동사의 애매성 해소를 위한 구문의미사전의 구축

송영빈, 채영숙, 박용일, 이정민, 설가영, 황혜리, 한나리, 최기선
한국과학기술원 전문용어언어공학연구센터
{ybsong,yschae,yipark,jmlee,kyseol,hrhwang,nrhan,kschoi}@world.kaist.ac.kr

Dictionary Making for Disambiguation

Young-Bin Song Young-Soog Chae Yong-il Park Jun-Min Lee
Kah-Young Seol Hye-ri Hwang Na-Ri Han Key-Sun Choi

요 약

동사의 애매성이란 동일 동사 내부에서 공기하는 명사의 상충적 의미의 분포에 의해 발생한다. 이는 동일한 동사라 하더라도 명사의 상위개념, 혹은 개개의 명사에 따라 동사의 의미가 달라진다는 것을 의미한다. 동사의 애매성 해소를 위한 구문의미사전은 동사가 갖는 격틀과 논항에 오는 명사의 단어 집합에 의해 구성된다. 기계용 사전에서의 동사의 애매성이란 명사의 상위개념, 혹은 개개의 명사에 관한 정보가 결여될 때 나타난다. 지금까지의 구문의미사전은 개개의 동사가 갖는 격틀을 중심으로 논항명사의 예만을 제시하거나 명사의 상위개념을 기술하는 형식으로 구성되어 왔다. 이는 형식적인 패턴의 추출에는 유용하지만 대역어 선정을 위한 구문의미사전과 같은 섬세한 의미 정보를 필요로 하는 사전에서는 거의 효력을 발휘하지를 못한다. 다국어를 전제로 한 동사 대역어의 추출을 목적으로 하는 구문의미사전에서는 동사와 공기하는 논항명사의 철저한 추출과 검증에 의한 명사목록의 구축이 애매성 해소와 정확한 동사 대역어의 선정에 전제가 된다. 본 논문에서는 KAIST Corpus를 기반으로 현재 구축 중인 한국어 구문의미사전의 개요와 구축 과정에서 얻어진 방법론을 소개한다. 이 연구개발 결과는 과학기술부 KISTEP 특정연구개발과제 핵심소프트웨어개발 국어정보처리 기술개발 중 "대용량 국어정보 심층 처리 및 품질 관리 기술 개발"의 지원을 받았다.

1. 머리글

동사의 애매성이란 동일 동사 내부에서 공기하는 명사의 상충적 분포의 차이에 의해 발생한다. 동일한 격틀을 갖는 경우에도 다음 예와 같이 N0, N1의 명사의 분포에 의해 [치다]의 의미는 달라진다.

치다1 N0 N1
N0=[인간]
N1=[소금, 후추, 식초…]
치다2 N0 N1
N0=[트럭, 승용차, 마차…]
N1=[사람, 동물…]

이와 같은 애매성은 모국어 화자에게 있어서는 경험과 학습에 의해 지식으로 축적되는 것으로 N0에 경험하거나 학습하지 않은 '기차'와 같은 새로운 명사가 등장하더라도 인간은 추론에 의해 그것이 [치다1]과 [치다2] 중에 어디에 들어갈 것인지를 쉽게 알 수 있다. 구문의미사전이 자연언어처리 분야나 외국어 교육의 분야에서 활발히 연구되는 것은 경험과 추론을 적용하기 힘든 특성 때문이다.

동사의 애매성 해소를 위한 표현 기법은 동사를 문법적 성질에 따라 자동사, 타동사, 자타동사와 같은 격틀을 설정하는 데서 출발한다. 여기에 논항명사의 목록을 추가함으로서 완성된다. 이와 같은 형식을 일반적으로 문형이라고 한다. 일반적인 문형은 논항명사에 구체적 명사의 예를 적어 주거나 상위개념을 기술하는 데 그치는 반면 현재 구축하고 있는 구문의미사전에서는 상위개념은 물론 동사와 공기하는 명사의 모든 목록을 구축하고 동사의 의미구분을 기존의 사전의 분류에 의존하지 않고 유형화한다는 점에서 기존의 문형사전과는 차이가 있다. 동사와 공기하는 명사의 모든 목록을 망라하는 이유

는 동일한 명사의 목록을 공유하는 명사 내부에서도 외국어로 동사의 대역어를 달아주면 서로 다른 동사와 대응하는 경우가 많기 때문이다. 이는 동사와 명사의 공기관계에 의한 의미 관계 파악이 각각의 언어마다 다르다는 것을 뜻한다.

본 논문은 다음과 같이 구성된다. 2장에서는 사전이 갖추어야할 이상적인 정보를 소개하고 본 연구에서 수록 대상으로 한 정보에 대해 작업 개요를 겸해서 제시한다. 3장에서는 표기정보에 대해 논하고 4장에서는 문법정보의 구성에 대해 언어학적인 논의를 검증한다. 5장에서는 의미정보를 명사와 동사로 나누어 논한다. 또한 구문의미사전의 효율성을 높이기 위한 중요성 정보 부과의 실제를 제시한다. 마지막으로 6장에서 결론을 맺는다.

2. 사전 기술에 있어서의 필요 정보

사전에는 크게 인간을 위한 사전과 기계를 위한 사전의 두 가지 종류가 있다. 인간용 사전은 인간의 다양한 경험과 지식, 추론을 전제로 만들어지기 때문에 비교적 간단한 정보만을 수록해도 문제가 없다. 그러나 기계용 사전의 경우 지식과 추론의 도움을 받을 수 없기 때문에 인간이 의미를 이해하고 추론하기 위한 갖가지 정보들을 정확히 기술할 필요가 있다.

인간이 의미를 이해하는 기준은 매우 다양한 요소들의 결합에 의해서 의미를 파악하고 이해한다. 이들 요소를 정보라고 했을 경우 기계를 위한 사전에서 필요로 하는 정보의 범위는 다음과 같다[1].

No	정보분류	수록정보의 내용
1	표기정보	(1)표제어 정보 (2)표준표기
2	발음정보	(1)음성정보 (2)음운정보 (3)액센트정보
3	형태정보	(1)어구성 정보 (어기, 약어, 병열어, 전성어, 파생어, 복합어, 연어, 관용어 등) (2)표기문자의 종류 (한글, 한자, 알파벳 등)
4	문법정보	(1)품사 정보 (2)활용 정보(활용형, 활용패턴 등) (3)동사유형 정보(자타동사, 본동사, 보조동사, 기능동사 등) (4)전후 접속 정보 (4)수동, 사동 정보 (5)Aspect 정보 등
5	의미정보	(1)단어 의미 속성 (2)어의 기술 정보 (3)어의 변별 정보(다의어의 의미 판정법) (4)용언의 격 정보 (5)용언 정보로의 pointer
6	위상정보	(1)위상 정보(일반어, 고유명사, 전문용어, 고유어, 한자어, 외래어, 혼종어, 현대어, 고어, 유행어, 사어 등) (2)문체 정보(문장체, 회화체, 속어체 등) (3)존비 정보(존경<주어/대상>, 겸양어, 멸시 등) (4)사용자 정보(남성어, 여성어, 노인어, 유아어 등) (5)감정 정보(멸시, 완곡, 선동, 욕 등) (6)어감 정보(+/0/-) (7) 분야 정보(전문분야, 사용장면)
7	중요성 판별 정보	(1)중요성 판별 정보(기초어, 기본어, 중요도 레벨 표시, 타사전 등재 여부 등) (2)통계적 정보(코퍼스 출현 빈도, 다른 단어와의 연접 확률, 공기 정보)
8	동형이의어 선택정보	(1)동형이의어 판별 정보(동형이의어 식별을 위한 판정법) (2)우선 순위 정보(동형이의어 사이에서의 선택 우선 순위)
9	관련어정보	(1)동의어와 그 위상 (2)상위어와 하위어와의 관계 (3)유의어와 이의 변별 특성 (4)반대어, 연상관계어 (5)전성어 (6)파생어 (7)복합어 (8)연어 (9)관용어 (10)약어
10	공기정보	(1)공기어의 위치(전방, 직전, 직후, 후방) (2)공기어 정보(품사, 의미속성 등) (3)공기관계(고유명사+접사, 명사+명사, 부사+용언)
11	사전검색 제어정보	(1)최장단어 정보(표제어를 포함한 보다 긴 단어가 없음을 표시) (2)동형이의어 최종어 정보(동형이의어의 최종어임을 표시) (3)표제어 내부의 단어 연쇄 정보(예 : [器用]<형용동사>, -器<명사>+用<접미사>)

위의 정보들은 각각 분리된 것이 아니라 서로 연관되어 있는 경우가 많다. [1.표기정보]의 [표제어 정보]를 "고빈도", "고유어"라고 정할 경우 이는 위의 정보분류에서 [7. 중요성 판별 정보]와 [6. 위상정보]를 포괄하게 된다. 실제적인 사전의 구축에 있어서는 기술 정보의 명확한 구분이 있는 것이 아니라 상호 연관된 것이라고 보는 것이 정확하다.

본 연구에서 기술 대상으로 삼은 정보는 다음과 같다.

1	표기정보	(1)표제어 정보
4	문법정보	(1)동사유형 정보(자타동사 정보)
5	의미정보	(1)어의 정보 (2)용언의 격 정보
6	위상정보	(1)위상 정보(일반어, 고유명사, 전문용어, 고유어, 한자어, 외래어, 혼종어, 현대어, 고어, 유행어, 사어 등)
7	중요성판별 정보	(1)중요성 판별 정보(기초어, 기본어, 타사전 등재 여부) (2)통계적 정보(코퍼스 출현 빈도, 다른 단어와의 연접 확률, 공기 정보)
10	공기정보	(1)공기어의 위치(전방, 직전, 직후, 후방) (2)공기어 정보(품사, 의미속성 등)

3. 표기정보

표기정보는 표제어와 관련된 정보를 포괄적으로 지칭하는 것으로 이에는 중요성 정보와 표준표기(규범적인 단어와 실제 사용 단어)가 해당된다. 중요성 정보는 기본어, 기초어, 타사전의 등재 여부 등 여러 가지 기준이 있을 수 있다. 현재 구축 중인 사전에서는 KAIST 코퍼스 빈도를 유일한 기준으로 삼았다. 동사의 애매성은 빈도와 표기의 형태(고유어, 한자어 외래어 등)와 비례한다. 고유어이자 고빈도어의 경우 애매성이 높다는 것은 일반적인 상식에 의거한 것이다.

표준표기에 관한 문제는 한국어에 존재하는 규범적인 단어와 언중이 인정하는 관용적 단어와의 차이를 뜻한다. 이는 경우에 따라서는 격틀의 유형에도 영향을 주는 문제로 언중들이 인식하고 사용하는 관용적 단어에 중점을 두되 규범적인 단어도 기술 대상으로 하고 있다. 대표적인 예가 '좇다'와 '쫓다', '떨다'와 '털다' 등

과, '가지다'와 '갖다' 같은 준말에서 오는 것이다[2].

'담배를 피다'에서, 이 표현을 그릇된 표현이라 생각지 않는 것이 일반적이고 일반 사전에서도 표제어로 등재되어 있다. 그러나 규범적인 입장에서 이는 문법에 맞지 않는 표현이 된다[3]. 접미사 '-우-'가 들어가서, 자동사인 '피다'에서 타동사인 '피우다'가 된 것이므로 자동사 '피다'가 목적어를 취하는 것은 옳지 않다는 주장이다. 이와 비슷한 예로 '하늘을 날다', '다리를 건너다'와 같은 예가 있는데 이와 같은 문제에 대한 언어학적인 논의는 다음으로 미루기로 하고 기계용 사전 구축이라는 관점에서만 논하자면 '담배를 피다'와 같은 표현을 인정하지 않을 경우 '피다'를 처리할 수 없게 되기 때문에 위와 같은 특수한 규범적인 논의는 무시하기로 한다.

4. 문법정보

문법정보 중에서 동사를 그 기술 대상으로 할 경우 가장 기본이 되는 것은 동사의 유형 정보이다. 동사의 유형 정보란 자동사냐 타동사냐 혹은 자타양용동사냐 하는 동사의 고유한 문법정보를 말한다. 이와 같은 정보는 각각의 동사가 갖는 고유한 문법적 특성으로 일반적으로 동사의 애매성 해소에 결정적인 역할을 하는 것은 아니다. 다만 다음과 같이 자동사와 타동사의 용법을 동시에 갖는 경우 이는 애매성 해소에 얼마간 영향을 준다.

치다3 N0
N0=[비바람, 태풍, 눈보라…]

치다4 N0 N1
N0=[트럭, 승용차, 마차…]
N1=[사람, 동물…]

여기서 [치다4]는 '치이다'와 같은 관련어가 존재하는 반면 [치다3]은 그와 같은 관련어가 존재하지 않는다는 점, 문법적으로도 [치다3]은 자동사인데 비해 [치다4]는 타동사라는 점에 의해 같은 동사로 보기 힘든 면이 있다. 이러한 문제에 대해서도 기계를 위한 의미사전의 관점에서는 문제삼지 않는다. 즉 어형이 같으면 동일한 표제어로 인정하는 것이다. 이는 기계처리를 위한 명확성을 확보하기 위한 것이다. 따라서 동일 표제어에는 본동사는 물론 다음과 같은 기능동사도 포함되게 된다.

[난리, 요동, 장난…]을 치다

구문의미사전은 앞서의 [치다3], [치다4]와 같이 각각의 논항에 들어가는 명사를 목록으로 제시함으로써 의미기술의 구체성을 확보하고 있다. 이는 대역구문사전을 구축하는 기본 자료로 활용할 수 있게 구성하기 위해서 명사 목록의 제시가 필수적이기 때문이다.

격틀 유형의 설정에서 문제가 되는 것은 논항의 수를 어디까지 인정을 하느냐는 것이다[4]. 이는 필수논항과 수의논항의 구별의 문제가 되는데 한국어에서, '비교하다'와 같은 동사는 원칙적으로 3 개의 필수 논항을 요구한다.

N0-가 N1-를 N2-와 비교하다

이 때, 위의 3 개 논항은, 술어 '비교하다'의 필수 성분들로서, 조사들은 술어 성분에 의해 내재적으로 이미 결정되어 있다. 따라서, 논항 자리에 나타날 수 있는 명사 성분과는 관계없이, 다음과 같은 형태의 격 형태들은 허용하지 않는다.

*N0-가 N1-에게 N2-를 비교하다
*N0-가 N1-를 N2-가 비교하다

또한, 앞의 3 개의 논항 형태가 '술어'에 의하여 필수적으로 요구된 형태들이기 때문에 다음과 같이 논항이 삭제된 문장은 비문이 된다.

*대의원단은 그 문제를 비교했다
*대의원단은 이 문제와 비교했다

한편, 위와 같은 구문적 특성은, 위와 동일한 형식 구조 속에 실현될 수 있는 '결정짓다'의 경우와 대조를 보인다.

대의원단은 그 문제를 지역 주민들과 결정지었다

위의 '결정짓다'는 '비교하다'의 경우처럼 'N0-가 N1-를 N2-와 Verb'의 통사 구조 속에 실현되었다. 그러나, 이 경우, 'N2-와'는 술어에 의해 요구된 필수논항이 아니기 때문에, 다음과 같이 생략이 가능하다[5].

대의원단은 그 문제를 결정지었다

이와 같은 필수논항과 수의논항의 범주 설정에 대해서 주관적이라는 비판이 있다. 일반적으로 수의논항은 통사 구조에 넣지 않고 필수논항으로만 문형을 구축하는

경향이 있다. 이는 수의논항이 동사의 의미와 직접적인 관련이 없다는 전제가 있기 때문이다. 그러나 수의논항이라고 하더라도 다음에 예시한 [치다5]의 N2처럼 [치다6]의 의미와의 구별을 위해 수의논항의 제시가 필요한 경우가 있다.

치다5 N0 N1 N2
N0=[사람]
N1=[소금, 후추, 식초…]
N2=[국, 고기, 생선…]

치다6 N0 N2 N1
N0=[사람]
N1=[초]
N2=[일]

[치다6]은 N1과 N2에 오는 명사의 목록이 제한적이라는 점과 어순이 [치다5]와 같은 일반문형과는 달리 자유롭지 못하다는 점에서 관용표현으로 분류되는 것이다. 일반적으로 관용표현은 소수의 특수한 언어적 현상으로 구문사전의 기술 대상에서 제외된다. 그러나 현재까지 작성한 3,005개 문형 중 관용표현문형이 255개로 11.78%를 차지하고 있다는 점, 또한 의미처리를 위한 기초 자료로서의 성격을 갖고 있다는 점에서 본 연구에서는 기술 대상에 포함시켰다.

5. 의미정보

구문의미사전에서 가장 중심적으로 다루어지는 의미정보는 명사의 의미 기술이다. 동사의 경우 코퍼스 분석을 통한 출현 빈도가 의미기술에 핵심이 된다.

5.1 명사

명사의 의미를 기술할 경우 의미의 명시성과 간결성을 확보하기 위해 의미속성을 사용하는 것이 일반적이다. 의미속성의 작성은 상하위 개념과 연상, 외연적 의미와 내연적 의미 등에 의해 여러 가지 분류가 가능하다. 많이 쓰이는 방법은 개념 중심의 분류와 연상관계를 이용한 분류이다.

개념 중심의 시소러스에서 가장 많이 쓰이는 방법은 'isa 관계'를 이용하는 방법이다. 'This is a pen', 'He is a teacher', 'She is a girl'과 같이 문장에서 'is a'를 사이에 두고 전항과 후항에 오는 명사를 각각 상위와 하위 관계에 있다고 보고 명사를 계층적으로 분류하는 것이다. 'pen'은 '사물(this)'의 하위개념이라고 할 수 있으며 'teacher', 'girl'은 '인간(he, she)'의 하위개념이라고 할 수 있다. 이는 비교적 간단 명확한 방법으로 지식을 계층화, 구조화하는데 유리하기 때문에 많이 사용되고 있다. 그러나 개념을 분류의 기준으로 삼기 때문에 최하위 노드에 나타나는 명사들의 의미적 공통성이 연상관계에 의한 분류보다 미약해서 의미적으로 상이한 단어가 동일 노드에 나타날 수 있다는 문제가 있다. 또한 추상명사와 같이 지칭하는 의미의 경계가 명확하지 않는 경우라든지 각 언어마다 동일 개념에 포함시키기 어려운 경우가 개별 단어 수준에서는 많다는 점, 동일한 형태의 명사의 의미가 동사에 의해 달라진다는 특성 때문에 하위개념으로 내려가면 갈수록 상위개념과의 정합성을 유지하기가 어려워진다는 점에서 만족할 만한 시소러스는 나오지 못하고 있다.

한편 연상을 이용한 시소러스는 인간의 연상을 근거로 작성이 되기 때문에 동일한 최하위 노드에 속하는 단어 사이의 유사도를 쉽게 확보할 수 있는 반면 상위개념과 하위개념의 관계가 계층구조를 이루기가 어려운 면이 있어서 효율이 떨어진다는 문제점이 있다.

이렇듯 어떠한 방법을 쓰더라도 시소러스 자체의 구축에는 여러 어려움이 수반되기 때문에 시소러스의 응용 분야라고 할 수 있는 동사의 애매성 해소에 시소러스를 쓰는 방법은 많은 한계가 있다. 명사의 의미는 동사와의 공기관계에 의해 구체적인 의미가 기술될 수 있으며 이럴 경우 한 개의 명사에 대해 복수의 의미속성이 주어져야 하는데 개개의 의미를 시소러스의 체계에 맞게 부여하는 것은 매우 힘들고 동시에 부여한다 하더라도 주관적일 수 있다. 하나의 예로써, 명사 '물'은, 다음과 같이 4 가지의 의미를 갖는 것으로 분석될 수 있다.

어휘	의미	예문
물1	산소 1과 수소 2의 화합물	그가 물을 마신다
물2	빛깔	옷에 파란 물이 들었다
물3	영향력	그 아이는 이미 나쁜 물이 들었다
물4	채소 등이 나오는 차례	포도가 이제 물이다

그러나, 여기서, [물2]와 [물3]의 '물'이 과연 두 가지의 의미로 나뉘어져 설명되어야 하는지, 아니면 하나의 원 뜻에 비유적인 정도가 가미되어 [물2], [물3]과 같은 의미로 사용된 것인지는, 판단하는 개인 주체에 따라 달라질 수 있다 (기존 사전에는 나뉘어져 있지 않다)[6].

그러나 외국어와의 대역어 부여라는 관점에서 위의 예문에 대역어를 부여하면 [물1]에서 [물4]까지의 의미적 차이는 비교적 명확하게 드러난다. 다음과 같이 [산소 1과 수소 2의 화합물]이란 의미에서의 '물'만이 일본어 대역어인 'みず(水)'와 대응하고 있을 뿐 나머지 [물

2]부터 [물4]까지는 각각 서로 다른 대역어가 대응하고 있음을 볼 수 있다.

어휘	의미	예문	일본어 대역문
물1	산소1과 수소2의 화합물	그가 물을 마신다	彼が水を飲む
물2	빛깔	옷에 파란 물이 들었다	服が青く染まった
물3	영향력	그 아이는 나쁜 물이 들었다	その子は悪に染まった
물4	채소 등이 나는 차례	포도가 이제 끝물이다	葡萄の旬が終わりに近づいている

한국어만 놓고 보면 의미의 전성과 파생의 문제가 결부되어 의미의 구별이 애매한 경우라도 언어 표현 형식이 다른 외국어와 비교를 통해 그것이 관습적으로 동일한 언어형식으로 한국어에서 굳어진 것일 뿐 의미적으로는 상당히 거리가 있는 단어들임을 알 수 있다. 이는 물리적 의미의 표현 형식이 각 언어마다 다르다는 것을 뜻함과 동시에 의미를 매개로 형식을 선택하는 외국어와의 대역어 관계를 통해 한국어의 의미를 객관적으로 파악할 수 있는 유용한 방법이 될 수 있다는 것을 보여준다. 특히 본 연구에서 대상을 일본어로 정한 것은 한국어와 구문적 특성이 비슷하고 고유어, 한자어, 외래어 등 다양한 어휘의 섬세한 의미적 구분이 발달되어 있는 언어이기 때문이다.

5.2 동사

동사의 의미는 공기하는 명사의 의미속성과 명사에 따라 달라진다. 특히 고유어이자 고빈도어의 경우 복수의 의미를 나타내는 것이 일반적이다. N0, N1을 논항으로 갖는 구문에서 다음과 같이 명사의 상위개념이 다르면 동사의 의미가 달라진다.

치다7 N0=인간 N1=악기
치다8 N0=인간 N1=막

[치다8]의 N1의 경우 한국어에서는 동일한 의미속성을 갖는 명사라고 하더라도 동사의 대역어 선택이라는 관점에서 보면 대역어가 달라지는 경우가 있다.

치다8	논항	명사	대역동사	대역동사의 한국어 의미
	N1 =	커튼	しめる	닫다
		병풍	めぐらす	두르다
		발	張る	치다
		금줄	張る	치다

이와 같은 현상은 매우 일반적으로 일어나는 것으로 대역 구문의미사전을 구축할 경우 명사의 의미속성만으로는 대역어 선정에 한계가 있으며 개개의 동사와 공기하는 모든 명사의 목록을 작성할 필요가 있다는 것을 의미한다. 각각의 언어마다 명사와 동사의 의미 관계가 다르게 파악되고 그 결과 대응하는 언어 형식도 달라진다. 따라서 공기하는 목록을 모두 망라함으로써 구문의미사전의 기본 요건이 갖추어지게 된다.

구문의미사전에서는 동사의 의미구별을 앞서 제시한 여러 예처럼 문형을 통해 구분한다. 의미의 구별은 공기하는 명사와의 관계를 일일이 분석하고 구별을 한다 하더라도 작성자의 주관이 들어가게 된다. 특히 고유어 고빈도어를 대상으로 할 경우 거의 모든 동사가 다의성을 갖고 각각의 동사에 따라 의미 구별의 분포가 상이해서 명사의 상위개념 못지 않게 동사의 의미 구분을 문형으로 표현한다는 것은 쉬운 일이 아니다. 다음은 현재까지 구축된 구문의미사전을 KAIST 코퍼스 출현 빈도와의 대비를 통해 각각의 동사의 의미구분의 특성을 조사한 것이다[7]. 여기서 '문형 유형'은 의미구분을 나타내고 '의미'는 문형을 구축할 시에 인간이 알기 쉽게 변별적인 특징만을 기술하는 정도에서 의미를 기술하고 있다. '출현 빈도'는 KAIST 코퍼스에서의 해당 동사의 의미 유형별 출현 빈도를 나타낸다.

문형 유형	의미	출현 빈도	출현 백분율
보이다1	시야에 들어오다	3478	41.76
보이다2	현상이 있다	175	2.10
보이다3	상태를 따다	34	0.41
보이다4	남에게 보게 하다	3750	45.02
보이다5	보여주다	308	3.70
보이다6	눈치를 받다	0	0.00
보이다7	태도가 나타나다	584	7.01
총계		8329	100.00

표 1. [보이다]의 문형별 출현 빈도

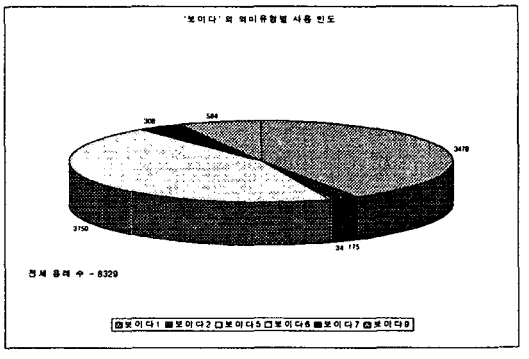

그림 1. [보이다]의 의미 별 사용 빈도

[보이다]의 경우 출현빈도 상으로 볼 때 [보이다1]과 [보이다4]에 의미가 집중되어 있다. 문법적으로 [보이다]는 '사동'과 '가능'을 나타내는 것으로 문법적인 의미를 나타내는 요소 [-이-]등이 들어 간 파생 동사의 경우 비교적 의미의 분포가 치중되어 나타난다는 것을 볼 수 있다. [보이다6]은 출현빈도가 0으로 되어 있는데 이는 코퍼스에 의해 작성한 것이 아니라 일반 사전, 혹은 인간의 머리에서 나온 문형이라는 것을 뜻한다.

한편, [쓰다]의 경우는 [보이다]와는 달리 '사동'이나 '가능'과 같은 문법적 의미가 들어가지 않은 동사로 비교적 의미 별 문형유형에 따른 의미의 분포가 고루 분포되어 있다. 이와 같이 선어말어미의 유무에 따라 분포가 달라지며 이는 파생형태도 표제어로 기술할 필요성이 있음을 알 수 있다. 일반적인 문형의 경우 파생형태를 표제어에서 제외하는 경우가 있으나 이와 같은 의미의 분포의 차이를 구문의미사전에서는 파생형태도 표제어로 기술하고 있다.

문형 유형	의미	출현 빈도	출현 백분율
쓰다1	글을 짓다	828	16.30
쓰다2	들다	0	0.00
쓰다3	덮다	1077	21.20
쓰다4	착용하다	293	5.77
쓰다5	억울하게 지명당하다	0	0.00
쓰다6	표를 만들다	0	0.00
쓰다7	사용하다	37	0.73
쓰다8	몸의 일부를 움직이다	529	10.42
쓰다9	능란하게 다루다	1270	25.00
쓰다10	부리다	0	0.00
쓰다11	소모하다	1	0.02
쓰다12	행사하다	446	8.78
쓰다13	마음을 쓴다	598	11.77
쓰다14	꾀를 쓰다	0	0.00
쓰다15	억지를 부리다	0	0.00
쓰다16	찌푸리다	0	0.00
TOTAL		5079	100.00

표 2. [쓰다]의 문형 별 출현 빈도

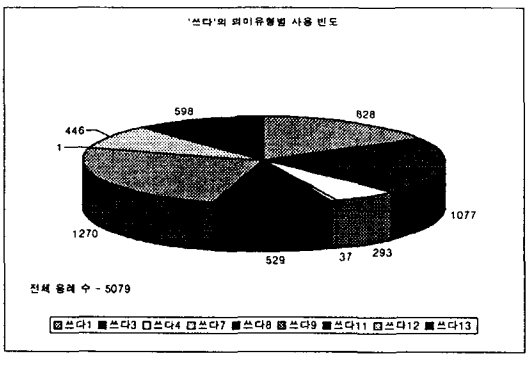

그림 2. [쓰다]의 의미 별 사용 빈도

동사의 이와 같은 출현 빈도는 의미의 중요성 판별 정보로 활용된다. 사전의 기술에 있어서 의미 별 문형의 순서를 출현 빈도에 의해 기술함으로써 보다 효율적인 구문의미사전을 만들 수 있다.

다음은 매우 다의성이 높은 [치다]와 [올리다]의 분석 결과이다. '치다'는 한국어에서 가장 다의성이 높은 동사로 일반문형이 32개, 관용표현문형이 12개로 총 44개의 문형이 작성되었다.

문형 유형	출현 빈도	출현 백분율	문형 유형	출현 빈도	출현 백분율
치다1	99	1.00	치다17	0	0.00
치다2	0	0.00	치다18	0	0.00
치다3	0	0.00	치다19	1	0.01
치다4	0	0.00	치다20	61	0.62
치다5	261	2.63	치다21	7	0.07
치다6	7	0.07	치다22	1223	12.34
치다7	0	0.00	치다23	1463	14.76
치다8	0	0.00	치다24	635	6.41
치다9	0	0.00	치다25	899	9.07
치다10	0	0.00	치다26	2648	26.72
치다11	349	3.52	치다27	10	0.10
치다12	0	0.00	치다28	1259	12.71
치다13	504	5.09	치다29	152	1.53
치다14	0	0.00	치다30	326	3.29
치다15	0	0.00	치다31	0	0.00
치다16	5	0.05	치다32	0	0.00
TOTAL				9909	100.00

표 3. [치다]의 문형 별 출현 빈도

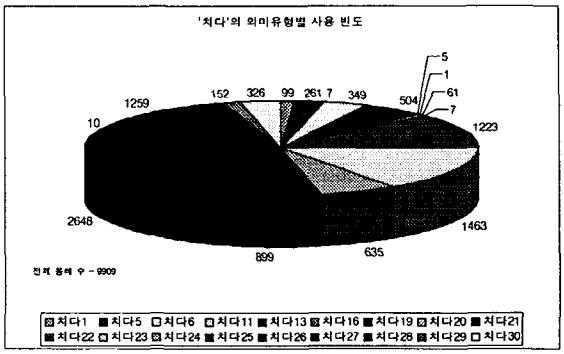

그림 3. [치다]의 의미 별 사용 빈도

'치다'는 표 3.에서 보듯 KAIST 코퍼스에 출현하지 않은 문형이 제일 많은 동사로 나타났다. 출현하지 않은 문형의 예는 다음과 같다.

치다2　　종이 치다
치다3　　트럭이 사람을 치다
치다4　　돼지가 새끼를 치다
치다8　　철수가 시험을 치다

'치다'의 경우 실제로 코퍼스에 나타나는 출현형태와 표제어의 대표형태의 차이가 두드러지는 동사이다. '치이다', '치르다' 등과 같은 실제표현에서 많이 쓰이는 표현형태를 함께 표제어에 기술함으로써 비출현의 문제를 해결할 수 있다. 또한 [치다2]에서처럼 '종이 울리다'와 같은 유사 표현이 존재할 경우 함께 관련어 항목으로 표제어에 기술함으로써 비출현의 문제를 해결할 수 있다.

다음은 다의성이 높으면서도 문형 별 출현 빈도가 고루 분포하고 있는 예이다.

문형 유형	출현 빈도	출현 백분율	문형 유형	출현 빈도	출현 백분율
울리다1	1959	17.80	울리다15	355	3.23
울리다2	624	5.67	울리다16	219	1.99
울리다3	537	4.88	울리다17	640	5.82
울리다4	55	0.50	울리다18	0	0.00
울리다5	596	5.42	울리다19	325	2.95
울리다6	584	5.31	울리다20	35	0.32
울리다7	546	4.96	울리다21	108	0.98
울리다8	717	6.51	울리다22	240	2.18
울리다9	312	2.83	울리다23	66	0.60
울리다10	711	6.46	울리다24	2	0.02
울리다11	28	0.25	울리다25	392	3.56
울리다12	304	2.76	울리다26	542	4.92
울리다13	0	0.00	울리다27	280	2.54
울리다14	25	0.23	울리다28	804	7.31
TOTAL				11006	100.00%

표 4. [울리다]의 문형 별 출현 빈도

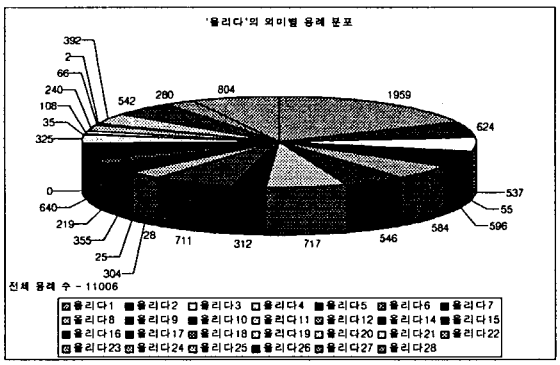

그림 4. [울리다]의 의미 별 사용 빈도

'치다'와 '울리다'가 이와 같이 문형 별 출현 분포에 차이가 나는 것은 '치다'가 비교적 제한된 명사 목록과 공기하는 것과는 대조적으로 '울리다'의 경우는 공기하는 명사의 범위가 고루 분포되어 있는 것과 관련이 있다. 이는 일본어 대역어를 통해서도 명백히 나타나는데, '울리다'의 경우 이에 대응하는 일본어는 'あげる', 'かける', '差し上げる', '申し上げる', '建てる' 등 총 13개의 동사와 대응하는데 비해 '치다'의 경우 '入れる', '打つ', '討つ', '生む', '起こす', '切る', 'する', '叩く', '造る', '付ける' 등 총 49개의 일본어 동사와 대응하고 있다. 이는 '치다'의 경우 주된 의미 영역이 '울리다' 보다 미약하고 명사와의 결합 관계가 비교적 강한 관용표현에 가까운 것이 많은데 비해서 '울리다'의 경우 '치다'와는 달리 일반적인 표현을 이루는 경우가 많고 따라서 대응하는 일본어 동사도 적게 나타난다는 데 따른 것이라고 할 수 있다. '치다'와 같은 경우는 코퍼스의 수록 분야 및 수록 내용에 따라 출현빈도가 달라질 가능성이 높기 때문에 문형에 중요성 정보를 기술하기가 매우 어렵다.

6. 결론

본 연구에서는 현재 개발 중인 구문의미사전에 대해 필요 기술 정보를 중심으로 그 구체적 실제에 대하여 논하였다. 그것은 기존의 구문사전과는 달리, 동사와 공기하는 명사의 모든 목록을 망라할 필요가 있다는 점에 특징이 있다. 이는 대역구문의미사전의 구축과 한국어를 대상으로 한 구문의미사전의 의미 기술의 명시성을 확보하기 위해서 중요한 의미를 갖는다. 또한 상위개념이라고 하는 시소러스에 의한 애매성 해소의 한계를 극복하는 데에도 유용한 정보를 제공하게 된다는 의미에서 매우 중요한 작업이라고 할 수 있다. 또한 기존의 구문사전과는 달리 표제어의 등재 형식도 관련어 정보, 출현 어형 정보 등의 기술이 필요하다는 것이 밝혀졌다.

현재 구축된 구문의미사전은 아직 실험을 통한 검증을 거치지 않은 불완전한 것이다. 앞으로 의미해석, 기계번역 등의 실험을 거쳐 유용성을 검증하고 실용적인 사전으로 개량해 갈 예정이다.

참고문헌

[1] [田中穗積 1999] 田中穗積監修, 自然言語処理-基礎と応用, 社団法人電子情報通信学会編, pages 172, 1999.

[2] [한나리 1999] 한나리, "한국어 구문사전 작성의 실제", 한국과학기술원 전산학과 여름 발표회 자료, 2쪽, 1999년 7월 23일

[3] [정희창 1999] 정희창, "'담배를 피우다'인가? '담배를 피다'인가?", 새국어생활, 3호, 국립국어연구원, 1-3쪽, 1999년

[4] [남지순 1999] 남지순, "동사의 의미 유형 분류", STEP2000 2차년도 보고서, 한국과학기술원,

1999년8월31일
[5] [4]와 동일
[6] [4]와 동일
[7] [채영숙 1999] 채영숙, 'KIBS I에서 구축한 코퍼스와의 비교 분석', STEP2000 2차년도 보고서, 한국과학기술원, 1999년 8월 31일.

감사의 글

본 연구는 과기부 특정연구개발과제 [대용량 국어정보 심층 처리 및 품질 관리 기술 개발]에 의해 수행되었다. 연구에서 사용한 한국어 구문의미사전의 1차 자료는 한국외국어대학교 대학원생인 박용일군, 동 대학교 대학원 졸업생인 이정민양, 연세대학교 대학원 불어불문학과 졸업생 설가영양, 이화여자대학교 국어국문학과 졸업생 황혜리양, 동 대학 대학원 재학 중인 한나리양에 의해 작성되었다. 또한 한국어 구문의미사전의 1차 교정본과 한일대역구문의미사전은 필자가 담당했다. 끝으로 한국과학기술원 전산학과 박사과정 이운재, 김창현군의 조언을 잊을 수 없다. 이 자리를 빌어 깊은 감사의 말씀을 드리고 싶다.