MỘT SỐ VẤN ĐỀ CƠ BẢN CỦA PHÂN TÍCH DIỄN NGÔN VỚI SỰ TRỢ GIÚP CỦA KHỐI LIỆU

PHAN THANH BẢO TRÂN[1]
NGUYỄN THỊ DIỆP NHƯ[2]
BÙI THANH TƯỜNG THỤY[3]

Abstract: This paper explores discourse analysis from an interdisciplinary perspective, focusing on the intersection between discourse analysis and corpus linguistics. It seeks to clarify the respective roles of corpus linguistics and discourse analysis in interdisciplinary linguistic research. The integration of computational tools and techniques into linguistic studies represents a significant advancement yielding promising results; however, such approaches cannot replace the necessity of qualitative, critical, and context-sensitive analyses. Research on language in use must be grounded in the theoretical foundations of linguistics and discourse analysis. This paper offers an introductory overview of the topic rather than an in-depth analysis and aims to lay the groundwork for further, more comprehensive investigations.

Keywords: *discourse analysis, corpus discourse analysis, corpus linguistics*

1. Giới thiệu phân tích diễn ngôn với sự trợ giúp của khối liệu

Phân tích diễn ngôn với sự trợ giúp của khối liệu (CADA) là phân tích diễn ngôn dựa trên một khối lượng dữ liệu điện tử quy mô lớn, song khối liệu và các công cụ (khối liệu, các phần mềm máy tính) và kĩ thuật của ngôn ngữ học khối liệu (các cách thức xây dựng và xử lí khối liệu, thống kê) chỉ đóng vai trò trợ giúp cho phân tích diễn ngôn mà hoàn toàn không thay thế nó hay thậm chí có tư cách tương đương với nó. Từ khóa trong tên gọi của lĩnh vực nghiên cứu này là "trợ giúp" (assisted). Xử lí diễn ngôn bằng máy tính mang lại hiệu quả cho nghiên cứu.

Xử lí diễn ngôn bằng máy tính là một bước tiến của phân tích diễn ngôn. Không thể phủ nhận các thành tựu nghiên cứu của nó, song phân tích khối liệu bằng máy tính vẫn chỉ là công cụ hay phương tiện của các nhà phân tích diễn ngôn. Phân tích khối liệu bằng máy tính không cung cấp cho nhà phân tích diễn ngôn câu trả lời, chúng hỗ trợ nhà nghiên cứu đưa ra những phán đoán đúng đắn và xây dựng lí thuyết dựa trên chứng cứ chắc chắn hơn. Nói cách khác, phân tích khối liệu bằng máy tính tự nó không thể phân tích diễn ngôn, song nó có thể hỗ trợ việc thực hiện phân tích diễn ngôn ở một số cách thức có giá trị [7, tr.41].

Trước hết các thuật ngữ liên quan như *phân tích diễn ngôn dựa trên khối liệu, phân tích diễn ngôn với sự trợ giúp của khối liệu* hoặc *phân tích diễn ngôn khối liệu* được sử dụng gần như đồng nghĩa, chỉ

[1] Trường Đại học Khoa học Xã hội & Nhân văn, ĐHQG Thành phố Hồ Chí Minh; Email liên hệ: baotrannnh83@gmail.com
[2] Viện Kĩ thuật và Khoa học Máy tính, Trường Đại học VinUni
[3] Trung tâm Đổi mới Sáng tạo, Trường Đại học VinUni

có khác nhau chút ít. Đó là thuật ngữ sau dùng với ý thức rõ hơn về vai trò của khối liệu và các công cụ, kĩ thuật liên quan đến phân tích diễn ngôn, không có khác biệt đáng kể nào.

Trên thực tế, *phân tích diễn ngôn dựa trên khối liệu*, cũng như *phân tích diễn ngôn với sự trợ giúp của khối liệu* CADA, cũng có hai cách tiếp cận, và không có lí do gì để chỉ bó hẹp việc nghiên cứu của nó ở một trong hai cách này [2, tr.16].

Thuật ngữ *nghiên cứu diễn ngôn với sự trợ giúp của khối liệu* (corpus-assisted discourse studies - CADS) lại là một vấn đề khác. Mới thoạt nhìn nó có vẻ đồng nghĩa với thuật ngữ CADA nhưng thực ra phạm vi bao quát của nó rộng hơn CADA rất nhiều.

2. Đặc điểm của phân tích diễn ngôn với sự trợ giúp của khối liệu

2.1. Về cách tiếp cận

Giống như ngôn ngữ học khối liệu, CADA cũng có hai cách tiếp cận đối với khối liệu là: (i) cách tiếp cận dựa vào khối liệu (corpus-***based*** approach) và (ii) cách tiếp được chỉ dẫn bởi khối liệu (corpus-***driven*** approach) [2, tr.16].

Cách tiếp cận dựa vào khối liệu là cách tiếp cận xuất phát từ các mô hình lí thuyết diễn ngôn đã có (rút ra từ phân tích một số lượng nhỏ văn bản hay cuộc thoại) hoặc các giả định lí thuyết dựa trên trực giác và dùng phân tích khối liệu để kiểm nghiệm, xác thực, điều tra các mô hình hoặc giả định lí thuyết đó trên quy mô lớn về ngữ liệu và lượng hóa kết quả nghiên cứu thành các con số thống kê cụ thể. Cách tiếp cận này gần với cách phân tích truyền thống của các nhà phân tích diễn ngôn, cho phép đi sâu vào các vấn đề thực chất của phân tích diễn ngôn, nhưng mất nhiều thời gian và công sức vì thường phải xử lí thủ công hoặc bán tự động khối liệu. Cách tiếp cận này chính là cách tiếp cận từ trên xuống (top-down approach) trong Biber, Connor & Upton [3, §6].

Cách tiếp cận được chỉ dẫn bởi khối liệu là cách tiếp cận ngược lại với *cách tiếp cận dựa vào khối liệu*. Nó không xuất phát từ các chỉ dẫn lí thuyết diễn ngôn đã có mà có tham vọng dùng các kĩ thuật phân tích của ngôn ngữ học khối liệu để phân tích tự động các khối liệu nhằm rút ra lí thuyết mới cho phân tích diễn ngôn. Cách tiếp cận này gần với cách phân tích theo truyền thống thực nghiệm và quy nạp của các nhà ngôn ngữ học khối liệu, đúng với tinh thần của việc xử lí khối liệu bằng máy tính hơn, tiết kiệm thời gian hơn, song do hạn chế của việc thiếu các phần mềm chuyên dụng hoặc thích hợp cho phân tích diễn ngôn, nên kết quả nghiên cứu theo hướng này thường chỉ dừng lại ở phương diện khảo sát hình thức ngôn ngữ (nếu không dùng phương pháp nghiên cứu định tính của phân tích diễn ngôn hoặc các phương pháp nghiên cứu liên ngành).Cách tiếp cận thứ hai này chính là cách tiếp cận từ dưới lên (bottom-up approach) trong Biber, Connor & Upton [3, §6].

2.2. Sự kết hợp giữa phương pháp định lượng và định tính trong nghiên cứu

CADA thực chất là sự kết hợp giữa phân tích diễn ngôn với ngôn ngữ học khối liệu nên phương pháp nghiên cứu của nó là tích hợp phương pháp của hai chuyên ngành này.

CADA là sự tích hợp của hai phương pháp nghiên cứu vừa nói mà tinh thần cơ bản của nó là sự *phối hợp giữa phương pháp định lượng* và *phương pháp định tính*. Sự phối hợp hai phương pháp này

giúp hoàn thành tốt nhiệm vụ phân tích diễn ngôn mà tự bản thân mỗi chuyên ngành chưa thể giải quyết, đem lại sự bao quát về bề rộng (cái mà chỉ ngôn ngữ khối liệu làm được) và chiều sâu (cái mà chỉ phân tích diễn ngôn làm được) cho mỗi kết luận rút ra [1, tr.174].

Vậy định hướng nghiên cứu nền tảng của CADA là quy nạp hay diễn dịch? Theo chúng tôi, điều này còn phụ thuộc vào cách tiếp cận nào (x. §2) mà các nghiên cứu CADA theo đuổi. Nếu nghiên cứu CADA đi theo cách tiếp cận *dựa vào khối liệu* thì nghiên cứu đó về cơ bản đi theo hướng diễn dịch, còn đi theo hướng *được chỉ dẫn bởi khối liệu* thì nghiên cứu đó về cơ bản đi theo hướng quy nạp.

Sự phối hợp các phương pháp nghiên cứu là khuynh hướng chủ đạo hiện nay trong lĩnh vực khoa học, cái mà các nhà nghiên cứu khoa học xã hội và phân tích diễn ngôn gọi là "triangualtion" (phép trắc đạc tam giác - kết hợp nhiều phương pháp nghiên cứu, nhiều cách tiếp cận khác nhau để cùng giải quyết một vấn đề), phù hợp với chủ nghĩa hậu cấu trúc "ủng hộ một cách tiếp cận chiết trung hơn để nghiên cứu, trong đó các phương pháp luận khác nhau có thể kết hợp với nhau, đóng vai trò hỗ trợ lẫn nhau" [2, tr.16].

Ngoài phương pháp nghiên cứu cơ bản là phối hợp *phân tích định lượng* và *phân tích định tính*, CADA còn sử dụng một phương pháp nghiên cứu bổ sung là *phương pháp so sánh*.

Phương pháp so sánh vốn được sử dụng tương đối phổ biến trong phân tích diễn ngôn truyền thống khi so sánh hai văn bản hoặc cuộc thoại với nhau. CADA tiếp tục truyền thống đó của phân tích diễn ngôn, song mở rộng phạm vi so sánh lên ở cấp độ khối liệu (ví dụ so sánh khối liệu thành phần với khối liệu toàn thể, so sánh khối liệu chuyên ngành với khối liệu tổng quát, so sánh các khối liệu tương đương có liên quan với nhau). Khác với phân tích diễn ngôn truyền thống, các kiểu so sánh này trong CADA phần lớn mang tính bắt buộc.

2.3. Một số kĩ thuật chung trong nghiên cứu

Các kĩ thuật nghiên cứu cơ bản của CADA, tức các kĩ thuật liên quan đến việc sử dụng máy tính để xử lí ngữ liệu, vốn có nguồn gốc từ ngôn ngữ học khối liệu, "gồm hai loại: xây dựng khối liệu và phân tích khối liệu" [2, tr.76].

Ngoài các kĩ thuật nghiên cứu cơ bản này, trong từng phạm vi nghiên cứu, hướng nghiên cứu cụ thể của CADA, các nhà nghiên cứu còn sử dụng các kĩ thuật nghiên cứu chuyên sâu khác, tùy theo mục đích nghiên cứu và khả năng phát triển các công cụ kĩ thuật của mình.

2.3.1. Xây dựng khối liệu

Trên thực tế, nhà phân tích diễn ngôn không nhất thiết phải tự mình xây dựng khối liệu để tiến hành nghiên cứu mà chọn dùng trong các nguồn khối liệu có sẵn. Khi các nhà phân tích diễn ngôn chọn dùng hoặc xin phép dùng các nguồn khối liệu sẵn có sẽ gặp một số trở ngại sau đây:

(i) Phải trả tiền phí sử dụng, và thường khoản phí này không phải ít;

(ii) Cách thức xây dựng khối liệu không phù hợp với mục đích nghiên cứu

(iii) Khối liệu không đủ số lượng hoặc đủ loại hình văn bản cần thiết;

(iv) Khối liệu quá cũ.

Vì thế cách thông thường nhất là nhà phân tích diễn ngôn tự mình thiết kế khối liệu để nghiên cứu.

Những điều cần chú ý nhất khi thiết kế khối liệu là loại khối liệu, kích thước khối liệu, tính đại diện, sự cho phép (của tác giả văn bản được dùng trong khối liệu) và chú thích các thành phần trong khối liệu. Trong bài này, chúng tôi chỉ đề cập hai vấn đề đầu tiên trong số đó, chúng liên quan nhiều đến CADA.

Về kích thước khối liệu, khối liệu chuyên biệt mà nhà nghiên cứu CADA cần xây dựng không cần phải là khối liệu hàng triệu từ: thường chúng có dung lượng tương đối nhỏ, khoảng 15.000 đến dưới 200.000 từ, bao gồm các văn bản hoàn chỉnh, chứ không phải các mẫu văn bản đại diện (như khối liệu trong ngôn ngữ học khối liệu). Vấn đề ở đây không phải là kích thước khối liệu bao nhiêu mà là chúng thể hiện đúng vấn đề cần nghiên cứu, "nghĩa là chất lượng hoặc nội dung của dữ liệu phải bằng hoặc ưu tiên hơn các vấn đề về số lượng" [2, tr.29].

Có một số loại khối liệu về cơ bản không liên quan đến phân tích diễn ngôn, ví dụ khối liệu song song (parallel/aligned corpus), khối liệu dành cho người học (learner corpora), khối liệu phương ngữ (dialect corpora) [2, tr.45].

2.3.2. Phân tích khối liệu

2.3.2.1. Phân tích danh sách từ (wordlist), tần số (frequency)

Danh sách từ (wordlist) chỉ đơn giản là danh sách tất cả các từ trong khối liệu (cùng với tần số của chúng và tỉ lệ phần trăm đóng góp mà mỗi từ tạo ra cho khối liệu). Danh sách từ được cung cấp bởi các phần mềm phân tích khối liệu như ***WordSmith Tools***.

Tần số (frequency) là một trong những khái niệm trung tâm nhất làm cơ sở cho việc phân tích khối liệu (...) một trong những công cụ cơ bản nhất của nhà ngôn ngữ học khối liệu, chúng là điểm khởi đầu tốt để phân tích bất kì loại khối liệu nào [2, tr.47].

2.3.2.2. Phân tích chỉ mục (concordance)

Chỉ mục (concordance) là một danh sách tất cả các lần xuất hiện của một từ được tìm kiếm cụ thể trong một khối liệu, được trình bày trong ngữ cảnh mà chúng xuất hiện; thường là một vài từ ở bên trái và bên phải của từ cần tìm kiếm. Hiện nay, theo Baker [2, tr.75], có một số phần mềm cho phép hiển thị thông tin chỉ mục như sau:

Concordancer	**Platform**	**Cost**	**Website**
Cone	Mac	Freeware	www.sil.org/computing/conc/conc.html
MicroConcord	PC	Freeware	www.liv.oc.uk/˜ms2928/software/
WordSmith Tools	PC		www.lexically.net/wordsmith/index.html
Concordance	PC		www.concordancesoftware.co.uk/
Mono Cone	PC		www.athel.com/mono.html
Simple Concordance	PC	Freeware	www.textworld.com/scp

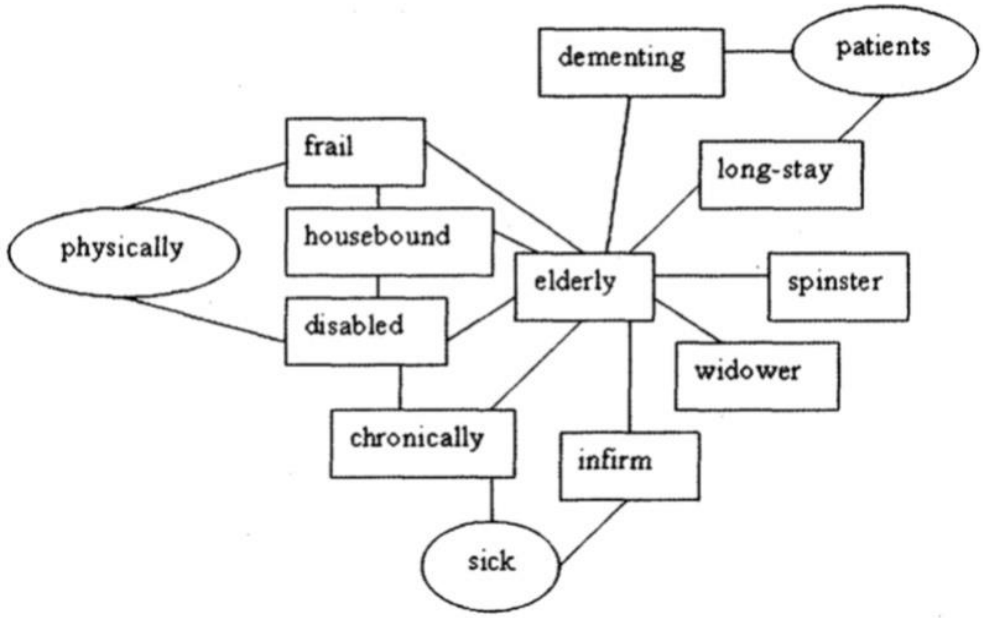
Các phần mềm này cho phép các nhà nghiên cứu sắp xếp và do đó xem xét dữ liệu theo nhiều cách khác nhau, song nhà phân tích vẫn có trách nhiệm nhận ra các mô hình ngôn ngữ và giải thích tại sao chúng tồn tại [2, tr.89]. Nghĩa là hiệu quả của phân tích chỉ mục phụ thuộc khá nhiều vào nhà phân tích, vì theo Baker thì các mô hình ngôn ngữ được tìm thấy (hoặc bị bỏ qua) có thể phụ thuộc vào lập

trường tư tưởng của chính nhà nghiên cứu [2, tr.89]. Và cách chúng được diễn giải cũng có thể được lọc qua quan điểm chủ quan của nhà nghiên cứu.

2.3.2.3. Phân tích kết ngữ/đồng hiện (collocate)

Khi một từ thường xuyên xuất hiện gần một từ khác và mối quan hệ ý nghĩa được thống kê theo một cách nào đó, chúng được gọi là kết ngữ/đồng hiện (collocates) và hiện tượng của một số từ nhất định thường xuyên xảy ra bên cạnh hoặc gần nhau là hiện tượng đồng hiện (collocation)" [2, tr.95-96].

Các kết ngữ rất hữu ích ở chỗ chúng giúp tóm tắt các mối quan hệ quan trọng nhất giữa các từ trong một khối liệu [2, tr.118]. Phân tích các kết ngữ có ý nghĩa quan trọng vì "Bạn sẽ biết rất nhiều về một từ từ đơn vị đồng hành đi cùng với nó". Phân tích kết ngữ cung cấp cho chúng ta những mô hình từ vựng nổi bật và rõ ràng nhất, xung quanh một chủ đề, từ đó có thể rút ra các diễn ngôn.



Hình 1. Mạng lưới kết ngữ của elderly (người cao tuổi)[4] Nguồn: [2, tr.117].

Ví dụ mạng lưới các kết ngữ của elderly (người cao tuổi) trong the British National Corpus (BNC) được Baker [2] biểu diễn thành sơ đồ ở Hình 1. Trong mạng lưới này, các kết ngữ *elderly- frail* (yếu ớt), *elderly-housebound* (không thể ra khỏi nhà), *elderly-disabled* (không thể cử động chân tay) gợi ra các diễn ngôn về thân thể, vẻ bề ngoài; các kết ngữ *eldely-chronically* (kinh niên), *elderly-infirm* (ốm yếu) gợi ra các diễn ngôn về bệnh tật *(sick)*; các kết ngữ *eldely-dementing* (sa sút trí tuệ), *elderly-longstay* (lưu trú dài hạn) gợi ra các diễn ngôn về người bệnh; các kết ngữ *elderly-widower* (người goá vợ), *elderlyy-spinter* (người không chồng) gợi ra các diễn ngôn về tình trạng hôn nhân hiện tại.

[4] Trong hình 1, hình tứ giác biểu thị các kết ngữ bậc 1, hình ô van biểu thị các kết ngữ bậc 2. Ví dụ: *elderly* đồng hiện với *chronically* và *infirm*, nhưng hai từ *chronically* và *infirm* không đồng hiện với nhau; trong khi *sick* đồng hiện với *chronically* và *infirm*, nên *sick* là kết ngữ bậc hai của *elderly*.

2.3.2.4. Phân tích từ khóa (keyword) và miền ngữ nghĩa (semantic domain)

Từ khóa (keyword) hay chính xác hơn là danh sách từ khóa (keyword list) là tất cả các từ xuất hiện thường xuyên hơn dự kiến trong một khối liệu khi so sánh danh sách tần số của nó với danh sách tần số của khối liệu khác. Từ khóa được cung cấp bởi các phần mềm như WordSmith Tools. Danh sách từ khóa là một công cụ giúp các nhà nghiên cứu tìm thấy sự khác biệt đáng kể về mặt từ vựng giữa các khối liệu, các loại diễn ngôn. Các từ khóa "không những chỉ ra sự tồn tại của các diễn ngôn, mà còn giúp tiết lộ các kĩ thuật tu từ được sử dụng để trình bày các diễn ngôn" [2, tr.148].

Miền ngữ nghĩa (semantic domain/ semantic field) là một vùng nghĩa chia sẻ một đặc trưng ngữ nghĩa chung và những từ được sử dụng để nói về nó. Miền ngữ nghĩa thường được đặt một tên theo một từ điển hình của miền. Các từ khóa đã truy xuất có thể được nhóm theo cách thủ công thành các miền ngữ nghĩa để xác định các đề tài và chủ đề chi phối trong tập hợp dữ liệu hay trong các diễn ngôn.

2.4. Hướng nghiên cứu phổ biến

Các tác giả Biber, Connor & Upton [3, tr.1], dựa theo Schiffrin, Tannen, and Hamilton [9, tr.1], nhóm các nghiên cứu của phân tích diễn ngôn thành ba loại (hay hướng nghiên cứu) chủ yếu sau đây: (1) nghiên cứu việc sử dụng ngôn từ (the study of language use); (2) nghiên cứu cấu trúc ngôn ngữ 'bên ngoài câu' (the study of linguistic structure 'beyond the sentence'); và (3) nghiên cứu thực tiễn xã hội và các giả thiết về ý thực hệ liên quan đến ngôn ngữ và/hoặc giao tiếp (the study of social practices and ideological assumptions that are associated with language and/or communication).Sự phân chia này áp dụng cho cả phân tích diễn ngôn theo quan điểm định tính và phân tích diễn ngôn dựa vào khối liệu (hoặc với sự trợ giúp của khối liệu).

2.4.1. Phân tích diễn ngôn với sự trợ giúp của khối liệu về cách sử dụng ngôn từ

Theo Biber, Connor & Upton, hầu hết các nghiên cứu dựa trên khối liệu là phân tích diễn ngôn theo nghĩa này, điều tra một cách hệ thống các mô hình sử dụng ngôn ngữ trong các bối cảnh diễn ngôn, được khái quát hóa trên tất cả các văn bản trong một khối liệu [3, tr.3]. Hoặc đó là nghiên cứu điều tra các khía cạnh về từ vựng, như các nghiên cứu về collocation (sự xuất hiện đồng thời của các từ cụ thể, ví dụ, Gledhill 2000, Ward 2007) hoặc các nghiên cứu xác định các mục từ vựng khác biệt về mặt thống kê trong ngữ vực (phân tích từ khóa, ví dụ Freddi 2005; Scott và Tribble 2006), hoặc các nghiên cứu về các tổ hợp từ (lexical bundle), cái thường xuất hiện thành chuỗi nhiều từ (xem Biber và cộng sự 1999, chương 13; Cortes 2004; Kim 2009; và Nesi và Basturkmen 2006) [6, tr.142].

Ngoài ra, các nhà nghiên cứu đang ngày càng dựa vào các phương pháp của ngôn ngữ học khối liệu để phát hiện ra các mô hình phân bố của các nhân tố ngữ pháp trong diễn ngôn, thể hiện qua nhiều báo cáo về các nghiên cứu khối liệu về ngữ pháp và diễn ngôn, ví dụ Collins (1991) về tách câu (cleft), Hunston và Francis (2000) về ngữ pháp mô hình (pattern grammar), và Römer (2005) về thể diễn tiến (progressives). Những nghiên cứu khác của CADA về cách sử dụng ngôn từ tập trung vào nghiên cứu các đặc điểm ngôn ngữ của diễn ngôn trong ngữ vực. Hoặc đó là các loại nghiên cứu tập trung vào việc mô tả các chức năng và sự phân bố của các đặc điểm ngôn ngữ riêng lẻ, hoặc các tập hợp nhỏ các đặc

điểm liên quan trong và xuyên qua các ngữ vực, hoặc mô tả các ngữ vực theo kiểu đồng xuất hiện của nhiều đặc điểm ngôn ngữ [6, tr.142].

2.4.2. Phân tích diễn ngôn với sự trợ giúp của khối liệu về cấu trúc ngôn ngữ 'bên ngoài câu'

Cấu trúc ngôn ngữ 'bên ngoài câu" là cách gọi khác đi của cấu trúc văn bản hay cấu trúc diễn ngôn. Phân tích cấu trúc ngôn ngữ 'bên ngoài câu' là phân tích cấu trúc văn bản hay phân tích cấu trúc diễn ngôn.

Theo Gray & Biber [6], nghiên cứu trong lĩnh vực CADA về cấu trúc ngôn ngữ 'bên ngoài câu' còn rất hạn chế. Công trình quan trọng nhất trong lĩnh vực này là cuốn sách của Biber, Connor & Upton [3] kết hợp phân tích đơn vị diễn ngôn chức năng với phương pháp của ngôn ngữ học khối liệu, phát triển một khung lí thuyết tổng quát để áp dụng phân tích dựa trên khối liệu cho phân tích cấu trúc văn bản. Đối với phương pháp tiếp cận từ dưới lên, trước tiên, văn bản được phân thành các đơn vị diễn ngôn dựa trên các tiêu chí ngôn ngữ như chuyển đổi cách sử dụng từ vựng. Sau đó, mỗi đơn vị diễn ngôn được phân tích về các đặc điểm ngôn ngữ của nó và được phân loại thành các loại xác định về mặt ngôn ngữ (thay vì chức năng), và các mô hình qua các văn bản của khối liệu được mô tả [3, tr.151].

Hai cách tiếp cận, từ trên xuống và từ dưới lên khác nhau về thứ tự các bước phân tích. Theo cách tiếp cận từ trên xuống, khung phân tích được phát triển ngay từ đầu: các loại đơn vị diễn ngôn được xác định trước khi bắt đầu phân tích khối liệu, và toàn bộ phân tích sau đó được thực hiện theo các giới hạn đó. Trong cách tiếp cận từ dưới lên, phân tích khối liệu được ưu tiên trước tiên và các loại đơn vị diễn ngôn hiện ra từ các mẫu khối liệu [3, tr.12]. Trái ngược với truyền thống nghiên cứu lâu đời áp dụng các phân tích diễn ngôn từ trên xuống, phương pháp từ dưới lên chỉ mới được phát triển gần đây, đặc biệt cho các phân tích dựa trên khối liệu về cấu trúc diễn ngôn. Cách tiếp cận này trước đây chưa được thực hành bởi các nhà phân tích diễn ngôn vì nó đòi hỏi các kĩ thuật tính toán tiên tiến. Cách tiếp cận từ dưới lên dùng các kĩ thuật tính toán tự động nên có thể dễ dàng áp dụng cho việc phân tích hàng trăm văn bản, không bị các hạn chế về vấn đề nhân lực để mã hóa thủ công các đơn vị diễn ngôn trong văn bản như cách tiếp cận từ trên xuống [3, tr.16-17].

Một điểm khác biệt chính giữa hai cách tiếp cận, theo Biber, Connor & Upton [3], là vai trò của các phân tích chức năng so với phân tích ngôn ngữ. Trong cách tiếp cận từ trên xuống, khung chức năng là chính.

2.4.3. Phân tích diễn ngôn với sự trợ giúp của khối liệu về thực tiễn xã hội và các giả thiết ý thức hệ

Hướng (/loại) phân tích diễn ngôn thứ ba này, theo Biber, Connor & Upton [3], có định hướng văn hóa xã hội, và nói chung không quan tâm đến việc mô tả các văn bản cụ thể hoặc phân tích cấu trúc và sử dụng ngôn ngữ. Các cách tiếp cận văn hóa xã hội đối với diễn ngôn đôi khi tập trung vào hành động của những người tham gia trong các sự kiện giao tiếp cụ thể, và có khi tập trung vào các đặc điểm chung của cộng đồng ngôn ngữ/diễn ngôn (speech/discourse community) liên quan đến các vấn đề như quyền lực và giới. Mặc dù các cách tiếp cận văn hóa - xã hội rõ ràng là quan trọng để hiểu

được vai trò rộng lớn hơn của văn bản trong văn hóa, nhưng chúng thường không quan tâm đến việc hiểu các hình thức ngôn ngữ được sử dụng trong các văn bản đó [3, tr.2].

2.5. Ưu nhược điểm của phân tích diễn ngôn với sự trợ giúp của khối liệu

2.5.1. Ưu điểm

Phân tích diễn ngôn với sự trợ giúp của khối liệu mang lại nhiều lợi thế so với cách tiếp cận truyền thống. Trước hết, khả năng mở rộng phạm vi khối liệu là ưu điểm nổi bật, nó giúp nhà nghiên cứu đi từ một khối dữ liệu nhỏ lẻ sang một tập hợp lớn các dữ liệu ngôn ngữ vừa đa dạng về loại, vừa đa dạng về bối cảnh. Điều này, giúp cho nhà nghiên cứu có thể so sánh, kiểm chứng các giả thuyết một cách hệ thống. Định hướng này không chỉ giúp cho việc phát hiện, khái quát các mô hình ngôn ngữ giữa các ngữ vực trở nên có độ tin cậy hơn mà còn cho thấy sự tiện lợi của việc sử dụng khối liệu, phân tích khối liệu lớn nhanh chóng mà vẫn có tính chính xác. Những con số, những định lượng có thể giúp người nghiên cứu định hướng lại quan điểm ban đầu và giảm thiểu tối đa những thiên kiến cá nhân, chủ quan trong nghiên cứu. Tuy nhiên, không thể nhấn mạnh định hướng này có tính khách quan tuyệt đối vì mọi phân tích, kể cả việc sử dụng khối liệu đều đã chịu sự ảnh hưởng bởi quan điểm, cách nhìn nhận vấn đề của người nghiên cứu khác nhau. Vì vậy, dù định hướng nghiên cứu này có sức thuyết phục hơn nhưng để thực sự hiệu quả thì nhà nghiên cứu cần tự phản biện, tự chứng minh và kiểm tra lại tất cả các kết quả trong toàn bộ quá trình nghiên cứu.

Rodney H. Jones [7] đã đánh giá về ưu điểm của *phân tích diễn ngôn với sự hỗ trợ của khối liệu* như sau: *Phân tích diễn ngôn với sự trợ giúp của khối liệu* độc đáo ở chỗ nó cho phép chúng ta vượt ra khỏi việc xem xét một số lượng nhỏ các văn bản hoặc các tương tác để phân tích một số lượng lớn chúng và có thể so sánh chúng với các văn bản và cuộc thoại khác được tạo ra trong những bối cảnh giống nhau hoặc khác nhau. Nó cũng cho phép chúng ta đưa ra phân tích của mình ở một mức độ 'khách quan' nào đó bằng cách cho chúng ta cơ hội kiểm tra các lí thuyết đã hình thành trong phân tích sâu của chúng ta về một số văn bản hoặc cuộc trò chuyện trên một lượng lớn dữ liệu một cách có hệ thống hơn [7, tr.40]. Việc dựa vào một lượng lớn dữ liệu và các kĩ thuật có sự hỗ trợ của máy tính không mang lại tính khách quan, nhưng nó có thể mang lại độ chính xác và trách nhiệm giải trình cao hơn [8, tr.7].

Vấn đề giảm thiểu thiên kiến của nhà nghiên cứu khi *phân tích diễn ngôn với sự trợ giúp của khối liệu* cần được hiểu đúng bản chất và mức độ. Burr (1995) lập luận rằng tính khách quan là không thể vì tất cả chúng ta nhìn nhận thế giới từ một số quan điểm nào đó (lập trường ‘khách quan’ vẫn là một lập trường). Thay vào đó, các nhà nghiên cứu cần phải thừa nhận sự tham gia của chính họ trong nghiên cứu và vai trò đối với kết quả được tạo ra [2, tr.10].

2.5.2. Nhược điểm

Phân tích diễn ngôn với sự trợ giúp của khối liệu còn tồn tại nhiều hạn chế do khối liệu chưa đủ lớn và bị tách khỏi ngữ cảnh. Xử lí ngữ liệu bằng máy tính thiên về *tính đại diện, thông tin tần suất, các mẫu lặp lại*, trong khi ngôn ngữ của các cá nhân trong giao tiếp lại đa dạng, không phải luôn có tính đại diện, và những thông điệp ngầm định hay điều không được nói ra gắn với ngữ cảnh là vô tận.

Điều này cho thấy rõ sự khác biệt về bản thể luận và nhận thức luận của ngôn ngữ học khối liệu và phân tích diễn ngôn, một bên nhấn mạnh vào tính đại diện, thống kê, một bên chú trọng tính toàn diện và chiều sâu ngữ nghĩa "các mô hình ngôn ngữ thường gặp (ngay cả khi được sử dụng bởi những người có uy tín) không phải lúc nào cũng bao hàm những cách suy nghĩ chính thống. *Đôi khi những gì không được nói hoặc viết còn quan trọng hơn những gì đang có* [2, tr.19].

Một hạn chế khác của cách tiếp cận này là cái mà Cape [4, tr.105] gọi là "điểm mù đa ký hiệu" hay đa phương thức (multisemiotic blind spots): "Khối liệu thuần tuý ngôn ngữ sẽ cho chúng ta biết một số câu chuyện, nhưng chúng sẽ không cho chúng ta biết toàn bộ câu chuyện..." [4, tr.85]. Đến nay, *phân tích diễn ngôn dựa vào khối liệu* chủ yếu tập trung vào ngôn ngữ dưới dạng ngôn từ, mà bỏ qua các nguồn lực ký hiệu khác như *hình ảnh, bố cục, màu sắc hay kiểu chữ, ...* những yếu tố góp phần quan trọng trong việc tạo nghĩa [7, tr.41].

2.6. Hướng phát triển của phân tích diễn ngôn với sự giúp của khối liệu

Theo Machi & Taylo, nghiên cứu khối liệu và diễn ngôn đã đến thời kỳ trưởng thành, đã đến lúc dừng lại và suy ngẫm những gì đã làm với cái nhìn phê phán, chỉ ra những điểm mù, góc khuất của nó và những gì phải thay đổi trong tương lai [8, tr.2].

Vấn đề đầu tiên cần làm trong số những thay đổi trong tương lai là việc xây dựng các khối liệu đa phương thức (multimodal corpora) và các phần mềm xử lí chúng phục vụ cho việc nghiên cứu các văn bản hiện đại (thường được xây dựng trên nền tảng hình ảnh). Thực ra hiện nay đã có các khối liệu đa phương thức và các phần mềm xử lí chúng và đã có sự manh nha của phân ngành nghiên cứu *phân tích diễn ngôn đa phương thức với sự trợ giúp của khối liệu* (corpus-assisted multimodal discourse analysis or CAMDA), một hướng tiếp cận do Bednarek & Cape (2014) phát triển, song quy mô hiện nay của nó còn tương đối nhỏ [4, tr.88]. Lí do là vì:

(…) việc phân tích quy mô lớn các văn bản đa phương thức là rất phức tạp, nó đòi hỏi sự phân tích với nhiều cách tiếp cận (triangulation) ở nhiều cấp độ và tập hợp nhiều chuyên gia về phân tích ngôn ngữ và đa phương thức/đa kí hiệu; các phần mềm phân tích khối liệu ngôn ngữ và đa phương thức và sự tích hợp của chúng với nhau; và sự phát triển của các công cụ để trình bày hoặc trực quan kết quả của những phân tích đó theo một cách có ý nghĩa đối với người đọc. [4, tr.89]

Vấn đề thứ hai cần làm là phát triển các phần mềm phân tích định tính dữ liệu (qualitative data analysis software), chứ không phải các phần mềm phân tích định lượng như phần lớn hiện nay, như các phần mềm Nvivo and Atlas.ti, những phần mềm cho phép máy tính phân tích sâu văn bản [1, tr.178-179]…

Vấn đề thứ ba cần làm đối với *phân tích diễn ngôn với sự trợ giúp của khối liệu* là chú ý tới các bối cảnh rộng hơn xung quanh văn bản [3, tr.259].

3. Kết luận

Ngôn ngữ học khối liệu và phân tích diễn ngôn thường bổ trợ cho nhau trong các nghiên cứu ngôn ngữ hiện nay. Có thể thấy, những khoảng cách về phương pháp, hướng tiếp cận giữa hai lĩnh vực này

ngày càng được thu hẹp. Chúng tôi đồng ý rằng phân tích diễn ngôn với sự trợ giúp của khối liệu có khả năng làm giảm những thiên kiến về vai trò, hiệu quả nghiên cứu của từng lĩnh vực. Vì, trong khi ngôn ngữ học khối liệu cung cấp bộ công cụ, kĩ thuật xử lí, giúp phát hiện các mô hình ngôn ngữ, đặc biệt hữu ích cho các nhà nghiên cứu không chuyên về công nghệ thông tin, thì phân tích diễn ngôn lại tập trung vào cung cấp cơ sở lí thuyết ngôn ngữ mà từ đó giúp thấy rõ rằng phải dựa vào tư duy của con người, dựa vào sự hành chức của ngôn ngữ trong hoàn cảnh giao tiếp cụ thể, mới khám phá được những khía cạnh mà máy móc không làm được. Tuy nhiên, vấn đề này, trong bài viết, chúng tôi cũng chỉ mới nhìn từ góc độ giao thoa, chưa đi sâu vào phân tích từ góc độ liên ngành, hoặc chuyên biệt để kiểm nghiệm tính hiệu quả. Chắc chắn rằng, chúng tôi sẽ tiếp tục tìm hiểu để mở rộng vấn đề và đặt ra các câu hỏi nghiên cứu tiếp theo.

TÀI LIỆU THAM KHẢO

1. Ancarno, C. *Corpus-Assisted Discourse Studies*. In A. De Fina & A. Georgakopoulou (eds.). *The Cambridge Handbook of Discourse Studies*. Cambridge University Press, 165-185. 2020.

2. Baker, P. *Using corpora in discourse analysis*. Continuum. 2006.

3. Biber, D., Connor, U. and Upton, T. *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins. 2007.

4. Caple, H. *Analysing the multimodal text*, In C. Taylor and A. Marchi (eds.). *Corpus approaches to discourse: A critical review*. Routledge, 85-109. 2018.

5. Flowerdew, L. *Corpus-based discourse analysis*. In J. P. Gee and M. Handford (eds.). *The Routledge Handbook of Discourse Analysis*. Routledge. 174-188. 2012.

6. Gray, B. and Biber, D. *Corpus approaches to the study of discourse*. In K. Hyland and B. Paltridge (eds.). *Bloomsbury Companion to Discourse Analysis*. Bloombury, 138-152. 2013.

7. Jones, R. H. *Corpus-assisted discourse analysis*. In Jones, R. H. *Discourse Analysis: A Resource Book for Students*. Routledge. 2012.

8. Marchi, A. and Taylor, C. *Introduction: Partiality and reflexivity*. In C. Taylor and A. Marchi (eds.). *Corpus approaches to discourse: A critical review*. Routledge, 1-15. 2018.

9. Schiffrin, D., Tannen, D., & Hamilton, H. (eds.). *The Handbook of Discourse Analysis*. Blackwell Publishers. 2001.

10. Charlotte Taylor, C. *Similarity*. In C. Taylor and A. Marchi (eds.). *Corpus approaches to discourse: A critical review*. Routledge, 19-37. 2018.

11. Taylor, C. and Marchi, A. (eds.). *Corpus approaches to discourse: A critical review*. Routledge, 2018.