한자용어로부터 한글색인어의 생성

최석두
이화여자대학교 문헌정보학과

A Generation of Hangul Index Term from Hanja Term

Suk Doo Choi
Dept. of Lib. & Inf. Sci., Ewha Womans Univ.

요 약

漢字로 기술된 용어를 한글로 자동변환하여 색인어로 사용하는 경우에 한글의 음운체계나 해당 시스템의 색인정책에 맞지 않는 일이 생기게 된다. 이런 문제가 생기는 원인은 해당 한자에 대응하는 정확한 한글을 입력하지 않고 변환하였을 경우, 해당 한자의 음이 없거나 한자와의 음운체계가 달라 생기는 경우 및 별도의 색인정책이 있는 경우 등을 생각할 수 있다. 본고에서는 KS C 5601 표준코드(이하 표준코드라 한다)를 기준으로 漢字의 多音字를 조사하였다. 多音字가 포함되어 있는 사전용어와 多音字파일을 이용하여 매핑파일을 구축함과 동시에 매핑파일을 보완함으로써 漢字로 기술된 용어의 바른 한글음을 자동생성하여 색인어로 사용할 수 있는 방안에 대하여 논한다.

1 서론

컴퓨터에 입력된 문헌의 書名이나 全文데이터는 상당부분 漢字를 포함하고 있으며, 최근에는 古書의 全文처리까지도 컴퓨터시스템에 의존하는 경우가 늘어나고 있다. 한자로 표기된 용어를 이용하여 한글색인어를 자동변환하게 되면 생성된 한글색인어가 원하는 바른 표기로 변환되지 않는 경우가 발생하게 된다. 또한 생성된 한글음이 바르다 할지라도 색인시스템의 정책상 다른 형의 한글색인어로 대체해야 하는 경우도 있다.

원하는 한글색인어가 생성되지 않는다는 것은 자동색인을 전제로 하고 있는 시스템에서는 대단히 심각한 문제로 대두되고 있다.

본고에서는 이를 해결하기 위하여 표준코드를 기준으로 漢字의 多音字를 조사하고, 이 多音字가 포함되어 있는 용어를 이용하여 매핑파일을 구축하며, 색인정책에 따른 대체어를 보완함으로써 한자용어에서 바른 한글색인어를 생성할 수 있는 방안에 대하여 논하고자 한다.

다만, 이 실험은 多音字파일을 매뉴얼로 만들고 한자·한글쌍의 매핑파일을 자동생성하며, 일부의 예외 용어를 보완하여 한글색인어의 생성 가능성을 시험한 것으로 완전한 데이터를 구축하기 위한 것은 아니다. 또한 漢字가 포함되어 있는 단위 후보색인어로서 한글음을 취한 후의 문자열을 처리의 대상으로 하며, 후보색인어 추출까지의 과정은 논외로 한다.

2 다른 한글색인어의 생성원인

한자가 포함되어 있는 다음과 같은 텍스트에서 키워드를 자동생성하는 경우를 보자.

> "……綜合的인 文獻探索을 必要로 할 것이므로 關聯分野의 網羅的 檢索 즉, 再現率이 높은 檢索을

원할 것이다. 反面에 製品生産이나 品質改善을 위한 特殊問題解決, 論文記事를 위한 特殊情報, 또는 短期硏究프로젝트 등을 수행하는 사람에게는 問題解決을 위한 特殊主題에 관한 正確한 少數의 最新情報 즉, 精度率이 높은 檢索이 더 重要할 것이다....."(尹龜鎬. 情報檢索效率에 관한 硏究. 圖書館學, 8:73-102(1981) 중 p.94의 일부분).

상기 텍스트를 대상으로 생성된 한글색인어에서 聯(연, 련), 羅(라, 나), 率(율, 률), 論(론, 논)이 포함되어 있는 용어가 다르게 표현될 수 있다는 것을 직감적으로 알 수 있을 것이다. 실제로는 보다 많은 경우의 수가 발생하게 된다.

한자용어에서 한글색인어의 변환시 다른 한글색인어가 생성되는 원인을 세분해 보면 다음 몇 가지로 나눌 수 있다.

첫째, 일반적으로 입력자는 한자 폰트의 획득이 목적일 뿐, 역으로 생성될 한글음에는 관심이 없다.
둘째, 정확한 한글표기를 몰라서 다른 한글음을 이용하여 변환한다.
세째, 표준코드에 해당 한자의 음이 없어서 다른 음의 한자를 빌려서 입력한다.
네째, 한자와 한글의 음운체계가 달라 생성된 한글음이 맞지 않는다.
다섯째, 한글맞춤법 1988년 개정안을 잘 모른다.
여섯째, 한글색인어의 일관성을 위한 통제정책이 시스템마다 다를 수 있다.

이와 같은 원인을 문자에 따른 문제와 색인어 통제정책에 따른 문제로 크게 구분하고 다시 각각의 경우를 세분하여 보면 다음과 같다. 각 예에서 기호←의 좌변은 색인어이며 우변은 이형, 기호→의 좌변은 이형이며 우변은 색인어이다.

2.1 문자에 따른 문제

1) 다른 음을 사용한다 : 해당음이 있지만 한자폰트만 얻으면 되므로 다른 음을 사용하거나 해당 한글음을 몰라서 입력자가 알고 있는 음으로 입력한 경우이며 에러의 대부분이 여기에 해당한다.

예) 樂山(요산) ← 樂山(악산, 낙산, 락산)
殺到(쇄도) ← 殺到(살도)
衛戍(위수) ← 衛戍(위술)
標識(표지) ← 標識(표식)

2) 해당 음이 없어서 다른 음을 빌려쓴다 : 한자폰트는 있으나 해당 한글음에 해당하는 한자코드가 없어서 다른 음을 빌려온 경우이다. 입력자는 동일한 한자폰트를 차용할 수밖에 없다.

예) 五六月(오뉴월) ← 五六月(오륙월, 오육월)
木瓜(모과) ← 木瓜(목과)
盟誓(맹세) ← 盟誓(맹서)
白魚(뱅어) ← 白魚(백어)
分錢(푼전) ← 分錢(분전)
林巨正(임꺽정) ← 林巨正(임거정)

3) 사이시옷을 표현할 수 없다 : 한자는 옳게 입력되어 있으나 한자에는 사이시옷을 표기할 수 없기 때문에 생성되는 한글이 달라지는 경우이다. 그러나 해당 문자가 2자이면서 모두 한자로 표기할 수 있는 용어는 다음 예1의 여섯 가지밖에 없으므로 이 용어만 포함시키면 된다(미승우, 1993).

다만 한글과 한자가 섞여서 하나의 용어로 사용되는 경우가 있다. 이 중 한글이 선행하는 경우에는 문제가 되지 않는다. 한글이 한자의 음에 영향을 주지 않기 때문이며, 예2의 "귓병(귓病)"과 같은 경우이다. 그러나 예2의 "탯줄(胎줄-태줄), 전깃불(電氣불-전기불)"과 같이 한자가 앞에 오는 경우에는 한자의 음에 영향을 주는 것이 있기 때문에 이를 처리하여야 한다. 또한 "마굿간"과 같은 경우도 있다.

예1) 庫間(곳간) ← 庫間(고간)
貰房(셋방) ← 貰房(세방)
數字(숫자) ← 數字(수자)
車間(찻간) ← 車間(차간)
退間(툇간) ← 退間(퇴간)
回數(횟수) ← 回數(회수)

예2) 귓病(귓병) "문제없음"
胎줄(탯줄) ← 胎줄(태줄)
電氣불(전깃불) ← 電氣불(전기불)
馬廐間(마굿간) ← 馬廐間(마구간)

4) 梵語 등의 영향으로 다른 한글음을 사용한다 : 해

당 음이 없어서 다른 음을 빌려쓰는 경우에 해당하나 주로 불교용어에서 빈번하게 나타난다.

예) 初八日(초파일) ← 初八日(초팔일)
婆羅(바라) ← 婆羅(파라)
陀羅尼(다라니) ← 陀羅尼(타라니)
南無阿彌陀佛(나무아미타불) ← 南無阿彌陀佛 (남무아미타불)

5) 관례대로 사용한다 : 특별하게 음을 붙여서 사용해 온 관례에 따르는 경우이다. 대부분 해당 음의 漢子가 없기 때문에 다른 음의 동일한 한자폰트를 차용하게 된다. 예를 들면 "寺"의 의미가 사원인 경우에는 "사"이나 직명을 나타낼 때는 "시"가 된다.

예) 內人(나인) ← 內人(내인)
司僕寺(사복시) ← 司僕寺(사복사)
奉常寺(봉상시) ← 奉常寺(봉상사)
宮商角徵羽(궁상각치우) ← 宮商角徵羽(궁상각징우)

6) 1988년 한글맞춤법 개정안과 다르다 : 1988년 1월 19일 문교부가 새롭게 개정고시하여 1989년 3월 1일부터 시행하도록 한 "標準語規定" 가운데 "標準語査定原則"(제11, 13항)에 규정되어 있는 내용과 상이한 경우이다. 다음 예와 같이 漢子 "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(주책) ← 主着(주착) "11항"
句節(구절) ← 句節(귀절) "13항"
文句(문구) ← 文句(문귀)
글句(글귀) ← 글句(글구) "귀를 쓴다"

7) 한자의 모양과 의미가 비슷하다 : 한자음과 관계없이 한자의 모양과 의미가 비슷해서 잘못 사용하는 경우이다. 예1)은 음이 같을 때이며, 예2)는 음이 다를 때이다. 잘못 입력된 용어가 사용되지 않는 용어이면 어느 정도 처리가 가능하나 별개의 의미를 갖는 용어라면 처리에 문제가 있다.

예1) 家政婦(가정부) ← 家庭婦(가정부)
錄音器(녹음기) ← 錄音機(녹음기)
辨證法(변증법) ← 辯證法(변증법)
十誡命(십계명) ← 十戒命(십계명)

예2) 容恕(용서) ← 容怒(용노)
罹災民(이재민) ← 羅災民(나재민)
薄命(박명) ← 簿命(부명)
辦務官(판무관) ← 辨務官(변무관)

2.2 색인어 통제정책에 따른 문제

색인어의 통제정책에 따른 문제는 시스템에 종속적인 문제이며, 대부분 그 분야에서 일반적으로 사용하고 있는 용어가 기준이 된다. 따라서 해당 분야의 시소러스가 있다면 상당 부분이 해결될 것이다. 다만, "색인어 통제정책에 따른 문제"의 용어에 "문자에 따른 문제"가 다시 영향을 주게 된다. 정책과 관련되는 것을 몇 가지 측면으로 나누어 보면 다음과 같다.

1) 외국의 국명 및 지명 등을 한자로 대응시킨 경우이다 : 언제나 문제가 되는 것은 아니나 시스템에 따라 색인어를 다음 예의 좌변과 같이 원음을 쓰기로 결정하면 문제가 된다. "오스트렐리아 ↔ 濠洲(호주)"와 같이 "호주"를 많이 쓰는 경우에는 용어정책에 따라 어느 쪽이라도 택할 수 있을 것이다.

예) 스페인 ← 西班牙(서반아)
프랑스 ← 佛蘭西(불란서)
네덜란드 ← 和蘭(화란)
인도네시아 ← 印尼(인니)
홍콩 ← 香港(향항)

"스페인어 ← 西班牙語"처럼 복합명사가 될 때도 동일하다. 그러나 佛蘭西語/(佛語, 프랑스語), 佛蘭西文學/(프랑스文學, 佛文學)과 같은 경우에는 명확한 기준이 있어야 할 것이다.

2) 일정한 원칙하에 다른 동의어로 바꾸어 색인어로 사용하는 경우이다 : 한자표기로서는 무방하나 그 한자의 한글음이 모호하거나 흔히 쓰이지 않는 등의 이유로 다른 한글색인어로 변환하고자 하는 경우이다. 이를 위하여 모든 대상 용어에 대하여 미리 동의어를 조사하고 우선어를 결정한다는 것은 매우 어려운 일이므로 기존의 시소러스가 있는 경우에는 "USE, UF"관계를 이용하는 것이 바람직할 것이다.

예) 馬(말 "동물") ← 馬(마)
海苔(김 "식물") ← 海苔(해태)
大豆(콩) ← 大豆(대두)
컴퓨터 ← 電子計算器(전자계산기)
컴퓨터 ← 電算器(전산기)

3) 약어를 사용하는 경우이다 : 고유명사와 일반용어를 불문하고 언어 전반적으로 한자용어에 대한 약어의 사용은 빈번하며, 사용된 약어 대신에 완전명을 한글색인어로 사용하고자 하는 경우에 발생한다. 마찬가지로 시소러스의 "USE, UF"관계를 이용하는 것이 바람직할 것이다.

예) 特別委員會(특별위원회) ← 特委(특위)
勤勞所得稅(근로소득세) ← 勤所稅(근소세)
勞動組合(노동조합) ← 勞組(노조)
國立科學搜査硏究所(국립과학수사연구소) ← 國科搜(국과수)

4) 한자문화권 자료내의 표기를 우리 식으로 변환하는 경우이다 : 중국, 일본 등의 자료를 처리하기 위해서는 표준코드로 변환해야 하며, 표준코드로 변환한 후 한글색인어로 통합하는 정책일 때는 다음 예와 같은 문제가 생기게 된다.

예1)은 고유명사, 예2)는 일반용어의 예이다. 본고에서는 한자가 포함되어 있지 않은 용어는 제외하였으나 일반용어가 대상이 될 때는 한자가 없는 것도 포함하여 다양한 측면을 고려해야 하므로 다국어시소러스를 만들어 일괄적으로 처리하는 것이 바람직할 것이다. 다만 이 문제는 다국어간 용어처리문제의 전체적인 측면에 볼 때 상당히 지엽적인 것이므로 본 연구의 매핑테이블에서의 시험대상에서는 제외한다.

예1) 獨島(독도) ← 竹島(죽도) "일본어"
美國(미국) ← 米國(미국) "일본어"
서울 ← 漢城(한성) "중국어"
캠브리지 ← 劍橋(검교) "중국어"
옥스포드 ← 牛津(우진) "중국어"

예2) 下女(하녀) ← 女中(여중) "일본어"
手票(수표) ← 切手(절수) "일본어"
컴퓨터 ← 電腦(전뇌) "중국어"
스프트웨어 ← 軟件(연건) "중국어"
핫도그 ← 熱狗(열구) "중국어"
칵테일 ← 鷄尾(계미) "중국어"

3 매핑파일의 구축

3.1 표준코드의 분석

두 가지 이상의 음을 갖는 한자가 포함되어 있는 용어의 처리대상이 가장 많으므로 우선 표준코드내에서 복수의 음을 갖는 漢子를 분석할 필요가 있다. 이춘택(1991)의 조사 결과를 보면 표준코드에 있는 한자로서 두 가지 이상의 음을 갖는 문자의 수는 다음과 같이 2,187자이다. 이 통계는 동아출판사 발행의 "漢韓大辭典"(李家源, 權五恂, 任昌淳 監修)을 기본으로 "大字源" (張三植 著)을 참고하였으며, 그 외 일부를 추가한 것이다.

2音字	805자	(1,610자)
3音字	139자	(417자)
4音字	28자	(112자)
5音字	7자	(35자)
6音字	1자	(6자)
7音字	1자	(7자)
계	980자	(2,187자)

多音字의 조사는 사용하는 한자사전에 따라 다소 달라지는 경향이 있으므로 대상 데이터에 따라 참조하는 사전을 달리 선택하여야 할 것이다. 본 연구에서는 상기 조사결과를 그대로 이용하였다.

또한 표준코드에 있는 漢子로서 두 가지 이상의 코드를 갖는 문자의 수(즉, 두 가지 이상의 음을 표준문자내에서 갖는 문자; 重出字)는 다음과 같다(이춘택, 1991).

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가 동시에 사용된 문자로서 암(岩, 巖)과 만(万, 萬) 등이 있으나 바른 한글음의 생성에는 영향을 주지는 않는다. 다만 어느 한쪽만을 쓰겠다는 시스템에서는 고려대상이 된다.

3.2 多音字 파일의 작성

파악된 표준코드내의 모든 多音漢子에 대하여 다음 예와 같이 각 音(표준코드에 없는 音도 포함한다)을 조사하여 多音字파일을 작성하였다. 조사결과로 多音字파일에 수록된 문자수는 두 가지 이상의 음을 갖는 문자 2,187자와 重出字 530자를 포함하여 도합 2,543자가 된다. 문자수가 줄어 든 것은 양쪽 집합에 속하는 문자가 있기 때문이다.

多音字파일에서 다음 예와 같이 첫 번째 숫자(356 및 10)는 글자의 그룹을 나타내는 일련번호이며 같은 숫자는 같은 문자군을 나타낸다. 이 그룹코드는 이형용어의 자동생성시 참조데이터로 사용한다. 두 번째 숫자가 1이면(예2 참조) 표준코드에 없는 음이라는 것을 나타낸다. 따라서 표준코드에 없는 음에 대한 한자코드는 같은 그룹의 문자를 임의로 사용하게 된다. 괄호내의 코드는 해당 한자의 코드(16진코드)이며 多音字파일에는 포함되지 않는다.

예1) 樂 낙 356 (E3E2)
樂 락 356 (E5A5)
樂 악 356 (ED55)
樂 요 356 (EF9B)

예2) 假 가 10 (CAA3)
假 하 10 1 (CAA3)

표준코드에 없는 음을 구별하는 것은 표준코드에 두 가지 이상의 음이 있을 때 어느 음의 한자를 차용했는지를 알 수 없으므로 차후 한자코드가 확장될 때를 대비하기 위한 것이다.

3.3 기본 매핑파일의 생성

어휘에서 多音字에 의해 생길 수 있는 이형을 조사하고 조사결과를 이용하여 이형 중 처리대상에서 제외시킬 수 있는 상식적인 규칙을 찾아낼 수 있을 것이다. 예를 들면, 다음과 같이 "살인, 여성, 희노애락, 김해" 등의 용어를 "쇄인, 녀성, 희노애요, 금해" 등으로 읽어 漢子로 변환할 확률은 매우 낮으며, 특히 "金"이라는 문자는 姓을 포함하여 고유명사이면 거의 "김"으로 읽고 있다(그렇지 않은 예도 많다. 예를 들면, "金川, 金華, 金村" 등의 지명은 "금천, 금화, 금촌"으로 읽으며, 북한에서는 "녀성"이라는 표기를 사용하고 있다).

예) 살인(殺人) 쇄인(殺人)
여성(女性) 녀성(女性)
희노애락(喜怒哀樂) 희노애요(喜怒哀樂)
김해(金海) 금해(金海)

그러나 상기 방법으로 각 이형에 해당하는 문자열을 하나 하나 생성한다는 것은 노동집약적인 일이며, 인간의 많은 지적 노력을 필요로 한다. 따라서 매핑파일이 방대하게 늘어나고 사용될 확률이 매우 낮은 문자열이 포함되더라도 일괄적으로 생성하는 것이 보다 효율적일 것이다. 다만, 한글맞춤법(1988년 1월 19일 문교부고시) "제5절 두음법칙"의 다음 규칙을 기본 매핑파일 작성시스템과 한글색인어 생성시스템에서 동일하게 참조한다면 생성되는 기본 매핑파일의 항목수를 대폭적으로 줄일 수 있을 것이다.

1) 한자음 "라, 려, 례, 료, 류, 리"가 단어의 첫머리에 올 적에는 두음법칙에 따라 "야, 여, 예, 요, 유, 이"로 적는다. 다만, 의존명사 "리(里), 리(理)"는 본음대로 적는다. 또한 "역이용(逆利用), 열역학(熱力學)"과 같이 접두사처럼 쓰이는 한자나 합성어의 뒷말도 두음법칙에 따라 적는다.

2) 한자음 "라, 래, 로, 뢰, 루, 르"가 단어의 첫머리에 올 적에는 두음법칙에 따라 "나, 내, 노, 뇌, 누, 느"로 적는다. "중노동(重勞動), 비논리적(非論理的)"과 같이 접두사처럼 쓰이는 한자의 뒷말도 두음법칙에 따라 적는다.

이형의 자동생성의 방법은 다음과 같이 1문자, 표준코드에 없는 문자, 2문자 이상의 세 가지가 경우가 생기게 된다.

1) 하나는 한 용어내에 多音字가 1문자만 포함되어 있을 때이다. "여성 女性"이란 용어의 예를 들어 보자. 多音字파일에

女 녀 503 (E443)
女 여 503 (EDFC)

가 있을 때, "여성 女性"이라는 용어가 입력되면 漢子부분만 한 문자씩을 잘라 多音字파일을 탐색한다. 즉 "女"

(여)와 "性"(성)을 차례로 탐색하게 된다. 우선 "女"(여)가 多音字파일의 "女 여 503"과 매칭되면 "여성 女性"을 같은 그룹에 있는 모든 문자로 각각 대체시킨다. 따라서 다음의 두 용어쌍이 생성된다.

녀성 女性 → 여성 女性
여성 女性 → 여성 女性

입력된 원래의 한자음은 "여, 녀" 어느 것이라도 좋다. "性"(성)은 매칭되는 문자가 없으므로 생성된 용어의 이형은 두 가지가 된다.

2) 표준코드에 해당 음이 없을 때이다. 多音字파일 등록시 어느 음에 해당하는 한자를 사용했는지 알 수 없는 때에도 동일하게 처리한다. 예를 들면 다음과 같다.

刺 자 459	(EDA9)
刺 척 459	(F4A7)
刺 라 459 1	(EDA9 혹은 F4A7)

상기 예에서 "刺"(라)는 표준코드에 없는 음이므로 "刺"(라)의 多音字파일 등록시 "刺"(자)의 음에 해당하는 한자를 따랐는지 "刺"(척)의 음에 해당하는 한자를 따랐는지 알 수가 없다. 그러나 어느 코드를 사용했는지에 관계없이 같은 문자군(459)의 모든 한자를 대체하면 된다. "水刺(수라): 임금에게 올리는 밥"의 예를 들면 다음과 같다.

수라 水刺 → 수라 水刺
수자 水刺 → 수라 水刺
수척 水刺 → 수라 水刺

3) 한 용어내에 多音字가 2문자 이상이 포함되어 있을 때이다. "樂山樂水"(요산요수)의 예를 들면, "樂"은 3.2 예1)과 같이 "낙, 락, 악, 요"의 4가지 음을 가지므로 첫 번째의 "樂"(요)에 의해 다음의 4가지 용어를 생성한다.

낙산요수 樂山樂水　락산요수 樂山樂水
악산요수 樂山樂水　요산요수 樂山樂水

두 번째 이후의 문자부터는 그 그룹에서 만들어진 용어 전체를 대상으로 대체한다. 3번째 "樂"(요)를 앞에서 만든 네 가지 용어에 대하여 적용시키게 되며 결과는 다음과 같이 16가지가 된다. 매핑파일을 만들 때 바른 용어는 제외시킬 수도 있고 포함시킬 수도 있으나 일괄적으로 포함시키는 쪽의 처리가 간단할 것이다.

낙산낙수 樂山樂水　락산낙수 樂山樂水
낙산락수 樂山樂水　락산락수 樂山樂水
낙산악수 樂山樂水　락산악수 樂山樂水
낙산요수 樂山樂水　락산요수 樂山樂水

악산낙수 樂山樂水　요산낙수 樂山樂水
악산락수 樂山樂水　요산락수 樂山樂水
악산악수 樂山樂水　요산악수 樂山樂水
악산요수 樂山樂水　요산요수 樂山樂水

하나의 용어내에 세 가지의 多音字가 있는 경우에도 동일한 방법으로 처리한다. 즉, 두 번째 처리완료 데이터가 다음 세 번째 처리의 대상정보가 된다. 표준코드에 없는 음인 경우에도 多音字파일에 漢子코드가 있으면 매칭된 음을 취하고 같은 문자군에 있는 다른 음과 漢子로도 대체시킨다. 이와 같은 과정으로 생성된 모든 이형에 대하여 바른 한자·한글쌍을 대응시켜 매핑파일을 만든다. 전자국어사전을 이용하여 모든 기입어의 한자를 모두 多音字파일과 비교함으로써 多音字가 포함되어 있는 용어를 찾아낼 수 있다.

본 연구에서는 공개된 전자국어사전이 없어서 이화여자대학교 중앙도서관이 사용하고 있는 한자음절변환사전(항목수 약 5만어)을 중심으로 하고 大漢韓辭典(1985)의 용례로 보완하여 실험하였다. 동 대학의 한자음절변환사전은 공개된 한자용어사전과 도서관에 구축된 동양서 데이터베이스의 서명류에서 추출한 한자용어를 포함하고 있다.

만들어진 한자·한글쌍이 각각을 구별할 수 있는 하나의 유일한 코드가 된다. 매핑파일의 예를 들면 다음 표 1과 같다. 매핑파일은 매뉴얼로 보완된 데이터를 포함하고 있다.

표 1 한자·한글쌍의 매핑파일의 예

바른표기(한자)	이형표기(한자)		대체표기(한자)
곳간 庫間	고간 AB	→	곳간 AB
구절 句節	구절 AB	→	구절 AB
	귀절 AB	→	구절 AB
나인 內人	내인 AB	→	나인 AB
쇄도 殺到	살도 A_1B	→	쇄도 AB
스페인 西班牙	서반아 ABC	→	스페인 ABC
시월 十月	십월 AB	→	시월 AB
오뉴월 五六月	오육월 AB_1C	→	오뉴월 ABC
	오륙월 ABC	→	오뉴월 ABC
요산 樂山	락산 A_1B	→	요산 AB
	낙산 A_2B	→	요산 AB
	악산 A_3B	→	요산 AB
	요산 AB	→	요산 AB
전깃불 電氣불	전기불 ABC	→	전깃불 ABC
초파일 初八日	초팔일 ABC	→	초파일 ABC

생성된 이형이 →의 좌변, 바른 표기가 우변이 된다. 바른표기를 매핑파일에 기술하는 것은 낭비이지만 본 연구의 시험에서는 처리의 단순화를 위하여 바른표기도 "악산요수 樂山樂水 → 요산요수 樂山樂水"와 같이 이형표기와 대체표기를 반복하여 포함시켰다. 표 1에서 이형표기와 대체표기 중 "(한자)"부분의 영문자는 그 위치에 있는 "바른 표기"의 한자임을, 첨자가 붙은 영문자는 코드가 다른 한자임을 의미한다.

대체표기에 한자를 부기하고 있는 것은 색인어파일에서의 동음이의어를 구별하여 처리하는 시스템에서는 매우 유용하기 때문이다. 검색어로 "조류"만을 입력한 이용자에게 다음과 같이 동음이의어를 디스플레이하여 원하는 주제를 선택하게 할 수 있다. 사용된 한자의 코드는 무엇이든 관계가 없다. 시스템의 이용자에게는 폰트가 필요할 뿐 코드값은 관계없기 때문이다.

예) 조류 (鳥類) "새"
조류 (潮流) "물"
조류 (藻類) "식물"

3.4 기본 매핑파일의 보완

전자국어사전을 이용하여 기본매핑파일을 만들어도 모든 용어가 망라될 수는 없다. 전술한 색인어 통제정책에 따른 관련 용어의 조사를 제외하더라도 추가로 여러 가지 전문용어사전을 이용하여 용어를 수집할 필요가 있다. 이 때 가장 효용성이 있는 것은 한자가 부기된 전문용어사전과 시소러스가 될 것이다. 전문용어사전이나 시소러스에 바른 한자코드가 입력되어 있을 필요는 없다. 최소한 바른 한글음이 한자와 함께 기술되어 있으면 된다.

그러나 漢子의 다른 음을 사용한 경우를 제외하고 상기 규칙을 그대로 적용해버리면 실제 상황에서는 사용하지 않는 용어가 다수 만들어지는 일이 생기게 된다. 또한 일정한 규칙에 의해 이형이 생기는 것이 아니라 의미에 따라 달라지는 음들이 있어서 맞지 않는 한글색인어가 생성될 수 있다.

이와 같이 용어의 수를 줄이고 불규칙적인 용어를 정확하게 처리하기 위하여 다른 음을 사용하는 경우를 제외하고는 예외에 해당하므로 별도로 조사하여 보완하여야 한다. 보완해야 하는 경우를 보면 다음과 같다.

첫째, 우리말은 대부분의 각 문자가 독립적인 개념을 갖고 있어서 한 문자로 하나의 단어가 되는 경우가 많으므로 상기 多音字파일의 모든 문자를 매핑파일에 포함시키는 것이 바람직할 것이다.

둘째, 多音字파일에 등록되지 않는 문자를 포함하는 용어의 보완이다. 예를 들면 "內, 寺"의 音은 "내, 사" 밖에 없으므로 특별한 용례에서의 音인 "나, 치"는 多音字파일에 등록되지 않으므로 기본매핑파일에 별도로 입력하여 보완한다. 이 때 대체표기에서 해당 음의 한자가 없고, 빌려와야 할 문자후보가 두 종류 이상일 때는 어느 것을 사용해도 좋다. 예를 들면, 표 1에서 "오륙월, 오육월"의 대체표기인 "오뉴월(五六月)"에서 "六"은 "륙, 육"의 어느 음을 따라도 좋다.

세째, 전술한 색인어 통제정책에 따라 대응용어를 다른 한글표기로 바꾸어 사용하는 시스템에서는 이와 같은 용어를 별도로 조사하여 보완하여야 할 것이다. 이것은 한자용어 뿐만 아니라 한자가 없는 용어도 포함되는 사항으로 명확한 표기 및 선정원칙이 있어야 할 것이다.

4 바른 한글색인어의 생성

4.1 자동생성

모든 경우의 수를 망라하는 시험을 위해서는 한자가 포함되어 있는 대단위 코퍼스가 필요하나 이와 같은 코퍼스를 구축한다는 것은 매우 어려운 일이다. 본 연구에서 간단한 생성의 시험은 문헌 "尹龜鎬. 情報檢索效率에 관한 硏究. 圖書館學, 8:73-102(1981)"를 대상으로 하였다. 이 문헌을 사용한 것은 내용이 정보검색에 관한 것으로 입력자의 전공분야와 일치하여 에러가 적을 것이라는 생각때문이었다. 입력은 문헌정보학과 3학년 5명이 분담하였다. 원문 그대로 입력하여 띄어쓰기와 괄호를 기준으로 한자부분만을 추출하고 이 한자부분을 한글로 자동생성한 후 결과를 검토하였다. 입력시 목차는 포함하였으며, 저자명, 저자의 소속, 각주, 그림, 표내의 용어 및 수식 등은 제외하였다. 입력된 문헌에서 추출한 대상 한자용어수는 2,269개였다.

시스템은 우선 처리대상이 되는 한자용어의 한자를 이용하여 한글·한자쌍을 생성한다(표 1의 대체표기에 해당한다). 생성된 한글·한자쌍은 그 용어를 식별할 수 있는

유일한 코드라고 가정하며 다음 경우 중의 하나가 된다.

첫째, 이형표기 문자가 없는 경우
둘째, 한글, 한자가 다 맞는 경우.
세째, 한글은 틀리나 한자가 맞는 경우
네째, 한글, 한자가 다 틀린 경우

이형표기 문자가 없는 용어는 매핑파일에서 매칭되지 않아 옳은 표기로 간주되므로 그대로 생성하게 된다. 나머지 세 가지는 매핑파일과 비교하여 매칭이 되면 대응되는 바른 표기의 코드쌍(표 1의 대체표기)으로 간단히 대체하게 된다.

전혀 다른 용어로 입력한 것(예를 들면, 凡函數, 測程, 不適合 등)과 문헌 자체에서 틀린 것을 제외하고, 입력측면을 보면, 1) 워드프로세서의 음절단위 한자변환기능이 있어서 전공분야에서 알고 있는 단어의 입력이 잘못되는 경우는 드물었다. 2) 가장 많이 틀리는 것은 워드프로세서에 등록되지 않은 복합명사의 입력이었다. 예를 들면, "再現率"을 "재현율, 재현률"로 변환하면 등록되어 있지 않으므로 "率"만을 선택하게 된다. 또한 網羅性(망라성→망나성)과 같이 "~性"이 붙는 경우도 동일하였다. 3) 그 외에 양쪽이 다 있는 論難(논란→논난) 등이 있었다. 워드프로세서의 음절단위 입력기능이 있고 동일 전공분야의 대학 3년생이 입력한 내용을 보아 워드프로세서의 음절단위 입력기능이 약하거나 타 전공분야이면 다르게 입력할 확률이 매우 높아질 것이라 생각된다.

생성측면을 보면, 입력된 한자가 잘못된 것을 제외하고 생성에서의 에러는 없었다. 그러나 입력측면에서 논한 바와 같이 근본적으로 입력에러가 적었기 때문으로 생각되며 결국은 매핑파일의 완전성 여부가 생성결과의 질을 좌우하게 될 것이다.

4.2 수동생성

본 연구의 실험에서 수동생성이 필요한 용어는 없었다. 그러나 여러 가지 경우의 문헌이 포함되면 매핑테이블을 보완하는 것 만으로 모두 옳게 생성할 수 있는 것은 아니다. 다음과 같이 의미에 따라 다르게 사용하는 경우에는 문제가 있다. 예1)의 "금슬(琴瑟)"이란 용어는, 악기를 나타낼 때는 "금슬(琴瑟)"이 맞으며, 부부의 사랑을 나타낼 때는 "금실(琴瑟)"이 맞다. 어느 쪽 한글을 생성할 것인지를 결정해야 한다.

예1) 금슬(琴瑟) → 금슬(琴瑟) "악기"
금슬(琴瑟) → 금실(琴瑟) "부부의 사랑"

차간(車間) → 차간(車間) "차간거리"
차간(車間) → 찻간(車間) "타는 곳"

이 부분에 대해서는 매핑파일의 탐색시점에서 색인자와 탐색자에게 필요정보를 제공하여 선택하게 하거나 자연언어처리시스템에서 문맥을 파악하여 내용을 유추하는 길이 있을 것이다. 본 연구의 실험에서는 상기 내용을 표시하여 선택하도록 하였다.

5 결론

漢字표기용어로부터 바른 한글색인어의 자동생성방법에 대하여 논하였다. 바른 한글음의 자동생성을 위해서는 多音字파일의 구축과 이형이 가능한 용어를 조사하는 일이 중요하며 어려운 일이다. 古書의 전산화가 진행되면 한자코드의 확장문제를 포함하여 全文이 아니라 서명만으로도 한글색인의 문제는 보다 심각해질 것이다. 조사된 용어는 근본적으로 망라적일 수 없으며, 실제 시스템에서 사용하면서 보완되어야 할 것이다. 또한 한자코드가 확장되거나 처리대상 분야가 확대되면 부수적으로 관련 多音字와 이형이 가능한 용어의 조사가 실시되어야 할 것이다.

또한 본고에서는 시험대상에서 제외하고 간단히 지적하였지만 현재 중국서와 일본서에 대한 한글색인어 생성, 한걸음 더 나아가 한자문화권의 약자까지를 모두 포함하고 있는 유니코드체제하에서의 한글색인어 생성에 관한 것도 간과해서는 안될 문제일 것이다.

참고문헌

미승우(1993). 새맞춤법과 교정의 실제. 증보판. 서울: 어문각.

이춘택(1991). 韓日 國家規格漢字코드의 統合硏究. 중앙대학교 대학원 도서관학과 자료조직전공 박사학위논문, 미간행