유로워드넷 기반의 어휘 데이터베이스 활용을 위한 한국어-독일어 ILI 대응 방법론 연구*

오장근(고려대)

1. 들어가는 말

다국어 정보처리를 위한 각종 시스템은 이를 위한 기초적인 작업의 일환으로 반드시 어휘 데이터베이스의 구축이 요구되는데, 이와 같은 경우 중요한 점은 어떻게 하면 원시언어와 목적언어가 자신의 고유한 특성들을 살리면서 동시에 기계번역이나 정보검색에 유용한 정보를 담아내느냐 하는 것이다. 이러한 점을 고려한다면 직접번역과 같은 단순 어휘의 대응이라는 차원을 뛰어넘는 고급의 어휘정보를 포함하는 어휘 데이터베이스가 구축되어야 할 것인데, 이와 관련하여 우리에게 요구되는 것은 워드넷이나 유로워드넷이 특징으로 하고 있는 어휘간의 상하위 관계를 기반으로 한 언어적 개념 구조가 주어져야 한다는 것이다.

본 논문은 특히 한국어와 독일어를 대상으로 하는 — 그러나 방법론적으로 어쩔 수 없이 일정부분 영어를 연구대상에서 배제할 수 없지만 — 다국어 정보처리 시스템에서 개별 언어 간의 센스 대응을 위한 중간언어[1] 방식의 ILI(Inter-Lingual-Index: 중간언어표지) 대응 방법론 개발에 그 목적을 두고 있는데, 이와 같은 경우 특히 미국의 워드넷(WordNet 1.5)을 기반으로 하는 유

* 본 논문은 2002 한국독일어문학회 추계학술대회에서 발표된 논문을 일부 수정, 보완한 것이다. 토론에 참여주신 한국전자통신연구원 최승권 팀장과 여러 선배 선생님들께 감사를 드린다. 또한 부족한 본고를 읽고 소중한 의견을 개진해 주신 익명의 심사자 세 분께도 감사를 드린다.

1) 자연언어(natürliche Sprache/natural language)란 기계언어나 인공언어와 대비되어 일상 사용되는 언어를 지칭하는 것으로, 이를 이해한다는 것은 사람이 아닌 기계가 자연언어를 주어진 알고리듬을 이용하여 이해하고 해석한다는 뜻으로 흔히 쓰인다. 중간언어란 'interlingua'를 옮긴 말로서 언어들 사이에서 공용되는 언어에 해당하며, 이를 달리 논리학이나 언어학에서 '언제어'나 '메타언어'로 불리어져 왔지만, 본 논문에서는 기계번역을 위해 구축된 언어 중립적인 의미표상을 뜻한다 (장석진 2001: 2 참조).

로워드넷은 다양한 언어의 개념구조 및 의미구조의 통일성을 염두에 두고 설계되었다는 점에서 우리에게 시사하는 바가 크다고 하겠다. 이제 다음에서는 간략히 다국어 정보처리를 위해 구축된 유로워드넷과 ILI의 구성원리를 살펴본 후, 이를 이용한 한국어-독일어 ILI 대응 방식에 대해 설명해 보고자 한다.

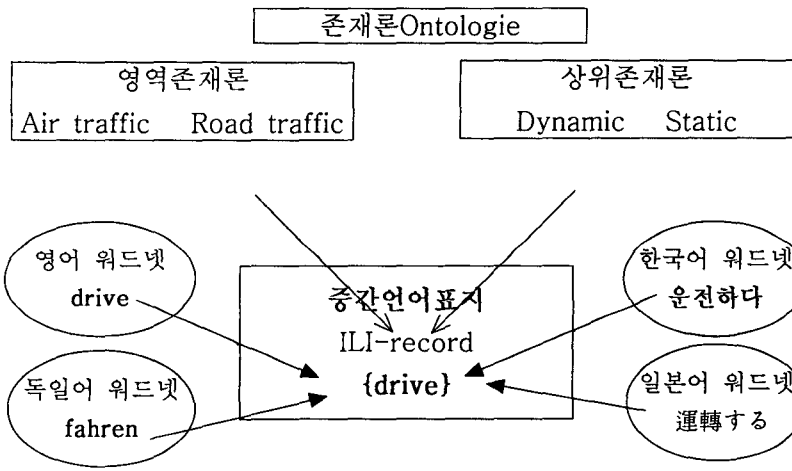
2. 유로워드넷(EUROWN)과 ILI 구성원리

2.1 다국어 어휘 데이터베이스로서 유로워드넷

유로워드넷은 프린스톤 대학의 밀러Miller를 중심으로 10여년 연구 끝에 개발된 언어심리학적 원리에 기반을 둔 온라인 데이터베이스 사전인 워드넷을 토대로 하여 유럽 8개국 언어를 대상으로 구축된 다국어 데이터베이스를 말한다. 이 작업은 1996년에 시작하여 1998년까지 지속되었다. 1차적으로 영어, 네덜란드어, 스페인어, 이탈리아어를 대상으로 하였으며, 2차적으로 독일어, 불어, 체코어, 에스토니아어를 대상으로 하여 데이터베이스를 구축하였다. 전체의 규모는 1차 작업을 통해 30,000 개념과 50,000 어휘의미를 구축하였고, 2차 작업에서는 15,000 개념과 25,000 어휘의미를 포함하게 되었다.

유로워드넷에서는 미국의 워드넷 1.5를 기초로 하여 개별 워드넷을 구축하였으며, 개별 워드넷에는 미국의 워드넷 1.5 보다 확장된 언어 내적 관계를 도입하였다. 언어 내적 관계에 자질을 첨가하였고, 품사 교차 관계를 도입하였으며, 계층관계를 세분화할 수 있는 새로운 관계를 도입하였다. 관계의 연접성(conjunction)과 이접성(disjunction), 인과관계의 비사실성과 의도성을 도입하고, 관계의 전도성(reversal)을 표시하였다. 특히, 영어 워드넷과는 교차품사관계(Cross Part of Speech Relation)를 도입하였다는 점이 큰 차이점이다. 영어 워드넷의 명사와 동사는 유사한 동의어집합('synset' oder 'synonym set')을 가지고 있음에도 품사가 다르기 때문에 연결이 되어 있지 않은 경우가 있다. 유로워드넷에서는 이러한 점을 보완하여 품사가 다른 단어 간에도 유의어나 상위어 관계를 연결시켰다. 즉, 품사가 다르고 의미가 유사한 단어들을 연결시킬 수가 있으며, 명사와 동사의 구별이 존재하지 않거나 매우 어려운 언어에도 이

와 같은 방법을 사용할 수 있다. 또 유사동의관계(near synonym relation)를 설정하여, 밀접한 관계를 가지는 단어들이지만 동의어 접합에 속할 수 없는 단어들을 연결하였다.



위의 그림에서와 같이 유로워드넷은 크게 언어내부단위(language-internal module)와 언어외부단위(language-external module)로 구성되어 있는데, 이때 언어외부단위에 해당하는 것이 ILI이다. ILI는 각각의 ILI-Record로 구성되어 있으며, 두개의 존재론(영역존재론Domain Ontology, 상위존재론Top Ontology)과 각각 연결되어 있다. 또한 이들 존재론은 ILI의 세부항목인 ILI-Record를 통해 개별 언어의 동의어집합들과 연결된다. 이럴 경우 상위존재론은 언어 독립적인 개념들의 위계적 구조이며, 영역 존재론은 주제나 스크립트에 의해서 분류될 수 있는 동일한 개념 영역에 대한 지식 구조이다. 이를 통해 일반적인 어휘와 특정 분야에서 사용되는 어휘를 구별하여 중의성을 해소할 수 있다.

1) Top concept ontology: 개별언어에 독립적인 개념들의 계층구조이다.
   Object and Substance, Location, Dynamic and Static
2) Domain label ontology: 의미를 주제에 따라 구분
   Traffic, Road-Traffic, Air-Traffic, Sports. Hospital, Restaurant.

상위존재론의 목적은 모든 워드넷 상에 나타나는 가장 중요한 개념들을 포괄하는 일종의 틀을 제공하는 것이다. 기본적인 63개의 의미구분으로 1300여개의 ILI-Record를 구분해 준다. 영역존재론은 정보검색(Information Retrieval) 분야에서 직접적으로 이용할 수 있다. 또한 언어학습 도구의 개발, 사전 출판 등에도 사용되고, 중의성 문제를 다루는데 있어서 유용하게 이용된다. 각각의 영역은 그 특성을 잘 나타내 줄 수 있는 어휘들로 구성된다. 현재 유로워드넷에서는 컴퓨터에 관련된 어휘들에 대한 작업이 이루어져 있다. 그 특징을 요약하면 다음과 같다.

1) 서로 다른 언어의 어휘의미망의 통일성과 양립성을 보장하면서 동일한 어휘 관계를 포착할 수 있다.
2) 최상위 개념을 이용하여 기본 개념들을 일관되고 체계적으로 분류한다.
3) 사용자가 다른 언어에 대한 지식이 없어도, 데이터베이스를 접근하고 다루는 것이 가능하다.
4) 언어 중립적인 존재론으로 계속 확장하는 것이 가능하다.

또한, 유로워드넷에서는 각각의 개별 언어 워드넷들의 내적 연결 구조에 관계없이 데이터 베이스 사용자들이 직접 영역에 단어들을 추가하거나 최상위 존재론을 수정할 수 있는 구조를 가지고 있다. 따라서 다른 존재론도 유로워드넷의 구조에 적용이 가능하거나, ILI와 적절히 연결만 가능하다면 같은 방법으로 수정이 가능하다. 유로워드넷은 '개별 워드넷의 구축 → 어휘 선정 → 언어 내적 관계 표시 → ILI 연결 및 등가 관계 표시'등의 과정으로 구축되었다. 이렇게 구축된 개별 워드넷들을 유로워드넷 데이터베이스에 통합하여 각각을 비교하고 재구조화하는 작업을 한 뒤, 중복된 부분은 제거하는 작업을 하였다.[2)]

2) 유로워드넷의 결과물은 Periscope을 통해서 검색이 가능하며, ELRA(the European Language Resources Association; http://www.icp.inpg.fr/ELRA)에서 배포하고, 인터넷사이트 http://www.hum.uva.nl/~ewn에서도 일부 샘플 자료 및 관련 문서들을 볼 수 있다.

2.2 ILI 구성원리

ILI는 구조화되지 않은 의미들의 리스트로, 워드넷 1.5(Princeton Wordnet 1.5)를 기반으로 하고 있다. 즉 개별 언어 워드넷의 동의어집합(synset)들은 같은 언어 내부에서 의미적 관계를 형성하고 있는 것과 동시에, ILI와도 의미적으로 연결되어 있다. ILI의 세부항목을 ILI-Record라 하는데, 이는 영어 워드넷의 동의어 집합을 그대로 가져왔으며, 해당하는 항목 기술(gloss)도 포함되어 있다. 따라서 개별 언어 워드넷을 구성하는 모든 동의어 집합들은 직접 또는 간접적으로 하나 이상의 ILI-Record와 연결된다.

워드넷은 동의어 집합인 신셋(synset), 상위어(hypernym), 하위어(hyponym), 부분어(meronym)와 같은 기본적인 의미 관계로 구축되어 있는 반면, 유로워드넷은 개개 동의어집합의 어휘의미에 대해 중간언어목록인 ILI에서 가장 가까운 의미를 찾아 등치관계를 설정하는 방법을 취한다. 영어 명사 <plant>의 워드넷과 유로워드넷의 실제 데이터 구성을 비교하면 다음과 같다:

the noun <plant> has 4 senses (워드넷1.6)

1. plant, works, industrial plant -- (buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles")
2. plant, flora, plant life -- (a living organism lacking the power of locomotion)
3. plant -- (something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant")
4. plant -- (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)

0@9@ ILI_RECORD (유로워드넷 ILI-Rekord)
1 PART_OF_SPEECH "n"
1 WORDNET_OFFSET 8894
1 GLOSS "a living organism lacking the power of locomotion & 03 1st Order Entity Group Living Natural Origin Plant"
1 VARIANTS

2 LITERAL "plant"
  3 SENSE 1
2 LITERAL "flora"
  3 SENSE 1
2 LITERAL "plant life"
  3 SENSE 1

유로워드넷의 ILI는 실제 데이터의 예에서 보는 바와 같이 고유의 일련번호(z.B. @9@), 품사정보, 워드넷 ID(offset number)가 부여되어 있으며, 또한 개별 어휘의 의미를 기술한 부분(gloss), 그리고 실제로 어떠한 영어 어휘로 나타나는지(literal)에 대한 항목이 포함되어 있다. 예를 들어 @9@에는 세 개의 LITERAL이 있는데, 각각 'plant', 'flora', 'plant life', 이들은 모두 같은 ILI를 갖는 동의어집합이다.

ILI를 이용하는 이러한 다국어 데이터베이스 구축방법은 여러 가지 이점을 제공하는데, 다음은 기계번역(Machine Translation)에서 ILI를 이용할 때의 장점들이다(Nirenburg 1987 참조):

1) 개별 언어간의 대응 관계(다대다 대응)를 일일이 명세할 필요가 없다. 각각의 개별 언어들은 단지 ILI와의 연결 관계만을 설정하면 된다.
2) 기존의 다른 언어들과의 연결에 신경 쓸 필요 없이, 새로운 개별 언어를 쉽게 추가할 수 있다.
3) ILI는 언어간의 대응 관계를 보다 효율적이고 정밀하게 보여주기 위해 수정이 가능하다.

결국 ILI를 통한 다국어 기계번역은 우선 의미해석 조건에 따라 입력어의 어휘의미가 결정되면 영어 워드넷(WN1.6)에서 대응치를 확인하게 되며, 그 대응치가 확인될 경우 영어 워드넷이 지니고 있는 ILI의 번호를 한국어나 그 밖의 언어 워드넷에 제시해 주고, 이를 통해 언어간의 의미적 대응관계가 설정되게 된다.

3. ILI를 통한 독일워드넷(GermaNet)의 유로워드넷 연결

3.1 GermaNet의 구성과 언어내적 관계

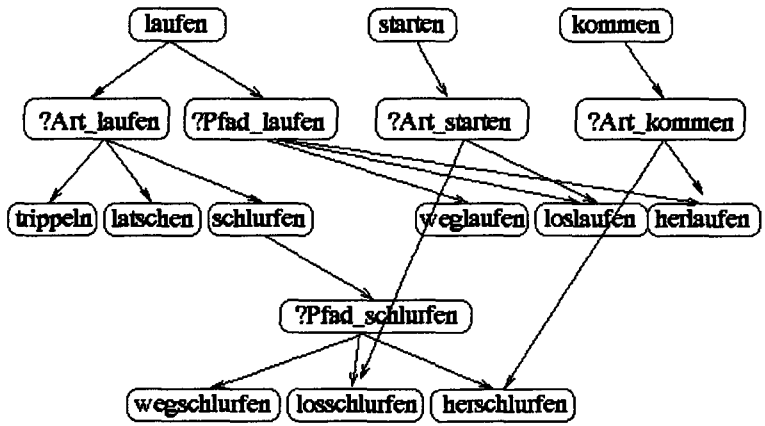
GermaNet은 이미 위에서 언급한 것처럼 유로워드넷을 구축하기 위해 프린스톤 워드넷1.5에 근거하여 구축된 독일어의 어휘-의미론적 개념망으로서, 1990년대 말부터 튀빙엔 대학을 중심으로 연구, 발전되었다.[3] GermaNet은 독일어의 기본 어휘를 동의어, 상위어, 하위어, 부분어 같은 의미론적 관계에 기초하여 구축하였는데, 현재 명사, 동사, 형용사와 같은 기본 품사들을 모델화시켰으며, 모두 25000단어의 어휘의미를 포함하고 있다. 이 작업은 현재 계속 보완되고 있는데, 2001년 10월 12일 현재 모두 41777개의 신셋이 구축되어 있다. GermaNet은 그러나 몇가지 원칙적 관점에서 프린스톤 워드넷과 구분된다. GermaNet은 완전히 새로운 자료로 구성되었다는 것이다. 다시 말해 워드넷의 단순한 번역물도 아니고, 개별 사전이나 시소러스에 기초해서 구축된 것도 아니다. 다음은 명사 'Atmungsorgan'과 동사 'laufen', 'starten', 'kommen'의 의미관계를 예시한 것이다;

Atmungsorgan 'respiratory organ'
a. HYPERONYM Organ 'organ'
b. SYNONYM Atemsystem 'repiratory system'
c. HYPONYM Lunge 'lungs'
d. MERONYM Luftröhre 'trachea'

---

3) 그러나 GermaNet의 발전은 이미 1996-1997에 펠트벡(H. Feldweg)과 힌리히스(E. Hinrichs)가 바덴-뷔르템베르크 주의 재정지원으로 행한 "의미적-어휘적 중의성 제거를 위한 연구 자료들과 방법론"이라는 SLD-프로젝트에서 시작되었다고 할 수 있는데, 이는 GermaNet의 근간이라고 할 수 있는 독일어 어휘의 의미소분석이 행하여졌기 때문이다 (Hinrichs et. all 1998).

3) 정보검색분야에서 사용되는 기존의 시소러스는 '후조합 검색을 위해 설계된 자연언어 용어들의 통제어휘집', '상위 및 하위 개념 사이의 전후관계를 명백하게 하기 위하여 공식적으로 조직·통제된 색인어의 어휘집', '구조적인 측면으로 보면 의미적, 종속적으로 관련된 용어들의 통제된 동적인 어휘집' 등으로 규정되고 있다 (이재윤/김태수 1999, 233 참조).



GermaNet에서의 개별 어휘들은 소위 동의어집합인 신셋으로 구성되어 있으며, 어휘적-개념적 관계를 통하여 결합되어 있는 것이다. 특히 GermaNet은 동사의 의미관계에 초점을 두었는데, 동사부분에서는 동의어, 반의어 관계와 더불어 상위어관계가 매우 중요한 구성성분이 되고 있으며, 그 외에도 부분관계, 사역관계, (의미론적) 파생관계, 하위사건관계, 함축관계, 선택제약관계, 일반적 다의어관계 등이 GermaNet의 하위범주로 구분되고 있다. 예를 들어 독일어 동의어집합 {öffnen 3, aufmachen 2}는 의미적으로 상위개념인 {wandeln 4, verändern 2, ändern 2}와 관련되어 있고, 4개의 하위개념인 {aufschieben 1}, {aufstoßen 2}, {aufbrechen 1}, {aufsperren 1}과도 연결되어 있다. 뿐만 아니라 사역적 의미의 기동동사로서 '열리다'의 의미를 지니는 {öffnen 1}, {aufgehen 1}과도 연결되어 있다. GermaNet의 구조가 프린스톤 워드넷의 구조와 구분되는 중요한 관점을 정리하면 다음과 같다:[5]

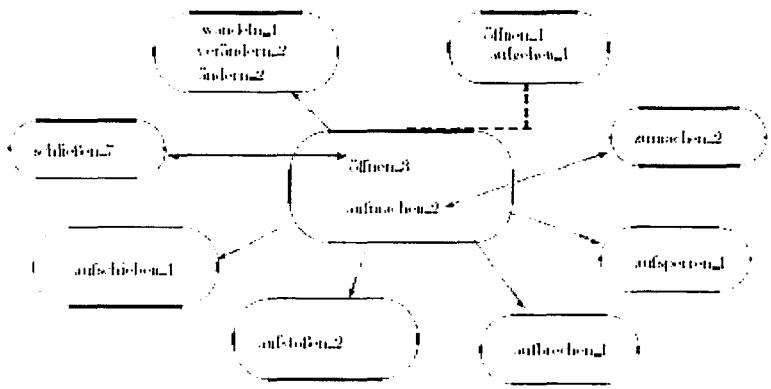
1. GermaNet에서는 작위적 개념들(어휘공백)이 논리적으로 의미있는 하위부류를 형성하기 위하여 사용되고 있으며, 그 자체로 명백하게 표시되고 있다.
2. 프린스톤 워드넷에서는 단지 산발적으로 쓰이고 있는 교체분류Kreuz-

5) Feldweg의 'GermaNet - ein lexikalisch-semantisches Netz für das Deutsch'의 개요 참고.

klassifikation가 GermaNet에서는 중요한 질서성분이 된다.

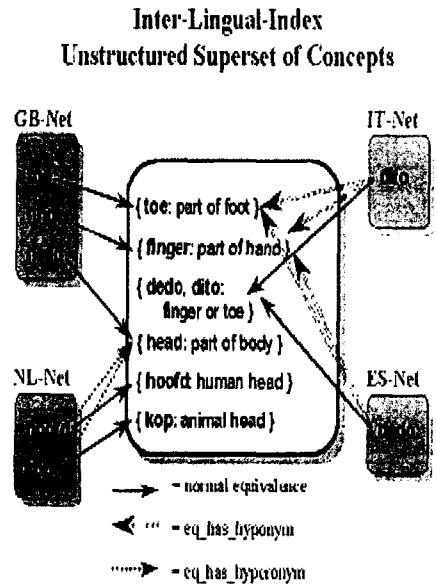
3. 보통의 다의어는 빈번히 출현하는 원형적인 형식의 목록과 더불어 '일반화'라고 할 수 있는 독특한 관계를 통해 모델화되고 있다.

다음은 독일어 동의어집합 {öffnen 3, aufmachen 2}의 GermaNet 내에서의 어휘적-개념적 관계를 도표로 나타낸 것이다:



3.2 독일워드넷GermaNet의 유로워드넷 대응관계

유로워드넷에서의 ILI와 개별 언어 워드넷의 대응관계를 살펴보면, 기본적으로는 각각의 개별 언어의 동의어집합과 ILI의 관계는 언어 내적인 동의어집합과 이 동의어집합간의 관계와 동일한데, 이를 그림으로 나타내면 다음과 같다:



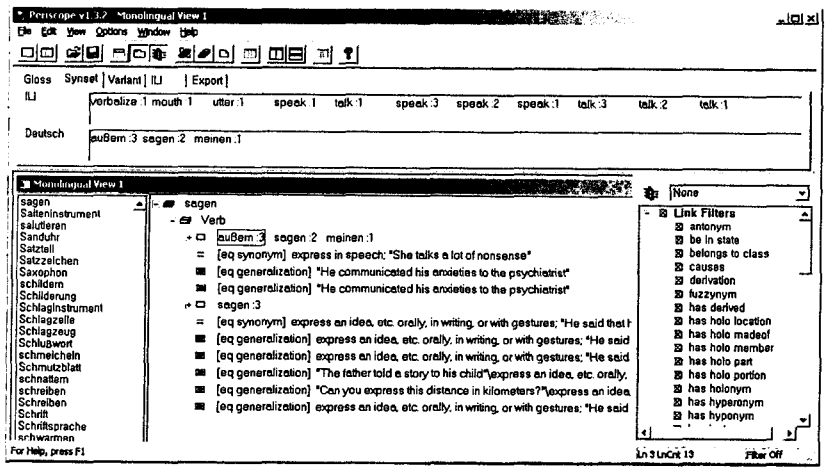
다국어 대응에 있어서 가장 중요한 관계는 등가관계(Equivalence relations)이다. 이는 두개의 동의어집합간의 일대일 대응관계를 나타내는데, ILI를 통해서 다국어를 대응할 때 나타나는 다국어 대응의 패턴은 단순하지가 않다. 유로워드넷에서의 다국어 대응은 4가지 패턴으로 작동된다.

(1) EQ_SYNONYM
(2) EQ_NEAR_SYNONYM when a meaning matches multiple ILI-records simultaneously
(3) EQ_HAS_HYPERONYM when a meaning is more specific than any available ILI-record
(4) EQ_HAS_HYPONYM when a meaning can only be linked to more specific ILI-records

(1)은 가장 이상적인 경우로 두 언어의 의미와 등치관계를 가지는 센스가

ILI에 실재하는 경우이다. 이 경우는 두 어휘가 ILI를 통하여 직접 링크된다. 예를 들어 독일어의 'Sportbekleidung'과 영어의 'sports equipment'는 동일한 동의어집합으로, 어휘의미 "sport garment"로 인해 직접적으로 연결될 수 있다.

(2)는 한 언어의 단일한 신셋이 동시에 여러 개의 ILI-records와 링크되는 경우이다. 예를 들어 독일어 동사 sagen은 워드넷에 적어도 2개 이상의 센스와 링크된다. 이 경우처럼 2개의 이상의 어휘의미와 동시에 연결되는 독일어 sagen을 기계번역에 응용할 때는 실제로 이들 중 어느 어휘의미를 택해서 번역을 해 주는가 하는 문제가 발생한다. 우리는 이 문제에 대한 답을 아직 갖지는 않은 상태지만, 잠정적으로 이들 중 어느 것을 택해도 문제가 되지 않는 상황이므로, 탐색되는 최초의 어휘의미를 기준으로 번역을 한다고 가정한다. 아래는 독일어 동사 sagen의 유로워드넷 대응관계를 표현해준다.



(3)은 두 언어의 워드넷 사이에 어휘적 공백이 있을 때 사용되는 방법이다. 예를 들어, 독일어의 Glühwein은 영어에는 대응되는 어휘가 없다. 이것은 일종의 포도주로 성탄절을 즈음하여 마시는 향료를 넣은 적포도주이다. 이 경우, ILI-records에는 독일어 Glühwein은 ILI-records의 EQ_HAS_HYPERONYM에서 wine/red wine의 신셋으로 링크된다.

(4)도 (3)과 마찬가지로 두 언어 사이의 어휘적 공백의 예이다. 그러나 그 방향이 반대인 경우다. 예를 들어, 스페인어에는 dedo라는 명사가 있다. 이것은 finger와 toe 모두를 가리킨다. 그러나 영어에는 이에 대응되는 어휘가 없다. 이와 같은 경우, ILI-records는 EQ_HAS_HYPONYM에서 finger 혹은 toe 모두에 동의어집합으로 링크된다.

그렇다면 이러한 유로워드넷 시스템의 문제점은 무엇일까? 특히 유로워드넷을 기반으로 하여 한국어와 같은 동아시아 언어를 포함한 월드워드넷 WorldWordNet(WWN)의 구축에 있어서 우리에게 관심을 끄는 것은 기존의 워드넷에 어떠한 문제점이 있고, 어떻게 변형을 줄 것인가를 살펴보는 것이며, 더 나아가 어떻게 하면 중간언어 번역 시스템에 활용할 수 있겠는가 하는 것이다.

1) 무엇보다 우선적으로 고려될 수 있는 것은, 유로워드넷에서의 ILI는 언어 내부적인 의미관계를 고려하지 않는다는 것이다. 예를 들어 현재 유로워드넷에서는 고유의 일련번호, 품사정보, 영어 워드넷 ID(offset number), 개별 어휘의 의미를 기술한 Gloss, 그리고 어떠한 영어 어휘로 나타나는지(Literal)에 대한 항목이 포함되어 있을 뿐 다른 언어의 의미관계에 대해서는 전혀 언급이 없다. 따라서 이것을 그대로 중간언어 기계번역 시스템에 사용하는 데는 한 가지 문제가 있다.

영어의 literal 정보뿐만 아니라, 관련된 모든 언어, 즉 한국어, 독일어 등의 모든 표제어 정보가 등재되어 있어야 한다. 그리고 언어간 표제어의 품사가 다를 경우는 그에 대한 정보도 요구된다.

예를 들어 영어 [fish]의 유로워드넷 출력을 보면 다음과 같다:

1 WORDNET_OFFSET : 4999301_w
1 PART_OF_SPEECH : n
1 GLOSS : the flesh of fish used as food & 03 13 1st Order Entity Comestible

Form Function Object Origin Substance
1 VARIANTS
2 LITERAL : [fish]
3 SENSE : 1

2) 워드넷의 개념구조를 그대로 따르는 유로워드넷과는 달리, 동아시아 언어인 한국어의 개념분류에 워드넷의 분류체계를 그대로 적용하는 것은 바람직하지 않다. 예를 들어 '날씨'라는 동사의 범주에서 독일어의 개념은 regnen, donnern, schneien 등의 단어로 표현되지만 한국어에서는 '비가 오다', '천둥이 치다', '눈이 오다' 등으로 단어만으로 표현될 수 없으며, 단어로 표현될 경우 이들은 대부분 '맑다', '흐리다' 등의 형용사로 나타나고 있다. 이와 같이 워드넷의 개념 분류체계를 한국어에 그대로 적용하기는 어려우므로 한국어 워드넷을 개발하기 위해서는 일부를 고쳐서 적용하거나 완전히 새로 개발하는 방법을 써야 할 것이다.

4. 한국어-독일어 ILI 대응에 대한 방법론적 제안

ILI의 목적은 개별언어 워드넷의 동의어 집합들간의 의미 관계를 보여주는 것이다. 원래 ILI-Rocord 자체를 워드넷 1.5에서 그대로 가져 온 것이기 때문에, 그 자체적으로도 의미적 연관 관계를 유지하고 있고, 그 관계를 사용자가 볼 수도 있기는 하다. 하지만 유로워드넷에서의 ILI는 그러한 내부적인 의미관계를 고려하지 않는다. 따라서, 워드넷 1.5에서 내부적 의미 관계에 관련된 정보는 ILI에는 포함되어 있지 않다. 완전히 개별 언어에서 독립된 존재론을 따로 구성하는 것은 그 구조가 매우 복잡해지고, 작업량이 지나치게 많아지기 때문에 불가능하고, 이처럼 기존에 구성된 자료를 이용하여 다국어 대응을 구성하는 것이 효율적일 것이다.

따라서 유로워드넷에 한국어 등의 동아시아 언어의 워드넷이 포함될 월드워드넷의 ILI는 이를 이용하는 사용자에게 일련번호(Offset number) 외에도 출발언어와 목적언어 또는 그 밖의 다른 언어와의 의미적 상관관계에 대한 정보를

제공함으로서 이 시스템이 보다 효과적으로 작동될 수 있게 하고자 한다. 이제 본 논문에서 제안하고자 하는 한국어-독일어 ILI 대응관계의 설계는 다음과 같이 진행될 수 있다:[6]

① 개별언어의 어휘의미정보를 비교한다.

'가다'의 어휘의미

의미	EnWN 대응치	GermaNet 대응치
sense1 -을 향해 이동하다	go, verb 2	gehen, verb 1
sense2 죽다	die, verb 1	sterben, verb 1
sense3 꺼지다	go out, verb 4	ausgehen, verb 6
sense4 변질되다	go bad, verb 1	verderben, verb 1
sense5 시간이 흐르다	go, verb 16	gehen, verb 8
sense6 노력이나 힘이 미치다	require, verb 1	gehen, verb 9
sense7 값이 나가다	cost, verb 1	kosten, verb 1
sense8 지탱하다	go, verb 15	gehen, verb 10
sense9 가능하다, 이루어지다	*have in mind*	*imstande sein*

② 어휘의미 분석이 끝나고 나면, 각국 출발언어의 워드넷에서는 개별어휘의 다양한 의미에 대해 번호가 결정된다.

영어워드넷 go (EnWN)	한국어워드넷 가- (KoWN)
1. travel, go, move	1. 진행하다, 나아가다, 움직이다
2. go, proceed, move	2. 지나다, 경과하다, 지나가다
3. go, go away, depart	3. 떠나가다. 떠나다, 가버리다
9. proceed, go	5. 변하다, 부패하다.....

6) 여기에서의 워드넷 센스 자료는 가설적으로 주어진 것이다. 이 논문의 ILI 활용 방안은 최호철/한정한 (2002)의 '다국어 기계번역을 위한 중간언어 모형과 방법론 연구' (고려대학교 민족문화연구원)의 보고서 내용을 독일어에 맞게 변용한 것이다.

독일어워드넷 gehen
(GermaNet)
1. fortbewegen, gehen, vorgehen
2. gehen, reisen, abgehen
5. gehen, verlassen, aufgeben
7. funtionieren, gehen....

③ 위의 어휘의미 번호는 다시 월드워드넷의 일련번호 @123456과 상호링크(inter-link)된다.

1 월드워드넷 WorldWordNet_Offset: 123456_w
<EnWN go 2>
<KoWN 가 1>
<GermaNet gehen 1>

* 이 경우 <EnWN go 2>는 영어 'go'와 영어워드넷인 EnWN에서의 센스번호를 의미한다.

④ ③번 작업 후, 월드워드넷의 일련번호는 중간언어 표현으로 바뀐다. 다시 말해 대역어 영어의 동사 'Go'가 논항 'He'를 가질 경우, 월드워드넷 일련번호는 Go의 센스번호와 He의 센스번호가 결합하여 다음과 같은 중간언어 표현으로 표출될 수 있을 것이다.

WorldWordNet_Offset: 123456(12) (이럴 경우, 123456 ⇒ Go 12 ⇒ He)

⑤ 만약 출력어가 독일어라면, 월드워드넷에 있는 독일어 대역어 정보를 이용하여 독일어 대역어와 그 대역어의 센스 번호를 알 수 있다. 따라서 위 중간언어는 각각 아래와 같이 대역된다.

123456 ⇒ gehen 2 12 ⇒ er 1

이 다음의 독일어 생성단계는 gehen 2의 <태voice>를 확인한 후 출력어 생성

단계로 들어간다.

이처럼 월드워드넷에 추가되는 개별 언어의 어휘는 영어 대역에서 원어의 뜻에 가까운 동의어집합을 찾아 해당하는 ILI 어휘의미 번호(Sense-ID)를 달고 등재하게 된다. 그러나 만약 독일어 워드넷의 어휘의미 정보와 다른 언어의 어휘의미 정보가 불일치할 경우, 이는 ILI에 새로운 어휘의미 정보를 등재함으로써 해결을 모색할 수 있을 것이다. 예를 들어 한국어 동사 '먹다'의 다의와 대응관계를 독일어의 동사와 비교해 설명하고자 한다. '먹다'에 대응하는 <essen>은 duden에 따르면 다음 3가지의 어휘의미 정보가 등재되어 있다:

<essen>의 어휘의미 분석
Sense1 essen-(feste Nahrung zu sich nehemen)
Sense2 essen-(als Nahrung zu sich nehemen)
Sense3 essen-(durch Essen in einen bestimmten Zustande bringen)

이에 따르면 독일어 <essen>의 기본 의미성분(Sense1)은 '고체음식을 취하는' 행위이다. 한편, 한국어의 '먹다'는 훨씬 대상의 폭이 넓어 독일어 <essen>이 지니는 대상물의 의미속성보다 상위에 속하는 개념을 취한다: '술을 먹다', '죽을 먹다', '공기를 먹다' 등. 이제 세종사전의 기록을 따르면 <superEntry>-'먹다'의 <entry>-'먹다'에는 다음의 5가지 어휘의미 정보가 등재되어 있다:

<먹다>의 어휘의미 분석
N0 N1-을 V
a. <먹다-1> N0=인물 N1=사물(밥/떡/과자/젖)
'목이 메어 밥을 먹지 못한다.'
b. <먹다-2> N0=인물 N1=식사(아침/점심/저녁/....)
'아침을 안 먹었더니 하루 종일 몸에 맥이 하나도 없다.'
c. <먹다-3> N0=동물(누에) N1=(뽕잎)
'누에는 뽕잎을 먹고 자란다.'

d. <먹다-4> N0=동물(염소/소) N1=(여물/풀)
'소가 여물을 잘 먹는다'
e. <먹다-5> N0=동물(닭) N1=(모이)
'닭이 모이를 아주 잘 먹는다.'

이러한 어휘의미의 세분은 독일어와는 달리 한국어의 경우 주어, 목적어의 논항에서 해당 어휘의 의미속성이 세분되어 표시되고 있음을 알 수 있다. 문제는 한국어와 독일어의 어휘의미를 의미속성의 어느 계층에서 동일한 동의어집합으로 묶느냐의 결정에 있다. 또한 이처럼 언어마다 다른 의미정보들의 대응관계를 찾아내서 유로워드넷과 ILI에 어휘의미 번호를 매기는 것이 다국어 정보처리를 위한 데이터베이스 구축 작업의 기초가 될 것이다. 따라서 한국어 '먹다'와 이에 대응하는 영어, 독일어의 동사들은 월드워드넷에 해당 어휘의 의미정보별로 ILI 번호(Sense ID)를 달고 기재된다. 그럴 경우 의미관계가 유사한 어휘의미에게는 동일한 ILI 번호를 주고, 다른 어휘의미에게는 새로운 ILI 어휘의미 번호를 줌으로써 기계번역의 의미 정확성과 적용 가능성을 향상시키게 된다. 이러한 ILI 정보는 다시 월드워드넷의 중간언어사전ILD에 입력되며, 여기서는 워드넷(1.6)의 등재정보와 어휘의미 번호를 잠정적으로 사용하게 된다.

그러나 월드워드넷에 참여하는 서로 다른 언어들 사이에는 서로 상이한 개념분류체계로 인해 출발언어에서는 한 단어로 표시되는 것이 목적언어에서는 두 품사의 결합으로 표시해야 하는 경우가 빈번히 출현하는데, 예를 들어 영어 동사 rain은 '비가 오다'로, 독일어동사 frühstücken은 '아침밥을 먹다'로 표현되고 있다. 유로워드넷의 경우, 이러한 문제는 ILI와 개별 언어 워드넷의 다양한 대응관계(동의어, 유사동의어, 상위어, 하위어 등)를 통해 설명하고자 했지만 대부분 올바른 번역에 실패를 겪게 된다. 따라서 한국어-독일어 정보처리를 위한 ILI 대응관계의 활용방안을 제안하고자 하는 본 논문에서는 이를 해결하기 위해, 이들 표현들을 하나의 센스를 지닌 '자유결합 또는 연어정보'로 처리하고자 하며, 이들의 데이터 베이스는 다음과 같이 정리된다:

```
<coll>
        <comb_v n=1>
                <comb_v_exp>비가 내리다</comb_v_exp>
                <comb_v_transE>rain</comb_v_transE>
                <comb_v_transDt>regnen</comb_v_transDt>
                <comb_v_transC>.....</comb_v_transC>
                <comb_v_transJ>.....</comb_v_transJ>
        </comb_v>
```

이와 같은 처리방법은 '관용어'와 '은유'의 경우에서도 적용될 수 있다:

```
<IdiomGrp>
        <idiom n=1>
        <idiom_exp>물 샐 틈 없다</idiom_exp> /*철저하고 빈틈이 없다
                <idiom_transE>.....</idiom_transE>
                <idiom_transDt>.....</idiom_transDt>
                <idiom_transC>.....</idiom_transC>
                <idiom_transJ>.....</idiom_transJ>

<MetaphorGrp>
        <meta n=1>
        <meta_exp>물(을) 먹다</meta_exp> /*어떤 일에 실패하다
                <meta_transE>.....</meta_transE>
                <meta_transDt>.....</meta_transDt>
                <meta_transJ>.....</meta_transJ>
```

6. 맺음말

각 나라마다 독립적으로 구축된 어휘자원들에서 정보를 추출하고, 이를 다

시 결합하여 어휘를 비교하는 유로워드넷과 같은 다국어 어휘 데이터베이스 모형 개발에서 중심적인 문제는 다양한 언어의 연결방법일 것이다. 이를 위해 유로워드넷은 ILI를 사용하고 있는데, ILI는 워드넷1.5로부터 출발하였지만, 개별 워드넷의 연결이 잘 이루어지도록 변형되었다. 즉 워드넷1.5에 존재하지 않는 새로운 어휘의미가 개별 워드넷에 나오면 ILI에 새로운 의미정보를 추가하고, 주석을 개선하였다. 그러나 이러한 유로워드넷의 ILI체계도 한국어와 같은 동아시아언어를 포함하는 월드워드넷을 구축하는 과정에서 볼 때, 몇 가지의 문제점들이 지적될 수 있을 것이다. 이를 위해 본 논문에서는 우선 유로워드넷 모형과 ILI 구성원리를 살펴보았으며, 그런 다음 독일워드넷(GermaNet)이 유로워드넷에 연결되는 기존의 ILI 모델을 비판·검토한 후, ILI의 변형을 통해 한국어와 같은 다양한 언어를 연결하는 보다 나은 방법론을 제안하고자 하였다.

본 논문은 사실 어떤 새로운 학문적 논리나 주장을 펼치기 위해 쓰여진 것은 아니다. 뿐만 아니라 상당 기간 연구·발전되어 온 완성된 이론도 아니고, 어쩌면 단순한 가설에 지나지 않을 수도 있을 것이다. 그럼에도 불구하고 이러한 졸고를 발표하게 된 것은, 다른 언어학의 분과에서 활발하게 논의되고 있는 다국어 정보처리 분야에 대한 이론적 접근에 독어학의 분과도 관심을 지니고, 또한 능동적으로 이러한 연구에 참여했으면 하는 바램이 있었기 때문이었다.

참고문헌

강범모/이유선/차재은 (2002): 다국어 어휘 데이터베이스 구축 방법론 연구 및 모형개발. 고려대학교 민족문화연구원.

권재일/김현권 (1998): 용언 전자 사전구축 지침서와 표본사전. 21세기 세종계획 1차년도 결과물.

남경완 (2000): "다의 분석을 통한 국어 어휘의 의미 관계 연구" 고려대 석사논문.

이두영/최석두 (1993): 지능형 정보검색에 관한 연구 -별책부록- 시소러스, 한국통신 연구개발단 93 장기기초연구과제 최종보고서 '93U019.

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

장석진 (2000a): '통합문법: 주제·초점과 시점' <언어와 정보사회> 1:145-175. 서강대학교 언어정보연구소.

_____ (2001): 자연언어 이해를 위한 중간언어 표상. 학술원논문집 40, 1-37.

최호철/한정한 (2002): 다국어 기계번역을 위한 중간언어 모형과 방법론 연구. 고려대학교 민족문화연구원, 2000년도 학술진흥재단 중점연구소 지원과제 보고서.

한국과학기술원 (1996): 『국어정보베이스』 (제2차년도 최종보고서).

Buitelaar, Paul (1998): Corelex: Systematic Polysemy and Underspecification. Ph.D Dissertation, Brandeis University.

Dorr, Bonnie (1993): 'The Use of Lexical Semantics in Interlingual Machine Translation', Machine Translation 7, 135-193

Dowty, David (1991): 'Thematic Proto-roles and Argument Selection', Language 67. 547-619.

Feldweg, Helmut (2001): GermaNet - ein lexikalisch-semantisches Netz für das Deutsche. Awailable 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.

Foskett, Douglsa J. (1980): 'Thesaurus' in: A. Kent, H. Lancour, and J. E. Daily (eds.) Encyclopedia of Library and Information Science, 30: 416-463. (New York: Marcel Dekker, Inc.)

Jackendoff, Ray S. (1990): Semantic Structures. Cambridge, MA: MIT Press.

Katz, Jerrold. and Jerry A. Fodor (1963): 'The Structure of a Semantic Theory', Language 39: 2. 170-210.

Kavi, Mahesh (1996): 'Ontology Development for Machine Translation: Ideology and Methodology', Technical Report MCCS 96-292, Computing Research Laboratory, New Maxico State University.

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>

Nirenburg, Sergei, Victor Raskin and Allen Tucker (1987): 'The Structure of interlingua in TRASLATOR', in Nirenburg, ed. 1987:90-113.

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

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 methodologische Untersuchung der koreanisch-deutschen ILI-Verbindung zur Anwendung der auf dem EuroNet basierten lexikalisch-semantischen Datenbasis

Oh, Jang-Geun(Korea Univ.)

EuroNet ist eine multilinguale Datenbasis mit WordNets für einige europäische Sprachen (holländisch, italienisch, spanisch, deutsch, französisch, tschechisch und estnisch). Die WordNets werden genauso wie das amerikanische WordNet für Englisch (Princeton WordNet, Miller et al. 1990) in Synsets (Zusammensetzen der synonymen Wörter) mit grundlegenden lexikalisch-semantischen Relationen zwischen ihnen ausgedrückt strukturiert. Jedes WordNet stellt also ein einzigartiges innersprachliches System für die lexikalischen und konzeptuellen Relationen dar. Zusätzlich werden diese auf dem Princeton WordNet basierten WordNets (z.B. GermaNet) mit einem Inter-Linguale-Index (kurz, ILI) verbunden. Über diesem Index werden die Sprachen zusammengeschaltet, damit zu gehen ist möglich, von den Wörtern in einer Sprache zu den ähnlichen Wörtern in jeder möglicher anderen Sprache. Der Index gibt auch Zugang zu einer geteilten Top-Ontologie von 63

semantischen Unterscheidungen. Diese Top-Ontologie stellt einen allgemeinen semantischen Rahmen für alle Sprachen zur Verfügung, während sprachspezifische Eigenschaften in den einzelnen WordNets beibehalten werden. Die Datenbasis kann, unter anderen, für einsprachige und multilinguale Informationsretrieval benutzt werden.

In der vorliegenden Arbeit handelt sich also um eine methodologische Untersuchung der koreanisch-deutschen ILI-Verbindung zur Anwendung der auf dem EuroNet basierten lexikalischen, semantischen Datenbasis. Dabei werden einzelnen Lexeme in koreanischen, deutschen WordNets zunächst mit Hilfe der Sense-Analyse semantisch differenziert, und dann durch lexikalische und konzeptuelle Relationen(ILI) miteinander verbunden. Die Equivalezverbindungen dienen, sprachspezifische Konzepte zum ILI abzubilden. Sie werden von einem anderen Synset der möglichen Relationen aus der Euronet-Spezifikation genommen. Wenn es keinen ILI-Rekord gibt, der ein direktes Equivalenz zu einem gegebenen Konzept darstellt, kann das Konzept in der Frage über EQ-Near-Synonymie, EQ-Hyperonymie oder EQ-Hyponymie Relationen verbunden werden.

[검색어] 유로워드넷 독일어워드넷 중간언어표지 동의어집합
EuroWordNet GermaNet Inter-Lingua-Indext Synset

오장근

153-821
서울시 금천구 시흥2동 268-46
탑스빌Apt. 101동 1304호
domplatz@hanmail.net