

동아시아 4개 언어 한자어 데이터베이스의 구축과 과제*

신 응 철**

국문초록

본 논문에서는 한국어, 중국어, 일본어, 베트남어의 한자어를 어휘의 의미를 중심으로 연결한 ‘동아시아 4개 언어 한자어 데이터베이스’의 구축 과정과 현황에 대해 소개하고 향후 보완 및 확장의 가능성에 대한 시론을 제시하였다. 이 데이터베이스는 한자어라는 공통 문명 자산을 통해 동아시아 한자문명의 성격을 분석하고 규명하기 위한 목적으로 구축되었다. 기존에 시도된 동아시아 여러 언어의 한자어를 망라하는 데이터베이스 구축에서는 표기를 매개로 서로 다른 언어의 어휘를 연결한 것이었기 때문에 동형 한자어가 각 언어에서 갖는 어휘적 위상의 차이가 적절히 반영되지 못하는 한계가 확인된다.

본 데이터베이스의 설계에서는 단어의 표기나 형태가 아닌 의미적 대응 관계에 주안점을 두었다. 우선 데이터베이스 구축의 기준점이 되는 한국어 표제어를 “한국어 기초사전”에 수록된 한자어로부터 선별하였다. 이어서 한국어 표제어에 의미적으로 대응할뿐만 아니라 사용역의 측면에서도 격차가 크지 않은 중국어, 일본어, 베트남어의 대역어를 선정하였다.

이렇게 선별하여 확보한 표제어 및 대역어는 언어별로 4개의 별도 테이블로 나누어 작성하였고 서로 대응하는 표제어와 대역어에는 동일한 ID를 부여하였다. 의미적 대응에 기반한 ID를 통해 각 언어별 테이블의 어휘가 관계형 데이터베이스의 구조로 연결된다. 또한 동아시아 네 언어의 각 어휘에는 각 언어 별로 일정 규모 이상의 사전으로부터 의미정보를 인용하여 입력하였고, 그것을 바탕으로 어휘의 의미속성을 대표할만한 태그를 별도의 필드로 수록하였다.

본 데이터베이스의 한계로는 개별 어휘라는 한정된 층위의 개체 정보를 평면적으로 수록한 닫힌 구조라는 점을 지적할 수 있다. 향후 RDF 형식으로 개별 어휘 간의 관계성을 입체적으로 기술한 데이터셋을 작성하여 데이터의 시각화와 타 데이터베이스와의 연계를 용이하게 함으로써 본 데이터베이스의 사용성과 학술적 이용 가치를 높일 수 있을 것이다. 또한 본 데이터베이스를 통해 창출된 동아시아 한자문명에 대한 연구 성과가 다시 데이터베이스의 확장 및 보완으로 반영되는 순환적 모델을 지향해야 한다.

[주제어] 한자어, 관계형 데이터베이스, 언어 횡단적 어휘 연구, 관계성, 의미속성

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A6A3A02043693)

** 경성대학교 한국한자연구소 HK연구교수 / shin.wngchL@gmail.com

— 목 차 —

- | | |
|-----------------------|-------------------------|
| I. 서론 : 데이터베이스 구축의 배경 | IV. 동아시아 한자어 데이터베이스의 확장 |
| II. 앞선 사례와 문제점 | V. 결 론 |
| III. 데이터베이스의 설계와 구축 | |

I. 서론 : 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 문명권²⁾을 형성하는 기능을 수행했다. 한자는 표의문자로서의 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 일방향이적이기만 했던 것은 아니며, 특히 서양 문명의 전면적 수용이 시작된 19세기 이후로는 보다 역동적이고 다방향적인 양상을 띠었다. 오늘날 동아시아 한자문명을 관통하는 문명적 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. 오늘날 동아시아 각 언어권 안에서 한자어가 갖는 의미적 특성에 주목하여, 그것을 바탕으로 문명의 전파와 변용의 궤적을 추적하여 동아시아의 문명 형성의 실체를 규명해 볼 수 있다. 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.

그러한 연구의 수행을 위해서는 우선 동아시아 4개 언어의 한자어가 갖는 의미를 단초로 삼아, 그것에 투사된 다양한 문명적/문화적 특성 고찰로 나아가는 것을 지향할 필요가 있다. 예컨대 동아시아 외부에서 도래한 불교적 개념을 나타내는 Bodhisattva는 음역되어 菩薩이라는 한자어로 동아시아 문명 내의 각 언어권으로 확산되었다. 각 언어 내에서 오늘날 이 한자어의 각 센스를 대표하는 키워드를 나열하면 다음과 같다.

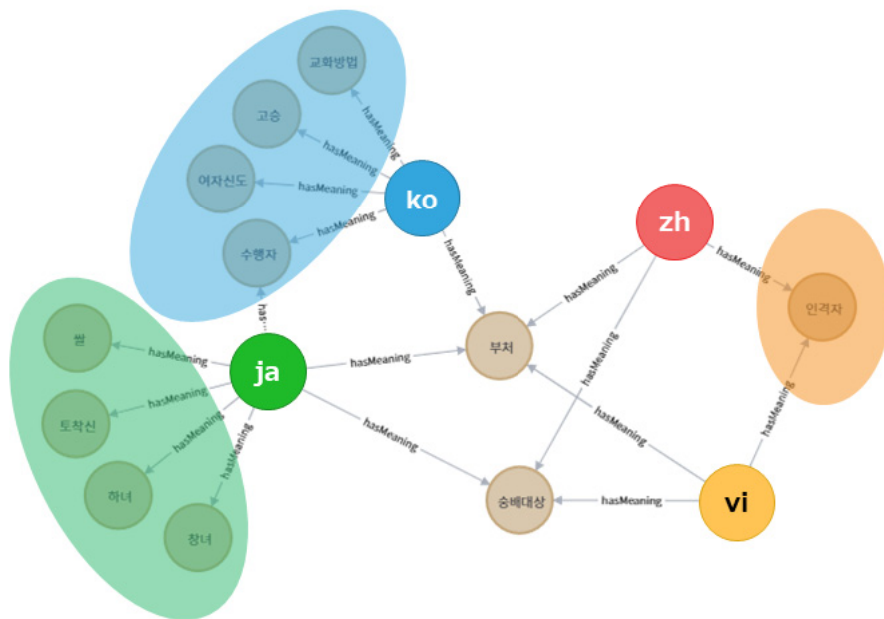
1) "Again, language does not exist apart from culture, that is, from the socially inherited assemblage of practices and beliefs that determines the texture of our lives." Sapir, Edward, *Language: An introduction to the study of speech*, New York: Harcourt, Brace and Company, 1921. pp.221~222, <https://www.gutenberg.org/ebooks/12629>

2) 이하 본 논문에서는 '문화'를 개별 언어 혹은 정치적 경계 내의 동질적 현상으로 정의하고, '문명'은 초언어적, 초국가적 규모의 동질적 현상으로 정의하여 사용하기로 한다.

〈동아시아 4개 언어의 한자어 菩薩의 센스 대표 키워드〉

- 한국어 菩薩(posal) : 부처; 수행자; 교화방법; 여자신도; 고승;³⁾
 중국어 菩薩(pusa) : 부처; 숭배대상; 인격자;⁴⁾
 일본어 菩薩(bosatsu) : 부처; 숭배대상; 토착신; 고승; 수행자; 딸; 하녀; 창녀;⁵⁾
 베트남어 菩薩(bồ tát) : 부처; 숭배대상; 인격자;⁶⁾

또한 이들 각 언어별 센스의 대표 키워드의 연결을 가시화하면 다음과 같다.



〈그림 1〉 동아시아 4개 언어의 한자어 菩薩의 센스 비교

- 3) 『표준국어대사전』 보살[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.
- 4) 『漢語大詞典』 菩薩: [1] 佛教名詞。梵文菩提薩埵(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。
- 5) 『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。大乘仏教が興って、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、地藏など。[2] 昔、朝廷から碩徳の高僧に賜った号。[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。
- 6) 『Từ điển tiếng Việt』 Bồ tát: (thường viết hoa) người tu hành đặc đạo trong đạo Phật, có hiểu biết rộng, có đức độ cao.

불교어휘로서 동아시아에 수용되었던 당초의 의미(부처, 숭배대상, 수행자)가 공통되게 분포하는 가운데, 개별 언어별로 특수한 의미 파생 또한 관찰된다. 특히 한국어(K)와 일본어(J)에서는 당초의 불교어휘로서의 의미만을 떠올렸을 때 선뜻 연결이 어려운 파생 의미(여자신도, 쌀, 하녀, 창녀 등)가 존재한다는 점은 주목된다. 이러한 파생 의미의 존재는 어휘 연구를 언어 외적인 현상으로 확장하는 단초를 제공한다. 그러한 연구 시야의 확장은 불교라는 종교가 각 언어권의 문화에서 어떻게 개별적으로 수용되었으며, 그것의 문화적 위상과 기능이 어떻게 변화하였는지에 대한 이해를 토대로 수행되어야 할 것이다. 菩薩이라는 어휘의 의미 파생이 그러한 문화적 현상을 어떻게 투영하고 있는가에 대해 고찰해 볼 수 있을 것이다.

이렇듯 한자어를 바탕으로 한 동아시아 문명 연구의 수행을 다음과 같은 세 단위(unit)로 나누어 시행하고자 한다. 각 단위 간의 관계에 대한 설명에 앞서 일러두자면, 그것은 선후관계라기보다는 순환적인 보완관계로 정의해 볼 수 있다.

〈한자어를 통한 동아시아 문명 연구의 수행 단위〉

- 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&이미지와와의 연계.
기초/문화/상업, 동식물/상업/과학/법률/제도, 종교/사상/번역, 기록물/특수어휘, 한자 사용의 역사/정책 특성
- 3) 문명연구
문화/철학 개념어, 불교/유교/기독교/도교 개념어, 근현대 번역어, 유행어/신조어, 한자문화 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 아카이빙 한다. 둘째, 수집된 한자어를 어휘 분야별로 나누어 개별 언어권 간의 공통성과 특수성을 비교 검토한다. 또한 아카이빙 된 기록물 및 이미지는 어휘와의 연계 정보를 부여하여 실제적 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 논문에서 소개하는 ‘동아시아 4개 언어 한자어 데이터베이스’(이하 동아시아 한자어DB로 약칭)는 위 세 단위 가운데 첫번째 단계에 해당하는 작업의 일환으로 구축된 것이다.

II. 데이터베이스 구축의 앞선 사례와 문제점

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”⁷⁾(이하 동형한자어DB로 약칭)를 들 수 있다. 이 동형한자어DB의 구축 데이터 항목을 정리하면 다음과 같다.⁸⁾

〈표 1〉 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

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 다른 기술, 둘째로 음운적 유사성 지표 설정을 통한, 외국어

7) 名古屋大学大学院国際言語文化研究科・玉岡賀津雄研究室, 「日韓中越同形二字漢字語データベース」, 2015, <https://kanjigodb.herokuapp.com/>

2014년 기준으로 일본어, 한국어, 중국어 동형한자어를 대상으로 한 품사 정보 중심의 데이터베이스가 먼저 구축되었으며(朴善嫻・熊可欣・玉岡賀津雄, 「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, 2014, 3~23쪽), 이후 베트남어 동형한자어 정보와 음운유사성 지표가 추가된 것으로 보인다(于劭贊・玉岡賀津雄・ホアーン ティ ラン フォン, 「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, 2019, 75~94쪽).

8) 朴善嫻・熊可欣・玉岡賀津雄, 「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, 2014, 5~6쪽.

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.

예컨대 동형한자어DB에서는 일본어에서 일상적으로 쓰이는 단어인 名刺(めいし, 명함)의 동형한자어로 중국어 名刺(mingci), 한국어 名刺(myengca)와 같은 일상적으로 사용되지 않는 어휘를, 베트남어는 名刺(danh thiếp)를 제시하고 있다.⁹⁾ 이것은 이들 동아시아 4개 언어의 한자어에 대해서 동형한자어를 기준으로 연결해 나열했을 때 각 언어 안에서의 어휘적 위상이 입체적으로 드러나지 못하는 문제점을 단적으로 보여준다.

2. 동형 표기어를 기준으로 한 어휘선정의 문제점

동아시아 4개 언어는 어휘적으로 한자와 한자어라는 요소를 갖는다는 공통점이 있지만, 개별 언어 안에서 의 운용에는 크고 작은 차이가 있다. 따라서 이들을 적절히 비교 고찰하기 위해서는 각 언어 안에서 개별 한자어의 어휘적 위상을 입체적으로 기술하는 지표로서 ‘사용역(register)’과 ‘난이도’에 대한 고민이 필요하다.

앞서 살펴본 동형한자어DB를 비롯하여 기존의 동아시아 개별 언어 간의 한자어 비교 연구는 같은 한자¹⁰⁾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 다른 한자로 표기되지만 의미적으로 동가에 가까운 어휘인 대역어(equivalent)가 분석의 시야에 들기 어렵다는 문제도 존재한다.¹¹⁾

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

9) 중국어의 경우 『现代汉语词典』(第7版)에 “《书》名片。”와 같이 서면에서 쓰이는 용어임을 약물로 명시하고 보다 일상적인 어휘인 名片으로 뜻풀이를 대신하였다. 한국어의 경우 『표준국어대사전』에 ‘명자²(名刺)’라는 표제어가 등재되어 있는 있으나 오늘날 ‘명함(名衡)’에 비해 현저히 사용빈도가 낮은 어휘이다. 한편 베트남어 동형한자어로 제시된 名刺(danh thiếp)는 어휘 자체는 일상적으로 사용되는 것으로 파악되지만 名帖으로 표기하는 것이 적당한 단어이기에 동형한자어로 간주하기 어렵다.

10) 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.

11) 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.

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.

Ⅲ. 데이터베이스의 설계와 구축

1. 데이터베이스의 설계

동아시아 한자어DB는 우선 각 언어별 평면 테이블을 기반으로 하는 관계형 데이터베이스(Relational Database, RDB)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 어휘 속성〉

한국어(ko) : [ID]-[ko-Hang]-[ko-Hant]-[ko-Latn]-[ko-def]-[ko-smntc]

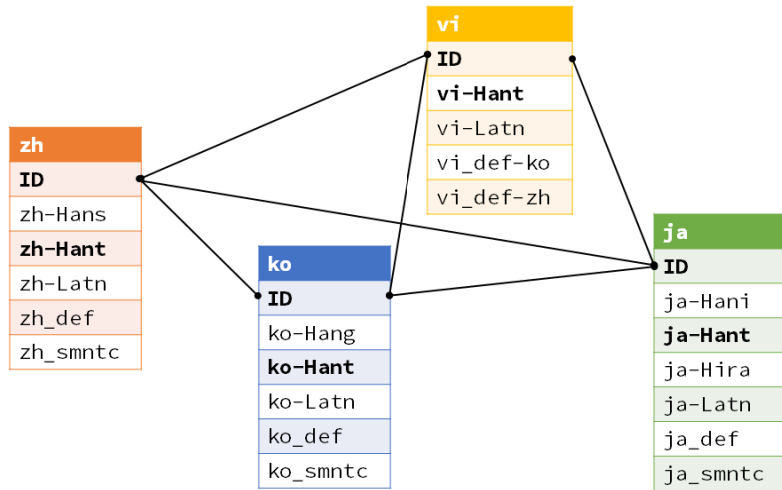
중국어(zh) : [ID]-[zh-Hans]-[zh-Hant]-[zh-Latn]-[zh-def]-[zh-smntc]

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.¹²⁾ 각 언어 테이블의 속성에 대해서는 후술하는 데이터 구축의 과정을 통해 설명하겠다.

12) 본고에서 제시하는 관계형 데이터베이스 설계의 합당성에 대해서는 ‘제12회 동양학연구원 사전학 학술회의: “동아시아의 사전학(XII)” 어휘 의미망 구축과 사전’에서 토론자 김바로 교수(한국학중앙연구원)로부터 문제에 대한 지적과 함께 보다 합당한 대안을 교시 받았다. 다만 기존 설계의 전면적 수정에 소요되는 시간과 비용의 규모를 가늠하기 어렵다 판단되어, 본 논문에서는 당초 설계에 따라 서술하고, 설계의 전면 수정에 대해서는 향후 추가적인 연구를 통해 공표하기로 한다.



〈그림 2〉 동아시아 4개 언어 한자어 DB 언어별 테이블의 속성

2. 기준언어 표제어와 비교언어 대역어 선정

국립국어원에서 제공 중인 “한국어기초사전”¹³⁾은 조남호(2002)를 비롯한 한국어 학습용 어휘목록과 『연세한국어사전』과 같은 학습자 사전의 표제어, 그리고 한국어 교재와 한국어 능력시험에 등장하는 단어를 말뭉치화한 것을 활용하여 약 5만여 표제어(단어, 구)를 수록하였다.¹⁴⁾ 이것은 한국어 교육 및 학습의 영역에서 일정한 사용빈도와 상용도가 인정되는 어휘의 목록으로 간주할 수 있다. 이와 같은 일정한 사용역의 어휘군을 기준 표제어로 삼아 다른 언어의 대역어와 비교하는 방식을 취함으로써, 앞서 동형한자어를 기준으로 접근했을 때의 문제점으로 지적한 서로 다른 언어의 어휘 간의 사용역 격차 발생을 최소화하고자 하였다.

본 데이터베이스에서는 그 “한국어기초사전”의 수록 표제어 가운데 한자어(단어)만을 추려 총 23,000여 단어를 추출하여 기준 표제어로 삼았다.¹⁵⁾ 기준 표제어 중에는 각 언어 간의 비교 대상으로서 유의미한 결과가 기대되지 않는 어휘¹⁶⁾를 배제하였고 최종적으로 21,763개 단어로 압축되었다. 확정된 표제어에 대해서는 다음과 같은 항목으로 표기 정보를 입력하였다.

13) 국립국어원, 「한국어기초사전」, 2016, <https://krdict.korean.go.kr/>.

14) 강현화·신자영·원미진, 「한국어 학습자 사전 표제어 선정을 위한 자료 구축 및 선정 방법에 관한 연구」, 『한국사전학』 16, 2010, 18~24쪽.

15) 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.

16) 배제된 어휘는 다음 세 가지 유형으로 분류된다. (1) 접사 및 파생어(-가(街), -증(症), 비정상적, 신경질적), (2) 고유명사(두만강, 신라, 전라남도), (3) 한국 특유 어휘(동고서저, 보신탕, 수학능력시험, 암행어사). 다만 접사 이외에 한자 한 글자로 구성된 어휘는 배제하지 않았다.

〈한국어(ko) 표제어의 어휘 속성 개요〉

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

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

[ko-Hangl] : 한국어 표제어의 한글 표기

[ko-Hant] : 한국어 표제어의 한자 표기

[ko-Latn] : 한국어 표제어의 문광부식 로마자 표기¹⁷⁾

이들 21,763개 표제어의 형태적 특성을 분석해 보면 아래 표와 같이 2자 한자어 12,666개, 3자 한자어 6,218개, 4자 한자어 1,782개로 2~4자 한자어가 전체의 95%를 차지한다.

〈표 2〉 한국어 표제어의 형태적 특성 : 한자수

한자수	1	2	3	4	5≤	합계
표제어수	634 (3%)	12,666 (58%)	6,218 (29%)	1,782 (8%)	463 (2%)	21,763

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)¹⁸⁾의 대역어를 우선적으로 활용하여 연구진과 원어 화자의 다중 검토 방식으로 대역어를 제의 또는 수정, 추가하였다.¹⁹⁾²⁰⁾ 한편 대역어 선정 작업을 진행하던 2019년 기준으로 “한국어기초사전”과 연계한 대역 학습사전이 공개되지 않았던²¹⁾ 중국어에 대해서는 “에듀월드 한한중사전”²²⁾과 “에듀월드 중중한사전”²³⁾의 원 데이터를 가공하여 연구진과 원어 화자의 다중 검토 방식으로 대역어를 확정하였다. 아울러 이들 문헌을 바탕으로 대역어를 선정하는 원칙은 다음과 같다.

- 17) 국립국어원, 「국어의 로마자 표기법」(문화체육관광부 고시 제2014-42호, 2014.12.05.), 2014, https://kornorms.korean.go.kr/regltn/regltnView.do?regltn_code=0004®ltn_no=444#a444.
- 18) 국립국어원, 「국립국어원 한국어-일본어 학습사전(国立国語院 韓国語-日本語学習辞典)」, 2016, <https://krdict.korean.go.kr/jpn/>. 국립국어원, 「국립국어원 한국어-베트남어 학습사전(Từ điển học tiếng Hàn-tiếng Việt của Viện Quốc ngữ Quốc gia)」, 2016, <https://krdict.korean.go.kr/vie/mainAction>.
- 19) 대역어 선정 작업도 한국어 표제어 추출 시와 마찬가지로 「한국어기초사전」의 사전 내려받기 기능을 활용하여 다운로드한 데이터를 기반으로 이루어졌다.
- 20) 「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.
- 21) 국립국어원 한국어-중국어 학습사전(国立国語院韓国語-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」(2020.05.22.), 2020, https://korean.go.kr/front/board/boardStandardView.do?board_id=6&mn_id=184&b_seq=754.
- 22) 차이나랩, 「에듀월드 표준한중사전」, 2018, <https://zh.dict.naver.com/>. 약 85,000단어를 표제어로 수록.
- 23) 차이나랩, 「에듀월드 표준중중한사전」, 2015, <https://zh.dict.naver.com/>. 약 150,000단어를 표제어로 수록.

〈동아시아 한자어DB의 대역어 선정 기본원칙〉

- 해당 언어를 전공하는 한국어를 모어로 하는 연구진과 해당 언어를 모어로 하는 연구진이 함께 참여하여 교차 검토
- 한국어 표제어의 각 센스를 최대한으로 포섭하는 대역어를 선정
- 대역어 후보 중에서는 동형 한자어-이형 한자어 순으로 우선순위를 부여
- 일본어의 경우에는 한자어(음독어) 이외에 한자로 표기할 수 있는 고유어(훈독어) 포함²⁴⁾
- 한자로 표기할 수 없는 고유어는 배제
- 구, 문장은 배제
- 복수 대역어 인정을 최소화

이와 같은 방식으로 중국어 23,091개, 일본어 19,517개, 베트남어 7,761개²⁵⁾의 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.

〈중국어(zh), 일본어(ja), 베트남어(vi) 대역어의 어휘 속성 개요〉

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

- [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 표기²⁶⁾

vi : [ID]-[vi-Hani]-[vi-Latn]

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

24) 일본어에서 한국어와 중국어로 유입된 한자어 중에는 본래 고유어(훈독어)로 분류되는 어휘가 적지 않다는 점을 감안한 조치이다. 만약 일본어의 한자어를 음독어만으로 한정할 경우, 고유어의 한자표기어가 주변 언어에 수용된 取消(とりけし), 割引(わりびき) 등의 어휘가 대상에서 배제되는 결과로 이어질 수밖에 없다.

25) 베트남어의 경우 일상적 범위의 사용역을 갖는 어휘에서 한자어의 비율이 높지 않은 것으로 보인다. 뿐만 아니라 비교적 일상 언어생활에서 사용 빈도가 높은 전문용어 중에 대해서도 비슷한 양상이 확인된다. 이에 대해서는 근대 이후 베트남어의 신어 수용 및 조어의 경향을 바탕으로 추가적인 고찰이 필요할 것으로 생각된다.

26) 日本国外務省, 「ヘボン式ローマ字綴方表」, 2017, <https://www.ezairyu.mofa.go.jp/passport/hebon.html>.

3. 기준언어 표제어와 비교언어 대역어 간의 대응 개요

이렇게 선정된 한국어 표제어와 중국어, 일본어, 베트남어의 대역어 간의 대응 양상은 다음과 같다. 한국어 이외의 3개 언어가 단수 대역어로 대응하는 것은 16,852개이며, 이 가운데 3개 언어 모두로부터 대역어를 수록한 어휘는 4,628개, 2개 언어로부터만 대역어를 수록한 어휘는 10,781개, 1개 언어로부터만 대역어를 얻은 어휘는 1,443개이다. 3개 언어, 2개 언어 대역어가 수록된 어휘에 대해 한자 표기의 같고 다를 것을 기준으로 유형화하면 아래 표와 같다.

〈표 3〉 동아시아 4개 언어 한자어의 한자표기 대응 유형(3개 언어 대역어 수록)

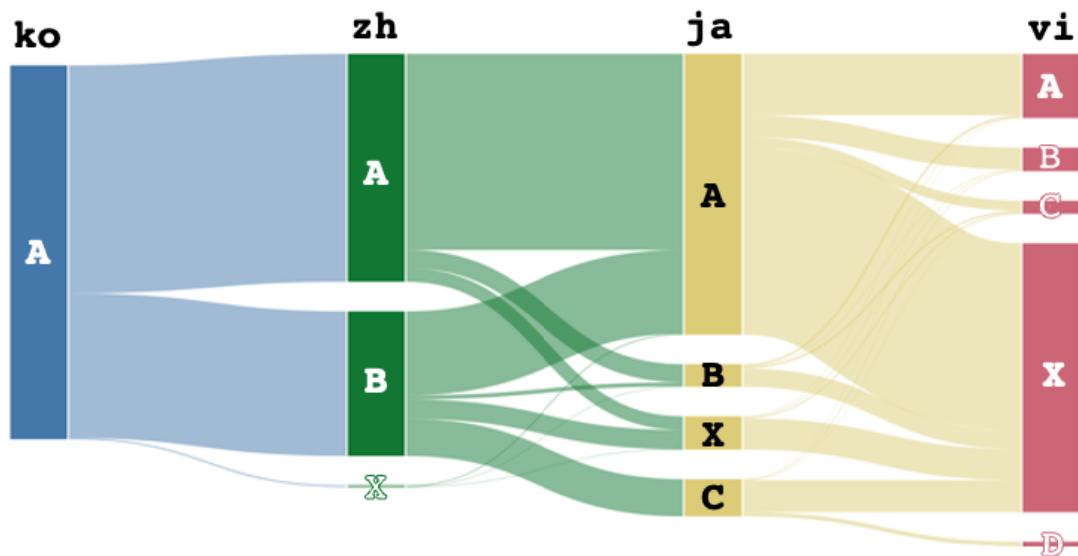
한국어	중국어	일본어	베트남어	대응어휘
A	A	A	A	2,632 (57%)
A	A	A	B	886 (19%)
A	A	B	A	88 (2%)
A	B	A	A	130 (3%)
A	A	B	B	8 (0%)
A	B	A	B	99 (2%)
A	B	B	A	2 (0%)
A	A	B	C	79 (2%)
A	B	A	C	447 (10%)
A	B	C	A	13 (0%)
A	B	B	B	15 (0%)
A	B	C	B	29 (1%)
A	B	C	C	9 (0%)
A	B	C	D	191 (4%)
합계				4,628

〈표 4〉 동아시아 4개 언어 한자어의 한자표기 대응 유형(2개 언어 대역어 수록)

한국어	중국어	일본어	베트남어	대응어휘
A	A	A	X	5,316 (49%)
A	A	X	A	20 (0%)
A	X	A	A	6 (0%)
A	A	B	X	644 (6%)
A	A	X	B	22 (0%)
A	B	A	X	3,107 (29%)
A	B	X	A	1 (0%)
A	X	A	B	2 (0%)

한국어	중국어	일본어	베트남어	대응어휘
A	X	B	A	0 (0%)
A	B	B	X	168 (2%)
A	B	X	B	1 (0%)
A	X	B	B	1 (0%)
A	B	C	X	1,439 (13%)
A	B	X	C	53 (0%)
A	X	B	C	1 (0%)
합계				10,781

위 표에서는 기준이 되는 한국어 표제어의 한자표기(A)와 같을 경우에는 A, 다를 경우에는 B, C, D 등으로 표시하였으며, 대응하는 대역어를 수록하지 못한 경우에는 X로 나타내었다. 예컨대 A-A-A-A로 표시된 것은 네 언어의 표제어 및 대역어의 한자 표기가 모두 같음을 의미하며, A-A-A-X로 나타난 것은 베트남어를 제외한 나머지 세 언어의 한자 표기가 같음을 뜻한다. 16,852개 표제어에 대한 대역어 대응 양상을 생키 다이어그램(Sankey Diagram)으로 시각화하면 다음과 같다.



〈그림 3〉 동아시아 4개 언어 한자어의 대응(Sankey Diagram)

한국어(ko) 표제어에 대해서 중국어(zh) 대역어는 이형 한자어의 비율이 두드러지는 반면, 일본어(ja) 대역어에서는 동형 한자어 비율이 현저히 높은 점을 확인할 수 있다. 이것은 예측 가능한 사실이라 할 수 있으나 본 데이터베이스 구축을 통해 계량적으로 다시 한번 확인할 수 있었다. 한편 베트남어(vi) 대역어에서는 한자어의 비율이 나머지 언어에 비해 낮은 점을 확인할 수 있다.²⁷⁾

4. 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”의 의미정보를 바탕으로 ‘학습기초어휘사전’의 쓰임에 걸맞게 가공된 것이라 소략한 경우가 있다. 또한 일본어, 베트남어로 번역된 의미정보는 ‘한국어 어휘’에 대한 의미정보를 각 언어로 번역한 것이라는 점에서 본 데이터베이스가 지향하는 방향에 부합하지 않는다.

본 데이터베이스가 대상으로 하는 네 언어의 어휘(표제어, 대역어)에 대한 뜻풀이를 다음의 각 문헌으로부터 추출하여 입력하였다.

〈표 5〉 동아시아 네 언어 어휘의 의미정보 인용 문헌

언어	이용 문헌명	수록 표제어 규모	입력항목
한국어(ko)	표준국어대사전	약 430,000개	[ko-def]
중국어(zh)	에듀월드 중중한사전	약 150,000개	[zh-def]
일본어(ja)	고지엔 일한사전	약 240,000개	[ja-def]
	日本国語大辞典 제2판	약 500,000개	
베트남어(vi)	Từ điển tiếng Việt	약 40,000개	[vi-def-ko] [vi-def-zh]

한국어 표제어에 대해서는 국립국어원의 “표준국어대사전”,²⁸⁾ 중국어 대역어 및 동형이의어에 대해서는 “에듀월드 중중한사전”, 일본어 대역어 및 동형이의어에 대해서는 “고지엔 일한사전”과 “日本国語大辞典” 제2판으로부터 의미정보를 확보하였다. 베트남어의 경우는 베트남 사회과학원 산하 언어연구원²⁹⁾의 “베트남어 사전(Từ điển tiếng Việt)”³⁰⁾³¹⁾의 베트남어 뜻풀이를 한국어와 중국어로 번역하여 추가하였다. 이들 문헌으로부터 의미정보를 추출하는 과정에서는 다음의 방침을 적용하였다.

27) 일본어, 베트남어에서 한자어 대역어를 찾을 수 없었던 경우에 대해서는 향후 비 한자어까지 시야에 넣은 종합적인 검토가 필요한데, 이는 향후 데이터베이스 확장의 한 방향으로서 중요한 부분이다.

28) 국립국어원, 『표준국어대사전』, 2008, <https://stdict.korean.go.kr/>.

29) Vietnam Institute of Linguistics(Viện Ngôn ngữ học), <http://www.vienngonnguhoc.gov.vn/>.

30) Hoàng Phê, Bùi Khắc Việt, and Viện ngôn ngữ học, *Từ điển tiếng Việt*, Đà Nẵng: Nhà xuất bản Đà Nẵng, 2015.

31) 베트남어 사전은 현존하는 베트남어 단일 언어사전으로는 최대 규모로 알려져 있으나, 한국어, 중국어, 일본어 사전과 비교했을 때 수록 표제어의 규모가 작을뿐만 아니라 의미정보의 심도 및 정밀함에서도 눈에 띄는 격차가 있는 것으로 확인되었다.

〈동아시아 네 언어 어휘의 의미정보 인용 방침〉

[한국어/중국어/일본어]

- 이용 문헌에 한국어로 기술된 의미정보 입력
- 단 日本國語大辭典은 일본어 원문에서 한국어로 번역하여 수록

[베트남어]

- 베트남어로 기술된 의미정보를 한국어와 중국어로 번역하여 수록
- 연구진(베트남어 모어 화자)의 주도로 베트남 현지 한국어 및 중국어 전공 인력 참여

이들 사전으로부터 추가한 의미정보를 바탕으로 각 언어별로 의미 속성을 나타내는 태그(smntc)를 추가하였다. 의미분류의 체계는 부산대학교 인공지능연구실과 한국어정보처리연구실에서 WordNet³²⁾ 기반으로 구축한 한국어 어휘의미망(KorLex)³³⁾을 참조하여 입력하였다. 먼저 한국어 표제어에 대한 입력 작업을 진행한 후, 나머지 세 언어의 대역어는 대응하는 한국어 표제어의 의미속성 태그를 토대로 사전에서 인용한 의미정보를 참고하여 추가 혹은 삭제가 필요한 부분을 점검하는 방식으로 진행하였다. 이에 따라 각 언어 별 테이블에 수록되는 정보는 최종적으로 다음과 같다.

〈동아시아 4개 언어 한자어의 의미정보 및 의미속성 태그에 대한 개요〉

ko : [ID]-[ko-Hang]-[ko-Hani]-[ko-Latn]-[ko-def]-[ko-smntc]

[ko-def] : 한국어 표제어의 “표준국어대사전” 의미정보

[ko-smntc] : 한국어 표제어의 KorLex 상위 의미범주([ko-def] 참고)

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

[zh-def] : 중국어 대역어의 “예듀월드 중중한사전” 의미정보

[zh-smntc] : 중국어 대역어의 KorLex 상위 의미범주([zh-def] 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([ja-def] 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)

32) Princeton University, 「Princeton University WordNet」, 2011, <https://wordnet.princeton.edu/>.

33) 부산대학교 인공지능연구실 · 한국어정보처리연구실, 「한국어 어휘의미망 (KorLex)」, 2009, <http://korlex.pusan.ac.kr/>.

서두에서 언급한 한자어 菩薩을 예로 들어 각 언어별 테이블 중 의미속성 태그를 제외한 속성을 총람하면 다음과 같다.

〈표 6〉 각 언어 별 테이블의 속성(菩薩)

테이블	속성	내 용
한국어	ID	11736
	ko-Hang	보살
	ko-Hani	菩薩
	ko-Latn	bosal
	ko-def	1. 부처가 전생에서 수행하던 시절, 수기를 받은 이후의 몸. 2. 위로 보리를 구하고 아래로 중생을 제도하는, 대승불교의 이상적 수행자상. 3. 삼승(三乘)의 하나. 보살이 큰 서원(誓願)을 세워 위로 보리를 구하고 아래로 중생을 교화하는 교법을 이른다. 4. 여자신도(信徒)를 높여 이르는 말. 5. ‘고승02’(高僧)을 높여 이르는 말. 6. 머리를 깎지 않고 절에서 사는 여자 신도.
중국어	ID	11736
	zh-Hans	菩薩
	zh-Hant	菩薩
	zh-Latn	púsà
	zh-def	1. 보살. 석가모니가 성불하기 전의 호칭으로 보리살타의 준말. 2. 넓게는 부처와 각종 신을 가리킴. 3. 자비로운 사람. 마음씨가 선량한 사람. 보살 같은 사람.
일본어	ID	11736
	ja-Hani	菩薩
	ja-Hant	菩薩
	ja-Hira	ぼさつ
	ja-Latn	bosatsu
	ja-def	1. <불교> 깨달음을 구해서 수행하는 사람. 원래 성도(成道) 이전의 석가모니 및 전세의 그것을 가리켜서 말했다. 나중에는 대승불교에서 자리(自利)·이타(利他)를 구하는 수행자를 가리키고 자리(自利)만의 소승(小乘)의 성문(聲聞; しょうもん)·연각(緣覺)에 대비하게 되었다. 또한 관세음·지장과 같이 부처에 버금가는 숭배 대상으로도 된다. 2. 조정에서 학덕이 높은 고승(高僧)에게 내린 호. 또는 세인들이 고승을 존경해서 사용하는 호. 3. 도착신에 부여된 칭호. 4. 쌀의 다른 이름. 5. 밥짓는 하녀. 7. 유녀(遊女)의 다른 이름.

테이블	속성	내 용
베트남어	ID	11736
	vi-Hani	菩薩
	vi-Latn	bồ tát
	vi-def-ko	【명】 지혜와 덕이 있는 불교의 득도한 사람
	vi-def-zh	【名】 佛教的修行得道者, 博古通今, 德高望重

의미속성 태그의 입력은 우선 한국어 표제어에 대해서 의미정보(ko-def)를 바탕으로 KorLex 1.5에서 적절한 어휘를 선택하여 검색결과 트리에서 제시하는 상위 속성어 및 병렬적 관계에 있는 어휘를 입력하는 방식으로 진행한다. 센스를 나누어 복수의 항목으로 제시된 경우에는 알파벳 소문자로 구분하였다. 이때 각 상위 속성의 층위는 세미콜론(;)으로 구분하여 나타내고, 같은 층위에서 병렬적 관계에 있는 어휘는 콤마(,)로 구분한다. 그 결과 다음과 같이 a~d 그룹의 의미속성 태그를 얻을 수 있었다.

〈한국어 어휘 菩薩의 의미속성 태그 그룹(ko-smntc)〉

- a : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확산1; 초자연적존재1; 상제1, 신5, 천제1, 하느님1, 하늘2; 보리살타1, 보살1, 진신1;
b : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 리터1; 정신적지주1, 지주2; 불승1, 스님1, 승5, 승가2, 승려1, 중1; 고승2; 보살5;
c : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확산1; 설1, 이즘1, 주의5; 가르침1, 교훈1; 보살2, 보살승1;
d : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 교도1, 교인1, 종교인1; 신도2, 신자1; 여신자1; 보살3, 보살할미1;

중국어와 일본어, 베트남어 대역어에 대해서는 이러한 한국어 표제어의 의미속성 태그를 기준으로 해당 언어의 의미정보와 대조하여 삭제 또는 추가, 수정하는 방식으로 진행하였다. 가령 여자신도라는 의미에 따른 d 그룹의 의미속성은 중국어, 일본어, 베트남어에서는 삭제하였고, 일본어에 대해서는 ‘딸’, ‘하녀’, ‘유녀’에 대응하는 의미속성 e~g 그룹을 추가하였다.

〈일본어 어휘 菩薩에 추가된 의미속성 태그 그룹(ja-smntc)〉

- e : 개체1, 실체1; 물질1, 성분1; 양분1, 영양분1, 영양소1; 식료품1, 식자재1; 곡류1, 곡물1, 곡식1; 미곡1, 쌀1;
f : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 근로자1, 노동자1, 노무자1; 하인1; 하녀1;
g : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 여자1; 여성2, 여인1, 여자1; 노계집1, 매음녀1, 매춘부1, 유녀2, 창녀1, 창부1;

〈표 7〉 각 언어 별 의미속성 태그의 내역 (菩薩)

테이블	속성	내 용
한국어	ko-smntc	a : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확신1; 초자연적존재1; 상제1, 신5, 천제1, 하느님1, 하늘2; 보리살타1, 보살1, 진신1; b : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 리더1; 정신적지주1, 지주2; 불승1, 스님1, 승5, 승가2, 승려1, 중1; 고승2; 보살5; c : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확신1; 설1, 이즘1, 주의5; 가르침1, 교훈1; 보살2, 보살승1; d : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 교도1, 교인1, 종교인1; 신도2, 신자1; 여신자1; 보살3, 보살할미1;
중국어	zh-smntc	a : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확신1; 초자연적존재1; 상제1, 신5, 천제1, 하느님1, 하늘2; 보리살타1, 보살1, 진신1; b : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 리더1; 정신적지주1, 지주2; 불승1, 스님1, 승5, 승가2, 승려1, 중1; 고승2; 보살5; c : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확신1; 설1, 이즘1, 주의5; 가르침1, 교훈1; 보살2, 보살승1;
일본어	ja-smntc	a : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확신1; 초자연적존재1; 상제1, 신5, 천제1, 하느님1, 하늘2; 보리살타1, 보살1, 진신1; b : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 리더1; 정신적지주1, 지주2; 불승1, 스님1, 승5, 승가2, 승려1, 중1; 고승2; 보살5; c : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확신1; 설1, 이즘1, 주의5; 가르침1, 교훈1; 보살2, 보살승1; e : 개체1, 실체1; 물질1, 성분1; 양분1, 영양분1, 영양소1; 식료품1, 식자재1; 곡류1, 곡물1, 곡식1; 미곡1, 쌀1; f : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 근로자1, 노동자1, 노무자1; 하인1; 하인1; 하녀1; g : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 여자1; 여성2, 여인1, 여자1; 노계집1, 매음녀1, 매춘부1, 유녀2, 창녀1, 창부1;
베트남어	vi-smntc	a : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확신1; 초자연적존재1; 상제1, 신5, 천제1, 하느님1, 하늘2; 보리살타1, 보살1, 진신1; b : 개체1, 실체1; 물건1, 물체1; 생물체1, 유기체1; 생물체1, 유기체1; 개인1, 사람1, 영장2, 인간1, 자11; 리더1; 정신적지주1, 지주2; 불승1, 스님1, 승5, 승가2, 승려1, 중1; 고승2; 보살5; c : 정신1, 정신적특징1; 노에시스1, 인식1, 인지1, 인지2, 지식1; 인지내용1; 교조2, 믿음1, 신념1, 확신1; 설1, 이즘1, 주의5; 가르침1, 교훈1; 보살2, 보살승1;

IV. 동아시아 4개 언어 한자어 데이터베이스의 확장을 위한 과제

1. 어휘의 사용빈도와 사용역 속성 추가

각 언어에서 한자어(표제어, 대역어, 동형어)를 보다 입체적으로 조감하기 위해서는 어휘의 사용빈도와 사

용역에 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 언어 횡단적인 활용에는 제약이 따를 것으로 예상된다. 유의어 관계에 있는 어휘 간의 상대적 사용빈도 비교도 하나의 대안으로 생각해 볼 수 있다.

사용역에 대해서는 어떠한 분류체계를 설정할 것인가가 문제이다. 도원영(2008)에 제시된 한국어사전의 표제어에 대한 사용역 분류표³⁴⁾를 바탕으로 가령 ‘북한어’ 등 나머지 3개 언어에 적용하기 어려운 부분을 제외한 분류체계를 고안할 필요가 있다. 한편 “표준국어대사전”의 어휘 속성으로 수록되어 있는 아래 67개 ‘전문분야’³⁵⁾ 정보 또한 어휘의 사용역을 판단하는 정보로서 유효하게 활용할 수 있을 것이다.

〈표준국어대사전 ‘전문분야’ 분류 (67개)〉

- 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

2. 전문용어의 확충

현재 동아시아 4개 언어 한자어 데이터베이스는 한국어 기초어휘 2만 여 개를 기준으로 구축한 것이기 때문에 분야별 전문용어가 충분하지 않다는 한계를 갖는다. 한편으로는 특수 분야에서만 통용되는 전문용어를 일정한 기준 없이 확장하는 것은 단순히 양적 확대 이상의 의미를 갖기 어렵다. 따라서 사용역에 대한 일정한 기준을 마련은 전문용어의 확장을 위한 출발점으로서 중요하다.

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 형식, 이하 교과서 편수용어로 약칭)³⁶⁾의 활용을 고려해 볼 수 있겠다. 교과서 편수용어는 11개 교과 별로³⁷⁾ 사용 가능한 어휘를 수록하고, 각 어휘에 대한 정보를 [용어]-[한자]-[외국어]-[비고]의 항목으로 정리한 어휘 목록이라는 점에서 앞서 언급한 ‘사용역에 대한 일정한 기준’을 충족하는 어휘 목록으로 간주할 수 있다.³⁸⁾ 이 자료에 수록된 어휘는 교육 현장에서 실용적인 필요에 맞춰 선별된 것으로 짐작되지만, 개별 어

34) 도원영, 「국어사전 표제어의 사용역 정보에 대한 고찰」, 『우리어문연구』 30, 2008, 43쪽.

35) 『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.

36) 한국검인정교과서협회, 「편수자료 Ⅱ, Ⅲ 수정 내용 반영본’ 안내」(2017.06.07.), 2017, https://www.ktbook.com/edata/PDS_AuthView.asp?num=327&pageno=1&startpage=1&keyword=%C6%ED%BC%F6%C0%DA%B7%E1.

37) 〈물리〉, 〈미술〉, 〈생명과학〉, 〈수학〉, 〈음악〉, 〈일반사회〉, 〈정보〉, 〈지구과학〉, 〈지리〉, 〈체육〉, 〈화학〉

휘의 선정 근거, 사용 빈도, 난이도 판정 등과 같은 정보가 결여되어 있다는 점은 주의를 요하는 부분이다.

아울러 위 교과서 편수용어와 앞서 언급한 “표준국어대사전”의 67개 ‘전문분야’ 정보를 교차하는 방법을 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”³⁹⁾에 수록된 ‘교과서 말뭉치 어휘표’⁴⁰⁾와 ‘교과 특징어 목록’⁴¹⁾을 들 수 있다. 이들 어휘표 및 어휘 목록은 교과서 말뭉치 구축을 통해 추출해 낸 것으로서 어휘의 난이도(초등학교, 중학교, 고등학교)와 사용 빈도까지 망라하여 유용하다. 한편 중국어와 베트남어에 대해서도 이와 유사한 성격의 어휘 목록의 유무를 점검할 필요가 있다.

3. 개별 어휘 간의 관계성 기술

현재 동아시아 한자어DB에 구축된 데이터는 개별 어휘의 속성에 국한된다. 물론 이러한 속성을 통해 언어 횡단적인 어휘의 상관관계를 추적할 수 있도록 설계하였으나, 관계형 데이터베이스라는 구조의 특성상 개별 어휘 간의 관계성을 정밀하게 확인하거나 기술하기에는 부족한 부분이 있다. 때문에 이러한 개별 어휘 간의 관계성을 언어 횡단적으로 확인하는 작업을 데이터베이스 내적으로 달성하기 어렵고, 결과적으로 데이터베이스에서 찾은 단서를 바탕으로 이용자 각각의 재량으로 추론해야 하는 몫이 커지는 구조이다. 따라서 향후 데이터베이스의 확장 및 보완의 과정으로서 이러한 어휘 간의 언어 횡단적인 관계성을 기계적으로 식별 및 분석할 수 있는 데이터(Machine-readable Data)로 명시적으로 기술할 필요가 있다.

앞서 소개한 동아시아 한자어DB는 단편 구조로 된 개별 어휘라는 한정된 층위의 개체(individual) 또는 접점(node)의 평면적인 목록의 단계에 머물러 있다. 즉, 이들 각 언어[lang] 별 테이블에 수록된 개별 어휘[word]에는 한자 표기[logogram], 표음문자 표기[phonogram]와 같은 형태 정보를 비롯하여 사전에서 인용된 의미 정보[def]와 의미속성의 태그[smntc]가 달려 있는 구조이다.

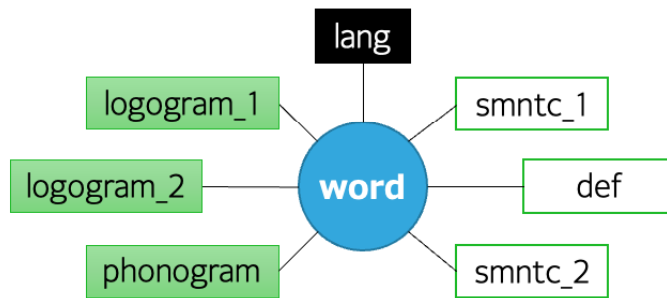
38) 다만 〈수학〉은 [용어] - [동어] - [외국어]로 한자 정보를 수록하지 않았다.

39) 国立国語研究所, 『言語政策に役立つ, コ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, 2011, <http://doi.org/10.15084/00003284>.

이 자료에 수록된 어휘표에 대한 간략한 설명은 홈페이지에서 확인 가능하다.

40) 教科書コーパス語彙表(Ver. 1.0) “2005년도에 사용된 초등학교, 중학교, 고등학교 모든 학년 모든 교과와 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コーパス)”의 도서관 서적(고정 길이 샘플)의 빈도와도 대조할 수 있습니다.” 国立国語研究所コーパス開発センター(Center for Corpus Development, NINJAL), 「教科書コーパス語彙表(Ver. 1.0)」, 2011, <https://ccd.ninjal.ac.jp/bccwj/freq-list.html>.

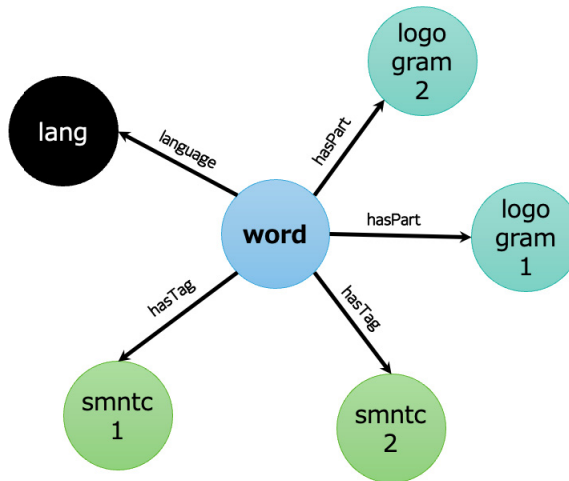
41) 教科書特徴語リスト(Ver. 1.0): “위 ‘교과서 말뭉치(教科書コーパス)’와 ‘도서관 서적(고정 길이 샘플)’의 어휘 빈도를 비교하여 교과 별로 특징어를 추출한 목록입니다. 중학교와 고등학교 각각에 대한 교과별 일람표로 구성됩니다.” 国立国語研究所コーパス開発センター(Center for Corpus Development, NINJAL), 「教科書コーパス語彙表(Ver. 1.0)」, 2011, <https://ccd.ninjal.ac.jp/bccwj/freq-list.html>.



〈그림 4〉 동아시아 한자어DB의 개별 어휘가 갖는 속성

개별 어휘 간을 연결(link)하는 관계성(relation)을 입체적으로 기술한 목록은 부재한 상태이다. 가령 RDF 형식의 개방적인 관계성 목록의 구축을 통해 인공지능 번역 및 데이터의 시각화⁴²⁾ 등으로의 활용을 용이하게 하는 방향으로 개선한다면 본 데이터베이스의 사용성과 학술적 이용가치를 높일 수 있을 것이다.

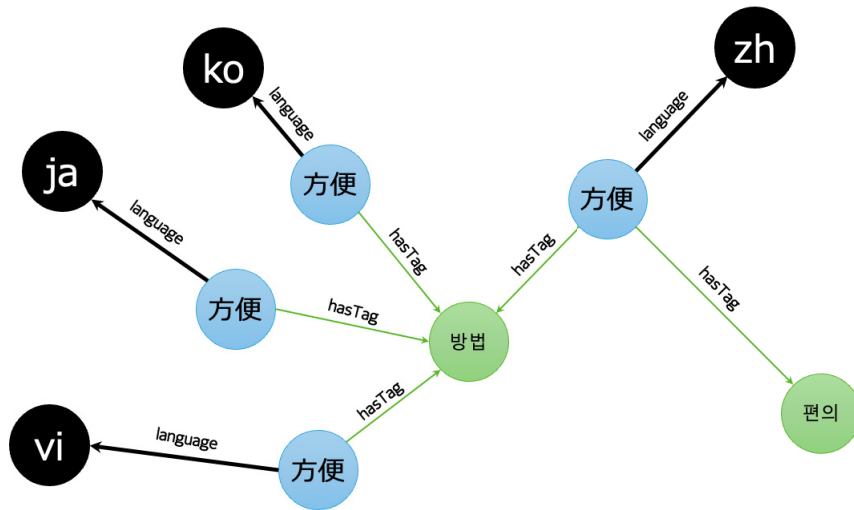
이러한 개선의 실제적 방안으로는 위 그림에서 개별 어휘[word], 즉 개체에 딸려 있는 속성으로 기술된 정보를 개체로 전환하는 방식을 생각해 볼 수 있다. 즉 ‘개체-속성’의 관계를 ‘개체-개체’의 관계로 전환하면서, [word]와 [lang; logogram; phonogram; smntc] 간의 관계성을 개체 대 개체로 기술하는 것이다.



〈그림 5〉 ‘개체-개체’의 관계로 기술한 개별 어휘의 속성

특히 의미 속성 태그는 아래 그림에서 예시한 方便의 경우와 같이 각 언어의 개별 어휘에 대해서 언어 횡단적인 관계성을 가시화할 수 있다.

42) 김현 · 임영상 · 김바로, 『디지털 인문학 입문』, 한국외국어대학교 지식출판원, 2016, 136~144쪽.



〈그림 6〉 의미 속성 태그를 통한 언어 횡단적 관계성의 가시화(方便)

동아시아 한자어DB 구축의 목적은 한자라는 공통의 문명 자산 위에 형성된 동아시아 한자문명을 규명하기 위한 기초 토대를 확보하는 것이다. 한편 연구 수행의 심화와 연구 성과의 집적이 데이터베이스의 보완 및 확장으로 연결되는 순환적 모델을 지향해야 할 것이다.

V. 결 론

이상 본 논문에서는 동아시아 4개 언어(한국어, 중국어, 일본어, 베트남어)의 한자어를 의미 중심으로 연결한 ‘동아시아 4개 언어 한자어 데이터베이스’의 구축 현황에 대해 소개하고, 향후 확장을 위한 과제에 대해서 열거하였다. ‘일-한-중-베트남 동형 2자 한자어 데이터베이스’의 사례를 들어 동아시아 한자어에 대한 언어 횡단적 연구에서 형태(한자 표기의 같고 다름) 중심으로 이루어진 접근법의 문제점으로 (1) 동일한 한자어가 개별 언어 안에서 갖는 어휘적 위상을 포착하기 어려운 점, (2) 의미적으로 높은 동가성을 갖는 대역어가 배제되는 점을 지적하였다. ‘동아시아 4개 언어 한자어 데이터베이스’는 의미 중심의 데이터베이스 구축을 통해 그러한 문제점을 극복하고 한자어의 언어 횡단적 연구를 보다 유의미하고 효율적인 방향으로 전개하기 위한 토대를 제공하는 것을 목표로 한다.

구체적인 데이터베이스 설계에서는 한국어를 기준언어로 삼아 일정한 사용빈도와 상용도가 인정되는 어휘 목록으로서 “한국어기초사전”에 수록된 한자어 21,763개를 이용하였다. 아울러 이들 한국어 어휘에 의미적으로 동가에 가까운 어휘로서 중국어 23,091개, 일본어 19,517개, 베트남어 7,761개의 대역어 선정의 과정에

대해 서술하였다. 각 언어의 어휘에 대해서 형태 정보와 의미 정보를 각각 입력하였고 그것을 기반으로 각 어휘의 의미와 형태가 대응하는 양상을 생키 다이어그램의 형식으로 가시화할 수 있었다. 다만 기준언어인 한국어 어휘에 대해서 한자어 대역어를 찾을 수 없는 것이 상당수 존재하는데, 이들 어휘에 대해서는 향후 비 한자어(=고유어)까지 시야에 넣은 종합적인 검토가 필요하다.

본 데이터베이스는 구축 과정에서 4개 언어의 대응 어휘들이 의미적으로 높은 동가성을 갖는다는 것을 검증하는 방식으로 설계하였다. 또한 높은 동가성에도 불구하고 서로 다른 언어의 어휘 간에 필연적으로 존재할 수밖에 없는 의미적 어긋남을 조감하기 위해 각 어휘의 의미 속성을 나타내는 키워드(태그)를 데이터화하였다. 또한 이러한 의미 속성 키워드를 통해 4언어 한자어의 의미적 대응을 언어 횡단적으로 조감하기 위해서는 관계형 데이터베이스(RDB)의 형식으로는 한계가 자명하다. 따라서 향후 이러한 개별 어휘들을 연결하는 관계성을 입체적으로 기술하기 위한 RDF 형식과 같은 개방적 목록의 구축 필요성을 제기하였다. 관계형 데이터베이스 안에서 어휘라는 개체의 속성으로 기술된 정보를 어휘와 동등한 개체의 층위로 전환시켜 기술하는 방식을 통해 언어 횡단적 어휘 연구의 기초 토대를 확보할 수 있을 것으로 기대한다.

〈참고문헌〉

- 김현 · 임영상 · 김바로, 『디지털 인문학 입문』, 한국외국어대학교 지식출판원, 2016.
- 中国社会科学院语言研究所词典编辑室, 『现代汉语词典』(第7版), 商务印书馆, 2016.
- 日本国語大辞典第二版編集委員会 · 小学館国語辞典編集部, 『日本国語大辞典』(第二版), 小学館, 2000~2002.
- 漢語大詞典編輯委員會 · 漢語大詞典編纂處 · 羅竹鳳, 『漢語大詞典』, 漢語大詞典出版社, 1986~1994.
- Hoàng Phê, Bùi Khắc Việt, and Viện ngôn ngữ học, *Từ điển tiếng Việt*, Đà Nẵng: Nhà xuất bản Đà Nẵng, 2015.
- 강현화 · 신자영 · 원미진, 「한국어 학습자 사전 표제어 선정을 위한 자료 구축 및 선정 방법에 관한 연구」, 『한국사전학』 16, 2010.
- 도원영, 「국어사전 표제어의 사용역 정보에 대한 고찰」, 『우리어문연구』 30, 2008.
- 신웅철, 「한국어와 일본어의 同綴한자어 비교 연구: 국립국어원 『한국어-일본어 학습사전』의 표제어와 대역어를 중심으로」, 『일어일문학연구』 115, 2020.
- 于劭贊 · 玉岡賀津雄 · ホアーン ティ ラン フォン, 「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, 2019.
- 朴善嫻 · 熊可欣 · 玉岡賀津雄, 「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, 2014.
- Sapir, Edward, *Language: An introduction to the study of speech*, New York: Harcourt, Brace and Company, 1921.

- 국립국어원, 『표준국어대사전』, 2008, <https://stdict.korean.go.kr/>(검색일 : 2021.08.24.).
- 국립국어원, 「국어의 로마자 표기법」(문화체육관광부 고시 제2014-42호, 2014.12.05.), 2014, https://kornorms.korean.go.kr/regltn/regltnView.do?regltn_code=0004®ltn_no=444#a444(검색일 : 2021.08.24.).
- 국립국어원, 「한국어기초사전」, 2016, <https://krdict.korean.go.kr/>(검색일 : 2021.08.24.).
- 국립국어원, 「국립국어원 한국어-베트남어 학습사전(Từ điển học tiếng Hàn-tiếng Việt của Viện Quốc ngữ Quốc gia)」, 2016, <https://krdict.korean.go.kr/vie/mainAction>(검색일 : 2021.08.24.).
- 국립국어원, 「국립국어원 한국어-일본어 학습사전(国立国語院 韓国語-日本語学習辞典)」, 2016, <https://krdict.korean.go.kr/jpn/>(검색일 : 2021.08.24.).
- 국립국어원, 「국립국어원 한국어-중국어 학습사전(国立国語院韓国語-漢語学習詞典)」, 2020, <https://krdict.korean.go.kr/chn/>(검색일 : 2021.08.24.).
- 국립국어원, 「《국립국어원 한국어-중국어 학습사전》 개통」(2020.05.22.), 2020, https://korean.go.kr/front/board/boardStandardView.do?board_id=6&mn_id=184&b_seq=754(검색일 : 2021.08.24.).
- 부산대학교 인공지능연구실·한국어정보처리연구실, 「한국어 어휘의미망(KorLex)」, 2009, <http://korlex.pusan.ac.kr/>(검색일 : 2021.08.24.).
- 차이나랩, 「에듀월드 표준중중한사전」, 2015, <https://zh.dict.naver.com/>(검색일 : 2021.08.24.).
- 차이나랩, 「에듀월드 표준한중중사전」, 2018, <https://zh.dict.naver.com/>(검색일 : 2021.08.24.).
- 한국검인정교과서협회, 「편수자료 Ⅱ, Ⅲ 수정 내용 반영본 안내」(2017.06.07.), 2017, https://www.ktbook.com/edata/PDS_AuthView.asp?num=327&pageno=1&startpage=1&keyword=%C6%ED%BC%F6%CO%DA%B7%E1(검색일 : 2021.08.24.).
- 株式会社ネットアドバンス, 「JapanKnowledge」, 2001, <https://japanknowledge.com/>(검색일 : 2021.08.24.).
- 国立国語研究所, 『言語政策に役立つ, コ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, 2011, <http://doi.org/10.15084/00003284>(검색일 : 2021.08.24.).
- 国立国語研究所コーパス開発センター(Center for Corpus Development, NINJAL), 「教科書コーパス語彙表(Ver. 1.0)」, 2011, <https://ccd.ninjal.ac.jp/bccwj/freq-list.html>(검색일 : 2021.08.24.).
- 名古屋大学大学院国際言語文化研究科・玉岡賀津雄研究室, 「日韓中越同形二字漢字語データベース」, 2015, <https://kanjigodb.herokuapp.com/>(검색일 : 2021.08.24.).
- 日本国外務省, 「へボン式ローマ字綴方表」, 2017, <https://www.ezairyu.mofa.go.jp/passport/hebon.html>(검색일 : 2021.08.24.).
- Princeton University, 「Princeton University WordNet」, 2011, <https://wordnet.princeton.edu/>(검색일 : 2021.08.24.).

* 이 논문은 2021년 11월 26일에 투고되어,
2021년 12월 22일에 심사위원을 확정하고,
2022년 1월 3일까지 심사하고,
2022년 1월 6일에 게재가 확정되었음.

■ Abstract ■

Establishment and Problems of CJKV Sinoxenic Words Database

Shin, Woongchul*

In this paper, we introduced the construction process and current status of the “CJKV Sinoxenic Words Database,” which connects Sinoxenic words in Korean, Chinese, Japanese, and Vietnamese, focusing on the meaning of vocabulary, and presented an attempt on the possibility of supplementation and expansion. This database was established for the purpose of analyzing and identifying the nature of East Asian civilization through common cultural assets called Chinese characters. In the existing database construction covering Sinoxenic words in various East Asian languages, the vocabulary of different languages was connected through notation, so it does not properly reflect the difference in register of Sinoxenic vocabulary in each language.

The design of this database focused on semantic correspondence, not on the notation or form of Sinoxenic words. First of all, the Korean entry words, which is the reference point for database construction, was selected from the “Basic Korean Dictionary”. Subsequently, not only did it respond meaningfully to the Korean words, but it also selected Chinese, Japanese, and Vietnamese equivalents that did not have a large gap in terms of vocabulary register.

The Sinoxenic words from four languages divided into four separate tables for each language, and the same ID was assigned to the corresponding Sinoxenic words. The vocabulary of each language table is connected to the structure of a Relational Database through an ID based on semantic response. In addition, in each vocabulary of the four East Asian languages, semantic information was cited and entered from a dictionary of a certain size or larger for each language. And based on that, tags representing the semantic attributes of vocabulary were included as separate fields.

It can be pointed out that the limitation of this database is that it is a closed structure that contains information on a limited layer of individual vocabulary in a plan manner. In the future, it will be possible to increase the usability and academic value of this database by creating a dataset that describes the relationship between individual vocabulary in the form of RDF to

* HK Research Professor, Center for the Study of Chinese Characters in Korea, Kyungsoong University

facilitate visualization of data and linkage with other databases. In addition, it is necessary to aim for a cyclical model in which the research results on East Asian civilization created through this database are reflected in the expansion and supplementation of the database.

[Keywords] Sinoxenic Words, Relational Database, Translingual Research on Vocabulary, Relationality, Semantic Attributes