ĐẠI HỌC HUẾ
TRƯỜNG ĐẠI HỌC KHOA HỌC
KHOA CÔNG NGHỆ THÔNG TIN

KHOÁ LUẬN
TỐT NGHIỆP ĐẠI HỌC

CHUYÊN NGÀNH CÔNG NGHỆ THÔNG TIN

Đề tài:

BẮT LỖI CHÍNH TẢ TIẾNG VIỆT TRÊN MÁY TÍNH

Sinh viên thực hiện: **Lê Viết Mẫn**
Lớp: **Tin K24B - Hệ Chính Quy**
Giáo viên hướng dẫn: **TS. Nguyễn Gia Định**

Huế 05-2004

MỤC LỤC

MỞ ĐẦU

Ngày nay, ở nước ta, việc soạn thảo văn bản bằng máy tính đã trở nên phổ biến. Tính thẩm mỹ và tốc độ cao đã quyết định nên điều đó. Nhưng các lỗi chính tả vẫn thường xuyên xuất hiện trong các văn bản. Điều đó làm cho các văn bản mất giá trị và cũng như sẽ gây ra những hiểu lầm rất nguy hiểm. Các lỗi có thể là do gõ nhầm hay do phát âm sai (một lỗi truyền thống hay gặp). Do đó, việc bắt lỗi chính tả tiếng Việt trên máy tính trở nên rất cấp thiết.

Bắt lỗi chính tả tiếng Việt cũng như xử lý ngôn ngữ tự nhiên lâu nay đã được quan tâm nghiên cứu ở nước ta. Nhiều sản phẩm bắt lỗi chính tả đã ra đời như: Vietkey Office, VietSpell… Nhưng vẫn còn nhiều lỗi bị bỏ qua trong quá trình phân tích mà đặc biệt là lỗi do cách phát âm. Những phần mềm này chỉ mới đáp ứng được phần nào sự mong đợi của người sử dụng.

Hiện nay, một hướng mới để bắt lỗi chính tả tiếng Việt đã ra đời và đang được nghiên cứu hoàn thiện. Đó là, bắt lỗi chính tả dựa trên mức cú pháp hay có thể nói là dựa trên cấu trúc cú pháp của các thành phần trong câu tiếng Việt để bắt các lỗi chính tả. Đây là một hướng giải quyết vấn đề có tính đột phá, một cơ hội để bắt được tất cả các lỗi chính tả đã mở ra. Dựa trên hướng giải quyết mới này, TS Phan Thị Tươi đã nghiên cứu và hoàn thành đề tài "Bắt lỗi chính tả tự động cho tiếng Việt bằng máy tính". Theo [8, 9], đề tài đã bắt được 95% các loại lỗi chính tả trong văn bản nhà nước phạm vi hành chính, xã hội. Để đạt được điều đó, phần mềm có 56.300 từ và khoảng 96 từ loại với 31.030 luật sinh hỗ trợ. Điều này cho thấy rõ khả năng bắt lỗi của phương pháp mới là cao hơn rất nhiều. Tuy nhiên, phương pháp này có những khó khăn như: vấn đề nhập nhằng ngữ nghĩa trong phân tích cú pháp và hiện tượng bùng nổ tổ hợp. Để giải quyết triệt để vấn đề lỗi chính tả, ta phải vượt qua những khó khăn trên.

Dựa trên phương pháp bắt lỗi chính tả tiếng Việt mới, lấy đề tài của TS Phan Thị Tươi làm cơ sở, khoá luận thực hiện nghiên cứu và phát triển phương pháp bắt lỗi chính tả tiếng Việt. Khoá luận sẽ đề cập đến một số vấn đề, như: từ điển tiếng Việt, quá trình thực thi bắt lỗi, phương pháp gợi ý, giải thuật Earley cho phân tích cú pháp, ngữ pháp tiếng Việt,… Đồng thời, khoá luận còn thực hiện cài đặt một số chương trình thử nghiệm, một số chương trình hỗ trợ quá trình nghiên cứu.

Khoá luận được trình bày trong 52 trang, ngoài phần mở đầu, kết luận và tài liệu tham khảo, khoá luận được chia thành ba chương:

Chương 1: **Bắt lỗi chính tả tiếng Việt**

- Trình bày tổng quan về các loại lỗi chính tả và phương pháp bắt lỗi chính tả.
- Trình bày những phân tích và thiết kế từ điển.
- Trình bày hai phương pháp bắt lỗi chính tả dựa trên âm tiết là: phương pháp dựa trên nguyên lý cấu tạo âm tiết và phương pháp dựa trên từ điển. Đồng thời còn trình bày phương pháp tạo gợi ý sửa lỗi ở mức âm tiết.
- Trình bày giải thuật Earley và những cải tiến
- Trình bày phương pháp bắt lỗi mức cú pháp: phân tích từ, phương pháp bắt lỗi chính tả dựa trên mức cú pháp, phương pháp gợi ý sửa lỗi.

Chương 2: **Ngữ pháp tiếng Việt với phân tích cú pháp**

- Trình bày khái quát về ngữ pháp tiếng Việt, bao gồm: từ loại, cụm từ và câu. Đồng thời, chỉ ra cách mã hoá từ loại và lập luật sinh cho chương trình bắt lỗi.

Chương 3: **Cài đặt chương trình**

- Trình bày hệ thống chương trình bắt lỗi chính tả và minh hoạ chương trình bắt lỗi.

Dù đã có nhiều cố gắng trong quá trình làm khoá luận, nhưng do hạn chế về mặt thời gian cũng như khả năng của bản thân nên không thể tránh khỏi những sơ suất. Mong nhận được sự chỉ giáo của quý thầy cô và bạn bè gần xa.

Nhân đây, em xin chân thành cám ơn thầy Nguyễn Gia Định đã tận tình giúp đỡ và hướng dẫn em trong suốt thời gian làm khoá luận. Thầy đã chỉ bảo và truyền đạt cho em nhiều kinh nghiệm quý báu.

Em cũng xin tỏ lòng biết ơn đến quí thầy cô khoa Công Nghệ Thông Tin - Trường Đại Học Khoa Học - Huế đã giảng dạy và truyền đạt những kiến thức cần thiết và bổ ích trong suốt thời gian học tập tại trường. Đồng thời, cũng xin cám ơn tất cả các bạn bè đã động viên, giúp đỡ tôi hoàn thành khoá luận này.

Gia đình và người thân luôn là nguồn động viên, cổ vũ hết sức to lớn, con xin ghi nhớ công ơn này.

Huế, tháng 5 năm 2004
Sinh viên thực hiện khoá luận
Lê Viết Mẫn

Chương 1

BẮT LỖI CHÍNH TẢ TIẾNG VIỆT

1.1. TỔNG QUAN BẮT LỖI CHÍNH TẢ TIẾNG VIỆT:

1.1.1. Lỗi chính tả:

Trong khi soạn thảo văn bản, người soạn thảo vẫn hay mắc các lỗi chính tả có thể là do gõ nhầm hay do cách phát âm. Từ những nguyên nhân đó, các lỗi chính tả được tạo ra là rất đa dạng. Nhưng nói chung (cũng theo những nghiên cứu gần đây), các lỗi chính tả có thể được xếp vào hai loại sau:

+ Từ bị sai không nằm trong từ điển – chúng tôi sẽ gọi là lỗi âm tiết:

Ví dụ:

- xổng → xỗng (do sai hỏi ngã)
- quỷ → qủy (sai do bỏ dấu)
- phát → pohát (gõ dính phím)
- đọc → đpjc (gõ nhầm ký tự)
- thăng → thưng (sai do gõ thiếu ký tự)
- cũng → ccũng (sai do gõ thừa ký tự)
- thật → thậc (sai do cách phát âm)
- tuyệt → tyuệt (sai do gõ nhầm vị trí ký tự)

+ Từ bị sai nhưng vẫn có trong từ điển – chúng tôi sẽ gọi là lỗi cú pháp:

Ví dụ:

- trái chanh → trái tranh (sai do cách phát âm)
- phiến diện → phiếu diện (sai do gõ nhầm)
- canh cửa → cánh cửa (sai do bỏ dấu)
- phát hoang → phác hoang (sai do phát âm)
- đầu đuôi → đầu đôi (sai do thiếu ký tự)
- thế thần → thết thần (sai do thừa ký tự)

1.1.2. Tổng quan phương pháp bắt lỗi chính tả tiếng Việt:

Từ hai loại lỗi đã chỉ ra ở trên, ta thấy rằng việc bắt lỗi chính tả chủ yếu như sau:

- Bắt lỗi các từ không có trong từ điển. (chúng tôi sẽ gọi giai đoạn này là bắt lỗi mức âm tiết.)
- Bắt lỗi các từ có trong từ điển nhưng vẫn sai. (chúng tôi sẽ gọi giai đoạn này là bắt lỗi mức cú pháp.)

Phương pháp bắt lỗi chính tả sẽ bao gồm hai giai đoạn như sau:

- Giai đoạn 1: Kiểm tra toàn bộ các âm tiết trong văn bản xem có hợp lệ không. Nếu đã hợp lệ thì sang bước hai, còn chưa thì gợi ý sửa lỗi và thay thế nếu muốn.

Để kiểm tra một âm tiết có hợp lệ hay không, ta có thể sử dụng một trong hai cách: hoặc là dựa trên nguyên tắc cấu tạo từ, hoặc là dựa trên từ điển. Phương pháp thứ hai dẫn đến việc phải lưu trữ nhiều từ không có nghĩa. Cả hai phương pháp này sẽ được bàn đến ở mục 1.4.

- Giai đoạn 2: Kiểm tra xem kết cấu câu có hợp lệ không; tức là, kiểm tra cú pháp câu có đúng với luật sinh quy định không. Ở giai đoạn này, câu được phân tích thành các chuỗi từ loại. Các chuỗi từ loại này được đưa qua bộ phân tích cú pháp. Bộ phân tích sẽ trả lại chuỗi từ loại đúng (tức là câu đúng) hoặc đưa ra vị trí từ loại nghi sai. Dựa trên vị trí từ loại nghi sai, ta có thể thực hiện gợi ý và sửa lỗi.

Trong giai đoạn này, bộ phân tích phải có giải thuật thích hợp cho phân tích cú pháp ngôn ngữ tự nhiên. Hiện nay, có bốn giải thuật phân tích cú pháp phù hợp là: thuật toán phân tích Top-Down, thuật toán phân tích Bottom-Up, thuật toán Cocke-Younger-Kasami (CYK) và thuật toán Earley. Trong khoá luận, chúng tôi sử dụng thuật toán Earley để làm bộ phân tích cú pháp. Vì giải thuật này dựa trên phương pháp lập bảng, tuy là phức tạp nhưng độ phức tạp của thuật toán lại nhỏ, $O(n^3)$. Giải thuật Earley và những cải tiến cho giải thuật sẽ được nói đến trong mục 1.5.

Giai đoạn này có bắt lỗi tốt hay không còn phụ thuộc vào việc chia nhỏ từ loại và lập luật sinh. Những điều này sẽ được nói đến trong Chương 2.

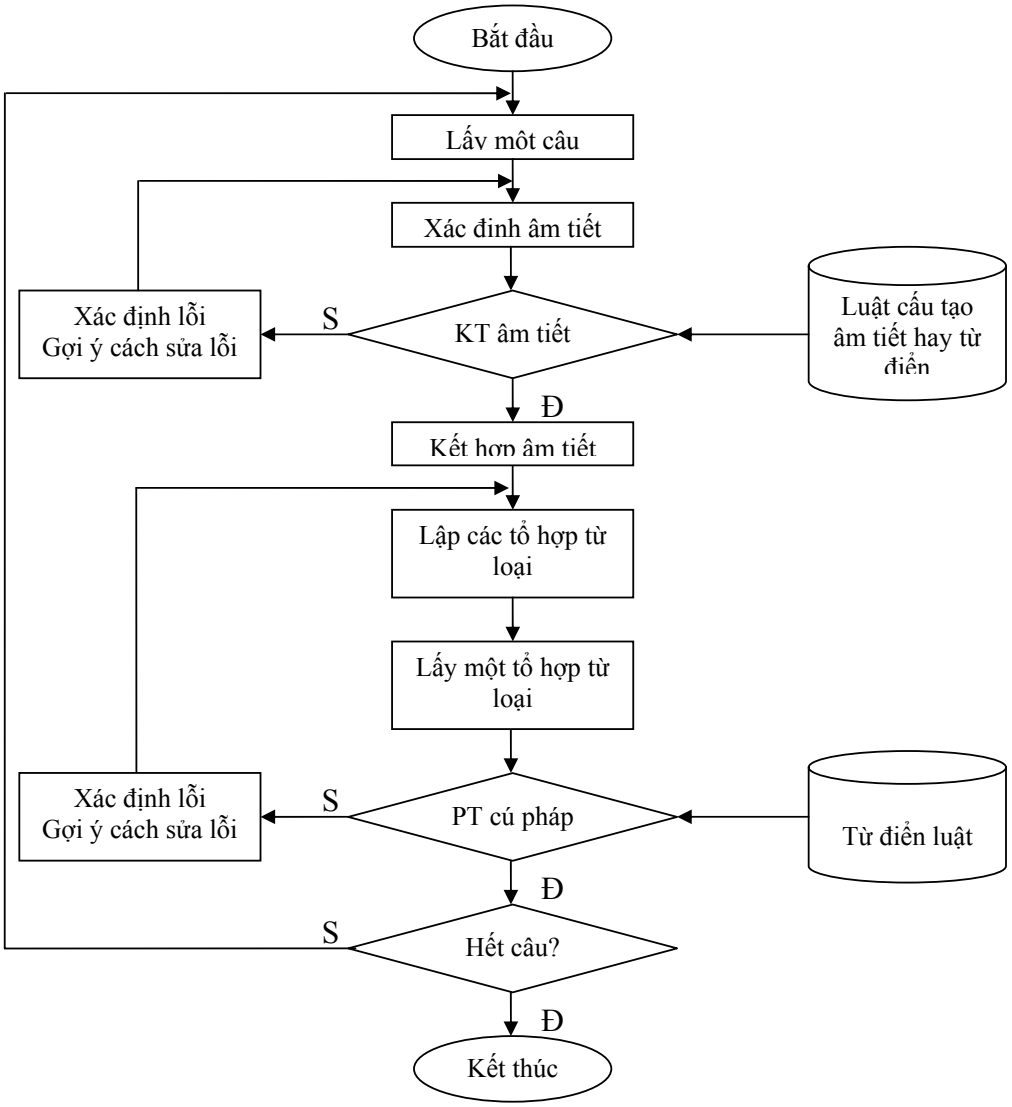
Phương pháp nói ở trên triệt để hơn rất nhiều, nó bắt lỗi được vừa lỗi âm tiết vừa lỗi cú pháp.

Nhưng do sự tinh tế của ngôn ngữ tiếng Việt, một số lỗi vẫn bị phương pháp này bỏ qua. Ví dụ, ta có cụm từ: “chính quyền lực này”, nhưng bị viết sai thành: “chính quyền lúc này”. Ở đây, bộ phân tích cú pháp vẫn báo đúng cho cụm từ sai ở trên. Có điều đó là do cụm từ sai này vẫn có chuỗi từ loại hợp với một luật sinh có sẵn và nó sẽ lọt qua trong phân tích cú pháp. Lớp lỗi này chỉ có thể bắt được thông qua ngữ nghĩa của câu, đoạn và văn bản mà thôi.

Trên thực tế, các lỗi loại này thường xuất hiện không nhiều trong tiếng Việt nên phương pháp trên vẫn có hiệu quả cao. Theo [8], thì phương pháp này có thể bắt lỗi đạt tới 95% tỷ lệ lỗi.

1.2. QUY TRÌNH BẮT LỖI CHÍNH TẢ:

Dựa trên tổng quan về phương pháp bắt lỗi chính tả tiếng Việt, chúng tôi đưa ra một qui trình bắt lỗi chính tả được trình bày bằng lưu đồ như sau:



1.3. TỔ CHỨC TỪ ĐIỂN:

Để hỗ trợ cho chương trình bắt lỗi chính tả hoạt động được tốt thì cần tổ chức một từ điển chính tả tiếng Việt chuẩn. Từ điển sẽ lưu trữ các từ và từ loại của từ. Từ điển phải đáp ứng các yêu cầu sau:

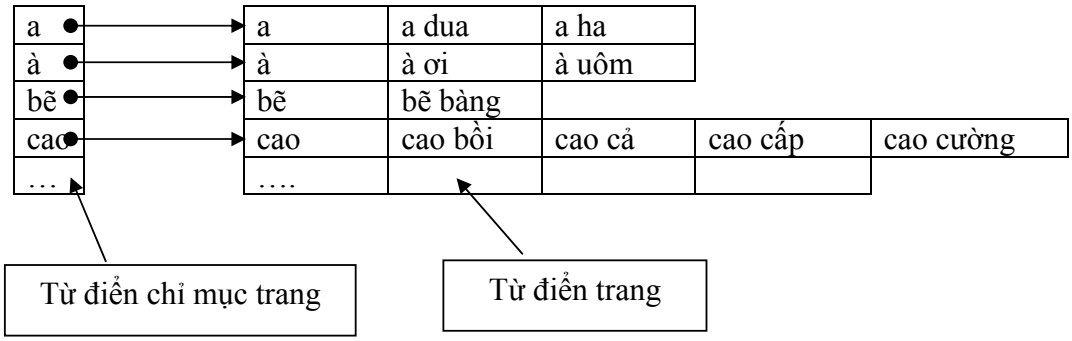
- Kích thước nhỏ, nhưng đầy đủ thông tin cần thiết.
- Truy xuất nhanh trong các tác vụ tìm kiếm, thêm, bớt.

Trong bắt lỗi chính tả, thao tác tìm kiếm từ trong từ điển là thao tác quan trọng nhất, nó ảnh hưởng đến tốc độ của giải thuật bắt lỗi. Do đó, từ điển sẽ được tổ chức theo dạng bảng băm có sắp xếp thứ tự và áp dụng thuật toán tìm kiếm nhị phân cho thao tác tìm kiếm trong từ điển. Cấu trúc từ điển sẽ như sau:

Từ điển được tổ chức thành nhiều trang. Các từ thuộc cùng một trang nếu có âm tiết đầu giống nhau. Nói một cách khác, từ điển sẽ được tổ chức thành hai danh sách:

- Một danh sách (từ điển chỉ mục trang) dùng để lưu chỉ mục, được sắp tăng dần. Danh sách này sẽ lưu các âm tiết đầu của từ.
- Một mảng các danh sách (từ điển trang) lưu nội dung các trang và cũng được sắp xếp tăng dần

Minh hoạ từ điển:



Như vậy, để tìm một từ trong từ điển, đầu tiên ta tìm xem âm tiết đầu tiên của từ đó có trong từ điển chỉ mục trang hay không. Nếu có tức là từ đó có thể có hoặc không có trong từ điển trang. Tiếp theo ta tìm kiếm nhị phân trong từ điển trang tại trang đã so khớp xem có từ cần tìm không.

Với cách tổ chức từ điển như trên, số lượng thành phần trong từ điển chỉ mục trang chính là số lượng âm tiết trong tiếng Việt, tức là khoảng hơn 5000 âm tiết. Còn số lượng từ trong mỗi trang nhiều nhất là 100 từ. Như thế, với thuật toán tìm kiếm như trên, ta có thời gian tìm kiếm một từ là: $\log_2 5000.\log_2 100$, tức xấp xỉ 86 đơn vị thời gian. Với cách xây dựng từ điển và thuật toán tìm kiếm như thế ta được số lần tìm kiếm nhỏ hơn nhiều so với kích thước từ điển.

1.4. BẮT LỖI CHÍNH TẢ MỨC ÂM TIẾT:

Để bắt lỗi chính tả mức âm tiết, ta có thể sử dụng một trong hai phương pháp sau đây:

- Bắt lỗi dựa trên nguyên lý cấu tạo âm tiết
- Bắt lỗi dựa trên từ điển

1.4.1. Bắt lỗi chính tả mức âm tiết dựa trên nguyên lý cấu tạo âm tiết:

Theo nguyên lý cấu tạo âm tiết, một âm tiết bao gồm ba thành phần: phụ âm đầu, nguyên âm và phụ âm cuối. Trong đó, thành phần nguyên âm là thành phần trung tâm kết hợp với hai thành phần còn lại. Thành phần nguyên âm này luôn có mặt trong các âm tiết tiếng Việt. Hai thành phần còn lại có thể có hoặc không.

Ví dụ:

- Đầy đủ ba thành phần: ruộng, vườn,…
- Không có phụ âm đầu: anh, em,….

- Không có phụ âm cuối: báo, chí,…
- Không có phụ âm đầu và cả phụ âm cuối: áo, eo,…

Như vậy, ta có thể tổ chức một cấu trúc lấy nguyên âm làm trung tâm, còn hai thành phần còn lại là tổ hợp của nhiều thành phần phụ có thể kết hợp với nguyên âm đó.

Trong ba thành phần của âm tiết thì hai thành phần phụ là có số lượng ít hơn cả và dễ quản lý nhất. Đối với thành phần phụ âm đầu, ta có tất cả là 26 phụ âm đầu: b, c, ch, d, đ, g, gh, h, k, kh, l, m, n, ng, ngh, nh, p, ph, q, r, s, t, th, tr, v, x. Do có trường hợp âm tiết không có phụ âm đầu nên ta phải quản lý cả trường hợp này. Trong lưu trữ, ta có thể quản lý bằng 27 bit dữ liệu, mỗi bit tương ứng với một phụ âm đầu ở trên.

Đối với thành phần phụ âm cuối, ta có tất cả 8 phụ âm cuối: c, ch, m, n, ng, nh, p, t. Cũng như trên, vẫn có âm tiết không có phụ âm cuối nên ta cũng phải quản lý trường hợp này. Trong lưu trữ, ta có thể quản lý phụ âm cuối bằng 9 bit tương ứng với từng phụ âm cuối.

Đối với nguyên âm, thành phần này rất phức tạp vì có cả dấu nên không thể lập mã như các thành phần kia. Riêng với thành phần này ta có thể giữ nguyên dạng chuỗi để lưu trữ.

Với cách mã hoá như trên, ta có thể lưu trữ theo một cấu trúc như sau:

- Thành phần phụ âm đầu, kiểu long
- Nguyên âm, kiểu string có độ dài 3
- Thành phần phụ âm cuối, kiểu integer

Với cấu trúc như trên, nhiều âm tiết sẽ được lưu trữ cùng một record. Đây là một cách để tiết kiệm không gian lưu trữ. Các record sẽ được lưu trữ vào một mảng được sắp xếp theo nguyên âm.

Sau đây là một ví dụ cho việc lưu trữ theo cấu trúc trên.

Ví dụ: Nguyên âm a có thể kết hợp với các phụ âm đầu sau: b, c và ch.

Nguyên âm a có thể kết hợp với các phụ âm cuối sau: n, m và nh.

Như thế, có thể tạo ra các âm tiết:

ban bam banh can cam canh chan cham chanh

Qua phân tích ta sẽ tạo thành một cấu trúc như sau:

- Nguyên âm: a
- Phụ âm đầu: 11100000000000000000000000 (viết dưới dạng nhị phân cho dễ thấy các bit nào được bật, với ví dụ này là các bit 1, 2, 3)
- Phụ âm cuối: 001101000 (với ví dụ này là các bit 3, 4, 6 được bật).

Như thế, để kiểm tra chính tả một âm tiết, ta chỉ cần tìm đến record có cùng thành phần nguyên âm. Sau đó, sử dụng phép toán AND để kiểm tra cho hai thành phần phụ.

Thuật toán kiểm tra chính tả của âm tiết:

1. Với một âm tiết vào, ta tách âm tiết ra làm ba thành phần.
2. Tìm kiếm nhị phân cho nguyên âm.

3. Nếu tồn tại record có cùng nguyên âm:
   i. Dùng phép toán AND để kiểm tra hai thành phần phụ;
   ii. Nếu cả hai đều khác 0, kết luận âm tiết đúng;
   iii. Ngược lại, kết luận âm tiết sai;
4. Nếu không tồn tại, kết luận âm tiết sai.

Với cách lưu trữ và thuật toán như trên, trong từ điển ta phải lưu trữ tất cả các quy tắc kết hợp của các âm tiết tiếng Việt mới đảm bảo hiệu quả của thuật toán.

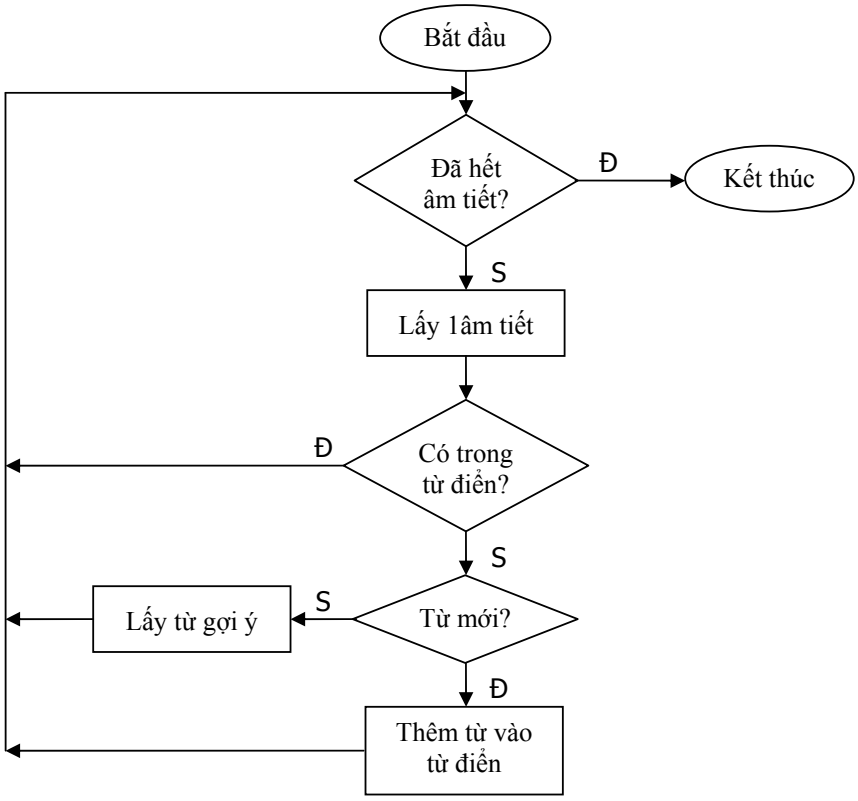
Tuy có nhiều ưu điểm trong lưu trữ và tốc độ nhưng phương pháp này chỉ phù hợp cho các âm tiết thuần Việt, còn các âm tiết được Việt hoá như: kilômét, ôxi… thì không quản lý được. Đây là một hạn chế do nguyên tắc cấu tạo của âm tiết tiếng Việt không xét đến những trường hợp này. Phương pháp này vẫn chưa hoàn chỉnh lắm.

Do hạn chế trên, khoá luận sẽ không sử dụng phương pháp này cho giai đoạn bắt lỗi chính tả mức âm tiết. Nhưng để minh hoạ phương pháp, chúng tôi cũng đã cài đặt phương pháp, phần cài đặt này sẽ được nói đến trong mục 3.1.2.

1.4.2. Bắt lỗi chính tả mức âm tiết dựa trên từ điển:

Phương pháp này sử dụng một từ điển để lưu trữ tất cả các âm tiết của tiếng Việt có thể có. Như thế, để kết luận một âm tiết là đúng chỉ cần kiểm tra xem âm tiết đó có trong từ điển hay không.

Thuật toán bắt lỗi chính tả mức âm tiết dựa trên từ điển:



Lợi dụng cấu trúc từ điển đã trình bày ở mục 1.3, chúng ta có thể kiểm tra các âm tiết có trong từ điển chỉ mục hay không. Như thế, từ điển phải lưu trữ thêm một lượng âm tiết không có nghĩa. Theo [8] lượng âm tiết này khoảng 1100. Cộng thêm số âm tiết của các từ có nghĩa thì lên đến khoảng 5000. (Trong thực tế cài đặt thì số lượng âm tiết cả có nghĩa và không có nghĩa lên đến hơn 5000.)

Như vậy, từ điển chỉ mục trang sẽ lưu trữ toàn bộ các âm tiết của tiếng Việt. Đồng thời, ta sử dụng thuật toán tìm kiếm nhị phân để tìm kiếm thì số phép toán cần để tìm một từ là $\log_2 5000 \approx 13$. Một con số khá nhỏ so với số lượng âm tiết trong từ điển.

Phương pháp này tuy tốn nhiều không gian lưu trữ hơn, nhưng lại giải quyết tất cả các kiểu âm tiết trong tiếng Việt. Ngoài ra, tốc độ tìm kiếm của phương pháp là rất nhanh. Chúng tôi sử dụng phương pháp này cho bắt lỗi chính tả mức âm tiết.

1.4.3. Phương pháp tạo gợi ý sửa lỗi mức âm tiết:

Khi bắt được một âm tiết sai nào đó, chương trình bắt lỗi cần phải đưa ra được những âm tiết đúng để gợi ý cho người sử dụng sửa lỗi.

Trong tiếng Việt, âm tiết dài nhất là "nghiêng" với bảy ký tự. Giả sử người gõ văn bản có thể gõ sai hai ký tự trên một âm tiết. Khi đó, độ dài âm tiết sai có thể lên đến chín ký tự.

Với quy ước số lượng ký tự có thể gõ sai là hai ký tự như trên, ta có thể định nghĩa, hai ký tự là "trùng" nhau nếu vị trí của chúng trong âm tiết chỉ chênh nhau nhiều nhất hai ký tự. Đồng thời, ta cũng có thể định nghĩa hai âm tiết là "so khớp" khi có ít nhất hai ký tự không "trùng".

Ví dụ:

Âm tiết "anh" và âm tiết "hai" có ký tự "a" trong hai âm tiết là trùng nhau.

Âm tiết "anh" và âm tiết "hanh" là hai âm tiết so khớp vì nó có ba âm tiết trùng nhau.

Qua thực tế nghiên cứu, định nghĩa trên không phù hợp cho các âm tiết có độ dài ký tự nhỏ hơn 4. Vì nó làm cho xuất hiện quá nhiều âm tiết hoàn toàn không phù hợp để gợi ý, do đó trong cài đặt, ta nên có những điều chỉnh thích hợp.

Ví dụ:

Ta muốn gõ âm tiết "ai" nhưng lại gõ sai thành "a8". Bộ bắt lỗi chính tả mức âm tiết sẽ bắt lỗi âm tiết "a8" và cho ra các âm tiết gợi ý như sau: a, ai, am, an, ai, au, ba, ca, da, ga, ha, la, ma, na, ra, sa, ta, va, xa, ùa, úa, đa, ưa, ỉa, ứa. (Phần gợi ý này đã có điều chỉnh cho các âm tiết nhỏ hơn 4 ký tự.)

Thuật toán tạo gợi ý sửa lỗi mức âm tiết:

1. Lấy từng âm tiết đúng có trong từ điển.
2. Nếu độ chênh lệch về độ dài âm tiết đúng với độ dài âm tiết cần sửa lỗi là nhỏ hơn hoặc bằng 2:
   i. Duyệt qua từng ký tự trong âm tiết đúng xem có trong âm tiết sai không;
   ii. Nếu không có, rõ ràng hai âm tiết không "so khớp", thoát khỏi vòng lặp;

iii. Ngược lại, xem hai ký tự có "trùng" vị trí không;
iv. Nếu có, tăng biến đếm ký tự "trùng" lên một;
v. Tiếp tục với ký tự tiếp theo;
vi. Ngược lại, thoát khỏi vòng lặp;
vii. Nếu số lượng ký tự "trùng" nhỏ hơn 2 so với độ dài âm tiết đúng thì đưa âm tiết đúng vào danh sách gợi ý;
viii. Ngược lại, tiếp tục vòng lặp ngoài;

3. Ngược lại, lấy âm tiết tiếp theo cho đến hết từ điển.

Thuật toán trên phụ thuộc vào cách tổ chức các từ đơn âm tiết trong từ điển. Theo cách tổ chức từ điển được trình bày ở mục 1.3, ta có các từ đơn âm tiết ở từ điển chỉ mục trang. Khoá luận sẽ thực hiện lấy các âm tiết gợi ý từ từ điển này. Thuật toán này có độ phức tạp $O(n^2)$.

Dù đã có cải tiến cho các âm tiết có độ dài nhỏ hơn bốn âm tiết, thuật toán trên vẫn còn đưa ra nhiều từ không phù hợp. Phương pháp này chưa có tính thông minh cao để chỉ đưa ra các âm tiết hợp lý nhất.

1.5. GIẢI THUẬT EARLEY CHO PHÂN TÍCH CÚ PHÁP:

Giải thuật Earley là một trong số các giải thuật được sử dụng để phân tích cú pháp trong xử lý ngôn ngữ tự nhiên. Đây là một giải thuật tổng quát, có thể phân tích bất kỳ văn phạm phi ngữ cảnh nào. Khoá luận sẽ sử dụng giải thuật Earley như bộ phân tích cú pháp trong giai đoạn bắt lỗi mức cú pháp.

1.5.1. Giải thuật Earley cơ bản:

Cho G=(V, W, S, P) là một văn phạm phi ngữ cảnh và $w=a_1 \ldots a_n \in V^*$. Khi đó, $A \rightarrow \alpha \bullet \beta$ là một luật có chấm khi $A \rightarrow \alpha\beta \in P$. Trong đó, luật có chấm giống như luật sinh bình thường nhưng có thêm một dấu chấm bên trong luật, thể hiện vị trí đang được phân tích trong luật đó. Giải thuật Earley được biểu diễn thông qua việc xây dựng bảng chứa tập các luật có chấm. Người ta xây dựng bảng Earley với các cột I_i (i=0..n), cột I_0 nhận giá trị khởi tạo, n là độ dài của chuỗi từ loại nhập. Mỗi ô sẽ có các giá trị: ***giá trị gốc*** để biết luật đó phát sinh từ cột nào và ***luật có chấm***.

Ví dụ: Giá trị gốc Luật có chấm

0 $S \rightarrow \bullet$CN VN
1 $VN \rightarrow$ ĐT$\bullet$ DT

1.5.1.1. Giải thuật:

Giải thuật bao gồm ba bước:

(1) Đoán nhận (Predict): Tại cột I_i:

Đối với các luật có ký tự không kết thúc ở bên phải dấu chấm, ta thêm các luật mới mà ký tự không kết thúc đó là vế trái của các luật. Giá trị gốc là i. Điều này có nghĩa là, với mỗi $[A \rightarrow \alpha \bullet B\beta, j]$ trong I_i, ta thêm $[B \rightarrow \bullet\gamma, i]$ vào I_i nếu $B \rightarrow \gamma \in P$.

(2) Duyệt (Scan): Tại cột I_i:

Đối với các luật mà ký tự kết thúc ở bên phải dấu chấm, luật này sẽ được chuyển sang cột I_{i+1} với dấu chấm được dịch ra sau ký tự kết thúc.

Tức là, với $[A \rightarrow \alpha \bullet a\beta, j]$ sẽ được đổi thành $[A \rightarrow \alpha a \bullet \beta, i]$ trong cột I_{i+1}.

(3) Hoàn thiện (Complete): Tại cột I_i:

Khi có luật [A→α•, j] thì sao chép và đổi [B→α•Aβ, k] trong cột I_j thành [B→αA•β, k] trong cột I_i.

1.5.1.2. Nhận xét:

- Đây là dạng phân tích từ trên xuống bởi vì ta bắt đầu với việc đoán nhận. Nếu ta thay đổi thứ tự trên, chúng ta sẽ có kiểu phân tích từ dưới lên.
- Thông thường, phân tích từ trên xuống có vấn đề với đệ qui trái, nhưng thuật toán Earley đã giải quyết bằng cách:

Mỗi luật giống nhau sẽ chỉ xuất hiện một lần trong mỗi cột. Có nghĩa là trong các bước thực hiện thuật toán, trước khi thêm một luật vào bảng thì phải kiểm tra xem nó có trùng với luật nào đã có trong cột cần thêm vào không. Nếu không thì thêm vào, còn có thì không thêm vào.

- Chuỗi từ loại là sai cú pháp khi ta đã duyệt qua hết các luật trong I_i mà I_{i+1} rỗng và chưa thể kết thúc bảng hợp lệ.
- Chuỗi từ loại là đúng cú pháp khi kết thúc chuỗi từ loại mà ta có luật khởi tạo tại cột cuối cùng.

Nói chung, chuỗi đúng khi tại điểm kết thúc chuỗi nhập, mà dấu chấm đã di chuyển ra sau ký tự bắt đầu S.

- Với việc sử dụng giá trị đoán nhận trước, có thể giúp ta tránh dư thừa.

Ví dụ, ta có luật VN → VN •BN tại vị trí kết thúc chuỗi nhập. Thông thường, ta sẽ đi đoán nhận BN, nhưng trong trường hợp này là không nên, ta chỉ nên làm như thế nếu còn từ trong chuỗi nhập.

Mặt khác, giá trị đoán nhận trước cũng gây ra sự phức tạp, và tăng số lượng luật được lưu trữ.

- Độ phức tạp của thuật toán là $O(n^3)$, với n là độ dài chuỗi nhập (bằng số lượng cột của bảng giảm đi một).

1.5.2. Những cải tiến cho giải thuật Earley:

Giải thuật Earley vẫn có những hạn chế nhất định trong xử lý ngôn ngữ tự nhiên. Sau đây là những cải tiến nhỏ để cải thiện tốc độ cho giải thuật Earley.

1.5.2.1. Giải quyết vấn đề luật dư thừa trong giải thuật Earley:

Với giải thuật đã nêu ở trên, ta có thể nhận thấy có rất nhiều luật dư thừa vẫn được lưu trữ trong bảng Earley. Như thế đồng nghĩa với việc phải duyệt qua quá nhiều luật dư thừa như Kilbury đã nhận xét [5]. Qua nghiên cứu, chúng tôi nhận thấy các luật dư thừa có dạng như sau:

- Thứ nhất, luật chỉ có một kí tự kết thúc ở vế trái mà không khớp với giá trị đoán nhận;
- Thứ hai, luật không dẫn đến đệ qui trái.

1.5.2.1.1. Dạng luật sinh:

Để giải quyết triệt để vấn đề này, khoá luận đã thực hiện lập luật sinh với dạng riêng. Tất cả các luật sinh đều thuộc vào một trong hai dạng sau:

- $A \rightarrow \alpha$, $\alpha \in W^*$
- $A \rightarrow a$, $a \in V$

Dạng luật trên vẫn phù hợp với văn phạm phi ngữ cảnh.

Với các luật sinh thuộc vào hai dạng trên thì các luật dư thừa sẽ là:

- Luật có vế phải chỉ là một ký tự kết thúc ($A \rightarrow a$);
- Luật có vế phải là các ký tự không kết thúc mà không dẫn đến đệ qui trái ($A \rightarrow B\alpha$).

1.5.2.1.2. Giải thuật cải tiến:

Với dạng luật như trên, giải thuật được cải tiến chỉ còn lại hai bước như sau:

(1) Đoán nhận:

Với mỗi $[A \rightarrow \alpha \bullet B\beta, j]$ trong I_i.

Lấy các luật trong từ điển luật sinh.

Duyệt qua các luật dạng $B \rightarrow a$, nếu khớp với giá trị đoán nhận thì đưa luật khớp cùng với các luật dạng $B \rightarrow B\alpha$ vào bảng Earley.

Ngược lại, nếu không so khớp với giá trị đoán nhận thì đưa toàn bộ các luật dạng $B \rightarrow \alpha$ vào bảng Earley.

(2) Hoàn thiện:

Như giải thuật Earley cũ.

Giải thuật cải tiến ở trên chỉ còn hai bước, bước quét trong giải thuật cũ đã được bỏ đi là do dạng luật sinh và giai đoạn đoán nhận mới đã giải quyết luôn bước này. Giải thuật cải tiến đã giải quyết được vấn đề luật dư thừa, nó không còn phải duyệt qua các luật không cần thiết trong phân tích cú pháp nữa. Như thế, sẽ cải thiện tốc độ của tiến trình xử lý nhiều hơn. Ngoài ra, còn giảm không gian lưu trữ. Nhưng giải thuật cải tiến vẫn có độ phức tạp thời gian là $O(n^3)$. Vì giải thuật chỉ mới thay đổi nội dung bên trong cấu trúc của giải thuật Earley chứ chưa thay đổi được cấu trúc của giải thuật nên độ phức tạp thời gian vẫn là như cũ.

1.5.2.2. Giải quyết vấn đề bùng nổ tổ hợp:

Hiện tượng bùng nổ tổ hợp xảy ra do một từ có thể thuộc vào nhiều từ loại khác nhau. Nhưng một đặc điểm dễ nhận thấy của các tổ hợp từ loại được sinh ra từ một câu là luôn có những đoạn con từ loại giống nhau trên các tổ hợp từ loại.

Ví dụ: (Ví dụ được lấy trong [10])

Đoạn câu cần phân tích: trong biên chế và hưởng lương từ ngân sách.

Có 12 chuỗi từ loại được xuất ra như sau:

A11 N22 L10 V43 N23 F10 N23 N50
A11 N22 L10 V43 N23 F11 N23 N50
A11 V40 L10 V43 N23 F10 N23 N50
A11 V40 L10 V43 N23 F11 N23 N50
F10 N22 L10 V43 N23 F10 N23 N50

F10 N22 L10 V43 N23 F11 N23 N50
F10 V40 L10 V43 N23 F23 N23 N50
F10 V40 L10 V43 N23 F10 N23 N50
F11 N22 L10 V43 N23 F10 N23 N50
F11 N22 L10 V43 N23 F11 N23 N50
F10 V40 L10 V43 N23 F10 N23 N50
F10 V40 L10 V43 N23 F11 N23 N50

Dựa vào đặc điểm này TS. Phan Thị Tươi trong [10] đã nêu một phương pháp để tăng tốc độ phân tích cú pháp như sau (trích từ [10]):

Nếu bộ phân tích thất bại khi đang kiểm tra một chuỗi, thì nó sẽ so trùng các chuỗi còn lại với đoạn vừa kiểm tra thành công và sẽ tiếp tục quy trình phân tích ở vị trí của một chuỗi khác có chuỗi con dài nhất trùng với đoạn đã phân tích. Quá trình này được lặp lại cho tới khi bộ phân tích duyệt qua hết một chuỗi nào đó. Lúc đó, câu nhập được xác nhận là đúng cú pháp. Ngược lại khi đi đến chuỗi cuối cùng mà vẫn không phân tích thành công thì bộ phân tích sẽ kết luận rằng câu nhập vào không đúng cú pháp.

Đây là một phương pháp hay, nó giúp ta tránh phải phân tích lại những gì đã phân tích rồi. Nhưng phương pháp trên lại chỉ hiệu quả trong trường hợp câu nhập vào là đúng, còn trong trường hợp câu nhập vào là sai thì không hiệu quả.

Qua nghiên cứu, chúng tôi còn nhận thấy còn có hiện tượng trùng một phần (chỉ một phần ngắn của đoạn đã kiểm tra thành công) bên cạnh hiện tượng trùng cả đoạn như đã nói ở trên. Điều này cho thấy ta có thể giảm thêm được một số bước phân tích nữa.

Thừa kế từ phương pháp trong [10] và bổ sung điều mới phát hiện, khoá luận đưa ra phương pháp như sau:

Ta sắp xếp các chuỗi tổ hợp từ loại theo thứ tự. Như thế, chuỗi ngay sau chuỗi đang xét sẽ là chuỗi có khả năng trùng đoạn đã phân tích. Khi kiểm tra một chuỗi thất bại ta chỉ việc so khớp đoạn vừa kiểm tra thành công với chuỗi ngay sau nó và lấy số từ loại so khớp liên tục bắt đầu từ đầu chuỗi. Việc phân tích sẽ được thực hiện tiếp với chuỗi ngay sau tại vị trí đầu tiên không so khớp.

Bước đầu tiên của giải thuật trên là sắp xếp các chuỗi tổ hợp từ loại, chúng tôi thực hiện bước này là để có thể sử dụng phương pháp trong [10]. Khi đã được sắp xếp, rõ ràng chuỗi từ loại ngay sau chuỗi vừa kiểm tra có xác suất trùng đoạn vừa kiểm tra thành công là lớn nhất. Như thế, ta chỉ việc thực hiện tiếp tục phân tích với chuỗi từ loại tiếp ngay sau. Giải thuật cải tiến đã thừa kế trọn vẹn phương pháp trong [10], đồng thời còn thay thế công đoạn tìm kiếm chuỗi trùng khớp bằng chỉ một bước đơn giản là tăng giá trị duyệt tổ hợp chuỗi từ loại lên một. Đây là một cải tiến đáng kể.

1.6. BẮT LỖI CHÍNH TẢ MỨC CÚ PHÁP:

Bắt lỗi chính tả mức cú pháp dựa trên việc phân tích cú pháp câu. Do đó, ta phải có các chuỗi tổ hợp từ loại để phân tích. Một bước quan trọng và là giai đoạn tiền xử lý của giai đoạn phân tích cú pháp là phân tích từ. Đây là giai đoạn quan trọng, nó tách các từ, lấy tổ hợp các câu có thể có và lấy tổ hợp ghép từ loại.

1.6.1. Phân tích từ:

Từ là đơn vị có sẵn của ngôn ngữ, là cái ngầm định và mọi người coi sự hiện diện của từ là tất nhiên. Hầu như không có quy tắc cấu tạo đối với từ. Do vậy, ta phải so sánh các từ tách được từ văn bản với các từ thực tế để kiểm nghiệm từ đó có phải là từ không.

Một vấn đề còn gây nhiều tranh luận trong việc phân tích từ là chọn độ dài tối đa của từ ghép. Trong tiếng Việt, ngoài các từ đơn, ta còn có các từ ghép hai âm tiết, ba âm tiết,… Nếu ta chọn giới hạn độ dài một từ quá lớn sẽ làm mất nhiều thời gian kiểm tra các từ không cần thiết; còn nếu chọn ngưỡng quá nhỏ, ta sẽ mất từ. Khoá luận sẽ chọn ngưỡng là 3 vì đối với các văn bản pháp quy hành chính thì từ dài nhất là 3 âm tiết.

Ta thực hiện tính số các khả năng ghép nối từ, từ loại:

- **Một câu có n âm tiết và ngưỡng ghép từ là 3 thì ta sẽ có tối đa 3n-3 từ khác nhau.**

Xét câu $a_1\ a_2\ a_3 \ldots a_n$,
Các từ có thể xuất phát từ từ a_1 là a_1, $a_1\ a_2$, $a_1\ a_2\ a_3$ (3 từ),
Các từ có thể xuất phát từ từ a_2 là a_2, $a_2\ a_3$, $a_2\ a_3\ a_4$ (3 từ),
…
Các từ có thể xuất phát từ từ a_{n-2} là a_{n-2}, $a_{n-2}\ a_{n-1}$, $a_{n-2}\ a_{n-1}\ a_n$ (3 từ),
Các từ có thể xuất phát từ từ a_{n-1} là a_{n-1}, $a_{n-1}\ a_n$ (2 từ),
Các từ có thể xuất phát từ từ a_n là a_n (1 từ),
Như vậy tổng số từ là: 3(n-2)+2+1 = 3n-3n từ.

- **Một câu có n âm tiết và ngưỡng ghép từ là 3 thì ta sẽ có 2^{n-1}-2n+7 cách tách từ khác nhau (n>3).**

Xét câu $s=a_1\ a_2\ a_3 \ldots a_n$,
Với câu có 1 âm tiết, số cách chia là: $s=a_1$ - 1 cách = 2^0,
Với câu có 2 âm tiết, số cách chia là: $s=a_1a_2$; $s=a_1\ a_2$ - 2 cách = 2^1,
Với câu có 3 âm tiết, số cách chia là:
$s=a_1a_2a_3$; $s=a_1a_2\ a_3$; $s=a_1\ a_2a_3$; $s=a_1\ a_2\ a_3$ - 4 cách = 2^2,
Với câu có 4 âm tiết, số cách chia là:
$s=a_1a_2a_3\ a_4$; $s=a_1a_2\ a_3a_4$; $s=a_1a_2\ a_3\ a_4$; $s=a_1\ a_2a_3a_4$;
$s=a_1\ a_2a_3\ a_4$; $s=a_1\ a_2\ a_3a_4$; $s=a_1\ a_2\ a_3\ a_4$ - 7 cách = 2^3-1,
Tương tự ta có với câu 5 âm tiết, số cách chia sẽ là: 13 cách = 2^4-3.
Như vậy, theo qui nạp toán học, ta có thể kết luận là:
+ Với n<=3, ta có 2^{n-1} cách tách từ.
+ Với n>3, ta có 2^{n-1}-2n+7 cách tách từ.

- **Một câu có n âm tiết, ngưỡng ghép từ là 3, mỗi từ có tối đa k kiểu từ loại, ta sẽ có tối đa $k(k+1)^{n-1}-(2n^2-13n+21)(k+1)$ tổ hợp từ loại.**

Câu có một âm tiết sẽ có k tổ hợp từ loại.

Dễ dàng nhận thấy, cứ bổ sung thêm một âm tiết (với ngưỡng tách không giới hạn) thì số tổ hợp sẽ tăng lên k+1 lần. Còn đối với ngưỡng tách là 3 thì số tổ hợp sẽ tăng lên theo hai trường hợp:

- Với n<=3, số tổ hợp sẽ tăng lên k+1 lần.
- Nhưng với n>3, số tổ hợp sẽ tăng lên $[2^{n-1} – (2(n-4)+1)](k+1)$ lần.

Như vậy, số tổ hợp từ loại sẽ là:

+ Với n<=3: $k(k+1)^{n-1}$

+ Với n>3: $k(k+1)^{n-1}-(2n^2-13n+21)(k+1)$.

Qua các phép tính trên, ta nhận thấy số cách tách từ và số tổ hợp từ loại sẽ tăng lên theo hàm mũ khi số lượng âm tiết tăng. Nó sẽ gây ra hiện tượng bùng nổ tổ hợp. Nhưng đây lại là điều không thể tránh khỏi. Ta chỉ có thể làm giảm bớt số lượng cách tách từ cần kiểm tra. Cụ thể, trong quá trình tìm kiếm các cách tách từ, nếu ta phát hiện một cách tách từ nào đó không phù hợp, ta phải loại bỏ tất cả các nhánh xuất phát từ cách tách từ đó.

Sau đây, khoá luận sẽ đưa ra một phương pháp để giải quyết vấn đề này.

Ta xây dựng một mảng hai chiều với 3 cột và n dòng, n là độ dài của câu. Trong mảng, cột thứ nhất chứa các từ có một âm tiết, cột thứ hai chứa các từ có hai âm tiết, cột thứ ba chứa các từ có ba âm tiết. Các từ sẽ được đưa vào mảng nếu là từ đúng.

Ví dụ với câu: “phó giáo sư là một chức danh”. Ta sẽ có mảng như sau:

phó		phó giáo sư
giáo	giáo sư	
sư		
là		
một		
chức	chức danh	
danh		

Dựa trên bảng trên, ta có thể có các câu sau (Các từ sẽ được gạch chân):

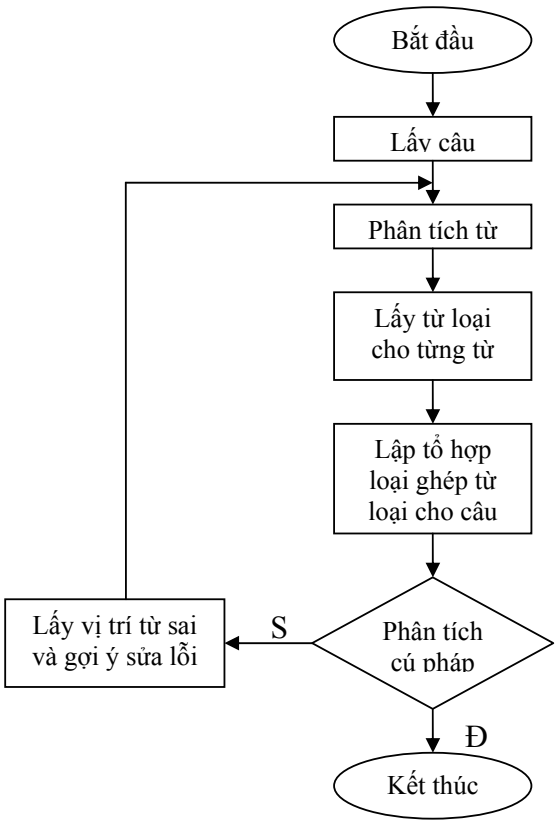
- phó giáo sư là một chức danh
- phó giáo sư là một chức danh
- phó giáo sư là một chức danh
- phó giáo sư là một chức danh
- phó giáo sư là một chức danh
- phó giáo sư là một chức danh

Một thuật toán duyệt qua bảng để lấy các câu như trên là đơn giản (sẽ không trình bày ở đây). Đây sẽ là một thuật toán đệ qui, nó loại bỏ được các nhánh ghép từ không phù hợp.

Phương pháp này có một điểm yếu là vẫn phải duyệt qua tất cả 3n-3 từ đề nghị để kiểm tra. Nhưng lại hạn chế được hiện tượng bùng nổ tổ hợp trong quá trình tách câu và có một thuật toán tách câu thành từ đơn giản hơn nhiều. Thuật toán có độ phức tạp là $O(n^3)$.

1.6.2. Thuật toán bắt lỗi chính tả mức cú pháp:

Qui trình bắt lỗi mức cú pháp bằng cách phân tích cú pháp một câu được trình bày bằng lưu đồ như sau:



Thuật toán này có độ phức tạp chính là độ phức tạp của thuật toán phân tích cú pháp.

1.6.3. Phương pháp tạo gợi ý sửa lỗi mức cú pháp

Phương pháp gợi ý sửa lỗi mức âm tiết đã trình bày ở mục 1.4.3 là phương pháp tạo gợi ý cho một âm tiết. Ở mức cú pháp, các từ sai không phải chỉ là các từ một âm tiết mà có thể là các từ hai âm tiết hoặc ba âm tiết, nên phương pháp trên không còn phù hợp nữa. Ta phải có phương pháp gợi ý thích hợp cho trường hợp này.

Về cơ bản, các từ tiếng Việt đều là sự kết hợp của một hoặc nhiều âm tiết. Các âm tiết trong một từ có thể nói là độc lập. Do đó, ta có thể thừa kế phương pháp gợi ý sửa lỗi mức âm tiết cho từng âm tiết một, rồi kết hợp các âm tiết gợi ý thành từ để được các từ gợi ý.

Ví dụ:

Ta muốn đánh từ đúng là "giáo sư" nhưng lại đánh sai thành "giao sư". Qua phân tích cú pháp câu có thể bắt lỗi được từ này.

Khi đó, với âm tiết "giao" ta có các âm tiết gợi ý: giao, giáo, ngoái,…

Và với âm tiết "sư" ta có các âm tiết gợi ý: sư, sưa, sưu,…

Lần lượt kết hợp các âm tiết trên lại, từ nào có nghĩa sẽ là từ gợi ý. Như thế, ta chỉ có từ "giáo sư" là từ duy nhất có nghĩa.

Thuật toán tìm từ gợi ý sửa lỗi mức cú pháp như sau:

1. Lấy số âm tiết của từ.
2. Tách từng âm tiết.
3. Duyệt qua từng âm tiết một.
    i. Lấy các âm tiết gợi ý.
4. Kết hợp các âm tiết gợi ý thành từ.
5. Duyệt qua các từ vừa tạo thành
    i. Nếu từ có nghĩa, đưa vào danh sách từ gợi ý.
    ii. Ngược lại, duyệt qua từ khác.

Thuật toán này có độ phức tạp là $O(n^3)$.

Chương 2

NGỮ PHÁP TIẾNG VIỆT VỚI PHÂN TÍCH CÚ PHÁP

Để bắt lỗi chính tả mức cú pháp ta phải phân tích cú pháp câu, xem xét các cách kết hợp các từ với nhau để tìm được các từ sai. Như thế, ngữ pháp tiếng Việt đóng một vai trò quan trọng trong giai đoạn này.

Nói đến ngữ pháp tiếng Việt là nói đến hai mặt từ pháp và cú pháp. Hai mặt này kết hợp với nhau, hỗ trợ lẫn nhau vô cùng chặt chẽ. Từ pháp nghiên cứu về cấu tạo từ và từ loại, còn cú pháp thì nghiên cứu về sự cấu tạo của cụm từ và câu. Nếu chỉ biết có từ và sự phân chia từ loại mà không biết cấu trúc cụm từ và câu thì không thể đặt câu. Và ngược lại, nếu chỉ biết cấu tạo của cụm từ và câu mà không biết về cấu tạo từ và từ loại thì đặt câu sẽ sai. Do đó, khoá luận đã thực hiện nghiên cứu ngữ pháp tiếng Việt thông qua nghiên cứu về sự phân chia từ loại và các đặc trưng cú pháp của từ loại, nghiên cứu cú pháp của các cụm từ và câu.

2.1. TỪ LOẠI:

Từ loại được coi là vấn đề thuộc phạm trù từ vựng - ngữ pháp, hiểu đơn giản là phạm trù ngữ pháp của các từ. Nói đến từ loại là nói đến sự phân lớp các từ vựng trong vốn từ vựng của một ngôn ngữ.

Cho đến hiện nay, vấn đề phân định các từ loại trong ngôn ngữ của chúng ta vẫn chưa được giải quyết triệt để. Do đó, vấn đề phân chia từ loại cho phù hợp, chính xác, thoả mãn các yếu tố cần xem xét vẫn là một vấn đề mở. Vấn đề này cần được kết hợp với những nhà ngôn ngữ học để giải quyết.

Theo GS. Diệp Quang Ban [1] thì từ loại trong tiếng Việt bao gồm các loại dưới đây:

2.1.1. Danh từ:

Danh từ là thực từ có ý nghĩa thực thể (ý nghĩa sự vật hiểu rộng) kết hợp được (về phía sau) với các từ chỉ định (này, nọ) và thường ít khi tự mình làm vị ngữ (thường phải đứng sau từ là).

Đặc điểm ngữ pháp của danh từ:

- Không trực tiếp làm vị ngữ. Chỉ có thể làm vị ngữ khi kết hợp với **Hệ từ** *là* (câu khẳng định), hoặc *không phải, không phải là* (câu phủ định). Không đặt sau các từ như: *đừng*, *hãy*, *sẽ*...
- Kết hợp được với **Đại từ chỉ định** (*này*, *nọ*,...), **Số từ** (*một*, *hai*,...), **Đại từ chỉ số lượng** (*tất cả*), **Định từ** (*những*, ...), **Danh từ chỉ loại** (*con*, *cái*).

Ví dụ:

- **Anh** này là **sinh viên**.
- Tôi mua tất cả những **quyển vở** này.
- Tôi nuôi hai **con vịt**.

2.1.2. Động từ:

Là loại từ biểu thị quá trình (sự hoạt động, động tác, hành vi, biến hoá và trạng thái).

Đặc điểm ngữ pháp:

- Có thể trực tiếp làm vị ngữ, không cần có **Hệ từ** *là*.
- Không đứng sau **Số từ**, **Định từ**, **Đại từ chỉ định**.

Ví dụ:

đọc, thực hiện, lấy (hoạt động), lo, kính nể, vui (trạng thái tâm lý), nghĩ ngợi, đau ốm, ngủ, nằm (trạng thái),…

2.1.3. Tính từ:

Tính từ là thực từ có ý nghĩa tính chất hiểu như là đặc trưng trực tiếp của sự vật, hiện tượng kết hợp được về phía trước với các từ *rất*, *cực*, *kỳ*, *hơi*, *khí*, *quá* hoặc về phía sau với các từ *lắm*, *quá*, *cực kỳ*, thường làm định ngữ và vị ngữ trong câu.

Đặc điểm ngữ pháp:

- Có thể trực tiếp làm vị ngữ.
- Không kết hợp được với *hãy*.

Ví dụ:

tốt, xấu, đẹp, vụng, nhiều, ít, đông, thưa, méo, tròn, xanh, đỏ, vang, dội, trầm, bổng, thơm, nồng, cay, ngọt,…

2.1.4. Số từ:

Là loại từ chỉ số lượng và thứ tự.

Đặc điểm cú pháp:

- Số từ số lượng phải đứng trước **Danh từ** và **Danh từ chỉ loại**.
- Số từ số lượng có thể làm vị ngữ trong trường hợp nói về tuổi tác, phân lượng.
- Số thứ tự bao giờ cũng đặt sau **Danh từ**. Nếu danh từ có Định ngữ phức hợp thì nó thường đứng sau **Tính từ**, trước **Đại từ chỉ định**.
- Số thứ tự có khả năng làm vị ngữ không cần Hệ từ *là*.

Ví dụ:

một, hai (số từ số đếm xác định), một vài, dăm ba (số từ số đếm phỏng định), thứ nhất, thứ nhì (số thứ tự),…

2.1.5. Đại từ:

Là loại từ không gọi tên gì cả mà chỉ dùng để trỏ những cái nói trên.

Đặc điểm ngữ pháp:

- Không có định ngữ và bổ ngữ do thực từ đảm nhận.

Ví dụ:

- **Anh ta** tự trách **mình.**
- **Chúng** gặp **nhau** trên đường quần ngựa. (Nam Cao) (nhau là đại từ tương hỗ)

2.1.6. Phụ từ:

Phụ từ là những từ chuyên đi kèm các từ khác, tự mình không có khả năng làm thành tố chính trong cụm từ chính phụ và cũng không có khả năng thay thế (như đại từ).

Căn cứ vào hoạt động ngữ pháp, có thể chia phụ từ thành hai lớp nhỏ:

- Phụ từ chuyên đi kèm danh từ, sẽ được gọi là **Định từ** (hay phụ danh từ).
- Phụ từ chuyên đi kèm vị từ (động từ và tính từ) sẽ được gọi là **Phó từ** (hay phụ vị từ).

2.1.6.1. Định từ:

Định từ chuyên đi kèm phía trước danh từ và chỉ quan hệ về số lượng. Các định từ thường gặp là *các*, *những*, *một*, *mọi*, *mỗi*, *từng*, *mấy*…

Ba từ *những*, *một*, *các* và dạng zêrô làm thành một đối hệ và được gọi là các **Quán từ**.

Các từ *mọi*, *mỗi*, *từng* có ý nghĩa phân phối.

Từ *mấy* dùng như *những*, *các* thường gặp ở phía nam và làm mất ranh giới của sự đối lập về thể thiết định.

2.1.6.2. Phó từ:

Phó từ chuyên đi kèm vị từ về phía trước và về phía sau.

Ví dụ:

đều, cũng, cùng, vẫn, cứ, từng, đã, mới, thường, hay, đang, rất, hơi, khí, quá, hãy, đừng, chớ, cho,…

2.1.7. Quan hệ từ:

Quan hệ từ là những hư từ dùng để liên kết các từ với nhau hoặc các vế trong câu. Trong tiếng Việt, quan hệ từ được phân biệt rõ thành **Giới từ** và **Liên từ** như sau:

2.1.7.1. Giới từ:

Giới từ dùng để nối danh từ - thành tố chính hoặc bổ ngữ gián tiếp với động từ - thành tố chính, một số bổ ngữ cảnh huống với động từ - thành tố chính.

Ví dụ:

của, bằng, do, vì, tại, bởi, để, từ, đến, ở, trong, ngoài, trên, dưới, đối với, với, như, về, cho, mà,…

2.1.7.2. Liên từ:

Liên từ dùng để nối các yếu tố ngôn ngữ có quan hệ bình đẳng với nhau về ngữ pháp hoặc quan hệ qua lại về ngữ pháp và về ý. Loại thứ nhất là **Liên từ bình đẳng**, loại thứ hai là **Liên từ qua lại** hoặc chính phụ.

Ví dụ:

và, với, cùng, cùng với, hay, hay là (liên từ bình đẳng), nhưng, song, mà, huống hồ, giả sử, hễ, vì, cho nên, do (liên từ qua lại),…

2.1.8. Tình thái từ:

Tình thái từ là hư từ chỉ mối quan hệ của người nói (sự nhấn mạnh, độ tin cậy, thái độ, đánh giá…) với nội dung nói hay người nghe. Khác với các phụ từ là những từ nằm trong cấu tạo cụm từ, tình thái từ thường chỉ xuất hiện ở bậc câu, tuy về mặt nội dung thì có thể liên hệ với một từ, một cụm từ hay cả câu.

Tình thái từ được chia thành hai nhóm là **Trợ từ** (nhấn mạnh) và **Tiểu từ tình thái**.

2.1.8.1. Trợ từ:

Trợ từ là những từ dùng để nhấn mạnh vào một từ, một cụm từ, một câu nào đó mà nó đi kèm. Trợ từ không có ý nghĩa riêng của mình. Hai trợ từ hay gặp là *mà* và *thì*. Với tư cách trợ từ, *mà* và *thì* chỉ có tác dụng nhấn mạnh hay đánh dấu một ranh giới nào đó vì có thể rút bỏ ra khỏi câu, và không gợi lên một kết từ tương ứng hay thay được bằng một kết từ, tương ứng.

Ví dụ:

- Trời hôm nay khi **thì** mưa khi thì nắng.
- Ai **mà** chẳng biết việc ấy.

Mà hay đi kèm với liên từ làm thành một khối, trong khối này nó có tác dụng nhấn mạnh. Ví dụ: *nếu mà*, *để mà*. Tuy nhiên, khi vắng liên từ *thì*, *mà* đảm nhiệm vai trò của liên từ.

Ví dụ:

- **Nếu mà** thầy biết thầy rầy chết!

2.1.8.2. Tiểu từ tình thái:

Tiểu từ tình thái là những từ dùng tạo dạng cho câu phân biệt theo mục đích nói (câu nghi vấn, câu cầu khiến, câu cảm thán) và bày tỏ quan hệ của người nói với nội dung câu nói hay với người nghe.

Ví dụ:

à, ư, nhỉ, nhé, chứ, chăng (câu nghi vấn), đi, nào, thôi, với, nhé (câu cầu khiến), thay, thật (câu cảm thán),…

ạ (sự kính trọng, thân thương), kia (chỉ một hướng khác của ý, không lường trước), vậy (miễn cưỡng, sự tất yếu), mà (phân trần, giải thích, nài nỉ), chỉ (hạn chế về lượng), những, những là (sự quá ngưỡng), chính, đích thị, ngay cả (nhấn mạnh ý xác tín), đây, ấy, này, nào (đại từ chỉ định), có thể, chắc hẳn, hình như,…

2.1.9. Thán từ:

Thán từ là từ - tín hiệu phản ánh các trạng thái tâm sinh lý.

Đặc điểm cú pháp:

- Có thể một mình làm thành câu, hoặc làm thành phần phụ biệt lập của câu.

Ví dụ:

Ôi, ối, a, ồ, ái, á, ái chà, úi, eo ôi, trời, mẹ cha ơi, tội nghiệp, khổ thân nó, gớm, hoan hô, muôn năm, hỡi, ơi, bớ, ừ, dạ, thưa, bẩm, lạy, trình, báo cáo, có tôi, có mặt…

2.2. PHÂN CHIA TỪ LOẠI VÀ LẬP MÃ:

Trong từng từ loại trên, người ta lại phân chia thành các tiểu loại nhằm đạt đến tính đúng cả về cú pháp và ngữ nghĩa. Việc chia nhỏ từ loại sẽ giúp cho ta càng đạt đến tỷ lệ bắt lỗi cao hơn.

Chúng tôi đã tiến hành chia nhỏ từ loại và lập mã như sau:

1. Danh từ - mã hoá dạng A__
   Ví dụ: A03 – Danh từ chỉ đồ vật
2. Động từ - mã hoá dạng B__
   Ví dụ: B06 - Động từ chỉ hoạt động
3. Tính từ - mã hoá dạng C__
   Ví dụ: C02 – Tính từ chỉ phẩm chất
4. Số từ - mã hoá dạng D__
   Ví dụ: D01 - Số từ số đếm
5. Đại từ - mã hoá dạng E__
   Ví dụ: E07 - Đại từ chỉ định
6. Định từ - mã hoá dạng F__
   Ví dụ: F01 – Quán từ
7. Phó từ - mã hoá dạng G__
   Ví dụ: G06 – Phó từ chỉ tần số
8. Giới từ - mã hoá dạng H__
   Ví dụ: H01 - Giới từ
9. Liên từ - mã hoá dạng I__
   Ví dụ: I03 – Liên từ lựa chọn
10. Trợ từ - mã hoá dạng J__
    Ví dụ: J01 - Trợ từ
11. Tiểu từ tình thái – mã hoá dạng K__
    Ví dụ: K01 - Tiểu từ tạo dạng câu nghi vấn
12. Thán từ - mã hoá dạng L__
    Ví dụ: L03 – Thán từ gọi đáp
13. Một số từ loại đặc biệt – mã hoá dạng N__
    Ví dụ: N01 - Dấu phẩy
14. Từ không có nghĩa – mã hoá dạng KN0

Ví dụ:

Với câu:

Nhưng các lỗi chính tả vẫn thường xuyên xuất hiện trong các văn bản.

Qua phân tích cú pháp ta sẽ có chuỗi từ loại đúng như sau:

I05 A10 A08 A08 G02 G06 B06 H01 A10 A02

2.3. CỤM TỪ:

Cụm từ là những kiến trúc gồm hai từ trở lên kết hợp tự do với nhau theo những quan hệ ngữ pháp hiển nhiên nhất định không chứa kết từ ở đầu.

Trong cụm từ người ta còn có khái niệm ngữ cố định và ngữ (còn gọi là cụm từ nửa cố định). Ngữ cố định là những kiến trúc cho sẵn gồm hai từ trở lên, có tính chất bền vững về từ vựng và ngữ pháp, thường được sử dụng như những khuôn dấu, không thay đổi, hoặc có thể thay đổi trong một khuôn khổ hạn hẹp.

Ví dụ:

- múa rìu qua mắt thợ.
- ông chằng bà chuộc.

Còn ngữ là một cụm từ chính phụ có thành tố chính cho sẵn và thành tố phụ thay đổi theo một khuôn ngữ pháp cố định.

Ví dụ:

- cái nhà, cây tre, con mèo,…
- màu đỏ, số bốn, ngày mai, hôm nay,…
- đi làm, đi săn, đi chơi,…

Ngoài ra, dựa trên kiểu quan hệ có tính chất chuyên môn trong việc nghiên cứu ngữ pháp. Trong tiếng Việt có ba kiểu quan hệ cú pháp phổ biến trong cụm từ sau đây:

- Quan hệ giữa chủ ngữ với vị ngữ, gọi tắt là *quan hệ chủ vị*.
- Quan hệ giữa thành tố chính với thành tố phụ về ngữ pháp, gọi là *quan hệ chính phụ*.
- Quan hệ giữa hai yếu tố chính bình đẳng với nhau, về ngữ pháp gọi là *quan hệ bình đẳng*.

Ví dụ:

- Cái xe này *máy hỏng*. (cụm chủ vị)
- *Tập thể dục* rất có ích. (cụm chính phụ)
- *Thầy giáo và cô giáo* đi tham quan với học sinh ngày mai. (cụm đẳng lập)

Nhưng nói chung, cụm từ thường được gọi tên theo từ loại của thành tố chính trong cụm. Trong tiếng Việt, chúng ta có thể gặp những loại cụm từ sau:

- Cụm danh từ - có danh từ làm thành tố chính
  Ví dụ: mấy *người* này, hai *người*, *người* nọ,…
- Cụm động từ - có động từ làm thành tố chính
  Ví dụ: đã *đọc* rồi, vừa *đọc*, *đọc* được,…
- Cụm tính từ - có tính từ làm thành tố chính
  Ví dụ: vẫn *tốt* hơn, rất *tốt*, *tốt* quá,…
- Cụm số từ - có số từ làm thành tố chính
  Ví dụ: hơn *ba mươi* một chút, độ *ba mươi*, *ba mươi* hơn,…
- Cụm đại từ - có đại từ làm thành tố chính
  Ví dụ: tất cả *chúng tôi* đây, hai *chúng tôi*,…

Trong đó, ba loại cụm từ đầu có cấu tạo đa dạng hơn nên sẽ được xem xét ở các mục sau.

2.3.1. Cụm danh từ:

Cụm danh từ là tổ hợp từ tự do không có kết từ đứng đầu, có quan hệ chính phụ giữa thành tố chính với thành tố phụ, và thành tố chính là danh từ.

Cấu tạo chung của cụm danh từ gồm ba phần: phần trung tâm, phần phụ trước, phần phụ sau. Phần trung tâm có thể là một danh từ hoặc một ngữ danh từ. Ngữ danh từ gồm một danh từ chỉ loại đứng trước và một danh từ chỉ sự vật hay một động từ, tính từ chỉ hoạt động, trạng thái, tính chất, quan hệ đứng sau (*cái* nhà, *cây* tre, *con* mèo, *người* thợ, *niềm* vui, *cuộc* họp, *vẻ* đẹp,…). Ngoài ra, ngữ danh từ có thể là các tổ hợp hai danh từ có quan hệ bình đẳng.

Theo Diệp Quang Ban [1], cụm danh từ có cấu trúc như sau:

tất cả	những	cái	con mèo	đen	ấy
-3	-2	-1	0	1	2

- **Phần trung tâm cụm danh từ (Vị trí 0):**
  Ở đây, ta có hai dạng:
  - Là danh từ đếm được tuyệt đối.
  - Ngữ danh từ: danh từ chỉ loại + danh từ|động từ|tính từ

- **Phần phụ trước của cụm danh từ:**
  - **Vị trí từ chỉ xuất (Vị trí -1):**
    Có thể thuộc vào các dạng sau:
    - từ *cái* đứng trước danh từ chỉ vật, sau danh từ thường kèm từ chỉ định *này*, *kia*, *ấy*,…
    - từ *cái* đứng trước danh từ chỉ loại trong phần trung tâm (*cái* cây tre này, *cái* con mèo này,…)

    Vị trí này còn có thể sử dụng một số danh từ chỉ loại khác.
  - **Vị trí từ chỉ số lượng (Vị trí -2):**
    Có thể thuộc vào các dạng sau:
    - Số từ xác định
    - Số từ phỏng định
    - Từ hàm ý phân phối: *mỗi*, *từng*, *mọi*
    - Quán từ: *những*, *các*, *một*
    - Từ *mấy*
  - **Vị trí từ chỉ tổng lượng (Vị trí -3):**
    Các từ chỉ tổng lượng là: *tất cả*, *hết thảy*, *tất thảy*, *cả*,…

- **Phần phụ sau của cụm danh từ:**
  - **Vị trí từ nêu đặc trưng miêu tả (Vị trí 1):**
    Tại vị trí này có thể gặp các dạng cụm từ sau: cụm danh từ (có thể là động từ hoặc tính từ) chính phụ, cụm danh từ (có thể là động từ hoặc tính từ) đẳng

lập, cụm chủ vị, số từ xác định, số từ thứ tự, đại từ, thời vị từ. Có thể bị ngăn cách với thành phần trung tâm bởi một kết từ.

- **Vị trí từ chỉ định (Vị trí 2):**

Các từ chỉ định như *này*, *kia*, *nọ*, *ấy*,… thường hay xuất hiện ở vị trí này.

2.3.2. Cụm động từ:

Cụm động từ là tổ hợp từ tự do không có kết từ đứng đầu, có quan hệ chính phụ giữa thành tố chính với thành tố phụ, và thành tố chính là động từ.

Cấu tạo chung của cụm động từ cũng gồm ba thành phần: phần trung tâm, phần phụ trước, phần phụ sau.

Cụ thể các yếu tố trong ba thành phần như sau:

- **Thành phần trung tâm:**

Tất cả các dạng từ loại động từ đều có thể làm thành tố chính của cụm động từ. Tại thành phần này, thành tố chính có thể là một động từ hoặc là một chuỗi động từ.

Ví dụ:

- Tôi đang *xem* sách.
- Tôi đang *ngồi xem* sách.

- **Thành phần phụ trước:**

Những từ có thể làm thành phần phụ trước của cụm động từ được xếp vào hai dạng: phó từ và thực từ. Phó từ là một từ loại đã được nói đến ở mục 2.1.6.2.

Trật tự của các phó từ được sắp xếp theo lược đồ sau:

<table>
<tr><td rowspan="3">Nhóm: đều, cũng, vẫn, cứ</td><td rowspan="2">Nhóm: từng, đã, đang, sẽ</td><td>Nhóm: không, chưa</td><td rowspan="3">Nhóm: hay, năng, ít</td></tr>
<tr><td>Nhóm: rất, hơi</td></tr>
<tr><td colspan="2">Nhóm: đừng, chớ</td></tr>
</table>

Trong lược đồ trên, những nhóm nằm trong cùng cột là những nhóm bài trừ lẫn nhau, không xuất hiện đồng thời trong một cụm từ.

Còn đối với thực từ, có thể gặp hai kiểu thực từ sau:

- Những 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: *ào ào* chảy, *lác đác* rơi, *khẽ* kêu, *ôn tồn* đáp,…
- Kiến trúc gồm một kết từ với một danh từ chỉ điểm xuất phát. Kiến trúc này thường đứng trước các động từ chỉ hướng. Các kết từ thường gặp là: *từ*, *ở*, *dưới*, *trên*, *trong*, *ngoài*,… Ví dụ: *từ* quê ra, *dưới* quê lên,…

- **Thành phần phụ sau:**

Thành tố phụ sau của cụm động từ thường là một từ hoặc một cụm từ đẳng lập, hay chính phụ, hay chủ vị. Tựu chung thành tố phụ sau có hai trường hợp sau:

- Thành tố phụ song hành: là trường hợp hai thành tố phụ đồng thời xuất hiện và cũng có những quan hệ xác định với động từ - thành tố chính. Thành tố phụ song hành có thể là hai danh từ chỉ đối tượng hoặc có thể là một danh từ nêu đối tượng và một động từ nêu đặc trưng hành động hay đối tượng. Ví dụ: đưa cho *bà cụ phong thư*, mượn *bạn quyển sách*, bảo *bạn* chép *bài* hộ, gọi *người ấy* bằng *anh*,…
- Thành tố phụ là cụm chủ vị. Ví dụ: chúng tôi cần *các bạn giúp cho một hôm nữa*, biết *bạn sắp đi xa*,…

2.3.3. Cụm tính từ:

Cụm tính từ là tổ hợp từ tự do không có kết từ đứng đầu, có quan hệ chính phụ giữa thành tố chính với thành tố phụ, và thành tố chính là tính từ.

Cấu tạo chung của cụm tính từ cũng gồm có ba thành phần: phần trung tâm, phần phụ trước, phần phụ sau. Cụ thể như sau:

- **Thành phần trung tâm:**

Thành phần trung tâm của cụm tính từ đặc biệt hơn so với các thành phần trung tâm của các cụm danh từ và cụm động từ là không có trường hợp tổ hợp tính từ làm thành phần trung tâm. Điều này giúp cho thành phần trung tâm cụm tính từ đơn giản hơn.

- **Phần phụ trước:**

Những từ làm thành tố phụ thường là: *rất*, *hơi*, *khí*, *cực*, *cực kỳ*, *tuyệt*, *quá*. Ví dụ:

rất đẹp, *cực* đẹp, *tuyệt* đẹp, *hơi* vụng, *khí* vụng,…

Ngoài những từ có tính chất chuyên dụng vừa nêu, tại phần phụ trước cụm tính từ, như đã nói, có thể xuất hiện hầu hết các phụ từ đi với động từ (trừ *hãy*, *đừng*, *chớ*).

- **Phần phụ sau:**

Phần phụ sau cụm tính từ có thể phân biệt:

- Những phụ từ làm thành tố phụ sau cụm tính từ: Những phụ từ thường làm thành tố phụ sau cụm tính từ là: *lắm*, *cực*, *cực kỳ*, *tuyệt*, *quá*. Ví dụ: đẹp *lắm*, đẹp *cực kỳ*, đẹp *tuyệt*,…
- Những thực từ làm thành tố phụ sau của cụm tính từ: tính từ chỉ lượng và tính từ chỉ tình trạng cộng với danh từ chỉ chủ thể; tính từ chỉ quan hệ định vị cộng với danh từ chỉ không gian, địa điểm; thực từ chỉ phương diện, nội dung trong quan hệ ý nghĩa của tính từ (trường hợp này có thể kết hợp với kết từ).

2.4. CÂU:

Câu là đơn vị cấu trúc lớn nhất trong tổ chức ngữ pháp của một ngôn ngữ. Câu có thể được phân loại theo hai phương diện: phương diện cấu tạo ngữ pháp và phương diện mục đích nói. Về phương diện cấu tạo ngữ pháp, câu được phân loại ra câu đơn hai thành phần và câu đơn đặc biệt, câu phức, câu ghép.

2.4.1. Câu đơn:

Câu đơn được chia thành các loại sau:

- Câu đơn hai thành phần
- Câu đơn đặc biệt
- Câu tĩnh lược

2.4.1.1. Câu đơn hai thành phần:

Câu đơn hai thành phần là câu đơn có một kết cấu chủ ngữ - vị ngữ và kết cấu ấy đồng thời cũng là nòng cốt câu. Các thành phần trong câu đơn hai thành phần bao gồm:

- **Chủ ngữ**: Về từ loại, thì ở vị trí chủ ngữ có thể là các từ thuộc danh từ, đại từ nhân xưng, tính từ, động từ, số từ, đại từ thay thế. Về cấu tạo cú pháp, chủ ngữ có thể được làm thành một từ hoặc cụm từ đẳng lập, cụm từ chính phụ, cụm từ chủ vị.

Ví dụ:

- **Mèo** là động vật ăn thịt. (một danh từ)
- **Mười** lớn hơn chín. (một số từ)
- **Tốt danh** hơn lành áo. (cụm động từ chính phụ)

- **Vị ngữ**: Về từ loại, có thể xuất hiện các từ thuộc từ loại động từ, tính từ. Về cấu tạo cú pháp, vị ngữ có thể được làm thành từ một từ, cụm từ đẳng lập, cụm từ chính phụ, cụm từ chủ vị.

Ví dụ:

- Gà **gáy**. (một động từ)
- Hoa **đẹp**. (một tính từ)
- Họ **mới đến hôm qua**. (cụm động từ chính phụ)

- **Bổ ngữ:**Về từ loại, bổ ngữ có thể được làm từ danh từ, số từ, động từ, tính từ, đại từ nhân xưng hoặc đại từ thay thế và cũng có thể được thể hiện bằng một số phụ từ: *hay*, *năng*, *ít*. Về cấu tạo cú pháp, bổ ngữ có thể là một từ, cụm từ đẳng lập, cụm từ chính phụ, và có thể có giới từ đứng trước.

Ví dụ:

- Mẹ rửa chân cho **con**.
- Họ cử ông ấy làm **giám đốc**.

- **Đề ngữ**: Về từ loại, đề ngữ có thể do danh từ, số từ, động từ, tính từ đảm nhiệm. Về phương diện cấu tạo, đề ngữ có thể được làm thành từ một từ, cụm từ đẳng lập, cụm từ chính phụ hoặc cụm từ chủ vị. Đề ngữ cũng có thể có quan hệ từ đứng trước. Sau đề ngữ và trước nòng cốt câu có thể thêm trợ từ *thì*, *mà* có khi là trợ là *là*.

Ví dụ:

- **Sách này** tôi đọc rồi.
- **Còn chị**, chị công tác ở đây à?

- **Trạng ngữ**: Về từ loại, trạng ngữ có thể được diễn đạt bằng danh từ, bằng động từ, bằng tính từ. Về cấu tạo cú pháp, trạng ngữ có thể được làm thành một từ, cụm từ đẳng lập, cụm từ chính phụ, và thường có quan hệ từ đưa trạng ngữ vào câu.

Ví dụ:

- **Ngoài sân**, hai con mèo đang vờn nhau.
- **Hôm qua**, Giáp đi câu cá.

- **Định ngữ**: Về cấu tạo, định ngữ có thể là một từ và cũng có thể là một tổ hợp từ như các loại cụm từ hoặc một tổ hợp từ có giới từ đứng đầu.

Ví dụ:

- Trăng **rằm** vừa tròn vừa sáng. (một danh từ)
- Nhà Giáp có một con mèo **rất đẹp**. (cụm tính từ)

- **Phần tình thái**: Đối với thành phần này không thể nghiên cứu theo dạng từ loại và cấu trúc tạo thành vì nó phức tạp hơn nhiều.

Ví dụ:

- **Kể** người ta giàu cũng sướng. (Nguyễn Công Hoan)
- **Chết thật**, tôi không nhận ra. (Nguyễn Đình Thi)

- **Phần phụ chú**: Về cấu trúc, phần phụ chú có thể diễn đạt bằng một từ hay một cụm từ (chính phụ, đẳng lập, chủ vị), nhưng cũng có khi được diễn đạt bằng nhiều câu.

Ví dụ:

- Bởi vì… bởi vì… (**San cúi mặt và bỏ tiếng Nam dùng tiếng Pháp**) người ta lừa dối anh. (Nam Cao)
- Ở thành thị thì trong xí nghiệp khác, trong trường học khác… **nghĩa là mỗi nơi có một nội dung cụ thể khác nhau**. (Phạm Văn Đồng)

- **Phần nối kết**: Về phương tiện diễn đạt, phần nối kết thường hay gồm có: quan hệ từ, tổ hợp đại từ và quan hệ từ, từ ngữ khác không chứa quan hệ từ.

2.4.1.2. Câu đơn đặc biệt:

Câu đơn đặc biệt là câu đơn được làm thành từ một trung tâm cú pháp chính (có thể có thêm trung tâm cú pháp phụ), không chứa hay không hàm ẩn một trung tâm cú pháp chính thứ hai có quan hệ với trung tâm cú pháp chính nói trên như là quan hệ giữa chủ ngữ với vị ngữ.

Về cấu tạo, câu đơn đặc biệt được làm thành từ một từ hoặc một cụm từ chính phụ hay cụm từ đẳng lập. Các từ thành tố chính thường gặp là danh từ hay động từ, tính từ. Câu đơn đặc biệt cũng có thể có một trung tâm cú pháp phụ đi kèm làm thành phần phụ trạng ngữ.

Ví dụ:

- Bom tạ. (danh từ) (Nguyễn Đình Thi)
- Một thứ im lặng ghê người. (cụm danh từ chính phụ) (Nam Cao)
- Nhiều sao quá. (cụm tính từ) (Nguyễn Đình Thi)

- *Ở làng này*, khó lắm. (trạng ngữ in nghiêng) (Nam Cao)

2.4.1.3. Câu đơn tĩnh lược:

Câu tĩnh lược không phải là một kiểu câu riêng. Trong phần lớn trường hợp, câu tĩnh lược gắn với câu đơn hai thành phần.

Có hai trường hợp cho câu tĩnh lược:

- Trường hợp tĩnh lược riêng chủ ngữ, riêng vị ngữ, hoặc riêng bổ ngữ.
- Trường hợp cùng một lúc tĩnh lược chủ ngữ và động từ thành tố chính của vị ngữ, tức là câu chỉ còn chứa bổ ngữ.

Ví dụ:

- Tiếng hát ngừng. Cả tiếng cười. (tĩnh lược vị ngữ) (Nam Cao)
- Tôi nghĩ đến sức mạnh của thơ. Chức năng và vinh dự của thơ. (tĩnh lược chủ ngữ và vị ngữ) (Phạm Hổ)

2.4.2. Câu phức:

Câu phức khác câu đơn ở chỗ trong câu phức có chứa hai hoặc hơn hai kết cấu chủ vị. Câu phức giống câu ghép ở chỗ trong cả hai kiểu câu này đều có chứa hai hoặc hơn hai kết cấu chủ vị, tuy nhiên chỗ khác nhau rất cơ bản giữa chúng là ở kiểu quan hệ giữa các kết cấu chủ vị với nhau.

Ở câu phức, tuy có hai (hoặc hơn hai) kết cấu chủ vị, nhưng trong số đó chỉ có một kết cấu chủ vị nằm ngoài cùng và làm nòng cốt của câu, các kết cấu chủ vị còn lại bị bao bên trong kết cấu chủ vị làm nòng cốt câu đó.

Ví dụ:

Nó	bảo Đ	nó [C	đi Đà Nẵng V]
C	V		

Câu phức bao gồm các kiểu sau:

- Câu phức có chủ ngữ là kết cấu chủ vị.
  Ví dụ: **Chuột chạy** làm vỡ đèn.
- Câu phức có vị ngữ là kết cấu chủ vị.
  Ví dụ: Cây này **lá vàng**.
- Câu phức có bổ ngữ là kết cấu chủ vị.
  Ví dụ: Nó bảo **nó đi Đà Nẵng**.
- Câu phức có định ngữ là kết cấu chủ vị.
  Ví dụ: Con mèo **Giáp mua** chạy mất rồi.
- Câu phức là câu bị động.
  Ví dụ: Thuyền được **người lái đẩy ra xa**.

2.4.3. Câu ghép:

Câu ghép là câu chứa hai (hoặc hơn hai) kết cấu chủ vị, trong số đó không kết cấu chủ vị nào bao kết cấu chủ vị nào; mỗi kết cấu chủ vị diễn đạt một sự việc (còn gọi là sự thể), và các sự việc này có quan hệ với nhau theo những mối quan hệ nào đó.

Câu ghép bao gồm các kiểu sau:

2.4.3.1. Câu ghép bình đẳng:

Câu ghép bình đẳng là câu ghép trong đó có quan hệ từ bình đẳng về ngữ pháp nối các vế câu ghép với nhau. Các quan hệ từ bình đẳng thường được nhắc đến trong câu ghép là *và*, *mà*, *còn*, *nhưng*, *rồi*, *hay*.
Ví dụ:

- Lốp xe nổ **và** chiếc xe dừng lại
- Tôi thích bóng đá **mà** bạn Giáp thì lại thích bóng chuyền.
- Tôi làm bài tập, **còn** bạn Giáp thì đang viết thư.
- Tôi thích bóng đá, **nhưng** bạn Giáp lại thích bóng chuyền.
- Bạn cứ làm như thế, **rồi** tôi sẽ chỉ cho mà làm tiếp.
- Mình đọc **hay** tôi đọc.

2.4.3.2. Câu ghép chính phụ:

Câu ghép chính phụ là câu ghép chứa vế câu trong đó có quan hệ từ phụ thuộc dẫn đầu. Vế chứa quan hệ từ phụ thuộc là vế phụ, vế còn lại là vế chính. Câu ghép chính phụ có các kiểu câu sau:

- **Câu ghép nguyên nhân**: là câu ghép chính phụ mà ở đầu vế phụ có chứa các quan hệ từ diễn đạt quan hệ nguyên nhân như *vì*, *do*, *tại*, *bởi*, *nhờ*,… Còn vế chính có thể xuất hiện các từ (*cho*) *nên*, *mà*.
- **Câu ghép điều kiện/giả thiết**: là câu ghép chính phụ mà ở đầu vế phụ có chứa các quan hệ từ diễn đạt quan hệ điều kiện/giả thiết như *nếu*, *hễ*, *miễn* (*là*), *giá*… Còn trong vế chính thường xuất hiện từ *thì*.
- **Câu ghép nghịch đối**: là câu ghép chính phụ mà ở đầu vế phụ có chứa các quan hệ từ phụ thuộc *tuy*, *mặc dầu*, *dù*,… Còn vế chính thường xuất hiện *những* (*mà*), *mà*.
- **Câu ghép mục đích**: là câu ghép chính phụ mà ở đầu vế phụ có chứa quan hệ từ diễn đạt quan hệ mục đích *để*. Còn trong vế chính thường xuất hiện từ *thì*.

2.4.3.3. Câu ghép qua lại:

Câu ghép qua lại là câu ghép dùng các cặp phụ từ hô ứng ở mỗi vế để nối kết hai vế câu lại với nhau; ngoài ra còn có trường hợp một vế câu chứa phụ từ *đang* và một vế câu chứa quan hệ từ *thì*. Câu ghép qua lại có các dạng sau:

- Câu ghép dùng cặp phụ từ vừa… vừa…
  Ví dụ:
  - Họ vừa đi, họ vừa hát.
  - Tôi vừa ngồi xuống, thì cái ghế vừa gãy.
- Câu ghép dùng cặp phụ từ vừa (mới)… đã…
  Ví dụ:
  - Chúng tôi vừa mới đến, thì xe đã chạy.

- Câu ghép dùng cặp phụ từ mới… đã
  Ví dụ:
  - Họ mới đến thì xe đã chạy mất.
  - Xe mới chạy đến đây đã nổ lốp.
- Câu ghép dùng cặp phụ từ chưa… đã…
  Ví dụ:
  - Tôi chưa nói gì, đứa bé đã khóc.
  - Bọn trẻ chưa kịp xì hơi, quả bóng đã nổ.
- Câu dùng phụ từ đang và quan hệ từ thì
  Ví dụ:
  - Tôi đang đứng chờ xe thì một cậu bạn chạy đến.
  - Xe đang chạy thì lốp xe bị xẹp.
- Câu ghép dùng cặp phụ từ còn (đang)… đã…
  Ví dụ:
  - Mọi người còn đang tắm dưới sông thì Giáp đã lên bờ.
  - Mọi người còn tắm dưới sông, Giáp đã lên bờ.
- Câu ghép dùng cặp phụ từ còn… còn…
  Ví dụ:
  - Anh còn đánh nó, nó còn không sợ anh.
  - Còn nước, còn tát.
- Câu ghép dùng cặp phụ từ càng… càng…
  Ví dụ:
  - Anh càng khoẻ, anh càng làm được nhiều việc.
  - Vì anh càng tỏ ra thích mua, nên họ càng ép giá.
- Câu ghép dùng cặp phụ từ chẳng những… mà còn…
  Ví dụ:
  - Cậu bé không những không bớt sốt, mà lại còn sốt cao hơn.
  - Nó chẳng những không nghe lời, mà nó còn cãi lại.
- Câu ghép dùng cặp đại từ phiếm định – xác định
  Ví dụ:
  - Ai làm, nấy chịu trách nhiệm.
  - Anh bảo sao, tôi làm vậy.

2.5. LUẬT SINH:

Dựa vào việc phân tích các cú pháp cụm từ và câu như trên, chúng ta có thể viết các luật sinh phục vụ cho phân tích cú pháp.

Để thuận tiện trong lưu trữ và xử lý, khoá luận tiến hành mã hoá cho một số thành phần như sau:

- Cụm danh từ - mã hoá thành dạng O__
- Cụm động từ - mã hoá thành dạng P__
- Cụm tính từ - mã hoá thành dạng Q__
- Câu – mã hoá thành dạng S__

- Chủ ngữ - mã hoá thành dạng R__
- Vị ngữ - mã hoá thành dạng T__
- Đề ngữ - mã hoá thành dạng U__
- Trạng ngữ - mã hoá thành dạng V__
- Định ngữ - mã hoá thành dạng W__
- Bổ ngữ - mã hoá thành dạng X__

Ví dụ:

Xét câu đơn giản nhất:

"Tôi cười."

Ta có các luật sinh sau:

<Câu> → <Chủ ngữ> <Vị ngữ>
<Chủ ngữ> → <Ngữ danh từ>
<Ngữ danh từ> → <Đại từ> tôi
<Vị ngữ> → <Ngữ động từ>
<Ngữ động từ> → <Động từ> cười

Ví dụ:

Xét câu nhiều thành phần hơn:

"Anh ta chạy rất nhanh ."

Ta phân tích thành các luật sinh sau:

<Câu> → <Chủ ngữ> <Vị ngữ>
<Chủ ngữ> → <Ngữ danh từ>
<Ngữ danh từ> → <Danh từ> <Đại từ> anh ta
<Vị ngữ> → <Ngữ động từ> <Bổ ngữ>
<Ngữ động từ> → <Động từ> chạy
<Bổ ngữ> → <Phó từ> <Tính từ> rất nhanh

Ví dụ:

Xét câu phức tạp hơn:

"Nhưng các lỗi chính tả vẫn thường xuyên xuất hiện trong các văn bản."

Ta phân tích thành các luật sinh sau:

<Câu> → <Liên từ> <Chủ ngữ> <Vị ngữ>
<Chủ ngữ> → <Danh từ chỉ loại> <Ngữ danh từ> các lỗi chính tả
<Ngữ danh từ> → <Danh từ> <Danh từ> lỗi chính tả
<Ngữ danh từ> → <Danh từ chỉ loại> <Danh từ> các văn bản
<Vị ngữ> → <Ngữ động từ> <Bổ ngữ>
<Ngữ động từ> → <Phó từ> <Ngữ động từ>
<Ngữ động từ> → <Phó từ> <Động từ> thường xuyên xuất hiện
<Bổ ngữ> → <Giới từ> <Ngữ danh từ> trong các văn bản

Ví dụ:

Xét câu ghép bình đẳng và câu tỉnh lược thành phần chủ ngữ:

"Điều đó làm cho các văn bản mất giá trị và cũng như sẽ gây ra những hiểu lầm rất nguy hiểm."

Ta phân tích thành các luật sau:

<Câu> → <Câu> <Liên từ> <Câu>
<Câu> → <Chủ ngữ> <Vị ngữ>
<Câu> → <Vị ngữ>
<Chủ ngữ> → <Ngữ danh từ>
<Ngữ danh từ> → <Danh từ> <Đại từ> điều đó
<Vị ngữ> → <Ngữ động từ> <Bổ ngữ>
<Ngữ động từ> → <Động từ> <Động từ> làm cho
<Ngữ động từ> → <Phó từ> <Ngữ động từ> cũng như sẽ gây ra
<Ngữ động từ> → <Động từ>
<Bổ ngữ> → <Ngữ danh từ> <Bổ ngữ>
<Bổ ngữ> → <Động từ> <Danh từ> mất giá trị
<Bổ ngữ> → <Phó từ> <Bổ ngữ> rất nguy hiểm

Ví dụ:

Xét câu ghép chính phụ:

"Nếu trời mưa thì đường trơn."

Ta phân tích thành các luật sau:

<Câu> → <Liên từ> <Câu> <Liên từ> <Câu> nếu… thì…
<Câu> → <Chủ ngữ> <Vị ngữ>
<Chủ ngữ> → <Ngữ danh từ>
<Ngữ danh từ> → <Danh từ> trời, đường
<Vị ngữ> → <Ngữ động từ>
<Vị ngữ> → <Ngữ tính từ>
<Ngữ động từ> → <Động từ> mưa
<Ngữ tính từ> → <Tính từ> trơn

Chương 3

CÀI ĐẶT CHƯƠNG TRÌNH

3.1. HỆ THỐNG CHƯƠNG TRÌNH BẮT LỖI CHÍNH TẢ:

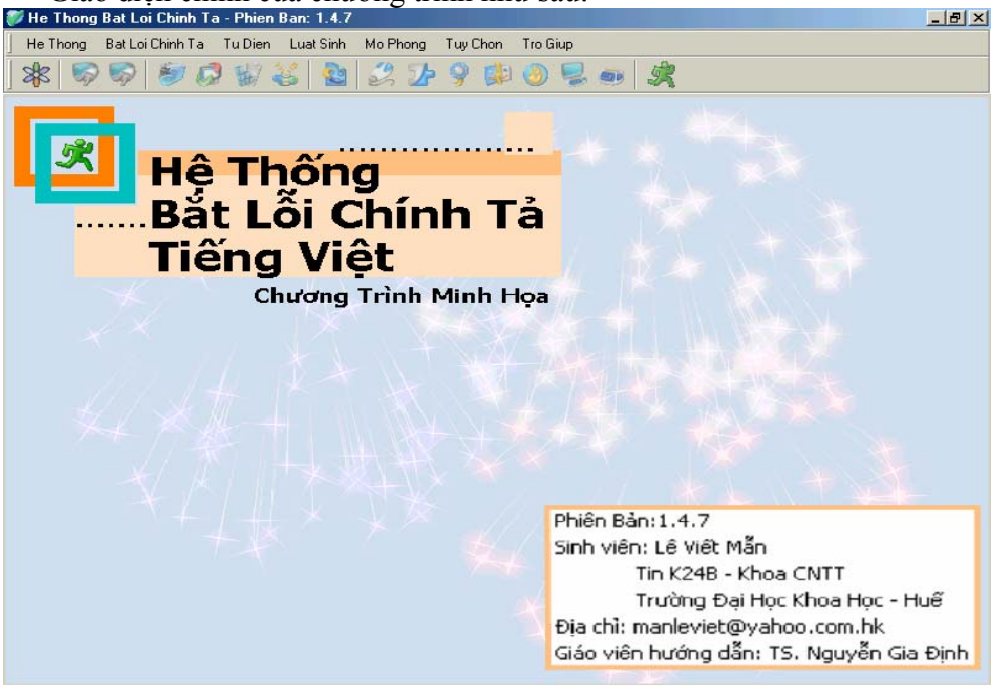
Với tính chất minh hoạ, khoá luận đã xây dựng chương trình bắt lỗi chính tả và một hệ thống các chương trình hỗ trợ bắt lỗi chính tả. Các chương trình được xây dựng với các ưu điểm sau:

- Vấn đề lưu trữ, hiển thị và xử lý đều được xử lý với mã Unicode.
- Chương trình có giao diện đẹp, dễ sử dụng.
- Hỗ trợ đủ chức năng cần thiết cho người sử dụng.
- Xây dựng theo mô hình hướng đối tượng.

Nhưng chương trình vẫn có một hạn chế lớn: Được cài đặt bằng VB, một ngôn ngữ không hỗ trợ các cấu trúc như cây, bảng băm, map,… và đặc biệt là không hỗ trợ hướng đối tượng hoàn toàn. Do đó, chương trình bị hạn chế về tốc độ.

Do thời gian thực hiện khoá luận là ngắn nên một số vấn đề chưa giải quyết triệt để, như: phân tích văn bản, tối ưu từ điển để có tốc độ tìm kiếm cao nhất (có thể xấp xỉ 58 đơn vị thời gian), cải tiến thuật toán Earley, khả năng học,…

Giao diện chính của chương trình như sau:

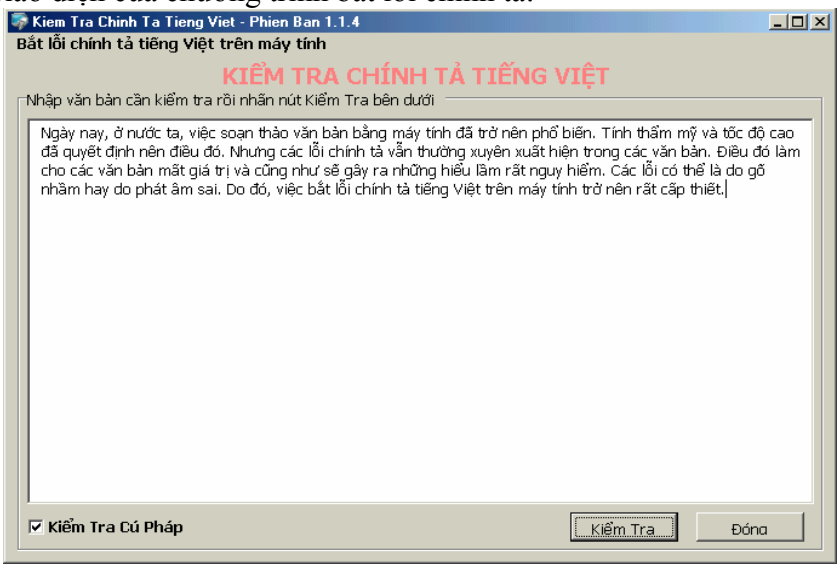


3.1.1. Chương trình bắt lỗi chính tả tiếng Việt:

Chương trình bắt lỗi chính tả được xây dựng độc lập (chưa tích hợp vào các chương trình soạn thảo văn bản). Do đó, để có thể bắt lỗi chính tả, người sử dụng phải nhập văn bản vào hoặc sử dụng chức năng Copy\Paste của hệ thống.

Chương trình hỗ trợ cả hai mức bắt lỗi chính tả là mức âm tiết và mức cú pháp. Chương trình mặc định sẽ không kiểm tra mức cú pháp cho văn bản. Nếu muốn kiểm tra mức cú pháp, người sử dụng có thể chọn tuỳ chọn Kiểm Tra Cú Pháp ở góc dưới cửa sổ chương trình.

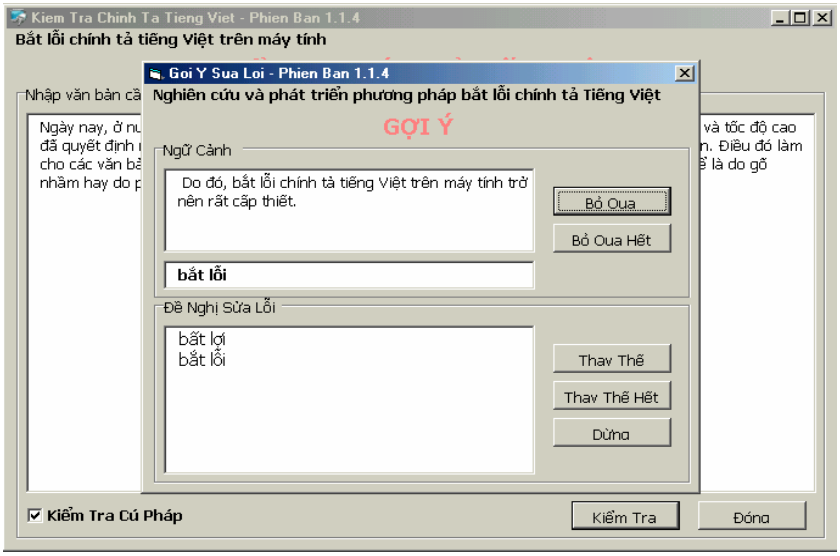
Giao diện của chương trình bắt lỗi chính tả:



Chương trình còn hỗ trợ gợi ý sửa lỗi cho các lỗi mức âm tiết và mức cú pháp. Khi gặp một lỗi, chương trình sẽ thực hiện gợi ý sửa lỗi và hiển thị cửa sổ Gợi Ý. Trên đó, người sử dụng có thể:

- Bỏ qua: Nếu cho rằng từ đó là đúng, kiểm tra từ tiếp theo.
- Bỏ qua hết: Nếu từ đó là hoàn toàn đúng và không muốn chương trình tự động bỏ qua không hỏi tới nếu gặp sau này thì chọn chức năng này.
- Thay thế: Chương trình sẽ thay thế từ sai trên văn bản bằng từ đúng trên cửa sổ Gợi Ý.
- Thay thế hết: Chương trình sẽ thực hiện thay thế tất cả các từ sai giống như vậy bằng từ đúng được chọn.
- Dừng: Chương trình sẽ dừng bắt lỗi.

Giao diện cửa sổ Gợi Ý bên trong Chương trình bắt lỗi chính tả tiếng Việt:



Chương trình sẽ không có chức năng học hay thêm từ mới. Vì đặc điểm của các từ được lưu trong từ điển là bao gồm cả từ loại, việc phân chia từ loại rất phức tạp mà bản thân người sử dụng sẽ gặp nhiều khó khăn, thậm chí là xếp từ loại sai. Do đó, chương trình sẽ không cho người sử dụng thêm từ mới vào. Chức năng này nên được đảm nhiệm bởi đội ngũ phát triển chương trình.

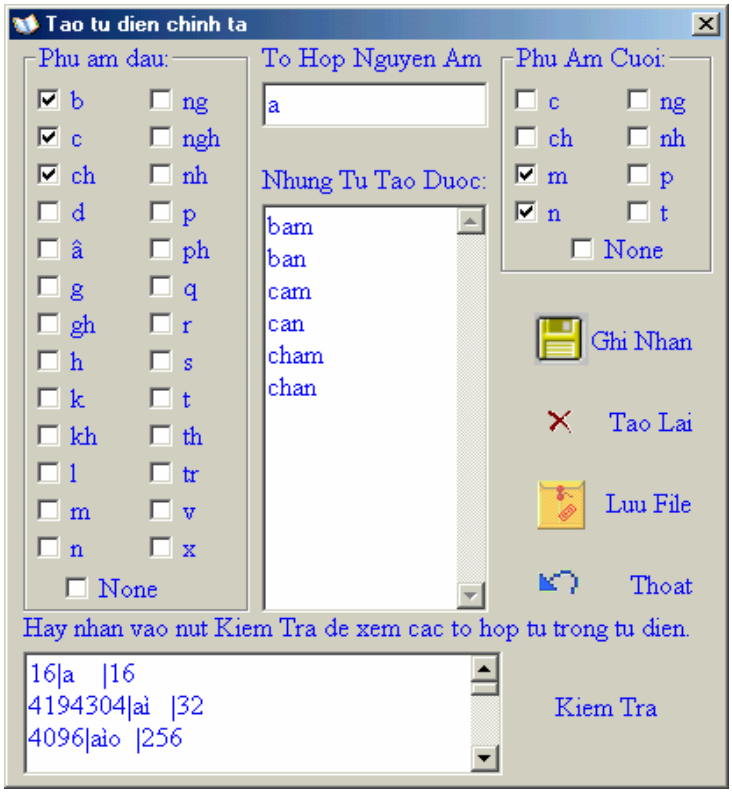
3.1.2. Chương trình bắt lỗi chính tả mức âm tiết dựa trên nguyên lý cấu tạo âm tiết:

Đây là một chương trình mô phỏng bắt lỗi mức âm tiết dựa trên nguyên lý cấu tạo âm tiết. Do không được sử dụng vào giải thuật bắt lỗi chung nên chương trình được xây dựng riêng và được đặt tên là Chương Trình Kiểm Tra Chính Tả.

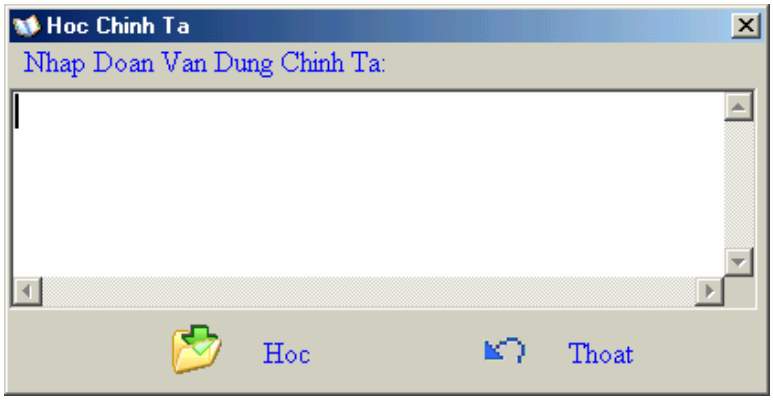
Chương trình có ba chức năng:

- Tạo từ điển chính tả: Nhập các âm tiết mới vào từ điển.
- Học chính tả: Chương trình tự động học âm tiết.
- Kiểm tra chính tả: Chương trình kiểm tra âm tiết.

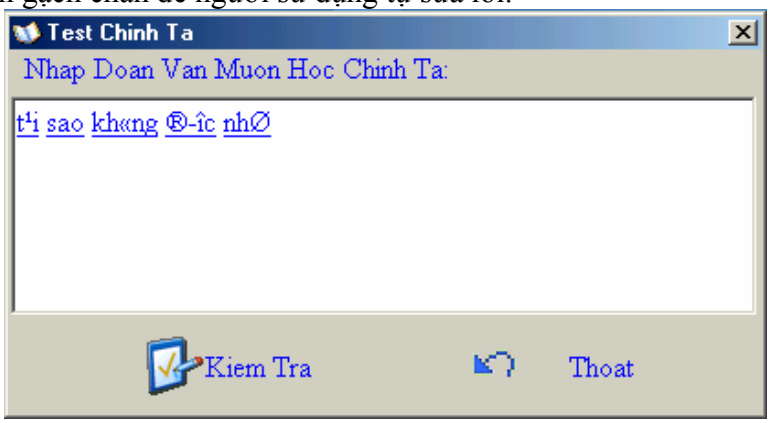
Trong chương trình Tạo từ điển chính tả, sau khi nhập xong tổ hợp các âm tiết cùng nguyên âm rồi thì có thể nhấn nút Hiển Thị để xem các âm tiết có thể tạo ra. Hay nhấn nút Nhập Mới để nhập mới tổ hợp các âm tiết khác. Sau khi xem tất cả các âm tiết có thể tạo ra, người sử dụng có thể nhấn Ghi Nhận để ghi vào mảng các âm tiết hay có thể nhấn Tạo Lại để tạo lại. Nút Lưu File là để lưu từ điển vào file lưu trữ. Nút Thoát là để quay trở lại giao diện chính. Nút Kiểm Tra là để xem tất cả các âm tiết trong từ điển.



Với chương trình Học chính tả, người sử dụng phải nhập một văn bản đúng chính tả rồi nhấn nút Học. Chương trình sẽ tự động phân tích văn bản để học các âm tiết.



Cuối cùng, với chương trình kiểm tra chính tả, người sử dụng cũng phải nhập một văn bản rồi nhấn nút Kiểm Tra. Chương trình sẽ phân tích văn bản ra từng âm tiết một và kiểm tra từng âm tiết xem có đúng không. Các âm tiết sai sẽ được chương trình gạch chân để người sử dụng tự sửa lỗi.

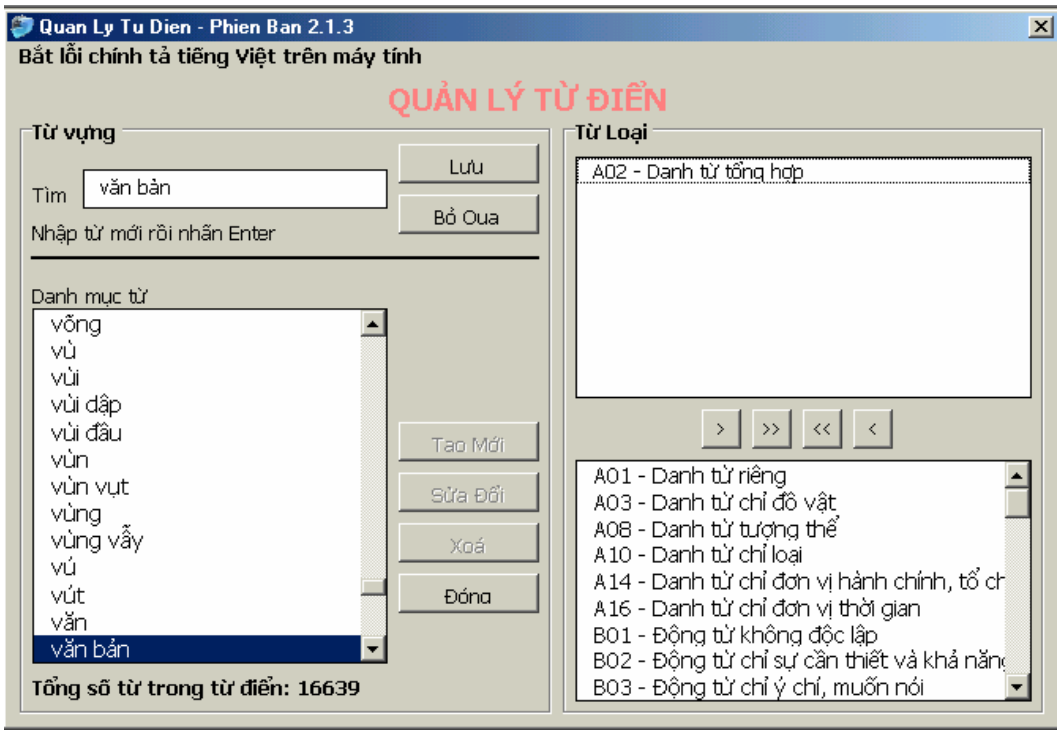


3.1.3. Các chương trình hỗ trợ và minh hoạ:

Sau đây sẽ là một số chương trình hỗ trợ cho chương trình bắt lỗi và vài chương trình minh hoạ một số giai đoạn trong quá trình phân tích cú pháp.

3.1.3.1. Quản lý từ điển:

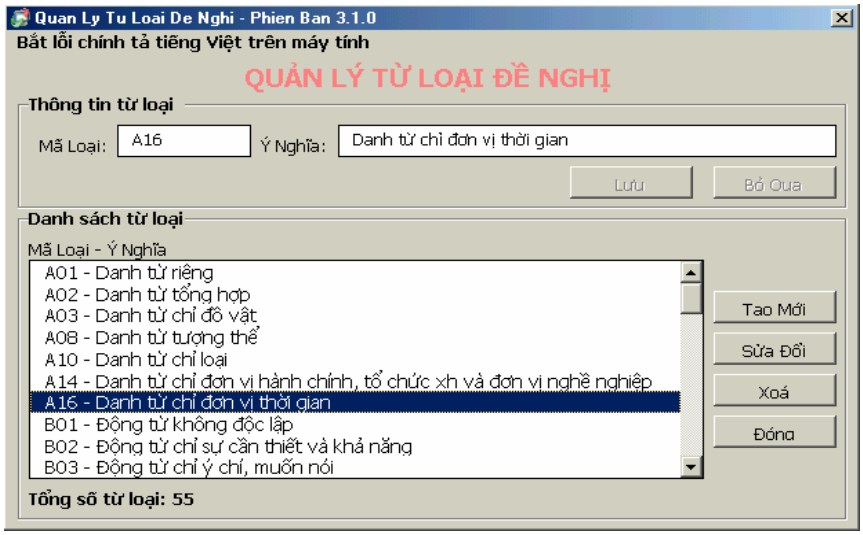
Đây là một trong những chương trình quan trọng hỗ trợ cho chương trình bắt lỗi. Chương trình cho phép thêm, bớt, sửa đổi và xoá các từ khỏi từ điển. Người sử dụng đồng thời có thể tìm kiếm các từ, xem từ loại và sửa đổi từ loại của các từ.



3.1.3.2. Quản lý từ loại:

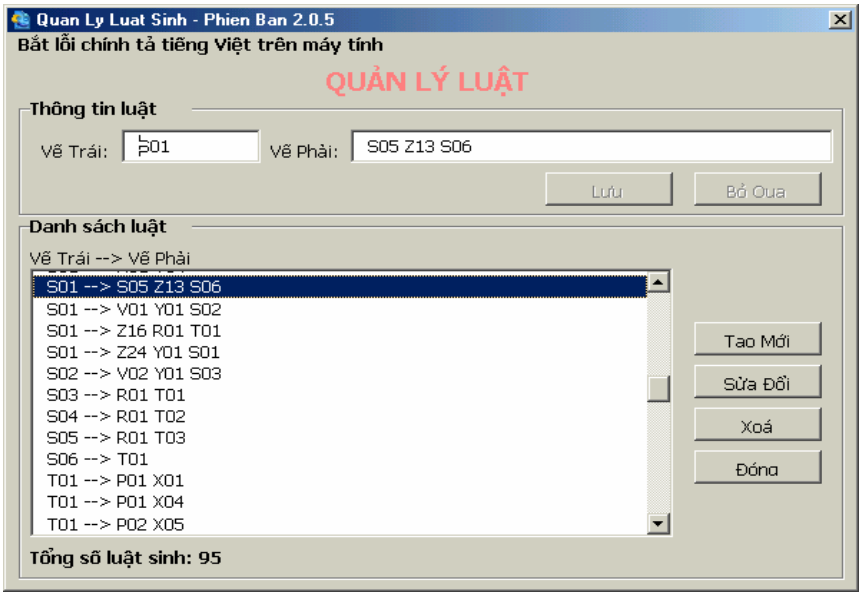
Đây là một chương trình nhỏ, hỗ trợ quản lý các từ loại được lập cho chương trình bắt lỗi. Chương trình có các chức năng sau:

- Tạo Mới: Cho phép nhập mã từ loại mới và ý nghĩa của từ loại đó.
- Sửa đổi: Cho phép sửa đổi mã và ý nghĩa của các từ loại đã có.
- Xoá: Cho phép xoá bỏ các từ loại đã có.



3.1.3.3. Quản lý luật sinh:

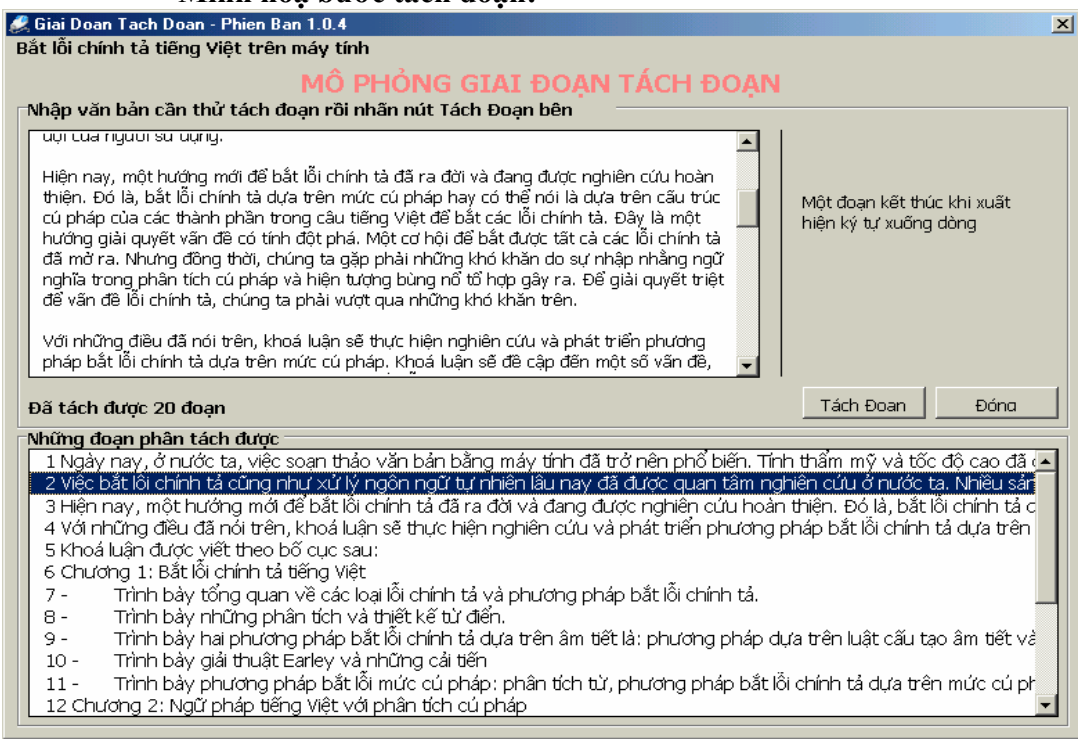
Để hỗ trợ cho phân tích cú pháp và quản lý bộ luật sinh, khoá luận đã xây dựng chương trình quản lý luật sinh. Chương trình hỗ trợ nhập luật mới, xem các luật đã có, sửa đổi các luật, xoá bỏ luật sai.



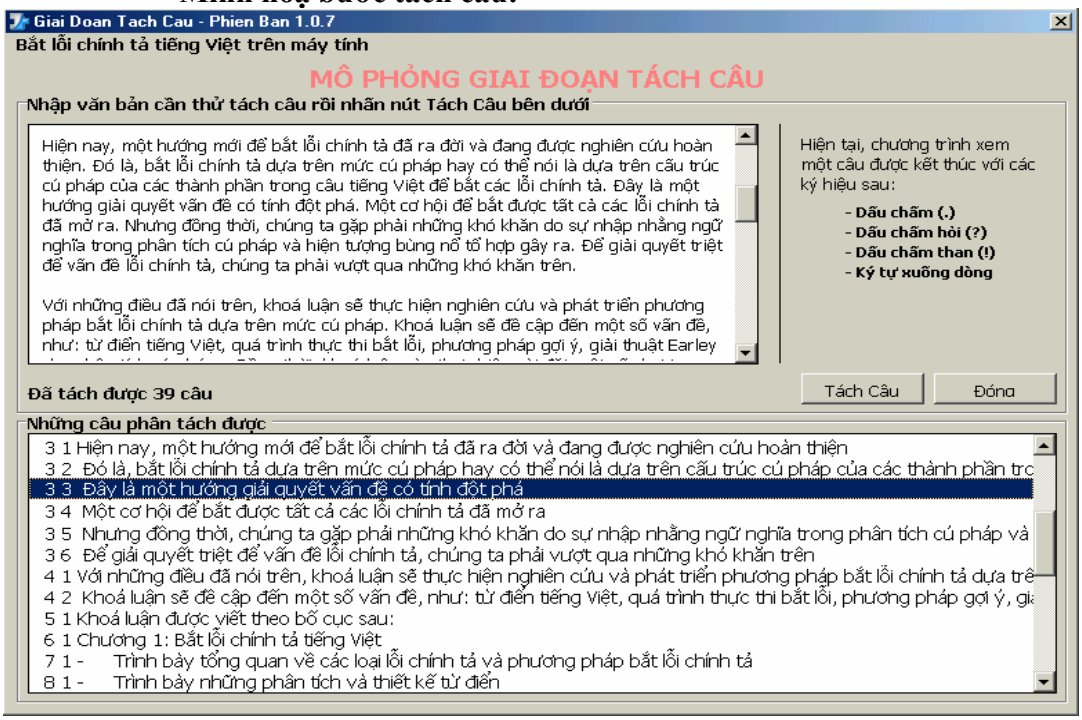
3.1.3.4. Các chương trình minh hoạ các giai đoạn phân tích cú pháp:

Để minh hoạ từng bước giai đoạn phân tích cú pháp, khoá luận xây dựng một vài chương trình minh hoạ.

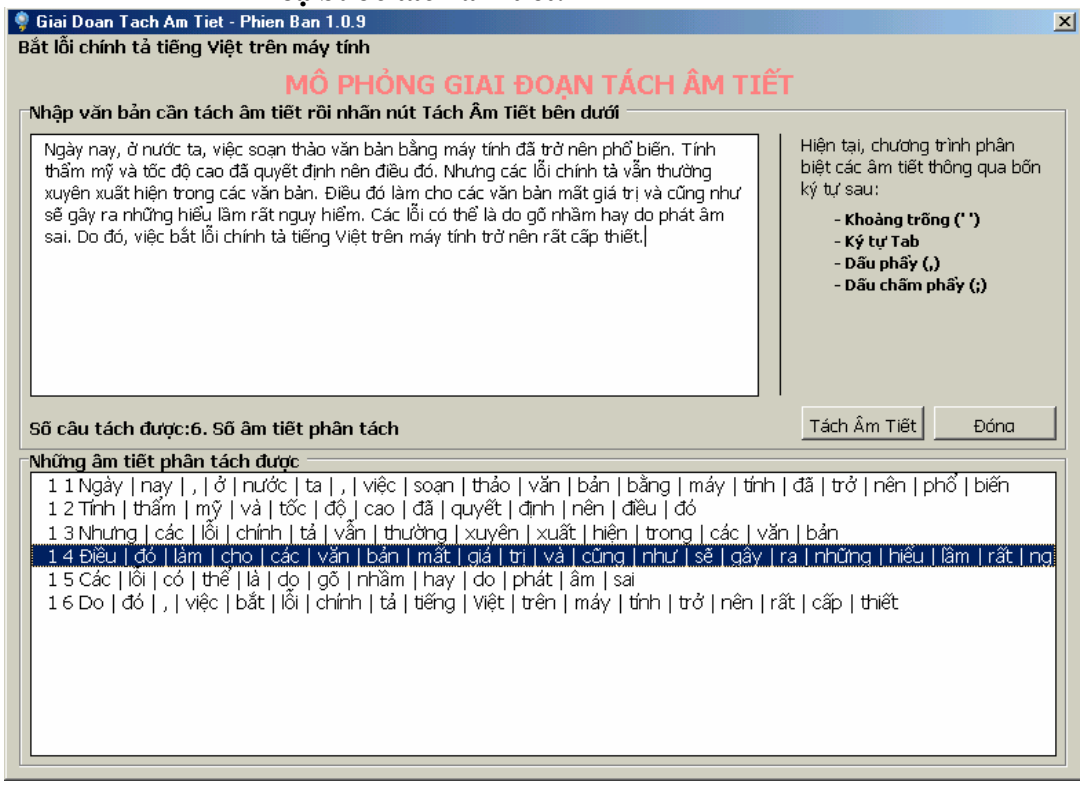
- **Minh hoạ bước tách đoạn:**



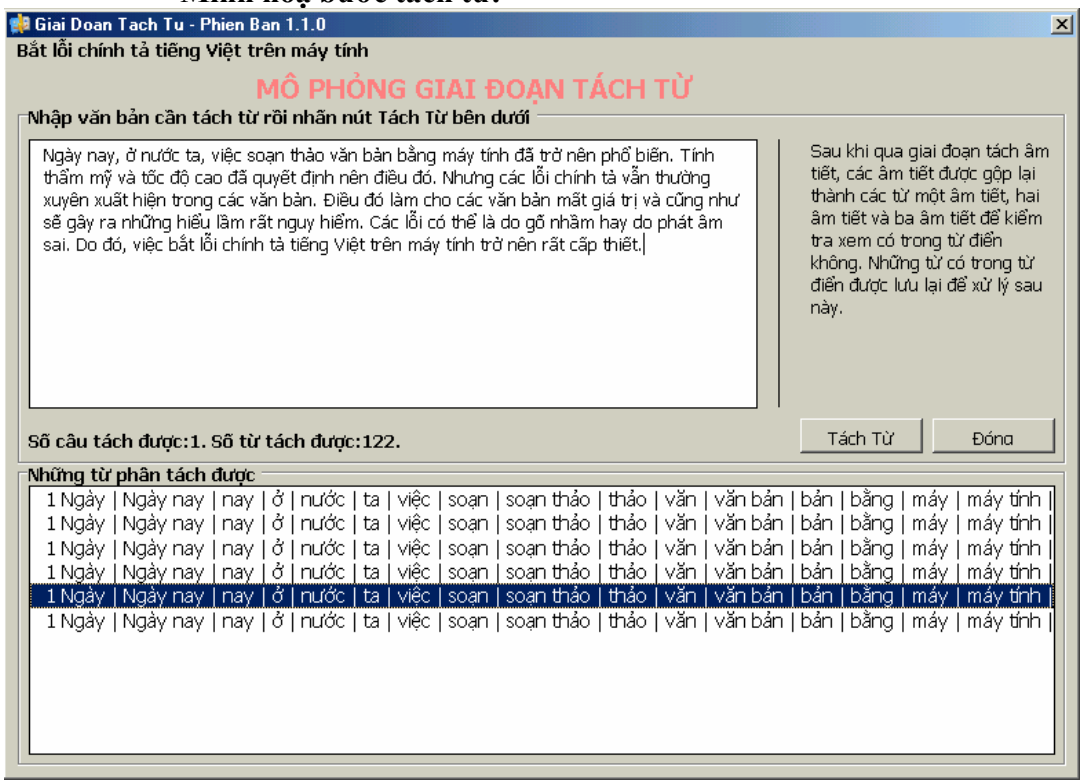
- **Minh hoạ bước tách câu:**



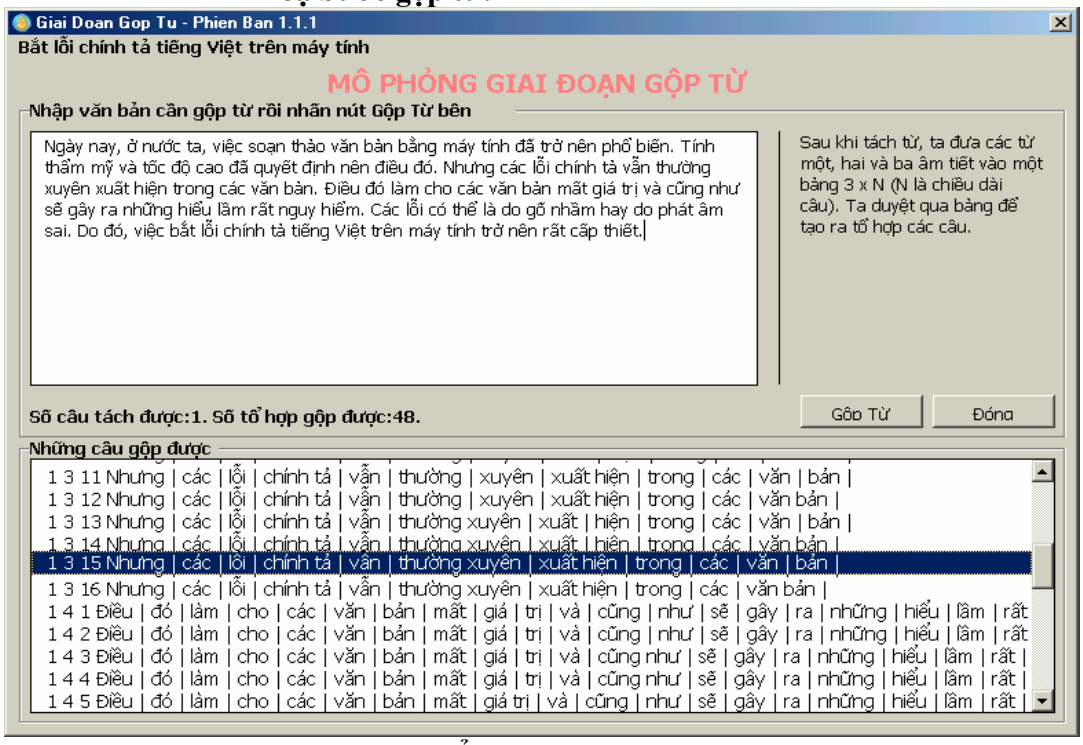
- **Minh hoạ bước tách âm tiết:**



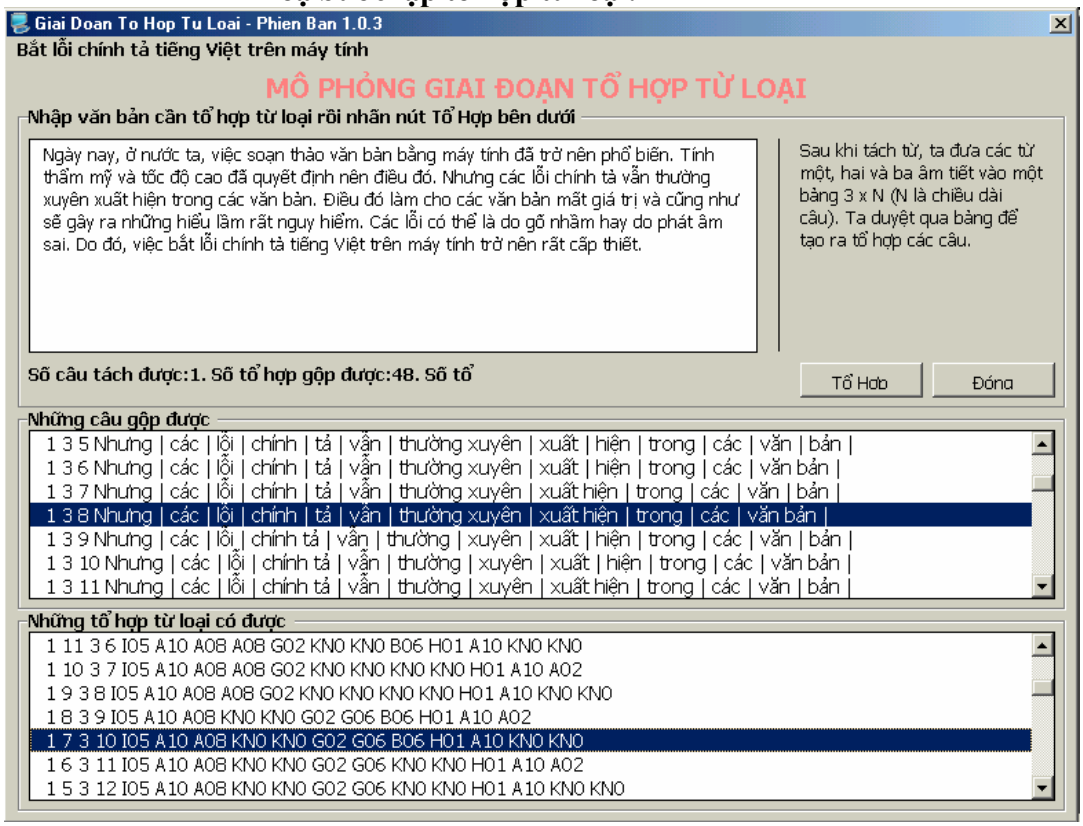
- **Minh hoạ bước tách từ:**



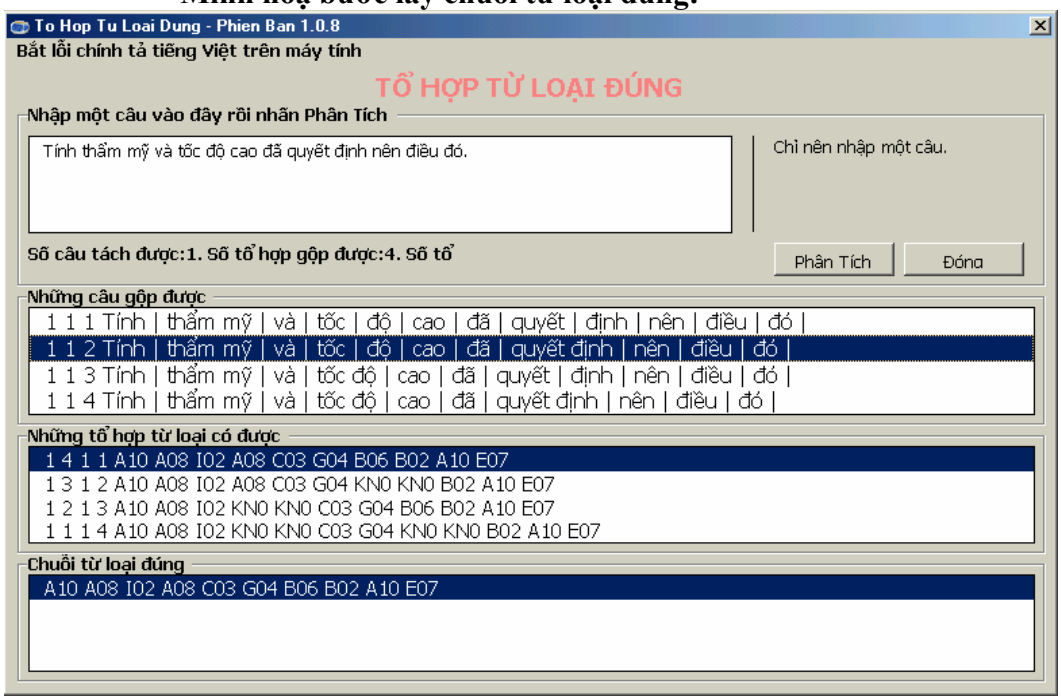
- **Minh hoạ bước gộp từ:**



- **Minh hoạ bước lập tổ hợp từ loại:**



- **Minh hoạ bước lấy chuỗi từ loại đúng:**



3.1.3.5. Chương trình học âm tiết và chương trình học từ:

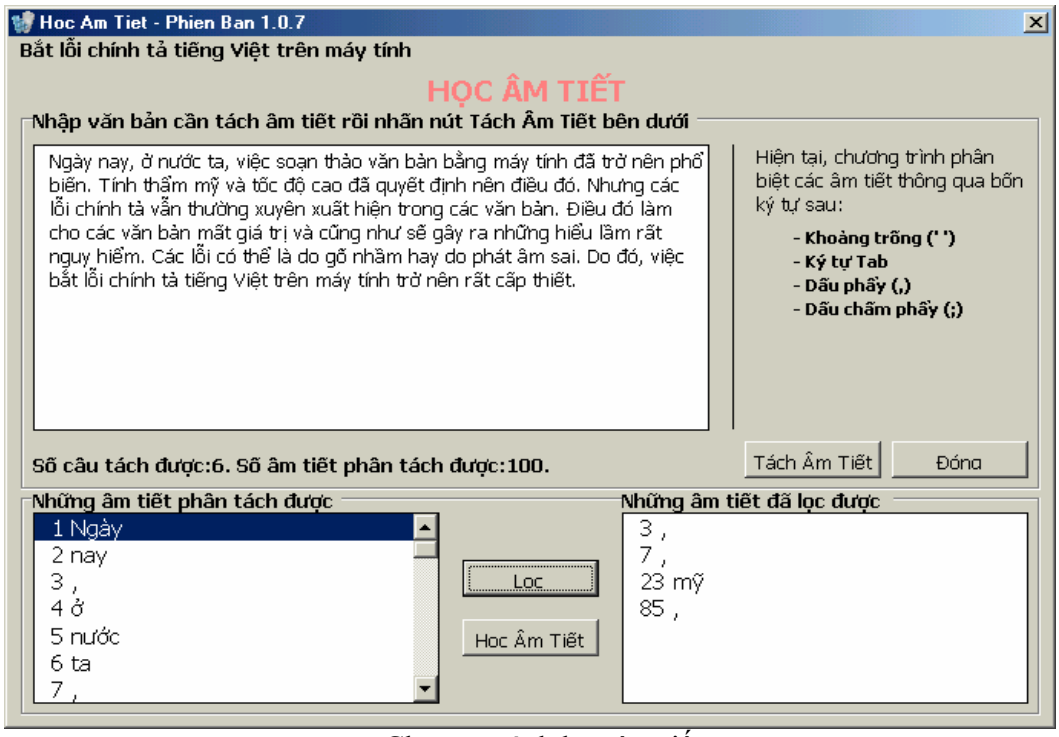
Nhằm tạo thuận tiện trong việc nhập âm tiết và từ mới vào từ điển, khoá luận cũng đã xây dựng hai chương trình tự động phân tích âm tiết và từ để người sử dụng có thể chọn các âm tiết và từ cần nhập vào từ điển.

- **Chương trình học âm tiết:**

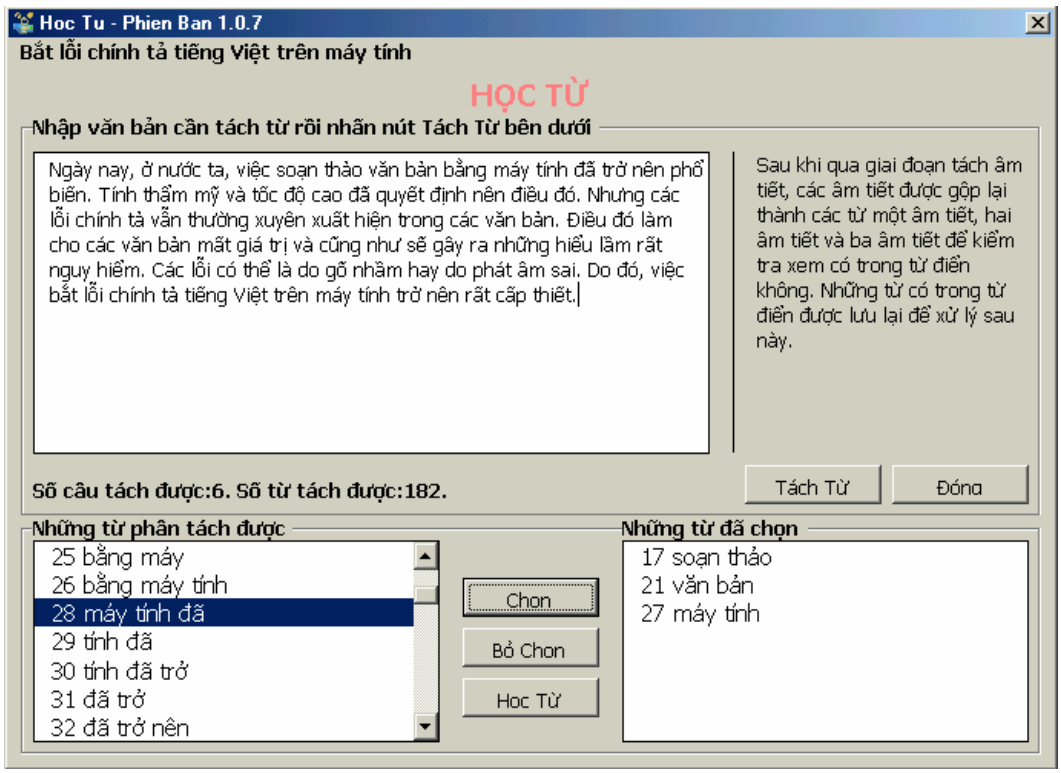
Người sử dụng nhập các âm tiết cách nhau một khoảng trống rồi nhấp nút Phân tích để chương trình tự động tách các âm tiết và liệt kê ở danh sách những âm tiết phân tách được. Đến đây, người sử dụng nhấn nút Lọc để lọc các âm tiết mới. Chương trình tự động lấy lần lượt các âm tiết đã tách được ra kiểm tra xem có trong từ điển hay chưa. Nếu âm tiết chưa có trong từ điển thì sẽ được liệt kê trong danh sách những âm tiết đã lọc được. Người sử dụng có thể nhấn nút Học Âm Tiết để nhập các âm tiết mới vào từ điển.

- **Chương trình học từ:**

Người sử dụng nhập các từ cách nhau một khoảng trống rồi nhấn nút Tách Từ. Chương trình sẽ tách các từ hai âm tiết và ba âm tiết, được liệt kê trong danh sách Những từ phân tách được. Các từ này có thể là đúng cũng có thể sai nên người sử dụng phải chọn từng từ đúng một sang danh sách Những từ đã chọn rồi mới nhấn nút Học Từ để nhập các từ được chọn vào từ điển.



Chương trình học âm tiết



Chương trình học từ

3.2. THỬ NGHIỆM:

Tiến hành phân tích đoạn đầu trong Lời nói đầu của khoá luận:

"Ngày nay, ở nước ta, việc soạn thảo văn bản bằng máy tính đã trở nên phổ biến. Tính thẩm mỹ và tốc độ cao đã quyết định nên điều đó. Nhưng các lỗi chính tả vẫn thường xuyên xuất hiện trong các văn bản. Điều đó làm cho các văn bản mất giá trị và cũng như sẽ gây ra những hiểu lầm rất nguy hiểm. Các lỗi có thể là do gõ nhầm hay do phát âm sai. Do đó, việc bắt lỗi chính tả tiếng Việt trên máy tính trở nên rất cấp thiết."

Đoạn văn trên có tất cả 75 từ với 29 từ loại. Cụ thể như sau:

1. Danh từ chỉ đơn vị thời gian: ngày
2. Danh từ chỉ đơn vị hành chính, tổ chức xã hội và đơn vị nghề nghiệp: nước
3. Đại từ chỉ thời gian: nay
4. Đại từ nhân xưng: ta
5. Giới từ: ở, bằng, trên, trong, như
6. Danh từ chỉ loại: việc, cái, tính, điều, tiếng
7. Danh từ tổng hợp: văn bản
8. Danh từ chỉ đồ vật: máy tính
9. Danh từ trừu tượng: thẩm mỹ, tốc độ, lỗi, giá trị, chính tả
10. Tính từ chỉ lượng: cao
11. Phó từ chỉ quan hệ thời gian đứng trước: đã, sẽ
12. Động từ chỉ sự bắt đầu, tiếp diễn, sự chấm dứt: trở nên
13. Tính từ chỉ quan hệ: nguy hiểm
14. Liên từ liên hợp: và
15. Động từ chỉ sự cần thiết và khả năng: nên, có thể
16. Đại từ chỉ định: đó
17. Động từ chỉ hoạt động: quyết định, gõ, phát âm, bắt, làm, soạn thảo, xuất hiện, mất, gây, hiểu
18. Tính từ chỉ phẩm chất: cấp thiết, sai, nhầm, lầm, phổ biến
19. Phó từ chỉ mức độ đứng trước: rất
20. Quán từ: những, các
21. Trợ động từ: là
22. Liên từ lựa chọn: hay
23. Liên từ kéo theo: do đó, do
24. Danh từ riêng: Việt
25. Phó từ chỉ hướng: ra
26. Phó từ chỉ sự tiếp diễn, sự tương tự: cũng, vẫn
27. Phó từ chỉ tần số: thường xuyên
28. Phó từ cho: cho
29. Liên từ tương phản: nhưng

Các luật sinh được sử dụng cho đoạn văn trên là:

<Câu> → <Liên từ> <Chủ ngữ> <Vị ngữ> nhưng…, do đó…

<Câu> → <Trạng ngữ>, <Chủ ngữ> <Vị ngữ>

<Câu> → <Chủ ngữ> <Vị ngữ>

<Câu> → <Câu> <Liên từ> <Câu>	… và …
<Câu> → <Vị ngữ>	
<Trạng ngữ> → <Trạng ngữ>, <Trạng ngữ>	
<Trạng ngữ> → <Ngữ danh từ>	ngày nay
<Trạng ngữ> → <Giới từ> <Ngữ danh từ>	ở nước ta
<Chủ ngữ> → <Danh từ chỉ loại> <Ngữ danh từ>	các lỗi chính tả
<Chủ ngữ> → <Ngữ danh từ> <Liên từ> <Ngữ danh từ>	… và …
<Chủ ngữ> → <Ngữ danh từ>	điều đó
<Ngữ danh từ> → <Danh từ> <Ngữ danh từ>	lỗi chính tả
<Ngữ danh từ> → <Danh từ chỉ loại> <Danh từ>	các văn bản
<Ngữ danh từ> → <Danh từ>	
<Ngữ danh từ> → <Động từ> <Ngữ danh từ>	soạn thảo…
<Ngữ danh từ> → <Ngữ danh từ> <Định từ>	
<Ngữ danh từ> → <Danh từ> <Tính từ>	tốc độ cao
<Ngữ danh từ> → <Quán từ> <Ngữ danh từ>	những hiểu lầm
<Ngữ danh từ> → <Danh từ> <Đại từ>	
<Định từ> → <Giới từ> <Danh từ>	bằng máy tính
<Vị ngữ> → <Ngữ động từ> <Bổ ngữ>	
<Ngữ động từ> → <Phó từ> <Ngữ động từ>	vẫn thường xuyên xuất hiện
<Ngữ động từ> → <Động từ> <Động từ>	quyết định nên, làm cho
<Ngữ động từ> → <Động từ> <Tính từ>	gõ nhầm, phát âm sai
<Ngữ động từ> → <Động từ>	trở nên
<Bổ ngữ> → <Giới từ> <Ngữ danh từ>	trong các văn bản
<Bổ ngữ> → <Tính từ>	
<Bổ ngữ> → <Ngữ danh từ>	điều đó
<Bổ ngữ> → <Ngữ danh từ> <Bổ ngữ>	
<Bổ ngữ> → <Động từ> <Ngữ danh từ>	mất giá trị
<Bổ ngữ> → <Phó từ> <Bổ ngữ>	rất nguy hiểm

Với phân tích về từ loại và luật sinh như trên, chương trình bắt lỗi sẽ kiểm tra và báo đúng cho đoạn văn trên. Những câu không phù hợp về cú pháp sẽ bị chương trình bắt lỗi và sẽ đưa ra gợi ý sửa lỗi.

KẾT LUẬN

Xử lý ngôn ngữ tự nhiên bao giờ cũng rất phức tạp và còn nhiều vấn đề mở cần tiếp tục nghiên cứu. Riêng về xử lý tiếng Việt thì phức tạp hơn vì cú pháp tiếng Việt tồn tại hiện tượng nhập nhằng (nhập nhằng cú pháp), vai trò của từ và cụm từ trong câu phụ thuộc vào ngữ cảnh rất nhiều. Vì thế, việc xác định cho được các từ và cụm từ trong câu để trợ giúp bắt lỗi chính tả không phải là đơn giản và hầu như không thể bắt lỗi với tỉ lệ 100%. Qua quá trình thực hiện, khoá luận đã đạt được một số kết quả như sau:

- Thực hiện nghiên cứu và cài đặt minh hoạ bắt lỗi chính tả tiếng Việt cả ở hai mức: mức âm tiết và mức cú pháp. Khẳng định khả năng có thể bắt lỗi lên đến 95% tỷ lệ bắt lỗi. Đồng thời, khẳng định phương pháp bắt lỗi dựa trên phân tích cú pháp là triệt để hơn.
- Thực hiện nghiên cứu ngữ pháp tiếng Việt để hỗ trợ cho phân tích cú pháp.
- Thực hiện nghiên cứu giải thuật Earley và cải tiến giải thuật cho phù hợp hơn với thực tế phân tích cú pháp. Giúp cho giải thuật Earley hoàn thiện hơn với xử lý ngôn ngữ tự nhiên.
- Từ điển được thiết kế nhỏ gọn và tối ưu trong tìm kiếm. Từ điển đã giải quyết được vấn đề lưu trữ Unicode.
- Xây dựng thuật toán phân tích từ mới hiệu quả hơn, nhưng vẫn còn hạn chế ở giá trị ngưỡng là 3.
- Xây dựng chương trình bắt lỗi chính tả tiếng Việt với phương pháp gợi ý khá hoàn chỉnh. Xây dựng một hệ thống các chương trình hỗ trợ cho chương trình bắt lỗi chính tả.

Bắt lỗi chính tả tiếng Việt là một vấn đề phức tạp. Khoá luận cũng chỉ mới là bước đầu đi vào giải quyết vấn đề mà thôi, cho nên còn nhiều hạn chế và chưa thể đáp ứng nhu cầu của người sử dụng. Trong thời gian tới, chương trình sẽ được phát triển theo hướng:

- Xây dựng chương trình phân tích văn bản ra các thành phần như: đoạn, câu, âm tiết, từ, kí hiệu, số… Chương trình có khả năng độc lập nền.
- Tối ưu hoá từ điển và tiếp tục cải tiến giải thuật Earley.
- Cải tiến nâng cao độ thông minh của phương pháp tạo gợi ý sửa lỗi.
- Tiếp tục nghiên cứu ngữ pháp tiếng Việt, đề ra các tiêu chuẩn cho phân chia từ loại, hoàn thiện hệ thống luật sinh cho câu tiếng Việt.
- Xây dựng hệ thống chương trình tiện ích hỗ trợ quá trình nghiên cứu. Phát triển khả năng học của chương trình.
- Nghiên cứu khả năng bắt lỗi mức ngữ nghĩa.

- Tích hợp vào các hệ thống soạn thảo văn bản hiện có trên thị trường.

Trong tương lai xa hơn, có thể sẽ phát triển:

- Hệ thống dịch tự động.
- Xây dựng các chương trình ứng dụng có khả năng hiểu dữ liệu nhập vào. Như trong quản lý thư viện, người sử dụng có thể nhập một từ nào đó liên quan đến sách cần tham khảo thì chương trình sẽ tự động liệt kê tất cả các sách cần thiết. Hay người sử dụng có thể nhập từ sai nhưng chương trình vẫn có thể hiểu được người sử dụng muốn gì.

TÀI LIỆU THAM KHẢO

[1]. Diệp Quang Ban, Giáo trình Ngữ pháp tiếng Việt, NXB GD 2001.

[2]. Jay Earley, *An efficient context-free parsing algorithm*. *Commun*. *ACM* 13, 2 (Feb. 1970) 94-102.

[3]. Jay Earley, *An Efficient Context-Free Parsing Algorithm*. PhD Thesis, Carnegie-Mellon University. 1968.

[4]. Lê Thanh Hương, Phân tích cú pháp tiếng Việt, Luận văn tốt nghiệp cao học, 1999.

[5]. J. Kilbury, *Earley-basierte Algorithmen für direktes Parsen mit ID/LP-Grammatiken*. KIT - Rep. 16, Institut für angewandte Informatik, TU Berlin, Berlin, June 1984.

[6]. Lê Mạnh Thạnh, Nhập môn Ngôn ngữ hình thức và Ôtômát hữu hạn, NXB GD 2000.

[7]. Phan Thị Tươi, Trình biên dịch, NXB GD 1996

[8]. Phan Thị Tươi, "Trợ giúp bắt lỗi chính tả tiếng Việt tự động bằng máy tính (giai đoạn 1)", đề tài cấp thành phố, Trường Đại học Bách Khoa TPHCM, 1998.

[9]. Phan Thị Tươi, "Trợ giúp bắt lỗi chính tả tiếng Việt tự động bằng máy tính (giai đoạn 2)", đề tài cấp thành phố, Trường Đại học Bách Khoa TPHCM, 2001.

[10]. Phan Thị Tươi, *Cải tiến một số giải thuật phân tích cú pháp trong xử lý ngôn ngữ tự nhiên*. Tạp chí Tin học và Điều khiển học, T.18, S.3 (2002), 279-284.

[11]. Như Ý, Từ điển tiếng Việt thông dụng, NXB GD 1996