<한국 漢字字形情報 통합데이터셋 구축을 위한 현황분석과 과제> 토론문

조성덕(단국대 한문교육연구소)

발표자는 우리나라의 '한자 자형 정보 통합데이터셋'을 구축하기 위하여 한국과 일본의 사례를 들어 설명하였습니다. 발표의 내용이 대부분 현황을 소개하고 문제점을 지적하한 것으로 보여서 별도의 질의는 하지 않겠습니다. 다만 우리나라의 한자 자형 정리의 문제점에 공감하면서 제가 알고 있는 관련 내용과 검색방식에 대해 간략하게 제안하는 것으로 질문을 대신할까 합니다.

우리나라의 한자 자형 정리는 한자를 모국어로 사용하는 중국과 한자를 모국어처럼 사용하는 일본에 비해 다소 늦은 감이 있습니다. 1993년 국립국어원에서 진행한 『漢字 略體 調査研究』를 시작으로 본다면 약 30년의 시간이 흘렀습니다. 그럼에도 불구하고 아직까지 국가 주도하에 명확하게 정리된 한자의 자형과 자의에 대한 표준이 없다는 것이 매우 안타까운 일입니다.

현재 KISA(한국인터넷진흥원)에서 ICANN(국제인터넷주소관리기구)의 주관 아래 세계에서 사용하는 언어의 문자셋을 인터넷 최상위도메인에 포함시키는 일을 진행하고 있습니다. 이 중 우리나라의 한글을 포함하여, 한자화권의 공용어였던 漢字가 포함되어 있습니다. 현재 韓中(대만、홍콩、마카오)日이 주축이 되어 최상위도메인에 사용할 '대표한자제정'을 진행하고 있는데, 三國이 이체자의 자형을 정의하는 과정에서 의견 충돌로 난행을 격었습니다. 이때 ICANN 농담반 진담반으로 "한국은 한자를 사용하지도 않는데 왜 이렇게 漢字에 목숨을 쓰는지 모르겠다."라고 했던 기억이 있습니다.

선생님이 언급한 대로 <유니코드한자검색시스템>에 수록된 신출자와 <이체자 정보>에 정리된 한자는 매우 미흡한 상태입니다. 그러나 단기간에 많은 양의 데이터를 구축해야 하는 기관과 업체로서는 세세한 부분까지 손쓸 여력이 없었을 것입니다. 그런 이유로 다양한 유형에서 문제가 드러나는 것이 오히려 당연하게 느껴집니다.

1. **자형 중복**: 발표문 [표1]에서 언급하신 <유니코드한자검색시스템>에

신출한자로 등록된 자형이 EXT_B에 수록되어 있는 경우와 기존 KC 코드와 중복되는 경우도 상당합니다. 많게는 동일 자형이 3개 이상 중복된 경우도 상당수 존재합니다.

보기)

- 艐 일기: KC10000 역통: U+27901
- 灤 KS00429: 코드 없음. 역통: KC06433

2. 폰트 제작: 사이트에 있는 폰트와 실제 원문과 자형이 서로 다른 경우도 많습니다. 이것 역시 우리가 풀어야 할 숙제라고 생각합니다. 모든 한자를 원문에 있는 자형 그대로 만들 수 없는 문제가 있습니다.

보기)

- 폰트  KS01052 원문이미지 

3. 자료 보완: 발표에서 언급된 관련 사이트의 수정 및 보완에 대한 문제입니다. <유니코드한자검색시스템>의 경우 자료는 업데이트 되지만, 기존의 정보가 수정되지는 않은 것으로 보입니다. <이체자 정보>는 2006년 사이트를 오픈한 후 15년이 지난 지금까지 업데이트 한 번도 업그레이드가 않았다는 것입니다.

- 고대 민족문화연구원의 <유니코드 한자검색기>와 2013년 『유니코드 한자정보사전』의 경우에도 가장 최근에 정리된 사전임에도 불구하고 여전히 대부분의 한자에 대한 독음을 수록하고 있지 않은 것이 현실입니다.

 # <한국 漢字 字形情報 통합 데이터셋 구축> 검색방법 제안

- 기존의 部件 및 構件 검색방식능 활용하면서 한국에 맞는 검색법을 항목으로 설정하면 어려운 한자를 좀더 효율적으로 검색할 수 있을 것입니다.
- 한자입력기 PM(팔만에디터)은 기존에 일부 팔만대장경과 역사통보통합시스템 등 한적입력에서 사용되었지만 한자를 모르는 사람이 입력하기 어려운 문제가 있어 일반적으로 사용되지는 못했습니다. 그러나 이 입력기의 장점은 214부수정도의 한자를 학습한 사람들이라면 한자의 구성요소를 한글 음가로 입력할 수 있는 장점이 있습니다. 우리나라가 한글을 만들었다는 감안하면 매우 합리적인 입력방법이라고 생각합니다.