# Sử dụng luật cấu tạo âm tiết tiếng Việt hai thành phần trong bài toán kiểm tra chính tả tiếng Việt

## Vietnamese Text Spelling Checking Based on Two-Ingredients Vietnamese Syllable Composition Rule

**Đinh Thị Phương Thu, Huỳnh Quyết Thắng, Nguyễn Văn Lợi**

***Abstract:*** *This article describes the approach for spelling checking of Vietnamese text. With the use of Vietnamese syllable composition rule the input text can be spelling checked. We proposed using the representation of Vietnamese syllable into two-ingredients form – first syllabe and rhyme. This presentation is more suitable and enhancing efficiency for spelling checking.*

***Keywords****: Spell checking, Vietnamese text syllable, text processing*

## I. MỞ ĐẦU

Bài toán kiểm tra chính tả tự động cho văn bản tiếng Việt đã được quan tâm nghiên cứu trong những năm gần đây, đặc biệt là với sự phát triển của CNTT cùng một khối lượng khổng lồ những văn bản điện tử. Ứng dụng của bài toán kiểm tra chính tả tự động có ý nghĩa thực tế rất lớn đối với những hệ thống xử lý văn bản và nhiều bài toán khác. Tại Việt Nam, những nghiên cứu về kiểm tra chính tả tiếng Việt hiện nay cũng đã thu được một số kết quả, tuy nhiên còn gặp nhiều khó khăn như: có nhiều chuẩn chính tả khác nhau và chưa thống nhất chung trên cả nước trong mọi lĩnh vực, có chuẩn chính tả nhưng chưa có chuẩn chính âm,... Một số phần mềm kiểm tra chính tả tiếng Việt cho văn bản điện tử cũng đã được công bố như: VietSpell, Unikey, tích hợp trong MSWord 2003,… tuy nhiên, ngoài VietSpell, hầu hết chúng đều chưa được áp dụng khả quan trong thực tế.

Để đảm bảo độ chính xác, một hệ thống kiểm tra chính tả sẽ cần kết hợp nhiều phương pháp khác nhau, tương ứng với đặc điểm của từng ngôn ngữ cũng như đảm bảo tuân theo các quy tắc chính tả đã được công nhận của bản thân ngôn ngữ đó. Trong khuôn khổ nội dung của bài báo này, chúng tôi tập trung đề cập đến một hướng tiếp cận sử dụng trong bài toán kiểm tra chính tả tiếng Việt, đó là hướng tiếp cận dựa trên luật cấu tạo âm tiết tiếng Việt để kiểm tra chính tả âm tiết tiếng Việt. Ngữ liệu về âm tiết tiếng Việt sử dụng trong bài báo được xây dựng dựa trên cụm công trình [1,2,3]. Đây là công trình đã nhận giải thưởng Nhà nước về khoa học và công nghệ năm 2005 và là kết quả nghiên cứu lâu dài, công phu và sâu sắc của của Giáo sư Hoàng Phê, một trong những chuyên gia hàng đầu nước ta về nghiên cứu ngữ nghĩa học, chính tả và từ điển tiếng Việt.

Cấu trúc của bài báo như sau: phần 2 của bài báo sẽ giới thiệu tổng quan về bài toán kiểm tra chính tả tiếng Việt; phần 3 là cái nhìn tổng quan về đặc điểm âm tiết tiếng Việt; phần 4 trình bày đề xuất hướng tiếp cận dựa trên luật cấu tạo âm tiết tiếng Việt hai thành phần, mô hình tổ chức mã hoá cấu trúc một âm tiết tiếng Việt và kiểm tra chính tả âm tiết dựa trên luật cấu tạo âm tiết hai thành phần này; cuối cùng là kết quả thử nghiệm, một số nhận xét và phần kết luận.

## II. TỔNG QUAN VỀ BÀI TOÁN KIỂM TRA CHÍNH TẢ TIẾNG VIỆT

Chính tả là sự chuẩn hoá hình thức chữ viết của ngôn ngữ. Đó là một hệ thống các qui tắc về cách viết các âm tiết, từ, các dấu câu, tên riêng, từ nước ngoài,… Quan niệm về chính tả không phải do bản thân ngôn ngữ quy định mà do xã hội quy định, và là các quy tắc được cộng đồng xã hội thừa nhận để viết.

Khác với các ngôn ngữ biến hình - ngôn ngữ mà các nội dung từ biểu hiện ngay ở mức từ khi biến đổi hình thái từ như các ngôn ngữ Châu Âu (tiếng Anh,

Pháp,...) - là chính tả ở mức "từ " thì chính tả tiếng Việt – ngôn ngữ đơn lập (nội dung của từ chỉ mang tính từ vựng) - lại là chính tả ở mức "âm tiết". Vì thế trong khi bước đầu tiên của bài toán kiểm tra chính tả cho các ngôn ngữ biến hình là kiểm tra chính tả "từ" thì với tiếng Việt sẽ phải tiến hành thêm một pha kiểm tra chính tả mức "âm tiết" ở trước pha kiểm tra mức "từ" này. Hay nói các khác mô hình tổng quan kiểm tra chính tả cho tiếng Việt sẽ bao gồm 3 pha (hình 3):

– Pha kiểm tra chính tả ở mức "âm tiết"

– Pha kiểm tra chính tả ở mức "từ"

– Pha kiểm tra chính tả ở mức "câu" (mức ngữ pháp)

Nội dung của bài báo này chỉ tập trung trình bày giải pháp luật cấu tạo âm tiết hai thành phần đề xuất cho pha kiểm tra chính tả ở mức âm tiết. Một số kết quả nghiên cứu về mức ngữ pháp có thể tham khảo trong [20].

## 1. Một số qui định về chuẩn hoá chính tả tiếng Việt

Một trong những yếu tố để giảm bớt khó khăn của công việc kiểm tra chính tả tiếng Việt là yêu cầu chúng ta phải có một chuẩn chính tả tiếng Việt thống nhất chung. Vì vậy trong những năm qua, để kịp thời thống nhất chính tả trong cải cách giáo dục và từng bước khắc phục tình trạng viết lộn xộn hiện nay, trên cơ sở những ý kiến thảo luận qua các hội nghị về chính tả và thuật ngữ, Bộ Giáo dục và Văn phòng chính phủ đã ban hành một số quy định mang tính pháp qui về chính tả tiếng Việt [5,6,7,8,9,10]. Muốn viết đúng chính tả tiếng Việt chúng ta phải tuân theo những qui định, qui tắc đã được xác lập này.

## 2. Các nguyên nhân gây ra lỗi chính tả

Có nhiều nguyên nhân khác nhau gây ra lỗi chính tả, tuy vậy có thể tổng hợp lại một số nguyên nhân như sau [14,15]:

– *Nguyên nhân do nhập liệu sai*: Lỗi này có thể do gõ sai/ thiếu/ thừa phím gây ra, do cách cài đặt bàn phím, loại bàn phím, do quy tắc gõ tiếng Việt của các kiểu gõ khác nhau (Telex, VNI, TCVN, Unicode,..)…

– *Nguyên nhân do phát âm sai*: Lỗi này do sự nhầm lẫn giữa cách đọc và cách viết của những từ đồng âm hoặc âm gần với nhau dẫn đến viết sai (như lỗi dấu hỏi/ngã, lối sai âm đầu s/x, tr/ch, r/d/gi/v...). Với tiếng Việt, do có nhiều khác biệt về cách phát âm giữa các vùng trong khi hệ thống chữ viết lại dựa trên hệ thống phát âm của thủ đô Hà Nội nên dễ dẫn đến các lỗi sai loại này.

– *Nguyên nhân do sử dụng từ vựng sai*: Lỗi này do khi sử dụng từ sai với ý nghĩa thực của nó. Đây là lỗi về vốn từ vựng của người sử dụng, nhưng nhiều khi vẫn đòi hỏi trình bắt lỗi chính tả phải tìm ra những lỗi này.

– *Các nguyên nhân khác*: Ngoài ra còn các loại lỗi chính tả khác như viết hoa, viết tên riêng, thuật ngữ, tên tiếng nước ngoài không đúng qui cách,…

## 3. Phân loại lỗi chính tả

Có nhiều cách phân loại lỗi chính tả theo các tiêu chí khác nhau. Ta có thể phân loại theo nguồn gốc sinh ra lỗi như ở trên. Nếu xét theo quan điểm của chương trình bắt lỗi chính tả ở mức từ thì lỗi chính tả có thể được phân làm hai loại là lỗi *non-word* và lỗi *real-word* :

– Lỗi *non-word* là lỗi tạo ra từ sai, nghĩa là từ đó hoàn toàn không có trong từ điển từ vựng tiếng Việt hoặc một số ngữ liệu đầu vào cho quá trình tiền xử lý văn bản như : từ điển tên riêng, từ điển viết tắt, từ điển vay mượn,... Đây là loại lỗi dễ phát hiện.

– Lỗi *real-word* là lỗi chính tả mà từ đó có trong từ điển nhưng sử dụng từ sai. Nếu không dựa vào ngữ cảnh xung quanh thì không thể xác định được đó có phải là lỗi chính tả hay không. Đây là loại lỗi rất khó phát hiện và xử lý.

## 4. Đánh giá tính hiệu quả của một trình bắt lỗi chính tả

Trình bắt lỗi chính tả có thể được đánh giá theo nhiều cách khác nhau. Nhưng chủ yếu vẫn dựa trên quan điểm của người dùng về khả năng phát hiện lỗi sai và khả năng đề nghị những từ thay thế cho lỗi sai đó. Có thể phân ra làm hai loại nhầm lẫn mà một trình chính tả thường gây ra: nhầm lẫn tích cực và nhầm lẫn tiêu cực.

– *Nhầm lẫn tích cực* xảy ra khi trình bắt lỗi chính tả

báo lỗi ở những từ hoàn toàn không sai chính tả.

– *Nhầm lẫn tiêu cực* xảy ra khi trình bắt lỗi chính tả bỏ qua những từ bị sai chính tả. Nhầm lẫn tiêu cực có thể xem là lỗi không phát hiện được. Phần nhiều những lỗi này đòi hỏi phải "hiểu" văn bản (ít nhất là một phần trong văn bản) để có thể phát hiện lỗi. Những dạng lỗi sử dụng sai nghĩa từ vựng và lỗi cú pháp thường rơi vào dạng này.

Trong hai loại nhầm lẫn thì nhầm lẫn tích cực thường gây khó chịu cho người sử dụng, dễ gây tâm lý không tin tưởng vào trình bắt lỗi chính tả. Ngược lại, nhầm lẫn tiêu cực phản ánh tính hiệu quả của trình bắt lỗi chính tả. Nhầm lẫn tiêu cực càng ít thì tính hiệu quả của trình bắt lỗi chính tả càng cao.

## III. ÂM TIẾT TIẾNG VIỆT

Về mặt ngữ âm, chuỗi lời nói của con người phát ra gồm nhiều khúc, đoạn dài ngắn khác nhau. Đơn vị phát âm ngắn nhất là âm tiết (*syllable*). Dù nói chậm đến đâu chăng nữa thì chúng ta cũng không thể tách được thành các đơn vị nhỏ hơn âm tiết [4]. Về mặt chính tả, âm tiết tiếng Việt được thể hiện là một chuỗi các kí tự liên tiếp và được giới hạn đầu và cuối chuỗi bằng kí tự cách hoặc các kí tự kết thúc câu như: ".", "!", "?",...

Âm tiết tiếng Việt có một số đặc điểm như: có tính độc lập cao, có cấu trúc chặt chẽ, số lượng âm tiết là hữu hạn,… Tính chất âm tiết của tiếng Việt đưa đến nhiều hệ quả quan trọng về ngữ âm cũng như về ngữ pháp. Trong khuôn khổ nội dung bài toán kiểm tra chính tả tự động cho các văn bản trên máy tính, chúng tôi sẽ chỉ đi vào các đặc điểm của âm tiết tiếng Việt xét trên phương diện chính tả.

Về mặt cấu trúc, âm tiết tiếng Việt có cấu trúc chặc chẽ. Quan điểm hiện nay của các nhà ngôn ngữ học Việt nam về cấu trúc âm tiết chia thành hai cấp độ:

– Cấp độ thứ nhất (hình 1): Âm tiết gồm âm đầu, khuôn vần và thanh điệu, ví dụ *"bạn" = [b] + <an> + [thanh nặng]*. Tiếng Việt có 25 âm đầu (kể cả âm đầu zero), các âm đầu này được ghi bằng các kí tự hoặc tổ hợp các kí tự. Thanh điệu tiếng Việt xét trong ngữ âm thì có 6 thanh điệu (*thanh huyền, thanh sắc, thanh hỏi, thanh ngã, thanh nặng, thanh bằng*) nhưng xét về mặt chính tả thì gồm 5 thanh điệu (*thanh huyền, thanh sắc, thanh hỏi, thanh ngã, thanh nặng*).

– Cấp độ thứ hai (hình 2): Âm tiết gồm âm đầu, âm đệm, âm chính, âm cuối và thanh điệu, ví dụ: *"chính" = [ch] + [ ] + <i> + [nh] + [thanh sắc]*

| [Thanh điệu] | |
|---|---|
| [Âm đầu] | <Khuôn vần> |

*Hình 1. Cấu trúc 3 thành phần của âm tiết tiếng Việt*

| [Thanh điệu] | | | |
|---|---|---|---|
| [Âm đầu] | Vần | | |
| | [Âm đệm] | <Âm chính> | [Âm cuối] |

*Hình 2. Cấu trúc 5 thành phần của âm tiết tiếng Việt*

*Ghi chú: Những thành phần nằm trong cặp dấu "<>" là bắt buộc phải có, những thành phần nằm trong cặp dấu " [ ]" thì có thể có hoặc không.*

Giữa các thành phần của âm tiết tiếng Việt cũng có những luật quan hệ ràng buộc lẫn nhau, ví dụ: độ dài của một âm tiết không quá 7 kí tự, âm đầu "*ngh*" chỉ đi với "*i*", "*e*", ..

## IV. SỬ DỤNG LUẬT CẤU TẠO ÂM TIẾT TIẾNG VIỆT HAI THÀNH PHẦN TRONG BÀI TOÁN KIỂM TRA CHÍNH TẢ

### 1. Đề xuất tiếp cận âm tiết tiếng Việt theo cấu trúc hai thành phần

Cấu trúc một âm tiết tiếng Việt gồm 3 hoặc 5 thành phần và giữa các thành phần này có các luật quan hệ phụ thuộc lẫn nhau [1,2,3]. Như vậy trong pha đầu tiên của bài toán kiểm tra chính tả tiếng Việt - pha kiểm tra ở mức âm tiết, với một âm tiết đầu vào ta hoàn toàn có thể tiến hành phân tích cấu trúc âm tiết theo các luật quan hệ thành phần, từ đó sẽ có thể kết luận âm tiết đó có sai chính tả hay không. Và chắc chắn việc sử dụng tập các luật quan hệ hữu hạn này so với sử dụng một từ điển âm tiết để kiểm tra sẽ tiết kiệm không gian lưu trữ từ điển hiệu quả hơn. Tuy nhiên việc tổng hợp các luật quan hệ 3 hoặc 5 thành phần trên lại không hề đơn giản, đòi hỏi chúng ta phải đảm bảo chắc chắn là đã tập hợp đầy đủ các luật tương ứng

với các trường hợp cụ thể, và dựa trên một cơ sở ngữ liệu đã được kiểm chứng. Tiếng Việt chúng ta hiện nay chưa có được một tập ngữ liệu đầy đủ như vậy cho âm tiết.

Sử dụng luật cấu trúc âm tiết tiếng Việt dựa trên thành phần là một cách tiếp cận rất tốt trong việc phân tích ngôn ngữ tiếng Việt nhưng cách tiếp cận theo quan điểm 3 và 5 thành phần hiện nay lại gây nhiều khó khăn trong việc kiểm tra lỗi chính tả trên máy tính. Vì thế, từ những khó khăn gặp phải trên và qua nghiên cứu các đặc điểm của âm tiết tiếng Việt dựa trên [1], chúng tôi đề xuất một hướng tiếp cận sử dụng luật cấu tạo âm tiết chỉ gồm **hai thành phần** ngắn gọn hơn cho bài toán kiểm tra chính tả. Cấu trúc hai thành phần này có thể tập hợp được tập luật cấu tạo để dễ dàng cho việc kiểm tra chính tả trên máy tính và nâng cao hiệu quả xử lý cho bài toán. Theo hướng tiếp cận đề xuất, âm tiết sẽ có cấu trúc bao gồm hai thành phần như sau:

**Âm tiết = [âm đầu] + <vần>**

Trong đó:

– *Âm đầu* là các phụ âm được ghi thể hiện bằng kí tự và tổ hợp các kí tự. Âm đầu gồm 25 phụ âm sau: **r, d, gi, v, ch, tr, s, x, l, n, qu, b, c/k, đ, g/gh, h, kh, m, ng/ngh, nh, p, ph, t, th** và **'-'** (phụ âm zero trong trường hợp âm tiết không có phụ âm đầu).

– *Vần* là sự kết hợp của khuôn vần cuối và thanh điệu như trong cấu trúc âm tiết 3 thành phần, ví dụ : *ăn, ậy,…* Âm tiết chính tả tiếng Việt có 126 khuôn vần cuối và 5 thanh điệu *(hỏi, sắc, huyền, ngã, nặng) [1,2,3]*.

## 2. Mô hình kiểm tra chính tả tiếng Việt mức âm tiết

Pha kiểm tra chính tả tiếng Việt ở mức "âm tiết" thực hiện kiểm tra từng âm tiết của một văn bản đầu vào có đúng chính tả không. Từ hướng tiếp cận trên, chúng tôi tiến hành mã hóa lưu trữ 6760 âm tiết chữ viết có nghĩa trong tiếng Việt [1] theo luật cấu trúc hai thành phần sau đó sử dụng các cấu trúc lưu trữ mã hóa này để tiến hành kiểm tra chính tả âm tiết cho văn bản đầu vào.

Mô hình kiểm tra chính tả mức âm tiết của chúng tôi (hình 3) bao gồm 4 bước:

***Bước 1.*** *Tổ chức mã hoá lưu trữ 6760 âm tiết chữ viết có nghĩa trong tiếng Việt theo luật cấu trúc hai thành phần đề xuất ở trên.*

***Bước 2.*** *Kiểm tra từng âm tiết của văn bản đầu vào để xác định âm tiết đó có sai chính tả hay không.*

***Bước 3.*** *Nếu âm tiết đó là sai chính tả, tiến hành tạo danh sách các âm tiết gợi ý thay thế cho âm tiết sai này.*

***Bước 4.*** *Sắp xếp danh sách âm tiết gợi ý vừa tạo ra theo thứ tự dựa vào khả năng có thể thay thế từ cao hơn đến thấp hơn để hỗ trợ người sử dụng lựa chọn thay thế.*

Kết thúc pha kiểm tra chính tả mức âm tiết, văn bản đầu vào sẽ gồm các âm tiết có nghĩa trong tiếng Việt. Kết quả này giúp tăng tính chính xác cao hơn rất nhiều cho pha kiểm tra chính tả mức từ ở pha sau (hình 3).

Trong phần tiếp theo chúng tôi sẽ đi vào mô tả chi tiết nội dung công việc thực hiện và các giải pháp xử lý đối với âm tiết tiếng Việt cho từng bước.

## 3. Giải pháp thực hiện kiểm tra chính tả tiếng Việt mức âm tiết

*a) Bước 1 - Tổ chức mã hoá âm tiết tiếng Việt áp dụng hướng tiếp cận cấu trúc hai thành phần*

Dựa trên [1,3], các âm tiết chữ viết có nghĩa trong tiếng Việt được chúng tôi tiến hành mã hóa lưu trữ trên hai tập dữ liệu:

– Tập thứ nhất lưu trữ từ điển 6760 âm tiết chữ viết mà tiếng Việt hiện đại sử dụng theo cấu trúc âm tiết hai thành phần (*syllable set*).

– Tập thứ hai lưu trữ các tập nhầm lẫn âm tiết chữ viết tương ứng cho các âm tiết có vấn đề chính tả (*confusion set*), ví dụ: *mỉ, mỷ, mĩ, mỹ, mị, mỵ,...* . Bên cạnh các tập nhầm lẫn được xây dựng chủ yếu là lỗi do phát âm (sai thành phần thứ nhất - 'phụ âm đầu' hoặc thành phần thứ hai - 'vần') dựa trên [1,3], chúng tôi còn kết hợp sử dụng luật cấu tạo âm tiết hai thành phần trên để tạo thêm tập nhầm lẫn cho các lỗi do nhập liệu sai.

Đối với tập ngữ liệu lưu trữ 6760 âm tiết chữ viết theo cấu trúc hai thành phần, chúng tôi sử dụng cấu

trúc cây tam phân để lưu trữ, mã hóa. Cây tam phân là một cây gồm một nút gốc và dưới mỗi nút trong cây sẽ có tối đa 3 nút khác. Giải thuật tìm kiếm trên cây tam phân (*Ternary Search Tree*) là một giải thuật về tìm kiếm từ điển đã được kiểm nghiệm và đánh giá là có tốc độ nhanh hơn so với cách xây dựng từ điển theo tổ chức tệp băm và cây nhị phân [19]. Quá trình duyệt và xây dựng cây tam phân được thực hiện như sau. Bắt đầu từ đầu xâu đến cuối xâu đang xét và duyệt từ nút gốc của cây, nếu kí tự đầu tiên của xâu nhỏ hơn giá trị tại nút gốc thì ta sẽ đi xuống nút bên trái của nút gốc, nếu nút này chưa có thì thêm vào. Tương tự nếu kí tự đang xét mà lớn hơn hoặc bằng kí tự tại nút gốc thì ta sẽ đi xuống nút bên phải hoặc nút giữa của nút gốc, nếu nút này chưa có thì thêm vào. Tiếp tục xét tiếp kí tự thứ hai với nút này và cứ như thế cho đến khi hết xâu. Tại nút mà xâu kết thúc ta sẽ gán một giá trị cho nút đó, ví dụ ID của xâu trong từ điển. Quá trình xây dựng cây xong khi tất cả các nút đều được đưa vào cây. Hình 4 minh họa một cây tam phân được xây dựng cho từ điển gồm các xâu *hàn, hành, hoàng, hùng, hang.*

Tập 6760 âm tiết chữ viết được tổ chức thành các cây tam phân với nút gốc của cây tương ứng một phụ âm đầu trong tiếng Việt (bao gồm cả phụ âm zero). Việc truy cập để kiểm tra một âm tiết đúng chính tả được thực hiện bằng cách xác định phụ âm đầu của âm tiết rồi truy cập cây tam phân tương ứng với phụ âm đầu đó. Giá trị được gán cho mỗi nút trong cây tam phân bao gồm một cặp số nguyên [*value1, value2*] để biểu diễn đặc điểm của âm tiết đó.

Các giá trị của *value1*:

*0: Không thể có cấu tạo âm tiết trong tiếng Việt hoặc không thể có dạng chính tả của âm tiết.*

*1: Có âm tiết sử dụng trong tiếng Việt.*

*2: Có yếu tố cấu tạo từ Hán-Việt trong tiếng Việt.*

Các giá trị của *value2*:

*0: Khi value1= 0 thì value2= 0.*

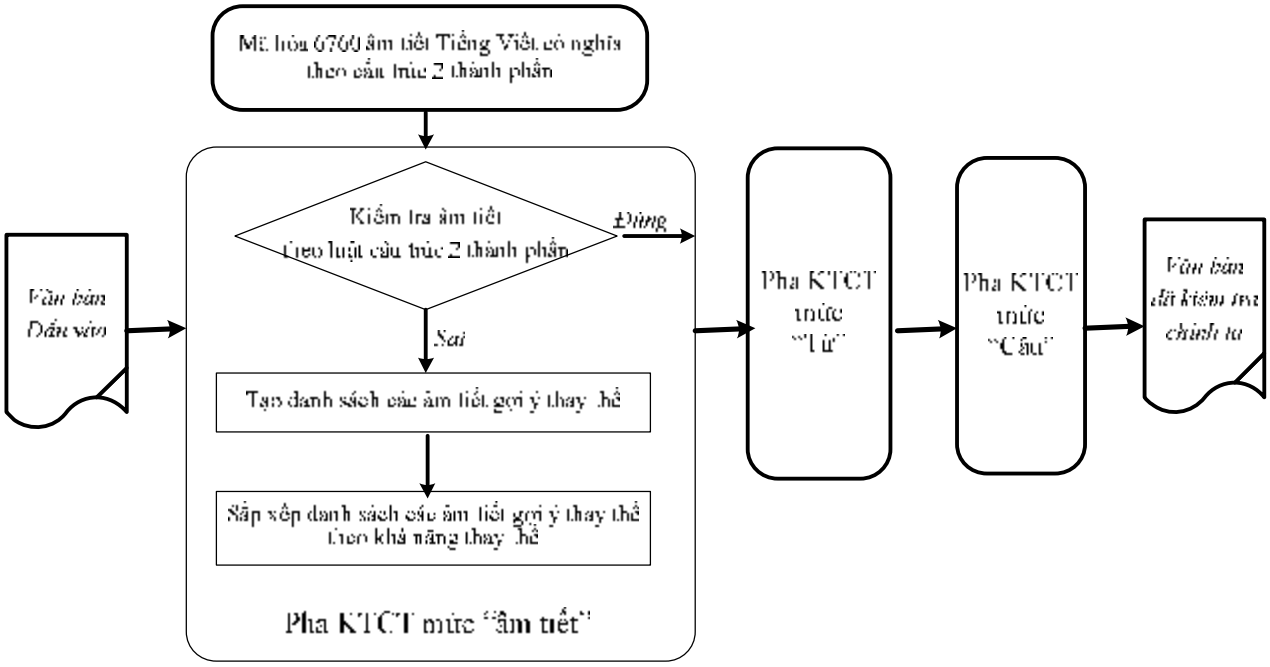
*1: Âm tiết có vấn đề chính tả.*

*2: Âm tiết không có vấn đề chính tả.*

*3: Biến thể ngữ âm-chính tả của yếu tố Hán-Việt.*

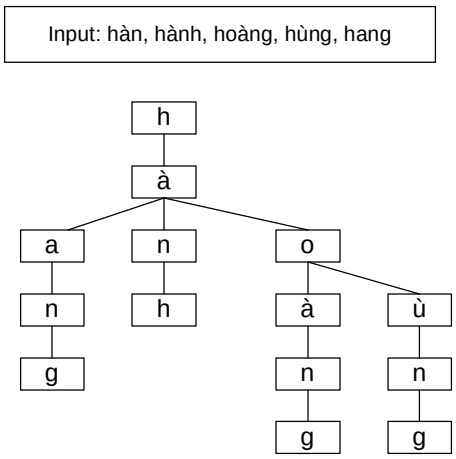
*4: Âm tiết chữ viết ghi cùng âm tiết ngữ âm đã có.*

Một số giá trị của *value1, value2* dù không cần dùng




*Hình 3. Mô hình kiểm tra chính tả tiếng Việt mức âm tiết*

đến trong mô hình thực hiện nhưng chúng tôi vẫn lưu để đảm bảo tính đầy đủ của ngữ liệu trong [1] và có thể sẽ phục vụ cho những mục đích khác trong tương lai.




*Hình 4. Minh họa một cây tam phân*

Đối với tập ngữ liệu thứ hai, tập lưu trữ các tập nhầm lẫn âm tiết chữ viết tương ứng cho các âm tiết có vấn đề chính tả, chúng tôi sử dụng bảng băm để lưu trữ. Cấu trúc của tệp "*Confusion Set*" này được tổ chức như sau:

- Mỗi dòng của tệp sẽ lưu một tập nhầm lẫn âm tiết, mỗi âm tiết cách nhau một dấu phẩy.
- Mỗi âm tiết chỉ thuộc về một tập nhầm lẫn duy nhất.

Hai tập ngữ liệu trên sau khi tiến hành xây dựng sẽ được đưa vào sử dụng để kiểm tra chính tả cho các âm tiết của văn bản đầu vào.

*b) Bước 2 - Kiểm tra, xác định âm tiết sai chính tả*

Quá trình kiểm tra chính tả âm tiết được thực hiện từ đầu văn bản (sau khi văn bản đã được tiền xử lý để loại nhiễu). Với mỗi âm tiết thực hiện kiểm tra kiểm tra xem âm tiết đó đúng chính tả không dựa trên tập ngữ liệu thứ nhất đã lưu trữ ở trên. Nếu xác định là âm tiết sai chính tả, chuyển sang bước 3.

*c) Bước 3 - Tạo danh sách các âm tiết gợi ý thay thế cho âm tiết sai chính tả*

Khi đã xác định một âm tiết là sai chính tả, trình kiểm tra chính tả cần thông báo cho người dùng trên giao diện tương tác và đưa ra một danh sách các âm tiết gợi ý thay thế cho âm tiết bị sai chính tả đó. Việc này được chúng tôi thực hiện bằng cách duyệt trong tập nhầm lẫn của âm tiết bị sai chính tả (tập ngữ liệu thứ hai), chọn ra các âm tiết có nghĩa và đưa vào danh sách gợi ý thay thế. Sau đó danh sách các âm tiết gợi ý thay thế này sẽ được chuyển sang bước 4 để thực hiện sắp xếp theo thứ tự ưu tiên thay thế.

Để tạo ra tập nhầm lẫn âm tiết này, chúng tôi tiến hành xây dựng dựa trên các nguyên nhân gây ra lỗi nhầm lẫn. Hầu hết các lỗi chính tả thường gặp là lỗi do nhập liệu và lỗi do phát âm [11,14]. Còn đối với một số lỗi chính tả ít gặp thì chúng ta chấp nhận có thể sẽ có những âm tiết mà chương trình báo lỗi nhưng không đưa ra gợi ý thay thế nào, ví dụ: *nguyyyyyyyên*,... Nguyên nhân là do nếu chúng ta quan tâm đến cả những lỗi chính tả ít gặp này thì kích thước tập nhầm lẫn sẽ tăng lên rất nhiều, đồng thời gây giảm độ chính xác về mức độ ưu tiên của từ gợi ý thay thế cũng như gây giảm tốc độ và hiệu năng của chương trình.

Tập nhầm lẫn âm tiết cho tiếng Việt này được chúng tôi xây dựng như sau:

1. Tạo danh sách các âm tiết gợi ý thay thế dựa trên lỗi phát âm: Lỗi phát âm gây ra do cách phát âm của một số vùng khác với các quy tắc chính tả đã quy định và do có một số âm vị, vần, âm tiết có cách đọc gần giống nhau [11]:

- Lỗi thanh điệu: Chủ yếu do nhầm lẫn giữa thanh *ngã/hỏi, ngã/nặng*.
- Lỗi về âm đầu: Thường do lẫn lộn trong một số nhóm âm sau: *c/k, g/gh, ng/ngh, ch/tr, s/x, v/d/gi/r*.
- Lỗi về âm chính. Thường lẫn lộn âm chính trong vần sau: *ai/ay/ây, ao/au/âu, ăm/âm, ăp/âp, iu/iêu/êu, im/iêm/em, ip/iêp/êp/ep, oi/ôi/ơi, om/ôm/ơm, op/ôp/ơp, ong/ông, oc/ôc, ui/uôi, um/uôm, up/uôp, ưi/ươi, ưu/ươu, ưm/ươm, ưp/ươp*.
- Lỗi về âm cuối. Thường lẫn lộn các âm cuối trong các vần sau: *an/ang, at/ac, ăn/ăng, ăt/ăc, ân/âng, ât/âc, en/eng, et/eng, et/ec, ên/ênh, êt/êch, in/inh, it/ich, iên/iêng, iêt/iêc, ơn/ơng, ơt/ơc, un/ung, ut/uc, uôn/uông, uôt/uôc, ưn/ưng, ưt/ưc, ươn/ương, ươt/ươc*.

2. Tạo danh sách các âm tiết gợi ý thay thế dựa trên lỗi

nhập liệu. Các lỗi nhập liệu chính bao gồm:

- Lỗi không tuân theo chuẩn quy định pháp quy mà trình chính tả đã lựa chọn, ví dụ : bỏ dấu vào nguyên âm chính khi âm tiết có hai nguyên âm như *tóan (toán)*,.. [8]
- Lỗi do gõ thừa/ thiếu phím 'cách ' (*spacebar*) gây tách/ghép âm tiết, ví dụ: "*nguy ên*", "*sinhviên*",…
- Lỗi do gõ sai thứ tự giữa một phím và phím 'cách' nên một kí tự thuộc âm tiết này bị đẩy sang âm tiết khác: "*sinhv iên*"
- Lỗi do gõ thêm/ thiếu một kí tự, ví dụ: "*ngguyên*" (nguyên)
- Lỗi do đảo ngược 2 kí tự gần nhau trong âm tiết, ví dụ: "*gnuyên*" (nguyên)
- Lỗi do nhầm kí tự này cho kí tự khác, ví dụ: "*hguyên*" (nguyên)
- Lỗi do cách cài đặt bàn phím, loại bàn phím, do quy tắc gõ tiếng Việt của các kiểu gõ khác nhau (Telex, VNI, TCVN, Unicode,..)…

Để sửa các lỗi nhập liệu này ta cần thực hiện ngược lại quá trình gây ra lỗi (*error reversal*). Đối với lỗi gõ thiếu/ thừa phím ta có thể thêm hoặc bớt đi một phím để tạo ra âm tiết mới. Đối với lỗi gõ sai thứ tự phím, ta duyệt âm tiết, lần lượt hoán vị 2 kí tự liên tiếp để tạo ra âm tiết mới. Đối với lỗi gõ nhầm phím, ta dựa vào bố trí của bàn phím để phát sinh ngược lại từ âm tiết lỗi,..

*d) Bước 4 - Sắp xếp danh sách các âm tiết gợi ý*

Sau khi đã tạo được một danh sách các âm tiết gợi ý thay thế cho âm tiết sai chính tả, tiếp theo ta sẽ sắp xếp danh sách này theo thứ tự khả năng có thể thay thế. Nói đơn giản nghĩa là âm tiết nào nhiều khả năng đúng hơn sẽ được sắp lên trước, âm tiết nào ít có khả năng đúng sẽ được sắp đứng sau. Việc này nhằm mục đích tăng tính tiện dụng cho người dùng khi lựa chọn các ứng viên thay thế để sửa âm tiết sai chính tả.

Có nhiều cách sắp xếp danh sách âm tiết gợi ý khác nhau, tuy vậy tổng quan lại là tổng hợp của 2 phương pháp dựa trên tần xuất xuất hiện của âm tiết và dựa vào ngữ cảnh. Ở đây, để sắp xếp danh sách âm tiết gợi ý ta cũng dựa vào cả 2 yếu tố trên. Để thực hiện ta đã có từ điển âm tiết, từ điển từ vựng và dữ liệu thống kê tần suất xuất hiện của từng âm tiết, từng từ trong ngữ liệu huấn luyện, ta tiến hành sắp xếp danh sách như sau:

- Với mỗi âm tiết trong danh sách gợi ý, lần lượt ghép với âm tiết đứng ngay sau và ngay trước nó. Nếu tạo thành một từ/ thành phần của một từ trong từ điển từ vụng ta sẽ xếp âm tiết đó lên phần trước, các âm tiết không ghép được sẽ bị xếp ra phần sau.
- Các âm tiết bị xếp ở phần sau: tiếp tục được sắp thứ tự dựa trên tần suất xuất hiện của âm tiết từ cao đến thấp.
- Các âm tiết ở phần trước: sẽ được xếp theo thứ tự tần suất xuất hiện của từ (mà có một thành phần của từ là âm tiết đang xét ghép được với âm tiết trước hay sau thành) theo thứ tự từ cao đến thấp

Cuối cùng ta sẽ thu được một danh sách các âm tiết gợi ý được sắp xếp theo thứ tự hợp lý hơn cho người dùng.

## V. CÀI ĐẶT THỬ NGHIỆM HƯỚNG TIẾP CẬN

### 1. Cài đặt thử nghiệm

Tiến hành cài đặt thử nghiệm mô hình kiểm tra chính tả ở mức âm tiết theo hướng tiếp cận luật âm tiết hai thành phần đề xuất trên, chúng tôi thu được một kết quả thực nghiệm khả quan. Dữ liệu thử nghiệm ban đầu được chúng tôi tiến hành trên các bản tin báo điện tử do tính chất ngôn ngữ ngắn gọn, ít phép ẩn dụ, sử dụng các từ vựng thông dụng... Hướng mở rộng trong tương lai sẽ tiến hành trên nhiều dữ liệu đa dạng hơn về thể loại văn bản và ở kích thước lớn hơn.

Chúng tôi sử dụng tập văn bản thử nghiệm *DataSet1* gồm 10 văn bản với số lỗi chính tả ngẫu nhiên đã được xác định, có tất cả hơn 300 lỗi chính tả gồm cả lỗi chính tả ở mức âm tiết và mức từ. Tập văn bản huấn luyện ngữ liệu *DataTraining1* bao gồm 800 văn bản khác nhau. Kích thước bài báo lớn nhất là 52KB, kích thước bài báo nhỏ nhất là 1KB, kích thước trung bình của một bài báo là 5KB. Kết quả thử nghiệm kiểm tra chính tả mức âm tiết thu được mô tả trong bảng 1.

*Nhận xét:*

– Độ chính xác trung bình trên tập 10 văn bản thử nghiệm khoảng 94% (bảng 1). Kết quả này phụ thuộc vào ngữ liệu đầu vào cho quá trình tiền xử lý văn bản như: từ điển tên riêng, từ điển viết tắt, từ điển vay mượn, khối lượng văn bản học cho hệ thống,... Quá trình tiền xử lý văn bản cần sử dụng các ngữ liệu này để phân loại âm tiết và lọc nhiễu trước khi thực hiện kiểm tra chính tả âm tiết. Với ngữ liệu đầu vào đầy đủ hiệu suất chương trình sẽ được nâng cao hơn.

– Các lỗi chính tả âm tiết không phát hiện được đều là các trường hợp âm tiết có trong từ điển âm tiết nhưng về mặt ngữ nghĩa thì không phù hợp với ngữ cảnh. Đây có thể xem như là lỗi chính tả ở mức từ - sự tổ hợp của các âm tiết đúng nhưng không phải là từ có nghĩa (*non-word*) hoặc không sử dụng đúng cho ngữ cảnh của câu (*real word*), ví dụ: *tự sử* (tự sự), '*Già gạo ngày càng tăng cao*' (Giá gạo ngày càng tăng cao),... . Lỗi chính tả loại này rất khó phát hiện và sẽ được thực hiện trong pha kiểm tra chính tả mức từ của mô hình. Một trong các giải pháp giúp nâng cao hiệu quả là bằng các dữ liệu học và huấn luyện cho hệ thống.

– Đối với các lỗi chính tả âm tiết sai do phát âm thì khả năng đề nghị sửa lỗi cao hơn so với các lỗi chính tả âm tiết sai do nhập liệu. Nguyên nhân do tập nhầm lẫn xây dựng cho lỗi chính tả phát âm được xây dựng đầy đủ theo luật cấu trúc âm tiết hai thành phần đề xuất dựa vào các dữ liệu thống kê của

*Bảng 1. Bảng kết quả thử nghiệm tập Dataset1*

| Văn bản | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Tổng hợp |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Thể loại | Tin tức khoa học | Tin tức khoa học | Tin tức kinh doanh | Tin tức kinh doanh | Tin tức giáo dục | Tin tức giáo dục | Tin tức pháp luật | Tin tức pháp luật | Tin tức thể thao | Tin tức thể thao | |
| Số lỗi chính tả âm tiết | 13 | 22 | 23 | 26 | 25 | 20 | 23 | 27 | 17 | 23 | 219 |
| Lỗi phát hiện đúng - gợi ý sửa đúng | 13 | 22 | 22 | 23 | 23 | 17 | 20 | 25 | 14 | 22 | 202 |
| Lỗi phát hiện đúng - chưa gợi ý sửa đúng | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 4 |
| Lỗi không phát hiện được | 0 | 0 | 0 | 2 | 1 | 2 | 3 | 2 | 3 | 1 | 13 |
| Tỉ lệ chính xác (%) | 100 | 100 | 100 | 92 | 96 | 90 | 87 | 93 | 83 | 96 | 94 |

*Bảng 2. Một số lỗi chính tả âm tiết hướng tiếp cận phát hiện được nhưng Vietspell không phát hiện được*

| TT | Câu thử nghiệm | VietSpell |
|---|---|---|
| 1 | Tại ĐH Sư phạm Hà Nội có chương trình đào tạo ráo viên chất lượng ***cap*** cũng với đầu vào tuyển chọn trong số thí sinh thi vào trường và được hưởng nhiều ưu đãi. | Không phát hiện được |
| 2 | Theo nhận định của toà án, Sở Y tế hoàn toàn có ***quỳn*** yêu cầu ông Thành thực hiện đúng hợp đòng hoặc huỷ bỏ hợp đồng; | Không phát hiện được |
| 3 | khẳng định họ sẽ không điều tra về chận hoà 2-2 của Thụy Điển và Đan Mạch sau khi có một xố nguồn ***ting*** cho rằng | Không phát hiện được |
| 4 | lớp trất lượng cao ở tất cả các trường, hkoa thành ciên, mỗi ***nghành*** một lớp. | Không phát hiện được |
| 5 | Đối với hẹ cử nhân tài năng, trỉ tuyển các ngành khoa học tự nhiên, ***đối tượngbao*** gồm những học sinh đã đoạt dải học sinh giỏi quốc gia, | Không phát hiện được |
| 6 | Ở Tây Ban Nha, mạng máy ***atính*** của các quan chức tòa án tối cao, | Không phát hiện được |
| 7 | Tại Brussels (Bỉ), dịch vụ liên lạc và báo chí cũng như văn phòng của một số cơ quan EU đều bị ảnh ***hưởn***. | Không phát hiện được |
| 8 | Gần đây nhất chúng tôi đã tham gia đấu giá và mua được cổ phần của Garmex Saigon và Công ty thương ***main*** dịch vụ ... | Không phát hiện được |

GS.Hoàng Phê [1], trong khi tập nhầm lẫn xây dựng cho lỗi chính tả nhập liệu hiện nay mới quan tâm đến một số lỗi phổ biến và tiêu biểu.

**2. So sánh đánh giá**

So sánh đánh giá khả năng phát hiện lỗi chính tả âm tiết với công cụ VietSpell 3.0 [18]: Khả năng phát hiện và đề nghị sửa lỗi *âm tiết* cao hơn so với công cụ VietSpell 3.0. Hầu hết trong các trường hợp phát hiệsn lỗi, chương trình đều đưa được ra âm tiết gợi ý đúng nằm trong 10 âm tiết đầu tiên của danh sách âm tiết gợi ý thay thế.

Bảng 2 mô tả một số ví dụ về lỗi chính tả âm tiết mà hướng tiếp cận phát hiện được nhưng Vietspell không phát hiện được (các lỗi gạch chân). Các lỗi chính tả khác trong bảng như: *'ráo viên'* (giáo viên), *'ưu đải'* (ưu đãi),… là các lỗi chính tả ở mức từ (*non-word* hoặc *real word*) nên sẽ không được xét ở đây. Pha kiểm tra chính tả mức từ ở mức 2 trong mô hình hệ thống sẽ thực hiện tiếp nhiệm vụ này.

## VI. KẾT LUẬN

Tiếp cận bài toán kiểm tra chính tả tiếng Việt ở mức âm tiết theo hướng sử dụng luật cấu trúc âm tiết hai thành phần đã giúp tiết kiệm được không gian lưu trữ từ điển, tốc độ truy cập và xử lý nhanh, góp phần tăng độ chính xác, giảm nhập nhằng cho kiểm tra chính tả mức từ ở pha sau của bài toán. Ngữ liệu sử dụng trong nghiên cứu được xây dựng dựa trên một công trình nghiên cứu toàn diện, có hệ thống, tập trung vào vấn đề chính tả tiếng Việt để đảm bảo tính chính xác cho các nghiên cứu và thử nghiệm.

Chúng tôi mong muốn sẽ kết hợp những kết quả khả quan này với những giải pháp khác [20] để mang lại hiệu quả cao hơn cho bước nghiên cứu tiếp theo ở hai pha sau của bài toán kiểm tra chính tả tiếng Việt, từ đó hướng tới một trong những mục tiêu mà chúng tôi kì vọng là tích hợp các kết quả nghiên cứu này vào các trình soạn thảo dùng tiếng Việt, như MS Office, OpenOffice hay các hệ thống tìm kiếm và trích chọn thông tin (SearchEngine),… để phục vụ cộng đồng.

## TÀI LIỆU THAM KHẢO


[1] HOÀNG PHÊ, *Từ điển chính tả*, Nhà xuất bản Đà nẵng - Trung tâm Từ điển học, 2006.

[2] HOÀNG PHÊ, *Chính tả tiếng Việt*, Nhà xuất bản Đà nẵng - Trung tâm Từ điển học, 1999.

[3] HOÀNG PHÊ, *Từ điển tiếng Việt*, Nhà xuất bản Khoa học Xã hội – Trung tâm từ điển học Hà Nội, 2006

[4] ĐOÀN THIỆN THUẬT, *Ngữ âm tiếng Việt*, NXB Đại học và Trung học chuyên nghiệp Hà Nội, 2003.

[5] *'Một số quy định về chính tả trong sách giáo khoa cải cách giáo dục'*, Bộ Giáo dục, 1980.

[6] *'Quy định về chính tả tiếng Việt và thuật ngữ tiếng Việt'*, Bộ Giáo dục, 1984.

[7] *'Quy định tạm thời về viết hoa trong văn bản của Chính phủ và của Văn phòng Chính phủ'*, Quyết định số 09/1998/QĐ-VPCP ngày 22/11/1998. http://www.luutruvn.gov.vn/

[8] *'Quy tắc chính tả tiếng Việt và phiên chuyển tiếng nước ngoài'*, Hội đồng Quốc gia chỉ đạo biên soạn Từ điển Bách khoa Việt Nam, 2000. http://www.bachkhoatoanthu.gov.vn

[9] *'Quy định tạm thời về chính tả trong sách giáo khoa mới'*, Bộ Giáo dục và Đào tạo, Nhà XBGD, 3/2002.

[10] *'Quy định tạm thời về viết hoa tên riêng trong sách giáo khoa'*, Bộ Giáo dục và Đào tạo, 2003.

[11] LÊ TRUNG HOA, *'Lỗi chính tả và cách khắc phục'*, Nhà xuất bản khoa học xã hội, 2002.

[12] KAREN KUKICH, *'Techniques for Automatically Correcting Words in Text'*, ACM Computing Surveys, Volume 24 , Issue 4 (December 1992) Pages: 377 – 439, 1992

[13] F. J. DAMERAU, *'A technique for computer detection and correction of spelling errors'*, Communications of the ACM, Volume 7 , Issue 3 (March 1964), Pages: 171 - 176

[14] CHAO-HUANG CHANG, *'A new approach for automatic chinese spelling correction'*, 1998. http://casper.beckman.uiuc.edu/~ctsai4/chinese/wordseg/chang.ps.

[15] SEBASTIAN DEOROWICZ and MARCIN G.CIURA, *'Correcting spelling errors by modeling their causes'*, International Journal of Applied Mathematics and Computer Science, 2005; 15(2):275–285

[16] ANDREW R. GOLDING and YVES SCHABES, *'Combining Trigram-based and Feature-based Methods for Context-sensitive spelling correction'*, http://citeseer.ist.psu.edu/golding96combining.html

[17] ANDREW R. GOLDING and DAN ROTH, *'A Winnow-based approach to Context-sensitive spelling correction'*, 1999. http:/citeseer.ist.psu.edu/golding99winnowbased.html

[18] http://www.vspell.com/

[19] JON BENTLEY, ROBERT SEDGEWICK, *'Fast Algorithms for Sorting and Searching Strings'***,** 1997. http://www.- cs.princeton.edu/~rs/strings/

[20] ĐINH THỊ PHƯƠNG THU, HOÀNG VĨNH SƠN, HUỲNH QUYẾT THẮNG. *Cải tiến giải thuật CYK cho bài toán phân tích cú pháp tiếng Việt*. Tạp chí tin học và điều khiển học, Tập 22, Số 4, 2006, Trang 325-338

**Ngày nhận bài : 2/12/06**

## SƠ LƯỢC TÁC GIẢ

**ĐINH THỊ PHƯƠNG THU**

Sinh ngày 31/01/1980.

Tốt nghiệp đại học năm 2002 và cao học Công nghệ thông tin năm 2004, ĐH Bách khoa Hà Nội. Hiện đang là Nghiên cứu sinh chuyển tiếp tại Khoa Công nghệ thông tin, ĐH Bách khoa Hà Nội.

Đang công tác tại công ty DICOM Việt Nam.

Lĩnh vực nghiên cứu quan tâm hiện nay: Xử lý văn bản tiếng Việt, Công nghệ phần mềm.

Email: thudtp@yahoo.com.

**HUỲNH QUYẾT THẮNG**

Sinh năm 1967.

Tốt nghiệp đại học năm 1990 tại Trường ĐH Điện-Máy Varna, CH Bulgaria. Bảo vệ luận văn Tiến sỹ tại CH Bulgaria năm 1995.

Hiện đang công tác tại Khoa Công nghệ thông tin, ĐH Bách khoa Hà Nội.

Các lĩnh vực quan tâm: Công nghệ phần mềm, Xử lý văn bản tiếng Việt, Đồ họa và đa phương tiện.

Email: thanghq@it-hut.edu.vn

**NGUYỄN VĂN LỢI**

Sinh năm 1947.

Bảo vệ luận văn Tiến sỹ chuyên ngành ngôn ngữ học tại Viện nghiên cứu Đông Phương - Cộng hòa Liên bang Nga (Liên xô cũ). Được phong chức danh Phó Giáo sư năm 1991, chức danh Giáo sư năm 1996.

Hiện đang công tác tại Viện Ngôn ngữ, Viện Khoa học xã hội Việt Nam.

Lĩnh vực quan tâm: Ngữ âm học, Lịch sử các ngôn ngữ ở Việt Nam, Xử lý ngôn ngữ tiếng Việt.

Email: vanloi201@yahoo.com