CÔNG TRÌNH
XÂY DỰNG LUẬN CỨ KHOA HỌC CHO VIỆC
"Biên soạn bộ từ điển tiếng Việt cỡ lớn"

MẠNG NGỮ LIỆU LONGMAN

Tài liệu dịch của Phòng Từ điển học
Hà Nội - 2001

MẠNG NGỮ LIỆU LONGMAN LÀ GÌ?

Vào thế kỉ 18, TS Samuel Johnson đã mất nhiều năm để biên soạn cuốn "Từ điển tiếng Anh". Từ điển của ông được dựa trên những đoạn trích của nhiều tác giả nổi tiếng, và những đoạn trích này được sao chép bằng tay trên những mảnh giấy để trở thành một phần nhỏ của một hệ thống gọt giũa vĩ đại. Sau hơn 200 năm, những điều như vậy đã trở nên dễ dàng hơn đối với các nhà từ điển học và những người cầm bút khác. Công nghệ máy tính đã đem lại sự phong phú của thông tin ngôn ngữ học khá dễ dàng (một động tác bấm phím), vì thế ngày nay chúng ta có những vấn đề khó khăn về ngôn ngữ trên màn hình máy vi tính để ủng hộ và trợ giúp chúng tả trong công việc.

Mạng ngữ liệu Longman là một thứ rất đa dạng, một nhóm các cơ sở dữ liệu có thể tiếp cận từ xa bao gồm hàng triệu từ. Mạng này cung cấp cho các nhà từ điển học và những người viết giáo trình một chiều sâu tri thức mới về từ, cách sử dụng, các khuynh hướng ngôn ngữ và các mô hình ngữ pháp qua công nghệ chính xác trong giây phút. Năm cơ sở dữ liệu ngôn ngữ mang tính phức tạp cao tạo thành tâm của Mạng: Ngữ liệu Longman / Lancaster với hơn 30 triệu từ bao trùm một lượng lớn các văn bản viết từ văn học đến các bảng thời gian biểu của xe buýt; Khối ngữ liệu người học Longman là khối ngữ liệu thu thập và giám sát những sản phẩm viết của sinh viên tiếng Anh và cho phép chúng ta xác định một cách chính xác những điều người học cần; Ngữ liệu văn bản viết Mĩ Longman bao gồm 100 triệu từ của các văn bản báo và sách Mĩ; Ngữ liệu văn bản nói Mĩ Longman là một tài nguyên duy nhất gồm 5 triệu từ trong ngôn ngữ nói hàng ngày của Mĩ; và Khối ngữ liệu văn bản nói Anh cung cấp cho ta những thông tin khách quan về tiếng Anh nói đầu tiên nó thực sự thế nào và làm sao nó lại khác với tiếng Anh-Anh viết. Năm sức mạnh kết nối này tạo nên Mạng Ngữ liệu Longman và cùng với Khối ngữ liệu quốc gia Anh đem lại sự phong phú về thông tin cho việc viết giáo trình và từ điển mà nó vừa đại diện chính xác cho tiếng Anh vừa thoả mãn nhu cầu của sinh viên ở mọi cấp độ.

Để có các thông tin thêm về cách sử dụng các khối ngữ liệu mà nó đang làm cuộc cách mạng về từ điển xem bài báo của Michael Rundell trong ấn phẩm số một của Phê bình Ngôn ngữ Longman.

* Khối ngữ liệu Longman/ Lancaster
* Khối ngữ liệu người học Longman
* Khối ngữ liệu văn bản nói BNC
* Mẫu của khối ngữ liệu văn bản nói BNC
* Khối ngữ liệu văn bản viết Mĩ Longman
* Khối ngữ liệu văn bản nói Mĩ Longman

Khối ngữ liệu Longman/Lancaster là gì?

Khối ngữ liệu Longman/Lancaster là một khối thu thập ngôn ngữ lớn đã được vi tính hoá tạo nên bởi những nguồn văn bản hiện hành với một phạm vi rộng lớn. Nó gồm 30 triệu từ ngôn ngữ viết được lấy từ văn học, tạp chí, báo và cả những vật liệu nhất thời như tờ rơi và bao bì. Khối ngữ liệu toàn cầu duy nhất được cấu trúc / xây dựng cẩn thận để càng đại diện cho ngôn ngữ viết càng tốt và là một sự phản ánh trung thực tiếng Anh thế kỉ 20, khối ngữ liệu Longman/Lancaster cung cấp cho các nhà từ điển, các chủ biên và các tác giả tất cả các thông tin họ cần biết để viết được những cuốn từ điển và những tài liệu EFL (tiếng Anh như một ngoại ngữ - ND) chất lượng cao nhất.

Nó trợ giúp như thế nào?

Từ có thể được xem trong ngữ liệu qua một chương trình bảng mục từ. Chúng ta gọi các từ mà chúng ta muốn kiểm tra và khối dữ liệu sẽ trình bày tất cả các hiện dạng (occurrence) của từ đó trong ngữ cảnh của ngân hàng dữ liệu 30 triệu từ. Từng ví dụ của từ đó hiện lên trên màn hình và từ hàng loạt ví dụ đó chúng ta có thể suy ra một số lượng lớn thông tin về một từ đang được sử dụng thế nào. Lấy từ **haunt** (thăm viếng (nơi nào), thường lui tới) làm ví dụ (hình 1). Hầu hết mọi người liên tưởng từ này với ma và nghĩa địa và nghĩ về nó như là nghĩa đầu tiên / khởi thuỷ. Tuy nhiên, từ các ví dụ trong khối ngữ liệu Longman/Lancaster, chúng ta có thể dễ dàng thấy rõ rằng trong tiếng Anh đương đại cách sử dụng thành ngữ của từ **haunt** là thông dụng nhất (hình 2). Thông tin này có thể được truyền cho những sinh viên sử dụng tài liệu học và từ điển của chúng tôi.

Hình 1

the theatre, stayed to **haunt** her for the rest of her life.

(cái rạp hát, vẫn còn **ám ảnh** cô ấy cho đến tận cuối đời.)

his assassin he will not **haunt** him. Ah, the past is filled

(kẻ ám sát anh ấy hắn sẽ không **ám ảnh** anh nữa. À, quá khứ đã được lấp đầy)

woman began to **haunt** him, and not only in his dream

(phụ nữ bắt đầu **ám ảnh** anh ta, và không chỉ trong giấc mơ của anh)

the eyebrow continued to **haunt** him, and the Coming, so

(đôi lông mày tiếp tục / vẫn còn **ám ảnh** anh ta, và điều sắp tới, như thế)

anxious fantasies will **haunt** him; but a witch he can push

(những ý nghĩ bồn chồn sẽ **ám ảnh** anh ta; nhưng một mụ phù thuỷ anh ta có thể chen lấn)

cased for a single day to **haunt** him. What they whispered in

(thư giãn một ngày để **thăm viếng** anh ấy. Các họ xì xào về)

vulnerable expression **haunt** his eyes, and wished I hadn't

(vẻ mặt dễ bị tổn thương dã **ám ảnh** đôi mắt anh ta, và cầu mong tôi không)

Hình 2

haunt [T *not in progressive forms*] **1** [*often pass.*] (of spirit, esp. of a dead person) to visit (a place), appearing in a strange form: *The ghost of a headless man haunts the castlea haunted house* **2** [usu. pass] (esp. of something strange or sad) to be always in the thought of (someone): *I was haunted by his last words to me. She had a haunted look, as if she were constantly anxious or afraid.* **3** *infml* to visit (a place) regularly; FREQUENT

haunt [T không ở dạng diễn tiến] **1** [thường ở thể bị động] (thuộc về tinh thần đặc biệt của một người chết) thăm viếng (nơi nào), xuất hiện trong một hình thức lạ kì: *Bóng ma một người đàn ông cụt đầu lảng vảng ở lâu đài / một ngôi nhà bị ma ám* **2** [thông thường ở thể bị động] (đặc biệt một cái gì đó cũ kĩ hoặc buồn bã) luôn ở trong suy nghĩ của (ai đó): *Tôi bị ám ảnh bởi những lời cuối cùng của anh ấy với tôi. Cô ấy có một cái nhìn ám ảnh, như thể cô ấy luôn luôn lo âu hay sợ hãi.* **3** *Không trang trọng* thăm viếng (nơi nào) đều đặn; HAY LUI TỚI

Khối ngữ liệu người học là gì?

Sinh viên và giáo viên trên toàn thế giới gửi những bài luận và những bài thi để giúp chúng tôi tạo khối ngữ liệu người học Longman, một cơ sở dữ liệu vi tính hoá gồm 10 triệu từ làm trọn vẹn thành ngôn ngữ do những sinh viên tiếng Anh viết. Mọi quốc tịch, mọi trình độ ngôn ngữ đều được làm mẫu điển hình trong khối ngữ liệu và điều này đưa lại khả năng thấu hiểu sự thật ẩn giấu bên trong duy nhất cho người học tiếng Anh.

Nó nói với chúng ta điều gì?

Mỗi bài luận của sinh viên được mã hoá bằng quốc tịch và trình độ ngôn ngữ giữa các bài khác (hình 1) - TUR, IN cho sinh viên trung cấp người Thổ Nhĩ Kì, SSP, EL cho sinh viên sơ cấp người Tây Ban Nha (từ Tây Ban Nha trái với từ Nam Mĩ), sau đó được nhập vào máy vi tính để tạo nên từng phần của khối ngữ liệu. Chúng ta có thể tập trung vào một nhóm các sinh viên đã được chọn lựa như là các sinh viên cao cấp người Pháp, từ đó hiểu được những địa hạt khó khăn đặc biệt của nhóm này (hình 2).

Một cách khác, chúng ta có thể tập trung vào một từ hay một đoản ngữ để xem những lỗi của tất cả các sinh viên.

Chúng ta sử dụng thông tin như thế nào?

Khối ngữ liệu người học Longman cung cấp rất nhiều thông tin quí giá về những lỗi mà sinh viên mắc phải và những điều họ đã biết, đó chính là một nguồn tài nguyên hoàn hảo cho các nhà từ điển học và các tác giả viết tài liệu - những người muốn sản xuất ra những cuốn từ điển và giáo trình đúng địa chỉ mà sinh viên đặc biệt cần. Khối ngữ liệu người học Longman đã được dùng để viết Những Ghi chú về Cách sử dụng trong cuốn Từ điển Học tập Tích cực Longman, căn cứ từ khối ngữ liệu cho biết rằng có 100% lỗi trong nghĩa và trong cách dùng của từ cloth (vải (mặc)): **My cloths and shoes were wet** (Vải và giầy của tôi đều ướt), **We have very good cloth stores** (Chúng tôi có những cửa hàng vải rất tốt), v.v. Xác định một cách chính xác vấn đề như vậy, các nhà từ điển học của chúng tôi đã viết một Ghi chú sử dụng phù hợp (hình 3).

Trở thành một phần của khối ngữ liệu người học

Cùng với sự lớn mạnh của khối ngữ liệu người học Longman, chúng tôi vẫn cần nhiều nhiều các bài làm viết tay của sinh viên. Vì thế, hãy dừng vứt các bài kiểm tra sinh viên của bạn cuối các năm học, hãy gửi nó đến cho chúng tôi tại Longman và chúng tôi sẽ gửi một Tiểu Chỉ dẫn mục từ Longman hay một bản Từ điển Longman tiếng Anh đương đại cho thư viện của trường bạn cho 100 bài kiểm tra hoặc hơn. Để biết thêm chi tiết về vấn đề này xin liên hệ Denise Denney.

Hình 1

TUR, IN pap which is in Leyden stowe to having nice

(TUR, In vấn đề tầm thường ở Leyden stowe thật hay ho)

TUR, IN e to havong noce drink there. They had nice

(TUR, IN e để có một bữa uống vui vẻ ở đó. Họ phải dễ chịu)

SSP, EL attic and loft space. The house is so nice

(SSP, EL khoảng trống gác mái và gác xép. Ngôi nhà thật là đẹp)

Hình 2

how the life should be **nice** and beautiful in peace. The

(cuộc sống thú vị và đẹp đẽ nhường nào trong cảnh hoà bình. Cái)

ned there. There was a **nice** and big place that Korosh a

(ở phía Bắc đó. Có một nơi thú vị và to lớn Korosh)

well and which are so **nice** and charmy. We might be a

(và chúng thật thú vị và quyến rũ. Chúng ta có thể)

from school and it is **nice** and clean with a qiut dista

(từ trường học và nó đẹp và sạch với một (khoảng cách?) khá)

your drink. Sunday is a **nice** and easy-going day. There a

(đồ uống của bạn. Chủ nhật là một ngày đẹp và dễ đi. Có một)

people that I've met are **nice** and friendly people and lik

(những người mà tôi đã gặp thật tử tế và thân thiện và)

Hình 3

USAGE Do not use **cloth** or **cloths** to mean "the thinhgs that people wear". Instead use **cloths** *a* **clothes** *shop. The guests all wore casual* **clothes**

CÁCH SỬ DỤNG không dùng *cloth* (vải, giẻ) hay *cloths* (vải, giẻ) để chỉ "thứ gì đó để con người mặc". Thay vào đó phải dùng *clothes* (quần áo) *a clothes shop* (một cửa hàng quần áo). Tất cả khách khứa đều mặc quần áo thường phục.

Khối ngữ liệu văn bản nói BNC là gì?

Là một phần của dự án nghiên cứu hợp tác quan trọng gọi là khối ngữ liệu quốc gia Anh đã thu thập hơn 100 triệu từ của văn bản viết và văn bản nói tiếng Anh, Longman đã phát triển một khối ngữ liệu văn bản nói gồm 10 triệu từ. Khối ngữ liệu văn bản nói bao gồm những đoạn hội thoại tự nhiên và thoải mái như những gì nghe được quanh ta, từ ngôn ngữ của các bài giảng, những buổi họp công việc, những lời nói sau bữa ăn tối và những buổi tán gẫu. Đây là lần đầu tiên tiếng Anh nói được ghi âm theo hệ thống với một tỉ lệ lớn, và đến nay các nhà từ điển học và các nhà ngôn ngữ có cơ hội đầu tiên nghiên cứu tiếng Anh như nó được nói, thứ tiếng Anh mà có thể nghe thấy ở trên đường phố.

Ngôn ngữ đã được thu thập như thế nào?

Một chi nhánh nghiên cứu thị trường độc lập đã được uỷ quyền làm cuộc thu thập xuyên khu vực những người nói tiếng Anh-Anh trong vương quốc Anh. Thu thập những cuộc đàm thoại thoải mái từ mẫu đại diện của quần chúng theo cách nói riêng biệt của tuổi tác, giới tính, nhóm xã hội và địa phương.

Thiết bị âm thanh

Mỗi người được chọn lựa được cấp cho một Walkman nhỏ với một microphone và giao cho đi ghi âm tất cả những lời nói mà anh (chị) ta nghe được hoặc tham gia trong thời gian một tuần. Vì họ được bảo đảm hoàn toàn vô danh, những người tham gia chẳng bao lâu quên đi những thiết bị ghi âm, một thực tế được phản ánh trong nội dung và phong cách của cuộc đàm thoại (thường hoàn toàn là khôi hài). Băng được gửi trở lại cho chúng tôi tại

Longman, thu âm lại và nhập vào máy vi tính. Tổng số những người tham gia là hơn hai nghìn.

Khối ngữ liệu văn bản nói chỉ ra cho chúng ta điều gì?

Sự phân tích thực sự khối ngữ liệu văn bản nói chỉ mới bắt đầu, nhưng trong khoảng thời gian tương đối ngắn này một số sự kiện quan trọng đã được tiết lộ. Cách sử dụng các thành ngữ, chẳng hạn, nó phổ biến trong ngôn ngữ nói hơn là trong ngôn ngữ viết. Một thành ngữ như **flash in the pan** (thành công trong chốc lát) đã xuất hiện vài lần trong khối ngữ liệu văn bản nói, nhưng không xuất hiện lần nào trong khối ngữ liệu 30 triệu từ Longman/Lancaster.

Cái độc đáo của kho ngữ liệu văn bản nói là nó chỉ ra cho chúng ta biết thực sự chúng ta sử dụng tiếng Anh thế nào, không phải là chúng ta được chỉ ra cách sử dụng tiếng Anh hay chúng ta sử dụng nó khi viết thế nào. Nó chỉ ra rằng từ trong lời nói rất khác biệt với từ trong văn viết như thế nào. Cuối cùng, sinh viên sẽ có thể học tiếng Anh trong một phạm vi thú vị mới của các tài liệu kiểm tra tiếng Anh (ELT) mà nó đại diện cho tiếng Anh như thực sự nó tồn tại. Lần đầu tiên ngôn ngữ nói *thực sự* đã tác động tới sự sáng tạo từ điển người học. Cuốn (The Longman Dictionary Of Contemporary English) Từ điển tiếng Anh đương đại Longman là cuốn từ điển duy nhất nhận ra tầm quan trọng của tiếng Anh nói, đưa ra các từ và đoản ngữ được sử dụng giao tiếp tự nhiên trong tiếng Anh nói.

Dự án Khối ngữ liệu quốc gia Anh là giai đoạn hợp tác ban đầu được thực hiện bởi Oxford University Press, Longman Group UK Ltd, Chambers, Lancaster University's Unit for Computer Research in the English Language (Tổ chức nghiên cứu tiếng Anh bằng máy vi tính của trường đại học Lancaster) Oxford University Computing Services and the British Library (Dịch vụ máy tính trường đại học Oxford và Thư viện Anh). Dự án đã nhận được tiền của từ Phòng thương mại và công nghiệp Anh quốc và Hội đồng nghiên cứu khoa học và kĩ thuật trong khung làm việc chung về Công nghệ Thông tin.

Trích trong thành phần văn bản nói của khối ngữ liệu quốc gia Anh.

Cuộc hội thoại trong phòng khách giữa hai đôi nói về chương trình vô tuyến ngày nghỉ trong vòng 60 giây.

Nhấn vào đây để nghe đoạn văn bản nói mẫu.

Speaker 1: We've just been watching [Saga].

Người nói 1: Chúng tôi mới xem [Saga - (truyện dân gian kể về các anh hùng)].

Speaker 2: [Saga].

Người nói 2: [Saga].

Speaker 1: on the holiday [programme].

Người nói 1: Vào ngày nghỉ [chương trình].

Speaker 3: [oh really, I videoet it].

Người nói 3: [Ồ thế à, Tôi đã thu nó đấy].

Speaker 1: Twenty-eight days in Fuengirola for er [two hundred and ninety-eight pounds].

Người nói 1: Hai mươi tám ngày ở Fuengirola với ừm [hai trăm chín tám bảng].

Speaker 2: [[unclear]] works out at twelve pound or [eighteen pound a]

Người nói 2: [[không rõ]] tính ra vào mười hai bảng hay [mười tấm bảng một]

Speaker 1: eighteen pound a day] - that [includes your flight and everything].

Người nói 1: mười tám bảng một ngày] - [bao gồm chuyến bay của bạn và mọi thứ].

Speaker 4: [with Saga]

Người nói 4: [cùng với Saga]

Speaker 2: [everything].

Người nói 2: [mọi thứ].

Speaker 3: Isn't it a shame that that nice girl's coming off it what's her name?

Người nói 3: Thật đáng xấu hổ, cái cô gái đó tên là gì í nhỉ?

Speaker 1: Anne Gregg.

Người nói 1: Anne Gregg.

Speaker 3: [Yes].

Người nói 3: [Ừ, đúng rồi].

Speaker 2: [[unclear] to them] she's lovely.

Người nói 2: [[không rõ] với họ] cô ấy thật đáng yêu.

Speaker 3: Stupid, [I mean she's so attractive].

Người nói 3: Ngốc, [Tôi muốn nói cô ấy quá quyến rũ].

Speaker 1: [[unclear] got that David Frost].

Người nói 1: [[không rõ] với David Frost].

Speaker 2: And that other bloke who's on there, Robinson.

Người nói 2: Và với một gã khác ở đó, Robinson.

Speaker 3: [I shall write up and complain].

Người nói 3: [Tôi sẽ ghi chép đầy đủ và phàn nàn].

Speaker 1: [Yeah Robert Robinson,] he's gone to [Hong Kong].

Người nói 1: [Ừ Robert Robinson,] hắn đã chuồn sang [Hong Kong].

Speaker 2: [She is so lovely]

Người nói 2: [Cô ấy thật đáng yêu]

Speaker 3: [I can't] stand Robert Robinson, but [Anne Gregg is so attractive].

Người nói 3: [Tôi không thể] chịu được Robert Robinson, nhưng [Anne Gregg thì thật hấp dẫn].

Speaker 4: That's right.

Người nói 4: Đúng đấy.

Khối ngữ liệu văn bản viết Mĩ Longman là gì?

Khối ngữ liệu văn bản viết Mĩ Longman là một khối ngữ liệu động 100 triệu từ gồm các văn bản nối tiếp từ báo, tạp chí, tiểu thuyết bán chạy nhất, các bài viết khoa học và công nghệ, sách khổ lớn có tranh minh hoạ. Sự kết hợp của khối ngữ liệu đang không ngừng được trau chuốt và tư liệu mới đang được bổ sung vào. Đề cương thiết kế của Khối ngữ liệu văn bản viết Mĩ Longman được dựa trên những nguyên lí thiết kế chung của khối ngữ liệu tiếng Anh Longman Lancaster và thành phần văn bản viết của Khối ngữ liệu quốc gia Anh. Cũng giống như các khối ngữ liệu khác trong Mạng ngữ liệu Longman, từ có thể được lập thành bảng chỉ mục, danh sách từ được tạo ra, và các đặc trưng thống kê được phân tích, cho phép nhà từ điển học so sánh và làm rõ sự khác biệt trong cách sử dụng tiếng Anh - Anh và tiếng Anh - Mĩ.

Khối ngữ liệu văn bản nói Mĩ Longman là gì?

Khối ngữ liệu văn bản nói Mĩ Longman là một sự khởi đầu mới tinh và gồm 5 triệu từ của văn bản. Trường đại học California tại Santa Barbara đã chịu trách nhiệm tập hợp các băng ghi cho Longman. Nó làm mẫu điển hình cho các cuộc hội thoại hàng ngày của hơn 1000 người Mĩ với những nhóm tuổi khác nhau, trình độ học vấn khác nhau, dân tộc khác nhau, và những người nói này từ hơn 30 bang của Mĩ. Những lời nói được ghi âm này sẽ được chuyển thu sang cơ sở dữ liệu của máy vi tính và được các nhà từ điển học của chúng tôi phân tích nhằm xác định tần số sử dụng, các nghĩa chính xác và những đoản ngữ tiêu biểu mà sinh viên cần học.

BNC LÀ GÌ?

Khối ngữ liệu quốc gia Anh là một khối ngữ liệu rất lớn của tiếng Anh hiện đại (hơn 100 triệu từ) cả ngôn ngữ nói và ngôn ngữ viết.

Dự án đã được thực hiện và quản lý bởi liên hiệp các tổ chức công nghệ/hàn lâm do Oxford University Press chỉ đạo, những thành viên khác của liên hiệp là các nhà xuất bản từ điển lớn: Addison - Wesley Longman và Larousse Kingfisher Chambers; các trung tâm nghiên cứu hàn lâm tại Dịch vụ Máy tính trường đại học Oxford (Oxford University Computing Services), Trung tâm nghiên cứu tiếng Anh trên máy tính của trường đại học Lancaster (Centre for Computer Research on the English Language), và Trung tâm nghiên cứu và đổi mới của Thư viện Anh (Research and Innovation Centre). Công việc xây dựng khối ngữ liệu bắt đầu năm 1991, và hoàn thành năm 1994. Dự án đã được tài trợ bởi các thành viên thương mại, Hội đồng Khoa học và kỹ thuật (hiện là EPSRC) và DTI của chương trình Khung hợp tác Công nghệ Thông tin (JFIT). Nguồn cung cấp hỗ trợ được cung cấp bởi Thư viện Anh và Viện hàn lâm Anh.

Khối ngữ liệu được thiết kế để đại diện cho tiếng Anh - Anh hiện đại với phạm vi càng lớn càng tốt. Chẳng hạn, phần văn bản viết (90%) gồm các đoạn trích từ báo địa phương và báo quốc gia, các tạp chí xuất bản định kì chuyên ngành và báo chí cho mọi lứa tuổi và sở thích, các sách học thuật và truyện viễn tưởng phổ biến, các lá thư và hồi ức đã xuất bản và chưa xuất bản, các bài luận của trường phổ thông và đại học, trong số rất nhiều các loại văn bản khác. Phần văn bản nói (10%) gồm một số lượng lớn các đoạn hội thoại không nghi thức chưa được văn tự hoá, được ghi bởi các tình nguyện viên với các độ tuổi khác nhau, các tầng lớp xã hội và tôn giáo khác nhau theo phương pháp cân đối nhân khẩu, cũng như ngôn ngữ nói được thu thập trong tất cả các ngữ cảnh khác nhau, được xếp sắp từ kinh doanh nghi thức hay các cuộc họp chính phủ đến buổi trình diễn trên đài và những chương trình trả lời thắc mắc do khán giả gọi đến hỏi bằng điện thoại.

Nguyên tắc chung này, và cách sử dụng những tiêu chuẩn thoả thuận mang tính quốc tế trong việc mã hoá chúng, đã cổ vũ chúng tôi tin tưởng rằng Khối ngữ liệu sẽ hữu ích cho những mục đích nghiên cứu khác nhau rất rộng lớn, trong các lĩnh vực đặc trưng như là từ điển học, trí tuệ nhân tạo, nhận diện và tổng hợp lời nói, nghiên cứu văn học, và tất cả các bộ môn khác nhau của ngôn ngữ học.

Khối ngữ liệu bao gồm 100.106.008 từ, chiếm khoảng 1,5 Gigabytes khoảng trống của đĩa - tương đương với hơn một nghìn chiếc đĩa mềm có sức chứa cao. Có thể hình dung những con số này như sau, trung bình một quyển sách bìa giấy thường 250 trang dày 1 centimetre; giả sử một trang có 400 từ, chúng ta có thể tính ra rằng toàn bộ khối ngữ liệu được in ra với cỡ chữ nhỏ trên giấy mỏng có thể chiếm tới khoảng mười mét khoảng trống kệ. Đọc to