BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

LÊ THỊ ANH ĐÀO

NGHIÊN CỨU XÂY DỰNG KHO NGỮ VỰNG SONG NGỮ VIỆT - KHMER

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.VÕ TRUNG HÙNG

Phản biện 1: TS. HOÀNG THỊ THANH HÀ

Phản biện 2: GS.TS NGUYỄN THANH THỦY

Luận văn đã được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ Kỹ thuật họp tại Đại học Đà Nẵng vào ngày 16 tháng 11 năm 2013

Có thể tìm hiểu luận văn tại:

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

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Ngày nay cùng với sự bùng nổ thông tin trên Internet mà trong đó văn bản là một trong những dạng chủ yếu thì nhu cầu xử lý ngôn ngữ tự nhiên trên máy tính là rất lớn. Làm thế nào để máy tính có thể hiểu được ngôn ngữ của con người vẫn là một trong những câu hỏi thách thức các nhà khoa học trong suốt lịch sử nửa thế kỷ của ngành trí tuệ nhân tạo.

Những năm gần đây, với sự tiến bộ về năng lực tính toán và khả năng lưu trữ của máy tính, các tiếp cận mới về xử lý ngôn ngữ tự nhiên đã thu được những thành công đáng khích lệ, đặc biệt là cách tiếp cận sử dụng phương pháp thống kê trên kho ngữ liệu lớn.

Trong xử lý ngôn ngữ tự nhiên, kho ngữ liệu là một nguồn tài nguyên quan trọng. Một mặt nó được dùng để huấn luyện các mô hình phân tích ngôn ngữ như tách câu, tách từ, gán nhãn từ loại, phân tích cú pháp. Mặt khác, nó còn được dùng để kiểm chứng độ tin cậy của các mô hình ngôn ngữ đó.Đồng thời nó hỗ trợ cho việc phát triển các ứng dụng như dịch máy thống kê, xây dựng từ điển song ngữ, tìm kiếm đa ngôn ngữ…

Xử lý ngôn ngữ tự nhiên là xử lý ngôn ngữ nói và ngôn ngữ viết của con người nên nó mang nét đặc thù riêng cho mỗi ngôn ngữ, mỗi quốc gia. Việt Nam với 54 dân tộc anh em, mỗi dân tộc có những đặc trưng văn hóa khác nhau, ngôn ngữ giao tiếp khác nhau nhưng hiện nay vẫn chưa có nhiều kho ngữ liệu đặc biệt là các kho ngữ liệu song ngữ và đa ngữ để hỗ trợ phát triển các hệ thống xử lý ngôn ngữ tự nhiên, phục

vụ xử lý tiếng Việt. Do đó, gây khó khăn trong việc giao lưu học tập, trao đổi văn hóa, phát triển giữa các dân tộc.

Dân tộc Khmer sống tập trung tại các tỉnh Sóc Trăng, Vĩnh Long, Trà Vinh,... là một tộc người trong cộng đồng các dân tộc Việt Nam. Người Khmer có tiếng nói và chữ viết riêng.

Sự cộng cư lâu đời và hòa hợp giữa hai dân tộc Việt và Khmer dẫn đến tình trạng là có nhiều người sử dụng cả hai loại ngôn ngữ này. Bên cạnh đó, nhu cầu học tiếng Việt của người Khmer hay học tiếng Khmer của người Việt ngày càng cao. Tuy nhiên, các công cụ hỗ trợ học tiếng Khmer hay các giáo trình học tập, cũng như các tài liệu tham khảo học tập tiếng Khmer rất ít. Do đó nhu cầu học tập, giảng dạy tiếng Khmer cho học sinh, sinh viên, giáo viên và đội ngũ cán bộ ngày càng trở nên bức thiết.

Để giải quyết những vấn đề nêu trên, tôi đề xuất đề tài: "Nghiên cứu xây dựng kho ngữ vựng song ngữ Việt – Khmer" để góp phần vào việc bảo tồn, quảng bá chữ viết cũng như một số đặc điểm về văn hóa, tín ngưỡng của người Khmer, tạo điều kiện thuận lợi cho việc giao lưu, học tập, trao đổi văn hóa giữa hai dân tộc.

2. Mục tiêu nghiên cứu

Mục tiêu của đề tài: nghiên cứu xây dựng một kho ngữ vựng song ngữ Việt – Khmer nhằm phục vụ việc giao lưu, học tập, trao đổi văn hóa của hai dân tộc.

3. Đối tượng và phạm vi nghiên cứu

- Đối tượng nghiên cứu:

+ Ngôn ngữ dân tộc Khmer;

+ Các giải pháp cập nhật CSDL;

+ Kho ngữ vựng;

+ Các công cụ xây dựng CSDL;

+ Các mô hình triển khai hệ thống.

- Phạm vi nghiên cứu:xây dựng kho ngữ vựng song ngữViệt – Khmer.

4. Phương pháp nghiên cứu

- Phương pháp nghiên cứu tài liệu:

+ Các tài liệu xuất bản, tài liệu phát thanh truyền hình tiếng Khmer;

+Các trang tin điện tử dân tôc Khmer;

+ Các luận văn và bài báo khoa học liên quan.

- Phương pháp thực nghiệm: sử dụng các công cụ thiết kế xây dựng kho ngữ liệu; thực nghiệm cập nhật, hiệu chỉnh kho ngữ liệu.

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

- Về khoa học: góp phần tạo ra một hướng nghiên cứu mới đi xây dựng các CSDL song ngữ, đặt biệt cho tiếng dân tộc ít người ở Việt Nam.

- Về thực tiễn: kho ngữ vựng song ngữ Việt – Khmer được tạo ra từ đề tài tạo tiền đề cho những nghiên cứu sau này.

6. Bố cục của báo cáo

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

Chương 1. Nghiên cứu tổng quan.

Chương 2. Phân tích thiết kế hệ thống.

Chương 3. Triển khai xây dựng.

CHƯƠNG 1
NGHIÊN CỨU TỔNG QUAN

1.1 TỔNG QUAN VỀ NGÔN NGỮ KHMER

1.1.1 Giới thiệu chung

a. Dân tộc Khmer

Đồng bào Khmer Nam Bộ là một bộ phận không thể tách rời trong cộng đồng 54 dân tộc Việt Nam. Dân tộc Khmer có 1,3 triệu dân, tập trung ở các tỉnh, thành phố thuộc khu vực Đồng bằng sông Cửu Long như: Sóc Trăng, Trà Vinh, Kiên Giang, An Giang, Bạc Liêu, Cà Mau, Vĩnh Long, Hậu Giang, Cần Thơ, Thành Phố Hồ Chí Minh và miền Đông Nam Bộ [1].

b. Tôn giáo, tín ngưỡng

Hình 1.1: Chùa Vàm Ray ở tỉnh Trà Vinh

Đa số, người Khmer theo Phật giáo Nam Tông (Theravada). Hiện nay có khoảng gần 500 chùa Khmer ở ĐBSCL đóng vai trò quan trọng trong đời sống văn hoá tinh thần của người Khmer. Chùa Khmer là trung tâm của cộng đồng Khmer ở các địa phương. Đàn ông Khmer đến tuổi thì thường đi tu một thời gian để tu thân và cũng là để trả hiếu cho cha mẹ. Đi tu làm cho đàn ông Khmer không những có giá trị

hơn, có đạo đức hơn, mà còn để học được chữ viết Khmer và tiếng Phạn[1].

c. Văn hóa Khmer

Người Khmer Nam Bộ có nhiều giá trị vật chất lẫn tinh thần, trang phục truyền thống của người Khmer cũng dễ phân biệt với các dân tộc khác và được sử dụng trong các dịp lễ Tết, đám cưới, ... Nhà ở hầu hết người Khmer làm nhà đất, mái lá rất đơn giản…

d. Văn học

Kho tàng văn học dân gian trong đồng bào Khmer Nam Bộ cũng phong phú, đa dạng ở cả nội dung lẫn hình thức...

e. Ngôn ngữ nói

Tiếng Khmer, còn gọi là tiếng Cambodia, tiếng Cam pu chia, là ngôn ngữ chính thức của Vương quốc Campuchia và người Khmer…

f. Ngôn ngữ viết

Chữ Khmer thuộc ngữ hệ Môn – Khmer, bộ chữ cái Khmer có 33 phụ âm và 40 nguyên âm. Các phụ âm được chia làm 2 loại : loại giọng or có 15 con chữ và loại giọng o có 18 con chữ. Nguyên âm gồm có hai loại: nguyên âm thường (là nguyên âm phải ráp với phụ âm mới có nghĩa, gồm có 25 con chữ và khi phát âm thì mỗi con chữ có 2 giọng âm khác nhau. Tức khi ráp vần với phụ âm có giọng or thì đọc khác, khi ráp vần với phụ âm có giọng o thì đọc khác) và nguyên âm độc lập (là nguyên âm không cần ráp vần với phụ âm cũng có nghĩa, gồm có 15 con chữ).

Văn bản tiếng Khmer được cấu tạo bởi các từ, mỗi từ được cấu tạo bởi các phụ âm và nguyên âm. Các từ được viết từ trái sang phải

và cách nhau bởi ký tự trống. Các dấu chức năng: ! "( ) [ ] { } : ; ? , cũng được sử dụng tương tự như trong các văn bản tiếng Việt, riêng dấu chấm câu tiếng Khmer sử dụng ký tự (។) [2].

g. Bộ chữ cái Khmer

Bảng chữ cái phụ âm tiếng Khmer: Gồm có 33 chữ và 32 chân, được chia làm hai giọng or/ɔ/ và o/o/.

ក[kɔ]	ខ[k'ɔ]	គ[ko]	ឃ[k'o]	ង[ŋo]	ច[cɔ]
ឆ[c'ɔ]	ជ[co]	ឈ[c'o]	ញ[ɲo]	ដ[dɔ]	ឋ[d'ɔ]
ឌ[do]	ឍ[d'o]	ណ[nɔ]	ត[tɔ]	ថ[t'ɔ]	ទ[to]
ធ[t'o]	ន[no]	ប[bɔ]	ផ[p'ɔ]	ព[po]	ភ[p'o]
ម[mo]	យ[yo]	រ[ro]	ល[lo]	វ[wo]	ស[sɔ]
ហ[hɔ]	ឡ[lɔ]	អ [ɔ]			

Hình 1.4: 33 phụ âm tiếng Khmer

Bảng chữ cái nguyên âm tiếng Khmer: Gồm có 25 nguyên âm thông thường và 15 nguyên âm độc lập:

- Nguyên âm thông thường:

អា	អាំ	អិ	អី	អុ
អូ	អឹ	អឺ	អេ	អែ
អៃ	អោ	អៅ	អួ	អៀ
អឿ	អុំ	អំ	អុះ	អះ
អេះ	អើ	អោះ	អិះ	អឺះ

Hình 1.7: 25 nguyên âm thông thường tiếng Khmer

- Nguyên âm độc lập:

ឥ	ឣ	ឤ	ឯ	ឬ
ឭ	ឮ	ឰ	ឱ	ឳ
឴	឵	ឲ	ា	ិ

Hình 1.8: 15 nguyên âm độc lập tiếng Khmer

h. Cách ráp vần tiếng Khmer

Ráp phụ âm với phụ âm

- ប(bo) ráp với ង(ngô) thành chữ បង (boong) Anh.

Ráp phụ âm với nguyên âm

- Chữ ក(co) đặt trước◌(a) thành កា (ca) cái ca.

Ráp 1 phụ âm có nguyên âm đi kèm với 1 phụ âm khác

- រ(rô) ghép với◌ៀ (ia)thành រៀ (ria), rồi lấy រៀghép với ន(nô)thành រៀន (riên) có nghĩa là học.

Ráp phụ âm có chân đi kèm với 1 nguyên âm

- Chữ ស(so)ghép với chân្គ(cô)thành ស្គ (sờ cô)lấy ស្គghép với ម(mô) thành ស្គម(sờ côm) Gầy.

i. Ngữ pháp Khmer

Ngữ pháp tiếng Khmer gần giống như ngữ pháp tiếng Việt Nam ví dụ:

Người Việt Nam nói:	*Sáng nay*	*tôi*	*đi*	*thành phố*	*Hồ Chí Minh*
Người Khmer nói	*Pô rức nís*	*kho nhum*	*tâu*	*ti co rông*	*Ho Chi Minh*

j. Cấu trúc câu và trật tự từ

Về cơ bản cấu trúc câu và trật tự từ của tiếng Khmer gần như tương đồng với tiếng Việt.

k. Thanh điệu

Ngôn ngữ Khmer không có thanh điệu, tức là tất cả các âm thanh đều là thanh bằng (tức là không có dấu).

1.1.2 Xử lý tiếng Khmer trên máy tính

a. Mã hóa chữ Khmer

Vì các ký tự tiếng Khmer không phải là ký tự La-tinh nên không có sẵn trong bảng mã ASCII. Hiện nay, để biểu diễn các ký tự

của tiếng Khmer người ta đã có thể sử dụng bảng mã Unicode. Việc xây dựng bộ mã và phông chữ Unicode tổ hợp cho tiếng Khmer nằm trong dự án phần mềm tiếng Khmer (KhmerOS) của tổ chức Open Forum of Cambodia.

b. Bộ gõ

Bảng 1.1: Cách tổ hợp các phụ âm tiếng Khmer

Chữ Khmer	Cách đánh máy	Chữ Khmer	Cách đánh máy
ក, គ	Được tổ hợp trong phím K	ប, ព	Được tổ hợp trong phím B
ខ, ឃ	Được tổ hợp trong phím X	ផ, ភ	Được tổ hợp trong phím P
ង	Được tổ hợp trong phím G	ម	Được tổ hợp trong phím M
ច, ជ	Được tổ hợp trong phím C	យ	Được tổ hợp trong phím Y
ឆ, ឈ	Được tổ hợp trong phím Q	រ	Được tổ hợp trong phím R
ញ	Được tổ hợp trong phím J*	ល, ឡ	Được tổ hợp trong phím L
ដ, ឌ	Được tổ hợp trong phím D	វ	Được tổ hợp trong phím V
ឋ, ឍ	Được tổ hợp trong phím Z	ស	Được tổ hợp trong phím S
ន, ណ	Được tổ hợp trong phím N	ហ	Được tổ hợp trong phím H

ត, ទ	Được tổ hợp trong phím T	អ	Được tổ hợp trong phím G*
ថ, ធ	Được tổ hợp trong phím F		

Bảng 1.2: Cách tổ hợp các nguyên âm thông thường tiếng Khmer

Chữ Khmer	Cách đánh máy	Chữ Khmer	Cách đánh máy
អា, អាំ	Được tổ hợp trong phím A	អៀ, អឿ	Được tổ hợp trong phím [
អិ, អី	Được tổ hợp trong phím I	អំ	Được tổ hợp trong phím M**
អុ, អូ	Được tổ hợp trong phím U	អុំ, អុះ	Được tổ hợp trong phím <
អឹ, អឺ	Được tổ hợp trong phím W	អះ	Được tổ hợp trong phím H**
អេ, អែ	Được tổ hợp trong phím E	អេះ	Được tổ hợp trong phím V**
អៃ	Được tổ hợp trong phím **	អើ, អោះ	Được tổ hợp trong phím ;
អោ, អៅ	Được tổ hợp trong phím S	អិះ	Là sự kết hợp của nguyên âm ិ và nguyên âm ះ
អួ	Được tổ hợp trong phím Y**	អឹះ	Là sự kết hợp của nguyên âm ឹ và nguyên âm ះ

Bảng 1.3: Cách tổ hợp các nguyên âm độc lập tiếng Khmer

Chữ Khmer	Cách đánh máy	Chữ Khmer	Cách đánh máy
អ	Tương đương với phụ âm អ là Shift G	ឫ	Được tổ hợp trong phím \
អា	Tương đương với phụ âm អ với nguyên âm ា là Shift G và A	ឬ	Được tổ hợp trong phím Shift \
		ឧ	Được tổ hợp trong phím =
		ឱ	Được tổ hợp trong phím]
ឥ	Được tổ hợp trong phím -	ឳ	Được tổ hợp trong phím Shift]
ឯ	Được tổ hợp trong phím Alt W	឴	Được tổ hợp trong phím Alt [(*)
ឫ	Được tổ hợp trong phím Shift R	឵	Được tổ hợp trong phím Alt]
ឬ	Được tổ hợp trong phím Alt R	ឮ	Được tổ hợp trong phím Alt P

1.2 CƠ SỞ DỮ LIỆU TỪ VỰNG, KHO NGỮ LIỆU, TỪ ĐIỂN

1.2.1 Cơ sở dữ liệu từ vựng

a. Khái niệm

Cơ sở dữ liệu được hiểu theo các định nghĩa kiểu kỹ thuật thì nó là một tập hợp thông tin có cấu trúc…

b. Cơ sở dữ liệu từ vựng đa ngữ

Một cơ sở dữ liệu được gọi là đa ngữ nếu chúng có thể làm việc trên CSDL đó với hai hay nhiều ngôn ngữ khác nhau.

1.2.2 Kho ngữ liệu

a. Một số khái niệm

b. Tổng quan về XML

c. Thu thập dữ liệu

Nguồn từ điển

Trong mỗi từ điển, ở mỗi mục từ, thường chứa các ví dụ hướng dẫn sử dụng từ đó. Hầu hết các ví dụ này đều là các câu thông thường.

Nguồn Internet

Đây là nguồn dữ liệu khổng lồ, nguồn ngữ liệu này có lợi thế là chúng đã tồn tại sẵn dưới dạng điện tử (nên không phải nhập liệu lại bằng tay).

Nguồn sách

Bao gồm các sách dạy tiếng Khmer, các mẫu câu đàm thoại Việt - Khmer, tự điển Việt - Khmer…

1.2.3 Phương pháp tách từ, tách câu

Để giải quyết những bài toán liên quan đến xử lý ngôn ngữ tự nhiên, xây dựng kho ngữ vựng thì các bài toán cơ bản nhất là bài toán tách từ, tách câu văn bản.

a. Bài toán tách từ

b. Bài toán tách câu

1.2.4 Một số giải thuật trong xử lý ngôn ngữ tự nhiên

a. Thuật toán liên kết từ

b. Thuật toán tách câu

1.2.5 Từ điển

a. Khái niệm

Từ điển là tập hợp từ (đôi khi cả hình vị hoặc cụm từ) sắp xếp theo trật tự nhất định, được dùng làm như cẩm nang giải thích nghĩa của các đơn vị miêu tả, cung cấp các thông tin khác nhau về các đơn vị đó hay dịch sang ngôn ngữ khác, hoặc cung cấp các thông tin về sự vật được các đơn vị miêu tả đó biểu đạt [5].

b. Phương pháp xây dựng từ điển

c. Một số từ điển Việt – Khmer

CHƯƠNG 2
PHÂN TÍCH THIẾT KẾ HỆ THỐNG

2.1 MÔ TẢ ỨNG DỤNG

2.1.1 Giới thiệu

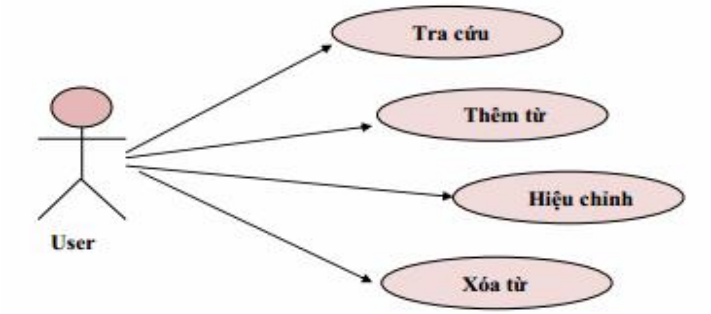
Kho ngữ vựng được xây dựng là tập hợp các cặp từ, cặp câu Việt – Khmer được thu thập từ nhiều nguồn dữ liệu khác nhau. Kho ngữ vựng này có thể giúp người sử dụng tra cứu các cặp từ Việt – Khmer, phiên âm tiếng Khmer, phát âm tiếng Khmer và các cặp câu ví dụ Việt – Khmer tương ứng.

2.1.2 Yêu cầu hệ thống

Bài toán đặt ra những yêu cầu xây dựng một kho ngữ vựng song ngữ Việt – Khmer có chức năng hỗ trợ tìm kiếm, tra cứu từ giữa tiếng Việt và tiếng Khmer.

2.2 PHÂN TÍCH, THIẾT KẾ HỆ THỐNG

2.2.1 Biểu đồ User – case



Hình 2.3: Biểu đồ ca sử dụng

2.2.2 Đặc tả User – case

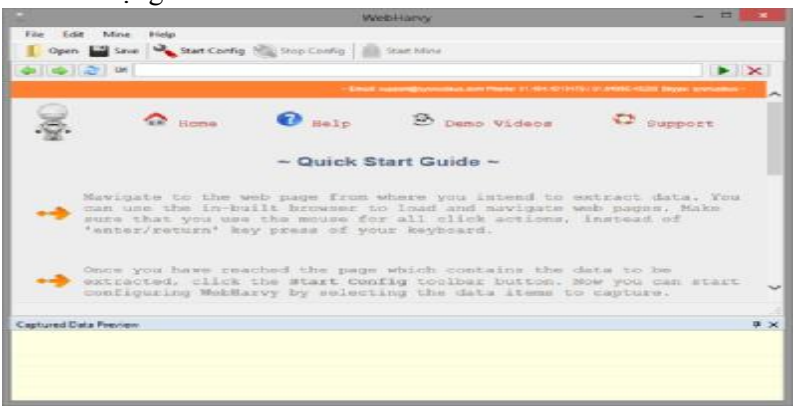
2.2.3 Biểu đồ trình tự

2.2.4 Biểu đồ hoạt động

2.3 GIẢI PHÁP XỬ LÝ DỮ LIỆU

2.3.1 Kỹ thuật trích lọc dữ liệu tự động bằng Web Scraping

Web Scraping là phần mềm khai thác dữ liệu Web, là một kỹ thuật được sử dụng để trích xuất một lượng lớn dữ liệu từ các trang web trên mạng.



Hình 2.12: Giao diện phần mềm trích dữ liệu WebHarvy

2.3.2 Kỹ thuật trích lọc dữ liệu file HTML

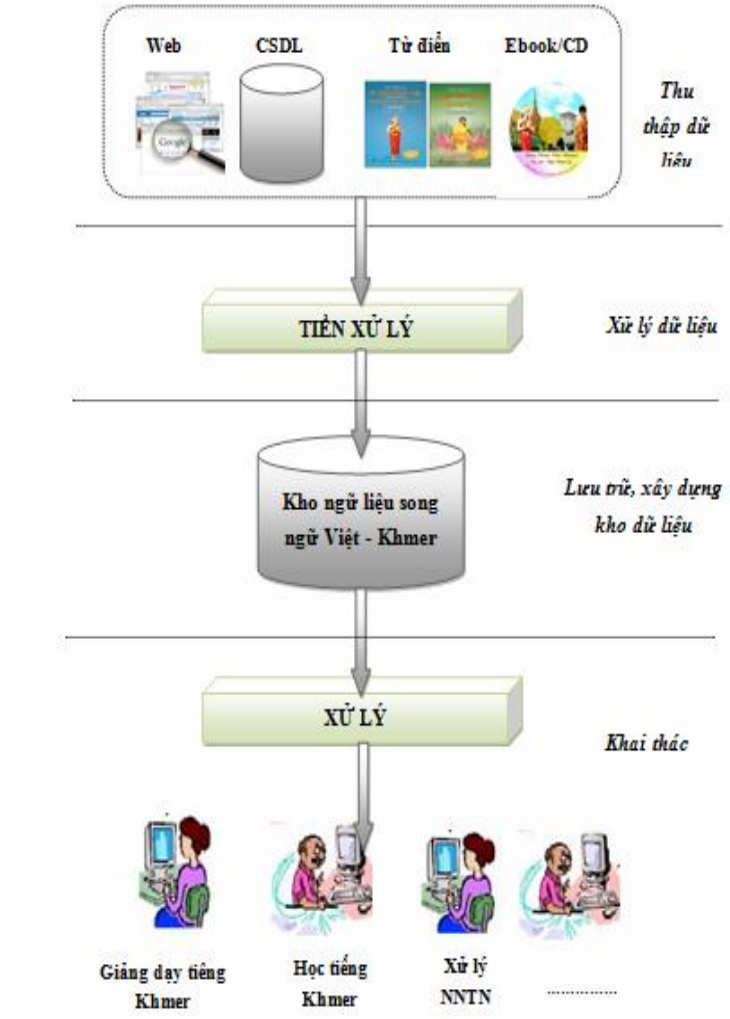
Một trang web sau khi được tải về để làm nguồn dữ liệu cập nhật kho, ta cần trích lấy nội dung cần thiết và phải làm sạch, bao gồm:

- Đọc nội dung văn bản đưa về định dạng chuỗi ký tự .
- Hủy bỏ dòng trắng không được hiển thị trên HTML.
- Hủy bỏ các khoảng trắng tab.
- Hủy bỏ các ký tự trắng liên tiếp trong HTML.
- Hủy bỏ thẻ HEAD.
- Hủy bỏ tất cả JavaScript.
- Thay thế các ký tự đặc biệt như &, <, >, "…
- Kiểm tra và thay thế ngắt dòng (
) hoặc khoản (<p>)

- Loại bỏ tất cả các thẻ HTML.

2.4 GIẢI PHÁP XÂY DỰNG

2.4.1 Mô hình tổng quát của hệ thống



Hình 2.1: Mô hình tổng quát của hệ thống

2.4.2 Giải pháp xây dựng kho ngữ vựng

a. Thu thập dữ liệu

Đầu tiên ta phải chuẩn bị dữ liệu trên nhiều ngôn ngữ khác nhau. Những dữ liệu đa ngữ này, ta có thể có được bằng cách sử dụng các tài liệu gốc có sẵn dưới nhiều ngôn ngữ khác nhau hoặc có thể dịch ra các ngôn ngữ khác từ một dữ liệu gốc ban đầu bằng các phần mềm dịch tự động trên mạng.

b. Xử lý dữ liệu

Dữ liệu thu thập về cần được chuẩn hóa trước khi đưa vào kho, có thể nhập trực tiếp dữ liệu, xử lý thủ công hoặc tự động.

Việc chuẩn hóa dữ liệu là việc chuyển đổi định dạng dữ liệu thành định dạng tương thích với mục đích của hệ thống. Nghĩa là, chúng ta cần phải lựa chọn các bộ gõ, hệ thống mã hóa và các hệ thống phông chữ phù hợp cho từng ngôn ngữ cần thể hiện. Đặc biệt cần lưu ý là nên sử dụng hệ thống mã hóa Unicode.

c. Lưu trữ, xây dựng kho dữ liệu

Chúng ta cần lựa chọn công cụ để lưu trữ dữ liệu đa ngữ ví dụ như XML, các hệ quản trị cơ sở dữ liệu như Access, Oracle… Đặc biệt, hiện nay thì XML được xem là một chuẩn rất tốt dành cho các dữ liệu đa ngữ.

d. Khai thác dữ liệu

Khai thác các CSDL từ vựng đa ngữ, tùy theo mục đích mà chúng ta có thể khai thác CSDL từ vựng đa ngữ theo các hướng và bằng nhiều công cụ khai thác dữ liệu khác nhau. Ở đây tôi xây dựng công cụ tra từ để đọc và truy xuất dữ liệu từ các file mô tả cơ sở dữ liệu đã được lưu trong kho.

CHƯƠNG 3
TRIỂN KHAI XÂY DỰNG

3.1 CÔNG CỤ HỖ TRỢ PHÁT TRIỂN HỆ THỐNG

3.1.1 Visual Studio.Net

3.1.2 SQL Server 2008

3.1.3 Ngôn ngữ lập trình C#.Net

3.2 THIẾT KẾ CƠ SỞ DỮ LIỆU

3.2.1 Đặc tả chi tiết các bảng

Bảng 3.1: Cấu trúc chi tiết bảng từ vựng tiếng Khmer

Tên trường	Kiểu dữ liệu	Diễn giải	Ghi chú
MaTuKhmerID	nchar(10)	Mã từ Khmer	Khóa chính
MaTuVietID	nchar(10)	Mã từ tiếng Việt	
TuKhmer	nvarchar(MAX)	Từ tiếng Khmer	
Phienam	nvarchar(MAX)	Phiên âm tiếng Khmer	
Phatam	nvarchar(MAX)	Phát âm tiếng Khmer	

Bảng 3.2: Cấu trúc chi tiết bảng câu tiếng Việt

Tên trường	Kiểu dữ liệu	Diễn giải	Ghi chú
MaCauTVID	nchar(10)	Mã câu tiếng Việt	Khóa chính
MaTuVietID	nchar(10)	Mã từ tiếng Việt	
CauTV	nvarchar(MAX)	Câu ví dụ tiếng Việt	

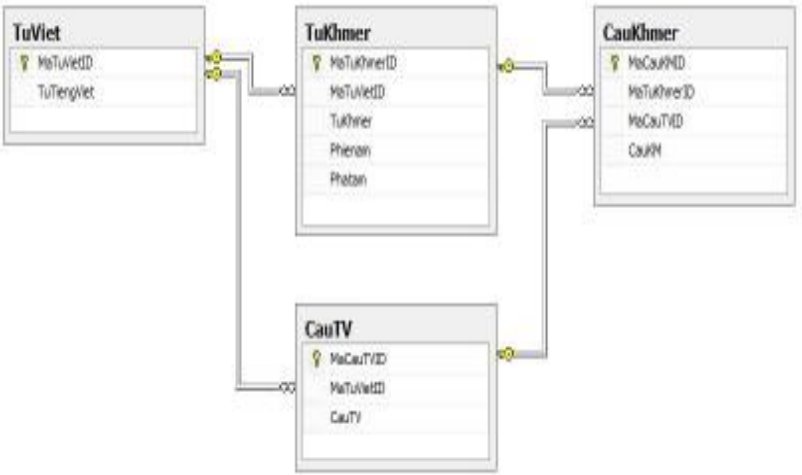
Bảng 3.3: Cấu trúc chi tiết bảng câu tiếng Khmer

Tên trường	Kiểu dữ liệu	Diễn giải	Ghi chú
MaCauKMID	nchar(10)	Mã câu tiếng Khmer	Khóa chính
MaTuKhmerID	nchar(10)	Mã từ tiếng Khmer	
MaCauTVID	nchar(10)	Mã câu tiếng Việt	
CauKM	nvarchar(MAX)	Câu ví dụ tiếng Khmer	

Bảng 3.4: Cấu trúc chi tiết bảng từ vựng tiếng Việt

Tên trường	Kiểu dữ liệu	Diễn giải	Ghi chú
MaTuVietID	nchar(10)	Mã từ tiếng Việt	Khóa chính
TuTiengViet	nvarchar(MAX)	Từ tiếng Việt	

3.2.2 Mô hình dữ liệu quan hệ



Hình 3.1: Mô hình dữ liệu quan hệ

3.3 CÁC BƯỚC TRIỂN KHAI

3.3.1 Thu thập dữ liệu

a. Nguồn dữ liệu

b. Cách trích dữ liệu

- Đối với cơ sở dữ liệu được cập nhật thủ công chúng ta xây dựng công cụ cập nhật:

Hình 3.2: Giao diện cập nhật dữ liệu vào kho

- Đối với cơ sở dữ liệu trích tự động từ những trang web chúng ta sử dụng công cụ WebHarvy để rút trích dữ liệu.

3.3.2 Xử lý dữ liệu

- Dựa vào ký hiệu ngắt câu của tiếng Khmer (។) và ký hiệu ngắt câu trong tiếng Việt (.), ta tiến hành tách trích từng cặp câu tương ứng.

- Hủy bỏ dòng trắng, khoảng trắng tab, các ký tự trắng liên tiếp trong HTML, các ký tự đặc biệt như &, <, >, "...và những phần không cần thiết

- Chuẩn hóa toàn bộ dữ liệu theo một chuẩn thống nhất. Trong phần này tôi chuyển đổi tất cả dữ liệu về phông chữ `Time new romand` thuộc bảng mã Unicode.

- Đối với các tập tin định dạng PDF tôi sử dụng phần mềm chuyển đổi sang định dạng .Docx để thuận tiện cho công việc tách lấy dữ liệu.

- Đa số dữ liệu lấy về là các cặp câu, cặp từ English – Khmer nên để trích lấy nguồn ngữ liệu này vào kho, tôi đã thông qua bộ máy dịch thuật tự động Google là một công cụ dịch thuật trực tuyến miễn phí được Google cung cấp để có thể dịch nhanh văn bản và các trang web,… với nhiều ngôn ngữ khác nhau.

Đồng thời, để đánh giá độ chính xác của các bản dịch này tôi cũng đã dùng một số trang dịch tự động khác như vdict.com/#, stars21.com/translator/, dict.vntranslate.net/,…, để kiểm chứng, so sánh độ chính xác của các kết quả dịch của nhau và từ đó rút ra, lựa chọn các bản dịch có độ chính xác cao hơn để đưa vào kho ngữ vựng.

3.3.3 Xây dựng kho ngữ vựng Việt – Khmer

- Dữ liệu được lưu trữ bằng định dạng Excel trước khi đưa vào kho với cấu trúc mô tả như sau:

Bảng 3.5: Sheet mô tả thông tin của từ tiếng Khmer

MỤC	NỘI DUNG
MaTuKhmerID	Mã của từ Khmer
MaTuVietID	Mã của từ tiếng Việt
TuKhmer	Từ tiếng Khmer
Phienam	Phiên âm tiếng Khmer
Phatam	Phát âm tiếng Khmer

Bảng 3.6: Sheet mô tả thông tin của từ tiếng Việt

MỤC	NỘI DUNG
MaTuVietID	Mã của từ tiếng Việt
TuTiengViet	Từ tiếng Việt

Bảng 3.7: Sheet mô tả thông tin của câu tiếng Việt

MỤC	NỘI DUNG
MaCauTVID	Mã câu tiếng Việt
MaTuVietID	Mã từ tiếng Việt
CauTV	Câu ví dụ TV

Bảng 3.8: Sheet mô tả thông tin của câu tiếng Khmer

MỤC	NỘI DUNG
MaCauKMID	Mã câu tiếng Khmer
MaTuKhmerID	Mã từ tiếng Khmer
MaCauTVID	Mã câu tiếng Việt
CauKM	Câu ví dụ tiếng Khmer

- Import dữ liệu từ tập tin mô tả tài liệu Excel vào cơ sở dữ liệu, với mỗi tập tin bằng định dạng Excel tương ứng là một bản ghi trong bảng dữ liệu, mỗi cột sẽ tương ứng với một trường trong bảng ghi đó.

Các bước thực hiện:

Bước 1: Thiết kế giao diện Import dữ liệu từ Excel sang SQL.

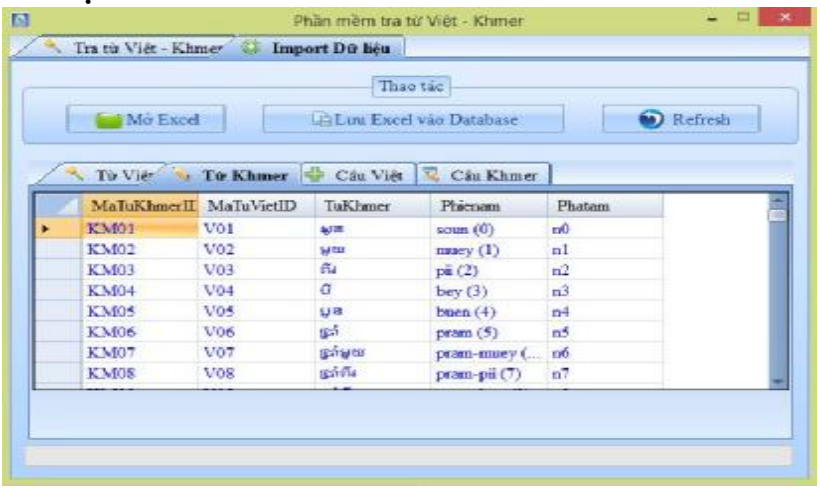
Bước 2: Viết code cho sự kiện Import dữ liệu.

3.3.4 Khai thác kho ngữ vựng song ngữ

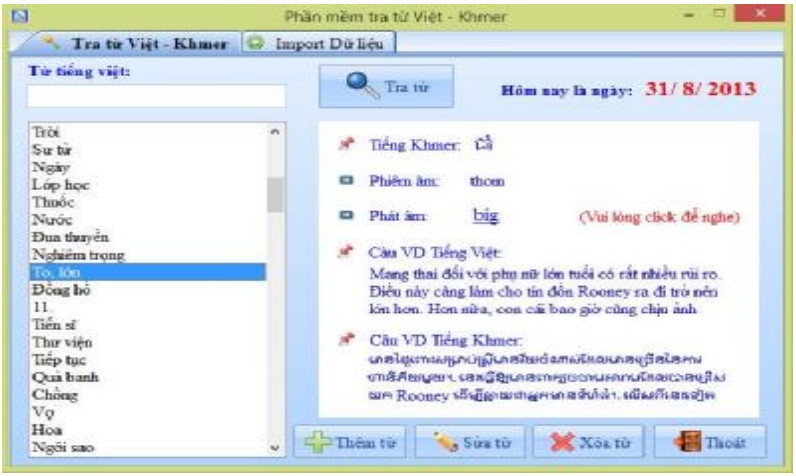
Để ứng dụng kho ngữ liệu song ngữ Việt – Khmer vào trong việc xây dựng từ điển, tôi thực hiện xây dựng một chương trình tra từ

để khai thác kho. Chương trình cho phép người sử dụng có thể tra từ, thêm từ, chỉnh sửa và xóa từ Việt – Khmer, đồng thời mỗi từ tra có các ví dụ, phiên âm và phát âm kèm theo.

3.4 MỘT SỐ DEMO CHƯƠNG TRÌNH



Hình 3.3: Giao diện Import dữ liệu từ Excel qua SQL



Hình 3.4: Giao diện màn tra từ

3.5 KẾT QUẢ ĐẠT ĐƯỢC

Việc triển khai xây dựng kho ngữ vựng song ngữ Việt – Khmer bước đầu đã ghi nhận được một số kết quả đạt được như sau:

Tìm hiểu hệ thống chữ viết tiếng Khmer, phương pháp trích lọc dữ liệu trên mạng, phương pháp xây dựng kho ngữ vựng áp dụng xây dựng kho ngữ vựng song ngữ Việt – Khmer.

Đã xây dựng được kho cơ sở dữ liệu từ vựng song ngữ với khoảng 2.000 từ thông dụng trong đời sống xã hội và đưa vào cơ sở dư liệu hơn 2.000 câu tiếng Khmer thông dụng.

Xây dựng công cụ tra từ vựng Việt – Khmer đáp ứng được nhu cầu học tập, giảng dạy của những người Việt muốn học tiếng Khmer và người Khmer muốn học tiếng Việt.

KẾT LUẬN

1. Kết quả đạt đươc

Về mặt khoa học:

Luận văn đã tiến hành nghiên cứu tìm hiểu về ngôn ngữ Khmer, các kiến thức về xử lý ngôn ngữ tự nhiên, kho ngữ liệu song ngữ, các vấn đề liên quan đến xử lý dữ liệu, các bước xây dựng kho ngữ vựng.

Về mặt thực tiễn

Luận văn đã nêu được giải pháp, kỹ thuật để xử lý dữ liệu và cập nhật kho ngữ liệu song ngữ Việt – Khmer.

Xây dựng thành công kho ngữ vựng song ngữ Việt – Khmer và công cụ tra từ vựng Việt – Khmer để khai thác kho tài liệu.

2. Về mặt hạn chế

Ngôn ngữ Khmer không được sử dụng và chia sẽ rộng rãi nên việc thu thập nguồn ngữ liệu gặp rất nhiều khó khăn. Vì thế số lượng ngữ vựng cập nhật trong kho chưa được nhiều.

Chất lượng các bản dịch của nguồn dữ liệu song ngữ chưa cao. Cơ sở dữ liệu sưu tập cho từng mục từ chưa đầy đủ về phần phát âm và ví dụ minh họa,…

3. Hướng phát triển

Tiếp tục sưu tập nguồn dữ liệu song ngữ Việt – Khmer cho kho ngữ vựng.