한독 워드넷 구축을 위한 기본 방법론 고찰*

남유선(원광대)

1. 들어가는 말

인간이 자연어 문장을 이해하기 위해 이용하는 지식베이스는 개념체계 Ontology 혹은 시소러스 Thesaurus라고 한다. 그 동안 이러한 지식 베이스를 컴퓨터에 도입하려는 많은 노력이 있었으며, 근래에 들어서는 자연어 처리 분야의 여러 연구들에서 광범위하고 완전한 어휘 지식 베이스의 필요성에 대한 요구가 높아지고 있다. 이에 대한 연구의 결과로서 자연언어처리 분야에서 특히 워드넷 WordNet이 주요 관심의 대상이 되고 있다(Miller 1990, Fellbaum 1998). 워드넷은 인간의 어휘지식에 대한 심리언어학적 연구의 성과를 토대로 1985년부터 Princeton 대학 인지과학연구실이 구축해온 영어어휘 데이터베이스이며,[1] 현재 자연언어처리와 정보검색의 여러 분야에서 널리 이용되고 있다. 또한 이를 바탕으로 미국이나 유럽, 인도, 일본, 중국, 대만 등 아시아 여러 나라와 국내에서도 워드넷에 대한 연구가 활발하게 진행 중에 있다(Fellbaum 1998, Kunze 1999, Feldweg 2001).

이와 더불어 다국어판 워드넷을 구현[2]하려는 시도가 잇따르고 있으며, 예컨대 유로워드넷(EuroWordNet)(Vossen 1997, 김현권 2000)이나 한국과학기술원 전문용어언어공학연구센터에서 개발한 KORTERM 한국어 워드넷(최기선 외 2003) 등이 인접 언어를 중심으로 그 범위를 확장해가고 있다. 그러나 아직까지 한국어와 독일어를 대상으로 하는 워드넷 구축은 이루어지고 있지 않다[3].

---

* 본 논문은 한국과학재단 특정기초연구(과제번호: R01-2003-000-11618-0)지원으로 수행되었음.

1) Princeton University Cognitive Science Laboratory. WordNet - a Lexical Database for English [WWW] <http://www.cogsci.princeton.edu/~wn/>.

2) University of Amsterdam Computer Centrum Letteren. EuroWordNet: Building a Multilingual Database with WordNets for Several European Languages. [WWW] <http://www.let_uva.nl/~ewn/>.

3) 독일어 분야에서 워드넷과 관련된 국내 연구는 아직까지 소수에 그치고 있으며, 이

본고를 출발점으로 하여 최종적으로 다다르고자하는 목표는 한독 워드넷 구축이다. 이를 위해 본고에서는 한독 워드넷 구축을 위한 기본적인 방법론을 고찰하는 것을 주요 목표로 삼겠다. 먼저 첫 단계로서 영어권에서 사실상 표준이 되어 가고 있는 Princeton 대학의 워드넷을 중심으로 워드넷의 주요 기본 개념체계에 관해 논해보겠다. 이어서 이를 기반으로 구축된 독일어 관련 워드넷과 한국어 관련 워드넷의 특징을 고찰해보고, 다국어판 워드넷 구축을 목표로 추진되어 현재 상당한 성과를 보이고 있는 KORTERM 한국어 워드넷의 특징과 방법론(최기선 외 2003)을 고찰해 봄으로써 이와 같은 다국어 워드넷을 기반으로 한 한독 워드넷 구축을 위한 기본적인 체계를 모색해보고자 한다. 본고의 최종적인 목표인 한독 워드넷은 한국어 개념체계를 출발점으로하여 한국어와 독일어의 관계로 확장하면서 구축될 것이다. 바로 이 점이 현재 상당한 진전을 보이고 있는 독일어 관련 다국어판 워드넷인 유로워드넷 보다는 KORTERM 한국어 워드넷을 한독 워드넷 구축 모델로 택한 이유이기도 하다. 본고의 목적이 한독 워드넷 구축을 위한 기본적인 방법론을 모색하는 데에 있기 때문에, 워드넷의 구체적인 구축 및 활용부분에 대한 논의는 추후 연구로 미루기로 하겠다.

2. 워드넷의 기본 개념체계

워드넷이란 본래 전산적인 이용을 목적으로 구축된 어휘망이며, 단어들의 의미 관계를 망으로 체계화한 것이다. 워드넷이란 용어는 맨 처음 Princeton 대학의 인지과학 연구소에서 George Miller 교수의 지도 하에 만들어진 "인간의 어휘 기억에 대한 심리언어학적 이론을 기반으로 디자인된 온라인 데이터베이스 사전으로 명사, 동사, 형용사, 부사에 대한 어휘데이터베이스이다. 영어 워드넷은 현재 2.0버전까지 구축되었으며 연구용으로 무료 공개되어 영어권에서 널리 이용되고 있다.

워드넷은 "사전을 단순한 자모순보다는 개념적으로 찾는데 이용하기 위해"

민행(1999, 2004)과 오장근(2002)이 이와 관련된 연구를 수행한 바 있다.

개발되었으며, "단어의 의미를 통해 어휘정보를 조직화하려는 시도"에서 비롯되었다(Miller 1990). 워드넷은 자연언어처리의 여러 분야에서 필요로 하는 아주 유용한 정보이다. 특히 정보검색, 기계번역 등과 같이 어휘의 의미를 다루는 데 있어서 워드넷은 지식 베이스로써 매우 중요한 역할을 한다. 워드넷은 단순하게는 관련 있는 단어들을 분류해서 유의어 정보를 제공하는 수준의 것도 있지만, 대개는 개념간의 여러 가지 관계 및 체계적인 의미풀이를 포함함으로써 다양한 자연언어처리 시스템의 지식 베이스로 이용되고 있다(최기선 외 2003).

2.1. 단어와 개념의 대응관계

워드넷의 기본 의미관계는 유의어관계 synonymy이며, 워드넷의 기본 구성단위는 유의어의 집합인 신셋 synset이다. 예를 들어 <표 1>의 {shot, pellet}와 {shot, injection}는 각각 하나의 의미를 표현하는 유의어 집합이다. 이 두 유의어 집합을 통해서 'shot'라는 유의어의 여러 의미가 실체화될 수 있다. 이와 같은 유의어집합을 신셋이라고 한다. 이러한 신셋은 표현하는 의미가 무엇인지 명시하지는 않으며 다만 그와 같은 의미가 존재하는 것만을 나타낸다(이재윤/김태수 1999 참조).

<표 1>

{shot, pellet}
{shot, injection}
.
.

대부분의 신셋에는 재래식 사전에서 볼 수 있는 것과 같은 종류의 뜻풀이가 수반되는데, 그러나 신셋이 사전 표제어와 동일한 것은 아니다. 특히 다의어가 사전 표제어인 경우에 사전에서는 서로 다른 뜻풀이가 여러 개 수반되는데, 워드넷에서는 신셋 하나에 뜻풀이가 하나만 주어진다. 즉 신셋 하나는 어휘화된 개념 하나를 나타낸다.

2.2. 어휘 계층

각각의 어휘는 일종의 계층구조를 형성하게 되는데 명사를 조직화하는 과정에서 가장 중요한 의미관계는 어휘화된 개념들 사이의 관계로서 하위어 관계 hyponymy를 들 수 있다. 예컨대 명사 robin은 명사 bird의 하위어이며, 이와는 반대로 bird는 robin의 상위어 hypernym이다. 워드넷에서 상하위어 관계는 신셋들 사이에 포인터 pointer를 둠으로써 나타낸다.

(i) 상위어 관계 hypernymy: S_s @→ S_g

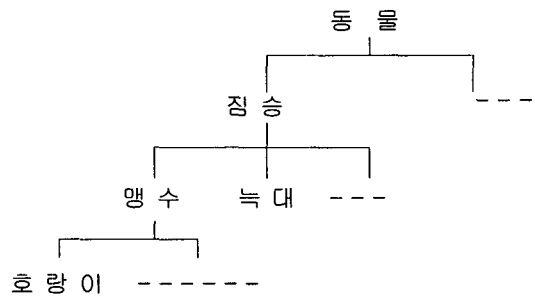
예) {robin, redbreast} @→ {bird} @→ {animal, animate_being} @→ {organism, life_form, living_thing}

(i)에서 @→는 'IS-A', 'IS-A-KIND-OF'를 의미하며, 상위어 관계는 특수한 것에서 일반적인 것으로의 의미관계를 나타낸다.

(ii) 하위어 관계 hyponymy: Sg ~→ S_s

(ii)에서 ~→는 'SUBSUMES'를 의미하며, 하위어 관계는 일반적인 적인 것에서 특수한 것으로의 의미관계를 나타낸다.

이러한 방식으로 상하위어 관계를 표상하게 되면 일종의 어휘계층구조 혹은 수형도 tree diagram가 생겨난다. 전산학에서는 이러한 계층구조를 '상속체계 inheritance system'라고 한다. 즉 하위의 것이 상위의 자질을 물려받음으로 이런 공유자질은 해당 항목마다 모두 기재할 필요가 없고 상위의 것에만 기재하면 된다는 것이다.



위의 수형도에서 '호랑이'에 대한 상위어는 '맹수'이고, '맹수'에 대한 상위어는 '짐승'이며, '짐승'에 대한 상위어는 '동물'이다.

상하위어 관계에 대한 모든 정보는 재래식 사전에도 포함되어 있으나 이를 찾아내기가 쉽지 않다. 이러한 정보 추출작업은 워드넷에서 주로 사람에 의해 수동으로 이루어져왔다. 개념체계를 구축할 경우 가장 확실하고 정확한 방법은 수동으로 구축하는 것이다. 그러나 이것은 많은 비용과 시간을 소비한다는 단점을 가지고 있다. 또한 어플리케이션에 따라 수동으로 구축된 것만큼의 정확도가 필요하지 않은 경우도 있다. 이러한 이유로 이미 구축되어진 많은 어휘정보들을 이용하여 대량의 어휘 지식을 자동 혹은 반자동으로 얻어내려는 많은 연구들(김민수 외 1995, 조평옥 1996, 이창기 외 1999, 2000)이 이어져 오고 있다.

2.3. 개념분류 체계

워드넷은 품사별로 구축되기 때문에 개념분류도 품사별로 이루어진다. 명사 워드넷에서는 모든 명사들이 단일 계층구조에 나타나지 않고 여러 계층구조에 나타난다. 각 계층구조의 최상위에는 더 이상 상위어를 취하지 못하는 최상위 계층이 존재한다. 이 다수의 계층구조들은 각각 어휘장 semantic field에 해당한다. 워드넷의 명사는 <표 2>와 같은 25개의 최상위 계층으로 나뉘어 계층구조를 이루고 있는데 계층의 깊이는 12단계를 넘지 않는다(G.A. Miller 1998: 29).

<표 2> 최상위 계층 (G.A. Miller 1998: 29)

List of 25 unique beginners for noun source files		
{act, activity}	{food}	{possession}
{animal, fauna}	{group, grouping}	{process}
{artifact}	{location}	{quantity, amount}
{attribute}	{motivation, motive}	{relation}
{body}	{natural object}	{shape}
{cognition, knowledge}	{natural phenomenon}	{state}
{communication}	{person, human being}	{substance}
{event, happening}	{plant, flora}	{time}
{feeling, emotion}		

동사 워드넷은 15개의 범주—신체기능과 치료 bodily functions and care, 변화 change, 커뮤니케이션 communication, 경쟁 competition, 소비 consumption, 접촉 contact, 인지 cognition, 창조 creation, 동작 motion, 감정/심리 emotion or psych, 상태 state, 지각 perception, 소유 possession, 사회 상호작용 social interaction, 날씨 weather—로 나누어진다. 이 중에서 14개 범주는 행동이나 사건을 기술하는 일반적인 동사로 이루어져 있고, 나머지 1개의 범주는 resemble, belong, suffice 등과 같이 상태를 표현하는 동사로 이루어져 있다.

형용사 워드넷은 의미적, 통사적 속성에 따라서 기술형용사 descriptive adjectives, 관계형용사 relational adjectives, 지시-수식형용사 reference-modifying adjectives, 색채형용사 color adjectives로 나누어지고, 반의관계와 유사관계를 중심으로 구성되어 있으며, 계층관계는 설정하지 않고 있다.

2.4. 개념 사이의 관계 표현

워드넷에서는 개념 사이의 관계를 이용하여 신셋을 정확히 표현한다. 워드넷에서 사용하는 관계 유형은 <표 3>과 같다.

유의관계 Synonymy는 신셋을 이루는 워드넷의 기본적인 관계에 속한다. 유의관계 설정이 pipe와 tube와 같이 명사에서는 비교적 쉽지만, 동사, 형용사, 부사의 경우에는 엄밀한 의미의 유의어가 많지 않으며 기준 적용의 강도에 따라 그 차이가 심하게 나타난다. 예를 들어 hide와 conceal은 둘 다 '숨기다'의 의미를 갖는 동의어이지만 'when have you hidden Dad's slippers?라는 문장에서 'hidden' 대신 'concealed'를 사용하면 어색하게 된다. 따라서 워드넷에서 동사의 유의관계는 반드시 어휘적인 유의어에만 적용된 것은 아니다.

<표 3> 워드넷에서 정의된 관계 유형 (이재윤/김태수 1999: 216)

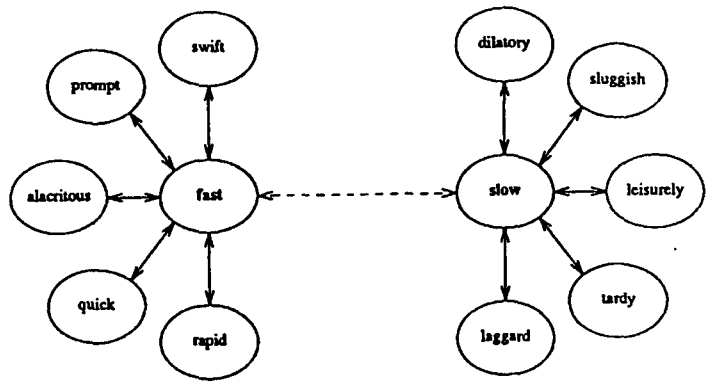
Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry
N = Nouns Aj = Adjectives V = Verbs Av = Adverbs		

반의관계 antonymy는 특히 형용사와 부사의 경우에 이용되는 관계이다. 하나의 형용사가 여러 개의 의미를 가지고 있는 경우 각각의 의미에 따라서 반의어가 달라 질 수 있기 때문에 형용사의 의미를 표현할 때에 반의어를 이용하는 것이 효과적이다.

하지만 형용사의 반의어를 규정할 때에는 다음과 같은 사항이 고려되어야 한다. 첫째, 비슷한 의미를 지닌 두 형용사가 서로 다른 반의어를 가질 수 있다(K.J. Miller 1998: 50f). 예를 들어 heavy와 weighty의 반의어는 각각 light와 weightless이지만 light와 weightless는 유의관계가 아니다. 따라서 신셋 {heavy, weighty}와 {light, weightless}를 반의어로 설정하는 것은 적절하지 않다.

둘째, 비슷한 의미를 지닌 두 형용사 중에서 한 쪽에만 반의어가 존재할 수 있다. 예컨대 heavy와 ponderous 중에서 heavy는 반의어가 있고 ponderous에 적합한 반의어는 없다. 워드넷에서는 비슷한 의미를 지닌 형용사끼리 동의관계를 무리하게 설정하기 보다는 유사관계 포인터 similar pointer를 설정하여 <그림 1>과 같이 양극관계의 두 형용사(fast와 slow)를 두고 비슷한 형용사를 각 극단의 형용사에 연결시켜주는 방식을 도입했다.

<그림 1> 형용사의 양극관계 구조 (K.J. Miller 1998: 51)

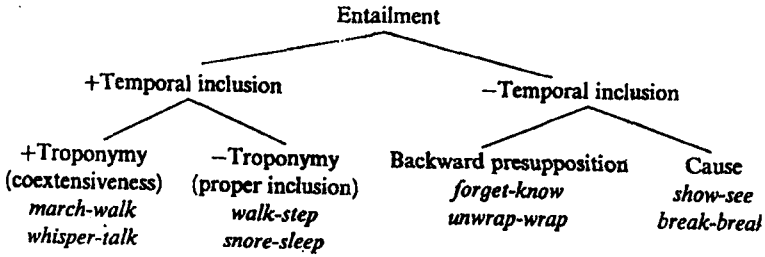


하의관계 hyponymy는 상의관계 hypernymy와 함께 명사 신셋 사이의 계층 관계를 표현한다. 부분관계 meronymy는 전체관계 holonymy와 함께 부분-전체 관계를 나타낸다.

함의관계 entailment는 동사 사이에서 내포관계(buy-pay) 또는 인과관계(give-have)를 표현하는데 사용되는데, 이러한 관계는 "동사 V1을 포함한 문장이 동사 V2를 포함한 문장을 함의하면 동사 V1과 V2는 함의관계에 있다"라고 정의된다. 동사 사이의 함의관계는 <그림 2>와 같이 세분화하고 있으며 여기에서 양식관계 troponomy는 별도로 구분하고 있다. 예컨대 snore는 sleep을 함의하고 있지만 snore가 sleep의 양식에 해당하지 않으므로 이 두 단어는 양식관계에 놓여 있지 않다.

양식관계 troponomy는 함의의 일종이지만 워드넷에서는 범주를 달리하고 있다. 동사 V1이 특정한 한 양식으로서 동사 V2를 행할 때 동사 V1은 동사 V2와 양식관계가 된다. 예컨대 limp(절뚝거리다)는 walk의 한 양식이다. 따라서 "He is limping"이라는 문장은 항상 "He is walking"의 의미를 내포하게 된다.

<그림 2> 동사 사이의 함의관계 (Fellbaum 1998: 84)



3. 독일어 관련 워드넷 연구

국내·외에서 워드넷에 관한 연구가 활발하게 진행되고 있다. 외국의 경우 오래 전부터 어휘 의미 체계에 대한 연구가 진행되어왔다. 오랜 기간 구축 및 보완 과정을 통해 상당한 수준의 워드넷이 구축되었고, 그 중 일부는 공개되어 자유롭게 연구용으로 이용되고 있다. 아래에서는 현재 진행 중인 독일어 관련 주요 워드넷의 연구를 살펴보겠다.

미국의 프린스턴 대학의 워드넷을 기반으로 유럽에서 활발하게 진행되고 있는 **유로워드넷(EuroWordNet)**[4] 프로젝트는 다국어 어휘의미 데이터베이스이며, 여기에는 유럽 8개국 언어들(1차 구축: 네덜란드어, 이탈리아어, 에스파냐어, 영어 ; 2차 구축: 독일어, 프랑스어, 에스토니아어, 체코어)이 참여하고 있다. 프린스턴 워드넷이 영어라는 한 언어로 구성된 어휘의미망인데 비해서 유로워드넷이 다국어를 기반으로 하고 있다는 점에서 그 특징을 찾아 볼 수 있다(김현권 2000: 145). 다국어 어휘의미 데이터베이스에는 단어들의 다양한 의미들이 코드화되어 있다. 이러한 단어 의미들은 어휘-의미적인 관계(예: 유의어, 반의어 또는 하위어)를 통해서 서로 연결되어 있다. 개념과 관계는 하나의 의미망을 형성하고 있다. 유로워드넷의 다국어적 데이터뱅크에는 개별언어의 다양한 의미망의 개념들이 하나의 중간언어 InterLingua에 연결된다. 이를 통해 하나의 개념을 여러 언어로 표현하는 단어들이 서로 연결되어 있다. 유

4) 유로워드넷 홈페이지: http://www.illc.uva.nl/EuroWordNet

로워드넷은 63개의 의미적 차이를 갖고 있는 "상위 개념체계 top-ontology"를 포함하고 있다. 이 상위 개념체계는 각 언어들의 공통적인 의미적 기본골격을 제공하고 있다. 반면에 언어특징적인 특성들은 개별 워드넷에 남아있다. 이러한 데이터베이스는 특히 단일어나 다국어 정보검색을 위해 유용하게 사용되고 있다.

현재 다국어 데이터베이스의 상태로 구축되어 있는 유로워드넷에는 독일어, 네덜란드어, 이탈리아어, 영어, 스페인어, 프랑스어, 체코어, 에스토니아어가 속해있다. 각 언어의 워드넷은 하나의 중간언어를 통해 서로 연결되어 있다. 각각의 워드넷 데이터베이스 구축은 다음과 같은 기관들이 책임지고 있다.

<표 4> 유로워드넷 관련 기관

워드넷	기 관
네덜란드어	Universität Amsterdam
스페인어	'Fundacion Universidad Empresa'
이탈리아어	Istituto di Linguistica Computazionale, C.N.R., Pisa
영어	Universität Sheffield
프랑스어	Universität Avignon und Memodata in Avignon
독일어	Universität Tübingen
체코어	Universität Masaryk in Brno, Tschechei
에스토니아어	University Tartu in Estland

또한 유로워드넷의 존속을 위한 기관으로서 The Global WordNet Association은 또 다른 워드넷 구축을 촉진시키고 개별 워드넷의 표준화와 연결 작업을 수행하고 있다. 이러한 기관들을 중심으로 지금까지 유로워드넷은 다양한 연구결과를 내놓고 있다.

- 각 언어의 워드넷을 영어 워드넷과 연결
- Princeton 워드넷에 존재하지 않았던 것을 중심으로 워드넷 1.5를 보완
- 유로워드넷 포맷으로 워드넷 1.5 보완
- 다양한 워드넷과 다른 개념체계를 연결하기 위해 워드넷 1.5에 기반 한 중간언어 InterLingua 개발

- Polaris: 유로워드넷에 접속되어야 하는 워드넷 개발을 위한 워드넷 편집기 개발
- 1차 작업에서는 30,000 개념과 50,000 어휘의미가 구축
- 2차 작업(1999년 6월 완료)에서는 15,000 개념과 25,000 어휘의미가 추가
- 데이터베이스 구축 설명 문안 작성
- 의미론적 관계를 위한 개념들과 시험문장 정리
- 유로워드넷에 대한 서적 출판.

독일에서 개발되고 있는 **게르마넷(GermaNet)** 역시 프린스턴 워드넷을 토대로 구축된 독일어의 어휘-의미망며, 1990년대 말부터 Tübingen 대학[5]을 중심으로 개발되고 있다. 게르마넷에서 사용되는 기본구조는 프린스턴 워드넷과 비슷하다. 하지만 몇몇 관점에서 게르마넷의 구조는 프린스턴 워드넷과 차이를 나타내고 있다.

첫째, 게르마넷에서는 작위적 개념들(어휘공백)이 논리적으로 의미있는 하위부류를 형성하기 위해 사용되고 있으며, 그 자체로 명백하게 표시되고 있다.

둘째, 프린스턴 워드넷에서는 단지 산발적으로 쓰이고 있는 교차분류 cross classification가 게르마넷에서는 중요한 질서성분이 되고 있다.

셋째, 보통의 다의어는 빈번히 출현하는 원형적인 형식의 목록과 더불어 '일반화'라고 할 수 있는 독특한 관계를 통해 모델화되고 있다.

게르마넷은 프린스턴 워드넷을 그대로 번역한 것도 아니고, 개별 사전이나 시소러스에 기초해서 구축된 것도 아니다. 즉 게르마넷은 완전히 새로운 자료로 구축되었다는 것이다. 게르마넷에서는 개별 어휘들이 신셋으로 요약되어 어휘적인 그리고 개념적인 관계를 통해 서로 연결된다. 유의어와 반의어 외에도 상위어 관계 hyperonymy relation가 게르마넷의 가장 중요한 정렬요소이다. 이 외에도 부분어 meronymy, 인과 causation, 파생 derivation, 내포 implication, 선택제한제한 selectional restriction, 다의어 polysemy가 적어도 부분영역에서는 일정한 역할을 한다. 게르마넷은 독일어의 기본어휘를 동의어, 상위어, 부분어와 같은 의미론적 관계에 기초하여 구축하였으며, 형용사, 명사, 동사와 같은

5) Tübingen대학의 GermaNet 홈페이지: http://www.sfs.nphil.uni-tuebingen.de/lsd

품사들을 모델화시켰으며, 모두 25,000단어의 어휘의미를 포함하고 있다. 이러한 작업은 지속적으로 보완되고 있으며, 2001년 10월까지 모두 41,777개의 신셋이 구축되어 있다.

이와 같은 독일어 관련 워드넷에 대한 선행연구가 한독 워드넷 구축 과정에서 효과적으로 활용될 수 있도록 그 가능성을 앞으로의 연구에서 좀 더 구체적으로 모색해 볼 것이다.

4. 한국어 관련 워드넷 연구

한국어 워드넷에 관한 본격적인 연구로는 문유진(2002)의 한국어 명사 워드넷을 들 수 있다. 이 연구에서는 초등학교 국어용 어휘집의 5,000단어를 대상으로 국어사전에서 자동으로 상위어 정보를 추출하고, 이를 자동 번역한 영어 워드넷과 합성하여 한국어 명사 워드넷 예비 리스트를 구축했다. 상위어 사전과 번역된 영어 워드넷을 합성할 때는 전산학자와 언어학자에 의해 수작업으로 가지치기 작업이 이루어졌다. 또한 이렇게 만들어진 예비 리스트를 기반으로 분야가 명시된 20,000개 정도의 명사를 추가로 입력하여 한국어 명사 워드넷 리스트를 구축했다. 여기에서 가지치기 작업은 일관성 있고 신중하게 수행되어야 하며 무척 까다롭고 시간이 많이 걸리는 수작업 단계로서, 상식과 언어학자의 어휘개념을 참조하여 수행되었다.

한국어 명사 워드넷의 설계에 있어서 중요한 관건을 기본 골격인 어휘개념을 표현하는 형식과 계층구조 관계의 설정으로 보았다. 어휘개념을 표현하는 형식에는 구성화 constructive 이론과 차별화 differential 이론이 있는데, 한국어 명사 워드넷의 어휘개념은 차별화 이론을 채택하여 개념별로 구별될 수 있는 상징에 의하여 표현되었다. 그리하여 어휘개념은 영어 워드넷의 동의어 집합으로 표현되고 영어 명사에 없는 어휘개념은 한국어 명사에 알맞은 어휘개념을 만들어 표현되었다. 그리고 한국어 명사 워드넷의 계층구조는 이 노드들 간의 상하위 개념 관계로 표현되었다. 이러한 의미론적 개념 분석을 토대로 도메인 선정 작업을 하고 한국어 명사 워드넷의 생성 시스템과 데이터베이스 구축 시스템을 설계하였다.

한국어 명사 워드넷을 크게 두 부분으로 구성하였는데, 그것은 응용 프로그램과 워드넷 데이터베이스이다. 응용 프로그램은 사용자 인터페이스를 위한 부분으로 명령어를 포함한 것이며 워드넷 데이터베이스는 워드넷에 필요한 어휘와 신셋에 관한 정보를 데이터베이스로 구축한 것이다. 한국어 명사 워드넷 데이터베이스는 명사의 인덱스 화일과 신셋 화일로 구성되어 있다. 인덱스 화일은 약 20,000개의 명사 단어를 엔트리로 가지고 있으며, 각 엔트리는 품사와 다의어의 뜻 그리고 다의어 별로 신셋들에 대한 포인터를 가지고 있다. 각각의 명사는 한 개 이상의 신셋에 대한 포인터를 가진다. 신셋 화일은 신셋들을 엔트리로 가지고 있으며, 각 엔트리는 상위어에 대한 포인터를 포함하고 있다.

또한 이 연구에서는 개념기반의 의미분석 기법을 제안하여 한영기계번역에서 동사의 번역 문제를 해결하는 데 한국어 명사 워드넷을 적용하였다. 그리하여 워드넷의 실용성을 검증한 것이다. 한영기계번역에서 동사의 번역 문제를 해결하는 이 기법은 연어번역 단계에서 사용되었다. 먼저 연어사전을 참고하여 동사 번역어를 찾아보고, 실패하면 연어사전에 있는 단어 용례들과의 유사도를 한국어 명사 워드넷을 사용하여 계산하였다. 계산결과 입력 문장과 가장 가까운 개념의 단어 용례를 참조하여 동사 번역어로 채택하였다.

이창기/이근배(1999, 2000)는 Princeton 대학의 워드넷을 기반으로 한영사전과 국어사전을 이용하여 한국어 명사의 개념체계를 자동으로 구축함으로써, 이미 구축되어진 다른 언어의 개념체계를 이용하여 새로운 언어의 개념체계를 자동으로 구축할 수 있음을 보였다. 한영사전과 국어사전으로부터 추출해낸 한국어 일부의 의미를 다양한 WSD(Word Sense Disambiguation) 방법을 적용시켜 워드넷의 신셋에 자동으로 연결시킬 수 있음을 보였고, 자동변환으로 나온 각각의 결과들에 대해서는 적용율과 정확도를 비교하도록 하였다. 그러나 이러한 연구들은 실험적인 수준에 그치고 있으며 현재 공개되어 이용되고 있는 것은 없다.

김민수 외(1995)에서는 한국어 MRD(Machine Readable Dictionary)의 명사의 정의를 이용하여 자동으로 한국어 명사 워드넷을 구축하는 방법을 제안하였

다. 또한 한국어 문장에서는 중심적인 말이 보통 뒤에 나타난다는 구조적인 특성과 명사의 정의문의 특수한 구조적인 특징을 분석하여 상위어를 추출하는 규칙을 제안하고 있다. 이 연구에서는 형용사, 동사, 부사 등의 단어 연관 관계를 트리구조로 구축하는 데에도 명사 워드넷 구축과 동일한 방법이 적용될 수 있기 때문에 시간과 비용의 절감효과가 크다는 점을 장점으로 내세우고 있다.

한편 한국과학기술원의 전문용어언어공학연구센터(KORTERM)에서는 컴퓨터 언어처리를 위하여 10여 년 전부터 구축해 온 코퍼스와 전자사전을 기반으로 자연언어처리에서의 의미 애매성 해소를 위한 한국어 개념 기반 워드넷을 구축하였으며, 이 한국어 워드넷은 현재 국내에서 가장 체계적으로 구축된 워드넷으로 평가받고 있다(5장 참조).

5. 한독 워드넷 구축 모델로서의 다국어판 워드넷 "KORTERM 한국어 워드넷"

KORTERM 한국어 워드넷은 다국어(한,중,일) 지향 워드넷으로서 다국어로의 확장이 용이하도록 설계되었다. 본 논문을 출발점으로하여 본 연구에서 목표로 하고 있는 한독 워드넷은 이 KORTERM 한국어 워드넷을 모델로 구축될 것이다. 우선 여기에서는 이 한국어 워드넷의 기본적인 특성을 살펴봄으로써 한독 워드넷 구축을 위한 기본적인 토대를 마련하고자 한다.

KORTERM 한국어 워드넷은 "자연언어처리에서의 의미 애매성 해소를 위한 개념체계 기반 어휘 의미망"이라는 특성을 갖고 있다. 개별 어휘 의미를 개념체계 중심의 망으로 형성하는 데에 만족하지 않고, 서술어의 경우 통사적 구조와 연결함으로써 명사, 동사, 형용사를 아우르는 입체적이고 종합적인 망을 구성하고 있다.

KORTERM 한국어 워드넷은 다음과 같은 특성을 가지고 있다.

첫째, KORTERM 한국어 워드넷은 컴퓨터 언어처리에서 발생하는 의미 애매성 해소를 목표로 다양한 의미적 정보를 제공하고 있다.

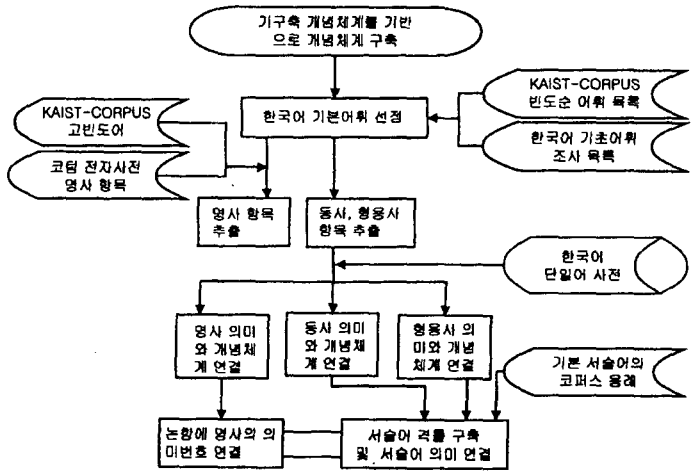
둘째, 실제로 사용되지도 않는 의미 가능성 보다는 실제로 인간 생활에 사용되는 어휘, 그리고 그 어휘의 의미만을 처리 대상으로 한다는 실제성과 경제성을 확보하려는 의도 하에 KORTERM 한국어 워드넷은 코퍼스를 기반으로 구축되었다.

셋째, KORTERM 한국어 워드넷은 한/중/일/영 한국어 기준 다국어로의 확장이 용이하도록 설계되었다. 실제로 중국어 명사 워드넷을 구축할 때 중국어 기본 명사의 워드넷을 구축하면서 한국어 대역어를 달아 주었다. 한국어 대역어는 한글학회의 "우리말 큰사전"의 의미구분에 따라 중국어와 연결되면서 자연스럽게 워드넷 번호와 연결하였다. 중국어와 한국어의 의미 일치가 안 되는 경우는 중국어와 한국어 모두에 능통한 언어학자와 중국의 전산언어학 전문가에 의해 상위 개념을 매칭시키거나 다른 개념으로 대입시켰다. 이 작업으로 중국어 워드넷과 한국어 워드넷이 연결되었으며, 이러한 방법론은 **독일어에도 적용**할 수 있다.

넷째, 종합적 체계를 가지고 있다. 보통 개념체계나 워드넷은 명사를 중심으로 이루어지고 있는데 반해서, KORTERM 한국어 워드넷은 명사와 구분되는 동사와 형용사의 특성을 잘 살리면서도 하나의 개념체계를 바탕으로 명사, 동사, 형용사를 연결하고 있다. 동사나 형용사가 명사에 비하여 문맥 의존적이기는 하지만, 각각의 서술어가 지니는 의미는 한정적일 것이므로 문맥을 고려하여 구문구조와 함께 다루되 각 구문구조에서의 명사, 형용사, 동사의 의미를 개념체계와 연결함으로써 그 동안 구분되어 처리되어 왔던 정보들을 모두 아우르고 있다.

KORTERM 한국어 워드넷은 대략 다음과 같은 3단계로 나뉘어 구축되었다. 첫 단계에서는 워드넷의 골격인 개념체계를 세웠고, 두 번째 단계에서는 한국어 어휘의 의미를 구분하여 개념체계와 연결함으로써 개념 기반 어휘 의미망을 구축하였다. 세 번째 단계에서는 코퍼스 용례를 기반으로 동사와 형용사의 격틀을 구축하고 논항의 목록과 의미를 기술하였으며 이를 개념체계와 연결함으로써 의미적 체계와 통사적 체계를 모두 고려할 수 있게 하였다.

<그림 3> KORTERM 한국어 워드넷 구축 과정 (최기선 외 2003: 9)



전산적인 이용을 목적으로 구축된 어휘망으로서 가장 잘 알려진 프린스턴 영어 워드넷과 국내에서 체계적으로 개발된 대표적인 워드넷인 KORTERM 한국어 워드넷의 특징적인 부분이나 문제점을 중심으로 살펴보면 다음과 같은 차별성과 독창성을 언급할 수 있다.

- 영어 워드넷의 명사와 동사는 유사한 동의어집합을 가지고 있음에도 불구하고 품사가 다르기 때문에 연결이 되어 있지 않은 경우가 있다.
- 일반적인 형태의 인쇄사전에서는 철자, 발음, 굴절형, 파생형, 어원 등의 많은 정보가 포함되어 있지만, 일반적으로 워드넷은 기계가독형태로 가공이 되어 있어서, 워드넷에서는 이러한 내용의 대부분이 생략되어 있다.
- 워드넷에서 명사의 계층 구조가 실현되는 방법 중에서 인간의 머리속 사전에서는 가능한 다양한 종류의 정보들이 워드넷에서는 가능하지 않는 경우도 있다. (예: IS-NOT-A-(KIND-OF) relation)
- 이 보다 더 심각한 문제는 하나의 개념이 실제로 하나 이상의 의미 관계를 표현한다는 것이다.
- 워드넷은 고유명사와 보통명사, 가산 명사와 집합 명사 사이의 정확한 구분을 하지 않으며 의미적 관계를 충분히 고려하지 않는다.

기본적으로 동의어 정보를 포함하며, 상하위 관계 이외에도 전체부분 관계, 반의어 관계 등을 반영하여 만들어진 영어 워드넷과 마찬지로, KORTERM의 한국어 워드넷 역시 상하위 관계와 전체부분 관계, 반의어 관계를 고려하였으나 반의어 관계의 체계는 아직 미흡한 편이다. 또한 영어 워드넷의 10여만 신셋에 비해서 KORTERM의 한국어 워드넷은 3,000여개의 개념노드로 구성되어 있어 개념 분류의 세분화가 필요한 상태이다. 이는 추후에 보완이 요구되는 부분이다. 하지만 영어 워드넷의 경우 각 신셋에 한해서 뜻풀이가 명시되어 있지만 KORTERM 한국어 워드넷은 어휘의 모든 의미에 대해서 뜻풀이가 명시되어 있다. 이를 이용하면 추후 개념노드를 세분화 하거나 개념노드 간의 새로운 관계를 찾는데 많은 도움을 줄 수 있다. 또한 품사별로 독립적으로 구축되지 않고 하나의 개념체계를 중심으로 연결되어 있어 서로 다른 품사들의 의미 관계까지 파악할 수 있다. 이를 바탕으로 향후 각 의미를 컴퓨터가 이해할 수 있는 논리 형태로 바꾸는 과정이 필요한데 현재 KORTERM 워드넷의 의미에는 이미 사전의 정의문이 붙어 있으므로 이를 최대한 이용하면 훌륭한 자원이 될 것이다. 위에서 언급한 바와 같이 KORTERM 한국어 워드넷은 자연언어처리에서의 의미 애매성 해소를 위하여 개념체계를 기반으로 어휘 의미를 망으로 구성한 것으로, 서술어의 경우 통사적 구조와 연결함으로써 명사, 동사, 형용사를 아우르는 입체적이고 종합적인 체계이다. 또한 어휘 선정에서부터 서술어 논항 목록까지 철저하게 코퍼스를 기반으로 구축되었다는 특징을 갖고 있다. 따라서 이러한 KORTERM 워드넷을 중심으로 한독 워드넷을 구축한다면 보다 발전된 모습의 워드넷을 구성할 수 있으리라 확신한다.

6. 맺는 말

본고에서는 먼저 한독 워드넷 구축을 위한 시발점으로서 프린스턴 워드넷을 중심으로 워드넷의 기본적인 특성을 살펴보는 데에 초점을 맞추었다. 또한 한독 워드넷에 구축에 중요한 자료가 될 수 있는 독일어 관련 워드넷 연구인 유로워드넷과 게르마넷을 간단히 고찰해봄으로서 본 연구와 관련된 독일어 워드넷 부분에 대한 가능성을 다소간 가늠해 볼 수 있었다. 이어서 한국어 워

드넷에 대한 다양한 연구결과들과 국내에서 가장 체계적으로 구축되었다는 평가를 받고 있는 KORTERM 한국어 워드넷의 기본적인 특성을 살펴보았다. 특히 KORTERM 한국어 워드넷은 명사, 동사, 형용사를 아우르는 종합적인 개념체계 갖추고 있어서 본 연구의 목표인 한독 워드넷 구축에 훌륭한 길잡이가 될 것이다. 또한 이 워드넷이 다국어 워드넷을 지향하면서 다국어로의 확장을 용이하게 하고 있기에 본 연구의 최종 목표인 "한독 워드넷" 구축을 위한 모델로 적합하다고 본다. KORTERM 한국어 워드넷을 바탕으로 구축될 한독 워드넷은 상호보완적인 측면에서 KORTERM 한국어 워드넷을 비롯하여 기존의 워드넷에 시사하는 바가 클 것이라 믿는다.

추후 연구과정에서 한국어와 독일어의 언어적 특성으로 인하여 여러 가지 문제점들이 나타나게 될 것이다. 이러한 문제점들을 보완해 가는 과정에서 보다 더 나은 워드넷 구축을 위한 제안점들이 나오게 될 것이라 믿는다.

참고문헌

김민수, 김태연, 노봉남 (1995): 국어사전을 이용한 한국어 명사에 대한 사위어 자동 추출 및 WordNet의 프로토타입 개발. 한국정보처리학회 논문집 제2권 제6호, 847-856.

김현권 (2000): EuroWordNet의 구성원리와 설계. 언어학 제27호, 145-177.

문유진 (2002): 한국어 명사 WordNet. 서울대학교 박사학위논문.

오장근 (2002): 유로워드넷 기반의 어휘 데이터베이스 활용을 위한 한국어-독일어 ILI 대응 방법론 연구. 독어학, 제 6집, 323-344.

이민행 (1999): 독일어 어휘부에 대한 연구. 독일문학 69집.

이민행 (2004): 독일어와 영어의 감정명사들의 의미관계에 대한 연구. 독일문학 89집.

이재윤, 김태수 (1999): WordNet과 시소러스. 언어탐구 1, 232-269.

이창기, 이근배 (1999): WordNet을 이용한 한국어 시소러스 자동 구축. 제11회 한글 및 한국어 정보처리 학술대회 논문집, 156-163.

이창기, 이근배 (2000): 의미 애매성 해소를 이용한 WordNet 자동 매핑. 제12회 한글 및 한국어 정보처리 학술대회 논문집, 262-268.

조평옥 (1996): 한국어 명사의 의미 계층 구조 구축, 울산대학교 교육대학원 석사학위논문.

최기선 외(편) (2002): 한국어 워드넷, "개념체계". 전문용어언어공학연구센터 한국과학기술원.

Feldweg, Helmut (2001): GermaNet - ein lexikalisch-semantisches Netz für das Deutsche. Available in Postscript format from Internet: <http://www.cl-ki.uni-osnabrueck.de/~petra/workshop/feldweg.htm>

Fellbaum, Christiane (1998): WordNet: An Electronic Lexical Database, Cambridge, MA: MIT Press.

Fellbaum, Christiane (1998): A Semantic Network of English Verbs. In: Christiane Fellbaum(ed.), WordNet: An Electronic Lexical Database, Cambridge, MA: MIT Press, 69-104.

Kunze, Claudia (1999): Semantics of Verbs within GermaNet and EuroWordNet. Seminar für Sprachwissenschaft Universität Tübingen.

Miller, Geroge A., et al. (1990): 'Introduction to WordNet: An On-line Lexical Database', International Journal of Lexicography3(4): 235-244. Also available in Postscript format from Internet : <ftp://ftp.cogsci.princeton.edu/ pub/wordnet>

Miller, Geroge A. (1998): Nouns in WordNet. In: Christiane Fellbaum(ed.), WordNet: An Electronic Lexical Database, Cambridge, MA: MIT Press, 23-46.

Miller, Katherine J. (1998): Modifiers in WordNet. In: Christiane Fellbaum(ed.), WordNet: An Electronic Lexical Database, Cambridge, MA: MIT Press, 47-68.

Princeton University Cognitive Science Laboratory: WordNet - a Lexical Database for English. <http://www.cogsci.princeton.edu/~wn/>.

Tübingen Uni GermaNet Homepage<ttp://www.sfs.nphil.uni-tuebingen.de/lsd>.

University of Amsterdam Computer Centrum Letteren: EuroWordNet. Building a Multilingual Database with WordNets for Several European Languages. <http://www.let.uva.nl/~ewn/>.

Vossen, Piek (1997): 'EuroWordNet: a Multilingual Database for Information Retrieval', Proceedings of DELOS Workshop on Cross-language Information Retrieval: 715-728. Also available in Postscript format from Internet: <http://www.let.uva.nl/~ewn/P011.ps> or in RTF format from Internet: <http://www-ir.inf.ethz.ch/ DELOS/Vossen/vossen.rtf.gz>

Zusammenfassung

Eine methodische Betrachtung für die Erstellung des koreanisch-deutschen WordNets

Nam, Yu-Sun(Wonkwang Univ.)

Das Ziel dieser Arbeit ist es, als eine methodische Grundlage zur Erstellung des koreanisch-deutschen WordNets das Grundwissen über das WordNet und einige bisherige Untersuchungen des WordNets darzulegen. Als erster Schritt wurde einige grundlegende Punkte für das WordNet im Rahmen des WordNets für Englisch in Betracht gebracht. Dabei ging es um lexikalische Hierarchie, und um semantische Relationen zwischen den Synsets(Zusammensetzen der synonymen Wörter) wie Synonymy, Antonymy, Hyponymy, Mronymy, Troponomy und Entailment.

Anschließend wurden EuroNet und GermaNet in kurzer Form vorgestellt, die auf dem Princeton WordNet basierten. EuroNet ist eine multilinguale Datenbasis mit WordNets für einige europäische Sprachen (holländisch, italienisch, spanisch, deutsch, französisch, tschechisch und estnisch). Dieses auf das Deutsch bezogenen WordNet kann wichtige Hinweise für die Erstellung des koreanisch-deutschen WordNets geben.

In Korea wurden auch verschiedene Untersuchungen über das WordNet für Koreanisch unternommen. Darunter kann insbesondere KORTERM WordNet für Koreanisch als ein umfassendes System erwähnt werden, in dem Nomen, Verben, Adjektive und Adverbien miteinander interagieren. KORTERM WordNet für Koreanisch ist eine multilinguale Datenbasis mit WordNets für einige asiatische Sprachen (koreanisch, japanisch und chinesisch) und versucht noch die weiteren Sprachen in diese multilinguale Datenbasis hineinzubringen. Nach diesem WordNet wird das koreanisch-deutsche WordNet erstellt.

[검색어] 워드넷, 유로워드넷, 게르마넷, 한국어 워드넷, 의미관계
Wordnet, EuroWordNet, GermaNet, Korterm, Semantische Relationen

남유선

570-749
전라북도 익산시 신용동 344-2
원광대학교 유럽지역어문학부 독일지역어문학 전공
nys@wonkwang.ac.kr