Phần Bảy

Từ Khoa học Kỹ thuật sang Khoa học Nhân văn (Ngôn ngữ học)

Góp phần cho Thế hệ Tương lai

Khoảng năm 1996, sau khi làm việc trên 15 năm tại khoa Hóa học Công nghệ, đại học UBC, Vancouver, Canada và gần 10 năm với công ty Điện lực Việt Nam, tôi nghĩ đến nghề cũ: dạy trong thời gian theo học bằng Cử nhân và có ý định về thăm lại trường cũ với bao nhiêu kỹ niệm trong tâm trí. Trước trào lưu phát triển tột độ của máy tính và Điện toán nói chung, tôi quyết định bỏ Khoa học Kỹ thuật để chuyển sang Khoa học Nhân văn, mặc dù tôi chưa có một định hướng rõ ràng.

Năm 1996 tôi có thực hiện một trang Web bằng tiếng Việt để phổ biến tin học. Nhiều bài về tin học căn bản và Thông minh Nhân tạo (Artificial Intelligence) được nhiều người ưa thích.

Thời gian này chính phủ Việt Nam đang trong vòng tranh luận: Nên hay Không Nên cho dân chúng tiếp cận với Internet. Mãi đến ngày 19 tháng 11 năm 1997, Việt Nam mới thực sự hòa mạng với Internet. Nhờ đó Việt Nam mới có cơ hội thấy được sự tiến bộ của thế giới.

Trong thời gian này chính phủ Canada và các Tổ chức Quốc tế có nhiều dự án giúp Việt Nam:

- Dự án đào tạo Tiến sĩ Đệ Tam Cấp / Cao học về Giáo dục (Master of Education) tại trường Simon Fraser University, do giáo sư Allan MacKinnon hướng dẫn.
- Dự án thực hiện Luật Biển, do giáo sư Ian Townsend-Gault Khoa Luật – UBC đảm trách.

Việt Nam có gởi nhiều phái đoàn gồm các giáo sư đại học, sang Canada để tìm đối tác. Trong số này có nhóm do giáo sư Hoàng Kiếm, giám đốc Trường Tin Học, đại học TP HCM hướng dẫn. Một trong những ước mơ của giáo sư Hoàng Kiếm là được trường British Columbia Institute of Technology (BCIT) hợp tác để mở những lớp học về Tin học và Kỹ thuật tại Sài gòn. Dự án này không thực hiện được vì quan niệm điều hành về giáo dục rất khác nhau giữa hai chính phủ.

Sau đó giáo sư Hoàng Kiếm có mời một giáo sư người Việt tại khoa Điện của trường UBC hợp tác, đồng thời góp ý với tôi: nếu có thể làm được gì để giúp sinh viên thì xin giúp đỡ, vì hiện nay tài liệu về kỹ thuật cao, đặc biệt là về Thông minh Nhân tạo, rất cần. Tự thấy mình không có đủ tư cách để thực hiện những dự án lớn, nhưng rất muốn giúp, nên tôi tự học và viết một số tài liệu ngắn, xúc tích và thực tiễn để giúp sinh viên của giáo sư Hoàng Kiếm.

Mùa Hè năm 1997, tôi về thăm trường cũ và được giáo sư Hoàng Kiếm giới thiệu với một số sinh viên của trường.

Nhân đó tôi có tặng trường một số tài liệu tôi thực hiện.

Thực ra ý tưởng thực hiện 24 tài liệu này đã có từ khi tôi còn làm tại khoa Hóa Học Công Nghệ - UBC. Phải mất một thời gian dài để đúc kết tài liệu và viết thành sách. Lúc đó tôi cũng đam mê về kỹ thuật In-Ấn bằng máy tính (Desktop Publishing) nên việc in tài liệu thành những cuốn sách nhỏ, như trong các hình trên, không khó khăn.

Giáo sư Hoàng Kiếm có ý định in và phổ biến những tài liệu này tại Việt Nam; nhưng vào thời điểm đó, việc xuất bản tài liệu giáo khoa hoàn toàn do nhà nước quyết định, nên ước muốn này không thực hiện được.

Tuy nhiên những tài liệu này đã giúp nhiều sinh viên trong học tập cũng như trong việc thực hiện các dự án tốt nghiệp. Đặc biệt sinh viên Cao học Lê Hoàng Thái đã dựa trên tài liệu của chúng tôi để viết sách Thuật Giải Di Truyền – cách Giải Tự Nhiên các Bài Toán Trên Máy Tính.

Nếu muốn đi sâu vào lãnh vực Thông minh Nhân tạo, như Luận lý Mờ (Fuzzy Logic), Lý luận dựa theo sự Tiến hóa (Genetic Algorithms), cần phải có một căn bản vững chắc về toán học và nhiều kinh nghiệm về lập trình. Tôi không có đủ hai yếu tố này, nên đã chọn Ngôn ngữ học Ngữ Liệu (Corpus Linguistics).

Do đó từ năm 1996 tôi bắt đầu làm việc bán thời gian cho Thống kê Canada (Statistics Canada) có sử dụng khả năng thông/phiên dịch. Tôi dành nhiều thì giờ cho Ngôn ngữ học, đặc biệt là Ngôn ngữ học Ngữ liệu (Corpus Linguistics). Từ đó tôi đi dần đến ý tưởng thực hiện Ngữ liệu Song ngữ (Parallel Corpora), Anh Việt và nhiều ngôn ngữ khác.

Dự án 1-triệu cặp câu song ngữ bắt đầu từ năm 2000 cho đến nay.

Ngữ liệu song ngữ của chúng tôi không ở dạng những bài bằng ngôn ngữ gốc (source language) và bản dịch của chúng sang ngôn ngữ đích (target language), ở dạng in ra giấy, vì như vậy tuổi thọ của chúng rất ngắn (in thành sách rồi về sau không dùng nữa thì bán như phế phẩm) và mức độ sử dụng rất hạn chế (chỉ để đọc hay tra cứu đơn giản).

Ngữ liệu do chúng tôi thực hiện được lưu ở dạng điện tử và được cắt ngắn thành câu (phrase), nhờ đó có thể xử lý bằng máy tính và được dùng cho nhiều ứng dụng khác nhau: thông/phiên dịch, lập từ điển song ngữ, lập bảng thuật ngữ chuyên ngành, quan sát trực tiếp ngữ pháp (grammar), tìm kiếm từ trong ngữ cảnh, khảo sát sự khác biệt giữa các ngôn ngữ v.v...

Việc thu thập các tài liệu song ngữ không khó khăn, nhưng việc xử lý để biến các văn bản thành câu ngắn rất nhiêu khê. Một trong những lý do chính là khi dịch, các nhà phiên dịch không giữ nguyên cấu trúc nguyên thủy của văn bản gốc. Phiên dịch, ngoài việc phải giữ nguyên ý nghĩa của văn bản gốc, còn phải diễn tả sao cho người đọc (ở ngôn ngữ đích) hiểu được nội dung tài liệu. Do đó người dịch có khi cắt một câu trong văn bản gốc thành hai hay nhiều câu trong văn bản địch; hay ngược lại gom nhiều câu trong văn bản gốc thành một câu trong văn bản dịch.

Trong hầu hết các ngôn ngữ, các từ thường có nhiều nghĩa, do đó để có thể hiểu một cách chính xác, người đọc phải tra cứu từ trong ngữ cảnh (word in context).

Ngày nay chúng ta đã có nhiều phương tiện để cắt văn bản (Căn chỉnh = Alignment) thành câu ngắn và phần mềm Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing).

Tiếng Việt không phải là ngôn ngữ chính thống trên thị trường, do đó nhiều phần mềm máy tính có sẵn không thể dùng cho tiếng Việt. Chúng tôi đã phải thực hiện những tiện ích (Utility) và những phần mềm thích nghi cho tiếng Việt. Cho đến nay chúng tôi xử dụng lối 10 phần mềm miễn phí hay với giá tương đối thấp và gần 10 phần mềm và tiện ích do chúng tôi thực hiện.

Trong công tác thực hiện 10 phần mềm và tiện ích này, tôi được sự giúp đỡ tích cực về kỹ thuật lập trình của Lê Quang Trường và của Bác sĩ Bảo Phi.

Bác sĩ Bảo Phi là một bác sĩ gia đình làm việc tại bệnh viện Long Khánh. Nhưng cũng như tôi, bác sĩ Bảo Phi đã dành nhiều thời gian cho Tin học hơn là công việc chính của mình. Tôi biết bác sĩ Bảo Phi qua Internet. Từ năm 2003 đến nay, mỗi khi tôi có ý tưởng mới và có kinh phí để đài thọ cho công tác, tôi phát họa ý tưởng rồi nhờ đến bác sĩ Bảo Phi lập trình.

Một số phần mềm chính mà chúng tôi đã dùng cho tiếng Việt là: Wordle, AntConc, Wordsmith Tools, ParaConc, Voyant Tools, SkELL, và Sketch Engine.

Một số phần mềm và Tiện ích (Utility) do nhóm chúng tôi (với bác sĩ Bảo Phi và Lê Quang Trường) thực hiện gồm: TMToolkit, TransMem, VinaConcordance, PcbT (Parallel Corpora-Based Translation), LQV, Notepad+.

Từ khi giao tiếp với thế giới, Việt Nam đã cơ hội giao lưu với trên 100 quốc gia và Tổ chức Quốc tế, do đó ngoài tiếng Anh chúng ta cũng có nhu cầu tìm hiểu và sử dụng nhiều ngôn ngữ khác, như: tiếng Pháp, Đức, Ý, Nhật, Trung, Nga, Hàn, Ấn độ, Tây Ban Nha, Bồ Đào Nha và đặc biệt là tiếng Á-rập (Arabic). Do đó ngoài Dự án 1-triệu cặp câu Anh-Việt, chúng tôi có thêm Dự án Trăm-ngàn câu Song ngữ cho những ngôn ngữ nêu trên. Trong khi thực hiện hai dự án trên, tôi sử dụng nhiều phần mềm khác nhau, nhờ đó thấy được sự đa dạng của Ngôn ngữ và dần dần đưa tôi đến với Ngôn ngữ học Ngữ liệu (Corpus Linguistics). Tôi rất đam mê lãnh vực Ngữ liệu Song ngữ vì những lợi ích mà nó đem lại :

- Giúp người dùng hiểu được nội dung văn bản gốc ;
- Giúp người học tiếng (Anh, Nhật, Trung, Hàn, Pháp, Ý, Đức, Tây Ban Nha, v.v.)
- Giúp các thông dịch viên / phiên dịch viên tạo các Bộ nhớ Phiên dịch ;
- Giúp các nhà Từ điển học thực hiện các Từ điển Song và Đa ngữ.

Tìm kiếm, Thu thập và Xử lý Ngữ liệu song Ngữ

Trong phần này chúng tôi sẽ trình bày 3 giai đoạn chính của việc Tìm kiếm, Thu thập và Xử lý.

A- Tìm tài liệu song ngữ : Trên Internet có rất nhiều bản dịch của các tài liệu khoa học, kỹ thuật, kinh doanh và đặc biệt là các tiểu thuyết. Đặc biệt trên trang Project Gutenberg (PG) chúng ta có thể tải xuống hàng triệu tài liệu miễn phí. Rất nhiều tài liệu này đã được dịch sang tiếng Việt. Nhiều quốc gia, tổ chức quốc tế, hội đoàn vô vụ lợi, cũng như nhiều cá nhân (trường hợp Ngô Bắc) đã tạo ra hàng nghìn bản dịch của những tài liệu đáng giá. Tuy nhiên việc tìm ra cả bản gốc và bản dịch là một công tác mất nhiều thời gian. Chúng tôi đã tìm được giải pháp để truy tầm cùng lúc văn bản gốc và bản dịch, hoặc tìm ngay tức thì bản gốc khi có bản dịch sang tiếng Việt.

B- Thu thập tài liệu song ngữ : Phần lớn các bản dịch được trình bày ở dạng PDF, Epub hay

HTML, rất ít khi ở dạng DOC (Word) hay TXT (Notepad). Dạng PDF được ưa chuộng bởi vì nó đảm bảo chất lượng của tài liệu ; tuy nhiên có nhiều trường hợp chúng ta tìm thấy nhiều tài liệu tiếng Việt ở dạng PDF, nhưng không đọc được. Lý do của việc này là vì người phiên dịch tài liệu (người Việt) và người biến đổi tài liệu từ DOC sang PDF (không phải là người Việt và không hiểu nguyên tắc cơ bản của việc chuyển đổi. Việc dùng các phông chữ không phải là phông chữ Unicode cũng góp phần tạo lỗi nêu trên. Chúng tôi đã gặp nhiều trang web với toàn bộ tài liệu có sai lỗi này và đã chuyển đổi sang Unicode để dùng.

C- Xử lý tài liệu song ngữ : Sau khi thu thập bản gốc và bản dịch ở dạng sử dụng được, chúng ta phải xử lý : loại bỏ những phần không cần thiết cho việc trình bày nội dung (thí dụ các tag trong tập tin dạng DOC), các ghi chú cũng như những chi tiết về tác giả, về tham chiếu, v.v. Giai đoạn quan trọng nhất của việc xử lý văn bản song ngữ là việc căn chỉnh (alignment). Mục đích chính của việc căn chỉnh là biến cặp tài liệu (gốc và dịch) thành tập hợp các cặp câu song ngữ : số câu trong tài liệu gốc bằng với số câu trong tài liệu dịch. Tuy việc căn chỉnh không quan trọng cho việc tìm hiểu nội dung hai tài liệu, nhưng nó rất quan trọng cho việc tạo Bộ nhớ Phiên dịch (Translation Memory), bản Thuật ngữ Song nữ (Bilingual Glossary) và Từ điển Song ngữ. Việc căn chỉnh văn bản phần lớn dựa trên dấu chấm ở cuối câu. Đối với những ngôn ngữ gốc La-tinh thì việc căn chỉnh rất đơn giản và chính xác (gần 100%). Tuy nhiên đối với các ngôn ngữ như tiếng Trung, tiếng Nhật, tiếng Hàn và đặc biệt là tiếng Á-Rập, việc căn chỉnh thường chỉ đạt ở mức 60 hay 70%. Nội dung văn bản và Hình thức thể hiện văn bản cũng góp phần cho sự thành công của việc căn chỉnh : tài liệu khoa học và kỹ thuật dễ căn chỉnh hơn tiểu thuyết ; văn bản có đánh số thứ tự, như Mục Lục, giúp nhiều cho việc thành công ; bản dịch có ghi thêm chi tiết bằng ngôn ngữ gốc cũng giúp phần mềm căn chỉnh làm việc chính xác hơn.

Sau khi đã căn chỉnh thành công, ngữ liệu song ngữ sẽ được trình bày ở nhiều dạng khác nhau tùy theo mục tiêu sử dụng. Ngữ liệu song ngữ được trình bày phổ thông nhất là dạng Excel, kế đó là TXT, DOC hay HTML. Tuy nhiên nó cũng được trình bày dưới dạng đặc thù của phần mềm đặc biệt, như dạng TMX, v.v.

Trong phần kế tôi sẽ trình bày kết quả nghiên cứu của chúng tôi, từ ứng dụng trong phiên dịch, sang sử dụng trong học tiếng (Anh, Trung, Nhật, Hàn, v.v.) sang so sánh ngôn ngữ. Chúng tôi sẽ

trình bày kết quả nghiên cứu qua hình ảnh và qua các trang web trình công tác.

1 - Trong trang Web : https://anchor.fm/vina-technology

Chúng tôi trình bày các bài nói chuyện về Ngôn ngữ học, Trí tuệ Nhân tạo và Blockchain.

2 – Trong trang web : https://dulieuphiendich.com

Chúng tôi trình bày các tập tin song ngữ : Việt-Anh, Việt-Trung , Việt- Nhật, Việt-Hàn. v.v

Hai trang web này đều được cập nhật để có thêm tài liệu hàng tuần

3 – Các trang web trên Facebook trình bày tài liệu phân tích ngữ liệu tiếng Việt:

https://www.facebook.com/1581278281/videos/10205139330434484

https://www.facebook.com/1581278281/videos/10205139357875170

https://www.facebook.com/photo.php?fbid=10209678708956110&set=pb.1581278281.-2207520000..&type=3

Tôi sẽ đưa quý vị và các bạn đến với Ngôn ngữ học Ngữ liệu theo cách "cưỡi ngựa xem hoa" trước khi đi vào lý thuyết và những ứng dụng thực tế của ngành này.

Ngôn ngữ học Ngữ liệu và Thiên văn học

Khi đọc tài liệu, chúng ta thường chỉ để ý đến nội dung của tài liệu chứ không quan tâm nhiều đến hình thức trình bày hay cách dùng chữ (từ) của tác giả. Một phần vì thời gian không cho phép, nhưng phần lớn là vì thiếu phương tiện và không cần thiết.

Người đọc khảo sát các từ (word) hay nhóm từ trong văn bản cũng giống như các nhà thiên văn học khảo sát các vì sao trên bầu trời.

Các nhà thiên văn học dùng các kính thiên văn (kính viễn vọng) để tìm hiểu về các vì sao.

Các nhà ngôn ngữ học ngữ liệu khảo sát các từ (word) hay nhóm từ trong văn bản nhờ phần mềm Xử lý Ngôn ngữ Tự nhiên (Natural Language Processing).

Nhìn sơ qua chúng ta thấy có sự tương đồng trong hai lãnh vực này. Tuy nhiên khảo sát kỹ hơn chúng ta thấy có nhiều khác biệt giữa hai ngành.

Đối với Thiên văn học, những nhà nghiên cứu phải có căn bản khoa học, đặc biệt là Vật lý. Họ là những chuyên viên có trình độ đại học và nhiều năm kinh nghiệm. Phương tiện nghiên cứu của họ, kính thiên văn, rất đắt tiền và chỉ có ở các đài thiên văn.

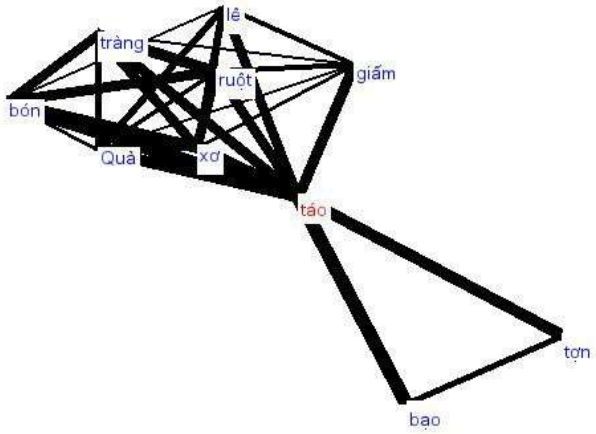
Ngoài ra đối tượng nghiên cứu của Thiên văn học, các vì sao, ở xa chúng ta hàng ngàn hay triệu năm ánh sáng.

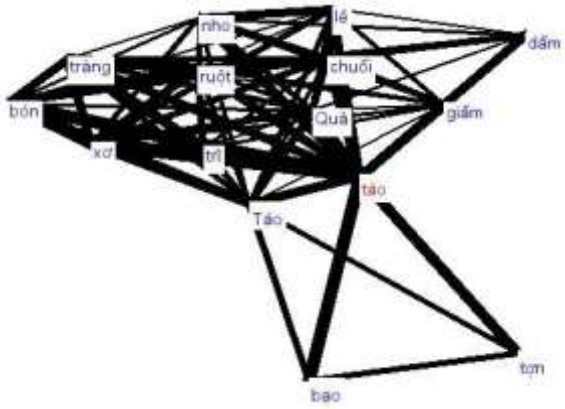
Đối với Ngôn ngữ học Ngữ liệu, các tài liệu để nghiên cứu ở ngay trước mặt, các phương tiện tìm hiểu Ngôn ngữ học Tự nhiên là những phần mềm miễn phí hay với giá tương đối rẻ. Đặc biệt hơn nữa là những người khảo sát Ngôn ngữ học Ngữ liệu không nhất thiết là những giáo sư, giáo viên có trình độ đại học hay nhiều năm trong nghề; họ có thể là sinh viên, hay bất cứ ai quan tâm đến ngôn ngữ. Tôi sẽ sử dụng những phần mềm xử lý ngữ liệu miễn phí để chứng minh điều này.

Máy tính và Con người

Máy tính và các phần mềm liên hệ không thể thay thế hoàn toàn con người, nhưng có thể biết được con người nghĩ gì và thể hiện ý nghĩ đó.

Thí dụ khi chúng ta nói đến từ (word) "táo", chúng ta có thể liên tưởng đến một loại trái cây, nhưng cũng có thể nghĩ đến nhiều nghĩa khác. Máy tính và phần mềm Phân tích Ngữ liệu giúp chúng ta thấy rõ sự khác biệt này.





Sự khác biệt được trình bày bằng hình ảnh giúp chúng ta thấy ngay các chi tiết: từ quan tâm với màu đỏ, các từ liên quan màu với xanh dương và các đường nối sự liên kết với màu đen.

Hình bên trái cho chúng ta thấy hai nhóm ý nghĩa chính của từ "táo".

Nhánh bên tay phải cho thấy ý nghĩa của sự táo tợn, táo bạo (bạo lực).

Nhánh bên tay trái cho thấy nhiều ý nghĩa khác, nhưng chúng đều có liên hệ với nhau (qua những đường nối).

Nếu cung cấp nhiều văn bản hơn thì ý nghĩa bên tay phải không thay đổi nhiều, trong khi đó nhóm bên tay trái trở nên phong phú hơn.

Nếu tiếp tục cung cấp thêm nhiều tài liệu nữa chúng ta sẽ có một hình mới với nhiều chi tiết hơn.

Như vậy càng có nhiều văn bản thì hiểu biết của chúng ta càng phong phú.

Đó là yêu cầu đầu tiên của Ngôn ngữ học Ngữ liệu.

Ngữ liệu (corpus, số nhiều là corpora) là tập hợp nhiều văn bản ở dạng điện tử.

Ý nghĩa "nhiều" rất mông lung. Trước năm 2000 phần lớn các văn bản đều ở dạng in trên giấy, do đó nếu muốn chuyển sang dạng điện tử, phải chép bằng tay (đánh máy) hay quét (scan). Công việc này mất nhiều thời gian và công sức, nên ngữ liệu chứa vài triệu từ đã được xem là rất lớn.

Trong thập niên 60, ngữ liệu chứa 1-triệu từ đã được xem là rất lớn. Nhưng sang thập niên 70 ngữ liệu chứa 10-triệu từ được xem là tạm đủ để soạn từ điển. Đến thập niên 90 thì đã có thể tạo ngữ liệu lớn hơn, như Ngữ liệu Quốc gia Anh (British National Corpus) chứa 100-triệu từ.

Nhưng từ đầu thế kỷ 21, tài liệu điện tử có rất nhiều trên Internet, nên ý nghĩa "lớn" của ngữ liệu đã thay đổi, và không biết phải chứa bao nhiêu mới được gọi là lớn. Các nhà Từ điển học đã dùng Ngữ liệu chứa vài tỷ từ để soạn từ điển. Hiện nay đã có thể tạo một cách dễ dàng những ngữ liệu chứa hàng chục tỷ từ.

Câu hỏi quý vị có thể đặt ra là: khả năng một người tốt nghiệp đại học cũng chỉ có thể biết tối đa 100,000 từ, vậy tại sao phải cần có những ngữ liệu chứa hàng tỷ từ?

Câu trả lời là vì trong khi dùng từ, chúng ta đã dùng đi dùng lại nhiều lần một số từ, đặc biệt là những "hư từ", tuy không có ý nghĩa gì nhưng là cầu nối giữa những "thực từ".

Thí dụ: trong Truyện Kiều của Nguyễn Du, từ "một" xuất hiện 308 lần; từ "đã" 253 lần; từ "người" 219 lần; từ "nàng" 197 lần.

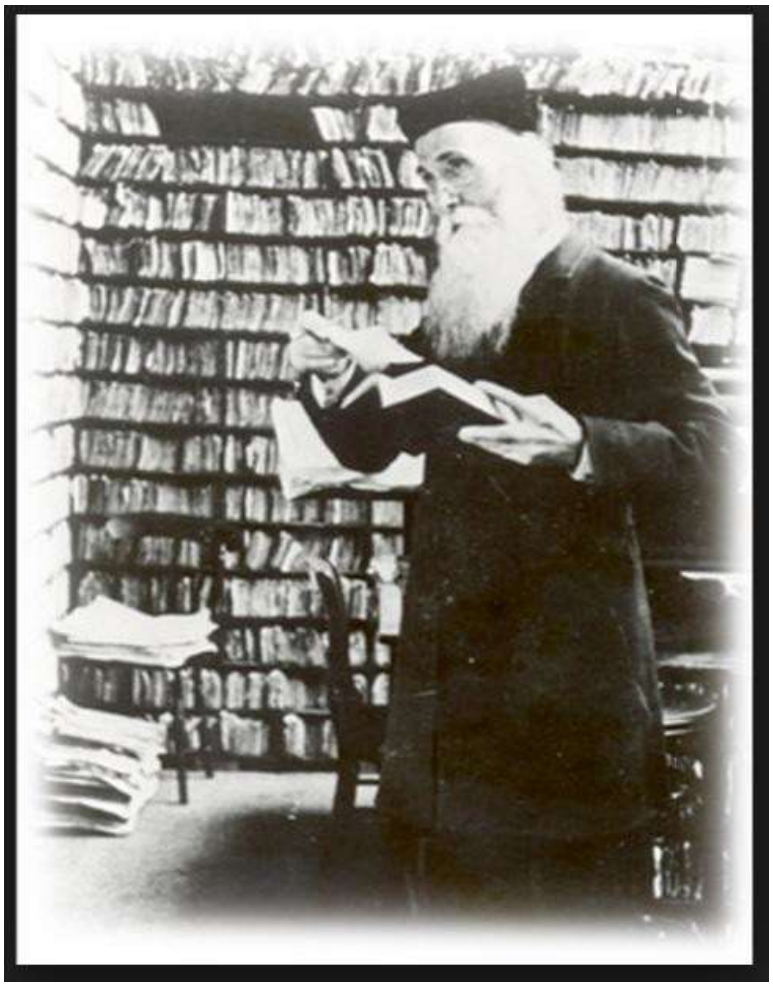
Trong ngữ liệu BNC chứa 100-triệu từ, có:

was	923,957	10 (đứng thứ 10, tính theo thứ tự nhiều đến ít)
at	478, 177	20
made	091,659	100
advice	010,316	1000
quiet	005,295	2000

Chúng ta thấy có một mối liên hệ giữa các từ này: “was” ở vị trí thứ 10 có gấp 2 lần từ “at” ở vị trí 20 và gấp 10 lần từ “made” ở vị trí 100, gấp 100 lần từ “advice” ở vị trí 1000, gấp 200 lần từ “quiet” ở vị trí 2000.

Như vậy giữa vị trí của một từ với số lượng của từ có một hệ thức, đó là định luật Zipf.
Định luật này gần đúng cho hầu hết các ngôn ngữ, và đây là định luật phổ quát.

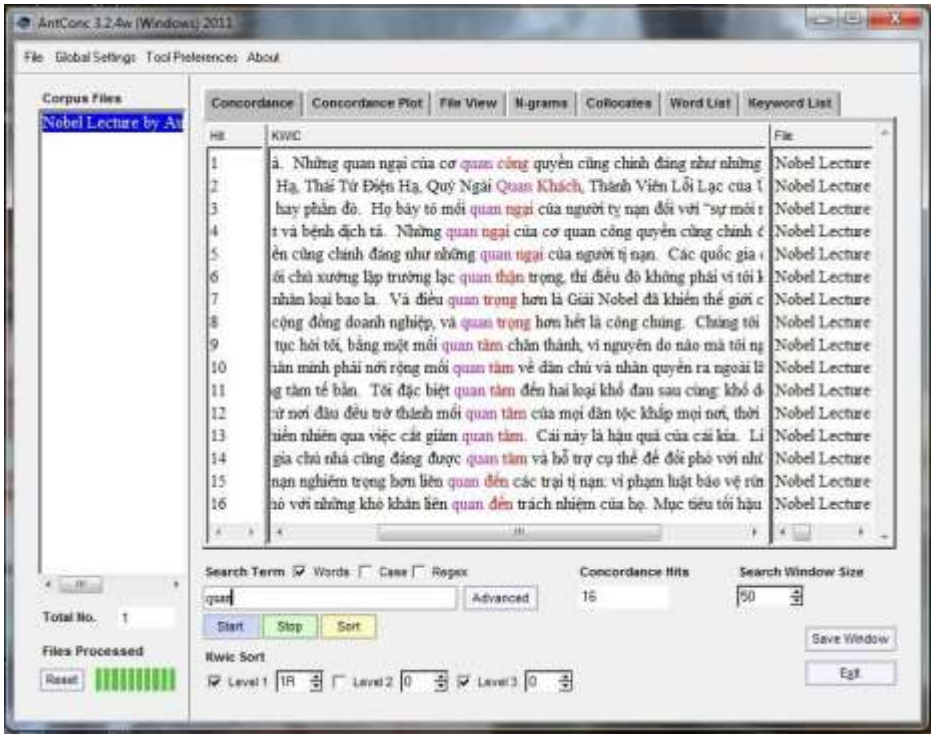
Nếu cộng tất cả 100 từ thường dùng nhất trong tiếng Anh (từ ở vị trí 1 cho đến từ ở vị trí 100), tổng số này chiếm đến gần 45% tổng số từ trong ngữ liệu nào. Thí dụ trong BNC, tổng số này gần 45 triệu từ. Như vậy những từ còn lại chỉ có mặt với số lần ít đi dần dần. Trong BNC, từ adjucate chỉ xuất hiện có 121 lần, từ inattentive chỉ xuất hiện 31 lần, và barnstorming chỉ xuất hiện 20 lần.



Kết quả này khiến chúng ta liên tưởng đến tin mới đăng trong tháng1-2015 về của cải của mọi người trên thế giới: Hơn nửa tài sản của nhân loại nằm trong tay của 2% dân số thế giới; bằng với của cải của 98% số người còn lại. Sự kiện này cũng giống như số từ trong mọi ngữ liệu.

Kết quả này có một ứng dụng rất quan trọng trong việc thực hiện từ điển.

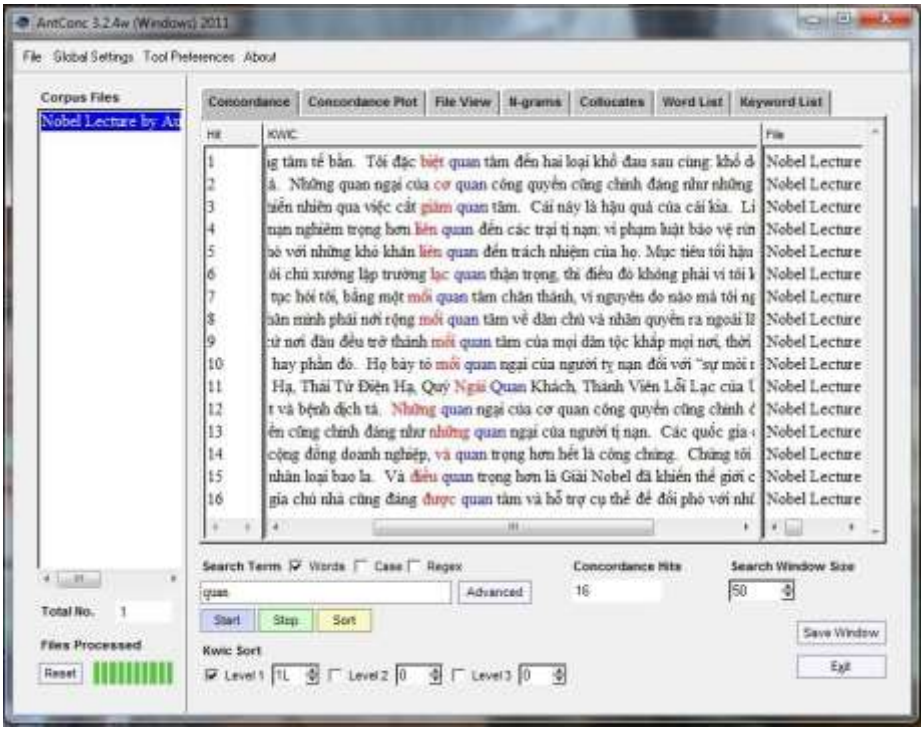
Trước đây, từ điển được thực hiện bởi các nhà từ điển học, thí dụ như James Murray đã dựa trên các Chỉ mục (Index) để có tài liệu soạn từ điển. Khi cần một từ nào đó, James Murray yêu cầu các phụ tá tìm trong các tủ (hình bên trên). Việc làm này mất thời gian và không chính xác. Có nhiều từ không còn được dùng, trong khi đó cần thêm từ mới nên kỹ thuật trên không hữu hiệu.



Ngày nay nhà làm từ điển, và ngay cả sinh viên, có thể tạo ra từ điển từ ngữ liệu lấy trên Internet. Vấn đề là phải dùng ngữ liệu chứa bao nhiêu từ mới đủ.

Câu trả lời là ngữ liệu càng lớn càng tốt, do đó ngày nay các nhà từ điển học đã dùng các ngữ liệu chứa hàng chục-tỷ từ.

Hình Ngữ cảnh 1



Hình Ngữ cảnh 2

Trước đây để tìm định nghĩa các từ, chúng ta phải dựa trên các câu có chứa từ liên hệ để biết nghĩa của từ đó. Việc này cũng có nghĩa như tìm từ trong ngữ cảnh (words in context), để biết từ đồng hiện (co-occurrence). Thí dụ đối với từ "quan", chúng ta có quan công, quan khách, quan ngại, quan tâm, v.v… đây là những nhóm từ do sự kết hợp giữa từ "quan" với từ đứng sau nó. Tuy nhiên cũng có những nhóm từ mà từ "quan" đứng sau các từ khác, thí dụ cảm quan, cơ quan, lạc quan, liên quan, v.v. …

Việc tìm kiếm những từ đứng trước hay đứng sau một từ quan tâm được thực hiện một cách dễ dàng với phần mềm: Trình bày Từ trong Ngữ cảnh (Concordancer).

Hình Ngữ cảnh 1 cho thấy từ "quan" được trình bày với màu tím lợt, những từ đứng sau nó được trình bày với màu đỏ lợt, trong khi những từ khác màu đen. Nhờ đó người dùng có thể thấy ngay sự khác biệt. Hình Ngữ cảnh 2 trình bày những từ đứng trước "quan". Cũng như hình trên, trong hình này "quan" được thể hiện với màu xanh, trong khi những từ đứng trước "quan" được trình bày với màu đỏ, còn các từ khác màu đen.

Các nhà nghiên cứu ngôn ngữ học cũng như những nhà soạn từ điển có thể dựa vào phần mềm này để tìm định nghĩa và thí dụ cho từng nghĩa của từ liên quan, như quan sát bằng mắt. Nhờ đó chúng ta có thể biết: Từ đứng trước, đứng sau một từ quan tâm; những tính từ đi với từ liên hệ cũng như những từ đi với nó sau từ "và" hay "hoặc".

Tuy nhiên trong nhiều trường hợp số hàng của phần mềm Trình bày Từ trong Ngữ cảnh quá lớn, trên 200 hàng, thì việc quan sát bằng mắt rất khó, ngoài ra khả năng nhớ của con người có giới hạn, không thể nhớ hết những gì quan sát suốt 2, 300 hàng. Do đó chúng ta cần có phần mềm có thể đúc kết tất cả các tính năng của một từ và trình bày ngắn gọn trong một trang (hay hai trang). Phần mềm đó là Sketch Engine.

Chúng tôi sẽ trình bày việc dùng Sketch Engine để nói lên sự khác nhau giữa hai từ "đàn bà" và "phụ nữ".

Chúng ta còn nhớ, lúc sinh thời nhạc sĩ Phạm Duy đã không chối cải sự gian díu tình ái với nhiều người nữ và đã kết luận bằng một câu: "Chỉ tại đàn bà". Như vậy "đàn bà" chỉ sự xấu xa, tội lỗi.

Tuy nhiên khi đề cập đến sự cao quý, trang trọng của người nữ, các tác giả dùng từ "phụ nữ".

Đối với phần đông chúng ta, sự phân biệt này xảy ra tức thì do trực giác. Nhưng đối với học sinh, đặc biệt là người nước ngoài học tiếng Việt thì khó có thể dùng đúng cách từ "đàn bà" và "phụ nữ".

Sketch Engine có khả năng trình bày các tính năng của từng từ, đồng thời liệt kê những tính từ hay nhóm từ nào thường đi với "đàn bà" nhưng không bao giờ được dùng với "phụ nữ". Trong hình ở trang sau, những từ trong các khung màu xanh lá cây thường được dùng với "đàn bà". Trong khi đó các từ hay nhóm từ trong các khung màu đỏ thường được dùng với từ "phụ nữ". Những từ/nhóm từ trong các khung màu trắng có thể dùng được cho cả "phụ nữ" và "đàn bà".

significant cooccurrences of táo:

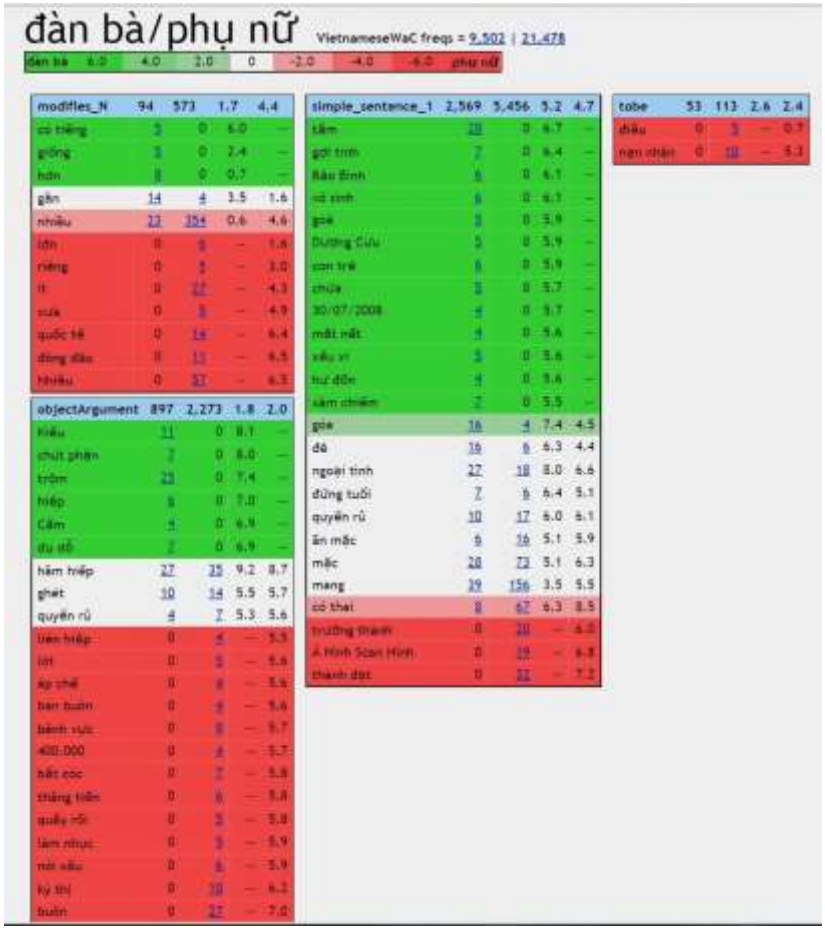
bao (153836), bón (136621), tính (61431.7), quả (18264.3), trọng (15124), ăn (12940.1), xơ (12247.7), Quả (10795.5), tiêu (10622.2), uống (8466.71), trái (8217.56), giảm (8165.96), lẻ (7881.51), . (7533.76), giúp (7223.79), trồng (7039.95), nuốt (6676.88), trị (5423.47), bưng (5006.77), bệnh (4771.84), cam (4753.78), chảy (4609.72), dầm (4533.77), đo (4507.05), Táo (4407.28), cây (4312.87), nhỏ (4295.28), Apple (4143.79), gây (4077.29), tiêu (3937), chuối (3923.83), chất (3904.06), ruu (3896.19), món (3857.39), cà (3815.59), óc (3765.96), khỏe (3606.85), đau (3544.38), ngâm (3534.52), chua (3450.26), chứng (3442.23), loại (3431.84), xé (3217.93), vitamin (3113.68), tây (3085.04), khổ (2914.97), cần (2903.32), ngủ (2881.76), gram (2868.13), vỏ (2767.04), nhuận (2732.38), chữa (2700.23), ép (2698.48), và (2639.44), máu (2580.1), chanh (2553.83), ong (2534.6), đun (2501.13), đe (2477.13), huyết (2467.56), mẫn (2466.23), sâm (2429.65), sữa (2405.06), thuốc (2399.6), xanh (2382.72), bị (2270.88), đầy (2267.51), cướp (2255.38), thể (2237.34), rượu (2216.87), tươi (2180.99), vày (2135.02), 12g (2134.45), sốt (2129.1), miếng (2096.86), ngọt (2095.65), chứng (2052.48), bưởi (2037.96), sống (2032.57), hạt (1975.05)

significant left neighbours of táo:

tính (170589), quả (34086.3), Quả (21680.2), tiêu (16675.4), giảm (14279.9), trái (10407.5), bị (10376.5), dầm (7494.63), chứng (4927.84), . (4633.58), chống (4271.54), xé (3894.48), cây (3864.16), ngâm (3714.37), trồng (3494.82), Giảm (2839.14), cướp (2607.45), nhuận (2586.89), gây (2197.75), vỏ (2055.68), rượu (1991.24), trang (1964.36), ăn (1933.31), ép (1913.25), bệnh (1834.98), khổ (1814.85), trị (1777.53), trành (1585.65), Dầm (1581.93), đại (1574.72), đầy (1304.86), chữa (1286.93), Cây (1108.34), đạnh (1032.75), miếng (953.24), Tính (907.58), Ăn (895.26), tiền (830.96), nuốt (782.17), Trái (775.47), vườn (772.51), sự (754.51), bổng (749.96), Chữa (741.26), như (733.07), Đại (727.25), cut-out (719.72), và (677.04), Toàn (668.73), uống (626.88), mứt (568.06), bánh (548.32), Khám (546.27), toàn (538.11), Cần (529.69), Vỏ (514.93), rất (495.63), cặp (494.25), lát (486.29), ngực (463.27), Đoàn (458.72), Tẩu (439.59), cắt (437.42), hỏ (432.11), trộn (431.15), khoét (415.22), hoặc (413.37), Cầu (397.43), hạt (382.99), gọt (382.72), Khôn (361.47), hái (357.05), phiến (341.53), hương (330.71), Chẩn (304.58), hoạch (292.91), Sự (264.39), neo (257.7), Nước (252.5), Trị (250.83)

significant right neighbours of táo:

bao (265877), bón (241812), tợn (25296.4), . (6983.95), méo (6711.94), đỏ (6087.38), tàu (4999.47), cần (4635.22), khuyết (3940.47), đế (3461.47), và (3370.08), " (2744.85), xanh (2588.85), bón… (2538.43), hoa (2461.23), bón? (2057.55), " (1543.33), bón. (1398.99), tây (1387.87), gai (1222.77), Fuji (1022.53), chín (876.43), tàu (837.24), ép (778.6), bạo… (714.17), nước (651.03), bạo. (645.21), thịp (613.67), bạo! (579.8), rượi (576.56), quân (519.03), bạo? (500.43), chứa (428.66), ty (415.1), rơi (385.8), mẫu (376), 12g (358.36), bónThúc (350.16), chua (340.81), rụng (290.32), hàn (268.59), Ambrosia (221.06), (207.6), khỏ (206.02), gọt (183.73), Tàu (180.33), bón! (177.78), hước (170.7), xen (167.85), thối (166.1), nghiền (164.79), thơm (163.9), giúp (162.79), bón… (148.79), 10g (141.5), tàu. (138.68), còn (138.02), Méo (132.41), Envy (126.47), 15g (123.43), thì (121.24), Trung Quốc (121), nhiệt (120.81), nhân (118), tàu… (117.23), Washington (116.29), 3-4g (115.35), xay (115), méo… (114.54), ngọt (106.43), Gala (104.9), lai (104.06), 10 (103.24), Ninh Thuận (102.99), đỏ… (101.42), Granny (99.13), 20g (97.12), ngon (90.8), vàng (86.07), hay (84.44)



Nghiên cứu tiếng Việt do người nước ngoài thực hiện

Một số nhóm nghiên cứu đã sử dụng những ngữ liệu chứa vài trăm triệu từ (word) tiếng Việt để nghiên cứu và tạo ra những kết quả rất hữu ích, như trình bày sau đây.

A- Nhóm nghiên cứu Wortschatz của Đức

Đối với bất cứ từ (word) nào, Wortschatz cũng có thể tạo ra một bản tóm lược trình bày 3 đặc điểm: những từ/nhóm từ đi cùng với từ quan tâm; những từ/nhóm từ đứng bên phải từ quan tâm, và những từ/nhóm từ đứng bên trái từ quan tâm. Từ đó tạo nên đồ họa giúp người dùng có một ý niệm tổng quát về từ quan tâm. Thí dụ dưới đây liên quan với từ "táo" mà chúng tôi đã có các hình trình bày ở trang 64.

Máy tính không thể suy nghĩ thay người, nhưng có thể dựa trên ngữ liệu để sắp xếp các văn bản theo từng ngữ cảnh (context). Thí dụ có thể cho thấy các nghĩa khác nhau của từ "táo":

1- Từ "táo" trong ý nghĩa "bạo", "tợn" nói về tính khí của con người

2- Từ "táo" liên quan đến "trái cây", hay "bệnh tật",

Ưu điểm của nhóm nghiên cứu Wortschatz là dùng phần mềm tự động để thu thập ngữ liệu của các ngôn ngữ (trên 200 ngôn ngữ) và tự động phân tích không cần phải gắn nhãn từ loại.

Hiện nay đã có nhiều chương trình máy tính tự động nhận diện các văn bản để biết văn bản đó là tiếng gì (Việt, Anh, Pháp, v.v…). Nhóm Wortchatz cũng có một phần mềm khác để tự động phát hiện tính năng của các từ/nhóm từ và xếp các từ/nhóm từ đi cùng thành nhóm.

B- Nhóm nghiên cứu Sketch Engine

Nhóm nghiên cứu Sketch Engine cũng có cùng mục đích như nhóm nghiên cứu Wortschatz nhưng đạt được một kết quả chi tiết và tiên tiến hơn, nhờ dùng phần mềm gắn nhãn từ loại. Như chúng ta đã biết một từ (word) có thể là danh từ, hoặc là động từ, tính từ, tùy theo cách dùng. Thí dụ: từ "bao", có thể là động từ trong nghĩa "bao che", "bao bọc", "bao vây"; nhưng cũng có thể dùng như danh từ "cái bao", như "bao giấy", v.v… Do đó phải gắn cho các từ này một nhãn thích nghi (động từ, danh từ, tính từ) thì việc phân tích mới chính xác. Năm 2006, giáo sư Adam Kilgarriff hoàn thành Sketch Engine và áp dụng phần mềm này để tạo ra các Sketch (bản phát họa), đặc biệt áp dụng cho tiếng Anh, Pháp, Đức, Ý, Trung, Nga.

Sketch – một bản tóm lược hay bản phát họa – là một tài liệu cỡ 1 hay 2 trang giấy liệt kê tất cả các tính năng của một từ (xem chi tiết trên đây của Sketch cho từ "phụ nữ" và từ "đàn bà" sẽ hiểu rõ hơn thế nào là một sketch).

Nhìn vào Sketch Engine, chúng ta thấy ngay sự khác biệt của hai từ và những tính chất đặc thù của từng từ, cả những từ thường dùng đến những từ không nên dùng cho từng trường hợp.

Lúc đó Sketch Engine chỉ mới được dùng cho vài ngôn ngữ chính, như: Anh, Pháp, Nga, Trung, Ý, Tây Ban Nha. Tôi rất muốn dùng Sketch Engine cho tiếng Việt, nhưng chúng ta chưa có bộ gắn nhãn hoàn hão. Tôi gởi cho giáo sư Adam Kilgarriff một tài liệu chứa 10,000 cặp câu Anh-Việt, để yêu cầu ông thử xem tiếng Việt có tương thích với phần mềm Sketch Engine không.

Kết quả: Sketch Engine trình bày chữ Việt rất tốt và tài liệu song ngữ Anh Việt này vẫn còn lưu trên máy chủ của Sketch Engine cho những nghiên cứu về Ngữ liệu Song ngữ.

Từ năm 2000 đã có nhiều nhóm nghiên cứu tại Việt Nam tìm cách viết phần mềm để gắn/gán nhãn tiếng Việt (nhóm của giáo sư Phạm Thị Tươi, VietLex). Năm 2005 tiến sĩ Nguyễn thị Minh Huyền trình luận án Tiến sĩ tại Loria với đề tài Phần Mềm Gắn Nhãn cho Tiếng Việt. Sau đó vài năm, nhiều nhà nghiên cứu ngoại quốc có dùng thử phần mềm của tiến sĩ Nguyễn thị Minh Huyền, nhưng kết quả không được như ý. Phải đợi đến năm 2010, khi Lê Hồng Phương tiếp tục chương trình của tiến sĩ Nguyễn thị Minh Huyền và đã hoàn tất văn bằng tiến sĩ với cùng chủ đề tại Loria.

Trước đó giáo sư Kilgarriff có đề nghị tôi hợp tác để tạo ngữ liệu tiếng Việt. Do bận rộn công việc dịch thuật và phải hoàn tất kế hoạch 1-triệu cặp câu song ngữ, và ở quá xa nhau không thể làm việc hữu hiệu, nên tôi đã từ chối. Tôi mất đi một cơ hội quý giá để tìm hiểu chi tiết về lãnh vực này. Sau đó giáo sư Kilgarriff hợp tác với một người Việt tại Anh để thực hiện một ngữ liệu tiếng Việt trên 200 triệu từ nhờ phần mềm thu thập tự động tài liệu lấy từ Internet.

Khi giáo sư Kilgarriff hoàn tất ngữ liệu tiếng Việt thì cũng là lúc tiến sĩ Lê Hồng Phương hoàn tất chương trình gắn nhãn với nhiều ưu điểm hơn. Cả hai, giáo sư Kilgarriff và tiến sĩ Lê Hồng Phương đều gặp may mắn đã hợp tác đúng lúc để sử dụng phần mềm Sketch Engine và bộ Gắn Nhãn tiếng Việt để tạo ra các Sketch cho tiếng Việt.

Đây là một may mắn cho tiếng Việt vì chúng ta đã có đủ phần mềm cần thiết cho Sketch Engine xử lý tiếng Việt một cách rộng rãi hơn. Tôi đang hợp tác với nhóm của giáo sư Adam Kilgarriff trong việc trích các nhóm từ đồng hiện từ các ngữ liệu song ngữ (Parallel Corpora). Hiện nay kỹ thuật này đã được dùng cho các ngữ liệu song ngữ, như Anh-Pháp, Anh-Đức, Anh-Trung, v.v…

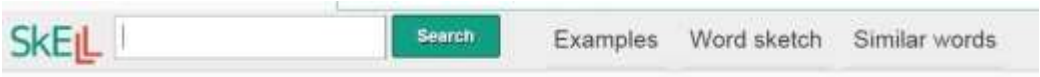
Có hay không có sự kỳ thị trong Dịch máy?

Dịch máy là dịch tự động bằng máy tính. Hiện nay chúng ta có những trang Web dịch tự động, như: Google Translate, Bing (của Microsoft) và nhiều phần mềm miễn phí khác. Phiên dịch tự động chưa hoàn hảo, nên phải hiệu đính. Tuy nhiên tính chính xác của bản dịch cũng tùy thuộc phần nào vào bản chất của văn bản gốc. Thí dụ một văn bản tiếng Anh dùng những từ thông dụng và cách cấu trúc đơn giản thì bản dịch sang Việt ngữ sẽ ít có lỗi, như trường hợp từ điển 120,000 từ do chúng tôi thực hiện và được trình bày ở trang 69. Nếu văn bản gốc dùng nhiều "tiếng lóng", "sáo ngữ", "thành ngữ", "đặc ngữ" thì bản dịch sẽ không thể nào đạt được mức chính xác cao. Tuy nhiên chúng ta không thể thực hiện văn bản tiếng Anh mà không dùng những nhóm từ thể hiện tính chất phong phú của ngôn ngữ khiến người đọc mất hứng thú khi tham khảo.

Điều này có thể thấy rõ khi đọc các bản tin Anh ngữ do các trang Web trong nước thực hiện. Nếu chúng ta dùng Google Translate để dịch những trang Web này sang tiếng Việt, chúng ta sẽ có những bản tiếng Việt gần như chính xác 100%.

ELL: Sketch Engine for Language Learning

Nhóm Sketch Engine đã thực hiện SkELL (phiên bản thu nhỏ của Sketch Engine) để cho mọi người dùng miễn phí một số tính năng của Sketch Engine. Dưới đây là hình ảnh trình bày cách dùng SkELL để tìm hiểu về từ "teacher"



Người dùng bắt đầu điền từ "teacher" vào ô trống, kế đó kích vào Search. SkELL, tự động tô màu phần Example và trình bày màn hình dưới đây.

Từ màn hình trên người dùng có thể chọn 6 tính năng của Word Sketch, như:

Verbs with teacher as object, (Động từ dùng với từ teacher, trong đó "teacher" là bổ ngữ)
Verbs with teacher as subbject, (Động từ dùng với từ teacher, trong đó "teacher" là chủ ngữ)
Adjective with teacher, (Những tính từ dùng với từ "teacher")
Modifiers of teacher, (Những từ bổ nghĩa cho từ "teacher")
Nouns modified by teacher, (Những danh từ được từ "teacher" bổ nghĩa)
Words and/or teached (Những từ dùng với từ "teacher" theo sau từ "and" hay từ "or")

Người dùng chỉ cần kích chuột vào bất cứ từ có gạch dưới để được đưa sang một màn hình mới.

Thí dụ về phân tích Ngữ liệu tiếng Việt: Truyện Kiều

Dưới đây là chi tiết xử lý Truyện Kiều với phần mềm trực tuyến miễn phí Text Analyzer của Online-Utility.org

Number of characters (including spaces) :	105074
Number of characters (without spaces) :	75628
Number of words :	22155
Lexical Density :	10.7154
Number of sentences :	1640
Number of syllables :	24029

Nếu đếm từ đầu đến cuối, truyện Kiều gồm có 22155 từ (word). Số này được gọi là Token. Token của Truyện Kiều là 22155.

Tuy nhiên khi đếm như vậy, chúng ta thấy nhiều từ xuất hiện nhiều hơn một (1) lần. Nếu chỉ tính một (1) lần thôi cho bất kỳ từ nào, dù xuất hiện một (1) hay nhiều lần, thì con số này được gọi là Type. Như vậy Type của Truyện Kiều là 2361. Đối với truyện Kiều: R = Type/Token = 10,6.

Tỷ số Type/Token có một ý nghĩa đặc biệt, cho biết cách dùng từ của tác giả hay thể loại tài liệu.

Phrases with 2 words	Occurencies
nàng rằng	32
một lời	24
bây giờ	23
một mình	20
nàng mới	19
tiểu thư	19
một nhà	18
làm chi	17
giác duyên	16
vội vàng	16
đoạn trường	15
bấy lâu	14
biết đâu	14
cùng nhau	14
nghe lời	13
một ngày	13
cũng là	12
hồng nhan	12
thế này	12
bao giờ	12

Đối với tác giả, nếu R lớn có nghĩa là tác giả dùng từ một cách phong phú; nếu R nhỏ, tác giả dùng đi dùng lại một số từ.

Đối với thể loạ i: ngữ liệu văn học nghệ thuật có R nhỏ; trái lại tiểu thuyết có R lớn (vì tác giả dùng nhiều tính từ, nhiều tên nhân vật, nhiều địa danh để trình bày sự kiện một cách văn hoa). Văn bản học thuật chứa ít tỉnh từ hơn nên thường có tỷ số R thấp.

Trong Truyện Kiều: Cặp hai từ: nàng rằng 32; một lời 24; bây giờ 23; một mình 20; nàng mới 19; tiểu thư 19; một nhà 18; làm chi 17; giác duyên 16; vội vàng 16; đoạn trường 15; bấy lâu 14; biết đâu 14; cùng nhau 14; nghe lời 13; một ngày 13; cũng là 12; hồng nhan 12; thế này 12; bao giờ 12; thì thôi 12; làm sao 12; làm cho 12; thong dong 12; những là 12; rõ ràng 11; phong trần 11; chút phận 11; thì cũng 11; thế nào 11.

Một số kết quả khác về Xử lý Truyện Kiều.

Một trăm năm mươi từ xuất hiện nhiều lần (tổng cộng 10190).

Một trăm bảy mươi lăm từ xuất hiện nhiều lần (tổng cộng 10893).

Hai trăm từ xuất hiện nhiều lần (tổng cộng 11529), tức 52%; trong khi 2161 từ còn lại chiếm 48%.

Tính năng thứ ba là chứng minh ý nghĩa của định luật Zipf

Một trăm từ xuất hiện nhiều lần (tổng cộng 8489): một 308; đã 253; người 219; nàng 197; cũng 167; là 166; cho 166; lời 165; lòng 161; có 157; rằng 152; lại 148; ra 147; hoa 131; tình 124; còn 118; mới 116; đâu 114; ai 113; chẳng 109; mình 108; thì 107; mà 106; biết 106; này 106; trong 105; đến 102; đường 99; nhà 95; càng 93; nào 93; thân 92; trời 92; ngày 88; khi 87; như 87; sao 84; vào 84; mặt 82; duyên 81; xa 78; vàng 77' về 75; sinh 75; tay 75; thôi 74; làm 73; trước 73; thấy 72; chàng 72; nghe 71; chi 71; những 70; sau 69; nỗi 68; hai 66; nước 64; từ 63; hương 60; nói 60; ta 57; thương 57; con 57; phải 57; hồng 56; trông 56; chưa 55; thế 55; tơ 54; gió 54; xuân 54; ở 53; tiếng 53; mấy 52; nghĩ 51; năm 51; xưa 50; đây 49; giờ 49; nhau 48; với 48; gì 47; đi 47; hay 47; công 47; chút 46; bao 46; ấy 46; bên 45; mây 45; phận 45; trăng 43; được 43; ngoài 43; điều 42; hỏi 42; đầu 42; riêng 41; bóng 41; cửa 41.

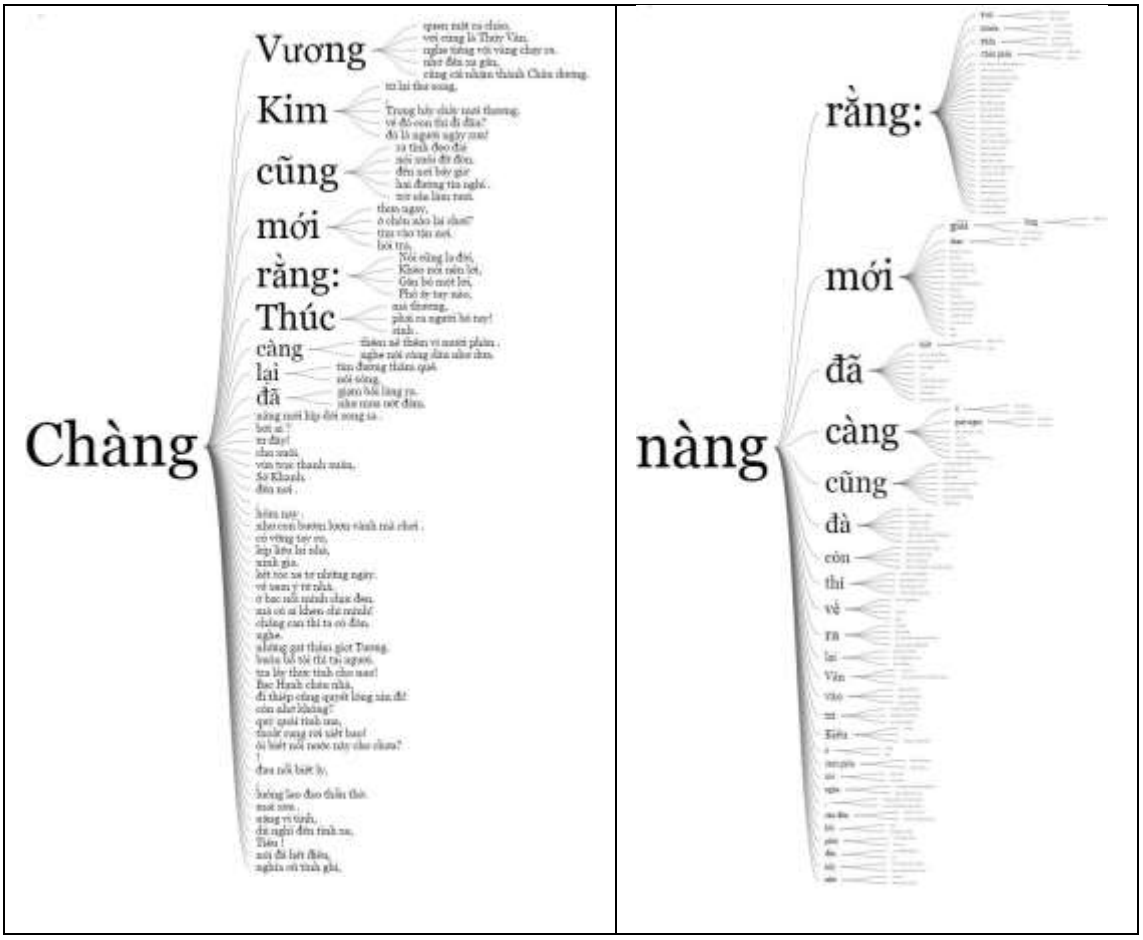
Trong danh sách này chúng ta thấy có nhiều hư từ chiếm những vị trí trên cùng. Từ càng xuất hiện nhiều lần thì cỡ càng lớn. Những từ cực nhỏ chỉ xuất hiện 1 hay 2 lần trong toàn truyện Kiều.

Nhờ phần mềm đồ họa chúng ta cũng có thể trình bày Truyện Kiều theo từng câu một với nhiều màu sắc như trong hình dưới đây.

Sau khi loại những hư từ (thì, mà, này, đến, càng, nào, khi, như, vào, xa, về, thôi, trước, chi, những, sau, nỗi, từ, phải, chưa, thế, mấy, đây, nhau, với, gì, đi, hay, chút, bao, ấy, bên, được, ngoài), chúng ta có hình mới ở trang 86. Trong hình mới này chúng ta thấy hầu hết các từ đều có kích cỡ gần bằng nhau và số lượng từ cũng nhiều hơn so với hình cũ ở trang 85.

Chúng ta cũng có thể dùng phần mềm thích nghi để trình bày các nhóm từ đứng sau từ quan tâm.

Hai hình dưới đây trình bày những nhóm từ đứng sau “Chàng” và đứng sau “Nàng”.



Như vậy chúng ta thấy Ngôn ngữ học Ngữ liệu đã giúp tìm hiểu nhiều tính chất của ngôn ngữ mà Ngôn ngữ học Cổ điển không làm được. Chúng tôi sẽ đi vào chi tiết sử dụng Ngữ liệu Song ngữ 1-triệu cặp câu và những phần mềm miễn phí và "Tự chế" (do chính nhóm chúng tôi thực hiện)

A) **Tạo Từ điển cho Học sinh/Sinh viên**: Học sinh, sinh viên cần từ điển để biết nghĩa của từ quan tâm, đồng thời cũng cần có những thí dụ về cách dùng các từ. Dựa trên phần Anh Ngữ của Ngữ liệu Song ngữ, chúng tôi tạo ra từ điển Anh ngữ. Sau đó dùng Google Translate để dịch sang tiếng Việt. Dĩ nhiên Google Translate không thể dịch chính xác 100%, nhưng nhờ dùng những từ anh ngữ đơn giản để tạo từ điển, nên không phải mất nhiều thời gian để hiệu đính phần tiếng Việt. Hiện nay chúng tôi có một từ điển Anh Việt chứa 120,000 từ. Với cùng phương pháp, chúng tôi có thể tạo ra những từ điển song ngữ tương tự Pháp-Việt, Trung-Việt, Hàn-Việt

apron tạp dề

apron (noun) - a garment of cloth or leather or plastic that is tied about the waist and worn to protect your clothing tạp dề (danh từ) - hàng may mặc vải hoặc da thuộc hoặc bằng nhựa được gắn liền với thắt lưng và đeo để bảo vệ quần áo của bạn

apron (noun) - (golf) the part of the fairway leading onto the green tạp dề (danh từ) - (golf) là một phần của fairway dẫn lên màu xanh lá cây

proscenium , apron , forestage (noun) - the part of a modern theater stage between the curtain and the orchestra (i.e., in front of the curtain) phía ngoài màn, tạp dề, forestage (danh từ) - một phần của một sân khấu nhà hát hiện đại, giữa bức màn và dàn nhạc (tức là, ở phía trước của bức màn)

apron (noun) - a paved surface where aircraft stand while not being used tạp dề (danh từ) - một bề mặt lát nơi máy bay đứng trong khi không được sử dụng

Hình trình bày từ "apron" của Từ điển Anh-Việt chứa trên 120,000 từ. Đây là kết quả do Google Translate thực hiện. Phần Anh ngữ (màu xanh dương) hoàn toàn chính xác, phần Việt ngữ (màu đen) cần hiệu đính, nhưng không phải mất nhiều thời gian.

B) **Khám phá cách dùng từ với Many Eyes**: Chúng tôi cung cấp những phương tiện học tập đặc biệt và miễn phí cho học viên ngôn ngữ. Đăc biệt chú trọng đến những ngôn ngữ mà hiện nay chưa có phương tiện hữu hiệu để giảng dạy, thí dụ tiếng Á-Rập (Arabic). Dựa trên Ngữ liệu Song ngữ Việt-Á-rập, chúng tôi trình bày cho học viên thấy cách thể hiện cũng như ngữ ngĩa của từng từ tiếng Việt được dịch như thế nào sang tiếng Á-Rập.

Hình trên có hai khung. Khung bên trái cho thấy các từ “đồng thời”, “đồng tiền”, “đồng minh”, “đồng ý”, “đồng nghĩa”, “đồng thuận” được dịch sang tiếng Á-Rập như thế nào.

Khung bên phải cho phép học viên tương tác với phần mềm: Click vào từ quan tâm sẽ có khung hình mới liên quan với từ quan tâm.

trị nội bộ và tình hình nhân khẩu học của quốc gia đó, đồng thời phải giải quyết các vấn đề an ninh của tất cả các nhóm bị ảnh hưởng một cách công bằng và bình đẳng.
ولكن إذا كانت القوات الأجنبية راغبة في المشاركة البنّاءة فلابد أن تتفهم السياسات والسمات الديموغرافية الداخلية للبلاد وأن تعالج المخاوف الأمنية ، لدى الجماعات المتضررة على قدم المساواة وبنزاهة. Nếu không làm như vậy, tất cả sẽ trở nên yếu đuối và dễ bị tổn thương. ومن الواضح أن الفشل في تحقيق هذه الغاية من شأنه أن يجعل الجميع ضعفاء وعُرضة للخطر Mặt tốt – mặt xấu của bất bình đẳng
التفاوت الطيب والتفاوت الخبيث Trong ngôi đền của các học thuyết kinh tế, nguyên tắc đánh đổi giữa sự bình đẳng và tính hiệu quả luôn chiếm một vị trí cao.
بريـنستون ـــ في معبد النظريات الاقتصادية، تعودت المقايضة بين المساواة والكفاءة على احتلال مكانة رفيعة سامية.
Nhà kinh tế học người Mỹ Arthur Okun, tác giả của cuốn sách kinh điển về chủ đề này có tên Equality and Efficiency: The Big Tradeoff (Bình đẳng và Hiệu quả: Một sự đánh đổi lớn), tin rằng các chính sách công chỉ quản

Chúng tôi dùng phương tiện thích nghi để tạo những ngữ liệu song ngữ do các thông dịch viên thực hiện. Ngoài ra chúng tôi cũng dùng phần mềm thích nghi để học viên có thể hiển thị được tiếng Á-Rập ngay trên máy tính của họ mà không phải cài đặt bất cứ Tiện ích hay Phần mềm bổ sung.

Hình bên cạnh liệt kê các cặp câu song ngữ với mức độ phiên dịch từng câu, để trình bày nghĩa của từ quan tâm trong ngữ cảnh (word in context). Khi học viên muốn biết cách dùng từ quan tâm, chỉ cần click chuột vào từ đó, phần mềm sẽ hiển thị một khung hình (khung bên trái) mới liên quan đến từ quan tâm.

Do khuôn khổ giới hạn, chúng tôi chỉ trình bày thí dụ cho Ngữ liệu Song ngữ Việt-Á-rập. Chúng tôi cũng có những Ngữ liệu Song ngữ cho các ngôn ngữ khác, như Việt-Trung, Việt-Nga, Việt-Pháp, Việt-Anh, Việt-Đức, Việt-Tây Ban Nha, Việt-Bồ Đào Nha. Các Ngữ liệu Song ngữ cho các cặp ngôn ngữ khác có thể được thực hiện nếu có yêu cầu.

Những lãnh vực chúng tôi quan tâm hàng đầu: Tin tức Thời sự, Nghiên cứu Biển Đông, Bảo vệ Sức khỏe, Hòa nhịp Cuộc sống Địa phương, Tài liệu Học thuật, Học đường và Luật pháp. Đây là những lãnh vực có nhiều tài liệu được phiên dịch sang Việt ngữ.

C) **Tạo Bộ nhớ Phiên dịch từ Ngữ liệu Song ngữ**: Có thể định nghĩa một cách đơn giản: Bộ nhớ Phiên dịch là những ngữ liệu song ngữ được dùng kèm với Phần mền Phiên/Thông dịch (PMTPD). Chúng được lưu trong máy tính và dính liền với PMTPD.

Mỗi khi dịch, các câu cần dịch được đưa vào ô ngôn ngữ nguồn. PMTPD sẽ tìm những câu đã dịch sẵn trong Bộ nhớ Phiên dịch và điền vào ô ngôn ngữ đích. Tùy theo mức độ chính xác, người phụ trách sẽ hiệu đính và sử dụng bản dịch hoàn tất. Tài liệu dịch hoàn chỉnh được lưu trong Bộ nhớ Phiên dịch để dùng cho những lần dịch sau. Dĩ nhiên Bộ nhớ Phiên dịch (BNPTD) không thể đáp ứng tất cả mọi trường hợp, nên việc hiệu đính là bước quan trọng, không thể thiếu. Để tăng hiệu quả cho công tác phiên dịch, người dùng có thể tạo những Bộ nhớ Phiên dịch cho từng lãnh vực, thí dụ BNPTD về Y tế, Luật Pháp, Học thuật. Có khi phải tạo những BNPTD đặc thù cho từng công ty lớn, thí dụ chúng tôi có BNPTD cho công ty Wells Fargo, Humana, v.v…

D) **Trích các nhóm từ từ Ngữ liệu Song ngữ**: Từ những Ngữ liệu Song ngữ đặc biệt theo ngành, chúng tôi có thể trích ra những thuật ngữ hay những nhóm từ đặc thù cho từng lãnh vực. Thí dụ, về Đạo đức trong Kinh doanh (Code of Conduct in Bussiness), chúng tôi có tài liệu từ hơn 100 công ty có COC đã được dịch sang tiếng Việt, gồm hơn 40,000 câu tương đương với 20 cuốn sách cỡ 400 trang). Với số lượng câu này chúng tôi thực hiện sách Thuật ngữ chuyên dụng về Đạo Đức trong Kinh doanh. Đây là một dự án quan trọng mà chúng tôi đang cố gắng thực hiện với nhóm Sketch Engine. Hiện nay nhóm này đã thành công trích nhóm từ thường đi chung với nhau (co-occurrence) cho các ngôn ngữ gốc La-tin. Chúng tôi muốn áp dụng cho ngữ liệu song ngữ Việt-Anh. Nếu thành công sẽ tiếp tục với các ngữ liệu song ngữ Việt-X (X có thể Hàn, Trung, Pháp, Ý, Tây Ban Nha)

Tóm lại lãnh vực Ngôn ngữ học Ngữ liệu tuy chỉ mới hình thành trong vòng 30 năm qua, nhưng nhờ những tiến bộ về phần cứng và phần mềm máy tính nên cộng đồng các nhà nghiên cứu Ngôn ngữ học đã đóng góp nhiều thành quả đáng kể. Chúng tôi có phương tiện để tiến hành một số công tác cho tiếng Việt và cần người hợp tác để triển khai những điều vừa trình bày trên.