# 제약기반 이론에 기반한 한국어 분석기 구축

(Building a Korean Parser with a Constraint-based Grammar)

김종복 · 양재형*

**요약**

제약기반(constraint-based) 언어이론에 근거하여 언어학적으로 타당하며 동시에 컴퓨터 구현의 실현성과 효율성을 갖춘 한국어구구조문법을 개발하고, 이를 LKB(Linguistic Knowledge Building) 시스템을 이용하여 구현하였다. 핵어중심구구조문법(Head-driven phrase structure grammar)의 문법적 틀 안에서, 유형화된 자질구조의 다중상속 위계에 근거하여 문법을 설계하였으며, 이러한 접근의 장점과 효율성을 한국어의 복합범주 현상을 중심으로 제시하였다. SERI 문법현상평가문을 이용한 실험의 결과는, 이론적으로 타당한 동시에 계산적으로도 실현성 있는 한국어 문법의 개발이 가능함을 보여주고 있다고 판단된다.

## 1. 서론

1990년대 이후의 자연언어처리 연구들은 대규모 언어자료를 이용하는 자동화된 학습 방법에 근거한 통계적 기법이 중심이 되어 왔다[Manning99]. 구문 분석에 있어서도 확률적 파싱 기법[Magerman95, Collins97]이라든가 청킹(chunking) 기법에 근거한 '얕은'(shallow) 분석을 시도함으로써[CoNLL00], 포착하는 언어 정보의 깊이가 얕은 대

* 경희대학교 영어학부 · 강남대학교 컴퓨터미디어공학부

신 견고하고 효율적인 시스템을 추구하였다. 이러한 연구들은 어느 정도의 성과를 거두어 몇몇 응용 영역에 유용한 결과물들을 산출한 것으로 평가할 수 있으나, 필연적으로 80년대 이후 전산언어학 분야에서 폭넓게 연구되어 온 엄밀한 전산적 문법이론의 연구 성과들로부터 멀어지는 계기가 되기도 했다.

한편 이러한 '얕은' 분석 방법만으로는 기계번역이나 대화번역과 같은 깊은 구문-의미적 지식을 필요로 하는 응용을 만족시킬 수 없다는 인식에 따라, 최근에는 다시 HPSG(Head-Driven Phrase Structure Grammar), LFG(Lexical Functional Grammar)나 LTAG (Lexicalized Tree Adjoining Grammar)와 같은 엄밀한 언어이론과 이에 기반한 제약기반적(constraint-based) 언어 처리에 대한 새로운 관심도 높아지고 있다[Oepen02]. 궁극적으로는 이러한 이론적 접근과 경험론적/통계적인 접근을 통합하여야 할 것으로 보이며, 일부에서는 그 필요성과 방법에 대한 심층적인 논의가 진행되고 있다 [Sparck Jones00].

한국어의 구문 분석을 시도한 기존의 연구들은 주로 특정 응용 영역을 고려한 특화된 방식이거나[김나리96, 정천영01], 분석 효율을 중요한 척도로 도입하는 공학적 접근 방법이었다[이현영99, 김미영00, 김광백02]. 포괄적인 시질기반 한국어 문법 및 이에 기반한 한국어 분석기를 제안한 [박소영99]의 연구도 엄밀한 기술적, 설명적 타당성을 갖는 언어이론에 근거하여 전개된 것은 아니다. 이와 같이 한국어 분석에 관한 기존의 연구들은 대체로 공학적인 측면에서 접근이 이루어져 어느 정도의 적용율(coverage) 및 효율성을 확보하는 데 성공하였지만, 언어학의 연구 결과들과 무관하게 개발자의 언어적 직관에

만 의존하거나 언어학의 연구 성과들을 단편적으로만 도입한 경우가 대부분이었고, 결과적으로 한국어에 대한 이론적 타당성을 갖춘 적절한 전산적 모형을 제시하지 못하였다.

본 연구는 제약기반 언어이론에 근거하여 언어학적으로 타당하며 컴퓨터 구현이 용이한 한국어구구조문법(Korean Phrase Structure Grammar, 이하 KPSG)의 개발 및 이에 기반한 분석기의 구현을 목표로 하였다. 한국어 문법을 위한 문법적 틀로는 핵어중심구구조문법(HPSG, Head-driven Phrase Structure Grammar)[Pollard94, Sag99]을 채택하였는데, HPSG는 문법에 대해 제약기반적 및 어휘중심적 접근을 취하며 언어를 유형화된 자질구조(typed feature structure)에 대한 제약의 시스템으로 모델링하고자 하는 이론이다. HPSG는 또한 어휘, 통사, 의미 등 여러 계층의 언어 정보를 통합적, 평행적으로 취급하여 이들 정보가 하나의 단일한 계층에서 상호 제약적으로 작용한다는 사실을 강조한다. 따라서 이러한 여러 종류의 정보들이 동일한 표현 형식에 의해 일관성 있게 통합적으로 기술되어야 한다는 점에서, 형태소분석, 구문분석, 의미분석 등을 서로 독립된 실행 단계로 간주하는 한국어 분석에 대한 기존의 여러 연구들과 구별된다.

문법의 컴퓨터 구현을 위한 도구로는 LKB(Linguistic Knowledge Building) 시스템[Copestake02]을 사용하였다.[1] LKB 시스템은 HPSG와 같은 제약기반 언어 이론의 구현을 위한 환경으로 Stanford 대학 CSLI(Center for the Study of Language and Information) 연구소에서 개발되었으며, 유형화된 자질 구조에 근거한 문법기술을 지원한

1) http://lingo.stanford.edu/에서 다운로드 가능함.

다. 현재 영어를 비롯한 여러 언어의 문법 구현을 위해 활발히 사용되고 있는데, 특히 영어에 대해서는 ERG(English Resource Grammar)라는 대규모 문법 개발에 이용되었으며 전자우편 자동응답 프로그램과 같은 상용 시스템을 개발하는 데 사용되어 그 유용성을 입증한 바 있다[Oepen02].

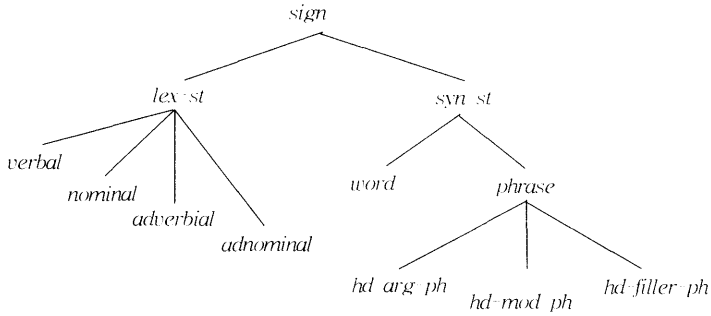
## 2. 한국어구구조문법

### 2.1 유형화된 자질구조와 유형 계층

제약기반 이론 중 대표적인 이론인 HPSG는 수학적 명시성과 전산언어학적 응용성을 문법기술의 척도로 채택하고 있으며, 언어를 유형화된 자질구조에 관한 제약 체계로 설명하려고 한다. 따라서 문법적 기술 대상이 되는 언어 정보가 유형화된 자질구조의 형태로 표현되는데, 이때 표현의 잉여성을 최소화하기 위해 유형의 계층(type hierarchy)을 설정하고 계층 하위의 유형이 상위 유형의 정보를 상속(inheritance)하도록 한다.

본 연구가 제안하는 KPSG는 유형 및 유형 계층의 정의, 어휘부, 어휘규칙, 문법규칙 등의 주요 구성 요소로 이루어진다. KPSG에서 모든 언어 정보는 *sign*의 하위 유형으로 표현되는데, *sign*의 유형위계의 상층부 일부를 단순화시켜 보이면 (1)과 같다.

(1)




어휘부의 기본 요소들을 이루는 *lex-st*(lexical structure)의 하위 유형들은 어휘규칙 등의 어휘적 프로세스에 의해 만들어지고, 이러한 어휘 요소들 가운데 일부분은 *word* 유형으로 실현되어 통사적 요소들로 기능하게 된다. *word* 유형으로부터 투사되는 구절들이 한국어의 기본적인 정형 구절(well-formed phrase)들을 이루게 되는데, 예를 들면 핵어(head)와 그 논항(argument)들로 이루어진 *hd-arg-ph* 나 핵어와 그 수식어(modifier)로 이루어진 *hd-mod-ph* 등이다. 한국어 문법은 이와 같이 언어의 기술에 필요한 유형들과 유형들 상호간의 계층 관계를 정의하며, 각 유형들의 정형성(well-formedness)에 대한 제약을 기술함으로써 언어 현상을 모형화하는 것이다.

## 2.2 어휘부의 구성

한국어의 어휘 구성에 있어서 동사의 예를 들면, 동사는 활용어미가 붙지 않은 상태로는 독립된 어휘로 통사부에서 사용되지 못한다. 또한 어미들이 임의로 어간에 붙지 못하고 정해진 순서에 따라 구성

된다. 따라서 흔히 한국어의 동사 형태소 형성 과정을 (2)와 같은 형틀(template)에 의해 나타내게 된다[Cho95].

(2) v-base + (passive/causative) + (honorific) + (tense) + mood + (comp)

즉, 동사 어미들은 특정한 순서에 따라 어간에 붙어 존칭, 시제, 서법 등의 기능을 담당한다. 많은 어미들이 생략 가능하지만 서법을 나타내는 어말어미는 반드시 나타나야 한다. 위의 형틀은 이러한 어미의 필수성과 더불어 엄격한 순서 제약을 나타내고 있다. 그러나 위의 형틀만으로는 (3)과 같은 틀린 어휘구성을 차단할 수 없다.

(3) a. *가+시+었+자
b. *가+시+었+(어)라

청유형 어미 '-자'와 명령형 어미 '-(어)라'는 시제어미와는 결합하지 못하는 특성을 갖고 있는데, 이는 이러한 어미들이 여타 어말어미들과는 다른 어휘형성 제약을 갖고 있음을 뜻한다. 따라서 이를 반영하기 위해서는 추가적인 장치가 필요하며, 이는 (2)와 같은 형틀에 기반한 접근 방법이 한국어 어휘 형성에 대한 적절한 모형이 되지 못함을 뜻한다.

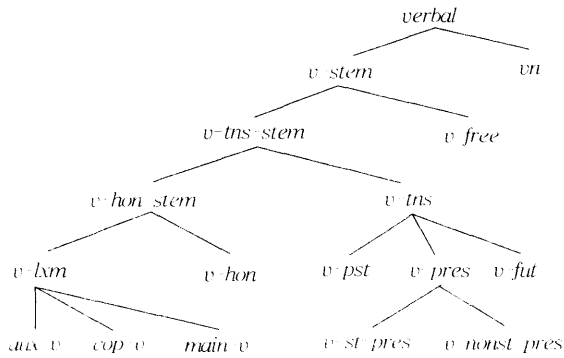
KPSG는 한국어 어휘의 형성을 유형화된 자질 구조에 기반한 유형 위계 및 유형들 간의 상호 제약 시스템으로 나타내고 있다. 어휘부 구조의 출발점은 품사의 분류라 할 수 있다. (1)에서 보인 바와 같이 KPSG에서는 전통적인 분류 방법과 크게 다르지 않게 어휘 요소

를 용언(verbal), 체언(nominal), 부사어(adverbial), 관형어(adnominal)로 구분한다.

### 2.2.1 용언 형성

KPSG에서 용언의 위계 구조의 일부를 간략히 살펴보면 아래 (4)와 같다[Kim98].

(4)



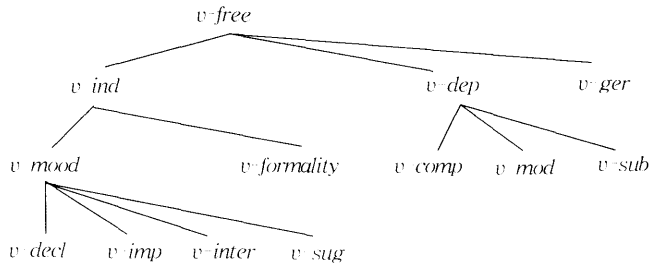

이와 같은 유형 위계는 한국어 용언 형성의 기본적 과정을 포착하기 위한 것이다. 굴절 언어인 한국어의 용언이 생성되는 과정은 엄격한 순차적 제약을 갖고 있으며, 통사부에 나타나기 위해서는 *v-free* 유형 단계까지 만들어져야 한다.

(5) 잡 + (히) + (시) + (었) + 다 + 고

*v-lxm* → *v-hon* (*v-hon-stem*) → *v-tns* (*v-tns-stem*) → *v-free* (*v-stem*) → *v-comp*

예를 들어 어휘소(lexeme) '잡'이 통사부에 나타날 수 있는 자격을 가지는 어말 어휘가 되기 위해서는 최소한 *v-free* 유형 단계에까지 가야 하는데, 그 전에 존칭접사나 시제접사 등과 결합하는 과정을 거칠 수도 있고 어말어미 뒤에 다시 보문자(complementizer) '고'가 붙을 수도 있다. (5)에 이러한 용언 형성의 각 단계에 거치게 되는 유형 위계 상의 유형들을 함께 보이고 있다. 통사부에 나타날 수 있는 유형 *v-free*의 종류를 다시 살펴보면 (6)과 같다.

(6)




통사부에 나타날 수 있는 용언은 먼저 주절의 용언이 될 수 있는 '먹다, 먹자'와 같은 *v-ind*(ependent), 종속절에 사용되는 '먹니, 먹어'와 같은 *v-dep*(endent)로 구분된다. 독립절의 용언들은 서법(mood)과 형식성(formality)에 따라 다시 하위 분류되는데, (6)에는 단순화하여 표시하고 있다. 각 유형들은 관련된 제약을 갖는데, LKB로 구현한 이들 유형 제약의 몇 가지 예를 보이면 (7)과 같다.[2]

2) LKB의 TDL(type description language)에 따라 기술된 것으로 자질기반 문법의 일반적인 표기와 유사하다. #로 표기된 것은 공지시(coindexing) 관계를 나타내는 인덱스이다.

```
(7) v-ind := v-free &
    [ SYN #syn & [ HEAD [ IC +,
                          NOMINAL -,
                          MOD <> ] ],
      SEM #sem,
      ARGS < [ SYN #syn, SEM #sem ] > ].

    v-decl := v-mood &
    [ SYN.HEAD.MOOD decl,
      SEM.MODE statement,
      ARGS < v-tns-stem > ].

    v-sug := v-mood &
    [ SYN.HEAD.MOOD suggest,
      SEM.MODE propose,
      ARGS < v-hon-stem > ].
```

ARGS는 유형에 규칙이 적용될 때 자식(daughter)들을 지칭하는 리스트 자질이다. 예를 들어 *v-decl* 유형은 *v-tns-stem* 유형에 (8)과 같은 적절한 어휘규칙이 적용됨으로써 얻어질 수 있다.[3]

```
(8) v-decl_irule :=
    %suffix (* -다)
    v-decl.
```

물론 모든 유형들은 계층상 상위 유형들이 갖는 모든 정보를 상속한다. 따라서 유형 *v-decl*에 해당하는 실제 개체는 *v-mood*와 *v-ind*

3) 이 경우에 대한 실제 어휘규칙은 형식성에 따른 세부 분류가 이루어지므로 본문에 기술된 것과 약간 다르다.

등의 유형 제약을 모두 상속하게 된다. 이러한 제약들이 상호작용하여 한국어 용언의 어말어미를 적절하게 형성할 수 있게 해준다. 특히 (7)에서 서술형 종결용언인 *v-decl* 유형은 *v-tns-stem*으로부터 구성되는 반면에 청유형 종결용언인 *v-sug*는 *v-hon-stem*으로부터 구성된다는 사실이 기술되어 있고, 이로써 '가+시+었+다'가 문제없이 구성될 수 있는 반면에 '*가+시+었+자'와 같은 용언은 차단되는 것이다.

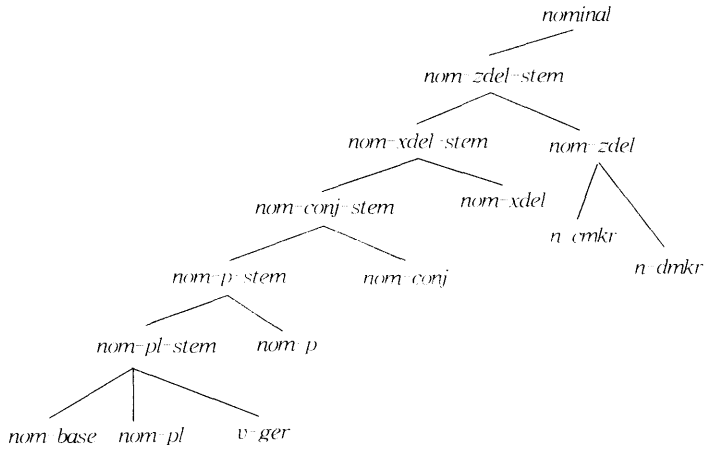
### 2.2.2 체언 형성

체언의 형성 과정도 용언의 형성과 근본적인 면에서 큰 차이가 없다. 용언의 어미와 달리 체언의 어미들은 생략될 수 있지만, 엄격한 순서를 지킨다는 점에서는 동일하다. (9)는 이러한 속성을 반영한 전통적인 체언 형성 형틀이며 (10)은 이러한 틀에 따라 형성한 예이다.

(9) n-base + (Hon) + (Plural) + (Postp) + (Conj) + (X-del) + (Z-del)
(10) 선생+(님)+(들)+(에게)+(만)+(은)
(n-base+Hon+Plural+Postp+X-del+Z-del)

이들 어미들은 나양한 문법적 기능을 표시하며, 전통적으로는 흔히 하나의 독립된 단어인 조사로 취급되어 왔다. 그러나 이들 조사가 혼자 사용될 수 없다는 점이나 여러 음운적 현상에서 독립적 단어보다는 접미사 성격을 더 많이 보여주고 있다는 점에서, KPSG에서는 어휘주의[Cho95,Kim98]의 입장을 따라 체언 어미를 용언 어미와 동일하게 어휘부에서 형성되는 것으로 분석한다. 이를 위해 체언의 유형을 다음과 같은 계층으로 세분화한다.

(11)




위 위계구조에 따르면 체언 형성은 *nom-base*에서 시작한다. 이 유형은 다시 *vn*(verbal noun), *n-bn*(bound noun), *n-cn*(common noun), *n-cl*(classifier), *n-prop*(proper noun) 등으로 하위분류되지만 그림에 보이지 않았다. 체언 형성은 위의 위계구조를 지키면서 단계적으로 일어나는데, 각 유형에 주어진 제약들이 유형들 간의 순서를 실질적으로 부여하는 효과를 가진다. (12)에 이들 몇몇 체언 어간들이 가지는 제약의 일부를 예시하였다. 용언의 경우와 마찬가지로 각 상위 유형의 형성은 적절한 어휘규칙의 적용에 의해 매개된다. 용언이 통사부에 나타나기 위해서는 *v-free* 유형이어야 하지만, 체언은 위의 모든 유형들이 통사부에 나타날 수 있다는 점에서 차이가 있다.

(12) nom-pl := nom-pl-stem &
[ SYN #syn & [ HEAD.PLU +],
SEM #sem,

```
    ARGS < nom-base & [ SYN #syn, SEM #sem ] > ].

nom-p := nom-p-stem &
[ SYN...
  SEM...
  ARGS < nom-pl-stem & [ ... ] > ].

nom-xdel := nom-xdel-stem &
[ SYN...
  SEM...
  ARGS < nom-conj-stem & [ ... ] > ].
```

이러한 제약들은 아래와 같은 비정형의 체언 생성을 방지하는 효과를 가져온다.

(13) a. *[ [*nom-zdel stem* 선생님+들+은]+에게]
b. *[ [*nom conj stem* 선생님+등+과]+에게]

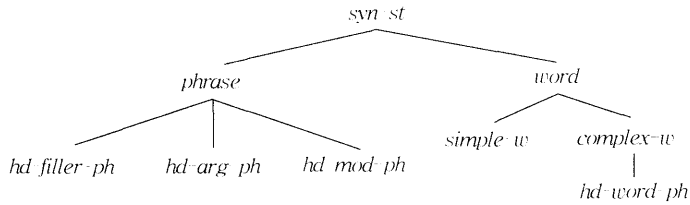
위 제약에서 전통적으로 후치사로 간주되는 '에게'가 첨가된 *nom-p* 유형은 어간의 유형이 *nom-pl-stem*이어야 하도록 명시되어 있고, (13)의 예에서 나타난 *nom-zdel-stem*, *nom-conj-stem* 등은 위계상 *nom-pl-stem* 혹은 그 하위 유형으로 간주될 수 없으므로 (13)과 같은 체언 형성이 차단된다.

지금까지 살펴 본 것처럼 어휘부의 유형화된 위계 구조는 어미 사이의 어순 제약을 포착할 수 있을 뿐만 아니라, 체계적으로 용언 및 체언을 생성할 수 있게 해준다.

## 2.3 통사부의 구성

어휘부와 마찬가지로 정형의 통사 구조도 유형에 따라 구(phrase)와 단어(word)로 대별된다.

(14)




위의 위계 구조에서 구 유형은 한국어에서의 정형의 구조를 나타내는 X-bar 구조를 의미한다. 즉, 한국어가 허용하는 구들의 종류를 의미하며 간단하게 나타내면 (15)와 같다.

(15) Korean X-bar Syntax (simplified) :
a. hd-arg-ph: XP → 1, H[ARG-ST <...1...>]
b. hd-mod-ph: XP → [MOD 1], H[1]
c. hd-filler-ph: XP → 1, H[GAP 1]
d. hd-word-ph: X[LEX +] →[word], H

구체적인 예와 이들 구들이 가지는 제약들을 자질구조로 나타내면 다음과 같다.

아래 (16)은 논항과 이를 선택하는 핵어로 이루어진 구(head-argument phrase)에 대한 스키마이다. ARG-ST(argument structure)

는 논항들의 목록을 나타내는 하위범주화 자질이며, (16)은 핵어가 요구하는 논항 중의 한 요소와 핵어가 결합하여 정형의 구를 형성할 수 있다는 것을 의미한다. 물론 이때 만들어진 구의 ARG-ST 값은 핵어의 ARG-ST 목록(□□)에서 결합된 항목(□□)을 제거한 것이 된다.

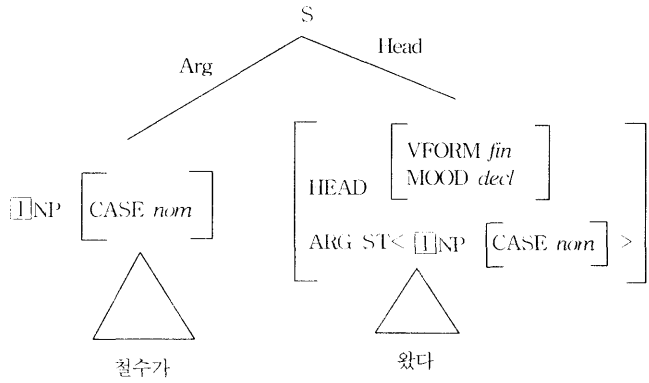
(16) *hd-arg-ph* ⇒ [ SYN.ARG-ST [A] ⊖ <[1]>
NON-HD-DTR [1]
HEAD-DTR [ SYN.ARG-ST [A]<...,[1],...> ]

이러한 통사부의 결합 과정이 LKB에서는 유형 위계와 문법 규칙에 의해 분리되어 기술되므로 (16)의 정의와는 약간 다른 형태를 가지나, 그 본질적인 내용은 동일하다. (17)은 LKB 내에 구현된 *hd-arg-ph*의 유형 정의인데, *headed-ph*와 *binary-pg*의 하위 유형으로 설정되어 이들로부터 제약을 상속한다. 그리고 (18)은 이 구에 해당하는 문법규칙의 한 예인데, 핵어의 ARG-ST의 첫 항목이 논항과 통합되어 *hd-arg-ph*를 형성하는 규칙이다.

```
(17) hd-arg-ph := headed-ph & binary-pg &
     [ ARGS < phrase & [ SYN [ HEAD.TOPIC -,
                               ARG-ST <> ] ],
              syn-st > ].
(18) head-arg-rule-1 := hd-arg-ph &
     [ SYN.VAL.ARG-ST #2,
       ARGS < #1 & [ SYN.HEAD.PRD - ],
              [ SYN.ARG-ST [ FIRST #1,
                             REST  #2 ] ] > ].
```

이러한 정의가 (19)와 같은 통사부 요소의 결합을 허용한다.

(19)




수식어와 핵어로 이루어진 구(head-modifier phrase)에 해당하는 *hd-mod-ph* 유형에 부여되는 기본적인 제약은 핵어를 수식하는 수식어구가 이 핵어와 결합하면 정형의 구를 이룬다는 것이다. 이를 위해 자질 MOD가 사용된다.

$$\text{(20) hd-mod-ph} \Rightarrow \begin{bmatrix} \text{NON-HD-DTR} \left[ \text{SYN.HEAD.MOD } \boxed{1} \right] \\ \text{HEAD-DTR } \boxed{1} \end{bmatrix}$$

충전소와 핵어로 이루어진 구(head-filler phrase)에 해당하는 *hd-filler-ph*는 주제화나 관계화 구문에서와 같이 충전소(filler)와 공소(gap)가 인접한 위치에 있지 않은 경우의 구를 생성하기 위한 유형이다. 이 경우 한 공소를 포함하고 있는 핵어가 이 공소와 동일한 통사, 의미적 정보를 가진 충전소와 결합했을 경우에 정형의 구가 된다

는 것을 의미한다. 자질 GAP은 공소를 나타낸다.

$$(21)\ \textit{hd-filler-ph} \Rightarrow \begin{bmatrix} \text{SYN.VAL.GAP} < > \\ \text{NON-HD-DTR } \boxed{1} \\ \text{HEAD-DTR } [\text{SYN.VAL.GAP} < \boxed{1} >] \end{bmatrix}$$

한국어에서는 영어와 달리 두 어휘 요소가 결합하여 구보다는 하위인 하나의 통사적 단위를 이루는 경우가 있다. 이러한 단위들은 "먹지 않았다"와 같은 복합술어 구문 등에서 쉽게 찾아볼 수 있는데, (22)는 이와 같은 어휘요소와 핵어로 이루어진 구(head-word phrase)에 대한 개략적 제약을 보여주고 있다.[4]

$$(22)\ \textit{hd-word-ph} \Rightarrow \begin{bmatrix} \text{SYN.LEX} + \\ \text{NON-HD-DTR } \textit{word} \\ \text{HEAD-DTR } \textit{word} \end{bmatrix}$$

## 3. 다중상속과 복합범주(mixed category)

앞장에서 전개한 유형화된 자질 체계에 의한 문법 기술의 중요한 특성 중의 하나로, 유형 위계상의 다중 상속(multiple inheritance)을 허용한다는 점을 들 수 있다. 이러한 유형의 다중 상속 위계는 한국어의 경동사(light verb)나 동명사와 같은 복합범주 현상을 직접적이고 효율적으로 포착할 수 있게 해 준다.

4) KPSG 내에서 복합술어 구문의 처리를 위한 자세한 사항은 [김종복(03)]을 참조.

## 3.1 동명사

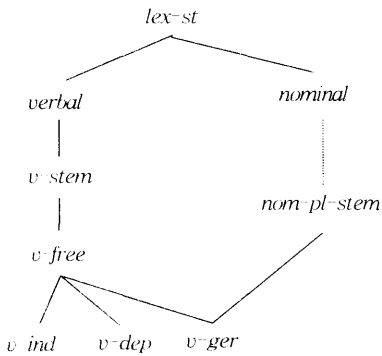
한국어에서 (23)의 '읽었음'과 같은 동명사구(verbal gerundive phrase, VGP)는 내부적으로는 동사의 특성을 보이지만 외부적으로는 명사의 특성을 보인다는 점에서 많은 주목을 받아 왔고 분석상의 어려움을 야기했다.

(23) 철수가 어제 그 책을/*의 읽었음이 확실하다.

내부적인 동사 특성은 동사 '읽다'와 마찬가지로 주어 및 목적어 논항을 취한다는 점에서 쉽게 알아 볼 수 있다. 이외에도 "책을 빨리 읽었음", "책을 안 읽었음", "책을 읽고 싶음", "그 책을 철수가 읽었음" 등과 같은 예에서 볼 수 있듯이, 부사어의 수식, 문장 부정어 '안'의 사용, 조동사의 사용, 어순의 변동성 등의 현상에서 동사 특성을 보이는 점을 알 수 있다. 한편 동명사구가 주어나 목적어, 혹은 후치사구의 위치에 나타날 수 있다는 점에서 동명사구의 외부 구조는 명사구와 유사하다.

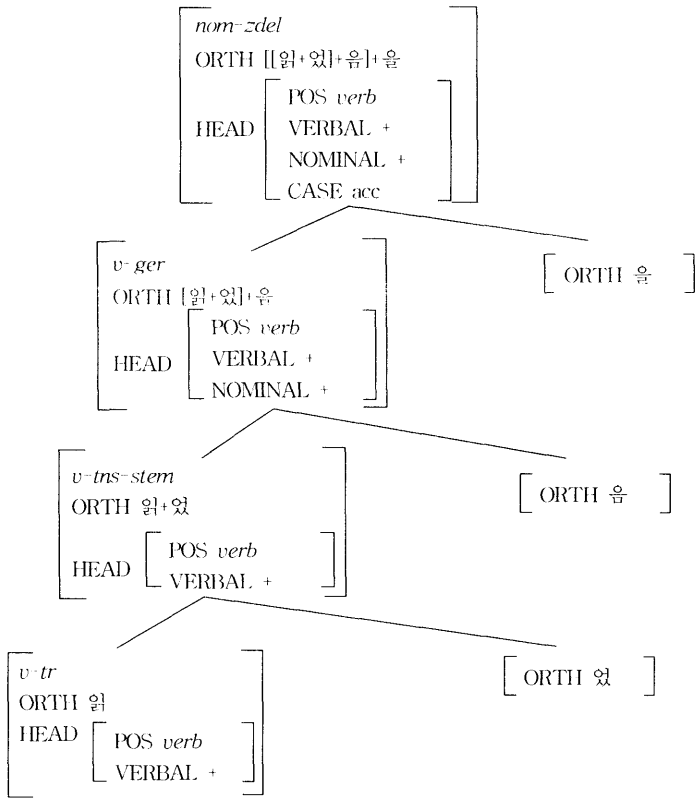
이러한 복합범주 현상은 통사분석 상의 어려움을 야기하여 기왕의 여러 시도들이 내연성(endocentricity), 어휘주의(lexicalism), 공범주 허가(null licensing) 등의 근본적인 이론적 원리를 포기하거나 변경하게 되는 경우가 많았다[Lapointe93, Yoon89, Malouf98]. KPSG에서는 유형 위계상의 다중상속 메카니즘을 이용하여, 동명사에 해당하는 유형 *v-ger*을 *v-free*와 *nom-pl-stem*의 하위 유형으로 설정함으로써 이 문제를 해결한다.

(24)




동시에 *v-free* 및 *nom-pl-stem*의 하위 유형이 된다는 것은 *v-ger*이 양쪽 모두로부터 특성을 상속함을 뜻한다. *v-free*의 하위 유형으로서 *v-ger*은 논항을 선택하고 거기에 격을 할당하는 등 동사와 같이 행동한다. 동시에 *nom-pl-stem*의 하위 유형으로서 *v-ger*은 명사와 동일한 굴절 과정을 거칠 수 있다. 이러한 유형 위계를 이용하면 '읽었음을'을 몇 가지의 어휘 규칙을 거쳐 생성할 수 있고, 이 과정을 약간 단순화시켜 보이면 (25)와 같다.

(25)




ORTH는 표층 발화의 스트링 형태에 대응하는 자질이며, CASE는 격 자질이다. 먼저 타동사 어간(v-tr)에서 과거시제 접사 '-었'이 부착되어 *v-tns-stem* 유형을 이루고, 여기에 명사화 접사 '-음'이 부착되어 동명사 *v-ger*을 이룬다. 이때 *nominal*의 하위 유형으로서 [NOMINAL +] 자질과 함께 여러 명사적 특성을 상속받게 되어 이후 후치사나 조사와의 결합을 허용하게 된다. *v-ger* 유형은 동사성(VERBAL) 및 명사성(NOMINAL) 자질이 모두 +가 되어 복합범주

임을 보이고 있다.

이와 같은 처리는 동명사를 단순히 동사나 명사로 보는 다른 방식에 비해 명백한 장점을 갖는다. 즉, 만약 이들을 동사로 간주한다면 명사적 굴절과정을 허용하기 위해 임의적인 장치를 두어야 할 것이며, 만약 명사로 간주한다면 이들이 시제나 존칭접사에 의해 동사와 동일하게 굴절되는 현상을 설명하기 어렵게 된다. 적절히 선언된 유형 자질과 함께 자질 구조의 다중상속 유형위계를 이용함으로써 이와 같은 문제를 피할 수 있다.

### 3.2 서술명사와 경동사

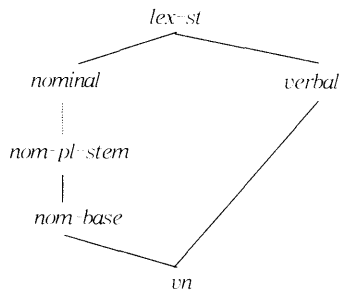
한국어의 서술명사(verbal noun, VN)들도 명사와 동사의 특성을 동시에 갖는 것으로 보인다. '연구'나 '수출'과 같은 VN들은 (26)의 예문에 보인 바와 같이 스스로 조사와 결합한다거나('연구+를') 소유격 논항과 결합하는("물리학의+연구") 점 등에서 명사적 성질을 갖지만, 논항을 선택하고 격을 할당하는 등의 동사적 성질도 갖는다. 가령 (26c)에서 '미국에'를 선택하는 것은 '수출'이며 이는 명백히 동사적 성질이라고 볼 수 있다.

(26) a. 철수가 물리학을/의 연구를 하였다.
b. 철수가 물리학을 연구 (중)
c. 철수가 그 제품을 미국에 수출(을) 하였다.

앞 절의 동명사 처리에서와 마찬가지로 다중상속 위계를 이용하여 이러한 복합범주 현상을 적절히 처리할 수 있다. KPSG에서 서술명

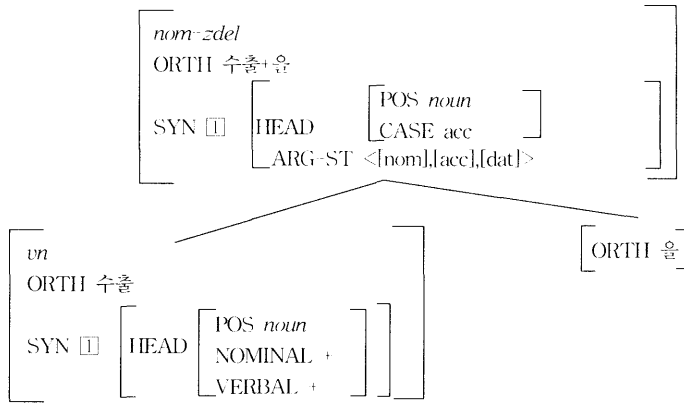
사 *vn*은 *nom-base*와 *verbal*의 하위 유형으로 정의된다.

(27)




따라서 *vn*은 *nom-base*와 *verbal* 양자로부터 속성을 상속받으므로 복합범주 특성을 띠게 된다. 예를 들어 KPSG 내에서 '수출을'의 자질 구조는 (28)과 같이 구성될 것이다.

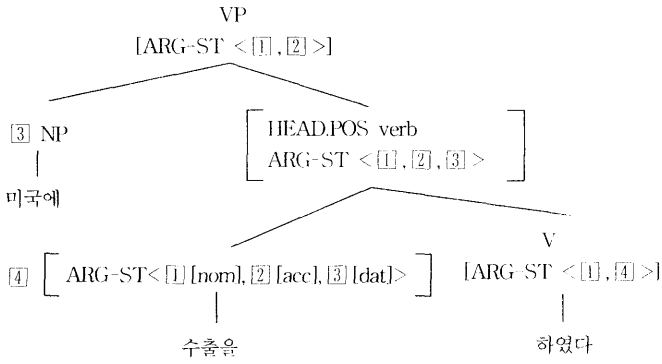
(28)




*vn*은 *nominal*의 하위 유형으로서 [NOMINAL +]를 상속받고,

*verbal*의 하위 유형으로서 [VERBAL +] 및 논항목록 ARG-ST를 상속받는다. 이러한 서술명사가 경동사(light verb) '하다'와 결합하게 되면 *hd-word-ph* 유형의 구조를 이루게 되는데, 이때 논항합성(argument composition) 원리에 따라 서술명사와 경동사의 논항이 합성되고 나서 이후의 통사적 결합과정이 이루어진다. 이와 같은 분석에 따라 KPSG 내에서 "미국에 수출을 하였다"가 결합되는 과정을 간략화된 수형도로 보이면 (29)와 같다. 주어와 서술명사를 논항으로 요구하는 경동사 '하다'가 서술명사 '수출을'과 결합하면서 두 성분의 논항이 합성되는데, 이 경우에 그 결과는 '수출'의 논항목록과 동일하게 된다.[5]

(29)




5) *hd word-ph*에 대해서는 본 논문의 (14)와 (22)를 참조할 것. 서술명사와 경동사의 결합을 위한 논항합성 원리는 본질적으로 집합연산이므로 공지시되어 있는 주격 논항은 하나만 남는다. 서술명사와 경동사의 결합에는 조동사 복합술어 구문의 처리를 위해 설정된 유형 및 규칙이 그대로 사용된다([김종복(03)] 참조).

## 4. 구현 및 평가

지금까지 설명한 바의 유형화된 자질구조에 근거한 한국어구구조 문법을 구현하기 위해 LKB 시스템을 이용하였다. 현재 한국어의 기본 구문에 대한 분석과 구현이 완료되었는데, 구현된 문법은 360여개의 유형 정의, 80여개의 어휘규칙, 30개의 문법규칙으로 구성되어 있다. LKB 시스템은 한국어에 대한 처리를 제공하고 있지 않으므로 형태소 분석을 위해 전단부에 MACH 시스템[심광섭04]을 통합하였다.[6]

KPSG의 평가를 위해 SERI Test Suite 97의 평가문 모음[성원경97]을 적용해 보았다. 이 평가문은 한국어 구문분석기가 포착해야 할 기본적이고 핵심적인 문법현상들을 포함하는 292개의 문법현상 평가문과 술어의 유형을 포괄하는 술어구문 평가문 180문장으로 이루어져 있으며, 구문분석기의 성능평가를 염두에 두고 개발되었다. 각 문장의 길이는 평균 4.1단어/문장, 9.2형태소/문장 정도로, 주로 짧은 단문 위주로 이루어져 있다. 현재 단계의 KPSG를 적용하여 분석한 결과는 표 1과 같았으며, 문장당 평균 트리수(mean parses)는 2.87개 정도였다. 이 결과는 완전히 독립된 검증 집합에 의한 엄밀한 성능측정이라고 할 수는 없으나, 현 단계의 KPSG가 한국어 주요 구문을 비교적 성공적으로 해결하고 있음을 보여준다.

---

6) MACH는 완전한 형태의 형태소분석기이지만, 본 연구는 형태, 통사, 의미 등 언어정보의 제 단계에 대해 동일한 원리에 의한 통합적 처리를 지향하기 때문에 MACH의 결과는 형태소 분할 및 원형복원까지만 이용되었다.

| Total | Parsed | Coverage |
|---|---|---|
| 472 | 423 | 89.6% |

표 1 평가 결과

## 5. 결론

본 연구는 핵어중심구구조문법의 이론적 틀 안에서 유형화된 자질구조에 근거하여 이론적 타당성 및 컴퓨터 구현을 위한 실현성을 동시에 만족하는 한국어구구조문법을 전개하였으며, 이를 LKB 시스템을 이용하여 구현하였다. 어휘부와 통사부 등 문법의 주요 구성 요소들을 다중상속 특성을 갖는 유형위계를 중심으로 설명하였으며, 특히 복합범주 현상의 예를 들어 이러한 유형화된 자질구조에 근거한 접근이 갖는 장점을 제시하였다.

구현된 KPSG 시스템에 대한 성능 평가는 현 단계의 KPSG가 한국어의 기본적인 현상들을 비교적 충실하게 해결하고 있음을 보여준다. 따라서 학문연계적인 접근을 통해 이론적 타당성과 구현의 실현성을 동시에 만족하는 한국어문법의 개발이 가능함을 보여주고 있다고 판단된다. 앞으로의 연구과제는 한국어의 언어현상들을 보다 포괄적으로 처리할 수 있도록 문법을 확장함과 동시에, 장문의 입력에서 발생하는 과도한 트리 발생을 억제하도록 문법을 다듬어 가는 것과, 실세계의 말뭉치를 이용하여 문법을 평가하고 수정해 나가는 것이 필요하다.

## 참고문헌

[김나리96] 김나리, 김영택, "한국어 동사 패턴에 기반한 한국어 문장 분석과 한영 변환의 모호성 해결," 정보과학회논문지, 23권 7호, pp.766-775, 1996.7.

[김광백02] 김광백, 박의규, 나동열, 윤준태, "구간 분할 기반 한국어 구문분석," 14회 한글 및 한국어정보처리학술대회 논문집, pp.163-168, 2002.10.

[김미영00] 김미영, 강신재, 이종혁, "단위 분석과 의존문법에 기반한 한국어 구문분석," 정보과학회 2000 봄 학술발표논문집, pp.327-329, 2000.

[김종복03] 김종복, 양재형, "조동사 복합술어 구문의 효율적 처리를 위한 구문, 의미분석기 구축," 제15회 한글및한국어정보처리학술대회 논문집, pp.191-197, 2003.

[박소영99] 박소영, 황영숙, 임해창, "X-바 이론의 중심어 개념을 도입한 형태소 단위의 한국어 자질기반 문법," 정보과학회논문지(B), 26권 10호, pp.1247-1259, 1999.10.

[성원경97] 성원경, 장명길, 박재득, 류법모, 이현아, 박동인, "SERI Test Suites '97," 제9회 한글및한국어정보처리학술대회 논문집, pp.320-326, 1997.

[심광섭04] 심광섭, 양재형, "인접 조건 검사에 의한 초고속 한국어 형태소 분석," 정보과학회논문지:소프트웨어 및 응용, 31권 1호, pp.89-99, 2004.

[이현영99] 이현영, 황이규, 배우성, 이용석, "문형을 제약조건으로 하는 CFG 기반의 한국어 구문분석," 정보과학회 '99가을 학술발표논문집, pp.190-192, 1999.

[정천영01] 정천영, 서영훈, "대화체 번역을 위한 논항 구조에 기반한 한국어 분석," 정보과학회논문지: 소프트웨어 및 응용, 28권 4호, pp.380-387, 2001.4.

[Cho95] Cho, Y.-M. Y., Sells, P., "A Lexical Account of Inflectional Suffixes in Korean," *Journal of East Asian Linguistics*, Vol.4, pp.119-174, 1995.

[Collins97] Collins, M., "Three Generative, Lexicalised Models for Statistical Parsing," *Proc of ACL*, pp.16-23, 1997.

[CoNLL00] Cardie, C., Daelemans, W., eds., *Proc of 4th Conference on*

*Computational Language Learning* (CoNLL-2000), 2000.

[Kim98] Kim, J.-B., "Interface between Morphology and Syntax: A Constraint-Based and Lexicalist Approach," *Language and Information*, Vol.2, pp.177-233, 1998.

[Lapointe93] Lapointe, S., "Dual Lexical Categories and the Syntax of Mixed Category Phrases," *Proc of ESCOL*, pp.199-210, 1993.

[Magerman95] Magerman, D., "Statistical Decision-Tree Models for Parsing," *Proc of ACL*, pp.276-283, 1995.

[Malouf98] Malouf, R., "Mixed Categories in the Hierarchical Lexicon," Stanford: CSLI Publications, 1998.

[Manning99] Manning, C.D., Schuetze, H., *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[Oepen02] Oepen, S., Flickinger, D., Tsujii, J., Uszkoreit, H., *Collaborative Language Engineering*, CSLI Publicatios, 2002.

[Pollard94] Pollard, C., Sag, I., *Head-driven Phrase Structure Grammar*, Stanford: CSLI Publications, 1994.

[Sag99] Sag, I., Wasow, T., Syntactic Theory: *A Formal Approach*, Stanford: CSLI Publications, 1999.

[SparckJones00] Sparck Jones, K., Gazdar, G., Needham, R., "Introduction: Combining Formal Theories and Statistical Data in Natural Language Processing," *in Computers, Language and Speech: Formal Theories and Statistical Data, Philosophical Transactions of the Royal Society*, Series A, Vol.358, No.1769, pp1227-1238, 2000.

[Yoon89] Yoon, J., *A Restricted Theory of Morphosyntactic Interaction and Its Consequences*, Ph.D Dissertation, Univ of Illinois, 1989.

김종복 (Jong-Bok Kim)
경희대학교 영어학부(School of English, Kyunghee Univ)
주소 : 서울 동대문구 회기1동 경희대학교
Email : jongbok@khu.ac.kr

양재형 (Jae-Hyung Yang)
강남대학교 컴퓨터미디어공학부(School of Computer Media Engineering, Kangnam Univ)
주소 : 경기도 용인시 기흥읍 강남대학교
Tel : 031) 280-3757
Email : jhyang@kangnam.ac.kr