Thống nhất chữ Việt trên máy tính theo chuẩn quốc tế Unicode

Đỗ Bá Phước
2000/06/17

Do đầu óc sáng tạo và tính độc lập của người Việt Nam, cho đến nay có trên dưới 43 bộ mã chữ Việt dùng cho máy tính các loại. Tuy nhiên, chỉ có vài bộ mã được dùng phổ biến nhất: TCVN 5712, VNI, VISCII. Tình trạng này xảy ra chủ yếu là vì khi máy tính được chế ra, mục đích chính là để xử lý chữ Anh (với bộ mã 7-bit), sau đó các loại chữ Tây Âu (với bộ mã 8-bit). Xử lý các loại chữ viết khác đòi hỏi những trò tiểu xảo với các hệ điều hành máy, có khi vi phạm nguyên tắc thiết kế (dẫn đến việc không hiện ký tự **ư** với TCVN 5712, và **ễ** với VISCII), nhưng mỗi giải pháp lại chỉ đáp ứng được một khía cạnh nào đó của việc xử lý chữ viết. Cách thực hiện TCVN 5712 và VISCII đòi hỏi bộ font thường và hoa, còn cách thiết kết VNI không phù hợp với nguyên tắc của ngôn ngữ học, gây phức tạp khi xử lý và cách trình bày theo font VNI không được chính xác. Vấn đề chế bản là phạm vi áp dụng chính của TCVN 5712, VNI, và VISCII.

Lẽ ra thì người dùng máy tính không cần biết đến những chi tiết kỹ thuật hỗ trợ cho việc xử lý chữ Việt, nhưng với tình trạng như hiện nay, cũng nên phân biệt khái niệm bộ mã với khái niệm cách thực hiện các *bộ mã* với khái niệm *cách thực hiện* các bộ mã. Bộ mã định nghĩa từng ký tự với một mã số nào đó. Cách thực hiện gồm *phương tiện đưa dữ kiện [input]* vào máy (như bộ gõ chữ VietKey, áp dụng cho các bộ mã khác nhau), *xử lý [processing]*, và *trình bày [output]* trên màn ảnh hoặc máy in (như bộ font ABC cho TCVN 5712). Sự lầm lẫn giữa những khái niệm này vẫn còn tồn tại, kể cả trong giới công nghệ thông tin.

Khi máy tính còn là những công cụ tách biệt nhau, tình trạng nhiều bộ mã không tương thích là một khó khăn lớn với người dùng. Nhưng khi đã bước qua thời kỳ Internet, tình trạng đa mã là một trở ngại lớn lao trong việc trao đổi thông tin, văn hoá, và thương mại. Tình trạng này đúng không những cho chữ Việt, mà cho các ngôn ngữ khác trên thế giới nữa. Chính vậy mà chỉ trong những năm gần đây mà bộ mã chuẩn quốc tế Unicode ngày càng quan trọng và được sử dụng phổ biến, tuy rằng Unicode là một công trình kéo dài hơn mười năm nay.

Bộ mã Unicode đa ngôn ngữ bảo đảm sự trao đổi thông tin thông suốt toàn cầu, vì lý do cơ bản là Unicode đặt cho mỗi ký tự trong các chữ viết trên thế giới một mã số 16-bit duy nhất. Như vậy, khi dữ kiện được trao đổi qua máy tính, sẽ tránh khỏi sự trùng lập giữa hai ký tự khác nhau -- một việc sẽ xảy ra nếu dùng bộ font chữ để phân biệt các ký tự khác nhau nhưng cùng mã số trong các bộ mã 8-bit khác nhau. Ký tự **ư** của TCVN 5712 (và **ễ** của VISCII) chỉ là ví dụ điển hình nhất của tình trạng này. Dữ kiện là dữ kiện (ký tự), không phù thuộc vào cách trình bày (font).

Từ đầu, **chữ quốc ngữ** đã có mặt trong Unicode, kể cả ký tự **đ** (đồng). **Chữ Nôm** (5065 chữ có mã số, 4234 chữ thuần Nôm đợi mã số, và khoảng 1000 chữ nữa), **chữ Chàm**, và **chữ Thái** của Việt Nam vẫn được tiếp tục đưa vào Unicode.

Do đòi hỏi mở rộng thị trường công nghệ thông tin ngoài Bắc Mỹ và châu Âu sang

những nước lớn như Trung quốc và Ấn độ, những công ty lớn như Microsoft, IBM, Apple, Oracle, ... đã đẩy mạnh việc thực hiện Unicode trong sản phẩm của mình. Như vậy, khả năng xử lý chữ Việt có sẵn trong các phần mềm này. Một ví dụ là Microsoft cho không những bộ font Unicode, trong đó có toàn bộ các ký tự Việt Nam, và các hệ điều hành Windows 98, NT và 2000 và bộ Office 2000 xử lý triệt để Unicode.

Người dùng máy tính **nên** chuyển về Unicode càng sớm càng tốt vì:

1. tất cả những bộ phận cần thiết đã được dựng sẵn trong Windows 98, NT, và 2000 -- mà không cần phải cài đặt những loại phần mềm khác. Tuy nhiên, để gõ chữ Việt cho tiện lợi, theo thói quen của từng người, phần mềm duy nhất cần cài thêm là VietKey.
2. khi soạn văn bản, không cần chuyển font hoa/thường, mà cũng không phân biệt font Việt hay không.
3. chữ Việt sẽ hiện lên màn ảnh rất đẹp và chính xác không cần đợi download font mỗi khi đọc trang web Việt, và sẽ không thiếu ký tự nào.
4. sẽ dùng được chữ Việt với mọi ngôn ngữ khác mà không ngại mâu thuẩn nào khi soạn và trao đổi văn kiện.

Nhà công nghệ thông tin cần chuyển về Unicode càng sớm càng tốt vì:

1. Unicode là một chuẩn quốc tế, được Nhà nước Việt Nam công nhận.
2. Unicode là một phần cơ bản của:
   * Windows 98, NT, 2000, CE (PocketPC), Mac OS X,
   * ngôn ngữ phổ biến nhất là Java,
   * XML dùng trong phạm vi Web, và là cơ sở của trao đổi thông tin, thương mại điện tử.
3. hiểu và ứng dụng được Unicode sẽ nới rộng thị trường của mình, trong lĩnh vực quốc tế hoá [internationalization (i18n)]và bản địa hoá [localization (l10n)].
4. có thể xử lý chữ Việt như bất cứ ngôn ngữ nào trong Unicode.
5. có thể nới rộng phạm vi xử lý chữ Việt đến những công cụ cầm tay như điện thoại di động, một thị trường đang bùng nổ khắp thế giới.

Với những điểm trên, câu hỏi cấp bách cần đặt ra là liệu giới công nghệ thông tin Việt Nam có đủ thông minh và sáng suốt nhìn thấy lợi ích chung, để thống nhất với một bộ mã duy nhất, nhằm cung cấp người dùng những giải pháp và công cụ hết sức tự nhiên và minh bạch để xử lý tiếng Việt trên máy tính./.