2008년도 제20회 한글 및 한국어 정보처리 학술대회 논문집

한국어 어휘의미망 KorLex 1.5의 구축방법론과 정보구조

윤애선○, 권혁철*, 이은령**, 황순희**
부산대학교 {불어불문학과○, 정보컴퓨터 공학부*}/인지과학협동과정, 인문학연구소**
{asyoon○, hckwon*, eunryounglee**, soonheehwang**}@pusan.ac.kr

Methodologies for Constructing KorLex 1.5 (a Korean WordNet) and its Semantic Structure

Aesun Yoon○, Hyuk-Chul Kwon*, Eun-Ryoung Lee**, Soon-Hee Hwang**
{Dept. of French○, Dept of Computer Science*}/Interdisciplinary Program for Cognitive Science,
Humanities Institute**, Pusan National University

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 참조모델로 구축되어 다국어처리의 기반을 제공할 뿐 아니라, 시맨틱 웹 이후 더욱 주목 받고 다양한 방식으로 활용되고 있다. 본고는 PWN을 참조 모델로 2004년부터 2007년까지 구축한 한국어 어휘의미망 KorLex 1.5를 소개하는 데 있다. 현재 KorLex은 명사, 동사, 형용사, 부사 및 분류사로 구성되며, 약 13만 개의 신셋과 약 15만 개의 어의를 포함하고 있다.

1. 서 론

밀러(G. Miller)의 워드넷(WordNet, 이하 PWN) 이후, 전세계적으로 다양한 어휘의미망(Lexical Semantic Network), 개념망, 시소러스, 온톨로지 등이 구축되어 오고 있다[1, 2, 3]. 한국어를 대상으로 구축하기 시작한 것은 90년대 중반부터인데[4], 이중 PWN을 참조한 것은 '한국어 시소러스'와 'KorLex (Korean Lexico-semantic Network)'이다([표1] 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 한국어에 고유한 어휘의미구조를 표상하고자 하며, 앞으로 지속적으로 확장될 예정이다.

본고의 목적은 2007년 11월에 발표된 KorLex 1.5를 소개하는 데 있다. 2장에서는 장점과 한계를 중심으로 PWN를 간략하게 소개하고, 3장에서는 구축 방법론 및 정보구조를 중심으로 KorLex 1.5의 특성을 설명하며, 4장에서는 향후 연구 및 개발 방향을 제시한다.

2. PWN의 장점과 한계

어휘의미 간의 관계를 표상하려는 PWN의 구축 대상은 영어 내용어(content words)였고 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), PWN 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 2.0이며, KorLex도 기본적으로 PWN 2.0을 모델로 삼는다. [표2]는 PWN 2.0, 2.1, 3.0의 자료 크기를 보여준다[2].

장점과 한계를 통해 PWN의 특성을 알아보자.

첫째, '개념=어휘의미'이라고 정의함으로써 PWN이 언어보편성을 갖기에는 개념의 크기가 지나치게 작고. 어휘와 개념 간의 구분이 명확하게 이루어지지 않는다는 비판을 받는다[15]. 하지만 개념을 메타언어로 새롭게 명명해야 하는 부담을 덜 수 있을 뿐 아니라, 영어로 기술된 텍스트에서 좀더 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 양의 언어정보를 담고 있다[16]. 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]. 테니스는 운동경기 하위노드에 위치할 뿐 테니스와 관계된 테니스 공, 테니스화, 테니스 선수, 라켓 등과 상호 연결정보가 없다는 것이다. 이 단점을 보완하기 위해서, PWN 정의문을 의미태깅하여 서로 연결함으로써 동일 어휘/개념장 내부의 관련 정보를 나타낼 수 있는 자질을 망 구조로 표시하는 방식이 제안되었고[17], 지속적으로 다른 어휘망/개념망/온톨로지의 정보를 PWN에 연동하는 시도가 있어 왔다[2]. 방사형 핵구조로 개발된 PWN 형용사의 경우, 반의어를 연상하는 심리적 실재를 반영하지만 그 결과를 바로 언어/지식처리에 활용하기가 쉽지 않다. 따라서 활용의 편의

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 데이터베이스 [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

[표2] PWN 버전별 구축 크기

버전	발표 연도	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

성을 위해 PWN을 참조한 독일어 어휘의미망(GermaNet)의 경우, 형용사의 구조를 계층적으로 재구성하였다[18].

넷째, PWN의 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. 따라서 PWN의 반의 관계에 대한 보완 연구와 정제가 필요하다. 또한 형용사의 참조(see also)와 동사군(verb group)은 좀더 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 계층구조와 방사형 핵구조에 연결되지 않은 단독 신셋(orphan node)의 정체성과 그 수의 적정성도 추후 논의할 대상이다.

다섯째, PWN은 영미 및 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 풍요롭지 않다.

일곱째, PWN은 1.5가 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 힘들다. 그 예로 일본어와 중국어를 대상으로 한 NTT어휘대계[19]나 HowNet[20]의 경우 자료의 크기는 PWN과 견줄 만하여, 상호 사상(mapping)을 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/개념망을 언어/지식처리에 활용하는 논의를 지속적으로 활발하게 벌이고 있다[22, 23, 24, 25].

3. KorLex의 특성

이상과 같은 특성을 가진 PWN을 참조 모델로 하여 2004년부터 2007년까지 한국어 어휘의미망 KorLex이 구축되었으며, 현재도 소규모로 진행 중이다. 현재 KorLex은 명사, 동사, 형용사, 부사 및 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.

본고에서는 KorLex 구축 방법론과 구축 결과의 개괄적인 소개를 하며, 각 품사별로 나타나는 어휘의미론적인 문제는 앞선 논문에서 좀더 상세히 다루었고[11, 12, 26, 27, 28], 앞으로도 더 발표할 예정이다. 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 있으나, 주관 구축기관인 국립국어원이 개선과 확장을 추진하고 있으므로, 앞으로 KorLex와 지속적인 상호보완 가능성이 가장 높다. 이에, KorLex를 구축하면서 의미세분화와 관련된 『표준』의 문제점을 검토하고, 부분적으로 그 해결 방식을 제안하고 있다[30].

3.1. KorLex1.0의 대역형 참조구축

대역형 참조구축의 장점은 구축 시간과 비용은 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 [illegible] 서 KorLex 1.0의 대역형 참조구축은 반자동으로 이루어졌다. KorLex 워크벤치에서 영-한 사전을 이용하여 전처리된 대역어 후보를 제공하면, 첫 단계로 10명의 어휘전문가 또는 해당전문 분야 전공자에 의해 대역어 선정이 이루어지고, 다음 단계에서 2명의 의미론 전공 박사가 검증하였다.

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

① PWN 2.0의 신셋, 신셋 간 계층구조 및 방사형 핵상구조는 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
 ⓑ 어형이 단의어로 쓰이며, 신셋이 2개 이상의 어의로 구성된 경우
 ⓒ 신셋이 1개의 어의로 구성되며, 해당 어의의 어형이 다의어로 쓰이는 경우
 ⓓ 신셋이 2개 이상의 어의로 구성되며, 해당 어의의 어형이 다의어로 쓰이는 경우
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 어휘공백(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 전문분야 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 분류에 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 어형는 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. 동의관계의 설정은 EWN에서 제시한 조건을 사용하였다[23].
 ⓐ 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[23].
 (KorLex 신셋n에 속하는) 어의n는 문맥n-1에서 (KorLex 신셋n-1에 속하는) 어의n-1를 내포해야 하며, 역은 성립하지 않는다. 이때 n은 KorLex의 계층을 나타낸다.(
⑧ 동일한 어형의 서로 다른 어의가 KorLex에서 상하위 또는 자매 관계가 성립하는 [illegible] 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 [표3]과 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. [표3]의 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 1.0 의 크기는 PWN을 참조모델로 대역한 다른 언어 어휘의미망의 크기와 유사하다[8, 10, 14, 22]

3.2 KorLex1.5의 확장형 참조구축 방식

KorLex 1.5의 구축은 어휘의미 추가확장과 계층구조 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 5회 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 사용한 어의나, '애시당초'처럼 오류어도 포함하지 않았다.

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

KorLexNoun 1.5 경우를 예로 들어 보자. [표4]의 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 새로운운 신셋을 생성한다. 예(3)-(5)처럼 명시적인 동의어(=), 유의어(≒)나, 정의문에 등가 표현이 제시되는 경우 그 어의에 해당되

[표3] KorLex 1.0 (대역형 참조 구축) 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	

[표4] 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.						

[표5] KorLex 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	

[표6] 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

는 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 [표5]와 같다. KorLexNoun/Verb 1.5에는 한국어로 대역되지 않은 7,316개(명사)와 102개(동사)의 신셋이 존재한다. 명사는 데이터베이스를 등록한 2007년 7월 이후에도 소규모로 확장을 계속하고 있다. 계층별 신셋 수를 참조모델이 된 PWN 2.0과 비교한 결과는 [표6]과 같다. 명사는 4단계-8단계가 주로 확장된 것을 볼 수 있는데, 이는 매우 추상적이고 광범위한 개념이 주로 명사의 1-3단계에 주로 분포하는 반면, 의미의 크기가 작고 구체적인 어의가 분포하는 층위이기 때문이다. 이에 비해 PWN에서부터 얕고 넓은 분포를 가진 동사는 2-5단계가 확장된 것도 언어의 실제 모습과 일치한다[10, 14].

3.3 KorLexClas 1.0의 직접 구축 방식

한국어에 발달한 분류사의 기능은 사물이나 사건을 범주화하고, 수량화하는 것으로 어휘의미망의 '분류'와 '개념화'라는 본질적인 특성을 갖고 있다. 이때 사물이나 사건은 분류사의 공기관계(cooccurrence)로 표상되므로,분류사와 명사 간 비교적강력한 공기 제약을 갖는다. KorLexClas 1.0은 한국어 언어자원을 이용한 직접구축 방식으로 개발하였다[24].

1단계로, 고빈도 분류사의 완전한 목록을 구성하고 공기명사

[표7] KorLexClas 1.0 (2008년 6월)

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

[표8] 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

정보가 함께 태깅된 자료의 확보하기 위해, 선행 언어학 연구, 『표준』의 정의문, 대용량 말뭉치의 문맥 정보를 이용하여 분류사 및 공기명사 목록을 수집한다. 2단계로, 분류사의 의미적 특성을 고려하여 ⓐ 도량성(mensural), ⓑ 개체성(sortal), ⓒ 중립성(neutral), ⓓ 사건성(event)으로 하위범주화하여 각각 분류사를 정의하고, 분류사의 의미자질 간 계층관계를 설정한다. 3단계로, 분류사-공기명사 간 선택제약 관계를 설정하기 위해 KorLexNoun과의 연동한다. [표7]과 같은 KorLexClas는 앞서 기술한 다른 품사 어휘망과는 달리 분류사 자체가 아니라 분류사를 구성하는 의미자질을 계층화하였다.

이상과 같이 구축된 KorLexNoun/Verb 1.5 및 KorLex Adj/Adv/Clas 1.0의 크기는 [표8]과 같다. PWN 2.0과 비교해 보았을 때 동사에서 가장 큰 차이를 보이는데, 영어에서는 중립동사 등으로 나타나는 단일 어형의 다의어가 한국어에서는 선어말 어미의 유무로 어형을 구분할 수 있는 단의어로 대역되었기 때문이다[26]. 한자어 어근을 많이 사용하는 한국어의 동형이의어 및 다의어 비율이 높다는 특성과는 달리 KorLex의 다의어 비율이 PWN과 유사하게 나타나는 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.

3.4. KorLex 1.5의 의미정보

기본적으로 KorLex의 신셋은 사상되는 PWN의 신셋이 가진 의미정보를 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 표상할 지에 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 것이지만 기존의 어떤 언어자원에서도 KorLex에 직접 사용할 수 있는 이러한 정보를 찾기 힘들다. KorLex이 어의 구분의 기준으로 삼은 『표준』은 [표9]에서 볼 수 있듯 격틀정보가 기술되지 않거나(타다1-1 ~ 타다1-5) 일부의 격틀 정보만이 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 KorLex은 『세종』, 『표준』, 등을 참조하되, 신셋을 구성하는 어의 단위로 격틀정보를 추가 기술하고, 논항의 선택제약(selectional restriction)은 KorLexNoun과 연동한다.

넷째, KorLex은 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 계층구조를 설정하고, 각 분류사와 공기하는 명사 정보로 KorLexNoun를 연동한다.

논항의 선택제약과 분류사의 공기명사를 연동하는 방식은 KorLexNoun에서 최하위 공통상위노드(Least Upper Bound Node, 이하 LUB)를 찾는 것이다. 예를 들어 '[N0]-이 [N1]-을 깁다'의 경우, [N1]에 '옷, 양말, 신' 등의 선택제약이 가해진다면, 이들의 공통상위 노드 중 최소공배수 격인 {피복류} 노드를 지정하고, 그 하위노드 전체에 제약규칙이 적용되는 것이다.

이상과 같은 KorLex의 모든 의미 및 관계정보는 XML로 정의되며, 다음과 같은 4개 테이블의 관계형 DB로 설계하였다. ① 신셋 정보 테이블 (PWN과 KorLex의 신셋 정보), ② 어의 정보 테이블 (어의 별 정보), ③ 신셋-어의 연관 정보 테이블 (각 신셋과 그 구성요소인 어의 간 관계 규정), ④ 신셋 및 어의 간 의미관계 정보 테이블이다. 또한 검색 인터페이스를 일반에게 공개하고 있다[34].

4. 활용 및 향후 개발 방향

본고에서는 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 같은 지에 대해 의심을 품었다. 후자는 자료 자체의 크기, 표상 방식의 체계성, 절차적 수행의 수월성에 주목하였다[35,36].

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 1.5단계로 확장해야 한다. 부사의 경우도 마찬가지다. 형용사와 부사는 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 KorLex1.0에 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.

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

이밖에도 기존의 언어자원에 수록된 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. 이밖에도 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 높다. 2004년 10월 KorLex 1.0의 공개에 이어 2007년 11월 KorLex 1.5를 공개하였으며, 일반사용자를 위해 검색 서비스(http://corpus.fr.pusan.ac.kr/korlex/)를 제공하고 있다. 다른 대규모 어휘의미망이 국가지원 연구비로 이루어진 데 비해 KorLex는 인간언어공학 벤처기업인 (주)나라인포테크에서 대부분의 구축비용을 지원하였고 확장과 보완의 일부는 한국과학재단의 지원으로 이루어졌다. 사용자들의 따갑지만 애정어린 피드백이 KorLex를 개선하고 지속적으로 확장하는 데 단비가 되리라고 기대한다.

참고 문헌

[1] Fellbaum, Ch.(ed.), WordNet: An Electronic Lexical Database, The MIT Press, Cambridge, 1998.
[2] http://wordnet.princeton.edu (PWN)
[3] http://www.globalwordnet.org/gwa/wordnet_table.htm (세계 워드넷 연합)
[4] 문유진, 의미론적 어휘 개념에 기반한 한국어 명사 워드넷의 설계와 초록, 서울대학교 컴퓨터공학과 박사학위 청구논문, 1996.
[5] [illegible](2000), "의미중의성 해소를 이용한 WordNet자동 매핑," 제 12회 한글 및 한국어정보처리 학술대회 발표논문집, pp.262-268.
[6] 임성신, 이은령, 권혁철, "한국어 워드넷 구축", 제16회 한글, 언어, 인지 학술대회 발표자료집, pp. 106-111. 2004.
[7] 이성헌, "사전편찬에 있어서의 어휘의미망의 역할과 기능", 한국어 어휘의미망 구축과 사전편찬 학술회의 자료집, 국립국어원, pp.77-90, 2007.
[8] 홍재성, 21세기 세종계획 전자사전 개발 연구보고서 (11-1370252-000063-10), 문화관광부, 국립국어원, 2007.
[9] 최호섭 외, "대규모 우리말 어휘지능망 구축 방법", 한글, 273, pp.125-141, 2006.
[10] 옥철영, "어휘의미망과 국어사전의 체계적 구성", 한국어 어휘의미망 구축과 사전편찬 학술회의 자료집, 국립국어원, pp. 35-53, 2007.
[11] 윤애선, "한국어 어휘의미망 구축의 현황과 과제", 한국어 어휘의미망 구축과 사전편찬 학술회의 자료집, 국립국어원, pp. 3-31, 2007.
[12] 윤애선, "국내・외 어휘의미망의 구축과 활용", 새국어생활, 17-3, pp. 5-25, 2007.
[13] 최경봉, 도원영, "한국어2] 동사 의미망 구축을 위한 상위 온톨로지 구성에 관한 연구", 한국어학, 28, pp. 217-244, 2005.
[14] 최기선 외, 다국어 어휘의미망(CoreNet), 3 vols, 한국과학기술원 전문용어언어공학연구센터, KAIST Press, 2005.
[15] Sowa, J., Knowledge Representation: Logical, Philisophical, and Computational Foundations, Brooks and Cole, 1999.
[16] 김양진, "국어 중사전의 전문어 표제어 선정에 대하여", 한국사전학, 7, pp.191-215.
[17] Evens, M.W.(ed.), Relational Models of the Lexicon, Cambridge University Press, Cambridge, 1988.
[18] http://www.sfs.uni-tuebingen.de/lsd/ (GermaNet)
[19] Ikehara, S. et al. The Semantic System, vol. 1 of Goi-Taikei, A Japanese Lexcion, Iwanami Shoten, 1997.
[20] Dong, Z. and Q. Dong, HowNet and the Computation of Meaning, World Scientific, 2006.
[21] www.google.com/adsense (구글 애드센스)
[22] Vossen, P., EuroWordNet: A Multilinugal Database with Lexical Semantic Network, The Kluwer Academic Publishers, 1998.
[23] http://www.illc.uva.nl/EuroWordNet/ (EuroWordNet)
[24] Pala, K. and R. Sedláček, "Enriching WordNet with Derivational Subnets", Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, pp. 305-311, 2005.
[25] http://www.ceid.uptras.gr/Balkanet/ (BalkaNet)
[26] Lee, E.R, et al., "Exploiting Morpho-syntactic Features for Verb Sense Distinction in KorLex", ICCS 2007, Lecture Notes in Computer Science, 4488, pp. 1170-1177, 2007.
[27] 황순희, 윤애선, "의미자질을 고려한 명사어휘의미망의 구축(1)", 한국어학, 29, pp. 309~338, 2005.
[28] Hwang, S.H. et al., "Semantic representation of Korean numeral classifier and its ontology building for HLT applications", Language Resources and Evaluation, 42-2, pp, 151-172, 2008.
[29] 국립국어원, 표준국어대사전 1.0, 두산동아, 2001.
[30] 이은령, 윤애선, "표준국어대사전의 동사정보 개선을 위한 연구", 한민족어문학, 51, pp. 157-194, 2007.
[31] Yablonsky, S. and A. Sukhonogov, "Semi-Automated English-Russian WordNet Construction", Proc. of the 3rd Int'l WordNet Conference, pp. 345-347, 2006.
[32] 국립국어연구원 현대 국어 사용 빈도 조사: 한국어 학습용 어휘 선정을 위한 기초 조사, 2002.
[33] 국립국어연구원 현대 국어 사용 빈도 조사2, 2005.
[34] http://corpus.fr.pusan.ac.kr/korlex/start.htm (KorLex)
[35] Dau, F. et al. (eds.), Conceptual Structures: Common Semantics for Sharing Knowledge, Springer, 2005.
[36] Schalley, A. and D. Zaefferer (eds.), Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts, Mouton de Gruyter, 2007.
[37] Hovy, E., "Methodologies for the Reliable Construction of Ontological Knowledge", LNAI, vol.3596, pp. 91-106, 2005.
[38] Nirenburg, S. and V. Raskin, Ontological Semantics, The MIT Press, 2004.
[39] http://www.memodata.com (Memodata)

*본 고에서 기술한 일부 구축 내용은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2007-000-20517-0).