한국어정보처리를 위한 한글 부호계의 적합성 평가

변 정 용
동국대학교 자연과학대학 전자계산학과
경북 경주시 석장동 707(우 780-350)

요 약

한글 부호계의 올바른 사용과 이들의 표준안 채택은 한국어정보처리 제분야에서의 적합성과 한글 문화창달에 기여성등을 먼저 검토해야 할 것이며, 그리고 한글 부호 제정사에서 갖는 위상과 정보교환 및 정보처리 기술에 어떤 영향을 끼칠 것인지에 관한 학문적 평가가 요구된다. 이러한 결과는 궁극적으로 한국어정보처리에 있어서 그 한계성의 제거와, 새로운 컴퓨터 기술의 개발에 원동력을 제공할 것이다.

1. 서론

한글 부호계의 개선에 관한 연구는 한글의 자소정보를 표현하는 낱자형 부호계를 근본으로 51자 부호계 또는 33자 부호계의 제안과 그에 따른 중요한 관련 알고리즘등의 제안이 이루어 진 바 있다.3)6)7)

본 논문은 한글 낱자형 부호계를 중심으로 조합형 부호계 및 완성형 부호계에 대하여 한국어정보처리와 정보교환 및 정보처리의 기술적 측면을 일련의 평가요소로써 조명하여 낱자형 부호계의 적합성을 밝히고, 이를 바탕으로 한국어정보처리에 최적한 한글부호계로의 개정을 추진하는 데 있어서 학문적 근거자료를 제공하는 데 그 근본적 목적이 있다.

2. 적합성 평가

평가 대상 부호계는 낱자형, 조합형, 완성형 부호계이다. 낱자형은 33자와 51자 중에 51자의 경우만 적용하고 33자의 경우는 51자와의 비교에서 33자가 우위인 부분을 별도로 평가하면 될 것이므로 본 논문에서는 다루지 않는다. 본 평가의 전제적인 관점은 "컴퓨터를 이용하여 한글을 처리한다"는 입장이다. 평가 요소는 한글 부호계의 8개항의 요구도를 중심으로 세가지 부호계를 평가 검토한다.

2-1. 한국어 정보처리 응용분야

이는 한글 자연언어처리로 통칭하는 데, 세분하면 한글의 자연언어 이해,

음성인식.합성, 문자인식, 기계번역, 자동전화통역등 많은 응용분야가 있다. 이분야의 응용분야는 한국어의 언어현상의 설명 또는 처리가 요구되므로 주지하는 바와 같이 음소정보처리가 많은 비중을 차지 한다. 완성형은 빈도수에 의해 선정된 글자에 코드를 부여한 것이므로 음소정보(자소정보)가 전혀없다. 예를 들어 'ㄱ'으로 시작하는 글자의 범위를 알 수 있는 규칙성이 없어서 이를 처리하려면 그에 관한 표를 별도로 만들어서 사용해야 한다. 이는 알고리즘상 계산량의 증가로 연결될 것이다.

조합형과 낱자형은 한글의 가장 중요한 정보인 자소정보를 내포하고 있으므로 이에 적합성을 가진다. 또한 현재의 컴퓨터 응용분야가 대부분 일반 자료처리 분야라는 점에서 완성형 부호계는 부분적 위상이 있겠으나, 증가일로에 있는 한국어 정보처리 응용분야에서 완성형 부호계는 그 한계 속성을 노출하고 있다.

2-2. 한글의 특성유지

한글은 오늘날 초성 19자, 중성 21자, 종성 28자(채움코드 포함)로 부터 11,172자를 조합할 수 있다. 이것에 대하여 제약을 가할 수 있는 방법은 한글 맞춤법이나 국어책 어느 곳에서도 찾아볼 수 없다. 그런데 로마자를 쓰는 기술자들이 먼저 컴퓨터를 만든 관계로 일차원꼴을 하고 있는 로마자 처리에 편리하도록 제작한 것이다. 그러면 일차원꼴을 읽기의 효율성을 높이기 위하여 이차원꼴로 표기하는 한글의 특성을 도외시하려는 여러가지 시도는 어쨌든 한글의 조합특성을 없애려는 노력에 불과한 것이다. 그런데 한글의 모아쓰기가 결정성 유한상태 오토마톤임은 주지의 사실이고, 이들은 이미 단말기에 구현, 사용하고 있다.

한글문자가 갖는 또하나의 특성으로 타이핑 수준에서 철자 오류를 걸러 낼 수 있다는 점이다. 이는 한글문자의 합성규칙에 의하여 글자로 구성이 안되는 글자를 자동조절에 의하여 음절인식 오류에 관한 신호를 해준다. 이것은 한글 자판에서 낱자를 입력하므로써 가능한 것이다.

최근에 국어학계에서 완성형 한글부호계가 우리의 어문생활에 장애를 제공하고 있음을 실예를 들어 지적하고 있다.2) 먼저 완성형은 훈민정음의 구조원리와 거리가 멀고 우리의 어문생활에 필요한 336 경우에 빠진 글자를 예시하였다. 즉 표준정서법에서 220자, 방언의 표기에서 83자, 국어교육을 위한 32자이다. 무엇 보다도

완성형의 국어에 대한 결정적 폐해는 우리의 글을 제대로 표기할 수 없게 만든 것이다.

고어는 15세기를 중심으로 할 때 초성이 37자, 중성이 29자, 종성이 26자이었다.4) 현행은 19자, 21자, 28자이다. 그 후에도 고어에는 많은 자음이 새로 나타난다. 현재 완성형에서 정의된 고어는 초종성에 오는 자음이 34자, 모음이 8자이다.5)9) 이들 유용한 고어의 조합 글자만 하더라고 약 1500자 라고 한다. 그런데 완성형에서는 풀어 쓰기만 하게 되어 있다. 이를 낱자형으로 한다면 별도의 단수 바이트 도형문자집단을 만들어서 확장법에 의하여 사용하면 될 것이다.

2-3. 입출력, 편집, 통신에서의 효율성

입출력에서 낱자형의 경우를 살펴보면, 33자 코드는 자판의 33자를 단지 한글 글자의 구성의 진위만 파악하고 아무런 변화없이 내부 표현한다. 51자 부호계는 오토마톤을 거치면서 복자모에 대한 단일 부호 생성을 하여 내부 표현한다. 한글 오토마톤의 계산량에 대해서는 O(n)임은 주지의 사실이다. 그리고 이들의 출력에서는 문자변환기(transducer)에 의하여 조합 폰트로 모아쓰게 된다. 조합형은 33자가 입력되어 2 바이트의 조합형 코드가 생성되고 출력에서 이들은 조합형 부호가 출력되거나 낱자형으로 풀어서 낱자형의 경우와 같게 된다. 완성형은 33자 입력에서 글자를 인식하고 2350자내에 있으면 완성형 부호를, 없으면 'FILL+초성+중성+종성'으로 8바이트로 표현한다. 출력에서 완성형 부호가 출력되어 완성형 폰트에 사상된다.

편집에서 평가의 중심은 에디터등에서 화면 커서와 텍스트 커서가 서로 대응하지 않으므로 이의 사상을 위한 것이다. 낱자형은 음절의 인식에 있어서 두가지 방법을 쓸 수가 있다. 하나는 한글의 글자는 반드시 모음을 중심으로 구성된다는 면에서 모음을 세는 것이고 다른 하나는 그림 1과 같이 칼럼계산 오토마톤을 사용하는 것이다. 이 경우 'CVCCV' 또는 'CCVCCC'은 해당 글자의 초성에, 'CVCVCV' 또는 'CVVV'는 모음위에 텍스트 커서가 놓인다. 두가지 모두 계산량은 O(n)에 해당한다. 조합형은 첫바이트의 8번째 빝을 1로 하여 한글과 영문을 구분하지만 두번째 바이트의 8번째 빝은 중성 모음을 표시하는데 사용되므로 0 또는 1이어서 로마자로 오도할 가능성이 있다. 그래서 대부분 처리의 편리성을 위하여 로마자를 모두 2 바이트로 표현하고 있다. 완성형은 한글을 8 빝환경에서 두 바이트의 8번째 빝에 1를 채우므로 만약 패턴 매치에서

앞글자의 뒷 바이트와 뒷글자의 앞바이트가 다른글자를 구성하는 밀려읽기 현상이 일어날 수 있다.

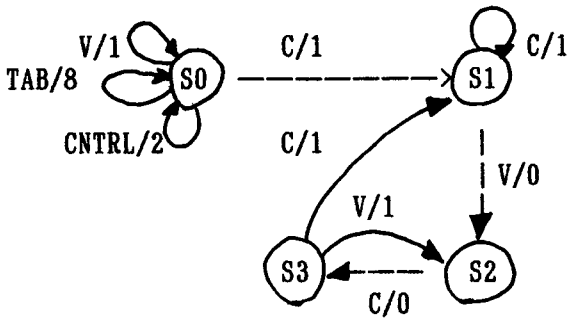

그림 1. 칼럼계산 상태도

상태	자음	모음	구분
0	1,1	0,1	0,1
1	1,1	2,0	0,1
2	3,0	0,1	0,1
3	1,1	2,1	0,1

그림 2. 칼럼계산 상태표

통신에서 완성형은 한글의 경우 2350자에 없는 글자인 경우 8 바이트로 풀어쓰기꼴로 전송해야 하고 특히 사용자 정의 구역을 이용한 글자의 경우에는 전혀 해결이 안된다. 조합형은 제어영역을 침범하므로 메타 문자가 발생되어 자료의 파괴 또는 시스템의 다운까지도 염려해야 하는 경우가 있다. 낱자형은 모든 문자집단을 전송할 수 있으며 또한 모든 한글 부호계를 상호간 번역할 경우에 그 축으로써 최적이다.

2-4. 소프트웨어 호환성

호환성 평가는 소프트웨어의 대다수가 어떤 부호계에서 개발되었느냐가 관건이 된다. 그런데 세계의 소프트웨어의 대부분은 로마자권에서 개발되었으며 또한 그들은 로마자를 사용하는 단수 바이트 부호계에 속한다. 한글 부호계의 완성형은 복수 바이트 부호계에 속하며, 따라서 그들과 부호개와 호환성 가진다. 복수 바이트 부호계로는 문자 집단의 크기가 94보다 큰 한자문화권의 나라 일본, 대만, 중국등에서 한자 수용을 위하여 채택하고 있다. 그렇다면 한글 부호계는 이들 나라의 부호계와 호환성을 가질 것이며 영문자계열의 부호계와는 호환성을 맞추기 위한 별도의 시간적 경제적 노력이 요구된다.

단수 바이트계 : 로마자 계열, 가나, 한글 낱자형

복수 바이트계 : JIS C 6226-1983, 한자, KS C 5601-1987

한글 낱자형은 로마자와 같은 단수 바이트 계열이며 로마자 부호계에서 영문 대문자 지역에 한글 자음 31자, 소문자 지역에 한글 모음 20자를 배치하고 이들은 7 빝계에서는 SI, SO를 사용하여 로마자와 혼용을 할 수 있으며, 8빝계에서는 8번째 빝에 1를

표시한다.

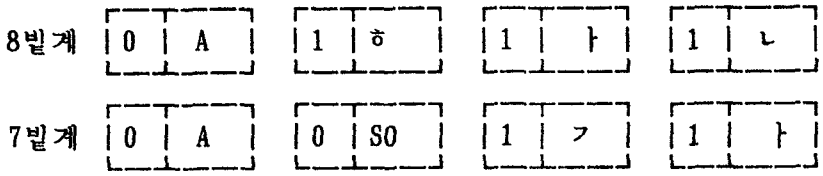

그림 3. 낱자형의 내부표현

한글 낱자형에서 내부 표현이 풀어쓰기 꼴로 내장되어 있기 때문에 영문자와는 달리 모아쓰기꼴로 출력하여야 하는 점, 한 글자의 구성요소가 일정하지 않다는 점, 글자식별을 위한 작업이 필요하다는 점등을 들어서 낱자형이 문제가 있다고 지적한다. 하지만 모아쓰기 틀은 화면 또는 인쇄기등의 단말기에 부착되어서 그 기능을 하고 있다. 그리고 문자열 정열 또는 검색 알고리즘의 계산 부담은 없음이 증명되었다.6)7) 결국 단수 바이트 부호계 계열의 소프트웨어가 복수 바이트 계열 보다 수적으로 많다는 점에서 완성형 보다는 낱자형 부호계는 호환성 문제에 있어서 우수하다는 점을 들 수 있다.

2-5. 한자문자의 수용성

한자를 한글 표기체계에 대하여 어떻게 볼 것인가 하는 문제는 국어학 분야에 있어서 한글전용론과 한자혼용론으로 오랜 논쟁을 하여왔다.1) 한자의 문자집단의 수용성은 다양한 적용성을 가져야 하고, 한자옥편에 있는 약 5, 6만자의
수용이 가능해야 한다. 물론 모든 문자를 모두 표현하도록 하는 문제는 부호계 문제만에 국한되지 않는다. 그들의 폰트문제 또한 함께 고려해야 할 것이다.

한자의 수용에 있어서 KS C 5601-1987은 4888자를 수용하고 있는 데 부족한 한자는 제 2 글자판을 제정할 것이라고 한다. 이것은 알고리즘을 모두 깨뜨릴 것이고 이를 해결하려면 알고리즘이 복잡해질 것이다. 그러므로 한자는 3 바이트 부호계로 하면 830,584자(94X94X94)를 수용할 수 있으며, 이는 현행의 모든 한자를 수용하고도 남는다. 실재의 내부 표현 방법은 그림 4과 같이 SS2로 시작하여 뒤에 3 바이트를 나열한다.

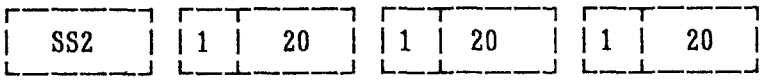

그림 4. 한자의 내부 표현방법

2-6. 다양한 문자집합의 수용성

나라마다 자국의 문자집단 부호계뿐만아니라 괘선문자, 비디오텍스, 고유한 도형 문자집단을 국제표준기구에 등록하고 있다. 그들 문자를 사용하려면 ISO에서 주어진 종단문자 F를 사용하여 부호계의 확장법에 적용하므로써 별도의 노력없이 사용할 수가 있는 것이다. 완성형 한글 부호계에는 이미 등록이 되어 있는 로마문자, 가다가나, 러시아문자, 그리스 문자등을 포함하고 있다. 이는 부호계 확장법을 무시한 결과이므로 완성형 부호계의 긍정적 존재 관점에서 본다면 분리시켜야 한다.

2-7. 알고리즘의 단순성

부호계의 제정에서 그 부호계가 알고리즘의 계산량에 미치는 영향 평가는 주요소 중의 하나다. 하지만 여기서도 기본 원칙은 설사 계산량이 늘어난다 하더라도 한글처리의 근본을 저해시켜서는 안될 것이다. 알고리즘의 단순화에 관한 평가는 한글 부호계 상호간의 우열에 관한 논점이 되고 있는 한글 문자열 정열, 한글 문자열 검색, 한글 오토마톤과 세가지 부호계의 계산량을 측정한다. 2-3의 입출력과 편집에서 언급한 바와 같이 계산량은 미미한 차이가 있을 뿐이고 단지 알고리즘의 복잡성을 본다. 검색에서 완성형은 밀려 읽기의 방지에서, 낱자형은 음절 인식 또는 계산에서, 조합형은 로마자와 한글의 두번째 바이트와의 구별에서 비슷하게 나타난다. 그런데 완성형은 2350 글자를 보완하기 위하여 사용자 정의 구역을 사용하였거나 제 2의 글자판을 만들었을 경우에는 기존 알고리즘으로는 처리가 곤란하다. 그리고 자연언어 처리와 같은 분야에서 자소정보를 요구할 때 자소정보에 관한 표를 가지게 되면 복잡도는 낱자형에 비하여 매우 증가하게 된다.

2-8. 국제규격의 적합성

한글 부호계의 제정에서 KS C 5601-1974와 KS C 5601-1987은 국제규격에 맞게 제정되었고 KS C 5601-1982만은 KS C 5620-1979가 있었음에도 불구하고 국제규격을 위반하였다. KS C 5601-1987 완성형은 복수 바이트계에 속하고 KS C 5601-1974는 낱자형(KS C 5601-1987의 부속서에 보조 부호로도 있음)으로 단수 바이트계에 속한다는 점이 다르다. 그러므로 완성형으로 개정 사유에 국제규격 운운5)만으론 설득력이 약하다.

2-9. 평가의 결론

이상의 8가지 평가 요소를 중심으로 적합성 평가를 하였다. 완성형 부호계는 한국어 정보처리와 같은 미래의 응용분야에 대하여 심각한 문제점이 지적되었고 또한 한글 특성의 유지면에서도 중차대한 결점을 가진 것으로 판명되었다. 알고리즘의 단순성 문제등에서도 만약 사용자 정의 구역에 각자 필요한 문자를 정의하여 쓰거나 제 2의 글자판을 만들 경우 역시 심각한 상황으로 빠지게 된다. 소프트웨어의 호환성 문제에 있어서도 로마자권 보다는 한자 부호계권인 가나 부호계 10)등에 더 부합할 것으로 나타났다. 한자 4,888자는 부족하므로 근본적인 대책이 요구되며, 제 2 글자판을 만들어 보완하여야 할 것이나 역시 알고리즘의 문제를 야기할 것이다. 완성형은 한글 폰트 설계 문제에 있어서도 조합형 폰트에 비하여 시간적 경제적으로 큰 부담이 된다.

낱자형의 경우 먼저 한글의 특성을 유지함에 뛰어나므로 한국어정보처리 분야에서 요구하는 요구도를 만족시킬 수 있다. 그리고 한글의 창제원리나 구성원리에 부합한다. 또한 단수 바이트계이므로 로마자권과 소프트웨어 호환성을 가지므로써 소프트웨어 수출입이나 개발에 있어서 유리하다. 통신에 있어서 여러 한글부호의 번역축 부호계가 되고 국제규격에 적합성을 가지므로 이를 한국어 정보처리의 부호계로 함이 합당하다고 본다.

3. 결론

한글문자는 음양오행의 바탕위에서 15세기의 수학적 개념이 도입된 우수한 문자라는 사실은 이제까지의 부호계의 제정관점을 가집에서 간과되어온게 사실이다. 그런 결과로 빚어진 한글 부호계의 혼란한 제정 및 개정사는 오늘날 한국의 정보산업 진입에 있어서 큰 짐이 되어 온게 또한 사실이다. 그런 시점에서 한글 부호계를 일련의 요건을 갖추어 한글문자의 구성원리적 관점에서 평가하였다. 평가의 결론부에서 볼 수 있는 여러가지 사실은 완성형 부호계를 더 이상 합리화하려는 노력을 포기하게 할 만한 이론적 사실적 근거로써 제시할 수 있다. 이는 앞으로 한글정보처리의 관점에서나 한국어정보처리의 관점에서 모든 이에게 배우기 쉽고, 읽기 쉽고, 쓰기 쉬운 부호계로 개선하는 방향의 제시이며, 한글문화 창달에 발전적 영향을 끼칠 부호계로의 개정이 이루어져야 할 것이다.

한글은 어느 특정계층의 문자가 아니다. 우리 국민전체가 어떠한 사상 어떠한 소리라도 적을 수 있도록 유이도를 크게 하여 가진 문자이다. 그러므로 완성형과 같이 한글의 원초적인 성질을 없애버림은 우리의 표현의 자유를 박탈하는 결과로 해석될 수도 있다. 한글부호계는 한국국민 모두의 것이므로 컴퓨터의 기술적 제약이나 국제규격의 한계성에 의한다 하더라도 한글 자체는 어떠한 제약을 받아서는 안될 것이다.

참고문헌

1. 김문창, 국어문자 표기론, 문학사, 1984
2. 김충회, "현행 KS 완성형 한글 코드의 문제점", 한글과 한국어 정보처리, 한글날기념학술대회,1989, P21-28
3. 변정용, 한글문자의 지적처리에 관한 연구, 한국전자통신연구소, 1989
4. 장영길, 15세기 국어의 자음체계 연구, 석사학위 청구논문, 1985
5. 박동순,이재현,최은만,강석, " 컴퓨터 한글코드의 사용과 표준화", 한국정보과학회지, 제6권, 제1호, 1988, P69-74
6. 변정용, 한국어정보처리에 최적한 한글코드에 관한 연구, 한글과 한국어 정보처리, 한글날기념학술대회,1989, P39-43
7. 변정용, 한글코드의 개선과 알고리즘의 개발 : 날자형 33자 코드, 동국대학교 논문집 제8집, 1989,P421-433
8. 정희성, 한글문자의 구조와 구성원리에 관한 과학적 고찰, 전자통신, 제10권 제4호, 1989, P99-117
9. 공업진흥청, KS C 5601-1974, 1982, 1987
10. 日本工業標準調査會, JIS C 6220, JIS 6226, JIS C 6225