

NGÔN NGỮ CÁC DÂN TỘC THIỂU SỐ Ở VIỆT NAM

TƯ LIỆU HÓA TIẾNG PÀ THÈN Ở VIỆT NAM BẰNG PHẦN MỀM ELAN: KINH NGHIỆM THỰC ĐỊA VÀ MỘT SỐ ĐỀ XUẤT TIẾP CẬN

NGUYỄN THU QUỲNH* - NGUYỄN VĂN HUÂN**

TÓM TẮT: Bài viết trình bày quy trình và kết quả bước đầu của việc tư liệu hóa tiếng Pà Thên - một ngôn ngữ thiểu số có nguy cơ mai một ở Việt Nam - thông qua việc ứng dụng phần mềm ELAN để phân lớp và chú thích đa tầng các dữ liệu đa phương tiện. Nghiên cứu cũng nêu bật vai trò của công nghệ số trong tư liệu hóa ngôn ngữ, đồng thời thảo luận những thuận lợi, khó khăn khi triển khai thu thập và xử lý dữ liệu theo các yêu cầu của phần mềm ELAN trong thực địa, từ thu thập, phiên âm, phiên dịch, chú thích ngôn ngữ đến tổ chức dữ liệu theo chuẩn quốc tế. Từ đó, bài viết đề xuất hướng tiếp cận thực địa ngôn ngữ phù hợp trong bối cảnh Việt Nam nhằm đảm bảo hiệu quả và tính bền vững của các dự án tư liệu hóa ngôn ngữ.

TỪ KHÓA: tư liệu hóa ngôn ngữ; tiếng Pà Thên; ngôn ngữ dân tộc thiểu số; ngôn ngữ nguy cấp; phần mềm ELAN; bảo tồn ngôn ngữ.

NHẬN BÀI: 08/05/2025.

BIÊN TẬP-CHỈNH SỬA-DUYỆT ĐĂNG: 16/07/2025

1. Đặt vấn đề

Theo *Catalogue of Endangered Languages (ELCat)*, hàng trăm ngôn ngữ đã không còn được sử dụng trong vòng 60 năm qua. Sự mất mát ngôn ngữ đang diễn ra nhanh hơn bất kì thời điểm nào trong lịch sử [Belew, A., & Simpson, S., 2018]. Hiện nay, có khoảng hơn 40% ngôn ngữ trên thế giới đang có nguy cơ tiêu vong [<https://www.ethnologue.com>], 46% trong tất cả các ngôn ngữ đang bị đe dọa [Campbell, L., & Okura, E., 2018]. Các hệ thống cảnh báo ngôn ngữ đều cho rằng nguy cơ ngôn ngữ bị đe dọa đang diễn ra mạnh mẽ và ảnh hưởng đến sự đa dạng ngôn ngữ trên toàn cầu [<https://www.unesdoc.unesco.org>], nguy cơ tiêu vong ngôn ngữ là một vấn đề nghiêm trọng, đặt ra những thách thức về mặt đạo đức, thực tiễn và khoa học [<https://www.endangeredlanguages.com/>].

Dân tộc Pà Thên với số dân là 8.248 người [<https://tongdieutradanso.vn/>] thuộc nhóm dân tộc thiểu số rất ít người ở Việt Nam. Tiếng Pà Thên [Pa-Hng] (ISO 639-3: pha), theo các nhà nghiên cứu trên thế giới và ở Việt Nam, cũng là một trong những ngôn ngữ có nguy cơ mai một cao. UNESCO đánh giá tiếng Pà Thên là ngôn ngữ “dễ tổn thương” (Vulnerable) [<https://www.unesdoc.unesco.org>]. Nguyễn Văn Lợi (2012) cho rằng, tiếng Pà Thên “chỉ dùng trong phạm vi gia đình”; Tạ Văn Thông (2017) xếp tiếng Pà Thên vào nhóm “ngôn ngữ nguy cấp”. Điều đó cho thấy, với số lượng người sử dụng ngôn ngữ ít (dưới 10.000 người), đặc biệt đang có xu hướng giảm dần, nhất là ở thế hệ trẻ, việc bảo tồn ngôn ngữ này ngày càng trở thành một yêu cầu cấp thiết hiện nay.

Tư liệu hóa ngôn ngữ (Language Documentation) là một lĩnh vực nhận được nhiều sự quan tâm của các nhà ngôn ngữ học. Một số nhà nghiên cứu đã đề cập đến vấn đề này như Woodbury, Anthony C. (2003), Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (2006), Harrison, K. David, David S. Rood & Arianne Dwyer (2008), Peter K. Austin (2010, 2013), Mari C. Jones, Sarah Ogilvie (2013), Henke, Ryan and Andrea L. Berez-Kroeker (2016),... Với việc thu thập, ghi chép, lưu trữ, xử lý, phân tích ngôn ngữ, tư liệu hóa ngôn ngữ được coi là một trong những giải pháp hữu hiệu để lưu giữ, bảo tồn ngôn ngữ trước nguy cơ ngôn ngữ, văn hóa của các dân tộc thiểu số đang bị mai một hiện nay. Việc tư liệu hóa tạo ra một công cụ quan trọng giúp lưu trữ các yếu tố văn hóa, truyền thống và tri thức bản địa của cộng đồng sở hữu đồng thời giúp cung cấp tư liệu cho các nhà ngôn ngữ học, dân tộc học có tư liệu nghiên cứu. Sản phẩm của hoạt động này cũng có thể được sử dụng để tạo ra các tài liệu giảng dạy, phục vụ cho hoạt động học tập của cộng đồng sử dụng ngôn ngữ và các chương trình hồi sinh ngôn ngữ. Nó cũng có thể tạo ra các nền tảng pháp lí quan trọng bằng việc

*TS; Trường Đại học Sư phạm - Đại học Thái Nguyên; Email: quynhn@tnue.edu.vn

**Trường Đại học Sư phạm - Đại học Thái Nguyên; Email: huannnguyen07082004@gmail.com

cung cấp chứng cứ cho việc xây dựng các chính sách bảo vệ ngôn ngữ và quyền của người nói ngôn ngữ đó.

Trên cơ sở bước đầu thu thập tư liệu tiếng Pà Thên trong giao tiếp gia đình tại các xã Tân Bắc, Tân Trinh, Yên Thành thuộc huyện Quang Bình, tỉnh Hà Giang, bài báo hướng đến mục tiêu thử nghiệm đề xuất phương án tư liệu hóa tiếng Pà Thên bằng phần mềm ELAN như một nghiên cứu trường hợp; trên cơ sở đó thảo luận về mục tiêu của việc xác lập cơ sở dữ liệu, phương pháp xác lập cơ sở dữ liệu, quy trình thu thập dữ liệu ngôn ngữ, xử lý dữ liệu phiên âm, phiên dịch, chú thích ngôn ngữ; nhập dữ liệu ngôn ngữ đã xử lý vào phần mềm ELAN.

2. Đối tượng, phạm vi, phương pháp nghiên cứu

Đối tượng nghiên cứu: Các vấn đề liên quan đến việc tư liệu hóa tiếng Pà Thên trong giao tiếp gia đình bằng phần mềm ELAN như: cơ sở dữ liệu ngôn ngữ, phương pháp xác lập cơ sở dữ liệu ngôn ngữ, quy trình xác lập cơ sở dữ liệu ngôn ngữ, kết quả xử lý dữ liệu phiên âm, phiên dịch, chú thích ngôn ngữ; kết quả nhập dữ liệu phiên âm, phiên dịch, chú thích ngôn ngữ vào phần mềm ELAN.

Phạm vi nghiên cứu: Phạm vi khảo sát được xác định là các cuộc nói chuyện, giao tiếp tự nhiên đời thường của người Pà Thên. Địa điểm khảo sát tại các xã Tân Bắc, Tân Trinh, Yên Thành, huyện Quang Bình, tỉnh Hà Giang. Đây là nơi tập trung đông người Pà Thên sinh sống nhất ở Việt Nam hiện nay. Thời gian khảo sát được thực hiện trong tháng 11/2024.

Cách thu thập: Dữ liệu ngôn ngữ được thu thập bằng cách ghi âm, quay video những cuộc nói chuyện, giao tiếp tự nhiên đời thường của người Pà Thên. Bên cạnh đó, nghiên cứu cũng tiến hành thu thập thông tin cá nhân và phỏng vấn người tham gia để hiểu rõ hơn về ngữ cảnh và ý nghĩa của cuộc hội thoại. Sau khi đã được cộng tác viên người bản địa đồng ý tham gia cung cấp tư liệu, quy trình thu thập tư liệu được thực hiện gồm 5 bước: (1). Chuẩn bị, lựa chọn môi trường ghi âm, ghi hình; (2). Thiết lập các thiết bị ghi âm; (3). Bố trí phương tiện ghi hình; (4). Đồng bộ âm thanh và hình ảnh; (5). Xử lý bản ghi.

Phương pháp nghiên cứu: Phương pháp nghiên cứu tài liệu được sử dụng để thu thập tư liệu, phân tích văn bản... liên quan đến nghiên cứu về tiếng Pà Thên và về vấn đề tư liệu hóa ngôn ngữ. Phương pháp điều tra, điền dã ngôn ngữ học được áp dụng trong quá trình thu thập dữ liệu để quan sát, ghi chép, ghi âm, ghi hình tư liệu ngôn ngữ sống trong phạm vi các cuộc nói chuyện, giao tiếp gia đình của người Pà Thên.

3. Kết quả nghiên cứu

3.1. Vấn đề sử dụng phần mềm ELAN trong tư liệu hóa ngôn ngữ

3.1.1. Giới thiệu chung về phần mềm ELAN

ELAN (EUDICO Linguistic Annotator) là một phần mềm chuyên dụng trong lĩnh vực ngôn ngữ học tài liệu do Viện Max Planck về Ngôn ngữ học Tâm lý (Max Planck Institute for Psycholinguistics) tại Nijmegen, Hà Lan phát triển. Ra mắt vào năm 2000, ELAN đã trải qua nhiều phiên bản cập nhật nhằm nâng cao tính năng, chất lượng sử dụng và khắc phục lỗi. Phiên bản mới nhất hiện nay là ELAN 6.9 (năm 2024) với nhiều cải tiến về hiệu suất và chức năng chú thích.

ELAN được xây dựng để trở thành công cụ cho phép người sử dụng chú thích (Annotation) trực tiếp trên các bản ghi âm thanh và video, phục vụ cho mục đích phân tích ngôn ngữ học, nghiên cứu nhân học và dân tộc học, nghiên cứu ngôn ngữ ký hiệu, ngôn ngữ học tâm lý,... Với phần mềm này, người dùng có thể chú thích đa tầng trên các bản ghi, tức là người dùng có thể tạo nhiều tầng chú thích. Các chú thích có thể ở nhiều cấp độ khác nhau như câu, từ, cụm từ, hình vị,... Bằng cách căn chỉnh theo thời gian (Timeline), các chú thích sẽ xuất hiện đúng với vị trí mà người dùng cần chú thích, các chú thích xuất hiện lần lượt từ trái qua phải và theo trình tự thời gian từ điểm bắt đầu đến điểm kết thúc của bản ghi.

3.1.2. Đặc điểm của phần mềm ELAN

- *Công cụ chú thích đa tầng*: ELAN là phần mềm chuyên dụng cho phép người dùng thực hiện chú thích thủ công trên các tệp âm thanh và video. Quá trình chú thích được thực hiện thông qua các thao tác trực tiếp trên từng “phân đoạn” của bản ghi. Phần mềm hỗ trợ tạo các chú thích đa tầng, trong đó mỗi tầng có thể được thiết lập quan hệ bao hàm hoặc độc lập tùy theo mục đích nghiên cứu. Ví dụ, tầng chú thích trên cùng là tầng chú thích ở cấp độ câu, tầng chú thích ở những cấp độ tiếp theo là cụm từ, từ, hình vị,... Các tầng chú thích cũng có thể được tạo ra để đối chiếu ngôn ngữ, chẳng hạn, tầng chú thích bên trên sử dụng tiếng Việt, tầng chú thích bên dưới sử dụng tiếng Anh, Đức, Pháp, Nga,...

- *Tính năng kiểm soát từ vựng*: ELAN tích hợp chức năng ghi nhớ từ vựng đã dùng ở các tầng chú thích trước đó (có thể ở các tầng khác nhau) và áp dụng lại với những đối tượng tương đương hoặc có liên quan. Điều này đảm bảo tính nhất quán trong quá trình chú thích, đồng thời các từ vựng này có thể được thay đổi, xóa bỏ tùy thuộc vào ý định của người dùng.

- *Khả năng chia sẻ và làm việc nhóm*: Các tệp chú thích của ELAN có thể dễ dàng chia sẻ giữa nhiều người dùng, miễn là họ có cài đặt phần mềm về máy tính, laptop và có lưu giữ các bản ghi (âm thanh, video) đã được đưa vào ELAN để chú thích. Tính năng này cho phép nhiều cá nhân cùng tham gia chỉnh sửa và phát triển trên cùng một tệp dữ liệu.

- *Yêu cầu định dạng*: ELAN không giới hạn số lượng tệp đưa vào để chú thích, tuy nhiên phần mềm có yêu cầu nghiêm ngặt về định dạng của dữ liệu đầu vào. Các tệp âm thanh bắt buộc phải là định dạng .WAV và các tệp video phải là định dạng .MP4. Chỉ khi tệp đáp ứng yêu cầu, phần mềm mới cho phép tiến hành các thao tác chú thích tiếp theo.

- *Khả năng liên thông với các phần mềm khác*: ELAN hỗ trợ xuất dữ liệu dưới nhiều định dạng khác nhau XML, CSV, CHAT, tương thích với nhiều phần mềm hỗ trợ nghiên cứu ngôn ngữ khác: PRAAT, EXMARaLDA, FLEX,... và hỗ trợ đa ngôn ngữ, người dùng có thể chú thích bằng bất kỳ hệ chữ viết nào kể cả chữ tượng hình.

3.1.3. Vai trò của phần mềm ELAN trong tư liệu hóa ngôn ngữ

Phần mềm ELAN được tạo ra nhằm cung cấp cơ sở khoa học trong việc chú thích, lập tài liệu và khai thác các bản ghi đa phương tiện. Các vai trò cụ thể của phần mềm ELAN trong việc tư liệu hóa ngôn ngữ được thể hiện như sau:

- *Đồng bộ hóa các dữ liệu đa phương tiện*: Phần mềm hỗ trợ liên kết bản ghi âm (audio) và ghi hình (video) tương ứng, đảm bảo dữ liệu được chú thích, lưu trữ một cách đồng bộ, có hệ thống.

- *Hỗ trợ nghiên cứu ngôn ngữ học kí hiệu*: Với khả năng chú thích theo khung hình, ELAN còn trở thành công cụ đắc lực trong lĩnh vực nghiên cứu ngôn ngữ học kí hiệu (Sign Language Documentation). Người dùng có thể đánh dấu, giải thích các yếu tố: cử chỉ, chuyển động của bàn tay, biểu cảm khuôn mặt,...

- *Hỗ trợ lưu trữ và chia sẻ dữ liệu chuẩn hóa*: Thông thường, trong quá trình làm việc các nhà nghiên cứu không sử dụng một phần mềm duy nhất mà sử dụng nhiều phần mềm cùng lúc. Chính vì vậy, việc liên kết, chia sẻ các dữ liệu cho những phần mềm khác nhau là thao tác thường gặp. Để hỗ trợ điều này, phần mềm xuất dữ liệu dưới dạng XML, CSV, CHAT,... tương thích với các phần mềm hỗ trợ nghiên cứu ngôn ngữ khác như PRAAT, FLEX (FieldWorks Language Explorer), Toolbox,... Ngoài ra, các tệp đã được chú thích bởi ELAN dễ dàng tích hợp vào các kho dữ liệu số, ví dụ như kho TLA (The Language Archive) hay ELAR (Endangered Languages Archive).

- *Ứng dụng trong bảo tồn và phục hồi ngôn ngữ*: Với sự hỗ trợ của ELAN, các nhà khoa học có thể tạo ra một bộ hồ sơ về ngôn ngữ nguy cấp bao gồm các bản ghi đã được chú thích, các bản ghi chép tay trong quá trình thu thập dữ liệu, các bản dịch,... Trên cơ sở đó, các nhà ngôn ngữ học có thể xây dựng từ điển hoặc bộ chữ cho những ngôn ngữ chưa có chữ viết kết hợp với các giải pháp để bảo tồn và phục hồi ngôn ngữ nguy cấp.

3.2. Nghiên cứu trường hợp: Tư liệu hóa ngôn ngữ trong giao tiếp gia đình của người Pà Thẻn ở huyện Quang Bình tỉnh Hà Giang bằng phần mềm ELAN

3.2.1. Mục tiêu của việc xác lập cơ sở dữ liệu ngôn ngữ

Việc thử nghiệm xác lập cơ sở dữ liệu ngôn ngữ trong giao tiếp gia đình của người Pà Thẻn hướng đến một số mục tiêu sau:

- *Khắc phục những hạn chế của phương pháp lưu trữ thủ công*: Việc xác lập cơ sở dữ liệu ngôn ngữ nhằm thay thế phương pháp lưu trữ thủ công vốn còn tồn tại nhiều bất cập như: dữ liệu trùng lặp, phân tán, dễ bị mất mát hoặc hư hỏng. Cơ sở dữ liệu số hóa cho phép lưu trữ tập trung, tối ưu hóa không gian lưu trữ, đồng thời hỗ trợ sao lưu, nhân bản, chia sẻ dữ liệu một cách dễ dàng và an toàn hơn.

- *Tích hợp các bản ghi đã được phiên âm, phiên dịch, chú thích ngôn ngữ*: Trong bối cảnh nghiên cứu và tư liệu hóa ngôn ngữ hiện nay, một cơ sở dữ liệu ngôn ngữ không chỉ đơn thuần là nơi lưu trữ các bản ghi âm thanh và video, mà còn là tập hợp các dữ liệu đã được xử lý, phân tích một cách có hệ thống nhằm nâng cao giá trị khoa học và ứng dụng thực tiễn. Việc thu thập và xác lập thử nghiệm cơ sở dữ liệu này hướng tới việc triển khai các bước phiên âm, phiên dịch và sử dụng phương pháp chú thích xen kẽ (Interlinear Glossing Text), một kỹ thuật thường được dùng trong ngôn ngữ học mô tả, nhằm phân tích ý nghĩa của từng hình vị trong chuỗi lời nói và xác định từ loại, dưới sự hỗ trợ của cộng tác viên người Pà Thẻn.

- *Số hóa tư liệu ngôn ngữ Pà Thẻn*: Việc xây dựng một cơ sở dữ liệu ngôn ngữ có cấu trúc khoa học là điều kiện tiên quyết để tiến hành nhập các tệp âm thanh và video vào phần mềm ELAN, từ đó thực hiện việc gán các lớp chú thích một cách hệ thống. Hơn thế nữa, cơ sở dữ liệu ngôn ngữ còn đóng vai trò là kho lưu trữ các dữ liệu đã được phân tích thông qua ELAN, tạo điều kiện thuận lợi cho việc truy xuất và chỉnh sửa thông tin khi cần thiết.

- *Hướng tới kết nối với hệ thống lưu trữ của khu vực và thế giới*: Một trong những mục tiêu dài hạn của việc tư liệu hóa tiếng Pà Thẻn là đảm bảo khả năng tích hợp dữ liệu với các kho lưu trữ ngôn ngữ nguy cấp trên thế giới như ELAR, PARADISEC, hoặc Corpus. Vì lẽ đó, trong quá trình thu thập, nghiên cứu thử nghiệm này cũng tiến hành xác lập cơ sở dữ liệu một cách bài bản, có quy trình và phù hợp với các tiêu chí đề ra của các kho dữ liệu.

- *Bảo tồn ngôn ngữ “nguy cấp”, “dễ bị tổn thương”*: Xây dựng cơ sở dữ liệu là một giải pháp bền vững để lưu giữ những bản ghi âm và ghi hình. Những dữ liệu quý giá này là cơ sở, nền tảng giúp các nhà khoa học, ngôn ngữ học tiếp tục nghiên cứu khám phá ra những đặc trưng của tiếng Pà Thẻn.

- *Phục hồi ngôn ngữ nguy cấp*: Sản phẩm của quá trình phân tích dữ liệu ngôn ngữ là bộ từ vựng, các đặc trưng ngữ âm, các quy tắc ngữ pháp,... Trên cơ sở đó, nghiên cứu hướng đến xây dựng bộ chữ Latin phù hợp với hệ thống âm vị của tiếng Pà Thẻn, biên soạn giáo trình, sách giáo khoa để đưa vào giảng dạy. Kết hợp với các biện pháp sư phạm, tổ chức dạy học cho học sinh dân tộc Pà Thẻn ở lứa tuổi tiểu học, phổ thông tạo điều kiện để tiếng Pà Thẻn được “sống” trong chính môi trường mà nó sinh ra.

- *Kết hợp công nghệ hiện đại*: Trong quá trình nghiên cứu, để phân tích ngôn ngữ cần sử dụng nhiều phần mềm chuyên dụng như PRAAT, FLEEx, ELAN,... Vì vậy, việc thiết lập một cơ sở dữ liệu giúp quản lý hiệu quả các tệp dữ liệu đã qua phân tích, đảm bảo tính nhất quán, đồng bộ trong quá trình nhập, xuất và chuyển đổi dữ liệu linh hoạt từ phần mềm này sang phần mềm khác. Hơn nữa, khi đã hoàn thiện hệ thống chữ viết Latin, cơ sở dữ liệu có thể được tích hợp với các công nghệ hiện đại nhằm phát triển bộ gõ tiếng Pà Thẻn, thậm chí tiến tới xây dựng phần mềm nhận dạng giọng nói. Nhờ đó, tiếng nói và chữ viết của cộng đồng dân tộc này không còn bị giới hạn trong môi trường sinh hoạt gia đình mà còn có cơ hội được sử dụng rộng rãi hơn, mở rộng phạm vi giao tiếp.

3.2.2. Phương pháp xác lập cơ sở dữ liệu ngôn ngữ

- *Thu thập thông tin cá nhân*: Nhóm điều tra đã tiến hành tiếp cận người dân và tiến hành thu thập thông tin cá nhân của người tham gia thông qua hình thức phỏng vấn với những câu hỏi ngắn (họ tên, tuổi, quê, quán, nghề nghiệp,...). Trong quá trình này, nếu phát hiện thông tin không rõ ràng, điều tra

viên sẽ xác minh lại bằng cách hỏi trực tiếp người tham gia hoặc người thân của họ để đảm bảo chính xác.

- *Thu thập dữ liệu hội thoại*: Nhóm điều tra tiến hành bố trí các thiết bị ghi âm và máy quay/ điện thoại thông minh để thu lại các cuộc hội thoại tự nhiên trong giao tiếp gia đình của người Pà Thẻn. Kết thúc buổi ghi âm, điều tra viên trao đổi lại với người tham gia, làm rõ nội dung, chủ đề của các cuộc hội thoại, giúp việc phân loại, xây dựng cơ sở dữ liệu thuận lợi hơn.

- *Ghi chép nhanh*: Trong quá trình ghi âm, ghi hình, điều tra viên đồng thời sử dụng sổ tay ghi chú nhanh các thông tin quan trọng như bối cảnh, địa điểm, thời gian và số lượng người tham gia hội thoại. Những thông tin này đóng vai trò quan trọng trong việc phân loại dữ liệu và tái hiện bối cảnh nghiên cứu một cách chính xác. Để đảm bảo tính khách quan, điều tra viên cần ghi chép trung thực và bảo mật các thông tin trước khi công bố nghiên cứu.

3.2.3. Quy trình thu thập dữ liệu ngôn ngữ tại hiện trường

Dựa trên kinh nghiệm thực địa từ quá trình thu thập dữ liệu, nhóm nghiên cứu đưa ra quy trình khuyến nghị cho công tác điều tra điền dã thu thập dữ liệu ngôn ngữ gồm 5 bước như sau:

a. Bước 1: Chuẩn bị, lựa chọn môi trường ghi âm

Điều tra viên nên chọn môi trường ghi âm lý tưởng, ưu tiên các phòng kín, có vật liệu giảm tiếng vang (thảm dày, rèm cửa, tủ quần áo,...) hoặc bố trí các vật dụng có thể giảm tiếng vang như hộp giấy đựng trứng, các tấm bìa các-tông cứng, dày,... Tiến hành đóng cửa sổ, cửa ra vào để hạn chế tiếng ồn ngoại cảnh (tiếng xe, tiếng động vật) làm giảm chất lượng bản ghi. Ngoài ra, chúng ta nên cố gắng giảm tối đa người trong phòng (tránh trẻ em và những người không tham gia trực tiếp vào công tác thu thập).

Thêm vào đó, chúng ta cần loại bỏ những nguồn gây nhiễu như: tiếng quạt, các thiết bị điện tử có thể phát ra âm thanh: điện thoại, tivi, đồng hồ,... và yêu cầu người tham gia không thực hiện hoạt động ăn, nhai trong quá trình ghi âm, ghi hình.

Tuy nhiên, rất khó để kiểm soát môi trường xung quanh, vì vậy các điều tra viên cần linh hoạt với điều kiện thực tế, cố gắng lựa chọn góc, khu vực yên tĩnh nhất, sử dụng các bộ lọc nhiễu nếu có để buổi ghi âm, ghi hình đạt hiệu quả, thu về những dữ liệu tốt nhất.

b. Bước 2: Thiết lập thiết bị ghi âm

Tránh tình trạng ghi âm nhưng không thể dùng được bản ghi đó, chúng ta cần phải kiểm tra và điều chỉnh trước khi ghi bằng những cách sau: thử nghiệm với vài câu nói ngắn để cân chỉnh âm lượng, đảm bảo bản ghi không bị rè, vỡ âm hoặc quá cao, đặt micro cách miệng người nói từ 20 - 50cm, tránh hướng trực tiếp vào hơi thở,... Sử dụng máy ghi âm chuyên dụng, điều chỉnh lại các thông số kỹ thuật: tần số mẫu 44.1 kHz, độ phân giải 16 bit, định dạng .wav.

c. Bước 3: Bố trí máy quay

Cần bố trí máy quay ngang tầm mắt người tham gia, góc quay bao quát tất cả người tham gia hội thoại và đảm bảo ghi lại tất cả biểu lộ trên gương mặt và các cử chỉ, chuyển động tay. Để tránh rung lắc, hiện tượng mờ tay khi ghi hình trong thời gian dài, chúng ta cần sử dụng chân máy (Tripod) để giữ khung hình cố định.

Trong trường hợp không có đầy đủ dụng cụ, thiết bị kỹ thuật số hiện đại, chuyên dụng, chúng ta có thể sử dụng điện thoại để thu thập dữ liệu. Điều tra viên cần phải đặt điện thoại ở một cự li thích hợp đảm bảo cả hai thao tác ghi âm và ghi hình. Tương tự khi sử dụng các thiết bị chuyên dụng, chúng ta nên thực hành ghi âm, ghi hình và đánh giá sản phẩm thử nghiệm, đặt điện thoại lên những mặt phẳng ổn định hạn chế rung lắc, đưa điện thoại về chế độ máy bay hoặc chế độ riêng tư để không có bất kỳ thông báo, cuộc gọi nào gián đoạn quá trình thu thập và hãy kết hợp với micro nếu có.

d. Bước 4: Đồng bộ âm thanh và hình ảnh

Để đạt được bộ dữ liệu chất lượng cao, các điều tra viên có thể kết hợp máy quay và máy ghi âm riêng, bởi lẽ micro của máy quay có phần hạn chế hơn so với máy ghi âm khi không thể giảm thiểu tiếng ồn và khi đặt ở xa, chất lượng âm thanh có thể không đủ to, rõ ràng. Chúng ta ghi lại tiếng vỗ tay, búng tay đánh dấu những đoạn ghi âm, ghi hình để dễ dàng ghép âm thanh và hình ảnh sau này.

đ. Bước 5: Xử lý bản ghi

Thông thường, máy quay và máy ghi âm có nhiều định dạng mà người dùng có thể lựa chọn, tuy nhiên trong trường hợp điều tra viên sử dụng điện thoại hoặc các thiết bị chưa đáp ứng các thông số kỹ thuật, chúng ta có thể sử dụng một số công cụ hỗ trợ sau: những phần mềm chuyển đổi định dạng: HandBrake, FFmpeg (giữ nguyên chất lượng bản ghi gốc), những phần mềm chỉnh sửa video và xử lý video: Audacity (loại bỏ tạp âm, cân bằng âm lượng), DaVinci Resolve, OpenShot,...

3.2.4. Quy trình xác lập cơ sở dữ liệu ngôn ngữ

a. Bước 1: Thiết kế khung cơ sở dữ liệu

Để hệ thống hóa dữ liệu thu thập từ quá trình điền dã, nhóm nghiên cứu đã xây dựng cấu trúc của hệ thống cơ sở dữ liệu, đảm bảo tính đồng bộ, dễ dàng truy xuất và chia sẻ dữ liệu giữa các thành viên. Cơ sở dữ liệu bao gồm các thành phần chính sau:

- Mỗi thư mục (pha001, pha002,...) đại diện cho một cuộc thoại riêng biệt, số thứ tự của thư mục tương ứng với thứ tự cuộc hội thoại trong thực tế.

- Tập âm thanh (pha001.wav, pha002.wav,...) và tập ghi hình (pha001.mp4, pha002.mp4,...) thu lại toàn bộ nội dung hội thoại, đã đồng bộ hóa âm thanh và hình ảnh.

- Tập hình ảnh (pha001-01.jpg, pha001-02.jpg,...) chứa ảnh chân dung của những người tham gia cuộc hội thoại đó.

- Tập chú thích (pha001_20250326.eaf) là tập dữ liệu ngôn ngữ đã được phiên âm, dịch nghĩa trong phần mềm ELAN.

- Thông tin siêu dữ liệu (Metadata) tập hợp các dữ liệu mô tả về cuộc hội thoại, bao gồm: thời gian thu thập, thời lượng, thông tin người tham gia, bối cảnh giao tiếp...

b. Bước 2: Thu thập cơ sở dữ liệu

+ *Tiếp cận người tham gia*: Nhóm điều tra liên hệ với chính quyền 3 xã Yên Thành, Tân Bắc, Tân Trịnh để giới thiệu mục đích nghiên cứu và xin phép tiến hành công tác điền dã. Trong quá trình làm việc, nhóm đã trình bày rõ mục tiêu của việc thu thập dữ liệu ngôn ngữ là nhằm phục vụ cho việc phiên âm, phiên dịch và chú thích các ngữ liệu, từ đó góp phần khám phá và làm sáng tỏ những đặc trưng của ngôn ngữ Pà Thẻn trong bối cảnh giao tiếp gia đình. Nhóm nghiên cứu cũng cam kết rằng toàn bộ bản ghi âm và ghi hình chỉ được sử dụng cho mục đích nghiên cứu khoa học, tuyệt đối không dùng vào bất kỳ mục đích nào khác. Người tham gia được thông tin đầy đủ, rõ ràng về nội dung và mục tiêu của nghiên cứu, đồng thời xác nhận sự tham gia của họ hoàn toàn dựa trên tinh thần tự nguyện.

+ *Thu thập dữ liệu*: Nhóm điều tra tiến hành ghi lại các cuộc hội thoại tự nhiên hoặc sử dụng một số chủ đề định hướng (nếu cần) trong bối cảnh sinh hoạt hằng ngày của người dân. Bố trí thiết bị ghi âm, ghi hình sao cho các nhân vật tham gia hội thoại đều có mặt đầy đủ trong khung hình và chất lượng âm thanh ở mức tốt nhất có thể, đảm bảo các thiết bị kỹ thuật không ảnh hưởng đến sự tự nhiên trong giao tiếp, tạo ra cảm giác gượng gạo cho người tham gia.

+ *Thu thập thông tin cá nhân*: Kết thúc buổi ghi hình, các điều tra viên thu thập các thông tin cơ bản của người tham gia (tên, tuổi, địa chỉ, nghề nghiệp, mối quan hệ với người cùng tham gia hội thoại) để phục vụ cho phân tích và lưu trữ dữ liệu.

+ *Phiên âm và phiên dịch*: Nhóm điều tra sử dụng những cộng tác viên (người dân tộc Pà Thẻn, có bố và mẹ đều là người Pà Thẻn, sinh sống tại huyện Quang Bình, tỉnh Hà Giang) nghe lại và phiên âm tiếng Việt nội dung hội thoại thành các văn bản. Nhằm đảm bảo tính chính xác của dữ liệu thu thập được, nhóm điều tra tiến hành kiểm tra chéo thông tin với hai cộng tác viên, nhằm hạn chế các sai lệch có thể xảy ra do cộng tác viên không nghe hoặc không hiểu được từ, cụm từ hay câu nói bất kỳ trong hội thoại. Trong quá trình trao đổi, nhiều trường hợp cộng tác viên Lưu Thị Chuyên, thuộc thế hệ trẻ, không nhận diện được một số từ ngữ do chúng ít được sử dụng hoặc ít xuất hiện trong giao tiếp hằng ngày. Để xử lý tình huống này, điều tra viên đã tiến hành tham vấn thêm cộng tác viên Sìn Thị Tuyết, người có vốn hiểu biết sâu hơn về ngôn ngữ dân tộc mình và có nhiều kinh nghiệm cộng

tác với các nhà ngôn ngữ học trước đó. Ngoài ra, sự khác biệt trong quá trình phiên âm và dịch nghĩa của 2 cộng tác viên cũng mở ra cơ hội nghiên cứu sâu hơn về ngữ âm, ngữ pháp của tiếng Pà Thẻn.

+ *Lưu trữ và mã hóa dữ liệu*: Để duy trì trạng thái tư liệu tốt nhất trong suốt quá trình nghiên cứu và lưu trữ sau này, nhóm tiến hành lưu bản ghi hình (video), ghi âm (audio) dưới dạng file chất lượng cao .MP4 và .WAV và đồng thời tạo thêm 1 bản sao để tránh trường hợp sơ suất làm mất, thất lạc dữ liệu.

Trước khi lưu trữ các dữ liệu, nhóm điều tra tiến hành mã hóa liệu theo quy định chung của ELAR. Mỗi tên tệp dữ liệu bao gồm tên cơ sở và định dạng tệp. Ví dụ như: pha001.mp4, pha001.wav, pha001.eaf,... Đối với tên cơ sở, chúng ta nên đặt tên theo mã ISO của ngôn ngữ đó (tiếng Pà Thẻn có mã ISO: pha). Trong trường hợp ngôn ngữ chưa có mã ISO, tên cơ sở có thể là tên viết tắt của ngôn ngữ đó. Để phân biệt các tệp của cùng một ngôn ngữ, chúng ta cần thêm số thứ tự vào sau mã ISO, ví dụ như pha001, pha002, pha003,... Tùy vào số lượng dữ liệu thu thập được mà chúng ta quyết định số chữ số thêm vào, có thể là 3 chữ số hoặc 4 chữ số.

Đặc biệt, đối với các bản ghi đã được chú thích bằng ELAN, chúng ta cần thêm ngày/ tháng/ năm, theo cấu trúc “tên cơ sở_yy/mm/dd” (ví dụ: pha001_20250326). Điều này giúp hỗ trợ tìm kiếm, kiểm soát bộ sưu tập, đánh dấu ngày cập nhật, bổ sung vào bản chú thích.

Các tệp bắt nguồn từ cùng một bản ghi âm hoặc video phải có cùng tên. Nếu tiến hành chú thích bản ghi bằng ELAN thì dữ liệu sau khi chú thích cũng được lưu ở thư mục chung này. Những tệp nằm trong cùng một thư mục phải có phần mở rộng (định dạng) khác nhau. Ví dụ:

Thư mục:	pha001.	Tệp video:	pha001.mp4
Tệp ghi âm:	pha001.wav.	Tệp chú thích:	pha001_20250326.eaf
Đối với các tệp ảnh sẽ có dấu gạch nối và một số thứ tự khác nhau (số thứ tự có 2 chữ số). Ví dụ:			
Thư mục:	pha001	Tệp ảnh:	pha001-01.jpg
Tệp ảnh:	pha001-02.jpg	Tệp ảnh:	pha001-03.jpg

+ *Đảm bảo đạo đức nghiên cứu*: Nhóm nghiên cứu cam kết chỉ sử dụng dữ liệu cho mục đích nghiên cứu, không công khai nếu không được cho phép, trong trường hợp người tham gia yêu cầu xóa bỏ những dữ liệu liên quan đến họ, nhóm nghiên cứu bắt buộc phải làm theo mong muốn của người tham gia.

c. Bước 3: Ghi chú và phân tích cơ sở dữ liệu

Ghi chú và phân tích cơ sở dữ liệu là thao tác then chốt để chuyển từ dữ liệu thô sang dữ liệu đạt chuẩn được sắp xếp, xây dựng có hệ thống. Dựa theo số lượng bản ghi thu thập được, nhóm nghiên cứu tiến hành phân tích theo ba giai đoạn sau:

+ *Phân tích mô tả (Descriptive)*: Sau khi kết thúc điền dã, điều tra viên tiến hành tổng hợp và phân loại các chủ đề xuất hiện trong các bản ghi âm thanh và video. Quá trình này bao gồm: tóm tắt, ghi lại những chủ đề, nội dung chính thường xuyên xuất hiện trong giao tiếp của người Pà Thẻn và thống kê tần suất xuất hiện chủ đề.

+ *Phân tích văn bản (Textual Analysis)*: Dựa trên văn bản dịch nghĩa do công tác viên cung cấp, người nghiên cứu tiến hành đối chiếu với chủ đề, so sánh với kết quả phân tích mô tả để phát hiện ra các chủ đề, nội dung chưa được ghi nhận và bổ sung thêm vào cơ sở dữ liệu.

+ *Phân tích khám phá (Exploratory Analysis)*: Nhóm tiến hành rà soát chất lượng dữ liệu thô để phát hiện và xử lý các vấn đề kỹ thuật. Cần kiểm tra các lỗi như dính tạp âm, âm lượng thấp, chất lượng khung hình chưa cao, khung hình không đầy đủ người nói,... để có những biện pháp khắc phục: lọc âm thanh, cân bằng âm, khuếch đại âm thanh, cắt bỏ những đoạn ghi âm, ghi hình chưa đạt yêu cầu hoặc có tiếng, hình ảnh của điều tra viên,... Đưa các tệp dữ liệu về định dạng tiêu chuẩn .MP4, .WAV, trong trường hợp thiết bị ghi âm và điện thoại không hỗ trợ điều chỉnh định dạng.

d. Bước 4: Quản lý cơ sở dữ liệu

Quản lý cơ sở dữ liệu không chỉ là thao tác kỹ thuật mà còn góp phần vào tính bền vững của lĩnh vực tư liệu hóa ngôn ngữ nói riêng và các ngành khoa học dữ liệu nói chung. Dưới đây là một số nguyên tắc chính được áp dụng để quản lý cơ sở dữ liệu ngôn ngữ trong giao tiếp ở phạm vi gia đình:

+ *Kiểm soát chất lượng*: Sau khi thiết kế và xây dựng cơ sở dữ liệu, điều tra viên cần kiểm tra tính thống nhất giữa các tệp đa phương tiện (tệp ghi âm, tệp ghi hình), tệp chú thích và thông tin thư mục.

+ *Bảo mật và truy cập*: Điều tra viên cần tiến hành chia sẻ quyền truy cập theo các cấp độ như: công khai, hạn chế (chỉ cho những nhà khoa học, nhà nghiên cứu truy cập), riêng tư (chỉ cho những người tham gia vào dự án truy cập).

+ *Tích hợp đa phần mềm*: Ưu tiên những định dạng tệp phổ biến, liên kết các tệp với những phần mềm ELAN, PRAAT, FLEX,... để tiếp tục phân tích và nghiên cứu.

+ *Xây dựng công cụ hỗ trợ*: Xây dựng cơ sở dữ liệu đi kèm với các công cụ hỗ trợ và bộ lọc để tìm kiếm dữ liệu bằng từ khóa, thời gian, địa điểm,...

3.2.5. Kết quả xử lý dữ liệu phiên âm, phiên dịch và chú thích

Một số quy ước trong quá trình chú thích xen kẽ đã được sử dụng theo bộ quy tắc chuẩn Leipzig Glossing Rules để chú thích ngôn ngữ:

N (noun): Danh từ; V (Verb): Động từ; ADJ (Adjective): Tính từ

PRO (Pronoun): Đại từ; CONJ (Conjunction): Kết từ; NUM (Numeral): Số từ

PART (Particle): Trợ từ; ADV (Adverb): Phó từ; 1-: Đại từ nhân xưng ngôi thứ nhất

2-: Đại từ nhân xưng ngôi thứ hai; 3-: Đại từ nhân xưng ngôi thứ ba

Trong nghiên cứu này, văn bản chú thích xen kẽ bao gồm 5 dòng:

- Dòng thứ nhất: câu gốc tiếng Pà Thẻn được phiên âm tiếng Việt.

- Dòng thứ hai: câu gốc tiếng Pà Thẻn phiên âm theo quy tắc.

- Dòng thứ ba: nghĩa từng hình vị.

- Dòng thứ tư: tách từng hình vị, gán nhãn ngữ pháp

- Dòng thứ năm: Câu được dịch nghĩa tiếng Việt

Dưới đây là kết quả phiên âm, phiên dịch và chú thích xen kẽ nội dung một đoạn hội thoại ngắn pha001, trong đó A, B lần lượt là tên người tham gia hội thoại:

A: hi chong bu gơ cung ho pờ cơ pe le vo
 hi3 tjo3 pu4 gơ4 *kũ3 ho4 pHơ21 kơ4 pe4 le4 vô4
 Thế năm nay mình cũng trồng lúa tốt đấy nhỉ?
 PART ? 1-PRO V N N ?

Thế, năm nay mình cũng trồng lúa tốt đấy nhỉ?

B: gơ tu xo po
 gơ4 tu4 shõ3 bo3
 Cũng được 30 bao.
 ADV V NUM N

Cũng được 30 bao.

A: hi gó ụ kẹ múng pê có he chung ngươ cộ tong
 hi3 gHo24 *ũ2? kễ2? mHũ24 pjhẽ4? kHơ24 he3 tjHũ3 gwơ3 kô2? tồ3
 Thế, thế bây giờ bà con gái đã tìm bạn trai?
 PART PRO 2-PRO N ADV V N

Thế bây giờ, con gái của bà đã tìm được bạn trai chưa?

B: u pà ngươ tu le
 ũ3 pHa21 gwơ3 tu4 *lHe21
 Không thấy tìm được ?
 ADV V V V ?

Không thấy tìm được.

A: kẹ nùng tơ si nhi thơ độ
 kễ2? nHũ21 tơ3 si3 njHĩ3 thơ4 *do2?
 Bây giờ nó đang đi học
 PRO 3-PRO ADV V V

Bây giờ, nó đang đi học.

A: chung nú thơ rô nú cung

tjũ4	*nHũ21	thơ4	jHũ21	nHũ21	ũ2?	kũ4
Thả	nó	học	xong	nó	về	việc
V	3-PRO	V	V	3-PRO	CONJ	N

Kệ nó, học xong việc của nó.

A: nú	a chợ	ngươ
nHũ21	a3 tjơ2?	gươ3
Nó	tự	tìm

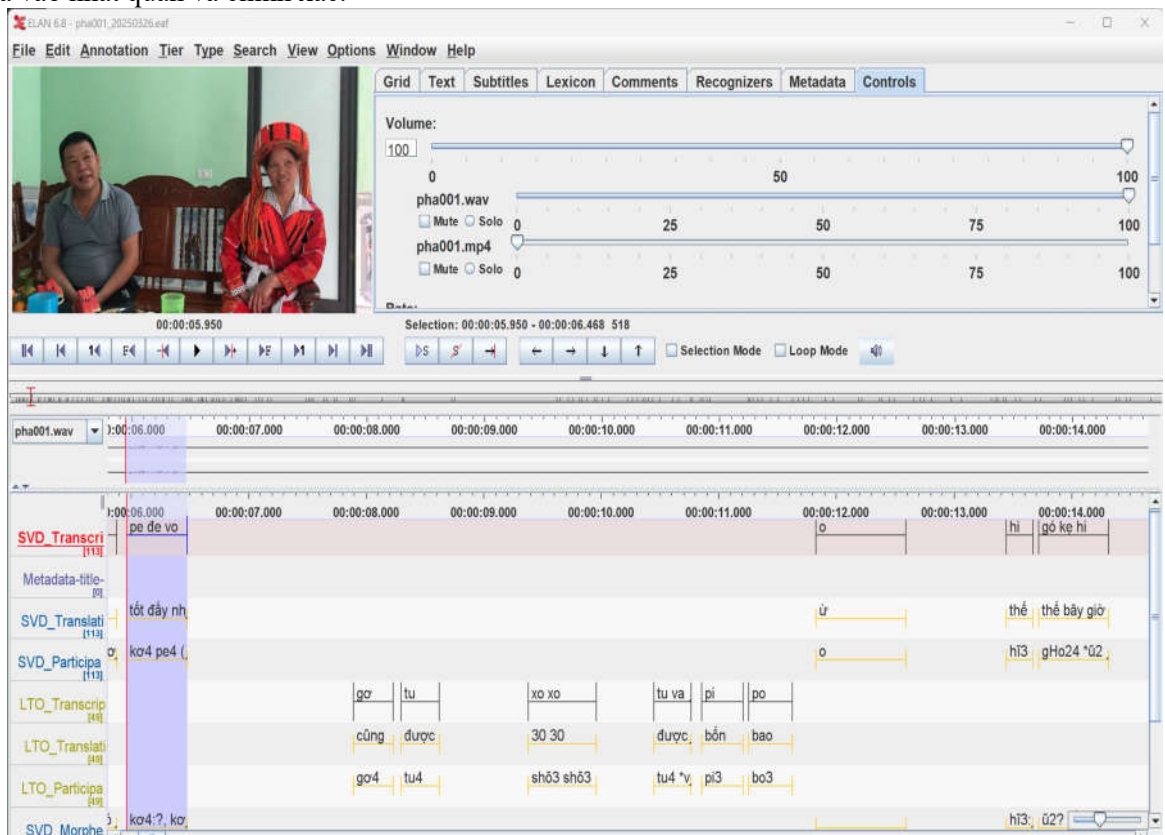
3-PRO PRO V

Nó tự tìm

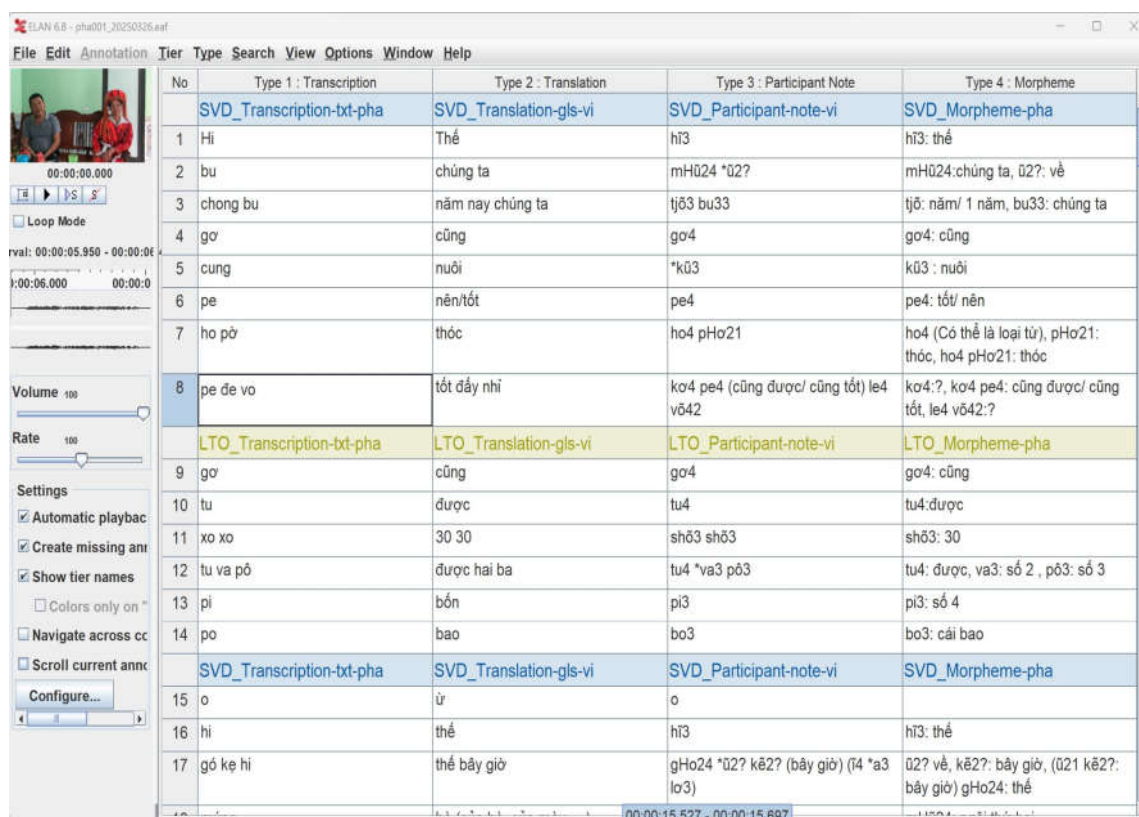
3.2.6. Nhập dữ liệu phiên âm, phiên dịch và chú thích vào phần mềm ELAN

Sau khi có kết quả phiên âm, phiên dịch và chú thích xen kẽ, người chú thích tiến hành nhập chú thích vào phần mềm ELAN thông qua các bước: tạo tệp chú thích, thay đổi các tầng, thêm người tham gia hội thoại và phân đoạn bản ghi. Trước khi tiến hành thao tác phiên âm, phiên dịch và chú thích bản ghi, người nghiên cứu cần lưu ý 2 điểm sau:

- Xây dựng hệ thống tầng chú thích và xác định mối quan hệ giữa chúng, đảm bảo mỗi tầng đảm nhận một chức năng cụ thể.
- Nhập thông tin chú thích vào các tầng tương ứng đã được thiết lập sao cho các thông tin được đưa vào nhất quán và chính xác.



Hình 1. Giao diện màn hình sau khi tiến hành kỹ thuật phân tầng (Tier) và phân đoạn (Segmentation) trên phần mềm ELAN



No	Type 1 : Transcription	Type 2 : Translation	Type 3 : Participant Note	Type 4 : Morpheme
	SVD_Transcription-txt-pha	SVD_Translation-gls-vi	SVD_Participant-note-vi	SVD_Morpheme-pha
1	Hi	Thế	hi3	hi3: thế
2	bu	chúng ta	mH024 *02?	mH024: chúng ta, 02?: về
3	chong bu	năm nay chúng ta	tj03 bu33	tj0: năm/ 1 năm, bu33: chúng ta
4	gơ	cũng	gơ4	gơ4: cũng
5	cung	nuôi	*kũ3	kũ3: nuôi
6	pe	nên/tốt	pe4	pe4: tốt/ nên
7	ho pờ	thóc	ho4 pHo21	ho4 (Có thể là loại từ), pHo21: thóc, ho4 pHo21: thóc
8	pe de vo	tốt đấy nhỉ	kơ4 pe4 (cũng được/ cũng tốt) le4 v042	kơ4?: kơ4 pe4: cũng được/ cũng tốt, le4 v042?:
	LTO_Transcription-txt-pha	LTO_Translation-gls-vi	LTO_Participant-note-vi	LTO_Morpheme-pha
9	gơ	cũng	gơ4	gơ4: cũng
10	tu	được	tu4	tu4: được
11	xo xo	30 30	sh03 sh03	sh03: 30
12	tu va pỏ	được hai ba	tu4 *va3 pỏ3	tu4: được, va3: số 2, pỏ3: số 3
13	pi	bốn	pi3	pi3: số 4
14	po	bao	bo3	bo3: cái bao
	SVD_Transcription-txt-pha	SVD_Translation-gls-vi	SVD_Participant-note-vi	SVD_Morpheme-pha
15	o	ừ	o	
16	hi	thế	hi3	hi3: thế
17	gỏ kỏ hi	thế bây giờ	gH024 *02? kỏ2? (bây giờ) (T4 *a3 lo3)	02? về, kỏ2?: bây giờ, (021 kỏ2?: bây giờ) gH024: thế

Hình 2. Giao diện màn hình sau khi tiến hành kỹ thuật phiên âm (Transcription) và chú thích (Annotation) trên phần mềm ELAN

5. Kết luận

Với tính năng là một công cụ chú thích đa tầng, ELAN cho phép đánh dấu tạo ra các phân đoạn giúp người dùng dễ dàng phiên âm, phiên dịch và chú thích ngôn ngữ. Đặc biệt, bằng cách tạo ra các tầng chú thích độc lập hoặc có mối quan hệ theo kiểu bao hàm, người dùng có thể cùng lúc phiên âm, dịch nghĩa hoặc chú thích ở nhiều cấp độ khác nhau như hình vị, từ, câu... Xét về ưu điểm, ELAN là một phần mềm có tính trực quan, dễ sử dụng, có thể xuất dữ liệu dưới nhiều định dạng và hỗ trợ phân tích ngôn ngữ có cấu trúc phức tạp. Xét về hạn chế, ELAN cũng đòi hỏi người dùng cần có những hiểu biết nhất định về font chữ IPA và cần đạt được những yêu cầu năng lực kỹ thuật nhất định. Tuy nhiên, nếu nhìn nhận một cách toàn diện ở các khả năng: chú thích đa tầng, đồng bộ hóa các tệp đa phương tiện, hỗ trợ nghiên cứu ngôn ngữ ký hiệu, hỗ trợ lưu trữ và chia sẻ dữ liệu chuẩn hóa... thì có thể khẳng định rõ ràng phần mềm ELAN là một công cụ hiệu quả để tư liệu hóa tiếng Pà Thên nói riêng và các ngôn ngữ dân tộc thiểu số ở Việt Nam nói chung, đặc biệt, đối với các ngôn ngữ chưa có chữ viết.

Dựa trên kinh nghiệm thực địa từ quá trình thu thập dữ liệu ngôn ngữ trong giao tiếp gia đình của người Pà Thên ở huyện Quang Bình, tỉnh Hà Giang, bài viết cũng đưa ra khuyến nghị cho quy trình 5 bước trong công tác thu thập, điều tra ngôn ngữ học nói chung và ngôn ngữ dân tộc thiểu số nói riêng tại hiện trường như sau: (1). Chuẩn bị, lựa chọn môi trường ghi âm; (2). Thiết lập thiết bị ghi âm; (3). Bố trí máy quay; (4). Đồng bộ âm thanh và hình ảnh; (5). Xử lý bản ghi. Đồng thời xác định quy trình 4 bước trong xác lập cơ sở dữ liệu ngôn ngữ: (1). Thiết kế khung cơ sở dữ liệu; (2). Thu thập cơ sở dữ liệu (tiếp cận người tham gia, thu thập dữ liệu, thu thập thông tin cá nhân, phiên âm, phiên dịch, lưu trữ và mã hóa dữ liệu; (3). Ghi chú và phân tích cơ sở dữ liệu (phân tích mô tả, phân tích văn bản, phân tích khám phá); (4). Quản lý cơ sở dữ liệu (kiểm soát chất lượng, bảo mật và truy cập, tích hợp đa phần mềm, xây dựng công cụ hỗ trợ).

Những nội dung nghiên cứu của bài viết về kinh nghiệm thực địa và một số đề xuất trong công tác tư liệu hóa tiếng Pà Thẻn ở Việt Nam đã mở ra hướng tiếp cận mới trong bảo tồn ngôn ngữ dân tộc thiểu số thông qua việc kết hợp giữa số hóa tư liệu và ứng dụng công nghệ hiện đại, tạo tiền đề cho các chương trình bảo tồn ngôn ngữ Pà Thẻn nói riêng và các ngôn ngữ dân tộc thiểu số khác tại Việt Nam nói chung. Những dự kiến triển khai tiếp theo có thể tập trung vào việc mở rộng phạm vi tư liệu hóa ngôn ngữ, với trọng tâm là các bối cảnh giao tiếp mang tính xã hội - văn hóa điển hình nhằm phản ánh đầy đủ hơn đặc điểm ngôn ngữ Pà Thẻn. Về định hướng ứng dụng, có thể hướng đến mục tiêu xây dựng hệ thống chữ viết cho tiếng Pà Thẻn theo định dạng Latin, tiến tới phát triển bộ gõ điện tử, qua đó hỗ trợ cộng đồng người Pà Thẻn sử dụng ngôn ngữ mẹ đẻ trong các môi trường số hóa hiện đại. Đây không chỉ là một bước tiến trong công tác bảo tồn ngôn ngữ, mà còn góp phần nâng cao vị thế văn hóa và tri thức của cộng đồng thiểu số. Những nỗ lực này có ý nghĩa thiết thực trong việc ngăn chặn nguy cơ mai một ngôn ngữ, đồng thời duy trì và làm phong phú thêm bức tranh đa dạng của ngôn ngữ các dân tộc thiểu số có nguy cơ mai một ở Việt Nam trong bối cảnh toàn cầu hóa hiện nay.

**Bài báo là sản phẩm của Đề tài KH&CN cấp Đại học Thái Nguyên, Mã số: DH2024-TN04-04*

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Nguyễn Minh Đức (1972), “Bước đầu tìm hiểu tiếng nói và vấn đề chữ viết Pà Hưng (Pà Thẻn)”, *Tìm hiểu ngôn ngữ các dân tộc thiểu số ở Việt Nam*, 1, tr.200-226.
2. Dương Thu Hằng, Nguyễn Thu Quỳnh (đồng chủ biên), Nguyễn Văn Lợi, Tạ Văn Thông (2022), *Ngôn ngữ với vấn đề bảo tồn văn hóa trong đời sống các dân tộc thiểu số*, Nxb Khoa học Xã hội, Hà Nội.
3. Nguyễn Văn Lợi (2012), “Công trình tra cứu về ngôn ngữ và vấn đề bảo tồn ngôn ngữ có nguy cơ tiêu vong”, *Tạp chí Từ điển học & Bách khoa thư*, số 2 (16).
4. Nguyễn Thu Quỳnh (2008), *Ngữ âm tiếng Pà Thẻn* (Luận văn Thạc sĩ, Trường Đại học Sư phạm - Đại học Thái Nguyên).
5. Nguyễn Thu Quỳnh (2009), “Vài nét về ngữ âm tiếng Pà Thẻn và phương án phiên âm tiếng Pà Thẻn”, *Tạp chí Ngôn ngữ và Đời sống*, 159 (1), tr.76-84.
6. Nguyễn Thu Quỳnh (2011), “Vài nét về hệ thống thanh điệu tiếng Pà Thẻn”, *Tạp chí Khoa học và Công nghệ*, 80 (4), tr.31-36.
7. Tạ Văn Thông, Tạ Quang Tùng (2017), *Ngôn ngữ các dân tộc ở Việt Nam*, Nxb Đại học Thái Nguyên.

Tiếng Anh

8. Austin, P. K. (2010), *Current issues in language documentation*. In P. K. Austin (Ed.), *Language Documentation and Description* (Vol. 7, pp.1-22). London: SOAS.
9. Austin, P. K. (2013), *Language documentation and meta-documentation*. In S. Ogilvie & M. C. Jones (Eds.), *Keeping Languages Alive* (pp.3-15). Cambridge University Press.
10. Belew, A., & Simpson, S. (2018), *Language extinction then and now*. In *Cataloguing the World's Endangered Languages* (pp.49-65). Routledge.
11. Bisang, W., Hock, H. H., Winter, W., Gippert, J., Himmelmann, N. P., & Mosel, U. (Eds.). (2006), *Essentials of Language Documentation*. Mouton de Gruyter.
12. Campbell, L., & Okura, E. (2018), *New knowledge produced by the Catalogue of Endangered Languages*. In *Cataloguing the World's Endangered Languages* (pp.79-84). Routledge.
13. Edmondson, J. A. (1992), *Pa-hng development and diversity*. In *Papers from the First Annual Meeting of the Southeast Asian Linguistics Society*, Arizona State University, pp.159-186.

14. Edmondson, J., & Lợi, N. V. (1997), *Tones and voice quality in modern northern Vietnamese: Instrumental case studies*. *Mon-Khmer Studies*, 28(35), pp.1-18.
15. Harrison, K. D., Dwyer, A., & Rood, D. S. (2008), *Lessons from documented endangered languages*.
16. Henke, R., & Berez-Kroeker, A. L. (2016), *A brief history of archiving in language documentation, with an annotated bibliography*.
17. Jones, M. C., & Ogilvie, S. (Eds.) (2013), *Keeping Languages Alive*. Cambridge University Press.
18. Woodbury, A. C. (2003). Defining documentary linguistics. *Language Documentation and Description*, 1, pp.35-51.

**Documenting the Pa Then Language in Vietnam with ELAN:
field experiences and methodological proposals**

Abstract: This article presents the process and initial results of documenting Pa Then language - an endangered minority language in Vietnam - by applying ELAN software to perform multi-layer segmentation and annotation of multimedia data. The article highlights the role of digital technology in language documentation, while discussing the advantages and challenges encountered during data collection and processing according to ELAN's requirements in fieldwork. This includes steps from data gathering, transcription, translation, linguistic annotation, to organizing data following international standards. Based on these experiences, the article proposes a suitable fieldwork approach for the Vietnamese context to ensure the efficiency and sustainability of language documentation projects.

Key words: language documentation; Pà Thên language; minority language; endangered language; ELAN software; language preservation.