**Đặng Minh Tuấn**

Mã dựng sẵn và mã tổ hợp

Chào mừng các bạn đón đọc đầu sách từ dự án sách cho thiết bị di động

*Nguồn:* [*http://vnthuquan.net/*](http://vnthuquan.net/)
Tạo ebook: Nguyễn Kim Vỹ.

**MỤC LỤC**

[Mã dựng sẵn và mã tổ hợp](%22%20%5Cl%20%22bm2)

**Đặng Minh Tuấn**

Mã dựng sẵn và mã tổ hợp

Mã dựng sẵn và mã tổ hợp, bàn thêm về Unicode và TCVN 6909

**Dựng sẵn và tổ hợp**
Có hai dòng giải pháp chính để xử lý tiếng Việt trên máy tính là mã dựng sẵn và mã tổ hợp. Cả hai đều có những ưu và nhược điểm riêng và cũng là chủ đề gây nhiều tranh luận.Trong bài này chúng ta sẽ phân tích một cách có hệ thống khía cạnh kỹ thuật của cả hai giải pháp này, đồng thời cũng bàn thêm một số điểm khác xung quanh Unicode vàTCVN5909.
Trước tiên cần làm rõ các khái niệm ký tự dựng sẵn và ký tự tổ hợp. Ký tự dựng sẵn là ký tự duy nhất không được tổ hợp từ những ký tự khác. Ký tự tổ hợp là ký tự được tổ hợp từ những ký tự cơ bản, ví dụ chữ/ký tự "ừ" là tổ hợp của ký tự cơ bản và ký tự dấu thanh huyền ` .Trong Unicode, ký tự tổ hợp còn có thể được tổ hợp từ những thành phần nhỏ hơn như ký tự ASCII, ký tự dấu nguyên âm và ký tự dấu thanh- khi đó ta có đang biểu diễn chính tắc (canonical) .
Vídụ chữ "ừ" trong các dạng biễu diễn nói trên sẽ là chuỗi ký tự sau:

Tổ hợp ký tự
Chuỗi mã Hex
Dựng sẵn
ư
1EAB
Tổ hợp
ư +
01B0,0300
Chính tắc
u + +
0075,031B,0300

Mã dựng sẵn và mã tổ hợp là quá trình mã hóa các ký tự thành các ký tự dựng sẵn hay thành các ký tự tổ hợp. Trong bài này, ta sẽ coi Canonical cũng là một dạng của mã hoá tổ hợp.
**ƯU ĐIỂM CỦA MÃ TỔ HỢP**
Mã tổ hợp có phần gọn nhẹ và chiếm ít mã hơn trong bảng mã, chỉ cần 20 vị trí cho ký tự thuần Việt ( ă, â, ê, ô, ơ, ư, ă, â, Ê Ô Ơ ư, các dấu thanh : huyền, hói, ngã, sắc, nặng và dấu tổ hợp nguyên âm : nón, mũ, râu cho dạng canomcal) trong khi mã dựng sẵn cần đến 134 cho ký tự thuần Việt.
Mã tổ hợp có phần gần với ngôn ngữ tự nhiên (Việt) hơn trong quá trình ghép chữ, ghép vần.
Mã tổ hợp sẽ dễ dàng hơn trong việc chuyển đổi chữ hoa/chữ thường, trong một số ứng dụng có thể dùng luôn tính năng Change case có sẵn để chuyển đổi .
Mã tổ hợp có vẻ như dễ dàng hơn trong việc sắp xếp tiếng Việt , nhưng thực ra không phải như vậy, lý do là các dấu thanh huyền, sắc ngã, hỏi, nặng - thứ tự trong bảng mã Unicode - lại nằm không đúng theo thứ tự sắp xếp tiếng Việt là huyền, hỏi, ngăn, sắc, nặng, do đó vẫn phải thiết kế thuật toán riêng để sắp xếp mà không thể dùng các hàm có sẵn trong tiếng Anh. Khi đã phải dùng thuật toán riêng thì việc sắp xếp cho mã dựng sẵn cũng không khó hơn, không phức tạp nhiều hơn so với việc sắp xếp mã tổ hợp .
Mã tổ hợp có phần dễ dàng hơn trong việc tìm kiếm tiếng Việt gần đúng, ví dụ những chữ tiếng Việt gần với âm "tha" chẳng hạn, thì các hàm tìm kiếm phổ thông sẽ tìm ra được các chữ thà, thá, thả, thà. tha... Nhưng nếu tìm những từ gần với âm "than" thì lúc ấy lại phải thiết kế thuật toán riêng, mà khi đã phải dùng thuật toán riêng thì giữa tổ hợp và dựng sẵn thuật toán không khó hơn nhau nhiều.
Trong thực tế, mã tổ hợp được hỗ trợ tốt hơn trong môi trường Windows 2000, và bộ MSoffice2000, ý tốt hơn ở đây là chuyển đổi chữ hoa/thường, sắp xếp Tiếng Việt được thiết kế ngay trong hệ điều hành và một số ứng dụng. Mã tổ hợp có thể hiện thì tốt hơn trong một số control có sẵn của Windows 2000,XP. Với WindowsXP, Microsoft đã hỗ trợ luôn cả mã dựng sẵn với tính năng sắp xếp tiếng Việt. Vietkey Group đã phối hợp với Microsoft việt Nam tiến hành một loạt thí nghiệm trên Window và rút ra kết luận như vậy. Các kết luận này không đúng cho Windows 95 và Windows 98 .
**NHƯỢC ĐIỂM CỦA BỘ MÃ TỔ HỢP**
Cài đặt mã tổ hợp khá phức tạp, số lượng môi trường cài đặt bị hạn chế hơn nhiều so với mã dựng sẵn, thông thường chỉ cài đặt được với font vector và bộ font cho phép định nghĩa các ký tự có độ rộng âm, khi đó 2 ký tự có độ rộng âm và dương tổ hợp lại sẽ cho ra ký tự cần hiển thị. Một khó khăn khá lớn nữa là phần lớn các công nghệ font phổ biến ngày nay như truetype,Opentype Typel...không cho phép thay đổi động vị trí nét trong hình chưa mà điều này lại rất cần thiết. Ví dụ chữ "à" , và "À" , thì vị trí của dấu huyền phải nằm ở 2 cao độ khác nhau tuỳ theo chữ cái cơ sở là chữ thường hay chữ hoa việc thay đổi động cao độ của dấu than theo ngữ cảnh là chưa thực hiện được bằng kỹ thuật font chữ hiện hành. Để khắc phục vấn đề này, VNI đã phải đề xuất 2 mã riêng cho từng dấu thanh : 2 dấu huyền, một mã cho chữ hoa và một mã cho chữ thường. Trong CP1258 và Unicode để đảm bảo tính đơn trị (tính một- một) các dấu thanh chỉ có một mã vì thế sẽ rất khó khăn trong hiển thị
Phương án thứ 2 mà Microsoft đưa ra để giải quyết vấn đề tăng giảm độ cao dấu thanh là dùng kỹ thuật Hook API thay đổi các hàm Display qua đó ánh xạ (Map) chuỗi ký tự tổ hợp về chuỗi ký tự dựng sẵn để hiển thị và in ấn.
Cơ chế này chỉ có trong windows95 tiếng Việt, Windows 2000, WindowsXP mà không có trong windows95, Windown98. Cơ chế này không phải bao giờ cũng thực hiện tốt, ngay cả trên Windows XP.
Từ việc cài đặt mã tổ hợp phức tạp như vậy dẫn đến một nhược điểm thứ hai khá nghiêm trọng, đó là tính tương thích của mã tổ hợp kém hơn. Có nghĩa là một văn bảng bằng mã tổ hợp ở môi trường này có thể không đọc được trong môi trường khác. Nhất là khi dùng font bitmap để làm font hệ thống thì hầu như không thể cài đặt được mã tổ hợp, cũng như trong dos,textconsole và nhiều môi trường unix,Linux. Mã tổ hợp cài đặt trên các hệ điều hành phổ biến hiện nay là Windows98 thì chữ rất xấu không thể chấp nhận được. Sẽ là một vấn đề lớn khi dùng mã tổ hợp phải nâng cấp phần cứng máy tính lên Windows 2000, XP (theo khuyến cáo của Microsoft để chạy mã tổ hợp tốt hơn) , như vậy sẽ cần kinh phí rất lớn để nâng cấp, đào tạo lại. . . .
Độ mỹ thuật của mã tổ hợp thường kém hơn nhiều so với mã dựng sẵn, lý do là một ký tự dấu thanh có vị trí và cao độ xác định trong font chữ thường được dùng chung cho nhiều nguyên âm khác nhau và chúng được tổ hợp tự động sau khi nhập đoạn text. Vị trí của dấu thanh có thể hợp và đẹp với nguyên âm này nhưng lại có thể không phù hợp với nguyên âm khác . Ví dụ độ rộng của nguyên âm A thì khác với độ rộng của nguyên âm I (độ rộng rất hẹp) vì thế nếu đẹp cho chữ A thì xấu cho chữ I và ngược lại, để khắc phục tình trạng này VNI phải định nghĩa riêng các mã cho các chữ ì, ì, ĩ, ị. Vì vậy chúng ta thường coi VNI là giải pháp khắc phục tình thế hơn là một bộ mã, vì nó không đảm bảo tính đơn trị, và nhất quán (có nhiều mã cho một dấu - với chữ lại có xử lý khác so với xử lý các nguyên âm khác). Trong khi đó mã dựng sẵn được thiết kế từ trước (dựng sẵn) nên có thể bố trí vị trí dấu thanh nhờ vào vị trí thích hợp nhất cho từng nguyên âm, nên bao giờ cũng có khả năng đẹp hơn nhiều so với mã tổ hợp .
Xử lý hiệu ứng với đoạn mã tổ hợp có nhiều vấn đề khó khăn hơn so với mã dựng sẵn. Trong nhiều trường hợp, một con chữ tiếng Việt trong lưu trữ và hiển thị với mã dựng sẵn lại không phải là một thể thống nhất (tổ hợp từ những ký tự rời rạc) cho nên khi thực hiện các hiệu ứng với đoạn văn bản như co dãn text, xoay, dồn chữ, canh đều hai bên. . . thì các dấu và chữ thường bị tách rời nhau, chữ đi một nơi và dấu đi một nơi, ảnh hưởng đến mỹ thuật và độ chính xác. Có thể thấy trên các tít báo dùng font VNI hay xuất hiện các hiện tượng xa rời dấu thanh.
Xử lý với các ký tự mã tổ hợp phức tạp hơn so với mã dựng sẵn, do mỗi chữ cái trong mã tổ hợp có độ rộng thay đổi, lúc có thể là một ký tự, lúc khác lại được tổ hợp từ nhiều byte khác nhau. Khi tách từ, tách ký tự (theo ngôn ngữ tự nhiên), thường dùng để phân tích cú pháp hay đánh chỉ số phải xây dựng thuật toán riêng khá phức tạp, trong khi mã dựng sẵn có độ rộng cố định nên việc rút ký tự từ đoạn text ra rất đơn giản, không cần xây dựng thuật toán riêng. Ngoài ra việc xử lý ký tự khác như: xóa ký tự, di chuyển cho trỏ di theo đơn vị ký tự thì thực hiện với mã tổ hợp khó khăn và phức tạp hơn : thường phải xóa 2 sẵn cho một chữ, di chuyển 2 lần con trỏ mới đi ra khỏi một chữ, điều này là xa lạ với ngôn ngữ tự nhiên.
Kích thước các tệp dữ liệu lưu ở dạng tổ hợp thường lớn hơn so với mã dựng sẵn khoảng 25-30% do đó nó chiếm nhiều không gian trong đĩa cứng, bộ nhớ hơn, và trên đường truyền mạng (intemet/intranet) ngốn nhiều băng thông hơn.
Trong cơ sở dữ liệu, thiết kế cấu trúc CSDL với mã tổ hợp thường phức tạp hơn. Vì mặc dù biết trước số chữ cái max nhưng lại khó đoán nhận chính xác độ dài chuỗi byte tương ứng lớn nhất, nếu thiết kế không khéo sẽ bị tràn bộ nhớ. Và khó khăn trong việc phân tách ký tự, phân tách từ cũng làm khó khăn thêm trong việc xử lý text trong lĩnh vực cơ sở dữ liệu.
Trong việc đánh chỉ số (index) , và tìm kiếm toàn văn (fulltextsearch), mà tổ hợp cũng gây nhiều khó khăn hơn (phân tách từ, phân tách ký tự) và các ký tự dấu thanh trong mã tổ hợp thường bị coi là dấu phân cách từ, dẫn đến việc đánh index bị sai và tìm kiếm toàn văn cũng không đúng. Hiện tượng này thường gặp với đa số các bộ search engine, công cụ tìm kiếm toàn văn trong Oracle và Lotus Notes 5.0 đều bị lỗi đánh chỉ số sai. Tuy rằng Oracle và Lotus Notes đều hỗ trợ Unicode trong phần encoding, nhưng đáng tiếc phần Full Text Search mua lại của hãng thứ 3 INSO và Verity đều thực hiện các phân cách từ sai với mã tổ hợp . Nhưng lỗi này không xảy ra với mã dựng sẵn.
Tính thực tế của mã tổ hợp kém hơn mã dựng sẵn : đa số ở Việt Nam cùng như ở nước ngoài, Unicode dựng sẵn được dùng rất phổ biến, các website của việt Nam như Vnexpress và VASCorient hàng ngày có gần 2 triệu lượt truy nhập chứng tỏ số lượng người dùng mã dựng sẵn rất lớn, trong khi các website dùng mã tổ hợp rất ít.
**ƯU VÀ NHƯỢC CỦA MÃ DỰNG SẴN**
Những ưu và nhược điểm của mã dựng sẵn cũng đã được phân tích khá kỹ và song song trong quá trình phân tích các ưu và nhược điểm của mã tổ hợp. Tôi chỉ xin được tổng kết lại một cách tóm tắt: mã dựng sẵn không bị những nhược điểm của mã tổ hợp, các file sử dụng mã dựng sẵn tốn ít không gian nhớ, cài đặt đơn giản hơn, chữ hiển thị đẹp hơn, mỹ thuật hơn, xử lý với xâu ký tự dựng sẵn dễ dàng hơn (phân tách từ, phân tách ký tự, xóa và di chuyển con trỏ) , chữ và dấu là một khối thống nhất nên khi co dãn text không bị hiện tượng xa rời dấu thanh, và đặc biệt tính tương thích của mã dựng sẵn cao hơn có thể chạy được trên nhiều môi trường khác nhau ( từ font vector đến bitmap, từ Windows, Macintosh đến Linux), vì chuyển mang văn bản dữ liệu giữa các môi trường không đồng nhất (Multiplatform) là một điều cũng rất cần thiết.
Tất cả những ưu điểm của mã tổ hợp, thì lại không phải là căn bản vả đều có thể giải quyết được tương đối khá dễ dàng, trong khi đó mã tổ hợp có những vấn đề về kỹ thuật khá phức tạp (cài đặt hiển thị...). Vẫn biết kỹ thuật phải theo chuẩn, nhưng tại thời điểm hiện nay có những giới hạn nhất định về kỹ thuật và công nghệ ví dụ cách đây 30 năm, mã hóa 8 bit là giới hạn chưa giải quyết được do không gian nhớ eo hẹp. . . ngày nay cài đặt mã tổ hợp trong nhiều môi trường chưa thể làm được hoặc phải làm rất khó khăn và đi vòng vo như ánh xạ từ tổ hợp về dựng sẵn.
Các nhược điểm của mã tổ hợp có thể khắc phục được một phần ở thời điểm hiện nay và có thể khắc phục hoàn toàn trong tương lai, nhưng theo ý kiến chúng tôi, chúng ta nên chọn những phương án đơn giản, dễ cài đặt, đẹp hơn...và nhiều ưu điểm ở trên là mã dựng sẵn bởi vì cái lợi của mã tổ hợp mang lại thì không đáng kể mà để thực hiện hoàn hảo nó thì lại có quá nhiều khó khăn. Chân lý nhiều khi lại ở chính sự đơn giản của vấn đề.
Cái có thể nói nhược điểm của mã dựng sẵn hiện nay là chưa được Microsoft chú ý hỗ trợ nhiều. Trong thời gian trước Microsoft có thiên hướng đi theo mà tổ hợp ( là phần nâng cấp từ mã 8-bit CP1258 lên). Những xử lý về tiếng Việt như sắp xếp, chuyển đổi chữ hoa/thường... được tích hợp luôn vào hệ điều hành như thế tốt hơn so với mã dựng sẵn, nhưng điều này chỉ đúng với Windows2000 ( không đúng với windows 95, 98 ), ngoài ra trên WindowsXP, Microsoft đã bắt đầu hỗ trợ thêm mã dựng sẵn vào trong hệ điều hành: phần sắp xếp tiếng Việt và việc trong tương lai Microsoft sẽ hỗ trợ hoàn toàn mã dựng sẵn vào hệ điều hành không phải là vấn đề phức tạp về kỹ thuật.
Ngoài ra trước kia với các mã VNI, ABC-TCVN 5712, chúng ta đâu có được Microsoft hỗ trợ tiếng Việt ở mức hệ điều hành, nhưng vẫn không bị ảnh hưởng máy, CNTT vẫn tiến triển tốt. Thêm nữa hiện đã có những hỗ trợ tiếng Việt ở mức thấp API (giống như Windows) do các đơn vị ở trong nước thực hiện và đáp ứng đầy đủ các yêu cầu về xử lý tiếng Việt ở tầng thấp. VietkeyGroup đã xây dựng bộ thư viện Vietkey API cho phép xử lý hầu hết các vấn đề về tiếng Việt: tìm kiếm, sắp xếp, chuyển đổi chữ hoa, chữ thường, chuyển mã, và cả kiểm tra chính tả tiếng Việt ở mức hệ thống, có thể nhúng trong rất nhiều ứng dụng và các công cụ lập trình trong môi trường Windows. Vietkey API xử lý tiếng Việt ở một số góc cạnh còn đầy đủ và hoàn hảo hơn nhưng hỗ trợ từ mức hệ điều hành của Windows.
Tóm lại mã dựng sẵn với nhiều ưu điểm nổi trôi có thể đáp ứng được nhu cầu trước mắt cũng như lâu dài và có thể sử dụng trong nhiều lĩnh vực : chế bản văn phòng, web, email, cơ sở dữ liệu và nhiều lĩnh vực khác, theo ý kiến của chúng tôi nên chọn mã dựng sẵn, và tổ hợp chỉ có thể dùng trong một số bài toán đặc thù như phân tích chuỗi ký tự hay để biểu diễn các hình chữ cho dấu thanh.
**UNICODE VÀ TCVN 6909**
Trước tiên cần phải nêu rõ mối quan hệ giữa Unicode và TCVN6909. TCVN6909 là một tập con của Unicode, nhưng đã được chọn lọc kỹ lưỡng để lấy ra đúng những ký tự sẽ dùng trong ngôn ngữ Việt. TCVN6909 quy định rõ ràng hơn và tính đơn trị một-một cao hơn trong Unicode. Ví dụ trong Unicode có đến 3 chữ Đ , 8 dấu nặng . nếu không quy định rõ sẽ gây nhầm lẫn, mà thực tế đã xảy ra nhầm lẫn chữ Đ trong bộ gõ Unicode : Vps của Việt Kiều Pháp. TCVN 6909 chứa tất cả các ký tự dựng sẵn và cùng có đầy đủ các mã để hiển thị cho dấu thanh, mã cho việc biểu diễn các dấu tổ hợp nguyên âm, như vậy TCVN 6909 là bộ mã rất đầy đủ ( trong đó có mã cho tất cả các thành phần tiếng Việt). TCVN 6909 không quy định biểu diễn dựng sẵn hay tổ hợp. Vì vậy nói TCVN6909 là mã dựng sẵn là không chính xác .
TCVN 6909 kế thừa các ưu điểm của bộ mã Unicode và ISO10646, cho nên nó hoàn toàn cho phép hội nhập tiếng Việt với các ngôn ngữ khác trong cộng đồng Unicode. Hiện nay đa số các ứng dụng phổ thông đều đa hỗ trợ Unicode : phần mềm văn phòng, cơ sở dữ liệu, web, e-mail...tuy nhiên cũng còn một số các phần mềm chuyên ngành chưa hỗ trợ Unicode.
Vấn đề kỹ thuật của Unicode khôngchỉ là bộ gõ Unicode (như một số người đã lầm tưởng) mà là một loạt các vấn đề khác, đặc biệt là các công cụ chuyển mã cho văn bản dữ liệu (MSoffice,Webpages,Database...) chuyển mã là vấn đề lớn hơn nhiều, vì khối lượng dữ liệu cần chuyển đổi sang Unicode rất lớn, đa dạng về khuôn thức loại hình, chủng loại. Ngoài ra còn các vấn đề sắp xếp, chuyển đổi chữ hoa, chữ thường, tìm kiếm toàn văn, kiểm tra chính tả tiếng Việt, đặc biệt hỗ trợ ở mức lập trình cho các công cụ phát triển cũng là điều rất cần thiết (hiển thị, lưu trữ.. ).
Vietkey Group trong những năm qua là một trong những đơn vị đi tiên phong trong vấn đề xử lý tiếng Việt Unicode, không chỉ có bộ gõ đa ngữ Unicode Vietkey2000 mà còn có các bộ công cụ khác như Vietkey Office tích hợp và biến bộ MS Office chuẩn thành bộ Office có những tính năng xử lý tiếng Việt (kiểm tra chính tả, sắp xếp, chuyển chữ hoa/thường cho tiếng Việt chuyển mã...). Trong bộ công cụ của Vietkey còn có bộ chuyển mã cho toàn bộ website, CSDL Oracle, SQL Server, Access... sang Unicode.
Về hỗ trợ tiếng Việt ở mức hệ thống, Vietkey Group còn có bộ thư viện Vietkey API cho phép so sánh xâu ký tự Việt, chuyển đổi chữ hoa/thường, sắp xếp tiếng Việt chuyển mã và kiểm tra chính tả tiếng Việt. Các công cụ này có thể nhúng được trong rất nhiều ứng dụng và các công cụ lập trình.
Hệ thống Vietkey Messenger là ATOL do VietkeyGroup phát triển hỗ trợ Unicode trong các môi trường, từ Windows 95 cho đến XP, tất cả mọi thành phần: Email, chat, Forum, FTP... đều được hiển thị và xử lý bằng tiếng Việt Unicode.
Cuối cùng, có thể nói chuyển đổi sang Unicode là một việc làm rất cần thiết và phải làm càng sớm càng tốt, vì càng để chậm, kho dữ liệu, các website càng ngày càng phát triển với dữ liệu càng lớn thì quá thì quá trình chuyển mã về sau càng phức tạp, càng tốn kém nhiều công sức và tiền của. Vấn đề Unicode, không chỉ là bộ gõ, mà phải nghĩ đó là một loạt các công cụ đằng sau chuyển mã và hỗ trợ tiếng Việt ở mức hệ thống, ứng dụng trong nhiều môi trường khác nhau.
Dưới đây là một minh hoạ cho sự lệch lạc dấu thanh của mã tổ hợp, dòng trên là soạn bằng mã tổ hợp, dòng dưới được soạn bằng mã dựng sẵn, được dùng cùng một font chữ Verdana, trong hệ điều hành Windows XP (là hệ điều hành mới nhất của Microsoft), và đoạn text được soạn trong MS PowerPoint 2000 (bộ soạn thảo văn phòng MS Office 2000), 2 dòng cuối được soạn bằng WordArt cũng trong Office 2000 và Windows XP. Tất cả các đoạn text trên chưa hề qua một hiệu ứng text nào , mà chữ và dấu thanh đã bị lệch và xa rời nhau trong khi mã dựng sẵn luôn hiển thị đúng và đẹp. Hiện tượng này còn bị phổ biến hơn với tất cả các ứng dụng (kể cả Word XP, Excel XP...) chạy trên hệ điều hành Windows 95,98

Lời cuối: Cám ơn bạn đã theo dõi hết cuốn truyện.
Nguồn: http://vnthuquan.net
Phát hành: Nguyễn Kim Vỹ.

Nguồn: Thái Nhi
Được bạn: Thành Viên VNthuquan đưa lên
vào ngày: 27 tháng 12 năm 2003