Dũng Vũ
Tiếng Việt và vấn đề dịch máy (3)

Đã có một kho dữ liệu từ vựng được soạn kỹ lưỡng, có phẩm chất và giàu thông tin, chúng ta có thể dùng cơ sở này cho những mục đích:

- Soạn từ điển tiếng Việt
- Soạn từ điển song ngữ.
- Cung cấp thông tin cho máy dịch tự động, hệ nhận diện ngôn ngữ tự nhiên.

4. Soạn từ điển tiếng Việt

Soạn từ điển là một công việc đòi hỏi nhiều thời gian và công sức vì lượng thông tin quá lớn. Tuy vậy công việc có thể trở nên nhẹ nhàng nhờ được tự động hóa bằng cách lập trình. Chương trình đọc dữ liệu từ kho dữ liệu từ vựng và chuyển dạng theo cách trình bày mong muốn [1]. Từ điển nhỏ, dữ liệu ít; từ điển lớn, dữ liệu nhiều. Từ điển in thì đơn giản. Từ điển điện tử thì phức tạp hơn, đa dạng hơn, giàu thông tin hơn. Tất cả tùy thuộc vào sự lựa chọn thuộc tính CAT, SEX, PERS, NUM, CASE, SEM, … Nói chung, nội dung văn bản, sự lựa chọn thuộc tính, cách trình bày, tiểu tiết là công việc của người thực hiện.

Xem vài ví dụ có thể được tự động hóa hoàn toàn bên dưới.

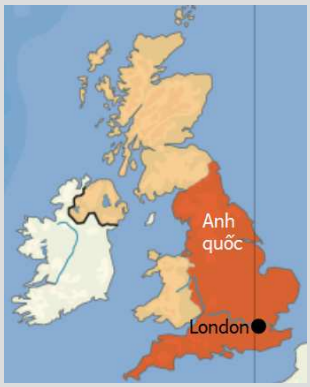
Từ điển bản in 1 là phiên bản đơn giản nhất bao gồm những dữ liệu rút ra từ các thuộc tính:
- SEM (semantic, ngữ nghĩa): phần giải thích ý nghĩa cho từng trường hợp sử dụng.
- EXP (example, ví dụ): phần ví dụ với chữ nghiêng (italic)
- GEO, IMG (image, hình ảnh): phần hình ảnh của IMG(1)

Từ điển bản in 2 được mở rộng thêm bằng các thuộc tính:
- CAT (lexical category, phạm trù từ vựng): <dt> (danh từ, noun)
- SUBCAT (lexical subcategory, tiểu phạm trù từ vựng): <nvđdt> (nhân vật đại danh từ, personal pronoun), <dtr> (danh từ riêng, proper noun). Xin lưu ý: Nếu chọn cả hai thuộc tính CAT và SUBCAT cùng lúc, CAT sẽ bị SUBCAT viết chồng lên (overwritten), nghĩa là *từ loại (part of speech)* của từ đó sẽ mang giá trị SUBCAT.
- REG (region, địa phương): được sử dụng ở ba miền: Nam, Trung, Bắc.
- MORE (chi tiết liên quan) với quy chiếu ký hiệu bằng mũi tên →. Bấm vào đây sẽ dẫn tới chi tiết liên quan.
- GEO (geometry, địa lý), IMG(2) (image 2, hình 2)

Đã định nghĩa xong những gì mong muốn (v.d. mục từ "anh", "Anh"), ta chỉ cần bấm chuột và kết quả sẽ hiện ra ngay:

anh
1. Tiếng xưng hô
a. thân mật, dùng trong quan hệ nam nữ, vợ chồng; người nam tự xưng *anh*; người nữ gọi lại cũng bằng *anh*. ***Anh** yêu em.*
b. thân mật, cũng dùng trong gia đình, họ hàng, xã hội; người nam tự xưng *anh*; người trong vai về em cũng gọi lại bằng *anh*.
c. lịch sự, ngang hàng, dùng trong giao tiếp xã hội để gọi một người nam đối diện từ tuổi trưởng thành trở lên. *Chào **anh** Hùng. Tôi tên là Cường. Hân hạnh được biết **anh**.*
d. lịch sự, dùng chỉ một người nam ngôi thứ 3, xác định được nhờ ngữ cảnh phía trước. *Hùng là giáo sư đại học. **Anh** hiện sống ở Hoa Kỳ.*
2. Người con trai trong gia đình lớn tuổi hơn các người con khác, hoặc trong họ hàng cùng thế hệ ở vai về cao hơn.
3. Người, người nam ngôi thứ 3. *Có người ví von: "Việt Nam, Cuba như là Trời Đất sinh ra. Một **anh** ở phía Đông. Một **anh** ở phía Tây. Chúng ta thay nhau canh giữ hòa bình cho thế giới. Cuba thức thì VN ngủ. Việt Nam gác thì Cuba nghỉ." (Nguyễn Minh Triết)*
4. Của anh (sở hữu từ dùng cho một người nam ngôi 1, 2). *Áo **anh** sứt chỉ đường tà. Vợ **anh** mất sớm, mẹ già chưa khâu (thơ Hữu Loan, Phạm Duy phổ nhạc)*

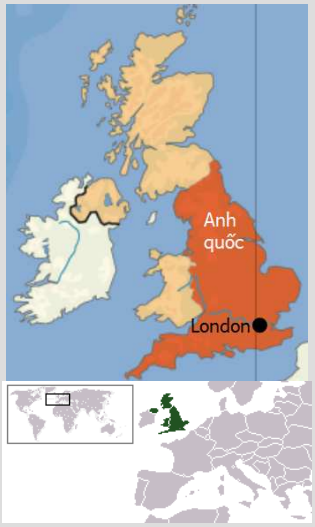
Anh Anh quốc, nước rộng lớn và đông dân nhất trong Liên hiệp Vương quốc Anh và Bắc Ireland, nằm ở phía Tây Bắc Âu châu, có thủ đô là Luân Đôn (London).



Từ điển bản in 1

anh
1. <nvđdt>: Tiếng xưng hô (ba miền)
a. thân mật, dùng trong quan hệ nam nữ, vợ chồng; người nam tự xưng *anh*; người nữ gọi lại cũng bằng *anh*. ***Anh** yêu em*
→ *em, mình, vợ, chồng, vợ chồng*
b. thân mật, cũng dùng trong gia đình, họ hàng, xã hội; người nam tự xưng *anh*; người trong vai về em cũng gọi lại bằng *anh*.
→ *em, chị*
c. lịch sự, ngang hàng, dùng trong giao tiếp xã hội để gọi một người nam đối diện từ tuổi trưởng thành trở lên. *Chào **anh** Hùng. Tôi tên là Cường. Hân hạnh được biết **anh**.*
→ *tôi*
d. lịch sự, dùng chỉ một người nam ngôi thứ 3, xác định được nhờ ngữ cảnh phía trước. *Hùng là giáo sư đại học. **Anh** hiện sống ở Hoa Kỳ.*
→ *anh ta, anh ấy, ảnh*
2. <dt>: Người con trai trong gia đình lớn tuổi hơn các người con khác, hoặc trong họ hàng cùng thế hệ ở vai về cao hơn.
→ *anh họ*
3. <dt>: Người, người nam ngôi thứ 3. *Có người ví von: "Việt Nam, Cuba như là Trời Đất sinh ra. Một **anh** ở phía Đông. Một **anh** ở phía Tây. Chúng ta thay nhau canh giữ hòa bình cho thế giới. Cuba thức thì VN ngủ. Việt Nam gác thì Cuba nghỉ." (Nguyễn Minh Triết)*
4. <tt>: Của anh (sở hữu từ dùng cho một người nam ngôi 1, 2, 3). *Áo **anh** sứt chỉ đường tà. Vợ **anh** mất sớm, mẹ già chưa khâu (thơ Hữu Loan, Phạm Duy phổ nhạc)*
→ *của*

Anh <dtr>: Anh quốc, nước rộng lớn và đông dân nhất trong Liên hiệp Vương quốc Anh và Bắc Ireland, nằm ở phía Tây Bắc Âu châu, có thủ đô là Luân Đôn (London).



Từ điển bản in 2

Cách thực hiện *từ điển điện tử* cũng theo nguyên tắc tương tự, càng ít lựa chọn (options)

càng đơn giản; càng nhiều lựa chọn, càng phong phú. Nói chung đó ý nghĩa của bài bản *lập đồ (configure, projektieren)*. Nó cho phép người thực hiện tự ứng dụng một cách uyển chuyển và nhanh chóng những tính chất mình mong muốn mà không cần phải lập trình (program, programmieren).

anh
1. <nvđdt>: Tiếng xưng hô (ba miền)
a. thân mật, dùng trong quan hệ nam nữ, vợ chồng; người nam tự xưng *anh*; người nữ gọi lại cũng bằng *anh*. ***Anh*** *yêu em.* → *em, mình, vợ, chồng, vợ chồng*

b. thân mật, cũng dùng trong gia đình, họ hàng, xã hội; người nam tự xưng *anh*; người trong vai vế em cũng gọi lại bằng *anh*. → *em, chị*

c. lịch sự, ngang hàng, dùng trong giao tiếp xã hội để gọi một người nam đối diện từ tuổi trưởng thành trở lên. *Chào* ***anh*** *Hùng. Tôi tên là Cường. Hân hạnh được biết* ***anh****.* → *tôi*

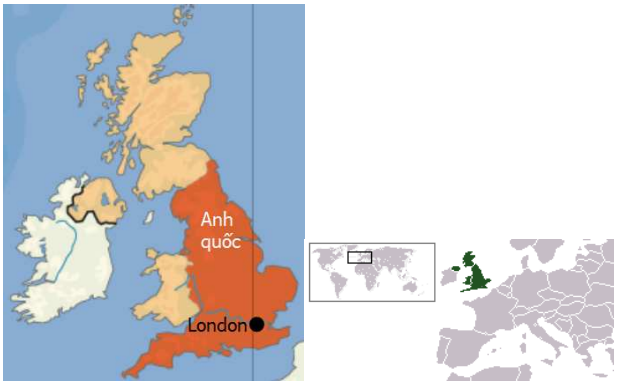
d. lịch sự, dùng chỉ một người nam ngôi thứ 3, xác định được nhờ ngữ cảnh phía trước. *Hùng là giáo sư đại học.* ***Anh*** *hiện sống ở Hoa Kỳ.* → *anh ta, anh ấy, ảnh*

2. <dt>: Người con trai trong gia đình lớn tuổi hơn các người con khác, hoặc trong họ hàng cùng thế hệ ở vai vế cao hơn. → *anh họ*

3. <dt>: Người, người nam ngôi thứ 3. *Có người ví von: "Việt Nam, Cuba như là Trời Đất sinh ra. Một* ***anh*** *ở phía Đông. Một* ***anh*** *ở phía Tây. Chúng ta thay nhau canh giữ hòa bình cho thế giới. Cuba thức thì VN ngủ. Việt Nam gác thì Cuba nghỉ." (Nguyễn Minh Triết).*

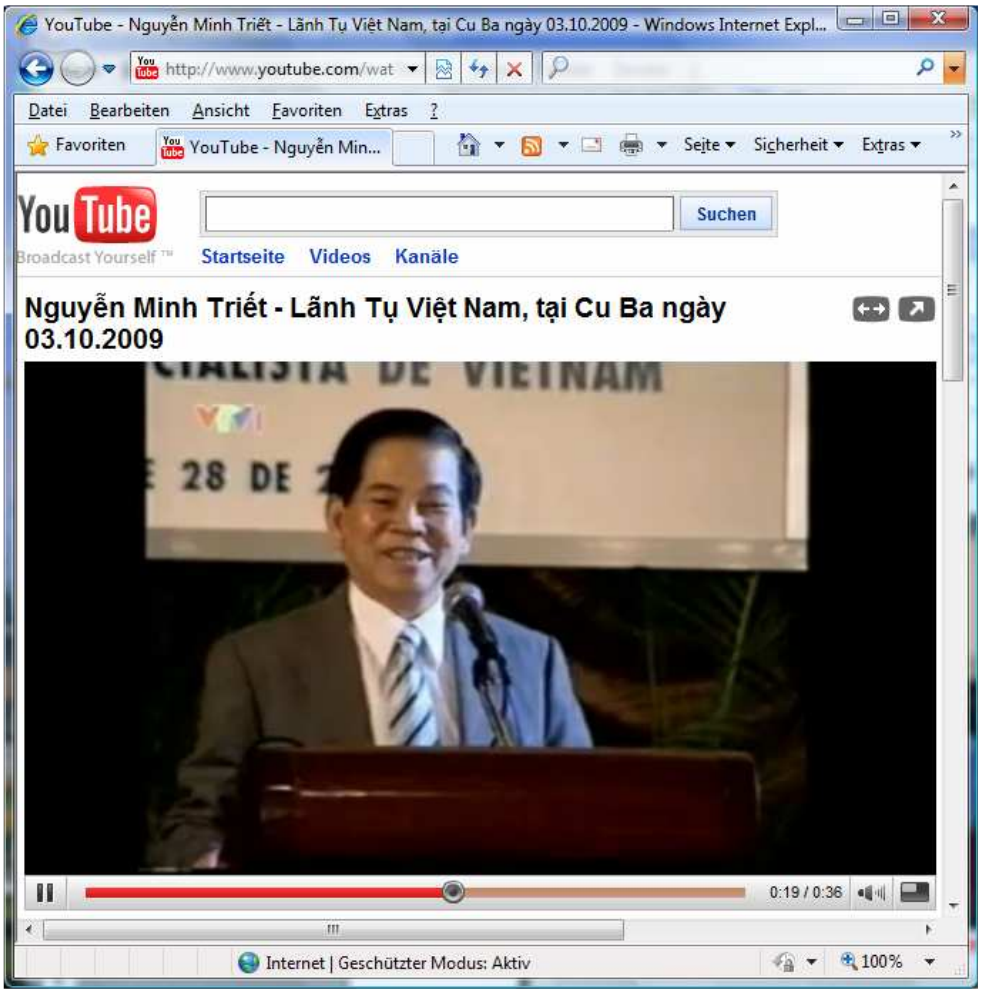
4. <tt>: Của anh (sở hữu từ dùng cho một người nam ngôi 1, 2, 3). *Áo* ***anh*** *sứt chỉ đường tà. Vợ* ***anh*** *mất sớm, mẹ già chưa khâu (thơ Hữu Loan, Phạm Duy phổ nhạc).* → *của*

Anh
<dtr>: Anh quốc, nước rộng lớn và đông dân nhất trong Liên hiệp Vương quốc Anh và Bắc Ireland, nằm ở phía Tây Bắc Âu châu, có thủ đô là Luân Đôn (London).
→ http://www.britannien.de/Geschichte/Geschichte.htm



Phiên bản điện tử bên trên được mở rộng từ bản *Từ điển bản in 2* bằng các lựa chọn:

- PRONUNC (Pronunciation, cách phát âm), ký hiệu bằng : Bấm vào đó sẽ nghe được cách phát âm chữ *"anh"*, rất tiện lợi cho người ngoại quốc muốn học tiếng Việt.
- MORE (more infos, chi tiết): ký hiệu bằng →, hyperlink dẫn đến (nguồn) chi tiết về nước Anh.
- AUDIO (âm thanh), ký hiệu bằng , hoặc VIDEO (phim ảnh), ký hiệu bằng : V.d. bấm vào đó sẽ dẫn đến YouTube xem chủ tịch nước Nguyễn Minh Triết phát biểu.



4.1. Thuộc tính khác

Còn nhiều thuộc tính khác có thể bổ sung vào kho dữ liệu từ vựng. Có những thuộc tính bổ ích mà trước nay chưa được để ý.

1) Cách dùng đúng *danh từ hình thức (formal noun)* cái, con, chiếc, ...: *"cái búa"* thay vì *"con búa"*; *"con gà"* thay vì *"cái gà"*, ... Những từ tiếng Việt này na ná như mạo từ *"the"* (Anh), *"le"*, *"la"* (Pháp), *"der"*, *"die"*, *"das"* (Đức), nhưng khá phức tạp và phong phú:

   Áng, Anh, Ánh, Bà, Bác, Bài, Bãi, Bản, Bao, Bận, Bầy, Bề, Bó, Bọn, Bộ, Bờ, Buồng, Bụm, Bức, Cách, Cánh, Cảnh, Căn, Cặp, Câu, Cậu, Cây, Chàng, Chị, Chiếc, Chiều, Chồng, Chú, Chùm, Chuỗi, Chương, Con, Cô, Cỗ, Cú, Cục, Cụm, Cuốc, Cuộc, Cuốn, Dãy, Dịp, Dòng, Đám, Đàn, Đạo, Đấng, Đóa, Đoàn, Đoạn, Đôi, Đồ, Đống, Đợt, Đứa, Đức, Em, Gã, Gian, Giọt, Góc, Gói, Hạt, Hòn, Hột, Hớp, Kẻ, Khẩu, Khóm, Khu, Khúc, Kiện, Kỳ, Làn, Lão, Loại, Loài, Loạt, Lô, Lời, Lũ, Luồng, Mảnh, Mảng, Mẩu, Mé, Miếng, Món, Mớ, Mụ, Mùa, Mụn, Nải, Nàng, Nắm, Nén, Nền, Ngọn, Ngôi, Ngụm, Người, Nhà, Nhóm, Niềm, Nỗi, Nụ, Ông, Phần, Pho, Quả, Quyển, Sự, Sợi, Tay, Tảng, Tấm, Tấn, Tập, Tên, Thang, Thanh, Thằng, Thày, Thể, Thứ, Thửa, Tiếng, Tính, Toán, Tốp, Tờ, Trái, Tụi, Viên, Vị, Việc, Vở, Vốc, Vũng, Xấp, Xó, ...

   Ví dụ: Áng mây, Anh lính, Bà hoàng hậu, Bài thơ, Bản nhạc, Bầy con nít, Bọn học trò, Bộ luật, Bức tranh, Cánh cửa, Cảnh đời, Cặp vợ chồng, Câu chuyện, Cây cau, Chàng chiến sĩ, Chị sinh viên, Chiếc lá, Chú học trò, Con bò, Cô y tá, Cục đá, Cuộc đời, Cuốn sách, Đám bèo, Đàn trẻ nhỏ, Đạo luật, Đấng trượng phu, Đống rác, Đôi lời, Đứa bạn, Đức Phật, Em gái, Gã thủy thủ, Giọt nước, Hạt cát, Hòn đảo, Hột đậu, Kẻ phản bội, Khẩu súng, Khúc

sông, Lão phường trưởng, Loại xe, Lời nhạc, Lũ quỷ, Mảnh vườn, Mảng đời, Mẩu truyện, Miếng cơm, Món cá chiên, Mụ tú bà, Nàng công chúa, Nền văn hóa, Ngôi chùa, Ngọn núi, Người lính, Nhà văn, Nhóm bạn, Nỗi lòng, Ông vua, Pho tượng, Quả địa cầu, Quyển truyện, Sự đời, Tay anh chị, Tấm hình, Tấn tuồng, Tên tài xế, Thanh gươm, Thằng em, Thứ trái cây, Thửa vườn, Tiếng chuông, Tờ giấy, Trái chuối, Viên Thống đốc, Vị tổng thống, ...

Muốn người ngoại quốc học tiếng Việt cho đúng, cần ghi chú thông tin này vào từ điển, v.d. *"cái búa".*

```
búa
   FORMNOUN: "cái"
```

2) Loại từ (type of word): thuần Việt [2], Hán-Việt, phiên âm, ...
3) Lĩnh vực: (tôn giáo, chính trị, hóa học, toán học, kinh tế, …)
4) Từ phản nghĩa (≠): *"dễ"* ≠ *"khó", "dễ dàng"* ≠ *"khó khăn".*
5) Từ đồng nghĩa (=), từ tương tự (≈), từ "thuần Việt" tương đương từ Hán-Việt (v.d. *"nước"* (thuần Việt) = *"quốc gia"* (Hán-Việt)).
6) Từ đồng nghĩa, tương tự (Bắc, Trung, Nam): *"má"* (Nam) = *"mạ"* (Trung) = *"mẹ"* (Bắc); *"chả giò"* (Nam) = *"nem rán"* (Bắc).
7) Nghĩa đen, nghĩa bóng, …
8) Ý nghĩa của từ viết tắt: CO2 = carbon dioxid. VN = Việt Nam.
9) Cách viết 'y', 'i'.
10) Cực tính mở rộng [3].
11) Từ nào thường đi với từ nào (v.d. "đẹp" thường đi với trạng từ "quá", "lắm")
12) Mức độ thông dụng (1, 2, 3, …) để làm từ điển. Tự điển nhỏ chứa từ thông dụng, từ điển lớn có thêm từ không thông dụng. Từ điển nhỏ cần ít ví dụ; từ điển lớn cần nhiều ví dụ.
13) Cách sử dụng từ theo chức năng từ vựng. V.d.: *"dễ dàng"*.

    (1) Công việc này rất *dễ dàng* (tính từ).
    (2) Nó có thể làm công việc này *một cách dễ dàng* (trạng từ).

    Nên nhớ không phải từ nào cũng có thể được sử dụng như vậy. V.d.: *"khó".*

    (1) Công việc này rất *khó* (tính từ).
    (2) * Nó có thể làm công việc này *một cách khó* (trạng từ). (* = Không nói được).

    Cho nên khi xét đến tính phản nghĩa/đồng nghĩa, cần để ý đến chức năng từ vựng (xem 4.)
14) Tính thực dụng của tiếng đệm "nhá", "nhé", …

…

Nhờ một cơ sở dữ liệu từ vựng giàu thông tin, được soạn kỹ lưỡng, có phẩm chất (như đã thấy), chúng ta có thể làm ra một cuốn từ điển tiếng Việt, bất kể dạng gì, bản in hay điện tử, đơn giản hay phức tạp. Nói chung là khả thi và có phẩm chất [4]. Còn quá nhiều thông tin bổ ích trong kho dữ liệu chưa dùng đến [5].

4.2. Vấn đề tối ưu hóa

Thử nghiệm một chữ *"anh"* duy nhất cũng đủ thấy công phu. Kho tàng tiếng Việt còn vô vàn từ ngữ khác. Từ ngữ càng nhiều, thông tin càng nhiều, càng phức tạp. Cái khó là làm sao tối ưu hóa việc xử lý thông tin. Cải cách là một hình thức tối ưu hóa. Tuy nhiên cần cân nhắc kỹ lưỡng, vì có thể được chỗ này nhưng lại mất chỗ khác. Phải trải qua kinh nghiệm thực tế thì mới thấy có những cải cách tưởng chừng có lợi nhưng hóa ra lại làm hại chính mình.

Chẳng hạn, cải cách lối viết 'y' thành 'i'.

Đây là đề tài từng gây tranh cãi. Người cải cách muốn thay thế y bằng i cho thống nhất. Thế nhưng không dễ thực hiện, bởi lẽ có người thích viết i dài, có người thích viết i ngắn theo thẩm mỹ của mình. Thực tế này đã kéo dài hàng thế kỷ chứ không phải mới đây. Không thể thay thế y bằng i một cách cứng nhắc. "Quý" đổi thành "quí" thì được nhưng "Thúy" cùng vần với "quý" và cùng lối viết, không thể đổi thành "Thúi".

Đối với vấn đề này, con người có thể quyết định trường hợp nào thay thế được, trường hợp nào không. Nhưng máy là một vật vô tri vô giác, muốn nó được như người, thì phải dạy nó và sẽ tốn rất nhiều công sức không đáng cho một chuyện nhỏ nhặt. Hơn nữa, nếu máy có làm được, chắc chắn nó cũng mất thời gian và làm giảm hoạt năng xử lý. Lại thêm vấn đề, chỉ vì cải cách mà ra cả.

Để đẹp lòng đôi bên và tiện lợi cho máy có lẽ chỉ còn cách là ghi chú thêm cách viết (cho i, y), mà thực ra cũng không nhiều lắm. V.d*. quý, quí.* Máy có thể nhận diện cách viết dễ dàng, không tốn nhều thì giờ tính toán.

Hoặc một trường hợp khác là cải cách lối đánh dấu thanh cho chữ Việt gần đây.

1) Mục đích đánh dấu thanh trong tiếng Việt xưa nay là gì ?

   Như đã biết, tiếng Việt là một ngôn ngữ thanh điệu; từ khi dùng chữ quốc ngữ, người Việt có thể ghi chú thanh điệu một tiếng bằng một ký tự nhất định. Quy tắc đánh dấu thanh chữ viết tiếng Việt xưa nay chỉ thuần quy ước, được đưa ra chỉ nhằm một mục đích duy nhất là ghi chú một âm tiết có thanh điệu gì, ngoài ra không ghi chú thêm gì khác.

   Trong quá khứ, ít nhất là trước năm 1975, đã có một quy tắc đánh dấu thanh nhất định và đã được sử dụng **tuyệt đối nhất quán**. Điều này có thể kiểm chứng qua **tất cả** các văn bản (báo chí, sách vở, tài liệu, từ điển [6]) đã được phát hành thời ấy, ít ra là ở miền Nam. Hai quy tắc ấy như sau:

   1: Gặp một chữ có 1 nguyên âm chứa dấu mũ, dấu ngoắc như Ă, Â, Ê, Ô, Ơ, Ư, thì đánh dấu lên đó. V.d.: *"Tuấn", "tập", "viết".* Nếu có hai (như ƯƠ), thì đánh dấu lên nguyên âm sau (Ơ). V.d.: *"đường", "được".*

   Không dùng được quy tắc 1 thì dùng quy tắc 2

   2: Gặp một chữ có phụ âm cuối, thì đánh dấu lên nguyên âm **chót**. V.d.: *"hoàng", "hoạt", "toán", "coóng"*. Nếu không có thì đánh dấu lên nguyên âm **áp chót**. V.d.: *"họa", "hòe", "hủy".* (Dĩ nhiên gặp một chữ chỉ có một nguyên âm thì chỉ còn cách là đánh dấu lên nguyên âm đó thôi. V.d.: *"gọn", "quá"*).

2) Khi xưa ta viết *"hóa",* nay có đề nghị viết *"hoá",* khi xưa ta viết *"háo",* nay vẫn viết *"háo"* mà không viết *"haó".* Tại sao các nhà ngôn ngữ học Việt Nam lại đề nghị như thế ? Điển hình là đề nghị của Vũ Xuân Lương, Hoàng Phê [7], đại ý là: "Với những âm tiết kết thúc bằng *"oa", "oe", "uy", dấu thanh được đặt vào con chữ nguyên âm chót.* V.d. *hoạ, hoè, huỷ, loà xoà, loé, suý, thuỷ* "

Đối với sự nhận diện của máy, sự khác biệt của cách viết y và i chỉ là 1, bởi vì chỉ tồn tại một trường hợp duy nhất: thế y bằng i. Nhìn chung không đáng kể, độ phức tạp (complexity grad) coi như bằng 1. Trong khi đó độ phức tạp của trường hợp *"oa", "oe", "uy"* tăng lên gấp 15 lần (3 âm tiết *"oa", "oe", "uy"* x 5 dấu thanh (sắc, hỏi, huyền, ngã, nặng)). Hoạt năng máy bị giảm đáng kể. Thử tưởng tượng, khi người dùng gõ chữ *Thúy* (dấu sắc trên u) vào máy và máy không tìm thấy (vì trong kho dữ liệu từ vựng chỉ có chữ *Thuý* (dấu sắc trên y)). Nó buộc lòng phải gọi một function (hàm số) để xem *Thúy* (dấu sắc trên u) có phải là trường hợp cần đổi cách đánh dấu không, đại để như sau:

```
private string GiveMeYourReformedWord(string input)
{
   // Bước 1: Bỏ phần phụ âm đầu, lấy phần nguyên âm.
   // Giả sử input = "Thúy". Bỏ "Th", lấy "úy", để được:

   string myInput = "úy";

   // Bước 2: đổi cách đánh dấu thanh (sắc, hỏi huyền, ngã, nặng)
   string myOutput = null;

   switch (myInput)
   {
      // case 1
      case "óa":
         myOutput = "oá";
         break;

      // case 2
      case "ỏa":
         myOutput = "oả";
         break;

      // case 3
      case "òa":
         myOutput = "oà";
         break;

      // case 4
      case "õa":
         myOutput = "oã";
         break;

      // case 5
      case "ọa":
         myOutput = "oạ";
         break;

      // case 6
      case "óe":
         myOutput = "oé";
         break;

      // case 7
      case "ỏe":
         myOutput = "oẻ";
         break;

      // case 8
      case "òe":
         myOutput = "oè";
         break;

      // case 9
      case "õe":
         myOutput = "oẽ";
         break;

      // case 10
      case "ọe":
         myOutput = "oẹ";
         break;
```

```
      // case 11
      case "úy":
         myOutput = "uý";
         break;

      // case 12
      case "ủy":
         myOutput = "uỷ";
         break;

      // case 13
      case "ùy":
         myOutput = "uỳ";
         break;

      // case 14
      case "ũy":
         myOutput = "uỹ";
         break;

      // case 15
      case "ụy":
         myOutput = "uỵ";
         break;
   }
   return ("Th" + myOutput);
}
```

Phải trải qua một function dài lê thê như thế [8], máy mới thông báo được phải viết làm sao cho đúng: À ! *Thúy* phải viết thành *Thuý*.

Đã thấy rắc rối chứ ? Nhưng chưa hết. Giả sử cái kho tàng từ vựng tiếng Việt chỉ có một chữ độc nhất là *Thuý*, thì không nói, chỉ cần gọi function `GiveMeYourReformedWord("Thúy")` một lần là xong, đàng này, cái khu rừng tiếng Việt còn muôn ngàn chữ có dấu khác nữa. Để đi tìm *Thúy* phải gọi muôn ngàn lần cái function ấy thì biết bao giờ mới tìm được Thúy ?

Cần nhớ, tốc độ là một trong những yếu tố quan trọng nhất của dịch máy.

(còn tiếp)

Chú thích:

[1] Đối với từ điển bản in, thì chuyển kết quả ra dạng DOC-file. Đối với từ điển điện tử, thì chuyển kết quả ra dạng XML, chẳng hạn.

[2] Thường gọi là "thuần Việt", chứ thực ra – theo ngôn ngữ học khảo cổ - tiếng Việt là một tổng hợp của nhiều kết quả ngôn ngữ trong vùng: tiếng Khmer, Mường, Hán, … Không thể xem kết quả nào trong đó là thuần Việt được.

[3] Đối với cách nhìn của giới ngôn ngữ học Âu châu, cực tính trước nay chỉ có ba giá trị: tốt (positive), xấu (negative), trung hòa (neutral). Từ có cực tính tốt như "*good*", "*pretty*" (Anh), *"gut"*, *"schön"* (Đức). Từ có cực tính xấu như *"stupid"* (Anh), *"dumm"* (Đức). Còn lại là từ có cực tính trung hòa, không tốt, không xấu.

Tuy nhiên ngôn ngữ tinh vi hơn nhiều. Tốt nhưng tốt thế nào; xấu nhưng xấu thế nào? Cái này con người thừa hiểu nhưng máy không hiểu. Muốn cho máy hiểu giống người thì phải dạy cho nó những tính chất tinh vi: lịch sự, thô kệch, tục tằn, mỉa mai, cay độc, dịu dàng, thân mật, …

[4] Thậm chí còn tốt hơn từ điển ngoại quốc, chẳng hạn so với từ điển Duden hoặc Blockhaus của Đức.

[5] Nếu biết tận dụng thông tin, chắc chắn giới từ điển học Việt Nam sẽ làm được những sản phẩm từ điển tốt chưa từng đạt tới. Một sản phẩm thông tin tốt tạo cơ hội kinh doanh tốt. Thử nhìn vào lĩnh vực từ điển trên thế giới hiện nay, nhiều nơi đang ra sức làm ra các sản phẩm *corpus*, *treebanks* (Penn treebank) cho mục đích thương mại. Phẩm chất hàng càng tốt, giá licence càng cao. Xem http://www8.informatik.uni-erlangen.de/IMMD8/Services/sammlung_korpora/Penn.html

[6] Riêng về từ điển, có thể xem:
Huình Tịnh Của (1895) *Đại Nam Quấc Âm Tự Vị*. Rey, Curiol & C^{ie}: Saigon
Đào Duy Anh (1957) *Hán Việt Từ Điển*. Trường Thi: Sài Gòn
Nguyễn Lương Ngọc (1971) (chủ biên) *Từ Điển Học Sinh*. Nxb Giáo Dục: Hà Nội. Điều này chứng tỏ cách đánh dấu thanh truyền thống cũng được dùng ở miền Bắc. Phiên bản được in lại lần thứ ba tại *sở giáo dục thành phố Hồ Chí Minh, 1997* cũng giữ nguyên cách đánh dấu thanh truyền thống.

[7] **Vũ Xuân Lương**: *Quy tắc đặt dấu thanh trong tiếng Việt*, Trung Tâm Từ Điển Học http://vietlex.com/vietnamese/quytacbodau.html,
Hoàng Phê chủ nhiệm trang Trung Tâm Từ Điển Học: http://www.vietlex.com/

[8] Cách lập trình *(switch … case)* chỉ nhằm mục đích cho dễ mường tượng rằng có nhiều trường hợp để quyết định. Dĩ nhiên có cách ngắn gọn hơn, chẳng hạn dùng *list, collection.* Tuy vậy, nhìn nội dung dưới dạng *assembly code*, độ phức tạp xử lý mã không thay đổi.