연구논문 정보화정책 제19권 제3호, 2012년 가을호, pp.50~63

유니코드 한자 검색의 문제점 및 개선방안*

이 정 현**

요 약 본고는 국내 한국학 관련 데이터베이스, 국내 도서관, 국내 학술 데이터베이스, 해외 도서관의 한자 검색 현황을 분석하여 문제점을 파악하고 개선 방안을 도출해 보고자 하였다. 유니코드 환경에서 한자 검색이 문제가 되는 주요한 이유를 '다중코드자', '간체자'와 '이체자'로 정리하고, 각각 3글자를 샘플로 검색하여 현황을 정리하였다. 한국학 관련 데이터베이스 13개, 국내 도서관 데이터베이스 5개, 국내 학술 데이터베이스 4개, 해외 도서관 데이터베이스 2개의 한자검색 현황을 조사하였다. 다중코드자에 대한 검색을 지원하기 위해서는 유니코드 컨소시엄에 공개된 소스를 적용해야 한다. 간체자와 이체자에 대한 검색 기능을 개선하기 위해서는 신뢰할 수 있는 매칭테이블을 표준화하여 유니코드 컨소시엄에 제안해야 한다.

주제어: 유니코드, 한자, 통합한자, 호환한자, 정규화

Problems with Chinese Ideographs Search in Unicode and Solutions to Them

Lee Jeong-hyeon

Abstract This thesis is designed to analyze how the search for Chinese ideographs is done in Koreanology-related domestic databases, domestic library databases, domestic academic databases, and overseas library databases, with a view to identifying problems and suggesting solutions to them. The major reasons that impede Chinese ideographs search in Unicode are classified as 'multicode characters', 'simplified characters', and 'variant characters', and three characters are chosen as samples to describe the current practice. Thirteen Koreanology-related databases, five domestic library databases, five domestic academic databases and two overseas library databases are analyzed in terms of Chinese ideographs search. To support search for multicode characters, the open source of the Unicode consortium must be applied. To improve search for simplified and variant characters, a matching table must be standardized and proposed to the Unicode consortium.

Keywords: unicode, ideographs, CJK unified ideographs, CJK compatibility ideographs, normalization

2012년 7월 17일 접수, 2012년 7월 18일 심사, 2012년 9월 14일 게재확정
* 이 연구는 한국한의학연구원 주요사업 '한의문헌의 창의적 해석을 통한 미래지식보감 구축(K12110)'의 지원을 받았음
** 한국한의학연구원 문헌연구그룹 연구원(han@kiom.re.kr)

Ⅰ. 서론

1990년대 〈조선왕조실록〉의 데이터베이스[1]를 필두로 수많은 한국학 관련 데이터베이스가 생겨났다. 대부분의 데이터베이스는 한국정보화진흥원에서 주도한 국가지식DB 사업의 결과물이며, 그 결과 누구나 손쉽게 데이터를 열람하고 검색할 수 있다. 이는 자료의 보존과 이용의 측면에서 무척이나 고무적인 일임은 분명하다.

그러나 한자 기반의 데이터베이스에서 열람의 경우라면 크게 상관없더라도, 검색의 경우라면 컴퓨터가 모아준 데이터를 검색의 전부라고 믿어서는 곤란하다. 검색엔진의 기본 기능은 입력된 글자를 코드로 인식하여 찾아주는 것이기 때문이다. 예를 들어 '樂(악)'과 '樂(락)'은 모양이 같은 글자지만 '음악(音樂)'을 검색한 결과와 '음락(音樂)'을 검색한 결과는 같지 않다. 컴퓨터의 문자코드는 '樂(악; U+6A02)'과 '樂(락; U+F95C)'을 다르다고 인식하기 때문이다. 이처럼 데이터베이스의 한자검색에 관한 문제는 바로 인간과 컴퓨터 간의 인식 차이에서 시작된다.

데이터베이스는 논리적으로 연관된 데이터들의 구조화된 집합인데(Hoffer, 2010: 7), 한자의 모양은 같지만 코드가 다를 경우 논리적 연결이 이루어질 수 없다. 이 말은 결국 인간이 같다고 인식하는 글자들이 코드가 다르다면 논리적 연관이 없기 때문에 컴퓨터에서 동일하게 인식할 수 없고, 검색에서 제외될 수밖에 없다는 의미이다. 물론 이러한 문제를 해결해주는 검색엔진의 기능은 이미 개발되었다. 본고의 목적은 새로운 기능을 제시하려는 것이 아니라, 실제 일어나고 있는 문제점을 분석하여 개선 방향을 제시하려는 것이다.

본고는 우선 유니코드 한자의 개괄과 코드 간의 관계를 살펴보고, 한자자료가 많은 한국학 관련 데이터베이스에서 일어나는 한자검색의 문제를 실제로 알아본 후, 한자검색과 관련 있는 국내의 도서관 데이터베이스와 학술 데이터베이스, 그리고 추가적으로 해외 도서관데이터베이스에서 한자검색 결과를 살펴보겠다. 또한 한자검색의 문제가 생기는 원인을 정리하여 개선 방향을 제시하도록 할 것이다.

Ⅱ. 유니코드 한자

유니코드는 1990년대 초 Apple, IBM, Microsoft 등 The Unicode Consortium이 문자코드 문제를 근본적으로 해결하고자 하는 취지에서 제안한 국제표준코드이다. 유니코드는 1991년 v.1.0이 발표된 이후로 1996년 v.2.0을 거쳐 2012년 v.6.1이 발표되었다(Unicode Consortium, 연도불명a).

유니코드에서 한자 관련 코드는 한중일 통합한자(CJK[2] Unified Ideographs; 이하 통합한자로 약칭), 한중일 호환한자(CJK Compatibility Ideographs; 이하 호환한자로 약칭), 한중일 부수(CJK Radicals/KangXi Radicals; 이하 부수로 약칭) 3가지 파트로 구성되어 있다. 통합한자는 다시 확장 A~D(CJK Extension-A~D) 구역이, 호환한자는 호환한자 보충(CJK Compatibility Ideographs Supplement)이, 부수는 한중일 부수 보충(CJK Radicals Supplement)과 한중일 자획(CJK Strokes)이 추가된다. 유니코드 중 한자 부분을 정리하면 〈표 1〉과 같다.

현재 유니코드를 지원하는 폰트은 대체로 통합한자, 확장A한자, 호환한자 등을 지원하며, 우리나라의 무료 폰트 중 '함초롬바탕'은 〈표 1〉에서 음영부분의 한자를 지원한다. '함초롬바탕'에는 총 64,138개의 글리프(Glyph)가 있고 한자는 통합한자 20,924자, 확장A한자 6,582자, 호환한자 중 467자(통합한자로 추후 지정된 12글자 포함) 등 총 28,304자가 포함되어 있다.

호환한자는 한중일 코드별로 분류된다. 구성을 살

1) 국사편찬위원회(연도불명). "조선왕조실록 소개." http://sillok.history.go.kr/intro/intro_info.jsp. (검색일: 2012.06.25).
2) Chinese Japanese Korean의 약칭

〈표 1〉 유니코드 한자

Class		Range _Start	Range _End	Range _Numbers	Assign _Start	Assign _End	Assign _Numbers	
CJK Unified Ideographs		4E00	9FCF	20944	4E00	9FBB#	20924#	(+12)*
	CJK Extension-A	3400	4DBF	6592	3400	4DB5#	6582#	
	CJK Extension-B	20000	2A6DF	42720	20000	2A6DF	42720	
	CJK Extension-C	2A700	2B73F	4160	2A700	2B73F	4160	
	CJK Extension-D	2B740	2B81F	224	2B740	2B81F	224	
CJK Compatibility Ideographs		F900	FAFF	512	F900	FAD9#	467#	
	CJK Compatibility Ideographs Supplement	2F800	2FA1F	544	2F800	2FA1F	544	
CJK Radicals / KangXi Radicals		2F00	2FDF	224	2F00	2FD5#	214#	
	CJK Radicals Supplement	2E80	2EFF	128	2E80	2EFF	128	
	CJK Strokes	31C0	31EF	48	31C0	31EF	48	
	Ideographic Description Characters	2FF0	2FFF	16	2FF0	2FFF	16	

* : CJK Unified Ideographs 추후지정 된 12글자(CJK Compatibility Ideographs에 배정되어 있음)
# : '함초롬바탕' 지원글자 기준

펴보면 통합한자에 배정된 글자 중에서 음가를 여러 개 갖는 글자가 대부분이고, 추후에 지정된 통합한자 12글자가 포함되었다. 호환한자를 정리하면 〈표 2〉와 같다.

호환한자는 통합한자와 모양은 같고 음가가 다르기 때문에 각각 통합한자의 코드가 매핑되어 있다. 예를 들어 호환한자의 즐거울낙(樂; U+F914), 즐거울락(樂; U+F95C), 좋아할요(樂; U+F9BF)는 모두 통합한자의 풍류악(樂; U+6A02)에 매핑되어 있다. 호환한자는 매핑 정보가 포함되어 있기 때문에 기계적으로 통합한자로 변환할 수 있고, 소스도 제공되고 있다(Unicode Consortium, 연도불명b).

다음 장에서는 검색에서 유니코드 한자를 사용할 때 나타날 수 있는 문제점과 현황을 살펴보도록 하겠다.

〈표 2〉 호환한자의 구성

Block	Assign _Start	Assign _End	Range _Numbers	notes	
Pronunciation variants from KS X 1001:1998	F900	FA0B	268	한국	
Duplicate characters from Big 5	FA0C	FA0D	2	중국	
The IBM 32 compatibility ideographs	FA0E	FA2D	32	IBM	12글자는 통합한자 (추후지정)
Korean compatibility ideographs	FA2E	FA2F	2	한국	
JIS X 0213 compatibility ideographs	FA30	FA6A	59	일본	
ARIB compatibility ideographs	FA6B	FA6D	3	일본	
DPRK compatibility ideographs	FA70	FAD9	106	북한	
Total			472*		

* : 배정된 한자는 472개지만 폰트마다 지원하는 숫자가 다르다. 〈표 1〉의 호환한자 467개는 '함초롬바탕' 폰트가 지원하는 한자의 개수이다.

Ⅲ. 한자검색의 문제점과 현황

1. 유니코드 환경에서 한자검색의 문제점

1) 다중코드자

호환한자 중에서 통합한자로 매핑된 한자는 모양은 같지만 음(音)이 다르거나 다수의 훈음(訓音)을 가진 한자로 이하 '다중코드자'라고 하기로 한다. 즉 다중코드자는 호환한자의 영역 중에서 통합한자로 추후 지정된 12글자를 뺀 460개의 한자를 말한다. 다중코드자에 대한 매핑소스는 위에서 언급한 것처럼 유니코드 컨소시엄에서 제공하고 있지만 실제 모든 검색엔진이나 프로그램에 적용되어 있지 않은 것이 현실이다. 다중코드자에 대한 검색을 지원하지 않을 경우 '여자(女子)'와 '녀자(女子)'처럼 한자의 모양은 같지만 음이 다른 경우 검색결과가 다르게 나오게 된다. 실제 일어나고 있는 검색의 문제점은 아래의 현황 부분에서 자세히 살펴보도록 하겠다.

2) 간체자와 이체자

간체자는 복잡한 한자의 획수를 간단하게 변형시켜 현대 중국에서 쓰고 있는 글자이다. 간체자가 문제가 되는 이유는 사용자가 간체자를 포함하여 검색어를 입력하거나, 데이터베이스에 간체자 일부가 입력되어 있기 때문이다.

이체자는 정자와 동일한 음과 뜻을 가지면서 정자와 자형이 유사한 글자로 고자, 와자, 통자, 약자, 이형자 등이 있다(이규옥, 2005: 534). 이체자는 대부분 텍스트 입력 당시 원본이미지와 최대한 같은 글자를 입력한다는 지침[3]때문에 데이터베이스 내에 존재하게 된다. 하지만 이런 지침을 근거로 입력된 텍스트는 대표자-이체자 매핑정보[4]가 없으면 전혀 검색할 수 없다.

실제 일어나고 있는 검색의 문제점은 아래의 현황 부분에서 자세히 살펴보도록 하겠다.

2. 데이터베이스 한자검색 현황

위에서 살펴본 한자검색의 문제점이 실제 사용되고 있는 데이터베이스에서 어떻게 나타나는지 살펴보도록 하겠다.

우선 간체자, 이체자, 다중코드자에서 각각 3글자씩 샘플로 설정하였다.[5] 각 데이터베이스에서 간체자 3글자(医, 与, 为), 이체자 3글자(毉, 與, 爲; 世, 仙, 來)를 검색어로 넣었을 경우와 대표자(醫, 與, 爲; 世, 仙, 來)로 검색했을 경우 결과의 값이 같다면 O, 같지 않다면 X로 표시하였다. 다중코드자는 '龜(균)', '樂(요)', '北(배)'로 검색했을 경우와 '龜(구)', '樂(악)', '北(북)'으로 검색했을 경우 결과의 값이 같다면 O, 같지 않다면 X로 표시하였다.

1글자 검색을 지원하지 않는 데이터베이스에서는 의원(醫員; 医員), 세계(世界; 世界) 등 검색어를 포함하는 단어로 확장하여 검색하였다.

1) 한국학 관련 데이터베이스

국내의 데이터베이스 중에서 한자로 된 자료가 많은 분야는 한국학분야이다. 많은 연구자가 한국학 관련 데이터베이스를 검색하여 자료를 정리하고 연구를 진행하고 있지만, 한자로 검색했을 경우 누락된 자료가 있다는 사실을 인지하고 있는 사람은 거의 없

3) 한국고전번역원, 국사편찬위원회 등 한국학 관련 데이터베이스 구축 사업의 제안요청서에는 대부분 Unicode 3.0을 제시하고 있다. 통상 Unicode 3.0은 CJK Extension-A까지를 의미한다.

4) 국가에서 지원한 한자의 매핑정보에 관한 연구는 과거 몇차례가 있었다. 하지만 연구 결과는 출력된 형태로 제공될 뿐이고 활용 가능한 파일형태로는 제공되지는 않는다(김홍규, 2002).

5) 글자의 선정 기준은 다음과 같다.
- 간체자 : 현대 간행물과 고문헌 모두 일반적으로 흔히 사용되며, 대표자와 형태의 차이가 뚜렷한 글자
- 이체자 : 고문헌에 자주 등장하며, 대표자와 형태의 차이가 뚜렷한 글자
- 다중코드자 : 음가의 종류가 가장 많은 글자와 두음법칙과 무관하게 다수의 음가를 갖는 글자

〈표 3〉 한국학 관련 데이터베이스의 간체자, 이체자, 다중코드자 검색 현황

사이트	기관	간체자			이체자			다중코드자					
		医	与	为	吾	僊	来	龜(균)	龜(구)	樂(요)	樂(악)	北(배)	北(북)
		U+533B	U+4E0E	U+4E3A	U+534B	U+50CA	U+6765	U+F908	U+9F9C	U+F9BF	U+6A02	U+F963	U+5317
국가지식포털	한국정보화진흥원	X	X	X	X	X	X	X		X		X	
유교넷*	한국국학진흥원	X	X	X	X	X	X	X		X		X	
호남기록문화시스템#	전북대학교	X	X	X	X	X	X	O		O		O	
한국역사정보통합시스템	국사편찬위원회	X	O	O	O	O	O	O		O		O	
조선왕조실록	국사편찬위원회	X	O	O	O	X	O	O		O		O	
승정원일기	국사편찬위원회	X	O	O	O	X	O	O		O		O	
장서각	한국학중앙연구원	X	O	X	X	X	O	O		O		O	
한국경학자료시스템	성균관대학교	X	O	X	X	X	O	O		O		O	
남명학고문헌시스템	경상대학교 도서관	X	O	X	O	X	O	O		O		O	
한국고전종합DB	한국고전번역원	X	O	X	O	X	O	O		O		O	
한의고전명저총서	한국한의학연구원	O	O	O	O	O	O	X		X		X	
규장각	서울대학교	O	O	O	O	O	O	X		X		X	
한국독립운동사정보시스템	독립기념관	O	O	O	O	O	O	O		O		O	

* : 같은 음의 한글까지 검색됨 # : 간체자, 이체자 지원하지 않음

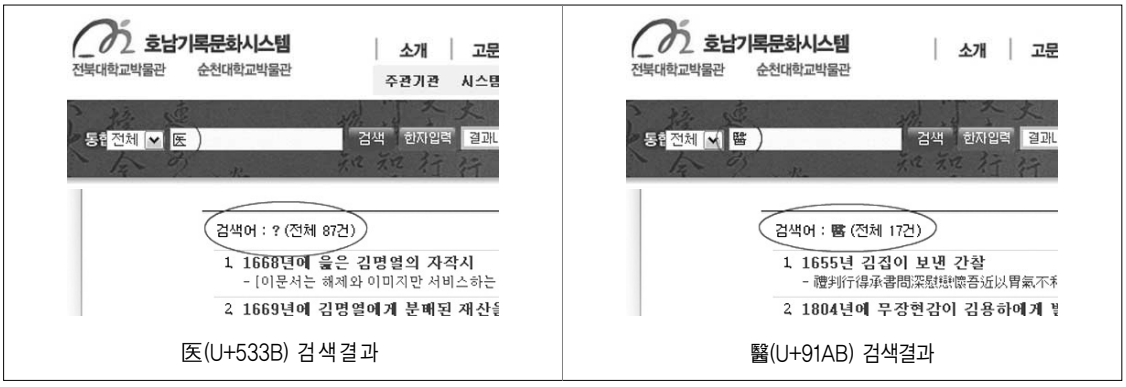
다. 또한 데이터베이스마다 검색되는 한자를 따로 기억하여 검색하기는 사실 불가능하다. 한국학 관련 데이터베이스 중 한국정보화진흥원의 국가지식포털(http://www.knowledge.go.kr/), 한국한의학연구원의 한의고전명저총서(http://jisik.kiom.re.kr/), 한국고전번역원의 한국고전종합DB(http://db.itkc.or.kr/), 서울대학교의 규장각(e-kyujanggak.snu.ac.kr/) 등 13종에 대해서 간체자, 이체자, 다중코드자를 검색어로 검색한 결과는 〈표 3〉과 같다.

국가지식포털은 대부분의 한국학DB로 연결되어 있지만 한자 검색에 있어서는 의외로 낮은 결과를 보여주었고, 유교넷은 같은 음의 한글까지 검색되어 원하는 자료를 찾기에 어려움이 있었다. 국가지식포털과 유교넷은 한국학DB와 수많은 콘텐츠를 제공하고 있으며, 연구자보다는 일반 사용자를 주요 대상으로 삼았기 때문에 이와 같은 결과가 나온 것으로 생각된다. 그러나 소스가 공개되어 있는 다중코드자에 대한 검색부분은 모든 데이터베이스에서 지원해야 할 기능일 것이다.

전북대학교의 호남기록문화시스템은 다중코드자에 대하여 같은 결과를 보여주지만 검색창에서 간체자나 이체자를 입력했을 경우 검색어가 '?'로 처리되며 글자에 상관없이 동일한 검색결과가 출력되는 에러가 나타났다. 검색 결과가 다르게 나오는 것보다 앞서 해결해야 할 문제이다(〈그림 1〉 참조).

한국역사정보통합시스템, 조선왕조실록DB, 승정원일기DB는 한 기관에서 서비스되고 있지만 결과는 약간 다르게 나왔으며, 일부 간체자와 이체자에 대해서는 참조검색이 되지 않고 있었다(〈그림 2〉 참조). 한국역사정보통합시스템은 조선왕조실록, 승정원일기와 성격이 다르기 때문에 관리도 별도로 되고 있다고 생각된다. 다만 한국역사정보통합시스템은 대부분의 한국학 데이터베이스로 연결되는 관문이므로 연계사이트에서 지원하는 글자에 대해서 재검토가 있으면 검색률이 향상될 수 있을 것으로 생각된다.

장서각, 한국경학자료시스템, 남명학고문헌시스템, 한국고전종합DB는 일부 간체자와 이체자에 대해서는 검색결과가 다르게 나타났고, 다중코드자에 대



医(U+533B) 검색결과 | 醫(U+91AB) 검색결과

〈그림 1〉 호남기록문화시스템의 간체자 검색 결과

医(U+533B) 검색결과 | 醫(U+91AB) 검색결과

〈그림 2〉 한국역사정보통합시스템의 간체자 검색 결과

医(U+533B) 검색결과 | 醫(U+91AB) 검색결과

〈그림 3〉 한국고전종합DB의 간체자 검색 결과

해서는 모두 동일한 결과를 나타냈다(〈그림 3〉 참조).

한의고전명저총서 DB와 규장각 DB에서는 간체자와 이체자에 대해 검색결과가 같게 나왔으나 다중코드자에 대해서는 참조검색이 되지 않고 있었다.

독립기념관의 한국독립운동사정보시스템에서는 간체자, 이체자, 다중코드자에 대해 모두 동일한 결과를 나타내었으며, 샘플이라는 한계는 있지만 양호한 검색결과를 나타냈다.

다중코드자의 검색결과가 같은 9개의 데이터베이스는 검색엔진에서 유니코드 정규화 알고리즘을 거친 결과이다. 유니코드 정규화에 대한 소스는 호환한자 지정 당시부터 제공되었으므로 결과가 같지 않은

4개의 데이터베이스에서는 수정이 필요하다고 보여진다.

흥미로운 사실은 위의 데이터베이스들이 대부분 같은 검색엔진을 사용하고 있다는 점이다(한국과학기술정보연구원, 연도불명). 대부분의 한국학 관련 데이터베이스에서 활용하고 있는 한국과학기술정보연구원(KISTI)의 크리스탈(KRISTAL)이 검색엔진으로 사용되었음에도 불구하고 위처럼 다른 결과가 나오는 까닭은 각 기관에서 검색엔진에서 기본으로 제공하는 이체자 테이블을 추가·삭제할 수 있기 때문이다(윤종웅, 2005: 557). 하지만 지금처럼 각 데이터베이스마다 다른 결과가 출력된다면 사용자에게 혼란을 주게 된다. 사용자가 원하는 자료를 빠뜨리지 않고 정확히 찾기 위해서는 국가수준에서 전문가들의 합의된 통일안이 절실히 요구된다. 또한 합의된 통일안은 공개하여 한자와 관계된 데이터베이스에서 쉽게 반영할 수 있도록 해야 한다.

2) 국내 도서관 데이터베이스

국내 도서관 내에도 많은 고문헌과 한자로 된 자료가 존재한다. 국내 도서관 데이터베이스 중 국립중앙도서관(http://www.nl.go.kr), 국회도서관(http://www.nanet.go.kr/), 고려대학교 도서관(http://library.korea.ac.kr/), 연세대학교 도서관(http://library.yonsei.ac.kr/), 계명대학교 도서관(http://library.kmu.ac.kr/) 5개 기관에 대하여 간체자, 이체자, 다중코드자를 검색어로 검색한 결과는 〈표 4〉와 같다.

대부분의 도서관은 간체자, 이체자, 다중코드자에 대한 검색을 지원하지 않는다. 한자로 된 자료뿐만 아니라 양적으로 훨씬 많은 한글, 영어 등의 자료를 주요 검색 대상으로 하기 때문으로 생각된다. 같은 음의 한글이나 한자까지 검색되는 이유도 아마 이와 같은 이유일 것이다.

국립중앙도서관의 소장자료 중에서는 271,586권(2012년 5월 31일 현재)의 고서가 있다(국립중앙도서관, 연도불명). 이는 전체 자료의 비율(약 3%)로는 얼마 되지 않을지 몰라도 단일 소장처로는 적지 않은 양이며, 국보 · 보물 등 가치 높은 책들도 포함되어 있다. 예를 들어 '洪武禮制'를 검색할 경우 '洪武禮(예; U+FB96)制'나 '홍무례제'로 검색할 경우 검색되지 않고, '洪武禮(례; U+79AE)制'나 '홍무예제'로 검색해야만 검색되는 문제가 있다(〈그림 4〉 참조).

국회도서관에서 '儀禮註疎'를 검색할 경우 '儀禮(례; U+79AE)註疎'나 '의례주소'로 검색할 경우 검색되지만, '儀禮(예; U+FB96)註疎'나 '의예주소'로 검색할 경우 검색되지 않는다(〈그림 5〉 참조).

국립중앙도서관의 '洪武禮制'를 검색한 경우와 국회도서관에서 '儀禮註疎'를 검색한 경우를 살펴보면 데이터 구축 당시 '禮(예; U+FB96)'로 입력하는 것과 '禮(례; U+79AE)'로 입력하는 것의 차이가 검색에 그대로 반영되어 있다. 다중코드자에 대한 유

〈표 4〉 도서관 데이터베이스의 간체자, 이체자, 다중코드자 검색 현황

사이트	간체자			이체자			다중코드자					
	医	与	为	𠚕	僊	来	龜(균)	龜(구)	樂(요)	樂(악)	北(배)	北(북)
	U+533B	U+4E0E	U+4E3A	U+534B	U+50CA	U+6765	U+F908	U+9F9C	U+F9BF	U+6A02	U+F963	U+5317
국립중앙도서관*	×	×	×	×	×	×	×		×		×	
국회도서관#	×	×	×	×	○	×	×		×		×	
고려대학교 도서관	×	×	×	×	×	×	×		×		×	
연세대학교 도서관#	×	×	×	×	×	×	×		×		×	
계명대학교 도서관	×	×	×	×	×	×	×		×		×	

* : 같은 음의 한글까지 검색됨 # : 같은 음의 한글, 한자까지 검색됨

洪武禮(예; U+FB96)制

홍무례제

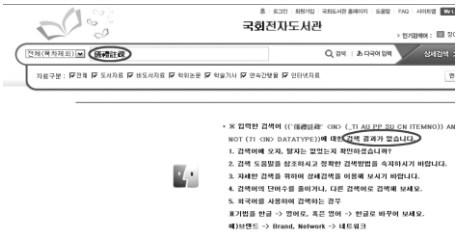
검색되지 않음

洪武禮(례; U+79AE)制

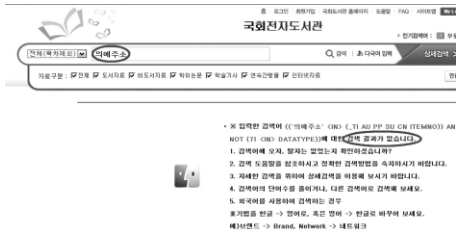
홍무예제

검색 됨

〈그림 4〉 국립중앙도서관의 다중코드자 및 한글 검색 결과

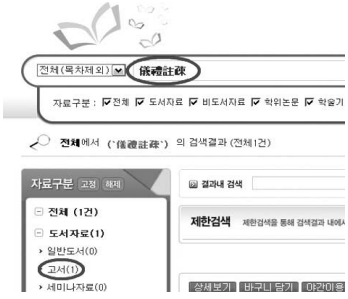


儀禮(예; U+FB96)註疏

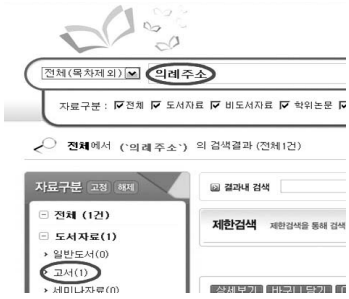


의예주소

검색되지 않음



儀禮(례; U+79AE)註疏



의례주소

검색 됨

〈그림 5〉 국회도서관의 다중코드자 및 한글 검색 결과

니코드 정규화 과정을 거친다면 '禮(예; U+FB96)'와 '禮(례; U+79AE)'에 상관없이 검색결과가 동일할 것이다.

고려대학교 도서관, 연세대학교 도서관, 계명대학교 도서관에는 각각 104,991책(12. 2. 29일자), 101,716책(11. 2. 28일자), 73,171책(12. 6월자)의 고서가 소장되어 있어 단일 기관으로는 상당히 많은 고전적 자료가 소장되어 있고, 이에 대한 검색도 지원하고 있다. 하지만 대부분 한글과 영어 검색에 주안점을 두고 있어 한자를 검색하는 기능은 한글 검색에 준하는 수준이다. 현재 한자의 음으로 참조검색하는 알고리즘에 간체자, 이체자에 대한 부분을 당장 추가할 수 없다면, 최소한 소스가 공개되어 있는 다중코드자에 대한 참조검색을 추가해야 한다. 다른 도서관도 같은 상황이므로 한자로 된 자료의 검색을 위해서 반드시 개선되어야 할 기능이라 하겠다.

3) 국내 학술 데이터베이스

국내 학술 데이터베이스 내에도 한자로 된 자료가 상당 수 존재한다. 국내 학술 데이터베이스 중 학술연구정보서비스(http://www.riss.kr/), 국가과학기술정보센터(http://www.ndsl.kr), 디비피아(http://www.dbpia.co.kr/), 한국학술정보(http://kiss.kstudy.com/) 4개 기관에 대하여 간체자, 이체자, 다중코드자를 검색어로 검색한 결과는 〈표 5〉와 같다.

대부분의 기관은 간체, 이체자에 대한 참조검색을 지원하지 않고 있으며, 국가과학기술정보센터와 디비피아에서는 다중코드에 대한 참조검색을 지원하고 있다(〈그림 6, 7〉 참조). 학술 데이터베이스는 고문헌에 대한 검색보다는 학위논문, 학술지, 단행본, 특허, 보고서 등의 검색을 위해 만들어졌기 때문에 간체자, 이체자에 앞서 다중코드자에 대한 부분만 지원해도 검색결과가 크게 향상될 수 있다.

〈표 5〉 학술 데이터베이스의 간체자, 이체자, 다중코드자 검색 현황

사이트	간체자			이체자			다중코드자					
	医	与	为	吾	僊	来	龜(균)	龜(구)	樂(요)	樂(악)	北(배)	北(북)
	U+533B	U+4E0E	U+4E3A	U+534B	U+50CA	U+6765	U+F908	U+9F9C	U+F9BF	U+6A02	U+F963	U+5317
학술연구정보서비스(RISS)*	X	X	X	X	X	X	X		X		X	
국가과학기술정보센터(NDSL)	X	X	X	X	X	X	O		O		O	
디비피아(DBpia)	X	O	X	X	X	X	O		O		O	
한국학술정보(KISS)*	X	X	X	X	X	X	X		X		X	

* : 같은 음의 한글까지 검색됨

균(龜;U+F908)열(裂;U+F9A0) 검색결과 구(龜;U+9F9C)렬(裂;U+88CE) 검색결과

〈그림 6〉 학술연구정보서비스의 다중코드자 검색 결과

균(龜;U+F908)열(裂;U+F9A0) 검색결과　　구(龜;U+9F9C)렬(裂;U+88CE) 검색결과

〈그림 7〉 국가과학기술정보센터의 다중코드자 검색 결과

〈그림 8〉 학술연구정보서비스의 검색 소외 자료 예시

학술연구정보서비스에서 '균열(龜裂)'과 '귀렬(龜裂)'로 검색한 결과 결과값이 다르게 나타난다. 위에서 살펴 본 것처럼 데이터베이스 구축 당시 입력한 코드대로 검색되는 현상을 볼 수 있는데, 이는 일부 자료가 검색에서 소외될 수 있는 결과를 가져온다.

예를 들어, 학술연구정보서비스에서 '연강의 열처리 온도와 두께 변화에 따른 피로귀렬성장거동에 관한 연구'라는 국내학술지 논문의 원문 이미지는 한자로 '균열(龜裂)'이라고 되어있지만 데이터베이스를 구축하면서 한글로 '귀렬(龜裂)'이라고 잘못 입력하여 '균열'이라는 검색어에는 전혀 검색되지 않는 문제가 있다(〈그림 8〉 참조).

학술연구정보서비스와 한국학술정보를 비롯해 다중코드자에 대한 참조검색을 지원하지 않는 학술 데이터베이스에서는 한자를 잘못 입력한 자료들이 검색에서 소외될 수 있다는 문제를 인식하고 기능을 개선해야 할 것이다.

4) 해외 도서관 데이터베이스

해외 도서관 데이터베이스 내에도 한자로 된 자료가 상당 수 존재한다. 해외 도서관 데이터베이스 중 일본의 東京大**学** 附**属図**書館(https://opac.dl.itc.u-tokyo.ac.jp/), 미국의 Harvard Library(http://lib.harvard.edu/) 2개 기관에 대하여 간체자, 이체자, 다중코드자를 검색어로 검색한 결과는 〈표 6〉과 같다.

한자로 된 자료는 국내 이외에도 여러 국가에 존재한다. 해외의 기관 중 일본의 東京大**学** 附**属図**書館은 간체자, 이체자, 다중코드자에 대하여 모두 결과가 같으며, 미국의 Harvard Library는 일부의 간체자와 이체자 검색을 지원하고 있으며 다중코드자에 대한 검색도 지원하고 있음을 알 수 있다(〈그림 9〉 참조).

해외의 데이터베이스 중 지극히 적은 두 곳의 결과이지만, 이 결과가 시사하는 점은 결코 가볍지 않다. 같은 한자문화권인 일본 東京大**学** 附**属図**書館은 샘플로 조사한 모든 글자에 대해 참조검색을 지원하고 있고, 영어문화권인 미국에서도 상당부분 글자에 대해 참조검색을 지원하고 있다. 기본적으로 다중코드자에 대한 참조검색을 지원하는 것은 두말할 나위가 없다. 국내외 연구자들이 웹에서 한자기반 데이터베이스를 탐색할 때 각각의 데이터베이스에서 지원하

〈표 6〉 해외 데이터베이스의 간체자, 이체자, 다중코드자 검색 현황

사이트	간체자			이체자			다중코드자					
	医	与	为	吞	僊	来	龜(균)	龜(구)	樂(요)	樂(악)	北(배)	北(북)
	U+533B	U+4E0E	U+4E3A	U+534B	U+50CA	U+6765	U+F908	U+9F9C	U+F9BF	U+6A02	U+F963	U+5317
(일본)東京大学 附属図書館	O	O	O	O	O	O	O		O		O	
(미국)Harvard Library	O	O	X	X	O	O	O		O		O	

〈그림 9〉 Harvard Library의 간체자 검색 결과

는 한자검색에 대한 정보가 없다면 원하는 검색결과를 얻지 못하거나 누락된 검색결과를 전체라고 착각할 소지가 있다.

Ⅳ. 개선방안

지금까지 살펴본 데이터베이스에서 한자 검색에 대한 문제점을 해결하는 가장 근본적인 방법은 신뢰할 수 있는 한자 매핑테이블을 만들어 표준화한 후 누구나 이용, 활용할 수 있도록 공개하는 것이다. 현재 검색엔진에서 다중코드자, 간체자, 이체자 검색을 지원하는 기능은 이미 존재한다(윤종웅, 2005: 557). 즉 기능적인 개선이 필요한 것이 아니라 내용적인 개선이 필요하다는 의미이다. 내용적인 개선은 각각의 한자에 대한 코드가 포함된 매핑테이블을 만드는 것이다. 매핑테이블은 검색엔진에서 참조검색의 소스로 활용하여 검색의 정확성을 높일 수 있으며, 정자로 입력하는 성격의 데이터베이스[6]의 경우 입력자가 이체자로 잘못 입력하더라도 기계적으로 모두 정자로 치환하는 프로그램의 소스로 활용할 수 있다.

현재 파일로 공개되어 있는 이체자 매핑자료는 한국역사정보통합시스템(http://www.koreanhistory.or.kr) 자료실에 있는 '한자이체자사전'이 유일하다. 각고의 노력으로 만든 파일을 공개했다는 점은 아주 높이 평가될 만하지만 이 파일 안에는 '검토요', '확인요', '폰트' 등 최종 완성본이 아니라는 것을 추측할 수 있는 내용이 포함되어 있으며, 한자의 코드는 전혀 포함되어 있지 않기 때문에 신뢰할 만한 공식자료로 사용하기에는 사실 어려움이 있다(〈그림 10〉 참조).

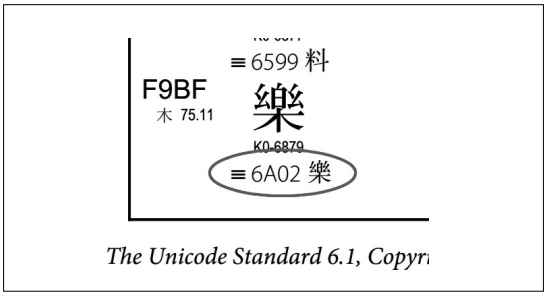
유니코드 호환한자의 매핑자료[7]는 〈그림 11〉, 〈그림 12〉와 같이 정확히 규정되어 있으며, 현재 여러 사이트나 프로그램에서 적용되고 있다. 국가에서 주도하

6) 데이터베이스를 구축하는데 있어서 가장 기초적인 문제는 대표자로 입력하는 것과 이체자를 살려서 입력하는 것의 문제이다. 지식정보화사업 초기에는 원래 자형을 살려주는 방향에서 정자로 입력하는 방향으로 선회하게 된다(윤종웅, 2005).

7) Unicode Consortium(연도불명b). "Publications and Data-Technical Reports-Unicode Standard Annex #15(UNICODE NORMALIZATION FORMS)-9 Detecting Normalization Forms." http://www.unicode.org/reports/tr15/; http://www.unicode.org/Public/UNIDATA/DerivedNormalizationProps.txt. (검색일: 2012.06.25).

연번	대표	부수	총획	이체1	구분	이체2	구분	통용자	비고1
27	于	二	3	亏		亏	고		확인요
28	五	二	4	乂*	고			伍*	음이 예일 때 별자
29	井	二	4	井	동				
39	享	亠	8	亯*	고			饗	내용으로 구분
40	亹	亠	22	亹	혹			娓*/𠔃**	힘쓰다의 뜻일 때 통용
45	仙	人	5	仚	속	仙*	별		가볍게 날다의 뜻일 때 별자
46	付	人	5	仅*	付동			衬	음이 근일 때 별자

〈그림 10〉 한국역사정보통합시스템의 한자이체자사전 일부



〈그림 11〉 호환한자의 매핑정보

〈그림 12〉 호환한자의 매핑정보 소스

연번	분류	대표_값	대표_코드	참조_값	참조_코드
1	이체자 등	弋	5F0C	一	4E00
2	이체자 등	亓	4E93	丌	4E0C
3	이체자 등	弍	5F0E	三	4E09
4	이체자 등	丄	4E04	上	4E0A
5	이체자 등	丅	4E05	下	4E0B
3986	다중코드	裂	88C2	裂	F9A0
3987	다중코드	說	8AAA	說	F9A1
3988	다중코드	說	F96F	說	F9A1
4254	간체	万	4E07	萬	842C
4255	간체	与	4E0E	與	8207
4257	간체	专	4E13	專	5C08
4258	간체	业	4E1A	業	696D
4259	간체	丛	4E1B	叢	53E2

〈그림 13〉 매핑정보 테이블의 예

여 이체자 연구로 끝낼 것이 아니라 한중일 유관기관이 협력하여 유니코드 컨소시엄에 제안할 정도의 공신력을 갖춘 표준 매핑테이블을 만들어야 한다. 그래야만 국가, 프로그램, 웹사이트에 관계 없이 활용할 수 있으며 한자 자료의 검색도 향상된 결과를 얻을 수 있을 것이다. 매핑테이블에는 최소한 속성(분류), 한자값, 유니코드값이 반드시 포함되어야 할 것이다(〈그림 13〉 참조).

기능적인 측면에서 생각해보면, 매핑테이블에 의한 참조검색을 할 경우 자칫 검색결과가 너무 많아질 수도 있다. 매핑테이블을 통하여 같은 의미나 같은 모양을 가진 한자를 검색할 수도 있지만, 반대로 일치검색과 같이 특정 한자만을 검색할 수도 있어야 한다. 즉 즐거울락(樂; U+F95C)을 검색하는 경우 매핑된 한자인 풍류악(樂; U+6A02), 좋아할요(樂; U+F9BF) 중에서 사용자가 원하는 한자를 선택하게 할 수도 있다면 한층 정확한 검색이 가능할 것이다.

또한 매핑테이블을 단순 소스로 활용하는 것에 그치지 않고 한자입력기와 같은 도구(Tool) 형태로 제공하고, 로그를 기록하여 지속적으로 발전 가능한 방향으로 사용된다면 한자 검색이 더욱 편리해질 것이다.

Ⅴ. 결론

본고는 한국학 관련 데이터베이스를 중심으로 국내 도서관, 학술 데이터베이스, 해외 도서관의 한자

검색 현황을 분석하여 문제점을 파악하고 개선 방안을 도출해 보고자 하였다.

먼저 유니코드 한자에 대한 범위, 수 등을 살펴보고, 통합한자와 호환한자의 관계 및 호환한자의 구성에 대하여 정리해 보았으며, 호환한자의 매핑 정보가 제공되고 있음을 살펴보았다. 이어서 유니코드 환경에서 한자 검색이 문제가 되는 주요한 이유 2가지를 제시하였다. 모양은 같지만 음이 다르거나 다수의 훈음을 가진 '다중코드자'는 매핑정보가 공개되어 있지만 검색엔진에서 적용하지 않는 경우가 있고, 현대 중국에서 쓰고 있는 '간체자'와 정자와 같은 음과 뜻을 가지면서 자형이 유사한 '이체자'는 매핑정보가 없거나 빈약하여 검색에 장애가 되는 경우가 있다.

이러한 문제점들이 실제 데이터베이스에서 검색 오류를 일으키게 되는데, 그 유형에 따라 4가지로 분류하여 살펴보았다. 간체자 3글자(医, 与, 为), 이체자 3글자(吞, 僊, 来), 다중코드자 3글자 (龜, 樂, 北)를 샘플로 설정하여 검색할 때 정자(대표자)와의 검색 결과가 같다면 O, 같지 않다면 X로 표시하였다.

한국학 관련 데이터베이스 13개를 대상으로 한자검색 현황을 살펴본 결과 소스가 공개되어 있는 다중코드자에 대한 검색을 지원하지 않는 경우도 4개의 데이터베이스가 있었다. 대부분의 사이트에서 간체자와 이체자에 대한 검색을 부분적으로 지원하고 있었다. 국가지식포털과 유교넷에서는 간체자, 이체자, 다중코드자에 대한 검색을 전혀 지원하지 않았고, 한국독립운동사정보시스템에서는 모두 지원하고 있었다.

국내 도서관 데이터베이스 5개를 대상으로 한자검색 현황을 살펴본 결과 5개 기관 모두 소스가 공개되어 있는 다중코드자에 대한 검색을 지원하지 않았고, 간체자와 이체자에 대한 검색도 대부분 지원하지 않았다.

국내 학술 데이터베이스 4개를 대상으로 한자검색 현황을 살펴본 결과 2개의 데이터베이스에서는 다중코드를 지원했으며 나머지 2개의 데이터베이스에서는 지원하지 않았다. 간체자와 이체자에 대해서는 4개의 데이터베이스 모두 지원하지 않았다.

해외 도서관 데이터베이스 2개를 대상으로 한자검색 현황을 살펴본 결과 2개 기관 모두 다중코드자에 대한 검색을 지원했으며, 간체자와 이체자에 대해서 대부분 지원하고 있었다.

현황조사 결과 소스가 공개된 다중코드자에 대한 검색을 지원하지 않는 데이터베이스도 상당히 많았다. 한자로 되어 있는 데이터가 있다면 반드시 지원해야하는 기능이며, 이 부분만 개선하더라도 어느 정도의 검색 기능이 향상될 수 있다.

간체자와 이체자에 대한 검색 기능을 개선하기 위해서는 유니코드 컨소시엄에서 제공하고 있는 호환한자-통합한자의 매핑정보와 같이 신뢰할 수 있는 매칭테이블을 표준화하는 것이다. 과거 비슷한 연구나 사업이 있었지만 코드가 빠져있거나 오류가 종종 발견되어 실제 공식자료로 활용하기에는 어려움이 있다. 매핑테이블은 한중일 유관기관이 협력하여 유니코드 컨소시엄에 제안해야 국가나 검색엔진에 상관없이 누구나 이용할 수 있을 것이다.

지금까지 유니코드 한자 검색의 문제점 및 개선 방안을 알아보았다. 유니코드를 아무리 정확하게 지정하고 매핑테이블을 만들어도 폰트 없이는 인간이 인식할 수 없다. 현재 유니코드를 지원하는 폰트마다 범위도 다르고 오류도 많이 발견된다. 유니코드 한자 매핑테이블의 작성과 동시에 폰트에 대한 연구도 동시에 진행되어야만 완전한 형태의 결과가 나올 수 있을 것이다.

■ 참고문헌

국립중앙도서관 (연도불명). "자료현황-소장자료." http://www.nl.go.kr/nl/havdata/havdataShow.jsp. (검색일: 2012.06.25).

국사편찬위원회 (연도불명). "조선왕조실록 소개." http://sillok.history.go.kr/intro/intro_info.jsp. (검색일 : 2012.06.25).

김홍규 (2002).「다국어 정보 처리를 위한 유니코드 (V3.0) 한자의 이체자 연구」. 서울: 고려대학교.
윤종웅 (2005).「고전 전산화 사업의 이체자 처리 현황-고전 전산화를 위한 이체자 인코딩 및 DB 구축 방안」. 서울: 고려대학교 민족문화연구원.
이규옥 (2005).「한문 고전 전산화에 있어서 이체자 처리 방안-고전 전산화를 위한 이체자 인코딩 및 DB 구축 방안」. 서울: 고려대학교 민족문화연구원.
한국과학기술정보연구원 (연도불명). "활용사이트." http://www.kristalinfo.com/kristal_sites.php. (검색일: 2012.06.25).
Hoffer, A. Jeffrey 저·서우종 역 (2010).「현대 데이터베이스 관리론」. 고양시: 사이텍미디어.
Unicode Consortium (연도불명a). "유니코드 컨소시엄." http://www.unicode.org. (검색일: 2012.06.25).
Unicode Consortium (연도불명b). "Publications and Data - Technical Reports - Unicode Standard Annex #15(UNICODE NORMALIZATION FORMS) - 9 Detecting Normalization Forms." http://www.unicode.org/reports/tr15/; http://www.unicode.org/Public/UNIDATA/DerivedNormalizationProps.txt. (Retrieved on June 25, 2012).