VẤN ĐỀ VỀ RANH GIỚI TỪ TRONG NGỮ LIỆU SONG NGỮ ANH-VIỆT

Đinh Điền, Hồ Bảo Quốc
Khoa CNTT, ĐH Khoa học Tự nhiên – ĐHQG Tp.HCM
(ddien, hbquoc)@fit.hcmuns.edu.vn

TÓM TẮT

Để dịch máy theo phương pháp thống kê, tra cứu xuyên ngôn ngữ, nghiên cứu so sánh đối chiếu các điểm tương đồng và dị biệt giữa ngôn ngữ tiếng Anh và tiếng Việt, chúng ta cần phải xây dựng được một kho ngữ liệu song ngữ Anh-Việt (English-Vietnamese parallel corpus). Kho ngữ liệu này phải qua các xử lý như: dóng hàng từ (word alignment), gán nhãn tự loại, cú pháp, ngữ nghĩa,..

Tuy nhiên, trước khi tiến hành các xử lý tự động trên, chúng ta nhất thiết phải xác định được các tiêu chí nhận diện ranh giới từ (word boundary) tiếng Anh cũng như tiếng Việt để làm cơ sở hình thái học cho các xử lý tự động đó. Trong bài báo này, chúng tôi sẽ trình bày một số vấn đề liên quan đến việc xác định ranh giới từ tiếng Anh và tiếng Việt một cách tự động trong song ngữ Anh-Việt.

Nội dung bài báo bao gồm 5 phần sau:

1. Giới thiệu: giới thiệu ngữ liệu song ngữ. Việc dóng hàng từ trong song ngữ. Nhu cầu xác định ranh giới từ cho bài toán dóng hàng từ.

2. Tổng quan: các quan điểm về ranh giới từ. Đơn vị "tiếng" và "từ" trong tiếng Việt.

3. Một số điểm khác biệt về hình vị giữa tiếng Anh và tiếng Việt.

4. Đề nghị tiêu chí ranh giới từ trong song ngữ Anh-Việt: nhằm phục vụ cho bài toán dóng hàng từ tự động.

5. Kết luận và hướng phát triển: nhận xét, khả năng ứng dụng và hướng phát triển trong tương lai.

1. GIỚI THIỆU

1.1 Giới thiệu về ngữ liệu song ngữ:

Thuật ngữ "ngữ liệu" được tạm dịch từ thuật ngữ tiếng Anh "corpus", có nghĩa là "kho dữ liệu, kho sưu tập tài liệu,.." (theo Từ điển Anh-Việt, ĐH Ngoại ngữ, NXB GD-2000 trang 368). "Ngữ liệu" ở đây có thể xem là những "dữ liệu, cứ liệu của ngôn ngữ", tức là những chứng cứ thực tế sử dụng ngôn ngữ. Ngữ liệu song ngữ

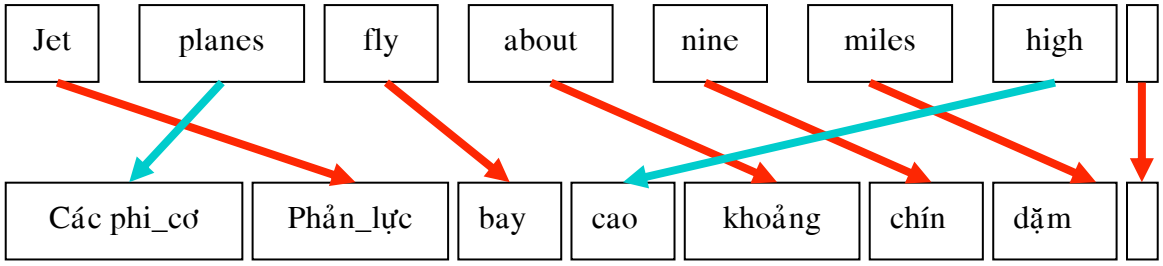
(dịch từ tiếng Anh là: bilingual corpus hay parallel text hay bitext) là ngữ liệu tồn tại dưới 2 ngôn ngữ và chúng là bản dịch của nhau.

Trong dịch máy theo phương pháp thống kê (Statistical Machine Translation), tra cứu xuyên ngôn ngữ (Cross-Lingual Information Retrieval), nghiên cứu so sánh đối chiếu các điểm tương đồng và dị biệt giữa ngôn ngữ tiếng Anh và tiếng Việt (English-Vietnamese contrastive linguistics), chúng ta không thể nghiên cứu trên lý thuyết, hay trên những câu do chúng ta nghĩ ra, mà phải nghiên cứu trên những câu có thật trong thực tế sử dụng. Điều này đòi hỏi chúng ta phải có các chứng cứ của ngôn ngữ, các ví dụ từ thực tế đã được nhiều người sử dụng và được xem là ngôn ngữ chuẩn [Tony McEnery, Andrew Wilson (1996)].

Với sự ra đời của máy tính điện tử và nhất là trong môi trường kết nối Internet toàn cầu như hiện nay, việc tập hợp ngữ liệu song ngữ đã được tự động hoá rất nhiều. Trên thế giới, người ta đã xây dựng được nhiều kho ngữ liệu song ngữ, như: Anh-Pháp, Anh-Hoa,… Trong bài báo này, chúng tôi sử dụng kho ngữ liệu song ngữ Anh-Việt điện tử 5 triệu từ được thu thập từ các tài liệu song ngữ thuộc lĩnh vực khoa học tự nhiên chủ yếu là tin học, điện tử viễn thông, y học,.. (Đinh Điền, 2002b).

1.2 Dóng hàng từ cho ngữ liệu song ngữ:

Dóng hàng từ là nhằm liên kết một từ tiếng Anh với một từ tiếng Việt tương ứng (Dien Dinh, 2002). Ví dụ:



Do sự khác biệt về loại hình ngôn ngữ (language typology) và loại hình văn hoá, nên trong bài toán dóng hàng từ tự động, chúng ta phải giải quyết nhiều vấn đề liên quan đến cơ sở ngôn ngữ học như:

- Sự khác biệt về từ vựng hoá (lexicalization)
- Sự khác biệt về phương tiện ngữ pháp: tiếng Anh thường dùng phương thức phụ tố, còn tiếng Việt thường dùng trật tự từ và từ hư
- Do đặc thù tiếng Việt: như phó danh từ, phó động từ, từ láy, …

Ngoài ra, còn có những yếu tố khác (như: sự khác biệt giữa cấu trúc cú pháp đề - thuyết của tiếng Việt và chủ vị của tiếng Anh,…) nhưng không liên quan đến ranh giới từ nên không được đặt ra ở đây.

1.3 Nhu cầu xác định ranh giới từ khi dóng hàng từ:

Trong bài toán dóng hàng từ nói trên, chúng ta nhất thiết phải xác định trước tiên đâu là từ để từ đó mới tính đến chuyện dóng hàng từ (Dinh Dien, 2005). Dẫu biết rằng việc xác định ranh giới từ trong tiếng Việt là bài toán cực kỳ khó và đến nay vẫn còn nhiều điều tranh cãi và chưa giải quyết được, nhưng do nhu cầu xử lý thực tế, chúng ta vẫn phải đưa ra một tiêu chí nhất quán nào đó (dù có thể chưa đúng hoàn toàn quan điểm về từ của ngôn ngữ học) để máy tính có thể dựa trên đó mà tiến hành xử lý tự động được (Dien Dinh, 2001).

Các tiêu chí đề nghị phải mang tính hình thức (để máy tính nhận diện tự động được) và tính định lượng cao (đo, đếm được). Các tiêu chí đó cũng phải xét đến nhu cầu sử dụng sau này đối với kho ngữ liệu song ngữ đã tách từ này.

2. TỔNG QUAN VỀ RANH GIỚI TỪ

2.1 Quan niệm về từ trong ngôn ngữ học đại cương:

- Theo L.Bloomfield, thì từ là "một hình thái tự do nhỏ nhất".
- Theo Solncev thì "Từ là đơn vị ngôn ngữ có tính hai mặt : âm và nghĩa. Từ có khả năng độc lập về cú pháp khi sử dụng trong lời".
- Theo B.Golovin, thì từ là "đơn vị nhỏ nhất có nghĩa của ngôn ngữ, được vận dụng độc lập, tái hiện tự do trong lời nói để xây dựng nên câu." . Đây cũng chính là định nghĩa mà trong ngôn ngữ học đại cương hay sử dụng.

Từ các định nghĩa trên, ta có thể rút ra những nét đặc trưng chính của từ như sau:

1. Về hình thức : từ phải là một khối về cấu tạo (mặt chính tả, mặt ngữ âm,…)
2. Về nội dung : từ phải có ý nghĩa hoàn chỉnh.
3. Về khả năng : từ có khả năng hoạt động tự do và độc lập về cú pháp.

Ngoài ra, ta còn gặp một số thuật ngữ khác mà S.E.Jakhontov đưa ra để nhận diện từ, như: từ ngữ âm, từ chính tả, từ hoàn chỉnh, từ từ điển học, từ biến tố,… Trên phương diện xử lý bằng máy tính, thì từ chính tả và từ từ điển là hai loại được nhận diện dễ nhất.

2.2 Đơn vị "tiếng" trong tiếng Việt:

"Tiếng" là đơn vị cơ bản trong tiếng Việt dùng để cấu tạo các đơn vị ngôn ngữ khác cao hơn. Số lượng tiếng trong tiếng Việt không lớn (khoảng 10.000), và chiều dài mỗi tiếng ngắn (không quá 7 chữ cái). Trong xử lý tiếng Việt tự động bằng máy tính, thì "tiếng" là đơn vị tự nhiên nhất mà máy tính dễ dàng lưu trữ, nhận diện và xử lý. Tiếng chính là "từ chính tả".

2.3 Vai trò của "tiếng" trong việc nhận diện "từ" tiếng Việt:

Đối với từ tiếng Việt, đến nay, chúng ta có thể điểm lại một số quan điểm sau:

1. Coi mọi tiếng đều là từ (Nguyễn Thiện Giáp). Điều này thuận tiện trong xử lý nhưng không đúng với các tiêu chí của ngôn ngữ học đại cương (vì có nhiều tiếng không có nghĩa, như: *phê* trong "cà phê", *bù* và *nhìn* trong "bù nhìn";…)

2. Coi tiếng chưa hẳn là từ (đa số các nhà Việt ngữ học). Trong số này, lại chia thành 3 nhóm sau:

a. "Xem tiếng là hình vị": quan niệm có thể chấp nhận được nếu hiểu khái niệm "hình vị" ở đây là hình vị tiếng Việt (gồm "tha hình vị" và "á hình vị" như phần dưới đây).

b. "Xem tiếng lớn hơn hình vị": (chỉ có một số ít người, như: Trần Ngọc Thêm, Lưu Văn Lăng) cho là trong tiếng có những hình vị (khuôn vần), như: "ch – v" có nghĩa là "đơn độc, không chắc chắn" như trong "chon von", "cheo veo",…

c. "Xem tiếng nhỏ hơn hoặc bằng hình vị": Đa số các tiếng đều là hình vị, ngoại trừ: "hấu" trong *dưa hấu*, "bù", "nhìn" trong *bù nhìn*,...vì những tiếng này không có nghĩa.

3. Xem tiếng Châu âu (tiếng Pháp, tiếng Anh,…) cái nào là từ, thì trong tiếng Việt cái đó là từ (bị ảnh hưởng bởi tư tưởng "dĩ Âu vi trung"). Quan niệm này chưa xét đến sự khác biệt về sự từ vựng hoá (lexicalization) giữa hai ngôn ngữ (do sự khác biệt về loại hình ngôn ngữ và loại hình văn hoá).

3. SO SÁNH HÌNH VỊ TIẾNG ANH VỚI TIẾNG VIỆT

Vì tiếng Anh (ngôn ngữ biến hình: inflection) và tiếng Việt (ngôn ngữ đơn lập: isolation) thuộc hai loại hình (typology) ngôn ngữ khác nhau, nên các phương thức ngữ pháp dùng để biểu thị các ý nghĩa ngữ pháp cũng như ý nghĩa từ vựng của hai ngôn ngữ sẽ khác nhau. Dưới đây ta sẽ phân tích những sự khác biệt này: (chỉ gồm các phụ tố mà chương trình có thể xử lý tự động).

3.1 So sánh hậu tố biến cách (inflectional suffixes)

Thay vì dùng phương thức phụ tố như tiếng Anh, thì trong tiếng Việt lại sử dụng phương thức từ hư (function words) để thể hiện các ý nghĩa ngữ pháp. Cụ thể như trong bảng 1 dưới đây:

Bảng 1: Hậu tố biến cách

	Ý nghĩa ngữ pháp	tiếng Anh		tiếng Việt	
		Phụ tố	Ví dụ	Từ hư	Ví dụ
1	Danh từ số nhiều	N + -s	books, two students	những, các	*những/các cuốn sách* hai sinh viên,
2	Động từ ngôi 3 số ít	V + -s	He sleeps	Φ	Anh ấy ngủ

3	Sở hữu cách	X's Y	John's book, teachers' books	của	cuốn sách của John, các cuốn sách của những giáo viên,
4	Hiện phân từ	V-ing	sleeping	đang	*đang ngủ*
5	Quá khứ/quá phân từ	V-ed	worked	đã	(đã) làm việc
6	So sánh hơn	Adj-er Adv-er	shorter slower	hơn	ngắn hơn chậm hơn
7	So sánh nhất	Adj-est Adv-est	shortest slowest	nhất	ngắn nhất chậm nhất

3.2 So sánh hậu tố dẫn xuất (derivational suffixes)

Tương tự như trên, thay vì dùng phương thức phụ tố như tiếng Anh, thì trong tiếng Việt lại sử dụng phương thức từ thực (tha hình vị tựa phụ tố) để thể hiện các ý nghĩa từ vựng. Ví dụ: *read*,v : đọc + *-able* (*có thể ~ được*) => *có thể đọc được*.

Bảng 2: Luật sinh của một số hậu tố dẫn xuất:

Stt	Hậu tố	Từ loại gốc	Từ loại mới	Loại	Nghĩa tiếng Việt	Ghi chú, Ví dụ
1.	able	V	A	2	có thể ~ được	readable
2.	al	A,N	A	3	(thuộc về) ~	national
3.	ate	N	V	3	làm cho ~	fascinate
4.	ed*	V	A-vpp	1/3	(đã được / bị) ~	closed-door
5.	en	N	A	1	làm bằng ~	golden
6.	er*	V	N	1/3	người/máy ~	teacher, printer
7.	ing*	V	Ger	1/3	(đang) ~	running car
8.	ise/ize	A,N	V	3	~ hoá	normalise, computerize
9.	ity	A	N-abs	3	sự ~	activity
10.	less	A, N	A	3	không có ~	careless
11.	like	N	A	3	giống như ~	humanlike
12.	ly	A	Adv	2	(một cách) ~	strongly
13.	ness	A	N-abs	3	sự ~	brightness
14.	tion	V	N-abs	3	sự ~	solution

Lưu ý:

- Các hậu tố đánh dấu * là những hậu tố bị trùng với hậu tố của biến cách.
- Loại 1: là loại chỉ nằm ở cuối từ, không thể thêm bất kỳ hậu tố nào.
- Loại 2: là loại nằm ở cuối từ, và chỉ có thể thêm hậu tố biến cách.
- Loại 3: là loại có thể thêm bất kỳ hậu tố nào.
- Loại 4: là loại chỉ gắn trực tiếp với thân từ mà thôi.

3.3 So sánh tiền tố dẫn xuất (derivational prefixes)

Ví dụ: president, : chủ tịch + vice- (phó ~) (phó chủ tịch.

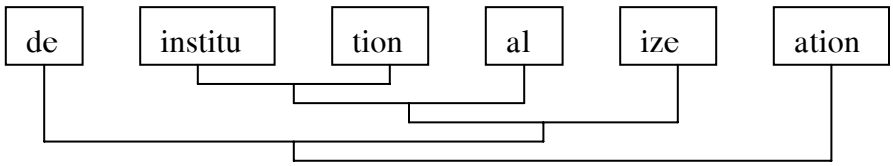
Bảng 3: Tiền tố dẫn xuất (POS là từ loại thường được kết hợp):

Stt	Tiền tố	POS	Nghĩa tiếng Việt	Ghi chú, Ví dụ
1.	anti	N	chống ~, kháng ~	antivirus
2.	co	N	đồng ~, liên ~	co-author
3.	dis	V	khử ~	discharge
4.	in, il, im, ir (*)	A	không ~, bất ~, vô ~	illegal, impatient, irregular
5.	re	V	~ lại	re-calculate
6.	un	A,V	không ~	unhappy

(*): "in-" biến thể thành "il-" khi đứng trước "l"; thành "im-" khi đứng trước "b", "m" hay "p" và thành "ir-" khi đứng trước "r". (xin xem thêm [Đỗ Đình Lan, 1993])

3.4 So sánh trật tự kết hợp các hình vị

Việc kết hợp các hình vị trong từ tiếng Anh theo nguyên tắc từ trong ra ngoài (xuất phát từ thân từ), từ trái sang phải đối với hậu tố và từ phải sang trái đối với tiền tố. Quá trình kết hợp phải tuân theo qui luật "phù hợp từ loại" (nghĩa là phụ tố nào kết hợp với từ loại nào). Ví dụ: Xét từ "deinstitutionalization", ta sẽ có qui cách kết hợp như sau:



Trong khi đó, đối với tiếng Việt, tuy việc kết hợp cũng xuất phát từ thân từ, nhưng trật tự lại được qui định riêng bởi từng phụ tố bởi vì trật tự các thành tố (âm tiết) này tuỳ thuộc vào loại từ "Hán-Việt" (ngược cú pháp tiếng Việt) hay "thuần Việt" (thuận cú pháp tiếng Việt) và thêm một số hư từ khác (đã/đang, được/bị,..). Các trật tự / hư từ này đã được ghi trong các bảng so sánh trên (bảng 1,2,3). Ví dụ: "un-program-able" => "*không (có) thể lập trình được*".

4. QUAN NIỆM VỀ TỪ TRONG VIỆC XỬ LÝ SONG NGỮ ANH-VIỆT

4.1 Quan niệm "hình vị" tiếng Việt:

Chúng tôi theo quan niệm "xem tiếng là hình vị". Tuy nhiên, hình vị ở đây phải hiểu là hình vị tiếng Việt, nghĩa là bên cạnh hình vị như trong ngôn ngữ học đại cương, ta còn phải có hình tố (là yếu tố thuần tuý hình thức biểu hiện những kiểu quan hệ bên trong giữa các thành tố trong từ, ta có thể gọi đây là những "tha hình vị" hay "á hình vị"). Như vậy, trong tiếng Việt ta sẽ có 3 loại hình vị ([Hoàng Văn Hành, 1998] trang 40-48) như sau:

- Hình vị gốc: là những nguyên tố, đơn vị nhỏ nhất, có nghĩa, chúng có thể là hình vị thực (từ vựng) hay hình vị hư (ngữ pháp), chúng có thể độc lập (tự do) hay hạn chế (ràng buộc).

- Tha hình vị: vốn cũng là hình vị gốc, song do mối tương quan với các thành tố khác trong từ mà chúng biến đổi đi về âm, nghĩa, …Tha hình vị bao gồm:

- Tha hình vị láy âm, như: *chúm chím, đo đỏ*, … ; nhưng phải cả chỉnh thể *lé đé, đủng đỉnh* mới được coi là hình vị vì ta không xác định được nghĩa của hình vị gốc.

- Tha hình vị láy nghĩa: trong các từ ghép hội nghĩa, như: *giá cả, hỏi han, tuổi tác,…; nhà cửa, yêu thương, ngược xuôi,…*

- Tha hình vị định tính: là các yếu tố phụ để miêu tả thuộc tính, như: *xanh lè, tối om, cười khẩy,…*

- Tha hình vị tựa phụ tố: là đơn vị hoạt động giống như những phụ tố (affix) trong các ngôn ngữ biến hình, như: *giáo viên, hiện đại hoá, tân tổng thống*,…

- Á hình vị: là những chiết đoạn ngữ âm được phân xuất một cách tiêu cực, thuần tuý dựa vào hình thức, không rõ nghĩa, song có giá trị khu biệt, làm chức năng cấu tạo từ. Ví dụ như: *dưa hấu, dưa gang, bí ử, đậu nành, cà niễng, bồ nông*, …

4.2 Quan niệm "từ" trong xử lý song ngữ Anh-Việt:

Về cơ bản, chúng tôi theo quan niệm cũng giống như trong ngôn ngữ học đại cương: nghĩa là từ được cấu tạo từ những hình vị mà đã được nêu ở phần trên. Tuy nhiên, để thuận tiện trong bài toán dóng hàng "từ" giữa tiếng Anh và tiếng Việt trong song ngữ Anh-Việt, chúng tôi còn tuân theo nguyên tắc sau:

a. Các hình vị dẫn xuất (derivation) trong tiếng Anh, khi dịch sang tiếng Việt được thể hiện bằng các tiếng tương ứng, chúng tôi xem các tiếng này như là những tha hình vị tựa phụ tố như đã định nghĩa trong phần 4.1. Ví dụ: *caller (người gọi), vice-president (phó tổng thống), normalize (bình thường hoá), non-government (phi chính phủ), …*

b. Các hình vị biến cách (inflection) trong tiếng Anh, khi dịch sang tiếng Việt được thể hiện bằng các tiếng tương ứng (phương thức từ hư), chúng tôi không xem các

tiếng này là những hình vị thuộc từ, mà xem chúng là những từ riêng rẽ (từ hư) để thể hiện ý nghĩa ngữ pháp của từ. Ví dụ: *books (những cuốn sách)*[số], *working (đang làm việc)* [thời]*, reached (đã đạt tới), won (đã thắng)* [thì]*, …Tương* tự cho các phó động từ chỉ đích/hướng của các động từ, như: *chạy ra/vào/lên/xuống; rơi xuống, rắc lên, tìm ra, nhận được, …* không được xem là hình vị của động từ.

c. Đối với các danh từ chỉ loại trong tiếng Việt, như: *cái, con, cuốn, lá, tấm,*… chúng tôi cũng xem nó là một từ độc lập để chỉ đơn vị cho danh từ. Ví dụ: *book (cuốn sách)*, *letter (lá thư / bức thư / cánh thư), house (ngôi nhà)* (NTCẩn, tr.187-239).

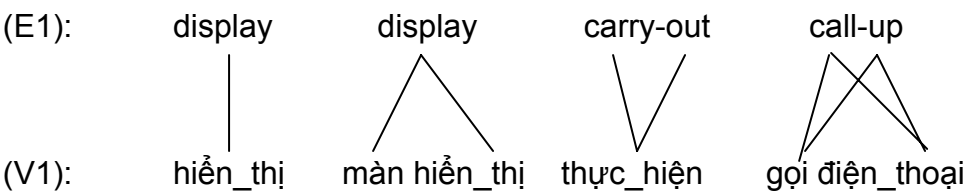
d. Ta cần phân biệt các danh từ chỉ loại với các danh từ đơn vị quy ước được dùng trong "cân, đong, đo, đếm" như: *tờ (giấy), đàn (gà), tạ (thóc),*… Đối với loại này, tiếng Anh dùng dạng danh ngữ như “sheet of” (tờ, tấm), “piece of” (miếng, mẩu), “pack of” (gói). Trong trường hợp này, chúng tôi xem các danh từ “sheet”, “piece”, “pack” tương đương các danh từ chỉ đơn vị quy ước trong tiếng Việt, chứ không tích hợp bên trong chính danh từ. Ví dụ: *sheet of paper (tờ giấy), piece of cake (miếng bánh), pack of cigarettes (gói thuốc lá)*,..

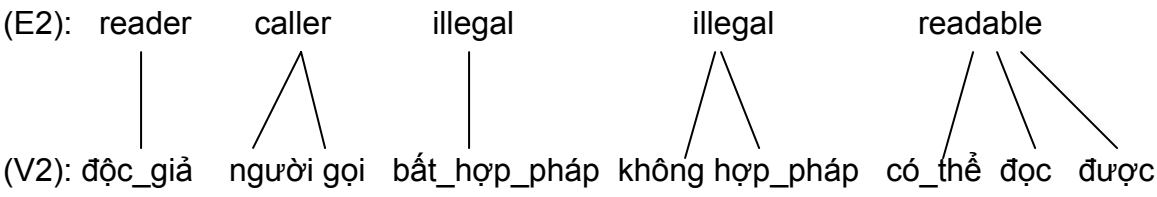
e. Ngoài ra, các từ chỉ chủng loại, như: *cây, máy, hoa, cá,* … đều được xem là các hình vị chỉ loại và được tích hợp bên trong chính danh từ đó. Ví dụ: *cây tre, cây chuối, trái chuối, trái hồng; máy in, máy tính; hoa hồng, hoa lan; cá hồng, cá rô; ...* . Chúng tôi xem đây là các từ ghép định danh bậc 1, đối với các từ ghép định danh bậc 2 trở lên, các thành tố hạn định không được xem là hình vị thuộc từ. Ví dụ: *máy_in tự_động* (2 từ), *máy_xay sinh_tố* (2 từ), *cây_tre lá nhọn* (3 từ),…

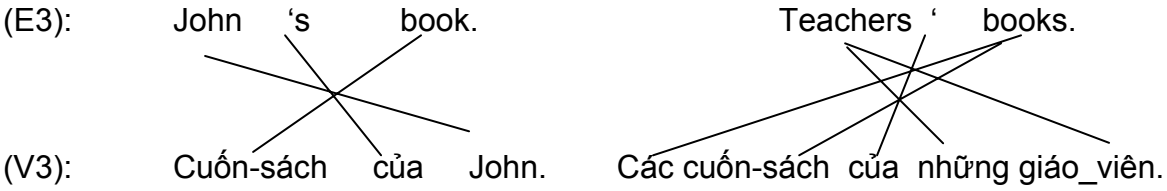
f. Nếu một khái niệm nào đó mà đã được từ vựng hoá trong tiếng Việt, nhưng ở tiếng Anh vẫn phải dùng cụm từ hay thành ngữ (idiom) thì khi liên kết với tiếng Việt, chúng tôi xem cụm từ / thành ngữ tiếng Anh đó là một “từ từ điển”. Ví dụ: “to *lead by the hand” (dìu), “black horse” (ngựa ô); carry out (thực hiện), make up one’s mind (quyết định), pick … up (đón), …*

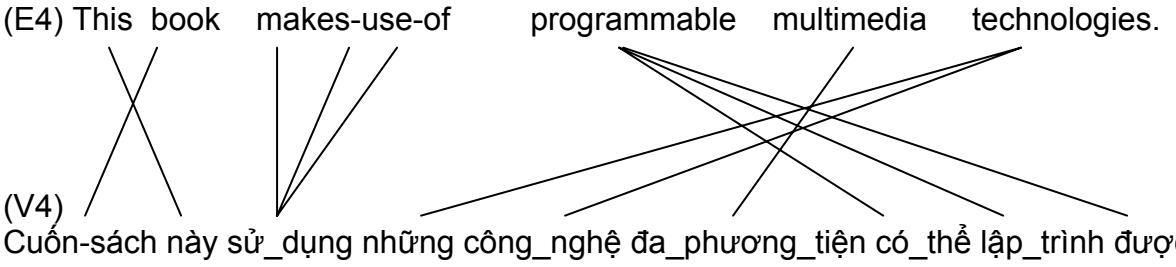
g. Đối với một số ít đơn vị tiếng Việt còn đang tranh cãi về tư cách từ của nó, chúng tôi sẽ dựa theo sự từ vựng hoá trong tiếng Anh. Chẳng hạn: *nhà_tranh (line), xe_đạp (bicycle), máy_tính(computer), đường_thẳng (line)*, *puppet (bù nhìn), watermelon (dưa hấu), hen (gà mái), waterpox (bệnh thuỷ đậu), to marriage (lấy vợ, lấy chồng)*… là từ; còn *nhà gạch (brick house), ..* không là từ.

Một số ví dụ minh hoạ:









5. KẾT LUẬN

Trong ngôn ngữ học, đã có cả hàng trăm định nghĩa về từ đã được đưa ra. Các định nghĩa ấy, ở mặt này hay mặt khác đều đúng, nhưng đều không đủ và không bao gồm hết được tất cả các sự kiện được coi là từ trong các ngôn ngữ và ngay cả trong một ngôn ngữ cũng vậy. Tuy nhiên, để thống nhất trong việc lựa chọn đơn vị nào là "từ" trong quá trình xử lý ngữ liệu song ngữ Anh-Việt, chúng tôi đã tạm đưa ra các tiêu chí lựa chọn trên đây. Các tiêu chí này có thể chưa thoả đáng về mặt ngôn ngữ học, nhưng vì yếu tố thuận lợi và nhất quán trong việc xử lý tự động ngữ liệu song ngữ Anh-Việt, nên các tiêu chí này vẫn có thể chấp nhận được. Ngoài ra, các tiêu chí này có thể được bổ sung, điều chỉnh ở một vài điểm nhỏ để phù hợp hơn với tình hình thực tế. Chúng tôi hy vọng rằng các tiêu chí này sẽ làm nền tảng cho mọi xử lý tiếng Việt tự động trên máy tính về sau.

Lời cảm ơn: đề tài này được thực hiện dưới sự tài trợ kinh phí trong chương trình KC-01. Chúng tôi xin chân thành cảm ơn các tổ chức đã tài trợ thực hiện dự án này.

Tài liệu tham khảo

1. Nguyễn Tài Cẩn (1998), Ngữ pháp tiếng Việt. NXB ĐHQG Hà Nội.

2. Đỗ Hữu Châu (1997), Các bình diện của từ và từ tiếng Việt. NXB ĐHQG Hà Nội.

3. Dien Dinh, Kiem Hoang, Toan Nguyen Van (2001), "Vietnamese Word Segmentation", *Proceedings of NLPRS'01* (The 6th Natural Language Processing Pacific Rim Symposium), Tokyo, Japan, 11/2001, pg 749-756.

4. Dien Dinh (2005), "Building an Annotated English-Vietnamese parallel Corpus", *MKS: A Journal of Southeast Asian Linguistics and Languages,* Vol. 35, pp. 21-36.

5. Đinh Điền (2002a), "Ứng dụng Ngữ liệu song ngữ Anh-Việt điện tử trong ngành ngôn ngữ học so sánh", Tạp chí Ngôn ngữ, Viện Ngôn ngữ học, số 3-2002, tr. 49-58.

6. Đinh Điền (2002b), "Xây dựng và khai thác kho ngữ liệu song ngữ Anh-Việt điện tử , luận văn tiến sĩ Ngôn ngữ học so sánh, trường ĐH Khoa học Xã hội & Nhân văn –ĐHQG TPHCM, tháng 2/2005.

7. Nguyễn Thiện Giáp (1996), Từ và Nhận diện từ tiếng Việt. NXB GD, Hà Nội.

8. Hoàng Văn Hành (chủ biên) – Hà Quang Năng – Nguyễn Văn Khang (1998), Từ tiếng *Việt: hình thái – cấu trúc – từ láy – từ ghép – chuyển loại. NXB KHXH. Hà Nội.*

9. Cao Xuân Hạo (1998), Tiếng Việt: mấy vấn đề về ngữ âm – ngữ pháp – ngữ nghĩa. NXB GD.

10. Đỗ Đình Lan (1993), Lexicology (tập 1 và 2). Trường CĐSP-TPHCM.

11. Viện ngôn ngữ học (2000), Loại từ trong các ngôn ngữ ở Việt Nam, NXB KHXH, Hà Nội.

12. McEnery T., Wilson A. (1996), *Corpus Linguistics*, Edinburgh University Press.