



ỨNG DỤNG KỸ THUẬT ĐO ĐỘ KHÓ TRONG XÂY DỰNG KHỐI NGỮ LIỆU ĐỌC TRÌNH ĐỘ A2, B1, B2 THEO “KHUNG NĂNG LỰC TIẾNG VIỆT DÙNG CHO NGƯỜI NƯỚC NGOÀI” TẠI TRƯỜNG ĐẠI HỌC NGOẠI NGỮ, ĐẠI HỌC HUẾ

Hồ Viết Hoàng*, Đinh Điền, Lương An Vinh, Trần Mai Phương,

Dương Thị Nhung, Trần Thị Xuân, Đặng Diễm Đông, Nguyễn Sơn

* Trường Đại học Ngoại Ngữ, Đại học Huế, 57 Nguyễn Khoa Chiêm, tp. Huế, Việt Nam

* Tác giả liên hệ: **Hồ Viết Hoàng** < hvhoang@hueuni.edu.vn >

(Ngày nhận bài: 06-03-2023; Ngày chấp nhận đăng: 12-04-2023)

Tóm tắt. Trên cơ sở tiếp cận, xử lý các tiêu chuẩn, tiêu chí của “Khung năng lực tiếng Việt dùng cho người nước ngoài” và “Định dạng đề thi đánh giá năng lực tiếng Việt theo Khung năng lực tiếng Việt dùng cho người nước ngoài”, chúng tôi đã vận dụng phần mềm, kỹ thuật đo độ khó văn bản của khoa học máy tính để: (1) phân tích, đánh giá các bài đọc hiện có trong các giáo trình, bài giảng, tài liệu phục vụ giảng dạy tiếng Việt cho người nước ngoài; (2) xây dựng 300 bài đọc và 1.000 câu hỏi tương ứng trình độ A2, B1, B2 để phục vụ biên soạn bài giảng, đề thi đánh giá năng lực tiếng Việt cho người nước ngoài tại Trường Đại học Ngoại ngữ, Đại học Huế.

Từ khóa: tiếng Việt; ngữ liệu đọc; độ khó; trình độ A2, B1, B2

APPLICATION OF TEXT READABILITY TECHNIQUES IN TEACHING VIETNAMESE WRITING SKILL BASE ON “VIETNAMESE COMPETENCY FRAMEWORK FOR FOREIGNERS” LEVEL A1, A2, B1

Ho Viet Hoang*, Dinh Dien, Luong An Vinh, Tran Mai Phuong,

Duong Thi Nhung, Tran Thi Xuan, Dang Diem Dong, Nguyen Son

University of Foreign Languages, Hue University, 57 Nguyen Khoa Chiem St., Hue, Vietnam,

*Correspondence to **Ho Viet Hoang** < hvhoang@hueuni.edu.vn >

(Received: March 06, 2023; Accepted: April 12, 2023)

Abstract. Based on “Vietnamese competency framework for foreigners” and “Format of test questions to assess Vietnamese language ability according to Vietnamese competency framework for foreigners”, we use text readability techniques to: (1) analyze and evaluate existing readings in textbooks, lectures and documents for teaching Vietnamese to foreigners; (2) collect 300 reading passages and 1.000 questions respectively to serve the preparation of lectures and tests to assess Vietnamese language ability for foreigners at University Foreign Languages and International Studies, Hue University.

Keywords: Vietnamese; corpus; readability; A2, B1, B2 level.

Mở đầu

Trong xu thế hội nhập và phát triển, Việt Nam đang từng bước khẳng định vị thế, vai trò, tầm ảnh hưởng mang tầm quốc tế (đặc biệt là tại Liên Hiệp quốc, APEC, ASEAN...) và được quốc tế công nhận. Cùng với đó, nhiều tổ chức, quốc gia trên thế giới đang thay đổi chiến lược hướng về Đông Nam Á và Việt Nam – điểm đến của cơ hội hợp tác, phát triển trên nền tảng an ninh, an toàn và bình đẳng, cùng có lợi.

Quá trình hội nhập, phát triển đã từng bước lan tỏa các giá trị của Việt Nam ra thế giới, trong đó có ngôn ngữ và văn hóa. Tiếng Việt đang trở thành ngôn ngữ – ngoại ngữ có vai trò quan trọng trong đời sống của nhiều quốc gia¹. Nhằm đáp ứng nhu cầu tìm hiểu, nghiên cứu, học tập về đất nước, con người Việt Nam, nhiều trung tâm, viện nghiên cứu, các cơ sở đào tạo Việt Nam học và tiếng Việt đã ra đời trên thế giới.

Từ thực tế về đào tạo tiếng Việt cho người Việt Nam ở nước ngoài (thế hệ con cháu người Việt) và người nước ngoài tại Việt Nam (đặc biệt đối tượng là người định cư, công tác, kết hôn với người Việt, học tập)... Bộ Giáo dục và Đào tạo đã ban hành nhiều Thông tư, Quyết định, Công văn nhằm thống nhất chương trình, mục tiêu, nội dung đào tạo và thi đánh giá theo

¹ Tiếng Việt là ngôn ngữ phổ biến thứ 14 của thế giới, với khoảng 67.662.000 người [Sil, 2017], tập trung tại các quốc gia có đông người Việt Nam sinh sống, định cư: theo số liệu thống kê chưa đầy đủ, hiện nay có khoảng 5.3 triệu Việt kiều sinh sống tại 130 quốc gia và vùng lãnh thổ [Kết luận số 12-KL/TW ngày 12/8/2021 của Bộ Chính trị về công tác người Việt Nam ở nước ngoài trong tình hình mới], chủ yếu tại Hoa Kỳ khoảng 2 triệu người; Pháp khoảng 350.000 người; Đài Loan khoảng 320.000 người; Úc khoảng 300.000 người; Nhật Bản khoảng 260.000 người; Canada khoảng 200.000 người; Hàn Quốc khoảng 150.000 người; Đức khoảng 130.000 người; Thái Lan, Lào, Campuchia...

Trên cơ sở này, Chính phủ đã chọn ngày 08/9 hàng năm làm ngày “Tôn vinh tiếng Việt trong cộng đồng người Việt Nam ở nước ngoài giai đoạn 2023-2030” Quyết định số 930/QĐ-TTg ngày 03/8/2022 của Thủ tướng Chính phủ về việc phê duyệt Đề án “Ngày Tôn vinh tiếng Việt trong cộng đồng người Việt Nam ở nước ngoài giai đoạn 2023-2030”.

khung năng lực tiếng Việt². Trên cơ sở chương trình khung của Bộ Giáo dục và Đào tạo, các cơ sở đào tạo tiếng Việt cho người nước ngoài ở Việt Nam và người Việt Nam ở nước ngoài đã tổ chức biên soạn giáo trình, bài giảng, xây dựng quỹ đề thi đánh giá năng lực tiếng Việt, tiếp cận chủ yếu theo hướng đơn ngành, đa ngành. Nhìn chung, các giáo trình, bài giảng và quỹ đề thi đánh giá năng lực tiếng Việt hiện có chủ yếu thiên về cảm tính, chủ quan, thiếu các đánh giá, đo lường khách quan, định lượng, nhất là ứng dụng kỹ thuật, phần mềm của khoa học máy tính.

Thông qua thực tiễn 18 năm đào tạo tiếng Việt và văn hóa Việt Nam cho người nước ngoài tại trường Đại học Ngoại ngữ, Đại học Huế và nhiều cơ sở giáo dục trên cả nước (kỳ thi đánh giá năng lực tiếng Việt 3 bậc dành cho người nước ngoài: A2, B1, B2); căn cứ các tài liệu hiện có phục vụ đào tạo tiếng Việt cho người nước ngoài (đại diện cho miền Bắc, miền Trung và miền Nam); trên cơ sở nhu cầu về việc thực hiện nhiệm vụ Bồi dưỡng và cấp chứng chỉ tiếng Việt cho người nước ngoài theo Công văn số 2563/BGDĐT-CQLCL ngày 13 tháng 6 năm 2019 của Bộ Giáo dục và Đào tạo và Quyết định số 778/QLCL-QLT ngày 02/6/2022 của Cục Quản lý Chất lượng, Bộ Giáo dục và Đào tạo về việc công nhận Trường Đại học Ngoại ngữ, Đại học Huế được phép tổ chức thi đánh giá năng lực tiếng Việt theo Khung năng lực tiếng Việt dùng cho người nước ngoài; bài nghiên cứu sẽ tiếp cận theo hướng liên ngành (và xuyên ngành giữa Ngôn ngữ học và Khoa học máy tính) nhằm *ứng dụng kỹ thuật đo độ khó trong việc xây dựng khối ngữ liệu đọc trình độ A2, B1, B2 theo “Khung năng lực tiếng Việt dành cho người nước ngoài” tại Trường Đại học Ngoại ngữ, Đại học Huế*, hướng đến sự khách quan, khoa học, định lượng trong giảng dạy, biên soạn giáo trình/bài giảng, kiểm tra và đánh giá năng lực tiếng Việt kỹ năng đọc trình độ A2, B1, B2.

1. Cơ sở thực hiện

1.1. Nội dung

- Hệ thống hóa cơ sở lý luận và thực tiễn bộ tiêu chuẩn, các tiêu chí theo “Khung năng lực tiếng Việt dùng cho người nước ngoài” và “Định dạng đề thi đánh giá năng lực tiếng Việt theo Khung năng lực tiếng Việt dùng cho người nước ngoài”, “Chương trình tiếng Việt cho người Việt Nam ở nước ngoài”... của Bộ Giáo dục và Đào tạo.

- Thu thập, thống kê các giáo trình, bài giảng, tài liệu về giảng dạy tiếng Việt cho người Việt Nam ở nước ngoài và tiếng Việt cho người nước ngoài; sử dụng phần mềm để phân tích, đo độ khó các bài đọc trình độ A2, B1, B2 từ các tài liệu để đưa ra kết quả, nhận định, đánh giá.

² Để cập cụ thể ở phần 1.4. Lịch sử nghiên cứu.

- Ứng dụng khoa học Máy tính:

+ Phần mềm ngữ liệu phục vụ việc xây dựng bộ tiêu chí đánh giá năng lực tiếng Việt 6 bậc, phục vụ biên soạn giáo trình tiếng Việt cho người nước ngoài ở trình độ Sơ cấp (tương đương bậc 1 - A1, bậc 2 - A2) và Trung cấp (bậc 3 - B1) do Trung tâm ngôn ngữ học Tính toán, Trường ĐH Khoa học Tự nhiên, ĐHQG Tp. Hồ Chí Minh biên soạn và đã chuyển giao quyền sử dụng cho Trường ĐH Ngoại ngữ, Đại học Huế với: (1) Danh sách 7.000 hình tiết phổ biến nhất trong tiếng Việt; (2) Từ điển tần số tiếng Việt đầy đủ (40.000 mục từ); (3) Từ điển âm thanh của tất cả các âm tiết tiếng Việt phát âm chuẩn (giọng người).

+ Ứng dụng kho ngữ liệu đơn ngữ tiếng Việt (Vietnamese Corpus) gồm 17.095.994 câu, 346.454.533 từ và 443.301.776 chữ (tiếng/âm tiết) thuộc 42 lĩnh vực và gom thành 18 chủ đề (khoa học, kinh tế, văn hóa...) được thu thập tự động tại các trang báo điện tử chính thống từ năm 2000 đến nay. Ngữ liệu này được gán nhãn tự động ranh giới câu và ranh giới từ.

+ Ứng dụng kỹ thuật đo độ khó nhằm xác định độ khó từ, câu, đoạn để xây dựng và hoàn chỉnh khối ngữ liệu đọc trình độ A2, B1, B2 theo các tiêu chuẩn, tiêu chí của “Khung năng lực tiếng Việt dùng cho người nước ngoài”, đảm bảo tính khách quan, khoa học, phục vụ giảng dạy và biên soạn đề thi đánh giá năng lực tiếng Việt.

- Trên cơ sở này, vận dụng kết quả nghiên cứu để định dạng và xây dựng khối ngữ liệu đọc trình độ A2, B1, B2 theo “Khung năng lực tiếng Việt dành cho người nước ngoài” phục vụ giảng dạy tiếng Việt, phù hợp với bộ tiêu chuẩn, các tiêu chí, chủ đề của Khung năng lực.

1.2. Đối tượng nghiên cứu

Tập trung vào các nguồn ngữ liệu đọc chính thống (mang tính pháp quy, phổ quát: Văn bản Nhà nước, báo chí, giáo trình, tác phẩm... đã được thẩm định, công bố) và phần mềm, kỹ thuật đo độ khó văn bản, tần suất xuất hiện từ để xây dựng khối ngữ liệu đọc trình độ A2, B1, B2 theo “Khung năng lực tiếng Việt dùng cho người nước ngoài”, nhằm phục vụ giảng dạy và biên soạn đề thi đánh giá năng lực tiếng Việt.

1.3. Phạm vi nghiên cứu

Bên cạnh các bài đọc được xây dựng theo tiêu chuẩn, tiêu chí, chủ đề theo Khung đánh giá năng lực tiếng Việt dùng cho người nước ngoài, bài nghiên cứu tập trung vào nội dung các chủ đề thuộc kỹ năng đọc có trong 5 giáo trình/bộ giáo trình đào tạo tiếng Việt cho người nước

ngoài: 1. Nguyễn Văn Huệ (cb) (2001, 2004), *Giáo trình tiếng Việt dành cho người nước ngoài (quyển 1, 2, 3, 4)*; 2. Nguyễn Việt Hương (2009), *Tiếng Việt cơ sở dành cho người nước ngoài (quyển 1, 2) – Elementary Vietnamese, use for non-Vietnamese speakers* và (2016, 2017), *Tiếng Việt nâng cao dành cho người nước ngoài (quyển 1, 2) – Intermediate Vietnamese, use for foreigners*; 3. Nguyễn Thiện Nam (1998), *Tiếng Việt nâng cao (cho người nước ngoài)*; 4. Đoàn Thiện Thuật (cb) (2005, 2014), *Thực hành tiếng Việt trình độ A, B, C (sách dùng cho người nước ngoài)*; 5. Viện Việt Nam học và Khoa học phát triển, ĐHQG Hà Nội (2014), *Tiếng Việt trình độ A, B, C...*. Đặc biệt, bài viết đã sử dụng kết quả nghiên cứu từ bộ giáo trình: Khoa Việt Nam học và Tiếng Việt, Trường ĐH KHXH và NV, ĐHQG Hà Nội (2019), “*Tiếng Việt cho người nước ngoài*” tập 1, 2, 3, 4 (tương ứng với trình độ A1, A2, B1, B2 của Khung năng lực). Từ kết quả nghiên cứu, bài nghiên cứu sẽ hướng đến xây dựng khối ngữ liệu đọc (đã được xử lý thông qua ứng dụng đo độ khó) theo Khung năng lực với 300 bài đọc, 1.000 câu hỏi tương ứng và phân theo trình độ A2, B1, B2.

1.4. *Lịch sử nghiên cứu*

Phân tích, áp dụng các Quyết định, Thông tư, Công văn... của Bộ Giáo dục và Đào tạo liên quan đến đào tạo tiếng Việt cho người nước ngoài và tiếng Việt cho người Việt Nam ở nước ngoài, làm nền tảng pháp lý, cơ sở triển khai xây dựng khối ngữ liệu đọc trình độ A2, B1, B2. Trong đó, tập trung vào: Thông tư số 17/2015/TT-BGDĐT ngày 01/9/2015 ban hành “*Khung năng lực tiếng Việt dùng cho người nước ngoài*”; Quyết định số 2097/QĐ-BGDĐT ngày 21/6/2016 ban hành “*Định dạng đề thi đánh giá năng lực tiếng Việt theo Khung năng lực tiếng Việt dùng cho người nước ngoài*”; Thông tư số 28/2018/TT-BGDĐT ngày 26/11/2018 ban hành “*Chương trình tiếng Việt cho người Việt Nam ở nước ngoài*”.

Nhóm tài liệu đơn ngữ và tài liệu song ngữ liên quan đến giảng dạy tiếng Việt cho người nước ngoài và người gốc Việt trên thế giới: Để giúp người nước ngoài, người nước ngoài gốc Việt dễ dàng tiếp cận và học ngôn ngữ tiếng Việt, nhiều cơ sở đào tạo đã biên soạn giáo trình, tài liệu song ngữ, gồm tiếng Việt và ngôn ngữ bản địa như Cho Jae Hyun, Jeon Hye Kyung, Song Jeong Nam, Nguyễn Văn Phúc (2005), *Tiếng Việt thực hành cho người Hàn Quốc (4 tập)*; Trần Thị Chung Toàn (2000), *Tiếng Việt cơ sở cho người Nhật*; Trần Văn Lâm, Tô Cẩm Duy (2005), *Tiếng Việt cho người Trung Quốc*; Bình Slaviká (2015), *Tiếng Việt thực hành (dành cho người Séc gốc Việt và người Séc học tiếng Việt)*... Đây là nguồn ngữ liệu quan trọng để đề tài hướng đến việc hệ thống hóa các chủ đề trong các giáo trình, bài giảng, làm cơ sở xây dựng bài đọc và quỹ đề thi đánh giá năng lực tiếng Việt khách quan, khoa học.

Nhóm tài liệu liên quan đến ngữ liệu tiếng Việt cho người nước ngoài: Nhóm tài liệu đã liệt kê trong phần *phạm vi nghiên cứu*. Đây là những nguồn tư liệu quan trọng (nhất là các chủ đề

giảng dạy trong từng cấp độ) để bài nghiên cứu tiếp cận, xử lý trong quá trình thu thập, ứng dụng kỹ thuật, phương pháp đo độ khó văn bản nhằm xây dựng khối ngữ liệu đọc trình độ A2, B1, B2 theo Khung năng lực tiếng Việt, phục vụ biên soạn bài đọc và đề thi đánh giá năng lực tiếng Việt khách quan, khoa học.

Nhóm tài liệu liên quan đến ứng dụng của hoa học máy tính trong giảng dạy tiếng Việt và kỹ năng đọc: Ngôn ngữ học Máy tính hay Ngôn ngữ học Tính toán là khoa học liên ngành giữa Ngôn ngữ học và Tin học nhằm sử dụng các công cụ tin học để nghiên cứu và xử lý ngôn ngữ của con người. Hiện nay có rất nhiều tài liệu liên quan đến nhóm vấn đề này. Trong đó, phải kể đến các nghiên cứu như Đinh Điền (2005), *Xây dựng và khai thác kho ngữ liệu song ngữ Anh – Việt điện tử*; Đinh Điền (2011), *Ngôn ngữ học máy tính và việc biên soạn từ điển*; Đinh Điền (2018), *Nghiên cứu liên ngành về Ngôn ngữ học tính toán và giảng dạy ngôn ngữ*; Đinh Điền (2018), *Ngôn ngữ học ngữ liệu*; Hoàng Phê (1980), *Từ điển tiếng Việt*; Hoàng Khuê, Nguyễn Thị Như Diệp, Đinh Điền, Nguyễn Thanh Thủy (2018), *Ứng dụng kho ngữ liệu song song đa ngữ trong giảng dạy tiếng Việt cho người nước ngoài*; Tony Mc, A. W. (2001), *Corpus Linguistics*; Sil. (2017), *Language Repot*, Retrieved from <https://www.ethnologue.com...> Các nghiên cứu đã chỉ ra rằng, ngôn ngữ học ngữ liệu nói chung và ngôn ngữ học ngữ liệu tiếng Việt nói riêng có vai trò quan trọng trong nghiên cứu, giảng dạy vì nó giúp xác định và xây dựng giáo trình, bài giảng với độ khó (phụ thuộc vào vốn từ vựng, cấu trúc ngữ pháp, kết cấu văn bản) phù hợp cho từng trình độ. Để thống kê, định lượng được độ khó, chúng ta cần phải có các kho ngữ liệu. Ngoài ra, khi giảng dạy ngôn ngữ – tiếng Việt, chúng ta không thể dạy từ vựng rời rạc mà phải đưa vào ngữ cảnh – được rút ra từ các kho ngữ liệu.

Từ kết quả nghiên cứu trước đó, nhóm tác giả xác định độ khó được biểu hiện trong các từ, câu, văn bản của giáo trình, tài liệu đã xuất bản, bao gồm những bài đọc, bài tập, từ vựng, giải thích từ vựng... Độ khó của văn bản là tập hợp tất cả các yếu tố nội tại của văn bản khiến cho văn bản dễ hay khó tiếp thu ở đa số người đọc ở một cấp độ cụ thể. Độ khó của văn bản phụ thuộc vào độ phổ biến của từ vựng (tần suất), độ phức tạp của cấu trúc câu (số lượng mệnh đề, chiều dài câu) và tổ chức văn bản (liên kết văn bản). Độ khó được đo trên chính nội tại văn bản (đây là yếu tố khách quan, không phụ thuộc vào người đọc). Còn độ đọc hiểu phụ thuộc vào trình độ, kinh nghiệm, sự say mê, tinh thần, sức khỏe... của người đọc (đây là yếu tố chủ

quan, phụ thuộc vào người đọc). Độ khó và độ đọc hiểu phụ thuộc vào nội dung của văn bản, còn độ rõ phụ thuộc vào hình thức của văn bản (font chữ, cỡ chữ, màu sắc...)³.

Từ việc định dạng về độ khó nêu trên, bài nghiên cứu đã tiếp cận, áp dụng những kết quả nghiên cứu, nhất là nghiên cứu cụ thể về kỹ năng đọc⁴ để ứng dụng kỹ thuật đo độ khó trong xây dựng khối ngữ liệu đọc theo Khung năng lực tiếng Việt nhằm phục vụ giảng dạy, biên soạn giáo trình và quỹ đề thi đánh giá năng lực tiếng Việt ở từng chủ đề cụ thể, các cấp độ khác nhau tùy vào tuần suất của từ, câu, đoạn, văn bản (dễ, trung bình, khó). Người dạy có thể rút ra từ khối ngữ liệu này những bài đọc, câu hỏi phù hợp với trình độ người học. Cách làm này mang tính định lượng, tránh trường hợp lựa chọn cảm tính, chủ quan của người biên soạn.

2. Phương pháp nghiên cứu

2.1. Cách tiếp cận

- Tiếp cận theo hướng liên ngành, nhất là tính xuyên ngành giữa Ngôn ngữ học và Khoa học máy tính.

- Tiếp cận dựa trên ngữ liệu đơn ngữ, đảm bảo một số tiêu chí về tính đại diện (ngữ liệu có nguồn uy tín, độ phổ biến cao, đa chiều: Giáo trình, bài giảng đã xuất bản phục vụ giảng dạy tiếng Việt ở Việt Nam và nước ngoài), tính cân bằng (vùng miền, giới tính, thể loại...) và niên đại (khoảng 20 năm trở lại đây).

2.2. Phương pháp nghiên cứu

Phương pháp thu thập và xử lý dữ liệu: Bao gồm thu thập các dữ liệu sơ cấp và thứ cấp để từng bước xây dựng và hoàn chỉnh khối ngữ liệu đọc trình độ A2, B1, B2 theo “Khung năng lực tiếng Việt dùng cho người nước ngoài”⁵. Đồng thời, vận dụng phương pháp thống kê, phương

³ Xem thêm: Nguyễn Sơn, Hồ Viết Hoàng, Nguyễn Thị Như Diệp, Đinh Điền, Lương An Vinh (2021), “So sánh các văn bản tiếng Việt theo độ khó”, *Tạp chí Khoa học Đại học Huế: Khoa học Xã hội và Nhân văn*, Tập 130, số 6A.

⁴ Như: DuBay, W. H (2007), *Unlocking language: The classic readability studies*; Bailin, A. & Grafstein, A. (2016), *Readability: Text and context*; An-Vinh Luong, Diep Nguyen, Dien Dinh (2017), *Examining the Text-length Factor in Evaluating the Readability of Literary Texts in Vietnamese Textbooks*; Nguyen Thi Nhu Diep, Luong An Vinh, Dinh Dien (2017), *Investigating some elements affecting the readability of Vietnamese texts - Primary level (in comparison to English ones) and Examining the Readability of proeses in the literature textbooks for Vietnamese students at primary and secondary schools*.

⁵ Chúng tôi đã thu thập nguồn ngữ liệu đọc trình độ A2, B1, B2 từ 15 bộ giáo trình phổ biến về dạy tiếng Việt cho người nước ngoài hiện nay, với 300 bài đọc tương ứng (trong đó, tập trung vào 05 giáo trình/bộ giáo trình đã đề cập ở phần Phạm vi nghiên cứu). Sau khi nạp ngữ liệu, hiệu chỉnh và làm sạch ngữ liệu, huấn luyện các công cụ xử lý ngôn ngữ tự động (như WordSmith Tools, Parallel Corpus, bộ phân lớp SVM)... , kết quả nghiên cứu đã phân tích độ khó văn bản/từng bài đọc, đánh giá trình độ phù hợp với từng bài đọc trong các giáo trình hiện có và xây dựng lại các bài đọc

pháp tổng hợp, phương pháp phân tích để hệ thống hóa nguồn ngữ liệu theo từng chủ đề, nội dung ở từng cấp độ cụ thể; thống kê tuần suất từ xuất hiện để đánh giá độ khó và phân cấp từ dễ đến khó...

Phương pháp học máy (Machine Learning) và phương pháp "học sâu" (Deep Learning) nhằm tìm kiếm văn bản, phân loại văn bản, tóm tắt văn bản, dịch văn bản, phân tích độ đo, hỗ trợ dịch thuật... Phương pháp nghiên cứu dựa trên ngữ liệu (*corpus-based approach*) là cách thức thu thập, xây dựng, xử lý, khai thác các kho ngữ liệu một cách hiệu quả và đúng mục đích sử dụng. Nhóm đã xử lý ngữ liệu thu thập bằng cách làm sạch ngữ liệu, chuẩn hóa ngữ liệu, kiểm lỗi ngữ liệu, phân đoạn câu cho ngữ liệu, gán nhãn ngôn ngữ (nhãn hình thái từ, nhãn ngữ pháp từ, nhãn ngữ nghĩa từ...), từ đó huấn luyện các công cụ xử lý ngôn ngữ tự động (huấn luyện phân đoạn từ, gán nhãn tự động, chuyển đổi cây cú pháp...).

Nhóm tác giả đã sử dụng một số công cụ để xử lý ngữ liệu, như:

+ WordSmith Tools⁶ (Concord – nhằm tìm chuỗi đồng hiện để liệt kê tất cả các xuất hiện của một từ hay ngữ kèm với ngữ cảnh trái và phải của từ ngữ đó; WordList – chức năng bảng từ để tạo ra một danh sách các từ có mặt trong các tập tin ngữ liệu mà chúng ta cần khảo sát như thứ hạng, tần suất, tỉ lệ phần trăm của từ trong văn bản, tỉ lệ phần trăm văn bản chứa từ đó...; và Keyword – rút trích từ khóa phục vụ tìm kiếm văn bản về chủ đề hay lĩnh vực nào đó).

+ ParaCor (Parallel Corpus)⁷: dùng để liệt kê, đối chiếu về hình thái, ngữ pháp, ngữ nghĩa và ngữ dụng của từ trong ngôn ngữ nguồn với ngôn ngữ đích trong một kho ngữ liệu (đơn ngữ và song ngữ). Quy trình thực hiện, gồm: Nạp ngữ liệu, chỉnh sửa ngữ liệu, tìm kiếm (tìm theo từ, theo thẻ, và kết hợp tìm từ + thẻ), thống kê (tần suất xuất hiện của tất cả các từ trong ngữ liệu và hiển thị các thông tin về tần suất, tỉ lệ, độ F – tần suất được chuẩn hóa trên kích thước ngữ liệu).

này theo trình độ A2, B1, B2 (trong đó có 65 bài có trình độ A2, 115 bài trình độ B1, 125 bài trình độ B2) trên cơ sở sử dụng phần mềm thay đổi độ khó từ, câu, đoạn, bài đọc.

⁶ Phần mềm này do Mike Scott (Đại học Liverpool) lập trình, dựa trên phần mềm MicroConcord của Mike Scott và Tim Johns, Nxb Đại học Oxford. Xem thêm: Scott, M. (2001). *WordSmith Tools version 5.0*. Retrieved from <http://www.lexically.net/wordsmith/>

⁷ Công cụ này do Trung tâm Ngôn ngữ học Tính toán, Trường ĐH Khoa học Tự nhiên, ĐHQG Thành phố Hồ Chí Minh xây dựng (CLC, 2007).

+ Phương pháp so sánh độ khó của các văn bản tiếng Việt với nhau, bằng việc sử dụng bộ phân lớp SVM. Bộ ngữ liệu sử dụng là các bài đọc (từ giáo trình, bài giảng tiếng Việt đã xuất bản và ngữ liệu mới được xây dựng trên cơ sở các tiêu chuẩn, tiêu chí của “Khung năng lực tiếng Việt dùng cho người nước ngoài”, trình độ A2, B1, B2) đã đánh giá độ khó tương quan với nhau thông qua từ, câu, đoạn, cấu trúc, hình thái... làm tiền đề cho việc so sánh và lựa chọn các văn bản/bài đọc phù hợp với trình độ của người học.

3. Kết quả nghiên cứu và thảo luận

- Nhìn chung, trước năm 2015, các cơ sở đào tạo tiếng Việt tại Việt Nam và nước ngoài đều xây dựng và tổ chức kỳ thi đánh giá năng lực tiếng Việt theo chuẩn đầu ra của từng chương trình đào tạo (Cử nhân, liên kết 2 + 2, 3 + 1, 1 năm, chứng chỉ/chứng nhận...). Tuy vậy, do phụ thuộc vào nhiều loại giáo trình khác nhau – thiếu thống nhất; đội ngũ, chất lượng giảng dạy không đồng đều; chưa có quy chuẩn rõ ràng về chuẩn đầu ra, mức độ đạt được trong từng chương trình đào tạo – thiếu tiêu chuẩn, tiêu chí, khung năng lực... nên nhiều khi chất lượng và kết quả đào tạo, đánh giá năng lực tiếng Việt mang tính chủ quan, cảm tính, thiếu cơ sở khoa học.

- Từ sau năm 2015, Bộ Giáo dục và Đào tạo đã ban hành các Quyết định, Thông tư, Công văn... nhằm chấn chỉnh, thống nhất, định hướng cho việc xây dựng chương trình, mục tiêu, nội dung đào tạo theo “Khung năng lực tiếng Việt dùng cho người nước ngoài”; tổ chức kiểm tra, đánh giá năng lực học tiếng Việt. Nhiều cơ sở đào tạo đã tiến hành việc xây dựng chương trình, nội dung, giáo trình phục vụ đào tạo tiếng Việt cho người nước ngoài theo Khung năng lực; thay đổi, hoàn thiện nội dung và hình thức các kỳ thi đánh giá năng lực tiếng Việt theo Định dạng đề thi đánh giá năng lực⁸. Tuy vậy, thực tế kết quả của các kỳ thi đã bộc lộ một số điểm cần cải thiện, thay đổi, nhất là quỹ đề thi vẫn còn mang tính chủ quan, thiếu tính phổ quát của vùng miền (chủ yếu sử dụng giáo trình, công cụ học tập theo phương ngữ miền Bắc – nhất là kỹ năng nghe, nói), chưa thể hiện rõ tính định lượng, khách quan, khoa học trong việc đánh giá đúng năng lực tiếng Việt theo từng cấp độ, dễ – khó.

Nguyên nhân của vấn đề này chính là việc các cơ sở đào tạo áp dụng “Khung năng lực tiếng Việt dùng cho người nước ngoài” và “Định dạng đề thi đánh giá tiếng Việt theo Khung

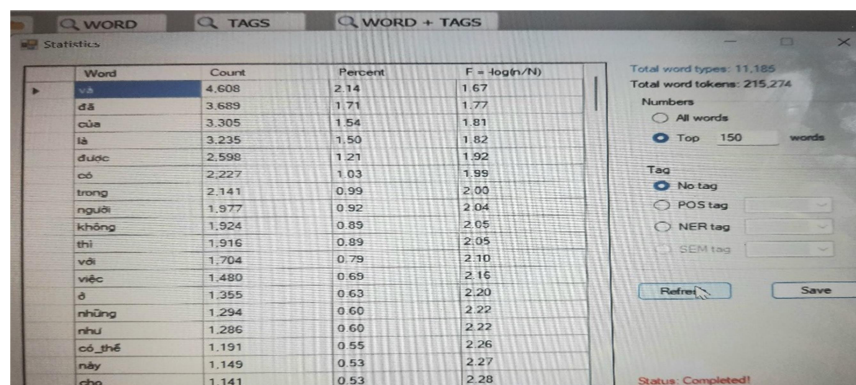
⁸ Đặc biệt tại các cơ sở đào tạo uy tín, tổ chức các kỳ thi đánh giá năng lực tiếng Việt mang tầm quốc tế, đáp ứng các nhu cầu về việc xin visa, nhập tịch tại Việt Nam như Trường Đại học Khoa học Xã hội và Nhân văn của ĐHQG Hà Nội và ĐHQG Thành phố Hồ Chí Minh; Đại học Ngoại ngữ Hàn Quốc (HUFs), Đại học Ngoại ngữ Pusan, Đại học Yongsan, Đại học Chungwoon, Hàn Quốc; Đại học Đài Loan, Đại học Sư phạm Chương Hóa, Đại học Quốc gia Thành Công, Đài Loan...

năng lực tiếng Việt dùng cho người nước ngoài” nhưng việc xây dựng khối ngữ liệu: (1) vẫn dựa vào cảm tính, chủ quan, thiếu cơ sở khoa học; (2) khối ngữ liệu còn hạn chế (vẫn sử dụng tài liệu – ngữ liệu cũ) nên việc truy xuất để biên soạn để thi chưa thể hiện sự phong phú, đa dạng và hiệu quả trong việc kiểm định, đánh giá khách quan theo từng cấp độ.

- Trên cơ sở thống kê, lựa chọn nguồn ngữ liệu từ các công trình đã công bố, bài nghiên cứu đã trích xuất các bài đọc tương ứng với trình độ A2, B1, B2 để đưa vào phần mềm xử lý⁹, phân loại độ khó của từ, câu đoạn, văn bản thông qua từ vựng (tần suất¹⁰), độ phức tạp của cấu trúc câu (số lượng mệnh đề, chiều dài câu), tổ chức văn bản (liên kết văn bản), hình thức văn bản... Kết quả phân tích cho thấy, những giáo trình, bài giảng, đề thi đánh giá năng lực tiếng Việt dùng cho người nước ngoài hiện nay chưa đảm bảo phân cấp độ phù hợp theo các tiêu chuẩn, tiêu chí của “Khung năng lực tiếng Việt dùng cho người nước ngoài” và “Định dạng đề thi đánh giá năng lực tiếng Việt theo Khung năng lực tiếng Việt dùng cho người nước ngoài”. Nhiều bài đọc mang tính chủ quan, cảm tính, sử dụng các từ, liên từ, đoạn văn bản có độ khó không phù hợp với trình độ (gồm cả hai chiều, độ khó thấp nhưng trình độ cao và độ khó cao nhưng thực tế ngữ liệu có trình độ thấp), kể cả bộ giáo trình biên soạn được xem là chuẩn nhất hiện nay dựa theo các tiêu chuẩn, tiêu chí của “Khung năng lực tiếng Việt dùng cho người nước ngoài”: Khoa Việt Nam học và Tiếng Việt, Trường ĐH KHXH và NV, ĐHQG Hà Nội (2019), “Tiếng Việt cho người nước ngoài” tập 1, 2, 3, 4 (tương ứng với trình độ A1, A2, B1, B2 của Khung năng lực).

⁹ Đã đề cập ở phần 1.1. và 2.2.

¹⁰ Ví dụ: từ phần mềm ngữ liệu hiện có (phần từ điển tần số tiếng Việt đầy đủ với 40.000 mục từ), chúng tôi đã truy xuất tần suất xuất hiện của 150 từ theo thứ tự từ nhỏ đến lớn: của, tôi, một, anh, ta, không, là, đã, có, được, những, cô, cho, trong, ấy, người, ông, đó, ở, và, các, này, làm, sẽ, với, phải, đi, chúng tôi, vào, về, khi, để, nó, đến, sự, bị, họ, có thể, cái, ra, lại, bạn, nhiều, rất, gì, nhà, việc, con, nhưng, chúng ta, bà, mình, đang, như, nào, trên, nói, mới, cuộc, hơn, vì, trước, muốn, biết, tiến, lên, từ, sau, thì, đây, chiếc, bằng, thấy, nếu, mà, nước, điều, hai, năm, chỉ, còn, không thể, hẳn, khác, ngày, lúc, rồi, nhất, rằng, nhau, tất cả, cũng, tới, quá, cách, cả, theo, ăn, công việc, cần, qua, thật, nữa, vậy, nghĩ, tiếng, xe, giờ, ai, lời, đường, thế, hay, mọi, nên, thích, phòng, xem, vấn đề, đối với, mắt, nghe, chưa, hàng, tìm, thứ, chúng, vẫn, tốt, chỗ, khỏi, lấy, dùng, sách, đưa, hoàn toàn, hãy, đưa, bao giờ, lớn, trẻ, gặp, chẳng, sống, cậu, tin, đúng, mua, đều, làm việc.



Word	Count	Percent	F = -log(n/N)
và	4,608	2.14	1.67
đã	3,689	1.71	1.77
của	3,305	1.54	1.81
là	3,235	1.50	1.82
được	2,598	1.21	1.92
có	2,227	1.03	1.99
trong	2,141	0.99	2.00
người	1,977	0.92	2.04
không	1,924	0.89	2.05
thì	1,916	0.89	2.05
với	1,704	0.79	2.10
việc	1,480	0.69	2.16
ở	1,355	0.63	2.20
những	1,294	0.60	2.22
như	1,286	0.60	2.22
có thể	1,191	0.55	2.26
này	1,149	0.53	2.27
cho	1,141	0.53	2.28

Total word types: 11,185
Total word tokens: 215,274

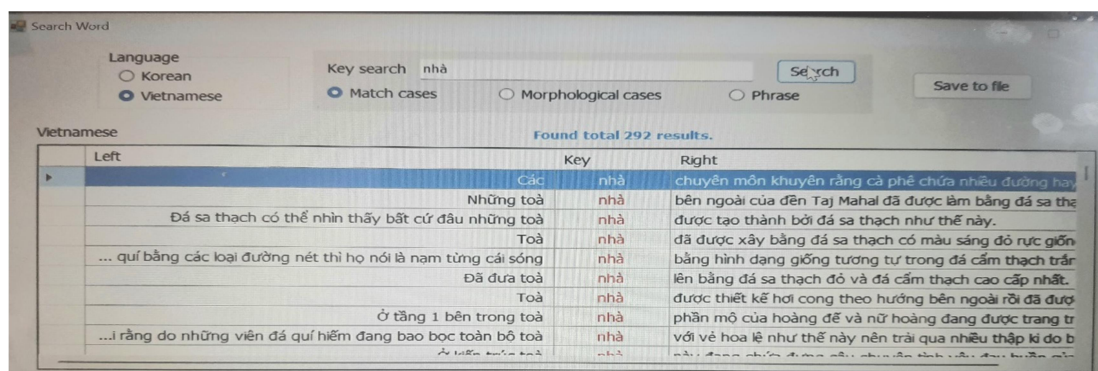
Numbers: ☐ All words ☒ Top 150 words

Tag: ☒ No tag ☐ POS tag ☐ NER tag ☐ SEM tag

Buttons: Refresh, Save

Status: Completed!

Hình ảnh 1: Tần suất xuất hiện từ trong phạm vi top 150 từ đầu tiên



Search Word

Language: ☐ Korean ☒ Vietnamese

Key search: nhà

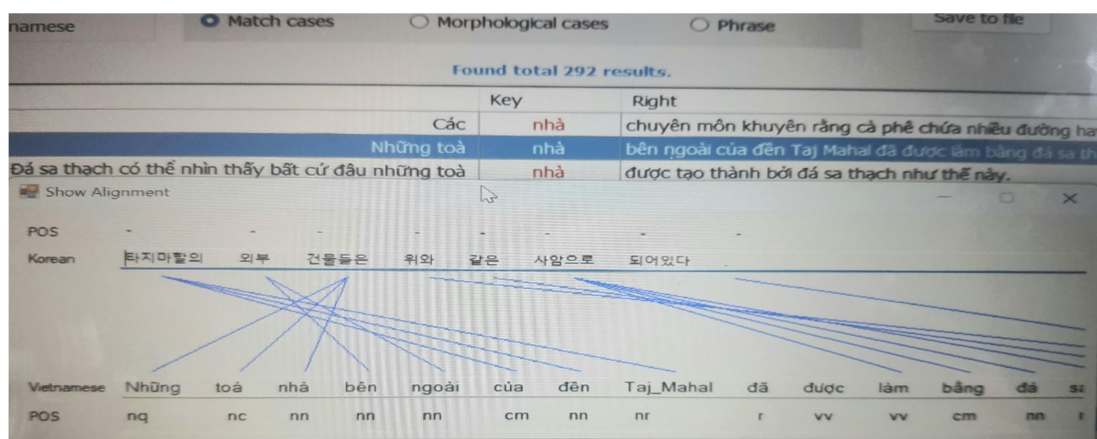
☒ Match cases ☐ Morphological cases ☐ Phrase

Buttons: Search, Save to file

Found total 292 results.

Left	Key	Right
Các	nhà	chuyên môn khuyên rằng cả phê chứa nhiều đường hay
Những toà	nhà	bên ngoài của đền Taj Mahal đã được làm bằng đá sa thạch
Đá sa thạch có thể nhìn thấy bất cứ đâu những toà	nhà	được tạo thành bởi đá sa thạch như thế này.
... quý bằng các loại đường nét thì họ nói là nam từng cái sống	Toà	nhà
Đã đưa toà	nhà	đã được xây bằng đá sa thạch có màu sáng đỏ rực giống
Toà	nhà	bằng hình dạng giống tương tự trong đá cẩm thạch trắng
ở tầng 1 bên trong toà	nhà	lên bằng đá sa thạch đỏ và đá cẩm thạch cao cấp nhất.
...i rằng do những viên đá quý hiếm đang bao bọc toàn bộ toà	nhà	được thiết kế hơi cong theo hướng bên ngoài rồi đã được
	nhà	phần mô của hoàng đế và nữ hoàng đang được trang tr
	nhà	với vẻ hoa lệ như thế này nên trải qua nhiều thập kỉ do b

Hình ảnh 2: Vận dụng tần suất xuất hiện của từ để xây dựng độ khó câu, đoạn



name: ☒ Match cases ☐ Morphological cases ☐ Phrase

Found total 292 results.

Left	Key	Right
Các	nhà	chuyên môn khuyên rằng cả phê chứa nhiều đường hay
Những toà	nhà	bên ngoài của đền Taj Mahal đã được làm bằng đá sa th
Đá sa thạch có thể nhìn thấy bất cứ đâu những toà	nhà	được tạo thành bởi đá sa thạch như thế này.

Show Alignment

POS

Korean: 타지마할의 외부 건물들은 위와 같은 사암으로 되어있다

Vietnamese: Những toà nhà bên ngoài của đền Taj_Mahal đã được làm bằng đá sa

POS: nq nc nn nn nn cm nn nr r vv vv cm nn i

Hình ảnh 3: Phân tích cấu trúc câu, độ khó câu, đoạn

Bảng thống kê ngữ liệu thu thập được sau khi xử lý

Ngữ liệu	Bài đọc		
Cấp độ	A2	B1	B2
Số bài đọc	65	115	125
Số từ trung bình	131,4	312,5	431,2
Số câu trung bình	3,1	4,7	6,3
Độ dài trung bình của câu tính theo từ	13,1	18,3	22,1
Độ khó theo tần suất xuất hiện từ	4.000	17.000	29.000

- Bài nghiên cứu đã kế thừa kết quả biên soạn các bài đọc trong các giáo trình, ứng dụng phần mềm, kỹ thuật đo độ khó từ, câu, đoạn, văn bản để định dạng lại theo trình độ tương ứng với A2, B1, B2. Nhóm tác giả đã biên soạn 300 bài đọc trên cơ sở sử dụng tần suất từ tùy từng cấp độ, tạm chia 40.000 mục từ đang có trong phần mềm, tương ứng cho 6 cấp độ trong Khung năng lực tiếng Việt dùng cho người nước ngoài để xây dựng câu, đoạn, văn bản phù hợp với từng chủ đề trong Khung năng lực, “Chương trình tiếng Việt của người Việt Nam ở nước ngoài”. Sau khi hoàn thiện bước 1, chúng tôi tiếp tục đưa đoạn văn bản/bài đọc vào phần mềm đo độ khó văn bản để chạy (với phần mềm được huấn luyện – máy học), kiểm tra, đánh giá tính khách quan, phù hợp. Quá trình này phát hiện một số lỗi, buộc chúng tôi phải chỉnh sửa lại văn bản và tiếp tục chạy phần mềm cho đến khi đáp ứng yêu cầu.

- Trên cơ sở văn bản/bài đọc được biên soạn theo phần mềm, kỹ thuật đo độ khó, chúng tôi tiếp tục biên soạn các câu hỏi tương ứng với các bài, mỗi bài có 03 câu hỏi, tương ứng với trình độ, năng lực theo “Khung năng lực tiếng Việt dùng cho người nước ngoài” và “Định dạng đề thi đánh giá năng lực tiếng Việt theo Khung năng lực tiếng Việt dùng cho người nước ngoài”.

4. Kết luận và kiến nghị

Kết quả nghiên cứu đã chỉ ra việc sử dụng ngữ liệu 300 bài đọc và 1.000 câu hỏi tương ứng được soạn mới theo tiêu chuẩn, tiêu chí, chủ đề, yêu cầu của “Khung năng lực tiếng Việt

dùng cho người nước ngoài”, “Chương trình tiếng Việt cho người Việt Nam ở nước ngoài”, “Định dạng đề thi đánh giá năng lực theo Khung năng lực tiếng Việt dùng cho người nước ngoài” và ứng dụng phần mềm, kỹ thuật đo độ khó đảm bảo tính phù hợp, định lượng, khách quan, khoa học.

Kiến nghị: Cần xây dựng, phát triển khối ngữ liệu đọc lớn hơn và ứng dụng thêm nhiều kỹ thuật, phần mềm của khoa học máy tính để đo độ khó, phân loại cấp độ theo “Khung năng lực tiếng Việt dùng cho người nước ngoài” nhằm phục vụ biên soạn giáo trình, bài giảng và quỹ đề thi đánh giá năng lực phù hợp, thống nhất.

TÀI LIỆU THAM KHẢO

1. An-Vinh Luong, Diep Nguyen, Dien Dinh (2017), “Examining the Text-length Factor in Evaluating the Readability of Literary Texts in Vietnamese Textbooks”; *9th International Conference on Knowledge and Systems Engineering (KSE), Hue, Vietnam, 2017*, pp. 36-41. DOI: 10.1109/KSE.2017.8119431.
2. Đinh Điền (2018), *Ngôn ngữ học ngữ liệu*, Nxb ĐHQG Tp. Hồ Chí Minh.
3. Nguyễn Văn Huệ (cb) (2001, 2004, 2013), *Giáo trình tiếng Việt dành cho người nước ngoài (quyển 1, 2, 3, 4)*; Nxb Giáo dục và Nxb ĐHQG Tp. Hồ Chí Minh.
4. Nguyễn Việt Hương (2013), *Tiếng Việt cơ sở dành cho người nước ngoài (quyển 1, 2)*, Nxb ĐHQG Hà Nội.
5. Nguyễn Việt Hương (2013), *Tiếng Việt nâng cao dành cho người nước ngoài (quyển 1, 2)*, Nxb ĐHQG Hà Nội.
6. Khoa Việt Nam học và Tiếng Việt, Trường Đại học Khoa học Xã hội và Nhân văn, ĐHQG Hà Nội (2019), *Tiếng Việt cho người nước ngoài (tập 1, 2, 3, 4)*, Hà Nội.
7. Tony Mc, A. W. (2001), *Corpus Linguistics*; Sil. (2017), *Language Repot*, Retrieved from <https://www.ethnologue.com>
8. Viện Việt Nam học và Khoa học phát triển, ĐHQG Hà Nội (2014), *Tiếng Việt trình độ A, B, C*, Nxb ĐHQG Hà Nội.