BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

ĐỖ GIA TRINH

XÂY DỰNG KHO DỮ LIỆU SONG NGỮ VIỆT - CƠ TU PHỤC VỤ TRA CỨU VĂN HÓA DÂN TỘC CƠ TU

Chuyên ngành: **Khoa học máy tính**
Mã số: **60.48.01**

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2013

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: PGS.TS. PHAN HUY KHÁNH

Phản biện 1: **TS. ĐẶNG BÁ KHẮC TRIỀU**

Phản biện 2: **TS. NGUYỄN MẬU HÂN**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 18 tháng 5 năm 2013.

Có thể tìm hiểu luận văn tại:

Trung tâm Thông tin - Học liệu, Đại Học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Việt Nam với 54 dân tộc anh em, trong đó dân tộc thiểu số sống rãi rác ở vùng rừng núi cao, dọc theo dãy Trường Sơn hùng vĩ. Đặc điểm địa lý vùng sâu, vùng xa, địa hình đi lại khó khăn, mỗi dân tộc có những đặc trưng văn hóa khác nhau, ngôn ngữ giao tiếp khác nhau tạo nên sự khó khăn trong việc giao lưu học tập, trao đổi văn hóa.

Đồng bào dân tộc Cơ Tu sống ở khu vực miền Trung, cư trú tập trung ở miền núi, vùng cao, vùng biên giới. Đây là vùng đặc biệt khó khăn, kinh tế chậm phát triển; giao thông cách trở; cơ sở hạ tầng còn quá nhiều thiếu thốn; tỷ lệ hộ nghèo cao; trình độ dân trí thấp; thông tin liên lạc còn nhiều hạn chế.

Văn hóa dân tộc Cơ Tu có từ lâu đời, đó là văn hóa Làng, văn hóa cộng đồng và văn hóa dân gian lành mạnh, trong sáng. Văn hóa dân tộc Cơ Tu nói chung, chữ viết của người Cơ Tu nói riêng là một trong những bộ phận cấu thành tạo nên một "Nền văn hóa Việt Nam đậm đà bản sắc dân tộc".

Hiện nay do nhiều nguyên nhân ảnh hưởng đến nên văn hóa và chữ viết dân tộc Cơ Tu dần bị mai một và có nguy cơ mất đi. Đặc biệt, thế hệ trẻ ngày nay đã tiếp cận với nền văn hóa hiện đại ngay từ nhỏ nên không biết tiếng mẹ đẻ. Nguy cơ thất truyền chữ viết của đồng bào Cơ Tu đang là vấn đề rất cấp thiết, rất cần một giải pháp nhằm bảo tồn chữ viết của đồng bào nơi đây.

Thời gian qua, nhiều đề tài nghiên cứu về tiếng Cơ Tu đã được thực hiện, tuy nhiên về mặt tin học thì còn hạn chế. Cho đến nay mới chỉ có đề tài xây dựng bộ gõ tiếng Cơ Tu do tác giả Phạm

Văn Tài, Cán bộ Trung tâm Công nghệ thông tin và Truyền thông thuộc Sở Thông tin và Truyền thông tỉnh Quảng Nam thực hiện.

Các công cụ hỗ trợ học tiếng Cơ Tu như băng, đĩa, từ điển giấy, từ điểm máy tính, giáo viên dạy tiếng Cơ Tu, cũng như số lượng người biết sử dụng tiếng Cơ Tu còn rất ít, đây là một trong những trở ngại lớn cho những người muốn quan tâm tìm hiểu, học tiếng Cơ Tu. Mặt khác, về giáo trình học tập, cũng như các tài liệu tham khảo học tập tiếng Cơ Tu còn hạn chế nên người học không có môi trường để rèn luyện khả năng đọc hiểu và viết tiếng Cơ Tu.

Với sự phát triển mạnh mẽ của công nghệ thông tin, các dịch vụ truyền thông ngày càng trở nên phổ biến và không thể thiếu của con người thì việc xây dựng kho dữ liệu song ngữ Việt – Cơ Tu phục vụ tra cứu văn hóa dân tộc Cơ Tu là điều cần làm nhằm hỗ trợ, phục vụ cho việc tìm hiểu về văn hóa dân tộc Cơ Tu, rút ngắn khoảng cách thông tin giữa đồng bằng và miền núi, giữa các dân tộc, đồng thời giới thiệu bản sắc văn hóa vùng đồng bào dân tộc Cơ Tu đến với đông đảo người dân trên mọi miền tổ quốc và cả thế giới.

Với lý do trên tôi chọn đề tài "Xây dựng kho dữ liệu song ngữ Việt – Cơ Tu phục vụ tra cứu văn hóa dân tộc Cơ Tu".

2. Mục tiêu của đề tài

Mục tiêu chính mà đề tài hướng đến là nghiên c ứu các vấn đề về xử lý ngôn ngữ tiếng Việt như phương pháp tách từ tiếng Việt, kho dữ liệu song ngữ Việt – Cơ Tu,…

Xây dựng kho dữ liệu song ngữ Việt – Cơ Tu phục vụ nhu cầu khai thác, tra cứu văn hóa dân tộc Cơ Tu về các lĩnh vực văn hóa – xã hội, kinh tế, an ninh – quốc phòng.

3. Đối tượng và phạm vi nghiên cứu

Để đáp ứng mục tiêu đã nêu, đề tài cần giải quyết những vấn đề chính sau:

Tìm hiểu lý thuyết

Tìm hiểu chữ viết, văn hóa và đặc trưng ngữ pháp của tiếng Cơ Tu.

Tìm hiểu về phương pháp tách từ tiếng Việt, cơ sở dữ liệu đa ngữ, cách tổ chức kho dữ liệu song ngữ bằng XML.

Xây dựng kho dữ liệu song ngữ

Phân tích cấu trúc cơ sở dữ liệu song ngữ, kho dữ liệu thô, chuyển đổi cơ sở dữ liệu từ dạng winword sang XML.

Cập nhật kho dữ liệu song ngữ Việt – Cơ Tu

Cập nhật kho dữ liệu bằng phương pháp thủ công, cập nhật tự động, tìm hiểu một số phương pháp tách từ tiếng việt.

Xây dựng ứng dụng

Xây dựng chương trình tra cứu song ngữ Việt – Cơ Tu phục vụ nhu cầu khai thác, tra cứu văn hóa dân tộc Cơ Tu của người dùng.

4. Phương pháp nghiên cứu

Phương pháp nghiên cứu lý thuyết

Nghiên cứu tài liệu, công cụ và công nghệ liên quan.

Tổng hợp các tài liệu, dữ liệu.

Phương pháp nghiên cứu thực tế

Tìm hiểu, đi thực tế nghiên cứu về văn hóa dân tộc Cơ Tu tại địa phương.

Phân tích yêu cầu, xây dựng ứng dụng.

Kiểm tra, thử nghiệm và đánh giá kết quả.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Ý nghĩa khoa học

Nắm bắt được các vấn đề cơ bản trong xử lý tiếng Việt. Đây là tiền đề cho các bài toán xử lý ngôn ngữ tự nhiên cho ngôn ngữ của các dân tộc thiểu số (như dịch, từ điển, phần mềm học tập, website đa ngữ…).

Ứng dụng những thành quả của công nghệ thông tin vào lĩnh vực văn hóa.

Ý nghĩa thực tiễn

Xây dựng kho dữ liệu song ngữ Việt - Cơ Tu tra cứu thông tin về văn hóa dân tộc Cơ Tu, giúp gìn giữ bản sắc văn hóa và chữ viết dân tộc Cơ Tu; đồng thời phục vụ cho nhu cầu dạy và học tiếng Cơ Tu cho các thế hệ người Cơ Tu. Góp phần xây dựng khối đại đoàn kết dân tộc, giữ gìn an ninh biên giới.

6. Cấu trúc luận văn

Báo cáo của luận văn được tổ chức thành 3 chương.

Chương 1. Tìm hiểu dân tộc Cơ Tu: Chương này tìm hiểu về văn hóa, chữ viết của dân tộc Cơ Tu.

Chương 2. Xây dựng kho dữ liệu song ngữ: Trình bày các vấn đề xử lý khi xây dựng kho dữ liệu song ngữ, xây dựng kho dữ liệu song ngữ Việt – Cơ Tu, cập nhật kho dữ liệu song ngữ.

Chương 3. Triển khai ứng dụng và đánh giá kết quả: Đề xuất giải pháp xây dựng chương trình tra cứu song ngữ Việt – Cơ Tu. Mô tả, phân tích và đánh giá kết quả chương trình.

CHƯƠNG 1
TÌM HIỂU DÂN TỘC CƠ TU

1.1. TÌM HIỂU VỀ VĂN HÓA VÀ TIẾNG CƠ TU

1.1.1. Giới thiệu dân tộc Cơ Tu

Trong 54 dân tộc ở nước ta, dân tộc Cơ Tu được xếp thứ 26 trong danh mục các tộc người ở Việt Nam. Theo điều tra năm 2003, người Cơ Tu ở Việt Nam có 56.569 người chủ yếu ở 03 tỉnh, thành phố: Quảng Nam, tập trung ở huyện Tây Giang, Đông Giang, 06 xã ở huyện Nam Giang (Thị trấn Thành Mỹ, xã Cà Di, Ta Bhing, Chà Vàl, Laê và thôn Công Tơ Rôn – xã Ladê); huyện Đại Lộc tại thôn Yều (Đại Hưng); thành phố Đà Nẵng, người Cơ Tu ở 02 xã Hòa Phú và Hòa Bắc – huyện Hòa Vang; Thừa Thiên Huế, tập trung ở huyện Alưới tại các xã: Hương Lâm, Hương Nguyên và người Cơ Tu sống xen kẽ với dân tộc Tà Ôi tại các xã Hồng Hạ, Ađớt, Hồng Thượng và tại huyện Nam Đông có người Cơ Tu sống ở các xã: Hương Hữu, Thượng Long, Thượng Nhật, Thượng Quảng, Thượng Lộ và Hương Sơn. Ngoài ra người Cơ Tu còn cư trú ở 02 huyện Đắc Chưng và Kà Lừm tỉnh Xê Công (Lào), có dân số trên một vạn người.

Hình 1.1 Phân bố dân cư - Dân tộc Cơ Tu tại tỉnh Quảng Nam

1.1.2. Giới thiệu văn hóa dân tộc Cơ Tu

a) Văn hóa làng

Văn hóa dân tộc Cơ tu có từ lâu đời, đó là văn hóa Làng – văn hóa cộng đồng và văn hóa dân gian lành mạnh, trong sáng, rất phong phú và đa dạng. Làng Cơ Tu thường quây quần bên nhau tạo thành một khối thống nhất trong cộng đồng.

Hình 1.2 Nhà Gươl – Dân tộc Cơ Tu

b) Hôn nhân và gia đình

c) Tục lệ ma chay

d) Trang phục

1.1.3. Tìm hiểu tiếng Cơ Tu

a) Lịch sử tiếng Cơ Tu

b) Một vài nét về tiếng Cơ Tu

c) Chữ viết Cơ Tu

d) Đặc điểm ngữ pháp tiếng Cơ Tu

1.2. GIAO THOA VĂN HÓA DÂN TỘC CƠ TU VỚI CỘNG ĐỒNG

1.2.1. Nguồn gốc văn hóa

1.2.2. Sự giao thoa văn hóa

Nghị quyết Trung ương 5 khóa VIII về "Xây dựng và phát triển nền văn hóa Việt Nam tiên tiến, đậm đà bản sắc dân tộc", có đề

ra nhiệm vụ cụ thể để bảo tồn, phát huy và phát triển văn hóa các dân tộc thiểu số. Để gìn giữ và phát triển văn hóa dân tộc Cơ Tu, cần đặt văn hóa dân tộc thiểu số trong bối cảnh chung của văn hóa Việt Nam hiện nay. Duy trì các lễ hội truyền thống của đồng bào với tinh thần gạn đục khơi trong, giúp cho người dân có ý thức tự hào về nền văn hóa của dân tộc mình, biết gìn giữ thuần phong mỹ tục, xóa bỏ những hủ tục, thói quen lạc hậu, biết tiếp thu một cách có chọn lọc tinh hoa văn hóa các dân tộc khác.

Các giá trị văn hóa truyền thống đã tạo ra sức sống, sự phong phú, đa dạng và nét độc đáo trong bức tranh toàn cảnh về văn hóa tộc người Cơ Tu. Những giá trị văn hóa đặc sắc của đồng bào dân tộc Cơ Tu được bảo tồn, phát huy và lưu truyền cho các thế hệ, sẽ làm phong phú thêm kho tàng văn hóa của đại gia đình các dân tộc Việt Nam; đồng thời mở ra khả năng khai thác tuyến du lịch sinh thái miền núi Quảng Nam gắn với những giá trị văn hóa của đồng bào Cơ Tu ở đây.

1.2.3. Phát triển tiếng nói người Cơ Tu

1.3. HIỆN TRẠNG, NHU CẦU HỌC TẬP VÀ BẢO TỒN VĂN HÓA DÂN TỘC CƠ TU

1.3.1. Hiện trạng

Hiện nay, văn hóa và chữ viết dân tộc Cơ Tu đang dần bị mai một và mất đi. Đặc biệt, thế hệ trẻ ngày nay đã tiếp cận với nền văn hóa hiện đại ngay từ nhỏ nên không biết tiếng mẹ đẻ. Nguy cơ thất truyền chữ viết của đồng bào Cơ Tu đang là vấn đề báo động, rất cần một giải pháp nhằm bảo tồn chữ viết của đồng bào nơi đây.

1.3.2. Nhu cầu học tập và bảo tồn văn hóa dân tộc Cơ Tu

Giảng dạy tiếng Cơ Tu nhằm mục đích bảo tồn bản sắc và văn hoá dân tộc Cơ Tu, là một chủ trương lớn của Đảng và Nhà

nước Việt Nam. Số lượng người Cơ Tu sử dụng song ngữ (Việt – Cơ Tu) đang ngày càng nhiều và cộng đồng dân tộc Cơ Tu cũng đang dần dần trở thành cộng đồng song ngữ Cơ Tu - Việt. Sự thành thạo trong nói viết tiếng phổ thông ngày càng nhiều, người Cơ Tu là tín hiệu đáng mừng. Song mặt khác, hiện tượng song ngữ không ý thức sẽ có nguy cơ xói mòn tiếng mẹ đẻ của họ. Điều đó cũng có nghĩa là "vốn quý của dân tộc Cơ Tu, tài sản văn hoá chung của cả nước" bị mai một. Vì vậy việc xây dựng kho ngữ vựng song ngữ Việt – Cơ Tu là vô cùng cấp thiết.

Để phục vụ cho nhu cầu học tập của đồng bào dân tộc Cơ Tu, chương trình phát sóng tiếng Cơ Tu tại các tỉnh Quảng Nam, Đà Nẵng, Huế chính thức đi vào hoạt động. Đặt biệt, ngày 12/10/2009, Chương trình phát thanh tiếng Cơ Tu của Đài Tiếng nói Việt Nam chính thức phát trên sóng FM, Hệ thời sự Chính trị - Tổng hợp (VOV1), có thời lượng 30 phút, được phát 3 lần trong ngày (6 giờ 30 phút, 11 giờ 20 phút, 19 giờ 30 phút) đã góp phần vào việc bảo tồn và phát huy tiếng nói, chữ viết của dân tộc Cơ Tu.

CHƯƠNG 2

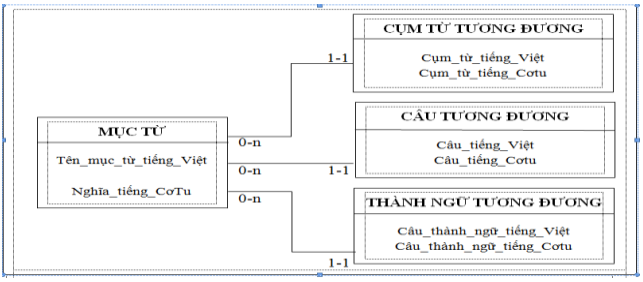
XÂY DỰNG KHO DỮ LIỆU SONG NGỮ

2.1. CƠ SỞ DỮ LIỆU SONG NGỮ VIỆT - CƠ TU

2.1.1. Khái niệm

2.1.2. Cấu trúc cơ sở dữ liệu song ngữ Việt - Cơ Tu

2.1.3. Mô hình thực thể - kết hợp của cơ sở dữ liệu



Hình 2.1 Sơ đồ biểu diễn mô hình ý niệm dữ liệu

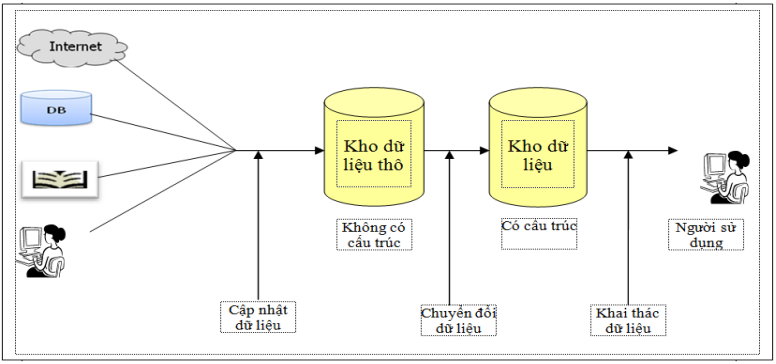
2.1.4. Mô hình logic

2.1.5. Xử lý tiếng Việt qua Unicode

2.1.6. Xử lý tiếng Cơ Tu

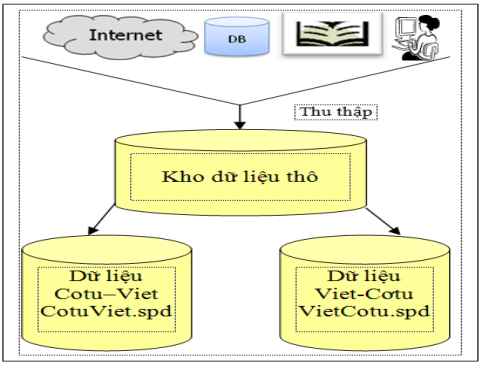
2.2. XÂY DỰNG KHO DỮ LIỆU SONG NGỮ VIỆT – CƠ TU

2.2.1. Tổng quan về quá trình xây dựng kho dữ liệu



Hình 2.2 Mô hình tổng quan xây dựng kho dữ liệu

2.2.2. Cấu trúc kho dữ liệu thô



Hình 2.3 Cấu trúc kho dữ liệu thô

Hai tập tin trên được tổ chức dưới dạng các mục từ.

2.2.3. Xây dựng Cơ sở dữ liệu song ngữ dạng Winword

Việc xây dựng cơ sở dữ liệu song ngữ là một công việc tốn nhiều thời gian, công sức, cũng như rất phức tạp về mặt cấu trúc ngữ pháp từ. Mặc khác, trên thực tế hiện nay từ vựng về tiếng Cơ Tu chưa hề có trên máy vi tính, để cập nhật tiếng Cơ Tu vào cơ sở dữ liệu, chủ yếu sử dụng từ điển giấy do Viện Ngôn Ngữ Học phối hợp với Sở Khoa học Công nghệ tỉnh Quảng Nam xuất bản gồm 9.500 từ. Trong đó gồm 4.500 từ Cơ Tu đối chiếu với nghĩa tiếng Việt và 5.000 từ tiếng Việt đối chiếu với nghĩa Cơ Tu. Phần lớn các từ trong quyển từ điển này là vốn từ cơ bản, thông dụng của tiếng Cơ Tu.

Dữ liệu dạng Winword được tổ chức thành 02 tệp văn bản, tiếng Việt – tiếng Cơ Tu và tiếng Cơ Tu – tiếng Việt. Cấu trúc tệp ngữ vựng gồm hai phần: Phần đầu là phần định dạng, phần thứ hai là phần hiển thị nội dung. Các yếu tố thuộc mục từ trong tệp RTF là các style trong Microsoft Word, một Style bao gồm các thành phần: tên kiểu (Stylename), tên Font (Fontname), kích cỡ chữ (Fontsize)...

Việc chèn tiếng Cơ Tu vào cơ sở dữ liệu song ngữ bằng cách thực hiện một cách thủ công, trước mắt chèn trực tiếp nghĩa tiếng Việt cũng như các cụm từ, câu ví dụ,... và tiếng Cơ Tu tương đương vào các tập tin RTF. Do nguồn từ điển tiếng Việt – Cơ Tu còn hạn chế, chủ yếu là trên từ điển giấy nên công việc này đòi hỏi mất rất nhiều thời gian.

2.2.4. Chuyển đổi sang XML

a) Giới thiệu XML

b) Tổ chức cơ sở dữ liệu Việt – Cơ Tu dưới dạng XML

Đầu tiên xây dựng phần tử gốc có tên là dictionary, trong dictionary có nhiều phần tử con như word chứa các thẻ dữ liệu tương ứng với các style được định nghĩa trong tệp RTF, đó là các phần tử con VietEntry. Mỗi phần tử con VietEntry chứa các thẻ dữ liệu EntryName; CotuEqu; CotuPron; VietPhr; CotuPhr; VietExp; CotuExp; VietIdi; CotuIdi.

Tên thẻ	*Nội dung hiển thị*
Word	Mục từ
EntryName	Tên mục từ
CotuEqua	Nghĩa tiếng Cơ Tu tương đương
CotuPron	Phiên âm tiếng Cơ Tu
VietPhr	Cụm từ tiếng Việt
CotuPhr	Cụm từ tiếng Cơ Tu tương đương
VietExp	Câu ví dụ tiếng Việt
CotuExp	Câu ví dụ tiếng Cơ Tu tương đương
VietIdi	Câu thành ngữ tiếng Việt
CotuIdi	Câu thành ngữ tiếng Cơ Tu tương đương

Hình 2.5 Mô tả các thẻ trong tệp XML

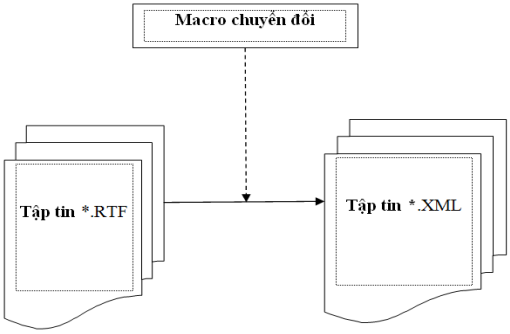
c) Ví dụ minh họa

d) Chuyển đổi cơ sở dữ liệu từ dạng Winword sang XML

Từ cơ sở dữ liệu dưới dạng tập tin Winword đã có, ta xây dựng macro chuyển đổi tập tin Winword sang dạng tệp XML, như sau:

Tên kiểu đoạn (trong tập tin RTF)	*Thẻ mở (trong tập tin XML)*	*Thẻ đóng (trong tập tin XML)*
EntryName	<EntryName>	</EntryName>
CotuEqua	<CotuEqua>	</CotuEqua>
CotuPron	<CotuPron>	</CotuPron>
VietPhr	<VietPhr>	</VietPhr>
CotuPhr	<CotuPhr>	</CotuPhr>
VietExp	<VietExp>	</VietExp>
CotuExp	<CotuExp>	</CotuExp>
VietIdi	<VietIdi>	</VietIdi>
CotuIdi	<CotuIdi>	</CotuIdi>

Hình 2.6 Các kiểu đoạn sử dụng trong CSDL



Hình 2.7 Sơ đồ chuyển đổi các tập tin RTF thành XML

Việc tạo ra cơ sở dữ liệu song ngữ dưới dạng tập tin XML tạo điều kiện thuận lợi khi mô tả cấu trúc một mục từ, dễ dàng thay đổi lại hay bổ sung thêm, hoàn toàn có tính mở. Có thể truy xuất dữ

liệu trực tiếp thông qua tên thẻ bằng cách dùng mã lệnh JavaScrip, nhất là khi định dạng thông qua các tập tin CSS, XSL. Đồng thời kích thước các tập tin nhỏ hơn nhiều lần so với định dạng DOC, RTF…

Bên cạnh đó, việc cập nhật, bổ sung thông qua giao diện khai thác vào các tập tin XML hiện tại còn khó khăn, vì nó là tập tin văn bản.

Thuật toán chuyển đổi tệp Winword sang dạng XML

```
Mở tệp RTF
Khai báo các biến
While not EOF(f)
    If stylename “entry” {từ tiếng Việt}
        Gán mục từ tiếng Việt cho biến filename
        Tạo tệp(filename.xml)
        Tạo tiêu đề tệp XML
        xml = roottag_mo + nội dung + roottag_dong
        {roottag phần tử thẻ gốc}
        While xstyle <> “entry” //đọc nội dung của mục từ
          If stylename = “các style được chọn”
          xml = xml+<tên_style> + xnoidung + </tên_style>
                Endif
        Loop
                Ghi_tệp(filename.xml)
    Endif
```

2.3. CẬP NHẬT KHO DỮ LIỆU SONG NGỮ

2.3.1. Cập nhật thủ công

Dựa trên cấu trúc kho dữ liệu được xây dựng, ta có thể sử dụng các công cụ hỗ trợ để thiết kế các giao diện cập nhật dữ liệu cho kho.

Giải pháp cập nhật thủ công như sau:

Căn cứ trên bộ từ điển giấy do Viện Ngôn Ngữ Học phối hợp với Sở Khoa học Công nghệ tỉnh Quảng Nam xuất bản gồm 9.500 từ. Trong đó gồm 4.500 từ Cơ Tu đối chiếu với nghĩa tiếng Việt và 5.000 từ tiếng Việt đối chiếu với nghĩa Cơ Tu. Tôi chia làm 2 cơ sở dữ liệu: Cơ sở dữ liệu Việt – Cơ Tu (Viet-Cotu) và cơ sở dữ liệu Cơ Tu – Việt (Cotu-Viet).

Tiến hành nhập liệu cho mỗi cơ sở dữ liệu dưới dạng file Winword.

Từ file dữ liệu ở dạng file Winword, xây dựng ứng dụng chuyển đổi định dạng ban đầu của kho dữ liệu và cập nhật vào chương trình từ điển để khai thác.

Ưu điểm của việc cập nhật thủ công là đơn giản. Tuy nhiên, việc cập nhật thủ công thường có nhiều sai sót, tốn nhiều công sức, việc kiểm dò rất quan trọng.

2.3.2. Cập nhật tự động

Sử dụng phương pháp hợp lưu, dựa trên các cơ sở dữ liệu có sẵn

Phương pháp này cho phép tạo ra kho ngữ vựng mới dựa trên các kho ngữ vựng đã có. Giả sử ta có kho ngữ vựng Cơ Tu – Anh, Anh - Cơ Tu và kho ngữ vựng Việt - Anh, ta có thể hợp lưu và cho ra kho ngữ vựng Việt – Cơ Tu.

Ưu điểm của phương pháp này là tạo được kho ngữ vựng nhanh chóng chính xác dựa trên các kho ngữ vựng đã có. Tuy nhiên

một số tiếng nói chưa được xây dựng kho ngữ vựng nên không thể thực hiện được.

Qua tìm hiểu hiện nay kho ngữ vựng liên quan đến tiếng Cơ Tu vẫn chưa được xây dựng, nên không thể sử dụng phương pháp trên.

2.3.3. Một số phương pháp tách từ

a) Phương pháp Maximum Matching (forward/bachward)

b) Phương pháp giải thuật học cải tiến (Transformation based learning, TBL)

c) Mô hình tách từ bằng WFST và mạng Neural

d) Phương pháp quy hoạch động (Dynamic Programming)

e) Phương pháp tách từ dựa trên thống kê từ trên Internet và giải thuật di truyền

2.4. ỨNG DỤNG CỦA KHO DỮ LIỆU SONG NGỮ

CHƯƠNG 3

TRIỂN KHAI ỨNG DỤNG VÀ ĐÁNH GIÁ KẾT QUẢ

3.1. THIẾT KẾ GIAO DIỆN

3.1.1. Các tiêu chí về thiết kế giao diện

3.1.2. Ý tưởng thiết kế giao diện, chương trình

3.1.3. Công cụ lập trình

3.2. MÔ TẢ HỆ THỐNG

3.2.1. Đối tượng sử dụng

3.2.2. Phân tích hệ thống

Hệ thống cung cấp cho người dùng các chức năng như:

Chọn giao diện tra cứu: Người sử dụng có thể chọn giao diện tra cứu bằng tiếng Việt hoặc tiếng Cơ Tu tùy theo nhu cầu của người

sử dụng. Nếu chọn giao diện hiển thị bằng tiếng Cơ Tu, người sử dụng kích chọn vào nút Tiếng Cơ Tu, giao diện tra cứu sẽ hiển thị bằng tiếng Cơ Tu. Ngược lại, nếu chọn giao diện hiển thị bằng tiếng Việt, người sử dụng kích chọn vào nút Tiếng Việt, ngay lập tức giao diện tra cứu hiển thị bằng tiếng Việt.

Hình 3.1 Giao diện tra cứu bằng tiếng Cơ Tu

Hình 3.2 Giao diện tra cứu bằng tiếng Việt

Chọn dữ liệu: Chọn dữ liệu cần tra cứu Việt – Cơ Tu hay Cơ Tu – Việt.

Tra từ điển: Người dùng có thể tra cứu từ điển Cơ Tu - Việt, Việt – Cơ Tu một cách nhanh nhất để phục vụ cho việc tra cứu của mình. Người dùng có thể nhập từ cần tra hay có thể chọn từ danh sách có sẵn.

Tra cứu đoạn văn bản: Người sử dụng có thể nhập đoạn văn bản cần tra cứu vào mục cần tra cứu và chọn kiểu dữ liệu tra cứu. Trong mục này có sử dụng chức năng tách từ trong đoạn văn bản để người sử dụng dễ dàng tra cứu từng từ đoạn văn bản cần tra.

Thao tác với kho dữ liệu: Hệ thống cho phép người dùng cập nhật và bổ sung kho dữ liệu một cách dễ dàng, làm cho kho dữ liệu ngày càng phong phú hơn.

Tìm kiếm thông tin về bản sắc dân tộc Cơ Tu: Khi người dùng có nhu cầu tìm hiểu văn hóa dân tộc Cơ Tu, người dùng vào mục "Giới thiệu dân tộc Cơ Tu" trên chương trình để xem các thông tin cần thiết.

3.3. PHÂN TÍCH CÁC CHỨC NĂNG CHƯƠNG TRÌNH

3.4. XÂY DỰNG CHƯƠNG TRÌNH

3.4.1. Chức năng tra từ

3.4.2. Chức năng quản lý kho

a) Chức năng thêm từ mới vào kho

b) Chức năng sửa từ

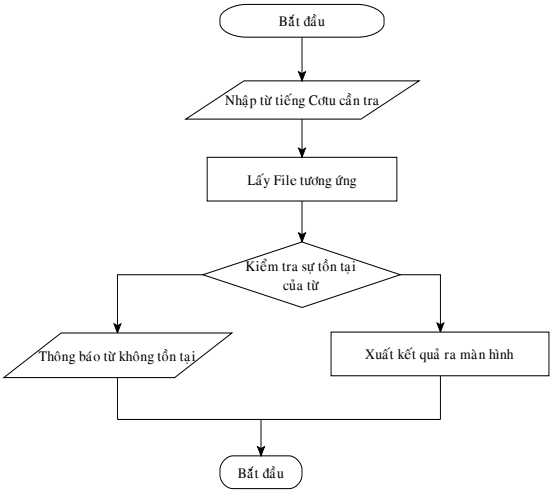
c) Chức năng xóa từ

3.5. THUẬT TOÁN TRA TỪ TRONG KHO DỮ LIỆU

Đầu vào: Nhập từ tiếng Cơ Tu cần tra.

Đầu ra: Giải nghĩa từ tiếng Cơ Tu gồm từ có: Phiên âm La_Tinh, nghĩa tiếng Việt tương đương, cụm từ tiếng Cơ Tu, NTVTĐ cụm từ tiếng Cơ Tu, câu ví dụ tiếng Cơ Tu, NTVTĐ câu ví

dụ tiếng Cơ Tu, ca dao tục ngữ tiếng Cơ Tu, NTVTĐ ca dao tục ngữ tiếng Cơ Tu.



3.6. ĐÁNH GIÁ KẾT QUẢ

Sau thời gian tìm hiểu, nghiên cứu, thực hiện đề tài, tôi đã tìm hiểu văn hóa, đặc điểm tiếng Cơ Tu. Dân tộc Cơ Tu có một bản sắc văn hóa rất phong phú, đa dạng, được lưu truyền qua nhiều thế hệ.

Tôi đã tìm hiểu về hệ thống chữ viết tiếng Cơ Tu, ngữ pháp tiếng Cơ Tu được áp dụng bộ chữ cải tiến của Viện Ngôn ngữ học vào các hoạt động nhằm bảo tồn, giới thiệu chữ viết, con người Cơ Tu đến với mọi miền Tổ quốc và trên thế giới.

Bên cạnh đó, tôi đã tìm hiểu các phương pháp tách từ tiếng Việt, các phương pháp xây dựng kho dữ liệu song ngữ, áp dụng xây dựng kho dữ liệu song ngữ Việt – Cơ Tu.

Thống kê kho dữ liệu đã cập nhật

Kho dữ liệu Việt – Cơ Tu

Số mục từ Việt	*Từ Cơ Tu (đối chiếu)*	*Dung lượng*	*Chiếm tỷ lệ*	*Ảnh minh họa*	*Dung lượng ảnh*
1145 từ	1175 từ	86 KB	25%	85 hình	452 KB

Bảng 3.1 Thống kê từ cập nhập nhật trong kho dữ liệu Việt-Cơ Tu

STT	*Chữ*	*Số mục từ*	*Chiếm tỷ lệ*
1	a	40	100.0%
2	ă	23	100.0%
3	â	26	92.9%
4	b	310	96.0%
5	c	245	40.2%
6	d	10	7.5%
7	đ	104	36.7%
8	e	9	100.0%
9	g	47	26.1%
10	h	9	5.8%
11	i	11	100.0%
12	k	15	7.3%
13	l	58	21.1%
14	m	46	18.3%
15	n	69	12.9%
16	o	11	52.4%
17	ô	8	21.6%
18	ơ	12	92.3%

19	p	5	6.4%
20	q	9	10.5%
21	r	12	5.5%
22	s	9	4.8%
23	t	1	0.2%
24	u	10	40.0%
25	ư	14	100.0%
26	v	10	5.6%
27	x	10	9.1%
28	y	12	100.0%

Kho dữ liệu Cơ Tu – Việt

Bảng 3.2 Thống kê từ cập nhập trong kho dữ liệu Cơ Tu-Việt

Số mục từ Cơ Tu	*Từ Việt (đối chiếu)*	*Dung lượng*	*Chiếm tỷ lệ*	*Ảnh minh họa*	*Dung lượng ảnh*
845 từ	890 từ	79 KB	20,9%	85 hình	452 KB

STT	*Chữ*	*Số mục từ*	*Chiếm tỷ lệ*
1	a	295	98.0%
2	ă	8	100.0%
3	b	11	3.2%
4	c	24	8.4%
5	d	20	24.4%
6	đ	43	17.5%
7	e	8	100.0%
8	ê	3	100.0%
9	g	20	12.7%

STT	*Chữ*	*Số mục từ*	*Chiếm tỷ lệ*
10	h	12	6.1%
11	i	9	100.0%
12	j	8	100.0%
13	k	55	8.6%
14	l	46	30.7%
15	m	43	32.6%
16	n	30	32.6%
17	o	10	100.0%
18	ô	7	100.0%
19	ơ	25	35.7%
20	p	50	11.8%
21	r	7	8.1%
22	t	12	2.5%
23	u	8	47.1%
24	v	9	19.6%
25	x	10	8.7%
26	z	72	64.3%

KẾT LUẬN

Là một dân tộc có nguồn gốc bản địa, người Cơ Tu đã tạo nên một nền văn hóa mang đậm bản sắc, đó là văn hóa làng-văn hóa cộng đồng và văn hóa dân gian lành mạnh, trong sáng. Nét độc đáo đậm đà bản sắc văn hóa của người C'tu được biểu hiện rất rõ với những giá trị văn hóa vật thể và phi vật thể như: Gươl, làng truyền thống, các loại nhạc cụ; lễ hội mừng lúa mới, chữ Cơ Tu; các trang phục; nghề dệt thổ cẩm, nghề đan lát, nghề rèn được bảo tồn và phát triển; những kinh nghiệm về mùa vụ, ẩm thực truyền thống; những sáng tạo trong văn học truyền khẩu dân gian như: truyện cổ tích, ca dao, dân ca; nghệ thuật biểu diễn các loại nhạc cụ, nói lý, hát lý, hát giao duyên, múa tung tung da dã; nghệ thuật điêu khắc, chạm trổ hoa văn...

Trong thời gian qua, Đảng và Nhà nước ta đã có những đầu tư, sự quan tâm đáng kể đến việc khôi phục, bảo tồn văn hóa, tiếng nói của các dân tộc thiểu số nói chung, trong đó có tiếng dân tộc Cơ Tu.

Theo đó, các công trình nghiên cứu về phát triển kinh tế xã hội, bảo tồn văn hóa dân tộc Cơ Tu được triển khai hầu hết ở những vùng có đồng bào dân tộc Cơ Tu sinh sống, sách viết về văn hoá Cơ Tu ngày càng nhiều. Đặc biệt, những người con của dân tộc Cơ Tu đã cho ra đời các cuốn sách nhằm bảo tồn các giá trị văn hoá, đồng thời giới thiệu hình ảnh, văn hoá, con người C'tu đến với công chúng và truyền dạy cho các thế hệ sau về văn hoá của dân tộc mình. Có thể kể đến các cuốn sách đó là: *"Tiếng thông dụng C'tu-Kinh và văn hoá Làng C'tu", "Văn hoá người C'tu"* của tác giả Bh'riu Liếc; *"Người C'tu ở Việt Nam"* của tác giả Trần Tấn Vịnh; *"Vóc dáng Tây Giang"*

của nhiều tác giả; *"Từ điển C'tu-Việt, Việt C'tu"* của Viện Ngôn ngữ học; tổ chức giảng dạy tiếng Cơ Tu cán bộ, công chức, con em đồng bào dân tộc đang làm việc, sinh sống tại đây; phát thanh - truyền hình tiếng Cơ Tu… phù hợp với Nghị quyết Trung ương 5 khóa VIII về "Xây dựng và phát triển nền văn hóa Việt Nam tiên tiến, đậm đà bản sắc dân tộc", có đề ra nhiệm vụ cụ thể để bảo tồn, phát huy và phát triển văn hóa các dân tộc thiểu số Việt Nam.

Trong Khuôn khổ của luận văn, tôi đã tìm hiểu, nghiên cứu và xây dựng kho dữ liệu song ngữ Việt – Cơ Tu phục vụ tra cứu văn hóa dân tộc Cơ Tu. Kho dữ liệu song ngữ Việt – Cơ Tu có ý nghĩa thiết thực với đồng bào dân tộc người Cơ Tu, những người có nhu cầu tìm hiểu, nghiên cứu về chữ viết, văn hóa và con người Cơ Tu; tạo điều kiện về tra cứu, dịch thuật các văn bản quan trọng từ tiếng Cơ Tu sang tiếng Việt và ngược lại, tạo tiền đề để xây dựng ứng dụng từ điển đa ngữ.

Hạn chế

Hiện nay, số người am hiểu thật sự về chữ viết Cơ Tu còn hạn chế nên việc nhờ các chuyên gia cập nhật dữ liệu vào phần mềm rất khó khăn.

Phần từ vựng trong cơ sở dữ liệu còn thiếu nhiều, chưa đầy đủ.

Chưa tìm hiểu sâu về các phương pháp cập nhật tự động, dữ liệu tiếng Cơ Tu chủ yếu là trên giấy nên chương trình mới cập nhật kho dữ liệu bằng phương pháp thủ công, chưa cập nhật kho dữ liệu song ngữ bằng các phương pháp cập nhật tự động.

Hướng phát triển

Từ những phân tích và hạn chế nêu trên, cần tiếp tục hoàn thiện cơ sở dữ liệu song ngữ Việt – Cơ Tu bằng cách nghiên cứu, bổ

sung đầy đủ các mục từ, nghĩa, hình ảnh, các thành ngữ... để kho dữ liệu được phong phú hơn.

Tiếp tục nghiên cứu, xây dựng và bổ sung các chức năng tra cứu khác từ chương trình như: Mục Hỏi – đáp tự động,…

Tìm hiểu sâu hơn về cấu trúc ngữ pháp của tiếng Cơ Tu để có thể dịch một câu hay một đoạn văn, một văn bản từ tiếng Việt ra tiếng Cơ Tu.

Tiếp tục nghiên cứu các phương pháp cập nhật tự động để xây dựng website giới thiệu con người, văn hóa, đời sống, cũng như phong tục tập quán của người Cơ Tu. Thông qua trang web này, ta có thể tra từ, thêm từ, xóa hoặc chỉnh sửa từ vào kho ngữ vựng, hoặc có thể cập nhật tự động vào kho dữ liệu từ các nguồn văn bản, các bài báo, … bằng tiếng Cơ Tu trên mạng.

Tiếp tục nghiên cứu, ứng dụng xây dựng từ điển đa ngữ./.