漢字結構情報를 이용한 漢字檢索 시스템 연구

申相賢*

• 목 차 •

Ⅰ. 문제제기

유니코드는 전 세계의 주요 언어를 포괄하고 있는 국제 표준 문자코드 체계로[1], 1995년에 제정된 이후 각종 운영 시스템(OS, Operating System)과 응용 프로그램에서 사용되고 있다. 그런데 유니코드 Ver.5.0을 기준으로 할 때, 전체 문자 가운데 약 87% 가량을 한자가 차지하고 있으며[2], 국제 표준 문자코드를 관할하는 ISO/IEC의 JTC1/SC2/WG2 산하 IRG[3]에서 '한중일통합한자 확장 C'와 D의 표준화가 진행 중에 있어 유니코드의 한자는 계속 늘어날 전망이다.[4]

이 가운데 전체 한자의 절반 이상을 차지하고 있는 것이 Ext.B 한자로 모두 42,711자에 달한다. 물론 이 한자는 실생활에서 많이 사용하지 않을 뿐만 아니라, 이를 구현하기 위한 프로그램 설계상의 문제와 시스템 자원의 처리 문제 등이 간단치 않은 것은 사실이다. 그렇지만 지난 2000년부터 시작된 한국역사정보통합시스템 구축사업의 고전 자료 전산화과정에서 약 17,000여 자에 이르는 신출한자가 발견되었는데, 이 가운데 절반 이상이 Ext.B 한자에 해당하였다.[5] 이와 같은 상황임에도 불구하고 현재 국내외적으로 Ext.B 한자를 효율적으로 처리할 수 있는 소프트웨어는 그리 많지 않은 실정이다.[6]

이렇게 된 데에는 여러 가지 이유가 있겠지만, 대표적으로 Ext.B 한자에 대

* 고려대 민족문화연구원 연구원 / partizan21@naver.com
** 이 연구에 참여한 연구자는 '2단계 BK21 고려대학교 한국어문학교육연구단'의 지원비를 받았음.

1) 유니코드는 'ISO/IEC 10646 國際符號化文字集合(Universal Multiple-Octet Coded Character Set)'에 수록된 문자를 바탕으로 전 세계의 언어를 컴퓨터에서 일괄적으로 처리하기 위한 국제 표준 문자코드이자 산업표준이다. 여기에는 전 세계 약 26개 주요 언어의 문자와 특수기호에 대해 코드값을 부여하고 있으며, 하나의 문자판(Plane)이 최대로 수용할 수 있는 문자수는 65,536자이다. 문자판은 기본다언어판인 BMP(Basic Multilingual Plane)와 Surrogate 확장에 의한 16개의 확장문자판으로 나뉘는데, Plane1을 보충다언어판(SMP, Supplementary Multilingual Plane)이라 하며, 한자전용 영역인 Plane2는 보충표의문자판(SIP, Supplementary Ideographic Plane)이라고 한다. Plane3부터 Plane13까지는 미사용 영역이며, Plane14를 특별보충판(SPP, Supplementary Special-purpose Plane)이라 하고, Plane15와 Plane16은 개인사용 영역(Private Use Area)으로 되어 있다.(The Unicode Consortium, "The Unicode Standard, Version 5.0", Addison-Wesley Professional, 2006과 유니코드 콘소시엄 홈페이지 "http://www.unicode.org/" 참조.)

2) 유니코드 Ver.5.0을 기준으로 현재 등록된 한자를 살펴보면 다음의 표와 같다.

문자판 영역	유니코드 영역	등록 자수
BMP(Plane0)	한중일통합한자	20,902자
	한중일통합한자 확장 A(Ext.A)	6,582자
SIP(Plane2)	한중일통합한자 확장 B(Ext.B)	42,711자
	한중일통합한자 확장 C(Ext.C)	4,219자(예정)
합 계		74,474자

이외에도 유니코드에는 '한중일 호환용 한자' 467자, '강희자전 부수' 214자, '한중일 부수 보충한자' 116자, '한중일 호환용 보충한자' 542자가 별도로 등록되어 있다.(The Unicode Consortium, *"The Unicode Standard, Version 5.0"*, Addison-Wesley Professional, 2006과 http://www.unicode.org/ 참조.)

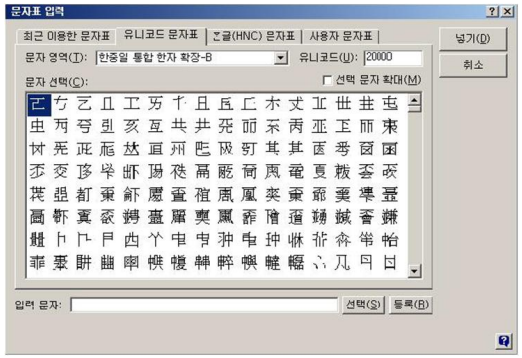
3) IRG는 'Ideographic Rapporteur Group'의 약자로 ISO/IEC 10646, 즉 유니코드에 등록할 한자를 선정하는 기초 작업을 담당하고 있으며, 이 작업에는 매년 2회에 걸쳐 한자문화권 국가인 한국·중국·대만·일본·베트남·북한 싱가폴 마카오 홍콩 등이 참여하고 있다.(IRG 홈페이지 "http://www.cse.cuhk.edu.hk/~irg/" 참조.)

4) '한중일통합한자 확장 C(Ext.C)'의 한자 4,219자는 2006년 11월 27일부터 12월 1일까지 중국 하이난[海南]에서 개최된 제27차 IRG 회의에서 국제 표준 규격에 새로 추가하기로 결정되었으며, 현재 ISO/IEC 10646에 추가하기 위해 JTC1/SC2에서 등록을 위해 관련 절차를 밟고 있다. 그리고 '한중일통합한자 확장 D(Ext.D)'를 결정하기 위해 각국으로부터 약 15,000여 자의 한자를 제안 받아 작업을 진행하고 있다.(IRG 홈페이지 "http://www.cse.cuhk.edu.hk/~irg/" 참조.)

5) 고려대학교 민족문화연구원·(주)솔트웍스 컨소시엄, 『'지식정보자원관리사업 신출한자 통합관리를 위한 목록 및 지침 작성' 과제 결과보고서』, 한국전산원, 2004년 12월, 8-9쪽 참조.

6) 현재 국내에서 출시된 소프트웨어 가운데 한글과 컴퓨터사의 '훈글 2004' 이후 버전과 마이크로소프트사의 'MS Office 2003' 버전부터 Ext.B 영역의 한자를 처리할 수 있다.

한 字音과 字義 정보 데이터베이스의 부재를 들 수 있다. 즉 자음과 자의 정보 데이터베이스를 바탕으로 대부분의 응용프로그램에서 '한글 자음에 의한 검색' 또는 '부수와 획수에 의한 검색'을 사용할 수 있지만,[7] Ext.B 한자에 대해서는 이러한 지원이 미비하였다. 그리고 Ext.B 한자는 주로 한자문화권 각국에서 인명과 지명으로 사용되던 한자와 기존 한자의 각종 異體字를 많이 포함하고 있는데, 특히 인명과 지명으로 사용되던 한자는 그 자음과 자의를 정의할 수 없는 것이 많기 때문에 검색과 입력에서 많은 어려움을 수반할 수밖에 없다. 이러한 까닭으로 현재 국내의 주요 워드프로세스에서 Ext.B 한자를 사용하기 위해서는 다음 그림에서 보는 바와 같이 문자표를 통해 해당 한자를 직접 선택하거나, 유니코드값을 이용하여 입력하는 방법밖에는 없다.



【'글 2005'의 문자표 입력창】
① 메뉴에서 '입력→문자표' 선택
② 문자표 입력창에서 '유니코드 문자표' 탭 선택
③ 유니코드 문자표 탭에서 '한중일 통합한자 확장-B' 선택
④ 해당 한자 입력

7) 우리나라는 1993년 유니코드의 한중일통합한자에 17,184자를 '한국측 제안한자'로 등록하였는데, 이것은 뒤에 국가표준 'KS C 5700-1995'(뒤에 KS X 1005-1로 개정, 2006년 8월 31일 폐지)로 수용되었다. 현재 대부분의 응용 프로그램에서는 KS C 5700에 수록된 한자에 대해서만 정확한 字音과 字義를 지원하고 있는데, 이것은 1998년에 문화관광부에서 KS C 5700의 한자에 대한 각종 정보를 데이터베이스화 하였기 때문이다.(서경호 등, 『國際 文字 코드 提案 漢字의 標準化에 대한 硏究(상·하권)』, 문화관광부, 1998.)

【'MS-Office'의 문자표 입력창】
① 메뉴에서 '삽입→기호' 선택
② 글꼴 항목에서 Ext.B를 지원하는 폰트 선택
③ 하위집합 항목에서 '한중일 통합한자 확장 B' 선택
④ 해당 한자 입력

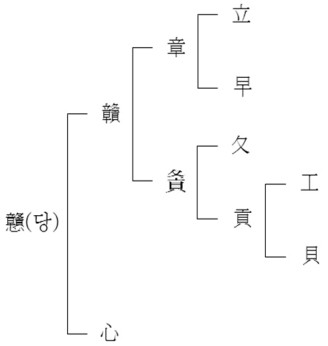
이상에서 살펴보았듯이 각종 고전 자료의 전산화과정에서 Ext.B 한자를 사용해야할 필요성이 대두되고 있음에도 불구하고 현재의 검색 및 입력 방법으로는 그 한계를 가질 수밖에 없다. 즉 한자의 자음과 자의에 의존하는 방법으로는 다량의 한자를 빠른 시간 내에 검색할 수 없을 뿐만 아니라 俗字·略字·簡體字 등 다양한 異體字와 각국 고유의 한자가 존재하고 있으며, 언어의 차이에 의해 같은 자형의 한자라도 그 의미가 전혀 달라지는 경우도 있다. 이와 같은 복잡한 현상으로 인하여 어떤 한자는 그 자음과 자의를 정의할 수 없어 '한글 자음과 자의에 의한 검색 및 입력 방법' 자체를 사용할 수 없게 만들기도 한다. 따라서 한자의 자음과 자의에 의존하지 않고서도 효율적으로 한자를 검색할 수 있고, 검색한 한자를 입력할 수 있는 방법을 강구할 필요가 있는 것이다.

이 논문은 이러한 문제의식을 바탕으로 漢字結構와 한자의 分解와 組合에 대한 이론을 응용하여 새로운 한자검색 방법을 모색한 것이다. 한자는 한자를 구성하는 結構에 의하여 특정한 개수의 部件으로 분해할 수 있으며, 분해된 부건은 이미 알고 있는 특정 한자[부수자 포함]일 수도 있고, 의미를 가지지 않는 하나의 필획일 수도 있다. 그리고 이렇게 분해된 부건을 결구와 함께 구조적으로 데이터베이스화하며, 이 데이터베이스를 이용하여 자음과 자의를 모르는 한자를 하나 이상의 분해된 부건의 조합으로 편리하게 검색할 수 있을 것이다.

이를 위해 우선 현대 한자학에서 제시하고 있는 한자의 분해와 조합, 그리고 결구에 대해서 살펴보고, 분해된 부건을 결구와 함께 구조적으로 데이터베이스한 '한자결구정보 데이터베이스'의 구성 방법과 검색시스템의 세부적인 요소와 역할에 대해 살펴보도록 하겠다.

Ⅱ. 漢字結構와 漢字의 分解 및 組合

중국의 현대 漢字學 가운데 漢字 構形學 이론에 의하면, 한자는 六書의 원리에 의해 字形을 구성하고, 모든 한자는 그 結構에 의해 獨體字[文]와 合體字[字]로 분류할 수 있다. 그리고 합체자는 일정한 개수의 部件으로 분해할 수 있으며, 분해된 부건은 다시 일정한 규칙인 '漢字結構'에 의하여 조합할 수 있음을 제시하고 있다. 여기서 독체자는 '一·丨·丶·丿·乀·㇀·亅'와 같이 더 이상 의미를 가지지 않는 최소 단위인 筆劃으로 구성되며, 육서 가운데 象形字와 指事字가 대부분으로 전체 한자의 3-4%에 지나지 않는다. 이와 다르게 합체자는 어원학적으로 의미를 가지는 최소 단위인 부건으로 구성되고, 부건은 '字根·字元·字素'라고도 하며, 육서 가운데 象形字와 指事字를 제외한 대부분의 한자가 合體字에 속한다.[8]



【'戇'자의 분해 예시】

8) 蘇培成 著, 李圭甲 譯, 『現代漢字學』, 學古房, 2007, 107-147쪽 참조.

그런데 한자를 분해하는 원칙과 어느 정도 수준까지 분해한 것을 부건으로 할 것인가, 그리고 조합 규칙인 한자결구를 몇 가지 유형으로 할 것인가는 연구자들 간에 많은 차이를 보인다. 이 가운데 일부 학자는 한자의 구조를 파괴할 정도로 분해하여 한자교육을 혼란시키기도 하는데, 이러한 폐해를 막기 위해 중국의 國家言語文字工作委員會는 1997년 12월에 "信息處理用GB13000.1字符集漢字部件規範"을 제정하기에 이른다. 이 규범은 현재 중국에서 한자 정보처리에서 규범적인 역할을 하고 있는데, 유니코드 BMP 영역의 한중일통합한자 20,902자의 한자를 분해한 '漢字基礎部件表'와 사용 규칙을 규정하고 있다.[9]

이렇게 분해된 한자의 자형은 일정한 규칙에 의해 구조화하여 다시 조합할 수 있으며, 이 규칙을 '漢字結構'라고 한다. 한자결구의 종류는 연구자들 간에 약간의 차이는 있으나 대체로 다음 표에서 보는 것과 같이 13가지의 결구가 주를 이루고 있으며, 일정한 개수로 분해된 합체자의 부건은 이 13가지 결구에 의해 모두 조합할 수 있다는 것이다.[10]

【漢字 構形學에서 정의하고 있는 13가지 조합 규칙】

좌우 구조	(1)좌우	形·波·保·貌·組·點·推 등
	(2)좌중우	班·辯·衢·街·術 등
상하 구조	(3)상하	是·貿·毒·藥·霜·買·要 등
	(4)상중하	曼·率·衰·愛·哀·器·葬 등
포위 구조	(5)전체포위	國·固·回·困·囚·圍 등
	(6)위쪽 삼면 포위	問·向·同·用·風 등
	(7)좌측 삼면 포위	匡·匣·匿·區·匠·臣·医 등
	(8)아래쪽 삼면 포위	凶·幽·函·畫·凹 등
	(9)위쪽과 좌측 포위	厄·壓·病·尿·居·君 등
	(10)위쪽과 우측 포위	句·匀·可·氣·匈 등
	(11)아래쪽과 좌측 포위	延·道·翹·勉·翅·德 등
	(12)아래쪽과 우측 포위	斗·頭 등
걸침 구조	(13)걸침 구조	承·乖·乘·夾·爽·坐·爽 등

이와 같은 한자의 분해와 조합에 대한 이론은 ISO/IEC 10646과 유니코드에 IDS(Ideographic Description Sequence)란 개념으로 수용되어 표준화되었다. IDS란 '유니코드에 등록된 한자에 대해 해당 한자를 구성하는 자형을 구조적으로 분

9) 상게서, 129-132쪽 참조.
10) 상게서, 145쪽 참조 재인용.

해하여 기술하는 원칙'을 말하며, '漢字結構情報' 또는 '漢字構造情報'라고 번역하고 있다.[11]

IDS는 結構에 의해 모든 한자를 일정한 개수의 部件으로 분해하며, 이 분해와 조합을 일종의 식으로 나타낸 것으로 한자, 합체자를 분해하고 조합한다는 기본 원칙은 기존의 이론과 대체로 동일하다. 다른 점이 있다면 현대 한자학에서는 13가지의 조합 규칙을 제시하고 있으나, IDS에서는 다음 표와 같이 12가지로 정의하고 있다는 것이다. 그리고 이 12가지 유형을 다시 특정한 형태의 기호로 표기하고 있는데, 이를 '結構形態'라고 한다.[12] 이처럼 IDS에서 조합 규칙인 결구를 부호화함으로써 분해된 부건을 구조적으로 데이터베이스화할 수 있으며, 이 데이터베이스로부터 찾고자 하는 한자의 일부 자형으로 해당 한자를 매칭하여 검색할 수 있는 기초를 제공하고 있다.

【유니코드 IDS에서 정의한 12가지 결구】

번호	결구형태	설명	글자 예
1	⿰	좌우 결구	校 = ⿰木交
2	⿱	상하 결구	忠 = ⿱中心
3	⿲	좌중우 결구	瀞 = ⿲氵青爭
4	⿳	상중하 결구	蓋 = ⿳艹去皿
5	⿴	완전 포위 결구	回 = ⿴口口
6	⿵	상삼면 포위 결구	問 = ⿵門口
7	⿶	하삼면 포위 결구	凷 = ⿶土口
8	⿷	좌삼면 포위 결구	医 = ⿷匚矢
9	⿸	좌상면 포위 결구	病 = ⿸疒丙
A	⿹	우상면 포위 결구	匈 = ⿹凶勹
B	⿺	좌하면 포위 결구	迄 = ⿺乞辶
C	⿻	겹침 결구	爽 = ⿻大玨

11) The Unicode Consortium, *"The Unicode Standard Version 4.0"*, Addison-Wesley Professional, 2003, pp.307-309과 'ISO/IEC 10646:2003 부속서 F'에서 IDS를 정의하고 그 문법을 기술하고 있다. 또 IRG 문서 N1154(Algorithm for Identifying the Duplicate Ideograph Characters by the IDS)에서 알고리즘 정의하고 있다.

12) 결구형태란 유니코드에서 정의하고 있는 'Ideographic Description characters'를 번역한 것으로 결구를 부호화한 것을 말한다.(The Unicode Consortium, *"The Unicode Standard Version 4.0"*, Addison-Wesley Professional, 2003, pp.307-309 참조.)

Ⅲ. 漢字結構情報 데이터베이스

앞에서 살펴본 것처럼 한자를 분해한 部件과 부호화된 조합 규칙인 結構形態를 이용하여 모든 한자를 구조적으로 데이터베이스화할 수 있는데, 이를 '漢字結構情報 데이터베이스'로 정의할 수 있다. 그리고 이 데이터베이스로부터 찾고자 하는 한자의 일부 자형인 부건으로 해당 한자를 매칭하여 검색할 수 있다.

그런데 문제는 중국의 "漢字部件規範" 제정 과정에서도 제기되었듯이 한자를 부건으로 분해하는 과정에서 어느 정도 수준까지 분해하느냐 하는 것이다. 한자는 俗字·略字·簡體字 등 다양한 異體字가 존재하고, 한자문화권에 속해 있는 각국마다 그 실정에 맞추어 새로 造字한 고유한자도 존재하고 있으며, 무엇보다도 사용하는 언어의 차이에 의해 같은 자형의 한자라도 그 의미가 전혀 달라지는 경우도 있다. 따라서 같은 의미의 한자를 서로 다르게 분해할 수 있는 미묘한 차이가 발생할 수 있는 것이다.

현재 한국에서 일반적으로 한자를 분해한다고 할 경우, 대부분의 사람들은 結構를 적용하여 분해하는 것이 아니라 지극히 개인적인 지식의 차원에서 임의로 행하는 것이 대부분일 것이다. 그러므로 어떤 경우에는 자음과 자의를 가지지 않는 자소 단위인 필획의 수준까지 과도하게 분해하여 키보드상에서 원활하게 입력할 수 없게 되고 만다. 물론 중국이나 대만의 경우에는 한자를 분해한 필획을 이용하여 입력이나 검색을 구현한 소프트웨어가 없는 것은 아니지만,[13] 이러한 방식은 우리 어문생활과 맞지 않을뿐더러 이를 사용하기 위해서는 사용자가 새로 학습을 해야 한다는 어려움을 가지고 있다.

따라서 본고에서는 이와 같은 상황을 고려하여 한자결구정보 데이터베이스를 구축하는 데에 있어서 다음과 같은 원칙을 적용하였다.

【한자결구정보 데이터베이스 구축 원칙】

① 데이터베이스 구축 범위는 유니코드에 등록된 한자 74,474자로 한다.

② 모든 한자를 부수자[변형된 部首字 포함]를 포함하여 字音과 字義를 가지는 하

13) 중국의 입력법은 크게 拼音에 의한 입력법과 필획의 조합에 의한 입력법으로 나눌 수 있는데, 필획에 의한 입력법 가운데 대표적인 것이 五筆字型法이라 할 수 있다. 이 입력법은 영문 자판 Z를 제외한 25개의 자판을 橫[一], 竪[丨], 撇[丿], 捺[㇏], 折[𠃌]의 5개 필획을 기준으로 5개의 區로 나누고, 각 자판에 130개의 자근을 각각 배열하여 이들 자근의 조합에 의해 원하는 한자를 입력할 수 있도록 하였다.(五筆字型輸入法 홈페이지 http://www.hongen.com/pc/newer/ime/wbx/imejswb1.htm 참조.)

나 이상 N개의 部件으로 분해한다.

③ 부건은 기본 운영체계에서 한글 음가로 입력 가능한 유니코드 Ext.A 영역의 한자까지로 제한한다.

④ '⺗[心], 巛[川], ㄴ[乙], ⺧[牛]' 등과 같이 변형된 部首字인 경우에만 별도의 입력창을 통해 입력한다.

⑤ 한자를 조합하기 위한 결구는 유니코드 IDS에서 정의한 12가지 종류의 결구를 사용한다.

데이터베이스 구축 원칙에서 보듯이 한자결구정보 데이터베이스 구축에서 가장 중요한 부분은 모든 한자를 부수자[변형된 부수자 포함]를 포함하여 字素 단위가 아닌, 字音과 字義를 가지는 복수 개의 부건으로 분해하는 것이다. 물론 변형된 부수자의 경우에는 어쩔 수 없이 별도의 입력창을 제공해야 하겠지만, 기본적으로 분해된 모든 부건은 한글 자음으로 입력할 수 있어야 하며, 유니코드 IDS에서 정의한 12가지 결구에 의해 원래의 한자로 조합할 수 있어야 한다.

이와 같은 원칙에 의해 구축된 한자결구정보 데이터베이스에는 검색된 한자의 자음이나 자의와 같은 각종 정보를 한 눈에 볼 수 있도록 다음의 정보를 함께 데이터베이스화 하였다. 이 가운데에는 일반 사용자에는 불필요한 정보도 있으나 향후 연구 목적이나 전자사전 개발 등에 도움을 줄 수 있을 것이다.

① 자형(Glyph) : 기본 자형으로 'SuperCJK 코드북 Ver14.0'의 자형을 이미지화하여 데이터베이스화

② 유니코드 코드값(Ucode) : 16진수로 구성된 유니코드값

③ 주요 자전 위치정보 : 康熙字典(KangXi)과 漢語大字典(Hydz) 위치정보 수록

④ 각국 제안 코드 정보 : 중국(Gcode), 일본(Jcode), 대만(Tcode), 한국(Kcode), 베트남(Vcode), 홍콩(Hcode), 북한(Kpcode) 등 유니코드에 한자를 제안한 국가의 최초 제안 각국 코드 정보

⑤ 잔여획수(Strok) : 부수자를 제외한 나머지 획수

⑥ 부수자(Radical) : 康熙字典 부수자 자형과 순서

⑦ 四角號碼(FourCorner) : 사각호마 검색법에서 사용하는 5자리의 사각호마값 수록

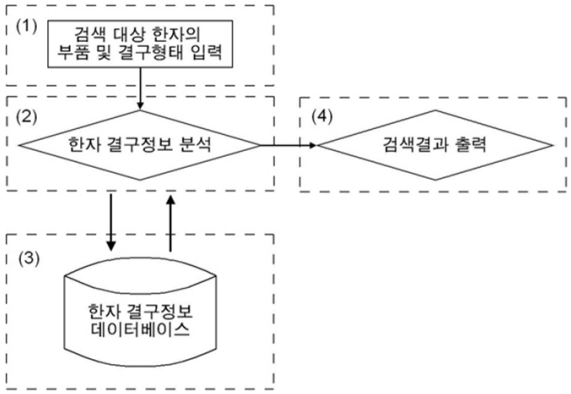
⑧ 한글 字音(Pronun)과 字義(Meaning) : 국내에서 간행된 교학사의 대한한사전, 명문당의 한한대자전, 삼성당의 한한대사전의 대표 한글 字音과 字義 수록

【한자결구정보 데이터베이스 구축 예】

영역	Glyph	Ucode	KangXi	Hydz	Gcode	Jcode	Tcode	Kcode	Vcode	Hcode	Kpcode	Strok	Radical	FourCorner	Pronun	Meaning	IDS
한중일 통합 한자	一	04E00	0075.010	10001.010	G0-523B	J0-306C	T1-4421	K0-6C69	V1-4A21		KP0-FCD6	0	一	10000	일	하나_같다(일)	一
	丁	04E01	0075.030	10002.050	G0-3621	J0-437A	T1-4423	K0-6F4B	V1-4A22		KP0-E8B9	1	一	10200	정	고무래_정간(정)	丁
	丂	04E02	0075.050	10004.010	G5-3021	J1-3021	T4-2126					1	一	10207	교	교묘하다(교)	丂
	七	04E03	0075.070	10003.030	G0-465F	J0-3C37	T1-4424	K0-7652	V1-4A23		KP0-EFA6	1	一	40710	칠	일곱(칠)	七
	丄	04E04	0076.010	10003.010	GE-2121	J1-3022	T3-2126					1	一	20100	상	위(상)	丄[上]
	丅	04E05	0076.020	10002.040	GE-2122	J1-3023	T3-2125					1	一	10200	하	아래(하)	丅[下]
	丆	04E06	0076.021	10003.021	G1-7D3D			K2-2121				1	一	10200			⿱一丿
	万	04E07	0076.030	10009.030	G0-4D72	J0-4B7C	T2-2126	K0-5832	V1-4A24		KP0-DAB9	2	一	10227	만	일만(만)	万
	<중간 생략>																
확장 A영역 한자	㐀	03400	0078.010	10015.030	G_KX	JA-2121	T6-222C					4	一	11100	구	언덕(구)	⿱丌一
	㐁	03401	0078.030	10019.020	G5-3024		T4-2224	K3-2121				5	一	10600	첨	핥다(첨)	㐁[⿱一四]
	㐂	03402	0078.101			JA-2122						5	一	40714			⿱七⿰七七
	㐃	03403	0079.021					K3-2122				2	丨	60200			㐃
	<중간 생략>																
확장 B영역 한자	𪛑	2A6D1	1538.380	74808.020	G_KX		T7-6349					9	龠		취	【대】(취)龡와 同字【명】(취)吹의 本字	⿰⿱火⿰彡儿龠
	𪛒	2A6D2	1538.390	32168.210	G_KX		T5-7B3B				KP1-9293	10	龠				⿰⿱竹龠⿱殳⿰几又
	𪛓	2A6D3	1538.410	74808.060	G_KX		T4-6E47					11	龠		록;각	【동】(각!록)龣(2220)과 同字【대】(록)龣과 同字【명】(록)龣과 同字【한】同龣 <集韻>龣, 東方音,或從鹿	⿰龠鹿
	𪛔	2A6D4	1538.420	42832.180	G_KX		T7-6625					14	龠		치;지	【교】(지)피리【동】(치)篪(2220)의 俗字【대】(치)篪의 俗字【명】(지)篪의 俗字【한】同篪 <正字通>, 俗篪字	⿰⿱竹龠⿱虎⿰乇儿
	𪛕	2A6D5	1538.430	74808.070	G_KX		T4-6E56					17	龠		소	【교】(소)籥(18927)와 同字【동】(소)籥(1339)와 同字【대】(소)籥의 古字【명】(소)퉁소	⿰龠龜
	𪛖	2A6D6	1538.440	74808.080	G_KX		T7-664C					21	龠		소	【교】(소)籥(18927)와 同字【동】(소)籥(1339)와 同字【대】(소)籥와 同字【명】(소)퉁소	⿰龠⿱龜灬

Ⅳ. 漢字檢索 시스템 구성

이상과 같은 방식으로 구성되는 한자결구정보 데이터베이스는 다음에서 도시한 한자검색 시스템 구성 개략도와 같이 한자검색 시스템을 구축하고, 시스템 내부에서 일련의 검색 진행과정을 거쳐 검색하고자 하는 한자를 검색할 수 있다. 먼저 시스템 개략도를 중심으로 그 구성을 살펴보면, (1)검색하고자 하는 한자를 구성하는 최소한 하나 이상의 부건과 결구형태를 입력하는 부분, (2)입력 요청된 결구정보를 분석하는 부분, (3)한자를 구성하는 부건을 결구형태와 함께 구조화한 한자결구정보 데이터베이스 부분, (4)한자결구정보 데이터베이스로부터 검색 결과를 출력하는 부분으로 구성된다.



【한자검색 시스템 구성 개략도】

위의 시스템 구성 개략도에서 (1)~(4) 부분을 다음과 같이 좀 더 상세하게 설명할 수 있겠다.

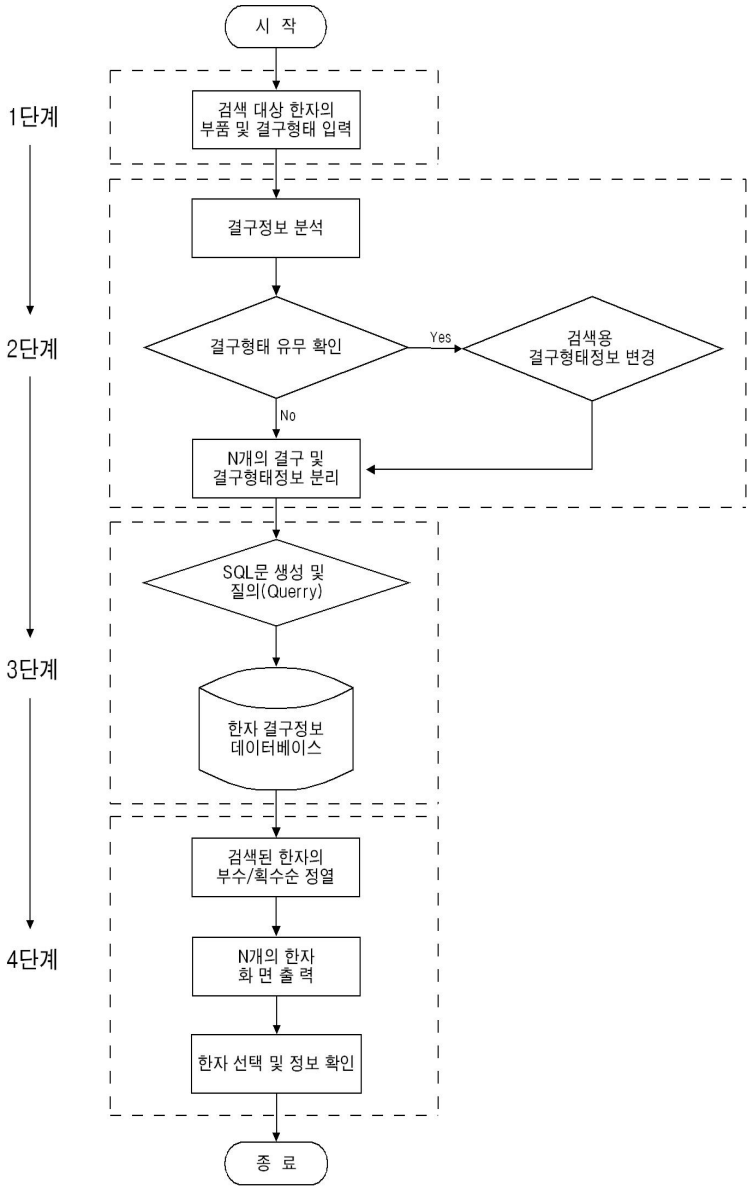
(1)은 검색하고자 하는 한자를 구성하는 최소한 하나 이상의 부건과 결구형태를 입력하는 부분으로 검색 대상 한자를 구성하는 하나 이상의 부건과 결구형태를 입력할 수 있는 입력창과, 변형된 부수자가 포함되어 있을 경우를 위해 부수자[변형된 부수자 포함]를 입력할 수 있는 부수테이블 창으로 구성된다.

(2)는 입력 요청된 결구정보를 분석하는 부분으로 내부 처리모듈로 구성되는데, 입력 요청된 결구정보를 분석하여 검색용 결구형태 정보로 변경하거나 결구형태 정보를 분리한다. 이를 위해서 먼저 입력 요청된 하나 이상의 부건과 결구형태로 구성되는 결구정보를 분석하여 결구형태의 유무를 확인한다. 이때 결구형태가 포함되어 있을 경우에는 검색용 결구형태 정보로 변경하여 복수 개의 결구 및 결구형태 정보로 분리하며, 결구형태가 포함되어 있지 않을 경우에는 검색용 결구형태 정보로 변경하지 않고 복수 개의 결구 및 결구형태 정보로 분리한다.

(3)은 한자를 구성하는 부건을 결구형태와 함께 구조화한 데이터베이스로 그 구성에 대해서는 이미 앞에서 언급하였으므로 자세한 설명은 생략한다.

(4)는 검색 결과를 화면상에 출력하는 부분으로 보편적인 화면 출력에 관한 기술을 사용하므로 자세한 설명은 생략한다.

이렇게 구성되는 한자검색 시스템의 검색 진행과정은 모두 4단계로 진행된다. 그 진행과정은 (1)검색하고자 하는 한자를 구성하는 최소한 하나 이상의 부건과 결구형태를 입력하는 제1단계, (2)입력 요청된 결구정보를 분석하여 검색용 결구형태 정보로 변경하거나 결구형태 정보를 분리하는 제2단계, (3)분석된 결구형태 정보를 바탕으로 SQL문을 생성하여 데이터베이스에 질의(Query)를 하는 제3단계, (4)데이터베이스로부터 질의 결과를 받아 부수와 획수 순으로 정렬하여 복수 개의 한자를 화면상에 출력하는 제4단계이다.



【한자검색 시스템 검색 진행과정】

제1단계에서는 먼저 찾고자 하는 한자를 부수자[변형된 부수자 포함]를 포함하여 자소 단위가 아닌, 자음과 자의를 가지는 복수 개의 부건으로 분해하며, 분해된 부건을 결구형태와 함께 부건으로 사용된 하나 이상의 한자를 검색창에 입력한다. 한자를 복수 개의 부건으로 분해할 때에는 반드시 부수자[변형된 부수자 포함]이거나 자음과 자의를 가지는 한자로 분해하여야 한다. 자음과 자의를 모르는 부건일 경우에는 입력을 생략할 수 있으며, 기본 입력기를 통해 부수자[변형된 부수자 포함]를 입력할 수 없을 경우에는 '부수자[변형된 부수자 포함] 입력창'을 통해서 입력할 수 있다. 또 더 이상 자음과 자의를 가지는 한자로 분해되지 않는 한자일 경우에는 해당 한자를 직접 입력할 수 있다. 이 과정에서 분해된 각 부건의 입력은 운영시스템에서 제공하는 유니코드 Ext.A 영역의 한자까지 한글 음가를 이용하여 변환 가능한 기본 입력기를 사용한다.[14]

그리고 12가지의 결구형태는 다음의 표와 같이 '+', '/', '++', '//', '*', '/*', '*/', '[*', '(*', '*)', '_*', '&' 등으로 대체하여 입력할 수 있고, 검색하고자 하는 한자의 결구형태를 파악할 수 없을 경우에는 ';(세미콜론)'과 같은 특정 부호나 문자로 대체 또는 생략할 수도 있다.[15]

결구형태	대응부호	글자 예	질의 방법
AB	+	校= 木交	+木交
AB	/	忠= 中心	/中心
ABC	++	瀞= 氵青爭	++氵青爭
ABC	//	蓋= ++去皿	//++去皿
AB	*	回= 口口	*口口
AB	(	問= 門口	(門口
AB	)	凶= 凵土	)凵土
AB	[	医= 匚矢	[匚矢
AB	<	病= 疒丙	<疒丙
AB	>	匈= 勹凶	>勹凶
AB	_	迄= 辶乞	_辶乞
AB	&	爽= 大玨	&大玨

14) 현재 Windows XP 혹은 리눅스와 같은 대부분의 운영체계에서 기본 입력기를 통해 유니코드 Ext.A 영역의 한자까지 한글 음가를 이용하여 변환하여 입력할 수 있다.

15) 표에서 제시한 대응부호 외에도 검색엔진 내부에서 얼마든지 특정한 문자나 부호로 대치하여 프로그래밍을 할 수 있으며, 결구형태를 생략하고 부건만 입력할 수도 있다.

제2단계에서는 먼저 제1단계에서 입력 요청된 하나 이상의 부건과 결구형태로 구성되는 결구정보를 분석하여 결구형태의 유무를 확인하는데, 결구형태가 포함되어 있을 경우에는 검색용 결구형태정보로 변경하여 복수 개의 결구 및 결구형태 정보로 분리하며, 결구형태가 포함되어 있지 않을 경우에는 검색용 결구형태 정보로 변경하지 않고 복수 개의 결구 및 결구형태 정보로 분리한다.

제3단계에서는 제2단계에서 분리된 복수 개의 결구정보와 결구형태 정보를 일정한 검색 조건을 포함하는 SQL문을 생성하고, 생성된 SQL문과 합치하는 복수 개의 한자를 한자 결구정보 데이터베이스로부터 호출한다.

제4단계에서는 제3단계에서 호출된 복수 개의 한자를 부수와 획수 순으로 정렬하고, 정렬이 완료된 검색 결과를 화면상에 출력한다.

Ⅴ. 과제와 전망

이상에서 漢字結構情報 데이터베이스를 이용한 한자검색 모델에 대해 살펴보았다.

한자는 일정한 結構에 의하여 일정한 개수의 部件으로 분해할 수 있으며, 분해된 부건은 이미 알고 있는 특정 한자[부수자 포함]일 수도 있고, 의미를 가지지 않는 하나의 필획일 수도 있다. 그리고 이렇게 분해된 부건을 결구와 함께 구조적으로 데이터베이스화하며, 이를 '漢字結構情報 데이터베이스'로 정의한다. 그리고 이 데이터베이스를 이용하여 자음과 자의를 모르는 한자를 분해된 부건의 組合만으로 편리하게 검색할 수 있도록 하였다.

이 한자검색 모델은 앞에서도 언급하였듯이 다량의 한자 가운데에서 원하는 한자를 빠르게 검색하고자 할 때, 유용하게 활용할 수 있을 것이다. 그리고 사용자는 찾고자 하는 한자의 字音과 字義를 정확하게 모르더라도 기본적인 한자의 분해와 조합에 대한 개념만 알고 있으면, 손쉽게 한자를 검색할 수 있는 것이 특징이다. 즉 기존에 알려져 있는 부수에 의한 한자검색이나 자음과 자의에 의한 한자검색이 아닌, 한자의 자형을 중심으로 한자를 검색하므로 난해한 한자를 쉽게 검색할 수 있도록 하였다.

이러한 특징을 충분히 살린다면, 이 한자검색 시스템은 신출한자를 비롯하여 異體字·簡體字·略字, 또는 草書나 古漢字 등과 같이 표준 코드체계에 등록되지 않는 비표준한자를 이미지 형태로 데이터베이스화한 자료로부터 원하는 한자를 검색할 수도 있을 것이다. 뿐만 아니라 한자의 조합에 의한 폰트 제작 시스템이나 전자사전, 그리고 한자 입력기와 같은 각종 소프트웨어에도 응용할 수 있을 것이다.

그러나 국내 학계에서는 아직 현대 한자학에 대한 연구 성과가 미흡한 까닭으로 한국의 어문생활에 완전하게 부합되는 한자결구정보 데이터베이스를 구축하였는지에 대한 검정 작업이 없었다. 향후 보다 정밀하게 한자결구정보 데이터베이스를 보완해야 하는 과제를 해결해야 할 것이다.

참고문헌

[단행본]

The Unicode Consortium, *"The Unicode Standard Version 4.0"*, Addison-Wesley Professional, 2003.

The Unicode Consortium, *"The Unicode standard 5.0"*, Addison-Wesley Professional, 2006.

蘇培成 著·李圭甲 譯, 『現代漢字學』, 學古房, 2007.

서경호 등, 『國際 文字 코드 提案 漢字의 標準化에 대한 硏究(상 하권)』, 문화관광부, 1998.

고려대학교 민족문화연구원·(주)솔트웍스 컨소시엄,『 '지식정보자원관리사업 신출한자 통합관리를 위한 목록 및 지침 작성' 과제 결과보고서』, 한국전산원, 2004.

[사이트]

유니코드 콘소시엄 홈페이지 http://www.unicode.org/

IRG 홈페이지 http://www.cse.cuhk.edu.hk/~irg/

五筆字型輸入法 홈페이지 http://www.hongen.com/pc/newer/ime/wbx/imejswb1.htm

국문초록

이 논문은 현대 漢字學의 構形學 이론에서 제시하고 있는 漢字結構와 한자의 分解와 組合에 대한 이론을 바탕으로 한자검색 시스템을 연구한 것이다. 한자는 일정한 結構에 의하여 일정한 개수의 部件으로 분해할 수 있으며, 분해된 부건은 이미 알고 있는 특정 한자[부수자 포함]일 수도 있고, 의미를 가지지 않는 하나의 필획일 수도 있다. 그리고 이렇게 분해된 부건을 구조적으로 데이터베이스화하며, 이 데이터베이스를 이용하여 자음과 자의를 모르는 한자를 분해된 부건의 組合만으로 편리하게 검색할 수 있다.

이와 같은 이론적 근거를 바탕으로 한자의 部件과 부호화된 조합 규칙인 結構形態를 이용하여 모든 한자를 구조적으로 데이터베이스화하며, 이를 '漢字結構情報 데이터베이스'로 정의한다. 데이터베이스 구축에는 몇 가지 원칙을 적용하였는데, ① 데이터베이스 구축 범위는 유니코드에 등록된 한자 74,474자로 하고, ② 모든 한자를 부수자[변형된 部首字 포함]를 포함하여 字音과 字義를 가지는 하나 이상 N개의 部件으로 분해하며, ③ 부건은 기본 운영체계에서 한글 음가로 입력 가능한 유니코드 Ext.A 영역의 한자까지로 제한하며, ④ '忄[心], 巛[川], 乚[乙], 牜[牛]' 등과 같이 변형된 部首字인 경우에만 별도의 입력창을 통해 입력한다.

이상과 같은 방식으로 구성된 한자결구정보 데이터베이스를 바탕으로 한자검색 시스템을 구축할 수 있는데, ① 검색하고자 하는 한자를 구성하는 최소한 하나 이상의 부건과 결구형태를 입력하는 부분과, ② 입력 요청된 결구정보를 분석하는 부분, ③ 한자를 구성하는 부건을 결구형태와 함께 구조화한 한자결구정보 데이터베이스 부분, ④ 한자결구정보 데이터베이스로부터 검색 결과를 출력하는 부분으로 구성된다.

이렇게 구성된 한자검색 시스템은 ① 검색하고자 하는 한자를 구성하는 최소한 하나 이상의 부건과 결구형태를 입력하는 제1단계, ② 입력 요청된 결구정보를 분석하여 검색용 결구형태 정보로 변경하거나 결구형태 정보를 분리하는 제2단계, ③ 분석된 결구형태 정보를 바탕으로 SQL문을 생성하여 데이터베이스에 질의(Query)를 하는 제3단계, ④ 데이터베이스로부터 질의 결과를 받아 부수와 획수 순으로 정렬하여 복수 개의 한자를 화면상에 출력하는 제4단계로 검색을 진행한다.

이 한자검색 모델은 한자의 字音과 字義를 정확하게 모르더라도 기본적인 한자의 분해와 조합에 대한 개념만 알고 있으면, 손쉽게 한자를 검색할 수 있는 것이 특징이다. 이러한 특징으로 인해 신출한자를 비롯하여 異體字·簡體字·略字, 또는 草書나 古漢字 등과 같이 표준 코드체계에 등록되지 않는 비표준한자를 이미지 형태로 데이터베이스화한 자료로부터 원하는 한자를 검색하는 데에 유용하게 이용할 수 있을 것이다. 뿐만 아니라 한자의 조합에 의한 폰트 제작 시스템이나 전자사전, 그리고 한자 입력기와 같은 각종 소프트웨어에도 응용할 수 있을 것이다.

주제어 : 한자검색 시스템, 한자결구, 결구형태, 한자분해, 한자조합, 한자결구정보 데이터베이스

【Abstracts】

The Study of Ideographic Characters Searching System, Using IDS(Ideographic Description Sequence)

Shin, Sang-hyun*

This article is the study of Ideographic Characters Searching System, using IDS(Ideographic Description Sequence) that based on the theory of Ideographic Structure Characters in Han Ideographic Characters.

Ideographic Characters can be analyze into some specific components, they are well known characters to include radicals or strokes to have no meanings. And we can construct database structurally by the union of these components with the kind of 12 IDC(Ideographic Description Characters). Due to using this database, we can search Ideographic Characters simply; nevertheless, it's the first meeting, never seen before.

Based on this theory of Ideographic Structure Characters in Han Ideographic Characters, we can construct a new type database called IDS Database. To structure of database be applied some rule, as follow four items.

① Number of Ideographic Characters are 74,474 in unicode.

② All of Ideographic Characters analyze into components the numbers of N having pronunciations and meanings.

③ Components are limited within CJK Unified Ideographics Ext.A, and must be input to change Ideographic Characters by Hangul.

④ Radicals changed original shape, example of 忄[心], 巛[川], 乚[乙], 牜[牛], are input through special input windows.

Based on this IDS Database, we can construct Ideographic Characters Searching System. It is consist of four part, ① the part of input one more components at

* Researcher, Department of Character Code Research Center, Institute of Korean Culture in Korea University

least with IDC, ② the part of analysis components and IDC, ③ the part of IDS Database, ④ the part of display the results of searching. And It has four step's processing for searching progress, ① the first step of input one more components at least with IDC, ② the second step of analysis components and IDC, or changing to IDS it is suitable for searching, ③ the third step of generating SQL Query to IDS Database, ④ the forth step of display the results of searching, getting from IDS Database.

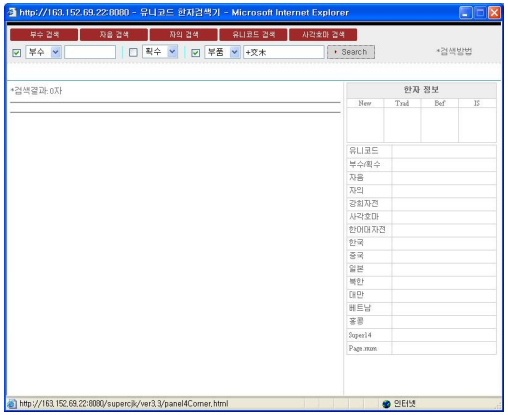
Thanks to use this Searching System, we can search Ideographic Characters simply, if you know that Ideographic Characters can be analyze into some specific components, and these components can be the union of its one; nevertheless, it's the first meeting, never seen before, and have no informations of it's pronunciations and meanings. Due to this feature, It will be use non standardization Ideographic Characters Searching System usefully, lake as Database of Variants Ideographic Characters, Database of Old Ideographic Characters, etc.

Key word : Ideographic Characters Searching System, IDS(Ideographic Description Sequence), IDC(Ideographic Description Characters), Ideographic Characters analysis, Ideographic Characters union, IDS Database

【부록】 한자검색 예

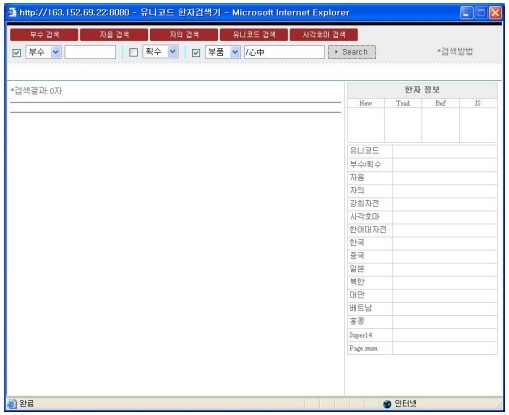
1. 결구형태를 이용한 검색 예

1) 「校= 木交 → +木交」 검색 화면



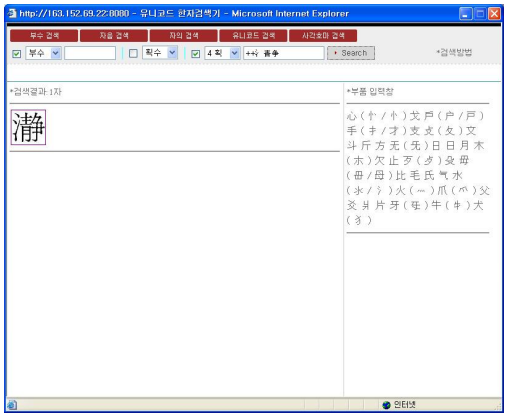
※ '+交木'의 경우에는 검색결과를 얻을 수 없다.

2) 「忠= 中心 → /中心」 검색화면

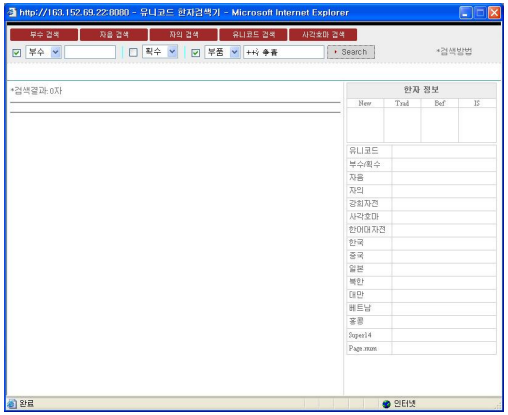


※ '/心中'의 경우에는 검색결과를 얻을 수 없다.

3) 「瀞= 氵青爭 → ++氵青爭」 검색 화면

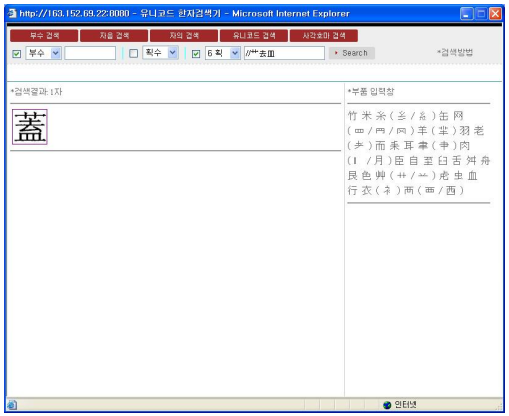


※ '氵'가 기본입력기에서 입력 불가능하므로 화면 오른쪽의 '부품입력창'을 통해 입력하였다.



※ '++氵爭靑' 또는 '++爭靑氵'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

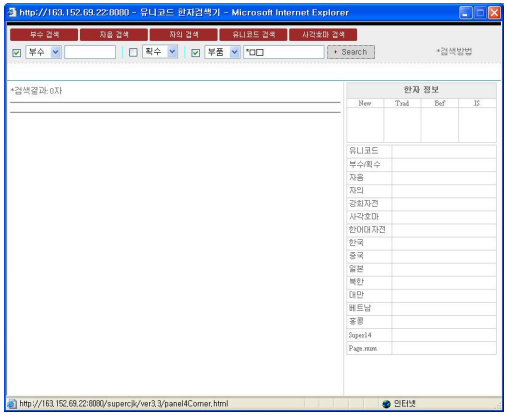
4) 「蓋= ⿱艹去皿 → //艹去皿」 검색 화면



※ '艹'가 기본입력기에서 입력 불가능하므로 화면 오른쪽의 '부품입력창'을 통해 입력하였다.

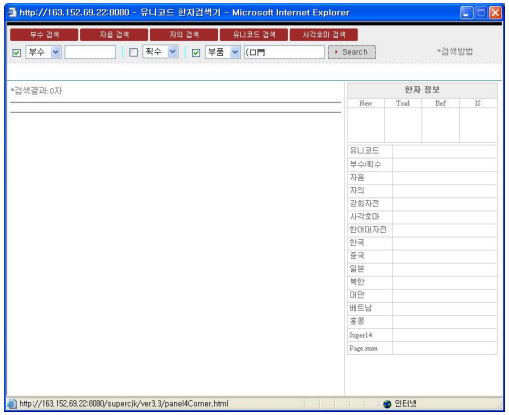
※ '//艹皿去' 또는 '//去皿艹'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

5) 「回= ⿴口口 → *口口」 검색 화면



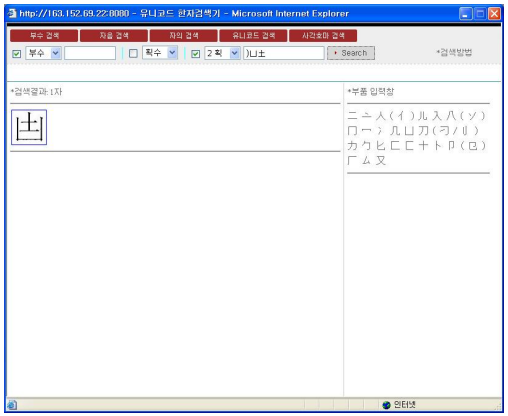
※ '*口口'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

6) 「問= 門口 → (門口」 검색 화면

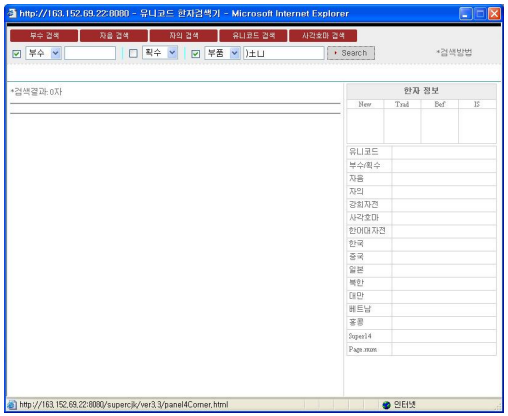


※ '(口門'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

7) 「凶= 土凵 →)凵土」 검색 화면

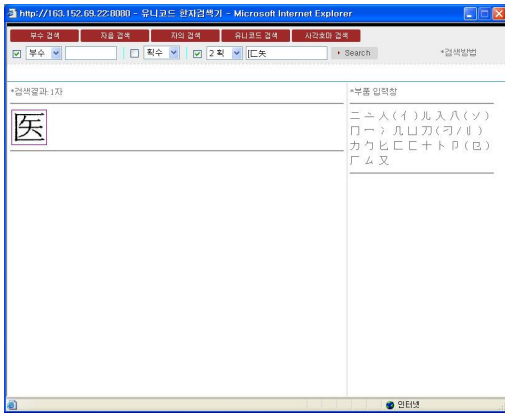


※ '凵'가 기본입력기에서 입력 불가능하므로 화면 오른쪽의 '부품입력창'을 통해 입력하였다.

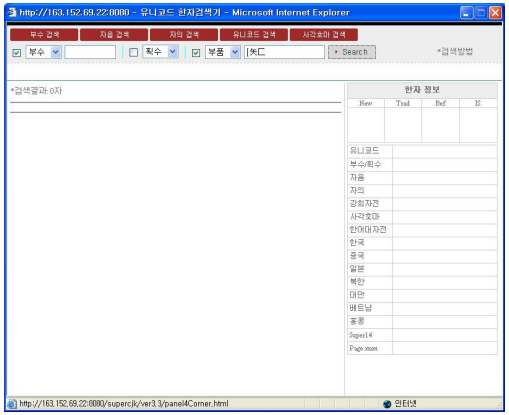


※ ')土凵'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

8) 「医= 匸匚矢 → [匚矢」 검색 화면



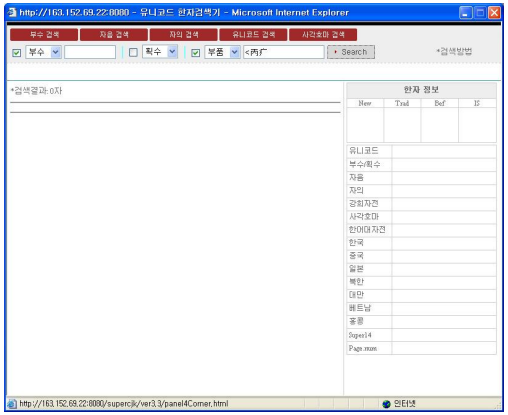
※ '匚'가 기본입력기에서 입력 불가능하므로 화면 오른쪽의 '부품입력창'을 통해 입력하였다.



※ '[矢匚'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

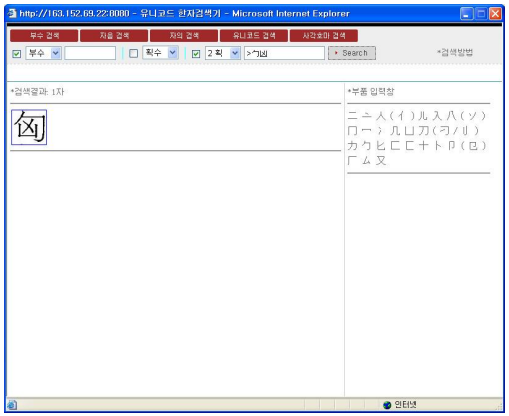
9) 「病= 疒丙 → <疒丙」 검색 화면

※ '疒'가 기본입력기에서 입력 불가능하므로 화면 오른쪽의 '부품입력창'을 통해 입력하였다.

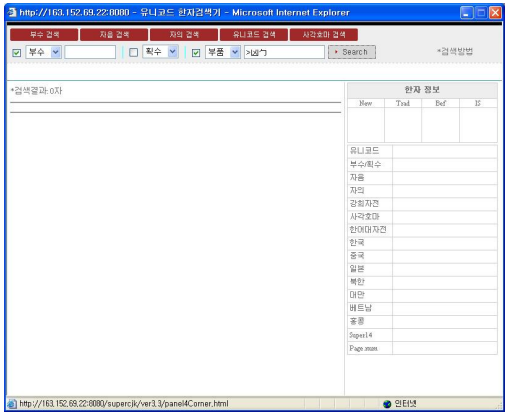


※ '<丙疒'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

10) 「匈= 勹勹凶 → >勹凶」 검색 화면

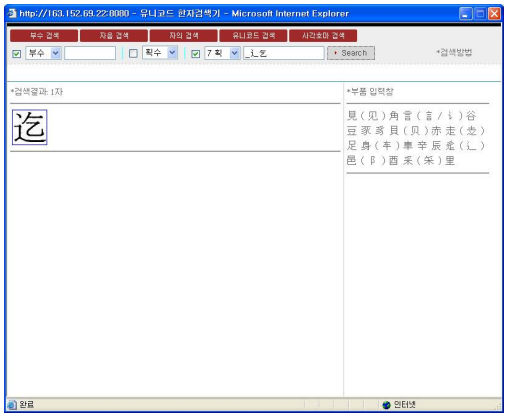


※ '勹'가 기본입력기에서 입력 불가능하므로 화면 오른쪽의 '부품입력창'을 통해 입력하였다.

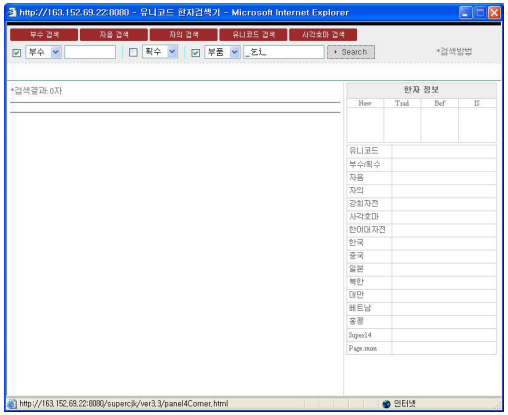


※ '>凶勹'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

11) 「迄= 𠂉乞 → _𠂉乞」 검색 화면

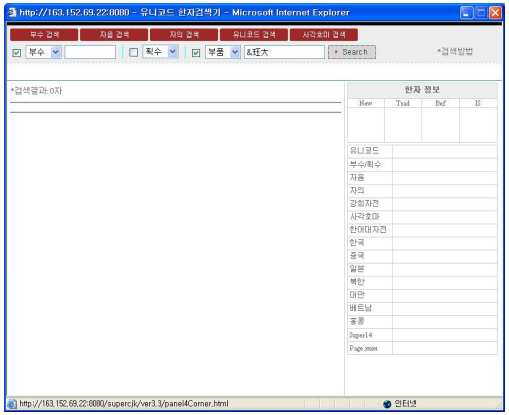


※ '𠂉'가 기본입력기에서 입력 불가능하므로 화면 오른쪽의 '부품입력창'을 통해 입력하였다.



※ '_乞𠂉'와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

12) 「爽= ⿰大玨 → &大玨」 검색 화면

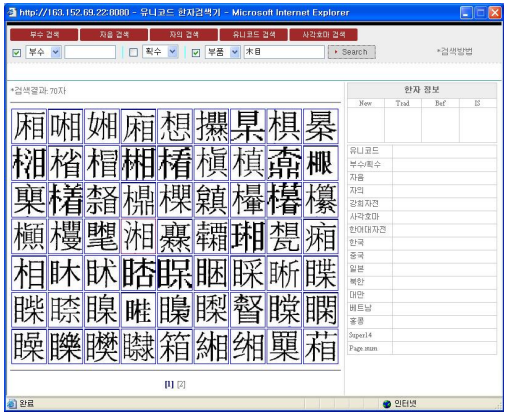


※ ‘&玨大’와 같이 입력순서가 틀리면 검색결과를 얻을 수 없다.

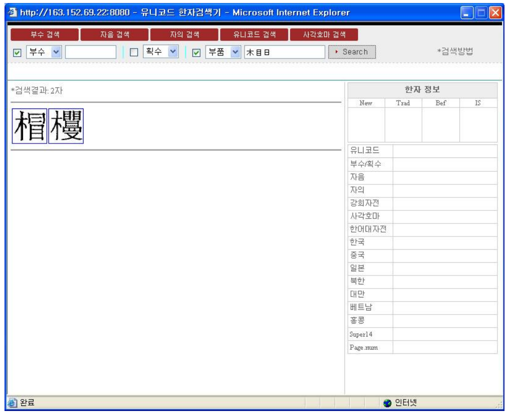
2. 결구형태를 대체[또는 생략]한 검색 예

○ ‘木’ 부건을 가지고 있는 모든 한자를 검색하면 3,866자의 결과를 얻을 수 있다.

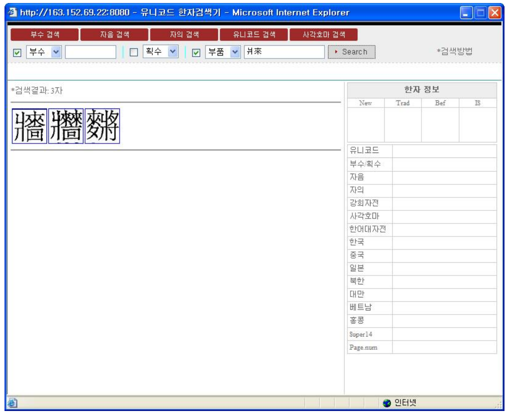
이 3,866자 가운데 ‘木’과 ‘目’ 부건을 동시에 가지는 모든 한자를 검색하면 70자의 결과를 얻을 수 있다. 이 때 ‘木目’ 또는 ‘目木’을 입력하여도 같은 결과를 얻을 수 있다.

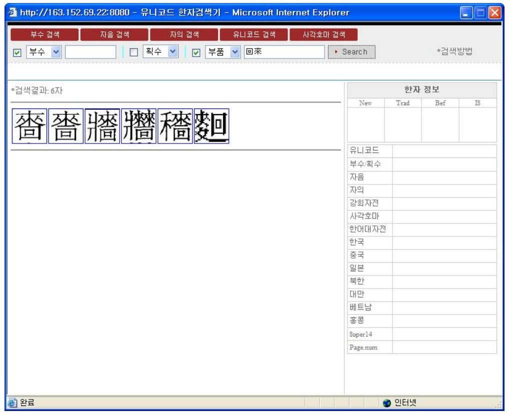


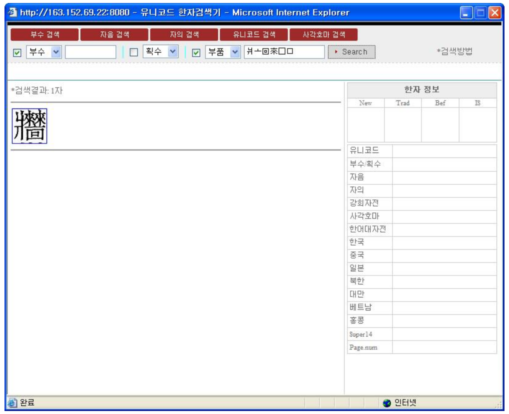
이 70자 가운데 ‘木’과 ‘目’, 그리고 ‘日’ 부건을 동시에 가지는 모든 한자를 검색하면 2자의 결과를 얻을 수 있다. 이 때 ‘木目日’ 또는 ‘木日目’, ‘目木日’ 또는 ‘目日木’, ‘日木目’ 또는 ‘日目木’을 입력하여도 같은 결과를 얻을 수 있다.



○ ‘ 𤖧U+245A7’은 결구정보가 ‘⿰爿⿱𣛧亩;⿱亠回;⿰來來;⿴囗口’인데, 이 가운데 ‘爿來’, ‘爿回’, ‘回來’, ‘爿亠回來囗口’, ‘亠囗口爿回來’ 등 다양한 부건의 조합으로 순서에 관계없이 다음과 같이 검색을 할 수 있다.(그림 순서대로 설명)




http://163.152.69.22:8080 - 유니코드 한자검색기 - Microsoft Internet Explorer
Search
*검색결과: 6자
한자 정보
유니코드
부수·획수
완료
인터넷


http://163.152.69.22:8080 - 유니코드 한자검색기 - Microsoft Internet Explorer
Search
*검색결과: 1자
한자 정보
유니코드
부수·획수
완료
인터넷