제12회 동양학연구원 사전학 학술회의

"동아시아의 사전학(XII)"

어휘 의미망 구축과 사전

❑ 일시 : 2021년 8월 27일(금) 13:00-18:30

❑ 주최 : 단국대학교 동양학연구원

❑ 장소 : 단국대학교 죽전캠퍼스 국제관 502호

❑ 회의 방식 : 대면 및 비대면 방식 겸용

온라인 zoom 활용

회의 ID : 686 288 5386

pass word : 1234

관련문의 : 12200283@dankook.ac.kr/ 031-8005-2632

발표문과 토론문은 동양학연구원 홈페이지에서 내려받을 수 있습니다.

모시는 글

단국대학교 동양학연구원은 1970년에 설립된 이후 반세기 동안 동아시아의 역사와 문화에 대한 연구를 진행해 왔습니다. 본 연구원에서는 1977년부터 한자사전의 편찬을 주력 사업으로 정하고, 그동안 『한국한자어사전』, 『한한대사전』, 『이두사전』을 완간하였으며, 『한국고유한자자전』의 편찬을 마무리하고 있습니다. 또한 정기적으로 사전학 학술회의를 개최하여 사전학 분야에서 이뤄지는 연구 성과들을 검토하고 새로운 사전을 편찬하기 위한 기초 조사를 수행하였습니다.

2021년 사전학 학술회의는 동양학연구원에서 축적해 온 한자와 한문에 관한 방대한 데이터를 활용하면서, 한자 어휘의 의미망 체계를 바탕으로 하는 새로운 사전을 편찬하기 위해 기획되었습니다. 앞으로 우리에게 필요한 사전은 종이사전의 정보를 인터넷에 서비스하는 정도가 아니라 컴퓨터를 통한 정보의 산출과 이해가 가능한 기계 가독형 사전이 되어야 한다고 판단합니다. 오늘의 학술회의는 기계 가독형 사전을 모색하는 작업의 일환으로 어휘 의미망 구축의 필요성과 이를 사전으로 연결시키는 방안을 검토하려 합니다.

이처럼 뜻깊은 학술회의에 발표와 토론을 맡아 주신 선생님들께 감사의 인사를 드리며, 동양학연구원의 미래를 향한 도전에 전문가 선생님들의 지혜와 힘을 모아주시기 바랍니다. 감사합니다.

2021년 8월

단국대학교 동양학연구원장 **김 문 식** 배상

일 정 표

❑ 등 록 13:00~13:30

❑ 개회식 13:30~14:00

사회 : 박영미(단국대학교)

• 개회사 김문식(동양학연구원장)

❑ 발표 1부 14:00~15:30

1. 어휘 의미망에 대한 인식과 사전의 구성 14:00~14:30
   • 발표 : 최경봉(원광대학교)

2. 어휘적 응집성과 한국어 어휘의미망(Lexical Cohesion and Korean WordNet)
   - 체계기능언어학의 관점(from Systemic Functional Linguistics point of view) 14:30~15:00
   • 발표 : 한정한(단국대학교)

3. 중국어 어휘 의미망 구축과 활용 - CCD와 MCD를 중심으로 15:00~15:30
   • 발표 : 강병규(서강대학교)

❑ 토론 1부 15:30~16:00

• 좌장 : 김우정(단국대학교)

• 토론 : 정성훈(목포대학교), 이동혁(부산교육대학교), 허철(단국대학교)

❑ 휴식 16:10~16:30

일 정 표

❑ 발표 2부 16:30~17:30

4. 디지털화된 이종 언어자원의 연계와 인문학 연구의 확장성
   -『통합디지털한한대사전』과 KorLex의 연동 가능성 검토를 통해 - 16:30~17:00
   • 발표 : 윤애선(부산대학교)

5. 동아시아 4개 언어(한 · 중 · 일 · 베트남) 한자어 데이터베이스의 구축과 활용, 그리고 확장 가능성 17:00~17:30
   • 발표 : 신웅철(경성대학교)

❑ 토론 2부 17:30~18:00
   • 좌장 : 김우정(단국대학교)
   • 토론 : 최호섭(단국대학교), 김바로(한국학중앙연구원)

❑ 총평 18:00~18:30
   • 좌장 : 김우정(단국대학교)

❑ 폐회식 18:30

목 차

어휘의미망에 대한 인식과 사전의 구성

최경봉(원광대학교)

1. 머리말

본고에서는 어휘의미망에 대한 언어사용자의 인식 양상을 살펴보고, 이를 근거로 국어사전의 구성 문제를 논의할 것이다.[1] 따라서 본고의 논의는 언어사용자가 인식하는 어휘의 관계망 정보를 국어사전에 어떻게 반영하여 기술할 것인지로 귀결될 것인데, 언어사용자가 인식하는 어휘의 관계망 정보를 구체화하는 논의는 어휘의미망을 구축하기 위한 실천적 연구와 관련지어 진행할 것이다.

어휘의미망 구축을 위한 실천적 연구는 20세기 후반부터 본격화하였지만, 근대 이후 편찬된 언어사전의 체재와 내용을 살펴보면, 어휘의미망에 대한 언어사용자의 인식을 반영하는 것이 사전 편찬에서 중요한 문제였음을 알 수 있다.[2] 이런 맥락에서 보면, 한국어 어휘의미망을 구축할 때 국어사전의 뜻풀이 및 관련어 정보를 활용한 것은 자연스러운 접근이었다.[3]

그런데 자연언어 처리에 활용하기 위한 어휘의미망의 구축은 기존 언어사전의 체제와 내용상 문제점을 드러냄[4]과 동시에 언어사전의 혁신 방안을 구체화하는 계기가 되었다. 이와 관련하여, 이동혁(2010: 12)에서는 “미래의 사전과 어휘의미망은 물리적인 대상으로서의 사전, 전자화된 사전, 자연언어 처리를 위한 시스템의 역할을 모두 할 수 있는 것이어야 한다.”고 밝히면서, ‘어휘의미 체계 기반 입체적 국어사전’(국립국어원, 2009)과 필모어(C. J. Fillmore)의 ‘프레임 망(FrameNet)’에 기대어 새로운 사전의 가능성을 모색한 바 있다.

이처럼 언어사전의 편찬과 어휘의미망의 구축이 상호 영향 관계에 있다면, 어휘의 관계망에 대한 언어사전의 기술 내용이 어휘의미망에 반영되고, 어휘의미망 구축의 성과가 언어사전의 어휘 관계망 정보를 정교화하는 선순환을 기대해 볼 수 있을 것이다. 그런데 어휘의미망 구축 연구로부터 영향을 받은 언어사전의 변화는, <우리말샘>에서 볼 수 있듯이, 관련어 정보를 확대하는 수준에 머물러 있다. 본고의 문제의식은 여기에서 비롯한다.

언어사전의 구조적 특성상, 어휘 관계망 정보를 정교화하는 것은 관련어를 체계화한다는 것만이 아니라, 언어사전의 미시구조 전반에 어휘 관계망 정보를 반영하는 것을 의미한다고 할 수 있다. 따라서 어휘 관계망 정보를 반영하는 차원에서 기존 사전의 미시구조를 재구성하는 방안을 모색해야 하는데, 이를 위해서는 본고의 논의 대상을 어휘 의미를 규정하고 구획하는 문제부터 시작하여 뜻풀이와 관련되는 미시구조 전반으로 확대할 필요가 있다.

---

1) 어휘의미망이 어휘 의미를 기반으로 하는 어휘의 관계망이란 점에서 음운 교체나 형태론적 작용이 개입하여 이루어지는 관계어나 능동사와 피동사 등과 같은 통사적 특성에 따른 관계어는 논의 대상에서 제외할 것이다.

2) 근대 이후의 사전뿐만 아니라, 근대 이전 제작된 사전에서도 어휘의미망에 대한 인식을 확인할 수 있다. 특히 자회류(字會類) 및 물명류(物名類)는 위계적 분류체계의 분류어휘집으로, 분류어휘집의 체제는 어휘의미망에 대한 언어사용자의 인식 양상을 보여준다고 할 수 있다.

3) 울산대에서 개발한 사용자 어휘지능망(U-WIN)은 『표준국어대사전』(이하 『표준』)에서 구분하는 어휘의미를 기본 구성단위로 삼았으며, 사전의 정의문을 비롯한 각종 어휘 정보를 통해 중심어 및 의미관계를 추출하여 어휘의미망을 구축하였다.

4) 어휘의미망의 구축과 관련하여 언어사전의 기술 문제를 논의한 것으로는 김진해(2007), 옥철영(2007), 차준경·임해창(2010) 등이 있다.

본고에서는 어휘 의미를 계열적 의미와 결합적 의미로 구분하여 규정하는 어휘의미론의 방법론에 따라, 어휘의미망을 계열적 관계망과 결합적 관계망으로 구분하고 이를 정교화하는 방안을 모색하고자 한다. 이와 더불어 은유적 의미에 따라 새롭게 형성되는 관계망에도 주목할 것이다. 은유적 의미는 근원영역과 목표영역의 관계망이 상호작용하며 형성된다는 점에서 상호작용 과정에서 형성되는 관계망 또한 의미 기술에서 중요하게 다룰 필요가 있다. 본고에서는 다음과 같은 순서로 논의를 진행할 것이다.

2장에서는 계열적 관계망이 탈문맥적, 즉 본질적 의미에 대한 언어사용자의 통찰을 보여준다는 점에 주목하면서, 이를 사전에 반영하는 방안에 대해 논의할 것이다. 3장에서는 결합적 관계망이 문맥적 의미, 즉 어휘 사용의 패턴에 따라 드러나는 의미를 보여준다는 점에 주목하면서, 이를 사전에 반영하는 방안에 대해 논의할 것이다. 4장에서는 은유적 개념화에 따라 형성된 어휘의 관계망에 대해 논의할 것이다. 문맥 작용에 따라 전경화되는 의미 속성 간의 관계망, 은유적 개념화에 따라 창조되는 비유적 의미 간의 관계망 등을 사전의 기술 내용과 비교하여 살펴볼 것이다.

2. 계열적 관계망으로서의 어휘의미망과 사전

2.1. 분류학적 관계망의 한계와 관계 차원의 확장

어휘의 존재론적 분류체계는 분류학적 관계(taxonomic relation)에 기반하는 만큼, 어휘의 분류체계가 언어 사용 양상을 제대로 반영하지 못하는 문제가 있다. 이런 점 때문에 어휘의미망의 구축에서는 의미 속성에 따른 다양한 관계를 분류체계에 반영하여 다차원적 어휘의미망을 구축하는 문제를 고민해 왔다. 이를 잘 보여주는 것이 명사의 기능적 관계(functional relation)를 명사의 분류체계에 포함하는 것이다. 이는 하나의 명사가 복수의 의미부류에 포함될 수 있음을 뜻한다.

(1) 분류학적 관계와 기능적 관계
ㄱ. '보리'
분류학적: 구체물-생물-식물-풀
기능적: 구체물-자연음식물- 곡식

ㄴ. '사과'
분류학적: 구체물-생물-식물-열매
기능적: 구체물-자연음식물- 과일

위의 분류는 언어 외적인 개념을 분할하여 체계화한 결과이지만, 언어 외적 개념이 분할되는 층위가 달라지면서 분류 양상이 달라짐을 알 수 있다. '풀'과 '열매'는 '식물'의 일종으로 분류되는 반면, '곡식'과 '과일'은 사물의 활용 영역과 관련하여 '자연음식물'의 일종으로 분류되는 것이다. 이러한 분류 층위의 차이는 문맥 내에서의 선택 제약과 같은 해당 어휘의 의미 작용을 원리적으로 설명하는 데 활용될 수 있다. 그런데 국어사전에서는 (1)의 어휘들을 다음과 같이 기술하고 있다.

⑵ ‘보리’와 ‘사과’에 대한 『표준』의 기술

ㄱ. 보리

[식물] 볏과의 두해살이풀. 줄기는 높이가 1미터 정도이고 곧고 속이 비었으며, 마디가 길다. 잎은 어긋나고 긴 선 모양으로 겉이 매끄러우며 나란히맥이 있다. 꽃은 5월에 수상(穗狀) 화서로 달리는데 이삭에는 까끄라기가 있다. 알이 껍질에서 잘 떨어지는지에 따라 쌀보리와 겉보리로, 파종 시기에 따라 가을보리와 봄보리로 나눈다. 보리쌀은 보리밥・맥주・된장・빵 따위의 원료이고, 줄기는 여름 모자・공예품・제지용・퇴비 따위에 쓴다. 서남아시아, 이집트가 원산지로 전 세계 온대 지방에서 재배한다.

ㄴ. 사과

사과나무의 열매.

(2)의 예시를 보면, 계열적 관계망에 대한 인식 양상이 국어사전에 충분히 반영되지 못했음을 알 수 있다. (2ㄱ)의 경우는 밑줄 친 부분처럼 그 활용 영역을 간접적으로 기술하고 있고 (2ㄴ)의 경우는 그러한 설명이 생략되어 있다. 이러한 문제는 ‘벼’와 ‘쌀’에 대한 기술 내용과 비교할 때 분명해진다.

(3) ‘벼’와 ‘쌀’에 대한 『표준』의 기술

ㄱ. 벼

①식물 볏과의 한해살이풀. 줄기는 높이가 1~1.5미터이고 속이 비었으며, 마디가 있다. 잎은 어긋나고 긴 선 모...

②‘「1」’의 열매. 가을에 영과(穎果)로 익는 것을 이르며, 이것을 찧은 것을 ‘쌀’이라고 한다. 쌀은 주식으...

ㄴ. 쌀

①벼에서 껍질을 벗겨 낸 알맹이.

②멥쌀을 보리쌀 따위의 잡곡이나 찹쌀에 상대하여 이르는 말.

③볏과에 속한 곡식의 껍질을 벗긴 알을 통틀어 이르는 말. 쌀, 보리쌀, 좁쌀 따위가 있다.

위의 기술 내용을 보면, 분류학적 관계와 기능적 관계에 대한 구분이 분명한 것을 확인할 수 있다. 위의 기술 내용을 기반으로 편찬된 <우리말샘>에서는 (3ㄱ①)의 상위어로 ‘볏과’를 (3ㄱ②)의 상위어로 ‘곡물’을 제시하고 있으며, (3ㄴ①)의 상위어로 ‘곡물’을, (3ㄴ②)의 상위어로 ‘쌀’을, (3ㄴ③)의 상위어로 ‘곡물’을 제시하고 있다. 이처럼 분류의 차원에 따른 관계망이 분명하게 제시된 것은 분류의 차원에 따라 각각 ‘벼’와 ‘쌀’로 어휘화가 되었기 때문에 가능한 면이 있다.

물론 어휘화가 되지 않았더라도 이러한 분류의 관점을 다의적 의미항목을 통해 구현할 수 있지만, <우리말샘>에서는 ‘보리’의 상위어로 ‘볏과’만을, ‘사과’의 상위어로는 ‘과일’만을 제시하고 있다. 이는 언어사용자의 분류 의식이 국어사전 기술에 체계적으로 반영되지 못했음을 말해준다. 그렇다면 분류 의식을 반영하여 개념의 분할을 체계적으로 할 수 있는 일정한 틀을 세우고 이를 국어사전의 의미 기술에 활용할 필요가 있다. 이런 점에서 최경봉(2015: 110-112)의 논의는 참고할 만하다.

⑷ 인간명사의 의미 분류

[-내포] 철수, 영희, 순희

[+내포] - 형상: 사람, 인간, 여자, 남자, 아가씨, 총각, 아줌마, 노인, 젊은이

- 구성: 부자, 천재, 바보, 미인, 영세민

- 기능: 선생님, 의사, 가수, 연기자, 국회의원, 시장, 군수, 대통령, 장관, 흡연자, 행인, 보행자, 환자, 제자, 일꾼, 지게꾼

- 작인: 아버지, 어머니, 아들, 딸, 형, 누나, 동생, 부하, 상관

위에서 보인 의미 분류 방식은 [내포] 자질을 기준으로 고유명사와 일반명사를 구분한 후, 생성어휘론에서 속성 구조를 기술할 때 사용하였던 네 가지 작용역(형상, 구성, 기능, 작인)을 틀로 삼아 '인간'을 나타내는 일반명사를 분류한 것이다. 이때 각 작용역에 할당된 명사들은 해당 작용역에 '의미적 중점'을 두는 명사들이다. 예를 들어, '선생님'을 '기능'에 포함하여 분류한 것은 '선생님'이라는 어휘가 '학생을 가르치는 사람'이라는 기능적 의미를 나타내기 위한 어휘라는 점을 고려한 것이다.

이러한 분류체계는 개념이 분할되는 현상을 설명하는 데 적절하게 활용할 수 있다. 즉, '선생님'은 의미적 중점은 '기능'에 있지만, '선생님'을 '나이가 어지간히 든 사람을 대접하여 이르는 말(『표준』)'로 사용하는 것은 '선생님'의 의미 속성에서 '형상'의 작용역이 활성화된 결과로 설명할 수 있기 때문이다. 작용역의 활성화에 따른 개념의 분할 원리를 (1)에 제시된 어휘들에 적용할 경우, '형상'의 작용역에 의미적 중점이 있는 '사과'는 '열매'로 분류되지만, '구성'의 작용역이 활성화될 경우엔 '나무', '기능'의 작용역이 활성화될 경우엔 '과일'로 분류된다고 설명할 수 있을 것이다. 이러한 점을 앞에 제시한 사전의 의미 기술에 반영할 경우 의미 분할이 이루어지게 될 것이다.

그렇다면 앞서 다차원적 분류 의식을 반영하기 위해 추가한 '기능적 관계'는 '형상, 구성, 기능, 작인'이라는 네 가지 작용역에 따라 세분화해 볼 수 있다. 이러한 시도는 언어사용자의 분류 의식을 반영하여 어휘의 관계망을 정치하게 구성하는 계기가 될 수 있을 뿐만 아니라, 사전에서의 의미 기술을 체계화하는 계기가 될 수 있다. 그러나 이러한 틀이 모든 명사의 의미를 기술하는 데 적용될 수 있는 것은 아니다. 언어사용자의 분류의식은 역사적으로 형성된 측면이 있기 때문이다.

⑸ '피[稷]'에 대한 『고려대 한국어대사전』(이하 『고려대』)의 풀이
[식물] 볏과에 속한 한해살이풀. 높이는 1미터 정도이며, 잎은 벼와 비슷하여 좁고 길다. 여름에 담녹색 또는 자갈색의 이삭으로 된 꽃이 피고, 가시랭이가 있는 열매는 먹거나 사료로 쓴다. 환경 적응성이 커서 척박한 땅에서도 잘 견디므로 옛날에는 구황 작물로서 많이 재배하여 왔지만 최근에는 그 재배를 거의 볼 수 없게 되었다.

'피'는 분류학적으로는 '풀', 기능적으로는 '곡물'로 분류할 수 있지만, 밑줄 친 부분은 '피'가 실질적으로 '곡물'의 역할을 하지 못하게 된 역사적 맥락을 보여준다. 이러한 경우, '보리, 벼, 사과'에 대한 분류 의식과 '피'에 대한 분류의식의 차이를 보여주는 것이 국어사전의 기술에서 중요해질 것이다.

이처럼 분류학적 차원을 넘어서는 분류체계는 어휘의 본질적 의미에 대한 언어사용자의 인식을 설명하는 데 역할을 할 수 있을 뿐만 아니라, 역사적으로 형성되는 언어사용자의 분류의식을 설명하는 데에도 일정한 역할을 할 수 있다. 그리고 네 가지 작용역의 역할은 부분전체의 관계망에 대한 인식을 설명하는 데에도 적용된다.

⑹ 부분어에 대한 『표준』의 풀이
ㄱ. 다리 사람이나 동물의 몸통 아래 붙어 있는 신체의 부분. 서고 걷고 뛰는 일 따위를 맡아 한다.
ㄴ. 배 사람이나 동물의 몸에서 위장, 창자, 콩팥 따위의 내장이 들어 있는 곳으로 가슴과 엉덩이 사이의 부위.
ㄷ. 손잡이 손으로 어떤 것을 열거나 들거나 붙잡을 수 있도록 덧붙여 놓은 부분.
ㄹ. 가지 나무나 풀의 원줄기에서 뻗어 나온 줄기.

ㅁ. 강변 강의 가장자리에 잇닿아 있는 땅. 또는 그 부근.

부분체를 가리키는 명사는 대체로 공간부분(강변)과 개체부분(다리, 손잡이, 가지)으로 나뉘는데, 이들은 전체를 가리키는 명사와의 관계 속에서 '형상'과 '기능'에 따라 개념의 분할이 이루어질 수 있는데, 위의 예에서 제시한 사전의 뜻풀이에서도 이러한 분할 양상을 확인할 수 있다. 즉, (6ㄱ)은 '형상'과 '기능', (6ㄴ)은 '형상'과 '구성', (6ㄷ)은 '기능', (6ㄹ)은 '형상', (6ㅁ)은 '형상'에 따른 개념의 분할을 보여준다. 그렇다면 '부분어'를 '기능적 부분어'(다리, 손잡이), '형상적 부분어'(다리, 가지, 배, 강변), '구성적 부분어(배)'로 분류해 볼 수 있다. 이러한 분류 의식을 사전의 의미 기술에 반영한다면 (6ㄱ, ㄴ)은 다음과 같이 재기술할 수 있을 것이다.

(6ㄱ') 다리 ①사람이나 동물의 몸통 아래 붙어 있는 신체의 부분.
②서고 걷고 뛰는 일 따위를 맡아 하는 신체의 부분.
(6ㄴ') 배 ①가슴과 엉덩이 사이의 부위
②사람이나 동물의 몸에서 위장, 창자, 콩팥 따위의 내장이 들어 있는 곳.

위에서 재기술한 것 중 (6ㄴ')은 『고려대』의 의미 분할 방식과 같다. 이러한 기술 방식은 복합적 의미를 지닌 표제어에 대한 의미 기술 방안으로 채택할 필요가 있다.

(7) 복합적 의미를 지닌 표제어에 대한 『표준』 의미 기술
ㄱ. 간식 끼니와 끼니 사이에 음식을 먹음. 또는 그 음식.
ㄴ. 식사 끼니로 음식을 먹음. 또는 그 음식.
ㄷ. 경비 ①도난, 재난, 침략 따위를 염려하여 사고가 나지 않도록 미리 살피고 지키는 일.
②경비의 임무를 맡은 사람.

위에서 '간식', '식사', '경비' 등은 '구체물'과 '사건'이란 두 가지 의미가 복합되어 있는 단어임을 알 수 있다. 따라서 이들의 어휘적 관계망을 나타내기 위해서는 복합적 의미를 분할하여 기술할 필요가 있다. 어휘의미망 기반의 언어사전을 지향한다면, (7ㄱ, ㄴ)은 (7ㄷ)의 기술 방식을 취해 재기술할 필요가 있는 것이다. 이러한 기술 방식을 취했을 때, '간식'과 '간식하다', '경비'와 '경비하다'의 관계도 분명하게 기술할 수 있는데, 이 문제에 대해서는 차준경·임해창(2010)에서 논의한 바 있다.

지금까지의 논의에서 볼 수 있듯이, 계열적 관계망은 주로 명사의 분류 체계를 중심으로 논의가 되어 왔다. 이는 계열적 관계망의 특성이 사물의 본질과 작용에 대한 체계적 인식을 보여주는 명사의 분류체계와 가장 잘 부합하기 때문이다. 이는 명사 어휘와 타 품사 어휘의 차이라고 할 수 있는데, 황순희(2010)에서 언급했듯이, 동사는 명사처럼 의미자질에 근거한 동사 고유의 분류가 가능함과 동시에 개별 동사가 갖는 논항 유형과의 관계에 의한 비본질적 분류도 가능하다는 특징이 있다. 그렇다면 동사와 관련한 모어화자의 분류 의식을 파악하기 위해서는 계열적 관계만이 아니라 결합적 관계에 대한 분석이 필요하다.

다만, 계열적 관계망을 구축하기 위해서는 동사에 대한 본질적 분류 의식을 확인하는 것이 필요한데, 최경봉·도원영(2005)에서 제안한 동사의 상위온톨로지는 동사의 존재론적 분류체계를 제안했다는 점에서 참고할 만하다.

(8) 동사의 분류 체계
상황 유형(상적 특성) : 기동, 과정, 결과, 완성
상황 성분(사건 유형)

사건	변화	양의변화
		질의변화
		소유변화
		존재변화
		장소변화
	작용	자연작용
		상호작용
		정신작용
		지각작용

위에 제시한 체계는 이글스 리포트와 유로워드넷에서 제시된 동사 관련 의미자질들을 '상황 유형'과 '상황 성분'으로 나누어 체계화한 것이다. 이러한 체계화는 (4)에서처럼 명사의 기본 분류체계를 분류학적 관계망에 기대어 설정한 후 여기에 작용역에 따른 관계를 추가하여 분류체계를 정교화한 것과 유사하다. 그렇다면 (8)의 분류체계는 동사에 대한 사전의 의미 기술에 어떻게 반영될 수 있는가?

'만들다'를 예로 들어보면, '만들다'의 '상황 유형(상적 특성)'은 '완성'에 해당하고, '상황 성분(사건 유형)'은 '존재변화'를 나타낸다고 설명할 수 있다. 이러한 존재론적 속성은 사전의 뜻풀이에 다음과 같이 반영되어 있다.

(9) '만들다'에 대한 『고려대』의 기술
①(사람이 어떤 물건을) 재료나 소재 따위에 노력이나 기술을 들여 이루어 내다.
②(어떤 사람이 다른 사람이나 사물을 어떤 지위나 상태로) 되게 하다.
③(어떤 것이 다른 것을 어떠하게) 되게 하다.
(…)
⑨(사람이 기회나 시간 따위를) 일부러 내다.
⑩(사람이 문제가 될 말이나 일 따위를) 꾸며내거나 일으키다.
⑪(사람이 흠집이나 상처 따위를) 생기게 하다.

(9)의 의미항목 ①-③을 보면 '이루어 내다'와 '되게 하다'라는 풀이말은 '만들다'의 '상적 특성'이 '완성'에 해당되고, 그 '사건 유형'이 '존재변화'임을 보여준다고 할 수 있다. 반면 의미항목 ⑨-⑪은 '상적 특성'은 '기동'이나 '결과'에 해당한다는 점에서 차이를 보인다. 이는 '만들다'가 기본적으로 '완성/존재변화'라는 존재론적 속성을 지니지만, 문맥 상황에 따라 '기동'과 '결과'의 상적 특성으로 실현된다는 것을 말해준다.

이처럼 상적 특성이 다양하게 실현되는 것은 '사건'이 기본적으로 '기동, 과정, 결과'의 구성을 지니면서 문맥 상황에 따라 특정한 상적 특성이 부각되는 것으로 설명할 수 있다. 그러나 '기동, 과정, 결과'가 모두 존재론적 속성의 일부라는 점에서, 상적 특성의 실현 양상에 따라 의미를 분할하고, 분할한 의미의 계열적 관계망을 구축할 수 있을 것이다.

이를 보면, 동사의 존재론적 속성에 기반한 분류체계는 계열적 관계망을 정치하게 구성하는 토대가 되고, 사전의 뜻풀이를 정교화하는 토대가 됨을 알 수 있다. 특히 계열적 관계망을 정치하게 구축하는 것은 관계망 정보로 정의적 정보를 대체하는 기계가독형 전자사전을 구축

하는 과제와 관련되는 만큼, 존재론적 속성에 따라 의미를 분할하는 것은 계열적 관계망 구축의 토대가 된다.

2.2. 계열적 관계망의 이원성

앞 절에서는 다차원적인 분류 의식을 반영하여 개념의 분할을 체계적으로 할 수 있는 일정한 틀이 필요함을 강조하였다. 위계적인 단일 분류체계로는 하나의 사물에 대한 개념화 양상을 반영하기 어렵기 때문이다. 그런데 이처럼 개념 분할을 위한 틀을 제시하더라도 이들을 위계적 체계 내에서 연결짓기는 쉽지 않다.

(10) '나무'에 대한 『표준』의 기술
①줄기나 가지가 목질로 된 여러해살이 식물.
②집을 짓거나 가구, 그릇 따위를 만들 때 재료로 사용하는 재목.
③땔감이 되는 나무.

위에서 '나무'는 분류학적으로 '식물'의 하위어이지만, '인공물'의 '자재'에 포함되기도 한다. '자재'는 '나무'의 속성을 나타내는 작용역 중 '기능'이 활성화된 결과로 볼 수 있는데, 이러한 '기능'과 관련한 분류 의식은, (10)에서 볼 수 있듯이, '나무'를 '연료'의 하위어에 포함하는 데까지 이어질 수 있다. 이러한 분류 의식은 언어 간 어휘장의 비교를 통해 체계적인 유형화가 가능할 수 있다.

(11)<나무>의 어휘장

한국어	나무	나무	나무(땔나무)
영어	tree	wood	wood(firewood)
스페인어	árbol	madera	leña

위의 어휘장을 비교해 보면, <나무>의 개념장이 '식물/목재/땔감'으로 분할되며 한국어의 어휘장에서는 하나의 어휘 '나무'가 이 개념장을 포괄하는 것임을 알 수 있다. 어휘장에서 개념의 분할 양상을 고려하면, (10)에서의 확인할 수 있는 분류의식은 분류체계에 반영할 수 있을 것이다.

그러나 '나무'의 활용 영역을 '자재'와 '연료'로 한정할 수 없다는 점에서, 분류의 체계성을 어떻게 확보하느냐의 문제는 여전히 남는다. 특히 분야에 따라 어휘가 나타내는 사물의 활용 영역이 달라지는 점을 감안하면, 이러한 활용 영역의 확장 양상을 분류체계에 반영하기는 어려울 것이다.

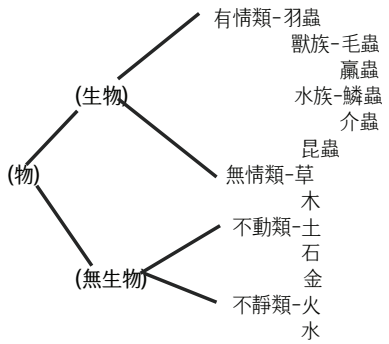
이러한 문제의식은 그간 온톨로지 구성 논의에서 주목한 부분이다. 즉, 유로워드넷의 경우 본질적이고 언어보편적인 개념체계를 표시하는 '상위 온톨로지'를 구성하고, 실제 생활에서 사물의 활용 영역과 관련되는 '영역 온톨로지'를 별도로 구성하였다. 또한 통사적 기준을 어휘의 분류 기준으로 삼는 대상부류 이론에서도 이원적 분류의 필요성을 확인할 수 있다. 대상부류 이론의 '신의미자질'은 의미자질로서의 역할을 하지만, 호응하는 '일반술어'를 근거로 정의된다는 점에서 특별하다. 그러나 신의미자질의 하위부류라고 할 수 있는 대상부류와 차원을 달리한다는 점에서는 상위온톨로지와 영역온톨로지를 구분하는 것[5]과 문제의식은 같다고 할 수

5) 최경봉·도원영(2005)에서는 '상위 온톨로지'와 '영역 온톨로지'의 이원적 분류체계를 제안하면서, 상위

있는 것이다.

의미자질의 체계로 구성된 상위온톨로지에 대한 인식은 전통적 분류어휘집에서도 찾아볼 수 있는데, 최경봉(2005)에서는 [물명고]의 어휘망 구성에서 확인할 수 있는 상하위 온톨로지에 대한 구분의식과 그 의미에 대해 설명한 바 있다.

(12)「물명고」의 분류체계(최경봉, 2005)



※ ( )안에 들어간 말은 물명고의 분류체계를 고려할 때 상위 분류 자질로 가정할 수 있는 것을 보이기 위해 해당 논문에 첨가한 것임.

위에 제시한 분류체계는 존재론적 인식 체계를 보여주는 것이기 때문에 그 관계가 위계적이라 할 수 있다. 그러나 이 분류체계의 자질들은 표제어에 해당하는 어휘들과 위계적으로 연결되지는 않는다. 한 예로 위 분류체계의 '우충(羽蟲)' 항에 배치된 어휘들을 보면 '유정류'라는 분류항과 이 분류항에 포함된 어휘 간 위계성이 분명하지 않음을 확인할 수 있다.

(13) '우충(羽蟲)'의 부류에 포함된 어휘

ㄱ. '봉황(鳳凰)', '치(雉, 꿩)', '백한(白鷳)', '이계(鵹鷄)', '반작(鴗雀)', '계(鷄, 닭)' 등.

ㄴ. '모(毛)', '미(尾, 꼬리)', '미파(尾把, 미우)' 등.

ㄷ. '난(卵, 알)'

ㄹ. '과(戈, 주살)', '위(罻, 새그물)', '부(罦, 덥치그물)', '견(絹, 올모)', '활구자(活扣子, 올모고)', '협자취(挾子觜, 창ᄋᆞ고동)', '협자구(挾子口, 창ᄋᆞ고패)' 등.

ㅁ. '알단(嗄蛋, 알겻다)', '하단(下蛋, 알낫타)', '비(飛, 날다)', '명제(鳴啼, 울다)', '혁(殛, 골다)' 등.

위에서 (13ㄱ)은 '우충'의 하위어, (13ㄴ)은 부분어, (13ㄷ)은 '우충'으로부터 얻는 산출물, (13ㄹ)은 새 사냥과 관련된 도구, (13ㅁ)은 새의 행위 및 산출물의 상태에 해당하는 어휘들이다. 이들은 '새'와 연관된 관련어들이란 점에서 공통점을 갖지만, 모두가 '새'의 상위 부류인 '유정류'에 포함되는 것은 아니다.

최경봉(2005)에서는 이러한 이원적 분류체계를 존재론적 분류체계를 정립함과 동시에 어휘망의 정보를 더욱 풍부하게 제시할 수 있는 요인으로 평가한 바 있다. (13)에 제시된 관련어

온톨로지의 구성 원칙으로, '구성 의미자질의 조합을 통해 사물의 특성을 설명할 수 있도록 할 것', '상위 온톨로지를 구성하는 의미자질 간의 체계적 관련성을 고려할 것', '상위 온톨로지와 영역 온톨로지는 위계적 관계가 아니지만, 이 둘의 연계성을 고려하여 상위 온톨로지를 구축할 것' 등을 제안한 바 있다.

들은 ‘새’와 관련된 의미영역의 어휘로서 제시되어 있지만, 다른 의미영역의 어휘로도 제시될 수 있는 것이다. 예를 들어 (13ㄹ)은 ‘새 사냥 도구’라는 점에서 ‘사냥도구’의 의미영역에 포함될 수도 있고, ‘사냥’의 의미영역에 포함될 수도 있는 것이다.

이러한 점을 볼 때, ‘상위 온톨로지’와 ‘영역 온톨로지’의 이원적 분류체계를 전제하면 분류학적 위계에 구애됨이 없이 특정 영역의 어휘적 관계망을 폭넓게 포착할 수 있는 장점이 있다. 특히 이원적 분류체계 내에서 정보의 상호작용을 전제하면 특정 어휘의 문맥 내 의미작용을 원리적으로 설명할 수 있다.

(14) ‘나무’의 언어 사용 양상
ㄱ. 할머니가 황칠나무를 끓여 먹으라고 보내주셨어요.
ㄴ. 나무를 태워 난방을 하는 구들 황토방
ㄷ. 불장난을 하다가 산의 나무를 태웠다.

위의 예에서 (14ㄱ)의 ‘나무’는 ‘자연음식물’보다는 ‘약재’에, (14ㄴ)의 ‘나무’는 ‘연료’에, (14ㄷ)의 ‘나무’는 ‘연료’보다는 ‘식물’에 해당할 것이다. 이처럼 맥락에 따라 ‘나무’의 의미가 달라지는 것은 ‘분류학적 관계망’과 별도로 ‘활용 영역 내의 관계망’을 설정하여 설명하는 것이 자연스러울 것이다. (14ㄱ)의 ‘황칠나무’는 ‘나무’의 하위어로 상위 분류체계의 ‘식물’과 연결되지만, ‘약’의 관계망에 포함되어 있기 때문에 ‘나무’가 ‘끓여 먹다’와 호응하는 의미작용을 설명할 수 있게 되는 것이다.

그런데 대상부류이론을 토대로 한 세종 의미부류 체계에서는 신의미자질과 대상부류의 구분을 없애고 단일한 분류체계를 구성하였다. 이는 의미부류 체계를 단순화하는 장점이 있으나, 앞서 논의한 바와 같이 단일한 분류체계는 상황 맥락에 따라 활용 영역이 확장되는 양상을 설명하는 데 한계가 있을 수밖에 없다.

예를 들어, 단일한 분류체계에서 <교통수단>(혹은 <이동수단>)은 ‘구체인공물’의 하위 부류로 설정될 수 있지만, <교통수단>으로 활용되는 것이 ‘자동차, 택시, 버스, 비행기, 배’ 등으로 한정되는 것이 아니란 문제가 있다. ‘말, 낙타’ 등은 문화에 따라 <교통수단>으로 자리 잡고 있다. 대상부류이론에 따른다며 ‘자동차’와 ‘말, 낙타’는 ‘타다’라는 적정술어를 공유한다는 점에서 하나의 의미부류로 묶일 수 있는 것이다. 그렇다면 <교통수단>이라는 관계망은 ‘구체인공물’이라는 사물의 존재론적 분류체계와 상관없는 ‘활용 영역’의 관계망이 되는 것이다.

(15) ‘타다’에 대한 국어사전의 기술
ㄱ. 타다 탈것이나 짐승의 등 따위에 몸을 얹다. 『표준』
cf) 탈것 자전거, 자동차 따위의 사람이 타고 다니는 물건을 통틀어 이르는 말.『표준』

ㄴ. 타다 (사람이 탈것을) 올라 몸을 싣다. 『고려대』
cf) 탈것 사람이 타고 다니는 것들을 통틀어 이르는 말. 가마, 말, 마차, 자전거, 자동차, 기차, 비행기, 배 따위가 있다.『고려대』

‘타다’에 대한 의미 기술을 보면, 『표준』과 『고려대』는 ‘탈것’의 범위 설정에서 차이를 보이고 있음을 알 수 있다. 『표준』은 ‘짐승’을 ‘탈것’에서 제외시키고 있지만, 『고려대』에서는 ‘짐승’을 ‘탈것’에 포함하여 기술하고 있는 것이다. 이는 위계적 분류체계에서 ‘탈것’을 ‘구체인공물’로 분류하느냐 그렇지 않느냐에 따른 차이로 볼 수 있다. 『고려대』의 기술 내용을 적용하게 되면 ‘탈것’으로 활용될 수 있는 짐승의 부류(말, 낙타 등)를 설정할 수 있고, ‘말, 낙타’를

‘동물’과 ‘탈것’으로 구분해 그 의미를 기술할 수 있다. 이 경우 ‘말을 타다’와 ‘개를 타다’에서 ‘타다’의 의미를 다음과 같이 구분할 수도 있을 것이다.

(15’) ‘타다’의 재기술
①(사람이 탈것을) 올라 몸을 싣다. ¶말을 타다
②사람이나 짐승의 등 따위에 몸을 얹다. ¶개의 등에 타다

지금까지의 논의를 통해, 계열적 관계망은 ‘존재론적인 인식 체계로서의 관계망(상위온톨로지)’과 ‘특정 영역에 대한 지식 체계로서의 관계망(영역온톨로지)’을 아우르는 것임을 알 수 있었다. 그리고 두 차원의 관계망을 설정함으로써, 어휘의 관계망을 정교하게 포착하여 사전에 기술할 수 있음을 알 수 있었다. 그런데 앞선 설명에서 살펴봤듯이, 언어사용자가 어휘의 관계망을 인식하는 토대는 탈문맥적인 의미와 문맥적 의미를 포괄하는 것이다. 따라서 계열적 관계망을 중심으로 한 어휘의미망에 대한 탐구는 논항 관계 및 선택 관계와 같은 결합적 관계망으로 확장될 수밖에 없다.

3. 결합적 관계망으로서의 어휘의미망과 사전

3.1. 어휘 간 상호 선택의 관계망과 사전의 기술

앞 장에서는 명사의 ‘분류학적 관계’와 ‘네 가지 작용역에 따른 속성 관계’에 따라 명사의 관계망을 파악해 보면서, 계열적 관계망의 다차원적인 특징에 대해 살펴본 바 있다. 그런데 이러한 논의는 적정술어와 관련하여 명사의 의미 부류를 파악하는 논의, 즉 대상부류 이론에서 대상 부류를 파악하는 논의와 관련된다. 특히 대상부류 이론은 어휘 간의 결합 관계 양상을 포착하여 의미 속성에 대한 인간의 인식 틀을 설명한다는 점에서, 결합적 관계망을 구축하는 방법론으로 주목을 받았다.

(16) ‘사과’와 술어의 결합 관계
ㄱ. 사과를 먹다
ㄴ. 사과를 사다/팔다
ㄷ. 사과를 따다
ㄹ. 사과를 심다

대상부류 이론에 따르면, ‘사과’는 그것이 함께 결합하는 술어와의 관련 하에 의미부류가 결정되는데, (16ㄱ)의 ‘사과’는 [과일], [음식물]의 부류에, (16ㄴ)의 ‘사과’는 [과일], [상품]의 부류에, (16ㄷ)의 ‘사과’는 [열매], [식물]의 부류에, (16ㄹ)의 ‘사과’는 [나무], [식물]의 부류에 포함될 수 있다. 대상과 서술어의 결합 관계는 어휘의미망의 중요한 요소가 된다고 할 수 있다. 이러한 의미부류는 앞서 살펴봤듯이 ‘사과’의 다의성을 파악할 수 있는 근거가 된다.

또한 같은 개념의 어휘들이더라도 그것이 어떤 논항과 선택 관계를 이루는지에 따라 그 부류를 달리 설정할 수도 있다. 한 예로 ‘상거래’의 의미 영역에서 ‘사다’와 관련한 동사들은 ‘구매하다, 구입하다, 매입하다, 수매하다…’ 등인데, 이 동사들은 그것이 어떤 논항과 함께 선택 관계를 이루는지에 따라 구분될 수 있다.

(17) '사다'의 유의어와 유의어에 따른 논항 선택
ㄱ. 구매하다, 구입하다; 제한 없음
ㄴ. 매입하다; 주식, 부동산
ㄷ. 수매하다; 곡식

이러한 결합 관계 정보는 개별 언어의 언어적 특징을 나타내는 정보로서 주목을 받았는데, 어휘의미론에서의 새로운 시도는 대부분 결합적 의미 관계 정보를 어떻게 포착하여 기술할 것인지에 집중되었다. 어휘의미망의 구축과 관련한 실천적 연구의 흐름도 이와 다르지 않은데, 윤애선(2012)에서 소개한 바와 같이, 한국어 어휘의미망인 'KorLex 2.0'은 한국어 의존적인 정보로 '수분류사와 명사 간 공기관계'와 '용언의 논항 정보와 선택 제약' 정보를 구축하였다. 이는 어휘 의미 연구에서 결합적 관계가 주목을 받는 것과 같은 맥락에서 이해할 수 있지만, 코퍼스 기반의 통계적 접근으로 포착한 결합적 관계망이 어휘의미망에 반영되지 않았다는 점에서, 어휘의미망의 구축이 결합적 관계에 대한 어휘의미론에서의 연구 성과와 연동되었다고 하기는 어렵다.

따라서 결합적 의미 관계를 포착하기 위한 통계적 접근의 성과를 언어사전에 반영하고 이를 기반으로 결합적 관계망 정보를 어휘의미망에 포함한 기계가독형 전자사전을 구축하는 절차를 고려할 필요가 있다. 어휘의미론에서 주목하는 결합적 의미 관계 중 기존의 언어사전에 제대로 반영되지 않은 것 중 하나가 연어 관계이다.

연어 정보의 체계적인 기술 방안은 기계가독형 전자사전의 구축을 위한 어휘함수 논의를 통해 구체화된 바 있다. 그렇다면 이러한 연어적 관계망은 언어사전에 어떻게 반영되고 있는가?

(18) 연어 정보와 관련한 국어사전의 기술
ㄱ. 짓다: 재료를 들여 밥, 옷, 집 따위를 만들다. ¶밥을 짓다. / 아침을 짓다. / 옷을 짓다. / 양복을 짓다. / 누에가 고치를 짓고 있다. / 그는 고향에 기와집을 지었다. 『표준』
ㄴ. 끓이다: 액체를 몹시 뜨겁게 해 소리를 내면서 거품이 솟아오르게 하다. '끓다'의 사동사. ¶물을 끓이다. / 국을 끓이다. / 차를 끓이다. / 저녁 반찬으로 찌개를 끓이다.『표준』
ㄷ. 거짓말: 사실이 아닌 것을 사실처럼 꾸며서 말함. 또는 그런 말. ¶그건 새빨간 거짓말이야. / 언니는 습관적으로 거짓말을 보태어 말한다. /요즘 정치인들을 보면 누가 거짓말을 잘하느냐로 실력을 겨루는 듯한 느낌을 받는다. 『고려대』
ㄹ. 사의: 감사하게 여기는 뜻. ¶사의를 나타내다. / 심심한 사의를 표하다.『표준』
ㅁ. 세금: 국가나 지방 단체가 필요한 경비를 충당하기 위해서 국민으로부터 거두어들이는 돈. ¶세금 포탈 / 세금 감면 / 세금을 내다 / 세금을 거두다 / 정부의 예산은 대부분 국민의 세금으로 충당된다. / 내수 경기가 호황을 보이면서 세금이 예상보다 훨씬 잘 걷히고 있다.『고려대』
ㅂ. 결정: 행동이나 태도를 분명하게 정함. 또는 그렇게 정해진 내용. ¶ 결정을 내리다 / 결정을 보다/ 결정을 짓다 / 결정이 나다 / 결정에 따르다.

(18ㄱ, ㄴ)은 연어의 결합적 구성에 대한 이해와 연어 구성을 이루는 어휘의 계열적 관계에 대한 이해가 동시에 이루어질 필요성을 말해준다. 먼저 '짓다'와 '끓이다'가 '조리어'의 영역에서 관계를 맺고 있음을 고려하면, '짓다'와 '끓이다'의 공통점을 의미 기술에 반영할 필요가 있다. 그리고 '짓다'와 '끓이다'의 연어적 결합에 따라 두 어휘의 의미를 구분하여 기술할 필요가 있다. 이를 반영하면 (18ㄱ, ㄴ)은 다음과 같이 재기술할 수 있을 것이다.

(18') ㄱ. 짓다 ① 밥이나 한 끼 음식을 만들다. ¶밥을 짓다. / 아침을 짓다. ② 재료를 들여 옷, 집 따위를 만들다. ¶ 옷을 짓다 / 양복을 짓다 / 누에가 고치를 짓고 있다. / 그는 고향에 기와집을 지었다.

ㄴ. 끓이다 ①액체를 몹시 뜨겁게 해 소리를 내면서 거품이 솟아오르게 하다. ‘끓다’의 사동사. ¶물을 끓이다. / 차를 끓이다. ②국이나 찌개 따위를 만들다. ¶국을 끓이다. / 저녁 반찬으로 찌개를 끓이다.

그런데 앞서 살펴본 (18ㄱ,ㄴ)을 포함하여 (18)의 기술 내용을 보면, 언어사전에서는 연어 관계를 용례로 보이는 방식을 취하고 있음을 알 수 있다. 이러한 처리 방식은 다음과 같은 문제가 있는데, 첫째는 용례로 선택되지 않은 연어 정보를 사전에서 확인하기는 어려울 수 있다는 것이고, 둘째는 결합하는 어휘 간의 제약 및 의존성을 나타내기가 어렵다는 것이다. 실제 (18ㄷ,ㄹ)의 경우, 용례로 제시된 ‘새빨간 거짓말’과 ‘심심한 사의’가 관습적으로 제약적인 결합 관계임을 나타내기 어렵다. (18ㅁ)의 경우, 다양한 용례가 제시되었음에도 ‘무거운 세금’, ‘가벼운 세금’ 등과 같은 결합 양상은 보이지 않는다. (18ㅂ)의 경우, ‘결정’이 ‘내리다’와 결합하는 것을 보여줄 수는 있지만 그 유의어인 ‘떨어뜨리다’나 반의어인 ‘올리다’와 결합하는 것이 불가능함을 나타내기는 어렵다.

그렇다면 언어사전의 경우 연어 정보의 관계망을 최대한 보여주면서 관계의 양상을 설명하는 시도가 필요할 것이다. 그간 이러한 시도가 적극적으로 이루어지지 않은 이유는 종이사전 편집 체계의 영향 때문으로 보이는데, 세종 전자사전에서는 다양한 어휘함수를 만들어 연어 관계의 양상을 기술한 바 있다.[6] 기계가독형사전에서 연어 관계 정보가 중요하게 다루어지는 상황에서, 웹사전이 종이사전의 한계를 극복하는 가능성에 주목하면, 사전의 기술 내용을 확장하기 위한 웹사전의 다양한 시도에 주목할 필요가 있다.

현재 네이버 영어사전 홈에서 영어 단어를 검색하면, 기존 영한사전의 내용에 ‘학습 정보’[7]가 추가적으로 제시되어 있다. 이중 연어 관계 정보는 *Oxford Collocations Dictionary for students of English*(2009)에서의 기술 내용을 제시하고 있다.

(19) ‘tax’의 연어 관계에 대한 사전의 기술

함께 사용되는 단어

[명사]로 tax이(가) 사용될 때

형용사

high, low | direct, indirect | flat | basic-rate, higher-rate(both BrE) | progressive, redistributive | regressive | stealth(BrE) | windfall | back(informal) | council(BrE), poll(esp. BrE) | federal, local, national, state | capital, capital gains, death(AmE), dividend(AmE), estate(AmE), income, inheritance, land, payroll, profits, property, social-security, wealth | consumption, luxury, purchase(BrE), sales, value added(= VAT)(BrE) | personal | company(BrE), corporate, corporation | car, road(BrE), vehicle | carbon, energy, fuel, gas(AmE), gasoline(AmE), petrol(BrE) | excise, import | environmental, green

> a windfall tax on the profits of the last few years
>
> The tax office demanded ' in back taxes.
>
> It's time to renew your car tax.

동사 + tax

pay | owe | charge, impose, introduce, levy, put | collect | deduct | calculate | increase, put up, raise | cut, keep down, lower, reduce | abolish, repeal | eliminate | overpay | claim back, reclaim(BrE) | offset sth against, set sth off against, write sth off against(all BrE) | avoid, escape | evade

> The government may put an indirect tax on books.

6) 김진해(2013)에 따르면, 21세기 세종계획 전자사전에서 연어사전은 모두 8,946개의 연어(다의어 항목까지 포함하면 9,521개)를 상세 기술하고 있는데, 이중에서 어휘함수가 기술되지 않은 것은 3,927개로 전체 연어 목록 중에서 41% 정도이다. 김진해(2013)에서는 이러한 상황을 의미 작용의 다양성을 기술하지 못하는 포괄하지 못하는 어휘함수의 한계로 보고 있다.

7) 학습 정보는 ‘유의어’, ‘문형/함께 쓰이는 말’, ‘함께 사용되는 단어’ 등으로 이루어져 있다.

이와 같은 체제는 복수의 사전을 연결해 놓은 것이지만, 한 표제어의 정보로 제시되면서 기존 사전의 내용을 보완한다는 점에서, 웹사전의 구조를 모색하는 데 참조할 수 있다. 위에 제시된 연어 관계 구성을 사전의 의미 기술에 포함하기 위해서는 어휘함수를 정교하게 설정하여 함수 정보를 기술하거나 결합 구성을 확장된 의미 단위로 기술하는 방안을 마련할 필요가 있을 것이다.

최근 연어 관계 연구는 통계적 접근법을 강화하는 경향을 띤다. 이러한 경향은 연어 관계의 양상을 다각도로 파악하는 계기가 되었는데, 이런 흐름 속에서 부각된 개념이 의미적 운율(semantic prosody)이다. 의미적 운율은 공기하는 언어를 선택하는 데 있어서의 선호도를 나타낸다고 할 수 있는데, 이 선호도는 코퍼스를 통해 확인할 수 있다. Sinclair(2004)에서는 확장된 의미단위의 개념을 '연어(collocation)', '연접(colligation)'[8], '의미적 선호(semantic preference)', '의미적 운율(semantic prosody)' 등의 네 개념과 관련지어 설명한 바 있는데, 이 개념들은 단어의 결합이 일정한 패턴을 이루면서 단어 결합 구성을 포괄하는 의미가 발생함을 포착하기 위한 개념이다. 즉 '의미적 선호'나 '연접'은 단어 결합의 패턴에 작용하고, '의미적 운율'은 그러한 패턴에서 포괄적 의미가 발생하는 데 작용한다.

(20) ㄱ. 그의 발명품은 일상생활에서 전혀 쓸모가 없었다.
ㄴ. 나는 너를 절대로 용서하지 않겠다.
ㄷ. 사람을 모함해도 유분수지 왜 나에 대해 그런 말을 하는 거야?
ㄹ. 우리 마을은 올해 극심한 물난리를 겪었다.
ㅁ. 나는 집을 나온 다음 편한 잠을 이루어 본 적이 없다.

(20ㄱ)에서 '쓸모'는 '있다'나 '없다' 혹은 '많다'나 '적다' 등과 제한적으로 어울린다는 점에서 그 결합이 긴밀하다고 볼 수 있는데, 애초 중립적이었던 '전혀'가 '없다'와 결합하여 전체 구성의 의미를 부정적으로 바꾼다. (20ㄴ)에서 '절대로'는 [반드시]라는 뜻의 중립적인 단어이지만 특정 문맥에서의 결합 패턴은 부정의 뜻을 지니게 된다.

(20ㄱ-ㄴ)이 특정 문법 범주가 관습적으로 공기하는 패턴을 보여준다면, (21ㄷ)은 의미적 선호를 보여주는데, '유분수'는 [분수가 있음]으로 중립성을 띠지만 실제 문맥에서 이것이 포함된 구성은 "<부정적 행동>(어)도 유분수지"의 패턴을 보이면서, 아무리 부정적인 행위를 하더라도 지켜야 할 선이 있다는 비판 의식을 드러낸다. 이러한 경향성은 언어사전에서 일정 정도 반영하고 있다.

(21) '전혀, 절대로, 유분수'에 대한 『고려대』의 기술
ㄱ. 전혀 부정어와 함께 쓰여, '절대로', '완전히'의 뜻을 나타내는 말.
ㄴ. 절대로 ① [부정어와 함께 쓰여] 어떤 일이 있더라도. ② [일부 단어와 함께 쓰여] 무슨 일이 있어도 반드시.
ㄷ. 유분수 [주로 '-어도 유분수이지'의 구성으로 쓰여] 마땅히 지켜야 할 분수가 있음.

(20ㄹ-ㅁ)은 앞의 예와 달리 '의미적 운율'을 확인할 수 있다. (20ㄹ)에서 '겪다'의 논항은 '물난리'인데, 다른 예에서도 '겪다'의 논항으로는 '가뭄, 전쟁, 고통...'과 같은 부류가 선호된다. '겪다'의 이러한 의미적 선호에 대해서는 남길임(2012)에서도 주목한 바 있다. 코퍼스에 나타난 이러한 선호 경향을 통해, '겪다'가 포함된 구성이 "<어려운 상황>을 겪다"와 같은 패턴을 보이고, 이 결합 구성의 의미적 운율이 [고생스럽다]라는 부정적 의미를 띤다고 할 수 있

8) 특정 어휘나 의미의 단위가 특정한 문법 범주, 즉 부정, 시제, 상 등과 관습적으로 공기하는 것.

다. 또한 (20ㅁ)의 '잠을 이루다'는 부정적 요소와 결합하는 경향성을 보이는데, 이는 남길임(2012)에서 입증한 바 있다.

의미적 운율이 어휘 결합 구성의 포괄적 의미를 나타내는 역할을 한다면, 부정 또는 긍정을 나타내는 데 관습적으로 쓰이는 어휘 및 문법형태소들을 포착할 수 있을 것이다. 그렇다면 '의미적 운율' 또한 결합적 특징으로 포착해 사전에 기술할 필요가 있다.

(22) '겪다'에 대한 사전의 기술
ㄱ.『고려대』
①기본의미 (사람이 일을) 당하여 치르다.
남편이 죽은 후에 그녀는 갖은 고초를 다 겪었다.
유의어: 경험하다 거치다1 치르다 먹다1 체험하다 맛보다
ㄴ.『표준』
①어렵거나 경험될 만한 일을 당하여 치르다.

두 사전의 기술 내용을 보면, 해당 표제어의 '긍정성'과 '부정성'을 나타내고 있음을 알 수 있다, 다만 통계적 근거가 분명하다면, (21ㄱ-ㄴ)에서처럼 '긍정성'과 '부정성'에 대한 표지를 보여줄 필요가 있다.

3.2. 논항 관계와 프레임

결합 관계 중 연어 관계가 언어사전에서 제한적으로 반영된 데 비해, 논항 관계는 언어사전에서 비중 있게 다루어졌다. 이는 논항 관계가 문형을 파악하는 것과 관련되는 정보이고, 문형 정보는 근대 사전에서 중요한 미시구조 정보로 취급되었기 때문이다.

(23) ㄱ. 그가 책을 샀다.
ㄴ. 그가 서점에서 책을 만 원에 샀다.

위의 두 예문에 대해 국어사전에서는 'NP1이 NP2를 사다'와 같은 하위범주화 틀을 '사다'의 문형 정보로 제시하면서, '서점에서', '만 원에' 등은 부가어로 취급한다. 'NP2'의 위치에서 '사다'와 결합하는 '책'은 사는 행위의 대상물이라는 점에서 국어사전에서 중요하게 취급되는데, 서술어가 나타내는 사건이나 사태의 대상이 되는 명사의 의미부류에 따라 서술어의 의미 항목을 구분하는 것은 사전의 전통적인 의미 기술 방식이다.

(24) '사다'에 대한 『고려대』의 기술
①(어떤 사람이 다른 사람에게 물건이나 권리를) 값을 치르고 자기 것으로 만들다. ¶할머니께서는 집에 오는 길에 제과점에 들러 팥빵을 사 오라고 하셨다.
②(어떤 사람이 다른 사람에게 자신에 대한 감정을) 자신의 말이나 행동으로 말미암아 가지도록 하다. ¶괜한 말을 해서 그에게 반감을 사고 말았다.
③(사람이 일을 할 사람을) 대가를 치르고 부리다. ¶사람을 사서 모내기를 하도록 합시다.
④((‘높이’와 함께 쓰여)) (사람이 대상을, 또는 그 대상의 장점을) 그 가치를 인정하다. ¶나는 자네의 그 패기를 높이 사는 바이네.
⑤(사람이 돈을) 가지고 있는 물건을 내주어서 마련하다. ¶아내는 장에 가서 쌀을 팔아서 돈을 샀다.
⑥((주로 ‘사서’의 꼴로 쓰여)) (사람이 고생을) 일부러 하다. ¶이 일은 쉽게 할 수도 있는데 네가 고생을 사서 하는구나.

⑦(어떤 사람이 다른 사람에게 음식 따위를) 함께 먹기 위하여 값을 치르다. ¶오늘은 내가 너희들한테 술 한잔 살게.

그런데 의미항목 ②와 ⑦은 'NP1이 NP2를 사다'와 같은 하위범주화 틀만으로는 설명하기 어려운 문형을 보여준다. ②와 ⑦에서 '에게(한테)'에 붙어 실현되는 논항의 의미와 역할은 문장이 나타내는 사건에 대한 이해에 기반하여 결정되기 때문이다. 따라서 의미항목 ②와 ⑦에 제시되는 문형에서는 '_에게' 논항을 필수 논항으로 표시하게 된다.

이런 점을 보면, 어휘의미와 구문의 상관성은 하위범주화의 틀보다 경험 지식을 반영한 사태의 프레임을 통해 분명해지는 측면이 있다. 이는 "문장의미란 전체로서의 사건과 발화상황이 그것을 구성하는 요소로서의 성분과 상관관계를 가지면서 생성.해석되는 의미인 것이다." (임채훈, 2012: 35)는 문제의식으로 이어진다. 사태의 프레임 안에서 프레임 요소를 환기하고 이를 통해 구문을 이해하는 것은, 서술어의 의미구조가 구문 구조로 투사된다는 관점의 한계를 극복하는 것이기도 하다.

따라서 프레임으로 문장을 파악하는 관점이 필요한데, '상거래'의 프레임에서는 '상품'과 같은 필수논항뿐만 아니라 '판매처', '판매자', '가격' 등과 같은 부가어도 프레임의 구성요소로서 중요하게 작용한다. 또한 프레임 요소들은 각각 경험 지식으로서의 프레임을 지닌다. 즉, '서점'에 대한 우리의 경험 지식은 "서점에는 팔 [책]을 전시하고 있고, [구매자]는 [서점]에서 [판매자]에게서 책을 산다."처럼 프레임화되었기 때문에 (23ㄴ)의 '서점에서'가 '장소'이면서 '판매자'로도 이해될 수 있는 것이다. 이러한 프레임은 ②처럼 정서적 상호작용을 상거래로 개념화하는 은유표현에서도 그대로 유지되면서 의미 해석을 유도하게 된다.[9)]

최경봉(2019)에서는 프레임을 기반으로 문장에서의 결합 관계를 파악할 경우, 프레임 요소가 특정 논항으로 전경화되거나 후경화되면서 통사구조가 달라지는 양상을 원리적으로 설명할 수 있음을 논의한 바 있다. 이에 따르면 문장의 생성과 해석에서는 프레임 요소의 선택과 더불어 선택된 프레임 요소의 속성과 작용이 중요해진다. 실질적으로 의미항목의 구분은 프레임 요소의 속성과 작용에 근거한 것이라 할 수 있기 때문이다. 의미항목 ⑦은 프레임 요소의 속성과 작용의 양상을 잘 보여준다.

(25) ㄱ. 오늘은 내가 너희들한테 술 한잔 살게.
ㄴ. 그가 고급 레스토랑에서 친구들에게 밥을 샀다.

위의 예문에서 '너희들한테'와 '친구들에게'란 논항은 의미항목 ①에 나오는 '다른 사람에게'란 논항과 그 성격이 다르다. 이는 다음과 같은 문장의 비교를 통해 확인할 수 있다.

(26) ㄱ. 나는 친구에게 밥을 샀다.
ㄴ. 나는 친구에게 책을 샀다.

위의 두 문장은 '돈을 지불하여 무엇인가를 취할 수 있는 상황'을 만든다는 점에서 공통적이다. 그러나 (26ㄱ)에서 '친구'는 '나와 함께 구입물을 공유하는 사람'이고, (26ㄴ)에서 '친구'는 '판매자'이다. 이러한 해석의 차이는 '사다' 프레임에서 '밥'과 '책'이라는 프레임 요소의 속성과 작용의 차이에서 비롯한다고 할 수 있다. '사다'의 프레임에서 '밥'이라는 프레임 요소의

9) '반감'이란 감정은 '그'가 가지고 있는 것이고, '나'는 '그'가 가지고 있는 감정인 '반감'을 '나'에게로 이끌어오는 것이다.

속성과 작용을 고려하면 (25ㄴ)의 문장은 다음과 같이 재구성해 볼 수 있을 것이다.

(27) 그가 고급 레스토랑에서 밥을 사서 (친구들에게 대접했다).

경험 지식상 '사다' 프레임에는 '구매자'뿐만 아니라 구매에 따른 '수혜자'도 설정될 수 있는데, 그 '수혜자'는 구매자 자신일 수도 있고 주위의 다른 사람일 수도 있다. 따라서 '사다' 프레임이 (25ㄴ)의 문장에서 (27)의 의미를 유추하는 데 작용하게 되는 것이다. 다만, 일반적인 '사다' 구문에서는 '수혜자'가 후경화되지만, (25)처럼 '본인이 돈을 지불하여 취한 결과를 공유하는 수혜자'가 있을 경우는 '구매 결과를 공유하는 수혜자'가 전경화된다고 할 수 있다. 이때 '구매 결과를 공유하는 수혜자'가 설정되는 경우는 '음식물'이 상품으로 설정되는 경우에 한정된다는 점에서, '사다' 프레임과 그 요소인 '밥'의 프레임[10] 간 상호작용이 부각되는 것이다. 이를 보면 프레임을 기반으로 결합 관계의 양상을 이해하는 것은 프레임 요소와 관련된 배경지식을 포착하는 데에서 시작한다고 할 수 있다.

(28) '치료하다'에 대한 국어사전의 기술
ㄱ. 병이나 상처 따위를 잘 다스려 낫게 하다. 『표준』
ㄴ. (사람이 다른 사람이나 상처 따위를) 잘 다스려 낫게 하다.『고려대』

'치료' 행위의 프레임을 구성하는 프레임 요소로는 '치료자, 환자, 환부, 병, 상처, 장소, 기간, 도구' 등이 설정될 수 있다. 이중 밑줄 친 요소는 치료 프레임의 핵심적인 요소가 되는데, 사전의 기술은 이 핵심적 프레임 요소를 중심으로 이루어진다.

사전에 제시된 '치료하다'의 예문은 "다리를 치료하다 / 상처를 치료하다 / 홍역을 치료하다 / 부상병을 치료하다 / 아이를 치료하다" 등인데, 여기에서 '홍역, 상처'는 '병'과 '상처'의 프레임 요소로, '부상병'과 '아이'는 '환자'의 프레임 요소로, '다리'는 '환부'의 프레임 요소로 목적어의 자리에 실현된다. 따라서 이러한 인식 작용을 사전에 기술하기 위해서는 목적어로 실현되는 프레임 요소의 관계망(환자, 환부, 병, 상처)을 제시하는 게 필요하다. 『표준』의 처리 방식은 '아이'가 목적어가 되는 것을 직접적으로 설명하지 않은 반면, 『고려대』의 처리 방식은 '아이'가 목적어가 되는 것을 직접적으로 설명한다는 차이가 있다. 그런데 모든 사전이 '환부'의 프레임 요소를 제시하지 않은 것은 '다리'를 '환자'로 이해하는 환유적 인식 양상을 전제한 것으로 볼 수 있다.

지금까지의 논의를 보면, 프레임은 언어사용자의 경험 지식을 반영한 것으로 볼 수 있는데, 특정 프레임은 하나의 의미 영역으로 해당 의미 영역과 관련한 어휘의 관계망을 호출한다고 할 수 있다. 결국 언어사용자의 경험 지식을 체계화하는 시도는 어휘의 다양한 관계망을 구성하는 논의를 촉발했다고 볼 수 있는데, 이중 통계적 접근법과 관련하여 깊이 있게 논의되는 것이 언어 사용에 기초한 화제(topic) 의미 관계이다. 화제 의미 관계처럼 코퍼스를 기반으로 한 관계망은 우리의 일반적 경험 지식을 가장 근접하게 보여줄 수 있다는 장점이 있다.

강범모(2017)에서는 언어 사용에 기초한 화제(topic) 의미 관계를 논의한 바 있는데, 화제 의미 관계는 하나의 화제 아래에서 함께 자주 사용되는 어휘들 사이의 관계를 말한다. 가령 '병원'은 '환자, 의사, 의료, 치료, 수술' 등과 자주 공기하여 나타나는데, 이는 '병원'이라는

10) 우리의 경험 지식에서 '밥'은 함께 먹는 행위를 수반하는 경우가 많다. 이는 곧 '밥'의 프레임으로 기억될 수 있다.

개념이 '환자, 의사, 의료, 치료' 등의 개념과 연관된다는 것을 뜻한다. 이처럼 공기어와 연어를 통해 한 어휘에 대한 관련어들을 추출하고 이들의 관계망을 구성하면, 특정 화제에 대한 관련어 네트워크를 구성할 수 있을 뿐만 아니라, 화제 간 관련어 네트워크를 구성할 수 있게 된다.

이러한 어휘망은 실제 언어 사용에서 공기하는 어휘 간의 출현 빈도에 기반한다는 점에서, 화제 의미 관계는 유의, 반의, 상하의, 부분·전체 관계 양상과도 다르며, 어휘 연상에 따른 의미 관계와 반드시 일치하는 것도 아니다. 따라서 화제 의미 관계는 사전에서 용례를 선택하고, 참고어 정보를 구체화하는 데 활용할 수 있을 것이다.

4. 은유적 개념화를 통한 관계망의 형성과 사전

4.1. 의미 속성의 관계망

다의어의 의미 항목들은 다른 어휘 항목과 의미 관계를 맺을 수 있다는 점에서 어휘의미망에서는 독립적인 어휘 항목으로 다루어진다. <우리말샘>에서의 표제어 처리 방식은 다의 항목의 이런 특성을 반영한 것이다.

(29) <우리말샘>의 표제어 기술 방식

나무 [나무]
- **나무** 「001」「명사」 줄기나 가지가 목질로 된 여러해살이 식물.
- **나무** 「002」「명사」 집을 짓거나 가구, 그릇 따위를 만들 때 재료로 사용하는 재목.
- **나무** 「003」「명사」 땔감이 되는 나무붙이.

나무 [나무]
- **나무** 「004」「명사」 소 장수들의 은어로, 팔백 냥을 이르던 말.

나무 [나무]
- **나무** 「005」「명사」『문학』 이양하가 지은 수필. 안분지족의 현인, 고독의 철인(哲人), 홀륭한 견인주의자(堅忍主義者)로 비유되는 나무의 속성을 통하여 지은이의 인생관을 표현하였다.

나무(躍舞) [나무]
- **나무** 「007」「명사」『무용』 마음을 깨끗이 하고, 도를 닦는 장소를 깨끗이 한다는 뜻으로 추는 불교 의식 무용. 양손에 바라를 쥐고 배꼽을 중심으로 하여 머리 위로 들어 올리거나 좌우로 돌리고, 빠른 동작으

위와 같은 기술 방식을 취하게 되면, 각 의미 항목마다 관련어를 제시하게 되는데, 의미 항목마다 관련어를 제시하는 것은 『고려대』의 의미 항목 처리에서도 확인할 수 있다. 그러나 모어화자들은 이러한 의미 항목들을 동음어처럼 독립적으로 인식하는 것은 아니기 때문에, 언어 사전에서는 의미 항목 간의 연관성을 기술하는 차원에서 다의 항목의 관계망을 파악할 필요가 있다. 이는 결국 해당 어휘 항목의 의미 속성에 대한 인식 양상을 파악하는 것과 같다.

(30) '손'에 대한 『표준』의 기술
①사람의 팔목 끝에 달린 부분. 손등, 손바닥, 손목으로 나뉘며 그 끝에 다섯 개의 손가락이 있어, 무엇을 만지거나 잡거나 한다. ¶손으로 잡다.
②손끝의 다섯 개로 갈라진 부분. 또는 그것 하나하나. ¶손에 반지를 끼다.
③일을 하는 사람. ¶손이 부족하다.
④어떤 일을 하는 데 드는 사람의 힘이나 노력, 기술. ¶나는 부모님이 돌아가셔서 할머니의 손에서 자랐다.

⑤어떤 사람의 영향력이나 권한이 미치는 범위. ¶손에 넣다.
⑥사람의 수완이나 꾀. ¶장사꾼의 손에 놀아나다.

위에서 ①과 ②는 ‘손’의 의미 속성, 즉 ‘손’의 작용역 중 ‘형상’과 ‘구성’ 그리고 ‘기능’에 대한 인식을 보여주는 것이다. 이러한 작용역은 문맥과 조응하여 활성화되는데, 이는 ‘손’과 부분전체 관계를 맺는 어휘들의 의미가 활성화되는 것으로 볼 수 있다. 그렇다면 여기에서 확인되는 의미 항목 간의 관계는 ‘나무’의 작용역에 따른 문맥 내 의미 작용과 같은 차원에서 볼 수 있을 것이다. 이처럼 ‘손’의 ‘형상’, ‘구성’, ‘기능’에 대한 인식에 천착하면, ‘손’의 의미를 확장적으로 인식하게 되는 이유를 설명할 수 있다. 즉, ‘손’이 ‘팔’의 구성 부분이라는 지식은 ‘손’을 ‘팔’의 의미로 확장할 수 있는 근거가 되는 것이다.[11] 이는 비유 표현 중 환유의 원리이기도 하다.

그런데 (30)을 보면 ③~⑥의 의미항목 간 관계망은 ①~②의 의미 항목 간 관계망과 차원이 다름을 알 수 있다. 이러한 차이는 ③~⑥까지의 관계가 ①~②의 관계와 달리 예측이 쉽지 않은 이유가 된다. 즉, ①~②의 의미적 관계는 환유적 전환에 의해 형성된 관계라면, ③~⑥의 의미적 관계는 은유적 전환에 의해 형성된 관계라 할 수 있다.

이때 은유적 전환에 의해 형성된 의미 항목 간의 관계망은 환유적 전환에 의한 관계망보다 복잡한 양상을 띠게 된다. ‘손’이 ‘사람’, ‘힘(노력)’, ‘기술’, ‘영향력 범위’, ‘꾀’ 등과 맺는 의미적 관계는 근원영역인 ‘손’의 속성 정보만으로는 설명할 수 없기 때문이다.

4.2. 의미 속성의 은유적 전환과 관계망

의미 속성의 은유적 전환은 목표영역을 근원영역의 관점에서 개념화하는 것이기 때문에 근원영역을 나타내는 어휘의 의미는 목표영역의 의미로 확장된다고 볼 수 있다. (30)에서 ③-⑥까지의 의미 확장은 은유적 개념화에 따른 의미 확장이다. ③의 경우는 ‘부분(손)’으로 ‘전체(사람)’을 개념화한다는 점에서 환유의 일종으로 볼 수도 있지만, 이는 ‘근원영역(손)’으로 ‘목표영역(노동력)’을 개념화한다는 점에서 궁극적으로는 은유적 개념화에 의한 의미라 할 것이다. 이러한 개념화를 바탕으로 하여 형성되는 근원영역과 목표영역의 관계망이 확장되면 의미항목의 관계망도 복잡하게 형성되는 것이다.

이처럼 관계망은 인간의 보편적 인식에 따라 이루어지는 측면도 있지만, 해당 언어권의 문화적 특수성을 반영하여 이루어지기도 한다. 이런 점은 ‘근원 영역과 목표 영역의 관계망’에 대한 논의를 촉발한 측면이 있다. 도원영 외(2018)에서는 은유 데이터베이스를 구축하여 이러한 관계망을 정리한 바 있다.

(31) 은유 데이터베이스 구축의 예시(도원영 외, 2018: 73)

위의 은유 데이터베이스는 한국어 은유표현에서 근원영역과 목표영역의 관계 양상을 파악하는 데 참조할 수 있을 것이다. 여기에서 주목할 것은 목표영역을 중심으로 은유 표현을 정리하고 분석하기 위해 ‘주제어’를 설정하여 목표영역을 분류하고 있는 점이다. 이러한 은유 데이터베이스에서는 근원영역이 관여하는 주제어의 관계망을 파악할 수도 있고, 주제어에 따라 근

11) (30)에는 기술되어 있지 않지만, “손을 번쩍 들다.”라는 문장에서 ‘손’은 ‘팔’과 대응되는 의미로 쓰이는 것은 이러한 인식의 작용이다. 실제 ‘손’과 ‘팔’의 연관성은 ‘손목시계’와 ‘팔목시계’의 관계에서도 확인할 수 있다.

	A	B	E	F	G	H	I
1	분류어	표현	유형	근원 영역	목표 영역	주제어	출처
1131	꽃	꽃 본 나비 (물 본 기러기)	1	꽃	사랑하는 대상	연인	표준
1132	꽃	향년 27세 '꽃이 지다'	2	꽃이 지다	죽다	죽음	비즈맨터, 2017.12.19
1133	꽃	나 머리에 꽃 달았다. 청줄 놨	2	꽃을 달다	미치다	광기	http://blog.nave
1134	꽃	푸른 하늘에 꽃 같은 새를 날리듯	3	꽃	새	새	우리교육 중등용 95년 1월호
1135	꽃	그때 백제의 꽃이 피었다.	2	꽃이 피다	전성기를 이루다	황금기	미주중앙일보, 2014.04.28.
1136	꽃	그녀는 우리 회사의 꽃이었다	1	꽃	아름다운 여성	여성	표준
1137	꽃	올림픽의 꽃	1	꽃	중요한 것	사물	국립극장 매거진 미르, 2017.09.01.
1138	꽃	꽃 본 나비 담 넘어가랴	1	꽃	사랑하는 대상	연인	표준
1139	꽃	꽃 본 나비 불을 헤아리랴	2	꽃 본 나비 불을 헤아리랴	남녀 간의 정이 깊으면 죽음을 무릅쓰고서라도 찾아가서 함께 사랑을 나눔.	열정	표준
1140	꽃	꽃은 꽃이라도 호박꽃이라	1	호박꽃	못생긴 여성	여성	표준
1141	꽃	꽃은 목화가 제일이다	2	꽃은 목화가 제일이다	겉치레보다는 실속이 중요함.	실속	표준

원영역의 관계망을 파악할 수도 있는데, 이때 주제어는 은유표현을 어휘의미망의 의미영역과 연결 짓는 고리가 되는 것이다. 특히 언어사전에서는 '사태'(꽃이 지다, 꽃을 달다 등)가 근원영역이 되는 은유표현을 관용표현으로 다루게 되는데, 이 데이터베이스의 주제어를 고리로 언어사전의 관련어 정보에 관용표현을 포함할 수 있을 것이다.[12]

위의 은유 데이터베이스에서는 주목하고 있지 않지만, 근원영역에 해당하는 어휘의 계열적 관계망은 목표영역에 해당하는 어휘의 계열적 관계망과 혼성되어 은유 표현을 확장하는 역할을 한다. 따라서 근원영역과 목표영역에서의 계열적 관계망 포착하는 것은 은유표현의 생성과 해석 원리를 설명하는 출발점이 된다.

정원용(1996)과 최경봉(2002) 등에서는 은유 표현의 생성과 해석 양상을 의미장 이론과 연결지어 설명한 바 있는데, '근원영역의 의미장이 목표영역을 개념화하면서 목표영역의 의미장과 상호작용하는 양상'을 살펴보는 일은 개념적 은유 이론이 제기된 이후의 모든 은유 연구에서 중심적인 탐구 사항이었다. 이때 개념 은유의 도식은 여러 방향으로 형성될 수 있는데, 개념 은유 도식의 확장과 근원영역 의미장의 상호작용 양상을 정리하는 것은 어휘 정보를 정교하게 기술하는 바탕이 될 것이다.

가령 <밋밋하다=특징이 없다>라는 개념도식은 <밋밋하다=재미없다>, <밋밋하다=맛없다> 등의 개념도식으로 확장될 수 있고, 이에 따라 '밋밋하다'가 포함된 <사물의 굴곡> 의미장은 '내용', '성격', '맛' 등의 평가와 관련된 의미장과 상호작용하여 은유 표현을 생성할 수 있다. 그렇다면 <사물의 굴곡>과 관련한 의미장에서 '울퉁불퉁하다, 들쑥날쑥하다, 튀어나오다, 밋밋하다…' 등의 관계망은 '내용', '성격', '맛' 등의 평가라는 목표영역을 개념화하는 데 활용될 수 있다. 이러한 개념화의 양상을 정리하면, 어휘의 관계망을 전제로 어휘의미를 기술해 볼 수 있을 것이다.

(32) '밋밋하다'에 대한 『표준』의 기술
①생김새가 미끈하게 곧고 길다.
②경사나 굴곡이 심하지 않고 평평하고 비스듬하다.
③생긴 모양 따위가 두드러진 특징이 없이 평범하다.

위의 기술 내용을 보면, 의미항목 ③이 은유적 개념화에 따른 의미를 반영하고 있지만, 개

12) 국어사전에서는 관용표현을 부표제어로 다루고 있지만, 부표제어인 관용표현에는 관련어가 제시되어 있지 않다. 또한 표제어마다 관련어는 제시되어 있지만, 그 관련어에 관용표현이 포함되어 있지는 않다.

념도식의 확장에 따른 관계망을 반영하고 있지는 않다. 위에서 거론한 개념도식을 적용한다면 『표준』보다는 『고려대』의 기술 내용이 개념도식의 확장을 수용할 수 있는 여지가 있다.

(33) '밋밋하다'에 대한 『고려대』의 기술
①(무엇이) 생김새가 튀어나온 곳 없이 미끈하게 곧다.
②(지형이) 굴곡이나 경사가 그다지 심하지 않고 평평하며 비스듬하다.
③(무엇이) 두드러진 특징이 없이 평범하다.

다만 어휘의 관계망을 구체화하기 위해서는 ③의 의미를 세분화할 필요가 있을 것이다. '내용'의 평가에서 '밋밋하다'는 '재미없다, 따분하다, 재미있다, 흥미롭다' 등과 관계망을 형성하지만, '맛'의 평가에서 '밋밋하다'는 '맛없다, 밍밍하다, 싱겁다, 맛있다, 맛깔나다' 등과 관계망을 형성할 것이기 때문이다.

5, 결론

참고문헌

강범모(2017),『한국어 명사의 화제 의미관계와 네트워크』, 한국문화사.
김숙정 외(2019). 은유의 범주와 유형 - 데이터베이스 구축의 관점에서-. 겨레어문학, 62, 221-249.
김진해(2007), 표준국어대사전의 관련어 정보와 어휘관계 기반 사전 기술『한국어의미학』24, 23-50.
김진해(2013), 연어 연구의 의미론적 함의-목록과 경향 사이에서-, 『국어학』68, 189-223.
남길임(2012), 어휘의 공기 경향성과 의미적 운율, 『한글』298, 135-164.
도원영 외(2018), 은유 데이터베이스 구축을 위한 시론, 『한국어 의미학』61, 55-79.
박만규(2002), 다의어의 의미 분할과 의미 부류, 『한글』257, 201-242.
옥철영(2007), 국어 어휘의미망 구축의 개념과 사전 편찬, 『새국어생활』17-3, 27-50.
유희정·최경봉(2020), 동물 부위 은유의 특징과 작용 원리, 『한국어의미학』68, 25-48.
윤애선(2007), 국내외 어휘의미망의 구축과 활용, 『새국어생활』17-3, 5-25.
윤애선(2012), 한국어 어휘의미망 KorLex 2.0 - 의미 처리와 지식 공학을 위한 기반 언어 자원-, 『한글』295, 163-201.
이동혁(2007), 의미 범주 체계의 구축과 사전에서의 활용, 『한국어 의미학』24, 51-82.
이동혁(2010), 사전과 어휘의미망에서의 의미 기술, 『한국사전학회 학술대회 발표논문집』, 3-19.
이성헌(2005), 전자사전 구축과 의미부류-세종 명사 의미부류 체계의 예, 『한국사전학』5, 103-138.
이성헌(2007), 세종 전자 사전의 어휘 의미 부류 체계, 『새국어생활』17-3, 51-67.
임근석(2011), 한국어 연어 연구의 전개와 쟁점에 대하여, 『국어학』61, 359-466.

임홍빈·임근석(2004), 21세기 세종계획 전자사전구축분과 연어사전의 정보구조와 기술내용, 『한국사전학』 4, 99-130.
정원용(1996), “隱喩의 意味와 構造”, 논문집 17-1, 경성대학교, 47-62.
차준경·임해창(2010), 어휘의미망의 형태 의미 관계 설정- 국어의 사건 명사를 중심으로 -, 『한민족문화연구』 34, 165-191.
최경봉(2001), 지식기반 구축을 위한 어휘의 의미 분류, 『담화와인지』 8-2호, 275 - 303.
최경봉(2002), 은유 표현에서 어휘체계의 의미론적 역할,『한국어학』 15, 283-306
최경봉(2005), ‘물명고(物名考)’의 온톨로지와 어휘론적 의의,『한국어의미학』 17, 21-42.
최경봉(2015), 『어휘의미론-의미의 존재 양식과 실현 양상에 대한 탐구』, 한국문화사.
최경봉(2017), 한국어 단어 단위와 의미-의미 단위의 인지적 실체를 중심으로-, 『한국어학』 77, 65-93.
최경봉(2019), 구문과 어휘의미의 상관성 고찰, 『국어학』 89, 255-283.
최경봉·도원영(2005), 한국어 동사 의미망 구축을 위한 상위온톨로지 구성에 관한 연구, 『한국어학』 28, 217-244.
최호섭·옥철영(2002), 한국어 의미망 구축과 활용-명사를 중심으로-, 『한국어학』 17, 301-329.
한정한·도원영(2005), 한국어 동사 의미망 구축을 위한 어휘의미관계 유형, 『한국어학』 28, 245 - 268.
황순희(2010), 인지동사의 의미분류와 어휘의미망 표상, 『언어연구』(한국현대언어학회)26-2, 373-406

Boas, Hans C.(ed.)(2009), *Multilingual FrameNets in Computational Lexicography*, Mouton de Gruyter.
Durkinn, Philip(2016), *The Oxford Handbook of Lexicography*, Oxford University Press.
Lakoff, G., & Johnson, M. (1980), *Metaphors we live by,* Chicago: The University of Chicago Press.
Trim, R.(2007), *Metaphor Networks: The Comparative Evolution of Figurative Language,* New York: Palgrave Macmillan.
Voessen, Piek(1998), *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Klwer Academic Publishers. 한정한 외 역(2004), 『유로워드넷』, 한국문화사

〈어휘의미망에 대한 인식과 사전의 구성〉에 대한 토론문

정성훈(목포대학교)

본 연구는 어휘의미망을 구축하기 위한 실천적 연구와 관련하여, 어휘의 관계망 정보에 대한 언어사용자의 인식 양상을 반영하는 국어사전의 구성 문제를 논의하였습니다. 어휘의미망 구축은 자연언어처리(NLP)에 활용하기 위한 기초 작업이기도 하면서 언어사전을 편찬하기 위한 기초 작업이기도 합니다. 또한 연구자께서도 언급하였다시피, 언어사전 편찬과 어휘의미망 구축은, 상호 영향 하에서 어휘의 관계망에 대한 언어사전의 기술 내용이 어휘의미망에 반영되고, 어휘의미망 구축의 성과가 언어사전의 어휘 관계망 정보가 정교화되어지는 선순환 관계입니다. 그런데 선생님께서는 어휘의미망 구축 연구로부터 영향을 받은 한국어 사전이 변화하고는 있으나, 선순환이 아닌, 관련어 정보를 확대하는 수준에 머물러 있다고 문제제기를 하셨습니다. 이에 한국어 사전의 어휘 관계망 정보를 정교화하기 위해서는 관련어를 체계화하고, 어휘 관계망 정보를 반영하도록 한국어 사전의 미시구조를 재구성해야한다고 주장하셨습니다.

특히 선생님께서는 어휘 의미를 계열적 의미와 결합적 의미로 구분하는 방법론에 따라, 어휘의미망을 계열적 관계망과 결합적 관계망으로 구분하고 이를 정교화하는 방안을 찾으셨습니다. 선생님께서도 말씀하셨지만, Saussure(1916)에서 Lyons(2011)로 이어지는 구조의미론에서는 어휘에 대한 의미적인 관계를 계열적 관계와 결합적 관계로 구분하고, 어휘망 구조 속에서 어휘 간의 관계와 대립으로 각 단어의 뜻(sense)이 나타납니다. 계열적 의미는 유의/동의, 반의, 상하의 등의 어휘 관계 속에서, 결합적 의미는 collocation이나 colligation 등과 같은 연속체의 관계 속에서 현저해지는 것입니다. 또한 선생님께서는 은유적 개념화에 주목하여 source domain과 target domain의 연결에 따라 창조되는 비유적 의미 간의 관계망도 살펴보셨습니다. 토론자는 선생님의 고민과 주장에 일견 동의하면서도 궁금한 점이 몇 가지 떠올랐습니다. 이에 본 토론자는 선생님께서 말씀하신 내용을 토대로 궁금한 점을 묻는 것으로 토론을 대신하고자 합니다.

1. 계열적 관계망에 대한 문제입니다. 선생님께서는 기능적 관계를 포함하여 다양한 층위(levels)의 어휘의미망을 구축한다면 문맥 내의 선택 제약 등과 같은 어휘 의미작용을 설명할 수 있다고 보았습니다. 따라서 한국어 사전에 분류학적 관계와 기능적 관계를 구분하여 기술할 것을 제안하셨습니다. 즉 '존재론적인 인식 체계로서의 관계망(상위온톨로지)'과 '특정 영역에 대한 지식 체계로서의 관계망(영역온톨로지)'을 아우르는 두 층위의 관계망을 설정하는 것이 한국어 사전의 어휘 관계망 정보를 정교하는 작업이라고 하셨습니다. 언어사용자의 분류의식을 반영할 수 있는 일정한 틀을 세우고 이를 활용하여야 한다고도 말씀하셨습니다. 그런데 일정한 틀(형상, 구성, 기능 작인/상황유형과 상황성분) 안에서 언어사용자가 어휘 관계망을 인식하는 토대, 특히 문맥적 의미를 포함하는 것을 사전에 기술하는 하는 것은 매우 어려워 보입니다. 언어 사용자가 사용할 수 있는 모든 언어적 문맥을 고려하여 그 단어의 계열적 관계망을 구축하는 작업은 많은 시간과 비용이 소용될 것으로 보이기 때문입니다.

   또한 선생님께서는 다차원적 분류체계가 언어사용자의 역사적 분류의식도 반영할 수 있

다고 하셨습니다. 그런데 '계열적'이라는 용어 자체가 공시적 언어체계를 가정한다고 본다면 어휘가 사용되는 환경에 대한 지식과 경험(문화 포함)까지 기술할 수 있는 있겠으나, '계열적 관계망'으로 역사적인 통시적 언어체계를 어떻게 반영할 수 있는 것인지 설명을 더 듣고 싶습니다.

2. 다음은 결합적 관계망의 문제입니다. 코퍼스언어학이나 어휘의미론에서는 결합적 의미 관계 정보를 어떻게 포착하여 기술할 것인지에 관심이 많습니다. 특히 어휘의미망 구축과 관련하여 어휘들의 결합적 관계에 주목을 하고 있습니다. 코퍼스언어학에서는 빈도 기반의 통계적 접근으로 단어 간 결합적 관계의 의미망을 포착하기도 하고 통사의미론에서는 논항 관계를 기초로, 문형정보를 제공하기도 합니다. 선생님께서는 '결합적 관계의 의미망을 포착하기 위한 통계적 접근의 성과를 언어사전에 반영하고 이를 기반으로 결합적 관계망 정보를 어휘의미망에 포함한 기계가독형 전자사전을 구축하는 절차를 고려할 필요가 있다'고 하시면서, 연어 관계에 주목하셨습니다. 기존 사전에서는 용례를 중심으로 연어 관계를 기술하고 있으나 모든 용례를 사전에 기술할 수 없고, 결합할 수 있는 어휘와 결합할 수 없는 어휘의 차이, 즉 제약과 의존성을 기술하지 못하였다는 한계가 있다고 하시면서, 연어 관계를 중심으로 한 전자사전과 그 구조의 가능성을 제시하였습니다. 그런데 결합적 관계망에 대한 통계는 결국 해당 어휘의미망이나 사전을 구축할 때 어떤 코퍼스를 사용하느냐의 문제로 귀결됩니다. 현재 범용 코퍼스인 세종코퍼스가 있으나 각 장르별, 구어와 문어별로 통계적으로 유의미하게 결합적 의미 관계 차이가 나타나기도 합니다. 즉 연어 관계를 중심으로 결합적 관계망을 적절하고 타당하게 포착하기 위해서는 보편타당한 대용량의 코퍼스가 필수적인데 이 문제를 어떻게 해결해야 한다고 생각하시는지요?

3. 마지막으로 은유적 개념화를 통한 의미 관계망 구축 문제입니다. 은유적 개념화는 근원영역의 어휘를 목표영역의 어휘로 사상(mapping)하여 형성된다는 점에서 사상 과정에서 형성되는 근원영역의 어휘와 목표영역의 어휘 간 관계망 또한 중요하게 다룰 필요가 있다는 점에서 선생님의 생각에 동의합니다. 그런데 일상적인 언어생활에서뿐만 아니라 문학에서도 은유적 개념화는 일어날 수 있다고 생각합니다. 즉 한국문학의 은유적 개념화는 한국어의 일반적인 은유적 개념화에 비해 보다 창의적이고 덜 진부할 것입니다. 반대로 우리가 은유 또는 은유적 개념화라고 생각하지도 않는 사은유(dead metaphor)도 있습니다. 이런 것들에 대한 분류나 의미관계망은 어떻게 구축하고 기술할 수 있을지에 대한 선생님의 고견을 듣고 싶습니다.

여기까지입니다. 혹시 제 짧은 지식으로 인해 선생님 논문의 주장이나 근거를 잘 이해하지 못한 점이 있다면 그것은 전적으로 제 잘못이니 넓은 마음으로 양해 부탁드립니다. 지금까지 제 토론을 경청해 주셔서 감사합니다.

어휘적 응집성과 한국어 어휘의미망
Lexical Cohesion and Korean WordNet
-체계기능언어학의 관점-
from Systemic Functional Linguistics point of view

한정한(단국대)

1. 서론

이 글의 목적은 두 가지이다. 첫째로, 체계기능언어학(Systemic functional linguistics)의 어휘적 응집성(Lexical Cohesion)의 개념을 소개하고, 이것이 서로 다른 국어 장르(텍스트)에서 어떤 차별적 실현 양상을 보이는지 밝혀 보는 것이다. 이를 통해 어휘적 응집성의 종류와 장르적 관련성을 검토해 본다. 둘째로, 어휘적 응집성의 종류와 그 특성을 잘 반영한 한국어 어휘의미망의 구축을 통해 중의성 문제를 해소할 수 있는 가능성을 제시해 보는 것이다.[1)]

체계기능언어학(SFL) 내에서 발표된 많은 담화 텍스트 분석 연구들은 Halliday & Hasan(1976)의 Seminar Book과 그 이후의 연구 업적에 크게 영향을 받아 왔다. Baker(1992, 180), Moreno(2003, Xi 2010)에서는 그들의 업적을 "지금까지 응집성에 관해 알려진 연구 중에서 가장 상세하다."[2)]는 찬사를 보냈다.

여러 가지 분류 기준에 따라서 달라질 수 있지만, Halliday & Hasan(1976)에 의하면, 응집성은 우선 세 가지 종류로 나누어 볼 수 있다. 각각 동일지시(co-referentiality), 동일종류지시(co-classification), 그리고 동일확장지시(co-extension)가 그것이다(Halliday & Hasan 1985, 73).

먼저 동일지시는 응집 관계를 형성하는 두 개 혹은 여러 개의 요소들이 지시적 동일성(referential identity)을 갖는 경우를 말한다. 아래 예문 (1)에서 'a little nut tree'와 대명사 'it'의 관계처럼 같은 형태를 지시하는 대명사, 또는 동일 형태를 반복하는 어휘들이 여기에 속한다. 둘째로, 동일확장지시는 위 (1)에서 silver(은)와 golden(금) 같은 경우로, 여기서 '은'과 '금'은 모두 '금속'에 포함된다. 이와 같이 동일확장 지시는 두 언어 요소들 간에 상위어-하위어, 또는 하위어-상위어의 어휘 관계로 나타난다. 셋째로, 동일종류지시는 (2)에서 'play the cello'와 생략 조동사 'does'처럼 동일종류의 다른 첼로(cello)를 가리킬 때 나타난다.

(1) I had a little nut tree. Nothing would it bear. But a *silver* nutmeg. And a *golden* pear.
(2) I paly the cello. My husband does, too. (Halliday & Hasan, 1985/1989: 73)

여기서 알 수 있는 것은, 개별 문장들의 응집성이 높으면, 읽는 사람은 낱낱의 문장들이 하

1) 첫 번째 목적이 핵심 내용이고, 두 번째 목적은 선행 연구의 소개 수준임을 미리 밝힌다.
2) "...the best known and most detailed model of cohesion available."

나의 텍스트를 구성하고 있다고 인식하게 된다는 것이다. 이것을 체계기능언어학에서는 텍스트성(Textuality)이라고 한다. 따라서 응집성이 높은 텍스트는 전체 의미를 이해하기가 더 쉽고, 따라서 번역하기도 쉽고, 그렇게 쓰인 텍스트들은 잘 된 텍스트(well-formed texts)라는 평가를 받는다.

응집성을 분류하는 또 다른 기준으로는 응집성 거리(cohesion ties)라는 게 있다(Halliday & Hasan, 1976). 이것은 응집성을 가지는 두 요소 사이의 거리를 기준으로 응집력을 구분하는 것이다. 응집성 거리는 먼저 동일 문장 내부 요소들 간의 응집성과 문장과 문장 사이의 요소들 간의 거리로 나누어진다. 본고에서는 전자를 WS(within-sentence), 후자를 AS(across-sentence)로 각각 분류하여 통계 처리하였다. 그리고 문장 간 거리 응집성(AS)은 다시 인접거리 유대(immediate ties), 원거리 유대(remote ties), 매개거리 유대(mediated ties)로 하위분류된다. 인접거리 유대는 아래 (3)의 'Alice[1]-she[2]'에서 보듯이 선행 요소와 후행 요소의 응집관계가 바로 인접한 문장에서 나타나는 것이다. 그리고 원거리 유대는 Alice rubbed her eyes[3]-Rub as she woul[7]처럼 두 요소 사이의 거리가 확장되어 여러 문장에 걸치는 유대를 말한다. 마지막으로 매개거리 유대는 먼저 선행어가 나오고(Alice[3]), 한 참 떨어져서 후행 응집어(she[5]가 나오는데, 그 중간에 서로를 매개하는 매개어(she[4])가 나타나는 경우를 말한다. 후술하겠지만 이런 응집성 유대의 차이는 응집성의 종류에 따라서 상당한 편차가 있고, 또 장르에 따라서도 차이가 있다.

(3) The last word ended in a long bleAS, so like a sheep thAS Alice quite started [1]. She looked AS the Queen, who seemed to have suddenly wrapped herself up in wool[2]. Alice rubbed her eyes, and looked again[3]. She couldn't make out whAS had happened AS all[4]. Was she in a shop[5]? And was thAS really- was it really a *sheep* thAS was sitting on the other side of the counter[6]? Rub as she would, she could make nothing more of it[7]. (Halliday & Hasan, 1976: 330)

그리고 앞의 매개거리 유대는 다시 동일성 사슬(identity chain)과 유사성 사슬(similarity chain)로 하위구분 된다. 동일성 사슬은 지시하는 표현과 지시대상 사이에 동일지시 관계(대명사, 반복어, 등가어 등)를 맺는 것을 말하고, 유사성 사슬은 지시하는 표현과 지시대상 개체 사이에 동일종류지시나 동일확장지시의 관계가 있는 경우를 말한다.

체계기능언어학에서 응집성(cohesion)과 일관성(coherence)의 관계도 언급해 둘 필요가 있다. 가장 잘 알려진 응집성과 일관성의 차이는 응집성이 텍스트의 내적 요소들 간의 결속 관계를 다루는 분야라면, 일관성은 텍스트의 외적 (또는 상황맥락적) 요소들 간의 결속 관계를 다루는 분야라는 것이다. 다시 말하면, 응집성이 문법적, 어휘적 자원들 간의 결속 관계를 통해 얻어지는 결속이라면, "일관성은 텍스트 외부에 존재하는 경험적 맥락(예, 화자, 청자), 사회문화적 맥락들 간의 결속 관계를 통해 얻어지는 결속력이다. 텍스트를 구성하는 개별 언어 표현들은 일관성을 통하여 하나의 통일된 전체로 인식하게 하는 성질을 갖는다. 다음 예를 보자.

(4) T(urn)1 : 가져가실 건가요, 드시고 가실 건가요?
T2 : 테이크아웃 잔에 주세요. (이관규 외, 2021: 218)

언어 표현 자체에만 집중하면 (4)의 두 발화는 서로 전혀 상관없는 이야기를 하는 듯하다. 그러나 이 대화 속의 언어 표현들이 함께 나타날 수 있는 어떤 일관성 있는 상황 맥락(예, 커피 전문점에서 음료를 주문하는 상황)을 가정할 수 있으면 (4)는 하나의 텍스트로 볼 수 있다. 카페에서 '마시고 갈 것인지 가져갈 것인지'를 묻는 이유는 음료를 일회용 잔에 담아서 줄지, 다회용 잔에 줄지 결정하기 위해서다. 이런 경우 (4T1)과 (4T2)는 하나의 텍스트로 볼 수 있다. 그런데 체계기능언어학 내부에서도 응집성과 일관성은 정확히 구별되는 개념이 아니다(실제로 생략이나 대용이 많은 텍스트에서 더욱 그렇다.) 본고에서는 이것을 정도상의 의미정보(gradual semantics)로 처리하기로 하겠다.

본고에서는 다양한 종류의 어휘적 응집성을 서로 다른 장르의 텍스트를 대상으로 분석해 보려고 한다. 본고의 연구 대상 텍스트는 세월호의 신문 기사문과 세월호 희생자들의 부모들이 쓴 편지글이다. 불특정 다수를 대상으로 하는 격식적인 신문 기사문과 다분히 개인적이고 사적인 편지글에 따라서 응징성이 어떤 차이를 보이는지를 검토해 볼 계획이다. 그리고 이러한 응집성 정보를 담고 있는 잘 구축된 한국어어희의망이 왜 필요한지도 지적하고자 한다.

2. 연구 대상 텍스트 소개

본고의 주요 분석대상은 세월호 침몰 사고[3]의 신문 기사와 세월호 유가족이 쓴 편지글이다. '세월호 사건'을 공통의 주제로 하여, 불특정 다수 독자를 대상으로 쓴 신문기사와 사별한 자녀들에게 보내는 개인적인 편지글이 보이는 어휘 응집성의 차이를 밝혀보려는 것이 목적이다. 신문기사 수와 편지글 수의 문장 수를 비슷하게 맞추기 위해서 양적 조정이 있었다. 그 내용은 다음과 같다.

<표 1> 세월호 침몰 사고 신문기사

번호	신문사	기사 제목	문장수
1	동아일보	여색선 세월호 침물	16
2	서울신문	여객선 세월호 안산 단원고 학생 등 460여명 사고	9
3	국민일보	여객선 침몰 임박 - 안산 단원고 홈페이지 마비	7
4	한겨레	진도 침몰 여객선 1명 사망 197명 구조	20
5	경향신문	여객션 진도 해상 좌초 - 정부, 대책본부 가동	13
6	조선일보	세월호 구조 작업 난항	9
7	중앙일보	진도 여객선 침몰 실종인원 290여명으로 번복 돼	17
8	한국일보	진도 여객선 침몰 중 안산단원고 학생 대부분	7
9	세계일보	침몰 세월호 수색 중단	8
10	문화일보	고등학생 등 459명 탄 여객선 침몰	15
계			121

3) 세월호 침몰 사고는 2014년 4월 16일 오전 8시 50분경 대한민국 전라남도 진도군 조도면 부근 해상에서 여객선 세월호가 전복되어 침몰한 사고이다. 이 사고로 안산시의 단원고등학교 학생들을 포함 304명이 사망하였다. (https://terms.naver.com/entry.naver?docId=2119309&cid=43667&cASegoryId=43667).

<표 2> 세월호 침몰 사고 편지글

편지글	번호	제목	문장수
1	T1	'생각하면 눈물이 앞을 가리고'	18
2	T2	'그곳에선 잘 지내고 있지?'	21
3	T3	'우리 딸, 태어나면서부터'	25
4	T4	'아기가 있는 집인지 모를 정도로'	19
5	T5	'유치원 다닐 때도 혼자'	12
6	T6	'학교 다니면서 공부도 잘하고'	30
계			125

3. 두 종류의 응집력

앞장에서 응집성은 텍스트의 내적 요소들 간의 결속 관계를 다루는 분야, 그리고 일관성은 텍스트의 외적 요소(상황맥락)들 간의 결속 관계를 다루는 분야라고 소개했다. 여기서는 좀 더 구체적으로 들어가 보자. 아래 <표 3>은 체계기능언어학의 응집성 유형을 정리한 것이다.

<표 3> 체계기능언어학의 응집성의 종류(Halliday & Metthiessen, 2014: 608)

General type		**Grammatical zone [(location in) grammatical unit]**	**Lexical zone [lexical item]**
transitions between messages		CONJUNCTION [unit: clause]	
statuses of elements	in meaning	REFERENCE [unit: nominal, adverbial group]	LEXICAL COHESION [synonymy, hyponymy]
	in wording	ELLIPSIS-&-SUBSTITUTION [unit (complex): clause, nominal group, adverbial group]	[repetition, collocation]

<표 3>에 따르면, 먼저, 응집성이 작용하는 일반적인 타입(가로 층위)은 세 가지이다. '메시지 간 이행'(transitions between messages) 차원에서 작용하는 응집성, '의미하기'(meaning) 차원에서 작용하는 응집성, '표현하기'(in wording) 차원에서 작용하는 응집성이 그것이다. 메시지 간 이행 차원은 주로 절과 절을 이어주는 접속사(conjunction)에 의해 실현된다.[4] '의미하기' 차원은 주로 문법적으로 '지시', 어휘적으로 '유의어', '상위어' 등에 의해 실현되며, '표현하기' 차원은 문법적으로 '생략'과 '대체', 어휘적으로 '반복', '연어' 등에 의해 실현된다.

그리고 응집성의 형성에 사용되는 언어 자원의 종류에 따라서(세로 층위) 응집성을 구분하면 '접속, 지시, 생략 및 대용 그리고 어휘' 이렇게 네 가지로 구분 할 수도 있다. 마지막으로 이들 네 가지는 다시 그것의 언어 자원이 '문법적 자원'이냐 '어휘적 자원'이나에 따라서 '문법적 응집성'과 '어휘적 응집성'으로 분류되기도 한다.

4) 본고의 핵심 연구 주제는 어휘적 응집성이므로 접속사에 의한 응집성에 대해서는 언급하지 않기로 한다 .

3.1. 문법적 응집성(grammatical cohesion)

앞장에서 설명했듯이, 응집성은 어떤 종류의 언어적 자원이 쓰였는가에 따라서 각각 문법적 응집성과 어휘적 응집성의 두 가지로 하위 구분된다. 먼저 문법적 응집성을 간단히 알아보자.[5)]

먼저 특정 개체가 앞에 언급된 개체를 의미적으로 가리키는 것을 '지시'(reference)라고 한다. 즉, 지시적 응집성은 가리키는 말과 가리켜지는 말 사이의 응집성이 문법적 성분인 인칭대명사, 지시대명사, 명사(구), 동사(구) 등에 의해서 발생하는 응집성이다. 이러한 '지시'는 다시 지시 대상이 지시하는 표현에 선행하는 '전방조응'(anaphoric), 지시하는 표현이 지시 대상에 후행하는 '후방조응'(cataphoric)으로 나뉜다. 아래 (5)에서 'the blind mice-they'의 관계가 전방조응의 예가 되고, (6)의 'they-the blind mice'가 후방조응의 예가 된다.

(5) Three blind mice, three blind mice. See how they run! See how they run. (Halliday & Hasan, 1976: 31)
(6) See how they run! See how they run. Three blind mice, three blind mice.

그런데 지시가 '의미하기'를 기준으로 한 문법적 응집성이라면, 대용과 생략은 '표현하기'를 기준으로 한 문법적 응집성이라고 할 수 있다. 그리고 아래 (7)에서 명사구 '*two poached eggs*'와 대명사 '*the same*'은 대용 응집성의 예가 되고, (8)에서 조동사 'have' 뒤에는 선행하는 'been swimming'이 생략된 것이므로 생략 응집성의 예가 된다.[6)]

(7) I'll have two poached eggs on toast, please.
I'll have the same. (Halliday & Hasan 1976: 105)
(8) Have you been swimming? Yest, I have. (Halliday & Hasan, 1976: 67)

3.2. 어휘적 응집성(lexical cohesion)

한편 어휘적 응집성은 응집에 사용되는 언어 자원들이 어휘적인 것들이다. 어휘적 응집성은 본고의 핵심 주제이므로 연구 대상 텍스트인 신문기사와 편지글을 중심으로 자세히 언급하기로 한다. 먼저 아래 <표 4>는 어휘적 응집성의 예들을 소개한 것이다. <표 3>, H & M(2014: 608)>의 분류법 대신에 Gómez(2018: 114)의 분류를 사용하기로 한다. 전자는 응집성과 일관성을 구별하였지만, 후자는 응집성과 일관성을 엄격하게 구별할 필요가 없다는 입장이다.

---

5) 본고의 핵심 주제는 어휘적 응집성이므로 문법적 응집성은 깊이 있게 논의되지 않을 것이다.
6) 응집성에 관한 국외 논문을 정리해 보면, 첫째로 응집성의 종류에 관한 연구로는 Morenno(2003), Sanders & Maat(2006), Taboada(2006), Dontcheva-Navratilova(2009)가 있고, 응집성의 표지들에 관한 연구로는 (전방조응, 추상 명사, 반복어, 담화 표지 등) Ochs Keenan(1977), Tadros(1985), Francis(1986), Norrick(1987), Schmid(2000), Fraser(2005), Tannen(2007), Flowerdew(2010) 등이 있다.

<표 4> 어휘적 응집의 유형 및 어휘적 관계 (Gómez, 2018: 114)[7]

<table>
<tr><th colspan="3">어휘적 응집어의 유형</th><th colspan="2">응집성 유대의 종류</th></tr>
<tr><td rowspan="3">반복어(R)</td><td colspan="2">일치 반복어(ER)</td><td>문장 내부(WT)</td><td>문장과 문장 사이(AS)</td></tr>
<tr><td rowspan="2">부분일치 반복어(IR)</td><td>굴절 반복어(IIR)</td><td rowspan="11"></td><td rowspan="11">인접거리 유대(IM)
매개거리 유대(MM)
원거리 유대(DM)</td></tr>
<tr><td>파생 반복어(DIR)</td></tr>
<tr><td rowspan="2">유의어(S)</td><td colspan="2">유사 유의어(NS)</td></tr>
<tr><td colspan="2">명제 유의어(PS)</td></tr>
<tr><td rowspan="3">반의어(O)</td><td colspan="2">상보 반의어(CpO)</td></tr>
<tr><td colspan="2">등급 반의어(AO)</td></tr>
<tr><td colspan="2">방향 반의어(DO)</td></tr>
<tr><td rowspan="2">포함어(I)</td><td colspan="2">상위어(GI)</td></tr>
<tr><td colspan="2">하위어(SI)</td></tr>
<tr><td colspan="3">연상어(AC)</td></tr>
</table>

위 <표 4>에서 보듯이 어휘적 응집은 크게 반복어(Repetition), 유의어(Synonymy), 반의어(Opposition), 포함어(Inclusion), 그리고 연상어(Associative Cohesion)로 대분류 된다. 그리고 반복어(R)는 다시 일치 반복어(ER), 부분일치 반복어(IR), 굴절 반복어(IIR), 파생 반복어(DIR)로 하위구분 된다. 일치 반복어(ER)란 형태가 동일한 반복어를 말하고, 굴절 반복어(IIR)는 조사, 어미를 제외한 나머지 부분의 반복, 파생 반복어(DIR)는 파생접사를 제외한 나머지 부분이 일치하는 반복어이다. 그리고 유의어(S)는 다시 유사-유의어(NS)와 명제-유의어(PS)로 나누어진다. 유사-유의어는 의미가 동일하거나 유사한 유의어이고, 명제-유의어는 어휘보다 큰 구나 절 표현이 낱개의 어휘와 의미적으로 유사한 관계를 맺는 경우를 말한다(cry-burst into tears).[8] 반의어(O)는 '남자-여자'처럼 A이면 B가 아니다가 성립하는 상보적 반의어(CpO), '작다-크다'처럼 등급적 반의어(AO), '아래-위'처럼 방향적 반의어(DO)[9]로 나누어진다. 포함어(I)는 특정한 어휘에서 일반적인 어휘로 가는 상위어(GI), 일반적인 어휘에서 특정한 어휘로 가는 하위어(SI)가 있다. 마지막으로 연상어(AC)는 특정 의미 프레임(Frame Semantics)에서 서로 연관된 어휘들 간의 관계이다. 예를 들어 아래 (9)와 같은 예문에서 인칭대명사 '그'가 가리키는 대상이 '철수'인지 '민수'인지 모르므로 (9)는 하나의 텍스트인지를 판단하기 어려운 중의적인 문장이다. 그러나 '장례식'과 '살해하다'를 '살인 사건'이라는 프레임의 틀요소(frame elements)(연상어)로 인식할 수 있으면 '그'가 '민수'를 가리키고 있다는 것을 추정할 수 있다.

(9) S1 : 화가 난 철수가 민수를 살해하였다.
S2 : 그의 장례식은 월요일에 진행되었다.

이러한 어휘 응집성의 분류 방식을 이용하여 아래 5장에서는 세월호 침몰 사고의 신문 기사와 세월호 유가족이 쓴 편지글을 중심으로 어휘적 응집성의 특징들을 차례대로 살펴보기로

7) 한국어에 맞게 일부 수정하였음을 밝힌다.
8) 드라마나 서정적 글쓰기 장르에서 자주 발견됨
9) 관계적 반의어라고도 함. e.g., teacher-pupil.

한다.

4. 분석 결과

이 장에서는 앞장 <표 1> 세월호 침몰사고 신문기사와 <표 2> 세월호 침몰 사고 편지글 목록에서 제시한 10개 신문기사와 6개의 편지글을 대상으로 3.2.장 <표 4>에서 소개한 15개의 어휘적 응집성 유형을 비교 검토해 보았다.

아래 <표 5>는 10개 신문 기사의 어휘 응집성 종류별 빈도를 종합한 결과이다 그리고 <표 6>은 6개 편지글의 어휘 응집성 종류별 빈도를 종합한 것이다.

<표 5> 신문 기사 어휘 응집성 종류별 빈도 (종합)

신문	ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계	카이제곱 χ^2	p-value
WS	10	1		8		2	9		4	44	41	119		
AS	185	28	30	38	8	7	3	2	3	16	77	397		
계	195	29	30	46	8	8	12	2	7	60	118	516	150.30	0.001

<표 6> 편지글 어휘 응집성 종류별 빈도 (종합)

편지	ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계	카이제곱 χ^2	p-value
WS	10	2	2	1	5		3		1	1	22	47		
AS	95	28	5	4	7	4	1	1	2	3	30	179		
계	105	30	7	5	12	4	4	1	3	3	52	226	36.84	0.00164

<표 5>와 <표 6>에서 볼 수 있듯이 전체적으로 신문 기사에서 516개, 편지글에서 226개, 총 742개의 어휘적 응집성 사례들이 분석 되었다. 이것은 동일 문장 내 어휘 응집성(WS)과 문장과 문장 사이 어휘 응집성(AS)의 사례를 모두 합산한 것이다.

여기서 제일 눈의 띄는 것은 신문 기사의 어휘 응집성 수가 편지글의 어휘 응집성 수보다 거의 2.3배 높다는 것이다. 이것은 통계적으로도 충분히 유의미한 값을 얻었다. <표 5>의 카이제곱값(chi square values)은 $\chi^2(7)$[10] = 150.30, $p < .001$ 이다. 그리고 <표 6>의 카이제곱값은 $\chi^2(8) = 36.84$, $p < .00164$ 이다.[11]

편지글에서 어휘적 응집어들이 상대적으로 적은 것은 테너(참여자) 변수, 즉 글쓴이의 사적이고 정서적인 감정 변화가 텍스트의 정보 전달적 요소보다 더 중요하게 작용한 것으로 보인다. 또 생략 어휘들이 상대적으로 많은 것도 영향이 있어 보인다. 이에 반해 신문 기사문은 글쓴이가 직업적인 전문가들이고 다수 독자를 대상으로 쓴 계획된 글이기 때문에 이러한 결과를 얻었지 않나 생각한다.

통계분석의 두 번째 특징은 유형별 출현 빈도에 일정한 순서가 발견된다는 것이다. 먼저 신문기사에서는 아래 (10), 편지글에서는 아래 (11)의 순위 결과가 나왔다.

---

10) 괄호 숫자는 자유도.

11) p값이 3가지 유의 수준 <.001, <.01, <.05 모두에서 유의미한 결과값으로 나왔다..

(10) 신문기사 : ER > AC > SI > NS > DIR > IIR > AO > CpO > PS > GI > DO
일치 반복어(ER, 195) > 연상어(AC, 118) > 하위어(SI, 60) > 유사 유의어(NS, 46) > 파생 반복어(DIR, 30) > 굴절 반복어(IIR, 29) > 등급 반의어(AO, 12) > 상보적 반의어(CpO, 9) > 명제 유의어(PS, 8) > 상위어(GI, 7) > 방향 반의어(DO, 2)

(11) 편지글 : ER > AC > IIR > PS > DIR > NS > AO = CpO > SI = GI > DO
일치 반복어(ER, 105) > 연상어(AC, 52) > 굴절 반복어(IIR, 30) > 명제 유의어(PS, 12) > 파생 반복어(DIR, 7) > 유사 유의어(NS, 5) > 등급 반의어(AO, 4)=상보 반의어(CpO, 4) > 하위어(SI, 3)=상위어(GI, 3) > 방향 반의어(DO, 1)

신문기사와 편지글 모두에서 반복어(R)는 가장 강력한 어휘 응집성이라는 것을 알 수가 있다. 반복어(R)는 신문 기사에서 37.7%(195/516)에 해당하고, 편지글에서도 46.4%(105/226)에 해당한다. 게다가 파생 반복어(DIR, 30)와 굴절 반복어(IIR, 29)를 일치 반복어(ER, 195)에 포함시키면 무려 신문기사는 49.2%(254/516)이고, 편지글도 파생 반복어(DIR, 7)와 굴절 반복어(IIR, 30)를 일치 반복어(ER, 105)에 합치면 무려 62.8%(142/226)가 반복어(R)가 된다. 아래 (12), (13)은 신문기사와 편지글에 나타나는 반복어를 보인 것이다.

(12) 신문기사(T1) - 동아일보
S1 : 전남 진도 해상에서 발생한 여객선 침몰사고의 사망자가 6명으로 늘었다.
S2 : 그러나 290여 명이 실종 상태여서 사망자 수는 앞으로 더 늘 것으로 우려된다.
S4 : 사망자 중 승무원 박지영(22•여)씨, 안산 단원고 2학년 정차웅(17)군, 권오현(17)군, 임경빈(17)군 등 4명은 신원이 확인됐으나…

(13) 편지글(T1) - '생각하면 눈물이 앞을 가리고'
S6 : 학교 다니면서 공부도 잘하고 혼자 다 해서 엄마, 아빠는 신경 쓸 게 없었는데…
S9 : 엄마, 아빠가 너무 무심했어.
S11 : 하영아, 엄마, 아빠는 잘 지내고 있어.
S16 : 나중에 엄마, 아빠가 꼭 찾아갈게.

반복어(R)에 의한 어휘 응집이 이렇게 높게 나타나는 데는 몇 가지 이유가 있어 보인다. 우선, 반복어는 공간적으로 (멀리) 떨어져 있는 동일한 주제를 다시 가져 오는 가장 확실한 방법이다. 둘째, 반복어는 매크로 레벨 주제(예, (12)의 세월호 사건)보다 마이크로 레벨(예, (12)의 사망자)에서의 정보 처리를 집중적으로 관리해 주는 데 더 효과적이다. 셋째, 반복어는 텍스트 내에 암묵적으로 동의되고 있는 아이디어에 대한 확인 및 부인을 반복 표현하는데도 자주 나타난다. 예컨대, 아래 (14)는 '사망하다'를 반복함으로써 이 사실을 재차 확인시켜 주고 있다.

(14) 신문기사(T6) - 조선일보
S1 : 전남 진도에서 발생한 여객선 세월호 침몰 사고로 현재까지 단원고 2학년 학생을 포함, 2명이 사망한 것으로 알려졌다.
S2 : 16일 오전 8시 55분쯤 청해진해운 소속 여객선 '세월호'가 전남 진도군 관매도 인근 해상에서 침몰해 오후 3시 30분 현재 선사 여직원 박지영(22)씨와 단

원고 2학년 정차웅(17)군이 사망한 것으로 전해졌다.

S6 : 침몰중 구조신고를 받은 해양경찰 등이 긴급 출동해 탑승객들을 구조했으나 2명이 사망하고 293명의 생사 여부가 확인되지 않는 등 대형 참사를 피할 수 없는 상황이다.

통계 분석에서 발견되는 세 번째 특징은 어휘적 응집어의 종류별 빈도 순서가 신문기사나 편지글에서 큰 차이가 나지 않는다는 것이다.

(15) 신문기사 순서
반복어(R, 254) > 연상어(AC, 118) > 포함어(I, 67) > 유의어(S, 54) > 반의어(O, 23)

(16) 편지글 순서
반복어(R, 142) > 연상어(AC, 52) > 유의어(S, 17) > 반의어(O, 9) > 포함어(I, 6)

특히 반복어(R) > 연상어(AC)의 순서가 압도적으로 동일하고, 포함어, 유의어, 반의어는 포함어(상하위어)를 제외하면 대동소이하다. 이것은 우리가 텍스트를 분석하거나 중의성을 해결할 때 반복어와 연상어가 가장 중요한 어휘적 응집어이고, 포함어, 유의어, 반의어는 텍스트의 장르적 차이를 반영하여 편차를 갖는 것으로 추정하게 해 준다.[12)]

통계 분석의 네 번째 특징은 연상어(AC)의 비중이 반복어(R) 다음으로 매우 높다는 것이다. 본고에서 말하는 연상어는 기존의 연어(連語, Collocates)와 연상어(Associates)를 포함한 개념이다. 연상어의 빈도가 이렇게 높은 것은 사전이나 어휘의미망을 구축할 때 포함어(상위어, 하위어), 유의어, 반의어보다 연상어가 더 중요하고 유용한 정보라는 사실을 알려준다. 다시 말해서 연상어 자체의 사전 의미 정보도 중요하지만 연상어의 어휘 의미망 정보도 매우 중요하다는 뜻이다.

주지하듯이 연어란 텍스트에서 우연히 발생할 확률보다 훨씬 높은 공기 가능성을 가진 어휘들의 묶음을 말한다. 이들은 대개 2개에서 6개 정도까지의 어절 사이에서 나타난다. 또 대부분 문장 내부에서 발생하며, 문장과 문장 사이에서는 거의 나타나지 않는다. 아래 (17)은 신문기사에서 발췌한 연어의 예, (18)은 편지글에서 발췌한 연어의 예들이다.

(17) 신문기사(T10) - 문화일보

S5 : 이 배는 15일 오후 9시쯤 인천여객터미널을 출항해 제주로 향하는 도중 사고를 당했다.

S10 : 또 화상 골절 등 중상을 입은 부상자 7명은 전남 목포시 목포한국병원 등으로 이송돼 치료를 받고 있다.

S14 : 김 실장은 청와대 위기관리센터에 머물며 사고 및 구조 현황을 파악하는 등 필요한 조치를 취하고 있는 것으로 전해졌다.

S15 : 특히 세월호에는 제주도로 수학여행을 가던 안산 단원고 325명이 교사들과 타고 있었는데 한때 전원 구조됐다는 소식이 들리기도 했지만 이후 사실이 아닌 것으로 밝혀져 부모들이 애를 태웠다.

12) 특히 포함어, 즉 상위어-하위어 정보는 격식적인 글쓰기에서 많이 발견된다.

(18) 편지글(T3) - 오늘도 아침 6시쯤 되니
S2 : 습관인지, 나이를 먹어서 그런지 눈이 떠진다.
S4 : 너를 학교에 보내기 위해 아침 6기면 일어나 아침 식사를 차려 주었지.
S24 : 엄마, 아빠 마음속 깊은 곳에 절대 성빈이를 잊지 않고 새겨 놓을게.

한편, 연상어(AC)는 이러한 위치 구분이 없이, 문장 내부에서도, 문장과 문장 사이에서도 나타난다. 연상어 개념의 기본적인 아이디어는 Filmore & Baker(2001)[13]의 프레임 의미론(Frame Semantics)에서 제기한 것이다.[14] 그들에 의하면 어떤 개별 어휘의 의미는 외부 세계와 연결시키는 통로 역할을 하는 본질적인 세계 지식(essential world knowledge)이 없으면 이해될 수 없다는 것이다. 예를 들어, '팔다'라는 어휘의 의미를 상업적인 '상거래'의 맥락을 이해하지 못하면서 이해할 수는 없다는 것이다. 즉 팔다, 구매자, 구입자, 상품, 돈 등의 연상어가 필수적으로 동원될 수밖에 없다. 이렇게 연상어는 촉발어(trigger)와 연상어(associates) 사이에 일어나는 일종의 인지적 추론 과정이다. 예를 들어 '세월호 사건'이라는 촉발어는 같은 경험을 공유하고 있는 사람들에게 자연스럽게 나머지 연상어들을 떠오르게 한다. 아래는 필자가 동아일보 기사(Text 1)에서 뽑아 본 연상어들의 목록이다.

(19) 동아일보(T1). 촉발어('세월호') - 연상어
사망자, 해양경찰청(해경), 승객, 승무원, 탑승자, 사고, 실종자, 세월호, 사고원인, 쿵하는 소리, 여객선, 구조, 침몰, 단원고

한편 명사가 아니라 동사도 촉발어가 될 수 있다. 촉발어가 동사인 연상어의 예를 편지글에서 가져와 보면 아래 (20)과 같다. 편지글은 사건 자체보다는 사별한 자녀들에게 보내는 부모님들의 아픈 심리적 상태가 더 지배적인 틀(Frame)이라고 보아서, 심리적 상태를 나타내는 동사와 그 동사의 참여자들을 연상어로 처리했다. 그래서 그런지 명사 촉발어가 거의 발견되지 않았다.

(20) 편지글(T1. 촉발어(태어나다) - 연상어
태어나다, 아기, 딸, 유치원, 학교, 가족, 엄마, 아빠

그러나 본고의 연상어 연구는 아직 구체적인 방법론을 논의할 만큼 연구가 깊지 않다. 다만 예를 들어, '세월호 사건'을 소개하는 위키피디어의 사건 개요에 나오는 어휘들이 연상어가 될 수 있지 않을까 생각하고 있다. 물론 이것은 틀의미론(Frame Semantics)의 분석 방법론에 따라서 그 결과가 달라질 수 있다.

통계적 분석 결과의 다섯 번째 특징은 포함어(I)이다. 포함어(I)는 신문기사에서는 세 번째

13) Filmore, Charles J. & Collin F. Baker(2001), "Frame semantics for thext understanding", *Proceedings of WordNet and Other Lexical Resources Workshop*, NAACL 2001.

14) FrameNet은 틀의미론(frame semantics)에 기반하여 International Computer Science Institute (UC Berkeley)에서 만든 어휘의미망(FrameNet lexical database)이다. FrameNet은 예를 들어, "John sold a car to Mary" 라는 문장이 관점을 달리하면 "Mary bought a car from John"을 의미한다는 것을 밝히는 작업이다. 여기서 틀의미는 사건, 관계, 대상, 참여자 등(이들은 모두 개념구조 안의 성분들)을 기술하는 개념구조의 하나로 이해된다. 위키피디어 참조.
https://en.wikipedia.org/wiki/FrameNet

순서로 다수 출현했지만 편지글에서는 여섯 번째 순서로, 마지막 순서로 나타났다. 이것은 상위어나 하위어처럼 논리적인 어휘 관계를 나타내는 포함어가 서정적인 글쓰기에서 잘 선호되지 않았기 때문이 아닌가 생각해 본다.

실제로 거의 모든 글쓰기에서 포함어(I)가 유의어나 반의어보다 빈번하게 출현한다는 보고가 있다. 영어 텍스트를 대상으로 어휘 응집성을 다룬 Gómes(2018: 119)에서는 다자간 토론(격식)과 양자 간 전화 통화(비격식) 모두에서 포함어(I)가 유의어(S), 반의어(O)의 빈도를 앞선다고 보고하고 있다.

<표 4>에서 설명했듯이 포함어(I)는 상위어(GI)와 하위어(SI)를 모두 포함한다. 아래 포함어(21)은 '승객과 승무원' → '탑승자'로의 상위어(GI) 관계를 보여준다. 그리고 (22)는 '잠수원' → '해경', '해군', '경찰'로의 하위어(SI) 관계를 보여준다. 한편 (22, ②)은 포함어를 이용하여 글쓴이가 주제 전개 방식 중 파생형 주제 전개(the derived Theme pattern) 방식을 사용하고 있음을 보여 준다. 즉 '당국'이 '해경'으로 파생되었다.

(21) GI(상위어)

① T1 - 11, 12 (신문기사)

여객선에는 승객과 승무원 475명이 탑승중이었다. 탑승자 가운데는 수학여행에 나선 안산 단원고 학생 325명이 포함돼 있었으며, 이날 낮 12시께 제주도 여객터미널에 도착할 예정이었다.

② T3 17, 19 (편지글)

엄마, 아빠는 다른 부모님과 함께, 모든 국민이 너희를 기억하게 하여 다시는 이런 일이 일어나지 않게 하기 위해 활동하고 있다.

(22) SI(하위어)

① T1 - 5 (신문기사)

현재 사고 현장에는 잠수원 178명(해경118명·해군 42명·경찰 18명), 선박 72척(해경 55척·해군 17척), 항공기 18대(해경 14대·해군 4대)가 동원돼 구조 작업을 벌이고 있다.

② T8 - 1, 2 (신문기사)

진도 해상서 학생 325명 등 460명이 탄 여객선이 침수 중이라는 조난신고가 들어와 당국이 구조에 나서고 있다.

T8 -2 해경은 경비정을 급파해 구조 중인 것으로 알려졌다.

신문기사의 경우, 총 67개의 포함어(I) 중에서 상위어(GI)가 7개(7%), 하위어(SI)가 60개(89%)로 하위어가 압도적으로 많이 쓰인다. 그리고 편지글의 경우는, 총 6개의 포함어(R) 중에서 상위어(GI)가 3개(50%), 하위어(SI)가 3개(50%)로 양쪽이 동등하게 쓰였다.

통계분석 결과의 여섯 번째 특징으로 유의어(S)를 보자. 우리는 Lyons(1977, 1981), Martin(1999)에 따라서 유의어(S)를 유사-유의어(NS)와 명제-유의어(PS)로 나누어 계산을 보았다. 그들에 의하면 이 둘은 척도적(scalar) 관계이지 이산적(either-or) 관계가 아니다. 유사-유의어는 원형적 어휘 의미와 동일하거나 점진적 척도를 갖는다. 반면에 명제-유의어(PS)는 같은 유의어지만 표현적 의미나, 문체적 의미 다른 경우를 말한다.

아래 (23)에서 '눈시울'-'눈물' 등이 유사 유의어(NS)의 관계라면, 아래 (24)에서 '생일'-'미역국을 먹다' 등이 명제 유의어(PS)에 해당한다.

(23) 유사-유의어(NS)

① T4 – 1,3 (편지글)
눈시울 – 눈물

② T4 – 14 (편지글)
분하다 - 억울하다

③ T7 – 1, 15 (신문기사)
여객선 – 선박

④ T7 – 11, 13 (신문기사)
탑승하고 – 타고 있다

⑤ T7 – 10 (신문기사)
실종되다 – 생사가 확인되지 않다

⑥ T7 – 6, 10 (신문기사)
파악하지 못했다 – 확인되지 않았다

(24) 명제 유의어(PS)

① T5 – 6 (편지글)
생일 – 미역국을 먹다

② T5 – 6 (편지글)
생일 –배 속에 있다가 세상 밖으로 나온

③ T5 – 7, 10 (편지글)
보고 싶다 – 빨리 만나서 꼬옥 안아 보고 싶고, 얼굴 비비고 싶다

④ T7 – 6, 8 (신문기사)
숫자를 파악하다 – 집계하다

신문기사의 경우, 총 54개의 유의어(I) 중에서 유사-유의어(NS)가 46개(85.1%), 명제-유의어(PS)가 8개(14.8%)로 유사-유의어가 훨씬 많이 나타났다. 그리고 편지글의 경우는, 총 17개의 유의어(I) 중에서 유사-유의어(NS)가 5개(29.4%), 명제-유의어(PS)가 12개(70.5%)로 명제-유의어가 훨씬 더 많이 나타났다.

통계분석 결과의 7번째 특징으로 반의어(O)를 살펴보자. 반의어(O)는 다시 세 가지 하위구분을 갖는다. 상보 반의어(CpO)는 '여자-남자'처럼 이분적(either-or)인 절대값을 가진다. 따라서 한쪽을 부정하면 다른 쪽인 되는 관계이다. 그리고 등급 반의어(AO)는 '어렵다-쉽다'의 경우처럼 점진적이고 비양립적인 어휘 관계를 말한다. 그리고 방향 반의어(DO)는 반대 방향을 가리키거나, '입사-퇴사'처럼 반대 관계를 가리키는 응집성이다. 아래 (25)는 상보 반의어의 예를, (26)은 등급 반의어의 예를, (27)는 방향 반의어의 예를 세월호 텍스트에서 가져 온 것이다

(25) 상보 반의어 (CpO)

① T7 – 9 (신문기사)
해경 관계자는 "선체가 기울며 침대 등에서 떨어져 부상을 입었거나 전기공급이 끊겨 승객들이 어둠 속에서 우왕좌왕하다 배 밖으로 탈출하지 못 했을 우려가 높

다"고 밝혔다.

② T4 - 4 (신문기사)
사망자 중 승무원 박지영(22·여)씨, 안산 단원고 2학년 정차웅(17)군·권오천(17)군·임경빈(17)군 등 4명은 신원이 확인됐으나 나머지 2명은 아직 확인되지 않았다.

(26) 등급 반의어 (AO)

① T3 - 17 (편지글)
엄마, 아빠는 다른 부모님과 함께, 모든 국민이 너희를 기억하게 하여 다시는 이런 일이 일어나지 않게 하기 위해 활동하고 있다.

② T3 - 5 (신문기사)
사고 배에는 안산 단원고 2학년 학생 324명과 교사 10명이 탑승하고 있다.

(27) 방향 반의어(DO)

① T1 - 10 (신문기사)
앞서 16일 오전 8시58분께 전남 진도군 조도면 병풍동 북방 1.8마일 해상에서 인천에서 출발해 제주로 향하던 6647t급 여객선 세월호가 침수 중이라는 신고가 해경에 접수됐다.

② T1 - 12 (신문기사)
탑승자 가운데는 수학여행에 나선 안산 단원고 학생 325명이 포함돼 있었으며, 이날 낮 12시께 제주도 여객터미널에 도착할 예정이었다.

반의어는 글쓴이가 글을 쓸 때 현재의 주제를 계속 유지시키려는 주제 전개 방식에서 많이 사용된다. 다만 같은 주제를 유지시켜 주는 방식인 '반복(R) 응집어'에 비해서 상대적으로 기피된다는 점이 다르다.

신문기사의 경우, 총 23개의 반의어(O) 중에서 상보 반의어(CpO)가 9개(39.1%), 등급 반의어(AO)가 12개(52.1%), 방향 반의어(DO)가 2개(8%)로 나타났다. 그리고 편지글의 경우는, 총 9개의 반의어(O) 중에서 상보 반의어(CpO)가 4개(44%), 등급 반의어(AO)가 4개(44%), 방향 반의어(DO)가 1개(11%)가 나왔다. 신문기사와 편지글 모두에서 방향 반의어(DO)의 빈도가 매우 낮았다.

통계분석 결과의 마지막 특징으로 신문기사와 편지글에서의 어휘 응집성 거리를 살펴보자. 응집성 거리는 두 요소 사이의 거리를 기준으로 응집력을 구분하는 것이다. 본고에서 이것은 다시 동일 문장 내부 요소들 간의 거리(WS)과 문장과 문장 사이의 요소들 간의 거리(AS)로 나누어 계산해 보았다. 후자는 다시 인접거리 유대(immediate ties), 원거리 유대(remote ties), 매개거리 유대(mediated ties)로 하위분류된다.

먼저 아래 <표 7>과 <표 8>은 신문기사의 응집성 거리와 편지글의 응집성 거리를 종합적으로 비교해 보인 것이다. <표 7>에서 보면 신문기사나 편지글을 막론하고 어휘적 응집어는 동일 문장 내부(WS)보다는 문장과 문장 사이(AS)에서 많이 나타나는 걸 확인할 수 있다. 신문기사는 WS가 119개, 23%이고, AS가 397개, 76.9%에 해당한다. 그리고 편지글에서도 WS가 47개, 20.7%이고, AS가 179개, 79.2%에 해당한다. 대부분의 어휘적 응집어들이 문장과 문장 사이에서 일어난다는 사실은 문장을 넘어 텍스트와 텍스트의 종류에 대한 정보가 얼마나 중요한 지를 잘 알려준다.

<표 7> 어휘 응집성 거리 (신문기사와 편지 종합)

신문	ER	IR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
WT	10	1		8		2	9		4	44	41	119
AT	185	28	30	38	8	7	3	2	3	16	77	397
계	195	29	30	46	8	9	12	2	7	60	118	516

편지	ER	IR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
WT	10	2	2	1	5	0	3		1	1	22	47
AT	95	28	5	4	7	4	1	1	2	2	30	179
계	105	30	7	5	12	4	4	1	3	3	52	226

그리고 아래 <표 8>부터 <표 10>까지는 신문기사를 기사별로 정리하고 어휘적 거리를 계산한 것이다.

응집성 거리 (신문 기사)

<표 8> 신문 기사 별 어휘응집성 빈도

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
1	동아	WS				1		1	1			8		19
	일보	AS	26	6	3	4	3	2		1	1	2	16	58

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
2	서울	WS	1			1						2		4
	신문	AS	14	5	2	2					1	4	7	35

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
3	국민	WS	1						1			3		8
	일보	AS	10	1	3	3		1				1	6	25

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
4	한겨	WS	1			1						5	5	12
	레	AS	28	5	4	6	2			1		1	12	59

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
5	경향	WS	3	1							2	5	5	16
	신문	AS	14	1	2	3	1		1			1	4	27

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
6	조선	WS	1			1		1	2			5	3	13
	일보	AS	15	2	3	2			1			1	8	32

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
7	중앙	WS	1			2			2			2	5	12
	일보	AS	20	1	6	8	1	1	1			3	7	48

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
8	한국	WS	1						1			2	3	7
	일보	AS	11	2	1	2		1				2	5	24

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
9	세계	WS				1			1			4	3	3
	일보	AS	14		2	2		1					6	6

			ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
10	문화	WS	1			1			1		2	8	6	19
	일보	AS	33	5	4	6	1	1			1	1	12	64

<표 9> 신문기사 별 동일 문장 내부 어휘 응집성 빈도 (WS)

신문		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
동아일보	WS				1		1	1			8	8	
서울신문	WS	1			1						2		
국민일보	WS	1						1			3	3	
한겨레	WS	1			1						5	5	
경향신문	WS	3	1							2	5	5	
조선일보	WS	1			1		1	2			5	3	
중앙일보	WS	1			2			2			2	5	
한국일보	WS	1						1			2	3	
세계일보	WS				1			1			4	3	
문화일보	WS	1			1			1		2	8	6	
계		10	1		8		2	9		4	44	41	119

<표 10> 신문기사 별 문장과 문장 사이 어휘 응집성 빈도 (AS)

신문		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
동아일보	AS	26	6	3	4	3	2		1	1	2	10	
서울신문	AS	14	5	2	2					1	4	7	
국민일보	AS	10	1	3	3		1				1	6	
한겨레	AS	28	5	4	6	2			1		1	12	
경향신문	AS	14	1	2	3	1		1			1	4	
조선일보	AS	15	2	3	2			1			1	8	
중앙일보	AS	20	1	6	8	1	1	1			3	7	
한국일보	AS	11	2	1	2		1				2	5	
세계일보	AS	14		2	2		1					6	
문화일보	AS	33	5	4	6	1	1			1	1	12	
계		185	28	30	38	8	7	3	2	3	16	77	397

그리고 아래 <표 11>부터 <표 13>까지는 개별 편지글에 대한 어휘적 거리를 계산한 것이다.

응집성 거리 (편지글)

<표 11> 편지글 수별 어휘 응집성 빈도

		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
T1	WS	1		1		1						4	9
	AS	15	5			3		1		1		4	29

		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
T2	WS	4	1	1				1				6	13
	AS	17	3	2	1	1	1				2	3	30

		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
T3	WS	3	1			1		1		1		3	10
	AS	23	7			1	1					7	39

		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
T4	WS				1			1				6	8
	AS	11	3	1	1	1			1			4	22

		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
T5	WS	1				2					1		4
	AS	7	2	1	1	1	1			1		4	18

		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
T6	WS	1				1						1	3
	AS	22	8	1	1		1					8	41

<표 12> 동일 문장 내 어휘 응집성 빈도 (WS)

편지글		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
T1	WS	1		1		1						6	
T2	WS	4	1	1				1				6	
T3	WS	3	1			1		1		1		3	
T4	WS				1			1				6	
T5	WS	1				2					1		
T6	WS	1				1						1	
계		10	2	2	1	5	0	3	0	1	1	22	47

<표 13> 문장과 문장 사이 어휘 응집성 빈도 (AS)

편지글		ER	IIR	DIR	NS	PS	CpO	AO	DO	GI	SI	AC	계
T1	AS	15	5			3		1		1		4	
T2	AS	17	3	2	1	1	1				2	3	
T3	AS	23	7			1	1					7	
T4	AS	11	3	1	1	1			1			4	
T5	AS	7	2	1	1	1	1			1		4	
T6	AS	22	8	1	1		1					8	
계		95	28	5	4	7	4	1	1	2	2	30	179

끝으로 아래 <표 14> - <표 16>은 문장 과 문장 사이(WS)의 하위 구분으로 각각 인접거리 유대(IM), 매개거리 유대(MM), 원거리 유대(DM)의 빈도를 제시한 것이다. 신문기사와 편지글의 장르적 차이는 찾기 어려웠다. 다만 예상과 달리, 상대적으로 원거리 유대(DM)의 빈도가 조금 더 높다는 점이 특이하다. <표 14>는 신문과 편지의 어휘적 응집어 거리 종합, <표 15> 눈 신문기사의 어휘적 응집어 거리 (종합), <표 16>은 편지글의 어휘적 응집어 거리 (종합) 빈도를 보여 준다.

<표 14) 신문기사와 편지글의 어휘적 응집어 거리 (종합)

텍스트	IM	MM	DM	계
신문	126	110	161	397
편지글	48	65	66	179
합계	174	175	227	576

<표 15> 신문기사의 어휘적 응집어 거리 (종합)

신문	IM	MM	DM	계
동아일보	8	15	35	58
서울신문	10	11	14	35
국민일보	17	6	2	25
한겨레	14	17	28	59
경향신문	4	10	13	27
조선일보	18	8	6	32
중앙일보	9	11	28	48
한국일보	11	7	6	24
세계일보	9	7	9	25
문화일보	26	18	20	64
합계	126	110	161	397

<표 16> 편지들의 어휘적 응집어 거리 (종합)

편지글	IM	MM	DM	계
T1	9	10	10	29
T2	4	11	15	30
T3	12	11	16	39
T4	5	9	8	22
T5	8	5	5	18
T6	10	19	12	41
합계	48	65	66	179

5. 어휘의미망(WordNet)과 중의성 해소

앞장에서 다룬 어휘적 응집성의 종류들, 즉 반복어(R), 연상어(AC), 포함어(I), 유의어(S), 반의어(O) 등은 대부분 어휘의미망(WordNet)의 관련어 성분들이다. 그러므로 본고에서 논의한 어휘적 응집성의 종류와 그 정보가 잘 반영된 한국어 어휘의미망이 구축될 수 있다면 국어 텍스트의 이해와 교육은 물론 텍스트의 기계번역 등에서 중의성 문제를 해소하는 방법으로 사용될 수가 있다. 여기서는 그 가능성 두 가지를 소개하기로 한다.

먼저 김민호 · 권혁철(2011)에서는 한국어 어휘의미망(KorLex)을 이용한 비감독 어의 중의성 해소 방법을 제안한다. 일반적으로 감독 중의성 해소가 미감독 중의성 해소보다 성능이 좋게 나타나지만, 대규모 의미 부착 말뭉치가 필요하다는 단점이 있다. 그들에 의하면 중의성 어휘의 주변 문맥에 나타나는 공기(하는) 어휘들[15]은 중의성 어휘의 의미를 판단하는 중요한 단서가 된다. 김민호 · 권혁철(2011: 557)에서 이렇게 대규모 의미 부탁 말뭉치의 도움 없이도 중의성을 해소하는 절차는 다음과 같다.

(28) KorLex를 이용한 중의성 해소 절차

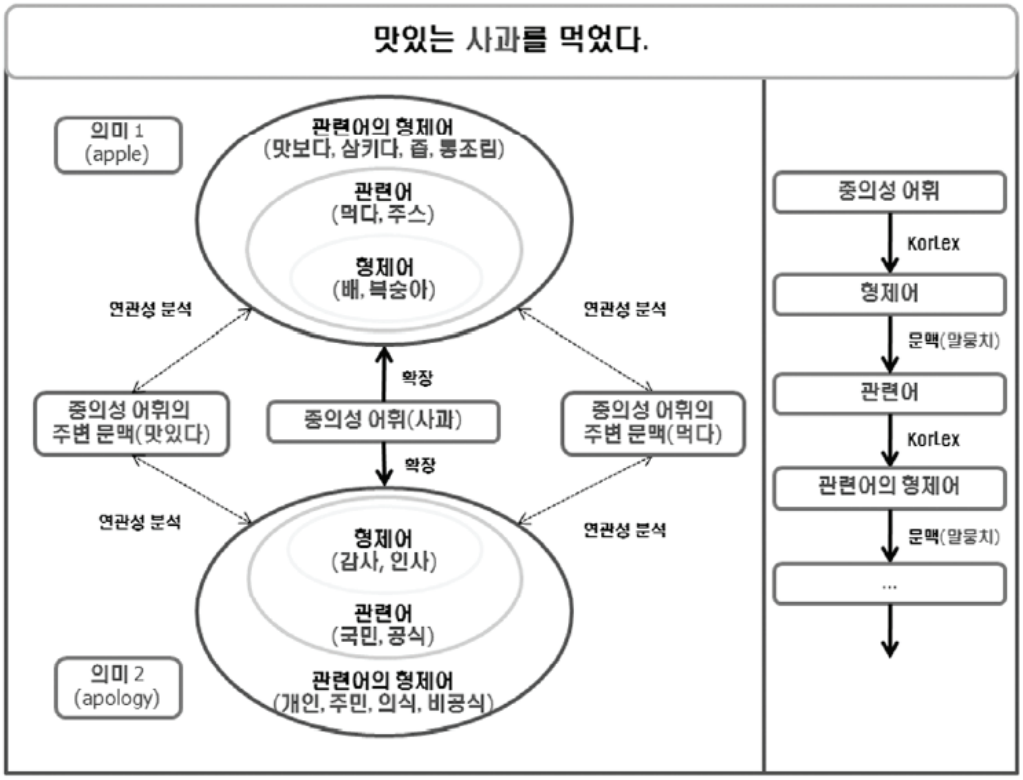
① 중의성 어휘의 관련어들을 한국어 어휘의미망(KorLex)에서 찾는다

② 중의성 어휘와 그 주변 공기 어휘 간의 공기 빈도를 대규모 말뭉치로부터 획득하여 카이 스퀘어(chi square) 값을 구한다.

③ 가장 큰 카이 스퀘어 값을 가지는 공기어와 일치하는 관련어가 중의성 어휘의 의미가 된다.

아래 <그림 1>은 중의성 어휘 '사과'(apple, apology)가 들어 있는 문장에서 KorLex를 이용한 중의성 해소 절차가 어떻게 진행 되는지를 보여준다. 간단히 설명하면, (28, ①)은 KorLex에서 중의성 어휘인 '사과'의 형제어를 먼저 찾고, 관련어를 찾고, 마지막으로 관련어의 형제어를 찾는다. (28, ②)는 중의성 어휘('사과')와 주변 문맥('맛있다', '먹다') 어휘와의 공기 빈도를 대규모 말뭉치로부터 카이 스퀘어 값을 구한다. (28, ③)은 가장 큰 카이 스퀘어 값을 가지는 '사과'(apple)이 중의성 어휘의 최종 의미가 된다.

15) 이것이 본고에서 언급한 어휘적 응집어들이다.



한편, 연상어(AC)를 이용하여 중의성을 해소할 수 있는 방법도 있다. 예를 들어 남경완 · 이동혁(2004)에서는 '틀의미론'(frame semantics)에 기반한 중의성 해소 방안을 소개하고 있다. 그들은 틀의미론에 기반하여 동사 '사다'의 의미분절 양상을 다음과 같은 절차를 이용하여 설명하고 있다.

(29) ① 실제 말뭉치 용례를 이용하여 '사다'의 각 단의(單意, seme)를 분석한다.
② '상거래'의 틀, 틀 요소, 틀 구조를 이용하여 '사다'의 의미 분절을 결정한다.
③ 그 결과 총 6개의 단의 의미가 결정된다.
F1-1 : BUYER가 MONEY를 주고 SELLER에게서 어떤 GOODS_01_Object를 넘겨받다. ¶딸기를 사다.
F1-2 : BUYER가 MONEY를 주고 SELLER에게 어떤 GOODS_02_Effort를 얻다. ¶노동력을 사다.
F1-3 : BUYER가 MONEY를 주고 SELLER에게 GOODS_03_Sex를 구하다. ¶여자를 사다.
F2 : DONOR가 THEME을 RECIPIENT에게 제공하다. ¶저녁을 사다.
F3 : RECEIVER의 TOPIC이 GIVER의 MINDS를 받다. ¶주위의 부러움을 사다.
F4 : VALUER가 VALUE를 평가하다. ¶그의 용기를 높이 사다.

본고의 입장에서 보면 위 (29)의 의미 분절 6개 중(F1-F4)에서 틀요소에 해당하는 BUYER, MONEY, SELLER, GOODS, DONOR, THEME, RECEIVER, TOPIC, VALUER, VALUE가 모

두 ‘사다’의 연상어(AC가 된다.

끝으로 국립국어원에서 구축한 ‘국립국어원 어휘 관계 자료(NIKLex)[16]를 소개한다. 이것은 인터넷 사전 서비스 중인 <우리말샘>에 등록된 비슷한말, 반대말, 상위어, 하위어 어휘 쌍을 대상으로 어휘 관계 강도를 5점 척도로 총 5만 명이 평가(집중도 문항을 포함하여 응답자별 1,005개의 어휘 쌍 평가)한 후 그 결과를 통계 정보로 제시한 것이다. 여기에는 어휘 관계 기초 자료 20만 쌍(비슷한말 60,000쌍, 반대말 10,000쌍, 상위어 70,000쌍, 하위어 60,000쌍)의 분량을 갖추고 있다.

6. 결론

응집성은 어떤 종류의 언어 자원이 쓰였는가에 따라서 문법적 응집성과 어휘적 응집성으로 구별 된다. 본고는 지금까지 세월호 침몰 사고에 대한 10편의 신문기사와 세월호 유가족이 쓴 7편의 편지글을 대상으로 하여, 11가지 종류의 어휘적 응집성과 5가지 종류의 응집성 거리의 출현 양상을 살펴보았다. 분석 결과 다음과 같은 사실들을 알게 되었다.

첫째, 신문기사의 어휘 응집성 빈도가 편지글의 그것보다 2.3배 높게 나왔다. 이것은 직업적인 글이나, 다수 독자를 대상으로 하는 글일수록 어휘적 응집도가 높다는 것을 보여준다. 둘째, 어휘적 응집성의 유형별 출현 빈도가 신문기사와 편지글에서 큰 차이를 보이지 않았다. 그 순서는 반복어(R) > 연상어(AC) > 포함어(I) >유의어(S) > 반의어(O)이다. 셋째, 신문기사와 편지글 모두에서 반복어(R)와 연상어(AC)가 가장 강력한 어휘적 응집어임을 알 수 있었다. 넷째, 연상어(AC)는 반복어(R)을 제외하면 사실상 가장 중요한 어휘의미망 자원이다. 다섯째, 포함어(I), 즉 상하위어는 신문기사에서는 유의어(S), 반의어(O)보다 높게 나타났지만, 편지글에서는 후자보다 낮게 나왔다. 그러나 대부분의 격식적인 글쓰기에서 포함어(I)가 유의어(S), 반의서(O)보다 빈도가 높다. 여섯째, 신문기사나 편지글 모두에서 유의어(S)의 빈도가 반의어(O)를 상회하였다. 일곱째, 어휘 응집성의 거리를 분석해 본 결과, 신문기사나 편지글을 막론하고 어휘적 응집성 거리의 빈도는 동일 문장 내부(WS, with-in-sentence)보다 문장과 문장 사이(AS, across-sentence)에서 높게 나타났다. 이것은 응집성을 파악할 때 문장 내부의 어휘 정보보다 문장 밖 어휘 정보들이 더 중요하다는 것을 알려준다.

그리고 본고에서 논의한 어휘적 응집성의 종류와 그 정보가 잘 반영된, 특히 연상어(AS) 정보가 잘 반영된 한국어 어휘의미망이 구축될 수 있다면 국어 텍스트의 중의성 문제를 해결 할 수 있다는 가능성을 언급하였다. 이를 위한 예시로 김민호 · 권혁철(2011)에서 제시한 KorLex를 이용한 중의성 해소 절차와 남경완 · 이동혁(2004)에서 제시한 틀의미론(Frame Semancis)을 이용한 동사 ’사다‘의 의미분절 절차를 소개했다.

끝으로 한국어 어휘의미망의 미래를 생각해 보면, 아직도 이상은 높고 현실은 아늑하여 지난한 고행의 길만 남아 있음을 깨닫게 된다. 미력한 힘이나마 도움이 되길 바랄 뿐이다.

16) (버전 1.0) 2020.8.25.일자. 관련사업 : 한국어 정보 처리를 위한 어휘 관계 기초 자료 구축(2019) https:corpus.korean.go.kr

참고문헌

김서형 · 유혜원 · 이동혁 · 이유진 · 정연주(역), 「체계기능언어학의 이해」, 역학. (원저) Suzanne Eggins(2004), An Introduction to systemic functional linguistics, (2nd ed.) London: Continuum.

남경완 · 이동혁(2004), “틀의미론으로 분석한 ‘사다’와 ‘팔다’의 의미 분절 양상”, 「한국어학」 29-1, 1-24p.

안경화(2001), “구어체 텍스트의 응결 장치 연구: 토론 텍스트를 중심으로”, 「한국어교육」 12(2), 국제한국어교육학회, 137-157.

유민애(2015), “장르-텍스트 기반 문법 교육 내용 연구: 논리적 응결 장치를 중심으로”, 「한국언어문화학」 12(1), 한국언어문화학회, 139-166.

이관규 · 김서경 · 노하늘 · 성수진 · 신희성 · 유상미 · 이현주 · 정려란 · 정지현 · 정혜현(2021), 「체계기능언어학개관」, 사회평론아카데미.

정희모(2019), “한국어 교육과 결속성(cohesion) 및 응집성(coherence)의 문제”, 「리터러시연구」 10(4), 한국리터러시학회, 89-123.

Baker, M.(1992), *In Other Words*. London: Routledge.

Dontcheva-Navratilova, O. & Povolná, R.(2009), *Coherence and Cohesion in Spoken and Written Discourse*. Cambridge: Scholars Press.

Fillmore, Ch. J. & Baker, C. F.(2001), Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, NAACL. Pittsburgh, June. Gzipped Postscript (112 KB) PDF (189 KB).

Flowerdew, J.(2010), Use of signalling nouns across L1 and L2 writer corpura. *International Journal of Corpus Linguistics* 1591): 36-55.

FrameNet https://en.wikipedia.org/wiki/FrameNet (Wikipedia)

FrameNet Project Offical Website, https://framenet.icsi.berkeley.edu/fndrupal/

Fraser, B.(2005), Towards a Theory of Discourse Markers. Retrieved from http://people.bu.edu.bfraser/

Gómez González, M. Á.(2018), Lexical cohesion revisited: A combined corpus and systemic-functional analysis, Quaderns de Filologia: Estudis Lingüístics XXIII: 105-127.

Halliday, M. A. K. & Hasan, R.(1976), *Cohesion in English*. London: Longman.

Halliday, M. A. K. & Hasan, R.(1985), *Language, Context, and Text: Aspect of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.

Halliday, M. A. K. & Matthiessen, M. I. M.(2014), An Introduction to Functional Grammar (4th ed.). London: Edward Arnold.

Morenno, A.(2003), The role of cohesive devices as textual constraints on relevance: A discourse-as-process view. *International Journal of English* Studies 3(1): 111-165.

Moreno, A.(2003), The role of cohesive devices as textual constraints on relevance: A discourse-as-process view. *International Journal of English Studies* 3(1):

111-165.

Ochs Keenan, E.(1977), Making it last: repetition in children's discourse. In Ervin-Tripp, S. & Mitchell-Kernan, C. (ed.) *Child Discourse*. New York: Academic Press, 26-39.

Sanders, T. & Pander Maat, H.(2006) *Cohesion and Coherence: Linguistics Approaches. Encyclopedia of Language and Linguistics* (2nd ed.). Elsevier: London.

Taboada, M.(2004), *Building Coherence and Cohesion: Task-Oriented Dialogue in English and Spanish*. Amsterdam / Philadelphia: John Benjamins.

Tadros, A.(1985), *Prediction in Text*. (Discourse Analysis Monograph 10). English Language Research, Birmingham: University of Bermingham.

〈어휘적 응집성과 한국어 어휘의미망(Lexical Cohesion and Korean WordNet)-체계기능언어학의 관점(from Systemic Functional Linguistics point of view)〉에 대한 토론문

이동혁(부산교육대학교)

이 글은 체계기능언어학에 바탕을 둔 어휘 응집성을 소개하고, 어휘 응집성이 신문 기사와 편지글이라는 별개의 텍스트 장르에서 어떻게 차별적으로 나타나는지를 통계적으로 분석하였으며, 어휘 응집성의 종류와 특성을 반영한 한국어 어휘의미망을 구축함으로써 어휘의 중의성 문제를 해소할 수 있는 가능성을 탐색해 본 것으로 이해하였습니다.
무엇보다 이 연구는 텍스트 장르별 어휘 응집성의 차이를 통계로 실증하였으며, 나아가 실증된 어휘 응집성의 특성을 한국어 어휘의미망 구축에 반영하여 어휘의 중의성을 해소하는 데 응용할 수 있다는 점을 제시한 것이 주목을 끌었습니다.
아래 몇 가지 질문과 의견을 제시함으로써 이 글의 논의에 발전적인 영향을 미치기를 기대해 봅니다.
이 글의 논의를 단계별로 정리해 보겠습니다.
1. 가설
1) '장르에 따라서' 응집성 장치(이 글에서는 '어휘 장치'로만 한정함 → 어휘적 응집성)의 사용 양상(주로 분포)이 다를 것이다.
2) '장르에 따라서' 응집성 거리의 양상이 다를 것이다.
2. 분석 방법
카이제곱값에 의한 통계적 분석
3. 분석 결과
가설 1)에 대하여: 장르에 따라서 응집성 장치의 사용 양상이 달랐다.
가설 2)에 대하여: 장르에 따라서 응집성 거리의 양상이 달랐다.
4. 해석
5. 응용
응용 대상: 어휘의미망을 활용한 중의성 해소
방법: 1) 중의성 어휘의 주변 문맥에 나타나는 공기어 활용, 2) 연상어 활용

2와 관련하여: 카이제곱값 분석의 용도를 설명해 주시기 바랍니다.
4와 관련하여: 이 글에서는 분석 결과에 대한 해석이 부족한 듯합니다. 이는 텍스트 장르에 대한 면밀한 분석이 선행되지 않았기 때문이라고 봅니다. 글에서 제시한 텍스트 장르에 관한 설명은
"'세월호 사건'을 공통의 주제로 하여, 불특정 다수 독자를 대상으로 쓴 신문기사와 사별한 자녀들에게 보내는 개인적인 편지글이 보이는 어휘 응집성의 차이를 밝혀보려는 것이 목적이다."
"편지글에서 어휘적 응집어들이 상대적으로 적은 것은 테너(참여자) 변수, 즉 글쓴이의 사적이고 정서적인 감정 변화가 텍스트의 정보 전달적 요소보다 더 중요하게 작용한 것으로 보인다.

또 생략 어휘들이 상대적으로 많은 것도 영향이 있어 보인다. 이에 반해 신문 기사문은 글쓴이가 직업적인 전문가들이고 다수 독자를 대상으로 쓴 계획된 글이기 때문에 이러한 결과를 얻었지 않나 생각한다."
이 두 설명뿐입니다. 텍스트 장르의 특성이 구체적으로 어떠하기에 응집성 장치의 사용 양상이 다를 수 있는지, 그리고 응집성 거리의 양상이 다를 수 있는지를 설명해 주셔야 한다고 생각합니다.
5와 관련하여: 1) 중의성 어휘의 주변 문맥에 나타나는 공기어 활용하고, 2) 연상어를 활용함으로써 중의성을 해소할 수 있다는 제안은 이미 여러 연구에서 해 왔습니다. 이 연구의 제안이 다른 연구와 차별적이기 위해서는 2~4장에서 논의한 '장르별' 어휘 응집성의 결과를 중의성 해소 방안에 활용했어야 하는데, 이를 해 내지 못한 것 같습니다.

기타 의견
본문에서 '연상어'를
"연상어(AC)는 특정 의미 프레임(Frame Semantics)에서 서로 연관된 어휘들 간의 관계이다."
"본고에서 말하는 연상어는 기존의 연어(連語, Collocates)와 연상어(Associates)를 포함한 개념이다."
와 같이 설명하였습니다. 그런데 두 설명의 연상어는 개념과 범위가 일관되지 못한 것으로 보입니다.

(19), (20)이 예가 되는 연상어의 추출 근거가 무엇인지 궁금합니다. "다만 예를 들어, '세월호 사건'을 소개하는 위키피디어의 사건 개요에 나오는 어휘들이 연상어가 될 수 있지 않을까 생각하고 있다."라고 하셨는데, 이는 과학적인 연상어 추출 방법은 아니라고 생각합니다.

"등급 반의어(AO)는 '어렵다-쉽다'의 경우처럼 점진적이고 비양립적인 어휘 관계를 말한다."라고 정의한 것에 비추어 보면, (26)의 예가 등급 반의어라고 보기 어렵습니다.

중국의 다국어 어휘 의미망 구축과 활용 - CCD와 MCD를 중심으로

강병규(서강대)

〈목차〉

1. 서론

본고는 WordNet에 기반을 둔 중국어 어휘의미망 'CCD(Chinese Concept Dictionary)'를 소개하고 이를 이용하여 중국어-영어-한국어-일본어 다국어 어휘의미망 구축 방법을 모색하는데 그 목적이 있다. 어휘의미망은 어휘의 의미 체계를 네트워크처럼 연결한 계층적 집합을 의미한다. 한 언어에서 사용되는 어휘는 일종의 개념적 네트워크라고 할 수 있다. 그런데 이 개념적 네트워크는 각각의 연결 노드가 존재하고 상위-하위, 동의-반의의 관계를 가진다. 한 언어에 존재하는 어휘 중에 상위 개념을 나타내는 것은 무엇인가? 그리고 그 상위 개념은 대개 몇 가지로 나누어지는가? 또한 그 상위 개념의 하위 개념은 몇 가지로 분류될 수 있는가? 각각의 어휘는 어떠한 유의 관계, 반의 관계를 가지는가? 이러한 질문에 대한 해답을 찾는 것이 어휘의미망 연구의 핵심이다. 어휘의미론적 관점에서 출발한 어휘의미망은 전산언어학 분야에서 어휘 의미를 표현하는 방법론으로 채택되면서 큰 주목을 받아 왔다. 자연언어(natural language:혹은 인간의 언어)를 이해하는 시스템 개발을 위해서는 대량의 언어 지식이 필요한데 그중에서 어휘 의미 정보를 담고 있는 언어자료가 중요한 역할을 하기 때문이다.

본고에서는 어휘의미망과 관련하여 다음의 몇가지 내용을 중심으로 논의를 진행하고자 한다. 첫째, 다국어 어휘 의미망 구축의 기초가 되는 온톨로지 형태의 어휘 자원에 대해 알아본다. 둘째, 다국어 어휘 의미망의 구축 방법을 논의한다. 본문에서는 특히 범언어적 상위 온톨로지 자원으로서 SUMO와, 영어의 WordNet, 중국어의 CCD(Chinese Concept Dictionary) 등의 언어자원을 결합하는 방법에 대해 주로 논의할 것이다. SUMO(The Suggested Upper Merged Ontology)는 세계적으로 통용되는 상위 온톨로지로서 총 1,000 여개의 개념노드가 설정되어 있다. 중국어 어휘 의미망인 CCD는 중국 북경대학 전산언어학연구소에서 구축한 어휘 의미망이다. 이 자원은 영어의 WordNet의 체계 위에 중국어의 개념 체계를 연결시킨 것으로서 영어-중국어 어휘 의미망(개념망)이라고 할 수 있다. CCD는 WordNet의 형식을 따르면서도 중국어의 개념 체계의 특징이 반영된 어휘의미망 DB이다. 즉, 단순한 중국어판 WordNet이 아니라 중국어의 특징이 반영된 자료이다. 셋째, 영어-중국어 어휘 의미망에 한국어-일본어-베트남어 대응어를 연결하는 방법을 논의한다. 다국어 어휘 의미망 구축에서 중요한 것은 범언어적 상위 온톨로지(upper ontology)의 설정, 어휘 개념 관계의 정확한 묘사, 기존

어휘 의미망 자원의 효율적인 통합인데 본고에서는 이 부분을 중점적으로 논의하고자 한다.

2. 다국어 어휘 의미망 구축을 위한 기초 자원

다국어 어휘 의미망 구축에 가장 필요한 자원은 개별 언어를 중심으로 구축된 단일어 어휘 의미망이라고 할 수 있다. 물론 기존의 일부 어휘 의미망에서는 두 언어 간의 대응관계가 기술되어 있기도 하다. 예를 들어 영어-중국어, 일본어-한국어, 영어-한국어 등의 이중 언어 자원이 그러하다. 두말할 나위 없이 이들은 다국어 어휘 의미망 구축에 필요한 가장 기초적인 자원이다. 이에 본 장에서는 이미 구축된 대표적인 언어 자원 중에서 온톨로지 형태의 데이터베이스 몇 가지를 소개하고자 한다.

2.1 영어의 어휘 의미망-워드넷(WordNet)

워드넷(WordNet)은 미국 프린스턴(Princeton) 대학의 인지과학 연구실(Cognitive Science Laboratory)에서 구축한 대규모의 어휘 자원으로서 현재 3.1판이 나와 있다. 심리언어학자 George A. Miller의 책임 하에 구축된 워드넷(WordNet)은 오늘날 전 세계에서 가장 영향력 있는 영어 어휘 의미망이라 할 수 있다. 특히 워드넷의 구조는 어휘 개념 관계를 정밀하게 기술한 형태로서 전산언어학과 자연언어처리에 다양하게 응용되고 있다. 이 전자사전 안에는 약 155,327개의 단어(word form)와 117,597개의 동의어 집합인 신셋(Synset)이 수록되어 있으며 검색이 가능하다.

WordNet은 영어의 명사, 동사, 형용사, 부사의 범주에 속하는 어휘를 수록하고 있다. 그리고 WordNet 안에는 이 4가지 범주에 속하는 어휘들이 나타내는 개념 부류가 계층적이고 다원적으로 연결되어 있다. 여기서 주의할 것은 '신셋(Synset)'이라는 개념이다. WordNet에서는 의미의 기본 단위를 동의어 집합이라고 할 수 있는 '신셋(Synset)'으로 삼아 모든 유의어들을 하나의 집합으로 분류하였다. 그리고 이것을 기초로 하여 어휘간의 의미 관계가 표현되어 있다[1].

여기서 우리는 WordNet 상의 의미 관계가 심리언어학적 이론에 기초한다는 점에 주목할 필요가 있다. WordNet의 기본 체계는 인간이 어휘를 기억하는 방식을 연구하고자 심리어휘론 (psycholexicology)에서 제기된 어휘 이론을 바탕으로 하고 있다. 즉 WordNet은 원래 심리언어학적인 관점에서 영어 모국어 화자들의 어휘 지식을 모델링하기 위해 구축되었다. 그리고 이 심리어휘론의 타당성을 검증하기 위해 일부 어휘에 한정하지 않고 일상생활에 사용되는 모든 어휘에 대해 확대 적용하여 거대한 어휘 의미 네트워크를 구축한 것이다.

WordNet의 영향력은 심리언어학의 영역에 국한되지 않았다. 그것은 전산언어학, 자연언어처리, 사전편찬학 등 다양한 분야에 영향을 미쳤다. 그리하여 애초에 영어 화자의 어휘 지식을 모델링하는 것을 목표로 구축된 WordNet이 이제는 어휘 의미망의 대명사로 자리를 잡기에 이르렀다. 실제로 많은 학자들이 어휘 의미망을 종종 '워드넷'이라는 말로 대신하기도 한다. 뿐만 아니라 WordNet의 체계는 다른 언어의 어휘 의미망 구축에도 영향을 주었다. 그 중에 가장 대표적인 것이 유로워드넷(EuroWordNet)이다. 네델란드 Amsterdam 대학 교수인 Vossen(1998)을 책임자로 한 유로워드넷 구축 프로젝트는 유럽의 주요 언어를 WordNet의

1) 예를 들어, WordNet Synset 체계에서 어휘 개념의 기본 관계는 동의어(synonymy), 반의어(antonymy), 상위/하위(hypernymy/hyponymy), 전체/부분(holonymy/meronymy), 속성(attribute), 내포(entailment) 관계 등으로 표현된다

체계에 기초하여 연결시킨 자료이다. 현재 유로워드넷(EuroWordNet)은 EU의 8개 국어인 영어, 이탈리아어, 스페인어, 네덜란드어, 독일어, 프랑스어, 에스토니아어, 체코어로 구성되어 있다.

2.2 중국어의 어휘 의미망-CCD(Chinese Concept Dictionary)

WordNet은 사실상 전 세계 어휘 의미망의 표준과 같은 역할을 하는데 이 체제를 수용한 중국어 어휘 의미망으로 들 수 있는 것이 CCD(Chinese Concept Dictionary)이다[2]. 중국어 어휘 의미망인 CCD는 중국 北京大學 計算語言學研究所에서 구축한 어휘 의미망이다(劉揚, 2003). 이 자원은 영어의 WordNet의 체계 위에서 중국어의 개념 체계를 묘사한 것으로서 영어-중국어 어휘 의미망(개념망)이라고 할 수 있다. 다음은 CCD에서 채용한 어휘 의미의 개념 관계를 정리한 것이다.

표2 CCD의 어휘 의미 관계(劉揚, 2003)

명사 개념 관계	동사 개념 관계	형용사 개념 관계	부사 개념관계
反義(Antonymy)	反義(Antonymy)	反義(Antonymy)	反義(Antonymy)
下位(Hyponymy)	下位(Hyponymy)	近義(Similar)	導出形式 (Derived Form)
上位(Hypernymy)	上位(Hypernymy)	關系性形容詞 (Relational Adj.)	
部分(Meronymy)	蘊涵(Entailment)	又見(Also See)	
整體(Holonymy)	致使(Cause)	屬性(Attribute)	
屬性(Attribute)	又見(Also See)		

위의 표에서 보이듯이 CCD는 기본적으로 WordNet의 의미관계를 그대로 계승하고 있다. CCD 프로젝트는 2000년 9월에 시작되었다. 처음에는 1000개의 중국어 개념을 설정하여 WordNet 1.6 판에 연결시키는 작업부터 시작되었다. 그리고 점점 그 자료의 규모가 커져 현재에는 99,642 개의 개념 노드 (명사 66,025 개, 동사 12,127개, 형용사 17,915개, 부사 3575개)와 107여 개의 의미 관계를 기술한 데이터베이스로 확장되었다[3]. 이 자료들은 모두 MS Access 데이터베이스 형식으로 저장되어 각 어휘들의 개념번호와 의미 관계들이 자동으로 검색된다.

중국어 어휘 의미망 CCD가 가지는 의의는 대체로 다음의 두 가지로 요약된다. 첫째, 기존의 어휘 의미망 자료인《同義詞詞林》이나 HowNet[4]과 달리 WordNet의 체제를 계승하였다는 것이다. 즉, 중국어 어휘들의 개념이 영어의 WordNet 체계와 대응이 됨으로서 영어와 유럽 언어들과도 자동으로 연결이 가능하다는 점이다. 둘째, 기본적인 틀은 WordNet을 따르면서도 중국어의 개념 체계의 특징을 최대한 반영하였다는 점이다. 즉, 단순한 중국어판 WordNet이 아니라 중국어의 특징이 반영된 자료라는 점이다.

---

2) 이밖에도 臺灣 中央研究院의 Sinica Bow 어휘 의미망을 들 수 있다. 언어학자 黃居仁교수의 주도하에 구축된 Sinica Bow 역시 기본적으로 WordNet 체계를 근간으로 하고 있다. 다만 여기에서는 그 체계가 비슷하여 일단 베이징대학의 CCD를 소개하는 것으로 그친다. 자세한 것은 대만 중앙연구원 Sinica Bow 사이트를 참고하기 바람.

3) 더 자세한 것은 劉揚(2003)을 참고하기 바람.

4) HowNet에 관한 자료는 http://www.keenage.com에서 자세히 소개되어 있다.

2.3 한국어의 어휘 의미망-코어넷(CoreNet)

본고에서는 한국어 어휘의미망으로 코어넷(CoreNet)을 사용하고자 한다. 코어넷(CoreNet)은 카이스트에서 구축된 개념 체계 기반 어휘 의미망이다.

CoreNet의 가장 큰 특징은 코퍼스를 기반으로 하고 있다는 점이다. 즉, 한국어 코퍼스에서 기본 개념이 되는 어휘들을 뽑아서 그것을 다시 개념체계에 맞추어 정리했다는 점이다. 실제로 CoreNet은 일상생활에서 자주 사용되는 어휘를 중심으로 구축되었다. 예를 들어 CoreNet의 최상위어도 일상에서 자주 사용되는 어휘들이 사용되었다. "공간, 과정, 힘, 관계, 물질, 속성" 등이 그러하다. 이처럼 CoreNet은 다른 어휘 의미망과는 달리 코퍼스에 기초하여 구축되었으므로 한국어 화자의 어휘 지식 모델링을 올바른 방향으로 이끌어 주는데 중요한 역할을 할 수 있다. 한편, 한국어 코퍼스에서 뽑은 어휘들을 분류함에 있어 근간이 되는 개념 체계는 일본어의《NTT 어휘대계》이다. CoreNet은 이 NTT 체계에 기초하여 총 2,954 개의 기본 개념으로 분류되어 있고, 각 개념의 하위 노드에 23,823개의 한국어 명사, 1,757 개의 동사, 804 개의 형용사가 연결되어 있다.

CoreNet의 또 한가지 특징은 다국어 체계를 지향하고 있다는 점이다. CoreNet은 기본적으로 일본어 NTT 개념 체계를 따르고 있기 때문에 일본어와 대응이 되는 체계이다. 그리고, 한국어 어휘에 기초하여 대응되는 중국어 대역어가 추가되어 있다. 중국어 대역어를 찾는 과정에서 주로 사용된 중국어 데이터베이스는 北京大學의《現代漢語語法信息詞典》이다. 이를 통해 현재 2만 여 개의 중국어 명사 대역어가 있고, 288개의 동사, 80개의 형용사에 대한 의미 기술이 이루어져 있다. 그러나 중국어와 한국어 연결 작업은 아직 만족할 만한 수준에 이르지 않고 있는 것이 사실이다. 누락된 어휘들이 상당수 존재하고, 추가하고 수정해야 할 것들이 많이 남아 있다.

2.4 일본어의 어휘 의미망-EDR 전자사전과 NTT 어휘 대계

일본의 대표적인 어휘 의미망 자원으로는 일본 국립국어연구소에서 발행한《분류어휘표(分類語彙表)》(2004)와 EDR 전자사전, 그리고 NTT를 들 수 있다. 전통적으로 사전 편찬에 많은 심혈을 기울여 온 일본 언어학계에서는 이러한 독자적인 어휘 의미 분류 사전들이 이미 90년대 이전부터 많이 사용되어 왔다. 이러한 사전들의 가장 큰 특징은 대체로 시소러스(thesaurus)형 온톨로지라는 점이다. 예를 들어《분류어휘표》만 보더라도 상하위 개념이 위계적으로 비교적 엄격하게 연결되어 있고 상위 레벨을 구성하는 개념 부분들은 존재론적 범주로 보았을 때 기초 개념들로 구성되어 있다. 이 점이 영어의 WordNet과 다른 점이다. WordNet은 어휘 개념들의 동의어 반의어 관계를 묘사하고 있기는 하지만 상하위 관계의 층차가 엄격하지 않은 편이다. 그러므로 분류의 측면에서 보았을 때는 일본어의《어휘분류대계》나 NTT 개념 사전이 훨씬 전형적인 온톨로지 자원에 가깝다고 할 수 있다[5].

5) 본문의 논의는 주로 영어-중국어-한국어 자원에 초점이 맞추어져 있기 때문에 일본어 자원에 대한 소개는 간략히 하기로 한다. 또한 일본어에 대한 분석은 필자의 능력 밖의 일이라 자세히 논의하지 않겠다. 다만 참고로 현재 일본어 연결 작업은 北京大學과 일본 와세다대학 간의 공동 연구로 진행되고 있다. 한국 국내에서 일본어 어휘 의미망에 대한 연구로 참고할 수 있는 자료도 적지 않은데 그 중의 한 가지로 한유석·설근수(2004)의 연구 결과를 들 수 있다.

3. 중국어 어휘 의미망의 구축 방법

2장에서 우리는 영어-중국어-한국어-일본어 어휘 의미망을 구축하기 위해 활용할 수 있는 기초 자원에 대해서 살펴보았다. 이 장에서는 이러한 기존의 온톨로지 형 어휘 의미망의 기초 위에서 어떻게 다국어 어휘 의미망을 효과적으로 구축할 것인가를 논하고자 한다.
사실 한-중-일 다국어 의미망 구축 프로젝트는 이전부터 한국과 중국의 전산언어학계에서 구상해 온 과제 중의 하나이다. 그리고 그 시작은 카이스트 최기선 교수 연구팀에서 CoreNet이라는 이름으로 이루어졌다. 그리고 중국 北京大學 計算語言學研究所에서도 'MCD (Multi-lingual Concept Dictionary)' 프로젝트를 시작하였다. 이 두 가지 사업은 모두 동북아시아의 주요 언어인 중국어, 한국어, 일본어와 영어의 어휘 의미망을 연결하고자 하는 필요에서 나온 것이다. 본문에서 논하고자 하는 다국어 어휘 구축론은 바로 이러한 연구 과제들과 밀접한 관계를 가진다[6]. CoreNet 계획은 이미 국내에서 발표된 연구 결과물들이 있기 때문에 생략하고 본문에서는 필자가 참여했던 北京大學의 MCD 어휘 의미망의 관점에서 몇 가지 원칙과 구축 과정을 소개하도록 하겠다.

3.1 기본 원칙 및 방법

다국어 어휘 의미망 구축의 핵심이 되는 것은 범언어적 상위 온톨로지(upper ontology)의 설정, 어휘 개념 관계의 기술 방식, 기존의 어휘 의미망 자원의 통합 방법 등으로 요약할 수 있다. 이 사항들에 대해 간단히 설명하면 다음과 같다.

(1) 범언어적 상위 온톨로지의 설정 : SUMO
2.1에서 언급했듯이 WordNet의 단점은 상위 개념들이 체계적으로 설정되지 못했다는 점이다. 이는 어휘 분류가 유의어 집합인 Synset을 토대로 상향식(Bottom-Up)방식으로 이루어졌기 때문이다. 이러한 단점을 보완하고자 만들어진 것이 상위 개념으로서의 온톨로지이다. 그 대표적인 것 중의 하나가 SUMO(The Suggested Upper Merged Ontology)이다. SUMO는 현재 세계적으로 널리 통용되고 있는 상위 온톨로지로서 총 1,000 여개의 개념노드(Conceptual Node)가 설정되어 있다[7].
필자의 판단으로는 상위 온톨로지 설정에 SUMO를 사용하는 것이 몇 가지 점에서 장점이 있다. 첫째, SUMO는 여러 언어들에서 나타나는 모든 개념들을 포괄할 수 있는 상위 집합(superset)을 형성하고 있다. 이는 다국어 데이터베이스를 관리하고 확장 보수하는데 유리한 방식이다. 또한 새로운 언어를 추가할 때 기존의 어휘의미망이나 인덱스에 대한 그들의 등가 관계에 최소한의 영향을 미친다. 둘째, SUMO는 WordNet과 완전히 사상(mapping)될 수 있다. 따라서 세계적으로 가장 널리 사용되고 있는 어휘 의미망인 WordNet과의 연결에 큰 어려움이 없다. 반면에 다른 시소로스형 개념체계, 예를 들어 로제의 시소러스나, 일본의 어휘대계, NTT시스템은 나름대로의 가치를 지니고 있음에도 불구하고 WordNet과의 직접적인 연결

6) 'MCD(Multi-lingual Concept Dictionary)' 프로젝트 구상은 北京大學의 兪士汶교수와 카이스트의 최기선 교수가 구상했다. 뿐만 아니라 대만 중앙 연구원의 유명한 전산언어학자인 黃居仁교수도 방법론 제시에 많은 도움을 주었고, 일본 와세다 대학의 砂崗和子교수도 일본어 구축 작업에 참여하였다.
7) 이에 대한 자세한 내용은 SUMO의 공식 사이트(http://www.ontologyportal.org/index.html)에 소개되어 있다. SUMO의 온톨로지 작업은 IEEE 표준 상위 온톨로지 연구 팀에서 작성한 것으로 세계의 여러 학자들이 공동으로 참여하였다. 이 작업에 참여하고 있는 아시아권의 대표적인 학자로는 대만 중앙연구원의 언어학자인 黃居仁 교수를 들 수 있다.

이 불가능하다. 실제로 이러한 장점 때문에 현재 어휘 의미망 구축 과정에서 SUMO가 상위 온톨로지로 많이 사용되고 있다(Chu-Ren Huang, 2004).

(2) 어휘 개념 관계의 기술 방식: WordNet의 체계

본고에서 상정하고 있는 다국어 어휘 의미망인 MCD(Multi-lingual Concept Dictionary)은 어휘 개념 관계를 기술함에 있어 가능한 WordNet의 체계를 유지하는 것을 원칙으로 한다. 즉, WordNet의 유의어 집합인 Synset을 만들고 이들 간의 의미 관계를 그물처럼 연결시키는 체계를 유지하는 것이다. 그리고 다국어 어휘 의미망은 영어의 WordNet을 모태로 하여 중국어-한국어-일본어를 한 어휘의미망으로 연결시키는 것을 목표로 한다.

(3) 중국어.한국어.일본어 어휘 개념을 WordNet에 연결시키기

다국어 어휘 의미망 구축에 있어서 가장 효과적인 방법 중의 하나는 기존에 구축된 언어 자원을 적극적으로 활용하는 것이다. 본고도 중국어 데이터베이스 외에 한국어 데이터베이스와 일본어 데이터베이스를 가능한 한 많이 활용하고자 한다. 이렇게 하면 자원 구축에 드는 시간과 노력을 줄일 수 있고 효율을 높일 수 있을 것이다. 이들 간의 관계는 (그림1.)과 같이 나타낼 수 있다.

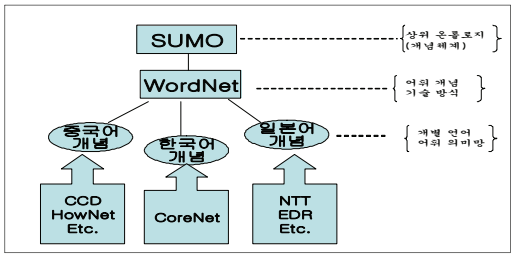


그림1. 다국어 어휘 의미망 구축의 기본 모형

3.2 구축 과정

다국어 어휘 의미망 구축의 과정은 크게 (1)SUMO와 WordNet의 연결 과정; (2)WordNet과 중국어.한국어.일본어 연결 과정으로 나뉜다.

3.2.1 SUMO와 WordNet의 연결 과정

SUMO의 개념과 WordNet은 기본적으로 컴퓨터로 자동 연결할 수 있도록 설계 되어 있다. 그리고 SUMO와 WordNet은 모두 인터넷에 자료를 무료로 공개하고 있다. 따라서 이 두 자료를 내려 받아 하나의 통합된 데이터로 연결시키고 그 개념의 중국어.한국어.일본어 번역을 하기만 하면 된다. 데이터베이스의 연결은 전산학 전공자가 주로 하고 언어학자들은 SUMO의 개념을 각각 중국어와 한국어와 일본어로 번역하는 일을 맡는다.

3.2.2 WordNet과 중국어.한국어.일본어 개념 연결 과정

이 과정은 다시 (1)영어-중국어 개념의 연결; (2) (영어)-중국어-일본어 개념의 연결; (3) (영어)-중국어-한국어 개념의 연결 과정으로 나누어진다.

(1) 영어-중국어 개념의 연결

이 과정은 영어와 중국어를 모두 잘 이해하고 있는 중국인 연구자들에 의해 이루어지는 것이므로 여기에서는 자세히 논하지 않겠다[8].

(2) (영어)-중국어-일본어 개념의 연결

이 과정은 영어의 WordNet과 중국어 개념을 참고하여 대응되는 일본어 개념을 연결하는 작업이다. 이 때 사용되는 일본어 어휘 자원은 NTT 어휘 데이터베이스와 EDR 어휘 개념 전자 사전이다. 그러나 이 작업 과정 역시 일본어와 중국어를 잘 이해하는 중국인과 일본인 연구자에 의해 이루어지므로 본문에서는 자세한 소개를 생략하도록 하겠다[9].

(3) (영어)-중국어-한국어 개념의 연결

이 작업은 영어와 중국어를 잘 이해하고 있는 한국인 연구자에 의해 이루어진다. 즉 연구자가 영어의 WordNet과 중국어 개념을 참고하여 대응되는 한국어 개념을 연결하는 것이다. 필자도 이 과정에 일부 참여하였다. 작업 과정에 주로 사용된 기초 자원은 중국어 WordNet인 北京大學의 CCD와 한국어 어휘의미망 데이터인 CoreNet이다. 이들은 모두 기계 가독형 사전(MRD:Machine Readable Dictionary)이다. CCD에는 영어-중국어 정보가 들어 있고, CoreNet은 한국어-일본어-중국어(일부) 어휘가 데이터베이스 형식으로 입력되어 있다. 이 두 가지 기계 가독형 전자 사전을 이용하여 중국어와 한국어의 연결(혹은 '사상(mapping)')을 하는 것이 본 작업의 주요한 과제이다. 이 과정은 다음 세 단계로 이루어진다.

첫째, 컴퓨터 프로그램을 이용하여 CCD의 중국어 개념과 CoreNet의 한국어 개념을 연결한다. CoreNet에는 중국어 정보가 들어 있다. 그래서 만약 이 중국어 정보가 CCD의 중국어와 일치하면 그 대역어인 한국어를 찾는다. 이 과정은 데이터베이스 연결 프로그램을 통해 자동으로 처리한다.

두 번째 단계는 자동 연결한 결과에 대한 적합성 검토이다. 자동 연결 작업은 방대한 어휘 의미 관계를 비교적 신속하게 처리할 수 있는 장점이 있다. 그러나 자동 연결 작업은 동형이의어 구분, 의미 세분화, 개념 체계 차이 등의 문제로 인해 오류가 발생할 가능성이 높다. 따라서 자동 연결 작업 이후에 오류 여부의 확인 및 선택된 대역어의 적합성 여부를 심도 있게 진행해야 한다.

세 번째 단계는 적합한 대역어를 찾지 못한 경우에 수동으로 진행하는 작업이다. CCD와 CoreNet은 의미 계층 구조가 정확히 일치하지 않는다. 더구나 두 언어의 차이가 존재한다. 따라서 자동 대역어나 연결 개념을 찾을 수 없는 경우가 발생한다. 현재로서는 이러한 문제점을 컴퓨터가 자동으로 해결하기 힘들다. 왜냐하면 CCD와 CoreNet의 정보가 불충분하기 때문이다. 결국 이 부분은 연구자의 개입이 필요할 수밖에 없다. 이 때 연구자는 중-한 사전이나 영-한 사전 또는 백과사전 등과 같은 다른 언어 자원을 참고하여 적절한 대역어를 선택하게 될 것이다. 이에 대한 처리 과정을 순서대로 요약하면 다음과 같다.

---

8) 이 부분에 대한 작업은 이미 많은 진행된 상태이다. 그 결과 臺灣의 中央研究院과 北京大學 計算語言學研究所의 연구 인력들에 의해 각각 Sinica Bow와 CCD라는 이름으로 영어-중국어 WordNet이 구축되어 있다.

9) 중국어와 일본어 연결 작업은 北京大學 計算語言學研究所의 연구원과 일본 와세다 대학의 砂崗和子교수가 공동으로 진행하고 있다.

<Begin>	비고
【제 1 단계】 1. 중국어 워드넷 CCD 파일을 불러온다. 2. CCD의 중국어 어휘를 CoreNet과 비교한다. 3. 만약 일치하는 중국어가 있으면 그 CoreNet 한국어 대역어와 개념번호를 연결한다. 4. 만약 CCD 상의 중국어 어휘를 CoreNet에서 찾을 수 없으면 동위어나 상위어와 대응되는 어휘가 있는지를 탐색한다. 만약 CoreNet에서 대응되는 상위어와 동위어가 존재한다면 그 의미에 대응되는 한국어와 개념번호를 찾아내어 연결한다. 5. 1부터 4의 과정을 반복한 다음 프로그램을 종료한다.	프로그램으로 자동 처리
【제 2 단계】 6. 연구자는 자동 연결 작업한 결과가 적합한 지 검토한다. 7. 만약 연결이 잘못되었거나 여러 대역어를 제시한 경우는 적합한 대역어를 선택한다. **【제 3 단계】** 8. 만약 컴퓨터가 자동 대역어나 연결 개념어를 못 찾아낸 경우에는 연구자가 중한, 영한 사전을 참고하여 적절한 대역어를 기입한다.	연구자의 수동 작업
<End>	

3.3 구축 도구: 프로그램 인터페이스

CCD 자료와 CoreNet 자료는 모두 데이터베이스 형식으로 되어 있다. 본고에서는 이 자료들을 MS Access 프로그램의 mdb 파일 형식으로 저장하였다. 그리고 데이터베이스 간의 자동 연결 작업은 Visual C++ 프로그램으로 처리하였다. 마지막으로 처리 결과를 열람하고 수정할 수 있는 편집 프로그램은 Visual Basic 프로그램으로 설계하였다. 다음은 그 처리 결과를 보여주는 프로그램 화면이다.

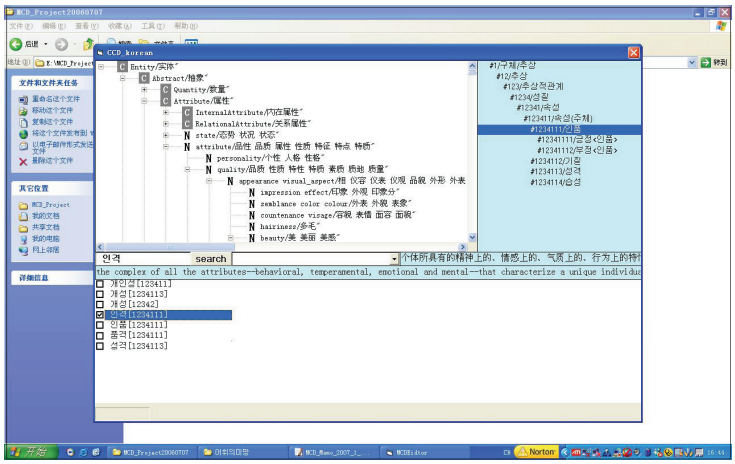


그림 2. 다국어 어휘 의미망 편집 프로그램

위의 그림에서 보이듯이 편집 화면은 크게 4 부분으로 나뉜다. 첫째, 좌측 상단은 중국어 워드넷인 CCD의 유의어 집합인 Synset의 체계를 나타낸다. 둘째, 우측 상단은 CoreNet의 체계를 나타낸다. 셋째, 가운데 편집 화면은 해당 어휘 개념, 예를 들어 "personality/個性 人格 性格"의 CCD상의 설명(gloss)을 나타낸다. 넷째, 화면 하단에는 체크 박스 형식으로 한국어 대역어 후보와 그 개념번호가 열거되어 있다. 편집자는 이 프로그램을 사용해서 적합한 대역어를 선택하고 저장하게 된다.

4. 영-중-한-(일) 어휘 의미망 구축의 실례 및 결과 분석

본 장에서는 필자가 직접 구축한 자료를 토대로 하여 실례를 소개하고 데이터베이스 안에 포함된 내용을 설명하고자 한다.

4.1 구축의 실례

영-중-한-일 다국어 어휘 의미망 데이터베이스 안에는 약 15개의 항목이 기술되어 있다. 이 자료는 각각 MS Access 파일 형식으로 15개의 필드(field)에 나누어 저장되어 있다. 그 중에 대표적인 항목으로는 MCD 개념번호, 품사 정보, Chinese.English.Korean_Synset(동의어 집합), Chinese.English_Gloss(주해), CoreNet 개념번호 등을 들 수 있다. 예를 들어 WordNet의 “teacher, instructor” 개념에 대응되는 중국어와 한국어 정보는 다음과 같이 기술되었다.

표2. “teacher instructor” 의 관련 어휘 정보

id	type	concept ch	concept en	concept ko	descript-ion_cn	description en	description_ko
M001002003003001003..	N	老師 講師 教師..	teacher instructor	교사 선생 강사	職業是教學的人	a person whose occupation is teaching	111131121

위의 그림에서 id는 해당 개념의 번호를 나타낸다. "00X"는 각 개념 노드의 단위이다. 그리고 "{老師, 講師, 教師,...}/ {teacher, instructor}/ {교사, 선생}"등은 동의어 집합(Synset:同義詞集)를 나타낸다. 그리고 "a person whose occupation is teaching/ 職業是教學的人"은 동의어들을 지시하는 개념을 설명한 주해(gloss)이다. 언뜻 보면 단순한 표의 나열에 불과하지만 이러한 자료는 개념 번호와 동의어 집합에 따라 상하위 관계, 부분 전체 관계 등을 나타낼 수 있다. 여기서 주목할 것은 주해의 기능보다는 동의어 집합과 개념번호가 훨씬 중요하다는 점이다. 주해의 기능은 사전을 이용하는 독자를 위한 편의 제공일 뿐, 자연언어처리에서 직접 이용할 수 없기 때문에 동의어 집합과 개념 번호의 중요성이 더 크다고 할 수 있다.

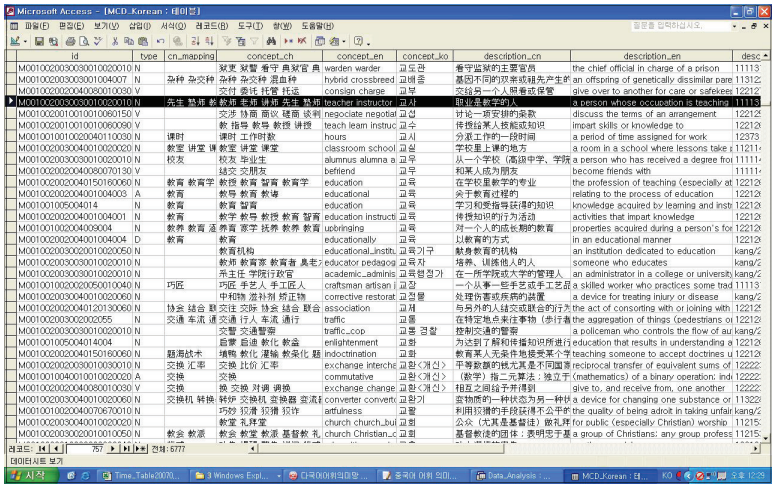


그림3. 다국어 어휘 의미망 데이터베이스의 실례

4.2 구축 결과 분석

본고에서 구축한 자료를 분석해 보면 전체적으로 상위 개념 위주로 되어 있음을 알 수 있다. 즉, 본 작업은 상위 개념들을 중심으로 하향식(top-down)방식으로 진행된 것이다. 위에서 말한 SUMO 개념은 상위 온톨로지에 해당하는 것이라고 할 수 있고, 나머지 어휘 개념들은 SUMO와 위계적으로 연결된다. 최상위 개념으로 설정된 것은 'Entity(실체)→ Abstract(추상) Physical(구체)'이다. 'Abstract(추상)'의 하위 개념은 'Attribute(속성), Relation(관계), Proposition(명제), Quantity(수량), SetOrClass(집합_또는_종류)' 등으로 나뉜다. 'Physical (구체)'의 개념은 'Agent(행위자), Object(물체), Process(과정), Phenomenon(현상), Region (지역)' 등의 하위개념으로 분류된다.

어휘 개념 체계를 표현하는 전형적인 방식은 개념 간의 위계(hierarchy)를 반영하는 것이다. 본문에서는 그것이 숫자로 표현되었다. 개념번호로 표현되는 계층적 개념 관계는 상-하 관계, 유-종 관계(일명 'is a' 관계)와 부분-전체 관계를 나타낸다. 그리고 상위 개념으로부터 하위 개념으로 갈수록 구체화되어 간다고 볼 수 있다. 예를 들어 “지진”이라는 어휘 개념은 다음과 같은 개념 체계를 가진다.

〔M001〕	{Entity} / {실체} / {實體}
〔M001002〕	{Physical} / {구체} / {物質}
〔M001002005〕	{phenomenon}/ {현상}/ {現象}
〔M001002005001〕	{natural_phenomenon nature} / {자연 현상} / {自然現象}
〔M001002005001001〕	{geological_phenomenon} / {지질 현상} / {地質現象}
〔M001002005001001001〕	{earthquake quake temblor seism} / {지진} / {地震}

위의 예에서 보이듯이 "지진(earthquake;地震)"이라는 개념은 "실체>구체>현상>자연현상>지질현상>"의 경로로 표현될 수 있다. 이것을 지칭하는 개념번호가 "M001002005001001001"인 것이다. 이처럼 전자화된 어휘 데이터베이스는 '지진(earthquake;地震)'이라는 하나의 어휘를 출발점으로 하여 그와 연결된 다른 개념들을 모두 확인할 수 있는 장점을 가진다. 자연언어처리의 입장에서 보자면 이러한 어휘 개념 정보는 컴퓨터가 의미를 추론할 때 유용한 자료로 활용될 수 있다. 물론 이 자료는 어휘 학습이나 어휘 개념 체계를 이해하고자 하는 전공자들에게도 도움이 될 수 있을 것이다.

한 가지 더 언급할 것은 어휘 의미 기술을 개별 어휘 중심으로 한 것이 아니라 개념 중심으로 한다는 것이다. 즉, 본 어휘 의미망 체계는 WordNet을 모형으로 하여 동의어집합(Synset)과 주해(Gloss)로 단어의 의미를 기술하였다. 동의어 집합(Synset)이란 하나의 개념을 지시하는 어휘들의 집합이다. 그리고 주해는 동의어들을 지시하는 개념을 설명한 것이다. 예를 들어 중국어의 '計算機'와 '電腦'는 의미적으로 동의 관계에 있으므로 하나의 'Synset'에 포함된다. 그리고 ‘a machine for performing calculations automatically (自動進行計算的機器)’는 이 동의어들의 개념을 설명한 주해이다. 동의어 집합은 '{X, Y, Z...}'와 같이 나타낸다. 예컨대, {計算機, 電腦} {丈夫, 老公,夫君} {妻子,老婆,內人} {老師, 教師, 教員} {春天, 春季}등과 같이 나타낸다.

4.3 구축상의 문제점 및 향후 해결 과제

어휘 의미망 구축 과정에서 직면한 문제점도 적지 않았다. 그 중에 몇 가지를 언급하도록 하겠다. 이러한 점은 앞으로 계속 수정·보완해야 할 것이다.

(1) 컴퓨터 자동 처리상에서 누락된 어휘들

컴퓨터로 자동 처리하는 과정에서 CoreNet의 해당 어휘와 개념번호를 찾지 못하는 경우가 종종 발생했다. 이것은 아마도 프로그램 설계 과정에서 한국어의 특징(예컨대 어미 변화, 띄어쓰기의 오류 등)을 고려하지 못하여 발생한 것이라고 생각된다. 차후에는 이 검색 방법에 약간의 규칙을 추가하여 자동 처리율을 높이도록 해야 한다. 만약 자동 처리 과정에서 누락된 어휘들이 많다면 사람이 일일이 중한 사전을 찾아서 수동으로 대역어와 개념번호를 기입해야 하므로 작업 시간이 몇 배나 늘어나게 된다.

(2) 한국어 대역어 선택의 제한

현재의 프로그램은 다중 매핑을 허용하지 않고 있다. 다만 영어와 중국어 동의어 집합에 대응되는 하나의 한국어만 선택할 수 있도록 설계되었다. 이것도 필자가 작업 중에 느끼게 된 어려운 점 중의 하나였다. 어느 것을 대표 어휘로 선택할 것인지에 대한 고민이 반복되었다. 앞으로 약간의 기술적인 문제를 해결하여 다중 매핑을 할 수 있도록 설계하는 것이 바람직할 것이다.

(3) 어휘의 등가성의 문제

어휘 의미의 구분은 문화에 따라 과잉으로 이루어질 수도 있고 과소로 이루어질 수도 있다. 예컨대, 영어에서는 아주 세밀하게 구분되는 어휘가 있다. 어휘의 종류도 많다. 반면에 중국어나 한국어는 상대적으로 그 개념에 일일이 대응되는 구체적인 어휘가 없을 수 있다. 동물이나 식물의 이름 등이 그러한 예이다. 현재로서는 이 경우에 해당 어휘 의미를 풀어서 해설하거나 그 어휘의 상위 개념어를 기입하는 방식으로 처리하고 있다. 그러나 앞으로는 번역하기 힘든 어휘 개념들을 처리할 좀 더 합리적이고 통일된 방법을 찾아야 하리라 생각된다.

(4) 결합관계 정보 누락

WordNet 체계의 의미망은 어휘 개념의 계열(paradigmatic) 관계 기술을 위주로 하고 있다. 반면 통사적 속성을 나타내는 통합(syntagmatic) 관계를 적절히 나타낼 수 없다는 한계가 있다. 그러나 동사나 형용사의 경우는 의미의 계열 관계도 중요하지만 통합(또는 결합) 관계도 중요한 정보이다. 따라서 통합 관계를 적절히 표현할 수 있는 보충 장치가 필요하다. 본고에서는 상위 개념어 중심으로 구축을 했기 때문에 문제가 덜 하지만 앞으로 하위 개념에 속하는 구체적인 어휘 집합의 특성을 기술할 때는 논항 정보, 연어 정보 등을 추가할 필요가 있다고 보인다.

5. 결론

본고에서 필자는 온톨로지 형태의 다국어 어휘 의미망 구축 방법과 실제에 대해 논하였다. 본문에서는 특히 범언어적 상위 온톨로지 자원으로서 SUMO(The Suggested Upper Merged Ontology)와, 영어의 WordNet, 중국어의 CCD(Chinese Concept Dictionary), 한국어의 CoreNet 등의 언어자원을 결합하는 방법에 대해 주로 논의하였다. 본 논의를 요약하자면 영-중-한-일 다국어 어휘 의미망의 기본 원칙은 SUMO에 기초하여 상위 개념 체계를 구성하고 WordNet 형식의 의미 기술 방식에 의거하여 영어-중국어-한국어-일본어 어휘데이터베이스를 연결한다는 것이다.

만약 본문에서 제시한 이상적인 다국어 어휘 의미망이 구축된다면 일반 언어학 연구뿐만 아니라 기계 번역, 정보 검색 등과 같은 자연언어처리 분야에서 중요한 기초 자원이 될 것이다. 첫째, 언어학적으로는 어휘 의미의 계열 관계를 기초로 하여 더 세밀한 통합(syntagmatic) 관계의 분석이 가능해진다. 그 중에 한 가지를 들자면 동사의 논항 구조(또는 격틀) 묘사에 중요한 정보를 제공할 수 있다. 오늘날 언어 구조를 동사를 중심으로 분석할 때, 동사가 취할 수 있는 논항구조 외에 논항의 의미 자질을 기술하는 것이 중시되고 있다. 예를 들어, 중국어 동사 '喜歡'은 기본적으로 두 개의 필수 논항(X=외부 논항, Y=내부 논항)을 요구한다. 이러한 동사의 논항 구조의 기술은 논항의 개수나 통사 범주만을 기술하는 것으로 끝나는 것이 아니라 이 논항의 의미 정보를 기술해야만 더 세밀한 분석이 가능하다. 만약 명사의 개념 체계가 이미 구축되어 있다면 의미정보의 기술은 훨씬 수월할 수 있다. 즉, '喜歡'의 통사 의미를 '喜歡: [X=동작주(인간|..), Y=대상역(인간|사물|식물|동물|자연경관|...)]'과 같이 기술할 수 있다. 따라서 다국어 어휘 의미망에 기초한 동사 정보의 분석은 통사정보, 의미정보, 개념 정보를 모두 포함할 수 있게 된다. 둘째, 자연언어처리의 관점에서 보자면 개념체계(온톨로지)에 기초한 어휘 의미망은 논리적 추론을 가능하게 함으로서 지식 처리를 가능하게 한다. 즉, 상위 개념의 특질이 하위 개념에 상속되는 관계를 통해서 어휘 의미의 중의성 해소(word sense disambiguation)나 자동적인 의미의 추론(semantic induction) 등이 가능해진다. 따라서 어휘의미망은 기계번역이나 정보 검색의 분야에서도 중요한 지식베이스로서의 역할을 할 것으로 기대된다.

이러한 작업은 기초 연구로서 계속 추진되고 있다. 그러나 그 구축이 쉽지는 않은 만큼 여러 가지 문제점들을 가지고 있는 것이 사실이다. 또한 미국의 WordNet과 같은 어휘 데이터베이스가 많은 언어학자와 전산학자들이 시간을 투자하여 이룩해 놓은 결과물인 것처럼 영-중-한-일 다국어 어휘 의미망 역시 많은 인력과 시간이 투자되지 않으면 안 된다. 이를 위해 앞으로도 중국 어학 연구자 및 전산학자들의 공동 연구를 통해 질적으로 우수한 어휘 의미망 구축 작업을 지속적으로 진행해야 할 것으로 생각된다.

【참고문헌】

최호섭·옥철영(2002) 한국어 의미망 구축과 활용, 한국어학 17

도원영·이봉원외(2004), 온톨로지에 기반한 한국어 동사 의미망 구축 시고, 한국어학 24.

최기선(2004), 코어넷(CoreNet) KAIST 개념기반 다국어 어휘의미망, 제2회 지식정보처리와 온톨로지 워크숍

한유석·설근수(2004) 한국어 시소러스 연구, 한국문화사

남유선·최병진(2005), 독한 워드넷 구축을 위한 독일어 한국어 동사격틀 분석 및 활용방안 모색, 독일언어문학 제29집

이동혁·오장근(2005), 다국어 어휘의미망에 한국어 관용표현을 연결하는 방법, 언어과학연구 36집

최경봉·도원영(2005) 한국어 동사 의미망 구축을 위한 상위 온톨로지 구성에 관한 연구, 한국어학 28

한국과학기술원 전문용어언어공학연구센터(2005), CoreNet 다국어 어휘의미망, 홍릉과학출판사

황순희·윤애선(2005), 의미자질을 고려한 명사어휘의미망 구축, 한국어학 29,

Vossen, P. ed.(1998), EuroWordnet : A Multilingual Database with Lexical Semantic Networks, John Benjamins Publishing Company, 1998. (한정한 외 6인 공역 유로워드넷 서울 한국문화사, 2004)

王惠 (2002) , 現代漢語名詞義位的組合分析研究, 北京大學博士學位論文。

劉揚(2003), 雙語WordNet語義知識庫的構造理論與工程實踐, 北京大學博士學位論文

Miller, G. A. et al. (1993) Introduction to WordNet: An On-line Lexical Database. Specification of WordNet.

Fellbaum, C. (1999) WordNet: an Electronic Lexical Database. Cambridge, Mass.: MIT Press.

Liu-Yang. (2002) Building a Bilingual WordNet-Like Lexicon. In Proceedings of COLING2002, P1243-1247, Taipei, China.

Yu,-Jiang Sheng (2003) The Specification of the Chinese Concept Dictionary. Journal of Chinese Language and Computing (in Singapore). Vol.13, No.2, P177-194.

Chu-Ren Huang, et al.(2004). Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. 4th International Conference on Language Resources and Evaluation(LREC2004). Lisbon. Portugal. 26-28 May, 2004.

〈중국의 다국어 어휘 의미망 구축과 활용 -CCD와 MCD를 중심으로〉에 대한 토론문

허철(단국대)

평소 존경해오던 강병규 교수님의 옥고를 잘 읽었습니다. 우리 인문학계에서 그동안 많이 주연구되지 않았던 어휘망에 관련된 해외 사례 소개와 선생님의 의견은 우리 연구에 큰 도움이 되리라 기대하며 선생님의 끊임없는 관련 연구를 부탁드립니다.
필요한 것만 취사선택하여 공부하던 제가 전체를 아우르지 못하기에 평소에 궁금했던 내용과 선생님의 원고를 읽으면서 생겨난 궁금증을 질문하고자 합니다.

1. 선생님께서는 워드넷에 기반한 “중국어-영어-한국어-일본어 다국어 어휘의미망 구축 방법”에 대해 논의하면서 “다국어 어휘 의미망 구축에 가장 필요한 자원은 개별 언어를 중심으로 구축된 단일어 어휘 의미망”의 중요성에 대해 말씀하셨습니다. 동의합니다. 기존에 각 언어의 어휘망이 존재하고 그 어휘망의 품질이 높다면 이것을 하나로 묶어 사용할 수 있습니다. 문제는 여기에 있는 것 같습니다. 각기 다른 기준과 특성, 그리고 양과 질의 차이가 있다면 기존의 개별적인 어휘망을 사용하기 보다는 구성하고자 하는 로우데이터에 맞는 기초자원을 새롭게 구축하는 것이 효과적이라고 보입니다. 특히 워드넷과 같은 형식을 공유하지만 내용은 표준국어대사전 등 우리말 사전의 정보를 이용하는 방법을 제안드립니다.

2. 다국어 어휘망이 영어-중국어-한국어-일본어로 이루어질 때 그 어휘 간의 관계에서 예상될 수 있는 문제는 언어의 특성에 관한 것이라 생각됩니다. 선생님께서도 말씀하셨던 한자문화권인 중국어와 한국어, 일본어의 경우 동음이의어가 발달해 있기에 이를 처리하는 것이 어렵다는 문제가 있습니다. 짧은 소견으로는 이를 해결하기 위한 여러 방법이 있지만 그 구현에는 여전히 어려움이 있는 것으로 알고 있습니다. 무엇보다 제가 궁금한 것은 이렇듯 각기 다른 어휘적 특성을 가진 어휘들을 단순 비교 혹은 일부 특정 대상의 비교가 아닌 전면적인 의미망 구축으로 가능하다고 보시는가이며, 이 때 초기 구성으로 생각해야 할 어휘 양과 이를 구축하기 위해 투입되어야 할 물리적인 노력의 정도는 얼마나 생각해야 하는가라는 현실적인 질문입니다.

3. 저 또한 동아시아 어휘 비교 연구를 할 때 기준어로 영어를 제시한 적이 있습니다. 기준어를 영어로 놓고 영어-중국어, 영어-한국어, 영어-일본어로 개별 구성하고, 이를 해당 영어 어휘 의미에 맞게 배열하는 방식이었습니다. 그런데 실질 구성에서 문제가 있었습니다. 개별 언어 셋트에 해당하는 어휘 구성은 인정될 수 있으나 이를 다시 중-한-일로 재구성했을 때도 이 관계를 등가적으로 볼 수 있는가의 문제입니다. 특히 모든 언어에 정통하지 않은 연구자이기에 이를 그대로 받아들이는 우를 겪은 적이 있습니다. 선생님께서는 이런 문를 어떻게 해결할 수 있다고 보시는지요? 두 번째 문제는 어휘의 형태적 특성입니다. 예로 드신 그림2를 보면 단음절 어휘와 다음절 어휘 중 다음절 어휘가 압도적으로 많지만, 실제 중국어에는 단음절 어휘도 상당수 사용됩니다. 이러한 단음절 어휘는 그 자체가 많은 의미

를 가지고 있는 경우가 대부분입니다. 동일 어휘가 다양한 의미를 가지고 사용되는 경우는 영어와 같은 서양어와 비교가 어렵지 않을까요?

4. 상위 온톨리지로 제안하셨던 SUMO가 동양 문화에 기인한 어휘들까지도 포함할 수 있는가에 대한 검증에 대한 연구 결과가 있는지 여쭙고 싶습니다.

5. 다국어 어휘의미망의 활용 범위와 가치에 대해 다시 한번 여쭙고 싶습니다. 선생님께서는 결론에서 언어 연구와 지식 베이스 연구(기계번역, 정보검색 등)을 말씀하셨습니다. 이를 좀 더 자세히 말씀해 주셔서 이 연구가 가져올 새로운 인문학 연구의 지평에 대해 말씀해 주시면 감사하겠습니다.

6. 아울러 드리는 질문입니다. 대부분의 온톨리지나 어휘의미망의 개발은 현대 사용 언어를 중심으로 이루어졌습니다. 그러나 일부 고전에서도 이를 활용한 지식베이스 연구를 시도하고 있습니다. 현대 언어의 경우 연구 성과가 비교적 많을 뿐 아니라 사용할 수 있는 자원도 있습니다만 고전어 같은 경우 매우 적고 단발적이며 자료의 축적 등이 긴밀하게 이루어지지 않고 있는 형편입니다. 만일 동양고전어를 중심으로 한 어휘망 설계, 보다 정확히 말해 단국대학교에서 출판한 한한대사전의 어휘를 대상으로 의미망을 구성하려고 한다면 이 때 고려해야 할 점은 무엇인지 선생님의 고견을 듣고 싶습니다.

이상입니다. 감사합니다.

디지털화된 이종 언어자원의 연계와 인문학 연구의 확장성
-『통합디지털한한대사전』과 KorLex의 연동 가능성 검토를 통해-

윤애선 (부산대학교)

- 차례 -

1. 들어가기

빅히스토리(Big History)의 관점에서 보면, 현생인류의 역사에서 지식의 축적에 혁혁한 공헌을 한 사건으로 음성언어, 문자, 인쇄술, 디지털 매체를 들 수 있다. 지식 축적의 형태와 방식은 다양하게 발달했는데, 그중에서 '사전(dictionary)'은 매우 독특한 위치를 차지한다. 아주 오랜 기간 동안 음성언어로 구전되던 지식이 문자의 발명과 종이처럼 보존할 수 있는 전달 매체를 통해 누적될 수 있었고, 15세기에 들어 인쇄술이라는 가히 혁명적인 수단을 통해 텍스트를 비교적 저렴한 가격에 대량 생산 및 유통이 가능하게 되었다. 인쇄술이 발전하면서 유럽에서 사전은 다양성과 양의 측면에서도 비약적으로 발전한다. 특히 18세기 중반부터 꽃을 피운 백과사전(encyclopedia)에는 인간의 언어로 표현될 수 있는 상당히 방대한 지식을 기술하고자 노력했다. 이러한 열정은 언어사전(language dictionary)의 편찬에도 경주되었고, 횡적으로는 다른 언어문화에도 급속도로 전파되었다. 다른 한 편으로는 사전의 방대한 내용을 체계적으로 담기 위해 기술형식(description format)에 대한 고려가 시작되었는데, 현대 사전학 용어를 빌리면 이른바 사전의 거시구조(macro-structure)와 미시구조(micro-structure)에 해당한다. 백과사전이든 언어사전이든 사전의 편찬은 짧게는 수 년, 길게는 수 십~수 백 년에 걸쳐 이루어진다. 200여 년간 종이에 인쇄되어 배포되었던 사전의 역할도 매우 컸다. 하지만, 20세기 후반부터 시작된 사전의 디지털화는 기존 사

전이 담고 있는 지식을 연동하여 질과 양 모두에서 확장성을 크게 제공할 수 있다는 점에서 새로운 장을 열었다. 디지털화된 사전은 컴퓨터/모바일기기와 인터넷을 통해 널리 유통되는 지식이 되었을 뿐만 아니라, '정보의 쓰레기통'에서 유의미한 지식을 추출할 수 있도록 시멘틱웹(semantic-web)에 유용한 그물망을 제공할 수 있었다(Pease et al. 2002).

국내에서도 1990년대 후반부터 종이사전을 디지털화하려는 노력이 시작되었다. 원본인 종이사전의 조판 방식에 따라 디지털화의 난이도와 구축 비용은 큰 차이를 보인다. 사전을 편찬하는 어떤 과정에도 컴퓨터를 사용할 수 없었던 육필본이나 활자를 심는 조판 방식으로 인쇄한 옛 사전의 경우에는, 디지털화하기 위해 사람이 사전 내용 전체를 입력해야 했다. 1990년대 초반부터 편찬되는 사전에는 컴퓨터 조판이 가능했고 점차 개인용 컴퓨터에서 워드프로세스를 이용할 수 있었기 때문에 디지털 파일의 형태로 존재하는 부분이 증가함으로써, 입력의 수고를 크게 덜 수 있었다. 하지만, 디지털화된 텍스트에서 정교한 지식을 추출할 수 있으려면, 사전의 거시구조와 미시구조를 분석하여 정규화된 형식(normalized format)을 갖춘 지식베이스(knowledge)로의 정제 및 가공 단계가 필요하다. 사전의 저작권을 가진 많은 기관, 단체, 개인이 다양한 목적을 위해 자신의 방식대로 이 작업을 수행했기 때문에, 표준화된 메타언어를 적용하지 못했다는 문제점이 드러났다. 하지만, 활용의 발판을 마련했다는 점에서 매우 큰 의의를 찾을 수 있다. 단국대학교 동양학연구원의 『통합디지털한한대사전(統合Digital漢韓大辭典)』(이하, 『통합』)은 전술한 종이사전 편찬과 디지털화의 역사를 그대로 보여주는 산 증인이다.

부산대학교 인공지능연구실의 한국어 어휘의미망 *KorLex*(Korean Lexical Semantic Network)는 완전히 다른 맥락에서 개발되었다. 그 모델은 영어 워드넷(Princeton WordNet, 이하 PWN)인데, 인간의 머리 안에 있는 지식이 '어휘' 단위로 저장되어 있다는 심리언어학의 이론을 근거로 20세기 중후반 미국 영어의 어휘의미 간 관계를 계층적 구조(hierarchical structure)로 설정한 것이다. 1980년대 중반부터 개발하기 시작한 PWN은 개발 단계부터 컴퓨터를 사용했다. 개발 초기에 상정했던 목적과는 달랐지만, PWN은 21세기 초반에 인터넷에서 유통되는 엄청난 양의 텍스트에서 '유의미한 지식'을 추출하는 데 아주 중요한 역할을 하게 되었다. 특히, 개발과 동시에 그 성과를 저작권 없이 공개한 무료배포 정책(copy-free policy)은 엄청난 파급 효과를 가져왔다. 각 언어에서 PWN을 이용한 파생 워드넷(derived WordNet)이 다수 만들어졌고, 해당 언어의 지식베이스가 되었을 뿐만 아니라, 등가의 어휘의미 연계를 통해 지식 구성 단위의 다국어 연동성(multi-lingual interoperability)을 확보하는 최초의 지식베이스가 되었다. PWN의 파생 워드넷인 *KorLex*는 다른 파생 워드넷들과 다국어 연동성을 확보하고 있었고, 한국어 내부에서는 『표준국어대사전』(이하, 『표준』)을 기반으로 다른 한국어 사전과의 연계성을 지향했다.

본 연구는 편찬 목적과 구축 환경에서 전혀 관련성이 없이 독자적으로 개발된 두 디지털화된 언어자원인 『통합』 및 *KorLex*의 연동 가능성을 검토해 보는 데 있다. 2절과 3절에서는 *KorLex* 및『통합』의 특성을 각각 살펴보고, 4절에서는 두 이종 언어자원 간 연계 가능한 요소가 무엇이며 현 단계에서 한계는 어떠한 지 검토하고, 이를 기반으로 미래의 인문학 연구를 활성화하기 위한 방향을 모색해 보겠다.[1)]

1) 본 발표는 제12회 동양학연구원 사전학학술회의를 주관하는 단국대학교 동양학연구원의 제안으로 시작하게 되었다. *KorLex*에 대한 소개와 함께 *KorLex*를 바탕으로 『통합』의 활용 방안을 제안해 달라는 것이었다. 필자의 과문 탓이겠지만, 『통합』의 존재를 필자가 그때 처음 알게 되었을 만큼 두 언어자원 간에는 공통점이 없어, 가시적인 연동 결과를 얻기 힘들 것이라는 점은 본 연구의 초기부터 예상했던 바이다. 특히 동양학연구원에서 많은 도움을 주었지만, 필자가 『통합』의 자료 전체를 직접 살

2. 한국어 어휘의미망 *KorLex*의 특성

*KorLex*를 구축한 부산대학교 인공지능연구실은 1990년대 초반부터 전산학자와 언어학자가 자연언어처리(Natural Language Processing, 이하 NLP) 분야의 공동연구를 수행하면서 이론을 연구할 뿐 아니라, 실용 NLP 프로그램을 개발해 왔다. 가장 대표적인 프로그램으로 '한국어 맞춤법/문법 검사기(Korean Spell and Grammar Checker, 이하 KSGC)'가 있는데, 30여 년간 지속적으로 성능이 향상된 버전을 출시하고 있다.[2] "*앓은, *졈, *같트니까"처럼 단순한 철자오류(non-word spelling error)만을 찾는 것이 아니라, 이것이 사용된 문맥에서 올바른 대치어를 제시할 수 있어야 하고, "감기가 {*낳으면 | 나으면} 전화해."나 "아기를 {낳으면 | *나으면} 전화해."에서처럼 문맥을 고려해야 오류인지 알 수 있는 문맥의존 철자오류(context- sensitive spelling error)를 검색하고 교정해야 하므로, 고도화된 NLP 기술과 다양한 언어자원(language resource)이 필요하다.[3] 따라서 형태소분석기(morphological analyser), 부분 문장분석기(partial sentence parser) 등을 개발했고, 각종 사전과 오류 관련 지식베이스를 구축해왔다. 하지만 맞춤법 검사기의 성능을 더 향상하기 위해서, 2000년도 중반부터 정교한 의미처리(semantic processing)의 필요성이 대두되었고, 연구팀이 범용으로 사용할 어휘망을 직접 개발하기로 한 것이 내부적인 동기다. 외부적인 상황으로는, 이때가 인터넷을 통해 유통되는 대규모 텍스트로부터 유용한 정보를 추출하고 이를 활용하기 위해 시맨틱웹 기술이 개발되기 시작한 시점이었다. 2.1절과 2.2절에서는 각각 PWN과 *KorLex*의 개발 목적과 정보구조를 소개한다. *KorLex*가 PWN을 모델로 한 파생 워드넷이지만 대다수 파생 워드넷과는 달리, 한국어 의미처리를 적합하도록 기존 정보를 수정하고 새로운 정보를 많이 추가했다. 2.3절에서는 *KorLex*가 다른 디지털 언어자원과 어떻게 연계되고, 어떻게 활용되는지 살펴본다.

2.1. Princeton WordNet의 개발 목적과 정보구조

1980년대 중반부터 시작한 PWN의 개발은 인지심리학적 배경에서 출발하였다.[4] 당시 미국 프린스턴 대학교의 인지심리학자 밀러(George Miller)는 '인간의 두뇌에 들어있는 지식이 어휘(word) 단위로 구동하며, 단위는 군집(chunk)을 이룰 수 있고, 계층적인 구조(hierarchical structure)를 형성하여 효율적인 방식으로 작동한다'는 이론 등 많은 연구로 상당한 명성을 쌓았다. 이를 바탕으로, 영어 화자의 머릿속에 들어있는 심상어휘집

펴볼 수 없었던 점과 한자/한문에 대한 필자의 무지가 가장 큰 걸림돌이었다. 따라서, 본 발표에서는 두 언어자원이 반드시 연동해야 한다는 필연성을 찾고 그 구체적인 성과를 제시하기보다는, *KorLex*라는 완전히 이질적인 언어자원의 관점에서 봤을 때 『통합』을 확장적으로 활용할 수 있는 방향을 모색함으로써 향후 새로운 특성을 가진 한국어 사전 편찬이나 사전학 연구의 지평을 넓히는 데 약간의 기여를 할 수 있다고 생각했다.

본고의 작성을 위해 『통합』과 *KorLex*의 원자료를 추출하고 다양한 통계를 내준 단국대학교 동양학연구원의 김지영 박사와 부산대학교 인공지능 연구실의 최성기 연구원에게 특별한 감사를 표한다.

2) KSGC의 최신 버전은 http://urimal.cs.pusan.ac.kr에 탑재되는데, 일반 사용자는 무료 이용이 가능하다. 이 사이트에는 KSGC 이외에도 본 연구진이 개발한 한국어 관련 소프트웨어와 자료를 공개하고 있다.

3) 문맥의존 철자오류의 검색과 교정에 관한 연구는 김민호 외(2014), 최현수 외(2015a, 2015b)를 참조하라.

4) PWN의 다운로드와 관련 자료는 웹사이트(http://wordnet.princeton.edu)를, PWN의 특성 소개는 Fellbaum (1998)과 윤애선 외(2009)를 참조하라.

(mental lexicon)을 재구성하겠다는 것이 PWN 개발의 목표였다.[5] 장기간에 걸쳐 파격적인 연구비를 지원받고, 심리학-언어학-전산학을 아우르는 공동 연구/개발팀을 구성하고, 당시로는 매우 놀랍게 컴퓨터를 PWN의 저장장치로 채택하고, 모든 결과물을 디지털 버전으로 도출하고자 했다. 1980년대 중반이면, 미국에서도 연구 현장에 개인용 컴퓨터를 아직 사용하지 못했던 때이고, 지금과 견주어 보았을 때 하드웨어의 성능도 매우 낮고, 그 하드웨어에 작동하는 DB나 컴퓨터 프로그램도 아주 원시적인 상황이었다. PWN에 담고자 하는 정보는 많았고, 컴퓨터 하드웨어와 소프트웨어의 성능은 매우 낮았기 때문에, 원 자료의 정보구조는 아주 복잡한 형식으로 표현되었다.[6]

앞서 언급한 것처럼, PWN의 관심사는 인간 지식의 표상으로서 심상어휘집을 재구성하는 것이다. 지식을 구성하는 것은 '개념(concept)'인데, 그것을 구체화한 언어적인 표현단위가 '어휘'라는 것이다. 그런데 동일한 '어휘형태(word form, 이하 어형)'는 1개 이상의 '어휘의미(word meaning, 이하 어의)'를 포함할 수 있고, 상이한 어형이 동일한(또는 유사한) 어의를 나타낼 수 있다. PWN에서는 '동일한 어의를 갖는 1개의 동의어 집합(synonym set, 이하 신셋)'이 1개의 개념을 표상한다고 정의한다. 예를 들어, [표1]에서 {paper2, report6}은 "an essay (especially one written as an assignment)"이라는 개념을 나타내는 신셋이고, {newspaper3, paper7}는 "a newspaper as a physical object"라는 개념을 나타내는 신셋이다. 여기에서 "paper, report, newspaper"는 어형이고, "paper2, report6, paper7, newspaper3"는 어의이다.

[표1] PWN의 신셋, 어의, 어형의 대응관계

어형 / 신셋{어의#}	paper	report	newspaper	...
{paper2, report6, ...}	○	○		
{paper7, newspaper3}	○		○	
{paper#, ...}	○			○

PWN에서 개념을 나타내는 신셋은 [그림1]처럼 계층구조를 이룬다.[7] {newspaper3, paper7}의 상위어(hypernym)는 {product2, production4}이고 자매어(sister nodes)는 {book2, volume3}, {book4} 등이다.[8]

---

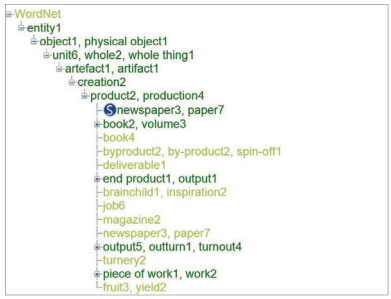
5) PWN의 구축 철학과 역사에 대해서는 Fellbaum(2005)과 Miller & Fellbaum(2009)를 참조하라.

6) *KorLex*를 개발하기 위해 2005년부터 PWN 및 이를 모델로 한 다른 언어 어휘의미망의 구축에 관한 선행 논문과 개발 문서를 살펴보았으나, PWN 원자료의 정보구조를 정확하게 파악하고 있는 선행 연구가 없었고, PWN에서 자신들이 필요한 일부 정보만 추출하여 사용하고 있었다. 암호문과 같은 PWN 원자료의 정보 표지와 구체적인 의미정보는 윤애선 외(2009:96)의 <표5>와 <표6>을 참조하라.

> 00935309 32 v 02 report_5 cover_2 008 @ 00831651 v 0000 + 06683784 n 0201 + 07217924 n 0101 + 06681551 n 0101 + 10521662 n 0101 + 06683784 n 0103 + 06683784 n 0102 $ 00967455 v 0000 03 + 08 00 + 09 00 + 22 01 | be responsible for reporting the details of, as in journalism; "Snow reported on China in the 1950's"; "The cub reporter covered New York City"

7) PWN의 계층구조는 품사별로 달리 적용되는데, 처음 구축된 명사와 동사는 계층 구조로 표현된 반면, 후에 구축된 형용사는 반의어를 중심핵으로 하는 방사형 구조로, 부사는 특별한 구조가 없는 목록(list)으로 표현되었다. PWN의 표현구조 변화에 대한 배경 설명은 Fellbaum (1998)의 ch. 2 "Modifiers in WordNet"을 참조하라.

8) [그림1]과 같은 PWN의 검색은 *KorLex*의 검색사이트(http://korlex.pusan.ac.kr)에도 함께 제공한다.



[그림1] PWN의 신셋 간 계층구조

PWN이 주목하는 것은 개념이었기 때문에, 어휘로 나타낼 수 있는 영어 품사 중 내용어(content word)인 명사, 동사, 형용사, 부사를 구축 대상으로 삼았는데, [표2]는 PWN의 버전별 발표 연도와 품사별 어형, 신셋, 어의 수를 보여준다. 또한 [표3]과 같이 신셋 및 어의 간에는 다양한 세부 의미정보가 기술되어 있다.[9)]

[표2] PWN 버전별 구축 크기

버전	발표 연도	명사			동사			형용사			부사			계		
		어형	신셋	어의	어형	신셋	어의	어형	신셋	어의	어형	신셋	어의	어형	신셋	어의
2.0	2003	114,648	79,689	141,690	11,306	13,508	24,632	21,436	18,563	31,015	4,669	3,664	5,808	152,059	115,424	203,145
2.1	2005	117,097	81,426	145,104	11,488	13,650	24,890	22,141	18,877	31,302	4,601	3,644	5,720	155,327	117,597	207,016
3.0	2006	117,798	82,115	146,312	11,529	13,767	25,047	21,479	18,156	30,002	4,481	3,621	5,580	155,287	117,659	206,941

[표3] PWN의 신셋 및 어의 간 의미관계

의미관계	관련 품사	예		표시단위	
				신셋	어의
동의	명,동,형,부	{board, plank} {rise, ascend} {sad, unhappy} {rapidly, speedily}			○
하의/상의	명,동	plant-〉tree-〉maple-〉sugar maple		○	
반의	명,동,형,부	wet 〈-〉 dry rapidly 〈-〉 slowly			○
유의	형	wet - watery, damp, moist, humid, soggy		○	
전의/분의	명	부분 구성소 물질	hat 〉 brim fleet 〉 ship milk 〉 protein	○	
함의	동	내포 양식 전제 인과	snore - sleep amble - walk divorce - marry kill - die	○	
속성	명-형	length - long, short		○	
영역	형,명,부,동	전문분야	chaotic-physics pas-ballet largo-music scroll-computer science	○	
		지역정보	blae-Scotland karate-Japan jolly-Britain scrimshank-Britain	○	

9) [표2]와 [표3]은 윤애선 외(2009:94-95)의 <표3>과 <표4>를 재수록했다. [표3]의 의미관계에 대한 상세한 설명은 윤애선 외(2009:95-96)를 참조하라. [표3]의 예에서 어의를 구분하는 아라비안 숫자 표시는 본고의 지면상 생략했다.

의미관계		관련 품사	예		표시단위	
					신셋	어의
			어법	commodious-archaicism bloomers-plural bang-colloquialism dandle-blend	O	
참조		형	true - correct, faithful, honest, sincere		O	
		동	pay - pay off			O
동일 어근	부속	명-형	icon – iconic, (hearing - auditive)			O
	파생	형-부	usual – usually, unusual - unusually			O
	관련	동-명	press-pressure, point-point			O
	분사	동-형	break-breaking, break-broken			O
동사군		동	{come to, resuscitate, revive}-{resuscitate, revive}		O	

PWN 및 파생 어휘의미망의 개발 현황과 이종 언어자원 및 지식베이스와의 연계정보는 워드넷 웹사이트(http://wordnet.princeton.edu)를 참조할 수 있고,[10] 무료로 배포되는 파생 어휘망 및 NLP 응용 프로그램은 Open Multilingual WordNet의 웹사이트(http://compling.hss.ntu.edu.sg/omw/)에서 다운받을 수 있다.[11] 2002년부터 격년으로 국제학술대회(Global WordNet Association Conferences)가 열려 관련 전문가가 연구 개발을 위해 지속적으로 교류하고 있다.[12]

2.2. *KorLex*의 개발 목적과 정보구조

한국어 어휘의미망 *KorLex*의 개발 목적과 배경은 PWN과 달랐다. 2004년에 *KorLex* 개발을 시작하기 전에 한국어를 대상으로 한 어휘망 내지 개념망에 연구와 개발이 시도되었으나, 시제품(prototype) 단계에 그쳐 실용성이 없거나 본 연구진의 필요성에 부응할 수 있을 만큼 상세하고 충분한 양의 한국어 정보를 제공하지 못하는 실정이었다.[13] 따라서 본 연구진이 직접 개발하기로 하였으며, 개발 방법론을 비교하고 기개발된 어휘망의 장단점을 살펴보았다. 그 결과 PWN 2.0을 기반 모델로 한 참조구축(reference-based) 방식을 채택했다.[14] 당시 본 연구진의 관점에서 PWN는 다음과 같은 장점을 가졌다고 판단했다.

① [표2]에서 보듯 약 15만 개의 어형과 약 20만 개의 어의는 중형 사전 이상의 규모로서, 실용 NLP 시스템에 적용될 때 높은 효율로 성능을 개선하는 효과를 가질 수 있다.

10) PWN 웹사이트에 연결된 논문 외에도 이종자원과의 연동을 통해 두 언어자원을 확장성을 넓힌 대표적인 연구로는 Reed & Lenat (2002), Gangemi et al. (2003a, 2003b), Ponzetto et al. (2009), Deng et al. (2009), Baccianella et al. (2010) 등을 들 수 있다.

11) PWN을 모델로 한 다국어 워드넷은 Vossen (1998), Tufis et al. (2004), Fellbaum & Vossen (2012), Rudnicka et al. (2018) 등을 참조하라. 프린스턴 대학교 외에 파생 워드넷 관련 통합 웹사이트가 유지되고 있는 곳은 Open Multilingual WordNet뿐이다(Bond & Ryan 2013, Bond & Paik 2012). 이 웹사이트는 일본에서 Japanese WordNet을 구축하고, 싱가포르의 Nanyang Technology University에 재직한 본드(F. Bond) 교수의 노력에 크게 힘입어 개발되었으나, 아쉽게도 2016년 이후 새로운 자료의 갱신이 이루어지지 않고 있다.

12) Global WordNet Association 홈페이지(http://globalwordnet.org)를 참조하라.

13) 2003년까지 한국어를 대상으로 추진된 어휘망 내지 의미망 개발과 그 특성에 대해서는 윤애선(2007)을 참조하라.

14) PWN의 여러 버전 중 2.0을 선택한 이유는 KorLex 개발 초기에 확보할 수 있던 최종 버전이었기 때문이다. 이후에 2.1버전과 3.0 버전이 출시되었으나, 2.0 버전과 내용에서 큰 차이가 나지 않고, PWN 홈페이지에서 세 버전의 신셋 간 사상(mapping) 표가 제공된다.

② 개념을 나타내는 신셋과 이를 구성하는 어의가 모두 어휘 단위로 작동하므로, 이를 규정하는 별도의 메타용어(meta-language) 설정이 필요하지 않다. 즉 개념과 어휘 간 괴리가 발생하지 않는다.[15]

③ 대부분의 의미망이나 개념망은 명사만을 이용하여 구축되었으나, PWN은 명사, 동사, 형용사, 부사를 모두 포함한다.

④ 계층적 구조는 NLP 시스템 등에서 언어규칙을 설정하거나 제어하는 데 효율적인 방식을 제공할 수 있다.

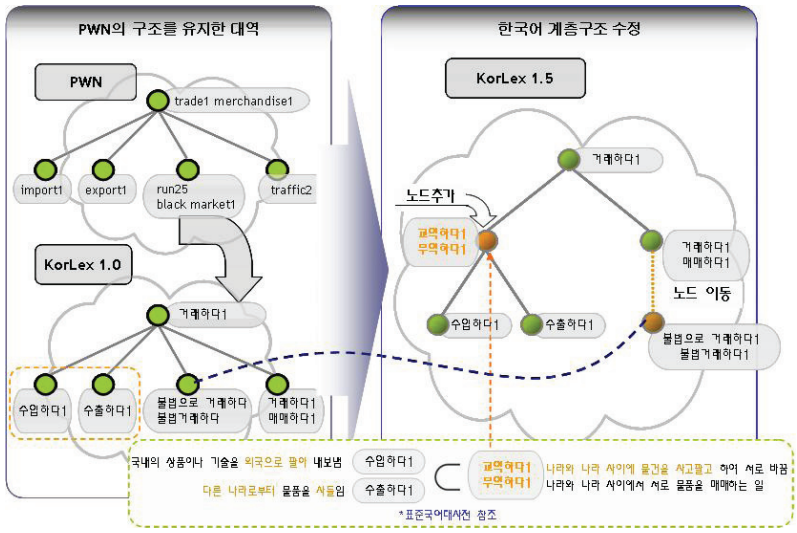
⑤ PWN의 최대 장점은 지적재산권을 주장하지 않은 무료 배포 정책이다. 이 정책은 파생 어휘망 구축과 다양한 분야에서 연구 및 활용을 크게 촉진한다.

⑥ PWN을 참조로 한 타 언어의 파생 어휘의미망이 다수 개발되었거나 개발 예정이라, 다국어 연계성이 가장 뛰어나다.

⑦ 다른 개념망이나 의미망과의 연동에 대한 연구가 활발히 진행되고 있었고, 상용 시스템에의 활용도도 가장 높다.

물론 본 연구진의 목적을 위해 PWN의 단점도 다음과 같이 분석하고, 이를 보완했다.

㉮ PWN이 영어 및 미국 문화에 경도되었으므로, 한국어 어휘의미망은 한국어와 한국문화를 반영할 수 있도록 수정해야 한다. 대부분의 PWN 파생 어휘의미망이 PWN의 정보구조를 그대로 유지한 채 다른 언어의 어휘로 대치하는 단순형 참조구축 방식을 사용한 것에 비해, KorLex는 한국어 및 한국문화를 반영할 수 있도록, [그림2]처럼 계층구조를 수정(추가, 삭제, 변경)하는 확장형 참조구축 방식을 적용했다.[16]

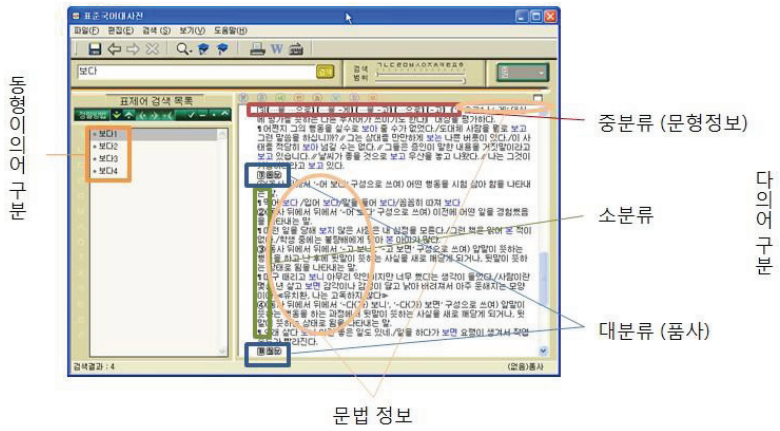


[그림2] KorLex의 확장형 참조구축 방식

15) 예를 들어, 『세종전자사전』에서 어휘집합을 정의하는 메타용어인 「세종의미부류」에는 '범위인간집단', '화시적장소' 등 실제 어휘가 아닌 표현을 만들어 사용한다.

16) [그림2]는 윤애선 외(2009:100)의 <그림2>를 재수록했다.

㉯ PWN에는 언어처리에 사용할 만큼 정교한 정보가 없다. 있다 하더라도 영어에 관련된 정보이기 때문에 한국어 정보처리에 도움이 될 수 있는 언어정보가 보완되어야 한다. ㉮와 ㉯의 문제점을 동시에 보완하기 위해, 『표준』의 어의 구분을 기준으로 삼았다.[17] 어의 단위에서 PWN와 『표준』을 사상(mapping)하여 『표준』의 언어정보를 최대한 활용할 수 있도록 했다. [그림3]은 『표준』에 실린 언어정보와 어의구분을 보여주고, [그림4]는 *KorLex*의 작업자용 워크벤치 중에서 PWN을 『표준』에 연동하는 과정을 보여준다.[18]



[그림3] 『표준』에서 "보다1" 의 언어정보와 어의 구분

㉠ 좌상단 창에서 수정 작업을 할 어휘망을 선택하고[예: KorLex], 검색할 어휘를 입력한다 [예: "이메일"].

㉡ 좌하단 창에 해당 어휘를 포함한 모든 신셋이 모두 출력되면, 원하는 신셋을 선택한다[예: {전자메일1, 전자우편1, 이메일1}].

㉢ 중상단 창에 검색된 신셋의 계층구조가 출력된다. 해당 신셋을 선택하면[예: {전자메일1, 전자우편1, 이메일1}],

㉣ 우측 창 상단에서 하단까지 다음 정보를 출력한다.

ⓐ 해당 신셋을 구성하는 KorLex 어휘;

ⓑ PWN의 신셋고유번호(Synest Offset);

17) *KorLex*가 어의 분할의 준거로 삼은 『표준』의 판본은 『표준국어대사전 1.0』(2001)이다.

18) [그림3]은 윤애선(2010:202)의 <그림1>을, [그림4]는 윤애선 외(2009:106)의 <그림7>을 재수록했다. 『표준』에 연계할 수 있는 적합한 어의가 없는 경우, [그림4]의 워크벤치에서 다른 사전 등의 어의와도 연동할 수 있으며, 초기 버전의 현황은 아래와 같다(윤애선 외 (2009:103) <표15> 재수록).

사전	KorLexNoun 1.5	KorLexVerb 1.5	KorLexAdj 1.0	KorLexAdv 1.0	KorLexClas 1.0	계
표준국어대사전	65,879	9,617	17,647	2,913	924	96,980
연세 한국어 사전	134	217	156	41	0	548
브리태니커 백과사전	34	0	58	2	0	94
프라임 영한사전	0	0	461	0	0	461
네이트 백과사전	15	0	32	0	0	47
네이버 백과사전	19	0	22	0	0	41
파스칼 백과사전	17	0	7	0	0	24
동의어 사전	1	0	0	0	0	1
기타	36,161	10,247	2,065	5	446	48,924
없음	98	52	457	162	7	776
계	102,358	20,133	20,905	3,123	1,377	147,896

ⓒ PWN의 정의문(Gloss);

ⓓ 다른 어휘망(Japanese WordNet, French WordNet 등)의 해당 신셋 정보;

ⓔ KorLex의 신셋 간 의미 정보;

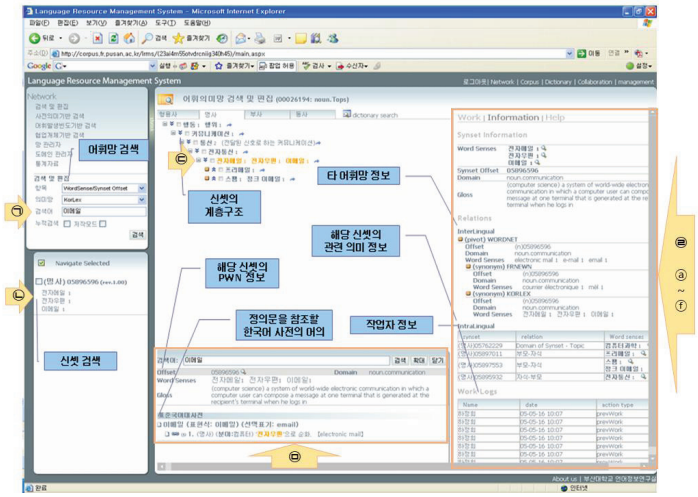
ⓕ 해당 신셋을 수정한 작업자 이력 정보;

㉤ ㉣-ⓐ에서 신셋을 구성하는 어휘를 선택하면, 다음 정보가 검색된다.

ⓐ 중하단 창에 해당 신셋의 PWN 정보;

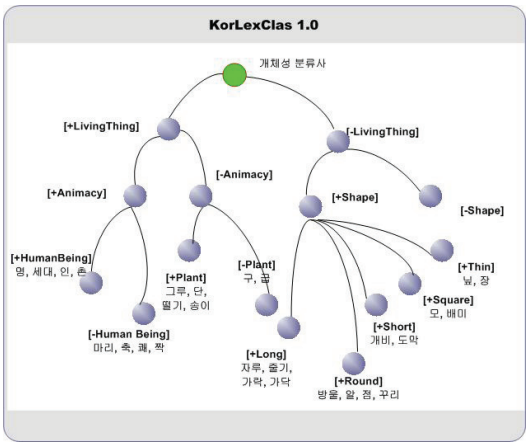
ⓑ 해당 어휘의 『표준』등의 어의별 정의문

㉥ 작업자는 PWN의 정의문에 가장 가까운 어의를 ㉤-ⓑ에서 선택하면 PWN과 *KorLex*를 연동된다.



[그림4] 작업자용 KorLex 구축 워크벤치

㉦ 한국어 의미처리에는 PWN를 구성하는 4개 품사의 내용어 외에도 추가적인 정보가 필요했는데, 기능어와 내용어의 역할을 모두 하는 수분류사(classifier)이다. *KorLex*는 [그림5]처럼 계층구조를 갖는 한국어 수분류사 어휘의미망(*KorLexClas*)을 새로 개발하고, [그림6]처럼 신셋 단위에서 명사 어휘의미망(*KorLexNoun*)과의 공기관계(co-occurrence) 정보를 구축했다.[19]



[그림5] *KorLexClas*의 계층구조

19) [그림5]는 윤애선 외(2009:102)의 <그림3>을, [그림6]은 윤애선(2012:185)의 <그림8>을 재수록했다.

[그림6] *KorLexClas* "대"의 신셋별 *KorLexNoun*과의 공기관계

㉮ 당초 *KorLex* 개발의 내부적인 동기가 KSGC의 성능 향상에 있었던 만큼, 한국어 문장분석과 관련하여 의미처리를 위한 상세 정보를 구축했다. 가장 중요한 것으로는 용언인 *KorLexVerb*와 *KorLexAdj*의 어의에 논항구조(argument structure) 정보를 추가하고 각 논항의 선택제약(selectional restriction) 정보를 *KorLexNoun*과 연동한다.[20] (예1)가 보여주듯이, *KorLexVerb* "갈다1" {*KorLex* 신셋번호 00076261}는 3개의 논항을 갖는데, 주어인 N0는 {인간1, 사람1}을, 도구격인 N2는 {기구1, 기계2}를, 목적어인 N1은 {고형 식품1}, {곡물1, 곡식1, 곡류1} 또는 {씨5, 씨앗1}이라는 선택제약을 갖는다는 정보를 추출하고, 이를 *KorLex* 2.0에는 (예1-ㄴ)과 같이 신셋번호로 저장한다.[21]

(예1) ㄱ. [{인간1, 사람1}]$_{N0}$-이 [{기구1, 기계2}]$_{N2}$-로 [{고형 식품1}|{곡물1, 곡식1, 곡류1}|{씨5, 씨앗1}]$_{N1}$-를 갈다
ㄴ. [00006026]$_{N0}$-이 [03443493]$_{N2}$-로 [07089248 | 07234211 | 07329605]$_{N1}$-를 갈다

[표4]는 *KorLex*의 버전별 각 품사의 어형, 신셋, 어의 수를 보여준다.[22] 단순한 언어정보의 양에서도 PWN을 상회하며, 본 연구진의 당초 목적에 따라 추가적인 한국어 정보가 구축되었으므로 언어정보의 질적인 면에서도 매우 높게 평가받고 있다.[23] *KorLex* 2.0 개발 이후에도 지속적인 확장을 하고 있는데, 주로 기개발되거나 개발 중인 전문분야의 어휘

20) 동사의 논항구조 및 선택제약 정보를 구축하는 방법에 대한 상세한 설명은 윤애선(2012)를 참조하라.
21) (예1)은 윤애선(2012:195)의 예문 (20)을 재수록했다.
22) [표4]는 윤애선.권혁철(2020:10)의 <表2>의 한국어 원본이다.
23) 언어학자로서 PWN 개발과 활용에 중심축 역할을 해온 펠바움(Christiane. Fellbaum)은 *KorLex*에 구축된 언어정보의 양과 질 덕분에 PWN 파생 어휘망 중 독보적인 역할을 하는 것으로 높이 평가했다(2018년 싱가포르에서 열린 제8회 GWA Conference에서).

망 또는 시소러스를 *KorLex*의 중하위 노드에 연결하여 그 상위에 있는 계층구조를 일반적인 개념 계층구조로 활용한다.

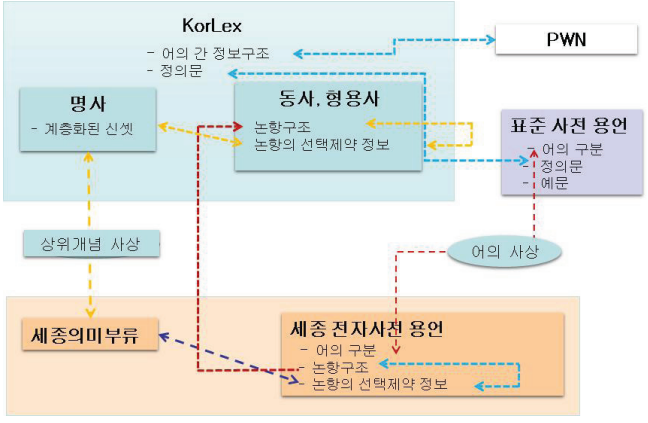
[표4] *KorLex*의 버전별 품사 및 언어정보의 양

버전 (개발 및 발표 연도)	단위	명사	동사	형용사	부사	수분류사	계
KorLex 1.0 (2004~2006)	어형	53,167	14,261	19,698	3,032	-	90,158
	신셋	58,565	13,429	18,558	3,651	-	94,203
	어의	59,405	14,700	20,905	3,123	-	98,133
KorLex 1.5 (2007~2009)	어형	90,909	17,957	19,694	3,032	1,285	132,877
	신셋	92,184	16,937	18,560	3,651	1,611	132,943
	어의	104,417	20,151	20,897	3,123	1,611	150,199
KorLex 2.0 (2010~2012)	어형	114,318	18,580	42,609	7,813	1,180	184,500
	신셋	106,366	17,354	18,660	3,668	1,377	147,425
	어의	139,836	24,932	55,346	9,634	1,377	231,125

2.3. 이종 언어자원과의 연계 및 활용

*KorLex*는 처음 설계할 때부터 PWN와의 연동성을 갖고 있었다. 신셋 고유번호를 이용하여 PWN와 *KorLex*를 연동하도록 개발했고, 이에 따라 PWN의 파생 어휘의미망뿐만 아니라 연계된 다양한 개념망과 자동으로 연동된다.

다른 한국어 언어자원과의 연동성은 *KorLex*에 부착된 『표준』을 이용하여 간접적으로 확보했다. 예를 들어, [그림7]은 『표준』을 이용하여 *KorLex*가 『세종전자사전』(이하, 『세종』)과 연동되는 방식과, 연계되는 세부정보를 보여준다.[24] 『세종』과 『표준』은 어의 단위에서 일치하는 정의문이 많기 때문에 『표준』에는 부족한 논항구조 정보를 『세종』에서 보완할 수 있었다.



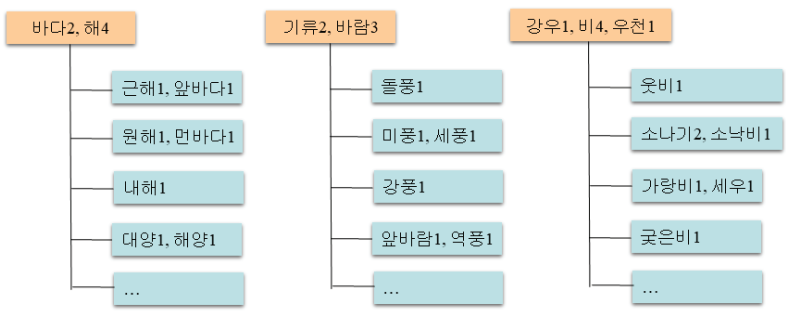
[그림7] *KorLex*와 이종 디지털 언어콘텐츠와의 연동 설계

24) [그림7]은 윤애선.권혁철(2020:12)의 [圖3]의 한국어 원본이다. 2.2절에서 설명한 (예1)과 같은 선택제약 정보는 『세종』의 세종의미부류와 *KorLexNoun* 간 사상을 통해 구축했다.

*KorLex*는 일차적으로 연구진이 개발한 각종 NLP 시스템에 적용되었고, 일반에 공개하여 한국어 의미처리와 지식처리가 필요한 다양한 연구 및 실용 시스템에 활용되었다.[25] 대표적인 예를 든다면, 한국전자통신연구원(ETRI)가 주도적으로 개발했던 다국어 기계통역 시스템인 지니톡(GenieTalk)과 지식처리를 기반으로 한 인공지능 문답 시스템(question & answer system)인 엑소브레인(ExoBrain)에 활용되었다. 대부분의 활용은 *KorLex* 개발을 계획했을 때 예상했던 것이었다.[26]

하지만, '한국어-한국수화 자동번역'(Korean-Korean_Sign_Language Machine Translation, 이하 K2KSL-MT) 시스템을 개발할 때, *KorLex*는 예상치 못했던 큰 수확을 거뒀다. K2KSL-MT는 본 연구진이 KBS와 함께 개발한 실시간(real-time) 자막 자동번역 시스템이다.[27] 뉴스의 자막에 적용되었는데, 뉴스에 전달되는 내용은 새로운 고유명사인 개체명(named entity)이 많이 등장하고, 다양한 어휘와 표현이 사용되므로 번역의 난도가 크게 올라간다. 하지만, 한국수화에는 한국어만큼 다양한 어휘의 세분화가 이루어지지 않았다. 적절한 번역어휘 및 표현의 선정은 실시간 번역 시스템에서 가장 중요한 과업 중 하나다. K2KSL-MT에서는 (예2)처럼 "앞바다, 돌풍, 웃비" 등은 『한국수어사전』을 포함한 수어 관련 사전에 등재되지 않았다. 한국어 어휘에 대응되는 한국수화를 찾기 위해서, [그림8]처럼 *KorLex*에서 계층구조 상 상위어(hypernym) 중 "바다, 바람, 비"로 대체함으로써 이러한 문제를 매우 효율적으로 해결할 수 있었다.[28]

(예2) 강릉 앞바다엔 돌풍이 불고 웃비가 내리는 곳도 있는데요
(a) 앞바다 => 바다
(b) 돌풍 => 바람
(c) 웃비 => 비



[그림8] *KorLex*에서 "앞바다, 돌풍, 웃비"의 의미 계층구조

25) 2021년 7월 21일 현재 부산대학교 산학협력단을 통해 배포된 KorLex(유료 버전) 중 연구용은 127개, 상업용은 72개이다.
26) 지니톡과 엑소브레인은 한국전자통신연구원의 대표성과 및 연도별(2017년, 2019년) 우수성과로 선정되었고, 홈페이지(https://www.etri.re.kr/korcon/sub5/sub5_0406.etri 및 https://www.etri.re.kr/korcon/sub5/sub5_0408. etri)에서 연구성과 파일을 내려받을 수 있다.
27) K2KSL-MT 개발의 배경은 2016년 8월부터 시행된 「한국수화언어법」이다. 이 법의 전문과 취지에 대한 소개 및 『한국수어사전』은 국립국어원의 해당 사전 홈페이지(https://sldict.korean.go.kr/front/main/main.do)를 참조하라. '한국수어'는 '한국수화언어'의 약칭이다.
28) (예2)와 [그림8]은 윤애선.권혁철(2020:21-22)의 (ex7)과 <圖9>의 한국어 원본이다.

3. 『통합디지털한한대사전』의 특성

대규모 사전의 편찬은 일반적으로 당초 계획보다 훨씬 오랜 기간이 소요된다. 이에 따라 비용도 기하급수적으로 증가한다. 공공기관에서 편찬하는 경우에도 해당 사전의 당위성을 계속 설득해야 겨우 유지할 수 있는데, 민간기관에서 편찬하는 경우에는 주요 추진자의 강력한 의지와 지원 능력이 수반되어야 큰 결실을 볼 수 있다. 단국대학교 동양학연구원이 오랜 세월에 걸쳐 편찬한 『한한대사전(漢韓大辭典)』(이하, 『한한』)과 『한국한자어사전(韓國漢字語大辭典)』(이하, 『한국』)은 후자에 해당한다. 또한, 기존의 자료를 디지털 콘텐츠화하고 이를 유지하며 확장하는 것에도 예상 밖의 시간과 노력을 경주해야 한다. 3.1절에서는 『통합』의 근간인 『한한』 및 『한국』의 편찬 목적과 배경을 소개하고, 3.2절은 디지털 콘텐츠로서의『통합』의 구성을 살펴보며, 3.3절에서는 현재 『통합』의 활용과 타 디지털 콘텐츠와의 연계성을 알아본다.[29)]

3.1. 편찬 목적과 배경

한반도 역사에서 지식 축적과 계승의 주요한 역할을 해온 한자 및 한문의 중요성을 주목하고, 우리와 같은 한자문화권인 일본과 대만이 이미 보유한 한자사전에 대한 부러움이 1970년대 후반에 『한한』 및 『한국』을 기획하게 된 가장 큰 원동력이었다. 일본에서는 『대한화사전(大漢和辭典)』이 1960년에 13권으로 출간되었고, 대만에서는 1968년에 『중문대사전(中文大辭典)』이 총 40권(색인 2권 포함)으로 편찬되었는데, 두 사전 모두 약 5만 개의 한자와 40~50만 개 내외의 어휘를 아우르는 대규모 사전이었다(김철웅, 2021:171).[30)] 1977년 당시 계획으로는, 동양학연구소(현 동양학연구원의 전신)를 편찬기관으로 하여 1978년부터 15년간 『한국』 4권 및 『한한』 17권의 발간을 목표했다. 두 사전의 편찬 목적은 긴밀히 연결되어 있다. 두 사전 중 규모가 더 작은 『한국』은 1992년부터 1996년까지 4권이 먼저 완간되었고, 규모가 훨씬 큰 『한한』은 집필이 지연되면서 16권(색인 1권 포함)으로 조정하고, 1999년부터 2008년까지 출간했다. 30년의 편찬 기간이 걸린 셈으로 계획보다 2배에 달했다. 완간의 지연은 사전의 편찬 과정에서 빈번히 접하는 광경이다. 이는 편찬자의 탓이 아니라 사전을 만든다는 작업이 그만큼 어렵고 끝이 없는 일이기 때문이다. 사전의 목적에 따라 사전 콘텐츠의 출처가 될 원전의 범위를 설정하고, 표제어 및 미시구조의 내용을 발췌하는 작업이 일직선으로 이루어지는 것이 아니고, 편찬 작업이 진행되면서 다시 처음부터 들여다봐야 하는 등 나선형으로 진행되는 경우가 흔하다. 또 종이 사전의 경우에는 인쇄하는 과정에서 조판과 폰트 등의 문제도 부딪힌다.

『한국』과 『한한』은 한국의 국학 및 동양학 연구에 필수적인 도구를 제공한다는 공통 목표를 갖고, 유기적으로 연결된다. 이중 『한국』은 한국에서 사용된 고유의 한자 및 한자어를 체계적으로 정리하고자 했는데,[31)] 이에 따라 『삼국사기』, 『삼국유사』, 『고려사』, 『조선

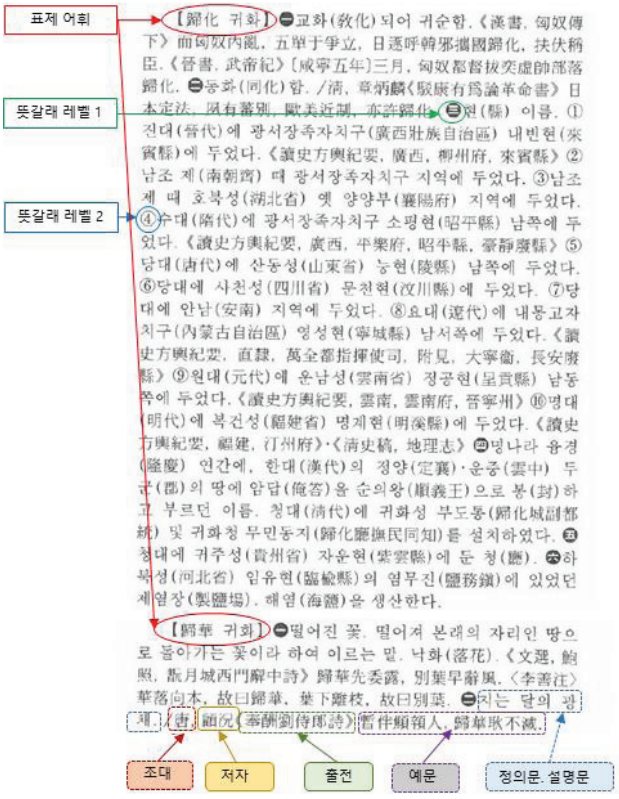
29) 본고의 3절은 단국대학교 동양학연구원의 선행연구 중 김지영(2013), 김철웅(2021), 윤승준(2012), 심경호(2018)의 내용을 요약한다. 이상 4개 논문에서 요약한 부분에는 인용을 따로 표시하지 않고, 직접 발췌한 경우에만 직접 인용 표시를 하겠다.

30) 『대한화사전(大漢和辭典)』은 2000년에 보충 1권 색인 1권을 추가하여 총 15권으로 구성되었고, 『중문대사전』은 1976년에 10권짜리 수정보급본이 출간되었다.

31) 『한국』에 수록된 표제자 중에는 한국어의 발음을 표시하기 위한 '돌(乭), 놀(弄)' 등 201개의 國子와 380여 개의 國儀字를 포함한다(김철웅 2021:173-174).

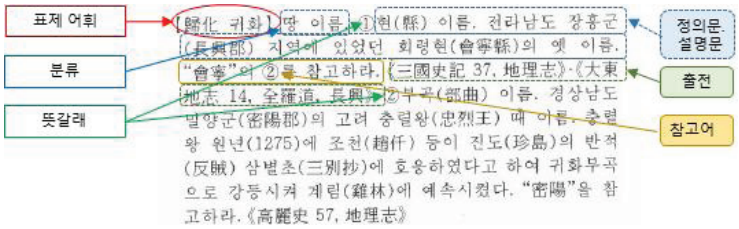
왕조실록』을 포함한 120종의 운서, 자전, 경서, 문학 등을 포함한 3,500여 권을 원전으로 삼았다. 인쇄 면수는 4,550쪽이고, 표제자 5,174개, 표제어휘 89,705개가 수록되었다.[32] 이에 비해 『한한』은 한국과 중국에서 사용된 모든 한자와 한자 어휘를 담으려고 했으며, 총 인쇄 면수는 20,786쪽이고, 표제자 54,964개와 표제어휘 420,269개를 포함한다. 두 사전 모두 한국 및 한자문화권의 고문헌의 독해에 도움이 될 수 있도록 일반어, 제도어, 전문어, 인명, 지명, 서명, 자호, 이두, 차자어, 동식물명, 성구, 속담 등을 포함했다고 함으로써, 백과사전적인 특성을 드러낸다.

우리의 관심사인 표제어휘의 미시구조가 어떤 정보로 구성되어 있는지 살펴보자. [그림9]와 [그림10]는 동일한 표제어휘인 "歸化" 등에 대해 『한한』과『한국』의 종이사전에서 제공하는 정보를 보여준다. 정보량이 많은 『한한』의 경우, 표제어휘는 한자와 한글로 제시되고, 정의문 또는 설명문은 뜻갈래(=어휘의미 구분)는 어의의 크기에 따라 레벨로 구분한다. 가장 큰 구분인 레벨1은 검은 원 안의 한자 숫자로, 레벨2는 흰 원 안의 아라비아 숫자로 나타낸다. 구분된 뜻에 따라 동의어나 참고어가 있다면 이를 제시하고, 조대, 저자, 출전, 예문(=용례)를 함께 제시하여 사용 맥락을 구체적으로 보여준다. 『한국』의 미시구조의 정보도 유사하나 정의문/설명문의 기술 형식과 뜻갈래 구분의 표시에 약간의 차이가 있다.



[그림9] 『한한』 표제어휘의 미시구조 구성정보

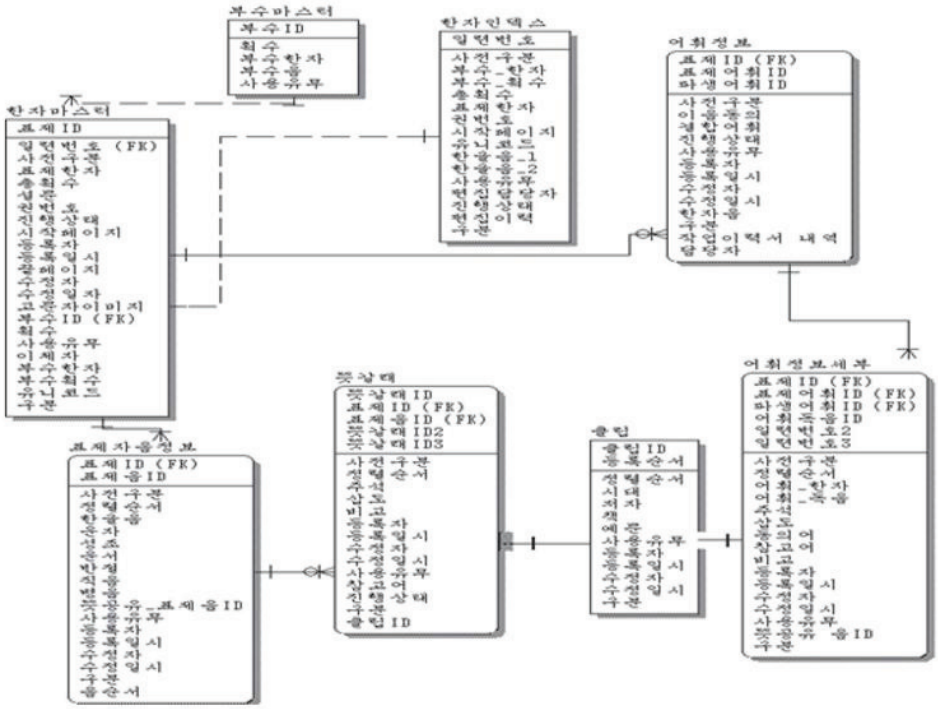
32) 사전에 따라 표제어와 그 구성요소를 부르는 명칭이 상이하다. 『한국』, 『한한』 등에서 "歸"나 "忠" 등 1자짜리 한자를 '표제자'로, 해당 표제자로 시작하는 "歸化"나 "忠烈" 등 2자짜리 이상의 한자어휘를 '표제어휘'로 칭한다. 본고에서도 '표제자'와 '표제어휘'라는 용어를 사용한다.



[그림10] 『한국』 표제어휘의 미시구조 구성정보

3.2. 디지털 언어콘텐츠로의 변환

『통합』은 물리적으로 분리되어 출간되었던 『한국』과 『한한』을 통합하여 디지털 콘텐츠화하면서 동시에 사전정보의 확장 가능성과 웹서비스를 고려했다. 크게 거시구조와 미시구조에 해당하는 내용을 구분하여 8개의 독립된 DB로 만들되, [그림11]처럼 각각의 DB 간 관련된 항목을 연결하는 관계형 DB로 구성했다. 또한 이 DB는 전용 통합사전편집기를 이용하여 온라인에서 『통합』의 콘텐츠를 수정, 확장, 삭제할 수 있다.[33]



[그림11] 『통합』의 관계형 DB 구성

앞 절 [그림9]와 [그림10]에서 보았던 표제어휘 "歸化"와 "歸華"의 미시구조 정보가 『통합』에 포함되어 있다.[34] [표5]는 [그림9]와 [그림10]에서 봤던 『한한』과 『한국』의 모든 어

33) [그림11]은 김지영(2013:225)의 <그림1>을 재수록했다. 각 DB의 세부 항목에 대한 설명과 전용 통합사전편집기에 대해서는 김지영(2013)을 참조하라.

의 정보를 포함하나, 미시구조를 구성하는 뜻갈래 수와 레벨을 수정하였고, '조대, 저자, 출전, 예문, 땅 이름, 현 이름' 등과 같은 태그명 또는 분류명과 같은 메타용어를 추가하였다. [표6]은 『통합』의 미시구조 정보가 『한한』과 『한국』의 어의와 어떻게 사상하는지 보여준다. "歸化"의 경우, 『한한』과 『한국』의 어의 중 어느 것도 누락되지 않고 재구조화되어 『통합』의 미시구조 정보를 구성한다.

[표5] 『통합』의 미시구조 정보

【歸化 귀화】
[1] 교화(敎化)되어 귀순함.
조대 저자 출전 漢書, 匈奴傳 下 예문 而匈奴內亂, 五單于爭立, 日逐呼韓邪攜國歸化, 扶伏稱臣.
조대 저자 출전 晉書, 武帝紀 예문 ［咸寧五年］三月, 匈奴都督拔奕虛帥部落歸化.
[2] 동화(同化)함.
조대 淸 저자 章炳麟 출전 駁康有爲論革命書 예문 日本定法, 夙有蕃別, 歐美近制, 亦許歸化.
[3] 하북성(河北省) 임유현(臨榆縣)의 염무진(鹽務鎭)에 있었던 제염장(製鹽場). 해염(海鹽)을 생산한다.
[4] 땅 이름.
[4_1] 현(縣) 이름.
[4_1_1] 진대(晉代)에 광서장족자치구(廣西壯族自治區) 내빈현(來賓縣)에 두었다.
조대 저자 출전 讀史方輿紀要, 廣西, 柳州府, 來賓縣 예문
[4_1_2] 남조 제(南朝齊) 때 광서장족자치구 지역에 두었다.
[4_1_3] 남조 제 때 호북성(湖北省) 옛 양양부(襄陽府) 지역에 두었다.
[4_1_4] 수대(隋代)에 광서장족자치구 소평현(昭平縣) 남쪽에 두었다.
조대 저자 출전 讀史方輿紀要, 廣西, 平樂府, 昭平縣, 豪靜廢縣 예문
[4_1_5] 당대(唐代)에 산동성(山東省) 능현(陵縣) 남쪽에 두었다.
[4_1_6] 당대에 사천성(四川省) 문천현(汶川縣)에 두었다.
[4_1_7] 당대에 안남(安南) 지역에 두었다.
[4_1_8] 요대(遼代)에 내몽고자치구(內蒙古自治區) 영성현(寧城縣) 남서쪽에 두었다.
조대 저자 출전 讀史方輿紀要, 直隸, 萬全都指揮使司, 附見, 大寧衛, 長安廢縣 예문
[4_1_9] 원대(元代)에 운남성(雲南省) 정공현(呈貢縣) 남동쪽에 두었다.
조대 저자 출전 讀史方輿紀要, 雲南, 雲南府, 晉寧州 예문
[4_1_10] 명대(明代)에 복건성(福建省) 명계현(明溪縣)에 두었다.
조대 저자 출전 讀史方輿紀要, 福建, 汀州府 예문 ·
조대 저자 출전 淸史稿, 地理志 예문

[4_1_11] 전라남도 장흥군(長興郡) 지역에 있었던 회령현(會寧縣)의 옛 이름. ⇨會寧의 [2]{3}.
참고어 會寧

조대 저자 출전 三國史記 37, 地理志 예문 ·

조대 저자 출전 大東地志 14, 全羅道, 長興 예문

[4_2] 명(明)나라 융경(隆慶) 연간에, 한대(漢代)의 정양(定襄)·운중(雲中) 두 군(郡)의 땅에 앙답(俺答)을 순의왕(順義王)으로 봉(封)하고 부르던 이름. 청대(淸代)에 귀화성 부도통(歸化城副都統) 및 귀화청 무민동지(歸化廳撫民同知)를 설치하였다.

[4_3] 부곡(部曲) 이름. 경상남도 밀양시(密陽市)의 고려(高麗) 충렬왕(忠烈王) 때 이름. 충렬왕 원년(1275)에 조천(趙仟) 등이 진도(珍島)의 반적(反賊) 삼별초(三別抄)에 호응하였다고 하여 귀화부곡으로 강등시켜 계림(雞林)에 예속시켰다. ⇨密陽.
참고어 密陽

조대 저자 출전 高麗史 57, 地理志 예문

[5] 청대(淸代)에 귀주성(貴州省) 자운현(紫雲縣)에 둔 청(廳).

【歸華 귀화】
[1] 떨어진 꽃. 떨어져 본래의 자리인 땅으로 돌아가는 것이라 하여 이르는 말. 낙화(落花).
조대 저자 출전 文選, 鮑照, 翫月城西門廨中詩 예문 歸華先委露, 別葉早辭風. 〈李善注〉 華落向本, 故曰歸華, 葉下離枝, 故曰別葉.
[2] 지는 달의 광채.
조대 唐 저자 顧況 출전 奉酬劉侍郎詩 예문 暫伴顦顇人, 歸華耿不滅.

34) [표5]는 [그림11]의 관계형 DB에 수록된 정보를 출력용으로 가공한 것이다.

[표6] “歸化”의『한한』과 『한국』의 어의 구분과 『통합』미시구조 정보의 재구조화

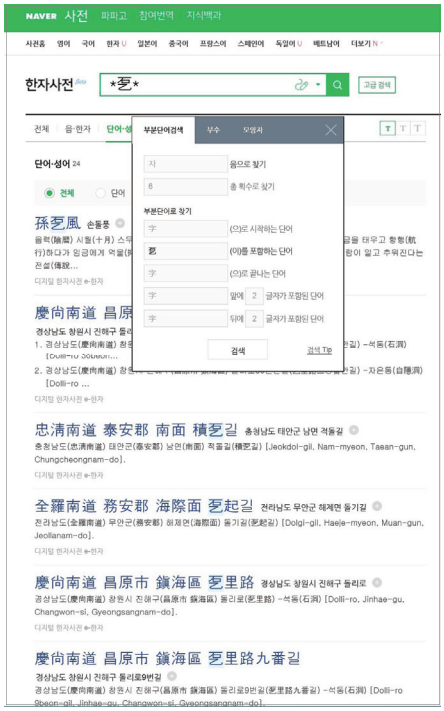
『통합』의 어의 구분				『한한』의 어의 구분		『한국』의 어의 구분	
[1]				一			
[2]				二			
[3]				六			
[4]					三 분류명	분류명	
	[4_1]			三			① 분류명
		[4_1_1]		三	①		
		[4_1_2]		三	②		
		[4_1_3]		三	③		
		[4_1_4]		三	④		
		[4_1_5]		三	⑤		
		[4_1_6]		三	⑥		
		[4_1_7]		三	⑦		
		[4_1_8]		三	⑧		
		[4_1_9]		三	⑨		
		[4_1_10]		三	⑩		
		[4_1_11]					①
	[4_2]			四			
	[4_3]						②
[5]				五			

3.3. 연계 및 활용

디지털 콘텐츠화한『통합』중에서『한국』은 [그림12]처럼 네이버의 사전 플래트폼에서 일반 사용자에게 서비스되고 있다. [그림12]에서 표제어휘 “歸化”가 인쇄본인 [그림10] 및 디지털콘텐츠화한 [표5]와 비교하면, 사소하지만 이 과정에서 누락되거나 오류인 정보를 찾을 수 있다. 예를 들어 ‘출전’ 정보인 “《大東地志 14, 全羅道, 長興》”은 ‘예문보기’ 기능을 선택해야 볼 수 있는데, 이는『한국』을 디지털 컨텐츠화하던 초기에 ‘예문’으로 태깅 오류가 있었던 것으로 보인다.[35] 이 사전 플래트폼에서는 검색창 우측의 ‘고급검색’ 기능을 통해 [그림13]처럼 간단한 콘코던스 프로그램(concordance)과 같은 검색 결과를 볼 수 있다. 하지만 아쉽게도 디지털 컨텐츠화된『한한』은 아직 일반에게 공개되지 않았다.

[그림12] 사전 플랫홈을 통해 서비스되는 『한국』

35) [표5]에서는 이 오류는 수정되었다.



[그림13] 사전 플랫홈이 제공하는 고급검색 기능

『통합』이 만들어지던 2011년 이후, 출간된 관련 논문은 [표7]과 같다.[36] 『통합』이 다른 대규모 언어자원과 연동 또는 연계를 통해 새로운 결과를 제시하는 유형의 논문은 없는 것 같다. 또한 소수의 연구자 및 동양학연구원 중심으로 연구가 이루어지고 있dj 확산이 제한적이다. 2017년 이후 석사학위 논문이 여러 편 발표되었으나, 동일 지도교수의 지도를 받은 논문이라 한 연구 집단의 결과물로 분류할 수 있을 것이다. 즉, 1차 연구자료로서 『통합』의 활용이 매우 제한적임을 알 수 있다.

[표7] 키워드 '한한대사전'로 검색된 학술논문

연도	저자	논문 명	학술지/호	주관기관
2011	최태훈	≪한한대사전(漢韓大辭典)≫ 일권(一卷)에 나타난 의미해석(意味解釋) 오류연구(誤謬研究)	중국어문논총/50	중국어문연구회
	최태훈	≪漢韓大辭典≫ 의 出典例文에 나타난 誤謬 研究	중국학/38	대한중국학회
	최태훈	《漢韓大辭典》 '甚' 字 誤謬研究	중국학/39	대한중국학회
2012	최태훈	『한한대사전(漢韓大辭典)』에 보이는 『김병매사화(金瓶梅詞話)』 관련 어휘 오류연구	비교문화연구/29	경희대학교 비교문화연구소
	윤승준	『한한대사전』의 편찬과정과 향후 계획	동양학/52	단국대학교 동양학연구원
2013	김영옥	한-중-일 한자자전(漢字字典)의 자형(字形)제시 기준 비교	한문교육연구/41	한국한문교육학회
	정재철	자전류의 역사와 한문 학습 자전의 필요성	한문교육연구/41	한국한문교육학회
	최태훈	『漢韓大辭典』에 보이는 『論語』 관련 어휘 연구	한중언어문화연구/33	한중언어문화연구

36) 검색은 부산대학교 도서관을 통해, '한한대사전'을 키워드로 하여 국내외 학술논문을 검색한 결과이다(검색일: 2021년 7월 16일). 『통합』과 관련된 논문이지만 해당 키워드를 부착하지 않은 논문은 검색에서 누락될 수 있다.

연도	저자	논문 명	학술지/호	주관기관
	김지영	『통합디지털한한대사전』의 DB구축과 온라인 사전편집기	동양학/54	단국대학교 동양학연구원
2014	최태훈	『한한대사전(漢韓大辭典)』에 보이는 명(明), 청대(淸代) 고백화어(古白話語) 오류연구(誤謬硏究)	중국언어연구/52	한국중국언어학회
2015	최태훈	『한한대사전(漢韓大辭典)』에 수록된 『사기(史記), 열전(列傳)』 관련어휘 오류연구(誤謬硏究)	비교문화연구/40	경희대학교 비교문화연구소
	최태훈	한자사전의 용례 선정 원칙과 실제 -'乾'자 표제어휘 용례를 중심으로 -	동양학/59	단국대학교 동양학연구원
	박진호	<漢韓大辭典>의 뜻풀이에 대하여	동양학/59	단국대학교 동양학연구원
	이건식	『통합 漢韓大辭典』의 國字 처리에 대하여	동양학/59	단국대학교 동양학연구원
	양찬진	『통합디지털한한대사전』 편찬 현황의 기계적 점검 방안 연구 - 표제자 및 표제어휘 통합 분석을 중심으로 -	동양학/60	단국대학교 동양학연구원
	엄미경	『碧巖錄』 飜譯을 통한 禪宗言語 '是甚마'에 대한 硏究	석사학위논문	동국대학교 대학원 (선학)
2016	최태훈	『한한대사전(漢韓大辭典)』에 기재된 『사기본기(史記 本紀)』 관련 어휘 오류연구(誤謬硏究)	동아시아 문화연구/64	한양대학교 동아시아문화연구소
	하정수	<한국한자어사전>의 음독구결	동양학/63	단국대학교 동양학연구원
	최태훈	『漢韓大辭典』에 보이는 몇 가지 誤謬分析과 그에 따른 『史記·世家』 번역서 校勘	중국언어연구/64	단국대학교 동양학연구원
2017	최태훈	한(韓)·중(中) 『사기(史記)·서(書)』 번역서 비교를 통한 『한한대사전(漢韓大辭典)』의 세 가지 관련 어휘 오류분석(誤謬分析)	동아시아 문화연구/70	한양대학교 동아시아문화연구소
	호홍휘	한·중 사전의 동음어와 다의어에 대한 연구: 동사를 중심으로	석사학위논문	중앙대학교 대학원 (국어학)
2018	심경호	한자사전의 현재적 의미와 개선 방안	동양학/71	단국대학교 동양학연구원
	곽가비	한국어 연결어미 '-자'와 '-자마자'의 중국어 대응 표현 연구	석사학위논문	동국대학교 대학원 (KSL)
	장원열	한국어 동자동의이음자(同字同義異音字)의 중국어 대응 양상: 한국 교육부 지정 상용한자 1800자를 중심으로	석사학위논문	동국대학교 대학원 (KSL)
	차오나	한국어 어미 '-으니'와 '-더니'의 중국어 대응 양상 연구	석사학위논문	동국대학교 대학원 (KSL)
2019	박덕영, 왕평	關于修訂《漢韓大辭典》及推出网絡辭典的建議 — 對比《大漢韓辭典》, 《ZON漢字辭典》等辭典予以探討	중국학/67	대한중국학회
	설홍	한국어 '아름답다', '예쁘다', '곱다'의 중국어 대응 표현 연구	석사학위논문	동국대학교 대학원 (KSL)
	장예운	한국어 다의어 '타다[乘]'의 중국어 대응 양상 연구	석사학위논문	동국대학교 대학원 (KSL)
2021	김철웅	동양학연구원 50년의 성과와 과제	동양학/82	단국대학교 동양학연구원

4. 『통합』과 *KorLex*의 연동 가능성 검토

『통합』과 KorLex는 편찬의 목적과 배경, 편찬 방법과 미시구조, 타 언어자원과의 연동 방식과 활용 분야도 완전히 다르다. 2.2절과 2.3절에서 살펴봤듯이 이종 언어자원의 연동 역시 많은 시간과 노력을 요한다. 따라서 그 목적이 뚜렷하면, 구체적인 계획을 설계하여 활용도를 크게 높일 수 있다. 본 논문은 『통합』과 *KorLex*의 연동가능성을 타진해 보는 데 그 목적이 있는 만큼, 두 언어자원의 중간 교량 역할을 하는 『표준』과의 연계성을 통해 살펴본다. 4.1절과 4.2절에서는 외형적으로 『통합』과 *KorLex*의 연계점이 될 수 있는 표제어휘의 한자어 정보의 사상 가능성을 거시적 관점과 미시적 관점에서 검토하고, 4.3절에서는 현 단계에서 『통합』의 활용을 확장하기 위한 방안을 알아보겠다.

4.1. 연계점

2절과 3절에서 『통합』과 *KorLex*의 미시구조를 살펴본 바, 두 이종언어자원의 연계점이 될 수 있는 단위는 한자어 정보를 가진 표제어휘이다. 『통합』의 모든 표제어휘가 한자어이나, *KorLex*는 각 어의에 연계된 『표준』 등의 어의가 있고, 그것이 한자어 정보를 포함한다는 매우 제한적인 조건을 만족해야 한다. 또한 『통합』과 *KorLex* 간 중간 매체가 『표준』이므로, 우선 『통합』과『표준』의 미시구조 상 한자어 정보를 추출하는 데 어떤 관련성과 한계가 있는지 살펴보자.

첫째, 어형 단위에서 사상되는 표제어휘의 수는 『통합』과 『표준』 간에는 73,187개, 『통합』과 *KorLex* 간에는 10,229개이다. 『통합』이 포함하는 표제한자어의 어형이 493,563개이므로, 어형 단위의 사상 비율이 각각 14.88%와 2.07%이다.[37]

[표8] 『통합』과 『표준』 및 KorLex의 연계 가능한 표제어휘의 어형

비교 사전 / 기준 사전	「표준」		KorLex	
「통합」	73,466/493,563	14.88%	10,228/493,563	2.07%

둘째, [표8]에서 제시한 어형을 대상으로 표제어휘의 글자 수에 따른 분포는 [표9]와 같다. 2자~4자가 거의 99%에 달한다.

[표9] 『통합』과 『표준』 및 KorLex 간 연계 가능한 어형의 글자 수별 분포

어형의 글자 수	「통합」-「표준」		「통합」- KorLex	
2	58,416	79.51%	9,405	91.95%
3	9,424	12.83%	641	6.27%
4	4,749	6.46%	174	1.70%
5	614	0.84%	8	0.08%
6	194	0.26%		
7	34	0.05%		
8	28	0.04%		
9	4	0.01%		
10	2	0.00%		
11	1	0.00%		
계	73,466	100.00%	10,228	100.00%

셋째, 동일한 어형의 표제어휘는 여러 개의 어의를 포함할 수 있는데, 이종 언어자원 간 정확한 사상을 하려면 어의 단위에서 이루어져야 한다. 하지만 이종 언어자원의 경우, 어

37) 『통합』의 정의 상 표제어휘는 2자 이상으로 구성된 어휘이다. 1자짜리 표제자의 통계는 별도의 작업이 필요하여 이번 발표문에는 포함하지 않았다. 또한 , 『표준』과 *KorLex*에는 어휘의 일부가 한자어로 형성된 합성어나 파생어의 경우(예, "自然스럽다, 발림性"도 [표8]의 연계가능한 표제 어휘 어형 수에 포함된다. 『통합』에는 품사 정보가 따로 표시되지 않기 때문에, 『통합』과 『표준』(*KorLex*) 간 사상을 하게 되면 이 점을 고려해야 한다.

의 분할 기준과 어의의 크기가 다르다. 앞서 [그림3]과 [그림9]에서 봤듯이, 『표준』과 『통합』이 어의 구분을 3단계로 하고 있으나, 어의의 크기나 갈래가 상이할 것이고, 한 어형이 포함하는 어의 수도 비대칭적인 경우가 많을 것이다. 또한 기준이 되는 사전을 어느 것으로 정하느냐에 따라 동일 표제한자로 연동 가능한 어의의 수도 상이하게 된다. [표10]과 [표11]은 [표8]에 제시된 동일한 어형 표제어휘의 어의 수와 비율을 보여준다. 일반적인 예상과는 달리 『통합』을 기준으로 했을 때, 동형 표제어휘의 어의 단위 사상 가능성은 1.6~2배 정도 높다. 또한 『표준』이나 *KorLex*을 기준으로 했을 때에는 그 사상 가능성이 33%~36%에 달해, 특히 『표준』의 어의 보완에 『통합』의 유용성을 미루어 짐작할 수 있다.

[표10] 『통합』과 『표준』의 연동 가능 어의

비교 사전 / 기준 사전	「표준」		「통합」	
「통합」	156,554/661,936	23.65%	-	
「표준」	-		138,444/382,222	36.22%

[표11] 『통합』과 KorLex의 연동 가능 어의

비교 사전 / 기준 사전	KorLex		「통합」	
「통합」	27,539/661,936	4.16%	-	
KorLex	-		20,953/63,409	33.04%

4.2. 연계점의 표본 분석

앞 절에서 『통합』과 『표준』 및 KorLex 간 연동 가능성과 문제점을 거시적인 관점에서 살펴봤다면, 이 절에서는 미시적인 관점에서 검토해 보겠다.

첫째, [표9]가 제시한 『통합』과 『표준』에 공통으로 나타나는 어형의 예를 글자 수 별로 살펴보자.

[표12] 『통합』과 『표준』 간 연계 가능한 어형의 글자 수 별 예

어형의 글자 수	「통합」-「표준」		표제어휘의 예
2	58,416	79.51%	家計, 加工, 假死, 假說, 佳友, 工事, 公敵, 過用, 耐性, 內衣, 雷雨, 大應, 料理, 螻蟈, 類推, 隆盛, 廩入, 理解, 立法, 麻姑, 漠漠, 網羅, 妄想, 妄言, 妹夫, 麥芽, 盲啞, 面皮, 防水, 妨害, 變例, 辨理, 變化, 屛黜, 竝行, 報復, 紗籠, 死法, 邪師, 思索, 邪心, 死地, 射侯, 山翁, 撒布, 三三, 杉板, 宣言, 說明, 纖細, 纖月, 深趣, 十翼, 雙生, 芽甲, 雅量, 牙拍, 啞然, 我執, 安保, 安穩, 斡旋, 暗礁, 壓膝, 仰訴, 艾虎, 夜天, 洋紗, 禦侮, 抑鬱, 億丈, 嚴昕, 逆水, 蜿蜒, 位高, 威嚴, 唯獨, 爭諫, 錚錚, 詛呪, 的當, 敵手, 積陳, 前房, 靑春 (일부 발췌)
3	9,424	12.83%	假監役, 加一年, 甘藷譜, 開土祭, 界三葉, 可否間, 金聖嘆, 急流水, 給事中, 那爛陀, 男妹間, 娘子軍, 茶食板, 唐琵琶, 大關節, 大頭瘟, 大小家, 德周寺, 道德經, 敦化門, 慕容皝, 木芍藥, 蒙古語, 無名指, 問安婢, 未亡人, 朴僉知, 百家語, 白玉壺, 別頭場, 奉先庫, 三體詩, 詳定例, 西突厥, 鼠目太, 閃刀紙, 細麻布, 水錦花, 壽庭木, 時波赤, 魚鱗圖, 緣覺乘, 長寧殿, 贓物罪, 裁判所, 全東屹, 諸宮調, 從父法, 村夫子, 親査頓, 駝酪粥, 婆婦草, 豹尾旛, 下棺布, 割衫婚, 許生傳 血見愁, 戶路國, 樺皮弓, 吽哆敎, 陰騭文 (일부 발췌)

[표12] 계속

어형의 글자 수	『통합』-『표준』		표제어휘의 예
4	4,749	6.46%	家家戶戶, 加減乘除, 可東可西, 佳人薄命, 刻舟求劍, 改過遷善, 孤雌寡鶴, 功虧一簣, 舊歲問安, 勸學條例, 今是昨非, 南五味子, 內傳消息, 能見難思, 多士濟濟, 達魯花赤, 大豆黃卷, 大書特筆, 都結兒匠, 同黨伐異, 蠹國病民, 亂世之音, 馬革裹屍, 滿城風雨, 猛健副尉, 明明白白, 猫項懸鈴, 聞一知十, 文質彬彬, 方面之任, 百年偕老, 法久弊生, 普通學校, 不得其所, 削奪官職, 三五之隆, 尙書庫部, 承安之樂, 食前方丈, 搖尾乞憐, 雨傘差備, 醫學博士, 紫門軍契, 壯元及第, 電光石火, 鮎魚上竹, 漸入佳境, 尊王攘夷 捉鷹別監, 擅句屬官, 風雲月露, 下筆成章, 肉頭文字 (일부 발췌)
5	614	0.84%	加德島海戰, 兼愛交利說, 高麗史節要, 交食推步法, 勸善指路歌, 南北村便射, 大都護府使, 讀書三品科, 龍鳳詩箋紙, 無量光明土, 發心修行章, 四溟堂實記, 三十三觀音, 瑞日和之曲, 選武軍官布, 禪門拈頌集, 宣部守三薦, 先進排後受, 惺所覆瓿藁, 稅關監視署, 續資治通鑑, 承宣院日記, 兒女英雄傳, 與天地偕亡, 五虎大將記, 元人百種曲, 咨離牟盧國, 族曾祖父母, 增正交隣志, 知子莫如父, 彩鳳感別曲, 天東象緯考, 婆娑尼師今, 判監察司事, 判中樞府事, 八道地理誌, 韓日新協約, 漢學上通事, 鄕藥集成方, 憲兵司令部, 惠民典藥局, 胡笳十八拍, 洪範十四條, 華山仙界錄, 畫禪室隨筆, 火砲式諺解, 皇極經世書, 會計檢查局, 訖解尼師今 (일부 발췌)
6	194	0.26%	嘉禮都監儀軌, 慶尙道地理誌, 館學儒生應製, 禁軍廳號令旗, 大越史記全書, 禱千手觀音歌, 東國通鑑提綱, 同知訓鍊院事, 萬波停息之曲, 武藝二十四般, 非想非非想處, 三十九餘甲幢, 小京餘甲幢主, 修城禁火都監, 阿尼大都唯那, 五倫全備諺解, 二十二史箚記, 鄭瓜亭三機曲, 左邊捕盜大將, 知尙書工部事, 參知門下府事, 春坊通事舍人, 判尙書兵部事, 韓日協商條約, 皇太孫講書院, 訓民正音韻解 (일부 발췌)
7	34	0.05%	各殿宮動駕儀節, 兼都評議使司事, 癸未搢紳風雨錄, 古今經驗活幼方, 高嶺鎭民善政歌, 國朝名臣言行錄, 閨中七友爭論記, 內史侍郎平章事, 唐宋八大家文鈔, 東萊接倭事目抄, 夢中老少問答歌, 門下侍郎平章事, 兵馬同僉節制使, 兵馬水軍節制使, 三道陸軍統禦使, 三道水軍統制使, 三士橫入黃泉記, 上輔國崇祿大夫, 商議式目都監事, 禪宗永嘉集諺解, 世宗實錄地理志, 續資治通鑑長編, 水軍同僉節制使, 藥師瑠璃光如來, 王太子宮侍講院, 帝室制度整理局, 帝室會計監査院, 中書侍郎平章事, 僉議侍郎贊成事, 淸江使者玄夫傳, 判都僉議使司事, 判都評議使司事, 八萬四千大藏經, 湖南丙子倡義錄
8	28	0.04%	各陵改修都監儀軌, 開國原從功臣錄券, 建炎以來繫年要錄, 慶尙道續撰地理誌, 經世訓民正音圖說, 寡婦處女推考別監, 俱利伽羅不動明王, 大匡輔國崇祿大夫, 都序科目幷入私記, 都摠中外諸軍事府, 督辦交涉通商事務, 東洋拓殖株式會社, 同中書門下平章事, 壁上三韓三重大匡, 分類杜工部詩諺解, 四庫全書簡明目錄, 四庫全書總目提要, 宣和奉使高麗圖經, 新刊增補三略直解, 新增東國輿地勝覽, 新編諸宗敎藏總錄, 十七史纂古今通要, 五洲衍文長箋散稿, 天上天下唯我獨尊, 統理軍國事務衙門, 特進輔國三重大匡, 協辦交涉通商事務, 華東正音通釋韻考
9	4	0.01%	東國新續三綱行實圖, 阿耨多羅三藐三菩提, 風雲雷雨山川城隍壇, 欽定勝朝殉節諸臣錄
10	2	0.00%	統理交涉通商事務衙門, 火者據執田民推考都監
11	1	0.00%	北漢山新羅眞興王巡狩碑
계	73,466	100.00%	

[표12]에서 볼 수 있듯이, 2자로 된 표제한자의 비율이 거의 80%이고, 3자짜리 표제한자를 합치면 92%가 넘는다. 이는 한국어 어휘에서 2음절과 3음절의 비율이 높고, 인명과 관련된 어휘가 많은 데 기인한다. 4자짜리 표제한자는 약 6.5%를 차지하는데, "가가호호(家家戶戶), 개과천선(改過遷善)"과 같은 4자 성어가 눈에 띈다. 5자 이상의 표제한자 수의 비율은 급격히 줄어들며, "가덕도해전(加德島海戰), 한일협상조약(韓日協商條約), 관학유생응제(館學儒生應製), 상보국수숭록대부(上輔國崇祿大夫), 팔만사천대장경(八萬四千大藏經), 동양척식주식회사(東洋拓殖株式會社), 동국신속삼강행실도(東國新續三綱行實圖), 北漢山新羅眞興王巡狩碑" 등 '사건명, 제도명, 직명, 서명, 지명' 등 이른바 개체명(named entity)이 대부분을 차지한다. [표13]은 『통합』과 KorLex 간 연동 가능한 2자 이상 표제한자 어형의 수와 비율을 보여주는데, 『통합』과 KorLex는 공유하지만 『표준』에 없는 어형이 3개의 DB를 비교했을 때 277개가 검색되나 상세한 검토가 필요하다.[38)]

[표13] 『통합』과 KorLex 간 연동 가능한 2자 이상 어형의 수

어형의 글자 수	「통합」 - KorLex		「표준」 비등재 어형†
2	9,405	91.95%	247
3	641	6.27%	23
4	174	1.70%	7
5	8	0.08%	0
계	10,228	100.00%	277

둘째,『통합』과『표준』을 사상하려면 어의 분할이 비대칭적인 부분에 대해서도 상세한 검토가 필요하다. [표14]에 제시된 몇 가지 예를 살펴보자.

[표14] 『통합』과 『표준』 간 어의 분할의 비대칭성

연동 한자 어형	어의 수		공통 어의 수	비연계 어의	
	「통합」	「표준」		개체명	기타
忠肅	75	1	1	시호(諡號) 74개	
忠烈	73	1	0	시호(諡號) 70개, 현(縣) 이름 1개	형용사 2개
忠毅	68	1	1	시호(諡號) 67개	
文獻	70	2	1	시호(諡號) 66개, 자(字) 1개, 인명 1개	명사 1개
太初	38	1	1 (△)	연호(年號) 6개, 자(字) 29개	명사 2개
上下	31	6	3		〖白〗 4개, 동사 6개, 형용사 1개, 명사 17개
發明	21	3	3	자(字) 1개	동사 13개, 형용사 1개, 명사 3개
影響	12	1	0		동사 1개, 형용사 4개, 명사 7개

38) "肝脈, 囊子, 木莖" 등처럼 실제로 『표준』에 등재되지 않은 어형이 있는가 하면, "昇進, 尼斯今, 益母草, 一擧二得" 등처럼 『표준』에 DB에 등재되었으나, 『통합』- KorLex 간 공통 자료에는 나타나지 않는 경우가 있다. 후자는『통합』과 『표준』이 오랜 기간 만들어지고 디지털 콘텐츠화 되면서 발생한 코드 불일치 문제로 보인다. 따라서 『통합』과 『표준』간에는 [표12]보다 더 많은 수의 공통 어형이 있을 것으로 추정하며, 이 부분은 『통합』과 『표준』을 기반으로 만들어진 다른 언어자원을 사상할 때 세부적으로 해결해야 할 문제이다.

『통합』과 『표준』에 공통으로 수록된 표제어휘의 어형 중 어의 수가 가장 많은 예는 “충숙(忠肅), 충렬(忠烈), 충의(忠毅)” 등으로 각각 70개 내외의 어의를 포함하고 있으나, 『표준』에는 1개의 어의만 있어 비연계 비율이 가장 높다. 비연계되는 어의는 고유명사에 해당하는 개체명(named entity)이 대부분이다. 일반명사나 용언의 어간으로 분류될 수 있는 “문헌(文獻), 태초(太初), 상하(上下), 발명(發明), 영향(影響)” 등의 경우도 두 사전 간 어의의 비대칭성이 큰데, 개체명에 해당하는 어의 수가 대다수를 차지하는 “문헌(文獻), 태초(太初)”보다 “상하(上下), 발명(發明), 영향(影響)” 등과 같은 부류의 표제어휘가 『표준』을 보완할 여지가 있다.

셋째, 이종 언어자원의 통합에서 어휘망이나 개념망이 상위 개념을 형성하고, 다른 언어자원이 좀더 세부적인 어의를 나타내는 경우가 많아, 이종 언어자원 간 자료강화(enrichment)의 방식으로 많이 사용된다.[39] 『통합』과 *KorLex* 간 연동되는 표제한자의 어의가 *KorLex*의 계층구조에서 어떤 분포를 그리나 알아보자. 기본적으로 상위 계층에 포진한 개념은 명사로 표현되므로 *KorLexNoun*과의 연동성을 살펴보겠다.

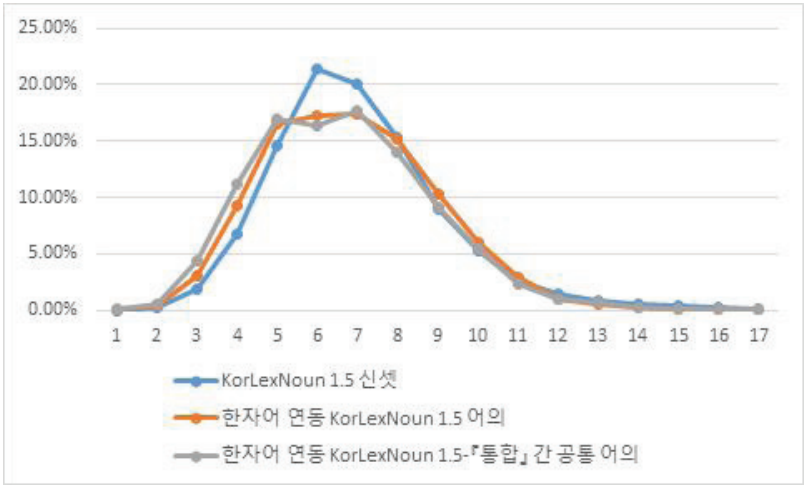
[표15] 한자어 정보가 연동된 KorLexNoun 1.5-『통합』 간 공통 어의의 계층별 분포

계층	Ⓐ KorLexNoun 1.5의 신셋 수†		Ⓑ 한자어가 연동된 KorLexNoun 1.5의 어의 수		Ⓒ 한자어가 연동된 KorLexNoun 1.5-『통합』 간 공통 어의 수	
1	9	0.01%	14	0.03%	8	0.05%
2	157	0.17%	178	0.33%	97	0.56%
3	1,653	1.83%	1,667	3.09%	763	4.41%
4	6,033	6.69%	4,979	9.23%	1,925	11.12%
5	13,129	14.57%	8,910	16.52%	2,934	16.95%
6	19,236	21.34%	9,291	17.23%	2,839	16.40%
7	18,079	20.06%	9,394	17.42%	3,058	17.66%
8	13,802	15.31%	8,178	15.17%	2,419	13.97%
9	8,053	8.93%	5,566	10.32%	1,584	9.15%
10	4,714	5.23%	3,264	6.05%	936	5.41%
11	2,305	2.56%	1,529	2.84%	397	2.29%
12	1,256	1.39%	546	1.01%	178	1.03%
13	733	0.81%	245	0.45%	113	0.65%
14	429	0.48%	97	0.18%	35	0.20%
15	346	0.38%	37	0.07%	13	0.08%
16	164	0.18%	18	0.03%	11	0.06%
17	36	0.04%	13	0.02%	2	0.01%
계	90,099.5	100.00%	53,914.5	100.00%	17,311.5	100.00%

『통합』과 한자어를 공유하는 *KorLexNoun*의 어의가 계층별로 나타나는 분포는 [표15]의 Ⓒ열과 [그림14]의 회색 그래프로 표시되는데, 한자어 정보가 부착된 *KorLexNoun* 1.5의 어의 분포(=[표15]의 Ⓑ열, [그림14]의 주황색 그래프) 및 한자어정보 유무와 관계없이

39) 실패한 경우에 대한 논문인 Poprat et al. (2013)을 참조해 보기 바란다.

KorLexNoun 1.5의 신셋이 계층별로 나타나는 분포(=[표15]의 Ⓐ열(=[그림14]의 하늘색 그래프)와 거의 유사하다. 이러한 분포의 유사성은 그동안 PWN이나 *KorLex*가 다른 언어자원을 보완하려고 했던 방식을 『통합』과의 연동에도 적용할 수 있는 가능성을 보여준다.



[그림14] 한자어 정보가 연동된 *KorLexNoun* 1.5-『통합』 간 공통 어의의 계층별 분포도 그래프

4.3. 『통합』 활용의 확장 가능성

본고의 목적은 편찬 동기, 개발 방식, 거시구조 및 미시구조가 완전히 다른 언어자원인 『통합』과 KorLex의 연동 가능성을 검토하여 각 언어자원의 활용성을 확대할 수 있는지를 알아보는 것이었다. 두 언어자원의 유일한 연계점은 한자로 된 표제어휘인데, 접점의 그물코가 매우 크고 성글기 때문에 현 상태의 그물로 당장 포착할 수 있는 상호 보완 결과는 풍성해 보이지 않는다. 하지만, 콘텐츠의 상호 보완을 위한 방안을 단계별로 실시하면서 그물망의 코를 점점 촘촘하게 만들 때, 양과 질에서 풍요롭고 다양한 결과를 얻을 수 있을 것이다.

첫째, 『통합』을 일반에게 공개하여 다양한 연구자와 필요한 사람들이 『통합』이 제공할 수 있는 정보를 소상히 알 수 있도록 하고, 자신들이 필요한 정보를 확장하도록 하는 것이다. 지금은 『한국』만 네이버의 사전 플래트폼에서 일반 사용자에게 서비스되고 있다. 내부 연구자가 아니면 『한한』을 포함한 『통합』을 사용할 수 없어, 『통합』의 활용성 확장에 가장 큰 걸림돌이 되고 있다. 사실 오랜 기간에 걸쳐 많은 수의 전문가 집단이 엄청난 노력을 기울여 만든 『통합』과 같은 귀중한 디지털콘텐츠를 조건 없이 공개하기란 쉽게 결정할 수 있는 문제가 아니다. 같은 문제로 고심했던 다른 사전이나 언어자원의 예를 참고한다면, 공개 여부의 결정과 공개 방식에 대한 계획을 짜는 데 도움이 될 것이다. 아무리 풍부한 정보를 가진 언어자원이라도 종이사전으로 형태로만 남거나 디지털화하였더라도 일반에게 공개하지 않은 사전은 디지털/모바일 시대에 더 이상 유통되지 않고 도서관 서고나 모니터 뒤편에 남아 우리의 기억에서 사라지고 있다는 점을 반면교사로 삼을 필요가 있다.

『표준』의 경우, 초기에는 저작권 문제 때문에 공개되지 못하다가, 그 문제를 해결하면서

일반 사용자에게는 국립국어원 누리집과 전자사전 플랫홈 등을 통해 검색 기능을 제공하고, 한국어처리 분야의 기술력 향상을 견인했을 뿐 아니라 『우리말샘』같은 후속 사전의 개발을 가속화할 수 있었다. *KorLex*는 PWN의 무료배포 정책에 힘입어 개발됐으나, 2000년대 중반부터 국내 대학의 산학협력단이 설치되고 연구결과물 규정과 같은 정책을 신설하면서 완전한 무료 배포에는 제동이 걸린 상황이었다. *KorLex* 연구진은 버전에 따라 무료와 유료 배포로 구분했고, 무료 배포 버전인 *KorLexNoun* 1.0은 아무 제한 없이 사용할 수 있도록 하고, *KorLex* 홈페이지나 PWN 관련 웹사이트를 통해 누구나 내려받을 수 있도록 했다. 유료 배포 버전인 *KorLex* 1.5는 사용 목적에 따라 연구용과 수익용으로 구분하여, 연구용은 큰 경제적인 부담 없이 디지털콘텐츠 전체를 사용할 수 있도록 했다. 또한 본 연구진이 *KorLex* 1.5의 검색 사이트를 개발하여 일반 사용자가 조건 없이 이용할 수 있도록 제공한다.

둘째, 기존에 편찬된 언어사전의 정보를 『통합』으로 보완하는 결과물 중 예상할 수 있는 유형은 일반 어휘의 역사적 정보를 추가하는 것이다. 본고 4.2절의 [표14]에서 보았듯이 『통합』과 『표준』 사이에는 1개 어형에 수록된 어의의 비대칭성이 크기 때문에, 국립국어원도 동양학연구원과 협력하면 『통합』에 수록된 정보를 이용하여 기존의 사전을 풍요롭게 보완할 수도 있고, 새로운 사전을 기획할 수도 있다는 장점이 있을 것이다.

예를 들어, [그림15]에서 보는 것처럼 『표준』에서 "발명(發明)"의 어의는 3개로, 정의문/설명문을 기준으로 할 때 [표16]의 『통합』에 수록된 21개 어의 중 [11] (또는 [16]), [20], [21]과 사상될 수 있다. 『통합』의 [20]과 『표준』"발명[2]"의 「1」처럼 정의문이 거의 동일하여 사상되는 어의에서는 15세기부터 19세기까지 용례로 역사성을 보완할 수 있을 것이다. 또한, 『통합』의 [11] (또는 [16])이 『표준』 "발명[1]"의 어원인지 밝힐 수 있는 자료를 제공할 수도 있을 것이다. 참고로, 두 사전과 전혀 다른 배경과 목적에서 시작된 18세기말~19세기초 개항기에 편찬된 사전에서 같은 표제어휘를 살펴보자.[40] [그림16]은 최초의 한국어 이중어사전인 『한불ᄌᆞ뎐』과 이 사전을 이어받은 『한영자뎐』의 검색 결과를, [그림17]은 『조선어사전』의 검색 결과를 보여준다.[41] 『한불ᄌᆞ뎐』에서는 "發明ᄒᆞ다"에 대응되는 프랑스어로 "pallier, s'excuser, se disculper, nier, disconvenir, se justifier"를 제시하여 [표16]의 [20]에 해당하는 어의를 설명한다. 하지만 『한영자뎐』에서는 그 대응어로 "to make clear"와 "to prove"를 제시하여 [표16]의 [5]/[10]과 [2]에 해당하는 어의를 나타낸다. 반면, [11]/[16]에 해당하는 어의는 영어 대응어로 "to discover"와 "to invent"를 제시하는 "新發明(ᄒᆞ다)"에서 찾아볼 수 있다. 『조선어사전』에는 2개의 어의가 나타나는데, (二)는 『표준』 "발명[2]"의 「1」에 대응되며, 유의어로 제시된 "暴白"은 『표준』의 동일 어의에도 제시되고, 예문인 "發明無路"가 『표준』에서는 독립된 표제어로 수록되었다. 『조선어사전』에서 어의 (一)의 뜻풀이인 "詳演辨析(='자세히 설명하고 꼼꼼히 분석한다.')"은 『통합』의 [2] 내지 [5]에 해당한다.[42]

40) "개항기"와 "개화기"는 19세기 중반부터 20세기 초반을 일컫는 명칭으로 함의하는 바가 다르나, 본고에서는 구분을 두지 않고 단순히 해당 시기의 명칭으로 사용하겠다.

41) 본 연구진의 일부는 부산대학교 인문학연구소를 도와 개항기에 발간된 초기 한국어 사전인 『한불ᄌᆞ뎐』(1880년 발간), 『한영자뎐』(1911년 발간), 『조선어사전』(1917년 추정, 조선총독부 편저)을 디지털화하여, 이를 웹(http://corpus.pusan.ac.kr)에 공개하고 유지관리하고 있다.

42) "詳演辨析"의 의미는 부산대학교 한문학과 교수의 자문을 받았다.

[그림15] 『표준』에서 “발명”의 어의 정보

[표16] 『통합』에서 “발명”의 어의 정보

【發明 발명】
[1] 눈과 귀를 밝게 함. 총명하게 함.

《文選, 宋王, 風賦》 淸淸泠泠, 愈病析酲. 發明耳目, 寧體便人. 〈呂延濟注〉 發, 開也. 言能開耳目之明.
《後漢書, 馬融傳》 若乃陽阿衰斐之晉制, 闡鼗華羽之南音, 所以洞蕩匈臆, 發明耳目.
[2] 설명함. 증명함. 표명함.

《史記, 商君傳》 且所因由嬖臣, 及得用, 刑公子虔, 欺魏將卬, 不師趙良之言, 亦足發明商君之少恩矣.
晉, 干寶 《搜神記序》 及其著述, 亦足以發明神道之不誣也.
明, 胡應麟 《少室山房筆叢, 史書佔畢 3, 冗篇 上》 采薇一歌, 足發明武未盡善.
朝鮮, 李光庭 《(挽章)又》 洛波輟餘響, 異論間汶颺, 後賢更發明, 雄辯互頡頏.
[3] 비교하고 대조하여 사실과 부합함을 증명함.

南朝梁, 沈約 《上注制旨連珠表》 連珠者, 蓋謂辭句連續, 互相發明, 若珠之結排也.
金, 王若虛 《五經辨惑 1》 經傳之間可以互相發明者多矣, 是故聞見貴乎博也.
[4] 수립함. 또는 건의함.

《史記, 張丞相傳》 高陵侯趙周等爲丞相. 皆以列侯繼嗣, 娖娖廉謹, 爲丞相備員而已, 無所能發明功名有著於當世者.
唐, 張九齡 《謝中書侍郎狀》 臣謬跡書府, 兼司綸翰, 思力淺近, 無所發明.
[5] 드러내어 밝힘. 명백히 논술함.

《史記, 孟子荀卿傳》 [愼到等] 皆學黃老道德之術, 因發明序其指意.
《後漢書, 徐防傳》 臣聞詩·書·禮·樂, 定自孔子, 發明章句, 始於子夏.
宋, 蘇轍 《歐陽文忠公神道碑》 公於六經長易·詩·春秋, 其所發明, 多古人所未見.
朝鮮, 李玄逸 《答丁君翊》 惟此訓釋, 已發明程子所言之義, 大煞分明.
[6] 생각이나 감정을 나타냄. 표현함.

《史記, 儒林傳》 寬爲人溫良, 有廉智, 自持, 而善著書·書奏, 敏於文, 口不能發明也.
《朝鮮太宗實錄 15, 8年3月戊午》 韓仲老敢以私物, 納諸櫃內, 見於朝廷, 恕於詰罪之際, 不卽發明, 回還本國, 又不啓達.

【發明 발명】

[7] 선양(宣揚)함. 현양(顯揚)함.

北齊, 顔之推 《 顔氏家訓, 文章 》 敷衍仁義, 發明功德, 牧民建國, 施用多途.

唐, 張九齡 《 爲何給事進亡父所著書表 》 臣亡父所論君臣之際, 必欲驗之行事, 非眞垂於空文, 誠宜上感宸衷, 由沒代而匡輔, 下藏秘府, 因聖君以發明.

宋, 曾鞏 《 贈職方員外郎蘇君墓誌 》 古之人亦不必皆能自見而卒有傳於後者, 以世有發明之者耳.

《 朝鮮太祖實錄 14, 7年5月己未 》 若上根之人, 不日成功, 發明大智.

[8] 분명함. 또는 분명히 앎.

唐, 張九齡 《 賀御注金剛經狀 》 臣雖愚昧, 本自難曉, 伏覽睿旨, 亦卽發明. 是知日月旣出, 天下普照, 誠在此也.

明, 方孝孺 《 答俞子嚴書 》 人苟能發明六經者, 大之於天下國家, 小之於善一已, 直易易耳.

[9] 들추어냄. 폭로함.

唐, 白居易 《 與昭義軍將士詔 》 ［盧從史］乃外示恭順, 內懷姦邪, 刻削軍中, 暴殄境內. 朕以君臣之道未忍發明, 爲之含容.

《 舊唐書, 憲宗紀 上 》 李錡屬列宗枝, 任居方伯……僚佐以獻規受屠, 王臣以傳命見脅. 朕切於含垢, 未忍發明, 累降中人, 令遵前旨.

[10] 명시함. 알려 줌.

唐, 李公佐 《 謝小娥傳 》 初, 父之死也, 小娥夢父謂曰, 殺我者, 車中猴, 門東草. 又數日, 復夢其夫謂曰, 殺我者, 禾中走, 一日夫. 小娥不自解悟……余備詳前事, 發明隱義, 暗與冥會, 符於人心.

清, 袁枚 《 隨園隨筆, 韓仲良碑 》 韓瑗之父仲良之碑, 唐書無傳……許敬宗等誣以不道, 至於削爵籍家, 子孫流竄. 所以碑亦斲削其姓名. 是不可不爲之詳考而發明之也.

[11] 새로운 사물이나 이치를 찾아냄.

《 三國志, 魏志, 和洽傳 》 洽同郡許混者, 許劭子也. 〈裴松之注〉 劭始發明樊子昭於鬻幘之肆, 出虞永賢於牧豎, 召李叔才鄉閭之間, 擢郭子瑜鞍馬之吏, 援楊孝祖, 擧和陽士.

《 舊唐書, 姚璹傳 》 則天又令洛州長史宋元爽, 御史中丞霍獻可等重加詳覆, 亦無所發明.

[12] 빛을 냄. 또는 능력을 발휘함.

宋, 梅堯臣 《 挑燈杖詩 》 油燈方照夜, 此物用能行. 焦首終無悔, 橫身爲發明.

宋, 張戒 《 歲寒堂詩話, 上 》 韓退之之文, 得歐公而後發明.

[13] 소서(小序).

宋, 文瑩 《 玉壺淸活 3 》 柳仲塗開知潤州, 胡旦祕監爲淮漕. 二人者具喜以名驁於時. 旦造漢春秋編年, 立五始先經·後經·發明·凡例之類, 切侔聖作.

[14] 어울려서 빛남. 서로 잘 어울림.

宋, 龐元英 《 談藪 3 》 平湖當前, 數十百頃. 其外連山橫陳, 樓觀森列, 夕陽返照, 丹碧紫翠, 互相發明.

[15] 생성하여 기름. 배양함.

《 西遊記, 1回 》 覆載羣生仰至仁, 發明萬物皆成善.

《 西遊記, 98回 》 如來對唐僧言曰, 此經功德不可稱量……蓋此內有成仙了道之奧妙, 有發明萬化之奇方也.

[16] 새로운 사물이나 방법을 창조함.

《 二十年目睹之怪現狀, 81回 》 不知某觀察的這個提油新法, 是那一國人·那一個發明的.

[17] 전설상의 신조(神鳥) 이름.

《 後漢書, 五行志 2 》 爲孽者四. 〈劉昭注〉 似鳳有四, 並爲妖, 一曰鸑鷟……二曰發明, 烏喙, 大頸, 大翼, 大脛.

[18] 새벽녘에 우는 봉황의 울음소리.

漢, 劉向 《 說苑, 辨物 》 天老曰, 夫鳳……晨鳴曰發明, 晝鳴曰保長, 飛鳴曰上翔, 集鳴曰歸昌.

《 宋書, 符瑞志 中 》 鳳凰者, 仁鳥也……晨鳴曰發明, 晝鳴曰上朔, 夕鳴曰歸昌, 昏鳴曰固常, 夜鳴曰保長.

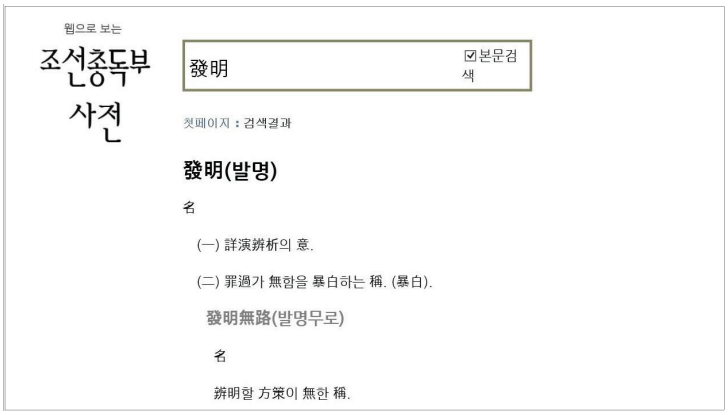
[19] 송(宋) 진개(陳開)의 자(字).

[20] 죄나 잘못이 없음을 말하여 밝힘.

《 朝鮮成宗實錄 240, 21年5月丁卯 》 照律時, 盡舍前日發明獄辭, 一從元推照律, 有乖哀矜之意, 要在聖上參情法, 酌輕重而已.

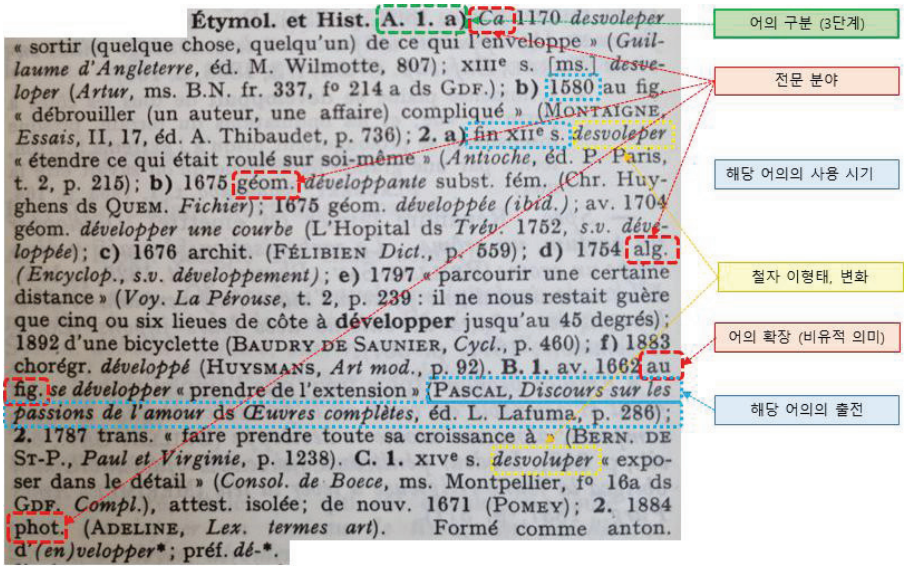
[21] 경서의 뜻을 스스로 깨달아서 밝힘.

[그림16] 『한불ᄌᆞ뎐』과 『한영자뎐』에서 "발명" 관련 표제어휘 및 어의



[그림17] 『조선어사전』에서 "발명" 관련 표제어휘 및 어의

어휘의 역사적 정보를 상세히 수록한 대표적인 사전 중 하나로 *Trésor de la langue française*(『프랑스어의 보고(寶庫)』, 이하 TLF)를 들 수 있는데, [그림18]은 TLF에서 표제어휘 "développer"의 어원과 역사적 정보가 수록된 부분이다. 최대 3단계로 세분화된 어의별로 사용되는 전문분야, 출전과 사용 시기, 철자의 변화, 비유적 의미로의 확장 여부 등의 정보를 상세하게 보여준다. 한국어를 대상으로 이러한 유형의 사전을 편찬하고자 한다면, 『통합』에 담긴 콘텐츠는 매우 중요한 역할을 할 것이다.



[그림18] TLF에서 표제어휘 "développer"의 어의별 어원 및 역사적 정보

셋째, 현재로선 『통합』과 *KorLex*를 연동해 활용의 범위를 늘이는 방법으로는 두 가지 정도를 생각해 볼 수 있다. 하나는 기존의 전문용어 사전과 PWN/*KorLex*를 연동할 때 사용했던 방법처럼, *KorLex*를 상위 온톨로지로 삼고, 『통합』의 풍부하고 상세한 개체명 등을 하위 노드로 구성하는 것이다. PWN과 항공 분야 전문용어 확장, *KorLex*와 현대 한국어의 개체명 확장, *KorLex*와 안전 분야 전문용어 확장 등 다양한 활용이 있었다. [표15]에서 한자어 정보를 이용한 『통합』과 *KorLex* 간 연계점 분포가 *KorLex*의 신셋 분포와 비례한다는 것을 보았다. *KorLex*와 현대 한국어의 개체명 간 연동할 때 *KorLexNoun*의 4단계~6단계를 상위 온톨로지로 사용했는데, *KorLex*와 『통합』 개체명을 연동해야 한다면 기존 작업의 절차와 방법론이 매우 유용할 것이다.[43]

다른 하나는 *KorLex*를 『통합』의 분류명이나 태그명을 정제하거나 정규화하는 참조 준거로 삼는 것이다. [표5]에서 봤듯이, 『통합』에는 '조대, 저자, 출전, 예문, 땅 이름, 현 이름'과 같은 분류명이 사용되었는데, 태그셋(tag set)이나 메타언어 정의가 정규화되지 않은 것 같다. 하지만 종이사전 편찬 시 일러두기나 상세한 지침 등이 있어, 이를 정비하여 분류명 등을 태그셋으로 명시적 정의를 한다면 추후 확장과 활용 및 표준화에 도움이 될 것이다.[44] 메타구조와 태그셋의 기술은 [표17]처럼 LexML을 사용할 수 있는데, LexML에서 정의되지 않은 어의의 세부적인 요소(element)와 속성(attribute)을 규정하는 데 *KorLex*를 이용할 수 있을 것이다.[45] 그 경우, PWN을 통해 다국어 연동성을 자동으로 확보할 수 있다.[46]

---

43) 언어자원 간 사상 방법에 대해서는 박흠.윤애선(2011), 배선미 외(2010)를 참조하라.
44) LexML 및 이를 이용한 메타언어 정의 방법에 대한 논문은 윤애선.정휘웅(2006), 윤애선(2009, 2011)을 참조하라. [표17]은 윤애선(2011:390-392)의 <표10> 중 일부를 발췌했다.
45) LexML의 전신인 LMF와 PWN의 연동을 위한 표준화 규정은 Vossen et al. (2013)을 참조하라.
46) 중국어 대상으로 개념망인 HowNet을 구축한 Dong & Dong (2006)의 연구도 참조하라. HowNet과 PWN을 연동했다고 하나, 아쉽게도 그 결과 자료를 구할 수 없다.

[표17] LexML을 이용한 『한불ᄌᆞ뎐(1880)』과 『한영자전(1911)』의 통합적 자료구조 정규화 규칙 (발췌)

(1)	<DictionaryEntry>	::==	ENTRYid <HeadwordGroup> <SenseGroup> [<DeveloperComment>]
(2)	<HeadwordGroup>	::==	<HeadWord> (<Transliteration> | <Pronunciation>) [<TranslationBlock|TranslationCtn>] [<InflectionBlock|InflectionCtn>] [<CaseMarkerBlock|CaseMarkerCtn>] [<OrthographyBlock|OrthographyCtn>] [<DeveloperComment>]
(3)	<SenseGroup>	::==	PartOfSpeech <SenseBlock|SenseCtn>$^{+}$ [<DeveloperComment>]
(4)	<AntonymBlock>	::==	<AntonymCtn>$^{+}$ [<DeveloperComment>]
(5)	<CaseMarkerBlock>	::==	<CasemarkerCtn>$^{+}$ [<DeveloperComment>]
(6)	<CollocationBlock>	::==	<CollocationCtn>$^{+}$ [<DeveloperComment>]
(7)	<DefinitionBlock>	::==	<DefinitionCtn>$^{+}$ [<DeveloperComment>]
(8)	<DerivationBlock>	::==	<DerivationCtn>$^{+}$ [<DeveloperComment>]
(9)	<ExampleBlock>		<ExampleCtn>$^{+}$ [<DeveloperComment>]
(10)	<ExlplationBlock>	::==	<ExlplationCtn>$^{+}$ [<DeveloperComment>]
(11)	<InflectionBlock>	::==	<InflectionCtn>$^{+}$ [<DeveloperComment>]
(중략)			
(27)	<LinkCtn>	::==	<Link> [<DeveloperComment>]
(28)	<OrthographyCtn>	::==	(<OrthographyVariant> | <FullForm> | <Abbreviation>)$^{+}$ [<LinkBlock|LinkCtn>] [<DeveloperComment>]
(29)	<RegisteCtn>	::==	<Register> [<LinkBlock|LinkCtn>] [<DeveloperComment>]
(30)	<SeeAlsoCtn>	::==	<SeeAlso> [<LinkBlock|LinkCtn>] [<DeveloperComment>]
(31)	<SenseCtn>	::==	SENSEid SenseGrainSize (<TranslationBlock|TranslationCtn> | <ExlplationBlock|ExplanationCtn> | <DefinitionBlock|DefinitionCtn>)$^{+}$ [<DerivationBlock|DerivationCtn>] [<CollocationBlock|CollocationCtn>] [<SynonymBlock|<SynonymCtn>] [<AntonymBlock|AntonymCtn>] [<SeeAlsoBlock|SeeAlsoCtn>] [<OrthographyBlock|OrthographyCtn>] [<RegisterBlock|RegisterCtn>] [<Reference>] [<Style>] [<Domain>] [<ExampleBlock|ExampleCtn>] [<DeveloperComment>]
(32)	<SynonymCtn>	::==	<Synonym> [<LinkBlock|LinkCtn>] [<DeveloperComment>]
(33)	<TranslationCtn>	::==	<Translation> [<DeveloperComment>]
(중략)			

5. 이어가기

본고의 목적은 편찬 동기, 개발 방식, 거시구조 및 미시구조가 완전히 다른 언어자원인 『통합』과 KorLex의 연동 가능성을 검토하여, 각 언어자원의 활용성을 확대할 수 있는지를 알아보는 것이었다. 두 언어자원의 유일한 연계점은 한자로 된 표제어휘인데, 접점의 그물코가 매우 크고 성글기 때문에 현 상태의 그물로 당장 거둬 올릴 수 있는 유용한 결과는 크지 않다. 하지만, 두 언어자원에 연계점을 제공하는 『표준』과 『통합』의 연동성은 좀더 가시적이고, 기존 한국어 사전의 보완 또는 편찬 방향에 시사점을 찾을 수 있었다. 예를 들어, 『통합』의 풍부한 정보를 이용하면, 한국어 사전에서 아직 결여된 부분인 어휘의 역사성을 상당히 보완할 수 있을 뿐 아니라, 앞으로 편찬될 사전에 새로운 방향을 제시할 수 있을 것이다. 이러한 새로운 시도에 디지털화된 다양한 한국어 관련 사전이 동참한다면 좀더 풍요로운 수확을 거둘 수 있을 것이다.

필자를 포함한 본 연구진은 지난 30여 년간 자연언어처리 연구와 관련 소프트웨어 개발을 하면서 우리에게 필요한 언어자원을 직접 구축하거나, 다른 전문가나 기관에서 만든 언어자원을 정제 및 가공해서 사용해 왔다. 또한 이종 언어자원의 연동을 통해 중요한 언어정보를 효율적으로 구축할 수 있었다. 이러한 경험이 처음부터 가시적인 결과를 얻기 힘들 것이라고 예상했던 이 논문 발표를 수락했던 계기다. 또한 '사전'이라는 공통 주제어로 다양한 사전을 편찬하고 연구하는 사전학회가 구심점이 되어 이룰 수 있는 여러 가지 과제 중 하나를 제안하고 싶기도 했다.

4절에서 잠시 언급했던 『한불ᄌᆞ뎐(1880)』, 『한영자뎐(1911)』, 『조선어사전(1917)』은 본 연구진이 부산대학교 인문학연구소와 협업하여 디지털 콘텐츠화한 사전이다. 세 사전 모두 원본인 종이사전이 인쇄된 상태이긴 했으나 문자자동인식 프로그램을 사용할 수 없는 경우라서 사람이 직접 입력해야 했다. 『한불ᄌᆞ뎐(1880)』과 『한영자뎐(1911)』은 입력부터 배포까지 부산대학교가 수행했다. 『조선어사전(1917)』의 경우, 원본의 1차 디지털화는 20년 전쯤 S 대학교에서 시작했고, 이를 이어받아 그로부터 10년 후쯤 Y 대학교에서 입력을 마치고, 부산대학교에 전해져 입력된 자료를 정규화하여 앞선 두 사전과 함께 웹에 공개한 것이다. 『조선어사전(1917)』의 디지털화가 3단계를 거치게 된 가장 큰 이유는 1, 2단계를 주도한 연구자가 정년퇴임하면서, 소속기관과는 관계없이 다음 단계로 나아가게 할 수 있는 가장 적합한 후임 연구자에게 과업을 넘겨주었기 때문이다. 세 번째 단계를 진행했던 필자도 다음 연구진을 찾아야 할 단계에 왔다. 앞에 언급한 3개 사전을 웹에 공개했지만, 활용도는 2017년 이후 점차 떨어지고 있다.[47] 한국의 개화기를 연구하는 전문가의 수가 많지 않기도 하나, 가장 큰 원인은 다른 언어자원과의 통합 검색 등 인문학 연구자들이 사용하는 데 편리한 연계 기능을 제공하지 못하기 때문으로 사료된다. 이는 비단 3개 사전만의 문제가 아니라 다수의 한국어 관련 사전이 안고 있는 공통적인 문제이기도 하다. 각각의 사전은 엄청난 노력과 시간과 비용이 투여된 지식의 산물이다. 대개는 그것을 편찬한 연구진이 내부 자료로 사용하고 있거나, 해당 연구진의 주도 하에 일반에 공개되고 있다. 하지만 그 활용도는 그리 높지 않고, 오히려 네이버나 다음의 사전 플랫홈을 통해 접할 수 있는 사전의 이용이 더 활발하다.

웹사이트를 통해 검색 서비스를 제공하는 방식의 결정은 한편으로 언어자원의 보호와 다른 한편으로 배포성 및 관리의 용이성이라는 두 가지 상반된 요소를 잘 고려해야 한다. 디지털 콘텐츠의 개발자가 검색 서비스를 제공한다면, 콘텐츠를 보호할 수는 있지만 지속적인 관리나 배포의 효율은 낮다.[48] 다른 전문적인 플랫홈에 맡긴다면 널리 배포될 수 있고 관리가 용이하겠지만, 자칫 플랫홈의 운영자가 콘텐츠의 주도권을 쥘 수도 있다. 차제에 사전학회에 건의하고 싶은 바는 윅쇼너리(Wiktionary)나 깃허브(GitHub)처럼 공통 검색하거나 자료를 공유할 수 있는 플랫홈(open source platform)을 장기적으로 운영할 수 있는 방안을 함께 모색하고 실천하는 것이다.[49]

47) 세 사전의 디지털화는 부산대학교 인문학연구소가 참여했던 인문한국(HK) 1차 사업(2007-2017)에서 지원을 받아 이루어졌다. 이 과제의 주요한 연구주제 중 하나가 '한국의 개화기'였으므로, 세 사전은 과제의 참여자들뿐 아니라 주변 연구자들에게 연구 자료로 활용되었다.

48) 이 발표논문을 준비하면서 PWN 및 파생 워드넷을 개발하여 제공해왔던 다수의 웹사이트가 더 이상 운영되지 않는 것을 확인했다. 대부분 연구와 개발을 주도하던 연구자가 은퇴하거나 이직하면서 자료 검색을 제공했던 웹사이트가 폐쇄된 것으로 보인다.

49) 이러한 점에서 『표준』, 『우리말샘』 등을 편찬하고, 독자적인 검색 플랫홈을 운영하는 동시에 다른 검색 사이트에 디지털콘텐츠를 제공하는 국립국어원과 긴밀하게 협력하면 상기 두 조건을 어느 정도 충족할 수 있을 것으로 보인다.

참고 문헌

1. 1차 자료가 탑재된 웹사이트

국립국어원 사전

『우리말샘』: https://opendict.korean.go.kr/main

『표준국어대사전』: https://stdict.korean.go.kr/main/main.do

『한국어기초사전』: https://krdict.korean.go.kr/mainAction

『한국수어사전』: https://sldict.korean.go.kr/front/main/main.do

네이버(Naver) 사전 플랫홈: https://dict.naver.com/

다음(Daum) 사전 플랫홈: https://dic.daum.net/

단국대학교 동양학연구원: https://www.dankook.ac.kr/web/ins1

『한한대사전(漢韓大辭典)』: https://hanja.dict.naver.com/#/main (구성사전 중 하나)

부산대학교 인공지능연구실 NLP 플랫홈: http://corpus.pusan.ac.kr/

『조선어사전(1917)』: http://corpus.pusan.ac.kr/csdic/ default.aspx (웹으로 보는 조선총독부 사전)

『한불즈뎐(1880)』&『한영자뎐(1911)』: http://corpus.pusan.ac.kr/dicSearch/Default.aspx (지능형개화기 한국어 사전)

한국어 맞춤법/문법 검사기(KSGC): http://urimal.cs.pusan.ac.kr (우리말 배움터)

KorLex : http://korlex.pusan.ac.kr

영어 워드넷 및 파생 워드넷 플랫홈

Global WordNet Association: http://globalwordnet.org

Open Multilingual WordNet: http://compling.hss.ntu.edu.sg/omw

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

2. 2차 자료

김민호・최현수・권혁철・윤애선(2014), “한국어 어휘의미망을 이용한 문맥 의존 철자오류 교정규칙의 일반화”, 『정보과학회논문지: 컴퓨팅의 실제 및 레터』 20(2) (한국정보과학회), 106-110.

김지영(2013), “『통합디지털한한대사전』의 DB구축과 온라인 사전편집기”, 『동양학』 54 (단국대학교 동양학연구원), 213-246.

김철웅(2021), “동양학연구원 50년의 성과와 과제”, 『동양학』 82 (단국대학교 동양학연구원), 169-187.

박흠・윤애선(2011), “Automatic Mapping Between Large-Scale Heterogeneous Language Resources for NLP Applications: A Case of Sejong Semantic Classes and KorLexNoun for Korean”, 『언어와 정보』 15(2) (언어정보학회), 23-45.

배선미, 임경업, 윤애선(2010), “인간언어공학에의 활용을 위한 이종 개념체계 간 사상 - 세종의미부류와 KorLexNoun 1.5-”,『인지과학』 21(1) (한국인지과학회), 95-126.

심경호(2018), “한자사전의 현재적 의미와 개선 방안”, 『동양학』 71 (단국대학교 동양학연구원), 51-74.

윤승준(2012), “『한한대사전』의 편찬과정과 향후 계획”, 『동양학』 52 (단국대학교 동양학연구원), 147-166.

윤애선(2007), “국내・외 어휘의미망의 구축과 활용”, 『새국어생활』 17(3) (국립국어원),

5-25.
윤애선(2009), “지식베이스 구축을 위한 『한불ᄌᆞ뎐(1880)』 <어휘부>의 미시구조 분석”, 『불어불문학연구』 78 (한국불어불문학회), 263-304.
윤애선(2010a), “인간언어공학에의 활용을 위한 표준국어 대사전과 세종전자사전 간 용언어의 사상”, 『언어학』 56 ((사)한국언어학회), 197-235.
윤애선(2010b), “단일어 워드넷을 넘어 다국어 워드넷으로”, 『일본어문학』 46 (한국일본어문학회), 3-31.
윤애선(2011), “LEXml을 이용한 『한영자전(1911)』의 지식베이스 설계 - 『한불ᄌᆞ뎐(1880)』과의 통합적 지식베이스 구축을 위하여-”, 『불어불문학연구』 87 (한국불어불문학회), 343-399.
윤애선(2012), “한국어 어휘의미망 KorLex 2.0 –의미 처리와 지식 공학을 위한 기반 언어자원”, 『한글』 295 (한글학회), 163-201.
윤애선(2016a), “근대 동아시아 어휘사전의 탄생과『한불ᄌᆞ뎐』”, 『한국사전학회 학술대회 발표논문집』, (한국사전학회), 63-99.
윤애선(2016b), “개항기 전후 한국어 사전의 지식베이스 구축－한불ᄌᆞ뎐(1880)에서 웹으로 보는 한불자뎐(2009)과 현대 한국어로 보는 한불자전(2014)으로”, 『한국사전학』 27 (한국사전학회), 7-48.
윤애선(2016c), “정보학, 인문학의 숨겨둔 서고를 여는 또 하나의 열쇠 – 인문정보학 센터의 성과와 전망”, 『코기토』 80, (부산대학교 인문학연구소), 36-66.
윤애선(2019), “『디지털 바벨탑』 세우기, 어디까지 왔나? -프-한 기계번역의 현황과 전망”, 『불어불문학연구』 117, (한국불어불문학회), 157-199.
윤애선・권혁철(2020), “以Princeton WordNet爲通用語言的韓語詞匯語義網KorLex的特性及應用”, 『한자연구』 26 (경성대학교 한국한자연구소), 1-30.
윤애선・정휘웅(2006), “전자사전 국제기술표준 LEXml의 정합성 및 확장성 - 단일어 및 다국어 사전에의 적용”, 『한국프랑스어논집』 54 (한국프랑스학회), 55-96.
윤애선・황순희・이은령・권혁철 (2009), “한국어 어휘의미망 「KorLex 1.5」의 구축”, 『정보과학회 논문지: 소프트웨어 및 응용』 36-1 (한국정보과학회), 92-108.
에이든, E.・미셸, J., (2015), 김재중 옮김, 『빅데이터 인문학－진격의 서막(Uncharted : Big Data As a Lens on Human Culture)』, 사계절.
이은령・황순희・윤애선(2004), “다국어 어휘의미망 구축의 현황과 문제점 - PWN과 EWN을 중심으로 -”, 『프랑스문화예술연구』 12 (프랑스문화예술학회), 369-400.
최현수・권혁철・윤애선(2015a), “동적 윈도우를 갖는 조건부확률 모델을 이용한 한국어 문맥의존 철자오류 교정규칙의 재현율 향상”, 『정보과학회논문지』 42(5) (한국정보과학회), 629-636.
최현수・윤애선・권혁철(2015b), “통합적 제약완화 방식을 통한 한국어 문맥의존 철자오류 교정규칙의 재현율 향상”, 『정보과학회 컴퓨팅의 실제 논문지』 21(6) (한국정보과학회), 412-417.
홍재성 외(2007), 『21세기 세종계획 전자사전 개발 연구보고서 (11-1370252-000063-10)』, 문화관광부, (국립국어원).
황순희・권혁철・윤애선(2010), “한국어 수분류사 어휘의미망 KorLexClas 1.5”, 『정보과학회논문지: 소프트웨어 및 응용』 37(1), (한국정보과학회), 60-73.
Baccianella, S., A. Esuli & F. Sebastiani (2010), “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining”. *in Proceedings of the 7th Conference on Language Resources and Evaluation* (LREC '10), 2200 – 2204.

Bond, F. & F. Ryan (2013), "Linking and Extending an Open Multilingual Wordnet", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1352 – 1362.

Bond F. & K. Paik (2012), "A survey of wordnets and their licenses", *Proceedings of the 6th Global WordNet Conference* (GWC 2012), 64 – 71.

Deng, J. & W. Dong, R. Socher, L. Li, K. Li & L. Fei-Fei (2009), "ImageNet: A Large-Scale Hierarchical Image Database", *in Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition,* (DOI: 10.1109/CVPR.2009.5206848).

Dong, Z. & Q. Dong (2006), *HowNet and the Computation of Meaning*, World Scientific.

Fellbaum, Ch. (ed.) (1998), *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge.

Fellbaum, Ch. (2005), "WordNet and wordnets", In: Brown, Keith et al. (eds.), *Encyclopedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, 665-670.

Fellbaum, Ch. & P. Vossen (2012), "Challenges for a multilingual wordnet". *Language Resources and Evaluation* 46 (2), 313 – 326.

Gangemi, A., R. Navigli & P. Velardi (2003), "The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet", *in Proceedings of International Conference on Ontologies, Databases and Applications of SEmantics* (ODBASE 2003), 820 – 838.

Gangemi, A., N. Guarino, C. Masolo, & A. Oltramari (2003), "Sweetening WordNet with DOLCE", *in AI Magazine* 24(3), 13 – 24.

Miller, G. A. & Ch. Fellbaum (2009), WordNet then and now, *Language Resources & Evaluation* 41, 209-214.

Pease, A., I. Niles & J. Li (2002), "The suggested upper merged ontology: A large ontology for the Semantic Web and its applications", *in Proceedings of the AAAI-2002 Workshop on Ontologies and the Semantic Web,* (available at https://www.aaai.org/Papers/Workshops/2002/WS-02-11/WS02-11-011.pdf).

Ponzetto, S. & R. Navigli (2009), "Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia", *in Proceedings of the 21st International Joint Conference on Artificial Intelligence* (IJCAI 2009), 2083 – 2088.

Poprat, M., E. Beisswanger & U. Hahn (2008), "Building a BIOWORDNET by Using WORDNET's Data Formats and WORDNET's Software Infrastructure – A Failure Story", *in Proceedings of the Software Engineering, Testing, and Quality Assurance for Natural Language Processing Workshop*, 31 – 39.

Reed, S. & D. Lenat (2002), "Mapping Ontologies into Cyc", *in Proceedings of AAAI 2002 Conference Workshop on Ontologies For The Semantic Web,* (available at https://www.aaai.org/Papers/Workshops/2002/WS-02-11/WS02-11-010.pdf).

Rudnicka, E. F. Bond, Ł. Grabowski, M. Piasecki & T. Piotrowski (2018), "Lexical Perspective on Wordnet to Wordnet Mapping", *in Proceedings of the 9th Global WordNet Conference* (GWC 2018), p. 210.

Tufis, D., D. Cristea & S. Stamou (2004), "Balkanet: Aims, methods, results and perspectives. A general overview", *Romanian J. Sci. Tech. Inform.* (Special Issue on Balkanet), 7(1-2), 9 – 43.

Vossen, P. (1998), *EuroWordNet: A Multilinugal Database with Lexical Semantic Network*, The

Kluwer Academic Publisher.

Vossen, P., Cl. Soria & M. Monachini (2013), “Wordnet-LMF: a standard representation for multilingual wordnets” (ch. 4.), *in LMF Lexical Markup Framework,* (ed.) G. Francopoulo, ISTE / Wiley (ISBN 978-1-84821-430-9).

〈디지털화된 이종 언어자원의 연계와 인문학 연구의 확장성 -『통합디지털한한대사전』과 KorLex의 연동 가능성 검토를 통해〉에 대한 토론문

최호섭(단국대학교)

먼저 자연언어처리, 국어정보학 분야에 큰 업적들을 남기고 계시는 윤애선 교수님의 발표에 토론을 맡게 되어 영광으로 생각합니다.

국내외 어휘의미망 관련 연구는 대부분 사전을 비롯한 대규모 언어자원을 기반으로 한 구축 및 활용 연구로서 '어휘통제집, 시소러스, 어휘망, 의미망, 개념망, 온톨로지, 지식베이스, 지식그래프' 등의 용어로 현재까지 다양한 분야에서 연구되고 있습니다. 본 발표문은 한국학 연구의 대표적인 성과인 『한국한자어사전』, 『한한대사전』을 통합한 『통합디지털한한대사전』과 2000년대 초반부터 연구개발되어 한국어정보처리에서 대표적인 의미적 언어자원으로 인정받고 있는 한국어 어휘의미망 *KorLex*의 연계 가능성을 모색하고 각 언어자원의 활용성 확대를 검토한 의미 있는 기초 연구로 생각합니다. 즉 본 발표문의 핵심은 기구축된 특정 대규모 사전과 어휘의미망을 실질적으로 연계·활용하기 위한 개별 자원의 특징 분석, 유사 사례 분석, 표본 분석, 확장 가능성 등을 제시하고 있다는 점입니다. 이러한 발표 내용과 제 경험에 비추어 간략하게 세 가지 의견과 질문을 드리고자 합니다.

첫째 특정 사전(자료)과 어휘의미망의 연계에서 가장 중요한 부분은 오랜 기간 연구·구축된 사전의 내적 데이터 구조를 의미적 데이터(Semantic Data) 구조로의 변환의 필요성입니다. 본 발표문에서도 언급하였듯이, 두 이종 언어자원의 일차적인 연계 방법은 표제어휘의 어형 중심의 사상(mapping)이지만 어의 분석(Semantic Analysis) 과정이 필요합니다. 이는 어휘의미망을 이용한 사전의 보완, 사전 내부의 의미적 탐색 등에도 활용될 수 있습니다. 이를 위해서는 개체명까지 포함한 의미 주석 뜻풀이 말뭉치(Sense Tagged Definition Corpus), 사전 내부의 특징을 고려한 통제어휘집(시소러스), 다양한 개념 및 의미 관계를 표현한 범용/도메인 온톨로지, 문헌정보학에서 많이 다루는 전거데이터(Authority Data) 등 사전의 의미적 데이터 구조를 마련한다면 어휘의미망과의 연계가 더욱더 체계적이고 정밀해질 수 있을 것입니다. 이와 관련하여 『통합디지털한한대사전』과 *KorLex*와의 연계, 나아가 다른 어휘의미망 또는 시소러스와 연계를 위해서 사전 내부를 어떻게 기계 가독형 의미적 구조화를 어떻게 하면 좋을지에 대한 의견을 주시면 관련 연구자에게 도움을 될 수 있을 것으로 생각합니다.

둘째 사전의 활용성 및 외부적 확장 가능성 모색 관점에서 『통합디지털한한대사전』과 *KorLex*와의 연계는 *KorLex*를 중심으로 한 이기종 데이터의 상호운용(Interoperability on Heterogeneous Data)의 가능성입니다. *KorLex*는 『표준국어대사전』을 비롯하여 『세종전자사전』, PWN 등과의 연계 즉 상호운용성을 지원하는 어휘의미망이자 온톨로지, MDR(Metadata Registry)이라 할 수 있습니다. 이러한 특징은 국사편찬위원회의 '한국역사용

어시소러스’를 통한 한국사데이터베이스, 한국역사정보통합시스템 등의 주요 데이터와의 검색에 활용된다는 점과도 유사하다고 할 수 있습니다. 그러나 이는 특정 사전을 중심으로 다른 사전과의 표제어 사상(mapping)이나 다른 자료와의 검색 연계가 아닌, 어휘의미망을 이용한 포괄적인 정보 연계 및 상호운용성을 지원할 수 있는 측면을 고려해야 한다고 생각합니다. 나아가 편찬 동기, 역사성 반영, 어의의 다양화 등 사전의 개별적 특성을 고려한 어휘의미망의 구축 및 연계 방법도 다를 수 있을 것으로 생각합니다. 이에 이기종 사전 데이터(나아가 관련 데이터) 연계 및 상호운용성을 고려함과 동시에 개별 사전을 특징으로 고려한 어휘의미망은 어떻게 구성되어야 하는지에 대한 의견을 부탁드립니다. 또한, 본 발표문에 디지털화한 사전으로 언급된 『한불ᄌᆞ뎐』, 『한영자뎐』, 『조선어사전』 등도 *KorLex*와의 연계가 어떻게 이루어진 것인지도 궁금합니다.

마지막으로 한 번 정도 고민해볼 만한 특정 사전마다 구축(직접 구축, 참조 구축 등)되는 어휘의미망의 필요성입니다. 관련된 연구 과제나 사업 등에서 발견되는 이 부분은 사전 내부의 의미적 연결 구조를 바탕으로 단순한 검색 이상의 사전의 활용성 증대, 정보 전달의 다양성 등 다양한 목적으로 사전 기반 어휘의미망 구축을 진행하고 있습니다. 즉 어휘의미망은 일반적인 단어의 의미 관계(Semantic Relation)뿐만 확장적 개념 관계(온톨로지의 다양한 관계 표현)까지 포함하여 어휘통제집(시소러스), 온톨로지 등과 같은 영역까지 구축 가능할 것이라는 생각이 많은 편입니다. 개별 사전을 중심으로 한 어휘의미망 직접 구축, 참조 구축에 대한 교수님의 의견을 부탁드립니다.

본 발표문에서 다루고 있는 『통합디지털한한대사전』과 어휘의미망(*KorLex*)의 연계 연구는 많은 시간과 노력이 요구되지만, 기초 연구에서부터 앞으로의 연구는 분명히 의미 있는 결과물, 다양한 확장・활용 방향으로 이어질 것으로 기대하고 있습니다.

동아시아 4개 언어 (한-중-일-베트남어) 한자어 데이터베이스의 구축과 활용, 그리고 확장 가능성

신웅철(경성대)

1. 들어가기: 데이터베이스 구축의 배경

언어의 배경에는 문화가 존재한다.[1] 한국, 중국, 일본, 베트남의 언어와 문화를 관통하는 요소를 꼽자면 한자어를 빼놓을 수 없다. 한자와 한자어는 방대한 문명적 자산을 실어나르는 매체로서, 동아시아 여러 언어권은 지리적, 사회적, 언어적 환경의 이질성을 뛰어넘어 공동의 문명권을 형성하는 기능을 수행했다. 한자는 표어문자로서의 특성을 바탕으로 동아시아 4개 언어권(한국어, 중국어, 일본어, 베트남어)을 하나의 문명권으로 잇는 보편적 의미를 보존하였으며, 다른 한편으로는 각 언어권의 문화적 특수성이 투영된 개별적 의미로 변형되기도 하였다. 이러한 문명적 자산의 확산과 공유가 반드시 일방향적이기만 했던 것은 아니며, 특히 서양 문명의 전면적 수용이 시작된 19세기 이후로는 보다 역동적이고 다방향적인 양상을 띠었다. 오늘날 동아시아 한자문명을 관통하는 문화적 보편성과 각 언어권에 보존된 개별성은 그러한 과정의 산물이다.

한자어는 그것을 실어날라 문명의 진보와 변화를 촉발해 온 도구임과 동시에, 문명의 보편성과 개별성이 투영된 실체적 요소이기도 하다. 따라서 우리는 동아시아 문명의 실체를 규명하기 위해 한자어에 주목할 필요가 있다. 경성대학교 한자문명연구사업단(이하 본 사업단으로 약칭)의 아젠다는 오늘날 동아시아 각 언어권 안에서 한자어가 갖는 의미적 특성에 주목하여, 그것을 바탕으로 문명의 전파와 변용의 궤적을 추적하여 동아시아의 문명 형성의 실체를 규명하고자 하는 것이다. 그것은 단순히 서로 다른 언어 간의 어휘적 유사성 혹은 차별성에 머무르는 것이 아니며, 그것을 바탕으로 언어 외적인 배경인 문화와 문명의 공시적, 통시적 현상을 내다보고자 하는 것이다.

본 사업단의 연구는 동아시아 4개 언어의 한자어가 갖는 의미를 단초로 삼아, 그것에 투사된 다양한 문화적 특성 고찰로 나아가는 것을 지향한다. 예컨대 동아시아 외부에서 도래한 불교적 개념을 나타내는 Bodhisattva는 음역되어 菩薩이라는 한자어로 동아시아 문명 내의 각 언어권으로 확산되었다. 각 언어 내에서 오늘날 이 한자어가 갖는 의미항을 대표하는 키워드를 나열하면 다음과 같다.

菩薩(K) : 부처; 수행자; 교화방법; 여자신도; 고승;[2]

---

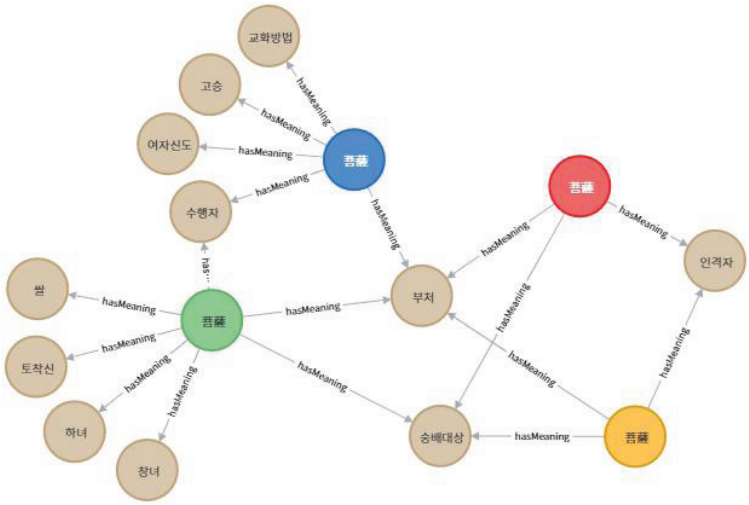
1) "Again, language does not exist apart from culture, that is, from the socially inherited assemblage of practices and beliefs that determines the texture of our lives." Sapir, Edward (1921), *Language: An introduction to the study of speech*. New York: Harcourt, Brace and Company. 221-222쪽. https://www.gutenberg.org/ebooks/12629

2) "표준국어대사전" **보살**[1][**菩薩**]: [1] <불교> 부처가 전생에서 수행하던 시절, 수기를 받은 이후의 몸. [2] <불교> 위로 보리를 구하고 아래로 중생을 제도하는, 대승 불교의 이상적 수행자상. ≒보리살타, 살타, 상사, 진신. [3] <불교> 삼승(三乘)의 하나. 보살이 큰 서원(誓願)을 세워 위로 보리를 구하고 아래로 중생을 교화하는 교법을 이른다. =보살승. [4] <불교> 여자 신도(信徒)를 높여 이르는 말. [5] <불교> '고승'(高僧)을 높여 이르는 말. [6] <불교> 머리를 깎지 않고 절에서 사는 여자 신도. =보살할미.

菩薩(C) : 부처; 숭배대상; 인격자;[3]

菩薩(J) : 부처; 숭배대상; 토착신; 고승; 수행자; 쌀; 하녀; 창녀;[4]

菩薩(V) : 부처; 숭배대상; 인격자;[5]



어휘 생성 당초의 본래의 의미(부처, 숭배대상, 수행자)가 공통되게 분포하는 가운데, 개별 언어별로 특수한 의미 파생 또한 관찰된다. 특히 한국어(K)와 일본어(J)에서는 본래의 의미를 떠올렸을 때 선뜻 이해하기 어려운 파생 의미(여자신도, 쌀, 하녀, 창녀 등)가 존재한다는 점에서 주목된다. 이것을 단초로 언어 외적인 현상으로 확장하여 해당 언어권의 문화 안에서 불교라는 종교가 어떻게 수용되어 어떠한 위상에서 문화적으로 기능하였는지, 그리고 그것과 菩薩이라는 어휘의 의미 파생과 어떤 상관성이 있는지 등에 대해 고찰해 볼 수 있을 것이다.

본 사업단에서는 한자어를 바탕으로 한 동아시아 문명 연구의 수행을 다음과 같은 세 단위로 나누어 시행하고자 한다.

1) [한자어 DB-아카이브 구축]

常用 한자어, 번역/유행/신조 한자어, 기록물 및 이미지 아카이브

2) [한자어 비교연구]

: 분야별 한자어의 의미 비교 및 해설 자료 작성 - 기록물&이미지와의 연계.

기초/문화/상업, 동식물/상업/과학/법률/제도, 종교/사상/번역, 기록물/특수어휘, 한자 사용의 역사/정책 특성

---

3) "汉语大词典" **菩$_2$薩**: [1]佛教名词。梵文菩提萨埵(Bodhisattva)之省，原为释迦牟尼修行而未成佛时的称号，后泛用为对大乘思想的实行者的称呼。 [2]指人们崇拜的神灵偶像。 [3]比喻心肠仁慈的人。

4) "日本国語大辞典"(第二版) **ぼ-さつ[菩薩]**: [1]仏語。もと、釈迦牟尼の前生における呼称。大乗仏教が興って、修行を経た未来に仏になる者の意で用いる。悟りを求め修行するとともに、他の者も悟りに到達させようと努める者。また、仏の後継者としての、観世音、彌勒、地蔵など。[2]昔、朝廷から碩徳の高僧に賜わった号。[3]本地垂迹説の勃興以後、神につけられた号。[4]菩薩に扮する雅楽の舞人。[5]米の異称。[6]転じて、飯炊きの下女。[7]遊女の異称。

5) "Từ điển tiếng Việt" **Bồ tát**: (thường viết hoa) người tu hành đắc đạo trong đạo Phật, có hiểu biết rộng, có đức độ cao.

3) [문명연구]

문화/철학 개념어, 불교/유교/기독교/도교 개념어, 근현대 번역어, 유행어/신조어, 한자 문화 특성 및 독자성, 한자교육과 학습, 인공지능/인지과학, 문명의 미래

첫째, 동아시아 4개 언어의 한자어에 대한 정보를 집적하여 데이터베이스화하는 한편 어휘와 관련한 역사적 기록물 및 이미지를 아카이빙한다. 둘째, 수집된 한자어를 어휘 분야별로 나누어 개별 언어권 간의 공통성과 특수성을 비교 검토한다. 또한 아카이빙된 기록물 및 이미지에는 어휘와의 연계 정보를 부여하여 실체적 문명 연구의 기반을 마련한다. 셋째, 한자어가 생성되어 이동, 변이하는 과정에 대한 규명을 통해 동아시아 문명 전반을 아우르는 보편성과 개별 언어권의 개별성의 변화 양상을 기술한다. 본 발표에서 소개하는 '동아시아 4개 언어 한자어 데이터베이스'(이하 동아시아 한자어DB로 약칭)는 위 세 단위 가운데 첫번째 단위에 해당하는 작업의 일환으로 구축된 것이다.

2. 데이터베이스 구축의 앞선 사례와 문제점

가. 1. '일-한-중-베트남 동형 2자 한자어 데이터베이스'의 사례

본 데이터베이스에 앞서 동아시아 4개 언어권의 한자어를 종합적으로 비교 분석하고자 시도한 사례로는 일본 나고야대학 다마오카(玉岡) 연구실을 중심으로 한 "일-한-중-베트남 동형 2자 한자어 데이터베이스"[6](이하 동형한자어DB로 약칭)를 들 수 있다. 이 동형한자어DB의 구축 데이터 항목을 정리하면 다음과 같다.[7]

[표] 일-한-중-베트남 동형 2자 한자어 데이터베이스의 데이터 항목

항목 번호	항목 내용
1	표제어 번호
2,3	일본어 표제어의 표기 정보: 한자, 히라가나
4~8	일본어 표제어의 품사 정보: 일본어 사전 5종
9,10	일본어 표제어의 사용빈도: 아사히 신문(1985-1998), 마이니치 신문(2000-2010)
11	일본어 표제어의 난이도: 일본어능력시험 출제기준 상의 등급
12,13	한국어 동형한자어의 표기 정보: 한자, 한글
14~17	한국어 동형한자어의 품사 정보: 표준국어대사전
18,19	중국어 동형한자어의 표기 정보: 간체자, 한어병음
20~22	중국어 동형한자어의 품사 정보: 중국어 사전 2종
추가	일본어-중국어 동형한자어의 의미적 관계 (Same, Overlap, Different, Nothing)
추가	베트남어 한자, 국어자 표기

---

6) 日韓中越同形二字漢字語データベース, https://kanjigodb.herokuapp.com/
2014년 기준으로 일본어, 한국어, 중국어 동형한자어를 대상으로 한 품사 정보 중심의 데이터베이스가 먼저 구축되었으며(朴善姸, 熊可欣, 玉岡賀津雄 (2014)「同形二字漢字語の品詞性に関する日韓中データベースの概要」『ことばの科学』27, 3~23쪽), 이후 베트남어 동형한자어 정보와 음운유사성 지표가 추가된 것으로 보인다(于劭贇, 玉岡賀津雄, ホアーン ティ ラン フォン (2019)「日韓中越4言語における2字漢字語の音韻類似性に関するデータベースおよび検索エンジンの構築」『ことばの科学』33, 75~94쪽).

7) 朴善姸, 熊可欣, 玉岡賀津雄 (2014)「同形二字漢字語の品詞性に関する日韓中データベースの概要」『ことばの科学』27, 5~6쪽.

동형한자어DB는 위 데이터 항목에서 보는 바와 같이 일본어의 2자 한자어를 기준으로 한국어, 중국어, 베트남어의 동형 한자어를 비교 대상으로 삼고 있다. 그러나 동형한자어DB의 주안점은 첫째로 품사로 대표되는 각 언어별 동형어가 갖는 문법적 기능의 같고 다름 기술, 둘째로 음운적 유사성 지표 설정을 통한, 외국어 학습과 교육 상의 편익이다. 따라서 동아시아 4개 언어의 한자어를 대상으로 한다는 점에서는 시사하는 바가 크다. 그러나 한자어의 의미적 특성을 통해 그것의 생성과 전파 및 변용 과정을 살피고, 나아가 그것에 투영된 문명적 현상을 추적하고자 하는 본 사업단의 연구 지향과는 부합하지 않는다고 볼 수 있다. 아울러 동형한자어DB는 일본어라는 단일 언어의 상용도를 기준으로 선정된 것에 대해 같은 한자로 표기되는 나머지 3개 언어의 어휘를 모았다는 점에서, 일본어 이외의 언어 안에서 각 한자어가 갖는 어휘적 위상에 대한 정보를 가늠하기 어렵다는 한계도 존재한다.

나. 2. 동형 표기어를 기준으로 한 어휘선정의 문제점

동아시아 4개 언어는 어휘적으로 한자와 한자어라는 요소를 갖는다는 공통점이 있지만, 개별 언어 안에서의 운용에는 크고 작은 차이가 있다. 따라서 이들을 적절히 비교 고찰하기 위해서는 각 언어 안에서 개별 한자어가 갖는 '사용역(register)'과 '난이도'에 대한 고민이 필요하다.

앞서 살펴본 동형한자어DB를 비롯하여 기존의 동아시아 개별 언어 간의 한자어 비교 연구는 같은 한자[8]로 표기되는 어휘(이하 동형어)를 시작점으로 삼아 의미적인 차이를 분석하는 방법이 주를 이루었다. 다만 그러한 방법에서는 동형어의 유무와 의미 차이에 주안을 두어, 사용역과 난이도 차이를 파악하기 어렵다는 문제점이 있다. 또한 전혀 다른 한자로 표기되지만 의미적으로 동가에 가까운 어휘인 대역어(equivalent)가 분석의 시야에 들기 어렵다는 문제도 존재한다.[9]

소규모 어휘를 대상으로 하는 연구에서는 이러한 맹점을 개별적으로 보완하여 진행할 수 있을 것이다. 그러나 대규모 어휘를 대상으로 하는 데이터베이스 구축을 통한 연구에서는 그러한 개별적 보완을 전면적으로 실시하기 어렵다. 따라서 언어별로 사용역의 격차가 작은 어휘를 중심으로 개별 언어 간의 대역어로 시야를 확장하기 쉬운 구조로 설계할 필요가 있다. 동아시아 한자어DB는 이러한 문제의식 위에 한국어를 기준 언어로 설정하여 기초어휘 목록에서 조건에 부합하는 표제어를 선정하고, 나머지 중국어, 일본어, 베트남어와의 대역사전에 수록된 어휘를 검토하여 대역어를 선정하는 방식을 취하였다.

---

8) 중국어의 간체자, 번체자, 일본어의 신자체, 구자체 등과 같은 구체적 실현 형태로서의 자체(字體)나 자형(字形)의 차이는 무시된다. 이러한 구체적 실현 형태의 차이를 추상화하여 통합하는 단위로서 자종(字種)이라는 단위를 설정해 볼 수 있을 것이다.

9) 별개의 자종으로 간주되지만 의미적으로 통용 가능한 글자에 대해서는 각 언어의 어휘의 표기에서 다르게 채용되는 경우가 있을 수 있다. 가령 紀念과 記念은 같은 의미의 단어이지만 중국어에서는 전자를 일본어에서는 후자를 일상적으로 사용한다. 한편, 특정 언어 안에서 동음 관계에 있는 글자로 대체되는 단어의 경우도 있다. 일본어에서는 洗滌(せんじょう)과 銓衡(せんこう)이 동음의 한자어 洗淨(せんじょう)과 選考(せんこう)로 대체되어 쓰이고 있으나 나머지 언어에서는 맥락에 따라 통용이 어려운 단어일 수 있다.

3. 데이터베이스의 설계와 구축

가. 1. 데이터베이스의 설계

1) 데이터베이스의 구조

동아시아 한자어DB는 우선 각 언어별 평면 테이블을 기반으로 하는 관계형 데이터베이스(Relational Database)로 구축한다. 한국어, 중국어, 일본어, 베트남어 각 언어별로 기본적으로 유사한 구조의 각기 독립된 평면 테이블을 작성하였으며, 각 언어별 테이블의 속성(attribute)은 다음과 같다.

<동아시아 4개 언어 한자어 DB 언어별 테이블의 속성>

한국어(ko) : [ID]-[ko-Hang]-**[ko-Hant]**-[ko-Latn]-[ko_def]-[ko_smntc]

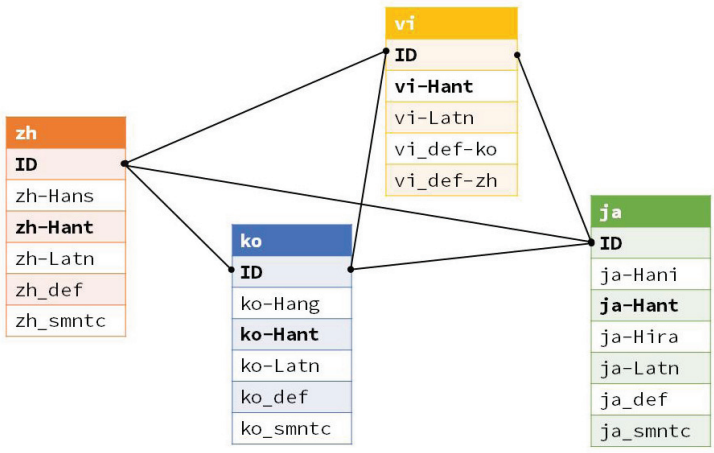
중국어(zh) : [ID]-[zh-Hans]-**[zh-Hant]**-[zh-Latn]-[zh_def]-[zh_smntc]

일본어(ja) : [ID]-[ja-Hani]-**[ja-Hant]**-[ja-Hira]-[ja-Latn]-[ja_def]-[ja_smntc]

베트남어(vi) : [ID]-[vi-Hant]-[vi-Latn]-[vi_def-ko]-[vi_def-zh]-[vi_smntc]

기준언어인 한국어 어휘에 일련번호[ID]를 부여하고 나머지 3개 언어의 대역어에 대해서 한국어 어휘와 동일한 일련번호[ID]를 부여하였다. 이것을 통해 4개 언어에 부여된 공통의 일련번호([ID])를 기본 키로 삼아 의미적으로 동가에 가까운 어휘들이 연결되는 관계형 데이터베이스로 설계하였다. 아울러 한자 표기를 기준으로도 연결될 수 있도록 각 언어의 한자 표기를 한국 정자체로 정규화한 속성([ko-Hant][zh-Hant][ja-Hant][vi-Hant])을 두었다. 각 언어 테이블의 속성에 대해서는 후술하는 데이터 구축의 과정을 통해 설명하겠다.

[그림] 동아시아 4개 언어 한자어 DB 언어별 테이블의 속성



나. 2. 기준언어 표제어와 비교언어 대역어 선정

국립국어원에서 제공 중인 "한국어기초사전"[10]에 수록된 기초어휘 가운데 한자어만을 추

10) 국립국어원, 한국어기초사전, https://krdict.korean.go.kr/

려 총 23,000여 단어를 기본 표제어로 추출하여 기준으로 삼았다.[11] 그 가운데 언어 간의 비교 대상으로서 유의미한 결과가 기대되지 않는 어휘, 즉 고유명사나 한국 특유의 문화적 사항을 나타내는 어휘 등을 배제하는 후처리 과정을 통해 최종적으로 21,763개 단어로 압축되었다. 확정된 표제어에 대해서는 다음과 같은 항목으로 표기 정보를 입력하였다.

ko: [ID]-[ko-Hang]-[ko-Hant]-[ko-Latn]

[ID]: 한국어 표제어의 고유 식별자 (이하 C, J, V도 같음)

[ko-Hang]: 한국어 표제어의 한글 표기

[ko-Hant]: 한국어 표제어의 한자 표기

[ko-Latn]: 한국어 표제어의 문광부식 로마자 표기[12]

이어서 기준언어로 지정된 한국어 표제어 21,763개를 바탕으로 이에 대응하는 대역어를 선정하였다. 일본어, 베트남어는 "한국어기초사전"의 언어별 대역 학습사전(한국어 표제어의 대역어, 한국어 뜻풀이의 해당 언어 번역 등 제공)[13]의 대역어를 우선적으로 활용하여 연구진과 원어 화자의 다중 검토 방식으로 대역어를 제외 또는 수정, 추가하였다.[14][15] 한편 대역어 선정 작업을 진행하던 2019년 기준으로 "한국어기초사전"과 연계한 대역 학습사전이 공개되지 않았던[16] 중국어에 대해서는 "에듀월드 한한중사전"[17]과 "에듀월드 중중한사전"[18]의 원 데이터를 가공하여 연구진과 원어 화자의 다중 검토 방식으로 대역어를 확정하였다.

이와 같은 방식으로 중국어 23,091개, 일본어 19,517개, 베트남어 7,761개의 대역어를 선정하였다. 아울러 중국어 1,611개, 일본어 988개의 동형이의어를 별도로 추가하였다. 위와 같은 내역으로 4개국 한자어 74,731개에 이르는 방대한 데이터베이스를 구축하였다. 선정된 대역어에 대한 표기(한자, 표음문자) 정보를 다음과 같은 속성으로 나눠 입력하였다.

zh: [ID]-[zh-Hans]-[zh-Hant]-[zh-Latn]

---

11) 어휘 추출 작업은 "한국어기초사전"의 사전 내려받기 기능을 활용하여 다운로드한 데이터를 기반으로 이루어졌다. 대상 어휘는 '한자어'뿐만 아니라 '혼종어' 중에서 동사 등의 어근으로 판단되는 것도 포함하였다.

12) 국립국어원, '국어의 로마자 표기법'(문화체육관광부 고시 제2014-42호, 2014. 12. 5.), https://kornorms.korean.go.kr/regltn/regltnView.do?regltn_code=0004&regltn_no=444#a444

13) 국립국어원 한국어-일본어 학습사전(国立国語院 韓国語-日本語学習辞典), https://krdict.korean.go.kr/jpn/ 국립국어원 한국어-베트남어 학습사전(Từ điển học tiếng Hàn-tiếng Việt của, Viện Quốc ngữ Quốc gia), https://krdict.korean.go.kr/vie/mainAction

14) 대역어 선정 작업도 한국어 표제어 추출 시와 마찬가지로 "한국어기초사전"의 사전 내려받기 기능을 활용하여 다운로드한 데이터를 기반으로 이루어졌다.

15) "국립국어원 한국어-일본어 학습사전"에서는 대역어 선정 원칙에서 동형어를 우선하는 경향이 관찰된다. 그러나 그 가운데는 사용역이나 난이도의 측면에서 비대칭적인 어휘가 포함되어 있어 원어 화자의 내성과 말뭉치 활용을 통한 수정이 필요하였다. 신웅철(2020) '한국어와 일본어의 同綴한자어 비교 연구: 국립국어원『한국어-일본어 학습사전』의 표제어와 대역어를 중심으로' "일어일문학연구" 115, 153-172쪽.

16) 국립국어원 한국어-중국어 학습사전(国立国语院韩国语-汉语学习词典, https://krdict.korean.go.kr/chn/)의 시범 운영은 2020년 5월 15일에 개시되었다. 국립국어원, 《국립국어원 한국어-중국어 학습사전》개통, https://korean.go.kr/front/board/boardStandardView.do?board_id=6&mn_id=184&b_seq=754

17) 차이나랩, 에듀월드 표준한한중사전, https://zh.dict.naver.com/

18) 차이나랩, 에듀월드 표준중중한사전, https://zh.dict.naver.com/

[zh-Hans]: 중국어 대역어의 간체자 표기
[zh-Hant]: 중국어 대역어의 번체자 표기
[zh-Latn]: 중국어 대역어의 한어병음 표기

ja: [ID]-[ja-Hani]-[ja-Hant]-[ja-Hira]-[ja-Latn]
[ja-Hani]: 일본어 대역어의 신자체 표기
[ja-Hant]: 일본어 대역어의 구자체 표기
[ja-Hira]: 일본어 대역어의 히라가나 표기
[ja-Latn]: 일본어 대역어의 헵번식 로마자 표기[19]

vi: [ID]-[vi-Hani]-[vi-Latn]
[vi-Hant]: 베트남어 대역어의 한자 표기
[vi-Latn]: 베트남어 대역어의 국어자(Chữ Quốc ngữ) 표기

다. 5. 표제어 및 대역어의 의미정보 및 의미 속성 추가

개별 언어 안에서 위의 한자어 간에 의미가 같고 다름을 비교하기 위해서는 신뢰할 수 있는 사전으로부터 어휘의 뜻풀이(definition)를 인용하여 데이터베이스에 편입하는 작업이 필수적이다. 앞서 밝힌 바와 같이 기준언어(한국어) 표제어 선정과 각 언어 간의 대역어 설정에는 "한국어기초사전"과 그것에서 파생된 개별 언어와의 대역 학습사전을 중심으로 활용하였다. 그러나 해당 사전의 한국어 뜻풀이는 "표준국어대사전"[20]의 의미정보를 바탕으로 '학습기초어휘사전'의 쓰임에 걸맞게 가공된 것이라 대체로 소략하다. 또한 일본어, 베트남어로 번역된 의미정보는 한국어 어휘에 대한 의미정보를 해당 언어로 번역한 것이라는 점에서 본 데이터베이스가 지향하는 방향에 부합하지 않는다.

이에 한국어 표제어에 대해서는 국립국어원의 "표준국어대사전", 중국어 대역어 및 동형이의어에 대해서는 "에듀월드 중중한사전", 일본어 대역어 및 동형이의어에 대해서는 "고지엔 일한사전"으로부터 각 표제어 및 대역어가 해당 언어에서 갖는 의미정보를 확보하였다. 베트남어의 경우는 베트남 사회과학한림원 산하 언어연구원[21]의 "베트남어 사전(Từ điển tiếng Việt)"[22]의 베트남어 뜻풀이를 한국어와 중국어로 번역하여 추가하였다.

이들 사전으로부터 추가한 의미정보를 바탕으로 각 언어별로 의미 속성을 나타내는 태그(smntc)를 추가하였다. 의미분류의 체계는 부산대학교 인공지능연구실과 한국어정보처리연구실에서 WordNet[23] 기반으로 구축한 한국어 어휘의미망(KorLex)[24]을 참조하여 입력하였다.

ko: [ID]-[ko-Hang]-[ko-Hani]-[ko-Latn]**-[ko_def]-[ko_smntc]**

19) 日本国外務省, 'ヘボン式ローマ字綴方表', https://www.ezairyu.mofa.go.jp/passport/hebon.html
20) 국립국어원, 표준국어대사전 https://stdict.korean.go.kr/
21) Vietnam Institute of Linguistics (Viện Ngôn ngữ học), http://www.vienngonnguhoc.gov.vn/
22) Hoàng Phê et al. (2015), *Từ điển tiếng Việt*, Đà Nẵng: Nhà xuất bản Đà Nẵng. http://www.vietlex.com/san-pham/42-TD_Tieng_Viet_2015
23) Princeton University WordNet, https://wordnet.princeton.edu/
24) 부산대학교 인공지능연구실, 한국어정보처리연구실, 한국어 어휘의미망 (KorLex), http://korlex.pusan.ac.kr/

[ko_def]: 한국어 표제어의 "표준국어대사전" 의미정보
[ko_smntc]: 한국어 표제어의 KorLex 상위 의미범주 ([K_definition] 참고)

zh: [ID]-[zh-Hans]-[zh-Hant]-[zh-Latn]**-[zh_def]-[zh_smntc]**
[zh_def]: 중국어 대역어의 "에듀월드 중중한사전" 의미정보
[zh_smntc]: 중국어 대역어의 KorLex 상위 의미범주 ([C_definition] 참고)

ja: [ID]-[ja-Hani]-[ja-Hant]-[ja-Hira]-[ja-Latn]**-[ja_def]-[ja_smntc]**
[ja_def]: 일본어 대역어의 "고지엔 일한사전" 의미정보
[ja_smntc]: 일본어 대역어의 KorLex 상위 의미범주 ([C_definition] 참고)

vi: [ID]-[vi-Hani]-[vi-Latn]**-[vi_def-ko]-[vi_def-zh]-[vi_smntc]**
[vi_def-ko]: 베트남어 대역어의 "Từ điển tiếng Việt" 의미정보 번역(한국어)
[vi_def-zh]: 베트남어 대역어의 "Từ điển tiếng Việt" 의미정보 번역(중국어)
[vi_smntc]: 베트남어 대역어의 KorLex 상위 의미범주 ([vi_def-ko], [vi_def-zh] 참고)

4. 동아시아 4개 언어 한자어 데이터베이스의 확장

가. 1. 어휘의 사용빈도와 사용역 속성 추가

각 언어에서 한자어(표제어, 대역어, 동형어)를 보다 입체적으로 조감하기 위해서는 어휘의 사용빈도와 사용역에 대한 속성이 필요하다. 사용빈도에 대해서는 각 언어의 말뭉치를 활용한 방식으로 수치화하는 방식이 유력하다. 다만 각 언어의 개별 말뭉치마다 규모나 특성이 균등하지 않을 것이기 때문에 사용빈도를 수치화하는 것만으로는 언어횡단적인 활용에는 제약이 따를 것으로 예상된다. 유의어 관계에 있는 어휘 간의 상대적 사용빈도 비교도 하나의 대안으로 생각해 볼 수 있다.

사용역에 대해서는 어떠한 분류체계를 설정할 것인지가 문제이다. 도원영(2008)에 제시된 한국어사전의 표제어에 대한 사용역 분류표[25]를 바탕으로 가령 '북한어' 등 나머지 3개 언어에 적용하기 어려운 부분을 제외한 분류체계를 고안할 필요가 있다. 한편 "표준국어대사전"의 어휘 속성으로 수록되어 있는 67개 '전문분야'[26] 정보 또한 어휘의 사용역을 판단하는 정보로서 유효하다.

나. 2. 전문용어의 추가

현재 동아시아 4개 언어 한자어 데이터베이스는 한국어 기초어휘 2만 여 개를 기준으로

---

25) 도원영(2008) '국어사전 표제어의 사용역 정보에 대한 고찰', "우리어문연구" 30, 43쪽.

26) "표준국어대사전"에 어종 '한자어'로 분류된 약 235,000개 어휘 가운데 '없음'을 제외한 나머지 67개 속성이 부여된 것은 약 129,000개로 55%에 이른다.
이들 67개 '전문분야'의 분류 내역은 다음과 같다. "가톨릭, 건설, 경영, 경제, 고유명 일반, 공업, 공예, 공학 일반, 광업, 교육, 교통, 군사, 기계, 기독교, 농업, 동물, 매체, 무용, 문학, 물리, 미술, 민속, 법률, 보건 일반, 복식, 복지, 불교, 사회 일반, 산업 일반, 생명, 서비스업, 수산업, 수의, 수학, 식물, 식품, 심리, 약학, 언어, 역사, 연기, 영상, 예체능 일반, 음악, 의학, 인명, 인문 일반, 임업, 자연 일반, 재료, 전기·전자, 정보·통신, 정치, 종교 일반, 지구, 지리, 지명, 책명, 천문, 천연자원, 철학, 체육, 한의, 해양, 행정, 화학, 환경."

구축한 것이기 때문에 분야별 전문용어가 충분하지 않다는 한계를 갖는다. 이러한 전문용어 확장을 위한 기초자료로서 한국검인정교과서협회의 “2015 개정 교육과정에 따른 교과용도서 개발을 위한 편수자료”의 Ⅱ(인문.사회과학/체육.음악.미술)와 Ⅲ(기초과학/정보)에 대한 편수용어 자료(xlsx, 이하 교과서 편수용어로 약칭)[27]의 활용을 생각해 볼 수 있다. 교과서 편수용어는 11개 교과별로[28] 사용 가능한 어휘를 수록하고, 각 어휘에 대한 정보를 [용어]-[한자]-[외국어]-[비고]의 항목으로 정리한 어휘 목록이다.[29] 다만, 개별 어휘의 사용 빈도나 난이도와 같은 정보가 없다는 점은 주의를 요한다.

위 교과서 편수용어와 앞서 언급한 “표준국어대사전”의 67개 ‘전문분야’ 정보를 교차하는 방법이 가능 방법도 생각해 볼 수 있다. 그것을 통해 각 어휘가 갖는 도메인 특성을 보다 세밀하게 유형화할 수 있을 것이다.

한국어 이외의 언어에서 이와 유사한 성격의 어휘 자료로는 일본 국립국어연구소의 “언어 정책에 도움이 되는, 말뭉치를 활용한 어휘표・한자표 등의 작성과 활용 부속 CD-ROM”[30]에 수록된 ‘교과서 말뭉치 어휘표’[31]와 ‘교과 특징어 목록’[32]을 들 수 있다. 이들 어휘표 및 어휘 목록은 교과서 말뭉치 구축을 통해 추출해낸 것으로서 어휘의 난이도(초등학교, 중학교, 고등학교)와 사용 빈도까지 망라하여 유용하다. 한편 중국어와 베트남어에 대해서도 이와 유사한 성격의 어휘 목록의 유무를 점검할 필요가 있다.

다. 3. 개별 어휘 간의 관계성 목록 구축

현재 동아시아 한자어DB에 구축된 데이터는 개별 어휘의 속성에 국한된다. 물론 이러한 속성을 통해 언어횡단적인 어휘의 상관관계를 추적할 수 있도록 설계하였으나, 관계형 데이터베이스라는 구조의 특성상 개별 어휘 간의 관계성을 정밀하게 확인하거나 기술하기에는 부족한 부분이 있다. 또한 그러한 관계성 확인에서 데이터베이스 이용자의 몫으로 상정된 부분이 크다. 따라서 데이터베이스 이용자의 몫으로 상정된 몫을 명시적으로 기술하여 기계적으로 식별 및 분석할 수 있도록 데이터화하는 작업이 뒤따를 필요가 있다.

현재의 동아시아 한자어DB를 개별 어휘라는 개체(individual) 또는 접점(node)에 대한 목록으로 본다면, 이들 간을 연결(link)하는 관계성(relation)을 기술한 목록을 별도로 구축하여 데이터를 시각화하고 활용도를 높일 수 있을 것이다.[33]

---

27) ‘편수자료 Ⅱ, Ⅲ 수정 내용 반영본’ 안내, 한국검인정교과서협회, 2017-06-07, https://www.ktbook.com/edata/PDS_AuthView.asp?num=327&pageno=1&startpage=1&keyword=%C6%ED%BC%F6%C0%DA%B7%E1

28) <물리>, <미술>, <생명과학>, <수학>, <음악>, <일반사회>, <정보>, <지구과학>, <지리>, <체육>, <화학>

29) 다만 <수학>은 [용어]-[동의어]-[외국어]로 한자 정보를 수록하지 않았다.

30) 国立国語研究所(2011), 『言語政策に役立つ，ユーパスを用いた語彙表・漢字表等の作成と活用』(国立国語研究所内部報告書; LR-CCG-10-07) 付属CD-ROM, http://doi.org/10.15084/00003284
이 자료에 수록된 어휘표에 대한 간략한 설명은 홈페이지에서 확인 가능하다.

31) 教科書ユーパス語彙表(Ver. 1.0) “2005년도에 사용된 초등학교, 중학교, 고등학교 모든 학년 모든 교과의 교과서 1종씩을 대상으로 한 ‘교과서 말뭉치’의 어휘 일람표입니다. 모든 교종(초, 중, 고) 모든 학년의 모든 교과를 종합한 빈도를 알 수 있으며, 교종 별, 학년 별, 교과 별 빈도를 알 수 있습니다. 또한 “현대일본어 글말 균형 말뭉치(現代日本語書き言葉均衡ユーパス)”의 도서관 서적(고정 길이 샘플)의 빈도와도 대조할 수 있습니다.” 국립국어연구소 말뭉치 개발센터(Center for Corpus Development, NINJAL), https://ccd.ninjal.ac.jp/bccwj/freq-list.html

32) 教科特徴語リスト(Ver. 1.0): “위 ‘교과서 말뭉치(教科書ユーパス)’와 ‘도서관 서적’(고정 길이 샘플)의 어휘 빈도를 비교하여 교과 별로 특징어를 추출한 목록입니다. 중학교와 고등학교 각각에 대한 교과별 일람표로 구성됩니다.” 국립국어연구소 말뭉치 개발센터(Center for Corpus Development, NINJAL), https://ccd.ninjal.ac.jp/bccwj/freq-list.html

33) 김현, 임영상, 김바로(2016), ‘데이터의 시각화’, “디지털 인문학 입문”, 한국외국어대학교 지식출판

〈동아시아 4개 언어 (한-중-일-베트남어) 한자어 데이터베이스의 구축과 활용, 그리고 확장 가능성〉에 대한 토론문

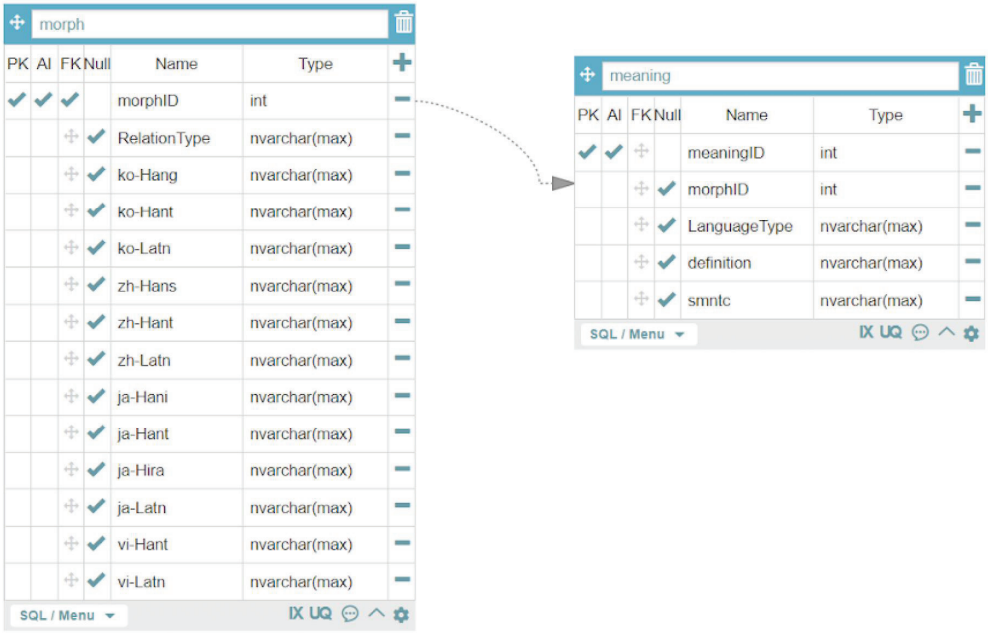
김바로(한국학중앙연구원)

본 발표는 의미를 중심으로 동아시아 4개 언어 (한-중-일-베트남어) 한자어 데이터베이스를 구축하는 배경과 방법에 대해서 상세하게 서술했기에, 경성대학교 한자문명연구사업단의 아젠다를 실현하는 아카이브로서의 비전이 보여 흥미롭게 읽었습니다. 토론자로서 "동아시아 4개 언어 한자어 데이터베이스"에 대한 다양한 생각이 떠올라서 이에 대해서 의견을 말씀드리고자 합니다.

1. 데이터 스키마 관련

1.1. 제안 RDB 데이터 스키마 및 샘플

기본적으로 현재의 아카이빙 내용이 다층적인 1:다의 관계로 상호 연결되기에 RDB 구조에서는 데이터 스키마가 지나치게 복잡해지는 문제가 분명 발생할 수 있다고 생각합니다. 그럼에도 불구하고, RDB 데이터 스키마는 최소한의 규칙을 지키면서도 실무적으로 활용하기 편한 다음의 RDB 데이터 스키마가 더 좋지 않을까 생각해 봅니다.



<동아시아 4개 언어 한자어 데이터베이스를 위한 실무형 RDB 스키마 설계 샘플>

원, 136-144쪽.

morphID	RelationType	ko-Hang	ko-Hant	ko-Latn	zh-Hans	zh-Hant	zh-Latn	ja-Hani	ja-Hant	ja-Hira	ja-Latn	vi-Hant	vi-Latn
1	동음이의어	보살	菩薩		菩萨	菩薩	pusa		菩薩	ぼさつ		菩薩	bồ tát

<형태(morph) 테이블 샘플 데이터>

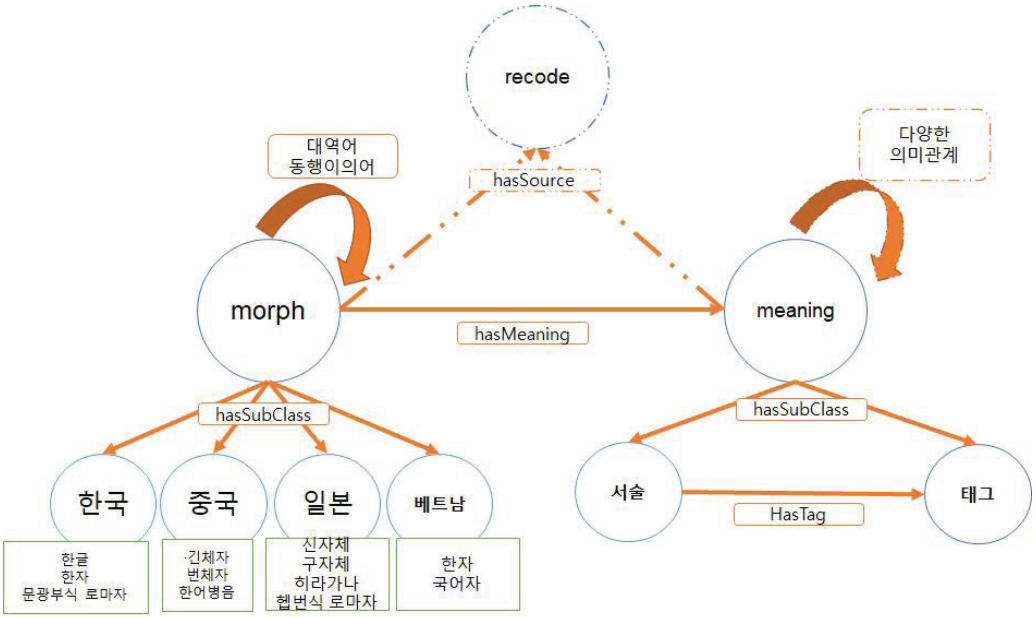
meaningID	morphID	LanguageType	LanguageType	smntc
1	1	ko	<불교> 부처가 전생에서 수행하던 시절, 수기를 받은 이후의 몸.	부처
2	1	ko	<불교> 위로 보리를 구하고 아래로 중생을 제도하는, 대승 불교의 이상적 수행자상. ≒보리살타, 살타, 상사, 진신.	수행자
3	1	ko	<불교> 삼승(三乘)의 하나. 보살이 큰 서원(誓願)을 세워 위로 보리를 구하고 아래로 중생을 교화하는 교법을 이른다. =보살승.	교화방법
4	1	ko	<불교> 여자 신도(信徒)를 높여 이르는 말.	여자신도
5	1	ko	<불교> '고승'(高僧)을 높여 이르는 말. [6] <불교> 머리를 깎지 않고 절에서 사는 여자 신도. =보살할미.	고승
6	1	zh	佛教名词。梵文菩提萨埵(Bodhisattva)之省，原为释迦牟尼修行而未成佛时的称号，后泛用为对大乘思想的实行者的称呼。	부처
7	1	zh	指人们崇拜的神灵偶像。	숭배대상
8	1	zh	比喻心肠仁慈的人。	인격자
9	1	ja	仏語。もと、釈迦牟尼の前生における呼称。大乗仏教が興って、修行を経た未来に仏になる者の意で用いる。悟りを求め修行するとともに、他の者も悟りに到達させようと努める者。また、仏の後継者としての、観世音、彌勒、地蔵など。	부처

<의미(meaning) 테이블 샘플 데이터>

1.2. 제안 RDF 데이터 스키마

발표문에서도 말씀하셨다 싶이, 다층적인 언어 관련 정보의 관계 설정을 위해서 RDB 에서 RDF 데이터로의 전환을 생각해보는 것도 분명 유의미하게 생각합니다. 발표문을 보고, 제가 떠올린 간략한 RDF 데이터 스키마는 다음과 같습니다.

파란색은 클래스(Class), 빨간색은 오브젝트 프로퍼티(Object Property), 녹색은 데이터 프로퍼티(Data Property)을 뜻합니다. 기본적으로 형태, 의미, 기록을 코어 클래스로 두고, 이들을 상호 연결하는 방식으로 의미망을 구성하는 안입니다. 물론 세부 내용은 "동아시아 4 개 언어 한자어 데이터베이스"의 데이터의 구축계획과 진행에 따라서 변경되어야할 것입니다.



<동아시아 4개 언어 한자어 데이터베이스를 위한 온톨로지 설계 개념도>

2. 미래 방향성 관련

2.1. 역사적 기록물 및 이미지 아카이빙

본 발표에서 소개하는 "동아시아 4개 언어 한자어 데이터베이스"는 동아시아 4개 언어의 한자어에 대한 정보를 집적하여 데이터베이스화하는 한편 어휘와 관련한 역사적 기록물 및 이미지를 아카이빙 하는 것"입니다. 그런데 저는 발표문에서 "어휘와 관련한 역사적 기록물 및 이미지 아카이빙"에 대한 내용을 찾지 못했습니다. 아마 2단계 계획이기에 생략된 부분이 있는 듯 합니다. 간략하게 나마, 아카이빙 대상 이미지와 기존 어휘 정보와의 연계에 대해서 설명해주셨으면 합니다.

2.2. 사용역과 전문용어

사용역과 전문용어를 추가하는 방안은 분명 유의미할 것으로 생각합니다. 다만, 조금만 더 구체적으로 실제 해당 정보를 입력하는 샘플과 이를 통해서 도달할 수 있는 목표 혹은 활용 방안에 대해서 서술했으면 하는 아쉬움이 있습니다.

2.3. 한자어의 생성, 이동, 변이

3단계의 한자어의 생성, 이동, 변이를 위한 아카이빙이 궁금합니다. 토론자의 짧은 생각으로는, 한자어를 시계열적으로 보기 위해서는 현 시점의 한자어 정보 이외에도 현대 이전의 모든 시기를 포괄하는 한자어 역사(시계열) 아카이브가 필요할 것으로 생각합니다. 이를 어떤 방식으로 구현하실 생각이신지 궁금합니다.