모빌구조와 표지 개념에 의한 지식기반적 한국어 구문분석기 개발

성결대학교 어문학부
우 순 조

Developing Knowledge-Based Korean Syntactic Parser In terms of Mobile Configuration and Marker Theory

Soon-Jo Woo
Department of Language and Literature, Sungkyul University, Anyang, Korea

요 약

이 글은 활용 개념과 수형도를 근간으로 기술되어 온 한국어 문법 모델에 대한 대안으로 표지 개념과 모빌 구조를 제시하고 이를 바탕으로 개발된 한국어 구문분석기의 특성을 소개하고자 한다. 먼저, 조사와 어미를 독자적인 통사 단위인 표지로 처리함으로써 국부 구조의 통사 범주와 문법적 기능을 명확하고 일관되게 구분할 수 있으며, 모빌 구조는 한국어의(상대적) 자유 어순 현상을 효과적으로 기술할 수 있다. 이에 의거한 문법 모형은 언어학적 지식과 구문분석 엔진 사이의 독립성을 향상시킴으로써 향후 구문분석기의 성능 개선을 보다 용이하게 한다. 이 글에서 소개하는 구문분석기는 언어학자에 의해 구축된 지식을 이용한다는 점에서 지식기반적이라고 할 수 있는데 여기에는 동사의 하위범주화 정보, 첨어 유형 정보, 의미정보가 핵심적인 언어 지식으로 이용된다. 모빌 구조에 의한 구문분석은 국부 구조를 단순화함으로써 구문적 중의성을 최소화하며, 의미정보는 주어진 술어의 논항적 자격을 검증하는 기준으로 작용하여 구문적 중의성을 감소시키고 정확한 분석을 가능하게 한다.

서 론

이 글은 활용 개념과 수형도적 구조에 입각하여 기술되어 온 한국어 문법 모형에 대한 대안을 제시하고 이를 바탕으로 개발된 한국어 구문분석기의 기본적인 특성을 소개하는 것을 목표로 한다. 이 글에서 소개하는 구문분석기는 언어학자에 의해 구축된 지식을 이용한다는 점에서 지식기반적이라고 부를 수 있다.

이를 위하여 2절에서는 지식기반적 접근법을 채택하는 동기를 간략히 설명할 것이다. 3절에서는 이 글에서 소개하는 한국어 문법 모형의 기본 개념으로서 표지 개념과 모빌적 형상 개념을 소개할 것이다. 표지 개념은 통사 범주와 문법적 기능을 선명하게 구분할 수 있으므로 언어 지식과 구문분석 엔진의 독립성을 증진시키는 데에 기여한다. 또한 모빌적 형상 개념은 한국어의 자유 어순 현상을 포착할 수 있으며 구조적 중의성 감소에 크게 기여한다. 4절에서는 3절에 소개된 문법 모형을 바탕으로 개발된 한국어 구문분석기의 특성을 설명할 것이다. 이를 언급의 편의상 문법 모델의 형상적 특성을 따서 KMSP(Korean Mobile Syntactic Parser)라 부르겠다. KMSP가 이용하는 중요한 언어지식으로는 용언의 하위범주화(subcategorization) 정보, 용언의 첨어 유형(adjunct type) 정보, 의미 자질(semantic feature) 정보가 있다. 마지막으로 5절에서 글을 마무리할 것이다.

E-mail : woosj@sungkyul.edu

지식기반적 접근법의 필요성

자연언어 처리의 성패는 정확한 언어 지식을 얼마나 효과적으로 컴퓨터로 구현하느냐에 달려 있다고 할 수 있다. 특히 무한한 데이터를 대상으로 하는 구문분석기의 개발을 위해서는 적절하고 타당한 언어 지식이 확보되어야 한다.

언어 지식을 확보하는 방안은 크게 기계적 방식과 수동적 방식으로 나눌 수 있다. 기계적 방식이란 태그드 코퍼스(tagged corpus)를 구축하고 단위 표현들 사이의 천이 확률을

추출하거나 구문구조 코퍼스(parsed tree corpus)에서 국부 구조(local structure)를 추출하고 각각의 국부 구조들 사이의 통합 확률를 추출하여 이용하는 방법 등을 말한다. 수동적 방식이란 언어학적 훈련을 받은 인력이 수작업을 통해 통사 규칙이나 의미 등의 연관된 언어 지식을 기술하는 방식을 말한다. 기계적 방식은 기본적으로 확률에 의존한다는 점에서 이를 확률 기반적이라고 부를 수 있다. 반면에 수동적 방식은 토박이 화자의 언어적 직관에 의존한다는 점에서 지식 기반적이라고 할 수 있다.

이 두 가지 접근법 가운데 어느 것이 더 적합한가에 대한 판단은 개발하고자 하는 시스템이 무엇을 목표로 하느냐에 따라 달라질 수 있으나 어떤 경우이든 타당성을 평가하는 기준은 필요하다. 그 기준으로는 개발의 용이성(efficiency)과 개선 가능성(improvability)를 생각할 수 있다.

개발의 용이성이란 성능 대비 개발 비용(performance per cost)의 측면과 개발에 걸리는 시간으로 요약할 수 있다(여기서 성능이란 입력 자료 대비 처리 시간이라는 일반적 의미를 포함하여 자료의 처리 범위 및 정확도까지를 포함하는 개념으로 쓴다.). 아무리 좋은 방안이더라도 그것을 개발하는 데에 지나치게 시간이 많이 걸리거나 많은 비용이 소요된다면 현실적인 방안으로 보기 어려울 것이다.

다음으로 고려되어야 하는 기준은 개선 가능성(improvability)이다. 이 기준은 다시 처리 범위(coverage)와 정확도(precision)로 나누어 생각할 수 있다.[1] 여기서 처리 범위란 주어진 자료 가운데 시스템이 처리할 수 있는 양을 말한다. 그리고 정확도란 시스템이 처리한 자료들 가운데 올바른 분석이 차지하는 비율을 말한다.

문법의 궁극적인 목표가 '자연언어가 허용하는 모든, 그리고 오직 문법적인 문장들을 기술하는 것'인 만큼 구문분석기의 궁극적인 목표 역시 '자연언어가 허용하는 모든, 그리고 오직 문법적인 문장들을 처리하는 것'이 되어야 한다. 이러한 관점에서 보자면 일정 수준의 처리 범위와 정확도에 도달하는 데에 어떤 접근법이 더 효과적이냐를 따지는 것은 무의미하다. 어떤 접근법을 채택하든지 초기 모델은 일정한 한계 안에 머물 수밖에 없으므로 원형(prototype)이 개발된 이후에는 개선의 문제가 중요한 이슈로 떠오르게 된다.

구문분석이 주어진 단위 표현과 이들 사이의 통합 관계를 포착하는 작업이라는 관점에서 볼 때, 개선 가능성의 핵심은 어휘와 문장 구조(또는 패턴)에 대한 대처 능력이 된다.

어휘와 관련된 정보는 학습을 통해서만 습득할 수 있으므로 확률기반적 접근법이나 지식기반적 접근법이나 별반 차이가 없다고도 할 수 있다.[2] 그러나 무한한 문장을 분석 대상으로 가지는 구문 분석에 있어서 확률기반적 접근법과 지식기반적 접근법은 개선 가능성이라는 기준에서 볼 때 큰 차이를 보인다.

확률기반적 접근법은 확률 추출에 사용되는 코퍼스에 의존하므로 데이터 영역에 따라 국부구조의 확률이 달라질 수 있으며, 이러한 이유에서 처리 능력이 영역 의존적(domain dependent)이라고 할 수 있는데, 이는 곧 사용 범위가 특정 영역으로 제한될 수 있다는 것을 의미한다. 보다 근본적인 문제는 코퍼스에서 추출한 확률은 사람의 손으로 조작할 수 없는 지식이라는 점에서 개선 가능성이 보장되지 않는다는 사실이다. 또한 확률의 속성상 아무리 코퍼스의 양을 증가시킨다고 하더라도 일정 규모 이상이 되면 확률은 안정된다는 한계를 안고 있다.[3]

이와는 대조적으로 지식기반적 접근법에 따라 구축되는 지식은 동일 언어의 모든 영역에 적용될 수 있다는 점에서 범영역적이라고 할 수 있으며, 인간의 언어적 직관에 기반하기 때문에 관리와 수정이 가능하다. 또한 확률기반적 접근법이(낮은 정확도와 개선 가능성에도 불구하고) 모든 분야의 자료를 처리할 수 있기 위해서는 어휘/통사적 특성을 공유하는 분야마다 새로이 코퍼스를 구축해야 하는 부담이 따른다. 이러한 측면을 감안하자면 궁극적으로는 지식기반적 접근법이 더 경제적일 수도 있다.[4]

1 이 두 기준은 상충적(conflicting)이다. 일반적으로 정확도를 높이려면 처리 범위가 줄어 들고, 처리범위를 늘리려면 정확도가 희생된다.

2 이는 어휘적 지식이 공시적으로 유한하다는 사실을 뜻한다. 그런데 형태소 분석과 구문 분석을 나누는 것은 편의에 의한 것일 뿐, 엄밀한 의미에서 이들을 구분하는 것이 쉽지는 않다. 형태소 분석은 크게 후보 형태소 구조 생성 단계와 적합 구조 선택 단계로 구분되며, 후자를 흔히 태깅이라 부르는데, 구문 구조를 배제한 상태에서 완전한 태깅이 이루어질 수는 없기 때문이다.

3 지금까지 개발된 한국어 구문분석기들은 크게 실험적 차원에서거나 기계번역과 같은 응용 프로그램의 모듈로서 개발되어 왔는데, 이들은 대개 확률기반적 방식에 의존하고 있다고 해도 과언이 아니다. 이러한 시스템들의 성능은 무작위 자료를 대상으로 했을 때에 대개 60%를 밑도는 것으로 알려져 있다.
사용 빈도가 높은 규칙을 파싱에 우선적으로 적용하자는 Bod(1998)의 제안은 속도의 개선이라는 측면에서는 고려할 만하나 정확도와는 직접적인 관련이 없다.

4 현재 상용화된 시스템들 가운데 Phrase structure rule로 구성된 문법을 탑재한 경우에 문법이 일정한 규모에 이르게 되면 유지와 보수 및 재사용에 어려움이 따르며, 규칙의 수가 커짐에 따라 실행 속도가 현저하게 감소한다는 평가가 있다.(Cole, Ronald et al., 1988, p.109). 그러나 이는 어떤 언어 모델을 사용하느냐에 따라 달리 평가될 수 있는 문제로서, 모든 지식기반적 접근법의 한계라고 단언할 수는 없다. 이 글에서 소개하는 모빌적 형상 개념에 의한 구문분석 시스템은 적은 수의 국부구조 규칙과 매우 제한된 수의 국부구조 통합 규칙들로 이루어지므로 문법이 매우 간결하다는 장점을 가진다. 하드웨어 기술 개발의 속도로 볼 때 실행 속도의 문제나 메모리 및 저장 장치에 소요되는 비용의 문제는 멀지 않은 미래에 해소되리라 기대된다.

한국어 문법 모형

한국어의 통사론에서 근간이 되는 부분은 교착형태소의 분포와 기능이라는 문제와 국부구조에서 구성 성분들 사이의 자리바꿈이 가능하다는(상대적) 자유어순 현상이다. 따라서 교착 형태소들의 기능과 분포 및 술어의 직접 구성 성분들이 보이는 자유 어순을 온전하게 포착해 낼 수 있는 적절하고도 타당한 문법 모델을 개발하는 것은 한국어 구문분석기 개발을 위한 대전제가 된다.

1. 교착형태소의 통사성

1) 기존 모델 검토

주시경(1910) 이래로 오랜 국어학적 전통에서 소위 조사로 통칭되는 교착형태소들은 어휘성을 인정받아 왔다. 반면에 최현배(1937)의 영향으로 동사의 끝바꿈은 소위 어미라는 형태론적 단위로 처리되어 왔다. 이는 교착형태소를 이질적으로 기술한다는 점에서 기술의 일관성을 결한 방식이다.

1980년대 말부터의 Chomsky적 생성문법의 분석은 이와 상반된 양상을 보인다. Pollock(1989)의 제안 이후 굴절범주의 통사성을 인정하려는 움직임이 호응을 얻으면서 한국어의 어미를 통사단위로 처리하는 분석이 지배적이다. 그럼에도 조사에 대해서는 여전히 격(case) 개념을 적용시켜 이를 형태론적으로 처리한다. 이러한 양상은 기술의 내용이 반전되었을 뿐이며, 비일관적 기술이라는 점에서 보면 전통 국어학적 입장과 동일한 문제를 안고 있다고 할 수 있다. 오히려 생성문법적 분석은 격조사에 대해서는 형태론적 단위로, 그밖의 사격성(obliqueness)을 표상하는 조사에 대해서는 후치사(postposition)라는 용어를 써서 이질적으로 기술하고 있는데 이는 분포와 기능이 동질적인 단일한 범주를 이론적 요구에 따라 분할하여 처리한다는 점에서 더 바람직하지 못하다.

조사의 어휘성은 널리 인정되므로 별다른 논의없이 이를 전제하기로 하고, 어미의 형태/통사적 지위를 결정하는 문제에 대해 간략히 언급하자면 다음과 같다.

2) 어미의 통사적 지위

구조 기술과 관련하여 어미를 형태론적 단위, 즉 어휘의 일부로 볼 때에 가장 문제가 되는 것은 이 개념이 명사형 어미, 관형형 어미, 부사형 어미의 경우에 필연적으로 범주미결정(categorical indeterminacy)의 문제를 야기한다는 것이다. 다음의 예를 보자.

(1) ① 철수가 이 일을 하였다.
② [철수가 이 일을 하-기]가 더 낫겠다.
③ [철수가 이 일을 하-는] 경우를 상상해 보자.
④ [철수가 이 일을 하-게] 네가 도와 주거라.

(1)의 ①는 동사 '하-'가 전형적인 서술어로 쓰인 예이다. ②, ③, ④는 동사 '하-'의 활용형으로서 각기 명사형, 관형형 그리고 부사형으로 불리운다. 이렇게 불리우는 이유를 살펴보면 먼저 (1②)의 경우는 명사류와 결합하는 조사가 부착되기 때문이고, (1③)는 영어의 형용사처럼 뒤따라 오는 명사를 수식/한정하기 때문이다. (1④)는 '-게' 형 자체가 문장의 필수성분이 아니고 마치 부사처럼 수의적 성분으로 기능한다는 이유에서이다.

(1)의 꼴바꿈을 활용으로 부른다는 것은 이들이 모두 동일한 어휘의 상이한 형태라는 판단을 반영한다. 그런데 주어진 언어 단위의(어휘) 범주를 규정하는 작업은 언제나 분포와 기능적 특성을 참조한다. '새 집'의 '새'에 관형사라는 범주를 할당하는 것은 이 형태소가 항상 명사를 선행하며 명사를 한정하기 때문인 것이다. 이러한 논리에서 (1)②-④의 '하-'는 그 통합 관계를 볼 때 동사이기도 하면서 각기 명사, 관형사, 부사이기도 하다고 해야 할 것이다. 이러한 현상을 두고 국어학에서는 소위 두자격법이라 일컫는데, 이 용어는 바로 이러한 판단을 반영하는 것이다.

그런데 사실의 인식과 설명은 별개라는 점을 유념할 필요가 있다. 두자격법이라는 용어는 단순히 문제의식을 반영할 뿐이며 결코 이 문제에 대한 해결책이 될 수는 없는 것이다. 하나의 어휘가 한 문장 안에서 두 가지 범주적 지위를 가진다는 분석은 필연적으로 범주 결정의 문제를 야기하게 된다.

예를 들어 (1②)의 '하기'를 명사로 보면 왜 명사가 동사의 논항을 이끄는지를 설명하기 어렵다. 반대로 '하기'를 동사로 분석하는 경우는 '철수가'라는 주어와 '이 일을'이라는 목적어를 이끄는 현상을 손쉽게 설명할 수 있으나, 왜 동사가 문장을 완성하는 데에 그치지 않고 문장의 성분으로서 기능하는지를 설명할 수 없다. 또한 이러한 분석은 한 문장 안에서 주어진 언어 단위는 반드시 하나의 범주로서만 기능할 뿐이라는 범언어적 관찰과도 상치된다. 아래 영어의 예를 보자.

(2) ① The enemy invaded the country.
② The enemy' s invasion of the country.
(3) ① *The enemy' s invaded of the country.
② *The enemy invasion the country.

(2)의 ②는 'invade'의 명사형이 핵어가 된 구성이다. 주어가 소유격을 취하고 목적어는 전치사에 이끌리어 핵어에

통합된다. 반면에 동사형에 명사형 논항이 통합된 (3①)나 명사형에 동사형 논항이 통합된 (3②)는 모두 비문이 된다. 이 자료들이 시사하는 바는 핵어의 어휘 범주에 따라 통합의 대상이 결정되며 한 어휘가 한 문장 속에서 복수의 범주로서 기능할 수 없다는 것이다.

이러한 사실은 다음과 같은 일반성을 반영하는 것이다. 즉, 형태론적 구조는 통사 규칙에 대하여 폐쇄적이라는 사실이다. 다시 말해서 어휘의 일부가 어휘 외부의 성분들과 통사적 관계를 맺을 수 없다는 말이다. (3②)에서 'invasion'이 'invade'라는 동사와 '-ion'이라는 명사형 어미의 결합형으로서 그 안에 동사를 포함하고 있을지라도 그 동사가 형태론적 층위를 벗어나서 동사의 논항으로서의 두 명사구, 즉 주어와 목적어를 거느릴 수 없는 것이다. 역으로 이러한 이유에서 'invade+ion'의 구성은 형태론적 구성이며, 통합의 결과인 'invasion' 전체는 명사라는 어휘 범주에 귀속되는 것이다.[5]

두자격법이 안고 있는 딜렘마는 통사 단위 판별에 있어서 '최소 자립 형식'만을 어휘로 보는 구조주의적 기준을 무비판적으로 수용한 결과이다. 그러나 통념과는 달리 한국어의 어미가 독자적인 통사 단위임을 지지하는 강력한 증거들이 존재한다. 자세한 논의는 우순조(1997, 2002)로 미루고 여기서는 어미가 동사 어간을 넘어서서 동사와 그 보어(complement)들로 이루어지는 동사구 전체와 통합됨을 보여주는 예를 제시하겠다.[6]

(9) ① 체중이 늘었다/줄었다.

② 입영 대상자들이 체중을 늘이었다/줄이었다.

(10) ① 배우들은 체중을 늘이거나 줄이-었-다.

② *배우들은 체중을 [늘-ø-거나 줄]-이-었-다.

(11) ① He outrun-s and outswim-s his rival.

② *He [outrun- ø and outswim]-s his rival.

(9①)와 (9②)에서 알 수 있듯이, '늘-'과 '줄-'은 일항 술어인데 반하여, 이들에 소위 파생접사 '-이'가 통합된 '늘이-'와 '줄이-'는 이항 술어로서 전혀 다른 어휘가 된다. 다시 말해서, '늘'과 '이'의 결합은 어휘적 구성을 이루는 것이다. 이러한 까닭에 '늘이'와 '줄이'가 대등 접속을 이룬 (10①)에서 대등접속항 모두에 파생접사 '이'가 등장하더라도 이들이 생략될 수 없으며, 생략되는 경우는 (10②)에서처럼 비문이 된다. 이는 (11①)의 'outruns'와 'outswims'에서 삼인칭 단수 현재 어미가 대등 접속되더라도 어느 하나가 생략되지 않으며, 생략될 경우 (11②)처럼 비문이 되는 현상과 평행하다.

이와는 달리, 한국어 어미는 파생접사와는 판이한 양상을 보인다. 시제 형태소의 분포를 통해 이를 확인해 보자.

(12) ① 허리를 늘이고, 기장을 줄이었다.

② 허리를 늘이었고, 기장을 줄이었다.

(13) ① [...늘이-ø]-고 [...줄이]-었-다.

② *[[...늘-ø-ø]-고 [...줄]]-이-었-다.

(13) ① old [man and woman]

② old man and old woman

(12a)는 (12b)와 같이 바꾸어 쓸 수 있는데, 이는 (12a)에서 '늘이고'의 '었'이 뒤따르는 대등접속항에 시제 형태소 '었'의 존재로 인하여 생략되었음을 암시한다. 이러한 현상은 (13a)가 (13b)와 같은 의미로 쓰일 수 있는 현상과 근본적으로 같다.

만일 기존의 분석에 따라 소위 어미로 통칭되는 형태소들을 형태론적 구성의 단위로 본다면 (13b)에 표시한 바와 같이 대등 접속 구성에서 보이는 파생 접사 '이'와 시제 형태소 '었'의 차이는 설명되지 않는다. 형태론적 구성의 경계밖에 있는 성분과 관계를 맺는 단위는 곧 통사적 단위인 것이다.

2. 어미의 범주

통사 단위가 확정되고 나면 주어진 통사 단위가 다른 통사 단위들과 맺는 의존 관계, 즉 분포를 기술하여야 하며, 그리고 그 때 주어진 통사 단위가 담당하는 언어 내적인 기능은 무엇인지를 규명하는 작업이 필요하다.

이 글은 우순조(1997, 1998, 2000, 2001)를 좇아 한국어 어미가 술어를 핵어로 하여 이루어지는 술어구에 통합된다고 보고, 이들을 통칭하여 표지(marker)라 부른다.[7] 이를 약식으로 표시하자면 아래와 같다.

(14) ① [[[그 배우가 몸무게를 늘이 V']-었 VPe]-다 VP].[8]

5 형태 층위와 통사 층위를 명확히 구분해야 한다는 이와 같이 주장은 Chomsky(1970)에 의해서 제안되었으며, 이후 Di Sciullo and Williams(1988)의 단원성 명제(atomicity thesis)로 이어진다.

6 우순조(1997)은 Postal(1969)와 Simpson(1995)에서 제안된 통사 단위 판별 기준을 정리한 Bresnan & Mchombo(1995)의 기준들을 한국어 어미에 적용시킨 것으로서, 한국어 어미가 독자적인 통사 단위임을 보여준다.

7 통사 단위들 간의 통합에 있어서 기준이 되는 어휘 범주를 핵어라고 한다. Pollock(1988)의 제안을 좇아 한국어 어미를 핵어로 분석하는 많은 논의들이 개진되고 있으나, 개별 언어의 기술은 철저히 자료에 입각해서 이루어져야 한다.
조사의 핵성(headness)에 대한 논의로는 임동훈(1991)이 있으며, 어미를 핵으로 보는 분석의 문제점에 대한 비판으로는 우순조(2002)를 참고하기 바란다.

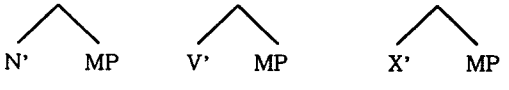
8 여기서 V',Vep, S 등과 같은 phrasal category 들은 projection level을 구분하기 위하여 편의상 사용되었다. 따라서 이들이 적극적인 언어학적 의미를 가지는 것으로 해석되어서는 안 된다.

② [[[마음이 따뜻한] 사람 N']-이 NP] 좋다.
(VP=S)

(14①)의 V'는 동사 '늘이'와 '늘이'의 하위범주들-주어와 목적어-이 함께 동사구를 형성함을 나타낸 것이다. (14②)의 N'는 핵어 '사람'과 그 의존 성분인 관계절이 통합되어 하나의 성분을 이루는 것을 나타낸다. 동사구에 어말어미가 통합되면 최대 투사 범주인 VP가 형성되고 명사구에 소위 격조사가 통합되면 역시 최대 투사 범주인 NP를 형성한다. 동사구의 최대 투사 범주를 어말어미 통합형으로 보는 중요한 근거는 한국어에서 어말어미가 문법적 기능을 표상하고, 이에 따라 통사적 의존 관계(=분포)가 결정되기 때문이다.

이러한 분석을 통해 우리는 그 동안 구분되지 않았던 통사 범주와 문법적 기능의 문제를 명확히 구분하여 다룰 수 있게 되며, 언어학적으로 의의있는 일반화를 도출해 낼 수 있다.

(15) ① NP ② VP ③ XP



(X = N, V, A, Adv
MP = (Sequence of) Case Marker or
(Sequence of) Verbal Endings)

즉, (15)에서 나타낸 바와 같이 MP(marker phrase) 왼쪽의 범주는 통사 범주를 나타내며, 여기에 부착된 MP에 의해 주어진 통사 범주의 문법적 기능과 그에 따른 분포가 결정되는 것이다.

앞서서 활용 개념에 따라 동사의 활용형을 기술하자면 필연적으로 '두자격법'이라는 용어로 지칭되는 범주적 미결정성의 문제가 파생된다는 것을 지적한 바 있다. 이 문제 역시 위의 분석에 따름으로써 간단히 해소된다. 어간과 어미가 별개의 통사 단위인 까닭에 어간인 동사는 언제나 자신의 하위 범주들과 통합되어 동사구를 형성한다. 그리고 이렇게 형성된 동사구의 최종적 분포는 여기에 부착되는 표지에 따라 결정되는 것이다.

3. 국부 구조의 형상(Configuration)

한국어 어미의 분포를 (14)와 같이 규정하면 동사구의 내부 구조를 어떻게 규정하느냐 하는 문제가 남는다.

문장의 구조 기술에 있어서 고려할 사항은 문장의 직접 구성 성분들 사이의 통사적 의존 관계를 포착하는 한편, 각각의 성분들이 담당하는 문법적 기능을 규명하는 일이다.

(15) ① S → NP VP
② VP → V NP

문제를 단순화해서 표현하자면, 생성 문법에 입각한 분석에서는 모든 문장이 (15①)와 같은 구조를 가지는 것으로 가정된다. 그리고 타동 구성(transitive construction)의 생성을 위해서는 (15②)와 같은 규칙이 상정된다. 이러한 규칙에 따르면 주어는 항상 문장을 직접 구성하는 명사구가 되고, 직접목적어는 동사구를 직접 구성하는 명사구가 될 것이다. 이 점에 착안하여 Chomsky(1965)는 문법적 기능을 형상(configuration)에 입각하여 정의하는 방안을 제안한 바 있다.

그러나 이는 단순한 제안으로서 경험적 자료에 의하여 검증되어야 할 문제이지, 자체로서 어떤 타당성도 보장되지 않는다. 주어진 문장의 구성 성분들이 어떤 형상을 이루느냐의 문제와 각각의 성분들이 어떤 문법적 기능을 가지느냐는 문제는 서로 별개이다.[9]

이 글은 문법적 기능이 형상에 의하여 이차적으로 규정될 수 있다는 Chomsky의 제안을 거부하고, 주어진 술어에 의하여 하위범주화되는 성분들은 각기 고유한 문법적 기능을 가진다고 본다.[10] 그리고 한국어에서 하위범주화되는 성분들은 술어와 함께 하나의 구를 형성하며 이들은 핵어를 선행해야 한다는 어순 규정을 따른다고 본다. 예컨대 (14①)의 '늘이'를 GPSG(1985)의 ID-LP format에 따라 표시하자면 아래 (16)과 같다.

(16) ① V' ← NP(subj), NP(obj), V (= H)
② NP(subj) < H, NP(obj) < H.

(16①)에서 '←'는 아래에서 소개할 구문분석기의 bottom-up parsing 방식을 나타낸다. Category label 뒤의 (subj), (obj)는 각 성분이 담당하는 문법적 기능의 명칭이다.[11] (16②)의 '<' 기호는 기호 좌측의 성분이 기호 우측에 있는 성분을 선행한다는 것을 의미한다. (16②)에서 보어들은 핵어와의 선후 관계만이 정의되어 있으므로 주어와 목적어는 어떤 순서로 배열되더라도 허용된다는 것을 의미한다. (어순이 고정된 경우에는 이를 LP statement에 명시하면 된다.) 따라서 자유어순이 자연스럽게 포착될 수 있다.

9 이러한 입장의 극단에 서 있는 분석으로 Larson(1988)을 들 수 있다. 그는 양분지 규칙만을 인정하고 형상적 위치에 따라 문법적 기능을 규정할 것을 제안하였다. 그러나 이러한 분석 방식은 영어에 존재하는 어순 변이 현상을 설명하는 데에 한계를 가질 수밖에 없다. 영어 전치사구의 자유 어순에 관해서는 Whitman(1979)를 참고할 것. 그리고 영어 동사구의 구조와 관련하여 McCauley(1982)의 논의를 참고할 것.

10 Frazier(1987)과 Rayner et al.(1984) 등은 thematic relation(대체로 이 글에서 말하는 하위범주화 내역)만이 구문분석기나 담화 모델 및 인간의 세계 지식과 공유되는 유일한 vocabulary라고 주장한 바 있다.

11 규칙에는 표시되지 않았으나, 각 성분에는 핵어(명사)가 가지는 의미자질과 함께, 각 성분의 문법적 기능을 표상하는 조사의 형태가 명시된다고 본다.

이러한 방식에 따르면 구문분석은 술어가 취하는 하위범주화 내역과 각각의 성분의 구성을 규정하는 국부구조 규칙들만으로 이루어질 수 있다.

KMSP

1. 알고리즘 개요

앞서서 소개한 한국어 문법 모델에 입각한 지식기반적 한국어 구문분석기의 알고리듬을 chart-parsing 방식에 따라 설명하자면 아래와 같다. 편의상 입력 문장의 어절별 tagging이 정확하게 이루어진 것으로 가정한다.

(17) KMSP 약식 알고리듬

```
ParseInput{
        init_Chart;
        FindTag{
                If(Tag eq S){ break;}
                ApplyRule(FindTag){
                If(meet XP){ ApplySubcat;
                        If(not match){
                            next;}}
                FindRule;
                CheckRule(FindRule){
                        If(match){IntoChart;
                        next;}}
        }
}
```

위에서 소개한 약식 알고리듬은 입력된 통사 단위(synt-actic unit)의 개수만큼 차트를 구성하고 각 통사단위별로 적용할 규칙을 검색하여 확인하는 면에 있어서는 일반적인 chart-parsing 방법과 동일하다.[12]

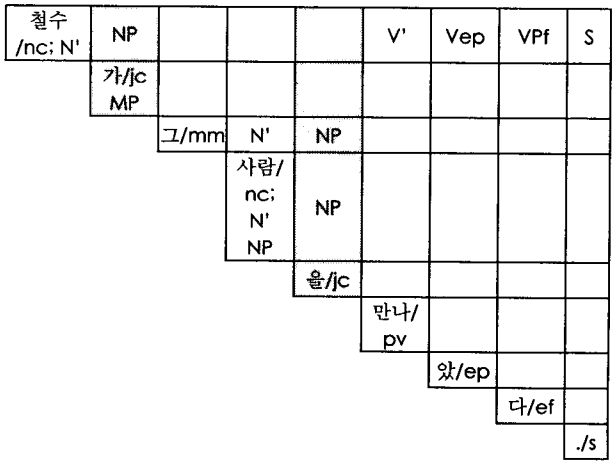
(17)에서 설명을 요하는 부분은 ApplyRule이라는 함수인데, 이 함수는 개념상 두 부류로 나뉜다. 제일 먼저 적용되는 규칙들은 아래 (18)과 같은 규칙들로서 이들은 술어에 의하여 하위범주화되는 성분들을 형성하는 국부구조 규칙들이다.

(18) ① N' ← nc
② N' ← mm n
③ NP ← N' MP
④ NP ← N'
⑤ MP ← jx*jc
⑥ MP ← jcjx*
(* : Kleene closure)

12 Parsing 방법에 관해서는 King(1997)을 참조할 것.

Chart에 국부구조 규칙들이 적용되어 최대 투사 범주(XP=NP, AdvP etc.)들이 형성되면 ApplySubcat이라는 함수가 활성화된다. 최대 투사 범주들은 순서에 상관없이 하위범주화 내역에 등재되어 있으면 술어와 통합하여 V' 를 형성한다. 이 때 국부구조에 미리 정의된 문법적 기능(grammatical role)이 할당된다.

(19) ① 철수가 그 사람을 만났다.
②



2. 표지 개념의 효과

표지 개념에 따라 조사와 어미를 통사단위로 분석하는 방식은 구문분석에 있어서 적용될 규칙의 종류와 적용 범위를 분할하는 효과를 발휘한다.

예를 들어 명사구의 경우에, 규칙(18③)에서 보이는 바와 같이, 국부 구조는 조사를 제외한 나머지 성분들로만 구성되며 여기에 조사가 부착될 때에 비로소 최대 투사 범주를 형성할 수 있게 된다. 이는(어말) 어미의 경우도 마찬가지이다. 따라서 조사와 어미를 통사 단위로 처리하는 분석은 구조적 중의성의 증가를 사전에 억제하는 효과를 가지며, 대등접속구성에서 나타나는 다양한 조합 가능성을 규칙으로 통제할 수 있다.

맺 음 말

자연언어처리는 컴퓨터에게 말을 가르치는 것이라 할 수 있다. 단순히 작동되는 시스템이 아니라 실용화될 수 있는 시스템의 개발은 적절하고 타당한 문법 모형에 기초할 개선을 위한 논의의 토대가 마련된다고 할 수 있다. 시스템의 성능 향상을 위해서는 공학적 관점에서의 탐구가 불가결하다. 언어, 컴퓨터, 인지 등 유관 분야 전문가들 사이의 학제적 협력이 요청된다고 하겠다.

REFERENCES

우순조(1994) : 한국어의 형상성과 관계표지의 실현 양상, 서울대학교 대학원 언어학과 박사학위논문

우순조(1997) : 국어어미의 통사적 지위, 국어학 *30*

우순조(1998) : 모빌구조와 표지이론에 의한 한국어 통사/의미 기술, 언어학 *28*집, 한국언어학회

우순조(2002) : *Syntactic Parsing and the Syntax of Verbal Endings in Korean,* 언어학 *34*호

임동훈(1991) : 격조사는 핵인가. 주시경학보 *8*

주시경(1910) : 대한국어문법, 한국문법대계, 탑출판사

최현배(1937) : 우리말본, 정음사

Bod, Rens(1998) : *Beyond Grammar,* 이강혁 역(2000) : 경진문화사, 서울

Bresnan & Mchombo(1995) : *The Lexical Integrity Principle : Evidence from Bantu. Natural Language and Linguistic Theory 13*

Cole, Ronald et al.(1998) : *Survey of the State of the Art in Human Language Technology, Cambridge University Press.*

Chomsky N(1965) : *Aspects of the Theory of Syntax, MIT Press, Massachusetts*

Di Sciullo, Anna-Maria and E. Williams(1989) : *On the Definition of Word, The MIT Press.*

Frazier L(1987) : *Syntactic Processing : Evidence from Dutch, Natural Language and Linguistic Theory 5 : 519-560*

King M(1997) : *Parsing Natural Language, Academic Press.*

Larson RK(1988) : *On the Double Object Construction, Linguistic Inquiry 19, pp335-392*

McCauley J(1982) : *Parentheticals and Discontinuous Constituent Structure, Linguistic Inquiry 13 : 91-106*

Pollock, Jean-Yves(1989) : *Verb Movement, Universal Grammar, and the Structure of IP, Linguistic Inquiry 20 : 365-424*

Postal P(1969) : *'Anaphoric Islands', in Robert I. Binnck, Alice Davison, Gerogeia M. Green and Jerry L. Morgan (eds.), CLS 5*

Rayner K, Carson M and Frazeir L(1984) : *The Interaction of syntax and Semantics in sentence Processing : Eye Movements in the An-alysis of semantically Biased Sentences. Journal of Verbal Learning and Verbal Behavior 22 : 358-374*

Simpson J(1995) : *Walpiri Morpho-syntax : A Lexicalist Approach, Kluer, Dordrecht*

Whitman J(1979) : *Scrambling, Over-easy or Sunny-side-up? Proceedings of Chicago Linguistics Society 15*