바른 한글음 생성을 위한 한자·한글 변환기 설계 및 구현

강 형 일[†] · 강 승 헌[†] · 장 수 민[†] · 유 재 수[††]

요 약

본 논문은 한자·한글 혼합 문서들을 정확한 우리말 문서로 자동 변환하는 한자·한글 변환기를 설계하고 구현한다. 구현된 한자·한글 변환기는 정보 검색 시스템에서 한자·한글 혼합 문서 검색시 사용자들이 한자에 해당하는 한글음을 파악하는데 걸리는 시간을 절약하고 쉽게 판독할 수 있도록 한다. 이를 위해 KS C 5601 표준코드를 기준으로 바르지 못한 한글음 생성의 원인을 조사하고 두 개 이상의 한글음을 갖는 한자들과 이형표기가 가능한 한자를 올바른 한글음으로 변환할 수 있는 한글단어 매핑사전을 구축한다.

Design and Implementation of a Hanja-Hangul Convertor for Generating Correct Hangul

Hyungil Kang[†] · Seung Hun Kang[†] · Su Min Jang[†] · Jae Soo Yoo[††]

ABSTRACT

In this paper, we design and implement a convertor that automatically transforms the mixed document with Hanja(chinese characters) and Hangul(korean characters) into a document with only Hangul. The Hanja-Hangul convertor reduces time to understand a Hnaja and reads the documents easily, when users retrieve mixed documents with Hanja and Hangul. To do this, we investigate the causes of incorrect Hangul generation based on KS C 5601 standard code and we construct a Hangul mapping dictionary that is used to convert Hanja with two or more Hangul sound or with different mark into the corresponding correct Hangul.

1. 서 론

현대 사회는 정보교류가 정치, 경제, 사회, 문화기술 등 다방면에서 급속히 이루어지고 있다. 이러한 상황에서 보다 효율적이고 다양한 정보 검색, 관리 및 교환을 위해 정보 검색 시스템이 사용되고 있다. 이와같은 정보 검색 시스템에서 우리가 사용하는 언어인 자연언어를 컴퓨터로 빠르게 처리해서 원하는 문서들을 찾아주는 기술을 갖는 것이 점차 필수적이 되어가고 있다.

정보 검색 시스템에서 처리되는 자연언어와 관리되는 많은 문서들은 우리 문화의 특성상 한자와 한글이 혼합되어 있다. "정보화 사회"로 불리우는 현대 사회에서 매일 기하급수적으로 쏟아져 나오는 정보가 많은 사람들이 읽거나 이해하기 어려운 한자가 섞

※본 연구는 연구개발정보센터 96년도 연구비 지원에 의해 수행되었음
† 준 회 원:충북대학교 정보통신공학과
†† 정 회 원:충북대학교 정보통신공학과
논문접수:1997년 7월 23일, 심사완료:1997년 12월 31일

여 있다면 정보를 해석하는데 많은 시간을 소비할 것이다. 따라서 정보 검색 시스템에서 검색시 사용자들이 한자 정보의 한글 음을 파악하는데 걸리는 시간을 절약하기 위해, 저장된 한자·한글 혼합 문서들을 사용자들이 이해하기 쉬운 우리말 문서로 제공될 수 있도록 해야한다. 즉, 한자·한글 혼합문서에서 한자를 한글화하는 한자·한글 변환 시스템 개발이 절대적으로 요구된다. 그러나 한자로 표기된 용어를 한글 용어로 자동 변환하게 되면 생성된 한글용어가 원하는 바른표기로 변환되지 않는 경우가 발생하게 된다. 자동 변환 시스템에서 올바른 한글용어가 생성되지 않는다는 것은 대단히 심각한 문제가 된다.

본 논문에서는 KS C 5601 표준코드를 기준으로 한자·한글 변환기 개발에 관한 연구를 수행하였다. 먼저 한자·한글 혼합 문서로부터 한자 추출기를 구현하며, 표준코드에 존재하는 한자 4,888자에 대한 한글사전을 구축하고 추출된 한자를 한글 코드로 변환하는 루틴을 설계한다. 동자이음에 해당하는 한자들은 현재 표준코드에서 각각의 한자 글꼴에 코드값이 반복적으로 배정되어 있는 점을 이용하여 처리한다. 또한 이형표기가 가능한 한자에 대해서 올바른 한글음 생성을 위하여 한글단어 매핑사전을 구축한다[3][6][8].

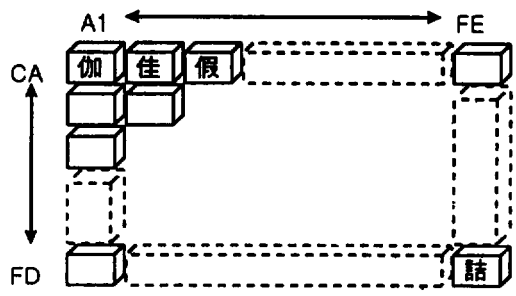
본 논문의 구성은 다음과 같다. 제 2장에서는 KS C 5601 표준코드를 기준으로 표준코드내에 존재하는 한자 분석 및 한글 변환시 바른 한글음 생성을 위해 사전에 고려하여야할 사항을 조사한다. 3장에서는 한자·한글 변환기를 설계하며 4장에서는 한자·한글 변환기 구현 및 분석한다. 마지막으로 본 논문의 결론과 향후 연구방향에 대해 5장에서 기술한다.

2. 한자 코드 체계

2.1 표준한자의 구성

KS 완성형 코드와 KS 조합형 코드에 완성형의 한자 4,888자가 포함되어 있다[4][7][9]. 차이점은 문자수는 같으나 코드 배열만 다를 뿐이다. 이것은 한자를 조합해서 사용하는 것보다 한자를 많이 사용하는 순으로 나열해 놓는 것이 훨씬 노력을 절감할 수 있기 때문이다.

한자의 출력에서 가장 중요하게 대두되는 문제는 한자 코드에 관한 부분이다. KS C 5601에서 표준으로 채택하여 활용하고 있는 표준 한자의 구성과 수는 (그림 1)과 같다.



(그림 1) 한자 코드의 구성
(Fig. 1) Architecture of Hanja Code

(그림 1)과 같이 한자는 한글과 동일하게 2바이트 체계로 구성되어 있다. 국제어를 위하여 확장되어 사용할 수 있는 문자의 수는 2바이트 체계 65536자에서 제어문자 영역을 제외한 것 중 D영역만 사용할 수 있다. 이중 한자에 할당되어 있는 코드영역은 상위 1바이트의 코드영역으로 ISO-2022의 부호확장 체계를 수용하여 처리할 수 있는 A1~FE이고 하위 1바이트의 코드영역은 A1~FE까지다. 이중에서 한글과 특수문자 그리고 사용자 정의 문자를 뺀 부분, 즉 상위 CA~FD 영역의 94개의 코드를 한 블럭으로 하여 52개의 블럭으로 구성되어 있다.

2.2 표준한자 코드의 분석

본 절에서는 한자·한글 변환시 원하지 않는 한글 색인어 생성에 따른 주요한 문제를 동자이음 처리 부분으로 보고 우선적으로 표준한자 코드내에서 복수의 음을 갖는 한자를 분석하였다[6][7]. 분석한 내용은 크게 표준한자 코드에 존재하는 한자로서 두 가지 이상의 음을 갖는 문자의 수 2,187자와 표준한자 코드에 있는 한자로서 하나의 한자 글꼴에 두 번 이상의 코드값을 갖는 문자의 수 268자임을 알 수 있었다.

표준한자 코드에 있는 한자로서 두 가지 이상의 음을 갖는 문자의 수는 다음과 같이 2,187자이다.

2음자	805자	(1,610자)
3음자	139자	(417자)
4음자	28자	(112자)

5음자	7자	(35자)
6음자	1자	(6자)
7음자	1자	(7자)
계	980자	(2,187자)

또한 표준한자 코드에 있는 漢子로서 두 가지 이상의 코드를 갖는 문자의 수(重出字)는 다음과 같이 268자이다.

2회	257종	514자	(257자)
3회	4종	12자	(8자)
4회	1종	4자	(3자)
계	262종	530자	(268자)

따라서 표준코드의 한자는 4,888자 중에서 268자를 제외한 4,620자를 수록한 셈이 된다. 그외에 正字와 略字, 本字, 古字, 俗字가 동시에 사용된 문자로서 표준한자 코드에 둘다 존재하는 경우 (예, 개(蓋, 盖), 년(年,秊), 만(万, 萬)) 바른 한글음의 생성에는 영향을 주지 않는다.

2.3 한자·한글 변환시 고려하여야 할 사항

정보 검색 시스템에서 관리하는 한자·한글 혼합 문서들을 순수한 우리말 문서로 변환하는 과정에서 올바른 한글음으로 변환되지 않는 경우가 발생하게 된다. 이에 본 절에서는 한자·한글 변환시 바른 한글음 생성을 위해 사전에 고려하여야 할 사항을 조사하였다[1][2][3][6][8].

2.3.1 동자이음어 및 두음법칙 적용 문자의 처리

한자는 대부분 한개의 한글 음을 갖고 있지만 두개 이상의 한글 음을 갖고 있는 문자들도 상당히 존재한다. 두개 이상의 한글 음을 갖는 한자를 정확한 한글로 변환하기 위해서 규칙을 발견하고 그 규칙을 이용하여 별도로 구현할 수 있겠지만 현재 KS C 5601의 표준 한자 코드 체계에서 두가지 이상의 한글 음을 갖는 한자에 대해서 각기 한자 코드에 반복적으로 배정되어 있는 점을 이용한다. (그림 2)에서는 한자 "不"에 대해서 각각 "불"과 "부"로 발음되는 한글음 모두에 대해서 각기 다른 한자 코드값을 가지고 있음을 알 수 있다.

(그림 2) 두개 이상의 한글 음을 갖는 한자 코드 배치
(Fig. 2) Hanja Code arrangement with two or more Hangul sound

표준한자 코드에 있는 한자로서 두 가지 이상의 코드값을 갖는 경우에는 큰 문제는 되지않으나 입력자가 한글음을 몰라서 알고 있는 다른 한글음으로 입력한 경우, 즉 한자폰트만 얻으면 되므로 다른 음을 사용하는 경우에는 심각한 문제가 된다. 한자·한글변환시 대부분의 에러는 여기에 해당한다.

예) 동자이음어

樂(악, 낙, 락, 요)→樂曲(악곡), 樂園(낙원), 娛樂(오락), 樂山(요산)

見(견, 현)→見地(견문), 謁見(알현)

예) 두음법칙 적용 문자

女(녀, 여)→長女(장녀), 女丈夫(여장부)

羅(라, 나)→網羅(망라), 羅列(나열)

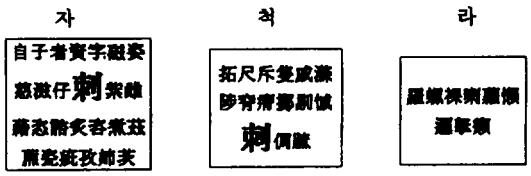
2.3.2 표준코드에 해당 한자음이 없는 경우

그러나 위에서와 같이 두개음 이상을 갖는 한자에 대해서 각각 다른 한자 코드값을 반복 배정하였다 하더라도 잘 쓰이지 않는 두개 이상의 한글 음을 갖는 한자 중 4888자에는 한자 글꼴이 존재하나 한자 코드값이 없는 관계로 한자를 입력할 수 없는 경우가 존

재한다.

예를 들어 "剌"는 (찌를)"자", (칼로 찌를)"척", (수라)"라"로 읽힌다. 그러나 "자"와 "척"은 4,888자에는 각각 존재하나 "라"는 없다. (그림 3)는 이 예를 보여주고 있다. 이렇듯 한자 코드값이 없는 글자를 입력자가 다른자를 이용하여 입력한 경우에는 한자·한글 변환시 심각한 문제가 될 것이다. 즉 다음 예에서처럼 수라상이 수자상 또는 수척상으로 변환하게 된다.

예) 水剌床(수라상)→ 수자상
수척상



(그림 3) 2개 이상의 한글음 모두를 표현하지 못하는 한자 코드체계

(Fig. 3) Hanja Code system that totally does not express two or more Hangul sound

2.3.3 한자와 한글의 음운체계의 차이

한자는 올바르게 입력되어 있으나 한자에는 사이시옷을 표기할 수 없기 때문에 변환시 생성되는 한글음이 달라지는 경우이다. 그러나 해당 문자가 2자이면서 모두 한자로 표기할 수 있는 용어는 다음 예의 여섯 가지밖에 없으므로 이 용어만 포함시키면 된다. 또한 '마굿간'과 같은 경우도 있다.

예) 貰房(세방)→ 貰房(셋방) 庫間(고간)→ 庫間(곳간)
車間(차간)→ 車間(찻간) 數字(수자)→ 數字(숫자)
回數(회수)→ 回數(횟수) 退間(퇴간)→ 退間(툇간)
馬廐間(마구간)→ 馬廐間(마굿간)

다만 한글과 한자가 섞여서 하나의 용어로 사용되는 경우가 있다. 이 중 한글이 선행하는 경우에는 한글이 한자의 음에 영향을 주지 않기 때문에 문제가 되지 않는다. '귓병(귓病)'과 같은 경우이다. 그러나 다음 예에서와 같이 한자가 앞에 오는 경우에는 한자음에 영향을 주는 것이 있기 때문에 이를 처리하여야 한다.

예) 胎줄(태줄)→ 胎줄(탯줄)
電氣불(전기불)→ 電氣불(전깃불)

1988년 한글맞춤법 개정안과 다른 경우로 '標準語規定' 가운데 '標準語査定原則'(제11, 13항)에 규정되어 있는 내용과 상이한 경우이다. 다음 예와 같이 한자 '着'은 모음의 발음변화를 인정하여 '책'으로, '句'는 단어의 일부가 될 때 '구'로 통일한다. 다만 '귀글', '글귀' 등에서는 '귀'를 쓴다는 예외 조항을 두고 있다.

예) 主着(주착)→ 주책(主着) 〈11항〉
귀절(句節)→ 구절(句節) 〈13항〉
문귀(文句)→ 문구(文句)
글句(글구)→ 글句(글귀) 〈예외〉

2.3.4 해당 한글음을 사용하지 않는 한자글꼴

梵語 등의 영향으로 해당 음이 없어서 다른 음을 빌려쓰는 경우에 해당하나 주로 불교용어에서 빈번하게 나타난다.

예) 初八日(초팔일)→ 初八日(초파일)
南無阿彌陀佛(남무아미타불)→ 南無阿彌陀佛(나무아미타불)

한자폰트는 있으나 해당 한글음에 해당하는 한자코드가 없어서 다른 음을 빌려온 경우이다.

예) 木瓜(목과)→ 木瓜(모과)
五六月(오륙월, 오육월)→ 五六月(오뉴월)

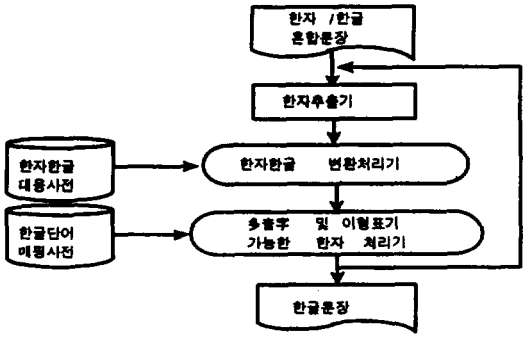
3. 한자·한글 변환기 설계

3.1 한자·한글 변환기 시스템 구성도

본 논문은 정보 검색 시스템에서 검색시 사용자들이 한자 정보에 대한 한글 음을 파악하는데 걸리는 시간을 절약하기 위해 한자·한글 혼합 문서들을 순수한 우리말 문서로 변환하는 한자·한글 변환기를 개발함으로써 한자에 익숙해 있지 않은 사용자들에게 정보를 쉽게 판독할 수 있도록 한다. (그림 4)는 개발하고자 하는 한자·한글 변환기의 구성요소를 보여 준다.

(그림 4)에서 보여주는 바와 같이 한자·한글 변환

기는 사용자들에게 빠른 정보 검색 서비스를 제공하기 위해 한자·한글 혼합 문장으로부터 한자를 추출하는 방법에 대한 연구를 수행하였으며, 추출된 한자를 대응하는 우리말로 변환하기 위해, 적합한 한자·한글 대응 사전을 구축하였다. 이를 위해 현재 환경에 적합한 한자 사전을 구축할 수 있는 방법을 유도한다. 또한 추출된 한자를 한자·한글 대응사전을 사용하여 정확한 우리말로 변환하는 처리기를 설계하고 개발하였다. 한자는 대부분 한개의 한글 음을 갖고 있지만 두개 이상의 한글 음을 갖고 있는 문자들도 상당히 있다. 두개 이상의 한글음을 갖는 한자들과 이형표기가 가능한 한자들에 대해서 올바른 한글음으로 변환할 수 있는 한글단어 매핑사전 구축에 관한 연구를 수행하며, 한글단어 매핑사전을 이용하여 多音字 및 이형표기 가능한 한자 처리기를 설계한다.



(그림 4) 한자·한글 변환기 시스템 구성도

(Fig. 4) Hanja/Hangul Converter System Architecture

3.2 한자·한글 변환기 설계

3.2.1 한자·한글 혼합 문서에서 한자 추출

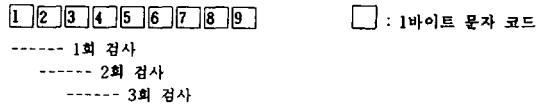
거의 모든 완성형 텍스트 화일이 한자와 한글만 혼합되어 문서가 작성되는 것이 아니라 영문과 사용자 정의 문자등 작성자가 사용할 수 있는 모든 문자들이 혼합되어 작성된 화일이 많다. 이러한 완성형 텍스트에서 한자나 한글은 2바이트 완성형으로 구성되어 있으며 각각의 코드는 한글의 두바이트 중 하위 바이트가 한자의 상위바이트와 겹치게 된다. 그러므로 한자 코드의 추출를 위해서는 두바이트를 동시에 검사하여야 한다. 다음은 완성형 텍스트 화일에서 각각의 문자들이 갖을 수 있는 코드형태는 〈표 1〉과 같다.

〈표 1〉 KS 완성형 텍스트에서 갖을 수 있는 코드 형태

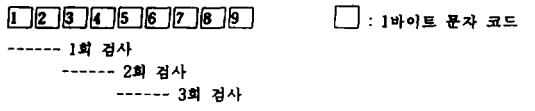
〈Table 1〉 KS text code type

상위바이트	하위바이트	코드형태
상 < 128		ASCII
0xb0 <= 상 <=0xc8	0xa1 <= 하 <=0xfe	글자한글
상=='/241'	하=='/241'	2바이트 공백 문자
0xa4 <= 상	0xa1 <= 하 <= 0xfe	자소 한글
0xa1 <= 상 <=0xac	0xa1 <= 하 <=0xfe	2바이트 특수 문자
0xca <= 상 <=0xfd	한자	

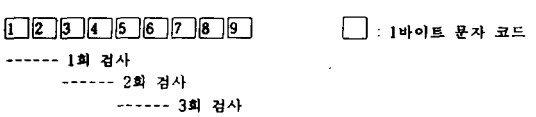
1회 검사시 상위 바이트(1번 문자코드)가 값이 128 이하 이면 ASCII 코드 형태의 1바이트 문자이므로 다음 바이트부터 검사(2회 검사)를 다시한다.



그러나 아래 그림에서와 같이 1회 검사시 상위 바이트(1번 문자코드)가 값이 128 이상 이면 한글 또는 한자코드이다. 상위 바이트가 한글 코드에 속하게 되면 다음 바이트(2번 문자코드)를 검사하지 않고, 2회 검사(3번과 4번 문자코드에 해당)를 다시 시작한다. 이와 같은 이유는 혼합 문서로부터 한자코드 추출을 위한것이다.



그렇지 않고 1회 검사시 상위 바이트가 값이 128이상 이며 상위 바이트가 한자 코드에 속하게 되면 다음 바이트(2번 문자코드)를 검사하고 그 데이타를 한자·한글 변환기로 넘긴다. 또 다시 2회 검사부터 텍스트 문서의 마지막까지 위의 수행과정을 반복 실행한다.



3.2.2 한자·한글 변환을 위한 한자·한글 대응 사전 구축

한자·한글이 혼합되어 있는 텍스트에서 한자를 추출하여 각 한자에 해당하는 한글를 출력하기 위해서는 표준 한자에 대응하는 한글 사전을 구축하여야 한다. 한글 사전은 표준 한자 4888자에 대해서 한자 코드값의 순서대로 저장된 화일에 해당하는 한글을 텍스트 화일로 만들어 사전을 구축한다. 다음 (그림 5)은 각 한자에 대응하는 한글 코드값을 가지고 있는 한글 화일을 보여주고 있다. A의 한자 코드값에 해당하는 B의 한글 코드 사전(화일)을 두어 1:1으로 대응되도록 하였다.

A(hanja.cod)

伽佳假價加可呵哥嘉嫁家暇架枷柯
歌珂痂稼苛茄街袈訶賈跏軻迦駕刻
却各恪慤殼珏脚覺角閣侃刊墾奸姦
干幹懇揀杆柬桿澗癎看磵稈竿簡肝
艮艱諫間乫喝曷渴碣竭葛褐蝎鞨勘
坎堪嵌感憾戡敢柑橄減甘疳監瞰紺
邯鑑鑒龕匣岬甲胛鉀閘剛堈姜岡崗
康强彊慷江畺疆糠絳綱羌腔舡薑襁
講鋼降鱇介价個凱塏愷愾慨改槪漑

B(hangul.cod)

가가가가가가가가가가가가가
가가가가가가가가가가가가각
각각각각각각각각간간간간
간간간간간간간간간간간간간
간간간간갈갈갈갈갈갈갈갈갈감
감감감감감감감감감감감감감
감감감감갑갑갑갑갑갑강강강강강
강강강강강강강강강강강강강
강강강강개개개개개개개개개개

(그림 5) 각 한자에 대응하는 한글 코드값을 갖는 화일
(Fig. 5) File with a Hangul code value corresponding to each Hanja

3.2.3 한자·한글 변환

혼합 문서에서 추출된 한자를 한글로 변환하기 위해서 한자 변환 테이블을 이용하게 된다. 한자 변환 테이블의 내용은 한자 코드순으로 한자를 한글 코드로 표현한 것이다. 추출된 한자 코드를 이용하여 한자 변환 테이블의 인덱스를 구해 한글 코드를 얻을 수 있다. 한자 코드 영역은 상위 1바이트 0xCA~0xFD과 하위 1바이트 0xA1~0xFE로 각각 94개의 한자 코드를 가진 53개의 블럭으로 구성된다.

한자 코드를 한글 코드로 변환하는 과정은 다음과 같다.

- 한자 코드의 상위 1바이트에서 0xCA를 감하여 블럭 번호를 구한다.
- 한자 코드의 하위 1바이트에서 0xA1를 감하여 그 블럭내 위치를 구한다.
- 블럭 번호에 94를 곱한 값과 블럭내 위치를 더해 한자 변환 테이블에서의 절대위치를 구한다.
- 이 절대위치를 변환 테이블의 인덱스로하여 한글 코드를 얻는다.

3.2.4 한글단어 매핑사전 구축에 관한 연구

한글단어 매핑사전은 3장에서 제시되었던 문제들를 고려하여 만들어졌다. 한글단어 매핑사전에는 3절에서 조사한 한자단어의 이형표기 또는 원음표기와 그러한 한자단어의 바른 한글표기가 등록된다. 그리하여 한자·한글 변환 과정중 표준코드에 의해 생성된 한자음은 한글단어 매핑사전의 이형표기 또는 원음표기를 검색하여 여기에 해당하면 바른 한글표기를 참조하여 올바른 한글음을 생성하게 된다. 3절에 제시되었던 각각의 문제점들을 고려하여 매핑사전에 한자단어를 등록한 예를 들면 다음과 같다.→의 좌변은 한자의 이형이나 고유의 한자음이며 우변은 한자를 한글로 변환했을 때 바른 한글표기이다.

예) 동자이음 및 두음법칙 적용 문자:
여러 가지 이형표기중 정확한 한자음만 바른 표기로 등록한다.
樂曲(악곡):낙곡, 락곡, 악곡, 요곡→악곡
見聞(견문):견문, 현문→견문 羅列(나열):나열, 라열→나열
女丈夫(여장부):여장부, 녀장부→여장부

예) 표준코드에 해당 한자음이 없는 경우:
누락된 한자음을 보충하여 바른 표기로 등록한다.

水剌床(수라상): 수자상, 수척상→수라상

예) 한자와 한글의 음운체계의 차이:

원음표기를 수정하여 바른 표기로 등록한다.

黃房(샛방): 세방→샛방 廊間(곳간): 고간→곳간

數字(숫자): 수자→숫자 回數(횟수): 회수→횟수

主着(주책): 주착→주책

句節(구절): 귀절→구절 文句(문구): 문귀→문구

馬廐間(마굿간): 마구간→마굿간

예) 해당 한글음을 사용하지 않는 한자글꼴:

원음표기를 수정하여 바른 표기로 등록한다.

初八日(초파일): 초팔일→초파일

南無阿彌陀佛(나무아미타불): 남무아미타불→나무아미타불

木瓜(모과): 목과→모과

五六月(오뉴월): 오륙월, 오육월→오뉴월

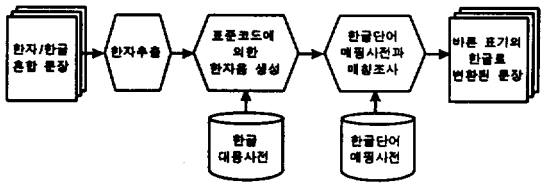
위의 예를 한글단어 매핑사전으로 만들면 다음 〈표 2〉와 같다.

〈표 2〉 한글단어 매핑사전

〈Table 2〉 Hangul mapping dictionary

한자	이형표기 및 원음표기	바른표기
樂曲	낙곡 락곡 악곡 요곡	악곡
見聞	견문 현문	견문
羅列	나열 라열	나열
水剌床	수자상 수척상 수척상	수라상
數字	수자	숫자
回數	회수	횟수
主着	주착	주책
句節	귀절	구절
文句	문귀	문구
南無阿彌陀佛	남무아미타불	나무아미타불
木瓜	목과	모과
五六月	오륙월 오육월	오뉴월

한자·한글 변환시 바른 한글음을 생성하기 위하여 KS C 5601 코드 체계를 벗어나지 않는 범위 내에서 사용할 수 있도록 한글단어 매핑사전을 구축하기 위하여 (그림 6)와 같은 방법으로 구현하였다.



(그림 6) 한글단어 매핑사전을 이용한 多音字 및 이형표기 가능한 한자처리 과정

(Fig. 6) Hanja process with two or more sound and different mark using Hangul mapping dictionary

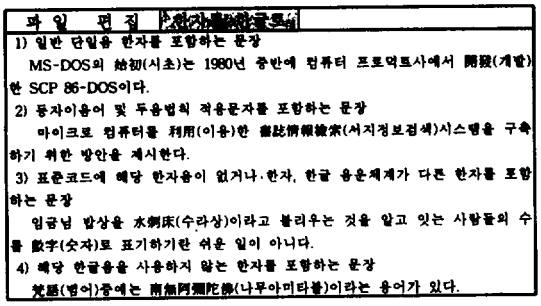
4. 한자·한글 변환기 구현 및 분석

한자·한글 변환기는 PC-586과 SUN SPARCstation에서 C언어를 사용하여 구현되었으며 한자·한글 변환기에서 사용한 테스트 데이터는 워드프로세서를 사용하여 작성하였다. 구현과정에서 多音字 및 이형표기 가능한 한자처리를 위한 한글단어 매핑사전 구축시 특히 한자를 한글로 변환하는데는 일정한 규칙성이 존재하지만 그렇지 않은 경우도 있다. 그렇기 때문에 규칙성을 이용하기보다는 이러한 모든 경우를 포괄할 수 있는 대응사전을 이용하여 특수한 한자를 한글로 변환하였다.

본 연구에서 간단한 변환의 시험은 (그림 7)에서 보여주고 있는 일반 단일음을 포함하는 문장, 한자코드를 따로 부여받은 동자이음어 및 두음법칙 적용문자를 포함하는 문장, 표준코드에 해당 한글음이 없거나 한글 음운체계가 다른 한자를 포함하는 문장, 해당 한글음을 사용하지 않는 한자를 포함하는 문장등과 같은 우리가 고려했던 문장들과 "金基泰, 李時鎔, 敎育史·哲學 중 第4章 20世紀 後半期 敎育哲學에서 實存主義의 背景 p.498-506를 대상으로 하였다. 입력한 데이터에서 한자용어의 수는 795개였다. 이중 단일음을 포함하는 한자 472자, 동자이음어 및 두음법칙에 해당하는 한자 178자, 기타로 구성하였다. 한글단어 매핑사전에 등록한 데이터는 300개의 한자단어로 주로 테스트에 사용한 한자 위주로 구성하였다.

변환된 한글용어중 多音字에 해당하지 않는 한자

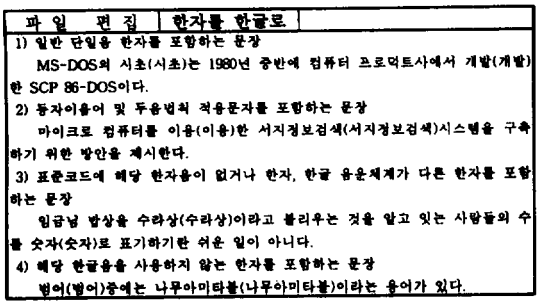
인 경우는 한글대응사전을 참조하여 그대로 생성하고 그렇지 않는 경우, 즉 이형가능한 문자를 포함하는 경우에는 한글단어 매핑사전을 참조하여 매칭이 되면 바른 표기로 대체하게 된다. 변환된 문서중에 입력된 한자가 잘못된 경우를 제외하고는 변환시 에러가 없었으며, 변환시 에러의 대부분은 한글단어 매핑사전에 등록되지 않는 동자이음어와 두음법칙에 해당하는 한자가 주류를 이루었다. 결국 완벽한 한자·한글 변환기 구현은 한글단어 매핑사전의 완전성 여부에 달려있다. (그림 8)는 (그림 7)와 같이 입력된 한자·한글 혼합 문서들이 위에서 설계한 처리기를 통하여 정확한 한글음을 갖는 문서로 변환되는 예를 보여주고 있다.


파 일 편 집 [illegible]

1) 일반 단일음 한자를 포함하는 문장
MS-DOS의 始初(시초)는 1980년 중반에 컴퓨터 프로덕트사에서 開發(개발)한 SCP 86-DOS이다.
2) 동자이음어 및 두음법칙 적용문자를 포함하는 문장
마이크로 컴퓨터를 利用(이용)한 書誌情報檢索(서지정보검색)시스템을 구축하기 위한 방안을 제시한다.
3) 표준코드에 해당 한자음이 없거나·한자, 한글 음운체계가 다른 한자를 포함하는 문장
임금님 밥상을 水刺床(수라상)이라고 불리우는 것을 알고 있는 사람들의 수를 數字(숫자)로 표기하기란 쉬운 일이 아니다.
4) 해당 한글음을 사용하지 않는 한자를 포함하는 문장
梵語(범어)중에는 南無阿彌陀佛(나무아미타불)이라는 용어가 있다.

(그림 7) 한자/한글 혼합 문장

(Fig. 7) Hybrid statements with Hanja/Hangul


파 일 편 집 | 한자를 한글로

1) 일반 단일음 한자를 포함하는 문장
MS-DOS의 시초(시초)는 1980년 중반에 컴퓨터 프로덕트사에서 개발(개발)한 SCP 86-DOS이다.
2) 동자이음어 및 두음법칙 적용문자를 포함하는 문장
마이크로 컴퓨터를 이용(이용)한 서지정보검색(서지정보검색)시스템을 구축하기 위한 방안을 제시한다.
3) 표준코드에 해당 한자음이 없거나 한자, 한글 음운체계가 다른 한자를 포함하는 문장
임금님 밥상을 수라상(수라상)이라고 불리우는 것을 알고 있는 사람들의 수를 숫자(숫자)로 표기하기란 쉬운 일이 아니다.
4) 해당 한글음을 사용하지 않는 한자를 포함하는 문장
범어(범어)중에는 나무아미타불(나무아미타불)이라는 용어가 있다.

(그림 8) 혼합 문장으로부터 변환된 문장

(Fig. 8) Statements with only Hangul translated from hybrid ones

5. 결론 및 향후 연구 방향

본 논문에서는 정보 검색 시스템에서 관리하는 한자·한글 혼합 문서들을 사용자들이 쉽게 읽을 수 있도록 올바른 우리말 문서로 변환하는 한자·한글 변환기를 설계하고 구현하였다. 이를 위해 KS 완성형 코드에 포함되어 있는 표준 한자코드에 대해서 조사 분석하였으며, 두개 이상의 한글 음을 갖는 한자들을 정확한 한글들로 변환할 수 있는 방법을 표준 한자코드에서 발견하였다. 또한 한자·한글 변환시 바른 한글음을 생성하기 위하여 사전에 고려하여야 할 사항으로 문자에 해당하는 문제 즉, 동자이음 처리, 두음법칙 적용 문자의 처리, 표준코드에 해당 한자음이 없는 경우, 한자와 한글의 음운체계의 차이에서 나타나는 표기상의 문제 등으로 나누어 조사하였다.

위와 같은 조사, 분석 및 연구된 내용들을 바탕으로 사용자들에게 보다 빠르고 정확한 정보 전달을 위해 한자·한글 혼합 문장으로 부터 한자 추출을 위한 한자 추출기를 설계하였으며 한자 추출기로부터 추출된 한자를 대응하는 한글로 변환하기 위해 한자·한글 대응 사전을 구축하였다. 한자·한글 대응 사전을 사용하여 한자·한글 혼합 문장으로 부터 추출된 한자를 우리말로 변환하는 변환 처리기를 설계하였다. 또한 한자·한글 변환시 올바르지 못한 한글음 생성을 방지하기 위하여 한글단어 매핑사전을 구축하였다.

한자·한글 변환시 바르지 못한 한글음 생성을 방지하기 위해 한글단어 매핑사전을 사용하였는데 향후 연구 방향으로는 처리분야에 따른 동자이음과 이형표기가 가능한 용어들의 조사가 다양하게 이루어져야 할 것이며 이를 실제 시스템에서 사용하면서 보완되어야 할 것이다. 또한 세계 표준으로 제정된 유니코드 전체영역 65,000여자 중에 한국, 중국, 일본 등 한자문화권의 나라에서 사용하는 한자 20,000여자에 대한 한자·한글 변환기를 설계하고 구현하는것이다.

참 고 문 헌

[1] 미승우, "새맞춤법과 교정의 실제", pp. 308-417, 어문각, 1993.

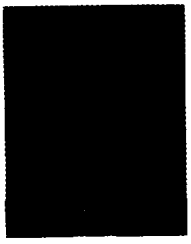
[2] 안병희, 이희승, "고친판 한글맞춤법 강의", pp. 55-116, 신구문화사, 1996.

[3] 유재수, 최종 연구 보고서, "한자·한글 변환기 원형 개발", 한국과학기술원연구개발정보센터. 1996.

[4] 이현호, "한글 라이브러리〈한〉", pp. 3-15, 가남사, 1992.

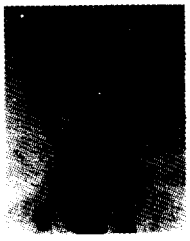
[5] 이준희, 정내권, "컴퓨터 속의 한글", pp. 17-107,

정보시대, 1991.
[6] 이춘택, "韓日 國家規格漢字코드의 統合硏究", 중앙 대학교 대학원 도서관학과 자료조직전공 박사 학위논문, 1991.
[7] 임현모, "컴퓨터와 한글의 만남", pp. 23-38, 정보문화사, 1992.
[8] 최석두, "한자용어로부터 한글색인어의 생성", 한글 및 한국어 정보처리학회 학술대회, pp. 51-58, 1996.
[9] 최원식, "한자한글 그래픽 데이터베이스", pp. 197-240, 가남사, 1992.



강 형 일

1996년 목포대학교 전산통계학과(학사)
1998년 목포대학교 전산통계학과(석사)
1998년 충북대학교 정보통신공학과 박사과정
1996년 9월~1997년 8월 목포대학교 전산통계학과 조교
관심분야: 멀티미디어 데이터베이스시스템, 정보검색, 저장구조, 분산객체 컴퓨팅 등



강 승 헌

1997년 목포대학교 전산통계학과(학사)
1997년 3월~현재 충북대학교 정보통신공학과 석사과정
관심분야: 데이터베이스시스템, 정보검색, 내용기반 이미지 검색 등

장 수 민

1997년 목포대학교 전산통계학과(학사)
1997년 3월~현재 충북대학교 정보통신공학과 석사과정
관심분야: 데이터베이스시스템, 정보검색, 개임소프트웨어 등

유 재 수

1989년 전북대학교 공과대학 컴퓨터공학과(학사)
1991년 한국과학기술원 전산학과(공학석사)
1995년 한국과학기술원 전산학과(공학박사)
1995년~1996년 8월: 목포대학교 전산통계학과 전임강사
1996년 8월~현재: 충북대학교 공과대학 정보통신공학과 조교수
관심분야: 데이터베이스 시스템, 정보검색, 멀티미디어 데이타베이스, 분산객체컴퓨팅 등