VIỆN CÔNG NGHỆ THÔNG TIN

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

NGHIÊN CỨU PHÁT TRIỂN PHẦN MỀM DỊCH MÁY VIỆT-ANH

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

6455-3
07/8/2007

HÀ NỘI- 2004

CHƯƠNG TRÌNH KH.01

ĐỀ TÀI MÃ SỐ KH01-03

NGHIÊN CỨU PHÁT TRIỂN CÔNG NGHỆ NHẬN DẠNG, TỔNG HỢP VÀ XỬ LÝ NGÔN NGỮ TIẾNG VIỆT

NĂM 2001-2003

CẤP QUẢN LÝ: Nhà nước
CƠ QUAN CHỦ TRÌ: Viện Công nghệ thông tin
CƠ QUAN THỰC HIỆN:

- Viện Công nghệ thông tin
- Trung tâm Ngữ âm học thực nghiệm – Viện Ngôn ngữ học
- Trung tâm kỹ thuật – Thông tấn xã Việt Nam
- Trung tâm Công nghệ Vi điện tử và Tin học – Viện Ứng dụng Công nghệ
- CSLU – Center of spoken language understanding, Viện sau đại học Oregon, Hoa kỳ
- Khoa Toán – Cơ – Tin học, Đại học Tự nhiên Hà nội

CHỦ NHIỆM ĐỀ TÀI: GS. TSKH. Bạch Hưng Khang
NHÁNH ĐỀ TÀI :
NGHIÊN CỨU PHÁT TRIỂN PHẦN MỀM DỊCH MÁY VIỆT–ANH

HÀ NỘI 2003

Tên Đề tài nhánh :

Nghiên cứu phát triển Phần mềm Dịch máy Việt–Anh

Nơi thực hiện :

Trung tâm CN Vi điện tử và Tin học, Viện Ứng dụng Công nghệ

Thời gian thực hiện :

2001 – 2003

Yêu cầu:

1. Bộ phân tích cho phép xử lý các tình huống phi ngữ cảnh và phụ thuộc ngữ cảnh (trong phạm vi hạn định - scope dependent)
2. Tốc độ biên dịch tự động đạt không dưới 5.000 từ / phút (tương đương với 10 trang A4).
3. Chất lượng dich thuật có thể xem hiểu những văn bản tiếng Việt đúng văn phạm (đối với những người hiểu tiếng Anh và không biết tiếng Việt).
4. Hệ văn phạm hình thức tiếng Việt bao gồm các yếu tố chính của luật hành văn tiếngViệt.
5. Kho mẫu câu tiếng Việt từ nhiều nguồn khác nhau và bao gồm những đặc trưng chính của các mẫu câu tiếng Việt thông thường.
6. Cơ sở tri thức bao gồm:
   - 5.000 qui tắc văn phạm tiếng Việt và dịch Việt - Anh.
   - 150.000 đơn vị từ vựng Việt – Anh.
   - 300.000 - 1.000.000 mẫu câu tiếng Việt thông dụng.

Các kết quả thực hiện:

I. LÝ THUYẾT VÀ CÔNG NGHỆ:

1. Đề xuất ***văn phạm định biên*** (*bound controlled grammar*) – một dạng mở rộng của mô hình văn phạm phi ngữ cảnh, chỉ ra một số tính chất của văn phạm, trong đó chứng minh được rằng lớp ngôn ngữ định biên là ***bao đóng*** của lớp ngôn ngữ phi ngữ cảnh đối với phép giao. Điều đó có nghĩa rằng văn phạm định biên là sự mở rộng đủ và tối thiểu cho lớp ngôn ngữ phi ngữ cảnh để thành một tập hợp đóng kín đối với phép hợp và phép giao. Ý nghĩa của văn phạm định biên là ở chỗ các kết quả lý thuyết và giải thuật trên lớp ngôn ngữ phi ngữ cảnh đều có thể áp dụng cho ngôn ngữ định biên. Nói riêng, các giải thuật phân tích văn phạm phi ngữ cảnh cũng như độ phức tạp của chúng được giữ nguyên gần như hoàn toàn trong văn phạm định biên.

2. Đề xuất ***văn phạm cảm ngữ đoạn*** (*phrase sensitive grammar*) – một phát triển tiếp tục của văn phạm định biên cho phép mô tả được nhiều tính chất phụ thuộc ngữ cảnh của ngôn ngữ tự nhiên, đặc biệt, đề xuất khái niệm ngữ đoạn như một yếu tố ràng buộc trọng tâm trong định nghĩa các cấu trúc của ngôn ngữ.

   Một số tính chất của văn phạm:

   - Các phần tử từ vựng, cú pháp, ngữ nghĩa và tập quy tắc được tổ chức thành hệ phân cấp (*dàn đại số*)
   - Đưa vào khái niệm "***phần tử được đánh dấu***" để thể hiện những ràng buộc ngữ nghĩa trong quy tắc văn phạm, đặc biệt, để biểu diễn các nút có số nhánh biến thiên trong cây phân cấp ngữ nghĩa. Bộ phân tích không dựng cây cú pháp mà dựng mô hình biểu diễn bên trong (*cây phân cấp ngữ nghĩa*) của câu văn trên cơ sở áp dụng các *quy tắc cảm ngữ đoạn*.

3. Đề xuất phương pháp giải quyết nhập nhằng ứng dụng trong xử lý ngôn ngữ tự nhiên dựa trên sự phân cấp của hệ luật sinh sử dụng một mô hình logic mới, trong đó miền giá trị không phải là nhị phân (*true, false* – như trong logic cổ điển) hay một đoạn liên tục (các số thực từ 0 đến 1 – như trong logic mờ) mà là một dàn đại số. Giải pháp đề xuất một mô hình hình thức cho sự "***lập luận theo lẽ thường"*** (*common-sense reasoning*) đối với tri thức ngôn ngữ.

   Mô hình phân cấp ngữ nghĩa áp dụng trong văn phạm cảm ngữ đoạn cho ta một công cụ để mô tả những quy tắc ngôn ngữ, vốn rất khó diễn đạt bằng toán học. Với cách tiếp cận được đề xuất, mỗi luật sinh đều có một phạm vi tác dụng trong khuôn khổ một hệ phân cấp miền tác dụng của tập luật. Tập các miền tác dụng của bộ luật tạo nên một ***phủ*** trên toàn bộ ngôn ngữ.

   Những kết quả nghiên cứu này tạo thành nền tảng để xây dựng một giải pháp dịch máy liên ngữ khả thi (*hiện đang được phát triển tại Viện Ứng dụng Công nghệ*). Cách tiếp cận có các đặc trưng cơ bản sau:

   - Bộ phân tích không dựng cây cú pháp mà dựng mô hình biểu diễn bên trong (*cây phân cấp ngữ nghĩa*) của câu văn
   - Bước *Tổng hợp* là quá trình đơn ngữ, được thực hiện hoàn toàn độc lập với quá trình *Phân tích*. Vì vậy, trong mô hình dịch máy được đề xuất, công đoạn tổng hợp văn bản khó hơn nhiều so với khâu phân tích, và văn bản được sản sinh ra sẽ tự nhiên, bản ngữ hơn, không phụ thuộc vào cách đặt câu của văn bản gốc.

4. Phát triển giải thuật phân tích văn phạm cảm ngữ đoạn.

   Xây dựng mô hình xử lý nhập nhằng cho kho ngữ liệu được tổ chức theo mô hình phân cấp dựa vào logic trên dàn và văn phạm cảm ngữ đoạn. Thuật toán phân tích theo sơ đồ dưới lên và từ phải sang trái (*bottom-up*

right-most analysis) dựng cây phân tích ngữ nghĩa không phụ thuộc ngôn ngữ và họ các bộ giá trị trạng thái liên ngôn ngữ.

5. Phát triển giải thuật tổng hợp văn phạm cảm ngữ đoạn.

   Xây dựng sơ đồ tổng hợp văn bản

6. Công trình.

   Một số kết quả nghiên cứu của đề tài đã được trình bày trên các hội nghị khoa học và đăng tải trên các tạp chí chuyên ngành:

   - Một báo cáo khoa học tại Hội thảo quốc gia về Nghiên cứu và Phát triển ICT-RDA, Hà Nội, 3, 2003.
   - Hai báo cáo khoa học tại Hội nghị toán học toàn quốc lần thứ 6, Huế, 09, 2002.
   - Hai bài báo đăng trên Tạp chí Bưu chính Viễn thông, Chuyên san số 8 và 10, 2002.
   - Một báo cáo khoa học tại Hội thảo Quốc gia Lần thứ 6 – Một số Vấn đề chọn lọc của Công nghệ Thông tin và Truyền thông, Chủ đề : Xử lý Ngôn ngữ và Đa phương tiện, (*Languague Processing and Multimedia*), Thái Nguyên, 8, 2003.
   - Một báo cáo khoa học tại Hội thảo quốc gia về Nghiên cứu và Phát triển Khoa học cơ bản, Hà Nội, 10, 2003.

II. THỰC HÀNH:

1. Ứng dụng một phần các kết quả lý thuyết và công nghệ được phát triển vào phần mềm dịch máy
2. Ứng dụng một số heuristics nhằm cải thiện tốc độ cho giải thuật phân tích văn phạm và biên dịch văn bản
3. Xây dựng hệ phân cấp từ loại tiếng Việt để đưa vào cơ sở tri thức tiếng Việt trên cơ sở mô hình ngữ nghĩa chung cho Hệ thống từ loại tiếng Việt, áp dụng lý thuyết dàn (*lattice*) làm mô hình ngữ nghĩa cho hệ thống từ loại tiếng Việt.
4. Khảo sát trên 400.000 mẫu câu song ngữ Việt-Anh thông dụng.
5. Xây dựng cơ sở tri thức dịch máy Anh Việt – Việt Anh bao gồm:
   - Trên 7.600 quy tắc văn phạm và biên dịch Anh-Việt và Việt-Anh
   - Trên 230.000 đơn vị từ vựng dịch Anh-Việt
   - Trên 260.000 đơn vị từ vựng dịch Việt-Anh

III. ỨNG DỤNG THỰC TIỄN

1. Đang thử nghiệm và tiếp tục hoàn thiện cơ sở tri thức để đưa ra sử dụng rộng rãi trong nửa đầu năm 2004 (EVTRAN 2.5 dịch hai chiều Anh-Việt, Việt-Anh). Một số đặc điểm của phần mềm:

- Dịch hai chiều Anh-Việt và Việt-Anh
- Chương trình tự động đoán nhận ngôn ngữ nguồn
- Có tính năng đa ngữ, có thể dễ dàng đưa một cặp ngôn ngữ mới vào hệ thống để biên dịch qua lại giữa hai ngôn ngữ mà không cần phải lập trình.
- Có khả năng vận dụng tri thức ngôn ngữ trong phân tích : kho ngữ liệu càng lớn thì tốc độ phân tích câu – và tương ứng – tốc độ biên dịch văn bản càng cao, trái với các giải thuật phân tích đơn định (*chẳng hạn đối với giải thuật Early thì thời gian phân tích tỷ lệ nghịch với bình phương kích thước của bộ quy tắc văn phạm*).
- Có các công cụ cập nhật tri thức ngôn ngữ và biểu diễn trực quan cây cú pháp để hỗ trợ việc hiệu chỉnh cơ sở tri thức
- Có kèm theo một số từ điển tra cứu thông dụng (Computing Dictionary, Thesaurus, Từ điển Anh-Việt và Việt-Anh, Oxford Advanced Learner's Encyclopedic Dictionary, Webster's Dictionary,...) để tiện việc cập nhật dữ liệu ngôn ngữ

2. Tiếp tục tích hợp những kết quả lý thuyết và công nghệ đã đạt được (trong khuôn khổ nghiên cứu của đề tài) cũng như bổ sung và hiệu chỉnh cơ sở tri thức ngôn ngữ vào sản phẩm để nâng cao chất lượng trong phiên bản tiếp theo (*dự kiến hoàn tất trong năm 2005*) và tiến tới bổ sung các ngôn ngữ khác vào hệ thống.

Báo cáo khoa học gồm 5 phần.

Phần I tổng quan các cách tiếp cận dịch máy hiện tại trên thế giới. Phần II giới thiệu những kết quả nghiên cứu của nhánh đề tài về một mô hình văn phạm mới, được sử dụng như công cụ để mô tả tri thức ngôn ngữ và giải quyết một số kiểu nhập nhằng. Văn phạm này cũng đặt cơ sở cho một giải pháp dịch máy liên ngữ mới, nội dung chi tiết được trình bày trong phần III. Các phần IV và V của báo cáo giới thiệu những kỹ thuật triển khai thực hành của đề tài.

I. DỊCH MÁY: MỘT SỐ TRÀO LƯU HIỆN NAY.

I.1. VĂN PHẠM VÀ PHÂN TÍCH CÚ PHÁP I-2
I.1.1. NGÔN NGỮ HÌNH THỨC VÀ VĂN PHẠM SINH I-3
I.1.2. MÔ HÌNH VĂN PHẠM DỰA TRÊN SỰ THỐNG NHẤT I-6

I.2. ÁP DỤNG VĂN PHẠM VÀ NHỮNG TRỞ NGẠI I-6
I.2.1. SỰ PHÂN CẤP KHÁI NIỆM. I-7
I.2.2. MỐI LIÊN HỆ GIỮA CÁC BỘ PHẬN TRONG CÂU. I-7
I.2.3. MỐI LIÊN HỆ GIỮA CÁC TẦNG CẤU TRÚC TRONG CÂU I-8
I.2.4. RÀNG BUỘC VĂN PHẠM VÀ THÔNG TIN DẪN XUẤT I-11
I.2.5. VĂN PHẠM CẢM NGỮ CẢNH YẾU I-11

I.3. CÁC KHUYNH HƯỚNG TRONG DỊCH MÁY I-15
I.3.1. CÁCH TIẾP CẬN DỰA THEO LUẬT I-15
I.3.2. PHƯƠNG PHÁP DỰA VÀO KHO NGỮ LIỆU I-18
I.3.3. MỘT SỐ HỆ DỊCH MÁY LIÊN NGỮ I-19
I.3.3.1. Dự án UNITRAN của MIT I-19
I.3.3.2. Dự án Dịch máy đa ngữ tại CICC. I-21
I.3.3.3. Dự án KANT của Trường đại học Carnegie Mellon I-21
I.3.4. CÁC PHƯƠNG HƯỚNG MỚI I-23

I.4. KẾT LUẬN I-23

Phần này trình bày một số khía cạnh của xử lý ngôn ngữ tự nhiên, các khía cạnh ngữ pháp, ngữ nghĩa học của ngôn ngữ; các phương hướng nghiên cứu và hiện trạng của lĩnh vực dịch máy như một bộ phận quan trọng của xử lý ngôn ngữ tự nhiên.

I.1. VĂN PHẠM VÀ PHÂN TÍCH CÚ PHÁP.

Hệ thống xử lý ngôn ngữ tự nhiên giữ một vai trò cốt yếu trong giao tiếp giữa con người với nhau hay với máy móc. Xử lý ngôn ngữ tự nhiên bao gồm nhận dạng tiếng nói, hiểu và sản sinh ngôn ngữ. Các hệ thống xử lý văn bản và biên dịch các thông báo rất hữu ích trong việc trích lọc thông tin từ kho ngữ liệu văn bản và tổ chức chúng thành dữ liệu theo nhiều khuôn dạng khác nhau để sử dụng về sau.

Xử lý đa ngôn ngữ đòi hỏi phải đi sâu vào các vấn đề đa ngôn ngữ như cung cấp thiết bị hỗ trợ biên dịch văn bản cũng như phiên dịch (*dịch nói*) ở một số lĩnh vực nhất định. Nghiên cứu về xử lý ngôn ngữ tự nhiên là nghiên cứu mô hình toán học về cấu trúc và chức năng của ngôn ngữ, sử dụng và sự tiếp nhận ngôn ngữ : cú pháp, ngữ nghĩa học, ngữ dụng học (nghĩa là một số khía cạnh nhất định trong mối quan hệ giữa người nói và người nghe, hay giữa người sử dụng và hệ thống trong hệ thống xử lý ngôn ngữ tự nhiên), cũng như các khía cạnh về mặt văn bản của ngôn ngữ. Đây là những nghiên cứu liên bộ môn và có liên quan đến một số chuyên ngành của khoa học máy tính bao gồm trí tuệ nhân tạo, ngôn ngữ học, logic học và tâm lý học.

Ngôn ngữ có cấu trúc tôn ti theo nhiều cấp độ khác nhau, đặc biệt ở cấp độ câu. Hầu hết mọi hệ thống xử lý ngôn ngữ tự nhiên đều có một hệ văn phạm và phân tích cú pháp tương ứng. *Văn phạm* là những đúc kết hữu hạn của một số lượng câu hầu như vô hạn, còn *phân tích cú pháp* là thuật toán để đưa ra một hay nhiều sự miêu tả cấu trúc cho câu theo văn phạm nếu câu đó có thể phân tích theo những đăc điểm ngữ pháp. Mô tả cấu trúc là sự ghi lại lịch sử nguồn gốc hình thành của câu theo văn phạm. Mô tả cấu trúc được xem là có vai trò quan trọng cho những nghiên cứu sâu hơn như hiểu văn bản hay dịch ngữ nghĩa[1].

[1] Tuy nhiên, có thể thấy rằng chính lịch sử áp dụng quy tắc trong văn phạm sinh lại cản trở việc nhận thức cấu trúc ngữ nghĩa (chi tiết trong phần II và III)

I.1.1. NGÔN NGỮ HÌNH THỨC VÀ VĂN PHẠM SINH

Vào cuối những năm 50, các kết quả nghiên cứu của nhà ngôn ngữ học Noam Chomsky [1] đã có ảnh hưởng sâu rộng đến toàn bộ lĩnh vực nghiên cứu về cú pháp. Nền tảng của những kết quả đó là Lý thuyết về ngôn ngữ hình thức, đặt nền móng cho khoa học máy tính lý thuyết và là khởi đầu cho việc xử lý ngôn ngữ tự nhiên. Ông đã xây dựng một mô hình hình thức mới về miêu tả văn phạm và đã phân tích một bộ phận đáng kể của tiếng Anh bằng các công cụ của mô hình mới này.

Nội dung quan trọng nhất trong lý thuyết của Chomsky là mô hình văn phạm sinh, trong đó những luận điểm chính bao gồm:

- Giả thuyết rằng cấu trúc ngôn ngữ phải đủ nhỏ để dễ dàng kiểm tra.
- Đối tượng nghiên cứu chính là hệ tri thức ẩn chứa đằng sau việc sử dụng ngôn ngữ.
- Có một nền tảng sinh học trong khả năng tiếp thụ tri thức ngôn ngữ của con người.

Chomsky cho rằng ngôn ngữ, đặc biệt là tổ chức văn phạm của nó có thể soi sáng cho chúng ta cấu trúc tư duy của con người. Theo ông, "*thực tế đáng chú ý nhất của ngôn ngữ loài người là sự tương phản kỳ lạ giữa sự phức tạp hiển nhiên của nó với sự dễ dàng mà trẻ em học tiếng*". Cấu trúc của bất kỳ ngôn ngữ tự nhiên nào cũng phức tạp hơn nhiều so với mọi ngôn ngữ nhân tạo hay những hệ thống toán học cao siêu. Nhưng lạ thay, học ngôn ngữ lập trình hay học toán đòi hỏi phải kinh qua những khóa đào tạo căng thẳng (mà không ít người rốt cuộc vẫn không tiếp thu được). Trong khi đó đứa trẻ lên ba đã gần như thành thạo ít nhất là một thứ tiếng.

Để giải thích nghịch lý này, Chomsky cho rằng phần lớn sự phức tạp của ngôn ngữ thì không cần phải học, vì con người khi sinh ra đã biết chúng; nghĩa là trong não người đã sẵn có khả năng học một loại ngôn ngữ nhất định. Khái quát hơn, ông cho rằng tư duy bẩm sinh của con người đã được môđun hóa cao độ. Nghĩa là chúng ta có những cơ quan tư duy chuyên dụng được thiết kế để thực hiện những loại bài toán đặc biệt theo những cách thức đặc biệt. Cơ quan ngôn ngữ (theo quan điểm của Chomsky, chứa một số môđun con tương đối độc lập) là đặc trưng riêng của loài người. Mọi người đều có tư duy ngôn ngữ, và không loài động vật nào có khả năng học bất cứ thứ gì tựa như tiếng người.

Một hệ quả từ giả thuyết về tri thức ngôn ngữ bẩm sinh của loài người là "*hầu hết các cấu trúc là chung cho mọi ngôn ngữ*". Thực tế là trẻ em nhanh chóng học nói thứ tiếng mà chúng tiếp xúc, không phụ thuộc vào nguồn gốc của bố mẹ chúng. Vì vậy tri thức ngôn ngữ bẩm sinh, nếu có, thì

chung cho mọi ngôn ngữ. Nếu tri thức này bao gồm các nguyên lý của cấu trúc văn phạm (theo như Chomsky quan niệm), thì "*mọi ngôn ngữ đều tương tự nhau*". Ông thậm chí sử dụng thuật ngữ ***Văn phạm phổ quát*** *(Universal Grammar)* để chỉ tri thức ngôn ngữ bẩm sinh trong mỗi người.[1]

Chomsky đã đưa ra hệ phân cấp các văn phạm và nghiên cứu sự tương ứng ngôn ngữ của chúng, trong đó đặc biệt quan trọng là văn phạm phi ngữ cảnh (*Context-Free Grammar*).

Văn phạm phi ngữ cảnh theo Chomsky bao gồm :

- Một tập hữu hạn các ***biến trung gian*** (ví dụ: ***C***: câu, ***DN***: danh ngữ***, ĐT***: động từ, ***TrT***: trạng từ),
- Một tập hữu hạn các ***từ cuối*** (ví dụ: ***Bích Thủy*** - tên riêng; ***ô mai*** – danh từ; ***thích*** – động từ; ***cực kỳ*** - trạng từ),
- Một tập ***quy tắc*** phân tích cấu trúc A thành ω, khi A là một ***biến trung gian*** còn ω là một chuỗi các ***từ cuối*** và ***biến trung gian***.
- S là một ***biến trung gian*** đặc biệt gọi là ký hiệu xuất phát.

Trên Hình 1 đưa ra một ví dụ đơn giản của văn phạm phi ngữ cảnh. Các quy tắc phân tích được gọi là các quy tắc cú pháp. Khởi đầu bắt nguồn từ S - ký hiệu xuất phát. Bằng việc áp dụng những quy tắc lên S, S được phân tích thành chuỗi các ***biến trung gian*** và các ***từ cuối***. Các ***biến trung gian*** mới lại được phân tích lại theo những quy tắc của chúng cho đến khi không thể phân tích thêm được nữa. Dễ dàng nhận thấy rằng câu: ***"Bích Thủy thích ô mai cực kỳ"*** có thể sản sinh ra từ văn phạm. Trên hình vẽ 1, sơ đồ thể hiện sự mô tả cấu trúc văn phạm của câu hình thành bởi những thành tố từ theo sơ đồ. Bắt đầu từ ký hiệu S. Ký hiệu này được phân tích thành chuỗi DN (danh ngữ) ĐN (động ngữ). Hai ký hiệu này lại được phân tích lại theo một thứ tự nào đó lần lượt thành chuỗi ***Bích Thủy*** và ĐN (động ngữ) TrT (trạng từ). Ký hiệu ĐN (động ngữ) lại được phân tích thành chuỗi ĐT (động từ) DN (danh ngữ); TrT (trạng từ) được phân tích thành ***cực kỳ.*** Cuối cùng, ĐT (động từ) được phân tích thành ***thích*** và DN (danh ngữ) được phân tích thành ***ô mai.*** Sơ đồ trên hình 1 là kết quả của sự phân tích này.

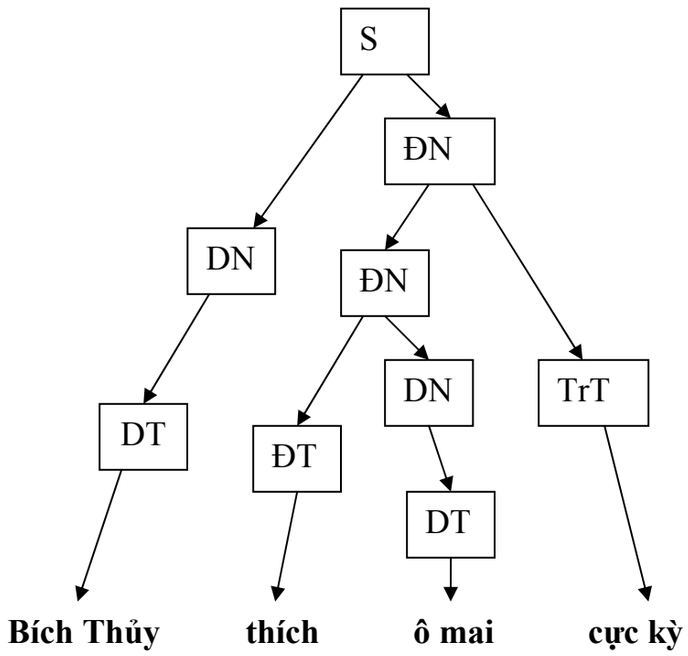
Quy tắc Cú pháp:

$S \rightarrow DN\ ĐN$	$DT \rightarrow$ ***Bích Thủy***
$ĐN \rightarrow ĐN\ TrT$	$DT \rightarrow$ ***ô mai***

[1] Cũng có người không chấp nhận quan điểm này. Chẳng hạn, trong bài "Một số biểu hiện của cách nhìn Âu châu đối với cấu trúc tiếng Việt" [40], có câu : "*Vả lại đến những năm 90 của thế kỷ không còn có ai mơ hồ đến mức tưởng rằng có những phạm trù ngữ pháp phổ quát cho ngôn ngữ toàn nhân loại*".? Chúng tôi cho rằng tất cả những ý tưởng của Chomsky vẫn giữ nguyên giá trị cho đến ngày nay. Nội dung phần này hoàn toàn không nhằm phản bác những giả thuyết về tri thức ngôn ngữ bẩm sinh và khái niệm *Văn phạm phổ quát* của Chomsky. Ở đây chỉ đưa ra những nghi vấn về cách mà chúng ta hiện đang vận dụng mô hình này trong thực tế. Giải pháp cụ thể cho một số vấn đề đặt ra được trình bày trong phần sau.

ĐN → ĐT DN
DN → DT

*ĐT → **thích***
*TrT → **cực kỳ***



Hình 1: Mô tả cấu trúc câu

Văn phạm chính quy giống văn phạm phi ngữ cảnh ngoại trừ quy tắc phân tích chỉ có các dạng A → **a**B hoặc A → **a**, trong đó A và B là biến trung gian, **a** là từ cuối. Người ta cho rằng văn phạm chính quy là quá thô sơ để mô tả cấu trúc ngôn ngữ tự nhiên. Văn phạm phi ngữ cảnh thường được chấp nhận trong thực tế.

Văn phạm cảm ngữ cảnh cũng giống văn phạm phi ngữ cảnh nhưng quy tắc phân tích biến trung gian phụ thuộc vào ngữ cảnh xung quanh cấu trúc, trong khi quy tắc phân tích văn phạm phi ngữ cảnh là không phụ thuộc vào ngữ cảnh. Văn phạm cảm ngữ cảnh có vẻ đầy đủ hơn khi mô tả cấu trúc ngôn ngữ tự nhiên. Tuy nhiên, toàn bộ lớp văn phạm cảm ngữ cảnh lại tỏ ra quá phức tạp để có thể áp dụng trong thực tế phân tích câu.

Có nhiều nghiên cứu xung quanh việc xây dựng những mô hình văn phạm mạnh hơn văn phạm phi ngữ cảnh nhưng thuận tiện hoặc đủ chuyên biệt để có thể áp dụng thực tế. Trong những năm 80 của thế kỷ 20 người ta đã đưa ra một số mở rộng văn phạm phi ngữ cảnh, nhúng thêm những ràng buộc hay những thỏa thuận về ngữ cảnh trong định nghĩa quy tắc. Những văn phạm được xây dựng theo xu hướng này được gọi chung là văn phạm

dựa trên sự thống nhất và ràng buộc (*unification- and constraint-based grammars*)

I.1.2. MÔ HÌNH VĂN PHẠM DỰA TRÊN SỰ THỐNG NHẤT

Một cấu trúc đặc biệt bao gồm các cặp mang giá trị thuộc ngữ khi một giá trị có thể là hạt nhân hay mang cấu trúc đặc trưng khác. Cấu trúc đặc trưng này có một thuộc tính thống nhất, giá trị của nó là những thuộc tính khác, (chẳng hạn sự phù hợp về số và ngôi). Quy tắc phân tích câu phi ngữ cảnh được coi như cách kết hợp chuỗi để thành câu.

Thao tác cơ bản trong kết hợp các cấu trúc đặc trưng được gọi là sự thống nhất. Với hai cấu trúc A và B, bằng cách kết hợp chúng, ta có thể tạo ra cấu trúc C mang đầy đủ những thông tin của A và B. Tất nhiên nếu A và B mang những thông tin mâu thuẫn với nhau, chúng sẽ không thể kết hợp với nhau được. Trong kiểu văn phạm văn phạm phi ngữ cảnh dựa vào sự thống nhất, văn phạm phi ngữ cảnh đóng vai trò như một bộ khung cho sự kết hợp chuỗi. Đối tượng cho sự vận dụng văn phạm là các cấu trúc đặc thù. Các cấu trúc đặc thù này được kết hợp bởi sự thống nhất đã nói ở trên. Vì vậy ở kiểu văn phạm thống nhất này, văn phạm tạo ra các chuỗi, còn sự thống nhất của các cấu trúc đặc thù phù hợp (bắt đầu là các cấu trúc đặc thù đi với các đơn vị từ vựng, ví dụ như các từ) thì tạo nên một cấu trúc đặc thù đi với chuỗi được tạo bởi văn phạm.

Nhiều kiểu văn phạm khác như văn phạm cấu trúc ngữ đoạn tổng quát (*GPSG - Generalized Phrase Structure Grammar*), văn phạm cấu trúc ngữ đoạn theo từ chủ (*HPSG - Head-Driven Phrase Structure Grammar*), Văn phạm Chức năng từ vựng (*LFG - Lexical Functional Grammar*) thực chất đều là kiểu văn phạm văn phạm phi ngữ cảnh dựa trên sự thống nhất. Các loại văn phạm này; nếu không có ràng buộc, nó có thể tương đương với máy Turing. Nhìn từ góc độ ngôn ngữ học, những kiểu văn phạm này cần được giới hạn để chức năng miêu tả của chúng chỉ đơn giản là cần và đủ chứ không hơn; còn nhìn từ góc độ tính toán, chúng cần được giới hạn để mang lại những thuật toán phân tích cú pháp có hiệu quả. Cả hai cách nhìn này là cơ sở cho những nghiên cứu tiếp theo trong lĩnh vực này.

I.2. ÁP DỤNG VĂN PHẠM VÀ NHỮNG TRỞ NGẠI

Những nghiên cứu về mô hình văn phạm Chomsky và ứng dụng nó trong xử lý ngôn ngữ tự nhiên về sau cho thấy có nhiều tình huống ngôn ngữ hoặc không thể diễn đạt được bằng mô hình Chomsky hoặc chỉ có thể diễn đạt theo cách không tự nhiên, rất khó hiểu đối với tư duy của con người [1]. Trong phần này ta sẽ xem xét những tình huống ngôn ngữ thực, trong đó mô

hình văn phạm Chomsky (cụ thể là các văn phạm phi ngữ cảnh và cảm ngữ cảnh) tỏ ra có những hạn chế nhất định.

I.2.1. SỰ PHÂN CẤP KHÁI NIỆM.

Trong văn phạm sinh, mỗi biến trung gian (*nonterminal*) đều là một ký hiệu riêng, không có sự liên hệ nào giữa chúng với nhau. Vì vậy, khi gán một tính chất nhất định cho một biến trung gian, ta không thể phân phối tính chất này cho các tên biến khác. Chẳng hạn „*Thêm đuôi _s để hình thành số nhiều của danh từ*“ là một ***quy tắc từ vựng*** chung cho lớp danh từ. Giả sử, trong lớp các danh từ, ta muốn phân loại thành các lớp con: danh từ khối, danh từ đếm, danh từ chỉ người, động vật, vật dụng, hiện tượng, khái niệm, ... tùy theo nhu cầu của ứng dụng và, giả sử, ta định đặt tên tương ứng khác nhau cho mỗi lớp con thông qua những biến trung gian khác nhau trong một hệ văn phạm sinh. Trong trường hợp này, ta sẽ không thể ngầm định tính chất về số nhiều cho tất cả các lớp con của danh từ. Khi đó, nếu ta muốn ***bộ phân tích từ vựng*** có thể tạo ra dạng số nhiều của loại danh từ thì ta phải quy định ra những quy tắc giống nhau cho tất cả các loại danh từ. Nếu trong mỗi lớp danh từ ta lại tiếp tục muốn chia ra thành những lớp con thì ta lại buộc phải tạo ra những quy tắc riêng cho những loại từ mới này nữa.

Văn phạm phi ngữ cảnh Chomsky không phân biệt hai loại quy tắc:

- $A \rightarrow \omega$ với ω có độ dài lớn hơn 1, và
- $A \rightarrow X$ với X là biến hoặc từ cuối

Quy tắc thứ nhất là một loại quy tắc gộp (khái niệm A được định nghĩa thông qua sự kết hợp của những khái niệm khác như những *thành phần* của nó), ta tạm gọi chúng là ***quy tắc sinh thực sự***. Trong khi đó quy tắc loại 2 là sự trừu xuất khái niệm (A là X). Như vậy có thể coi loại quy tắc này không phải là một quy tắc sinh, chúng có thể được sử dụng để xây dựng hệ phân cấp các khái niệm dưới dạng một giàn đại số. Khi đó, bộ quy tắc chỉ chứa những quy tắc thực sự, và một sự áp dụng quy tắc sẽ luôn luôn thay đổi độ dài của dạng câu.[1]

I.2.2. MỐI LIÊN HỆ GIỮA CÁC BỘ PHẬN TRONG CÂU.

Trong các tài liệu dạy hay khi truyền đạt kiến thức ngoại ngữ ta thường gặp những câu chỉ dẫn về ngữ pháp như:

- *Khi trong một cấu trúc Z có mặt X thì có nghĩa là ...*

[1] Như vậy có thể coi ***dạng chuẩn Chomsky*** là khởi đầu cho việc tách hai loại quy tắc.

mà không nói rõ X nằm trong ngữ cảnh cụ thể nào (nghĩa là không quan tâm đến việc bên cạnh X có những từ ngữ gì) mà chỉ có chỉ dẫn về việc X nằm trong ngữ đoạn Z (chẳng hạn, nếu X nằm trong một danh ngữ, một trạng ngữ, hay một động ngữ, ...). Ngữ cảnh cụ thể bên cạnh X tỏ ra không có ý nghĩa quan trọng (*hoặc thậm chí không thể liệt kê hết ra được*). Loại chỉ dẫn như thế này thường có mục đích để giải quyết nhập nhằng: Giả sử X có các ngữ nghĩa $\aleph_{X1}, \aleph_{X2,...}\aleph_{Xn}$. Khi X nằm trong ngữ đoạn Z thì ngữ nghĩa của nó sẽ nhận giá trị cụ thể $\aleph_{Xz,}$. Kiểu ràng buộc này có *tính cảm ngữ cảnh khái quát* (*generic context-sensitivity*). Áp dụng văn phạm theo mô hình phân cấp của Chomsky, ta sẽ buộc phải tạo ra một tập (vô hạn tiềm năng) các quy tắc cảm ngữ cảnh để mô tả tình huống văn phạm như ở trên.

Đối với những mối liên hệ ngữ nghĩa loại này, ta phải cần có một sự mở rộng nhất định về dạng của quy tắc sinh để mô tả chúng. Trong quy tắc sinh ngoài hai vế $A \rightarrow \omega$ còn cần có thêm một biến B như một *ngữ đoạn (phrase)*, hay *phạm vi* (*scope*) để chỉ rõ điều kiện mà quy tắc $A \rightarrow \omega$ được áp dụng.

I.2.3. MỐI LIÊN HỆ GIỮA CÁC TẦNG CẤU TRÚC TRONG CÂU.

Các *ngữ đoạn (phrase)* trong câu thường bao gồm nhiều thành phần, chẳng hạn, đối với *Danh ngữ*, bên cạnh danh từ chính, có thể còn có các danh từ, tính từ, định ngữ, v.v.. bổ nghĩa cho nó. Các mô hình phân tích dựa trên văn phạm Chomsky thường đặt các phần tử phụ nghĩa này theo một thứ tự phân cấp chặt chẽ tuân thủ nghiêm ngặt hệ các quy tắc sinh cho danh ngữ đó. Trong khi đó, chẳng hạn, để nhận thức một cụm danh ngữ, người ta phân tích sự liên hệ giữa danh từ chính với mỗi phần tử phụ nghĩa cho nó, *không phụ thuộc vào vị trí tương đối* của chúng so với vị trí của danh từ chính trong cụm từ. Đó là hạn chế do hình dạng của quy tắc sinh: vế phải của quy tắc phải có một độ dài nhất định. Chẳng hạn quy tắc

Noun $\rightarrow$ Noun Noun (1)

(tổ hợp hai danh từ đứng cạnh nhau trong tiếng Anh hình thành một danh từ) không chỉ rõ danh từ nào là chính, còn danh từ nào là phụ, bổ nghĩa cho danh từ kia.

Trong tiếng Việt, cụm danh từ (với hai danh từ đứng cạnh nhau) được biểu diễn dưới dạng:

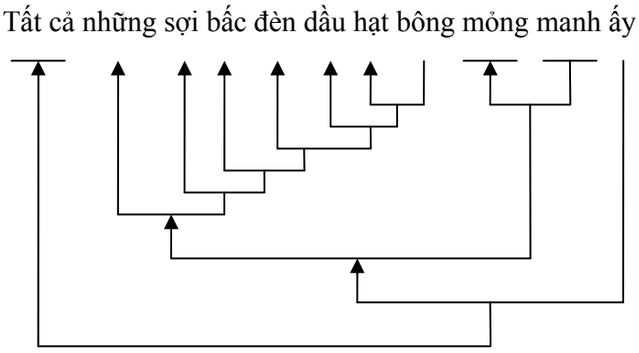
Danh_từ $\rightarrow$ Danh_từ Danh_từ (2)

Về mặt hình thức, hai quy tắc (1) và (2) trên đây có dạng thức hoàn toàn giống nhau. Quy tắc sinh không cho ta thấy trật tự khác nhau giữa tiếng Việt và tiếng Anh trong việc hình thành cụm danh từ : trong tiếng Anh danh

từ chính thường đứng sau danh từ bổ nghĩa cho nó còn trong tiếng Việt, danh từ chính lại đứng trước.

Sự không nhất quán giữa cấu trúc ngữ đoạn và biểu diễn hình thức (qua cây cú pháp) còn thể hiện ở một khía cạnh khác. Trong một tài liệu về tiếng Việt [40] dẫn ra một ví dụ phân tích cụm từ : „*Tất cả những sợi bấc đèn dầu hạt bông mỏng manh ấy*“ (Hình 1).

Ở đây ta thật khó hình dung ra mối liên hệ giữa đâu là danh từ chính. đâu là những phần tử phụ nghĩa cho nó, cây cú pháp như thế này không phản ánh sự phụ thuộc về ngữ nghĩa sẽ rất khó khăn[1] để nhận thức và vì vậy, vô dụng, mà nguyên nhân lại nằm ở chỗ sử dụng một cách máy móc mô hình văn phạm sinh Chomsky để dựng ra cây cú pháp.



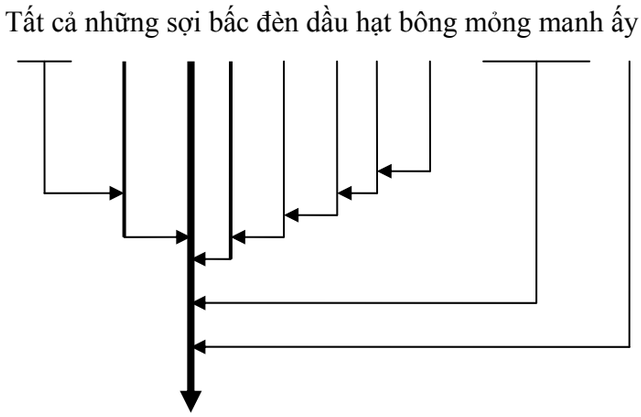
Hình 1. Cây cú pháp của danh ngữ theo [2]

Dễ thấy rằng những từ „*tất cả*“, „*những*“, „*bấc đèn dầu hạt bông* “, „*mỏng manh*“ ,“*ấy*“ đều bổ nghĩa cho „*sợi* “, và xét về mặt ngữ nghĩa, chúng đều bình đẳng với nhau, và thứ tự của chúng trong câu về thực chất là không quan trọng đối với nhận thức của chúng ta, mặc dù trật tự này là bắt buộc đối với *hành văn* tiếng Việt. Một cách phân tích hợp lý và dễ hiểu cho phát biểu này có thể thấy trong Hình 2. Số lượng mũi tên trỏ trực tiếp đến danh từ “*sợi*” là một *đại lượng biến thiên* tùy theo độ phức tạp của danh ngữ. Nghĩa là khi thêm các phần tử bổ nghĩa cho danh từ chính thì có thêm

1 Trong cụm từ đã nêu có một sự phụ thuộc hàm mà chúng ta đều cảm nhận một cách rõ ràng. Ở đây, những từ „tất cả“, „những“, „bấc“, „mong manh“, „ấy“ đều là phụ nghĩa cho danh từ chính „sợi“; từ „đèn“ phụ nghĩa cho „bấc“, từ „dầu“ phụ nghĩa cho „đèn“, từ „hạt bông“ phụ nghĩa cho „dầu“. Trong cấu trúc phụ thuộc hàm, số lượng các từ phụ nghĩa không cố định, đồng thời trật tự của chúng cũng không quan trọng. Thực tế là trật tự các từ chỉ bị chi phối bởi nhu cầu diễn đạt trên một ngôn ngữ cụ thể do tính tuyến tính bắt buộc của mọi ngôn ngữ tự nhiên, và vì vậy, trật tự này chỉ đúng cho từng ngôn ngữ cụ thể với những quy ước riêng của cộng đồng những người sử dụng ngôn ngữ đó.

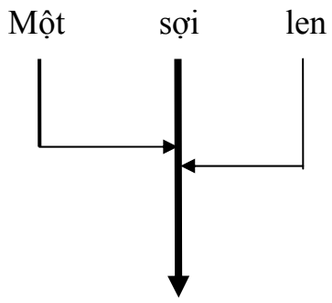
một mũi tên trỏ đến nó. Để diễn đạt tình huống này không thể sử dụng các quy tắc văn phạm thông thường như định nghĩa của Chomsky được[1].

Trong Hình 2, ta thấy danh từ chủ đạo được đánh dấu riêng (tô đậm - danh từ *sợi*). Danh ngữ, như một cụm từ, mang trong mình mọi thuộc tính của danh từ chính *(từ chủ)* của nó.



Hình 2. Dạng cây cú pháp của danh ngữ theo trực cảm
(Cây phụ thuộc ngữ nghĩa)

Bằng cách đó, ràng buộc ngữ cảnh giữa một thành phần nào đó (chẳng hạn, động từ) với một ngữ đoạn (chẳng hạn, danh ngữ) có thể đưa về sự ràng buộc ngữ cảnh giữa thành phần đó với *từ chủ* của ngữ đoạn. Đây cũng chính là cách thức mà con người liên tưởng khi đọc hiểu hay đặt câu.



Hình 3. Dạng cây cú pháp của danh ngữ

[1] Các giải thuật phân tích đều xây dựng một tổ chức bên trong (cây cú pháp) tương ứng với các quy tắc sinh và với lịch sử áp dụng chúng, vì vậy cây cú pháp luôn luôn bị gắn chặt với cách thức biểu diễn các quy tắc sinh của văn phạm được áp dụng.

Như vậy, sự phụ thuộc nghĩa theo trực cảm có một tính chất mà mô hình Chomsky không mô tả được, đó là các nút với số nhánh ***biến thiên*** (như trong ví dụ nêu trên, số lượng các mũi tên trỏ tới danh từ *sợi* có thể thay đổi, xem Hình 3).

I.2.4. RÀNG BUỘC VĂN PHẠM VÀ THÔNG TIN DẪN XUẤT

Xét ví dụ về dạng câu hỏi ***Tag-question*** trong tiếng Anh:

„*Your old* ***friend*** *from south Đaklak* ***likes*** *coffee,* ***doesn't he***?“

Phần ***Tag*** (như *isn't it?, won't you?, aren't they?*) được xác định tùy thuộc vào từ chính trong chủ ngữ (ở đây: ***friend***) và tùy thuộc vào *thì* và loại của động từ (ở đây: ***likes***) của câu hỏi. Khi phân tích câu, bộ phân tích cần phải kiểm tra tính tương thích của phần *Tag* (ở đây là "***doesn't he"***) ứng với mệnh đề chính của câu hỏi. Khi tổng hợp câu, phần *Tag* phải được tạo ra từ nội dung của phần mệnh đề chính của câu hỏi để có thể xây dựng được kiểu câu hỏi đúng văn phạm. Sử dụng mô hình văn phạm sinh Chomsky, ta sẽ phải tạo ra một họ các quy tắc tương tự nhau ứng với một loạt các tình huống khác nhau của phần *Tag*. Thêm vào đó, những quy tắc này phải được dẫn xuất trong hệ phân cấp để đi đến các thành phần sâu hơn trong cây cú pháp (từ chính của chủ ngữ, trợ động từ hoặc động từ chính,...). Phần *Tag* trong câu hỏi tiếng Anh là bộ phận dư thừa, nhưng luật hành văn đòi hỏi cần phải được tổng hợp đúng văn phạm, mặc dù nó không mang thông tin nội dung nào (*ngoài ý nghĩa giúp nhân mạnh và khẳng định rằng đây là một câu hỏi chứ không phải là một thông báo*).

Để xử lý tình huống này, trong mô hình hình thức cần có công cụ để mô tả sự tương quan giữa các thành phần của một quy tắc sinh thông qua các *thỏa thuận*.

I.2.5. VĂN PHẠM CẢM NGỮ CẢNH YẾU

Trong bất cứ kiểu văn phạm mang tính tính toán chính xác nào, người ta đều phải mô tả mối liên hệ ràng buộc giữa các thành tố văn phạm khác nhau. Sau đây là một vài ví dụ:

- Sự phù hợp về ngôi, số, giống. Chẳng hạn, trong tiếng Anh, động từ phải phù hợp với chủ ngữ về ngôi và số.
- Sự phân loại nhỏ các động từ trong đó mỗi động từ định rõ một hay nhiều khung phân loại nhỏ cho các bổ ngữ của mình. Chẳng hạn, động từ ***ngủ*** không cần có bổ ngữ (***Việt Dũng ngủ),*** động từ ***thích*** cần có một bổ ngữ (***Bích Thủy thích ô mai***), động từ ***đưa*** cần có hai bổ ngữ (như ***Việt Dũng đưa Bích Thủy gói ô mai)*** vv...

- Đôi khi mối liên hệ giữa các tham tố không hiện ra ở các vị trí thường thấy. Trong câu: ***Who1 did John invite*** e1.[1] ở đây, e1 thay thế cho who1, who1 là từ điền vào chỗ trống. Từ được điền và chỗ trống không càn thiết phải ở một vị trí cố định. Vì vậy trong câu: ***Who***i ***did Bill ask John to invite e***i. Từ điền vào chỗ trống và chỗ trống ở khoảng cách xa nhau hơn so với câu trên.
- Đôi khi mối liên hệ này bị ẩn đi. Ví dụ trong tiếng Đức, người ta có thể nói: ***Hans***i ***Peter***j ***Marie* schwimmen*lassen***j ***sah***i , ***(Hans saw Peter make Marie swim)*** ở đây, danh từ và động từ ở thứ tự bị ẩn đi, như ký hiệu dưới các từ đã chỉ ra.
- Tuy nhiên, trong tiếng Đức, những mối liên hệ này được xen kẽ móc nối vào nhau, như trong ví dụ:

Jani ***Piet***j ***Marie***k ***zag***i ***laten***j ***zwemmen***k***.***

Tất nhiên, có những tình huống mà mối liên hệ này ở dạng phức tạp hơn. Mô hình toán học của những mối liên hệ này là một trong những vấn đề cơ bản của xử lý ngôn ngữ tự nhiên. Nhiều mối liên hệ (chẳng hạn như mối liên hệ chéo như đã đề cập ở trên) không thể trình bày bằng kiểu văn phạm phi ngữ cảnh. Có thể dễ dàng nhận ra điều này từ một thực tế được công nhận rộng rãi là văn phạm phi ngữ cảnh thì tương đương với ô tô mát đẩy xuống. Vì vậy ô tô mát đẩy xuống có thể phân tích được các mối liên hệ ẩn này.

Trong kiểu văn phạm văn phạm phi ngữ cảnh như trên hình 1, mối liên hệ giữa động từ ***(thích***) và hai tham tố (chủ ngữ (CN) và tân ngữ (TN) được định bởi hai quy tắc văn phạm. Không thể làm rõ mối liên hệ này với một quy tắc duy nhất mà không bỏ động ngữ trên sơ đồ. Nghĩa là, nếu chúng ta đưa ra quy tắc S → DN ĐT DN, chúng ta có thể biểu diễn mối liên hệ chỉ bằng một quy tắc, nhưng nếu vậy chúng ta không thể có động ngữ trong văn phạm. Vì thế nếu chúng ta coi mỗi một quy tắc của văn phạm phi ngữ cảnh là định rõ một lĩnh vực khu biệt, thì một phạm vi khu biệt của văn phạm phi ngữ cảnh lại không thể khu biệt mã hoá mối liên hệ giữa động từ và các tham tố của nó, và vẫn xuất hiện động ngữ trên nút của sơ đồ (mô hình ***văn phạm cảm ngữ đoạn*** có thể giải quyết được các tình huống ngôn ngữ này).

Còn trong kiểu văn phạm kết nối cây (*Tree-Adjoining Grammar*), mỗi từ (từ đóng vai trò như là điểm tựa cho sơ đồ) đi với một cấu trúc (sơ đồ) mã hoá mối liên hệ giữa từ và tham tố của nó (và vì thế sự phụ thuộc không trực tiếp vào các từ khác là điểm tựa cho cấu trúc sẽ lấp đầy các vị trí của các tham tố). Vì vậy, với ***thích,*** sơ đồ tương ứng của nó mã hoá các tham tố (là 2 nút danh ngữ trên sơ đồ của ***thích)*** đồng thời cũng tạo ra các khoảng trống

[1] Các ví dụ lấy từ [2].

thích hợp trong cấu trúc. Sơ đồ của ***Bích Thủy*** và ***ô mai*** có thể lần lượt thay thế cho chủ ngữ và tân ngữ trong sơ đồ cho ***thích***. Sơ đồ cho ***cực kỳ*** có thể điền vào vị trí động ngữ trên sơ đồ của ***thích***. Xuất phát điểm của kiểu văn phạm kết nối cây hơi khác so với kiểu văn phạm phi ngữ cảnh. Trong kiểu văn phạm kết nối cây, toàn bộ văn phạm bao gồm các thành tố từ và các cấu trúc đi kèm với nó. Có những sự thay thế, tiếp nối và vận hành phổ biến miêu tả cách các cấu trúc có thể kết hợp với nhau bằng cách nào.

Trong kiểu văn phạm kết hợp vô điều kiện, mỗi từ được quy là một loại, đơn hoặc kép. Trong kiểu văn phạm ràng buộc ngữ cảnh (*Context Constrained Grammar*), toàn bộ hệ thống văn phạm bao gồm các thành tố từ và các loại từ được quy định cho chúng. Có 2 chức năng phổ biến mô tả sự kết hợp của các mục từ loại, chức năng ghép và chức năng kết hợp. văn phạm ràng buộc ngữ cảnh cũng cho phép sự tăng loại. Nguồn gốc nghiên cứu của văn phạm ràng buộc ngữ cảnh là lịch sử tạo thành chuỗi bằng việc vận dụng thành công chức năng ghép và kết hợp. Một văn phạm ràng buộc ngữ cảnh thì không nhất thiết phải định ra một cấu trúc từ ngữ duy nhất. Cấu trúc này phụ thuộc vào cách thức và thứ tự bản thân nó được sử dụng. Cách thức và thứ tự sử dụng khác nhau sẽ cho ta những kết quả mô tả cấu trúc từ ngữ khác nhau, thậm chí cho cả những câu mang nghĩa rõ ràng.

Cả văn phạm ràng buộc ngữ cảnh và văn phạm kết nối cây đều có khu vực khu biệt rộng hơn văn phạm phi ngữ cảnh, bởi vì trong mọi trường hợp, tham tố của động từ ***thích*** được mã hoá trong cấu trúc đi với động từ, và vì thế mà có động ngữ. Khu vực khu biệt rộng hơn cho phép văn phạm kết nối cây hoàn toàn loại bỏ sự quay lại của các khu vực liên hệ, vì thế khu biệt hoá các mối liên hệ trên sơ đồ chính.

văn phạm kết nối cây và văn phạm ràng buộc ngữ cảnh có nhiều điểm giống nhau. Trên thực tế, chúng tỏ ra tương đương nhau (chú ý khả năng sinh sản hạn chế của chúng, nghĩa là các cặp câu mà chúng tạo ra). Chúng mạnh hơn văn phạm phi ngữ cảnh và nằm trong hệ thống văn phạm cảm ngữ cảnh yếu. Hệ thống này mang nhiều đặc điểm cơ bản của văn phạm phi ngữ cảnh và vì thế có thể đủ mạnh để phát hiện những mối liên hệ trong cấu trúc ngôn ngữ, chẳng hạn như mối quan hệ chéo như chúng ta đã nói ở trên. Một vài cách tiếp cận hình thức trong thời gian gần đây như Văn phạm Chỉ mục tuyến tính (*Linear Indexed Grammar*) và Văn phạm từ chủ (*Head Grammar*) cũng tỏ ra giống với văn phạm kết nối cây. Sự tương đồng giữa một số kiểu văn phạm thuần tuý ngôn ngữ dựa trên sự khác biệt về bản chất trong cấu trúc ngôn ngữ đã dẫn đến sự tìm kiếm sự bất biến trong các kiểu văn phạm thuộc loại này, mà xét về một khía cạnh nào đó, những sự bất biến này còn quan trọng hơn bản thân từng kiểu văn phạm. Văn phạm học về văn phạm cảm ngữ cảnh yếu (*Mildly Context-sensitive*) và những nghiên cứu các

tương đồng với nó là một trong những lĩnh vực nghiên cứu năng động nhất trong ngôn ngữ học chính xác trong thập niên 80.

Chúng ta đã kết luận rằng văn phạm đưa ra một kiểu cấu trúc duy nhất cho một câu (giả sử câu đó mang nghĩa rõ ràng). Vì thế, ví dụ: ***Bích Thủy thích ô mai*** sẽ được đưa vào trong ngoặc như sau (bỏ qua tên các cụm từ và một số ngoặc đơn không cần thiết cho mục đích nghiên cứu trong tình huống này của chúng ta)

(a) (Bích Thủy (thích ô mai))

Trong kiểu văn phạm ràng buộc ngữ cảnh, như đã nói ở trên, chúng ta có thể đưa ra nhiều cấu trúc cho các câu mang nghĩa rõ ràng. Vì vậy văn phạm ràng buộc ngữ cảnh đưa ra nhóm câu sau cho câu ***Bích Thủy thích ô mai.***

(b) (Bích Thủy (thích ô mai))

(c) ((Bích Thủy thích) ô mai)

Chứng minh cho những cấu trúc như vậy là cách sử dụng chúng trong câu ghép (chẳng hạn với ***và, nhưng, còn...***) và trong cụm từ có ngữ điệu rõ ràng. Vì thế, cách ghép ngoặc trong câu (b) là cần thiết cho câu (d), (c) cho (e).

(d) (Bích Thủy ((thích ô mai) nhưng lại (ghét mận)))

(e) (((Bích Thủy thì thích) còn (Việt Dũng thì ghét)) ô mai)

Cũng như vậy (b) tương đương với cụm từ mang ngữ điệu nếu ngữ cảnh trên là (f) và tương đương với (c) nếu ngữ cảnh là (g).

(f) Ai thích ô mai? (Bích Thủy (thích ô mai))

(g) Bích Thủy thích gì? ((Bích Thủy thích) ô mai)

Sự linh hoạt trong sự phân định một cấu trúc có được nhờ bỏ đi khái niệm về một cấu trúc chuẩn. Tuy nhiên không cần phải bỏ đi khái niệm về một cấu trúc chuẩn. Ta có thể vẫn duy trì một cấu trúc cố định ở một cấp độ nhất định (chẳng hạn như trong sơ đồ cơ bản ở văn phạm kết nối cây) và vẫn có thể có được sự linh hoạt cần thiết như trong các ví dụ ở phần trên. Trong nghiên cứu HPSG ta cũng có thể thu được các kết quả tương tự.

Trên thực tế, những nghiên cứu về văn phạm cảm ngữ cảnh yếu vẫn chưa kết tinh thành các ứng dụng xử lý ngôn ngữ tự nhiên thực tiễn thuyết phục.

Trong chương sau, ta sẽ đề cập cách giải quyết những vấn đề này theo một hướng khác – bằng một công cụ được gọi là *văn phạm cảm ngữ đoạn*.

I.3. CÁC KHUYNH HƯỚNG TRONG DỊCH MÁY

Các phương pháp được sử dụng trong nghiên cứu về dịch tự động đã trải qua nhiều sự thay đổi. Phần này giới thiệu tổng quan hai cách tiếp cận dịch máy chính hiện nay là dựa theo luật và dựa trên kho ngữ liệu.

Tùy thuộc vào việc kiểu kiến thức bổ sung tích hợp trong dịch máy, người ta phân biệt ba kiểu hệ thống [12]:

1. Những hệ thống sử dụng thuật ngữ được tổ chức theo mô hình chuyên ngành kỹ thuật. Những hệ thống này không chứa đựng cơ sở tri thức theo lĩnh vực.

2. Những hệ thống sử dụng những kiến thức về khái niệm hoặc những sự kiện cho những nhiệm vụ đặc biệt như giải quyết nhập nhằng cú pháp, ngữ nghĩa.

3. Những hệ thống có sự biểu diễn ngữ nghĩa sâu (thường là các hệ thống liên ngữ) bằng việc sử dụng kiến thức bổ sung của một thể loại nào đó.

I.3.1. CÁCH TIẾP CẬN DỰA THEO LUẬT

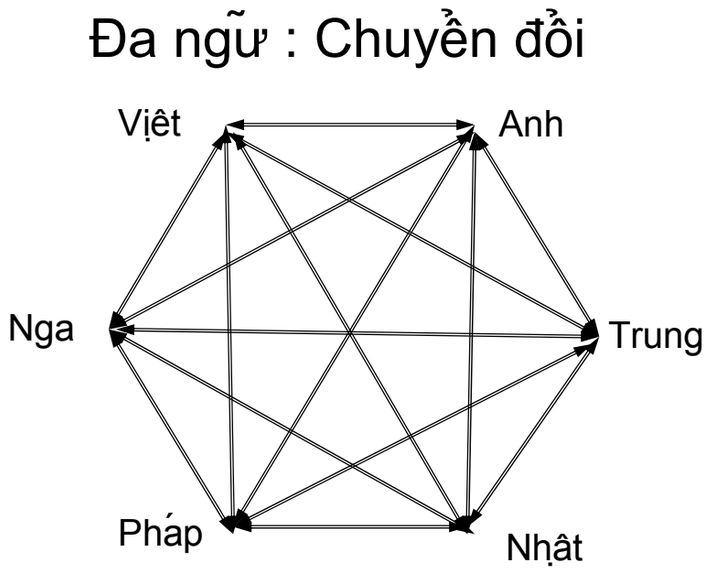
Trong những năm 1980, phương hướng chủ đạo trong nghiên cứu dịch máy thực chất là cách tiếp cận dựa trên quy tắc ngôn ngữ theo nhiều kiểu: quy tắc phân tích cú pháp, quy tắc từ vựng, quy tắc chuyển đổi từ vựng, hình thái học, quy tắc tổng hợp cú pháp, v.v... Những hệ thống chuyển đổi chiếm đa số [4, 7, 9, 11, 12, 13, 14, 22] (chẳng hạn Ariane, Metal, SUSY, Eurotra, SITE, LMT,...), có một số hệ thống liên ngữ (DLT và Rosetta), một vài hệ có cách tiếp cận trên nền kiến thức, sử dụng thông tin phi ngôn ngữ liên quan đến các lĩnh vực của văn bản cần phải dịch [9].

Phương pháp liên ngữ [3, 6, 8, 9, 11, 13, 15, 17, 19, 21, 23] được đánh giá là cách tiếp cận tiên tiến hơn do hứa hẹn bản dịch có chất lượng cao hơn cũng như giảm chi phí khi xây dựng hệ dịch máy đa ngữ so với phương pháp chuyển đổi.

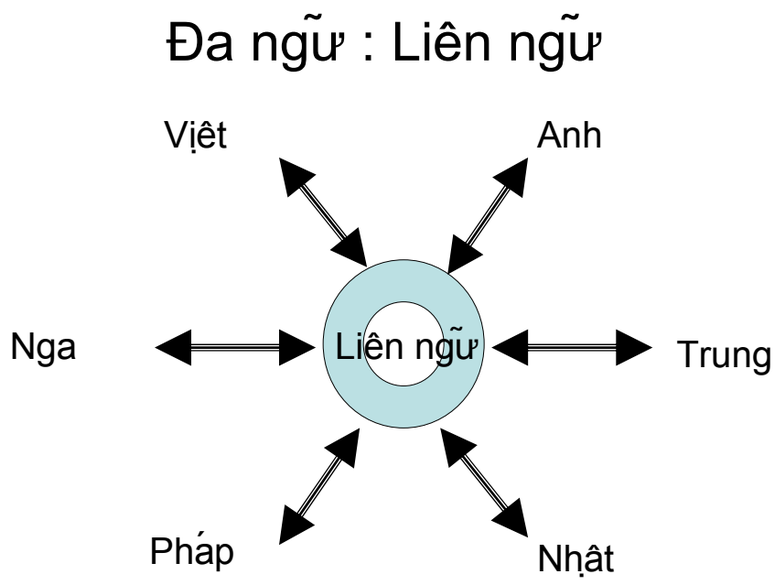
Một đặc tính điển hình của những hệ thống dựa trên quy tắc là sự biến đổi hoặc ánh xạ của biểu diễn cây được gắn nhãn [13], từ một cây hình thái học vào một cây cú pháp, từ một cây cú pháp vào một cây ngữ nghĩa, từ một cây giao diện của ngôn ngữ nguồn sang cây tương đương của ngôn ngữ đích, v.v...

Sự chuyển đổi quy tắc yêu cầu thỏa mãn những điều kiện chặt chẽ: cây phải có cấu trúc đặc biệt và chứa đựng những tiết mục từ vựng hoặc đặc tính cú pháp hay ngữ nghĩa đặc biệt. Ngoài ra, mỗi cây được kiểm tra bởi

những quy tắc hình thành; chẳng hạn một *văn phạm* xác nhận tính chấp nhận được của cấu trúc của nó và những mối quan hệ mà nó đại diện. Quy tắc văn phạm và chuyển đổi chỉ rõ những sự ràng buộc xác định khả năng chuyển đổi từ mức này sang mức khác và cuối cùng - chuyển đổi văn bản ngôn ngữ nguồn tới văn bản ngôn ngữ đích [11, 13, 14].

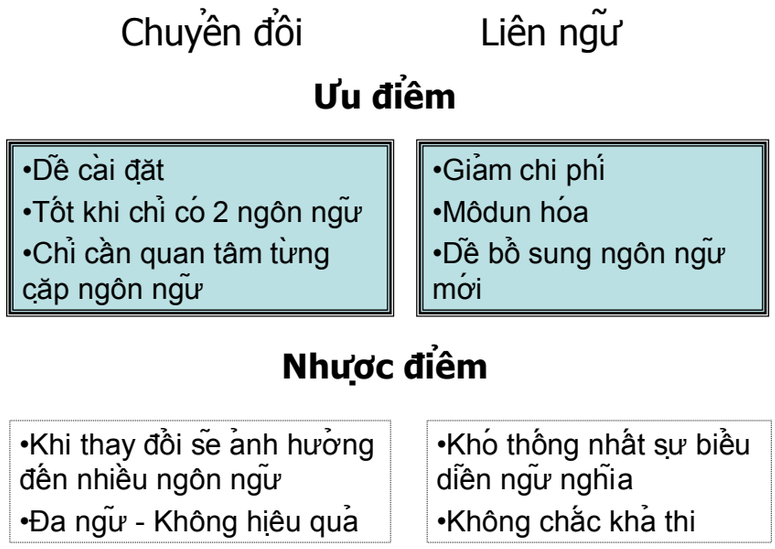


Hình 4 : Dịch máy Chuyển đổi



Hình 5 : Dịch máy Liên ngữ

Từ giữa những năm 1980 có một xu hướng chung sử dụng mô hình hình thức "dựa trên sự hợp nhất" (*unification-based*) và "dựa trên ràng buộc" "*constraint-based*" [9, 21]. Ưu điểm chính của cách tiếp cận này là sự đơn giản hóa các quy tắc (và dẫn đến sự đơn giản hóa quá trình tính toán) để phân tích, biến đổi và tổng hợp. Thay vì một dãy biểu diễn nhiều mức phức tạp và tập hợp lớn các quy tắc chuyên biệt (mà chỉ được áp dụng trong hoàn cảnh và cấu trúc riêng), tồn tại cách biểu diễn một lớp và với một tập hợp hạn chế các quy tắc trừu tượng, cùng với những điều kiện và ràng buộc gắn kết với mục từ vựng đặc biệt. Đồng thời, những thành phần của các văn phạm này, về nguyên tắc, đều có thể đảo ngược, sao cho không còn cần thiết phải xây dựng các văn phạm khác nhau để phân tích và tổng hợp cho cùng một ngôn ngữ.



Cách tiếp cận từ vựng dần dần thay thế cho sự định hướng cú pháp mô tả đặc điểm giao tiếp của những hệ thống trước đây, với một sự gia tăng thông tin gán cho những đơn vị từ vựng từ điển: tương đương dữ liệu và phiên dịch hình thái học, kèm theo thông tin về hầu hết những ràng buộc và thông tin phi ngôn ngữ cũng như nhận thức cú pháp và ngữ nghĩa học. Sự mở rộng dữ liệu từ vựng được thể hiện rõ ràng nhất trên hệ thống liên ngữ, bao gồm một số lượng lớn thông tin phi ngôn ngữ.

I.3.2. PHƯƠNG PHÁP DỰA VÀO KHO NGỮ LIỆU

Trong những năm gần đây, người ta đã đưa ra những cách tiếp cận mới: dịch theo thống kê (*Statistical-based Translation*), dịch theo ví dụ (*Example-based Translation*) [1], dịch nhớ (*Translation Memory*). Những công cụ này có tác dụng hỗ trợ việc tự động hóa khâu thu thập tri thức ngôn ngữ trên cơ sở duyệt một khối lượng lớn văn bản (*đơn ngữ, song ngữ,...*) cũng như xử lý thành ngữ, những cụm từ ổn định thường gặp,... Những cố gắng này giúp giảm bớt chi phí thu thập, xử lý cơ sở tri thức ngôn ngữ trong các hệ dịch máy.

Từ 1989 bắt đầu hình thành những phương pháp dựa vào kho ngữ liệu, sau khi một nhóm nghiên cứu của IBM công bố kết quả thí nghiệm trên hệ thống Candide với một cách tiếp cận thuần túy thống kê [24]. Trong hệ thống này phương tiện duy nhất để phân tích và tổng hợp là thống kê (không sử dụng bất kỳ một *quy tắc ngôn ngữ* nào). Kho ngữ liệu là biên bản chính thức về các cuộc họp của nghị viện Ca-na-đa. Phương pháp của IBM có thể mô tả vắn tắt như sau:

- Dóng hàng câu, nhóm từ và từ đơn lẻ của văn bản song ngữ,
- Tính toán xác suất mà bất kỳ từ nào trong ngôn ngữ này có quan hệ với một từ hoặc một cụm từ trong câu dịch tương ứng với nó ở ngôn ngữ kia.

Kết quả thử nghiệm rất hứa hẹn: non nửa số câu được dịch chính xác hoàn toàn với bản dịch trong kho ngữ liệu, hoặc thể hiện cùng một nội dung với từ ngữ hơi khác, hoặc đưa ra bản dịch gần như tương đương.

Phương pháp *kho ngữ liệu* [2, 3, 5, 10, 11, 16, 18, 20, 23, 24] với việc tham khảo nhanh chóng một khối lượng dữ liệu văn bản lớn mang bản chất của cách tiếp cận *trên nền ví dụ,* hay *trên nền kí ức:* việc dịch thường là kết quả tìm kiếm hoặc nhớ lại những ví dụ tương tự, tìm hiểu hoặc suy diễn xem có cách diễn đạt đặc biệt hoặc có mệnh đề tương tự nào đó đã được dịch từ trước hay không.

Cách tiếp cận dựa vào ví dụ (là phương hướng đang được Microsoft Research (*Công ty Microsoft*) thực hiện) [2, 10, 20] cũng được thể hiện qua quá trình tích lũy và lựa chọn mệnh đề hoặc nhóm từ tương đương trong kho văn bản song ngữ, được sắp xếp bằng phương pháp thống kê (tương tự cách thức của nhóm IBM) hoặc bằng nhiều phương pháp phân tích trên nền quy tắc truyền thống. Chất lượng dịch thuật [10] được các tác giả đánh giá là so sánh được với các hệ dịch máy dựa theo luật hiện có (SYSTRAN, BABELFISH, http://world.altavista.com/ , và L&H, http://officeupdate .lhsl.com/), nghĩa là chưa có những tiến triển rõ rệt.

Để tính toán sự tương đồng, một số nhóm sử dụng phương pháp ngữ nghĩa, như mạng ngữ nghĩa hoặc sự *phân cấp thuật ngữ chuyên ngành*. Một số nhóm khác sử dụng thông tin thống kê về những tần số từ vựng trong ngôn ngữ đích [9]. Lợi thế chính của cách tiếp cận là một khi văn bản đã được rút ra từ ngân hàng dữ liệu của những bản dịch thực tế trước đó do những người dịch chuyên nghiệp thực hiện thì sẽ cho kết quả chính xác và trơn tru.

Tuy nhiên, những kết quả ứng dụng thực tiễn cho thấy chất lượng của các hệ thống dịch máy (dù là dựa theo luật hay thống kê) chưa đáp ứng được những kỳ vọng của xã hội.

I.3.3. MỘT SỐ HỆ DỊCH MÁY LIÊN NGỮ

Phương pháp liên ngữ giả thiết rằng tồn tại một dạng biểu diễn trung gian độc lập ngôn ngữ. Văn phạm phân tích của ngôn ngữ nguồn được sử dụng để đưa câu văn về dạng biểu diễn tri thức ngôn ngữ (chung cho mọi ngôn ngữ tự nhiên). Sau đó sử dụng văn phạm tổng hợp của ngôn ngữ đích để dịch từ liên ngữ sang ngôn ngữ đích.

Mô hình dịch máy liên ngữ có những ưu điểm sau:

- Độc lập ngôn ngữ: trong khi phân tích ta chỉ cần quan tâm đến ngôn ngữ nguồn, khi tổng hợp – ngôn ngữ đích.
- Dễ dàng bổ sung ngôn ngữ mới vào hệ dịch máy. Để thêm một ngôn ngữ vào hệ thống, ta chỉ cần xây dựng các bộ văn phạm phân tích và tổng hợp cho ngôn ngữ mới. Trong khi đó, với mô hình chuyển đổi, ta phải xây dựng các hệ văn phạm chuyển đổi từ ngôn ngữ mới sang tất cả các ngôn ngữ đã có và ngược lại.

Tuy nhiên, cho đến nay, những hệ dịch máy phổ biến hiện có trên thị trường đều được xây dựng theo phương pháp chuyển đổi, chưa có hệ dịch máy liên ngữ thương phẩm nào. Thực tế cho thấy rất khó xây dựng một mô hình biểu diễn tri thức ngôn ngữ *không phụ thuộc ngôn ngữ* như đòi hỏi đối với *Liên ngữ*.

Trong phần này giới thiệu sơ lược một vài hệ dịch máy liên ngữ được nhắc tới nhiều trong thời gian gần đây.

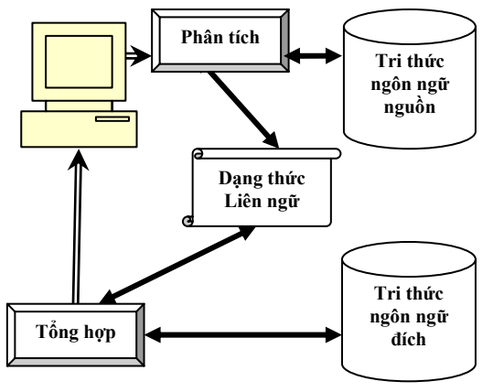
I.3.3.1. Dự án UNITRAN của MIT

Trong cách tiếp cận UNITRAN [11], các tác giả đã đề xuất mô hình dịch máy xử lý các tương quan giữa các ngôn ngữ mà không dựa trên những quy tắc phi ngữ cảnh phụ thuộc ngôn ngữ. Theo quan điểm của nhóm tác giả, nhiều hệ thống dịch máy không dựa trên mô hình liên ngữ phụ thuộc nặng nề vào các bộ quy tắc phi ngữ cảnh. Cách tiếp cận của UNITRAN đề

xuất một mô hình tính toán được gọi là hệ thống dựa trên nguyên lý (principle-based system). Trong UNITRAN, văn phạm được xem xét như một hệ thống các đơn thể – gọi là những nguyên lý – thay vì một tập lớn các quy tắc phụ thuộc ngôn ngữ.

Quá trình dịch trong UNITRAN chủ yếu là quá trình cú pháp, không có hệ thống 'hiểu' toàn cục nào. Hệ thống biên dịch từng câu rời rạc. Ngữ nghĩa chỉ áp dụng đối với việc tham chiếu tới những đại từ (chẳng hạn việc tương ứng giữa anh ấy với người đàn ông) hoặc việc gán vai trò ngữ nghĩa cho một số phần tử cụ thể trong câu, đặc biệt là các đối tượng của động từ (tân ngữ). Việc xác định ánh xạ giữa các động từ tương đương về ngữ nghĩa được xem là công việc không tầm thường. Chẳng hạn, mặc dù động từ нравиться trong tiếng Nga được xem là tương đương với động từ like trong tiếng Anh, nhưng các cấu trúc đối tượng của hai động từ này không giống nhau. Người được thích trong tiếng Anh là tân ngữ, còn trong tiếng Nga lại là chủ ngữ.

Mô hình dịch máy của UNITRAN thiết kế dựa trên các nguyên lý



Hình 1. Sơ đồ dịch máy UNITRAN

(principle-based). Ngôn ngữ nguồn đưa về dạng biểu diễn độc lập với mọi ngôn ngữ. Một bộ phân tích và một bộ tổng hợp duy nhất sử dụng chung cho mọi ngôn ngữ. Bộ phân tích và bộ tổng hợp có thể được lập trình (thông qua việc thiết đặt các thông số) để xử lý câu nguồn và câu đích. Chẳng hạn, người mô tả văn phạm có thể chỉ rõ rằng câu tiếng Anh đòi hỏi luôn luôn phải có chủ ngữ, còn câu tiếng Việt thì không nhất thiết phải có. Khi đó chỉ cần thiết đặt thông số null subject trong tiếng Việt giá trị true và trong tiếng Anh giá trị false. Mỗi ngôn ngữ đều có một bộ từ điển.

Quá trình biên dịch bao gồm 3 bước:

- Bộ phân tích thực hiện công việc phân tích từ vựng và tạo ra cấu trúc cây thể hiện mối liên hệ giữa các bộ phận của câu nguồn (Cấu trúc này là biểu diễn liên ngữ chung cho cả hai ngôn ngữ).

- Các thủ tục chọn và thay thế các phần tử ngữ vựng bằng những mục tương ứng của ngôn ngữ đích.
- Bộ tổng hợp thực hiện công việc tổng hợp ngữ vựng và sắp xếp lại trật tự câu cho ngôn ngữ đích.

Trong bước phân tích, thành phần xây dựng cấu trúc - một sự cài đặt của giải thuật phân tích Early (1970) - thực hiện việc dự đoán và phân tích ngữ vựng. Cấu trúc câu được tạo ra không chứa những thông tin về các thỏa thuận cú pháp, về vai trò ngữ nghĩa, về cấu trúc các thông số...Thành phần xác định kiến trúc ngôn ngữ sẽ hạn chế hoặc biến đổi các cấu trúc câu tuân thủ các nguyên lý để lọc ra các thỏa thuận ngôn ngữ, tình huống, điều kiện vai trò ngữ nghĩa...Cách thiết kế này thỏa mãn một số các nghiên cứu gần đây cho rằng con người khảo sát ngôn ngữ bằng cách gán sự phân tích cấu trúc sơ bộ (thường là nhập nhằng và chưa cụ thể) cho mệnh đề và sau đó mới thực hiện việc quyết định về từ vựng và ngữ nghĩa của nó.

Theo các tác giả, vì các ràng buộc ngôn ngữ luôn có sẵn trong quá trình phân tích, kích thước của văn phạm rất nhỏ gọn (không quá 150 quy tắc). Thuật toán Early có thể tăng thời gian thực hiện của nó lên 4 lần khi kích thước văn phạm tăng gấp đôi.

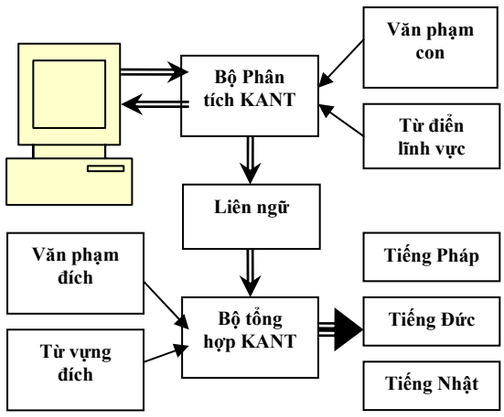
I.3.3.2. Dự án Dịch máy đa ngữ tại CICC.

ELT, CICC và chính phủ các nước Trung quốc, Malaisia, Indonesia, Thái lan đã hỗ trợ phát triển một hệ thống phiên dịch tự động đa ngữ [10]. Trong bài mô tả nội dung của Liên ngữ được chấp nhận của dự án. Đó là một đề án sáu năm bắt đầu từ 1987. Dự án nhằm xây dựng hệ dịch tự động cho các thứ tiếng Trung, Anh, Nhật, Thái Lan, Malaysia và Indonesia. Giải pháp Liên ngữ được chọn để thu được hiệu suất và chất lượng tốt cho hệ thống phiên dịch tự động đa ngữ. Liên ngữ được đặc trưng bởi những khái niệm (concepts) để loại bỏ sự phụ thuộc ngôn ngữ. Những khái niệm quan hệ và khái niệm thuộc tính được đề cập trong dự áni có những tính chất dùng chung cao giữa các ngôn ngữ. Mỗi khái niệm được phân loại vào trong cấu trúc có cấp bậc.

Dự án đã kết thúc năm 1994 mà chưa đưa ra được sản phẩm cuối cùng.

I.3.3.3. Dự án KANT của Trường đại học Carnegie Mellon

KANT (Knowledge-based, Accurate Natural-language Translation) là một bộ công cụ phần mềm phân tích văn bản nguồn và sản sinh bản dịch tự động hoặc tương tác, Hình 2. Nó được thiết kế để biên dịch văn bản kỹ thuật. KANT sử dụng những quy tắc từ điển, văn phạm, và ngữ nghĩa để thực hiện bản dịch. KANT là một hệ thống liên ngữ, sử dụng dạng biểu diễn trung gian như một "*Trục quay*" giữa ngôn ngữ nguồn và đích.



Hình 2. Kiến trúc của KANT

Có ba lợi điểm chính trong cách tiếp cận của KANT:

- Bản dịch có độ chính xác cao hơn.
- Hỗ trợ nhiều ngôn ngữ đích.
- Sự tách biệt của mã và cơ sở tri thức.

Toàn bộ phần mềm trong KANT (mô đun phân tích và tổng hợp) đều độc lập với cặp ngôn ngữ cần dịch. Không giống những hệ thống chuyển đổi, việc thêm ngôn ngữ đích mới không yêu cầu thiết kế lại dữ liệu.

Phạm vi ứng dụng của KANT:

- Biên dịch chính xác cho văn bản với văn phạm chặt chẽ;
- Văn bản được dịch tập trung trên một lĩnh vực hẹp thông tin kỹ thuật;
- Khi có đòi hỏi cao về độ chính xác.
- Khối lượng văn bản cần dịch đủ lớn để việc phát triển một hệ thống phiên dịch máy tự động là một đầu tư đáng giá;
- Những văn bản được tạo ra bởi một tổ chức đặc biệt, để thực hiện một tiêu chuẩn ngôn ngữ miền;
- Những văn bản cần phải được dịch sang hơn một ngôn ngữ đích.

Khi kiểm tra văn phạm, KANT có khả năng đoán nhận sự nhập nhằng trong văn bản nguồn (có thể dẫn dắt tới bản dịch không chính xác). KANT có một API (giao diện chương trình ứng dụng) cho phép nó thông báo tới người soạn thảo về sự vị trí và kiểu nhập nhằng, để hỏi về việc giải quyết nhập nhằng. Kiểm tra API văn phạm cho phép bộ phân tích KANT chạy trong khi xử lý, và giao tiếp với bất kỳ công cụ nào hỗ trợ API.

Hiện nay hệ dịch máy KANT và phiên bản hướng đối tượng của nó – KANTOO – vẫn chỉ đang được ứng dụng cho lĩnh vực chuyên môn hẹp mà chưa tìm được ứng dụng rộng rãi do chất lượng chưa được khẳng định.

I.3.4. CÁC PHƯƠNG HƯỚNG MỚI

Nhiều chuyên gia cho rằng những hệ thống dịch máy tương lai sẽ kết hợp phương pháp dựa vào kho ngữ liệu với cách tiếp cận trên nền quy tắc - chúng sẽ là những hệ thống lai [11, 14, 18, 22]. Chẳng hạn, dữ liệu ngôn ngữ của các hệ thống truyền thống được cung cấp dựa trên một ngân hàng kiến thức chuyên biệt, số liệu thống kê và ví dụ của văn bản được dịch sẵn. Trong cách tiếp cận này, quy tắc ngôn ngữ sẽ đơn giản hơn so với các hệ thống hiện thời, tức là sự phân tích cú pháp có thể hạn chế trong việc đoán nhận những cấu trúc và phần phụ thuộc mệnh đề bề mặt, sự phân tích ngữ nghĩa học sẽ hạn chế hơn, và thông tin từ vựng sẽ được lấy chủ yếu từ những nguồn chuẩn mực như từ điển đa dụng. Phương pháp trên nền kho ngữ liệu sẽ được sử dụng để tinh lọc việc phân tích các quy tắc cơ bản, để cải thiện sự chọn lọc từ vựng và để phát sinh văn bản mang tính thành ngữ hơn của ngôn ngữ đích. Cần phải nhấn mạnh rằng cách tiếp cận trên nền kho ngữ liệu còn phải được kiểm chứng đầy đủ, và chưa thể có một hệ thống thương mại được sử dụng rộng rãi sớm xuất hiện.

Một số dự án tham vọng nhất hiện nay là những hệ ***phiên dịch*** tiếng nói hạn chế trong lĩnh vực hẹp. Dự án ATR của Nhật là một hệ thống phục vụ liên lạc bằng điện thoại ở hội nghị quốc tế và phục vụ đăng ký chỗ khách sạn bằng điện thoại. Dự án Verbmobil của Đức nhắm vào việc phát triển một công cụ hỗ trợ xách tay phục vụ người Đức và người Nhật có thể đàm phán thương mại bằng tiếng Anh mà không cần phải biết tiếng Anh trôi chảy.

Dự án JANUS - một dự án hợp tác giữa ATR, Trường đại học Carnegie Mellon và Karlsruhe - cũng chỉ hạn chế trong lĩnh vực giao tiếp mang tính chất đàm phán và đăng ký hội nghị. Mỗi nhóm phát triển các mô đun nhận dạng và tổng hợp tiếng nói riêng rẽ cho từng ngôn ngữ (Nhật, Anh, Đức) và chương trình phiên dịch liên kết ngôn ngữ của họ với hai ngôn ngữ còn lại.

I.4. KẾT LUẬN

Hiện trạng của lĩnh vực dịch máy sau 50 năm nghiên cứu và phát triển trên thế giới cho thấy vẫn còn rất nhiều việc phải làm. Một số trở ngại chính trên con đường xây dựng các hệ dịch máy chất lượng cao bao gồm:

II. MỞ RỘNG MÔ HÌNH VĂN PHẠM

II.1. NHU CẦU MỞ RỘNG MÔ HÌNH VĂN PHẠM.....II-2

II.2. VĂN PHẠM ĐỊNH BIÊN (BOUND-CONTROLLED GRAMMAR) II-5

II.2.1. ĐỊNH NGHĨA.....II-5

II.2.2. ĐỊNH LÝ 1.....II-6

II.2.3. ĐỊNH LÝ 2.....II-8

II.3. VĂN PHẠM CẢM NGỮ ĐOẠN II-9

II.3.1. HỆ PHÂN CẤP KHÁI NIỆM.....II-10

II.3.2. TÍNH KHÔNG LIÊN TỤC NGỮ CẢNH.....II-10

II.3.3. RÀNG BUỘC NGỮ CẢNH – TÍNH CẢM NGỮ ĐOẠN.....II-11

II.3.4. ĐỊNH NGHĨA.....II-11

II.3.5. DẠNG MỞ RỘNG CỦA QUY TẮC CẢM NGỮ ĐOẠN.....II-14

II.3.6. SO SÁNH VỚI VĂN PHẠM CẢM NGỮ CẢNH.....II-15

II.3.7. XỬ LÝ NHẬP NHẰNG TRONG VĂN PHẠM CẢM NGỮ ĐOẠN II-17

II.4. KẾT LUẬN.....II-18

Phần này giới thiệu những đề xuất về văn phạm phục vụ việc dịch máy được phát triển tại Viện Ứng dụng Công nghệ.

II.1. NHU CẦU MỞ RỘNG MÔ HÌNH VĂN PHẠM

Những hạn chế của mô hình Văn phạm phi ngữ cảnh đã được đề cập nhiều [41], [42], [43], [44], [45], [46], [32], [47]. Trong [32] chúng tôi đã đề xuất ngôn ngữ định biên (được xác định bởi văn phạm định biên) như một bao đóng của lớp ngôn ngữ phi ngữ cảnh đối với phép giao. Một số tính chất của nó cho thấy đây là lớp ngôn ngữ có sức mạnh mô tả lớn hơn lớp ngôn ngữ phi ngữ cảnh nhưng lại có một đặc tính rất hữu ích là có thể kế thừa nhiều tính chất của ngôn ngữ phi ngữ cảnh, nhất là những kết quả liên quan đến độ phức tạp của các giải thuật phân tích văn phạm.

Ngôn ngữ tự nhiên là một thực thể hết sức phức tạp. Nhiều vấn đề hiển nhiên trong thực hành sinh ngữ lại rất khó, có khi không thể phát biểu dưới dạng các quy tắc của văn phạm sinh Chomsky.

Trước hết, văn phạm sinh không phải chỉ là công cụ cho phép "*sản sinh ra tất cả các câu thuộc một ngôn ngữ và không sản sinh ra gì ngoài những câu thuộc ngôn ngữ đó*", nó cần phải chỉ ra được (*một cách đúng đắn*) mối liên hệ giữa các thành phần của mỗi câu mà nó sản sinh ra.

Ta có thể quan sát việc áp dụng văn phạm vào phân tích và dễ dàng nhận thấy rằng cấu trúc cú pháp (như chúng ta hình dung một cách vô thức) thường khác với loại *cây cú pháp* được tạo thành khi vận dụng một văn phạm phi ngữ cảnh (xem [32]).

Để thể hiện được những đặc tính của ngôn ngữ tự nhiên, ta cần một công cụ hình thức mạnh hơn để :

- Mô tả sự liên hệ giữa các bộ phận khác nhau trong câu [32].
- Xây dựng mô hình cấu trúc câu với tổ chức gần gũi hơn với quan niệm trực quan (1) của con người
- Đưa vấn đề nhập nhằng cú pháp vào mô hình hình thức của văn phạm.

([1]) Bằng việc chấp nhận cây cú pháp trong đó các nút có số nhánh không hạn định (với mô hình Chomsky mỗi quy tắc đều có vế phải tất định, vì vậy số nhánh của mỗi nút đều xác định, cái biến thiên là độ sâu của cây cú pháp)

Để minh họa việc văn phạm có thể ảnh hưởng đến cách chúng ta xử lý tri thức ngôn ngữ như thế nào, ta khảo sát một ví dụ. Với mục đích làm cho kích thước của ví dụ minh họa nằm trong khuôn khổ hạn chế, ví dụ được trích dẫn ở đây không thuộc ngôn ngữ tự nhiên. Tuy nhiên, minh họa nhỏ gọn giúp ta hình dung được vấn đề.

Ví dụ 1.

Giả sử ta cần xây dựng văn phạm cho biểu thức số học với các phép toán *nhân* và *cộng* chẳng hạn :

a+b*c (1)

a*(b+c*e). (2)

. . . . .

Văn phạm thứ nhất G1 có tập quy tắc P1 bao gồm:

S → S + S | S * S | (S)

S → a | b | c | ...

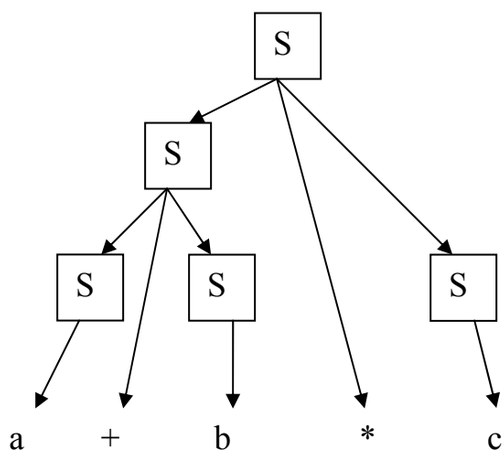
Văn phạm thứ hai G2 có tập quy tắc P2 bao gồm:

S → T | S + T

T → F | T * F

F → (S)

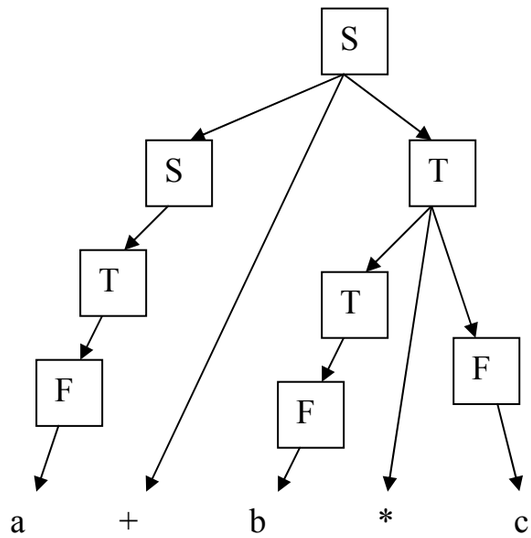
F → a | b | c | ...



Hình 1: Một cây cú pháp theo G1

Cả hai văn phạm đều mô tả đúng ngôn ngữ biểu thức số học (theo định nghĩa hình thức của văn phạm sinh). Tuy nhiên văn phạm G1 hoàn toàn

vô ích để ứng dụng vì các quy tắc của nó giải thích sai cấu trúc của các biểu thức số học (ở đây là trật tự ưu tiên các phép toán : *nhân chia trước, cộng trừ sau*). Trong khi đó, G2 phản ánh chính xác trình tự tính toán của biểu thức qua sự phân cấp của cây cú pháp.



Hình 2: Cây cú pháp theo G2

Hình 1 và Hình 2 mô tả cây cú pháp cho biểu thức a + b * c sử dụng các văn phạm tương ứng.[1] Dễ nhận thấy cây cú pháp trên Hình 1 thể hiện hoàn toàn sai cấu trúc của biểu thức. G1 không thể sử dụng trong các trình biên dịch (*compiler*) để phân tích biểu thức số học được.

Nội dung tiếp theo của phần này bao gồm :

Mục 1 giới thiệu văn phạm định biên, một mở rộng tất yếu của văn phạm phi ngữ cảnh để hình thành một lớp ngôn ngữ đóng kín với nhiều tính chất chung (*kế thừa các tính chất của ngôn ngữ phi ngữ cảnh*). Văn phạm định biên được định nghĩa dựa trên tính chất cảm ngữ đoạn chặt (*strict phrase-sensitivity*)

Mục 2 mô tả văn phạm cảm ngữ đoạn – một phát triển tiếp tục của văn phạm định biên với những tính chất hữu dụng trong xử lý ngôn ngữ tự nhiên.

[1] Đối với G1 mỗi biểu thức số học đều có thể tồn tại nhiều cây cú pháp khác nhau (đúng và sai).

II.2. VĂN PHẠM ĐỊNH BIÊN (BOUND-CONTROLLED GRAMMAR)

Trong thực hành ngôn ngữ tính phụ thuộc ngữ cảnh thường được nhắc đến để chứng tỏ rằng văn phạm phi ngữ cảnh là công cụ không đủ mạnh đối với ứng dụng xử lý ngôn ngữ tự nhiên.

Có nhiều phương pháp mở rộng mô hình văn phạm phi ngữ cảnh [44, 32, 45, 46, 47]. Mô hình văn phạm định biên đưa ra một cách tiếp cận trực tiếp : xây dựng bao đóng của lớp ngôn ngữ phi ngữ cánh đối với phép giao. Đây thực chất là mô hình hình thức của ý tưởng về nguyên lý văn phạm động được hình thành để phục vụ việc phân tích văn phạm [34]. Nội dung phần này là bản chỉnh sửa của [31], vì trong tài liệu đó có một sai sót đáng tiếc liên quan đến ***Định lý 2***.

II.2.1. ĐỊNH NGHĨA

Văn phạm định biên là bộ G = (Σ,N,S,P), trong đó:

- Σ là tập các từ cuối
- N là tập các từ trung gian
- S là từ xuất phát, S ∈ N
- P là tập các quy tắc định biên

Quy tắc định biên được định nghĩa đệ quy như sau:

- Quy tắc phi ngữ cảnh dạng A → ω là một quy tắc định biên
- Biểu thức A[R], trong đó R là quy tắc định biên và A là biến trung gian, là một quy tắc định biên.

Ta viết A[B → ω] và phát biểu rằng biến A kiểm tra biên của quy tắc B → ω nếu quy tắc này chỉ được áp dụng khi với mọi m ∈ Σ* sao cho B ⇒*ω⇒*m trong G thì m thuộc ngôn ngữ sinh bởi văn phạm G_A = (Σ,N,A,P \ {B → ω}).

Trong quy tắc định biên A[R], biến trung gian A được gọi là ***biến kiểm tra biên***, còn R là ***quy tắc sản xuất***.

Ngôn ngữ định biên là ngôn ngữ sinh bởi một văn phạm định biên.

Theo định nghĩa thì mọi quy tắc phi ngữ cảnh đều là quy tắc *định biên* (khi không có mặt biến kiểm tra biên).

II.2.2. ĐỊNH LÝ 1.

Giao của hai ngôn ngữ định biên là một ngôn ngữ định biên.

Chứng minh:

Giả sử ta có hai ngôn ngữ định biên L1 và L2 được sinh bởi các văn phạm

G1 = (Σ1, N1, S1, P1) và
G2 = (Σ2, N2, S2, P2) tương ứng, với điều kiện N1 ∩ N2 = Ø. Có thể thay đổi cách gọi *tên* các biến trung gian của một trong hai văn phạm để thỏa mãn điều kiện này.

Xây dựng ngôn ngữ định biên L với văn phạm được xác định như sau:

G = (Σ, N, S, P)

Trong đó:

- Σ = Σ1 ∪ Σ2
- N = N1 ∪ N2 ∪ {S}
- P bao gồm P1 ∪ P2, ngoài ra được bổ sung thêm các quy tắc sau:

S2[S → S1] (1)
S1[S → S2] (2)

Giả sử S1 ⇒* m (xâu m thuộc L1). Khi đó

- Nếu m thuộc L2 (S2 ⇒* m) thì khi áp dụng quy tắc (1), ta có :

  S ⇒* m, vì vậy m thuộc L.
- Nếu m không thuộc L2 thì quy tắc (1) không áp dụng được, vì vậy m không thuộc L.

Lập luận tương tự đối với trường hợp m thuộc L2. Như vậy L chính là giao của L1 và L2.

Hệ quả 1.1.

Giao của một tập hữu hạn các ngôn ngữ phi ngữ cảnh là một ngôn ngữ định biên.

Hệ quả 1.1 là kết quả trực tiếp của định lý 1 vì mỗi ngôn ngữ phi ngữ cảnh là một ngôn ngữ định biên.

Hệ quả 1.2.

Lớp ngôn ngữ phi ngữ cảnh là tập con thực sự của lớp ngôn ngữ định biên.

Chứng minh:

Hiển nhiên mỗi ngôn ngữ phi ngữ cảnh đều là một ngôn ngữ định biên.

Để chứng minh ta chỉ cần nêu ra sự tồn tại của ngôn ngữ định biên mà không phải là ngôn ngữ phi ngữ cảnh.

Ta biết rằng ngôn ngữ L = $\{a^n b^n c^n\}$ không phải là một ngôn ngữ phi ngữ cảnh.

Văn phạm G0 với các quy tắc sau:

U[S → T]
U → AJ
A → a | aA
J → bJc | bc
T → KC
K → aKb | ab
C → c | cC

là một văn phạm định biên.

Từ định nghĩa văn phạm G0 ta thấy:

- Việc áp dụng các quy tắc phân tích cho biến trung gian J luôn bảo tồn số lượng b và c bằng nhau.
- Việc áp dụng các quy tắc phân tích cho biến trung gian K luôn bảo tồn số lượng a và b bằng nhau.
- Quy tắc U[S → T] bảo đảm mọi xâu m thuộc ngôn ngữ đều có tính chất T $\Rightarrow^*$ m và U $\Rightarrow^*$ m.

Từ đây ta kết luận mọi xâu thuộc L có số lượng các chữ a, b, c luôn luôn bằng nhau, vì vậy ngôn ngữ sinh bởi G0 chính là $\{a^n b^n c^n\}$.

Hệ quả 1.2 cũng có thể chứng minh bằng cách xây dựng ngôn ngữ định biên có văn phạm G từ giao của hai ngôn ngữ phi ngữ cảnh sinh bởi G1 và G2 như sau:

Văn phạm G1 có các quy tắc:

S1 → AT
A→ a | aA
T → bc | bTc

Văn phạm G2 chứa các quy tắc

S2 → HC
H→ ab | aHb
C → c | cC

Áp dụng định lý 1, ta xây dựng văn phạm G = (Σ1 ∪ Σ2, N1 ∪ N2 ∪ {S}, S, P) bằng cách bổ sung các quy tắc

S2[S → S1]
S1[S → S2]

Ngôn ngữ L1 chứa các xâu $a^m b^n c^n$, còn ngôn ngữ L2 chứa các xâu $a^n b^n c^m$. Ngôn ngữ L sinh bởi G là giao của L1 và L2, vì vậy L chứa các xâu $a^n b^n c^n$.

II.2.3. ĐỊNH LÝ 2.

Mọi ngôn ngữ định biên đều là giao của một số hữu hạn các ngôn ngữ phi ngữ cảnh (*Cụ thể hơn, nếu văn phạm của một ngôn ngữ chứa n quy tắc định biên thì ngôn ngữ này có thể thể hiện được dưới dạng giao của không nhiều hơn 2^n ngôn ngữ phi ngữ cảnh*).

Chứng minh:

Giả sử L là ngôn ngữ sinh bởi văn phạm định biên G có n quy tắc định biên. Chọn một quy tắc định biên bất kỳ $P_i : A_i[B_i \to \omega_i]$ trong G, ta xây dựng các văn phạm sau:

- Gi là văn phạm G trong đó quy tắc $P_i : A_i[B_i \to \omega_i]$ được thay thế bằng quy tắc $P_i' : B_i \to A_i$
- Gi là văn phạm G trong đó quy tắc $P_i : A_i[B_i \to \omega_i]$ được thay thế bằng quy tắc $P_i' : B_i \to \omega_i$

Giả sử $s \in L_G$, khi đó dễ thấy rằng $s \in L_{Gi'}$ và $s \in L_{Gi'}$;

Giả sử $s \in L_{Gi}$ và $s \in L_{Gi'}$, giả sử trong Gi' ta có $S \Rightarrow^* ...Bi... \Rightarrow^* s$.

Khi đó vì $s \in L_{Gi}$ nên trong Gi tồn tại cách phân tích sao cho $S \Rightarrow^* ...Ai... \Rightarrow^* s$. Theo định nghĩa suy ra $s \in L_G$.

Từ đây có thể kết luận L_G là giao của L_{Gi} và $L_{Gi'}$. Mặt khác số quy tắc định biên trong Gi và Gi' đều là n-1 (giảm đi 1 so với G).

Thực hiện quá trình tương tự đối với mỗi văn phạm Gi và Gi' ta nhận được 4 văn phạm mới với số quy tắc định biên trong mỗi văn phạm là n–2. Sau n bước, ta nhận được 2^n văn phạm không còn chứa quy tắc định biên nào (nghĩa là đều trở thành văn phạm phi ngữ cảnh).

Văn phạm định biên kế thừa những tính chất của văn phạm phi ngữ cảnh. Các giải thuật ứng dụng trong khuôn khổ văn phạm phi ngữ cảnh đều có thể áp dụng cho văn phạm định biên với những điều chỉnh không đáng kể. Độ phức tạp của các giải thuật phân tích cho ngôn ngữ định biên tương đương với ngôn ngữ phi ngữ cảnh. Cùng với những tính chất khác của nó, ta

có thể áp dụng những kết quả toán học (chẳng hạn về lý thuyết dàn) vào việc khảo sát ngôn ngữ này.

Tuy nhiên, mô hình văn phạm định biên cũng tỏ ra chưa đủ tinh tế để mô tả các tính chất của ngôn ngữ tự nhiên. Trong định nghĩa của quy tắc định biên ta thấy bên cạnh một quy tắc phi ngữ cảnh thông thường có kèm theo một (hoặc nhiều) biến trung gian. Những biến này đóng vai trò kiểm soát việc áp dụng một quy tắc cụ thể. Tính chất cảm ngữ cảnh được thể hiện một cách gián tiếp. Quy tắc định biên xác định ranh giới ngữ đoạn và ràng buộc việc phân tích ngữ đoạn đó. Thực tế sử dụng ngôn ngữ đòi hỏi một hạn định ranh giới mờ hơn, ít chặt hơn so với tính định biên. Văn phạm cảm ngữ đoạn chính là sự mở rộng tự nhiên tiếp tục của.

II.3. VĂN PHẠM CẢM NGỮ ĐOẠN

Trong mục trước đã giới thiệu lớp ngôn ngữ định biên (được xác định bởi văn phạm định biên) như một bao đóng của lớp ngôn ngữ phi ngữ cảnh đối với phép giao. Một số tính chất của văn phạm này cho thấy đây là một mô hình có sức mạnh mô tả vượt ra ngoài phạm vi văn phạm phi ngữ cảnh, đồng thời kế thừa nhiều tính chất của văn phạm phi ngữ cảnh, nhất là những kết quả liên quan đến các giải thuật phân tích văn phạm.

Ngôn ngữ tự nhiên là một thực thể phức tạp. Nhiều vấn đề hiển nhiên trong thực hành sinh ngữ lại rất khó phát biểu dưới dạng các quy tắc văn phạm. Ta có thể quan sát việc áp dụng văn phạm vào phân tích câu và dễ dàng nhận thấy rằng cấu trúc ngữ pháp (như chúng ta hình dung một cách vô thức) thường khác xa với loại cây cú pháp được tạo thành khi vận dụng một văn phạm hình thức (xem [32]), cho dù đó là văn phạm cảm ngữ cảnh hay văn phạm phi ngữ cảnh. Văn phạm định biên cũng tỏ ra còn nhiều hạn chế khi mô tả các tính chất của ngôn ngữ tự nhiên.

Phần này mô tả một lớp văn phạm mới – *văn phạm cảm ngữ đoạn* – có khả năng mô tả được một số tính chất thường thấy trong ngôn ngữ tự nhiên mà các mô hình văn phạm quen biết hoặc không thể diễn đạt, hoặc diễn đạt dưới một hình thức không từ nhiên, hoặc, tệ hơn, dưới một hình thức phi lý, trái ngược hẳn với trực cảm của con người.

Văn phạm cảm ngữ đoạn được phát triển như một cố gắng xây dựng công cụ hình thức để :

- Mô tả hai khía cạnh trực giao của tri thức ngôn ngữ (cấu trúc sinh và trạng thái, xem [27]), và từ đó, mô tả được một số liên hệ giữa các câu trong bài văn.
- Mô tả sự liên hệ giữa các bộ phận (tách rời nhau) trong câu [32].

- Đưa vấn đề nhập nhằng cú pháp vào mô hình hình thức của văn phạm.
- Xây dựng mô hình cấu trúc câu có tổ chức gần gũi hơn với quan niệm trực quan [1] của con người

II.3.1. HỆ PHÂN CẤP KHÁI NIỆM

Trong văn phạm phi ngữ cảnh (*ta chỉ xét văn phạm không chứa quy tắc rỗng*), quy tắc sinh có hai dạng

$A \rightarrow m_1 m_2 ... m_n$; và (1)

$A \rightarrow m_0$; trong đó m_i là một từ cuối hoặc là một biến trung gian. (2)

Trong quy tắc loại (1), biến trung gian A được định nghĩa như một khái niệm mới, có các thành phần là m_1, m_2..., m_n . Trong khi đó quy tắc (2) xác định một phép *gán tên* cho một sự vật : biến A là sự khái quát hóa của m_0.

Ta cần phân biệt hai loại quy tắc này vì hai mục đích: hiệu năng tính toán và hiệu năng mô tả.

Tất cả các quy tắc loại 2 trong văn phạm có thể được tổ chức thành một dàn (*lattice*), sau đó có thể loại bỏ hoàn toàn chúng khỏi danh sách các quy tắc. Điều này dễ hiểu vì từ quan hệ phân cấp trong dàn các khái niệm, ta có thể dễ dàng sử dụng chúng để dựng *cây phân cấp ngữ nghĩa* tạo bởi chỉ các quy tắc loại 1.

Tất cả các quy tắc loại 1 cũng có thể tổ chức thành một dàn sao cho những quy tắc so sánh được với nhau là những quy tắc trong đó mỗi ký hiệu tương ứng thì so sánh được vợi nhau và có cùng tương quan.

II.3.2. TÍNH KHÔNG LIÊN TỤC NGỮ CẢNH

Trong [47] đưa ra một dạng thức khái quát hóa của văn phạm phi ngữ cảnh : văn phạm không liên tục ngữ cảnh (*Contextual Discontinuous Grammar*) trong đó các quy tắc phi ngữ cảnh được áp dụng hợp lệ khi chúng đồng thời có mặt trong cây cú pháp [2]

[1] Bằng việc chấp nhận cây cú pháp trong đó các nút có số nhánh không hạn định (với mô hình Chomsky mỗi quy tắc đều có vế phải tất định, vì vậy số nhánh của mỗi nút đều xác định, cái biến thiên là độ sâu của cây cú pháp)

[2] Chẳng hạn trong quy tắc không liên tục ngữ cảnh $A \rightarrow \omega$; $B \rightarrow \varphi$ có 2 quy tắc phi ngữ cảnh; việc áp dụng $A \rightarrow \omega$ và $B \rightarrow \varphi$ chỉ hợp lệ nếu tồn tại $\alpha\beta$ sao cho $S \Rightarrow^* \alpha A B \beta$ hoặc tồn tại $\alpha\beta\gamma\delta\varepsilon$ và C sao cho $S \Rightarrow^* \alpha A \beta C \gamma$ và $C \Rightarrow^* \delta B \varepsilon$

Một nhược điểm của văn phạm không liên tục ngữ cảnh là không có ràng buộc gì về văn cảnh đối với nhóm quy tắc phi ngữ cảnh trong một quy tắc không liên tục ngữ cảnh.

Về một khía cạnh nào đó, văn phạm cảm ngữ đoạn kế thừa mô hình văn phạm ngữ cảnh không liên tục [47]. Tuy nhiên, trong [47] không thấy bất cứ sự ràng buộc nào đối với tính không liên tục ngữ cảnh, chúng tôi cho rằng đây là một giả thiết trái với thực hành sinh ngữ : *sự ràng buộc lẫn nhau giữa các thành phần khác nhau (nằm cách xa nhau) trong thực tế chỉ có tác dụng trong phạm vi một ngữ đoạn cụ thể.*

II.3.3. RÀNG BUỘC NGỮ CẢNH – TÍNH CẢM NGỮ ĐOẠN

Ở đây ta đưa ra một mở rộng của tính định biên : tính xác định ngữ đoạn.

Quy tắc hạn định ngữ đoạn được định nghĩa đệ quy như sau:

- Quy tắc phi ngữ cảnh dạng $A \rightarrow \omega$ là một quy tắc hạn định ngữ đoạn
- Biểu thức A(R), trong đó R là quy tắc hạn định ngữ đoạn và A là biến trung gian, là một quy tắc hạn định ngữ đoạn.
- Ta viết $A(B \rightarrow \omega)$ và nói rằng biến A kiểm tra ngữ đoạn của quy tắc $B \rightarrow \omega$ nếu quy tắc này chỉ được áp dụng khi tồn tại $\alpha, \beta, \gamma, \delta$ sao cho $S \Rightarrow^* \gamma A \delta \Rightarrow^* \gamma\alpha B\beta\delta$.

Khác với quy tắc định biên, quy tắc hạn định ngữ đoạn chỉ ràng buộc việc áp dụng quy tắc trong *phạm vi* (*scope*) của một ngữ đoạn, mà không bắt buộc phải là biên của chính ngữ đoạn đó. Tính hạn định ngữ đoạn có thể được hiểu như tính chất ***cảm ngữ đoạn*** (*phrase-sensitivity*) của văn phạm hay là như tính chất ***cảm ngữ cảnh tổng quát*** (*generic context-sensitivity*) vì để thể hiện một ràng buộc hạn định ngữ đoạn, ta buộc phải thay thế bằng một họ (vô hạn tiềm năng) các ràng buộc cảm ngữ cảnh.

II.3.4. ĐỊNH NGHĨA

Chuỗi ký hiệu

- Mỗi *từ cuối* hoặc *biến trung gian* là một ký hiệu
- B (s) là một ký hiệu nếu B là biến trung gian và s là một chuỗi ký hiệu.

Văn phạm Cảm ngữ đoạn là bộ $G = (\Sigma, N, A, S, P)$, trong đó:

- Σ là tập các từ cuối.

- N là tập các biến trung gian.
- A là tập các tham đối (thuộc tính).
- S là từ xuất phát, S ∈ N.
- P là tập các quy tắc cảm ngữ đoạn.

Từ Σ, N và A ta xây dựng không gian trạng thái Ω như sau.

- Phần tử ⊤ (*khái niệm bất kỳ*) thuộc Ω,
- Phần tử ⊥ (*không gì cả*" thuộc Ω.
- Σ ⊂ Ω, N ⊂ Ω, A ⊂ Ω.
- X(a1,a2,…an) ∈ Ω nếu X ∈ N, ai ∈ A (i = 1...n), n >= 0.
- Tập Ω là một giàn đại số (< là phép toán so sánh trong dàn) với các tính chất:
- Với mọi m ∈ Σ tồn tại A ∈ Ω sao cho m < A.
- Không tồn tại m ∈ Σ và A ∈ N nào thỏa mãn A < m.
- Với mọi x ∈ Ω ta đều có ⊥ < x và x < ⊤.

Quy tắc cảm ngữ đoạn:

- Quy tắc dạng A → ω là một quy tắc cảm ngữ đoạn nếu ω là chuỗi các ký hiệu thuộc (Σ U N), trong đó mỗi ký hiệu có thể là một ký hiệu thông thường hoặc là một ký hiệu được *đánh dấu*
- Biểu thức A(R) là một quy tắc cảm ngữ đoạn nếu R là quy tắc cảm ngữ đoạn và A là biến trung gian.
- R1; R2 là một quy tắc cảm ngữ đoạn nếu R1, R2 là quy tắc cảm ngữ đoạn
- Ta viết A(B → ω) và nói rằng biến A là ***ngữ đoạn*** của quy tắc B → ω nếu quy tắc này chỉ được áp dụng khi tồn tại α, β sao cho A ⇒* αBβ

Phần tử thông thường có thể là hằng hoặc biến. Phần tử được đánh dấu là một biến. Để chỉ ký hiệu được đánh dấu trong quy tắc sinh, ta sử dụng dấu gạch dưới. Một phần tử được đánh dấu ở vế phải kế thừa những thuộc tính (hợp lệ, nếu có) của biến M ở vế trái (*khi phân tích từ trên xuống*), và biến M ở vế trái kế thừa các thuộc tính của tất cả các phần tử được đánh dấu ở vế phải (*khi phân tích từ dưới lên*).

Ví dụ quy tắc sinh:

- CụmDanhTừ → <u>CụmDanhTừ</u> TínhTừ
- CụmDanhTừ → <u>CụmDanhTừ</u> DanhTừ

- *CụmDanhTừ → DanhTừ - quy tắc (*suy biến*) này tương đương với phép toán DanhTừ < CụmDanhTừ

Trong ví dụ trên DanhTừ hoặc TínhTừ có thể bổ nghĩa cho CụmDanhTừ với số lượng không hạn chế. Với việc ***đánh dấu*** khái niệm CụmDanhTừ, ta tạo ra được một nút với số nhánh *biến thiên* (Hình 1).

TagQuestion → Subject VerbPhrase Tag

TagQuestion (Subject(Human) VerbPhrase (Op("***have***")) Tag("***haven't he***")

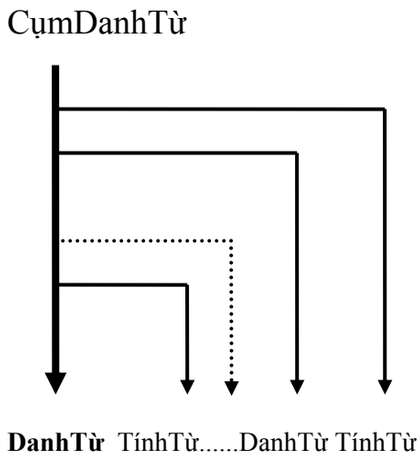
TagQuestion (Subject(Human(Plural)) VerbPhrase (Op("***aren't***")) Tag("***are they***")

TagQuestion (Subject(Human) VerbPhrase (Op("***didn't***")) Tag("***did he***")

...

Quy tắc thứ nhất là quy tắc sinh (phi ngữ cảnh) thông thường theo mô hình Chomsky.

Các quy tắc sau xác định sự thỏa thuận (*agreements*) giữa các thành phần trong câu.



Hình 1. Ví dụ cây con với số nhánh biến thiên (Sử dụng quy tắc có ký hiệu được đánh dấu)

Khác với quy tắc sinh Chomsky, quy tắc có đánh dấu chỉ rõ phần tử kế thừa trong chuỗi phần tử. Chẳng hạn, hai quy tắc sau

DanhTừ → DanhTừ DanhTừ

Noun → Noun Noun

có phần tử đánh dấu nằm ở các vị trí khác nhau (*mặc dù có dạng tương tự nhau*) phản ánh sự khác nhau giữa cách tổ chức cụm danh từ trong tiếng Việt và tiếng Anh – ở đây là phản ánh sự khác nhau về ngữ nghĩa giữa hai ngôn ngữ.

II.3.5. DẠNG MỞ RỘNG CỦA QUY TẮC CẢM NGỮ ĐOẠN.

Tương tự như trường hợp văn phạm phi ngữ cảnh, có thể sử dụng ký pháp mở rộng đối với quy tắc cảm ngữ đoạn (thông qua các ký hiệu siêu ngôn ngữ '{','}','[',']').

Dạng mở rộng cho phép mô tả các quy tắc văn phạm một cách trực quan; tuy nhiên nó không mở rộng hiệu lực mô tả. Có thể chứng minh được rằng mọi quy tắc cảm ngữ đoạn dạng mở rộng đều có thể đưa về một tập hữu hạn các quy tắc cảm ngữ đoạn dạng chuẩn.

Giả sử quy tắc văn phạm cho danh ngữ (tiếng Anh) có thể viết dưới dạng tổng quát (dạng quy tắc cảm ngữ đoạn mở rộng) như sau:

Np → {PreDet} {CentralDet} {PostDet} {PreModifier} N [PostModifier] (1)

Dễ thấy rằng phần tử chính xác định danh ngữ ở đây là danh từ N được đánh dấu.

Ta có thể xây dựng giải thuật để đưa quy tắc này về dạng quy tắc cảm ngữ đoạn dạng chuẩn (bằng cách đưa thêm những biến phụ $1, $2,…):

Các quan hệ giàn : N < $1 < $2 < $3 < $4 < $5 < Np

Các quy tắc :

- $1 → PreModifier $1
- $2 → PostDet $2
- $3 → CentralDet $3
- $4 → PreDet $4
- $5 → $4 PostModifier

Dễ thấy rằng tập quy tắc dạng chuẩn nêu trên xác dịnh cùng một ngôn ngữ con như ngôn ngữ con xác định bởi quy tắc (1).

Ví dụ văn phạm cảm ngữ đoạn :

G = ({a,b,c}, {S,A,B,C}, {}, {}, S, R) với R gồm có các quy tắc:

A → a

$B \rightarrow b$

$C \rightarrow c$

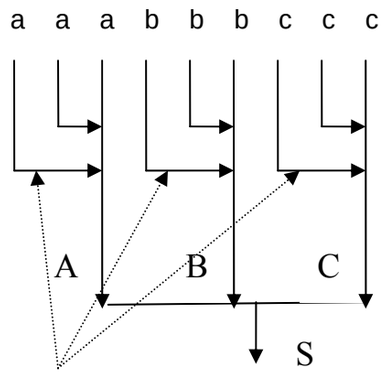
$A(A \rightarrow a\ A)$; $B(B \rightarrow b\ B)$; $C(C \rightarrow c\ C)$

$S \rightarrow ABC$

G là văn phạm sinh cho ngôn ngữ $L = a^n b^n c^n$. Thật vậy, các quy tắc a), b), c) chỉ có thể áp dụng một lần; quy tắc d) luôn đảm bảo số lượng các chữ a, b, c được sinh ra là như nhau.

Cây cú pháp cho câu "a a a b b b c c c" được biểu diễn trên Hình-1 (So sánh với [5]). Ví dụ này cho thấy quy tắc cảm ngữ đoạn bảo tồn mối liên hệ chéo (vốn phổ biến trong ngôn ngữ tự nhiên) mà mô hình văn phạm phi ngữ cảnh (và cảm ngữ cảnh) không thể thể hiện được[1].

Văn phạm trước hết cần phải thể hiện được cấu trúc ngữ nghĩa của văn bản chứ không phải chỉ để trả lời được câu hỏi liệu một câu cụ thể có thuộc ngôn ngữ đã cho hay không.



Hình 2. Cây cú pháp của $a^n b^n c^n$

Ví dụ này cho thấy văn phạm cảm ngữ đoạn có một ý nghĩa thực dụng nhất định do khả năng mô tả những ràng buộc ngôn ngữ tinh tế.

II.3.6. SO SÁNH VỚI VĂN PHẠM CẢM NGỮ CẢNH

Quy tắc cảm ngữ cảnh có dạng

$\alpha\ A\ \beta \rightarrow \alpha\ \omega\ \beta$

[1] Văn phạm định biên có thể mô tả được ngôn ngữ L như đã được trình bày trong phần trước, nhưng cách thức mà nó mô tả hoàn toàn không phản ánh được mối liên hệ ngữ nghĩa giữa các bộ phận trong câu.

ở đây, α và β là những xâu ký tự cụ thể. Ta phải làm thế nào khi có hàng loạt (vô hạn) các cặp α_i và β_i khác nhau đều cho phép áp dụng $\alpha_i A\beta_i \rightarrow \alpha_i\omega\beta_i$, miễn là có B, sao cho $B \Rightarrow^* \alpha_i A\beta_i$, mà chính tình huống này mới là hiện tượng thường gặp trong ngôn ngữ tự nhiên. Trong những bài giảng về ngữ pháp ta thường nghe thấy những hướng dẫn có dạng : nếu trong một cụm từ (một ngữ x cụ thể nào đó có mặt x thì x và y mang một nghĩa cụ thể...), nghĩa là yếu tố quan trọng để hiểu câu trúc ngữ cảnh ở đây là cụm từ x bao chung quanh phần tử x chứ không phải những ngữ cảnh cụ thể bên cạnh x. Hướng dẫn không hình thức trên đây có thể viết dưới dạng một phát biểu cảm ngữ đoạn như sau:

$$X\ (x \rightarrow \omega)$$

Để chuyển sự tương quan này sang dạng quy tắc cảm ngữ cảnh, ta phải xác định tất cả các tình huống ngôn ngữ để tìm ra họ các ngữ cảnh cụ thể α_i,β_i, cũng như các ngữ cảnh cụ thể γ_j,δ_j để xây dựng các bộ quy tắc:

$$\alpha_i x\beta_i \rightarrow \alpha_i\omega\beta_i$$

và các tập quy tắc sao cho $X \Rightarrow^* \gamma_j x\delta_j$

Vì vậy ta có thể coi văn phạm cảm ngữ đoạn là một loại văn phạm có tính cảm ngữ cảnh tổng quát (*generic context-sensitivity*).

Mặt khác, dạng thức quy tắc của văn phạm cảm ngữ đoạn được xây dựng trên nền tảng mô hình văn phạm phi ngữ cảnh. Vì vậy nhiều kỹ thuật quen thuộc có thể được tùy biến để tính toán với mô hình văn phạm này.

Đối với văn phạm cảm ngữ đoạn, việc đưa vào các biến (*nonterminal*) mới cần được dựa trên cơ sở thể hiện một khái niệm ngữ pháp cụ thể, không phải là một ẩn số phụ để xây dựng hệ quy tắc mô tả ngôn ngữ.

Xét ví dụ: Danh ngữ bao gồm một danh từ đi cùng với các từ bổ nghĩa tuân thủ một thứ tự nhất định. Để văn phạm mô tả được trật tự đó thường phải đưa vào những biến phụ trung gian, từ đó có thể dựng được một tổ hợp đa dạng nhiều cây cú pháp có dạng nhiều tầng với kiến trúc khác nhau mà không một cây nào thực sự là cây cú pháp mà ta hình dung một cách trực quan về danh ngữ [32].

Một ví dụ khác, xét mệnh đề:

"***gán cho các biến x, y, z giá trị a, b, c tương ứng với a = 1, b = 2, c = 3.***".

Trong câu này ta đều hiểu rằng *biến x phải được gán giá trị a, biến y phải được gán giá trị b và biến z phải được gán giá trị c*.

Câu có mối liên hệ chéo dạng này không thể diễn đạt bằng công cụ văn phạm phi ngữ cảnh, đồng thời cũng rất khó biểu diễn trong khuôn khổ văn

phạm cảm ngữ cảnh[1]. Áp dụng văn phạm cảm ngữ đoạn ta có thể dễ dàng mô tả dưới dạng các quy tắc sau:

S → "***gán các biến"*** VL "***giá trị"*** PL "***tương ứng với"*** CL (1)

VL → V (2-1)

PL → P (2-2)

CL → C (2-3)

VL(VL → V VL) ; PL(PL → P, PL) ;CL(CL → C, CL) (3)

Việc áp dụng quy tắc (3) bảo đảm x ứng với a, y ứng với b và z ứng với c và việc áp dụng quy tắc (2-i) bảo đảm số lượng các ký hiệu tương đương với nhau.

II.3.7. XỬ LÝ NHẬP NHẰNG TRONG VĂN PHẠM CẢM NGỮ ĐOẠN

Hệ thống giàn các đối tượng văn phạm tạo nên một môi trường nhất quán và tự nhiên để xử lý nhập nhằng. Quan hệ giữa các phần tử trong giàn các ký hiệu xác định quan hệ ưu tiên trong hệ luật sinh. Chẳng hạn ta xét các quan hệ :

look < Vp3a, (1)

for < Preposition (2)

và nhóm quy tắc (các từ cuối (terminal) được đánh dấu đậm):

1. VerbPhrase → Vp3a Preposition Object
2. VerbPhrase → **look for** Object

Ta phát biểu chuỗi ký hiệu "***look for*** *Object*" là thành ngữ hơn chuỗi ký hiệu "*Vp3a Preposition Object*"

Quy tắc 2. có độ ưu tiên cao hơn quy tắc 1. Ta viết 2. < 1. (dấu < là phép toán trong giàn các đối tượng). Kết luận này được dẫn xuất từ các quan hệ 1. và 2 (vì **look <** Vp3a và **for** < Preposition) nên ta có

VerbPhrase → **look for** Object < VerbPhrase → **look** Preposition Object

và

VerbPhrase → **look** Preposition Object< VerbPhrase → Vp3a Preposition Object

cho nên

[1] Văn phạm cảm ngữ cảnh mô tả cấu trúc trên một cách không tự nhiên và không tương ứng về ngữ nghĩa.

VerbPhrase → **look for** Object < VerbPhrase → Vp3a Preposition Object

Tương tự, hệ phân cấp các tham đối cũng đưa ra một phương thức ngầm định để giải quyết nhập nhằng. Chẳng hạn, trong khi phân tích một đoạn văn bản, hệ thống đã xác định được lĩnh vực cụ thể (được biểu diễn bằng ký hiệu A trong giàn các tham đối). Khi đó, để phân tích câu hiện thời, nếu như một từ cụ thể là thuật ngữ thuộc nhiều lĩnh vực khác nhau (với ngữ nghĩa khác nhau) thì bộ phân tích có thể chọn được lĩnh vực thích hợp (được biểu diễn bằng ký hiệu B) nếu thỏa B < A hoặc B = A.

Từ mô hình giải quyết nhập nhằng có thể xây dựng những ***hàm định giá*** dưới dạng tổ hợp của những phép so sánh như đã nêu trên. Khả năng lựa chọn phương án (từ vựng và cú pháp) của bộ xử lý nhập nhằng càng tinh tế khi dàn các tham đối càng được xây dựng chi tiết và phong phú hơn.[1]

II.4. KẾT LUẬN

Mô hình hình thức của văn phạm có tác dụng để lý giải những cách thức con người thu thập và xử lý tri thức ngôn ngữ, từ đấy đưa ra những công nghệ xử lý ngôn ngữ tự nhiên. Theo quan điểm của Chomsky về kiến trúc văn phạm phổ quát thì con người xử lý theo một cách thức giống nhau đối với mọi ngôn ngữ, và vì vậy, có thể phần nào mô hình hóa để thao tác bằng máy tính. Mô hình văn phạm cảm ngữ đoạn là một mở rộng tự nhiên của văn phạm Chomsky, và cho phép mô tả nhiều tình huống ngôn ngữ thực một cách tự nhiên và trực tiếp. Vì vậy, nó có thể là công cụ hình thức tốt cho công nghệ ngôn ngữ nói chung cũng như dịch máy nói riêng. Như được trình bày trong [27], văn phạm cảm ngữ đoạn còn cho ta một phương tiện khả thi sáng sủa để cài đặt một mô hình dịch máy được xem là tiên tiến hiện nay – mô hình dịch máy liên ngữ.

Có một quan niệm chung của cộng đồng những người hoạt động trong lĩnh vực xử lý ngôn ngữ tự nhiên là *cần tìm cách hạn chế mô hình văn phạm để hạn chế độ phức tạp tính toán.* Trong khi đó chúng ta đều cảm nhận rằng *khối lượng tính toán* cần thiết để phân tích ngôn ngữ là không nhiều. Bài toán dịch máy khác một cách cơ bản với những bài toán nặng về xử lý (như chơi cờ, nén ảnh,...). Một mặt, với cùng một bài văn, một trăm người dịch sẽ cho ra một trăm bản dịch khác nhau và tất cả các bản dịch ấy, nói chung, đều đúng. Mặt khác, một người dịch có kinh nghiệm sẽ luôn luôn thực hiện công việc dịch thuật nhanh hơn hẳn người mới vào nghề : tri thức càng nhiều sẽ giúp cho người dịch càng nhanh chóng ***tạo ra*** được bản dịch đúng

[1] Sự phân cấp này trong thực tế được tổ chức theo một mô hình phức tạp hơn trên cơ sở chọn giàn con tối thiểu, với nội dung chi tiết sẽ được trình bày trong một báo cáo khác.

chứ không phải làm cho việc dịch thuật phải *xử lý nhiều thông tin hơn* trước khi đưa ra bản dịch đúng. Điều đó có nghĩa *người dịch* không phân tính nhiều mà thường căn cứ vào các *mẫu đoạn* có sẵn theo kinh nghiệm (trong vốn hiểu biết của mình) để sản sinh bản dịch thích hợp.

Như vậy, bản chất của công việc biên dịch mang nhiều tính chất ***lựa chọn*** phương án chấp nhận được trong số các phương án. Vì vậy, mô hình văn phạm cũng cần cung cấp công cụ hình thức để *so sánh, đánh giá* các cấu trúc ngôn ngữ – các *ngữ đoạn*. Từ đây, ta có thể nhận thấy rằng văn phạm sinh không phải chỉ "*sản sinh ra tất cả các câu thuộc một ngôn ngữ và không sản sinh ra gì ngoài những câu thuộc ngôn ngữ đó*". Nó có thể sản sinh ra ***những câu không hoàn toàn đúng*** bên cạnh những câu đúng. Tri thức ngôn ngữ sẽ cho phép (con người hoặc ứng dụng xử lý ngôn ngữ tự nhiên) so sánh và chọn ra được những câu đúng hơn trong số các câu đồng nghĩa nhưng khác nhau về cách đặt câu, thành ngữ hay lỗi văn phạm.... Văn phạm cảm ngữ đoạn tích hợp công cụ hình thức để so sánh, lựa chọn mẫu như một thành phần của mô hình.

III. GIẢI PHÁP DỊCH MÁY

III.1. BIỂU DIỄN TRI THỨC NGÔN NGỮ III-2

III.2. NHỮNG YÊU CẦU ĐỐI VỚI LIÊN NGỮ III-3
III.2.1. TÍNH ĐẦY ĐỦ (VẠN NĂNG) III-3
III.2.2. TÍNH KHẢ NGHỊCH (ĐỐI XỨNG) III-3
III.2.3. TÍNH ĐƠN TRỊ (KHÔNG NHẬP NHẰNG) III-4
III.2.4. TÍNH TỐI TIỂU III-5

III.3. KIỂM CHỨNG LIÊN NGỮ III-6
III.3.1. NGÔN NGỮ TỰ NHIÊN III-6
III.3.2. CẤU TRÚC CÚ PHÁP (SYNTAX STRUCTURE) III-7
III.3.3. LOGIC MỆNH ĐỀ (PROPOSITIONAL LOGIC) III-7
III.3.4. KHÁI NIỆM PHỔ QUÁT (CONCEPTUAL UNIVERSALS) III-7

III.4. TỔ CHỨC TRI THỨC ĐA NGÔN NGỮ III-7
III.4.1. HAI KHÍA CẠNH CỦA NGÔN NGỮ III-8
III.4.2. CÂY PHÂN CẤP NGỮ NGHĨA III-10
III.4.3. PHÂN HOẠCH LIÊN NGỮ III-15

III.5. PHƯƠNG PHÁP DỊCH MÁY III-21
III.5.1. SƠ ĐỒ DỊCH MÁY III-22
III.5.2. HÌNH THỨC HÓA III-24
III.5.3. NHỮNG ÍCH LỢI CỦA PHƯƠNG PHÁP III-25

Phần này này đề cập cách tiếp cận về dịch máy đang được phát triển tại Viện Ứng dụng Công nghệ (NACENTECH). Phương pháp được đề xuất dựa trên sự phân hoạch tri thức ngôn ngữ thành nhiều khối dựa theo sự tương đồng về cách biểu diễn ngôn ngữ. Công cụ hình thức để mô tả ngôn ngữ là văn phạm cảm ngữ đoạn. Phương pháp giống với cách tiếp cận dịch máy liên ngữ ở chỗ khi biên dịch, câu văn của ngôn ngữ nguồn được đưa về một dạng biểu diễn trung gian không phụ thuộc ngôn ngữ, sau đó từ dạng thức trung gian này mới tổng hợp thành câu văn của ngôn ngữ đích. Nội dung phần này trình bày cơ sở lý thuyết của cách tiếp cận, kiến trúc của liên ngữ và sơ đồ mô hình dịch máy được đề xuất.

III.1. BIỂU DIỄN TRI THỨC NGÔN NGỮ

Trong các hệ dịch máy (và các hệ xử lý ngôn ngữ tự nhiên) đều đòi hỏi phải có một dạng thức biểu diễn tri thức ngôn ngữ nào đó để tiện cho việc xử lý bằng máy.

Bản thân ngôn ngữ tự nhiên cũng chính là một cách biểu diễn tri thức ngôn ngữ. Tuy nhiên, chọn ngôn ngữ nào? Tiếng Việt, tiếng Anh, tiếng Nga,... đều là không thích đáng và không đặc trưng. Dạng thức biểu diễn tri thức ngôn ngữ phổ biến hiện nay là cây cú pháp với các biến thiên khác nhau (với các nút được gán thông tin từ vựng và ngữ nghĩa).

Cho dù cách thức biểu diễn thông tin ngôn ngữ ra sao thì cũng cần phải thỏa mãn những tính chất nhất định nếu muốn sử dụng để biểu diễn cho nhiều ngôn ngữ khác nhau.

Cây cú pháp, như ta đã biết, tỏ ra không thích hợp vì quá ràng buộc vào trình tự đặt câu cũng như cấu trúc cú pháp của một ngôn ngữ cụ thể.

Trước hết, ta sẽ xem xét về những thuộc tính cần thiết của liên ngữ (nếu tồn tại). Việc khảo sát những thuộc tính này có thể sẽ giúp chúng ta xây dựng được mô hình liên ngữ đáp ứng tốt nhất những yêu cầu xử lý ngôn ngữ của chúng ta1. Khi đó liệu việc lấy một ngôn ngữ tự nhiên nào đó (chẳng hạn tiếng Việt), hay một ngôn ngữ nhân tạo (như esperanto) làm liên ngữ thì có phải là cách lựa chọn tốt hay không?2

[1] Như vậy trước khi xây dựng kiến trúc của phương pháp biểu diễn tri thức ngôn ngữ , ta cần tìm hiểu xem nếu kiến trúc bên trong đó tồn tại thì nó cần phải thỏa mãn những tính chất gì?.
[2] Trong thực tế có một dự án chọn Esperanto với tư cách là liên ngữ [2].

III.2. NHỮNG YÊU CẦU ĐỐI VỚI LIÊN NGỮ

Những tính chất mong muốn của liên ngữ bao gồm các đề xuất về những thuộc tính cần có của một hệ thống để hệ thống đó có thể đóng vai trò như một liên ngữ. Chúng tôi đã cố gắng tìm hiểu nhưng chưa thấy có tài liệu nào đề cập đến vấn đề này. Những tính chất dưới đây được đưa ra với mục đích làm căn cứ để xây dựng một công cụ hình thức mô tả ngôn ngữ hiệu quả và dễ điều khiển hơn

III.2.1. TÍNH ĐẦY ĐỦ (VẠN NĂNG)

Một hệ thống biểu diễn tri thức có thể được coi là đầy đủ nếu nó chứa trong mình đủ tri thức (đủ các khái niệm (*concepts*)) để biểu diễn mọi tri thức ngôn ngữ của mọi ngôn ngữ tự nhiên.

Bất kỳ ngôn ngữ nào cũng đều có đủ phương tiện để biểu diễn bất kỳ tri thức nào cho nên ngôn ngữ tự nhiên là đầy đủ. Tuy nhiên, mỗi ngôn ngữ có một cơ sở từ vựng hạn định và vì vậy nhiều khái niệm cơ sở của ngôn ngữ này khi diễn đạt trên ngôn ngữ khác thì phải dùng tổ hợp từ. Tình huống này được gọi là lỗ hổng từ vựng (*lexical holes*). Chẳng hạn trong tiếng Anh các từ *brother, sister* được sử dụng chung cho cả "*anh, em trai*" hay "*chị, em gái*" trong tiếng Việt. Trong khi đó, từ *em* trong tiếng Việt lại cũng được sử dụng chung cho các khái niệm tổ hợp từ chỉ em trai hay em gái trong tiếng Anh[1]. Như vậy ta phải có các khái niệm phân biệt *anh (trai)*, *em trai, chị (gái), em gái* trong liên ngữ giả định.

Đây chính là cách quan niệm phổ biến về liên ngữ hiện nay. Rõ ràng thực hiện điều này là vô cùng khó (thực tế là *không thể*). Tuy nhiên, nếu không thỏa mãn được tính chất này thì không thể có được hệ dịch máy hoàn hảo.

Một cách hiểu khác của tính vạn năng là khi hệ thống có đủ *phương tiện* để có thể biểu diễn bất kỳ tri thức ngôn ngữ nào cho bất kỳ một ngôn ngữ nào. Như vậy, liên ngữ theo cách hiểu này không phải là một ngôn ngữ mà là một khung - một hệ thống ký pháp để biểu diễn tri thức ngôn ngữ. Đi theo cách tiếp cận này ta có thể làm giàu kho tri thức tùy thuộc vào vốn tri thức của từng ngôn ngữ được đưa vào hệ thống.

III.2.2. TÍNH KHẢ NGHỊCH (ĐỐI XỨNG)

Hệ thống ký pháp mô tả ngôn ngữ khả nghịch khi có thể sử dụng cùng một công cụ để chuyển đổi văn bản ngôn ngữ tự nhiên sang liên ngữ và ngược lại. Nếu hệ thống không khả nghịch thì với mỗi ngôn ngữ cần có một

[1] Và nhiều khái niệm khác (*em* là đại từ nhân xưng tiếng Việt).

bộ văn phạm phân tích và một bộ văn phạm tổng hợp riêng rẽ. Các hệ dịch máy theo phương pháp chuyển đổi thường đặt trọng tâm vào văn phạm phân tích : ứng với mỗi quy tắc văn phạm của ngôn ngữ nguồn đều có chỉ dẫn tương ứng về cách tổng hợp câu văn của ngôn ngữ đích (luật dịch). Không tồn tại *Văn phạm tổng hợp* như một hệ ký pháp độc lập[1].

Tính khả nghịch phụ thuộc vào cả kiến trúc của liên ngữ lẫn mô hình văn phạm và tùy thuộc vào sức mạnh mô tả của văn phạm cũng như mức độ tương đồng của ngôn ngữ với liên ngữ. Vì liên ngữ được giả thiết là vạn năng (nó đứng trung gian giữa mọi ngôn ngữ tự nhiên) nên tính khả nghịch tùy thuộc trước hết vào kiến trúc của chính liên ngữ.

III.2.3. TÍNH ĐƠN TRỊ (KHÔNG NHẬP NHẰNG)

Hệ thống đơn trị nếu mỗi cấu trúc chỉ có thể được hiểu theo một cách duy nhất. Mọi ngôn ngữ tự nhiên đều nhập nhằng (đều không đơn trị) theo nghĩa một câu văn có thể hiểu theo vài cách khác nhau.

Một hệ thống ký pháp hình thức để mô tả ngôn ngữ cần có phương tiện hiển ngôn để diễn đạt và xử lý tính nhập nhằng của ngôn ngữ tự nhiên.

Nếu liên ngữ là đơn trị thì mỗi câu văn nhập nhằng trên một ngôn ngữ tự nhiên nào đó sẽ có thể được chuyển sang một số cấu trúc liên ngữ phân biệt, mà mỗi cấu trúc đó chỉ có thể hiểu theo một cách duy nhất. Nghĩa là một câu văn nhập nhằng khi *dịch* sang liên ngữ thì thành một tập hữu hạn các *bản dịch liên ngữ* khác nhau (mà mỗi bản dịch đó đều không nhập nhằng) chứ không phải thành **một** *bản dịch liên ngữ* nhập nhằng. Mỗi cấu trúc liên ngữ này, khi dịch ngược lại sang ngôn ngữ ban đầu, sẽ lại sinh ra một tập hữu hạn các câu văn khác nhau mà mỗi câu đều có thể nhập nhằng.

Ví dụ 1.

Mệnh đề *"Những nơtron và điện tử tích điện âm"* có thể hiểu theo các dạng là:

1. *((Những nơtron) và điện tử) tích điện âm – những nơtron tích điện âm, điện tử tích điện âm*
2. *Những (nơtron và (điện tử tích điện âm)) – những nơtron không tích điện âm, những điện tử tích điện âm*
3. *Những ((nơtron và điện tử) tích điện âm) – những nơtron tích điện âm, những điện tử tích điện âm*

[1] Như vậy, với phương pháp chuyển đổi, cần phải có những bộ luật dịch, chẳng hạn, từ tiếng Anh và từ tiếng Pháp sang tiếng Việt riêng rẽ; đồng thời bộ văn phạm tiếng Việt để dịch sang tiếng Anh cũng không giống với bộ văn phạm tiếng Việt để dịch sang tiếng Pháp.

4. *(Những (nơtron)) và (điện tử tích điện âm) – những nơtron không tích điện âm, điện tử tích điện âm*

Để so sánh giữa các bản dịch (*tương ứng với cùng một câu văn*) với nhau, ta có thể dựa trên sự nhập nhằng. Ta sẽ gọi một bản dịch là ***bản dịch tốt*** nếu bản dịch đó chuyển tải được sự đa nghĩa của văn bản gốc, nghĩa là nó giữ nguyên được sự nhập nhằng của câu nguồn. Trong trường hợp sử dụng một liên ngữ không nhập nhằng, sự lựa chọn *bản dịch tốt* có thể được thực hiện một cách tường minh theo giải thuật sau:

- Dịch câu văn ngôn ngữ nguồn *s* sang liên ngữ, nhận được các *bản dịch liên ngữ* $l_1,..., l_i,..., l_n$
- Từ mỗi *bản dịch liên ngữ* l_i dịch sang ngôn ngữ đích nhận được *mi* bản dịch $t_{i1}, t_{i2},..., t_{imi}$, là những câu văn của ngôn ngữ đích.
- Dịch các câu văn của ngôn ngữ đích $t_{11}, t_{12},..., t_{1m1},..., t_{i1}, t_{i2},..., t_{imi},..., t_{n1}, t_{n2},... t_{nmn}$ sang liên ngữ (sau khi loại bỏ những phiên bản trùng nhau), chọn câu văn t_j nào có tập bản dịch liên ngữ ($l_{tj1}, l_{tj2},..., l_{tjk}$) trùng hoặc gần nhất với tập $\{l_1, l_2,...l_n\}$

Theo định nghĩa trên, về hình thức, bản dịch tốt (tiếng Anh) của mệnh đề *"Những nơtron và điện tử tích điện âm"* sẽ là *"Negatively charged electrons and neutrons"* hoặc *"Negatively charged electron and neutrons"* chứ không phải là *"Negatively charged neutrons and electrons"* hoặc *"Neutrons and negatively charged electrons"* hoặc*"Neutron and negatively charged electrons"* ... vì các nhập nhằng của hai mệnh đề đầu có ít khác biệt với những nhập nhằng của câu tiếng Việt hơn so với các câu sau.

III.2.4. TÍNH TỐI TIỂU

Hệ thống được hiểu là tối tiểu nếu những mệnh đề đồng nghĩa thì có cùng một dạng biểu diễn liên ngữ. Điều mong muốn là những sự khác biệt nhỏ về ngữ nghĩa sẽ được thể hiện bằng những sự khác biệt cũng nhỏ trong liên ngữ bất kể sự khác biệt đáng kể trong cách đặt câu của ngôn ngữ tự nhiên. Chẳng hạn rằng ta muốn hai câu sau có biểu diễn liên ngữ giống nhau (hoặc gần giống nhau) mặc dù trật tự cũng như cấu trúc câu (*cách đặt câu*) trong tiếng Việt là hoàn toàn khác nhau.

Ví dụ 1:

- *Hôm qua con chó đuổi con mèo*
- *Con mèo bị con chó nó đuổi hôm qua.*

Hoặc các cụm từ "*Bàn bằng gỗ màu vàng to tướng*", "*Bàn màu vàng to tướng bằng gỗ*", "*Bàn to tướng bằng gỗ màu vàng*",... đều có chung một

nội dung, chẳng qua trình tự sắp xếp từ ngữ hơi khác nhau. Ta mong muốn có một biểu diễn liên ngữ chung cho chúng.

Một đặc trưng nữa của tính tối tiểu là kiến trúc của liên ngữ phải đơn giản, dễ hiểu, trực quan : Ta mong muốn rằng biểu diễn liên ngữ của một câu tiếng Việt đơn giản thì cũng phải đơn giản.

Một hệ thống ký pháp (độc lập) được sử dụng như liên ngữ thì phải có khả năng mô tả được tri thức của mọi ngôn ngữ tự nhiên. Từ đây có thể nghĩ rằng nó phải chứa những phần tử bên ngoài một ngôn ngữ cụ thể, bên ngoài nhóm ngôn ngữ cụ thể; và những phần tử đó thì không cần thiết để biên dịch văn bản giữa các ngôn ngữ trong nhóm được chọn. Trong khi đó ta biết rằng mỗi ngôn ngữ đều có thể diễn đạt được bất kỳ điều gì mà một ngôn ngữ khác có thể diễn đạt. Như vậy, các cách diễn đạt khác nhau của những ngôn ngữ khác nhau (hay của cùng một ngôn ngữ) về cùng một nội dung thì cùng có biểu diễn liên ngữ giống nhau.

Ngoài bốn thuộc tính trên, có thể kể thêm những thuộc tính khác, hữu ích cho việc xử lý bằng máy tính, như tính độc lập ngôn ngữ, khả năng mô tả sự liên hệ giữa các câu khác nhau trong đoạn văn, mức độ tuân thủ các mô hình hình thức hiện có,…nhưng bốn tính chất trên là là nền tảng để kiểm chứng một hệ ký pháp có thể là một liên ngữ *tốt* hay không.

III.3. KIỂM CHỨNG LIÊN NGỮ

Với việc nêu ra những yêu cầu trên, ta có thể sử dụng chúng để kiểm chứng các mô hình được vận dụng như liên ngữ.

III.3.1. NGÔN NGỮ TỰ NHIÊN

Một thực tế là mỗi ngôn ngữ đều có thể mô tả được bất kỳ khái niệm nào của mọi ngôn ngữ khác. Như vậy, ngôn ngữ tự nhiên là đầy đủ. Ngôn ngữ tự nhiên, nói chung, không đối xứng, không đơn trị và không tối tiểu. Như vậy, Esperanto cũng như bất kỳ một ngôn ngữ tự nhiên nào đều không thích hợp để làm liên ngữ. Ta có thể nghĩ rằng một hệ thống nếu tối tiểu thì không khả nghịch vì nếu liên ngữ là tối tiểu thì từ một câu văn, chẳng hạn, tiếng Việt, sau khi chuyển sang liên ngữ (giả định nào đó), ta sẽ không thể khôi phục lại câu ban đầu do tiếng Việt là không đơn trị. Tuy nhiên, sự thực không phải như vậy; vì từ cấu trúc liên ngữ ta có thể có được một tập các câu tiếng Việt, trong đó (chắc chắn) có câu ban đầu, cho nên nó vẫn có thể là khả nghịch.

III.3.2. CẤU TRÚC CÚ PHÁP (SYNTAX STRUCTURE)

Vì cấu trúc cú pháp có thể được xây dựng cho bất kỳ mệnh đề nào của mọi ngôn ngữ nên nó đương nhiên là đầy đủ. Cấu trúc cú pháp hoàn toàn phụ thuộc ngôn ngữ nên nó không đối xứng. Với mỗi câu đa nghĩa có thể dựng được nhiều cấu trúc cú pháp khác nhau và ứng với mỗi cấu trúc cú pháp chỉ có một câu xác định nên cấu trúc cú pháp là đơn trị (không nhập nhằng về cú pháp). Bên cạnh đó, có thể diễn đạt cùng một nội dung bằng nhiều cấu trúc câu khác nhau, (và vì vậy, bằng nhiều cấu trúc nổi hay cấu trúc sâu dạng Chomsky khác nhau) nên tổ chức dạng cấu trúc cú pháp là không tối tiểu.

III.3.3. LOGIC MỆNH ĐỀ (PROPOSITIONAL LOGIC)

Xuất phát từ quan điểm cho rằng liên ngữ là phương tiện để biểu diễn nội dung, vì vậy logic mệnh đề được xem là công cụ để biểu diễn nội dung của văn bản thông qua các biểu thức logic. Mặc dù nói chung có thể suy diễn ra các biểu thức logic từ mỗi câu nguồn, rất khó lựa chọn bản dịch thích hợp và cũng rất khó hình thành hệ thống khái niệm cơ sở của tri thức. Logic mệnh đề đồng thời cũng là một mô hình rút gọn và phiến diện của tri thức nên không thể là đầy đủ.

III.3.4. KHÁI NIỆM PHỔ QUÁT (CONCEPTUAL UNIVERSALS)

Hệ thống các concepts [11] được xây dựng với mục đích mô tả được mọi đơn vị ngữ nghĩa nhỏ nhất của mọi ngôn ngữ tự nhiên. Như vậy, nhìn chung có thể đạt tới sự đầy đủ[1]. Hệ khái niệm phổ quát là không tối tiểu vì từ vựng của các ngôn ngữ khác nhau thì khác nhau. Nói chung, có thể tổ chức các khái niệm sao cho thỏa mãn các tính chất đối xứng và đơn trị. Tuy nhiên cho đến nay trên thế giới chưa có một hệ thống dịch máy thương phẩm dựa trên hệ thống concept nào được thực hiện trọn vẹn vì việc xây dựng toàn bộ các khái niệm tri thức cho mọi ngôn ngữ là không khả thi.

III.4. TỔ CHỨC TRI THỨC ĐA NGÔN NGỮ

Dữ liệu về ngôn ngữ – *Cơ sở tri thức* – là phần chủ đạo trong mọi hệ dịch máy. Tri thức ngôn ngữ bao gồm hệ từ vựng, các quy luật biến đổi và sử dụng từ, hệ thống phân loại từ ngữ, văn phạm, thành ngữ...Việc tổ chức

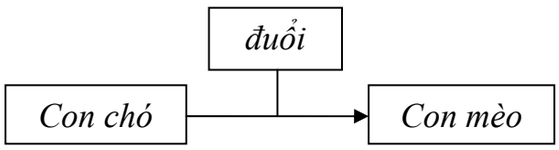
[1] Hệ thống concept là dư thừa vì khi dịch cho những ngôn ngữ rất gần nhau (chẳng hạn từ tiếng Hán sang tiếng Việt) thì vẫn phải chuyển từ vựng tiếng Hán (mặc dù có tương đương trong tiếng Việt) sang tập các khái niệm rồi sau đó mới chọn nghĩa tiếng Việt.

tri thức ngôn ngữ thế nào quyết định chết lượng của mọi hệ xử lý ngôn ngữ tự nhiên.

III.4.1. HAI KHÍA CẠNH CỦA NGÔN NGỮ

Để có thể xây dựng mô hình liên ngữ thỏa mãn các tính chất mong muốn đã nêu trên, ta cần phân tích kỹ hơn một số đặc trưng chung của các ngôn ngữ tự nhiên. Câu hỏi đầu tiên đặt ra là :*Liệu có tồn tại văn phạm phổ quát (một thứ văn phạm có thể sử dụng chung cho mọi ngôn ngữ) hay không?* Đây là một vấn đề tranh cãi trong giới ngôn ngữ học.

Phương pháp dịch máy được trình bày ở đây giả định rằng một văn phạm như vậy là *tồn tại!* Chomsky đã tin tưởng sâu sắc vào điều đó và hơn thế, còn cho rằng cấu trúc của văn phạm phổ quát theo một cách nào đó được tích hợp sẵn trong não của người [1]. Cũng có người, trái lại, hoàn toàn phủ định giả thuyết này. Chẳng hạn, trong bài "Một số biểu hiện của cách nhìn Âu châu đối với cấu trúc tiếng Việt" [40], có câu : "*Vả lại đến những năm 90 của thế kỷ không còn có ai mơ hồ đến mức tưởng rằng có những phạm trù ngữ pháp phổ quát cho ngôn ngữ toàn nhân loại*"?. Tuy nhiên, không thấy có những luận cứ thuyết phục nào làm chỗ dựa cho lập luận cực đoan này.



Hình 3: Cấu trúc của phát biểu

Quay trở lại ví dụ 1, ta thấy rằng đây là hai câu hoàn toàn đồng nghĩa với nhau. Sự đồng nghĩa nằm ở chỗ cùng một hành động được thực hiện (*con chó đuổi con mèo*), tại cùng một thời điểm (*hôm qua*). Điểm khác biệt duy nhất là đối tượng nào được chọn đưa lên đầu câu (với mục đích hướng sự chú ý của người nghe hay người đọc) thì khác nhau.

Việc thay đổi trật tự các từ trong câu ở đây tạo nên những *sắc thái*, *khung cảnh* khác nhau cho nội dung thông điệp (gần như) trùng nhau. Như vậy ta thấy trong một phát biểu luôn luôn bao gồm hai khía cạnh:

- Nội dung chính của phát biểu
- Khung cảnh trong đó nội dung được trình bày

Phần nội dung chính của cả hai câu trong ví dụ 1 có thể biểu diễn bằng cùng một sơ đồ dưới đây (Hình 3). Mũi tên trong hình vẽ (chỉ hướng

của hành động từ tác nhân đến đối tượng hoặc nhận định) được sử dụng hoàn toàn vì mục đích trực quan (không phải thể hiện mối quan hệ phụ thuộc của biểu diễn bên trong).

Phần khung cảnh ở đây bao gồm những thông tin sau:

- Sự kiện đã xẩy ra trong quá khứ (cụ thể là *hôm qua*)
- Đối tượng được hướng sự chú ý tới (ở câu thứ nhất là *con chó*, còn ở câu thứ hai là *con mèo*)

Trong phần nội dung chính ta thấy cấu trúc của phát biểu (hành động, hoặc nhận định). Thông tin này đặc trưng cho phát biểu hiện thời, nó ít hoặc không có vai trò gì đối với các phát biểu trước và sau nó. Đây là thành phần không phụ thuộc ngôn ngữ.

Ngược lại, trong phần khung cảnh là những thông tin chỉ ra hoàn cảnh mà hành động hoặc nhận định được thực hiện. Thông tin về phần khung cảnh có thể có vai trò kiểm tra tính hợp lệ của văn bản theo đòi hỏi về hành văn của một ngôn ngữ cụ thể. Đây là thành phần phụ thuộc ngôn ngữ. Chẳng hạn, trong tiếng Anh, thông tin về thì của câu trước thường chỉ dẫn cho việc tổng hợp câu sau đó.

Như vậy ngôn ngữ có hai khía cạnh tách biệt nhau : Phần nội dung với cấu trúc câu của phát biểu (*độc lập ngôn ngữ*) và phần khung cảnh với các giá trị trạng thái (*phụ thuộc ngôn ngữ*). Khía cạnh nội dung của phát biểu được gọi là *cây phân cấp ngữ nghĩa* (*hierarchical semantic tree*), còn khía cạnh về khung cảnh là được gọi là *không gian trạng thái* (*state space*).

Cây phân cấp ngữ nghĩa của câu xác định những tương quan cú pháp của các thành phần trong câu, trong khi đó không gian trạng thái xác định các hiệu ứng mặc định (về thời gian, không gian, lĩnh vực...). Những thông tin của không gian trạng thái có thể khá bền vững (trong một đoạn văn hoặc trong cả bài), vì vậy có thể kế thừa từ câu văn này sang câu văn khác; trong khi đó, cây phân cấp ngữ nghĩa chỉ tồn tại đối với từng mệnh đề riêng rẽ. Để minh họa ta xét ví dụ:

Hôm qua con chó đuổi con mèo. Nó đớp một cái trúng đuôi.

Trong câu thứ hai ta biết được đại từ *nó* chỉ *con chó* chứ không phải là *con mèo*. Tiếng Việt không có thì, nhưng trong ví dụ trên, ta biết rằng khi dịch sang tiếng Anh (hay tiếng Nga) động từ *đớp* phải ở *thì quá khứ* vì từ câu trước, có thể xác định được rằng sự kiện xảy ra *hôm qua*.

Để so sánh, ta xem ví dụ :

Hôm qua con mèo bị chó đuổi. Giờ đuôi nó giống cái chìa khóa.

Trong đoạn văn này khi muốn dịch sang tiếng Anh ta cần phải lưu tâm rằng đại từ *nó* trong câu thứ hai chỉ có thể trỏ tới *con mèo*, và động từ *giống* phải ở *thì hiện tại* và *ngôi thứ ba, số ít* vì thời gian đã thay đổi , từ *hôm qua* trong câu trước, sang câu sau đã là (*bây*) *giờ*.

Để máy tính có thể xử lý được những tình huống tương tự như trên ta có thể lưu giữ thông tin về đối tượng được hướng tới cũng như thông tin về thời điểm xảy ra sự kiện trong các biến của không gian trạng thái. Mỗi hình trạng của không gian trạng thái sẽ xác định một tổ hợp các tình huống văn cảnh nhất định; từ đó ta có được những chỉ dẫn ngữ pháp *(những chỉ dẫn về các thoả thuận giữa các bộ phận trong câu của ngôn ngữ đích*) để tổng hợp văn bản đúng văn phạm (trong trường hợp này là đúng *thì, số,* hay *ngôi* của tiếng Anh).

Mô hình văn phạm cảm ngữ đoạn - PSG (*Phrase-sensitive grammar*) cho phép mô tả hai khía cạnh của ngôn ngữ và có thể được sử dụng để tách chúng trong quá trình phân tích văn bản, (xem [25]). Trong văn phạm PSG ta có thể mô tả hình thức :

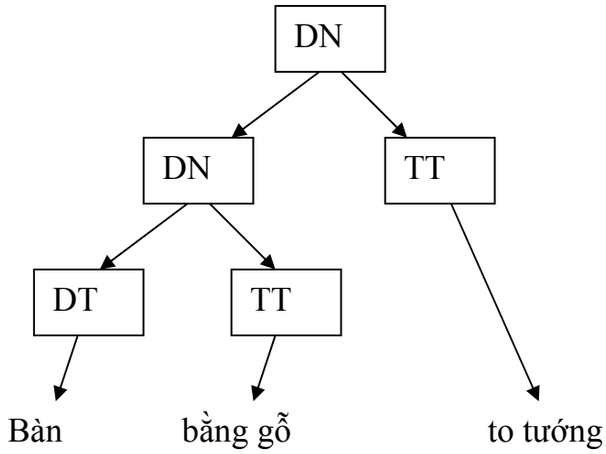
- Thỏa thuận (*agreements*) về văn phạm
- Ràng buộc ngữ cảnh tổng quát (*generic context sensitivity*)
- Tính không liên tục ngữ cảnh (*contextual discontinuity*)
- Sự phân cấp các khái niệm ngữ pháp
- Kế thừa và lan truyền thuộc tính
- Cơ chế giải quyết nhập nhằng cú pháp và từ vựng

Văn phạm cảm ngữ đoạn là một mở rộng của mô hình văn phạm sinh Chomsky. Bằng cách đưa vào những quy tắc đánh dấu, có thể chỉ rõ những mối quan hệ chính, phụ cũng như tính kế thừa giữa các khái niệm ngữ pháp. Mô hình văn phạm cũng cho ta một phương pháp biểu diễn tri thức ngôn ngữ có tính nghịch đảo: ta chỉ cần một bộ văn phạm cho mỗi ngôn ngữ, nó sẽ được dùng để phân tích cũng như tổng hợp văn bản.

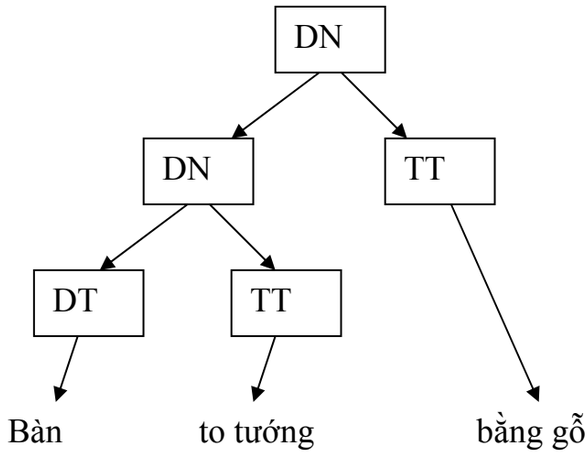
III.4.2. CÂY PHÂN CẤP NGỮ NGHĨA

Cấu trúc trọng tâm trong phân tích cú pháp là *cây cú pháp*. Cây cú pháp thể hiện mối liên hệ phân cấp giữa các bộ phận trong câu. Qua cây cú pháp có thể thấy được các quy tắc văn phạm được áp dụng như thế nào cũng như lịch sử việc áp dụng các quy tắc. Từ một câu có thể dựng được nhiều cây cú pháp (*tính nhập nhằng*), nhưng từ một cây cú pháp chỉ có thể tổng hợp được một câu văn *duy nhất*. Việc phản ánh ***đúng và đầy đủ*** trình tự áp dụng các quy tắc văn phạm lại chính là thiếu sót của kiến trúc này : những câu đồng nghĩa (nhưng khác nhau về sắp xếp từ ngữ) luôn luôn có biểu diễn

cây cú pháp khác nhau. Trên Hình 4 và Hình 5 vẽ cây cú pháp cho các cụm từ "*Bàn bằng gỗ to tướng*" và "*Bàn to tướng bằng gỗ*".

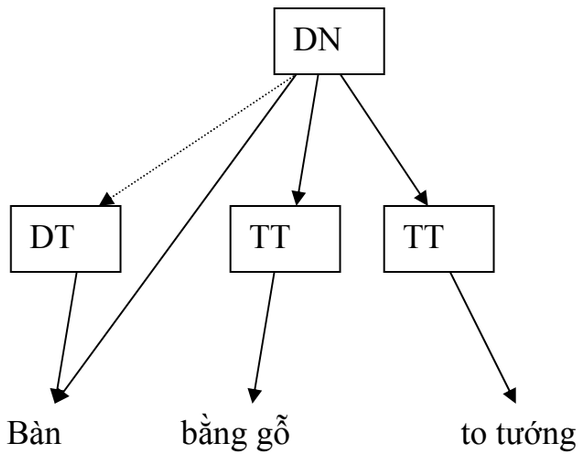


Hình 4. Cây cú pháp

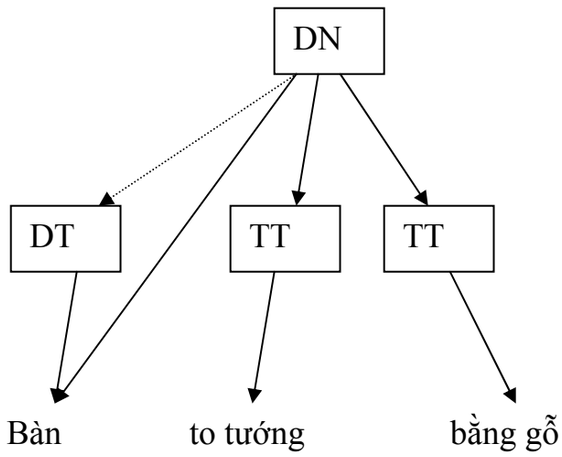


Hình 5. Cây cú pháp

Trong Hình 4 ta không nhận thấy sự liên hệ giữa "*bàn*" và "*to tướng*", còn trong Hình 5 ta không nhận thấy sự liên hệ giữa "*bàn*" và "*bằng gỗ*" mặc dù trong cả hai câu thì "*bằng gỗ"* và *"to tướng*" đều ***bổ nghĩa trực tiếp*** cho "*bàn*". Thêm vào đó, trật tự các nhánh trong cây cú pháp cũng cố định. Điều này cản trở rất nhiều việc phân tích câu, đặc biệt là khi cần phải xét những yếu tố phụ thuộc ngữ cảnh; mà trong ngôn ngữ tự nhiên thì yếu tố ngữ cảnh là không thể thiếu.



Hình 6. Cây phân cấp ngữ nghĩa

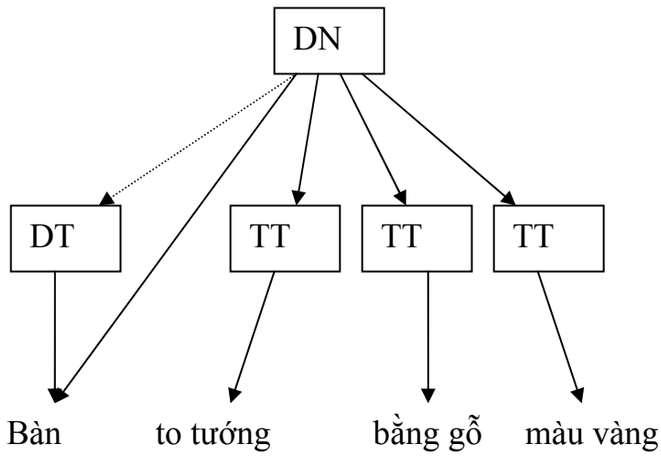


Hình 7. Cây phân cấp ngữ nghĩa

Cây phân cấp ngữ nghĩa (*hierarchical semantic tree*) cũng được hình thành trong quá trình phân tích câu. Tuy nhiên, nó không phản ánh lịch sử áp dụng các quy tắc sinh : Qua cây phân cấp ngữ nghĩa ta không thể khôi phục chính xác trình tự áp dụng các quy tắc văn phạm. Nghĩa là từ cây phân cấp ngữ nghĩa, trong nhiều tình huống, ta không thể tái hiện được câu văn ban đầu. Một điểm đặc biệt là trật tự các nhánh trong cây phân cấp ngữ nghĩa là không có ý nghĩa. Hình 6 và Hình 7 vẽ các cây phân cấp ngữ nghĩa cho các ví dụ trên.

Nếu xét đến quy định rằng trật tự các nhánh trong cây phân cấp ngữ nghĩa không có giá trị thì các cây trên Hình 6 và Hình 7 là hoàn toàn đồng nhất với nhau.

Điều này có nghĩa rằng hai mệnh đề "*Bàn bằng gỗ to tướng*" và "*Bàn to tướng bằng gỗ*" đều có chung một cây phân cấp ngữ nghĩa. Như vậy, ta có thể chỉ lưu giữ phần tương quan nội dung của câu văn mà loại bỏ đi phần hình thức (do tính chất tuyến tính của ngôn ngữ tự nhiên đòi hỏi).



Hình 8. Cây phân cấp ngữ nghĩa

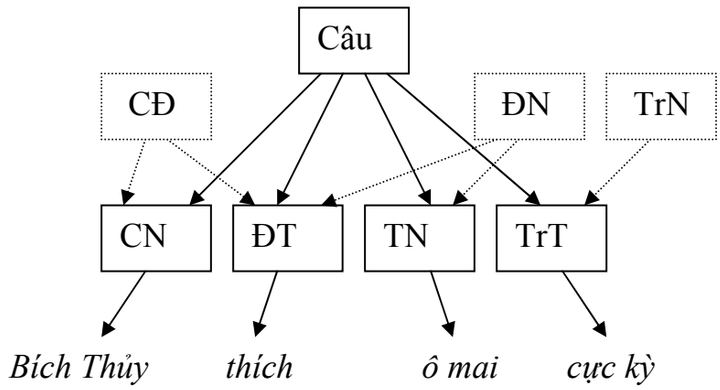
Trên Hình 8 là cây phân cấp ngữ nghĩa, chung cho tất cả các cụm từ như "*Bàn bằng gỗ màu vàng to tướng*", "*Bàn màu vàng to tướng bằng gỗ*", "*Bàn to tướng bằng gỗ màu vàng*",...Các cụm từ này đều đồng nghĩa với nhau và được đưa về cùng một cấu trúc.

Ta có thể nhận thấy trên Hình 8, nút DN (Danh ngữ) có 4 nhánh, còn trên Hình 7 chỉ có 3 nhánh. Trong cây phân cấp ngữ nghĩa, các nút có thể có *số nhánh biến thiên* mặc dù các quy tắc sinh có độ dài vế phải là cố định. Để có được tính chất này, cây phân cấp ngữ nghĩa, trong khi là cấu trúc cú pháp; được tạo thành khi áp dụng các quy tắc nhưng không phản ánh cụ thể lịch sử áp dụng chúng. Vì vậy, khi tổng hợp văn bản, chẳng hạn, từ cây phân cấp ngữ nghĩa trên Hình 8 có thể sinh ra bất kỳ mệnh đề đồng nghĩa cụ thể nào trong danh sách đã nói (mà không thể sinh ra đúng mệnh đề ban đầu).

Cú pháp – theo định nghĩa của từ điển Webster – là *bộ phận của ngữ pháp giải quyết cách thức mà các ký hiệu ngôn ngữ (như các **từ**) được sắp xếp với nhau để hình thành những tổ hợp (như các **ngữ đoạn** hoặc mệnh đề)*. Theo định nghĩa này thì cây phân cấp ngữ nghĩa cũng chính là cây cú pháp. Không có lý do gì để phát biểu rằng các cấu trúc trên Hình 4 và Hình

5 là chuẩn hơn các cấu trúc trên Hình 6 và Hình 7 tương ứng (*thậm chí có lý do để phát biểu ngược lại*).

Trên Hình 8, ta thấy nút DN (danh ngữ) trỏ thẳng tới từ "*bàn*". Có thể làm như vậy khi ta sử dụng loại quy tắc với các *ký hiệu được đánh dấu.* Cũng chính loại quy tắc này cho phép ta xây dựng cây cú pháp có những nút với số nhánh biến thiên.



Hình 9. Cây phân cấp ngữ nghĩa

Như vậy, cây phân cấp ngữ nghĩa có một tính chất quan trọng mà cây cú pháp theo quan niệm thông thường không có là tồn tại nhiều mệnh đề khác nhau có cùng chung một cây phân cấp ngữ nghĩa. Những mệnh đề đó được coi là đồng nghĩa với nhau. Điều đó có nghĩa rằng, bằng cách thiết kế văn phạm thích hợp, ta có thể định nghĩa ngôn ngữ sao cho những cấu trúc ngôn ngữ đồng nghĩa với nhau được đưa về cùng một dạng biểu diễn của cây phân cấp ngữ nghĩa; và công việc này có thể thực hiện một cách toàn cục – đối với tất cả các ngôn ngữ. Cấu trúc phân cấp ngữ nghĩa phản ánh mối liên hệ nội dung của các thành phần ngôn ngữ, không phải mối liên hệ về *luật hành văn tuyến tính* của từng ngôn ngữ cụ thể.

Tập hợp tất cả các cấu trúc phân cấp ngữ nghĩa (mà ta có thể xây dựng khi khảo sát các ngôn ngữ trong một hệ thống dịch máy đa ngữ cụ thể) tạo thành một bộ khung cho kiến trúc liên ngữ. Đó chính là nền tảng của giải pháp dịch máy liên ngữ đang được phát triển tại NACENTECH.

Sử dụng văn phạm cảm ngữ đoạn, cây phân cấp ngữ nghĩa của mệnh đề ***"Bích Thủy thích ô mai cực kỳ"*** *(xem ví dụ (1) trong phần I)* có thể có dạng như ở Hình 9.

Ở đây, Chủ ngữ (CN), Động từ hay Tân ngữ (TN) đều có thể có mối nối trực tiếp đến khái niệm mức cao hơn (Câu), và vì vậy, có thể đưa vào những quy tắc ràng buộc, chẳng hạn giữa danh từ chính của *Chủ ngữ* và

Động từ cũng như giữa *Động từ* và danh từ chính của *Tân ngữ* (So sánh Hình 1, Phần I).

Điều đáng lưu ý là ở đây, việc dựng cây phân cấp ngữ nghĩa được thực hiện trực tiếp một cách hình thức từ văn bản nguồn, sử dụng văn phạm cảm ngữ đoạn chứ không phải có thể nhận được bằng cách biến đổi cây cú pháp

III.4.3. PHÂN HOẠCH LIÊN NGỮ

Trong Ethnologue language family index [49] đưa ra một sự phân loại chi tiết các ngôn ngữ trên thế giới, tổng cộng 6,809 thứ tiếng, bao gồm 108 họ ngôn ngữ. Theo tài liệu này thì ở Việt nam có gần 70 triệu người nói các ngôn ngữ thuộc họ Austro – Asiatic; hơn 2 triệu người nói các ngôn ngữ họ Daic; ngoài ra còn có các ngôn ngữ thuộc họ Miao-Yao, Austronesian và Tibeto-Burman. Tiếng Việt được xếp trong nhóm Việt-Mường, dòng Môn-Khơ me, họ Austro – Asiatic theo cây phân cấp (*trích đoạn*) như trên Hình 10. Tiếng Anh được xếp trong nhóm West dòng Germanic họ Indo-European theo trích đoạn cây phân cấp như trên Hình 11.

Theo sơ đồ này thì tiếng Việt và tiếng Mường rất gần nhau vì cùng nhóm Việt Mường; còn tiếng Việt và tiếng Khơ me tuy không thuộc cùng nhóm nhưng thuộc cùng dòng Môn-Khơ me nên cũng khá gần nhau. Tương tự, tiếng Anh và tiếng Scots rất gần nhau; còn tiếng Anh và tiếng Đức thì cùng thuộc một nhóm West. Trong khi đó tiếng Anh và tiếng Thụy điển xa nhau hơn : cùng thuộc dòng Germanic.

Austro-Asiatic (168)
- Mon-Khmer (147)
  - Eastern Mon-Khmer (67)
    - Bahnaric (40)
      - Central Bahnaric (6)
      - North Bahnaric (14)
      - South Bahnaric (9)
      - West Bahnaric (11)
    - Katuic (19)
      - Central Katuic (5)
      - East Katuic (8)
      - West Katuic (6)
    - Khmer (2)
      - KHMER, CENTRAL [KMR] (Cambodia)
      - KHMER, NORTHERN [KXM] (Thailand)
      - ....
  - Viet-Muong (10)
    - Chut (3)
      - AREM [AEM] (Viet Nam)

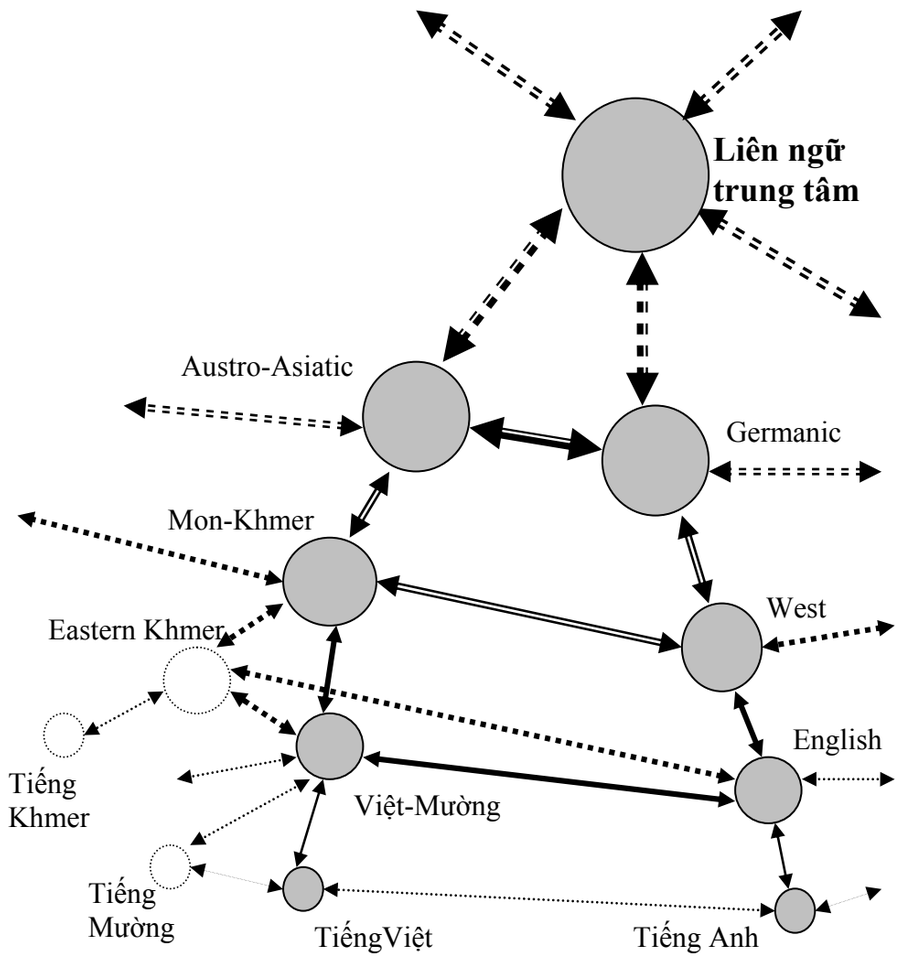
MALENG [PKT] (Laos)
CHUT [SCB] (Viet Nam)
Cuoi (2)
HUNG [HNU] (Laos)
THO [TOU] (Viet Nam)
Muong (3)
BO [BGL] (Laos)
MUONG [MTQ] (Viet Nam)
NGUÔN [NUO] (Viet Nam)
Thavung (1)
AHEU [THM] (Thailand)
Vietnamese (1)
VIETNAMESE [VIE] (Viet Nam)

Hình 10. Cây phả hệ ngôn ngữ họ Austro-Asiatic

Indo-European (443)
Germanic (58)
East (1)
GOTHIC [GOF] (Ukraine)
North (14)
East Scandinavian (8)
Danish-Swedish (8)
West Scandinavian (6)
FAROESE [FAE] (Denmark)
ICELANDIC [ICE] (Iceland)
JAMSKA [JMK] (Sweden)
NORN [NON] (United Kingdom)
NORWEGIAN, NYNORSK [NRN] (Norway)
TRAVELLER NORWEGIAN [RMG] (Norway)
West (43)
English (5)
CAYMAN ISLANDS ENGLISH [CYE] (Cayman Islands)
ENGLISH [ENG] (United Kingdom)
ANGLOROMANI [RME] (United Kingdom)
SCOTS [SCO] (United Kingdom)
YINGLISH [YIB] (USA)
Frisian (3)
FRISIAN, WESTERN [FRI] (Netherlands)
FRISIAN, NORTHERN [FRR] (Germany)
FRISIAN, EASTERN [FRS] (Germany)
High German (19)
German (17)
Yiddish (2)
Low Saxon-Low Franconian (16)
Low Franconian (3)
Low Saxon (13)
.....

Hình 11. Cây phả hệ ngôn ngữ họ Indo-European

Giữa tiếng Anh và tiếng Việt không có gì chung trong phả hệ ngôn ngữ. Tuy nhiên, như đã đề cập trong phần II, có một nền tảng chung cho tư duy ngôn ngữ của loài người. Phần chung đó chính là ***Liên ngữ***.



Hình 12. Phân hoạch ngôn ngữ

Theo cách diễn đạt này thì thực tế ta có một lớp các *Liên ngữ* khác nhau được trừu xuất theo các cấp độ khác nhau tùy thuộc vào các nhóm, dòng, họ ngôn ngữ khác nhau. Điều này một phần lý giải vì sao không thể xây dựng một *Liên ngữ* đầy đủ, hoàn toàn độc lập ngôn ngữ được.

Dựa theo sự phân cấp ngôn ngữ, ta có thể thấy được mức độ gần gũi hay khác nhau của các ngôn ngữ cùng nhóm, cùng dòng và cùng họ, cũng như sự khác biệt lớn giữa những ngôn ngữ khác họ. Sự gần gũi hay khác

biệt thể hiện ở hệ thống từ vựng, văn phạm và ngữ dụng của các cộng đồng người sử dụng ngôn ngữ. Hình 12 biểu diễn một vài tương quan đó.

Các vòng tròn biểu thị phần tri thức ngôn ngữ chung giữa các ngôn ngữ thuộc cấp thấp hơn trong cây phân cấp. Liên ngữ là vòng tròn lớn nằm ở trung tâm. Liên ngữ chứa tất cả những tri thức chung cho mọi ngôn ngữ trên thế giới. Vòng tròn có nhãn Austro-Asiatic chứa tất cả những tri thức ngôn ngữ chung cho họ các ngôn ngữ Austro-Asiatic (*bên ngoài liên ngữ*). Tương tự vòng tròn có nhãn Germanic chứa tất cả những tri thức ngôn ngữ chung cho họ các ngôn ngữ Germanic không thuộc liên ngữ. Cứ như thế, vòng tròn có nhãn Tiếng Việt chứa phần tri thức riêng cho tiếng Việt mà không thuộc nhóm Việt Mường, Mon-Khmer, Austro-Asiatic và Liên ngữ; tri thức đó bao gồm những cấu trúc, thành ngữ, từ vựng,... của tiếng Việt cần phải lưu tâm khi dịch sang các ngôn ngữ (cụ thể, không phải nhóm, dòng ngôn ngữ) khác; nghĩa là những nhóm đơn vị tri thức rời rạc ứng với từng ngôn ngữ cụ thể (tiếng Mường, tiếng Khmer, tiếng Anh,...).

Tổ chức theo sơ đồ này cho phép ta phân lập được tri thức ngôn ngữ một cách có hệ thống. Ta luôn luôn có vừa đủ tri thức ngôn ngữ (của một cặp ngôn ngữ cụ thể) để xử lý :

Chẳng hạn, khi muốn dịch từ tiếng Anh sang tiếng Việt, ta cần đến những tri thức sau thuộc những khối có mũi tên có thể trỏ tới tiếng Việt, gồm:

- Tri thức Liên ngữ, Germanic, West, English, tiếng Anh (phân tích)
- Tri thức Austro-Asiatic, Môn-Khơ me, Việt Mường, Tiếng Việt;
- Tri thức về họ Germanic khác biệt với họ Austro-Asiatic.
- Tri thức về dòng West khác biệt với dòng Môn-Khơ me
- Tri thức về nhóm English khác biệt với nhóm Việt Mường
- Tri thức về tiếng Anh khác biệt với tiếng Việt

Nếu như ta muốn dịch từ tiếng Mường sang tiếng Việt thì ta chỉ cần tri thức của :

- Tri thức Liên ngữ, Austro-Asiatic, Môn-Khơ me, Việt Mường, Tiếng Mường (phân tích)
- Tri thức Tiếng Việt;
- Tri thức về Tiếng Mường khác biệt với Tiếng Việt.

Phương pháp dịch máy Chuyển đổi giả thiết không có gì chung giữa các ngôn ngữ; còn phương pháp dịch máy Liên ngữ thì giả thiết là mọi tri thức ngôn ngữ đều chung cho các ngôn ngữ nên đều có thể chuyển về một

Liên ngữ duy nhất. Xét về mặt tổ chức tri thức ngôn ngữ, từ việc tồn tại mối liên hệ giữa các ngôn ngữ với nhau theo một trật tự phân cấp, *cả hai phương pháp đã nói đều là dạng suy biến của mô hình nhiều tầng này*.

Tất nhiên, thực tế ngôn ngữ nói chung sẽ không tuân thủ trật tự phân cấp theo phân loại [49] vì các ngôn ngữ, bất kể nguồn gốc của mình, trong quá trình phát triển lịch sử hàng nghìn năm, trong sự giao lưu giữa các cộng đồng dân cư mà tiếp thu lẫn nhau các phương tiện biểu diễn tri thức ngôn ngữ. Vì vậy việc tổ chức hệ phân cấp liên ngữ có thể khác với cách phân loại của [49]. Sự hình thành hệ phân cấp có thể thực hiện một cách tự động thông qua sự tổ chức theo dàn phân cấp của hệ quy tắc và bảng đối chiếu đa ngữ (*bảng đồng nghĩa*)

Việc tổ chức Thông tin đa ngôn ngữ cần được thực hiện sao cho sự tra cứu, cập nhật số liệu bằng tay và xử lý tự động dễ dàng và thuận tiện.

Dữ liệu đa ngôn ngữ được lưu giữ trong cùng một tập tin. Một bộ chỉ mục chung được thành lập cho tất cả các mục từ của mọi ngôn ngữ trong hệ. Để có thể sử dụng từ điển đồng thời với hai mục đích, vừa để cho người dùng tra cứu và cập nhật (có khả năng thay thế cho từ điển thông thường), lại vừa để cho chương trình máy tính sử dụng để phân tích và tổng hợp ngôn ngữ (trong đó có dịch tự động), dữ liệu từ điển (bao gồm *từ vựng, ngữ pháp, thành ngữ* và *ngữ dụng*) được tổ chức thành các thành phần độc lập sau:

- *Từ điển ngôn ngữ* – gồm những mục từ, cụm từ hoặc cấu trúc văn phạm (tuân thủ văn phạm cảm ngữ đoạn) được đánh số. Từ điển được tổ chức thành một dàn đại số (bằng cách bổ sung thêm các phần tử $\top$ và $\bot$)
- *Bảng quy tắc văn phạm* – được sử dụng để định nghĩa các mục từ thông qua các *khái niệm*, và định nghĩa các mẫu đoạn thông qua các *khái niệm.* Dàn các quy tắc văn phạm được xây dựng dựa trên dàn xác định trong văn phạm như sau:

  $A1 \rightarrow B1 < A2 \rightarrow B2$ *khi và chỉ khi*

  $A1 < A2$ *và* $B1 < B2$
- *Bảng đồng nghĩa* – gồm những bộ số nguyên (dãy số hiệu của các mục trong *từ điển ngôn ngữ*). Mỗi phần tử của bảng đồng nghĩa là danh mục những từ, cụm từ hoặc cấu trúc văn phạm có nghĩa tương đương với nhau (Mỗi từ hoặc cụm từ là định nghĩa của các từ hoặc cụm từ khác của danh mục trong ngôn ngữ khác hay trong cùng ngôn ngữ).
- *Bảng cận nghĩa* – gồm những bộ số nguyên (dãy số hiệu của các mục trong *từ điển ngôn ngữ*). Mỗi phần tử của bảng cận nghĩa là

danh mục những từ, cụm từ hoặc cấu trúc văn phạm có nghĩa gần với nghĩa của từ đã cho.

- *Bảng trái nghĩa* – gồm những bộ số nguyên (dãy số hiệu của các mục trong *từ điển ngôn ngữ*). Mỗi phần tử của bảng cận nghĩa là danh mục những từ, cụm từ hoặc cấu trúc văn phạm có nghĩa trái ngược với nghĩa của từ đã cho.

Khi con người cần tra cứu thông tin ngôn ngữ (từ, cụm từ hoặc cấu trúc văn phạm được mô tả theo văn phạm cảm ngữ đoạn, thuộc ngôn ngữ bất kỳ), ứng dụng phần mềm sẽ lọc các mục thông tin liên quan đến mục từ hoặc cụm từ trên *từ điển ngôn ngữ*, sau đó, dựa theo các số hiệu thu được, thu thập các phần tử thông tin từ *bảng đồng nghĩa* (hoặc *bảng trái nghĩa*) và *bảng cận nghĩa* để lọc ra (theo ngôn ngữ được lựa chọn để hiển thị [1]) những số hiệu của các mục từ điển liên quan và hiển thị kết quả theo định dạng thích hợp. Dữ liệu được tổ chức để có thể chứa đầy đủ thông tin như từ điển điện tử thông thường.

Để phân tích văn bản, ứng dụng phần mềm sử dụng các mục thông tin liên quan đến mỗi mục từ hoặc cụm từ trong văn bản trên *từ điển ngôn ngữ*, từ đó xác định được các yếu tố từ vựng, ngữ pháp và ngữ nghĩa của mục từ. Những thông tin này giúp cho bộ phân tích có thể dựng được các cách phân tích văn bản (bằng cách áp dụng các quy tắc có trong *từ điển ngôn ngữ và Bảng quy tắc văn phạm*) và xây dựng các ***cấu trúc phân tích*** của văn bản đó.

Để tổng hợp văn bản, từ *cấu trúc phân tích* của một văn bản, ứng dụng phần mềm dựa theo các số hiệu thu được, thu thập các phần tử thông tin từ *bảng đồng nghĩa* (hoặc *bảng trái nghĩa*) và *bảng cận nghĩa* để lọc ra (theo ngôn ngữ được lựa chọn, bằng cách thay thế những số hiệu mục từ điển bằng những số hiệu trong *bảng đồng nghĩa* hoặc *bảng cận nghĩa* của các mục ứng với số hiệu của mục đó) những số hiệu của các mục từ điển liên quan, sử dụng *bảng quy tắc văn phạm* để thực hiện việc tổng hợp văn bản.

Để dịch tự động một văn bản, ứng dụng phần mềm thực hiện lần lượt các bước sau:

- Phân tích văn bản (dựa vào *từ điển ngôn ngữ* và *bảng quy tắc văn phạm*) để dựng các cấu trúc phân tích của văn bản đó – các cây phân cấp ngữ nghĩa.
- Dựa vào *bảng đồng nghĩa* (hoặc *bảng trái nghĩa*) và *bảng cận nghĩa* để lọc ra (theo ngôn ngữ được lựa chọn để hiển thị) những

[1] là ngôn ngữ của người dùng, chẳng hạn, tiếng Việt cho người Việt nam

số hiệu của các mục từ điển liên quan để thực hiện việc xác định tương ứng khái niệm giữa ngôn ngữ nguồn và ngôn ngữ đích.

- Tổng hợp văn bản (dựa vào *từ điển ngôn ngữ* và *bảng quy tắc văn phạm*) để dựng các cấu trúc tổng hợp văn bản từ cấu trúc phân tích đã cho.

Với việc tổ chức dữ liệu đa ngữ trong cùng một kho ngữ liệu, và dữ liệu ngôn ngữ bao gồm các *quy tắc văn phạm* cũng như *bảng đồng nghĩa* (hoặc *bảng trái nghĩa*) và *bảng cận nghĩa*, ứng dụng phần mềm phân tích và dịch tự động có thể thực hiện công việc mà *không cần biết văn bản gốc thuộc ngôn ngữ nào*.

III.5. PHƯƠNG PHÁP DỊCH MÁY

Thông thường, trong các hệ dịch máy theo phương pháp chuyển đổi chấp nhận sơ đồ dịch gồm các bước sau:

- Ngắt câu để từ đoạn văn chọn ra một câu.
- Phân tích từ vựng : xử lý tiếp đầu, tiếp đuôi, ghép từ (đối với những ngôn ngữ biến hình thì phần ghép từ là suy biến, còn đối với những ngôn ngữ đơn lập thì phần xử lý tiếp đầu, tiếp đuôi là suy biến)
- Phân tích văn phạm : xây dựng tập các cây cú pháp của câu nguồn
- Xử lý nhập nhằng : chọn ra cây cú pháp thích hợp nhất theo một tiêu chí nào đó.
- Chuyển đổi cây cú pháp : Thông thường là ứng với mỗi luật sinh của ngôn ngữ nguồn có kèm theo một quy tắc dịch (*chọn luật tương ứng trong ngôn ngữ đích để xây dựng cây cú pháp của ngôn ngữ đích từ cây cú pháp của ngôn ngữ nguồn*).
- Tổng hợp từ vựng và phát sinh bản dịch.

Ta có thể nhận thấy một vài đặc điểm của sơ đồ trên :

- Sự phụ thuộc nặng nề của quá trình dịch đối với ngôn ngữ nguồn. Cây cú pháp của ngôn ngữ nguồn quyết định cách thức biên dịch văn bản sang ngôn ngữ đích. Điều này dẫn đến sự suy biến của bước tổng hợp : ta không thấy có khối tổng hợp cú pháp của ngôn ngữ đích. *Công đoạn phức tạp nhất chính là phân tích cú pháp*. Kết quả là phải cần rất nhiều quy tắc dịch (cho những tình huống khác biệt giữa hai ngôn ngữ) kéo theo rất nhiều quy tắc phân tích

văn phạm (có dạng tương tự nhau trên ngôn ngữ nguồn nhưng khác nhau về luật dịch sang ngôn ngữ đích)[1]

- Dữ liệu chỉ sử dụng được cho dịch một chiều và cho một cặp ngôn ngữ. Để dịch ngược lại ta phải xây dựng lại toàn bộ hệ quy tắc và từ vựng.

Con người dịch ngôn ngữ theo một cách hoàn toàn khác. Việc đọc hiểu *đúng* câu văn (*phân tích*) không chiếm nhiều thời gian và công sức. Khó khăn chính mà người dịch thường gặp là khi chuyển ngữ : tổng hợp câu văn của ngôn ngữ đích. Chất lượng bản dịch phụ thuộc chủ yếu vào công việc tổng hợp này.

III.5.1. SƠ ĐỒ DỊCH MÁY

Mô hình trung gian độc lập ngôn ngữ của sơ đồ dịch máy liên ngữ được đề xuất là cấu trúc dạng cây gọi là *Cây phân cấp ngữ nghĩa* (*Hierarchical Semantic Tree*) với mỗi nút có gán *thông tin trạng thái* như đã được trình bày ở trên. Ứng với mỗi câu của văn bản, cây phân cấp ngữ nghĩa được xây dựng để biểu diễn sự ràng buộc chức năng logic (chứ không phải ràng buộc cú pháp) giữa các thành phần trong câu; còn thông tin trạng thái biểu diễn mối liên hệ giữa các thành phần bên trong câu cũng như mối liên hệ giữa các câu trong bài văn.Sơ đồ dịch liên ngữ bao gồm hai khối cơ bản : Phân tích (*Analysis*) và Tổng hợp (*Generation*) (Xem Hình 9).

Bộ phân tích đọc câu nguồn, thực hiện các bước phân tích từ vựng và cú pháp để:

- Tạo ra cây phân cấp ngữ nghĩa của câu.
- Cập nhật thông tin trạng thái của văn bản.

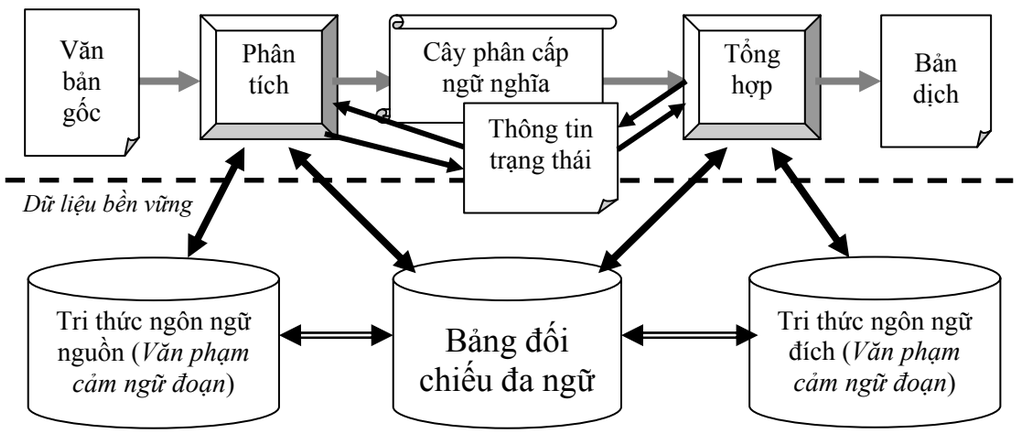
Trong khi phân tích, hệ thống có thể đối chiếu các giá trị trạng thái để kiểm tra tính hợp lệ của câu nguồn cũng như cập nhật các giá trị mới. Thông tin về sự phân cấp ngữ nghĩa và thông tin về bối cảnh (từ cơ sở tri thức đơn ngữ) được mô tả bằng các quy tắc văn phạm cảm ngữ đoạn. Cây phân cấp ngữ nghĩa của câu chỉ tồn tại trong quá trình phân tích cho từng câu nguồn, trong khi đó các giá trị trạng thái được lưu giữ sau khi đã xử lý xong mỗi câu.

Bộ tổng hợp sử dụng thông tin từ *bảng đối chiếu* (cơ sở tri thức dịch đa ngữ, bao gồm *đồng nghĩa* và *cận nghĩa*) ứng với mỗi *mẫu đoạn* phân tích của ngôn ngữ nguồn có thể tìm thấy một tập các *mẫu đoạn* tương ứng của

[1] Con người thực hiện việc dịch theo cách hoàn toàn khác : Chỗ khó nhất chính là tổng hợp cú pháp cho ngôn ngữ đích chứ không phải là phân tích câu nguồn

ngôn ngữ đích. Các *mẫu đoạn* là chuỗi thông tin tri thức ngôn ngữ cảm ngữ đoạn.

Khác với các hệ dịch máy thông dụng, trong đó một quy tắc văn phạm của ngôn ngữ này tương đương với một quy tắc văn phạm của ngôn ngữ khác[1]; trong cách tiếp cận này sự tương đương chỉ được định nghĩa giữa các *mẫu đoạn.* Đặc tính này cho phép hệ dịch máy có sự biểu diễn độc lập ngôn ngữ trên cơ sở *liên kết muộn*: không cần phải tổ chức các quy tắc biên dịch thành từng ***cặp*** (song ngữ) hay từng bộ (đa ngữ).



Hình 1. Sơ đồ dịch máy liên ngữ của NACENTECH.

Quá trình tổng hợp vì vậy phức tạp hơn nhiều so với phân tích vì từ mỗi nút trong cây phân cấp ngữ nghĩa ta phải chọn và thay thế các nút con của nó dựa vào văn phạm của ngôn ngữ đích, đồng thời số các cây phân cấp ngữ nghĩa được tạo ra có thể rất lớn. Không có một sự gợi ý nào từ giai đoạn phân tích (ngôn ngữ nguồn). Điều này khá giống với cách mà con người – dịch giả thường gặp : chọn cách diễn đạt bản ngữ nào thích hợp nhất cho một câu văn cần phải dịch khi họ đã hiểu đầy đủ *nội dung* của câu văn?

Như vậy, một cây phân cấp ngữ nghĩa của một câu cụ thể có thể có những bản dịch khác nhau tùy theo bộ giá trị của không gian trạng thái. Phương pháp dịch liên ngữ cho phép tích lũy tri thức về bối cảnh của văn bản (lĩnh vực, không gian, thời gian, thể loại,…tùy thuộc tính chất của từng ngôn ngữ cụ thể [2]) để biên dịch văn bản với chất lượng cao hơn.

[1] Chẳng hạn quy tắc **Preposition** → *into* tương đương với quy tắc **Giới_từ** → *vào*; cách tổ chức đối sánh đa ngữ theo các cặp quy tăc như thế này có vẻ rất trực quan và tự nhiên. Tuy vậy, nó không hữu dụng đối với ngôn ngữ tự nhiên vì cấu trúc các ngôn ngữ tự nhiên khác xa nhau; không tồn tại sự tương ứng quy tắc một một.

[2] Đối với tiếng Việt, ta không cần quan têm đến cách chia động từ nên không cần phải nhớ thì động từ của câu trước để tạo ra câu sau; còn đối với nhiều ngôn ngữ thì điều này là quan trọng.

Việc không có sự tương ứng *Quy tắc – Quy tắc* (mà chỉ có sự tương ứng *Mẫu đoạn – Mẫu đoạn*) giữa các ngôn ngữ với nhau trong mô hình dịch sẽ dẫn đến việc tổng hợp bản dịch trở thành khối xử lý phức tạp nhất trong quy trình dịch máy. Nhưng cũng chính điều này làm cho mô hình liên ngữ trở nên khả thi trong ứng dụng thực tế : sự tương ứng *Mẫu đoạn – Mẫu đoạn* có thể cập nhật thủ công dễ dàng bởi những người không chuyên (về máy tính) hoặc thu thập tự động từ kho ngữ liệu song ngữ hoặc đa ngữ.

III.5.2. HÌNH THỨC HÓA

Khái niệm *biên dịch*, hay *phiên dịch* hay đơn giản, *dịch,* thường không được định nghĩa chặt chẽ. Chúng tôi chưa tìm thấy ở đâu có định nghĩa toán học của khái niệm này.

Trong từ điển tiếng Anh "*Oxford English Dictionary*", 1997, mục từ "***translation"*** được giải thích là :

> "*The action or process of turning from one language into another; also, the product of this; a version in a different language*"*1*

và mục từ "***translate"*** được giải thích là :

> *"To turn from one language into another; 'to change into another language retaining the sense' (J.); to render; also to express in other words, to paraphrase. (The chief current sense)"2*

Phần này giới thiệu một phác thảo định nghĩa hình thức của khái niệm dịch thông qua mô hình dịch liên ngữ dựa trên văn phạm cảm ngữ đoạn được đề xuất.

Giả sử Gi = (Σi, Ni, Ai, Si, Pi) với i = 1, 2,…n là văn phạm cảm ngữ đoạn xác định các ngôn ngữ Li.

Định nghĩa 1.

Họ văn phạm là bộ $\mathcal{T}$ = (Σ, N, A, S, P, E), trong đó :

- Σ = Σ1 U Σ2 U…U Σn
- N = N1 U N2 U…U Nn
- A = A1 U A2 U … U An
- S = {S1, S2,… Sn}, Si ∈ Ni; S là một phần tử của E
- P = P1 U P2 U … U Pn

---

[1] Hành vi hoặc quá trình chuyển hóa từ ngôn ngữ này sang ngôn ngữ khác; còn là sản phẩm của hành vi hoặc quá trình đó; một phiên bản trên một ngôn ngữ khác

[2] Chuyển hóa từ ngôn ngữ này sang ngôn ngữ khác; thay đổi sang ngôn ngữ khác vẫn giữ nguyên ý nghĩa; tái hiện; còn là, phát biểu bằng những từ ngữ khác, viết lại

- Mỗi văn phạm Gi = (Σi, Ni, Ai, Si, Pi) xác định một ngôn ngữ – Li
- E là một bảng (gọi là *bảng đồng nghĩa*), trong đó mỗi phần tử là một danh sách các *mẫu đoạn* (được gọi là các mẫu đoạn tương đương với nhau).

Họ văn phạm xác định một họ các ngôn ngữ và sự liên hệ của chúng thông qua bảng đồng nghĩa E.

Định nghĩa 2.

Ta gọi chuỗi ký hiệu t là một ***phép viết lại*** trên ngôn ngữ Lj của một chuỗi ký hiệu s cho trước nếu tồn tại i sao cho Si $\Rightarrow*$ s và bằng cách thay thế từng phần của s sử dụng các chuỗi ký hiệu trong cùng một phần tử từ bảng đồng nghĩa E và cuối cùng nhận được t sao cho Sj $\Rightarrow*$ t.

Phép viết lại t trên ngôn ngữ Lj của một chuỗi ký hiệu s cũng được gọi là ***bản dịch sang ngôn ngữ*** Lj của s.

Ta gọi bản dịch t_1 là có ***tính thành ngữ*** hơn t_2 nếu các chuỗi ký hiệu tương ứng có tính thành ngữ hơn.

Lựa chọn t trong số các ***bản dịch tốt*** sang Lj của s sao cho: với mọi bản dịch t_k của s (sang Lj) thì t có tính thành ngữ hơn hoặc tương đương t_k (t là bản dịch của s sang Lj có tính thành ngữ nhất).

Trong định nghĩa *dịch,* hay *viết lại*, ta không nhắc nhở đến sự phân tích hay tổng hợp; phân tích là việc xác định i trong họ văn phạm $\mathcal{T}$ sao cho Si $\Rightarrow*$ s (*dựng cây phân cấp ngữ nghĩa từ chuỗi ký hiệu s*), còn tổng hợp chính là từ Si $\Rightarrow*$ s xây dựng Sj $\Rightarrow*$ t (bằng các phép thay thế trong cây phân cấp ngữ nghĩa các mẫu đoạn thuộc Gi bởi các mẫu đoạn thuộc Gj thông qua bảng đối chiếu đa ngữ – *bảng đồng nghĩa*).

Trong trường hợp i = j sơ đồ *dịch* suy biến thành *hiệu chỉnh văn phong*.

III.5.3. NHỮNG ÍCH LỢI CỦA PHƯƠNG PHÁP

Mô hình dịch máy của NACENTECH có những ưu điểm chính sau:

- *Cơ sở dữ liệu tinh gọn*, dễ dàng bổ sung ngôn ngữ mới vào hệ dịch đa ngữ : chỉ cần thêm một bộ tri thức đơn ngữ (bộ quy tắc cảm ngữ đoạn) và thông tin trong bảng đối chiếu đa ngữ thì sẽ có hệ dịch từ ngôn ngữ mới đến tất cả các ngôn ngữ đã có trong hệ và ngược lại (là ưu điểm chung của cách tiếp cận liên ngữ). Hơn thế, với việc sử dụng văn phạm cảm ngữ đoạn, cơ sở tri thức đơn ngữ (cho mỗi ngôn ngữ) chỉ cần chứa một bộ văn phạm (*bao gồm hệ*

quy tắc từ vựng, cú pháp và ngữ dụng của ngôn ngữ). Do tính chất nghịch đảo của văn phạm cảm ngữ đoạn, bộ văn phạm này được sử dụng vừa để phân tích vừa để tổng hợp.

- *Độc lập ngôn ngữ*. Văn bản dịch chỉ phụ thuộc vào văn phạm của ngôn ngữ đích, không bị ảnh hưởng bởi văn phạm hay trình tự đặt câu (lối viết) của ngôn ngữ nguồn. Biểu diễn bên trong không phụ thuộc ngôn ngữ, quá trình phân tích và tổng hợp diễn ra độc lập với nhau. Việc đưa một ngôn ngữ mới vào hệ thống được đơn thể hóa: Phần tri thức ngôn ngữ là hoàn toàn đơn ngữ. Mối liên hệ giữa các ngôn ngữ được cập nhật thông qua việc tìm các cấu trúc đồng nghĩa của các khái niệm được định nghĩa trong bảng đối chiếu đa ngữ.
- Tính *tùy biến*. Hệ thống có thể được phát triển để thích nghi với *thể loại* hoặc với *phương ngữ* của văn bản tùy theo nhu cầu diễn đạt của người sử dụng.
- Hệ thống biên dịch trên cơ sở *toàn bài văn* : khi dịch mỗi câu đều đối chiếu với những thông tin cần thiết của các câu trước (thông qua *những giá trị được tích lũy trong không gian trạng thái*) cũng như bổ sung những thông tin cần thiết (*thông qua việc cập nhật các giá trị của không gian trạng thái*) để giúp cho việc biên dịch sát đúng hơn với ngữ cảnh : Một câu cụ thể có thể được dịch ra thành những câu *khác nhau* tùy theo văn cảnh.
- Cơ chế *giải quyết nhập nhằng* (chọn văn bản tốt nhất) dựa trên cơ sở mô hình dàn đại số có thể được ứng dụng trong các bài toán xử lý ngôn ngữ tự nhiên nói chung (*trong đó có dịch tự động*), chẳng hạn khi ngôn ngữ nguồn của văn bản là tiếng Việt và ta chọn ngôn ngữ đích cũng là tiếng Việt thì động cơ dịch trở thành một bộ kiểm tra và hiệu chỉnh văn phạm.
- Hệ thống *khả thi*. Mô hình cho phép bổ sung tri thức ngôn ngữ trong quá trình vận hành. Chất lượng sẽ phụ thuộc vào cơ sở tri thức. Hệ thống được mở rộng tự động khi bổ sung tri thức mới của ngôn ngữ cũng như khi đưa ngôn ngữ mới vào.
- Khả năng *đoán nhận ngôn ngữ*. Sơ đồ dịch cho phép hệ thống tự nhận biết ngôn ngữ nguồn trên cơ sở thực hiện những đánh giá khi phân tích câu. Tính năng này cho phép xây dựng hệ dịch những *văn bản đa ngữ* (khi một bài văn đồng thời chứa văn bản của nhiều ngôn ngữ khác nhau), hoặc biên dịch những văn bản mà người sử dụng không xác định được (*hoặc không đòi hỏi người sử dụng phải xác định*) ngôn ngữ nguồn. Tính chất này đặc biệt hữu

ích trong những ứng dụng trực tuyến (duyệt internet, thư điện tử hoặc tán gẫu đa ngữ). Người sử dụng chỉ cần chọn ngôn ngữ đích (là ngôn ngữ mà mình muốn sử dụng để xem) mà không cần biết văn bản của bên đối thoại thuộc ngôn ngữ nào.

IV. CÔNG CỤ CẬP NHẬT DỮ LIỆU

IV.1. CƠ SỞ TRI THỨC (KNOWLEDGE BASE) IV-2
IV.1.1. TRA CỨU TRONG CƠ SỞ TRI THỨC IV-3
IV.1.2. TÌM KIẾM TỪ THEO BIỂU THỨC CHÍNH QUY. IV-7
IV.1.3. XEM DANH SÁCH CÁC TỪ ĐÃ TRA CỨU IV-8
IV.1.4. BỔ SUNG, XÓA, SỬA CHỮA MỘT MỤC TỪ IV-9
IV.1.5. BỔ SUNG MỘT MỤC TỪ VÀO CƠ SỞ TRI THỨC IV-9
IV.1.6. XÓA MỘT MỤC TỪ TRONG CƠ SỞ TRI THỨC IV-14
IV.1.7. THÊM HOẶC THAY THẾ MỘT MỤC TỪ VÀO CSTT IV-16

IV.2. TRỢ LÝ NGÔN NGỮ (LANGUAGE ASSISTANT) IV-17
IV.2.1. GIỚI THIỆU CHUNG IV-18
IV.2.2. CÁCH SỬ DỤNG TRỢ LÝ NGÔN NGỮ IV-20
IV.2.3. DỊCH CÂU TRONG TRỢ LÝ NGÔN NGỮ IV-22
IV.2.4. TRA CỨU CHÉO IV-23

IV.3. BỔ SUNG FILE VĂN BẢN VÀO DỮ LIỆU IV-25

Việc cập nhật dữ liệu đóng vai trò thiết yếu trong mọi phần mềm dịch máy. Sự linh hoạt và khả năng đối chiếu, so sánh dữ liệu khi cập nhật cơ sở tri thức có ý nghĩa quyết định đến chất lượng dịch thuật của các hệ quy mô lớn vì một khi cơ sở tri thức đủ lớn thì người dùng có thể không còn kiểm soát được kho dữ liệu của mình nữa. Phần này mô tả hướng dẫn sử dụng cơ sở tri thức.

IV.1. CƠ SỞ TRI THỨC (KNOWLEDGE BASE)

Cơ sở tri thức bao gồm tất cả thông tin về ngôn ngữ mà chương trình dịch sử dụng để phân tích câu tiếng Anh cũng như để tổng hợp câu tiếng Việt khi dịch văn bản, bao gồm các quy tắc văn phạm, quy tắc dịch và từ điển.

Để chương trình dịch sát nghĩa hơn theo từng yêu cầu chuyên biệt, từng ngữ cảnh văn phong riêng, người sử dụng cần có toàn quyền kiểm soát và tra cứu dữ liệu.

Bộ soạn thảo cơ sở tri thức của phần mềm EVTRAN có những tính năng linh hoạt (*trong đó có những đặc điểm không thấy trong các phần mềm dịch máy khác của thế giới*) như:

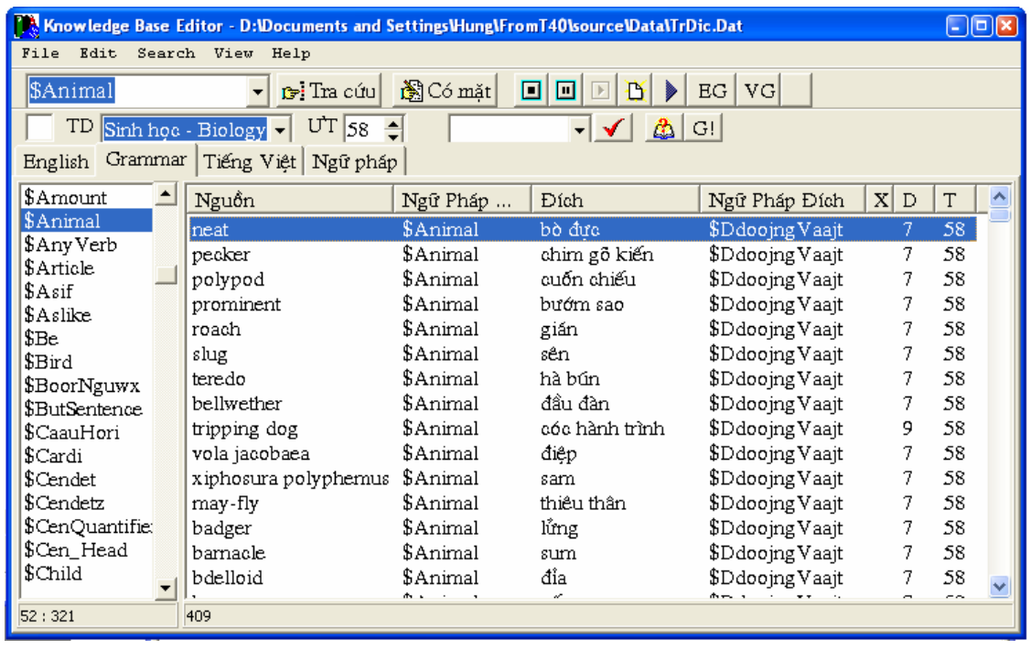
- Có một hệ chỉ mục đa dạng, giúp tra cứu xuôi, ngược, tổ hợp từ, khái niệm văn phạm, tiếp đầu, tiếp đuôi,...
- Tra cứu ***chéo*** vào các từ điển đơn ngữ, song ngữ khác để hỗ trợ người nhập tri thức (*knowledge engineers*) trong việc cập nhật thông tin ngôn ngữ
- Phân tích cú pháp cho mỗi mẫu đoạn trong cơ sở tri thức để hỗ trợ việc tổ chức một kho tri thức không mâu thuẫn, trực quan (đây là khả năng đặc biệt, chưa thấy trong các hệ dịch tự động phổ biến nào)
- Có công cụ theo dõi trực quan cây cú pháp được phân tích cho mỗi câu trong cả tiếng Anh lẫn tiếng Việt.
- Đặt nền tảng ban đầu cho việc tổ chức một cơ sở tri thức duy nhất cho cả dịch máy tự động (*bằng chương trình máy tính*) lẫn tra cứu (*bằng con người*)

IV.1.1. TRA CỨU TRONG CƠ SỞ TRI THỨC

Trước hết, khi mở cửa sổ Knowledge Base Editor sẽ xuất hiện giao diện cửa sổ soạn thảo cơ cở tri thức như trên Hình 1.

Phần xem nội dung các quy tắc dịch gồm có các thực đơn, các nút tra cứu và 4 trang hiển thị thông tin tra cứu :

- ***English*** để hiển thị kết quả khi tra từ tiếng Anh
- ***Grammar*** để hiển thị kết quả khi tra khái niệm văn phạm (*nonterminal*) tiếng Anh
- ***Tiếng Việt*** để hiển thị kết quả khi tra từ tiếng Việt
- ***Ngữ pháp*** để hiển thị kết quả khi tra khái niệm văn phạm (*nonterminal*) tiếng Việt



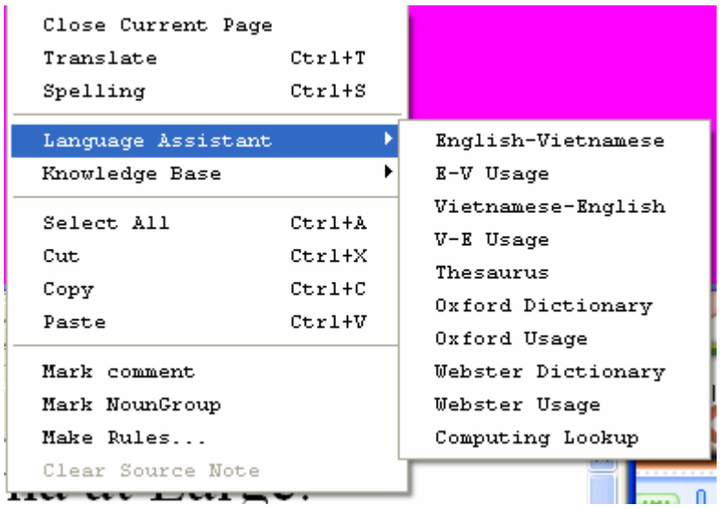
Hình 1: Cửa sổ xem và cập nhật từ điển dịch máy

Mở cơ sở tri thức để tra cứu có thể bắt đầu vào từ thanh menu, hay có thể xuất phát từ bất kỳ vị trí nào trong chương trình dịch, bằng cách chọn từ cần tra và sau đó gọi menu thả xuống, trên menu này xuất hiện tất cả các loại tra cứu, dịch ... chọn mục cơ sở tri thức và loại từ muốn tra. Khi đó cửa sổ giao diện chính của cơ sở tri thức sẽ xuất hiện, cho phép tra cứu, hiệu chỉnh, tuỳ biến theo yêu cầu của người sử dụng (Hình 2).

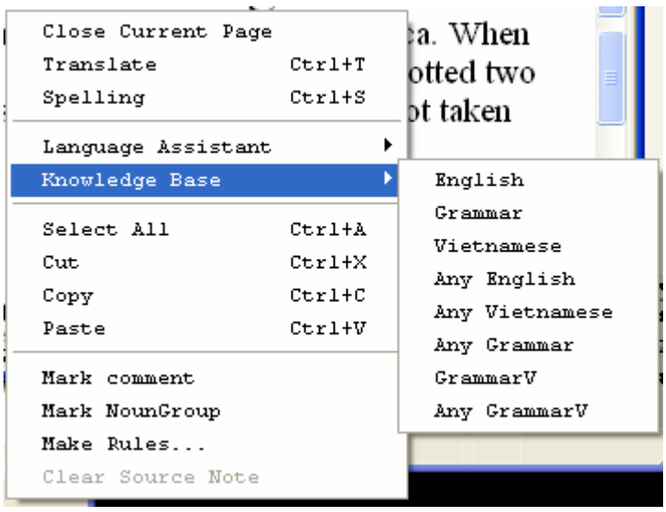
Đây là cách thuận tiện cho việc tra cứu. Nếu ta đang tra muốn tra cứu một từ có sẵn trên máy mà phải lần mở từng trang từ điển trên giấy thì thật

vất vả. Bằng cách này đã rút ngắn lại các công đoạn đó, giúp cho việc tra cứu dễ dàng thuận tiện hơn nhiều. Ngoài ra, còn cho ta chọn lựa tùy theo công việc muốn tra cứu trong cơ sở tri thức hay trong trợ lý ngôn ngữ.

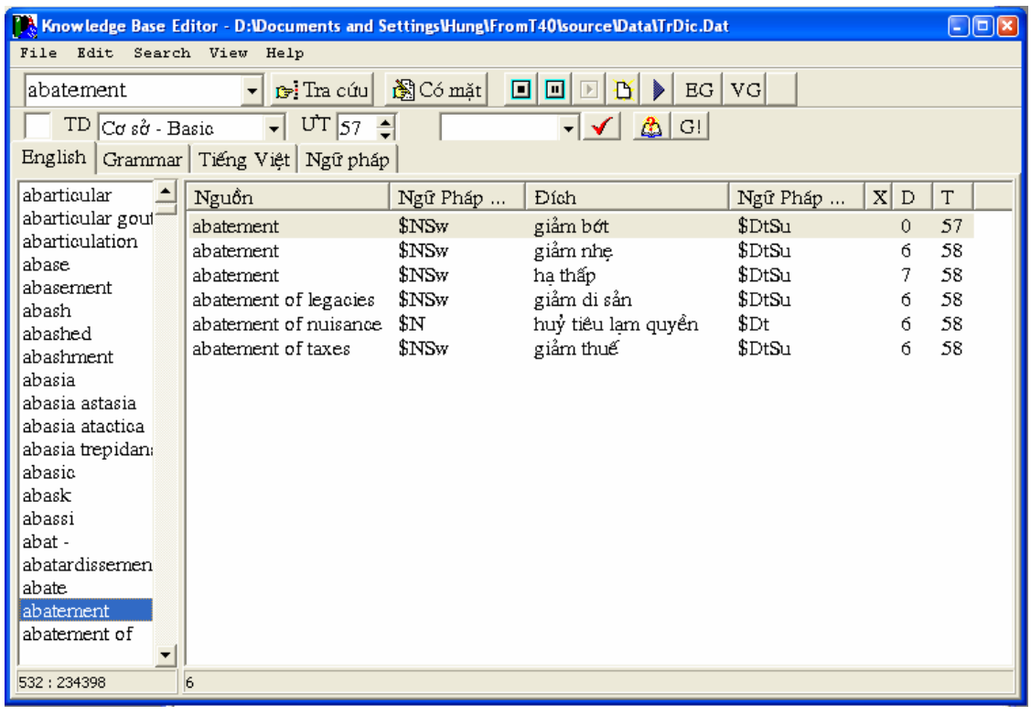
Cơ sở tri thức của chương trình dịch bao gồm các quy tắc văn phạm Anh – Việt và Việt – Anh. Để cung cấp cho người sử dụng một cách nhìn tổng quát về các quy tắc này, cũng như mối quan hệ giữa từ vựng tiếng Anh và nghĩa tiếng Việt của chúng, có bốn trang hiển thị các quan hệ khác nhau: Anh-Việt, Việt-Anh, Văn phạm Anh-Việt và Văn phạm Việt Anh, người sử dụng có thể tra cứu xuôi ngược tùy theo nhu cầu sử dụng để tra trên các trang thích hợp (Hình 2.1, 2.2).



Hình 2.1 : Thực đơn thả xuống

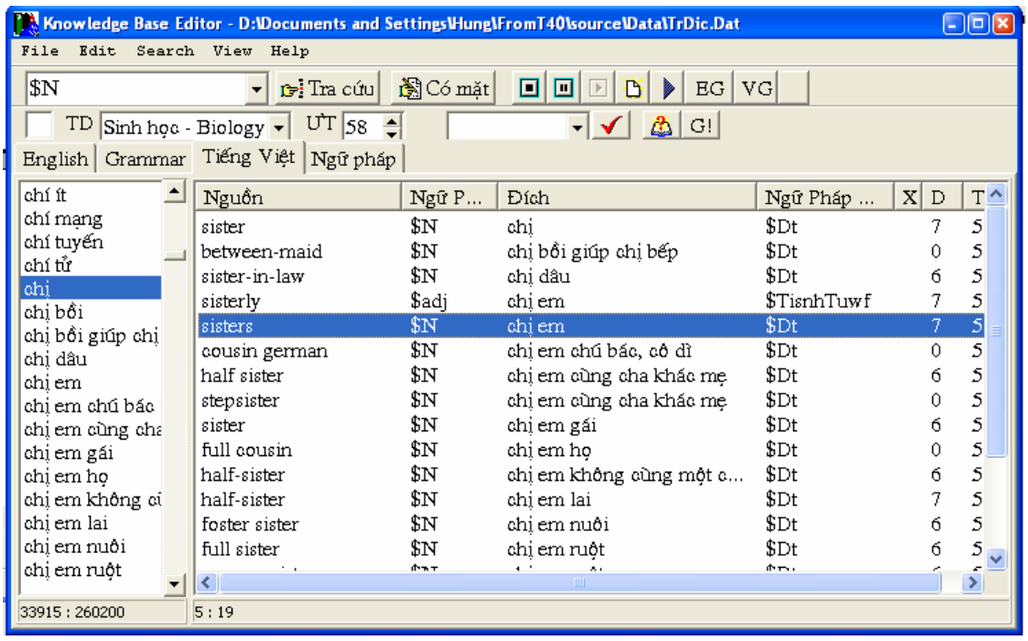


Hình 2.2 : Thực đơn thả xuống



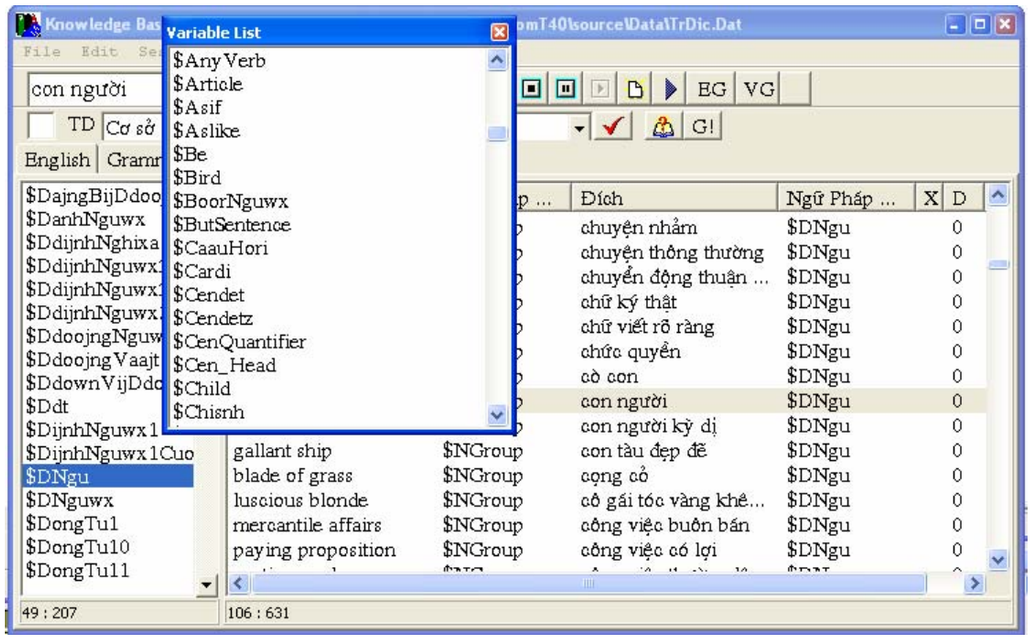
Hình 3: Xem và cập nhật từ điển dịch máy: tra từ tiếng Anh

Sau khi chọn trang từ điển muốn tra, nhập từ muốn tra cứu. Hộp danh sách từ sẽ tự động cuốn theo tới từ tương ứng, nếu từ nhập vào mà chưa biết chính xác chính tả, cửa sổ các từ gần sẽ cho các gợi ý chọn lựa, để người sử dụng chọn từ muốn tra cứu.

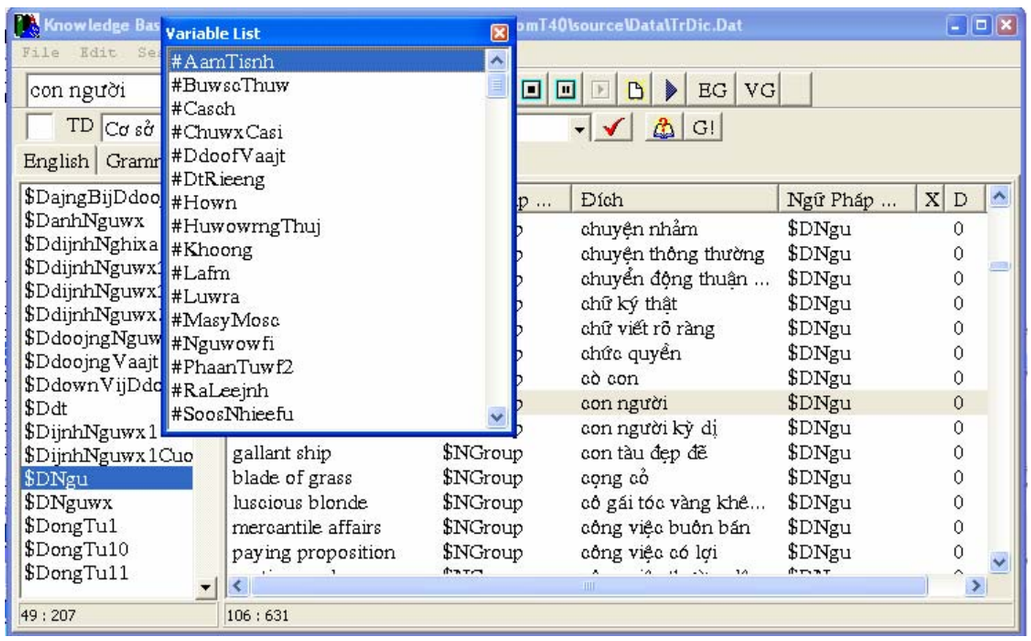


Hình 4: Xem và cập nhật từ điển dịch máy: tra từ tiếngViệt

Nghĩa của từ cần tra sẽ hiển thị trong cửa sổ giải nghĩa. Trên cửa sổ giải nghĩa ta không những tra được nghĩa của từ mà còn xem được loại từ, các cụm từ thường đi kèm ...



Hình 5: Xem và cập nhật từ điển dịch máy:Danh sách biến trung gian (nonterminals) gợi ý tiếng Anh



Hình 6: Xem và cập nhật từ điển dịch máy:Danh sách biến trung gian (nonterminals) gợi ý tiếng Việt

Bên cạnh việc tra cứu để biết ngữ pháp, nghĩa tiếng Việt. Nút **Có mặt** còn cho biết xem sự có mặt của một mục từ trong cửa sổ giải nghĩa hay không (có thể nằm trong các phần Nguồn, Ngữ pháp, Đích... của cửa sổ giải nghĩa).

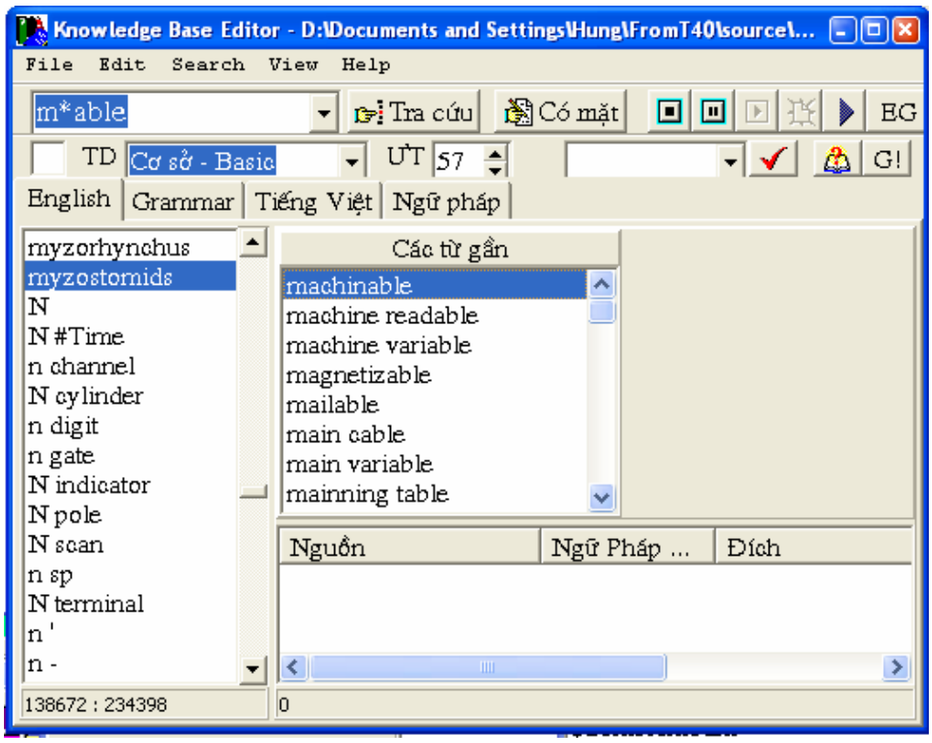
IV.1.2. TÌM KIẾM TỪ THEO BIỂU THỨC CHÍNH QUY.

Đôi khi việc tìm kiếm gây ra mất thời gian, đặc biệt là các từ tìm kiếm không nhớ chính xác chính tả, hay tìm kiếm theo từ loại chỉ biết biết tiếp đầu ngữ hoặc tiếp vị ngữ của từ đó. Cơ sở tri thức cung cấp một chức năng tìm kiếm đáp ứng cho yêu cầu tìm kiếm này. Để tìm kiếm từ trong vùng nhập từ người sử dụng cần chỉ ra các tiêu chuẩn tìm kiếm từ với các qui ước như sau:

Ký tự ý nghĩa

***** Thay thế cho một số lượng ký tự bất kỳ (≥ 0).

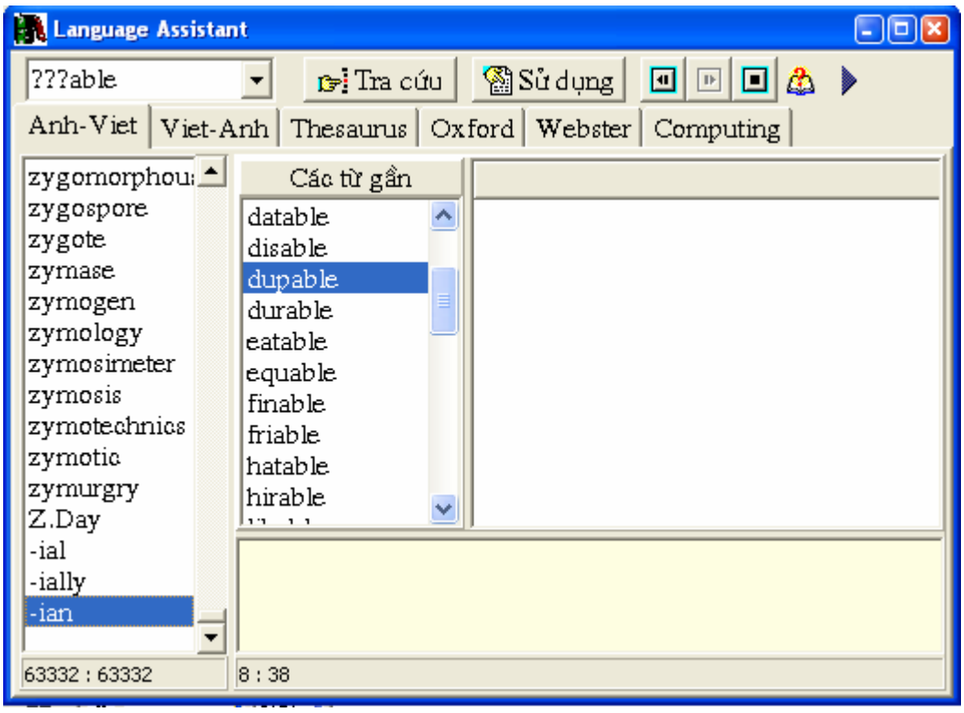
? Thay thế cho một ký tự bất kỳ.



Hình 7: Sử dụng biểu thức chính quy

Sau khi đã chỉ ra tiêu chuẩn tìm kiếm trong ô nhập từ. Chương trình sẽ tìm kiếm tất cả các từ đáp ứng đúng tiêu chuẩn tìm kiếm đã đưa ra. Danh sách các từ tìm thấy sẽ được hiển thị trong hộp danh sách các từ gần giống. Người sử dụng có thể xác định từ muốn tìm trong danh sách kết quả này. Điều kiện tìm kiếm được chỉ ra càng chi tiết, danh sách kết quả tìm kiếm sẽ

càng ngắn và sẽ dễ dàng hơn khi xác định từ cần tra trong danh sách kết quả tìm kiếm.



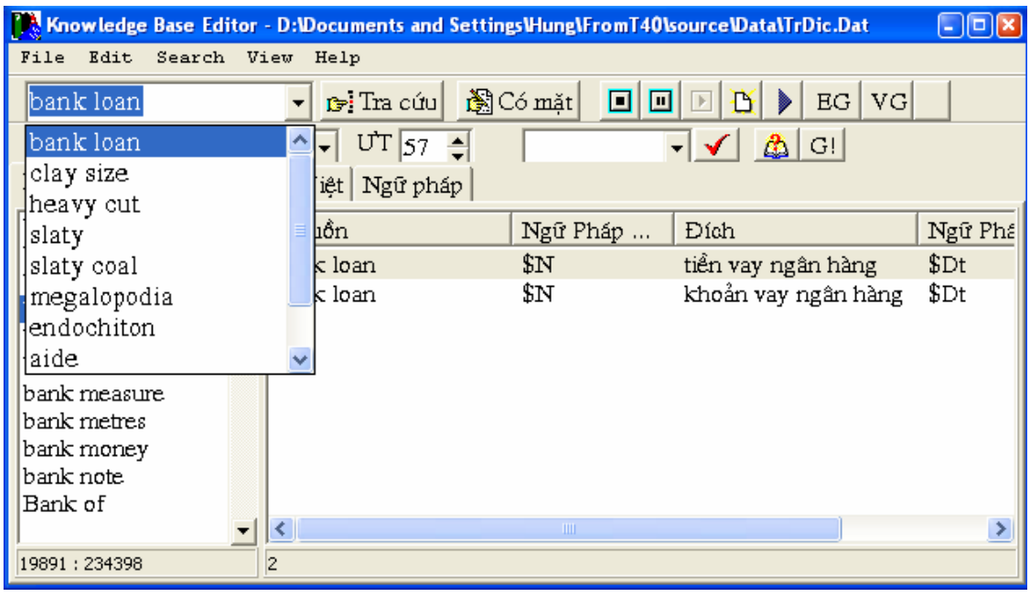
Hình 8 : Sử dụng biểu thức chính quy

Ví dụ: Muốn tra các từ bắt đầu là "`m`" và kết thúc là "`able`" ta có danh sách các từ gần hiện lên như trên Hình 7.

Ví dụ: Muốn tra các từ gồm 7 ký tự và kết thúc là "`able`" ta có danh sách các từ gần hiện lên như trên hình 8.

IV.1.3. XEM DANH SÁCH CÁC TỪ ĐÃ TRA CỨU

Khi muốn tra cứu lại một từ, mà không muốn gõ lại từ đó vào cửa sổ nhập từ, thì ta kích chuột lên biểu tượng cạnh ô nhập từ, dùng con trỏ chuột để tìm từ cần xem lại. Chương trình đã tự động lưu lại các từ mới tra, để người sử dụng tiện tra cứu lại. Và biết được danh sách các từ mình đã tra.



Hình 9 : Xem danh sách những từ đã tra cứu

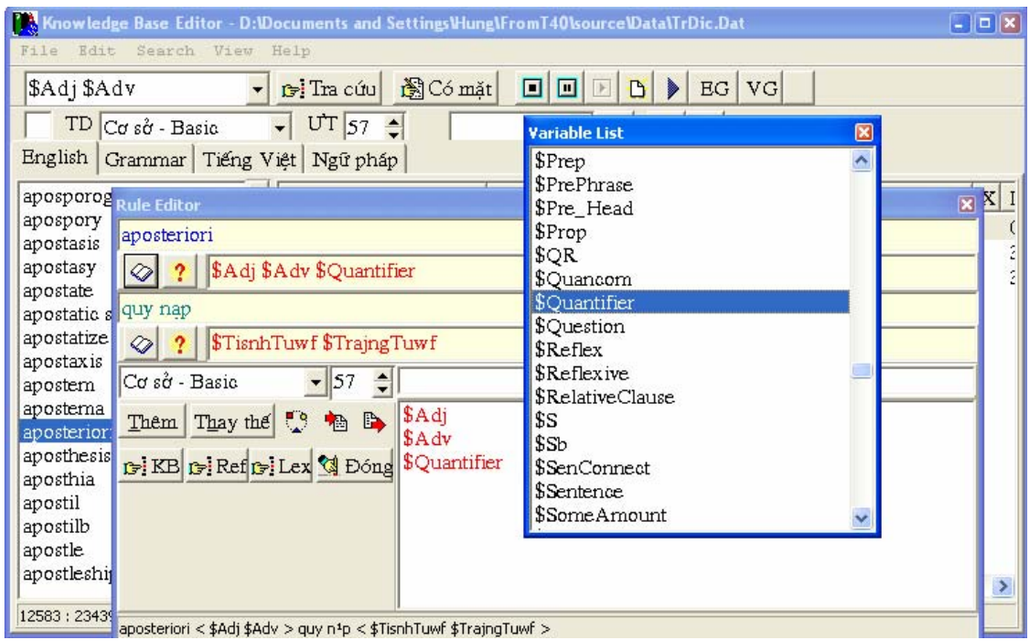
IV.1.4. BỔ SUNG, XÓA, SỬA CHỮA MỘT MỤC TỪ

Trong cơ sở tri thức cho phép bổ sung, xóa bỏ hay sửa đổi các mục để phù hợp với chuyên môn, và yêu cầu của từng người sử dụng khác nhau. Đây là bước cập nhật dữ liệu, tạo cho chương trình dịch có một cơ sở tri thức linh hoạt, và sẽ dịch tốt hơn.

IV.1.5. BỔ SUNG MỘT MỤC TỪ VÀO CƠ SỞ TRI THỨC

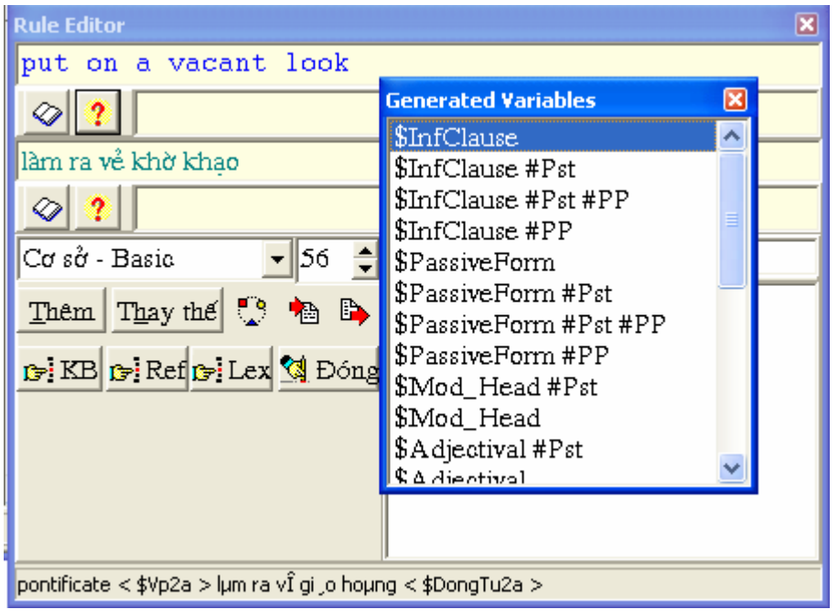
Nhập cụm từ muốn bổ sung, chọn ngữ pháp, điền nghĩa của từ, chọn tên chuyên ngành phù hợp với cụm từ được bổ sung, hoặc sửa đổi, chọn mức độ ưu tiên (mục nào có số thứ tự ưu tiên càng nhỏ thì độ ưu tiên càng cao). Từ lúc đó, chương trình dịch có thêm một quy tắc mới để xử lý, lựa chọn, và cho kết quả dịch tùy theo yêu cầu người sử dụng.

Nút **KB** (Knowledge Base) để tra cứu mục từ trong ô nhập từ, cửa sổ giải nghĩa của cơ sở tri thức hiện ra, cho chúng ta xem phần mình muốn bổ sung có chưa, hay là sửa đổi ngữ pháp cho đúng. Nút **Ref** cũng giúp ta tra cứu tham khảo nhưng trong từ điển Trợ lý ngôn ngữ. Ngoài ra, nếu muốn tham khảo từ vựng chọn nút **Lex** (Lexical), sẽ xuất hiện những cụm từ vựng liên quan.



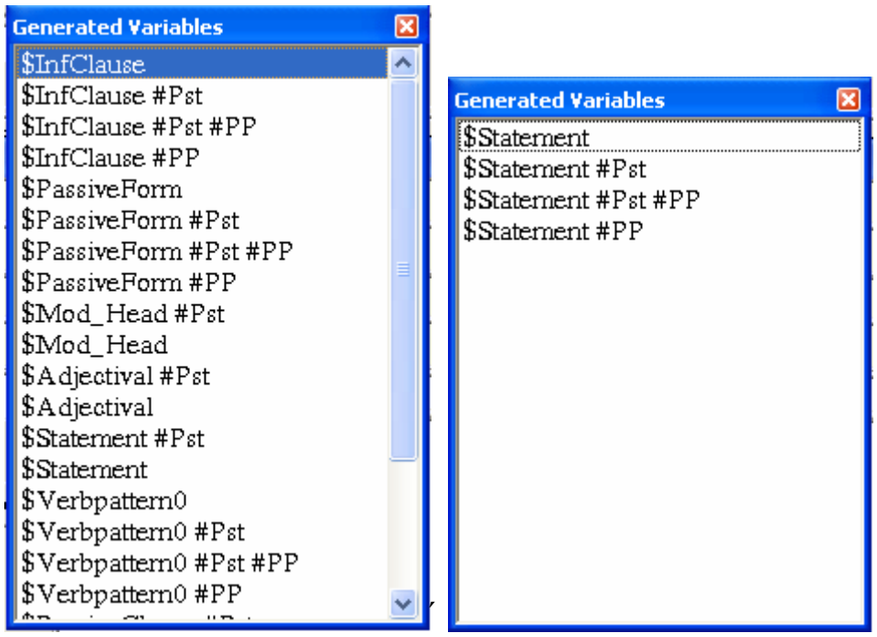
Hình 10 : Chọn khái niệm ngữ pháp

Ví dụ: Đang thêm vào cơ sở tri thức từ "get up" nhưng chưa biết nghĩa của nó, thì có thể tra trong từ điển xem nghĩa, hoặc cơ sở tri thức xem đã có chưa bằng cách kích chuột tương ứng vào nút Ref hoặc KB



Hình 11 : Tra cứu phân tích văn phạm

Ví dụ: Ta bổ sung vào cơ sở tri thức "put on a vacant look" và nghĩa trong ô tiếng việt (ô thứ 3) "làm ra vẻ khờ khạo" có cửa sổ bổ sung như trên Hình 11:

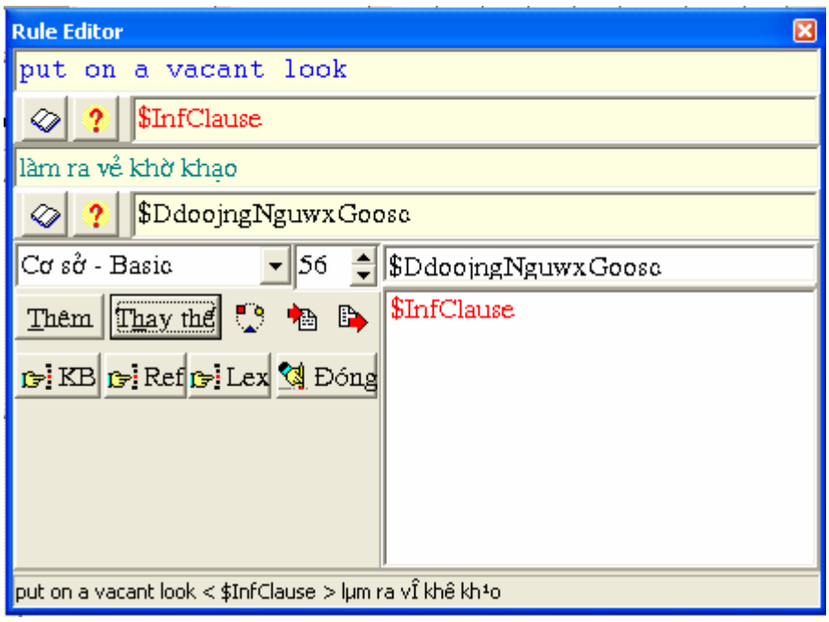


Hình 12 : Tra cứu phân tích văn phạm

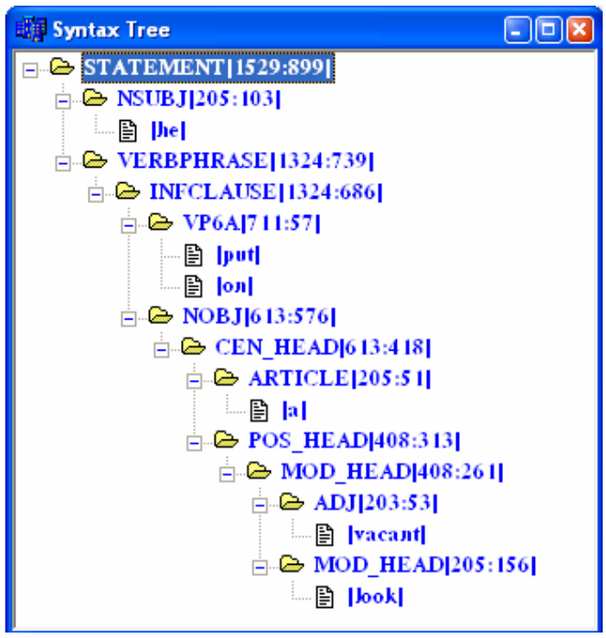
Khi kích chuột vào nút  (Get Support): Chương trình sẽ phân tích cấu trúc văn phạm của mục từ đó, cho ra lời khuyên đối với người dùng nên sử dụng quy tắc nào. Đây là tính năng trợ giúp quan trọng để người sử dụng có thể tham khảo và lựa chọn cấu trúc văn phạm thích hợp cho từng cụm từ hay cấu trúc đưa vào cho hợp lý (Hình 12, phía trái)

Cũng làm như vậy nhưng thêm từ "he" vào trước tức là thêm câu "he put on a vacant look", thì sau khi phân tích lại đưa ra các lựa chọn chỉ ngắn ngọn (xem Hình 12, phía phải).

Sau khi đã chọn các mục ngữ pháp tương ứng như trên Hình 13. Nếu dùng chương trình dịch, cho hiện cây cú pháp lên ta có cấu trúc cây "he put on a vacant look" như trên Hình 14.

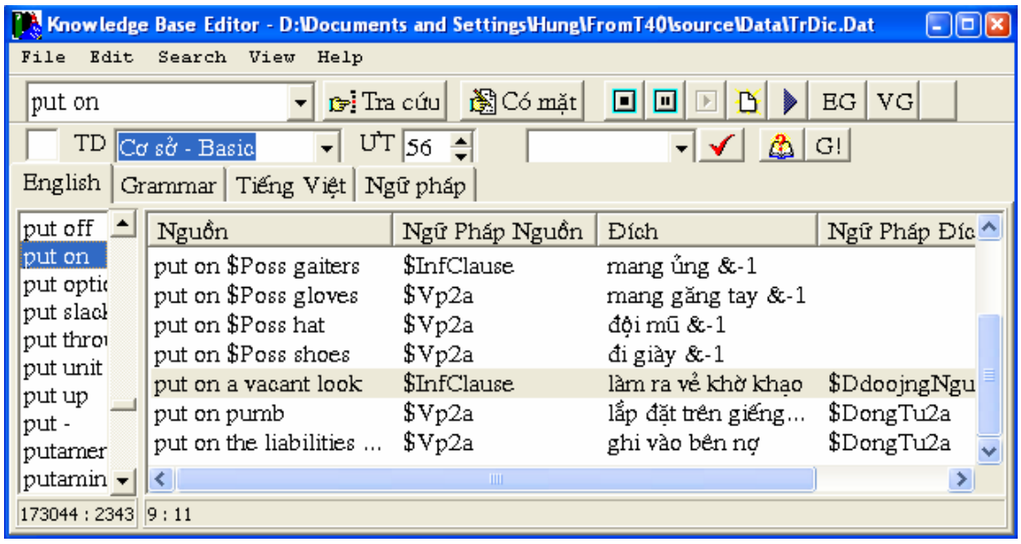


Hình 13 : Bổ sung quy tắc văn phạm

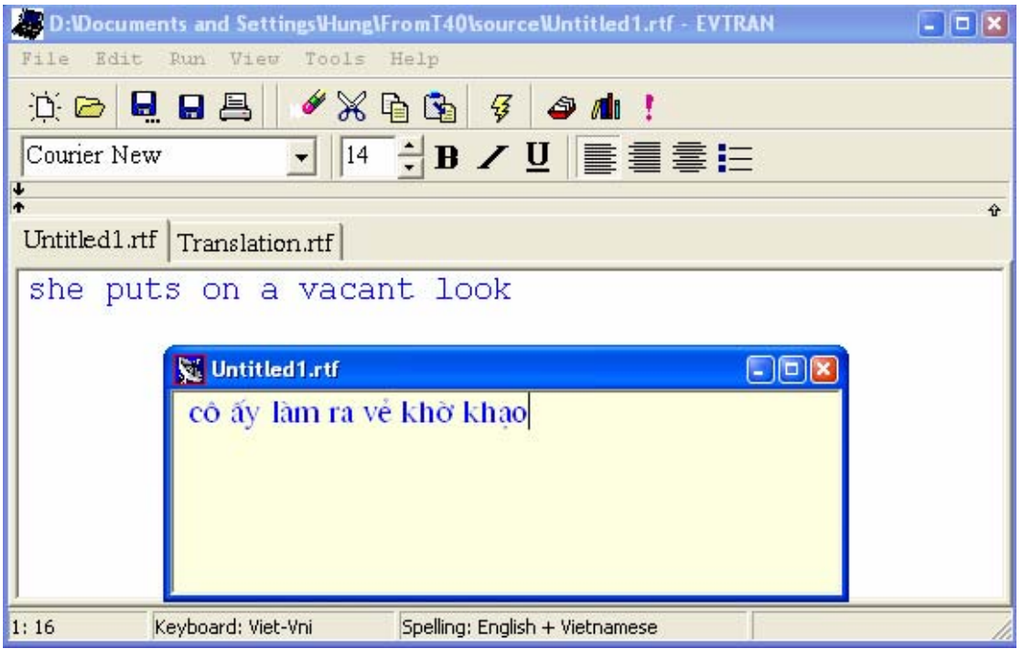


Hình 14 : Xem cây cú pháp

Sau đó chọn chuyên ngành để thêm vào cơ sở tri thức, xem lại trong cơ sở tri thức tra từ "`put on`" ta thấy trong cơ sở tri thức đã có "`put on a vacant look`" như được trình bày trên Hình 15.

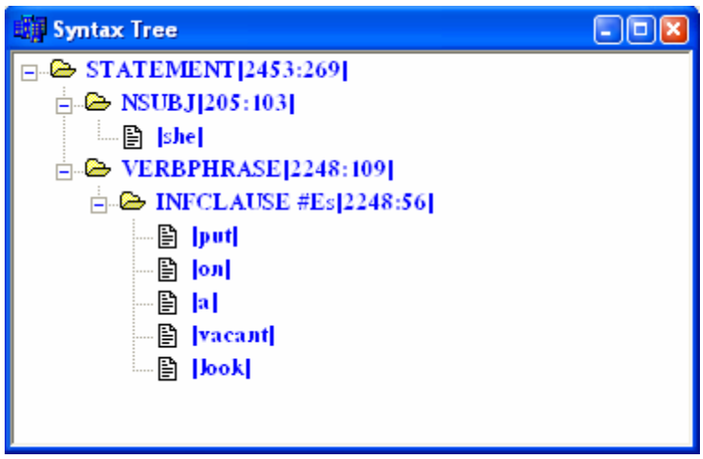


Hình 15 : Tra cứu



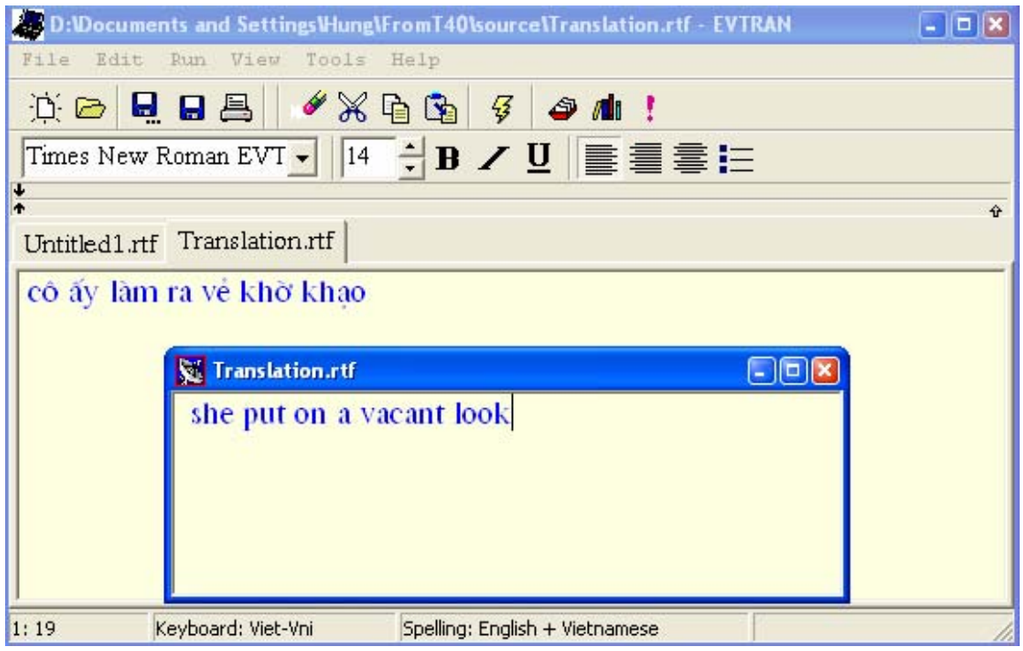
Hình 16 : Dịch thử Anh-Việt

Sau đó dùng chương trình dịch hai câu sau "she puts on a vacant look." "he have put on a vacant look." thành "cô ấy làm ra vẻ khờ khạo." "anh ta đã làm ra vẻ khờ khạo." tương ứng.



Hình 17 : Cây cú pháp sau khi cập nhật quy tắc

Bây giờ, có thể xem phần dịch ngược

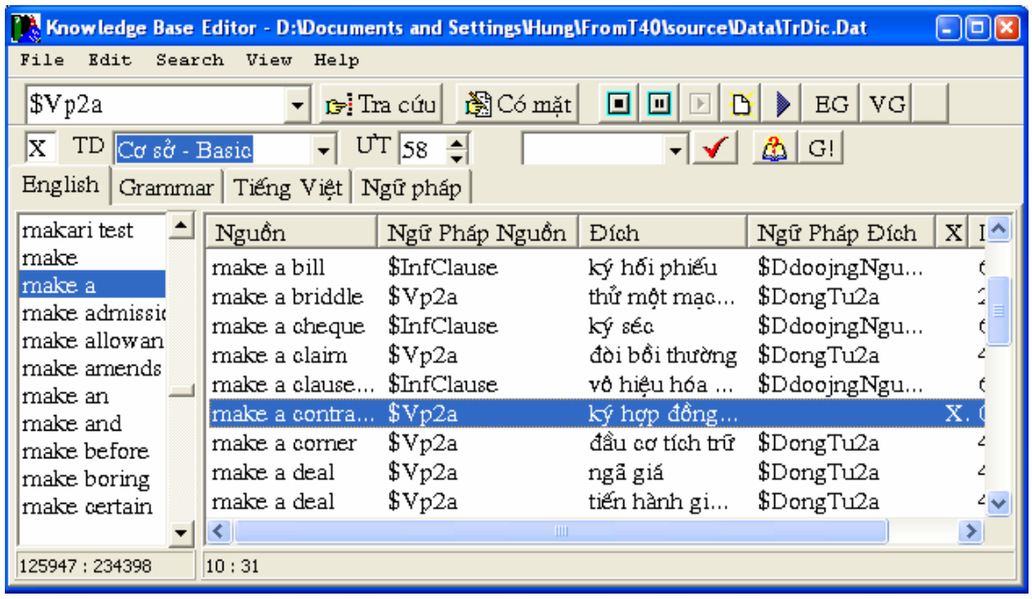


Hình 18 : Dịch thử Việt-Anh

IV.1.6. XÓA MỘT MỤC TỪ TRONG CƠ SỞ TRI THỨC

Chương trình cho phép bổ sung, và cho phép loại bỏ những mục từ, những ngữ pháp không phù hợp với người sử dụng trong cơ sở tri thức. Không phức tạp như bổ sung vào cơ sở tri thức, cần phải xem ngữ nghĩa, từ

loại, việc xóa khỏi cơ sở tri thức dễ dàng, chỉ cần kích phải chuột vào từ cần xóa và chọn **Delete.**



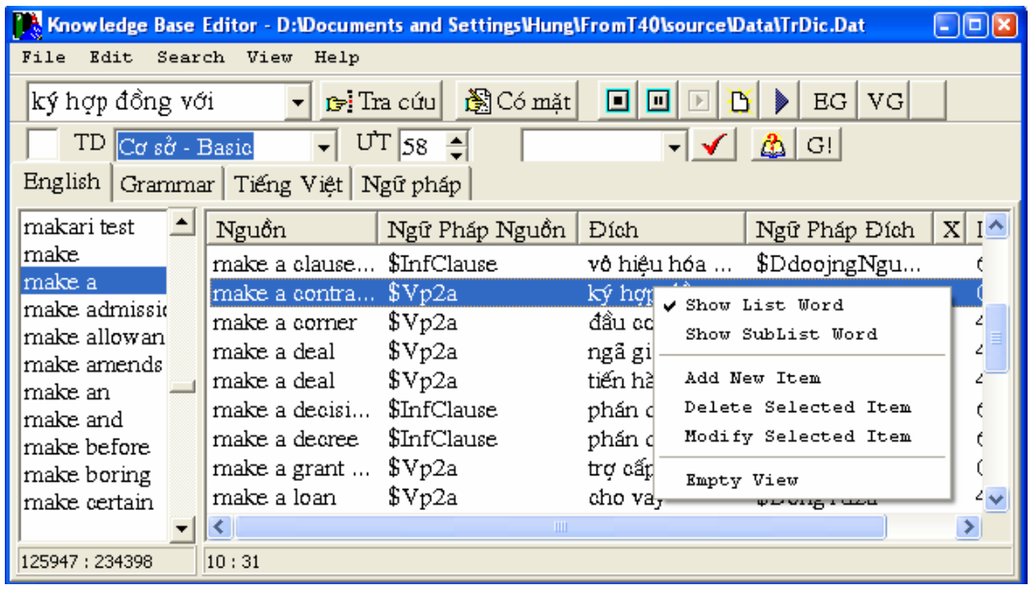
Hình 19 : Xóa một mục từ

Trong ô hiển thị trạng thái xóa (bên dưới ô nhập từ) hiển thị dòng chữ *Đã xóa*. Trong cửa sổ giải nghĩa tại cột X, mục từ bị xóa đã được đánh dấu x (hiển thị từ này đã bị xóa khỏi cơ sở tri thức).

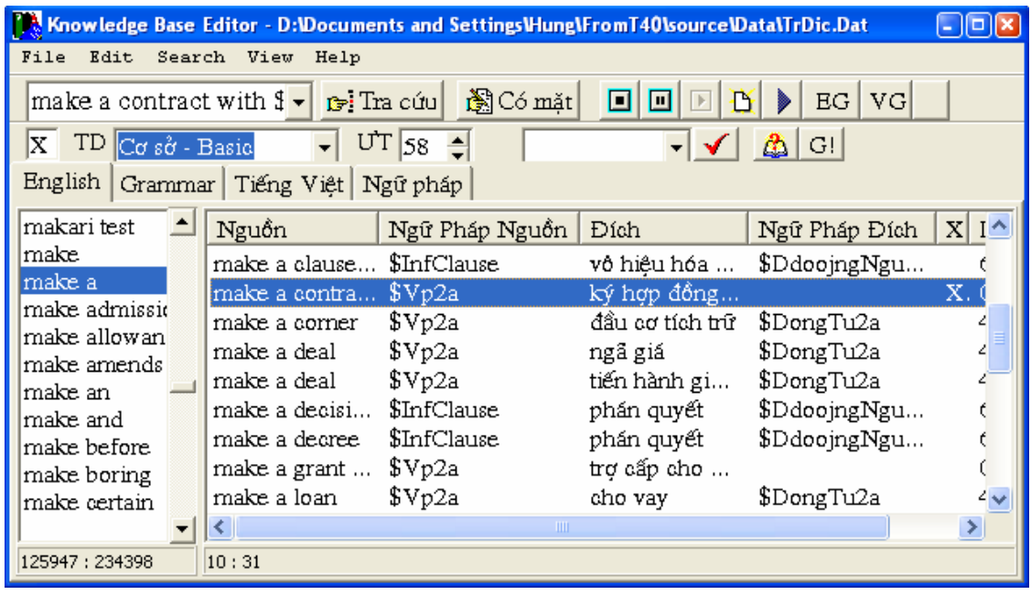
Các mục từ bị xóa trong cơ sở tri thức (không tham gia vào việc dịch văn bản) được đánh dấu x trong cột trạng thái xóa X của phần giải nghĩa, và trong ô hiển thị trạng thái xóa (bên dưới ô nhập từ) có hiển thị dòng chữ *Đã xóa*.

Để khôi phục lại từ này, thực hiện: Kích chuột trái lên mục từ đã bị xóa nếu muốn khôi phục lại, kích chuột phải lên mục từ đó, trên thực đơn lệnh thả xuống kích chuột lên mục **Delete Selected Item** (hoặc bấm phím **Del**).

Khi đó dấu x tại cột X trong cửa sổ giải nghĩa bị xóa bỏ, đồng thời dòng chữ hiển thị trạng thái *Đã xóa* khỏi ô hiển thị trạng thái xóa (bên dưới ô nhập từ).



Hình 20 :Khôi phục lại một mục từ



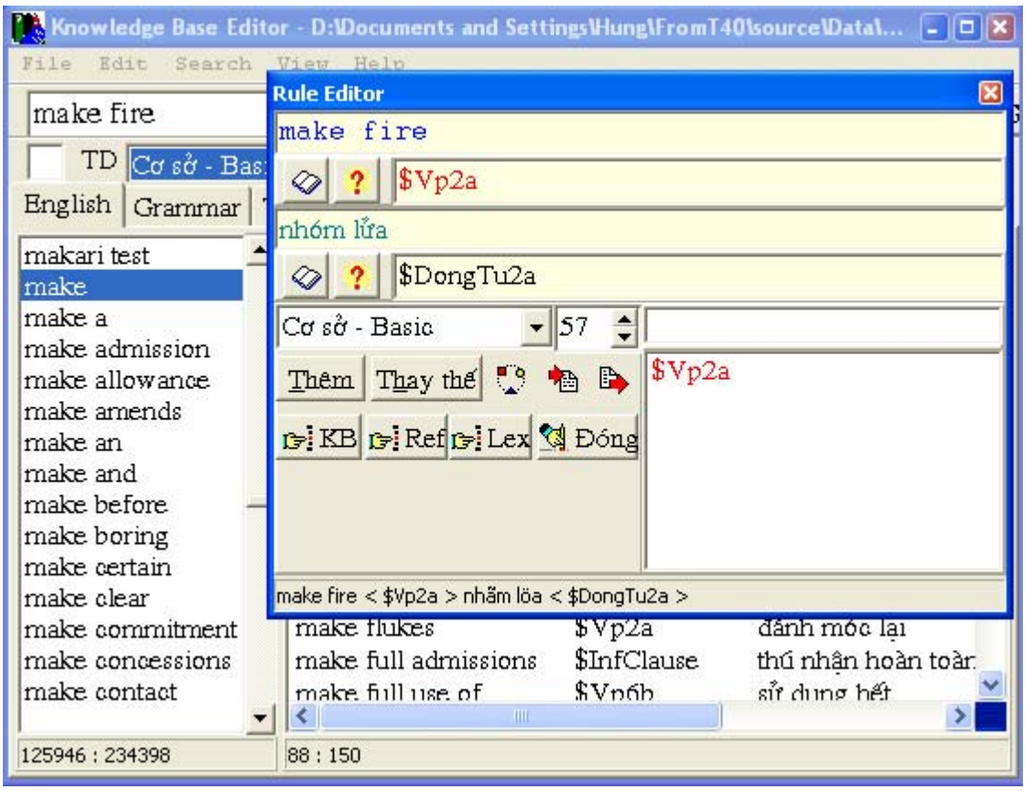
Hình 21 : Mục từ đã bị xóa

IV.1.7. THÊM HOẶC THAY THẾ MỘT MỤC TỪ VÀO CSTT

Trong cửa sổ giải nghĩa của từ, chọn mục **Modify Selected Item** để thay đổi nội dung mục đã được chọn. Gõ cụm từ cần sửa đổi bổ sung vào ô nhập từ. Chương trình sẽ hiển thị một cửa sổ bổ sung và sửa đổi cho phép ta soạn thảo thay đổi lại nội dung mục từ đó.

Sau khi sửa chữa các nội dung của mục đó: Để bổ sung vào Cơ sở tri thức như một mục từ mới, kích chuột trái lên nút **Thêm**. Nếu muốn thay thế mục từ này vào mục từ đang được chọn lựa, kích chuột trái lên nút **Thay thế**. Nếu muốn hủy bỏ việc thay đổi này, kích chuột lên nút **Đóng**.

Trên thực tế, để thay thế một mục từ vào cơ sở tri thức ta làm tương tự như việc thêm vào cơ sở tri thức một mục từ mới, nên ở đây không đưa ra minh họa cụ thể.



Hình 22 : Thêm hoặc thay thế một mục từ

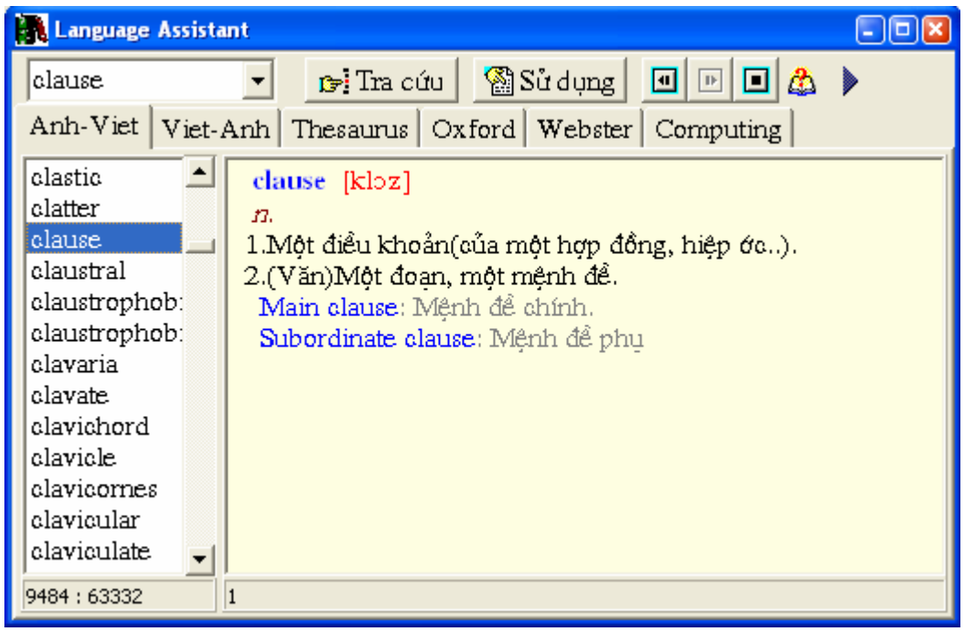
IV.2. TRỢ LÝ NGÔN NGỮ (LANGUAGE ASSISTANT)

Ngoài chức năng biên dịch tự động, phần mềm có một số công cụ hỗ trợ về ngôn ngữ (gọi là ***Trợ lý Ngôn ngữ***) gồm những chức năng tìm kiếm thông tin về ngôn ngữ (tiếng Anh và tiếng Việt) giúp cho người sử dụng tiện tra cứu khi hiệu đính bản dịch cũng như khi biên soạn văn bản. (Chúng tôi tạm dùng thuật ngữ *biên soạn* để chỉ việc tạo ra một văn bản chưa tồn tại, phân biệt với thuật ngữ *soạn thảo* thường được quen dùng để chỉ việc đánh máy và sắp chữ một hoặc nhiều đoạn văn bản có sẵn). Nội dung của Trợ lý ngôn ngữ có nhiều bổ sung so với bản EVTRAN 2.0.

IV.2.1. GIỚI THIỆU CHUNG

Một người phiên dịch hay biên soạn văn bản thường cần tra cứu từ điển không phải chỉ do không hiểu nghĩa một từ cụ thể mà phần nhiều là để chọn được một từ ưng ý nhất. Tuy nhiên, khi tra từ điển không phải lúc nào ta cũng thoả mãn với lời giải thích của một mục từ như cách mà ta thường thấy trong một ***cuốn*** từ điển. Bên cạnh đó, không phải khi nào ta cũng nhớ chính xác thuật ngữ mà ta muốn sử dụng khi đang biên soạn văn bản. Những lúc đó, ta muốn tìm hiểu các cách thức mà thuật ngữ cụ thể được sử dụng như thế nào cũng như tìm hiểu những thuật ngữ tương tự. Để làm được việc đó, ta có thể buộc phải lật từng trang để lần tìm trong suốt cuốn từ điển. ***Từ điển điện tử -*** khác với một cuốn từ điển bằng giấy thông thường, cần phải giúp ta thực hiện những công việc như thế này một cách tức thời. Tiện ích Trợ lý ngôn ngữ có trong phần mềm giúp chúng ta tìm hiểu nhanh chóng thông tin đa dạng về từ vựng trong những văn cảnh khác nhau cũng như những minh họa thực tế trong văn học và đời sống...

Nguồn thông tin tra cứu gồm một bộ từ điển Anh-Việt/Việt-Anh, từ điển đồng nghĩa và phản nghĩa tiếng Anh (*Thesaurus*) và hai bộ từ điển tiếng Anh lớn hiện nay là *Oxford Advanced English Encyclopedic Dictionary* (Anh-Anh) và *Webster's Dictionary* (Anh-Mỹ)[1].



Hình 23 : Trợ lý ngôn ngữ

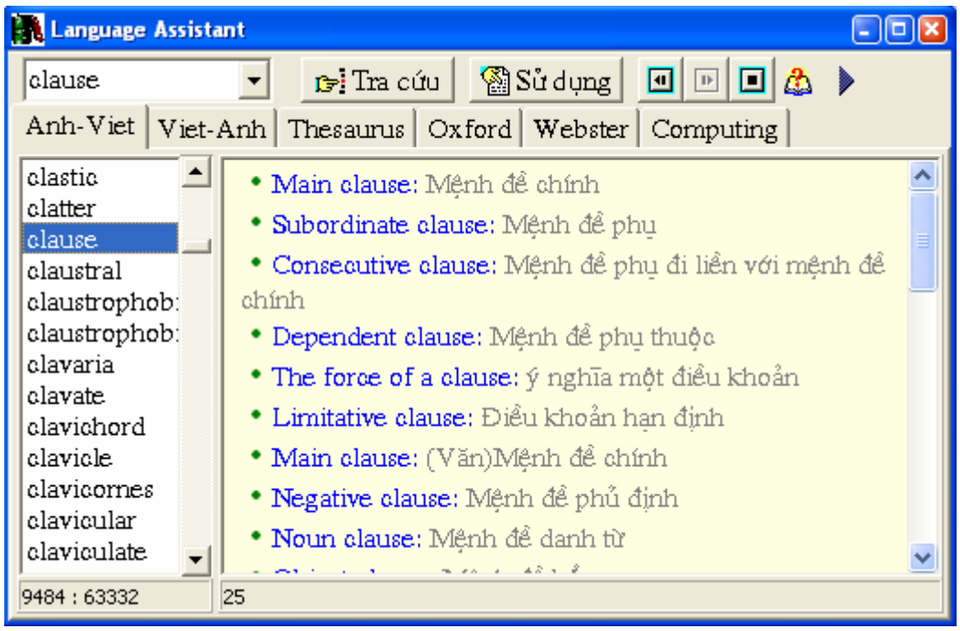
Khác với những từ điển điện tử hiện có, Từ điển Anh-Việt và Việt-Anh được tổ chức trên cơ sở một ***cuốn*** duy nhất nghĩa là chỉ có một bản từ

[1] Trong phần mềm có bổ sung một từ điển máy tính (tiếng Anh) để hỗ trợ những người sử dụng máy tính

điển với các chức năng tra cứu xuôi, ngược, minh họa sử dụng. Nguyên lý từ điển ***hai trong một*** này chỉ có thể thực hiện được dưới dạng sách điện tử; nó cho phép chỉ cần có duy nhất một tập tin dữ liệu cho cả hai từ điển. Điểm lợi của phương pháp là ta chỉ cần biên soạn một cuốn từ điển: trong khi ta soạn thảo hay hiệu chỉnh một mục từ, chẳng hạn Anh-Việt, hiệu ứng sẽ có ngay cho mục từ Việt-Anh tương ứng.

Các bộ từ điển tiếng Anh có trong phần mềm cũng cho phép tra cứu linh hoạt với việc xem duyệt tất cả các ví dụ sử dụng một từ nào đó. Từ điển sẽ đưa ra các tình huống ngữ pháp và thành ngữ hoặc các trích dẫn của những nhà văn kinh điển (Anh, Mỹ...). Những minh họa đó có thể nằm rải rác trong toàn bộ từ điển mà bình thường ta không thể dễ dàng thu gom lại được.

Đây là một cố gắng đầu tiên để thực hiện một bộ *Trợ lý ngôn ngữ điện tử* (không chỉ đơn thuần mô phỏng việc *tra cứu từ điển trên trang sách)*. Sự ra đời của các từ điển này nảy sinh từ nhu cầu nội bộ: Chúng tôi đã rất mất thời gian vì phải giở sách từ điển quá nhiều khi cập nhật cơ sở tri thức cho chương trình dịch. Từ khi sử dụng Trợ lý ngôn ngữ, tốc độ cập nhật và chất lượng của cơ sở tri thức đã được cải thiện nhanh chóng rõ rệt; và chính trình độ tiếng Anh của chúng tôi cũng được nâng cao đáng kể.



Hình 24 : Tìm kiếm toàn văn bản trong Trợ lý ngôn ngữ

Tuy nhiên, do phải đưa vào một khối lượng khổng lồ thông tin về ngôn ngữ, nên không tránh khỏi còn có nhiều thiếu sót trong bộ trợ lý ngôn ngữ này. Những ý kiến phản hồi của người sử dụng sẽ giúp cho sản phẩm có chất lượng ngày càng cao hơn.

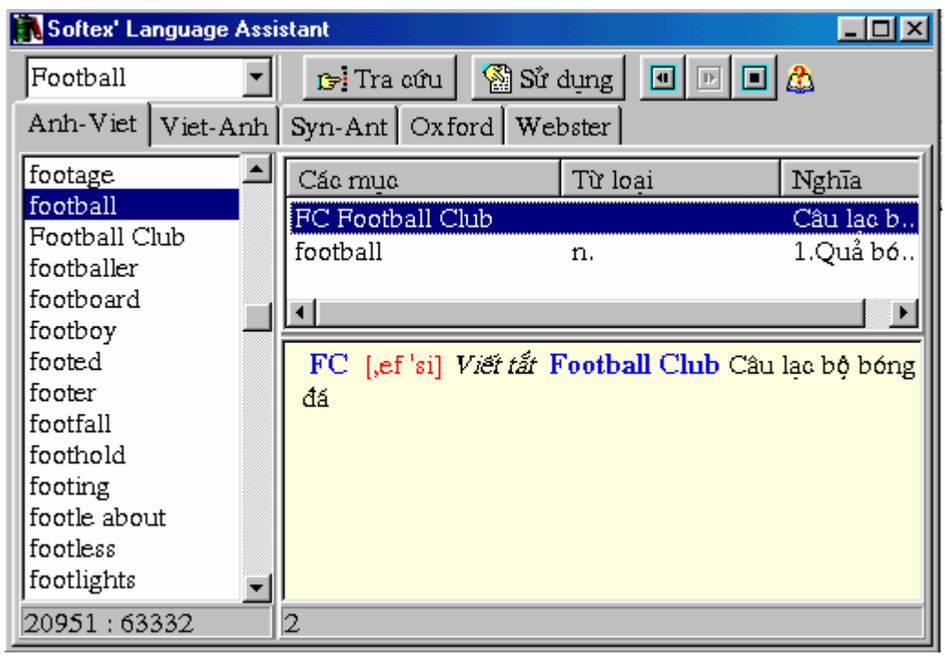
IV.2.2. CÁCH SỬ DỤNG TRỢ LÝ NGÔN NGỮ

Sau khi mở trợ lý ngôn ngữ ta thấy xuất hiện giao diện như sau:

Ví dụ: Như tra từ "*clause*" nhấn vào nút **Tra cứu** ta có được nghĩa của từ như hình trên. Khi nhấn vào nút **Sử dụng** thì ta có thể xem được cách sử dụng của từ đang tra trong các câu tiếng Anh như thế nào, và có giải thích nghĩa tiếng Việt tương ứng như minh hoạ Hình 19.

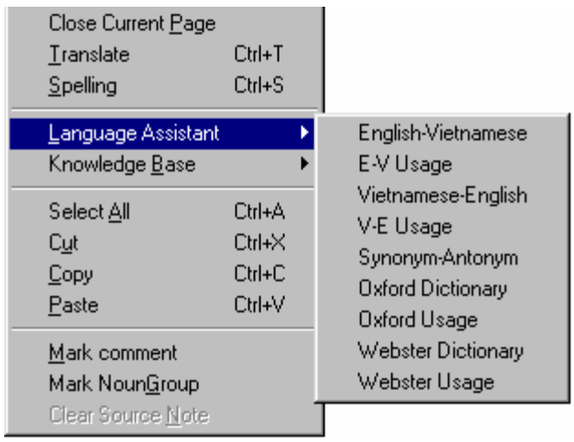
Để tra một từ trong của sổ giải nghĩa ta không cần phải gõ lại vào ô nhập từ mà chỉ cần kích đúp chuột lên từ đó trong cửa sổ giải nghĩa. Nguyên lý hoạt động của nó là, khi bôi đen một cụm từ nào trong cửa sổ giải nghĩa, nó lập tức được sao chép **copy** lên ô nhập từ và được tra nếu ta kích chuột thêm một lần nữa.

Ví dụ: Từ "Football" đã có trong cửa sổ giải nghĩa bên trên ta chỉ cần kích đúp chuột lên từ này ta sẽ có nghĩa của từ này như sau:



Hình 25 : Tra từ trên cửa sổ hiển thị của Trợ lý ngôn ngữ

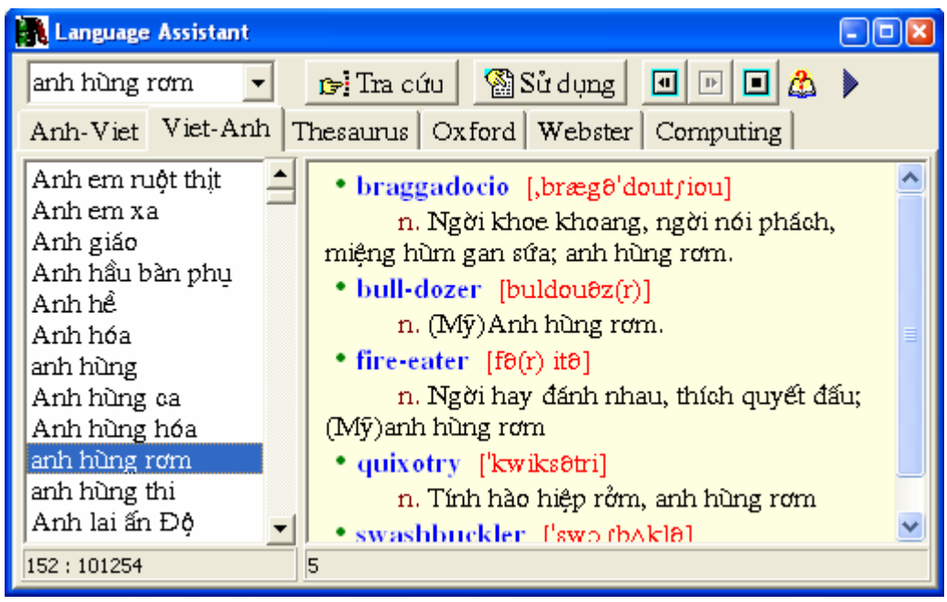
Để tra một từ từ chương trình dịch ta chọn loại từ điển cần tra từ menu thả xuống:



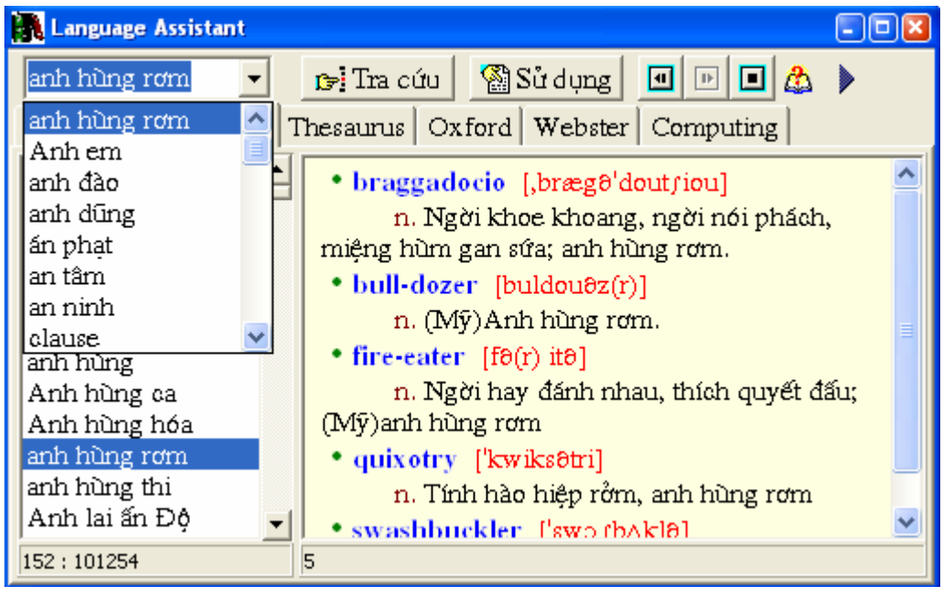
Hình 26 : Tra từ trên thực đơn nổi

Ngoài ra, còn có các từ điển đồng nghĩa, phản nghĩa: "Synonym-Antonym". từ điển Anh – Anh: "Oxford", từ điển Anh – Mỹ "Webster", tiện cho người sử tra cứu và sử dụng theo từng yêu cầu của mình.

Chúng tôi giải thích thêm về phần tra từ điển Việt-Anh. Từ điển Việt – Anh của chúng tôi ngoài chức năng tra nghĩa, nó còn là từ điển trợ giúp cho người dùng trong biên soạn văn phạm bằng tiếng Anh. Đang có một cụm từ tiếng Việt, muốn có một cụm từ tiếng Anh ngắn gọn diễn đạt đủ ý nghĩa của cụm từ tiếng Việt. Hãy tra ngay từ điển Việt – Anh người sử dụng sẽ có một từ tiếng Anh như ý muốn.



Hình 27 : Tra cứu Cụm từ

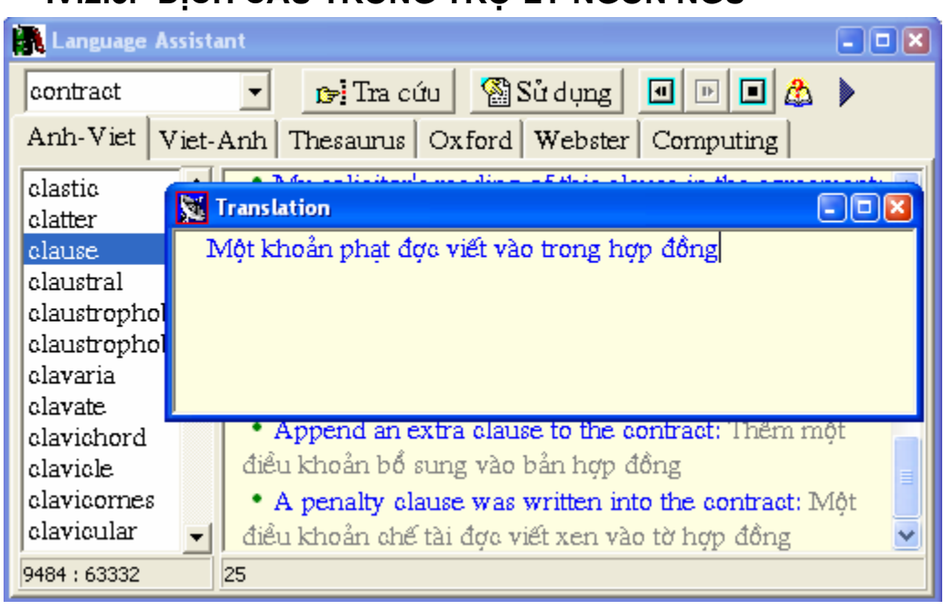


Hình 28 : Chọn từ đã tra cứu

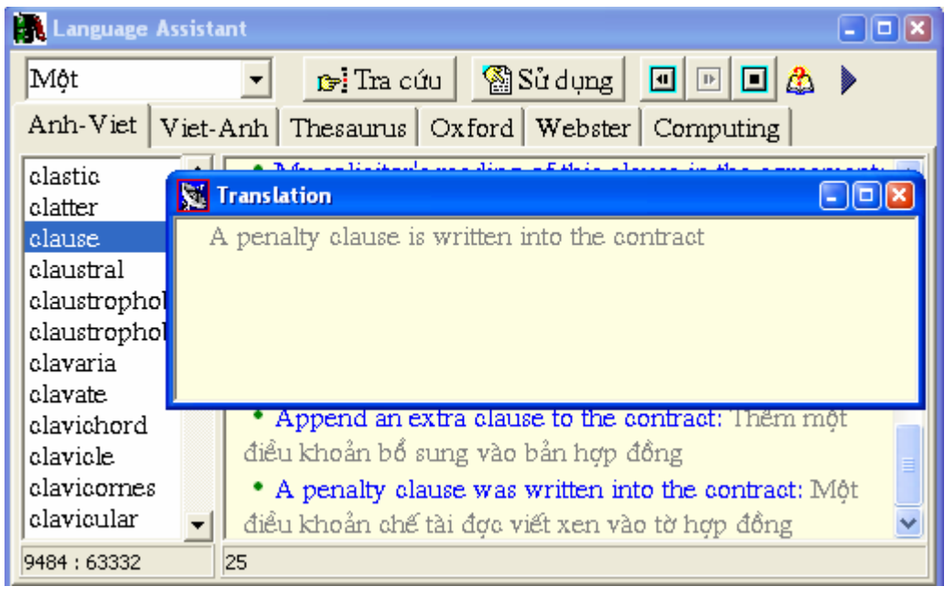
Ta nhận thấy trong tiếng Anh có nhiều từ diễn đạt được nghĩa của cụm từ này (Hình 19). Vì vậy, khi soạn thảo văn bản tiếng Anh, từ điển Việt – Anh là trợ lý đắc lực.

Để xem lại các từ đã tra, chỉ cần kích chuột vào biểu tượng cạnh ô nhập từ, một danh sách thả xuống sẽ xuất hiện ngay lập tức cho phép người sử dụng chọn mục từ muốn xem lại.

IV.2.3. DỊCH CÂU TRONG TRỢ LÝ NGÔN NGỮ



Hình 29 : Dịch Anh-Việt từ Trợ lý ngôn ngữ

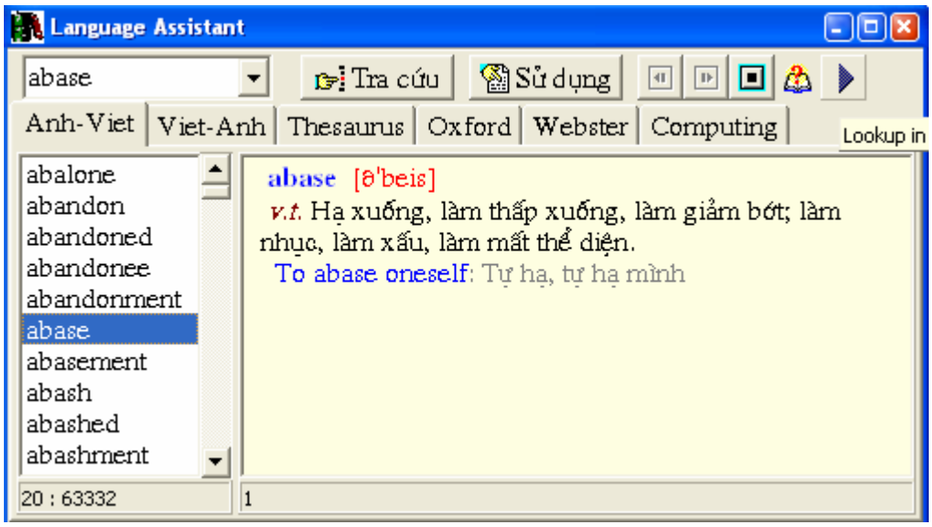


Hình 30 : Dịch Việt-Anh từ Trợ lý ngôn ngữ

Có thể dịch bất kỳ câu tiếng Anh nào trong cơ sở tri thức bằng cách bôi đen, bấm phím phải chuột và chọn **Translate**. Cửa sổ giải nghĩa sẽ hiện nghĩa của câu cần dịch theo minh hoạ dưới đây.

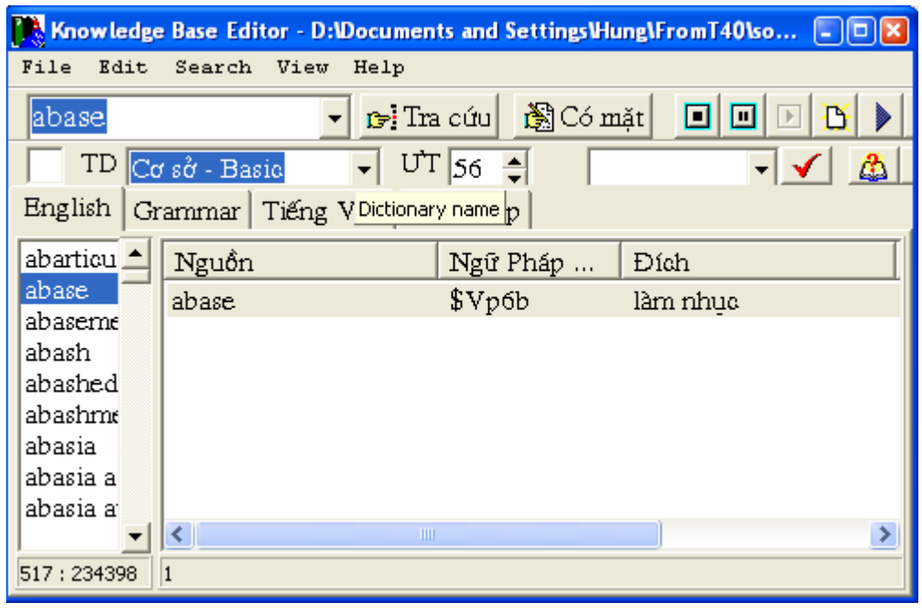
Khi chọn văn bản tiếng Anh thì ứng dụng sẽ dịch sang tiếng Việt, còn khi chọn văn bản tiếng Việt thì ứng dụng sẽ tự động dịch sang tiếng Anh.

IV.2.4. TRA CỨU CHÉO

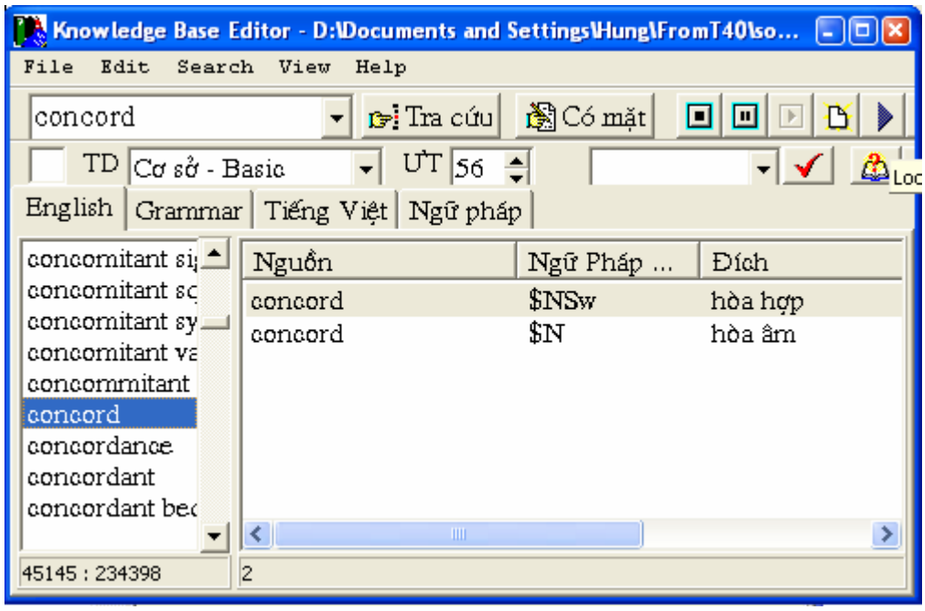


Hình 31 : Tra cứu chéo từ Trợ lý ngôn ngữ

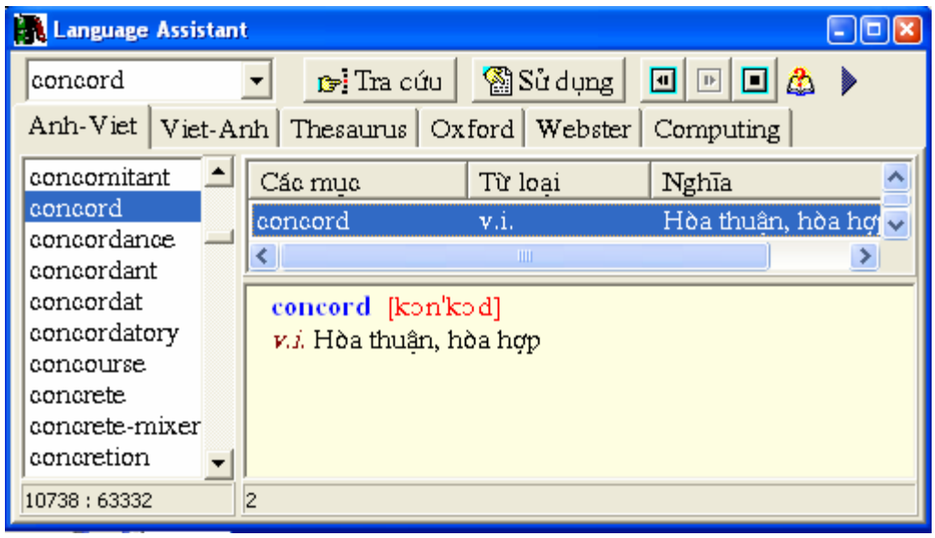
Từ Hệ soạn thảo cơ sở tri thức có thể tra cứu sang Trợ lý ngôn ngữ cũng như từ Trợ lý ngôn ngữ có thể tra cứu sang Hệ soạn thảo cơ sở tri thức chỉ bằng một thao tác nhấn chuột (nút ). Việc chuyển đổi qua lại giữa Cơ sở tri thức và trợ lý ngôn ngữ rất hữu ích cho thao tác cập nhật cơ sở tri thức cũng như cho việc học ngôn ngữ.



Hình 32 : Kết quả Tra cứu chéo từ Trợ lý ngôn ngữ



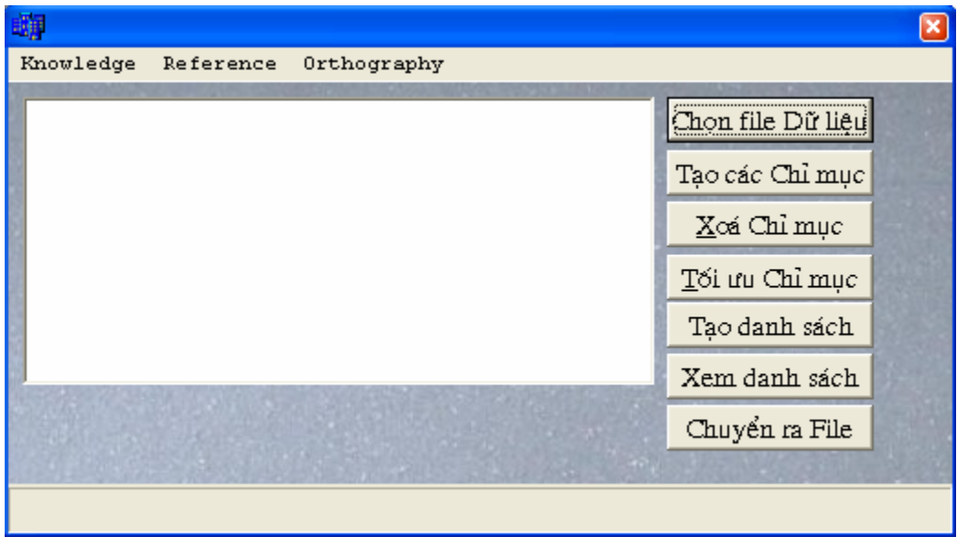
Hình 33 : Tra cứu chéo từ Hệ soạn thảo cơ sở tri thức



Hình 34 : Kết quả Tra cứu chéo từ Hệ soạn thảo cơ sở tri thức

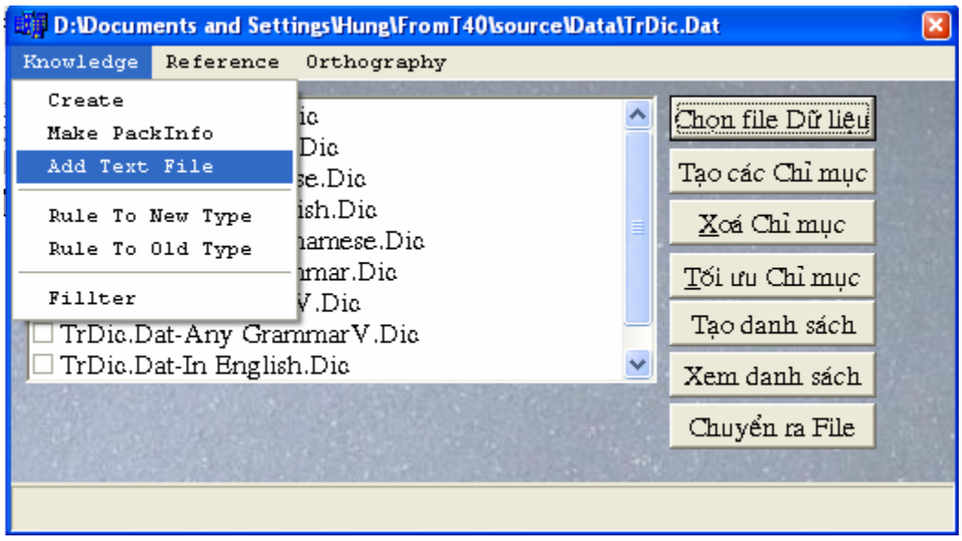
IV.3. BỔ SUNG FILE VĂN BẢN VÀO DỮ LIỆU

Trong EVTRAN có công cụ để bổ sung một hoặc nhiều tập tin văn bản vào *Cơ sở tri thức* và *Trợ lý ngôn ngữ*. Sử dụng tính năng này, ta có thể cập nhật đồng thời nhiều mục từ điển hoặc quy tắc văn phạm, hay thập chí toàn bộ cơ sở tri thức dịch thay vì phải vào từng quy tắc hay từng mục từ điển.



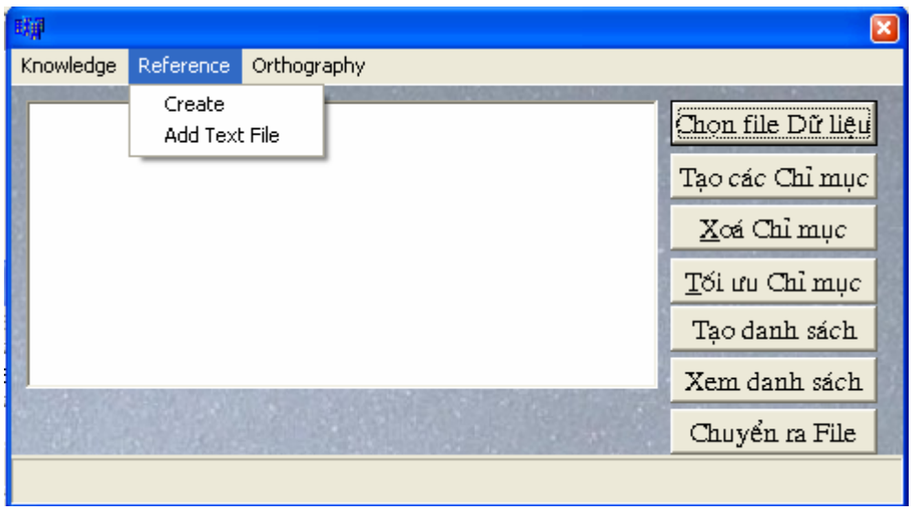
Hình 35. Cửa sổ Bổ sung một file văn bản

Để mở cửa sổ công cụ này, từ cửa sổ chính của EVTRAN, vào **Tools**, sau đó vào **Dictionary Administrator**... ta sẽ thấy một cửa sổ như trên Hình 22.



Hình 36. Cửa sổ Bổ sung một file văn bản vào cơ sở tri thức

Nếu muốn bổ sung file vào cơ sở tri thức ta chọn **Knowledge** rồi **Add Text File.**

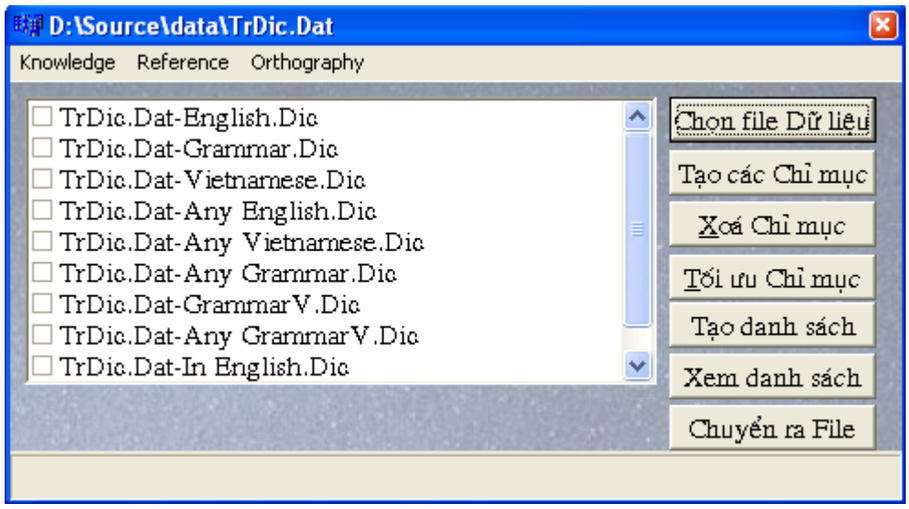


Hình 37. Cửa sổ Bổ sung một file văn bản vào Trợ lý ngôn ngữ

Nếu muốn bổ sung file vào cơ sở tri thức ta chọn **Reference** rồi **Add Text File.**

Trước khi cập nhật ta phải chọn file dữ liệu bằng cách nhấn **"chọn file dữ liệu".** Lúc đó việc cập nhật file mới sẽ được thực hiện cho từ điển được chọn.

Sau đó chọn file cần cập nhật. Khi đó dữ liệu được tự động cập nhật theo khuôn dạng quy định.



Hình 38. Tạo chỉ mục cho Trợ lý ngôn ngữ

Sau khi cập nhật dữ liệu, cần phải tạo các file chỉ mục để phục vụ việc tìm kiếm nhanh thông tin từ điển. Các chỉ mục bao gồm: chỉ mục tra từ tiếng Anh, tra từ tiếng Việt, tra từ theo khái niệm văn phạm tiếng Anh, tra từ theo khái niệm văn phạm tiếng Việt, tìm kiếm văn bản trong từng mục,...

V. CÁC KỸ THUẬT

V.1. TỪ VỰNG.....V-2
V.1.1. HỆ THỐNG TỪ LOẠI TIẾNG VIỆT.....V-2
V.1.2. BỘ QUI TẮC TỪ VỰNG TIẾNG VIỆT. V-5
V.1.3. QUI TẮC PHÂN TÍCH TIẾNG VIỆT.....V-8
V.1.4. QUI TẮC TỔNG HỢP TỪ VỰNG TIẾNG ANH V-8

V.2. CÁC KỸ THUẬT TĂNG TỐC ĐỘ PHÂN TÍCH.....V-10
V.2.1. ỨNG DỤNG VĂN PHẠM CẢM NGỮ ĐOẠN. V-10
V.2.2. CÁC YÊU CẦU ĐỐI VỚI GIẢI THUẬT PHÂN TÍCH V-13
V.2.3. PHÂN TÍCH QUAY LUI, SÂU DẦN. V-14
V.2.4. HẠN ĐỊNH ĐỘ SÂU CÂY CÚ PHÁP.....V-15
V.2.5. KẾT QUẢ V-15

Phần này giới thiệu các kỹ thuật và giải thuật chính áp dụng trong xây dựng hệ dịch máy Việt-Anh.

V.1. TỪ VỰNG

V.1.1. HỆ THỐNG TỪ LOẠI TIẾNG VIỆT

Do chưa tìm thấy một hệ thống văn phạm sinh tiếng Việt từ bất kỳ nguồn nào nên việc xây dựng hệ thống từ loại hình thức cho tiếng Việt phải thực hiện từ đầu. Vì ngôn ngữ tiếng Việt hết sức phức tạp nên nhóm đề tài chỉ giới hạn công việc của mình trong phạm vi văn viết tiếng Việt. Và ngay cả trong giới hạn này chúng tôi cũng chỉ xây dựng hệ thống từ loại tiếng Việt cho một lớp đủ lớn từ vựng tiếng Việt.

Từ loại đông đảo nhất trong tiếng Việt là danh từ. Nguyên lý phân loại đặt nền tảng trên hệ thống khái niệm của tiếng Việt (tương đối phổ quát cho một số ngôn ngữ Đông Á). Cách phân loại danh từ dựa trên sự phân lớp theo tập hợp. Hệ thống kiểu của từ loại tạo thành một nửa giàn (Semi-Lattice). Cách phân loại này giúp giảm thiểu độ phức tạp của cây phân tích khi tăng số khái niệm dẫn xuất miêu tả ngôn ngữ (số lượng nonterminals). Mô hình Chomsky không có sự phân biệt gì các biến trung gian với nhau, cũng như giữa các quy tắc văn phạm. Vì vậy rất khó để mô tả những quy luật văn phạm chung cho từng nhóm biến cụ thể.

Trong các tài liệu về ngữ pháp ta thường thấy có 8 nhóm từ chính, trong đó mỗi nhóm có thể được chia nhỏ tiếp tục thành những nhóm con :

- Danh từ: danh từ chung, riêng, danh từ đếm được, không đếm được, danh từ khối, đơn vị...
- Tính từ: tính từ chỉ chất liệu, tính từ chỉ trạng thái,...
- Động từ: ngoại động từ, nội động từ, trợ động từ...
- Đại từ: đại từ nhân xưng, đại từ sở hữu...

Tuy nhiên, tiếng Việt có những tính chất đặc trưng khác biệt với những ngôn ngữ Ấn Âu làm cho việc phân loại từ vựng thông thường có những điều không ổn:

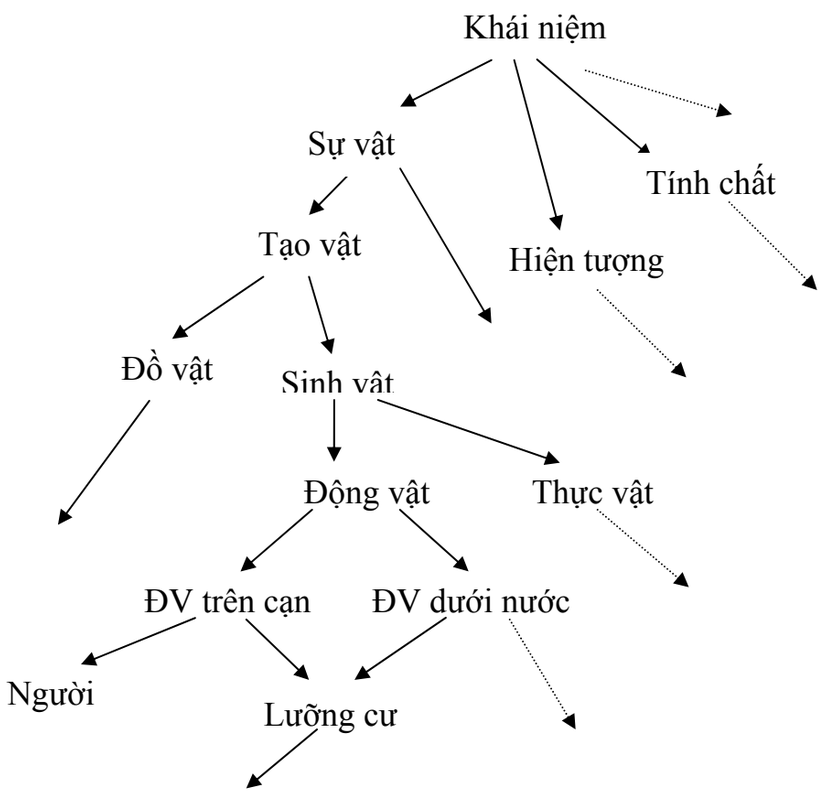
- Danh từ chỉ sự vật (đồ vật, động vật, ...) trong tiếng Việt thường được sử dụng một cách khái quát (để chỉ loài, hoặc chỉ số nhiều, hoặc được sử dụng như tính từ)

- Có một loạt danh từ được sử dụng để chỉ những sự vật cụ thể và đứng trước danh từ chỉ loài (*con, cái, chiếc, tấm, cục, bầy, đàn, bộ, dòng, bức, cặp, cú, bó, dàn, giọt, tấm.....*)
- Tính từ trong tiếng Việt thường được sử dụng như vị ngữ
- Danh từ trong tiếng Việt thường được sử dụng như đại từ (*ông, bà, đồng chí,..*)

Đó là một số vấn đề gây trở ngại cho việc xử lý tự động ngôn ngữ.

Hiện nay chưa thấy có tài liệu nào đề cập việc phân tích từ loại tiếng Việt một cách đầy đủ, chi tiết.

Như một cố gắng bước đầu, chúng tôi hiện đang xây dựng mô hình ngữ nghĩa chung cho Hệ thống loại từ tiếng Việt trên cơ sở cách thức mà người Việt nam sử dụng chúng. Hệ phân cấp khái niệm bao gồm trên 1.500 nhóm được tổ chức thành 14 lớp. Các loại từ tiếng Việt được sắp xếp theo phân loại ngữ nghĩa của chúng.



Hình 1. Ví dụ dàn khái niệm

Cách sắp xếp này là một cố gắng tạm thời nên có thể còn nhiều khiếm khuyết và sẽ còn được thay đổi trong tương lai.

Hệ phân loại khái niệm là một giàn đại số (*lattice*) . Sự phân cấp của các đối tượng dựa trên tính khái quát của các khái niệm. Chẳng hạn trong khái niệm sự vật có các khái niệm thấp cấp hơn là đồ vật, sinh vật. Trong khái niệm sinh vật có động vật, thực vật... (Hình 1). Phân loại theo khái niệm khá tương đồng với việc sử dụng danh từ tiếng Việt đặc trưng bởi những danh từ đếm được chỉ đơn vị (như đã nói ở trên).

Phân loại tính từ được thực hiện dựa theo tính chất ngữ pháp của chúng. Theo các nhà ngôn ngữ học, trong tiếng Việt, tính từ chiếm một vị trí quan trọng vì rất khó phân biệt giữa tính từ và vị ngữ.

Ví dụ 1: Trong hai câu sau:

- *Đó là một ngôi nhà đẹp*
- *Ngôi nhà này đẹp*

Từ *đẹp* ở câu thứ nhất là tính từ còn từ *đẹp* ở câu thứ hai lại được xem là *vị ngữ*?. Cũng vì vậy, người ta đưa thêm vào khái niệm ***vị từ.*** Coi từ *đẹp* trong cả hai ví dụ trên đều thuộc một loại từ duy nhất : vị từ. Khi đó trong tiếng Việt sẽ không tồn tại tính từ. Vị từ được sử dụng như động từ hoặc như tính từ ở tiếng nước ngoài.

Quan điểm đáng tranh cãi này dường như dựa trên việc từ *đẹp* được dịch sang tiếng Anh thành hai cách khác nhau : "*beautiful"* và "*to be beautiful"* tùy theo việc trong câu có động từ hay không; nếu chưa có động từ thì tính từ được xem như có vai trò của động từ. Từ đây có thể coi các tính từ tiếng Việt vừa là động từ vừa là tính từ. Mà như thế thì luôn luôn có các cặp từ giống nhau nhưng thuộc hai loại từ khác nhau. Vì thế nên người ta nghĩ ra loại "*vị từ*"

Có thể phân tích các tình huống sử dụng tính từ và đi đến một cách kiến giải về tính từ cũng như cách sử dụng tính từ trong câu tiếng Việt. Phần lớn các ví dụ thường gặp trong câu tiếng Việt đều có thể giải thích (và phân tích câu) trên cơ sở phương pháp được đề cập. Theo cách phân loại này thì trong cả hai ví dụ trên, chữ đẹp đều có thể coi là tính từ. Và như vậy, nếu cách giải thích này đứng vững thì tính chất của tính từ trong tiếng Việt là (gần như) tương đồng với cách mà chúng được hiểu trong các ngôn ngữ Ấn Âu.

Điểm trung tâm của phương pháp này là ở giả thiết : "*Người ta thường có xu hướng loại bỏ một phần tử nào đó trong câu (vốn được xem là chuẩn mực) nếu không ảnh hưởng đến việc hiểu nghĩa của câu*". Nói cách khác, một số từ bổ trợ thường được bỏ hẳn trong cách đặt câu của tiếng Việt

(như các giới từ *ở, trong, đối với,..*, động từ *thì, là, có,...* liên từ *mà, rằng, sao cho,...*)

Ví dụ 2:

- Ta viết "*Ngôi nhà này đẹp"* chứ không viết "*Ngôi nhà này thì đẹp"*
- Ta viết "*Nước ta nhiều núi đồi"* chứ không viết "*Ở nước ta có nhiều núi đồi"*
- Ta viết "*Tôi đau chân"* chứ không viết "*Tôi bị đau ở chân"*
- ...

Với giả định như thế thì từ *đẹp* của tiếng Việt trong cả hai câu của Ví dụ 1 ở trên đều có thể coi là tính từ, chẳng khác gì các ngôn ngữ khác.

Đại từ là loại từ phức tạp trong tiếng Việt. Hầu như mọi danh từ chỉ người đều có thể sử dụng như đại từ : "*anh, chị, ông, bà, cô, dì, chú, bác, bạn*, ...". Có những đại từ có thể dùng thay cho hầu hết các loại từ khác, danh từ, tính từ, động từ, và cả đại từ, như những đại từ "*ấy, chi, gì*" mà không thấy trong tiếng Anh hay một số ngôn ngữ khác.

Trong hệ phân cấp từ vựng, chúng tôi đã chia nhóm đại từ ra thành các loại danh đại từ, tính đại từ, động đại từ,... và nhận thấy rằng cách phân loại này có thể cho phép phân tích hình thức lớp đại từ tiếng Việt một cách tương đối chi tiết.

Trạng từ được coi như một lớp từ loại không được định nghĩa rõ ràng: nó có thể bổ nghĩa cho tính từ, động từ, danh ngữ, câu, hay trạng từ khác...

Để dễ hình thức hóa, chúng tôi phân loại trạng từ theo loại từ mà nó có thể bổ nghĩa. Như vậy, một trạng từ mà có thể đồng thời bổ nghĩa cho cả tính từ và động từ thì được xem như đó là hai từ khác nhau (chẳng hạn các trạng từ "*rất*, *cũng*..." trong "rất đẹp", "cũng cao" và "rất muốn nói...", "cũng đi"..., nhưng "*hơi, quá..."* thì chỉ bổ nghĩa cho tính từ,...). Việc phân biệt từ loại theo các khía cạnh sử dụng chúng cho phép mô tả văn phạm một cách nhất quán.

V.1.2. BỘ QUI TẮC TỪ VỰNG TIẾNG VIỆT.

Về bản chất, tiếng Việt không cần hệ quy tắc từ vựng. Bộ quy tắc từ vựng tiếng Việt chủ yếu để phục vụ dịch máy. Chúng tôi đã chọn cách tiếp cận xuất phát từ hệ thống quy tắc từ vựng tiếng Anh để xây dựng quy tắc từ vựng tiếng Việt. Hệ quy tắc này được soạn thảo tuân thủ tính nghịch đảo toàn phần (phục vụ dịch hai chiều Anh-Việt và Việt-Anh). Nhiều vấn đề từ vựng trong tiếng Anh, thì ở tiếng Việt có thể coi là vấn đề cú pháp. Trong thực tế, mỗi quy tắc từ vựng tiếng Anh có thể có sự tương đương với một

quy tắc văn phạm (quy tắc tạo từ – từ ghép) tiếng Việt hoặc tương đương với một dạng cú pháp nhất định (bổ sung một *đơn vị văn phạm* (thường được gọi là *hư từ*) vào một *đơn vị từ vựng* (thường được gọi là *thực từ*)

Tiếng Việt là ngôn ngữ đơn lập. Trong tiếng Việt, một khái niệm với nghĩa xác định, nguyên thủy không phải luôn luôn có một *từ* (từ rời, bao gồm một tiếng duy nhất) dành riêng cho nó. Theo các nguyên tắc tạo một *tiếng* của tiếng Việt (phụ âm đầu, âm phụ, âm chính, âm cuối và dấu) thì có tối đa không tới 16.000 tiếng (có nghĩa hoặc vô nghĩa), còn trong thực tế sử dụng không tới 7.000 tiếng. Vốn từ vựng thông thường của một ngôn ngữ dao động trong khoảng 30.000 – 60.000 tùy thuộc ngôn ngữ. Từ đây suy ra vốn từ vựng (từ gốc) trong tiếng Việt chủ yếu là những từ bao gồm hai âm tiết trở lên.

Trong tiếng Việt không có từ dẫn xuất theo cách thay đổi đầu hoặc đuôi của một từ cho trước. Vì vậy các từ gốc hoặc từ dẫn xuất được thành lập theo cách :

- Tổ hợp từ với một nghĩa xác định được gán cho một nghĩa bền vững
- Vài tiếng được ghép với nhau để có được một từ với một nghĩa cụ thể (không căn cứ vào nghĩa của mỗi thành phần)
- Ghép một từ với một hoặc một vài từ phụ để hình thành một từ với nghĩa khác.

Trong trường hợp thứ nhất ta nhận được một từ ghép với ngữ nghĩa ổn định.

Ví dụ:

- Bến đò, xe đạp, bất công ...

Trong trường hợp thứ hai ta nhận được một từ ghép với ngữ nghĩa khác với các từ thành phần. Chúng ta sẽ gọi trường hợp này là sự tổng hợp từ mới.

Ví dụ:

Cộng hòa, sư tử, đồng hồ ...

Trong trường hợp thứ ba, từ mới kế thừa ngữ nghĩa của từ gốc nhưng đóng vai trò ngữ pháp khác với từ gốc.

Ví dụ:

*sự **thành lập**, có thể **thực hiện** được, tính hay **thay đổi**, một cách **chính thức**,...*

Bằng cách đối chiếu sự hình thành từ nghữ trong tiếng Anh và tiếng Việt ta có thể xây dựng được hệ thống từ tiếng Việt có ý nghĩa từ vựng tương đương với những từ đơn lẻ hoặc tổ hợp từ bền vững trong tiếng Anh.

Mặc dù ta có thể thấy sự tương quan rõ ràng giữa cách thức tạo từ ghép tiếng Việt với các cơ chế hình thành từ hoặc từ dẫn xuất trong tiếng Anh, nhưng hai ngôn ngữ không hoàn toàn tương đồng. Cụm từ tiếng Việt có thể diễn giải theo một vài cách bằng tiếng Anh, chẳng hạn cụm từ "*có thể chấp nhận được*" trong tiếng Anh được diễn đạt theo hai cách : "*acceptable"* hoặc "*can be accepted*". Trong khi đó, chẳng hạn "*tính có thể chấp nhận được*" trong tiếng Anh chỉ là "*acceptability"*. Hệ dịch máy Việt Anh cần phải cung cấp những lựa chọn để tùy theo mỗi tình huống ngữ cảnh, cụm từ tiếng Việt có thể được diễn giải như một từ hoặc như một cụm từ tiếng Anh.

Cụm từ tiếng Việt được sử dụng để phản ánh hiện tượng từ dẫn xuất (*derivation*) cũng như hiện tượng biến đổi từ (*inflection*) của tiếng Anh. Tuy nhiên, do văn phạm quy định, một hiện tượng từ dẫn xuất hoặc biến đổi từ trong tiếng Anh có thể được diễn đạt theo những cách khác nhau. Chẳng hạn Ing-Clause trong tiếng Anh có thể tương đương với việc thêm từ phụ (không hiểu vì sao thường được gọi là hư từ) "*việc*" hoặc "*mà*" vào một động ngữ trong tiếng Việt tùy theo văn cảnh sử dụng cụm từ đó.

Từ ghép tiếng Anh có thể tương ứng với cụm từ tiếng Việt trên cơ sở ghép những yếu tố rời rạc (ví dụ "*có thể phân hủy được*" - *decomposable*) hay tổ hợp từ bền vững (thường là từ gốc Hán, ví dụ "*khả chuyển*" - *transferable*) hay một từ ổn định (ví dụ "đáng kể" – *considerable*). Để giải quyết nhập nhằng trong tạo từ ghép tiếng Việt có thể xây dựng văn phạm cụm từ để phân tích sơ bộ các cụm từ như bước tiền xử lý cho bộ phân tích văn phạm.

Do tính chất các cụm từ (thành ngữ) thường có độ dài khống chế nên ta có thể áp dụng một mô hình văn phạm phức tạp để mô tả mà không lo lắng nhiều về độ phức tạp tính toán của các giải thuật phân tích câu.

Một quan sát thực nghiệm : *Trong thực hành ngôn ngữ, độ phức tạp của văn phạm nghịch biến với độ dài của câu.* Kết luận này mang tính chất tiên nghiệm nhưng rất giống với cách chúng ta tư duy về ngôn ngữ. Khi đặt câu, chúng ta thường có xu hướng chọn những cấu trúc phức tạp, hàm ẩn trong những đoạn văn ngắn (cụm từ) và có xu hướng diễn đạt rõ ràng với những cấu trúc câu đơn giản cho những câu dài. Vì vậy, khi phân tích ta có thể sử dụng những mô hình văn phạm tổng quát hơn, phức tạp hơn trong những phạm vi hẹp và văn phạm đơn giản (chẳng hạn phi ngữ cảnh) đối với những câu dài. Rộng hơn, trong phạm vi đoạn văn hay toàn bài văn, mối quan hệ cú pháp (giữa các câu) lại càng tuân thủ những quy tắc nghiêm ngặt

và đơn giản hơn, cho nên có thể áp dụng mô hình văn phạm còn chặt chẽ hơn nữa (và vì vậy, hiệu quả hơn). Quan sát này lý giải việc dịch những câu dài tỏ ra không phức tạp một cách bùng nổ so với dịch các câu ngắn – nghĩa là có thể tìm được giải pháp phân tích sao cho độ phức tạp tính toán tiến dần đến tuyến tính khi có đủ tri thức ngôn ngữ.

V.1.3. QUI TẮC PHÂN TÍCH TIẾNG VIỆT.

Như đã nói trên, từ tiếng Việt được tạo thành thông qua việc ghép các *từ* khác với nhau. Mỗi câu tiếng Việt có thể ngắt từ theo một số cách khác nhau. Sau đây là vài ví dụ:

- Cụm từ "*trọng tài trọng tài*" có thể ngắt là "*trọng | tài | trọng tài*" hoặc là "*trọng tài | trọng | tài*"
- Mệnh đề "*Trưởng phòng phòng phòng cháy chữa cháy*" có thể ngắt thành một danh ngữ là "*Trưởng phòng (của) phòng (có tên là)| phòng cháy chữa cháy*" hoặc thành một câu là "*Trưởng phòng (của) phòng (có tên là)| phòng cháy (đang)|chữa cháy*". . .

Bộ phân tích từ vựng có thể thực hiện việc *phân tích nông* (*shallow analysis* để phân tích các cụm từ) như đa số các hệ dịch máy vẫn làm để đưa ra mộc phương án ngắt từ cho mỗi câu tiếng Việt.

Có một cách tiếp cận khác là tổ chức tất cả các kiểu ngắt từ cho toàn bộ mệnh đề cần phân tích thành một dàn (*với nút bé nhất là nút nằm trước từ đầu tiên và với nút lớn nhất là nút nằm sau từ cuối cùng*) mà chưa bận tâm đến việc ngắt từ theo một cách nào đấy có làm cho câu trở nên vô nghĩa hoặc sai văn phạm hay không. Sau đó bộ phân tích cú pháp không phải thực hiện trên một chuỗi kí hiệu mà là trên dàn các *từ*.

Cách tiếp cận này có tính khái quát cao: có thể xây dựng một bộ phân tích từ vựng chung cho mọi ngôn ngữ – những ngôn ngữ biến hình như tiếng Anh, tiếng Nga, tiếng Pháp hoặc những ngôn ngữ đơn lập như tiếng Việt, tiếng Hoa,... đều có mô hình xử lý giống nhau.

V.1.4. QUI TẮC TỔNG HỢP TỪ VỰNG TIẾNG ANH

Bộ quy tắc tổng hợp tiếng Anh là nghịch đảo của bộ quy tắc phân tích từ vựng tiếng Anh. Nội dung mục này được thực hiện tích hợp chặt chẽ với nội dung 2.

Từ vựng Tiếng Anh bao gồm các từ gốc (root words), từ biến đổi (inflection) từ dẫn xuất (derivation) và từ ghép (compound words).

Để tổng hợp từ vựng tiếng Anh, có hai cách tiếp cận:

- Sử dụng văn phạm chính quy và mạng chuyển trạng thái

- Sử dụng văn phạm bất kỳ và tích hợp vào bộ phân tích văn phạm như một bộ tiền xử lý vạn năng.

Với phương pháp thứ nhất, việc phân tích (và tổng hợp) từ vựng phải tách ra khỏi quá trình phân tích (hay tổng hợp) cú pháp.

Cách tiếp cận thứ hai làm phức tạp hóa bộ phân tích văn phạm nhưng cho phép xử lý thống nhất cho mọi ngôn ngữ. Độ phức tạp tính toán tùy thuộc vào giải thuật phân tích (và tổng hợp) được áp dụng.

Về mặt tính toán, không có gì khác biệt giữa các từ dẫn xuất (*derivation*) và biến đổi từ (*inflection*) trong tiếng Anh, sự khác nhau hình thức nằm ở vị trí của sự biến đổi từ: tiếp đầu (prefixes), tiếp đuôi (suffixes) hay tiếp giữa (infixes).

Quy tắc biến đổi từ vựng gồm có 5 thành phần:

- ***Từ_dẫn_xuất***
- ***Gốc_trái***
- tiếp_tố_dẫn_xuất
- tiếp_tố_gốc
- ***Gốc_phải***

Trong đó ***Gốc_trái***, ***Gốc_phải***, ***Từ_dẫn_xuất*** là những biến trung gian (khái niệm ngữ pháp) còn tiếp_tố_dẫn_xuất, tiếp_tố_gốc là một tổ hợp chữ hoặc là một từ.

Có các tình huống áp dụng quy tắc:

- Quy tắc cho tiếp đầu : vắng mặt ***Gốc_trái***.
- Quy tắc cho tiếp đuôi : vắng mặt ***Gốc_phải***.
- Quy tắc cho tiếp giữa : có đủ tất cả các thành phần
- Quy tắc cho những từ bất quy tắc :vắng mặt ***Gốc_phải,*** đồng thời tiếp_tố_gốc là một từ nguyên.[1]

Heuristics chung để xử lý cho cả ba loại tiếp tố là : *cực đại* tổng độ dài các tiếp_tố_dẫn_xuất, tiếp_tố_gốc của *cực tiểu* tổng số lần áp dụng quy tắc từ vựng. Heuristics đơn giản này tỏ ra hữu hiệu để giải quyết mọi nhập nhằng trong tổng hợp từ vựng tiếng Anh khi có những từ có thể tổng hợp theo nhiều luật khác nhau.

Để minh họa quy tắc từ vựng trong ứng dụng thực tiễn, ta xét một vài ví dụ.

[1] Khái niệm từ ở đây được hiểu không phải như một chuỗi ký tự (chữ cái) mà như một mục từ điển được định nghĩa (chẳng hạn trong tiếng Anh, *lay* là dạng quá khứ của *lie – nằm* chứ không phải của *lie – nói dối*)

Các quy tắc:

1. ***Plural→ Noun*** _s _
2. ***Plural→ Noun*** _es _
3. ***Plural→ Noun*** _es _e
4. ***Plural→ Noun*** _ies _y

Quy tắc 1. áp dụng cho mọi từ, nếu kết thúc là es thì ưu tiên áp dụng quy tắc 2. Khi từ gốc có đuôi là e thì áp dụng quy tắc 3 trước hết. Nếu quy tắc 4 áp dụng được thì đó là quy tắc được ưu tiên nhất.

V.2. CÁC KỸ THUẬT TĂNG TỐC ĐỘ PHÂN TÍCH

V.2.1. ỨNG DỤNG VĂN PHẠM CẢM NGỮ ĐOẠN.

Xây dựng và giới thiệu mô hình Văn phạm định biên, nêu những đặc tính hữu ích của nó trong xử lý ngôn ngữ tự nhiên. Nhóm thực hiện đề tài đã phát triển mô hình văn phạm định biên theo hướng mở rộng hiệu năng mô tả. Mô hình văn phạm cảm ngữ đoạn được phát triển và định hình trên cơ sở văn phạm định biên. Lớp văn phạm này tỏ ra hữu hiệu để mô tả được những tính chất đặc biệt của ngôn ngữ tự nhiên như: sự phụ thuộc giữa các phần tử trên khoảng cách, các mối liên hệ ngôn ngữ không lồng nhau, sự phụ thuộc giữa các tầng phân tích khác nhau.

Để đảm bảo hiệu năng tính toán khi phân tích văn phạm, giải thuật duyệt theo chiều rộng, phân tích sâu cục bộ đã được phát triển. Giải thuật này có thể áp dụng cho nhiều loại văn phạm, kể cả những loại văn phạm phức tạp trong phân cấp Chomsky. Đối với văn phạm cảm ngữ đoạn, giải thuật có một đặc tính quan trọng là : khi tri thức ngôn ngữ càng nhiều, tốc độ phân tích càng được cải thiện.

Nội dung nghiên cứu trong khuôn khổ đề tài đã được giới thiệu trong các báo cáo khoa học:

- Lê Khánh Hùng (2003) Văn phạm cảm ngữ đoạn, Báo cáo khoa học tại hội thảo quốc gia lần thứ sáu “Một số vấn đề chọn lọc của CNTT và TT”, Thái nguyên, 8-2003.
- Lê Khánh Hùng, Trần Cảnh (2003) Về một số hạn chế của mô hình văn phạm Chomsky, Tạp chí Bưu chính Viễn thông, , Chuyên san 10, 2003.
- Lê Khánh Hùng (2003) Một Phương pháp Dịch máy Liên ngữ. Kỷ yếu Hội thảo Khoa học Quốc gia lần thứ nhất về Nghiên cứu, Phát triển và Ứng dụng CNTT&TT, Hà nội, 2003.

- Lê Khánh Hùng (2003) Văn phạm phụ thuộc phạm vi, các tính chất và ứng dụng, Báo cáo khoa học tại hội nghị ICT 2003, Hà Nội, 03-2003.
- Lê Khánh Hùng (2002) Văn phạm định biên và một số tính chất, Tạp chí Bưu chính Viễn thông, Chuyên san 8, 11-2002.
- Báo cáo khoa học tại Hội thảo quốc gia về Nghiên cứu và Phát triển Khoa học cơ bản, Hà Nội, 10, 2003.

Bên cạnh những nghiên cứu trong khuôn khổ các nội dung đăng ký của đề tài, chúng tôi hiện đang tập trung thử nghiệm một công nghệ hoàn toàn mới trong dịch máy, đó là Dịch máy theo mô hình Liên ngữ. Chính những kết quả nghiên cứu về mô hình Văn phạm của đề tài đã mở ra triển vọng đi xa hơn trong việc phát triển một công nghệ dịch máy tiên tiến. Nếu những kết quả nghiên cứu là khả quan, chúng tôi sẽ có những đề nghị hiệu chỉnh nội dung nghiên cứu của đề tài để sớm có được một sản phẩm dịch máy với chất lượng cao hơn một bậc.

Đã xây dựng hệ phân cấp các khái niệm ngôn ngữ tự nhiên dựa trên mô hình toán học giàn đại số (*lattice*). Hệ phân cấp này cho phép mô tả được những đặc tính ngữ nghĩa và các cấu trúc ngôn ngữ chuyên biệt trên cơ sở xây dựng những giải thuật phân tích tiên nghiệm. Kết quả này sẽ là tiền đề để xây dựng hệ dịch không phụ thuộc cặp ngôn ngữ với chất lượng hứa hẹn cao hơn. Đây cũng là mô hình hình thức cho hệ ngữ vựng của ngôn ngữ tự nhiên cho phép tìm được những ứng dụng bên ngoài dịch máy (các hệ hiểu ngôn ngữ tự nhiên, kiểm tra và sửa chữa lỗi văn phạm và ngữ nghĩa tiếng Việt, tìm kiếm toàn văn theo nội dung, nhận dạng và tổng hợp văn bản...). Về nội dung này, chúng tôi đã tổng hợp lại và công bố thành một báo cáo khoa học và gửi đăng tại một tạp chí chuyên khảo về ngôn ngữ học.

Ngoài ra, trong quá trình thực hiện đề tài, chúng tôi đã đưa vào nhiều bộ từ điển song ngữ lớn để chuẩn bị cho phần mềm dịch máy hai chiều Anh-Việt / Việt-Anh, trong đó có các từ điển kỹ thuật tổng hợp Anh-Việt (95.000 mục từ), từ điển toán học (75.000 mục từ), từ điển y, sinh học (65.000 mục từ), từ điển kinh tế thương mại (40.000 mục từ)...tổng cộng 230.000 mục từ Anh-Việt và trên 260.000 mục từ Việt-Anh. Ngoài ra, hiện đang cập nhật bộ từ điển Việt-Anh với trên 200.000 cụn từ thông dụng và thành ngữ tiếng Việt.

Với cách tiếp cận mở rộng mô hình hình thức về văn phạm thì những giải thuật phân tích quen thuộc, hiệu quả cho văn phạm phi ngữ cảnh (giải thuật Early, Cock-Young-Casami,...) không thể sử dụng được, ít ra là trong dạng nguyên thủy của nó. Trong EVTRAN 2.0, để xử lý những ràng buộc ngữ cảnh bên ngoài khung văn phạm phi ngữ cảnh, giải thuật đã không còn giữ được độ phức tạp đa thức trong những tình huống nhất định, (hàm mũ

cho những trường hợp xấu nhất). Một điều đáng ngại là độ phức tạp tính toán đồng biến với dung lượng của cơ sở tri thức, nghĩa là càng có nhiều tri thức về ngôn ngữ thì hệ thống càng trở nên chậm chạp hơn. Điều này trái ngược với cách thức của con người: khi càng có nhiều tri thức, người phiên dịch càng thực hiện công việc dịch thuật nhanh chóng và đúng đắn hơn. Tư duy ngôn ngữ của con người là : Thay vì làm giảm tốc độ do phải đối chiếu với nhiều tình huống lựa chọn hơn, con người thường nhanh chóng chọn ngay những tình huống riêng (chuyên biệt) khả dĩ nhất mà bỏ qua, không xem xét đến những tình huống chung (phổ quát). Cách phân tích của con người làm cho khi người dịch càng có nhiều tri thức chuyên biệt (tri thức riêng) thì càng nhanh chóng xử lý bài toán hơn.

Phỏng theo cách tư duy của con người, ta có thể xây dựng giải thuật phân tích dựa trên cơ sở ưu tiên áp dụng các quy tắc riêng trước khi áp dụng các quy tắc chung. Khái niệm trực giác quy tắc riêng và quy tắc chung cần được định nghĩa để có thể phân loại và trên cơ sở đó, có một sự sắp xếp đối với tập quy tắc văn phạm. Mô hình văn phạm Chomsky không đề xuất một tiêu chí nào để so sánh hai quy tắc với nhau.

Để so sánh các quy tắc văn phạm với nhau, ta căn cứ vào tính phổ quát của chúng. Quy tắc được coi là chung hơn so với một quy tắc khác khi có một thành phần chung hơn một thành phần tương ứng của quy tắc khác. Mô hình văn phạm cảm ngữ đoạn đã đưa ra một phương thức để sắp thứ tự hệ quy tắc.

Bộ quy tắc được sắp thứ tự hình thành một dàn (bằng cách bổ sung thêm phần tử cực đại và cực tiểu – hai khái niệm trừu tượng để thể hiện *quy tắc bất kỳ* và *quy tắc rỗng* tương ứng)

Cần có giải thuật phân tích đủ nhanh để ứng dụng văn phạm cảm ngữ đoạn trong thực tế. Giải thuật phân tích theo chiều rộng với sự ưu tiên quy tắc riêng (phép tính nhỏ hơn trong quan hệ dàn đại số của các phần tử cú pháp trong văn phạm cảm ngữ đoạn) cho phép phân tích nhiều lần duyệt với heuristics sau:

- Ưu tiên quy tắc nhỏ hơn
- Dừng ngay khi gặp đích đầu tiên

Với chiến lược phân tích theo chiều rộng và ưu tiên quy tắc chi tiết; Giải thuật phân tích sẽ thực hiện cành nhanh khi càng có nhiều quy tắc riêng được áp dụng. Trong trường hợp suy biến, khi quy tắc riêng bao gồm toàn bộ câu nguồn thì quá trình phân tích dừng ngay lập tức, và giải thuật là một dạng "dịch theo nhớ" (translation memory).

Nhờ việc áp dụng giải thuật tiên nghiệm theo đó việc phân tích thực hiện tùy theo mức độ ưu tiên của các quy tắc, sự kết thúc thành công đầu tiên bao giờ cũng thuộc lớp các phân tích tốt nhất của câu nguồn.

Do tính chất của văn phạm cảm ngữ đoạn, độ sâu của cây cú pháp không bao giờ vượt quá n-1 (với n là độ sâu tối đa của dàn từ vựng của câu nguồn) và độ sâu trung bình của cây cú pháp là log2n (Phương pháp phân tích sử dụng thông tin kế thừa của các Nonterminal để giảm độ sâu của cây cú pháp).

Mô hình văn phạm cảm ngữ đoạn cũng cho phép đảm bảo giải thuật phân tích trên xuống luôn luôn dừng bất kể tính chất đệ quy trái của các quy tắc văn phạm (chỉ cần đếm và đánh dấu độ sâu của các nút được tạo thành trong cây cú pháp và không tiếp tục phát triển những nhánh có độ sâu k-1 với k là độ dài của phần đuôi câu nguồn kể từ nút hiện thời).

Mô hình văn phạm cảm ngữ đoạn đồng thời cũng đảm bảo giải thuật phân tích dưới lên luôn luôn dừng theo cách mà mô hình này đảm bảo tính dừng cho giải thuật phân tích trên xuống.

V.2.2. CÁC YÊU CẦU ĐỐI VỚI GIẢI THUẬT PHÂN TÍCH

Thời gian phân tích văn phạm tùy thuộc vào giải thuật được áp dụng. Lý thuyết về phân tích văn phạm được phát triển tương đôi sâu rộng từ những năm 60 – 70 của thế kỷ trước chủ yếu cho lớp ngôn ngữ phi ngữ cảnh. Một số kết quả chính bao gồm:

- Độ phức tạp của các giải thuật phân tích (dựng cây cú pháp) *trong trường hợp xấu nhất* là e^n với n là độ dài câu.
- Có thể đạt được độ phức tạp n^3 (dừng sau khi dựng được cây cú pháp đầu tiên). Tuy nhiên độ phức tạp của *giải thuật Early* tỷ lệ với bình phương của k (với k là số lượng quy tắc văn phạm). Điều này có nghĩa rằng tri thức về ngôn ngữ càng có nhiều thì bộ phân tích hoạt động càng chậm.
- Giải thuật với độ phức tạp n chỉ có thể đạt được cho một tập con rất hạn chế của lớp ngôn ngữ phi ngữ cảnh.

Văn phạm cảm ngữ đoạn là một mở rộng của văn phạm phi ngữ cảnh nên mọi sự khái quát hóa các giải thuật để thích nghi với mô hình đều không thể cải thiện được tốc độ phân tích, mà ngược lại, chỉ có thể làm cho tình trạng càng trở nên tồi tệ hơn. Do văn phạm cảm ngữ đoạn còn được sử dụng để giải quyết nhập nhằng nên số lượng quy tắc có thể rất lớn (hàng chục nghìn, thậm chí hàng trăm nghìn đơn vị cho các hệ dịch máy quy mô đủ lớn). Mặt khác, ta cũng không thể thỏa mãn với việc dừng phân tích ngay sau khi dựng được cây cú pháp bất kỳ. Thêm vào đó, ngôn ngữ phi ngữ cảnh

là mô hình giản lược với rất nhiều hạn chế khi đem so sánh với bất kỳ ngôn ngữ thực nào.

Với những lý do nêu trên, ta không thể chờ đợi một độ phức tạp nào khả quan hơn giá trị e^n khi thực hiện các giải thuật phân tích đối với ngôn ngữ tự nhiên.

Mặc dù vậy, giá trị độ phức tạp e^n là quá lớn để có thể ứng dụng trong thực tế.

Như vậy, rõ ràng cần phải xây dựng các heuristics để tăng tốc độ xử lý. Yêu cầu đối với giải thuật tiên nghiệm ở đây là:

- Có xu hướng chọn được cây cú pháp có giá trị của hàm định giá là tối thiểu (đối với phép tính so sánh trong dàn đại số của hệ phân cấp khái niệm)
- Dừng khi dựng được cây cú pháp đầu tiên (nếu không mọi heuristics đều đưa đến độ phức tạp tương đương hàm mũ của độ dài câu).
- Có độ phức tạp đa thức.
- Độ phức tạp của giải thuật không tăng lên khi tăng số lượng quy tắc trong văn phạm.

Đối với văn phạm cảm ngữ đoạn, yêu cầu cuội cùng là hết sức quan trọng vì văn phạm không những mô tả cú pháp của ngôn ngữ mà còn đưa ra các *luật hành văn* bao gồm hệ phân cấp ngữ nghĩa, các quy tắc thành ngữ cũng như các chỉ dẫn giải quyết nhập nhằng cho nên số lượng quy tắc phải rất lớn.

V.2.3. PHÂN TÍCH QUAY LUI, SÂU DẦN.

Các giải thuật phân tích quen thuộc đều có tính chất chung là độ phức tạp tính toán càng tăng khi cơ sở tri thức càng lớn : với việc bổ sung tập quy tắc văn phạm từ một tập con "*trò chơi*" cùa ngôn ngữ lên thành một bộ văn phạm quy mô lớn, chương trình trở nên chậm chạp một cách không thể chấp nhận được. Càng học, bộ phân tích càng đứng trước nhiều lựa chọn và càng khó tìm lời giải hơn. Đây là nghịch lý của các giải thuật tất định. Con người phiên dịch ứng xử hoàn toàn khác: càng tích lũy được kinh nghiệm, người dịch càng nhanh chóng tìm ra bản dịch đúng đắn nhất. Ta có thể lý giải nghịch lý này như sau: Khi biên dịch một văn bản, chúng ta không phân tích sâu văn bản đó mà thường dựa vào những *mẫu câu* (những *cấu trúc* hoặc những *cụm từ, thành ngữ*) tương tự đã từng gặp và cố gắng thử ghép chúng lại với nhau. Nếu thành công thì câu văn coi như đã được dịch xong, và đó lại thường là những bản dịch tốt nhất. Trong trường hợp lần thử này không dựng được cây cú pháp, ta sẽ *xem lại câu văn* kỹ hơn (*thử phân tích sâu*

hơn) để nhận được cách phân tích đúng. Quá trình dịch là một chu trình ***thử và sai*** được lặp lại với mỗi lần phân tích được đào sâu thêm một số bước nào đó.

Điều đáng lưu ý ở đây là bằng cách ghép các thành ngữ như vậy, ta đã chọn chính những cấu trúc có giá trị hàm định giá là tối thiểu trong ***văn phạm cảm ngữ đoạn*** được sử dụng để mô tả ngôn ngữ, nghĩa là ta luôn luôn có xu hướng chọn được bản dịch tốt nhất ngay trong lần thử đầu tiên (*nếu có thể*). Lược đồ phân tích như đã được mô tả ở trên chính là gợi ý cho giải thuật tiên nghiệm đang được vận dụng trong hệ dịch máy. Lần thứ nhất duyệt văn bản và dựng cây cú pháp với độ sâu bằng giá trị n_1 đủ nhỏ. Nếu dựng được cây cú pháp cho toàn bộ câu thì kết thúc quá trình phân tích. Duyệt lần hai với độ sâu $n_2 > n_1$. Quá trình này được lặp lại cho đến khi dựng được cây cú pháp đầy đủ hoặc với n_i đủ lớn. Việc lựa chọn các giá trị n_1, n_2..., n_i sẽ có ảnh hưởng đến tốc độ phân tích trung bình của giải thuật. Nếu các giá trị này quá nhỏ thì có thể việc tính toán phải quay lui nhiều lần trước khi đi tới đích. Mặt khác, với một cơ sở tri thức ngôn ngữ càng phong phú và chi tiết hơn thì có thể chọn n_1, n_2..., n_i với những giá trị càng nhỏ hơn (vì rằng lúc đó càng có nhiều khả năng sớm dựng được cây cú pháp đầy đủ hơn cho câu văn).

Khó khăn chính của giải thuật này nằm ở việc tổ chức tri thức ngôn ngữ. Hệ luật cần được tổ chức theo mô hình phân cấp ngữ nghĩa với việc sắp xếp độ ưu tiên trên cơ sở hệ phân cấp khái niệm (*dàn đại số khái niệm và dàn đại số luật sinh*).

V.2.4. HẠN ĐỊNH ĐỘ SÂU CÂY CÚ PHÁP

Một trong những đặc tính quan trọng của văn phạm cảm ngữ đoạn là mọi quy tắc sinh đều có tính chất là vế phải có độ dài lớn hơn vế trái. Nhờ tính chất này độ phức tạp của mọi giải thuật phân tích đều có thể đạt được giá trị không quá e^n. Vì vậy vấn đề đệ quy trái không bao giờ tồn tại trong các bộ phân tích trên xuống đối với văn phạm cảm ngữ đoạn: chỉ cần khống chế độ sâu phân tích không quá k-1 (với k là độ dài phần đuôi của *dạng câu* (*sentential form*) kể từ điểm đang phân tích).

Mặt khác, do độ dài vế phải của mọi quy tắc phi ngữ cảnh (trong bộ quy tắc cảm ngữ đoạn) đều lớn hơn hoặc bằng 2 nên độ sâu phổ biến của cây cú pháp có thể chờ đợi trong khoảng giá trị $\approx \log_2 n$. Giá trị này là chỉ dẫn tốt cho việc chọn các bộ giá trị độ sâu quay lui khi phân tích (bộ giá trị được chọn hiện nay 1, 2, 4, ..., $\log_2 n$, $2\log_2 n$, $4\log_2 n$).

V.2.5. KẾT QUẢ

Việc áp dụng đồng thời hai giải thuật tiên nghiệm nêu trên đã cho những kết quả rất khả quan. Tốc độ biên dịch tăng đột biến. Điều đặc biệt là

khi có nhiều tri thức, giải thuật phân tích không có xu hướng chậm lại (thời gian phân tích từ vựng tăng lên nhưng thời gian phân tích cú pháp giảm do giảm được độ sâu phân tích, kết quả là tổng thời gian xử lý giảm đáng kể). Tất nhiên ở đây mọi tính toán về độ phức tạp đều không cho những kết quả khả quan (*nói chung là hàm mũ*), nhưng thực tế có thể đạt tốc độ tuyến tính chỉ bằng cách bổ sung đủ cơ sở tri thức ngôn ngữ. Hiện nay chúng tôi đang tiến hành những khảo sát thống kê về hiệu năng tính toán của phương pháp tùy thuộc vào kích thước của cơ sở tri thức.

VI. TÀI LIỆU THAM KHẢO

[1] Noam Chomsky, On certain formal properties of grammars, Inform Control, vol 2, p.137-167, 1959.

[2] Christian Boitet (2002) A rationale for using UNL as an Interlingua and more in various domains, Geta, Clips, Imag, 385, av. de la bibliothèque, BP 53, F-38041 Grenoble cedex 9, France, Christian.Boitet@ imag.fr, LREC-02 First International Workshop on UNL, other Interlinguas and their Applications, 1 June 2002

[3] Bonnie Dorr and Nizar Habash (2002) Interlingua Approximation: A Generation-Heavy Approach, University of Maryland, Institute for Advanced Computer Studies, {bonnie,habash} @umiacs.umd.edu (UNITRAN)

[4] John Hutchins W. (2003) Machine translation: half a century of research and use, UNED summer school at Ávila, Spain, July 2003], http://ourworld.compuserve.com/ homepages/

[5] Stephen D. Richardson (2002) Achieving commercial-quality translation with example-based methods, Stephen D. Richardson, William B. Dolan, Arul Menezes, Jessie Pinkham, Microsoft Research, One Microsoft Way, Redmond, WA 98052, {steveri, billdol, arulm, jessiep}@ microsoft. com

[6] Arturo Trujillo (1999) Translation Engines: techniques for Machine Translation. Springer-Verlag, Berlin, 1999.

[7] Kevin Knight (1995) Integrating Knowledge Bases and Statistics in MT, Kevin Knight, Ishwar Chander, Matthew Haines, Vasileios Hatzivassiloglou, Eduard Hovy, Masayo Iida, Steve K. Luk, Akitoshi Okumura, Richard Whitney, Kenji Yamada, USC Information Science Institute, 4676 Admiralty Way, Marina del Rey, CA 90292

[8] DeryleW. Lonsdale, Alexander M. Franz, and John R. R. Leavitt (1994) Large-Scale Machine Translation: An Interlingua Approach, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, Pa., USA, 15213, Email: lonz@cs.cmu.edu, amf@cs.cmu.edu, jrrl@cs.cmu. edu (KANT)

[9] Michele Banko and Eric Brill (2002) Scaling to Very Very Large Corpora for Natural Language Disambiguation, Microsoft Research, 1 Microsoft Way, Redmond, WA 98052 USA, {mbanko, brill}@ microsoft.com

[10] Unification and Some New Grammatical Formalisms, Aravind K. Joshi, Department of Computer and Information Science, University of Pennsylvania (Nguồn : Internet)

[11] ISHIZAKI Shun, UCHIDA Hiroshi, (1998) On Interlingua for Multilingual Machine Translation, 1998, IPSJ SIGNotes Natural Language Abstract No.070 – 003

[12] Lê Khánh Hùng (2003) Văn phạm cảm ngữ đoạn, Báo cáo khoa học tại hội thảo quốc gia lần thứ sáu "Một số vấn đề chọn lọc của CNTT và TT", Thái nguyên, 8-2003.

[13] Lê Khánh Hùng, Trần Cảnh (2003) Về một số hạn chế của mô hình văn phạm Chomsky, Tạp chí Bưu chính Viễn thông, Chuyên san, 10, 2003.

[14] Lê Khánh Hùng (2003) Một Phương pháp Dịch máy Liên ngữ. Kỷ yếu Hội thảo Khoa học Quốc gia lần thứ nhất về Nghiên cứu, Phát triển và Ứng dụng CNTT&TT, Hà nội, 2003.

- Thiếu một Công cụ hình thức đủ mạnh và tổng quát để mô tả tri thức ngôn ngữ.
- Chưa có một giải pháp hình thức hữu hiệu cho vấn đề xử lý nhập nhằng
- Chưa tồn tại (và chưa rõ liệu có tồn tại) một liên ngữ đủ phong phú và thuận tiện làm trung gian cho mọi ngôn ngữ.