한국어 어휘의미망 「KorLex 1.5」의 구축

(Construction of Korean Wordnet 「KorLex 1.5」)

윤 애 선 † 황 순 희 †† 이 은 령 ††† 권 혁 철 ††††

(Aesun Yoon) (Soonhee Hwang) (Eunryoung Lee) (Hyuk-Chul Kwon)

요 약 1980년대 중반부터 지난 20여 년간 구축해 온 영어 워드넷(PWN)은 인간의 심상어휘집을 재현하려는 목적으로 개발되기 시작하였으나, 그 활용 가능성에 주목한 것은 자연언어처리와 지식공학 분야다. 컴퓨터 매개 의사소통(CMC), 인간-컴퓨터 상호작용(HCI)에서 인간 언어를 자연스럽게 사용하여 필요한 정보를 획득하기 위해서는 의미와 지식의 처리가 필수적인데, 그 해결의 실마리를 어휘라는 실체를 가진 언어단위에서 찾을 수 있기 때문이다. 이후 전 세계적으로 약 50개 언어의 어휘의미망이 PWN을 참조 모델로 구축되어 다국어처리의 기반을 제공할 뿐 아니라, 시맨틱 웹 이후 더욱 주목받고 다양한 방식으로 활용되고 있다. 이 논문은 PWN을 참조 모델로 2004년부터 2007년까지 구축한 한국어 어휘의미망 KorLex 1.5를 소개하는 데 있다. 현재 KorLex는 명사, 동사, 형용사, 부사 및 분류사로 구성되며, 약 13만 개의 신셋과 약 15만 개의 어의를 포함하고 있다.

키워드 : 워드넷, 어휘의미망, 코렉스, 한국어정보처리, 다국어처리, 지식공학, 온톨로지

Abstract The Princeton WordNet(PWN), which was developed during last 20 years since the mid 80, aimed at representing a mental lexicon inside the human mind. Its potentiality, applicability and portability were more appreciated in the fields of NLP and KE than in cognitive psychology. The semantic and knowledge processing is indispensable in order to obtain useful information using human languages, in the CMC and HCI environment. The PWN is able to provide such NLP-based systems with 'concrete' semantic units and their network. Referenced to the PWN, about 50 wordnets of different languages were developed during last 10 years and they enable a variety of multilingual processing applications. This paper aims at describing PWN-referenced Korean Wordnet, KorLex 1.5, which was developed from 2004 to 2007, and which contains currently about 130,000 synsets and 150,000 word senses for nouns, verbs, adjectives, adverbs, and classifiers.

Key words : WordNet, Lexical Semantic Network, Korean Language Processing, Multilingual Processing, Knowledge Engineering, Ontology

· 이 논문의 작성은 2007년 정부(교육과학기술부)의 지원(과학재단 R01-2007-000-20517-0)의 지원을 받음

† 비 회 원 : 부산대학교 불어불문학과/인지과학협동과정 교수
asyoon@pusan.ac.kr
†† 비 회 원 : 부산대학교 인문학연구소 연구교수
soonheehwang@pusan.ac.kr
††† 비 회 원 : 부산대학교 인문학연구소 HK연구교수
eunryounglee@pusan.ac.kr
†††† 종신회원 : 부산대학교 정보컴퓨터공학부 교수
hckwon@pusan.ac.kr

논문접수 : 2008년 8월 18일
심사완료 : 2008년 11월 7일

정보과학회논문지: 소프트웨어 및 응용 제36권 제1호(2009.1)

1. 서 론

인간은 주변 환경을 어떻게 인지하고, 그것을 지식화하며, 그 지식을 다른 사람과 공유하는가? 이미 획득된 지식을 바탕으로 새로운 지식을 어떻게 추론하고, 새로운 상황에 그 지식을 적용하는가? 언어는 그 지식을 어떤 방식으로 추상화하는 데 이바지하는가? 인간이라는 종의 개체생존에 직결되는 근원적인 문제다. 고대 철학에서부터 현대 인지과학(cognitive science)에 이르기까지 이에 대해 단편적이지만 다양한 답을 제시하고 있다.

그 시도 중 하나가 심리학 분야에서 인지과학의 초석을 제공한 밀러(G. Miller)의 워드넷(WordNet, 이하 PWN)이다[1,2]. 지식을 구성하는 기본단위가 개념(concept)이며, 단어 또는 어휘를 통해 그 개념을 언어화할 수 있고, '어휘가 심리학적 실재를 가진 기억의 최소 단

위'라는 자신의 이론을 바탕으로, 지식의 기본 단위 간 다양한 의미관계를 계층적 망(hierarchical network) 형태로 설정한 것이다. 1985년부터 영어를 대상으로 본격적으로 시작된 연구와 구축은 20여 년이 지난 지금까지도 계속되고 있으며, 자연언어처리(Natural Language Processing)의 의미연구에 초석을 마련하였다. 2008년 현재 전 세계적으로 약 50개 언어의 어휘의미망(lexical semantic network)이 PWN을 모델로 삼아 구축되어 다국어 처리(multilingual processing)의 기반을 제공하고 있다[3]. 또한, 개별 언어의 특성을 넘어선 보편적 개념망과의 사상(mapping)이 매우 활발하게 이루어져 왔고, 구글의 의미검색 시스템에 적용되기도 하며, 시맨틱 웹(Semantic Web)이 등장하면서 더욱 주목받기 시작했다. 폭발적으로 늘어나고 다양화되는 컴퓨터 매개 의사소통(CMC), 인간-컴퓨터 상호작용(HCI)에서 인간 언어를 자연스럽게 사용하여 필요한 정보를 획득하려면 의미와 지식의 처리가 필수적인데, 그 해결의 실마리를 '어휘(word)'라는 실체를 가진 언어 단위에서 찾으려고 한다.

한국어를 대상으로 어휘의미망을 구축하기 시작한 것은 90년대 중반부터다[4]. 이중 PWN을 참조한 것은 '한국어 시소러스'와 'KorLex(Korean Lexico-semantic Network)'다([표 1] 참조). 1997년-2000년에 개발된 전자는 PWN 중 일부 명사를 대역(translation)한 시제품적 특성을 띠었다[5]. 2004년부터 개발되기 시작한 후자는 2004년 10월 PWN의 명사를 대역한 KorLexNoun 1.0을 공개한 데 이어[6], PWN의 구축 범위를 포괄하는 동시에 한국어에 특히 발달한 내용어(content words) 범주인 분류사(classifier)를 추가하고, 구축 방법론에서도 대역 단계를 넘어 한국어의 의미 특성을 잘 반영할 수 있는 어휘의미망을 구축하고자 하며, 앞으로 이 노력은 지속적으로 확장될 예정이다.

이 논문의 목적은 2007년 11월에 발표된 KorLex 1.5를 소개하는 데 있다. 2장에서는 KorLex의 모델이 된 PWN의 개발 배경, 정보 구조, 활용 현황을 소개하고, 3장에서는 KorLex 1.5의 구축 방법론 및 정보 구조를 설명하며, 4장에서는 향후 연구 및 개발 방향을 제시한다.

2. PWN의 개발 배경과 현황

어휘가 표상하는 의미 간의 관계를 표상하려는 PWN의 구축 대상은 영어 내용어였고 그중에서도 명사와 동사에 주된 초점을 맞추었다. 첫 결실인 1.0은 1991년도에 발표되었으며, 1995년의 1.5은 EWN의 참조모델이 되면서 다국어처리 가능성을 열어 놓게 된다. 거의 같은 시기에 발표되는 어휘의미망, 시소러스, 개념망 등과의 사상(mapping)이 활발하게 일어나고, 2003년에 2.0이 발표된다. 이후 소규모의 수정과 보완 작업이 반영된 2.1(2005년), 2.1의 Unix용인 3.0(2006년)이 발표되면서, 다양한 분야에서 그 활용 가치를 인정받고 있다. PWN의 버전 중 다른 어휘망이나 개념망에 영향을 많이 끼친 것은 1.5과 2.0이다. 2004년도에 구축하기 시작한 KorLex는 기본적으로 PWN 2.0을 모델로 삼고, PWN에서 발표한 2.0·2.1·3.0의 신셋 간 사상표을 함께 제공한다. 이 논문에서는 KorLex 1.0와 1.5의 참조모델인 PWN 2.0 및 2.1의 특성을 중심으로 기술하겠다.

2.1 신셋: 개념의 표상 단위

PWN에서 개념을 표상하는 최소 단위를 '동일한 어휘의미(word meaning, 이하 어의)를 가지는 동의어집합(synonym set, 이하 신셋)'으로 규정하면서, '개념=어휘의 세분화된 의미'라는 등식이 성립하게 된다. 예를 들어, 표 2와 같이 다의어 'report'의 여러 어의가 각각 'paper, story, study 등'의 특정한 어의와 동의관계를 이룬다면, 동일한 어의를 {report6, paper2} 등으로 묶어 표현함으로써 중의성이 없이 하나의 개념을 표상한다[1]. 이때 다의어의 어의 구분은 어휘형태(이하, 어형) 뒤에 아라비아 숫자로 표시한다. PWN은 '어형:어의'의 多:多 관계를 최대한 세분화하여 표시할 수 있으며, 개념을 명명하는 데 자연언어와 구분되는 메타언어를 따로 설정해야 하는 부담이 없다는 장점이 있다[1]. PWN에서 어의는 ① 신셋이 표현하는 개념과 ② 신셋 집합을 구성하는 원소를 모두 의미할 수 있다. 이하 글에서는 용어의 혼동을 피하기 위해 전자는 '신셋'으로, 후자는 '어의'로 구분하여 사용하겠다. 따라서 어의는 특정 어형과 밀접한 관련을 맺는다. 예를 들면, {report1,

표 1 대표적인 국내 어휘의미망(발췌)

명칭	중심구축기관	중심구축자 전공	구축 기간	구축방식/참조모델	의미/개념(n) vs 어의(w) 수	구축 품사
한국어 명사워드넷[4]	호남대학교	전산학	1994-1995	직접	20,000w	명
세종 전자사전[7,8]	서울대학교	언어학	1998-2007	직접	581n vs. 540,000w	모든 품사
U-Win[9,10]	울산대학교	전산학	2002-2007	직접	46,339n vs. 약250,000w	모든 품사
한국어 시소러스[5]	포항공과대학	전산학	1997-2000	참조/PWN	18,362n vs. 21,390w	명
KorLex 1.5[11,12]	부산대학교	전산학/언어학	2004-현재	참조/PWN	130,639n vs. 147,906w	명,동,형,부,분류사
다국어 어휘 데이타베이스[13]	고려대학교	언어학	2000-2006	참조/EWN	5,500w	명
CoreNet[14]	KAIST	전산학/언어학	1995-2004	참조/NTT어휘대계	2,938n vs. 62,632w	명,동,형

표 2 어의와 어형의 대응관계

어형 / 신셋{어의}	report	paper	story	study	...
{report6, paper2}	○	○			
{report2, story5}	○		○		
{report1, study3}	○			○	
{report, ...}	○				○

study3}은 “a written document describing the findings of some individual or group”이라는 개념을 나타내는 신셋이고, ‘report1’과 ‘study3’는 이 개념을 각각 ‘report, study’라는 어형으로 실현하는 어의이다.

PWN 1.5까지 품사별 신셋, 어형, 어의의 구체적인 통계자료는 남아있지 않다. 다만 워드넷의 검색시스템을 개발한 텐지(R.I. Tengi)에 의하면 1992년에 발표된 PWN 1.2판에는 약 52,000개 신셋과 102,000개 어의가 개발되고, 1995년의 PWN 1.5판에는 약 91,600개 신셋과 168,000개 어의를 포함한다고 한다[1]. 표 3은 통계자료가 남아있는 PWN 2.0판, 2.1판, 3.0판의 자료 크기를 보여준다[2].

2.2 의미표상구조

PWN의 큰 장점 중의 하나로 신셋 간, 어의 간의 의미관계를 표 4처럼 매우 다양하고 풍요롭게 표현한 점을 들 수 있다. (표 4에서 편의상 다의어 구분 번호는 표시하지 않는다.)

동의(synonymy)는 PWN의 가장 기본적인 관계로 한 신셋을 구성하는 2개 이상의 어의를 맺는 전제조건이며, 모든 품사에 적용된다. 명사와 동사의 경우, 각 신셋은 하의(hyponymy)는 상의(hypernymy)와 함께 쌍을 이루어 IS-A 방식의 계층관계로 나타낸다. 상위어는 총체적이고 보편적 의미자질을 하위어에 물려주고, 하위어는 이를 승계(inherit)하고 직접 상위어를 구별해 줄 자질을 적어도 하나 이상 추가하여 가지는 방식이다. PWN 2.0에서는 11개의 최상위 개념(unique beginners)에서 출발하여 실제 최대 17개 층위, 동사는 최대 12개 층위로 구성된다. (PWN을 소개하는 문서에서는 명사와 동사의 최대 층위를 각각 12개, 4개라고 하나 실제 자료와는 다르다.) PWN 2.1 명사는, 최상위 개념을 {entity}라는 1개로 묶는 시도를 하였다.

또한, 명사와 동사의 신셋은 각각 25개와 15개 의미분류(semantic domain)로 구분한다. 명사가 1개의 최상위 개념에서 시작하는 대신, 동사는 다수의 상층 개념에서 시작하여 넓고 얕게 분포되어 있다. 또한 1개의 의미분류가 여러 개의 최상위 개념을 갖는 것이 일반적이다. 예를 들어, ‘possession’에 분류되는 동사의 어의는 {transfer5}, {get1, acquire1}, {have1, hold6}처럼 3개의 최상위 신셋과 연결된다. PWN에서는 이런 의미분류를 ‘사전편찬자 파일(lexicographers′ file)’이라고 부른다. 이러한 명칭과 처리방식은 PWN이 개발되던 시기인 80년대 후반-90년대 초반의 낮은 컴퓨터 저장 및 처리 능력과도 관계가 있다.

반의(antonymy)는 명사와 동사에도 표현되기도 하나, 형용사와 부사에서는 방사형 핵 구조를 형성한다. 즉 반의 관계를 갖는 2개 또는 3개 어의가 핵(head)을 이루고 각각의 핵은 유의(similar) 관계 신셋과 방사형 구조를 갖는다. 이는 심리학의 단어 연상 실험에서 형용사가 제시되었을 때 많은 사람들이 반의어를 떠올리는 것을 관찰한 심리학 연구 결과를 반영한 것이다. PWN에서 반의는 신셋이 아닌 어의 간 관계로 정의한다.

전의(holonymy)와 분의(meronymy)는 짝을 이루어 명사의 전체-부분 관계를 표현한다. 분의는 구체적으로 부분(component), 집합의 구성소(member), 물질(substance) 등 3개의 종류로 구분한다. 전체-부분 관계를 명사에서 전의/분의로 표현한다면 동사에서는 함의(entailment)로 나타낸다[1]. 함의는 내포(proper inclusion), 전제(presupposition), 양식(troponymy), 인과(cause) 등을 단방향적 관계로 표현한다. 즉, 한 행위(V1=snore, amble, divorce, kill)가 다른 행위(V2=sleep, walk, marry, die)를 내포, 실현, 전제하고 결과로 삼지만(V1->V2), 역은 참이 아니며(V2/->V1), 동시에 V2가 성립하지 않으면 V1도 성립하지 않는다(¬V2 -> ¬V1). 내포와 양식 관계에서 V1과 V2는 동시성을 갖고, 전제 관계에서는 V2가 V1에 선행하고, 인과 관계에서는 V1이 V2에 선행한다[1]. 이러한 의미 관계는 계층 구조나 핵 구조와는 별도로 기술된다.

속성(attribute)은 ‘length - long, short’처럼 명사-형용사 간 속성자질과 그 자질 값(value)과의 관계를 연결한다. 영역(domain)은 모든 품사에서 해당 신셋의 전문분야(topic), 지역(region), 어법(usage) 정보를 표현한

표 3 PWN 버전별 구축 크기

버전	발표연도	명			동			형			부			계		
		어형	신셋	어의	어형	신셋	어의	어형	신셋	어의	어형	신셋	어의	어형	신셋	어의
2.0	2003	114,648	79,689	141,690	11,306	13,508	24,632	21,436	18,563	31,015	4,669	3,664	5,808	152,059	115,424	203,145
2.1	2005	117,097	81,426	145,104	11,488	13,650	24,890	22,141	18,877	31,302	4,601	3,644	5,720	155,327	117,597	207,016
3.0	2006	117,798	82,115	146,312	11,529	13,767	25,047	21,479	18,156	30,002	4,481	3,621	5,580	155,287	117,659	206,941

표 4 신셋 및 어의 간 의미관계

의미관계	관련 품사	예		표시단위	
				신셋	어의
동의	명,동,형,부	{board, plank} {rise, ascend} {sad, unhappy} {rapidly, speedily}			○
하의/상의	명,동	plant -> tree -> maple -> sugar maple		○	
반의	명,동,형,부	wet <-> dry rapidly <-> slowly			○
유의	형	wet - watery, damp, moist, humid, soggy		○	
전의/분의	명	부분 구성소 물질	hat > brim fleet > ship milk > protein	○	
함의	동	내포 양식 전제 인과	snore - sleep amble - walk divorce - marry kill - die	○	
속성	명-형	length - long, short		○	
영역	형,명,부,동	전문분야	chaotic-physics pas-ballet largo-music scroll-computer science	○	
		지역정보	blae-Scotland karate-Japan jolly-Britain scrimshank-Britain	○	
		어법	commodious-archaicism bloomers-plural bang-colloquialism dandle-blend	○	
참조	형	true - correct, faithful, honest, sincere		○	
	동	pay - pay off			○
동일 어근 부속	명-형	icon - iconic (hearing - auditive)			○
동일 어근 파생	형-부	usual - usually unusual - unusually			○
동일 어근 관련	동-명	press-pressure point-point			○
동일 어근 분사	동-형	break-breaking break-broken			○
동사군	동	{come to, resuscitate, revive}-{resuscitate, revive}		○	

다. 참조(also see)는 형용사에서는 핵을 이루는 신셋 간의 의미 관계를 표현하나, 동사에서는 동사 어의 간 관계를 나타내며 동사의 특정한 의미와 이와 관련을 맺고 있는 '동사+전치사/부사'로 구성된 연어를 연결한다.

의미가 승계되는 품사 간 연계는 명->형, 형->부, 동->명, 동->형 네 경우에 표시되고, 이를 각각 부속(pertain), 파생(derive), 관련(relate), 분사(participle)로 칭한다. 전자는 명사에서 파생된 형용사나(icon -> iconic), 어근이 다르나 동일한 참조물을 지칭하는 명사와 형용사의 관계를 나타내나(hearing - auditive), 후자 3개의 경우는 어근이 동일한 파생관계만을 나타낸다. 분사는 동사와 이것의 현재분사나 과거분사에서 파생된 형용사와의 관계를 표현한다.

이 밖에도 동사군(verb group)은 통계적으로 유사한 의미를 나타내는 동사 신셋을 연결한다[2].

2.3 개별언어 의존적 정보

PWN에는 개념 관계 이외에도 영어에만 적용되는 범주·파생·통사 정보를 포함하는데, 다른 언어에서는 유효성을 갖지 못한다. 첫째, PWN의 기본 골격을 형성하는 개념을 명·동·형·부사 4개의 문법 범주로 구분한다. 이러한 범주화는 언어에 따라 달리 구현될 수 있는데, 예를 들어 한국어에서는 동사와 형용사 일부를 용언

으로 통합할 수도 있을 것이다. 둘째, 앞 절에서 살펴본 부속, 파생, 관련, 분사 관계 및 동사 참조 관계의 대부분 예는 어의 단위로 동일한 어근의 파생관계에 기초한다(예: determine-determination, number-numeral, usual-usually, break-breaking, give-give off). 셋째, 동사 신셋의 각 어휘의미에 35개의 매우 간략한 문장 격틀(sentence frame) 정보를 제공한다(예: Somebody ----s something to somebody; It ----s that CLAUSE). 넷째, 형용사의 사전편찬자 파일은 명사나 동사와는 달리 의미분류가 아닌 파생관계로 구분되어 있다. 형용사의 사전편찬자 파일은 2개(adj.all, adj.pert)로 나누는데, 후자는 명사에서 파생된 형용사를 모은 것이고 전자는 그 밖의 형용사 집합이다. 이를 보완 수정하기 위해, PWN에 기반한 독일어 어휘의미망에서는 형용사의 의미분류를 15개로 구분한다[15]. 다섯째, 예의 수가 매우 제한적이기는 하지만, 형용사와 피수식어 간 어순에 관한 정보(예: 술어 위치, 명사 앞, 명사 뒤)를 제공한다.

2.4 신셋의 정보구조

신셋의 정보는 표 5처럼 표현되며, 이를 구성하는 표지는 표 6과 같이, ID번호, 의미분류, 품사, 해당 신셋을 구성하는 어의의 수, 의미식별자가 달린 어의, 해당 신셋이 맺고 있는 의미관계의 수, 각 의미관계를 구체적 내용, 문형 정보의 수 및 각 문형정보 내용과 함께 수의적인 요소로 정의문과 예문을 나타낸다.

2.5 장점과 한계

위와 같은 PWN의 특성은 동시에 장점과 한계로 작용한다.

첫째, '개념=어휘의미'이라고 정의함으로써 PWN이 언어보편성을 갖기에는 개념의 크기가 지나치게 작고, 어

표 5 PWN 신셋 정보 표지

00935309 32 v 02 report_5 cover_2 008 @ 00831651 v 0000 + 06683784 n 0201 + 07217924 n 0101 + 06681551 n 0101 + 10521662 n 0101 + 06683784 n 0103 + 06683784 n 0102 $ 00967455 v 0000 03 + 08 00 + 09 00 + 22 01 | be responsible for reporting the details of, as in journalism; "Snow reported on China in the 1950's"; "The cub reporter covered New York City"

표 6 PWN 신셋 정보 표지의 내용

구분	표지	내용	
신셋 ID	00935309		
의미 분류	32	verb.communication	
신셋의 품사	v	동사	
신셋 구성 어의 수	02	2개	
신셋 구성 어의	report_5 cover_2	report	어형
		5	해당 품사에서 다의어를 구분하는 식별자
의미관계 수	008	8개의 의미관계 가짐	
의미관계	@ 00831651 v 0000	@	해당 신셋의 상위관계임
		00831651	상위 신셋ID
		v	상위 신셋의 품사
		0000	이 의미관계는 신셋 간 관계임
	+ 06683784 n 0201	+	해당 신셋과 파생관계임.
		06683784	파생관계 신셋 ID
		n	파생관계 신셋의 품사
		0201	파생관계는 해당 신셋 2번째 어의(cover_2)와 파생 신셋 1번째 어의(coverage_0) 간의 관계임.
문형정보 수	03	이 신셋에 속하는 어의는 총 3개의 문형정보를 가짐	
구분자	+	문형정보 시작	
문형정보	08 00	08	적용 문형 (Somebody ----s something)
		00	해당 신셋의 첫 번째 어의(00)에 적용됨
	22 01	22	적용 문형 (Somebody ----s PP)
		01	해당 신셋의 두 번째 어의(01)에 적용됨
구분자	|	정의문 시작	
정의문	be responsible for reporting the details of, as in journalism		
구분자	;	예문 시작	
예문	"Snow reported on China in the 1950's"		

휘와 개념 간의 구분이 명확하게 이루어지지 않는다는 비판을 받는다[16]. 하지만 개념을 메타언어로 새롭게 명명해야 하는 부담을 덜 수 있을 뿐 아니라, 영어로 기술된 텍스트에서 좀 더 직접적인 방식으로 의미와 지식을 추출할 수 있다.

둘째, PWN의 표제어 수가 약 15만 개이고 한 표제어당 다의어 수가 약 1.4개 정도 되는 중형사전에 해당한다. 중형사전은 해당 언어를 모국어로 사용하는 보통 화자가 일반적인 텍스트를 이해하는 데 필요한 언어정보를 담고 있다[17]. 아무리 정교한 언어/지식 정보라고 하더라도 충분한 양이 구축되지 않으면, 자연언어처리나 지식처리에 실제로 사용하기 어렵다는 점을 감안할 때, PWN의 크기는 지식처리에 필요한 배경지식이나 상식을 구성하거나 자연언어처리 분야의 실용적인 시스템을 개발하는 데 유용하다. 이러한 범용성 때문에 PWN 자체에는 전문분야가 적게 포함되어 있으나, 특정 전문분야 온톨로지나 어휘의미망을 만들 때 PWN은 초기 상위구조를 제공할 수 있다.

셋째, PWN 명사와 동사의 계층적 구조는 상위노드의 의미자질을 하위노드가 계승하게 함으로써, 언어/지식처리의 효율성을 기할 수 있다. 하지만 계층적 구조는 잘 알려진 '테니스 문제(tennis problem)'를 안고 있다[1]. 테니스를 칠 때 일어나는 사건(event)에는 테니스 채, 공, 선수, 심판, 관객, 코트, 의복 등 다양한 구성요소가 개입하게 되는데, PWN의 계층성으로 이들이 서로 관련있다는 점을 표현할 수 없다. 이 단점을 보완하기 위해서, PWN 정의문을 의미 태깅하여 서로 연결함으로써 동일 어휘/개념장 내부의 관련 정보를 나타낼 수 있는 자질을 망 구조로 표시하는 방식이 제안되었고[18], 지속적으로 다른 어휘망/개념망/온톨로지의 정보를 PWN에 연동하는 시도가 있었다[2]. 방사형 핵 구조로 개발된 PWN 형용사의 경우, 반의어를 연상하는 심리적 실재를 반영하지만 그 결과를 바로 언어/지식처리에 활용하기가 쉽지 않다. 따라서 활용의 편의성을 위해 PWN을 참조한 독일어 어휘의미망(GermaNet)의 경우, 형용사의 구조를 계층적으로 재구성하였다[15].

넷째, 표 4에 정리된 신셋/어의 간 의미관계가 다양하고 풍요롭지만, 일부 의미관계는 불투명하고 부분적이다. 우선, 반의를 어형과 밀접한 관계를 맺고 있는 어의 단위로 설정한 것은 PWN이 직간접적으로 참고한 사전의 전통과 의미세분화의 결과에서 비롯된 임시방편적인 정의다. 예를 들어 "move downward and lower, but not necessarily all the way"로 정의되는 신셋 {fall2, descend1}의 반의어로 각각 'rise1("move upward")'과 'ascend1("travel up")'을 연결해 놓았다. 하지만 반의가 의미나 개념의 차원에서 정의될 수 있는 것이라면, ① 이 두 어의는 다른 신셋에 속하므로 다른 개념을 표현하며, ② 역으로 각각의 어의와 동일한 신셋을 이루는 어의와의 연계성은 고려의 대상이 되고 있지 않다는 점에서 논리적 모순을 야기할 수 있다. 따라서 PWN의 반의 관계에 대한 보완 연구와 정제가 필요하다. 또한 형용사의 참조(see also)와 동사군(verb group)은 좀 더 명확한 정의를 필요로 한다. 동사군의 관계 설정이 제한된 범위에만 적용된다는 언급만 있을 뿐, PWN의 문서나 논문에 동사 간 유사한 의미를 측정하는 통계적인 방법이나 형용사의 참조 관계를 검증하는 방식에 대한 명시적인 기술이 없다. 그리고 동사와 형용사의 기본 의미관계인 계층 구조와 방사형 핵구조에 연결되지 않은 단독 신셋(orphan node)의 정체성과 그 수의 적정성도 추후 논의할 대상이다. PWN에는 아직도 {travese3, deny6} ("deny formally (an allegation of fact by the opposing party) in a legal suit")과 {wet2} ("supporting or permitting the legal production and sale of alcoholic beverages") 등과 같이 동사와 형용사의 단독 신셋이 각각 223개와 4,828개 있다.

다섯째, PWN은 언어와 문화는 불가분의 관계에 있다는 일반적인 언어의 속성을 반영하듯이, 영미 중심의 서유럽 문화에 편향적이고, 구축 주체의 주관성과 시공간적 한계가 드러난다. 독자적인 문화와 밀접한 의·식·주생활 관련 어휘뿐 아니라, 비교적 보편성을 갖는 국가·정부·종교·축제의 하위분류 등에서 이러한 특성을 쉽게 찾아볼 수 있다.

여섯째, PWN은 영어 의존적이지만, 언어처리에 다각적으로 사용될 만큼 상세한 언어정보를 제공하지는 못한다. 예를 들어, 동사 문형정보는 그 자체가 불완전할 뿐 아니라, 격틀구조, 논항의 종류나 논항의 의미자질 세분화와 같은 언어정보는 정교하지도 풍요롭지도 않다.

일곱째, PWN은 1.5판이 공개된 이래, 50개 정도의 참조방식 어휘의미망이 구축되었으므로, 다국어 처리로의 응용이 매우 용이하다[3]. 언어마다 정교한 어휘의미망이 개발되었다 하더라도 서로 다른 원칙과 기준이 적용되었다면, 각 구성단위와 구조 간의 연계성을 확보하기 어렵다. 그 예로 일본어와 중국어를 대상으로 한 NTT 어휘대계[19]나 HowNet[20]의 경우 자료의 크기는 PWN과 견줄 만하여, 상호 사상을 시도하였으나, 다국어 처리에 직접 이용할만한 결과를 도출하지는 못했다.

여덟째, PWN은 추후 개발된 다른 어휘의미망/개념망과의 사상이 가장 많이 시도되어 활용도가 높으며, 구글의 애드센스(AdSense)는 PWN을 이용하여 정보검색 분야에서 수익모델을 제시하였다[21]. 또한 2002년부터 격년으로 국제학술대회(Global WordNet Conference)를 열어 2008년 제4회 대회를 개최하였으며, 어휘의미망/개

넘망을 언어/지식처리에 활용하는 논의를 지속적으로 활발하게 벌이고 있다.

3. KorLex 구축방식 및 특성

이상과 같은 특성을 가진 PWN을 참조 모델로 하여 2004년부터 2007년까지 한국어 어휘의미망 KorLex가 구축되었으며, 현재도 소규모로 진행 중이다. 현재 KorLex는 명사, 동사, 형용사, 부사, 분류사로 구성되며, 약 13만 개의 신셋과 약 15만 개의 어의를 포함하고 있어, 자연언어처리와 지식공학 시스템에 적용할 수 있는 단계이다. PWN을 참조하였다고 하더라도 KorLex를 구축하는 것은 단순한 작업이 아니며, 어휘의미론과 전산언어학에서 나타나는 유사한 문제에 봉착하게 된다. 3.1절에서 3가지 어휘의미망 구축방식의 특성을 알아보고, 3.2절에서는 대역형 참조구축, 확장형 참조구축, 직접구축 방식을 적용한 KorLex 구축의 구체적인 원칙과 지침을 살펴본다. 3.3절에서는 그 결과로 구축된 KorLex의 의미정보 구조를 소개하고, 3.4절에서는 이를 구축하기 위한 개발자 워크벤치 및 사용자 인터페이스를 간략하게 소개한다. 이 논문에서는 KorLex 구축 방법론과 구축 결과의 개괄적인 소개를 하며, 각 품사별로 나타나는 어휘의미론적인 문제는 앞선 논문에서 좀 더 상세히 다루었고[20-23], 앞으로도 더 발표할 예정이다.

3.1 어휘의미망 구축방식

어휘의미망/개념망을 구축하는 방식으로는 크게 직접구축과 참조구축으로 구분할 수 있다. 새롭고 독립적인 의미체계를 갖추려는 직접구축방식은 하향식(top-down) 또는 상향식(bottom up)으로 이루어질 수 있다. 하향식은 최상위 개념/의미에서 출발하여 하위 단계로 분화하는 과정을 거치는데, 고도의 배경지식을 가진 전문가가 특정 분야 온톨로지나 소규모의 개념망을 수동으로 개발할 때 사용된다. 상향식은 하위어에서 출발하여 더 포괄적인 의미/개념을 가진 상위어를 찾아가는 방식이다. 주로 대규모 어휘의미망을 구축할 때, 그림 1처럼 인간 지식이 집적된 사전의 정의문을 분석하여 상위어를 반자동으로 추출하는 방법을 사용하기도 한다. 이러한 방법은 사전의 정의문이 통제된 경우 효율성을 기할 수 있지만, 기존 사전의 정의문은 어휘가 통제되거나 정형적 표현구조를 갖추지 못한 실정이다. 직접구축 방식은 활용 목적에 맞는 독자적인 어휘의미망/개념망을 만들 수 있다는 장점이 있으나, 개발자의 주관이나 기존 사전에 치우치기 쉽고, PWN이나 NTT 어휘대계와 같이 일반 목적에 쓸 수 있는 범용성을 갖추려면, 구축 범위가 크고 개발 시간과 노력이 매우 많이 소요된다. 참조방식은 직접 방식 등으로 이미 만들어진 어휘의미망을 근간으로 이를 다른 언어로 번역하는 방법을 사용함으로써

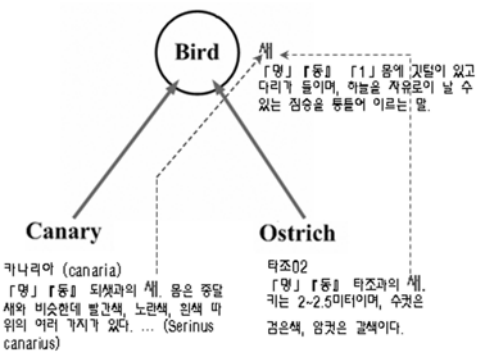

그림 1 사전정의문 기반 상향식 직접구축

개발기간을 단축한다. PWN과 NTT 어휘대계를 참조모델로 한 파생 어휘의미망인 EWN[25,26], BWN[27,28], CoreNet 등이 그 예인데, 파생 어휘의미망은 피참조 어휘망에 경도되어 있다는 단점에도 불구하고, 피참조 어휘망과 파생 어휘망 간에 개념의 표상 단위가 동일하여, 다국어 연계성을 확보하는 데 유리하다[12]. 물론 프랑스어 워드넷(FWN) 등과 같이 피참조 어휘망을 대역하는 단순한 경우도 있으나, 본질적으로 이질적인 자연언어가 동일한 의미체계로 표상될 수 없을 뿐만 아니라, 참조 어휘망과 파생 어휘망 사이의 언어계보 및 언어문화적 역사성에서 공유점이 적으면 적을수록 대역형 참조구축의 문제점은 더욱 커진다. 이를 보완하기 위해, 일반적으로 자국어 의미구조와 사용목적에 맞도록 변환하는 확장형 참조구축 방식을 택한다. 실제 어휘망 구축 과정에서는 이상과 같이 분류한 방식 중 어느 하나를 배타적으로 적용하기보다는 상호의 단점을 보완하는 통합적(hybrid) 방식을 사용한다.

3.2 KorLex 구축방법론

KorLex 1.0 단계에서는 PWN 2.0을 대역한 후, KorLex 1.5부터는 기존 신셋의 삭제/변경과 새로운 신셋의 생성에 상/하향 직접구축 방식을 통합하여 적용한다. PWN의 신셋에 적합한 한국어 어휘의미를 사상하는 1단계와, 이를 바탕으로 확장과 변환을 모색하는 2단계에서 모두 고려해야 할 사항은 일관성을 유지하는 것이다. 이를 위해 KorLex는 한국어에 적용될 의미세분화의 기준을 『표준국어대사전』([29] 이하 『표준』)에 두었다. 『표준』은 어느 사전에나 나타나는 거시적·미시적 구조의 부분적 결함을 갖고 있다. 하지만, 특정한 언어학 이론에 치우치지 않았으며, 실제 말뭉치를 이용하여 예문을 제공하고, 비교적 의미세분화의 기준을 명시적으로 공표하였다. 또한 주관 구축기관인 국립국어원이 개선과 확장을 추진하고 있으므로, 앞으로 KorLex와 지속적인 상호보완 가능성이 가장 높다. 이에, KorLex를 구

축하면서 의미세분화와 관련된 『표준』의 문제점을 검토하고, 부분적으로 그 해결 방식을 제안하고 있다[30].

3.2.1 KorLex1.0의 대역형 참조구축

대역형 참조구축의 장점은 구축 시간과 비용을 대폭 단축하는 것이므로, 이중어 사전과 단일어 사전 등을 이용한 (반)자동 구축이야말로 이러한 장점을 극대화할 수 있는 방법이다[31]. EWN의 FWN 등의 구축에서 실제로 적용되었다. 하지만, 영어의 어휘 중 70%는 프랑스어 어원을 가지나, 형태를 기준으로 한 단순한 어휘대치가 많아 FWN의 결과는 그리 탐탁하지 않다. 그 결과 EWN에서도 함께 구축된 이탈리아어, 네델란드어, 스페인어 어휘망과는 달리 FWN의 활용 가능성이 낮다고 평가된다. 한국어는 영어와 언어 계통이 다르며 공유하는 문화가 적다. 더욱이 한자어를 어원으로 하는 동형이의어가 많으므로, 참조구축 방식에서 영-한 사전을 이용한 대역어의 (반)자동 선택은 그 정확도가 매우 떨어지며, 피대역어와 대역어 관계가 1:1이라는 매우 제한된 경우에만 적용될 수 있을 뿐이다. 따라서 KorLex 1.0의 대역형 참조구축은 반자동으로 이루어졌다. KorLex 워크벤치(그림 7)에서 영-한 사전을 이용하여 전처리된 대역어 후보를 제공하면, 첫 단계로 10명의 어휘전문가 또는 해당전문분야 전공자에 의해 대역어 선정이 이루어지고, 다음 단계에서 2명의 의미론 전공 박사가 검증하였다.

KorLex 1.0을 구축하면서 적용한 원칙과 지침은 다음과 같다. 단, 계층 구조를 갖지 않는 KorLexAdj 1.0과 KorLexAdv 1.0에는 ①-⑥이 적용된다.

① PWN 2.0의 신셋, 신셋 간 계층 구조 및 방사형 핵상 구조는 변경하지 않는다.

② PWN의 대역의 방향은 말단노드에서 상위노드로 향하는 상향식(bottom-up) 구축을 원칙으로 하되, 대역 순서는 다음과 같이 그룹화하여 진행한다.
   ⓐ 어형이 단의어로 쓰이며, 신셋이 1개의 어의로 구성된 경우(표 7의 A)
   ⓑ 어형이 단의어로 쓰이며, 신셋이 2개 이상의 어의로 구성된 경우(표 7의 C-A)
   ⓒ 신셋이 1개의 어의로 구성되며, 해당 어의의 어형이 다의어로 쓰이는 경우(표 7의 B-A)
   ⓓ 신셋이 2개 이상의 어의로 구성되며, 해당 어의의 어형이 다의어로 쓰이는 경우(표 7의 다의어∩(~B))

③ 대역어 선정은 각 신셋을 대상으로 한다.
   ⓐ KorLex 신셋의 구성은 어의 이외의 단위인 영(zero)형태, 접사, 어휘, 관용표현, 구, 절 등으로 나타낼 수 있다. (KorLex의 신셋이 영형태가 되는 어휘 공백(lexical blank)의 경우에는, PWN의 신셋을 그대로 유지한다.)
   ⓑ PWN과 KorLex의 동일 신셋을 구성하는 어의 수는 일치하지 않을 수 있다.

④ 대역어 후보의 검색은 PWN의 어형을 기준으로 한다.
   ⓐ PWN 신셋의 의미관계 중 전문 분야 및 어법 정보는 대역어 선정에서 우선적으로 고려한다.
   ⓑ PWN 신셋이 단일 어의로 구성된 경우, 해당 어형의 대역어 후보 중 다수 사전에 출현한 빈도에 따라 대역어(들)을 선택한다.
   ⓒ PWN 신셋이 2개 이상의 어의로 구성된 경우, 모든 어의에 대응하는 어형의 대역어 후보 중 빈도에 따라 대역어(들)을 선택한다.
   ⓓ 동형이의어 및 다의어를 구분하기 위해 각 대역어 어의별로 『표준』의 세분화된 의미와 사상한다. 『표준』에 수록되지 않은 어의는 출처, 정의문, 예문과 함께 KorLex 사전에 새로 등재한다. (이때 정의문과 예문은 필수적 구성 요소가 아니다.)

⑤ 대역어 선정은 PWN의 품사별, 의미분류별로 진행하며, 해당 부류에 따라 대역어 후보를 검색할 영-한 사전의 참조 순위를 결정한다.
   ⓐ 하위노드의 신셋은 PWN의 의미분류(동물, 식물 등) 및 전문분야(컴퓨터, 무기, 화학, 선박, 음악, 해부학, 미술 등) 정보에 따라 해당 영-한/한-영 전문용어 사전을 우선 순위로 참조하며, 상위노드로 갈수록 범용 영-한/한-영 사전을 참조한다.
   ⓑ 상위노드의 신셋 및 일반 어형은 범용 영-한/한-영 사전을 우선 순위로 참조한다.
   ⓒ 사전에서 등재되지 않은 어형은 공신력을 가진 웹사이트를 참조한다.

⑥ KorLex의 한 신셋을 구성하는 대역어 후보들의 동의관계는 다음 중 한 조건을 만족해야 한다. 동의관

표 7 PWN 2.0에서 단일 어의로 구성된 신셋과 단의어 및 다의어 크기

품사	A=B∩C	단일 어의로 구성된 신셋(B)	단의어(C)	다의어		(C+E)/D	E/D
				어형(D)	어의(E)		
명사	12,822	40,755	99,524	15,124	42,325	1.23	2.79
동사	1,418	7,855	6,256	5,050	18,522	2.17	3.66
형용사	5,795	11,366	16,103	5,333	14,979	1.44	2.80
부사	1,570	2,323	3,901	768	1,913	1.24	2.49
계	21,605	62.299	125,784	26,275	77,739	1.24	2.49

계의 설정은 EWN에서 제시한 조건('Word1 in contextc entails and is entailed by Word2 in contextc.')을 사용하였다[25].

ⓐ PWN 신셋의 예문을 한국어로 옮겼을 때, 그 예문 내에서 대역어 후보들은 의미를 크게 변화하지 않고 교체 가능해야 한다.

ⓑ 영-한 사전 또는『표준』에 대역어 후보의 어의에 적합한 한국어 예문이 있다면, 그 예문 내에서 대역어 후보들은 의미를 크게 변화하지 않고 교체 가능해야 한다.

ⓒ 하지만, '동일한 시니피에가 2개 이상의 시니피앙으로 표현되지 않는다.'는 언어의 경제성 원리에 따라 엄밀한 의미에서 동의관계를 교체로만 판단할 수는 없다. 특히 동사의 경우, 문형 및 논항의 제약 조건의 차이로 교체로서 동의관계를 검증하기 어렵다는 점을 감안해야 한다[1].

⑦ 상하위 관계는 다음과 같은 내포 조건을 만족해야 한다[25].

(KorLex 신셋n에 속하는) 어의n는 문맥n-1에서 (KorLex 신셋n-1에 속하는) 어의n-1를 내포해야 하며, 역은 성립하지 않는다. 이때 n은 KorLex의 계층을 나타낸다.("Wordn (which belongs to KorLex-Synsetn) in contextc-1 entails Wordn-1 (which belongs to KorLex-Synsetn-1), and the reverse is not allowed (where n stands for a level of the KorLex hierarchical structure), where n represents for the KorLex hierarchy.")

⑧ 동일한 어형의 서로 다른 어의가 KorLex에서 상하위 또는 자매 관계가 성립하는 경우, 다음 조건을 만족해야 한다.

ⓐ 동일 사전에서 상위 신셋 어의는 하위 신셋 어의보다 더 하위 단계에서 선택할 수 없음을 원칙으로 한다.

ⓑ 자매관계에 놓인 대역어 어의는 같은 세분화 단계에 제시된 어의를 선택함을 원칙으로 한다.

이상의 원칙과 지침에 따라 2004년 1월 - 2007년 4월 동안 대역형 참조구축 방식으로 이루어진 KorLex 1.0의 결과는 표 9와 같다. 결과의 효용성에 따라 명 -> 동 -> 형, 부 순으로 수행하였다. 위 원칙 ①에 따라 신셋 A는 KorLex 1.0에 그대로 유지되는 PWN 2.0의 신셋 수이며, 신셋 B와 어형, 어의의 수는 한국어로 대역된 경우만을 나타낸다. 표 3의 PWN 2.0과 크기를 비교해 볼 때, 대역이 되지 않은 명사의 비율이 현저하게 높다. 이는 명사에 영어(권 문화)와 한국어(권 문화) 간의 차이에서 생긴 개념 공백이나 어휘 공백이 있고, 대역어를 선정하는 초기 작업의 미숙함에도 기인하나, 한국어 대역어를 찾기 어려운 동식물명·병명 등과 같은 전문용어와 인명·지명·개체명 등 고유명사의 수가 많았기 때문이다. KorLexNoun 1.0에서는 전문용어나 고유명사 등은 대역하지 않았으나, KorLexNoun 1.5에서는 이중 한국어 음차 표기가 있는 경우 대역하였다. KorLexNoun 1.0의 크기는 PWN을 참조모델로 대역한 다른 언어 어휘의미망의 크기와 유사하다[8,10,14,26].

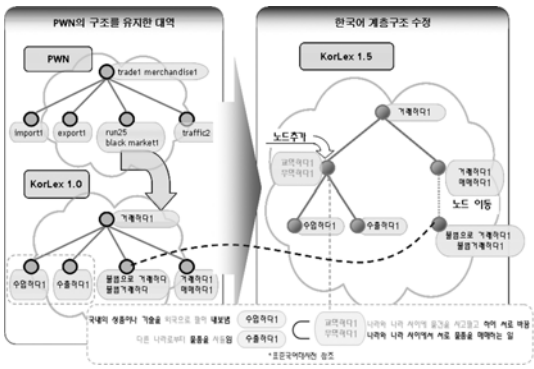

그림 2 KorLex의 확장형 참조구축 방식

3.2.2 KorLex1.5의 확장형 참조구축 방식

KorLex 1.5의 구축은 어휘의미 추가확장과 계층 구조 변환이라는 두 가지 측면에서 수행되었으며, 현재까지 명사와 동사에만 적용되었다. 우선 어휘의미 추가확장은 KorLex 1.0에 결여된 어휘형태를 보완하는 것에서 시작한다.

대역을 통해 구축한 KorLex 1.0은 자주 쓰이는 '밥, 그저께, 파탄, 하수인, 하극상, 넘다, 삼다, 인하다, 비롯하다, 쓰이다'와 같은 어휘를 포함하지 못했다. 이를 보완하기 위해, KorLex1.5의 1단계 확장은 국립국어원의 현대국어 사용빈도를 조사한 자료[32,33]에서 명사는 5

표 8 KorLex 1.0 (대역형 참조구축) 결과

품사	어형	신셋			어의	구축 시기
		A (PWN 2.0)	B (유대역)	A-B (무대역)		
KorLexNoun 1.0	53,167	79,789	58,565	21,224	59,405	2004년 9월
KorLexVerb 1.0	14,261	13,508	13,429	79	14,700	2006년 2월
KorLexAdj 1.0	19,698	18,563	18,558	5	20,905	2007년 4월
KorLexAdv 1.0	3,032	3,664	3,651	13	3,123	2007년 4월
계	87,126	115,524	90,552	24,972	95,010	

표 9 KorLexNoun 1.5에 추가된 어의 정보 및 추가 방식

	추가 대상 어형	추가 대상 어의의 『표준』 사전적 어의 정보		상위어	하위어	동의어, 유의어	반의어	한영 대역	KorLex 추가 및 확장 방법
(1)	돈	정의문	무게의 단위. 귀금속이나 한약재 따위의 무게를 잴 때 쓴다.	(무게의) 단위					'(무게의) 단위'의 하위어로 새로운 신셋 생성
(2)	집	정의문	칼, 벼루, 총 따위를 끼거나 담아 둘 수 있게 만든 것		칼집				'칼집'의 상위어로 새로운 신셋 생성
		예문	칼을 잘 닦은 후 집에 넣어 보관해라.						
(3)	소리	정의문	사람의 목소리.			목소리			'목소리'와 동일 신셋의 구성요소로 추가
		예문	소리가 너무 크니 조용히 말해라.						
(4)	소리	정의문	여론이나 소문.			여론, 소문			'여론, 소문'과 동일 신셋의 구성요소로 추가
		예문	주민들 사이에 이상한 소리가 돌고 있다. 침묵하는 다수의 소리에 귀를 기울여 보라.						
(5)	눈	동의어	시력01(視力).			시력01			'시력'과 동일 신셋의 구성요소로 추가
		예문	눈이 나빠 안경을 쓴다.						
(6)	온기	정의문	따뜻한 기운.			난기03	냉기03		'냉기'의 자매 노드에 '온기, 난기'라는 새로운 신셋 생성
		유의어	난기03(暖氣)						
		반의어	냉기03(冷氣)						
(7)	날	정의문	하루 중 환한 동안					daylight, daytime	'낮, 대낮, 백주, 하루해, 하루'와 동일 신셋의 구성요소로 추가
		예문	날이 새면서 주위가 밝아 온다.						

회 이상, 동사는 3회 이상 출현한 표제어(어형)를 그 대상으로 삼았다. 이 자료는 동형이의어나 다의어를 구분하지 않고 품사와 어형을 기준으로 빈도를 제시하였으므로, 이중 KorLex 1.0에 포함되지 않은 어형을 선정하고, 『표준』의 의미세분화에 기대어 이 표제어의 어의를 되도록 충실히 추가하되, 고어, 지방어, 특수 전문용어 및 사용빈도가 현저하게 낮은 어의는 확장 대상에서 제외하였다. 이 자료에서 4회 이하 출현하는 어휘는 대부분 '어름치'처럼 동식물명과 같은 전문용어, '십정(十停)'과 같은 특정 시대의 기관/관직명이었다. 범용적 지식을 구성하려는 KorLex의 개발 원칙에 따라, 이상의 특수 분야 어휘는 1.5버전에 포함하지 않았다. 또한 빈도가 5가 넘더라도, '사정(司正, 조선 시대에, 오위(五衛)에 속한 정칠품 벼슬)'처럼 특정한 시대에만 사용한 용어 또는 어의나, '애시당초'처럼 오류어도 포함하지 않았다.

추가되는 어의는 신셋의 구성요소로 추가될 수도 있고, 또는 새로운 신셋을 만들 수도 있다. 첫 번째 경우는 이 논문 3.2.1 절의 ⑥번 조건을 만족해야 하며, 좀 더 신중함이 필요한 두 번째 경우는 ⑦번과 ⑧번 조건을 충족해야 한다.

KorLexNoun 1.5 경우를 예로 들어 보자. 예(1)은 그 정의문에서 '(무게의) 단위'라는 중심어를 추출하고, 그 하위 노드에 새로운 신셋을 생성한다. 예(2)처럼 정의문보다 예문이나 복합어 등에서 '칼집'과 같은 하위어 정보를 추출하고 이것이 이미 어휘망에 존재하고 있다면 그 상위어로 '집'이라는 새로운 신셋을 생성한다. 예(3)-(5)처럼 명시적인 동의어(=), 유의어(≒)나, 정의문에 등가 표현이 제시되는 경우 그 어의에 해당되는 기존 신셋의 구성요소로 추가한다. 예(6)의 반의어인 '냉기'가 기구축되어 있다면 그 자매 노드에 추가대상 어의와 유의어 '온기, 난기'를 새로운 신셋으로 생성한다. 예(7)과 같이 정의문과 예문으로부터 단서를 찾을 수 없고 다른 의미정보도 주어지지 않는다면, 영-한 사전을 이용하여 '날'에 해당하는 PWN의 신셋을 찾아 그 구성요소로 추가한다.

확장형 참조구축은 4명의 어휘전문가가 1차 추가확장을 하고, 그 결과 전체를 상호교차 검토한 후, 2명의 의미론 전공 박사가 검증하였다. 그 결과 KorLexNoun/Verb 1.5는 표 10과 같다. KorLexNoun/Verb 1.5에는 한국어로 대역되지 않은 7,316개(명사)와 102개(동사)의 신셋이 존재한다. 명사는 데이타베이스를 등록한 2007년 7월 이후에도 소규모로 확장을 계속하고 있다. 계층별 신셋 수를 참조모델이 된 PWN 2.0과 비교한 결과는 표 11과 같다. 명사는 4단계-8단계가 주로 확장된 것을 볼 수 있는데, 이는 매우 추상적이고 광범위한 개념이 주로 명사의 1-3단계에 주로 분포하는 반면, 의미의 크

표 10 KorLexNoun 1.5와 KorLexVerb 1.5 (확장형 참조구축) 결과

품사	어형	신셋		어의	개발 시기
		A (PWN 2.0)	B		
KorLexNoun 1.5	89,125	79,689	90,134	102,358	2007년 7월
KorLexVerb 1.5	17,956	13,508	16,923	20,133	2007년 4월
계	107,081	93,197	107,057	122,491	

표 11 PWN 2.0과 KorLex 1.5 계층별 신셋 수 비교

계층	PWN 명사 2.0	KorLexNoun 1.5	PWN 동사 2.0	KorLexVerb 1.5
1	9	9	554	600
2	158	157	3,210	3,864
3	1,307	1,653	3,819	4,896
4	4,489	6,033	2,962	3,759
5	10,297	13,129	1,598	2,040
6	17,536	19,236	737	985
7	15,336	18,079	363	462
8	12,225	13,802	146	180
9	7,605	8,053	41	50
10	4,793	4,714	41	44
11	2,501	2,305	25	30
12	1,444	1,256	11	11
13	852	733	1	2
14	477	429		
15	415	346		
16	206	164		
17	39	36		
계	79,689	90,134	13,508	16,923

기가 작고 구체적인 어의가 분포하는 층위이기 때문이다. 이에 비해 PWN에서부터 얕고 넓은 분포를 가진 동사는 2-5단계가 확장된 것도 언어의 실제 모습과 일치한다[10,14].

3.2.3 KorLexClas 1.0의 직접 구축 방식

영어, 프랑스어 등의 인구어(Indo-European languages)와 달리 대부분의 아시아어, 아프리카어 등처럼 한국어도 정교한 분류사 체계를 지닌 것으로 알려져 있다. 분류사의 기능은 사물이나 사건을 범주화하고, 수량화하는 것으로 어휘의미망의 '분류'와 '개념화'라는 본질적인 특성이 있다. 이때 사물이나 사건은 분류사의 공기 관계(co-occurrence)로 표상되므로, 분류사와 명사 간 비교적 강력한 공기 제약을 갖는다. KorLexClas 1.0은 한국어 언어자원을 이용한 직접구축 방식으로 개발하였다[24].

1단계로, 고빈도 분류사의 완전한 목록을 구성하고 공기 명사 정보가 함께 태깅된 자료를 확보하기 위해, 선행 언어학 연구, 『표준』의 정의문, 대용량 말뭉치의 문맥 정보를 이용하여 분류사 및 공기 명사 목록을 수집한다. 2단계로, 분류사의 의미적 특성을 고려하여 ⓐ 도량성(mensural), ⓑ 개체성(sortal), ⓒ 중립성(neutral),

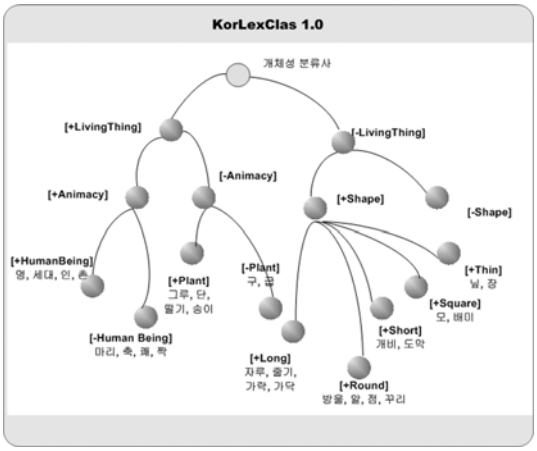

그림 3 KorLexClas 1.0의 분류사 의미자질 계층성

표 12 KorLexClas 1.0 (직접 구축) 결과

품사	어형	신셋, 어의		개발 시기
KorLexClas 1.0	1,181	도량성	856	2007년 4월
		개체성	424	
		중립성	4	
		사건성	93	
		계	1,377	

ⓓ 사건성(eventuality)으로 하위범주화하여 각각 분류사를 정의하고, 그림 3처럼 분류사의 의미자질 간 계층관계를 설정한다. 3단계로, 분류사-공기 명사 간 선택제약 관계를 설정하기 위해 KorLexNoun과 연동한다. 표 12에서 볼 수 있듯 분류사의 모든 신셋은 단일 어의로 구성된다. KorLexClas는 앞서 기술한 다른 품사 어휘망과는 달리 분류사 자체가 아니라 분류사를 구성하는 의미자질을 계층화하였다.

이상과 같이 구축된 KorLexNoun/Verb 1.5 및 KorLexAdj/Adv/Clas 1.0의 크기는 표 13, 14와 같다. 표 7에 제시한 PWN 2.0과 비교해 보았을 때 동사에서 가장 큰 차이를 보이는데, 영어에서는 중립 동사 등으로 나타나는 단일 어형의 다의어가 한국어에서는 선어말 어미의 유무로 어형을 구분할 수 있는 단의어로 대역되었기 때문이다[22]. 한자어 어근을 많이 사용하는 한국어의 동형이의어 및 다의어 비율이 높다는 특성과는 달리 KorLex의 다의어 비율이 PWN과 유사하게 나타나

표 13 KorLex 1.5 구축 현황 (총괄표)

품사	어형	신셋		어의	개발 시기
		A (PWN 2.0)	B		
KorLexNoun 1.5	89,125	79,689	90,134	102,358	2007년 7월
KorLexVerb 1.5	17,956	13,508	16,923	20,133	2007년 4월
KorLexAdj 1.0	19,698	18,563	18,558	20,905	2007년 4월
KorLexAdv 1.0	3,032	3,664	3,651	3,123	2007년 4월
KorLexClas 1.0	1,181	-	1,377	1,377	2007년 4월
계	130,992	115,424	130,643	147,896	

표 14 KorLex 1.5의 단의어/다의어 크기

품사	단의어(C)	다의어		(C+E)/D	E/D
		어형(D)	어의(E)		
KorLexNoun 1.5	80,953	8,172	21,405	1.15	2.62
KorLexVerb 1.5	16,437	1,519	3,696	1.12	2.43
KorLexAdj 1.0	18,695	99	2,202	1.06	2.20
KorLexAdv 1.0	2,958	74	165	1.03	2.23
KorLexClas 1.0	1,083	98	294	1.17	3
계	120,126	9,962	27,762	1.13	2.56

는 이유는, 확장형 참조구축 시 KorLex 1.0에 없는 어형을 우선 추가 대상으로 삼았기 때문이다. 따라서 KorLex 구축이 더 진행될 때 한국어에서 다의어 비율이 높은 어형을 추가 대상으로 고려해 봐야 한다. 표 15는 KorLex 어의별 정의문 출처를 보여준다. 『표준』의 정의문을 이용하여 구축한 U-Win과 사상을 추후에 시도할 때 고려해야 할 부분이다.

3.3 KorLex 1.5의 의미정보

기본적으로 KorLex의 신셋은 사상되는 PWN의 신셋이 가진 의미정보(표 4의 신셋 간 의미관계)를 승계한다. 하지만 PWN의 영어 의존적 정보인 문법 범주, 파생 관계, 문장 격틀은 다른 언어에서 유효하지 않으므로, KorLex 1.5에 새롭게 구축되었고, 향후 지속적으로 구축되어야 할 대상이다.

첫째, 내용어를 명사, 동사, 형용사, 부사로 나누고 첫 2개 범주는 계층적 구조로, 형용사는 방사형 구조, 부사는 목록으로 제시한 PWN의 구분에 KorLex 1.5와 KorLex 1.0은 아직 수정을 가하지 않았다. 하지만 한국어의 경우 동사와 형용사로 구분하기보다 용언으로 통합하는 것과 통합했을 때 개념 간 관계를 어떤 구조로 표상할지에 대한 논의가 필요하다.

둘째, 파생 정보로는 KorLexNoun/Verb 1.5 중 『표준』에 수록된 '확장-확장하다, 명령-명령하다' 등과 같이 '어근 명사'+'기능동사(-하다, -되다)'의 관계가 표시된다.

셋째, 용언의 격틀정보와 논항의 의미자질/분류를 명사 어휘망과 연결한다면 자연언어처리 제 분야에서 매우 유용하게 사용될 수 있을 것이다. 하지만 기존의 어떤 언어자원에서도 KorLex에 직접 사용할 수 있는 이러한 정보를 찾기 힘들다. KorLex가 어의 구분의 기준으로 삼은 『표준』은 표 16에서 볼 수 있듯 격틀정보가 기술되지 않거나(타다1-1 ~ 타다1-5) 일부의 격틀정보만이 매우 거친 상태로 제시되며(타다4-1 ~ 타다4-2), 논항의 의미정보는 명시적으로 기술되지 않아 정의문이나 예문을 통해 추정해야 한다. '세종전자사전'(이하 『세종』 [7,8])의 경우도 용언의 격틀정보와 논항의 의미부류를 명세화하고 있지만, 어의 구분 및 논항의 의미분류 구분기준이 『표준』이나 PWN 및 KorLex와는 완전히 다르므로, 『세종』에 담긴 정보를 손쉽게 사상할 수 없다. 따라서 KorLex는 『세종』, 『표준』 등을 참조하되, 신셋을 구성하는 어의 단위로 격틀정보를 추가 기술

표 15 KorLex 어의의 한국어 정의 출처

사전	KorLexNoun 1.5	KorLexVerb 1.5	KorLexAdj 1.0	KorLexAdv 1.0	KorLexClas 1.0
표준국어대사전	65,879	9,617	17,647	2,913	924
연세 한국어 사전	134	217	156	41	0
브리태니커 백과사전	34	0	58	2	0
프라임 영한사전	0	0	461	0	0
네이트 백과사전	15	0	32	0	0
네이버 백과사전	19	0	22	0	0
파스칼 백과사전	17	0	7	0	0
동의어 사전	1	0	0	0	0
기타	36,161	10,247	2,065	5	446
없음	98	52	457	162	7
계	102,358	20,133	20,905	3,123	1,377

표 16 『표준』과 『세종』의 격틀 및 논항 의미부류 정보

『표준』의 의미세분화	정의문	예문	『표준』의 격틀정보	『세종』의 격틀와 논항의 선택 제약
타다1-1	불씨나 높은 열로 불이 붙어 번지거나 불꽃이 일어나다.	담배가 타다; 벽난로에서 장작이 활활 타고 있었다.	-	[구체물]N0-이 [구체자연물(불)]N1-에 타다
타다1-2	피부가 햇볕을 오래 쬐어 검은색으로 변하다.	땡볕에 얼굴이 새까맣게 탔다.	-	[신체부위]N0-이 [기상관련물(볕|햇볕|봄볕)]N1-에 타다
타다1-3	뜨거운 열을 받아 검은색으로 변할 정도로 지나치게 익다.	고기가 타다	-	[음식물|자연음식물]N0-이 타다
타다1-4	마음이 몹시 달다.	입술이 마르고 심장이 탄다.	-	[신체부위(입|입술|목)]N0-이 타다
타다1-5	물기가 없어 바싹 마르다.	긴장이 되어 입술이 바짝바짝 탄다; 오랜 가뭄으로 농작물이 다 타 버렸다.	-	[식물(농작물|벼)]N0-이 타다
타다4-1	몫으로 주는 돈이나 물건 따위를 받다.	회사에서 월급을 탄다; 형에게서 용돈을 탄다; 경연 대회에서 상을 탄다;	[…에서/에게서/에게/으로부터 …을]	[인간]N0-이 [단체|인간]N2-에게서 [돈(|상)]N3-를 타다.
		병원에서 정기적으로 약을 타 먹다.		-
타다4-2	복이나 재주, 운명 따위를 선천적으로 지니다.	좋은 팔자를 타고 태어나다; 그녀는 아버지의 음악적 소질을 타고 태어났다.	[…을]	[인간]N0-이 [추상적대상(정기|사주|팔자|운명|재주)]N1-를 타다.

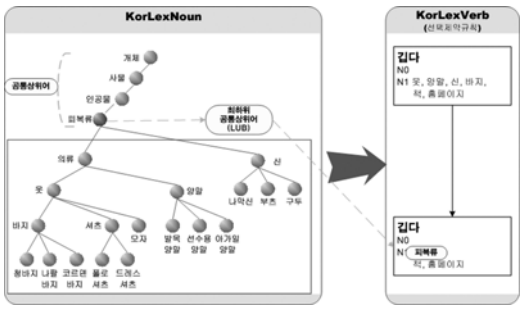

그림 4 KorLexNoun과 KorLexVerb 간 논항의 선택 제약 연동1

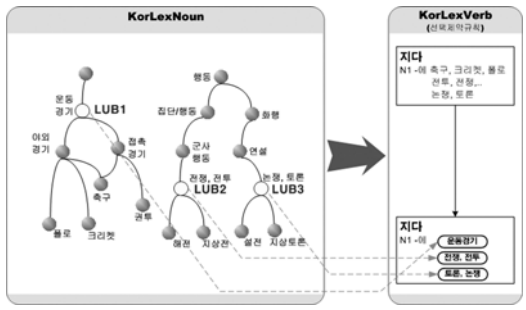

그림 5 KorLexNoun과 KorLexVerb 간 논항의 선택 제약 연동2

하고, 논항의 선택 제약(selectional restriction)은 KorLexNoun과 연동한다.

넷째, KorLex는 영어에는 잘 쓰이지 않으나 한국어에는 매우 발달한 분류사 어휘의미망을 새롭게 구축하였다[24]. 『표준』과 『세종』 등 사전과 분류사에 관한 선행연구를 통하여 1,377개의 분류사 신셋을 구성하고, 이들 간 계층 구조를 설정하고, 각 분류사와 공기하는 명사 정보로 KorLexNoun를 연동한다.

논항의 선택 제약과 분류사의 공기 명사를 연동하는 방식은 그림 4처럼 KorLexNoun에서 최하위 공통 상위 노드(Least Upper Bound Node, 이하 LUB)를 찾는 것이다. 예를 들어 '[N0]-이 [N1]-을 깁다'의 경우, [N1]에 '옷, 양말, 신' 등의 선택 제약이 가해진다면, 이들의 공통 상위 노드 중 최소공배수 격인 {피복류} 노드를 지정하고, 그 하위노드 전체에 제약규칙이 적용되는 것이다. LUB는 복수로 지정이 가능한데, 그림 5처럼 '[N0]-이 [N1]-에 지다'에서는 [N1]의 LUB는 {운동경기}, {전쟁, 전투}, {토론, 논쟁}이 된다. 마찬가지로 LUB가 지정되었을 때 하위노드 중 선택 제약의 적용이나 공기되지 못하는 경우는 Neg-LUB(Negative LUB)으로 지정할 수 있다. 예를 들어, 분류사 '마리'가 공기명사로 {동물}이라는 상위계층의 노드를 지정하되, 그 하위계층에 속하는 {인간, 사람}을 제외해야 한다면 후자를 Neg-LUB으로 지정하는 식이다.

이상과 같은 KorLex의 모든 의미 및 관계정보는 XML로 정의되며, 그림 6처럼 다음과 같은 4개 테이블의 관계형 DB로 설계하였다. ① 신셋 정보 테이블(PWN과 KorLex의 신셋 정보), ② 어의 정보 테이블

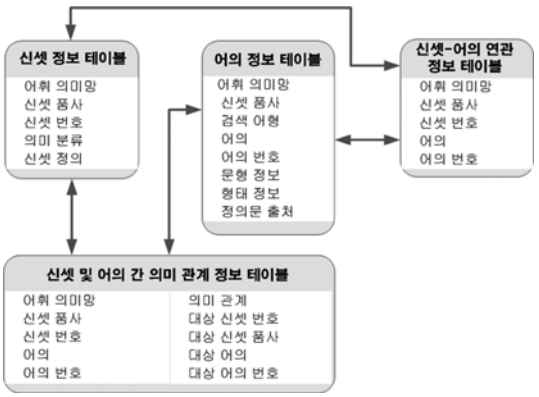

그림 6 KorLex 의미정보 구조

(어의 별 정보), ③ 신셋-어의 연관 정보 테이블 (각 신셋과 그 구성요소인 어의 간 관계 규정), ④ 신셋 및 어의 간 의미관계 정보 테이블 (신셋 간 또는 어의 간 표 4와 같은 의미정보 규정). 표 17은 신셋 정보 테이블의 신셋 정의 필드에 수록된 신셋 {지다1, 패배하다1, 패하다1}(동사 01064957)의 예를 보여준다.

3.4 KorLex 구축 및 검색 서비스 도구

KorLex 1.0 및 1.5 모두 자동적으로 수행될 수 있는 부분보다 어휘전문가의 정밀한 판단을 요구하는 경우가 많으며, 이때 PWN, 이중어/단일어 전자사전, 말뭉치 등 다양한 온라인/오프라인 언어자원을 참조하고, 참조한 언어자원의 출처를 자동으로 수록해야 한다. 동시에 3그룹의 어휘전문가가 '대역-> 대역 검토 -> 확장 및 변환 -> 확장 및 변환 검토' 등 적어도 4단계 이상의 작업을 공동으로 진행해야 하므로, 모든 작업 내역을 기록해야 한다. 또한 불필요한 정보나 지나치게 많은 언어정보는 오히려 구축의 효율성을 크게 떨어뜨린다. 따라서 효율적인 방식으로 필요한 정보를 어휘전문가에게 제공해야 한다. 이에, 본 연구진은 그림 7처럼 개발자용 워크벤치인 LRMS(Language Resource Management System)를 자체 개발하여 사용하며 일반 사용자를 위해 그림 8와 같은 KorLex/PWN 검색 Browser를 제공하고 있다[34].

LRMS에서는 기구축된 어휘망의 각종 정보를 검색할 수 있으며, 효율적인 확장/변환 작업이 용이하다. 이 워크벤치에서 PWN이나 KorLex의 특정 신셋을 선택하면, 해당 신셋의 PWN 정의, 계층 구조와 관련 의미 정보, 이와 연동된 다른 참조 어휘망의 정보, 사상된 『표준』의 어의, 작업자 및 작업 내역 기록을 검색할 수 있다. 변경/확장의 경우는 다음과 같은 절차로 수행된다. ① 고빈도 어휘목록과 KorLex1.0을 비교하여 결여된 어휘형태를 검색하여, ② 우선 순위대로 사전을 검색하고 적합한 어의를 등록한 후, ③ 정의문의 중심어 등에 기반하여 상위어를 찾는다. ④ 해당 어의가 ⓐ 기존 신셋에 추가하는 경우와 ⓑ 새로운 신셋을 만드는 경우에 따라 입력창을 띄우고, 이와 관련된 의미정보를 추가/삭제한다.

4. 활용 및 향후 개발 방향

이 논문에서는 1980년대 중반부터 20여 년간 구축한 영어 어휘의미망 PWN과 이를 참조하여 구축한 한국어 어휘의미망 KorLex를 소개하였다. 심상어휘집(mental lexicon)임을 표방하며, 지식이 인간의 뇌에 어떤 방식으로 저장되며 처리되는지를 살펴보기 위한 시발점으로 만들기 시작한 PWN은 인지심리학보다 자연언어처리와 지식공학에 훨씬 더 큰 반향을 불러 일으켰다. 인지심리학자인 밀러는 이 점을 매우 아쉬워 하나[1], 동일한 자료를 대하는 두 분야의 시각 차이를 극명하게 드러낸다. 전자는 근본적으로 PWN의 의미표상 방식이 인간이 의미를 처리하는 실제와 같은지 의심을 품었다. 후자는 자료 자체의 크기, 표상 방식의 체계성, 절차적 수행의 수월성에 주목하였다[35,36].

표 17 동사 01064957 {지다1, 패배하다1, 패하다1}의 신셋 정보

신셋 정보	설명
<SYN pos="v" lexfn="verb.competition" soff="01064957" descendent="0">	PWN 품사, 의미분류, 신셋번호, 하위노드 유무
<POINTER symbol="Topic-Domain-of-Synset" tsoff="00407449" tpos="n" />	PWN의 신셋의 의미관계: 영역(전문분야)
<POINTER symbol="parent" toff="01064559" tpos="v" />	PWN의 신셋의 의미관계: 상위노드
<GLOSS>lose (a game); "The Giants dropped 11 of their first 13"</GLOSS>	PWN의 정의문
<DOMAIN>verb.competition</DOMAIN>	KorLex 의미분류
<WORD senseid="1" seq="0">지다</WORD>	KorLex 신셋을 구성하는 어의 1: '지다1'
<WORD senseid="1" seq="1">패배하다</WORD>	KorLex 신셋을 구성하는 어의 2: '패배하다1'
<WORD senseid="1" seq="2">패하다</WORD>	KorLex 신셋을 구성하는 어의 3: '패하다1'
<POINTER symbol="parent" tsoff="01064559" tpos="v" />	KorLex 신셋의 의미관계: 상위노드
<POINTER symbol="child" tsoff="02691569" tpos="v" />	KorLex 신셋의 의미관계: 하위노드
<POINTER symbol="child" tsoff="02692052" tpos="v" />	KorLex 신셋의 의미관계: 하위노드
</SYN>	

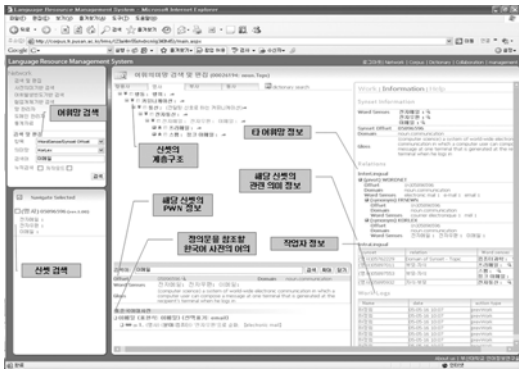

그림 7 개발자용 KorLex 구축 워크벤치

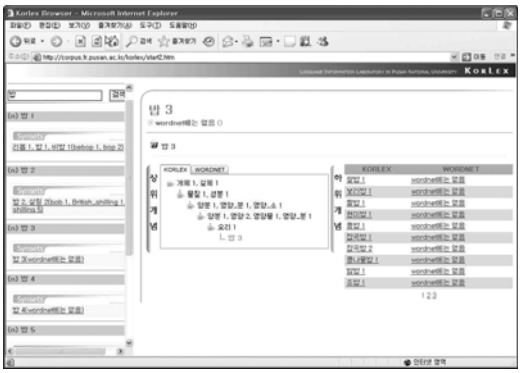

그림 8 일반 사용자용 KorLex 검색 브라우저

세상에 존재하는 또는 존재한다고 믿는 사물, 생명체, 추상체를 명명하고, 분류하고, 범주화하는 것은 절대성을 띠거나 보편화할 수 있는 것은 아니며, 자료의 양과 질에서 완결성을 기대하기는 더욱더 어렵다. 그것이 사전, 시소러스, 어휘의미망, 온톨로지 등 어떤 이름으로 어떤 형식으로 나타나든 시간, 공간, 분야, 문화, 목적, 개발자 등 수많은 주관성과 제약을 태생적으로 갖고 있다[37,38]. KorLex도 이와 같은 본질적인 한계에서 자유롭지 못하다. 다만 PWN을 참조 모델로 삼고 있다는 점, 가장 큰 사용 목적이 다국어 처리와의 연계성을 가진 한국어 분석과 생성이라는 점에서 살펴봤을 때, 적어도 앞으로 시급히 보완해야 할 부분은 다음과 같다.

첫째, 한국어에서 용언으로 문장을 구성하는 데 중요한 기능을 하는 형용사를 KorLexAdj 1.5로 확장해야 한다. 부사의 경우도 마찬가지다. '기특하다, 새삼스럽다, 퉁명스럽다. 나지막이, 살금살금, 기막히게' 등과 같은 형용사와 부사는 감정이나 화행을 표현하므로 자연언어처리 기반 감정 분석, 화행 분석에서 없어서는 안 될 요소이다.

둘째, KorLex 1.5 구축 단계에서 어의 확장의 1차 후보는 빈도가 높은 명사와 동사 중 KorLex 1.0에 나타나지 않는 어형이었다. 따라서 다의어 중에서 1개의 어의라도 KorLex 1.0에 등재되어 있다면, 다른 사용 빈도가 높은 어의가 누락되었더라도 KorLex 1.5에 확장되지 않을 수 있다. 이는 『표준』에서 다의어 비율이 높은 어형을 대상으로 KorLex 1.5에서 어의 분포를 비교해 봄으로써, 확장 대상 어의를 선정할 수 있다. 또한 명사의 경우, KorLex와 마찬가지로 『표준』의 정의문 등을 이용하여 어휘의미망을 구축하여 어의의 크기(grain size)가 유사한 U-Win과 교차 비교하여 상호 보완할 예정이다.

셋째, 한국어의 문장 분석과 생성에는 용언 및 서술성 명사의 논항구조와 각 논항의 선택 제약 정보가 필수적이다. KorLexVerb 1.5에는 매우 제한된 범위만 수록되어 있으나, 동사를 보완할 뿐 아니라 형용사와 서술성 명사에도 이러한 정보가 포함되어야 한다. 논항구조와 논항의 선택 제약 정보는 『세종』에 상세히 표현되어 있으나, 표 15에서도 밝혔다시피 용언의 어의 구분 등에서 『세종』과 『표준』은 어의 세분화 기준과 어의 크기에서 큰 차이가 있어 조정이 필요하며, 명사에서 LUB를 지정하기 위해서는 『세종』의 의미부류(object class)와 KorLex의 계층 구조 간 사상을 해야 한다. 후자는 KorLexClas에서 공기 명사의 LUB를 설정한 방식을 이용할 예정이다.

이 밖에도 기존의 언어자원에 수록된 '경어, 큰말/작은말, 지역방언, 등' 한국어에 존재하는 신셋 간, 어의 간 의미관계를 추가할 필요가 있다.

본 연구진은 KorLex를 이용하여 어휘중의성 해결(word sense disambiguation)과 문장 분석의 성능을 실험하여, 띄어읽기 시스템과 상용 한글 맞춤법 검사/교정기인 '바른한글'에 적용한 바 있으며, 위에서 언급한 바와 같이 정보의 보완과 함께 지속적으로 적용될 예정이다. 이 밖에도 소규모이기는 하지만 다국어 검색 기능을 강화하기 위해 검색 엔진에 적용된 예, 호텔 예약전화 음성인식을 위한 개체분류의 상위온톨로지 구성, 전문 분야의 상위 온톨로지 구현 등에 적용되고 있으며, 영-한/한-영 기계번역의 성능 개선에도 활용될 예정이다. 국외에서는 EWN 및 PWN 공식 딜러인 Memodata에서는 KorLex를 EWN과 사상하여 자사 홈페이지에서 다국어 검색 기능을 제공하고 있다[39].

KorLex는 다듬어지고 보완되어야 할 부분이 많지만, 현재 상태로도 언어와 직접적인 관련이 있는 자연언어처리, 지식공학, 음성공학, 언어학뿐 아니라 심리학, 감성공학, 뇌공학 등 사용할 수 있는 학문 분야도 광범위하고, 실용 시스템에 활용 가능성도 매우 크다. 2004년 10월 KorLex 1.0의 공개에 이어 2007년 11월 KorLex

1.5를 공개하였으며, 사용자들의 따갑지만 애정어린 피드백이 KorLex를 개선하는 데 단비가 되리라고 기대한다. KorLex는 특정한 연구비의 지원을 지속적으로 받지 않은 채 구축되어 왔으나, 이상과 같은 보완과 확장에는 좀 더 안정적인 연구지원 환경이 요구된다[40].

참 고 문 헌

[1] Ch. Fellbaum (ed.), WordNet: An Electronic Lexical Database, The MIT Press, Cambridge, 1998.
[2] PWN: http://wordnet.princeton.edu.
[3] 세계워드넷 연합: http://www.globalwordnet.org/gwa/wordnet_table.htm.
[4] 문유진, 의미론적 어휘 개념에 기반한 한국어 명사 워드넷의 설계와 초록, 서울대학교 컴퓨터공학과 박사학위 청구논문, 1996.
[5] 이창기・이근배, "의미애매성 해소를 이용한 WordNet 자동 매핑", 제12회 한글 및 한국어정보처리 학술대회 발표논문집, 2000, pp. 262-268.
[6] 임성신, 이은령, 권혁철, "한국어 워드넷 구축", 제16회 한글, 언어, 인지 학술대회 발표자료집, 2004, pp. 106-111.
[7] 이성헌, "사전편찬에 있어서의 어휘의미망의 역할과 기능", 한국어 어휘의미망 구축과 사전편찬 학술회의 자료집, 국립국어원, 2007, pp. 77-90.
[8] 홍재성, 21세기 세종계획 전자사전 개발 연구보고서 (11-1370252-000063-10), 문화관광부, 국립국어원, 2007.
[9] 최호섭 외, "대규모 우리말 어휘지능망 구축 방법", 한글, 273, 2006, pp. 125-141.
[10] 옥철영, "어휘의미망과 국어사전의 체계적 구성", 한국어 어휘의미망 구축과 사전편찬 학술회의 자료집, 국립국어원, 2007, pp. 35-53.
[11] 윤애선, "한국어 어휘의미망 구축의 현황과 과제", 한국어 어휘의미망 구축과 사전편찬 학술회의 자료집, 국립국어원, 2007, pp. 3-31.
[12] 윤애선, "국내・외 어휘의미망의 구축과 활용", 새국어생활, 17-3, 2007, pp. 5-25.
[13] 최경봉, 도원영, "한국어 동사 의미망 구축을 위한 상위 온톨로지 구성에 관한 연구", 한국어학, 28, 2005, pp. 217-244.
[14] 최기선 외, 다국어 어휘의미망(CoreNet), 3 vols, 한국과학기술원 전문용어언어공학연구센터, KAIST Press, 2005.
[15] GermaNet: http://www.sfs.uni-tuebingen.de/lsd/.
[16] J. Sowa, Knowledge Representation: Logical, Philisophical, and Computational Foundations, Brooks and Cole, 1999.
[17] 김양진, "국어 중사전의 전문어 표제어 선정에 대하여", 한국사전학, 7, 2006, pp. 191-215.
[18] M.W. Evens (ed.), Relational Models of the Lexicon, Cambridge University Press, Cambridge, 1988.
[19] S. Ikehara et al. The Semantic System, vol. 1 of Goi-Taikei, A Japanese Lexcion, Iwanami Shoten, 1997.
[20] Z. Dong, Q. Dong, HowNet and the Computation of Meaning, World Scientific, 2006.
[21] Google AdSense: http://www.google.com/adsense.
[22] E.R. Lee, A.S. Yoon, H.C. Kwon., "Exploiting Morpho-syntactic Features for Verb Sense Distinction in KorLex," ICCS 2007, Lecture Notes in Computer Science, 4488, 2007, pp. 1170-1177.
[23] 황순희, 윤애선, "의미자질을 고려한 명사어휘의미망의 구축(1)," 한국어학, 29, 2005, pp. 309-338.
[24] S.H. Hwang, A.S. Yoon, H.C. Kwon., "Semantic representation of Korean numeral classifier and its ontology building for HLT applications," Language Resources and Evaluation, 42-2, 2008, pp, 151-172.
[25] P. Vossen, EuroWordNet: A Multilinugal Database with Lexical Semantic Network, The Kluwer Academic Publishers, 1998.
[26] EuroWordNet: http://www.illc.uva.nl/EuroWordNet/.
[27] K. Pala, R. Sedláček, "Enriching WordNet with Derivational Subnets," Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, 2005, pp. 305-311.
[28] BalkaNet: http://www.ceid.uptras.gr/Balkanet/.
[29] 국립국어원, 표준국어대사전 1.0, 두산동아, 2001.
[30] 이은령, 윤애선, "표준국어대사전의 동사정보 개선을 위한 연구", 한민족어문학, 51, 2007, pp. 157-194.
[31] S. Yablonsky, A. Sukhonogov, "Semi-Automated English-Russian WordNet Construction," Proc. of the 3rd Int'l WordNet Conference, 2006, pp. 345-347.
[32] 국립국어연구원 현대 국어 사용 빈도 조사: 한국어 학습용 어휘 선정을 위한 기초 조사, 2002.
[33] 국립국어연구원 현대 국어 사용 빈도 조사2, 2005.
[34] KorLex: http://corpus.fr.pusan.ac.kr/korlex/start.htm.
[35] F. Dau, M.L. Mugnier, G. Steumme (eds.), Conceptual Structures: Common Semantics for Sharing Knowledge, Springer, 2005.
[36] A. Schalley, D. Zaefferer (eds.), Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts, Mouton de Gruyter, 2007.
[37] E. Hovy, "Methodologies for the Reliable Construction of Ontological Knowledge," LNAI, Vol.3596, 2005, pp. 91-106.
[38] S. Nirenburg, V. Raskin, Ontological Semantics, The MIT Press, 2004.
[39] Memodata: http://www.memodata.com.
[40] KorLex: http://korlex.cs.pusan.ac.kr

윤 애 선

1982년 이화여자대학교 불어불문학과 학사. 1984년 이화여자대학교 불어불문학과 석사. 1989년 (프) Paris-Sorbonne 대학교 언어학과 박사. 1992년~1993년 (미) Stanford 대학교 CSLI 방문 교수. 1987년~현재 부산대학교 불어불문학과, 인지과학협동과정 교수. 관심분야는 자연언어처리, 지식처리, 언어자원 표준화

황 순 희

1986년 이화여자대학교 불어불문학과 학사. 1988년 (프) Rouen 대학교 언어학과 석사. 1993년 (프) Paris 8대학교 언어학과 박사. 2006년~2008년 부산대학교 U-Port IT 산학공동사업단, 전임연구원 2008년~현재 부산대학교 인문학연구소, 연구교수. 관심분야는 전산어휘의미론, 온톨로지

이 은 령

1991년 부산대학교 불어불문학과 학사 1992년 프랑스 Paris 7대학 언어학 석사 2004년 프랑스 국립고등사회과학원 언어학 박사. 2004년~2007년 부산대학교 언어정보연구실 선임연구원. 2008년~현재 부산대학교 인문학연구소 HK연구교수 관심분야는 지식처리, 언어자원구축

권 혁 철

1982년 서울대학교 컴퓨터 공학과 학사 1984년 서울대학교 컴퓨터 공학과 석사 1987년 서울대학교 공학과 박사. 1992년~1993년 (미) Stanford 대학교 CSLI 방문 교수. 1987년~현재 부산대학교 정보컴퓨터공학부, 인지과학협동과정 교수 관심분야는 인간언어공학, 정보검색, 인공지능