VIEGRID

KỶ NIỆM 1000 NĂM THĂNG LONG
HÀ NỘI

ITI

VIEGRID JSC

VIỆN CNTT, ĐHQGHN

"Hãy cùng chúng tôi giữ gìn tiếng Việt"

BÁO CÁO VỀ TÌNH HÌNH CHÍNH TẢ TRONG VĂN BẢN TIẾNG VIỆT

Đợt đánh giá Tháng 6/2010

NGUYỄN ÁI VIỆT
NGUYỄN TẤN TÔN THẤT ĐỖ VŨ

HÀ NỘI 2010

"Hãy cùng chúng tôi giữ gìn tiếng Việt"

BÁO CÁO VỀ TÌNH HÌNH CHÍNH TẢ TRONG VĂN BẢN TIẾNG VIỆT

Đợt đánh giá Tháng 6/2010

NGUYỄN ÁI VIỆT VÀ NGUYỄN TẤN TÔN THẤT ĐỖ VŨ

HÀ NỘI 2010

Các tác giả

Nguyễn Ái Việt, Phó Viện trưởng Viện Công nghệ thông tin, Đại học Quốc gia Hà nội.

Nguyễn Tấn Tôn Thất Đỗ Vũ, Phó Giám đốc Trung tâm Nghiên cứu Phát triển Công nghệ GRID tại Huế, Công ty VIEGRID JSC.

Khuyến nghị của các tác giả

Báo cáo này vừa là báo cáo thống kê cho các mục đích xã hội và truyền thông vừa có những nội dung nghiên cứu chuyên ngành. Độc giả có thể bỏ qua những nội dung không quan tâm mà không ảnh hưởng tới việc theo dõi nội dung của báo cáo.

Các số liệu, nội dung trong báo cáo dựa trên hiểu biết và cố gắng cao nhất của các tác giả. Độc giả sử dụng các kết quả công bố trong báo cáo phải tự chịu trách nhiệm về việc áp dụng các kết quả này. Yêu cầu các độc giả trích rõ nguồn trong các bài viết của mình khi sử dụng kết quả của báo cáo.

Cảm ơn

Tác giả Nguyễn Ái Việt chân thành cảm ơn các đồng nghiệp tại Viện Công nghệ thông tin (ITI), Đại học Quốc gia, đặc biệt là PGS. TS. Nguyễn Đình Hóa, TS. Phạm Bảo Sơn đã góp ý cho Báo cáo. Các chuyên gia ngôn ngữ tại Đại học Xã hội và Nhân văn, đặc biệt là PGS.TS Vũ Đức Nghiệu đã đóng góp những ý kiến quý báu.

Các tác giả cám ơn các cộng sự tại VIEGRID là các kỹ sư phần mềm Trần Đăng Hòa, Hoàng Ngọc Tường Vy đã trực tiếp góp sức cho kết quả báo cáo. Đóng góp ý kiến của các chuyên gia CNTT và các chuyên gia ngôn ngữ là hết sức quan trọng đối với kết quả cuối cùng của bản Báo cáo.

Các tác giả chân thành cảm ơn Công ty VIEGRID đã cho phép công bố các thông tin và đã tài trợ cho nghiên cứu này.

TÓM TẮT

Trong giai đoạn phát triển trước mắt, xã hội Việt nam đang phải đối diện với nhiều thách thức mới. Chính tả tiếng Việt đang bị xao lãng so với các thách thức về kinh tế xã hội khác. Bản "Báo cáo về tình hình chính tả văn bản tiếng Việt" này là một đánh giá nhỏ về chất lượng chính tả văn bản tiếng Việt và cũng là lời kêu gọi thiết tha "Hãy cùng chúng tôi giữ gìn tiếng Việt". Nhóm tác giả mong mỏi rằng với một nhận thức mới mẻ. toàn thể cộng đồng người Việt nam sẽ cùng tuyên chiến với vấn nạn xã hội này.

Trước khi đánh giá, chúng tôi đã tiến hành một cuộc điều tra nhỏ trong hai nhóm Chuyên gia ngôn ngữ và Chuyên gia CNTT. Nhóm chuyên gia ngôn ngữ yêu cầu tỷ lệ lỗi chính tả trong văn bản Việt phải là dưới 1%. Nhóm chuyên gia CNTT chấp nhận tỷ lệ này trong khoảng 2.5-5%. Hai nhóm chuyên gia đều nhất trí cho rằng báo chí truyền thông có trách nhiệm nhiều nhất đối với chính tả tiếng Việt. Tuyệt đại đa số các chuyên gia cũng cho rằng tỷ lệ 10% là ngưỡng báo động đối với các lỗi chính tả và tỷ lệ 30% là ngưỡng mà một từ đã không còn là lỗi chính tả.

Đợt xếp hạng tháng 6/2010 đã đánh giá 177 đơn vị và xếp hạng 132 đơn vị trong 7 khu vực: 1) Bộ và Văn phòng Trung ương; 2) Ủy ban nhân dân Tỉnh và Thành phố trực thuộc Trung ương; 3) Cơ quan thuộc Chính phủ và thuộc Bộ; 4) Đại học và Viện nghiên cứu; 5) Báo chí, nhà xuất bản và cơ quan truyền thông; 6) Doanh nghiệp Việt nam; 7) Tổ chức và doanh nghiệp nước ngoài tại Việt nam. Chúng tôi đã chọn phương pháp thống kê với tập lỗi điển hình, là phương pháp phù hợp với điều kiện hạn chế về nguồn lực. Với 67 nghìn mẫu thống kê, tỷ lệ lỗi chính tả trung bình của văn bản tiếng Việt là 7.79%, cao hơn nhiều so với mức yêu cầu tối thiểu. Nhóm tác giả đã quan sát được tỷ lệ lỗi chính tả cao nhất như sau: "**soi** mói" 74.33%, "**sáng** lạn" 41.66%, "cọ **sát**" 28.38% và "**thăm** quan" 20.61%. Khu vực báo chí và truyền thông có tỷ lệ lỗi chính tả cao nhất, gần mức báo động 10%. Khu vực Đại học và Viện nghiên cứu có tỷ lệ lỗi xấp xỉ mức trung bình của xã hội, chưa phát huy được tính mẫu mực và tiên phong trong vấn đề dùng chữ nghĩa. Điều đáng nói là trong cả hai khu vực này đều có các đại diện có tỷ lệ lỗi vượt mức 30%. Khu vực các chính quyền địa phương, và các cơ quan thuộc chính phủ, thuộc Bộ có tỷ lệ lỗi chính tả khá cao. Đặc biệt, có đơn vị có tỷ lệ lỗi gần 40%. Ngay cả các khu vực khá nhất là doanh nghiệp và các Bộ vẫn còn phải tiếp tục cải thiện chất lượng để có thể đạt được mức 1%. Các kết quả đánh giá chi tiết được công bố trên trang Web www.xephangvanban.com khai trương ngày 26/7/2010.

Các kết quả nói trên là một cố gắng của các tác giả để giúp toàn thể xã hội và các đơn vị đã được xếp hạng bước đầu nhận thức được về vấn đề chính tả tiếng Việt. Các đợt đánh giá tiếp sau sẽ được tiến hành 3 tháng một lần và sẽ liên tục được mở rộng về quy mô để hậu thuẫn cho một chiến dịch cộng đồng về quét lỗi chính tả.

Việc giới thiệu các sản phẩm soát lỗi chính tả tiếng Việt một cách khách quan cũng sẽ là cần thiết cho chiến dịch quét lỗi chính tả. Nhóm tác giả đã phân tích thống kê dựa trên khối liệu chủ yếu là báo chí và đã ước lượng được tỷ lệ giữa lỗi phi từ và lỗi thực từ trong tiếng Việt là 31.70%: 68.30%. Khác với quan niệm của một số chuyên gia CNTT và tình trạng trong tiếng Anh, lỗi thực từ chiếm đa số lỗi chính tả trong tiếng Việt. Điều đó có thể giải thích tại sao các sản phẩm không có khả năng quét lỗi thực từ đã không nhận được hưởng ứng mạnh mẽ từ phía người dùng.

Một kết quả đánh giá khác cho thấy, trái với nhận thức của nhiều người, hiện đã có một số sản phẩm Việt nam đạt được độ nhận biết lỗi thực từ vượt xa mức 33% của Microsoft Word 2007. Đặc biệt có sản phẩm đã đạt tỷ lệ 92.5%. Kết hợp với giải pháp hạ tầng quét lỗi cho cơ quan xí nghiệp, việc đạt tới độ nhận biết lỗi trên 99% đã là hiện thực.

Để có đánh giá khách quan, chúng tôi đề nghị sử dụng các độ đo như độ nhận biết, độ chính xác và khả năng gợi ý để đánh giá hiệu năng cho các phần mềm soát lỗi. Chúng tôi cũng đưa ra độ đo VIE độ đo cân bằng giữa các yếu tố nói trên và tỷ lệ với tần số xuất hiện các lỗi phi từ và lỗi thực từ.

Trong tương lai, các doanh nghiệp, chuyên gia và người sử dụng có thể tiếp tục giới thiệu các sản phẩm mới của mình với cộng đồng trên trang Web www.xephangvanban.com.

Nỗ lực đơn độc dù cố gắng đến đâu cũng không thể mang lại kết quả như mong muốn. Với công trình này, chúng tôi hy vọng sẽ có các nhà quản lý, chuyên gia ngôn ngữ và nhà văn hóa sẽ cùng vào cuộc để dẫn dắt cho chương trình cộng đồng này. Đồng thời, hy vọng rằng sẽ có nhiều bạn đồng nghiệp trong lĩnh vực CNTT cho ra các sản phẩm cùng giúp ích cho cộng đồng. **HÃY CÙNG CHÚNG TÔI GIỮ GÌN TIẾNG VIỆT.**

NỘI DUNG

MỞ ĐẦU

Ngôn ngữ hình thành từ nhu cầu chia sẻ thông tin là động lực giúp con người suy nghĩ, đứng thẳng dậy trên đôi chân của mình và làm chủ thế giới. Loài người tìm ra chữ viết và dùng các văn bản để chuyển tải tri thức tới cho các thế hệ tương lai. Nhờ đó, ngày nay loài người đã có được cuộc sống tốt đẹp hơn và có thể ngẩng cao đầu để nhìn về những khoảng không cách xa hàng triệu năm ánh sáng.

Văn bản là tinh hoa tư duy của nhiều thế hệ, là nguyên khí dân tộc, quyết định sự tồn vong của quốc gia. Trong đêm đen nô lệ ngoại bang dài thăm thẳm, biết bao thế hệ người Việt đã không ngừng trút tim óc của mình lên giấy mực với niềm hy vọng bi thiết "tiếng ta còn thì nước ta còn". Kho thư tịch tiếng Việt đã bị mai một nhiều do binh hỏa và thời gian, nhưng tiếng Việt, gia tài của cha ông để lại vẫn còn đây, để cho chúng ta được làm người Việt và có một nước Việt nam toàn vẹn. Trong những năm chiến tranh, các nhà lãnh đạo ở cả hai miền vẫn đau đáu kêu gọi "giữ gìn sự trong sáng trong tiếng Việt" và "chỉnh đốn Việt ngữ" vì một tương lai tươi sáng của dân tộc.

Ngày nay, chúng ta như quá mải miết trong những thăng trầm của cuộc sống và quên rằng bao nhiêu đế chế hùng mạnh đã lụi tàn, bao nhiêu nền văn hóa rực rỡ đã chìm vào quên lãng, bao nhiêu sinh ngữ đã trở thành tử ngữ trong những cuộc va đập khốc liệt giữa các nền văn minh. Để có nước non và tiếng nói Việt này, các thế hệ tiền nhân đã phải quên thân đền nợ nước và "dưới trăng bao thu bạc đầu mài kiếm".

Hội nhập kinh tế văn hóa là một hành trình gian nan, đòi hỏi ngôn ngữ phải đủ chính xác để thu nhận được tri thức của nhân loại vừa đủ sức phổ cập để biến tri thức đó thành sức mạnh của cộng đồng người Việt. Tiếng Việt đáng lẽ phải được quan tâm chăm sóc, chỉnh đốn và phát triển hơn nữa. Tiếc thay, hiện nay tiếng Việt đang bị xuống cấp, lỗi chính tả tràn lan trên phương tiện thông tin đại chúng, hoành hành trong học đường, len lỏi cả vào các văn bản pháp quy và gieo mầm bệnh trong sách vở như một đại dịch. Việc sử dụng các phương tiện soạn thảo cắt dán, thư điện tử và Internet lại càng tạo điều kiện cho các loại vi rút này nảy nở sinh ra nhiều chủng loại kỳ dị khó chữa và truyền nhiễm nhanh hơn.

Một số chuyên gia ngôn ngữ cho rằng đây không còn là nguy cơ tiềm ẩn chỉ cần cảnh báo mà đã một đại nạn như cháy hay vỡ đê phải huy động sức mạnh của cả cộng đồng mới giải cứu được. Một vấn nạn xã hội lớn như vậy chỉ có thể đối phó nếu có một chiến dịch truyền thông lớn với mọi người Việt cùng vào cuộc.

Nhóm tác giả của "Báo cáo về tình hình chính tả Tiếng Việt: Đợt đánh giá Tháng 6/2010" là các chuyên gia về công nghệ xử lý tiếng Việt tâm huyết với tương lai của tiếng Việt. Chúng tôi hy vọng đây sẽ là lời tuyên chiến đầu tiên của tất cả chúng ta với vấn nạn chính tả tiếng Việt. Chúng tôi rất mong sẽ có các nhà quản lý, các chuyên gia ngôn ngữ,

các nhà báo, trí thức, các doanh nghiệp và các bạn đồng nghiệp sẽ cùng tham gia để chủ trương và chỉ dẫn thêm cho đại cục. Cũng mong các đơn vị đã được xếp hạng sẽ thấy được thiện chí đó, không hiềm hơn kém để cùng chỉnh đốn lại tiếng Việt của chúng ta.

CHƯƠNG 1: VẤN NẠN CHÍNH TẢ TIẾNG VIỆT

Đầu thế kỷ XX, chữ Quốc ngữ đã trở nên phổ biến để ghi tiếng Việt bằng các ký tự La tinh. Đây là một điều may mắn lớn giúp tiếng Việt phát triển, trở thành một ngôn ngữ có khả năng chuyển tải các tư tưởng mới, có sức phổ cập, có thể đánh vần theo ký âm, dễ học và dễ đọc. Tiếng Việt thừa hưởng được các cách dùng chữ viết hoa để phân biệt danh từ riêng, dấu ngắt câu và các dấu biểu lộ tình thái do ký âm La tinh đem lại.

Là một ngôn ngữ đơn âm vị, tiếng Việt đòi hỏi xử lý phức tạp hơn so với tiếng Anh hoặc các ngôn ngữ đa âm vị khác. Để quyết định một âm vị hay một từ sai chính tả, phải xét chúng trong tương quan văn cảnh. Chẳng hạn, "trưa" và "phải" đều là các âm vị có trong từ điển, nhưng tổ hợp "trưa phải" lại là một lỗi chính tả tiềm năng. Tuy nhiên, câu "Anh làm đến trưa phải không?" lại không có lỗi chính tả. Nói một cách khác, trong tiếng Việt, giữa lỗi chính tả và lỗi ngữ pháp không có một ranh giới rõ ràng được đánh dấu bằng dấu cách (ký tự trắng) như trong tiếng Anh.

Phần lớn lỗi chính tả tiếng Việt bắt nguồn từ việc thiếu hệ thống chuẩn hóa cách ký âm và phát âm. Đến nay, các cách hài dấu, sử dụng "i" hay "y", "d" hay "gi", phiên âm địa danh, tên người vẫn là đề tài tranh cãi và chưa đạt được sự đồng thuận. Có lẽ không có quốc gia nào trên thế giới có nhiều cách phát âm vùng miền nhiều như Việt nam. Mỗi vùng miền đều mang vào tiếng Việt những lỗi chính tả đặc trưng của mình như "l-n", "ch-tr", "s-x" ở miền Bắc, "t-c", "n-ng" ở miền Nam, nhầm lẫn về các dấu "sắc" và "huyền" ở miền Trung. Từ đó mà có những chuyện tiếu lâm về "lờ thấp lờ cao", "xờ bướm sờ chim" hay "cá có cuống cà có đuôi". Các lỗi chính tả khó sửa nhất xuất hiện còn do thói suy luận mò trên cơ sở một âm vị sai ngẫu nhiên trùng nghĩa như "soi mói", "cọ sát", "thăm quan"… Trong văn bản được soạn thảo trên máy tính, lỗi chính tả xuất hiện do việc gõ sai chệch sang ký tự bên cạnh, đảo lộn thứ tự ký tự, dính từ do bỏ quên ký tự trắng hoặc sự xuất hiện các ký tự "w", "s", "j" không đúng chỗ.

Vào những năm 60-70 của thế kỷ trước, mục Dọn vườn của Báo Văn nghệ là chuyên mục giúp ích nhiều cho chính tả tiếng Việt. Cũng trong thời gian đó, các ấn phẩm ít lỗi chính tả và sử dụng chữ nghĩa của công chức cũng có trách nhiệm hơn nhiều so với ngày nay. Rồi bẵng đi một dạo, vườn văn Việt không có ai chăm lo, lỗi chính tả đã nảy sinh nhiều như sâu bệnh mùa lụt trên sách vở, phương tiện thông tin đại chúng và nơi công cộng. Biểu ngữ "bánh trưng" ở lễ hội vua Hùng, hay "đất nớc" bên cạnh Ủy ban nhân dân thành phố Hà nội chỉ làm được dư luận xôn xao vài ngày đã trở thành chuyện "biết rồi khổ lắm nói mãi". Lỗi chính tả nhiều đến mức chữa không xuể, trám được chỗ này lại bục ở chỗ kia, làm nản chí ngay cả các thức giả.

Sẽ sai lầm vô cùng nếu nghĩ rằng chính tả là chuyện nhỏ hay quá cao xa. Chính tả có tầm quan trọng đặc biệt trong đời sống văn hóa xã hội. Văn bản pháp luật sai chính tả sẽ ảnh hưởng tới lòng tin của công dân đối với các cơ quan công quyền. Báo chí sách vở sai

chính tả sẽ làm méo mó thông tin, để lại mầm độc vào ngôn ngữ và tư duy của thế hệ trẻ. Có lẽ vì lỗi chính tả quá nhiều, chúng ta đã trở nên chai lì với chúng đến mức thờ ơ. Hãy nhớ rằng quan tâm tới chính tả cũng còn là quan tâm đến quyền lợi thiết thực của mỗi người. Đơn thư sai chính tả có thể ảnh hưởng tới con đường tiến thân và chế độ lương bổng của người lao động. Chào hàng, giới thiệu sản phẩm sai chính tả sẽ làm tổn hại tới hình ảnh và doanh thu của doanh nghiệp.

Nhiều tờ báo lớn trên thế giới yêu cầu tỷ lệ lỗi chính tả tiêu chuẩn là dưới mức 0.1%. Với tình trạng lỗi chính tả trầm trọng như ở nước ta hiện nay, nhiều chuyên gia cho rằng lỗi chính tả cũng phải ở mức dưới 1%. Tuy vậy, kết quả khảo sát của chúng tôi cho thấy có rất ít tổ chức đạt được mức này. Thậm chí còn có những tổ chức có tỷ lệ sai lỗi chính tả tới mức 30-40% là ranh giới giữa sai và đúng đã bị xóa nhòa và mọi từ điển đều trở nên vô dụng.

Khi việc đảo lộn sai đúng và thiếu trách nhiệm với chất lượng công việc trở thành một thái độ sống, chúng ta sẽ phải đối mặt với một vấn nạn xã hội còn lớn hơn nhiều. Sửa lỗi chính tả cũng là bước đầu để chấn chỉnh kỷ cương quốc gia, nâng cao chất lượng công việc và tinh thần trách nhiệm của công dân.

Công nghệ thông tin có thể đem lại những sản phẩm xử lý văn bản, hỗ trợ biên dịch và máy tìm kiếm. Trong số đó, soát lỗi chính tả là nhu cầu thiết thực hàng đầu. Người dùng đòi hỏi phải có những sản phẩm đúng quy trình chất lượng và dễ mất lòng tin với các sản phẩm sơ sài. Họ thường không biết rằng trên thị trường đã có những sản phẩm có thể đáp ứng nhu cầu soát lỗi chính tả. Nhiều nhà sản xuất cho rằng thị trường cho sản phẩm soát lỗi tiếng Việt quá nhỏ hẹp, do đó đã ngừng phát triển hoặc không đầu tư hơn nữa vào nghiên cứu phát triển. Thực ra, thị trường các phần mềm nâng cao chất lượng văn bản tiếng Việt là vô tận với hơn 5 triệu máy tính dùng được để soạn thảo văn bản hàng ngày.

Công nghệ thông tin chỉ có thể đem lại các công cụ trợ giúp, việc quét sạch lỗi chính tả trong văn bản tiếng Việt phải nhờ đến một chiến dịch cộng đồng rộng lớn, trong đó báo chí, truyền thông và các trường đại học phải đi đầu. Trong chiến dịch này, trước hết các nhà ngôn ngữ, các nhà văn hóa phải lên tiếng tạo thành dư luận xã hội và thuyết phục các cơ quan nhà nước ủng hộ để sớm có được một chương trình tái thiết tiếng Việt thậm chí một đạo luật về sử dụng tiếng Việt. Các chuyên gia công nghệ thông tin sẽ phải cố gắng để có những sản phẩm thiết thực phục vụ cho công cuộc tái thiết này.

Việc xếp hạng văn bản sẽ nâng cao nhận thức xã hội về lỗi chính tả này bằng những con số, để giúp mỗi người chúng ta cảm nhận được vấn nạn này một cách cụ thể hơn để cùng nhau giữ gìn tiếng Việt

CHƯƠNG 2: MỤC TIÊU VÀ PHƯƠNG PHÁP

2.1 Mục tiêu tổng quát

Xếp hạng văn bản là hoạt động khởi đầu cho một chiến dịch truyền thông về vấn đề chính tả tiếng Việt với lời kêu gọi "Hãy cùng chúng tôi giữ gìn tiếng Việt"

2.2 Các mục tiêu cụ thể

a. Tạo dư luận xã hội để hỗ trợ cho các hoạt động truyền thông
b. Giúp phát hiện vấn đề, tránh thành tích hình thức hay tâm lý đấu chọi
c. Nhiều đợt đánh giá thường xuyên và liên tục
d. Lôi kéo được sự tham gia rộng rãi của cộng đồng.

2. 3 Tiêu chí đánh giá

a. Có kết quả định lượng và khách quan
b. Phù hợp với điều kiện nguồn lực với số lượng mẫu đủ thuyết phục.
c. Dựa trên tập lỗi phổ biến
d. Phân chia theo khu vực, để tránh so sánh không cân xứng.

2.4 Một số khái niệm về đánh giá chất lượng chính tả văn bản tiếng Việt

Chính tả Cách viết được cho là chuẩn. Trong Báo cáo này, chúng tôi tạm coi một số từ điển như từ điển Hoàng Phê và từ điển của Ủy ban Khoa học Xã hội làm gốc, tuy chưa có một quy định nào chính thức.

Lỗi chính tả Sai lệch so với cách viết chuẩn. Trong tình trạng chưa có chuẩn được công bố chính thức, bên cạnh một số chuẩn mặc định theo một số từ điển được tạm chọn là gốc, các cơ quan có thể tạm quy định một số chuẩn dùng trong nội bộ. Lỗi chính tả chia làm hai loại *lỗi phi từ (nonword)* và *lỗi thực từ (real-word).*

Lỗi phi từ (nonword) Trong ngôn ngữ đa âm vị, lỗi phi từ là lỗi chính tả không trùng với bất cứ từ nào có trong từ điển. Trong ngôn ngữ đơn âm vị, lỗi phi từ là bất cứ lỗi nào có chứa âm vị không có trong từ điển.

Lỗi thực từ (real-word) Trong ngôn ngữ đa âm vị, lỗi thực từ là các lỗi chính tả ngẫu nhiên trùng với các từ khác có trong từ điển. Trong ngôn ngữ đơn âm vị, lỗi thực từ là lỗi chính tả chỉ chứa các âm vị có thể tìm thấy trong từ điển. Khác với tiếng Anh, tiếng Việt có lỗi thực từ nhiều hơn lỗi phi từ. Microsoft Word từ phiên bản 2007 trở đi mới chú trọng đến lỗi thực từ.

Tập lỗi Danh sách lỗi chính tả và các từ gốc tương ứng. Tập lỗi có thể sinh bởi một số quy tắc sinh lỗi, hoặc được thống kê từ thực tế. Chúng tôi tập trung đánh giá lỗi thực từ là nhóm lỗi chính tả chiếm đa số và khó soát trong văn bản tiếng Việt.

Máy tìm kiếm Máy tìm kiếm là một hệ thống tìm thông tin được thiết kế để tìm kiếm thông tin lưu trên một hệ máy tính. Máy tìm kiếm Google, nổi tiếng về việc tìm kiếm thông tin trên Internet, chứa lượng thông tin lớn nhất.

Thống kê Là việc thu thập và phân tích các số liệu để tìm ra quy luật. Thống kê sẽ có tính khách quan nếu quy mô mẫu số liệu thu thập được đủ lớn và việc lựa chọn mẫu không có tính chủ quan.

Tỷ lệ lỗi Với một từ lỗi cho trước, trong một thống kê, nếu số lần xuất hiện của lỗi là L, số lần xuất hiện của từ viết đúng là Đ, tỷ lệ lỗi sẽ được tính bằng L/(L+Đ).

Số mẫu Số mẫu là tổng số các số liệu được thu thập trong mỗi thống kê. Trong thống kê lỗi chính tả, số mẫu là L+Đ bằng tổng số các lỗi và các từ đúng thu thập được.

2.5 Phỏng vấn chuyên gia

Để định hướng cho việc đánh giá, chúng tôi đã tiến hành phỏng vấn hai nhóm chuyên gia: chuyên gia ngôn ngữ và chuyên gia CNTT. Một số kết quả đáng chú ý như sau:

a. *Tỷ lệ lỗi chính tả chấp nhận được trong tiếng Việt*: Nhóm chuyên gia CNTT chấp nhận tỷ lệ 2-5%. Chúng tôi đồng tình với ý kiến của nhóm chuyên gia ngôn ngữ chỉ chấp nhận tỷ lệ dưới 1%.

b. *Mức độ một lỗi chính tả còn được coi là lỗi*: Cả hai nhóm chuyên gia đều thống nhất ở mức 30%. Như vậy, khi một lỗi chính tả đạt mức 30-70%, theo các chuyên gia, đã có thể được chấp nhận như một cách viết khác của từ tương ứng có trong từ điển. Cũng theo đó, khi một lỗi chính tả truyền thống vượt mức 70%, bản thân nó đã trở thành từ đúng và có thể thay thế từ trong từ điển.

c. Tầm quan trọng của vấn đề chính tả: Cả hai nhóm chuyên gia đều nhất trí chính tả là vấn đề văn hóa, xã hội và kinh tế quan trọng và phải là mục tiêu hàng đầu của việc nghiên cứu xử lý tiếng Việt.

d. Vai trò của các chủ thể trong xã hội đối với lỗi chính tả và sửa lỗi chính tả: Đa số các chuyên gia trong cả hai nhóm đều nhất trí là báo chí và truyền thông có ảnh hưởng lớn nhất tới tình trạng lỗi chính tả. Một số chuyên gia khác nhấn mạnh vai trò của nhà nước, các trường đại học và viện nghiên cứu.

e. Nhận thức về các sản phẩm soát lỗi chính tả: Đa số các chuyên gia biết rất ít về các sản phẩm soát lỗi chính tả tiếng Việt. Một số chuyên gia bi quan về thị trường của các

phần mềm này và cho rằng nhà nước cần phải đầu tư cho việc phát triển các phần mềm này.

2.6 Phương pháp đánh giá

Đánh giá sẽ được tiến hành theo các đợt. Sau đợt đầu tiên vào tháng 6/2010, dự kiến các đợt đánh giá sẽ được tiến hành thường xuyên ba tháng một lần. Kết quả của các đợt sẽ được sắp xếp theo tiến trình thời gian giúp người sử dụng quan sát được diễn biến của lỗi chính tả.

Do thời gian, kinh phí và nguồn lực hạn chế, việc sử dụng phương pháp tải về và quét toàn bộ các văn bản có liên quan tới một trang Web để đánh giá "vét cạn" không khả thi. Chúng tôi chọn phương pháp thống kê với các bước như sau:

a. Chọn một tập lỗi của những từ "phổ biến" có tần suất xuất hiện cao trong những văn bản chúng tôi đã thu thập được trong quá trình xây dựng khối liệu tiếng Việt. Trong tương lai, tập lỗi này sẽ được mở rộng theo những nguyên tắc thống nhất và sẽ được cộng đồng bổ sung thêm để có tính khách quan.

b. Sử dụng máy tìm kiếm của Google và một chương trình phần mềm tự động để tìm kiếm số lần một lỗi nhất định và số lần từ đúng của nó xuất hiện trên một Website cho trước. Thông tin này được xử lý để tránh các trường hợp nhập nhằng và trùng lặp. Kết quả tìm kiếm của Google bao gồm cả những từ xuất hiện trên các văn bản doc, pdf và một số định dạng khác có thể tải về từ các trang Web.

c. Lựa chọn một danh sách các Website của các đơn vị tiêu biểu trong các khu vực khác nhau như:

- Các Bộ và văn phòng Trung ương
- Các Ủy ban nhân dân Tỉnh và thành phố trực thuộc Trung ương
- Các cơ quan thuộc Chính phủ và thuộc Bộ
- Các trường đại học và viện nghiên cứu
- Các báo, nhà xuất bản và cơ quan truyền thông
- Các doanh nghiệp trong nước
- Các doanh nghiệp nước ngoài và tổ chức quốc tế ở Việt nam.

d. Thống kê và tính tỷ lệ lỗi đối với tập lỗi đã chọn cho từng đơn vị và sắp xếp theo thứ tự chung và thứ tự riêng trong mỗi khu vực; Thống kê tỷ lệ của từng lỗi; Phân tích và đưa

ra một số kết luận. Để đảm bảo tính thống kê, các Website không đủ 1000 mẫu sẽ không được xếp hạng.

2.7 Hạn chế của phương pháp và phương hướng khắc phục

Phương pháp thống kê trên đây có ưu điểm là có thể cho một đánh giá nhanh, tương đối khách quan trong điều kiện thời gian và nguồn lực hạn chế. Các kết luận về đặc trưng về lỗi của mỗi khu vực hoặc tần suất của các lỗi có thể xem là tương đối khách quan vì số lượng mẫu đã đạt đến con số hàng chục nghìn.

Tuy nhiên, phương pháp này chưa thể coi là một cách đánh giá đầy đủ và việc xếp hạng không phải là tuyệt đối chính xác với các lý do sau đây:

a. *Tập lỗi còn hạn chế*. Phương án khắc phục: Trong tương lai, tập lỗi này sẽ không ngừng được mở rộng, nhờ một thuật toán tìm kiếm các lỗi phổ biến trong khối liệu của chúng tôi. Mặt khác, người dùng có thể đề xuất các lỗi mới. Mỗi lỗi nhận được đủ một số lượng đề xuất nhất định từ các địa chỉ IP khác nhau sẽ được đưa vào tập lỗi để đánh giá.

b. *Số lượng các Website còn hạn chế và chưa thực sự tiêu biểu*. Phương án khắc phục: Người dùng có thể đề xuất các Website mới và yêu cầu đánh giá. Các Website có đủ một số yêu cầu nhất định từ các địa chỉ IP sẽ được đưa vào danh sách xếp hạng lần tới.

Như vậy số lượng Website và số lượng mẫu thống kê sẽ ngày một tăng và việc đánh giá chất lượng văn bản tiếng Việt sẽ ngày một toàn diện hơn

CHƯƠNG 3: KẾT QUẢ ĐỢT ĐÁNH GIÁ THÁNG 6/2010

Đợt đánh giá tháng 6/2010 khảo sát 177 tổ chức thuộc 7 khu vực. Tập lỗi dùng để đánh giá trong đợt này được chọn từ một số lỗi "phổ biến" gồm các lỗi do phát âm (như "bổ **xung**", "**sử** lý", "**xử** dụng",…) và các lỗi có âm vị sai tình cờ có nghĩa gần với nghĩa của cả từ (như "**sáng** lạn", "cọ **sát**", "**soi** mói", "**thăm** quan",…).

Trong số các tổ chức đã khảo sát, chúng tôi quyết định chỉ xếp hạng 132 tổ chức có số mẫu hơn 1000 để đảm bảo tính khách quan. Kết quả cụ thể cho từng tổ chức có thể tra cứu được trên trang www.xephangvanban.com. Kết quả đánh giá và xếp hạng của các tổ chức tiêu biểu có tỷ lệ lỗi cao nhất và thấp nhất cho một số khu vực và cho toàn đợt đánh giá được hiển thị để người dùng có một hình dung cụ thể về tình trạng chính tả tiếng Việt. Bên cạnh đó, các lỗi có tỷ lệ cao nhất cũng được thống kê để nghiên cứu nguyên nhân.

3.1 Kết quả đánh giá theo khu vực

Trên tổng số gần 67 nghìn mẫu, tỷ lệ lỗi chính tả trong văn bản tiếng Việt là 7.79%, một con số cao hơn nhiều so với tiêu chuẩn 1% do các chuyên gia ngôn ngữ đề ra và rất cao so với tiêu chuẩn quốc tế 0.1%.

Báo chí, xuất bản và truyền thông là khu vực phạm lỗi nhiều nhất, có tỷ lệ lỗi là 9.58% phù hợp với dự báo của các chuyên gia. Các cơ quan nhà nước có tỷ lệ phạm lỗi khá đa dạng. Trong khi khu vực Bộ và văn phòng Trung ương có tỷ lệ lỗi chỉ là 4.24%, các Ủy ban nhân dân Tỉnh và các cơ quan trực thuộc chính phủ hoặc thuộc Bộ có tỷ lệ lỗi khá cao lần lượt là 8.15% và 8.63%. Khu vực Đại học và Viện nghiên cứu có tỷ lệ lỗi xấp xỉ như mức trung bình của cả nước chưa chứng tỏ được vai trò đi đầu và gương mẫu về chữ nghĩa. Khu vực tổ chức và doanh nghiệp nước ngoài, được đánh giá để làm tiêu bản so sánh, có mức phạm lỗi dưới mức trung bình của cả nước là 5.68%, xếp thứ 3 trong các khu vực. Điều đó chứng tỏ rằng quy trình chất lượng đã ảnh hưởng tốt tới chất lượng văn bản, tuy sự quan tâm cụ thể về chính tả tiếng Việt trong khu vực này chưa chắc đã cao.

Các doanh nghiệp được đánh giá, chủ yếu là các Tổng công ty lớn và các ngân hàng, có số lỗi ít nhất. Có thể giả thiết trong khu vực này do yêu cầu kinh doanh, sự quan tâm đến chất lượng văn bản lớn hơn các khu vực khác.

STT	***Khu vực***	***Số đơn vị đánh giá***	***Số đơn vị xếp hạng***	***Tỷ lệ lỗi***	***Số mẫu***
1	Doanh nghiệp Việt Nam	35	13	2.96%	39867
2	Bộ và cơ quan Trung ương	19	17	4.28%	79873
3	Tổ chức & Doanh nghiệp nước ngoài	18	7	5.68%	29594
4	Đại học và Viện nghiên cứu	23	18	7.13%	54157
5	Chính quyền địa phương	32	29	8.15%	139583
6	Cơ quan thuộc Chính phủ và thuộc Bộ	14	15	8.63%	58954
7	Báo chí, xuất bản và truyền thông	36	33	9.58%	263435
TỔNG		177	132	7.79%	665463

3.2 Kết quả tiêu biểu của các Bộ và cơ quan trung ương

Khu vực đánh giá này gồm 17 Bộ và các Văn phòng Quốc hội, Trung ương Đảng và Chính phủ. Bộ Nội vụ, cơ quan hướng dẫn về văn bản hành chính, là Bộ có tỷ lệ lỗi 1.20%

khá gần với mức 1%. Bộ Tư pháp và Bộ Tài chính có tỷ lệ lỗi cao khá bất ngờ. do trong lĩnh vực pháp luật và tài chính có yêu cầu cao về chất lượng văn bản.

Các lỗi chứa âm vị sai có nghĩa gần đúng nghĩa gốc có tỷ lệ phạm lỗi rất cao. Có thể thấy các từ lỗi như "soi mói", "cọ sát", "sáng lạn" đã đạt mức độ sử dụng đủ có thể coi là một cách viết mới của cùng với các từ "xoi mói", "cọ xát" và "xán lạn". Lỗi "thăm quan" cũng có tỷ lệ đến 17.80% vượt xa ngưỡng báo động.

a. Các Bộ ít lỗi nhất

STT	*Bộ và cơ quan trung ương*	*Tỷ lệ lỗi*	*Số mẫu*
1	Bộ Nội vụ	1.20%	4259
2	Bộ Y tế	2.04%	4561
3	Bộ Giáo dục Đào tạo	2.10%	2667
4	Bộ Khoa học và Công nghệ	2.46%	3136
5	Bộ Giao thông	2.99%	4110

b. Các Bộ nhiều lỗi nhất

STT	*Bộ và cơ quan trung ương*	*Tỷ lệ lỗi*	*Số mẫu*
1	Bộ Văn hóa, Du lịch, Thể thao	7.47%	4672
2	Bộ Tư pháp	7.30%	6042
3	Bộ Tài chính	6.97%	6611
4	Bộ Tài nguyên và Môi trường	5.38%	5355
5	Văn phòng Trung ương Đảng	4.50%	6997

c. Những lỗi chính tả có tỷ lệ cao

STT	*Lỗi*	*Tỷ lệ lỗi*	*Số mẫu*

1	soi mói	56.38%	149
2	sáng lạn	38.28%	128
3	cọ sát	36.71%	346
4	thăm quan	17.80%	5370

3.3 Kết quả tiêu biểu của các địa phương

Trừ Tuyên Quang xếp ở vị trí số 5, các tỉnh có tỷ lệ phạm lỗi ít nhất đều ở phía Nam với tỷ lệ lỗi tương đương như trong khu vực Bộ và cơ quan trung ương.

Một kết quả đáng ngạc nhiên là trong số 5 địa phương có tỷ lệ lỗi cao nhất lại có 3 thành phố lớn là TP Hồ Chí Minh, Đà Nẵng, Hải Phòng và các Tỉnh tương đối phát triển là Bắc Ninh và Đồng Nai.

a. Các địa phương ít lỗi nhất

STT	Tỉnh	Tỷ lệ lỗi	Số mẫu
1	Tỉnh Lâm Đồng	2.08%	3083
2	Tỉnh Trà Vinh	2.62%	3708
3	Tỉnh Bình Dương	2.90%	4275
4	Tỉnh Tiền Giang	2.93%	5699
5	Tỉnh Tuyên Quang	2.99%	1839

b. Các địa phương nhiều lỗi nhất

STT	***Tỉnh***	***Tỷ lệ lỗi***	***Số mẫu***
1	TP Hồ Chí Minh	18.98%	9703
2	Đồng Nai	17.31%	7185
3	Đà Nẵng	15.83%	6423
4	Bắc Ninh	11.69%	7496
5	Hải Phòng	11.19%	5441

c. Những lỗi có tỷ lệ cao nhất

STT	*Lỗi*	*Tỷ lệ lỗi*	*Số mẫu*
1	soi mói	41.67%	120
2	sáng lạn	37.16%	183
3	cọ sát	30.50%	505
4	thăm quan	16.41%	9091

3.4 Kết quả tiêu biểu của các trường Đại học và Viện nghiên cứu

Các trường Đại học và Viện nghiên cứu là nơi có các trí thức đầu ngành làm việc và là nơi đào tạo ra thế hệ tương lai của đất nước. Tình trạng lỗi chính tả trong khu vực này cũng phần nào nói lên chất lượng của đội ngũ tinh hoa của Việt nam. Việc có tỷ lệ lỗi chính tả tương đương với tỷ lệ phạm lỗi trung bình của xã hội đáng báo động vì khu vực này không giữ được tính tiên phong và khuôn mẫu.

Đặc biệt, Viện Năng lượng Nguyên tử có tỷ lệ lỗi rất cao ở mức 31.49%, vượt ranh giới mà năng lực về chính tả còn cho phép phân định lỗi nếu có trợ giúp. Việc các trường đại học danh tiếng như Đại học Bách Khoa, Đại học Cần Thơ, Đại học Sư phạm Hà Nội và Đại học Đà Nẵng rơi vào tốp cuối với tỷ lệ lỗi báo động đáng để toàn xã hội phải suy nghĩ.

a. Các Đại học và Viện nghiên cứu có nhiều lỗi nhất

STT	*Đại học và Viện nghiên cứu*	*Tỷ lệ lỗi*	*Số mẫu*
1	Viện Năng lượng Nguyên từ	31.49%	1826
2	Đại học Đà Nẵng	21.67%	2576
3	Đại học Bách khoa Hà Nội	11.83%	3768
4	Đại học Cần Thơ	8.98%	6531
5	Đại học Sư phạm Hà Nội	8.76%	4479

b. Các Đại học và Viện nghiên cứu ít lỗi nhất

STT	*Đại học và Viện nghiên cứu*	*Tỷ lệ lỗi*	*Số mẫu*
1	Bách khoa Tự điển	0.45%	2425
2	Đại học Vinh	1.25%	1381
3	Viện Khoa học Xã Hội	1.47%	1457
4	Đại học Đà Lạt	2.07%	2763
5	Đại học Sư phạm HCM	2.64%	2510

c. Những lỗi có tỷ lệ cao nhất

STT	*Lỗi*	*Tỷ lệ lỗi*	*Số mẫu*
1	soi mói	56.92%	130
2	cọ sát	42.59%	479
3	thăm quan	33.89%	3503
4	sáng lạn	20.55%	73

3.5 Kết quả tiêu biểu của lĩnh vực báo chí và truyền thông

Báo chí, nhà xuất bản và cơ quan truyền thông là các tổ chức dễ bị tổn thương nhất bởi lỗi chính tả. Do đó, việc khu vực này có tỷ lệ lỗi cao nhất cũng là dễ hiểu. Tuy nhiên, cần lưu ý do khu vực này có tác động lớn nhất đối với xã hội, vấn đề chính tả trên phương tiện truyền thông cần phải được đặc biệt chú trọng. Hơn nữa, các văn bản trong khu vực này chủ yếu là sản phẩm của các nhà báo với đội ngũ biên tập viên chuyên nghiệp, việc có các đơn vị có tỷ lệ lỗi cao trên 20% dù thế nào đi nữa cũng không thể chấp nhận.

Đặc biệt, Đài tiếng nói Việt nam là cơ quan truyền thông lớn thuộc Chính phủ, có tỷ lệ lỗi hơn 30% đứng đầu về tỷ lệ lỗi. Các báo điện tử VNExpress, Báo điện tử 24h và Việt báo cần phải có nỗ lực lớn để hạn chế lỗi. Việt báo USA, là một tờ báo của người Việt hải ngoại có mức lỗi 21.15%, do điều kiện để đảm bảo chất lượng văn bản ở hải ngoại có nhiều khó khăn.

Có thể thấy rằng các báo lớn có truyền thống lâu đời có tỷ lệ lỗi tương đối thấp hơn so với các báo mới ra đời. Đặc biệt. Nhà Xuất bản Chính trị Quốc gia, báo An ninh Hải

Phòng, có tỷ lệ lỗi chính tả đạt mức dưới 1%. Tạp chí Cộng sản cũng đạt được gần với mức này.

Đáng chú ý, theo tiêu chuẩn sử dụng của khu vực này, từ "xoi mói" trong từ điển đã trở thành từ sai chính tả do không đạt mức 30%.

a. Các đơn vị truyền thông có nhiều lỗi nhất

STT	*Báo*	*Tỷ lệ lỗi*	*Số mẫu*
1	Đài Tiếng nói Việt Nam	30.15%	2040
2	VNExpress	28.40%	17740
3	Việt báo USA	21.15%	10211
4	Báo điện tử 24h	21.00%	16093
5	Việt báo Việt Nam	19.85%	20017

b. Các đơn vị truyền thông có ít lỗi nhất

STT	*Báo chí, Nhà Xuất bản*	*Tỷ lệ lỗi*	*Số mẫu*
1	Nhà Xuất bản Chính trị Quốc gia	0.49%	1432
2	Báo An ninh Hải Phòng	0.89%	3033
3	Tạp chí Cộng sản	1.05%	4296
4	Báo Sài Gòn Tiếp thị	1.43%	8576
5	Báo Đầu tư	1.66%	1203

c. Những lỗi có tỷ lệ cao nhất

STT	*Lỗi*	*Tỷ lệ lỗi*	*Số mẫu*
1	soi mói	76.07%	5812
2	sáng lạn	43.54%	2905
3	cọ sát	25.35%	7499

4	thăm quan	19.77%	22644

3.6 Đánh giá chung toàn xã hội

Trong số các đơn vị có lỗi cao nhất có 3 đơn vị có tỷ lệ lỗi vượt 30%. Đặc biệt, Cục Vệ sinh An toàn Thực phẩm của Bộ Y tế có tỷ lệ lỗi gây choáng váng cho các chuyên gia, tới 38.46%. Các đơn vị có tỷ lệ lỗi cao nhất lại nằm trong các khu vực được chờ đợi là phải có chất lượng văn bản cao như Cơ quan nhà nước, Trường Đại học - Viện nghiên cứu và Báo chí – Truyền thông.

Năm đơn vị có chất lượng chính tả cao nhất đều đạt yêu cầu tỷ lệ lỗi dưới 1%. Đặc biệt, trong số này có tới 3 ngân hàng, chứng tỏ yêu cầu chất lượng văn bản của ngành Tài chính-Ngân hàng rất cao, mặc dù bộ quản lý ngành chưa chú trọng tới vấn đề này.

Theo thống kê tổng thể toàn xã hội và dựa trên các tiêu chí chuyên gia đã đề nghị, từ "**soi** mói" đã trở thành từ đúng với tỷ lệ sử dụng lên hơn 74%, "**sáng** lạn" có thể xem như một cách viết tương đương với "xán lạn" do đạt tới tỷ lệ sử dụng gần 42%. Các lỗi "cọ sát" và "thăm quan" đều đã đạt đến mức độ báo động đỏ.

a. Các đơn vị có lỗi nhiều nhất

STT	*Các đơn vị nhiều lỗi nhất*	*Tỷ lệ lỗi*	*Số mẫu*
1	Cục Vệ sinh An toàn Thực phẩm	38.46%	9758
2	Viện Năng lượng Nguyên tử	31.49%	1826
3	Đài Tiếng nói Việt Nam	30.15%	2040
4	VNExpress	28.40%	17740
5	Đại học Đà Nẵng	21.67%	2576

b. Các đơn vị có lỗi ít nhất

STT	Đơn vị tiêu biểu	Tỷ lệ lỗi	Số mẫu
1	Ngân hàng ACB	0.34%	1490
2	Nhà Xuất bản Chính trị Quốc gia	0.49%	1432
3	Ngân hàng BIDV	0.50%	1004

4	Ngân hàng Nhà nước	0.81%	5064
5	Báo An ninh Hải Phòng	0.89%	3033

c. Các lỗi có tỷ lệ cao nhất

STT	*Lỗi*	*Tỷ lệ lỗi*	*Số mẫu*
1	soi mói	74.33%	6742
2	sáng lạn	41.66%	3725
3	cọ sát	28.38%	10011
4	thăm quan	20.61%	48548

CHƯƠNG 4: GIỚI THIỆU CÁC CÔNG CỤ SOÁT LỖI TIẾNG VIỆT

Trong chiến dịch làm sạch chính tả tiếng Việt, các công cụ phần mềm soát lỗi có một vai trò quan trọng. Tuy việc đánh giá văn bản không sử dụng đến các công cụ này, nhóm chuyên gia vẫn thấy cần thiết phải giới thiệu các sản phẩm này một cách khách quan với cộng đồng.

4.1 Một số khái niệm về soát lỗi chính tả

Phần mềm soát lỗi Chương trình phần mềm giúp người dùng phát hiện ra lỗi chính tả trong văn bản và gợi ý cách sửa. Do tính không đơn nghĩa trong văn bản, mỗi lỗi có thể có nhiều gợi ý sửa khác nhau, phụ thuộc vào văn cảnh hoặc cá nhân người sử dụng.

Hạ tầng soát lỗi Một giải pháp cơ quan xí nghiệp gồm các phần mềm soát lỗi cài đặt trên máy trạm được kết nối với các máy chủ có năng lực soát lỗi sâu hơn, tốc độ nhanh và được cài các tập lỗi mới do cơ quan xí nghiệp quy định để thống nhất chính tả nội bộ.

Tỷ lệ lỗi phi từ và lỗi thực từ T Kết quả nghiên cứu của chúng tôi cho thấy tỷ lệ giữa lỗi phi từ và thực từ trong tiếng Việt T=31.70%:68.30%. Các đặc trưng cho phần mềm soát lỗi đều sẽ được tính riêng cho lỗi phi từ và lỗi thực từ, sau đó sẽ lấy trung bình cộng với trọng số tính theo tỷ lệ này.

Độ nhận biết (recall) R là đặc trưng quan trọng hàng đầu đối với chất lượng soát lỗi, đo khả năng nhận biết lỗi của phần mềm. R cho mỗi loại lỗi bằng tổng số lỗi mỗi loại do phần mềm soát lỗi nhận biết được chia cho tổng số lỗi loại đó có trong văn bản. Đối với lỗi phi từ các phần mềm soát lỗi thông thường nhất đều có độ nhận biết R=1. Đối với lỗi thực

từ, Microsoft Word 2007 mới đạt độ nhận biết 22.5-25.6% [1] là mức mà một số phần mềm soát lỗi tiếng Việt đã vượt qua khá xa.

Độ chính xác (precision) P đặc trưng cho mức độ chính xác của việc báo lỗi. P của một loại lỗi là số lần báo đúng chia cho tổng số lần báo lỗi đối với loại lỗi đó. Cần lưu ý rằng khi văn bản có nhiều lỗi mà phần mềm chỉ phát hiện được một lỗi với độ chính xác 100% cũng không có ý nghĩa gì. Các phần mềm chỉ sửa lỗi phi từ thường có độ chính xác cao chính vì như vậy.

Độ gợi ý đúng S đặc trưng cho khả năng gợi ý sửa lỗi của phần mềm. Chất lượng gợi ý của phần mềm phụ thuộc vào cách sắp xếp ưu tiên các gợi ý. Chúng tôi đề nghị, với mỗi lần gợi ý, các phương án gợi ý đúng xếp vị trí số 1, 2, 3, 4, 5 sẽ được tính điểm gợi ý lần lượt là 1, 0.8, 0.6, 0.4 và 0.2. Các gợi ý ở vị trí thứ 5 trở lên hoặc không có gợi ý đều không tính điểm do quá khó khăn trong việc lựa chọn của người sử dụng. S cho mỗi loại lỗi sẽ bằng tổng số điểm gợi ý của loại lỗi đó chia cho số lần gợi ý loại lỗi đó.

Độ đo VIE Phần mềm soát lỗi có chất lượng cao phải dung hòa được cả ba yêu cầu nói trên về chất lượng. Chúng tôi đề nghị sử dụng độ đo VIE là trung bình điều hòa của cả ba độ đo trên để đánh giá: Độ đo VIE= $3/(1/R + 1/P + 1/S)$ thay thế độ đo F1 mà nhiều tác giả trên thế giới hay sử dụng, nhưng bỏ qua độ gợi ý đúng.

4.2 Nhận thức về hiệu năng của các phần mềm soát lỗi tiếng Việt

Đa số người sử dụng Việt nam biết chút ít tiếng Anh đều biết đến chức năng soát lỗi chính tả của Microsoft Word và đều đánh giá cao ích lợi của việc sử dụng chức năng này. Họ không hề biết rằng hiện đã có các phần mềm soát lỗi chính tả tiếng Việt có hiệu năng đã vượt xa chức năng này của Microsoft Word.

Tình trạng nói trên có thể bắt nguồn từ nhận thức chủ quan của một số chuyên gia công nghệ thông tin, dựa trên kết luận của một số chuyên gia quốc tế trong tiếng Anh có hơn 80% số lỗi văn bản là lỗi phi từ. Vì vậy, họ chủ trương xây dựng các phần mềm chỉ sửa lỗi phi từ, và để lại các lỗi thực từ như "bổ xung", "xử dụng", "sử lý", "thăm quan",… Chúng tôi cho rằng, do đặc trưng đơn âm vị, tiếng Việt có tỷ lệ lỗi chính tả thực từ cao hơn rất nhiều.

Để có bằng chứng khách quan về vấn đề này, nhóm tác giả đã tiến hành nghiên cứu thống kê các nhóm khối liệu được trích ngẫu nhiên từ kho khối liệu tiếng Việt thô hơn 20 triệu câu chưa qua xử lý chính tả của VIEGRID JSC được thu thập từ nhiều nguồn khác nhau. Bằng cách lấy một số mẫu có quy mô 20 nghìn câu và dùng phép ngoại suy, nhóm tác giả đã thấy được tỷ lệ giữa lỗi phi từ và lỗi thực từ trong tiếng Việt là 31.70 %:68.30%. Đối với những người soạn thảo văn bản chịu ít áp lực về tốc độ so với báo chí, tỷ lệ này còn thấp hơn nhiều. Như vậy, có thể kết luận lỗi thực từ chiếm phần lớn lỗi chính tả trong

các văn bản tiếng Việt. Như vậy, hiệu năng phần mềm soát lỗi tiếng Việt chủ yếu là do năng lực soát và sửa lỗi thực từ quyết định.

Nhìn chung, các phần mềm soát lỗi tiếng Anh và tiếng Việt hiện nay đều có thể phát hiện tới 100% lỗi phi từ. Microsoft Word 2007, theo đánh giá của chuyên gia quốc tế chỉ có thể phát hiện tới 33% lỗi thực từ trong văn các bản tiếng Anh. Trong khi đó, một số phần mềm soát lỗi tiếng Việt có khả năng phát hiện lỗi thực từ tiếng Việt như Công Cụ Việt, Cọp Con, Cú Mèo, VietSpell,… đều vượt xa mức này.

4.3 Các phần mềm soát lỗi tiếng Việt

Sau đây là một số phần mềm soát lỗi tiếng Việt có trang thông tin hỗ trợ và có thể tải về dùng thử. Một số phần mềm khác tuy có tiếng tăm, nhưng không có đủ thông tin để đánh giá, nhóm tác giả đành tạm thời không giới thiệu trong báo cáo này. Một tình trạng đáng báo động là đa số các phần mềm soát lỗi đều đã ngừng phát triển.

Sản phẩm	*Nhà phát triển*	*Tình trạng*	*Dung lượng*	*Tính chất*	*Site*
Công Cụ Việt 1.4	VIEGRID JSC	Đang phát triển	177MB	Thương mại, cho dùng thử	www.viegrid.com
Cọp Con 3.1	Mai Tuấn Khôi	Ngừng phát triển	40MB	Miễn phí, ngừng cung cấp	chinhta.bacthangban.com
Cú Mèo Pro 2.0.2	SOBIC	Ngừng phát triển	8.32 MB	Thương mại, ngừng cung cấp	www.sobic.com.vn
VCatSpell	TTX Công giáo VN	Ngừng phát triển	0.8MB	Miễn phí, cho tải	www.vietcatholic.net
Vietkey for Office	Vietkey Group	Ngừng phát triển	6MB	Thương mại	www.vietkey.net
VietSpell	Lưu Hà Xuyên	Ngừng phát triển	1.01 MB	Thương mại, dùng thử hạn chế	N/A

4. 4 Các đặc trưng về hiệu năng của các phần mềm soát lỗi tiếng Việt

Các đặc trưng về hiệu năng là các đặc trưng có thể đo định lượng về chất lượng công nghệ của phần mềm soát lỗi. Việc nâng các đặc trưng hiệu năng lên một vài phần trăm cũng đòi hỏi những nỗ lực công nghệ và nghiên cứu rất lớn. Đôi khi chính nỗ lực đó sẽ quyết định tới lựa chọn của người dùng.

Nhóm tác giả tính các đặc trưng hiệu năng trên cơ sở các đặc trưng tính riêng rẽ cho lỗi phi từ và lỗi thực từ như sau:

Độ nhận biết: R = 31.70% x R(phi từ) + 68.30% x R(thực từ)

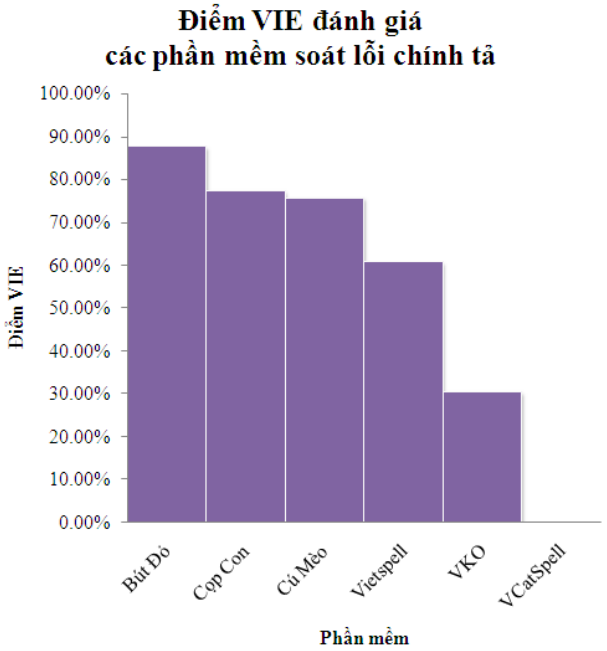
Độ chuẩn xác: P= 31.70% x P(phi từ) + 68.30% x P(thực từ)

Năng lực gợi ý: S= 31.70% x S(phi từ) + 68.30% x S(thực từ)

Nhóm tác giả để nghị dùng độ đo VIE = 3/(1/R+1/P+1/S) là trung bình của ba đặc trưng này để đánh giá các phần mềm soát lỗi. Kết quả đánh giá cuối cùng như sau:

	Công Cụ Việt (Bút Đỏ)	Cọp Con	Cú Mèo	VietSpell	VKO	VCatSpell
Độ nhận biết R	92.50%	67.77%	62.65%	62.76%	26.26%	26.52%
Độ chuẩn xác P	87.51%	87.61%	95.82%	69.81%	96.25%	22.94%
Năng lực gợi ý S	83.24%	78.85%	75.42%	51.96%	19.79%	0%
VIE	87.59%	77.21%	75.65%	60.60%	30.30%	0%

Độ đo VIE cho các phần mềm soát lỗi chính tả tiếng Việt được mô tả trong biểu đồ sau



Biểu đồ 1 *Đánh giá chất lượng soát lỗi của các phần mềm kiểm tra chính tả*

4.5 Một số tính năng người sử dụng và đặc trưng khác

Thực tế cho thấy thành công của một phần mềm không chỉ phụ thuộc vào những yếu tố thuần túy công nghệ. Các tính năng người sử dụng và các đặc trưng khác nhiều khi ảnh hưởng mạnh hơn đến người sử dụng và giúp các phần mềm soát lỗi tiếng Việt đi vào cuộc sống dễ dàng hơn.

Các tính năng người sử dụng sẽ giúp người sử dụng lựa chọn phần mềm theo yêu cầu công việc của mình. Một số tính năng quan trọng của các phần mềm soát lỗi tiếng Việt được thống kê trong bảng sau:

Các tính năng \ Các phần mềm	Công Cụ Việt	Cọp Con	Cú Mèo	VietSpell	VKO	VCátpell
Kiểm tra lỗi phi từ	√	√	√	√	√	√
Kiểm tra lỗi thực từ	√	√	√	√		
Kiểm tra viết Hoa đầu câu, chuẩn dấu cách và dấu thanh	√	√	√	√		
Chuẩn hóa y/i	√					
Gợi ý từ thay thế chính tả	√	√	√	√		
Soát lỗi chính tả theo chế độ tương tác	√	√	√	√	√	
Soát lỗi chính tả theo chế độ tự động	√					√
Tích hợp với phần mềm chuyển mã văn bản	√		√	√		
Báo trước bảng mã	√					
Tự động phát hiện và chuyển nhiều bảng mã trong văn bản	√					
Tích hợp Từ điển chính tả Việt	√					
Công cụ biên soạn từ điển người dùng	√	√				
Kết nối với máy chủ trong hạ tầng soát lỗi	√					
Có phiên bản quét lỗi với các định dạng khác (HTML)	√					

Các đặc trưng khác như tốc độ xử lý, mức độ sử dụng tài nguyên, mức độ thân thiện, độ ổn định, độ tùy biến, tính tích hợp, hỗ trợ và hướng dẫn, đóng gói, nhãn mác, phương thức phân phối, quảng bá,… là những yếu tố cũng có thể ảnh hưởng lớn tới ấn tượng của người

dùng đối với phần mềm. Các phần mềm soát lỗi Việt nam, đa số là do các nhà khoa học thiết kế và phát triển, thường ít chú trọng tới các đặc trưng thương mại và công nghiệp như vậy. Trong tương lai, nhóm tác giả hy vọng sẽ được các nhà phát triển và người sử dụng hỗ trợ để đưa ra được các đánh giá khách quan về các đặc trưng này.

KẾT LUẬN

Trong đợt đánh giá này, do thời gian, nguồn lực và kinh nghiệm còn hạn chế, các kết quả chưa thể phản ánh đầy đủ năng lực thực sự của các tổ chức được đánh giá. Tuy vậy, chúng tôi hy vọng rằng, các con số dù sơ sài nhưng trung thực sẽ đem lại cho cộng đồng những hình dung tương đối cụ thể về tình trạng lỗi chính tả trong văn bản tiếng Việt. Đánh giá sơ bộ về các phần mềm soát lỗi tiếng Việt sẽ giúp các nhà phát triển nâng cao chất lượng sản phẩm để giúp ích cho cộng đồng. Mặt khác, kết quả đánh giá này cũng phần nào giải tỏa ấn tượng chưa tốt của người sử dụng đối với các phần mềm Việt nam. Với các đợt xếp hạng tiếp sau, việc đánh giá sẽ ngày càng chính xác hơn. Nhóm tác giả hy vọng rằng, với nỗ lực bền bỉ và liên tục của toàn xã hội, tiếng Việt của chúng ta sẽ sớm được chuẩn hóa, phát triển và ngày càng đẹp đẽ hơn.

TÀI LIỆU THAM KHẢO

[1] G. Hirst, http://ftp.cs.toronto.edu/pub/gh/Hirst-2008-Word.pdf University of Toronto (2007)