동사 어휘의미망의 반자동 구축을 위한 사전정의문의 중심어 추출

김혜경·윤애선*†
부산대학교

Hae-Gyung Kim, Aesun Yoon. 2006. The Extraction of Head words in Definition for Construction of a Semi-automatic Lexical-semantic Network of Verbs. *Language and Information 10.1*, 47–69. Recently, there has been a surge of interests concerning the construction and utilization of a Korean thesaurus. In this paper, a semi-automatic method for generating a lexical-semantic network of Korean '-ha' verbs is presented through an analysis of the lexical definitions of these verbs. Initially, through the use of several tools that can filter out and coordinate lexical data, pairs constituting a word and a definition were prepared for treatment in a subsequent step. While inspecting the various definitions of each verb, we extracted and coordinated the head words from the sentences that constitute the definition of each word. These words are thought to be the main conceptual words that represent the sense of the current verb. Using these head words and related information, this paper shows that the creation of a thesaurus could be achieved without any difficulty in a semi-automatic fashion. (**Pusan National University**)

Key words: 시소러스 (thesaurus), 워드넷 (WordNet), 어휘의미망 (lexical-semantic network), 중심어 (head word), '-하' 동사류 ('-ha' verb), 부가어 (additional word), 서술성 명사 (predicative noun)

1. 머리말

최근 들어 어휘의미망(Lexical-Semantic Network) 및 시소러스(thesaurus)에 대한 관심이 높아지고 있다. 외국에서는 물론, 국내에서도 다양한 방법으로 이에 대한 연구가

* 주저자(김혜경): 부산대학교 인지과학협동과정, (609-735) 부산 금정구 장전동 산 30 부산대학교. E-mail: haegyungk@gmail.com
교신저자(윤애선): 부산대학교 불어불문학과, 인지과학 협동과정, (609-735) 부산 금정구 장전동 산 30 부산대학교. E-mail: asyoon@pusan.ac.kr
† 이 논문의 어휘의미망에 사용된 개념명은 '코어넷(CoreNet)'에 기초한다. 데이터 작업을 위해 '코어넷'의 활용을 허락해주신 한국과학기술원(KAIST) 최기선 교수님과 코텀(KORTERM)에 깊이 감사드린다.

진행되고 있으며, 다년간의 연구를 바탕으로 한 시스템 개발이 활발히 이루어지고 있다. 대표적인 것으로는 미국의 프린스턴(Princeton) 대학에서 영어를 대상으로 구축한 '워드넷(Wordnet)' (Fellbaum, 1998)과 유럽에서 이 워드넷의 1.5 버전을 모형으로 유럽 8개 국어를 대상으로 구축한 다국어 어휘의미망인 '유로워드넷(EuroWordNet)' (Vossen, 2005) 등이 있다. 아시아에서는 일본과 중국에서 구축한 어휘의미망이 대표적이다. '하우넷(HowNet)' (Dong and Dong, 2006)은 중국어와 영어를 대상으로 하여 보편적인 의미체계를 구성하고자 한 개념망이며, 일본의 NTT사(Nippon Telegraph Telephone Corporation: 일본 전신 전화사)의 '어휘대계'는 기계번역 시스템의 번역사전 중에서 일본어 의미사전에 관한 부분으로, 단어의미속성체계, 단어의미사전, 구문의미사전으로 구성되어 있다. 이러한 어휘의미망은 명사뿐만이 아니라 동사나 형용사, 부사 등의 여타 다른 품사를 이용해서 개념체계(conceptual system)를 구축하고 그 결과물을 발표하고 있다.

어휘의미망은 일반적으로 명사, 동사, 형용사 등 품사적 기준에 따라 개념체계를 분리하고 각 품사가 하나의 체계를 이루고 있다. '워드넷' 명사의 경우, 약 8만 개의 '동의어집합(synsets)'이 12개의 층위(level)를 이루고 있으며, 동사는 명사와 별도로 약 1만 3천여 개의 동의어 집합이 4개의 층위로 형성되어 있다. '어휘대계'의 명사는 12개 층위의 2,710개 개념이 어휘의미망을 이루고 있으며, 37만개의 명사가 각 개념에 군집화(grouping)된 형태로 이루어져 있다. '어휘대계'에서도 동사는 품사적 기준에 의해 별도로 분류되어 있으며, 4개 층위의 26개 개념 노드(node)를 지니고 있다. 즉 근래의 어휘의미망에 대한 대부분의 연구 결과물에서 어휘가 1차적으로는 품사를 나타내는 형태적 기준으로 구분되고, 다음으로 그 각각의 범주 내에서 개념을 중심으로 다루어지고 있다.

어휘의미망은 자연언어처리(Natural Language Processing) 과정에서 발생할 수 있는 어휘 간의 개념으로 인한 문제를 효율적으로 극복하기 위해 구상된 체계이다. 개념체계는 특정분야, 혹은 일반분야에서 지구상에 존재하는 언어로 표현될 수 있는 다양한 개념에 명칭을 붙이고 그 개념명의 서로 유기적인 관계에 대한 설명을 나무구조(tree-structure) 등으로 설명하는 체계이다. 여기서 개념체계의 구성요소가 되는 개념명의 '개념(concept)'은 실세계의 언어인 단어가 지닐 수 있는 '의미(sense)'이다. 따라서 서로 다른 품사일지라도 같은 뜻을 내포하고 있는 단어의 쌍이라면, '의미'는 개념체계 내에서의 위치가 변하거나 서로 다를 수 없다. 어휘의미망에서 먼저 고려되어야 하는 것은 어휘에 대한 품사적 기준이 아니라 어떤 의미를 갖고 있느냐 하는 개념의 문제이다. 만약 두 어휘가 품사 구분은 서로 다르지만 같은 개념을 지니고 있다면, 이 두 어휘는 하나의 개념체계 내에서 같은 개념명을 부여받을 수 있어야 한다. 이는 특히 서술성명사(predicative noun)와 기능동사(support verb)의 쌍에서 그 개념적 특성으로 인해 더욱 확연히 드러난다. 서술성명사와 기능동사의 쌍에서 동사는 한 문장을 이루기

위해 사용되는 형태, 통사적인 도구로써 시제(tense)나 인칭(person), 수(number), 그리고 상(aspect) 등을 나타내는 표지 역할만을 담당하는 반면, 서술성명사는 전체 서술성명사와 기능동사 쌍의 의미를 대표하는 개념의 핵요소(nuclear element)로 작용한다 (Gross, 1981; 김혜경, 1996).

본 연구에서는 서술성명사와 기능동사의 쌍 중에서 가장 대표적인 '[-하]동사류'를 중심으로 하나의 개념체계 아래에 명사와 동사를 모두 통합할 수 있는 범(凡)품사적인 어휘의미망을 구축하고자 한다. 또한 기존의 어휘의미망 구축방식이 거의 모든 과정을 작업자의 직관이나 수작업에 의존해 왔다면, 본 연구에서 구현하고자 하는 동사 어휘의미망은 사전정의문의 중심어(head word) 개념을 통해 반자동으로 보다 효율적인 방식으로 구축하고자 한다.

먼저 2장에서는 어휘의미망에 대한 선행 연구를 분류기준에 따라 나누어 설명하고, 대표적인 국내외의 어휘의미망을 구체적으로 살펴봄으로써 발표된 연구 결과에서 나타나는 여러 가지 문제점을 짚어본다. 3장에서는 본 연구에서 구축하게 될 동사 어휘의미망의 연구 내용에 대해 소개하고, 최종 결정된 데이터의 범위와 그 근거에 대해 기술하고자 한다. 다음으로 4장에서는 동사 사전정의문의 특성을 분석하여 정제(filtering)하고 각각의 사전정의문에 알맞은 형태적 제약 정보를 이용하여 중심어를 추출해 가는 과정에 대해 기술할 것이다. 이 과정에서 문장의 유형에 따라 제시된 여러 가지 규칙과 준거를 설명한다. 마지막으로 5장에서는 결론과 함께 본 연구가 갖는 의의와 구축된 동사 어휘의미망을 통한 향후 연구 방향에 대해 소개할 것이다.

2. 선행 연구 및 문제점

어휘의미망에 대한 연구는 [표 1]에서 보는 바와 같이 분류기준에 따라 다양하게 나누어 볼 수 있다.

분류기준	종류
언어별	단일어 어휘의미망
	다국어 어휘의미망
구축방식	직접 구축방식
	간접 구축방식
	하향식(top-down) 구축방식
	상향식(bottom-up) 구축방식
지역별	국내 어휘의미망
	국외 어휘의미망

[표 1] 분류기준에 따른 어휘의미망의 종류

지역적으로는 국내 혹은 국외의 연구에 따라 국내 어휘의미망과 국외 어휘의미망으로 나눌 수 있고, 그 언어별로는 한국어나 영어처럼 단일어(monolingual)를 대상으로

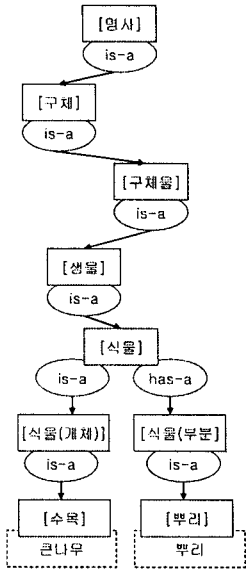
하느냐 혹은 두 개 이상의 다국어(multilingual)를 대상으로 하느냐에 따라 단일어 어휘의미망과 다국어 어휘의미망으로 나눌 수 있다. 또한 연구의 기초 자료로 기존의 다른 어휘의미망을 사용하는지에 따라 직접 혹은 간접 구축방식으로 나눌 수도 있다. 이 중 직접 구축방식은 최상위 개념에서 시작하여 각 개념에 대한 정의를 내리고 그것에 대한 하위개념체계를 구성해 나가면서 결과물을 만들어내는 하향식(top-down)과 데이터를 기반으로 구축해나가는 상향식(bottom-up)으로 구분할 수 있다. 상향식 구축방식 중 하나는 현 시점에서 사용되고 있는 모든 용어를 담고 있는 사전이나 말뭉치를 이용하는 것이다. 말뭉치나 사전정의문을 이용하여 의미를 추출하고, 그 의미에 기초하여 최종적인 개념체계의 모습을 완성하는 것이다. 이는 실제 데이터에 기초하여 추출한 자료를 토대로 하므로, 개념체계의 실용적인 타당성을 가지고 허실이 없다는 장점을 지닌다.

국외의 가장 대표적인 어휘의미망에 대한 연구인 미국 프린스턴 대학의 '워드넷은 하향식 구축방식에 의해 만들어진 어휘의미망이다. 워드넷은 사전편찬가에 의해 사전편찬명(lexicographer's file name)을 시작으로 수동으로 구축되었다. 예를 들어 명사에서는 act, animal, artifact 등 25개의 '원천개념(unique beginners)'을 이용하여 11개의 '최상위개념'을 먼저 단계별로 구성하고, 다음으로 이를 공유하는 하위어를 구성하는 방식이다. 동사는 body, change, cognition 등의 14개 원천 개념을 지닌다 (이은령·황순희·윤애선, 2004). 워드넷은 또한 동의어집합으로 개념체계의 관계를 정의하고 있는데, [표 2]에서 보는 바와 같이 2.1버전을 기준으로 명사는 81,426개, 동사는 13,650개의 동의어 집합을 지니고 있다.

품사	동의어집합	원천 개념 수	최대 계층 수
명사	81,426개	25	12
동사	13,650개	14	4
형용사	18,877개	-	-
부사	3,644개	-	-
계	117,597개		

[표 2] 워드넷 2.1의 동의어집합 구성

하향식 구축방식은 일본어 어휘대계에서도 마찬가지로 나타난다. 일본어 어휘대계는 일본어-영어 기계 번역시스템인 'ALT-J/E(Automatic Language Translator - Japanese to English)'의 번역사전 중에서 일본어 의미사전에 관한 부분을 출판 형태로 정리한 것이다. 일본어 의미사전 중 명사 개념체계에 해당하는 단어의미속성체계는 2,710개의 노드로 구성되어 있다. 이 체계의 특징은 명사의미체계에서 상-하위 관계(is-



[그림 1] 일본어 어휘대계의 관계 유형

a 관계)와 전체-부분 관계(has-a 관계)가 구분되어 있다는 것이다. [그림 1]은 그 예이다 (Ikehara and others, 1997).

국내의 대표적인 어휘의미망에 대한 연구는 한국어와 일본어, 중국어의 다국어에 기반하여 '전문용어언어공학연구센터(KORTERM, 이하 '코텀'으로 표기함.)'에서 구축한 '다국어 어휘의미망' (한국과학기술원 전문용어언어공학연구센터, 2005)인 '코어넷(CoreNet)'이다. 또한 워드넷 2.0 버전을 바탕으로 한국어에 맞게 수정하고 보완되어 구축 중인 부산대의 'KorLex'와 한국어사전에서 상위어 개념을 자동으로 추출하고 이를 이용하여 의미 계층 구조를 만들어낸 울산대의 'U-WIN(UOU-Word Intelligent Network)'이 있다 (옥철영, 2005; 최호섭·옥철영, 2002). 'U-WIN'이 한국어 사전의 사전정의문을 이용한 단일어의 상향식 구축방식을 사용한데 비해, '코어넷'과 'KorLex'는 각각 '한국어-중국어-일어'와 '한국어-영어'의 다국어를 바탕으로 일본어 '어휘대계'와 영어의 '워드넷'에 기반한 간접 구축방식을 취하고 있다.

'KorLex'는 영어의 '워드넷'에서 시작된 것으로 명사와 동사 등이 각 품사별로 서로 다른 개념체계를 이루고 있으며, 어휘의미는 동의어집합 관계로 아주 세밀하게 연결고리를 구성하고 있다. 현재 명사 약 3만, 동사 약 5천 개 어휘의미에 대한 한국어 어휘의미망의 1차 버전이 완성 단계에 있으며, 계속해서 대역에서 누락된 어휘의미를 추가 확장할 뿐 아니라, 한국어에서 미분화된 어휘의미를 분화하고 영어에서 과분화된 어휘의미를 통합하며 중복된 어휘의미를 조정하는 등의 여러 단계의 작업을 통해 한국어 어휘의미망의 모습을 완성해 나가고 있다 (이은령·윤애선, 2005).

'코어넷'은 초기에 일본어 '어휘대계'의 명사의미체계 부분의 2,710개의 개념명과 노드 규칙을 한국어로 번역하는 것부터 시작되었다. 이후 한국어에 맞게 개념명을 적절하게 수정하기도 하고 추가하여 보완하기도 하여, 최종 2,937개의 개념명을 지닌 한국어 어휘의미망을 만들었다. '코어넷'에서 눈에 띄는 점은 명사, 동사, 형용사가 총 2,937개의 개념명을 지닌 하나의 어휘의미망에 개념을 중심으로 분포된다는 것이다 (이주호·은광희·최기선, 2001; 한국과학기술원 전문용어언어공학연구센터, 2005). 예를 들어 명사 '감'과 동사 '가다', '공부'와 '공부하다' 등의 동사의 명사형과 동사, 서술성명사와 동사 쌍은 물론, '연습'과 '(솜씨 따위를) 익히다' 등의 단어쌍에서 동사 '익히다'도 명사 '연습'과 같은 노드에 포함되어 있다. 이 예의 명사는 모두 '추상'의 개념에 속하며, 동사 어휘의미망에서 구축되는 어휘 일부가 수정과 추가의 과정을 거치기는 했지만 모두 의미적으로 각 명사쌍과 함께 '추상'의 개념에 포함되어 품사 구분을 배제한 '개념'으로 연결되어 있다.

그러나 코어넷에 있어서도 개념명 부여의 가장 결정적인 역할은 사전편찬가에 의해 이루어졌으며 거의 모든 구축과정이 최종 공정인 수동적 검토와 작업자의 개념명 부여에 의존해야 했다. 이 과정에서 대부분의 개념명 부여 작업이 전문가의 주관적 판단에 의존했을 뿐만 아니라, 객관적인 판단 기준이나 테스트셋(test-sets)을 제시하지 못했던 점이 코어넷의 가장 큰 문제점으로 남는다. 또한, 코어넷의 동사는 명사와는 달리 코퍼스에서 기본어휘를 추출하여 개념명을 부여하였기 때문에 그 개수가 크게 부족하므로,[1] 동사 어휘의미망의 확장이 절실히 요구된다.

3. 연구 내용과 범위

본 절에서는 본 연구를 위해 다루게 될 데이터베이스의 주된 연구 내용과 그 범위에 대해 기술할 것이다. 3.1절에서는 명사와 동사가 하나의 범품사적인 어휘의미망으로 통합되기 위한 '[-하]동사류'의 연구 내용을 제시할 것이다. 3.2절에서는 다양한 기준을 통해 데이터베이스의 구체적인 연구 범위를 제시할 것이다.

3.1 연구 내용

본 연구는 '[-하]동사류'[2]를 연구대상으로 함으로써, 명사어휘의미망과 통합되어 범품사적인 어휘의미망을 이룰 수 있는 동사어휘의미망을 구축하고자 한다.

동사와 명사 모두를 같은 의미 계층으로 구현하면 명사와 동사 간에 패턴의 유사성을 파악할 수 있을뿐더러, 언어 생성의 측면에서도 명사와 동사가 품사라는 형식적인 면과는 상관없이 자유롭게 문장을 생성해 낼 수 있다 (Choi and Bae, 2004). 특히 한국어의 '[-하]동사류'가 그러하다.

(1) ㄱ. 그 선수가 운동을 한다.

---

[1] 명사는 21,401개 어휘(51,607개 어휘의미)인데 반해, 동사는 1,758개 어휘 (5,290개 어휘의미), 형용사는 813개 어휘 (2,801개 어휘의미)로 동사는 명사의 어휘로는 8.2%, 어휘의미로는 10.3% 수준이다.

[2] '[-하]동사류'의 명사가 서술성명사인지의 여부는 오래전부터 언어학에서 논의와 연구가 진행되고 있다. 프랑스에서는 1960년대 말 Gross (1981)를 중심으로 자연언어의 기계적 처리를 위한 언어 자료 축적을 목적으로 시작된 경험주의적 언어 이론인 어휘문법(lexique-grammaire)에서, 국내에서는 1970년대 서정수 (1975)의 '하'에 대한 여러 가지 변형테스트를 통한 연구를 시작으로 연구가 진행되고 있다. 예를 들어 다음의 1)과 2)의 예를 보자.

1) 그 선수가 운동을 한다.

2) 그 분이 머리를 한다.

1)은 다시 '관계절화 변형'이나 '명사절화 변형'을 통해 1′)와 1″)로의 변형이 가능하지만 2)는 2′)와 2″)에서 보듯이 명사절화 변형이 성립되지 않는다.

1′) 그 선수가 하는 운동

1″) 그 선수의 운동

2′) 그 분이 하는 머리

2″) *그 분의 머리

즉, 1)의 '운동'은 서술성명사로 쓰인 예이고 2)의 '머리'는 일반명사이다.

ㄴ. 그 선수가 하는 운동

'[-하]동사류'의 '-하' 기능동사와 그 서술성명사로 연관되어진 (1ㄱ)과 (1ㄴ)은 명제와 명사구의 서로 다른 문장 구조이지만 동일한 의미를 지닌다. 개념적으로 명사와 동사 간에 연계되어 있는 '[-하]동사류'의 어휘의미망을 구축하는 것은 전체 동사 어휘의미망을 확장할 수 있는 토대가 될 뿐만 아니라, 동시에 '[-하]동사류' 어휘의미망 내에서 동사적 요소를 제외함으로써 명사어휘의미망을 확장하고 보충할 수 있는 계기가 된다.

본 논문은 어휘의미망 구축을 위해 사전정의문의 중심어 개념을 통한 상향식 방식을 채택한다. 어휘의 의미별로 분화된 사전정의문을 형태소 분석하여 구문별 혹은 단어별로 분석하고 분석된 결과에 근거하여 중심어를 추출한다. 추출된 중심어에 기반하여 개념명을 부여하고 부여된 개념명을 기반으로 하여 동사 어휘의미망을 구현한다. 연구 대상이 되는 사전정의문은 한글학회 <우리말큰사전>의 데이터베이스[3]에 기초한다.

3.2 연구 범위

본 논문의 데이터베이스를 위해 제공되는 <우리말큰사전>에 등재된 '[-하]동사류'는 총 35,100개이다. 다음 기준에 의해 최종 3,656개의 데이터 리스트를 결정하고 연구 범위를 제한한다.

첫째, 의성어와 의태어는 연구대상에서 제외한다. 본 연구에서처럼 사전정의문으로 중심어를 추출하여 개념명을 부여하는 방식에서는, 의성어와 의태어가 개념체계의 특정 개념명 몇 군데에 집중되어 분포된다. 다음의 [표 3]에서 그 예를 살펴보자.

표제어	표제어번호	품사정보	의미번호	사전정의문
삐걱하다	0	자동사	0	크고 딴딴한 물건이 서로 닿아 가볍게 갈리는 소리가 한 번 나다.
왱왱하다	0	자동사	0	벌이나 돌팔매 따위가 빨리 날아가는 소리가 자꾸 나다.

[표 3] 의성어의 예

[표 3]의 두 표제어는 의성어인 '삐걱하다/0/자동사/0'[4]과 '왱왱하다/0/자동사/0'로, 4장에서 다루게 될 중심어 추출 방식에 따르면 그 중심어가 '소리가 나다.'가 된다. 의성어와 의태어로 추출된 3,839개의 목록 중에서 2,350개의 목록이 '소리'와 관련된 중심어를 지니는 의성어였다. 의태어는 의성어처럼 일관된 하나의 중심어를 지니지는 않는다. 하지만, 몇몇 개념명으로 의태어를 모두 설명하는 것이 가능하다. 예를 들어 [표 4]에 나타나는 의태어의 두 예를 서로 비교해 보자.

[3] '코텀'에서는 1998년 한글학회로부터 연구용으로 <우리말큰사전>의 파일을 받았다. 이후 2000년도에 그것을 데이터베이스화하여 총 448,430개의 어휘의미가 있는 <우리말큰사전> 데이터베이스를 구축하였다.

[4] 본문에서는 '표제어'와 함께 나오는 '표제어번호', '품사정보', '의미정보'에 대한 정보는 지면상 '표제어/표제어번호/품사정보/의미번호'와 같은 형태로 표기한다.

표제어	표제어번호	품사정보	의미번호	사전정의문
번적번적하다	0	자타동사	0	큰 빛이 잠깐 악하게 자꾸 빛나다.
북적북적하다	0	자동사	0	사람들이 많이 모여 움직이며 매우 수선스럽게 자꾸 들끓다.

[표 4] 의태어의 예

[표 3]의 의성어가 '~한 소리'를 나타내는 어휘의 집합이라면, [표 4]의 의태어는 '~한 모습' 혹은 '~한 모양'을 나타내는 어휘의 집합이다. 즉 '소리'와 '모습', 또는 '모양'의 개념명에 의성어와 의태어를 하위개념으로 분화시킬 수 있다.

둘째, <우리말큰사전>의 데이터베이스에 나타나는 총 35,100개의 '[-하]동사류' 중 기본어휘에 속하는 어휘를 추출하여 대상으로 삼았다. 사전에는 일상생활에서는 거의 사용되지 않으며, 사전정의문을 살펴보아도 그 뜻을 쉽게 이해할 수 없는 단어가 많이 있다. 다음 [표 5]의 '가반하다/0/타동사/0'과 같은 예가 여기에 속한다.

표제어	표제어번호	품사정보	의미번호	사전정의문
가반하다	0	타동사	0	절에서, 음식을 여러 몫에 도르고 나서 남을 때에 다시 그것을 더 도르는 일.

[표 5] '가반하다/0/타동사/0'의 예

[표 5]의 표제어가 나타내고자 하는 의미는 사전정의문만으로는 파악하기 힘들다. 사전정의문에 쓰인 어휘를 다시 사전을 통해 검색해봐야 의미를 파악할 수 있는 일상생활에서는 찾아보기 힘든 단어이다. 구축하게 될 동사어휘의미망은 현대 한국어 화자가 실질적으로 사용가능한 어휘의미망을 구현하고자 한다. 따라서 대상이 되는 어휘 리스트를 기본어휘에 속하는 어휘로 제한하기로 한다.

기본어휘라 함은 빈도순과 기초어휘의 두 가지 측면을 고려해 볼 수 있다. 한 나라의 언어를 말하고 이해하는 데에 필수적인 기초어휘와 고빈도 어휘를 통틀어 기본어휘라고 한다 (한국과학기술원 전문용어언어공학연구센터, 2000, 15쪽)는 가정하에, 대상 어휘 리스트를 실제 말뭉치의 빈도수(frequency)에 기반하여 고빈도의 어휘를 추출하는 방식과 국립기관에서 권장하는 교육용 어휘에 기반하여 기초어휘를 추출하는 두 가지 방식에서 추출된 데이터의 합집합을 사용하는 방식을 택했다. 빈도수를 위해 사용된 말뭉치는 '카이스트'에서 보유하고 있는 4000만 어절 형태소 분석 말뭉치(tagged corpus)와 '21세기 세종계획'(이하 '세종'으로 표기함.)의 결과물인 형태소 분석 말뭉치 1000만 어절(공개용 550만 어절과 비공개용 450만 어절)이다. 그 중 빈도수 '3' 이상인 어휘를 대상으로 대상 항목을 선정하였다.

사용된 말뭉치는 의미분석된 말뭉치(sense-tagged corpus)가 아니므로 표제어 대

표제어의 대응으로 빈도순 리스트를 완성하였다. 예를 들어 다음의 [표 6]에서 하나의 표제어인 '발전하다'는 서로 다른 두 개의 어휘의미 모두가 대상 항목이 된다.

표제어	표제어 번호	품사 정보	의미 번호	사전정의문1	사전정의문2[5]	카이스트 말뭉치 빈도수	세종 말뭉치 빈도수
발전하다	1	자동사	0	발전1	더 잘 되거나 나아지거나 활발해지거나 하는 일.	1146	864
발전하다	2	자동사	0	발전2	전기를 일으킴		

[표 6] '발전하다'의 서로 다른 두 개의 의미

즉 위의 [표 6]에서 '발전하다'는 <우리말큰사전>에 따르면 '발전하다/1/자동사/0'과 '발전하다/2/자동사/0'의 두 가지 의미를 지니지만, 본 연구에서는 두 가지 경우 모두를 '카이스트' 말뭉치의 빈도수 '1146'과 '세종' 말뭉치의 빈도수 '864'로 본다.

다음으로, 기초어휘 리스트는 국립국어원에서 2003년 5월에 발표하여 배포한 '한국어 학습용 어휘 목록'을 사용하였다.[6] '한국어 학습용 어휘 목록' 역시 일부 의미구분은 하였으나 <표준국어대사전>의 동음이의어 구분만 하고 있고 구분정보는 한자만 있는 등, 아직 충분한 구분정보를 담고 있지 못하다. 따라서 기초어휘도 빈도수와 마찬가지로 의미구분을 하지 않았으며 표제어별로 그 대상 리스트를 선정하였다. '한국어 학습용 어휘 목록'에는 의미구분을 기준으로 총 5,965개의 단어의 기초어휘가 수록되어 있으며, 동사는 1,345개, 그 중에서도 '[-하]동사류'는 424개였다.

마지막으로, 기구축된 어휘의미망에서 서술성명사를 통해 자동으로 개념명을 부여할 수 있는 경우는 제외한다. 본 연구를 통해 구축되는 '[-하]동사류' 어휘의미망은 '코어넷'의 어휘의미망을 확장하는 연구이다. 기구축된 '코어넷'의 어휘의미망에는 다수의 서술성명사 항목을 포함하고 있다. 기구축된 '코어넷'의 서술성명사에 부여된 개념명으로 '[-하]동사류'의 개념명을 대체할 수 있는 경우는 간접적이긴 하지만 동사의 개념명이 이미 구축되었다 할 수 있으므로 본 논문의 연구대상에서 제외한다. 의성어와 의태어를 제외하고 기본어휘로 추출된 총 8,489개의 '[-하]동사류' 중 이 같은 경우는 4,833개 어휘의미였다. 따라서 본 논문의 대상 데이터는 최종 3,656개 어휘의미의 '[-하]동사류'로 제한된다.

[5] '사전정의문1'과 '사전정의문2'의 구분은 사전정의문이 명사로의 링크만을 담고 있는 경우, 명사로의 링크정보만으로 구성된 사전정의문을 '사전정의문1', 링크된 명사로 사전정의문을 다시 찾아 최종 문장 형태로 연결된 사전정의문을 '사전정의문2'라 명명한다. 즉, '발전하다/1/자동사/0'는 '사전정의문1'에서 '→발전1'로 구성되어 있으며, 최종 '사전정의문2'의 정보를 통해 '더 잘 되거나 나아지거나 활발해지거나 하는 일.'이라는 정의문을 얻게 되었다.

[6] http://www.korean.go.kr/

4. 사전정의문을 이용한 중심어 추출

명사의 사전정의문은 대부분 특질소(differentia)[7]와 상위어로 구성되며, 상위어가 중심어 역할을 한다. 그러나 동사의 사전정의문은 명사와는 달리 상위어 성분을 지니지 않는다. 4.1절에서는 동사 사전정의문의 중심어에 대한 정의를 내리고, 4.2절에서는 본 논문의 대상 데이터가 될 '[-하]동사류'의 사전정의문에 나타나는 특성을 통해 사전정의문의 수의적인(arbitrary) 요소를 정제하는 방법에 대해 기술한다. 다음으로 4.3절에서는 정제된 사전정의문을 형태소 분석하고 문장의 역방향(reverse) 분석을 이용한 여러 가지 형태적 제약 정보를 통해 중심어를 추출하는 과정에 대해 설명한다.

4.1 동사의 사전정의문 분석

사전에서 단어는 화자가 그 단어의 의미를 명시하는 방식에 의해 또 다른 단어로 정의된다. 명사의 사전정의문은 그 표제어의 의미적 특징(semantic feature)을 설명하는 특질소와 해당 표제어에 대한 상위어(superordinate)의 쌍으로 이루어진 것이 보통이다 (Fellbaum, 1998). 한국어에서 그 예로 [표 7]에서처럼 <우리말큰사전>에 나타나는 '사전/12/명사/0'의 사전정의문을 살펴보자.

표제어	표제어번호	품사정보	의미번호	사전정의문
사전	12	명사	0	어떤 언어의 낱말 들을 모아 일정한 차례로 벌여 그 맞춤법, 발음, 말밑, 말본 형태, 뜻, 쓰임 따위를 보이는 책.

[표 7] 명사의 사전정의문

명사의 사전정의문에서 중심어란 사전정의문이 전달하고자 하는 의미의 핵심을 이루는 단어 혹은 구를 말한다. [표 7]에서, '사전/12/명사/0'은 '책'의 여러 종류 중의 하나로서 '책'의 하위어에 속하며, 또한 '책'은 표제어 '사전'의 상위어인 동시에 사전정의문 전체의 중심어가 된다. 그 외의 '어떤 언어의 낱말을 모아 일정한 차례로 벌여 그 맞춤법, 발음, 말밑, 말본 형태, 뜻, 쓰임 따위를 보이는'은 '사전'이라는 표제어의 특징을 열거하는 말이 된다. 즉, 일반적으로 명사의 사전정의문에 있어 중심어란, 사전정의문의 끝부분에 오는 표제어의 상위어이다 (문유진, 1996).

그러나 동사의 중심어는 명사의 경우와는 다르다. <우리말큰사전>에 등재된 '[-하]동사류'의 사전정의문을 모두 살펴보는 과정에서 경험적으로(heuristic) 얻은 결과로는, 동사 사전정의문에서는 명사 사전정의문의 특질소 자리에 양태낱말(manner word)을 사용하며, 동사 사전정의문의 중심어는 표제어의 상위어가 아닌 '풀이말'에 가까운 성격을 지니고 있다.

[7] 사전정의문에서의 '특질소'란 상위개념과 그 개념을 다른 개념과 구분지어 주는 요소로, 상위어에 의미론적 특성(semantic feature)을 가미한 단어들을 말한다 (문유진 (1996)).

표제어	표제어번호	품사정보	의미번호	사전정의문
간택하다	2	타동사	0	여럿 가운데서 특별히 가리다.

[표 8] 동사의 사전정의문

위의 [표 8]의 사전정의문에서 '여럿 가운데서 (특별히)'는 양태소이며, '가리다'는 표제어 '간택하다'와 동의관계에 있으면서도 의미전달이 좀 더 용이한 기본어휘에 속하는 단어(general word)[8]인 '풀이말'이다. 따라서, 동사 사전정의문에 있어서는 사전정의문의 끝부분에 오는 표제어의 '풀이말'을 '중심어'라 정의하기로 한다.

4.2 사전정의문의 정제

본 절에서는 사전정의문에서 정제될 수의적인 요소로 접속부사와 부가어(additional word)에 대해 살펴볼 것이다. 이 요소는 사전정의문이 나타내고자 하는 의미 전달에서 잉여적(redundant)이거나 잉여적임을 판별하는 척도가 된다.

가. 접속부사 '또는'

'[-하]동사류' 사전정의문에 나타나는 가장 먼저 정제되는 요소는 접속부사 '또는'을 중심으로 두 개의 문장이 선택적으로 병렬하여 나열되는 경우이다. 사전정의문에서 이 부사는 의미적으로 부사의 양쪽에 쓰인 내용이 모두 같은 정도의 비중을 지닌다. 다음의 [표 9]는 그 중 하나의 예이다.

표제어	사전정의문1	사전정의문2
가늠하다/0/타동사/0	가늠	목표나 기준에 맞고 안 맞음을 헤아려 봄. 또는, 그렇게 헤아려 보는 목표나 기준.

[표 9] '[-하]동사류' 사전정의문 중 접속부사 '또는'이 쓰인 예

[표 9]의 '사전정의문2'는 '또는'을 중심으로 앞과 뒤가 마침표('.')를 지닌 완전한 두 개의 사전정의문으로 이루어져 있다. '또는'을 중심으로 하나는 서술적 의미를, 다른 하나는 명사적 의미를 지니면서 서술적 의미를 행하는 일이나 내용 등을 지칭한다. 예를 들면 [표 9]에서 '~을 헤아려 봄'과 '~는 목표나 기준'의 예이다. '또는'이라는 접속부사는 '그렇지 않으면'[9]이라는 의미의 부사로, 흔히 둘 중 하나에 대한 선택의 의미를 나타낸다. 즉, 두 가지 사전정의문 중 하나만으로도 그 의미 전달이 가능하다는 뜻이 된다. 의미 전달에 있어서 반복적인 요소는 수의적인 요소로 보아 삭제하되 본 논문의 목

[8] 워드넷에서도 똑같은 방식의 경험적 정의가 이루어졌다 (Fellbaum, 1998). "...예를 들어 명사는 '(x is) a kind of'의 공식에 의해 보통 상위어로 정의된다....(중략) 동사는 좀 더 일반적인 또 다른 동사로 정의된다."

[9] <표준국어대사전>, <연세한국어 사전>, 그리고 <우리말큰사전>의 부사 '또는'에 대한 사전정의문을 참조하였다.

적이 동사의 어휘의미망 구현을 위한 것이므로 4절의 중심어 추출시 효율성을 높이기 위해 중심어가 서술성을 지닌 '명사형 전성어미'로 되어 있는 것을 선택하기로 한다. 따라서 [표 9]의 예에서는 '목표나 기준에 맞고 안 맞음을 헤아려 봄'이 최종으로 남겨진 '사전정의문2'가 된다.

나. 부가어

사전정의문에서 부가어란, 사전정의문에서 표현하고자 하는 의미를 전달하는 데 부차적인 요소로서 사전정의문의 중심어를 부연하고 설명해 주는 단어[10]를 일컫는 말이다. 즉, 사전정의문의 의미적인 면에서는 결정적인 역할을 담당하지 않고 중심어의 기능적인 면을 부연해 주는 어휘적 요소를 일컫는다. 다음의 [표 10]에 쓰인 '~을 비유하는 말' 따위이다.

표제어	사전정의문1	사전정의문2
개화하다/1/자동사/0	개화1	어떤 사물이 한창 번성함을 비유하는 말.

[표 10] '개화하다/1/자동사/0'에 나타난 부가어의 예

[표 10]에서 '개화1'에 대한 정의는 '어떤 사물이 한창 번성함'까지로 그 의미전달이 충분히 이루어졌으며, '~을 비유하는 말'은 중심어를 떠받쳐주는 역할만을 담당할 뿐이다. 따라서 이러한 부가어는 사전정의문의 의미 전달을 하는 데 있어 수의적인 요소라 할 수 있다. '[-하]동사류' 사전정의문 전체에서 발견할 수 있는 단일 형태의 부가어는 다음의 [표 11]과 같다.

동사류	사전정의문에 나타난 부가어
'[-하]동사류'	~을 비유하는 말 ~의 뜻 ~이라는 뜻 ~는 뜻 ~의 비유 ~는 말 ~을 비유하는 말 ~ 일컫는 말 ~의 낮은 말 ~을 이르는 말

[표 11] '[-하]동사류'에 나타난 단일 형태의 부가어 목록

[표 11]의 단일 형태 외에 복합 형태의 부가어도 있는데, 다음의 [표 12]의 예를 보자.

[0] 문유진 (1996)에서는 '기능어(functional word)'로 명명한다.

표제어	사전정의문1	사전정의문2
하야하다/0/자동사/0	하야1	'시골로 내려간다'는 뜻으로, '관직이나 정계에서 물러남'을 이르는 말.

[표 12] '하야하다/0/자동사/0'에 나타난 복합 형태의 부가어

[표 12]의 '하야하다/0/자동사/0'의 예는 선행절과 후행절을 지니는 복문 구조를 보이며, 후행절에서는 '~을 이르는 말'이 부가어이다. 선행절은 보충적 의미를 전달하려는 부사절임을 알 수 있는데, 이러한 보충절에 쓰인 '~는 뜻으로'를 포함한 절 전체가 중심어 추출에 영향을 미치지 못하는 수의적인 요소이다. '[-하]동사류'의 사전정의문에 나타난 이와 같은 복합 형태의 부가어는 아래의 [표 13]에서 나타나는 예이다.

동사류	사전정의문에 나타난 복합형태 부가어
'[-하]동사류'	~이란 뜻으로, ~을 비유하는 말 ~서 나온 말로, ~을 비유하는 말 ~는 뜻으로, ~을 비유하는 말 ~는 말로, ~의 뜻 ~의 뜻으로, ~을 이르는 말 ~의 뜻으로, ~을 일컫는 말 ~의 뜻으로, ~의 형용 ~의 뜻으로, ~의 비유 ~이라는 뜻으로, ~을 일컫는 말 ~이라는 뜻으로, ~을 이르는 말 ~이라는 뜻으로, ~이라는 말 ~이라는 뜻으로, ~의 비유

[표 13] '[-하]동사류'에 나타난 복합 형태의 부가어 목록

다. 준부가어

[표 14]에서 보는 바와 같이 '절구질하다/0/자타동사/0'의 사전정의문은 '~는 일'이라는 어휘로 끝맺는다.

표제어	사전정의문1	사전정의문2	형태소분석결과
절구질하다/0/자타동사/0	절구질	절구에 곡식 따위를 넣고 찧거나 빻거나 하는 일.	절구/ncn+에/jca 곡식/ncn 따위/nbn+를/jco 넣/pvg+고/ecc 찧/pvg+거나/ecc 빻/pvg+거나/ecc 하/pvg+는/etm 일/ncn+./sf

[표 14] '절구질하다/0/자타동사/0'의 형태소 분석

하지만 [표 14]에 제시된 사전정의문을 [표 15]의 '헛일하다/0/자동사/0'이 지니는

사전정의문과 비교해 보자.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
헛일하다/0/자동사/0	헛일	쓸데없는 일	쓸데없/paa+는/etm 일/ncn+./sf	일

[표 15] '헛일하다/0/자동사/0'의 예

위 [표 15]의 일반적인 사전정의문의 예와는 달리 [표 14]에서 '일'은 '~는'과 같은 관형형어미(etm)[11]와 함께 쓰여, 바로 앞에 나오는 일반동사(pvg)인 '찧거나 빻다'를 '그러한 일'의 의미를 지닌 명사 '찧거나 빻는 일'로 전환해 주는 역할을 한다. 품사적 역할을 변형하는 면에서는 본 절의 '나'에서 이미 기술한 바 있는 부가어와 유사하다. 하지만, '~는 일'의 어휘를 제외한다면 문장의 중심적인 의미는 일부 동일하다는 특징을 지닌다. 이 같은 '~는 일'의 형태를 '나'에서 기술한 부가어와 구분하여 '준부가어'라 명명하기로 한다. 준부가어가 나오는 사전정의문에서는 준부가어를 제외한 문장으로 중심어를 추출하도록 한다. 준부가어로는 '~는 일' 외에 '~는 짓', '~는 것' 등이 있다.

4.3 형태적 제약 정보를 이용한 중심어 추출

본 장의 중심어 추출 방식은 '사전정의문2'를 포함한 '[-하]동사류'의 최종 사전정의문을 이용한다. 중심어 추출의 대상이 되는 사전정의문은 4.2절에서 제시한 사전정의문의 정제 과정을 통해 한 문장 단위로 제시되었으며, 그 문장을 '카이스트 형태소 분석기'[12]를 이용하여 형태소 분석한 결과물을 사용한다.

중심어를 자동으로 추출하는 기법은 [그림 2]에서 보는 바와 같이 형태소 분석된 사전정의문의 마지막 마침표를 기준으로 역방향으로 형태소를 분석하는 방식으로 진행된다.

[그림 2]의 가로행은 형태소 분석된 사전정의문에 나타날 수 있는 일련의 태그[13]를 지닌 형태소이며, 진하게 표시된 항목이 중심어로 추출되는 형태소이다. [그림 2]에 대한 자세한 설명은 4.3.1절에서 4.3.4절까지 기술되며, 여기서 제시하는 여러 가지 형태적 제약정보에 만족하는 단어를 중심어로 추출하게 된다.

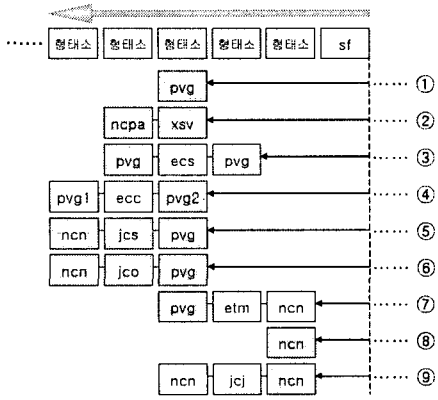
4.3.1 기본 형태. 중심어 추출은 형태소 분석된 사전정의문을 이용한 정보이다. 형태소 분석된 사전정의문은 역방향으로 형태소를 분석해 나가는 방식을 거치게 되며 그 과정에서 기술되는 여러 가지 제약 규칙에 의해 중심어가 추출된다. 본 절에서는 중심어 추출을 위한 가장 기본적인 과정에 나타나는 규칙을 소개한다.

---

[11] '~는' 외에 그 변이형(variant)으로 '~ㄴ', '~ㄹ'이 있다.

[12] http://bola.or.kr

[13] 이때의 태그는 카이스트 태그셋을 사용한다. 카이스트 형태소 태그와 분석 방식에 대한 자세한 사항은 박석문 (2000)을 참조하라.



[그림 2] 사전정의문의 역방향 분석을 통해 중심어를 추출하는 방식

규칙 가-1

형태적 제약 정보를 이용한 중심어 추출의 가장 기본이 되는 제약정보는, 사전정의문의 역방향으로 형태소를 분석하는 과정에서 제일 처음 출현하는 서술어 — '일반동사(pvg)'나 '서술성명사+동사파생접미사 쌍(ncpa+xsv)'[14] — 혹은 명사(ncn)가 그 사전정의문의 중심어가 된다는 것이다. 이 정보는 이하 기술되는 모든 중심어 추출 규칙에서의 기본 규칙이 된다. 예를 들어, [표 16]에서 보는 '망각하다/1/타동사/0'의 '사전정의문2'를 형태소 분석하고, 역방향 분석 과정을 거치게 된다. 이 과정에서 제일 처음 출현하는 서술어는 '잊다'이며, 이 '잊다'가 바로 '망각하다/1/타동사/0'의 사전정의문의 중심어가 된다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과
망각하다/1/타동사/0	잊어버리다①	죄다 잊다.	죄다/mag 잊/pvg+다/ef+./sf

[표 16] 기본적인 제약 정보

규칙 가-2

규칙 가-1에 의해 추출할 수 있는 서술어가 둘 이상으로 이루어진 경우가 있다. 다음 [표 17]에서처럼 '가감하다/1/타동사/0' 혹은 '가격하다/0/타동사/0'는 각각 '더하다'와 '덜다', '치다'와 '때리다'로 서술어 형태가 둘 이상으로 나타난다.

이때는 둘 모두를 중심어로 선택하여 준다. 이 규칙에 준하는 사전정의문은 형태소 분석된 문장을 역방향으로 분석하는 과정에서 추출된 중심어 바로 앞에 대등적 연결어

[14] 카이스트 형태소 분석 방식에서는 서술성의 유무에 따라 '일반동사(pvg)'의 위치에 '서술성명사+동사파생접미사 쌍(ncpa+xsv)'이 쓰인 경우가 있다. 4.3.1절의 제약 정보에서 쓰이는 '일반동사(pvg)'의 위치에는 항상 '서술성명사+동사파생접미사 쌍(ncpa+xsv)'도 같이 쓰일 수 있다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
가감하다/1/타동사/0	없음	더하거나 덜다.	더하/pvg+거나/ecc 덜/pvg+다/ef+./sf	더하다, 덜다
가격하다/0/타동사/0	가격1	치거나 때림	치/pvg+거나/ecc 때리/pvg+ㅁ/etn+./sf	치다, 때리다

[표 17] 둘 이상의 중심어

미(ecc)가 나오며, 이 대등적 연결어미 앞에 또 다른 서술어가 나오는 형태를 띤다.

4.3.2 필수격 정보. 형태소 분석된 사전정의문을 역방향으로 분석하는 과정에서 첫 번째 분석 대상이 되는 서술어 중 몇몇은 사전정의문의 의미를 전달하기 위해 반드시 필요로 하는 논항이다. 주어, 목적어, 보어 등이 그것인데, 이와 같은 필수격 논항은 사전정의문의 의미전달을 위해 중심어에 포함한다. 다음의 '규칙 나-1'에서부터 '규칙 나-5'까지는 이러한 논항 정보로 인해 파생되는 여러 가지 규칙과 준거이다.

규칙 나-1

주어와 목적어, 보어는 동사의 필수격 논항이다 (서충원, 2001). 만약 중심어가 되는 서술어 앞에 '주격조사(jcs)'나 '목적격조사(jco)', '보격조사(jcc)'의 태그를 동반하고 '주어'나 '목적어', '보어'가 함께 나온다면 그 '주어'나 '목적어', '보어'는 중심어에 포함한다. 제약정보로는 첫 번째 분석 대상인 서술어 앞에 '주격조사(jcs)'나 '목적격조사(jco)', '보격조사(jcc)' 태그가 나온다. 다음 [표 18]의 '매개하다/0/타동사/0'가 그 예이며, 여기서 '목적어+서술어'의 두 어절 형태인 '관계를 맺다'가 중심어로 추출된다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
매개하다/0/ 타동사/0	매개/2/명사/1	사이에 들어 양편의 관계를 맺어 줌.	사이/ncn+에/jca 들/pvg+어/ecs 양편/ncn+의/jcm 관계/ncn+를/jco 맺/pvg+어/ecx 주/pvg+ㅁ/etn+./sf	관계를 맺다

[표 18] 중심어 추출에 있어서의 필수격 정보

추출된 필수격 논항은 중심어 중 서술어의 애매성(ambiguity)을 제거하는 데도 도움을 준다.

[표 19]에서 '전사하다/6/타동사/0'과 '옷갓하다/0/자동사/0'은 동일하게 '쓰다'를 중심어의 서술어로 추출하였다. 그러나 어휘 '쓰다'가 갖는 중의적인 의미로 인해 '붓, 연필 따위로 글씨를 적다.'의 의미인지 '모자 따위를 머리 위에 얹어 덮다.'의 의미인지 판단하기가 힘들다. '옷갓하다/0/자동사/0'은 '갓을'이라는 목적격 논항 정보로 인해 '쓰다'의 의미가 후자임을 알 수 있다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
전사하다/6/타동사/0	전사4	서로 돌려 가며 베끼어 씀.	서/ncn+로/jca 돌리/pvg+어/ecx 가/px+며/ecc 베끼/pvg+어/ecx 쓰/pvg+ㅁ/etn+./sf	쓰다
옷갓하다/0/자동사/0	없음	웃옷을 입고 갓을 쓰다.	웃옷/ncn+을/jco 입/pvg+고/ecc 갓/ncn+을/jco 쓰/pvg+다/ef+./sf	웃옷을 입다, 갓을 쓰다

[표 19] 중심어 '쓰다'에 나타난 의미의 애매성

규칙 나-2

필수격을 중심어로 추출할 때 필수격 논항을 수식하는 다른 형태소가 있다면, 그 수식어구는 중심어에 포함시키지 않는다. 즉, 다음 [표 20]의 '경하하다/2/타동사/0'의 예에서 '경사로/paa+운/etm 일/ncn'에서와 같은 예이다. 목적어인 '일/ncn'을 수식하기 위해 쓰인 관형사 '경사로/paa+운/etm'은 중심어에 넣지 않는다. 따라서 '경하하다/2/타동사/0'의 중심어는 '일을 치하하다'가 된다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
경하하다/2/타동사/0	경하4	경사로운 일을 치하함.	경사로/paa+운/etm 일/ncn+을/jco 치하/ncps+하/xsm+ㅁ/etn+./sf	일을 치하하다

[표 20] 수식어구를 지닌 중심어의 필수격 논항

규칙 나-3

필수격의 논항이 '또는' 등과 같은 접속부사(maj)나 '-과/와' 등의 공동격조사(jct)로 연결되어 논항을 이루는 경우가 있다. 이때의 '또는'은 4.2절의 접속부사와는 달리 형태소 분석 결과 '또는' 앞에 마침표('.')가 쓰이지 않았다. 아래 [표 21]의 '접하다/1/자동사/3'에서처럼 주어나 목적어 자리에 나오는 명사가 '또는'으로 연결되어 2개 이상이 나온다면 다수의 명사를 모두 중심어로 추출한다. 따라서 [표 21]의 '접하다/1/자동사/3'의 중심어는 '직선, 곡선이 만나다'가 된다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
접하다/1/자동사/3	없음	직선 또는 곡선이 다른 곡선과 한 점에서 만나다.	직선/ncn 또는/maj 곡선/ncn+이/jcs 다르/paa+ㄴ/etm 곡선/ncn+과/jcj 한/nnc 점/nbu+에서/jca 만나/pvg+다/ef+./sf	직선, 곡선이 만나다

[표 21] 공동격조사로 연결된 중심어의 필수격 논항

규칙 나-4

필수격 논항은 간혹 '~따위' 등의 의존명사가 함께 나오기도 한다. '~따위' 등은 의미전달에는 필요하지 않은 요소이므로, 이때는 [표 22]에서의 예와 같이 '~ 따위'를 제외한 명사를 열거 형식으로 중심어에 포함하여 추출한다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
거역하다/0/ 타동사/0	거역2	윗사람의 뜻이나 명령 따위를 항거하여 거스름.	윗사람/ncn+의/jcm 뜻/ncn+이나/jcj 명령/ncn 따위/nbn+를/jco 항거/ncpa+하/xsv+어/ecs 거스르/pvg+ㅁ/etn+./sf	뜻, 명령을 거스르다

[표 22] '~따위'가 쓰인 사전정의문의 예

'~따위'와 같은 의미로 쓰인 [표 23]의 '~같은'과 '~ 등'이 있다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
개원하다/3/ 자타동사/0	개원2	병원, 양로원, 학원 같은 것을 처음으로 엶.	병원/ncn+,/sp 양로원/ncn+,/sp 학원/ncn 같/paa+은/etm 것/nbn+을/jco 처음/ncn+으로/jca 열/pvg+ㅁ/etn+./sf	병원, 양로원, 학원을 열다
논파하다/1/ 타동사/0	논파1	학설, 이론 등을 논하여 깨뜨림.	학설/ncn+,/sp 이론/ncn 등/nbn+을/jco 논하/pvg+어/ecs 깨뜨리/pvg+ㅁ/etn+./sf	학설, 이론을 깨뜨리다

[표 23] '~같은'이 쓰인 사전정의문의 예

준거(criterion)[15] 나-1

사전정의문에서 지시대명사(npd)가 나오는 경우가 있다. 사전정의문 내에서 이미 한번 나온 명사를 지시할 때는 그 지시하는 명사를 찾아서 완전한 중심어를 만들어준다. 필수 논항 중에 대명사를 형태소 분석 결과로 구분하여 별도로 수작업하였다. [표 24]의 '그것을'과 같은 예이다. [표 24]에서 중심어는 '그것을 아끼다, 사랑하다'가 된다. '그것'은 지시대명사이며 사전정의문내에 '물건'을 가리키므로 최종 사전정의문을 '물건을 아끼다, 사랑하다'로 완성해 준다.

준거 나-2

필수격 논항이 쓰였더라도 그 논항의 주격조사, 목적격조사, 보격 조사 등의 자리에 통용보조사(jxc)가 대신 쓰인 경우가 있다. 이때는 통용보조사가 쓰인 경우만을 모아

[15] 자동으로 추출되는 '규칙'에 비해, 수작업을 필요로 하는 경우에, 규칙이 아닌 일련의 정련에 사용되는 기준을 말하는 것을 '준거'라 명명하기로 한다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
애착하다/0/타동사/0	애착	어떤 사물과 떨어질 수 없게 그것을 사랑하고 아낌.	어떤/mmd 사물/ncn+과/jct 떨어지/pvg+ㄹ/etm 수/nbn 없/paa+게/ecs 그것/npd+을/jco 사랑/ncpa+하/xsv+고/ecs 아끼/pvg+ㅁ/etn+./sf	사물을 아끼다, 사랑하다

[표 24] 지시대명사가 쓰인 사전정의문

서 수작업으로 검수하여 적합한 격조사로 바꾸어 주고 최종 중심어를 완성하였다. [표 25]에서 '고갱이'는 목적어로 쓰였으며 조사 '만'은 '목적격조사' 자리에 대신 쓰인 통용 보조사이다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
수집하다/4/타동사/0	수집4	고갱이만 뽑아 모음.	고갱/ncn+이/jcs+만/jxc 뽑/pvg+어/ecs 모으/pvg+ㅁ/etn+./sf	고갱이를 모으다

[표 25] 통용보조사가 쓰인 사전정의문

준거 나-3

중심어로 추출될 수 있는 서술어의 필수격 논항은 간혹 조사가 생략되어 나타나기도 한다. [표 26]의 '이름하다/0/타동사/0'의 사전정의문은 필수격 논항인 목적어 '이름'의 조사가 누락되어 기술되어 있다. 이 경우에는 목적격 조사를 명시하여 중심어를 완성한다. 따라서 [표 26]에서 '이름하다/0/타동사/0'의 중심어는 '이름을 짓다'가 된다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
이름하다/0/타동사/0	없음	이름 짓다.	이름/ncn 짓/pvg+다/ef+./sf	이름을 짓다

[표 26] 목적격 조사가 생략된 사전정의문

규칙 나-5

사전정의문이 복문의 형식으로 이루어진 경우는 두 개의 문장에서 각각 중심어를 추출하여 다수의 중심어를 취한다. [표 27]의 '보국하다/2/자동사/0'은 그 일례로, 사전정의문은 '나라의 은혜를 갚다'와 '나라를 위하여 충성을 다하다'라는 두 개의 문장으로 이루어져 있다. 전자의 문장에서 '은혜를 갚다'라는 중심어가 추출되며, 후자의 문장에서는 '충성을 다하다'라는 중심어가 추출된다. 따라서 '보국하다/2/자동사/0'은 '은혜를 갚다, 충성을 다하다'라는 2개 이상의 중심어를 갖는 어휘이다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
보국하다/2/자동사/0	보국2	나라의 은혜를 갚거나, 나라를 위하여 충성을 다함.	나라/ncn+의/jcm 은혜/ncn+를/jco 갚/pvg+거나/ecc+,/sp 나라/ncn+를/jco 위하/pvg+여/ecs 충성/ncn+을/jco 다하/pvg+ㅁ/etn+./sf	은혜를 갚다, 충성을 다하다

[표 27] 복문으로 이루어진 사전정의문

4.3.3 연결어미 구문.

규칙 다-1

정의문의 서술어 형태가 복합형을 보이기도 한다. 첫 번째 형태는 종속적 연결어미(ecs)나 보조적 연결어미(ecx)가 쓰인 경우이다. 보조적 연결어미가 쓰였을 때는 '본용언+보조적 연결어미+보조용언'의 쌍이므로, 전자에 나오는 '본용언'만을 중심어로 선택한다. 다음의 [표 28]과 같은 경우로, '거들다'라는 본용언이 중심어로 추출된다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
가공하다/1/자동사/0	가공2	옛날의 형률에서, 범죄 행위를 거들어 줌.	옛날/ncn+의/jcm 형률/ncn+에서/jca+,/sp 범죄/ncn 행위/ncn+를/jco 거들/pvg+어/ecx 주/px+ㅁ/etn+./sf	범죄 행위를 거들다

[표 28] 보조적연결어미가 쓰인 사전정의문

그러나, 서술어의 또 다른 복합형인 '서술어+종속적연결어미+서술어'의 형태에서는 후자의 서술어를 중심어로 택한다. 종속적 연결어미의 의미적 특성상 후자의 서술어가 문장전체의 전달하고자 하는 의미를 지닌다. 다음 [표 29]의 '감사하다/5/타동사/0'의 예에서 보듯이, 사전정의문 전체의 주요 의미는 '바로잡다'이다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
감사하다/5/타동사/0	감사7	감별하여 조사함.	감별/ncpa+하/xsv+어/ecs 조사/ncps+하/xsm+ㅁ/etn+./sf	조사하다

[표 29] 종속적연결어미가 쓰인 사전정의문

4.3.4 기타.

준거 라-1

사전정의문의 어순이 바르지 못하다면, 주어, 목적어, 서술어 등이 순차적으로 나오는 기본형의 어순으로 만든 후에 중심어를 추출한다.

[표 30]의 '중창하다/2/자타동사/0'은 어순이 주어, 목적어, 서술어 순이 아니라

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
중창하다/2/ 자타동사/0	중창4	둘 이상의 성부를 한 사람이 한 성부씩 동시에 노래 부르는 일.	둘/nnc 이상/ncn+의/jcm 성부/ncn+를/jco 한/nnc 사람/ncn+이/jcs 한/nnc 성부씩/ncn 동시/ncn+에/jca 노래/ncn 부르/pvg+는/etm 일/ncn+./sf	사람이 성부를 노래 부르다

[표 30] 어순이 바르지 못한 사전정의문

목적어가 문두에 나오고 주어와 서술어가 그 뒤에 나오는 역순으로 되어 있다. 이 어순으로는 같은 중심어를 지니는 다른 목록과 일관된 중심어를 취할 수가 없다. 따라서 일반적으로 사용되는 주어, 목적어, 서술어의 어순을 기본형으로 삼고 이 어순에 맞게 중심어를 추출한다. 따라서 '중창하다/2/자타동사/0'의 중심어는 '성부를 사람이 노래 부르다'가 아니라 '사람이 성부를 노래 부르다'가 된다.

준거 라-2

다음은 부사어구 포함된 관용적 표현으로 반드시 부사어가 중심어에 포함되어야 사전정의문이 나타내고자 하는 의미를 제대로 전달할 수 있는 경우이다.

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
기소하다/2/ 타동사/0	기소2	속이어 우습게 봄.	속이/pvg+어/ecs 우습/paa+게/ecs 보/pvg+ㅁ/etn+./sf	우습게 보다

[표 31] 관용적 표현으로 이루어진 중심어

[표 31]의 '기소하다/2/타동사/0'에서 중심어는 서술어 '보다'가 아닌 '우습게 보다'이다. '우습게 보다'는 관용적인 표현으로 어구 고유의 뜻을 지니게 된다. <우리말 큰사전>에서 살펴보면 '보다/1/타동사/1'은 '눈으로 느끼다'라는 뜻과 함께, '보다/1/타동사/3'에서 '어떻게 여기거나 평가하다.'라는 의미가 쓰인다. [표 31]의 사전정의문에서는 관용적 표현으로 쓰인 후자의 뜻이다. 중심어로 관용적 표현이 추출될 때는 그 표현 전체를 중심어로 본다.

준거 라-3

준거 라-2의 관용적 표현 외에도 의미전달을 위해 반드시 필요한 논항일 경우에는 4.3의 필수격 정보가 아니더라도 중심어에 포함시켰다. 다음의 [표 32]와 같은 예이다.

[표 32]에서 추출된 중심어의 서술어 '가다'는 '이 곳에서 다른 곳으로 움직이다.'의 일차적인 의미가 아니다. '근무를 위해 어떤 곳으로 임명되어 감'의 의미로 '사신으로'라는 명사와 연결될 때 비로소 그 의미가 명확해 진다. 이렇게 중심어의 의미전달을 위

표제어	사전정의문1	사전정의문2	형태소 분석 결과	중심어
출강하다/1/자동사/0	출강1	왕명을 받아 외국에 사신으로 감.	왕명/ncn+을/jco 받/pvg+아/ecs 외국/ncn+에/jca 사신/ncn+으로/jca 가/pvg+ㅁ/etn+./sf	사신으로 가다

[표 32] 필수격 논항외의 중심어의 논항

해 반드시 필요하다고 판단되는 논항에 대해서는 중심어에 포함하도록 한다. 준거 라-2와 준거 라-3과 같은 작업은 수동으로 이루어졌다.

5. 결론 및 향후 연구

본 논문에서는 '[-하]동사류'의 사전정의문 3,656개에 대해 여러 가지 규칙과 준거에 의한 중심어 추출 방식에 대해 소개했다. 추출된 중심어는 향후 연구를 통해 동사 어휘의미망을 본격적으로 구축하는 데 활용될 것이다. 구축되는 동사 어휘의미망은 두 가지 면에서 의의를 가진다. 첫째, 본 논문을 통해 구축되는 어휘의미망은 중심어를 통한 반자동 구축방식이므로 기존의 작업자의 주관적인 판단에 의존하는 방식과 달리 효율적인 구축 및 확장이 가능하다. 즉, 사전정의문을 통해 중심어를 선정하고 선정된 중심어를 이용한 어휘의미망 구축 방식으로써, 실험 대상이 되는 동사어휘에 국한되지 않고 비교적 짧은 시간과 인력을 활용하여 효율적으로 전체 한국어 동사어휘의미망으로의 확장이 가능하다. 둘째, 기존의 '워드넷'이나 일본의 '어휘대계'에서 구축하였던 동사의 통사적인 측면에 초점을 맞춘 어휘의미망이 아니라, 동사의 의미적인 분류 기준을 정하려 노력하였다. 따라서 동사에만 한정되지 않고 기존의 명사 어휘의미망에 통합되며 나아가 형용사나 부사와 같은 다른 'open class word'를 포함하는 범품사적인 어휘의미망으로의 구축이 가능할 것이다.

앞으로 남은 과제는 구축된 중심어를 활용하여 본격적인 동사의 어휘의미망 구축과 구축된 동사어휘의미망을 다양한 자연언어처리 시스템에 직접 활용하여 시스템의 성능 향상을 보임으로써[16] 동사어휘의미망 구축의 정당성을 수립하는 데 있다.

<참고문헌>

Choi, Key-Sun and Hee-Sook Bae. 2004. Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy. In *Proceedings of the Second International WordNet Conference*, pp. 91–96.

Dong, Zhendong and Quiang Dong. 2006. *HowNet and the Computation of Meaning.* World Scientific Publishing.

Fellbaum, Christiane. 1998. *Wordnet: An Electronic Lexical Database.* MIT Press.

[16] 본 연구를 통해 구축된 동사 어휘의미망은 향후 단어클러스터링 시스템에 활용되었으며, 활용된 단어클러스터링 시스템의 향상을 통한 동사 어휘의미망의 평가 및 실효성 의 검증을 연구 중에 있다.

Gross, Maurice. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages* 63, 7–52.

Ikehara, Satoru et al. 1997. *The Semantic System, volume 1 of Goi-Taikei — A Japanese Lexicon.* Iwanami Shoten.

Vossen, Piek. 2005. EuroWordNet General Document. Technical report, University of Amsterdam.

김혜경. 1996. 불어의 술어기능명사와 한국어 변환. 석사학위 논문, 부산대학교 불어불문학과.

문유진. 1996. *의미론적 어휘개념에 기반한 한국어 명사 WordNet의 설계와 구축.* 박사학위 논문, 서울대학교 컴퓨터공학과.

박석문. 2000. *코퍼스 품사 태깅 매뉴얼.* 한국과학기술원.

서정수. 1975. *동사 "하-"의 문법.* 형설출판사.

서충원. 2001. 용언의 필수격 정보를 이용한 한국어 단문의 의존 구조 분석. 석사학위 논문, 한국과학기술원 전자전산학과.

옥철영. 2005. 한국어 Wordnet 구축: 명사를 중심으로. *한국언어정보학회 2005 정기 학술대회* 발표 논문집에서, 1–15쪽.

이은령·윤애선. 2005. 피동 정보를 통한 한국어 동사 어휘의미망 정제. *한국어학* 28, 139–165.

이은령·황순희·윤애선. 2004. 다국어 어휘의미망 구축의 현황과 문제점. *프랑스문화예술연구* 6.2, 369–401.

이주호·은광희·최기선. 2001. 기계가독사전을 이용한 한국어 시소러스 구축. *제13회 한글 및 한국어 정보처리 학술대회*에서, 273–278쪽.

최호섭·옥철영. 2002. 한국어 의미망 구축과 활용. *한국어학* 17, 301–329.

한국과학기술원 전문용어언어공학연구센터. 2000. 대용량 국어정보 심층처리 및 품질관리 기술개발. 기술보고서, 한국과학기술원.

한국과학기술원 전문용어언어공학연구센터. 2005. *다국어 어휘의미망.* KAIST PRESS.

<참고 웹사이트>

국립국어원	http://www.korean.go.kr (2005년3월~2005년5월)
연세한국어사전	http://dic.yonsei.ac.kr/ (2004년9월~2006년5월)
우리말큰사전	http://nlpweb.kaist.ac.kr/Urimal/ (2004년9월~2006년5월)
카이스트 형태소분석기	http://bola.or.kr (2005년3월~2006년2월)
표준국어대사전	http://www.korean.go.kr/000_new/50_dic_search.htm (2004년9월~2006년5월)

접수 일자: 2006년 4월 20일
게재 결정: 2006년 7월 27일