ISO 14651에 의한 한글 ordering의 문제점과 그 해결 방안

김종휘*, 김경석
부산대학교 대학원 멀티미디어협동과정
kjw6305@sungsim.ac.kr, gimgs@hangeul.pnu.edu

Some Problems on the Ordering of Hangeul by ISO 14651 and a Proposal for their Resolutions

Kim, Jong-whee*, Gim, Gyeongseog
Dept. of Multimedia Cooperation Course, Graduate School of Pusan National University

요약

문자열 간추리기(string ordering)에 관한 국제 표준인 ISO 14651의 내용 중 공통틀표(CTT)의 한글 관련 규정은, 첫가끝 조합형과 완성형 어느 쪽으로도 부호화가 가능한 한글 문서의 특성을 무시하여 이들을 분리하여 기술함으로써 두 부호값 체계에 대한 상호 연관성과 통일성을 잃고 있다. 또한 ordering에 필수적인 UCS 완성형 글자마디의 무게값(weight)을 할당하지 않음으로써 형식적 완결성과 내용적 명료성을 잃고 있다. 이에 따라 본 논문은 CTT의 규정을 한글 문서의 부호화 방법에 따라 유형별로 검토하여 그 문제점을 지적하고, 이와 관련하여 어떤 한글 문서이든 이를 일관성있게 ordering할 수 있도록 하는 'CTT 한글 부분의 개정 방안'을 제안함으로써 한글 ordering과 관련된 현 CTT 상의 여러 문제점들을 해결하고자 하였다.

1. 개관

ISO 14651은 ISO(국제표준화기구)/IEC(국제전자기술위원회)에서 세계 모든 언어에 대한 ordering/sorting을 체계적이고 명시적으로 가능하게 하도록 제정한 표준화 규약으로서 그 정식 명칭은 'ISO/IEC 14651 - International string ordering and comparison - Method for comparing character strings and description of the common template tailorable ordering'이다.

ISO 14651 문서는 내용적 범위의 정의, 용어상의 정의, 문자열 비교 방법 등의 내용을 중심으로 한 본문과 몇 개의 부록(annex)으로 구성되어 있는데, 특히 문자열 비교 방법에서는 간추리기(ordering)에 필요한 key 형성과 참조 비교 방법(reference comparison method), 공통틀표(common template table : CTT[1])의 형성과 맞춤(tailoring) 등의 핵심 사항이 다루어지고 있다. 특히 공통틀표(CTT)는 규범적 사항으로서 그 모든 내용이 부록 A에 제시되어 있다.[1]

공통틀표는 ordering을 위해 모든 언어의 문서 자료가 공통적으로 참조할 수 있는 표로서 ISO 14651의 핵심적 참조 규범이다. 내용적으로는, 각 수준(level)별로 간추림 기호(collating symbols)에 대해 정의한 다음, 역시 각 수준별로 이들 심볼에 대해 무게값을 할당(weight assignments)하고 이를 바탕으로, UCS[2]의 모든 부호값에 대한 무게값 목록(weight list)을 명시해주고 있다. 이에 따라, 문자열 간추리기(string ordering)는 각 수준에서 subkey의 무게값(weight)을 비교함으로써 가능해진다.

ISO 14651에 의해 ordering되는 문서는 원칙적으로 UCS로 부호화되어 있어야 한다. 한글 문서에 사용되는 자모와 글자마디도 모두 UCS 내에서 고유한 부호값(code position)을 가지고 있으며[2], 특히 한글 문서는 각각의 글자마디에 대해 첫가끝 조합형과 완성형을 포함한 두 가지 이상의 방식으로 부호화될 수 있어서 그 부호값이 서로 다를 수 있음에 유의해야 한다.[3]

문자열을 간추릴 때, 특별한 제약이나 조건이 없는 경우에는 공통틀표만으로도 올바른 결과물을 얻을 수 있어야 한다. 그러나 한글의 경우, 예를 들어, 첫가끝 조합형과 완성형으로 부호화된 문자열이 동일한 문서 내에 존재하게 되면 한글의 특성을 올바로 반영하지 못하고 있는 현 공통틀표로는 ordering 자체가 불가능하거나 기대한 결과물을 얻을 수 없다.

따라서, 본 논문에서는 논리적 차원의 접근을 통해 ISO 14651의 공통틀표가 지니고 있는 한글 문서의 ordering과 관련된 문제점을 지적하고 그 원인과 해결 방안을 제시해 보고자 한다.

2. 문제의 제기

1) CTT의 정식 명칭은 'ISO14651_2000_TABLE1.txt'로서 URL:http://www.iso.ch/ittf/ISO14651_2000_TABLE1.htm에서 확인할 수 있다.

2) ISO/IEC 10646 표준의 이름은 Universal Multiple-Octet Coded Character Set인데 간략히 UCS라고도 하며, 일반적으로는 Unicode[4]로 통용되기도 한다.

한글 문서와 관련된 공통틀표(CTT)의 주된 문제점은 UCS 부호화에서 첫가끝 조합형과 완성형이 모두 가능한 한글 부호화 체계의 특성을 종합적으로 고려하지 못한 데에서 비롯된다. 각 경우에 따른 문제점은 다음과 같다.

2.1. 첫가끝 조합형 문서

현 공통틀표를 참조할 때, 한글 UCS 문서 내에 U+AC00부터 U+D7A3까지의 한글 완성형 부호값에 의한 문자열이 들어있지 않으면 아무런 문제가 발생하지 않는다. 즉, 첫가끝 조합형 또는 호환적 동등(compatability equivalence) 부호값[5]으로 표현된 한글 문서에서는 기대하는 ordering 결과값을 별 문제 없이 얻을 수 있다.

2.2. 완성형 문서

현 공통틀표에서 완성형 부호값의 한글 문서를 ordering하는 데에는 몇 가지 문제점이 있다.

우선, 1st-level collating symbol로 이미 정의된 <SAC00> ~ <SD7A3>에 대해 단순 무게값(simple weight)을 할당하여야 한다. 공통틀표에서 이 단순 무게값을 할당하지 않은 것은 첫가끝 조합형과의 충돌을 염려한 때문으로 추측되는데, 중요한 것은 완성형 문서를 ordering하기 위해서는 어떠한 경우라도 이 단순 무게값 부분이 없어서는 안 된다는 점이다.

<SAC00> ~ <SD7A3>이 삽입될 곳은 1-st level 내에서는 임의의 장소라도 가능하나 편의상 첫가끝 조합형 자모 뒤에 두면 그 전후를 포함한 형태는 다음과 같다.

```
:
<S11F9> % HANGUL JONGSEONG YEORINHIEUH
<SAC00>..<SD7A3> % Hangul syllables
<S30A2> % KATAKANA LETTER A
:
```

이와 더불어, 각 완성형 부호값에 대한 기호 무게값(symbol weight)의 list를 할당해 두어야 한다. 예를 들어 '우리'라는 문자열(string)을 다른 것과 비교하여 ordering하기 위해서는 '우리'의 무게값을 확인하여야만 하는데, 이 때, 글자마디 '우'와 '리' 각각의 무게값 목록(weight list)이 반드시 필요한 것이다. 그러나 현 공통틀표에서는 한글 완성형 부호값에 대한 무게값 목록이 빠져 있고 다만, 완성형을 ordering할 때 이 무게값 목록을 활용할 수 있다는 코멘트만 제시되어 있을 뿐이다. 따라서,

'%<UAC00>..<UD7A3>
<SAC00>..<SD7A3>;<BASE>;<MIN>;<UAC00>..<UD7A3>'를 decomment하여

'<UAC00>..<UD7A3>
<SAC00>..<SD7A3>;<BASE>;<MIN>;<UAC00>..<UD7A3>'로 바꿔 주어야 한다.

이상과 같이, 완성형 문서의 ordering에 대한 사항을 코멘트로만 처리해 두어, 실제 ordering에 필요한 각 글자마디 부호값의 무게값을 구할 수 없게 한 것은 공통틀표의 문제점이라 할 수 있다. 공통틀표가 모형적 규범으로서 모든 UCS 부호값을 ordering할 수 있도록 규정하는 것이라면, 한글 완성형에 대해서만 기본적으로 어떤 ordering도 불가능하게끔 제한하여 기술하고 있는 현 공통틀표는 논리적 모순을 지닌 것이다.

한글 완성형 문서를 ordering하기 위해 모든 한글 완성형 부호값의 단순 무게값(simple weight)과 기호 무게값 목록(symbol weight list)을 공통틀표에 첨가하는 것은 공통틀표를 맞춤(tailoring)하는 것이 아니라 공통틀표를 새롭게 바꾸는 것이다. 맞춤은 ordering 대상에 대해 부분적 또는 제한적 조작으로서 가할 수 있는 변형일 뿐이다.

2.3. 첫가끝 조합형과 완성형이 혼합된 문서

현 공통틀표에서는 혼합형 문서를 ordering할 수 있는 방법[3]으로서 완성형 부호값을 첫가끝 조합형 부호값으로 바꾸어 놓는 것[4]에 대해 코멘트로서 기술하고 있다. 즉, 각 완성형 글자마디를 U+U1100부터 U+11F9 사이의 첫가끝 조합형 자모로 분해하고 이 때의 부호값에 따라 무게값을 설정할 수 있다는 것이다. 예를 들어 '각 U+AC01'은 'ㄱ U+1100, ㅏ U+1161, ㄱ U+11A8'로 분해되고, 이에 따라

<UAC01>
"<S1100><S1161><S11A8>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";"<U1100><U1161><U11A8>" 로 바뀌는 것이다.

이와 관련하여, 혼합형 문서를 올바로 ordering하기 위해서는 완성형 문서만을 위한 '<SAC00> ~ <SD7A3>와 관련 무게값 할당'을 공통틀표 내에 포함시켜서는 안 된다는 사실도 중요하다. 이들은 첫가끝 조합형과 조화를 이루지 못하고 오히려 중복과 혼란을 야기하여 전혀 예기치 않은 결과를 가져오기 때문이다. 달리 말하면 혼합형 문서의 올바른 ordering을 위해서는, 완성형 글자마디를 U+U1100부터 U+11F9 사이의 첫가끝 조합형 자모로 분해하여 그 부호값에 따라 무게값을 설정하는 방식만이 유효할 뿐이다.

더구나 이러한 첫가끝 조합형 변환 방식은 '<SAC00> ~ <SD7A3>의 무게값을 할당하기' 방식을 포괄하게 된다. 즉, 완성형 부호값을 첫가끝 조합형 부호값으로 바꾸는 것만으로 모든 한글 문서의 ordering이 가능해지므로 굳이 공통틀표 내에서 완성형만을 위한 사항(<SAC00> ~ <SD7A3>의 무게값 할당)에 대해 거론할 필요가 더 이상 없어지게 되는 것이다.

이와 같은 몇 가지 문제에도 불구하고 굳이 완성형만을 위한 '<SAC00> ~ <SD7A3>와 관련 무게값 할당'을 코멘트에 포함시킨 것은, 공통틀표가 한글에 대한 충분한 이해와 검토 없이 작성된 것이 아닌가 하는 의심을 가지게 하는 점이다.

---

3) 공통틀표(CTT)는 한글 완성형 부호값에 대해 1) 완성형 글자마디를 첫가끝 조합형 자모로 변환하기 2)완성형 부호값에 weight 할당하기라는 두 가지 ordering 방법을 동시에 제시하고 있다.

4) 실제적인 orderig 절차로서는 프로그래밍에 의해 원래의 부호값을 첫가끝 조합형으로 미리 바꾸어 둠으로써 문제를 근원적으로 해결할 수도 있다.

이러한 문제점에 덧붙여, 틀(template) 문서로서의 공통틀표 형식 또한 재론의 여지가 많다. 완성형 부호값의 문자열 문서와 관련하여 완성형 부호값을 첫가끝 조합형 부호값으로 바꾸면 올바른 ordering이 가능하다고 코멘트로만 지적하는 데 그쳐, 완성형 문서의 ordering에 대한 방법적 표본을 공통틀표 자체적으로 제시해주지 못하고 있다. 이러한 기술은 틀이라기보다는 해결 방법(solution)에 가까운 것으로서 공통틀표의 형식으로는 부적절하다 할 것이다.

3. CTT 한글 부분의 개정 방안

앞서 살펴본 바와 같이, 한글과 관련된 공통틀표(CTT)의 주요 문제점은 첫가끝 조합형과 완성형을 통합적으로 처리할 수 있는 방법을 제시하지 못하고 있다는 점과 한글 관련 조항을 틀(template) 문서로서의 형식에 적합하게 기술해주지 못하고 있다는 점이다. 그리고 이들 두 사항은 서로 관련된 문제점이기도 하다.

본 논문은 한글 부호값의 유형에 관계없는 통합적 ordering의 방법을 제시하고 또한 이를 통해 현 공통틀표 문서의 형식적 간결성을 제고해 보고자 한다. 이제부터는 이 제안을 '공통틀표 한글 부분의 개정 방안"이라고 하기로 한다.

우선 현 공통틀표에서 <SAC00>부터 <SD7A3>까지 collating symbol로 정의하고 이에 대해 코멘트한 부분을 다음과 같이 간략히 한다. 즉,

'collating-symbol <SAC00>..<SD7A3> % Symbols for Hangul (weights must be constructed)'을

'collating-symbol <SAC00>..<SD7A3> % Symbols for Hangul'로 수정한다.

한글 완성형 부호값에만 기술된 '(weights must be constructed)'는 '공통틀표 한글 부분의 개정 방안'에서는 더 이상 불필요해지며 이에 따라 공통틀표도 형식적 통일성을 띠게 된다.

이와 함께 현 공통틀표 끝부분의 한글 완성형 부호값에 대한 소리마디 무게값(syllable weight) 할당을 삭제하고 그 대신 각 소리마디를 첫가끝 조합형 부호값으로 환원한 소리마디 무게값 목록(syllable weight list)을 할당한다. 보기를 들면,

'% <UAC00>..<UD7A3>

<SAC00>..<SD7A3>;<BASE>;<MIN>;<UAC00>..<UD7A3> % Hangul

% A Hangul tailoring for a system which does not use combining jamos may choose simply weight the Hangul syllables directly as shown above.'를

'<UAC00>

"<S1100><S1161>";"<BASE><BASE>";"<MIN><MIN>";"<U1100><U1161>" **% Hangul 가 ga**

<UAC01>

"<S1100><S1161><S11A8>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";"<U1100><U1161><U11A8>" **% Hangul 각 gag**

⋮

<UD7A3>

"<S1112><S1175><S11C2>";"<BASE><BASE><BASE>";"<MIN><MIN><MIN>";"<U1112><U1175><U11C2>" **% Hangul 힣 hih'**로 수정한다.

이렇게 함으로써 한글 문서에 포함된 완성형 문자열도 ordering에 필요한 무게값을 모두 첫가끝 조합형으로 환원된 값으로 가지게 되어 다른 첫가끝 조합형 문자열과 아무런 혼란을 일으키지 않는다. 예를 들어, 완성형 '가구'를 첫가끝 조합형 '가보', 'ㄱㅜㄹ'과 함께 ordering하더라도 우리의 직관에 부합하는 결과를 얻을 수 있다.

그러나 현 공통틀표로는 이러한 혼합형 문서를 어떻게 해서든 ordering한다 하더라도 첫가끝 조합형 '가보', 'ㄱㅜㄹ'이 완성형 '가구'보다 앞서게 된다.

4. 마무리

현 공통틀표 문서는 한글과 관련하여 불필요하게 복잡한 제약을 많이 가하고 있다. 이에 본 논문은 '공통틀표 한글 부분의 개정 방안'을 제안하여 그 문제를 해결하고자 하였다.

완성형 부호값의 문서를 ordering할 때에는, 공통틀표에 delta를 추가로 정의하는 맞춤(tailoring)이 많이 필요할 뿐 아니라 그 결과값도 첫가끝 조합형이 공존할 경우에는 신뢰할 수 없다. 이는 ISO 표준 문서로서 내용적 형식적 측면에서 문제점으로 지적될 수 있다.

'공통틀표 한글 부분의 개정 방안'은 한글 문서 내에서 완성형 부호값으로 표현된 문자열에 대해 그 글자마디의 무게값 목록을 해당 첫가끝 조합형 부호값의 ordering key로 환원하여 설정함으로써 이러한 문제들을 해결하였다.

참고문헌

[1]. ISO/IEC JTC1/SC 22WG, *ISO/IEC 14651 - International string ordering and comparison*, ISO/IEC JTC1/SC 22WG 20 N 731, 2000

[2]. The Unicode Consortium, *The Unicode Standard, version 3.0*, Addison-Wesley, 2000

[3]. 김경석, *컴퓨터 속의 한글 이야기-둘째 보따리*, 부산대학교 출판부. 1999

[4]. "Unicode Collation Algorithm", *Unicode Technical Standard* #10, http://www.unicode.org/unicode/reports/tr10

[5]. "Unicode Normalization Forms", *Unicode Standard Annex* #15, http://www.unicode.org/unicode/reports/tr15