대한한의학원전학회지
2019년 02월 32권 1호

pISSN 1229-8328
J Korean Med Classics 2019:32(1) 105-116
https://doi.org/10.14369/jkmc.2019.32.1.105

JKMC
Journal of Korean Medical Classics

고문헌 벽자(僻字) 입력을 위한 한자 자형 부호화 방법

부산대학교 한의학전문대학원 교수
김기왕*

A Character Shape Encoding Method to Input Chinese Characters in Old Documents

Kim Kiwang*
Professor at Pusan National University School of Korean Medicine

Objectives : There are many secluded Chinese characters - so called Byeokja (僻字) in ancient classic literature, and Chinese characters that are not registered in Unicode and Variant characters (heterogeneous characters) that cannot be found in the current font sets often appear. In order to register all possible Chinese characters including such characters as units of information exchange, this study attempts to propose a method to encode the morphological information of Chinese characters according to certain rules.

Methods : This study suggests the methods to encode the connection between the nodules constituting the Chinese character and the coordinates of the nodules. In addition to that, rules for expressing information about curves, expressions of aspect ratios of characters, rules for minimizing coordinate lines, and rules for expressing aggregation status of character components are added.

Results : Through the proposed method, it is possible to generate codes of a certain length by extracting only information expressing the morphological configuration of characters.

Conclusions : The method of character encoding proposed in this study can be used to distinguish variant characters with small variations in Byeokja, new Chinese characters and character strokes and to store and search them.

Key words : Chinese characters, old classics, encoding, shape based code

* Corresponding author : Kim Kiwang.
Pusan National University School of Korean Medicine
Tel : 051) 510-8466. E-mail : kimgiwang@hanmail.net
Received(February 7, 2019), Revised(February 12, 2019), Accepted(February 12, 2019)

Ⅰ. 서론

한자로 이루어진 고문헌을 보다 보면 수많은 생소한 글자들, 즉 벽자(僻字)를 마주하게 된다. 또한 비교적 잘 알려진 글자라 하더라도 현재의 표준 자형에서 벗어난 다양한 이체자(異體字)로 이를 표기한 사례를 적지 않게 볼 수 있다. 이들을 전자 문서로 입력하려 하거나 활자화된 문서로서 출판하고자 할 때는 갖가지 어려움을 경험하게 된다. 이러한 문자는 상용 문자 목록에서는 검색이 되지 않는데다 어떤 방식으로 검색해야 하는지 가늠하기도 어렵다. 어렵게 검색이 되더라도 해당 한자의 입력을 위해 추가적인 글꼴 파일의 설치가 필요할 때도 있다. 더구나 아예 현재의 문자 목록 어디에도 등록되어 있지 않은 글자를 다룰 경우도 있는데 이 때는 기존 문자의 조합으로 이 글자를 지시하거나 문자가 아닌 그림으로서 해당 문자를 표현하는 수밖에 없다.

그러나 한자를 구성하는 필획의 구조를 일정한 규칙에 따라 부호로 표현할 수 있다면 어떠한 문자가 나타나든지 이를 특정 부호(코드)로 표시하여 컴퓨터에 기록하거나 문서로 출판할 수 있을 것이다. 이렇게 한자의 형태를 부호화하는 일정한 방식이 확립된다면 검색이 쉽지 않은 한자를 찾는 데 도움이 될 뿐만 아니라 형태를 기준으로 실존하는 갖가지 한자들의 실태를 정비하고 한자를 통한 정확하고도 유효한 정보 교환이 이루어지게 하는 데 기여할 수 있다.

따라서 본 논고에서는 한자의 형태를 부호화하는 한 가지 방법을 제안하고자 한다.

II. 본론

가) 현황 - 어디에 문제가 있는가

한자는 열린 집합이다. 즉 정해진 수량의 문자 집합으로 규정할 수 없다. 로마자(알파벳)가 26종, 일본어 가나가 96종(탁음, 요음 표시 등 변형 요소 제외), 한글이 11,172종인 것과 달리 한자는 문자 총수가 한정되어 있지 않으며 오늘날에도 새롭게 만들어지고 있다. 예를 들어 원소(元素)를 나타내는 한자는 지난 한 세기에 대다수가 새로 만들어졌으며 최근 화학계의 발표(2016년)에 뒤따라 니호늄(Nh), 모스코븀(Mc), 테네신(Ts), 오가네손(OG)이라는 새로운 원소를 나타내는 4 가지 한자(鉨, 鏌, 𥑮, 𨨏)가 원소 명칭으로서 추가되기도 하였다. 정보 교환을 위한 세계 문자의 부호 목록인 유니코드(Unicode)에서도 한자 영역은 여러 차례의 확장을 거쳐 현재는 확장 영역 F(CJK Unified Ideographs Extension F)까지 만들어진 상태다.

과거 동아시아에서는 활자 인쇄가 아닌 목판 인쇄 또는 필사(筆寫)의 방법을 통해 서적을 만드는 일이 많았기 때문에 문헌에 새로운 글자를 기입하는 데 아무런 제약이 없었고, 이 때문에 수많은 이체자(異體字)가 지속적으로 만들어졌다.

이처럼 한자의 수량이 한정되어 있지 않다는 특성 때문에 한자로 표기된 문자를 출판하기 위해서는 종종 거대한 글꼴(font) 파일을 확보해야만 한다. 예를 들어 위에서 언급한 유니코드 확장 영역 F를 온전하게 표현할 수 있는 글꼴인 화원명조체(花園明朝體, HanaMinA.ttf, HanaMinB.ttf)에는 모두 107,518자의 문자가 포함되어 있다(한자는 97,712자).

더 어려운 문제는, 이러한 글꼴을 확보해도 표현할 수 없는 글자가 고문헌에 출현한다는 것이다. 현재의 어떤 글꼴 목록에서도 찾을 수 없는 이러한 한자를 입력하려면 결국 문자가 아닌 그림으로 그 글자를 표시할 수밖에 없다. 실제로 고문헌을 다루는 학술 논문에서는 종종 손으로 쓴 글자가 섞여 있는 것을 볼 수 있다.

하지만 이렇게 그림으로 표시된 글자들은 검색을 통해 전송과 가공이 가능한 정보 요소로서 역할을 할 수가 없다.

결국, 새로운 한자가 그림이 아닌 정보의 매체로서 기능을 하려면 일정한 규칙에 따라 부호화되어야만 한다.

나) 기존의 자형 부호화 방법과 한계

한자에 부호를 부여하여 한자 정보를 체계적으로 정리하려는 시도는 오래 전부터 존재하였다. 전기통신이라는 목표에 국한한 개발 사례이기는 하나

1872년 개발된 한자 전보(電報) 코드[1]는 그러한 시도의 초기 사례라 할 수 있다.

한자의 부호화는 음이나 획수에 따라 이루어진 예도 있으나 자형을 기준으로 한 부호화 시도가 많았다. 일찍이 1925년 중국의 왕운오(王雲五, 1888-1979)는 사각호마(四角號碼)를 제창[2]하여 한자의 네 모퉁이에 자리잡은 필획의 형태에 따라 한자에 4자리(중복문자 식별부호를 포함하면 5자리)의 십진수 부호를 부여하였다. 사각호마의 도입 목적은 한자의 검색을 쉽게 하려는 데 있었으나 20세기 후반, 컴퓨터가 보급됨에 따라 한자 검색 이외에 한자의 입력이나 정보 교환을 위한 포괄적 목적에서 한자의 자형을 부호화하려는 시도가 이어졌다. 대표적인 예가 1976년 주방복(朱邦復, 1937-)이 발표한 창힐입력법[倉頡輸入法]에 의한 부호화 사례다. 이후로 자형을 토대로 한 수많은 한자 부호화 체계(소위 "形碼" 체계)가 제안되었다.

그러나 지금까지 제안된 부호화 방식은 자형 분해 과정이 자의적인 경우가 많고(Fig. 1), 자형을 보고 문자 코드를 도출할 수는 있으나 문자 코드로부터 하나의 자형이 결정되지는 않는다(Fig. 2). 이는 과거의 한자 부호화 체계가 한정된 문자 집합 안에서 코드와 문자의 일대다 대응을 최소화하는 데만 목적을 두어 만들어진 것이기 때문이다.

Fig. 1. An example of difficulties in Changjie encoding method. Stipulated code of Han (漢) in current Changjie system is ETLO (a), but it is possible for someone to input ETLK code (b) or ETLI code (c).

그림1. 자형 기반 한자 부호화의 어려움. 창힐입력법(倉頡入力法) 체계에서 한(漢) 자는 水(氵)+廿+中(+手)+人(𠆢)와 같은 자형 요소가 결합된 것으로 볼 수 있으나(a) 어떤 사람은 水(氵)+廿+中(+十)+大의 형태로 분석을 시도할 수도 있으며(b) 이 글자의 다른 자형을 기준으로 할 경우 水(氵)+廿+中(+手)+戈(丶)의 형태로 분석하게 된다(c).

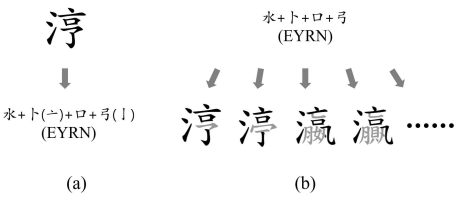

Fig. 2. Irreversibility in current shape based Chinese character encoding system. Although it is possible to get a Changjie code from a character in most cases, but, sometimes, it is impossible to reconstruct a unique character from a Changjie code.

그림2. 현재의 자형 기반 한자 부호화 체계에서 나타나는 비가역성. 형(淳) 자에 대한 창힐입력법의 코드는 EYRN(水+卜(亠)+口+弓(亅))이지만 EYRN이란 코드에 대응하는 한자는 형 자 이외에도 여러 글자가 있다.

따라서 본 논고에서는 한자에 대한 지식이 없는 사람도 간략한 규칙만으로 한자의 자형 코드를 생성할 수 있는 방법, 아울러 하나의 자형 코드가 정확히 하나의 글자에 대응되는 방식의 자형 코드를 만드는 방법에 대해 제안하고자 한다.

다) 제안

본 논고에서 제안하는 한자 부호화 방법의 핵심은, 어떤 한자의 자형을 부호화하기 위해 해당 한자

1) 威基謨. 電報新書. 上海. 電機司. 1872.
2) 王雲五. 號碼檢字法. 東方雜誌. 1925. 22(12). pp.82-98.

를 그와 위상동형(位相同型, homeomorphic)인 가장 간단한 도형으로 바꾼다는 것이다(Fig. 3). 이렇게 단순화된 도형에서 각 선의 점(시작점, 종지점, 연결점)과, 점 사이의 관계(어느 점과 연결되는가)를 최소한의 정보로 표현해 보자는 것이 필자의 제안이다.

(a) (b) (c)

Fig. 3. Three types of homeomorphic form for Jia (家) character. A character used in stamp (a), a gothic type character (b) and it's homeomorphic simplified form (c).

그림3. 인장(印章)에 사용하는 가(家) 자(a)와 고딕체 가 자(b). 오른쪽(c)은 이와 이를 간략하게 변형한 위상동형 도형.

그러나 한자를 포함한 대다수 문자 체계 안에서 위상동형인 문자가 서로 다른 글자로 간주되는 경우가 많다. 예를 들어 부수글자인 두(亠, 돼지해머리) 자와 면(宀, 갓머리) 자는 위상동형이지만 같은 글자가 아니다.

이러한 문제로 인해 한자의 형태를 부호화할 때는 가장 간단한 위상동형의 도형으로 축약하는 방법을 일률적으로 적용할 수 없고, 적당한 제한 조건 아래에서 비교적 간단한 위상동형 도형으로 변형한 후 그 형태 정보를 부호화하는 방법을 적용해야 한다.

1) 결절, 관계

다수의 문자 체계에서 선이 꺾이는 부분은 문자를 구별해 주는 중요한 표지가 된다. 따라서 도형의 위상동형 여부를 구분해 주는, 선과 선이 만나는 점뿐만 아니라 선이 꺾이는 점도 자형의 부호화 과정에서 의미 있는 요소로 인정해야 한다.

이와 같이 꺾인 점을 포함하여 시작점, 종지점, 연결점의 4종류 점을 의미 있는 결절(結節, node)로 보고 이들 결절의 상대적 위치와 결절 간의 연결 상태를 표현하면 이것이 한자 자형의 정보를 함축한 코드가 될 수 있다고 본다.

아래에서 아(亞) 자를 예로 들어 필자가 제안한 한자 자형 부호화의 방법을 설명하기로 한다(Fig 4).

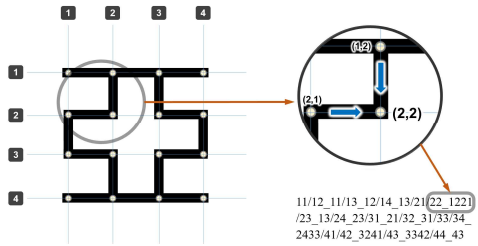

Fig. 4. Shape encoding of Ya (亞) character.

그림4. 아(亞) 자의 자형 부호 생성.

일단 결절 각각에 좌표를 부여한다. 결절이 위치한 행이 a, 결절이 위치한 열이 b일 때 이 결절의 좌표를 'ab'의 형태로 표시한다. 예를 들어 Fig. 4에서 맨 왼쪽 위의 결절은 11, 그 오른쪽의 결절은 12로 표시된다.

이어서 각 결절이 자신과 이어지는 상·좌측 결절들과 어떤 관계를 갖는지 표시한다. Fig. 1에서 2행 2열에 있는 결절은 위쪽의 결절(1행 2열의 결절)과 왼쪽의 결절(2행 1열의 결절)과 연결된다. 이를 "22_1221"로 표시한다.

이상의 방법에 따라 맨 왼쪽 위 결절로부터 맨 오른쪽 아래의 결절까지 그 위치와 관계를 표시하고 각 결절 정보 사이를 사선(/)으로 구분하면 Fig. 4의 오른쪽 아래에 나타낸 것과 같은 자형 코드가 생성된다.

2) 곡선의 처리

한자는 다른 문자에 비해 곡선이 적은 편이지만 곡선이 없는 것은 아니기에 한자의 자형 부호화 과정에서도 곡선에 관한 형태 정보를 표현할 방법이 필요하다.

필자가 제안하는 방법은 Fig. 5에 나타낸 바와 같이 곡선에 연결된 결절에 선의 비틀림 방향을 표시해 주는 것이다. 인(人) 자의 왼쪽 아래 결절(3행 1열의 결절)은 위쪽 결절(2행 2열의 결절)과 연결되어 있는데 두 점을 직선으로 연결한 후 3행 1열 결

절에서 이를 시계 방향으로 비틀면 인(人) 자의 왼쪽 획 형태가 된다. 이를 수직선(|)을 이용하여 “31_22|0” 이와 같은 형태로 표시한다. 수직선 오른쪽의 숫자 0은 0도 방향으로 획 종단을 꺾는 것을 말하며 숫자 1은 -90도(270도) 방향으로 획 종단을 꺾는 것, 숫자 2는 180도 방향으로 획 종단을 꺾는 것, 숫자 3은 90도 방향으로 획 종단을 꺾는 것을 표현한다. 이와 같은 규칙에 따라 Fig. 5에서 3행 3열의 결절이 가지는 형태 정보는 “33_22|2”로 표시된다.

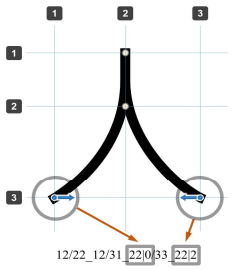

Fig. 5. Shape encoding of Ren (人) character.
그림5. 인(人) 자의 자형 부호 생성.

이런 방법을 통해 유사하지만 다른 형태를 갖는 곡선 획들을 서로 다른 코드로 부호화할 수 있다. 일례로 복(卜) 자의 오른쪽 획과 입(入) 자 의 마지막 획은 주행 방향이 서로 다른데 위에서 제시한 방법을 통해 이 둘을 잘 구분할 수 있다(Fig. 6).

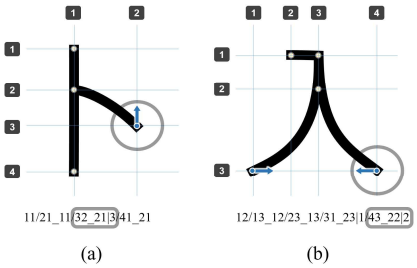

Fig. 6. Shape encoding of Bu (卜) character (a) and Ru (入) character (b).
그림6. 복(卜) 자(a)와 입(入) 자(b)의 자형 부호화. 복 자의 오른쪽 획은 “32_21|3”으로, 입 자의 마지막 획은 “43_22|2”로 부호화된다.

한자의 어떤 지점에서는 여러 개의 곡선이 교차하기도 한다. 이 경우에도 위에서 설명한 방법에 따라 연결되는 결절마다 꺾임 방향을 첨부해 주면 글자의 형태 정보를 빠짐없이 나타낼 수 있다. 예를 들어 Fig. 7에서 여(女) 자의 아래 중간 부분에는 두 개의 곡선 획이 교차하는 지점이 있는데 우상방 결절(그림의 2행 6열 결절)로부터 내려오는 획은 0도 방향으로 비틀림을 받고 있고 좌상방 결절(그림의 3행 2열 결절)로부터 내려오는 획은 90도 방향으로 비틀림을 받고 있다. 이를 부호화하면 “45_32|032|3”으로 표현된다.

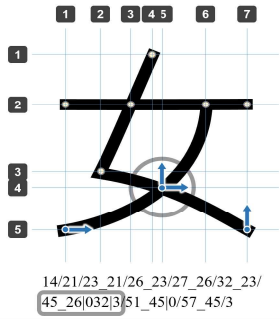

Fig. 7. Shape encoding of Nu (女) character.
그림7. 여(女) 자의 자형 부호 생성.

인간이 사용하는 문자 중에는 곡선이 길게 이어지며 곡률이 여러 차례 바뀌는 글자도 있다. 예를 들어 로마자 알파벳 S가 그러하다. 이런 문자에서 결절 부분에만 부호를 부여하면 글자 또는 필획의 전체 형태를 제대로 표현할 수 없는 경우가 생긴다. 이 문제를 해결하기 위해 필획의 주행 방향이 곡선 주행 시작점을 기준으로 90도 이상 바뀌었을 때는 그 지점에 좌표점, 즉 결절을 새로이 지정하는 것이 필요하다. 을(乙) 자의 두 번째 획은 주행 방향이 크게 회전하는 부분을 가지고 있는데 이곳에 중간 결절을 하나 추가해야 한다(Figure 8). 곡선 주행 중 추가되는 결절은 밑줄(_) 대신 줄 표(-)로 좌·상방 결절과의 관계를 표시하여 곡선 주행부의 결절임을 나타낸다.

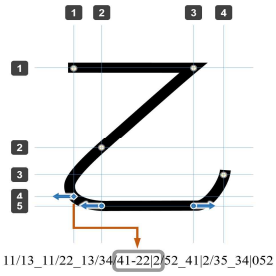

Fig. 8. Shape encoding of Yi (乙) character.
그림8. 을(乙) 자의 자형 부호 생성. 4행 1열에 곡선 주행부 중간에 결절이 추가되어 있다.

매우 드물지만 한자의 곡선 필획 가운데 결절, 즉 시작점, 종지점, 분기점 및 꺾인 점을 가지고 있지 않은 곡선 필획이 존재한다. Fig. 9에 나타낸 강(㪚) 자의 아래 획(한글의 이응에 해당하는)이 그러한 예다.

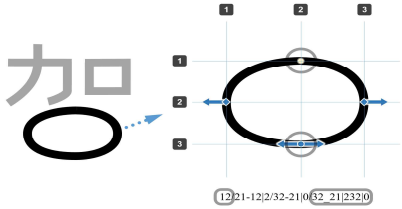

Fig. 9. Shape encoding of bottom stroke of Kang (㪚) character.
그림9. 강(㪚) 자 아래 획(O)의 자형 부호 생성.

이러한 폐곡선 획에서는 필획의 최상방 지점과최하방 지점에 결절을 각각 하나씩 부여한다.

3) 종횡비

점, 선의 배치와 연결이 동일한데도 불구하고 서로 다른 한자로 간주되는 경우가 있다. Fig. 10에 나타낸 왈(曰) 자와 일(日) 자가 그러하다. 이 두 글자는 순전히 종횡비(縱橫比)에 의해서만 구분 가능하다.

이처럼 종횡비를 특정해 주어야 하는 한자 또는 한자 구성 요소는 그 글자의 마지막 결절에 역사선(\)을 추가하고 종횡비를 표시한다. 세로로 긴 글자는 0을 가로로 긴 글자는 1을 부여한다. 종횡비가 의미를 갖지 않는 대부분의 한자는 이러한 정보를 기입하지 않는다.

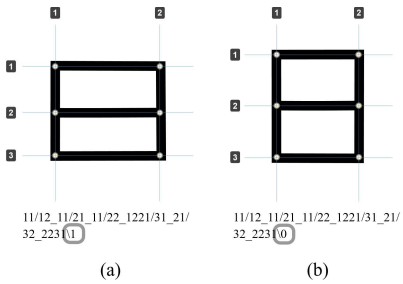

Fig. 10. Shape encoding of Yue (曰, a) and Ri (日, b).
그림10. 왈(曰)과 일(日)의 자형 부호 생성. 두 글자는 종횡비에 의해서만 구분 가능하므로 끝에 종횡비 표시 코드를 추가해야 한다.

4) 결절 위치의 세분화 수준

손으로 쓴 한자에서는 당연한 문제겠으나, 인쇄된 문자에서도 한자의 결절 지점은 한 줄로 정렬되지 않는 경우가 많다. 즉 하나의 세로획 또는 하나의 가로획 안에 존재하는 결절이 아닌 한 행이나 열이 일치하는 결절은 많지 않다. 우리의 인지 과정에서 동일한 형태의 한자로 인지되는 한자를 유일한 코드로 부호화하기 위해서는 결절 위치의 미세한 차이를 되도록 제거하여 가능한 한 적은 수의 좌표선을 통해 한자가 부호화되도록 해야만 한다.

업(業) 자의 경우 이 글자를 이루는 4개의 가로선은 모두 약간씩 다른 길이를 가지고 있어 그 끄트머리(결절)가 하나의 세로 좌표선 위에 정렬되지 않는다. 그러나 이들 가로 획의 장단이 글자를 변별하는 데 의미를 가지고 있는 것일까? 그렇지 않다. Fig. 11에 나타낸 바와 같이 표준형(a)에서 벗어나 (b)와 같은 형태로 이 글자를 변형해도 우리는 이를 업(業) 자로 인지한다. 이처럼 결절의 상호 위치 비교가 문자 식별에 의미를 갖지 않을 때는 해당 결절을 모두 하나의 좌표선 상에 정렬한다.

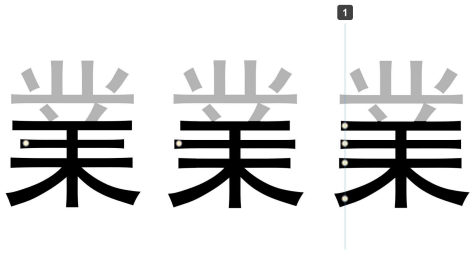

Fig. 11. Recommended allocation rule for nodes and coordinate lines. As far as current Chinese character system is concerned, (a) and (b) is identical, so it is good for character shape encoding that left five nodes in Ye (業) character (c) align in one coordinate line.

그림11. 형태 구성이 동일한 한자들이 유일한 코드로 부호화되기 위해서는 가급적 적은 수의 좌표선을 통해 결절의 좌표가 정해져야 한다. 업(業) 자의 표준적 자형은 (a)와 같으나 (b)와 같이 가로 획의 길이 비율을 바꾸어도 우리는 이 글자를 업(業) 자로 인지한다. 이처럼 상호 위치 비교가 무의미한 결절들은 모두 같은 좌표선에 정렬(c)하는 것이 좋다.

그러나 사(士) 자와 토(土) 자(Figure 12) 또는 말(末) 자와 미(未) 자와 같이 획의 상대적 길이 비교가 글자의 변별에 필수적인 경우가 있다. 이 경우에는 결절의 좌표를 구분할 수 있는 만큼의 좌표선을 부여한다.

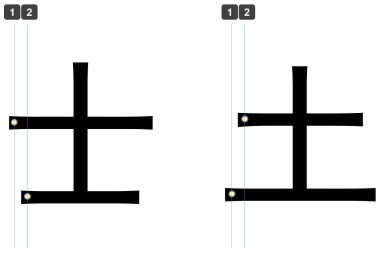

Fig. 12. Coordinate lines for Shi (士) character (a) and Tu (土) character (b).

그림12. 사(士) 자와 토(土) 자의 가로획 양단은 서로 다른 좌표선에 정렬한다.

5) 다수의 구성 요소로 구성된 한자의 자형 부호화

이상에서 설명한 방법으로 어떠한 한자라도 그 자형을 나타내는 부호(코드)를 생성할 수 있다고 생각된다. 그러나 한자 사용자가 동일한 자형으로 간주하는 여러 형태의 사례가 상이한 코드로 부호화되거나 한자 자형을 부호화하는 과정에서 사람의 자의적 판단에 의해 동일한 자형에 대해 상이한 코드가 생성되면 안 될 것이다. 아쉽게도 위에서 제안한 방법으로 한자의 자형을 부호화해 보면 자의적 판단을 내릴 수밖에 없는 상황에 도달하게 된다. 이는 두 개 이상의 자형 요소(character component, 서로 연결된 필획의 집합)로 구성된 한자에서 발생하는 문제인데, 예를 들어 명(明) 자의 경우 이 글자를 구성하는 부분인 일(日) 자를 월(月) 자에 대해 좀 낮게 배치하여 그 자형을 인코딩할 수도 있고(Figure 13의 (a)) 같은 높이로 배치하여 인코딩할 수도 있다(Figure 13의 (b)). 그 밖의 배치도 가능하다. 이 경우 서로 다른 코드가 생성되어 버리는데 이 가운데 어느 것을 명(明) 자의 자형 코드로 채택할 것인지 결정하기 어렵다. 더구나 더욱 복잡한 구성을 갖는 다수의 한자에서 다른 자형 요소에 속한 결절들 사이의 위치 관계를 자의성 없이 결정하는 것은 보다 어렵다.

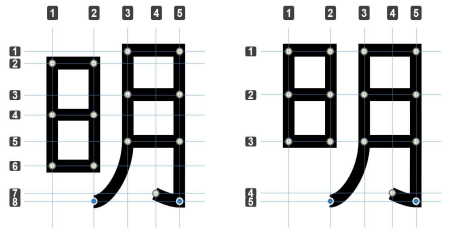

Fig. 13. A problem in encoding the Ming (明) character by suggested procedure. Both (a) and (b) are possible encoding.

그림13. 제안된 방법으로 명(明) 자의 자형을 부호화할 발생하는 문제. 부호화하는 사람의 주관에 따라 일(日) 자 부분을 낮게 할 수도(a), 높게 할 수도(b) 있는데, 이에 따라 생성된 자형 부호는 완전히 다르게 된다.

이는 인간이 한자의 자형을 인지하고 기억할 때 자형 요소 간의 대략적 위치 관계만을 감지하고 기억할 뿐 서로 멀리 떨어진 결절들의 세세한 상하좌

우 관계에 대해서는 주목하지 않기 때문이다. 따라서 이러한 인간의 인지적 구조에 맞는 자형 부호화가 필요하다.

필자는, 서로 연결된 필획의 집합, 즉 자형 요소(character component)에 대해서는 위에서 제시한 부호화 방법을 적용하되 자형 요소와 자형 요소의 관계에 대해서는 그들 사이의 형태적 관계를 규정하는 3가지 요소, 즉 상호 위치, 상대적 크기, 중첩 여부를 기입하는 방법으로 한자 자형을 인코딩할 것을 제안한다.

먼저 상호 위치 코드는 Fig. 14에 설명한 바와 같이 정한다. 후행 요소를 기준으로 선행 요소가 좌상단에 위치하면 1을, 상단에 위치하면 2를, 우상단에 위치하면 3을, 좌측에 위치하면 4를 부여하고 선행 요소가 후행 요소를 감싸고 있으면 0을 부여한다.

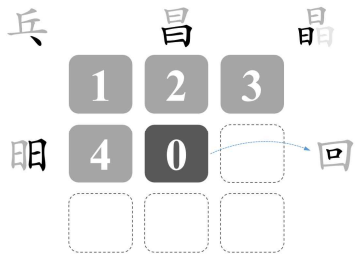

Fig. 14. Codes for relative position between character components.

그림14. 한자 구성 부분 사이의 위치 관계를 나타내는 부호의 결정 방법. 선행 부분이 좌상단에 있으면 1, 상단에 있으면 2, 우상단에 있으면 3, 좌측에 있으면 4를 부여한다. 선행 부분과 해당 부분의 중심이 겹치면 0을 부여한다.

이어서 상대적 크기를 나타내는 코드는 Fig. 15에 나타낸 방법으로 정한다. 후행 요소를 기준으로 선행 요소의 크기가 그와 대등하면 1을, 선행 요소가 확연하게 크다면 2를, 선행 요소가 확연하게 작다면 3을 부여한다.

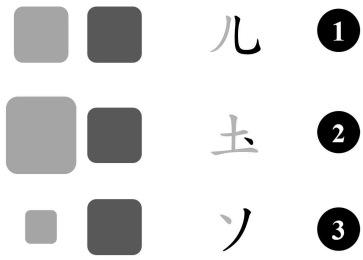

Fig. 15. Codes for relative size between character components.

그림15. 한자 구성 부분 사이의 상대적 크기를 나타내는 부호의 결정 방법. 선행 부분이 대등한 크기면 1, 비교되는 부분보다 크면 2, 작으면 3을 부여한다.

마지막으로 자형 요소 사이의 중첩(포함) 여부를 지시하는 코드는 Fig. 16에 도시한 것과 같이 중첩이 없으면 1, 중첩이 있으면 2를 부여한다. 여기서 중첩이 있다는 것은 선행 요소의 중심(무게중심)이 후행 요소의 외곽 결절을 직선으로 이은 다각형 안에 위치한다는 뜻이다.

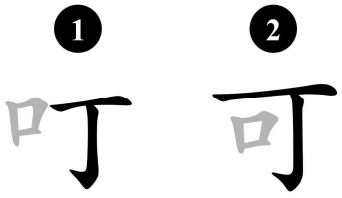

Fig. 16. Codes for overlapping information between two character components.

그림16. 한자 구성 부분의 중첩 여부를 나타내는 부호. 선행 부분과 중첩이 없으면 1, 중첩이 있으면 2를 부여한다.

이러한 3종류의 상호 관계 코드를 이용하여 자형 요소의 코드를 결합하면 한자의 자형 코드가 완성된다.

Figure 17에서 국(國)자를 예로 들어 자형 요소의 분해와 상호 관계 코드 설정 과정을 설명한다. 먼저 국 자의 자형 요소를 구분해야 하는데, 서로 연결되지 않는 필획 집합을 나누어 보면 5개의 요소로 나뉨을 알 수 있다. 이어 이들 자형 요소의 순서를 정해야 하는데, 여기에는 상부 요소가 먼저,

동일 수평선에 존재할 경우 좌측 요소가 먼저라는 원칙이 적용된다. 외곽을 둘러싼 요소가 있어 선행 요소와 후행 요소의 중심이 겹칠 때는 외곽을 둘러싼 요소를 먼저 인코딩한다. 또한 복수의 자형 요소가 하나의 선행 자형 요소와 한꺼번에 관계를 맺는 경우가 있는데(예를 들어 國에서 口과 或의 관계) 이러한 경우에는 후행 요소를 묶어서 선행 요소와의 관계를 나타내는 코드를 부여한다. Fig. 17에서 국(國) 자는 (a)~(d) 순서로 분해되며 그림에 설명한 바와 같이 선행 요소와 관계 코드가 설정된다.

이 방법을 적용한 실제 자형 인코딩 사례는 본론의「라) 적용실례」에서 설명한다.

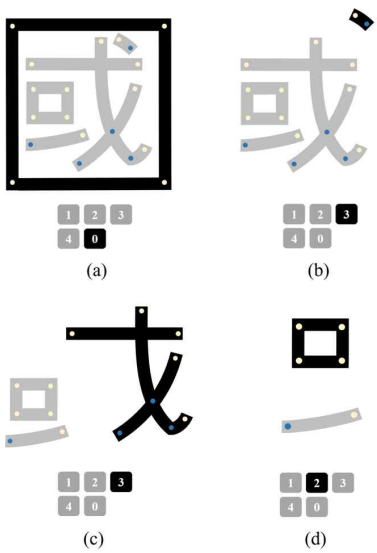

Fig. 17. A dividing example of Chinese character for shape encoding.

그림17. 한자 구성 부분의 분해 사례. 서로 연결된 필획을 하나의 요소로 간주할 경우 국(國) 자는 5개의 요소로 나뉜다. 에운담(口)과 혹(或)은 중심이 겹치고(위치 관계 코드 0 부여, a) 크기는 혹 자가 작다. 혹은 다시 점(丶)과 나머지 부분으로 나뉘는데 후자에 대해 점은 우상방에 위치하고(위치 관계 코드 3 부여, b) 크기는 후자가 크다. 후자의 부분은 다시 (c)와 같이 분해되며 상호 위치 코드는 3, 상호 크기 코드는 2가 부여된다. 마자막으로 구(口) 자와 비스듬한 가로획으로 구성된 부분은 (d)와 같이 분해된다. 상호 위치 코드는 2, 상호 크기 코드는 1이 부여된다. (a), (b), (c)는 구성 요소 사이에 중첩이 존재하며 (d)는 중첩이 존재하지 않는다. 이에 따라 (a), (b), (c)에는 중첩 여부 코드 2를 (d)에는 중첩 여부 코드 1을 부여한다.

6) 점유 면적과 편위(偏位)

한자 자체에는 해당되지 않는 사항이지만, 문장 부호나 구결 문자 가운데 글자 크기를 다른 문자보다 작게 표기해야 하거나 일반적인 글자가 차지하는 직사각형 구역의 일부만을 점유하도록 표기해야 하는 문자가 있다. 예를 들어 중국어 문장의 마침표(“。”, 句號)는 원(○)과 다른 크기로, 행의 아래쪽에 치우치게 표기해야만 한다(중화민국의 경우에는 중앙). 또한 세계 각국의 문자 가운데 크기를 달리하거나 편위(偏位)된 문자를 사용하는 예가 상당수 존재하므로 향후 본 논고에서 제안한 자형 부호화 방법의 적용 범위를 확장하려면 문자의 점유 면적 비와 편위 형태에 대한 부호화 방식을 고안할 필요가 있다.

7) 필획 장식(serif)

인쇄 매체에 등장하는 다양한 글꼴을 보면 필획의 기시점, 종지점과 필획이 꺾이는 부분에서 다양한 장식 요소가 사용된 것을 볼 수 있다. 이를 세리프(serif)라고 하는데, 본 논고에서 제안하는 자형 부호화 과정에서 이런 장식요소는 모두 무시하는 것을 원칙으로 한다. 하나의 자형 코드에서 여러 서체가 생성될 수 있는 만큼 각 서체별로 달리 만들어질 필획 장식을 자형 부호에 포함시킬 수는 없기 때문이다.

그러나 어떤 것이 장식 요소이고 어떤 것이 자형 구분에 유효한 필획인지 구분하기 어려운 경우가 있을 수 있는데 이에 대해서는 추후 심도 있는 검토를 통해 구분 방식이 확립되어야 할 것이라 생각한다.

8) 정보 갱신 웹페이지 운영

본 논고에서 제안한 한자 자형 부호화 방법은 한자에 대한 지식이 없는 사람도 한자의 자형을 하나의 코드로 전환하는 데 목적을 두고 제안한 것이지만 실제로 수많은 사람들이 이 방식을 통해 한자의 자형 코드를 입력하다 보면 동일한 형태의 글자인데도 서로 다른 코드가 부여되거나 서로 다른 형태의

글자인데도 같은 코드가 부여되는 사례가 나타날 수 있으리라 생각된다. 따라서 이들에 대한 감별과 정리를 지속적으로 해나갈 정보 공유 매체가 필요하다.

한자의 자형 정보를 한 곳에서 관리할 자형 부호 등록 웹사이트를 개설, 운영하는 것이 바람직하리라 생각된다.

라) 적용 실례

이제 고문헌에 출현하는 벽자(僻字)를 부호화하는데 제안한 방식을 사용하는 것을 실례를 통해 설명해 보기로 한다.

규장각 소장『동의보감』목판본에서 쓸개에 대해 설명한 부분을 펼쳐 보면 Fig. 18과 같은 담(膽) 자가 보인다.

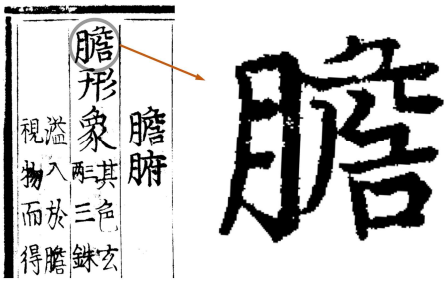

Fig. 18. Tan (膽) character in *Dongyibogam*.
그림 18.『동의보감』목판본에 보이는 담(膽) 자.

이 글자는 담(膽) 자의 이체자임이 분명하지만 유니코드 확장판에도 등록되어 있지 않으며 한자 자형에 관한 방대한 정보를 담고 있는「글리프위키」에조차 수록되어 있지 않다. 이러한 글자를 전자매체에 검색 가능하게 하려면 이 글자에 새로운 문자 코드(유니코드)를 부여하는 것을 생각할 수 있으나 그렇게 하려면 문자가 새로 등록될 때까지 많은 시간 지체가 있고 사소한 변형을 가진 이체자들에 모두 다른 문자 코드를 부여하는 것은 현재의 유니코드 시스템에서 지양하고 있으며 경우에 따라 등록하려는 한자의 정체(미지의 의미나 용법이 있는지, 단순히 誤記에 의해 만들어진 것인지 등)를 알 수 없는 경우도 있기에 문헌을 다루는 현장에서 일률적으로 적용할 수 있는 방법은 아니다. 따라서 순수하게 자형 정보만을 일정한 방법에 따라 신속하게 등록하고 공유할 필요가 있으며 본 논고에서 제안한 방법이 이 목적에 유용하게 사용될 것이라 본다.

Figure 19에 이 글자의 자형 정보를 부호화하는 과정을 차례로 보인다. 먼저 글자의 결절(node)위치를 확정한다(a). (b)는 세선화(細線化, thinning)를 거친 모습이다. 이어 필획의 연결 상태를 점검하여 분리할 곳을 분리하고 연결할 곳은 연결한다(c). 이 과정에서 한자에 대한 사전 지식이 다소 필요하나 이 과정을 도와줄 소프트웨어가 개발된다면 비교적 쉽게, 일정한 방식으로 이 과정을 수행할 수 있을 것이다. 이러한 자형 정비가 완료되면 각 자형 요소를 분리한다(d). 분리된 각 자형 요소 각각에 대해 본 논고 본론의「제안」단락 1)~5)절에서 설명한 방법에 따라 코드를 부여한다(e). 이어 자형 요소 사이의 관계를 나타내는 코드(f에서 굵은 고딕체 숫자로 나타낸 것)를 정하고 요소의 층차에 맞추어 결합({}를 이용)한 후 연결한다.

이런 과정을 거치면 담(膽) 자의 자형 코드 "{12/14_11/22_12/24_1422/32_22/34_2432/43/51_32|0/54_3443|2}411{{14/22_14/27_22/32_23|0/42/45_42/46_27/47_46/49_47/59/63_45|0/67_47/68_59|0/72_42/81_72|0}122{{12/13_12/21/23_13/24_23}211{11/12_11/21_11/22_1221}}}"을 얻게 된다.

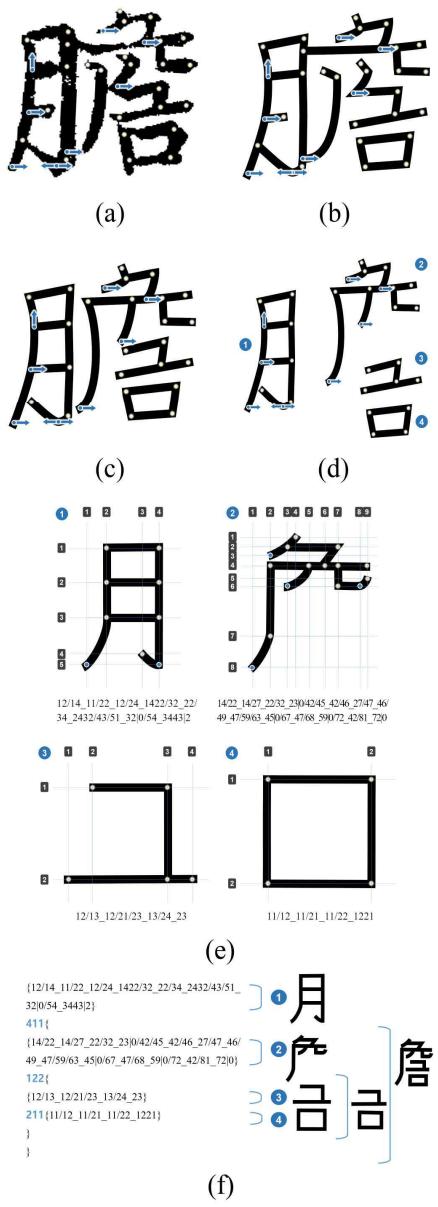

Fig. 19. Encoding procedure of Tan (膽) character.
그림19. 담(膽) 자의 자형 부호화 과정.

마) 문자 코드와 자형 코드의 관계

현존 한자 중에는 수많은 이체자가 존재하는데 이들 가운데 일부 문자에 새로 유니코드가 부여되고 있다. 그러나 모든 이체자에 새로운 코드를 부여하는 것은 아니며 기존 한자의 사소한 변형 결과라 생각되는 글자는 별도 코드 없이 기존의 코드를 사용하도록 하고 있다. 이 때문에 국가에 따라, 그리고 사용자가 사용하는 글꼴 파일에 따라 같은 코드의 한자가 꽤 다양한 형태로 표시되고 있는 실정이다.

현재 하나의 한자 코드에 소속된 다양한 자형의 한자들은 따로 분류, 등록되어 관리되지 못하고 있다. 이들 한자 각각에 문자 코드와 별개의 식별 코드를 부여(Fig. 20)한다면 한자 자형을 관리하는 데 큰 도움이 될 것이다.

나아가 하나의 자형 코드에서 다양한 서체를 자동 생성하는 소프트웨어가 개발된다면 거대한 저장 공간을 사용할 필요 없이 필요에 따라 다양한 서체로 표현된 한자를 빈용자(頻用字), 벽자(僻字) 여부에 무관하게 컴퓨터에서 자유롭게 사용할 수 있게 될 것이다.

Ⅲ. 결론

본 논고에서는 한자의 자형을 부호화하기 위한 한 가지 방법을 제안하였다.

이 방법은 기본적으로 한자를 구성하는 필획의 결절, 즉 기시점(起始點), 종지점(終止點), 연결점(連結點)과 필획이 꺾이는 점을 “결절”로 규정하여 이들의 상대 좌표와 이들 사이의 관계를 숫자로 나타내는 방식이다.

여기에 곡선 필획의 형태 정보를 부호화하기 위한 몇 가지 규정과 글자 또는 글자 요소의 종횡비를 표기하는 규정, 결절의 위치를 표시하는 기준이 되는 좌표선 수의 최소화를 위한 규칙과 자형 요소의 상호 관계를 표시하는 규칙을 추가로 제시하여 현존하는 모든 한자를 부호화 가능하도록 하였다.

본 논고에서 제안한 방법은 면적을 가진 필획으로 구성된 문자, 여러 종류의 선(실선, 파선 등)으로 이루어진 문자, 세선화(細線化)된 자형에는 표시되지 않는, 획의 주행 방향이 의미를 갖는 문자 등을 제외하면 세계의 대다수 문자에 그대로 적용할 수 있다. 지속적으로 확장되고 있는 유니코드를 체계적이

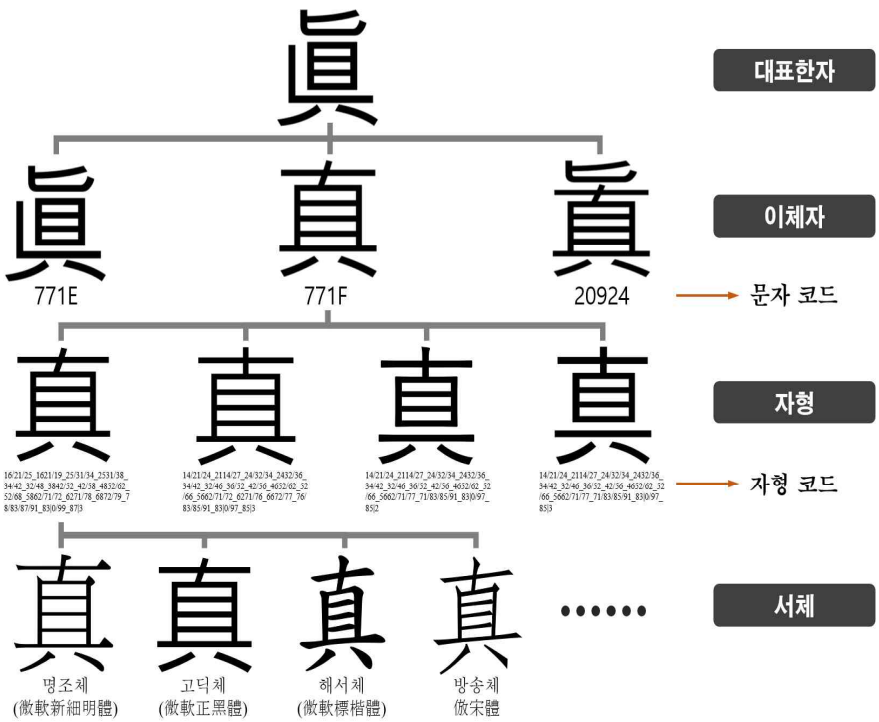

Fig. 20. The hierarchy of representative characters, identical characters, morphological variation of characters and individual fonts. To distinguish the morphological variations of Chinese characters a new encoding system is required apart from current character code system (Unicode).

그림20. 대표한자, 이체자, 자형, 서체의 관계. 그림에 보이듯 진(真, 유니코드 771F) 자의 구체적 형태는 여러 가지가 존재한다. 이들을 구분하기 위해서는 문자 코드와 별도로 자형 코드가 필요하다.

고 효율적으로 관리하는 데도 본 논고에서 제안한 방법이 도움을 줄 수 있을 것이다.

감사의 글

이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

References

1. Septime Auguste Viguier. New book for the telegraph. Shanghai. department of electric machinery. 1872.
威基謁. 電報新書. 上海. 電機司. 1872.
2. Wangyunwo. A Chinese Character searching method by codes. Dongfangzazhi. 1925. 22(12). pp.82-98.
王雲五. 號碼檢字法. 東方雜誌. 1925. 22(12):82-98.