

BỘ KHOA HỌC VÀ CÔNG NGHỆ
CHƯƠNG TRÌNH KH&CN CẤP NHÀ NƯỚC KC01/06-10
Đề tài KC01.01/06-10
“Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý
tiếng nói và văn bản tiếng Việt”

PHỤ LỤC SẢN PHẨM

SP7.3 - KHO NGỮ LIỆU CÂU TIẾNG VIỆT
CÓ CHÚ GIẢI (VIETREEBANK)
(Quyển 2)

Thời gian thực hiện: 5/2007- 5/2009

Chủ trì thực hiện: TS. Nguyễn Phương Thái
Đơn vị công tác: Khoa Công nghệ thông tin,
Đại học Công nghệ,
Đại học Quốc gia Hà nội

HÀ NỘI, 05/2009

MỤC LỤC

Nghiên cứu nội dung và cấu trúc cho ngân hàng câu tiếng Việt được chú giải ngữ pháp.....	3
<i>Nguyễn Phương Thái, Vũ Xuân Lương</i>	
Mã hóa dữ liệu treebank tiếng Việt.....	25
<i>Lê Hồng Phương, Nguyễn Thị Minh Huyền, Phan Thị Hà</i>	
Công cụ hỗ trợ xây dựng treebank tiếng Việt – SynAF.....	34
<i>Lê Hồng Phương, Lưu Văn Tăng, Nguyễn Thị Minh Huyền</i>	
Hướng dẫn sử dụng công cụ soạn thảo treebank (TBE).....	46
<i>Nguyễn Phương Thái, Lê Anh Cường</i>	
Hướng dẫn tách câu tiếng Việt.....	53
<i>Đình Diên</i>	
Hướng dẫn tách từ tiếng Việt.....	60
<i>Nguyễn Thị Minh Huyền, Hoàng Thị Tuyền Linh, Vũ Xuân Lương</i>	
Hướng dẫn gán nhãn từ loại tiếng Việt.....	72
<i>Nguyễn Phương Thái, Vũ Xuân Lương, Nguyễn Thị Minh Huyền</i>	
Hướng dẫn gán nhãn cú pháp tiếng Việt.....	78
<i>Nguyễn Phương Thái, Vũ Xuân Lương, Nguyễn Thị Minh Huyền, Nguyễn Thị Minh Ngọc, Đào Minh Thu, Lê Kim Ngân</i>	

Nghiên cứu nội dung và cấu trúc cho ngân hàng câu tiếng Việt được chú giải ngữ pháp

Nguyễn Phương Thái¹, Vũ Xuân Lương², Nguyễn Thị Minh Huyền³
SP 7.3 - VLSP

Giới thiệu

Quá trình xây dựng treebank có một số bước cơ bản là: *tìm hiểu*, thiết kế, xây dựng công cụ, thu thập ngữ liệu thô, và gán nhãn dữ liệu. Trong tài liệu này chúng tôi trình bày kết quả của giai đoạn tìm hiểu. Tài liệu được tổ chức thành hai phần chính. Phần thứ nhất trình bày về các loại treebank, tiếp cận xây dựng, kinh nghiệm xây dựng treebank của các ngôn ngữ khác. Ở phần hai chúng tôi trình bày về các đặc điểm ngữ pháp tiếng Việt.

Nội dung

Tìm hiểu các Penn Treebank.....	4
<i>Nguyễn Phương Thái, Vũ Xuân Lương, Nguyễn Thị Minh Huyền</i>	
Tìm hiểu ngữ pháp tiếng Việt.....	12
<i>Vũ Xuân Lương</i>	

¹ JAIST

² Trung Tâm Từ Điển Học

³ ĐH KHTN, ĐH QGHN

Tìm hiểu các Penn Treebank

Nguyễn Phương Thái, Vũ Xuân Lương, Nguyễn Thị Minh Huyền
SP 7.3 – Dự án VLSP

Nội dung:

- *Xây dựng tập nhãn từ loại*
- *Xây dựng tập nhãn cú pháp*
- *Công cụ*
- *Chọn văn bản thô*
- *Kích thước corpus*
- *Mã hóa cây cú pháp*
- *Gán nhãn*
- *Quá trình xây dựng tài liệu hướng dẫn gán nhãn*

Giới thiệu

Trong tài liệu này chúng tôi trình bày tiếp cận xây dựng treebank của Marcus và cộng sự (1993). Các vấn đề được trình bày bao gồm: tập nhãn, tài liệu hướng dẫn gán nhãn, công cụ, cách tiến hành quá trình gán nhãn. Đây là một tiếp cận đã được kiểm chứng qua việc áp dụng cho nhiều ngôn ngữ khác nhau như: tiếng Anh, một ngôn ngữ thuộc họ Ấn-Âu; tiếng Trung, một họ ngôn ngữ riêng; tiếng Hàn; tiếng Ả-rập. Do đó kinh nghiệm xây dựng treebank của các ngôn ngữ này cũng được đề cập khi có thể.

Tiếp cận xây dựng treebank này cho phép thể hiện nhiều loại thông tin ngôn ngữ học khác nhau trên cây cú pháp như từ loại, cụm từ, mệnh đề, chức năng ngữ pháp (chủ ngữ, vị ngữ, trạng ngữ, đề ngữ, v.v.), v.v. Thêm vào đó, nó đưa ra những tiêu chí kỹ thuật về thiết kế tập nhãn, tài liệu hướng dẫn gán nhãn, cũng như qui trình gán nhãn dữ liệu. Tuy tiếng Việt có những đặc thù riêng nhưng khi xây dựng treebank tiếng Việt chúng ta cần học hỏi từ những ngôn ngữ khác.

1. Xây dựng tập nhãn từ loại

1.1 Các thông tin có thể chứa trong nhãn từ loại

Về nguyên tắc, các thông tin về từ có thể được chứa trong từ loại bao gồm: từ loại cơ sở (danh từ, động từ, v.v.), thông tin hình thái (số ít, số nhiều, thì, ngôi, v.v.), thông tin về phân loại con (ví dụ động từ đi với danh từ, động từ đi với mệnh đề that, v.v.), thông tin ngữ nghĩa, hay một số thông tin cú pháp khác. Ví dụ nhãn NNS của Penn Treebank (PTB) cho biết từ loại danh từ ở số nhiều, nhãn VBZ cho biết từ loại động từ ở ngôi thứ ba số ít. Có một điểm đáng chú ý là nhãn từ loại của các treebank thường chỉ chứa thông tin từ loại cơ sở và thông tin hình thái (phần 1.2 và 1.3 sẽ phân tích tại sao lại như vậy). Như vậy tập nhãn của treebank sẽ nhỏ gọn hơn rất nhiều các tập nhãn mà các nhãn thành phần chứa cả các thông tin khác. Ví dụ tập nhãn của PTB có 6 từ loại động từ:

- + VB: động từ nguyên mẫu
- + VBZ: động từ ngôi thứ ba số ít, thì hiện tại
- + VBP: động từ ở thì hiện tại và không phải là ngôi thứ ba số ít
- + VBD: động từ ở thì quá khứ
- + VBN: động từ ở thì quá khứ phân từ
- + VBG: danh động từ hoặc động từ ở thì hiện tại phân từ

Tập nhãn của CTB chỉ có 4 từ loại động từ (chú ý là tiếng Trung không biến hình từ):

- + VA: tính từ vị ngữ. Ví dụ câu: “Cô ấy đẹp”, thì “đẹp” có nhãn là VA
 - + VC: động từ nối. Ví dụ câu “Anh ấy là sinh viên”, thì “là” có nhãn là VC
 - + VE: dành cho động từ trong các câu như “có năm sinh viên trong lớp”, khi đó “có” được gán nhãn là VE.
 - + VV: các động từ khác (nội động từ, ngoại động từ, động từ tình thái, v.v.)
- (Trong các ví dụ về loại động từ của CTB, tôi dùng ví dụ tiếng Việt cho dễ hiểu).

Một ví dụ về tập nhãn được phân loại mịn là từ điển OALD (Oxford Advanced Learner Dictionary), tập nhãn của nó chứa tới hơn 30 nhãn động từ do các nhãn này có cả thông tin về phân loại con (subcategorization). Từ điển COMLEX⁴ phân loại mịn nhất với số nhãn từ loại động từ lớn gấp nhiều lần.

1.2 Tính gia tăng trong các vấn đề của XLNNTN

Phần này tìm cách giải thích cho câu hỏi tại sao không đưa thông tin ngữ nghĩa hay thông tin phân loại con vào nhãn từ loại?

Trước hết cần chú ý là các vấn đề trong XLNNTN được tổ chức theo kiểu tăng dần độ phức tạp:

- Phân đoạn từ
- Gán nhãn từ loại
- Phân tích cú pháp nông
- Phân tích cú pháp đầy đủ
- Phân tích ngữ nghĩa

⁴ <http://nlp.cs.nyu.edu/comlex/index.html>

Các nghiên cứu hiện tại⁵ cho thấy cấu trúc phân cấp này là hiệu quả. Khi giải quyết vấn đề ở mức i , thông thường các kết quả của các mức trước đó được sử dụng. Chẳng hạn như khi phân tích ngữ nghĩa, người ta có thể giả sử câu đã được phân tích cú pháp đầy đủ. Ngược lại, nếu có thông tin ngữ nghĩa thì có cải tiến được phân tích cú pháp hay gán nhãn từ loại không? Câu trả lời thường là cải tiến rất ít hoặc thậm chí mang lại kết quả ngược với mong muốn. Đó là lý do người ta không đưa ngược thông tin ở các mức trên vào mức dưới. Đến đây ta có thể hiểu tại sao các treebank đã không đưa thông tin ngữ nghĩa (mức phân tích ngữ nghĩa) hay thông tin phân loại con (mức phân tích cú pháp đầy đủ) vào nhãn từ loại (mức gán nhãn từ loại).

Ngay cả trong các nghiên cứu ngôn ngữ học, nhiều tác giả phân loại từ dựa vào cả thông tin ngữ nghĩa. Tuy nhiên phổ biến hơn vẫn là quan điểm phân biệt các phạm trù ngữ pháp, ngữ nghĩa, và ngữ dụng. Khi đã phân biệt như thế ta có thể nghiên cứu từng lĩnh vực một cách độc lập tương đối.

1.3 Tính khôi phục được

So với một số corpus khác, PTB có tập nhãn từ loại đã được đơn giản hóa. Ngoài lý do được nêu trong phần 1.2, còn có một lý do quan trọng khác là làm giảm hiện tượng dư liệu thừa⁶. Chiến lược chủ yếu để làm giảm kích thước tập nhãn là cân nhắc cả thông tin từ vựng và thông tin cú pháp. Bằng cách sử dụng thông tin từ vựng, PTB tránh dùng các nhãn được đặt ra chỉ cho một từ cụ thể. Ta có thể lấy từ “have” làm ví dụ. Từ này vừa có thể là động từ, vừa có thể là trợ động từ. Mới nhìn qua thì ta thấy nên đặt ra 2 nhãn khác nhau cho nó. Tuy nhiên chỉ cần gán nhãn động từ cho mọi trường hợp là xong, bởi vì việc từ này có thể là trợ động từ không có thể xác định dựa vào ngữ cảnh và vào thông tin từ vựng (tức là nếu cần, ta có thể dùng một thủ tục đơn giản để chuyển đổi nó sang nhãn trợ động từ một cách tự động). Tương tự như vậy, những từ loại mà có thể khôi phục sử dụng thông tin về cấu trúc cú pháp ta có thể bỏ đi. Các ví dụ có thể có là về đại từ, giới từ, hoặc động từ với các phân loại con như ở phần 1.1 (chỉ cần đặt ra một loại thay vì chia thành nhiều loại).

1.4 Tính nhất quán

Một tập nhãn tốt giúp cho việc gán nhãn có tính nhất quán cao. Giảm thiểu các trường hợp nhập nhằng mà người gán nhãn cảm thấy có nhiều hơn một lựa chọn đúng. Một ví dụ là tập nhãn của PTB không có nhãn RN như của Brown Corpus (RN là một loại phó từ) mà chỉ có một nhãn duy nhất là RB cho phó từ. Nếu dùng RN thì các từ như “here” và “then” khi thì được gán nhãn RB khi thì được gán nhãn RN – thậm chí trong các ngữ cảnh cú pháp giống hệt nhau.

1.5 Chức năng ngữ pháp

⁵ XLNNTN bằng tiếp cận thống kê

⁶ Vì treebank phục vụ cho các nghiên cứu về ngôn ngữ và xử lý ngôn ngữ bằng tiếp cận thống kê.

Có một số trường hợp, nhãn từ loại được xác định dựa vào chức năng cú pháp của từ. Ví dụ như trong cụm từ “the one”, “one” được gán nhãn là NN (danh từ) thay vì CD (số từ). Lý do là “one” là từ trung tâm của cụm từ “the one”.

1.6 Các trường hợp không xác định

Cho dù tập nhãn đã được thiết kế thỏa mãn tất cả các tiêu chí kể trên, vẫn có thể có những trường hợp người gán nhãn không thể xác định một nhãn duy nhất cho một từ nào đó. Đối với trường hợp này cần liệt kê các nhãn hợp lý cho từ phân cách nhau bởi dấu ‘|’ thay vì chọn ngẫu nhiên chỉ một nhãn.

2. Xây dựng tập nhãn cú pháp

2.1 Nhãn thành phần cú pháp

Loại nhãn này mô tả các thành phần cú pháp cơ bản là cụm từ và mệnh đề. Nhãn thành phần cú pháp là thông tin cơ bản nhất trên cây cú pháp, nó tạo thành xương sống của cây cú pháp⁷. Tập nhãn cú pháp của các ngôn ngữ khác nhau là khác nhau (ở một tỉ lệ nhất định) vì hai nguyên nhân. Nguyên nhân cơ bản nhất là do sự khác biệt về ngôn ngữ. Chẳng hạn như trong tiếng Trung, từ chỉ loại có chức năng làm bổ nghĩa trước cho danh từ. Từ chỉ loại lại có thể được kết hợp với số từ trong phần phụ trước của cụm danh từ. Vì vậy nhóm thiết kế Chinese Treebank (CTB) đã đặt ra nhãn cụm từ chỉ loại. Đây là một điểm khác biệt với treebank tiếng Anh. Nguyên nhân thứ hai là do kỹ thuật thiết kế tập nhãn. Chẳng hạn như với các cụm từ nghi vấn, PTB có 4 loại nhãn là WHNP, WHPP, WHADJP, WHADVP. Trong khi CTB lại chỉ đặt ra một nhãn chức năng là WH. Nhãn này sẽ được dùng kèm với nhãn cụm từ khi trong cụm từ đó có từ dùng để hỏi. Như vậy vẫn đủ để mô tả các cụm từ nghi vấn (NP-WH, PP-WH, ADJP-WH, ADVP-WH).

2.2 Nhãn chức năng cú pháp

Nhãn chức năng của một thành phần cú pháp cho biết vai trò của nó trong thành phần cú pháp mức cao hơn. Nhãn chức năng cú pháp được gán cho các thành phần chính trong câu như chủ ngữ, vị ngữ, tân ngữ. Nhờ thông tin do nhãn chức năng cung cấp ta có thể xác định các loại quan hệ ngữ pháp cơ bản sau đây:

- Chủ-vị
- Đề-thuyết
- Phân chêm
- Bổ ngữ
- Phụ ngữ
- Sự kết hợp

⁷ Nhiều lý thuyết về cú pháp dựa trên cấu trúc xương sống này.

Ngoài ra nhãn chức năng cũng có thể tương ứng với một loại phụ ngữ nào đó, ví dụ thời gian, nơi chốn, hay mục đích. Như vậy loại nhãn chức năng này chứa thông tin ngữ nghĩa “nông” của một thành phần cú pháp. Hình 1 chỉ ra một ví dụ, trong đó cụm danh từ “the committee” có nhãn chức năng là SBJ cho biết nó là chủ từ trong câu, còn mệnh đề trạng ngữ “while eating lunch” có nhãn chức năng là TMP cho biết nó chỉ thời gian.

2.3 Nhãn thành phần rỗng

Đây là một loại thành phần khá đặc biệt. Nó chỉ ra sự tồn tại (được ngầm hiểu) của một thành phần cú pháp cho dù nó không xuất hiện ở vị trí đó. Thông thường thành phần rỗng được gán chỉ số của thành phần mà nó đại diện. Hình 1 chỉ ra một ví dụ:

```
(S (NP-SBJ-1 The committee)
  (VP continued
    (NP its meeting)
    (SBAR-TMP while
      (S (NP-SBJ *-1)
        (VP eating
          (NP lunch))))))
```

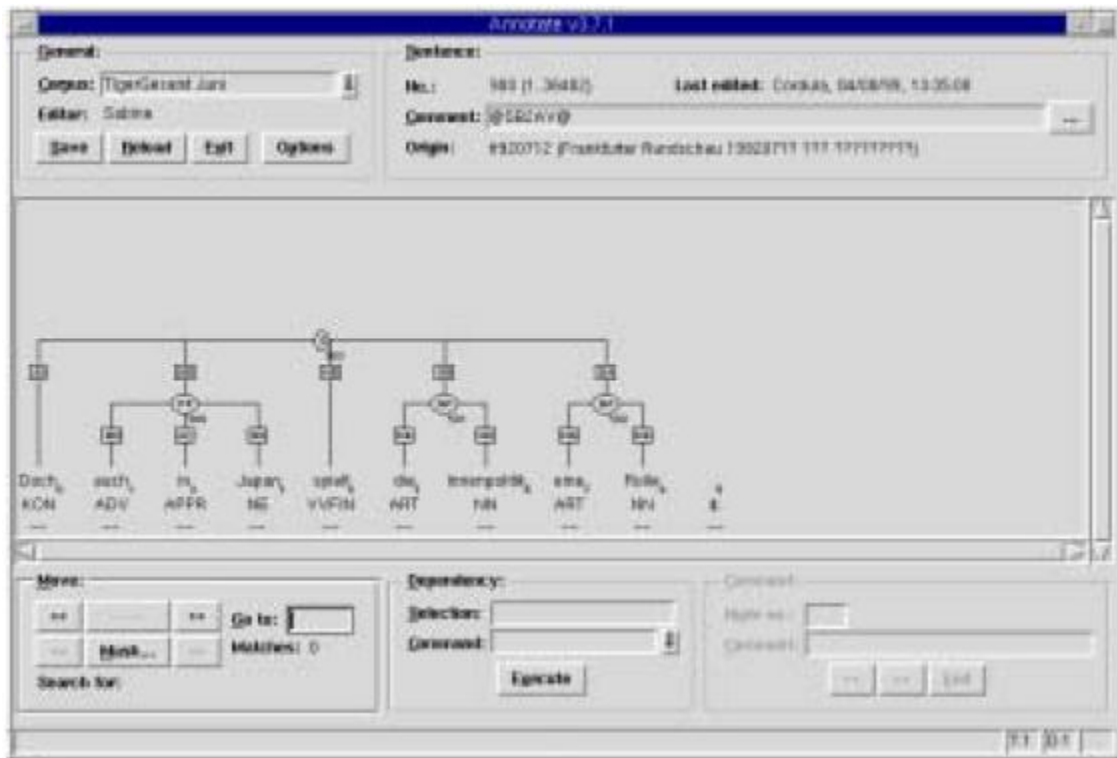
Hình 1. Một cây cú pháp tiếng Anh

3. Công cụ

Công cụ hỗ trợ những người gán nhãn làm việc hiệu quả hơn. Có hai nội dung chính là hỗ trợ soạn thảo cây cú pháp (giao diện) và gán nhãn trước, sau đó người sẽ sửa lại. Kinh nghiệm xây dựng treebank đã cho thấy là công cụ giúp tăng tốc độ gán nhãn lên rất nhiều. Hình dưới đây là của công cụ soạn thảo cây cú pháp của Tiger Treebank, một treebank tiếng Đức (Sabine Brants và cộng sự, 2003).

Tùy điều kiện mà ta lựa chọn công cụ gán nhãn tự động thích hợp. Chẳng hạn với việc gán nhãn từ loại, nếu đã có sẵn chương trình gán nhãn từ loại thì ta sử dụng nó làm công cụ luôn. Nếu không thì ta chấp nhận việc phải gán nhãn từ đầu (bằng tay hoàn toàn) cho một phần ngữ liệu thô. Sau đó huấn luyện hệ gán nhãn từ loại⁸ dựa trên phần này rồi dùng nó làm công cụ xử lý phần còn lại của kho ngữ liệu thô. Việc này có thể được lặp lại trong quá trình làm việc.

⁸ Trên Internet có sẵn một số hệ mã nguồn mở, ta có thể tùy biến nó để dùng cho ngôn ngữ mới



Hình 2. Công cụ của Tiger Treebank

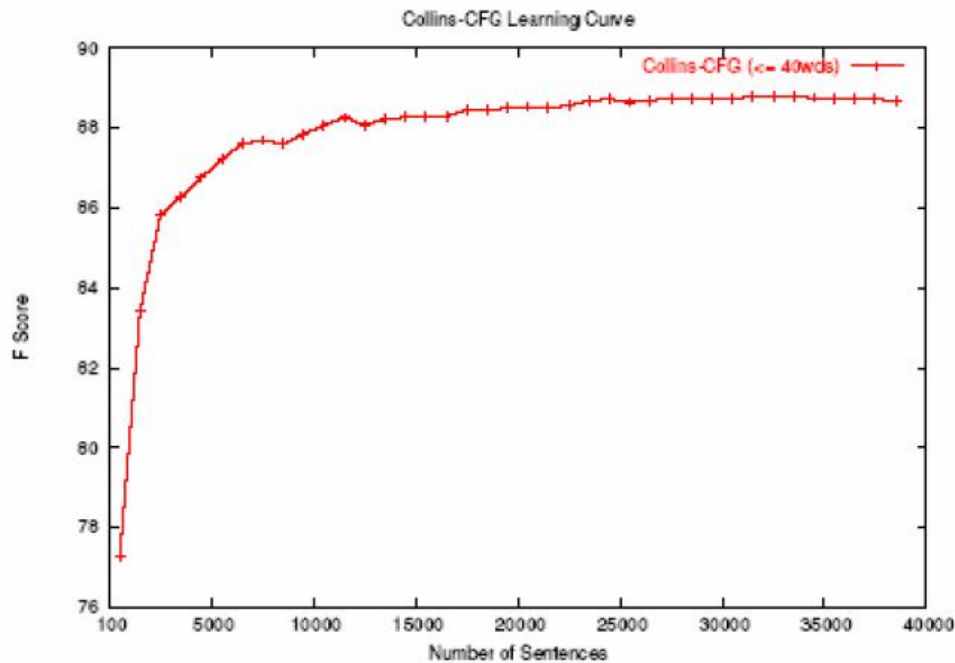
4. Chọn văn bản thô

Đối với các corpus văn bản không được gán nhãn phục vụ cho nghiên cứu từ vựng, từ điển thì thường được lấy mẫu trên phạm vi rộng, bao trùm nhiều chủ đề. Tuy nhiên với các corpus gán nhãn cú pháp đầy đủ thì kích thước corpus nhỏ hơn và chủ đề cũng hẹp hơn. Thông thường lấy trên một chủ đề, nếu như corpus chỉ có kích thước vài chục ngàn câu. Chẳng hạn như treebank tiếng Trung là báo XinHua (Fei Xia và cộng sự, 2000). Treebank tiếng Anh (Marcus và cộng sự, 2003) thì gồm nhiều chủ đề, đây là treebank lớn nhất và được xây dựng công phu nhất. Giới nghiên cứu phân tích cú pháp hay sử dụng phần Wall Street Journal của corpus này. Giả sử ta chọn một báo nào đó, lấy theo một chủ đề nào đó thì cũng lấy các bài trong một khoảng thời gian nhất định. Cách làm này giảm hiện tượng dữ liệu thừa.

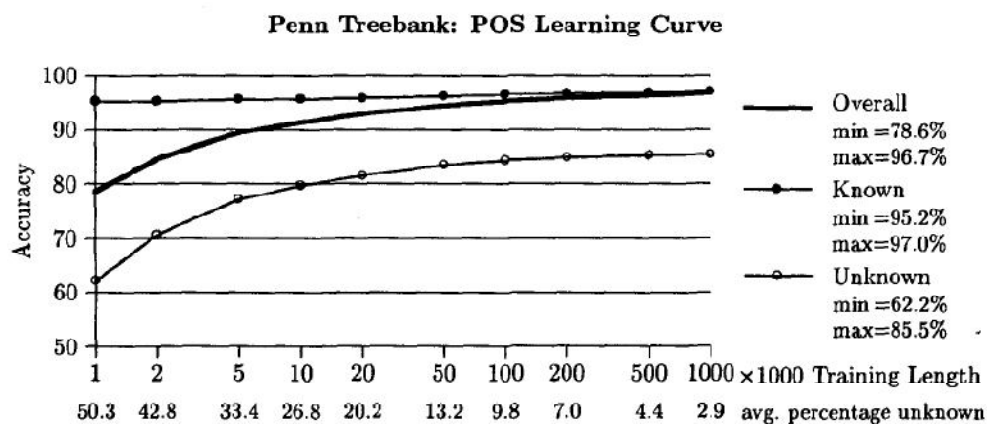
5. Kích thước corpus

Kích thước của corpus cũng là một vấn đề. Về lý thuyết, càng nhiều dữ liệu càng tốt, đặc biệt đối với các tool mà được huấn luyện dựa vào treebank. Tuy nhiên do các giới hạn về thời gian và kinh phí, trong thực tế các corpus khá hạn chế về kích thước. Các nghiên cứu về phân tích cú pháp tiếng Anh đã cho thấy một số điều khá thú vị. Độ chính xác test trên Penn Treebank của các hệ phân tích cú pháp tiếng Anh tốt nhất hiện nay đạt khoảng 90%. Đường cong trong Hình 3 chỉ ra sự tương quan giữa số câu huấn luyện và độ chính xác của hệ phân tích cú pháp (Steedman và Osborne, 2003). Theo hình vẽ đó để đạt chất lượng gần 88%, chỉ cần khoảng 10000 câu huấn luyện. Đối với gán nhãn từ loại tiếng Anh, độ chính xác tối đa khoảng vào khoảng 97%. Theo Hình 4, nếu ta có 10000 câu (ứng với 200000 từ tổ, độ dài trung bình một

câu khoảng 20 từ tổ), thì chất lượng có thể đạt 95% (Brants, 2000). Như vậy với tiếng Anh, trong cả hai trường hợp ta đều có thể đạt xấp xỉ 98% độ chính xác tối đa với 10000 câu huấn luyện. Đây là một căn cứ của việc chọn 10000 câu làm mục tiêu cho giai đoạn 2007-2009 của xây dựng treebank tiếng Việt.



Hình 3. Tương quan số câu huấn luyện và độ chính xác phân tích cú pháp



Hình 4. Tương quan số câu huấn luyện và độ chính xác gán nhãn từ loại

6. Mã hóa cây cú pháp

Có hai cách thường được sử dụng để mã hóa cây cú pháp. Cách thứ nhất⁹ đơn giản sử dụng cấu trúc dấu ngoặc như trong Hình 1. Theo cách này mỗi thành phần cú pháp sẽ có một cặp dấu ngoặc bao quanh. Ngay sau dấu ngoặc đầu tiên là ký hiệu ngữ pháp và các thuộc tính (nếu có). Sau đó sẽ là danh sách các thành phần cú pháp con. Cách thứ hai là sử dụng lược đồ mã hóa XML. Cách này đã được nghiên cứu kỹ lưỡng và được áp dụng vào một số dự án về xử lý

⁹ Vì tính đơn giản mà cách này được sử dụng rộng rãi khi xây dựng treebank

ngôn ngữ của Châu Âu¹⁰. Sau đây là ví dụ về biểu diễn cây cú pháp của câu “I love you” bằng lược đồ này:

```
<struct type="S" >
  <struct type="NPB" >
    <word type="PRP"> I </word>
  </struct>
  <struct type="VP" >
    <word type="VBP"> love </word>
    <struct type="NPB">
      <word type="PRP"> you </word>
      <word type="PUNC."> . </word>
    </struct>
  </struct>
</struct>
```

7. Gán nhãn

Quá trình gán nhãn một câu gồm ba bước: tách từ, gán nhãn từ loại, và phân tích cú pháp. Qui trình thực hiện gán nhãn là tương tự nhau, tuy nhiên mỗi bước yêu cầu những kiến thức và có những đặc trưng riêng. Trước tiên, những người gán nhãn cần được huấn luyện về cách gán nhãn, tập nhãn, và cách sử dụng công cụ. Sau đó họ sẽ gán nhãn cho từng phần của corpus thô. Sau mỗi phần là qui trình test và kiểm tra chéo để biết được mức độ đồng thuận. Cách kiểm tra chéo là so sánh xem kết quả gán nhãn cùng một văn bản của 2 người (hay nhóm người) khác nhau là bao nhiêu. Nếu sự khác biệt là quá lớn thì có vấn đề hoặc về phía người gán nhãn hoặc do bản tài liệu hướng dẫn (thiết kế). Để so sánh, ta có thể dùng người hoặc dùng một phương pháp tự động nào đó, ví dụ Parseval. Ngoài ra, trong quá trình gán nhãn cần có tương tác chặt chẽ giữa nhóm gán nhãn và nhóm thiết kế bởi vì có những hiện tượng ngữ pháp chưa có trong bản hướng dẫn.

8. Quá trình xây dựng tài liệu hướng dẫn gán nhãn

Đây là một tài liệu rất quan trọng. Nó bao gồm không chỉ các thông tin về tập nhãn, mà còn hướng dẫn gán nhãn cho các hiện tượng cụ thể với các ví dụ minh họa. Để xây dựng tài liệu này, nghiên cứu các tài liệu về ngữ pháp và về kinh nghiệm xây dựng treebank đã có là việc đầu tiên cần làm. Ngoài ra còn cần cộng tác chặt chẽ với các nhà ngôn ngữ để xử lý các hiện tượng khó. Khi gặp hiện tượng khó và có một vài lựa chọn, chủ động chọn một cái và khi cần thì chuyển đổi sang cái kia. Tham gia hoặc tổ chức các workshop về vấn đề liên quan. Nếu có điều kiện thì mời các chuyên gia nước ngoài cố vấn. Những người gán nhãn được khuyến khích đưa ra các câu hỏi trong quá trình làm việc.

Khi xây dựng phiên bản đầu tiên của tài liệu này, nhóm thiết kế cần tự tay phân tích trên một tập câu mẫu lấy từ sách ngữ pháp, vừa phân tích vừa viết tài liệu. Kết quả sẽ bao trùm các cấu

¹⁰ <http://www.xml-ces.org/>

trúc và hiện tượng ngữ pháp cơ bản nhất. Bước kế tiếp là phân tích các câu lấy từ ngữ liệu thực tế (kết quả của bước chọn văn bản thô). Việc này rất quan trọng, nó giúp nhóm thiết kế đưa ra được tài liệu sát với thực tế hơn là chỉ dựa vào các câu mẫu trong sách. Các vấn đề ngôn ngữ phát sinh khi xây dựng treebank đa dạng và phức tạp hơn nhiều so với những cái cơ bản được chỉ ra trong các sách ngữ pháp (Han và cộng sự, 2002). Do đó tài liệu hướng dẫn còn được chỉnh sửa, nâng cấp, và bổ xung trong quá trình gán nhãn văn bản.

Tài liệu tham khảo

- [1] Thorsten Brants, 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA
- [2] Sabine Brants et al. The TIGER Treebank. 2003. COLING.
- [3] Vũ Dũng. 2003. Tiếng Việt và ngôn ngữ học hiện đại sơ khảo về cú pháp. VIET Stuttgart – Germany, 2004.
- [4] Chung-hye Han et al. Development and Evaluation of a Korean Treebank and its Application to NLP. 2002. LREC.
- [5] Cao Xuân Hạo. 2006. Tiếng Việt sơ thảo ngữ pháp chức năng. NXB KHXH, 2006.
- [6] Mitchell P. Marcus et al. Building a Large Annotated Corpus of English: The Penn Treebank. 1993. Computational Linguistics.
- [7] Mark Steedman, Miles Osborne. 2003. Bootstrapping Statistical Parsers from Small Datasets. EACL 2003.
- [8] Mark Steedman, Rebecca Hwa. 2003. Example Selection for Bootstrapping Statistical Parsers. NA-ACL 2003.
- [9] Fei Xia et al. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. 2000. COLING.
- [10] Nianwen Xue et al. Building a Large-Scale Annotated Chinese Corpus. 2002. COLING.

TÌM HIỂU NGỮ PHÁP TIẾNG VIỆT

Vũ Xuân Lương – Trung tâm Từ điển học

I. GIỚI THIỆU

1. Để miêu tả một ngôn ngữ đòi hỏi phải xác định được đặc điểm ngữ pháp của ngôn ngữ đó. Ngữ pháp của bất kì một ngôn ngữ nào cũng là một hệ thống bao gồm các đơn vị, các kết cấu và các quan hệ thuộc nhiều tầng bậc khác nhau. Trong ngôn ngữ, nếu như từ, ngữ và câu thường biểu hiện những gì là cụ thể thì ngữ pháp lại có tính khái quát cao. Ngữ pháp hướng đến các quy tắc về cấu tạo từ, kết hợp từ và các quy tắc tạo câu của một ngôn ngữ. Ngữ pháp của một ngôn ngữ thường có tính bền vững, hay nói đúng hơn, nếu có biến đổi thì biến

đổi cũng rất chậm. Ngữ pháp của ngôn ngữ này có những điểm giống và khác nhau với các ngôn ngữ khác, do đó không thể có chung một bộ quy tắc ngữ pháp cho mọi ngôn ngữ.

2. Nghiên cứu ngữ pháp có hai bộ phận là từ pháp học (hình thái học) và cú pháp học. Từ pháp học nghiên cứu các quy tắc *cấu tạo từ, hình thái của từ và từ loại*. Cú pháp học nghiên cứu các quy tắc về *kết hợp từ* thành những *đơn vị lớn hơn từ* (cụm từ, câu), và đặc điểm, chức năng của chúng.

3. Phân tích ngữ pháp của một ngôn ngữ là nhằm chỉ ra cơ cấu tạo nên hệ thống các quy tắc ngữ pháp. Với lập luận là: Từ một tập hợp (corpus) đủ lớn các mẫu câu khác nhau lấy từ các thể loại văn bản khác nhau, dựa vào các thành tựu nghiên cứu về ngữ pháp và ngữ nghĩa của tiếng Việt, chúng ta có thể phân tích câu tiếng Việt ra thành những đơn vị nhỏ hơn, tiến hành phân loại và mô hình hoá, tổ chức lại thành cơ sở ngữ liệu. Dựa vào cơ sở ngữ liệu đó có thể xây dựng một công cụ phân tích tự động văn phạm tiếng Việt trên máy tính. Công việc cụ thể của quá trình đó được hình dung như sau:

- Dựa vào bảng từ của một cuốn từ điển, xây dựng một bảng từ vựng tiếng Việt có gán nhãn từ loại;

- Căn cứ vào các quy tắc ngữ pháp và các mối quan hệ ngữ nghĩa của tiếng Việt, xây dựng một chương trình trên máy tính có nhiệm vụ phân tích câu trong các văn bản mẫu ra thành những *đơn vị từ vựng*;

- Xây dựng một chương trình trên máy tính có nhiệm vụ sử dụng bảng từ (có gán nhãn từ loại) để *gán nhãn từ loại tự động* trở lại cho đơn vị từ vựng ở các văn bản mẫu;

- Dùng tri thức chuyên gia kiểm tra lại kết quả gán nhãn từ loại tự động. Khôi phục lại văn bản dưới dạng gồm các đơn vị *câu*;

- Phân tích câu đã được gán nhãn từ loại ra thành những đơn vị ngữ pháp nhỏ hơn câu là *ngữ*; phân tích ngữ ra thành những đơn vị nhỏ hơn ngữ là *từ*. Mã hoá chúng dưới dạng mô hình;

- Tổ chức đơn vị *câu* và *ngữ* thành cơ sở ngữ liệu, thống kê và đưa ra mẫu các mô hình câu và mô hình ngữ;

- Từ những đơn vị hữu hạn là *mô hình câu* và *mô hình ngữ*, xây dựng một chương trình phân tích văn phạm tiếng Việt.

Thao tác phân tích ra thành tổ trực tiếp và thao tác mở rộng được áp dụng trong việc phân tích câu. Quy tắc viết lại của Chomsky được áp dụng để miêu tả và mã hoá câu và đơn vị dưới câu.

II. XÁC ĐỊNH ĐƠN VỊ NGỮ PHÁP TIẾNG VIỆT

A. TỪ PHÁP HỌC

Do tiếng Việt là một ngôn ngữ đơn lập, mỗi từ chỉ có một hình thức và không thể biến đổi bằng sự biến dạng hoặc sự phái sinh, nên trong phạm vi của báo cáo, chúng tôi không đề cập đến phân nghiên cứu về hình thái của từ. Chúng tôi chỉ tập trung đề cập tới bộ phận nghiên cứu về từ vựng và từ loại.

1. Từ và từ vựng

Từ vựng là vốn từ của một ngôn ngữ. Vốn từ là tập hợp tất cả các từ và các đơn vị tương đương với từ (cụm từ cố định / ngữ cố định) của một ngôn ngữ. Thông thường, từ vựng được

phản ánh trong từ điển. Từ điển là khoa học tập hợp vốn từ cho những mục đích thực dụng về ngôn ngữ. Từ của tiếng Việt, trong cấu tạo, không có căn tố và phụ tố; trong ngữ nghĩa, không có các ý nghĩa thuộc phạm trù hình thái; trong hoạt động tạo câu, các mối liên hệ ngữ pháp không biểu hiện ở sự biến hình mà biểu hiện bằng trật tự từ.

Trong tiếng Việt, có một đơn vị dễ nhận biết mà trước nay quen gọi “tiếng” hay “chữ”. Gọi là “tiếng” là căn cứ vào ngữ âm, ví dụ: *nói dần từng tiếng một*; gọi “chữ” là căn cứ vào văn tự, ví dụ: *“Chữ tài liền với chữ tai một vần”, câu thơ có bảy chữ*.

- Tiếng là đơn vị phát âm tự nhiên nhỏ nhất, được coi như trùng với *âm tiết*.
- Tiếng là đơn vị nhỏ nhất mang nghĩa, ở góc độ ngữ pháp được gọi là *hình vị*, ở góc độ cấu tạo từ được gọi là *từ tố*.
- Tiếng là đơn vị nhỏ nhất, đóng vai trò làm đơn vị dùng để tạo thành phần câu, được coi như là vai trò của *từ*.

Từ ba nhận xét trên ta có thể phân tiếng trong tiếng Việt ra thành 3 loại sau:

- a) Những tiếng mang ý nghĩa thực như *sông, núi, đi, đứng, nhớ, thương...* có thể độc lập làm thành phần của câu và có đầy đủ tư cách ngữ nghĩa, ngữ pháp thì được gọi là *từ điển hình*.
- b) Những tiếng như *nhưng, mà, tuy, nên...* tuy không độc lập làm thành phần câu nhưng được sử dụng với chức năng tạo thành phần câu và có ý nghĩa ngữ pháp như từ điển hình thì được gọi là *từ công cụ*.
- c) Những tiếng gốc Hán như *son, thủy, gia, bắt...* và những tiếng mờ nghĩa, thường không đứng một mình mà tổ hợp với một tiếng khác như *cộ* (xe cộ), *đẽ* (đẹp đẽ), *vẻ* (vui vẻ)... là những đơn vị có chức năng tạo từ, và có thể lâm thời được sử dụng như từ.

Như vậy, từ tiếng Việt là đơn vị nhỏ nhất có nghĩa hoàn chỉnh và cấu tạo ổn định, dùng để tạo thành phần câu. Từ vừa là đối tượng nghiên cứu của từ vựng-ngữ nghĩa học, vừa là đối tượng nghiên cứu của ngữ pháp học.

- Chỉ có những tiếng loại (a) và loại (b) được coi là từ. Đó là những từ đơn tiết. Từ đơn tiết là đơn vị từ vựng cơ bản của tiếng Việt, có tần số sử dụng cao, nên có khả năng chuyển nghĩa và khả năng tạo từ đa tiết rất lớn.
- Từ đa tiết là từ có trên 1 tiếng, gồm hai loại: từ láy và từ ghép.
 - + Từ láy là những từ có 2 hoặc trên 2 tiếng, được cấu tạo theo dạng thức đặc thù của tiếng Việt, đó là dạng thức hoà phối ngữ âm từ đơn vị đã có. Từ được tạo ra theo dạng láy thường đồng nghĩa (nhưng không hoàn toàn) với đơn vị gốc, do giữa chúng có sự khác nhau ít nhiều về sắc thái ngữ nghĩa hoặc khả năng tổ hợp. Ví dụ: *trắng - trắng trắng; đẹp - dẽm đẹp, bé - be bé*, v.v.
 - + Từ ghép là những từ có 2 hoặc trên 2 tiếng, có quan hệ ghép nghĩa trong cấu tạo. Những từ ghép do hai yếu tố song kết liên hợp lại và có ý nghĩa thuộc cùng phạm trù, thì gọi là *từ ghép đẳng lập* (vd. *thầy trò, giảng dạy, ăn mặc, ...*). Những từ ghép gồm một yếu tố chính làm trung tâm ngữ pháp, ngữ nghĩa và một yếu tố phụ hạn định hoặc bổ sung nét nghĩa khu biệt cho nó, thì gọi là *từ ghép chính phụ* (vd. *dưa lê, dưa gang, đậu đũa, xe lửa, làm duyên, ăn cánh, ...*). Những tiếng ghép lại với nhau nhưng không rõ là ghép nghĩa và cũng không theo quy luật hoà phối ngữ âm nào, để tạo ra những từ *ngẫu kết* thì gọi là *từ ghép ngẫu kết* (vd. *bù nhìn, bỏ kết, bùng nhùng, mặc cả, tắc kè, ...*).
- Ngoài từ đơn, từ láy và từ ghép ra, trong tiếng Việt còn có những tổ hợp từ ổn định về cấu tạo, có nghĩa và được dùng như một đơn vị để tạo thành phần câu. Đó là những *tổ*

hợp từ cố định (cụm từ cố định / ngữ cố định). Tổ hợp từ cố định có hai loại, *thành ngữ* và *quán ngữ*.

+ Thành ngữ là tổ hợp từ cố định đã quen dùng, nghĩa thường không giải thích được một cách đơn giản bằng nghĩa của các yếu tố tạo nên nó cộng lại, mà thường có nghĩa bóng, có tính biểu cảm (vd. *áo gấm đi đêm, ăn hương ăn hoa, ...*). Tuy gọi là "ngữ", nhưng thành ngữ có thể có kết cấu chủ-vị, và chức năng vẫn chỉ là để tạo thành phần câu, như chức năng của từ.

+ Quán ngữ là tổ hợp từ cố định đã dùng lâu thành quen, như những công thức có sẵn, nghĩa có thể suy ra từ nghĩa của các yếu tố tạo thành (vd. *lên lớp, lên mặt, nghĩ cho cùng, nói tóm lại, ...*).

Tóm lại, khi phân tích một văn bản, ta lần lượt thu được những đơn vị ngữ pháp sắp xếp theo thứ bậc thấp dần. Cái đơn vị có thể tìm ra được sau câu là *ngữ*, sau ngữ là *từ*. Từ là chính thể tự nhiên hữu hạn trong ngôn ngữ.

2. Từ loại

Từ tuy là đơn vị hữu hạn, nhưng số lượng có thể lên tới hàng vạn. Mỗi từ tuy có một nét nghĩa riêng, nhưng có thể tìm thấy những nét giống nhau về ý nghĩa khái quát, về khả năng kết hợp với các từ ngữ khác trong câu. Phân loại từ theo đặc điểm về ý nghĩa khái quát và khả năng hoạt động cú pháp, ta sẽ có các từ loại.

Từ loại chỉ ra phạm trù ngữ pháp bao gồm đặc điểm ngữ pháp, quan hệ cú pháp và ý nghĩa khái quát của đơn vị từ vựng. Căn cứ vào từ loại có thể nhận ra được chức năng của đơn vị từ vựng trong hoạt động ngôn ngữ, chẳng hạn chức năng chủ ngữ đối với danh từ, vị ngữ đối với động từ, v.v. Do đặc thù của tiếng Việt có thể có những đơn vị từ vựng chưa xác định được từ loại.

Những đặc điểm có tính chất khái quát nêu trên về từ vựng và từ loại được phản ánh tương đối rõ ràng trong các cuốn từ điển tiếng Việt. Vì vậy, thay vì đi xây dựng một bảng từ vựng từ đầu, chúng tôi dựa vào danh sách từ vựng của một cuốn từ điển tiếng Việt cụ thể, có đưa thêm vào những đơn vị từ vựng mới xuất hiện, và tiến hành gán nhãn (tag) từ loại cho từng đơn vị từ vựng. Chúng tôi cũng dựa vào cách phân chia từ loại trong từ điển và đưa ra danh sách các từ loại cần phải gán nhãn như sau:

từ loại	ý nghĩa từ vựng	quan hệ cú pháp
1. danh từ 2. động từ 3. tính từ 4. đại từ	có ý nghĩa thực (thực từ)	có khả năng làm trung tâm của thành phần câu
5. phụ từ 6. kết từ 7. trợ từ 8. cảm từ	không có ý nghĩa thực (hư từ)	không có khả năng làm trung tâm của thành phần câu

B. CÚ PHÁP HỌC

Trong ngôn ngữ, bên cạnh từ, còn có những đơn vị khác cũng có khả năng hoạt động độc lập như từ. Đó là cụm từ (ngữ)* và câu.

1. Khái lược về ngữ

Trong hoạt động ngôn ngữ, từ có thể một mình làm thành tổ cú pháp, hoặc có thể kết hợp với một số từ khác làm thành tổ cú pháp. Ví dụ:

a) Nó đang đọc sách.	sách một mình làm thành tổ cú pháp
b) Nó đang đọc sách văn học.	sách kết hợp với thực từ làm một thành tổ cú pháp: sách văn học
c) Nó đang nói về sách văn học.	sách kết hợp với hư từ làm một thành tổ cú pháp: về sách văn học

Thông thường, trong thực tế sử dụng ngôn ngữ để giao tiếp, chúng ta ít dùng loại câu có các thành tố cú pháp là một từ, mà chủ yếu là dùng câu có các thành tố cú pháp là ngữ. Chẳng hạn câu:

Những bông hoa trong vườn đang nở thắm
 (1) (2) (3)

Ngữ (1) và (2) mang đặc điểm ngữ pháp của danh từ *hoa* và *vườn*; ngữ (3) mang đặc điểm ngữ pháp của động từ *nở*, mỗi ngữ đều có một chức năng trong câu.

Như vậy, ngữ là một đơn vị cú pháp trung gian giữa từ và câu, có cấu tạo gồm một từ trung tâm liên kết với các thành phần phụ bằng quan hệ chính phụ. Từ trung tâm quy định đặc điểm ngữ pháp và chức năng của toàn kết cấu.

1.1. Cấu tạo của ngữ

Ở dạng đầy đủ, ngữ gồm 3 thành phần: *phần phụ trước* - *trung tâm* - *phần phụ sau*. Ví dụ:

tất cả những bông hoa vừa mới hái ấy
 phần phụ trước trung tâm phần phụ sau

- Trung tâm là thành tố chi phối sự xuất hiện các thành tố phụ trước và sau ngữ. Từ đóng vai trò trung tâm phải là thực từ, chứ không thể là hư từ. Từ trung tâm thuộc từ loại nào thì ngữ sẽ mang đặc điểm ngữ pháp và chức năng của từ loại ấy.

- Phần phụ của ngữ, về mặt ngữ pháp là những thành tố phụ có tác dụng bổ sung ý nghĩa cho từ làm trung tâm. Chúng là kết quả của sự chi phối về đặc điểm ngữ pháp của từ trung tâm và nhu cầu giao tiếp. Ví dụ:

con mèo
 con mèo đen
 một con mèo đen
 một con mèo đen ấy
 hầu hết những con mèo đen ấy

* Từ đây trở đi chúng tôi gọi là ngữ.

Các ví dụ trên cho thấy sự có mặt của thành tố trung tâm trong ngữ là bắt buộc. Ngữ ở dạng đầy đủ gồm 3 phần, nhưng ở dạng khuyết có thể chỉ xuất hiện thêm một trong hai phần phụ.

1.2. Chức năng của ngữ

Ngữ là kết quả của thao tác mở rộng theo quan hệ chính phụ của từ trung tâm. Do đó, ngữ mang đặc điểm ngữ pháp và chức năng của từ trung tâm. Từ trung tâm là danh từ thì ngữ mang đặc điểm và chức năng của danh từ, và gọi là *ngữ danh từ* (danh ngữ). Từ trung tâm là động từ thì ngữ mang đặc điểm và chức năng của động từ, và gọi là *ngữ động từ* (động ngữ). Từ trung tâm là tính từ thì ngữ mang đặc điểm và chức năng của tính từ, và gọi là *ngữ tính từ* (tính ngữ).

1.2.1. Ngữ danh từ

Ở dạng đầy đủ, ngữ danh từ có 3 phần: *phần phụ trước - trung tâm - phần phụ sau*. Phần phụ trong danh ngữ được gọi là *định tố*. Ngữ danh từ có chức năng làm thành tố trong ngữ (vd. Cháu yêu *chú bộ đội*), hoặc làm thành phần câu (*Những dòng sông đỏ nặng phù sa*).

a) Trung tâm của danh ngữ

Trung tâm của danh ngữ là danh từ. Việc xác định trung tâm của danh ngữ về cơ bản là thuận lợi, chỉ khó khăn khi có hai từ đứng liền nhau, một danh từ chỉ đơn vị và một danh từ chỉ nội dung cụ thể của đơn vị. Ví dụ:

hai con dao ấy
những quyển sách này
mười quả cam kia

Có 4 quan điểm để xác định từ trung tâm:

- Quan điểm thứ nhất cho danh từ đứng sau là trung tâm vì xác định nó là trung tâm ngữ nghĩa, đồng thời là trung tâm ngữ pháp của ngữ.
- Quan điểm thứ hai cho cả hai danh từ liên hợp với nhau làm trung tâm ghép của ngữ.
- Quan điểm thứ ba cho rằng ở đây chỉ có một danh từ làm trung tâm, tức cho rằng *con dao*, *quyển sách*, *quả cam* là danh từ trung tâm.
- Quan điểm thứ tư cho danh từ đứng trước là trung tâm của ngữ vì nó phù hợp với cách nhìn của người bản ngữ khi nhận thức hiện thực khách quan, cũng như phù hợp với trật tự quan hệ chính-phụ thông thường trong tiếng Việt.

Chúng tôi chọn quan điểm thứ tư, vì:

Nếu chọn quan điểm thứ nhất thì có nhiều trường hợp rất khó xác định ranh giới giữa từ trung tâm với các thành tố phụ của ngữ. Ví dụ: “*khe đá nứt*” thì hiểu là *đá nứt ra thành khe* hay *khe có toàn là đá nứt*? Việc hiểu như thế nào sẽ quyết định cách miêu tả.

Nếu chấp nhận quan điểm thứ hai thì sẽ mâu thuẫn với quan điểm xác định trung tâm của ngữ thường là do một từ hoặc trên một từ có quan hệ đẳng lập với nhau cùng đảm nhiệm (vd. *Hà Nội*, *Hải Phòng* và nhiều *thành phố* khác).

Nếu chấp nhận quan điểm thứ ba thì phải chấp nhận *con dao*, *quyển sách*, *quả cam* là từ. Điều này sẽ dẫn đến quan niệm lại về cấu tạo từ của tiếng Việt.

b) Định tố của danh ngữ

- Định tố trước của danh ngữ thường là những từ chỉ lượng, chia làm 2 nhóm. Nhóm 1 gồm các đại từ và danh từ chỉ tổng số: *tất cả, cả thấy, cả, toàn bộ...* Nhóm 2 gồm các danh từ chỉ số và phụ từ chỉ lượng: *một, những, các, mọi, vài, mỗi, từng...*

- Định tố sau của danh ngữ tương đối phức tạp, nó có thể là một từ, một ngữ, một kết cấu chủ-vị. Ví dụ:

cột <i>tre</i>	(định tố là từ)
cột <i>bằng tre</i>	(định tố là ngữ)
cột <i>tre gãy hôm qua ấy</i>	(định tố là một kết cấu chủ-vị)

Định tố sau có tác dụng hạn định loại cho trung tâm, thường là danh từ không đếm được hoặc danh ngữ có trung tâm là danh từ không đếm được (vd. một cân *thịt*, một cân *thịt nạc vai...*). Định tố sau có tác dụng hạn định đặc trưng cho trung tâm, thường là động từ, động ngữ, tính từ, tính ngữ (vd. nhân viên *bảo vệ*, nhân viên *bảo vệ sân bay*, màu *vàng*, màu *vàng no ấm...*). Định tố sau nhằm xác minh cho trung tâm, thường có cấu tạo là một kết cấu có quan hệ từ hoặc kết cấu chủ-vị (vd. sách *mẹ mua hôm qua...*). Định tố sau nhằm chỉ định cho trung tâm thường là đại từ chỉ định và thường nằm ở vị trí cuối cùng của ngữ danh từ (vd. cái con người *bạc ác ấy...*).

1.2.2. Ngữ động từ

Ở dạng đầy đủ ngữ động từ cũng có 3 phần: *phần phụ trước - trung tâm - phần phụ sau*. Phần phụ trong động ngữ được gọi là *bổ tố*.

a) Trung tâm của động ngữ

Trung tâm của động ngữ là động từ. Việc xác định trung tâm của động ngữ, nói chung, là tương đối dễ dàng. Chỉ khó khăn khi xác định trong trường hợp có hai động từ đứng liền nhau, ví dụ: *ngồi xem phim, định đọc sách...* Có nhiều quan điểm khác nhau về trường hợp này.

- Quan điểm thứ nhất cho động từ đứng sau là trung tâm, vì cho rằng động từ đứng sau là trung tâm ngữ pháp của ngữ. Động từ đứng trước được coi là không hoạt động độc lập, hoặc nếu có hoạt động độc lập thì chỉ bổ sung một ý nghĩa nào đó cho hoạt động chính được biểu thị ở động từ đứng sau. Chẳng hạn: khi đang nằm xem phim, có người hỏi: *Anh đang làm gì đấy?* thì có thể trả lời: *đang nằm xem phim*, hoặc *đang ngồi xem* đều được. Như vậy, trọng tâm thông báo là “*đang xem phim*”, chứ không phải là “*ngồi xem phim*” hay “*nằm xem phim*”.

- Quan điểm thứ hai xác định trung tâm là động từ đứng trước, vì cho rằng nó phù hợp với cảm nhận của người bản ngữ và phù hợp với trật tự quan hệ chính-phụ thông thường trong tiếng Việt. Chúng tôi chấp nhận quan điểm thứ hai, vì:

Nếu theo quan điểm thứ nhất thì khó lí giải được các trường hợp sau:

cần	làm		cần	tiền	muốn	ăn		muốn	cam
↓	↓		↓	↓	↓	↓		↓	↓
BT	TT		TT	BT	BT	TT		TT	BT

(TT: trung tâm; BT: bổ tố)

Chấp nhận quan điểm thứ hai sẽ thuận lợi trong thao tác phân tích ngữ động từ.

b) Bổ tố của động ngữ

- Bộ tổ trước của động ngữ thường do những phụ từ chỉ tình thái đảm nhận. Gồm:

Phụ từ chỉ sự cầu khiến: *hãy, đừng, chớ...*

Phụ từ chỉ sự khẳng định hay phủ định: *có, không, chưa, chẳng...*

Phụ từ chỉ thời gian: *đã, từng, đang, sẽ, sắp...*

Phụ từ chỉ sự so sánh: *cũng, vẫn, cứ, còn, luôn, luôn luôn, mãi, mãi mãi...*

Phụ từ chỉ mức độ: *rất, hơi, hết sức...*

- Bộ tổ sau của động ngữ, về số lượng là không hạn chế, vì do nhu cầu giao tiếp chi phối. Tuy nhiên, do sự chi phối về đặc điểm ngữ pháp của động từ trung tâm, sự xuất hiện một số bộ tổ và vị trí xuất hiện của chúng là xác định được. Ví dụ:

(Bộ đội) kéo *pháo*.

sang *sông*

(Tôi) hiểu *những điều anh nói*.

có *một mùa hoa cải*.

bị *cám cúm*.

dạy *con học hát*

gửi *lại cho anh một nửa vầng trăng*, v.v...

Cũng như định tổ sau trong danh ngữ, bộ tổ sau trong động ngữ cũng có thể là *từ, ngữ, cụm chủ-vị*, thậm chí là một *liên hợp chủ-vị*. Ngữ động từ có chức năng làm thành tổ trong ngữ (vd. Cầu thủ *đoạt giải quả bóng vàng* là một người Braxin), và làm thành phần câu (*Chết vinh còn hơn sống nhục*).

1.2.3. Ngữ tính từ

Ở dạng đầy đủ ngữ tính từ có 3 phần: *phần phụ trước - trung tâm - phần phụ sau*. Phần phụ trong tính ngữ cũng được gọi là *bổ tổ*.

a) Trung tâm của tính ngữ

Trung tâm của tính ngữ là tính từ (vd. rất *sành* âm nhạc, *giỏi* hùng biện, *xanh* một màu xanh hi vọng). Việc xác định trung tâm của tính ngữ có khó khăn khi thành tổ trung tâm của tính ngữ có liên quan đến thành tổ trung tâm của ngữ động từ. Ví dụ:

bình tĩnh *bám vào* và *bám vào* *bình tĩnh*

hăng hái *tiến công* và *tiến công* *hăng hái*

- Quan điểm thứ nhất căn cứ theo trật tự của quan hệ chính phụ trong tiếng Việt và cho rằng, thành tổ đứng trước là trung tâm.

- Quan điểm thứ hai căn cứ về mặt ngữ nghĩa và cho rằng, tính từ *bình tĩnh*, *hăng hái* trong các tổ hợp trên chỉ có tác dụng bổ nghĩa cho động từ *bám* và *tiến công*. Cho dù vị trí của các thành tố có thể thay đổi nhưng quan hệ ngữ pháp và ý nghĩa vẫn không thay đổi.

- Quan điểm thứ ba xử lý tương tự như quan điểm thứ hai, nhưng lại thừa nhận có sự thay đổi về ngữ nghĩa khi vị trí các thành tố thay đổi.

Chúng tôi chấp nhận quan điểm thứ nhất.

b) Bộ tổ của tính ngữ

- Bộ tổ trước của tính ngữ cũng giống như bộ tổ trước của động ngữ. Tuy nhiên cần lưu ý thêm một số đặc điểm sau:

+ Hầu hết các tính từ đều có khả năng kết hợp với các phụ từ chỉ mức độ, và khả năng xuất hiện của phụ từ loại này là rất thường xuyên (vd. *rất* anh hùng, *hơi* béo, *hết sức* thông minh...).

+ Chỉ có một số tính từ là có khả năng kết hợp được với các phụ từ chỉ mệnh lệnh, cầu khiến (vd. *chớ* dại dột thế, *hãy* dừng cảm lên, *đừng* xanh như lá bạc như vôi).

- Bỏ tố sau của tính ngữ: Tính từ không chi phối sự xuất hiện số lượng các bỏ tố sau, mà số lượng bỏ tố sau phụ thuộc vào nhu cầu và mục đích giao tiếp. Cũng giống như động ngữ, bỏ tố sau của tính ngữ có thể là *từ*, *ngữ*, *kết cấu chủ-vị*, *liên hợp kết cấu chủ-vị*.

Ngữ tính từ có chức năng làm thành tố trong ngữ (vd. *đỏ rực một màu lửa*), và làm thành phần câu (*Hèn nhát như thế* là điều không thể tưởng tượng nổi, *Biển bạc đầu thương nhớ*).

2. Khái lược về câu

Câu là đơn vị cơ bản của lời nói, do từ hoặc ngữ tạo thành, có ngữ điệu nhất định, diễn đạt một ý trọn vẹn. Trong hoạt động lời nói, ít khi chúng ta sử dụng câu có thành phần câu là từ, mà chủ yếu là ngữ. Câu do các ngữ tạo thành gồm có một nòng cốt và các thành phần phụ bổ sung cho nòng cốt. Nòng cốt câu gồm hai thành phần chính, chủ yếu là *chủ ngữ* và *vị ngữ*. Trong thực tế sử dụng, câu có thể có đầy đủ thành phần, hoặc có thể được rút gọn. Thông thường là câu một nòng cốt đơn có phần đề và phần thuyết. Tuy nhiên, do nhu cầu trong quá trình tư duy, giao tiếp mà câu có thể có cấu tạo đơn giản hay phức hợp, có nòng cốt đơn hay nòng cốt ghép.

2.1. Các thành phần chính của câu

2.1.1. Chủ ngữ

Chủ ngữ là một trong hai thành phần chính yếu của một câu đơn thông thường, nêu đối tượng mà hành động, tính chất, trạng thái sẽ được nói rõ trong vị ngữ. Như vậy, về mặt ngữ pháp, chủ ngữ là thành phần chi phối sự xuất hiện của vị ngữ. Trong tiếng Việt, sự chi phối ấy thể hiện bằng trật tự chủ - vị. Về mặt ý nghĩa, chủ ngữ là cái được thông báo, còn gọi là phần *đề*.

Chủ ngữ trong tiếng Việt rất đa dạng. Nói chung, tất cả các kết cấu ngữ pháp đều có thể trực tiếp làm chủ ngữ. Tuy nhiên, do đặc trưng là phân nêu đối tượng, nên phần lớn câu tiếng Việt là do danh ngữ đảm nhiệm. Về mặt thông báo, do chủ ngữ thường là cái đã biết sẽ được nói rõ trong vị ngữ, nên trong những tình huống giao tiếp cụ thể, nó có thể được rút bớt cho gọn. Ví dụ: *Có gì đâu mà phải sợ! Vả lại làm quan mà không ăn lộc, thì ai làm quan làm quái gì?*

2.1.2. Vị ngữ

Cũng như chủ ngữ, vị ngữ là thành phần chính yếu của một câu đơn, nói rõ hành động, tính chất, trạng thái của đối tượng được nêu ở chủ ngữ. Trong tiếng Việt, về mặt ngữ pháp, vị ngữ thường đứng sau chủ ngữ. Về mặt thông báo, do vị ngữ là phần nêu rõ cái được nói tới ở chủ ngữ, nên còn được gọi là phần *thông báo* hoặc phần *thuyết* (thuyết minh cho phần đề). Tuy nhiên, trong những tình huống sử dụng ngôn ngữ cụ thể, do mục đích dụng pháp, mà có sự thay đổi về trật tự quan hệ của chủ-vị. Ví dụ:

Từ xa *tiến lại* một người cao to, vạm vỡ.
Nhớ nước đau lòng con cuộc cuộc

Thương nhà mỗi miệng cái gia gia.

Vị ngữ trong tiếng Việt cũng rất đa dạng, tất cả các kết cấu ngữ pháp đều có khả năng đảm nhiệm thành phần này, nhưng phổ biến vẫn là do động từ hoặc ngữ động từ và tính từ hoặc ngữ tính từ đảm nhiệm. Các kết cấu ngữ pháp khác khi đảm nhiệm chức năng vị ngữ thường phải có điều kiện, chẳng hạn phải có mặt các từ chỉ quan hệ. Ví dụ:

Cháu là cháu cứ nói.
Chúng nó thì vợ chồng gì.
Chồng gì anh. Vợ gì tôi.

2.2. Các thành phần phụ của câu

2.2.1. Trạng ngữ

Trạng ngữ là thành phần phụ quan trọng nhất trong câu, biểu thị ý nghĩa tình huống như thời gian, địa điểm, nguyên nhân, mục đích, phương tiện... cho thông báo của câu. Vị trí của trạng ngữ nằm ở đầu câu, giữa câu và cuối câu. Nhưng phổ biến nhất là nằm ở đầu câu. Ví dụ:

Hiện nay tôi đang ở Hà Nội.
Tàu hiện đang đỗ ở ga Hà Nội.
Hai cậu bé đang tiến lại từ đằng xa.

Vai trò của trạng ngữ thường chỉ liên quan đến toàn câu hoặc phần chủ ngữ. Đảm nhiệm chức năng trạng ngữ thường do phụ từ (trạng từ) hay tính từ.

2.2.2. Khởi ngữ

Trong tiếng Việt có một thành phần phụ khá đặc biệt, luôn nằm ở trước một nòng cốt, và được gọi là *khởi ngữ*, *khởi ý* hay *đề ngữ*. Về ý nghĩa, khởi ngữ thường là thành phần nêu lên một ý mở đầu. Giá trị thông báo được tập trung ở thành phần đó. Về cấu tạo, thành phần khởi ngữ có thể được đưa trở lại làm phụ tố cho một thành phần trong nòng cốt, hoặc được lặp lại trong nòng cốt bằng chính nó hay bằng đại từ. Ví dụ so sánh:

Tám áo ấy, con thường vẫn mặc - Con thường vẫn mặc tám áo ấy.
Nhà, ông có hàng dây ở phố - Ông có hàng dây nhà ở phố
Giàu, tôi đã giàu rồi.
Thần tốc và táo bạo, đó chính là khẩu hiệu tiến công của quân đội ta.

2.2.3. Hô ngữ

Hô ngữ là thành phần dùng để than gọi, thường do đại từ xưng hô và danh từ riêng kết hợp với một cảm từ, hoặc do các từ khác như *bầm, thưa, kính, này...* tạo thành. Ví dụ:

Người ơi, người ở đừng về.
Em ạ, Cuba ngọt lịm đường.
Bầm cụ, Cụ cho cho gọi con ạ !
Này, anh nói gì thế ?

Vị trí của hô ngữ có thể ở đầu câu, giữa câu và cuối câu. Giữa hô ngữ và các thành phần khác của câu phải có quãng ngắt khi nói và có dấu phẩy khi viết. Ví dụ:

Anh em ơi, vì nhân dân quên mình.
Tình dậy em ơi, qua rồi cơn ác mộng.

2.2.4. Thành phần chú thích

Thành phần chủ thích dùng để giải thích thêm một thành phần trong nòng cốt, hoặc cho một yếu tố của thành phần đó, hoặc bổ sung một ý nghĩa tình thái nào đó cho cả câu. Ví dụ:

Người lớn – *chắc chắn rồi* – luôn luôn đúng.

Ngày tôi sinh, *ngày 19 tháng 8*, là một ngày rực nắng.

Đẹp quá, một đàn cò trắng đang bay qua đồng.

Vị trí của thành phần chủ thích thường nằm giữa nòng cốt, hoặc đứng trước nòng cốt.

2.2.5. Thành phần chuyển tiếp

Thành phần chuyển tiếp dùng để dẫn vào nội dung thông báo với tác dụng tiếp ý phần trước, hoặc với tác dụng đưa đẩy. Thành phần này thường do các quán ngữ đảm nhận. Do đóng vai trò chuyển tiếp trong câu, nên thành phần này thường đứng ở đầu câu đơn, đôi khi xen vào giữa. Ví dụ:

Tóm lại, chúng ta cứ thực hiện theo kế hoạch đã định.

Nhìn chung, nên nhìn nhận lại tất cả những gì đã xảy ra.

Dù sao đi nữa, hấn vẫn quyết tâm thực hiện ý định của mình.

Tương là tốt, *trái lại*, ngày càng tồi hơn.

2.3. Phân loại câu

Căn cứ theo cấu tạo ngữ pháp của câu để phân loại câu trong Việt là một hướng phân loại quan trọng. Cách phân loại như vậy dẫn đến việc dễ dàng nhận ra cấu tạo ngữ pháp của câu, và quan trọng hơn, là có thể tạo được câu theo mô hình cho trước. Câu tiếng Việt nhìn chung được phân thành ba loại: *câu đơn*, *câu phức* và *câu ghép*.

2.3.1. Câu đơn

Câu đơn là câu được cấu tạo bằng một kết cấu chủ-vị, hay là một “nòng cốt đơn”, còn được gọi là *câu đơn bình thường*. Ví dụ:

Khi mẹ có hai khi con.

CN VN

Bữa nọ người ta rượt đuổi khi mẹ.

CN VN

Câu đơn mà nòng cốt chỉ do một ngữ tạo thành thì gọi là *câu đơn đặc biệt*. Ví dụ:

Vợ với chả con!

Chết thật! (Ai xui dại nó thế không biết).

Do các ngữ khi trở thành câu đơn đặc biệt phải phụ thuộc vào bối cảnh giao tiếp và mục đích thông báo, nên không thể xác định được đâu là chủ ngữ, đâu là vị ngữ như ở câu đơn bình thường. Vì vậy, khi miêu tả, chúng tôi miêu tả theo cấu trúc và quan hệ của ngữ.

Câu đơn bình thường, do bối cảnh giao tiếp và mục đích thông báo cho phép, nhờ quy luật tiết kiệm của ngôn ngữ, có thể lược bớt đi một thành phần câu thì gọi là *câu đơn rút gọn*. Ví dụ:

Thế nào, xong đám cưới rồi chứ !

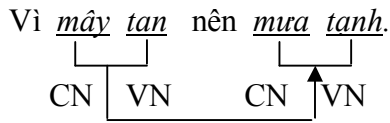
Không muốn ăn à ?

Không ăn thì uống vậy !

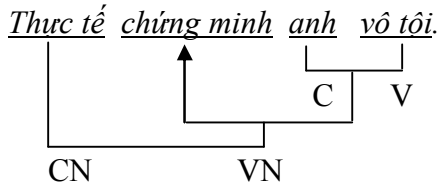
Khác với câu đơn đặc biệt, câu đơn rút gọn có thể xác định được đâu là chủ ngữ, đâu là vị ngữ. Khi miêu tả, chúng tôi miêu tả như câu đơn bình thường, nếu thành phần câu nào được rút gọn thì để trống giá trị.

2.3.2. Câu phức

Câu phức là câu có trên một kết cấu chủ-vị, mà ngoài kết cấu chủ-vị nòng cốt còn có ít nhất một kết cấu chủ-vị (C-V) làm thành một vế thuộc câu phụ, vd:



hoặc làm phần phụ trong một ngữ thuộc một thành phần câu, vd:



Câu phức được biểu hiện qua các kết cấu ngữ pháp có quan hệ chính phụ. Về ý nghĩa, nó biểu thị một phán đoán phức hợp, một suy lí suy ra từ trật tự lôgic. Nếu thay đổi trật tự lôgic ấy thì nội dung suy lí sẽ bị thay đổi. So sánh: Khi nói “*Mây tan, mưa tạnh*” thì “*mây tan*” là nêu nguyên nhân dẫn đến kết quả là “*mưa tạnh*”. Khi đảo thành “*Mưa tạnh, mây tan*” thì nội dung của câu đã bị thay đổi và không còn tính chất “suy lí” nữa. Tức là dẫn đến một nội dung phi lí so với nhận thức thông thường: *kết quả dẫn đến nguyên nhân*.

Cần phân biệt câu phức với câu đơn bình thường có các thành phần phụ, ví dụ:

Khi có tiền, tôi sẽ mua xe.

Làm đúng như thế, ông sẽ thưởng

Những câu kiểu như: *Thế thì thôi* ; *Vậy cũng được...* được coi là câu phức đặc biệt. Khi miêu tả, sẽ gặp rất nhiều khó khăn, có thể cần phải nghiên cứu thêm.

2.3.3. Câu ghép

Câu ghép được biểu hiện bằng hai kết cấu chủ-vị trở lên có quan hệ đẳng lập với nhau. Thực chất là liên hợp các kết cấu chủ-vị thuộc bậc câu với câu. Về quan hệ ngữ nghĩa giữa các kết cấu chủ-vị có thể là quan hệ liệt kê, quan hệ nối tiếp, ... Dấu hiệu phân biệt được biểu hiện bằng quan hệ từ, hoặc dấu câu. Ví dụ:

Anh đến tôi hay tôi đến anh.

Con cá, yêu nước ; con chim cá, yêu trời.

Câu ghép cũng có hình thức gây cảm giác giống với câu đơn bình thường có nhiều vị ngữ, nên cần phải chú ý phân biệt. Ví dụ:

Ông bình tĩnh bám vào khe đá và leo lên nhẹ nhàng.

Chị yêu chồng, thương con hết mực

Những câu kiểu như: *Nghèo nhưng vui* ; *Không chồng thì vợ ...* được coi là câu ghép đặc biệt, và cũng gặp khó khăn khi miêu tả giống như câu phức đặc biệt.

Kết luận

Trên đây chúng tôi đã trình bày hướng phân tích câu tiếng Việt ra thành những *đơn vị ngữ pháp* và những *kết cấu ngữ pháp*. Có thể áp dụng *quy tắc cấu trúc ngữ đoạn* của Ngữ pháp

tạo sinh để miêu tả câu tiếng Việt. Tổ chức câu đã miêu tả thành cơ sở ngữ liệu để rút ra danh sách mô hình cấu tạo của *câu* và *ngữ* trong tiếng Việt. Từ những mô hình hữu hạn đó có thể xây dựng một chương trình phân tích văn phạm tiếng Việt.

Kết quả phân tích văn phạm đạt chất lượng như thế nào là phụ thuộc vào số lượng câu phân tích mẫu, cũng như phụ thuộc vào thao tác phân tích thành phần câu có chính xác hay không. Tuy nhiên, đi theo hướng chúng tôi đề cập cũng sẽ không tránh khỏi những nhược điểm thường gặp, chẳng hạn sẽ rất khó khăn khi miêu tả một số mẫu câu đơn và câu ghép đặc biệt trong tiếng Việt. Kiểu như:

Gió ! Mưa ! Bão bùng!
Thế thì thôi ! v.v...

Nhìn về xa hơn, khi đã có được một cơ sở ngữ liệu mẫu đủ lớn, cùng với kỹ thuật tin học, hi vọng là công việc mà chúng tôi đang thực hiện sẽ giúp được một phần nào đó làm sáng tỏ quy tắc ngữ pháp của tiếng Việt.

TÀI LIỆU THAM KHẢO

1. Ủy ban Khoa học Xã hội Việt Nam, Ngữ pháp tiếng Việt, Nhà xuất bản KHXH, Hà nội, 1983.
2. Ủy ban Khoa học Xã hội Việt Nam, Viện Thông tin Khoa học Xã hội, Ngôn ngữ học khuynh hướng – lĩnh vực – khái niệm, Tập 1, Nhà xuất bản KHXH, 1984.
3. Nguyễn Tài Cẩn, Từ loại danh từ trong tiếng Việt, Nhà xuất bản KHXH, Hà Nội, 1975.
4. Nguyễn Tài Cẩn, Ngữ pháp tiếng Việt, Nhà xuất bản Đại học Quốc gia, Hà Nội.
5. Nguyễn Thiện Giáp (chủ biên) – Đoàn Thiện Thuật – Nguyễn Minh Thuyết, Dẫn luận ngôn ngữ học, Nhà xuất bản Giáo dục, 1995.
6. Cao Xuân Hạo, Tiếng Việt - mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa, Nhà xuất bản Giáo dục, 1998
7. Cao Xuân Hạo (chủ biên) – Hoàng Xuân Tâm – Nguyễn Văn Bằng – Bùi Tất Tươi, Câu trong tiếng Việt (quyển 1), Nhà xuất bản Giáo dục, 1982.
8. Bùi Tất Tươi (chủ biên) – Nguyễn Văn Bằng – Hoàng Xuân Tâm – Nguyễn Thị Quy – Hoàng Diệu Minh, Giáo trình tiếng Việt, Nhà xuất bản Giáo dục, 1994.
9. Hồ Lê, Cú pháp tiếng Việt (quyển 3), Nhà xuất bản KHXH, Hà Nội, 1993.
10. Nguyễn Văn Hiệp, Các thành phần phụ trong câu tiếng Việt, Luận án phó tiến sĩ khoa học ngữ văn, Hà Nội 1992.
11. Noam Chomsky, Topics in the Theory of Generative Grammar, The Hague – Paris, 1966.
12. Nancy Ide and Jean Véronis, Text Encoding Initiative, Kluwer Academic Publishers (Reprinted from Computer & the Humanities, volume 29, Nos. 1,2 & 3 (1995)). (Bản dịch tiếng Việt của Ngô Trung Việt).

MÃ HOÁ DỮ LIỆU TREEBANK TIẾNG VIỆT

SP 7.3 - Dự án VLSP

Người thực hiện: Lê Hồng Phương, Nguyễn Thị Minh Huyền, Phan Thị Hà
Viết báo cáo: Phan Thị Hà

Nội dung báo cáo

1. Giới thiệu	25
2. Mô hình Meta cho SynAF	26
2.1. Giới thiệu	26
2.2. Các phần tử SynAF	26
3. Định dạng chú giải cho các nhãn cú pháp tiếng Việt theo XML:	28
3.1. Chú giải XML cho các nút kết thúc:	28
3.2. Chú giải XML cho các nút không kết thúc (NT):	28
3.3. Chú giải XML cho các cung (egde):	29
3.4. Danh sách các bộ nhãn sử dụng cho quá trình mã hoá	29
3.5. Các ví dụ minh hoạ	31

1. Giới thiệu

Việc mã hoá kho văn bản được gán nhãn ngôn ngữ phải đạt được các yêu cầu sau đây:

- Dễ chuyển đổi sang các định dạng khác nhau
- Dễ khai thác các thông tin ngôn ngữ đã được đánh dấu
- Dễ bổ sung nhãn ngôn ngữ mới
- Dễ đối sánh với ngôn ngữ khác.

Các yêu cầu này cũng chính là các yêu cầu mà các dự án chuẩn hoá tài nguyên ngôn ngữ của tiêu ban kỹ thuật ISO/TC 37/SC 4 hướng tới.

Trong khuôn khổ SP 7.3, chúng tôi quan tâm đến vấn đề mã hoá kho văn bản được gán nhãn cú pháp được phát triển trong dự án SynAF của ISO/TC 37/SC 4. Dự án này nhằm phát triển mô hình gán nhãn chuẩn XML, dựa trên cơ sở những dự án dự án lớn về ngân hàng cây cú pháp (treebank): Penn Treebank cho tiếng Anh, French treebank cho tiếng Pháp, Negra/Tiger cho tiếng Đức, ISST cho tiếng Ý, Prague Treebank cho tiếng Tiệp, v.v. Nhìn chung trong các dự án này việc gán nhãn cú pháp chủ yếu đều chứa thông tin về cấu trúc thành phần (constituent structure) và cấu trúc phụ thuộc (dependency structure).

2. Siêu mô hình cho SynAF

2.1 Giới thiệu

Siêu mô hình chú giải cú pháp SynAF (SynAF Metamodel) là một sơ đồ biểu diễn đầy đủ mối tương quan giữa chú giải đa tầng với chú giải cú pháp thành phần và phụ thuộc. Sau đây là siêu mô hình chú giải cú pháp được biểu diễn bằng UML:

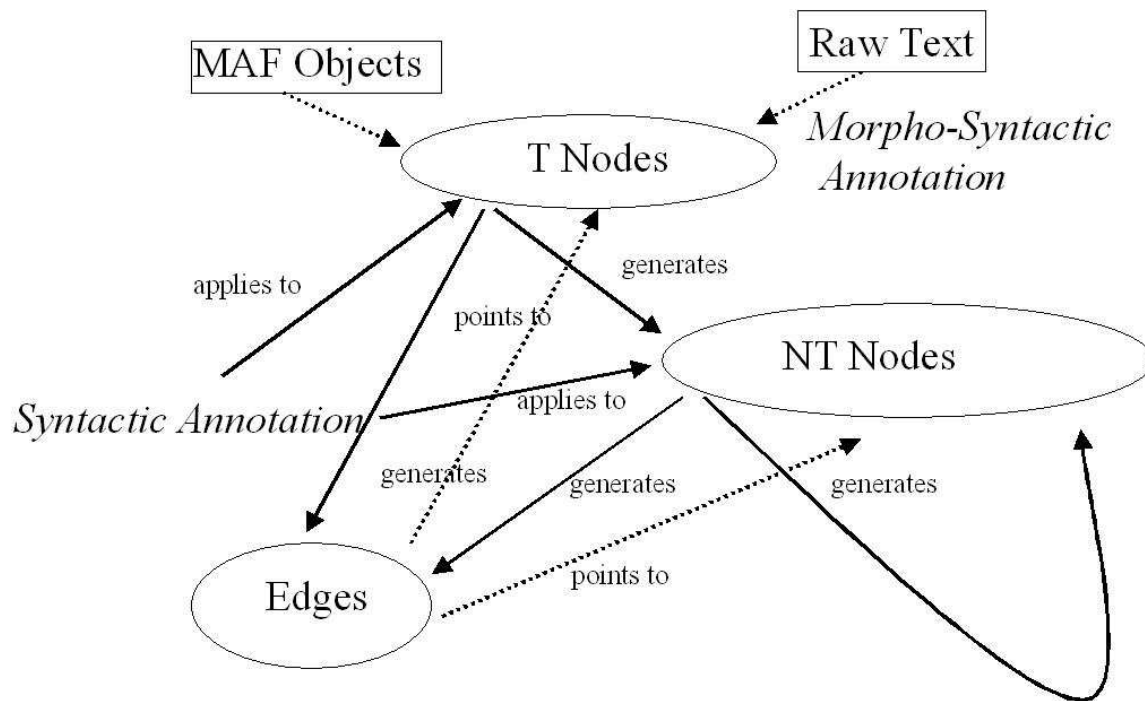


Figure 1: The SynAF metamodel

2.2 Các phần tử SynAF

T Nodes : Biểu diễn các nút kết thúc của cây cú pháp :

- Mỗi nút gồm có các từ đã được chú giải cú pháp , các phần tử rỗng là cho phép.
- Ranh giới giữa các nút T được xác định thông qua một khoảng (span), có thể là nhiều span (dùng cho việc giải thích các thành phần không liên tục).
- Giá trị các phạm trù cú pháp là được gán nhãn ở mức từ.

NT Nodes : Biểu diễn các nút không kết thúc của cây cú pháp :

- Mỗi nút gồm có các nút kết thúc T và không kết thúc NT , phần tử rỗng là cho phép.
- Ranh giới giữa các nút NT cũng được xác định thông qua một khoảng (span), có thể là nhiều span.
- Giá trị các phạm trù cú pháp là được gán nhãn ở mức cụm từ (thành ngữ) và mức cao hơn (mệnh đề, câu).

Edges :

Biểu diễn quan hệ phụ thuộc giữa các nút (cả hai loại nút kết thúc và không kết thúc), đây là quan hệ đôi. Một cung (edges) bao gồm tên nhãn và cặp nút nguồn và nút đích.

Syntactic Annotation (SA) :

Biểu diễn việc đưa các thông tin cú pháp tới đầu vào chú giải- MAF (Morphosyntactic Annotation Framework). SA có thể là một tài liệu hoặc một ứng dụng tự động. Khi chú giải cú pháp được gắn vào các nút (T hoặc NT), khi đó sẽ sinh ra một nút (NT) mới hoặc một cung (egde) phụ thuộc

Danh sách các phạm trù thành phần và phụ thuộc:

Cung cấp danh sách giá cơ bản cho nội dung các nút

Các phạm trù thành phần của SynAF

Label	Meaning	Label	Meaning
AP	adjective phrase	AVP	adverbial phrase
CAC	coordinated adposition	CAP	coordinated adjective phrase
CAVP	Coordinated adverbial phrase	CCP	Coordinated complementiser
CH	Chunk(non-recursive constituent)	CNP	Coordinated noun phrase
CO	coordination	CPP	Coordinated adpositional phrase
CVP	Coordinated verb phras	NP	noun phrase
PN	proper noun Sentence	PP	PP adpositional phrase
S	verbal nucleus with finite tense and all	VP	verb phrase
IBAR	adjoined elements like clitics, adverbs and negation	SV2	infinitival claus
SV3	participial clause	SV5	gerundive clause
FINT	+wh interrogative sentence	F2	relative clause
CP	dislocated or fronted sentential adjuncts	COM	copulative/predicative complement
		PC	and some
and some specific labels for German and Italian			

Các phạm trù phụ thuộc SynAF

Label	Meaning
mod	the word introducing the dependent in a head-modifier relation
cmod, xmod, ncmo	clausal and non-clausal modifiers may (optionally) be distinguished by the use of cmod/xmod, and ncmo respectively, each with the same slots as mod
Subj	the subject in the grammatical relation subjectpredicate
csubj, xsubj, ncsbj	the Grammatical Realtions (RL) csubj and xsubj may be used for clausal subjects, controlled from within, or without, respectively. ncsbj is a non-clausal subje
Dobj	the object in the grammatical relation between a predicate and its direct object
iobj	relation between a predicate and a non-clausal complement introduced by a preposition

dependent	the most generic relation between a head and a dependent dependent (introducer,head,dependent)
-----------	---

3. Định dạng chú giải cho các nhãn cú pháp tiếng Việt theo XML:

Chúng ta dựa vào mô hình SynAFMeta làm cơ sở cho việc định dạng chú giải các nhãn cú pháp tiếng Việt theo XML.

3.1 Chú giải XML cho các nút kết thúc:

Tất cả các nút kết thúc T được biểu diễn bên trong cặp thẻ <terminals></terminals>. Trong đó mỗi một từ nguyên dạng sẽ có một nút T, tương ứng với một thẻ thành phần <t...../>, mỗi thẻ bao gồm các thuộc tính:

- Địa chỉ id: địa chỉ này được định nghĩa ở thuộc tính tar của egd tương ứng.
- Từ nguyên dạng wordform: đây là từ gốc nguyên dạng được lấy từ câu vào.
- Từ loại Pos: Từ loại tương ứng của từ, ví dụ: danh từ, động từ, tính từ.....được lấy từ danh sách nhãn từ loại.

```
<terminals>
<t id="diachi egd 1" wordForm="từ nguyên dạng1" pos="nhãn từ loại" />
<t id="diachi egd 2" wordForm="từ nguyên dạng2" pos="nhãn từ loại" />
.....
<t id="....." ....."" pos="....." />
</terminals>
```

3.2 Chú giải XML cho các nút không kết thúc (NT):

Tương ứng với một câu đầu vào sẽ có nhiều nút (theo sơ đồ biểu diễn cú pháp hình cây). Ở đây chúng tôi dùng cặp thẻ <Nonterminals>...</Nonterminals> để đánh dấu cho việc biểu diễn tất cả các nút NT trong cây cú pháp. Trong đó, mỗi nút NT được biểu diễn bằng một cặp thẻ <nt....các thuộc tính>...E...</nt>. Các thuộc tính biểu diễn cặp thẻ của mỗi một nút bao gồm:

- Địa chỉ ID của nút :được xác định tại phần thuộc tính tar của cung đi tới nút đó , nếu nút đó là nút gốc thì sẽ được xác định tại thuộc tính root của thẻ đồ thị (graphs)
- Nhãn Label của mỗi nút: Chính là tên của một nút (NT) chỉ hạng mục tương ứng được lấy trong danh sách nhãn cụm từ
- E là danh sách các thẻ biểu diễn các cung (Egd) đi ra từ nút NT, tương ứng với mỗi một nút có thể có nhiều thẻ thành phần.

```
<nonterminals>
<nt id="địa chỉ egde1" label="tên nhãn gốc">
<edge id="địa chỉ 1" label="nhãn cụm từ1" tar="địa chỉ đích" />
<edge id="địa chỉ 2" label="nhãn cụm từ2" tar="địa chỉ đích" />
.....
</nt>
<nt id="địa chỉ egd1" label="tên nhãn gốc">
<edge id="địa chỉ 1.1" label="nhãn cụm từ1" tar="địa chỉ đích" />
<edge id="địa chỉ 1.2" label="nhãn cụm từ2" tar="địa chỉ đích" />
.....
```

</nt>

.....

</nonterminals>

3.3 Chú giải XML cho các cung (egde):

Cung (egde) là một đoạn đi từ nút nguồn-nút cha (nút T) đến nút đích (nút T hoặc NT), để biểu diễn mối quan hệ phụ thuộc giữa các nút, mỗi cung được biểu diễn bằng một thẻ thành phần :<edge id="địa chỉ nguồn " label="nhãn cụm từ" tar="địa chỉ đích" />. Các thuộc tính của thẻ <edge/> bao gồm:

- Địa chỉ id : Được đánh tùy ý, tuy nhiên khi đánh địa chỉ nên dùng kí hiệu có liên quan đến nút cha của cung.
- Nhãn Label: Chính là các nhãn được lấy từ danh sách nhãn chức năng, danh sách nhãn phân loại phụ ngữ của động từ, nhãn phần tử rỗng. Có thể có những thẻ không cần có nhãn này (trong trường hợp không phải là nhãn phân loại phụ ngữ của động từ, không muốn cụ thể chi tiết hơn các thông tin đã có trong cây cú pháp).
- Địa chỉ đích tar: được đánh tùy ý, tuy nhiên các cung mà có nút đích là nút dạng NT thì nên đánh kí hiệu có liên quan đến nút nguồn (nút cha).

3.4 Danh sách các bộ nhãn sử dụng cho quá trình mã hoá

Được lấy từ “ Thiết kế tập nhãn cú pháp và hướng dẫn gán nhãn-*Nguyễn Phương Thái, Nguyễn Xuân Lương, Nguyễn Thị Minh Huyền*”. Trong đó: Bộ nhãn từ loại được gán nhãn thành phần trong thuộc tính workForm của thẻ <t...../> cho các nút không kết thúc T. Bộ nhãn cụm từ, nhãn mệnh đề được sử dụng để gán nhãn thành phần vào thuộc tính label của thẻ <nt...../> cho các nút không kết thúc (NT). Còn lại bộ nhãn chức năng cú pháp, bộ nhãn phân loại phụ ngữ của động từ được sử dụng để gán nhãn phụ thuộc cho các cung tại thuộc tính của thẻ <egde...../>. Nhãn rỗng được sử dụng để gán nhãn thành phần cho các nút T và nút NT. Bộ nhãn khác được sử dụng để gán nhãn cho các nút T.

Sau đây là toàn bộ các bộ nhãn:

Nhãn từ loại:

STT	Tên	Chú thích
1	N	Danh từ
2	Nc	Danh từ chỉ loại
3	V	Động từ
4	A	Tính từ
5	P	Đại từ
6	D	Định từ
7	M	Số từ
8	R	Phụ từ
9	S	Giới từ
10	C	Liên từ
11	I	Thán từ
12	T	Trợ từ, tiểu từ, từ tình thái
13	U	Từ đơn lẻ
14	Y	Từ viết tắt

15	X	Các từ không phân loại được
----	---	-----------------------------

Nhãn cụm từ:

STT	Tên	Chú thích
	NP	Cụm danh từ
	VP	Cụm động từ
	AP	Cụm tính từ
	RP	Cụm phụ từ
	PP	Cụm giới từ
	QP	Cụm từ chỉ số lượng
	WHNP	Cụm danh từ nghi vấn (ai, cái gì, con gì, v.v.)
	WHAP	Cụm tính từ nghi vấn (lạnh thế nào, đẹp ra sao, v.v.)
	WHRP	Cụm từ nghi vấn dùng khi hỏi về thời gian, nơi chốn, v.v.
	WHPP	Cụm giới từ nghi vấn (với ai, bằng cách nào, v.v.)

Nhãn mệnh đề:

STT	Tên	Chú thích
	S	Câu trần thuật (khẳng định hoặc phủ định)
	SQ	Câu hỏi
	SE	Câu cảm thán
	SC	Câu mệnh lệnh
	SBAR	Mệnh đề phụ (bổ nghĩa cho danh từ, động từ, và tính từ)
	SF	Câu mà chỉ có thể được giải thích hợp lý dưới quan điểm ngữ pháp chức năng

Nhãn chức năng cú pháp:

STT	Tên	Chú thích
	SBJ	Nhãn chức năng chủ ngữ
	OBJ	Nhãn chức năng tân ngữ trực tiếp
	IO	Nhãn chức năng tân ngữ gián tiếp
	TPC	Nhãn chức năng chủ đề
	PRD	Nhãn chức năng vị ngữ không phải cụm động từ
	LGS	Nhãn chức năng chủ ngữ logic của câu ở thể bị động
	EXT	Nhãn chức năng bổ ngữ chỉ phạm vi hay tần suất của hành động
	TH	Nhãn phân thuyết của câu SF

Nhãn phân loại phụ ngữ của động từ:

STT	Tên	Chú thích
	TMP	Nhãn chức năng phụ ngữ chỉ thời gian
	LOC	Nhãn chức năng phụ ngữ chỉ nơi chốn
	DIR	Nhãn chức năng phụ ngữ chỉ hướng
	MNR	Nhãn chức năng phụ ngữ chỉ cách thức
	PRP	Nhãn chức năng phụ ngữ chỉ mục đích hay lý do

Các nhãn khác:

STT	Tên	Chú thích
	T	Nhãn phần tử rỗng

Các nhãn quy ước trong tài liệu này:

STT	Tên	Chú thích
	.	Nhãn dấu chấm câu, bao gồm: . ? !
	,	Nhãn dấu phẩy
	:	Nhãn dùng cho cả dấu hai chấm và dấu gạch ngang chú thích

3.5 Các ví dụ minh họa

Ví dụ 1: Mã hoá câu trần thuật: Tôi đi học

<pre> graph TD S[S] -- SBJ --> NP1[NP] S -- SBJ --> VP1[VP] S -- SBJ --> P1[.] NP1 --> T1[Tôi] VP1 --> V1[đi] VP1 --> VP2[VP] VP2 --> V2[học] VP2 --> P2[.] </pre>	<pre> <SynAF> <head>...</head> <body> <s id="s1"> <graph root="s1_0"> <nonterminals> <nt id="s1_0" label="S"> <edge id="s1_50" label="SBJ" tar="s1_1" /> <edge id="s1_51" tar="s1_2" /> <edge id="s1_52" tar="t4" /> </nt> <nt id="s1_1" label="NP"> <edge id="s1_53" tar="t1" /> </nt> <nt id="s1_2" label="VP"> <edge id="s1_54" tar="t2" /> <edge id="s1_55" tar="t3" /> </nt> </nonterminals> <terminals> <t id="t1" wordForm="tôi" pos="P" /> <t id="t2" wordForm="đi" pos="V" /> <t id="t3" wordForm="học" pos="V" /> <t id="t4" wordForm="." pos="." /> </terminals> </graph> </s> ... </body> </SynAF> </pre>
--	---

Ví dụ 2: Mã hoá cụm danh từ Quả bóng màu xanh – đã được gán nhãn như sau

(NP (Nu quả) (N bóng) (N màu xanh))	<pre> <s id="s1"> <graph root="s1_0"> <nonterminals> <nt id="s1_0" label="NP"> <edge id="s1_50" tar="t1" /> <edge id="s1_51" tar="t2" /> <edge id="s1_52" tar="t3" /> </nt> </nonterminals> </pre>
---	--

	<pre> <terminals> <t id="t1" wordForm="quả" pos="Nc" /> <t id="t2" wordForm="bóng" pos="N" /> <t id="t3" wordForm="màu xanh" pos="N" /> </terminals> </graph> </s> </pre>
--	---

Ví dụ 3: Ngày mai tôi đi thi

<p>(S (NP-TMP(N Ngày mai) (S (NP-SBJ(P tôi)) (VP (V đi) (R thi)))))</p>	<pre> <s id="s1"> <graph root="s1_0"> <nt id="s1_0" label="S"> <edge id="s1-50 " tar="s1-1" /> <edge id="s1-51 " tar="s1-2" /> </nt> <nt id="s1_1" label="NP"> <edge id="s1-52 lable="TMP"" tar="t1" /> </nt> <nt id="s1_2 "bel="S"> <edge id="s1-56 tar="s1-5> <edge id="s1-57 " tar="s1-6 /> </nt> <nt id="s1_5 label="NP"> <edge id="s1-57 "label="SBJ" tar="t3" /> </nt> <nt id="s1_6 label="VP"> <edge id="s1-57 " tar="t4" /> <edge id="s1-58 " tar="t5" /> </nt> </nonterminals> <terminals> <t id="t1" wordForm="Ngày mai" pos="N" /> <t id="t3" wordForm="tôi" pos="R" /> <t id="t4" wordForm="đi" pos="V" /> <t id="t5" wordForm="thi" pos="R" /> </terminals> </graph> </s> </pre>
--	---

CÔNG CỤ HỖ TRỢ XÂY DỰNG KHO NGỮ LIỆU TREEBANK – SynAF

SP 7.3 - Đề tài VLSP

Xây dựng công cụ: Lê Hồng Phương, Lưu Văn Tăng, Nguyễn Thị Minh Huyền

Viết báo cáo: Lưu Văn Tăng

Nội dung báo cáo

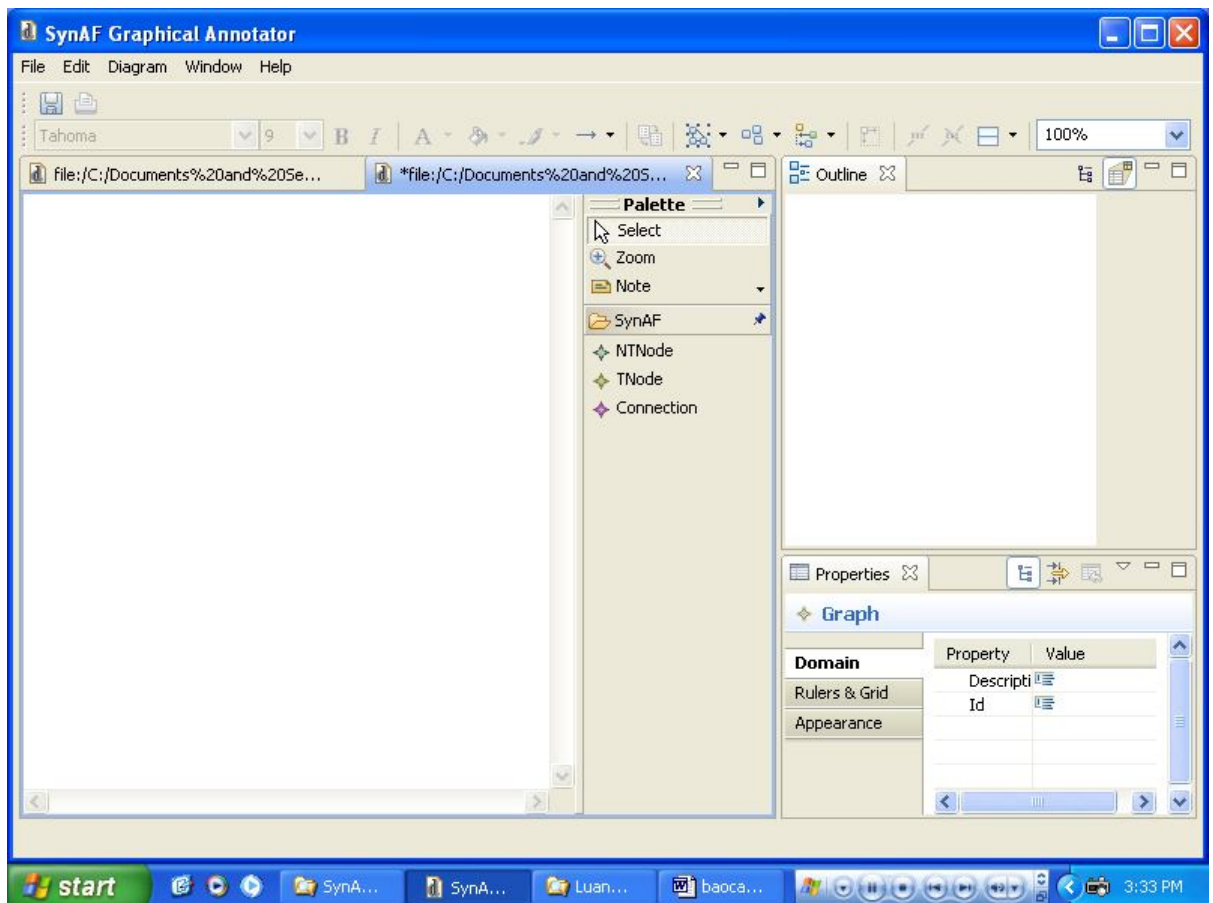
I - Giới thiệu sơ lược	34
II - Mô tả chương trình SynAF	35
III. Hướng dẫn cách sử dụng chương trình.....	37
3.1. Các công cụ chính.....	37
3.2. Sử dụng SynAF.....	39
3.2.1. Vẽ cây	40
3.2.2. Sửa cây cú pháp.....	43

I - Giới thiệu sơ lược

SynAF là bộ công cụ được xây dựng dựa trên nền tảng Eclipse, một môi trường hỗ trợ các công cụ phát lập trình java và phát triển các công cụ cho việc xây dựng các ứng dụng khác.

SynAF có tích hợp nhiều modun cho phép người sử dụng thực hiện được nhiều khả năng xây dựng, chỉnh sửa cây cú pháp một cách mềm dẻo. Với giao diện đồ họa giúp người sử dụng dễ dàng thao tác chỉ với một số động tác kích chuột và nhập từ bàn phím. Các khả năng thực hiện việc xây dựng, chỉnh sửa cây cú pháp nằm trong thanh menu, trên các biểu tượng của thanh công cụ (tool bar).

Dưới đây là giao diện chương trình SynAF:



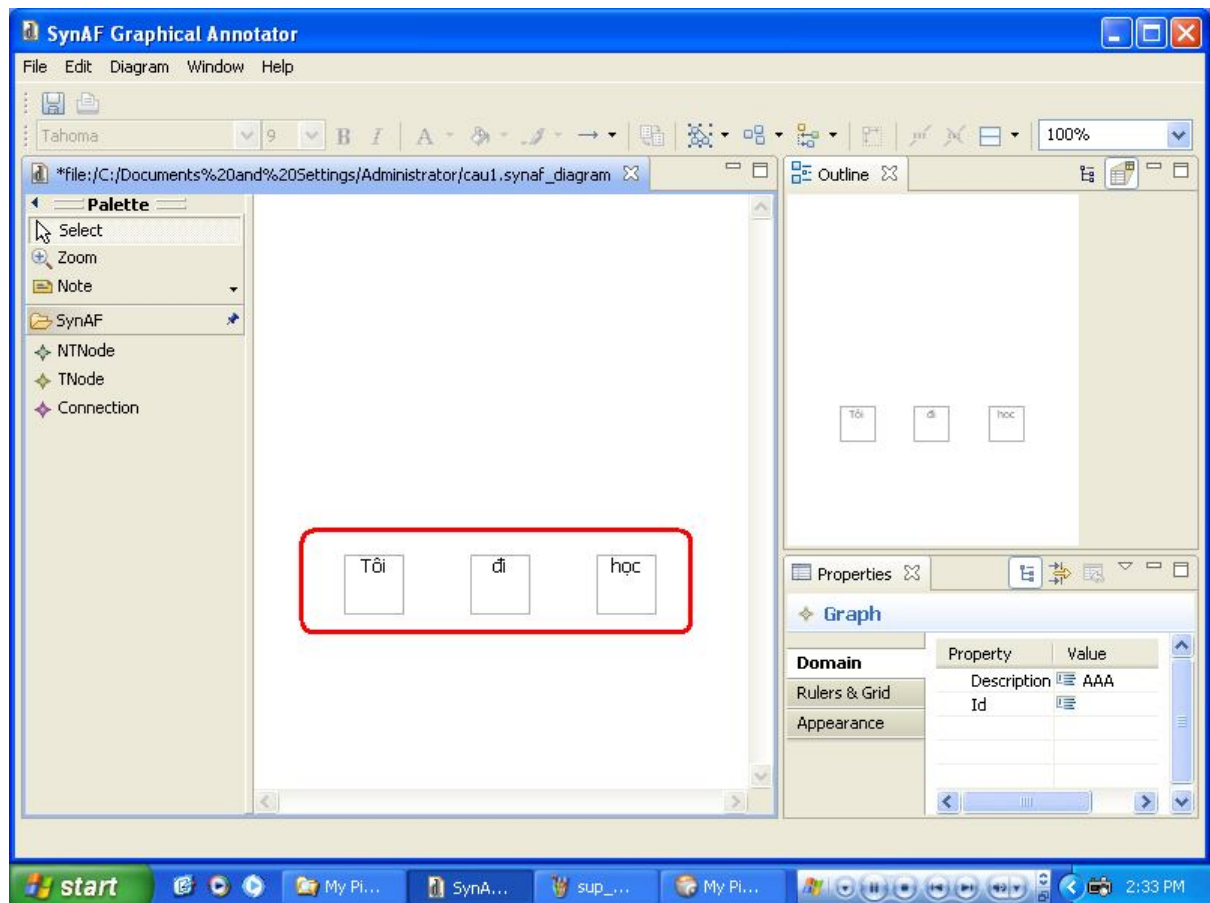
II - Mô tả chương trình SynAF

Chương trình SynAF cung cấp các công cụ xây dựng ngân hàng kho ngữ liệu.

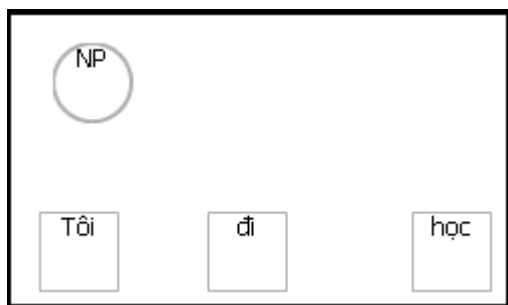
- Dữ liệu đầu vào được hỗ trợ ở một số mức mức sau:

- a. Đầu vào là một file văn bản dạng text chứa tập văn bản (có 4 lựa chọn):
 - Tập văn bản thô
 - Tập văn bản đã được tách từ
 - Tập văn bản đã được gán nhãn từ loại
 - Tập văn bản đã được gán nhãn cú pháp ở mức nông (ngữ)
- b. Đầu vào là một tập văn thô được nhập trực tiếp từ một cửa sổ của bộ công cụ SynAF
- c. Đầu vào rỗng, ta sử dụng các công cụ vẽ trực tiếp các thành phần của cây phân tích cú pháp (sử dụng vùng đồ họa 3 + các công cụ hỗ trợ).

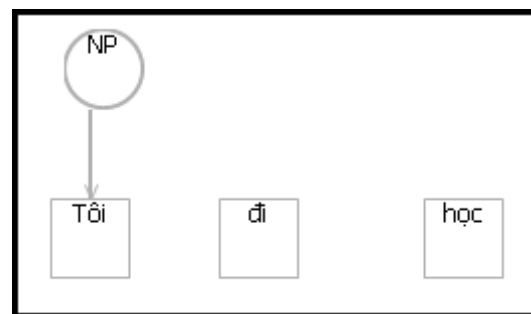
- Sau khi nhập dữ liệu đầu vào (với trường hợp a. và b.), chương trình cho hình ảnh cây phân tích của các câu ở mức lá. Ví dụ câu: “Tôi đi học”, ta có kết quả như sau:



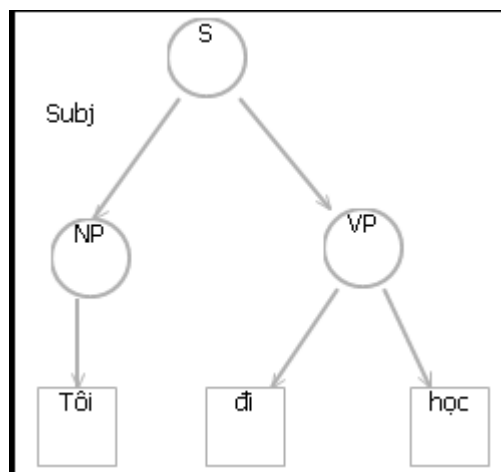
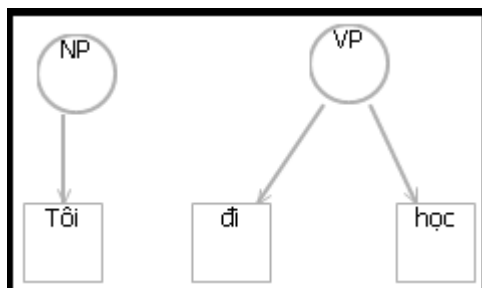
- Người sử dụng (nhà ngôn ngữ học) sẽ tiếp tục thực hiện xây dựng cây cú pháp của câu đó bằng tay:



1)



2)



3)

4)

- Quá trình xây dựng từng bước một cây phân tích cú pháp bằng đồ họa
- 4) là hình ảnh một cây phân tích cú pháp hoàn thiện.

- Đầu ra hay kết quả của quá trình xây dựng cây cú pháp là một tập tin (file) **“tên_project.synaf”** chứa các câu được phân tích về cú pháp biểu diễn dưới dạng XML.

Dưới đây là nội dung của file synaf về các câu được phân tích cú pháp:

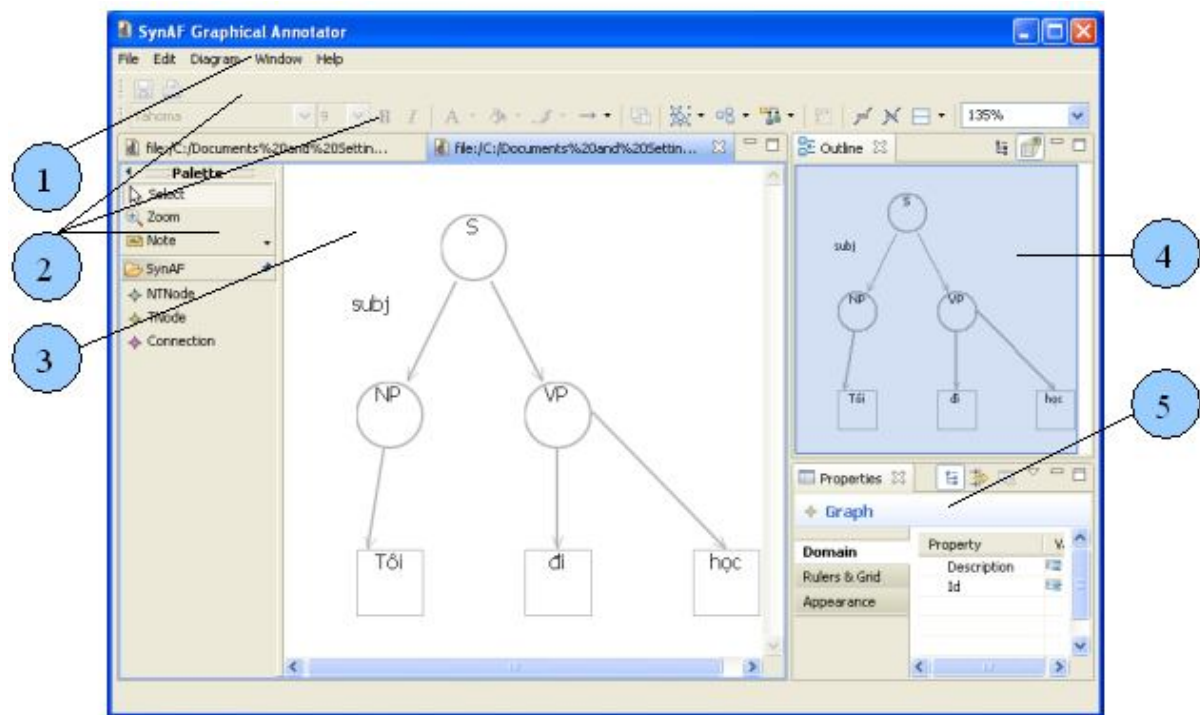
```
<synaf:Graph xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI"
xmlns:synaf="http://www.hus.edu.vn/synaf" description="AAA">
  <nonterminals label="S"/>
  <nonterminals label="NP"/>
  <nonterminals label="VP"/>
  <terminals label="Tôi"/>
  <terminals label="đi"/>
  <terminals label="học"/>
  <edges label="subj" source="//@nonterminals.0" target="//@nonterminals.1"/>
  <edges label=" " source="//@nonterminals.1" target="//@terminals.0"/>
  <edges label=" " source="//@nonterminals.0" target="//@nonterminals.2"/>
  <edges label=" " source="//@nonterminals.2" target="//@terminals.1"/>
  <edges label=" " source="//@nonterminals.2" target="//@terminals.2"/>
</synaf:Graph>

<synaf:Graph xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI"
xmlns:synaf="http://www.hus.edu.vn/synaf" description="BBB">
  .....
  .....
</synaf:Graph>

.....
```

III. Hướng dẫn cách sử dụng chương trình

3.1. Các công cụ chính



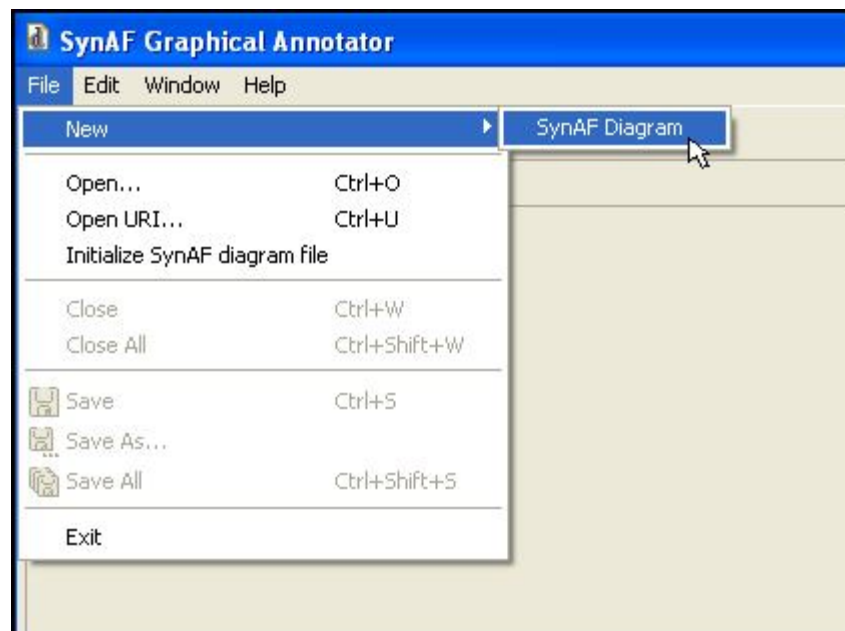
Giải thích:

- ① Thanh menu: chứa các chức năng hỗ trợ xây dựng cây cú pháp
 - ② Các thanh công cụ: chứa các biểu tượng chức năng hỗ trợ xây dựng cây cú pháp.
 - ③ Đồ hoạ cây cú pháp, người dùng có thể chỉnh sửa về các nhãn, hình dáng của cây, zoom, ...
 - ④ Hình previews cây một cách tổng thể
- 5 - Cửa sổ thể hiện các thuộc tính của các thành phần của cây cũng như hình hoạ. Người dùng có thể chỉnh sửa các thuộc tính ở đây.

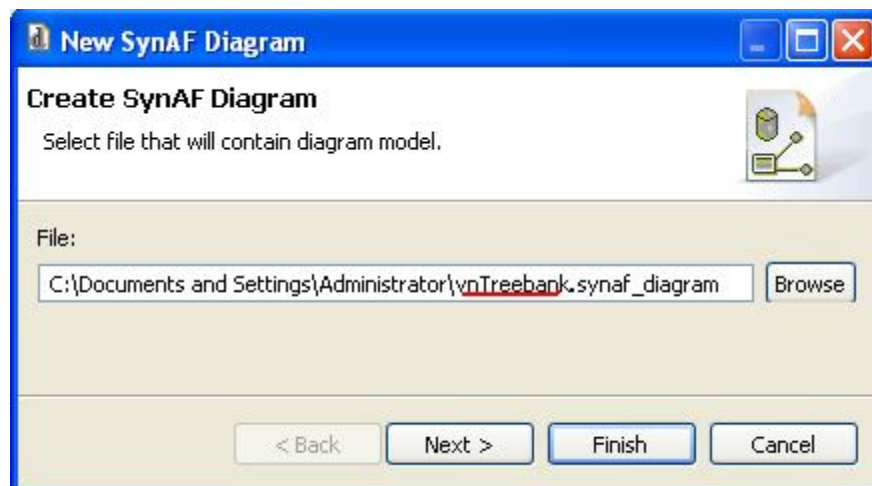
3.2. Sử dụng SynAF

Tạo lược đồ SynAF mới

- File | New | SynAF Diagram



- Đặt tên lược đồ (vnTreebank)



- Kết thúc thao tác tạo lược đồ: Finish

3.2.1. Vẽ cây

- Nhập câu cần “phân tích”

- Nhập từ file chứa các câu (đã được tách từ):

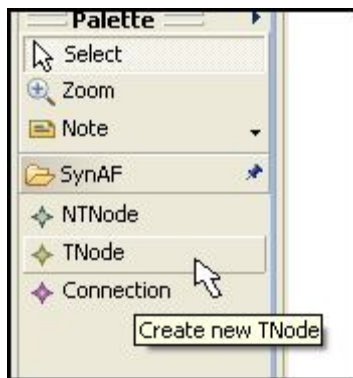
File | Insert | File Sentences Input

- Nhập câu từ hộp thoại:

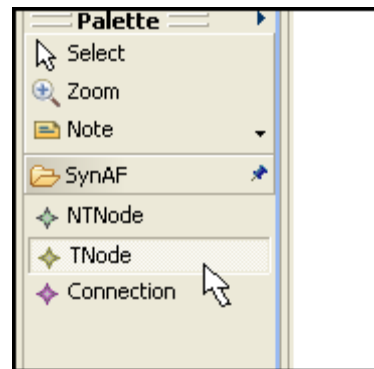
File | Inser | Sentence

- Thao tác tay xây dựng cây cú pháp

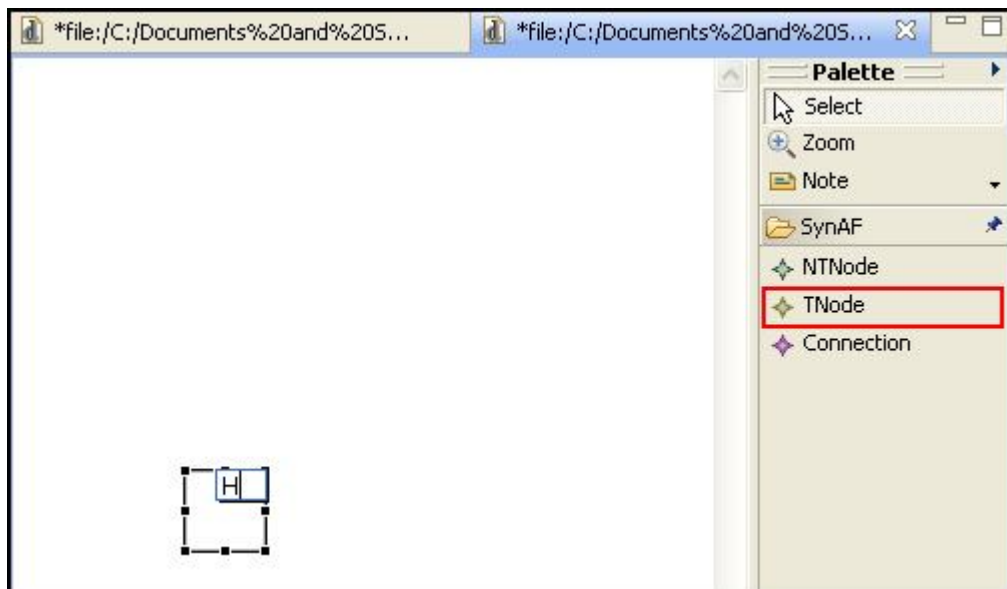
+ Vẽ nút lá (TNode):



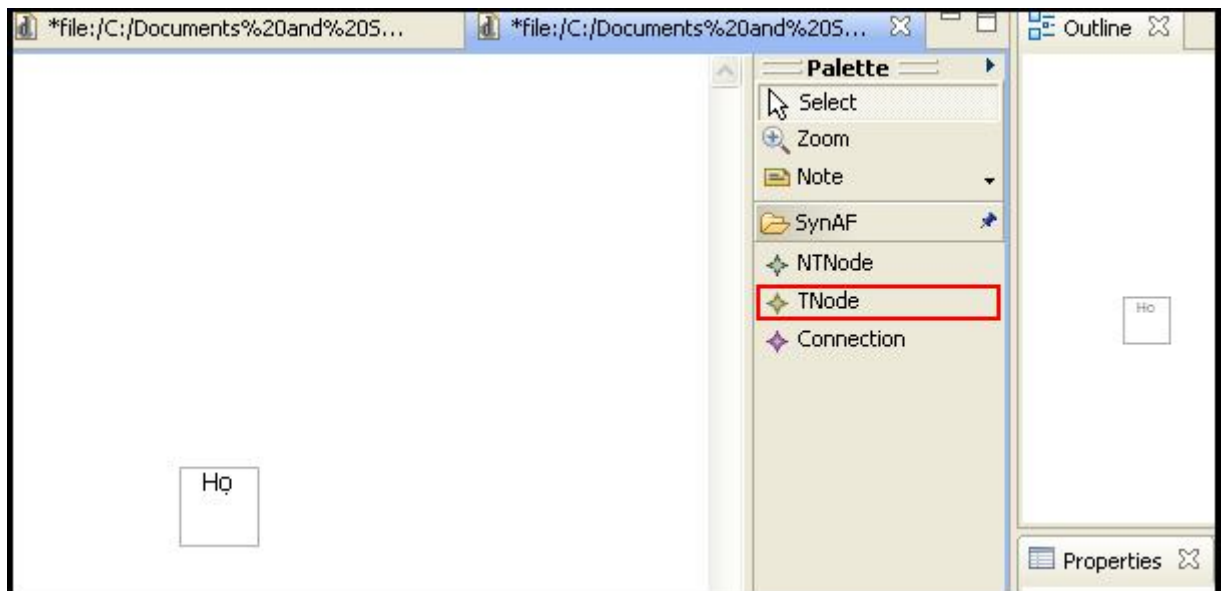
a) Lựa chọn TNode



b) Nhấn chuột

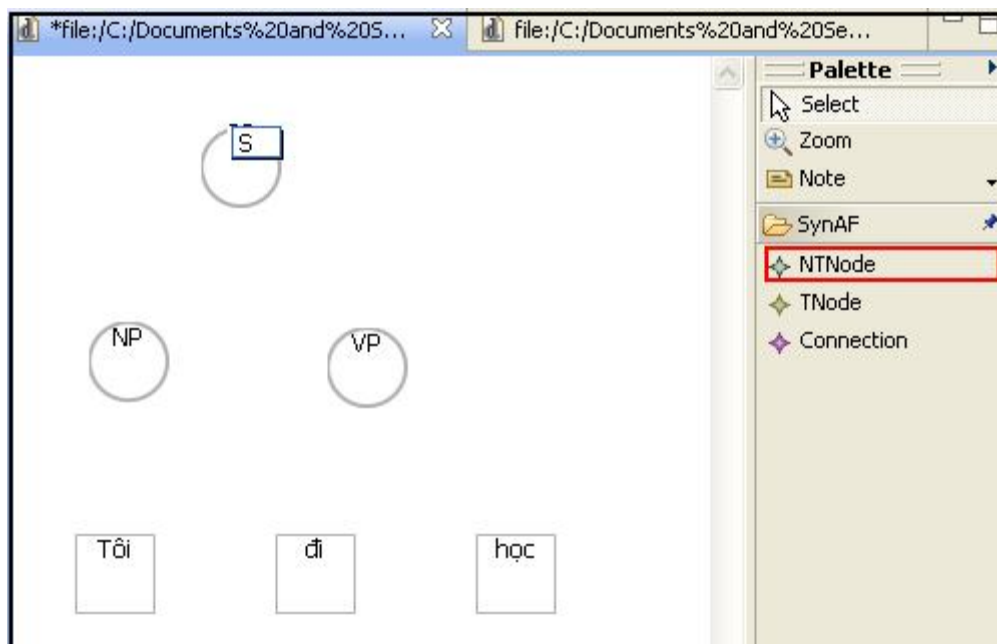


c) Nhấn chuột tại vùng vẽ cây, ghi tên nút



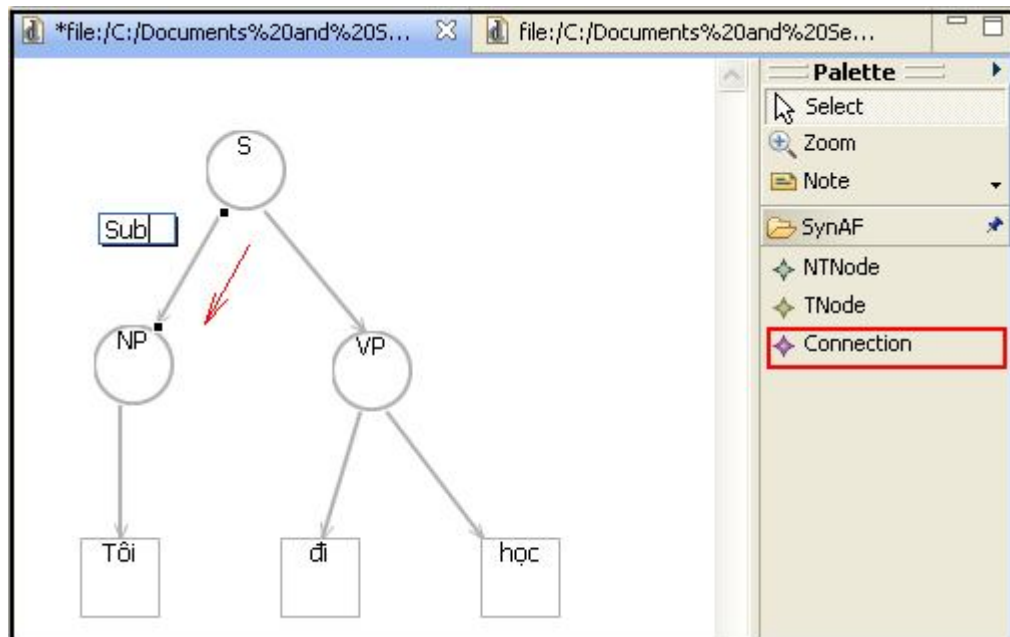
d) Kết thúc

+ **Vẽ nút giữa (NTNote):** Cũng giống như vẽ nút lá, để vẽ nút giữa ta chọn NTNode, rồi nhấn chuột vào vùng vẽ cây, điền tên nút. Ta được:



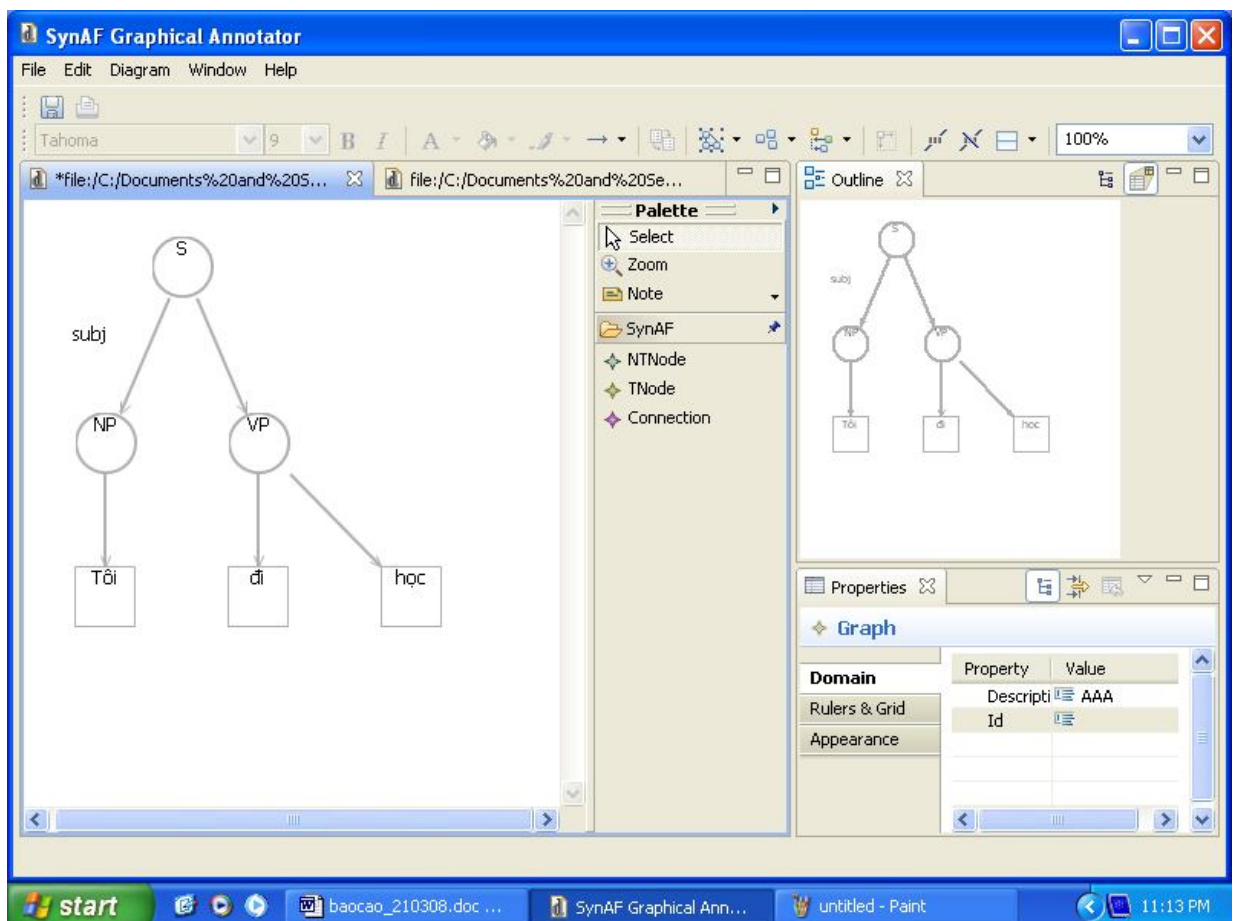
+ **Vẽ cung (Connection):** Cung hay đường nối được xuất phát từ một nút giữa đến một nút giữa hoặc từ một nút giữa đến một nút lá.

Để vẽ đường nối, ta chọn kiểu vẽ Connection, nhấn chuột và di từ nút này (nút đầu) đến nút kia (nút cuối) rồi thả chuột, sau đó ghi nhãn cho cung nếu có. Ví dụ minh họa bằng hình dưới đây:



Chọn Connection, nhấn chuột từ nút S đi đến nút NP rồi thả, ghi nhãn Subj

+ **Kết thúc ta được cây phân tích hoàn chỉnh cho câu “Tôi đi học” như sau:**



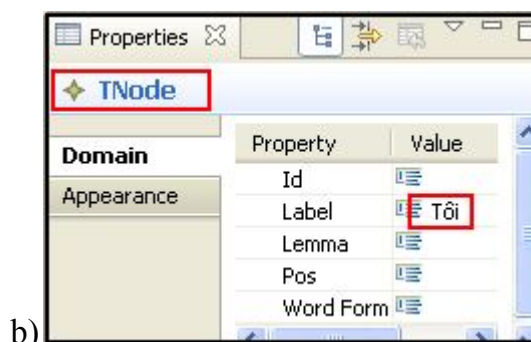
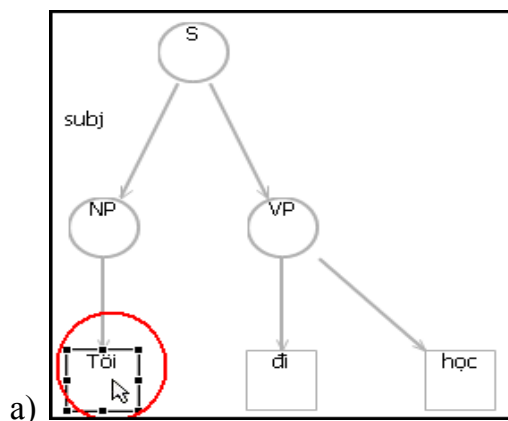
+ **Và cuối cùng** thực hiện lưu lược đồ lại ta sẽ được file **vnTreebank.synaf** có phần bổ xung đoạn phân tích cú pháp cho câu “Tôi đi học” dưới dạng mã XML.

3.2.2. Sửa cây cú pháp

Với mỗi cây cú pháp ta có thể chỉnh sửa các thuộc tính của các nút và các cung như: tên nút, nhãn của cung (label); mã nút, mã cung (id),

- Các thuộc tính này thể hiện qua cửa sổ thuộc tính *Properties*:

+ Thuộc tính nút lá *TNode*:



a) Nhấn chuột lên nút có nhãn “Tôi”

b) Cửa sổ *Properties* thể hiện các thông số thuộc tính tương ứng:

Id (mã nút):

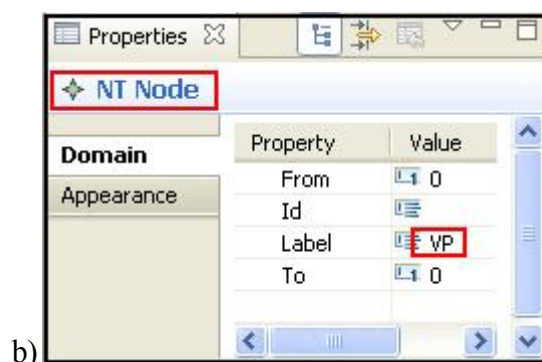
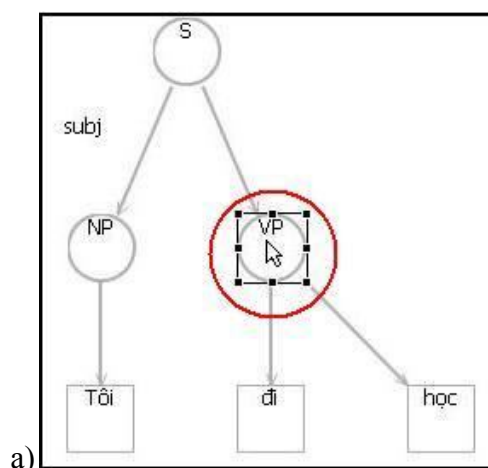
Label (nhãn): Tôi

Lemma (ghi chú):

Pos (vị trí):

Word Form:

+ Thuộc tính nút giữa *NTNode*:



a) Nhấn chuột lên nút giữa (vp)

b) Cửa sổ *Properties* hiện các thông số tương ứng:

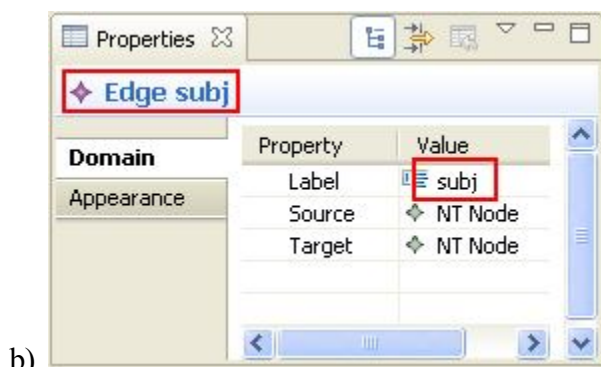
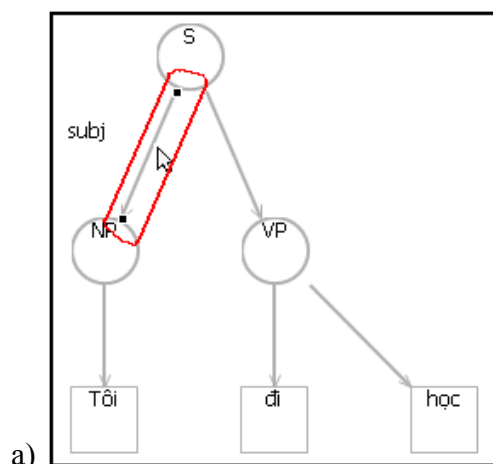
From:

Id (mã nút):

Label (nhãn): vp

To:

+ Thuộc tính cung **Connection**:



a) Nhấn chuột lên một cung (cung nối S tới NP)

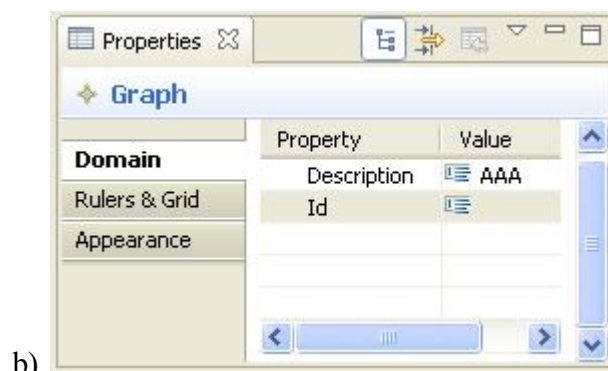
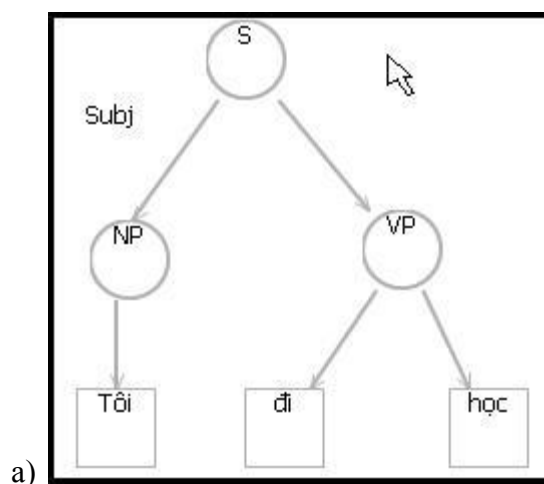
b) Cửa sổ Properties hiện các thông số tương ứng:

Label (nhãn): subj

Source (điểm xuất phát): NTNode

Target (điểm đến): NTNode

+ Thuộc tính của cây **Graph**:



a) Nhấn chuột lên vị trí nền bất kỳ

b) Cửa sổ Properties hiện các thông tin thuộc tính tương ứng:

Description (mô tả): AAA

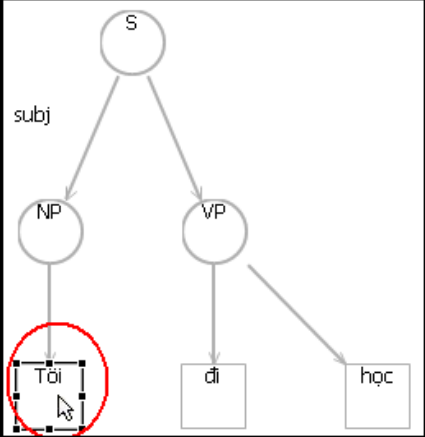
Id (mã câu):

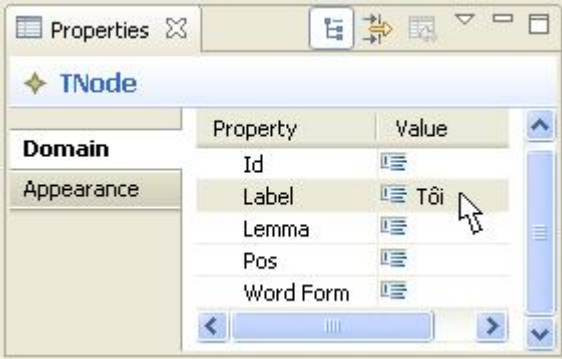
- Sửa thuộc tính bất kỳ của cây:

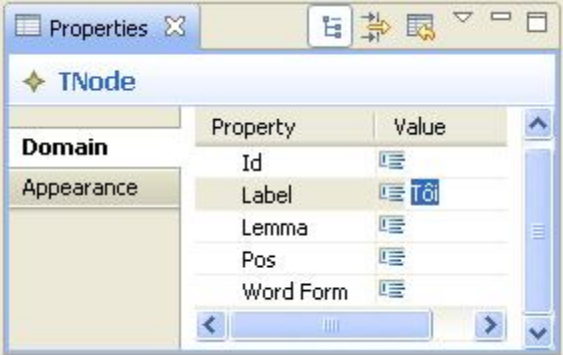
+ Nhấn chuột lên thành phần cây: nút lá, nút giữa hay là cung có thuộc tính cần sửa.

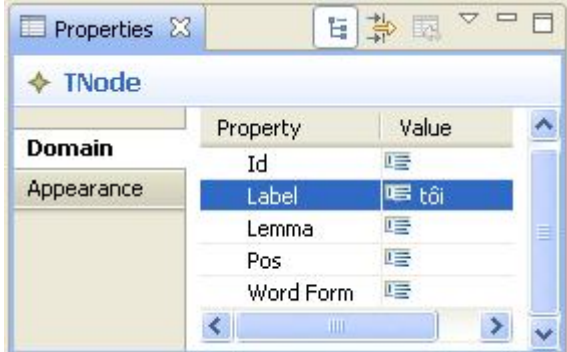
- + Nhấn chuột lên trường Value của Property tương ứng của cửa sổ Properties
- + Thay đổi giá trị

Ví dụ: Sửa nút lá có nhãn “Tôi” thành “tôi”:

a) 

b) 

c) 

d) 

- Nhấn chuột lên nút lá tôi trong phần đồ họa
- Nhấn chuột lên trường Value của thuộc tính Label trong cửa sổ Properties
- Sửa nhãn **Tôi** thành **tôi** rồi nhấn phím Enter.
- Kết quả thay đổi được thể hiện trên cửa sổ Properties và cây đồ họa.

Sau các thao tác vẽ hoặc sửa cây biểu diễn cú pháp, lưu lược đồ lại ta sẽ thu được kết quả biểu diễn các cây phân tích cú pháp trong file **.synaf** và ở minh họa trên đây là **vnTreebank.synaf**

Hướng dẫn sử dụng treebank editor (TBE)

Nguyễn Phương Thái, Lê Anh Cường

Giới thiệu

TBE là công cụ trợ giúp người làm dữ liệu gán nhãn cho câu ở nhiều mức độ khác nhau (tách từ, gán nhãn từ loại, xây dựng cây cú pháp). Các sửa đổi trên dữ liệu sẽ được lưu lại dạng file log (để thống kê, theo dõi tiến độ). Hiện tại TBE chưa được tích hợp các công cụ phân tích tự động¹¹.

Contents

I.	Các chế độ làm việc	46
II.	Soạn thảo cây	47
III.	Gán nhãn từ loại	48
IV.	Tách từ	49
V.	Xem file log	50
	V.1 Xem log cây cú pháp	50
	V.2 Xem log gán nhãn từ loại	50
VI.	Tìm kiếm và thống kê	51
VII.	Chế độ làm việc kiểm tra đồng thuận	51
VIII.	Tab Câu và tab Tách câu	52

I. Các chế độ làm việc

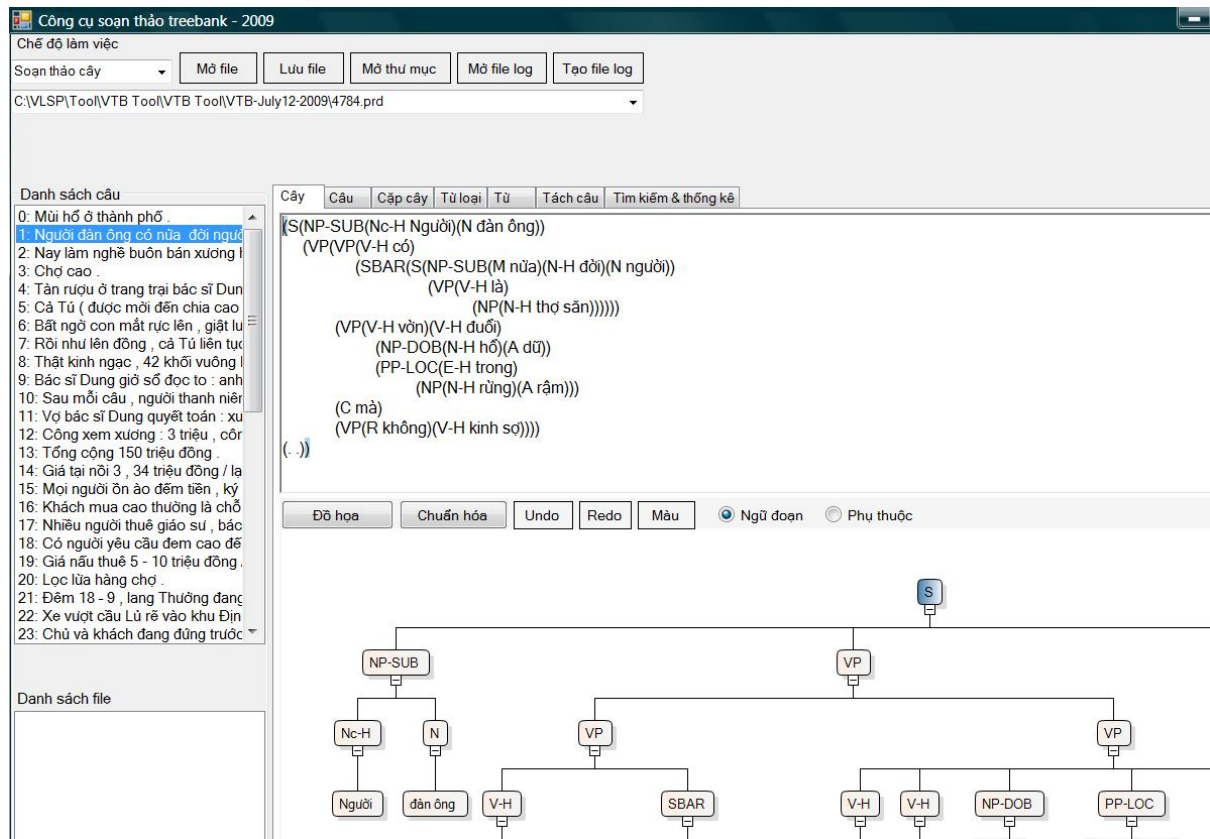
Có bảy chế độ làm việc:

- Soạn thảo cây (*)
- Gán nhãn từ loại (*)
- Tách từ (*)
- Xem log cây
- Xem log từ loại
- Xem log tách từ
- Kiểm tra đồng thuận

Việc chọn chế độ làm việc được yêu cầu thực hiện ngay khi chạy TBE. Các chế độ có dấu * (chế độ soạn thảo) sẽ ghi ra file log các sửa đổi trên dữ liệu của người làm dữ liệu. Do đó người dùng sẽ phải tạo file log mới hoặc mở file log đã có trước khi mở file dữ liệu.

¹¹ Các công cụ này sẽ được dùng để tiền xử lý dữ liệu

II. Soạn thảo cây



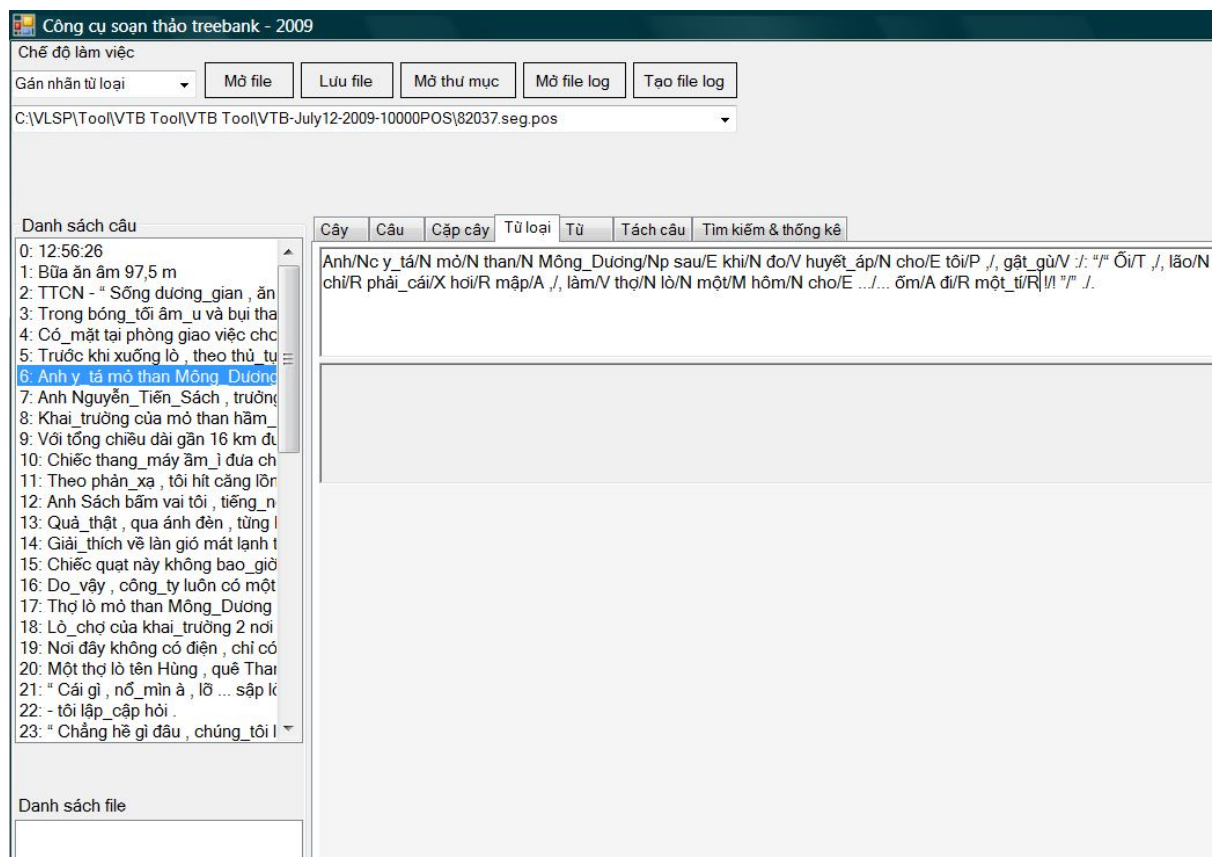
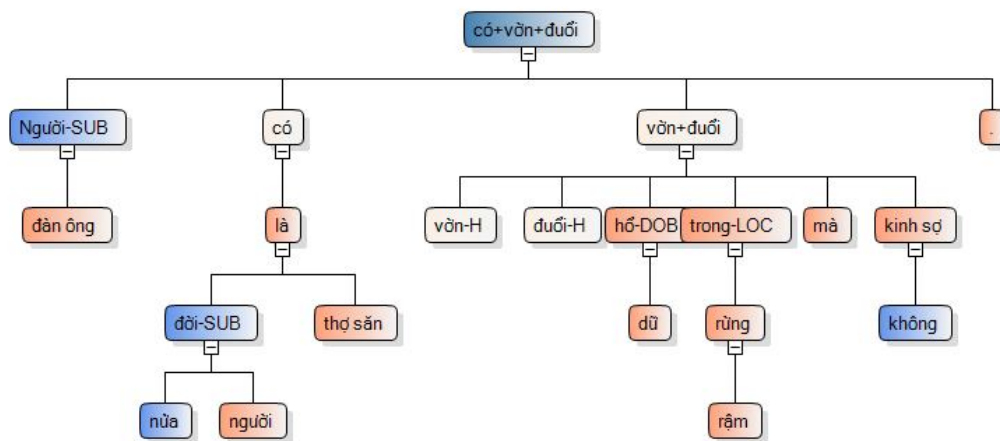
Hình 1. Chế độ soạn thảo cây cú pháp với file dữ liệu mở

Hình 1 minh họa TBE ở chế độ soạn thảo cây cú pháp. Ở góc trái trên là combo box chứa danh sách chế độ làm việc và hiện tại đang là “Soạn thảo cây”. Ngay bên dưới là combo box khác chứa tên file đang mở “C:\VLSP\Tool\VTB Tool\VTB Tool\VTB-July12-2009\4784.prd”. Dưới nữa là ba cửa sổ quan trọng khác:

- (1) Cửa sổ danh sách câu (bên trái): liệt kê các câu trong file dữ liệu đang mở
- (2) Cửa sổ hiển thị cây ở chế độ văn bản (phải-trên): cho phép người làm dữ liệu soạn thảo cây. Các cặp dấu ngoặc tương ứng với nhau sẽ được highlight.
- (3) Cửa sổ hiển thị cây ở chế độ đồ họa (phải-dưới)

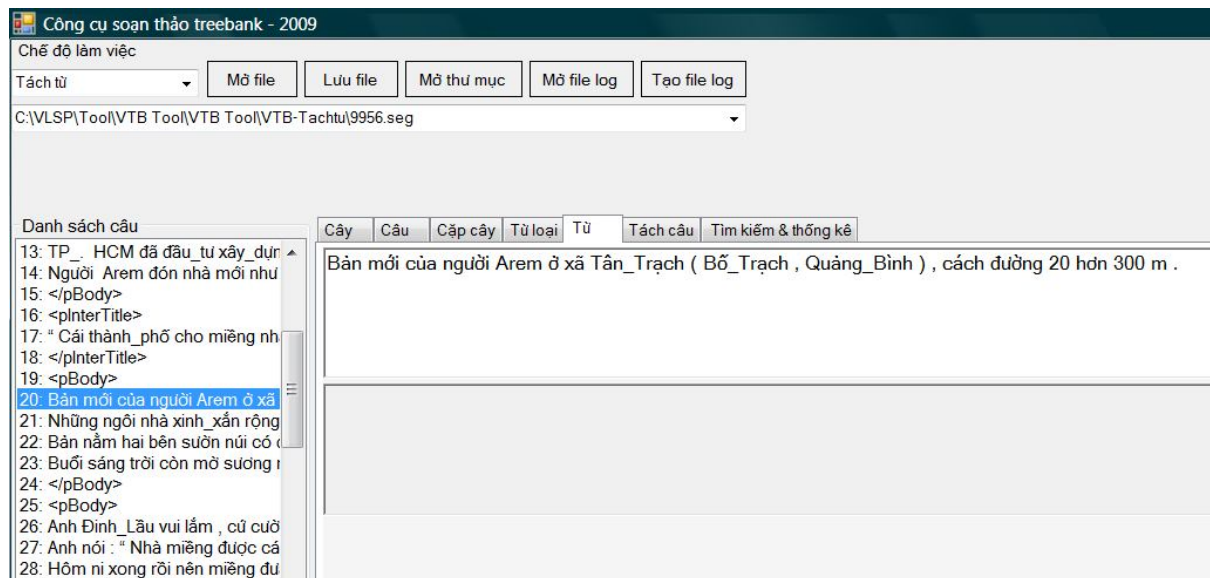
Giữa hai cửa sổ hiển thị cây là một số nút:

- Nút “Đồ họa”: tạo cây đồ họa ở cửa sổ (3) tương ứng với câu text đang có trong cửa sổ (2)
- Nút “Chuẩn hóa”: Chuẩn hóa qui cách (nút con phải lùi vào so với nút cha, các nút con phải thẳng hàng nhau, v.v.) cho cây dạng text ở cửa sổ (2)
- Nút “Undo”: hủy bỏ các hành động soạn thảo gần đây nhất
- Nút “Redo”: khôi phục các hành động soạn thảo gần đây nhất
- Nút “Màu”: Chọn màu đen cho văn bản ở cửa sổ (2)
- Nút “Ngữ đoạn”: Cây đồ họa sẽ ở dạng ngữ đoạn
- Nút “Phụ thuộc”: Cây đồ họa sẽ ở dạng phụ thuộc (xem ví dụ ở Hình 2). Cây phụ thuộc là cây mà nhãn của nút là từ (có thể là một hoặc nhiều) chứ không phải phân loại cú pháp, nhãn ở nút con là từ phụ thuộc vào từ ở nút cha, màu xanh để chỉ từ bên trái, màu hồng chỉ từ bên phải (của từ trung tâm ở nút cha).



- Chọn chế độ “Gán nhãn từ loại”
- Mở hoặc tạo file log (nút “Mở file log” hoặc nút “Tạo file log”)
- Mở file dữ liệu có tên dạng *.pos (nút “Mở file”)
- Chọn tab “Từ loại”
- Gán nhãn từ loại cho câu ở cửa sổ thuộc tab “Từ loại”
- Ấn "Lưu file" sau khi sửa mỗi câu

IV. Tách từ



Hình 4. Chế độ tách từ với file dữ liệu đang mở

Qui trình làm việc:

- Chọn chế độ làm việc là “Tách từ”
- Mở hoặc tạo file log (nút “Mở file log” hoặc nút “Tạo file log”)
- Mở file dữ liệu có tên dạng *.seg (nút “Mở file”)
- Chọn tab “Từ”
- Gán nhãn từ loại cho câu ở cửa sổ thuộc tab “Từ”
- Ấn "Lưu file" sau khi sửa mỗi câu

Chú ý: Trong danh sách câu có cả các nhãn XML (ví dụ </pBody>, <pInterTitle>) thì bỏ qua, không sửa gì các nhãn đó.

V. Xem file log

V.1 Xem log cây cú pháp

Hình 5. Xem log cây

Quy trình làm việc:

- Chọn chế độ làm việc “Xem log cây”
- Mở file log có tên dạng *.log (nút “Mở file”)
- Chọn tab “Cặp cây”
- Xem danh sách câu đã được sửa (nhấp đúp chuột vào từng câu trong danh sách câu): Ở tab “Cặp cây” có hai cửa sổ là “Cây đã sửa” và “Cây ban đầu”. Các nút được sửa sẽ có màu vàng, nút mới hoàn toàn thì màu xanh (ở Hình 5 có một nút vàng là NP-LOC trong cửa sổ “Cây đã sửa”).

V.2 Xem log gán nhãn từ loại

Quy trình làm việc:

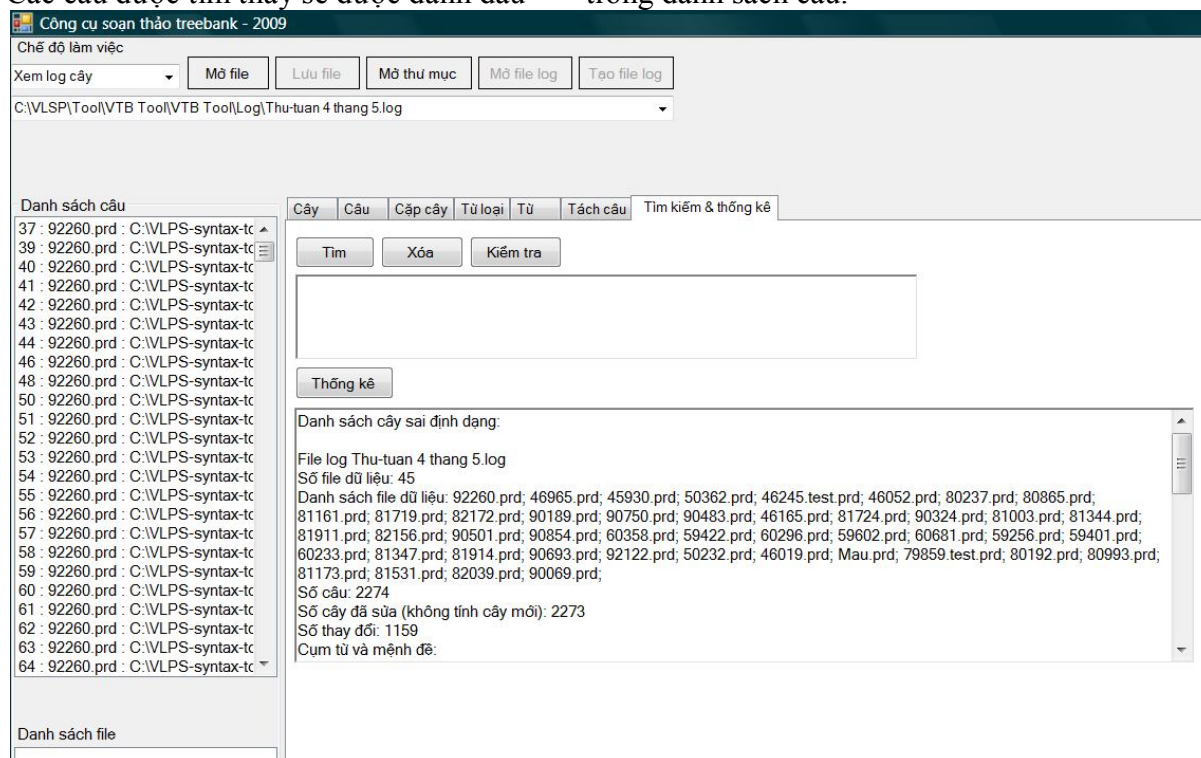
- Chọn chế độ làm việc “Xem log từ loại”
- Mở file log có tên dạng *.log (nút “Mở file”)
- Chọn tab “Từ loại”
- Xem danh sách câu đã được sửa (nhấp đúp chuột vào từng câu trong danh sách câu): Ở tab “Từ loại” có hai cửa sổ, cái trên hiển thị câu đã sửa (từ màu đỏ là đã được sửa), cái dưới hiển thị câu gốc.

VI. Tìm kiếm và thống kê

Đây là tính năng đi kèm với các chế độ soạn thảo (“Soạn thảo cây”, “Gán nhãn từ loại”, “Tách từ”). Ta có thể tìm:

- Theo từ: ở cả 3 chế độ soạn thảo
- Theo nhãn thành phần cú pháp, nhãn nút con của nó, và từ trung tâm: chỉ dành cho chế độ “Soạn thảo cây”. Ví dụ:
 - o Nếu khóa là “t=PP-LOC” thì các cây chứa nhãn PP-LOC sẽ được tìm ra
 - o Nếu khóa là “t=VP;ct=SBAR” thì các cây có cụm VP mà con của nó là SBAR sẽ được tìm ra
 - o Nếu khóa là “t=VP;hw=lấy” thì các cây có cụm VP mà trung tâm là từ “lấy” sẽ được tìm ra

Các câu được tìm thấy sẽ được đánh dấu “*” trong danh sách câu.



Hình 6. Thống kê file log cây

Tính năng thống kê khi xem log cây (Hình 6) sẽ đưa ra cho chúng ta:

- Danh sách file đã sửa
- Số nút mới, bị sửa, bị xóa (nút có thể là từ, từ loại, hay cụm từ và câu)

VII. Chế độ làm việc kiểm tra đồng thuận

Ở chế độ làm việc này, qui trình làm việc như sau:

- Chọn chế độ làm việc “Kiểm tra đồng thuận”
- Mở file của người thứ nhất
- Mở file của người thứ hai
- Chọn tab “Cặp cây”
- Xem các cặp cây giống chế độ “Xem log cây cú pháp”

VIII. Tab Câu và tab Tách câu

Tab Câu hiển thị toàn bộ các câu trong văn bản đang xét và highlight câu hiện thời. Người làm dữ liệu có thể dùng tab này trong các chế độ soạn thảo.

Tab Tách câu dùng để tách lại câu và hoạt động với chế độ Gán nhãn từ loại và Tách từ.

Người làm dữ liệu sẽ trực tiếp soạn thảo trên file nguồn được mở trong cửa sổ thuộc tab Tách câu. Họ cần chú ý giữ đúng khuôn dạng file. Tách xong cần ấn nút Lưu file ngay.

HƯỚNG DẪN TÁCH CÂU TIẾNG VIỆT

1. Tiền đề cơ sở để tách câu:

Theo sách ngữ pháp tiếng Việt của Ủy ban Khoa học Xã hội (1980): “ *Câu là đơn vị dùng từ hay đúng hơn dùng ngữ mà cấu tạo nên trong quá trình tư duy, thông báo; nó có nghĩa hoàn chỉnh, có cấu tạo ngữ pháp, và có tính chất độc lập*”. Dựa vào quan điểm này về câu ta sẽ xét một đơn vị ngôn ngữ có phải là câu hay không.

2. Mục tiêu:

- Xác định ranh giới rõ ràng và nhất quán giữa các câu tiếng Việt. Phân biệt đơn vị câu với các đơn vị nhỏ hơn câu (từ, ngữ...) và lớn hơn câu (đoạn, văn bản).
- Làm tiêu chí chính để xây dựng nên câu tiếng Việt trong ngữ liệu tiếng Việt.
- Làm cơ sở để gán các nhãn ngôn ngữ cao hơn (tách từ, gán nhãn từ loại, phân tích cú pháp...)
- Làm tiền đề cho các bài toán khác, như: đóng câu trong song ngữ Anh-Việt, Pháp-Việt, dịch tự động Việt-Anh ...

3. Phân tích và nhận diện câu:

3.1. Phân tích câu :

Xét về cấu tạo có câu đơn, câu ghép.

3.1.1 Câu đơn:

Một câu đơn cơ bản gồm có một nòng cốt đơn. Nòng cốt đơn gồm có hai phần phân đề và phần thuyết (theo quan điểm ngữ pháp chức năng) mà quan điểm ngữ pháp truyền thống gọi là chủ ngữ và vị ngữ.

Ví dụ 1:

Bão Lekima cấp 11 / đang hướng vào Nghệ An - Hà Tĩnh.

Mọi chuyện / rồi sẽ qua đi .

Trong cấu tạo câu đơn có thể có những thành phần ngoài nòng cốt như thành phần than gọi, thành phần chuyển tiếp, thành phần chú thích, thành phần tình huống, thành phần khởi ý.

Ví dụ 2:

Nhiều lúc , tôi cũng muốn gào thét thật to , đập tung , phá vỡ tất cả ...

Con người, đó là cái vốn quý nhất.

Chao, đường còn xa lắm!

Riêng với thành phần than gọi thì ta chỉ xét nó thuộc nòng cốt câu khi nó đứng ở cuối hoặc ở giữa câu.

Ví dụ 3 :

*Chúng ta đi về đi, **bà con ơi!***

Khi thành phần than gọi đứng ở đầu câu thì ta xem nó là một câu. Vì vốn dĩ thành phần than gọi đã có tính chất độc lập. Hơn nữa, nó được ngăn cách với nòng cốt câu bằng dấu (!) nên ta xem nó như một câu đặc biệt.

Ví dụ 4 :

***Trời!** Nó lại quay trở về.*

→ Tách thành 2 câu :

Trời!

Nó lại quay trở về.

Câu đơn đặc biệt là câu mà nòng cốt đơn chỉ có một thành phần.

Ví dụ 5 :

Chỉ còn lại những ngày cuối cùng ...

Điều chỉnh lại mình đi !

3.1.2 Câu ghép:

Về cấu trúc, câu ghép được tạo nên bởi ít nhất hai vế, mỗi vế là một nòng cốt đơn. Câu ghép cũng có thể có những thành phần ngoài nòng cốt như câu đơn.

Về cấu trúc câu ghép thì có hai loại câu ghép cơ bản là câu ghép song song (câu ghép đẳng lập) và câu ghép qua lại (câu ghép chính phụ).

Câu ghép song song (câu ghép đẳng lập)

Nếu cấu trúc câu đơn giản, ngắn gọn (gồm 2 vế mà mỗi vế là một nòng cốt đơn) thì ta giữ nguyên cấu trúc câu của ngữ liệu.

Ví dụ 6 :

Giọng của cháu đôi lúc đã nghẹn lại trong quá trình phiên dịch cho tổng thống và Chủ tịch nước, cháu đã cố kiềm chế những giọt nước mắt của mình vì quá xúc động.

Nếu cấu trúc câu ghép song song có hơn hai vế và quá phức tạp (gồm nhiều nòng cốt đơn) thì ta có thể tách thành những câu đơn. Bởi vì quan hệ giữa các vế trong câu ghép song song không thật chặt chẽ và tách ra càng đơn giản thì việc xử lý dữ liệu sẽ càng dễ dàng.

Ví dụ 7:

Mong ước của tôi là : đấu tranh cho đến khi đất nước giành được độc lập và sau đó lập quan hệ ngoại giao và bình thường hóa quan hệ giữa VN và Mỹ , được như vậy thì tôi có thể mỉm cười mà nhắm mắt xuôi tay bất cứ lúc nào cũng thỏa lòng rồi " .

→ Câu trên là một câu ghép đẳng lập gồm nhiều nòng cốt đơn. Ta có thể tách thành:

Mong ước của tôi là : đấu tranh cho đến khi đất nước giành được độc lập và sau đó lập quan hệ ngoại giao và bình thường hóa quan hệ giữa VN và Mỹ.

Được như vậy thì tôi có thể mỉm cười mà nhắm mắt xuôi tay bất cứ lúc nào cũng thỏa lòng rồi” .

Ví dụ 8:

Đa số bà con ủng hộ chủ trương xây dựng khu đô thị mới Thủ Thiêm và họ sẵn sàng giao đất để thực hiện dự án , nhưng họ muốn phải đảm bảo quyền lợi và cuộc sống sau khi di dời .

→ Theo ngữ nghĩa thì câu này có thể tách :

*Đa số bà con ủng hộ chủ trương xây dựng khu đô thị mới Thủ Thiêm.
Họ sẵn sàng giao đất để thực hiện dự án , nhưng họ muốn phải đảm bảo quyền lợi và cuộc sống sau khi di dời .*

Tuy nhiên ta nên hạn chế việc tách câu này, đặc biệt là với những câu ghép đẳng lập mà các vế câu được nối với nhau bằng kết từ (và, rồi, hay, còn). Vì việc tách câu này có thể làm cho câu cú gọn gàng nhưng ý nghĩa tự nhiên của ngữ liệu ít nhiều đã bị thay đổi.

Câu ghép qua lại (câu ghép chính phụ)

Câu ghép chính phụ là câu ghép mà các vế trong câu phụ thuộc lẫn nhau, không thể tách ra được.

Có thể nhận biết câu ghép chính phụ qua các cặp từ quan hệ như: nếu...thì, tuy...nhưng, do...mà, ...

Ví dụ 9:

- Dù họ là nhà thầu Nhật Bản nhưng nếu họ vi phạm pháp luật VN thì vẫn xử họ theo qui định của pháp luật VN .

- Và lại , đây là loại tội phạm mới thuộc về lĩnh vực khoa học kỹ thuật , vì vậy ngoài lực lượng điều tra của ngành công an , chúng tôi cần phải phối hợp với các ngành chuyên môn khoa học kỹ thuật khác để tìm ra nguyên nhân .

Giả sử mẹ nắm 60% vốn của công ty con ; vậy mẹ phải cử đại diện dự các phiên họp của ĐHCD của công ty con và biểu quyết theo số vốn góp .

3.2. Nhận diện câu :

3.2.1 Nhận diện chung:

Với các kiểu câu bình thường như trên ta có thể nhận biết câu qua dấu câu: dấu chấm (câu tả, câu trần thuật, câu kể), dấu chấm than (câu cảm, câu cầu khiến), dấu chấm hỏi (câu hỏi).

3.2.2 Nhận diện câu trong hội thoại:

Trong hội thoại dấu 2 chấm (:) báo hiệu cho lời nói trực tiếp, và lời nói trực tiếp này nằm trong dấu ngoặc kép (“...”) hoặc bắt đầu sau dấu gạch đầu dòng (-). Trong trường hợp này, ta sẽ tách câu (nhận diện câu qua dấu hai chấm (:)).

Ví dụ 10:

Ông cho biết:

- Căn cứ vào kết quả kiểm tra , khảo sát và những chứng cứ thu thập ban đầu từ các đơn vị nghiệp vụ , tôi nhận thấy đây là một vụ án đặc biệt nghiêm trọng, gây hậu quả lớn về người và của .

Hắn nói : “Mày chạy trước đi.”

→Tách thành hai câu:

Hắn nói :

“Mày chạy trước đi.”

Đối với đoạn hội thoại có vế trích dẫn nằm ở cuối câu thì ta cũng sẽ tách câu. Vì trong lời nói trực tiếp có nhiều câu, khi ta tách chúng ra thành những câu riêng biệt, vế trích dẫn cuối cùng sẽ gắn với câu cuối cùng làm thành một câu khác có ý nghĩa khác thì câu sẽ trở nên sai. Vì vậy ta sẽ tách vế này ra thành một câu.

Ví dụ 11:

"CSGT có nhìn thấy cũng chịu chết vì đâu có len vào được mà xử phạt . Nếu bắt dừng xe thì kẹt đường ngay" , một CSGT chốt tại đây nói .

→Tách thành ba câu:

CSGT có nhìn thấy cũng chịu chết vì đâu có len vào được mà xử phạt .

Nếu bắt dừng xe thì kẹt đường ngay.

Một CSGT chốt tại đây nói .

Ví dụ 12:

“Điều khác lạ ở VN so với nhiều nước châu Âu là các doanh nghiệp sản xuất có thể tham gia phân phối , các qui định trong kinh doanh được hiện có không qui định nhiệm vụ cụ thể của từng tổ chức trong dây chuyền phân phối ” - ông Andre nhận xét .

→ tách thành 2 câu:

“Điều khác lạ ở VN so với nhiều nước châu Âu là các doanh nghiệp sản xuất có thể tham gia phân phối , các qui định trong kinh doanh được hiện có không qui định nhiệm vụ cụ thể của từng tổ chức trong dây chuyền phân phối”.

Ông Andre nhận xét .

3.2.3 Nhận diện câu sau dấu chấm phẩy (;)

Dấu chấm phẩy (;) thường dùng để chỉ ranh giới giữa các vế trong câu ghép song song. Vì vậy ta có thể tách câu giống như câu ghép song song.

Ngoài những tiêu chí nhận diện câu qua câu ghép song song ta có những trường hợp khác sau:

Không nên tách câu khi sau dấu (;) là “thì”, “và”, “nên”

Ví dụ 13:

Giả sử , theo bản điều lệ , HĐQT có sáu thành viên ; thì công ty mẹ phải thuyết phục các cổ đông trong ĐHCĐ bầu bốn người đại diện của họ vào HĐQT. → không tách câu.

Nói một cách khác theo ngôn từ ta thường dùng , cơ quan chủ quản ra lệnh cho công ty con (1) qua số vốn mình nắm và theo quyền biểu quyết đa số tương đối hay tuyệt đối trong ĐHCĐ của công ty con ; và (2) có người đại diện của mình nắm đa số thành viên trong HĐQT . → không tách câu

Sau dấu (;) không phải là “thì”, “và”, “nên” thì ta có thể tách câu được. Riêng trường hợp sau “và” không phải là động từ, không phải là sự liệt kê thì cũng có thể tách được.

Ví dụ 14:

Vốn của nó do Nhà nước bỏ vào ; nó hoạt động theo chỉ thị của cơ quan chủ quản ; và cơ quan này là người nắm vốn duy nhất .

→ Nên tách thành:

Vốn của nó do Nhà nước bỏ vào .

Nó hoạt động theo chỉ thị của cơ quan chủ quản .

Và cơ quan này là người nắm vốn duy nhất .

Sau dấu (;) là cặp từ “nhưng (để/ nếu/ muốn)...thì” thì cũng có thể tách câu được vì cặp từ này có khả năng tạo thành một câu có đủ ý nghĩa và hoạt động độc lập được

Ví dụ 15:

Đối với chiến lược của tập đoàn , việc nâng cao hiệu quả sử dụng đất là cần thiết ; nhưng để cho các công ty con thực hiện thì đại diện của PetroVietnam tại ĐPM phải họp ĐHCĐ hay HĐQT để ra quyết định .

→ Nên tách thành:

Đối với chiến lược của tập đoàn , việc nâng cao hiệu quả sử dụng đất là cần thiết .

Nhưng để cho các công ty con thực hiện thì đại diện của PetroVietnam tại ĐPM phải họp ĐHCĐ hay HĐQT để ra quyết định .

Sau dấu (;) là một cụm từ có đầy đủ chủ vị và có khả năng độc lập thì cũng nên tách câu

Ví dụ 16:

Một nghiên cứu đã chứng minh rằng đối với các nước có trình độ phát triển thấp , mức độ phát triển xã hội là một nhân tố thích ứng với tăng trưởng ; ở một trình độ cao hơn , mức độ này dẫn đến thay đổi về phát triển cơ sở hạ tầng và các thể chế kinh tế ...

→ Nên tách thành:

Một nghiên cứu đã chứng minh rằng đối với các nước có trình độ phát triển thấp , mức độ phát triển xã hội là một nhân tố thích ứng với tăng trưởng .

Ở một trình độ cao hơn , mức độ này dẫn đến thay đổi về phát triển cơ sở hạ tầng và các thể chế kinh tế ...

Ví dụ 17:

Theo đó , chủ xe khách 63L-5796 Võ Hồng Xuân bị phạt 2,1 triệu đồng; tài xế Đặng Hữu Thành (con bà Xuân) bị phạt 2,6 triệu đồng; tài xế xe khách 63L-

5691 Lê Ngọc Trân bị phạt 2,1 triệu đồng; tài xế xe khách 63L-5634 Nguyễn Văn Thủy bị phạt 430.000 đồng .

→ Nên tách thành:

Theo đó , chủ xe khách 63L-5796 Võ Hồng Xuân bị phạt 2,1 triệu đồng.

Tài xế Đặng Hữu Thành (con bà Xuân) bị phạt 2,6 triệu đồng .

Tài xế xe khách 63L-5691 Lê Ngọc Trân bị phạt 2,1 triệu đồng .

Tài xế xe khách 63L-5634 Nguyễn Văn Thủy bị phạt 430.000 đồng .

3.2.4 Nhận diện câu sau dấu ngang (-):

Dấu ngang dùng để chỉ ranh giới của thành phần chú thích, đặt trước những lời đối thoại, liệt kê.

Đối với câu có dấu ngang dùng để chỉ thành phần chú thích thì ta không nên tách câu.

Ví dụ 18 :

Cơn sốt vé trong năm nay không còn nghi ngờ gì nữa phải thuộc về ngôi sao nhạc nhẹ mới 14 tuổi Miley Cyrus , diễn viên ngôi sao của bộ phim truyền hình Hannah Montana trên Disney Channel - bộ phim nói về cuộc sống thú vị của một cô nàng vừa là sinh viên vừa là ngôi sao nhạc nhẹ .

TTO - Sau một thời gian chạy thử nghiệm , Công ty VinaGame sẽ chính thức giới thiệu Zing MP3 - công cụ tìm kiếm âm nhạc trực tuyến đầu tiên tại Việt Nam vào đầu tháng tới .

Trên đây là những trường hợp thông thường và một số trường hợp đặc biệt mà công việc tách câu thường gặp phải (đặc biệt là đối với ngữ liệu lấy từ báo chí).

3.2.5 Thực tế nhận diện câu và một số vấn đề lưu ý khác:

- Nhận diện câu trong văn bản thơ:

Khi trích dẫn thơ xuất hiện dấu / chúng ta phải tách câu.

Ví dụ 19:

“Tôi muốn tắt nắng đi / Cho màu đừng nhạt mất / Tôi muốn buộc gió lại / Cho hương đừng bay đi”

Chúng ta phải tách thành:

*“Tôi muốn tắt nắng đi
Cho màu đừng nhạt mất
Tôi muốn buộc gió lại
Cho hương đừng bay đi”*

- Nhận diện câu qua dấu hai chấm, ngay sau đó có đánh số:

Ví dụ 20:

Người ta tổng kết có năm nguyên nhân bỏ học : (1) kinh tế gia đình khó khăn ; (2) cha mẹ không quan tâm ; (3) quản lý của nhà trường kém , chưa tập trung bồi dưỡng HS yếu ; (4) phối hợp giữa nhà trường và gia đình chưa chặt chẽ , thường xuyên ; (5) HS thiếu chuyên cần , học lực kém .

Tách thành:

Người ta tổng kết có năm nguyên nhân bỏ học :

- (1) kinh tế gia đình khó khăn ;*
- (2) cha mẹ không quan tâm ;*
- (3) quản lý của nhà trường kém , chưa tập trung bồi dưỡng HS yếu ;*
- (4) phối hợp giữa nhà trường và gia đình chưa chặt chẽ , thường xuyên ;*
- (5) HS thiếu chuyên cần , học lực kém .*

Như vậy, gặp trường hợp hai chấm (số 1, 2, 3...) chúng ta cần tách câu. Nếu không có dấu hai chấm, chỉ có (số 1,2,3...) thì chúng ta không tách.

Ví dụ 21:

Nói một cách khác theo ngôn từ ta thường dùng , cơ quan chủ quản ra lệnh cho công ty con (1) qua số vốn mình nắm và theo quyền biểu quyết đa số tương đối hay tuyệt đối trong ĐHCĐ của công ty con ; và (2) có người đại diện của mình nắm đa số thành viên trong HĐQT .

Trường hợp này không tách.

Tp.HCM ngày 15 tháng 1 năm 2008
TM Nhóm biên soạn VCL

Đinh Điền

HƯỚNG DẪN NHẬN DIỆN ĐƠN VỊ TỪ TRONG VĂN BẢN TIẾNG VIỆT

Nguyễn Thị Minh Huyền, Hoàng Thị Tuyền Linh, Vũ Xuân Lương
Báo cáo SP8.2

I. Nguyên tắc tách từ

1. Hướng tới chuẩn tách từ - ISO/TC37/SC4/WG2/WordSeg

Trong các hoạt động về chuẩn hoá tài nguyên ngôn ngữ của ISO/TC37/SC4 có nhóm làm việc WG2/WordSeg[1-3] về vấn đề chuẩn hoá tách từ cho các ngôn ngữ trong đó ranh giới giữa các từ không thể xác định rõ ràng chỉ dựa vào hình thức in ấn (như sử dụng dấu cách trong tiếng Anh).

Cho đến nay, nhóm làm việc này đã đưa ra một số bản thảo (trang web <http://tc37sc4.org>) hướng dẫn nguyên tắc chung về việc đưa ra chuẩn tách từ.

2. Đặc trưng cấu tạo từ tiếng Việt

Các phương thức cấu tạo từ tiếng Việt:

Từ đơn:

- Từ có ý nghĩa từ vựng.
- Từ có ý nghĩa ngữ pháp (từ công cụ).
- Từ tượng thanh.
- Từ cảm thán.

Từ phức:

- Từ ghép.
 - Từ ghép đẳng lập (tổng hợp).
 - Từ ghép chính phụ.
 - Từ ghép phụ gia (yếu tố ghép trước hay ghép sau để tạo từ hàng loạt).
- Từ láy.
- Dạng lặp.

Ngữ cố định:

- Thành ngữ (cao chạy xa bay, tránh vỏ dừa gặp vỏ dừa...).
- Quán ngữ (nói tóm lại, đáng chú ý là, mặt khác thì...).

Ngoài ra, trong văn bản còn có các thành phần sau:

- Tên riêng (người, địa danh, tổ chức).
- Các dạng ngày – tháng – năm.
- Các dạng số – chữ số – kí hiệu.
- Dấu câu, dấu ngoặc.
- Từ tiếng nước ngoài.
- Chữ viết tắt.

3. Đề xuất nguyên tắc tách từ cho tiếng Việt

Nguyên tắc tách từ cho tiếng Việt xét các loại đơn vị từ vựng sau đây:

Từ đơn.
 Từ ghép đẳng lập.
 Từ ghép chính phụ.
 Từ ghép phụ gia (kết hợp với yếu tố cấu tạo từ: *bất, vô, hoá, phi, viên, v.v.*).
 Từ láy, dạng lặp.
 Thành ngữ.
 Quán ngữ.
 Tên riêng.
 Ngày – tháng – năm, số – chữ số – kí hiệu.
 Dấu câu, ngoặc.
 Từ tiếng nước ngoài.
 Chữ viết tắt.

II. Hướng dẫn cụ thể

Coi là một đơn vị từ khi thực hiện tách từ đối với các đơn vị có những đặc điểm sau đây:

1. Từ đơn.

a. Từ đơn là thực từ:

- Những từ một tiếng có ý nghĩa từ vựng độc lập, có chức năng định danh (gọi tên các sự vật, hiện tượng, hành động, phẩm chất, thuộc tính, quan hệ trong thực tại khách quan).
- Đa số đều nằm trong vốn từ cơ bản của tiếng Việt, đã có từ lâu đời: *cha, mẹ, chân, tay, cơm, nước, lợn, gà, ăn, uống, cười, nói, xấu, đẹp, v.v.*; hoặc những từ gốc Hán hay gốc Ấn-Âu đã được Việt hoá: *tim, gan, buồng, phòng, cón, xăng, xăm, lớp, v.v.*; hoặc những từ Hán-Việt được dùng độc lập (do không có từ thuần Việt đồng nghĩa tương đương): *tuyệt, bút, học, đáp, cao, thấp.*
- Có một số vốn là dạng nói tắt của từ ghép: *rô* (cá rô), *chim* (cá chim), *thu* (cá thu), *nhụ* (cá nhụ), *đé* (cá đé), v.v.

b. Từ đơn là hư từ:

- Những từ một tiếng không có ý nghĩa từ vựng độc lập, không có chức năng định danh.
- Không có khả năng độc lập làm thành phần câu.
- Dùng để biểu thị các quan hệ ngữ pháp giữa các thực từ.
- Gồm phụ từ, liên từ, giới từ: *đã, sẽ, đang, vừa, mới, từng, vẫn, là, của, bằng, vì, bởi, cùng, với, nếu, tuy, nên, v.v.*

c. Từ đơn là từ tình thái:

- Những từ một tiếng đã mất ý nghĩa từ vựng và ý nghĩa ngữ pháp cụ thể, có chức năng như một phương tiện biểu thị tình thái.
- Không có khả năng độc lập làm thành phần câu.
- Biểu thị mối quan hệ giữa người nói với thực tại phát ngôn.

- Gồm thán từ và trợ từ: *à, ư, nhỉ, nhé, ời, hử, sao, a, ạ, ối, ái, thế, nào, đâu, vậy, v.v.*

2. Từ ghép đẳng lập

- Do hai thành tố (A và B) có ý nghĩa thực kết hợp với nhau theo quan hệ bình đẳng về nghĩa.
- Hai thành tố bao giờ cũng thuộc cùng một phạm trù ngữ nghĩa hoặc có quan hệ logic với nhau.
- Trật tự giữa hai thành tố nói chung có thể thay đổi được (AB hoặc BA): *quần áo – áo quần, chung riêng – riêng chung, đồ đen – đen đồ, ốm đau – đau ốm, v.v.*

2.1. Từ ghép đẳng lập gốc Việt

- Từ ghép đẳng lập gốc Việt là từ ghép trong đó hai thành tố đều là từ gốc Việt.

a. Từ ghép đẳng lập gốc Việt gồm hai thành tố có sự *gần nhau về nghĩa*:

đất nước – trời đất – đất cát – ruộng đất – ruộng vườn – ruộng nương; ẩm chén, bát đĩa, bố con, cày cuốc, chồng con, cướp phá, dệt thêu, làng xã, lúa gạo, nương vườn, râu tóc, tài sức, thác ghềnh, thầy cô, thiếu kém, thu đông, vá may, vải sợi, vườn trại, xinh đẹp, v.v.

b. Từ ghép đẳng lập gồm hai thành tố có sự *trái nhau về nghĩa*:

đỏ đen, may rủi, trong ngoài, trước sau, trên dưới, tháo lắp, cao lớn, chung riêng, hay dở, khen chê, v.v.

2.2. Từ ghép đẳng lập gốc Hán

- Từ ghép đẳng lập gốc Hán là từ ghép trong đó hai thành tố đều là từ gốc Hán.

a. Từ ghép đẳng lập gốc Hán gồm hai thành tố đã được Việt hoá hoàn toàn (được dùng độc lập như những từ gốc Việt khác):

ân nghĩa, công tư, đầu não, đấu tranh, học tập, lợi lộc, thuận lợi, v.v.

b. Từ ghép đẳng lập gốc Hán gồm hai thành tố chưa được Việt hoá hoàn toàn (không dùng độc lập như những từ gốc Việt khác):

chung thuỷ, giang sơn, kiến thiết, mỹ lệ, quốc gia, tao nhã, tranh chấp, v.v.

c. Ngoài ra còn có những từ ghép đẳng lập gồm một thành tố gốc Việt và một thành tố gốc Hán (in nghiêng là gốc Hán):

bình lính, bụng dạ, gan dạ, lính tráng, nuôi dưỡng, v.v.

3. Từ ghép chính phụ

- Do hai thành tố (A và B) trực tiếp kết hợp với nhau theo quan hệ không bình đẳng. Đó là sự phối hợp giữa một thành tố chính có ý nghĩa khái quát (A) và một thành tố phụ (B) có ý nghĩa hạn định.

- Ý nghĩa từ vựng do thành tố chính (A) quyết định; thành tố phụ (B) có vai trò bổ sung, phân loại, chuyên biệt hoá, sắc thái hoá cho thành tố chính.

- Thành tố A có thể dùng thành từ, còn thành tố B thì có thể không có tư cách ngữ pháp đó. Trật tự giữa hai thành tố A và B là không thể thay đổi được. So sánh: *xe máy – máy xe; không quân – quân không, v.v.*

3.1. Từ ghép chính phụ gốc Việt

- Vị trí của hai thành tố A và B trong cấu tạo từ ghép chính phụ gốc Việt là *chính trước – phụ sau* (AB: *xe máy, xe đạp, xe tăng*).

a. Từ ghép chính phụ bậc 1, trong đó thành tố A là từ đơn và thành tố B là một từ đơn, hoặc một từ ghép, hoặc một tổ hợp từ:

- + cá (A): cá mè, cá rô, cá trắm, cá quả, cá hồng, cá voi, cá heo, cá chai, cá bột, cá nhà táng, cá sần sật, cá thồn bơn, v.v.
 - + chim (A): chim gáy, chim khuyên, chim ngói, chim hát bội, chim cánh cụt, chim phượng chèo, chim thầy bói, v.v.
 - + hoa (A): hoa hồng, hoa nhài, hoa lan, hoa li, hoa sói, hoa mõm sói, hoa mép dê, hoa cứt lợn, hoa loa kèn, v.v.
 - + rau (A): rau má, rau sam, rau răm, rau sống, rau húng, rau thơm, rau tập tàng, v.v.
 - + cà (A): cà chua, cà bát, cà pháo, cà tím, cà dái dê, cà độc dược, v.v.
 - + máy (A): máy bay, máy bơm, máy sát, máy xay, máy kéo, máy cày, máy gặt đập, máy phát điện, máy quay đĩa, máy thu hình, v.v.
 - + xe (A): xe đạp, xe tăng, xe cút kít, xe cứu hoả, xe cứu hộ, xe cứu thương, v.v.
 - + bếp (A): bếp dầu, bếp điện, bếp gas, bếp từ, v.v.
 - + nồi (A): nồi hầm, nồi hấp, nồi hơi, nồi supde, nồi áp suất, nồi cơm điện, v.v.
 - + bàn (A): bàn đọc, bàn giấy, bàn thờ, bàn cờ, v.v.
 - + làm (A): làm bếp, làm biếng, làm công, làm giàu, làm việc, v.v.
 - + đen (A): đen đúa, đen giòn, đen hắc, đen ngòm, đen nhẻm, đen sì, v.v.
- v.v...

b. Từ ghép chính phụ bậc 2, trong đó thành tố A là một từ ghép và thành tố B là một từ đơn, hoặc một từ ghép (gốc Việt hoặc gốc Hán), hoặc một tổ hợp từ:

- + cá mè (A): cá mè hoa, cá mè trắng, v.v.
 - + máy bay (A): máy bay bà già, máy bay trực thăng, máy bay lên thẳng, máy bay cường kích, máy bay khu trục, máy bay không người lái, v.v.
 - + máy xay (A): máy xay sinh tố, máy xay thịt, v.v. (???)
 - + động cơ (A): động cơ diesel, động cơ đốt trong, động cơ điện, động cơ vĩnh cửu, v.v.
- v.v...

3.2. Từ ghép chính phụ gốc Hán

a. Trường hợp thông thường, hai thành tố A và B trong từ ghép chính phụ gốc Hán được sắp đặt theo trật tự *phụ trước – chính sau*. Trong đó, thành tố A là từ đơn được dùng độc lập hoặc không độc lập và thành tố B là một từ đơn, hoặc một từ ghép.

- + ca (A): dân ca, đồng ca, xướng ca, khái hoàn ca, v.v.
- + dân (A): bình dân, cư dân, ngư dân, nông dân, v.v.
- + học (A): bác học, văn học, kinh tế học, cổ sinh vật học, v.v.

- Chú ý: Có trường hợp thành tố B là từ gốc Việt, gốc Anh.

môi hoá, nhót ké, ampe ké, logic học, v.v. (*môi, nhót, ampe, logic* là B)

b. Có trường hợp hai thành tố A và B trong từ ghép chính phụ gốc Hán được sắp đặt theo trật tự *chính trước – phụ sau*; trường hợp này A là động từ và B là từ đơn gốc Hán được dùng độc lập hoặc không độc lập.

- + đá (A): đá đảo, đá động, đá kích, đá phá, v.v.
- + thuyết (A): thuyết giảng, thuyết lí, thuyết minh, thuyết phục, v.v.

CHÚ Ý:

1. Với loại từ ghép chính phụ, khi thành tố A là danh từ chỉ đồ vật (vật vô sinh: *máy, xe, bếp, nồi, bàn*, v.v.), thì thường thành tố B là động từ hoặc tính từ biểu thị ý nghĩa công

dụng, mục đích, cách thức, tính chất. Theo đó, các tổ hợp kiểu: *nồi đồng* (*nồi bằng đồng*), *nồi đất* (*nồi bằng đất*), *mâm nhôm* (*mâm bằng nhôm*), *bàn gỗ* (*bàn bằng gỗ*), *ghế đá* (*ghế bằng đá*), v.v. không có tư cách là một từ ghép chính phụ. B ở đây (*đồng, đất, nhôm, gỗ, đá*) là những danh từ chỉ chất liệu.

2. Trong tiếng Việt còn có những từ có nhiều tiếng (bao gồm cả từ vay mượn đã được Việt hoá, hoặc có hình thức phiên âm gần giống với tiếng Việt), xét theo phương thức cấu tạo thì không thuộc loại từ ghép cũng không thuộc loại từ láy. Chúng bao gồm những tiếng không có nghĩa hoặc mờ nghĩa (có thể do chưa biết được nghĩa gốc), phải cả khối gồm nhiều tiếng hoà quyện làm một chỉnh thể chặt chẽ mới có nghĩa: *bỏ nông, bỏ hóng, bù nhìn, mặt chược, ca la thầu, ba lô, béc giê, cà phê, căng tin, xi măng, xích lô*, v.v. Những từ này cũng được xếp chung vào nhóm từ ghép.

3. Cũng coi là từ ghép với các tổ hợp gộp (của hai từ ghép) biểu thị ý nghĩa tổng hợp:

- Kết hợp giữa hai, ba thành tố đầu trong mỗi từ ghép: công nông (*công nhân* và *nông dân*), công nông binh (*công nhân, nông dân* và *binh lính*), v.v.
- Cả hai từ ghép đều có chung thành tố chính A (đứng cuối): *y bác sĩ* (*y sĩ* và *bác sĩ*), *ưu nhược điểm* (*ưu điểm* và *nhược điểm*), *khám chữa bệnh* (*khám bệnh* và *chữa bệnh*), *binh công xương* (*binh xương* và *công xương*), v.v.
- Dạng viết đầy đủ: *phòng cháy chữa cháy, phòng bệnh chữa bệnh*, v.v.

4. Trong những trường hợp lưỡng lự có thể xét đến các lí do sau đây:

- a) Những tổ hợp có cấu tạo tương đương như các từ đã được thu thập trong *Từ điển công cụ* (từ điển dùng làm công cụ tách từ), nhưng không được hoặc chưa được thu thập (trong ngoặc là từ có trong *Từ điển công cụ*):

anh hồn (*anh linh*), *chao ơi* (*chao ôi*), *chúng bay* (*chúng mày*), *chúng nó* (*chúng tôi, chúng ta*), *con ở* (*người ở*), *công dân quyền* (*quyền công dân*), *đánh tâm* (*đang tâm*), *đôi lúc* (*đôi khi*), *giời ơi* (*trời ơi*), *giời phật* (*trời phật*), *hai thân* (*song thân*), *khăn tay* (*khăn mùi soa*), *khốn nỗi* (*khốn một nỗi*), *không thể nào* (*không thể*), *luật phép* (*luật pháp*), *oai tín* (*uy tín*), *quan binh* (cũ, như *quan quân*), *sốt tiết* (*điên tiết*), *sức của* (*vật lực*), *sức người* (*nhân lực*), *tấm gương* (như *gương*), *thang thuốc* (*thuốc thang*), *tín tâm* (*lòng tin*), *thiệt ra* (*thật ra*), *tổng sản phẩm trong nước* (*tổng sản phẩm quốc nội*), *xem trọng* (= *coi trọng*), v.v.

- b) Chưa được thu thập trong *Từ điển công cụ*, nhưng đã được thu thập ở một vài quyển từ điển khác:

giá trị gia tăng (NLân), *khách hàng* (TĐ2008), *khu công nghiệp* (TĐ 2008), *kiến trúc sư trưởng* (NLân), *kim tiêm* (Đại TĐ, NLân, VTân), *lưu toan* (NLân, VTân), *nghe nga* (NLân), *nhà ở* (NLân, VTân), *như vậy* (NLân, VTân, LVĐức, TNghị), *như thế* (NLân, VTân, LVĐức, KTrí, TNghị), *quan binh* (LVĐức), *quan tư* (Đại TĐ, VTân, KTrí, ĐVTập), *quốc công tiết chế* (Đại TĐ, NLân, VTân), *rẻ rẻ* (LVĐức, ĐVTập), *thù hiềm* (LVĐức), *tự đại* (ĐVTập, TNghị, LVĐức), v.v.

- c) Đơn vị từ vựng mới xuất hiện (từ mới hoàn toàn, hoặc từ cũ nay dùng lại):

ảnh điểm, bưu báo, bo chủ, giả lập, lục sì, máy để bàn, máy tính để bàn, nguyên lão nghị viện, quan năm, tác vụ, trình khách, tư tủng, v.v.

d) Các cụm từ kiểu: *đáp lễ, đến nỗi, làm sao, như ai*, v.v.

4. Từ láy, dạng lặp

4.1. Từ láy

- Từ láy phổ biến là từ gồm hai tiếng (song tiết, hai âm tiết), trong đó một tiếng có hình thức lặp lại âm của tiếng kia. Các tiếng kết hợp với nhau vừa có sự hài hoà về ngữ âm, vừa có giá trị biểu cảm, gọi tả.

- Thường chỉ có một tiếng có nghĩa và một tiếng mờ nghĩa: *chậm chạp* (*chậm* có nghĩa), *long lanh* (*long* có nghĩa), *lúng túng* (*túng* có nghĩa), *long tong* (*tong* có nghĩa); hoặc cả hai tiếng đều mờ nghĩa: *khấp khểnh, lênh đênh, lênh khênh, lêu nghêu, lung linh*, v.v.

a. Kiểu AA' (A là tiếng gốc, tiếng chính; A' là tiếng láy của A):

chậm chạp, lành lặn, nhanh nhẩu, vừa vặn, v.v.

b. Kiểu A'A (A là tiếng gốc; A' là tiếng láy của A):

b.1. *bành bạch, bì bạch, long tong, lộp bộp, lúng túng, rôm rộp*, v.v.

b.2. *đềm đẹp, đo đỏ, lành lạnh, nho nhỏ*, v.v.

c. Kiểu AA:

c.1. Lặp hoàn toàn âm của tiếng gốc, phần lớn là từ tượng thanh: *ào ào, âm âm, au au, ặc ặc, âm âm, bành bành, độp độp, êm êm* (không phải tượng thanh), *ha ha, khao khao, khặc khặc*, v.v.

c.2. Lặp hoàn toàn âm của tiếng gốc một cách đơn giản (nghĩa không biến đổi gì nhiều): *cau cau, chau chau, đen đen, lấm lấm, quen quen, run run, xanh xanh*, v.v.

d. Kiểu ABB (B là thành tố của từ ghép chính phụ AB):

đen sì sì, đỏ lòm lòm, nông choèn choèn, tối om om, xanh lè lè, v.v.

e. Kiểu AB'B (B' là tiếng láy của B; AB là từ ghép chính phụ):

đen trùi trùi, đỏ hoen hoét, đỏ hơn hơn, cao lêu nghêu, dài đuôn đuôn, v.v.

f. Kiểu ABC (có sự biến đổi về thanh điệu) – nghiên cứu thêm:

dừng dưng dưng, sạch sành sanh, v.v.

g. Kiểu AA'AB (A là tiếng đầu của từ ghép AB; A' là tiếng láy của A; A' có cấu tạo dạng xa, trong đó x là phụ âm đầu của A, a là phần vần có giá trị hoà phối ngữ âm cho cả khối):

ấm a ấm ỨC, đủng đa đủng đỉnh, long la long lanh, nhí nha nhí nhánh, v.v.

CHÚ Ý:

1. Các kiểu b.2 (của b), c.2 (của c), d, e, f, g có tài liệu phân thành *dạng láy*. Khái niệm “dạng láy” không chỉ ra được sự khu biệt với khái niệm “láy”. Vả lại, *láy* bản thân là một *dạng* của phương thức cấu tạo từ, cũng như *ghép, lặp*. Vì những lẽ đó, tài liệu này không phân biệt *từ láy* và *dạng láy* của từ.

2. Các tổ hợp dạng *ba ba, cào cào, châu châu, chuồn chuồn*, (quả) *đu đủ*, (quả) *su su, thần lằn, thường luông*, v.v. xét về mặt ý nghĩa, chúng không có giá trị biểu cảm, gọi tả như các từ láy, nhưng xét về hình thức ngữ âm thì chúng có cấu tạo giống như từ láy, vì vậy tài liệu này xếp chung vào loại từ láy.

4.2. Dạng lặp

a. Kiểu AA (lặp hoàn toàn tiếng gốc để chỉ số lượng nhiều, hoặc chỉ mức độ cao; cả hai thành tố đều là danh từ):

ai ai, đầu đầu, đêm đêm, lớp lớp, ngày ngày, người người, nhà nhà, sáng sáng, tháng tháng, tối tối, v.v.

b. Kiểu AAA (thường là tượng thanh):

âm âm âm, ha ha ha.

c. Kiểu AABB (AB là từ ghép đẳng lập, trong đó A ngược nghĩa với B)

đi đi lại lại, hư hư thực thực, lên lên xuống xuống, quần quần áo áo, ra ra vào vào, v.v.

d. Kiểu ABAC (B và C thường tạo thành từ ghép đẳng lập, trong đó B ngược nghĩa với C, nhưng đôi khi cũng có thể B đồng nghĩa với C; A là yếu tố chen vào đầu và giữa tổ hợp BC).

chạy ngược chạy xuôi, chẳng nói chẳng rằng, dần đi dần lại, đá đi đá lại, đảo đi đảo lại, khát quanh khát quẩn, khoảng lầy khoảng để, khua đi khua lại, người này người nọ, trông trước trông sau, về lâu về dài, v.v.

5. Từ ghép phụ gia

- Đây là kiểu tạo từ hàng loạt bằng cách ghép các yếu tố có khả năng cấu tạo từ cao (như *bất, vô, phi...*) vào trước hay sau một từ ghép khác. Có một số tổ hợp được tạo ra từ phương thức này do không có sự ổn định cao nên có thể chưa được thu thập trong từ điển giải thích ngôn ngữ, chẳng hạn *cổ bộ trưởng, cựu bộ trưởng, cố giáo sư, nguyên giáo sư, v.v.*

5.1. Danh sách các yếu tố đứng trước

bán + N = N: *bán bình nguyên, bán nguyên âm, bán sơn địa, bán thành phẩm.*

bán + A = A: *bán tự động, bán vũ trang.*

bất + A = A: *bất bình đẳng, bất đắc chí, bất hợp lí, bất khả thi.*

bất + V = V: *bất bạo động, bất hợp tác.*

bất + N = N: *bất đẳng thức, bất động sản, bất phương trình.*

cổ + N = N: *cổ bộ trưởng, cố giáo sư, cố thủ tướng, v.v.*

cựu + N = N: *cựu bộ trưởng, cựu giám đốc, cựu thủ tướng, v.v.*

đa + N = N: *đa phương tiện, đa tác vụ*

đại + N = N: *đại bản doanh, đại bộ phận, đại công nghiệp, đại gia đình.*

hữu + N = A: *hữu hạn, hữu hình, hữu sự, hữu thần.*

hữu + V = A: *hữu dụng, hữu khuynh, hữu sinh, hữu trách*

liên + N = N: *liên bang, liên bộ, liên ngành, liên cầu khuẩn, liên chi uỷ, v.v.*

nguyên + N = N: *nguyên bộ trưởng, nguyên thủ tướng, nguyên trưởng phòng, v.v.*

nhà + V = N : *nhà cung cấp, nhà phê bình* (chú ý: tách phần bổ ngữ tiếp sau, nếu có: *nhà phê bình / văn học ; nhà phê bình / điện ảnh, ...*).

phi + N = A: *phi chính phủ, phi lợi nhuận, phi nhân đạo, phi nông nghiệp, phi windows.*

phó + N = N: *phó chủ nhiệm, phó chủ tịch, phó viện trưởng, phó giám đốc.*

siêu + N = N: *siêu giai cấp, siêu hạng, siêu sao, siêu cầu thủ, siêu lợi nhuận, siêu trầm*

siêu + V = V: *siêu dẫn, siêu thoát, siêu thắng*

siêu + A = A: *siêu thực, siêu trường, siêu trọng*

tái + V = V: *tái cơ cấu, tái đầu tư, tái định cư, tái sản xuất*

tiểu + N = N: *tiểu bang, tiểu công nghệ, tiểu gia súc, tiểu khí hậu, tiểu loại, tiểu vương quốc*

trưởng + N = N: *trưởng ban, trưởng phòng, trưởng thôn, trưởng tộc*

tối + A = A: *tối đại đa số, tối thông minh*

vô + N = A: *vô chủ, vô đạo, vô đạo đức, vô gia cư, vô nhân đạo, vô thần, vô văn hoá*

vô + V = A: vô can, vô địch, vô học,...

vô + V = P: vô kể, vô luận

5.2. Danh sách các yếu tố đứng sau

N + hoá = V: lao động *hoá*, công nông *hoá*, trí thức *hoá*

N + kiêu = N: Ân *kiêu*, Hoa *kiêu*, Việt *kiêu*

N + trưởng = N: đại đoàn *trưởng*, phân viện *trưởng*, tiểu đoàn *trưởng*,

V + viên = N: cộng sự *viên*, lập trình *viên*, điều tra *viên*

N + viên = N: công an *viên*

6. Tổ hợp có tính thành ngữ, quán ngữ

6.1. Danh sách các đơn vị thành ngữ

anh hùng áo vải	muốn gì được vậy
ăn cần nói rõ	muru sâu kẻ giỏi
ăn cơm chúa múa tối ngày	năm chừng mười hoa
ăn đói mặc rách	người quen kẻ thuộc
buốt như kim châm	như muối bỏ bể
bụng chứa vượt mặt	nhức đầu sổ mũi
bữa rau bữa cháo	nhức như búa bổ
chân mây ngọn sóng	no đói có nhau
chia ba xẻ bảy	nổi như cồn
chủ quan khinh địch	nước mắt nhà tan
có thực mới vực được đạo	quanh đi quẩn lại
con Lạc cháu Hồng	quân nào tướng nấy
cứu khổ cứu nạn	sĩ nông công thương
dân chi phụ mẫu	suy đi nghĩ lại
dầu sương dãi nắng	tan nhà nát cửa
đánh ngay thắng ngay	tán gia bại sản
đi nắng về mưa	thâm sơn cùng cốc
đồng chu cộng tế	thiên kinh vạn quyển
đủ ăn đủ mặc	thuật kỳ phép lạ
đường đi nước bước	tiền nghìn bạc vạn
giết người cướp của	tối mù tối mịt
hoá chính vi linh	trao tứ chiếng gái giang hồ
huynh đệ chi bang	trời cao đất dày
hương lạnh khói tàn	trời xanh nước biếc
hữu tiến vô thoái	trường xuân bất lão
khoanh tay chờ chết	tuổi già sức yếu
lai vô ảnh khứ vô tung	tư thù tư oán
lầu son gác tía	vay quanh mượn quản
mắt to hơn bụng	vắt cam vứt xác
một cổ đôi trùng	vợ đẹp con khôn
một mất một còn	...
một sống một chết	

6.2. Danh sách các đơn vị quán ngữ

lễ với nghĩa
vợ với con
đáng chú ý là
mặt khác thì
nói cho cùng
nói một tiếng
nói tóm lại = tóm lại
v.v...

7. Tên riêng

* Tên người, tên địa danh, tên tổ chức được coi là một đơn vị từ vựng: tách theo quy định tách từ thông thường, riêng danh từ riêng thì gộp làm một.

- Tên tổ chức:

báo - Tuổi trẻ
Công ty - Cao su - Đồng Nai
Điện lực - Bến Tre
Công ty - tàu biển - Simexco
Tập đoàn - dệt may - Khatoco
Công an - Thành phố - Hà Nội
Bộ - giáo dục - đào tạo
Trường - Đại Học - Quốc Gia - Hà Nội
Công ty - TNHH - AIVIETNAM
Công ty - cổ phần - Traphaco

- Tên địa danh:

+ Tách riêng phần danh từ chung và tên riêng địa danh
xã - Xuân Thanh
huyện - Long Khánh
tỉnh - Đồng Nai
Nông trường - Cẩm Đường
TP. - HCM
sông - Nhơn Mỹ
chợ - Phương Lâm
đảo - Hoàng Sa

+ Không tách đối với những trường hợp những tên địa danh có số lượng rất hạn chế:

Châu Á, châu Âu, châu Phi, châu Đại dương, châu Mỹ Latin

+ Không tách đối với những tên địa danh chỉ một thực thể được cấu tạo ghép:

Đông Nam Á, Bắc Mỹ, Bắc Triều Tiên, Đông Âu.

(riêng các trường hợp tên địa danh có cấu tạo gộp như Châu Á – Thái Bình Dương thì tách)

Đề thống nhất thì các trường hợp sau cũng tách¹²:

Chợ Hôm, Chợ Viềng, Chợ Si, Chợ Sắt, Chợ Âm Phủ - Chợ 19-2, Chợ Chà, Chợ Nôn, Sao Hôm, Sao Mai, Sao Thổ, Phố Hiến, Làng Vòng, Làng Tó (Tó Thôn), Cống Mọc, Hồ Tây, Hồ Bảy Mẫu, Hồ Thiên Quang, Hồ Hạ-le, Hồ Than Thở, Biển Chết, Biển Đen, Biển Đỏ, Sông Hồng, Sông Mã, Sông Chảy, Công viên Lenin, ...

- Tên người:

+ Tách riêng phần danh từ chung chỉ địa vị, tư cách, ... với tên riêng chỉ người

bạn đọc - Nguyễn Hữu Ngọc Anh

bạn đọc - Nguyễn Thừa Nghiệp

Bạn đọc - Phan Văn Chiến

Thủ tướng - Nguyễn Tấn Dũng

Chủ tịch - Hồ Chí Minh

Cầu thủ - Nguyễn Hồng Sơn

+ Tách riêng phần danh từ chung chỉ địa điểm, ... với tên riêng chỉ người

Thành phố - Hồ Chí Minh

Đường - Nguyễn Trãi, đường - Nguyễn Chí Thanh, đường - Phạm Văn Đồng

Công Viên - Lê Nin

+ Các trường hợp không phải tên riêng, nhưng gián tiếp chỉ người, thì tách như tách từ thông thường:

Chủ tịch - nước - Việt Nam

Quả bóng vàng - 2008

8. Ngày – tháng – năm, số – chữ số – kí hiệu

8.1. Ngày – tháng – năm.

- Giữ nguyên cả khối với các dạng (trong ngoặc không tính đến):

30-4-1975; 30-04-1975; 30-4-75; 30-04-75

(Ngày) 1-6; 01-06;

(Quốc khánh) 2-9; 02-09

- Tách thành từng đơn vị số, dấu, chữ như quy định thông thường:

tháng / 6 / – / 2003, Năm / 1997

8.2. Số – chữ số – kí hiệu

- Công thức hoá học, biểu thức toán học giữ nguyên cả khối:

$H + O_2 = H_2O$; $100 - x + 5 = 50$; $x - 23 < 23$

- Biểu hiện liên tục một con số chính xác bằng số (có dấu chấm: 1.500, không có dấu chấm 23000, VII, hay có dấu cách 1 000) hoặc bằng chữ (VD: hai mươi vạn, hai mươi phẩy hai, ba phần tư).

¹² Nhiều trường hợp dùng độc lập nhưng vẫn hàm chứa tên địa danh: *Phùng, Nhón, Mẹt, Trôi*.

- Biểu hiện đặc biệt cả số và kí hiệu một cách liên tục (không có dấu cách) như: 19g25, 50%, 20ha.

- Trường hợp kí hiệu đơn vị đứng trước hoặc sau (không chen vào giữa) số thì tách:

20ha → 20 – ha

15\$ → 15 – \$

£12 → 12 – £

- Biểu hiện hỗn hợp cả số và chữ thì tách riêng từng phần:

60	hai mươi nghìn	2	121, 8
phần trăm	tấn	-	tỉ
	rưỡi	3	
		triệu	
100			
phần			
100			

9. Dấu câu

- Tách riêng toàn bộ các loại dấu câu.

10. Từ tiếng nước ngoài

- Với các từ, thuật ngữ, khái niệm thì tách theo từng khối kí tự viết liền.

- Đối với tên riêng (tên người, tên địa danh) viết theo dạng đầy đủ thì tách theo mục 7.

- Trường hợp tên người và tên đệm viết tắt thì vẫn giữ nguyên cả khối:

V. E. Lênin,

11. Chữ viết tắt

- Tách theo từng khối kí tự viết liền:

ADSL, CNXH

- Chữ viết tắt là một bộ phận của tên riêng thì xử lí giống như tên riêng, tức là giữ nguyên cả khối:

Đại học KHXH&NV Hà Nội, Cty TNHH Rạng Đông

Một số lưu ý khi thực hiện công việc tách từ vòng 2:

1) Các cụm từ chỉ ngày, tháng, năm một cách chính xác vòng 1 đã gộp thì bây giờ tách ra theo quy định trên.

2) Các trường hợp có dấu cách trước và sau dấu chấm ở chữ số (13 . 000), trước và sau dấu phẩy ở số thập phân: (3 , 5) thì xoá dấu cách và để thành 1 đơn vị.

3) Có những biểu thức có dấu ‘/’ như phân số, nhưng thực chất không phải thì phải tách: VD ở file 1019.txt: chỉ – có – 1 – / – 7 – khu – đã – khởi công – xây dựng (1/7 ở đây đọc là “một trên bảy” hoặc “một trong bảy” chứ không đọc “một phần bảy”); có những biểu thức có dấu ‘,’ như số thập phân, nhưng thực chất không phải thì phải tách: VD ở file 1019.txt: phát triển

– 1 . 447 – km – đường ống – cấp – 1 – , – 2 – , – 3 (dấu phẩy ở đây là dấu câu chứ không phải dấu trong số thập phân).

3) Các trường hợp có dấu cách sau các chữ viết tắt (TP .) thì xoá dấu cách và để thành một đơn vị.

4) Hiện tượng nhập nhằng về nghĩa: Rất nhiều trường hợp từ được tách đúng về mặt hình thức, nhưng sai về nghĩa trong ngữ cảnh cụ thể, đòi hỏi người tách từ phải nhận ra và sửa lại cho đúng:

- rút xuống sông vì / cầu sập	- cử phóng viên làm / tin	- về vụ cháy / chợ Phương Lâm	- anh giật / dây cầu cứu	- Hầm đi / sâu vào lòng núi	- nhận được / cái lắc đầu
-------------------------------	---------------------------	-------------------------------	--------------------------	-----------------------------	---------------------------

- thừa ủy nhiệm kiến trúc sư trưởng TP	- Trường hợp căn / số 365 PNL được cấp GPXD	- do Công ty TNHH Hai Thành làm / chủ đầu tư		
--	---	--	--	--

5) Các lưu ý khác:

- Khi gặp những trường hợp rất khó xác định hoặc khi quyết định những đơn vị từ không có trong Từ điển thì phải ghi chú lại, thảo luận để tìm cách giải quyết thống nhất trong nhóm và đảm bảo tính nhất quán trong tư liệu.

- Có những trường hợp vòng 1 gộp nhưng bây giờ thấy nên tách ra thì đúng hơn: **một / nửa; mà / còn.**

- Có những trường hợp vòng 1 tách nhưng bây giờ thấy nên tách ra thì đúng hơn: **nhà ở (phân biệt với nhà xưởng).**

Hướng dẫn gán nhãn từ loại

Nguyễn Phương Thái, Vũ Lương, Nguyễn Thị Minh Huyền, và nhóm dữ liệu

SP 7.3 – VLSP

Nhóm VTB lựa chọn tiêu chí *phân loại từ dựa trên khả năng kết hợp và chức vụ ngữ pháp của từ*. Chẳng hạn danh từ thì thường có chức vụ ngữ pháp là chủ ngữ hoặc bổ ngữ trong câu, thêm vào đó là khả năng kết hợp với số từ (hai, ba) và định từ (mỗi, mọi). Khi gán nhãn từ loại nhóm dữ liệu thường tham khảo từ điển tiếng Việt và các sách ngữ pháp. Tuy nhiên nếu không có một tài liệu hướng dẫn thì vẫn có trường hợp mọi người đưa ra các quyết định khác nhau cho cùng một tình huống. Tài liệu này được xây dựng trong quá trình gán nhãn từ loại cho ngữ liệu thô.

Nội dung

1. Bảng từ loại.....	73
2. Nhập nhằng động từ - kết từ chính phụ	73
3. Nhập nhằng động từ - trợ từ.....	74
4. Nhập nhằng động từ - danh từ.....	74
5. Nhập nhằng động từ - tính từ.....	74
6. Nhập nhằng động từ - phó từ.....	74
7. Nhập nhằng danh từ - kết từ chính phụ.....	74
8. Nhập nhằng tính từ - danh từ.....	74
9. Nhập nhằng kết từ đẳng lập – phụ từ.....	75
10. Các định từ ở vị trí -3[NTCần77, DQBan05]	75
11. Các cụm danh từ dẫn xuất (hình thái dẫn xuất)	75
12. Các cụm từ “sau khi”, “trong khi”, v.v.....	75
13. Các từ “ra”, “vào”, “lên”, “xuống”, v.v.....	75
14. Danh sách các từ chức năng không có trong từ điển nhưng lại có trong sách ngữ pháp [DQBan05].....	76
15. Danh sách từ (hay tổ hợp từ) chức năng có trong từ điển nhưng lại không được phân loại	76
16. Nhóm từ chỉ vị trí: “trong”, “ngoài”, “trên”, “dưới”, v.v.....	76
17. Từ bổ nghĩa cho số từ	76
18. Trợ từ “một cách”	77
Tài liệu tham khảo	77

1. Bảng từ loại

STT	Nhãn	Tên	Ví dụ
1	N	Danh từ	tiếng, nước, thủ đô, nhân dân, đồ đặc, cây cối, chim muông
2	Np	Danh từ riêng	Nguyễn Du, Việt Nam, Hải Phòng, Trường Đại học Bách khoa Hà Nội, Mộc tinh, Hoả tinh, Phật, Đạo Phật
3	Nc	Danh từ chỉ loại	con, cái, đứa, bức
4	Nu	Danh từ đơn vị ¹³	mét, cân, giờ, năm, nhóm, hào, xu, đồng
5	V	Động từ	ngủ, ngồi, cười; đọc, viết, đá, đặt; thích, yêu, ghét, giống, muốn
6	A	Tính từ	tốt, xấu, đẹp; cao, thấp, rộng
7	P	Đại từ	tôi, chúng tôi, hắn, nó, y, đại nhân, đại ca, huynh, đệ
8	L	Định từ ¹⁴	mỗi, từng, mọi, cái; các, những, mấy
9	M	Số từ	một, mười, mười ba; dăm, vài, mười; nửa, rưỡi
10	R	Phó từ	đã, sẽ, đang, vừa, mới, từng, xong, rồi; rất, hơi, khi, quá
11	E	Giới từ ¹⁵ (kết từ chính phụ)	trên, dưới, trong, ngoài; của, trừ, ngoài, khỏi, ở
12	C	Liên từ (kết từ đẳng lập)	và, với, cùng, vì vậy, tuy nhiên, ngược lại
13	I	Thán từ	ôi, chao, a ha
14	T	Trợ từ, tình thái từ (tiểu từ) ¹⁶	à, a, á, à, ấy, chắc, chẳng, cho, chứ
15	B	Từ tiếng nước ngoài (hay từ vay mượn)	Internet, email, video, chat
16	Y	Từ viết tắt	OPEC, WTO, HIV
17	S	Yếu tố cấu tạo từ	bất, vô, gia, đa
18	X	Các từ không phân loại được	

Khi gán nhãn ngữ liệu, nhãn từ viết tắt sẽ là nhãn kép. Chẳng hạn nếu từ viết tắt là HIV thì nhãn của nó là Ny vì HIV viết đầy đủ thì là danh từ. Tương tự, nhãn từ vay mượn cũng là nhãn kép, ví dụ: email/Nb, chat/Vb

2. Nhập nhằng động từ - kết từ chính phụ

Trong câu sau “vào” là kết từ chính phụ (E) chứ không phải động từ.

Vào/E dịp/N nghỉ/V hè/N rất/R đông/A học sinh/N phổ thông/N theo/V học/V ./.

¹³ Mới bổ xung vào tháng 5/2008

¹⁴ Trong tài liệu hướng dẫn cũ là D

¹⁵ Trong từ điển (SP7.2) là O

¹⁶ Trong từ điển (SP7.2) phân ra hai loại là trợ từ (T) và cảm từ (E)

Hay “lên” trong ví dụ sau:

Rmah Rô/Np khoác/V gùi/N lên/E vai/N ./.

3. Nhập nhằng động từ - trợ từ

Bé đang tập đi/V .

trông sạch quá đi/T

4. Nhập nhằng động từ - danh từ

Hiện tượng nhập nhằng này khá phổ biến. Một số trường hợp là:

- Động từ chỉ hành động như “cày”, “cuốc”, “đọc”, v.v. chuyển thành danh từ chỉ đồ vật tương ứng “(cái) cày”, “(cái) cuốc”, “(cái) đọc”, v.v.

Tôi mượn *cuốc/N* để *cuốc/V* đám đất sau nhà.

- Động từ chỉ hành động như “suy nghĩ”, “đẩn đo”, v.v. chuyển thành danh từ chỉ khái niệm hay sự vật trừu tượng “(những) suy nghĩ”, “(những) đẩn đo”, v.v.

Vài suy nghĩ/N về Toán học Việt Nam

Hôm nay, đọc lại hồi kí Hồi ức và suy nghĩ/N của Trần Quang Cơ trên tinh thần “ôn cố tri tân”, bạn đọc Thông Luận có thể nhận ra nhiều điều thú vị.

5. Nhập nhằng động từ - tính từ

Đồng hồ này chạy/V rất chính xác .

Hàng bà ấy dạo này bán không *chạy/A* .

6. Nhập nhằng động từ - phó từ

Nhằm “được” là động từ trong trường hợp sau:

Mình/P không/R nuôi/V chúng nó/P *được/R* ./.

Từ này có độ nhập nhằng cao vì nó vừa có thể là động từ, vừa có thể là tính từ, phó từ, và trợ từ. Trong ví dụ trên “được” là phó từ vì nó biểu thị điều vừa nói đến (động từ “nuôi”) là có khả năng thực hiện.

7. Nhập nhằng danh từ - kết từ chính phụ

Nhà ấy nhiều *của/N* lắm !

sách *của/E* thư viện

8. Nhập nhằng tính từ - danh từ

Từ chỉ tính chất như “khó khăn”, “gian khổ”, v.v. chuyển thành danh từ chỉ sự vật trừu tượng “(những) khó khăn”, “(những) gian khổ”, v.v.

Tuy nhiên, sức chống đỡ trước *khó khăn/N* của Việt Nam đã tốt lên rất nhiều.

Tăng giá sách, *khó khăn/N* sẽ thêm chồng chất

9. Nhập nhằng kết từ đẳng lập – phụ từ

Cháu Ngọc rất bé mà/C rất khỏe .
Mẹ đã bảo mà/T !

10. Các định từ ở vị trí -3 [NTCần77, DQBan05]

Danh sách : "tất cả", "tất thủy", "toàn bộ", "hầu hết", "phần lớn"
tất thủy/L học sinh trường này, hầu hết/L những giáo sư này

11. Các cụm danh từ dẫn xuất¹⁷ (hình thái dẫn xuất)

Có nhiều cụm danh từ có cấu tạo gồm động từ hoặc tính từ đi sau danh từ chỉ loại. Ví dụ như: "sự nóng hổi", "vụ hổi lộ", v.v. Các từ này phải được gán nhãn là động từ hay tính từ.
sự/Nc nóng hổi/A; vụ/Nc hổi lộ/V
Trường hợp của một số cụm từ như: "những vất vả", "những thành công", v.v. Thì ta gán từ loại danh từ cho "vất vả", "thành công", v.v.

12. Các cụm từ "sau khi", "trong khi", v.v.

Các cụm từ này gồm hai từ, một là kết từ chính phụ (chẳng hạn "sau"), từ còn lại là danh từ thời gian (chẳng hạn "khi").
sau/E khi/N; trong/E khi/N; trước/E khi/N
Ngoài ra còn có: "đến khi", "tới khi", "trước lúc", "trong lúc", "đến lúc", "tới lúc", v.v.; "trong đó" cũng có thể được xếp vào nhóm này (trong/E đó/P).

13. Nhóm từ "ra", "vào", "lên", "xuống"

Danh sách : ra, vào, lên, xuống, đến, tới, sang, qua, lại, về
Các từ này có thể được gán nhãn là động từ, phó từ, và kết từ chính phụ (giới từ) tùy thuộc vào ngữ cảnh:

- Khi không có vị từ là thực từ (như "đi", "hiểu", "béo") đứng trước, các từ này có tư cách động từ:

Tôi/P ra/V sân/N. Tôi/P vào/V lớp học/N.

- Khi có vị từ là thực từ đứng trước, các từ này làm phó từ:

đi/V ra/R, bước/V xuống/R, đẩy/V xe/N ra/R, kéo/V xe/N lên/R

- Khi có động từ đằng trước¹⁸ và danh từ (thường là chỉ vị trí) đằng sau hoặc khi ở đầu câu, các từ này làm kết từ chính phụ:

đi/V xuống/E Hải Phòng/Np, đi/V lên/E Kontum/Np, bê/V bàn/N ra/E vườn/N

- Khi động từ đứng trước có mẫu NP-VP thì các từ này là động từ :

bảo/V nó/P xuống/V đây/P

¹⁷ Trong tiếng Anh hậu tố dẫn xuất làm thay đổi từ loại (happy/A → happiness/N), trong tiếng Việt thì lại sử dụng từ chức năng.

¹⁸ Chú ý là giữa từ ngữ cảnh (động từ, tính từ) và từ đang xét phải có quan hệ ngữ pháp

14. Danh sách các từ chức năng không có trong từ điển nhưng lại có trong sách ngữ pháp [DQBan05]

Kết từ chính phụ (giới từ): “cho đến”, “cho tới”, “để mà”, “để cho”, “cùng với”

Kết từ đẳng lập: “cũng như”

15. Danh sách từ (hay tổ hợp từ) chức năng có trong từ điển nhưng lại không được phân loại

Kết từ chính phụ: “đến nỗi”

khỏe/A đến nỗi/E có thể/V dùng/V hai/M tay/N nắm/V hai/M sừng/N trâu/N ghì/V xuống/E đất/N

Phó từ: “đúng ra”¹⁹, “lẽ ra”, “mới đây”, “vừa rồi”, “vừa qua”; “hình như”, “đường như”; “ít nhất”

đúng ra/R, /, nó/P phải/V bị/V kỷ luật/V

Các từ này có đặc điểm là có thể xuất hiện ở đầu câu hoặc giữa chủ ngữ và vị ngữ.

Kết từ đẳng lập: “chẳng hạn”; “ngoài ra”; “mặt khác”; “tức”, “tức là”; “nghĩa là”; “ngược lại”; “nói chung”, “nhìn chung”; “nói riêng”; “nói tóm lại”

Mặt khác/C, thông thường các sản phẩm tẩy trang đều chứa thành phần dầu.

Một số từ khác: “do đó”, “do vậy”, “tuy vậy”, “dẫu vậy”

Dẫu vậy/C, cả hai ông và bất kỳ người nào khác đều không thể thống nhất về những phẩm chất cần có, chính xác của một ông chủ Nhà Trắng.

16. Nhóm từ chỉ vị trí: “trong”, “ngoài”, “trên”, “dưới”, v.v.

Có một số trường hợp ta không làm theo hướng dẫn trong từ điển. Nhóm danh từ này thuộc một trong các số ít đó. Xét từ “trên” như một ví dụ.

(1) Danh từ theo từ điển → giới từ:

Máy bay/N lượn/V trên/E thành phố/N ./.

Nhà/N anh/P ở/V trên/E tầng/N năm/M ./.

(2) Danh từ theo từ điển → tính từ

hàng/N ghế/N trên/A ./.

đọc/V lại/R mấy/L trang/N trên/A ./.

Có hai lý do để ta qui định lại:

- Hành vi của từ “trên” chả khác gì giới từ trong trường hợp (1) và tính từ trong trường hợp (2)
- Qui định lại sẽ giúp việc gán nhãn đơn giản hơn (nhất quán hơn)

17. Từ bổ nghĩa cho số từ

Danh sách từ: “hơn”, “trên”, “dưới”, “gần”, “khoảng”, v.v.

hơn/A 200/M đại biểu/N, trên/A mười/M suất/N học bổng/N, khoảng/A 50/M bảng/Nu

Yêu cầu: Phân loại cho các từ này là tính từ (A).

¹⁹ [DQBan05] phân loại từ này là trợ từ

18. Trợ từ “một cách”

Nước/N xã/V Miwon/Np bỗng nhiên/R trong/A một cách/T kỳ lạ/A !/.

Trợ từ này có tác dụng nhấn mạnh. Nó luôn đứng ở trước bộ phận cần được nhấn mạnh [LBiên99].

19. Mẫu “từ .. sang ..”

Đó/P là/V hải_trình/N lớn/A nhất/R từ/E tây/N sang/E đông/N
“từ” và “sang” cần được gán nhãn là E. Đây là các thành phần của cụm giới từ chỉ hướng.

Tài liệu tham khảo

[DQBan05] Diệp Quang Ban. 2005. Ngữ pháp tiếng Việt (2 tập). NXB Giáo dục.

[LBiên99] Lê Biên. 1999. Từ loại tiếng Việt hiện đại. NXB Giáo dục.

[NTCần77] Nguyễn Tài Cần. 1977. Ngữ pháp tiếng Việt: tiếng – từ ghép – đoản ngữ. NXB ĐH & THCN

[UBKH83] Ủy ban Khoa học Xã hội Việt Nam. 1983. Ngữ pháp tiếng Việt. NXB Khoa học Xã hội.

[Vietlex08] Trung tâm Từ điển học. 2008. Từ điển tiếng Việt. NXB Đà Nẵng.

Thiết kế tập nhãn cú pháp và hướng dẫn gán nhãn

Nguyễn Phương Thái, Vũ Xuân Lương, Nguyễn Thị Minh Huyền

Đào Minh Thu, Đào Thị Minh Ngọc, Lê Kim Ngân

SP 7.3 – Dự án VLSP

Giới thiệu

Đây là tài liệu hướng dẫn gán nhãn cú pháp cho treebank tiếng Việt. Tập nhãn từ loại và hướng dẫn gán nhãn từ loại được trình bày trong một tài liệu khác. Với mỗi hiện tượng ngữ pháp, chúng tôi trình bày cách nhận diện, cách gán nhãn, cùng với các ví dụ cụ thể để minh họa. Các ví dụ được lấy từ sách ngữ pháp hoặc từ ngữ liệu thực tế. Chúng tôi cố gắng trích dẫn tài liệu tham khảo đầy đủ để người đọc có thể tự tìm hiểu thêm khi cần. Tài liệu này liên tục được chỉnh sửa và bổ sung trong quá trình thực hiện dự án.

Mục lục

1. Toàn bộ tập nhân	81
2. Chú ý chung khi gán nhãn cụm từ	83
2.1 Nhân phần tử trung tâm H	83
2.2 Cấu trúc với liên từ đẳng lập	84
3. Cụm danh từ	84
3.1 Cấu trúc chung	84
3.2 Nhập nhằng gán nhãn cụm danh từ cơ sở	85
4. Cụm động từ	86
4.1 Phần phụ trước	86
4.2 Phần phụ sau: bổ ngữ	87
4.3 Phần phụ sau: phụ ngữ	88
5. Cụm tính từ	89
6. Cụm phó từ	90
7. Cụm giới từ	90
8. Cụm từ chỉ số lượng	90
9. Ngữ tình thái	91
10. Câu trần thuật	91
11. Mệnh đề phụ kết (subordinate clause)	93
12. Câu hỏi	94
13. Câu cảm thán	95
14. Câu mệnh lệnh	96
15. Câu đặc biệt và tín báo	96
16. Các nhãn chức năng	97
16.1 Nhãn chức năng chủ ngữ	97
16.2 Nhãn chức năng tân ngữ trực tiếp	98
16.3 Nhãn chức năng tân ngữ gián tiếp	99
16.4 Nhãn chức năng khởi ngữ	100
16.5 Nhãn chức năng dành cho vị ngữ không phải cụm động từ	100
16.6 Nhãn chức năng của chủ ngữ logic	101
16.7 Nhãn chức năng bổ ngữ chỉ phạm vi hay tần suất của hành động	101
17. Nhãn phân loại phụ ngữ của động từ	102
17.1 Phụ ngữ thời gian	102
17.2 Phụ ngữ nơi chốn	103
17.3 Phụ ngữ chỉ hướng	103
17.4 Phụ ngữ chỉ cách thức hay phương tiện	103
17.5 Phụ ngữ chỉ mục đích hay lý do	104
17.6 Phụ ngữ chỉ điều kiện	105
17.7 Trạng ngữ chỉ ý nhượng bộ	105
17.8 Trạng ngữ	106
18. Nhãn phần tử rỗng	106
18.1 Nhãn *T*	107
18.2 Nhãn *0*	108
18.3 Nhãn *RNR*	108
19. Các cấu trúc sử dụng liên từ đẳng lập	108

19.1	Các từ trung tâm không có chung bổ ngữ.....	109
19.2	Các từ (hay cụm từ) có chung bổ ngữ	109
20.	Kết từ đẳng lập (C) và kết từ chính phụ (E)	109
21.	Thành phần chú thích hoặc trích dẫn	110
21.1	Thành phần chú thích	110
21.2	Thành phần trích dẫn.....	111
22.	Câu phức.....	112
22.1	Câu phức chủ ngữ.....	112
22.2	Câu phức vị ngữ	113
22.3	Câu phức bổ ngữ	113
22.4	Câu phức định ngữ	114
23.	Câu ghép	115
22.1	Câu ghép song song (UBKHXXH, 1983)	115
22.2	Câu ghép qua lại (UBKHXXH, 1983)	115
22.3	Phân biệt câu ghép với câu đơn có thành phần trạng ngữ	116
24.	Tính lược.....	117
25.	Câu bị động.....	122

1. Toàn bộ tập nhãn

Nhãn từ loại:

STT	Tên	Chú thích
	N	Danh từ
	Np	Danh từ riêng
	Nc	Danh từ chỉ loại
	Nu	Danh từ đơn vị ²⁰
	V	Động từ
	A	Tính từ
	P	Đại từ
	L	Định từ ²¹ (lượng từ)
	M	Số từ
	R	Phụ từ
	E	Giới từ ²²
	C	Liên từ
	I	Thán từ
	T	Trợ từ, tiểu từ, từ tình thái ²³
	U	Từ đơn lẻ
	Y	Từ viết tắt
	X	Các từ không phân loại được

Khi gán nhãn, nhãn từ viết tắt sẽ là nhãn kép. Chẳng hạn nếu từ viết tắt là HIV thì nhãn của nó là Ny vì HIV viết đầy đủ là danh từ.

Nhãn cụm từ:

STT	Tên	Chú thích
	NP	Cụm danh từ
	VP	Cụm động từ
	AP	Cụm tính từ
	RP	Cụm phụ từ
	PP	Cụm giới từ
	QP	Cụm từ chỉ số lượng
	MDP	Cụm từ tình thái
	UCP	Cụm từ gồm hai hay nhiều thành phần không cùng loại được nối với nhau bằng liên từ đẳng lập
	LST	Cụm từ đánh dấu đầu mục của danh sách
	WHNP	Cụm danh từ nghi vấn (ai, cái gì, con gì, v.v.)
	WHAP	Cụm tính từ nghi vấn (lạnh thế nào, đẹp ra sao, v.v.)
	WHRP	Cụm từ nghi vấn dùng khi hỏi về thời gian, nơi chốn, v.v.
	WHPP	Cụm giới từ nghi vấn (với ai, bằng cách nào, v.v.)

²⁰ Mới bổ sung vào tháng 5/2008

²¹ Trong tài liệu hướng dẫn cũ là D

²² Trong từ điển (SP7.2) là O

²³ Trong từ điển (SP7.2) phân ra hai loại là trợ từ (T) và cảm từ (E)

Nhãn mệnh đề:

STT	Tên	Chú thích
	S	Câu trần thuật (khẳng định hoặc phủ định)
	SQ	Câu hỏi
	S-EXC	Câu cảm thán
	S-CMD	Câu mệnh lệnh
	SBAR	Mệnh đề phụ kết (bổ nghĩa cho danh từ, động từ, và tính từ)
	SF	Câu mà chỉ có thể được giải thích hợp lý dưới quan điểm ngữ pháp chức năng

Nhãn chức năng cú pháp:

STT	Tên	Chú thích
	H	Phần tử trung tâm của cụm từ
	SUB	Nhãn chức năng chủ ngữ
	DOB	Nhãn chức năng tân ngữ trực tiếp
	IOB	Nhãn chức năng tân ngữ gián tiếp
	TPC	Nhãn chức năng chủ đề (khởi ngữ)
	PRD	Nhãn chức năng vị ngữ không phải cụm động từ
	LGS	Nhãn chức năng chủ ngữ logic của câu ở thể bị động
	EXT	Nhãn chức năng bổ ngữ chỉ phạm vi hay tần suất của hành động
	VOC	Nhãn chức năng thành phần than gọi
	TH	Nhãn phần thuyết của câu SF

Nhãn phân loại phụ ngữ của động từ:

STT	Tên	Chú thích
	TMP	Nhãn chức năng phụ ngữ chỉ thời gian
	LOC	Nhãn chức năng phụ ngữ chỉ nơi chốn
	DIR	Nhãn chức năng phụ ngữ chỉ hướng
	MNR	Nhãn chức năng phụ ngữ chỉ cách thức
	PRP	Nhãn chức năng phụ ngữ chỉ mục đích hay lý do
	CND	Nhãn chức năng phụ ngữ chỉ điều kiện
	CNC	Nhãn chức năng phụ ngữ chỉ ý nhượng bộ
	ADV	Nhãn chức năng trạng ngữ (khi không sử dụng được một trong các loại cụ thể trên)

Các nhãn khác:

STT	Tên	Chú thích
	T	Nhãn phần tử rỗng (lưu vết trong phạm vi câu)
	E	Nhãn phần tử rỗng ứng với hiện tượng tỉnh lược
	0	Nhãn phần tử rỗng ở vị trí tác tử phụ ngữ hóa

Các nhãn qui ước trong tài liệu này:

STT	Tên	Chú thích
	.	Nhãn dấu chấm câu, bao gồm: . ? !

,	Nhãn dấu phẩy
÷	Nhãn dùng cho cả dấu hai chấm và dấu gạch ngang chú thích

2. Chú ý chung khi gán nhãn cụm từ

2.1 Nhãn phần tử trung tâm H

Mô tả: Phần tử trung tâm của một cụm từ (ngữ đoạn) có các thuộc tính sau (CXHạo, 2007):

- Nó là yếu tố mang tất cả các thuộc tính ngữ pháp của ngữ đoạn
- Nó là yếu tố duy nhất của ngữ đoạn có thể có quan hệ ngữ pháp và ngữ nghĩa vượt ra ngoài ngữ đoạn
- Các yếu tố khác của ngữ đoạn chỉ có quan hệ phụ thuộc trực tiếp hay gián tiếp với trung tâm ngữ đoạn mà thôi (chứ không có bất cứ quan hệ gì ra ngoài phạm vi ngữ đoạn)

Cụm từ có một từ trung tâm:

(NP (N-H cô gái) (A đẹp))

(NP (N-H võ) (N dân tộc))

(NP (N-H nước) (Np Việt Nam))

Cụm từ có nhiều từ trung tâm:

(NP (NP-H (N-H chén))

(C và)

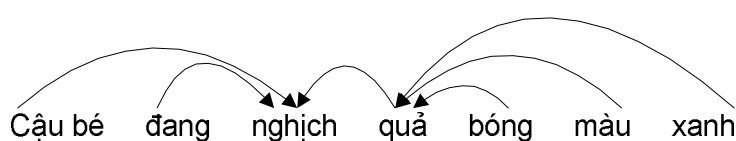
(NP-H (N-H đĩa) (A sạch)))

Cụm từ có nhiều từ trung tâm chung bỏ ngữ:

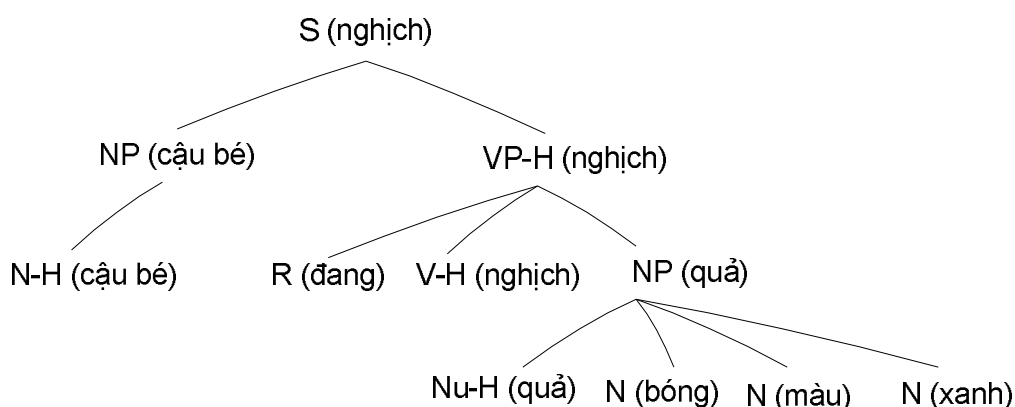
(NP (N-H chén) (C và) (N-H đĩa) (A sạch))

Hình 1 thể hiện một cây phụ thuộc trong đó các mũi tên đi từ từ bỏ nghĩa đến từ trung tâm.

Xét cụm danh từ “quả bóng màu xanh”, ta dễ dàng thấy là chỉ từ trung tâm “quả” là có liên hệ ra bên ngoài mà cụ thể là với động từ “nghịch”.



Hình 7. Cây phụ thuộc



Hình 8. Cây được từ vựng hóa

Có sự liên hệ chặt chẽ giữa cây phụ thuộc và cây cú pháp thành phần. Hình 2 biểu diễn cây cú pháp thành phần ứng với cây phụ thuộc trong Hình 1.

Thủ thuật xác định phần tử trung tâm (ĐMThu&NPThái, 2008):

- Lược: trung tâm là yếu tố mà nếu lược bỏ nó sẽ làm cho nghĩa và thuộc tính ngữ pháp của ngữ đoạn thay đổi
- Mở rộng văn cảnh: để xác định xem phần tử nào trong ngữ đoạn có quan hệ với văn cảnh
- Đặt câu hỏi: thay yếu tố phụ bằng từ nghi vấn
- Dựa vào các mô hình danh ngữ, động ngữ, tính ngữ, v.v. đã được tổng kết và các thành tố thường có của chúng để xác định nhanh thành tố trung tâm
- Chèn: để xác định xem có một hay nhiều thành tố trung tâm
- Qui tắc chính trước phụ sau: để nhận diện nhanh thành tố trung tâm

2.2 Cấu trúc với liên từ đẳng lập

a) Tính đối xứng

Khi đã có liên từ đẳng lập thì một cụm từ hay mệnh đề sẽ có nhiều phần tử trung tâm. Cần chú ý tính đối xứng của một cấu trúc có liên từ nối. Chẳng hạn như:

N C N
NP C NP

Chứ không nên:

N C NP
Hay NP C N

b) Nhãn UCP

Trường hợp liên từ đẳng lập được dùng để nối hai từ (hay cụm từ) không cùng loại, ta có thể dùng nhãn UCP.

(NP sản phẩm
(UCP (AP rẻ)
và
(NP chất lượng tốt)))

Không nhất thiết phải luôn dùng UCP trong mọi trường hợp, chẳng hạn như khi câu có hai vị ngữ, một cái là cụm tính từ và cái còn lại là cụm động từ. Tuy nhiên với những trường hợp tương tự ví dụ trên, việc dùng UCP là bắt buộc vì nếu không nhóm các thành phần đó lại sẽ gây nhập nhằng cấu trúc.

Ta sẽ quay lại nội dung liên quan đến liên từ đẳng lập trong một phần khác của tài liệu.

3. Cụm danh từ

Ký hiệu: NP

3.1 Cấu trúc chung

Cấu trúc cơ bản của một cụm danh từ như sau [1, trg24]:

<phần phụ trước> <danh từ trung tâm> <phần phụ sau>

Ví dụ: “một mái tóc đẹp” thì danh từ chỉ loại “mái” là trung tâm, “một” là phần phụ trước, còn danh từ “tóc” và tính từ “đẹp” thuộc phần phụ sau.

(NP (M một) (Nc-H mái) (N tóc) (A đẹp))

Một cụm danh từ có thể thiếu phần phụ trước hay phần phụ sau nhưng không thể thiếu phần trung tâm.

Phần phụ trước:

Phần này có tối đa hai thành phần²⁴:

<vị trí -2> <vị trí -1>

tất cả những chiếc kẹo
(NP (L tất cả)
(L những)
(Nc-H chiếc)
(N kẹo))

Ở vị trí -2 là định từ chỉ tổng lượng như “tất cả”, “hết thảy”, v.v. Ở vị trí -1 là số từ (hoặc cụm số từ) và định từ. Chi tiết cấu tạo từng thành phần xin tham khảo thêm trong [1, trg45].

Phần phụ sau:

Nói chung phần phụ sau của cụm danh từ có cấu tạo phức tạp hơn phần phụ trước. Bổ ngữ sau có thể là danh từ, cụm tính từ, cụm động từ, số từ, đại từ chỉ định, cụm giới từ, hay mệnh đề phụ. Đại từ chỉ định, nếu có, thì thường được đặt sau cùng. Sau đây là một số ví dụ:

Ví dụ 1: Cụm danh từ đơn giản (không có bổ ngữ là cụm giới từ, cụm động từ, hay mệnh đề phụ):

quả bóng màu xanh
(NP (Nc-H quả)
(N bóng)
(N màu)
(A xanh))

Ví dụ 2: Cụm danh từ phức tạp với bổ ngữ sau là cụm giới từ²⁵:

cái máy tính của cơ quan
(NP (Nc-H cái)
(N máy tính)
(PP của cơ quan))

Ví dụ 3: Cụm danh từ phức tạp với bổ ngữ sau là mệnh đề phụ:

cái máy tính mà tôi mới mua hôm qua
(NP (Nc-H cái)
(N máy tính)
(SBAR mà tôi mới mua hôm qua))

// Nên trao đổi về nhập nhằng khi gán nhãn NP ở đây

3.2 Nhập nhằng gán nhãn cụm danh từ cơ sở

Xét hai cụm từ “quả bóng xanh” và “bản đồ hàng lậu”:

(NP (Nc-H quả)
(N bóng)
(A xanh))

(NP (N-H bản đồ)
(NP (N-H hàng) (A lậu)))

²⁴ Cũng có quan điểm coi danh từ chỉ loại thuộc phần phụ trước, như vậy sẽ có tối đa ba thành phần

²⁵ Để đơn giản chúng tôi chưa mô tả cấu trúc cụ thể của PP và SBAR, chỉ nêu cụm từ tiếng Việt

Tại sao “quả bóng xanh” lại có cấu trúc phẳng hơn “bản đồ hàng lậu”? Nói cụ thể hơn, tại sao “bóng xanh” không được nhóm lại thành NP trong khi “hàng lậu” lại được? Lý do là vì cả “bóng xanh” và “quả xanh” đều chấp nhận được, do đó cả “quả” lẫn “bóng” đều có liên hệ với văn cảnh²⁶. Trong khi đó ta không thể nói “bản đồ lậu” vì nghĩa bị thay đổi. Như vậy nếu coi “hàng lậu” là cụm danh từ thì đảm bảo yêu cầu là chỉ có từ trung tâm (“hàng”) có liên hệ với văn cảnh.

4. Cụm động từ

Ký hiệu: VP

Cấu trúc chung:

Giống như cụm danh từ, cấu tạo một cụm động từ về cơ bản như sau:

<phần phụ trước> <động từ trung tâm> <phần phụ sau>

4.1 Phần phụ trước

Phần phụ trước của cụm động từ thường là phụ từ. Những nhóm tiêu biểu là²⁷:

- Những từ chỉ ra sự tồn tại của hoạt động trong thời gian và diễn tiến của hoạt động đối với thời gian: *đã, sẽ, đang, từng, còn, chưa, sắp, v.v.*
- Tiêu chí phủ định hay khẳng định: *không (chẳng, chả), có, chưa*
- Các từ chỉ ra khả năng diễn tiến của hoạt động, trạng thái: *cũng, vẫn, đều, lại, cứ, chỉ*
- Các từ với ý nghĩa mức độ của các đặc trưng vận động, tính chất: *rất, hơi, khi, quá*
- Các từ chỉ tần số (số lần) khái quát của sự xuất hiện hoạt động trạng thái: *thường, hay, năng, ít, hiếm, v.v.*

Ví dụ:

“đang ăn cơm”

(VP (R đang) (V-H ăn)
(NP-DOB (N-H cơm)))

“vẫn còn chưa ăn cơm”

(VP (R vẫn) (R còn) (R chưa)
(V-H ăn)
(NP-DOB (N-H cơm)))

Trong đó: *đang, vẫn, còn, chưa...* là các phụ từ.

Ngoài ra còn có các từ tượng thanh, tượng hình và một số tính từ có tác dụng miêu tả hành động, trạng thái nêu ở động từ. Ví dụ: tí tách rơi, ào ào tuôn, khẽ khàng đáp, tích cực đóng góp, cơ bản hoàn thành, v.v.

tí tách rơi

(AP (A-H tí tách) (V rơi))

²⁶ Từ “xanh”

²⁷ Các từ chỉ ra tình thái ngăn cấm, khuyên bảo: đừng, chớ, hãy, phải, cần, nên được coi là động từ trong treebank tiếng Việt.

4.2 Phần phụ sau: bổ ngữ

Động từ có khả năng kết hợp với các từ loại khác một cách rất đa dạng. Mỗi cách kết hợp có thể coi như một mẫu cú pháp của động từ: nội động từ, động từ đi với danh từ, động từ đi với cụm giới từ, động từ đi với mệnh đề, v.v. Ta xét các ví dụ sau:

Nội động từ:

đi
(VP (V-H đi))

Bổ ngữ là cụm danh từ:

yêu cô ấy
(VP (V-H yêu)
(NP-DOB (N-H cô) (P ấy)))

Bổ ngữ là cụm giới từ:

chuyển hàng xuống thuyền
(VP (V-H chuyển)
(NP-DOB (N-H hàng))
(PP-LOC/DIR (E-H xuống)
(NP (N-H thuyền))))

thanh toán bằng tiền mặt
(VP (V-H thanh toán)
(PP-MNR (E-H bằng)
(NP (N tiền mặt))))

Bổ ngữ là hai cụm danh từ :

Nhóm động từ trao nhận (LBiên, 1999) thường có cấu trúc này.

tặng bạn hai quyển sách
(VP (V-H tặng)
(NP-IOB (N-H bạn))
(NP-DOB (M hai) (Nc-H quyển) (N sách)))

Bổ ngữ là cụm danh từ và cụm giới từ :

pha cà phê với sữa
(VP (V-H pha)
(NP-DOB (N-H cà phê))
(PP-MNR (E-H với)
(NP (N-H sữa))))

Bổ ngữ là cụm động từ:

Nhóm động từ tình thái (LBiên, 1999) thường có cấu trúc này.

cần viết thư
(VP (V-H cần)
(VP (V-H viết)
(NP-DOB (N-H thư))))

Bổ ngữ là cụm danh từ và cụm động từ :

Nhóm động từ gây khiến (LBiên, 1999) thường có cấu trúc này.

nhờ bạn chép bài
 (VP (V-H nhờ)
 (NP-DOB (N-H bạn))
 (VP (V-H chép)
 (NP-DOB (N-H bài))))

Bổ ngữ là mệnh đề:

Nhóm động từ nói năng, cảm nghĩ trạng thái (LBiên, 1999) thường có cấu trúc này.

nói rằng cô ấy đẹp
 (VP (V-H nói)
 (SBAR-DOB (C rằng)
 (S (NP-SUB (N-H cô) (P ấy))
 (AP-PRD (A-H đẹp))))))

Bổ ngữ là cụm tính từ:

Thường áp dụng cho các động từ chỉ quan hệ biến hóa (LBiên, 1999).

Cô ấy trở nên hung dữ.
 (S (NP-SUB (N-H cô) (P ấy))
 (VP (V-H trở nên)
 (AP (A hung dữ))))
 (. .))

4.3 Phần phụ sau: phụ ngữ

Ngoài bổ ngữ, góp phần cấu tạo nên cụm động từ còn có phụ ngữ. Phụ ngữ có thể là phụ từ, cụm tính từ, cụm danh từ chỉ thời gian, cụm giới từ, hoặc mệnh đề phụ. Các ví dụ về phụ ngữ là cụm giới từ hoặc mệnh đề phụ xin xem trong phần 17.

Phụ ngữ là cụm tính từ²⁸ :

đi rất nhanh
 (VP (V-H đi)
 (AP (R rất) (A-H nhanh)))
 Cô ấy hát không hay.
 (S (NP-SUB (N-H cô) (P ấy))
 (VP (V-H hát)
 (AP (R không) (A-H hay)))
 (. .))

Các nhóm phụ từ:

- Nhóm từ chỉ ý kết thúc: *rồi, đã*
- Nhóm từ chỉ ý cầu khiến (mệnh lệnh, mời mọc, rủ rê) dùng với người ngang hàng hoặc bề dưới gồm có: *đi, nào, thôi*
 học đi, nghỉ nào, ăn thôi, v.v.
- Nhóm từ chỉ kết quả gồm:
 - o chỉ sự vừa ý: *được*
 chơi được, cưới được, yêu được, v.v.
 - o chỉ sự tiếc : *mất*
 chết mất, đánh mất, làm mất, v.v.

²⁸ Thường là MNR

- chỉ ý không mong muốn: *phải*
gặp phải kẻ trộm, mua phải hàng giả, v.v.
- chỉ sự tự lực: *lấy*
nấu ăn lấy, đóng lấy, viết lấy, v.v.
- Nhóm từ chỉ sự cùng chung: *với, cùng*
cho nó đi với!; để bạn học cùng
- ~~Nhóm từ chỉ sự qua lại, tương hỗ: *nhau*
gửi thư cho nhau, làm việc cùng nhau, v.v.~~

5. Cụm tính từ

Ký hiệu: AP

Cấu trúc chung:

Cấu tạo một cụm tính từ về cơ bản như sau:

<phần phụ trước> <tính từ trung tâm> <phần phụ sau>

Phần phụ trước:

- Phần phụ trước của tính từ thường là phụ từ chỉ mức độ: *rất, hơi, khá, cực, tuyệt, quá*

Ví dụ:

rất đẹp

(AP (R rất) (A-H đẹp))

khá xinh

(AP (R khá) (A-H xinh))

- Thêm vào đó nhiều phụ từ đi với động từ (xem 4.1) cũng có thể đi với tính từ
lúa còn xanh, nhà đang bận, đèn chưa sáng

Phần phụ sau:

Bổ ngữ sau có thể là phụ từ chỉ mức độ như trong ví dụ sau:

xinh quá

(AP (A-H xinh) (R quá))

đẹp lắm

(AP (A-H đẹp) (R lắm))

Bổ ngữ sau có thể là cụm danh từ:

mỏng cùi

(AP (A-H mỏng)

(NP (N cùi)))

xa trung tâm thành phố

(AP (A-H xa)

(NP (N-H trung tâm) (N thành phố)))

gần ngày Tết Hàn thực

(AP (A-H gần)

(NP (N-H ngày)

(NP (N-H Tết) (Np Hàn thực))))

Bổ ngữ sau có thể là cụm giới từ:

giỏi về thể thao

(AP (A-H giỏi)

(PP (E-H về)

(NP (N thể thao))))

Nhóm các phụ từ *ra, lên, đi, lại* khi kết hợp với tính từ, không chỉ hướng mà chỉ ra các kết quả diễn biến của đặc trưng.

béo ra, béo lên

Các phụ từ chỉ mức độ: *lắm, quá, cực, tuyệt*

tốt lắm, thơm lắm, đẹp cực (khẩu ngữ)

6. Cụm phó từ

Ký hiệu: RP

Cụm phó từ chủ yếu tạo bởi sự kết hợp giữa các phó từ với nhau. Ví dụ như :

vẫn chưa²⁹

(RP (R vẫn) (R chưa))

7. Cụm giới từ

Ký hiệu: PP

Cấu trúc chung :

<giới từ> <cụm danh từ>

hoặc <giới từ> <cụm động từ>

Ví dụ :

vào Sài Gòn (trong "đi vào Sài Gòn")

(PP (E-H vào)

(NP (Np-H Sài Gòn))))

để tìm cơ hội du học (trong "tôi hay vào mạng để tìm cơ hội du học")

(PP (E-H để)

(VP (V-H tìm)

(NP-DOB (N-H cơ hội) (V du học))))

8. Cụm từ chỉ số lượng

Ký hiệu : QP

²⁹ Gán nhãn H ở đầu cũng được, chọn phần từ đầu cho tiện

Cấu trúc chung :

Thành phần chính của QP là các số từ. Có thể là số từ xác định, số từ không xác định, hay phân số. Ngoài ra còn có thể có phụ từ như "khoảng", "hơn", v.v. QP thường đóng vai trò là thành phần phụ trước của cụm danh từ (vị trí -1).

Ví dụ 1:

năm trăm
(QP (M-H năm) (M trăm))

Ví dụ 2:

hơn 200
(QP (R hơn) (M-H 200))

9. Ngữ tình thái

Ký hiệu: MDP (modal phrase)

Mô tả: Tình thái ngữ là thành phần phụ của câu, có nhiệm vụ bổ sung những ý nghĩa về tình thái cho câu. Khi gán nhãn cần chú ý phân biệt tình thái ngữ với: thán từ làm thành một vế của câu ghép, tiểu từ nhấn mạnh, và vị ngữ (NMThuyết và NVHiệp, 1999; tr 235-242).

Cô ta hồi hận thì có³⁰.

(S (NP-SUB Cô ta)
(VP (V-H hồi hận))
(MDP (T thì có))
(. .))

10. Câu trần thuật

Ký hiệu : S

Cấu trúc chung :

Theo quan điểm coi cấu trúc chủ-vị là cấu trúc chủ đạo của câu tiếng Việt [1], câu trần thuật sẽ có cấu trúc sau:

<chủ ngữ> <vị ngữ>

Trong đó chủ ngữ thường là cụm danh từ, còn vị ngữ thường là cụm động từ hoặc cụm tính từ. Với một số ngôn ngữ như tiếng Anh, vị ngữ luôn là cụm động từ.

Ví dụ :

Anh yêu em .
(S (NP-SUB (N-H Anh))
(VP (V-H yêu)
(NP-DOB (N-H em))))
(. .))

Nhãn chức năng chủ từ cho ta biết đâu là chủ ngữ của câu (chủ ngữ bề mặt). Cụm động từ theo sau chủ từ sẽ là vị ngữ. Nếu vị ngữ không phải cụm động từ thì sẽ được gán nhãn chức năng PRD.

³⁰ Trong Từ điển Tiếng Việt, “thì có” được phân loại là *khẩu ngữ*.

Chủ ngữ :

Chủ ngữ là cụm danh từ:

“Việc dậy đúng giờ thật khó.”

(S (NP-SUB (N-H Việc)

(VP (V-H dậy)

(AP-MNR (A-H đúng) (N giờ))))

(AP-PRD (R thật) (A-H khó))

(. .))

Chủ ngữ là cụm động từ:

“Dậy đúng giờ thật khó .”

(S (VP-SUB (V-H dậy)

(AP (A-H đúng) (N giờ)))

(AP-PRD (R thật) (A-H khó))

(. .))

Chủ ngữ cũng có thể là cụm chủ vị:

(S (S-SUB (NP-SUB (N-H Anh))

(VP (V-H nói) (P thế)))

(AP-PRD (R không) (A-H đúng) (T đâu)))

(. .))

Vị ngữ :

Vị ngữ là cụm động từ:

Tôi đi học .

(S (NP-SUB (P-H Tôi))

(VP (V-H đi)

(VP (V học)))

(. .))

Vị ngữ là cụm tính từ:

Nhà anh ấy xa .

(S (NP-SUB (N-H nhà)

(NP (N-H anh) (P ấyyy)))

(AP-PRD (A-H xa))

(. .))

Vị ngữ cũng có thể là cụm danh từ:

Em bé 7 tuổi.

(S (NP-SUB (N-H em bé))

(NP-PRD (M 7) (Nu-H tuổi)))

(. .))

Sự đa dạng trong cấu trúc của cụm động từ và cụm tính từ khiến cho cấu trúc của câu trần thuật cũng rất đa dạng. Về các khuôn hình câu đơn³¹, DQBan (2005) đã mô tả khá đa dạng.

Câu với động từ “có”:

Có con chuột trong góc nhà.

(S (VP (V-H có)

(NP-SUB (Nc-H con) (N chuột))

(PP-LOC (E-H trong)

³¹ Chú ý là nhân S không chỉ dùng cho câu đơn.

(NP (N-H góc) (N nhà))))
(. .))

Động từ này đặc biệt ở chỗ nó đứng đầu câu và danh từ theo sau là chủ ngữ của câu. Động từ này chỉ sự tồn tại.

// Bổ sung một số động từ tương tự: lấp lánh (?)

// Thêm mẫu câu (những cái chưa được nói đến ở phần cụm động từ)

11. Mệnh đề phụ kết (subordinate clause)

Ký hiệu : SBAR

Cấu trúc và chức năng:

Mệnh đề phụ kết đóng vai trò bổ nghĩa cho danh từ, động từ, hay tính từ. Bản thân nó không thể đứng độc lập làm thành một câu. Về cơ bản cấu trúc của mệnh đề phụ bao gồm một liên từ phụ kết và một mệnh đề (ký hiệu S).

Bổ nghĩa cho danh từ :

Quyển sách mà anh mượn
(NP (Nc-H Quyển) (N sách)
(SBAR (C mà)
(S (NP-SUB (N-H anh))
(VP (V-H mượn))))))

Bổ nghĩa cho động từ:

không đi đá bóng vì bạn gái ốm
(VP (R không) (V-H đi)
(VP-PRP (V-H đá)
(NP-DOB (N-H bóng)))
(SBAR-PRP (E-H vì)
(S (NP-SUB (N-H bạn) (N gái))
(AP-PRD (A-H ốm))))))

Trong ví dụ này mệnh đề phụ kết "vì bạn gái ốm" chỉ nguyên nhân của hành động "không đi đá bóng", vì thế có thêm nhãn PRP.

Bổ nghĩa cho tính từ :

khỏe vì chơi thể thao đều đặn
(AP (A-H khỏe)
(SBAR-PRP (C vì)
(S (NP-SUB *T*)
(VP (V-H chơi)
(NP-DOB (N-H thể thao))
(AP-MNR (A-H đều đặn))))))

Chú ý:

- Nhãn này chỉ dùng để thể hiện mệnh đề tính ngữ (bổ nghĩa cho danh từ), mệnh đề trạng ngữ (bổ nghĩa cho động từ hoặc câu), hoặc mệnh đề danh ngữ (bổ ngữ của động từ, tính từ).

- Không dùng SBAR ngay sau giới từ trung tâm của cụm giới từ. Không gán nhãn SBAR cho câu đơn mà làm thành phần của một câu phức (phân biệt với trường hợp mệnh đề trạng ngữ).

12. Câu hỏi

Ký hiệu : SQ

Khi ta đã thành thạo việc gán nhãn câu trần thuật, việc gán nhãn cho câu hỏi sẽ trở nên đơn giản hơn. Ta xem xét các dạng câu hỏi chính dưới đây :

Câu hỏi chuyên biệt (wh-question):

Loại câu hỏi này được dùng để hỏi về người, vật, địa điểm, thời gian, v.v.

Hỏi người, vật:

Ai đang ở trong nhà ?
 (SQ (WHNP-SUB (P-H Ai))
 (VP (R đang) (V-H ở)
 (PP-LOC (E-H trong)
 (NP (N-H nhà))))
 (. ?))

Cụm danh từ nghi vấn (WHNP) được sử dụng trong loại câu hỏi này. Cụm danh từ nghi vấn có thể là một đại từ nghi vấn (ai) hoặc là một cụm danh từ có đại từ nghi vấn làm bổ ngữ sau (cái gì, con gì).

Hỏi thời gian:

Bao giờ anh đi hội nghị ?
 (SQ (WHADV (P-H Bao giờ))
 (NP-SUB (N-H anh))
 (VP (V-H đi)
 (NP (N-H hội nghị))))
 (. ?))

Hỏi cách thức:

Anh sẽ giải bài toán này bằng cách nào ?
 (SQ (NP-SUB (N-H anh))
 (VP (R sẽ) (V-H giải)
 (NP-DOB (N-H bài toán) (P này))
 (WHPP (E-H bằng)
 (NP (N-H cách) (P nào))))
 (. ?))

Cụm giới từ nghi vấn (WHPP) là do giới từ kết hợp với cụm danh từ nghi vấn tạo ra.

Hỏi về trạng thái:

Bàn tay của cô ấy mềm mại ra sao ?
 (S (NP-SUB (N-H bàn tay)
 (PP (E-H của)
 (NP (N-H cô) (P ấy))))
 (WHAP-PRD (A-H mềm mại)
 (VP (V-H ra) (P sao))))

(. ?))

Cụm tính từ nghi vấn (WHAP) là do tính từ kết hợp với đại từ nghi vấn tạo ra.

Câu hỏi có-không (yes-no question):

Loại sử dụng cặp phụ từ trái nghĩa “có... không”, “đã... chưa”, v.v.

Ví dụ:

Em có đi chơi không ?
(SQ (NP-SUB (N-H em))
(VP (R có) (V-H đi)
(VP (V chơi))
(R không))
(. ?))

Cũng có thể chỉ sử dụng một phụ từ:

Ví dụ³²:

Mai anh đi chưa ?
(SQ (NP-TMP (N-H Mai))
(NP-SUB (N-H anh))
(VP (V-H đi) (R chưa))
(. ?))

Hoặc dùng tiểu từ tình thái:

Ví dụ :

Cô ấy chưa về nhỉ ?
(SQ (NP-SUB (N-H Cô) (P ấy))
(VP (R chưa) (V-H về))
(T nhỉ)
(. ?))

// Phụ từ thì ở mức VP còn tiểu từ thì mức câu (?)

13. Câu cảm thán

Ký hiệu : S-EXC

Cấu trúc chung :

Câu cảm thán dùng để thể hiện tình cảm của người nói, người viết, đối với hoặc bên cạnh sự tình được nêu. Loại câu này cũng có những đặc trưng về mặt hình thức, chẳng hạn như sử dụng thán từ (ôi, ơi, ơi là), tiểu từ (thay), trợ từ (lạ, thật), v.v.

Câu cảm thán sử dụng thán từ :

Ôi sức trẻ !
(S-EXC (I Ôi)
(NP (N-H sức) (N trẻ))
(. !))

³² Trong ví dụ này, TMP là nhân phụ ngữ chỉ thời gian.

Câu cảm thán sử dụng tiểu từ "thay" :

Vinh quang thay những vị anh hùng dân tộc !

(S-EXC (AP-PRD Vinh quang) (T thay))

(NP-SUB (L những) (Nc-H vị) (N anh hùng) (N dân tộc))

(. !))

Trong ví dụ này cụm tính từ vị ngữ đứng trước cụm danh từ chủ ngữ. Chúng được nối với nhau bằng tiểu từ "thay".

Câu cảm thán dùng trợ từ:

Con này gớm thật !

(S-EXC (NP-SUB (N-H Con) (P này))

(VP (V-H gớm) (T thật))

(. !))

14. Câu mệnh lệnh

Ký hiệu : S-CMD

Cấu trúc chung :

Câu mệnh lệnh của tiếng Việt được cấu tạo nhờ những phụ từ tạo ý mệnh lệnh, bằng ngữ điệu mệnh lệnh, và chỉ được chứa những từ liên quan đến nội dung của lệnh (đảm bảo tính ngắn gọn) [1]. Các phụ từ mệnh lệnh hay dùng là : hãy, đừng, chớ, đi, thôi, v.v.

Ví dụ 1 :

Không được làm ồn !

(S-CMD (VP (R không) (V-H được)

(VP (V-H làm) (A ồn)))

(. !))

Ví dụ 2 :

Đi đi, em !

(S-CMD (VP (V-H đi) (R đi))

(, ,)

(NP-SUB (N-H em))

(. !))

15. Câu đặc biệt và tín báo

15.1 Câu đặc biệt

Nhãn chức năng: SPL

Mô tả : Trong (UBKHXH, 1983) các tác giả phân biệt nhiều loại câu đặc biệt, trong đó có loại câu xác định trạng thái tồn tại của sự vật. Có một điều cần chú ý là ta không coi câu với động từ “có” là câu đặc biệt.

Tiếng reo.

(S-SPL (NP (N-H Tiếng)(V reo))

(. .))

Giỏi thật!

(S-SPL-EXC(AP(A-H Giỏi)(T thật))

(. !))

15.2 Tít báo

Nhãn chức năng: TTL

Mô tả : Nhãn này được dùng để gán cho cấu trúc cú pháp làm tiêu đề của một bài báo. Tiêu đề có thể là cụm từ hay câu. Sau đây là ví dụ về cụm danh từ và câu làm tiêu đề:

Cuộc đời sau tấm màn nhung

(NP-TTL (N-H Cuộc đời))

(PP (E-H sau)

(NP (Nc-H tấm) (N màn) (N nhung))))

Con thi, cha mẹ cũng thi

(S-TTL (S (NP-SUB (N-H Con))

(VP (V-H thi)))

(,,)

(S (NP-SUB (N-H cha mẹ))

(VP (R cũng) (V-H thi)))

(. .))

16. Các nhãn chức năng

Thông tin cú pháp cơ bản nhất được thể hiện trong cây cú pháp qua các nhãn từ loại, cụm từ, và mệnh đề. Tuy nhiên, trong các ứng dụng của treebank [] nhiều trường hợp cần thông tin cụ thể hơn nữa. Do đó nhãn chức năng được sử dụng để làm giàu thông tin thể hiện trong cây cú pháp.

16.1 Nhãn chức năng chủ ngữ

Ký hiệu : SUB

Mô tả : Nhãn này được dùng để gán cho cụm từ làm chủ ngữ ở trong câu. Xét về vị trí, chủ ngữ thường đứng trước vị ngữ.

Ví dụ :

Anh này là sinh viên .

(S (NP-SUB (N-H Anh) (P này))

(VP (V-H là)

(NP-DOB (N-H sinh viên))))

(. .))

Trong ví dụ này nhãn chức năng vị ngữ không được sử dụng vì sau cụm từ có nhãn chức năng chủ ngữ, nếu tồn tại cụm động từ (không phải phụ ngữ) thì cụm đó là vị ngữ. Trong trường hợp vị ngữ không phải cụm động từ, khi đó ta mới dùng nhãn chức năng vị ngữ PRD (phần 16.5).

16.2 Nhãn chức năng tân ngữ trực tiếp

Ký hiệu : DOB

Mô tả : Nhãn này được dùng để gán cho cụm từ làm tân ngữ (object) của động từ trong cụm động từ. DOB ứng với vai bị thể trong [NVHiệp, 2008]. Các dấu hiệu để nhận biết DOB :

- Nó trả lời câu hỏi: ai, cái gì.
- Nó có thể làm chủ ngữ của câu bị động
- Về vị trí, nó thường đứng sau động từ

Ví dụ :

Tôi lái **ô tô**.
(S (NP-SUB (P-H Tôi))
(VP (V-H lái)
(NP-DOB (N-H ô tô)))
(. .))

Hỏi : Ai lái ô tô ?

Dạng bị động : Ô tô được tôi lái.

Vị trí : sau động từ "lái"

Các trường hợp đặc biệt:

Có một số động từ mà theo sau là danh từ nhưng danh từ đó không được gán nhãn DOB. Xét động từ “là”:

Tôi là sinh viên.
(S (NP-SUB (P-H Tôi))
(VP (V-H là)
(NP (N-H sinh viên)))
(. .))

Rõ ràng “sinh viên” là danh từ đi sau động từ “là” nhưng không thể gán cho nó nhãn chức năng DOB, vì nó không phải là đối tượng bị tác động bởi chủ thể “tôi”. Như vậy câu không có dạng bị động.

Tương tự:

bằng
Cái âm này bằng nhôm.
(S (NP-SUB (Nc-H cái) (N ấ m) (P này))
(PP-PRD (E-H bằng)
(NP (N-H nhôm)))
(. .))

tại
Việc này tại anh ấy.
(S (NP-SUB (N-H Việc) (P này))
(PP-PRD (E-H tại)
(NP (N-H anh) (P ấ y)))
(. .))

của

Cái áo này của tôi.

(S (NP-SUB (Nc-H Cái) (N áo) (P này))
(PP-PRD (E-H của)
(NP (P-H tôi))))
(. .))

như

Anh ấy như người ốm.

(S (NP-SUB (N-H Anh) (P ấy))
(AP-PRD (A-H như)
(NP (N-H người) (A ốm))))
(. .))

(Có lẽ câu đầy đủ là “Anh ấy trông như người ốm.”?)

có

Anh ấy có chiếc xe mới.

(S (NP-SUB (N-H Anh) (P ấy))
(VP (V-H có)
(NP (Nc-H chiếc) (N xe)
(AP (A-H mới))))))
(. .))

lên

Em bé này lên 10 tuổi.

(S (NP-SUB (N-H em) (N bé) (P này))
(VP (V-H lên)
(NP (M 10) (Nu-H tuổi))))
(. .))

Một trường hợp khác, xin xem phần 16.7.

16.3 Nhân chức năng tân ngữ gián tiếp

Ký hiệu : IOB

Mô tả : Nhân này được dùng để gán cho cụm từ làm tân ngữ gián tiếp (indirect object) của động từ trong câu. IOB ứng với vai tiếp thể và vai kẻ hưởng lợi trong [NVHiệp, 2008]. Các dấu hiệu để nhận biết :

- Tồn tại giới từ "cho" (cho ai, cho cái gì) trước cụm danh từ đang xét
- Có thể chèn "cho" vào trước cụm danh từ đang xét

Ví dụ:

Tôi tặng bạn quyển sách .

(S (NP-SUB (P-H Tôi))
(VP (V-H tặng)
(NP-IOB (N-H bạn))
(NP-DOB (Nc-H quyển) (N sách))))
(. .))

Động từ “tặng” trong ví dụ này có hai tân ngữ. Tân ngữ trực tiếp là “sách” thì được gán nhãn chức năng DOB, còn tân ngữ gián tiếp “bạn” thì có nhãn IOB.

16.4 Nhãn chức năng khởi ngữ

Ký hiệu : TPC

Mô tả : Khởi ngữ còn được gọi là chủ đề, thành phần khởi ý, v.v. Có nhiều nhà ngôn ngữ học đã nghiên cứu về vấn đề này. Chúng tôi coi sách của Nguyễn Minh Thuyết và Nguyễn Văn Hiệp (1999; tr 187-214) và Cao Xuân Hạo (2006) là các tài liệu tham khảo chính. Khi gán nhãn cần chú ý phân biệt khởi ngữ với các thành phần khác của câu như trạng ngữ và vế của câu ghép (NMThuyết và NVHiệp, 1999; tr 200-204).

Ví dụ:

Vấn đề này chúng tôi đang bàn .

(S-TC (NP-TPC-1 (N-H Vấn đề) (P này))
(NP-SUB (P-H chúng tôi))
(VP (R đang) (V-H bàn)
(NP-DOB-1 *T*))
(. .))

Trong ví dụ này, khởi ngữ (chủ đề) của câu là “vấn đề này”. Chú ý là nhãn chức năng TC (đề-thuyết hay topic-comment) được dùng để gán cho câu (nhãn S) có thành phần khởi ngữ.

Giàu thì tôi đã giàu rồi.

(S-TC (AP-TPC (A-H Giàu))
(C thì)
(S (NP-SUB (P-H tôi))
(AP-PRD (R đã) (A-H giàu) (T rồi)))
(. .))

Tiền mất 3 triệu đồng, lại cũng mệt mỏi.

(S (S-TC (NP-TPC(N-H Tiền))
(S (NP-SUB *E*))
(VP (V-H mất)
(NP-DOB (M 3) (M triệu) (Nu-H đồng))))
(, ,)
(S (NP-SUB *E*))
(AP-PRD (R lại) (R cũng) (A-H mệt mỏi)))
(. .))

16.5 Nhãn chức năng dành cho vị ngữ không phải cụm động từ

Ký hiệu : PRD

Mô tả : Nếu vị ngữ của câu không phải là một cụm động từ thì nó được gán nhãn PRD. Nói chung ngoài cụm động từ, cụm tính từ và cụm danh từ cũng có thể làm vị ngữ trong câu. Trong tiếng Việt, cụm tính từ làm vị ngữ là hiện tượng phổ biến.

Ví dụ 1:

Cô gái đẹp .
(S (NP-SUB (N-H Cô) (N gái))
(AP-PRD (A-H đẹp))
(. .))

Ví dụ 2:

Nhà này 60 mét vuông .
(S (NP-SUB (N-H Nhà) (P này))
(NP-PRD (M 60) (Nu-H mét vuông))
(. .))

16.6 Nhãn chức năng của chủ ngữ logic

Ký hiệu : LGS (logical subject)

Mô tả : Với một câu bị động tiếng Việt được viết đúng ngữ pháp [1, trg149], ta không cần đến nhãn này. Tuy nhiên hiện nay có hiện tượng viết sai ngữ pháp do ảnh hưởng của tiếng Anh. Nếu gặp những câu như vậy thì ta dùng thêm nhãn chức năng LGS. Dấu hiệu nhận biết:

- Tồn tại giới từ “bởi” trước cụm từ đang xét.

Ví dụ 1³³:

Yahoo! 360⁰ có thể bị thay thế bởi Yahoo! Mash
(S (NP-SUB (Np-H Yahoo! 360⁰))
(VP (V-H có thể)
(VP (V-H bị)
(SBAR (S (NP-SUB-1 *T*)
(VP (V-H thay thế)
(PP-MNR (E-H bởi)
(NP-LGS-1 (Np-H Yahoo! Mash)))))))))
(. .))

Ví dụ này được lấy từ tiêu đề của một bài báo gần đây trên báo Tuổi Trẻ Online.

Ví dụ 2:

Yahoo! 360⁰ có thể bị Yahoo! Mash thay thế
(S (NP-SUB (Np-H Yahoo! 360⁰))
(VP (V-H có thể)
(VP (V-H bị)
(SBAR (S (NP-SUB (Np-H Yahoo! Mash))
(VP (V-H thay thế))))))
(. .))

Câu trong ví dụ 1 được sửa cho đúng với ngữ pháp tiếng Việt hơn. Khi đó ta không dùng nhãn LGS nữa.

16.7 Nhãn chức năng bổ ngữ chỉ phạm vi hay tần suất của hành động

Ký hiệu: EXT

Mô tả: Nếu cụm danh từ chỉ phạm vi hay tần suất làm bổ ngữ sau cho động từ thì được gán nhãn EXT. Chú ý là trong trường hợp này cụm danh từ không phải tân ngữ (DOB).

Ví dụ:

Anh ấy chạy 5 km .
(S (NP-SUB (N-H Anh) (P ấy))

³³ <http://www.tuoiitre.com.vn/Tianyon/Index.aspx?ArticleID=220683&ChannelID=16>

(VP (V-H chạy)
(NP-EXT (M 5) (Nu-H km)))

(. .))

Thành phần than gọi

Nhãn chức năng: VOC

Mô tả: Thành phần này nêu lên một lời than hay lời gọi (UBKHXH, 1983). Nó thường do một cảm từ hay một ngữ có tác dụng như cảm từ đảm nhiệm.

Chao, đường còn xa lắm!

(S-EXC (MDP-VOC (I Chao))

(, .)

(NP-SUB (N-H đường))

(AP-PRD (R còn) (A-H xa) (R lắm))

(. !))

Trong ví dụ trên “chao” là cảm từ.

Anh ơi, chờ em với !

(S-EXC (NP-SUB-VOC (N-H Anh) (I ời))

(, .)

(VP (V-H chờ)

(NP-DOB (N-H em))

(R với))

(. !))

Trong ví dụ trên “ơi” là cảm từ và được phân tích như thành phần của cụm danh từ. Như vậy “anh ời” có hai chức năng chủ ngữ và than gọi.

17. Nhãn phân loại phụ ngữ của động từ

Phụ ngữ (hay trạng ngữ) là thành phần câu đóng vai trò thiết lập tình huống diễn ra hành động hay trạng thái mà động từ chính mô tả. Về hình thức, phụ ngữ có thể là từ, cụm từ, hay mệnh đề. Về ý nghĩa, phụ ngữ thường diễn tả: thời gian, nơi chốn, cách thức, nguyên nhân, mục đích, hay điều kiện.

17.1 Phụ ngữ thời gian

Ký hiệu: TMP

Ví dụ:

Ngày mai tôi đi thi .

(S (NP-TMP (N-H Ngày) (N mai))

(NP-SUB (P-H tôi))

(VP (V-H đi)

(VP (V thi)))

(. .))

Tôi hay nghe nhạc trong khi làm việc.

(S (NP-SUB (P-H Tôi))

(VP (R hay) (V-H nghe)

(NP-DOB (N-H nhạc))
 (PP-TMP (E-H trong)
 (NP (N-H khi)
 (VP (V-H làm)
 (NP (N-H việc))))))
 (. .))

Tôi mới lên Hà Nội được ba tháng.

(S (NP-SUB (P-H Tôi))
 (VP (R mới) (V-H lên)
 (NP-LOC (Np-H Hà Nội))
 (VP-TMP (V-H được)
 (NP (M ba) (N-H tháng))))
 (. .))

17.2 Phụ ngữ nơi chốn

Ký hiệu: LOC

Dấu hiệu: Là cụm danh từ chỉ địa điểm hoặc cụm giới từ chỉ địa điểm (giới từ có tác dụng đánh dấu vai địa điểm).

Ví dụ:

Tôi sẽ đi nghỉ ở Tokyo .
 (S (NP-SUB (P-H Tôi))
 (VP (R sẽ) (V-H đi)
 (VP (V-H nghỉ)
 (PP-LOC (E-H ở)
 (NP (Np-H Tokyo))))))
 (. .))

17.3 Phụ ngữ chỉ hướng

Ký hiệu: DIR

Mô tả: Vai này cho biết chuyển động diễn ra theo đường nào hay hướng nào. Dấu hiệu nhận biết:

- Tồn tại các giới từ như “từ”, “ra”, “vào”, “lên”, “xuống”, “đến”, “tới”, “sang”, “qua”, “lại”, “về” trước cụm từ đang xét

Ví dụ:

Anh ấy sẽ bay từ Sài Gòn ra Hà Nội .
 (S (NP-SUB (N-H Anh) (P ấy))
 (VP (R sẽ) (V-H bay)
 (PP-DIR (PP (E-H từ)
 (NP (Np-H Sài Gòn)))
 (PP (E-H ra)
 (NP (Np-H Hà Nội))))))
 (. .))

17.4 Phụ ngữ chỉ cách thức hay phương tiện

Ký hiệu: MNR

Mô tả: Vai này cho biết một hành động được thực hiện như thế nào. Dấu hiệu nhận biết:

- Đặt câu hỏi *thế nào*

- Tồn tại các giới từ như “bằng” (chỉ phương tiện) ở trước cụm từ đang xét

Ví dụ:

Cô gái ăn chè bằng thìa .
 (S (NP-SUB (N-H Cô) (N gái))
 (VP (V-H ăn)
 (NP-DOB (N-H chè))
 (PP-MNR (E-H bằng)
 (NP (N-H thìa))))

(. .))

ăn đứng ăn ngồi

(VP (VP (V-H ăn)
 (VP-MNR (V-H đứng)))
 (VP (V-H ăn)
 (VP-MNR (V-H ngồi))))

Nó vẽ quá đẹp.

(S (NP-SUB (P Nó))
 (VP (V-H vẽ)
 (AP-MNR (R quá) (A-H đẹp)))

(. .))

kêu ới á

17.5 Phụ ngữ chỉ mục đích hay lý do

Ký hiệu: PRP

Mô tả: Có một số dấu hiệu nhận biết sau

- Tồn tại các giới từ như “để”, “cho”, “mà” (mục đích), “vì”, “do” , “tại” , “bởi” (lý do, nguyên nhân) ở trước cụm từ hay mệnh đề đang xét
- Có thể chèn các giới từ đó vào trước mệnh đề đang xét
- Đặt câu hỏi *để làm gì* hoặc câu hỏi *tại sao* cho cụm chủ vị chính

Ví dụ:

Nó không đi làm được vì ốm .

(S (NP-SUB-1 Nó)
 (VP không
 đi
 (VP làm)
 được
 (SBAR-PRP vì
 (S (NP-SUB-1 *T*)
 (AP-PRD ốm))))

(. .))

// “đi làm” nên được ghép lại?

Con nên mượn sách của bạn mà học tiếng Anh.

(S (NP-SUB (N-H Con))
 (VP (V-H nên)
 (VP (V-H mượn)
 (NP-DOB (N-H sách)
 (PP (E-H của)
 (NP (N-H bạn))))
 (C mà)

(VP-PRP (V-H học)
(NP-DOB (N-H tiếng) (Np Anh))))))
(. .))

17.6 Phụ ngữ chỉ điều kiện

Ký hiệu: CND

Dấu hiệu nhận biết: Tồn tại các từ/cụm từ “nếu”, “giá mà”, “miễn là”, “hễ” đi trước cụm từ hay mệnh đề đang xét.

Nếu thời tiết đẹp thì lớp chúng tôi sẽ thăm rừng Cúc Phương vào chủ nhật này.

(S (SBAR-CND (C Nếu)
(S (NP-SUB (N-H thời tiết))
(AP-PRD (A-H đẹp))))
(C thì)
(S (NP-SUB (N-H lớp) (P chúng tôi))
(VP (R sẽ) (V-H thăm)
(NP-DOB (N-H rừng) (Np Cúc Phương))
(PP-TMP (E-H Vào)
(NP (N-H chủ nhật)
(NP-TMP (N-H tuần) (P này))))))
(. .))

Hễ ông Hòa đến thì anh gọi tôi đây.

(S (SBAR-CND (C Hễ)
(S (NP-SUB (Nc-H ông) (Np Hoà))
(VP (V-H đến))))
(C thì)
(S (NP-SUB (N-H anh))
(VP (V-H gọi)
(NP-DOB (P-H tôi))
(VP (V-H đây))))
(. .))

17.7 Trạng ngữ chỉ ý nhượng bộ

Ký hiệu: CNC³⁴

Dấu hiệu nhận biết: Tồn tại các từ/cụm từ “tuy”, “dù”, “mặc dù” đi trước cụm từ hay mệnh đề đang xét.

Tuy anh ấy bị hỏng mắt nhưng anh ấy vẫn sống rất lạc quan.³⁵

(S (SBAR-CNC (C Tuy)
(S (NP-SUB (N-H anh) (P ấy))
(VP (V bị)
(VP (V-H hỏng)
(NP-DOB (N mắt))))))
(C nhưng)
(S (NP-SUB (N-H anh) (P ấy))
(VP (R vẫn) (V-H sống)
(AP-MNR (R rất) (A-H lạc quan))))

³⁴ Viết tắt của concession

³⁵ Ví dụ này được lấy từ (LBiên, 1999; tr166)

(. .))

17.8 Trạng ngữ

Nhân chức năng: ADV

Mô tả³⁶: Nhãn này được sử dụng khi thành phần trạng ngữ không thuộc một trong các loại cụ thể đã được mô tả ở phần trên. Nếu một trạng ngữ được phân loại cụ thể hơn (chẳng hạn TMP) thì cũng được hiểu là có loại ADV.

Không có cách nào hơn, cô ấy lại phải tiếp tục đề nghị lên trên.

(S (VP-ADV (R không) (V-H có)
(NP (N-H cách) (P nào)
(AP (A-H hơn))))
(, ,)
(NP-SUB (N-H cô)(P ấy))
(VP (R lại) (V-H phải)
(VP (V-H tiếp tục)
(VP (V-H đề nghị)
(PP (E-H lên)
(NP (N-H trên))))))

(. .))

Y hện, hấn ta đến để giao hàng.

(S (VP-ADV (V-H Y) (N hện))
(, ,)
(NP-SUB (P-H hấn)(P ta))
(VP (V-H đến)
(PP-PRP (E-H để)
(VP (V-H giao)
(NP-DOB(N-H hàng))))

(. .))

Định thần sau sự cố, tôi bình tĩnh nhìn lại túi xách của mình.

(S (VP-ADV (V-H định thần)
(PP (E-H sau)
(NP (N-H sự cố))))
(, ,)
(NP-SUB (P-H tôi))
(VP (V-H bình tĩnh)
(VP (V-H nhìn) (R lại)
(NP-DOB (N-H túi xách)
(PP (E-H của)
(NP (P-H mình))))))

(. .))

³⁶ Tài liệu hướng dẫn gán nhãn cú pháp phần 1 đã phân biệt các loại trạng ngữ cụ thể.

18. Nhãn phần tử rỗng

18.1 Nhãn *T*

Mô tả: Nhãn phần tử rỗng *T* được dùng để thể hiện cấu trúc sâu của một số hiện tượng ngữ pháp³⁷ như bị động, khởi ngữ, mệnh đề phụ kết, cụm động từ làm bổ ngữ.

Ví dụ 1:

Thuyền được đẩy ra xa .

(S-PV (NP-SUB-1 (N-H Thuyền))

(VP (V-H được)

(SBAR (S (NP-SUB *E*)

(VP (V-H đẩy)

(NP-DOB-1 *T*)

(VP-MNR (V-H ra) (A xa))))))

(. .))

Đây là một câu bị động trong đó tân ngữ của động từ “đẩy” được đưa lên đầu làm chủ ngữ của câu.

Ví dụ 2:

Vấn đề này chúng tôi đang bàn .

(S-TC (NP-TPC-1 (N-H Vấn đề) (P này))

(S (NP-SUB (P-H chúng tôi))

(VP (R đang) (V-H bàn)

(NP-DOB-1 *T*)))

(. .))

Ví dụ 3:

Tôi đã mua quyển sách mà thầy giáo giới thiệu .

(S (NP-SUB-1 (P-H Tôi))

(VP (R đã) (V-H mua)

(NP-DOB (Nc-H quyển) (N sách)

(SBAR (C mà)

(S (NP-SUB (N-H thầy giáo))

(VP (V-H giới thiệu)

(NP-DOB-1 *T*))))))

(. .))

Câu này có hai mệnh đề, trong đó mệnh đề phụ kết bổ nghĩa cho từ “quyển sách”. Ở mệnh đề này, tuy tân ngữ không trực tiếp xuất hiện sau động từ “giới thiệu” nhưng ta ngầm hiểu đó là cụm từ “quyển sách”. Do đó ta cần đưa vào ký hiệu cụm danh từ rỗng giá trị chỉ số là 1, đồng chỉ số với cụm danh từ “quyển sách”³⁸. Một cụm danh từ rỗng vẫn được gán nhãn chức năng như bình thường, trong trường hợp này là DOB.

Ví dụ 4:

Anh ấy khỏe vì chơi tennis đều đặn .

³⁷ Nhãn *T* dùng cho A-movement

³⁸ Nếu cụm từ không có phần tử rỗng tương ứng thì không cần gán chỉ số.

(S (NP-SUB-1 (N-H Anh) (P ấy))
 (AP-PRD (A-H khỏe)
 (SBAR-PRP (C vì)
 (S (NP-SUB-1 *T*)
 (VP (V-H chơi)
 (NP-DOB (N-H tennis))
 (AP-MNR (A-H đều đặn))))))
 (. .))

Ở câu này thì phần từ rỗng lại là chủ ngữ của mệnh đề phụ kết bỏ nghĩa cho tính từ vị ngữ của mệnh đề chính.

18.2 Nhãn *0*

Mô tả: Nhãn này dùng để gán cho những mệnh đề phụ kết mà thiếu tác tử phụ ngữ hóa (complementizer). Dấu hiệu nhận biết:

- Có thể chèn “rằng”, “mà” vào trước mệnh đề đang xét (mệnh đề này có thể là bổ ngữ của động từ hoặc định ngữ của danh từ)

(S (NP-SUB (P-H Chúng))
 (VP (V-H dẫn)
 (AP-MNR (A-H thẳng))
 (PP-DIR (E-H lên)
 (NP (N-H núi)
 (. .)
 (NP-LOC (N-H nơi)
 (SBAR *0*
 (S (NP-SUB (N-H hàng) (A lậu))
 (VP (V-H được)
 (SBAR-DOB (S (NP-SUB *E*)
 (VP (V-H chuyển) (R đến))))))))))
 (. .))

18.3 Nhãn *RNR*

Mô tả:

Anh ấy vừa ăn vừa nói trong bữa tiệc .
 (S (NP-SUB Anh ấy)
 (VP (VP vừa ăn
 (PP-TMP *RNR*-1))
 (VP vừa nói
 (PP-TMP *RNR*-1))
 (PP-TMP-1 trong
 (NP bữa tiệc))))
 (. .))

19. Các cấu trúc sử dụng liên từ đẳng lập

Liên từ đẳng lập được dùng để nối (hay thể hiện quan hệ đẳng lập giữa) hai từ hay cụm từ.

19.1 Các từ trung tâm không có chung bổ ngữ

Nếu hai (hay nhiều) từ được nối với nhau bằng liên từ độc lập thì ta có thể chỉ cần gán như sau:

Bố, mẹ, và con

(NP (N-H BỐ) (, ,) (N-H mẹ) (, ,) (C và) (N-H con))

Chú ý là ở ví dụ này cụm danh từ có 3 danh từ trung tâm.

Trường hợp có ít nhất một thành phần là cụm từ thì ta gán cho chúng nhãn cụm từ:

Hai bút chì và một quyển sách

(NP (NP hai bút chì)

và

(NP một quyển sách))

Các ví dụ khác:

Cấu trúc cú pháp và ngữ nghĩa

(NP Cấu trúc

(NP cú pháp và ngữ nghĩa))

Đã, đang và sẽ thực hiện mua sách, giấy và bút

(VP (RP đã, đang và sẽ)

(VP thực hiện

(VP mua

(NP sách, giấy và bút))))

19.2 Các từ (hay cụm từ) có chung bổ ngữ

Họ thường chỉ nghĩ tới nghĩa vụ đóng góp của Việt kiều, chú trọng tới khía cạnh “khai thác”.

(S (NP-SUB (P-H Họ))

(VP (R thường) (R chỉ)

(VP (V-H nghĩ) (R tới)

(NP-DOB (N-H nghĩa vụ) (V đóng góp)

(PP (E-H của)

(NP (Np-H Việt kiều))))))

(VP (V-H chú trọng)

(PP (E-H tới)

(NP (N-H khía cạnh) (“”) (V khai thác) (“”))))))

(. .))

20. Kết từ đẳng lập (C) và kết từ chính phụ (E)

Qui ước:

- Dùng nhãn C với kết từ đẳng lập (và, hoặc, v.v.)
- Dùng nhãn E với kết từ chính phụ mà theo sau là mệnh đề (vì, do, v.v.), khi đó nút cha sẽ là SBAR
- Dùng nhãn E với kết từ chính phụ mà theo sau là cụm danh từ hoặc cụm động từ (của, cho, để, v.v.), khi đó nút cha sẽ là PP

Một tài liệu tham khảo tốt về các loại kết từ là (DQ Ban, 2007; Tập 1, tr132-).

Ví dụ:

Tôi yêu anh vì những nguyên nhân sâu xa hơn.

(S (NP-SUB (P-H Tôi))
(VP (V-H yêu)
(NP-DOB (N-H anh))
(PP-PRP (E-H vì)
(NP (L những) (N-H nguyên nhân)
(AP (A-H sâu xa) (R hơn))))))
(. .))

Người ta nhìn anh vì anh giống Rô-mê-ô của họ.

(S (NP-SUB (P-H Người ta))
(VP (V-H nhìn)
(NP (N-H anh))
(SBAR-PRP (C vì)
(S (NP-SUB (N-H anh))
(VP (V-H giống)
(NP-DOB (Np-H Rô-mê-ô)
(PP (E-H của)
(NP (P-H họ))))))))))
(. .))

Trong cả hai câu “vì” đều có từ loại E. Trong câu thứ nhất, “vì” là phần tử trung tâm của cụm giới từ PP. Trong câu thứ hai, “vì” lại là phần tử trung tâm của mệnh đề phụ SBAR.

21. Thành phần chú thích hoặc trích dẫn

21.1 Thành phần chú thích

Về hình thức, thành phần chú thích có thể được đặt trong cặp dấu gạch ngang, cặp dấu ngoặc, hay dấu phẩy. Về chức năng, phần chú thích có thể bổ nghĩa cho cụm từ hoặc cho cả câu. Ta xét các ví dụ sau:

Phần chú thích bổ nghĩa cho cụm danh từ:

Bà Hiền, giám đốc công ty X, cho biết ...
(S (NP (Np-H Bà Hiền)
(, ,)
(NP (N-H giám đốc)
(NP (N-H công ty) (Np X))))
(, ,)
(VP (V-H cho) (V biết)(... ...))
(. .))

Phần chú thích bổ nghĩa cho cả câu:

Em học sinh ấy – thật là gan dạ – đã nỗ lực diệt cả tổp địch.
(S (NP-SUB (N-H Em) (N học sinh) (P Ấy))

(- -)
 (AP (A-H thật)
 (C là)
 (AP (A-H gan dạ)))
 (- -)
 (VP (R đã) (V-H nỗ)
 (NP-DOB (N-H mình)))
 (VP-PRP (V-H diệt)
 (NP-DOB (P cả) (N-H tốp) (N địch))))
 (. .))

Nếu thành phần chủ thích được bao trong cặp dấu ngoặc thì dấu ngoặc trái ‘(‘ sẽ được chuyển thành LBKT còn dấu ngoặc phải ‘)’ thành RBKT. Chú ý là cặp dấu chủ thích nên được đặt ngang bậc với nhau trong cây cú pháp. Chẳng hạn như câu trên không nên được gán nhãn như sau :

(S (NP-SUB (N-H Em) (N học sinh) (P ấy))
 (- -)
 (AP (A-H thật)
 (C là)
 (AP (A-H gan dạ))
 (- -))
 (VP (R đã) (V-H nỗ)
 (NP-DOB (N-H mình)))
 (VP-PRP (V-H diệt)
 (NP-DOB (P cả) (N-H tốp) (N địch))))
 (. .))

Trong đó dấu gạch kết thúc chủ thích được đặt ở mức sâu hơn dấu gạch bắt đầu chủ thích.

21.2 Thành phần trích dẫn

Thành phần này thường là câu nói, được đặt trong cặp dấu nháy kép, đi sau các động từ như "nói", "cho biết", v.v.

Cô Lan nói thêm : "chúng tôi lên đây được 6 năm rồi".

(S (NP-SUB (Nc-H Cô) (Np Lan))
 (VP (V-H nói) (V thêm)
 (: :)
 (SBAR *0*
 (“ “)
 (S (NP-SUB (P-H chúng tôi))
 (VP (V-H lên)
 (NP-LOC (P-H đây))
 (VP-TMP (V-H được)
 (NP-TMP (M 6) (N-H năm))
 (T rồi))))
 (“ ”))
 (. .))

Chú ý cặp dấu nháy kép cũng được đặt ngang bậc nhau trong cây cú pháp.

22. Câu phức

Ngoài câu đơn, các nhà nghiên cứu còn phân câu trong tiếng Việt thành hai loại khác nữa là câu phức và câu ghép. Về thực chất, câu phức và câu ghép được cấu tạo từ các câu đơn. Cách thức tổ chức, sắp xếp và quan hệ giữa các câu đơn này làm thành những loại câu phức và câu ghép khác nhau, và dựa vào đó mà các nhà nghiên cứu chia các câu phức và câu ghép thành những loại khác nhau.

Câu phức và câu ghép đều được cấu tạo từ các câu đơn (từ hai câu đơn trở lên) nhưng cần phân biệt được sự khác nhau giữa hai loại câu này. Câu phức chỉ có một nòng cốt chủ ngữ - vị ngữ chính, còn trong nòng cốt chủ ngữ và vị ngữ ấy, chủ ngữ, vị ngữ (hoặc định ngữ, bổ ngữ) có thể là một hoặc nhiều câu đơn. Trong khi đó, một câu ghép có thể có hai hoặc nhiều nòng cốt chủ ngữ - vị ngữ *tồn tại ngang nhau*. Các nòng cốt chủ ngữ - vị ngữ này không cái nào bao chứa cái nào.

Trong cuốn *Ngữ pháp tiếng Việt*, 2005, tác giả Diệp Quang Ban đã đưa ra quan điểm của mình về câu ghép như sau: “Câu ghép là câu do hai (hoặc hơn hai) câu đơn kết hợp với nhau theo kiểu không câu nào bao chứa câu nào; mỗi câu đơn trong câu ghép tự nó thoả mãn định nghĩa về câu”.

Dựa trên những tổng kết về nòng cốt câu phức và câu ghép của các tác giả, dựa trên việc khảo sát, phân tích các ngữ liệu tiếng Việt, chúng tôi tạm thời chia câu phức và câu ghép ra một số loại như sau:

22.1 Câu phức chủ ngữ

Là câu có chủ ngữ là một nòng cốt chủ ngữ - vị ngữ.

Anh làm như vậy là không đúng.

(S (S-SUB(NP-SUB (N-H Anh))

(VP (V-H làm)

(C như)

(NP (P-H vậy))))

(C là)

(AP-PRD (R không) (A-H đúng))

(. .))

Cháu khỏi bệnh là nhờ các bác sĩ.

(S (S-SUB(NP-SUB (N-H Cháu))

(VP (V-H khỏi)

(NP (N bệnh))))

(C là)

(VP (V-H nhờ)

(NP-DOB (L các) (N-H bác sĩ))))

(. .))

Anh ấy về quê đã được năm ngày.

(S (S-SUB (NP-SUB (N-H Anh) (P ấy))
 (VP (V-H về) (N quê)))
 (VP (R đã) (V-H được)
 (NP (M năm) (N-H ngày)))
 (. .))

22.2 Câu phức vị ngữ

Là câu có vị ngữ là một nòng cốt chủ ngữ - vị ngữ.

Xe của tôi, máy vẫn chạy tốt.

(S (NP-SUB (N-H Xe)
 (PP (E-H của)
 (NP (P-H tôi))))
 (, ,)
 (S-PRD (NP-SUB (N-H máy))
 (VP (R vẫn) (V-H chạy) (A tốt)))
 (. .))

Cái áo ấy, giá là một trăm ngàn.

(S (NP-SUB (Nc-H Cái) (N áo)(P ấy))
 (, ,)
 (S-PRD (NP-SUB (N-H giá))
 (VP (V-H là)
 (NP (M một) (M trăm) (M ngàn))))
 (. .))

22.3 Câu phức bổ ngữ

Là câu có bổ ngữ của động từ làm vị ngữ là một nòng cốt chủ ngữ - vị ngữ.

Tôi thấy cô ấy đi với một người đàn ông lạ.

(S (NP-SUB Tôi))
 (VP-PRD thấy cô ấy đi với một người đàn ông lạ)
 (. .))

trong đó “cô ấy đi với một người đàn ông lạ” là một S, được phân tích thành:

(S (NP-SUB (N-H cô) (P ấy))
 (VP (V-H đi)
 (PP (E-H với)

(NP (M một) (Nc-H người) (N đàn ông) (A lạ))))

(. .))

Năm em học sinh được ban giám hiệu nhà trường tuyên dương.

(S (NP-SUB Năm em học sinh)

(VP-PRD được ban giám hiệu nhà trường tuyên dương)

(. .))

trong đó: “ban giám hiệu nhà trường tuyên dương” là một S, được phân tích thành:

(S (NP-SUB (N-H ban giám hiệu) (N nhà trường))

(VP (V-H tuyên dương))

(. .))

22.4 Câu phức định ngữ

Là câu có định ngữ của danh từ là một nòng cốt chủ ngữ - vị ngữ.

Quyển sách mà anh cho tôi mượn đã bị mất.

(S (NP-SUB Quyển sách mà anh cho tôi mượn)

(VP-PRD đã bị mất)

(. .))

trong đó: “anh cho tôi mượn” là một S, được phân tích thành:

(S (NP-SUB (N-H anh))

(VP (V-H cho)

(NP-DOB (P-H tôi))

(VP (V-H mượn)))

(. .))

Ngày anh phải đi công tác sắp đến rồi.

(S (NP-SUB Ngày anh phải đi công tác)

(VP-PRD sắp đến rồi)

(. .))

Trong đó, “anh phải đi công tác” là một S, được phân tích thành:

(S (NP-SUB (N-H anh))

(VP (V-H phải)

(VP (V-H đi)

(VP (V công tác))))

(. .))

Anh ấy đã mua quyển sách mà thầy giáo giới thiệu.

(S (NP-SUB (NP Anh ấy))

(VP-PRD (đã mua quyển sách mà thầy giáo giới thiệu))

(. .))

trong đó “... quyển sách mà thầy giáo giới thiệu”, được phân tích thành:

(NP-DOB (N-H quyển sách)

(SBAR (C mà)

(S (NP-SUB (N-H thầy giáo))

(VP (V-H giới thiệu))))))

23. Câu ghép

Trong các tài liệu về ngữ pháp tiếng Việt (UBKHXXH, 1983; DQBan, 2007) câu ghép được nhận biết và phân loại dựa vào các dấu hiệu: số thành phần chủ-vị, cách dùng phụ từ, và cách dùng kết từ. Thêm vào đó quan hệ ngữ nghĩa giữa các thành phần câu cũng là căn cứ phân loại câu ghép và phân biệt câu ghép với câu đơn.

22.1 Câu ghép song song³⁹ (UBKHXXH, 1983)

Dấu hiệu hình thức là câu ghép này có từ hai cụm chủ vị trở lên. Các cụm chủ vị này được nối với nhau bằng dấu phẩy hoặc kết từ đẳng lập.

Chim kêu, vượn hú.

(S (S (NP-SUB (N-H Chim))

(VP (V-H kêu)))

(, ,)

(S (NP-SUB (N-H vượn))

(VP (V-H hú)))

(. .))

Tôi chưa làm kịp, hay anh làm giúp tôi vậy?

(SQ (S (NP-SUB (P-H Tôi))

(VP (R chưa) (V-H làm) (A kịp)))

(, ,)

(C hay)

(S (NP-SUB (N-H anh))

(VP (V-H làm)

(VP-PRP (V-H giúp)

(NP-DOB (P-H tôi)))

(T vậy)))

(. ?))

22.2 Câu ghép qua lại (UBKHXXH, 1983)

Phụ từ được dùng trong vị ngữ của các vế: vừa...đã, chưa...đã, mới...đã, v.v.

³⁹ Còn gọi là câu ghép đẳng lập

Thầy giáo vừa dạy xong cậu đã quên rồi à.

(S (S (NP-SUB (N-H Thầy giáo))
(VP(R vừa) (V-H dạy) (R xong)))
(S (NP-SUB (N-H cậu))
(VP(R đã) (V-H quên) (T rồi) (T à)))
(. .))

Kết từ được dùng ở đầu mỗi vế: vì...nên, vì...mà, nếu...thì, v.v.

Vì trời mưa *nên* tôi không đi chơi nữa.

(S (C *Vì*
(S (NP-SUB (N-H trời))
(VP (V-H mưa)))
(C *nên*)
(S (NP-SUB (P-H tôi))
(VP (R không) (V-H đi)
(VP (V chơi))
(R nữa)))
(. .))

22.3 Phân biệt câu ghép với câu đơn có thành phần trạng ngữ

Sau khi tham khảo các tài liệu (NXBKHXH, 1983; NMThuyết và NVHiệp, 1999; DQBan, 2007) chúng tôi thấy rằng chưa có một cách phân biệt thống nhất giữa các tác giả. Tuy nhiên chúng tôi thấy rằng (NMThuyết và NVHiệp, 1999, tr314-316) có đề xuất một cách giải quyết vấn đề thiên về mặt cú pháp hình thức và khá rõ ràng, thuận tiện cho việc áp dụng. Do đó chúng tôi lựa chọn cách đó cho gán nhãn Treebank. Tiêu chuẩn như sau:

- Nếu liên từ thứ hai vắng mặt thì ta sẽ có một câu đơn có trạng ngữ chỉ nguyên nhân

(S (SBAR-PRP *Vì*
(S (NP trời)
(VP mưa)))
(S (NP tôi)
(VP không đi chơi nữa))
)

- Các trường hợp còn lại ta có câu ghép (3 trường hợp)

(S *Vì*
(S (NP trời)
(VP mưa))
nên
(S (NP tôi)
(VP không đi chơi nữa))
)

(S (S (NP trời)
(VP mưa))
nên
(S (NP tôi)
(VP không đi chơi nữa))
)

(S (S (NP trời)
 (VP mưa))
 (S (NP tôi)
 (VP không đi chơi nữa))
 .))

24. Tình lược

Nhãn: *E*

Mô tả: Trong nhiều trường hợp cần phải dùng nhãn phần tử tình lược để mô tả đầy đủ hơn cấu trúc ngữ pháp của một câu.

1. Tình lược chủ ngữ (TLCN)

Dạng tình lược này xuất hiện trong các trường hợp sau:

- TLCN trong câu mệnh lệnh

Ví dụ:

Ông chủ lại nói to:

A! Nó đói! Ø Đi bắt cho nó vài con châu chấu!

(S-CMD (NP-SUB *E*)
 (VP (V-H Đi)
 (VP (V-H bắt)
 (PP-IOB (E-H cho)
 (NP (P-H nó)))
 (NP-DOB (L vài) (Nc-H con) (N châu chấu))))
 (. .))

- Câu có dạng lời cầu chúc / cầu mong / lời chào...

Ví dụ:

Mời ông vào trong này. Chúng tôi đợi mãi. Ø Mời ông vào thưởng trồng.

(S (NP-SUB *E*)
 (VP (V-H Mời)
 (NP-DOB (N-H ông))
 (VP (V-H vào)
 (VP-PRP (V-H thưởng)
 (NP-DOB (N-H trồng))))
 (. .))

- Câu chứa các từ tình thái “cần, nên, phải, hãy”.

Ví dụ:

Nó chết phải đem chôn. Ø Phải mua cỗ gỗ. Ø Phải mời xóm, mời làng.

(S (NP-SUB *E*)
 (VP (V-H phải)
 (VP (V-H mua)
 (NP-DOB (N-H cỗ) (N gỗ))))
 (. .))

(*) kết ngôn: phát ngôn được liên kết với phát ngôn bị tình lược

Ø: kí hiệu cho yếu tố bị tình lược

- Câu có chủ ngữ mang ý nghĩa khái quát

Ví dụ:

Cụ Tiên chỉ làng Vũ Đại nhận ra rằng: Ø Hãy ngấm ngấm đẩy người ta xuống sông nhưng rồi dắt nó lên để đền ơn.

(Chí Phèo – Nam Cao)

Xin giải thích thêm đối với trường hợp này: chủ ngữ ở đây có thể tự do tùy vào hoàn cảnh sử dụng (tính lược trong trường hợp này thường gặp trong các câu thành ngữ, các câu nêu kinh nghiệm: Ăn quả nhớ kẻ trồng cây, uống nước nhớ nguồn...)

(S (NP-SUB (Nc-H Cụ) (N Tiên chỉ)
 (NP (N-H làng) (Np Vũ Đại)))
 (VP (V-H nhận) (R ra)
 (SBAR (C rằng)
 (: :)
 (S (NP-SUB *E*)
 (VP (VP (R Hãy) (A ngấm ngấm) (V-H đẩy)
 (NP-DOB (P-H người ta))
 (PP-DIR (E-H xuống)
 (NP (N-H sông))))
 (C rồi)
 (VP (V-H dắt)
 (NP-DOB-1 (N-H nó))
 (R lên)
 (SBAR-PRP (E-H để)
 (S (NP-1 *T*)
 (VP (V-H đền ơn))))))))))
 (. .))

- Chủ ngữ trong câu bị tính lược đồng chức năng với chủ ngữ ở các kết ngôn (*)
 Ví dụ này cho thấy yếu tố đóng vai trò là chủ ngữ trong phát ngôn bị tính lược “anh” cũng đồng thời là chủ ngữ bị tính lược trong các phát ngôn đăng sau.

Ví dụ:

Anh cứ hát. Gò ngực mà hát.

(S (NP-SUB *E*)
 (VP (V-H gò) (N ngực)
 (C mà)
 (VP-PRP (V-H hát)))
 (. .))

- Chủ ngữ trong câu bị tính lược không cùng chức năng trong các kết ngôn

Ví dụ:

(1) Sau cùng thì y gạ THAI cổ vườn. (2)Ø Không muốn cổ thì Ø xoay tiền mà trả y.

(S (S (NP-SUB *E*)
 (VP (R không) (V-H muốn)
 (VP (V cổ))))
 (C thì)

(S (NP-SUB *E*)
 (VP (V-H xoay)
 (NP-DOB (N-H tiền))
 (C mà)
 (VP-PRP (V-H trả)
 (NP-DOB (P y))))))
 (. .))

Trong ví dụ trên, “Thai” là bổ ngữ trong phát ngôn (1) nhưng sang đến (2) có chức năng là chủ ngữ của các phát ngôn đó.

- Chủ ngữ là yếu tố được ngầm định
- Chủ ngữ là lời của chính tác giả

Ví dụ:

Cái đầu ông ngoẹo xuống, như đầu một thằng bé khi nó cúi. Trông thật là thiếu não.

(S (NP-SUB *E*)
 (VP (V-H Trông)
 (AP (T thật) (C là) (A-H thiếu não))))
 (. .))

Chủ ngữ là nhân vật được nói đến trong truyện

Ví dụ:

Nguyên là bà ấy béo quá – gớm! Ø Béo đến nỗi bụng sẽ xuống.

(S (NP-SUB *E*)
 (AP-PRD (A-H Béo)
 (XP (X-H đến nỗi)
 (NP (N-H Bụng)
 (VP (V-H sẽ) (R xuống))))))
 (. .))

- Chủ ngữ là một trong những người đang đối thoại

Ví dụ:

Rồi đổi giọng, cụ thân mật hỏi:

Ø Về bao giờ thế?

(S (NP-SUB *E*)
 (VP (V-H Về)
 (WHADV-TMP (P-H bao giờ))))
 (T thế)
 (. ?))

- Chủ ngữ ẩn

Ví dụ:

Ông Bình cho biết sắp tới sẽ đề nghị trung ương phân cấp cho TP, phân cấp cho tổng công ty để đẩy nhanh hơn tốc độ cổ phần hoá.

(S (NP-SUB (Nc-H Ông) (Np Bình))
 (VP (V-H cho) (V biết)
 (SBAR-DOB (S (VP-TMP (R sắp) (V-H tới))
 (NP-SUB *E*))

(VP (R sẽ) (V-H đề nghị)
 (NP-DOB (N-H trung ương))
 (VP (VP (V-H phân cấp)
 (PP (E-H cho)
 (NP (Ny-H TP))))
 (, .))
 (VP (V-H phân cấp)
 (PP (E-H cho)
 (NP (N-H tổng công ty))))
 (PP-PRP (E-H để)
 (VP (V-H đẩy)
 (AP (A nhanh) (R hơn))
 (NP-DOB (N-H tốc độ) (V cổ phần hóa))))))
 (, .))

Câu tỉnh lược trên đã lược bỏ thành phần chủ ngữ trong mệnh đề thứ nhất. Mà khó có thể khôi phục chính xác thành phần chủ ngữ. Do đó gán nhãn *E* nhằm ngầm thông báo rằng đã có một thành phần bị tỉnh lược.

- Chủ ngữ bị tỉnh lược khi câu có mô hình : Là + danh từ
- Mô hình hiển ngôn từ “là”

Cấu trúc “Danh là Danh” biểu thị quan hệ đồng nhất giữa các sự vật, hiện tượng.

Ví dụ:

Ông là người cha nghiêm khắc của lũ con ích kỉ, đần độn. Ø Là người chồng đáng kính của các bà vợ tâm thường...

(S (NP-SUB*E*)
 (VP (V-H Là)
 (NP (Nc-H người) (N chồng)
 (VP (V-H đáng) (V kính))
 (PP (E-H của)
 (NP (L các) (N-H bà vợ) (A tâm thường))))
 (, .))

- Mô hình ẩn từ “là”

Ví dụ:

Mà bọn cô đầu thì ác quá. Họ cậy có quần áo đẹp, tóc uốn quăn. Họ cứ nhìn cái đầu thợ nhà quê xén vụng của tôi mà cười. Huống chi lại còn có bao nhiêu khách của ô. H nữa. Ø Toàn những phú thương cả.

Nếu khôi phục câu này sẽ là: Họ toàn là những phú thương cả.

(S (NP-SUB *E*)
 (NP-PRD (R toàn) (L những) (N-H phú thương) (T cả)
 (, .))

2. Tỉnh lược vị ngữ

Khi tỉnh lược vị ngữ thì nòng cốt còn lại trong câu là chủ ngữ. Ngoài ra còn có thể có những thành phần phụ khác nếu chúng thuộc phần báo. Khi đó, giữa chủ ngữ với các thành phần phụ này thường có dấu phẩy (hoặc dấu ngang nói...) ngăn cách – đây là một hình thức đánh dấu vị trí tỉnh lược.

Ví dụ:

Văn Viện kiểm sát: Bộ có chủ trương cho thứ trưởng ký xác nhận vào công văn vào công văn Công ty tiếp thị không?

(S (NP-SUB (R Vẫn) (Np Viện Kiểm sát))
 (VP (V-H *E*))
 (: :)
 (“ “)
 (SBAR (SQ (NP-SUB Bộ)
 (VP (V-H có)
 (VP (V-H chủ trương)
 (VP (V-H cho)
 (NP-DOB (N-H thứ trưởng))
 (VP (V-H ký)
 (NP-DOB (N-H xác nhận))
 (PP-LOC (E-H vào)
 (NP (N-H công văn)
 (NP (N-H Công ty) (V tiếp thị)))))))))
 (R không))
 (? ?)))
 (" "))
 (. .))

Nếu các thành phần phụ cũng thuộc phần nêu thì chúng có thể tỉnh lược cùng với vị ngữ. Khi đó chủ ngữ sẽ là thành phần duy nhất còn lại trong câu bị tỉnh lược.

Ví dụ:

Lan vừa bước vào nhà. Ø Cả bố và mẹ.
 (S (NP-SUB (P cả) (N-H bố) (C và) (N-H mẹ))
 (VP *E*))
 (. .))

3. Tỉnh lược C-V

Ví dụ:

Thoáng chốc, Quyên nhớ đến mọi nét mọi vẻ của Cà My. Ø1 Ø2 Cả cái cử chỉ khi Cà My ôm cô mà hôn thiết là kêu.

Ø= Quyên nhớ đến...

(S (NP-SUB *E*))
 (VP (V *E*))
 (NP (P cả) (Nc-H cái) (N cử chỉ)
 (NP-TMP (N-H khi)
 (SBAR *Ø*
 (S (NP-SUB (Np-H Cà My))
 (VP (V-H ôm)
 (NP-DOB (N-H cô))
 (C mà)
 (VP-PRP (V-H hôn)
 (NP-DOB (N-H cô))
 (AP-MNR (A-H thiết)
 (C là)
 (VP (V-H kêu)))))))))))))
 (. .))

Trong tiếng Việt, một cấu trúc khá quan trọng là C-V-B. Nên khi lược bỏ C-V thì vẫn còn một thành phần khá quan trọng là bổ ngữ. Điều này thể hiện rõ trong các câu tồn tại. Ví dụ:

Nhìn lại đằng sau, Dũng có cả một khu gang thép. Và Ø1 Ø2 một gia đình sau bao nhiêu năm tan tác đã dần dần đoàn tụ.

Ø= Dũng có cả...
(S (C Và)
(NP-SUB *E*)
(VP (V *E*)
(NP-DOB (M một) (N-H gia đình)
(PP-TMP (E-H sau)
(NP (P bao nhiêu) (N-H năm) (A tan tác)))
(VP (R đã) (R dần dần) (V-H đoàn tụ)))
(. .))

Đây là những thành phần dễ bị tỉnh lược nhất. Còn các thành phần khác như: V-B, C-V-B thương ít gặp. Khi khôi phục sẽ khó khăn và mang tính chủ quan. Khi làm mọi người nên gán nhãn cho các thành phần đã được mô tả trên đây.

25. Câu bị động

I. Cấu trúc chung của câu bị động tiếng Việt

Tiếng Việt không biến hình từ, nên động từ không có dạng chủ động và bị động. Tuy nhiên tiếng Việt cũng có cách diễn đạt ý bị động một cách đều đặn như các quy tắc ngữ pháp, bằng hai phương thức ngữ pháp hư từ và trật tự từ. Với hai phương thức này dạng bị động của câu tiếng Việt được xác định bằng một số yếu tố hữu hạn có quan hệ cấu trúc khá chặt chẽ làm thành ba điều kiện cần và đủ cho việc tạo nên một kiến trúc bị động ổn định.

- Chủ ngữ bị động, về mặt nghĩa chịu ảnh hưởng của động từ ngoại động trong câu bị bao (điều kiện cần để phân biệt nó với chủ ngữ chủ động).
- Có mặt trợ động từ bị động “bị” hay “được” (điều kiện cần để phân biệt câu bị động với câu trung tính).
- Vị tổ là một câu bị bao (trong đó chủ ngữ chủ động có thể vắng mặt, vị tổ là động từ ngoại động, thực thể nêu ở chủ ngữ chủ động của câu bị bao không trùng với thực thể nêu ở chủ ngữ bị động của câu. Đây là điều kiện cần để phân biệt bị, được là trợ động từ bị động với bị, được là động từ tình thái).

Do đó ta có mô hình cấu trúc cú pháp chung của câu bị động tiếng Việt

CN1 + TĐT bị động + Vị tổ 1(câu bị bao)
(bị động) (bị, được) (NC2 + Vị tổ 2(ĐT ngoại động) + Bỏ ngữ)

Ví dụ: Thuyền được (người lái) đẩy ra xa.

Em bé được mẹ rửa chân cho.

II. Gán nhãn cú pháp cho câu bị động tiếng Việt

Việc gán nhãn cho câu bị động tiếng Việt dựa trên vai trò của từng thành phần trong câu. Nó cũng giống như việc gán nhãn cho các câu chủ động trong tiếng Việt, cụ thể như sau:

- Nhãn chức năng bị động cho câu là: PV (passive voice)
- CN1 (bị động): đóng vai trò là chủ ngữ của câu, gán nhãn SUB
- Trợ động từ bị động “bị” , “được”: đóng vai trò là động từ trong câu, nên gán nhãn V cho “bị” và “được”
- Vị tổ 1(câu bị bao): là một mệnh đề đóng vai trò bỏ ngữ của câu, nên gán nhãn là SBAR

Ví dụ 1: Thư được Giáp gửi cho Tị.

(S-PV (NP-SUB-1(N-H Thư))
(VP (V-H được)
(SBAR(S(NP-SUB(Np-H Giáp))

(VP(V-H gửi)
 (NP-DOB-1 *T*)
 (PP-IOB (E-H cho)
 (NP(Np-H Tị))))))

(. .))

Ví dụ 2: Nó bị cảnh sát bắt.

(S-PV (NP-SUB-1(P-H Nó))
 (VP (V-H bị)
 (SBAR (S (NP-SUB (N-H cảnh sát))
 (VP (V-H bắt)
 (NP-DOB-1 *T*)))))

(. .))

Ví dụ 3: Mái nhà bị gió lật.

(S-PV (NP-SUB(N-H Mái nhà))
 (VP (V-H bị)
 (SBAR (S (NP-SUB(N-H gió))
 (VP (V-H lật)
 (NP-DOB-1 *T*)))))

(. .))

III. Phân biệt trợ động từ bị động (bị, được) với động từ thực và động từ tình thái
 Trong tiếng Việt thường có sự nhập nhằng và nhầm lẫn giữa trợ động từ bị động "bị", "được" với động từ thực "bị", "được" và động từ tình thái "bị", "được". Vì vậy để gán nhãn chính xác điều đầu tiên là cần xác định chính xác đó có phải là câu bị động hay không? Sau đây là những điều kiện dùng của động từ thực "bị", "được" và của động từ tình thái "bị", "được", và trợ động từ bị động "bị", "được" thể hiện ở chu cảnh cú pháp và chức năng cú pháp riêng biệt trong cách dùng hai từ này.

1. Chức năng và chu cảnh cú pháp của trợ động từ bị động “bị”, “được”
 - Làm tác tố bị động, không tham gia vào vị tố
 - Đứng trước vị tố là câu bị bao (câu này có thể vắng chủ ngữ), vị tố của câu bị bao là động từ ngoại động tác động lên thực thể nêu ở chủ ngữ của toàn câu
 - Chủ ngữ của câu bị bao và của “bị”, “được” không trùng nhau.

Ví dụ 1: Họ bị kẻ gian lấy mất tiền.

(S-PV (NP-SUB(P-H Họ))
 (VP(V-H bị)
 (SBAR(S(NP-SUB(N-H kẻ gian))
 (VP(V-H lấy)(V mất)
 (NP-DOB(N-H tiền))))))

(. .))

Ví dụ 2: Tường được treo tranh.

(S-PV (NP-SUB(N-H Tường))
 (VP(V-H được)
 (SBAR(S(NP-SUB *E*)
 (VP(V-H treo)
 (NP-DOB(N-H tranh))))))

(. .))

2. Chức năng và chu cảnh cú pháp của động từ thực “bị”, “được”
 - Với tư cách một thực từ, tức là từ mang ý nghĩa từ vựng đầy đủ các từ “bị”, “được” dễ dàng làm vị ngữ và có chu cảnh cú pháp sau đây:
 - + Bỏ ngữ là một danh từ hay cụm danh từ

Ví dụ 1 :

Con thả bị đạn .

(S (NP-SUB (Nc-H Con) (N thả))
(VP (V-H bị)
(NP-DOB (N-H đạn)))
(. .))

Ví dụ 2:

Cậu bé được cái bút rất đẹp.

(S (NP-SUB (N-H Cậu bé))
(VP (V-H được)
(NP-DOB (N-H cái bút)
(AP (R rất) (A-H đẹp))))
(. .))

+ Bỏ ngữ là một câu bị bao với hai điều kiện

- Chủ ngữ 1(của toàn câu) không chịu tác động của vị tổ 2 trong câu bị bao. Vị tổ 2 của câu bị bao có thể là động từ nội động hay ngoại động.
- Thực thể của chủ ngữ 2(của câu bị bao) không trùng với thực thể ở chủ ngữ 1 (chủ ngữ của toàn câu)

Ví dụ:

Em này bị bố mẹ mất sớm.

(S (NP-SUB(N-H Em)(P này))
(VP (V-H bị)
(SBAR (S (NP-SUB (N-H bố mẹ))
(VP (V-H mất)
(AP(A-H sớm))))))
(. .))

Kết luận: Chức năng và chu cảnh cú pháp của 2 động từ thực “bị”, “được”:

- Làm vị ngữ; đứng trước bỏ ngữ do danh từ hoặc cụm danh từ đảm nhiệm.
 - Làm vị tổ; đứng trước bỏ ngữ do một câu đảm nhiệm với điều kiện:
- + Chủ ngữ của toàn câu không chịu tác động của vị tổ trong câu bị bao.
- + Thực thể ở chủ ngữ của toàn câu không trùng với thực thể ở chủ ngữ của câu bị bao.
3. Chức năng và chu cảnh cú pháp của động từ tình thái “bị”, “được”
- Các động từ tình thái đích thực có nét chung là chúng đứng trước một động từ khác có chủ ngữ là một thực thể trùng với thực thể ở chủ ngữ của câu.
 - Chức năng của 2 động từ tình thái “bị”, “được” trong câu diễn đạt tính tình thái được phép, bắt buộc, nằm trong phân tình thái của câu, chứ không giữ vai trò vị tổ diễn đạt sự thể như động từ thực từ. Vì vậy mặc dù động từ tình thái khi gán nhãn treebank được gán là V nhưng không phải là động từ chính của vị tổ trong câu.
 - Mặt khác hai động từ tình thái này không thực hiện chức năng biến câu thành câu bị động như trợ động từ bị động.

Ví dụ 1 :

Nó được đi xem kịch.

(S (NP-SUB (N-H Nó))
(VP (V được)
(VP (V-H đi)
(VP (V-H xem)
(NP-DOB (N-H kịch))))
(. .))

Ví dụ 2:

Bạn ấy bị ốm.

(S (NP-SUB (N-H Bạn) (P ấy))

(VP (V-H bị)

(VP (V-H ốm))))

(. .))

Kết luận: Chức năng và chu cảnh cú pháp của 2 động từ tình thái “bị”, “được”.

- Làm yếu tố tình thái, không tham gia vào vị tổ.

- Đứng trước vị tổ là động từ nội động, ngoại động, tính từ hay quan hệ từ; các từ này có chủ ngữ trùng với chủ ngữ của bị, được

IV. Phân tích một vài cách dùng bị, được có thể gây lẫn lộn

Ví dụ 1:

(A) Cầu thủ X bị phạm lỗi.(câu này dùng trong thuyết minh bóng đá hiện nay có quan hệ nghĩa với câu sau đây: Cầu thủ Y phạm lỗi đối với cầu thủ X.)

(B) Em này bị phạm lỗi chính tả trong bài viết.

Trong hai câu (A) và (B) thì câu (A) là câu bị động vì chủ ngữ của toàn câu khác với chủ ngữ của “phạm lỗi”. Nếu diễn đạt khác đi sẽ là “ Cầu thủ X bị cầu thủ Y phạm lỗi”. Về mặt nghĩa cầu thủ X là người bị hại. Ở câu (B) chủ ngữ của “bị” và “phạm lỗi” đều là “em này”. “Bị” là động từ tình thái, nó không giữ chức năng vị tổ của câu, có thể bỏ “bị” mà không làm thay đổi nghĩa sự việc của câu .

- Gán nhãn cho hai câu trên sẽ là

(A) Cầu thủ X bị phạm lỗi.

(S-PV (NP-SUB (N-H Cầu thủ X))

(VP (V-H bị)

(SBAR (S (NP-SUB *E*))

(VP (V-H phạm)

(NP-DOB (N-H lỗi))))))

(. .))

(B) Em này bị phạm lỗi chính tả trong bài viết.

(S (NP-SUB (N-H Em này))

(VP (V-H bị)

(VP (V-H phạm)

(NP-DOB (N-H lỗi) (N chính tả)

(PP (E-H trong)

(NP (N-H bài viết))))))

(. .))

Ví dụ 2 :

(C) Các nhà báo được chất vấn.

Được đặt trong mối quan hệ nghĩa với câu

(D) Ông cố vấn bị chất vấn.

Trong hai câu trên thì câu (C) không phải là câu bị động. “Bị” là động từ tình thái, nó không tham gia vào vị tổ của câu. chủ ngữ của toàn câu và của “chất vấn” là một “các nhà báo”.

Còn câu (D) là câu bị động, chủ ngữ của câu là “ông cố vấn”, còn chủ ngữ của “chất vấn” là “các nhà báo”.

- Gán nhãn cho hai câu trên sẽ là
- (C) Các nhà báo được chất vấn.
 (S (NP-SUB (L Các) (N-H nhà báo))
 (VP (V được)
 (VP (V-H chất vấn))))
 (. .))
- (D) Ông cố vấn bị chất vấn.
 (S-PV (NP-SUB-1 (N-H Ông) (N cố vấn))
 (VP (V-H bị)
 (SBAR (S (NP-SUB *E*)
 (VP (V-H chất vấn)
 (NP-DOB-1 *T*)))))
 (. .))

Tài liệu tham khảo:

Tiếng Việt

- [1] Diệp Quang Ban. Ngữ pháp tiếng Việt. 2005. *NXB Giáo dục*.
- [2] Lê Biên. Từ loại tiếng Việt hiện đại. 1999. *NXB Giáo dục*.
- [3] Vũ Tiên Dũng. Tiếng Việt và ngôn ngữ học hiện đại sơ khảo về cú pháp. 2003. *VIET Stuttgart – Germany*.
- [4] Cao Xuân Hạo. Tiếng Việt sơ thảo ngữ pháp chức năng. 2006. *NXB Khoa học xã hội*.
- [5] Cao Xuân Hạo. Tiếng Việt mấy vấn đề ngữ âm ngữ pháp ngữ nghĩa. 2007. *NXB Giáo dục*.
- [6] Nguyễn Văn Hiệp. *Vài nét về lịch sử nghiên cứu cú pháp tiếng Việt. Tạp chí Ngôn ngữ, Hà Nội, số 10/2002*.
- [7] Nguyễn Văn Hiệp. Cơ sở ngữ nghĩa phân tích cú pháp. 2008. *NXB Giáo dục*.
- [8] Đào Minh Thu, Nguyễn Phương Thái. Thủ thuật phân tích câu và cụm từ. 2008. *Tài liệu nội bộ nhóm VTB*.
- [9] Nguyễn Minh Thuyết, Nguyễn Văn Hiệp. Thành phần câu tiếng Việt. 1999. *NXB Đại học Quốc gia Hà Nội*.

Tiếng Anh

- [1] Peter Sells. Lectures on Contemporary Syntactic Theories. 1987. CSLI.
- [2] Mitchell P. Marcus et al. Building a Large Annotated Corpus of English: The Penn Treebank. 1993. Computational Linguistics.
- [3] Fei Xia et al. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. 2000. COLING.
- [4] Nianwen Xue et al. Building a Large-Scale Annotated Chinese Corpus. 2002. COLING.
- [5] Chung-hye Han et al. Development and Evaluation of a Korean Treebank and its Application to NLP. 2002. LREC.
- [6] Sabine Brants et al. The TIGER Treebank. 2003. COLING.