국내·외 어휘 의미망의 구축과 활용

윤애선 • 부산대학교 불어불문학과/인지과학협동과정 교수

1. 들어가는 말

‘먹을 수 있는 것’, ‘안전한 곳’, ‘피해야 할 것’, ‘약한 것’, ‘비가 오는 때’ 등, 인지 능력을 가진 생물체라면 자신의 생존을 위해 외부 세계를 분류하고 이를 지식화한다. 지구상 생물체 중 가장 발달된 지능과 인지능력을 가진 것으로 평가받는 ‘인간’의 지식은 어떻게 구축되고 어떤 방식으로 사용될까? 기원이 오래된 철학적 질문이었고, 20세기 후반에 조명을 다시 받게 된 논제다. 지식의 거처는 ‘뇌’지만, 뇌의 활동을 분석할 수 있는 기기나 방법은 이제 겨우 첫 발을 내디딘 정도다. 그렇다면 블랙박스 같은 뇌에서 형성되는 지식을 들여다 볼 수 있는 창은 없을까?

언어, 그중에서도 어휘(또는 단어)를 ‘아리안느의 실’로 삼아 그 미지의 동굴을 탐험하려는 노력이 사전(dictionary), 분류체계(taxonomy), 시소러스(thesaurus), 어휘 의미망(lexical semantic network), 온톨로지(ontology)라는 다양한 이름으로 시도되어 왔다.[1] 이 중 어휘 의미망은 미국 프린스

1) 이들 용어가 같은 시기, 같은 분야에 사용되지 않았다는 점에서 단순 비교를 할 수 없다.

톤(Princeton) 대학교의 인지심리학자인 밀러(G. Miller)[2]가 주창하여 언어학자 및 전산학자와 함께 1985년부터 지금까지 구축하고 있는 워드넷(WordNet, 이하 PWN)을 지칭하는 용어다. '어휘가 심리학적 실재를 가진 기억의 최소 단위'라는 자신의 이론을 바탕으로 이전에 존재했던 사전, 분류체계, 시소러스의 특성을 수용하였고, 2000년대 전후로 활발해진 전산학 분야의 온톨로지 구축에 중요한 모형을 제공하였다는 점에서, 이들 용어의 교차점에 놓여있다.

최근 어휘 의미망과 온톨로지에 대해 급증하는 관심은 시맨틱웹(Semantic-web)으로 대변되는 의미기반 웹서비스와 지식 처리, 이 두 가지와 밀접히 관련된다. 인간에게 가장 유용한 기계는 인간의 말을 이해하고, 적절히 반응하는 '똑똑한 기계'다. 이런 기계를 만들기 위해서는 인간의 언어 및 지식 표상 모형의 명세화가 선행되어야 하는데, 기왕에 만들어진 어휘 의미망이 중요한 역할을 할 수 있으리라고 기대하면서, 다양한 응용 분야에서 파생과 활용을 시도하고 있다. 국내에서도 10여 년 전부터 연구를 시작하여, 각기 다른 방식으로 구축된 여러 어휘 의미망이 존재한다.

본고에서는 국내·외 어휘 의미망의 개발 배경, 구축 방식, 구조와 활용 현황을 살펴보려고 한다. 2장에서는 일반론으로 어휘 의미망을 포함한 지식표상 체계의 종류와 구축 방법론을 소개하고 3장과 4장에서는 각각 국외와 국내의 대표적인 어휘 의미망에 초점을 맞춰 그 특성을 알아볼 것이다. 그중에서도 인간의 상식이나 넓은 범위의 배경 지식을 구성하는 데 필요한 일반 목적용이며, 실제 자연 언어 처리를 통해 지식에 접근할 수

물론 용어 자체가 다른 만큼 기본 단위의 특성과 크기, 담고 있는 정보의 형식과 구조, 사용 목적 등에서 차이점을 찾을 수 있지만, 모두 '어휘화된 개념'이나 '어휘'를 단초로 인간 지식을 표상하려는 큰 공통점을 갖고 있다. 이들의 비교는 윤애선(2007:8-11)을 참조할 것.

2) 1956년에 발표한 유명한 논문 "The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information"에서 인간 기억의 처리 능력을 제시한 밀러는 1960-1970년대 연구를 통해 언어학 이론의 심리학적 실재를 논증하면서 어휘가 언어처리의 기본 단위임을 주장한다. PWN은 바로 영어 어휘로 구성된 심상 어휘집(mental lexicon)의 구조와 작동 메커니즘을 재현하기 위해 지난 20여 년간 약 300만 달러가 투입된 방대한 작업이다.

있을 정도로 충분히 크고 상세한 정보를 포함하는 어휘 의미망을 중심으로 검토할 것이다.[3)]

2. 지식 표상 체계의 종류와 구축 방식

2.1 지식 표상 체계의 종류

지식은 다양한 방식으로 존재하며 끊임없이 변화하므로, 그 표상 체계의 수는 실로 무한하다. 지식공학자 하비(Hovy, 2005)는 구축자의 전공, 크기, 목적과 지식의 범위, 언어 의존성, 구축 방식에 따라 다음 5가지 종류로 구분한다.

첫째, 아리스토텔레스로부터 시작한 전통적인 철학 영역의 온톨로지에서는, 표상할 수 있는 최소의 의미 자질 집합으로 개념을 정의하고, 논리적으로 의미 자질을 결합하여 개념 간의 계층적 위계 구조를 구축한다. 이 구조에 새로운 개념을 추가하려면 기존의 것과 다른 새로운 차이점(differentiae)을 명시하면 된다. 이러한 지식 표상 체계는 언어 독립적이라는 장점이 있으나, 온톨로지 최상부(upper model)에 위치한 소수(500개 이하)의 추상적 개념을 정의하는 데만 유용하다는 제한점을 갖는다.

둘째, 인지심리학에서는 일반 사람에게 친숙한 직관에 주목한다. 가장 오래되고 자연적인 방법으로 사람들은 새로운 개념을 정의할 때, 기존에 정의된 어떤 개념과도 같지 않다는 직관적 느낌을 가지면, 그 개념을 분리하여 새로운 개념으로 창조해 낸다. 한 언어를 대상으로 실제로 사람들이 사용함직한 대규모(약 10만 개 내외) 심상 어휘집을 재현하는 PWN, NTT 일본어 어휘대계(이하 NTT LN), 중국어 하우넷(HowNet) 등이 이 부류에 속한다. 그러나 사람들은 일관성 있게 개념의 분리 작업을 수행하지 못하고, 자신의 관심사, 배경 지식, 업무 혹은 기타 요인에 따라 분리

3) 이 글에서 소개하는 어휘 의미망의 특성은 각 구축자의 발표 자료 및 배포 홈페이지에 기댔다. 이러한 1차 자료를 매번 인용하는 대신 참고문헌에 어휘 의미망 별로 정리하는 형식을 취한다.

의 기준이 임의적으로 바뀐다. PWN 구축에서도 개념 구분에 적용할 수 있는 엄밀한 준거 부재가 항상 비판받아 왔으나, 이를 극복하기 위해 인지과학자들은 사람들이 유사 개념을 어떻게 구별하는지에 대한 실험을 통해 지식 표상 체계 구축 방법론을 증명하려고 노력한다.

셋째, 동일한 개념이 여러 언어에서 표현될 수 있으면, 이를 기반으로 다국어 지식 표상 체계를 구축할 수 있다. 이는 기계 번역과 같은 다국어 처리에 유용한 언어 자원으로 사용될 수 있어, 전산언어학의 많은 주목을 받아 왔다. 특히 PWN 1.5를 참조하되, 다국어 중계 인덱스(Inter-Lingual Index)라는 공통 개념 층위를 새로 설정하여 8개 유럽어 간 교차 언어적 연계성을 확보하고 이를 6개 동유럽어에 적용한 유로워드넷(Euro WoerNet, 이하 EWN)이나 발카넷(BalkaNet, 이하 BWN)[4]이 이 부류에 속한다. 이 방법은 어휘로는 존재하지 않는 개념이 많고, 언어 간 개념-어휘쌍의 분포는 상상을 초월할 정도로 다양하다는 한계에도 불구하고, 의미를 표상하는 가장 풍부하고 중심적인 도구가 어휘라는 점에서 기계 번역 등 자연 언어 처리 분야에서 활용 가능성이 높다. 이에, 교차 언어적 개념 관계에 기반한 온톨로지 구축 연구가 활발하다.

넷째, 지식공학 분야에서 추론에 사용할 지식 표상 체계를 구축하려면, 데이터 모델을 만들어 한 분야에서 유사하게 사용되는 항목들을 그룹화하고 유형화하여 동일하게 처리할 필요가 있다. 이러한 전산학 영역의 온톨로지는 해당 분야의 메타데이터나 시스템 변수를 반영하여 구축되며, 그 결과에 대한 검증도 단순하고, 정확도도 높다. 그러나 시스템 구현을 위해 구축된 온톨로지는 형식논리적 엄밀성은 매우 뛰어나나, 철학, 인지심리학, 언어학 등에서 주목했던 실제 세계의 불연속적이고 중의적인 정보나 개념의 본래적 속성을 외면하는 경우가 많다.

다섯째, 전문 분야에서는 실질적 필요에 의해, 개념을 구분하고 구조화하여 온톨로지를 구축한다. 따라서 해당 분야의 관점을 잘 반영하고 있으며 활용도도 높지만, 타 분야에서 사용하기 어렵다. 의료분야의 메드(MED)가 대표적이며 자동자 엔진 디자인, 비행기 예약, 포도주 등 실용성이 높고, 극히 세분화된 분야에서 급속도로 구축되고 있다. 그 예로, OntoSelect

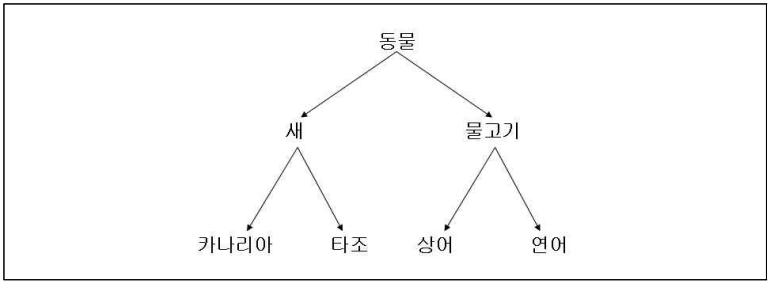
4) BWN은 유로워드넷 프로젝트 2단계로 기획되었다.

(http://views.dfki.de/ontologies)나 DAML(http://www.daml.org/ontologies/keyword.html) 웹페이지를 방문하면 수백 개가 넘는 전문분야 온톨로지의 목록을 볼 수 있고, 하루가 다르게 증가하고 있다.

2.2 지식 표상 체계의 구축 방식

이상과 같은 지식 표상 체계를 구축하는 방식은 크게 개발자의 직관이나 전문 지식, 다른 형태의 언어 자원을 이용하여 직접 만드는 방식과 기존 체계에 기대어 비교적 단기간에 구축하는 참조 방식으로 구분할 수 있다. 참조 방식에 비해 상대적으로 장기간 개발을 필요로 하는 직접 방식은 다시 하향식과 상향식으로 나눌 수 있다.[5)]

첫째, PWN을 처음 개발할 때 사용한 하향식 구축 방식은 개발자의 직관이나 전문 지식을 이용하여 상위 개념을 더 작은 하위개념으로 분화함으로써 세분화된 계층 관계의 설정이 가능하다는 장점이 있고, 소규모 전문 분야의 지식 표상 체계를 구축하는 데 적합하다. 하지만 구축 대상이 큰 경우, 구축 기간이 오래 걸리며 구축 비용이 많이 들면서 동시에 일관성이 떨어진다는 단점이 있다. 또한, 전문가에 의해 개념이나 의미 간 계층 관계가 설정된다고 해도 참여한 전문가가 가진 전문 지식과 학설의 차



<그림 1> 하향식 구축 방식

5) 실제 대규모 지식 표상 체계를 개발할 때는 이중 어느 한 방식만을 고집하지 않고, 주된 방식은 선택하되 그것의 단점을 다른 방식의 장점으로 보완하는 방식의 통합적(hybrid) 구축이 이루어진다.

<그림 2> 사전정의문을 이용한 상위어 추출의 예

이로 인해 구축 결과가 주관적이거나 비일관적일 수 있으며 검증이 어렵다.

둘째, 상향식 구축 방식에서는 기존 언어 자원이나 지식에 기대어 과정과 결과물의 객관성과 일관성을 유지하려고 한다. 예를 들어, 사전 정의문에 사용된 중심어(headword)를 추출하여 <그림 2>처럼 연쇄적으로 어휘 의미의 상·하위 관계를 설정하는 방법이다. 사전 정의문이 통제된 경우, 상·하위 관계를 추출하는 작업이 효과적이고 효율적으로 진행되며, 하향식보다 구축 시간이 짧게 걸린다. 그러나 일반적으로 인간을 위한 사전은 정의문이 통제되어 있지 않고, 순환적 정의가 많고, 중심어를 추출하기가 용이하지 않다.

셋째, 참조 구축은 이미 만들어진 지식 표상 체계를 이용하여 새로운 시스템을 간접적으로 개발하는 방식이다. 예를 들어, 한 언어로 구축된 어휘 의미망의 최소 단위를 대역하여 사상(mapping)하거나, 대역어 특성을 고려하여 새로운 어휘 의미망으로 정제할 수 있다. 2.1절의 세 번째 부류에 해당하는 이 유형은 비교적 짧은 시간에 적은 수의 분야 전문가로 어휘 의미망이 구축 가능하고, 개념/용어 간 계층 구조의 정합성에 대한 논란이 적다. 또한, 적절한 대응 용어를 선택할 때, 기구축 개념망과의 비교를 통해 의미범주와 맞지 않는 상위어 구조, 의미 세분화가 필요한 구조를 재분석하게 되므로, 계층 구조 및 의미 관계의 일관성 유지, 기구축 어휘 의미망에 결여된 의미 관계 등을 보완하여 구축할 수 있다는 장점이

있다. 반면, 참조한 온톨로지에 경도될 위험이 있으므로 개별어 특성을 고려하면서 동시에 일관된 기준을 가지고 참조 구축된 어휘 의미망을 정제하는 연구가 수반되어야 한다.

참조 구축의 다른 유형들로는 기구축된 지식 표상 체계의 일부를 이용하여 상위 온톨로지(upper ontology)로 삼고 하위 구조를 구축하거나, 전문 분야 온톨로지의 초기 틀로 삼아 촘촘한 그물망을 엮는 방식 등을 들 수 있다. 또한 상이한 방식으로 구축된 지식 표상 체계를 사상하여 개념 중립성을 찾아내려는 실험이 다양하게 시도되었다.

3. 국외 어휘 의미망의 구축과 활용

반세기 남짓한 자연 언어 처리 역사에서 개념망과 어휘 의미망이 처음 만들어지기 시작하던 1980년대 중반은 1960~1970년대에 불어 닥친 기계 번역의 열풍과 역풍이 모두 휩쓸고 지나간 후 학문적 반성의 시기를 거치며, 자연 언어 처리에서 의미의 중요성을 다시금 인식하고 철학, 심리학, 언어학, 전산학 등이 따로 또 같이 연구하던 시기이다. 또한 반도체 발견에 힘입어 컴퓨터 하드웨어가 비약적으로 발전하여, 연구를 위해 중대형 컴퓨터를 사용하는 것이 용이해졌을 뿐 아니라 개인용 컴퓨터도 막 출현하던 시기이다.

이 시기에 학계는 인간이 언어로 의사소통을 하려면 상당량의 배경 지식(background knowledge)을 가져야 한다는 인식을 널리 공유하면서 그 문제를 해결하고자 하였다. 본고 2장에서 기술한 바와 같이 그 지식의 문을 여는 한 가지 열쇠로 '개념'을 표현하는 단위로 '어휘'에 주목을 하였고, 그 결과로 개발되기 시작한 대표적인 국외 어휘 의미망의 특성을 <표 1>에서 볼 수 있다.[6]

6) <표 1>에 소개된 어휘 의미망이 '명사' 부류로 언어화된 개념의 관계를 표현하는데 치중하는 경향을 가졌다면, 개념 노드와 그 관계를 프레임(frame) 형식으로 표시하려는 또 하나의 흐름이 있다. 후자의 대표적인 예로 프레임넷(FrameNet)을 들 수 있으며, <표 1>의 개념망을 이용하여 프레임 형식을 기술하는 방식(Sowa, 2000, Nirenburg & Raskin 2004)도 민스키(M. Minsky)와 위노그라드(T. Winograd)의 전통을 잇는 지식공학 분야에서는

<표 1> 대표적인 국외 개념망/어휘 의미망(발췌)

구분	명칭	구축 시기	개념(n,s)/어휘(w) 수	구축언어(품사)
개념망	Mikrokosmos	1995-1997	약 5,000n	-
	SUMO	2000-2001(?)	약 1,000n	-
	CYC	1984-2001	약 6,000n	-
어휘 의미망	PWN	1984 - 현재	117,417s/155,297w	영어(명, 동, 형, 부)
	NTT LN	(?) - 1997	2,710n/400,000w	일본어(명)
	HowNet	1988 - 현재	95,690s/81,062w	중국어(명, 동, 형)
	EWN	1996 - 1999	277,068s/484,466w	서유럽어(명, 동)
	BWN	2001 - 2004	78,165s/125,604w	동유럽어(명, 동)
전문분야 온톨로지	특정 목적에 따라 만드는 세부 전문분야 온톨로지 (OntoClean, DAML 웹사이트참조)			

특정 언어에 의존하지 않고 인간이 가진 보편적 개념을 찾고 그것을 정의하고, 그것을 통해 지식의 표현에 이용하는 Mikrokosmos, SUMO, CYC[7]와 같은 개념망이나, OntoClean, DAML 웹사이트에 등록된 수백 개의 전문분야 온톨로지에 비해, 이들의 중간에 위치한 어휘 의미망은 언어 의존적이고, 자료의 크기가 커서 일부 일관성이 결여된 부분도 발견되지만, 일반인의 실세계 지식을 가장 잘 반영할 수 있어 실제 시스템에 적용할 수 있는 활용도가 높다. 이 글에서는 지면의 제한으로 가장 먼저 개발되기 시작하고 가장 많은 참조 어휘망을 파생한 PWN을 소개한다.

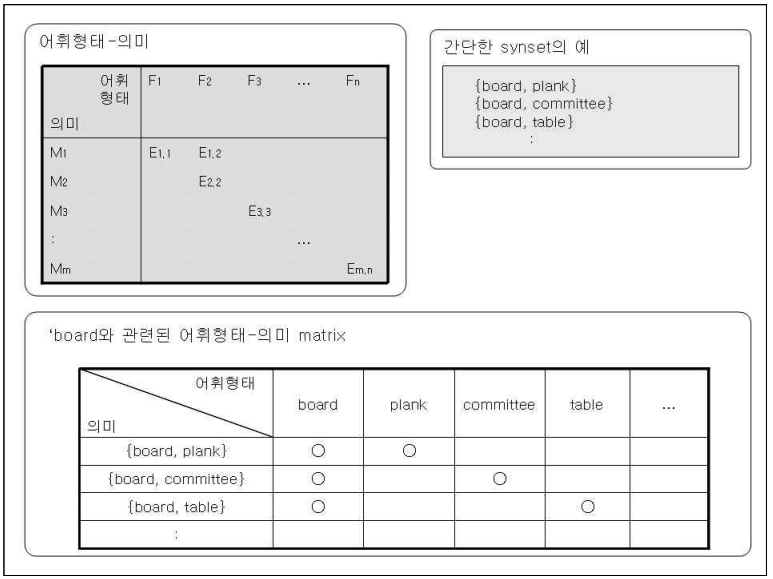
밀러의 어휘 처리 연구에 이론적 기반을 둔, PWN은 1985년에 구축 프로젝트로 만들어져, 1991년 1.0버전이 발표된 이래 꾸준히 결과물이 보완되고 갱신되어, 1995년에 1.5, 2003년에 2.0, 2006년에는 <표 2>와 같은

활발하게 진행되고 있다.

7) 개념망을 처음 제안할 때는 언어 독립적이고 보편적인 사실을 기술하기 위한 '개념'을 정의하고, 이를 이용하여 지식을 표현할 수 있으리라는 가정에서 출발하였다. <표 1>의 개념망 크기가 어휘 의미망에 비해 아주 작고, 극단적으로 Goddard & Wierzbicka(1994)가 주장하는 의미 원소(semantic primitives)의 수는 50개 미만이다. 하지만 이것이 현실에 적용되기에 충분한가 하는 점은 여전히 의문이다. CYC를 이어받은 Cycorp이 2006년에 발표된 OpenCYC의 경우 개념의 수가 47,000개, 적용의 수는 30만 개를 상회한다. 일반적으로 한 언어의 중형 사전이 약 5만 개 내외의 표제어를 가진 것으로 정의한다는 점에 비춰 볼 때, 47,000개의 개념이 과연 언어 독립적일 수 있을까?

<표 2> WordNet 3.0 구축 결과물

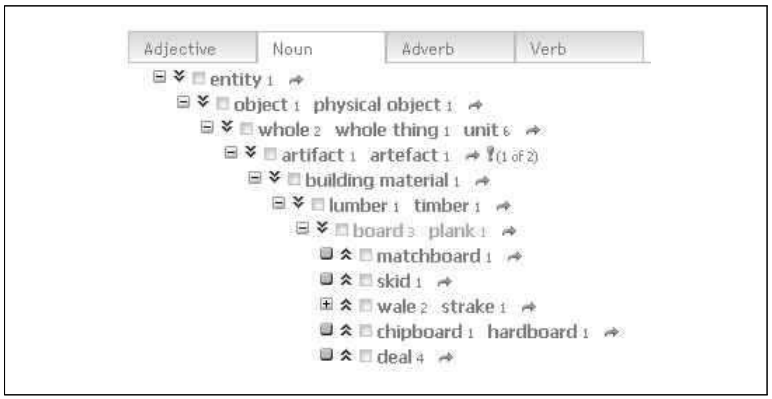
품사	분화된 어휘형태	개념(신셋)
명사	117,798	82,115
동사	11,529	13,767
형용사	21,479	18,156
부사	4,481	3,621
계	155,287	117,659



<그림 3> 어휘형태와 의미 간 多 대 多 관계

크기의 3.0버전이 배포되어 현재에 이르고 있다.

PWN을 구성하는 최소 단위는 '동일한 의미를 가지는 어휘 의미를 하나의 집합으로 묶은 신셋(synset), 즉 동의어 집합(synonym set)'이다. 예를 들어, 동일한 의미를 가지는 어휘인 'board'와 'plank'를 묶어서 {board3, plank1}으로 표현함으로써 중의성이 없는 하나의 개념을 표현한다. 동일한 어휘 형태에 붙인 번호는 다의성을 구분하며, <그림 3>과 같



<그림 4> {board3, plank1}의 계층성

이 다의성과 동의 관계를 이용하여 '어휘(형태) 대 의미가 多 대 多'인 관계를 최대한 세밀하게 표현할 수 있다.

개념 간에는 '상의/하의, 전체/부분, 반의, 속성, 원인, 함의, 참조, 유사, 관련' 관계가 유기적으로 연결되어 있다. 이중 모든 개념을 묶는 <그림 4>와 같은 상의/하의 관계는 자질 승계 체계(inheritance system)를 형성한다. 상위어는 총체적이고 보편적 의미 자질을 하위어에 물려주고, 하위어는 승계받은 의미 자질과 자신과 직속 상위어를 구별해 줄 자질을 적어도 하나 이상 추가하여 가지는 방식이다. 명사는 최대 12층위, 동사는 최대 4층위의 계층구조를 이룬다. 이와는 별도로 '동물, 행위, 신체, 인지활동, 접촉, 소유' 등 명사와 동사를 대상으로 각각 25개와 15개의 의미 범주로 분류해 놓아, 후속 연구자가 특정 분야 온톨로지를 만들 때 출발점으로 삼을 수 있다.

PWN은 명사, 동사, 형용사, 부사 등 내용어 품사를 모두 포함하지만 의미의 세분화, 계층성, 관계 설정 등 모든 면에서 명사에 치중하고 있다. 동사는 넓고 얕은 형태로 계층구조화되어 있고, 형용사는 핵심어를 중심으로 관련어를 표시하는 방사형 클러스터 구조로 되어 있으며 아직 부사는 단순한 목록으로만 제시하고 있다. 품사 간 연계성은 '분사형, 파생' 등의 최소 관계만 설정되어 있다. 또한 일부 격틀 구조가 동사에 표현되

지만 논항의 종류나 논항의 의미 구분과 같은 언어학적 정보는 정교하거나 풍요롭지 않다. 또한 '어휘의미=개념'이라고 정의함으로써, 언어 보편성을 갖기에는 개념의 크기가 지나치게 작으며, 어휘와 개념 간의 구분이 명확하게 이루어지지 않는다는 비판을 받고 있다.

그럼에도 불구하고, PWN은 다음과 같은 장점을 갖고 있어 상당기간 영향력을 발휘할 것으로 예상된다. 첫째, 어휘 의미망의 크기가 실제 자연언어 처리 시스템에 활용할 수 있을 정도로 크다. 둘째, 구성단위인 개념(어휘 의미)이 충분히 세분화되어 다른 언어의 어휘 의미와의 등가성을 설정하는 데 비교적 용이하다. 셋째, 일반 목적의 범용적 지식 표상 체계이므로 인간의 상식이나 일반적 배경 지식을 추론하는 데 유리하다. 넷째, 가장 많은 참조 방식의 어휘 의미망을 파생하였으므로 다국어 연계성을 확보하기가 쉽다.[8] 다섯째, 다른 어휘 의미망/개념망과의 사상이 가장 활발히 이루어지고 있어 활용 가능성이 크다. 실제로 구글 애드센스(AdSense)는 워드넷을 이용하여 검색된 정보에 가장 가까운 광고를 찾아 검색 결과와 같이 제공하여 수익 모델을 제시하였다. 2002년부터 격년으로 국제 학술 대회(Global WordNet Conference)를 열어 구축, 확장, 활용에 대한 논의를 활성화하고 있다.

4. 국내 어휘 의미망의 구축과 활용

국내의 자연 언어 처리 연구가 1980년대 중반에 시작한 만큼 지식 표상 체계의 구축도 1990년대 중·후반에 출발한다. 자연 언어 처리 시스템에 필요한 의미 분석을 해야 한다는 실질적인 목적을 가진 전산학자들에 의해 주도되었고, 지금도 대부분의 실용 시스템 구축을 주도하고 있다. 따라서 개념망보다는 일반 목적의 어휘 의미망이나 전문 분야 온톨로지 개발에 주력하고 있으며, 전자의 경우 <표 3>처럼 주로 참조 구축 방식을 택하고 있다.[9] 국내 최초로 시도된 어휘 의미망인 한국어 명사워드넷,[10]

8) http://www.globalwordnet.org/gwa/wordnet_table.htm에 있는 약 30개 언어의 50여 개 어휘 의미망 목록을 참조하라.

<표 3> 대표적인 국내 개념망/어휘 의미망(발췌)

명칭	중심구축 기관	중심구축자 전공분야	구축기간	참조 모델	의미/개념 vs 어휘 수	구축 품사
한국어 명사워드넷	호남대학교	전산학	1994-1995	(직접)	20,000w	명
세종전자사전	서울대학교	언어학	1998-2007	(직접)	540,000w	모든 품사
U-Win	울산대학교	전산학	2002-2007	(직접)	46,339s/ 약250,000w	모든 품사
한국어 시소러스	포항공과대학	전산학	1997-2000	PWN	18,362s/ 21,390w	명
KorLex	부산대학교	전산학/ 언어학	2004-현재	PWN	126,653s/ 143,479w	명, 동, 형, 부, 분류사
다국어 어휘 데이터베이스	고려대학교	언어학	2000-2006	EWN	5,500w	명
CoreNet	KAIST	전산학/ 언어학	1995-2004	NTT LN	2,938n/ 62,632w	명, 동, 형

국책과제로 기획된 세종전자사전, 울산대학교의 U-Win이 직접 구축 방식으로 개발되었다.[11]

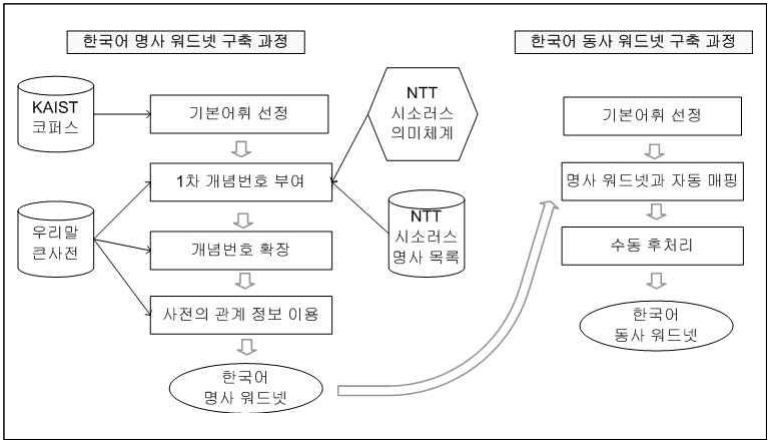
이 중에서 자료의 크기 면에서 실용성을 가지며, 지속적으로 관리 및 보완되고 있는 4개 어휘 의미망인 CoreNet, 세종전자사전, U-Win, KorLex를 개발 순서에 따라 살펴보겠다.

카이스트(KAIST)에서 구축된 CoreNet은 국내 최초의 실용화 시스템이고, 체계화된 방식으로 자료와 인쇄물이 일반에 공개되었다. 일본의 NTT LN를 참조하고 언어 전문가에 의해 수동으로 정교화하는 참조 구축 방식을 따랐다. NTT LN의 계층화된 개념 구조에 200개 가량의 개념을 추가

9) <표 3>에 제시된 국내 어휘 의미망의 현황은 가능한 한 구축 기관의 최근 발표 자료 및 홈페이지를 참조하고 각 구축 기관에 확인을 거쳤으나, 응답이 없는 경우 저자가 구할 수 있는 최신 자료에 기댔다. 한국어 명사 워드넷의 구축기간과 다국어 어휘 데이터베이스의 결과물 크기가 이에 해당한다.

10) 국내에서 '워드넷'은 어휘 의미망(lexical semantic network)을 일컫는 보통명사('word net')화하여, 여러 어휘 의미망 명칭에 사용되고 있다.

11) 이와는 다른 연구 방향으로, 세종전자사전의 동사 일부를 미크로 코스모스의 프레임 형식으로 변환하려는 연구(신효필, 2007)를 시도하였고, 언어중립적 온톨로지 구축을 위한 기반 연구 등이 진행 중이다.



<그림 5> CoreNet 명사 워드넷과 동사 워드넷 구축 과정

하고, 기본 어휘를 포함한 고빈도 약 6만 개의 어휘 의미를 약 3,000개 개념에 할당하였다. 이때 어휘 의미의 분화는 한글학회의 『우리말 큰사전』을 기준으로 하였고, 사전 정의문에서 핵심어나 중요 의미 자질을 추출하여 관련된 개념에 연결하였다.

이 일련의 과정은 자동 추출과 어휘 전문가가 개입하여 판단하는 반자동으로 이루어진다. 동사, 형용사, 동작성 명사의 경우 격틀 정보가 제공되는데, 논항의 의미 자질을 CoreNet 명사부와 연동하여, 자료 내부의 선순환 구조를 잘 유지하고 있다. 같은 방식으로 중국어 어휘 의미망을 개발하여 일·한·중 3개어 간 연계성을 가지나, 영어나 유럽어와의 연동성은 낮다. 또한, 한 개의 개념에 때로는 2백 개가 넘는 어휘 의미가 할당되어, 최소 구성 단위의 크기(grain size)가 크다보니 세밀한 의미 구분이 필요한 검색 등에 이용하기에는 어려움이 있다.

21세기 세종계획(1998-2007) 사업의 연구 프로젝트의 일환으로 개발되어온 세종전자사전은 G. Gross의 '적정 술어(appropriate predicate)'와 '대상 부류(object class)' 분석 방법론을 이용하여 어휘 의미의 계층 구조를 직접 구축하는 방식으로 이루어졌다. 언어학자의 직관과 기구축된 언어 자원을 모두 이용하여 <표 4>와 같이 5개의 최상위 부류를 포함한 총

<표 4> 세종사전 명사의미부류 체계

최상위부류명	총 의미부류 수
구체물	196
집단	29
장소	52
추상적대상	149
사태	155
합 계	581

581개의 의미 부류를 추출하고 구조화하였다.

여기에서는 최상위 부류 5개를 대상으로 하향식으로 의미 영역을 분화한다. 최상위 부류로는 <구체물>, <집단>, <장소>, <추상적대상>, <사태> 등 5개의 부류가 설정되었는데 이중 첫 4개는 논항 명사의 의미 부류이고, 나머지 <사태> 부류는 술어명사의 의미 부류이다. 또한 최상위노드를 기점으로 최소 2층위에서 최대 7층위까지의 깊이를 갖는 위계적 구조를 갖는다. 이것을 이용하여 명사 등의 어휘 의미를 세분화하고, 술어기능을 하는 명사와 용언의 논항 의미 자질을 기술한다.

세종전자사전은 일반적인 전자사전과 어휘 의미망이 수록한 정보를 모두 포함할 만큼 방대한 양의 체계적인 언어 정보를 제공하며, 일반에게 무료로 공개되므로 활용 가능성이 높다. 다만, 개념 분류 체계를 독자적으로 구축하였기 때문에 다른 어휘 의미망이나 개념망과의 사상이나 연동이 어렵고 다국어 연계성이 아주 낮다.

울산대학교에서 개발한 U-WIN(User-Word Intelligent Network)은 사전, 말뭉치 등의 기초 자원을 이용하여 상위어를 추출하는 직접 구축 방식을 따랐다. 『표준국어대사전』에서 다의어를 구분하는 어휘 의미를 기본적인 구성 단위로 삼아 그 단위에 해당하는 사전 정의문에서 중심어를 추출하여 상향식으로 구축하되, 필요에 따라 하향식 또는 참조 구축 방식을 일부 수용하였다. 어휘 의미망의 구조는 최대한 의미론적 상하관계로 맺어진 계층적 구조를 가지며, <표 5>처럼 명사·동사·형용사와 같은 내용어 뿐 아니라 부사나 기능어를 모두 포함한다는 점에서 한국어의 특성

<표 5> U-WIN의 구축 결과물

규모	품사	대상언어
30여만 어휘	핵심적 대상 – 명사, 동사, 형용사 부수적 대상 – 부사, 관형사, 대명사, 감탄사, 조사, 수사, 의존명사 등	한국어

을 세밀하게 기술한다.

또한, 면역학 온톨로지, 국가과학기술 R&D 기반정보 온톨로지 구축에 활용되고 있어, 국책연구기관에서 실질적인 활용도가 가장 높은 편에 속한다. 다만, 직접 구축한 어휘 의미망이 가지는 공통적 단점인 다국어 연계성이 낮고 다른 개념망이나 어휘 의미망과 연동하기 어렵다는 점을 U-Win도 공유한다. 공개된 자료는 일부분이라 아직 일반인들이 활용할 수 없다.

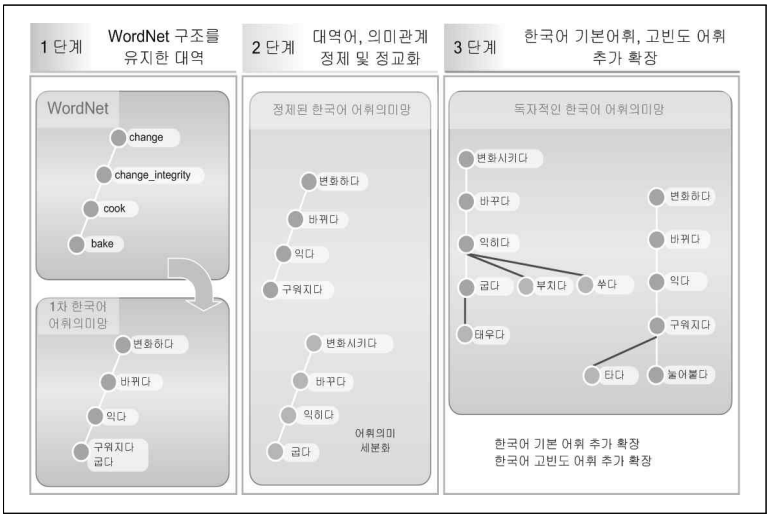
부산대학교의 KorLex는 PWN 2.0(2003)을 기반으로 참조 구축 방식으로 이루어지고 있다. 3장에서 소개한 것처럼 PWN은 세계 30여 개 언어의 50여 개 어휘 의미망의 참조 모델이 되므로, 이들의 피봇(pivot) 역할을 할 수 있다. 이러한 다국어 연계성은 다국어사전 편찬과 기계 번역에 직접 활용된다. 실제로 PWN과 EWN의 배포를 담당하는 메모 데이터(Memodata)는 이런 방식으로 다국어사전과 기계 번역 서비스를 하고 있고, 이때 사용하는 한국어 모듈이 KorLexNoun1.0이다.

참조 구축 방식의 잘 알려진 단점은 참조 모델에 경도되어, 새로 구축하는 언어의 의미 체계를 제대로 반영하지 못한다는 것이다. 이를 보완하기 위해 KorLex는 <그림 6>과 같이 3단계를 거침으로써 한국어 의미 체계의 독자성을 확보하려고 한다. PWN과 사상하는 1단계부터 각 신셋에 『표준국어대사전』의 어휘 의미를 연결하고 PWN의 명사, 동사, 형용사, 부사에 추가하여 한국어의 특성인 분류사를 포함하여 <표 6>과 같이 구축하고 있다.

하지만 KorLex는 명사와 동사, 그중에서도 명사를 중심으로 구축되고 있고, 형용사와 부사의 경우 아직 <그림 6>의 구축 1단계에 머무르고 있다.[12] 또한 PWN의 약점인 격틀 정보의 정교성 보완은 KorLex에서도 하

<표 6> PWN 2.0과 KorLex 1.5 비교

	PWN 2.0		KorLex 1.5	
품사	분화된 어휘형태	개념(신셋)	분화된 어휘형태	개념(신셋)
명사	114,648	79,689	97,292	86,185
동사	11,306	13,508	20,135	16,922
형용사	21,436	18,563	21,552	18,558
부사	4,669	3,664	3,123	3,611
분류사	–	–	1,377	1,377
계	152,059	115,424	143,479	126,653



<그림 6> KorLex 구축의 3단계

루바삐 수행되어야 할 부분이다.

12) 2007년 9월 공개한 KorLex의 버전은 명사와 동사가 1.5이고, 형용사, 부사, 분류사가 1.0이다.

5. 맺는 말

기계가 인간이 사용하는 언어를 이해하고, 지식을 재현하는 것이 가능할까? 전망은 엇갈린다. 인간과 동일한 방식과 수준으로 이루어질 것이라는 기대는 매우 낮지만, 그리 머지않은 미래에 제한된 수준에 도달하고 그 수준이 단계별로 높아질 것이라는 예측이 일반적이다. 마치 비행기가 새가 나는 것과 다른 방식의 기술을 갖추었지만 '이카루스의 꿈'을 실현시켰듯이 말이다. 물론 이런 예측이 현실이 되려면 다양한 실험과 끊임없는 노력이 선행되어야 한다.

이 글에서 살펴본 어휘 의미망은 인간의 지식과 언어를 이해하려는 수많은 지식 표상 체계 중 한 방편이다. 미국이 주도하는 국외 지식 표상 체계의 경우, 연구의 씨앗이 되었던 아이디어도 달랐고, 만드는 방식도 다양했으며, 따라서 각각의 장점과 단점도 다르다. 하지만, 장기적인 연구로 진행될 수 있었고 그 결과가 공개되었으므로, 이를 바탕으로 한 수많은 확장, 개선, 활용이 가능하다.

국내의 자연 언어 처리 역사의 출발점은 1980년대 초반으로, 미국·유럽·일본에 비해 20~30년 늦게 시작했다. 공학 분야에서는 기술 개발에 늦게 참여함으로써 초기의 문제점을 겪지 않고 건너뛰는 다행스러운 경우도 있지만, 자연 언어 처리 분야에서는 그런 행운을 기대하기 어렵다. 또 비행기를 만들 수 없으면 외국에서 만든 것을 사다 쓰면 되는 것과 같이 완제품 형태의 한국어 처리 시스템을 수입할 수 있는 상황이 일어날 것 같지도 않다. 한국어가 국제어로서 위상을 갖추어, 시장에서 재화 가치를 충분히 가지는 경우가 아니라면, 한국어 처리 기술은 국내 연구진에 의해 이루어질 것이고, 그럴 수밖에 없다.

언어 처리 기반 기술과 기초 자료 구축에 대한 국외(특히 미국)의 연구 지원에 비해, 지난 20여 년간 국내 한국어 처리 기술에 대한 지원 현실은 흡사 안팎 곱사 등이 같은 형국이다. 대규모 장기간 국책 과제는 21세기 세종계획이 전부라고 해도 과언이 아닐 정도이고, 대부분 전자사전, 말뭉치, 언어 분석 도구 등은 대학이나 국책연구소 일부 연구진에 의해 단기간에 부분부분 만들어졌다. 신자유주의적 규칙의 적용을 받는 공학 분야

의 단기 목적성 연구비는 투자액보다 몇 배의 결실을 얼마나 빨리 회수할 수 있느냐는 경제적 효용성에 따라 지원이 결정되므로, 결실이 더딘 한국어 처리 기반 기술과 기초 자료 연구는 언제나 뒷전으로 밀린다. 포함된다고 하더라도 투자가치 높은 상품을 만드는 데 필요한 수많은 요소 기술 중 하나로만 여겨진다. 이와는 대척점에 선 기초 학문 진흥 관점에서, 지난 5년간 인문학 분야에 집중 투자된 대규모 국책 연구는 한국어 기초 자료의 체계적인 구축에 거의 눈길을 준 적이 없다. 문·사·철이 인문학의 본류라는 믿음은 언어학을 인문학의 주변에 위치시켰고, '전산언어학'이나 '자연 언어 처리'는 경계선 밖에 머물게 했다.

4장에서 소개한 국내 어휘 의미망은 국외 지식 표상 체계 구축에 비교해 볼 때, 두 방향으로 치우쳐 있고, 그 결과의 활용이 아직 저조하다. 자료 자체가 가진 단점과 부족한 부분이 많이 있지만, 이런 현실적인 어려움 속에서 구축되어 왔다는 점을 간과해서는 안 된다. 다른 관점과 방법론으로 또 다른 어휘 의미망을 구축할 수도 있지만, 현 시점에서는 기왕에 만들어진 어휘 의미망을 보완하고, 통합하고, 개선하고, 활용하는 데 더 많은 노력을 쏟아야 한다. 바로 여기에서 한국어에 대한 책임과 권한을 모두 갖고 있는 국립국어원의 역할이 기대된다. 다양한 방식으로 만들어져서 결과물이 상이하나, 그만큼 각각의 장점과 활용도를 가진 기존 어휘 의미망을 수용하여, 통합하고, 가공하여, 한국어의 또 다른 공용 기초 자료로 일반에게 제공함으로써, 한국어 의미 처리 연구뿐 아니라 다양한 이론 및 응용 연구의 파생과 시스템 구현을 촉진하게 될 것을 기대한다.

| 참고 문헌 |

1차 자료

▷ 국내 어휘 의미망 소개 자료 및 운영 웹사이트

<CoreNet, http://bola.kaist.ac.kr/>

최기선 외(2005),『다국어 어휘 의미망(CoreNet)』1, 2, 3권, 한국과학기술원 전문용어언어공학연구센터, KAIST Press.

<세종전자사전, http://www.sejong.or.kr>

이성헌(2007), 사전편찬에 있어서의 어휘 의미망의 역할과 기능, “한국어 어휘 의미망 구축과 사전편찬 학술회의 자료집”, 국립국어원, pp.77-90.

홍재성(2004), “21세기 세종계획 전자사전 개발 연구보고서” (11-1370000-000089-14), 문화관광부, 국립국어원.

<U-Win, http://nlplab.ulsan.ac.kr/>

옥철영(2005), U-Win, “Kiponto 2005 발표 자료집”, no pagination.

옥철영(2007), 어휘 의미망과 국어사전의 체계적 구성, “한국어 어휘 의미망 구축과 사전편찬 학술회의 자료집”, 국립국어원, pp.35-53.

최호섭 외(2006), 대규모 우리말 어휘지능망 구축 방법, “한글”, 273, pp.125-141

<KorLex, http://corpus.fr.pusan.ac.kr/korlex>

윤애선(2007), 한국어 어휘 의미망 구축의 현황과 과제, “한국어 어휘 의미망 구축과 사전편찬 학술회의 자료집”, 국립국어원, pp.3-31.

이은령·윤애선(2005), 피동 정보를 통한 한국어 동사 어휘 의미망의 정제, “한국어학”, 제28권, pp.139-166.

황순희·윤애선(2005), 의미자질을 고려한 명사어휘 의미망의 구축(1), “한국어학”, 제29권, pp.309-338.

<한국어 명사 워드넷>

문유진·노봉남·윤평현(1995), "한국어 명사 워드넷의 구축 방안 및 구현에 관한 연구", 한국과학재단 과제결과보고서 (과제번호: 94-0100-11-01-1).

문유진(1996), 한국어 명사를 위한 WordNet 설계와 구현, "정보과학회 논문지(C)", 제2권 제4호, pp.437-445.

<다국어 어휘데이터베이스>

강범모·이유선·차재은(2002), 다국어 어휘데이터베이스 구축 방법론 연구 및 모형 개발, 고려대학교민족문화연구원.

최경봉, 도원영(2005), 한국어 동사 의미망 구축을 위한 상위 온톨로지 구성에 관한 연구, "한국어학" 28, pp.217-244.

<한국어 시소러스>

이창기·이근배(2000), 의미애매성 해소를 이용한 WordNet자동 매핑, "제12회 한글 및 한국어정보처리 학술대회 발표논문집", pp.262-268.

▷ 국외 어휘 의미망 소개 자료 및 운영 웹사이트

<PWN(WordNet), http://wordnet.princeton.edu/>

Fellbaum, Ch. et al.(1998), *WordNet: An Electronic Lexical Database*, The MIT Press.

<EWN(EuroWordNet), http://www.illc.uva.nl/EuroWordNet/>

Vossen, P.(1998), *EuroWordNet: A Multilinugal Database with Lexical Semantic Network*, The Kluwer Academic Publishers.

<BWN(BalkaNet), http://www.ceid.uptras.gr/Balkanet/>

Pala, K. & R. Sedláček(2005), Enriching WordNet with Derivational Subnets, *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing*, pp.305-311.

<HowNet, 웹사이트??>

Dong, Z. & Q. Dong(2006), *HowNet and the Computation of Meaning*, World Scientific.

<CYC, http://www.cyc.com/>

<FrameNet, http://framenet.icsi.berkeley.edu/>

<Memodata, http://www.memodata.com>

<Mikrokosmos, http://crl.nmsu.edu/Research/Projects/mikro/>

<SUMO, http://www.ontologyportal.org/>

2차 자료

신효필(2007), 온톨로지를 이용한 사전의 기술과 활용, "한국어 어휘 의미망 구축과 사전편찬 학술회의 자료집", 국립국어원, pp. 393-114.

Dau, F. & al. eds.(2005), *Conceptual Structures: Common Semantics for Sharing Knowledge*, Springer.

Hovy, E.(2005), Methodologies for the Reliable Construction of Ontological Knowledge, *LNAI*, vol. 3596, pp.91-106.

Nirenburg, S. & V. Raskin(2004), *Ontological Semantics*, The MIT Press.

Schalley, A. & D. Zaefferer eds.(2007), *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*, Mouton de Gruyter.

Sowa, J.(1999), *Knowledge Representation: Logical, Philisophical, and Computational Foundations*, Brooks and Cole.