정보처리를 위한 한일 양언어 데이터베이스 구축*

韓 有 錫**

目 次

1. 연구의 의의와 국내외 현황

데이터베이스로서의 한일 병렬 시소러스의 개발이라고 하는 본 연구가 갖는 의의를 살펴보면 대략 다음의 다섯 가지로 정리할 수 있다.

첫째, 한국어 시소러스 구축을 위한 기반 데이터로서의 역할이다. 한일 병렬 시소러스의 한 축을 이루는 한국어 분류어휘표는 한국어의 기초어휘(연세한국어사전의 표제어)를 분류한 하나의 한국어 시소러스이다. 일본의 분류어휘표의 체계에 맞추어 분류했기 때문에 분류체계에 약간의 공백은 있지만, 분류체계를 우리말의 어휘 체계에 맞도록 수정을 가함으로써 한국어 시소러스 개발의 시간과 비용을 단축할 수가 있다.

둘째, 양언어 어휘 비교는 물론, 언어학의 제분야(음운, 형태, 표기, 문법, 문체 등)의 비교연구가 용이하다는 점이다. 나고야대학의 다지마(田島) 연구실에서는 분류어휘표를 이용하여 유학생들이 동일 작품을 대상으로 일본어와 모국어간의 어휘 비교를 한 바 있었는데, 만일 한일 병렬시소러스와 같은 데이터가 각 언어에서도 구축이 된다면 자국어에 대한 의미코드 부여의 수고를 크게 절감할 수 있어 비교어휘연구가 매우 용이해 질 것이다.

셋째, 다언어 정보검색·자동번역 등의 정보처리 분야에 응용할 수 있다. 형태소분석기, 문장해석시스템 등 각종 정보처리의 도구가 효율적으로 컴퓨터상에서 자연언어를 처리하기 위해서는 한국어 시소러스, 전문용어 시소러스, 다언어 시소러스와 같은 각종 시소러스

* 본 과제는 정보통신부의 IT학술연구지원사업(정보통신연구진흥원)으로 수행한 연구결과물입니다.
** 동신대학교 부교수 일본어학

가 정교하면서 동시에 방대하게 구축이 되어져야 한다.

넷째, 대역사전의 대응어 선정 및 국어사전의 의미 기술의 정교화에 기여한다. 과거에 국어사전에서는 의미 정의를 다람쥐 쳇바퀴 도는 식으로 하거나, 유의어간의 의미 정의가 같거나 해서 사용자에게 불편을 주었다. 시소러스의 개발로 유의어간의 관계가 명확해지면 국어사전의 의미도 좀더 정확하게 기술할 수 있게 될 것이다.

다섯째, 일본어 교육, 한국어 교육에 응용할 수 있다. 병렬시소러스는 일본어교육, 한국어교육에 있어서의 학습평가, 교과서 개발, 단계별 학습계획 등에 응용할 수 있다.

다음으로 현재 개발, 공개된 시소러스의 현황을 일본과 한국 순서로 정리해 본다. 먼저 일본에서 전문용어를 제외한 일반어 시소러스로서는, 林大의 원판 분류어휘표(1964, 36,700어), 大野進·浜西正人의 유어신사전(1981년, 50,000어), EDR의 EDR 개념사전(1995년, 41만어), NTT의 일본어 어휘대계(1997년, 40만어), 荻野綱男의 명사 시소러스(개발중, 50,000어), 柴田武·山田進의 유어대사전(2002년, 79,000어), 山口翼의 일본어 대시소러스(2003년, 개별어 20여만어), 국립국어연구소의 증보개정판 분류어휘표(2003년 96,000어), 이밖에 정보 검색 또는 전자사전용으로 개발된 언어공학연구소의 디지털 類語辭典(22만어), 學習研究社의 學研시소러스(85,000어)가 있다.

각 시소러스의 특징은 다음과 같다. 먼저 원판과 증보개정판의 분류어휘표는 연구용에 적합한 시소러스이고, 유어신사전은 학습용에 잘 맞는 시소러스이며, 유어대사전은 유어신사전을 확대, 발전시킨 것이라고 할 수 있다. 일본어 대시소러스는 뚜렷한 특징은 없지만, 분류어수가 많다는 점에서 학습용이나 정보검색용에 적합한 것처럼 보인다. EDR 개념사전이나 NTT의 일본어 어휘대계는 일영 기계번역을 위한 시소러스이며, 기계번역을 위한 것은 아니지만, 荻野의 명사 시소러스, 디지털 유어사전, 學研시소러스도 정보 검색이나, 전자사전용으로 개발되었다는 점에서는 개념사전이나 어휘대계와 같은 기계용 시소러스라고 할 수 있다.

한편, 한국의 시소러스 개발 현황은 일본과는 큰 차이를 보이고 있다. 전문용어 시소러스는 최근 정보처리기술의 발달과 함께 연구가 활발한 편이지만, 일반어 시소러스는 표준국어대사전의 간행과 함께 연구된 임홍빈의 시소러스, 로제의 시소러스를 참고로 하여 만든 지식공학의 K-시소러스밖에 없으며, 이들 시소러스도 분류어수가 적고, 분류가 정교치 않다는 단점이 있다.

2. 한국어 시소러스의 보완

본 연구에서 한국어 시소러스를 구축함에 있어 분류어휘표를 기저 시소러스로 삼는 이유는 다음과 같다.

먼저 분류어휘표는 품사 분류가 1차 분류로 되어 있고, 각 품사 간 의미가 비슷한 것들은 동일한 코드 체계를 갖게 하여 상호 검색을 가능케 하였으며, 세부 항목내의 분류어는 의미의 유사성, 친소성을 고려하여 정밀하게 분류하여 배치한 세계 유일의 시소러스이라는 점이다.

또한 원판 분류어휘표는 5만부가 판매되고, 수백편의 연구 이용 논문이 있는 등 일본의 국어학 분야에서 최고의 문헌자료이라는 것은 자타가 공인하는 사실이라는 점이다. 단, 용례가 없다는 점에서 학습용으로는, 상하 관계가 세밀하지 않다는 점에서 정보처리용으로는 각각 약간 부적합하다는 평을 받고 있다.

그러나 이와 같은 분류어휘표의 단점은 문제가 되지 않는다. 어떠한 시소러스이건 모든 용도에 적합한 완벽한 범용 시소러스를 만든다고 하는 것은 불가능하기 때문이다. 분류어휘표는 연구용이라고 하는 특징 하나만으로도 충분히 개발 목적을 달성할 수 있지만, 용례를 추가하고, 상하관계에 대한 표시를 좀더 명확히 한다면 학습용, 정보처리용으로서도 손색이 없는 시소러스가 될 것이다.[1)]

한국어 시소러스에서는 분류어휘표를 기저시소러스로 하면서 분류어휘표의 장점은 그대로 살리고, 단점은 보완하기 위해서 몇 가지 방안을 도입하는데 이에 대해서 논하면 다음과 같다.

2-1. 분류어의 증대와 분류체계의 세분화

http://www.lexfn.com에 들어가면 영어에서의 단어와 단어 간의 여러 관계정보를 얻을 수 있다. 가령 train이라는 단어를 검색어로 입력하면 train의 각 의미항에 대한 동의어, 유의어는 물론, 상위어, 하위어, 전문어, 일반어, 반의어, 방언, 속어뿐만 아니라 연상어에 이르기까지 수백 개의 정보가 총 망라되어 검출된다. 한국어나 일본어에서는 이와 같은 정보검색이 불가능한데, 이는 이와 같은 모든 관련어(유의어, 방언, 연상어 등)에 대한 의미연구의 기반구축이 불충분하기 때문이다.

가령, '기차'에 대해서, 상위어는 수송기관, 운송수단, 교통시설 등이 되겠고, 하위어는 일등칸, 기관실, 동의어는 열차, 기관차, 유의어는 전차, 전동차, 화차, 탄광차, 객차, 식당차 등이 되겠고, 연상어는 칙칙폭폭, 호남선, 완행열차, 새마을호, 철도, 기차역, 고속철도 등을 생각할 수 있을 것이다. 또한 '어머니'에 대해서 '엄마', '모친', '자친', '자당', '어멈', '모', '엄니', '에미', '선대부인', '어마마마' 등은 동의어이고, '아버지', '아빠' 등은 반의어, '어마이', '어머이', '어무이', '오마니', '어메' 등은 '어머니'의 방언이다. 그런데 이러한 것들은 지금 즉흥적으로 생각해낸 것들이고, 우리나라에서는 아직 이와 같은 관련어를 한번에 검

1) 이러한 일은 지면 제약상 책자형으로는 무리이지만 전자 출판이라면 얼마든지 가능하다.

색하여 나타낼 수 있는 자료가 구축되어 있지 않다. 지금으로서는 방언사전, 속어사전, 동의어사전 등을 일일이 따로따로 검색해야만 되는데, 이러한 사전들은 내용의 정교성은 차치하고라도 우선 제공할 수 있는 정보량이 충분하지 않다. 보다 정밀한 정보를 얻기 위해서는 관련어에 대한 정보량이 대용량이지 않으면 안 된다. 참고로 1999년에 작성된 상기 Lexical FreeNet에서는 800여만 개의 관련어 정보가 데이터베이스로 구축되어 있다.

한국어에서도 이와 같은 정보검색이 가능하기 위해서는 먼저 대규모의 대용량시소러스를 구축할 필요가 있는데, 현재 한국어에서 공개된 시소러스는 3만~10만어 정도의 규모이다.(옥철형의 UOU온톨로지, 최기선의 코어넷, 지식공학의 K시소러스) 한국어 시소러스 작성의 모델로 되어 있는 일본의 분류어휘표 또한 97,000어에 불과하다. 시소러스에서는 유의어, 동의어, 상위어, 하위어 반의어 정보를 얻을 수 있으므로, 먼저 이러한 시소러스의 규모를 50만~100만어 정도로 확대할 필요가 있다. 최근 일본에서는 20만어, 영어권에서는 50만어 규모의 시소러스가 공개되고 있는데, 한국어에서도 국어정보 기반구축연구사업으로서의 시소러스의 확대사업을 적극 추진해야 할 것이다.

아울러 방언이나, 속어, 연상어, 고어, 북한어, 전문어에 대해서도 일부 연구 자료나 사전 등이 나와 있지만, 금후 본격적이고 대대적인 조사를 통해서 자료를 보강하는 작업이 이루어져야 한다.

또한 시소러스를 확대하기 전에 시소러스를 어떠한 용도로 쓸 것인지에 대한 분류의 목적을 분명히 하고, 분류체계를 정비할 필요가 있다. 일본의 분류어휘표에서도 36,000어의 원판을 97,000어의 증보판으로 확대했을 때 분류어휘표의 분류체계를 다소 수정하였다. 현재 본 연구에서 작성중인 시소러스는 97,000어의 분류체계인데, 이를 50~100만어로 확대하려면, 분류체계의 기본틀은 바꾸지 않는다고 해도, 분류체계의 하위부분을 좀더 세분화할 필요성은 있다. 분류어휘표에서는 분류어간의 상하관계가 명시되어 있지 않은데, 분류체계의 세분화는 이러한 분류어휘표의 단점을 보완하는 방식으로 작업이 진행되어야 한다.

예를 들어 1.4650의 교통수단(육상)에서 1.4650-01의 세부항에는 '교통 기관'이 나오고, 1.4650-13의 세부항에는 '버스', '관광버스', '고속버스', '직행버스' 등이 나오는데, 이를 '교통 기관'은 1.4650의 상위항에, ' 관광버스', '고속버스' 등은 1.4650-13-01처럼 1.4650-13의 하위항을 새로이 신설하여 배치를 하는 방법이다.

이와 같이 금후 시소러스를 대용량으로 확대하는 과정에서는 단어와 단어간의 개념관계를 어떻게 하면 효과적으로 구축할 수 있는지에 대한 방안이 함께 고려되어져야 할 것이다.

2-2. 빈도수 정보

한정어 정보가 책자형 출판을 위한 것이라면, 빈도수 정보는 전자출판을 위한 정보라고 할 수 있다.

여기서는 먼저 빈도수 정보에 대해서 서술을 하겠는데, 빈도수 정보는 한국어 시소러스의 분류어가 모두 연세한국어사전의 표제어이기 때문에 연세한국어사전 구축 시 보유하고 있는 빈도수 정보를 붙이면 된다. 하지만 다의어에 따른 중복 분류어를 처리하는 문제는 그리 간단하지만은 않다. 연세한국어사전의 빈도수 정보가 표제어에 대한 빈도수를 나타낸 것이어서 표제어가 다의어일 경우 각 의미항에 대한 빈도수는 확인할 수가 없기 때문이다. 한편, 한국어 시소러스에서는 다의어의 각 의미항을 하나하나 분류어로 삼았기 때문에 단의어와 다의어의 구별 없이 산출한 빈도수를 그대로 일률적으로 적용할 수는 없다. 따라서 빈도수 정보를 효율적으로 이용하기 위해서는 다의어에 대한 빈도수 연구가 선행되어야 할 것이다. 그러나, 다의어에 대한 빈도수 연구가 이루어졌다고 해도, 그 연구가 연세한국어사전의 다의어의 의미항에 기초한 조사연구가 아니라면 그 조사연구를 그대로 적용하는 것은 곤란하다. 다의어는 사전에 따라 그 기술 방법이 제각기 다르기 때문이다.

가령, '활활'이라는 단어가 A사전에 입각하여 작성한 다의어의 실제 빈도수조사를 적용했을 때 다음과 같이 나타났다면, 그리고 연세한국어사전에서는 '활활'이라는 단어가 17의 빈도수를 갖는다면 연세한국어사전의 '활활'이라는 다의어의 가상 빈도수는 다음과 같은 방법으로 산정이 가능할 것이다.

A사전의 '활활'의 의미항	다의어의 실제 빈도수 조사	연세한국어사전의 '활활'의 의미항	연세한국어사전의 예상 빈도수
의미항 ①	18	의미항 ①	12(=18/25×17)
의미항 ②	4	의미항 ②	3(=4/25×17)
의미항 ③	2	의미항 ③	2(=3/25×17)
의미항 ④	1		
합계	25	합계	17

'활활'의 A사전의 의미항 ①이 연세한국어사전의 의미항 ①과 서로 같다고 할 때, A사전의 '활활'의 전체 빈도수와 의미항 ①의 빈도수가 각각 25와 18이고, 연세한국어사전의 '활활'의 전체 빈도수는 17이라면, 연세한국어사전 '활활'의 의미항 ①의 빈도수를 계산하는 방법은 18/25×17이 되어 12라고 하는 빈도수를 예상할 수 있다는 것이다.

이와 같은 방법으로 한국어 시소러스의 모든 분류어에 빈도수를 붙이게 되면 의미와 빈도수에 관련된 각종 연구가 활발히 이루어질 것이다. 여기서 각종 연구란 어휘, 의미, 음운, 문법, 사전학, 타언어와의 비교, 대조연구와 같은 것들을 말한다.

2-3. 한정어 정보

한정어 정보는 책자형으로 출판을 했을 때[2] 검색자가 쉽게 검색어의 의미를 구별할 수 있도록 분류어 뒤에 기입하는 의미정보를 말한다. 한국어 시소러스의 기저 시소러스인 일본의 분류어휘표에서도 한정어가 가끔 보이는데, 이는 다의어에 한해서 혼동을 피하기 위해 최소한으로 기입한 것으로 보여진다. 그러나 한국어 시소러스에서는 다의어뿐만 아니라 단의어에 대해서도 그 뜻이 쉽게 파악되기 어려울 것으로 생각되는 것들에는 모두 한정어를 붙이도록 했다. 일본어처럼 한자를 쓰지 않기 때문에 한국어에서는 표제어만으로 의미판정이 어려운 동음이의어들이 많기 때문이다. 또한 동음이의어와 같은 문제는 아니지만 단어의 뜻이 어려워 사전을 다시 찾아보아야 하는 단어들에도 한정어를 붙였다.

한정어의 부가 방법은 가급적 연세한국어사전의 용례를 그대로 이용하도록 하지만, 경우에 따라서는 작례를 부가하기도 한다. 또한 한정어로 쓰이는 용례는 의미가 통하는 범위에서 그 길이를 최소한으로 줄이도록 했다.

<2.1560-23의 한정어 부가 예>
뜨다1[장판이~] 들뜨다[도배지가~] 벌다2[밤송이가~] 벌어지다1[틈/격차가~]
벌리다1[다리를~] 두다1[거리를~]

3. 분류의 방법

분류의 기본 원칙은 다음과 같다.

(1) 분류의 대상은 1차적으로 연세한국어사전의 표제어 대략 52,000 어와 부표제어 약 5,000어 중 어미, 조사, 준꼴, 조음소, 보조동사, 보조형용사를 제외한 모든 품사이다. 병렬 시소러스에서는 최종적으로는 분류어휘표의 분류어수 79,000어(총어수 96,000어)와 비슷한 규모의 어구를 분류할 것을 목표로 한다.[3]

(2) 다의어는 가능한 한 의미 항목을 전부 분류하나, 사전의 기술에 문제가 있거나, 분류에 문제가 있을 경우에는 분류에서 제외한다. 또한, '부사적으로 쓰이어'와 같은 품사적 파생 항목은 독립 파생어로 인정하여 이를 해당 품사류에 분류한다.

---

2) 전자출판이나 온라인상으로 제공되는 데이터에는 의미정의 및 용례를 함께 제공하기 때문에 따로 한정어를 부가할 필요는 없다.

3) 2차 분류의 증보어 선정 시에는 연세의 표제어의 불균형을 해소할 필요가 있다. 예를 들어 연세한국어사전에서는 [교위]라는 단어는 있으나 [교육위원회]라는 단어는 없으며, [그끄러께]는 단어는 있는데 [그러께]라는 단어는 없는 등, 표제어의 불균형이 많은데 이들은 대개 동일 유의어항목에 분류되므로 분류어휘표에서는 이들 분류어에 대한 결락을 쉽게 찾아낼 수 있다.

(3) 분류는 병렬 시소러스의 작성이라는 목적의 실현을 위해 증보개정판 분류어휘표의 분류 체계와 분류 방법을 그대로 따른다.

분류를 함에 있어서 방법상의 문제를 열거하자면 너무나 많아 한 곳에 다 열거할 수가 없다. 그래서 여기서는 일본어와 한국어가 어형이 같지만 의미나 품사가 달라서 분류에 주의를 요하는 것들만 정리해 본다. 당연히 고유어는 해당이 안 되고, 형태가 같은 한자어나 외래어가 여기에 해당된다.

3-1. 의미는 같으나 품사가 다른 것

(1) 품사가 다르지만 대응어에 따라 분류

다음 어구는 대응 일본어와 품사는 다르지만(한국어는 명사, 일본어는 형용동사 또는 부사), 의미가 같으므로 대응 일본어에 따라 상의 류에 분류한 것들이다.

불연, 천정부지(~로), 즉시, 오랫동안, 시기상조, 적시, 정기2, 이원적1, 개별, 개별적1, 공동체적2, 수직적1, 합법적1 ,수동적1 피동적1, 공허, 필연적1, 자발적, 보편적, 개성적, 본능적, 최고2, 결국, 정기적, 정기, 졸속, 즉시, 예상 외로/밖으로, 논외, 고액, 고율, 고속, 최대한, 최소한, 따름, 십중팔구, 단칸, 불량, 과민, 무감각, 둔감, 무상, 우울, 불편, 침울, 겸양, 겸허, 겸손, 순정, 불가사의, 변칙, 중용, 가공2, 태연, 안이, 경건, 교만, 저자세, 고자세, 악취미, 충실, 불성실, 성실, 열심, 공평, 엄정, 평등, 무책임, 비겁, 비굴, 만성, 만성적, 언감생심, 무분별, 부지불식간에, 무의식, 불가해, 불확실, 불문가지(불문가지), 대중이 없다, 신예, 불특정 무방비, 미시적, 거시적, 고명, 공통, 추신, 핑계 김에, ~기 일쑤이다, 대1, 이색, 이색적, 지리멸렬, 주름투성이 돌투성이 먼지투성이 피투성이, -씩2

---

(2) 동형어이지만 대응어가 아닌 것

가령 [무책임]은 [無責任]과 동형어이지만 한국어는 명사, 일본어는 형용동사이다. 즉 일본어의 [無責任]은 한국어의 [무책임](명사)보다는 [무책임하다(형용사)]에 대응한다고 할 수 있다. 따라서 [무책임]을 상의 류에 분류해서는 안 되고, 체언류에 분류해야 한다. [유익](有益)이나 [유덕](有德)도 이와 같은 예이다.

한편 [급박하다], [오목하다], [우월하다]는 각각 대응하는 일본어가 [急迫する], [へこむ/くぼむ], [優越する]인데, 한국어는 형용사이고 일본어는 동사이므로 대응어에 관계없이 상의 류에 분류한다.

3-2. 기본의가 다른 것

(1)의미가 미묘하게 다른 것

[전과] 초등학교 전 과목을 풀이해 놓은 참고서	[全科] すべての教科. 全科目.
[정사] 정통적인 역사 체계에 의하여 서술된 역사. 왕조를 중심으로 하여 편찬된 역사.	[正史] 国家によって正式に編纂された歴史書
[타자] 타인, 남	[他者] 前者, 後者
[사무관] 행정 업무를 담당하는 5급의 공무원	[事務官] 国の行政機関で, 一般事務を担当する公務員.技官·教官などに対していう. 文部事務官·通商産業事務官など.
[용달] ① 물건이나 짐 따위를 배달하는 것, 또는 그 일. ② '용달차'의 준말.	[用達] (1)用事をすませること. (2)役所·会社などに出入りして品物を納めること. またそれをする商人.御用達. (3)大小便をすること.
[필부] 보통 사람	[匹夫] 身分の低い男.また,道理をわきまえない卑しい男.

한국어에서 [전과]는 <전 과목 참고서>의 뜻이지만, 일본어에서는 <전과목>이라는 뜻이다. 한국어에서 [정사]는 <역사>의 뜻이지만, 일본어에서의 [正史]는 <역사서>의 뜻이다. 이와 같이 양언어 동일 형태의 한자어라고 해도 그 쓰임이 미묘하게 달라 일본어와 동일한 분류항에 분류해서는 안 되는 것들이 있다. 다음의 예들도 그와 같은 예이다.

[意匠]과 [의장], [전과]와 [全科], [분과]와 [分科], [야사]와 [野史], [답례]와 [答礼], [타이프]와 [タイプ], [別途]와 [별도], [촌민]과 [村民], [작법]과 [作法], [수포1]과 [水泡], [착의]와 [着衣]

(2)의미가 완전히 다른 것

[미혹] 무엇에 홀려 정신을 못 차리는 것	[迷惑] 人のしたことで不快になったり困ったりする·こと (さま).
[유사4] ① 옛날부터 전하여 내려오는 이야기를 모은 책. ② 죽은 사람에 관한 널리 알려지지 않은 이야기 모음. 일화집.	[遺事] (1)昔から伝わって残っている事柄. (2)故人のし残した事柄. (3)計画などでもれ残された事柄.

한국어의 [책(冊)]은 일본어에서는 [本]이라고 하고, [공부]는 [勉强]이라고 한다. 이와 같이 쉬운 단어들은 분류에 있어 오류를 범하지 않지만, [미혹]이나 [유사4]와 같은 어구에서는 자칫 동형어를 대응어로 삼아 잘못된 분류를 행할 수 있다. [수배]와 [手配], [지행]과 [知行], [치기]와 [稚氣], [증]과 [証]과 같은 동형어도 의미가 다르므로 분류에 주의해야 한다.

3-3. 파생의가 다른 것

[풍물]에는 <지방의 경치와 생활하는 모습>이라는 뜻 외에 <악기>라는 뜻이 있지만, 일본어에서는 <악기>라는 뜻은 없고, 대신 <어떤 지역이나 계절의 특징을 나타내는 사물>이라는 뜻이 있다. 한국어를 분류할 때는 이와 같이 파생의가 다른 점을 고려하여 분류해야 한다. [도발]이나 [팔방미인]의 파생의미도 마찬가지이다.

[풍물] ①지방의 경치와 생활하는 모습 ② 악기	[風物] (1)目にはいるながめ風景,またそれを形作っている個々の景物. (2)ある土地や季節の特徴を表している事物.
[도발] ① 일부러 남의 화를 돋우거나 싸움을 거는 짓. ② 전쟁이나 전쟁의 위험이 있는 사태나 사건 따위를 일으키는 것.	[挑発] (1)相手を刺激して向こうから事を起こすようにしむけること. (2)刺激をえて色情をそそりたてること.
[팔방미인] ① 어느 모로 보나 아름다운 미인. ② 여러 방면에 재주가 있는 사람	[八方美人] (1)どこから見ても欠点のない美人. (2)だれに対しても如才なく振る舞う人.

3-4. 일본어의 의미가 확대된 것

[복상] 상복을 입는 것	[服喪] 喪に服すること. 近親者の死後, 一定期間外出などを控え身を慎むこと.
[의형제] 남남끼리 의로 맺은 형제 관계	[義兄弟] (1)互いに交わした約束で兄弟の交わりをする人. (2)妻や夫の兄弟.義兄や義弟など. 義理の兄弟.
[불복] 복종하거나 동의하지 아니하는 것.	[不服] (1)納得できないこと 不満に思うこと.また,そのさま. (2)服従しないこと.

[조합] (약재나 물감 같은 원료를) 일정한 비율로 한데 섞는 것. 일정한 분량대로 두루 배합하는 것.	[調合] (1)幾種類かの薬品をきめられた分量でまぜ合わせること. (2)香料·調味料などをまぜて,一定の香りや味を作り出すこと.
[지척] 아주 가까운 거리.	[咫尺] (1)距離がきわめて近いこと (2)貴人に接近すること.
[산발] (여자의) 풀어헤친 머리.	[散髪] (1)髪を刈り,形を整えること. 調髪. (2)元結(もとゆい)を結わずに下げた,乱れた髪. ちらし髪. (3)「斬髪」に同じ. (4)「散切り」に同じ.
[귀천] (어떤 일, 지위 따위의) 귀한 것과 천 한 것.	[貴賤] 身分の高い人と低い人. 貴いことと卑しいこと.

일본어의 [服喪]에는 <상복을 입는 것> 외에 <상중에 근신하는 것>의 의미도 있다. 또한 [義兄弟]에도 <의로 맺은 형제 관계> 외에 <아내의 형제 또는 남편의 형제>라는 뜻도 있다. 이와 같이 일본어와 한국어의 의미가 서로 동일하면서 일본어에서 의미가 확대된 것으로 표로 나타낸 것 외에도 다음과 같은 것들이 있다.

[무심]과 [無心], [촌1]과 [村], [의장2]와 [意匠], [국가2]와 [国歌], [자신]과 [自身], [순례]와 [巡礼], [존비]와 [尊卑], [참예]와 [参与], [배달]과 [配達], [육사]와 [陸士], [징역]과 [懲役], [전임]과 [前任], [탈모]와 [脱帽], [철저]와 [徹底], [만족]과 [満足]

동형 한자어는 아니지만, 다음과 같은 어구들도 기본적인 의미는 동일하면서 일본어 쪽에서는 한국어에 없는 파생의를 갖는 것들이라고 볼 수 있다.

[개]와 [いぬ], [봄]과 [春], [일출]과 [日の出][4]

3-5. 한국어의 의미가 확대된 것

반대로 적극적으로 조사한 것은 아니어서 그 용례는 적으나, 동형 한자어(또는 외래어)

5) 일본어의 [いぬ]에는 <스파이>라는, [春]에는 <색정>이라는, [日の出]에는 <기세가 등등함>이라는 각각 한국어에 없는 파생적인 의미가 있다.

중에 한국어 쪽이 의미가 확대된 것으로 [토픽], [대파]와 같은 어구가 있다.

[토픽] 사람들의 흥미를 끄는 화제, 또는 그 화제를 다룬 신문이나 방송의 기사	[トピック] 話題. 論題.
[대파] ① 크게 부서지는 것. ② 크게 이기는 것.	[大破] 修理できないほど大きく破損すること.

4. 연세한국어사전의 문제

한국어의 시소러스는 연세한국어사전의 표제어를 대상으로 삼아 분류를 하고 있으며, 이때 연세한국어사전의 의미정의와 용례를 참조로 하고 있다. 그런데 이 연세한국어사전을 이용하다 보면 다음과 같은 문제점들이 발견된다.

4-1. 잘못된 뜻풀이

연세한국어사전의 표제어를 1차 분류어의 대상으로 삼음에 따라 분류 과정에서 참고로 하는 의미 기술(=뜻풀이, 의미 정의)이나 용례도 연세한국어사전의 그것들을 참고로 하는 때가 많다. 그렇지만 연세한국어사전의 의미 기술에는 더러 잘못된 곳들이 있어서 조금이라도 의심이 갈 때는 표준국어대사전이나 일본의 국어사전 등을 참고로 하지 않으면 안 된다.

가령 [차입2]의 의미는 연세한국어사전에서 다음과 같이 되어 있다.

> ▣ 차입2 「구치소나 교도소에 갇혀 있는 사람에게 바깥에서 옷이나 음식, 돈 따위를 들여보내는 것. 옥바라지 ¶ 석방이 안 되면 차입을 넣으려고 먹을 것을 좀 싸 왔어요.」

우리말에는 [수혜]의 의미가 혜택을 받는 것이라는 의미가 있는데 실제 문장에서는 '수혜를 받다'와 같이 의미를 중복적으로 사용한다. 위의 예의 '차입을 넣다'도 그와 같은 예로서 이럴 경우 [차입]의 의미는 연세한국어사전의 <옷, 음식, 돈 따위를 들여보내는 것> 외에 <또는 그 옷이나 음식, 돈>이라는 의미를 추가해야 할 것이다.

[먹성]의 의미항 ②는 <음식을 먹는 양>으로 되어 있는데, 이는 <음식을 많이 먹는 성향>으로 바뀌어 기술되어야 한다.

[대타]는 <~는 사람>으로 되어 있으나 분류어휘표의 [代打]처럼 <~는 것>이라는 뜻도 포함되어야 한다.

[정통]의 의미항 ①의 정의는 표준국어대사전, 大辭林처럼 <올바른 계통>으로 기술해야 한다.

[잠자리2]에는 표준국어대사전처럼 <이부자리>의 뜻도 기술되어야 한다.

[코맹맹이]는 다음과 같이 연세한국어사전과 표준국어대사전의 의미 기술에 차이가 있으나, 이는 표준국어대사전처럼 해야 한다.

▣ 코맹맹이 「코가 막혀서 제대로 말소리를 내지 못하는 사람, 또는 그런 소리.」『연세』
▣ 코맹맹이 「코가 막혀서 소리를 제대로 내지 못하는 상태. 또는 그런 사람.」 『표준』

[행장1]은 <여행할 때 쓰이는 여러 가지 물건>으로 되어 있는데, 이는 표준국어대사전에서처럼 <여행할 때 쓰는 물건과 차림>으로 하는 것이 옳다.

[종사1]이 연세한국어사전에서는 <나라에 관한 중요한 일>이라고 되어 있는데, 표준국어대사전에서는 <나라를 이르는 말>이라고 의미 기술이 다르다. 표준국어대사전에 따라 1.2530-01에 분류해야 한다.

[궐기]는 <여러 사람이 어떤 목적을 위해 결심하고 함께 행동으로 의사를 나타내는 것>이라는 기술보다는 大辭林의 [蹶起]처럼 <勢いよく立ち上がること>(힘차게 일어나는 것)으로 정의되어야 한다.

[공2]는 ① 훌륭한 일을 이룩하는 데에 들인 노력과 정성. ② 무엇을 하는 데 들인 힘이나 노력처럼 기술되어 있는데, 이와 같이 <~한 노력, 정성>보다는 大辭林의 동형어에서처럼 <성취한 일. 공적>이라고 해야 한다.

[천신2]의 의미 ②는 표준국어대사전처럼 방언으로 처리해야 하는 다른 뜻의 단어이다.

4-2. 의미용법의 쓰임이 불확실하거나 너무 세밀히 기술된 의미항

연세한국어사전의 [발고(發告)]의 의미 ②는 표준국어대사전에는 안 나오는 불확실한 의미이다. 일단 표준국어대사전에 기술이 되어 있지 않은 것은 일반적인 의미가 아니라고 판단하고 분류를 하지 않는다.

[종두2]의 의미 ②는 <종의 윗부분>인데, 이는 너무 전문적인 세밀한 의미이므로 분류하지 않는다. 참고로 의미 ①은 <교당에서 종을 치는 사람>이다.

4-3. 연세한국어사전에 기술이 없더라도 그런 용법이 실제 가능할 경우

예를 들어 [中堅]은 분류어휘표에서 1.2440-15와 1.2450-08에 분류되어 있다. 이와 달리 연세한국어사전에는 야구 용어로서의 [중견]의 의미 기술은 없다. 그러나 한국어에서 [중견]은 실제 야구 용어로 쓰이는 표현이므로 1.2440 외에 1.2450에도 분류를 한다.

또 다른 예로서 연세한국어사전에서 [상주3]의 뜻은 하나이다. 그러나 분류어휘표에서는 <불교>의 용어로도 분류되어 있다. 이를 표준국어대사전을 통해 확인해 보니 역시 <불교>의 용어로서의 기술이 있다. 따라서 [상주3]은 연세한국어사전의 기술 누락으로 보고 의미에 따라 2가지로 분류한다.

연세한국어사전, 표준국어대사전에는 [신문지]가 종이의 뜻으로만 나와 있는데, 분류어휘표에서 1.3160(문헌, 도서)에도 분류한 이유는 '신문지를 구독하다'처럼 도서의 의미로도 쓰일 수 있기 때문일 것이다. 따라서 분류어휘표의 분류 체계에 맞춰 1.3160-24에도 분류한다.

4-4. 의미정의와 예문의 불일치

연세한국어사전에는 의미 정의와 예문이 일치하지 않는 때가 종종 있다. 예를 들어 [출세 가도]를 사전의 의미 정의대로 분류한다면 <변화>의 항에 분류할 수밖에 없지만, 실제 용례를 보면 [출세 가도]는 [~를 달리다]와 같은 문형 속에서 출현한다.

> ▣ 출세 가도(出世街道) 「사회적으로 높이 평가되거나 유명해짐 그는 돈 많은 아버지의 후광에 힘입어 거침없는 출세 가도를 달렸다.」

따라서 <변화>의 항에 [출세 가도]를 분류해 버리면 위와 같은 문형이 실현될 수가 없고, 분류는 오류를 범하게 된다. 이와 같이 사전의 표제어를 분류할 때 1차적으로는 사전에서의 의미 정의를 참고로 하지만, 의미 정의에 문제가 있는지를 항시 용례를 통해 확인을 하여야 한다. 만일 의미 정의가 잘못되었을 때는 분류를 그 잘못된 의미 정의에 의존해서는 안 되고, 용례나 다른 사전의 의미 정의를 참고로 분류를 해야 한다.

4-5. 불필요한 파생의

[글줄]의 의미는 <①줄이어 씌어 있는 글 ②(비꼬는 말로) 그리 깊지 않은 학문 글줄이나 읽어 소위 학식이 있다는 젊은이들이...>로 되어 있다. 연세한국어사전은 실제 문헌에 나타난 용례에 의거하여 의미를 정의하고 있지만, [글줄]의 ②의 정의는 그 밑에 제시한 용례에 맞지 않는다. ②에 대해서는 표준국어대사전에서처럼 <약간의 글>로 정의하는 것이 옳을 것이다. 만일 ②의 의미를 <학문>으로 파악하여 분류하면 이와 같은 연세한국

어사전의 잘못된 정의를 그대로 분류에 반영하는 꼴이 될 것이다.

[결합]의 ②, [가늠]의 ①, [마력]의 ②, [이동1]의 ②, [헛손질] ①, [흥정]의 ③, [공개]의 ②도 삭제되어야 할 불필요한 파생의이다.

4-6. 단순한 오류

연세한국어사전에는 같은 단어가 각기 다른 해설로 두 번 나오거나(팔자소관), 한자가 잘못 되어 있거나(비준, 도량2, 중과, 교환하다), 품사가 잘못 표시된 것(유물2)이 있다. 이와 같은 연세한국어사전의 오류들은 분류작업을 할 때 적절히 수정을 하면서 분류해야 한다.

5. 분류어휘표의 문제

기저 시소러스인 분류어휘표에도 다음과 같은 문제점을 안고 있다.

5-1. 분류어휘표에서의 분류의 오류

2003년에 국립국어연구소의 소내 배포용으로 출간된 분류어휘표(연구자료집 14)에서는 다음과 같이 분류가 잘못된 것으로 보여지는 것들이 있다.

예를 들어 [布陣]은 1.3560-12에 분류되어 있는데, 이는 1.1513-09에 분류되어야 한다. 또한 1.1100-03의 [品種]은 1.5300-10(生物)에도 분류해야 한다. 일본어에도 [担任]에는 <교사>라는 뜻이 있는데, 1.3400-09에만 분류되어 있고, 1.2410-03에는 분류되어 있지 않다. [同一性]은 [아이덴티티] 옆에도 분류해야 되나 분류어휘표에는 그렇게 되어 있지 않다.

이밖에도 [法務], [与件], [ずつ], [闘士], [代金], [法律上], [海辺], [一石二鳥], [年内], [有徳], [上層部] 등도 그 분류의 위치가 수정되어야 하는 것들이다.

5-2. 분류항목명의 통일성

분류어휘표의 분류항목명은 형태적으로 보통 명사이다. 그런데 1.3091과 1.5160에서는 각각 見る(보다), 物質の変化(물질의 변화)처럼 동사이거나, 명사구의 형태이거나 하여 통일적이지 못 하다.

5-3. 파생어의 단락 내 배치 방법에 일관성 결여

1.3021-07항에는 [信賴性], [信憑性]이 각각 [信賴], [信憑] 옆에 분류되어 있다. 그러나 1.3041-01항에는 [自尊心], [自負心]의 분류가 [自尊], [自負]와 각기 떨어져 분류되어 있어 단락 내 파생어를 배치하는 방법에 대한 일관성이 결여되어 있다.

07　信用　信賴　信賴性　信憑(しんぴょう)　信憑性
01　自主　自信　自尊　自任
　　自認　自負　矜持　自恃
　　自尊心　自負心　プライド　エリ　ト意識

6. 맺음말

본고의 머리말에서는 병렬시소러스를 개발하는 의의를 기술하고, 양언어의 시소러스의 개발 현황을 개괄적으로 설명하였다.

2장에서는 먼저 일본의 분류어휘표의 장단점을 논한 뒤, 이와 관련하여 한국어 시소러스에서는 어떠한 점이 보완되어야 하는가를 논하였는데, 이를 요약하면 분류어를 50만~100만어의 대용량으로 확대한다는 점과 분류체계를 세분화한다는 점, 빈도수 정보와, 한정어 정보를 제공한다는 점이다.

그 다음 3장~5장에서는 분류작업을 하는 과정에서 일어나는 여러 가지 문제점을, 한일 양언어 사이에 형태가 같은 동형 한자어나 동형 외래어에 대한 처리 방법, 분류작업 시 연세한국어사전의 잘못된 뜻풀이나 용례를 어떻게 처리할 것인가에 대한 것, 분류어휘표의 문제점으로 나누어 서술을 했다.

끝으로 금후의 병렬시소러스의 작업의 진척과 전망에 대해서 간단히 언급해 둔다.

(1) 병렬시소러스의 완성 : 36,700어 체언류 병렬시소러스의 완성(2005년), 52,000어 한일 병렬시소러스의 완성(2006년), 96,000어 한일 병렬시소러스의 완성(2007년)

(2) 분류체계 개선 : 한국어 대용량 시소러스(메크로시소러스) 완성을 위한 분류체계의 보완(2007년)

(3) 한국어 시소러스의 확대 : 20만어 한국어 대시소러스의 완성(2010년), 이후 해마다 新語, 未知語(사전 미등록어)의 추가 및 유지, 보수

(4) 방언(북한어 포함), 고어, 속어, 고유어, 전문용어를 포함하는 50만어 규모의 한국어 대용량 시소러스 구축(2015년)

【參考文獻】

- 김광해(1993), 『유의어·반의어 사전(개정판)』, 한샘출판
- 연세대학교 언어정보개발연구원 편(1998), 『연세한국어사전』, 두산동아
- 池原悟 外(1997), 『日本語語彙大系』, 岩波書店
- 日本電子化辭書研究所(1995), 『EDR電子化辭書仕様說明書』
- 國立國語研究所(1964), 『分類語彙表』, 秀英出版
- 國立國語研究所(2003), 『分類語彙表』(改訂增補版), 國立國語研究所
- 宮島達夫·小沼悅(1992), 「言語研究におけるシソーラスの利用」, 『研究報告集』13, 國立國語研究所
- 長尾眞(1996), 『自然言語處理, 岩波講座ソフトウェア科學15』, 岩波書店
- 田島毓堂(2003), 「『分類語彙表』コードからの發展－言語教育のための語彙詳細コードの提案－」, 한국시소러스연구회 국제학술포럼 발표논문집, p17

要 旨

本稿では、はじめに並列シソーラスの開發に對して學問的、敎育的、産業的側面等五つの側面からその意義を叙述し、次に日本と韓國のシソーラスの開發現況を列擧しながら、特に韓國では國語全体を對象にするマクロシソーラスの開發が非常に遅れていることを指摘した。

次に、2章「韓國語シソーラスの補完」では、分類語彙表は研究用、敎育用としては向いているが、情報處理用としてはあまり向いていないという、分類語彙表の長所と短所を述べ、その分類語彙表の改善策として分類語を50万～100万語に擴大すること、それとともに分類体系を細分化して情報處理用としても使いやすくすることを主張した。 また分類語彙表にはない頻度數情報を記入する方法、限定語情報を付け加える方法についても記述した。

3章の「分類の方法」では、まず分類の基本原則を提示し、分類上起きられやすい諸問題の中で、特に日本語と韓國語が語形は同じだが、意味が微妙に違っていて間違いやすい、漢語と外來語の分類方法を五つの点に分けて說明した。

4章では分類の根據を提供してくれる延世韓國語辭典の利用上の問題点を述べ、5章では分類の基底シソーラスである分類語彙表が抱えている問題点を述べた。

6章の「最後に」では、本文の内容をもう一度要約し、最後に今後の本作業の進行を展望した。

キーワード：韓國語シソーラス · データベース · 情報處理 · 分類語彙表 · 延世韓國語辭典 · 大容量シソーラス · 頻度數情報 · 限定語情報

투　　고：2005. 11. 30
1차 심사：2005. 12. 10
2차 심사：2005. 12. 31

住　所：(520-714) 전남 나주시 대호동 252 동신대학교 일본어학과
電　話：010-3644-3641
e-mail：yusukhan@hanmail.net