자연언어처리를 위한 한국어 어휘 자원과 언어 교육에의 응용*

신효필**

<차 례>

Ⅰ. 서론

인간의 언어를 컴퓨터가 이해하고 처리할 수 있는 방법을 개발하는 자연언어처리(natural language processing), 또는 컴퓨터언어학(computational linguistics)은 컴퓨터 및 인터넷의 발달로 인한 정보의 홍수 속에서 그 중요성이 날로 증대하고 있다. 한 언어를 다른 언어로 번역하는 기계번역 시

* 이 논문은 국어교육학회 제46회 학술대회의 기획 발표 논문을 수정한 것이다. 학술 대회에서의 발표는 자연언어처리, 컴퓨터언어학 분야에서 구축된 한국어 어휘자원에 대한 소개가 주된 내용이었으며, 언어 교육의 적용은 본 논문에서 새로 첨가되었다. 그러나 자연언어처리가 전공인 저자가 언어 교육 내지 국어 교육에의 응용이라는 점을 논의한다는 것은 한계가 있다. 언어 처리라는 관점에서 구축된 자원의 특징을 주로 하는 이 글을 바탕으로 하여 언어 교육 전공자들의 더 심도 있는 언어 교육으로의 응용을 기대한다.

** 서울대학교 인문대학 언어학과, hpshin@snu.ac.kr

스템이나, 필요한 정보를 효율적으로 찾게 해주는 정보 검색 시스템, 문서의 내용을 분류하고 요약하는 시스템 등은 실제로 우리 주변에서 쉽게 찾아볼 수 있고 널리 활용되는 자연언어처리 시스템 중 일부이다.

자연언어처리는 인간의 언어를 다루는 만큼 그 근간이 되는 것은 언어의 정보를 체계화한 어휘 자원의 구축이다. 기계가독형 전자사전(machine readable dictionary), 온톨로지(Ontology), 그리고 어휘 의미망(lexical meaning network) 등은 대표적인 어휘자원이다. 그 중에서도 어휘 의미망은 기존의 종이사전이나 전자사전의 평면적인 어휘 기술에서 벗어나 동의어, 반의어, 상위어 등의 여러 의미관계를 서로 연결하여 입체적인 언어 정보를 구축한다는 점에서 의의가 있다. 이런 어휘 의미망의 대표적인 것으로 미국 프린스턴 대학의 G, Miller에 의해 구축된 영어 워드넷(WordNet, 이하 워드넷)을 들 수 있다. 워드넷은 원래 심리학적, 인지적 관점에서 개념(concept)의 언어화된 표상인 어휘들이 지니는 다양한 의미관계를 계층적인 망으로 구축한 것으로, 이를 근간으로 하여 EuroWordNet, MultiWordNet, BalkaNet 등 약 50개의 언어들의 어휘 의미망이 구축되고 있다(신효필 2010).

한국어의 경우도 이런 어휘 자원을 구축하려는 작업이 오랫동안 행해져 왔다. 10년에 걸친 21세기 세종계획은 코퍼스 및 전자사전 구축을 포함하여 대규모로 한국어 자원을 구축하는 대표적인 작업이다. 어휘 의미망 관련 연구로는 한국어 시소러스(포항공대), 다국어 DB(고려대학교), U-Win(울산대학교), CoreNet(KAIST), KorLex(부산대학교) 등이 있다. 이 중 CoreNet은 다국어 범용 어휘망을 지향하여, 명사, 동사, 형용사들을 모두 하나의 개념 체계 안에서 분류한다는 점에서 기존의 워드넷과 차이가 있다. 울산대학교의 U-Win은 한국인의 보편적인 인지 체계와 개념 관계를 파악하여 개념 구조를 형성하려는 노력으로 명사, 동사, 형용사, 부사, 관형사, 감탄사, 조사, 수사 의존명사 그리고 북한어, 방언, 옛말, 전문용어, 고유명사, 어근 어미 등을 포함한, 한국어 전 어휘를 대상으로 하고 있다. 부산대학교의 KorLex는 영어 워드넷의 한국어로의 대역에서부터 출발하여 동의어 집합인 기존의 신셋(synset)을 확장하여, 현재 KorLex1.5에 이르고 있다. KorLex는 대규모의 본격적인 한국어 어휘의미망 구축이라는 점에서 의의

가 있으며 현재 여러 분야에서 활용되고 있다.[1)]

본 논의에서는 이런 어휘 자원 외에, 이미 존재하는 개념체계에 한국어 어휘를 사상시켜 개념망에 따르는 어휘 의미망을 구축하려는 KOLON (the KOrean Lexicon ONtology)에 대해 설명하고 이 어휘자원이 어떻게 한국어 어휘 교육에 활용될 수 있는지를 살펴본다. 신효필(2007, 2010)에서는 워드넷이 "어휘의미=개념"이라는 등식을 설정한 것과 달리 개념과 어휘의미는 구분되어야 하며, 이를 분리시켜 놓을 때 언어적으로 의의있는 자료가 될 수 있음을 보이고 있다. 개념과 어휘를 분리시키고 어휘를 개념체계에 사상(mapping)시키게 되면, 워드넷과 같이 세밀한 동의어 그룹이 형성되는 것과는 달리, 인지적, 개념적으로 연결된 폭넓은 동의어 그룹이 형성된다. 또한 KOLON은 어휘를 개념에 사상시킬 뿐만 아니라 기존의 잘 체계화된 세종전자사전으로부터 통사적, 의미적 정보들을 추출하고 이를 다시 프레임 기반의 어휘부에 통합시킨다. 이런 정보들은 비단 자연언어처리 뿐만 아니라 한국어 교육에도 활용될 수 있다.

본 논의는 KOLON의 구축과 그 특징에 대해 기술하고 이것이 한국어 다른 어휘 자원과 어떤 점에서 차이가 나는지를 우선 살펴본다. 그리고 KOLON이 어떻게 언어 교육에 활용될 수 있는지, 특히 어휘 교육에 어떻게 적용될 수 있는지에 대해 논의하도록 한다. 그러나 논의는 어휘 교육, 그 중에서도 유의어 교육에 한정된다. 이는 컴퓨터 언어학을 전공하는 필자가 언어 교육의 포괄적인 내용을 다 파악하기 어려울 뿐만 아니라 언어처리와 언어 교육이라는 관점의 차이에서 오는 불일치도 크기 때문이다. 따라서 언어 처리에서 체계화된 자질을 활용하는 관점에서의 어휘 교육이 주된 논의 대상이 된다.

1) 한국어 어휘 자원에 관한 비교는 윤애선(2009), 신효필(2007, 2009) 참조

Ⅱ. KOLON

1. 미크로코스모스 온톨로지

KOLON은 미국 뉴멕시코 주립대학 CRL(Computing Research Laboratory)에서 개발된 미크로코스모스 온톨로지(Mikrokosmos Ontology)를 근간으로 하고 있다. 미크로코스모스 온톨로지는 1990년대 초 Lynn Carlson과 Sergei Nirenburg에 의해 시작된 자연언어처리를 위해 개발된 온톨로지로 총 5,449개의 개념으로 되어 있다.[2] 미크로코스모스 온톨로지의 개념들은 다른 온톨로지와 마찬가지로 계층관계를 이루고 있다. 최상위의 ALL을 기점으로 하여 OBJECT, EVENT, PROPERTY의 삼분 구조로 출발한다. 개념은 메타언어인 영어 대문자로 표시되며, 이를 이루고 있는 여러 속성들의 집합인 프레임(frame)으로 구조화된다. OBJECT는 주로 대상을 지칭하는 개념이며 EVENT는 행위, 상태를 나타내는 개념이다 PROPERTY는 대상과 행위의 특질을 기술하기 위한 개념으로 다시 어떤 속성을 나타내는 ATTRIBUTE와 관계를 나타내는 RELATION으로 하위 구분된다. EVENT에 속하는 개념 중의 하나인 DISPLAY 개념의 프레임 구조를 살펴보면 다음과 같다.

〈표 1〉 DISPLAY 개념 구조

CONCEPT	SLOT	FACET	FILLER
DISPLAY	DEFINITION	VALUE	"to show to public view"
	IS-A	VALUE	PHYSICAL-EVENT
	SUBCLASSES	VALUE	INDICATE, SIGNIFY
	AGENT	SEM	ANIMAL, ENERGY, FORCE

2) 미크로코스모스 온톨로지에 관해서는 Maheshi(1996), Nirenburg and Raskin(2004), 신효필(2007, 2010) 참조. 여기서는 신효필(2010)의 기술을 근간으로 하였다. 여기서 개념 수는 매핑 작업의 결과 새로 획득된 개념을 포함한 것이다.

CONCEPT	SLOT	FACET	FILLER
DISPLAY	BENEFICIARY	SEM	OBJECT
	INSTRUMENT	DEFAULT	ANIMAL-PART, DEVICE
	LOCATION	SEM	PLACE
	THEME	SEM	PHYSICAL-OBJECT

슬롯(SLOT)은 DISPLAY 개념의 속성을 기술하기 위한 것으로 PROPERTY에 해당하는 개념들이 사용되며 패싯(FACET)은 그 채워지는 값들이 어떤 종류인지를 나타난다. 슬롯의 값으로 채워지는 값들은 OBJECT나 EVENT에 속하는 개념들이거나 아니면 숫자, 아니면 아직 개념화되지 않은 문자열(string)들이다. DISPLAY의 경우 상위 개념(IS-A)은 PHYSICAL- EVENT이며 하위개념(SUBCLASSES)으로는 INDICATE, SIGNIFY를 가지며, 행위의 주체(AGENT)로는 ANIMAL, ENERGY, FORCE를, 그리고 행위의 대상(THEME)으로는 PHYSICAL-OBJECT라는 것을 취한다는 것이 프레임으로 기술되어 있다.

Nirenburg and Raskin(2004)에 의하면, 개념은 텍스트 의미표상(Text Meaning Representation, TMR)에 있어 어휘부와 연결된다. 텍스트 의미표상이란 문장들 그리고 이것의 집합인 텍스트의 의미를 형식적으로 기술하기 위한 장치이다. 텍스트 의미표상은 문장의 중심이 되는 서술어를 중심으로 이 서술어가 사상되는 개념과 서술어에 의해 나타나는 통사, 의미적 구조를 기술하는 방법으로 이루어진다. 어휘가 특정 개념에 사상되면 사상된 개념의 속성을 어휘부에 전달하여 어휘부에도 개념의 속성이 반영된 구조를 형성할 수 있다. 다음은 개념 BUY의 사상된 어휘 *buy*의 예이다.[3)]

(1)
buy-v1
cat v
morph stem-v bought v+past
bought v+past-participle

3) 이 예는 Nirenburg and Raskin(2004)를 기초로 한 신효필(2010)에서 그대로 발췌하였다.

```
anno  def "when A buys T from S. A acquires possession of T previously
        owned by S, and S acquires a sum of money in exchange"
      ex  "Bill bought a car from Jaen"
syn   syn-class trans
      syn-structure     root      buy
                        subj      root $var1
                                  cat n
                        obj       root $var2
                                  cat n
                        oblique   root from
                                  cat prep
                                  opt +
                                  obj      root     $var3
                                  cat      noun
      sem-structure
                        BUY
                        agent     value    ^$var1
                                  sem      HUMAN
                                  theme    value ^$var2
                                           sem   OBJECT
                                           source value ^$var3
                                                  sem HUMAN
```

*buy*의 형태, 통사 정보는 특별할 것이 없으나, 의미 정보는 차별성을 지닌다. 우선 어휘 *buy*에 해당되는 개념이 BUY임이 명시되어 있고 이 개념이 지닌 속성 중 agent, theme, source를 그대로 상속해서 그 값들이 각각 HUMAN, OBJECT, HUMAN임을 어휘부에 명세한다. 또한 그 value로 통사적 주어, 목적어 등의 변수 $var1, $var2를 각각 연결하여 통사구조와 의미정보가 연결될 수 있는 기제를 마련해 놓았다.

KOLON은 이러한 어휘부 구축을 지향한다. 개념체계의 정보를 어휘부로 전달하여 통사와 의미정보가 결합된 어휘기술을 추구한다. 결과적으로 영어의 FrameNet[4]과 비슷한 기술이 이루어진다.

2. KOLON의 구축

KOLON은 2007년부터 시작한 1차 작업, 2009년에 진행된 2차 작업 및 교열작업으로 이루어진다. 1차 작업은 국립국어원(2002)에서 2003년 5월에 발표한 한국어 학습용 어휘 목록 중에서 1283개의 동사를 단순히 미크로코스모스 개념에 대응시켰다. 전체적으로 4,818개의 어의(sense)가 구별되어 기술되었다.[5] 2차 작업은 한국어 동사의 항목 수를 늘이고 명사와 형용사도 포함하는 등 기술 항목을 확대하였다. 단순한 어휘의 개념 사상에서 벗어나 세종 사전의 통사, 의미 정보를 추출하여 결합하는 작업도 병행하였다. 또한 1차 작업에서 나타난 오류와 잘못된 사상을 교정하는 작업도 이루어졌다. 이렇게 하여 명사 25,459개, 동사 15,180개, 그리고 4,075개의 형용사가 개념에 사상되었고 더불어 논항 정보 및 의미제약에 관한 정보들도 추가되었다. 전체적으로 65,326개의 어의가 기술되었다. 이 수는 기술된 34,714어휘의 모든 어의를 다 포함하는 것이 아니며 앞으로의 지속적인 작업에 의해 더 보충되어야 한다.[6]

KOLON은 어휘 사상 작업에 의한 통합적인 어휘부를 구축할 뿐만 아니라 구축된 어휘부를 관리하고 웹으로 검색할 수 있는 툴도 제공하고 있다. KOLON의 데이터베이스는 온톨로지 프레임 시스템과 어휘부 프레임 시스템으로 구축되어 있다. 어휘부 프레임에는 개별 어휘의 사상된 개념과 세종 사전에서 추출된 정보들이 프레임으로 기술되어 있다. 다음은 어휘부가 프레임으로 표시되는 예이다.

---

4) FrameNet은 미국 버클리대학에서 구축되고 있는 어휘 자원으로 Fillmore의 격문법(case grammar), 프레임 의미론(frame semantics)에 기반한 연구이다. FrameNet에서는 각 문장 성분들의 의미역(thematic role)과 그 통사적 구조를 프레임 요소(Frame Element)를 사용하여 프레임으로 기술하고 있다.

5) 자연언어처리에서 어의(sense)는 동철자이음어, 다의어 등을 구별하지 않고 다 포함하는 개념으로 쓰인다.

6) 사상의 구체적인 과정에 대해서는 신효필(2007, 2010)을 참조.

〈표 2〉 '먹다'의 다의 의미 기술 예

Lex	Slot	Facet	Filler
먹다	SEM-CLASS	SEM	INGEST
_01_01_VV	THEME	SEM	FOOD, FOODSTUFF
		CASE-MARKER	을(object)
	AGENT	SEM	HUMAN
		CASE-MARKER	이(subject)
	FRAME	VALUE	X=N_0-이Y=N_1-을 V
	EXAMPLE	VALUE	**우리는**〈AGENT〉 어머니가 해주신 **요리를**〈THEME〉 배불리 먹었다.

<표 2>는 '먹다'의 의미 중 하나를 보여주는 것으로 온톨로지 개념으로는 INGEST에 사상되었으며 개념 INGEST가 가지는 슬롯 정보를 어휘부에 계승하여 적절히 제약이 가해진 정보를 보여주고 있다. 이 경우 INGEST의 AGENT는 HUMAN이어야 하며, THEME으로는 FOOD, FOODSTUFF가 해당된다는 것이 명시되어 있다. 또한 논항 정보를 표시하기 위해 FRAME 슬롯에 세종 전자 사전에서 추출한 격틀정보를 변환하여 명시한다. 격틀 구조에 있어 각 논항(N0, N1)이 조사 정보에 의해 각각 AGENT와 THEME에 결속됨이 명시되어 있다. 따라서 AGENT, THEME과 같은 슬롯에는 한국어의 특성을 반영한 CASE-MARKER라는 패싯을 설정하여 조사 정보를 명시하였다. AGENT와 THEME에 채워지는 값들은 세종 사전에서는 단순히 어휘로 나열되어 있는 것을 일일이 미크로코스모스 개념으로 사상하였다. 따라서 FOOD, FOODSTUFF라는 개념에 사상된 어휘들이 실제 논항에 나타난다. 이 어휘 구조를 예시하기 위해 예문은 해당되는 논항과 그 의미역을 태깅하여 보여주는 방식으로 이루어져 있다.

또 다른 프레임인 온톨로지 프레임 시스템은 미크로모스 개념 구조를 프레임 형식으로 보여주기 위한 장치이다. 온톨로지에 기술된 정보들이 기본적으로 다 명시되지만 첨가적으로 각 언어에 사상된 어휘정보를 보여주기 위해 사상된 어휘정보를 함께 명시한다. 한국어의 경우 LEXK라는 슬롯에 의해 사상된 어휘들이 온톨로지 브라우저에 표시된다. 다음은 표

2의 '먹다' 어휘부 프레임에서 THEME의 제약으로 사용되었던 FOOD의 온톨로지 프레임이다.[7)]

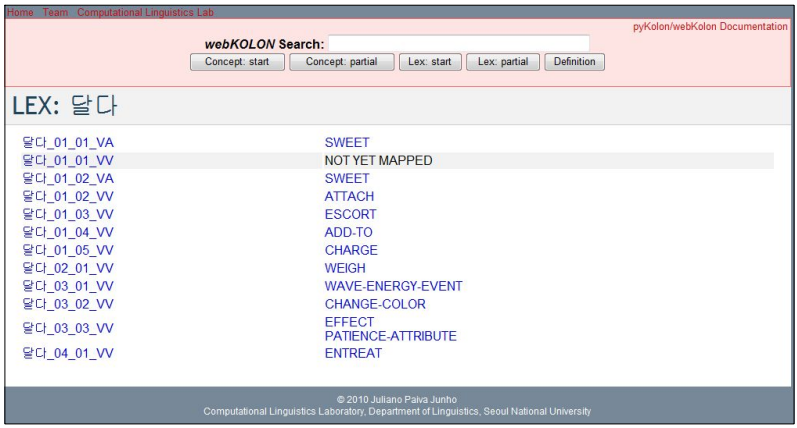
〈표 3〉 FOOD의 온톨로지 프레임

Concept	Slot	Facet	Filler
F O O	DEFINITION	VALUE	"a substance take in and assimilated by an organism to maintain life and growth : nourishment"
D	IS-A	VALUE	INGESTIBLE
	LEXK	LEXK	관식_01_01_NN 김치_01_01_NN 두부_01_01_NN 식품_01_01_NN(food)
	MADE-OF	INV	FAT
	SUBCLASSES	VALUE	FOODSTUFF,
	THEME-OF	INV	PREPARED-FOOD ACQUIRE-GROCERY

KOLON은 온톨로지와 어휘를 검색할 수 있는 웹 인터페이스를 제공하고 있다. http://word.snu.ac.kr/kolon에서 개념과 어휘를 찾을 수 있다. 다음은 이 브라우저를 통하여 '달다'라는 어휘를 찾은 경우이다. 이 브라우저를 통해, 개념과 어휘를 검색할 수 있으며 슬롯에 사용된 개념이나 그 값으로 채워진 개념들을 클릭하면 해당 개념의 구조를 볼 수 있다.

<그림 1>에서는 '달다'의 어의 구분된 리스트들과 사상된 개념들이 일차적으로 표시되어 있다. 해당 어의를 클릭하면 개별 어휘정보가 <표 2>와 같은 어휘부 프레임 구조가 표시되며, 사상된 개념을 클릭하면 해당되는 개념을 구조와 사상되어 있는 어휘들이 표시된다.

7) 여기에 사상된 어휘, LEXK는 지면상 일부만을 예시하였다.



〈그림 1〉 KOLON 브라우저

3. 어의 크기(Granularity)

영어 워드넷에서는 동의관계를 기본적인 의미관계로 규정하고 개념을 표상하는 최소 단위로 삼는다. 따라서 어휘의 세분화된 의미가 개념이라는 등식이 성립된다. 이 세분화된 의미는 신셋(synset, synonym set)으로 표시되며 이를 구성하고 있는 원소들은 어의(sense)가 된다(윤애선 외 2009 : 93) 워드넷에서 *car*의 신셋은 다음과 같이 이루어진다.

(2) 어의WN(car)={{car, auto, automobile, machine, motorcar},
{car, railcar, railwaycar, railroadcar},
{cablecar, car}
{car, gondolar}
{car, elevatorcar}}

*car*의 경우 다섯 가지 동의어 그룹이 이루어지면 각 신셋은 별도의 상위 신셋에 연결된다. 이 중에서 두 개의 신셋으로 이루어진 경우, *car*를 제외하고는 단지 하나의 동의어로만 이루어져 있어 지극히 세분된 분류라고

할 수 있다. 이렇게 세분된 신셋은 자연언어처리에 있어 정밀한 의미 기술을 제공한다는 점에서는 유리할 수도 있으나 단어 중의성 해소(word sense disambiguation) 측면에 있어서는 별로 유리하지 않다(Ide and Wilks, 2006). *car*의 경우 다섯 가지 어휘 의미를 자동으로 파악해야 하기 때문이다. 또한 이렇게 규정되는 신셋은 엄밀한 의미에서 동의어라기보다는 유사어(similar words)에 더 가깝다.

KOLON에서는 이와 달리 개념적으로 규정되는 상당히 넓은 범위의 동의어가 형성된다. *car*의 경우 이에 대응하는 개념 AUTOMOBILE에는 '자동차', '경차', '자가용', '장갑자동차', '차', '차량', '차종', '화학자동차' 등 상당히 넓은 부류의 어휘들이 사상되어 있을 뿐만 아니라 행위를 나타내는 EVENT 개념에도 어휘들이 사상되기도 한다. '하차하다', '증차되다' 등이 그 예이다. 이는 다음 절에서 살펴 볼 다중 사상에 의해 나타난 결과로 어휘의 의미를 더 정교하게 명시하기 위해 여러 개념들에 동시에 사상된 결과이다.

3. 다중 사상(Multiple Mapping)

KOLON은 한 어휘가 여러 개념에 사상되는 다중 사상을 허용한다. 이 다중 사상에 의해 여러 부모로부터의 정보를 물려받는 다중 상속(multiple inheritance)이 이루어진다. 다중 사상은 어휘 사상에 있어, 새로운 개념을 가급적 도입하지 않으려고 할 때나 어휘의 의미를 더 상세하게 기술하려 할 때 주로 발생한다. 온톨로지 구축 원리 중의 하나는 가능한 제한된 개념 체계를 유지하는 것이다. 만일 어휘 의미 그대로의 개념이 필요하다고 하다면, 어휘 수만큼 개념이 필요하고 이는 어휘와 개념의 구분을 모호하게 한다. 예를 들어 '정확도'라는 어휘의 경우, 이에 직접 해당하는 개념은 없다. 이 경우 새로운 개념을 도입하는 대신, ABSTRACT-OBJECT와 PRECISION-ATTRIBUTE로 다중 사상하여 해당 어휘의 의미가 파악되도록 하였다.

이런 다중 사상은 개념적으로 관련된 어휘들을 파악하는데도 도움이 된다. 속성을 나타내는 개념인 SUCCESS-ATTRIBUTE가 다중 사상으로 명시되어 있는 명사들을 추출해 보면, '성공', '성사', '승리', '합격' 등이 있다. 이 어휘들은 모두 '성공'이라는 속성을 공유하고 있는 어휘들로, 워드넷과 같이 세분된 어휘 의미에 기반한 동의어 집합으로는 파악되기 어렵다.

Ⅲ. 언어 교육에의 적용

KOLON을 비롯한 어휘 자원은 일차적으로 자연언어처리를 위해 구축되었지만 언어 교육을 비롯한 여러 분야에도 활용될 수 있다. 언어 교육 중 한국어 어휘 교육에 초점을 맞추어서 이런 언어 자원들이 어떻게 활용될 수 있는지를 살펴보도록 한다. 한국어 어휘 교육은 여러 기준에 의해서 다양한 방법으로 이루어질 수 있지만, 여기서는 유의어 교육에 있어 개념으로 규정되는 유의어 그룹과 논항 정보와 선택 제약 등을 활용하는 방법에 초점을 맞춘다.

1. 한국어 교육에서의 어휘 교육

한국어 교육에 있어 어휘 교육은 주로 교육용 어휘를 선정하고 등급별로 분류하거나, 어휘의 의미적 속성에 따른 교육 방법을 구축하는 방식으로 이루어지고 있다. 조형일(2009 : 13)에 의하면, 어휘 교육 관련 연구의 경향은 1) 어휘 자체의 분석 논의, 2) 어휘의 등급화와 관련한 논의, 3) 어휘의 의미, 문법 정보 기술과 관련한 논의, 4) 어휘의 시소러스 구축과 전자 사전과 관련한 논의, 5) 어휘의 교육적 활용 방안과 관련한 논의로 구분할 수 있다. 이 중에서 어휘의 의미에 관련해서는 유의어, 동의어, 반의어, 상위어, 하위어, 다의어 등과 같은 의미관계를 교육에 활용하는 방법

에 관한 논의가 많이 이루어지고 있다.

어휘의 의미 관계를 교육하기 위해서는 학습자용 어휘를 선정하고 이를 위계화하는 것이 필요하다. 그래서 많은 연구들이 학습자 말뭉치를 활용하여 어휘목록을 선정하고 이를 등급화하고 있다. 실제로 송현주, 최현(2008)은 2002년 문화관광부 국어 정책공모 과제로 구축된 학습자 말뭉치를 사용하여 유의어 범위와 목록을 선정하고 교육용 유의어를 추출하였다. 이 말뭉치는 50만 어절 규모의 한국어 학습자 원시 말뭉치와 10만 어절 규모의 한국어 학습자 오류 말뭉치로 이루어져 있다.

이런 학습 말뭉치로 규정되는 의미 관계 중 하나인 유의어의 규정에 대해 살펴보자. 송현주・최현(2008)은 의미적 유사성에 따른 유의어와 사용역 차이에 따른 유의어를 구분하고 있다. 다음 문장은 의미적 유사성에 따른 유의군 오류를 보이는 예문이다.

(3) ㄱ. 지금은 저도 자기 가족을(√가정을) 일우었습니다.
ㄴ. 그 사람의 집(√가정) 생활만을 이해할 수 있다.
ㄷ. 어떤 친구는 제가 한국에서 취직하고 싶다고 말했으면 정부를 주고나 같이 PC방에 가서 찾고나 언제나 육친처럼(√가족처럼) 생각하고나.
ㄹ. 버스를 타면 아주 어렵고(√힘들고) 불편해요.

위 예문 ㄱ~ㄷ은 '가족 : 가정 : 육친 : 집'의 유의군에서의 오류를 ㄹ은 '어렵다 : 힘들다' 유의군의 오류 유형을 보여주는 것으로 '어휘의미의 유사성'에 기인한다. 이는 기본 어휘 의미의 차이가 크지 않거나 거의 없어 오류에 대한 근본적인 원인을 줄일 수 없다는 점에서 표제항 정보의 정밀성이 요구된다고 한다. 그러나 이런 문제는 사전적 의미 규정을 받아들이기 보다는 실제 학습자 말뭉치에서 나타나는 의미를 중심으로 유의어를 규정하는 것에서 비롯된다. 어휘의 다의 구분을 명확히 하지 않고서는 학습자 말뭉치의 나타나는 예로 표제항의 정밀성을 구하기는 쉽지 않다. '어렵다 : 힘들다'의 경우 '어렵다'의 다의 중 하나로 '힘이 들다'라는 뜻이 있기 때문에 다음과 같은 경우는 ㄹ과 달리 두 어휘다 사용가능하다.

(4) 버스를 타면 서있기가 아주 어렵고/힘들고 불편해요.

따라서 의미 관계를 규정할 때는 어의(sense)에 대한 명확한 구분이 필요하다. 그러나 한국어 어휘 교육에 있어서는 이런 명시적인 어의 구분을 하지 않은 경우가 많다. 사전기술이나 워드넷과 같은 언어 자원에서처럼 단어에 어의에 따라 번호를 붙여 구별하여 기술하는 방법은 실제 언어교육에 적용하기 어렵기 때문으로 보인다.

한국어 어휘 교육을 위한 자료 구축에 있어서도 어의를 명시적으로 구분하지 않는 것은 등급별 유의 관계 분류에서도 나타난다. 조형일(2009)는 동사 2,668개와 형용사 921개를 추출하여 시소러스를 구축하기 위해 동사와 형용사를 등급별 유의 관계로 구분하고 있다. 이때 가장 기본이 되는 어휘를 1등급 어휘로 하고 이를 출발점으로 하여 동일 등급 내 유의어와 2등급 3등급 유의어를 분류하고 있다. 다음은 조형일(2009 : 97)에서 제시된 '튼튼하다'와 '안전하다'의 등급별 유의어이다.

〈표 4〉 '튼튼하다'와 '안전하다'의 등급별 유의어

<table>
<tr><th>1등급 형용사 어휘</th><th colspan="2">형용사 유의어</th><th>1등급 형용사 어휘</th><th colspan="2">형용사 유의어</th></tr>
<tr><td rowspan="8">튼튼하다</td><td rowspan="4">1등급</td><td>건강하다</td><td rowspan="5">안전하다</td><td rowspan="3">1등급</td><td>괜찮다</td></tr>
<tr><td>씩씩하다</td><td>좋다</td></tr>
<tr><td>좋다</td><td>튼튼하다</td></tr>
<tr><td>안전하다</td><td>2등급</td><td></td></tr>
<tr><td rowspan="3">2등급</td><td>강하다</td><td>3등급</td><td></td></tr>
<tr><td>탄탄하다</td><td colspan="2" rowspan="3"></td></tr>
<tr><td>든든하다</td></tr>
<tr><td>3등급</td><td></td></tr>
</table>

조형일(2009)은 출발어휘의 방향성이 중요하여 한 어휘의 유의어로 제시된 어휘가 역방향으로도 완전히 그리고 항상 유의어가 되지 않는다고 한다. <표 4>에서 1등급 어휘인 '건강하다'의 유의어로 제시된 '튼튼하다'는 '안전하다'의 유의어로도 제시되어 있지만, '건강하다'와 '안전하다'

의 유의 관계는 인정되지 않는다고 한다. 이는 '건강하다'나 '안전하다'의 의미적 교집합의 영역에 '튼튼하다'가 올 수는 있지만 '건강한 것'과 '안전한 것'의 의미적 거리는 분명히 드러나기 때문이라고 한다.

그러나 이 '의미적 거리'가 무엇인지는 분명히 명시하지 않는다. 이 의미적 거리는 세종전자사전이나, KOLON 그리고 다른 사전 구분에 의하면 다의 속성에 의한 어의의 차이다. 실제로 '튼튼하다'는 KOLON에서 HEALTH-ATTRIBUTE와 INTENSITY로 사상되어 있다. 따라서 '건강하다', '안전하다'가 의미적 교집합을 갖고 있는 것은 '건강하다'는 '튼튼하다'와 HEALTH-ATTRIBUTE에서 '안전하다'는 '튼튼하다'의 INTENSITY에서 서로 관련을 맺고 있기 때문이다. 따라서 개념에 따른 다의 기술을 받아들인다면 이는 서로 다른 어의와 관련을 맺는 것으로 구분할 수 있다. 이를 방향성 등의 불분명한 개념으로는 설명하기 어려워 보인다. 결국 <표 4>에서 각 등급별 차이가 바로 해당 어휘의 다의적 속성에 따른 차이가 된다.

2. 어휘 교육을 위한 의미 관계 비교

한국어 처리를 위해 구축된 어휘 자원이 어휘 의미 관계를 어떻게 규정하고 있는지를 대표적인 어휘 자원인 KorLex와 낱말밭에서 제공하고 있는 한국어 유의어 대사전, 그리고 KOLON을 중심으로 살펴보도록 하자. KorLex는 영어의 워드넷을 대역하는 작업에서 시작하여 한국어에 맞게 그 신셋을 확장하고 있는 어휘 자원이다. 낱말밭[8]은 한국어 반의어, 방언, 유의어 사전 등 한국어 자원을 집대성하여 온라인 상에서 제공하고 있는 어휘 자원이다. 그 중에서 유의어 사전은 김광해 교수의 1987년부터의 연구 결과를 집대성하고 있는 것으로 101,781개의 표제어를 중심으로 1차 유의어 283,733개, 2차 유의어 2,001,129개가 수록된 대규모의 사전이다. 단순

8) http://www.natmal.com

히 유의어만 제공하는 것이 아니라 등급별로 어휘를 분류하는 난이도 사전이기도 하며 의미망을 제공하는 종합적인 사전이다. 논의를 신효필(2010)에서 제시된 '흡수하다' 어휘로 이루어지는 동의어 또는 유의어에 한정하도록 한다.[9] KOLON에서 '흡수하다'가 사상된 개념 ABSORB에 사상되어 있는 어휘들은 다음과 같다.[10]

(5)
먹다_01_16, 먹다_01_17, 먹히다_01_07, 물먹다_01_01 ,방음되다_01_01, 배어들다_01_01, 불리다_03_03, 빨다_01_01, 빨다_01_02, 빨리다_02_02, 빨리다_03_01, 빨리다_03_02, 빨아들이다_01_01, 빨아먹다_01_01, 삼투되다_01_01, 소화하다_01_06, 스며들다_01_01, 스며들다_01_02, 스며들다_01_04, 스며들다_01_05, 스미다_01_01, 잠그다_02_01, 잠기다_02_02, 젖어들다_01_01, 젖어들다_01_02, 철벅하다_01_01, 침윤시키다_01_02, 투수하다_01_01, 흡수되다_01_01, 흡수하다_01_01, 흡수하다_01_02, 흡열하다_01_01, 흡인되다_01_01, 흡인하다_01_01, 흡입되다_01_01, 흡입되다_01_02, 흡입시키다_01_01, 흡입하다_01_01, 흡착되다_01_01, 흡착시키다_01_01, 흡착하다_01_01, 흡착하다_01_02

다음은 KorLex에서 '흡수하다'로 이루어지는 신셋들이다.

(6)
흡수하다 1, 흡수시키다 1(absorb 5)
동화하다 1, 흡수하다 1(imbibe 4)
흡수하다 1(absorb 8)
빨아들이다 1, 흡수하다 1, 빨아들이다 2, 흡수하다 2(absorb 4, suck 4, imbibe 1, soak_up 1, sop_up 1, suck_up 1, draw 24, take_in

9) 동의어, 유의어는 엄밀한 의미에서 서로 다르다. 그러나 실제 어휘 자원에서는 이를 명시적으로 구분하지 않는 경우가 많다. 유의어는 시소로스 관점에서 어휘를 분류할 때 선호되며, 동의어는 사전 기술에서 여러 의미 관계 중 하나로 규정된다. 워드넷에서 신셋은 동의어를 나타내고 낱말밭에서는 유의어로 비슷한 의미를 지니는 단어들을 분류하고 있다. 이 논문에서는 이 둘을 엄밀히 구분하지 않는다.

10) 표제어 뒤에 첫 번째 숫자는 동음이의어 구분 번호이며, 그 다음 숫자는 다의 구분 번호이다. 번호 다음에는 품사표지가 붙어 있으나 지면상 여기서는 생략하였다.

13, take_up 11)
빨아들이다 1, 흡수하다 1, 빨아들이다 2, 흡수하다 2(absorb 4, suck 4, imbibe 1, soak_up 1, sop_up 1, suck_up 1, draw 24, take_in 13, take_up 11)
흡수하다 2(take_in 12, take_up 10)
빨아들이다 2, 흡수하다 2(absorb 6, take_in 3)
흡수하다 3, 흡착하다 1(sorb 1, take_up 8)

'흡수하다'의 동의어로 KorLex에서 추출될 수 있는 어휘는 '동화하다, 흡수시키다, 빨아들이다, 흡수하다, 흡착하다' 등이다. 이는 어휘들의 미세한 동의 관계인 신셋에서 출발한 워드넷의 특성이기도 하다. 신효필(2010)은 KOLON에서 ABSORB에 사상되어 있는 어휘들을 KorLex에서 해당되는 신셋을 찾거나 아니면 근접한 신셋을 찾아 각각 어떻게 분류되고 있는지를 보이고 있다. 지면의 제약 상 그 일부만을 예시하도록 한다..

〈표 5〉 신효필(2010)에서의 어휘 비교 중 일부 발췌

KOLON	KorLex 신셋	KorLex 해당(근접) 신셋
먹다1_16	• 늙다1, 먹다1 • **먹다2, 섭취하다1** • 먹다2(eat1) • 먹다2(feed6, eat3) • 따 2, 따먹다1, 먹다3	• 먹다2, 섭취하다1 ↳ 술을_과하게_마시다1, 술을 많이 마시다1, 과하게_음주하다1, 지나치게_음주하다1 사람의_고기를_먹다1, 인육을_ 먹다1 습관적으로_먹다1, 습관적으로_섭취하다1 식사하다1 먹다2 마시다1 끽주하다1, 음주하다1 손대다1 먹다2 포식시키다1

KOLON	KorLex 신셋	KorLex 해당(근접) 신셋
먹혀들다1_2	• 먹혀들다1(Wordnet에는 없음)	• 인정하다1, 받아들여지다1 ↳ 먹혀들다1
먹히다1_7	**• 먹히다1(Wordnet에는 없음)** • 들다5, 먹히다2	• 실행하다1 ↳ 관여하다1, 참여하다1 ↳ 경험하다1, 겪다1 ↳ 경험하다1, 겪다1, 당하다1 ↳ 먹히다1
물먹다1_1	• 물먹다1(Wordnet에는 없음)	• 실패하다2, 실패하다1 ↳ 물먹다1
방음되다_1_1	해당 어휘 없음	

신효필(2010)에 의하면 ABSORB에 사상된 어휘들이 KorLex에는 없는 경우가 많다. '먹다'(화선지가 물을 먹다와 같은 경우)에서 파생되는 의미는 제대로 포착되지 않는다. '먹다, 먹혀들다, 먹히다, 물먹다, 빨아먹다' 등에서는 스며드는 의미가 나타나지 않는다. 특히 '물먹다'는 기본의미보다는 비유적인 의미(실패하다)로만 기술되어 있다. 이는 KorLex가 영어의 대역에 기반하기 때문이다. 즉, 영어의 eat이 한국어의 '스며들다'의 의미를 갖지 못하기 때문으로 보인다.

우리말 유의어 대사전에서 '흡수하다'의 유의어로 '빨다01-01', '받아들이다02', '빨아들이다03', '섭취하다03' 등이 일차적으로 도출된다. 우리말 유의어 대사전도 워드넷과 같이 어휘의미에 기반한 유의어 그룹을 형성하기 때문에 세밀한 유의어 그룹을 갖게 된다. 이 어휘들의 유의어를 각각 선택하면 관련된 다른 유의어 그룹으로 확장될 수 있다. 예를 들어, '섭취하다03'의 유의어를 선택하면 '먹다02-01', '받아들이다02', '빨아들이다03, 흡수하다02-03', '취하다01-01'로 다시 확장된다. 이 경우 '취하다01-01', '먹다02-01'의 새로운 유의어들이 획득된다. 그러나 이 경우에도 어디까지 확장해야 관련된 유의어를 얻게 되는지는 확실하지 않다. 확장된 유의어 '취하다01-01'의 유의어를 선택하며, '해석하다, 가지다, 나타내다, 보이다, 빌리다, 꾸다, 섭취하다, 융통하다'로 확대되어 '섭취하다' 외에는 더 이상의 확장이 어려워 보인다.

3. 어휘 교육을 위한 문법 관계에 따른 유의어 어휘 군집

KOLON에 의해 형성되는 유의어 집합은 워드넷 기반의 동의어 보다 더 넓은 범위의 어휘들을 포함하고 있음을 살펴보았다. 이제 이렇게 넓게 형성된 유의어들을 어떻게 세분하여 한국어 어휘 교육에 활용할 수 있는지를 살펴보도록 하자.

조형일(2009)를 비롯한 여러 연구에서 유의어 교육은 어휘를 단순히 비교해서 설명하는 교육방법에서 벗어나 문형과의 연계를 고려한 교육 방안이 필요함을 지적하고 있다. 실제로 조형일(2009)는 'A-군요, A/V-(으)르까요? A/V-지요?, A/V-아요/어요, A/V-(으)르거에요' 등의 격틀과 등급화된 형용사, 동사들을 구분하여 제시하고 있다.

KOLON은 앞에서 서술한 바와 같이, 개념체계에 사상되어 있을 뿐만 아니라 논항과 그 의미제약을 포함한 다양한 통사, 의미적 정보를 포함하고 있는 어휘부이다. 따라서 한 유의어 그룹에서 나타나는 어휘들을 여러 기준에 의해서 더 세분하여 분류할 수 있고 이를 한국어 어휘 교육에도 활용할 수 있다. 일례로 앞 절에서 살펴 본, ABSORB에 사상된 어휘들이 논항구조나 의미 제약에 따라 어떻게 더 세분될 수 있는지를 살펴보도록 하자. 우선 ABSORB에 사상되어 있는 넓은 부류의 유의어들을 격틀을 기준으로 세분해 보면 다음과 같다.

〈표 6〉 격틀에 의해 구분된 ABSORB에 사상된 어휘들

어휘	격틀제약
물먹다_01_01, 방음되다_01_01, 빨리다_02_02, 젖어들다_01_01, 철벅하다_01_01,	X=N0-이 V
배어들다_01_01, 잠기다_02_02, 젖어들다_01_02, 흡수되다_01_01, 흡입되다_01_01, 흡착되다_01_01, 흡착하다_01_02	X=N0-이 Y=N1-에 V
삼투되다_01_01, 스며들다_01_01, 스며들다_01_02, 스며들다_01_04, 스며들다_01_05, 스미다_01_01, 투수하다_01_01, 흡인되다_01_01,	X=N0-이 Y=N1-에|로 V

어휘	격틀제약
먹다_01_16, 빨다_01_01, 빨다_01_02, 빨아들이다_01_01, 빨아먹다_01_01, 소화하다_01_06, 흡수하다_01_01, 흡수하다_01_02, 흡열하다_01_01, 흡인하다_01_01, 흡입하다_01_01, 흡착하다_01_01	X=N0-이 Y=N1-을 V
빨리다_03_01, 빨리다_03_02, 흡입시키다_01_01	X=N0-이Z=N2-에게 Y=N1-을 V

격틀은 논항구조를 보여주는 것으로 같은 유의어 그룹에서도 문형에 따라 더 세분된 어휘들로 분류할 수 있게 한다. 흡수되는 대상만을 요구하는 어휘들(물먹다, 방음되다 등)과, 어떤 대상과 그 흡수되는 구조를 요구하는 어휘들(배어들다, 잠기다 등), 그리고 어떤 주체가 대상을 흡수하여 취하는 구조로 쓰일 수 있는 어휘들(먹다, 흡수하다 등)로 세분된다. 또한 어휘들이 취하는 논항들의 의미적 속성에 따라 구분할 수도 있다. 다음 표는 AGENT, DESTINATION, THEME과 그 값들의 제약에 따른 구분이다.

〈표 7〉 의미적 속성에 따른 어휘들의 세분화

어휘들	SLOT	FILLER
불리다_03_03, 빨다_01_01, 빨다_01_02, 빨리다_03_01, 빨리다_03_02, 흡입시키다_01_01, 잠그다_02_01	AGENT	HUMAN
흡수되다_01_01, 흡수하다_01_01, 흡인되다_01_01, 흡착되다_01_01, 흡착시키다_01_01	DESTINATION	PHYSICAL-OBJECT
스며들다_01_04, 흡입되다_01_01		ANIMAL-PART
먹다_01_16, 먹다_01_17, 먹히다_01_07, 빨다_01_01, 스며들다_01_01, 흡수되다_01_01, 흡수하다_01_01	THEME	LIQUID

행위의 주체인 AGENT가 인간(HUMAN)이어야 하는 어휘들로 '불리다, 빨다, 빨리다, 잠그다' 등이 해당된다. '아기가 젖을 빨다'와 같은 문장이 그 예이다. 이런 제약은 THEME이 액체적 속성을 지닌 LIQUID일 경우 더 명확해진다. '먹다, 먹히다, 빨다, 스며들다, 흡수되다, 흡수하다' 등의 유의어 집합이 이에 해당된다.

논항정보, 의미 제약 등은 서로 결합하여 더 세분된 유의어 그룹을 형성할 수도 있다. 다음은 THEME으로 PHYSICAL-OBJECT를 갖는 어휘들이 격틀에 따라 다시 하위 유형으로 나뉘는 경우이다.

〈표 8〉 논항정보와 의미 제약의 결합에 의한 분류

어휘	SLOT	FILLER	격틀
소화하다_01_06, 흡열하다_01_01, 흡인하다_01_01, 흡착하다_01_01	THEME	PHYSICAL-OBJECT	X=N0-이 Y=N1-을 V
물먹다_01_01			X=N0-이 V
불리다_03_03, 잠그다_02_01			X=N0-이 Y=N1-을 Z=N2-에 V
흡착되다_01_01, 흡착하다_01_01, 잠기다_01_01			X=N0-이 Y=N1-에 V
흡인되다_01_01			X=N0-이 Y=N1-에|로 V

Ⅳ. 결론

지금까지 자연언어처리를 위한 어휘 자원과 KOLON에 대해 살펴보고 이를 어휘 교육에 적용하는 방법에 대해 논의하였다. 자연언어처리를 위한 어휘 자원은 세밀한 어의 구분과 풍부한 정보 기술을 지향하고 있다. 한국어 어휘 교육은 특히 유의어 교육에 있어서는 어의에 기반한 유의어보다는 어휘 항목의 기본 의미에 초점을 맞추고 있는 것으로 보인다. 이런 접근은 다의적 기술을 단순화하여 교육에 적용할 수 있는 장점이 있지만 각 어휘가 지니는 다양한 의미적 속성을 체계적으로 분류하여 설명하기는 부족해 보인다. 이런 점에서 세밀한 다의 기술에 기반하여 의미 관계를 체계화하는 여러 어휘 자원들이 도움이 될 수 있다.

KOLON은 개념과 어휘 의미를 구분하고 어휘를 개념 구조에 사상시켜 개념이 지니는 다양한 정보를 어휘부에 계승하여 종합적인 어휘부를

구축할 수 있도록 하였다. 논항정보와 의미 제약 등을 포함한 정보를 포함시키기 위하여 세종전자사전 작업 결과를 통합하고, 논항구조에 나타나는 명사부류를 개념으로 다 전환하였다. KOLON에 의해 나타나는 동의어 내지 유의어 그룹은 다른 어휘 의미망 연구에 비해 포괄적인 집합으로 나타난다. 이를 한국어 어휘 교육에 적용함에 있어서는 격틀이나 의미 제약과 같은 정보에 따라 더 세분하여 유형화할 수도 있다. 그러나 이런 어휘 교육에 적용은 실제 언어 교육 전문가들에 의해 더 검증되어야 한다. 논문의 한 심사자가 지적한 바와 같이 완전하지 않은 어휘망을 가지고 어휘 교육에 적용하는 것은 시기상조일 수도 있다. 그러나 어느 어휘망이나 사전 기술도 완벽할 수는 없으며, 기존의 방법과 달리 의미 관계를 논항 구조나 의미 제약 등으로 체계화하여 어휘 교육에 적용할 수 있는 방법을 모색해 보는 것도 충분히 의의가 있는 일이다.*

* 본 논문은 2010. 10. 31. 투고되었으며, 2010. 11. 9. 심사가 시작되어 2010. 11. 29. 심사가 종료되었음.

▣ 참고문헌

국립국어원(2002), 『현대 국어 사용 빈도 조사』, 국립국어원.

송현주 · 최준(2008), "한국어 교육용 유의어 사전 편찬을 위한 표제어 선정 및 기술방안에 대한 연구", 『어문논총』 48, 한국문학언어학회.

신효필(2004), "온톨로지(Ontology)를 기반으로 하는 개념구조와 어휘기술", 『어학연구』 4-3, 서울대학교 언어교육원.

신효필(2007), "미크로코스모스 온톨로지로의 한국어 기본 동사의 사상", 『언어학』 49, 305-324.

신효필(2010), "KOLON(the KOrean Lexicon mapped onto the Mikrokosmos ONtology) : 한국어 어휘의 미크로코스모스 온톨로지로의 사상과 언어 자원의 결합", 『언어학』 56, 159-196.

옥철영(2006), "한국어 워드넷 개발과 상위 온톨로지", 『제1회 언어 중립적 온톨로지 워크샵』, 한국외국어대학교.

윤애선 · 황순희 · 이은령 · 권혁철(2009), "한국어 어휘의미망 「KorLex1.5」의 구축", 『정보과학회논문지 : 소프트웨어 및 응용』 36-1, 92-108, 한국정보과학회.

홍재성(2007), 『세종전자사전 최종보고서』, 문화관광체육부.

조형일(2009), "시소러스 기반 한국어 어휘 교육 연구", 서울대학교 국어교육과 박사학위 논문.

BalkaNet, http://www.ceid.uptras.gr/Balkanet

EuroWordNet, http://www.illc.uva.nl/EuroWordNet/

Ide, N. and Wilks, Y.(2006), Making Sense about Sense, Word Sense Disambiguation, *Algorithms and Applications*, 47-73, Springer, Heidelberg.

Mahesh, K.(1996), Ontology Development for Machine Translation : Ideology and Methodology, Memorandom in Computer and Cognitive Science MCCS-96-292, CRL, New Mexico State University.

MultiWordNet, http://multiwordnet.fbk.edu/english/home.php

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

WordNet, http://wordnet.princeton.edu/

<초록>

자연언어처리를 위한 한국어 어휘 자원과 언어 교육에의 응용

신효필

한국어 처리의 근간이 되는 어휘 자원을 구축하려는 시도가 그동안 행해져 왔다. KOLON은 어휘와 개념을 구분하여 어휘를 미크로코스모스 개념체계에 사상시키고, 세종전자사전의 통사, 의미정보를 추출하여 통합적인 어휘부를 구축하려는 시도이다. 또한 개념체계가 지니는 다양한 정보를 어휘부에 명시하여 통사와 의미 구조가 제약으로 작용할 수 있는 기재를 마련하였다. 명사, 동사, 형용사를 포함하여 현재 65,326개의 어의가 기술되어 있다. KOLON에 의해 개념별로 사상된 어휘들은 인지적 관점의 넓은 부류의 동의어 그룹을 형성한다. 격틀, 의미제약, 그리고 이 둘의 결합과 같은 조건으로 동의어 그룹을 더 세분화하여 문형 정보 및 의미 제약이 반영된 어휘 교육에 활용할 수 있다.

【핵심어】 한국어 처리, 어휘자원, KOLON, 언어 교육, 유의어

<Abstract>

The Korean Language Resources for Natural Language Processing and Their Applications to the Korean Language Education

Shin, Hyo-pil

Much effort has been spent on the construction of Korean language resources for Korean language processing. KOLON was one of these works. It created a unified lexicon for Korean by mapping Korean words onto Mikrokosmos concepts and combining the result with information from the Sejong Dictionary. The lexicon can obtain syntactic and semantic constraints on the argument structures of a word through inheritance of constraints originating from the conceptual structure. Currently KOLON contains 65,326 word senses collected from nouns, verbs and adjectives. Synonyms link to a single concept and cover a wide range of words. We can reclassify synonyms into small subgroups by putting syntactic and semantic constraints on their argument structures. These classifications can be a good resource for Korean language education as well as for Korean language processing.

【Key words】 Korean Language Processing, Language Resource, KOLON, Language Education, Synonyms