

# 한국어 문법 검사기의 기능 확장을 위한 서술어와 논항의 통사·의미적 관계 분석

남현숙, 손훈식, 최성필, 박용욱, 소길자, 권혁철  
부산대학교 전자계산학과 인공지능 연구실

Analysis of Predicate/Arguments Syntactico-Semantic Relation for the  
Extension of a Korean Grammar Checker

AI Laboratory, Department of Computer Science, Pusan National University

## 요약

언어의 내적 특성을 반영하는 의미·문체의 검사 및 교정은 언어의 형태적인 면과 관련된  
는 단순한 철자 검사 및 교정에 비해 더 난해하고 복잡한 양상을 띤다. 본 논문이 제안하는  
의미 정보를 이용한 명사 분류 방법은 의미와 문체 오류의 포착과 수정 기능을 향상시키기  
위한 방법의 하나이다. 이 논문은 문맥상 용법이 어긋나는 서술어를 교정하기 위해 명사 의  
미 분류방법을 서술어/논항의 통사·의미적 관계 분석에 이용하여 의미 규칙을 세우는 과정  
을 서술한다. 여기서 논항인 명사의 의미 정보를 체계적으로 분류하기 위해 시소러스 기법  
과 의미망을 응용한다. 서술어와 논항 사이의 통사·의미적 관계에 따라 의미·문체 오류를  
검사하고 교정함으로써 규칙들을 일반화하여 구축하게 하고 이미 존재하고 있는 규칙을 단  
순화함으로써 한국어 문법 검사기의 기능을 보완한다.

## I. 서론

철자 오류와는 달리 의미·문체 오류는 그 판별 기준  
을 가늠하기 어렵고 체계적으로 설명하기도 까다롭다.  
그 이유는 검사 단어에는 기계가 포착할 수 있는 형태상  
의 오류는 없고, 다만 어휘의 용법에 문맥상 하자가 있  
기 때문이다. 따라서 기계에 인간의 문맥 파악 능력에  
비견가는 기능을 부여해야 통사·의미 오류의 포착과 교  
정이 가능하다. 이와 같은 문제를 해결하기 위해서는 의  
미론적 차원에서 더욱 세밀한 언어분석이 이루어져야 한  
다.

현재 사용하는 문법 검사기[1, 2]는 서술어의 통사 자  
질에 따른 서술어/논항의 연어관계에 기반한 규칙들을  
구축하여 활용하고 있다. 그러나 이 구성소 사이의 의미  
적 연관 관계를 밝히는 정보는 하나의 규칙으로 체계화  
되어 있지 못하고, 각 서술어에 지배를 받는 의존소<sup>1)</sup>의  
예를 하나하나 열거함으로써 개별 단어에 대응하는  
방식으로 오류를 검사하고 있다.

본 논문은 명사를 이 논항들의 의미 정보에 따라 하위

범주화하고 각각 하나의 목록으로 만드는 방법을 제안한  
다. 그리하여 논항이 속한 상위 범주와 서술어 사이의  
의미적 논리 관계를 규칙으로 체계화하여 일반화한다.

본 논문은 다음과 같은 단계를 거쳐 전개된다.

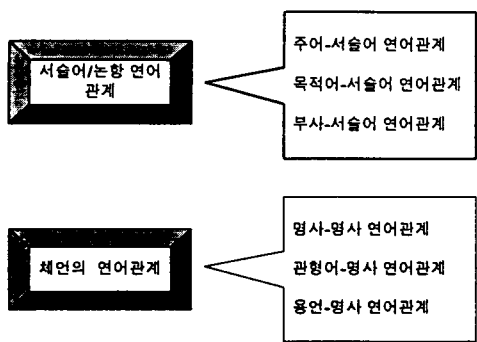
- 1) 이 논문에서 다루게 될 연구의 범위를 제한하고 주요 초점  
이 되는 내용을 소개한다.
- 2) 현재 문법 검사기의 의미·문체 오류에 대한 처리 기능을  
간략하게 설명함으로써 문제점을 밝힌다.
- 3) 현재 시스템이 해결할 수 없는 자료를 처리하기 위한  
방법론을 소개한다.
- 4) 이 분석으로 얻어진 결과를 이용하여 새로운 의미 규칙  
을 세워서 추가한다.
- 5) 이 과정을 통하여 드러나는 새로운 문제점을 앞으로 나  
아갈 연구의 방향으로 고찰한다.

## II. 논문의 초점

현재 한국어 문법 검사기의 구문 분석은 의존 문법에  
의거한다[1, 2]. 먼저 검사 교정할 오류 유형을 조사하고,  
단어 사이의 연관 관계에 기반한 처리 규칙을 만들어 지

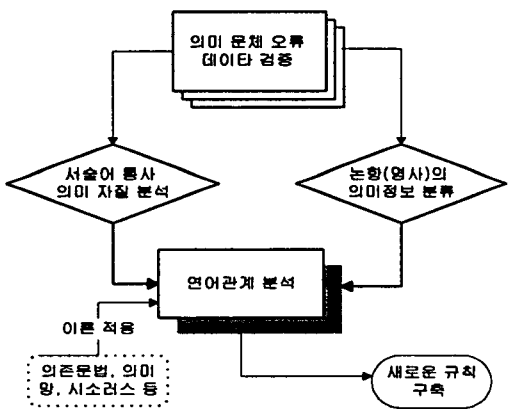
1 본 논문에서는 지배소인 서술어와 의존소인 논항을 주로 다루기 때  
문에 이 후 '의존소' 대신 '논항'이란 용어를 사용할 것이다.

식 베이스화한 후, 이를 기반으로 검사·교정을 한다[1]. 규칙 처리부의 언어관계(collocation) 정보는 일반적으로 다음과 같다.



[그림 1] 단어 사이의 언어 관계 정보

본 논문은 이 언어관계 정보 중에서 서술어/논항 사이의 언어관계에 연구의 초점을 둔다. 아래 그림은 새로운 의미 규칙을 세워 현재 시스템의 규칙 처리부에 추가하는 과정을 보여준다.



[그림 2] 새로운 규칙 추가 과정

우선 의미 오류를 일으키는 예를 열거해 놓은 데이터를 검색한다. 여기서 발견되는 공통된 사항을 탐색하여 규칙으로 정립한다. 이 때 시소러스 기법과 의미망을 이용하여 서술어의 논항 역할을 하는 명사들을 의미 정보에 따라 분류한다. 하위 범주화된 명사의 정보 목록은 의미·문체 오류의 검사 및 교정에 필요한 부분 사전 정보를 체계화하는 데에 이용된다.

의미·문체 오류의 여러 가지 유형 중에서도 가장 빈번히 일어나는 오류는 서로 비슷한 발음을 가진 동사들에 기

인한다. 몇 가지 예를 들어 보자.

- ① 통사자질로 구별할 수 있는 동사
  - 안다: 타동사 + 목적어(대격: -을/-를)
  - 알다: 자동사 + 부사어(장소격: -에)
  - 그외 '낫다/날다', '뾰다/뿔다' 등
- ② 논항의 의미자질로 구별할 수 있는 동사
  - '다리다/달이다', '붙이다/부치다', '엮기다/엮키다', '저리다/절이다' 등

이 예들은 한국어 동사의 특성상 독특하게 드러나는 오류의 한 유형이다. 첫 번째 항은 두 동사를 통사자질로써 구별할 수 있기 때문에 본 논문에서 다루지 않는다. 두 번째 항은 동사의 통사 자질이 동일하므로 본 논문의 서술어/논항의 의미적 언어 관계 분석에 적절한 자료이다.

### III. 논항(명사)의 하위 범주화

#### III-1. 현 시스템의 사전 체계

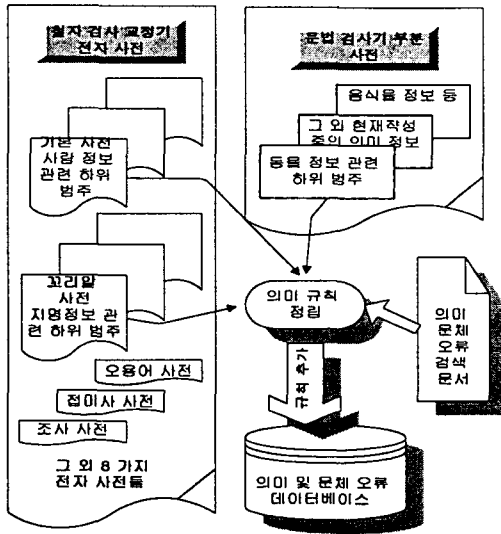
현 시스템의 철자 검사/교정기는 12 가지 종류의 전자 사전을 보유하고 있다. 기본 사전, 조사 사전, 어미 사전, 접미사 사전, 꼬리말 사전, 오용어 사전 등이 그 예이다. 특히, 기본 사전 과 꼬리말 사전 내의 어휘 정보 중에는 의미와 관련있는 몇 가지 정보들이 삽입되어 있다. 하나의 어절을 이루는 구성소 사이에는 때때로 의미적 연관 관계가 성립한다. 즉, 형태소 분석이나 단어의 생성을 설명하는 규칙을 세우는 데에 의미관계가 관여한다. 하위 범주화된 의미 정보는 원래 형태론의 제반 규칙들을 세우기 위한 제약 정보로 이용하였다. 예를 몇 가지 들어 보자.

- ▶ 기본 사전:
  - ① 사람 정보를 나타내는 어절의 6 가지 하위범주: '-가', '-사', '-자' 등과 같은 사람 정보를 나타내는 접미사가 붙은 어절, 성씨, 인명, 사람명사, 직명 등.
  - ② 지명 정보를 나타내는 하위범주: '나라', '시', '군' 등.

- ▶ 꼬리말 사전:
  - 나라 이름 뒤에는 '-사람', '-땅', '-령', '-어'와 '-말' 등이 온다.

최근에 문법 검사기를 개발하면서 이 정보를 문법 검사기에서 의미·문체 오류를 검색하는 데에도 활용하고 있다. 이에 비해 동물 정보②의 하위 범주화는 의미·문

체 오류 검사 및 교정을 목적으로 따로 만들어져서 부분 사전을 구성한다. 이와 같은 사전 정보 체계를 그림으로 나타내면 다음과 같다.



[그림 3] 의미 규칙 정립에 쓰이는 사전 정보

기존 문법 검사기는 각 시스템에서 필요에 따라 하위 범주화한 정보들을 참고로 한다. 의미·문체 오류의 검사 및 교정을 위한 부분 사전의 구성은 특정 오류와 관련있는 명사의 의미 정보에서부터 일반적인 의미 정보의 하위 범주화로 나아간다. 현재 문법 검사기에는 명확한 의미 계층구조를 가진 동물 정보가 먼저 하위 범주화되어 있다. 명사의 하위 범주화는 일반화한 지식 표현의 방법으로 필요하다. 본 논문은 이를 위해 시소러스 작성 기법과 의미망 구조를 이용한다.

### III-2. 시소러스<sup>3)</sup> 기법 응용

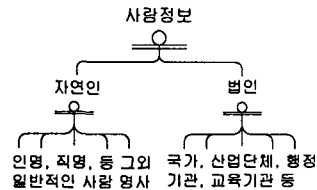
시소러스 기법은 단어의 의미 정보에 따라 단어를 체계적으로 분류하는 데에 응용될 수 있다. 대개 시소러스는 하향식(Top-Down)방식으로 의미를 분류하는 것이 원칙이다. 그러나 현재 문법 검사기에서 구현하는 방식은 일종의 상향식(Bottom-Up)방식으로서 구체적인 용례에서부터 상위 범주를 밝혀 나간다. 그 이유는 정보 검색 분야와는 달리 현재 문법 검사기의 Corpus는 한 특정 분야의 어휘 집합체가 아니므로 추출된 어휘가 속하는

2 동물 정보는 '포유류', '조류', '파충류', '어류' 등 10 가지이다.

3 시소러스란 '통제된 색인 용어의 어휘집, 개념간의 특정관계를 형식적으로 조직화하여 명시한 것'으로 데이터베이스나 책자 형태의 색인 또는 목록류 등의 정보검색 시스템에서 이용되고 있다.

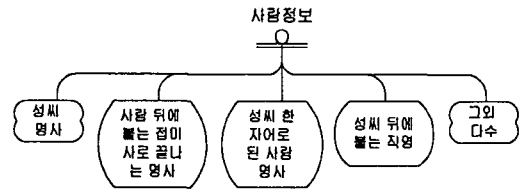
주제 분야를 정확하게 정의할 수 없기 때문이다. 또한 비교 대상인 동시에 따라 서로 변별력을 가지는 하위 범주의 쌍이 시소러스의 계층구조처럼 조직적으로 나타나지 않는 경우가 많다. 그러므로 우리는 시소러스 기법의 '이용' 대신에 '응용'이라는 표현을 쓴다.

아래의 두 그림[4, 5]에서 일반적인 시소러스 작성 기법과 철자 검사/교정기의 기본 사전에 이용된 기법을 비교한다.



[그림 4] 시소러스의 일반적인 기법

시소러스 작성에 있어서 두 가지 원칙은 첫째, 상위어와 하위어의 체계적인 계층화이다. 둘째, 동일 계층 어휘의 등가성이다. 현재 철자 검사/교정기의 기본 사전에 삽입된 사람 정보 분류 방법(그림 5)은 위의 두 가지 원칙을 자연언어 처리에 적절하게 활용한다.

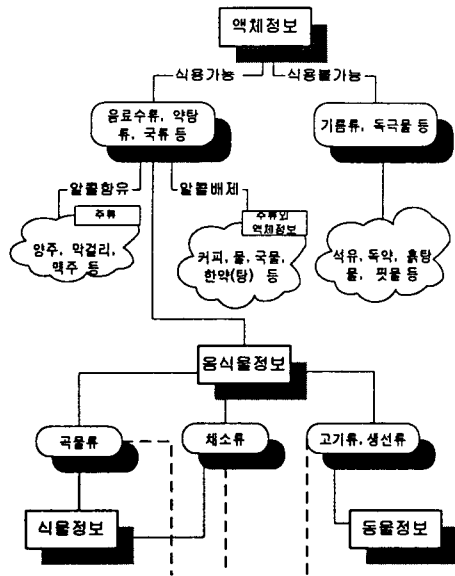


[그림 5] 사람 정보 하위 범주화 적용의 예

위의 그림에 제시된 분류방법은 먼저 형태소 분석기에서 활용되었기 때문에 단어의 형태론적 특성을 고려한 것이다. 이것이 정보 검색 시스템에서 사용하는 시소러스 기법과 구별되는 독창성이다. 다음 페이지에서 의미·문체 오류를 검사/교정하는 데에 쓰이는 하위 범주화의 유형을 설명하겠다.

### III-3. 시소러스 기법을 응용한 명사의 하위 범주화 구현

이 논문은 문법 검사기에 시소러스 작성의 기본 원칙과 언어의 특성을 동시에 반영하여 의미 정보를 하위 범주화하고자 한다. 이를 위해 각 단어가 함축하고 있는 의미자질을 최대한 활용한다.



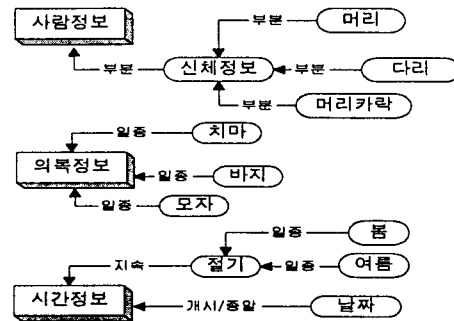
[그림 6] 의미자절을 이용한 시소러스 작성

어휘 정보의 하위범주화를 위해서는 자연언어의 특성을 살려 시소러스 기법을 활용해야 한다. 단어는 문맥상의 쓰임에 따라 여러 가지 상위 범주를 가질 수 있다. 예를 들면, 음료수류, 국물류 등은 액체 정보 뿐만 아니라 음식물 정보에도 속한다. 공간 정보를 가진 명사들 중에서 '국가', '기업', '학교' 등은 생명 정보(+Animate)를 나타내는 명사들처럼 문장에서 동작주 보어(Agent) 기능을 하기도 한다. 그러므로 이 명사들은 장소 정보이면서 동시에 법인 정보를 가지는 의미 범주에 포함된다. 어휘 정보의 시소러스는 상위 범주와 하위 범주의 연결 관계가 순환구조를 이루지 않을 뿐 그 범주들이 서로 그물망처럼 연결되어 있다. 이런 이유에서 '의미망(semantic networks)' 이론을 도입한다. 이 의미망 구조도 시소러스의 일종으로서 특정 명사의 하위 범주화에 유용한 정보를 제공한다.

### III-4. 의미망(semantic networks)의 활용

의미망은 심리학에 그 근원을 둔다. 이 이론도 언어의 의미적 문제를 완전하게 해결하지는 못한다. 그러나 단어 사이의 의미적 관계를 다양하게 정의하는 데에 유용하다. 의미망에서 이용되는 상위 범주와 하위 범주 사이의 관계 유형의 예를 몇 가지 소개해 보자. 의미 계층 구조를 나타내는 데에 쓰이는 '일종의\_\_', 어느 한 범주의 일부를 표현하기 위해 사용되는 '\_\_\_의 부분' 등이 있으며, 공간적 관계의 유형으로는 '방향', '근접과 이원', '시간적

관계의 유형으로는 '지속', '시작과 끝' 등과 같은 관계 정의가 있다[5, 10]. 다음의 그림을 보면 그 구체적인 쓰임을 이해할 수 있다.



[그림 7] 의미망에서 관계 유형의 예

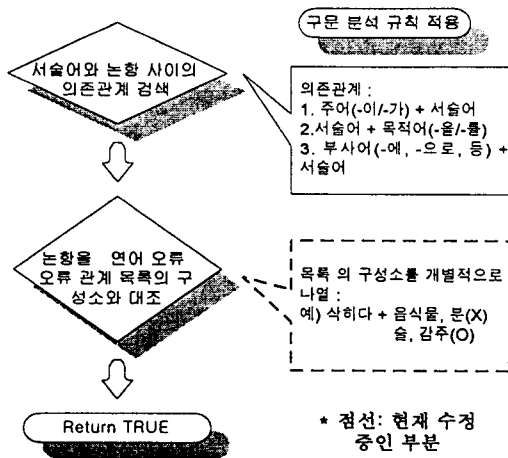
현재 문법 검사기의 Corpus에 있는 어휘를 이와 같은 과정을 거쳐 맹목적으로 하위 범주화할 수는 없다. 문법 검사기 시스템에 맞도록 가장 경제적으로 이 작업을 수행하는 것이 중요하다. 하위 범주를 지나치게 세분화하면 오히려 의미 규칙을 복잡하게 할 수 있다. 현실적으로 시스템에 적용하는 것이 중요하다. 그리하여 본 논문에서 서술어의 의미에 따른 부분적인 명사의 하위 범주화 방법을 제안한다. 다음 장에서는 명사의 하위 범주화가 서술어/논항의 의미 관계 분석에 어떻게 활용되는 지를 살펴 본다.

## IV. 서술어/논항의 통사·의미적 관계 분석

### IV-1. 서술어/논항 의미 규칙 구현 과정

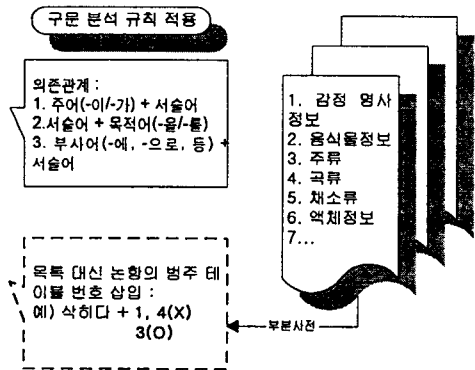
해당 문서 속에 검사할 대상이 되는 서술어가 나타나면 의미·문체 오류 처리는 검사 단어의 연어관계 검사 규칙에 의거하여 실행된다<sup>4)</sup>. 이 때 구문 분석을 통해 자동사와 타동사는 의존소인 논항이 목적어인가 주어인가에 따라 구별될 수 있다. 그러나 검사어와 대치어가 서로 동일한 통사 자질을 가지면 이 단계에서 검색이 불가능하다. 불가피하게 동사의 논항이 가지는 의미 정보를 고려해야 한다. 현 시스템에서 논항인 명사들은 연어 관계/연어 오류 관계의 목록에 개별적으로 등록되어 있다. 바로 이런 이유로 논항의 의미적 하위 범주 테이블을 작성하는 일이 중요하다. 아래의 두 그림을 비교해 보면 논항의 하위 범주화에 의해 달라진 의미 규칙 내용을 알 수 있다.

4 이에 대한 상세한 내용은 이 논문의 참고 문헌 [1]에서 다루고 있다.



[그림 8] 의미 문체 오류의 규칙 적용 과정

아래의 그림은 현재 수정 중인 부분(점선 내)이 어떻게 달라지고 있는지를 구체적으로 나타낸다. 아래의 번호는 임의로 붙여졌다.



[그림 9] 부분 사전 정보를 이용한 의미 규칙 단순화

여기서 각 범주에 번호가 매겨진 논항의 의미 분류 테이블에 주목해야 한다. 현재 시스템에서는 논항의 예가 개체별로 나열되어 있어서 직접 규칙 데이터베이스에 들어가서 규칙을 수정하거나 추가해야 한다. 그러나 명사들이 하위 범주화되어 고유 번호를 가지면 부분 사전에서 정보를 추가 또는 삭제함으로써 이와 관련한 모든 의미 규칙의 적용 환경이 자동으로 변경된다. 이 체계에서는 규칙 처리부와 부분 사전이 독립되므로 의미 규칙의 수정이나 추가 과정이 단순화된다.

동사의 의미에 따라 '액체 정보', '사람 정보'처럼 상위

의 범주로도 구별이 가능한 것이 있다. 또 일부 동사의 경우는 '주류', '기름류'처럼 세분된 하위 범주만이 변별력을 가진다. 다양한 용례들을 참고로 하여 명사의 하위 범주화와 의미 규칙의 안정화 작업을 차근차근 수행해 나가야 한다. 다음 페이지에서 이에 대한 구체적인 예를 설명하겠다.

## VI-2. 변별 자질에 따른 하위 범주화

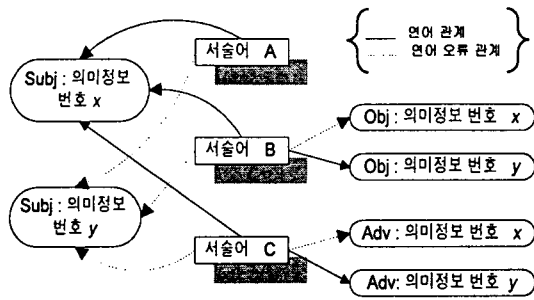
그림 8에 나타난 서술어 '삭히다'는 주류(술) 정보를 가진 명사를 목적어로 취하며, 대체 단어인 '삭이다'는 일반적인 음식물 정보나 감정 명사인 '분'을 목적어로 가진다. 이 경우 음식물 정보의 하위 범주는 세분되어야 한다. 반면 '엷기다'와 '엷기다'를 검색하는 데에는 상위 범주인 '액체 정보'가 변별 자질로서 충분하다.

서술어/논항 의미 관계	변별 자질
(술, 감주, 막걸리, 등)을 삭히다 (음식물, 분)을 삭이다	[±주류 정보] / [±음식물 정보] [±감정 정보]
(실, 줄, 머리카락, 등)이 엷기다 (피, 우유, 기름, 등)이 엷기다	[±액체 정보]

이미 밝힌 바와 같이 단어의 하위 범주화 작업은 서술어와 논항의 구체적인 용례를 자료로 하여 시소러스 기법과 의미망을 적용하였다. 의미 정보 분류의 기준은 서술어의 통사·의미 자질에 의거했다. '삭히다'와 '삭이다'의 예를 보자. 먼저 상위 범주인 '음식물 정보'에서 하위 범주인 '주류 정보'를 세분한다. 반면, 현재로서는 추상 명사 '분'처럼 분류하기 까다로운 정보를 하나의 상위 범주 '감정 정보' 속에 포함시켜도 변별자질로서는 충분하다. '엷기다'와 '엷기다'의 경우를 보자. '엷기다'의 논항의 의미 정보가 일반적인 상위 범주 '액체 정보'에 등록되어 있으면, '엷기다'의 논항이 어떤 범주에 속하는가와 무관하게 두 동사는 구별된다. 서술어의 의미에 따른 분류의 장점은 이러한 작업의 경제성에 있다.

## VI-3. 서술어/논항 통사·의미 관계 유형

서술어/논항의 연어 관계 규칙에 논항의 의미 정보를 변별 자질로서 추가하면 일반적인 세 가지 규칙 패턴을 세울 수 있다.



[그림 9] 서술어/논항 언어 관계의 일반 패턴

위의 그림에서 먼저 동사를 통사 자질에 따라 나누고 논항의 의미 정보를 이용한 서술어/논항의 통사·의미 관계를 유형별로 분류했다. 각 패턴에 해당하는 동사의 예를 들면 다음과 같다.

주어	목적어	부사어	서술어 교정
[+액체 정보 번호]	없음	임의 요소	영키다(X) 영기다(O)
[+사람 정보 번호] [+동물 정보 번호]	[+음료수 정보 번호]	임의 요소	들이키다(X) 들이켜다(O)
[+사람 정보 번호]	[+신체 내부 기관 정보 번호]	임의 요소	의식하다(X) 이식하다(O)
[+사람 정보 번호]	없음	[+공간 정보 번호]	들리다(X) 들르다(O)

언어의 내면 구조에는 의미 그 자체로 이해할 수 없는 심리적 요소가 다양하게 작용하고 있다. 이미 알려진 바와 같이 명사를 의미 정보에 따라 체계적으로 분류하는데 무리가 따른다. 그러므로 이와 같은 작업을 하는 데는 많은 시간과 다양한 지식이 필요하다. 분류하기 쉬운 어휘에서부터 시작하여 더 복잡한 어휘의 하위 범주화 작업으로 나아가야 한다.

### V. 앞으로의 과제

자연 언어 처리 분야에서 이제 언어의 정확한 의미 분석은 한 시스템의 수준을 판가름하는 중요 관건이다. 단번에 만족할 만한 성과를 얻기는 어렵다. 의미·문체 오류의 검사/교정에 필요한 의미 규칙 정립을 위한 단어의 하위 범주화가 의미 분석의 기초 작업이다. 의미론 단계에서 해야 할 과제는 아직도 많이 남아 있다. 추상 명사의 의미 정보는 심리학과 밀접한 관련이 있기 때문에 분

류하기가 가장 복잡한 어휘 정보이다. 발음이 유사하여 오류를 저지르게 하는 동사 뿐만 아니라 동음이의어나 다의어의 처리도 풀어야 할 숙제이다. 문체론 단계로 나아가면 은유(metaphor) 문제를 결코 간과할 수 없다. 화용론 단계의 자연 언어 처리도 미개척 분야로 눈앞에 있다.

현재 문법 검사기의 특성상 흥미를 끄는 또 하나의 연구 과제가 있다. 의미·문체 오류에 대해 주석 기능을 하는 도움말 처리이다. 의미·문체 오류는 그 언어에 대한 지식 부족과 관련이 있기 때문에 특히 사용자에게 오류의 근거를 명확하게 밝혀줄 구체적인 설명이 필요하다. 이는 교정 기능 뿐만 아니라 우리말을 바르게 사용하고 정확한 문서 작성이 가능하도록 도와주는 학습 지침서 기능을 문법 검사기에 부여하는 효과를 준다. 현 문법 검사기는 오류 교정과 학습 지도를 동시에 수행할 수 있는 시스템으로 발전할 수 있다.

### 참고 문헌

- [1] 김영진, 최성필, 손훈석, 박용욱, 권혁철, 박동인, “단어의 하위 범주화 정보를 이용한 한국어 문법 검사기”, 한국정보과학회, 인공지능연구회, 1997년도 춘계 학술발표 논문집, pp.72-75, 1997
- [2] 김현진, 어절 간 의존 관계와 부분 문장 분석을 이용한 한국어 문법 검사기 구현, 석사학위 논문, 부산대학교 대학원 전자계산학과, 1997.
- [3] 김영택, 자연언어처리, 교학사, 1994
- [4] 서정수, 국어문법, 한양대학교 출판원, 1994
- [5] 송도규, 인지 언어학과 자연 언어 자동 처리, 홍릉과학출판사, 1996
- [6] J. AITICHISON & A. GILCHRIST (전학병 옮김) 시소러스의 작성법, 산업기술정보원, 1991
- [7] J. ALLEN, Natural Language Understanding, Second Edition, The Benjamin/Cummings Publishing Company, INC., 1994
- [8] Andrew.R. GOLDING, Combining Trigram-Based and Feature-Based Methods for Context-Sensitive Spelling Correction, Annual Meeting of the ACL, 1996
- [9] FUCHS(C.), Linguistique et Traitements Automatiques des Langues, Hachette, 1993
- [10] NOGIER(F.) Génération Automatique de Langage et Graphes Conceptuels, HERMES, 1991
- [11] FRANCOIS(R.), Pour l'Analyse Sémantique, Masson, 1994