Regular paper http://dx.doi.org/10.13067/JKIECS.2015.10.3.375

한글의 정보처리 및 통신용 부호 최적화를 위한 한국어 분석

홍완표*

Analysis of Korean Language to Optimize the Hangul Character Coding for Information Processing and Communication

Wan-Pyo Hong*

요 약

본 논문은 정보처리 및 전송용으로 사용되는 한글의 부호화를 최적화할 수 있도록 하기 위하여 한국어를 연구하였다. 본 논문은 한국어 구성하고 있는 한글의 구성현황과 그 한글들에 대한 각각의 사용빈도를 분석하였다. 본 논문은 본 연구결과 분석된 한글의 구성현황을 한국 KS 문자 표준과 국제 문자표준인 유니코드로 부호화되어 있는 한글 문자와 비교하였다. 연구를 위해 사용된 한국어는 국립국어원의 "현대국어사용빈도조사 결과"를 대상으로 하였다. 이 보고서에 수록된 한국어는 총 58.437개이다. 분석결과 한국어 총58,437국어를 구성하고 있는 한글은 총1,540개였다. 이 총1,540개 한국어 중에서 사용빈도가 가장 높은 글자는 "다"로서 전체 사용빈도의 15%였다. 사용빈도가 가장 낮은 글자는 "훳"으로서 전체사용빈도의 0.00003%였다. 한국어를 구성하고 있는 한글 글자수는 유니코드 한글문자 부호를 구성하고 있는 한글 수 보다 약 7.2배, KS X 1001 한글 문자 부호를 구성하고 있는 한글 수보다 약 1.5배 적은 것으로 나타났다.

ABSTRACT

This paper is studied the Korean language to optimize the Hangul character coding for information processing in information terminal device and transmission in network. The paper analyzed Hangul character in Korean language and use frequency of each character. The paper also compared the analysis result to Hangul characters which are coded in standard in Korean character and Unicode. This study referred "Modern Korean Use Frequency Rate Survey Result" issued by The National Institute of the Korean Language. There are total 58,437 Korean words in the report. As a result of this paper, the Korean word 58,437ea are consisted of Hangul character total 1,540ea. The highest use frequency character is "다" and its use frequency to total use frequency rate is 15%. The lowest use character is "훳"and its use frequency to total use frequency rate is 0.00003%. The number of analyzed Hangul character 1,540 is less 7.2 times and 1.5 times than Korean and Unicode standard respectively.

키워드

Hangul Character Coding, Information Processing, Network, Use Frequency, Unicode
한글 문자 부호화, 정보 처리, 통신망, 사용 빈도, 유니코드

* 교신저자(corresponding author) : 한세대학교 정보통신공학과(wphong@hansei.ac.kr)
접수일자 : 2014. 12. 20 심사(수정)일자 : 2015. 03. 13 게재확정일자 : 2015. 03. 23

Ⅰ. 서 론

1990년대 말을 전후로 한 국내외의 정보통신네트워크의 초고속화는 인터넷의 확산을 가속화 시켰다. 이것은 인터넷을 통한 다양한 컨텐츠의 제공을 가능하게 만들었고 이로 인하여 폭발적인 데이터 트래픽의 증가를 촉발시켰다. 특히 인터넷의 관문 역할을 하는 웹페이지를 구성하는 문자기반의 트래픽은 웹사이트의 증가와 함께 전송 데이터 트래픽을 증가시키는 하나의 큰 요인이 되었다. 또한 문자기반의 이메일, 검색, 파일전송 및 SNS서비스 등은 문자 트래픽의 증가를 촉발시키고 있다.

본 논문은 인터넷상에서 사용되고 있는 한글의 부호화에 대한 문제점을 분석하였다. 그리고 그 문제를 해결하기 위하여 기초가 되는 한국어를 구성하고 있는 한글을 분석하였다. 이것을 토대로 하여 이 한글들에 대한 사용빈도를 분석하였다. 본 연구결과는 한글의 정보처리와 전송분야에서 다양하게 활용할 수 있다. 정보처리분야에서는 음성인식, 데이터처리속도의 개선, 문자처리 소프트웨어의 최적화 등에 적용될 수 있다. 문자전송분야에서는 원천부호화, 회선부호화, 스크램블링 기술, 오류검색 및 정정 등에서 활용될 수 있다.

Ⅱ. 한글문자부호체계와 문제점

2.1. 문자부호화 현황

현재 인터넷 상에 전송되는 문자부호의 구성포맷은 ASCII부호 체계[1]의 개념을 기반으로 진화되어 온 EBCDIC, UNICODE[2] 등을 기반으로 하고 있다. 이들 문자의 원천부호체계는 통신용 부호체계에서 요구하는 조건들을 만족시키지 못하고 있다. 예를 들어 문자부호의 시초라 할 수 있는 모르스 부호 체계[3]는 문자의 사용빈도를 토대로 하여 문자를 부호화함으로서 문자의 전송시간을 단축함으로서 정보의 전송효율을 높이고자 하였다. 이것은 통신의 신속성을 고려한 것이다. 허프만 부호체계는 문자의 사용빈도에 대한 적용을 정보기기 내에서의 디지틀 이진비트의 문자부호화에 적용하였다[4]. 즉 이 두 개의 부호체계는 사용빈도가 높은 문자에 짧은 부호를 부여하고 반면에 사용빈도가 낮은 문자에 긴 부호체계를 부여하여 문자의 전송효율을 제고시키고자 하는 가변길이 부호체계이다. 반면에 ASCII코드, 유니코드는 고정길이 부호체계를 가지고 있다.

2.2. 한글의 문자부호표준과 한글문자부호특성

한글의 문자부호표준은 크게 두 가지가 있다. 하나는 국제문자부호 표준인 유니코드표준이고 다른 하나는 국내문자부호 표준인 KS X 1001[5]이다. 한글문자의 부호체계는 문자의 특성상 다른 나라의 문자부호체계와는 다르게 아주 독특한 부호체계를 가지고 있다. 세계의 모든 문자는 중국문자 등과 같은 독특한 경우를 제외하고는 자음과 모음의 낱자 체계로 부호화되어 있다. 그러나 한글의 부호체계는 자음과 모음의 낱자 부호체계와 자음과 모음을 조합한 한글자의 형태로 된 부호체계를 가지고 있다.

글자는 자음과 모음의 조합에 의하여 만들어지기 때문에 자음과 모음의 부호체계에 따라 그 글자의 수가 다르게 된다. 한글은 훈민정음의 경우에 자음과 모음을 합하여 28개의 자소로 되어 있다. 현대 한글의 경우에는 자음14개, 모음10개의 24개 자소를 가지고 있다. 한글은 초성, 중성, 종성의 형태로 조합되어 완성된다. 즉 초성과 종성은 자음, 중성은 모음으로 이루어진다. 그러므로 24개의 한글낱자는 총1,960개의 한글글자를 만들 수 있다. 한글문자부호 표준인 KS X 1001에는 초성, 중성 및 종성이 각각 19개, 21개 및 27개로 부호화되어 있다. 따라서 총 10,772개의 글자를 만들 수 있다. 동일한 국내표준인 KS X 1001 문서에 부호화되어 있는 완성형 한글 문자 부호수는 총 2,350개이다.

표 1. 한글문자부호 체계
Table 1. Hangul coding system

Hangul Code systems	Jamo (Initial consonant, Vowel, Final consonant)	Number of using character	Number of total character
Basic Hangul	19, 21, 27	-	10,772
Unicode	124, 94, 138	11,172	1,608,528
KS X 1001	44, 21, 51	2,350	47,124

국제문자부호체계인 유니코드에는 한글 낱자가 초성, 중성 및 종성 각각 124개, 94개 및 138개로 부호화되어 있다. 따라서 이 낱자부호체계에 의해 총 1,608,528개의 글자를 만들 수 있다. 그러나 실제 유니코드 상에 부호화되어 있는 완성형 한글문자 부호는 총11,172이다. 이와 같이 표준에 따라 한글낱자와 완성형글자의 수가 다르게 부호화되어 있다. 표 1은 이것에 대한 현황을 보여주는 것이다.

2.3. 현행 문자부호체계의 문제점

현재의 ASCII, UNICODE 등의 부호체계는 디지틀 이진비트로 구성된 문자부호가 통신망을 통하여 전송될 때 발생하는 중대한 문제점을 전혀 고려하지 않고 있다. 예를 들어 이진비트 0이 일정개수 이상 연속하여 구성되어 있는 문자부호가 전송로 상에 전송될 때 수신측에서 이 문자의 비트 간 동기를 잃게 되어 디코딩을 할 수 없도록 만든다. 또한 2진 비트 1이 일정개수 이상 연속하여 구성되어 있는 문자부호가 장거리 전송로에 전송될 때 전송로 상에 비주기 직류성분의 에너지가 있는 것과 같은 현상을 갖도록 한다. 연속된 2진비트 1로 구성된 문자 부호로 인하여 장거리 전송로상에 발생하는 직류성분의 에너지 발생을 막기 위해 사용하는 AMI회선 부호화방법, 연속된 2진비트 0으로 인하여 발생하는 수신측에서의 비트동기를 상실하지 않도록 하는 스크램블링 기술의 적용이 그것들이다[6]. 문자통신 프로토콜로서 HDLC의 프로토콜을 사용하는 경우에는 플래그부호와 정보문자부호간의 혼란을 방지하기 위해 정보문자부호에 플래그부호와 유사한 부호가 있을 때에는 다른 비트부호로 변환하여 주고 있다[8]. 즉, 이러한 사항들은 모두 문자의 원천부호의 특성에 의하여 발생되는 것이다. 현재까지는 이렇게 원천부호로 인하여 발생하는 통신상의 문제점을 통신측면에서 해결하여 오고 있다.

만약에 사용빈도가 높은 문자부호가 이러한 문제를 일으키도록 부호화된다면 더욱 전송효율을 감소시키게 된다. 이상에서 언급한 두 가지의 예의 경우에 첫 번째 사항은 물리계층의 회선부호화과정에서 해결하고 있고 두 번째 사항은 데이터링크계층에서 해결하고 있다. 즉 두 가지 사항 모두 소프트웨어측면에서 발생한 문제점을 하드웨어적으로 해결해 오고 있다. 본 논문은 이러한 문제를 응용계층에서 소프트웨어적으로 해결하여 전송장비의 운용과 데이터의 전송효율을 제고 시키도록 하고자 하는 것이다. 그러한 방법으로 원천부호체계를 개선 또는 새로운 통신에 적합한 원천부호체계를 만들 수 있도록 하고자 하는 방안을 제시하기 위해 기본적으로 요구되는 한글사용현황을 연구하였다.

III. 한글분석

3.1. 한글분석내용

본 논문에서는 한글의 분석을 위하여 한국어는 국립국어원의 연구보고서 “현대국어사용빈도조사”을 대상으로 하였다[7]. 본 연구보고서는 국어의 사용빈도와 사용개수를 교재, 교과, 교양, 문학, 신문, 잡지, 대본, 구어 및 기타 등 9개 분야로 나누어 기술하고 있다. 이 파일에 수록된 한국어는 총58,437개 이다. 9개 분야 중에서 사용빈도가 가장 많은 순서대로 기술하면 교양, 신문, 문학, 잡지, 교과, 기타, 교재, 구어, 대본분야 순이다.

3.2. 한글분석절차 및 방법

한글분석은 사용빈도수가 많은 순서로 정렬을 하였다. 그리고 사용빈도를 5단계로 나누었다. 1단계는 사용빈도수가 1,000개 이상, 2단계는 사용빈도수가 500-999개, 3단계는 사용빈도수가 100-499개, 4단계는 사용빈도수가 10-99 그리고 5단계는 사용빈도수가 1-9개로 하였다. 각 단계별 한국어 단어수는 1단계 192개, 2단계 202개, 3단계 1,792개, 4단계 11,826개 그리고 5단계 44,425개이다. 분석방법[9]은 첫 번째 단계로 사용빈도가 많은 순서대로 각 단계별로 각각의 단어를 구성하고 있는 글자를 한 글자씩 모두 나누어 가, 나, 다 순서인 내림차순으로 정렬을 하였다. 두 번째로 동일한 글자들만의 사용빈도를 합하였다. 그리고 동일한 글자 중에서 사용빈도수가 더하여져 있는 한 글자만 남기고 나머지는 모두 삭제하였다. 세 번째로 한글자 씩만 남은 글자를 가, 나, 다 순서로 정렬하였다. 이렇게 각 단계를 분석한 다음에 단계를 색으로 구분하였다. 예를 들어 사용빈도가 가장 많은 1단계는 흰색, 2단계는 노랑색, 3단계는 청색 등으로 구분하였다. 네 번째로 세 번째 단계에서 정렬된 단계별 글자

들을 하나로 합하여 정렬하였다. 그리고 이것을 가, 나, 순으로 정렬하였다. 다섯 번째로 이 정렬된 글자들의 동일한 글자들에 대한 사용빈도 합을 계산하였다. 그리도 동일한 글자중에서 사용빈도수가 더하여져 있는 한글자만 남기고 모두 삭제하였다. 마지막 여섯 번째로 사용빈도가 더하여진 글자들만을 가, 나, 다 순으로 정렬하였다[10].

3.3. 한국어 분석결과

한국어 단어 총 58,437개를 분석한 결과 이 단어들을 구성하고 있는 글자수는 총 1,540개로 나타났다. 표 2는 한국어의 자음을 기준으로 한 글자수를 보여주는 것이다. "ㄱ"열이 194개로 가장 많고 다음은 "ㅅ"열로 194개이고 가장 적은 글자는 "ㅎ"으로 64개이다.

표 2. 한글 1,540자의 자음기준 글자 수
Table 2. Number of Hangul character in Hangul character 1,540ea

Consonant Character	Number of Character	Consonant Character	Number of Character
ㄱ	194	ㄱ	194
ㄴ	96	ㅅ	154
ㄷ	144	ㅇ	148
ㄹ	89	ㄷ	144
ㅁ	90	ㅈ	131
ㅂ	127	ㅂ	127
ㅅ	154	ㅎ	98
ㅇ	148	ㄴ	96
ㅈ	131	ㅁ	90
ㅊ	67	ㄹ	89
ㅋ	69	ㅋ	69
ㅌ	69	ㅌ	69
ㅍ	64	ㅊ	67
ㅎ	98	ㅍ	64
계	1540		

표 3은 한국어의 자음을 기준으로 사용빈도별 글자수를 보여주는 것이다. 총 사용빈도수는 3,368,465로서 "ㄷ"열이 사용빈도가 725,586로서 사용 빈도 율이 전체의 22%로서 가장 많고 다음이 "ㅇ"열 순으로 사용빈도 율이 전체의 15%로 나타났다. 사용빈도율이 가장 적은 것은 "ㅋ"열로서 사용빈도율이 전체의 약 1%로 나타났다.

표 3. 한글 1,540자의 자음기준 글자 사용빈도
Table 3. Use frequency of Hangul character in Hangul character 1,540ea

Cons-onant Chara-cter	Use freque-ncy	Frequ-ency rate (%)	Conso-nant Charac-ter	Frequ-ency rate (%)
ㄱ	403,952	12	ㄷ	22
ㄴ	141,830	4	ㅇ	15
ㄷ	725,586	22	ㄱ	12
ㄹ	171,343	5	ㅅ	9
ㅁ	165,004	5	ㅈ	9
ㅂ	161,801	5	ㅎ	8
ㅅ	305,115	9	ㄹ	5
ㅇ	505,570	15	ㅁ	5
ㅈ	291,176	9	ㅂ	5
ㅊ	114,136	3	ㄴ	4
ㅋ	18,775	1	ㅊ	3
ㅌ	43,242	1	ㅍ	1
ㅍ	45,032	1	ㅌ	1
ㅎ	275,903	8	카	1
계	3,368,465	100	계	100

표 4는 분석단계별 단어수와 이 단어를 구성하고 있는 글자 수이다. 각 분석단계별내 단어를 구성하고 있는 글자 수는 1단계 170개, 2단계 194개, 3단계 632개, 4단계 1,187개 그리고 5단계 1,475개로 나타났다.

표 4. 사용빈도별 단어 및 글자 수
Table 4. Number of word and character per Use frequency

Step	Use Frequency	Number of word	Number of character
1	1,000~	192	170
2	500~999	202	194
3	100~499	1,792	632
4	10~99	11,826	1,187
5	1~9	44,425	1,475

표 5는 한국어 총 58,437개를 구성하고 있는 총 1,540개 글자에 대한 사용빈도율을 보여 주는 것이다. 이 표에서 보는 바와 같이 사용빈도율이 가장 많은 글자는 “다”로서 전체사용빈도의 약15.2%로 나타났다. 다음으로 “하”가 4.6%, “이”가 2.0% 순으로 나타났다. 사용빈도가 전체사용빈도의 1.0%에 해당되는 글자수는 총12개로 나타났다.

표 5. 한글 1,540자중 사용빈도율 1%이상의 글자 및 사용빈도율
Table 5. Hangul character and Use frequency rate in Hangul character 1,540ea

Character	Use Frequency rate (%)	Remark
다	15.2	
하	4.6	
이	2.0	
어	1.5	
그	1.5	
지	1.4	
나	1.3	
리	1.3	
사	1.2	
기	1.2	
있	1.1	
가	1.1	

표 6은 본 논문의 연구결과와 표 1의 한글문자부호체계와 비교한 것을 보여주고 있다. 이 표에 의하면 연구결과와 한글기본문자의 낱자 구성이 매우 유사함을 알 수 있다. 연구결과로 나타난 사용문자수는 유니코드상의 한글문자 부호수 보다 약7.2배 적고 KS X 1001의 한글문자 부호수보다는 약1.5배 적은 것으로 나타났다.

표 6. 연구결과와 현행 한글부호체계와의 비교
Table 6. Comparison between study result and existing Hangul code system

Section	Number of Character	Number of total character	Jamo (Initial Consonant, Vowel, Final Consonant)
Study result	1,552	9,576	19, 21, 24
Basic Hangul	-	10,772	19, 21, 27
Unicode	11,172	1,608,528	124, 94, 138
KS X 1001	2,350	47,124	44, 21, 51

Ⅳ. 결 론

본 논문은 정보처리 및 네트워크에 전송되는 한글문자의 부호화를 최적화할 수 있는 한국어를 분석하였다. 본 논문은 한국어의 한글 구성현황과 그 한글에 대한 사용빈도를 분석하였다. 이 연구는 국립국어원의 “현대국어사용빈도조사결과” 보고서를 기반으로 하였다. 분석결과 사용빈도가 가장 많은 1단계 순으로 각 국어를 구성하고 있는 글자수는 1단계 170개, 2단계 194개, 3단계 632개, 4단계 1,187개 그리고 5단계 1,475개이다. 이를 종합하여 분석한 결과 총58,437개의 국어를 구성하고 있는 한글은 총1,522개였다. 이 총 1,522개 글자중에서 사용빈도가 가장 높은 글자는 “다”로서 전체사용빈도의 15%였다. 사용빈도가 전체사용빈도의 1.0%에 해당되는 글자수는 총12개였다.

본 논문의 연구결과와 한글기본문자의 낱자가 매우 유사하였다. 반면에 사용문자수는 유니코드 한글문자수 보다 약7.2배 적고 KSX 한글문자보다 약1.5배 적음을 나타내고 있다. 본 연구결과에 의하면, 본 연구에서 분석대상이었던 국립국어원의 한글사용빈도조사 보고서에 있는 한국어 단어들이 현재 인터넷에서 일반적으로 널리 사용되고 있는 글자들임을 고려할 때, 현행 유니코드 한글문자부호와 KSX 한글문자 부호는 통신용으로는 과다하게 부호화되어 있는 것으로 나타났다. 특히 유니코드 한글문자 부호는 특별히 과다하게 부호화되어 있는 것으로 분석되었다. 향후 본 논문의 결과는 현행 한글부호체계에 대한 개선 또는 새로운 통신용 한글문자부호의 제정 등에 활용될 것으로 기대된다. 또한 본 연구결과는 음성인식기술, 능률적 자판문자배열 등에도 활용이 기대된다.

참고 문헌

[1] Kan Laitman, *A Natural Introduction Computer Programming with C++*. St.Victoria Canada :

Trafford Publishing, 2002.
[2] Jukka Korpela, *Unicode Explained.* California U.S.A, O'Reilly Media Inc, 2006.
[3] E. Desurvire, *Classical and Quantum Information Theory.* New York: Cambridge university press, 2009.
[4] Eugene. S. Schwartz, "An Optimum Encoding with Minimum Longest Code and Total Number of Digits," *Information and control* vol 7, 1964, p.37.
[5] C. Kwon, "International Standardization of Code for Information Interchange," *KIOA review,* vol.1, no.2, 1995, pp.119-138.
[6] B. A. Forouzan, *Data communications and Networking. 4th ed.,* New York: McGraw Hill, 2007.
[7] W. "An Analysis on the Korean Language for Optimum Transmission of Hangul Code," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 10, no. 1, 2015, pp. 33-38.
[8] Y. Han, "A study on motion prediction and subband coding of moving pictuers using GRNN," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 5, no. 3, 2010, pp. 256-261.
[9] K. Lee and Y. Son, "Fast Encoding Algorithm of Low Density Codes," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 9, no. 4, 2014, pp. 403-408.
[10] Y. Kim, "A Study on Fractal Image Coding," *J. of the Korea Institute of Electronic Communication Sciences,* vol. 7, no. 3, 2012, pp. 559-566.

저자 소개

홍완표(Wan-Pyo Hong)

1991년 서울과학기술대학교 전자공학과(공학사)
1994년 연세대학교 공학대학원 전자공학전공(공학석사)
1999년 광운대학교 대학원 전자공학과(공학박사)
1990년 전기통신기술사합격
1991년 정보통신부 5급특별채용고시합격 본부 통신정책실, 전파방송관리국, 정보화기획실
1997년 삼성전자(주) 통신사업부 전송영업그룹장
1999년 광운대학교 연구전담교수
2000년 한국정보통신기술사협회장
2002년 한세대학교 정보통신공학과 교수
2014년 USC 동북아언어문화학과 방문학자
※ 관심분야 : 위성통신방송, 문자코딩, 통신정책