한국어 수분류사 어휘의미망 KorLexClas 1.5

(KorLexClas 1.5: A Lexical Semantic Network for Korean Numeral Classifiers)

황 순 희 † 권 혁 철 †† 윤 애 선 †††
(Soonhee Hwang) (Hyuk-Chul Kwon) (Aesun Yoon)

요 약 본 연구의 목적은 한국어 수분류사 체계를 설정하고, 수분류사와 공기명사 간 의미관계 정보를 제공하는 한국어 수분류사 어휘의미망 「KorLexClas 1.5」의 정보구조와 구축방식을 소개하는 데 있다. KorLex 명사, 동사, 형용사, 부사가 영어 워드넷(Princeton WordNet)을 기반으로 참고구축 방식으로 개발된 것에 비해, KorLexClas 1.0버전과 이를 확장한 1.5버전은 직접구축 방식으로 개발하였다는 점에서, 수분류사의 계층구조와 언어단위 간 의미관계 설정은 매우 방대한 시간과 정교한 구축 방식을 요구한다. 따라서 작업의 효율성을 기함과 동시에, 구축된 어휘의미망의 신뢰성 및 확장성을 높이기 위해, ① 다양한 기구축 언어자원을 활용하되 상호 검증하는 절차를 거치고, ② 부분문장 분석방법을 이용하여, 수분류사 및 공기명사 목록을 확장하며, ③ 언어학적 준거를 기준으로 수분류사의 계층구조를 설정하고, ④ 수분류사와 공기명사 간 의미관계 정보를 제공하되 확장성을 확보하기 위해, KorLexNoun 1.5에 '최하위 공통상위노드(LUB : Least Upper Bound)'를 설정하는 방식을 택한다. 이러한 특성을 가진 KorLexClas 1.5는 기계번역을 비롯한 한국어정보처리의 제 분야에 응용될 수 있다.

키워드 : 수분류사, 어휘의미망, 공기명사, 의미범주, 최하위 공통상위노드, 코렉스 클래스 1.5

Abstract This paper aims to describe KorLexClas 1.5 which provides us with a very large list of Korean numeral classifiers, and with the co-occurring noun categories that select each numeral classifier. Differently from KorLex of other POS, of which the structure depends largely on their reference model (Princeton WordNet), KorLexClas 1.0 and its extended version 1.5 adopt a direct building method. They demand a considerable time and expert knowledge to establish the hierarchies of numeral classifiers and the relationships between lexical items. For the efficiency of construction as well as the reliability of KorLexClas 1.5, we use following processes: ① to use various language resources while their cross-checking for the selection of classifier candidates; ② to extend the list of numeral classifiers by using a shallow parsing techniques; ③ to set up the hierarchies of the numeral classifiers based on the previous linguistic studies; and ④ to determine LUB(Least Upper Bound) of the numeral classifiers in KorLexNoun 1.5. The last process provides the open list of the co-occurring nouns for KorLexClas 1.5 with the extensibility. KorLexClas 1.5 is expected to be used in a variety of NLP applications, including MT.

Key words : Numeral Classifier, Lexical Semantic Network, Co-occurring Nouns, Semantic Category, LUB(Least Upper Bound), KorLexClas 1.5

· 이 논문은 2007년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2009-0083761)

† 정 회 원 : 부산대학교 인문학연구소 연구교수
soonheehwang@pusan.ac.kr
†† 종신회원 : 부산대학교 정보컴퓨터공학부 교수
hckwon@pusan.ac.kr
††† 정 회 원 : 부산대학교 불어불문학과 교수
asyoon@pusan.ac.kr

논문접수 : 2009년 4월 29일
심사완료 : 2009년 10월 16일

정보과학회논문지: 소프트웨어 및 응용 제37권 제1호(2010.1)

1. 서 론

전 세계 모든 언어, 특히 문법체계를 가진 모든 언어에는 분류사(classifier)가 존재하는데[1-5] 영어나 프랑스어 등에 비해 한국어, 일본어, 동남 아시아어, 아프리카어 등 소위 '분류사 언어(classifier language)'에서 더 정교한 체계를 보인다. 분류사는 크게 명사 분류사(nominal classifier), 수분류사(numeral classifier), 일치적 분류사(concordial classifier), 술어 분류사(predicate classifier) 등의 네 가지 유형으로 구분할 수 있는데, 한국어에는 수분류사가 복잡한 양상으로 실현된다. 수분류사의 두 가지 주요 기능은 사물 또는 사건의 ① 부류화(classification) 또는 범주화(categorization)와 ② 수량화(quantification)이다. 수분류사는 수량사 또는 수량 표현과 공기하며, 특정 명사 범주에 대해 비교적 명확한 공기제약(co-occurrence restriction)을 갖는다.

분류사는 인지과학, 언어학, 전산학 등의 여러 학문 분야에서 중요한 연구 대상이 되어 왔다. 먼저, 인지과학적 관점에서 볼 때, 분류사는 부류화나 범주화와 밀접한 관련이 있다. 부류화란 사물 또는 사건을 유형화하고 하나의 범주로 묶는 정신활동인데, 이는 개인의 개별적 경험을 일반적 개념으로 추상화하는 과정으로 동질화(identification), 이질화(differentiation) 등과 함께 인간의 주요한 심리작용으로 이해된다[6]. 분류사는 인간의 부류화 작용을 언어로 표상한 전형적인 형태로, 인간의 인지 및 인식세계를 근원적으로 표상하는 장치이기도 하다. 언어학 특히 의미론 분야에서는 분류사의 선택이 공기명사가 지닌 대표적 의미자질(semantic feature)에 따라 결정된다는 강력한 공기제약(selectional restriction)에 주목한다. 전산학, 특히 자연언어처리(Natural Language Processing) 분야에서는 정확한 문장 분석과 자연스런 문장 생성에 분류사-공기명사 간 공기제약 정보의 구축과 처리가 반드시 필요하다[7,8].

본 논문에서는 2006년부터 2009년 3월 현재까지 구축한 한국어 수분류사 어휘의미망(lexical semantic network) KorLexClas 1.5의 정보구조와 구축방식을 소개하고자 한다. 명사, 동사, 형용사, 부사의 어휘의미 구조 정보를 표상하는 KorLex가 영어 워드넷(Princeton WordNet[9], 이하 PWN)을 기반으로 참고구축 방식으로 개발된 것에 비해, 분류사가 발달한 한국어에 필요한 KorLexClas는 직접구축 방식으로 개발되었으므로 매우 방대한 시간과 전문적 지식이 요구된다. 따라서 작업의 효율성을 기함과 동시에, 구축된 어휘의미망의 신뢰성 및 확장성을 높이고자, ① 다양한 기구축 언어자원을 활용하되 상호 검증하는 절차를 거치고, ② 부분문장 분석 방법을 이용하여, 수분류사 목록을 확장하였으며, ③ 언어학적 준거를 기준으로 수분류사의 계층구조를 설정하고, ④ 수분류사와 공기명사 간 의미관계 정보를 제공하되 확장성을 확보하기 위해 KorLexNoun 1.5에 '최하위 공통 상위 노드(Least Upper Bound, 이하 LUB)'를 설정하는 방식을 사용했다.

본 논문의 구성은 다음과 같다. 2장에서는 분류사 관련 국내외 선행연구를 검토하고, 3장에서는 본 연구에 사용된 언어자원과 자연언어처리 방법을 논의한다. 수분류사와 공기명사의 추출을 위해 선행 언어학 연구에서 추출한 분류사 목록, 사전의 정의문, 고빈도 어휘 목록, 대용량 말뭉치로부터 추출한 문맥정보를 상호 보완적으로 이용한다. 4장은 의미적 특성을 기초로 6가지 수분류사의 유형 구분과 하위 범주화를 제시한다. 5장은 수분류사 유형별로 KorLexNoun 1.5와의 의미범주 설정 방법을 제안하고, 이를 기반으로 설정된 의미범주의 결과를 제시한다. 6장에서는 본 연구의 최종 결과와 의의, 제한점을 밝히고, 향후 연구 방향을 제시한다.

2. 선행연구와 KorLexClas 1.5 구축단계

어휘의미망과 온톨로지는 개념과 의미를 계층적으로 구조화하여, 지식과 정보를 효과적으로 체계화하기에 용이하므로, 인공지능(AI), 자연언어처리, 정보통합(information integration), 전자 상거래(e-commerce) 등 다양한 분야에 이용될 수 있다. 특히 분류사는 공기명사의 의미자질과 특성에 의존하여 선택되기 때문에, 어휘의미망이나 온톨로지 구성을 위한 '개체'와 '관계' 설정이 비교적 용이하다. 분류사가 매우 발달한 한국어에서 ① 자연언어처리에서 미확인 개체명(unknown named entity) 예측에 단서를 제공할 수 있고, ② 자동번역에 활용되어 번역률 향상 및 개선에 기여할 수 있으며, ③ 기구축 온톨로지 및 어휘의미망, 시소러스(thesaurus) 등과 연동될 경우 정보추출 및 검색(information extraction or retrieval)의 성능 향상 등으로 활용될 가능성이 크다. 2.1은 분류사 어휘의미망 및 온톨로지와 관련된 선행연구의 장점과 한계를 논의하고, 2.2에서는 분류사 어휘의미망 구축을 위해 선결되어야 할 문제점을, 2.3은 KorLexClas 1.0 및 1.5의 구축과정을 기술한다.

2.1 선행연구 검토

분류사 전반에 관한 선행연구는 다음 네 가지 방향에서 수행되었다. 이론언어학 분야에서는 ① 개별어 분류사의 유형조사 및 유형화[1-5], ② 개별 분류사의 의미기술[10], ③ 개별 분류사의 의미분석[11,12]이 이루어졌다. 또한 ④ 자연언어처리 분야에서 활용할 수 있는 언어자원의 구축과 평가로는 [13-23] 등을 들 수 있다. 본 논문과 직접적인 관련이 있는 ④를 중심으로 선행 연구를 살펴본다. 자연언어처리 분야 중 기계번역(MT)이 분

류사와 관련된 언어자원을 가장 많이 필요로 한다. 개별 언어마다 분류사의 수와 종류가 상이하고, 공기하는 명사의 특성이 다르다. 따라서 주된 연구 대상은 개별 언어별로 ⓐ 분류사 목록을 작성하고, ⓑ 각 분류사가 공기하는 명사의 목록을 구축하며, ⓒ 다른 언어 분류사와의 대응관계를 구축하는 방법론 및 언어자원 개발하는 것이다.

F. Bond의 일련의 연구 중 [8,16,17,19]는 시소러스와 사전의 의미범주를 이용하여 일본어와 한국어 수분류사 체계의 자동 생성 방법을 연구했다. 한국어의 경우, 매우 적은 수의 수분류사를 대상으로 삼았으므로, 그 결과를 실제 활용하는 데는 제한이 있고 온톨로지의 기본적 관계를 형식화하지 않았다. [20]은 일본어 수분류사 체계를 영어와 같은 비분류사 언어에 사상(mapping)하기 위해, 문장정렬 방법(phrase alignment method)과 코퍼스 기반 추출 방법을 제안하였고, [21]은 분류사-공기명사 간 의미관계를 이용한 타이 어 수분류사 생성 알고리즘을 제안하였다. [20]과 [21] 역시 제한된 수의 수분류사-공기명사 쌍을 대상으로 하여 실제 응용은 힘들며, 분류사 선정 규칙과 정확한 평가 결과를 제시하지 않았다는 점이 문제점으로 지적된다. [22]는 한국어 의미부착 말뭉치(tagged corpus)와 의미코드 체계를 이용하여 자동번역에 이용될 수분류사 모듈의 반자동 구축 방법을 제안하였다. 구축에 사용된 자료는 과학기술분야 문서에서 추출한 600만 문장과 9,021개 수분류사-공기명사 쌍으로, 8개 레벨로 구성된 416개 의미코드('계층화된 명사범주'를 의미)가 수분류사에 할당되고 구축되었다. 그러나 특정 수분류사와 공기명사(또는 명사 범주)는 중복적인 의미 관계 설정이 가능하고, 그 규모가 상당히 크다는 점을 감안한다면, 이 연구의 수분류사-명사 의미범주는 상당히 제한된 수로 설정되어 있다. [23]은 952개 수분류사를 데이터베이스화 하려는 시도로 자연언어처리 분야에의 활용 가능성을 제시하였으나 500개 이상의 분류사를 뚜렷한 의미적 기준이 없이 동일한 범주에 분류하였고, 수분류사-공기명사 간 상관관계 설정 및 의미적 특성에 대해 충분히 기술하지 않았다. 자연언어처리 분야 연구에 속하는 본 연구진의 선행연구로 [13-15]가 있다. KorLexClas 1.0의 전반적 구축 방법의 효율성은 [13]에서, OWL을 이용한 KorLexClas 1.0의 온톨로지 형식화는 [14]에서, 기계번역에 활용될 수분류사 처리 모듈인 KCL-SYS의 개발과 성능 평가는 [15]에서 부분적으로 소개한 바 있다.

이밖에 이론 언어학 분야의 한국어 수분류사 연구 [24-29]는 동일 분야 여러 선행연구의 단점을 보완하였다. [24]는 동음이형어인 한국어 수분류사의 대부분이 중국어 수분류사에서 차용되어 온 점에 착안, 의미를 기반으로 한·중 수분류사 목록을 제시했고, [25]는 수분류사 중 사건성 수분류사의 특성을 분석했으며, [26]은 의존명사의 개별적 의미 분석을 통해 어휘 간 변별성(distinctiveness)을 분석했는데, 이중 일부가 수분류사의 의미정보에 해당된다. 또한, [27]은 자립명사의 수분류사적 용법을, [28]은 수분류사를 양수사와 서수사로 분류하고 의존명사에 따른 양수사의 하위범주를 시도했고, [29]는 수분류사 '개'의 특징을 논의하였다.

2.2 KorLexClas 1.5 구축을 위한 선결과제

한국어 수분류사 어휘의미망을 구축하려면 다음 3가지 문제를 해결해야 한다.

첫째, 수분류사는 형태적 관점 또는 기능적 관점에 따라 그 범위가 달라질 수 있다. 형태적 관점에서 수분류사는 의존명사에 국한되나, 기능적 관점에서 보면 공기명사의 수량화와 의미 속성의 명세화(specification)를 가능케 하는 의존명사 또는 일부 일반명사를 모두 포함한다.

둘째, 수분류사 및 공기명사의 추출이다. 앞에서 소개한 선행연구에서 이러한 목록이 산발적으로 소개되었으나, 각 목록의 크기가 매우 작으며, 각 목록 간 일치하는 비율도 낮다. 비교적 수분류사 목록이 큰 사전의 경우도 마찬가지다. 예를 들어, 『표준국어대사전』[30]에는 수분류사와 관련된 품사적, 통사적, 의미적 정보가 체계적으로 들어 있지 않으므로, 일관된 방식으로 수분류사와 공기명사를 추출하는 데 어려움이 있다. 『세종전자사전』[31]에는 통사정보 중 <단위표현> 항목이 들어 있어, 수분류사-공기명사 쌍을 쉽게 추출할 수 있으나, 그 수가 제한되어 있고, '놈-명', '의식-번, 차례, 회', '인신매매-차', '하늘-조각', '재능-개' 등과 같이 부자연스러운 결합이 다수 발견된다.

셋째, 수분류사-공기명사의 의미범주 설정이다. 예를 들어, '벌' (예: *바지 두 벌*), '켤레' (예: *장갑 세 켤레*) 등에서 '벌-바지', '켤레-장갑' 간의 의미관계를 밝히는 것이다. 각 수분류사와 공기하는 명사를 목록에 일일이 열거하는 것보다 의미범주를 설정하는 것이 수분류사-공기명사 목록의 확장성을 보장하며, 결과적으로 목록의 완전성을 획득할 수 있는 방법이다. 즉, '벌'과 공기하는 명사를 '바지, 스웨터, 한복, 양복, 재킷' 등으로 열거하면, 이 목록에 들어 있지 않은 '스판 바지, 목폴라, 스키니 진, 마고자, 카디건, 스리피스' 등은 '벌'과 공기관계를 갖지 못하게 된다. 따라서 '의류' 또는 '의복' 등의 의미범주를 할당하되, 동시에 의미범주에 포함되는 공기명사의 확장성을 확보해야 한다.

2.3 KorLexClas 1.5의 구축과정

그림 1에서 소개된 한국어 수분류사 어휘의미망 KorLexClas의 구축 과정은 앞에서 언급한 3가지 문제의 해결에 초점을 맞추고 있다.

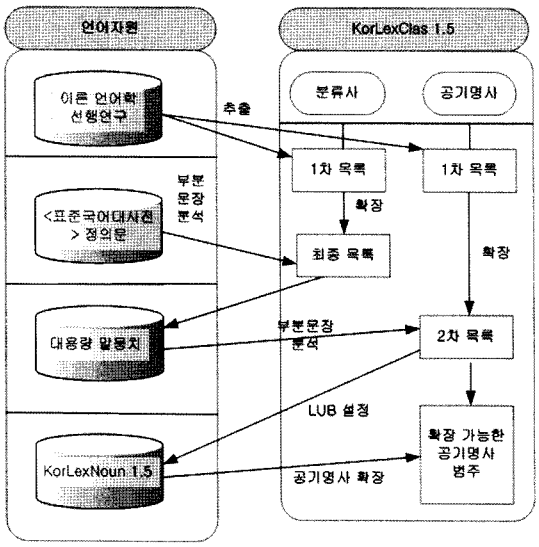

그림 1 수분류사 및 공기명사 목록 추출을 위한 언어자원 활용단계

첫째, 수분류사의 범위와 관련하여, KorLexClas 1.0에서는 형태적 관점에 따라 의존명사인 수분류사를 연구 대상으로 한정하였으나, 1.5버전에서 기능적 관점을 도입하여 그 범위를 확장하면서 일반 명사가 수분류사로 사용된 유형(예: '공기, 잔, 트럭, 더미, 덩어리, 뭉치, 조각' 등)을 추가했다. 두 버전의 크기 차이는 표 3에 기술하였다.

둘째, 수분류사 및 공기명사를 추출하고 유형을 구분하기 위해, 1단계로 선행 언어학 연구[25,26], 기구축 사전의 공기관계 설정 정보(『세종전자사전』) 및 정의문(『표준국어대사전』), 대용량 말뭉치의 문맥 정보를 이용하여 수분류사 목록과 공기명사 목록을 수집했다.

수분류사 목록 추출과 관련하여 한 가지 염두에 두어야 할 사실은 분류사 언어에 속하는 한국어를 비롯한 중국어, 일본어 등의 분류사는 대부분 동형이의어(homograph)에 해당하며, 중의성(ambiguity) 정도가 높다는 점이다. 따라서 중의적 수분류사의 경우 『표준국어대사전』의 정의문을 기초로 각각의 분화된 의미를 기준하여, 개별적인 수분류사로 처리한다. 가령, '대'는 다음과 같이 7가지 수분류사적 용법을 가지며 이들을 개별적 수분류사로 처리한다. 다음 표 1은 '대'의 사전 정의문과 분화된 의미이다.

2단계로, 수분류사의 의미적 특성을 고려하여, 개별 수분류사를 유형화하고, 하위 범주화한다. 그 결과 KorLexClas 1.0에서는 ① 도량성(mensural), ② 개체성(sortal), ③ 중립성(neutral), ④ 사건성(event)으로 나누었다. KorLexClas 1.5에서는 개체성 수분류사의 일부를 ⑤ 용기(container), ⑥ 준중립성(semi-neutral) 수분류사로 독립시켰다.

셋째, 수분류사-공기명사 간 의미범주를 설정하고 확장성을 확보하기 위해, 3단계로 한국어 명사어휘의미망(KorLexNoun 1.5)[32,33]과 연동하였다. 본 논문의 3, 4, 5장에서 각 단계의 구축 과정과 방법론을 구체적으로 기술한다.

3. 수분류사와 공기명사의 추출

선행 연구와 기구축 사전을 이용하여 수분류사 및 수분류사-공기명사 목록을 추출하는 데는 어려움이 있다. 선행연구의 목록은 일부에 지나지 않고, 기구축 사전의 정의문이 일관된 방식으로 기술되어 있지 않기 때문이다. 따라서 보다 완전한 수분류사 및 수분류사-공기명사 목록을 체계적으로 추출하기 위해 다음과 같은 언어자원과 자연언어처리 기법을 상호 보완적인 방법으로 이용한다.

첫째, 언어학 분야의 선행연구는 수분류사의 의미기술에 초점을 맞췄기 때문에, 여기에서 제시한 수분류사 및 공기명사의 목록은 제한적이며, 서로 다른 분류 기준을 사용하고 있다. [25]는 200여 개 수분류사를 제시했고, [26]은 수분류사를 위한 독립적 연구가 아니어서 이 역시 완전한 목록 추출에는 한계가 있다. 따라서 이러한 목록을 참고하되 일관된 목록의 추출이 필요하였다.

둘째, 수분류사 및 공기명사 목록을 보완하기 위해 『표준국어대사전』의 정의문을 이용했다. 예를 들어 표 2에서 볼 수 있듯, 수분류사로 쓰이는 표제어의 정의문에는 '수량', '~단위', '~세다(는)' 등의 정보가 들어 있으므로,

표 1 중의적 수분류사 '대'의 사전 정의문과 분화된 의미

의미 분화된 수분류사 '대'	『표준국어대사전』의 정의문
대01	「1」 담배통에 채워 넣는 담배의 분량이나 담배를 피우는 횟수를 세는 단위.
	「2」 때리는 횟수를 세는 단위.
	「3」 주사를 놓는 횟수를 세는 단위.
대06	「2」 가계나 지위를 이어받은 순서를 나타내는 단위.
대10	「3」 수량을 나타내는 말 뒤에 쓰여 편제된 무리를 세는 단위.
대11	「1」 두 짝이 합하여 한 벌이 되는 물건을 세는 단위.
대15	(의존명사) 차나 기계, 악기 따위를 세는 단위.

표 2 수분류사와 사전적 정의문 예시

분류사	『표준국어대사전』의 정의문
번	번04「의존명사」 ① 일의 차례를 나타내는 말. 둘째 번/다음번 면담은 너이다. ② 일의 *횟수*를 **세는 단위**. 여러 번/누구나 한 번은 겪는 일/몇 번을 그 앞을 왔다 갔다 하여 보았지만 (...). ③ 어떤 범주에 속한 사람이나 사물의 차례를 나타내는 단위. 4번 타자/1학년 2반 34번/1번 버스.
벌	벌02「의존명사」 ① *옷*을 **세는 단위**. *두루마기* 한 벌/*드레스* 두 벌. ② *옷*이나 *그릇* 따위가 두 개 또는 여러 개 모여 갖추는 덩어리를 **세는 단위**. *바지저고리* 한 벌/*반상기* 세 벌/*공구* 몇 벌...
갑	갑05「명사」 ② (수량을 나타내는 말 뒤에 쓰여) 작은 물건을 ①(갑)에 담아 그 *분량*을 **세는 단위**. *담배* 한 갑/*분필* 세 갑.
가지	가지04「의존명사」 사물을 그 성질이나 특징에 따라 종류별로 낱낱이 헤아리는 말. 두 가지 방법/그 예를 몇 가지 들어 보면 다음과 같다.
마디	마디01「명사」 ④ *말, 글, 노래* 따위의 한 도막. 몇 마디 이야기를 건네다/나는 그와 한두 마디 말만 했을 뿐 잘 아는 사이는 아니다.

정의문을 부분문장 분석하여 '수량', '~단위', '~세다' 등이 포함된 수분류사를 추출한다. 하지만, '가지, 마디'에서 볼 수 있듯이, 정의문의 통사적, 의미적 정보가 일관적이지 않으므로 위 정보만으로 수분류사나 공기명사를 자동 추출할 수는 없다.

셋째, KorLexClas 1.0에서는 전 단계에서 추출한 수분류사 목록이 <고빈도 어휘 목록>[34]에 들어있는지 확인하고, 최종 수분류사 목록을 마련하였다. <고빈도 어휘목록>은 한국어 학습에 필수적인 기초 어휘를 선정한 목록으로 어휘 학습 및 사용에 필요한 어휘 빈도를 제시하고 있는데, 그 중요도에 따라 1단계 982개, 2단계 2,111개, 3단계 2,872개, 총 5,965개 어휘가 포함되어 있다. 이중 수분류사 추출과 직접 관련된 어휘는 수사와 의존명사로 분류된 177개이다. 이 목록에는 대부분의 도량성 수분류사, 일반명사가 수분류사로 사용된 용법의 준중립성, 용기 수분류사가 포함되어 있지 않다는 점을 고려하면, 다수의 개체성 수분류사의 확인 및 추출에 상대적으로 유효한 어휘 목록이라 할 수 있다. 그러나 KorLexClas 1.5에서는 1.0버전에서 이미 기초적인 수분류사 목록은 추출되었기에, <고빈도 어휘목록>과의 매칭은 하지 않았다.

넷째, 이상의 과정에서 수분류사 목록과 일부 수분류사-공기명사 목록이 확보되면, 후자를 보완하기 위해, 본 연구진이 보유하고 있는 대규모 말뭉치(신문, 중학교 교과서, 과학 및 문학 텍스트, 법률 문서 등으로 구성된)를 대상으로, 수분류사와 공기 가능한 명사 목록을 반자동으로(semi-automatically) 구축한다. 대용량 말뭉치는 비구조화(unstructured text)된 경우가 대부분이어서 원하는 자료의 자동 추출이 어렵고, 또한 자동 추출하더라도 노이즈(noise : 비관여적 예)가 많기 때문이다. 수분류사와 관련된 내용으로 7,778,848개 어절, 450,000개 예문을 추출하여, 좌우 3어절 내 문맥과 의미를 확인하여 개별 수분류사의 공기명사 목록을 확보한다. 또한, 특징적으로 준중립성 수분류사의 경우 대규모 말뭉치[35]에서 수사적 용법과 그렇지 않은 예가 다수 섞여서 추출되기 때문에 결과를 직접 확인하고 공기명사 범주 설정에 참고하였다.

4. 수분류사의 유형 구분 및 하위범주화

수분류사의 유형으로, KorLexClas 1.0에서는 [4]를 참조하여 '도량성', '개체성', '중립성'을, [11]을 참조하여 시간성과 동작성이 대표적 특성으로 포함된 '사건성'을 구분하였다. 반면, '용기 수분류사', '준중립성 수분류사'는 ① 특정 명사 범주 또는 용언과 공기한다는 공기 제약을 갖는 다른 유형의 수분류사와 달리 매우 광범위한 명사 범주와 공기가 가능하고, ② 공기명사를 담는 용기 여부, 또는 공기명사 자체의 '모양'이나 '형태'가 주요 속성으로 작용하기 때문에, KorLexClas 1.5에서는 따로 구분하였다. 특히 용기 수분류사는 KorLexClas 1.5에서 기능적 관점에서 수분류사의 범위를 일반명사로 확장함에 따라 세분화된 유형이다.

개체성 수분류사("sortal classifier: individuating whatever it refers to in terms of the kind of entity that it is"[4])는 공기명사의 종류를 분류하는 특성이 강하며,

표 3 KorLexClas 1.5의 구축현황

유형	예	KorLex Clas 1.0	KorLex Clas 1.5
개체성	명, 축, 잎, 송이, 마리, 환, 권...	424	313
중립성	개, 가지, 종류, 종	4	4
사건성	번, 대, 건, 발, 방, 차례...	93	93
용기	그릇, 대접, 박스, 봉지, 컵...	-	65
준중립성	꾸러미, 다발, 도막, 등분, 쌍, 겹..	-	22
도량성	리터, 되, 미터, 센티미터, 마일..	856	1,115
합계		1,377	1,612

대부분의 수분류사가 이에 속한다. 개체성 수분류사의 하위범주화를 위해서는 [±living thing], [±animacy], [±human being], [±plant], [±shape] 등의 의미자질을 이용하여 계층구조화 하였는데, 이것은 수분류사 어휘의미망 및 온톨로지 구축에 필수적인 의미적 재범주화 과정이기도 하다. 대표적인 예로 '명, 축, 잎, 송이, 마리, 환, 권…' 등을 들 수 있다.

이에 비해, 중립성 수분류사는 특징적으로 대단히 넓은 의미 범주의 명사와 공기할 수 있다. 가령 '개'는 [+animate being] 자질을 가진 명사와 추상명사를 제외한 대부분의 명사와 공기하며, '개, 가지, 종류, 종'이 이에 속한다. '중립성'이라는 용어에 관해 여러 이견이 있을 수 있는데, 본 연구는 대부분의 수분류사가 특정 부류의 개체 부류하고만 공기하는 의미상 특화된(semantically specialized) 것과는 달리 중립성 수분류사는 거의 모든 종류의 개체 부류와 공기 가능하고, 의미상 중립적(semantically neutral)이라는 의미로 사용한다. 또한, 중립성 수분류사는 대부분의 수분류사 언어에서 공통적으로 발견된다[36].

사건성 수분류사는 공기명사가 동작성(행위)과 시간성을 지닌다는 점에서 통사적, 의미적 특성을 갖는다. 대표적인 예로 '번, 대, 건, 발, 방, 차례, 통' 등이 있다. 사건성 수분류사의 판단 기준은 2가지의 의미적 기준과 2가지의 통사적 기준이 사용될 수 있다[11]. 의미적 기준으로는 사건성 수분류사가 시간성(temporally anchored)을 지닌다는 점, 소유격으로 된 사격 논항(oblique argument) 구조(예: 친구의 전화 한 통, 간호사의 주사 한 방)로 환원할 수 있다는 점이며, 통사적 기준은 특정 술어와 공기하는 성질이 강하다는 점, 중립성 수분류사와 공기할 수 없다는 점이다.

수분류사는 기본적으로 의존명사이지만 '잔, 병, 방울, 더미' 등과 같은 자립명사도 수량 표현 뒤에 쓰이면 수분류사가 된다. 이때 '커피 세 잔'의 '잔'은 단순히 그릇으로서의 '잔'을 의미하는 것이 아니라, 문맥적 영향 또는 화용적인 요인에 의해 어떤 물질의 양과 관련된 크기가 문제되는 대상을 부각시킨다. 용기 수분류사는 주로 '공기, 그릇, 대접, 박스, 봉지, 숟가락, 자루, 잔, 차, 컵, 트럭' 등과 같이 담는 기능이 있는 '용기'인 일반 자립명사가 수분류사로 사용된 경우로, '공기, 그릇, 대접, 숟가락, 잔, 젓가락' 등에는 음식물하고만 사용된다는 제약이 추가되어야 한다.

준중립성 수분류사는 공기명사의 '모양' 또는 '형태'와 관련된 유형의 수분류사로 공기명사가 ⓐ 뭉쳐서 이루어진 것, ⓑ 나누거나 자른 것, ⓒ 두 개의 개체가 하나를 구성하는 것, ⓓ 면 또는 선이 거듭된 것으로 구성된다. 준중립성 수분류사는 주로 특정 모양이 정해지지 않은 대상을 수량화하기 때문에 공기명사의 범주 설정이 용이하지 않다. 예로 ⓐ '꾸러미, 다발, 더미, 뎡어리, 무더기, 묶음, 뭉치', ⓑ '도막, 동강, 등분, 조각', ⓒ '쌍, 짝', ⓓ '겹' 등이 있다.

도량성 수분류사("mensural classifier: individuating in terms of quantity"[4])란 특정 지시물의 수량 측정과 관계된 수분류사로, 시·공간과 관련된 단위, 계량단위, 통화단위 등으로 구성되며, 하위범주로 길이, 넓이, 무게, 부피 등을 추가할 수 있다. 대표적인 도량성 수분류사로 '넌, 리터, 되, 미터, 센티미터, 분, 도….' 등을 들 수 있다. 이처럼 도량성 수분류사는 공기명사를 범주화하는 것이 아니라 양을 표시하는 것으로 '럭스'(빛의 밝기 단위), '광년'(천체 사이의 거리 단위), '미터'(미터법에 의한 길이 단위) 등 대부분 측정단위(metric unit)이므로 별도의 의미분석과 재범주화, 의미범주의 설정을 하지 않고, KorLexNoun 1.5의 계층구조를 수용하였다.

5. 수분류사-공기명사 간 의미범주 설정

수분류사-공기명사 간 선택제약 관계는 자연언어처리 분야에서 가장 필요로 하는 언어자원이다. 하지만, 닫힌 목록(closed list)의 활용 범위는 매우 제한적이므로, 본 연구에서는 목록의 확장성을 확보하기 위해 KorLexNoun 1.5와 연동한다.

KorLexNoun 1.5는 PWN을 기반으로 1차 반자동 번역, 2차 수작업 수정 및 확장한 한국어 어휘의미망이다[33]. 약 9만 개의 신셋(synonym set: '동의어 집합'이라는 뜻으로 '어휘 단위로 표현될 수 있는 작은 개념'에 해당한다[9])과 10만 2천 개 정도의 다의어를 포함한다. 신셋 간 계층구조가 설정되어 있어, 상위노드와 하위노드 간에는 'is-a' 관계나 '전체-부분' 관계가 성립하며, 상위노드는 하위노드를 포함한다. 따라서 본 연구에서는 단순하게 수분류사의 공기명사 목록을 만드는 대신, KorLexNoun의 의미 계층구조를 이용하여 최하위 공통 상위노드를 설정함으로써, 사용빈도가 높은 분류사의 열린 목록(open list)을 제시하였다.

5.1 개체성 수분류사

(1) {피아노/ 컴퓨터/ ?책상} 한 대를 들여놨다.
(2) {책/ 잡지/ *신문}을 세 권만 추천해 주십시오.
(3) 친구라고 해서 {수박/ 양배추/ *사과} 한 통 싸게 주는 법이 없다.

사전적 정의에 의하면 '대'는 '차나 기계, 악기를 세는 단위'이며, '권'은 '책을 세는 단위', '통'은 '수량을 나타내는 말 뒤에 쓰여 배추나 박 따위를 세는 단위'로 '대'와 '권'은 상이한 의미 속성을 지닌 명사 범주와 공기한다. 즉, '대'를 '차, 기계, 악기' 등의 명사범주와 체계적으로 연결시켜 자동처리에 적합한 형태로 만드는 것이

관건이다. '권'은 '책' 이외의 다른 의미 범주와는 연결될 수 없기 때문에 LUB를 설정하는 것이 비교적 용이하다. 반면 '통'의 경우 이론적으로 예문정보, 대용량 말뭉치에서 추출한 문맥정보를 이용하면 '통'과 공기하는 모든 명사의 자동 추출 및 범주화가 가능하지만, 실제로 수분류사-공기명사 간의 노이즈, 어의의 중의성 해소(WSD) 등의 문제가 빈번히 발생한다. 따라서 공기명사의 LUB의 설정이 까다로운 경우에 해당하며, 표 4는 '통'의 LUB의 설정 과정을 요약한다.

이렇게 분류사의 공기명사(범주)를 LUB로 설정하는 이유는 가령, 특정 분류사 A의 공기명사를 개별 명사 단위가 아닌 범주로 설정하고, 그 모든 하위어에 동일한 의미제약을 가하여 최종적으로 A와 공기 가능한 명사를 특정 범주로 체계적으로 설정할 수 있다는 LUB의 유용성 때문이다. 이는 술어가 하위 범주화하는 논항의 선택제약을 설정하는 것과 유사한 효율성을 갖는다고 할 수 있다.

개별 수분류사와 공기명사를 연동하는 방식은 그림 2처럼 KorLexNoun에서 LUB를 찾는 것이다. '통'과 공기 가능한 명사인 '양배추, 멜론, 파인애플' 등의 명사 목록을 추출하고 선택 제약을 가한다. 그런 다음 공통 상위 노드 중 최소공배수 격인 LUB를 지정하고 그 하위 노드 전체에 제약규칙을 적용한다. '양배추, 멜론, 파인애플'의 LUB는 {청과물1}이지만 {청과물1}이 '양배추', '멜론', '파인애플'의 직속 상위 노드가 아니므로 최종적으로 LUB는 {양배추1}, {멜론1}, {파인애플1} 자신이 된다. 따라서 '통'과 공기하는 명사는 결국 {양배추1}, {멜론1}, {파인애플1}과 그 모든 하위어로 정해진다. 이것은 어휘의미망 KorLexNoun의 규모 및 의미 입자(semantic granularity)에서 비롯한다고 볼 수 있다.

수분류사가 공기하는 명사의 의미범주 범위(scope)를 기준으로 나눠보면 ① '개', '명', '마리', '군데' 등과 같이 공기 명사의 의미범주가 대단히 넓어서, 광범위한 명사 범주를 LUB로 설정할 수 있는 경우와 ② '갑'(담배, 성냥, 시계), '구'(시체, 시신), '쌈'(바늘), '보시기'(김치, 깍두기), '종지'(간장, 고추장, 기름) 등과 같이 공기 명사의 의미범주가 상당히 협소하여, 몇몇 특정 명사 범주에 국한된 LUB를 갖는 경우처럼 양 극단이 존재한다. '통'은 ①과 ②의 중간에 위치하는 유형이며, 대부분의 수분류사가 이 경우에 해당한다. 이와 관련된 LUB 설정의

표 4 '통'의 공기명사 설정 단계

① 1단계 : '통'과 공기하는 명사목록과 공기하지 않는 명사목록을 자동 생성한다.
예. '통(+)', '통(-)'으로 분류한다.
▶ 통(+) : {수박1, 파인애플1, 배추1, 양배추1...}
▶ 통(-) : {사과2, 포도1, 감1, 무2...}
② 2단계 : 1단계에서 추출된 명사를 KorLexNoun 목록에 사상시켜 지정된 노드의 모든 최하위 공통상위노드(LUB)를 추출한다.
만일 추출된 LUB가
▶ 구성 명사의 직속 상위어면, 구성 명사들의 LUB로 설정하며,
▶ 구성 명사의 직속 상위어가 아니면, 구성 명사 각각을 LUB로 설정한다.
③ 3단계 : 2단계에서 추출한 명사 목록으로부터 사상된 노드의 LUBs를 찾는다.
예. {양배추1}, {멜론1}, {파인애플1}이 LUBs 로 추출되고, 각 노드를 '통'과 공기하는 의미 범주로 결정한다.
④ 4단계 : LUB로 결정한 명사 각각을 '통'과 연결하고, '통'의 공기명사 의미범주로 확정한다.

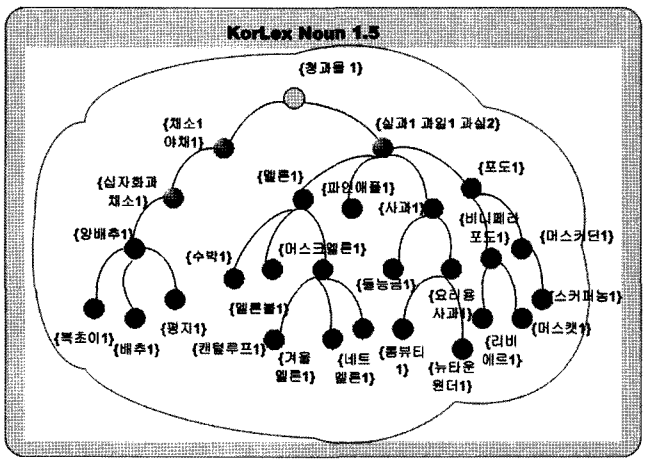

그림 2 '통'의 공기명사 의미범주

효율성 검토는 6절에서 논의한다.

한편, 설정한 LUB의 일부분이 해당 분류사와 공기관계를 맺고 있지 않아, Neg-LUB로 지정해야 할 경우가 있다. 예를 들어 수분류사 '마리'는 {동물1} 및 그 모든 하위어와 공기 가능한데, 그중 {인간1, 사람1} 및 그 하위어는 제외해야 한다. 이 경우 Neg-LUB를 지정한다. 이상의 방법을 통해 LUB가 설정되면 그 모든 하위어에도 특정 수분류사와 공기할 수 있으므로, 대부분의 경우 신조어 등이 어휘의미망에 추가·확장되어도 수분류사-공기명사 간의 관계를 자동으로 확장할 수 있다.

5.2 중립성 수분류사

다음은 사전적 정의로 유추해 본 중립성 수분류사의 공기명사 범주이다.

원칙적으로 명사 범주는 특정한 의미 속성을 지닌 수분류사 하고만 공기하지만, 상당수의 명사 범주는 중립성 수분류사 '개, 가지, 종류, 종'과도 공기한다. 이중 '가지, 종류, 종'은 KorLexNoun 1.5의 모든 노드와 공기한다. '개'는 수량화의 대상이 되는 명사를 범주화시키지 않는다는 의미로 중립성 또는 일반적 수분류사로 명명되곤 한다[29]. 또한, 중립성 수분류사는 다양한 부류, 대부분의 명사와 공기하는 것으로 알려져 있지만, '개'의 사용에는 표 6과 같은 제약조건이 따른다. 따라서 '개'의 공기명사 의미범주 설정에는 이러한 제약을 고려해야 한다.

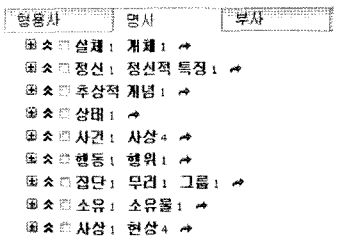

그림 3 KorLexNoun 1.5 최상위 노드

'개'의 공기제약 (a)-①에 의해 {실체1 개체1}, {집단1 무리그룹1}, {소유1 소유물1}를 LUB로 지정하되, 그중 {유기체 1 생물체1} 및 그 모든 하위어는 Neg-LUB로 지정한다. (a)-②의 신체 일부는 KorLexNoun 1.5 상 {실체> 사물> 부분> 몸부위> 신체외부기관> 부속지> 부기관...}으로 {유기체1 생물체1}과는 다른 계층 구조에 속하므로 따로 Neg-LUB가 필요치 않다. (b)는 그림 3의 {정신 1 정신적 특징1}, {추상적 개념1}, {상태1}, {사건1 사상1}, {행동1 행위1}, {사상1 현상1} 등에 나타나므로 모두 제외한다. 특히 (a)-②, (c), (d)은 '개'의 화용상 용법에 속하므로 공기명사 의미범주 설정에는 반영하지 않는다. 최종적으로 '개'의 공기명사 범주는 그림 4처럼 KorLexNoun 1.5의 9개 최상위 노드 중 {실체1 개체1}, {집단1 무리1 그룹1}, {소유1 소유물1} 노드와 그 모든

표 5 중립성 수분류사의 공기명사 범주

중립성 수분류사	사전적 정의문	정의문으로부터 유추 가능한 공기명사 범주
개	[의존명사]① 낱으로 된 물건을 세는 단위. 예. 사탕 한 개/사과 몇 개.	물건; 낱으로 셀 수 있는 것; 개체
가지	[의존명사]① 사물을 그 성질이나 특징에 따라 종류별로 낱낱이 헤아리는 말. 예. 두 가지 방법/그 예를 몇 가지 들어 보면 다음과 같다.	사물; 낱낱이 셀 수 있는 것, 생물체, 추상명사
종	[명사]③ 수량을 나타내는 말 뒤에 쓰여 종류를 세는 단위. 예. 서너 종의 견본/다섯 종의 서적	사물; 셀 수 있는 것, 생물체, 추상명사
종류	[명사]② 수량을 나타내는 말 뒤에 쓰여 갈래의 수를 세는 단위. 예. 서너 종류/이 옷은 부드러운 흰색의 음과 면, 두 종류로 만들었다.	사물; 셀 수 있는 것, 생물체, 추상명사

표 6 수분류사 '개'의 공기제약과 공기명사 의미범주

유형	중립성 수분류사 '개'의 공기제약	KorLexNoun 내 공기명사의 의미범주
(a)	① 사람을 포함한 모든 유기체, 생물체와는 공기할 수 없음. 예) '시체'와 사용할 수 없음 ② 신체의 일부(손가락, 이, 여드름...), 알, 식물(뿌리 또는 뿌리처럼 변한 줄기(감자, 고구마, 양파), 열매(사과, 오이, 수박)...)와는 공기할 수 있음	1. LUB : {실체1 개체1}, {집단1 무리그룹1}, {소유1 소유물1}의 모든 하위어 2. Neg-LUB : {유기체1 생물체1} 노드 및 모든 하위어
(b)	① 행위명사, ② 추상명사, ③ 물질명사, ④ 사건명사, ⑤ 자연현상, ⑥ 초자연적 대상 등과 공기할 수 없음	
(c)	고정된 형태를 갖춘 개체하고만 공기	
(d)	'개'는 공기명사 자체의 의미자질보다는 화용적 상황에 좌우됨. 물질명사도 일정한 형태와 크기를 갖춘 단위체로 바뀐 경우는 '개'와 공기할 수 있음. 예. 커피, 비빔밥, 냉면, 피자 등도 일정한 형태를 갖춘 경우 공기할 수 있음	

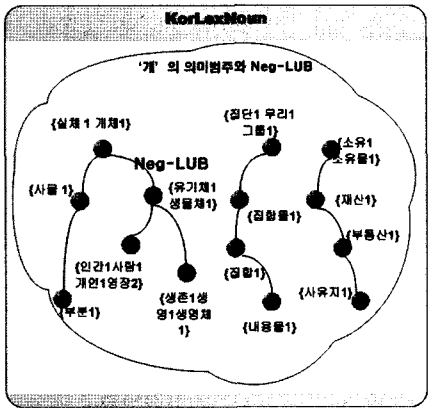

그림 4 '개'의 공기명사 의미범주

하위어가 되며, '개'가 유기체, 생물체와는 공기할 수 없기 때문에 {유기체1 생물체1} 이하 모든 하위어는 제외하며 이를 Neg-LUB로 지정한다.

5.3 용기 수분류사

용기 수분류사는 '공기, 그릇, 대접, 박스, 봉지, 숟가락, 자루, 잔, 차, 컵, 트럭' 등의 담는 기능이 있는 '용기'류 일반명사가 수분류사로 사용된 경우이다. 용기 수분류사는 후보어휘를 『표준국어대사전』에서 반자동으로 추출하여, 사전 정의문에 '담다, 넣다, -로 뜨다, 움켜쥐다, 집다, 쥐다, -에 싣다' 등이 포함되었는지를 확인하여 선정한다. 용기 수분류사는 표 7처럼 6가지 유형으로 분류된다. (e)에 속하는 수분류사는 '액체'만을 공기명사로 허용하며, 표 7에 제시된 LUB 및 모든 하위어가 공

표 7 용기 수분류사 유형과 LUB

구분	공기명사 범주 유형	개수	분류사 예	LUB
(e)	{액체}	11	구기, 국자, 대접, 동이, 배, 잔, 종구라기, 종발, 종지, 초롱, 탕기	1. {액체1 14090688}, {액체1 14090852} 2. {체액1} 3. {윤활제1 기름5}, {기름4}, {기름7}, {기름6} 4. {원유1} 5. {물1 07457438}, {물2}, {물1 14000512}, {물1 14002073}, {물3} 6. {음료1 드링크1} 7. {주정음료1} 8. {세면화장품류1} 9. {비4 강우1}
				1. '종지'는 {간장1}, {고추장1}, {기름3} 만을 LUB로 지정
(f)	{물질 성분}	12	가마, 가마니, 광주리, 기, 바구니, 움큼, 움큼, 자루, 접시, 젓가락, 주먹, 줌	1. {성분1 물질1}
				1. '기', '그릇', '젓가락'은 {영양분1} 만을 LUB로 지정
(g)	{액체}, {물질 성분}	24	갸자, 공기, 그릇, 바가지, 박스, 병, 봉, 봉지, 사발, 상자, 숟가락, 술, 스푼, 식기, 잎, 짝, 차, 차판, 캔, 컵, 통, 트럭, 포, 포대	1. 유형 (e)와 (f)의 모든 LUB
(h)	{액체}, {기체}	1	모금	1. {담배1} 2. {기체2} 3. 유형 (e)의 LUB 전체
(i)	차량·수단	3	바리, 짐, 짝	* '바리', '짝'의 LUB 1. {장작1 나무1} 2. {볏짚1} 3. {곡식1 07329605}, {곡식1 11394209} 4. {식물1} 5. {풀1 초본1} 6. {채소1 11464258}, {채소1 07235951} 7. {속씨식물1} 8. {목본식물1} 9. {꽃1}
(j)	특정물만 허용	14	갑, 궤, 곽, 다래끼, 달구지, 모숨, 목기, 목판, 보시기, 삼태기, 삽, 쌈지, 자밤, 지게	* '갑'의 LUB 1. {담배1 04271155}, {담배2 12157799} 2. {분필1, 백묵1} 3. {성냥1}
합계		65		

* 참고 : 위 표에서 '14090688'와 같은 8자리 일련번호는 KorLexNoun 1.5내 신셋의 ID 번호이며, 이것은 PWN 내 해당 신셋의 ID 번호와도 동일하다. {액체1 14090688}, {액체1 14090852}와 같이 동일 어휘이나 신셋이 다름을 표시하기 위해 필요한 경우 병기한다.

기명사가 된다. 단 '종지'의 경우는 '간장, 고추장, 기름'하고만 공기하므로 {간장1}, {고추장1}, {기름3}을 개별적으로 LUB로 지정한다.

(f)는 '물질 성분'만을 공기명사로 허용하므로, {성분1 물질1} 및 그 모든 하위어를 공기명사로 허용한다. 유형 (g)은 '액체'와 '성분 물질'의 모든 관련 명사군을 공기명사로 동시에 허용한다. 따라서 유형 (e)과 유형 (f)에 연동된 LUB가 모두 공기명사 범주가 된다. (h)는 '액체'와 '기체'의 모든 관련 명사군을 공기명사 범주로 동시에 허용하며, '모금'이 이에 해당한다. LUB로 (e)에 연동된 모든 공기명사와 '기체'의 모든 관련 명사군이 LUB가 된다. (i)는 '차량·수단의 개념이 포함된 수분류사로 '짐, 바리, 짝' 등이 이에 해당하며, (j)는 특정물하고만 공기하는 것으로 각 수분류사에 대한 의미 분석을 통한 LUB의 설정이 요구된다. 가령, '갑'의 LUB는 {담배1}, {성냥1}, {분필1}을 '보시기'는 {김치1}, {깍두기1}를, '삼태기'는 {흙1}, {쓰레기1}, {거름1}이 각각 공기명사로 지정되어 LUB를 설정한다.

또한 (f)와 (g)에 속하는 수분류사 중 '그릇, 기, 사발, 숟가락, 술, 스푼' 등은 표 8의 (k)처럼 '음식물'하고만 공기하므로, 이 경우 {영양분1}을 LUB로 설정하며 '젓가락'은 '음식물' 중 '액체'를 제외한 명사 범주를 LUB로 지정한다. 따라서 영양분을 LUB로 지정한 후, {음료1}, {물1}, {유미즙1}을 Neg-LUB로 설정한다.

5.4 준중립성 수분류사

용기 수분류사 이외에 개별적 분류가 요구되는 또 다른 유형의 수분류사가 준중립성 수분류사이다. 이 유형은 표 9처럼 크게 (l) 뭉쳐서 이루어진 것, (m) 나누거나 자른 것, (n) 두 개의 개체가 하나를 구성하는 것, (o) 면 또는 선이 거듭되는 것을 세는 수분류사로 나뉠 수 있다.

(l)은 주로 특정 모양이 정해지지 않은 대상을 뭉쳐진 모양으로 세는 데 사용되며, '꾸러미, 다발, 덩어리, 무더기, 묶음' 등이 이에 해당된다. (m)은 어떤 물건을 나누

표 8 '음식물'하고만 공기하는 용기 수분류사와 LUB

구분	공기명사 범주 유형	개수	분류사 예	LUB
(k)	(f), (g) 중 음식물만 해당	24	가자, 국자, 그릇, 기, 대접, 목기, 목판, 배, 보시기, 사발, 숟가락, 술, 스푼, 식기, 잎, 자반, 접시, 젓가락, 종구라기, 종발, 종지, 캔, 탕기	1. {영양분1}

표 9 준중립성 수분류사 유형과 개별 수분류사

구분	유형	개수	분류사 예	LUB
(l)	뭉쳐서 이루어진 것	11	꾸러미, 다발, 더미, 덩어리, 덩이, 동, 무더기, 묶음, 뭉치, 보따리, 아름	* '꾸러미'의 LUB 1. {성분 물질1} 2. {선물2 12509957}, {선물2 12513647} 3. {짐1} 4. {책1 02768681}, {책1 02769059}, {책2} 5. {편지1}
(m)	나누거나 자른 것	6	도막, 동강, 등분, 조각, 쪽, 토막	* '도막'의 LUB 1. {종이1 14123366}, {종이1 05874049} 2. {사진1 03531852}, {사진2 03777247} 3. {서류1} 4. {신문2} 5. {김밥1} 6. {수생척추동물1} 7. {장작1} 8. {이야기5 05972518}, {이야기1 05974336}, {이야기3 06672772}, {이야기3 06695510}, {이야기3 06699035}, {이야기1 06778553}, {이야기3 06779435}
(n)	두 개의 개체가 하나를 구성하는 것	4	조, 쌍, 짝, 켤레	* '켤레'의 LUB 1. {양말1} 2. {장갑1} 3. {신1 신발1}
(o)	면 또는 선이 거듭되는 것	1	겹	1. {종이1 14123366}, {종이1 05874049} 2. {서류1 문서1} 3. {사진1 03531852}, {사진2 03777247} 4. {신문2} 5. {천1}
합계		22		

거나 자르고 난 일부를 세는 수분류사로 '도막, 동강, 조각, 토막' 등이 있다. (l)과 (m)의 수분류사는 '아름, 쪽, 다발'을 제외한 나머지의 경우 수사 없이 일반 명사와 결합하여 사용되는 예가 많아 대규모 말뭉치에서 수사적 용법을 반자동으로 추출하는 것은 사실상 불가능하다. 가령, '토막'은 (...꼼장어 한 토막을 집어 먹었다.../...칼로 찌르고 토막 살인해 내버린 사건이.../...관계없이 말을 토막내는 일도 잦다.../...내가 살아온 이야기 중의 한 토막이어서 당연히 시시할 수밖에..) 등과 같이 다양한 구조로 실현된다. 따라서 용례추출기[35]를 이용하여, 수사적 용법을 직접 확인하고 공기명사 범주 설정에 참고하였다.

(l)의 '다발'은 '꽃, 푸성귀, 돈, 장작' 부류를, '아름'은 '꽃, 푸성귀' 부류를 공기명사로 취하므로 이를 LUB로 설정한다. '다발, 아름'을 제외한 (l)과 (m)에 속하는 나머지 수분류사의 공기명사 범주는 다소 차이가 있다. '더미, 꾸러미, 보따리'는 '선물, 짐, 소포, 신문, 책, 원고, 문서, 생선, 식품, 약, 과일, 도시락, 돈, 이불' 등이 공기명사로 추출되며, LUB는 {성분 물질1}, {선물2}, {짐1}이 된다. '덩어리, 덩이'는 '밥, 빵, 묵, 수박, 참외, 유기물, 땅, 흙, 똥' 등이 공기명사로 {성분 물질1}이 LUB이다. '무더기'는 {책, 편지, 국화, 꽃, 비둘기, 똥, 사람} 등이 공기명사로 {성분 물질1}과 더불어 생물체인 {꽃1}, {동물1}, {사람1} 등의 생물체도 공기명사로 사용될 수 있다. '묶음, 뭉치'는 '볏짚, 책, 짚, 바나나, 꽃, 꽁치, 생선, 낱말' 등이 공기명사로 추출되어, {성분 물질1}, {초본식물1} 등과 '추상적 실체(예: 낱말)'도 사용될 수 있다.

(m)의 '동강'은 '섬, 몸, 나라, 국토, 문학' 등 주로 추상적 실체 명사와 공기할 수 있고, '도막'은 '김밥, 나무, 이야기' 등의 구체물과, '등분'은 '옷, 모조지, 트랙, 케이크, 빵, 시간, 음력, 시기' 등의 구체물 및 추상물과 모두 공기할 수 있다. '조각'은 '빵, 떡, 헝겊, 과일' 등의 구체물, '쪽'은 '과일, 빵, 마늘' 등의 구체물과 공기한다. 또한 '토막'은 '생선, 고기, 나무, 연필, 양파, 필름, 이야기, 역사, 풍경' 등 구체물, 추상물과 모두 공기한다.

(n)의 '쌍'과 '짝'은 공히 '둘이나 한 벌의 것'을 세는 단위로 각각 '하나로 묶어' 세거나 '각각'을 세는 단위이다. '쌍'은 '부부, 연인, 동물, 물고기, 곤충', '반지, 가락지' 등의 '보석장식품', '총', '(몸의) 부속지' 등의 명사 범주를 공기명사로 취하므로 각 명사 범주를 LUB를 설정하고 그 이하 모든 하위어를 공기명사로 설정할 수 있다. '짝'은 '양말, 장갑, 신, 신발, 젓가락, 부속지(다리...)' 및 그 이하 모든 하위어를 공기명사로 가지며, 역시 이 노드를 LUB로 설정한다.

(o)의 '겹'은 '면과 면, 선과 선이 거듭됨을 세는 단위'로 {종이1}, {서류1 문서1}, {사진1}, {사진2}, {신문2}, {천1} 등이 LUB로 지정될 수 있다.

5.5 사건성 수분류사

사건성 수분류사는 '추상적 사건을 수량화'하는 수분류사로 [±time] 자질 유무에 따라, [+event]와 [+attribute]를 지닌 범주로 나뉜다. 또한, [+event]를 갖는 수분류사는 '반복성(repetition)' 이라는 의미 속성이 두드러진 경우를 [+repetition], 그렇지 않은 경우는 [+action]인 수분류사로 하위범주화한다. 그 자체로 '시간성'을 지니고 있지만, 보다 특징적인 대표적 의미자질을 고려하여 표 10처럼 동작의 횟수, 반복성, (단순) 시간성, 추상성([-time])이 포함된 수분류사로 재분류한다.

사건성 수분류사는 크게 3부류로 구분할 수 있는데, ① 다른 분류사와 마찬가지로 특정 유관명사와 공기하는 양상을 보이거나, ② 특정 범주의 공기명사 없이 단독 사용되거나, ③ 특정 술어하고만 공기하는 경우다.

① 유형으로, '게임, 등, 승, 패'는 {운동경기1}, {스포츠1}, {게임1}, {콘테스트1}와 '곡'은 {노래1}, {노래4}'와, '교시'는 {수업1}과, '급'은 {바둑1}, {태권도1}과, '끼, 끼니'는 {식사1}, {밥1}과, '단'은 {바둑1}, {태권도1}, {장기7}, {검도1}, {유도2}와, '대'는 {주사1}과, '박자'는 {노래1} 등과 LUB로 공기관계를 설정한다.

② 유형으로는 아래 예 (4)의 '번'처럼 사건성 명사 전체와 공기 관계를 갖는 것으로, 공기 용언의 범주도 매우 크다.

③ 유형의 사건성 수분류사는 예 (5)처럼 공기하는 동사의 수량 범위나 횟수를 명세화하여, 동사를 한정하거나 수식하는 부사어의 역할을 하는 경우다[25]. 도량성, 개체성, 중립성 수분류사 등은 공기명사와의 사이에 선택제약(selectional restriction)이 존재하지만, 이러한 사건성 수분류사는 동사와의 대응관계에 선택 제약이 존재한다. 따라서 '바퀴'는 '돌다, 달리다', '대'는 '때리다, 치다' 등의 동사와 '발'은 '쏘다'와, '방'은 '뀌다'와 '끼, 끼니'는 '먹다, 때우다' 등의 동사와 공기한다. 우리는 사건성 수분류사의 이러한 특성을 이용하여 사건성 수분류사와 KorLexVerb 1.5의 특정 용언을 연동하였다.

(4) 한 **번** {봤다, 먹었다, 갔다, 쳤다, 찌르다, 때우다...}
(5) 쌀이 한두 **끼니** {먹을, *할} 정도밖에 안 남았다.

5.6 도량성 수분류사

도량성 수분류사 추출은 KorLexNoun 1.5를 이용하였다. 우선, {단위1 12818586}과 그 모든 하위어를 추출한다. 다만, {단위1}의 하위어에 개체성 수분류사(예: 가구, 가락, 가래...)가 상당수 포함되어 있기 때문에, 다른 언어자원에서 추출한 개체성 수분류사와의 비교를 통해 모든 어휘 엔트리와 의미를 확인하여 노이즈를 제거한다. 한편, KorLexNoun 1.5에는 도량성 수분류사가 {단위1} 이외의 여러 노드에 다수 분포한다. {척도2}, {기본

표 10 사건성 수분류사의 LUB와 공기용언

구분	의미자질		개수	분류사 예	LUB	공기 용언
(p)	[+시간성]	[+동작]	22	걸음, 끼, 끼니, 대(매), 대(주사), 바퀴, 발(운동경기), 발(총포), 발(걸음), 발자국, 발짝, 방(방귀), 방(총포), 방(때리기), 방(사진), 배, 합, 회, 획...	* '대'의 LUB 1. {매1 01098386}, {매1 04202948} 2. {채찍1 04400273}	* '대'의 공기 용언 1. {맞다6} 2. {때리다2 01356759}, {때리다2 01369678}, {때리다2 01359510} 3. {치다1 00104050}, {치다3 01046227}
(q)		[+반복]	3	고팽이, 배, 벌	* '배'의 LUB 1. {새끼1 01248130}	* '배'의 공기 용언 1. {낳다1}
(r)		[+시간]	37	거리(연극), 거리(탈), 건, 게임(운동), 게임(정구), 격, 경, 곡, 교시, 대, 번, 조, 집, 차례, 쿼터, 회전, 퀘, 탕(다녀오는 횟수), 탕(일)...	* '게임'의 LUB 1. {운동경기1} 2. {스포츠1} 3. {게임1} 4. {콘테스트1}	* '게임'의 공기 용언 1. {하다2}
(s)	[-시간성]	[+추상성]	31	격, 급, 끗, 단(변속단계), 단(바둑, 유도 등), 도, 동, 등, 등급(별), 등급(단계), 번, 세, 순, 위, 점, 탕(목욕), 학년, 학점, 호(번지), 호(순서, 차례)...	* '단'의 LUB 1. {바둑1}, 2. {태권도1} 3. {장기7} 4. {검도1} 5. {유도2}	* '단'의 공기 용언 1. 특정 공기 용언 없음
합계			93			

량1}, {자기화도1}, {시간단위1}, {돈4}, {속도1}, {정수3} 등이 이에 속하며, 이것의 의미노드와 모든 하위어를 추출하여 도량성 수분류사로 선정하였다.

6. 결론 및 향후 연구

본 논문은 한국어 수분류사 어휘의미망 KorLexClas 1.5를 소개하고, KorLexClas 1.5의 구축 및 기계번역에서 사용되는 수분류사 모듈 KCL-SYS의 개발 시 선결해야 할 사안인 수분류사와 공기명사 의미범주의 효율적인 설정방법을 제안하였다. 이를 위해 수분류사와 공기명사 사이에 특정한 의미관계가 존재한다는 점에 착안하여 유형별 수분류사의 의미범주 설정을 개별적으로 시도하였고, 자연언어처리 기법을 이용하여 보다 효과적인 설정방법을 제시하였다. 수분류사의 체계적 추출을 위해서 『표준국어대사전』의 정의문을 기반으로 부분문장 분석기법과 반자동 추출방법을 병행하여 수분류사를 추출하였고, 공기명사의 의미범주 설정을 위해서는 의미범주와 의미적 계층구조가 포함된 적절한 언어자원과 이를 기반으로 생성된 알고리즘을 이용하였다. 또한, 공기명사의 최하위 공통상위노드인 LUB 개념을 도입, 사용하면 해당 명사의 상위어, 또는 하위어를 일일이 추가할 필요 없이 효율적으로 의미범주가 설정된다는 장점을 논의하였다.

분류사의 공기명사가 KorLexNoun 1.5에서 보이는 분포는 다음 표로 정리할 수 있다.

표 11은 분류사의 LUB로 설정된 노드가 실제로 KorLexNoun 상에 어떻게 분포하는지를 추출한 결과로, 설정된 LUB의 분포를 검토하기 위해 KorLexNoun의 lexicographer's file(어휘편찬자 파일)에 따라 어떤 분포를 보이는지 살펴보았다. lexicographer's file이란 KorLexNoun과 그 모델 어휘망인 PWN에서 개별 신셋을 의미에 따라 범주화한 것으로 특정 의미영역을 9개

표 11 공기명사의 KorLexNoun 1.5 내 분포 유형

분류사 유형	KorLexClas 1.5 개수	[1] LUB가 동일한 lexicographer's file에 속함		[2] LUB가 상이한 lexicographer's file에 속함
		① LUB가 구성 공기명사의 직속 상위어임	② LUB가 구성 공기명사의 직속 상위어가 아님	
개체성	313	138	134	41
중립성	4	0	0	4
용기	65	13	48	4
준중립성	22	0	9	13
사건성	93	31	27	35
합계	497	182	218	97

cf. 도량성 1,115개는 LUB를 설정하지 않음

로 분할한 최상위 노드(Top-node)를 의미하며, {실체1 개체1}, {정신1 정신적 특징1}, {추상적 개념1}, {상태1}, {사건1 사상4}, {행동1 행위1}, {집단1 무리1 그룹1}, {소유1 소유물1}, {사상1 현상4} 등이 9가지 의미범주에 해당한다. LUB가 동일한 lexicographer's file에 속해 있다는 것은 분류사의 LUB 각각이 개별 범주나 어휘 단위로 지정된 것이 아니라, 일정한 부류에 속한 의미범주로 설정되었고 이는 LUB의 필요성을 뒷받침한다.

LUB가 동일한 lexicographer's file에 속한 경우(표 11의 [1])는 '분류사' - {LUB} 간 연결 유형을 크게 ① '그루'(한 해에 같은 땅에 농사짓는 횟수를 세는 단위) - {농사1}; ② '건'(사건, 서류, 안건, 조항 따위를 세는 단위) - {문서1, 안건1, 조항1} 로 양분할 수 있다. 유형 ①은 LUB인 {농사1}이 구성 공기명사인 '논농사, 밭농사, 벼농사....' 등의 직속 상위어로 구성되었고, LUB를 구성 공기명사의 직속상위어로 설정한 경우가 이에 해당한다. 유형 ②는 '통' - {멜론1, 양배추1, 파인애플1}과 유사한 경우로 LUB를 구성 공기명사의 직속상위어가 아닌 개별 노드 각각으로 설정한 경우이다.

LUB가 상이한 lexicographer's file에 속한 경우(표 11의 [2])는 본 논문에서 제시한 '통' - {양배추, 멜론, 파인애플}의 예와는 달리 LUB 간 의미범주가 상이한 경우를 의미한다. 가령, 분류사 '가닥'(수량을 나타내는 말 뒤에 쓰여 한 군데에서 갈려 나온 낱낱의 줄이나 줄기 따위를 세는 단위)은 개체의 의미범주를 분류하는 본래적 속성보다는 개체의 특정한 모양과 같은 세부적 특질과 관련된다. 따라서 '모발, 빛, 물줄기, 실, 끈, 줄...' 등의 다양하고 상이한 의미범주에 속하는 명사와 공기하며, 다양한 의미범주에 분포된 여러 개의 노드를 LUB로 설정하게 된다.

수분류사 어휘의미망 구축은 다음 몇 가지를 향후 연구과제로 남긴다. 첫째, 일반명사가 수분류사로 사용되는 경우를 체계적으로 추출하고 공기명사 추출과 의미범주의 설정이 필요하다. 또한, 세부적 의미분석과 용기수분류사, 준중립성 수분류사 등과의 관련성 또한 검토해야 한다. 둘째, 본 연구에서 설정된 수분류사의 공기명사 범주는 전적으로 PWN과 KorLexNoun 1.5의 의미범주 및 계층구조에 의존되어 있다. 따라서 PWN과 KorLexNoun 1.5의 상하의 관계를 재고하여 부적절한 의미관계가 발견될 경우 KorLexClas 1.5에 반영해야 한다. 셋째, 우리는 KorLexClas 1.0을 기반으로 개발된 자동번역 시스템의 하위모듈인 수분류사 시스템(KCL-SYS)에 본 연구에서 제안한 공기명사의 의미범주를 반영하고 보완해야 한다. 아울러 대용량 데이터를 이용한 실험과 성능평가 등이 진행될 것이다. KCL-SYS는 검색하고자 하는 분류사를 입력하면 공기 가능한 명사 목록이 제시되며, 공기 가능한 분류사를 알고자 하여 해당 명사를 입력하면 공기 가능한 분류사가 출력되도록 설계되었다. KorLexClas 1.5의 성능 평가 실험을 위해 기개발된 KCL-SYS을 KorLexClas 1.5를 기초로 확장·보완하고, 대용량 말뭉치인 British National Corpus (BNC)를 이용할 예정이다.

참 고 문 헌

[1] K. Allan, "Classifiers," *Language*, 53(2), pp.285-311, 1977.

[2] W. Croft, "Semantic Universals in Classifier System," *Word*, 45(2): pp.145-71, 1994.

[3] C. Goddard, *Semantic Analysis: A Practical Introduction*, Oxford University Press, Oxford, 1998.

[4] J. Lyons, *Semantics*, 2 vols., Cambridge University Press, Cambridge, 1979.

[5] A. Wierzbicka, *Semantics: Primes and Universals*, Oxford University Press, Oxford, 1996.

[6] G. Lakoff, "Classifiers as a Reflection of Mind," (In C. Craig. ed.), *Noun Classes and Categorization*(pp.13-51), John Benjamins Publishing Company, Amsterdam/Philadelphia, 1986.

[7] K. Allan, *Natural Language Semantics*, Blackwell, Oxford, 2001.

[8] F. Bond, K. Paik, "Classifying Correspondence in Japanese and Korean," *Proc. of the 3rd PACLING 97*, pp.58-67, 1997.

[9] C. Fellbaum, ed. *WordNet - An Electronic Lexical Database*, MIT Press, Cambridge, 1998.

[10] P. Downing, "Pragmatic and Semantic Constraints on Numeral Quantifier Position in Japanese," *Linguistics*, 29, pp.65-93, 1993.

[11] C.R. Huang, and K. Ahrens, "Individuals, Kinds and Events: Classifier Coercion of Nouns," *Language Sciences*, 25, pp.353-373, 2003.

[12] Y. Matsumoto, "Japanese Numeral Classifiers: A Study of Semantic Categories and Lexical Organization," *Linguistics*, 31(4), pp.667-713, 1993.

[13] S.H. Hwang, A.S. Yoon, H.-C. Kwon, "Semantic Representation of Korean Numeral Classifier and Its Ontology Building for HLT Applications," *Language Resources and Evaluation*, 42-2, Springer-Verlag, pp.151-172, 2008.

[14] Y.I. Jung, S.H. Hwang, A.S. Yoon, H.-C. Kwon, "Formalization of Ontological Relations of Korean Numeral Classifiers," *Lecture Notes in Computer Science* 4304, Springer-Verlag, pp.1106-1110, 2006.

[15] S.H. Hwang, A.S. Yoon, H.-C. Kwon, "Semantic Feature-Based Korean Classifier Module for MT Systems," *Proc. of the 6th International Conference on Advanced Language Processing and Web Information Technology*, pp.146-154, 2007.

[16] F. Bond, K. Ogura, S. Ikehara, "Classifiers in Japanese-to-English Machine Translation," *Proc.*

of the 16th COLING 96, pp.125-130, 1996.
[17] F. Bond, K. Paik, "Reusing an Ontology to Generate Numeral Classifiers," *Proc. of the 18th Conference on Computational Linguistics*, pp.90-96, 2000.
[18] H. Guo, H. Zhong, "Chinese Classifier Assignment Using SVMs," *Proc. of the 4th SIGHAN Workshop on Chinese Language Processing*, pp.25-31, 2005.
[19] K. Paik, F. Bond, "Multilingual Generation of Numeral Classifiers Using a Common Ontology," *Proc. of the 19th ICCPOL 2001*, pp.141-147, 2001.
[20] M. Paul, E. Sumita, S. Yamamoto, "Corpus-based generation of numeral classifier using phrase alignment," *Proc. of the 19th COLING*, pp.779-785, 2002.
[21] V. Sornlertlamvanich, W. Pantachat, S. Meknavin, "Classifier assignment by corpus-based approach," *Proc. of the 14th COLING*, pp.152-159, 1994.
[22] 이기영·최승권·김영길, "영한 자동번역에서의 한국어 분류사의 반자동 구축방법", *제20회 한글 및 한국어 정보처리 학술대회 논문집*, pp.134-138, 2008.
[23] 남지순, "수량표현 명사구의 자동 불·한 번역을 위한 한국어 단위명사 유형에 대한 연구", *「한국프랑스학 논집」*, 54, 한국프랑스학회, pp.1-28, 2006.
[24] 곽추문, "한·중 양국 한자 분류사의 쓰임 비교", *「한국어의미학」* 7, 한국어의미학회, 1-28, 2000.
[25] 곽추문, 한국어 분류사 연구, 성균관대학교 국어국문학과 박사학위논문, 1996.
[26] 안정아, 현대 국어 의존 명사의 의미 연구, 고려대학교 국어국문학과 박사학위논문, 2007.
[27] 임홍빈, "국어 분류사의 성격에 대하여", 국어학의 새로운 인식과 전개, 김완진 선생 회갑 기념논총, pp.586-611, 1991.
[28] 유재원, "자연어 처리를 위한 수사의 하위 범주 분류", *「언어와 언어학」*, 제24호, 한국외국어대학교 외국어종합연구센터 언어연구소, pp.103-110, 1999.
[29] 채완, "국어의 분류사 '개'의 차용 과정과 의미", *「진단학보」*, N◦82, 진단학회, pp.193-215, 1996.
[30] 국립국어원, 표준국어대사전 1.0, 두산동아, 2001.
[31] 국립국어원, 세종사전. 2007.
[32] KorLex: http://corpus.fr.pusan.ac.kr/korlex/start.htm
[33] 윤애선·황순희·이은령·권혁철, "한국어 어휘의미망 「KorLex 1.5」의 구축", *「정보과학회지」*, 제36권 1호, 한국정보과학회, pp.94-110, 2009.
[34] 국립국어원, 현대 국어 사용 빈도 조사: 한국어 학습용 어휘 선정을 위한 기초 조사. 2002.
[35] 고려대 민족문화연구원 전자텍스트연구소 용례추출기, http://ikc.korea.ac.kr/cgi-bin/kwic/kwic.cgi
[36] K.O. Lee, "Role of Semantic and Syntactic Knowledge on the First Language Acquisition of Korean Classifiers," *Journal of Korean Association of Child Studies*, 18-2, pp.73-85, 1997.

황 순 희

1986년 이화여자대학교 불어불문학과 학사. 1988년 (프)Rouen 대학교 언어학과 석사. 1993년 (프)Paris 8대학교 언어학과 박사. 2006년~2008년 부산대학교 U-Port IT 산학공동사업단, 전임연구원 2008년~현재 부산대학교 인문학연구소, 연구교수. 관심분야는 전산어휘의미론, 온톨로지

권 혁 철

1982년 서울대학교 컴퓨터 공학과 학사 1984년 서울대학교 컴퓨터 공학과 석사 1987년 서울대학교 공학과 박사. 1992년~1993년 (미)Stanford 대학교 CSLI 방문 교수. 1987년~현재 부산대학교 정보컴퓨터공학부, 인지과학협동과정 교수 관심분야는 인간언어공학, 정보검색, 인공지능

윤 애 선

1982년 이화여자대학교 불어불문학과 학사. 1984년 이화여자대학교 불어불문학과 석사. 1989년 (프)Paris-Sorbonne 대학교 언어학과 박사. 1992년~1993년 (미)Stanford 대학교 CSLI 방문 교수 1987년~현재 부산대학교 불어불문학과, 인지과학협동과정 교수. 관심분야는 자연언어처리, 지식처리, 언어자원구축