국어 정보화의 방향 요약본

권성우 (2023-82030)

02/10/2023

본 논문은 국어 정보화, 특히 소리와 문자 중 후자의 정보화의 역사와 당면 과제에 대해 다루고있다. 정보화란, 어떠한 자료를 컴퓨터에서 효과적으로 구현할 수 있는 방안을 모색하는 것이고, 일반적으로 입력, 내부처리, 출력의 세 단계로 대표 될 수 있으며 그 대상은 소리나 이미지, 나아가서 데이터 형태로 존재할 수 있는 모든 것이다. 이것을 문자언어에 대입해 보면, 키보드와 같은 장치를 통해 기계에 문자 정보를 입력하고, 내부에서는 문자를 정해진 문자-2진수(컴퓨터가 이해할 수 있도록) 매핑을 통해 2진수로 변환 후 메모리(하드디스크)에 텍스트 파일의 형태로 저장한다. 출력을 해야할 때는 반대로 저장된 정보를 다시 문자로 매핑하고 사용자가 원하는 출력장치에 시각적 형태로 재현한다. 이러한 과정에서 볼 수 있듯, 내부처리를 할 때 사용되는 '정해진 규약', 즉 문자코드가 미리 엄격하게 정의 되어있는 것이 매우 중요하다.

대표적인 문자코드로는 미국에서 발명된 아스키(ASCII)가 있다. 아스키는 대문자, 소문자, 숫자등 표준영어에서 사용되는 문자들을 1바이트(8비트, 2의 8승으로 최대 256개의 기호)로 나타낼 수 있게해준다. 아스키는 전 세계, 특히 로마자를 사용하는 나라에서는 표준처럼 사용되는 경우가 빈번하다. 하지만 로마자를 사용하지 않는 언어나 1바이트로는 부족해 2바이트를 사용해야하는 언어(ex.한글)의 경우, 경제성과 효율성의 충돌하는 일이 자주 일어나 문자 코드의 표준을 정하는 일이 단순하지 않다.

한글의 음절형 글자 특성을 고려하면서 효율적인 문자코드를 만드는 방식은 많이 연구되어 온 주제이다. 경제적 측면에대한 고민이 부상한 후에는 크게 조합형 한글 코드와 완성형 한글 코드 두 가지를 예로 들 수 있는데, 조합형은 16비트(2바이트)중 첫번째 비트를 제외한 초성-중성-종성 코드에 각각 5비트씩 할당하여 조합, 결합하는 방식인 데 비해, 완성형은 한글로 구현할 수 있는 11,172자 중, 현대에서 사용되는 2,350자에 한자 4,888자를 추려서 만든 시스템이다. 경제적 요소를 제외하고도 비교적 단순하고 사용자 친화적인 조합형을 제치고 국가표준으로 채택된 것은 완성형이었고, 이는 몇 몇 문제점을 야기하였는데, 배제되었던 글자를 추가하는 과정에서 하위 호환성을 보존하려다 보니 문자 코드 값과 한글 자모 순서 사이에 배열의 문제가 발생하게 됨과 더불어 자소 분해에서도 불편함이 관찰되었다.

이렇듯 많은 국가들이 인코딩을 자체 개발하다 보니 각 국가간 사용되는 문자코드에서 코드값을 배정하는 방식이 천차만별로 달라져 버렸고, 전 세계적으로 통일된 문자 코드 체계를 만들려는 노력 끝에 유니코드라 불리는 인코딩 방식이 탄생하게 되었다. 유니코드는 세계의 모든 문자들을 표현할 수 있음과 더불어 체계 내에서 각 문자가 서로 겹치지 않는다. 하지만 경제성과 효율성의 충돌은 피해 갈 수 없는데 , UTF-32부터 16, 8까지 세 가지의 인코딩 각자 문자의 일관성/비용, 아스키와의 호환성/일관되지 않은 바이트 수 와 같은 장단점이 있다. 유니코드 내에서 한글 문자들은 사용빈도가 높은 글자들에 속하고, 조합식을 따르지는 않으나 자모순대로 배열되어 있어 간단한 계산만으로 초성, 중성, 종성을 알아내는 것이 가능하다. 각 자소도 따로 코드를 배정 받았다는 점이 눈 여겨 볼 만하다.

유니코드 전 존재했던 기법으로는 옛한글에서 사용했던 특수문자를 표상할 수 없었지만, 20세기말 워드프로세서를 서비스하던 기업들이 옛한글에 코드를 배당하여 표상한 것을 시작으로 21세기초에 유니코드 기반으로 제품들을 수정하는 과정에서 옛한글 처리방식을 통일하고, 글꼴까지 제작하였다. 하지만 이때 옛한글과 구결자를 임시방편적으로 국제적 통용성이 없는 사용자 정의 영역에 배당했고, 이는 후에 조합식으로 바뀌는 과정에서 저장공간과 직결되는 글자들의 바이트 수의 차이와 구/신버전 호환성 문제로 연결되었다.

역시 한글을 사용하는 북한에서는 국규9566라는 이름의 코드 표준을 제작해 한국의 완성형 코드처럼 문자를 추가해가며 사용해왔고, 유니코드가 전 세계적으로 통용된 후에는 두 가지를 모두 사용하고있다.

디지털 시대에서 국어 정보화가 어떤 방향으로 변화 해 왔는지는 앞으로 나아가야 할 방향에 대해 시사하는 바가 크다. 한글의 특성을 세밀하게 반영할 수 있어야 하고 미래지향적이되 이제까지 존재해온 방식과 호환이 가능해야 한다. 또한 어느 때보다 범세계적 정보통신이 활발한 오늘날, 국제표준과도 일맥상통하는 방식이 채택되는 것이 바람직 할 것으로 여겨진다.