Tạp Chí Tin Học và Điều Khiển Học
Tập IX(1993), Số 2, 28-35

# Chữ Việt Trong Xử Lí và Truyền Thông

Ngô Trung Việt
Viện Tin Học
Viện Khoa Học Việt Nam

## 1. Mở đầu

Trong bài này chúng tôi đề cập đến các vấn đề sau:

- Các yêu cầu cơ bản cuả bộ mã chuẩn cho chữ Việt.
- Bộ mã chuẩn và việc biểu diễn đầy đủ cho chữ Việt.
- Bộ mã chuẩn trong quan hệ giữa chữ Việt và xử lí máy tính.
- Bộ mã chuẩn trong quan hệ giữa chữ Việt và truyền thông.
- Đề nghị cụ thể về bộ mã chuẩn.

## 2. Các yêu cầu cơ bản cuả bộ mã chuẩn cho chữ Việt

Đã có nhiều tranh luận về các yêu cầu cơ bản đối với bộ mã chuẩn chữ Việt dùng trong trao đổi và xử lí thông tin trên máy tính. Nhiều ý kiến nêu ra các yêu cầu có tính chất cục bộ hay quá kỹ thuật, không phù hợp với tính chiến lược, tổng quát cuả bộ mã. Tuy nhiên sau đó các ý kiến đã dần thống nhất lại theo những chủ đề chỉnh phản ánh khía cạnh nền tảng, cơ sở cuả việc ấn định ra bộ mã chuẩn:

- thể hiện đầy đủ chữ Việt.
- tương hợp với khả năng xử lí và ghi nhớ cuả máy tính.
- thích ứng với các quy tắc và phương tiện truyền thông hiện nay trong việc truyền các thông tin có chữ Việt.

Yêu cầu thứ nhất phản ánh đặc thù cuả chữ Việt mà bộ mã phải mô tả được. Các yêu cầu thứ hai và ba là các ràng buộc thực tế về phương tiện xử lí và truyền tin mà chữ Việt cũng chỉ là một loại thông tin được xử lí hay truyền thông qua đó.

Bộ mã chuẩn phải là nền tảng cho các quá trình đưa chữ Việt vào/ra máy tính, xử lí và truyền thông. Bộ mã chuẩn phải tạo điều kiện để có thể đưa được đầy đủ các thông tin trong các văn bản chữ Việt vào máy tính; có thể tái tạo lại trung thành các văn bản đó trên các thiết bị ra: màn hình, máy in; có thể thuận lợi cho việc dùng các phần mềm xử lí và ghi nhớ dạng văn bản chữ Việt trong máy tính và thuận tiện cho việc truyền các văn bản đó giữa các máy tính trên thế giới.

Do đặc thù về cách biểu diễn chữ Việt và đòi hỏi cuả khả năng kỹ thuật hiện thời, không có bộ mã chữ Việt nào đáp ứng được đầy đủ cho ba yêu cầu này. Tính không giải được này bắt nguồn từ chỗ không gian mã hóa 8 bít có hạn, mà tổng mã hóa cận xử lí, truyền thông và biểu diễn chữ Việt thì vượt quá giới hạn đó. Các bộ mã hiện có, do các nhóm nghiên cứu ở các nơi xây dựng ra, mới chỉ đáp ứng được nhu cầu cục bộ cuả từng nhóm người xử dụng nhỏ trong các ứng dụng đặc thù. Do vậy, để đáp ứng nhiều nhất đến mức có thể cho các yêu cầu này, cần phải tìm sự thoả hiệp trong cách bố trí bộ mã nhưng cũng phải duy trì khả năng mở rộng bộ mã cho những yêu cầu đặc biệt. Các đòi hỏi và hạn chế cuả khiá cạnh kĩ thuật sẽ bắt buộc việc xây dựng bộ mã chuẩn phải có một nền tảng thống nhất trên cơ sở phân tán theo ba hướng yêu cầu trên.

## 2. Bộ mã chuẩn và việc biểu diễn đầy đủ cho chữ Việt

Trong thực tế sử dụng tiếng Việt hàng ngày, trong cách viết thông thường, chúng ta hay tạo ra các từ tiếng Việt theo kiến trúc dấu rời: ghép các phụ âm, nguyên âm và dấu thanh thành các đơn vị từ. Do đó khi dùng máy tính để biển diễn cho tiếng Việt, theo thói quen, kiến trúc dấu rời có ảnh hưởng tới cách cấu thành biểu diễn cho kí tự Việt trong máy tính. Kiến trúc này đi theo hướng mã hoá cho các thành phần sẽ cấu tạo nên một dạng biểu diễn tổ hợp dấu thanh và nguyên âm dựa trên thành tố cuả bộ mã nào đó (ASCII chẳng hạn), sau đó dùng các thuật toán tái tạo để tạo dựng hình ảnh chính xác cuả kí tự mang dấu đó. Như vậy, để biểu diễn cho một chữ Việt, cần dùng từ một đến vài mã cùng với thuật toán tái tạo hình ảnh thật cuả con chữ đó mặc dầu dạng thật đó không có trong bộ mã. Phương án này còn được gọi là cách biểu diễn chữ Việt theo kiến trúc tái tạo hay mã hoá theo nhiều bai. Nhược điểm chính cuả phương án này là việc bộ mã không phản ánh được hình ảnh thực cuả con chữ Việt cùng với vị trí tương ứng cuả dấu. Mọi hình dạng thấy được cuả con chữ tiếng Việt đều phải là kết quả cuả quá trình xử lí máy tính. Như vậy cấu tạo bộ mã ngay từ đầu bị lệ thuộc vào người cài đặt hệ thống cụ thể, cấu hình máy tính cụ thể, do đó gây khó khăn cho việc chuyển chữ Việt lên các hệ thống máy tính khác. Quan trọng hơn nữa, việc mã hoá như vậy tự bản thân nó đã mang mầm mống không chuẩn, mang tính cách riêng cuả từng người thực hiện.

Tuy nhiên, do đặc trưng cuả kỹ thuật máy tính có nguồn gốc cuả tiếng Anh, mỗi kí tự mang một dạng hình (glyph) thể hiện riêng và gói gọn được trong bộ mã 7 bit ISO 646, nên có gây ra khó khăn khi ta định thể hiện chữ Việt trên máy tính. Khả năng kỹ thuật hiện tại cho phép ta đưa vào máy tính, ngoài các dạng con chữ Việt thông thường, cả các dạng nguyên âm có mang theo dấu. Với cách tạo ra các mẫu tự có sẵn có kèm các dấu thanh, xử lí trong máy tính sẽ thuận lợi hơn nhiều. Do đó xuất hiện kiểu kiến trúc dựng sẵn tất cả các dạng có thể có trong tổ hợp

dấu thanh và các nguyên âm chữ Việt, còn gọi là kiến trúc con chữ dựng sẵn, và đó là kiểu mã hoá theo một bai. Kiến trúc này có ưu điểm là phù hợp với xu hướng xây dựng các bộ mã trên thế giới: mỗi dạng biểu diễn khác biệt mang một mã riêng. Mặt khác, do hình thức xuất hiện của mọi dạng chữ Việt đều được ấn định chính xác trong bộ mã nên bản thân cách dựng mã là mang tính qui định chung, thống nhất độc lập với mọi xu hướng sử dụng. Nhược điểm của nó, như sẽ phân tích ở sau, là cần thêm 134 dạng biểu diễn cho các nguyên âm mang dấu, làm khó khăn cho việc truyền thông cũng như xử lí. Như vậy, chính kỹ thuật đã tạo ra một yêu cầu kiến trúc mới, bên cạnh kiến trúc tái tạo thông thường.

Sự khác biệt chủ yếu trong hai kiểu kiến trúc nói trên thể hiện ở thời gian và không gian của cách xử lí. Kiến trúc kiểu dấu rời cần có thời gian để tái tạo lại mang dạng hình mẫu tự theo các thông tin có sẵn về kí tự và dấu thanh. Nói cách khác ta cần có thuật toán để cấu thành mẫu tự. Và để dùng được phổ cập trên máy tính, kết hợp với các phần mềm khác được, thì các thuật toán này phải được cài đặt ở lớp sâu của hệ điều hành, nơi chuyển đổi các mã số vào ra thành dạng mẫu tự. Đây quả là một công việc khó khăn nếu ta không có cách nào can thiệp sâu ở mức các nhà sản xuất máy tính.

Kiến trúc kiểu dựng sẵn cần có không gian bộ nhớ lưu giữ sẵn các dạng mẫu tự có thể có của ngôn ngữ để khi cần đến mẫu tự nào thì chỉ việc cho thể hiện ra, không cần đi qua bất kỳ thủ tục sửa đổi nào. Với kiến trúc kiểu dựng sẵn, thích hợp cho xử lí bên trong máy tính, ta có thể dùng được nhiều loại phần mềm có bán trên thị trường, do đó đẩy nhanh việc phổ cập tin học. Nhược điểm của kiểu kiến trúc này là không thích hợp cho việc truyền thông quốc tế vì để mã hoá được hết các kí tự thì không có chỗ cho các kí tự điều khiển truyền in. Trái lại, kiến trúc kiểu rời dấu lại thích hợp cho việc truyền thông và không thuận tiện cho việc xử lí bên trong máy tính (do thường xuyên phải thực hiện các thuật toán tái tạo ở mức thấp). Việc mã hoá theo kiến trúc dấu rời rất thích hợp cho việc truyền thông nhưng không có lợi cho việc ghi nhớ và xử lí dữ liệu.

Vấn đề đặt ra là làm sao tận dụng được ưu thế của cách định mã của hai kiểu kiến trúc dấu rời và dựng sẵn và hạn chế được nhược điểm của riêng mỗi cách. Kiến trúc dấu rời, ngoài lợi thế cho truyền thông, do gần với thói quen, cho nên sẽ có ảnh hưởng rất nhiều tới cách bố việc đưa thông tin vào máy tính (tổ hợp dấu thanh và con chữ để cho ra các mẫu tự). Kiến trúc dựng sẵn sẽ giúp ích cho việc ghi nhớ nguyên dạng thông tin, xử lí tính toán, đưa thông tin ra một cách nhanh chóng.

Điều rõ ràng là chỉ một trong hai kiểu kiến trúc này sẽ không thể giải quyết được vấn đề mã hoá chữ Việt thích hợp cho cả truyền thông lẫn xử lí. Hai trường phái đối lập nhau này cần phải đi đến sự dung hòa thì mới mong tìm được giải pháp hữu hiệu cho vấn đề chữ Việt trên máy tính. Việc xây dựng một bộ mã tiềm ẩn cả hai cách kiến trúc này sẽ giúp giải quyết vấn đề mã hóa cho chữ Việt.

### 3. Bộ mã chuẩn trong quan hệ giữa chữ Việt và xử lí máy tính

Như các nghiên cứu đã chỉ ra, để mã hoá được cho mọi dạng thể hiện của chữ Việt, ngoài 128

mã thường dùng cuả bộ mã ASCII 7 bit, còn cần thêm 134 mã nữa. Chỉ chú trọng tới mặt mã hoá cho chữ Việt thì bộ mã 8 bit hoàn toàn đủ và ta có thể bố trí các mã chữ Việt vào đâu cũng được. Nhưng ràng buộc về việc chữ Việt phải thể hiện được trên máy tính và tương hợp với cách xử lí máy tính, đặc biệt là phải tương hợp với các phần mềm hoạt động dựa trên bộ mã ASCII, kéo theo việc phải giữ nguyên 128 mã đầu cuả ASCII. Điều này dẫn tới khó khăn đầu tiên là với 128 vị trí còn lại thì không đủ mã bố trí cho thêm 134 chữ Việt.

Phương án xây dựng bộ mã chữ Việt bằng cách hạn chế khả năng biểu diễn chữ Việt, thay thế các chữ ít dùng bằng các chữ khác, để giữ sự tương hợp với các phần mềm dựa trên ASCII, thực tế không được chấp nhận. Bộ mã phải có khả năng biểu diễn cho mọi chữ Việt, không được loại bỏ bất cứ dạng nào dù trong tương lai có thể có những thay đổi trong cách viết.

Phương án thoả hiệp giữa hai yêu cầu đối chọi trên đáng chú ý hơn cả. Theo cách bố trí bộ mã này thì thực chất ta có hai bộ mã đều chứa cả bộ mã ASCII 7 bit và mã các kí tự Việt, trùng nhau ở hầu hết mọi chỗ. Các vị trí không trùng nhau về mã là các vị trí có ý nghĩa nhất trong biểu diễn cho chữ Việt cũng như trong bộ mã ASCII. Tùy theo yêu cầu xử dụng, nếu cần in ấn, soạn thảo chữ Việt thì chọn bộ mã có đầy đủ mọi con chữ. Khi đó một số giá trị mã cho chữ Việt sẽ che lấp các mã khác, không quan trọng trong bộ mã ASCII, và vẫn đảm bảo các chương trình liên quan tới xử lí soạn thảo, in ấn làm việc bình thường. Khi cần dùng các chương trình xử lí thì dùng bộ mã thứ hai, trong bộ mã này mọi mã chữ Việt trùng với mã cuả ASCII sẽ bị mã ASCII che khuất. Khi đó phải chấp nhận không xử lí cho một số con chữ Việt ít dùng hay có cách khác thay thế mà vẫn bảo toàn ý nghĩa.

Thực tế hiện nay trong nước hầu hết mọi bộ phận tin học đều dùng công cụ xử lí là các máy vi tính cùng phần mềm dựa trên ASCII. Điều này làm nảy sinh một nhu cầu nữa khi dùng các phần mềm này cùng chữ Việt là làm sao bảo toàn được các kí tự nửa đồ họa trong bộ mã ASCII mở rộng. Giải pháp thoả hiệp trên có thể được áp dụng nhưng với một hạn chế thêm nữa: trong khi xử lí cho chữ Việt, chỉ dùng các con chữ chỏ có dấu còn con chữ hoa không có dấu. Nhiều bộ mã cuả nhiều nơi đưa ra đã lựa chọn cách thoả hiệp này. Tuy vậy, nhược điểm chính cuả cách lựa chọn này là ở chỗ chưa giải quyết được vấn đề truyền thông trên phương diện bộ mã.

## 5. Bộ mã chuẩn trong quan hệ giữa chữ Việt và truyền thông

Theo quy định phân vùng mã cho việc truyền thông thì 32 mã đầu tiên trong cả hai nửa cuả bộ mã 8 bit được dành làm các mã điều khiển truyền tin, do đó không được phép sử dụng để mã hoá cho các đối tượng khác. Điều này mâu thuẫn nghiêm trọng với việc biểu diễn đủ cho chữ Việt vì thiếu quá nhiều chỗ. Có thể tất cả các bộ mã đã có trước đây đều không đáp ứng được yêu cầu cuả truyền thông với việc vi phạm nặng nề nào quy tắc chung cuả bộ mã quốc tế dành cho truyền thông này.

Điều đó thực ra cũng có nguyên nhân vì trong tất cả các ứng dụng đã được thực hiện, chưa có ứng dụng nào cần tới việc mã hóa thích hợp cho việc truyền thông. Tin học ở Viet nam theo một nghĩa nào đó vẫn còn bị cô lập với thế giới. Trong các cộng đồng sử dụng tiếng Việt trên thế giới tin học cũng mới phát triển một cách cục bộ, độc lập mà chưa có sự liên kết rộng.

Đề nghị về bộ mã chuẩn [2] nêu ra trong hội thảo chuẩn hóa chữ Việt trên máy tính năm 1988 đã có chú ý tới vấn đề truyền thông, đã đề nghị một hình thức mã hóa phù hợp với yêu cầu truyền thông. Tuy nhiên, độ khác biệt giữa bộ mã đầy dủ cho chữ Việt và bộ mã truyền thông còn rất lớn. Cách tách bạch phần nguyên âm và các dấu đi kèm cuả chữ Việt phần nào vẫn còn mang tính chất kỹ thuật và chưa chú trọng tới đặc thù về hệ thống nguyên âm và dấu cuả chữ Việt.

Trong [3] đã đưa ra một đề nghị có thể giải quyết về cơ bản vấn đề mã hóa cho chữ Việt trên máy tính phù hợp với truyền thông. Xuất phát từ việc định nghĩa các đơn vị chính tả cơ bản cuả chữ Việt bao gồm các phụ âm, nguyên âm và dấu thanh, ngoài việc duy trì bộ mã mô tả đầy đủ mọi dạng tổ hợp con chữ và dấu, bộ mã còn chứa đủ các thành tố mã cơ bản cho các đơn vị chính tả. Các đơn vị chính tả sẽ đượcx bố trí trong bộ mã để tránh vùng các mã điều khiển. Khi thực hiện truyền thông sẽ chuyển đổi từ các mã chữ Việt thông thường thành các dạng các mã cho đơn vị chính tả. Các con chữ được khôi phục trở lại từ mã cuả các đơn vị chính tả theo khuôn mẫu đã được quy định thống nhất trong bộ mã chung. Về thực chất, đây là một cách thoả hiệp giữa hai yêu cầu mô tả đầy đủ cho chữ Việt và truyền thông. Chúng ta cũng có hai bộ mã, một cho đầy đủ các chữ Việt thì không đáp ứng yêu cầu truyền thông, một phù hợp với quy tắc truyền thông thì không có đủ dạng biểu diễn chữ Việt. Hai bộ mã này được bố trí trùng nhau ở các mã mô tả cho các đơn vị chính tả. Các mã chữ Việt rơi vào vùng điều khiển truyền thông sẽ được bố trí cho các con chữ có khả năng dễ dàng khôi phục lại theo quy tắc tạo phông chữ thông thường. Khi dùng trong xử lí, chúng ta có thể không bị hạn chế bởi các quy tắc cuả truyền thông và vẫn dùng các mã đầy đủ cho chữ Việt.

## 6. Đề nghị cụ thể về bộ mã chuẩn

Do vậy giải pháp thoả đáng nhất là phải thiết kế ra một bộ vài bảng mã, mỗi bảng mã có thể đáp ứng đầy đủ cho một yêu cầu nào đó. Tổ hợp sức mạnh cuả các bảng mã này chúng ta có được bộ mã đáp ứng được cho cả ba yêu cầu trên. Đương nhiên bộ các bảng mã này phải không được quá xa nhau, trái lại, chúng phải hỗ trợ cho nhau, phải là hình ảnh cuả nhau trong các tình hướng khác nhau.

Bộ mã chuẩn cho chữ Việt bao gồm một bộ mã chính mang số 2, mã hóa cho tất cả con chữ với dấu trong tiếng Việt và hai biến thể cuả bộ mã chính này. Bộ mã chính chứa tất cả các dạng thể hiện hợp pháp cuả chữ Việt có mang dấu. Nó là cơ sở để thống nhất các dạng hiển thị chữ Việt trên các phương tiện xử lí thông tin và là một gợi ý cho việc tạo từ diển chính tả các bộ phông dùng kỹ thuật PostScript.

Biến thể dành cho việc xử lí trong máy tính, mang số 3 cho phép người xử dụng lựa chọn các mã chữ Việt cần cho mình và phần mềm, dùng được một số kí tự nửa đồ họa.

Biến thể dành cho truyền thông mang số 1 tuân thủ mọi yêu cầu cuả truyền thông và được dự định đệ trình cho Tổ chức tiêu chuẩn quốc tế, dùng cho việc truyền thông các tài liệu chữ Việt. Toàn bộ ba biến thể cuả bộ mã đều được đệ trình trong một khuôn khổ bộ mã chuẩn thống nhất ở các cấp sử dụng quốc gia và quốc tế.

Các đơn vị chính tả biểu thị cho dấu thanh được xếp vào một vùng mã liên tục để tạo điều kiện thuận lợi cho việc kiểm tra dấu và để phù hợp với yêu cầu bố trí mã truyền thông.

Các đơn vị chính tả cho các nguyên âm thuần Việt được bố trí trong một vùng mã liên tục khác, theo yêu cầu bố trí mã truyền thông và để thuận lợi cho việc kiểm tra các nguyên âm thuần Việt.

Các con chữ mang dấu được sắp xếp theo thứ tự thanh. không dấu. huyền, hỏi, ngã, sắc, nặng.

Như vậy, ta nên quan niệm bộ ba bảng mã không phải là biệt lập mà là ba hình thái thể hiện cho tư tưởng về một bộ mã chung. Tư tưởng cơ bản nằm ở chỗ bảng mã 2 thể hiện đầy đủ mọi dạng hình cuả các chữ Việt với các dấu thanh, tuân theo nguyên tắc kiến trúc dựng sẵn. Ở đây tính dân tộc lấn át mọi tính chất xử lí và truyền thông, tuy nhiên vẫn bảo toàn phần kí tự GO để có thể cùng song tồn với tiếng Anh và do đó dùng được một số sản phẩm phần mềm cuả thế giới. Bảng mã này đáp ứng cho yêu cầu in ấn các văn bản chữ Việt và các hệ soạn thảo văn bản có thể dùng chúng làm công cụ tạo văn bản trong chữ Việt. Nói riêng, cách bố trí các mã đều cố gắng trong chừng mực có thể được theo trật tự các con chữ và dấu thanh đã quy định.

Khi cần truyền thông các văn bản chữ Việt, chúng ta chuyển sang dùng bảng mã 1 với quy tắc chuyển đổi từ các mã một byte sang hai byte cho các nguyên âm có dấu. Ở nơi nhận sẽ có chương trình trả lại dạng mã một byte cho các kí tự Việt. Như vậy, bảng 1 chỉ dùng để đáp ứng cho truyền thông, tính chất thích hợp cho truyền thông lấn át tính chất cuả mã một byte theo kiến thức dựng sẵn. Đặc điểm cuả bảng 1 là hoàn toàn tuân thủ luật quy định cuả ISO về việc để trống các vùng C), C1. Cách nói bảng 1 tuân thủ ISO 8859 cũng hàm ý về góc độ này. Phần còn lại cuả G1 được bố trí theo quan niệm thể hiện đầy đủ các kí tự Việt chữ con, các đơn vị chính tả Việt và một số các chữ hoa vốn cách vẽ không có tính đối xứng, cần phải có hình dạng định sẵn (như dấu huyền, hỏi đặt ở bên trái cuả dấu mũ, dấu sắc đặt ở bên phải cuả dấu mũ). Các chữ hoa có mã lọt vào vùng C1 sẽ là những chữ có thể tái tạo lại được theo cách chồng chất mẫu chữ và dấu, dấu luân vào chính giữa.

Các chữ con được bố trí để cố gắng phản ánh trung thành trật tự các con chữ Việt và trật tự dấu thanh, ưu tiên cho tính chất dân tộc này vượt lên trên cách bố trí mã cuả bảng chuẩn ISO 8859. Cách bố trí một số kí tự trong phần G1 cuả ISO 8859 chỉ phản ánh cho một số kí tự cuả một số nước khác. Ta cần phải bố trí mã trong phần này phải phản ánh được đặc thù cấu tạo các từ chữ Việt và cho phép tận dụng thêm một số kí tự đồ họa do đó làm cho chữ Việt có thể đi đôi được với một số phần mềm có bán trên thị trường. Sở dĩ bảng mã này được mang tên bảng 1 là vì còn có ngụ ý sẽ dùng nó để xin đăng kí làm một mã chuẩn thế giới. Dạng thể hiện không vi phạm các vùng điều khiển C0, C1 cho phép nó dễ dàng được ISO chấp thuận và do đó kéo theo cả bảng 2 và bảng 3 cũng được chấp thuận.

Cần phải nói thêm là thực ra để truyền thông văn bản tiếng Việt, có nhiều cách thức khác nhau gắn với nhiều mức độ kỹ thuật và phương tiện khác nhau. Hiện nay trong nước đang dùng phổ biến bộ mã điện tín, loại 5 bit hay 6 bit, cho việc truyền thông các văn bản chữ Việt. Đây là bộ mã mặc định (de facto) phát sinh do thực tế cần truyền các văn bản Việt từ thời phương tiện truyền thông mới chỉ có loại xử lí 5,6 bit. Đề nghị về việc dùng bộ mã 7 bit để mã hóa cho chữ

Việt là một phương án có thể xem xét tới trong việc lựa chọn chuẩn 7 bit cho biểu diễn chữ Việt [4]. tuy nhiên vấn đề ở đây là chúng ta cần có bộ mã 8 bit thích hợp cho truyền thông 8 bit thông qua các phương tiện truyền thông 8 bit. Các phát triển mới cuả kỹ thuật cũng yêu cầu phải có bộ mã 8 bit thích hợp cho việc truyền thông giữa các thiết bị, chẳng hạn như kĩ thuật chồng phông để tạo mẫu chữ từ các phông khác. Do đó nếu né tránh việc truyền thông 8 bit mà đi vòng qua các cách mã hóa theo số bit ít hơn thì tuy rằng ta có thể vẫn truyền thông được ngữ nghĩa văn bản Việt nhưng sẽ không đáp ứng được các yêu cầu phát triển mới cuả tin học và truyền thông

Đối với việc xử lí bên trong máy tính, bảng mã 2 là bảng mã kiểu một byte dựng sẵn cho chữ Việt và được khuyến khích sử dụng chừng nào chưa có xung đột với các mã điều khiển dùng trong các phần mềm. Trong trường hợp có xung đột thì ưu tiên nhường cho các mã điều khiển, không có mã chữ Việt vào vị trí đó nữa. Tuy nhiên để tránh được hoàn toàn mọi rắc rối, chỉ nên dùng chữ Việt nhỏ và chữ hoa không mang dấu thanh vì chúng được bố trí vào vùng chắc chắn không có xung đột với các kí tự điều khiển. Tính chất phù hợp cho xử lí được ở đây là mỗi byte ứng với một dạng thể hiện khác nhau cuả chữ Việt và bảo toàn vùng GO, vùng cơ sở cho các công cụ phần mềm.

Bảng mã 3 phát sinh do yêu cầu cuả anh chị em làm tin học trong nước. Hiện tại trong nước dùng một số phần mềm cuả thế giới trong việc quản lí cơ sở dữ liệu, nảy sinh vấn đề phải giữ lại cả một số mã cho các kí tự nửa đồ họa vẽ đường thẳng. Do đó bắt buộc phải hi sinh hầu hết các chữ Việt hoa và chỉ giữ được chữ con cùng một số kí tự đồ họa thôi. Bảng này qua thử nghiệm đã chạy được với hầu hết các phần mềm hiện có ở trong nước.

Tính thống nhất cuả ba bộ mã này được thể hiện khi ta đặt chồng cả ba bộ mã này lên nhau, phần nào khác biệt thì bỏ đi. Kết quả thu được là tập kí tự tối thiểu chứa toàn bộ tiếng Anh, vùng GO và các chữ Việt con có dấu cùng với các đơn vị chính tả Việt trong vùng G1. Về mặt ngữ nghĩa thì rõ ràng chỉ cần có bộ chữ con là đủ thể hiện cho mọi văn bản chữ Việt. Như vậy về mặt biểu diễn cần có cách tạo lập chữ hoa từ các chữ con và một kí tự báo hiệu bắt đầu, kết thúc chữ hoa.

**Tài liệu tham khảo**

1. Hội thảo về các xử lí văn bản tiếng Việt, Hà nội 19-20/1/1987, UBKHKTNN.
2. Proceeding of the 3rd National symposium on Standardization for Vietnamese Character Set Coding with Diacritics, Hochiminh-city, July 26-30 1988.
3. A Proposal for Vietnamese Character Encoding Standards in a Unified Text Processing Framework, James Do, Ngo Thanh Nhan, Nguyen Hoang July 7 1991.
4. Một khuôn khổ thống nhất cho việc xử lí dữ kiện Việt ngữ, Nhóm nghiên cứu tiêu chuẩn tiếng Việt, tháng giêng 1992, Viet-Std.

**Abstract**

Vietnamese Characters in Computer Processing and Communication

*In this paper, the relation between representing of Vietnamese characters, computer processing and communication, concerning with the Vietnamese standard code, will be discussed. Two architectures for representing Vietnamese characters: Precomposed and separated are presented. Based on that, some arguments about setting up a Vietnamese standard code for processing and communication are discussed. This standard code proposal consists of three separated code tables but unified in the same framework and satisfied all requirements about Vietnamese characters in computer processing and communication.*