을 것이다.

(11)



4.4. 부사, 관형사, 감탄사의 자질구조

7) 이 글에서 논의해야 할 부분이 아니기 때문에 여기서는 간단히 그 가능성만을 보이기 위해 '추상, 구체..'라고 제시하였는데, 더 엄밀한 하위 분류를 통해 체언의 하위부류들이 설정되야 할 것이다. 이것은 아래 다른 품사의 경우도 마찬가지다.

8) '토'에 1이라고 되어 있는 것은 지면상 자질 '토'의 '값'을 제시할 수 없어 구조 공유의 표시이다. 5장에서 이 1을 확인할 수 있다.

부사, 관형사, 감탄사 등은 불변화사에 해당되는 것으로 사실 한국어에서 굴곡의 모습이 많이 나타나는 것은 아니나, 실제 문서에서 그 예가 보이므로, 일단은 HEAD 자질에 자질 '토'를 설정해 두었다.

(12) 부사의 자질 구조

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{MAJ} & \text{부사} \\ \text{토} & \boxed{1} \end{bmatrix} \\ \text{SUBCAT} & \text{동사수식/형용사수식/부사수식} \\ \text{LEX} & + \end{bmatrix} \\ \text{CTT} & \begin{bmatrix} \text{IND} & \\ \text{RESTR} & \text{빈도/성상/...} \end{bmatrix} \\ \text{CTX} & [\text{BKG} \quad] \end{bmatrix}$$

(13) 관형사의 자질 구조

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{MAJ} & \text{관형사} \\ \text{토} & \boxed{1} \end{bmatrix} \\ \text{SUBCAT} & \\ \text{LEX} & + \end{bmatrix} \\ \text{CTT} & \begin{bmatrix} \text{IND} & \\ \text{RESTR} & \text{지시/사람/...} \end{bmatrix} \\ \text{CTX} & [\text{BKG} \quad] \end{bmatrix}$$

(14) 감탄사의 자질 구조

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \text{MAJ} & \text{감탄사} \\ \text{토} & \boxed{1} \end{bmatrix} \\ \text{SUBCAT} & \\ \text{LEX} & + \end{bmatrix} \\ \text{CTT} & \begin{bmatrix} \text{IND} & \\ \text{RESTR} & \end{bmatrix} \\ \text{CTX} & [\text{BKG} \quad] \end{bmatrix}$$

4.5. 한국어 품사 분류(MAJOR)

언어 처리를 하는데 있어서 범주 구분은 기본이 되는 작업일 것이다. 따라서 여러 응용 분야에서 다양한 용도로 사용할 수 있도록 계층적인 분류가 필요하다. 4장에서는 그러한 계층적인 분류 가운데서 가장 상위의 분류를 제시하였다. 정리하면, (15)와 같다.

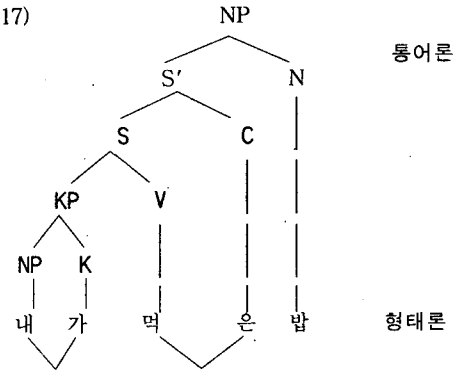
(15) 체언, 용언, 부사, 관형사, 감탄사

5. 토와 단어의 결합

'토'란 한국어의 조사와 어미를 묶어서 이르는 말이다. (16)의 '은'은 (17)에서와 같이 형태론적인 측면과 통어론적인 측면에서 보면 불일치를 보인다.

(16) 내가 먹은 밥

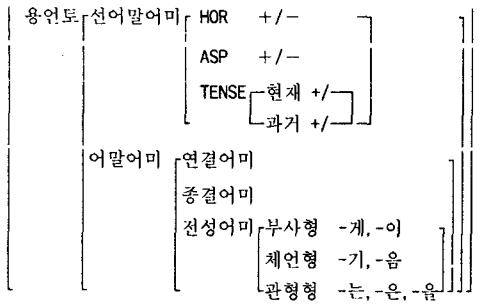
(17)



(16)에서 '먹은'의 경우를 보자. '은'은 형태적으로는 '먹'에, 통어적으로는 '내가 먹'에 결합해서 뒤에 오는 '밥'이라는 성분을 수식하게 하는 성분이다. 따라서 형태적 분석인 품사 분류에서는 용언에 결합하는 용언토라는 것을 통합을 통해 각 문장 성분이 결합하는 통어적 분석이 필요할 때는 '내가 먹'을 관형사의 기능을 하는 것으로 나타내 주어야 한다. 따라서 이러한 특성을 가진 토는 형태론적으로 분석할 때와 통어론적으로 분석할 때 각각 다르게 분석해야 한다. 이 글에서는 품사 분류가 목적이기 때문에 사실은 통어론적 분석은 크게 고려하지 않았다. 단지 토의 특성 때문에 통어론적 분석을 시사하는 정도만을 제시하였다. 토는 형태론적으론 주로 체언이나 용언뿐만 아니라 부사나 관형사와 같은 단어와 결합하는데, 이 글에서는 이러한 언어적 현실을 반영하여 각 품사가 이 토와 결합할 수 있도록 설정하였다. 그리고 자질 '토'의 값은 (17)과 같다.

(17)

$$\boxed{1}\begin{bmatrix} \text{체언토} & \begin{bmatrix} \text{격조사} & \text{이, 가/을, 를} \\ \text{보조사} & \end{bmatrix} \end{bmatrix}$$



6. 맺음말

6.1. 언어학이나 전산학의 요구를 함께 충족시킬 수 있는 문법 모델 개발의 필요성에 따라 이 글에서는 HPSG의 문법 모형을 참고로 하여 한국어 품사 분류에 대해 살폈고, 이를 자질구조로 도식화하였다.

6.2. 한국어의 특성상 '조사'와 '어미'는 형태론적 분석과 통어론적 분석이 다름을 살폈고, 따라서 이 글에서는 이들을 독립된 품사로 인정하지는 않고 '토'라고 묶어서 각 품사마다 이 토와 결합 가능하도록 설정하였다.

6.3. 이 글에서 살핀 한국어 품사분류의 상위 분류는 '체언, 용언, 부사, 관형사, 감탄사'로 설정하여 살폈다.

6.4. 이러한 논의를 통해 남은 문제는 4.5.에서 설정한 기본 품사 분류에서 다시 세밀한 부분까지 적절한 기준에 따라 하위분류를 해야 하며, 한정된 범위에서라도 이러한 품사 분류의 적절성을 검토해야 한다는 것이다.

참 고 문 헌

[1] 김두봉, ≪깁더조선말본≫, 역대문법대계 제1부 8책, 탑출판사, 1922.

[2] 김윤경, ≪나라말본≫, 역대문법대계 제1부 8책, 탑출판사, 1948

[3] 서상규 · 한영균, ≪국어정보학 입문≫, 태학사, 1999

[4] 서영훈, '의미정보를 이용하는 중심어 주도의 한국어 파싱', 서울대 컴퓨터공학과 박사논문,1991.

[5] 서정수, ≪국어문법≫, 한양대학교 출판부, 1996.

[6] 시스템공학연구소, ≪우리말 정보처리 S/W 기술개발에 관한 연구≫, 최종결과보고서, 1997.

[7] 주시경, ≪국어문법≫, 역대문법대계 제1부 4책, 탑출판사, 1910.

[8] 최규수, '주시경의 토를 다루는 방식과 그 계승', 한글 232, 한글학회, 1996.

[9] 최규수, '통어론과 형태론의 관계에 대하여', 언어과학 8-1. 동남언어학회, 2001.

[10] 한국과학기술원, ≪통합 국어정보베이스≫, 3차년도 최종보고서, 1997.

[11] Carl Pollard & Ivan A. Sag, *Information -Based Syntax and Semantax,* vol. 1, CSLI, 1987.

[12] Carl Pollard & Ivan A. Sag, *Head-Driven Phrase Structure Grammar,* CSLI, 1994.