Thân tặng ngành ngôn ngữ học điện toán Việt Nam. Thân tặng riêng các anh trong nhóm nghiên cứu tiếng Việt của cố giáo sư Cao Xuân Hạo: Hoàng Dũng, Bùi Mạnh Hùng, Nguyễn Đức Dương, Hoàng Xuân Tâm, Nguyễn Văn Hiệp.

Stuttgart, 10.2010

Tác giả

STUTTGART

2 0 1 0

Dũng Vũ

Tiếng Việt và vấn đề dịch máy (1)

Dịch máy (machine translation) đã trở thành một tiện ích phổ biến trong thời đại internet. Đây là một trong những đề tài phức tạp nhất của việc xử lý ngôn ngữ con người trong lĩnh vực *trí thông minh nhân tạo (artificial intelligence)* mà giới khoa học đã miệt mài nghiên cứu để đạt nhiều kết quả đáng kể ngày hôm nay: tự điển song ngữ tự động (Anh-Đức, Anh-Pháp, Anh-Nhật, Anh-Việt, …), dịch cụm từ, câu (như Google Translator, Babel Fish, …). Tuy nhiên vẫn còn nhiều khó khăn.

Dịch máy cũng là đề tài được giới khoa học Việt Nam quan tâm và còn gặp nhiều khó khăn hơn nữa, ít nhất vì những nguyên nhân khách quan sau đây:

Thứ nhất, tiếng Việt là một ngôn ngữ rất phức tạp (so với ngôn ngữ Âu châu).

Thứ hai, dịch máy là một chủ đề còn mới mẻ đối với giới ngôn ngữ học điện toán Việt Nam. Thời gian nghiên cứu chưa lâu (ước khoảng một thập niên trở lại đây). Thiếu nhân sự, kỹ thuật, kinh nghiệm kỹ nghệ, tài chính, … Thậm chí thiếu kiến thức ngôn ngữ học thuần túy và đôi khi vẫn chưa nắm được đặc điểm tiếng Việt.

Bài viết sơ lược này sẽ chỉ ra một số vấn đề dịch máy tiếng Việt và đồng thời gợi ý giải quyết. Tuy đây là một đề tài chuyên môn nhưng bài viết chọn cách trình bày bình dân, dễ hiểu thay cho hình thức một tiểu luận khoa học khô khan [1].

1. Mục đích dịch máy

Nhiều người nghĩ, đã có máy dịch, thì từ đây trở đi không cần người dịch nữa; máy sẽ dịch như người; muốn dịch gì cũng được; dịch một cuốn truyện từ tiếng Anh sang tiếng Việt cho người không biết tiếng Anh đọc thoải mái; dịch thơ Nguyễn Du cho người ngoại quốc thưởng thức. Chỉ cần bấm nút là xong.

Thực ra mục đích của dịch máy – tính đến ngày hôm nay - không phải nhằm thỏa mãn nhu cầu dịch thuật văn chương mà là nhu cầu thực tế bình thường, ví dụ dịch từ điển song ngữ, dịch các câu văn, lời nói theo lối hành văn đơn giản: một câu chào hỏi, một mẩu tin tức, một

cẩm nang hướng dẫn dùng thuốc tây, ... Phức tạp hơn là một tài liệu hướng dẫn sử dụng máy móc, một tài liệu khoa học kỹ thuật, v.v.

2. Vấn đề dịch máy

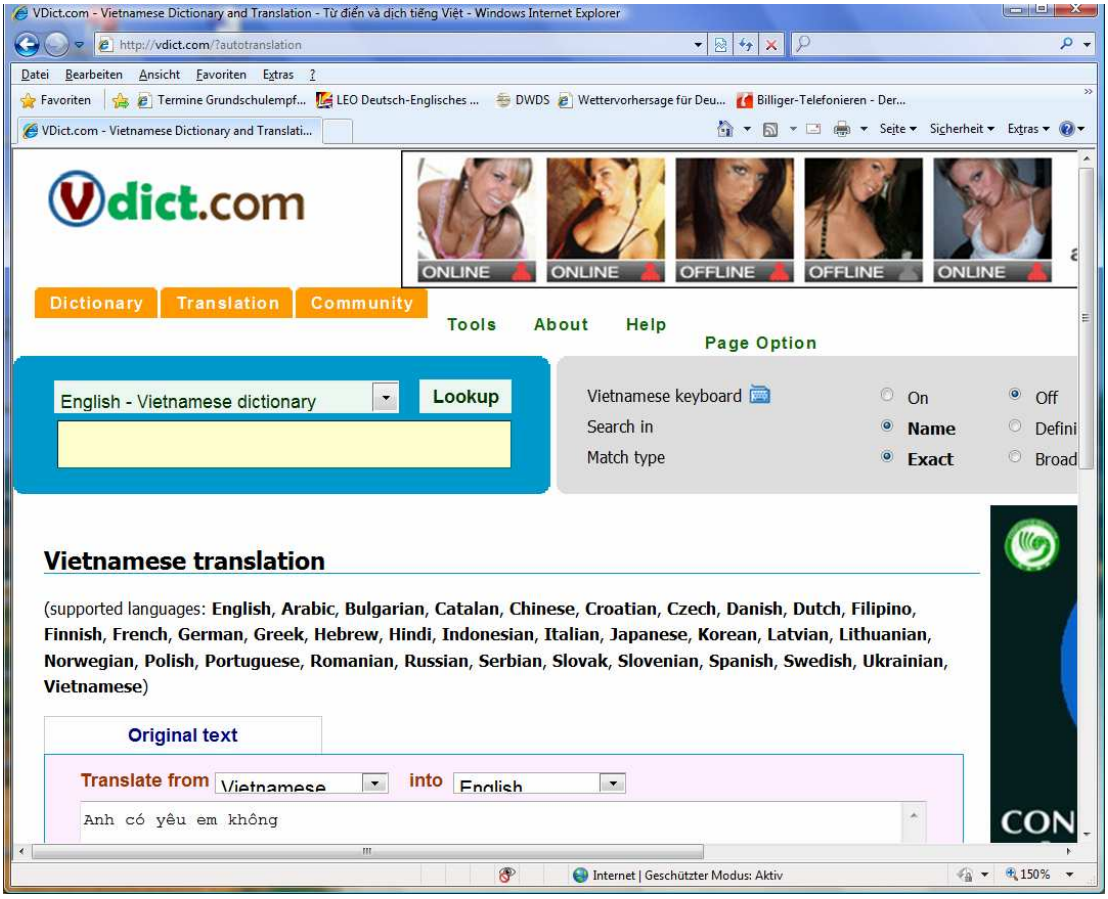
Vấn đề dịch máy phụ thuộc vào nhiều yếu tố: độ phức tạp của ngôn ngữ (cú pháp, ngữ nghĩa, …), cách sử dụng ngôn ngữ (tùy văn hóa, phong cách, mục đích, …). Chẳng hạn:

Về mặt cú pháp: Dịch một từ dễ nhất. Dịch một cụm từ khó hơn. Dịch một câu càng khó nữa.

Về mặt ngữ nghĩa: Nếu một từ gốc chỉ có một nghĩa *tương đương (equivalent)* với một từ dịch, thì không thành vấn đề, ví dụ *"vườn"* (Việt) tương đương với *"garden"* (Anh), *"Garten"* (Đức). Nhưng gặp trường hợp đa nghĩa, ta sẽ thấy sự *"nhập nhằng" (disambiguation),* v.d. *"anh"* tương đương với *"you", "he/him", "brother"* của tiếng Anh. Máy không hiểu *"anh"* trong câu phải dịch theo nghĩa nào. Hoặc dịch sang tiếng Đức, sẽ còn có thêm trường hợp. *"Anh"* cũng có thể là *"Sie"* (tiếng xưng hô) đối với người nam trưởng thành đáng kính trọng hoặc chưa thân, ngược lại cũng có thể là *"Du"* đối với trường hợp một cô gái gọi bạn trai mình, hoặc một người vợ gọi chồng mình. Trong trường hợp ngược lại, người đàn ông Đức có thể dùng tiếng *"Du"* (như "you") gọi vợ mình, người yêu của mình, thế nhưng người đàn ông Việt không thể dùng tiếng *"anh"* để gọi tương tự vậy mà phải gọi là *"em"*. Đó là yếu tố văn hóa, cách hành ngôn.

Sau đây hãy xem một ví dụ ứng dụng dịch máy của công ty Vdict [2].

Vdict sử dụng hai máy dịch, một máy của Google, một máy của Việt Nam có tên là EVtran (http://vdict.com/?autotranslation).

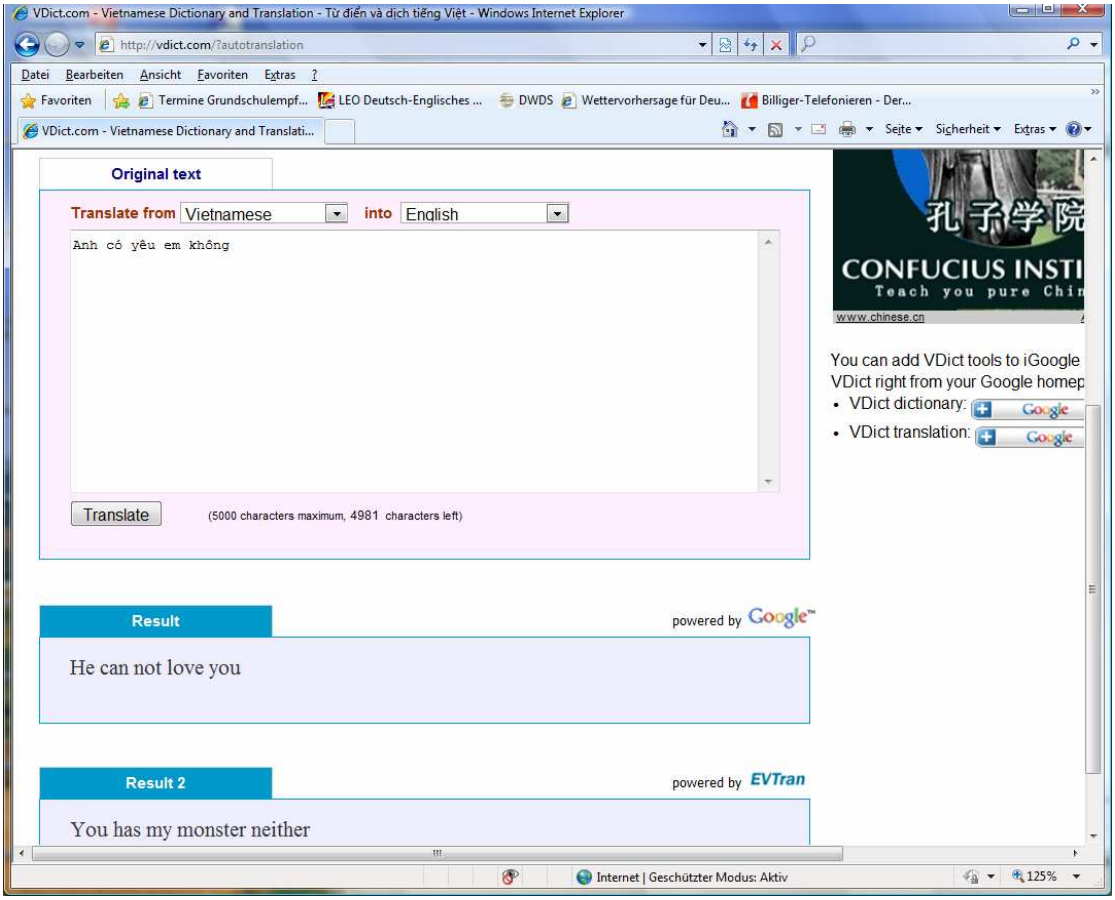


Nhân đang nhắc đến tiếng xưng hô thân mật *"anh", "em",* hãy cho thử cho dịch một câu

thường gặp trong tình yêu nam nữ, ví dụ: *"Anh có yêu em không".*

Bấm nút Translate, máy sẽ cho ra kết quả (xem hình bên dưới):

Result powered by Google (ô trên): He can not love you
Result 2 powered by EVtran (ô dưới): You has my monster neither



"yêu" được EVtran dịch thành *"monster".* Không đúng ! *"Yêu"* có nhiều nghĩa nhưng máy lại hiểu như thể là "yêu tinh" (monster).

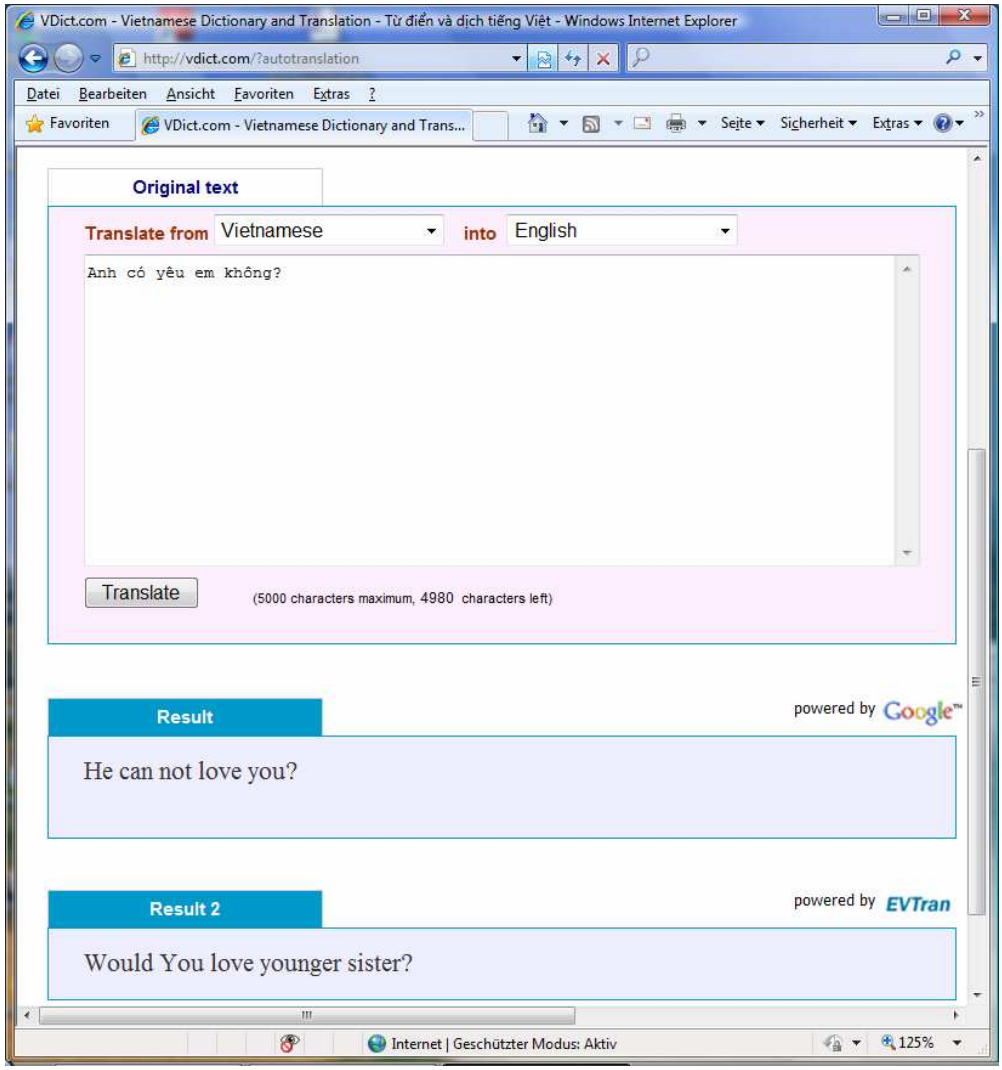
"anh" đáng lẽ phải được dịch thành *"you"* nhưng Google dịch thành *"he",* như thể máy hiểu nhầm thành *"anh ta"* vậy. Đây là vấn đề ***nhập nhằng ngữ nghĩa*** thường thấy trong dịch máy.

Mới xét từng từ riêng lẻ đã thấy vấn đề không đơn giản huống gì dịch một cụm từ, một câu. Câu *"Anh có yêu em không"* đáng lẽ phải được dịch thành *"Do you love me"* nhưng kết quả dịch máy lại là:

Google: He can not love you
EVtran: You has my monster neither

Thí nghiệm này đã cố tình không đặt dấu hỏi sau câu và được kết quả như trên. Nếu cho thêm dấu hỏi vào, EVtran sẽ cho một kết quả khác:

Input: Anh có yêu em không ?
Output: **EVtran:** Would You love younger sister?



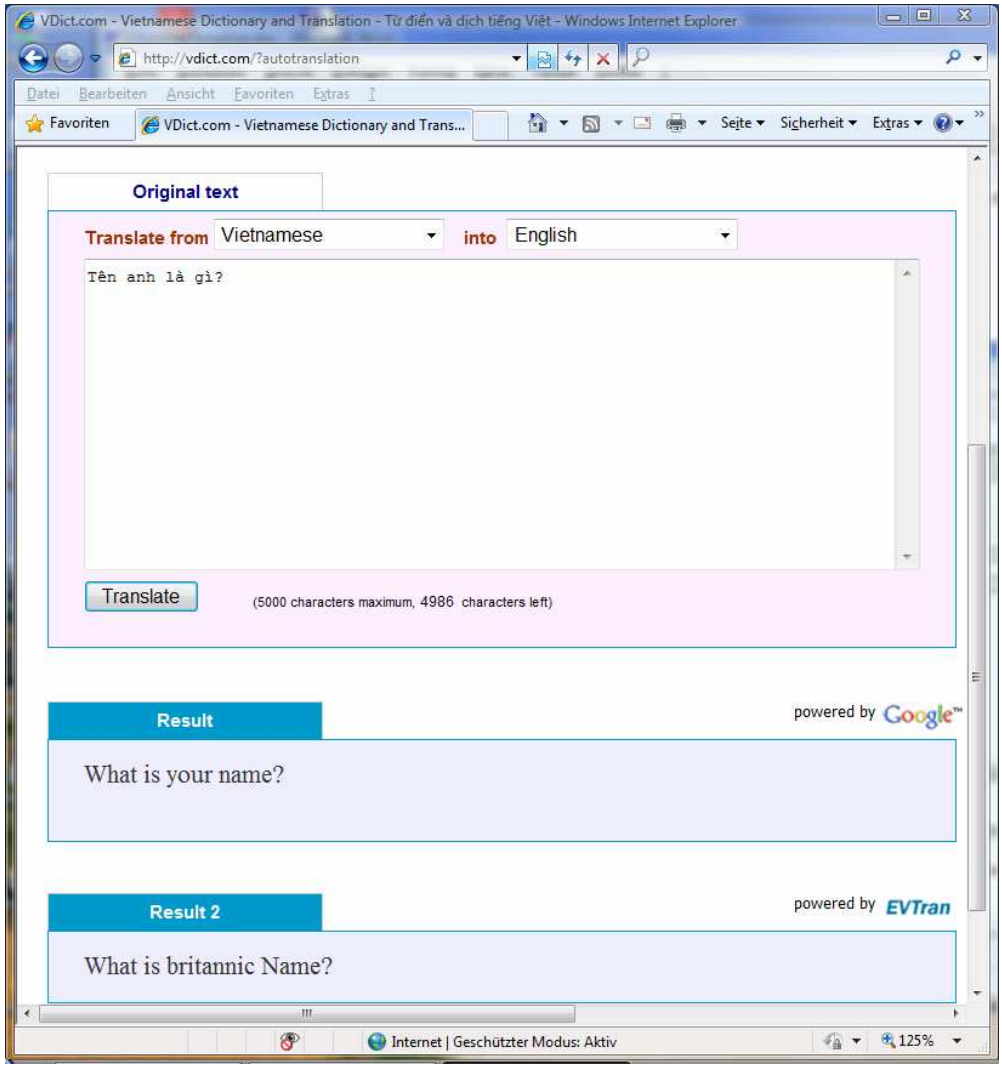
"Em" ở đây đáng lý phải dịch thành *"me",* thay vì *"younger sister".*

Nhìn chung, từ ngữ càng dài, ngữ nghĩa càng bất định song điều đáng nói hơn vẫn là vấn đề ***cấu trúc***. Cấu trúc một cụm từ, một ngữ đoạn, một câu càng phức tạp, máy càng dễ dịch sai.

Tuy nhiên vẫn có trường hợp đúng. Chẳng hạn kết quả dịch câu: *"Tên anh là gì ?"*:

Google: What is your name?
EVtran: What is britannic Name?

Google dịch đúng. (EVtran dịch sai, không cần bàn).



Qua cách dịch câu *"Tên anh là gì ?"*, có thể đoán được cách *diễn giải (interpretation)* của Google:

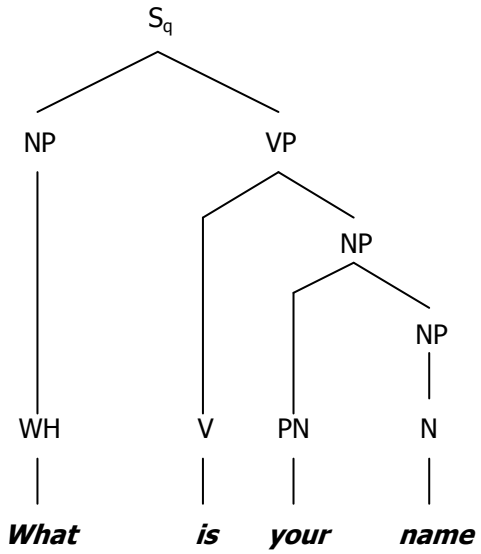
"X là gì" được dịch thành "What is X" dựa vào *quy tắc sản sinh (production rules)* theo mô hình *General Grammar (ngữ pháp tạo sinh)* của Chomsky. Quy tắc sản sinh tiếng Anh, v.d. cho câu *"What is your name"* là:

- S_q → NP VP
- VP → V NP
- NP → WH | PN NP | N

- WH → *what*
- V → *is*
- PN → *your*
- N → *name*

(S_q = Question, VP = verb phrase, NP = noun phrase, WH = WH-word, PN = pronoun, V = verb, N = noun, | = phép OR trong logics)

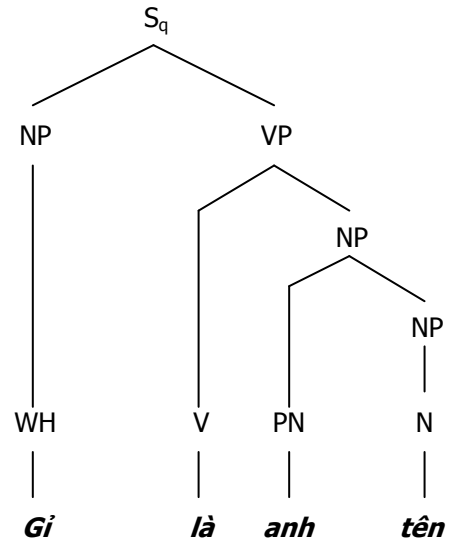
Mô tả cấu trúc câu hỏi trên bằng *cây cú pháp (syntax tree),* ta có:



Áp dụng quy tắc sản sinh của tiếng Anh cho tiếng Việt và thay thế những *nút tận cùng (terminal node)* bằng tiếng Việt, theo hình thức dịch 1-1, ta có:

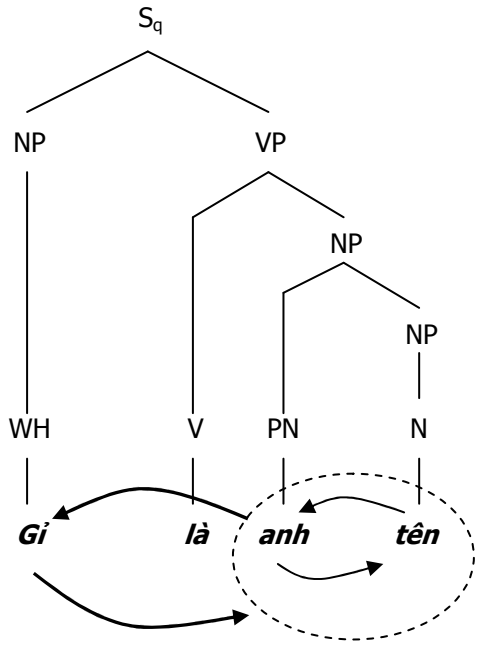
- `WH` → *gì*
- `V` → *là*
- `PN` → *anh*
- `N` → *tên*

Mô tả bằng cây cú pháp:



Kết quả *"Gì là anh tên?"* bên trên sai ngữ pháp tiếng Việt vì dịch 1-1.

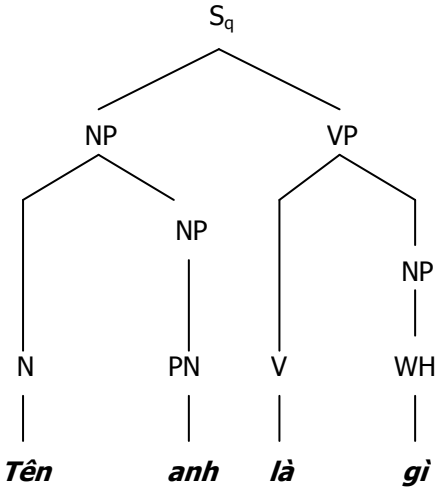
Cấu trúc câu hỏi của tiếng Anh và tiếng Việt trái ngược, do đó cần thay đổi vị trí từ:



Thay đổi vị trí từ đòi hỏi thay đổi quy tắc sản sinh cho tương xứng:

- S_q → NP VP
- VP → V NP
- NP → WH | NP PN | N

Và ta được cây cú pháp:



Mặc dù câu *"Tên anh là gì ?"* không sai ngữ pháp tiếng Việt nhưng đó không phải là lối hành ngôn tiêu biểu của người Việt. Thay vì vậy, người Việt thường hỏi: *"Anh tên là gì ?"* (hoặc ngắn hơn: *"Anh tên gì ?"*).

Cấu trúc câu *"Anh tên là gì ?"* chứa 2 chủ ngữ: *"Anh"* và *"tên"*. *"Anh"* là *chủ ngữ tâm lý (psychological subject)*, *"tên"* là *chủ ngữ ngữ pháp (gammatical subject)*. [3]

Cấu trúc này cũng thường thấy ở cách hành ngôn tương tự:

Anh nhà ở đâu ?

Chị quê ở đâu ?

Cháu Hương nó mấy tuổi rồi ?

v.v.

Nếu thử cho dịch một câu có cấu trúc 2 chủ ngữ như *"Anh tên là gì ?",* ta sẽ được kết quả:

Google: He called what?
EVtran: What is name Great Britain?

Phân tích hai kết quả, chúng ta có thể đoán ra cách dịch:

Google theo cách dịch (gần như) 1-1:

- `anh` → *`he`*
- `tên` → *`called`*
- `là gì` → *`what`*

Máy dịch Google có vài điểm sai. *"Anh"* dịch thành *"he"*, như thể máy hiểu là *"anh ta"*. ***"Là gì"*** được hiểu là *"what",* thay vì *"what **is**",* mâu thuẫn với trường hợp *"What is your name"* (với *"is"* là một động từ (verb)).

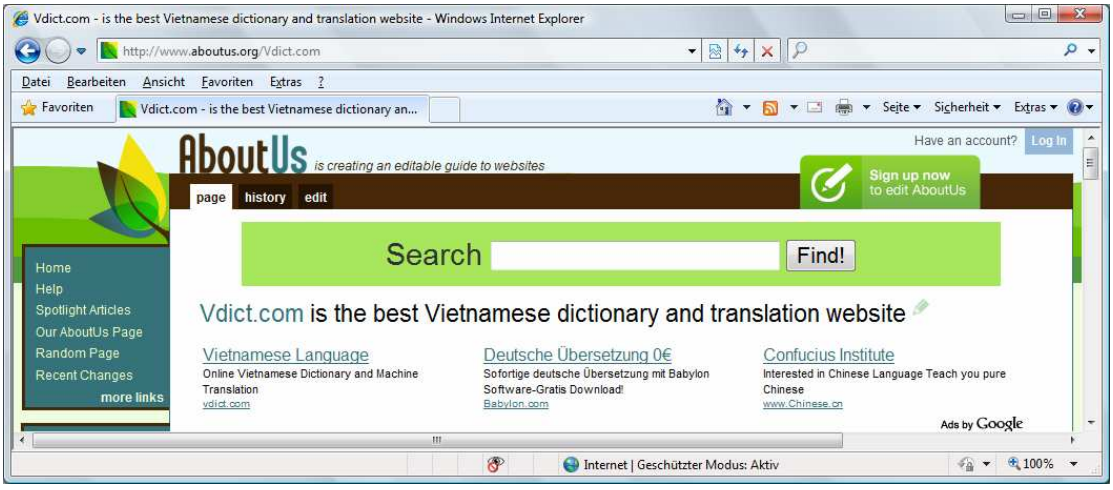
Máy dịch EVtran cũng có điểm sai. Hai *subject "Anh", "tên"* biến thành thành hai *object: "name"* và *"Great Britain"* (ngoài ra, *"Anh"* ở đây cũng không có nghĩa là *"Great Britain"*).

Nói chung, cả hai máy, Google và EVtran đều dịch sai vì không hiểu cách hành ngôn của người Việt, đặc biệt là cấu trúc Đề-Diễn hay Đề-Thuyết (Theme-Rheme) [4]. Không những vậy mà còn sai ngữ pháp tiếng Anh.

(còn tiếp)

Chú thích

[1] Chi tiết được trình bày trong cuốn *Tiếng Việt và Ngôn ngữ học hiện đại - Dịch máy.* Dũng Vũ, VIET: Stuttgart (sẽ phổ biến)
[2] Theo đánh giá của trang AboutUs (http://www.aboutus.org/Vdict.com), Vdict.com hiện nay (2009) là website dịch máy tiếng Việt tốt nhất.



[3] xem **Halliday M.A.K** (1994) *An Introduction to Functional Grammar.* London: Arnold. (tr. 32)

[4] Tôi gọi là Đề-Diễn (Chủ đề và diễn đạt). Cao Xuân Hạo gọi là Đề-Thuyết (Chủ đề và thuyết minh). Xem:
Cao Xuân Hạo (2004) *Tiếng Việt. Sơ thảo ngữ pháp chức năng.* TP.HCM: Nxb Giáo Dục. (tr. 22)
Dũng Vũ (2003) *Tiếng Việt và ngôn ngữ học hiện đại - Sơ khảo về cú pháp*. Stuttgart: VIET. (tr. 36) http://www.talawas.org/talaDB/suche.php?res=3244&rb=08