인지과학, 제21권 제1호
Korean Journal of Cognitive Science
2010, Vol. 21, No. 1, 95~126

인간언어공학에의 활용을 위한 이종 개념체계 간 사상 - 세종의미부류와 KorLexNoun 1.5 -*

배 선 미 임 경 업 윤 애 선†
부산대학교 인문학연구소 부산대학교 정보컴퓨터공학부 부산대학교 불어불문학과

본 연구에서는 인간언어공학에서의 활용을 위해 매우 이질적인 세종전자사전의 의미부류(SJSC)와 KorLexNoun 1.5(KLN)의 상위노드 간의 사상을 목표로, '의미 입자(sense grain)가 작은 개념체계(fine-grained ontology)' 간 귀납적이며 상향적인 수동 사상 방법론을 제안하였다. 동시에 이종 자원 간의 사상에 있어 각 의미체계의 이질성 때문에 발생하는 여러 가지 문제점을 살펴보고, 그 해결방안도 제안하였다. 두 이종 개념체계 간의 사상 방법은 SJSC의 단말 노드와 KLN의 Least Upper Bound(LUB)를 기본단위로 하여, 첫째, 어휘 분포를 이용하여 사상 후보군을 결정하고, 둘째, 계층 관계와 정의문과 용례를 이용하여 후보군들 간의 정확한 의미구분을 하며, 셋째, 상·하위-자매노드에 SJSC의 적정술어 및 정의문을 적용하여 LUB의 단계를 결정하고, 넷째, 양 의미체계의 계층관계를 비교함으로써 SJSC의 단말 노드와의 사상 여부를 판단하며, 마지막으로 KLN의 오류 및 전문용어 후보군은 사상에서 제외하였다. 이와 같이 본 연구에서는 단계별 사상 준거의 설정에 있어 각 의미체계에 기술되어 있는 다양한 언어정보를 적극 이용하였는데, 이는 세밀한 수동 사상의 장점이라 할 수 있다. 본 연구에서 제안한 방법으로 사상한 결과, SJSC의 474개의 단말 및 비단말 노드와 KLN의 신셋(synset) 간에는 중복을 제외하고 6,487개의 LUB가 사상되었으며, 각 LUB의 하위노드를 포함해서는 모두 88,255개의 KLN 신셋이 사상되어 전체적으로는 97.91%가 사상되었다. 본 연구의 결과는 정교한 한국어 통사 및 의미 분석에 활용될 수 있을 것이다.

주제어 : 의미부류, 어휘의미망, 언어공학, 세종전자사전, KorLex, 상향적 귀납적 사상 방법론, 온톨로지 수동 사상

---

* 이 논문은 2007년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2009-0083761). 이 논문은 2009년 '한글 및 한국어 처리 학술대회에서 발표했던 내용을 보완한 것임.
† 교신저자: 윤애선, 부산대학교 불어불문학과/인지과학협동과정, 연구분야: 전산언어학
E-mail: asyoon@pusan.ac.kr

1. 서 론

본 연구는 한국어를 대상으로 기구축된 매우 이질적인 개념체계인 세종 의미부류체계(이하, SJSC)와 한국어 명사어휘의미망 KorLexNoun 1.5(이하, KLN)의 상위 노드 간의 수동 사상의 준거와 절차를 제안하고, 두 이종 자원 간의 사상에 있어 각 개념 구조와, 분류의 상이함으로 인해 발생하는 문제점을 살펴보고, 그 해결방안을 모색하는 것을 연구 목표로 한다.

이와 같은 이종개념 체계의 사상은 다음과 같은 목적과 배경을 갖는다. 첫째, SJSC가 밝히고자 했던 한국어의 어휘의미구조[1]와, Princeton WordNet(이하, PWN)[2]을 참조로 한 KLN에 여전히 영향을 미치는 영어 어휘의미구조[3]를 비교함으로써 공통점과 차이점을 파악할 수 있고, 이를 바탕으로 언어 독립적인 개념체계를 구축하는 데 기여할 수 있다[4]. 둘째, 향후 KorLex의 용언에 기술되어 있는 문형정보와 세종전자사전의 용언의 격틀 정보를 통합 구축하여 구문분석에서 이용할 때, 논항의 의미정보를 결정하는 세종 의미부류 이용가능성을 확보하여 추후 논항의 일반화된 선택제약규칙을 기술할 수 있다. 셋째, 전문가의 수작업에 기반한 이와 같은 연구는 향후 이종 자원의 자동 사상 연구에서 사상 방법론 연구에도 크게 기여할 수 있다. 넷째, 세종 의미부류 체계와 KLN이라는 두 이종 자원의 사상을 통해 두 자원의 결함을 상호 보완하여 보다 완전한 언어자원으로써 구문분석이나 의미분석 등에 이용하여 실용성의 증대를 꾀할 수 있다.

본 연구의 사상 대상은 SJSC의 노드와 KLN 1.5의 상위 노드이다. 보다 구체적으로는 SJSC 단말 노드를 중심으로 KLN에서 공통 상위 노드들 중 가장 구체적인 개념/의미를 갖는 최하위 공통 상위 노드(Least Upper Bound Node, 이하 LUB)를 찾아 사상하는 것으로, 사상 대상의 개념/의미 입자의 크기가 작다.[1)]

전문가의 수작업에 의해 이종 개념분류 체계 간의 사상에 관한 대표적 연구로는 PWN과 SUMO[2)]를 사상한 사례가 있으며, 대부분은 자동 사상에 대한 연구가 주를 이루고 있다. [5]에서 제안하는 PWN의 신셋[3)]을 SUMO의 개념에 수작업으로 사상

1) PWN과 KLN의 계층별 노드 수는 본고 표 2를 참고하라.

2) [5]의 연구에 사용된 SUMO(Suggested Upper Merged Ontology)는 2003년 2월 버전으로 1,000개의 개념을 가진 상위 온톨로지이다.

하는 방법은 의미의 등가관계(synonymy)를 우선으로 고려하고, PWN의 신셋이 SUMO에서 적절한 용어를 찾지 못하였을 때는 그 신셋보다 상위개념(hypernymy)의 SUMO 개념에 사상하며, 의미의 등가관계나 포함관계가 아닌 경우에는 SUMO에서 가장 근접한 인스턴스(instance)를 찾아 사상한다. 이러한 방법은 [5]에서도 밝힌 바와 같이, SUMO의 개념 입자의 크기가 크므로, PWN에 있는 대부분의 상위 단계의 개념들은 SUMO에서 등가 개념들을 용이하게 찾을 수 있다. 따라서 두 자원 간의 사상에 있어 실질적으로든 이론적으로든 큰 문제를 야기하지 않는다. 의미 입자가 큰 개념체계인 SUMO에 의미 입자가 작은 PWN을 사상한다는 점에서 본 연구에서 대상으로 하는 SJSC에 KLN을 사상하는 것과 유사하다. 하지만, SJSC는 SUMO에 비해 의미 입자가 훨씬 작아 더 세밀한 사상 준거를 필요로 한다. 또한 SJSC의 경우, 의미부류가 어휘 간 선택제약을 표현하는 데 사용되고, 따라서 적정술어[4]가 의미부류의 설정에 있어서 중요한 하나의 기준이 되기 때문에, SJSC에서는 각 개념이 맺는 관계가 KLN과는 매우 다르며, 하위단계로 내려갈수록 SJSC와 KLN의 의미체계와는 상이한 의미부류가 많아 SUMO와 PWN의 사상 방법을 본 연구에 적용하기가 어렵다.

또한 자동 사상에 대한 연구로는 이개어 온톨로지를 구축하기 위해 다국어 정보검색과 기계번역을 목적으로 코퍼스에 기반하여 중국어 개념체계 HowNet과 PWN을 연결한 연구[6]가 있으며, 다국어 사전을 구축하여 기계번역에 이용할 목적으로 Goi-Taikei 개념체계와 일-영 워드넷(JWN)을 연결한 연구[7]가 있다. 그 이외에도 한국어 워드넷을 구축하기 위해 WSD(Word Sense Disambiguation)를 이용하여 한-영 MRD(Machine Readable Dictionary)와 워드넷을 자동으로 사상한 연구[8]와 중국어 워드넷을 구축하기 위해 중-영 MRD와 워드넷을 사상한 연구[9] 등이 있다. 그러나 이와 같은 자동 사상 방법의 적용은 [10]에서도 보여 주듯이, 본 연구의 대상과 같이 의미 입자의 크기가 크게 다를 경우에는 사상의 범위와 정확도에서 한계

3) PWN에서는 개념을 표상하는 최소 단위를 '동일한 어휘의미를 가지는 동의어 집합(신셋)'으로 규정하고 있다. 보다 자세한 것은 2.2를 참조하라.

4) [1]에 의하면, 적정 술어(appropriate operators)란 특정 의미 영역에 속하는 어휘들과 제한적으로 결합하는 속성을 가지므로 의미부류를 정의하는 데 형식적인 근거가 되는 어휘들을 말한다. 예컨대, 동사 '먹다'나 명사 '맛'처럼 음식을 칭하는 명사들과 특징적으로 결합하는 속성으로 인해 <음식>이라는 의미부류 설정의 형식적인 기준이 되는 어휘들을 <음식> 부류의 적정 술어라고 한다.

가 있다.

본 논문은 다음과 같이 구성된다. 2절에서는 사상 대상이 되는 SJSC와 KLN 1.5의 특성을 각각 살펴본 다음, 3절에서는 이 두 이종 자원 간의 사상에 있어 사상 절차와 준거를 제안하고, 4절에서는 사상 결과 및 개념 체계의 이질성에서 비롯된 문제점과 해결 방안에 대해 논의한다. 5절에서는 연구내용을 요약하고, 향후 연구 방향을 살펴본다.

2. 대상 언어자원의 특성 분석

SJSC와 KLN 1.5는 모두 계층적 의미구조를 가지며, 정의에 따라 상위 노드의 개념/의미는 자식 노드에 상속된다는 공통점을 갖고 있다. 하지만, 구축 철학이 다른 이질적인 개념/의미 체계를 사상할 때 구체적인 문제가 발생하는 이유는 ① 각 개념/의미가 맺는 관계가 상이하고, ② 개념/의미 입자의 크기가 다르며, ③ 구축자의 모국어에 경도되어 있기 때문이다.

2.1 세종 전자사전의 의미부류체계

SJSC는 개념을 지칭하는 메타 용어 집합으로 각 노드의 입자의 크기가 크며, 세

표 1. SJSC의 노드(의미부류) 수

최상위 부류명	최대 계층 수	최상위 노드를 포함한 전체 노드 수	단말 노드 수
구체물	7	198	150
집단	5	29	24
장소	4	54	40
추상적 대상	5	151	118
사태	7	191	152
합계	-	623	484

종전자사전에 수록된 체언, 용언 등 구체적인 어휘의미의 부류를 구분하고, 어휘 간 선택제약을 표현하는 데 사용한다. 최상위 노드를 기점으로 최소 2단계에서 최대 7단계의 깊이를 갖는 계층적 구조를 갖는다. 이렇게 구성된 의미부류의 수는 현재 최상위 노드를 포함하여 총 623개이며, 5개 최상위 부류 별 단말 노드 수는 표 1과 같다.

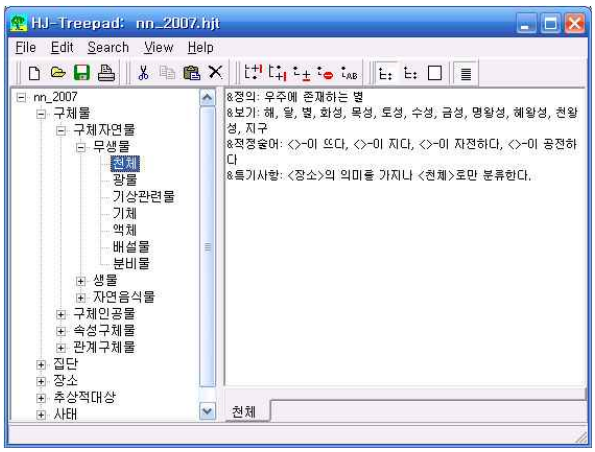

그림 1. <천체>[1]의 TreePad에서의 기술 예

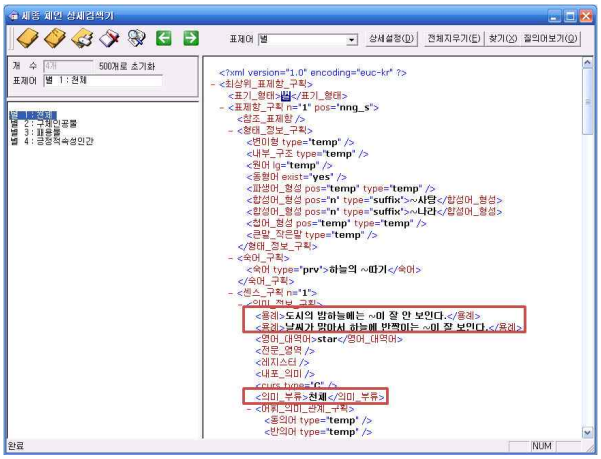

그림 2. SJND의 표제어 기술의 예-'별'

각 의미부류의 특성을 추출할 수 있는 정보로는 ① 그림 1처럼 세종 TreePad viewer를 통해 볼 수 있는 '정의, 적정술어, 보기'와 계층적 구조, ② 그림 2처럼 세종전자 상세사전에서 어휘의미를 구분하는 데 사용하는 의미부류명과 표제어 쌍과 예문이다. 2007년도에 최종 공개된 세종 전자사전에서 체언은 동형이의어 수준에서 25,458개의 표제어가 세종 의미부류에 따라 35,854개의 의미로 구분되어 있다[11].

2.2 KorLexNoun 1.5

이에 비해 '개념=어휘의미'로 정의한 PWN과 KLN 1.5에서 각 노드를 나타내는 신셋은 실제 1개 이상의 어휘(의미)로 구성되며, 의미 입자의 크기가 아주 작다. 하지만, 상위 노드에 위치한 어휘일수록 개념/의미 입자가 크며, 따라서 메타 용어로 사용되는 어휘가 분포되어 있다. PWN Noun 2.0과 KLN 1.5는 그림 3처럼 9개의 최상위 노드(unique beginer)로 시작하며, 25개 의미범주(lexicographer's files)로 구분한다. 계층별 하위노드의 수는 표 2와 같다.[5)]

그림 3. KLN의 최상위 노드

각 노드의 의미적 특성을 추출할 수 있는 정보로는 그림 4, 5에서 볼 수 있는

5) SJSC와는 달리 PWN과 KLN에서는 2개 이상의 상위 노드의 의미속성을 1개의 하위노드가 상속받을 수 있기 때문에, SJSC의 특성을 파악하는 데 필요한 표 1과 같은 통계는 큰 의미를 갖지 못한다.

표 2. PWN Noun 2.0과 KLN 1.5 계층별 노드 수

계층	PWN 명사 2.0	KLN 1.5
1	9	9
2	158	157
3	1,307	1,653
4	4,489	6,033
5	10,297	13,129
6	17,536	19,236
7	15,336	18,079
8	12,225	13,802
9	7,605	8,053
10	4,793	4,714
11	2,501	2,305
12	1,444	1,256
13	852	733
14	477	429
15	415	346
16	206	164
17	39	36
계	79,689	90,134

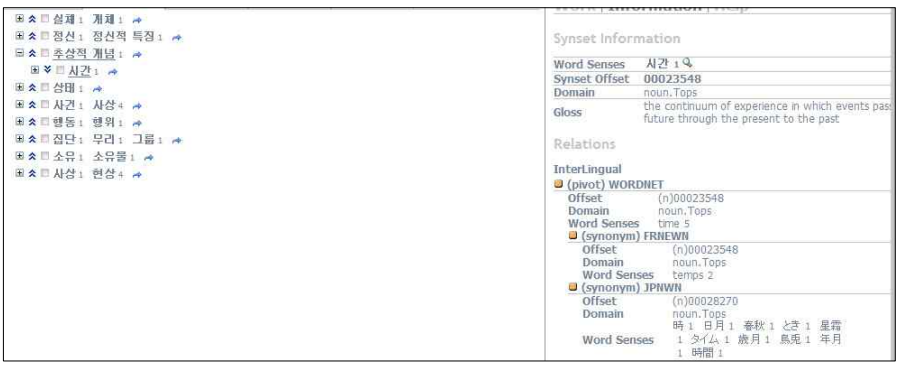

그림 4. PWN 2.0과 KLN 1.5의 공통 의미정보

그림 5. KLN 1.5의 뜻풀이 정보

것과 같이 ① PWN과 KLN에 공통으로 제공되는 영어 정의문, 예문, 신셋, 반의어, 전의/분의 관계 등과 ② KLN에 제공되는 한국어 신셋과 정의문(표준국어 대사전 뜻풀이)이 있다. 이때 KLN 1.5의 90,134개의 신셋에는 102,358개의 어휘의미가 포함되어 있다[3].

3. SJSC와 KLN의 사상 방법론

이상에서 살펴본 것처럼 매우 이질적인 체계와 의미정보를 가진 SJSC와 KLN을 어떤 기준에서 어느 정도 범위까지 사상할 것이며, 이를 위해 어떤 방법론을 사용할 것인가? 이는 제 인간언어에 공통으로 존재하는 보편적 개념과 특정한 개별언어에 유효한 개념을 구분하고, 동시에 두 개념체계의 사상을 통해 얻어지는 통합적 정보를 인간언어공학의 제 분야에 활용하고자 하는 본 연구의 목적과 밀접히 관련된다. 3절에서는 이를 위한 사상의 전제조건과 준거 및 절차를 검토한다.

3.1 사상의 전제조건

SJSC와 PWN/KLN는 한국어와 영어를 분석대상으로 하였지만, 모두 연역적/하향적 방식으로 구축된 개념체계이다[3, 7]. 두 개념체계를 비교 통합하는 데, 구축할

때와 동일한 방식을 채택하는 것은 차이점만을 확인할 뿐이다. 또한, 두 개념체계는 계층적 구조를 갖고 있으나, 상·하위 관계가 모두 엄격하게 IS-A 관계로 구성되지는 않으며, 자매 노드 간 의미 크기 및 유형의 불균형도 자주 발견되므로, 완전 자동 사상 방법을 적용하는 데는 정확도와 범위에 한계가 있다[10]. 따라서 본 연구에서는 두 개념체계의 특성을 드러내는 언어정보를 이용하여 귀납적/상향적 방식으로 공통성을 찾고자 한다. 이를 위해 부분적으로 통계와 자동 사상 결과를 이용하여 사상의 후보군을 제시하되, 전문가가 다른 언어정보를 이용하여 수동 검증하거나 직접 사상하는 방식을 채택한다.

귀납적/상향적 방식을 이용한 사상의 기본 단위는 SJSC의 단말 노드로 설정하고, 이를 기준으로 KLN에서 동일한 개념(군)을 찾는다. 이는 SJSC와 KLN의 계층적 구조에서 부모 노드의 의미적 자질이 자식 노드에 상속됨을 전제로 하며, 이러한 상속성 원칙에 위배되는 노드는 사상의 범위에서 제외한다. 이는 본 연구의 결과로 사상된 SJSC와 KLN의 노드(들)의 상위 노드와 하위 노드는 추후 자동적으로 사상하고자 위해서다.

메타 용어로 구성된 SJSC와 실제 어휘 집합인 KLN은 추상성의 정도와 개념/의미 입자의 크기에서 큰 차이를 보이므로, 직접 비교할 수 있는 대상이 되지 못한다. 따라서 세종 체언상세사전(이하, SJND)에 들어 있는 표제어-의미부류 정보 쌍을 추출하여, 사상의 대상이 되는 의미부류에 속한 SJND의 표제어와 KLN의 신셋을 구성하는 어휘를 비교하고, 이 정보를 바탕으로 KLN에서 해당 어휘형의 상위 노드를 SJSC의 단말 노드와 사상한다.

3.2 사상의 절차 및 준거

본 연구에서 제안하는 두 이종 개념체계 간의 세밀한 사상 방법론은 SJSC의 단말 노드와 KLN의 LUB를 기본단위로 하여, 그림 6과 같은 사상 절차를 따른다.

사상의 준거로는 ① 사상 여부 및 사상 후보군을 제시해 주는 어휘의 분포, ② 사상 후보군의 정확한 의미구분을 하는 상·하위-자매 관계 및 정의문과 용례, ③ LUB의 단계를 결정하기 위해 상·하위-자매 노드에 SJSC의 적정술어 및 정의문의 적용, ④ SJSC의 단말 노드와의 사상 여부를 결정하기 위한 양 체계의 계층관계

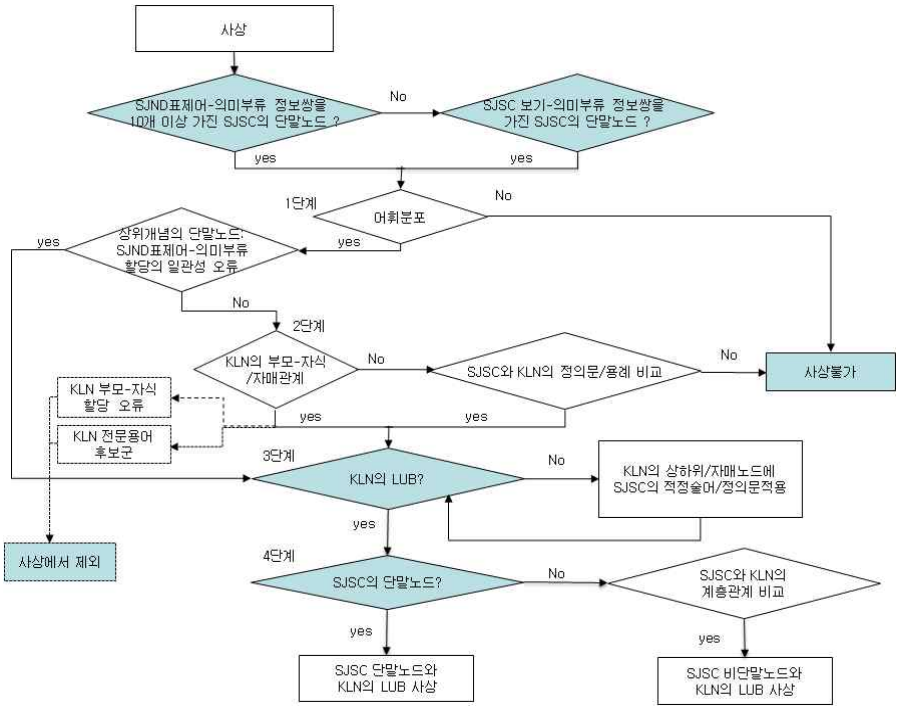

그림 6. SJSC와 KLN 1.5 간의 사상 절차

비교, 그리고 ⑤ 사상에서 제외할 신셋을 결정해 주는 준거 등이 있다. 이러한 준거를 적용한 예는 다음에서 좀 더 면밀히 살펴보자.

3.2.1 사상 후보군 결정의 준거: 어휘의 분포

두 이종 개념체계의 사상에 있어 가장 먼저 고려해야 할 것은 SJSC의 단말 노드가 KLN의 신셋 중 사상이 가능한 신셋이 있는지, 만약 사상이 가능한 후보 신셋군이 있다면 그 범위가 어느 정도인지를 파악하는 것이다. 이를 위해, 본 연구에서는 SJND에서 체언을 의미분류 하는 데 사용된 표제어-의미부류 정보 쌍을 이용하여 SJND의 표제어가 KLN에서 얼마나 분포하는지 그 어휘분포를 살펴본다. 이 과정은 SJSC의 의미부류로 분류된 SJND 표제어와 KLN의 신셋 간의 자동 사상 결과를 이용하여 사상의 후보군을 제시함으로써, 동형이의 수준에서 SJND 표제어의

어휘 분포가 있는지를 파악한다[6]. 이때, SJND의 표제어-의미부류 정보 쌍이 10개 미만이거나 전혀 없다면 표제어-의미부류 정보 쌍이 충분치 않은 것으로 간주하고 SJSC의 각 의미부류에 기술되어 있는 '보기'를 추가로 이용하여 보기-의미부류 정보 쌍의 KLN에서의 어휘 분포를 살펴본다[7]. 만약 SJND의 표제어나 SJSC의 보기의 KLN에서의 어휘분포가 전혀 없다면 해당 SJSC의 단말 노드와 KLN의 사상은 불가능하다. 그러면, SJND 표제어의 어휘분포와 SJSC의 보기의 어휘분포의 예를 각각 들어 보자.

예를 들면, 세종 의미부류 <기상관련물>로 분류되는 SJND의 표제어로는 '꽃비, 농무, 눈, 눈발, 눈송이' 등이 있는데, KLN과의 자동 사상을 이용하여 그 분포 여부를 파악할 수 있다. 그림 7은 세종 의미부류 <기상관련물>과 이 의미부류에 속하는 표제어 정보 쌍의 KLN에서의 어휘분포 일부를 보여주는 예이다.[8]

그림 7의 왼쪽 박스는 SJSC의 TreePad의 계층구조와 정의, 적정술어, 보기 등의 정보를 담고 있고, 중앙 박스는 해당 의미부류로 분류되는 SJND의 표제어 수와 목록을 KLN과 사상된 것과 그렇지 않은 것을 구분하여 제시하고 있다. 오른쪽 박스는 KLN의 신셋과 사상된 수와 사상된 노드들을 제시하고 있다. 그림 7에 의하면, <기상관련물>에 속하는 SJND의 표제어 수는 의미부류 자신을 포함하여 모두 26개이고, 그 중 KLN과 사상된 표제어는 17개이며, 동형이의어 수준에서 36개의 KLN의 신셋과 사상되어 있음을 알 수 있다. 예를 들면, '농무 2, 눈 9, 눈발 2, 눈송이 1, 먹구름 3, 백설 1, 비구름 1, 빗물 1, 빗방울 1, 서리 1[9]'등은 KLN에서 사

6) SJND의 표제어-의미부류 정보 쌍이 가장 많은 단말 노드는 <직업인간>으로 536개의 정보 쌍이 있으며, 그다음으로 많은 의미부류는 <직위인간>으로 303개의 정보 쌍이 있지만, 대부분의 단말 노드에 속하는 의미부류의 정보 쌍은 수십 개에서 많아도 200개를 넘지 않는다.

7) SJSC의 484개의 단말 노드 중에서 SJND의 표제어-의미부류 정보 쌍이 전혀 없는 노드는 4개의 노드가 있으며, 10개 미만인 노드는 105개가 있다. 후자는 '보기'를 이용하여 추가로 어휘분포를 파악할 수 있다. SJSC의 단말 노드 가운데 '보기'가 있는 노드는 모두 441개이며, 293개의 노드가 10개 미만의 보기가 기술되어 있고, 10개~19개의 보기가 기술되어 있는 노드는 126개, 20~29개의 보기가 기술되어 있는 노드는 19개, 30개 이상의 보기가 기술되어 있는 노드는 2개이다.

8) 그림 7은 본 연구를 위해 만든 사상 워크벤치의 일부를 보여주는 화면이다.

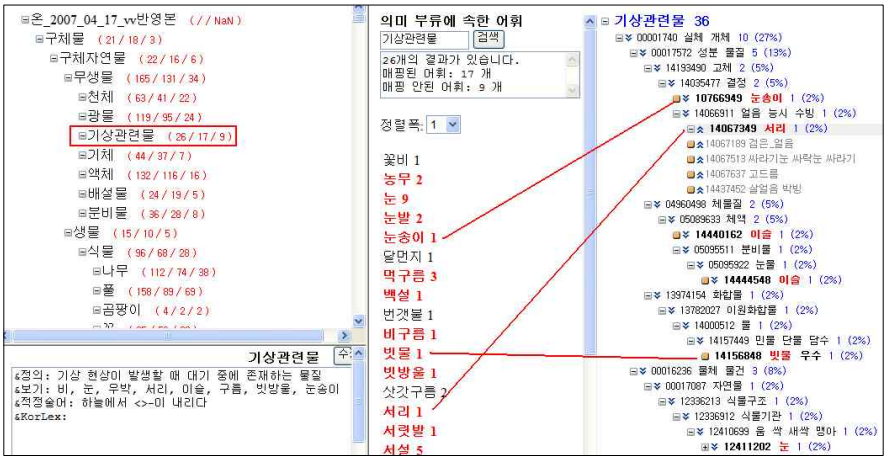

그림 7. SJND 표제어-의미부류 정보 쌍의 KLN에서의 어휘분포의 예

상이 가능한 신셋 후보군이 있는 표제어들이다[10]. 물론, SJND 표제어-의미부류 정보 쌍의 KLN에서의 어휘분포가 있다고 해도 이와 같은 자동 사상을 이용한 어휘분포는 동형이의어 수준에서 후보군의 제시단계이므로, 3.2.2절 이하에서 소개하는 다른 준거들을 이용한 정확한 의미구분, 사상에서 제외할 노드의 결정, LUB의 단계의 결정, 두 체계 간의 계층 비교를 통한 SJSC의 단말 노드와의 사상 여부 등을 고려하여 최종적으로 KLN의 신셋과 사상해야 한다.

하지만, SJND 표제어-의미부류 정보 쌍이 충분하지 않은 경우에는 SJSC의 각 의미부류에 기술되어 있는 '보기'의 어휘 분포를 살펴본다. 예를 들어, 의미부류 <종료>에는 '절판 1, 절품 1, 종식 1, 종언 1, 출하 2, 종료 0' 등 5개의 표제어-의미부류 정보 쌍이 있지만, 이 중 4개만 KLN에서 찾을 수 있으므로 사상의 근거로

9) SJND에서 표제어 뒤에 있는 숫자는 의미부류에 따른 '센스(SJND에서 중분류에 해당하는 다의어 구분 메타 용어)' 구분의 표시로, SJND에서는 그 숫자만큼 다의어 구분이 되어 있다.

10) 지면상의 제약으로 SJND 표제어-의미부류 정보 쌍의 자동 사상 결과를 모두 제시하기는 어렵지만, 예를 들면, '꽃비 1, 달먼지 1, 번갯불 1, 샛갓구름 2' 등은 KLN과 자동 사상되는 결과가 없고, '눈송이 1'에는 신셋 '10766949 눈송이'가, '서리 1'에는 신셋 '서리 14067349'가, '빗물 1'에는 신셋 '1156848 빗물' 등이 자동 사상되어 있다.

는 충분하지 않다. 이와 같은 경우에는 '보기'를 이용하여 추가적으로 KLN에서의 어휘분포를 검토한다. 그리고 동형이의어의 구분이나 LUB의 결정 등의 사상절차는 표제어의 어휘분포를 검토하는 과정과 동일하다.[11] SJND의 표제어-의미부류 정보 쌍이 전혀 없는 노드 중 <무게단위>는 보기를 이용하여 사상이 가능하다. 그러나 1개의 표제어-의미부류 쌍이 있지만 KLN과의 자동 사상이 전혀 없는 <미각적행위>와 3개 노드 <고가도로>, <밝기단위>, <음악기호>는 '보기'를 이용해도 KLN에서 사상되는 신셋이 없어 이들은 아예 사상 자체가 불가능하다[12].

그런데, SJND의 표제어/보기-의미부류 정보 쌍의 어휘분포를 살펴보면 SJND 표제어-의미부류 쌍이 수십 개로 충분히 사상이 가능하지만 표제어들의 성격을 볼 때 일관성이 없는 경우들이 있다. 즉, 같은 의미부류로 분류된 SJND의 표제어들이 해당 단말 노드가 아닌 그 상위 노드로 분류되어야 하거나 심지어 다른 의미부류로 분류되어야 마땅한 표제어들이 해당 단말 의미부류로 분류되어 있는 것이다[13]. 이와 같은 경우는 SJSC에서 주로 개념이 큰 의미부류일 때 나타나는 현상으로 단

갑충 1 강 2 경지 3 계보 2 계열 1 계통 2 고유계정 1 공유폴더 1 과 2 과거 3 금융계 1 기종 1 기종 2 기타 2 논의 2 단속대상 1 단위 1 동종 1 문 5 물장군과 1 바늘꽃과 1 밖 2 반열 1 범죄형 2 범주 1 부류 1 분꽃과 1 사회갈래 1 산업별 1 세목 2 속 9 수종 1 어류 1 어류 2 어족 1 어족 2 어종 1 어종 2 업종 1 연령층1 영장류 1유 2 유형 2 음 3 음 4 인간계 1 인기순위 1 일반 1 일종 2 장르 1 정치제도1 정치질서 1 정치체제 1 정치풍토1 조류 2 조류 4 족보 2 종 4 종류 1 종류별 1 종목 1 종별 2 주류 4 지렁이강 1 지연 1 차종 1 친고죄1 카테고리 1 코너 3 타입 1 테두리 3 토종 1 특정 1 특종 1 팔각목 1 포유류 1 폴더 1 품종 1 현재 2 화폐단위 1

그림 8. <범주>로 분류되는 SJND의 표제어의 예

11) 본고 그림 6을 참조하라.

12) 예를 들면, SJSC 의미부류-보기쌍인 '<고가도로>-고가도로, 고가차도', '<밝기단위>-룩스, 촉', '<음악기호>-높은음자리표, 사분음표, 셋잇단음표, 숨표, 도도리표, 반올림표, 반내림표', '<미각적행위>-음미' 등은 보기마저 KLN에서 사상되는 신셋이 전혀 없어 아예 이들 단말 노드는 사상이 불가능하다.

13) 추후 이들 표제어의 의미할당 오류는 SJSC에서 수정 · 보완되어야 할 사항이다.

그림 9. KLN의 신셋 '범주'와 하위노드

말이 아닌 1-3단계 정도의 상위 노드에서 많이 나타나는데, <방법>이나 <범주>와 같은 2단계 상위개념의 단말 노드에서 발견된다. 이 경우에는 KLN이 분류학적 기준에 의해 분류가 잘 되어 있다면, 일관성 없이 분류된 일부 SJND 표제어들의 KLN에서의 자동 사상된 신셋 후보군들을 일일이 고려하지 않고, 해당되는 SJSC의 단말 의미부류 (예를 들면, <범주>)에 대응되는 KLN의 의미부류에 기반하여 KLN의 적합한 LUB단계(예를 들면, 신셋 '범주 2 범주 3')를 정하여 사상하도록 한다[14].

3.2.2 의미구분의 준거

3.2.1에서 SJND의 표제어-의미부류 정보 쌍과 KLN의 신셋과의 자동 사상을 이용하여 동형이의어 수준에서 사상 후보군을 한정하였으나 사상해야 할 신셋을 결

14) 본고 그림 8과 9를 참조하라. 예를 들면, 그림 8에서 '범죄형, 친고죄' 등과 '계보, 계열, 계통, 종류, 종, 유형' 등의 SJND의 표제어를 동일한 의미부류인 <범주>로 나란히 분류할 수 있을 지 의문스럽기 때문에, 이와 같은 경우에는 SJND 표제어들의 자동 사상 결과에 대해 의미구분을 하기 위한 KLN의 부모-자기 노드나 SJSC와 KLN의 정의문이나 용례 비교는 무의미하며, KLN의 어느 LUB와 사상시킬 것인지가 문제가 되므로 그림 6의 사상절차에서 바로 KLN에서의 LUB를 선택하는 단계로 넘어간다.

정하기 위해서는 정확한 의미구분이 필요하다. 이를 위해, 일차적으로 상 · 하위 관계 및 자매관계를 고려하고, 부모와 자식 노드만으로 의미구분이 되지 않는 경우에는 이차적으로 SJSC의 각 의미부류에 기술되어 있는 정의문과 용례 및 KLN의 각 어의에 기술되어 있는 표준국어대사전의 정의문과 용례를 비교·검토하여 정확한 의미구분을 한다.

3.2.2.1 의미구분의 준거 I: 상 · 하위 관계 및 자매관계. 일차적으로 SJND 표제어/보기-의미부류 정보 쌍의 사상 후보군들의 부모 또는 자식 노드 및 자매 노드를 의미구분의 준거로 삼아 의미 중의성을 해소한 다음에, SJSC의 의미부류에 적합한 KLN의 신셋을 찾는다. 이때, SJSC의 적정술어나 정의문이 있다면[15] 이를 동시에 상 · 하위 노드에 참조적으로 적용할 수도 있다.

예를 들어, 의미부류 <종교분파>로 분류되는 어휘를 사상할 때에 신셋 '09811005 청교도'는 그 신셋의 의미나 적정술어만으로는 종교분파로 간주하여 이 의미부류에 사상할 수 있다. 그러나 그 부모 노드인 '09139230 금욕주의자, 고행자'와 자매 노드인 '09972181 주상고행자'가 모두 종교인을 뜻하므로 이 경우에는 청교도가 '청교도인'을 의미한다. 따라서 '09811005 청교도'를 <종교분파>에 사상하지 않는다[16]. 자식 노드가 준거가 되는 한 예를 보면, 의미부류 <혈연집단>으로 분류되는 표제어 '가정'이나 '집'은 신 셋 '13687178 가정, 집'과 사상 가능하나 그 자식 노드가 '농어가, 전업농가, 초상집, 소가' 등으로 이들은 <혈연집단>으로 볼 수 없으므로 사상하지 않는다.

15) SJSC의 단말 노드 484개 중 정의와 적정술어/분포가 있는 노드 수는 다음 표와 같다.

정보	적정술어	분포	정의	적정술어/분포 ∩정의
단말 노드 수	230	63	212	115

16) 이때, 정의와 적정술어/분포는 추가로 의미구분을 확인시켜 주는 역할을 할 수도 있으나 반드시 필요한 것은 아니다. 예를 들면, 신셋 '09811005 청교도'의 부모 노드인 신셋 '09139230 금욕주의자, 고행자'를 <종교분파>의 '정의'와 '적정술어'에 적용시켜 보아도 맞지 않음을 확인할 수 있지만, 부모-자식 노드와 자매 노드와의 의미관계만을 보아도 충분히 의미구분을 할 수 있다.

3.2.2.2 의미구분의 준거 II: 정의문과 용례. 부모와 자식 노드만으로 의미구분이 되지 않을 때에는 이차적으로 SJSC의 각 의미부류에 기술되어 있는 정의문과 용례 및 KLN의 각 어의에 기술되어 있는 표준국어대사전의 정의문과 용례를 비교・검토하여 의미구분을 한다.

우선, SJSC의 의미부류에 기술된 정의와 KLN의 각 어의에 기술된 정의가 서로 부합하지 않을 때에는 사상하지 않는다. 예를 들어, 의미부류 <장소>의 자식 노드인 <바다>[17]로 분류되는 표제어 '바다'와 사상된 KLN의 신셋 '08762162 바다'는 신셋 '00022625 장소'의 자식 노드인 '0878162 권역, 지역'의 자식 노드로 부모-자식의 관계를 고려할 때 사상이 가능한 것으로 보인다. 그러나 의미부류 <바다>의 정의는 '지상 장소로 둘러싸이지 않은 물 장소'이며, 신셋 '08762162 바다'의 정의는 '달이나 화성 표면의 검게 보이는 부분'으로 서로 정의가 들어맞지 않으므로

표 3. 언어정보를 기반으로 한 사상 후보 쌍의 의미 등가성 판단

사상 후보 노드		의미 구분을 위한 언어 정보	
		유형	제공된 언어 정보
SJND <전기 3>		용례	- 센스_구획 n="1"> - <의미_정보_구획> <용례>자세한 내용은 ~의 사항을 참고하시오.</용례> <용례>~를 검토했을 때, 이것은 사실이 아닙니다.</용례> <의미_부류>텍스트의부분</의미_부류>
KLN	'144386733 전기'	정의문 용례	어떤 대목을 기준으로 하여 그 앞부분에 씀. 또는 그런 기록. [예] 자세한 내용은 {전기의} 사항을 참조하시오. [예] 종교 사상으로서 {전기} 여러 저술에 못지않은 중대한 의의를 가지고 있다. «안병욱, 사색인의 향연»
	'14438733 전기'	정의문	전하여 듣고 기록함

17) <장소>의 자식 노드는 <물장소>이며, <물장소>의 자식 노드는 <바다>이다.

이 경우는 사상하지 않는다.

SJSC에는 정의가 없는 의미부류도 많은데, 이 경우는 SJND의 용례와 KLN의 용례를 면밀히 비교 · 검토하여 들어맞지 않으면 사상하지 않는다. 예를 들어, 의미부류 <텍스트의 부분>[18]로 분류되는 표제어 '전기 3'은 동형이의 수준에서 KLN에서 많은 신셋과 사상이 되는데, 그중에서 부모 노드를 '기록'으로 갖는 신셋으로는 '144386733 전기'와 '14438733 전기'가 있다. 이 같은 경우에 부모 노드가 같으므로 이 둘 모두가 사상이 가능한 것으로 보인다. 그러나 SJND에 기술된 <전기 3>의 용례와 KLN에 기술된 '144386733 전기'와 '14438733 전기'의 정의문과 용례를 표 3과 같이 비교·검토한다면 <텍스트의 부분>과 KLN의 신셋 '144386733 전기'는 사상해야 하지만, '1443873 전기'는 사상할 수 없음을 알 수 있다[19].

3.2.3 LUB 결정의 준거: 상 · 하위 · 자매 노드에 적정술어/정의문의 적용

SJSC의 단말 의미부류를 중심으로 KLN에서의 사상 후보군을 설정하고, 후보군 중에서 상 · 하위관계나 자매관계를 고려하고 이차적으로 정의문과 용례를 이용하여, 이들의 의미구분을 했다면 과연 그 후보 신셋을 LUB로 해서 사상할 것인지 아니면 더 상위나 하위 노드를 LUB로 해야 할지, 사상할 LUB의 단계를 결정해야 한다. 물론, 자동 사상되어 의미구분이 된 KLN의 신셋이 자매 노드나 자식 노드가 없는 경우라면 의미구분 이후에 바로 그 신셋을 LUB로 해서 사상할 수 있지만, 비교적 상위 단계에 사상된 신셋이라면 상 · 하위 및 자매 노드가 다소 복잡하여 상·하위 및 자매 노드, 자신 중에서 LUB를 결정해야 한다. 이때에는 각 의미부류에 기술되어 있는 적정술어와 정의문 등의 정보를 이용하여 사상된 KLN의 신셋 노드뿐만 아니라, 그 신셋의 자식 노드와 자매 노드, 그리고 부모 노드 모두에 이들을 적용해 서 LUB를 결정하게 된다. 이러한 과정은 적절한 LUB를 찾을 때까지 부모, 자식, 자매 노드, 자매 노드의 자식 노드 등을 계속해서 순환하면서 반복된다[20].

---

18) <텍스트의 부분>은 정의가 없는 단말 의미부류이다.

19) 본고 표 3을 참조하라.

20) LUB의 단계를 결정하기 위한 상 · 하위 · 자매 노드에 적정술어와 정의문을 적용하는 것은 3.2.2의 의미구분의 준거로 사용된 상 · 하위 · 자매 관계를 고려하는 것과는 그 목

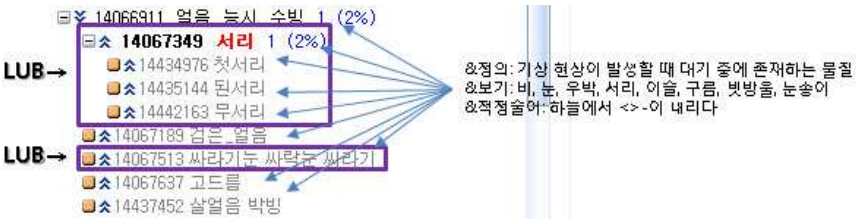

그림 10. 적정 술어/정의문을 상·하위·자매 노드에 적용하여 LUB을 결정하는 예

예를 들면, <기상관련물>에서 자동으로 사상된 신셋 '14067349 서리'의 세 개의 자식 노드 '첫서리, 된서리, 무서리'는 모두 <기상관련물>의 '기상 현상이 발생할 때 대기 중에 존재하는 물질'이라는 정의와 '하늘에서 < >-이 내리다'라는 적정 술어를 만족하므로 '14067349 서리'를 LUB로 잡을 수 있다. 그러나 '14067349 서리'의 자매 노드인 '14067189 검은 얼음', '14067637 고드름', '14437452 살얼음, 박빙' 등은 정의문과 적정술어를 만족하지 않으므로 사상에서 제외된다. '14067349 서리'의 부모 노드인 '14066911 얼음, 능시, 수빙' 역시 이들 준거를 만족시키지 못하므로 LUB가 될 수 없다. 따라서 <기상관련물>과 사상할 수 있는 LUB는 '14067349 서리'와 자매 노드 중에서는 '14067513 싸라기눈, 싸락눈, 싸라기' 등이다.

3.2.4 SJSC의 사상 노드 결정의 준거: SJSC와 KLN의 계층관계 비교

본 연구에서 사상의 기본단위는 SJSC의 단말 노드와 KLN의 LUB로 설정하였다. 그러나 세종 의미부류의 위계에 따라 단말 노드의 직접 상위 노드인 부모 노드 즉, 비단말 노드와 사상해야 할 경우가 발생하기도 한다. 이는 세종의 의미부류체계의 위계와 KLN의 위계해 다른 데서 주로 비롯되는데, 이 경우에는 SJSC와 KLN의 계층관계를 서로 비교·검토하여 SJSC의 단말 노드의 직접 상위 노드인 비단말 노드와 KLN의 LUB를 사상한다.

예를 들면, 세종 의미부류 <기상현상>의 자식 노드에는 <강우>, <강설>,

적이 다르다. 3.2.2에서는 상·하위·자매 노드의 의미관계 자체가 의미구분의 준거이며, 적정술어와 정의문의 적용은 이와 같은 정보가 있다면 추가로 검증하는 차원에서 적용해 볼 수 있다는 점에서 차이가 있다.

<바람>, <안개>, <구름>, <대기>, <날씨>가 있고, KLN에는 신셋 '10782227 날씨, 날, 일기, 기상'의 자식 노드로써 '10759408 비, 강우', '10766265 눈, 강설' 등의 신셋이 있다. 이 경우에는 KLN의 신셋 '10782227 날씨, 날, 일기, 기상'은 SJSC에서 <기상현상>에 해당하는 상위개념의 신셋이다. 따라서 SJSC의 의미부류 <기상현상>의 자식 노드인 <날씨>와 KLN의 신셋 '10782227 날씨, 날, 일기, 기상'을 사상하면, SJSC의 <강우>나 <강설> 등에 사상되는 신셋들이 다시 <날씨> 안에 사상되어 상속성의 원칙에 위배된다. 이 경우에는 <날씨>에는 KLN의 신셋을 사상하지 않고 비워두고, 부모 노드인 <기상현상>과 KLN의 신셋 '10782227 날씨, 날, 일기, 기상'을 사상하고, 신셋 '10759408 비, 강우'는 <강우>에, 신셋 '10766265 눈, 강설'은 <강설>에 사상한다.

3.2.5 사상 제외 준거: KLN의 오류 및 전문용어 후보군

마지막으로 사상에서 제외할 경우를 판단하는 준거를 살펴보자. 사상에서 제외하는 경우는 KLN 자체의 내부적인 부모-자식 할당 오류가 있는 경우와 KLN의 전문용어 후보군 두 경우다.

두 언어 자원을 사상하면서 양쪽 모두 의미부류 할당 오류가 발견되었다. 먼저, SJND 표제어의 의미부류 할당 오류의 한 예를 살펴보면, 표제어 '햇병아리'가 <새>가 아니라 <짐승>[21]으로 의미부류가 잘못 할당되어 있다. 이러한 오류는 SJSC 내부적으로는 오류로 추후 수정해야 하지만 두 자원의 사상에 있어서는 사상 오류를 야기하지는 않는다. 왜냐하면, 이러한 오류는 사상 준거 3.2.3에 따라 KLN의 상 · 하위 및 자매 관계를 살펴보는 과정에서 오류가 발견되지만, 부모 노드의 의미자질 상속이 위배되면 결국 사상되지 못하고, 결과적으로는 KLN의 의미분류 기준에 따라 올바르게 사상되기 때문이다. 다시 '햇병아리'의 예를 들자면, SJSC에서 <짐승>으로 잘못 분류되어 있어도 KLN에서 '햇병아리'는 신셋 '14436121 햇병아리'에 자동 사상되고, 그 부모 노드는 '01535520 새끼새'이므로 부모 노드를

21) SJSC에서 <짐승>과 <새>는 자매 노드이다. 의미부류 <짐승>의 정의는 '육지에서 사는, 날지 못하는 동물'이며, 그 적정술어는 '<>-이 새끼를 낳다'이다. <새>의 정의는 '날개 있는 척추동물'로 적정술어는 '<새>-(의) 날개, <새>-(의) 부리, <새>-(의) 둥지, <새>-가 날다'이다.

검토하면 '햇병아리'가 <짐승>의 의미자질이 아니라 <새>의 의미자질을 지니므로 사상하지 않게 된다. 한편, 의미부류 <새>에는 표제어 '병아리'가 분류되어 있어 자동으로 사상되는 신셋은 '01711406 병아리'이고, 그 부모 노드는 '01535520 새끼새'이다. 사상 준거에 따라 신셋 '01535520 새끼새'는 의미부류 <새>와 사상되는 LUB중 하나가 되므로 자동으로 '14436121 햇병아리'는 그 부모 노드를 통해서 <새>와 사상된다. 이와 같은 SJND 표제어의 의미부류 할당 오류는 <국가>로 잘못 분류된 표제어 '연방'이나 <신>으로 잘못 분류된 표제어 '스타킹' 등에서도 발견되지만, 이들은 결과적으로 KLN의 의미분류 기준에 따라 올바르게 사상된다.

이와는 달리, KLN에서는 사상 후보가 되는 신셋 중에서 부모-자식 노드의 의미할당에서 오류를 보이는 신셋은 사상 오류를 야기하므로 사상에서 제외한다. 예를 들면, 의미부류 <건반악기>의 표제어 '건반악기'와 '피아노'는 KLN에서 '02930011 클라비어, 건반악기'와 '03780029 피아노, 피아노포르테'에 각각 자동 사상되는데, 그 부모 노드가 각기 '04171373 현악기, 현'과 '03767801 타악기'로 부모-자식 관계가 잘못 분류되어 있다. 따라서, 이와 같이 KLN 자체의 부모-자식 노드 할당의 오류가 있는 신셋들은 부모의 의미자질이 상속되지 않아 상속의 원칙 즉, 사상의 전제조건에 위배되므로 사상에서 제외한다[22].

사상에서 제외하는 두 번째 경우는 자동 사상된 KLN의 후보군이 전문용어일 때이다. 예를 들면, 의미부류 <액체>는 KLN의 다수의 전문용어 신셋이 후보군인데 전문용어의 경우는 KLN에서도 정의문이나 용례가 없는 경우가 많아 해당 분야의 전문가가 아니라면 이해하기 어려운 수준의 신셋들이 많이 있다. <액체>로 분류되는 SJND의 표제어들과 자동 사상된 신셋 중에는 '14063488 타르'와 그 자식 노드인 '14063647 콜타르, 석탄타르, 타르'[23], '13932062 벤졸, 벤진, 벤젠' 등의 전문용어 신셋들이 있는데 이들의 부모 노드를 살펴보면, '14063488 타르'의 부모 노드는 '14063332 역청'이며, '13932062 벤졸, 벤진, 벤젠'의 부모 노드는 '13925570 방향족탄화수소'인데, '14063332 역청'과 '13925570 방향족탄화수소'는 자매 노드로

22) SJND의 표제어의 의미부류 할당 오류나 KLN의 부모-자식 할당 오류는 추후 수정되어야 하며, 특히 KLN의 부모-자식 할당 오류는 추후 수정 이후에 사상해야 한다.

23) SJND의 표제어 '타르'와 KLN의 신셋 '14063488 타르'와 그 자식 노드인 '14063647 콜타르, 석탄타르, 타르'는 동형이의어 수준의 자동사상에서는 모두 사상된다.

써 그 부모 노드는 공통적으로 '14062859 탄화수소'이다. 이 같은 경우에 전문용어 신셋들이 부모-자식 또는 자매 관계를 이루고 있어, '14063488 타르', '14063332 역청', '13932062 벤졸, 벤진, 벤젠', '13925570 방향족탄화수소', '14062859 탄화수소' 중에서 어떤 신셋을 LUB로 결정해야 할지 해당 분야 전문가가 아니라면 판단하기 어렵다.

또한, KLN은 다수의 번역되지 않은 영어 전문용어도 포함하고 있는데, 예를 들면, 의미부류 <미생물>과 자동 사상된 KLN의 신셋에는 '12242807 균'이 있고, 그 자매 노드로는'01308478 식물플랑크톤', '10787125 endemic', '10792309 acrogen', '10787270 야생식물' 등이 있다. 의미부류 <미생물>에 기술된 정보를 이용하여 '01308478 식물플랑크톤'은 <미생물>에 사상시킬 수 있지만 '10787270 야생식물'에 사상할 수 없다. 그러나 '10787125 endemic', '10792309 acrogen' 등과 같은 영어 용어는 전문가가 아닌 이상 그 의미 자체를 파악하기 어려울 수밖에 없다. 그러나 따라서 한국어를 모국어 말하는 사람이 이해하는 수준이 아닌 전문분야 용어는 LUB 결정 판단이 어려우므로 사상에서 제외한다.

4. 사상의 결과 및 논의

4.1 사상 결과

세종 의미부류 중 단말 노드를 중심으로 KLN의 LUB를 사상한 결과, 484개의 단말 노드 중 470개의 단말 노드와 일부 비단말 노드와 사상된 결과는 표 4와 같

표 4. SJSC 노드와 사상된 LUB의 수

세종의미부류	세종의미부류와 사상된 KLN의 LUB 수	하위노드 수	전체 노드 수 (LUB + 하위노드 수)
SJSC 단말	7039 (6438)	91226 (81550)	98265 (87988)
SJSC 비단말	53	225	278
전체	7092 (6487)	91451 (81768)	98543 (88255)

고, 표 5는 SJSC와 사상된 KLN의 LUB의 단계와 수를 제시하고 있다.

표 4에서 괄호 안의 수치는 중복을 제외한 수를 뜻하는데, 여기서 중복은 하나의 신셋이 여러 의미부류에 사상된 경우를 뜻한다. 예를 들면, SJSC에서는 유사한 의미부류들인 <가방>과 <주머니>, <그릇>과 <항아리> 등이 자매관계를 이루고 있지만 KLN에서는 '가방'과 '주머니'가 '02676701 주머니, 자루, 가방, 낭탁'으로 신셋을 이루고 있으며, '그릇'과 '항아리'는 부모-자식 관계를 이루고 있다. 따라서 이들 의미부류와 KLN의 신셋 간의 사상에 있어서는 KLN의 노드들이 유사한 의미부류들에 중첩되게 사상되는 경우가 발생한다. SJSC의 일부 비단말 노드를 포함하여 단말 노드와 사상된 KLN의 전체 노드 수는 중복을 제외하고 88,255개로 KLN의 전체 노드 수 90,134개의 97.91%를 차지하고 있음을 알 수 있다. 또한, 484개의 단말 노드 중 470개의 단말 노드가 사상되어 14개의 단말 노드가 사상이 불가능했는데 그 이유는 크게 세 가지이다. 첫째, SJND 표제어/보기-의미부류 정보쌍의 어휘분포를 참조해도 대응되는 KLN의 신셋이 부재하여 '<밝기단위>, <음악기호>, <고가도로>, <미각적행위>' 등의 의미부류는 사상되지 못하였는데, 이들은 KLN 신셋의 추가적인 보완을 통해 해결되어야 할 것이다.[24] 둘째, 의미체계의 이질성 때문에 <날씨> 등의 단말 노드는 직접 상위 노드인 비단말 노드와 사상이 되어 결국 단말 노드인 이들은 사상이 되지 못하였다.[25] 마지막으로 '<추상적 부분>, <관계속성값>, <외향적심리상태>, <내재적 심리상태>' 등은 개념의 구분이 어려워 결국 사상 준거를 찾지 못하여 사상하지 못하였다. 이들 의미부류에 대해서는 이차적인 두 자원 간의 검증작업을 통해 해결 방안을 모색해 보아야 할 것이다.

표 5에서는 주로 5-8단계에 있는 LUB에 많이 사상되어 있는데 이는 표 2의 KLN의 계층별 하위노드 수와 밀접히 관계된다. KLN의 5-8단계는 가장 신셋 수가 많은 단계로 이와 같은 사상결과가 나온 것으로 판단된다.[26] 또한, 상위단계일수록 사상된 LUB 아래 포함된 하위노드의 개수가 많으므로 예를 들어 2단계에서 사상된 LUB는 비록 그 개수는 35개에 불과하지만 실제로 그 LUB의 속성을 상속하는

24) 본고 3.2.1을 참조하라.

25) 본고 3.2.5을 참조하라.

26) 표 2에서 보았듯이, 5-8단계에는 단계별로 약 13,000～19,000개의 신셋이 할당되어 있다.

표 5. 단계별 SJSC와 사상된 LUB의 수

KLN의 단계	세종의미부류와 사상된 KLN의 LUB 수
2단계	35
3단계	395
4단계	832
5단계	1246
6단계	1363
7단계	1197
8단계	1166
9단계	566
10단계	194
11단계	80
12단계	14
13단계	4
계	7092

하위 노드 수가 매우 많다고 할 수 있다. 또 한 가지 흥미로운 점은 SJSC에서는 단말 노드라고 할지라도 2단계에서 최대 7단계까지 분포되어 있는데, SJSC에서 2-3단계에 위치한 단말 노드들은 KLN의 2-3단계의 LUB와 많이 사상되어 있고, SJSC에서 7단계에 있는 단말은 KLN에서도 5단계 이하, 주로는 6-7단계 이하의 LUB에 사상되는 경향을 보이고 있다. 예를 들어, 의미부류 <관리>, <상황>, <사실명제>, <부분시간>, <능력>, <모양>, <운> 등은 SJSC의 2-3단계에 분포되어 있는 단말 노드인데, KLN에서 2단계에 사상된 LUB들은 이들 의미부류에 많이 사상되어 있다. 반면, 의미부류 <긍정적신체속성인간>, <부정적신체속성인간>, <긍정적정신속성인간>, <부정적정신속성인간>, <종교적추종자>, <정치적추종자>, <에술적추종자> 등은 7단계에 분포된 단말 노드들인데 이들과 사상된 LUB는 주로는 6-7단계 이하의 LUB들과 사상되고 5단계 이상의 LUB도 거의 사상되지 않고 있으며, 3단계 이상의 LUB와는 전혀 사상되지 않는다.

표 6. SJSC와 KLN간 사상 상위 15순위의 의미부류와 KLN의 신셋 수

의미부류	SJND 표제어 수	LUB 수	하위노드 수	전체 노드수 (LUB+하위노드)
<직업인간>	535	155	5664	5819
<인간이름>	14	28	3485	3513
<질병및증세>	271	35	1979	2014
<일시적역할인간>	174	145	1841	1986
<나무>	111	8	1835	1843
<액체>	131	42	1694	1736
<사회계급인간>	49	22	1650	1672
<행정구역>	39	21	1547	1568
<기계>	139	43	1430	1473
<짐승>	133	23	1389	1412
<소속인간>	11	8	1401	1409
<능력>	140	63	1220	1283
<일>	118	22	1171	1193
<풀>	157	16	1157	1173
<모양>	123	27	1024	1051

표 6은 SJSC와 KLN 간에 사상이 많이 된 15개의 SJSC의 의미부류와 사상이 된 KLN의 신셋의 수를 나타내고 있다.

의미부류 <직업인간>, <인간이름>, <질병 및 증세> 등과 KLN의 신셋이 가장 많이 사상되고 있음을 알 수 있는데, 특히 SJND의 표제어 수가 적은 <인간이름>, <사회계급인간>, <행정구역>, <소속인간> 등은 KLN과의 사상을 통해 그 표제어를 많이 보완할 수 있음을 확인할 수 있다.

그 이외에도 SJSC와 KLN 간에 사상이 많이 되지 않은 의미부류로는 <하차>, <비교>, <철교>, <탑승>, <포장행위> 등이 있었는데, 이들 의미부류로 분류된 SJND의 표제어 수도 2-3개 수준으로 현저히 적었을 뿐만 아니라 이들과 사상된 LUB

역시 거의 단말 노드들이어서 이들과 사상된 전체 노드 수도 2-3개에 불과했다.

표 7은 SJND의 표제어 수가 많아 KLN에서 사상 가능한 후보군이 많음에도 KLN에서 표제어 수보다 적게 사상된 의미부류의 예이다.

이들은 양 의미부류 체계가 사상이 제대로 되지 않은 경우인데, SJSC의 오류나 KLN의 오류 및 결함 등에 기인한다. 예를 들면, <범주>의 경우에는 SJSC의 2단계

표 7. SJND의 표제어 수> 사상된 KLN의 신셋 수

의미부류	SJND 표제어 수	LUB 수	하위노드 수	전체 노드수 (LUB+하위노드)
<범주>	80	1	46	47
<사회운동>	44	9	34	43
<현재>	33	14	9	23
<정치유파>	33	5	14	19
<탈퇴>	27	11	10	21
<순서>	27	16	1	17
<방송물>	26	4	21	25
<논쟁>	24	10	12	22
<공중장소>	22	8	10	18
<관습>	21	5	18	23
<시각적행위>	19	6	11	17
<긍정적신체속성인간>	18	10	5	15
<천문현상>	16	11	3	14
<화시적장소>	11	10	0	10
<요일>	11	7	2	9
<중간정도>	11	7	0	7
<경연대회>	10	3	6	9
<학술모임회의>	10	5	1	6
<포상>	8	4	0	4

의 단말 노드로 그 표제어의 의미부류 할당에 다소 오류가 있어 KLN의 의미부류에 사상한 경우이다. 또 다른 예를 들자면, <정치유파>의 경우에는 KLN의 다의분할 오류에 의한 것으로 KLN에서는 개인과 집단의 다의적 관계에서 집단부분의 신셋이 없는 경우로, '강경파', '급진파' 등은 <정치유파>에도 속하는 표제어인데, KLN에서는 집단으로는 분류되어 있지 않고, 개인만을 뜻하여 사상되지 못했다. 표 7과 같은 의미부류들은 왜 표제어 수보다 적게 사상되었는지 그 원인을 분석하여 각 의미체계의 수정 보완 시에 참조해야 할 것이다.

4.2 개념부류의 이질성에서 비롯된 문제점과 해결방법

그 이외에도 SJSC와 KLN의 의미부류 체계가 달라서 발생하는 문제점들이 있는데, 이를 어떻게 해결하였는지 몇 가지 유형별 사례를 살펴보자.

첫째, SJSC에서는 대등한 단말 의미부류이지만 KLN에서는 부모-자식 관계로, SJSC의 대등한 단말 노드를 KLN의 부모 노드에 사상했을 때 상속성의 원칙에 위배된다면, KLN의 부모 노드와는 사상하지 않는다. 예를 들면, <대칭적친족>과 <비대칭적친족>은 SJSC에서는 대등한 노드들인데, KLN에서는 <대칭적친족>은 대체로 총칭적인 뜻을 나타내는 부모 노드(예: 친척, 여자 친인척, 남자 친인척 등)이고, <비대칭적친족>은 하나하나의 예들로 주로 자식 노드들(예: 큰어머니, 여동생, 질녀 등)에 해당하는 것들이 많다. <대칭적친족>으로 사상되는 신셋의 예로는 '친족, 친척, 부부, 사촌, 오누이, 형제, 이복형제, 친사촌' 등이 있고, <비대칭적친족>으로 사상되는 신셋의 예로는 '여동생, 남동생, 고모, 고모부, 누나, 언니, 질녀, 며느리' 등이 있다. 이 같은 경우에 <대칭적친족>에 해당하는 신셋 '09576199 여자 친인척'을 사상하면 되지만, 이렇게 되면 <비대칭적친족>에 속하는 자식 노드에 있는 신셋들이 문제가 된다. 따라서 이 같은 경우에, <대칭적친족>과 '09576199 여자 친인척'을 사상해야 마땅하지만, 그 하위노드가 <대칭적친족>에 속하지 않아 상위 노드의 의미자질을 계승하는 데 문제가 있으므로 이 경우에는 SJSC의 <대칭적친족>과 KLN의 '09576199 여자 친인척'을 사상하지 않는다. 반면, <비대칭적친족>의 경우에는 <대칭적친족>의 하위노드에 해당되므로 하위노드는 사상

한다.

둘째, 신셋 중 하나의 어의는 적정술어에 따라 사상이 되고, 다른 하나의 어의는 적정술어에 맞지 않아 사상이 안 되는 경우, 다수 또는 중요한 개념을 중심으로 사상한다. 예를 들면, 의미부류 <모자>로 분류되는 표제어 중 하나인 '왕관'에는 왕이 쓰는 왕관 이외에도 '보기'에서 제시하는 '월계관'과 유사한 의미의 '왕관'과도 사상이 가능한데, 그 신셋이 '06293451 우승기 왕관'이다[27]. '우승기'를 쓰거나 벗을 수는 없지만, 개념상으로는 '우승기', '왕관', '월계관' 모두 유사하다. 이 같은 경우에는 다수 또는 중요한 개념을 중심으로 사상한다. 따라서 '우승기'는 무시하고 '06293451 우승기, 왕관'을 <모자>에 사상한다.

셋째, SJSC에서 서로 상반되는 두 의미부류가 KLN에서 부모-자식 노드에 각각 사상되는 경우에는 사상 원칙에 어긋나지 않는다면 사상한다. 예를 들면, 의미부류 <비용>과 <소득>은 서로 상반되는 의미의 완전히 다른 노드이다. 그러나 SJND 표제어-의미부류 정보 쌍과 KLN 신셋 간의 자동 사상 결과를 보면, <비용>으로 분류되는 표제어 '계약금'은 KLN에서 부모 노드 '12593229 계약금, 계약보증금, 증거금, 체약금, 약조금'에, <소득>으로 분류되는 표제어 '마진'은 그 자식 노드인 '14435848 마진'에 사상되어 있다[28]. 이것은 지불하는 입장에서는 비용이 되고, 지불을 받는 입장에서는 소득이 되어서이다. 이 경우는 SJSC와 KLN의 분류 기준의 차이에서 오는 것이므로 사상 원칙에 따라 <비용>은 부모 노드인 '12593229 계약금, 계약보증금, 증거금, 체약금, 약조금'에, <소득>은 자식 노드인 '14435848 마진'에 사상한다.

넷째, KLN의 신셋이 열거형의 고유명사인 경우는 KLN의 특성에 기인한 것으로 일반적인 경우와 같이 사상한다. 예를 들면, <종교적역할인간>에 사상되는 신셋 '09774028 포프, 로마교황, 교황'의 자식 노드들은 모두 '10108436 알렉산더 육세',

---

27) 의미부류 <모자>에 대해 SJSC에 기술되어 있는 정보는 다음과 같다.
&보기: 모자, 갓, 월계관
&적정술어: <>-을 쓰다, (<>-을 벗다)

28) '12593229 계약금, 계약보증금, 증거금, 체약금, 약조금'과 '14435848 마진'은 KLN에서 부모-자식 관계를 이루고 있다.

표 8. 언어정보로 사상 후보 쌍의 의미 등가성 판단이 안 되는 예-'격하'

사상 후보 노드		의미 구분을 위한 언어 정보	
		유형	제공된 언어 정보
SJND	<격하 1>	용례	<센스_구획 n="1"> <용례>이번 사태로 인한 한국의 국제적인 위상의 ~는 불보듯 뻔하다.</용례> <영어_대역어>degradation</영어_대역어> <의미_부류>감소</의미_부류>
	<격하 2>	용례	<센스_구획 n="2"> <용례>그런 행위는 스스로의 가치를 ~를 하는 것과 다름없다.</용례> <영어_대역어>degradation</영어_대역어> <의미_부류>감소행위</의미_부류>
KLN '격하 0018875'		정의문	자격이나 등급, 지위 따위의 격이 낮아짐. 또는 그것을 낮춤

'10171736 칼릭투스 삼세' 등 교황의 이름들이다. SJSC에서는 고유명사를 다루고 있지 않지만, 이와 같은 고유명사는 KLN의 개념체계의 특성이므로 일반명사와 마찬가지로 사상한다.

다섯째, SJND에서는 매우 섬세한 의미구분으로 인해 다른 의미부류들에 속하는 다의적 추상명사인데 반해, KLN에서는 단의로 기술되어 정의문이나 용례 등으로도 의미구분이 어려운 신셋의 경우에는 KLN의 신셋을 의미에 따라 유사한 SJSC의 의미부류들에 모두 사상한다. 예를 들어, SJND에서 '격하 1'은 <감소>로 분류되고, '격하 2'는 <감소행위>로 분류되지만 KLN에서는 단의로 기술되므로 신셋 '격하 0018875'는 <감소>와 <감소행위> 모두에 사상한다.

5. 결 론

본 연구에서는 인간언어공학에서의 활용을 위해, 의미입자가 작은 SJSC와 KLN 1.5의 사상을 목표로 전문가의 수작업을 위한 귀납적/상향적 사상방법론을 제안하였다. 두 이종 개념체계 간의 사상 방법은 SJSC의 단말 노드와 KLN의 LUB를 기본 단위로 하여, ① 사상 후보군 결정의 준거, ② 후보군들 간의 정확한 의미구분의 준거, ③ LUB의 단계를 결정하기 위한 준거, ④ SJSC의 단말 노드와의 사상 여부를 결정하기 위한 준거, 마지막으로 ⑤ 사상에서 제외할 신셋을 결정해 주는 준거 등 적용하여 사상하였다. 또한, 양 개념체계의 이질성에서 비롯된 문제점들을 그 유형과 사례를 중심으로 살펴보고 그 해결방안에 대해 논의도 하였다. 본 연구에서 제안한 방법으로 사상한 결과, SJSC의 474개의 단말 및 비단말 노드와 KLN의 신셋 간에는 중복을 제외하고 6,487개의 LUB가 사상되었으며, 각 LUB의 하위노드를 포함해서는 모두 88,255개의 KLN 신셋이 사상되어 전체적으로는 97.91%가 사상되었다.

본 연구는 이종적인 개념체계인 SJSC와 KLN의 상위 노드 간의 사상을 시도한 첫 작업으로 앞으로 많은 과제를 남기고 있다. 사상이 되지 않은 14개의 단말 노드 중 SJND 표제어/보기-의미부류 정보 쌍의 어휘분포를 참조해도 대응되는 KLN의 신셋이 부재하여 '<밝기단위>, <음악기호>, <고가도로>, <미각적행위>' 등의 의미부류도 사상되지 못한 채 남아 있다. 이 문제는 향후 KLN의 보완을 통해 해결할 수 있을 것이다. 또한, 의미부류 '<추상적부분>, <관계속성값>, <외향적심리상태>, <내재적심리상태>' 등은 여전히 사상 준거를 찾지 못하여 사상하지 못하였다. 이들에 대해서는 이차적인 두 자원 간의 검증작업을 통해 해결 방안을 모색해 보아야 할 것이다.

이 이외에도 본 연구에서는 다루지 못하였으나 양 의미체계를 사상하면서 하나의 자원만을 고려할 때는 명시적으로 드러나지 않았던 문제점들이 파악되었다. 예를 들면, SJSC의 경우에는 KLN의 분류체계와의 비교를 통해 분류체계의 문제가 발견되었고(예: <전기전자기구>→ <조명기구>, <구체물의부분>→ <의복부분>, <기관건물>→ <교도소> 등 하나의 하위노드만 있음), SJND의 표제어에 대한 의미부류 할당 오류도 발견되었다. KLN에서는 SJSC와의 비교 · 검토 과정에서 다의

분할 오류나 부모-자식 노드 할당 오류, 문화적 차이로 인한 한국어 신셋 부족 등이 발견되었다. 이들 문제는 향후 두 의미체계가 지닌 장점을 극대화함과 동시에, KLN은 SJSC를 통해서, SJSC는 KLN을 통해 각 개념체계의 단점을 상호 보완할 수 있을 것으로 보인다. 이들이 서로 상호 보완된다면 추후 보다 완전한 개념체계로써 구문분석이나 의미분석 등에 이용하여 실용성의 증대를 꾀할 수 있을 것이다. 아울러 본 연구에서 제안한 자동 사상방법론을 토대로 효율적인 자동 사상 방법 및 이를 이용한 LUB의 광역화 가능성을 모색해야 할 것이며, 사상된 LUB의 자동 검증 방안도 연구되어야 할 것이다. 마지막으로 KorLex의 용언에 기술되어 있는 문형정보와 세종 전자사전의 용언의 격틀 정보를 통합 구축하여 구문분석에서 이용할 때, 본 연구의 결과를 이용하여 논항의 일반화된 선택제약규칙의 기술을 시도해 보고, 실제 구문분석기에서 의미영역에 따라 어느 단계의 LUB를 설정하는 것이 가장 효율적인지도 모색해 보아야 할 것이다.

참고문헌

[1] 이성헌 (2007). “세종 전자 사전의 어휘 의미 부류 체계”, **새국어생활** 17-3, pp.51-67.

[2] PWN: http://wordnet.princeton.edu/

[3] 윤애선, 황순희, 이은령, 권혁철 (2009). “한국어 어휘의미망 KorLex 15.의 구축”, **정보과학회 논문지: 소프트웨어 및 응용, 36**(1), pp.92-108.

[4] Piek Vossen and Christiane Fellbaum (2009). “Universals and idiosyncrasies in multilingual WordNets”, *Multilingual FrameNets in Computational Lexicography: Methods and Applications,* Hans C. Boas Ied) Mouton de Gruyter, pp.319-345.

[5] Niles, I., and Pease, A. (2003). “Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology”, *Proceedings of the IEEE International Conference on Information and Knowledge Engineering,* pp.412-416.

[6] Carpuat, M., Ngai, G., Fung, P., and Church, K. W. (2002). “Creating a Bilingual

Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet", *Proceedings of GWC 2002, the 1st Global WordNet conference*, pp.284-292, Mysore, India.

[7] Asanoma, A. (2001). "Alignement of Ontologies: WordNet and Goi-Taikei", *Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pp. 84-89, Pittsburgh, U.S.A.

[8] Lee, Changki and Geunbae Lee and Jungyun Seo (2000). "Automatic WordNet mapping using Word Sense Disambiguation." *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 2000)*, pp.142-147, Hong Kong.

[9] Changhua Yang, Sue J. Ker (2002). "Considerations of linking WordNet with MRD", *Coling 2002: Proceedings of the 19th International Conference on Computational Linguistics*, pp.1-7, Taipei, Taiwan.

[10] 소길자, 윤애선, 권혁철 (2009). "세종 의미 부류와 KorLex 명사 어휘 의미망 자동 맵핑", **2009 한글 및 한국어정보처리 학술대회 발표논문집**, pp.92-96.

[11] 홍재성 외 (2000, 2001, 2002, 2003, 2004, 2005, 2006). **"21세기 세종계획 전자사전 개발 연구 보고서"**, 문화관광부.

1 차원고접수 : 2009. 11. 12
2 차원고접수 : 2010. 3. 26
최종게재승인 : 2010. 3. 26

(*Abstract*)

Mapping Heterogenous Ontologies for the HLP Applications – Sejong Semantic Classes and KorLexNoun 1.5 –

Sun-Mee Bae Kyoungup Im Aesun Yoon

Pusan National University

This study proposes a bottom-up and inductive manual mapping methodology for integrating two heterogenous fine-grained ontologies which were built by a top-down and deductive methodology, namely the Sejong semantic classes (SJSC) and the upper nodes in KorLexNoun 1.5 (KLN), for HLP applications. It also discusses various problematics in the mapping processes of two language resources caused by their heterogeneity and proposes the solutions. The mapping methodology of heterogeneous fine-grained ontologies uses terminal nodes of SJSC and Least Upper Bounds (LUB) of KLN as basic mapping units. Mapping procedures are as follows: first, the mapping candidate groups are decided by the lexfollocorrelation between the synsets of KLN and the noun senses of Sejong Noun Dfotionaeci(SJND) which are classified according to SJSC. Secondly, the meanings of the candidate groups are precisely disambiguated by linguistic information provided by the two ontologies, i.e. the hierarchicllostructures, the definitions, and the exae les. Thirdly, the level of LUB is determined by applying the appropriate predicates and definitions of SJSC to the upper-lower and sister nodes of the candidate LUB. Fourthly, the mapping possibility ic inthe terminal node of SJSC is judged by che aring hierarchicllorelations of the two ontologies. Finally, the ituorrect synsets of KLN and terminologiollocandidate groups are excluded in the mapping. This study positively uses various language information described in each ontology for establishing the mapping criteria, and it is indeed the advantage of the fine-grained manual mapping. The result using the proposed methodology shows that 6,487 LUBs are mapped with 474 terminal and non-terminal nodes of SJSC, excluding the multiple mapped nodes, and that 88,255 nodes of KLN are mapped including all lower-level nodes of the mapped LUBs. The total mapping coverage is 97.91% of KLN synsets. This result can be applied in many elaborate syntactic and semantic analyses for Korean language processing.

Keywords : Semantic class, Lexical semantic network, Language engineering, Sejong electronic dictionary, KorLex, Bottom-up and inductive mapping methodology, Manual ontology mapping