MỘT SỐ ĐỀ XUẤT HỖ TRỢ CHUYỂN ĐỔI VĂN BẢN TIẾNG VIỆT SANG VĂN BẢN TIẾNG DÂN TỘC THIỂU SỐ Ở VIỆT NAM

PROPOSED SOLUTIONS FOR SUPPORTING THE CONVERSION OF TEXTS FROM VIETNAMESE TO ETHNIC MINORITY'S LANGUAGES IN VIETNAM

Trương Đình Huy[1], Lương Chi Mai[2], Huỳnh Công Pháp[3]

[1]*Trường Đại học Duy Tân; Email: huy.truongdinh@gmail.com*
[2]*Viện Hàn lâm Khoa học và Công nghệ Việt Nam; Email: lcmai@ioit.ac.vn*
[3]*Trường Cao đẳng Công nghệ Thông tin, Đại học Đà Nẵng; Email: hcphap@gmail.com*

Tóm tắt - Nhằm gìn giữ, phát huy bản sắc văn hóa, phát triển kinh tế-xã hội (KT-XH) của các vùng đồng bào dân tộc thiểu số (DTTS), Đảng và Nhà nước ta đã có rất nhiều chủ trương, chính sách, chiến lược phát triển đối với đồng bào các DTTS [7] [8]. Tuy nhiên, quá trình cụ thể hóa, thể chế hóa các quan điểm, chủ trương, đường lối của Đảng và Nhà nước đến với đồng bào DTTS cũng như việc nghiên cứu tiếng của đồng bào DTTS đang gặp khó khăn do sự khác biệt về mặt ngôn ngữ giữa tiếng Việt và tiếng của các đồng bào DTTS. Bên cạnh các kết quả nghiên cứu về xử lý tiếng DTTS, xử lý tiếng Việt, các phần mềm, các website, các tiện ích…đang được sử dụng, nhóm tác giả nêu một số đề xuất hỗ trợ chuyển đổi văn bản tiếng Việt sang văn bản tiếng của một số đồng bào DTTS ở Việt Nam, nhằm nêu ra những vấn đề cần giải quyết, giúp định hướng nghiên cứu trong xử lý tiếng của DTTS, giúp giải quyết những khó khăn do sự khác biệt về ngôn ngữ giữa tiếng Việt và tiếng của đồng bào DTTS ở Việt Nam…

Từ khóa - chuyển đổi văn bản; chuyển đổi ngôn ngữ; xử lý tiếng dân tộc thiểu số; dịch tự đông; công cụ xử lý tiếng dân tộc.

Abstract - In order to preserve and promote the cultural identity, the socio - economic development of the areas inhabited by minorities, the Party and the State have issued a lot of guidelines, policies and strategies for the development of the ethnic minorities. However, the process of specifying, institutionalizing viewpoints, policies and guidelines of the Party and the State for ethnic minority languages as well as the study of ethnic minority languages are facing difficulties due to differences between Vietnamese and ethnic minority languages. Besides the results of research on minority language processing, Vietnamese processors, software, websites, gadgets,... in use, the authors put forward some proposals to support the conversion of texts from Vietnamese to a number of ethnic minorities languages in Vietnam with a view to highlighting problems to be solved, orientating research in minority language processing, helping to solve the difficulties caused by the differences between Vietnamese and the languages of ethnic minorities in Vietnam...

Key words - Converting texts; language conversion; ethnic language processing; machine translation; ethnic language processing tools.

1. Đặt vấn đề

Hơn 50 năm qua, các nhà ngôn ngữ học và các nhà tin học đã giải quyết được rất nhiều bài toán về xử lý ngôn ngữ tự nhiên (XLNNTN) và đã đạt được nhiều kết quả khả quan [1][2], giải quyết được nhiều vấn đề trong quá trình sử dụng máy tính để xử lý ngôn ngữ, góp phần quan trọng trong quá trình hội nhập quốc tế, đặc biệt giải quyết vấn đề khác biệt về mặt ngôn ngữ giữa các quốc gia trên thế giới. Các kết quả đó có thể liệt kê như sau: *Kiểm lỗi chính tả, kiểm lỗi văn phạm, từ điển đồng nghĩa, phân tích văn bản, tóm tắt văn bản, tổng hợp tiếng nói, nhận dạng tiếng nói, dịch tự động, xử lý đa ngôn ngữ trên cùng một phần mềm* [1], [2], [10], [13] …

Tại Việt Nam những năm gần đây, XLNNTN đã trở thành một lĩnh vực khoa học công nghệ được coi là mũi nhọn [9], đặc biệt trong xử lý tiếng Việt đã có nhiều kết quả của nhiều công trình nghiên cứu làm cơ sở để xây dựng các ứng dụng, tiện ích…nhằm giải quyết bài toán phục vụ nhu cầu xử lý tiếng Việt.

Việt Nam có 54 thành phần dân tộc khác nhau, trong đó dân tộc Kinh (Việt) chiếm gần 90% tổng số dân cả nước, hơn 10% còn lại là dân số của 53 dân tộc thiểu số [3], [5], [7], [8], [11]. Trong đó, 26 dân tộc đã có chữ viết riêng, có bản sắc văn hóa riêng. Quá trình cụ thể hóa, thể chế hóa các chủ trương, chính sách của Đảng và Nhà nước đến với các vùng đồng bào DTTS đang gặp khó khăn do "rào cản" về mặt ngôn ngữ giữa tiếng Việt và tiếng của đồng bào DTTS.

Việc xử lý tiếng dân tộc thiểu số (DTTS) ở Việt Nam, đặc biệt xử lý tiếng DTTS ở khu vực miền Trung và Tây Nguyên, hiện nay cũng đã có một số kết quả nghiên cứu của các cá nhân, tổ chức…Tuy nhiên, số lượng các công trình còn rất ít và còn rất nhiều vấn đề về xử lý tiếng DTTS cần phải giải quyết.

Trong bài báo này, chúng tôi phân tích, đánh giá tổng quan tình hình nghiên cứu, xử lý tiếng DTTS hiện nay, từ đó nêu ra một số đề xuất giúp hỗ trợ chuyển đổi văn bản tiếng Việt sang văn bản của một số tiếng DTTS. Đồng thời, phân tích những khó khăn, lợi ích về mặt KT-XH và tính khả thi của các đề xuất nêu ra nhằm giúp định hướng nghiên cứu, xử lý tiếng DTTS ở Việt Nam.

2. Tổng quan tình hình nghiên cứu, xử lý tiếng DTTS ở Việt Nam

2.1. Tổng quan tình hình nghiên cứu

Hưởng ứng các chủ trương, chính sách của Đảng và Nhà nước đối với đồng bào DTTS, nhiều nhà khoa học, nghiên cứu sinh, học viên cao học, sinh viên các trường đại học [5] trong nước chọn hướng nghiên cứu xử lý tiếng DTTS. Đặc biệt, tại Trung tâm Nghiên cứu và Ứng dụng công nghệ thông tin (DATIC) thuộc Trường Đại học Bách khoa, Đại học Đà Nẵng, đã có nhiều công trình nghiên cứu về xử lý tiếng DTTS đã công bố, trong đó có thể kể đến các giả: PGS. TS. Phan Huy Khánh, PGS.TS. Võ Trung Hùng, TS. Huỳnh Công Pháp. Bên cạnh đó, nhiều học viên cao học chuyên ngành khoa học máy tính của Đại học Đà Nẵng cũng chọn hướng nghiên cứu xử lý tiếng DTTS. Ngoài ra, còn có một số công ty, trung tâm trong nước và ở khu vực Tây Nguyên như: Trung tâm CNTT-TT Sở Thông tin và

Truyền thông Gia Lai, Công ty TNHH Công nghệ Tin học tuổi trẻ Lạc Việt đã đầu tư xây dựng một số công cụ, tiện ích giúp soạn thảo, học tập, nghiên cứu tiếng các DTTS như: Các bộ gõ tiếng DTTS, các từ điển điện tử phương ngữ giúp tra cứu nghĩa giữa tiếng Việt và tiếng của một số DTTS như Jrai, Bahnar…

2.2. Một số công trình, sản phẩm đã công bố

Xử lý tiếng DTTS là mảng nghiên cứu mới tại Việt Nam, tập trung chủ yếu các nhà khoa học, nhà nghiên cứu ở khu vực miền Trung và Tây Nguyên. Hiện đã có một số công trình, sản phẩm đã công bố và đưa vào sử dụng, làm cơ sở, đặt nền tảng quan trọng trong hướng nghiên cứu xử lý tiếng DTTS còn mới mẻ và còn nhiều bài toán, nhiều vấn đề cần phải được giải quyết. Các kết quả các công trình, sản phẩm, tiện ích có thể nêu ra như sau:

2.2.1. Một số Bộ gõ tiếng DTTS

Một trong các sản phẩm phần mềm đầu tiên phải kể đến đó là các Bộ gõ tiếng DTTS, đây là tiện ích giúp quá trình soạn thảo các văn bản tiếng DTTS, hỗ trợ quá trình nhập dữ liệu "đầu vào" cần thiết trong quá trình nghiên cứu, xử lý tiếng DTTS. Một số bộ gõ tiếng DTTS đang được sử dụng có thể kể ra như sau: Bộ gõ TayNguyenKey do nhóm tác giả gồm: TS. Tiến sĩ Y Ghi Niê, Kỹ sư Võ Ngọc Hiệp, Ths. Trần Cát Lâm xây dựng, đây là đề tài *"Nghiên cứu hoàn thiện chương trình hỗ trợ xử lý chữ viết của một số dân tộc thiểu số vùng Tây Nguyên bằng phần mềm TayNguyenKey"* (2006). *TayNguyenKey* cho phép gõ được 6 thứ tiếng DTTS Tây Nguyên: Êđê, Jarai, Bahnar, Sédang, KơHo và M'Nông. Tương tự, còn có Bộ gõ VNKEY do tác giả Trần Thanh Bình, Công ty Cao Su Đăk Lăk xây dựng, VNKEY cũng cho phép gõ tiếng Việt và một số ngôn ngữ của dân tộc thiểu số Việt Nam như: Êđê, Jarai, M'Nông, K'Ho, Xê đăng, Sán Chỉ…

Ngoài ra, còn một số bộ gõ tiếng DTTS khác đã có, tuy nhiên hầu hết các bộ gõ hiện nay vẫn còn một số nhược điểm: Các bộ gõ thường sử dụng bộ mã chuẩn có sẵn là ASCII 8 bit và tạo phông chữ theo bộ mã này. Do phương pháp mã hóa sử dụng các ký tự dựng sẵn nên bộc lộ nhiều hạn chế. Nhiều bộ gõ chưa xây dựng chức năng bỏ dấu bằng phím tắt và sử dụng các vĩ lệnh (macro) để tra từ thông dụng theo nguyên tắc hệ thống mở, để cho phép người sử dụng cập nhật dữ liệu trong các chức năng soạn thảo văn bản (STVB) quen thuộc như AutoText, AutoCorrect trong WinWord [6]. Khắc phục nhược điểm này, Nhóm tác giả Phan Huy Khánh, Hoàng Thị Mỹ Lệ, Vilavong Souksan đưa ra giải pháp: *Mã hóa tiếng Êđê sử dụng Unicode ứng dụng trong soạn thảo văn bản tiếng dân tộc* (Tạp chí Hội thảo Khoa học CNTT và Ứng dụng trong các lĩnh vực, Trường Cao đẳng CNTT, Đại học Đà Nẵng, Số 1 [6/2012]), trong đó đề xuất giải pháp mã hoá sử dụng Unicode và xây dựng bộ phông chữ tương ứng, từ đó đề xuất giải pháp sử dụng các bộ gõ tiếng Việt thông dụng trong môi trường MS Windows Office như Vietkey, Unikey… phục vụ STVB tiếng DTTS nói chung và tiếng Êđê nói riêng [6].

2.2.2. Từ điển điện tử phương ngữ

Từ điển điện tử ngoài việc hỗ trợ tra cứu từ, giúp học tập, tìm hiểu văn hóa cũng như một ngôn ngữ của một quốc gia, nó còn là một trong những cở sở dữ liệu (CSDL) cơ bản, cần thiết cho việc xử lý ngôn ngữ tự động bằng máy tính. Việc XLNNTN bằng máy tính bao gồm nhiều bài toán khác nhau, như: Phân tích hình thái, cú pháp, ngữ nghĩa cho các cấp độ từ, ngữ, câu, văn bản. Nhưng tất cả các công việc xử lý đó đều cần truy cập đến CSDL từ điển điện tử (CSDL về từ trong ngôn ngữ đó) [1]. Như vậy, việc tiên quyết cho mọi bài toán xử lý ngôn ngữ chính là cần phải xây dựng được từ điển điện tử có thể "đọc" được (MRD: Machine Readable Dictinary) [1] [11].

Hiện tại, Trung tâm CNTT-TT Sở Thông tin và Truyền thông Gia Lai - Công ty TNHH Công nghệ Tin học Tuổi trẻ Lạc Việt đã xây dựng được một số từ điện điện tử phương ngữ, như: *Từ điển điện tử phương ngữ Việt – Jrai, Từ điển điện tử phương ngữ Việt - Bahnar (Bana)*. Các từ điển phương ngữ này chủ yếu đang được phục vụ cho mục đích học tập ngôn ngữ, nghiên cứu văn hóa các đồng bào DTTS. Cụ thể, các bộ Từ điển điện tử phương ngữ này gồm 3 từ điển chính là tiếng Việt – tiếng của DTTS, tiếng DTTS – tiếng Việt và từ điển hình ảnh được cấu thành từ các thành phần như bộ từ vựng gồm tập hợp của những từ có cùng tính chất ngữ pháp, kèm theo đó là các phương ngữ, các từ trái nghĩa, từ đồng nghĩa, các ví dụ mẫu minh họa cho từ tra cứu và các hình ảnh, đoạn phim về văn hóa của đồng bào DTTS [14] [14] [15].

2.2.3. Kho ngữ liệu song ngữ Việt – DTTS

Trong XLNNTN một tài nguyên rất cần thiết đó là các kho ngữ liệu song ngữ song song (parallel corpus). Các kho ngữ liệu song ngữ song song này được sử dụng cho nhiều mục đích khác nhau, như: Nghiên cứu ngôn ngữ học so sánh, tìm kiếm thông tin xuyên ngôn ngữ, dịch tự động… Đây là nguồn tài nguyên để các ứng dụng có thể học các tương ứng của các đơn vị ngôn ngữ (từ, ngữ, câu, đoạn, văn bản…) của hai ngôn ngữ, từ đó giải quyết các vấn đề liên quan. Kết quả của các bài toán về XLNNTN ở trên phụ thuộc rất nhiều vào độ lớn và chất lượng của kho ngữ liệu song song được sử dụng [1] [10] [13]. Để nghiên cứu, xử lý tiếng DTTS tất nhiên cần phải có các kho ngữ liệu song song tiếng Việt – tiếng DTTS.

Hiện nay, đã có một số công trình nghiên cứu, đề xuất các giải pháp xây dựng kho ngữ liệu song song tiếng Việt – tiếng DTTS, cụ thể như sau: Phan Huy Khanh, *Nghiên cứu tích hợp phát triển các giải pháp xây dựng, khai thác và ứng dụng các kho ngữ liệu tiếng Việt-Kinh và tiếng Việt thiểu số*, Đề tài cấp Bộ, mã số 203706 (2008). Hoàng Thị Mỹ Lệ - Phan Huy Khánh, *Giải pháp xây dựng kho ngữ liệu đa ngữ Việt – Êđê gán nhãn theo ngữ cảnh*, Tạp chí KH&CN, Đại học Đà Nẵng – Số 1 (74).2014. Quyển II. Lê Thị Anh Đào, *Nghiên cứu xây dựng kho ngữ vựng song ngữ Việt – Khmer,* Luận văn thạc sĩ chuyên ngành Khoa học máy tính, Đại học Đà Nẵng (2013). Đỗ Thị Thuận, *Nghiên cứu và xây dựng hệ thống dịch tự động Jrai - Việt và Việt – Jrai,* Luận văn thạc sĩ chuyên ngành Khoa học máy tính, Đại học Đà Nẵng – 2012. Bên cạnh các giải pháp xây dựng kho ngữ liệu song song Việt – DTTS, các nhà nghiên cứu còn có các chương trình minh họa, đánh giá kho ngữ liệu song song Việt – DTTS đã xây dựng.

Nhìn chung, các công trình nghiên cứu, giải pháp xây dựng kho ngữ liệu song song Việt – DTTS có gán nhãn, các chương trình minh họa, đánh giá kho ngữ liệu song

song Việt - DTTS đã đạt được một số kết quả quan trọng, làm cơ sở và nền tảng quan trọng cho việc nghiên cứu, xử lý tiếng DTTS ở Việt Nam, đặc biệt tiếng của đồng bào các DTTS ở khu vực miền Trung và Tây Nguyên. Kết quả đạt được của các công trình này có thể nêu ra như sau:

1. Tương tác với kho ngữ liệu tiếng Việt để tạo kho ngữ liệu đa ngữ Việt – DTTS.

2. Gán nhãn theo ngữ cảnh và tần suất xuất hiện với mỗi từ trong kho ngữ liệu đa ngữ Việt – DTTS.

3. Bổ sung vào kho ngữ liệu tiếng Việt các từ chưa có góp phần nâng cao chất lượng của kho ngữ liệu.

4. Xây dựng công cụ tra từ vựng Việt – DTTS đáp ứng nhu cầu học tập, giảng dạy tiếng DTTS cũng như người DTTS muốn học tiếng Việt.

5. Cài đặt các cơ sở lý thuyết về các phương pháp dịch tự động, đặc biệt là phương pháp dịch máy thống kê trên kho ngữ liệu song song Việt – DTTS, đây là phương pháp được áp dụng rất nhiều trong các hệ thống dịch tự động hiện nay.

6. Cài đặt thành công một số phần mềm mã nguồn mở: Moses, Giza++, SRILM để xây dựng mô hình dịch máy thống kê và ứng dụng dịch máy cho cặp ngôn ngữ Việt - DTTS.

2.3. Đánh giá chung

Hiện tại, việc nghiên cứu, xử lý tiếng DTTS ở Việt Nam, đặc biệt việc nghiên cứu, xử lý tiếng DTTS ở khu vực miền Trung và Tây Nguyên đã có một số kết quả quan trọng, làm cơ sở, nền tảng cho quá trình nghiên cứu xử lý tiếng DTTS. Tuy nhiên, quá trình nghiên cứu, xử lý tiếng DTTS vẫn còn một số hạn chế, có thể nêu ra như sau:

1. Quá trình nghiên cứu còn độc lập, chưa có tính kế thừa.

2. Hầu hết bộ gõ tiếng DTTS chưa đưa được bộ mã chữ tiếng DTTS vào bảng mã Unicode, cũng như chưa có bộ mã chuẩn tiếng DTTS theo bảng mã Unicode.

3. Các từ điển điện tử phương ngữ Việt - DTTS hiện có chủ yếu dành cho người sử dụng, chưa có từ điển điện tử dành cho máy tính sử dụng trong xử lý tiếng DTTS.

4. Kho ngữ liệu song song Việt - DTTS chỉ dừng lại ở một số kết quả nghiên cứu của một số trường đại học, đó là những đề tài tốt nghiệp đại học, thạc sĩ, mang tính chất tìm hiểu, chưa hệ thống và định hướng rõ ràng [5]. Việc nghiên cứu, xây dựng kho ngữ liệu chưa hướng đến việc chuẩn hóa và giải pháp thu thập dữ liệu cho kho ngữ liệu này.

5. Một số chương trình minh họa dịch tự động, khai thác kho ngữ liệu song song Việt – DTTS đã xây dựng chủ yếu sử dụng các bộ công cụ mã nguồn mở có sẵn.

Trên đây là những đánh giá chung về tình hình nghiên cứu, xử lý tiếng DTTS, qua đó cho thấy việc nghiên cứu, xử lý tiếng DTTS còn rất nhiều vấn đề cần phải giải quyết, đặc biệt các vấn đề hỗ trợ chuyển đổi giữa tiếng Việt sang một số tiếng của các đồng bào DTTS.

3. Đề xuất

Việc xử lý tiếng DTTS ở Việt Nam, đặc biệt xử lý tiếng DTTS ở miền Trung và Tây Nguyên hiện nay theo như đánh giá chung còn nhiều vấn đề cần phải quyết. Với mục tiêu *giúp hỗ trợ chuyển đổi giữa tiếng Việt sang một số tiếng của đồng bào DTTS*, nhóm tác giả nêu một số đề xuất thực hiện để có thể đạt được mục tiêu vừa nêu như sau:

1. Cần phải đưa được toàn bộ bộ mã chữ tiếng các DTTS vào bảng mã unicode.

2. Thừa kế các kết quả nghiên cứu, xử lý tiếng DTTS để giải quyết các bài toán cơ bản trong xử lý tiếng DTTS, cụ thể các bài toán như: *Tiền xử lý, phân tích hình thái, phân tích ngữ pháp, phân tích ngữ nghĩa và phân tích ngữ dụng* [1].

3. Cần xây dựng và tổ chức CSDL giúp quá trình "DTTS hóa các phần mềm", giúp tự động chuyển đổi giao diện phần mềm bằng tiếng Việt sang giao diện tiếng DTTS.

4. Cần xây dựng từ điển điện tử dành cho máy tính sử dụng trong xử lý tiếng DTTS.

5. Cần có giải pháp và cách thức thu thập cụ thể để xây dựng kho ngữ liệu song song Việt – DTTS bao gồm nhiều lĩnh vực, có chú thích, được gióng hàng theo nhiều mức khác nhau, chuẩn hóa, có cấu trúc, định dạng…dễ dàng khai thác, sử dụng cho nhiều mục đích khác nhau.

6. Xây dựng hệ thống dịch tự động văn bản giữa tiếng Việt – tiếng DTTS.

4. Bàn luận

Xử lý tiếng DTTS đang là vấn đề nóng bỏng, với mục đích xóa đi "rào cản" về mặt ngôn ngữ giữa tiếng Việt và tiếng của các DTTS, giúp việc tuyên truyền các chủ trương, chính sách của Đảng và Nhà nước đến với đồng bào DTTS một cách nhanh chóng và thuận lợi hơn. Để thực hiện được các mục tiêu đó, cần thiết phải giải quyết một số đề xuất đã nêu trong nội dung bài báo.

4.1. Khó khăn

XLNNTN nói chung, xử lý tiếng DTTS nói riêng là một trong những bài toán khó khăn nhất của ngành khoa học máy tính trong hơn 50 qua [1]. Xử lý tiếng DTTS có những khó khăn riêng, có thể nêu ra như sau:

1. Nguồn dữ liệu hiện có của từ điển Việt – DTTS ở dạng tài liệu giấy, hoặc ở dạng tệp tin văn bản, các tác giả không dùng chung phông chữ unicode có hỗ trợ tiếng Việt, hầu hết đều xây dựng bộ phông chữ riêng để sử dụng [5] [7] [8], đều này sẽ gặp khó khăn trong việc xây dựng từ điển điện tử dành cho máy tính để xử lý tiếng DTTS, khó khăn trong quá trình sử dụng lại nguồn dữ liệu từ điển hiện có.

2. Nguồn tài liệu song ngữ Việt – DTTS chủ yếu trên các văn bản giấy, tài liệu điện tử được tải lên mạng Internet của các bài giảng tiếng DTTS đa số ở định dạng. PDF. Hiện nay, vẫn chưa có website song ngữ Việt – DTTS hoặc sử dụng tiếng DTTS, nhiều văn bản giấy và văn bản điện tử tiếng DTTS hiện có ở một số cơ quan, tổ chức chưa được chia sẻ và cũng chưa có nơi để tập hợp, chia sẻ các tài liệu này… Điều này làm khó khăn trong quá trình thu thập dữ liệu để xây dựng kho ngữ liệu song song Việt – DTTS.

3. Kết quả nghiên cứu, xử lý tiếng DTTS còn ít, nhiều bài toán cơ bản về xử lý tiếng DTTS chưa được giải quyết (nêu trong nội dung *Đánh giá chung* về tình hình nghiên cứu tiếng DTTS).

4. Việc kế thừa các kết quả nghiên cứu, xử lý tiếng DTTS còn khó khăn, điều này do nhiều nguyên nhân, trong

đó nguyên nhân chính là do quá trình nghiên cứu chưa được hệ thống và chưa có định hướng rõ ràng, các kết quả nghiên cứu chưa hướng đến vấn đề chuẩn hóa, chưa xác định mục tiêu đầy đủ (cụ thể như kho ngữ liệu song song Việt - DTTS). Hiện vẫn chưa có nơi để trao đổi, chia sẻ, phát triển các kết quả nghiên cứu, xử lý tiếng DTTS.

4.2. Tính khả thi

Mặc dù việc nghiên cứu, xử lý tiếng DTTS, đặc biệt giải quyết bài toán về chuyển đổi văn bản tiếng Việt sang văn bản tiếng DTTS sẽ gặp khó khăn như đã nêu, tuy nhiên vẫn có những mặt thuận lợi, giúp quá trình nghiên cứu, xử lý tiếng DTTS đạt được những kết quả mong muốn, mang tính khả thi cao. Các yếu tố thuận lợi đó có thể liệt kê như sau:

1. Tiếng của một số DTTS (Êđê, Bahnar…) có nhiều đặc trưng giống với tiếng Việt, cụ thể: Ngôn ngữ đơn lập, từ đơn tiết, ngữ pháp mang đặc điểm cơ cấu ngữ pháp của ngôn ngữ đơn lập, các dấu câu (dấu phẩy, dấu chấm, dấu chấm hỏi…) giống tiếng Việt. Bên cạnh đó, do vốn từ tiếng DTTS không nhiều, nên thường mượn một số từ của tiếng Việt để biểu thị…[3] [14]. Đây là những điểm giống nhau, giúp thừa kế được những kết quả nghiên cứu, xử lý tiếng Việt vào xử lý tiếng DTTS.

2. Hiện nay, có nhiều kết quả nghiên cứu, xử lý tiếng Việt thành công, nhiều bài toán cơ bản trong xử lý tiếng Việt đã được giải quyết…bên cạnh các điểm tương đồng giữa tiếng Việt và tiếng DTTS, có thể thừa kế và áp dụng cho nghiên cứu và xử lý tiếng DTTS.

3. Trên thế giới, việc XLNNTN cũng có rất nhiều công trình, đạt được những kết quả mong đợi. Đặc biệt, các kết quả nghiên cứu này được công bố và chia sẻ cho mục đích nghiên cứu…điều này giúp thừa kế, áp dụng trong nghiên cứu, xử lý tiếng DTTS.

4.3. Lợi ích về mặt kinh tế - xã hội

Nghiên cứu, xử lý tiếng DTTS, đặc biệt nghiên cứu hỗ trợ chuyển đổi văn bản tiếng Việt sang văn bản tiếng DTTS, sẽ mang lại nhiều lợi ích to lớn về mặt KT-XH cho nước ta, nhất là đối với đồng bào DTTS. Các kết quả nghiên cứu này giúp việc tuyên truyền các chủ trương, chính sách của Đảng và Nhà nước đến với đồng bào DTTS thuận lợi và dễ dàng hơn. Bên cạnh đó, giúp việc nghiên cứu văn hóa các DTTS cũng như giúp đồng bào DTTS bảo tồn và phát huy bản sắc văn hóa của từng dân tộc, giúp nâng cao trình độ dân trí và chất lượng cuộc sống, phát triển KT-XH ở các khu vực đồng bào DTTS.

5. Xây dựng Hệ thống dịch tự động văn bản Việt – Êđê

Nhằm minh họa tính khả thi của các đề xuất trên, trong khuôn khổ bài báo này, nhóm tác giả tập trung triển khai giải pháp để thực hiện một trong các nội dung đã đề xuất, đó là: *Xây dựng hệ thống dịch tự động văn bản giữa tiếng Việt – tiếng DTTS* (cụ thể là tiếng DTTS Êđê). Các giải pháp để thực hiện các nội dung đề xuất còn lại, nhóm tác giả sẽ thực hiện và công bố trong các công trình khác.

5.1. Giải pháp xây dựng hệ thống dịch tự động

Hệ thống dịch tự động Việt – Êđê do nhóm tác giả xây dựng dựa trên hệ thống mã nguồn mở Moses, sử dụng ưu điểm vượt trội của cách tiếp cận dịch máy dựa trên thống kê [1] [11] [12].

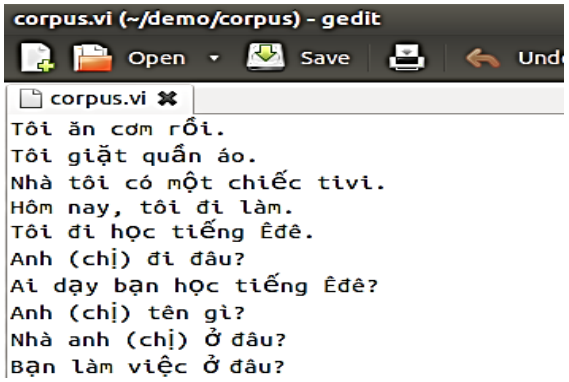
Moses kết hợp với sự đóng góp của các công cụ hoàn chỉnh khác, đang được sử dụng rộng rãi [12], nhiều chức năng phù hợp cho cặp ngôn ngữ Việt – Êđê và có khả năng mở rộng được cho nhiều cặp ngôn ngữ Việt - DTTS khác.

5.2. Các bước thực hiện

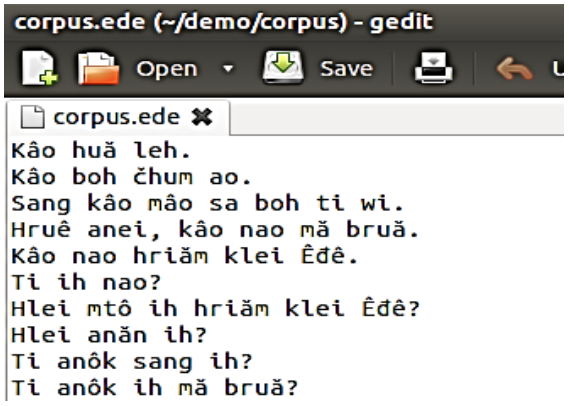
5.2.1. Xây dựng kho ngữ liệu song song Việt – Êđê

Hiện nay, nguồn tài liệu song ngữ Việt – Êđê hầu hết có trên tài liệu giấy, chưa có kho ngữ liệu đơn ngữ cũng như song ngữ Việt – Êđê được chia sẻ. Do vậy, quá trình thu thập dữ liệu để xây dựng kho ngữ liệu song song Việt – Êđê được thực hiện thủ công: Từ các giáo trình, sách song ngữ Việt – Êđê…tác giả chọn lọc, nhập vào máy tính và tổ chức thành 2 tệp tin:

- *Tệp tin thứ nhất*: Gồm các câu bằng tiếng Êđê.
- *Tệp tin thứ hai*: Lưu trữ bảng dịch bằng tiếng Việt các câu tương ứng trong tệp tin thứ nhất.



Hình 1. *Nội dung tệp tin tiếng Việt*



Hình 2. *Nội dung tệp tin tiếng Êđê*

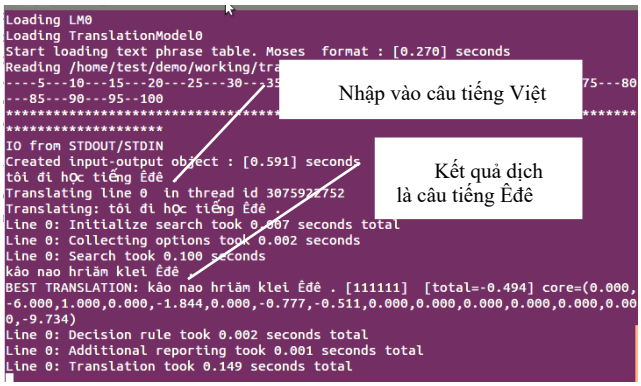
5.2.2. Cài đặt Moses và các công cụ đi kèm

- Cài đặt và cấu hình Hệ thống Moses.
- Cài đặt công cụ xây dựng mô hình ngôn ngữ: IRSTLM.
- Cài đặt công cụ gióng hàng và mô hình dịch: GIZA++, mkcls.

5.3. Kết quả thử nghiệm

Hệ thống dịch tự động Việt – Êđê, cùng với kho ngữ liệu song song do nhóm tác giả xây dựng (khoảng 10.000 cặp câu Việt – Êđê) có khả năng dịch được các câu đơn giản từ tiếng Việt sang tiếng Êđê.

Hình 3 là giao diện của Hệ thống dịch Việt – Êđê, được xây dựng dựa trên Hệ thống dịch máy thống kê mã nguồn mở Moses.



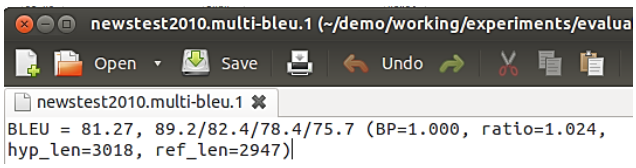
Hình 3. *Giao diện của Hệ thống dịch Việt - Êđê*

5.4. Đánh giá chất lượng hệ thống dịch đã xây dựng

Để đánh giá hệ thống dịch Việt – Êđê đã xây dựng được, chúng tôi đánh giá chất lượng thông qua phương pháp đánh giá BLEU.

Moses cung cấp tiện ích EMS (Experiment Management System) giúp tự động đánh giá chất lượng của hệ thống dịch dựa trên phương pháp đánh giá BLEU. Quá trình thực hiện đánh giá gồm nhiều bước: Chuẩn bị dữ liệu huấn luyện, xây dựng mô hình ngôn ngữ và mô hình dịch, điều chỉnh, kiểm tra, hiển thị và phân tích kết quả [4] [11] [12].

Hình 4 là giao diện Kết quả đánh chất lượng hệ thống dịch Việt – Êđê do nhóm tác giả xây dựng.



Hình 4. *Giao diện kết quả đánh giá theo BLEU*

6. Kết luận

Nghiên cứu, xử lý tiếng DTTS, đặc biệt nghiên cứu chuyển đổi văn bản tiếng Việt sang văn bản tiếng DTTS đang là vấn đề nóng bỏng, mặc dù sẽ gặp nhiều khó khăn, phải giải quyết rất nhiều bài toán đặt ra. Tuy nhiên, việc thừa kế các kết quả nghiên cứu trong lĩnh vực XLNNTN đã có trên thế giới, các kết quả xử lý tiếng Việt, dựa vào những điểm tương đồng giữa tiếng DTTS và tiếng Việt… điều này thể hiện tính khả thi rất cao. Minh họa cho tính khả thi của bài toán xử lý tiếng DTTS, nhóm tác giả đã xây dựng thành công Hệ thống dịch tự động Việt – Êđê cùng với kho ngữ liệu song song (khoảng 10.000 cặp câu Việt – Êđê). Kết quả của bài toán nghiên cứu, xử lý tiếng DTTS giúp giải quyết khó khăn do sự khác biệt về ngôn ngữ giữa tiếng Việt và tiếng DTTS, giúp gìn giữ, phát huy bản sắc văn hóa của các dân tộc. Bên cạnh đó, giúp thực hiện dễ dàng và thuận tiện hơn các chủ trương, chính sách của Đảng và Nhà nước đối với đồng bào DTTS, đồng thời giúp nâng cao dân trí, phát triển KT-XH ở các vùng đồng bào DTTS.

TÀI LIỆU THAM KHẢO

[1] Đinh Điền, 2006, *Xử lý ngôn ngữ tự nhiên*, NXB Đại học Quốc gia TP. Hồ Chí Minh.

[2] Phan Thị Tươi, 2012, "Xử lý ngôn ngữ tự nhiên", NXB Đại học Quốc gia TP. Hồ Chí Minh.

[3] Đoàn Văn Phúc, Tạ Văn Thông, "Ngữ pháp tiếng Êđê", NXB Giáo dục Việt Nam.

[4] Võ Trung Hùng, "Phương pháp và công cụ đánh giá tự động các hệ thống dịch tự động trên mạng", *Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng*, số 1 (18), 2007.

[5] Hoàng Thị Mỹ Lệ, Phan Huy Khánh, Giải pháp xây dựng kho ngữ liệu đa ngữ Việt – Êđê gán nhãn theo ngữ cảnh, *Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng* – Số 1(74).2014. Quyển II, 2014.

[6] Hoàng Thị Mỹ Lệ, Phan Huy Khánh, Vilavong Souksan, Mã hóa tiếng Êđê sử dụng Unicode ứng dụng trong soạn thảo văn bản tiếng dân tộc, *Tạp chí Hội thảo Khoa học CNTT và Ứng dụng trong các lĩnh vực, Trường Cao đẳng CNTT, Đại học Đà Nẵng*, Số 1 [6/2012]).

[7] Phan Thị Thu Nhân, 2013, *Xây dựng kho ngữ vựng song ngữ Việt-Êđê trong xử lý tiếng Êđê*, Đại học Đà Nẵng.

[8] Đỗ Gia Trinh, 2013, *Xây dựng kho dữ liệu song ngữ Việt - Cơ Tu phục vụ tra cứu văn hóa dân tộc Cơ Tu*, Đại học Đà Nẵng.

[9] Phan Thị Hà, 2013, *Nghiên cứu việc xây dựng, chuẩn hóa và khai thác kho ngữ liệu từ nguồn Internet cho xử lý tiếng Việt, Học viện Bưu chính Viễn thông.*

[10] Hồ Tú Bảo, Lương Chi Mai, *Về xử lý tiếng Việt trong công nghệ thông tin*, http://www.jaist.ac.jp/~bao/Writings/VLSPwhitepaper%20-%20 Final.pdf

[11] Hồ Bảo Quốc, Đinh Điền, Đặng Bác Văn, Lương Vỹ Ninh, Phạm Đào Duy Vũ, 2008, *Xây dựng kho ngữ liệu song ngữ Anh – Việt.*

[12] http://www.statmt.org/

[13] http://vietlex.com/

[14] http://violet.vn/buonmathuot/present/show/entry_id/9223067

[15] http://tudienjrai.gialai.gov.vn/Default.aspx

[16] http://tudienbahnar.gialai.gov.vn/

(BBT nhận bài: 27/08/2014, phản biện xong: 11/11/2014)