NGÔN NGỮ VÀ BẢN SẮC TRONG KỶ NGUYÊN AI: ĐỊNH HƯỚNG GÌN GIỮ VÀ PHÁT HUY GIÁ TRỊ TIẾNG VIỆT

Phạm Bình Phương My*, Dương Vân Anh
Trường Đại học Đồng Tháp
***Email: pbpmy@dthu.edu.vn**

Tóm tắt: *Trong kỷ nguyên AI, tiếng Việt vừa đứng trước cơ hội mở rộng không gian sử dụng, vừa đối diện nguy cơ "chuẩn hoá lệch" do dữ liệu huấn luyện thiên lệch, dịch máy áp đặt cấu trúc ngoại lai và thói quen giao tiếp số rút gọn. Bài viết tiếp cận tiếng Việt như một thực thể văn hoá - xã hội, nơi bản sắc được kết tinh qua chuẩn mực, biến thể vùng miền, diễn ngôn cộng đồng và năng lực sáng tạo ngôn ngữ. Trên cơ sở đó, nghiên cứu đề xuất ba định hướng: (1) xây dựng hệ sinh thái dữ liệu tiếng Việt có kiểm soát chất lượng, tôn trọng đa dạng; (2) phát triển chuẩn đánh giá đầu ra AI theo tiêu chí ngôn ngữ học ứng dụng (độ đúng, độ tự nhiên, tính phù hợp ngữ dụng); (3) tăng cường giáo dục năng lực số - ngôn ngữ để người dùng biết "đồng kiến tạo" với AI.*

Từ khoá: *tiếng Việt; bản sắc ngôn ngữ; trí tuệ nhân tạo; ngữ dụng học; chuẩn hoá dữ liệu.*

LANGUAGE AND IDENTITY IN THE AI ERA: ORIENTATIONS FOR PRESERVING AND PROMOTING THE VALUE OF VIETNAMESE

Abstract*: In the AI era, Vietnamese has new opportunities to expand its domains of use, while also facing the risk of "skewed standardization" resulting from biased training data, machine translation that imposes exogenous structures, and shortened patterns of digital communication. This paper conceptualizes Vietnamese as a socio-cultural entity in which linguistic identity is crystallized through norms, regional varieties, community discourses, and language creativity. On this basis, the study proposes three orientations: (1) developing a Vietnamese language data ecosystem with quality control that respects diversity; (2) establishing evaluation standards for AI outputs grounded in applied linguistics criteria (accuracy, naturalness, and pragmatic appropriateness); and (3) strengthening digital-linguistic literacy so that users can "co-construct" meaning with AI.*

Keywords: *Vietnamese language; linguistic identity; artificial intelligence; pragmatics; data standardization.*

Nhận bài: 05/03/2026 *Phản biện: 26/03/2026* *Duyệt đăng: 29/03/2026*

I. ĐẶT VẤN ĐỀ

Sự phát triển nhanh của trí tuệ nhân tạo (AI) và các nền tảng số đã mở rộng mạnh mẽ không gian sử dụng ngôn ngữ, làm chuyển dịch hoạt động giao tiếp từ môi trường trực tiếp sang hệ sinh thái số đa phương thức. Trong bối cảnh đó, ngôn ngữ không chỉ tồn tại trong tương tác người với người mà còn hiện diện ngày càng phổ biến trong giao tiếp người - máy và máy - người thông qua các hệ thống sinh ngôn ngữ tự động, dịch máy và trợ lý hội thoại. AI vì vậy không còn đơn thuần là công cụ xử lý ký hiệu mà đã trở thành một chủ thể trung gian tham gia vào quá trình tạo lập và phân phối diễn ngôn. Các mô hình ngôn ngữ lớn với khả năng tạo văn bản lưu loát ở quy mô lớn đang tác động trực tiếp đến cách viết, đọc và diễn đạt của người dùng. Tuy nhiên, bên cạnh cơ hội mở rộng năng lực biểu đạt, AI cũng đặt ra nguy cơ tái cấu trúc thói quen ngôn ngữ, làm xuất hiện xu hướng "chuẩn hóa lệch", suy giảm sự đa dạng diễn đạt và ảnh hưởng đến bản sắc tiếng Việt do dữ liệu huấn luyện thiên lệch hoặc chịu tác động mạnh từ ngoại ngữ. Trên cơ sở đó, bài viết tập trung phân tích mối quan hệ giữa tiếng Việt, bản sắc ngôn ngữ và AI; làm rõ tác động của AI đến chuẩn mực và biến thể tiếng Việt; nhận diện cơ hội, thách thức trong gìn giữ bản sắc ngôn ngữ; đồng thời đề xuất định hướng xây dựng hệ sinh thái dữ liệu tiếng Việt chất lượng, tiêu chí đánh giá đầu ra AI và nâng cao năng lực số - ngôn ngữ cho người sử dụng.

II. NỘI DUNG NGHIÊN CỨU

2.1. Tổng quan lý thuyết

2.1.1. Khái niệm bản sắc ngôn ngữ

Trong ngôn ngữ học xã hội và ngôn ngữ học văn hoá, bản sắc ngôn ngữ được xem là kết quả tương tác giữa hệ thống chuẩn, thực hành giao tiếp và ý thức cộng đồng về ngôn ngữ. Bản sắc không chỉ nằm ở cấu trúc ngữ pháp hay từ vựng, mà được kiến tạo qua chuẩn mực sử dụng, qua lựa chọn biến thể và qua cách cộng đồng gán giá trị cho các hình thức biểu đạt. Theo quan điểm của Bucholtz và Hall, "identity is the social positioning of self and other" - bản sắc là sự định vị xã hội của chủ thể trong diễn ngôn, được hình thành thông qua thực hành ngôn ngữ cụ thể (Bucholtz & Hall, 2005)

Ở bình diện thứ nhất, bản sắc thể hiện qua chuẩn mực ngôn ngữ, bao gồm chính tả, ngữ pháp, phong cách chức năng và các quy ước diễn đạt được thể chế hoá trong giáo dục và truyền thông.

Ở bình diện thứ hai, bản sắc thể hiện qua biến thể vùng miền và xã hội, phản ánh lịch sử, địa lý và cấu trúc cộng đồng nói năng. Sự đa dạng phương ngữ không làm suy yếu hệ thống, mà góp phần tạo chiều sâu bản sắc. Ở bình diện thứ ba, bản sắc được kiến tạo qua diễn ngôn cộng đồng - nơi các nhóm xã hội sử dụng ngôn ngữ để thể hiện lập trường, giá trị và quan hệ quyền lực. Ở bình diện thứ tư, bản sắc còn gắn với năng lực sáng tạo ngôn ngữ, bao gồm ẩn dụ mới, lối nói mới và các hình thức chuyển nghĩa trong môi trường truyền thông số. Cách nhìn này cho phép tiếp cận bản sắc tiếng Việt như một cấu trúc động, đa tầng và có khả năng thích ứng.

2.1.2. Ngôn ngữ trong hệ sinh thái công nghệ số

Trong hệ sinh thái công nghệ số, ngôn ngữ vận hành đồng thời ở ba cấp độ: dữ liệu - mô hình - thuật toán. Ngôn ngữ vừa là phương tiện giao tiếp, vừa trở thành đối tượng được số hoá, gắn nhãn và đưa vào kho dữ liệu huấn luyện cho các hệ thống AI. Các mô hình ngôn ngữ lớn được xây dựng trên nguyên lý học từ mẫu thống kê trong tập văn bản quy mô lớn, qua đó tạo ra văn bản mới dựa trên xác suất chuỗi ký hiệu. Như Bender và cộng sự nhận định, mô hình ngôn ngữ "do not understand language; they generate text by modeling form, not meaning" - các mô hình này mô phỏng hình thức ngôn ngữ chứ không nắm giữ nghĩa theo cách của con người (Bender & cs., 2021).

Trong bối cảnh đó, ngôn ngữ trở thành dữ liệu huấn luyện và chịu tác động của lựa chọn nguồn dữ liệu, phương pháp tiền xử lý và tiêu chí tối ưu hoá mô hình. Một hệ quả đáng chú ý là hiện tượng trung gian hoá ngôn ngữ qua máy: nhiều văn bản được tạo, biên tập hoặc dịch thông qua hệ thống tự động trước khi đến người đọc. Quá trình này có thể làm tăng tốc độ sản xuất diễn ngôn, nhưng cũng có nguy cơ làm phẳng phong cách, giảm sắc thái ngữ dụng và tái phân bố chuẩn mực biểu đạt theo logic của mô hình.

2.1.3. Ngữ dụng học và tiêu chí đánh giá giá trị ngôn ngữ

Để đánh giá chất lượng và giá trị của văn bản trong môi trường AI, cần bổ sung góc nhìn ngữ dụng học bên cạnh tiêu chí hình thức. Ngữ dụng học nhấn mạnh mối quan hệ giữa phát ngôn, ngữ cảnh và mục đích giao tiếp, xem nghĩa là kết quả của suy diễn trong tình huống cụ thể. Theo Levinson, ngữ dụng học nghiên cứu "the relations between language and context that are basic to an account of language understanding" (Levinson, 1983).

Từ khung này, có thể xác lập ba nhóm tiêu chí đánh giá đầu ra ngôn ngữ trong hệ thống AI. Thứ nhất là độ đúng (accuracy), tức mức độ phù hợp với quy tắc ngôn ngữ và nội dung sự thật có thể kiểm chứng. Thứ hai là độ tự nhiên (naturalness), thể hiện ở tính lưu loát, quen thuộc và phù hợp với thói quen diễn đạt của người bản ngữ. Thứ ba là tính phù hợp ngữ cảnh (pragmatic appropriateness), tức mức độ tương thích với vai giao tiếp, thể loại văn bản và mục đích phát ngôn. Ba tiêu chí này cho phép chuyển trọng tâm đánh giá từ "đúng cấu trúc" sang "đúng trong sử dụng", phù hợp với định hướng ngôn ngữ học ứng dụng.

2.1.4. Khung phân tích tiếng Việt trong môi trường AI

Trên cơ sở các tiếp cận trên, bài viết đề xuất khung phân tích tiếng Việt trong môi trường AI theo hướng tích hợp chuẩn mực và đa dạng. Tiếng Việt được xem là một hệ thống vừa có chuẩn hoá chức năng, vừa có phổ biến thể vùng miền và phong cách diễn ngôn. Việc đánh giá và phát triển dữ liệu tiếng Việt cho AI vì thế không chỉ nhắm tới "một chuẩn duy nhất", mà cần phản ánh cấu trúc đa dạng có kiểm soát.

Trong tương tác người-máy, tiếng Việt không còn chỉ là công cụ biểu đạt của người nói, mà là phương tiện đối thoại với hệ thống sinh ngôn ngữ tự động. Nghĩa của phát ngôn được hình thành qua chuỗi trao đổi, hiệu chỉnh và phản hồi. Từ góc độ này, có thể vận dụng quan điểm "meaning as use" của Wittgenstein - nghĩa nằm trong cách dùng - để lý giải cơ chế đồng kiến tạo nghĩa giữa người dùng và AI trong quá trình tương tác (Wittgenstein, 1953/2009). Khung tiếp cận này cho phép đặt trọng tâm vào vai trò chủ động của người sử dụng tiếng Việt trong việc định hướng, hiệu chỉnh và làm giàu giá trị ngôn ngữ khi làm việc cùng hệ thống AI.

2.2. Tác động của AI đến tiếng Việt và bản sắc ngôn ngữ

2.2.1. Cơ hội mở rộng giá trị và phạm vi sử dụng tiếng Việt

Sự phát triển của trí tuệ nhân tạo và các nền tảng số đã tạo điều kiện mở rộng đáng kể không gian hiện diện của tiếng Việt trong môi trường trực tuyến. Các hệ thống tìm kiếm, trợ lý hội thoại, dịch máy và công cụ tạo sinh văn bản giúp tiếng Việt gia tăng mật độ xuất hiện trong kho dữ liệu số toàn cầu, từ đó góp phần củng cố vị thế của ngôn ngữ trong giao tiếp xuyên biên giới. AI đồng thời nâng cao năng lực tiếp cận tri thức cho người

dùng tiếng Việt thông qua tự động dịch, tóm tắt và diễn giải tài liệu chuyên môn, giúp rút ngắn khoảng cách giữa nguồn tri thức quốc tế và cộng đồng bản ngữ. Trong lĩnh vực giáo dục và sáng tạo nội dung, công cụ AI hỗ trợ thiết kế học liệu, gợi ý diễn đạt và đa dạng hoá phong cách văn bản, tạo thêm điều kiện cho người học và người viết phát triển năng lực ngôn ngữ. Các nghiên cứu về ứng dụng mô hình ngôn ngữ lớn trong giáo dục cho thấy AI có thể đóng vai trò như công cụ hỗ trợ nhận thức và diễn đạt nếu được sử dụng có định hướng (Kasneci & cs., 2023).

2.2.2. Nguy cơ chuẩn hoá lệch từ dữ liệu huấn luyện

AI đặt ra nguy cơ chuẩn hoá lệch đối với tiếng Việt khi dữ liệu huấn luyện không phản ánh đầy đủ phổ đa dạng ngôn ngữ. Các mô hình ngôn ngữ học từ tập văn bản lớn nhưng không trung tính: nguồn dữ liệu có thể thiên lệch về thể loại, vùng miền, tầng lớp xã hội hoặc phong cách diễn đạt. Khi các mẫu phổ biến được ưu tiên về mặt xác suất, đầu ra của hệ thống có xu hướng tái sản xuất một dạng "chuẩn trung bình", làm mờ các biến thể địa phương và sắc thái phong cách. Nghiên cứu phê bình về mô hình ngôn ngữ đã nhấn mạnh rằng thiên lệch dữ liệu không chỉ tạo sai lệch nội dung mà còn ảnh hưởng đến chuẩn biểu đạt được tái tạo tự động (Bender & cs.,). Đối với tiếng Việt, nếu kho ngữ liệu số tập trung chủ yếu vào văn bản hành chính - báo chí - truyền thông đại chúng, thì phương ngữ, khẩu ngữ và diễn ngôn cộng đồng có nguy cơ bị giảm mức đại diện, dẫn đến khuynh hướng trung tính hoá văn phong và thu hẹp phổ bản sắc.

2.2.3. Ảnh hưởng của dịch máy và cấu trúc ngoại lai

Dịch máy và các hệ thống sinh văn bản đa ngữ tạo thuận lợi cho giao tiếp liên ngôn ngữ, nhưng đồng thời cũng làm gia tăng ảnh hưởng của cấu trúc ngoại lai lên tiếng Việt. Trong nhiều trường hợp, văn bản dịch tự động giữ lại trật tự cú pháp và lối kết hợp từ của ngôn ngữ nguồn, tạo nên hiện tượng "dịch bám cấu trúc" thay vì chuyển đổi theo quy tắc tự nhiên của tiếng Việt. Điều này có thể dẫn đến lai hoá cấu trúc câu, mở rộng các mẫu diễn đạt không điển hình và làm suy giảm sắc thái ngữ dụng. Các nghiên cứu về dịch máy thần kinh cho thấy dù độ trôi chảy tăng, hệ thống vẫn thường mắc lỗi ở tầng diễn ngôn và ngữ dụng, đặc biệt với ngôn ngữ ít tài nguyên (Koehn & Knowles, 2017). Với tiếng Việt, nếu văn bản dịch máy được sử dụng rộng rãi mà thiếu khâu hiệu đính ngôn ngữ học, các mẫu cấu trúc ngoại lai có thể dần được "bình thường hoá" trong thực hành viết.

2.2.4. Biến đổi thói quen giao tiếp số

Môi trường giao tiếp số, kết hợp với công cụ AI, đang thúc đẩy biến đổi trong thói quen sử dụng ngôn ngữ theo hướng rút gọn và ký hiệu hoá. Lối viết tắt, thay thế bằng ký hiệu và emoji, cũng như cấu trúc câu tối giản xuất hiện ngày càng phổ biến trong tương tác trực tuyến. Khi công cụ gợi ý và sinh văn bản tự động cung cấp phương án diễn đạt nhanh, người dùng có xu hướng ưu tiên tốc độ hơn chiều sâu lập luận. Một số nghiên cứu về truyền thông số cho thấy áp lực ngắn gọn và tức thời có thể làm giảm độ phức hợp cú pháp và mật độ lập luận trong văn bản thường nhật (Tagg, 2015). Trong bối cảnh học thuật, nếu phụ thuộc quá mức vào văn bản sinh tự động và lối diễn đạt mẫu, người học có nguy cơ suy giảm năng lực tổ chức lập luận và diễn đạt chuyên sâu bằng tiếng Việt. Do đó, tác động của AI đối với bản sắc ngôn ngữ cần được nhìn nhận như một quá trình hai mặt: mở rộng năng lực sử dụng, nhưng đồng thời đòi hỏi cơ chế định hướng và hiệu chỉnh từ góc độ ngôn ngữ học.

2.3. Phân tích các trục bảo tồn và phát huy giá trị tiếng Việt trong kỷ nguyên AI

Trên cơ sở các phân tích về tác động hai chiều của AI đối với ngôn ngữ, việc bảo tồn và phát huy giá trị tiếng Việt trong kỷ nguyên số cần được triển khai theo các trục đồng bộ: dữ liệu - đánh giá - năng lực người dùng. Cách tiếp cận này xem ngôn ngữ không chỉ là đối tượng được xử lý bởi công nghệ, mà là hệ giá trị cần được quản trị bằng tri thức ngôn ngữ học, chuẩn mực học thuật và thực hành giáo dục.

2.3.1. Hệ sinh thái dữ liệu tiếng Việt có kiểm soát chất lượng

Một hệ sinh thái dữ liệu tiếng Việt bền vững cho AI cần dựa trên nguyên tắc chuẩn hoá nguồn dữ liệu và kiểm soát chất lượng ngữ liệu. Chuẩn hoá ở đây không đồng nghĩa với đồng nhất hoá, mà là thiết lập tiêu chí tuyển chọn, phân loại và mô tả dữ liệu rõ ràng về thể loại, phong cách và ngữ cảnh sử dụng. Ngữ liệu cần được gắn nhãn ngôn ngữ học ở các tầng cơ bản (từ loại, cú pháp, diễn ngôn, thể loại) để tăng khả năng truy vết và đánh giá. Các nghiên cứu về quản trị dữ liệu ngôn ngữ nhấn mạnh rằng tài liệu hoá dữ liệu (data documentation) là điều kiện then chốt để giảm thiên lệch và tăng độ tin cậy của mô hình (Gebru & cs., 2021).

Bên cạnh đó, hệ sinh thái dữ liệu cần bảo đảm mức đại diện của đa dạng vùng miền và phong cách, bao gồm văn bản chuẩn, văn bản học thuật, khẩu ngữ, phương ngữ và diễn ngôn cộng đồng. Vai trò của cơ sở đào tạo và viện nghiên cứu ngôn ngữ đặc biệt quan trọng trong việc xây dựng kho ngữ liệu chuẩn, thẩm định chất lượng và cung cấp siêu dữ liệu ngôn ngữ học, thay vì để dữ liệu tiếng Việt phụ thuộc hoàn toàn vào nguồn thu thập tự phát trên Internet.

2.3.2. Chuẩn đánh giá đầu ra AI theo tiêu chí ngôn ngữ học

Đánh giá đầu ra ngôn ngữ của hệ thống AI không thể chỉ dựa trên độ trôi chảy hình thức, mà cần dựa trên bộ tiêu chí của ngôn ngữ học ứng dụng. Các tiêu chí cốt lõi gồm: độ đúng ngôn ngữ và nội dung, độ tự nhiên trong diễn đạt, và mức độ phù hợp ngữ dụng theo ngữ cảnh giao tiếp. Từ góc nhìn ngữ dụng học, một phát ngôn được xem là đạt yêu cầu khi tương thích với vai giao tiếp, thể loại văn bản và mục đích phát ngôn (Levinson, 1983).

Do đó, cần xây dựng bộ tiêu chí đánh giá đầu ra AI theo ngữ cảnh sử dụng tiếng Việt: văn bản giáo dục, hành chính, truyền thông, học thuật hay sáng tạo. Quá trình đánh giá nên có sự tham gia của chuyên gia ngôn ngữ nhằm kiểm định sắc thái nghĩa, tính phù hợp văn phong và giá trị biểu đạt. Cách tiếp cận "đánh giá con người trong vòng lặp" (human-in-the-loop evaluation) được nhiều nghiên cứu AI khuyến nghị như một cơ chế đảm bảo chất lượng và trách nhiệm ngôn ngữ (Floridi & Cowls, 2019).

2.3.3. Giáo dục năng lực số - ngôn ngữ cho người dùng

Song song với dữ liệu và tiêu chí đánh giá, yếu tố quyết định vẫn là năng lực của người sử dụng tiếng Việt trong môi trường AI. Giáo dục năng lực số - ngôn ngữ cần hướng đến khả năng đọc - viết cùng AI, tức biết khai thác gợi ý, so sánh phương án diễn đạt và lựa chọn hình thức phù hợp. Người học cần được trang bị kỹ năng hiệu chỉnh đầu ra AI: phát hiện lỗi ngữ pháp, lệch nghĩa, sai ngữ dụng và dấu vết cấu trúc ngoại lai. Các nghiên cứu về năng lực số nhấn mạnh vai trò của tư duy phản biện và năng lực siêu ngôn ngữ trong tương tác với hệ thống tự động (UNESCO, 2018).

Bên cạnh đó, đạo đức sử dụng AI trong ngôn ngữ cũng cần được đưa vào đào tạo: minh bạch nguồn hỗ trợ, không đánh đồng văn bản sinh tự động với năng lực cá nhân, và tôn trọng chuẩn mực học thuật. Đây là điều kiện để AI trở thành công cụ hỗ trợ phát triển tiếng Việt, thay vì làm suy giảm trách nhiệm diễn đạt của người dùng.

2.4. Vai trò của cơ sở đào tạo và nhà nghiên cứu ngôn ngữ

2.4.1. Vai trò của trường đại học sư phạm - ngôn ngữ

Các trường đại học sư phạm và cơ sở đào tạo ngôn ngữ có vai trò hạt nhân trong việc bảo đảm chất lượng ngôn ngữ tiếng Việt trong kỷ nguyên AI. Trước hết, cần chủ động xây dựng và chia sẻ kho ngữ liệu chuẩn phục vụ nghiên cứu và huấn luyện mô hình, có chú giải ngôn ngữ học và phân loại thể loại rõ ràng. Đồng thời, các đơn vị đào tạo có thể tham gia phát triển bộ tiêu chí đánh giá đầu ra AI theo chuẩn tiếng Việt, dựa trên ngữ pháp, phong cách và ngữ dụng. Việc đào tạo năng lực ngôn ngữ số cho sinh viên sư phạm và sinh viên ngành ngôn ngữ cũng cần được tích hợp vào chương trình, nhằm hình thành đội ngũ có khả năng giảng dạy và hiệu chỉnh ngôn ngữ trong môi trường công nghệ.

2.4.2. Liên kết ngôn ngữ học - công nghệ - giáo dục

Bảo tồn và phát huy giá trị tiếng Việt trong hệ sinh thái AI đòi hỏi liên kết liên ngành giữa ngôn ngữ học, khoa học dữ liệu và khoa học giáo dục. Các dự án ngữ liệu tiếng Việt cần có sự phối hợp giữa nhà ngôn ngữ học và kỹ sư dữ liệu để bảo đảm vừa đúng chuẩn học thuật vừa phù hợp xử lý máy. Mô hình phòng thí nghiệm ngôn ngữ số (digital language lab) trong trường đại học có thể là không gian thử nghiệm, đánh giá và cải tiến công cụ AI tiếng Việt. Kinh nghiệm quốc tế cho thấy hạ tầng nghiên cứu ngữ liệu mở và liên ngành là yếu tố thúc đẩy phát triển công nghệ ngôn ngữ bền vững (Bird, Klein & Loper, 2009).

2.4.3. Đề xuất mô hình tham gia của giảng viên ngôn ngữ Việt Nam

Giảng viên ngôn ngữ Việt Nam có thể tham gia trực tiếp vào hệ sinh thái AI qua ba vai trò: biên soạn và thẩm định dữ liệu chuẩn; phản biện và hiệu chỉnh ngôn ngữ đầu ra AI; và hướng dẫn người học thực hành đồng kiến tạo nghĩa với hệ thống. Đồng kiến tạo ở đây được hiểu là quá trình người dùng đặt yêu cầu, đánh giá, hiệu chỉnh và tái diễn đạt văn bản sinh tự động, qua đó giữ vai trò chủ thể của hoạt động ngôn ngữ. Mô hình này giúp chuyển AI từ vị trí "thay thế" sang "đối tác ngôn ngữ" có kiểm soát.

2.5. Định hướng chính sách và khuyến nghị thực tiễn

Ở cấp học thuật - nghiên cứu, cần ưu tiên chương trình xây dựng và chú giải kho ngữ liệu tiếng Việt đa thể loại, đa vùng miền, có kiểm định chuyên môn. Ở cấp giáo dục - đào tạo, cần tích hợp năng lực ngôn ngữ số và kỹ năng làm việc với AI vào chương trình dạy học tiếng Việt và đào tạo giáo viên. Ở cấp nền tảng công nghệ, cần khuyến khích doanh nghiệp phát triển mô hình tiếng Việt minh bạch dữ liệu và có cơ chế đánh giá ngôn ngữ học. Ở cấp cộng đồng sử dụng, cần nâng cao nhận thức về giá trị bản sắc ngôn ngữ, khuyến khích sử dụng tiếng Việt chuẩn mực nhưng linh hoạt, có ý thức phản biện đầu ra tự động.

III. KẾT LUẬN

Bài viết đã chỉ ra rằng AI tạo ra cả cơ hội mở rộng lẫn thách thức chuẩn hoá đối với tiếng Việt và bản sắc ngôn ngữ. Tác động của AI không chỉ nằm ở công nghệ, mà ở cách dữ liệu được lựa chọn, cách đầu ra được đánh giá và cách người dùng tương tác. Tiếng Việt trong kỷ nguyên AI cần được tiếp cận như một hệ thống chuẩn kết hợp đa dạng, được bảo vệ bằng tri thức ngôn ngữ học và phát triển thông qua giáo dục năng lực số. Khẳng định vị thế tiếng Việt trong môi trường AI không phải bằng cách khước từ công nghệ, mà bằng cách tham gia chủ động, có kiểm soát và có định hướng học thuật. Hướng nghiên cứu tiếp theo có thể tập trung vào xây dựng bộ tiêu chí đánh giá ngữ dụng cho đầu ra AI tiếng Việt và mô hình kho ngữ liệu chú giải phục vụ huấn luyện mô hình ngôn ngữ.

TÀI LIỆU THAM KHẢO

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT ’21) (pp. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. *O’Reilly Media*. https://www.nltk.org/book/

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies, 7*(4-5), 585-614. https://doi.org/10.1017/S0047404505050201

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Minds and Machines,* 29, 689-707. https://doi.org/10.1007/s11023-018-9482-5

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM,* 64(12), 86-92. https://doi.org/10.1145/3458723

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Instruction,* 103, Article 101780. https://doi.org/10.1016/j.learninstruc.2023.101780

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *Computational Linguistics*, 43(2), 369-387. https://doi.org/10.1162/COLI_a_00294

Levinson, S. C. (1983). Pragmatics. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511813313

Tagg, C. (2015). Exploring digital communication: Language in action. *Cambridge University Press.* https://doi.org/10.1017/CBO9781139341127

UNESCO. (2018). *A global framework of reference on digital literacy skills.* https://unesdoc.unesco.org/ark:/48223/pf0000265403

Wittgenstein, L. (2009). *Philosophical investigations (Rev. 4th ed.).* Wiley-Blackwell. https://doi.org/10.1002/9781444315011.ch66