Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00026

GIẢI PHÁP CHUYỂN ĐỔI VĂN BẢN TIẾNG Ê ĐÊ DÙNG PHÔNG CHỮ RIÊNG SANG UNICODE

Hoàng Thị Mỹ Lệ [1], Phan Huy Khánh[2]

[1] Trường Cao đẳng Công nghệ, Đại học Đà Nẵng

[2] Trường Đại học Bách khoa Đà Nẵng, Đại học Đà Nẵng

htmle@dct.udn.vn, phkhanh@dut.udn.vn

TÓM TẮT: *Trong soạn thảo văn bản việc sử dụng nhiều bảng mã khác nhau trong cùng một nước là một trở ngại lớn trong việc phát triển các hệ thống thông tin lớn. Giải pháp toàn vẹn nhất cho sự không tương thích giữa các bảng mã, đó là sử dụng Unicode. Unicode không chỉ giải quyết về mặt kĩ thuật hiển thị phông chữ mà còn tạo tiền đề cho sự phát triển kĩ thuật xử lý ngôn ngữ trên máy tính, xây dựng các giải pháp tự động sửa lỗi chính tả và ngữ pháp trên máy tính và là xu hướng tất yếu trong sự phát triển của internet hiện nay... Việc sử dụng Unicode là giải pháp quốc tế cho mọi ngôn ngữ trên thế giới trong việc trao đổi và sử dụng thông tin. Các văn bản, tài liệu tiếng Ê Đê hiện nay chủ yếu vẫn dùng bộ phông chữ TayNguyenKey. Vì vậy, nếu trên máy tính không có bộ phông chữ này thì sẽ không hiển thị được chữ viết tiếng Ê Đê. Trong nghiên cứu xử lý tiếng Ê Đê không sử dụng được các tài liệu điện tử tiếng Ê Đê hiện có. Hiện nay, việc chuyển từ dùng phông chữ riêng sang Unicode trong soạn thảo văn bản tiếng Ê Đê là công việc chưa thể thực hiện được trong ngày một ngày hai. Trước mắt, để giải quyết những khó khăn trong việc trao đổi các văn bản tiếng Ê Đê dùng phông chữ riêng trên internet hay giữa các máy tính. Bài báo đề xuất giải pháp chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang phông chữ Unicode.*

Từ khóa: *dân tộc Ê Đê, xử lý tiếng Ê Đê, mã hóa Unicode, soạn thảo văn bản.*

I. GIỚI THIỆU

Tiếng nói và chữ viết là hai yếu tố cơ bản của ngôn ngữ tự nhiên. Vấn đề mã hóa ký tự cho từng ngôn ngữ để lưu trữ và hiển thị thông tin trên máy tính là vấn đề đặt ra đầu tiên trong lĩnh vực xử lý ngôn ngữ tự nhiên. Việc sử dụng nhiều bảng mã khác nhau là một trở ngại lớn trong việc phát triển các hệ thống thông tin. Giải pháp xử lý cho sự không tương thích giữa các bảng mã là sử dụng Unicode. Giải pháp này đã được nhiều quốc gia chấp nhận và được đưa vào làm chuẩn cho việc trao đổi và sử dụng thông tin [15].

Tất cả các ký tự của tiếng Việt (dân tộc Kinh) đều có mặt trong Unicode. Vì vậy, để hiển thị tiếng Việt Unicode máy tính cần phải cài đặt phông chữ Unicode. Để cài đặt phông chữ Unicode chỉ cần cài đặt một trong các phần mềm Internet Explore hoặc MS Office trong Windows hoặc cài đặt Windows. Khi cài đặt một trong những phần mềm này các phông Unicode có hỗ trợ tiếng Việt sẽ được tự động cài đặt vào máy tính. Các phông chữ cơ bản của Microsoft đi kèm với các phần mềm trên đã hỗ trợ tiếng Việt Unicode như: Times New Roman, Arial, Courier, Tahoma. Ngoài ra có thể tải các phông Unicode có hỗ trợ tiếng Việt khác như Verdana, Arial Narrow, Arial Black, Bookman Old Style, Garamond, Impact, Lucida Sans, Comic Sans,... từ internet. Bộ gõ Unikey và bộ gõ Vietkey là sự thành công của việc đưa được bộ chữ Việt vào Unicode, cũng như việc chọn Unicode cho bộ mã chuẩn tiếng Việt và xây dựng bộ gõ chữ tiếng Việt trên các phông chữ Unicode [7], [14].

Tuy vậy, bộ chữ cái của số tiếng các dân tộc thiểu số (DTTS) sử dụng bộ chữ cái Latinh ở Việt Nam nói chung và tiếng Ê Đê nói riêng vẫn còn một số chữ cái không có mặt trong Unicode. Vì vậy, để hiển thị chữ viết tiếng các DTTS trên máy tính các tác giả thường tập trung xây dựng bộ gõ và tạo bộ phông chữ đi kèm với bộ gõ. Bộ phông chữ được tạo ra thường dựa trên bộ mã chuẩn có sẵn là ASCII 8 bit. Cũng đã có những giải pháp hiển thị chữ viết tiếng các DTTS không cần bộ gõ mà chỉ sử dụng phông chữ riêng, vì khi tạo phông chữ đã ghi đè lên các ký tự có trên bàn phím tiếng Anh. Đây là một hình thức ghi đè hay sử dụng phương pháp "lấp chỗ trống" [8] lên mã ASCII hay mã Unicode. Tuy nhiên đây không phải là những giải pháp khả thi, vì gây tranh chấp giữa các bộ phông chữ khác nhau.

Việc tạo bộ phông chữ riêng có nhiều hạn chế, cụ thể như: các văn bản tài liệu tiếng DTTS khi soạn thảo chưa hoà nhập vào phông chữ Unicode như tiếng Việt, không hiển thị được chữ viết tiếng DTTS trên các trang Web nếu không có phông chữ tiếng DTTS đi kèm. Hay người sử dụng muốn sao chép một đoạn văn bản tiếng DTTS trên trang web xuống để dùng thì không hiển thị được tiếng DTTS nếu trên máy không có phông chữ tiếng DTTS, cụ thể: các bản tin tiếng Ê Đê của Hệ phát thanh dân tộc VOV4, hệ thống dịch tự động song ngữ tiếng Việt-tiếng Ê Đê [4].

Cho đến nay, vấn đề xử lý chữ viết tiếng các DTTS trên máy tính chỉ mới giải quyết cục bộ địa phương cho từng dân tộc, chưa thống nhất theo quy chuẩn mang tính quốc gia. Bên cạnh đó, các bài toán xử lý tiếng các DTTS rất ít được sự quan tâm của các nhà khoa học. Trong bối cảnh bùng nổ sử dụng internet và các dịch vụ trên internet đã tương đối quen thuộc và có mặt hầu khắp mọi miền của Tổ quốc. Cùng với nhu cầu phát triển văn hoá, hội nhập của cộng đồng các DTTS ở Việt Nam. Lúc này nhu cầu xử lý tiếng các DTTS lại càng đặt ra bức thiết hơn bao giờ hết.

Trong khuôn khổ nội dung bài báo, chúng tôi tập trung hướng tiếp cận mã hóa Unicode bộ chữ cái tiếng Ê Đê áp dụng cho giải pháp chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode. Phần tiếp theo bài báo sẽ

trình bày tóm lược một số vấn đề về thực trạng xử lý chữ viết tiếng các DTTS ở Việt Nam nói chung và tiếng Ê Đê nói riêng, vấn đề mã hóa Unicode bộ chữ cái tiếng Ê Đê, giải pháp chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode. Cuối cùng là kết quả thực nghiệm và kết luận.

II. MÃ HÓA UNICODE TIẾNG Ê ĐÊ

A. Thực trạng xử lý chữ viết tiếng các dân tộc thiểu số ở Việt Nam

Vấn đề xử lý chữ viết tiếng Việt đã được triển khai khá sớm và đã có nhiều kết quả đáng kể. Bên cạnh đó, xử lý chữ viết tiếng các DTTS ở Việt Nam đã có được các kết quả sau:

Bộ gõ STVB tiếng Chăm, bộ gõ này đã sử dụng phần mở rộng của ASCII để mã hóa, dùng phần mềm Corel Draw để vẽ 65 ký tự của bộ chữ Chăm. Sau đó dùng phần mềm FontCreator để tạo phông và sử dụng các kiểu gõ tiếng Việt thông dụng như kiểu gõ Telex và kiểu gõ VNI [13].

Bộ phông chữ Ê Đê, tác giả đã xây dựng bộ phông chữ Ê Đê dùng phần mềm FontoGrapher để ghi đè lên một số kí tự trong phông chữ VntimeNewRoman [1].

Bộ gõ STVB tiếng dân tộc, bộ gõ này đã sử dụng phần mở rộng trong bảng mã ASCII để mã hóa các chữ cái có dấu chung cho cả 4 ngôn ngữ Ê Đê, Gia Rai, Ba Na, M'Nông. Phông chữ được tạo bằng công cụ FontoGrapher. Kiểu gõ được dùng là các kiểu gõ Telex và VNI. Bộ gõ này được dùng trong STVB cho bốn tiếng DTTS Ê Đê, Gia Rai, Ba Na, M'Nông và tiếng Việt [5].

Phần mềm chữ Chăm, tác giả đã xây dựng bộ phông chữ bộ phông chữ Chăm Unicode [3].

Chương trình TayNguyenKey, hỗ trợ gõ chữ viết của 6 tiếng DTTS Tây Nguyên Ê Đê, Gia Rai, Ba Na, Xơ Đăng, Cơ Ho và M'Nông. Ngoài ra, còn gõ được tiếng Việt và tiếng Anh. Chương trình này xây dựng bộ phông chữ TayNguyenKey trong STVB 6 tiếng DTTS Tây Nguyên [6].

Bộ gõ STVB VnKey, hỗ trợ gõ tiếng Việt và tiếng các DTTS ở Việt Nam như: Ê Đê, Gia Rai, M'Nông, Cơ Ho, Xê Đăng, Sán Chày [9].

Bộ gõ WinVNKey là bộ gõ để gõ tiếng Việt và hơn 30 tiếng các nước khác [11].

Bộ gõ STVB Standar Thai Son La, có thể gõ văn bản chữ Thái mà không dùng chương trình hỗ trợ bên ngoài. Bằng cách, bộ gõ này tạo bộ phông chữ Thái riêng và ghi đè ký tự chữ Thái lên các chữ cái trên bàn phím tiếng Anh. Đây là một hình thức ghi đè lên các kí tự trong bảng mã ASCII. Giải pháp này gõ được chữ Thái ngay trên bàn phím tiếng Anh, vì số ký tự chữ Thái nhỏ hơn số các phím ký tự trên bàn phím [10].

Bộ gõ STVB của dự án số hóa chữ Thái, sử dụng hệ thống mã hóa Unicode 2 byte, dùng kỹ thuật keyboard hook để quản lý bàn phím thực như Unikey, VietKey. Vấn đề này mới xét trên khía cạnh lý thuyết mà chưa tính tới thực tế cho người sử dụng và chưa được ứng dụng vào thực tiễn [10].

Các kết quả nghiên cứu về xử lý chữ viết tiếng các DTTS ở Việt Nam có những ưu điểm và nhược điểm sau:

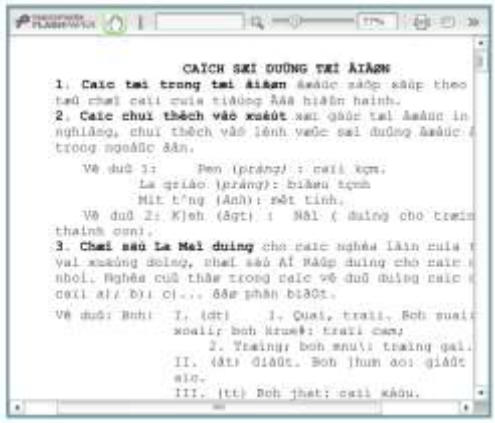
Ưu điểm: góp phần tin học hóa văn bản tiếng các DTTS, giải quyết được vấn đề hiển thị chữ viết tiếng DTTS trên máy tính cho các dân tộc Chăm, Thái, Ê Đê, Gia Rai, Ba Na, M'Nông, Ê Đê, Xơ Đăng, Cơ Ho, Xê Đăng, Sán Chay. Tạo điều kiện cho đồng bào các DTTS tiếp cận với những ứng dụng trong lĩnh vực Công nghệ Thông tin, cũng như các ứng dụng khoa học công nghệ mới.

***Nhược điểm*:** chưa sử dụng Unicode trong soạn thảo các văn bản tiếng DTTS.

Chính nhược điểm này mà các văn bản tiếng DTTS giữa các cơ quan nhà nước chưa trao đổi được với nhau qua internet. Chữ viết tiếng DTTS không hiển thị được trên các trang web với phông Unicode hay trên các máy tính không có phông chữ riêng tương ứng.

Hình 1 và Hình 2 minh hoạ nhược điểm về vấn đề hiển thị chữ viết tiếng Ê Đê dùng phông chữ riêng.

Hình 1. Minh họa hiển thị chữ viết tiếng Ê Đê dùng phông chữ riêng trên máy tính



Hình 2. Minh họa hiển thị chữ viết tiếng Ê Đê trên trang web khi dùng phông chữ riêng

Tóm lại, các kết quả nghiên cứu trên vẫn chưa có tính hệ thống, chưa định hướng rõ ràng, còn rời rạc, thiếu chia sẻ, mang tính địa phương, chỉ phục vụ cho cộng đồng DTTS từng huyện, từng tỉnh.

Từ những thực trạng xử lý chữ viết tiếng các DTTS ở Việt Nam trên máy tính, vấn đề mã hóa Unicode trong STVB tiếng các DTTS là rất cần thiết và phải được triển khai càng sớm càng tốt. Bởi vì, các kho ngữ liệu, các website ngày càng phát triển với dữ liệu càng lớn thì sau này quá trình chuyển mã sẽ phức tạp và tốn kém.

B. Sử dụng Unicode cho tiếng Ê Đê

Bộ chữ cái tiếng Ê Đê cũng được xếp vào họ Latinh, gồm 76 chữ cái (kể cả kí tự hoa và kí tự thường) như tromg Bảng 1. Trong đó, 68 chữ cái đã có trong Unicode, còn 8 chữ cái **(ẽ, õ, ở, ử, Ẽ, Ỏ, Ơ, Ư)** chưa có trong Unicode. Vì vậy, cho đến nay vẫn chưa có giải pháp nào đưa 8 chữ cái này vào Unicode.

Bảng 1. Bảng chữ cái Ê Đê

	Phụ âm										**Nguyên âm**									
Chữ hoa	B	Ɓ	Č	D	Đ	G	H	J	K	L	A	Ă	Â	E	Ĕ	Ê	Ẽ	I	Ĭ	O
	M	N	Ñ	P	R	S	T	W	Y		Ŏ	Ô	Ỏ	Ơ	Ơ	U	Ŭ	Ư	Ư	
Chữ thường	b	ɓ	č	d	đ	g	h	j	k	l	a	ă	â	e	ĕ	ê	ẽ	i	ĭ	o
	m	n	ñ	p	r	s	t	w	y		ŏ	ô	õ	ơ	ở	u	ŭ	ư	ử	

Áp dụng giải pháp sử dụng Unicode trong mã hóa tiếng DTTS ở Việt Nam [2] cho việc mã hóa Unicode chữ viết tiếng Ê Đê như sau:

- Bước 1: nhóm các chữ cái tiếng Ê Đê theo ba nhóm (Bảng 2).
  Nhóm 1 gồm 54 chữ cái có trong bảng chữ cái tiếng Việt và có trong Unicode.
  Nhóm 2 gồm 14 chữ cái không có trong bảng chữ cái tiếng Việt, mà có trong Unicode.
  Nhóm 3 gồm 8 chữ cái không có trong bảng chữ cái tiếng Việt, cũng không có trong Unicode.

Bảng 2. Bảng phân nhóm bảng chữ cái tiếng Ê Đê

Nhóm	**Chữ cái tiếng Ê Đê**																	
1	A	a	Ă	ă	Â	â	E	e	Ê	ê	I	i	O	o	Ô	ô	Ơ	ơ
	U	u	Ư	ư	B	b	D	d	Đ	đ	G	g	H	h	J	j	K	k
	L	l	M	m	N	n	P	p	R	r	S	s	T	t	W	w	Y	y
2	Ɓ	ɓ	Č	č	Ĕ	ĕ	Ĭ	ĭ	Ñ	Ñ	Ŏ	ŏ	Ŭ	ŭ				
3	Ê	ẽ	Ỏ	õ	Ơ	ở	Ư	ử										

- Bước 2: đặt chữ cái nhóm 2 và nhóm 3 vào Unicode trong các phạm vi:
  Kí tự Latinh bổ sung (H00A0:H00FF)
  Kí tự Latinh mở rộng (H0100:H024F)
  Dấu phụ kết hợp (H0300:H036F)

Kết quả đặt chữ cái nhóm 2 và nhóm 3 vào Unicode thể hiện trong Bảng 3.

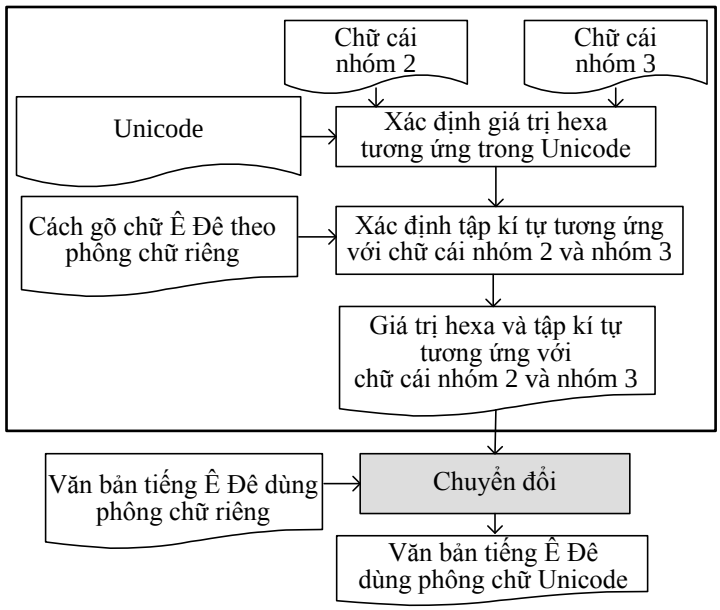
Bảng 3. Đặt chữ cái nhóm 2 và nhóm 3 vào Unicode

Nhóm	Giá trị hexa của chữ cái tiếng Ê Đê trong Unicode							
2	Ƀ H0243	ƀ H0180	Č H010C	č H010D	Ĕ H0114	ĕ H0115	Ǐ H012C	ǐ H012D
	Ñ H00D1	ñ H00F1	Ŏ H014E	ŏ H014F	Ŭ H016C	ŭ H016D		
3	Ê H00CA H0306	ê H00EA H0306	Ô H00D4 H0306	ô H00F4 H0306	Ơ H01A0 H0306	ơ H01A1 H0306	Ư H016C H0306	ư H016D H0306

III. CHUYỂN ĐỔI VĂN BẢN TIẾNG Ê ĐÊ DÙNG PHÔNG CHỮ RIÊNG SANG UNICODE

A. Giải pháp chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode

Hiện nay các văn bản tiếng Ê Đê không dùng phông chữ Unicode mà dùng phông chữ riêng. Việc dùng phông chữ riêng trong STVB là một khó khăn trong việc trao đổi và phát triển các hệ thống thông tin. Để góp phần giải quyết những khó khăn trong việc trao đổi, sử dụng các văn bản tiếng Ê Đê dùng phông chữ riêng, bài báo đã đề xuất giải pháp chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode. Giải pháp chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode được thể hiện trong hình 3.



Hình 3. Mô hình chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode

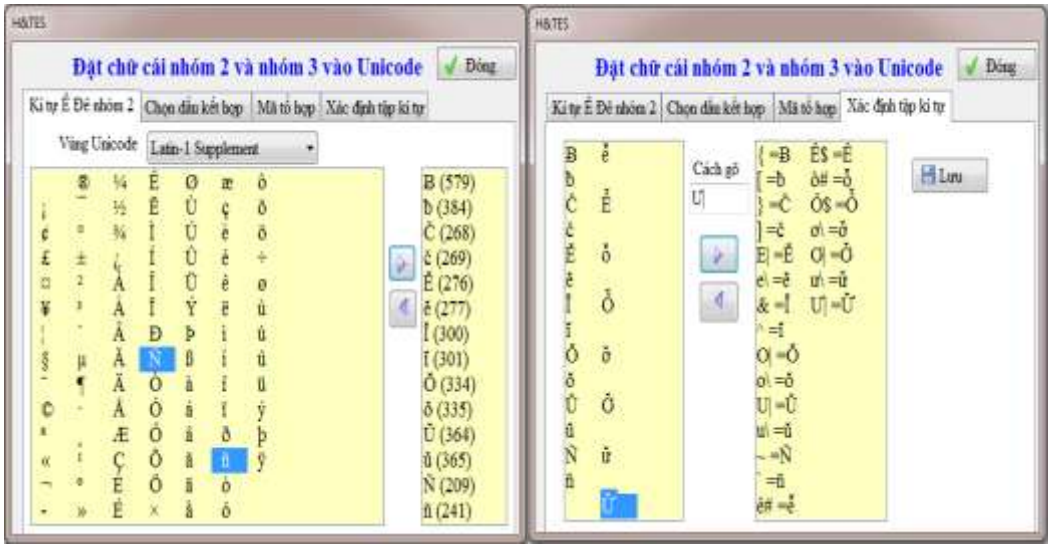
Hoạt động trong mô hình:

- Áp dụng giải pháp mã hóa Unicode chữ cái Ê Đê để xác định giá trị hexa tương ứng trong Unicode.
- Dựa vào cách gõ chữ cái tiếng Ê Đê theo phông chữ riêng để xác định tập kí tự tương ứng với chữ cái thuộc nhóm 2 và nhóm 3.
- Xây dựng bộ công cụ H&TES (hình 4) cho phép lưu giá trị hexa và tập kí tự tương ứng với các chữ cái nhóm 2 và nhóm 3 đã sử dụng trong văn bản cần chuyển đổi vào tập tin cơ sở dữ liệu. Đây chính là nguồn dữ liệu đầu vào cho chức năng chuyển đổi văn bản tiếng DTTS dùng phông chữ riêng sang Unicode.
- Chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode. Chức năng chuyển đổi văn bản sẽ thực hiện lần lượt tìm kiếm tập kí tự trong tập tin cơ sở dữ liệu có trong văn bản và thay thế bằng kí tự tương ứng với giá trị hexa.

B. Kịch bản xây dựng bộ chuyển đổi

Dựa vào giải pháp chuyển đổi văn bản tiếng DTTS dùng phông chữ riêng sang phông chữ Unicode, bài báo đề xuất xây dựng bộ công cụ chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode, được đặt tên là CEDU (Convert EDe text Unicode).

Sử dụng bộ công cụ H&TES (hình 4), cho việc tạo tập tin cơ sở dữ liệu để lưu giá trị hexa và tập kí tự được gõ tương ứng với các chữ cái nhóm 2 và nhóm 3.



Hình 4. Bộ công cụ H&TES

Tệp lưu giá trị hexa và tập kí tự được gõ tương ứng với các chữ cái nhóm 2 và nhóm 3 theo bộ phông chữ TayNguyenKey với kiểu gõ VNI thể hiện trong Bảng 4.

Bảng 4. Tệp lưu giá trị hexa và tập kí tự được gõ theo bộ phông chữ TayNguyenKey với kiểu gõ VNI

c	Nội dung tệp	Kí tự	Nội dung tệp	Kí tự	Nội dung tệp
ă	a\ =H0103	ĭ	^ =H012D	ễ	ê# =H00EA+H0306
Ă	A\ =H0102	Ĭ	& =H012C	Ễ	Ê$ =H00CA+H0306
ƀ	[=H0180	Ŏ	O| =H014E	õ	ô# =H00F4+H0306
Ƀ	{ =H0243	ŏ	o\ =H014F	Ô	Ô$ =H00D4+H0306
č	] =H010D	Ñ	~ =H00D1	ỡ	o\ =H01A1+H0306
Č	} =H010C	ñ	` =H00F1	Ơ	O| =H01A0+H0306
ĕ	e\ =H0115	Ŭ	U| =H016C	ử	u\ =H01AF+H0306
Ĕ	E| =H0114	ŭ	u\ =H016D	Ư	U| =H01B0+H0306

Hoạt động của mô đun chuyển đổi văn bản tiếng Ê Đê trong clipboard hay trong tệp có phần mở rộng (.TXT, .DOC, .DOCX, .RFT) dùng phông chữ riêng sang Unicode trong CEDU:

Bước 1: đọc tệp văn bản hoặc nội dung có trong clipboard

Bước 2: đọc tệp (f_HTF.txt) qua việc chọn bộ phông chữ sử dụng và kiểu gõ trên giao diện của CEDU.

Bước 3: lần lượt đọc từng dòng trong f_HTF vào st

Bước 3.1: tách st thành 2 phần st1 và st2. st1 là tập kí tự được gõ tương ứng với các chữ cái nhóm 2 và nhóm 3 theo bộ phông chữ TayNguyenKey, st2 là kí tự tương ứng với giá trị hexa tách được trong st

Bước 3.2: thay thế trong văn bản tất cả tập kí tự st1 bằng chữ cái tiếng Ê Đê tương ứng với giá trị hexa của st2.

C. Kết quả thực nghiệm

Bài báo đã tiến hành thực nghiệm bộ chuyển đổi CEDU với nguồn dữ liệu đầu vào là các bản tin của Hệ phát thanh dân tộc VOV4. Các bản tin này đảm bảo được gõ đúng kỹ thuật và đã được kiểm tra. Phông chữ dùng trong bản tin là phông chữ TayNguyenKey, kiểu gõ VNI và bộ gõ UniKey.

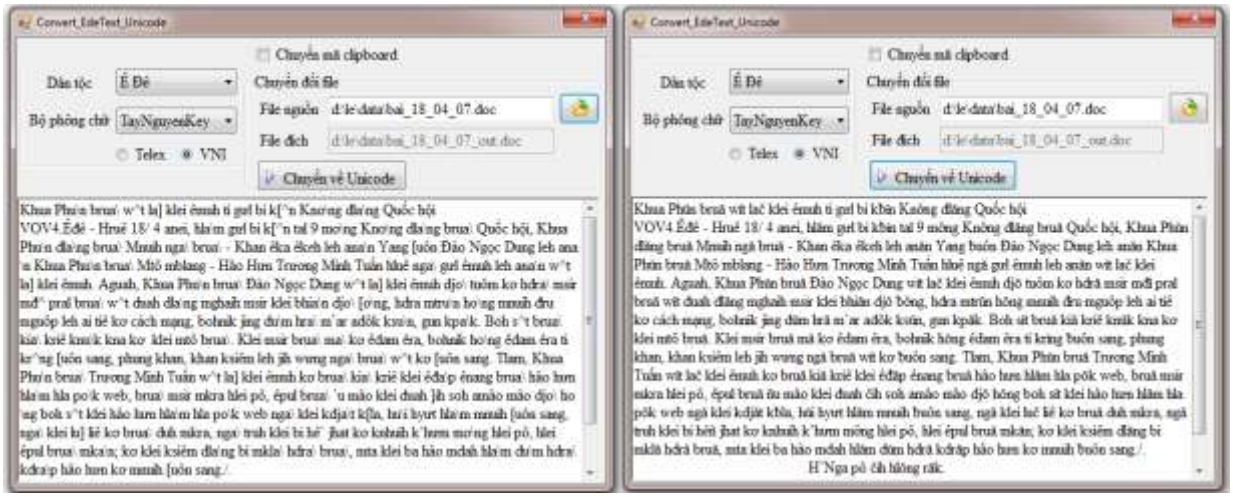
Kết quả thực nghiệm, bộ chuyển đổi CEDU đã chuyển đổi tất cả các tập kí tự được gõ theo phông chữ TayNguyenKey về chữ cái tiếng Ê Đê dùng phông chữ Unicode. Bảng 5 trình bày kết quả thực nghiệm trên 10 bản tin được lưu trên 10 tệp cho mỗi phần mở rộng.

Bảng 5. Kết quả thực nghiệm chuyển đổi văn bản phông chữ riêng sang Unicode

Tệp văn bản có phần mở rộng	Dung lượng	Tập kí tự gõ theo phông chữ TayNguyenKey	Kí tự được chuyển đổi sang Unicode
DOC	301 Kb	2239	2239
DOCX	168 Kb	2239	2239
RFT	433 Kb	2239	2239
TXT	53 Kb	2239	2239

Nguồn dữ liệu đầu vào của bộ chuyển đổi CEDU được lấy từ 20 bản tin của Hệ phát thanh dân tộc VOV 4. Các bản tin này dùng phông chữ TayNguyenKey, kiểu gõ VNI và bộ gõ UniKey. Kết quả có được sau khi qua bộ chuyển đổi CEDU đã được chúng tôi kiểm tra thủ công trên 20 bản tin gốc và nhận thấy rằng bộ chuyển đổi CEDU đã chuyển đổi được tất cả các tập kí tự được gõ theo phông chữ TayNguyenKey về chữ cái tiếng Ê Đê với phông chữ Unicode.

Hình 5 là giao diện của bộ công chuyển đổi CEDU, minh họa văn bản tiếng Ê Đê dùng phông chữ riêng và văn bản tiếng Ê Đê đã được chuyển sang Unicode.



Hình 5. Bộ chuyển đổi CEDU

D. So sánh đánh giá

Bộ chuyển đổi CEDU thực hiện cho các tệp có phần mở rộng .TXT, .DOC, .DOCX, .RTF, .XML, chứ không chỉ giới hạn các tệp chỉ có phần mở rộng .TXT hoặc .RTF như chức năng chuyển đổi bảng mã tiếng Việt của Unikey ToolKit trong bộ gõ Unikey. CEDU đã góp phần giải quyết những bất cập trong việc trao đổi các văn bản tiếng Ê Đê dùng phông chữ riêng trên internet hay giữa các máy tính với nhau, cũng như sử dụng lại các nguồn dữ liệu điện tử hiện có trong nghiên cứu xử lý tiếng Ê Đê.

IV. KẾT LUẬN

Các văn bản tiếng Ê Đê phần lớn không dùng phông chữ Unicode mà dùng phông chữ riêng. Việc dùng phông chữ riêng trong STVB là một khó khăn trong việc trao đổi và phát triển các hệ thống thông tin. Từ thực trạng này, để góp phần giải quyết những khó khăn trong việc trao đổi và sử dụng các văn bản tiếng Ê Đê dùng phông chữ riêng. Bài báo đã đề xuất được giải pháp chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode và đã xây dựng bộ công cụ CEDU thực hiện chuyển đổi văn bản tiếng Ê Đê dùng phông chữ riêng sang Unicode. Bộ chuyển đổi CEDU đã góp phần giải quyết những bất cập trong việc trao đổi các văn bản tiếng Ê Đê dùng phông chữ riêng và sử dụng nguồn tài liệu điện tử đã có trong nghiên cứu xử lý tiếng Ê Đê hiện nay.

TÀI LIỆU THAM KHẢO

[1] Hoàng Thị Mỹ Lệ, "*Xây dựng hệ thống xử lý tin học tiếng Ê Đê trong soạn thảo văn bản*", Luận văn Thạc sĩ ngành Khoa học Máy tính, Đại học Đà Nẵng, 2002.

[2] Hoàng Thị Mỹ Lệ, Phan Thị Bông, Phan Huy Khánh, "*Building a Machine Translation System in a Restrict Context from Ka-Tu Language into Vietnamese*", Proceeding of the International Conference on Knowledge and System Engineering, KSE 2012, Danang, pp. 167-172, 2012.

[3] Hương Giang, "*Đưa vào ứng dụng phần mềm chữ Nôm, chữ Thái, chữ Chăm*", Liên hiệp các Hội Khoa học và Kỹ thuật tỉnh Thừa Thiên-Huế, Địa chỉ: http://www.husta.org/tin-khoa-hoc-cong-nghe/dua-vao-ung-dung-phan-mem-chu-nom-chu-thai-chu-cham.html, 2012.

[4] Lê Quang Hùng, *Khảo sát quan hệ song ngữ tiếng Việt-tiếng dân tộc Ê Đê và xây dựng hệ dịch tự động Việt-Ê Đê*, Đề tài Khoa học và Công nghệ cấp Bộ, mã số B2014-28-07, 2016.

[5] Nguyễn Đức Khanh, "*TayNguyenKey - Chương trình hỗ trợ gõ chữ các dân tộc thiểu số Tây Nguyên, Sở giáo dục Đăk Lăk*", Địa chỉ: http://thpt-ngogiatu-daklak.edu.vn/taynguyenkey-chuong-trinh-ho-tro-go-chu-cac-dan-toc-thieu-so-tay-nguyen.htm, 2010.

[6] Nguyễn Hằng, "*Dự thi TTVN: Xây dựng bộ gõ tiếng dân tộc, Việt Báo*", Địa chỉ: http://vietbao.vn/Vi-tinh-Vien-thong/Du-thi-TTVN-Xay-dung-bo-go-tieng-dan-toc/10844303/226/, 2003.

[7] Phạm Kim Long, "*Bộ gõ tiếng Việt – Unikey"*, http://unikey.vn/vietnam/

[8] Phan Huy Khánh, "*Contribution à Informatique Mulitilingue. Extension D'un Éditeur de Documents Structurés",* Thèse Docteur en Informatique, L'Universitté des Sciences et Techniques de LILLE, 1991.

[9] Quang Huy, Nam Thủy, "*Đưa ngôn ngữ dân tộc thiểu số lên bộ gõ VnKey*", Địa chỉ: http://www.vietnamplus.vn/Home/Dua-ngon-ngu-dan-toc-thieu-so-len-bo-go-VnKey/20112/78629.vnplus 2011.

[10] Sương Cầm, "*Vấn đề bộ gõ và font tiếng Thái*", Học tiếng Thái, Địa chỉ: http://learntaidam.blogspot.com/2012/05/van-e-bo-go-va-font-tieng-thai.html, 2006.

[11] Trần Tư Bình "*Vietnamese & multilingual Keyboard Driver for Windows*", Địa chỉ: http://winvnkey.sourceforge.net/

[12] Trang tin điện tử của Ủy ban Dân tộc Việt Nam CEMA, "*Người Ê Đê"* http://cema.gov.vn/modules.php?name=Content&op=details&mid=498

[13] Trương Kỳ Quốc, Trần Xuân Dũng, "*Xử lý tiếng Chăm - Xây dựng hệ thống soạn thảo văn bản đa ngữ Anh-Việt-Chăm*", Luận văn tốt nghiệp Kỹ sư CNTT, trường ĐHBK, Đại học Đà Nẵng, 2003.

[14] *Vấn đề bộ gõ và font tiếng Thái,* http://learntaidam.blogspot.com/2012/05/van-e-bo-go-va-font-tieng-thai.html

[15] *Vietnamese Unicode*, http://vietunicode.sourceforge.net/main.html.

THE SOLUTION FOR CONVERTING EDE DOCUMENT WITH OWN FONTS TO UNICODE

Hoang Thi My Le, Phan Huy Khanh

ABSTRACT: *In the text editing, using many different character sets in the country is a major obstacle in the development of large information systems. The most complete solution for incompatibility between character sets is to use Unicode. Unicode does not only solve the technical issues of displaying fonts but also creates the foundation for the development of language processing techniques, builds the solutions correcting misspelling and grammar automatically on the computer. Unicode is the indispensable trend in the development of the internet today and the international solution for the every world language in exchanging and using information. So far, the Ede documents are still using the TayNguyenKey font set. Therefore, Ede letter will no display if the computer does not have TayNguyenKey font set. In Ede language processing research do not use the existing Ede electronic documents. Currently, to change using own fonts to using Unicode fonts in Ede language text editing do not be done in a day or two. In the immediate, to solve the difficulties of exchanging Ede documents using own fonts on the internet or between computers. The paper proposes the solution for converting Ede document with own font to Unicode. n the immediate, to solve the difficulties of exchanging Ede documents using own fonts on the internet or between computers. The paper proposes the solution for converting Ede document with own font to Unicode.*