PHÂN TÍCH DIỄN NGÔN TIẾNG ANH: HƯỚNG TIẾP CẬN KHỐI LIỆU

ENGLISH DISCOURSE ANALYSIS: A CORPUS-BASED APPROACH

TRẦN HỮU PHÚC
(TS; Đại học Ngoại ngữ, Đại học Đà Nẵng)

Abstract: Corpus linguistics, with the method of discourse analysis based on data collected from actual use of language, has become a popular approach in several research on English discourse. This paper presents a general view to corpus linguistics, methods of collecting and building a research corpus; introduces tools for corpus-based analysis such as key words, frequency list and concordance lines; illustrates ways of using software packages for researching a specific issue of English discourse in accordance with the method of corpus-based analysis.

Key words: corpus; discourse; corpus-based analysis; text linguistics; concordance.

1. Đặt vấn đề

1.1. Nghiên cứu ngôn ngữ đã và đang phát triển vượt xa những vấn đề cơ bản của lí thuyết ngôn ngữ học truyền thống. Ngoài các phạm trù cơ bản về ngữ âm, hình thái, cú pháp, ngữ nghĩa,...nghiên cứu ngôn ngữ hiện nay hướng đến các cấu trúc thông tin của diễn ngôn ở đơn vị trên câu, quan hệ giữa các câu, các phương tiện liên kết ngữ pháp, từ vựng và ngữ nghĩa trong một đoạn văn, một văn bản hay liên văn bản. Chính những đòi hỏi phức tạp này đã hình thành nên xu hướng nghiên cứu ngôn ngữ học văn bản (text linguistics) hay phân tích diễn ngôn (discourse analysis). Bài viết giới thiệu hướng tiếp cận ngôn ngữ học văn bản dựa trên khối liệu nghiên cứu và các phần mềm tiện ích để phân tích diễn ngôn. Phương pháp phân tích khối liệu (corpus-based analysis) hiện đang được nhiều nhà nghiên cứu ngôn ngữ sử dụng nhằm tìm hiểu các đặc trưng cụ thể của diễn ngôn dựa trên các khối liệu chuyên biệt được thu thập từ thực tế ứng dụng của ngôn ngữ.

1.2. Khối liệu ngôn ngữ học là một tập hợp các văn bản viết hoặc nói của cùng một loại thể được sử dụng trong ngôn ngữ giao tiếp tự nhiên, được cấu trúc một cách có hệ thống và được thiết kế để phục vụ mục đích nghiên cứu các phương diện cụ thể về cấu trúc và ứng dụng của ngôn từ.

Meyer (2002) định nghĩa "khối liệu là tập hợp các văn bản hay bộ phận của một loại hình văn bản mà dựa vào đó việc phân tích diễn ngôn được thực hiện". Khối liệu là một tập hợp văn bản đọc được bằng máy tính, có dung lượng rất lớn, dễ dàng được truy xuất nhờ sự hỗ trợ của các phần mềm chuyên dụng (xem Kennedy, 1998; Hunston, 2002; Baker, 2006).

Khối liệu chứa đựng nhiều thuộc tính tiềm ẩn, các ứng dụng ngôn ngữ tự nhiên; cung cấp các thông tin về mô hình phân bố và các thực thể từ vựng, cấu trúc của ngôn ngữ trong một loại hình văn bản. Khối liệu bao gồm nhiều loại khác nhau, phục vụ những mục đích cụ thể trong nghiên cứu ngôn ngữ như: khối liệu chuyên biệt, khối

liệu tổng hợp, khối liệu so sánh, khối liệu song song, khối liệu dành cho người học, khối liệu đồng đại, lịch đại,...Các nhà nghiên cứu dựa vào một loại hình khối liệu cụ thể để tìm kiếm các minh chứng phục vụ mục tiêu mô tả, phân tích ngôn ngữ hay đưa ra những luận giải về các vấn đề được nghiên cứu.

2. Ngôn ngữ học khối liệu (Corpus linguistics)

2.1. Những vấn đề chung

Ngôn ngữ học khối liệu nghiên cứu ngôn ngữ trên cơ sở ngữ liệu xác thực được xây dựng thành khối liệu văn bản. Theo Aijmer and Altenberg (1991) phương pháp nghiên cứu này xuất phát từ 2 sự kiện lớn: một là công bố của Randolph Quyrk (1959) về khảo sát ứng dụng của tiếng Anh (SEU) với mục đích thu thập một khối liệu lớn và đa dạng về phong cách diễn đạt trong tiếng Anh, mô tả một cách có hệ thống văn phong nói và viết của ngôn ngữ này; hai là sự ra đời của các phần mềm máy tính có thể lưu trữ và truy xuất khối lượng lớn dữ liệu văn bản. Sau Quyrk, hàng loạt các tập khối liệu tiếng Anh đã ra đời, phục vụ các mục đích nghiên cứu ngôn ngữ. Tiêu biểu là các công trình như: *Brown Corpus* (Francis & Kucera 1961),*Lancaster-Oslo/Bergen (LOB) Corpus* 1970-1978, *London Lund Corpus (LLC)* 1975. Đến những năm 1980 hàng loạt các tập khối liệu được biên soạn, phục vụ các mục đích nghiên cứu chuyên biệt. Các khối liệu với dung lượng hàng trăm triệu từ đã được biên soạn như: *Cobuild Corpus*, hệ thống *Longman Corpus* (LLELC, LSC and LCLE), và tiêu biểu là tập khối liệu quốc gia Anh - *British National Corpus* (BNC) (xem Aston and Burnard 1998, Leech và đồng sự 2001).

Phương pháp khối liệu phân tích các mô hình và ứng dụng của ngôn ngữ thực tế thông qua tập hợp văn bản được điện toán hoá. Phương pháp này được sử dụng trong nhiều lĩnh vực nghiên cứu ngôn ngữ khác nhau như: biên soạn tự điển (Longman Dictionary of Contemporary English 1995, Collins COBUILD English Dictionary 1995); sách tham khảo ngữ pháp (Longman Grammar of Spoken and Written English 1999).

Biber (1994) chỉ ra 4 thuộc tính của phương pháp phân tích dựa trên khối liệu: (1)phân tích các mô hình ứng dụng của ngôn ngữ dựa trên các văn bản tự nhiên; (2) tập hợp các văn bản tự nhiên thành một khối liệu làm cơ sở để phân tích; (3) sử dụng rộng rãi các phần mềm máy tính, bao gồm cả các kỹ thuật tự động lẫn tương tác để phân tích; và (4) áp dụng cả phương pháp định lượng lẫn định tính trong phân tích.

2.2. Thiết kế khối liệu

2.2.1. Thu thập dữ liệu

Dữ liệu được thu thập để xây dựng một khối liệu phải đầy đủ và phù hợp với mục đích nghiên cứu. Hunston (2002) chỉ ra 4 nguyên tắc cơ bản trong việc thu thập dữ liệu bao gồm: *dung lượng* (size), *nội dung* (content), *sự cân xứng* (balance) và *tính đại diện* (representativeness).

Các khối liệu có dung lượng lớn thường được thiết kế và xuất bản nhằm mục đích sử dụng chung, so sánh với dữ liệu phân tích từ các khối liệu chuyên biệt. Nội dung của khối liệu được quyết định bởi mục đích của nhà nghiên cứu trong việc sử dụng khối liệu. Sự cân xứng của khối liệu được thể hiện qua việc lựa chọn thể loại văn bản.

Chẳng hạn văn bản viết bao gồm các loại hình sách báo, tạp chí, bài viết, thư từ,...Văn bản nói bao gồm các cuộc thoại, phóng vấn,

tranh luận, bài phát biểu, bản tin,...Loại hình văn bản được lựa chọn phục vụ cho mục tiêu nghiên cứu phải là đại diện cho đối tượng văn bản đang được nghiên cứu. Ví dụ để thực hiện một nghiên cứu về ngôn ngữ báo chí, khối liệu được tập hợp phải bao gồm đa dạng các loại hình báo: báo khổ lớn hay tin vắn (broadsheet or tabloid), các thể loại bài báo khác nhau như điểm tin, bình luận, phân tích hay quảng cáo,... Mỗi loại đều phải chọn đầy đủ các đối tượng để đảm bảo tính đại diện của khối liệu.

2.2.2. Xây dựng khối liệu

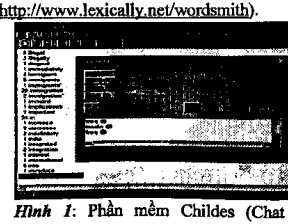
Yếu tố tiên quyết đối với việc xây dựng khối liệu là dữ liệu phải đầy đủ và phù hợp. Dữ liệu sau khi đã được tập hợp sẽ được chuyển thể thành văn bản được máy tính hoá. Văn bản nói phải được chuyển thể cho phù hợp với yêu cầu về định dạng của phần mềm ứng dụng để phân tích khối liệu. Chẳng hạn, trong một nghiên cứu về ngôn từ sử dụng trong báo chí nước Anh về vấn đề nhập cư và tị nạn (immigration and asylum), tác giả bài viết, Gabrielatos và Baker (2008), đã xây dựng một tập khối liệu với dung lượng 140 triệu từ gồm tập hợp các bài báo viết được xuất bản từ năm 1996 đến 2005 về đề tài này. 19 tờ báo ở nước Anh đã được lựa chọn để thu thập dữ liệu, bao gồm 5 tờ báo khổ lớn (*Business, Guardian, Herald, Independent và Telegraph*); 2 tờ ra ngày Chủ Nhật (*Observer, Independent on Sunday*), 6 tờ nhật báo khổ nhỏ (*Sun, Daily Star, People, Daily Mirror, Daily Express, Daily Mail*), 4 tờ tin Chủ Nhật (*Sunday Express, Mail on Sunday, Sunday Mirror, Sunday Star*), và 2 tờ báo địa phương. Sự lựa chọn đa dạng các thể loại báo viết nêu trên đảm bảo tính đại diện của khối liệu được xây dựng như Bảng 1 dưới đây:

Loại báo	Số lượng bài báo		Số từ	
	Số lượng	Tỉ lệ	Số lượng	Tỉ lệ
khổ lớn	100.242	57,24%	87.001.072	62,36%
khổ nhỏ	50.476	28,82%	29.883.001	21,42%
địa phương	24.421	13,94%	22.625.964	16,22%
Tổng cộng	175.139	100%	139.510.037	100 %

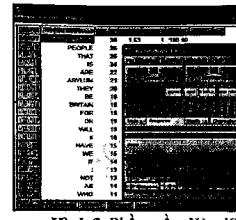
Bảng 1: Khối liệu về báo viết ở nước Anh

2.2. Sử dụng phần mềm hỗ trợ phân tích khối liệu

Có nhiều phần mềm và công cụ chuyên dụng khác nhau được sử dụng cho nghiên cứu cụ thể về ngôn ngữ bằng phương pháp phân tích khối liệu. Bài viết này giới thiệu hai phần mềm được sử dụng khá phổ biến trong các nghiên cứu bằng phương pháp khối liệu là CHAT & CLAN (http://childes.psy.cmu.edu/clan) và WordSmith 5.0 (http://www.lexically.net/wordsmith).

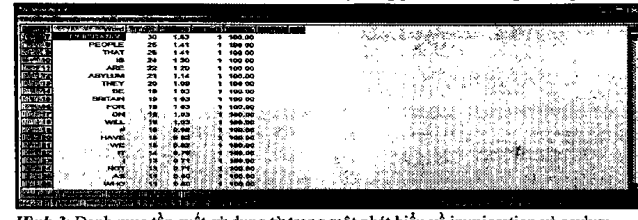


Hình 1: Phần mềm Childes (Chat & Clan)



Hình 2: Phần mềm WordSmith 5.0

Dữ liệu cung cấp cho phương pháp phân tích khối liệu chủ yếu dựa trên các công cụ cơ bản như từ khoá (key word) danh mục từ (wordlist) tần suất sử dụng từ (frequency list) và cấu trúc liên kết (concordance lines). Bài viết này phân tích phát biểu của Michael Howard vào ngày 22/9/2004 về vấn đề nhập cư và tị nạn (*immigration và asylum*) ở nước Anh nhằm minh hoạ tiện ích của các công cụ này trong phân tích diễn ngôn tiếng Anh.



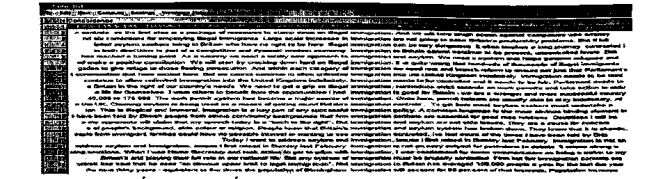
Hình 3: Danh mục tần suất sử dụng từ trong một phát biểu về immigration và asylum

2.2.1. Công cụ danh mục tần suất (frequency list)

Công cụ này cung cấp các thông tin thống kê về tần số xuất hiện của từ trong khối liệu (Hình 1, Hình 2), số lượng từ đếm được trong một văn bản cụ thể (types) và số lượng các từ khác nhau được sử dụng trong văn bản đó (tokens). Với công cụ này, người nghiên cứu có thể tìm thấy tần số xuất hiện của từ xếp theo danh mục giảm dần hay theo thứ tự ABC; xác định từ khoá của văn bản; phân tích cấu trúc liên kết và ý nghĩa của từ khoá trong một văn bản; so sánh ngôn từ trong nhiều văn bản cùng thể loại. Hình 3 minh hoạ danh mục về tần suất sử dụng từ trong bài phát biểu của Michael Howard được thực hiện bằng phần mềm WordSmith 5.0.

2.2.2. Công cụ phân tích tính kết hợp (concordance lines)

Sau khi đã xác lập được danh mục các từ khoá của văn bản (keyword list) người nghiên cứu có thể tìm hiểu ngữ cảnh và cấu trúc liên kết của một từ khoá bất kì trong văn bản bằng công cụ cấu trúc liên kết (concordance lines). Công cụ này cung cấp dữ liệu thống kê cả về cấu trúc ngữ pháp lẫn ngữ nghĩa của một từ cụ thể trong văn bản. Dữ liệu thống kê do các công cụ tiện ích này cung cấp sẽ giúp người nghiên cứu thực hiện việc phân tích các đối tượng ngữ pháp, từ vựng, ngữ nghĩa, ... của một ngôn bản. Ví dụ hình 4 dưới đây minh hoạ các cấu trúc liên kết của từ khoá "immigration" xuất hiện trong bài phát biểu nêu trên.



Hình 4: Các cấu trúc liên kết (concordance lines) của từ khoá "immigration"

Trong các cấu trúc liên kết của từ khoá "immigration" ở Hình 4, các nghiên cứu về từ vựng, ngữ pháp và ngữ nghĩa liên quan đến thuật ngữ "immigration" sẽ được thực hiện. Tác giả sử dụng thủ thuật ẩn dụ ý niệm để ngụ ý về tình trạng tị nạn và nhập cư ở nước Anh hiện nay như hiểm hoạ tự nhiên (natural disaster) thông qua các cấu trúc diễn đạt như: *the tide of immigration* (làn sóng nhập cư), *the meltdown in the immigration system* (sự lan tràn trong hệ thống nhập cư), *the chaos in the immigration system* (sự hỗn độn trong hệ thống nhập cư), *immigration crisis that has engulfed the Government* (khủng hoảng nhập cư đã nhấn chìm chính phủ), *this new influx that is about to engulf us* (sự tràn ngập mới này sắp nhấn chìm chúng ta), *the flood of asylum seekers* (dòng lũ những người tị nạn),...

Từ dữ liệu thống kê cấu trúc liên kết của từ khoá "immigration", có thể dễ dàng xác định ngụ ý của tác giả về tình trạng "*nhập cư và tị nạn*" được diễn đạt trong bài phát biểu theo các ý niệm sau đây:

(1) Tình trạng nhập cư ở nước Anh hiện nay đã vượt ra khỏi tầm kiểm soát (out of control): 1/ Britain's immigration controls today are neither firm nor fair;2/ People know that Britain's immigration and asylum system has broken down;3/ They know that it is chaotic, unfair and out of control;4/ Immigration needs to be controlled and it needs to be fair;5/(...) immigration to Britain cannot continue at its present, uncontrolled levels.

(2) Nhập cư bất hợp pháp đang gia tăng và ngày càng trở nên nguy hiểm: 1/ We need to get a grip on illegal immigration;2/ A Conservative Government would re-introduce embarkation controls, as the first step in a package of measures to clamp down on illegal immigration;3/Illegal immigration can be very dangerous.

(3) Phản đối đề xuất không kiểm soát vấn đề nhập cư hợp pháp: 1/David Blunkett may believe that there is "no obvious upper limit to legal immigration";2/Blunkett has said that he sees "no obvious upper limit to legal immigration";3/They seemed to believe that British people from immigrant families could have no possible interest in wanting to see immigration controlled.

(4) Kiểm soát chặt chẽ vấn đề nhập cư là thiết yếu: 1/Firm but fair immigration controls are essential for good race relations; 2/Any system of immigration must be properly controlled.

3. Kết luận

Phân tích diễn ngôn là một trong những phương diện quan trọng của ngôn ngữ học. Trong thời đại bùng nổ thông tin ngày nay, văn bản được thể hiện bằng nhiều dạng thức và thể loại phong phú thông qua các phương tiện giao tiếp khác nhau. Với sự hỗ trợ của các chương trình máy tính và các phần mềm chuyên dụng, việc xây dựng các khối liệu đã trở nên thuận lợi và nhanh chóng hơn rất nhiều.

Chính nhờ các khối liệu văn bản, việc tìm hiểu một đối tượng nghiên cứu trong ngôn

ngữ đã trở nên đơn giản và chính xác hơn nhiều thông qua các công cụ phần mềm. Khối liệu đã trở thành công cụ tập hợp và mang tính đại diện cao cho một loại hình văn bản cụ thể để nghiên cứu. Phân tích diễn ngôn bằng phương pháp phân tích dựa trên khối liệu đã trở thành một hướng tiếp cận chính cho các nghiên cứu ngôn ngữ, từ các nghiên cứu về từ vựng đến ngữ pháp, ngữ nghĩa. Khối liệu có thể được xem như kho tàng tiềm ẩn rất nhiều các vấn đề thú vị về ngôn ngữ. Phân tích khối liệu vì vậy đã trở thành phương pháp phổ biến nhất đối với phân tích diễn ngôn.

Như vậy, phân tích khối liệu có thể được xem là một trong những phương pháp cung cấp thông tin hữu dụng để thực hiện việc nghiên cứu một vấn đề cụ thể trong ngôn ngữ thông qua đối chiếu và minh hoạ lí thuyết bằng dữ liệu thu thập từ ngôn ngữ giao tiếp thực tế. Vì vậy, kết quả nghiên cứu một vấn đề cụ thể của ngôn ngữ được quyết định bởi việc xây dựng khối liệu nghiên cứu. Bài viết giới thiệu phương pháp nghiên cứu diễn ngôn trên cơ sở phân tích khối liệu bao gồm: thu thập văn bản về vấn đề nghiên cứu, tích hợp dữ liệu trên máy tính, phân loại dữ liệu đã được tích hợp và hoàn chỉnh khối liệu, sử dụng phần mềm phù hợp để cung cấp số liệu thống kê phục vụ việc mô tả và phân tích một đối tượng ngôn ngữ bất kì.

TÀI LIỆU THAM KHẢO

1. Baker, P. (2006), *Using corpora in discourse analysis*. London: Continuum Discourse Series.

2. Biber, D., Conrad, S. & Reppen, R. (1998), *Corpus linguistics investigating language structure and use*. Cambridge University Press.

3. Brown, G. & Yule, G. (1983), *Discourse analysis*. Cambridge University Press.

4. De Beaugrande, R. & Dressler, W. (1981), *Introduction to text linguistics*. London & New York: Longman.

5. Gabrielatos, C. & Baker, P. (2008), *A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005*. Journal of English Linguistics 36 (1): 5-38

6. Halliday, M.A.K. & Hasan, R. (1976), *Cohesion in English*. London: Longman.

7. Halliday, M.A.K. (1991), *Corpus studies & probabilistic grammar*, in Aijmer, K. & Altenberg, B. 1991. English Corpus Linguistics. London & New York: Longman.

8. Hoey, M. (2001), *Textual interaction: an introduction to written discourse analysis*. London & New York: Routledge.

9. Hunston, S. (2002), *Corpora in applied linguistics*. Cambridge University Press.

10. Jaworski, A. & Coupland, N. 2nd edition (2006), *The discourse reader*. London & New York: Routledge.

11. Kennedy, G. (1998), *An introduction to corpus linguistics*. London & New York: Longman.

12. McCarthy M. (2006), *Explorations in corpus linguistics*. Cambridge University Press.

13. McEnery T. & Wilson A. (2001), *Corpus linguistics*. Edinburgh University Press.

14. Meyer, F. C. (2004), *English corpus linguistics*. Cambridge University Press.

15. Nguyễn Thiện Giáp (2004), *Dụng học Việt ngữ*. Nxb ĐHQG Hà Nội.

16. Reah, D. (2002), *The language of newspapers*. London & New York: Routledge.

17. Schiffrin, D. (1994), *Approaches to discourse*. Oxford: Blackwell Publishers Ltd.

18. Stubbs, M. (1983), *Discourse analysis the sociolinguistic analysis of natural language*. Oxford: Blackwell Publishers Ltd.

19. Stubbs, M. (1996), *Text and corpus analysis: Computer-assisted studies of language and culture*. Cambridge: Blackwell Publishers Inc.