VIỆN CÔNG NGHỆ THÔNG TIN

BÁO CÁO TỔNG KẾT KHOA HỌC VÀ CÔNG NGHỆ
ĐỀ TÀI NHÁNH

XÂY DỰNG MÔ HÌNH TỪ ĐIỂN ĐIỆN TỬ CHO TIẾNG VIỆT

THUỘC ĐỀ TÀI CẤP NHÀ NƯỚC
"NGHIÊN CỨU PHÁT TRIỂN CÔNGNGHỆ NHẬN DẠNG, TỔNG HỢP VÀ XỬ LÝ NGÔN NGỮ TIẾNG VIỆT"

Mã số: KC 01.03

Chủ nhiệm đề tài: GS.TSKH . BẠCH HƯNG KHANG

6455-4
07/8/2007

HÀ NỘI- 2004

Đề tài KC01 - 03:

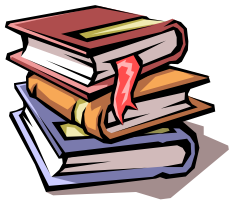
BÁO CÁO KỸ THUẬT

VỀ MÔ HÌNH TỪ ĐIỂN ĐIỆN TỬ VMTD

Người thực hiện:

GS. TSKH Hồ Tú Bảo, Japan Advanced Institute of Science and Technology

KS. Nghiêm Anh Tuấn, Viện Công Nghệ Thông Tin.



MỤC LỤC

Giới thiệu

Một trong các mục tiêu quan trọng của ngành Công nghệ thông tin là làm cho máy tính có khả năng giao tiếp với con người bằng ngôn ngữ của con người (ngôn ngữ tự nhiên). Tương tự việc con người cần đến từ điển khi học và sử dụng một ngôn ngữ, máy tính cần có từ điển của riêng mình để có thể hiểu và sử dụng các từ trong một ngôn ngữ tự nhiên. Từ điển điện tử cung cấp nguồn tri thức giúp máy tính có thể hiểu được ngôn ngữ con người và đóng vai trò nền tảng cho các nghiên cứu về ngôn ngữ tự nhiên.

Khác với các từ điển trên máy tính dành cho con người như Lạc Việt Từ điển, Click and See hay Kim từ điển… từ điển điện tử được thiết kế riêng cho các ứng dụng xử lý ngôn ngữ tự nhiên như dịch máy, trả lời tự động… Vì vậy hệ thống ngữ nghĩa (cách biểu diễn nghĩa của từ) trong từ điển điện tử không được lưu trữ dưới dạng ngôn ngữ tự nhiên như trong từ điển thông thường mà phải ở một số dạng đặc biệt để máy tính có thể xử lý được như mạng ngữ nghĩa, frame…

Để có thể thấy rõ hơn vai trò của từ điển điện tử ta hãy xét một số ví dụ sau đây:

Xây dựng engine tìm kiếm dựa trên ngữ nghĩa: với các engine tìm kiếm thông dụng như Google hay Yahoo, ta có thể tìm được những văn bản có chứa một từ khóa nào đó. Tuy nhiên, với các từ khóa đa nghĩa như table (là "bàn" hoặc "bảng biểu") và nếu người dùng chỉ muốn tìm các văn bản có chứa từ "table" với nghĩa "bảng biểu" thì các engine tìm kiếm hiện nay sẽ trả về rất nhiều tài liệu không liên quan. Trong trường hợp này nếu ta thực hiện việc chỉ mục các văn bản không phải theo sự xuất hiện của từ khóa mà theo nghĩa của từ thì ta có thể dễ dàng giải quyết vấn đề nêu trên.

Xây dựng hệ quản trị cơ sở dữ liệu cho phép truy vấn dựa trên ngữ nghĩa: Giả sử ta có câu truy vấn sau: "Hãy tìm tất cả những người trí thức đang sống trong khu phố X". Với một hệ quản trị cơ sở dữ liệu thông thường trong điều kiện ta chỉ có trường mô tả nghề nghiệp, ta không thể thực hiện được câu truy vấn này bởi trong cơ sở dữ liệu không lưu trữ bản ghi nào có giá trị trường nghề nghiệp là "trí thức" cả. Tuy nhiên, với sự hỗ trợ của từ điển điện tử, ta có thể biết rằng "bác sỹ", "kỹ sư", "nhà văn", "nhà thơ"… là những nghề nghiệp của giới trí thức. Vì vậy ta có thể tìm ra tất cả các bản ghi có chứa những từ này.

Trên thế giới, đã có rất nhiều dự án lớn kéo dài nhiều năm nghiên cứu về từ điển điện tử như dự án WORDNET tại Đại học Princeton, dự án Cyc phát triển bởi công ty CYCORP, dự án EDR của Viện nghiên cứu về từ điển điện tử của Nhật bản. Tại Việt Nam, từ điển điện tử cũng đã bắt đầu được sử dụng trong một số ứng dụng xử lý ngôn ngữ tự nhiên tiếng Việt. Mặc dù vậy, các từ điển này được thiết kế chuyên biệt cho từng ứng dụng cụ thể nên chúng khó có thể được áp dụng một cách rộng rãi. Hơn nữa, việc thiếu những nghiên cứu chuyên sâu về từ điển điện tử đã phần nào ảnh hưởng đến chất lượng của các từ điển này. Chính vì vậy, yêu cầu đặt ra là cần tiến hành nghiên cứu các mô hình từ điển điện tử trên thế giới, từ đó đề xuất một mô hình phù hợp cho từ điển điện tử tiếng Việt và cuối cùng là đưa ra quy trình thực hiện việc xây dựng từ điển.

Tài liệu này giới thiệu một mô hình của từ điển điện tử tiếng Việt phát triển trong khuôn khổ đề tài KC01-03. Tài liệu tập trung giới thiệu cấu trúc của từ điển điện tử cho tiếng Việt VMTD, gồm bốn phần chính như sau:

1. Giới thiệu cấu trúc chung của VMTD: các từ điển con cùng mối liên hệ giữa chúng.
2. Giới thiệu chi tiết cấu trúc từng bản ghi của các từ điển con.
3. Xác định các bước cần thực hiện cũng như các vấn đề cần giải quyết để xây dựng VMTD.
4. Kết luận

1. Cấu trúc chung của từ điển VMTD

VMTD bao gồm các từ điển con sau:

- Từ điển từ.
- Từ điển khái niệm.
- Từ điển song ngữ.
- Từ điển đồng hiện diện.
- Corpus.

Mỗi từ điển con có hai phiên bản cho tiếng Anh và tiếng Việt.

1.1. Từ điển từ

Chứa các thông tin về mặt cấu tạo từ và đặc tính ngữ pháp của từ. Bên cạnh đó, từ điển từ còn chứa các con trỏ khái niệm liên kết từ với nghĩa (khái niệm) tương ứng của nó trong từ điển khái niệm.

1.2. Từ điển khái niệm

Biểu diễn các khái niệm của con người dưới dạng mạng ngữ nghĩa. Từ điển khái niệm gồm có hai từ điển con: từ điển phân loại khái niệm và từ điển miêu tả khái niệm.

Từ điển miêu tả khái niệm lưu trữ tất cả các mối liên hệ giữa các khái niệm. Nó là một mạng ngữ nghĩa trong đó các khái niệm được liên kết với nhau thông qua 18 loại mối liên hệ khác nhau. Các mối liên hệ này được lựa chọn sao cho việc sử dụng chúng có thể biểu diễn được hầu hết mối liên hệ giữa các khái niệm trong một câu.

Ví dụ trong câu "Tôi ăn cơm", giữa các khái niệm "tôi", "ăn" và "cơm" ta có hai mối liên hệ sau: ("Tôi" <- tác nhân- "ăn"), ("cơm" <- đối tượng – "ăn"). Từ điển phân loại khái niệm là một cấu trúc cây trong đó các khái niệm được liên kết với nhau thông qua mối quan hệ "cha-con". Ví dụ "chim" là một khái niệm con của khái niệm "động vật". Từ điển này giúp giảm bớt khối lượng lưu trữ số mối liên hệ trong từ điển miêu tả khái niệm thông qua sự kế thừa. Trong ví dụ trên, do "chim" là một khái niệm con của "động vật" nên nó thừa hưởng mọi mối liên hệ của khái niệm "động vật" với các khái niệm khác.

Thông thường, từ điển khái niệm được sử dụng để biểu diễn ngữ nghĩa của câu, để xác định tính giống nhau về mặt ngữ nghĩa giữa các câu, hoặc để biến đổi một nội dung ngữ nghĩa này về nội dung ngữ nghĩa khác gần tương đương (Ví dụ như trong dịch tự động khi một khái niệm của ngôn ngữ gốc không có khái niệm tương ứng trong ngôn ngữ đích thì ta phải tìm một khái niệm khác trong ngôn ngữ đích gần tương đương với nó).

1.3. Từ điển đồng hiện diện

Chứa các cặp từ có mối quan hệ phụ thuộc lẫn nhau về mặt ngữ pháp cũng như ngữ nghĩa trong các câu thực tế. Ví dụ người ta hay nói "tra từ điển" chứ ít khi nói "tìm trong từ điển", hoặc "xem phim" chứ không "nhìn phim". Từ điển này được sử dụng trong một số ứng dụng sau:

- Sản sinh tự động câu trong ngôn ngữ tự nhiên: (ví dụ như các hệ thống trả lời tự động) giúp cho câu được tạo ra gần giống ngôn ngữ của con người hơn.
- Xây dựng từ điển với sự trợ giúp của máy tính: xác định tự động những cụm từ hay xuất hiện cùng nhau để liệt kê trong từ điển.
- Hỗ trợ việc giải quyết nhập nhằng trong quá trình phân tích cấu trúc ngữ pháp của câu: những cấu trúc nào có chứa nhiều cặp từ giống với ngôn ngữ tự nhiên hơn sẽ được ưu tiên hơn.

1.4. Từ điển song ngữ

Cũng giống như từ điển song ngữ thông thường, từ điển này liệt kê sự tương ứng về mặt từ trong các ngôn ngữ khác nhau. Để phục vụ cho mục đích dịch tự động, từ điển này cung cấp sự tương ứng tốt nhất về mặt từ giữa hai ngôn ngữ.

1.5. Corpus

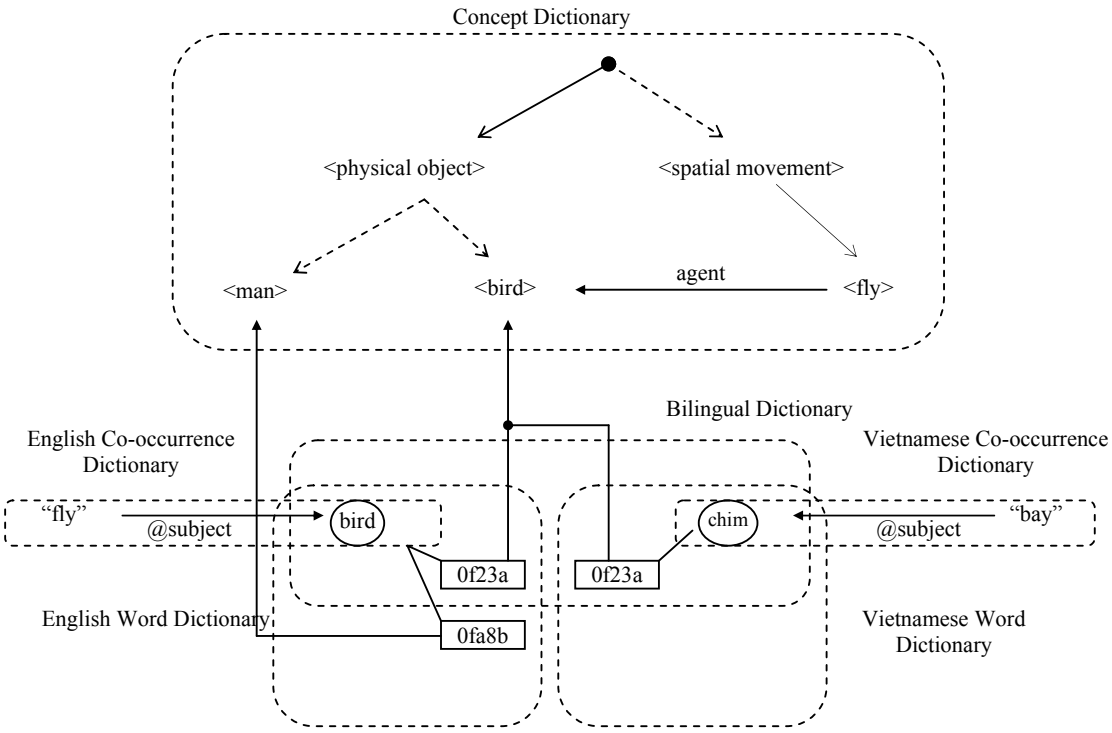
Là một tập các câu được phân tích đến mức ngữ nghĩa. Với mỗi câu, corpus lưu trữ thông tin về mặt hình thái cấu tạo từ, cấu trúc ngữ pháp và mối liên hệ giữa các khái niệm trong câu. Corpus được sử dụng chủ yếu để xây dựng từ điển đồng hiện diện, từ điển khái niệm và từ điển từ:

- Mối liên hệ giữa các khái niệm trong câu được sử dụng để xây dựng từ điển khái niệm.

- Cấu trúc ngữ pháp của các câu trong corpus được sử dụng để tìm ra mối quan hệ đồng hiện diện cho từ điển đồng hiện diện.
- Nghĩa của từ cũng như cách sử dụng từ trong từ điển từ được kiểm chứng thông qua corpus. Corpus cũng giúp xác định những từ mới để bổ sung vào từ điển từ.

1.6. Mối quan hệ giữa các từ điển con trong VMTD

Mỗi từ điển con trong VMTD bao gồm hai phần cho tiếng Anh và tiếng Việt. Hình 1 thể hiện mối liên hệ giữa các từ điển con trong VMTD.



Hình 1: Mối liên hệ giữa các từ điển con trong VMTD

2. Cấu trúc các từ điển con trong VMTD

Sau đây là nội dung của các từ điển con. Phần tiếng Anh và tiếng Việt của mỗi từ điển con sẽ được trình bày chung, chỉ khi nào có sự khác biệt thì hai phần này sẽ được trình bày riêng.

2.1. Từ điển từ

Đơn vị của từ điển từ là một mục từ. Mỗi mục từ bao gồm thông tin từ đầu mục, thông tin ngữ pháp, thông tin ngữ nghĩa và các thông tin thêm. Thông tin từ đầu mục bao gồm từ đầu mục, phân chia âm tiết và cách phát âm. Thông tin ngữ pháp bao gồm từ loại, các thuộc tính ngữ pháp và thông tin từ chức năng. Thông tin ngữ nghĩa là một con số dùng để xác định khái niệm tương ứng của từ trong từ điển khái niệm. Ta gọi số đó là định danh khái niệm. Thông tin thêm bao gồm cách sử dụng (đối với các từ viết tắt và tiếng lóng) và tần suất xuất hiện của từ. Tần suất xuất hiện của từ là một phân số mà tử số là số lần xuất hiện của từ với khái niệm chỉ bởi định danh khái niệm và mẫu số là số lần xuất hiện của từ trong corpus.

Bảng 1: Cấu trúc chung của một mục từ trong từ điển từ.

Thông tin từ đầu mục	Thông tin ngữ pháp	Thông tin ngữ nghĩa	Thông tin thêm
Từ đầu mục Phân chia âm tiết Cách phát âm	Từ loại (danh từ, động từ, tính từ...) Thuộc tính ngữ pháp. Thông tin từ chức năng.	Định danh khái niệm	Cách sử dụng Tần suất

Một số đặc điểm riêng của từ điển từ tiếng Anh

Trong phần từ đầu mục thì một từ đầu mục tiếng Anh là một danh sách các thành tố khác nhau mà mỗi thành tố bao gồm thân từ (Notation) và các thuộc tính kế cận. Ví dụ soon(Adverb with Initial Consonant Sound, Adverb - Inflection Pattern er).

Trong phần thông tin ngữ pháp có thêm hai mục cây cú pháp và biến tố. Cây cú pháp là cấu trúc ngữ pháp của các cụm từ cố định hoặc các thành tố. Biến tố là cách biến đổi của từ khi sử dụng trong những trường hợp cụ thể, ví dụ khi động từ “go” chia ở ngôi thứ ba số ít sẽ thêm hậu tố và trở thành “goes”.

Bảng 2: Cấu trúc của một mục từ tiếng Anh

Thông tin từ đầu mục	Thông tin ngữ pháp	Thông tin	Thông tin

		ngữ nghĩa	thêm
Từ đầu mục Các thành tố Thân từ Thuộc tính kế cận Phân chia âm tiết Cách phát âm	Từ loại (danh từ, động từ, tính từ...) Thuộc tính ngữ pháp. Thông tin từ chức năng. Cây cú pháp Biến tố	Định danh khái niệm	Cách sử dụng Tần suất

Ví dụ về một mục từ trong từ điển từ tiếng Anh

<Thông tin từ đầu mục>

<Từ đầu mục>: soon

<Thành tố và các thuộc tính kế cận>: soon(Adverb with Initial Consonant Sound, Adverb - Inflection Pattern er)

<Phân chia âm tiết>: soon

<Cách phát âm>: s'u:n

<Thông tin ngữ pháp>

<Từ loại>: trạng từ

<Cây cú pháp>:

<Biến tố>: trạng từ - mẫu biến tố “er” (Adverb - Inflection Pattern er)

<Thuộc tính ngữ pháp>: Có thể đứng sau bổ ngữ (object hoặc complement).

<Thông tin từ chức năng>:

<Thông tin ngữ nghĩa>:

<Định danh khái niệm>: 0ea98d

<Thông tin thêm>:

<Cách dùng>:

<Tần suất>: 209/892.

Một số đặc điểm riêng của từ điển từ tiếng Việt

Thành phần của một mục từ thuộc từ điển từ tiếng Việt giống với cấu trúc chung của mục từ đã trình bày ở trên. Tuy nhiên tiếng Việt cũng có những đặc trưng riêng ảnh hưởng đến cấu trúc của một mục từ:

- Trong tiếng Việt, các âm tiết được phân cách bằng khoảng trắng.
- Cách đọc của tiếng Việt không có trường hợp ngoại lệ nên nếu ta biết một âm tiết được viết như thế nào thì ta cũng có thể biết cách đọc âm tiết đó.

Do vậy, mục phân chia âm tiết và cách phát âm trong phần thông tin từ đầu mục chỉ dành cho các từ mượn như “Braxin”, “taxi”...

Ví dụ một mục từ của từ điển từ tiếng Việt

<Thông tin từ đầu mục>

<Từ đầu mục>: từ điển

<Phân chia âm tiết>

<Cách phát âm>

<Thông tin ngữ pháp>

<Từ loại>: Danh từ

<Thuộc tính ngữ pháp>: Danh từ chỉ vật

<Thông tin từ chức năng>

<Thông tin ngữ nghĩa>

<Định danh khái niệm>: 0f6f4b

<Thông tin bổ trợ>

<Cách dùng>

<Tần suất>: 73/73

2.2. Từ điển khái niệm

Như đã trình bày ở trên, từ điển khái niệm bao gồm từ điển phân loại khái niệm và từ điển miêu tả khái niệm. Tuy nhiên, trong hai từ điển này khái niệm được biểu diễn dưới dạng những con số (định danh khái niệm). Vì vậy để giúp con người có thể phân biệt các khái niệm với nhau, cần phải có thêm từ điển giải thích khái niệm trong đó các định danh khái niệm đều được giải thích bằng ngôn ngữ tự nhiên.

2.2.1. Từ điển giải thích khái niệm

Từ điển giải thích khái niệm bao gồm một tập các mục giải thích khái niệm, mỗi mục giải thích khái niệm tương ứng với một khái niệm cụ thể. Cấu trúc của một mục giải thích khái niệm được thể hiện trong bảng 3.

Bảng 3: Cấu trúc của một bản ghi giải thích khái niệm

Định danh khái niệm	Từ biểu diễn khái niệm	Giải thích khái niệm

Số hexa đại diện cho khái niệm	Từ tiếng Anh Từ tiếng Việt	Giải thích bằng tiếng Anh Giải thích bằng tiếng Việt

Từ biểu diễn khái niệm là một từ mà nghĩa của nó gần với khái niệm đang xét nhất. Phần giải thích khái niệm là một câu giải thích rõ nghĩa của khái niệm bằng ngôn ngữ tự nhiên. Sau đây là một ví dụ về mục giải thích khái niệm.

<Định danh khái niệm>: 3d0ecb

<Từ biểu diễn khái niệm>

    <Từ tiếng Anh>: borrow

    <Từ tiếng Việt>: mượn

<Giải thích khái niệm>

    <Giải thích bằng tiếng Anh>: to use a person's property after promising to return

    <Giải thích bằng tiếng Việt>: sử dụng tài sản của người khác sau khi đã hứa sẽ trả lại.

2.2.2. Từ điển phân loại khái niệm

Từ điển phân loại khái niệm bao gồm một tập các bản ghi phân loại khái niệm. Mỗi bản ghi phân loại khái niệm là một cặp định danh của khái niệm cha và định danh của khái niệm con. Sau đây là ví dụ của một bản ghi phân loại khái niệm.

<Định danh của khái niệm cha>: 4445bc (khái niệm chỉ một văn bản)

<Định danh của khái niệm con>: 4445a0 (khái niệm chỉ một bức thư)

2.2.3. Từ điển mô tả khái niệm

Từ điển mô tả khái niệm bao gồm một tập các bản ghi mô tả khái niệm. Sau đây là ví dụ của một bản ghi mô tả khái niệm.

<Loại mô tả>: E

<Mô tả>

    <Định danh khái niệm 1>: 0d0ecb (Định danh của khái niệm “mượn”)

<Loại quan hệ>: object

<Định danh khái niệm 2>: 0e5097 (Định danh của khái niệm “sách”)

<Nhân tố chắc chắn>: 1

Trường “Loại mô tả” có thể nhận một trong hai giá trị là “I” và “E”. “E” có nghĩa là trong corpus có chứa ít nhất một câu trong đó hai khái niệm này liên kết với nhau bởi mối quan hệ object. “I” có nghĩa là mối liên hệ giữa hai khái niệm này được xây dựng dựa trên trực quan của con người.

Nhân tố chắc chắn có thể nhận một trong 2 giá trị 0 hoặc 1. Nếu nhân tố chắc chắn có giá trị 0 thì có nghĩa là không thể có một quan hệ như vậy giữa hai khái niệm. Có thể thấy sự cần thiết của giá trị này trong ví dụ sau.

Do khái niệm “chim cánh cụt” là một khái niệm con của khái niệm “chim” nên nó có thể thừa hưởng mọi đặc tính của khái niệm “chim”. Nhưng giữa “chim” và “bay” có mối liên hệ agent bởi “chim” thì có thể “bay”. Điều này là không đúng với khái niệm “chim cánh cụt”. Để thể hiện rằng “chim cánh cụt” thì không biết “bay” ta sẽ thêm vào từ điển miêu tả khái niệm một bản ghi với nhân tố chắc chắn nhận giá trị 0.

2.3. Từ điển song ngữ

Mỗi bản ghi của từ điển song ngữ bao gồm thông tin từ đầu mục ở ngôn ngữ gốc và thông tin từ tương ứng ở ngôn ngữ đích. Thông tin từ đầu mục bao gồm từ đầu mục, từ loại và định danh khái niệm. Thông tin từ tương ứng là một danh sách các cặp (loại tương ứng, từ tương ứng). Từ tương ứng chỉ ra từ gần giống nghĩa với từ đầu mục trong ngôn ngữ đích, loại tương ứng chỉ ra mối quan hệ tương đương giữa từ đầu mục và từ tương ứng. Loại tương ứng có thể nhận các giá trị:

- Tương đương: từ đầu mục và từ tương ứng cùng biểu diễn một khái niệm
- Khái niệm con: từ tương ứng biểu diễn một khái niệm hẹp hơn khái niệm của từ đầu mục.
- Khái niệm cha: từ tương ứng biểu diễn một khái niệm rộng hơn khái niệm của từ đầu mục.
- Giải thích: Trong ngôn ngữ đích không tìm được một từ tương đương về mặt nghĩa với từ đầu mục. Ví dụ như từ đầu mục nói về một lễ hội đặc biệt nào

đó trong ngôn ngữ gốc. Khi đó từ tương ứng sẽ là một câu giải thích khái niệm của từ đầu mục.

Sau đây là một ví dụ về một bản ghi của từ điển song ngữ Việt – Anh.

<Thông tin từ đầu mục>
<Từ đầu mục>: thông cáo
<Từ loại>: Danh từ
<Định danh khái niệm>: 0b13c9
<Thông tin tương ứng>
<Thông tin từ tương ứng>
<Loại tương ứng>: tương đương
<Từ tương ứng>: announcement

<Loại tương ứng>: tương đương
<Từ tương ứng>: notice

2.4. Từ điển đồng hiện diện

Từ điển đồng hiện diện bao gồm một tập các bản ghi về từ đồng hiện diện. Mỗi bản ghi lưu trữ thông tin về một cặp (từ chính, từ phụ thuộc). Từ chính là từ quyết định xem đứng cạnh nó có thể là những từ nào. Sau đây là một ví dụ về một bản ghi từ đồng hiện diện tiếng Việt cho cặp (ăn, cơm).

<Từ chính>
<WN> <HW> <POS> <C>
{1 ăn VERB 3bc6f0}
<Quan hệ >: @object
<Tần suất>: 12
<Từ phụ thuộc>
<WN> <HW> <POS> <C>
{2 cơm NOUN 2bec74}

WN: thứ tự của từ trong câu thực tế. Trong bản ghi nói trên giá trị WN của từ “ăn” là 1 và từ “cơm” là 2 nên từ “ăn” sẽ đứng trước từ cơm.

HW: từ

POS: từ loại.

C: định danh khái niệm.

Quan hệ: mối quan hệ giữa hai từ. Trong ví dụ trên là mối quan hệ giữa động từ và bổ ngữ.

Tần suất: số lần xuất hiện của cặp từ này trong corpus.

Cấu trúc của một bản ghi về từ đồng hiện diện tiếng Anh cũng gần tương tự như với tiếng Việt. Sau đây là một ví dụ về bản ghi từ đồng hiện diện tiếng Anh cho cặp (eaten, lunch).

<Từ chính>

<WN>	<M>	<HW>	<POS>	<C>
{1	eaten	eat	VERB	3bc6f0}

<Quan hệ >: @object

<Tần suất>: 12

<Từ phụ thuộc>

<WN>	<M>	<HW>	<POS>	<C>
{2	lunch	lunch	NOUN	2bec74}

M: dạng biến tố của từ trong câu thực tế.

2.5. Corpus

Corpus là một tập hợp các câu đã được phân tích đến mức ngữ nghĩa. Sau đây là một ví dụ về một câu tiếng Việt đã được phân tích trong corpus.

<Thông tin về câu>

<Số hiệu câu>: 0020000026cd

<Nguồn>: Báo Nhân Dân

<Câu>: Việt Nam có tiềm năng du lịch to lớn.

<Thông tin về từ>

<WN>	<HW>	<POS>	<C>
1	Việt Nam	NOUN	"Đất nước Việt Nam"
2	có	VT	2dc2fd
3	tiềm năng	NOUN	2dc2fd
4	du lịch	NOUN	3cdfda
5	to lớn	ADJ	2fcd3a
6	"."	PUNC	2dc2e5

<Cây cú pháp>

S[NP[Việt Nam]VP[V[có]NP[NP[N[tiềm năng]ADJ[du lịch]]ADJ[to lớn]]]]

<Cây ngữ nghĩa>

(Biểu diễn ngữ nghĩa của câu này)

[<Nhãn quan hệ><Thứ tự từ trong câu><Từ><Định danh khái niệm>]

[[main 2:có:0e910d] [agent 1:Việt Nam:2dc304][object [main [main 3:tiềm năng:3d0797] [object 4: du lịch: 31123]][object 5: to lớn]

Sau đây là ví dụ về một câu tiếng Anh đã được phân tích trong corpus.

<Thông tin câu>

<Số hiệu câu> 0020000026cd

<Nguồn> Japan Times

<Câu> He's a very promising young man.

<Thông tin về từ>

<WN>	<M>	<HW>	<POS>	<C>
1	he	he	PRON	2dc304
2	's	be	BE	2dc2f8
3	_	_	BLNK	2dc2ed
4	a	a	ART	2dc2f3
5	_	_	BLNK	2dc2ed
6	very	very	ADV	0f847a
7	_	_	BLNK	2dc2ed
8	promising	promising	ADJ	3ce992
9	_	_	BLNK	2dc2ed
10	young	young	ADJ	0e2544
11	_	_	BLNK	2dc2ed
12	man	man	NOUN	0c7a38
13	.	.	PUNC	2dc2e5

<Thông tin hình thái>

/1:he/2:'s /3: /4: a /5: /6:very /7: /8:promising/9: /10:young/11: /12:man

/13: . /

<Cây cú pháp>

S[PRON[He]][VP[Verb

be['s]][COMP[[ADV[very]][NP[[ADJ[promising]][NP[[ADJ[young]][PRON[man]]]]]]]]

<Cây ngữ nghĩa>

[main 12:"man":0c7a38]

[which [[main 10:"young":0e2544]

[object 12:"man":0c7a38]]]

[a-object [[main 1:"he":2dc304]

[attribute topic]]]

[modifier [[main 8:"promising":3ce992]

[manner 6:"very":0f847a]]]]

3. Các bước xây dựng từ điển VMTD

Chúng tôi đề xuất các bước cần thực hiện để xây dựng từ điển như sau:

- Bước 1: Xây dựng từ điển giải thích khái niệm và phân loại khái niệm.
- Bước 2: Xây dựng từ điển từ.
- Bước 3: Xây dựng corpus.
- Bước 4: Xây dựng các từ điển khác dựa trên corpus.

Từ điển giải thích khái niệm cần phải xây dựng đầu tiên bởi tất cả các từ điển khác cần phải được kết nối thông qua từ điển giải thích khái niệm. Sau khi có từ điển giải thích khái niệm thì ta có thể tiến hành xây dựng từ điển phân loại khái niệm bằng cách import dữ liệu của một từ điển khác sẽ nói kỹ hơn ở phần sau. Từ điển giải thích khái niệm cần phải xây dựng trước từ điển từ vì nó giúp cho quá trình kết nối giữa từ điển từ và từ điển khái niệm được tiến hành dễ dàng hơn.

Sau khi có từ điển từ thì có thể sử dụng nó để xây dựng corpus, thành phần cơ bản giúp xây dựng nên từ điển.

Sau khi corpus đã được xây dựng thì nó được sử dụng để xây dựng dữ liệu cho các từ điển khác.

3.1. Xây dựng từ điển giải thích khái niệm và phân loại khái niệm

Về nguyên tắc, do từ điển khái niệm là thành phần tương đối độc lập với ngôn ngữ nên ta có thể sử dụng lại từ điển khái niệm của EDR. Tuy nhiên, nếu ta làm như vậy thì kết quả là các ứng dụng khó có thể sử dụng VMTD vì giá thành cao.

Trong số các từ điển mà VMTD có thể sử dụng được dữ liệu thì WordNet là một từ điển miễn phí chất lượng cao, được xây dựng tại đại học Princeton. Hơn nữa, cấu trúc từ điển phân loại khái niệm của WordNet về cơ bản tương đối giống so với từ điển phân loại khái niệm của EDR. Vì vậy ta có thể sử dụng dữ liệu của WordNet cho VMTD.

3.2. Xây dựng từ điển từ

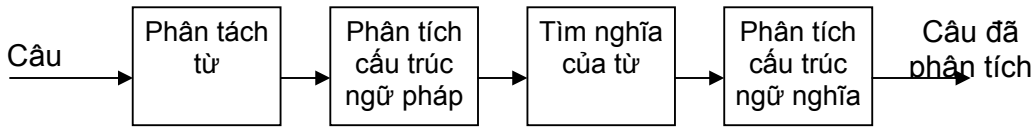
Nội dung của từ điển từ không có gì đặc biệt. Tuy nhiên một yêu cầu đặt ra là mỗi từ trong từ điển từ phải được liên kết với những khái niệm mà từ đó diễn tả trong từ điển khái niệm. Để thực hiện được điều này thì người nhập liệu phải hiểu toàn bộ

cấu trúc của từ điển khái niệm. Khi kích thước của từ điển khái niệm lên đến hàng chục nghìn thì quá trình nhập dữ liệu sẽ rất phức tạp. Do đó cần phải tự động hóa quá trình này.

Sau khi đã import dữ liệu của WordNet, ngoài từ điển phân loại khái niệm, ta còn có thêm từ điển từ tiếng Anh, mặc dù các thông tin về từ còn chưa đầy đủ, nhưng điều quan trọng nhất là các từ tiếng Anh đã được liên kết với các khái niệm tương ứng. Vì vậy để giải quyết vấn đề với các từ tiếng Việt, ta có thể thông qua từ điển song ngữ để kết nối từ tiếng Việt với khái niệm tương ứng như một số nơi đã làm cho các thứ tiếng khác.

3.3. Xây dựng corpus

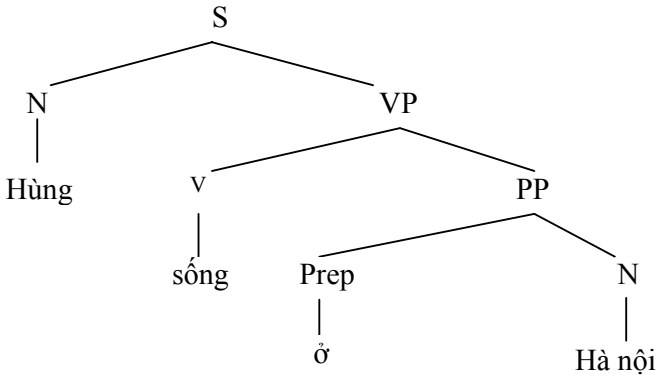
Corpus là một tập hợp các câu đã được phân tích đến mức ngữ nghĩa. Quá trình phân tích một câu được thể hiện trên hình 2.



Hình 2: Quá trình phân tích câu

3.3.1. Phân tách từ

Trong tiếng Anh, các từ được phân tách bởi khoảng trắng. Tuy nhiên, trong tiếng Việt không tồn tại một biên giới rõ ràng giữa các từ. Vì vậy việc đầu tiên trong quá trình phân tích là phải biết được câu đang xét được cấu tạo nên từ những từ nào. Hiện thời một thuật toán tách từ hiệu quả được nhiều người công nhận vẫn còn chưa tồn tại. Do đó, đây cũng còn là một vấn đề cần được quan tâm.



Hình 3: Cấu trúc ngữ pháp biểu diễn ở dạng cây

3.3.2. Phân tích cấu trúc ngữ pháp

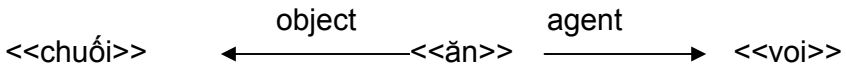
Phân tích cấu trúc ngữ pháp có nghĩa là phải tìm ra các từ đã được kết hợp với nhau như thế nào để tạo ra câu hoàn chỉnh. Thông thường để biểu thị cấu trúc ngữ pháp của một câu, người ta dùng cây phân tích. Ví dụ cấu trúc ngữ pháp của câu "Hùng sống ở Hà Nội" được biểu diễn như ở hình 3.

3.3.3. Tìm nghĩa của từ

Một từ có thể mang một vài ý nghĩa khác nhau. Nhưng vấn đề là phải tìm ra trong một câu cụ thể thì từ đó mang ý nghĩa gì. Vấn đề này mang tính chất quyết định đối với các hệ dịch tự động. Đây vẫn còn là một vấn đề mở, đặc biệt trong trường hợp tiếng Việt.

3.3.4. Phân tích cấu trúc ngữ nghĩa

Cũng giống như phân tích cấu trúc ngữ pháp, việc phân tích cấu trúc ngữ nghĩa nhằm tìm ra mối liên hệ giữa các nghĩa đơn lẻ của từ để tạo nên nghĩa của toàn bộ câu. Ví dụ như cấu trúc ngữ nghĩa của câu "Voi ăn chuối" được thể hiện trong hình 4.



Hình 4: Một ví dụ về cấu trúc ngữ nghĩa

Việc tự động phân tích cấu trúc ngữ nghĩa vẫn còn là vấn đề ít được nghiên cứu. Tuy nhiên, do VMTD được xây dựng dựa trên EDR, ta có thể học cách phân tích cấu trúc ngữ nghĩa từ EDR.

4. Kết luận

Trong tài liệu kỹ thuật này, chúng tôi đã mô tả cấu trúc từ điển điện tử VMTD cho tiếng Việt. Mô hình VMTD được xây dựng dựa trên từ điển điện tử EDR của Nhật bản với hi vọng nó có thể giúp cho sự phát triển của các ứng dụng xử lý ngôn ngữ tự nhiên của Việt Nam. Sau đó chúng tôi đã đề xuất những bước cần thực hiện cũng như những vấn đề cần giải quyết để xây dựng nên một từ điển hoàn chỉnh.

Mô hình VMTD mới chỉ là bước đầu trong quá trình xây dựng một từ điển điện tử thực sự. Quá trình này đòi hỏi phải có sự đầu tư nghiên cứu lâu dài của nhiều chuyên gia về ngôn ngữ học cũng như về tin học để có thể xây dựng được một từ điển điện tử chất lượng cao.

Tài liệu tham khảo

[1] Cheng-Ming Guo. *Machine Tractable Dictionaries, Design and Construction,* Ablex Publishing Corporation, Northwood, New Jersey 1995.

[2] Donald E. Walker, Antonio Zampolli, Nicoletta Calzolari. *Automation the lexicon,* Oxford University Press 1995.

[3] Douglas B. Lenat, R.V. Guha. *Building large knowledge-based systems: representation and inference in the CYC project,* Addison-Wesley Pub. Co., 1989, c1990.

[4] Fellbaum, Christiane. *WordNet: An electronic lexical database,* MIT Press 1998.

[5] German Rigau, Eneko Agirre. Disambiguating bilingual nominal entries against WordNet. In *Proceedings of The Computational Lexicon Workshop. Seventh European Summer School in Logic, Language and Information, ESSLLI'95*, Barcelona, Spain, 1995.

[6] Japan Electronic Dictionary Research Institute, Ltd. *EDR Electronic Dictionary Technical Guide*, 1993.

[7] Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, Horacio Rodriguez. Combining multi methods for the Automatic construction of multilingual WordNet, in *Proceeding of RANLP*, Bulgaria, 1997.

[8] Latifur R. Khan, Eduard H. Hovy. Improving the Precision of Lexicon-to-Ontology Alignment Algorithms, in *Proceedings of the AMTA/SIG-IL First Workshop on Interlinguas*, San Diego, CA October, 1997.

[9] Piek Vossen. *EuroWordNet: a multilingual database with lexical semantic network,* Dordrecht: Kluwer Academic, 1998.

[10] Vincent B.Y.Ooi. *Computer Corpus Lexicography,* Edinburgh University Press, 1998.

[11] Yorick A. Wilks, Brian M. Slator, and Louise M. Guthrie. *Electric Words,* MIT Press, 1996.

[12] Diệp Quang Ban. *Ngữ Pháp Tiếng Việt,* NXB Giáo Dục 2000 (in Vietnamese).

[13] Nguyễn Thị Quy. *Vị từ hành động tiếng Việt và các tham tố của nó*, NXB Khoa học Xã hội, 1995 (in Vietnamese).

[14] Trung tâm Khoa học Xã hội và Nhân văn Quốc Gia. *Ngữ Pháp Tiếng Việt*, NXB Khoa học Xã hội, 2000 (in Vietnamese).

Phụ lục A: Bảng mã từ của từ điển từ tiếng Anh

Bảng từ loại tiếng Anh

Category	Part of speech	Code	Example
Nouns	Common Noun	1	book
	Proper Noun	2	Tokyo
	Cardinal Number	3	one, two
	Ordinal Number	4	first
	Classifier	5	piece, amount, bit
Pronouns	Personal Pronoun	6	I, my, me, mine
	Interrogative Pronoun	7	who, what
	Demonstrative Pronoun	8	this, that
	Indefinite Pronoun	9	some, anyone
	Relative Pronoun	10	who, whose, that
Verbs	Verb	11	run
	Be-verb	12	am, are, is
Adjectives	Adjective	13	beautiful
Adverbs	Relative Adverb	14	whenever
	Interrogative Adverb	15	how
	Adverbial Particle	16	off, up, back, round
	Conjunctive Adverb	17	because, since
	Common Adverb	18	very, actually
Prepositions	Preposition	19	in, on, at
	Preposition Equivalent	20	in front of, according to, regarding
Determiners	Demonstrative Determiner	21	this, that
	Indefinite Determiner	22	any, both, either, such
	Article	23	a, an, the
	Auxiliary Verbs	24	will, must
	Auxiliary Verb Equivalent	25	have to, would rather
Interjections	Interjection	26	ah, oh
Conjunctions	Coordinate Conjunction	27	and, but
	Coordinate Conjunction Equivalent	28	Equivalent
	Subordinate Conjunction	29	whether
	Subordinate Conjunction Equivalent	30	even if, so that
To-Infinitive	To-Infinitive	31	to, not to

Affixes	Suffix	32	semi-
Word	Noun Ending	33	(book)s
Endings	Verb Ending	34	(turn)s, (turn)ed, (turn)ing
	Adjective Ending	35	(small)er, (small)est
	Adverb Ending	36	(hard)er, (hard)est
Phrases and	Noun Phrase	37	
Sentences	Verb Phrase	38	kick the bucket
	Adjective Phrase	39	green with envy
	Adverb Phrase	40	all in good time
	Preposition Phrase	41	under the counter
	Independent Phrase	42	no names, no pack drills
	Sentence	43	Time flies like an arrow.
Other	Unit	44	cm, kg
	Symbol	45	A,B,C,a,b,c,?,&

Thông tin dạng thức từ tiếng Anh (Word form Infomation)

Category	Code	Word form	Comment/Explanation
Verb	1	Invariable Portion	Applicable to verbs, be-verbs, auxiliary verbs and verb endings
	2	Stem Form	
	3	3rd Person Singular	
	4	Present Tense	
	5	Past Tense	
	6	Past Participle	
	7	Present Participle,	
	8	Gerund Form	
Noun	9	Invariable Portion	Applicable to common nouns, proper nouns, noun classifiers, and noun endings
	10	Singular Form	
	11	Plural Form	
	12	Common	
	13	Possessive Case	
Adjective	14	Invariable Portion	Applicable to adjectives and adjective endings
	15	Positive Degree	
	16	Comparative Degree	
	17	Superlative Degree	
Adverb	18	Invariable Portion	Applicable to adverbs and adverb endings
	19	Positive Degree	
	20	Comparative Degree	
	21	Superlative Degree	

Giống và số cho đại từ (Case and Number For Pronouns :Word form Information)

Code	Comment	Example
22	Subjective	I
23	Determinative possessive	My
24	Objective	Me
25	Independent possessive	Mine
26	Reflexive pronoun	Myself
27	1st person singular	I
28	1st person plural	We
29	2nd person singular	You
30	2nd person plural	You
31	3rd person singular	It
32	3rd_person_plural	They

Dạng biến tố của động từ (Verb inflection Pattern)

Regular inflection

Inflection pattern	Code	Example	Word form	Infinitive ending	3 sg form ending	Past form ending	Past participle ending	Present participle ending
s-d	1	agree	2	-	s	d	d	Ing
s-ed	2	turn	2	-	s	ed	ed	ing
es-ed	3	watch	2	-	es	ed	ed	ing
e	4	hik	1	E	es	ed	ed	ing
y	5	stud	1	Y	ies	ied	ied	ying
ie	6	d	1	Ie	ies	ied	ied	ying
s-tt	7	bat	2	-	s	C + ed	C + ed	C + ing

Partially Irregular Inflection

s-ing	8	see	2	-	s	(*)	(*)	ing
es-ing	9	go	2	-	es	(*)	(*)	ing
e-ing	10	writ	1	E	es	(*)	(*)	ing
y-ying	11	fl	1	Y	ies	(*)	(*)	ying
s-irreg.	12	hit	2	-	s	(*)	(*)	(*)
Irregular	13	have	2	(*)	(*)	(*)	(*)	(*)

Chú ý 1: '*' chỉ một biến tố bất quy tắc. Mẫu biến tố bất quy tắc được ghi trong từ điển như một từ đầu mục độc lập.

Chú ý 2: 'C + ed' và 'C + ing' chỉ việc nhân đôi phụ âm cuối cùng của từ đầu mục và thêm vào -ed và -ing.

Chú ý 3: Thông tin biến tố được cung cấp cho mọi động từ và mọi dạng của động từ 'be'.

Dạng biến tố của danh từ (Noun Inflection Pattern)

Inflection Code Example Word Singular Plural noun

pattern			form	noun ending	ending
S	14	boy	10	-	s
Es	15	box	10	-	es
Y	16	lad	9	Y	ies
Fe	17	wi	9	Fe	ves
F	18	lea	9	F	ves
s & es	19	potato	10	-	s, es
's	20		10	-	's
s &'s	21	NP	10	-	s, 's
Irregular	22		10		

Dạng biến tố của tính từ (Adjective Inflection Pattern)

Inflection pattern	Code	Example	Word form	Positive form ending	Comparative form ending	Superlative form ending
er	23	hard	15	-	er	Est
r	24	pale	15	-	R	St
ier	25	eas	14	y	ier	Iest
tt	26	big	15	-	C + er	C + est
Irregular	27	good	15	-	(*)	(*)

Dạng biến tố của trạng từ (Adverb Inflection Pattern)

Inflection pattern	Code	Example	Word form	Positive form ending	Comparative form ending	Superlative form ending
er	28	deep	19	-	er	est
r	29	late	19	-	r	st
ier	30	earl	18	y	ier	iest
tt	31	hot	19	-	C + er	C + est
Irregular	32	well	19	-	(*)	(*)

Thuộc tính ngữ pháp của động từ (Verb: Grammatical attributes)

Abbreviation		Code	
EVIT0		1	Takes neither a direct object nor an indirect object (SV/SVC/SV+ADV) Note: [EVSC0]is coded together with [EVIT0] if a subjective complement is required.
EVIO0		2	Necessarily takes an indirect object (SVOO) Note: 'Indirect object' is an object which is not a direct object. If [EVIO0] is assigned [EVDO0] must also be assigned.
	EVIO1	3	[Indirect Object (IO) + Direct Object (DO)] is replaceable with [Direct Object (DO) +'to' + noun phrase].
	EVIO2	4	[Indirect Object (IO) + Direct Object (DO)] is replaceable with [Direct Object (DO) + 'for' + noun phrase]. Note: [EVIO1] and [EVIO2] cannot be assigned together.
EVDO0		5	Necessarily takes a direct object (SVO/SVOO/SVOC/SVOC+ADV)
	EVDO1	6	DO = noun phrase (includes proper nouns and pronouns)
	EVDO2	7	DO = that - clause
	EVDO3	8	DO = subordinate wh / if - clause
	EVDO4	9	DO = wh - to - infinitive phrase
	EVDO5	10	DO = bare infinitive phrase
	EVDO6	11	DO = to - infinitive phrase
	EVDO7	12	DO = -ing phrase
	EVDO8*	13	DO = for to-infinitive phrase
	EVDO9*	14	DO = noun phrase + -ing phrase Note: [EVDO0] must be accompanied by one and only one from [EVDO1 - 9]. * To be included in next version
EVSC0		15	Necessarily takes a subjective complement (SVC)
EVOC0		16	Necessarily takes an objective complement (SVOC)
	EVC10	17	C = noun phrase (including proper nouns and pronouns)
	EVC20	18	C = adjective phrase
	EVC30	19	C = 'to be' + noun phrase
	EVC40	20	C = 'to be' + adjective phrase
	EVC50	21	C = bare infinitive phrase
	EVC60	22	C = to - infinitive phrase
	EVC70	23	C = past participle
	EVC80	24	C = -ing phrase

	EVC91*	25	C = like + noun phrase
	EVC92*	26	C = as + noun phrase
	EVC93*	27	C = for + noun phrase Note: Both [EVSC0] and [EVOC0] must be accompanied by one and only one from [EVC10 - 93]. * To be included in next version
EVSA0		28	Necessarily takes a prepositional phrase (SV+ADV/SVO+ADV)
	EVSA2	29	ADP= preposition + noun phrase (A code from Table 3-15 must be assigned.) Prepositions that can occur at the same in a sentence are given. Prepositions that cannot occur at the same time, are given on separate records.
	EVSA4*	30	ADP= prepositional phrase or adverb that indicates place, time, direction or manner
	EVSA5	31	ADP = to-infinitive
	EVSA6	32	ADP = ing phrase Note: [EVSA0] must be accompanied by one and only one from [EVSA2-6]. * To be included in next version
EVSI1		33	Prop 'it' sentence structure ('It' in subject position and has no referent) Ex. It rains.
EVSI2		34	Occurs in 'it-that' sentence structure, with 'It' as subject Ex. It seems that the wind has blown the trees over. It is said that humans have lived on earth for millions of years.
EVSI3		35	Occurs in 'as-if' sentence structure, with 'It' as subject Ex. It appears as if the movie will start late.
EVSI4		36	Occurs in 'whether' sentence structure, with 'It' as subject Ex. It doesn't matter whether you take the medicine in the morning or in the evening.
EVSI5*		37	Occurs with to-infinitive phrase, with 'It' in subject posi- tion Ex. It appears to be the only one left.
EVSI6*		38	Occurs with wh-to infinitive phrase, with 'It' in subject position

		Ex. It wasn't apparent how to solve the problem.
EVSI7*	39	Occurs with for-to infinitive phrase, with 'It' in subject position Ex. It was considered impossible for anyone to reach the top. Note: the sentence pattern information for 'consider' is: EVIT0;EVSC0;EVC20;EVSI7;EVEXPASS
EVSA1	40	Sentence structure in which 'there' is the subject of the sentence Ex. There seems to be a misunderstanding. Note: all applicable codes given when appropriate. * To be included in next version.
EVTH2	41	That' clause following verb can be replaced with 'so' Ex. He believed that he could run the hundred meter dash in under nine seconds. -> He believed so.
EVTH3	42	That clause' following verb can be replaced with 'not' Ex. He believed that he could run the hundred meter dash in under nine seconds. ->_He_believed_not.
EVNOPASS	43	Allows no passive Ex. He lacks motivation. * Motivation is lacked by him.
EVEXPASS	44	Occurs only in the passive form Ex. John was said to be a good teacher. * They said him to be a good teacher. * Indicates an ill-formed construction.
EVNOPRG	45	Does not occur in the progressive tense Ex. * I am knowing him for a long time.
EVEXPRG	46	Occurs only in the progressive tense Ex. The baby is teething

Code Combinations

Note: The codes on the left must be accompanied by a code indicated on the right.

[EVIO0]->[EVDO0]
[EVOC0]->[EVDO0]
[EVDO0]->[EVDO1-9] One code from EVDO1-9
[EVSC0]->[EVC10-93] One code from EVC10-93
[EVOC0]->[EVC10-93] One code from EVC10-93
[EVSA0]->[EVSA2-6] One code from EVSA2-6
[EVSA2]-> Specific Preposition Code

Thuộc tính ngữ pháp của danh từ (Nouns: Grammatical Attributes)

Attribute	Code	Comment/Explanation
Countability	47	Countable
	48	Uncountable
	49	Uncountable noun that can be instantiated Note: Only one code is assigned. Word form and right adjacency attribute are given based on countable usage for ENUC records. All other coding for ENUC nouns is based on noun when noun is uncountable.
Collectivity	50	Collective noun (Ex. people) Note: Code is given only when applicable.
Gender	51	Referent of noun is male (Ex. man)
	52	Referent of noun is female (Ex. woman)
	53	Referent of noun is neutral (Ex. book, baby)
	54	Referent of noun can be either male or female (Ex. student, baby) Note: More than one code can be assigned when applicable.
Verb Agreement	55	Always treated as singular in subject - verb agreement
	56	Always treated as plural in subject - verb agreement
	57	Treated either singular or plural in subject -verb agreement
	58	Note: For non-count nouns, only one code is given. Verb agreement coding is given to count nouns only when the noun in the singular form takes either a plural form verb or a singular form word.
Co-occurrence with Articles	59	Does not have restrictions on the article
	60	Always takes an article
	61	Must be preceded by a definite article
	62	Must be preceded by an indefinite article
	63	Never occurs with a definite article
	64	Never occurs with an indefinite article
	65	Never occurs with an article Note: One code only is given to each noun. A count noun that is not coded is interpreted as ENWAR and a non-count noun that is not coded is interpreted as ENNOINF
Word Form Restrictions	66	Nouns occurs in the singular form only
	67	Nouns occurs in the plural form only

Thuộc tính ngữ pháp của tính từ (Adjectives: Grammatical attributes)

Code	Explanation
68	Does not occur in the positive degree form
69	Does not occur in the comparative degree form
70	Does not occur in the superlative degree form

Thuộc tính ngữ pháp của trạng từ (Adverbs: Grammatical attributes)

Code	Explanation
71	Does not occur in the positive degree form
72	Does not occur in the comparative degree form
73	Does not occur in the superlative degree form

Chức năng và vị trí của hạn định từ (Determiner: Function and Position Information)

Code	Explanation
74	May follow an indefinite article
75	May not follow a definite article Note: Code is given only when applicable.
76	May be followed by a countable singular noun
77	May be followed by a countable plural noun
78	May be followed by an uncountable noun Note: Code is given only when applicable.
79	May be followed by a noun phrase beginning with an indefinite article
80	May be followed by a noun phrase begining with a definite article

Bảng mã từ chức năng
Function Word Codes: Preposition Equivalents

Code	Preposition Equivalent	Code	Preposition Equivalent
1	concerning	22	due to
2	considering	23	out of
3	excepting	24	by means of
4	excluding	25	by way of
5	following	26	in front of
6	including	27	in respect to
7	involving	28	in terms of
8	pending	29	in view of
9	regarding	30	on account of
10	respecting	31	on behalf of
11	according to	32	on top of
12	along with	33	with regard to
13	as for	34	for the benefit of
14	as regards	35	for the purpose of
15	as to	36	for the sake of
16	based on	37	in the course of
17	based upon	38	in the matter of
18	because of	39	in the middle of
19	consisting of	40	in the way of
20	down to	41	on the basis of
21	prior to	42	up to

Function Word Codes: Be-Verb, Auxiliary Verbs, Auxiliary Verb Equivalents

Code	Function Word	Code	Function_Word
43	be	55	should
44	can	56	will
45	cannot	57	would
46	could	58	be to
47	do	59	had better
48	dare	60	have to
49	have	61	ought to
50	may	62	used to
51	might	63	be able to
52	must	64	be about to
53	need	65	be going to
54	Shall		

Function Word Codes:Coordinate Conjunctions, Corrdinate Conjunction Equivalents, Subordiate Conjunctions, Subordinate Conjunction Equivalents, and Conjunctive Adverbs

Code	Function Word	Code	Function_Word

66	After	91	only
67	Against	92	once
68	Also	93	otherwise
69	And	94	or
70	As	95	provided
71	Because	96	providing
72	Before	97	since
73	Beside	98	so
74	Besides	99	suppose
75	Both	100	supposing
76	But	101	than
77	Directly	102	that
78	Either	103	then
79	Else	104	though, although
80	Except	105	till, until
81	If	106	unless
82	Immediately	107	while
83	Instantly	108	yet
84	Lest	109	as if
85	Like	110	as though
86	Moreover	111	even if
87	Namely	112	even though
88	Neither	113	in order that
89	Nor	114	so that
90	now		

Function Words: Relative Pronouns, Interrogative Pronouns, Relative Adverbs, Interrogative Adverbs

Code	Function Word	Code	Function Word
115	how	123	whether
116	that	124	which
117	whenever	125	whichever
118	what	126	who
119	whatever	127	whoever
120	when	128	whom
121	where	129	whose
122	wherever	130	why

Other Function Words

Code	Function Word	Code	Function Word
131	never	135	that
132	not	136	to (verb infinitive)
133	more	137	not to
134	most		

Phụ lục B: Bảng mã từ loại của từ điển từ tiếng Việt

Bảng từ loại tiếng Việt

Loại	Từ loại	Mã	Ví dụ
Danh từ	- Danh từ riêng	1	Hà Nội, Hồ Chí Minh
	- Danh từ chung		
	+) Danh từ chỉ loại thể	2	cái,con,cơn,quyển,sự,cuộc, ...
	+) Danh từ chỉ đo lường		
	Chính xác	3	thước, trặc, phần, lít...
	Không chính xác	4	cục, miếng, mẩu, đoạn, mảnh, toán, dãy, tốp, mớ,...
	+) Danh từ chỉ chất liệu	5	sắt than, chì, mỡ, thịt, muối...
	+) Danh từ chỉ người		
	Chỉ quan hệ thân thuộc	6	cha, mẹ,anh,cậu,cô...
	Chỉ chức vụ, nghề nghiệp	7	bác sĩ, công nhân, thanh niên. giám đốc, giáo sư...
	+) Danh từ chỉ vật		
	Chỉ đồ vật	8	bàn, ghế, như
	Chỉ động vật	9	chó, mèo,gà,...
	Chỉ thực vật	10	cam, quýt, tre...
	+) Danh từ chỉ khái niệm trừu tượng	11	thiên nhiên, xã hội...
	- Danh từ chỉ hiện tượng thiên nhiên	12	trời,mây,gió,bão...
	- Thuật ngữ chuyên môn	13	tế bào, mạng,...
Động từ	-Động từ ngoại hướng	14	làm, ươm, mua, bán,ăn, ra, vào. lên, tăng. biến, nộp, vay,...
	-Động từ nội hướng	15	thức, ngủ, cười, đùa, nhìn, nằm, bò...
	-Động từ gây khiến	16	làm, để, bắt...
	-Động từ xuất hiện, tồn lại, tiêu tan	17	còn, có, hết, mất, xuất hiện. nảy, mọc, nổi...
	-Động từ chỉ trạng thái tiếp thu	18	bị, được, chịu
	-Động từ cảm nghĩ, nói năng	19	biết, thấy, khen, chê, bảo, nhận định, tin tưởng, ...
	-Động từ tình thái	20	muốn, toan, định, nên, dám phải, chịu,...
	-Động từ quan hệ	21	là, làm, hoá, giống, khác
Tính từ	Tính từ chỉ đặc điểm bên ngoài của sự vật		
	+) Màu sắc	22	xanh, đỏ, tím,...
	+) Hình thể	23	to nhỏ, tròn, vuông...
	+) Dung lượng	24	nhẹ, nặng, căng...
	+) Kích thước	25	dài, ngắn, cao, thấp,
	-Tính từ chỉ đặc tính bên trong và trạng thái	26	tốt, xấu, hiền, to gan, nhanh ...
	-Tính từ miêu tả mức độ	27	đầy, vơi, nhiều,ít,dày, thưa...
Số từ	- Số từ chỉ số lượng chính xác	28	một, hai phần ba,...
	- Số từ chỉ số lượng áng chừng	29	mấy, vài ba, dăm, một vài,...
	- Số từ chỉ số thứ tự	30	nhất, nhì, thứ mười,...

Đại từ	- Đại từ nhân xưng - Đại từ chỉ định sự vật - Đại từ chỉ định vị trí không gian, thời gian - Đại từ chỉ trạng thái - Đại từ chỉ số lượng - Đại từ để hỏi +) Hỏi về sự vật +) Hỏi về vị trí không gian +) Hỏi về hoạt động, trạng thái +) Hỏi về số lượng	31 32 33 34 35 36 37 38 39	tôi, chúng tôi, nó,... này, nọ, kia, ấy, ... đây,đấy, đó. kia, nay, bây giờ thế, vậy,... bấy nhiêu, hết thảy, cả, tất cả,... ai, chi, gì, nào đâu nào, bao giờ,... thế nào, sao,... mấy, bao nhiêu,...
Phó từ (trạng từ)	- Biểu thị số lượng toàn thể hay riêng lẻ - Biểu thị ý nghĩa thời gian - Biểu thị ý nghĩa phủ định - Biểu thị ý nghĩa yêu cầu, sai khiến, khích lệ - Biểu thị ý nghĩa đồng nhất hay liên tục - Biểu thị mức độ - Biểu thị sự diễn biến - Biểu thị sự kết thúc hành động	40 41 42 43 44 45 46 47	những, cái, mọi, mỗi, từng, ... đang,đương, sẽ, đã, vừa, mới,... không, chưa, chẳng,... hãy, đi, đừng, chớ cũng, đều, vẫn, còn, lại, cứ .. rất, khá, hơi, khí, quá, lắm,... càng, lại, luôn,mãi,bèn,bỗng,. . xong, rồi
Quan hệ từ (kết từ)	- Quan hệ từ chính phụ - Quan hệ từ liên hợp +)Song song +)Phụ thuộc +)Sau động từ cảm nghĩ, nói năng +)Từ nối, cặp từ nối +) Lựa chọn	48 49 50 51 52 53	của, bằng, với, về, đến, hỏi, bởi, để, và, cùng, với,... rằng, là,... thì, mà...thì,... vì... do, tuy... nhưng.. . hay, hoặc,....
Trợ từ	- Trợ từ cho từ - Trợ từ cho cụm từ - Trợ từ cho câu +) Nhấn mạnh +) Hoài nghi +) Ngạc nhiên +) Cầu mong +) Dứt khoát +) Nũng nịu	54 55 56 57 58 59 60 61	thì, là, cả,... chính,tự,những,cái,thì,cả, ngay kia, đâu, đấy,... chăng, hử,... nhỉ, ư,... đi nào, thôi, với,... đâu, đấy cơ, kia, nhé
Thán từ	- Thán từ làm tiếng gọi - Thán từ làm tiếng đáp - Thán từ làm tiếng than	62 63 64	hỡi, đi, ê, này vâng, dạ, ừ, phải ôi, chao, khiến, trời, đất....
Cụm động từ	Cụm từ với động từ làm trung tâm	65	chạy thục mang
Cụm danh từ	Cụm từ với danh từ làm trung tâm	66	xe máy
Cụm tính từ	Cụm từ với tính từ làm trung tâm	67	đỏ hon hỏn

Thuộc tính ngữ pháp của động từ

Mã	Giải thích

1 2 3 4 5 6 7	-Không cần kết hợp với phụ tố: động từ nội hướng - Phải kết hợp với phụ tố: +) Phụ tố là danh từ: động từ ngoại hướng +) Phụ tố là cụm CV: động từ gây khiến, cảm nghĩ, nói năng +) Phụ tố là động từ: động từ tình thái - Có hoặc không kết hợp với phụ tố: động từ xuất hiện, tồn tại, tiêu tan. -Làm vị ngữ trong câu

Thuộc tính ngữ pháp của danh từ

Mã	Giải thích

8	-Kết hợp với danh từ
9	+)Danh từ chỉ quan hệ xã hội, gia đình: danh từ riêng
10	+)Danh từ khác: danh từ chỉ loại thể, đo lường
11	+)Danh từ chỉ chất liệu: danh từ chỉ đo lường
12	+)Danh từ chỉ đơn vị đo lường: danh từ chỉ chất liệu
13	-Không kết hợp với danh từ
14	+)Chỉ loại thể: danh từ chỉ chất liệu, khái niệm trừu tượng (ít)
15	-Kết hợp với đại từ chỉ định: danh từ chỉ loại thể, chất liệu, thời gian, người, phương hướng, vị trí (trừ đông, tây, nam, bắc), vật, khái niệm trừu tượng.
16	- Không kết hợp với đại từ chỉ định: danh từ riêng
17	- Kết hợp với số từ: danh từ chỉ đơn vị đo lường, thời gian, chỉ người, vật, khái niệm trừu tượng.
18	- Không kết hợp với số từ: danh từ riêng, chỉ đơn vị đo lường không chính xác, phương hướng vị trí (trừ phía, phương, bên, hướng).
19	- Kết hợp với đại từ chỉ số lượng: danh từ chỉ người (chức vụ nghề nghiệp), chỉ vật, loại thể, chất liệu, khái niệm trừu tượng.
20	- Không kết hợp vớt đại từ chỉ số lượng: danh từ riêng, đo lường, thời gian, người (quan hệ thân thuộc).
21	- Kết hợp với định từ: danh từ chỉ loại thể, chỉ người, vật. .
22	- Kết hợp với tính từ: danh từ chỉ người, vật, chất liệu, khái niệm trừu tượng.
23	- Làm chủ ngữ

Thuộc tính ngữ pháp của tính từ

Mã	Giải thích
24	-Kết hợp với trạng từ: tất cả trừ “công", "tư", "riêng", chung"
25	- Kết hợp với danh từ: tính từ chỉ đặc điểm bên ngoài, đặc điểm bên trong và trạng thái.
26	- Kết hợp với động từ thành cụm động từ: tính từ chỉ đặc điểm bên ngoài
27	- Kết hợp với tính từ chỉ tính chất: tính từ chỉ màu sắc
28	- Làm vị ngữ: trừ “công”, "tư", "riêng", "chung" .

Thuộc tính ngữ pháp của số từ

Mã	Giải thích
29	- Số từ làm tiền tố phụ trong cụm danh từ
30	- Số từ làm vị ngữ trong câu: số thứ tự, số lượng chính xác+"là"
31	- Số từ chỉ số lượng phỏng chừng
32	- Số từ chỉ số lượng tượng trưng: ba, trăm. nghìn....

Thuộc tính ngữ pháp của đại từ

Mã	Giải thích

33 34 35 36	-Đại từ làm chủ ngữ: đại từ nhân xưng, chỉ định sự vật (ấy, này), để hỏi -Đại từ làm định tố: chỉ định sự vật, để hỏi -Đại từ thay thế cho một đơn vị ngữ pháp: chỉ định sự vật trạng thái -Đại từ thay thế cho số từ chỉ số lượng: đại từ chỉ số lượng

Thông tin từ chức năng của phó từ

Mã	Giải thính
1 2 3 4 5	-Đi kèm với danh từ: phó từ biểu thị số lượng toàn thể hay riêng lẻ -Đi kèm với động từ, tính từ: phó từ biểu thị ý nghĩa trung gian, phủ định, diễn biến. - Trước động từ: phó từ biểu thị ý nghĩa yêu cầu, sai khiến, đồng nhất hay liên tục. - Trước tính từ: phó từ chỉ mức độ, đồng nhất hay liên tục - Sau động từ: phó từ biểu thị kết thúc hành động, phó từ biểu thị diễn biến.

Thông tin từ chức năng của quan hệ từ

Mã	Giải thích
6 7 8 9 10	- Trong cụm danh từ: của, bằng (chính phụ) - Trong cụm động từ: của (chính phụ), với, đến, về - Trong cụm tính từ: về (chính phụ) + động từ: chỉ mục đích, đối tượng: ở đâu, để, cho, đến - Trong câu ghép: quan hệ từ liên hợp

Thông tin từ chức năng của trợ từ

Mã	Giải thích
11 12 13	- Trợ từ cho từ - Trợ từ cho cụm từ - Trợ từ cho câu

Thông tin từ chức năng của thán từ

Mã	Giải thích
14	Thành phần phụ biệt lập trong câu

Cách sử dụng

Cách sử dụng	Mã	Giải thích	Ví dụ
Abbreviation	1	Hình thức rút gọn của một từ	ĐHQG
Slang	2	Không phù hợp với khi nói ở nơi công cộng hoặc trong văn bản	mổ

Phụ lục C: Các bài báo liên quan

- Nghiem Anh Tuan, Ho Chi Kien, Ho Tu Bao. Issues in Construction of a Vietnamese Machine Tractable Dictionary. in *Proceeding of APF*, Japan, 2002.