북쪽 가나다 차례를 지원하는 로캘과 활용 프로그램 개발

정일동° 김경석
부산대학교 전자계산학과
{idjung°, gimgs}@asadal.pnu.edu

Development of Locale and Application Program Supporting North Korean Collating Sequence

Il-dong Jung° Kyongsok Kim
Dept. of Computer Science, Pusan National University

요 약

ISO 14651을 이용하면 간추리는 차례가 공통 틀 표 (Common Template Table) 에 들어 있기 때문에 글자의 간추리는 차례를 바꾸더라도 간추리기 프로그램을 바꿀 필요가 없다. 남쪽과 북쪽의 가나다 차례를 통일하지 않고 순서가 다른 문제를 해결할 수 있는 방법은 ISO 14651 을 활용해야 한다. 본 논문에서는 북쪽의 한글 가나다 차례를 남쪽에서 활용할 수 있도록 하기 위해서 북쪽의 한글 가나다 차례를 포함하는 북쪽 로캘을 만들고, 입력된 글자떼를 남쪽 혹은 북쪽의 한글 가나다 차례에 따라 간추리기 할 수 있는 프로그램을 개발하였다.

1. 서 론

UCS (=ISO/IEC 10646, =Unicode) 는 앞으로 점점 더 많이 쓰게 될 것이고, 얼마 지나지 않아서 남북도 모두 UCS 를 쓰게 될 것으로 보인다. 한국이 일찍 통일되지 않는다면 남북이 각각 UCS 를 쓰다가 통일이 되면서 계속하여 UCS 를 쓰게 될 것이다. 그러나 만일 통일이 이른 시일 안에 이루어진다면, 통일이 된 뒤에 UCS 를 널리 쓰게 될 것으로 보인다 [1].

남북쪽의 가나다 차례가 다르다는 것은 잘 알려진 사실이며, 이 때문에 북쪽은 ISO/IEC JTC1/SC2/WG2 회의에 참석하여 UCS 에 있는 한글 가나다 차례에 대하여 북쪽의 의견을 말하였다. 그러나 거의 모든 나라 대표들은 북쪽의 이 제안이 받아들일 수 없다고 느꼈으며, 따라서 WG2 에서 받아들여지지 않았다 [1].

WG2 에서는 한글 가나다 차례의 해결 방안으로 북쪽에 ISO/IEC 14651 을 활용할 것을 제안하였다. 국제 표준 ISO/IEC 14651:2000 (International String Ordering) 은 2000 년 4월에 최종 국제 표준안으로 발표되었으며, 여러 나라 글자계 (script) 가 섞여 있을 때, 모든 글자의 차례를 정하고 간추리는 틀에 관한 표준이다 [2].

ISO/IEC 14651 은 2000 년에 국제 표준으로 확정되었지만, 아직 KS 규격으로 되지 않았다. ISO/IEC 10646 에 있는 한글 글자와 글자마디의 차례는 남쪽 가나다 차례를 따랐다. 그래서 북쪽의 가나다 차례를 지원하려면 ISO/IEC 14651 을 활용해야 하며, ISO/IEC 14651 연구는 앞으로 정보 기술 분야의 남북 통일에 크게 아바지할 것이다 [1].

현재 리눅스에는 ISO/IEC 14651 을 지원하는 strcoll()/strxfrm() 함수가 있다. 이 함수는 로캘 (Locale) 을 바탕으로 작동하는데, 로캘은 프로그램 속으로 하드코딩을 하기 어려운 언어/문화와 관련한 사항을 다룬다. 리눅스에서 다양한 로캘을 만들고 설치하여 여러 나라의 언어에 관한 설정을 할 수가 있다.

본 논문에서는 북쪽의 한글 가나다 차례를 남쪽에서 활용할 수 있도록 하기 위해서 북쪽의 한글 가나다 차례를 포함하는 북쪽 로캘을 만들고, 입력된 글자떼를 남쪽 혹은 북쪽의 한글 가나다 차례에 따라 간추리기 할 수 있는 프로그램을 개발하였다.

2. 관련 연구

2.1 ISO 14651

일반적으로 간추리기 (=정렬, =sort, =collating) 프로그램에 글자를 간추리는 차례가 들어 있다. 따라서 새로운 글자를 발견하거나, 기존의 간추리는 차례를 변경해야 하는 경우에 간추리기 프로그램 자체를 변경해야 한다. ISO 14651을 이용하면 간추리는 차례가 공통 틀 표 (Common Template Table) 라는 표에 들어 있다. 글자의 간추리는 차례를 바꿀 경우 간추리기 프로그램을 바꿀 필요 없이 공통 틀 표만 수정하여 글자의 간추리는 차례를 바꿀 수 있다.

ISO 14651은 글자떼 (=문자열) 의 순서를 결정하기 위하여 각 글자를 정의하는 상징 (=symbol) 을 이어 붙여서 글자떼의 간추리기 키 (=collating key) 를 만든다. 각 글자의 상징은 보통 4 단계 (=level) 로 정의하는데, 한글은 1 단계만 사용해도 된다. 따라서 나머지 단계에 대한 내용은 [2]을 참고하기 바란다.

"가다" 부호값(code):	U+1100 1161 1103 1161
"ㄱ "	<U1100> <S1100>
"ㅏ "	<U1161> <S1161>
...	
1단계 키:	<S1100><S1161><S1103><S1161>

[그림 1] ISO 14651의 보기

논문에서 사용하는 ISO 14651 관련 용어는 다음과 같다.

- 간추리기 상징 (=Collating Symbol): 글자를 대신해서 사용할 상징으로, 각 글자의 순서를 지정한다. 보기를 들면, [그림 1] 의 <S1100>이다.
- 간추리기 요소 (=Collating Element): 비교할 때 1개의 단위로 취급하는 글자 혹은 글자떼이다. 보기를 들면, [그림 1]의 "ㄱ"을 표시하는 <U1100>이다.
- 간추리기 표 (=Collation Table): 간추리기 상징의 순서를 정의하고, 간추리기 요소를 해당하는 상징으로 정의한 표를 말한다. 공통 틀 표의 가장 중요한 부분이다.

2.2 리눅스에서 ISO 14651 지원

리눅스에서는 로캘을 지원하는데, 로캘은 화폐 단위 및 표기, 연월일의 순서와 같이 지역에 따라 바뀌는 사항을 정의한다. 리눅스 라이브러리 glibc 2.1.94 부터 ISO 14651을 지원하는 함수인 strcoll()/strxfrm() 가 추가되면서 LC_COLLATE 라는 부분이 로캘에 추가되었다.

이 로캘 변수 부분은 공통 틀 표를 포함하고 있으므로 LC_COLLATE 변수에 원하는 글자 순서의 로캘을 설정하면 글자 순서를 바꿀 수 있다. 구체적인 보기를 들어, 남쪽 가나다 차례를 원하면 남쪽 가나다 차례를 LC_COLLATE 로캘에 넣어둔 뒤 strcoll()/strxfrm() 을 쓰면 남쪽 가나다 차례대로 간추릴 수 있으며, 북쪽 가나다 차례를 원하면 북쪽 가나다 차례를 LC_COLLATE 로캘에 넣어둔 뒤 strcoll()/strxfrm() 을 쓰면, 프로그램을 전혀 바꾸지 않고 북쪽 가나다 차례대로 간추릴 수 있다.

남쪽의 가나다 차례를 반영한 ISO 10646의 부호값을 기준으로 ISO 14651의 공통 틀 표를 정의하였으며, 리눅스의 로캘 파일에는 북쪽의 로캘을 포함하고 있지 않기 때문에 북쪽의 가나다 차례로 간추리는 프로그램을 만들기 위해서는 북쪽 로캘 파일을 만들어야 한다.

3. 간추리기 프로그램 개발

3.1 북쪽의 로캘 개발

로캘의 글자 순서는 글자의 부호계와 연관되어 있는데, 북쪽의 부호계인 EUC-KP는 준비된 것이 없다. 또한 EUC-KP 부호계를 따르는 글자때는 만들 수 있어도 남쪽에서는 그 결과를 쉽게 확인할 수 없다. 따라서 본 시스템에서는 EUC-KR 부호계를 기반으로 북쪽의 가나다 차례를 반영한 로캘을 이용하였다. [그림 2]와 같이 EUC-KR 부호값이 간추리기 상징에 할당되어 있다.

```
...
<UAC00>     /xb0/xa1     가
<UAC01>     /xb0/xa2     각
<UAC04>     /xb0/xa3     간
...
```

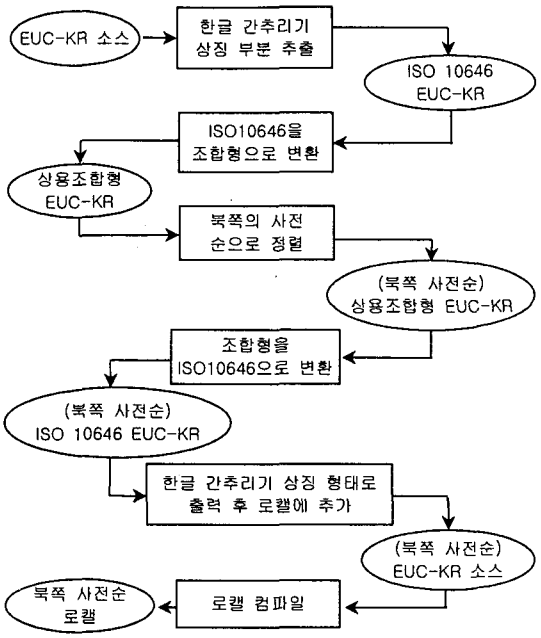
[그림 2] 부호값과 간추리기 상징의 대응

북쪽의 가나다 순서는 [그림 3]과 같다.

```
첫소리 글자(19 개): ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, ㄲ, ㄸ, ㅃ, ㅆ, ㅉ, ㅇ
가운뎃소리 글자(21 개): ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ, ㅐ, ㅒ, ㅔ, ㅖ, ㅚ, ㅟ, ㅘ, ㅝ, ㅙ, ㅞ
끝소리 글자(27 개): ㄱ, ㄲ, ㄴ, ㄵ, ㄶ, ㄷ, ㄹ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅁ, ㅂ, ㅄ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ, ㄳ, ㅆ
```

[그림 3] 북쪽의 가나다 순서

북쪽의 로캘을 만든 순서는 [그림 4]의 순서와 같다. 이를 요약하면, 처음에 ASCII로 만들어진 EUC-KR 소스 파일에서 한글 부분의 간추리기 상징을 추출한다. 다음에 2진값으로 저장하여 부호계를 바꾼다. 세 번째로 조합형 한글의 각 글자를 북쪽 가나다 차례에 따라 글자의 순서를 바꾼 다음, 다시 간추리기 상징의 형태로 만들어 낸다. 마지막으로 로캘을 컴파일하여 새로운 로캘을 만들어 낸다.



[그림 4] 북쪽 로캘 만든 순서. 간추리기 상징은 ASCII로 작성되어 있으므로 2진수로 변환하는 과정이 필요.

3.2 프로그램 개발 및 운영 환경

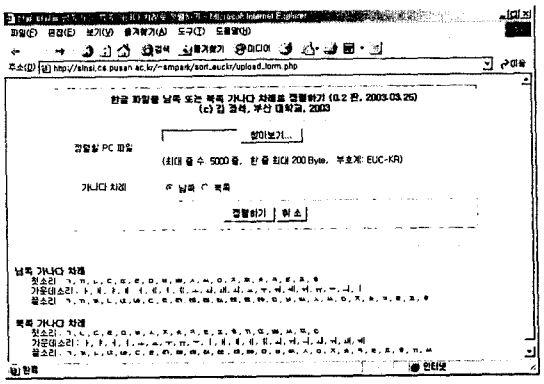
- 플랫폼: 인텔 펜티엄3
- 운영체제: 맨드레이크 리눅스 8.2
- 개발 언어: C, PHP
- 사용자 환경: 웹
- 제약 사항:
  ㄱ) 입력 파일 부호계: EUC-KR
  ㄴ) 입력 파일 크기: 5000줄, 한 줄은 200 바이트

3.3 시스템 구성

1) 남북쪽의 로캘
   남쪽의 로캘과 본 논문의 3.1에서 만든 북쪽 로캘
2) 파일을 입력받는 프로그램 (upload_form.php)
   폼에서 입력한 파일을 받아서 files/ 디렉토리에 저장하고, 정렬 프로그램인 sorting을 실행한다. sorting의 실행 결과로 작성된 파일을 읽어 브라우저에 보여준다.
3) 글자의 순서를 결정하는 프로그램 (sorting)
   2) 에서 받은 사용할 로캘과 정렬할 파일 이름을 인자로 받아서 실행한다. setlocale() 함수로 사용할 로캘을 지정하고, strcoll() 함수로 글자때의 순서를 결정한다. 정렬한 결과는 파일로 저장하여 웹에서 보이도록 한다.

3.4 실행

1) 브라우저 주소 입력창에 프로그램의 URL을 입력한다.
2) 정렬할 지역 (local) 파일을 찾아보기 버튼을 눌러 선택한다.
3) 정렬 기준 (남쪽 또는 북쪽 가나다 차례) 을 선택한다.
4) 정렬하기 버튼을 누른다.
5) 선택한 가나다 차례에 따라 정렬된 파일이 보일 것이다.



[그림 5] 실행 화면

3.5 실행 결과

입력 파일	출력 파일	
	남쪽 가나다 차례	북쪽 가나다 차례
셔틀버스	가족	가족
카세트	개미	건빵
가족	건빵	귤
떡볶이	귤	기차
건빵	기차	개미
빵	까마귀	나비
타조	나비	다락방
나비	다락방	라면
파전	떡볶이	마늘
사진	라면	바람
개미	마늘	사진
학교	바람	셔틀버스
아기	빵	자동차
자동차	사진	카세트
까마귀	셔틀버스	타조
마늘	아기	파전
기차	자동차	학교
귤	짜장면	까마귀
다락방	카세트	떡볶이
라면	타조	빵
짜장면	파전	짜장면
바람	학교	아기

[그림 6] 입력 파일을 남쪽 가나다 차례와 북쪽 가나다 차례로 간추린 결과

4. 결론

ISO 14651을 이용하면 간추리는 차례가 공통 틀 표 (Common Template Table) 에 들어 있기 때문에 글자의 간추리는 차례를 바꾸더라도 간추리기 프로그램을 바꿀 필요가 없다. 남쪽과 북쪽의 가나다 차례를 통일하지 않고 순서가 다른 문제를 해결할 수 있는 방법은 ISO 14651 을 활용해야 한다.

본 논문에서는 북쪽의 한글 가나다 차례를 남쪽에서 활용할 수 있도록 하기 위해서 북쪽의 한글 가나다 차례를 포함하는 북쪽 로캘을 만들고, 입력된 글자떼를 남쪽 혹은 북쪽의 한글 가나다 차례에 따라 간추리기 할 수 있는 프로그램을 개발하였다.

본 연구에서 북쪽의 가나다 차례를 반영한 로캘을 EUC-KR을 사용하였으나, 향후 연구에서는 북쪽의 EUC-KP도 함께 지원하는 로캘을 만들어야 할 것이다. 또한 남쪽과 북쪽에서 ISO 14651을 효과적으로 활용할 수 있는 방안을 모색해 봐야할 것이다.

5. 참고 문헌

[1] 김 경석, "국제 표준에 따른남북 가나다 차례 지원 방안 연구", 한국 표준 협회, 2001
[2] ISO/IEC. "ISO/IEC 14651 - International String Ordering and Comparison", ISO/IEC, 2000
[3] 옥 제영, 정 일동, 김 경석, "ISO/IEC 14651 틀에서 한글 간추리기 문제점에 대한 해결 방안", 한국 멀티미디어 학회 추계 학술 발표 논문집, 제 5권 2호, pp. 59~64, 11월, 2002
[4] 옥 제영, "ISO/IEC 14651 틀에서 한글 간추리기 문제점에 대한 해결 방안", 부산대학교, 2003