ĐỊNH HƯỚNG XỬ LÝ TIẾNG DÂN TỘC HRÊ VÀ DÂN TỘC CO. ỨNG DỤNG XÂY DỰNG KHO NGỮ VỰNG SONG NGỮ VIỆT-HRÊ, HRÊ-VIỆT, VIỆT-CO VÀ CO-VIỆT

ORIENTATING FOR THE PROCESSING OF ETHNIC LANGUAGE HRE AND ETHNIC LANGUAGE CO. APPLICATING TO BUILD BILINGUAL VOCABULARY DATABASE VIET-HRE, HRE-VIET, VIET-CO AND CO-VIET

Tiến sĩ Hoàng Thị Mỹ Lệ

Phó Trưởng khoa, Phụ trách Khoa Công nghệ số, Trường Đại học Sư phạm Kỹ thuật – Đại học Đà Nẵng

Tóm tắt - Hiện nay, nước ta có trên 20 ngôn ngữ có chữ viết trên tổng số 55 dân tộc. Vấn đề xử lý tiếng Việt đã được nghiên cứu từ năm 1990, đã có nhiều kết quả và vẫn đang được tiếp tục duy trì. Tuy nhiên, các kết quả nghiên cứu tiếng dân tộc thiểu số (DTTS) đạt được vẫn còn những mặt hạn chế như: chưa được liên kết, thiếu chia sẻ và không có tính kế thừa, chỉ tập trung vào các sản phẩm cho người dùng, ít nghiên cứu phát triển hạ tầng cơ sở như từ điển máy tính, kho ngữ liệu, đây là những thứ không thể thiếu trong xử lý ngôn ngữ tự nhiên. Những hạn chế này là do không có sự đồng thuận giữa các đơn vị nghiên cứu, giữa các nhà khoa học về xử lý ngôn ngữ tự nhiên. Cho đến nay, các vấn đề xử lý tiếng DTTS Hrê và tiếng DTTS Co ở Việt Nam cho đến nay vẫn chưa có các nhà khoa học quan tâm. Trong bối cảnh bùng nổ sử dụng internet, cùng với nhu cầu phát triển văn hoá và hội nhập của cộng đồng các DTTS Hrê và Co ở Việt Nam, thì lúc này nhu cầu xử lý tiếng DTTS Hrê và DTTS Hrê Co đặt ra càng bức thiết hơn bao giờ hết. Xuất phát từ những thực trạng về nghiên cứu xử lý tiếng Việt nói chung, tiếng Hrê và tiếng Co nói riêng, bài báo đề xuất định hướng qui trình xử lý tiếng dân tộc Hrê và dân tộc Co. Trên cơ sở đó, tập trung nghiên xây dựng các kho ngữ vựng song ngữ Hrê-Việt, Việt-Hrê, Việt-Co và Co-Việt làm hạ tầng cơ sở cho xử lý tiếng Hrê và tiếng Co.

Từ khóa - DTTS Hrê, DTTS Co, dân tộc thiểu số, kho ngữ vựng song ngữ, Việt-Hrê, Hrê-Việt, Việt-Co, Co-Việt.

Abstract - Currently, our country has over 20 written languages out of a total of 55 ethnic groups. The issue of Vietnamese language processing has been researched since 1990, has had many results and is still being maintained. However, the results of research on ethnic minority languages still have limitations such as: not being linked, lacking sharing and not inheriting, only focusing on products for ethnic minorities. used, little research is done to develop infrastructure such as computer dictionaries and corpuses, which are indispensable in natural language processing. These limitations are due to the

lack of consensus among research units and scientists on natural language processing. Up to now, the issues of dealing with the Hre ethnic minority language and the Co ethnic minority language in Vietnam have not received any attention from scientists. In the context of the explosion of internet use, along with the need for cultural development and integration of the Hre and Co ethnic minority communities in Vietnam, at this time the need to handle the Hre and Hre Co ethnic minority languages is even more urgent. more necessary than ever. Starting from the current situation of research on processing Vietnamese in general, Hre and Co languages in particular, the article proposes orientations for the processing of Hre and Co ethnic languages. On that basis, focus on building bilingual vocabulary Hre-Vietnamese, Vietnamese-Hre, Viet-Co and Co-Viet as the basic infrastructure for processing Hre and Co languages.

Keywords – Hre ethnic minorities, Co ethnic minorities, ethnic minorities, bilingual vocabulary database, Vietnamese-Hre, Hre-Vietnamese, Vietnamese - Co, Co- Vietnamese.

1. Đặt vấn đề

Hiện nay đã có một số ngôn ngữ, chữ viết DTTS được sử dụng trên các phương tiện thông tin đại chúng từ trung ương tới địa phương, như: Tày, Thái, Dao, Mông, Gia Rai, Ê Đê, Ba Na, Chăm, Khơ Me… Nhiều địa phương đã triển khai thực hiện dạy tiếng dân tộc cho học sinh trong trường phổ thông. Từ đó, những vấn đề xử lý Tin học các tiếng DTTS tương tự Việt đặt ra bức thiết hiện nay.

Trong những năm qua việc nghiên cứu xử lý tiếng DTTS đã đạt được những kết quả sau:

- Bộ gõ tiếng Chăm Multilingual Edit của Trương Kỳ Quốc, lớp 20TLĐ, Đà Lạt, 2002.
- Bộ gõ tiếng dân tộc của kĩ sư Trương Đình Tú, STVB bốn ngôn ngữ DTTS Ê Đê, Gia Rai, Ba Na, M'Nông và tiếng Việt [1].
- Phần mềm chữ Chăm sử dụng phông chữ Chăm Unicode của Phan Anh Dũng tại Trung tâm CNTT Thừa Thiên-Huế, phần mềm và website có thể xử lý chữ Thái và chữ Chăm [2].
- Trang web Tự học tiếng Chăm-Inrasara [3].
- Bàn phím chữ Chăm: Gần như cùng lúc với chữ Thái Việt Nam, Unicode đã đặt chữ Chăm ở khu vực từ U+AA00 tới U+AA5F, liền trước chữ Thái Việt Nam; bàn phím Chăm [4].
- Đã có các bộ gõ tiếng Chăm, tiếng Ê Đê, tiếng Gia Rai, tiếng Cơ Tu, tiếng M'Nông, tiếng Ba Na.
- TayNguyenKey là chương trình hỗ trợ gõ chữ các dân tộc thiểu số Tây Nguyên của nhóm các tác giả: Tiến sĩ Y Ghi Niê, Kỹ sư Võ Ngọc Hiệp, Thạc sĩ Trần Cát Lâm. TayNguyenKey đã thành công với bộ phông chữ TayNguyenKey xử lý chữ viết của một số dân tộc thiểu số vùng Tây Nguyên. Với bộ phông chữ TayNguyenKey có thể gõ được 6 thứ tiếng dân tộc thiểu số Tây Nguyên: Ê Đê, Gia Rai, Ba Na, Xê Đăng, Cơ Ho và M'Nông. Ngoài ra còn gõ được tiếng Việt và tiếng Anh [5].
- Bộ gõ VnKey của tác giả Trần Thanh Bình hỗ trợ gõ tiếng Việt và ngôn ngữ của dân tộc thiểu số Việt Nam như: Ê Đê, Gia Rai, M'Nông, Cơ Ho, Xê Đăng, Sán Chay…
- Từ điển điện tử Việt-Ê Đê do Đài tiếng nói Việt Nam khu vực Tây Nguyên, nhằm phục vụ việc tra cứu trong quá trình dịch thuật từ tiếng Việt sang tiếng Ê Đê và ứng dụng trong công tác dịch, đọc của chương trình phát thanh tiếng Ê Đê tại Cơ quan thường trú.
- Từ điển điện tử phương ngữ Gia Rai-Việt, đề tài khoa học của sở Thông tin-Truyền thông tỉnh Gia Lai.
- Từ điển điện tử M'Nông-Việt, Việt-M'Nông của sở Khoa học và Công nghệ Đăk Nông.
- Xây dựng môi trường xử lý tiếng Ê Đê ứng dụng trong dạy và học tiếng Ê Đê

- Những đề tài Thạc sĩ về tiếng DTTS tại các trường Đại học, Cao đẳng trong cả nước.

Với những kết quả đạt được về nghiên cứu xử lý tiếng DTTS trên, những khó khăn để tiếp tục nghiên cứu từ các kết quả nghiên cứu xử lý tiếng DTTS:

- Chưa có các kết quả nghiên cứu cho xử lý tiếng DTTS Hrê và tiếng DTTS Co.
- Việc nghiên cứu xử lý ngôn ngữ DTTS thiếu sự liên kết, hợp tác với bên ngoài như: các Đài Phát thanh-Truyền hình, các công ty, các cơ quan chuyên trách về DTTS trong cả nước.

2. KHẢO SÁT THỰC TRẠNG TIẾNG DÂN TỘC THIỂU SỐ Ở VIỆT NAM

2.1. Ngôn ngữ dân tộc Việt Nam

Nước ta, tiếng Việt bao gồm cách phát âm tiếng Việt và chữ Quốc ngữ để viết là ngôn ngữ của người Việt (người Kinh) đang được dùng chính thức trong toàn quốc. Tiếng Việt được chính thức ghi nhận trong hiến pháp là ngôn ngữ quốc gia của Việt Nam. Đây là tiếng mẹ đẻ của gần 86% dân cư Việt Nam, cùng với hơn bốn triệu người Việt ở nước ngoài. Tiếng Việt còn là ngôn ngữ thứ hai của các DTTS Việt Nam và là một phương tiện giao tiếp trong các cơ quan của đại chúng; trong các hoạt động nghiên cứu khoa học, sáng tác, xuất bản văn học nghệ thuật. Mặc dù các dân tộc đều có ngôn ngữ riêng nhưng vẫn xem tiếng Việt là ngôn ngữ của mình. Chính sách song ngữ là một biểu hiện tính thống nhất và đa dạng trong ngôn ngữ của các dân tộc Việt Nam.

Với đặc điểm đa dạng về tộc người nên Việt Nam cũng là quốc gia đa ngôn ngữ. Dân tộc Việt Nam nói các ngôn ngữ khác nhau. Ngoài dân tộc Kinh là dân tộc chiếm gần 86% dân số, còn có 54 dân tộc khác, thuộc các ngữ hệ khác nhau thể hiện trong bảng 1

Bảng 1. Các ngôn ngữ dân tộc Việt Nam

Ngữ hệ	Ngôn ngữ
Nam Á	Kinh (Việt), Mường, Nguồn, Poọng, Thổ, Cuối, Đan Lai, Li hà, Rục, Mày, Sách, Mã Liềng, Kri (Phọong), Aream, Mảng, Khơ Mú, Xinh Mun, Kháng, Ơ Đu, Bru-Vân Kiều, Pacô, Tà Ôi, Cơ Tu, Ba Na, Co, Ca Dong, Ha Lăng, Giẻ, Triêng, Xơ Đăng, Rơn Gao, Ta Kua, Hrê, Mơ Nâm, Ve, Rơ Mân, Tơ Drạ, Brâu, Cơho, M'Nông, Mạ, Xtiêng, Chơ Ro, Khơ Me Nam Bộ.
Thái-Ka Đai	Tày, Nùng, Cao Lan, Thu Lao, Thái Đen, Thái Trắng, Thái Đỏ, Thái Thanh, Thái Dọ, Thái Hàng Tổng, Lào, Lự, Tày Nặm, Pa Dí, Giáy, Bố Y, Tu Dí, Pu Nà, Tống, Thuỷ, Laha, La Chí, Pu Péo, Cơ Lao, Nùng Vẻn.
Mèo-Dao	Mông, Na Mẻo, Pà Thẻn, Miền (Dao Đỏ, Dao Đeo Tiền,

Ngữ hệ	Ngôn ngữ
	Dao Cooc Ngáng, Dao Ôgang, Dao Quần Chẹt, Dao Đại Bản, Dao Tiểu Bản...), Mùn (Dao Quần Trắng, Dao Thanh Y, Dao áo Dài, Dao Họ, Dao Tuyển, Dao Làn Tẻn,...)
Nam Đảo	Chăm Đông (Chăm Ninh – Bình Thuận), Chăm Tây (Chăm An Giang, Tây Ninh), Ê Đê, Gia Rai, Ra Glai, Hroi, Chu Ru.
Hán-Tạng	Hoa, Lô Lô, Hà Nhì, La Hủ, Si La, Cống, Xá Phó, Phù Lá

Trong các ngôn ngữ trên, một số ngôn ngữ có chữ viết cổ truyền như: chữ Nôm Tày; các loại chữ Thái cổ ở Tây Bắc, Quỳ Châu, Man Thanh, Lai Pao; chữ Hán; chữ viết tự dạng Sanscrit của Khơ Me; chữ Nôm Nùng; chữ Chăm cổ; chữ viết tự dạng Sanscrit của Lào; chữ Nôm Dao; chữ Nôm Cao Lan.

Đặc điểm nổi bật của các DTTS Việt Nam là sống đan xen nhau làm cho trạng thái đa ngữ xã hội là trạng thái phổ biến ở các vùng DTTS. Tiếng Việt được xem là ngôn ngữ giao tiếp giữa các dân tộc. Tuy nhiên, ngôn ngữ của các DTTS vẫn có vị trí và tác dụng trong mỗi vùng. Ví dụ ngôn ngữ dân tộc Chăm có vai trò không nhỏ trong đời sống xã hội ở miền Nam, Trung Bộ, Thành phố Hồ Chí Minh và miền Tây Nam Bộ. Dân tộc Chăm có nền văn hoá nghệ thuật rực rỡ, có ngữ ngôn - văn tự lâu đời. Ngôn ngữ dân tộc Chăm không chỉ dùng trong gia đình, trong sinh hoạt lễ hội, tôn giáo mà còn là phương tiện bảo tồn và phát triển văn hoá dân tộc. Bên cạnh đó, một số ngôn ngữ như tiếng Thái, tiếng Tày, tiếng Nùng... cũng được coi là ngôn ngữ vùng, tức là phương tiện giao tiếp giữa các dân tộc cùng chung sống trong vùng nào đó.

2.2. Nguy cơ mai một ngôn ngữ dân tộc thiểu số

Với 55 dân tộc trên đất nước Việt Nam và có khoảng hơn 90 ngôn ngữ khác nhau. Mỗi dân tộc đều có ngôn ngữ của riêng mình. Tuy nhiên, xu hướng hội nhập quốc tế là nguy cơ giảm các ngôn ngữ DTTS.

Chữ viết của mỗi dân tộc thể hiện sự phát triển cao về mặt văn hoá, trình độ phát triển tư duy và nền văn minh. Văn hoá của các dân tộc Việt Nam có nhiều nét tương đồng, nhưng về cơ bản các DTTS vẫn tồn tại một nền văn hoá mang bản sắc riêng; trình độ phát triển kinh tế, văn hoá giữa các dân tộc không đồng đều. Một số DTTS có chữ viết từ rất lâu đời, nhưng nhiều dân tộc khác lại không có chữ viết riêng. Ngôn ngữ DTTS mất dần sự trong sáng vốn có và bị pha tạp tiếng Việt.

Một số nguyên nhân dẫn đến nguy cơ mai một ngôn ngữ DTTS:

- Số người nói các ngôn ngữ DTTS so với tiếng Việt là không nhiều.
- Số lượng người nói một ngôn ngữ trong một đơn vị địa lý, hành chính không cao và không tập trung vì các DTTS ở Việt Nam thường sống đan xen nhau.
- Sự nói bồi lẫn nhau làm cho ngôn ngữ các DTTS nghèo đi và sẽ dẫn tới nguy cơ mai một. Điều này thể hiện rất rõ ở chỗ ngày càng có nhiều người nói được bằng lời nhưng lại không hiểu được văn bản khi đọc, dẫn tới tư duy

chậm chạp.

- Số người nói được các ngôn ngữ DTTS thường thuộc lứa tuổi già và trung niên, còn lứa tuổi thanh niên biết tiếng mẹ đẻ ít hơn, thậm chí còn rất nhiều trẻ em không biết tiếng mẹ đẻ của mình.
- Các hệ thống ngôn ngữ DTTS có phạm vi sử dụng rất hẹp và có chưa được nhiều người biết đến. Phần lớn các ngôn ngữ DTTS không được truyền dạy có tổ chức mà chỉ được truyền dạy tự phát, hay dùng trong phạm vi gia đình, bản làng...
- Vì nhiều nguyên nhân khác nhưng chủ yếu vì lý do kinh tế, các bậc cha mẹ DTTS phải hướng con em mình nắm vững tiếng Việt và các ngoại ngữ (Anh, Pháp, Trung, Nhật Bản...) để tìm kiếm việc làm, bảo đảm đời sống.
- Trước thực trạng tiếng nói của các DTTS đang đứng trước nguy cơ mai một, cộng đồng các dân tộc Việt Nam và Chính phủ cần có những chương trình như khuyến khích, vận động nhân dân các dân tộc giao tiếp hằng ngày bằng tiếng mẹ đẻ.

Trong Giáo dục - Đào tạo, cần xuất bản nhiều hơn nữa các loại sách song ngữ. Khuyến khích thế hệ trẻ thuộc đồng bào các dân tộc thiểu số học tập, hiểu biết và sử dụng thành thạo tiếng nói, chữ viết của dân tộc mình. Đào tạo đội ngũ trí thức thuộc đồng bào các dân tộc thiểu số và tạo điều kiện để trí thức, cán bộ dân tộc thiểu số trở về phục vụ quê hương. Đưa vào chương trình giảng dạy trong các trường phổ thông, trường phổ thông dân tộc nội trú, trung tâm giáo dục thường xuyên, trung tâm học tập cộng đồng, trường dạy nghề, trung học chuyên nghiệp, cao đẳng và đại học phù hợp với địa bàn vùng dân tộc

Về mạng lưới truyền thông đại chúng, cần mở thêm nhiều kênh phát thanh, truyền hình bằng tiếng DTTS. Cần xây dựng thêm nhiều chương trình nội dung phong phú, chất lượng cao và định giờ phát sóng để mọi người dân có thể theo dõi.

2.3. Khó khăn và thách thức

Trong xử lý ngôn ngữ DTTS khó khăn đặt ra đầu tiên là phải mã hóa thích hợp hệ thống chữ viết các DTTS trong Unicode và phải phù hợp với bàn phím tiếng Anh, bởi vì các DTTS thường có hệ thống chữ viết của riêng mình.

Xử lý ngôn ngữ DTTS thường xuyên phải đối mặt với khó khăn đầu tiên đó là bộ chữ cái tiếng DTTS đã có trong Unicode hay chưa, tiếp theo là thiếu nguồn tài nguyên dữ liệu chuẩn hóa dưới dạng điện tử, chuyên môn. Chính sự khan hiếm nguồn tài nguyên dữ liệu là một hạn chế cho phương pháp tiếp cận hướng dữ liệu trong xử lý ngôn ngữ DTTS. Khó khăn cũng phải được kể đến đó là thiếu sự hỗ trợ về tài chính dành cho các hoạt động nghiên cứu xử lý ngôn ngữ DTTS.

Mặt khác, xử lý ngôn ngữ DTTS còn phải vượt qua một số khó khăn phát sinh từ những thực trạng đặc biệt của ngôn ngữ DTTS vì chỉ có nhóm ít người dùng, không có đủ nguồn nhân lực chuyên môn, rất ít các nhà ngôn ngữ học DTTS và các nhà khoa học máy tính là người DTTS. Chính vì vậy, việc áp dụng các

phương pháp tiếp cận dựa trên luật để gán nhãn, phân tích cú pháp… có thể rất khó khăn.

3. ĐỊNH HƯỚNG QUI TRÌNH XỬ LÝ TIẾNG DTTS HRÊ VÀ TIẾNG DTTS CO

Trong xử lý ngôn ngữ tự nhiên nói chung và xử lý ngôn ngữ DTTS Hrê và DTTS Co ở Việt Nam nói riêng, việc xây dựng hạ tầng cơ sở cho xử lý ngôn ngữ là rất cần thiết nhằm tạo ra bất kỳ một công cụ kỹ thuật hay ứng dụng liên quan đến xử lý ngôn ngữ. Trong xử lý ngôn ngữ DTTS việc xây dựng hạ tầng cơ sở, xây dựng các công cụ kỹ thuật và triển khai các ứng dụng phải được thực hiện từng bước và phối hợp với nhau để có được kết quả tốt nhất. Thông qua các hoạt động nghiên cứu của các nhóm SALTMIL, MILLE, EMILLE, xử lý ngôn ngữ Basque [6], [7], [8], [9], [10] cho thấy qui trình nghiên cứu xử lý ngôn ngữ DTTS thường được thực hiện qua bốn giai đoạn:

Giai đoạn đầu tiên là xây dựng hạ tầng cơ sở, cụ thể: mã Unicode hệ thống chữ viết, xây dựng cơ sở dữ liệu từ vựng và xây dựng từ điển máy tính.

Giai đoạn thứ hai là xây dựng các công cụ kỹ thuật trong XLNNTN nói chung và xử lý ngôn ngữ DTTS nói riêng, cụ thể: công cụ thống kê trong xây dựng kho ngữ liệu, công cụ phân tích hình thái học, công cụ kiểm tra và sửa lỗi chính tả, công cụ xử lý tiếng nói ở mức từ, công cụ gán nhãn từ loại trong các kho ngữ liệu.

Giai đoạn thứ ba là xây dựng các công cụ kỹ thuật và các ứng dụng nâng cao, cụ thể: môi trường để tích hợp các công cụ, thu thập dữ liệu từ website, kiểm tra ngữ pháp, nâng cấp các phiên bản từ điển, kho ngữ vựng đa ngữ, xử lý tiếng nói ở mức câu.

Giai đoạn thứ tư là vấn đề về đa ngữ và các ứng tổng quát, cụ thể: tìm kiếm và khai thác thông tin, dịch máy, từ điển trực tuyến và các ứng dụng liên quan đến mối quan hệ giữa từ vựng và ngữ nghĩa đa ngữ.

Qua đó cho thấy rằng, khi thực hiện triển khai hệ thống xử lý ngôn ngữ cho DTTS không nên bắt đầu phát triển các ứng dụng nếu chưa có hạ tầng cơ sở cho xử lý ngôn ngữ. Các nguồn tài nguyên CSDL nên thiết kế theo hướng mở và có thể được sử dụng lại cho bất kỳ các công cụ và các ứng dụng khác.

Tóm lại, vấn đề chia sẻ các kết quả nghiên cứu trong xử lý ngôn ngữ DTTS cũng là một yếu tố quan trọng, nhằm tận dụng tất cả sự hợp tác có thể nảy sinh giữa các nhà nghiên cứu về xử lý ngôn ngữ DTTS.

4. XÂY DỰNG KHO NGỮ VỰNG SONG NGỮ VIỆT-HRÊ, HRÊ-VIỆT, VIỆT-CO VÀ CO-VIỆT

Hiện nay, các nguồn dữ liệu song ngữ của DTTS ở Việt Nam nói chung và dân tộc Hrê, dân tộc Co nói riêng chủ yếu là ở dạng từ điển giấy. Vì vậy, trong xử lý ngôn ngữ DTTS, việc hợp nhất các nguồn dữ liệu từ điển giấy trong xây dựng KNV song ngữ Việt-Dân tộc thiểu số là thật sự cần thiết.

4.1. Tổ chức kho ngữ vựng song ngữ

4.1.1. Tiêu chí dữ liệu KNV song ngữ Việt-Hrê và Hrê-Việt

Với mục tiêu, xây dựng KNV song ngữ Việt-Hrê và Hrê-Việt làm hạ tầng cơ sở cho môi trường xử lý tiếng Hrê. Các tiêu chí dữ liệu được đặt ra trong kho ngữ vựng như sau:

Các từ tiếng Hrê chủ yếu được thu thập và ghi theo tiếng Hrê địa phương vốn được xem dễ nghe và dễ hiểu nhất. Các mục từ tiếng Hrê phản ảnh phần nào vốn văn hóa truyền thống của người Hrê. Tiếng Hrê được ghi bằng chữ Hrê.

Các từ tiếng Việt là từ tiếng Việt phổ thông và được ghi bằng chữ Quốc ngữ.

Các ví dụ được đưa vào để làm rõ nghĩa và cách sử dụng của mục từ hay còn gọi là ngữ cảnh của mục từ.

Các mục từ được gán nhãn từ loại: gán nhãn N cho danh từ, gán nhãn V cho động từ, gán nhãn A cho tính từ, gán nhãn O cho các mục từ không phải là danh từ, động từ hay tính từ.

Từ đa nghĩa được ghi nhận, dịch và đối chiếu với các từ khác nhau tương đương trong ngôn ngữ đích.

Khi gióng hàng từ của ngôn ngữ nguồn, tìm từ tương đương trong ngôn ngữ đích, trên cơ sở nghĩa cơ bản, nghĩa thường dùng hiện nay ở cả hai ngôn ngữ.

Dữ liệu được lưu trên máy với phông chữ Unicode.

4.1.2. Tiêu chí dữ liệu KNV song ngữ Việt-Co và Co-Việt

Với mục tiêu, xây dựng KNV song ngữ Việt-Co và Co-Việt làm hạ tầng cơ sở cho môi trường xử lý tiếng Co. Các tiêu chí dữ liệu được đặt ra trong kho ngữ vựng như sau:

Các từ tiếng Co chủ yếu được thu thập và ghi theo tiếng Co địa phương vốn được xem dễ nghe và dễ hiểu nhất. Các mục từ tiếng Hrê phản ảnh phần nào vốn văn hóa truyền thống của người Hrê. Tiếng Hrê được ghi bằng chữ Hrê.

Các từ tiếng Việt là từ tiếng Việt phổ thông và được ghi bằng chữ Quốc ngữ.

Các ví dụ được đưa vào để làm rõ nghĩa và cách sử dụng của mục từ hay còn gọi là ngữ cảnh của mục từ.

Các mục từ được gán nhãn từ loại: gán nhãn N cho danh từ, gán nhãn V cho động từ, gán nhãn A cho tính từ, gán nhãn O cho các mục từ không phải là danh từ, động từ hay tính từ.

Từ đa nghĩa được ghi nhận, dịch và đối chiếu với các từ khác nhau tương đương trong ngôn ngữ đích.

Khi gióng hàng từ của ngôn ngữ nguồn, tìm từ tương đương trong ngôn ngữ đích, trên cơ sở nghĩa cơ bản, nghĩa thường dùng hiện nay ở cả hai ngôn ngữ.

Dữ liệu được lưu trên máy với phông chữ Unicode.

4.1.3. Cấu trúc kho ngữ vựng

Tổ chức cấu trúc KNV là bước quan trọng trong xây dựng KNV. Trong

nghiên cứu này, KNV được thiết kế theo mô hình CSDL quan hệ. CSDL quan hệ được sử dụng như một tập hợp các bảng lưu trữ dữ liệu và lưu trữ một tập hợp các thực thể có quan hệ với nhau. Các bảng CSDL tương tự như một KNV, được lưu trữ hoàn toàn độc lập về cấu trúc cũng như về dữ liệu. Mô hình CSDL quan hệ có những ưu điểm và nhược điểm sau:

Ưu điểm: CSDL quan hệ là một KNV riêng biệt, có khả năng linh hoạt rất cao, ít lập trình để truy cập CSDL hơn các CSDL khác. Độc lập về cấu trúc CSDL, do đó, người sử dụng và người thiết kế hoàn toàn không phải quan tâm tới cấu trúc CSDL. Dễ tạo ra một giao diện thích hợp với người sử dụng.

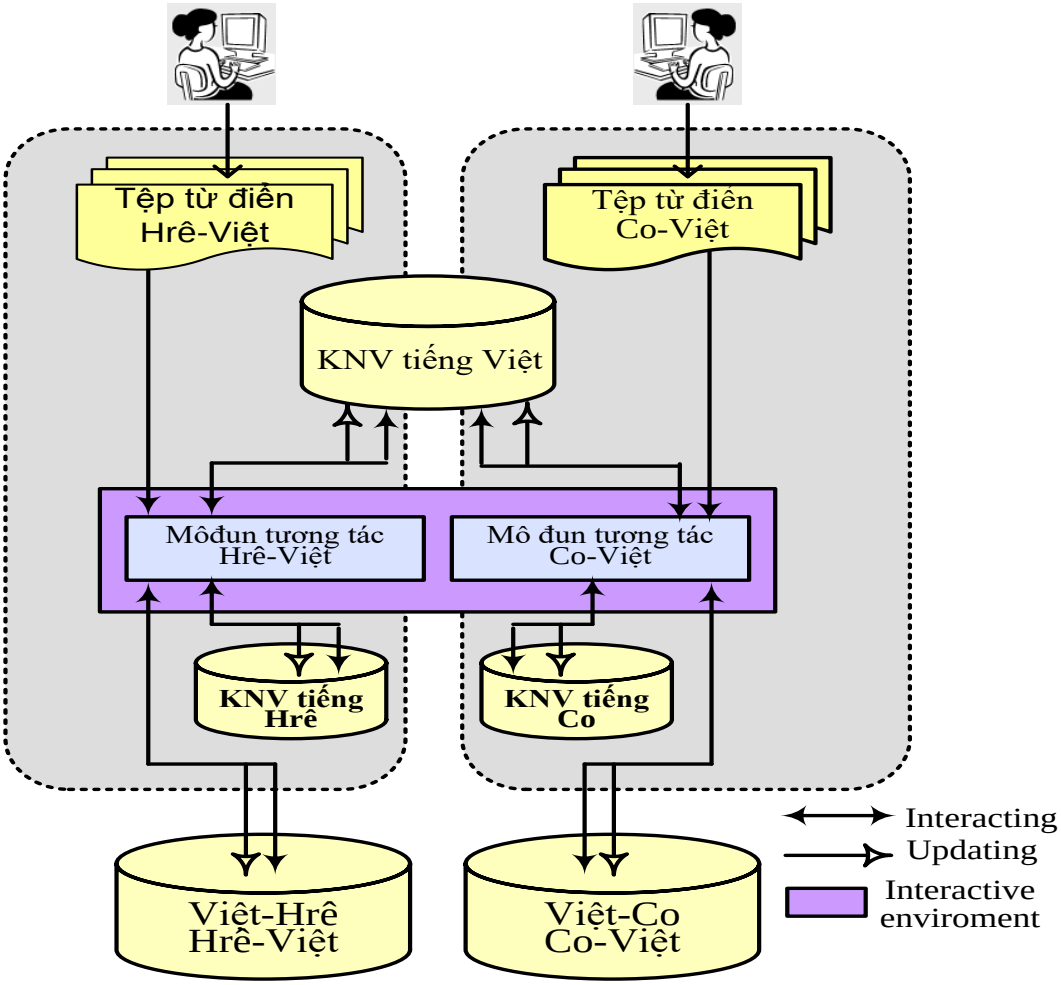
Nhược điểm: CSDL quan hệ che hết gần như toàn bộ cấu trúc vật lý của CSDL. Do đó, cần phải có phải có hệ điều hành và một hệ thống máy tính đủ mạnh để hỗ trợ cho việc thực hiện những thao tác truy cập dữ liệu.

Tuy nhiên, các KNV song ngữ Việt-Hrê, Hrê-Việt, Việt-Co, Co-Việt với số mục từ không quá lớn, cùng với cấu hình máy tính ngày càng được phát triển và sự hỗ trợ của công nghệ cao, thì nhược điểm này cũng được chấp nhận.

4.2. Mô hình hợp nhất nguồn dữ liệu song ngữ

Xuất phát từ thực trạng KNV Việt-DTTS ở Việt Nam nói chung và các KNV song ngữ Việt-Hrê, Hrê-Việt, Việt-Co, Co-Việt nói riêng, nhằm góp phần giải quyết bài toán xây dựng các KNV song ngữ Việt-Hrê, Hrê-Việt, Việt-Co, Co-Việt với nguồn dữ liệu đầu vào chủ yếu là các từ điển giấy Hrê-Việt, Co-Việt. Bài báo đề xuất mô hình hợp nhất nguồn dữ liệu song ngữ từ điển giấy Hrê-Việt, Co-Việt trong xây dựng các KNV song ngữ Việt-Hrê, Hrê-Việt, Việt-Co, Co-Việt.

Mô hình hợp nhất nguồn dữ liệu song ngữ được thể hiện trong Hình 1.



Hình 1. *Mô hình hợp nhất nguồn dữ liệu song ngữ*

4.2.1. Hoạt động của mô đun tương tác Hrê-Việt

- Dữ liệu vào:
  - Tệp từ điển Hrê-Việt,
  - KNV tiếng Việt,
  - KNV tiếng Hrê,
  - KNV Việt-Hrê
  - KNV Hrê-Việt
- Dữ liệu ra:
  - KNV tiếng Việt
  - KNV tiếng Hrê
  - KNV Việt-Hrê
  - KNV Hrê-Việt
- Trình từ thực hiện

Bước 1: đọc dữ liệu trên mỗi hàng trong tệp từ điển Hrê-Việt (từ tiếng Hrê, tập các từ tiếng Việt, từ loại và các ví dụ Hrê:Việt).

Bước 2: kiểm tra từ tiếng Hrê trong KNV Hrê, nếu chưa có thì bổ sung vào.

Bước 3: đọc chỉ số của từ tiếng Hrê

Bước 4: tách từ tiếng Việt trong tập các từ tiếng Việt đọc được ở cột thứ hai của hàng trong tệp từ điển. Thực hiện lần lượt cho mỗi từ tách được:

Bước 4.1: kiểm tra từ tiếng Việt tách được trong KNV tiếng Việt, nếu chưa có thì bổ sung vào và ghi chú cho việc xác định từ mới được bổ sung vào KNV tiếng Việt.

Bước 4.2: đọc chỉ số của từ tiếng Việt

Bước 4.3: trích trong tập ví dụ các ví dụ Hrê:Việt tương ứng với từ tiếng Việt đã tách được ở bước 4 trong tập ví dụ đọc được ở bước 1. Chuyển ví dụ Hrê:Việt thành Việt:Hrê.

Ví dụ Hrê-Việt: "Gu d'ruh Hrê ta lêu d'ha ra ngot: Cô gái Hrê hát cũng hay."

chuyển thành ví dụ Việt-Hrê: "Cô gái Hrê hát cũng hay: Gu d'ruh Hrê ta lêu d'ha ra ngot."

Bước 4.4: kiểm tra bộ ba giá trị (chỉ số từ tiếng Việt, chỉ số từ tiếng Hrê, từ loại) trong KNV Việt-Hrê:

Nếu chưa có thì bổ sung bộ ba giá trị và các ví dụ Việt:Hrê có được từ bước 4.3 vào KNV Việt-Hrê.

Nếu đã có thì kiểm tra các ví dụ đã trích trong bước 4.3 trong tập các ví dụ tương ứng với bộ ba giá trị, nếu ví dụ nào chưa có thì bổ sung vào.

Bước 5: quay lại bước 1 lần lượt đọc dữ liệu trên mỗi hàng trong tệp từ điển Hrê-Việt cho đến hết.

4.2.2. Hoạt động của mô đun tương tác Co-Việt

- Dữ liệu vào:
  - Tệp từ điển Co-Việt,
  - KNV tiếng Việt,
  - KNV tiếng Co,
  - KNV Việt-Co
  - KNV Co-Việt
- Dữ liệu ra:
  - KNV tiếng Việt
  - Tiếng Co
  - KNV Việt-Co
  - KNV Co-Việt
- Trình từ thực hiện

Bước 1: đọc dữ liệu trên mỗi hàng trong tệp từ điển Co-Việt (từ tiếng Co, tập các từ tiếng Việt, từ loại và các ví dụ Co:Việt).

Bước 2: kiểm tra từ tiếng Co trong KNV Co, nếu chưa có thì bổ sung vào.

Bước 3: đọc chỉ số của từ tiếng Hrê

Bước 4: tách từ tiếng Việt trong tập các từ tiếng Việt đọc được ở cột thứ hai của hàng trong tệp từ điển. Thực hiện lần lượt cho mỗi từ tách được:

Bước 4.1: kiểm tra từ tiếng Việt tách được trong KNV tiếng Việt, nếu chưa có thì bổ sung vào và ghi chú cho việc xác định từ mới được bổ sung vào KNV tiếng Việt.

Bước 4.2: đọc chỉ số của từ tiếng Việt

Bước 4.3: trích trong tập ví dụ các ví dụ Co:Việt tương ứng với từ tiếng Việt đã tách được ở bước 4 trong tập ví dụ đọc được ở bước 1. Chuyển ví dụ Co:Việt thành Việt:Co.

Ví dụ Co-Việt: "Tamoi Kool êp e rôt hmât chêêk?: Người Co các anh có thích đánh chiêng không?"

chuyển thành ví dụ Việt-Co: "Người Co các anh có thích đánh chiêng không?: Tamoi Kool êp e rôt hmât chêêk?"

Bước 4.4: kiểm tra bộ ba giá trị (chỉ số từ tiếng Việt, chỉ số từ tiếng Co, từ loại) trong KNV Việt-Co:

Nếu chưa có thì bổ sung bộ ba giá trị và các ví dụ Việt:Hrê có được từ bước 4.3 vào KNV Việt-Co.

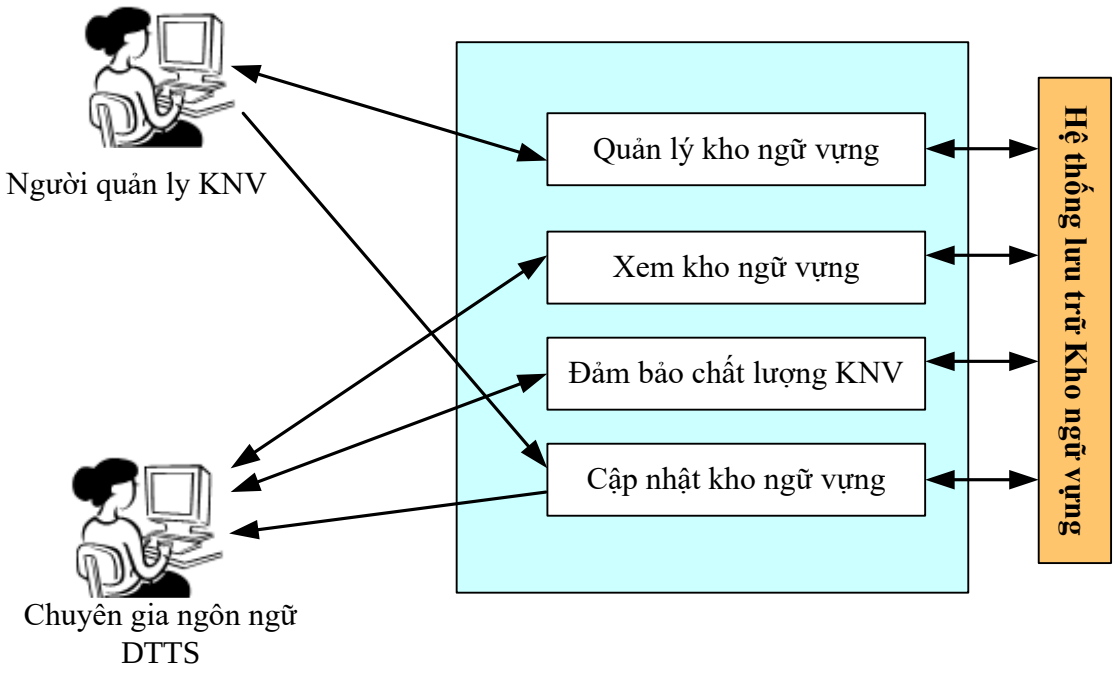
Nếu đã có thì kiểm tra các ví dụ đã trích trong bước 4.3 trong tập các ví dụ tương ứng với bộ ba giá trị, nếu ví dụ nào chưa có thì bổ sung vào.

Bước 5: quay lại bước 1 lần lượt đọc dữ liệu trên mỗi hàng trong tệp từ điển Hrê-Việt cho đến hết.

4.2.3. Hệ thống quản lý kho ngữ vựng

Viêc chia sẻ KNV cho các hoạt động nghiên cứu là điều cần thiết. Để quản lý dữ liệu trong kho ngữ vựng đề tài tập trung giải pháp sử dụng những lợi thế của việc sử dụng công nghệ Blockchain mục tiêu tạo lập một nền tảng chia sẻ, trao đổi dữ liệu an toàn, tính toàn vẹn của dữ liệu và chất lượng dữ liệu. Blockchain và các công nghệ được cung cấp bởi blockchain có thể giải quyết những thách thức này. Hệ thống quản lý kho ngữ vựng làm sao để có thể truy cập kho ngữ vựng được xác nhận bất cứ lúc nào, có thể sử dụng dữ liệu đã lưu trữ. Đồng thời các chuyên gia ngôn ngữ dân tộc thiểu số có thể xem thông tin lưu trữ trên hệ thống, không chỉ là tra cứu từ vựng mà còn cả quá trình các mục từ được cập nhật cụ thể trong Blockchain. Viêc chia sẻ kho ngữ vựng cho các nhà nghiên cứu tiếng DTTS nói chung và tiếng Hrê, tiếng Co nói riêng dễ dàng.

Blockchain là giải pháp phù hợp để góp phần nâng cao chất lượng KNV đảm bảo chất lượng các mục từ được cập nhật vào KNV.



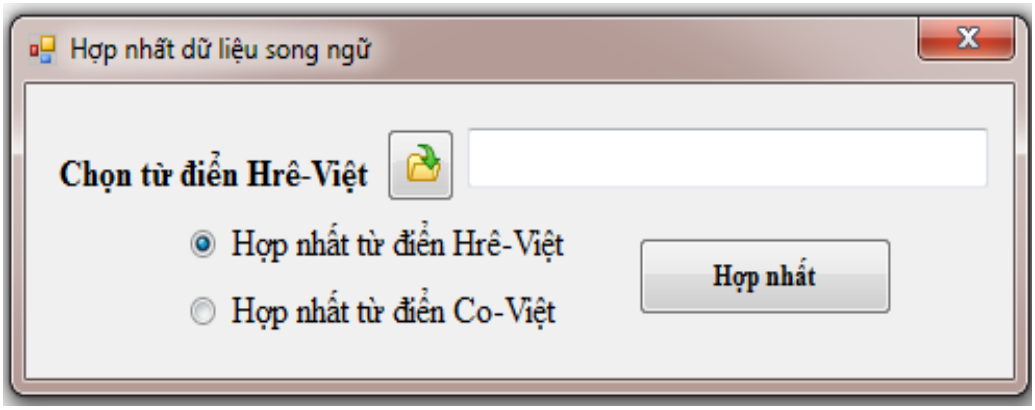
Hình 1. *Kiến trúc việc đảm bảo chất lượng các mục từ được cập nhật vào KNV*

Hình trên là một cách tiếp cận đối với sử dụng nền tảng Bảo mật, đảm bảo tính toàn vẹn của dữ liệu và chất lượng dữ liệu của KNV. Tất cả các thuộc tính bắt buộc không thể được đáp ứng bởi một cơ chế. Phạm vi quản lý KNV:

- Bộ dữ liệu: là dữ liệu thực tế phải được chuyển cho các chuyên gia ngôn ngữ DTTS. Những thách thức liên quan đến việc chuyển dữ liệu thực tế là cung cấp tính toàn vẹn của dữ liệu và truyền an toàn. Một lợi thế lớn của việc sử dụng blockchain là nó có thể được sử dụng để cung cấp bằng chứng giả mạo vì bất kỳ dữ liệu nào trên blockchain là bất biến. Do đó, tính toàn vẹn của dữ liệu có thể được xác minh nếu nó nằm trên blockchain. Một khả năng là lưu các dấu thời gian của bộ dữ liệu trên blockchain để chúng không thể bị giả mạo. Dữ liệu sẽ được lưu trữ trên blockchain như thế nào và nó sẽ được chuyển đến chuyển cho chuyên gia ngôn ngữ DTTS như thế đó. Việc sử dụng blockchain để truyền dữ liệu thực tế có thể hữu ích theo nhiều cách, khả năng triển khai sẽ là một phần trong công việc trong tương lai.
- Chất lượng dữ liệu: chuyên gia ngôn ngữ DTTS có thể kiểm tra chất lượng dữ liệu mà không cần xem dữ liệu thực tế và người quản lý KNV cũng không thể xem yêu cầu của chuyên gia ngôn ngữ DTTS. Để kiểm tra chất lượng của bộ dữ liệu, một chức năng đảm bảo chất lượng KNV được đề xuất. Chức năng này sử dụng thực hiện kiểm tra mục từ được cập nhật trong KNV.

5. KẾT QUẢ THỰC NGHIỆM

Việc cập nhật cập nhật mục từ vào các kho ngữ vựng thông qua bộ công cụ hợp nhất nguồn dữ liệu song ngữ được đề xuất xây dựng thể hiện trong hình 3.



Hình 3. *Bộ công cụ cập nhật mục từ vào kho ngữ vựng*

Kết quả, các mục từ được cập nhập vào trong các KNV sau khi thực hiện chuyển lần lượt các tệp từ điển Hrê-Việt và các tệp từ điển Co-Việt vào môi trường hợp nhất, được thống kê trong Bảng 2.

Bảng 2. *Số mục từ được cập nhật vào các KNV*

Kho ngữ vựng	**Mô đun tương tác**	
	Hrê-Việt	**Co-Việt**
Hre	3.020	
Hre-Vietnamese	3.971	
Vietnamese -Hre	4.345	
Co		1.042
Co-Vietnamese		1.957
Vietnamese -Co		2.112

6. KẾT LUẬN

Trên cơ sở định hướng qui trình nghiên cứu xử lý tiếng DTTS Hrê và DTSS Co ở Việt Nam. Các KNV song ngữ Việt-Hrê, Hrê-Việt, Việt-Co, Co-Việt được cập nhật từ mô hình hợp nhất nguồn dữ liệu song ngữ từ điển giấy Hrê-Việt, Co-Việt được đề xuất.

Các KNV đươc xây dựng là cở sở hạ tầng trong xử lý tiếng DTTS Hrê và DTTS Co. Từ cở sở hạ tầng này sẽ tiếp tục phát triển các ứng dụng như tra cứu từ vựng, dịch văn bản, kiểm tra lỗi chính tả, …

Giải pháp đề xuất có tính thiết thực, vì đã góp phần khắc phục những hạn chế trong các KNV song ngữ Việt-Hrê, Hrê-Việt, Việt-Cơ, Đồng Việt mà các nghiên cứu trước đây chưa thực hiện được.

Tài liệu tham khảo

[1] http://www.vietlex.com/kho-ngu-lieu.

[2] Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Le-Minh Huyen, and Quang-Thuy Ha. Vietnamese word segmentation with CRFs and SVMs: An investigation. In 20th Pacific Asia Conference on Language, Information

and Computation (PACLIC 2006)
[3] Lưu Tuấn Anh và Yamamoto Kazuhide. Ứng dụng phương pháp Pointwise vào bài toán tách từ cho tiếng Việt, http://viet.jnlp.org/dongdu
[4] Hong Phuong Le, Thi Minh Huyen Nguyen, Azim Roussanaly, Tuong Vinh Ho. A Hybrid Approach to Word Segmentation of Vietnamese Texts. 2nd International Conference on Language and Automata Theory and Applications - LATA 2008, Mar 2008, Tarragona, Spain.
[5] Nguyễn Đức Khanh. "TayNguyenKey - Chương trình hỗ trợ gõ chữ các dân tộc thiểu số Tây Nguyên", Sở giáo dục Đăk Lăk,2010, địa chỉ: http://thpt-ngogiatu-daklak.edu.vn/taynguyenkey-chuong-trinh-ho-tro-go-chu-cac-dan-toc-thieu-so-tay-nguyen.html.
[6] A. Diaz de Ilarraza, A. Gurrutxaga, I. Hernaez,N. Lopez de Gereñu and K. Sarasola, "Integrating language engineering resources and tools into systems with linguistic capabilities", Proceeding of TALN (Traitement Automatique de Langues Naturelles), pp. 243-252, 2003.
[7] Briony Williams, Mikel L. Forcada, Kepa Sarasola, "6th SaLTMiL Workshop on: Collaboration: interoperability between people in the creation of language resources for less-resourced languages", SALTMIL proceeding, Morocco, 2008.
[8] Kepa Sarasola, Francis M. Tyers, Mikel L. Forcada, "7th SaLTMiL Workshop on: Creation and use of basic lexical resources for less-resourced languages", Proceeding of SALTMIL, Malta, 2010.
[9] Mikel L. Forcada, Guy De Pauw, Gilles-Maurice de Schryver, Kepa Sarasola, Francis M. Tyers, Peter Waiganjo Wagacha, "Language technology for normalisation of less-resourced languages", proceeding of SALTMIL, Turkey, 2012.
[10] Mikel L. Forcada, Kepa Sarasola, Francis M. Tyers, "Free/open-Source Language Resources for the Machine Translation of Less-Resourced Languages", SALTMIL procceding, Iceland, 2014.