### CẤU TRÚC CỦA MỘT CHƯƠNG TRÌNH DỊCH TỰ ĐỘNG

Có nhiều chương trình dịch tự động khác nhau và các chương trình dịch này cũng có cấu trúc chi tiết khác nhau, tuy nhiên, về mặt cấu trúc tổng thể, tất cả các chương trình đều được chia làm 3 khối chính: khối xử lý hình thái, khối xử lý ngữ pháp và khối xử lý ngữ nghĩa. Sự khác nhau giữa các chương trình nằm ở cách thức và mức độ xử lý hình thái, ngữ pháp hay ngữ nghĩa khác nhau. Dù vậy, hầu hết các chương trình đều làm việc một cách đơn giản theo sơ đồ sau:

(Đại từ)

(Danh từ)

(Trợ động từ)

(Ngoại động từ)

(Danh từ)

(Tính từ)

(Danh từ)

Xử lý hình thái

Xử lý ngữ pháp

(Obj)

(Obj)

(S+A2)

(S,J)

(Sub)

Mạo từ Phó từ Tính từ Danh từ

Tân tử

Trợ động từ Động từ

Đại từ

Câu nguồn: I will give him a very difficult exam

Từ loại: I will ……………….. difficult exam

Cú pháp:

 I will give him a very dificult exam

Xử lý ngữ nghĩa

Câu đích: Tôi sẽ cho anh ta một bài thi rất khó

h ta một bài thi rất khó

*Hình 1: Quá trình xử lý của một chương trình dịch tự động*

**Khối xử lý hình thái**

Khối xử lý hình thái là khối đầu tiên của chương trình dịch tự động, khối nhận dữ liệu đầu vào là một câu của ngôn ngữ nguồn và phân tích hình thái của câu đó. Nói chung, một khối phân tích hình thái thường làm các nhiệm vụ sau:

(1) Chia cắt một câu trong ngôn ngữ nguồn thành các từ độc lập.

(2) Tìm tất cả các hình thái và thông tin bổ sung của các từ trong câu để phục vụ quá trình xử lý về sau.

**Khối xử lý ngữ pháp**

 Khối xử lý ngữ pháp xây dựng cây phân tích ngữ pháp cho câu đầu vào dựa trên văn phạm của ngôn ngữ nguồn. Khối này nhận danh sách cách hình thái từ đã được sinh ra bởi khối xử lý hình thái, khối này sẽ tiến hành phân tích văn phạm của câu dựa trên văn phạm của ngôn ngữ nguồn và thông tin bổ sung - thường là các văn phạm bổ sung cho phép sinh cây phân tích chính xác hơn. Nói chung, có thể áp dụng các phương pháp phân tích cú pháp của các ngôn ngữ nhân tạo cho ngôn ngữ tự nhiên, tuy nhiên vấn đề không suôn sẻ như với các ngôn ngữ nhân tạo, có những vấn đề chính sau đây thường gặp phải khi chúng ta sử dụng các bộ phân tích văn phạm thông thường:

 (1) Văn phạm của các ngôn ngữ tự nhiên thường rất phức tạp.

(2) Trong ngôn ngữ nói, người sử dụng thường nói tắt, thậm chí nói sai ngữ pháp – không thể phân tích ngữ pháp được.

(3) Xuất hiện hiện tượng có nhiều cây phân tích ứng với một câu – dẫn đến nhiếu khả năng dịch khác nhau cho một câu.

**Khối xử lý ngữ nghĩa**

Xử lý ngữ nghĩa là khối cuối cùng trong dịch tự động, khối nhận đầu vào là cây phân tích từ khối xử lý ngữ pháp và thực hiện việc sinh câu ở ngôn ngữ đích. Việc thực hiện sinh câu được thực hiện dựa trên 4 nguồn thông tin chính: Cây phân tích ngữ pháp (lấy từ khối xử lý ngữ pháp), nghĩa của các từ tố trong câu văn nguồn (lấy từ dãy hình thái từ tố của khối xử lý hình thái), cú pháp điều khiển (là thông tin bổ sung của văn phạm, được sinh bởi khối xử lý hình thái) và thông tin ngữ cảnh (được tổng hợp từ việc xử lý ngữ nghĩa các câu trước và nhận vào từ người sử dụng chương trình). Nhiệm vụ chính của khối xử lý ngữ nghĩa là:

(1) Thực hiện việc đánh giá các cây phân tích cú pháp (nếu khối xử lý ngữ pháp phân tích được nhiều cây phân tích) và lựa chọn cây tốt nhất.

(2) Thực hiện việc chọn nghĩa thích hợp nhất cho các từ tố của ngôn ngữ nguồn trong ngôn ngữ đích.

(3) Sinh câu trong ngôn ngữ đích đúng với văn phạm của ngôn ngữ đó (vấn đề xử lý thứ tự các từ).