TRUNG TÂM KHOA HỌC XÃ HỘI VÀ NHÂN VĂN QUỐC GIA
VIỆN NGÔN NGỮ HỌC

BÁO CÁO TỔNG QUAN

XÂY DỰNG LUẬN CỨ KHOA HỌC CHO VIỆC BIÊN SOẠN BỘ TỪ ĐIỂN TIẾNG VIỆT CỠ LỚN

ĐỀ TÀI NGHIÊN CỨU ĐỘC LẬP CẤP NHÀ NƯỚC

HÀ NỘI - 2002

Mục lục

PHẦN THỨ NHẤT

BÁO CÁO TỔNG QUAN

Tên đề tài: **"Xây dựng luận cứ khoa học cho việc biên soạn bộ Từ Điển tiếng Việt cỡ lớn"**.

I. Nhiệm vụ được giao

(thể hiện trong bản "Thuyết minh đề tài" do Viện ngôn ngữ học kí với lãnh đạo Trung tâm Khoa học xã hội và Nhân văn Quốc gia ngày 6-12-1999"):

1. Nội dung công việc:

1.1. Chuẩn bị về lí luận.

Tiến hành nghiên cứu cơ sở lí thuyết để xây dựng bộ maket và 10 bản thử nghiệm với yêu cầu: cụ thể, khoa học, có thể áp dụng được để mở rộng phạm vi, số lượng.

1.2. Chuẩn bị về bảng từ.

Nghiên cứu hệ tiêu chí để xây dựng bảng từ và bảng từ chuẩn thử nghiệm trên ba vần (dạng sơ thảo) với yêu cầu: đầy đủ, khoa học, có tính hệ thống cao, phản ánh được hệ thống từ vựng đa dạng và phong phú của tiếng Việt, có thể áp dụng để lập bảng từ toàn bộ.

1.3. Xây dựng chương trình thử nghiệm.

Xây dựng hệ chương trình thử nghiệm ngân hàng dữ liệu tiếng Việt với yêu cầu: nhập, cập nhật và khai thác nhanh chóng, thông suốt, có tính hệ thống; khai thác nhanh theo yêu cầu của người lập bảng từ và biên soạn TĐ (chạy thử trên một số dữ liệu).

2. Kinh phí:

800.000.000đ (tám trăm triệu đồng) cho thực hiện đề tài + 100.000.000đ (một trăm triệu đồng) cho đoàn đi khảo sát tại Cộng hoà Pháp.

3. Thời gian:

Đề tài thực hiện trong hai năm, từ tháng 4-2000 đến 4-2002.

II. Các công việc đã hoàn thành

1. Cơ sở lí thuyết để xây dựng bộ maket và các maket đã xây dựng

Hoàn thành tập bài viết (21 bài = 372 trang khổ A4) về các vấn đề cơ bản của từ điển tiếng Việt cỡ lớn là: Cấu trúc vi mô, cấu trúc vĩ mô, vấn đề phương ngữ, từ ngữ lịch sử, xử lí từ đồng nghĩa, ví dụ trong từ điển, xử lí từ ngữ khẩu ngữ, xử lí từ loại, thông tin ngữ pháp, xử lí thành ngữ, xử lí trợ từ, v.v.

- Hoàn thành bộ Maket (hơn 30 mẫu) định nghĩa cho các từ loại lớn là: Tính từ, trợ từ, thành ngữ và từ ngữ khẩu ngữ.

Đánh giá chung: Cơ sở lí thuyết là một trong những vấn đề cơ bản của công trình. Phần nghiên cứu này là một cố gắng lớn của tập thể tác giả. Nó đã tiếp thu được những kết quả mới nhất của từ điển học trên thế giới, đã phản ánh được những thành tựu chung của giới từ điển học Việt Nam cho đến nay. Việc ứng dụng để xây dựng bảng từ cũng như biên soạn cuốn TĐ tiếng Việt cỡ lớn còn tuỳ thuộc nhiều vào sự tiếp thu cũng như quan điểm của chủ biên và Ban biên tập, nhưng những cơ sở lí thuyết được đưa ra đủ mạnh để làm chỗ dựa cho bộ TĐTV cỡ lớn.

2. Hệ tiêu chí để xây dựng bảng từ và bảng từ chuẩn thử nghiệm

- Đưa được hệ tiêu chí để xây dựng bảng từ, có thể áp dụng để xây dựng bảng từ cho TĐTV cỡ lớn.

- Vào máy đầu mục từ của kho phiếu cũ để làm bảng từ thô: 1200 hộp, và xử lí toàn bộ lần thứ nhất, tổng số 118 000 đơn vị.

-Tập hợp được 125.000 đơn vị đầu mục đã được thu thập vào các từ điển giải thích tiếng Việt (trên tổng số 371.000 đầu mục trong 8 quyển từ điển đã nhập).

- Làm bảng từ thử nghiệm 3 vần ABC: Trên cơ sở các nguồn (từ điển đã có, ngân hàng dữ liệu tiếng Việt thử nghiệm, từ mới giai đoạn 2000-2002) đã chọn được 33.994 đơn vị mục từ (dự kiến 20.000 mục).

Đánh giá: Với các bảng từ thô đã tập hợp được từ các nguồn khác nhau, lần đầu tiên chúng ta có thể có toàn bộ kho từ vựng tiếng Việt. Từ đó, dựa vào hệ tiêu chí đã xác định và kho dữ liệu tiếng Việt có thể dựng được bảng từ cho cuốn TĐTV cỡ lớn (thường được đánh giá là một phần quan trọng nhất trong toàn bộ công việc biên soạn mới một cuốn từ điển). Thử lập bảng từ cho ba vần đầu của cuốn TĐTV cỡ lớn để hình dung khối lượng và rút kinh nghiệm.

3. Hệ chương trình thử nghiệm Ngân hàng dữ liệu tiếng Việt:

3.1. Chương trình lưu trữ và khai thác các từ điển tiếng Việt đã có:

Hoàn thành chương trình, nhập vào chương trình 8 cuốn Từ điển tiếng Việt (tổng số: 371.000 mục từ, 10.789 trang hầu hết là khổ lớn, đang nhập tiếp một cuốn 596 trang khổ lớn). Kết quả tốt, đáp ứng được các yêu cầu tra cứu và khai thác đề ra.

Với chương trình này, có thể tra cứu, khai thác được các thông tin theo cả hai chiều dọc và ngang:

* Theo chiều dọc - tra cứu riêng từng cuốn từ điển, có thể:

- Thống kê được danh sách (theo a-b) tất cả các mục từ trong cuốn từ điển đó.

- Tra cứu được từng mục từ với tất cả các thông tin của từ đó kèm theo.

- Thống kê được số lượng và hiển thị danh sách (theo a-b) các từ cùng từ loại (danh từ, động từ, tính từ, v.v.) trong cuốn từ điển đó (nếu từ điển đó có chú giải từ loại)

- Thống kê được số lượng và hiển thị danh sách các từ cùng phong cách (kng., phg., vch., v.v.) trong cuốn từ điển đó (nếu từ điển đó có chú giải phong cách).

- Có phần tìm kiếm để khi đánh vào một từ bất kì, có thể biết được từ đó đã có trong từ điển hay chưa.

* Theo chiều ngang - tra cứu, đối chiếu các từ điển với nhau:

- Thống kê được bảng từ chung của tất cả các từ điển có trong chương trình.

- Tra cứu được từng mục từ với các thông tin của từ đó đồng thời trong nhiều cuốn từ điển cùng một lúc. Công việc này sẽ vô cùng thuận lợi và tiết kiệm thời gian cho người biên soạn và nghiên cứu nói chung.

3.2. Chương trình quản lí và tra cứu ngữ liệu tiếng Việt (ngân hàng ngữ liệu tiếng Việt):

Hoàn thành chương trình, nhập dữ liệu (tổng số 13.241.000 âm tiết, so với kế hoạch đặt ra là 10 triệu âm tiết), quản lí và khai thác tốt, đáp ứng các yêu cầu tra cứu đề ra.

Với chương trình này, ta có thể:

- Bước đầu tìm cách khắc phục "vấn đề ranh giới từ trong tiếng Việt", phản ánh phần nào mô hình kết hợp cú pháp của từ, giúp người biên soạn khai thác tư liệu ngôn ngữ nhanh với số lượng lớn gấp bội.)

- Tìm kiếm được tất cả các ngữ cảnh của một từ bất kì trong tiếng

Việt (trong phạm vi ngữ liệu đã nhập vào chương trình) với các thông tin kèm theo về xuất xứ của ngữ cảnh đó. Các ngữ cảnh này khi hiển thị trên máy sẽ được chương trình tự động sắp xếp theo một trật tự nhất định.

- Tìm kiếm được tất cả các ngữ cảnh của một từ trong một tác phẩm/ của một tác giả/ trong một khoảng thời gian/ thuộc một loại văn bản... (trong phạm vi ngữ liệu đã nhập vào chương trình) với các thông tin về xuất xứ kèm theo.

- Thống kê được và cung cấp được danh sách các âm tiết "mới", các từ "lạ", tức là các âm tiết, các từ chưa có trong danh sách mục từ chuẩn của chương trình.

3.3. Chương trình trợ giúp biên soạn:

Hoàn thành chương trình, đáp ứng được các yêu cầu đề ra.

Với hệ chương trình này, người soạn có thể soạn định nghĩa từ điển trên máy tính. Các thông tin trong một định nghĩa sẽ được đưa vào các trường khác nhau. Cách xử lí chương trình tốt, không hạn chế số lượng với đồng âm, đa nghĩa... Mỗi một trường có định dạng riêng để có thể gọi, đếm, copy và in các thông tin theo các trường đã nhập. Toàn bộ nội dung định nghĩa sau khi hoàn thành có thể chuyển sang Microsoft Word để in ấn thành từ điển theo các quy cách đã định sẵn.

Với tổng thể 3 chương trình nói trên, công tác biên soạn từ điển đã được chuẩn bị về kĩ thuật tin học để có thể hoàn toàn tiến hành trên máy tính. Người biên soạn có thể tiến hành tất cả các thao tác: tra cứu ngữ liệu, tham khảo các định nghĩa của các từ điển đã có, soạn thảo định nghĩa từ điển, sửa chữa định nghĩa, trao đổi thông tin với những người cùng nhóm biên soạn,... trên máy tính, chuyển kết quả soạn thảo sang Microsoft Word để in ấn thành từ điển. Đây là kĩ thuật biên soạn từ điển hoàn toàn mới ở nước ta, lần đầu tiên được ứng dụng ở Viện ngôn ngữ học. Với kĩ thuật này, chắc chắn công tác biên soạn từ điển (trước mắt là từ điển giải thích tiếng Việt) sẽ thuận lợi hơn nhiều so với trước đây, góp phần tiết kiệm được công sức cũng như rút ngắn được thời gian biên soạn từ điển (tất nhiên, chất lượng từ điển là hoàn toàn phụ thuộc vào trình độ người biên soạn).

4. Trang thiết bị kĩ thuật:

- Sắm trang thiết bị theo dự kiến, gồm: hai máy tính cá nhân, 1 máy chủ, 1máy quét, 1máy in Lase, 1máy ghi CD.

- Mua phần mềm nhận diện chữ Việt, đã sử dụng để quét và nhận diện văn bản đầu vào.

- Thiết lập mạng cục bộ. Nối mạng cục bộ cho 8 máy.

- Nối mạng Internet và khai thác tư liệu trên mạng.

5. Chuẩn bị về tư liệu:

5.1. Nguồn tài liệu dịch để tham khảo về lí thuyết (tổng số 648 trang in khổ A_4):

- Tập bài dịch các vấn đề liên quan đến lí thuyết biên soạn từ điển và từ điển học, gồm 14 bài, 267 trang (dịch từ các ngữ: Anh, Nga, Trụng quốc, Pháp)

- Tập tài liệu giới thiệu về Mạng ngữ liệu Longman (Ngân hàng ngữ liệu tiếng Anh hiện đại), 60 trang.

- Tác phẩm "Từ điển học và việc phân tích khái niệm" của Anna Wierzbiska, in tại Hoa Kỳ, 1985, (321 trang dịch).

5.2. Tư liệu

- Mua 60 cuốn sách lí luận cơ bản và các từ điển ở trong nước và nước ngoài để tham khảo (có danh sách kèm theo).

- Đã mua được 40 cuốn sách tiếng Việt chọn lọc, thuộc loại có chất lượng tốt để làm tư liệu đầu vào

- Đã mua được 471 tác phẩm các loại văn bản dưới dạng chế bản để nhập vào Ngân hàng ngữ liệu, trong đó đã đưa vào 234 tác phẩm (có danh sách kèm theo).

- Đã chọn nhập vào kho phiếu 616 tác phẩm văn học và 120 xuất bản phẩm định kì (có danh sách kèm theo).

- Liên hệ với 20 toà soạn báo và nhà xuất bản để thường xuyên mua hoặc trao đổi văn bản dưới dạng chế bản để cập nhật cho Ngân hàng ngữ liệu (có danh sách kèm theo).

6. Đoàn khảo sát tại Cộng hoà Pháp

Tổ chức đoàn khảo sát (3 người) về công tác từ điển học tại Cộng hoà Pháp trong thời gian 10 ngày (tháng 5-2001) tại hai nhà xuất bản từ điển lớn của Pháp là Hachette và Larousse. Kết quả: học tập được kinh nghiệm của Pháp trong công tác xây dựng kế hoạch và tổ chức biên soạn các loại từ điển, kinh nghiệm về việc ứng dụng thành tựu của công nghệ

thông tin vào việc xây dựng ngân hàng dữ liệu ngôn ngữ cũng như vào việc biên soạn từ điển. Những kinh nghiệm và bài học đó đã được ứng dụng vào quá trình thực hiện đề tài cũng như sẽ được áp dụng vào quá trình xây dựng kế hoạch, tổ chức biên soạn các loại từ điển sắp tới (có các báo cáo kèm theo, in trong tập Cơ sở lí luận cho việc biên soạn bộ TĐTV cỡ lớn).

III. Đánh giá, đề xuất

- Công trình đã hoàn thành các nhiệm vụ được giao với tiến độ thực hiện tương đối tốt, đều ở các phần việc; một số phần việc hoàn thành vượt mức kế hoạch (như tư liệu cho thử nghiệm ngân hàng dữ liệu tiếng Việt: vượt mức 3.200.000 âm tiết). Các hạng mục công việc đã hoàn thành đạt chất lượng tốt, đảm bảo yêu cầu đề ra. Các kết quả đạt được trong công trình này có thể dùng để biên soạn cuốn TĐTV cỡ lớn trong các năm tới.

- Thực hiện thành công công trình này không những thu được những kết quả cụ thể nêu trên mà còn là bước tập dượt, đào tạo đối với các cán bộ tham gia công trình về các bước tiến hành, việc tổ chức thực hiện,...

- Một số đề xuất về việc xây dựng ngân hàng dữ liệu tiếng Việt: Qua chuyến khảo sát việc làm từ điển và ứng dụng công nghệ tin học ở hai nhà xuất bản lớn của Pháp Hachette và Larousse, chúng tôi nhận thấy: Việc xây dựng một kho dữ liệu tiếng Việt hiện đại là rất cần thiết, nhưng đây thực sự là việc làm hết sức khó khăn, đòi hỏi rất nhiều thời gian, tiền của và tri thức. Việc làm một ngân hàng dữ liệu thử nghiệm như chúng ta đang làm hiện nay là hoàn toàn đúng hướng và đã bước đầu ứng dụng được những thành tựu mới nhất về tin học trong công tác từ điển. Nhưng với điều kiện của chúng ta hiện nay, việc đặt ra mục đích cuối cùng là một ngân hàng dữ liệu tiếng Việt nhằm đại diện cho tiếng Việt hiện đại tương tự như những ngân hàng dữ liệu của Anh, Pháp và các nước tiên tiến khác trong một vài năm là chưa thể đạt được (Ngân hàng dữ liệu tiếng Anh của Longman có hơn 100 triệu từ, gồm cả chữ viết và âm thanh, được thực hiện với kinh phí đầu tư hơn 1 triệu USD). Do vậy, để tiến tới một ngân hàng dữ liệu tiếng Việt đầy đủ, hiện đại, cần phải có những kế hoạch tiếp theo, kế thừa và phát triển kết quả của công trình thử nghiệm này.

PHẦN THỨ HAI

MỤC TIÊU, NHIỆM VỤ CỦA TỪ ĐIỂN TIẾNG VIỆT CỠ LỚN

1. *Từ điển tiếng Việt cỡ lớn* là công cụ cần thiết cho công cuộc chuẩn hoá tiếng Việt trong giai đoạn hiện nay.

Nhiệm vụ hàng đầu của *Từ điển tiếng Việt cỡ lớn* là làm công cụ để tra cứu, giúp cho người sử dụng hiểu đúng ý nghĩa và dùng đúng từ ngữ tiếng Việt ở giai đoạn hiện nay. Nó sẽ cung cấp cho cán bộ giảng dạy, các nhà báo, nhà văn, các nhà nghiên cứu và cán bộ hoạt động trong các cấp, các ngành,... hệ thống thuật ngữ khoa học chính xác và hiện đại của tiếng Việt. Cũng qua từ điển này, người dùng có thể tìm thấy những dạng viết đúng chính tả của những từ cần phân biệt về chính tả, cách phiên âm từ có gốc nước ngoài, v.v.

2. *Từ điển tiếng Việt cỡ lớn* là công trình phản ánh diện mạo từ vựng tiếng Việt ở những năm đầu thế kỉ XXI.

- Để thực hiện được nhiệm vụ này, việc xây dựng ngân hàng dữ liệu tiếng Việt phục vụ cho việc biên soạn từ điển là rất quan trọng. Từ kho dữ liệu phong phú, *Từ điển tiếng Việt cỡ lớn* sẽ có điều kiện phản ánh được kho từ vựng cơ bản của tiếng Việt ở những năm đầu thế kỉ XXI. Thực hiện được nhiệm vụ này, *Từ điển tiếng Việt cỡ lớn* còn có tác dụng tích cực cho việc nghiên cứu tiếng Việt nói chung.

- Là công cụ tra cứu, *Từ điển tiếng Việt cỡ lớn* còn là cơ sở cần thiết, là chỗ dựa để xây dựng từ điển các loại như từ điển song ngữ, từ điển học sinh, từ điển tiếng Việt thông dụng, từ điển đồng nghĩa, trái nghĩa, từ điển bách khoa và các từ điển chuyên ngành, v.v, góp phần hình thành nên hệ thống các loại từ điển tiếng Việt cần thiết cho xã hội.

3. Các công việc cần tiến hành

3.1.Tổ chức lực lượng cán bộ

Cán bộ biên soạn từ điển gồm:

- Các nhà nghiên cứu, biên soạn từ điển.

- Lực lượng cộng tác viên gồm chuyên gia đầu ngành các ngành khoa học kĩ thuật, khoa học tự nhiên và khoa học xã hội - nhân văn.

- Các kĩ thuật viên: các chuyên gia thông thạo việc ứng dụng tin học vào ngôn ngữ học để quản lí mạng nội bộ, khai thác tin trên mạng Internet,..., kĩ thuật viên có kiến thức ngôn ngữ học và sử dụng thông thạo một số phần mềm chuyên dụng trên máy tính, để giúp sửa bản nhận dạng chữ Việt, nhận diện từ mới, chọn ngữ cảnh điển hình,...

3.2. Xây dựng ngân hàng dữ liệu từ điển tiếng Việt bằng công nghệ thông tin

Khác với nhiều công trình biên soạn khác, từ điển không phải là một công trình sáng tác. Người biên soạn từ điển không tự sáng tác ra các từ và cách dùng của từ. Công việc của người biên soạn từ điển là: trên cơ sở lí luận ngôn ngữ học, dựa vào thực tế sử dụng phong phú của từ ngữ trong cuộc sống mà phân tích và khái quát, vạch ra nghĩa của từ để giúp cho người sử dụng hiểu và sử dụng đúng từ ngữ. Cái "thực tế sử dụng phong phú của từ ngữ" đó chính là nguồn dữ liệu cho công việc biên soạn từ điển.

Việc xây dựng kho dữ liệu là rất quan trọng bởi các lí do:

- Kho dữ liệu là cơ sở cần thiết để xây dựng bảng từ cho từ điển. Kho dữ liệu càng phong phú, vốn từ được thu thập càng nhiều, càng đảm bảo cho sự lựa chọn, xây dựng bảng từ được khách quan, đầy đủ, không bị sót.

- Kho dữ liệu là chất liệu giúp cho người biên soạn từ điển tìm ra các nghĩa của từ một cách đầy đủ và khách quan; đồng thời giúp cho việc định nghĩa từ chính xác.