한국불교전서 전산화에 따른 누락문자처리와 활용방안

홍 영 식*, 이 금 석*, 이 용 규*, 한 보 광**

* 동국대학교 컴퓨터·멀티미디어공학과 교수
** 동국대학교 선학과 교수

목 차

1. 서론

한국불교전서를 비롯한 한적 위주의 고문헌 디지털화에서 한자표기를 위한 코드와 폰트는 필수적인 사항이다. 특히, 코드관련 문제를 해결하기 위해서 한국, 대만, 중국 및 일본에서 국가별로 독자적인 노력을 해오다가 한국, 대만 및 일본이 중심이 되어 CJK 코드가 정

해졌고, 이것이 현재 표준코드로 정해진 유니코드의 기반이 되었다.

국내의 경우 해인사 대장경연구소에서 추진한 대장경 디지털화를 위한 4바이트 코드체계가 시도되었고, 동국대학교 전자불전연구소에서는 한국불교전서 전산화를 위해서 2바이트 유니코드를 채택했다[4,5,6,10]. 2000년 2월에 발표된 유니코드 3.0의 경우 27,786개의 한자용 코드가 할당되었지만, 현재 국내 폰트제작 업체에서 공급하는 폰트 수는 유니코드 2.1에 기반한 21,204 개의 코드에 대한 한자 폰트가 제공되고 있는 실정이다[1].

이와 같은 배경에서 한국불교전서를 위시한 고문헌의 디지털화에 필요한 한자 폰트의 수가 절대적으로 부족하여 코드가 할당되지 않은 누락문자에 대한 처리방안이 필요하게 되었고, 전자불전연구소는 동국대학교 100주년 기념사업의 일환으로 1999년부터 시작된 한국불교전서 전산화 사업에서 인터넷 환경을 고려해서 누락문자에 대한 이미지 파일을 생성하고 HTML로 표현된 웹 문서에 포함시켜서 누락문자 문제를 해결했다[2,7,9].

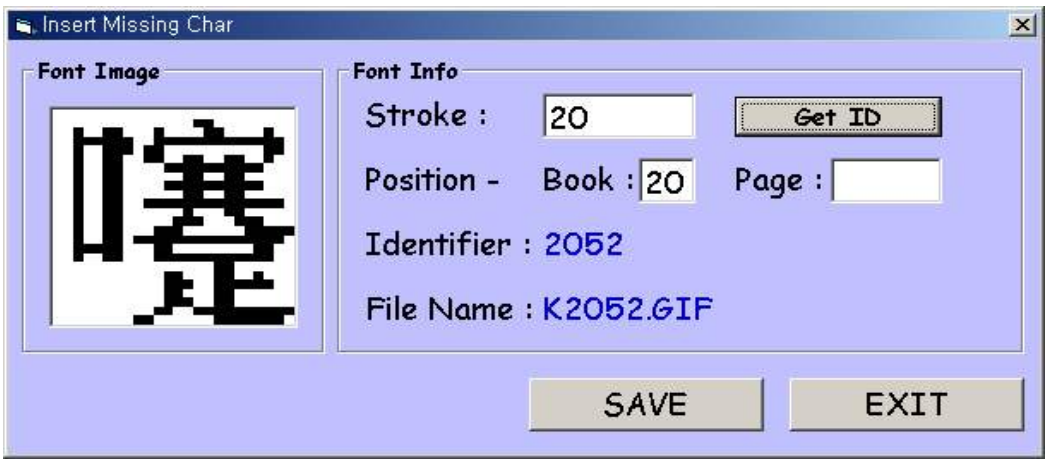
그러나, 누락문자 문제를 해결하는 여러 가지 시도가 있을 수 있지만, 한국불교전서를 위시한 고문헌의 디지털화 작업은 계속되고 있기 때문에 누락문자 처리 방안에 대한 계속적인 연구와 현재까지 만들어진 누락문자 이미지들의 활용방안의 마련이 필요하다 할 것이다.

2. 누락문자 처리절차 및 통계

한국불교전서 문서의 데이터베이스 구축과정에서 일차적으로 초기입력이 필요하며, 초기입력은 워드프로세서로 직접 입력하거나, 혹은 스캐너를 사용해서 입력파일을 생성한다. 그 다음에 이것을 원문과 대조하여 누락문자로 판단된 글자에 대해서 누락문자 데이터베이스를 검색하고, 누락문자 데이터베이스에 아직 존재하지 않는 글자의 경우 이미지 폰트파일을 생성하여 누락문자 데이터베이스에 등록한

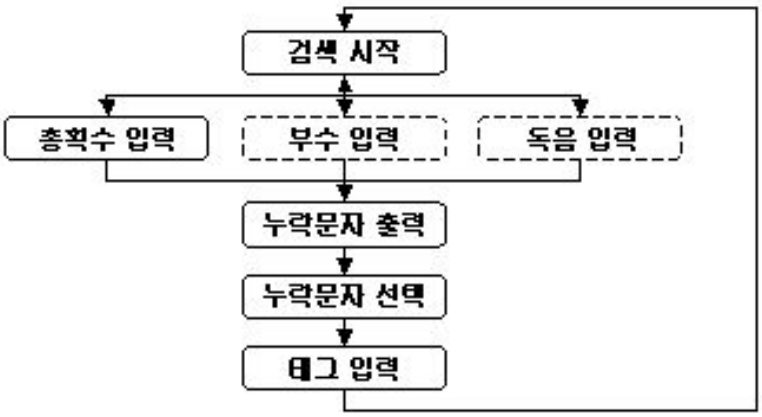
다. [그림 1]은 누락문자를 데이터베이스에 등록하는 절차를 도시한 것이고, [그림 2]는 이미지 폰트를 생성하여 누락문자 데이터베이스에 입력하는 예를 보인 것이다[2,7].

[그림 1] 누락문자 데이터베이스 등록절차

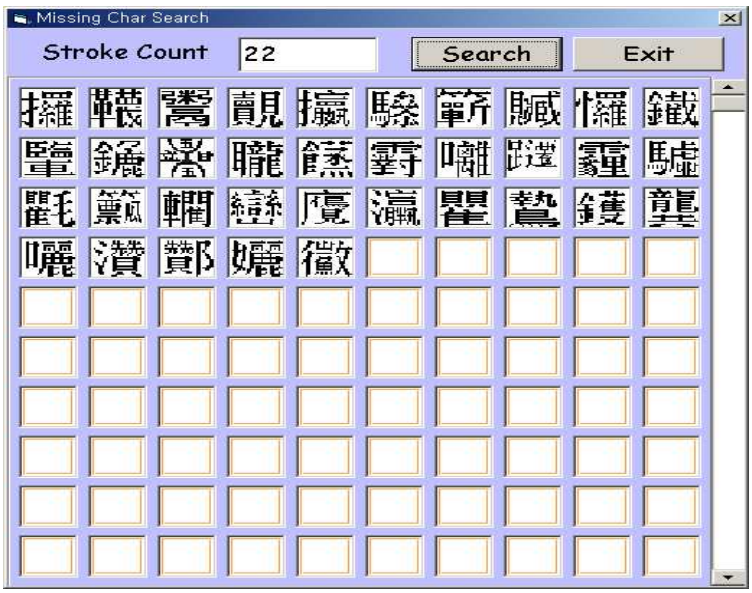


[그림 2] 누락문자 입력 예

그리고, 누락문자 데이터베이스에 이미 등록되어 있는 누락문자인 경우에는 데이터베이스를 검색한다. 누락문자 검색 창에서 총획수를 입력하고, 출력된 문자들 가운데서 해당문자를 선택한다. [그림 3]은 누락문자 검색절차를 도시한 것이고, [그림 4]는 누락문자의 총 획수 22로 검색된 누락문자들을 나타낸 것이다.



[그림 3] 누락문자 검색 절차



[그림 4] 누락문자 검색 예

한국불교전서를 전산화하면서 지금까지 데이터베이스에 등록된 누락문자의 총수는 이체자를 포함하여 2,313자로서 획수별 통계는 [표 1]과 같다.

[표 1] 획수별 누락문자 통계(2003.10 현재)

획수	개수	획수	개수
1	17	18	106
2	29	19	94
3	22	20	51
4	27	21	48
5	50	22	35
6	52	23	35
7	96	24	20
8	117	25	3
9	130	26	9
10	153	27	4
11	202	28	3
12	196	29	0
13	170	30	1
14	181	31	1
15	156	32	0
16	174	33	1
17	130	합계	2,313

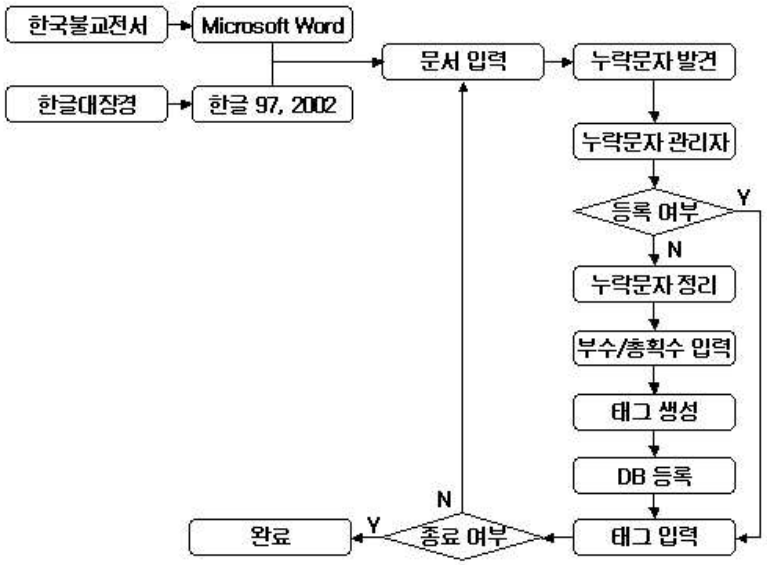
지금까지 발견된 누락문자 2,313자 가운데서 822자는 유니코드에 포함된 글자로 판명되었으나 입력시점에 폰트가 존재하지 않았던 글자에 해당한다. [표 2]는 누락문자의 연도별 발견 빈도수를 나타낸 것으로서 2002년에 1,042자를 발견하고, 2003년에 930자를 발견했으나 입력문서의 내용에 따라 영향이 있는 것으로 판단되며 통계적인 의미는 별로 없는 것으로 판단된다.

[표 2] 연도별 누락문자 발견 빈도수

년도	개수
2000	124
2001	217
2002	1042
2003	930
합계	2313

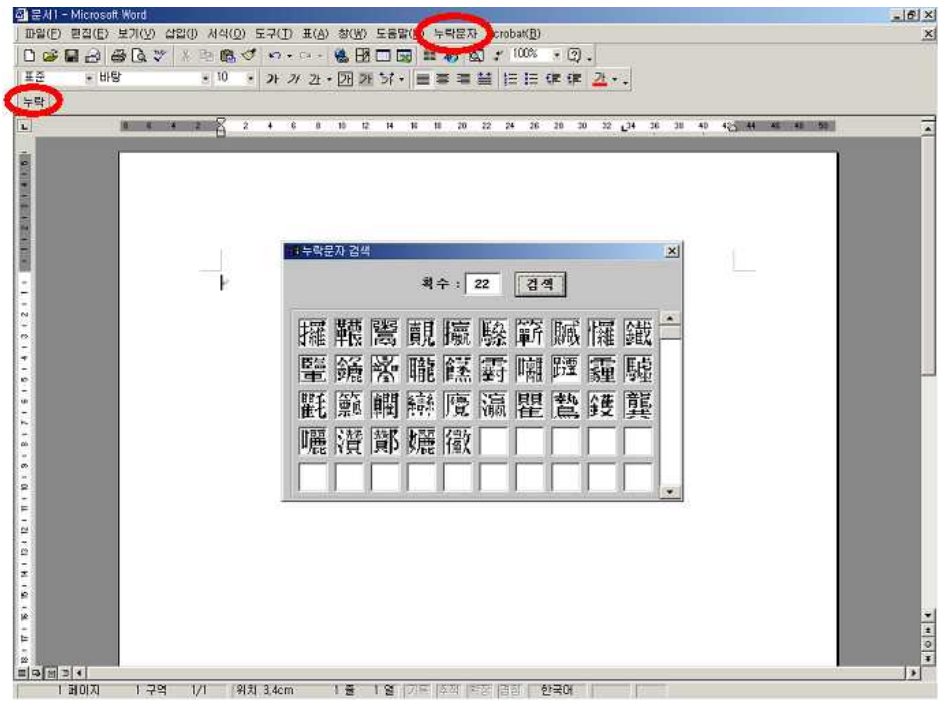
3. 타 시스템과의 연관성

2001년부터 문화관광부의 지원으로 역경원과 전자불전연구소가 진행하고 있는 한글대장경 전산화 사업과정에서도 누락문자에 대한 처리가 필요하게 되었다[8]. 한국불교전서 전산화 사업 과정에서 생성된 누락문자 데이터베이스가 한글대장경 전산화 사업에서 활용됨은 물론 한글대장경 전산화 사업에서 추가적으로 발생하는 누락문자는 기존의 누락문자 데이터베이스에 등록되어 활용된다. [그림 5]는 한국불교전서 전산화 사업과 한글대장경 전산화 사업에 연관된 누락문자 처리시스템을 도시한 것이다.



[그림 5] 누락문자처리와 타 시스템과의 관계

한국불교전서 전산화 사업을 시작한 1999년 당시에는 유니코드에 기반한 한자폰트가 한글워드프로세서에서 지원되지 않아서 전자불전연구소에서 유니코드 기반 한문 입력 및 편집기를 자체 개발하여 입력작업을 진행했다[3]. 현재는 유니코드용 확장 한자 폰트가 설치된 마이크로소프트 MS워드나 한글과컴퓨터(주)의 한글2002와 같은 한글워드프로세서에서 유니코드 폰트가 지원되고 있다. 전자불전연구소에서는 이러한 워드프로세서를 누락문자처리기와 연계시켜서 문서편집이 가능하게 했다. [그림 6]은 누락문자처리 기능이 한글워드프로세서인 MS워드에 아이콘으로 추가된 것을 보여준다.



[그림 6] 누락문자처리 기능이 추가된 한글워드프로세서

4. 누락문자 처리 시스템 활용과 개선방향

앞에서 기술한 바와 같이 누락문자 처리시스템은 한국불교전서 전

산화와 한글대장경 전산화에서 모두 사용되고 있으며, 이것은 인터넷 검색을 기반으로 한 여타 고문헌 전산화에서도 활용될 수 있음을 입증하고 있다. 또한, 누락문자 처리에 소요되는 시간이 고문헌 입력 및 편집시간에 미치는 영향이 크기 때문에 보다 효율적인 누락문자 처리시스템으로 개선할 필요가 있다.

누락문자 처리시스템의 개선에 고려할 사항은 다음과 같다.

1) 사용자 인터페이스의 개선
2) 다양한 누락문자 검색방법 제공
3) 검색속도 향상
4) 다양한 크기의 폰트 제공

현재 누락문자 처리 시스템이 마이크로소프트 MS워드와 한글97 및 한글2002와 연계해서 사용하는 것이 가능해졌지만 일반 사용자들도 사용하기 쉽게 사용자 인터페이스를 개선할 필요가 있다. 누락문자 검색방법도 총 획수에 의한 검색 외에 부수에 의한 검색이나 독음에 의한 검색방법의 지원이 필요하다. 다양한 검색이 가능하려면 기존의 누락문자 데이터베이스의 자료구조를 고쳐야 한다. 또한, 누락문자 데이터베이스의 크기가 증가함에 따라 문자검색 속도가 느려질 것을 대비해서 보다 빠른 검색이 가능한 검색알고리즘과 자료구조의 변경이 필요하다. 그리고, 다양한 크기의 이미지 폰트를 생성하여 저장함으로서 누락문자 처리시스템의 효율성을 제고할 수 있을 것이다.

5. 결론

한국불교전서의 전산화 과정에서 구축된 누락문자처리 시스템은 기존의 상용 한글 워드프로세서와 연계해서 사용할 수 있기 때문에

인터넷 검색을 전제로 한 여타 고문헌 전산화에 활용될 수 있음을 살펴보았다. 또한, 현재까지 출판된 한국불교전서 총 10책 가운데 8책에 대한 본문 데이터베이스가 구축된 현 시점에서 지금까지 발견된 2,313자의 누락문자가 저장된 누락문자데이터베이스에 앞으로 발견될 누락문자를 계속 추가함으로써 향후의 고문헌 디지털화에 활용할 수 있다.

인터넷 사용이 보편화된 현 시점에서 한국불교전서와 같은 고문헌의 전산화는 인터넷 검색을 전제로 해야 하고, 인터넷 검색은 한자 표현에 유니코드의 채택이 불가피하다. 따라서, 한국불교전서의 전산화나 한글대장경 전산화에서 누락문자의 처리는 필수적이며, 기존의 누락문자처리 시스템을 보다 효과적인 누락문자처리 시스템으로 발전시킴으로써 향후의 고문헌 전산화에 크게 기여할 것으로 판단된다.

참고 문헌

[1] Richard Cook , “UniHan Variation: Issues and Solutions”, 23rd Inernationailzation & Unicode Conference, 2002.

[2] Y.S.Hong, et al. “Searching Missing Characters from the Hanguk Pulgyo Chonso Database”, 2001 EBTI International Conference, 2001.

[3] Y.S.Hong, et al. “Development of a Syntax-directed SGML Editor for Processing Korean Ancient Documents”, Prodeedings of 1999 EBTI, ECAI, SEER & PNC Joing Meeting in Taipei, 1999.

[4] In Sub Hur, “Report on the Digital Tripitaka Korean 2001”, 2001 EBTI International Conference, 2001.

[5] 김재성, “고려대장경 전산화 현황”, 동국대학교 전자불전연구소 제4회 세미나, 2002.

[6] 윤용석, “고려대장경의 인터넷 검색 및 열람”, 동국대학교 전자불전연구소 설립기념 세미나, 1999.

[7] 이금석, 이용규, 홍영식, 한태식, “한국불교전서 전산화를 위한 누락문자 처리방안”, 동국대학교 전자불전연구소 제2회 세미나, 2000.
[8] 이금석, 이용규, 홍영식, 한태식, “한글대장경 검색시스템”, 동국대학교 전자불전연구소 제4회 세미나, 2002.
[9] 이용규, 이금석, 홍영식, 한보광, “한국불교전서 전산화”, 동국대학교 전자불전연구소 설립기념 세미나, 1999.
[10] 허인섭, “전산화본 고려대장경 2000 완성의 학술적 의미와 미래전망”, 동국대학교 전자불전연구소 제2회 세미나, 2000.

키워드(Keyword)

유니코드, 누락문자처리시스템, 한국불교전서, 한글대장경, 웹 기반 검색시스템
Unicode, Missing Character Handling System, Hanguk Pulgyo Chonso, Tripitaka Koreana, Web-based Retrieval System