KIPS Tr. Comp. and Comm. Sys.
Vol.5, No.1 pp.7~16 pISSN: 2287-5891

http://dx.doi.org/10.3745/KTCCS.2016.5.1.7

A Study on a Method of Selecting Variant Groups to be Reviewed for LGR (Label Generation Rule) of Internet Top-Level Hanja Domain

Kyongsok Kim†

ABSTRACT

This paper discusses a method of selecting variant groups to be reviewed for LGR (Label Generation Rule) of Internet Top-Level Hanja Domain. The most difficult problem in setting up LGR of Internet Top-Level Hanja Domain is how to treat Hanja variants. If domains containing variants (e.g.: 東海國, 東海国) are directed to different addresses, confusion will arise. Therefore, it is desirable that such domains are directed to the same address. Since variant groups of Korea and China are not same, we need to unify variant groups of Korea and China. In the process of reviewing 3093 Chinese variant groups, the author found that Korea does not need to review Chinese variant groups which include no or just one Korean Hanja character. Korea only need to review Chinese variant groups which include two or more Korean Hanja characters. By doing so, the author could reduce the number of Chinese variant groups to be reviewed by Korea from 3093 to 303, which is only one-tenth of the original number of Chinese variant groups. After Korea finishes reviewing 303 Chinese variant groups selected according to the method suggested in this paper, the job of setting up LGR of Internet Top-Level Hanja domain will be accelerated by negotiating with China.

Keywords : Variant Group, Hanja, Label Generation Rule, Top-Level Domain, Korea-China

인터넷 최상위 한자 도메인의 국제 생성 규칙(LGR)을 위한 검토 대상 이체자 묶음 선정 방안 연구

김 경 석†

요 약

이 논문은 인터넷 최상위 한자 도메인의 국제 생성 규칙(LGR)을 위하여 검토 대상 이체자 묶음 선정 방안을 연구한다. 한자 도메인의 국제 생성 규칙을 정하는데 있어 가장 어려운 점은 한자 이체자를 어떻게 처리할 것인가이다. 만일 이체자가 들어간 도메인들(보기: 東海國, 東海国)이 각각 다른 주소로 접속된다면 혼란이 일어나므로 같은 주소로 접속되는 것이 바람직하다. 그런데 한중의 이체자 묶음이 같지 않으므로 한중의 이체자 묶음을 통합할 필요가 있다. 이런 이유로 저자는 중국 이체자 묶음 3093개에 대한 검토 작업을 진행하다가 한국 한자가 전혀 없거나 한 글자만 있는 중국 이체자 묶음은 한국이 검토할 필요가 없음을 발견하였다. 한국은 중국 이체자 묶음 가운데 한국 한자가 2자 이상 있는 것만 검토하면 된다. 이렇게 함으로써 저자는 한국이 검토해야 할 중국 이체자 묶음의 개수를 3093개에서 303개로 줄일 수 있었는데, 이는 원래 중국 이체자 묶음 개수의 1/10 이다. 본 논문에서 제안한 방법에 따라 중국 이체자 묶음 303개에 대한 검토가 마무리되면 중국과의 협의를 통해 인터넷 최상위 한자 도메인의 국제 생성 규칙을 정하는 작업이 가속화될 것이다.

키워드 : 이체자 묶음, 한자, 국제 생성 규칙, 최상위 도메인, 한중

1. 서 론

현재 ICANN(Internet Corporation for Assigned Names and Numbers, 국제 인터넷 주소 기구)에서는 최상위 한자 도메인 생성을 위한 국제 생성 규칙(LGR, Label Generation Rule)을 만드는 작업이 진행 중인데[1, 2], 이는 2013년부터 시작된 작업이다. 저자는 한국측 위원장으로 이 작업에 지속적으로 참여하고 있다. 현재 .中国/.中國, .香港, .台灣/.台湾 처럼 ccTLD(국가 코드 최상위 도메인, country code Top-Level Domain)인 한자 도메인은 쓰고 있지만, gTLD(일반 최상위 도메인, generic Top-Level Domain)인 한자 도메인은 쓰고 있지 않다.

ICANN은 한자에 관심을 가지는 주요국인 Korea(한국), China(중국), Japan(일본)에게, 세 나라가 합의한 한자 최상

† 정 회 원 : 부산대학교 정보컴퓨터공학부 교수
Manuscript Received : October 16, 2015
First Revision : November 25, 2015
Accepted : December 7, 2015
* Corresponding Author : Kyonsok Kim(gimgs@pnu.kr)

위 도메인 생성 규칙(안)을 제출하여 줄 것을 요청하였다. 과거에도 한중일 도메인에 관한 RFC 문서[3]와 한자 도메인에 관한 RFC 문서[4]가 나온 바 있다.

일본 말은 일본에서만 쓰지만, 중국 말은 중국뿐만 아니라, 타이완, 홍콩, 마카오 등에서도 쓰며, 한자에 대한 중국의 한자 목록과 이체자 목록은 중국, 타이완, 홍콩, 마카오가 참여하는 CDNC(Chinese Domain Name Consortium)[5]에서 공동으로 만든 안이다.

한국 말(Korean Language)은 남과 북에서 쓰고 있지만, 현실적으로 북의 의견을 반영할 길이 많지 않아서 사실상 남쪽 주도로 진행되고 있다. K(Korea) 한자 목록은 남북의 주요 한자 목록을 합집합 하였기에 남의 의견뿐만 아니라 북의 의견까지 반영하였다고 볼 수 있다. 하지만, 이체자 목록은 북의 의견을 반영할 접촉 창구나 방법이 없어서 남쪽의 의견만 반영하였다. 아래에서 K, Korea, 또는 한국이라고 할 때 한자 목록은 남북의 의견이 반영되었고, 이체자 목록은 남쪽 의견만 반영된 것임을 밝혀둔다.

ICANN의 요청에 따라 2014년부터 한중일 세 나라가 1년에 세 번 열리는 ICANN 회의 및 그 밖의 다른 회의에서 계속적으로 만나서 협의하고 있으며, 저자도 여러 번 회의에 참석하였다. 아마도 2016년 또는 2017년쯤에 합의한 한자 생성 규칙(LGR)이 만들어질 것으로 보인다.

한국에서는 한자, 중국에서는 한쯔(Hanzi), 일본에서는 간지(Kanji)라고 불리는 한자에 관한 문제는 크게 두 가지 주제가 있다. 하나는 각 나라에서 최상위 도메인에 쓰고자 하는 한자 목록이고, 또 하나는 한자의 이체자 목록이다. 한자 목록은 각 나라의 실정에 따라 최상위 도메인에서 쓰고자 하는 한자 목록을 만들기만 하면 되기 때문에 세 나라 사이의 협의도 필요하지 않고 별로 어려운 점이 없다.

그런데 이체자 목록은 각 나라가 이체자 목록을 만들기만 하면 되는 것이 아니라, 세 나라의 이체자 목록을 통합해야 하므로 복잡한 문제가 생긴다. 이 논문에서는 한중일의 이체자 목록이 주어졌을 때 이체자 목록 통합을 위하여 한국이 검토해야 할 중국 이체자 묶음을 선정하는 방안을 다루고자 한다.

2장에서는 이체자란 무엇이며, 인터넷 도메인에서 왜 이체자가 문제 되는지 살펴보자. 3장과 4장에서는 이 논문의 핵심 내용인 중국과 한국의 이체자 목록 통합하기 위하여 검토해야 할 이체자 묶음을 선정하는 방안을 살펴본 뒤, 5장에서는 맺음말을 보겠다.

2. 한자 이체자의 개념과 인터넷 도메인에서의 문제점

이 논문에서 다룰 주제는 한중일의 이체자 목록 통합을 위하여 한국이 검토해야 할 이체자 묶음 선정 방안이므로, 2장에서는 이체자란 무엇인지, 그리고 인터넷 도메인에서 왜 이체자가 문제 되는지 살펴보겠다.

2.1에서는 이체자란 무엇인지 살펴본 뒤, 2.2에서는 중국만의 특수한 이체자인 간화자와 번체자를 살펴보고, 간화자 가운데 특별한 간화자인 간번자를 2.3에서 살펴보겠다. 2.4에서는 인터넷 도메인에서 왜 이체자가 문제 되는지 알아보겠다.

2.5에서는 중국의 한자 목록과 이체자 묶음을, 2.6에서는 일본의 간지 목록과 이체자 묶음을, 2.7에서는 K(한국) 한자 목록과 이체자 묶음을 각각 살펴보겠다.

2.1 이체자란 무엇인가?

한자에는 모양(꼴, glyph)은 달라도 뜻이 같은 이체자(variant)가 있다. 한국에서 사용하는 이체자의 보기를 보자.

峰 (U+5CF0) (“U+”는 ISO/IEC 10646[6]의 부호 자리 앞에 붙이는 기호로 ISO/IEC 10646에서 한자 峰을 나타내는 부호 자리가 16진법으로 5CF0라는 뜻이다)

峯 (U+5CEF)

위의 한자 두 자는 음(소리 값)과 훈(새김)이 모두 “봉우리 봉”으로 같으며 두 자는 서로 이체자 관계에 있다. 이 한자 두 자를 자세히 살펴보면 한자를 이루는 구성 요소는 같은데 뫼 산(山)자가 왼쪽에 있는지 위쪽에 있는지만 다르다. 다시 말하여 뫼 산(山)자가 나머지 구성 요소와의 상대적인 자리가 다를 뿐이다. 다른 이체자 보기를 보자.

弔 (U+5F14)

吊 (U+540A)

위의 한자 두 자는 소리 값과 새김이 모두 “조상할 조”로 같으며 두 자는 서로 이체자 관계에 있다. 얼핏보면 다른 한자로 볼 수도 있지만, 이 두 자는 같은 한자로 본다.

2.2 중국만의 특수한 이체자: 간화자와 번체자

위의 2.1에서 본 이체자 상황이 중국에도 있다. 그런데 중국에는 이 밖에도 중국만의 독특한 이체자 상황이 하나 더 있는데, 그건 바로 간화자(簡化字, jiǎnhuàzì 지앤화쯔) - 번체자(繁体字, fántǐzì, 판티쯔) 문제이다.

중국은 1956년에 한자 간화 방안(漢字 簡化 方案)을 발표하고, 그 뒤 몇 년 동안의 연구를 거쳐 1964년에 간화자 총표(簡化字 總表)를 발표하였는데 여기에는 간화자 2235자가 들어있다[7].

중국에서 간화자에 대응하는, 간화되기 전의 복잡한 한자를 번체자라고 부른다. 번체자는 우리가 보통 말하는 정자와 비슷한 개념이고, 간화자는 약자와 비슷한 개념이다. 다만 중국의 간화자는 간화자 총람에 나오는 2235자만 가리키며, 각 간화자에 대응하는 번체자도 정해져 있다. 또한 간화자 2235자 가운데 약 1700자는 그 전에는 쓰지 않던 완전히 새로 만든 한자이고, 나머지 약 500자는 그 전부터 쓰던 한자를 간화자로 지정하였다. 완전히 새로 만든 간화자 1700

Table 1. The Number of Simplified Characters and Corresponding Traditional Characters

간화자 1자에 대응하는 번체자 수 (가)	간화자 수 (나)	간화자에 대응하는 모든 번체자 수 (가) * (나)
1자	2211자	2211자
2자	22자	44자
3자	2자	6자
합계	2235자	2261자

자는 한국에서는 거의 쓰지 않고 있다. Table 1에서 보듯이 간화자 한 자에 대응하는 번체자는 보통 한 자이지만, 대응하는 번체자가 두 자 또는 세 자인 경우도 있다.

이제 중국의 간화자와 번체자 사이의 이체자 보기를 보자.

東 (U+6771)
东 (U+4E1C)

위의 한자는 모두 “동녘 동” 자인데, 東(U+6771)은 우리에게 익숙한 번체자이고, 东(U+4E1C)은 간화자인데 우리에게 익숙하지 않다. 중국에서는 이 두 자를 이체자로 본다.

2.3 간번자 (簡繁字, simplitional character)

중국의 간화자 가운데 보기를 들어 东(U+4E1C)은 그 전에 전혀 쓰지 않던, 간화자 총람에 새로 나온 한자이다. 그러나 중국의 간화자 가운데에는 그 전부터 이미 쓰던 한자인 경우도 있는데, 아래 보기를 보자.

機 (U+6A5F)
机 (U+673A)

중국에서는 위의 한자 두 자가 모두 “틀(machine) 기”인데, 机(U+673A)는 번체자 機(U+6A5F)의 간화자이다. 그런데 간화자 东과는 달리 간화자 机는 그 전부터 오랫 동안 이미 써 오던 한자라는 데 문제가 있다. 이처럼 간화자 2235자 가운데 약 500자는 그 전부터 이미 오랫 동안 써오던 한자를 간화자로 지정하였다(이 주제에 대하여는 다른 논문에서 자세히 다루겠다). 중국 공항에 내리면 눈에 띄는 것이 机場(ji chang, 지창)이라는 한자인데, 이는 공항을 뜻하며 機場의 간화자이다.

그런데 한국에서 机는 “책상(desk) 궤”이고, 機는 “틀 기”로 완전히 다른 한자이며, 이 두 자는 한국과 중국 모두에서 오래 전부터 써오던 한자이다. 한국 한자에 익숙한 사람은 중국의 机場을 보고 기장(지창)이 아니라 궤장이라고 읽게 될 것이고, 그 뜻이 무엇일까 하고 고개를 갸우뚱할 것이다.

위의 내용을 요약하면 机라는 하나의 한자에 대해 한국과 중국에서 뜻이나 용법이 완전히 다르다. 이처럼 간화자 총표 2235자 가운데, 새로 만든 한자가 아니라 그 전부터 써오던 한자인데 간화자로 지정된 한자를 간번자(simplitional character)라고 부르겠다 (“간번자”라는 용어는, 현재는 “간” 화자이지만, 오래 전부터 써오던 “번”체자라는 뜻에서 “간” + “번” + “자”라고 만든 용어이고, simplitional이라는 용어는 simplified + traditional에서 따온 용어이다).

Table 2. Two Situations where www.東海 and www.东海 Point to the Same or Different IP Addresses

도메인 \ 상황	상황 1	상황 2
www.東海	111.10.100.1 사이트 접속	111.10.100.1 사이트 접속
www.东海		222.20.200.2 사이트 접속

2.4 인터넷 도메인에서 왜 이체자가 문제가 되는가?

이제 인터넷 도메인에서 왜 이체자가 문제 되는지 위에서 본 동녘 동 자를 포함하는 아래의 도메인 두 개를 구체적으로 생각해보자.

東海
东海

그리고 東海와 东海를 최상위 도메인(TLD, Top-Level Domain)으로 하는 아래의 URL 주소 두 개를 생각해보자.

www.東海 (번체자 주소)
www.东海 (간화자 주소)

위의 주소를 웹 브라우저 주소 창에 쳤을 때 Table 2에서 볼 수 있듯이 다음과 같은 두 가지 상황을 생각해볼 수 있다. 상황 1에서는 東海와 东海가 같은 IP 주소로 가고(東海와 东海의 소유자가 같을 가능성이 높음), 상황 2에서는 다른 IP 주소(東海와 东海의 소유자가 다를 가능성이 높음)로 간다.

중국 사용자의 경우 東海와 东海는 같은 도메인으로 생각하기 때문에 상황1은 그런대로 무난하지만, 상황 2는 혼란스러울 수 있다. 상황 2이라면 가짜 누리집(home page)을 만들어 사기를 칠 수도 있는 등 혼란스럽게 된다.

이러한 이유로 한자 도메인을 하나 만들면 그 도메인의 한자 가운데 이체자가 있으면 이체자로 바꾼 도메인까지 묶어서 그 모든 도메인을 하나의 묶음으로 처리하는 것이 바람직하다. 위의 보기에서 중국 사용자가 东海라는 도메인을 신청하면 东海뿐만 아니라 東海까지 도메인 두 개를 하나의 묶음으로 처리한다는 것이다.

여기서 “하나의 묶음으로 처리한다”라는 말은 몇 가지 가능성을 열어둔다. 첫째, 신청자에게 东海뿐만 아니라 東海까지 같이 쓸 수 있게 하되, 꼭 같은 IP 주소로 가게 한다. 둘째, 신청자에게 东海는 쓸 수 있게 하고 東海는 아무도 쓰지 못하게 한다(아마도 중국 신청자인 때). 셋째, 신청자에게 東海는 쓸 수 있게 하고 东海는 아무도 쓰지 못하게 한다(아마도 타이완 신청자인 때).

Table 3. The difference between when two characters, 机 and 機, are bundled in a variant group and when they are not

도메인	(机, 機)가 이체자 묶음으로 묶였을 때			(机), (機)가 이체자 묶음으로 묶이지 않고, "외톨이" 한자인 때
	가능성 1	가능성 2	가능성 3	
wx机yy	소유자 1이 wx机yy, wx機yz 두 개 모두 소유	소유자1이 wx机yy만 씀.	wx机yy는 아무도 쓰지 못 함	소유자 1
wx機yz		wx機yz는 아무도 쓰지 못 함	소유자1이 wx機yz만 씀.	소유자 2

다른 보기로 东海國이라는 도메인에는 동녘 동 东과 나라 국 國에 각각 이체자가 있으므로 조합 가능한 도메인은 2*1*2=4개가 된다: 东海国, 東海國, 东海國, 東海国. 다만 이 가운데 셋째와 넷째 도메인에는 간화자와 번체자가 섞여 있으므로 도메인으로 쓰지 않을 가능성이 높다. 신청자가 네 개 가운데 몇 개를 실제 도메인으로 쓰든, 이 도메인 네 개는 신청자가 아닌 다른 사람은 쓸 수 없게 된다. 긴 한자 도메인이라면 이체자가 있는 한자가 많을 수 있고 더욱이 한자 한 자에 대해 이체자가 최대 여섯 개까지 더 있을 수 있어서 하나의 묶음으로 다루어야 할 한자 도메인 개수는 수십 개, 수백 개가 될 수도 있다.

이처럼 한중일 사이에 이체자 묶음을 어떻게 결정하느냐에 따라, 어떤 도메인을 신청했을 때 묶음으로 처리될 도메인의 개수가 결정되므로 한중일 사이의 도메인 묶음 통합 문제는 아주 중요하다. Table 3에서 보듯이 机와 機가 이체자 묶음으로 묶였을 때 机가 들어간 도메인이 쓰이고 있으면, 机 대신 機가 들어간 도메인을 다른 사람은 쓸 수 없게 된다. 그렇지만 机와 機가 도메인 묶음으로 묶이지 않았을 때 机가 들어간 도메인이 쓰이고 있더라도, 机 대신 機가 들어간 도메인을 다른 사람이 쓸 수 있게 된다.

机와 機의 보기에서, 한국은 이 두 자를 이체자 묶음으로 묶지 않고 외톨이 한자(이체자가 없는 한자)로 두는 것이 바람직하고, 중국은 이 두 자를 이체자 묶음으로 묶는 것을 좋아할 것이다. 이런 상충되는 이해 때문에 한중일 사이에 이체자 묶음 통합은 아주 어려운 문제이다.

2.5 중국의 한자 목록(12563자)과 이체자 묶음(3093개)

이제 중국의 이체자 묶음을 살펴보자. 이체자 묶음이란 상호 이체자 관계에 있는 한자를 하나로 묶은 것이다. 중국이 2015.04.30.일에 발표한 한쯔(한자) 목록[8]을 바탕으로 저자가 분석한 결과 중국 한자 목록의 한자 수는 12563자였고, 이체자 목록을 분석하니 이체자 묶음 3093개가 있었으며, 그 이체자 묶음 안의 한자(이체자) 총 수는 6889자로, 이체자 묶음 당 평균 이체자 수는 2.2자였다. 중국의 이체자 묶음 보기 6개가 아래에 나와 있다. 아래에서 맨 마지막 이체자 묶음에는 이체자가 7자나 들어있다.

(U+4E1C 东) (U+6771 東): 동녘 동의 이체자 묶음
(U+673A 机) (U+6A5F 機): 틀 기의 이체자 묶음
(U+7ADC 竜) (U+9F8D 龍) (U+9F99 龙): 용 룡의 이체자 묶음
(U+4E00 一) (U+58F1 壱) (U+58F9 壹) (U+5F0C 弌): 한 일의 이체자 묶음
(U+56EF 国) (U+56FD 国) (U+5700 圀) (U+570B 國): 나라 국의 이체자 묶음
(U+5B81 宁) (U+5BCD 寔) (U+5BD5 寧) (U+5BD7 寗)
(U+5BDC 寜) (U+5BE7 寧) (U+752F 甯): 寧(편안할 녕)/寗(차라리 녕) 등의 이체자 묶음

2.6 일본의 간지 목록(6356자)과 이체자 묶음(0개)

2015.03.23.일에 발표한 일본의 한자 목록과 이체자 목록에 따르면, 한자 수는 6356자이고, 이체자는 전혀 없다. 실제로는 일본에 이체자가 있다고 볼 수 있다. 다시 말하여 1947년까지 썼던 구자체(旧字体, kyujitai)라고 불리는 정자, 그리고 1947년부터 쓰는 신자체(新字体, shinjitai)라고 불리는 약자가 있는데, 일본의 간지(한자) 도메인에서는 이체자로 인한 혼란이 크게 없을 것으로 보고, 한자 최상위 도메인에 관한 한 일본은 이체자가 없다고 발표하였다. 물론 앞으로 한중일 사이에 이체자 묶음이 확정되기 전에 일본이 입장을 바꾸어 이체자를 발표할 가능성이 없는 것은 아니다.

2.7 K(한국) 한자 목록(4819자)과 이체자 묶음(37개)

2015.08.13.일에 발표한 K(한국) 한자 목록 v0.3판에는 한자 4819자가 있으며 이체자 묶음은 37개가 있다[9]. 앞으로 한자 목록은 살짝 바뀔 가능성이 있으며, 이체자 목록은 중국과의 협상 결과에 따라 바뀔 가능성이 높지만, 아주 많이 바뀔지 아니면 조금만 바뀔지는 알 수 없다.

3. C(중국)의 이체자 묶음 3093개 가운데 한국이 검토해야 할 이체자 묶음을 선정하는 방안

이제 중국이 2015.04.30.일에 발표한 한자 목록과 이체자 목록, 그리고 한국이 2015.08.13.일에 발표한 한자 목록과 이체자 목록(0.3판)을 바탕으로 중국의 이체자 묶음 가운데 한국이 검토해야 할 이체자 묶음을 선정하는 방안을 살펴보겠는데, 좋은 검토 대상 이체자 묶음 선정 방안을 찾아내는 것이 이 논문의 핵심적인 부분이다.

처음 중국 이체자 목록 검토를 시작할 때에 가장 큰 고민은 중국은 이체자 묶음이 3093개이고, 한국은 37개이므로 비율이 100:12인데 이를 어떻게 하나로 통합할 수 있을까 였다. 이제 이 큰 고민을 상당히 해결한 과정을 아래에서 살펴보겠다.

3.1에서는 이체자 목록이 만족시켜야 하는 세 가지 성질을 먼저 살펴본 뒤, 3.2에서는 이체자 검토 작업의 네 단계를 살펴보겠다.

3.3에서는 C(중국) 이체자 묶음의 크기와 이체자 묶음 안의 K(한국) 한자 수별 표에 대해서 살펴보고, 3.4에서는 K 한자 수가 0이나 1 인 C 이체자 묶음은 검토하지 않아도 된다는 점을 살펴본 뒤, 3.5에서는 K 한자 수가 2 이상인 C 이체자 묶음은 반드시 검토해야 하며, 이런 묶음만 검토하면 된다는 점을 알아보겠다.

3.4와 3.5를 바탕으로 3.6에서는 C 이체자 묶음 3093개 가운데 303개만 검토하면 된다는 점을 알아보겠다. 3.7에서는 3.3의 Table 4에 대한 상세한 표(Table 5)를 살펴보고, 마지막으로 3.8에서는 3.7의 Table 5에 대한 요약한 표(Table 7)를 살펴보겠다.

3.1 이체자 목록이 만족시켜야 하는 세 가지 성질

이체자 목록은 reflexivity(재귀성, 반사성, 되돌아오기), symmetry(대칭성), transitivity(이행성, 옮겨가기)의 세 가지 성질을 만족시켜야 하는데 각 성질이 무엇을 뜻하는지 알아보자.

1) reflexivity: 왼쪽 표제 한자("[...]" 안에 있는 한자)에 대응하는 오른쪽의 이체자들 ("(...)" 안에 있는 한자들)에 표제 한자 자신이 반드시 나와야 한다는 것을 뜻한다. 아래 보기에서는 표제 한자 [U+4E1C 東]가 오른쪽에 나오지 않기 때문에 reflexivity를 어긴 것이며 --> 아래와 같이 바로잡아야 한다. 다시 말하면 a -> a가 반드시 있어야 한다는 것이다.

```
[U+4E1C 东] (U+6771 東) # reflexivity 어김
-->
[U+4E1C 东] (U+4E1C 东) (U+6771 東) # reflexivity 지킴
```

2) symmetry: 아래 보기에서 표제 한자 [U+4E1C 东]의 이체자에는 (U+6771 東)이 나오는데, 거꾸로 표제 한자 [U+6771 東]의 이체자에는 (U+4E1C 东)이 나오지 않기 때문에 symmetry를 어긴 것이며, --> 아래와 같이 바로잡아야 한다. 다시 말하면 a -> b이면 b -> a도 반드시 있어야 한다는 것이다.

```
[U+4E1C 东] (U+4E1C 东) (U+6771 東)
[U+6771 東] (U+6771 東)
-->
[U+4E1C 东] (U+4E1C 东) (U+6771 東)
[U+6771 東] (U+4E1C 东) (U+6771 東)
```

3) transitivity: 아래 보기에서 표제 한자 [U+7ADC 竜]의 이체자로 (U+9F8D 龍)이 있고, 또한 표제 한자 [U+9F8D 龍]의 이체자로 (U+9F99 龙)이 있는데, 표제 한자 [U+7ADC 竜]의 이체자로 (U+9F99 龙)이 없기 때문에 transitivity를 어긴 것이며, --> 아래와 같이 바로잡아야 한다. 다시 말하면 a -> b, b -> c이면 a -> c도 반드시 있어야 한다는 것이다.

```
[U+7ADC 竜] (U+7ADC 竜) (U+9F8D 龍)
[U+9F8D 龍] (U+7ADC 竜) (U+9F8D 龍) (U+9F99 龙)
[U+9F99 龙] (U+7ADC 竜) (U+9F8D 龍) (U+9F99 龙)
-->
[U+7ADC 竜] (U+7ADC 竜) (U+9F8D 龍) (U+9F99 龙)
[U+9F8D 龍] (U+7ADC 竜) (U+9F8D 龍) (U+9F99 龙)
[U+9F99 龙] (U+7ADC 竜) (U+9F8D 龍) (U+9F99 龙)
```

3.2 이체자 검토 작업 네 단계: 이체자 묶음 3093개(이체자 6889자)를 다 검토해야 하는가?

이체자 검토 작업은 아주 많은 단계를 거치지만 크게 보면 다음 네 단계를 거쳤다.

첫째 단계는 중국 이체자 목록이 reflexivity, symmetry, transitivity를 만족하는지 확인하는 것이었다. 아마도 중국은 프로그램으로 확인하지 않았는지 저자가 프로그램으로 확인하자 symmetry/ transitivity 문제 2건이 나왔다[10]. 이에 대해 중국에 확인하자 잘못을 인정하고 다음 판 이체자 목록에서는 바로잡겠다고 약속하였다. 그리고 약 600건의 reflexivity 문제가 있었는데[10] 이에 대해 중국에 확인하자 xml로 된 자료는 정확한데 이를 excel로 바꿀 때 실수로 그렇게 되었는데 이는 단순한 실수라고 하였으며, 사실 이는 큰 문제가 되는 건 아니다.

둘째 단계는 이체자 목록을 이체자 묶음으로 바꾸는 것이었다. 한 일 자에 대한 이체자 목록을 아래에서 생각해 보자. 한 일 자에는 이체자가 넉 자 있는데, 이체자 목록에는 다음과 같이 왼쪽의 표제 한자 넉 자 ("[...]" 안에 있는 한자) 각각에 대하여 오른쪽에 이체자 넉 자("(...)" 안에 있는 한자)가 나온다.

```
[U+4E00 一] (U+4E00 一) (U+58F1 壱) (U+58F9 壹) (U+5F0C 弌)
[U+58F1 壱] (U+4E00 一) (U+58F1 壱) (U+58F9 壹) (U+5F0C 弌)
[U+58F9 壹] (U+4E00 一) (U+58F1 壱) (U+58F9 壹) (U+5F0C 弌)
[U+5F0C 弌] (U+4E00 一) (U+58F1 壱) (U+58F9 壹) (U+5F0C 弌)
```

이런 상태에서 그대로 검토하게 되면 꼭 같은 이체자 묶음에 대하여 4번 검토하게 되어 시간을 낭비할뿐만 아니라, 네 번 검토한 결과가 반드시 일치하지 않을 수도 있게 된다. 그래서 위와 같은 경우 표제 한자를 없애고 나면 네 줄이 꼭 같아지므로 다음과 같이 하나만 남도록 바꾸었는데, 이를 이체자 묶음(variant group)이라고 한다.

```
--> (U+4E00 一) (U+58F1 壱) (U+58F9 壹) (U+5F0C 弌)
```

이제 이체자 목록에서 네 번 검토하는 대신 이체자 묶음

Table 4. Table showing the number of C variant groups for each of the size of C variant groups and the number of K characters

C 이체자 묶음 크기	C 이체자 묶음 안에 있는 K 한자 수					
	0	1	2	3	C 기준 소계	
1	3166개	2508개	--	--	5674개	5674개
2	1059개	1336개	161개	--	2556개	3093개 (C 이체자 묶음 크기 >= 2)
3	42개	281개	78개	12개	413개	
4	8개	50개	30개	6개	94개	
5	3개	9개	8개	1개	21개	
6		1개	3개	2개	6개	
7		1개	1개	1개	3개	
K 기준 소계	4278개	4186개	281개	22개	8767개	8767개
			303개			

을 한 번만 검토하면 되게 되었다.

중국 이체자 목록 전체적으로 보면 이체자 묶음 3093개에 이체자 6889자가 있으므로, 6889개를 검토하는 대신 3093개만 검토하면 되었으니 검토 대상이 55% 줄어들었다.

그리고 나서는 이체자 묶음을 검토하기 시작했다. 이 때에는 한자 전문가가 3093개의 이체자 묶음에 있는 한자 이체자 6889자를 모두 검토해야 한다고 생각하였다. 그러나 막상 시작해보니 그렇게 많은 이체자 묶음을 일일이 검토한다는 것이 쉬운 일이 아니었고, 성과가 별로 좋지 않았다. 이체자 묶음 한 개 검토에 1분만 잡아도 3093개를 검토하는데 50시간이 걸리며, 결국 하루 5시간씩 두 주간 검토해야 한다는 얘기가 된다.

이에 컴퓨터 전공인 저자는 여러 가지 방법으로 프로그램을 만들어 좀 더 짧은 시간 안에 검토할 수 있는 방안을 찾고자 하였다. 그래서 셋째 단계로 아래에서 보듯이 C 이체자 묶음에 있는 각 한자(이체자)가 K 한자 목록에 있는지 없는지를 O, X로 나타내었다.

(U+4E05 丌) (U+4E0B 下)
-->
(X U+4E05 丌) (O U+4E0B 下)

위의 보기에서 “(X U+4E05 丌)”에서 “X”는 “(U+4E05 丌)” 한자가 K(한국) 한자 목록에 없다는 것을 나타내며, “(O U+4E0B 下)”에서 “O”는 “(U+4E0B 下)“ 한자가 K(한국) 한자 목록에 있다는 것을 나타낸다.

이렇게만 해도 한자 전문가는 검토하기가 훨씬 쉬워졌다고 하였다. 그렇지만 아직도 검토 작업 진행이 너무 더디어서 다른 방안을 찾아야만 했다.

많은 시간을 들여서 이런 저런 방안을 찾다가 드디어 넷째 단계로 3.2의 Table 4를 만들고 검토 시간을 획기적으로 줄일 수 있는, 검토해야 할 중국 이체자 묶음 수를 10분의 1로 줄일 수 있는 선정 방안을 찾게 되었다. 이체자 묶음 검토 연구를 시작할 때에는 전혀 생각하지도 못했던 획기적으로 좋은 선정 방안을 찾은 것이다.

3.3 C(중국) 이체자 묶음의 크기와 이체자 묶음 안의 K(한국) 한자 수

Table 4에서 각 칸에 있는 수의 뜻을 살펴보기 전에 먼저 알아두어야 할 것이 있다. 아래에 나오는 이체자 묶음 세 개에서 각 줄의 처음에 2가 나오는데, 이는 C(중국) 이체자 묶음 크기가 2라는 것을 나타낸다. C 이체자 묶음 크기가 2라는 것은 C 이체자 묶음 안에 C 한자가 두 자 있다는 것인데, 아래 세 줄의 각 줄에 한자가 두 자씩 있음을 알 수 있다.

2 0 (X U+4F21 佚) (X U+4FE5 俥)
2 1 (X U+4E05 丌) (O U+4E0B 下)
2 2 (O U+4E07 万) (O U+842C 萬)

위의 첫째 줄의 “(X U+4F21 佚)”에서 “X”는 “(U+4F21 佚)” 한자가 K(한국) 한자 목록에 없다는 것을 나타내며, “(O U+4E0B 下)”에서 “O”는 “(U+4E0B 下)“ 한자가 K(한국) 한자 목록에 있다는 것을 나타낸다.

위의 세 줄의 각 줄에 나오는 둘째 수는 각각 0, 1, 2인데 이는 그 C 이체자 묶음 안에 있는 K 한자 수, 다시 말하여 “O”로 표시된 한자의 수를 나타낸다. 첫째 줄의 경우, “X X”이므로 둘째 수가 0이며, 둘째 줄의 경우 “X O”이므로 둘째 수가 1이고, 셋째 줄의 경우 “O O”이므로 둘째 수가 2이다.

Table 4를 보면 C 이체자 묶음의 크기가 2인 때, 그 안에 K 한자 수가 0인 이체자 묶음 수는 1059개이고, K 한자 수가 1인 이체자 묶음 수는 1336개이고, K 한자 수가 2인 이체자 묶음 수는 161개임을 알 수 있다.

아래에는 C 이체자 묶음의 크기가 3이고, 이체자 묶음 안의 K 한자 수가 각각 0, 1, 2, 3인 이체자 묶음 보기가 나와 있다. 위의 풀이를 다 이해하였다면 아래 이체자 묶음에 대한 풀이는 필요 없으리라고 본다.

3 0 (X U+4D19 鶊) (X U+9DFF 鷿) (X U+9E0A 鶊)
3 1 (X U+4E21 两) (X U+4E24 两) (O U+5169 兩)
3 2 (O U+5186 円) (X U+5706 圓) (O U+5713 圓)
3 3 (O U+4E30 丰) (O U+8C4A 豐) (O U+8C50 豐)

Table 4를 보면 C 이체자 묶음의 크기가 3인 때, 그 안에 K 한자 수가 0인 이체자 묶음 수는 42개이고, K 한자 수가 1인 이체자 묶음 수는 281개이고, K 한자 수가 2인 이체자 묶음 수는 78개이고, K 한자 수가 3인 이체자 묶음 수는 12개임을 알 수 있다.

아래에는 C 이체자 묶음의 크기가 7이고, 이 묶음 안에 K 한자는 1자뿐인 보기가 나와 있는데, 이런 이체자 묶음은 1개뿐임을 Table 4에서 알 수 있다. 참고로 C 이체자 묶음의 최대 크기는 7이다.

7 1 (X U+7877 鹼) (X U+78B1 城) (X U+9669 险)
(X U+967A 険) (O U+96AA 險) (X U+9E78 鹼)
(X U+9E7C 鹼)

이제 Table 4에서 C 이체자 묶음의 크기가 1인 줄을 살

펴보자. 이체자 묶음의 크기가 1이라는 것은 이체자가 없다는 뜻이며, 이런 한자는 “외톨이” 한자이다.

1 0 (X U+34E4 刮)
1 1 (O U+4E0A 上)

“1 0”에 속하는 C 외톨이 한자는 3166자인데, 이 한자는 K 한자 목록에 없는 C 한자이다.
“1 1”에 속하는 C 외톨이 한자는 2508자인데, 이 한자는 K 한자 목록에 있는 C 한자이다.

3.4 K 한자 수가 0이나 1 인 C 이체자 묶음은 검토하지 않아도 된다

위에서 본 C 이체자 묶음 가운데 K 한자 수가 0인 이체자 묶음 보기가 아래에 두 개 나와 있다.

2 0 (X U+4F21 侐) (X U+4FE5 俥)
3 0 (X U+4D19 鶊) (X U+9DFF 鷺) (X U+9E0A 鸊)

이 C 이체자 묶음에는 K 한자가 한 자도 없기 때문에, K(한국)에서는 전혀 검토할 필요가 없다. C 이체자 묶음 안에 있는 모든 한자를 K(한국)가 쓰지 않기 때문에, 보기를 들어 (X U+4F21 侐) (X U+4FE5 俥) 이 두 자가 묶음으로 다루어지든 다루어지지 않든 K에는 아무런 영향이 없다. K에서는 (X U+4F21 侐) (X U+4FE5 俥) 이 두 한자를 어차피 도메인에 쓰지 않기 때문이다.

비슷하게 위에서 본 C 이체자 묶음 가운데 K 한자 수가 1인 C 이체자 묶음 보기가 아래에 두 개 나와 있다.

2 1 (X U+4E05 丅) (O U+4E0B 下)
3 1 (X U+4E21 两) (X U+4E24 两) (O U+5169 兩)

이 이체자 묶음에는 K 한자가 한 자뿐이기 때문에, K(한국)에서는 전혀 검토할 필요가 없다. K(한국)가 쓰는 한자가 한 자뿐이기 때문에, 보기를 들어 (X U+4E05 丅) (O U+4E0B 下) 이 두 자가 묶음으로 다루어지든 다루어지지 않든 K에는 영향이 없다. K에서는 (O U+4E0B 下) 한자만 도메인에 쓰고, (X U+4E05 丅) 한자는 어차피 도메인에 쓰지 않기 때문이다.

3.5 K 한자 수가 2 이상인 C 이체자 묶음은 반드시 검토해야 하며, 이런 묶음만 검토하면 된다.

K 한자 수가 2 또는 3인 이체자 묶음 3개가 아래에 나와 있다.

2 2 (O U+4E07 万) (O U+842C 萬)
3 2 (O U+5186 円) (X U+5706 圆) (O U+5713 圓)
3 3 (O U+4E30 丰) (O U+8C4A 豊) (O U+8C50 豐)

첫째 보기인 “2 2 (O U+4E07 万) (O U+842C 萬)” 이체자 묶음에 K 한자가 두 자 있으므로 이 한자 두 자를 K에서도 이체자로 보는지 않는지 반드시 확인해야 한다. 참고로, K에서는 이 두 자를 이체자로 보지 않기 때문에 이 두 자를 이체자로 볼지 말지 한국과 중국 사이에 협상을 해야 한다.

둘째 보기인 “3 2 (O U+5186 円) (X U+5706 圆) (O U+5713 圓)” 이체자 묶음에 K 한자가 두 자 있으므로, (O U+5186 円) 한자와 (O U+5713 圓) 한자를 K에서도 이체자로 보는지 않는지 반드시 확인해야 한다. 참고로, K에서는 이 두 자를 이체자로 보지 않기 때문에 이 두 자를 이체자로 볼지 말지 한국과 중국 사이에 협상을 해야 한다.

한편 (X U+5706 圆) 한자는 K에서 쓰지 않는 한자이므로 K에서 이체자인지 아닌지 확인할 수도 없고, 따라서 확인하지 않아도 된다.

셋째 보기인 “3 3 (O U+4E30 丰) (O U+8C4A 豊) (O U+8C50 豐)” 이체자 묶음에 K 한자가 석 자 있으므로, 이 세 자를 K에서도 이체자로 보는지 않는지 확인해야 한다. 참고로, K에서는 이 석 자 각각을 모두 외톨이 한자로 보기 때문에 이 석 자를 이체자로 볼지 말지 한국과 중국 사이에 협상을 해야 한다.

그러면 K 한자 수가 2 이상인 C 이체자 묶음을 K에서도 꼭 같이 이체자 묶음으로 보고 따라서 C 이체자 묶음을 그대로 받아들인 경우를 살펴보자.

2 2 (O 5CF0 峰) (O 5CEF 峯)

위의 이체자 묶음은 C 이체자 묶음이다. 그런데 K에서도 위의 한자 두 자를 이체자로 보고 K 이체자 묶음 37개 가운데 하나로 나온다. 따라서 이 C 이체자 묶음은 K와 C 사이에 꼭 같으므로 아무런 문제가 없으므로, 중국의 이체자 묶음을 그대로 받아들이면 된다.

또 다른 경우를 살펴보자.

4 2 (X 5956 奖) (X 5968 獎) (O 596C 奬) (O 734E 奬)

위의 이체자 묶음은 C 이체자 묶음이다. 그런데 이 C 이체자 묶음에서 K 한자는 두 자뿐인데, K에서도 (O 596C 奬)와 (O 734E 奬)를 이체자로 보고 K 이체자 묶음에 넣었다. 따라서 이 C 이체자 묶음은 K 한자가 아닌 (X 5956 奖) (X 5968 獎)를 무시하면 K와 C 사이에 꼭 같으므로 아무런 문제가 없으므로, 중국의 이체자 묶음을 그대로 받아들이면 된다.

3.6 C 이체자 묶음 3093개 가운데 303개만 검토하면 된다

위에서 살펴보았듯이, C 이체자 묶음 3093개 가운데 K 한자 수가 2이상인 C 이체자 묶음만 검토하면 한중 사이의 C 이체자 검토는 마칠 수 있다는 것을 알게 되었다. Table 4에 보면 K 한자 수가 2이상인 C 이체자 묶음의 개수는 303개임을 알 수 있다.

결론적으로 C 이체자 묶음 3093개를 다 검토할 필요가 없으며, 3093개 가운데 303개만 검토하면 된다. 이는 303 / 3093 = 9.8%로 거의 10분의 1로 줄어든 것이다.

만일 3093개를 검토하는 데 열흘 걸린다면 303개를 검토하는 데는 하루면 될 것이다. 이는 검토 시간을 엄청나게 줄

Table 5. Table Showing the Number of C Variant Groups and Characters for each of the Size of C vAriant Group and the Number of K Characters (Detailed Table)

C 이체자 묶음 크기	C 이체자 묶음 안에 있는 K 한자 수					
	0	1	2	3	C 기준 소계	
1	3166개 (C 3166자 (K 0자	2508개 (C 2508자 (K 2508자	--	--	5674개 (C 5674자 (K 2508자	5674개 (C 5674자 (K 2508자
2	1059개 (C 2118자 (K 0자	1336개 (C 2672자 (K 1336자	161개 (C 322자 (K 322자	--	2556개 (C 5112자 (K 1658자	3093개 (C 6889자 (K 2306자
3	42개 (C 126자 (K 0자	281개 (C 843자 (K 281자	78개 (C 234자 (K 156자	12개 (C 36자 (K 36자	413개 (C 1239자 (K 473자	
4	8개 (C 32자 (K 0자	50개 (C 200자 (K 50자	30개 (C 120자 (K 60자	6개 (C 24자 (K 18자	94개 (C 376자 (K 128자	
5	3개 (C 15자 (K 0자	9개 (C 45자 (K 9자	8개 (C 40자 (K 16자	1개 (C 5자 (K 3자	21개 (C 105자 (K 28자	
6		1개 (C 6자 (K 1자	3개 (C 18자 (K 6자	2개 (C 12자 (K 6자	6개 (C 36자 (K 13자	
7		1개 (C 7자 (K 1자	1개 (C 7자 (K 2자	1개 (C 7자 (K 3자	3개 (C 21자 (K 6자	
K 기준 소계	4278개 (C 5457자 (K 0자	4186개 (C 6281자 (K 4186자	281개 (C 741자 (K 562자	22개 (C 84자 (K 66자	8767개 (C 12563자 (K 4814자	8767개 (C 12563자 (K 4814자
		* K-C 같이 쓰지만 C 이체자 묶음(>=2)에 없는 K 한자 수: 4186자	C 이체자 묶음 (>=2) 3093개 가운데 303개(= 281 + 22), 628자(= 562 + 66)만 검토하면 C 이체자 묶음 검토는 다 될 것으로 보임 [붙임 1]		* K 한자 4819자 가운데 5 (= 4819 - 4814) 자는 C에 없음	

일 수 있다. 한자 전문가에게 이 접근 방식으로 검토하면 이체자 검토가 완전하게 되는지 여러 번 확인하여 맞다는 답변을 받았다. 이체자 검토에 관한 연구를 시작할 때는 전혀 예상하지도, 알 수도 없었던 아주 좋은 성과를 낸 것이다.

3.7 C 이체자 묶음 크기와 그 묶음 안의 K 한자 수별 C 이체자 묶음 개수 (상세 표)

위의 3.2에 나오는 Table 4에서는 C 이체자 묶음 크기와 그 이체자 묶음 안의 K 한자 수별 C 이체자 묶음 개수를 보여주었는데, Table 5에서는 C 이체자 묶음 개수뿐만 아니라 C 이체자 묶음 안에 있는 C 한자 수와 K 한자 수까지 보여준다.

Table 5에서 C 이체자 묶음 크기가 2인 부분만을 떼어낸 것이 Table 6에 나와 있는데 이를 보기로 하여 C 한자 수와 K 한자 수가 계산되는 과정을 살펴보자.

C 이체자 묶음의 크기가 2이므로, K 한자 수와 무관하게 C 한자 수는 늘 이체자 묶음 개수 * 2가 된다. K 한자 수가 0인 칸에서 C 한자 수는 1059 * 2 = 2118자가 된다. 마찬가지로 K 한자 수가 1인 칸에서 C 한자 수는 1336 * 2 = 2672자가 되고, K 한자 수가 2인 칸에서도 C 한자 수는 161 * 2 = 322자가 된다. C 한자 수 합계는 2118 + 2672 + 322 = 5112자인데 이 값이 소계에 나와 있다.

한편 K 한자 수가 0인 경우, C 이체자 묶음 개수가 몇 개이든 무관하게 K 한자 수는 당연히 0이다. 마찬가지로, K 한자 수가 1인 경우 K 한자 수는 1336 * 1 = 1336자이고, K 한자 수가 2인 경우 K 한자 수는 161 * 2 = 322자이다.

Table 6. The Row of Table 5 when the size of C variant group is 2

C 이체자 묶음 크기	C 이체자 묶음 안에 있는 K 한자 수				C 기준 소계
	0	1	2	3	
2	1059개 (C 2118자 (K 0자	1336개 (C 2672자 (K 1336자	161개 (C 322자 (K 322자	--	2556개 (C 5112자 (K 1658자

K 한자 수 합계는 0 + 1336 + 322 = 1658자인데 이 값이 소계에 나와 있다.

C 이체자 묶음 크기가 2라는 것은 이 이체자 묶음 안에 C 한자가 2자라는 것을 말한다. 따라서 K 한자 수는 0, 1, 또는 2만 될 수 있으며 3은 될 수 없으므로 K 한자 수가 3인 열에는 "--"으로 나타내었다.

C 이체자 묶음 크기가 2인 때 C 한자 수와 K 한자 수가 계산되는 과정을 보았는데, 이것만 잘 이해하면 Table 5의 나머지 부분에서 K 한자 수, C 한자 수, 소계 값 등이 계산된 과정을 쉽게 이해할 수 있으리라고 본다.

3.8 C 이체자 묶음 크기와 그 묶음 안의 K 한자 수별 C 이체자 묶음 개수(요약 표)

Table 5를 다음과 같이 바꾼 표가 Table 7에 나와 있다.

- C 이체자 묶음 크기를 1과 2 이상인 두 가지로 나눔

Table 7. Table showing the number of C variant groups and characters for each of the size of C variant group and the number of K characters (Summarized Table)

C 이체자 묶음 크기	C 이체자 묶음 안에 있는 K 한자 수			
	0	1	>= 2	C 기준 소계
1	3166개 (C 3166자 (K 0자	2508개 (C 2508자 (K 2508자	-- -- --	5674개 (C 5674자 (K 2508자
>= 2	1112개 (C 2291자 (K 0자	1678개 (C 3773자 (K 1678자	303개 (C 825자 (K 628자	3093개 (C 6889자 (K 2306자
K 기준 소계	4278개 (C 5457자 (K 0자	4186개 (C 6281자 (K 4186자	303개 (C 825자 (K 628자	8767개 (C 12563자 (K 4814자

- K 이체자 묶음 크기를 0, 1, 2 이상인 세 가지로 나눔

Table 7에 대한 몇 가지 풀이를 살펴보자.

- C 이체자 묶음 크기가 1이고 K 한자 수는 0인 때:
  - 이 한자 3166자는 K는 쓰지 않고 C만 쓰는 C 외톨이 한자임
- C 이체자 묶음 크기가 1이고 K 한자 수는 1인 때:
  - C 이체자 묶음 수는 2508개이고, 그 안의 C 한자 수는 2508자, K 한자 수는 2508자.
  - 이 한자 2508자는 K와 C가 공통으로 쓰며, 더욱이 K와 C에서 모두 외톨이 한자임.
- C 이체자 묶음 크기가 2이고 K 한자 수는 0인 때:
  - C 이체자 묶음 수는 1112개이고, 그 안의 C 한자 수는 2291자, K 한자 수는 0자.
  - 이 C 이체자 묶음 1112개는 C 한자만으로 이루어진 C 이체자 묶음으로 K에 영향을 주지 않기 때문에 K에서는 검토할 필요가 없음
- C 이체자 묶음 크기가 2이고 K 한자 수는 1인 때:
  - C 이체자 묶음 수는 1678개이고, 그 안의 C 한자 수는 3773자, K 한자 수는 1678자.
  - 이 C 이체자 묶음 1678개에는 K 한자가 한 자만 들어있으므로 K에 영향을 주지 않기 때문에 K에서는 이 C 이체자 묶음을 검토할 필요가 없음
- C 이체자 묶음 크기가 2이고 K 한자 수는 >= 2인 때:
  - C 이체자 묶음 수는 303개이고, 그 안의 C 한자 수는 825자, K 한자 수는 628자
  - 이 C 이체자 묶음 303개는 K 한자가 2자 이상 들어있는 C 이체자 묶음으로 K에 영향을 줄 수 있기 때문에 K에서는 자세히 검토할 필요가 있음.

4. K(한국) 이체자 묶음 37개를 검토하는 방안

2015.08.13.일에 발표한 K(한국) 한자 목록 v0.3판에는 한자 4819자가 있으며 이체자 묶음은 37개가 있다[9]. 중국의 한자 목록과 이체자 묶음에 대한 자료가 Table 4에 나와 있는데, 그와 꼭 같은 형식으로 한국의 한자 목록과 이체자 묶음에 대한 자료가 Table 8에 나와 있다.

Table 8. Table showing the number of K variant groups for each of the size of K variant groups and the number of C characters

K 이체자 묶음 크기	K 이체자 묶음 안에 있는 C 한자 수			
	0	1	2	K 기준 소계
1	5개 (K 5자 (C 0자	4740개 (K 4740자 (C 4740자		4745개 (K 4745자 (C 4740자
2	0개 (K 0자 (C 0자	0개 (K 0자 (C 0자	37개 (K 74자 (C 74자	37개 (K 74자 (C 74자
C 기준 소계	5개 (K 5자 (C 0자	4740개 (K 4740자 (C 4740자	37개 (K 74자 (C 74자	4782개 (K 4819자 (C 4814자

Table 8에 대한 몇 가지 풀이를 살펴보자.

- K 이체자 묶음 크기가 1이고 C 한자 수는 0인 때:
  - 이 K 한자 5자는 C는 쓰지 않고 K만 쓰는 K 외톨이 한자임
- K 이체자 묶음 크기가 1이고 C 한자 수는 1인 때:
  - 이 한자 4740자는 C와 K가 공통으로 쓰며, 더욱이 C와 K에서 모두 외톨이 한자임.
- K 이체자 묶음 크기가 2이고 C 한자 수는 2인 때:
  - K 이체자 묶음 수는 37개이고, 그 안의 K 한자 수는 74자, C 한자 수는 74자.
  - 이 K 이체자 묶음 37개는 C 한자가 2자 들어있는 K 이체자 묶음으로 C에 영향을 줄 수 있기 때문에 K에서 자세히 검토할 필요가 있음.

위에서 보듯이 K 이체자 묶음 37개 각각에는 C 한자가 2자씩 들어있으므로, 한중 사이에 이체자 통합 때 문제가 생길 수 있다. 따라서 K에서는 반드시 K 이체자 묶음 37개를 모두 자세히 검토해야 한다.

5. 맺음말과 앞으로 할 일

5.1 맺음말

이 논문에서는 인터넷 최상위 한자 도메인의 국제 생성 규칙(LGR, Label Generation Rule)을 위하여 한국이 검토해야 할 중국 이체자 묶음 선정 방안을 연구하였다. ICANN의 요청으로 이 작업이 2013년부터 시작하여 아마도 2016/2017년쯤에 한자 생성 규칙(LGR)이 마무리될 것으로 보인다. 저자는 한국 쪽 위원장으로 이 작업에 지속적으로 참여하고 있다.

그런데 이체자 목록은 한중일의 이체자 목록을 통합해야 하므로 복잡한 문제가 생긴다. 이 논문에서는 한중일의 이체자 목록이 주어졌을 때 이체자 목록 통합을 위하여 한국이 검토해야 할 중국 이체자 묶음을 선정하는 방안을 찾았으며, 현재 이 선정 방안에 따라 한국 위원회가 작업하고 있다.

처음 중국 이체자 목록 검토를 시작할 때에 가장 큰 고민은 중국은 이체자 묶음이 3093개이고, 한국은 37개이므로 비율이 100 : 12인데 한중 이체자 묶음을 어떻게 통합하느냐 였다. 이

연구에서 제시한 선정 방안에 따라 큰 고민이 해결되었다.

이 연구에서 제시한 선정 방안에 따라 중국 이체자 묶음 3093개 가운데 K 한자가 두 자 이상 들어간 중국 이체자 묶음 303개만 한국에서 검토하면 되기 때문에, 검토 대상 중국 이체자 묶음 수를 10 분의 1로 줄일 수 있었다. 이체자 검토에 관한 연구를 시작할 때는 전혀 예상하지 못했던, 검토해야 할 중국 이체자 묶음 수를 획기적으로 줄일 수 있는 선정 방안을 찾아낸 것이 이 논문의 가장 중요한 기여이다.

5.2 앞으로 할 일

앞으로 몇 가지 할 일이 남아 있다.

1) 한중 사이의 이체자 통합

한중 사이의 이체자 통합은 이제 시작 단계이다. 1차적인 C 이체자 묶음 목록 검토를 마쳤으며 각 이체자 묶음에 대하여 다음 두 가지 방안 가운데 하나를 택하게 된다.

가) K 한자에 국한해서 보면 한국과 중국의 이체자 목록이 같아서 아무런 문제가 없고, 이런 경우 그대로 받아들이면 된다 (acceptable).

나) 한국에서는 다른 한자인데 중국은 이체자로 보는 이체자 묶음에 대해서는 일단 중국에 받아들일 수 없다고 (unacceptable) 1차로 연락하였다.

나-1) 만일 중국이 한국의 의견을 받아들여서 이체자로 보지 않고 외톨이 한자로 바꾸겠다고 하면, 그 이체자 묶음은 더 이상 이체자 묶음이 아니게 되므로 더 이상 아무런 문제가 되지 않는다.

나-2) 그런데 만일 중국도 한국의 의견을 받아들일 수 없다고 답장이 오면, 그 다음에는 한중 사이에 협상을 하여 이체자로 볼지 외톨이 한자로 볼지 결정해야 한다. 이 과정의 결과는 예측할 수 없다.

2) 한국 또는 중국의 이체자 목록이 바뀔 경우

한중 모두 한자 목록과 이체자 목록이 바뀔 가능성이 있다. 특히 중국의 경우 한자 목록을 곧 바꿀 예정이라고 한다. 다만 이체자 목록(묶음)은 바꾸지 않고, 외톨이 한자만 더 넣을 것이라고 한다.

한국의 한국 목록은 크게 바뀌지는 않겠지만 조금 바뀔 가능성은 있다. 또한 한국 이체자 목록은 어느 정도 바뀔 것으로 예상한다.

특히 중국의 이체자 묶음 가운데 일부는 중국이 한국의 의견을 받아들여서 외톨이 한자로 바꾸겠지만, 일부 C 이체자 목록은 한국이 받아들여야 할 것으로 보인다. 한중의 한자 목록이나 이체자 목록이 바뀌면 이체자 검토 작업을 또 해야 할 것이다.

3) 일본의 이체자 목록

아직까지 일본은 이체자가 없다고 하지만, 앞으로 일본이 이체자 목록을 발표할 가능성은 있다. 그렇게 되면 어떻게 해야 할까? 이미 한국과 중국의 이체자 목록을 비교하는 접근 방법이 정립되어 있기 때문에, 꼭 같은 접근 방법으로 한국과 일본의 이체자 목록을 비교하면 될 것이다.

본 논문에서 제안한 방법에 따라 중국 이체자 묶음 303개에 대한 검토가 마무리되면 중국과의 협의를 통해 인터넷 최상위 한자 도메인의 국제 생성 규칙을 정하는 작업이 가속화될 것이다.

References

[1] Guidelines for Developing Script-Specific Label Generation Rules for Integration into the Root Zone LGR [Internet], https://www.icann.org/en/system/files/files/Guidelines-for-LGR-20150424.pdf.

[2] Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels [Internet], https://www.icann.org/en/system/files/files/lgr-procedure-20mar13-en.pdf.

[3] K. Konishi, K. Juang, H. Qian, and Y. Ko. RFC 3743, Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese, and Korean. Apr., 2004.

[4] X. Lee, W. Mao, E. Chen, N. Hsu, and J. Klensin. RFC 4713, Registration and Administration Recommendations for Chinese Domain Names, Oct., 2006.

[5] CDNC Chinese Domain Name Consortium [Internet], http://www.cdnc.org/english/introduction/index.html.

[6] ISO/IEC 10646, Information technology – Universal Coded Character Set (UCS).

[7] The Whole Table of Simplified Characters (簡化字 總表), 1964, China Character Reform Committee, China Ministry of Culture, China Ministry of Education [Internet], http://www.china-language.gov.cn/wenziguifan/managed/002.htm.

[8] CGP MSS 20150430 (Chinese repertoire of 12563 Hanzi characters and variants, 2015.04.30. Meeting document).

[9] K-LGR v0.3, Korean repertoire of Hangeul syllables and Hanja characters and variants. Document number klgp171_4. 2015.08.13.

[10] KIM Kyongsok. Possible errors in C-LGR-1 (2015.04.30.). Document number 165_31. 2015.07.16.

김 경 석

e-mail : gimgs@pnu.kr

1988년 일리노이 주립대학교(어바나-샴페인) 전자계산학(박사)

1988년~1992년 미국 노스다코타 주립대학교 전자계산학과 조교수

1992년~현 재 부산대학교 정보컴퓨터공학부 교수

관심분야 : 데이터베이스, 한글/한말 정보처리, 인터넷 컴퓨팅 등