MỘT SỐ VẤN ĐỀ CƠ BẢN CỦA PHONG CÁCH HỌC KHỐI LIỆU

Nguyễn Thế Truyền
Trường Đại học Sư phạm Tp. HCM
nguyenthetruyen2004@yahoo.com
Ngày nhận bài: 21/5/2018, Ngày duyệt đăng: 7/8/2018

Tóm tắt

Phong cách học khối liệu (PCHKL) là một lĩnh vực nghiên cứu mới của phong cách học phương Tây đương đại, ứng dụng những kỹ thuật hiện đại của khoa học máy tính để xử lý văn bản ngôn ngữ với dung lượng lớn. PCHKL có nhiều giá trị ứng dụng vào thực tiễn nghiên cứu của Việt Nam và hứa hẹn sẽ tạo ra một bước đột phá về phương pháp và kỹ thuật nghiên cứu, đem lại cho phong cách học tiếng Việt vốn dựa trên nền tảng trực giác và phân tích thủ công những luồng sinh khí mới của sự định lượng ngữ liệu khối dựa vào sự trợ giúp của máy tính.

Nhằm giúp bạn đọc Việt Nam có những kiến thức cơ bản của PCHKL để họ tiếp tục đi sâu vào lĩnh vực này hoặc ứng dụng vào nghiên cứu trên dữ liệu thực tế, đặc biệt là dữ liệu bằng tiếng Việt, bài viết này giới thiệu những thành tựu nghiên cứu cơ bản của PCHKL phương Tây. Bài viết gồm năm phần chính: định nghĩa, cách tiếp cận và phương pháp nghiên cứu, các phần mềm công cụ, những công trình nghiên cứu tiêu biểu, thế mạnh và giới hạn của PCHKL.

Từ khóa: *Phong cách học khối liệu, phân tích định lượng, phân tích định tính, cách tiếp cận dựa vào khối liệu, cách tiếp cận được chỉ dẫn bởi khối liệu.*

Some basic issues of corpus stylistics

Abstract

Corpus stylistics is a recent field of study in the contemporary Western stylistics, applying modern computational techniques to process large articles. Corpus-based approach has many applicable values in conducting research in Vietnam and has prospects of creating a breakthrough in research methodology and techniques. It will bring to Vietnamese language learning, which is based mainly on human intuition and manual analysis, a new trend of corpus quantitative analysis with the aid of computers. In order to help Vietnamese readers have basic knowledge of corpus-based approach so that they will have an in-depth thorough understanding of this field or apply it to research on the actual corpora, especially the corpora in Vietnamese language. This article introduces the basic research achievements of Western corpus-based approach. It has five main sections, including definition, approaches, research methodology, software tools, typical research works, strengths and limitations of corpus-based approach.

Keywords: *Corpus stylistics, quantitative analysis, qualitative analysis, corpus-driven approach.*

1. Phong cách học khối liệu là gì?

Trong khoảng thời gian những năm đầu của thế kỷ XXI đến nay, cùng với sự phát triển của ngôn ngữ học khối liệu (corpus linguistics), người ta thường nói đến "bước ngoặt khối liệu" trong phong cách học ("corpus turn" in stylistics; Leech và Short 2007: tr. 286). Thuật ngữ "corpus stylistics" (PCHKL) từ đó cũng xuất hiện và cùng với nó là sự ra đời của một phân ngành phong cách học mới, hay theo một cách nhìn nhận khác, là một phương pháp luận mới trong nghiên cứu của phong cách học đương đại.

Theo Mahlberg (2016: tr. 144) thì mặc dầu thuật ngữ "corpus stylistics" chỉ mới phổ biến rất gần đây, nhưng sự ứng dụng các phương pháp nghiên cứu với sự trợ giúp của máy tính để phân tích văn bản văn chương đã có một truyền thống khá lâu liên quan với tin học văn chương (literary computing) và phong cách học máy tính (computational stylistics).

Hiện nay đang có một số vấn đề về định

nghĩa khái niệm "PCHKL". Hoặc là cách định nghĩa PCHKL khá hẹp như là sự phân tích văn bản văn chương sử dụng kỹ thuật của ngôn ngữ học khối liệu. PCHKL "quan tâm tới việc ứng dụng phương pháp khối liệu vào nghiên cứu văn chương bằng những miêu tả ngôn ngữ liên quan với sự đánh giá văn chương" (Mahlberg, 2013: tr. 5). Theo cách định nghĩa này thì PCHKL, đơn giản là ngôn ngữ học khối liệu với một đối tượng nghiên cứu khác (văn chương như một sự tương phản với ngôn ngữ phi văn chương). Mahlberg (2016: tr. 139) cho rằng: "Đối tượng nghiên cứu của PCHKL là văn bản văn chương". Thực ra, từ trước đến nay, phong cách học (bao gồm kiểu khối liệu và phi khối liệu) quan tâm không chỉ riêng phân tích văn chương. Vì thế, định nghĩa PCHKL như là sự nghiên cứu ngôn ngữ học khối liệu về ngôn ngữ văn chương là một sự thu hẹp không thể chấp nhận được" (McIntyre, 2015: tr. 61). Một cách hiểu khác: "Phong cách khối liệu là sự ứng dụng lý thuyết, mô hình, khung lý luận của phong cách học vào phân tích khối liệu" (McIntyre, 2015: tr. 61). Cách hiểu của McIntyre có phần lướt qua vai trò của hệ phương pháp ngôn ngữ học khối liệu trong PCHKL, vì một trong những vấn đề cơ bản là PCHKL dựa trên những công cụ lý thuyết và kỹ thuật của ngôn ngữ học khối liệu để nghiên cứu những vấn đề liên quan đến phong cách ngôn ngữ. PCHKL nối kết các nguyên lý của ngôn ngữ học khối liệu với phong cách học và phong cách học văn chương. PCHKL quan tâm tới việc nhận thức các khuôn mẫu sử dụng ngôn ngữ thông qua nghiên cứu định lượng một khối lượng lớn dữ liệu ngôn ngữ. PCH KL có thể được định vị trong ngữ cảnh rộng hơn là khoa học nhân văn kỹ thuật số (digital humanities) (Mahlberg, 2016: tr. 214). Ba thành tố cơ bản tạo nên diện mạo của PCHKL là:

(1) Các khối liệu văn bản điện tử,

(2) Máy tính và các phần mềm xử lý định lượng,

(3) Đích nghiên cứu phong cách học.

Nếu chỉ có hai thành tố đầu thì đó là ngôn ngữ học khối liệu, và nếu chỉ có hai thành tố sau thì đó là phong cách học máy tính.

Như nhiều người đã biết, phong cách học truyền thống tìm kiếm một giải thích nghiêm ngặt về văn bản bằng các thủ pháp phân tích thủ công với ngữ liệu rời, dung lượng nhỏ. Còn trong PCHKL, ngữ liệu dùng để phân tích là khối liệu[1] văn bản điện tử có dung lượng thường hàng chục vạn từ trở lên (gồm nhiều văn bản hoàn chỉnh, hoặc những trích đoạn dài của những văn bản lớn) và được xử lý tự động hoặc bán tự động bằng các chương trình máy tính.

Xử lý định lượng để đi đến kết quả nghiên cứu trong phong cách học là con đường của lập luận quy nạp và là sự khái quát hóa theo xác suất (probabilistic generalization), mà trong PCHKL đó là "sự khái quát hóa từ những khối văn bản dung lượng lớn" (McIntyre, 2015: tr. 59). Cách tiếp cận định lượng nhấn mạnh tầm quan trọng của chứng cứ ngôn ngữ và sự đòi hỏi của cách phân tích phong cách học khách quan và khoa học hơn. Trong những cách xử lý định lượng đã có của phong cách học thì cách xử lý định lượng dựa trên khối liệu và bằng phần mềm máy tính, hiện nay là cách xử lý tối ưu hơn cả, vì theo nhiều nhà phong cách học máy tính thì "sự quan sát thông thường của con người không thể nắm bắt được tiến trình của nhiều nhân tố và đặc điểm khác nhau tạo nên một phong cách, và những nghiên cứu dựa trên sự quan sát không có sự trợ giúp của thiết bị kỹ thuật rất dễ bị ảnh hưởng bởi thiên kiến của người quan sát" (Steward, 2006: tr. 1758). Khối liệu "giúp xem xét các hiện tượng ngôn ngữ từ điểm nhìn số liệu tần số và sự liên hệ của các mô hình (trong văn bản, tập hợp văn bản)" (Mahlberg, 2013: tr. 7). Như vậy, PCHKL giúp xử lý các vấn đề của văn bản dung lượng lớn trong một khoảng thời gian ngắn. Nó mang lại tính khách quan mà các nhà phong cách học tìm kiếm. PCHKL giúp xác nhận hoặc phủ nhận những nhận định trực giác và thẩm tra những nhận định phê bình văn học. Nó còn giúp chỉ ra những đặc tính của văn bản mà phân tích thủ công bỏ sót hoặc do sức người

[1] Bất kỳ một khối liệu nào cũng có thể dùng để tìm kiếm cho một từ rời (ví dụ, "ngày"), một tập hợp từ (ví dụ, "ngày, tháng, năm"), một chuỗi từ (ví dụ, "ngày hôm sau"), hoặc để tìm kiếm một từ đã được dán nhãn từ loại, chức vụ cú pháp, hoặc một chuỗi nhãn, như "giới từ + danh từ + định ngữ", hoặc tìm kiếm một từ theo sau hay đứng trước một lớp từ đã cho nào đó. Những khối liệu đã có sự phân tích từ loại, cú pháp (tự động hay thủ công) sẽ cho phép tìm kiếm một mệnh đề cụ thể hay cấu trúc câu cụ thể, chẳng hạn tìm những câu chứa mệnh đề *If* đứng trước hay đứng sau mệnh đề chính. Khối liệu có thể được dán nhãn cho những loại thông tin khác (phục vụ cho việc xử lý của máy tính) như các phạm trù ngữ nghĩa, liên kết, lời nói nghệ thuật, phép tu từ,...

không thể giải quyết được.

Theo một số nhà nghiên cứu, có ba kiểu PCHKL:

(1) Dùng công cụ ngôn ngữ học khối liệu và kỹ thuật phân tích để kiểm nghiệm (test) các nhận định phê bình văn chương.

(2) Dùng một mô hình phong cách học để tạo ra các thông tin dán nhãn (cho khối liệu), sau đó phân tích kết quả bằng cách dùng khung lý luận phong cách học.

(3) Nghiên cứu phong cách học truyền cảm hứng từ khối liệu.

Về tư cách của PCHKL, có tác giả xem nó là một phân ngành phong cách học, nhưng có tác giả chỉ xem nó như là một phương pháp luận nghiên cứu hay một hướng tiếp cận, hoặc đơn giản xem nó là một lĩnh vực nghiên cứu mới của phong cách học. Trong bài viết vào năm 2011, Mahlberg và McIntyre (tr. 205) xem phong cách học như là một phân ngành: "Quả thực, trong những năm gần đây, việc dùng kỹ thuật khối liệu trong phong cách học đã trở nên phổ biến hơn, đem đến sự xuất hiện của một *phân ngành phong cách học mới* (new sub-branch of stylistics) ngày càng được biết đến rộng rãi là PCHKL". Nhưng đến năm 2015, thì chính tác giả McIntyre (tr. 59) lại cho rằng "xem PCHKL và phong cách học tri nhận (cognitive stylistics) như những phân ngành của phong cách học sẽ có tác dụng ngược, lợi ít hại nhiều". Theo McIntyre, tốt hơn hết là nên xem chúng như một kiểu phân tích phong cách học mà mỗi cái chỉ tập trung vào một phạm vi đặc thù. Phong cách học tri nhận quan tâm tới tiến trình và bộ máy giải mã trong quá trình đọc, còn PCHKL thì quan tâm tới vấn đề phong cách trên một khối liệu văn bản cấp độ lớn. Còn Studer (2008: tr. 6) cũng như Semino và Short (2004) chỉ xem PCHKL "như một hướng tiếp cận rất hữu ích khác, thêm vào cho kho công cụ phân tích của các công trình nghiên cứu phong cách học".

2. Cách tiếp cận, phương pháp và kỹ thuật nghiên cứu

2.1. Cách tiếp cận

Giống như ngôn ngữ học khối liệu, PCHKL cũng có hai hướng tiếp cận là cách tiếp cận dựa vào khối liệu (corpus-based approach) và cách tiếp cận được chỉ dẫn bởi khối liệu (corpus-driven approach).

Cách tiếp cận dựa vào khối liệu bắt đầu với những giả định và mô hình ngôn ngữ đã có, vì thế trong trường hợp này, kỹ thuật khối liệu được dùng chủ yếu để kiểm nghiệm và xác nhận những mô hình đã tồn tại; trong khi đó, cách tiếp cận được chỉ dẫn bởi khối liệu trao sự ưu tiên cho khối liệu và tìm kiếm các phạm trù và mô hình ngôn ngữ trên cơ sở những kiểu mẫu xuất hiện trong dữ liệu (Mahlberg, 2013: tr. 13).

Hai cách tiếp cận này đều có vai trò riêng của chúng trong nghiên cứu phong cách học (vai trò kiểm nghiệm và vai trò khám phá). Tuỳ theo mục đích nghiên cứu mà nhà phong cách học lựa chọn cách tiếp cận nào cho phù hợp với công việc của mình.

2.2. Phương pháp nghiên cứu

PCH KL sử dụng hai phương pháp nghiên cứu chính là: (a) phân tích và so sánh định lượng, (b) phối hợp phân tích định lượng và phân tích định tính.

2.2.1. Phân tích và so sánh định lượng

Theo Mahlberg (2014: tr. 382) thì "Phong cách học văn chương quan tâm đến chức năng nghệ thuật của văn bản và ấn tượng mà văn bản tạo ra trong ý nghĩ người đọc. Trọng tâm của nó là về những cách dùng ngôn ngữ đặc biệt. Những hình thể đặc biệt như thế về hình thức và nghĩa sẽ được nghiên cứu tốt nhất trong một văn bản đơn lẻ hoặc ngay cả trong một trích đoạn văn bản, vì thế phong cách học văn chương, trong nghĩa này, liên hệ về bản chất với phương pháp đọc sâu (close reading)". Nói rộng ra, trong phong cách học truyền thống, phương pháp nghiên cứu cơ bản là phân tích định tính thủ công.

Khi chuyển từ phong cách học truyền thống sang PCHKL thì phương pháp nghiên cứu của phong cách học có sự dịch chuyển trọng tâm từ định tính sang định lượng, ưu tiên nhiều hơn cho phương pháp định lượng. Trong PCHKL, phân tích định lượng bao gồm một phạm vi rất rộng (rộng hơn trong ngôn ngữ học khối liệu hay phong cách học máy tính, trắc học phong cách) từ từ vựng, ngữ pháp, tu từ, đến diễn ngôn,… nói chung là tất cả các phương diện ngôn ngữ có thể khảo sát bằng máy tính nhằm tìm ra các đặc điểm phong cách của một tác giả, một giai đoạn lịch sử hay của một biến thể chức năng nào đó. Công việc định lượng đó thường dựa vào các tính năng có sẵn trong các phần mềm ngôn ngữ học khối liệu hay trắc học phong cách để

thực hiện tự động, hoặc dán nhãn thủ công sau đó nhờ máy tính xử lý (bán tự động). Với trình độ kỹ thuật máy tính hiện nay thì phần việc thủ công khi nghiên cứu phong cách khối liệu còn tương đối nhiều.

Vì so sánh là nền tảng cơ bản của nghiên cứu phong cách học nên phân tích định lượng trong PCHKL luôn gắn liền với so sánh định lượng. Trong PCHKL, khối liệu dùng để phân tích định lượng được gọi là khối liệu đích (target corpus; hay cũng gọi là khối liệu mẫu – sample corpus), còn khối liệu dùng để đối chứng gọi là khối liệu tham chiếu (referenced corpus) trong trường hợp là một khối liệu tổng quát hơn. Trong trường hợp khối liệu tham chiếu ngang cấp với khối liệu đích thì thường gọi là khối liệu so sánh (comparative corpus) hay khối liệu đối chiếu (contrastive corpus). Theo Mahlberg (2014: tr. 388) thì "Điều cốt yếu đối với phân tích PCHKL là so sánh một khuôn hình cụ thể hay một văn bản cụ thể với dữ liệu tham chiếu". Việc so sánh một trường hợp cụ thể với một mẫu tổng quát hơn như một toàn thể (được thể hiện trong khối liệu tham chiếu) liên quan đến khái niệm "lệch chuẩn" (deviation) được dùng trong phong cách học để đo đạc khuôn hình sáng tạo của một tác giả trong văn bản văn chương và giải thích hiệu quả của sự lạ hóa (foregrounding). Theo Ho (2011: tr. 6) thì "Những hiện tượng nổi bật về tần số hoặc hiếm thấy hơn trong văn bản so với những văn bản khác được xác định như những "chỉ dấu phong cách" (style maker) cho văn bản đó. Tỷ lệ cao hơn của một hiện tượng ngôn ngữ xuất hiện đi xuất hiện lại trong văn bản văn chương biểu thị sự nhấn mạnh thẩm mỹ nào đó. Chúng đưa lại cho văn bản tính chất độc đáo về phong cách. Những "chỉ dấu phong cách" là những vật chuyển tải nghĩa hoặc là dấu vết của những mã bị dấu kín trong văn bản. Một số chỉ dấu phong cách có sự điều khiển có ý thức của tác giả trong khi một số chỉ dấu phong cách khác lại được dùng một cách vô thức.

2.2.2. Phối hợp phân tích định lượng với phân tích định tính

PCHKL không phải thuần túy là một sự nghiên cứu định lượng về văn chương và về văn bản các loại nói chung. Nó "vẫn là cách tiếp cận phong cách học định tính (qualitative stylistic approach) đối với nghiên cứu ngôn ngữ của văn chương, kết hợp với hoặc được ủng hộ bởi kỹ thuật và phương pháp định lượng dựa trên khối liệu" (Ho, 2011: tr. 10). Như vậy trong PCHKL, "Định lượng và thống kê phải luôn luôn được sử dụng như một phương tiện hơn là một mục đích, để xác nhận hoặc để bác bỏ những phân tích dựa trên trực giác (intuition-based analysis) của chúng ta" (Ho, 2011: tr. 11).

Quan điểm chung của các nhà nghiên cứu là phương pháp nghiên cứu của PCHKL phải phối hợp chặt chẽ giữa phân tích định lượng và phân tích định tính, trong đó phương pháp phân tích định tính là gốc rễ, còn phương pháp phân tích định lượng là công cụ để kiểm nghiệm hay củng cố kết quả của phương pháp định tính. Như Carter[2] đã nhận định: "Phân tích PCHKL là một thủ tục phương pháp luận khá khách quan mà khâu tốt nhất của nó được một tiến trình giải thích khá chủ quan hướng dẫn". Phân tích PCHKL phải chú trọng không chỉ một mình các nhân tố định lượng của văn bản đích, mà còn phải nhạy cảm với những phương diện của văn bản cần được phân tích định tính nhiều hơn. Đi theo hướng tiếp cận này, Simino và Short cho rằng, "Đã có một số công trình nghiên cứu thú vị, kết hợp một cách nhuần nhuyễn công việc phân tích định tính một văn bản cụ thể với việc phân tích dựa trên khối liệu" và phân tích định tính văn bản vẫn luôn là "trung tâm của lĩnh vực phân tích phong cách học" (Simino và Short, 2004: tr. 7).

Một vấn đề đặt ra ở đây là vai trò của trực giác trong phân tích PCHKL. Milic từng cho rằng, phương pháp nghiên cứu của phong cách học là "bắt đầu với trực giác" và kết thúc với "dữ liệu cụ thể"[3]. Trực giác phong cách học có tác dụng đặt ra những giả thuyết, ý tưởng phân tích. Công việc còn lại của nhà nghiên cứu là "phát triển và thao tác hóa những trực giác phong cách học"[4] đó. PCHKL cần phải kết hợp "kỹ thuật phân tích dựa trên khối liệu với những cách tiếp cận dựa nhiều vào trực giác" (Simino and Short, 2004: tr. 8). Một sự tương tác hữu

[2] Carter R. (2010). *Methodologies for stylistic analysis: Practices and pedagogies*. Dẫn theo: Mahlberg (2014: tr. 379).

[3] Milic L. (1967). *A quantitative approach to the style of Jonathan Swift*. The Hague: Mouton. Dẫn theo: Steward (2006: tr. 1758).

[4] van Peer W. (1989). Quantitative studies of literature: a critique and an outlook. *Computers and the Humanitie*s, 23 (4/5). Dẫn theo: Ho (2011: tr. 205).

hiệu giữa hai đường hướng nghiên cứu này phải không dẫn tới làm mất đi hoặc là tính nghiêm ngặt của phương pháp luận khoa học, hiện đại hoặc là sự tinh tế của trực giác[5] nhạy cảm của não bộ con người.

Sự kết hợp chặt chẽ phương pháp định tính và phương pháp định lượng trong PCHKL sẽ giúp chúng ta đạt được cấp độ cao hơn trong hiểu biết tác phẩm văn chương hoặc bất cứ dữ liệu nào được lựa chọn. Sự kết hợp của hai phương pháp đó cũng là nguyên lý hạt nhân của phong cách học dòng chủ lưu.

2.3. Kỹ thuật nghiên cứu

PCHKL sử dụng một số kỹ thuật nghiên cứu mang tính chất đặc trưng. Sau đây, bài viết giới thiệu một số kỹ thuật định lượng (thực hiện tự động) mà PCHKL thường sử dụng. Một số kỹ thuật đó vốn có sẵn trong ngôn ngữ học khối liệu hay trắc học phong cách.

Wordlist (Danh sách từ): xác định danh sách tất cả các từ được dùng trong khối liệu và tần số của chúng. Nếu khối liệu là toàn bộ tác phẩm của một tác giả thì danh sách từ chính là vốn từ vựng của tác giả đó.

Keywords (Từ khóa): xác định những từ xuất hiện với tần số nhiều hơn hay ít hơn so với một khối liệu tham chiếu. Từ khóa nói lên khuynh hướng nghệ thuật về lựa chọn chủ đề, đề tài hay phong cách tác giả, phong cách thể loại hoặc biến thể chức năng. Có ba loại từ khóa cơ bản là tên riêng, từ chức năng và từ nội dung. Trong đó, theo một số tác giả, từ chức năng có nhiều giá trị trong nghiên cứu phong cách tác giả, vì từ nội dung xuất hiện phụ thuộc nhiều vào đề tài và chủ đề văn bản. Từ khóa được các máy tính tự động tạo ra bằng cách so sánh danh sách từ của khối liệu đích và khối liệu tham chiếu. Vì khối liệu tham chiếu được xem là cung cấp chuẩn (norm), nên "từ khóa là những từ có tần số lệch so với chuẩn" (Mahlberg và McIntyre, 2011: tr. 207), và sự lệch chuẩn như thế rất được quan tâm nghiên cứu trong phong cách học nơi mà khái niệm "lạ hóa" được dùng để miêu tả như những hiệu quả hình thành từ sự lệch chuẩn ngôn ngữ. Việc nghiên cứu từ khóa trong so sánh với chuẩn khối liệu tham chiếu sẽ giúp nhận diện những hiện tượng ngôn ngữ mang những đặc điểm phong cách có giá trị cho những phân tích định tính sâu hơn. Từ khoá, theo Mahlberg và McIntyre (2011: tr. 207), "cần được xem như những dấu hiệu về việc xây dựng thế giới hư cấu cũng như những nút bấm về chủ đề của tiểu thuyết".

Concordance (Chỉ mục): liệt kê tất cả các ngữ cảnh xuất hiện của từ ngữ cần phân tích. Chỉ mục cho phép chúng ta biết tất cả các câu có từ ngữ cần phân tích, xếp theo thứ tự ABC của từ bên trái hay bên phải của từ cần phân tích. Chỉ mục giúp nhà nghiên cứu dễ dàng hình dung ra các mô hình cần thiết qua các ngữ cảnh xuất hiện đó. Trọng tâm của việc phân tích chỉ mục là các khuôn mẫu mang tính chức năng và ngữ cảnh hơn là những con số thống kê về tần số. Văn bản, ở ngoài khối liệu, được đọc theo chiều ngang, nhưng văn bản trong khối liệu, qua sự trình diễn chọn lọc của chỉ mục, được đọc theo chiều dọc (Mahlberg, 2013: tr. 7). Phân tích chỉ mục "bổ sung và truyền sức mạnh cho việc đọc sâu một văn bản" (Mahlberg, 2013: tr. 8).

Cluster[6] (Chuỗi từ lặp lại): xác định các chuỗi từ lặp lại trong khối liệu (chuỗi 3 từ hay 4 từ, 5 từ, cho đến chuỗi 9 từ, 10 từ). Chuỗi từ lặp lại phản ánh phong cách ngôn ngữ của nhân vật, tác giả, thời đại hoặc phản ánh những công thức, khuôn mẫu diễn đạt của các biến thể chức năng. Chuỗi từ càng dài thì tần số xuất hiện càng thấp. Sự khác biệt giữa các chuỗi từ còn ở chỗ chúng xuất hiện trong một hay nhiều tác phẩm hoặc thể loại khác nhau.

Sentence length (Độ dài của câu): thống kê tổng số lượng câu trong khối liệu và độ dài của các câu đó (tính theo đơn vị từ). Thông tin này cho người nghiên cứu biết khuynh hướng cấu tạo câu của tác giả hay một lĩnh vực giao tiếp (dài, trung bình hay ngắn).

Key semantic domain (Miền ngữ nghĩa chìa khoá): xác định những miền ngữ nghĩa quan trọng, nổi bật của khối liệu đích (tác phẩm hay loạt tác phẩm cần nghiên cứu). Miền ngữ nghĩa

[5] "Một trong những mối quan tâm lớn nhất của phong cách học là kiểm tra và xác nhận tính hợp lệ của trực giác bằng những phân tích chi tiết, nhưng phong cách học cũng là cuộc đối thoại giữa người đọc văn chương và nhà quan sát ngôn ngữ, trong đó sự thấu hiểu, không hoàn toàn khách quan, là mục đích. Phân tích ngôn ngữ không thay thế được cho trực giác của người đọc, cái mà Spitzer gọi là cú nhấp chuột (click) trong tâm trí, nhưng nó có thể thúc đẩy, điều khiển, định hình trực giác vào trong một cách hiểu " (Leech và Short, 2007: tr. 4).

[6] cluster: cũng gọi là n-gram hoặc lexical bundle.

(semantic domain) là những trường được đặc trưng bởi những từ liên hệ chặt chẽ với nhau về mặt từ vựng trong văn bản. Ví dụ: *hoa, cây, sự quang hợp* thuộc miền ngữ nghĩa thực vật. Miền ngữ nghĩa được phần mềm Wmatrix dán nhãn tự động. Sau đó, miền ngữ nghĩa chìa khoá sẽ được Wmatrix tạo ra trên cơ sở so sánh miền ngữ nghĩa của khối liệu đích với một khối liệu tham chiếu (chẳng hạn so sánh với phần văn xuôi hư cấu trong khối liệu quốc gia Anh – BNC). Như vậy, miền ngữ nghĩa chìa khoá được hiểu là những miền ngữ nghĩa nổi trội hơn so với khối liệu tham chiếu. Miền ngữ nghĩa chìa khoá gợi ra các tuyến chủ đề, các liên hệ phát triển cốt truyện và giúp cho người nghiên cứu có những định hướng phân tích về nội dung hay cách thức biểu hiện của tác phẩm.

Các kỹ thuật định lượng vừa liệt kê ở trên cũng được một số nhà phong cách học gọi là "bảng liệt kê những mục cần kiểm tra" (checklist) của PCHKL (Mahlberg và McIntyre, 2011: tr. 206), hay nói cách khác, là những đối tượng khảo sát của PCH KL.

3. Một số phần mềm công cụ

Hiện nay, PCHKL chỉ mới sử dụng các phần mềm máy tính vốn được thiết kế cho ngôn ngữ học khối liệu hay trắc học phong cách. Sau đây là một số phần mềm được nhiều nhà PCHKL hiện nay sử dụng.

3.1. WordSmith Tools

WordSmith Tools là phần mềm phân tích từ vựng. Nó là một bộ tích hợp nhiều chương trình giúp khảo sát hoạt động của từ trong văn bản. Trong các chương trình tích hợp đó thì nòng cốt là ba chương trình WordList, Keywords, và Concord; trong đó, WordList cung cấp danh sách của tất cả các từ trong văn bản (hoặc một loạt văn bản), sắp xếp chúng theo tần số hay theo thứ tự ABC; Keywords gọi ra các từ khóa trong khối liệu theo một tiêu chuẩn thống kê nào đó (chẳng hạn xuất hiện ít hơn hay nhiều hơn mức quy định so với một khối liệu tham chiếu) bằng cách đối chiếu hai WordList của hai khối liệu; Concord cho phép chúng ta quan sát bất cứ từ hay cụm từ cần nghiên cứu nào đó trong tất cả các ngữ cảnh xuất hiện của chúng bằng các đường chỉ mục (concordance line). Ngoài ra, WordSmith còn có các chương trình công cụ tiện dụng khác. WordSmith Tools do nhà ngôn ngữ học người Anh - Mike Scott xây dựng nên và công bố vào năm 1996. Phiên bản hiện nay của WordSmith Tools là 7.0 (2018). WordSmith Tools là gói phần mềm xử lý được nhiều ngôn ngữ (82 ngôn ngữ, trong đó có tiếng Việt) và chạy trong hệ điều hành Windows.

WordSmith Tools là phần mềm vận dụng phổ biến trên thế giới cho những công trình dựa trên hệ phương pháp của ngôn ngữ học khối liệu và cũng là công cụ quan trọng để các nhà phong cách học dùng phân tích các văn bản điện tử về phương diện từ vựng nhằm tìm ra các đặc trưng phong cách của một đối tượng nghiên cứu nào đó.

3.2. Wmatrix

Wmatrix là một công cụ phần mềm dùng cho việc phân tích và so sánh khối liệu. Wmatrix cung cấp một giao diện web cho hai công cụ chú thích khối liệu tiếng Anh CLAW và USAS (nhằm dán nhãn cho các yếu tố cần nghiên cứu trong văn bản), và những kỹ thuật ngôn ngữ học khối liệu như danh sách tần số, chỉ mục. Nó cũng mở rộng kỹ thuật từ khóa cho những phạm trù ngữ pháp chìa khóa và miền ngữ nghĩa chìa khóa. Wmatrix cho phép người dùng chạy hai công cụ chú thích trên những trình duyệt web như Chrome, Firefox hoặc Internet Explorer. Wmatrix do Paul Rayson sáng chế.

Trong Wmatrix, CLAW là công cụ để dán nhãn từ loại cho từ (POS – part of speech). USAS là hệ thống chú thích ngữ nghĩa. Nó gán mã biểu thị đặc điểm ngữ nghĩa hoặc trường ngữ nghĩa của từ trong văn bản. Wmatrix cho phép người dùng tải lên dữ liệu khối liệu của chính họ và chạy hai công cụ chú thích này qua một trình duyệt web. Mỗi khi người dùng tải dữ liệu của mình lên thì Wmatrix tự động thực hiện các bước dán nhãn ngữ nghĩa và từ loại, sau đó tạo ra một tập hợp danh sách tần số các từ, các nhãn từ loại và các nhãn ngữ nghĩa cho dữ liệu.

Chức năng quan trọng nhất của Wmatrix là cho phép người dùng so sánh các danh sách tần số. Việc so sánh này giúp thấy sự khác nhau về từ khóa (keyword) ở cấp độ từ, sự khác nhau về lớp từ khóa (key word class) ở cấp độ từ loại (POS), hoặc sự khác nhau về khái niệm chìa khóa (key concept) ở cấp độ ngữ nghĩa. Hiện tại, Wmatrix là công cụ chính để phân tích miền ngữ nghĩa chìa khóa (key semantic domain). Wmatrix hiện chưa có phiên bản dùng cho tiếng Việt.

3.3. Signature

Signature là một phần mềm được thiết kế phục vụ cho việc phân tích trắc học phong cách (stylometric analysis) một cách thuận tiện với sự nhấn mạnh về việc nhận diện tác giả qua những dấu hiệu đặc trưng (như tên gọi của chương trình này – signature). Signature cho phép người dùng lựa chọn một số công cụ để khám phá văn bản văn chương và các loại hình văn bản điện tử khác. Các thông số phân tích như tần số độ dài của từ, độ dài của câu, độ dài đoạn văn, ký tự chữ cái, dấu câu, danh sách từ khóa trong văn bản sẽ được hiển thị và trực quan hóa qua các đồ thị hai chiều hay ba chiều.

Việc phân tích tần số của những từ cụ thể cũng được Signature tiến hành một cách hiệu quả qua các công cụ Wordlists, Key words. Hiện nay, Signature chưa có phiên bản dùng cho xử lý văn bản tiếng Việt.

4. Những công trình nghiên cứu tiêu biểu

Trong phần này, chúng tôi giới thiệu 4 công trình tiêu biểu của PCHKL (đã in thành sách) bàn về những vấn đề chung và những vấn đề chuyên biệt của PCHKL, xuất bản từ 2004 đến 2013.

Corpus Stylistics and Dickens's Fiction (Phong cách học khối liệu và tiểu thuyết của Dickens)

Tác giả quyển sách là Michaela Mahlberg (Đại học Nottingham, Anh), Routledge xuất bản năm 2013, 221 trang. Sách vừa giới thiệu chung về PCHKL, vừa là một nghiên cứu trường hợp về tiểu thuyết Dickens. Mục đích chính của quyển sách là "khám phá mối quan hệ giữa các tập hợp ngôn ngữ với phần đóng góp chúng có thể tạo ra đối với hiệu quả mà văn bản có ở người đọc" (Mahlberg, 2013: tr. 6).

Cách tiếp cận của quyển sách là cách tiếp cận được chỉ dẫn bởi khối liệu và khối liệu dùng để khảo sát chủ yếu là khối liệu không dán nhãn. Phần mềm được sử dụng là WordSmith Tools, phiên bản 5.0. Khối liệu đích là 23 tác phẩm của Dickens, trong đó có 15 tiểu thuyết, gồm 4,5 triệu từ. Khối liệu đối chiếu: 29 tiểu thuyết của 18 tác giả thế kỷ XIX, gồm 4,5 triệu từ. Nội dung chủ yếu của sách là phân tích những chuỗi từ lặp lại (cluster) mà Dickens thường khai thác như những chỉ dấu về nghĩa và chức năng văn bản. Mahlberg tập trung phân tích những chuỗi 3 từ hoặc 4 từ, 5 từ được lặp lại trong tác phẩm của Dickens, như *as if he, I don't know what, you don't mean to say, his hands in his pockets, young lady with black* ở 5 phương diện: danh xưng, lời nói của nhân vật, ngôn ngữ cơ thể, lời bình luận của người trần thuật, thời gian và nơi chốn. Chẳng hạn theo Mahlberg, chuỗi từ lặp lại trong lời nói của nhân vật nêu lên công thức lịch sự (*am delighted to see you*), thông tin thương lượng (*you don't mean to say*), sự đối đầu (*what do you want here*), sự chiếm giữ lượt lời (*I was going to say*). Chúng có thể mang hiệu quả hài hước khi diễn tả sự lặp lại hay hành động theo thói quen của nhân vật. Những chuỗi từ này liên quan đến kỹ thuật ngoại hiện (*externalisation techniques*) trong xây dựng nhân vật của Dickens.

Công trình *Corpus Stylistics and Dickens's Fiction* là một trong những tác phẩm tiên phong trong việc dùng hệ phương pháp của ngôn ngữ học khối liệu để làm rõ mối tương quan giữa khuôn mẫu diễn đạt với nghĩa và chức năng tổ chức văn bản, với việc miêu tả nhân vật, xây dựng thế giới hư cấu và phong cách ngôn ngữ tác giả.

Corpus stylistics: speech, writing and thought presentation in a corpus of English writing (Phong cách học khối liệu: Việc miêu tả lời nói và ý nghĩ qua một khối liệu tiếng Anh)

Công trình này của Elena Simino và Mick Short (Đại học Lancaster, Anh), Routlege xuất bản năm 2004, 256 trang đã ứng dụng phương pháp luận khối liệu để khám phá và kiểm nghiệm các khuôn mẫu trình bày lời nói và ý nghĩ (của nhân vật) trong ba thể loại là văn xuôi hư cấu, bản tin trên báo, tiểu sử/ tự truyện.

Trước khi đi vào nghiên cứu cụ thể, các tác giả giới thiệu cấu trúc khối liệu và hệ thống chú thích (dán nhãn)[7] khối liệu mà họ phát triển. Theo đó, khối liệu mà hai nhà nghiên cứu dùng để khảo sát gồm 120 mẫu văn bản, dài tổng cộng là 258.348 từ của văn viết tiếng Anh thế kỷ XX, chia làm ba phần cho ba thể loại trần thuật vừa nói ở trên. Khác với Mahlberg (2013), khối liệu nghiên cứu của hai nhà nghiên cứu này là khối liệu có dán nhãn. Cách tiếp cận của hai tác giả này là cách tiếp cận dựa vào khối liệu (corpus-based approach), vì xuất phát điểm của họ là các

[7] annotation or tagging.

mô hình lý thuyết của Leech và Short (1981)[8] về việc miêu tả lời nói và ý nghĩ của nhân vật trong văn xuôi tự sự. Khối liệu được hai tác giả này dùng để kiểm nghiệm và cải biên mô hình của Leech và Short.

Theo Simino và Short thì trong ngôn ngữ học khối liệu, việc chú thích khối liệu cho những hiện tượng ngôn ngữ như cấu tạo từ, từ loại, một số phạm trù ngữ nghĩa thường được thực hiện một cách tự động (do máy tính tham chiếu file dữ liệu chứa thông tin có sẵn và tự động gắn nhãn cho sự kiện ngôn ngữ tương ứng xuất hiện trong khối liệu). Tuy nhiên, một số phạm trù ngôn ngữ phức tạp, trừu tượng mà không có tiêu chuẩn hình thức xác định thì phải thực hiện việc dán nhãn một cách thủ công hoặc bán tự động. Cho nên, hai tác giả này cho biết, trong phong cách học, để chú thích (dán nhãn) khối liệu, trước hết, nhà phong cách học phải xây dựng hệ thống phân loại cho phạm trù mà mình nghiên cứu. Bước tiếp theo là kiểm nghiệm và hoàn thiện hệ thống phân loại trên một số mẫu thí điểm. Bước thứ ba là xây dựng nguyên tắc chỉ đạo và bộ công cụ cho việc dán nhãn. Bước thứ tư là dán nhãn thủ công hoặc bán tự động. Trong trường hợp của hai tác giả Simino và Short là dán nhãn thủ công (nhưng hai tác giả cũng hy vọng trong tương lai sẽ có người sáng chế được phần mềm dán nhãn tự động cho việc nghiên cứu này). Việc dán nhãn này hơi khác so với công việc tương tự trong ngôn ngữ học khối liệu cả về sản phẩm lẫn tiến trình vì đối tượng nghiên cứu là những hiện tượng diễn ngôn ở cấp độ cao và phức tạp (Simino và Short 2004: tr. 26). Việc nhận diện các đối tượng cần dán nhãn phải luôn được đặt trong một môi trường văn bản cụ thể, trên cơ sở tham chiếu về những yếu tố khá tinh tế về dụng học và ngữ cảnh.

Corpus Stylistics in Principles and Practice (Phong cách học khối liệu: Nguyên lý và thực tiễn)

Tác giả sách là Yufang Ho (Đại học Huddersfield, Anh), Continuum xuất bản năm 2011, 254 trang. Mục đích chính của quyển sách là khám phá những khác biệt về phong cách giữa hai phiên bản của tiểu thuyết hậu hiện đại, siêu trần thuật *The Magus* (Pháp sư) của nhà văn John Fowles (Anh) qua phân tích, so sánh giữa hai bản in (1966 và 1977). Sách gồm ba phần:

Phần I: *Những nguyên lý chung*. Những vấn đề quan trọng nhất, có tính chất nền tảng về PCHKL được Ho giới thiệu trong chương này.

Phần II: *PCHKL trong thực tiễn – một phân tích so sánh về The Magus*. Phần II gồm sáu chương với những nội dung quan trọng như: đo lường mức độ tương đương giữa các phiên bản; so sánh mật độ và mô hình ngôn ngữ hình tượng; sự khác biệt về phong cách giữa *The Magus* và phiên bản của nó.

Phần mềm dùng để đo lường mức độ tương đương về văn bản giữa hai bản in của *The Magus* (bản đầu tiên và bản sửa chữa lại) là TESAS/Crouch[9] và Wcopyfind[10] nhằm tìm ra những thay đổi giữa hai phiên bản và những vấn đề liên quan đến phong cách và sự lựa chọn. Ngoài ra, phần mềm Wmatrix cũng được tác giả dùng để thực hiện "việc so sánh ngôn ngữ vĩ mô ở cấp độ ngữ nghĩa", để kiểm nghiệm giả thuyết liên quan đến "điểm nhìn trần thuật" và "sự phong chiếu thế giới văn bản" giữa hai phiên bản (Ho, 2011: tr. 200). Phần mềm WordSmith Tools cũng được Ho dùng để kiểm nghiệm giả thuyết của mình về những thay đổi của ngôn ngữ hình tượng[11] giữa hai phiên bản. Cách làm của Ho là dùng Concordance phát hiện các dữ liệu ngôn ngữ liên quan (dựa trên những tiêu chuẩn hình thức mà người phân tích xác lập), sau đó trích dẫn, sắp xếp và trình diễn trong một cách thức để người phân tích dễ dàng phát hiện ra các mô hình của ngôn ngữ hình tượng.

Phần III: *Những vấn đề sâu hơn về PCHKL.* Trong phần này, tác giả cho biết, để chọn dùng một cách tiếp cận PCHKL đối với nghiên cứu văn chương nói chung, có một số nguyên tắc tiên quyết, trong đó:

"PCHKL gắn bó với sự quan tâm tối đa đối với nghiên cứu thống kê và định lượng về phong cách. Nhà nghiên cứu phải có những hiểu biết

[8] Leech G and Short M. (1981). *Style in Fiction*, London: Longman. Dẫn theo: Simino and Short (2004: 9).

[9] TESAS/Crouch vốn là những công cụ được xây dựng để nghiên cứu việc tái sử dụng văn bản trong lĩnh vực báo chí (Ho 2011: 201).

[10] Wcopyfind: là phần mềm phát hiện đạo văn bằng cách so sánh điểm tương tự giữa các văn bản, do Louis Bloomfield, Đại học Virginia, xây dựng vào năm 2002 (Ho, 2011: tr. 79).

[11] Ngôn ngữ hình tượng quá trừu tượng, phức tạp và là lĩnh vực khó giải quyết nhất đối với công việc định lượng (Ho, 2011: tr. 202).

cơ bản về công cụ thống kê và những hiểu biết sâu sắc về dữ liệu ngôn ngữ mà họ quan sát; và phải ghi khắc trong lòng rằng thống kê học cũng có thể sai lạc như bất cứ phương pháp truyền thống nào về phong cách. Như Ardat (1986)[12] đã chỉ ra rằng: thống kê phục vụ cho phong cách học với ba điều kiện sau: (a) lúc nó được dùng như một phương tiện, chứ không phải là một mục đích; (b) lúc khám phá của nó có hiệu lực thực tiễn như hiệu quả thẩm mỹ và nhấn mạnh tác dụng nghệ thuật; và (c) lúc dữ liệu quan sát của nó là thỏa đáng cả về phương diện định tính và định lượng[13].

Nguyên tắc cuối cùng và thách thức lớn nhất là phát triển và thao tác hóa trực giác phong cách học, đó là cách phát triển phương pháp đáng tin cậy về định lượng và làm nổi bật sự cân bằng giữa phân tích định tính và định lượng. Và đó là cách tiếp cận hợp nhất giữa khoa học thực nghiệm với nghệ thuật thông diễn học". (Ho, 2011: tr. 204-205)

Cách tiếp cận của Ho trong công trình này là cách tiếp cận dựa vào khối liệu để kiểm nghiệm những nhận định của các nhà phê bình về *The Magus*.

Historical corpus stylistics (Phong cách học khối liệu lịch sử)

Công trình "Historical corpus stylistics" của Patrick Studer (Đại học Zurich, Thụy Sĩ), Nhà xuất bản Continuum, 2008, 267 trang, đi theo hướng nghiên cứu khối liệu lịch đại (diachronic corpus). Công trình này nghiên cứu phong cách báo chí và các tác phẩm truyền thông đại chúng qua các giai đoạn lịch sử của diễn ngôn báo chí hiện đại, thời kỳ đầu. Mục đích của quyển sách là "đưa ra nhận thức thấu đáo về các nguyên tắc phong cách nền tảng của diễn ngôn báo chí bằng cách vạch ra những đặc điểm cơ bản của các thể loại truyền thông thời kỳ đầu và các nguyên lý biến đổi của phương tiện truyền thông đại chúng" (Studer, 2008: tr. 1). Khối liệu để phân tích là các xuất bản phẩm báo chí tiếng Anh thế kỷ XVIII và XIX, in ở Luân Đôn, chủ yếu lấy từ nguồn dữ liệu kỹ thuật số ZEN do khoa tiếng Anh, trường đại học Zurich (Thụy Sĩ) sưu tập.

Phong cách, trong công trình này, được xem xét theo quan điểm lịch đại của phong cách học lịch sử, kết hợp với phương pháp phân tích khối liệu dựa vào sự trợ giúp của máy tính. Trọng tâm nghiên cứu của quyển sách là "nhận diện và phân tích thực nghiệm các nhân tố phong cách đóng vai trò tích cực trong việc kích thích sự thay đổi của diễn ngôn báo chí trong giai đoạn đầu", mà khái niệm then chốt là "sự lạ hóa phong cách" (stylistic foregrounding). Công trình này nhấn mạnh việc xem xét mối tương quan giữa ngữ cảnh (bao gồm các nhân tố về văn hóa, thể chế, xã hội, kỹ thuật và bối cảnh thông tin) với sự đổi mới về phong cách (lệch hoặc vi phạm các quy ước và chuẩn phong cách), và phát triển các công cụ PCHKL tinh tế phục vụ cho việc phân tích mối liên hệ này. Công trình của Studer cũng bàn về các tiêu chuẩn sưu tập khối liệu lịch đại, phục vụ cho nhu cầu nghiên cứu phong cách học.

Công trình của Studer có thể xem thuộc dạng "nghiên cứu phong cách học truyền cảm hứng từ khối liệu", vì về mặt kỹ thuật, tác giả không dùng các phần mềm công cụ máy tính để xử lý thông tin như cách mà các nhà PCHKL thường làm, mà chỉ dùng thống kê thủ công.

5. Quan hệ với các phân ngành liên quan

5.1. Với ngôn ngữ học khối liệu

Ngôn ngữ học khối liệu và PCHKL là "anh em họ hàng". Hiển nhiên không có ngôn ngữ học khối liệu thì cũng không có PCHKL. Ngôn ngữ học khối liệu với kỹ thuật tinh tế về định lượng và xây dựng mẫu của nó là nền tảng khởi đầu của PCHKL, nhưng PCHKL vừa có những điểm chung vừa có những nét đặc thù.

Trong ngôn ngữ học khối liệu, các yếu tố ngôn ngữ thường được phân tích trong giới hạn của sự phân bố thông qua các loại hình văn bản hay ngữ vực khác nhau. Ngôn ngữ học khối liệu quan tâm chủ yếu đến vấn đề khái quát các hiện tượng ngôn ngữ, dựa trên thông tin định lượng. Ngôn ngữ học khối liệu nhấn mạnh sự tương tự giữa các văn bản trong ngữ vực. Văn bản văn chương trong ngôn ngữ học khối liệu được nhìn nhận với tư cách là một ngữ vực, mà không phải với tư cách những văn bản cá nhân. Ngôn ngữ học khối liệu quan tâm đến mối tương liên giữa hình thức và nghĩa, nhưng trọng tâm của nó vẫn là hình thức. Ngược lại, PCHKL quan tâm nghiên cứu nghĩa của văn bản cá nhân, những

[12] Ardat, 1986. "Stylostatistics: pros and cons in theory and application". Dẫn theo: Ho (2011: tr. 204).

[13] Tức là cần nắm vững "những nguyên tắc có tính chất phương pháp luận về sự kết hợp giữa quan sát ngôn ngữ với đánh giá văn chương" (Mahlberg, 2013: tr. 6).

khuôn mẫu liên quan đến một văn bản cụ thể, hoặc ngay cả những hiện tượng ngôn ngữ là độc nhất của một văn bản. PCHKL "không chỉ miêu tả các hiện tượng ngôn ngữ, mà còn giải thích chức năng của chúng trong việc tạo ra nghĩa của văn bản" (Mahlberg, 2016: tr. 145). PCHKL kết hợp các kỹ thuật phân tích khối liệu (của ngôn ngữ học khối liệu) với các phương pháp phân tích khác của phong cách học. PCHKL phối hợp chặt chẽ với lý thuyết, mô hình và phương pháp của phân tích phong cách học định tính (qualitative stylistic analysis) để gia cố, tăng cường hiệu lực của kỹ thuật máy tính. Sự tích hợp của kỹ thuật khối liệu với các khung lý luận, mô hình, lý thuyết phong cách học tiền khối liệu là điểm phân biệt PCHKL với ngôn ngữ học khối liệu nói chung.

5.2. Với phong cách học máy tính và trắc học phong cách

Phong cách học máy tính (computational stylistics) xuất hiện vào khoảng cuối thập kỷ 60 (thế kỷ XX), dùng sự phân tích và phương pháp trợ giúp của máy tính và thống kê để nghiên cứu những vấn đề khác nhau về phong cách. Phong cách học máy tính được xem là một phân ngành của ngôn ngữ học máy tính (Wales, 1989: tr. 85).

Tương tự ngôn ngữ học khối liệu, phong cách học máy tính dùng những phương pháp nghiêng về thống kê để định lượng và so sánh các hiện tượng ngôn ngữ. Thông qua so sánh, phong cách học máy tính mô tả và phân loại văn bản. Phương pháp của PCHKL liên quan với những kỹ thuật định lượng được ứng dụng trong ngôn ngữ học khối liệu và phong cách học máy tính. Tuy nhiên, PCHKL có sự khu biệt, vì nhấn mạnh việc dùng các mục tiêu văn chương và định tính để hướng dẫn phân tích và giải thích kết quả (Mahlberg, 2016: tr. 153).

Phong cách học máy tính là bước ngoặt tiến bộ của phong cách học với nghĩa là sự đối lập với phong cách học thủ công. PCHKL được truyền sức mạnh của các phương pháp máy tính, nhưng nó bước lên một cấp độ cao hơn là thao tác trên các khối liệu và hướng tới các định hướng về nghĩa và giải thích về phong cách.

Trắc học phong cách (stylometry hoặc stylometrics)[14] là một phân ngành của phong cách học. "Trắc học phong cách dùng phân tích thống kê để khám phá các khuôn mẫu phong cách nhằm xác định nguồn gốc tác giả của văn bản" (Wales, 1989: tr. 85). Trắc học phong cách quan tâm nhiều tới phong cách ở phương diện vốn từ vựng riêng của tác giả, và xem phong cách như một loại "dấu vân tay" (fingerprint) hoặc dấu vết DNA để nhận diện tác giả.

Các hiện tượng ngôn ngữ thường được khảo sát trong trắc học phong cách bao gồm độ dài của từ, độ dài của câu, từ ngữ liên kết, các ngữ kết hợp (các đồng hiện từ – collocation). Các hiện tượng này không cần thiết phải là sự lạ hóa, nhưng phải được tác giả dùng một cách vô thức, và vì thế, khá ổn định trong suốt sự nghiệp của tác giả đó (Wales, 1989: tr. 139).

Trắc học phong cách khá gần gũi với PCHKL ở phương diện phương pháp định lượng và các mục tiêu khảo sát. Nhưng trắc học phong cách có một vấn đề trung tâm, song không nằm trong phạm vi nghiên cứu của PCHKL, là xác định phong cách của một tác giả chưa biết trên những văn bản còn có những tranh cãi về nguồn gốc tác giả. Trắc học phong cách có thể thực hiện theo lối thủ công hay dựa vào kỹ thuật máy tính, khảo sát trên khối liệu điện tử, nhưng trọng tâm nghiên cứu của nó khác với PCHKL.

6. Thế mạnh và giới hạn

6.1. Thế mạnh của phong cách học khối liệu

Thế mạnh lớn nhất và dễ dàng thấy nhất của việc dùng phương pháp khối liệu trong nghiên cứu phong cách học là có được các kết quả thông tin định lượng rõ ràng, nhanh chóng, chính xác, và đó là những thông tin không dễ dàng (hoặc không thể nào) thu được bằng phân tích thủ công. Với những khối liệu lớn hoặc cực lớn (hàng chục triệu từ trở lên), việc phân tích thủ công gần như là bất lực (ít nhất cũng từ phương diện định lượng), và khi đó, PCHKL phát huy được thế mạnh của mình.

Một thế mạnh khác của PCHKL là dữ liệu định lượng có thể làm nổi bật các hiện tượng ngôn ngữ mà người đọc có thể không nhận thức được do dung lượng quá lớn của văn bản. Máy tính sẽ chỉ ra những dấu vết ngôn ngữ mà bộ óc con người bỏ sót đó.

PCHKL dựa trên những chứng cứ ngôn ngữ trung thực, khách quan do máy móc xử lý nên tránh được cách phân tích thủ công bị ảnh

[14] stylometry, stylometrics hoặc stylostatistics (thống kê học phong cách); chúng tôi tạm dịch là trắc học phong cách, với nghĩa là khoa học đo đạc để xác định phong cách.

hưởng nhiều bởi thiên kiến của người phân tích.

Một thế mạnh khác nữa của PCHKL là kết quả định lượng do máy tính thực hiện được liệt kê, lưu trữ trong những định dạng dễ dàng truy xuất, tính toán, hiển thị và kiểm chứng.

Phương pháp luận khối liệu trong phong cách học không đơn giản chỉ dùng để xác nhận (hay phủ nhận) giả thuyết ban đầu của người nghiên cứu; nó còn gợi ra giả thuyết. Xuất phát từ "điểm không" ban đầu, bằng cách dùng những phần mềm máy tính xử lý khối liệu, nhà phong cách học có thể phát hiện ra những ý tưởng nghiên cứu, từ đó tìm thêm những chứng cứ để chứng minh giả thuyết của mình trong một khối liệu lớn hơn hoặc thông qua phân tích định tính bằng cách vận dụng các khái niệm và kỹ thuật cao hơn của phong cách học.

6.2. Giới hạn của phong cách học khối liệu

Phương pháp khối liệu nói chung tập trung vào hình thức của hiện tượng ngôn ngữ được phân tích và cung cấp thông tin định lượng về sự xuất hiện và mối quan hệ của chúng trong văn bản, nên các phương diện nghĩa, chức năng của hiện tượng ngôn ngữ không được đề cập (hoặc chú trọng). Vì "tập trung vào hình thức và bề mặt văn bản, PCHKL bị cho rằng đã dẫn chúng ta trở lại cuộc tranh luận trong phong cách học về cái bị nghi ngờ là giá trị của thông tin định lượng[15] trong phân tích văn bản" (Mahlberg, 2016: tr. 140). Đó cũng là "giới hạn của việc ứng dụng tiêu chuẩn ngôn ngữ bề mặt để đo lường nội dung của văn bản", và việc dùng những tiêu chuẩn hình thức[16] như thế sẽ có nguy cơ là "đo lường những cái không thể đo lường được" (Ho, 2011: tr. 84). Cấp độ càng cao hơn và bình diện càng trừu tượng hơn của tổ chức ngôn ngữ, thì càng khó định lượng hơn. Vì vậy, đã có "nhiều khuynh hướng kháng cự tất cả các nghiên cứu văn chương mang tính thực nghiệm, toán học và kỹ thuật hóa, vì các thủ pháp máy tính và cách tiếp cận định lượng dường như phá hủy tính văn chương thực sự của văn bản và là hình ảnh thu nhỏ của các tiếp cận phi nhân văn đối với văn chương" (Ho, 2011: tr. 9).

Một thách thức cho sự phát triển của PCHKL là cho đến nay, nó chưa có các công cụ xác lập mặc định chuyên dùng cho chính nó, chứ không phải là vay mượn từ ngôn ngữ học khối liệu, trắc học phong cách hoặc một số chuyên ngành kỹ thuật. Mặt khác, nhà phong cách học phải biết thiết kế và dán nhãn cho một khối liệu chuyên dụng khi đi vào nghiên cứu các vấn đề chuyên biệt của phong cách học, chứ không thể đơn giản chỉ sử dụng lại các nguồn khối liệu có sẵn. Một vấn đề khác là hầu hết các học giả nghiên cứu văn chương không hiểu các nguyên tắc, kỹ thuật thống kê và họ không thực sự tin tưởng dùng nó để giải quyết các vấn đề văn chương. Do "cách đào tạo của phương pháp truyền thống về nghiên cứu khoa học nhân văn (nặng về phán đoán trực giác và phân tích thủ công), những người thực hành phong cách học có thể không có kỹ năng hoặc được trang bị những thứ cần thiết cho việc dùng hiệu quả máy vi tính trong nghiên cứu của họ" (Wynne, 2006: tr. 10453).

7. Kết luận

Trong nghiên cứu ngôn ngữ, đặc biệt, trong lĩnh vực phong cách học, máy móc không thể thay thế bộ óc con người, chúng chỉ là công cụ hỗ trợ. Kỹ thuật khối liệu không thể thay thế hoàn toàn cho phân tích phong cách học thủ công. Chạy một chỉ mục hoặc tạo ra một danh mục từ khóa, tự nó chưa phải là một sự phân tích. PCHKL tạo ra kỹ thuật tân tiến cho nghiên cứu phong cách, nhưng nó cũng không phải là giải pháp cho mọi vấn đề của phong cách học. Cũng như bất cứ phương pháp luận nghiên cứu nào, PCHKL cũng có sở trường và sở đoản của riêng nó. Điều quan trọng là phải biết ứng dụng nó đúng chỗ. Phân tích PCHKL và phong cách học truyền thống phải cộng tác cùng nhau và bổ sung cho nhau, vì nếu "không có nền tảng lý thuyết và những quan tâm giải thích thì việc ứng dụng phương pháp khối liệu cho một văn bản cụ thể có thể chỉ đưa đến kết quả là một quan sát ngây thơ" (Mahlberg, 2016: tr. 153).

[15] "Kết quả tìm thấy chỉ là những con số lạnh lẽo, xa lạ với lĩnh vực khoa học nhân văn" (Ardat,1986. "Stylostatistics: pros and cons in theory and application", dẫn theo: Ho, 2011: tr. 11).

[16] Vấn đề then chốt của việc định lượng bằng công cụ máy tính là các dấu hiệu hình thức (formal criteria) để máy tính nhận diện. Cho nên, một trong những vấn đề chìa khóa của PCH KL là tìm ra các dấu hiệu hình thức của các hiện tượng phong cách học trừu tượng và phức tạp để máy tính xử lý.

Tài liệu tham khảo

Ho, Y. (2011). *Corpus Stylistics in Principles and Practice.* New York: Continiuum.

Leech, G. and Short, M. (2007). *Style in fiction: a linguistic introduction to English fictional prose (Second edition).* London: Pearson/ Longman.

Louw, B. and Milojkovic, M. (2016). *Corpus Stylistics as Contextual Prosodic Theory and Subtext.* Amsterdam and Philadelphia: John Benjamins.

Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction.* London and New York: Routledge.

Mahlberg, M. (2014). Corpus stylistics. In: Burke Michael (ed.). *The Routledge Handbook of Stylistics.* London: Routledge, pp. 378-392.

Mahlberg, M. (2016). Corpus stylistics. In: Sotirova V (ed.) *The Bloomsbury Companion to Stylistics.* London and New York: Bloomsbury, pp. 139-156.

Mahlberg, M. and McIntyre, D. (2011). A case for corpus stylistics: Ian Fleming's Casino Royale. *English Text Construction*, 4 (2), pp. 204-227.

McIntyre, D. (2015). Towards an intergrated corpus stylistics. *Topics in Linguistics*, 16 (1), pp. 59-68.

Semino, E. and Short, M. (2004). *Corpus stylistics: speech, writing and thought presentation in a corpus of English writing.* London: Routledge.

Steward, L. (2006). Computational stylistics. In: Brown K (ed.) *Encyclopedia of language and linguistics.* New York: Elsevier, pp. 1757-1764.

Studer, P. (2008). *Historical Corpus Stylistics: Media, Technology and Change.* London and New York: Continuum.

Wales, K. (1989). *A Dictionary of Stylistics.* London and New York: Longman.

Wynne, M. (2006). Stylistics: Corpus approaches. In: Brown K (ed.) *Encyclopedia of language and linguistics.* New York: Elsevier, pp. 10451-10455.