



ĐẠI HỌC BÁCH KHOA HÀ NỘI  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

# Dịch máy

**Viện CNTT & TT – Trường ĐHBKHN**

# Ví dụ

- Au sortir de la saison 97/98 et surtout au debut de cette saison 98/99...
- With leaving season 97/98 and especially at the beginning of this season 98/99...

# Các vấn đề

## 1. Xử lý sự giống và khác nhau giữa các ngôn ngữ

- Hình vị: # số âm tiết/từ:
  - *Ngôn ngữ đơn âm tiết ( tiếng Việt, Trung Quốc) – 1 tiếng/từ*
  - *Ngôn ngữ đa âm tiết (Siberian Yupik), 1 từ = cả 1 câu*
- Mức độ phân chia âm tiết

# Các vấn đề

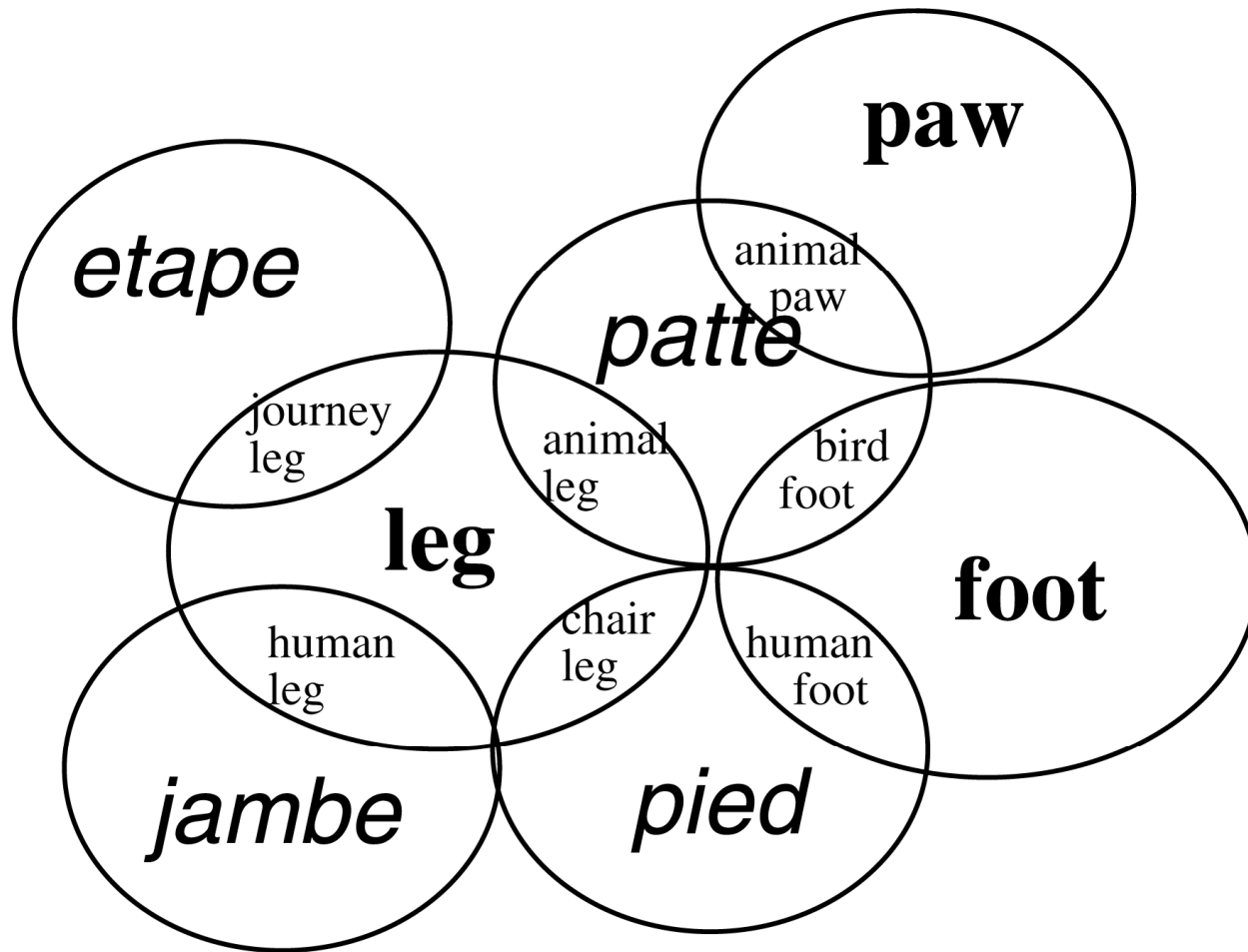
## 2. **Cú pháp:** trật tự từ trong câu

- *To Yukio; Yukio ne*
- Tiếng Anh – tiếng Việt:
  - *The* (affix1) *red* (affix2) *flag* (head)
  - *Lá cờ* (head) *đỏ* (affix2) *ấy* (affix1)

## 3. **Các nét riêng biệt**

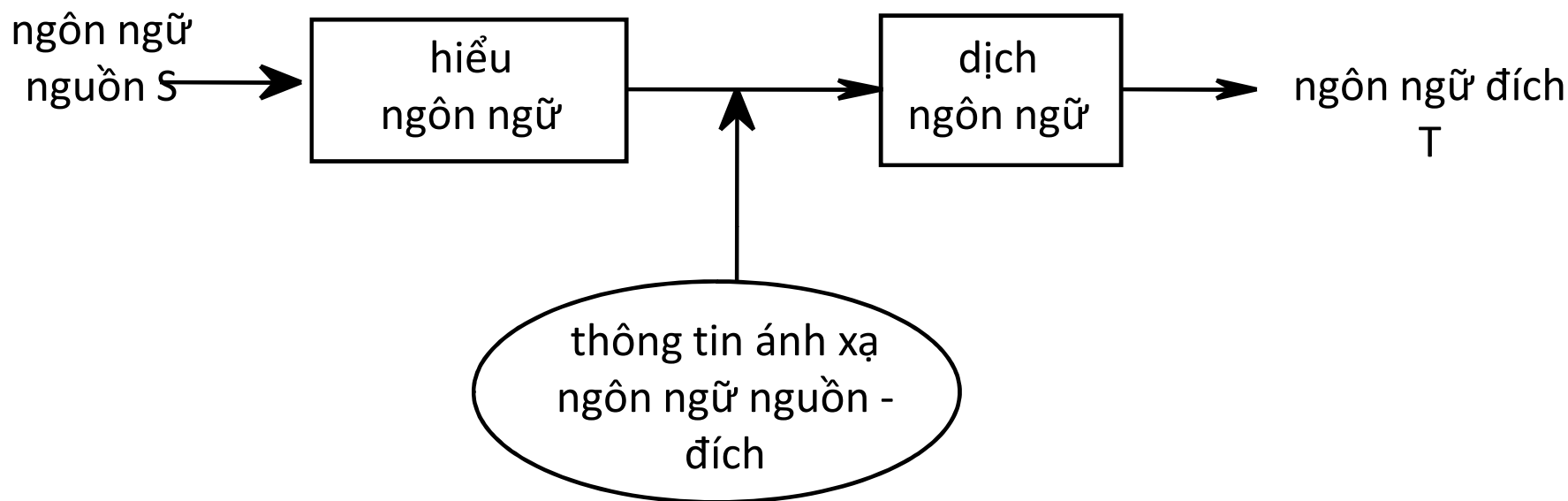
|         |         |            |                                 |
|---------|---------|------------|---------------------------------|
| English | brother | Vietnamese | anh<br>em                       |
| English | wall    | German     | wand (inside)<br>mauer(outside) |
| German  | berg    | English    | hill<br>mountain                |

# Không gian khái niệm



Khoảng trống từ vựng: tiếng Nhật không có từ nào nghĩa *privacy*;  
tiếng Anh không có từ ứng với *yakoko* (lòng hiếu thảo)

# Ba khối chính trong dịch máy



# Hiểu ngôn ngữ

## 1. Nhập nhằng từ vựng:

English: *book* - Spanish *libro, reservar*


⇒ Sử dụng thông tin cú pháp

## 2. Nhập nhằng cú pháp:

*I saw the guy on the hill with the telescope*

### 3. Nhập nhằng ngữ nghĩa:

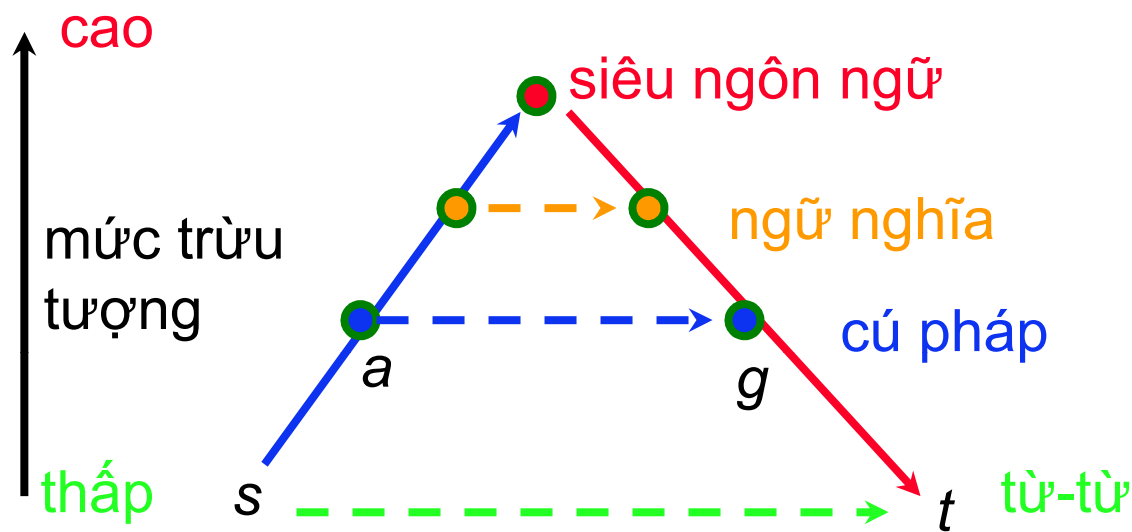
- *E: While driving, John swerved & hit a tree*



John's car

- *S: Minetras que John estaba manejando, se desvio y golpeop con un arbo*

# Các phương pháp dịch máy



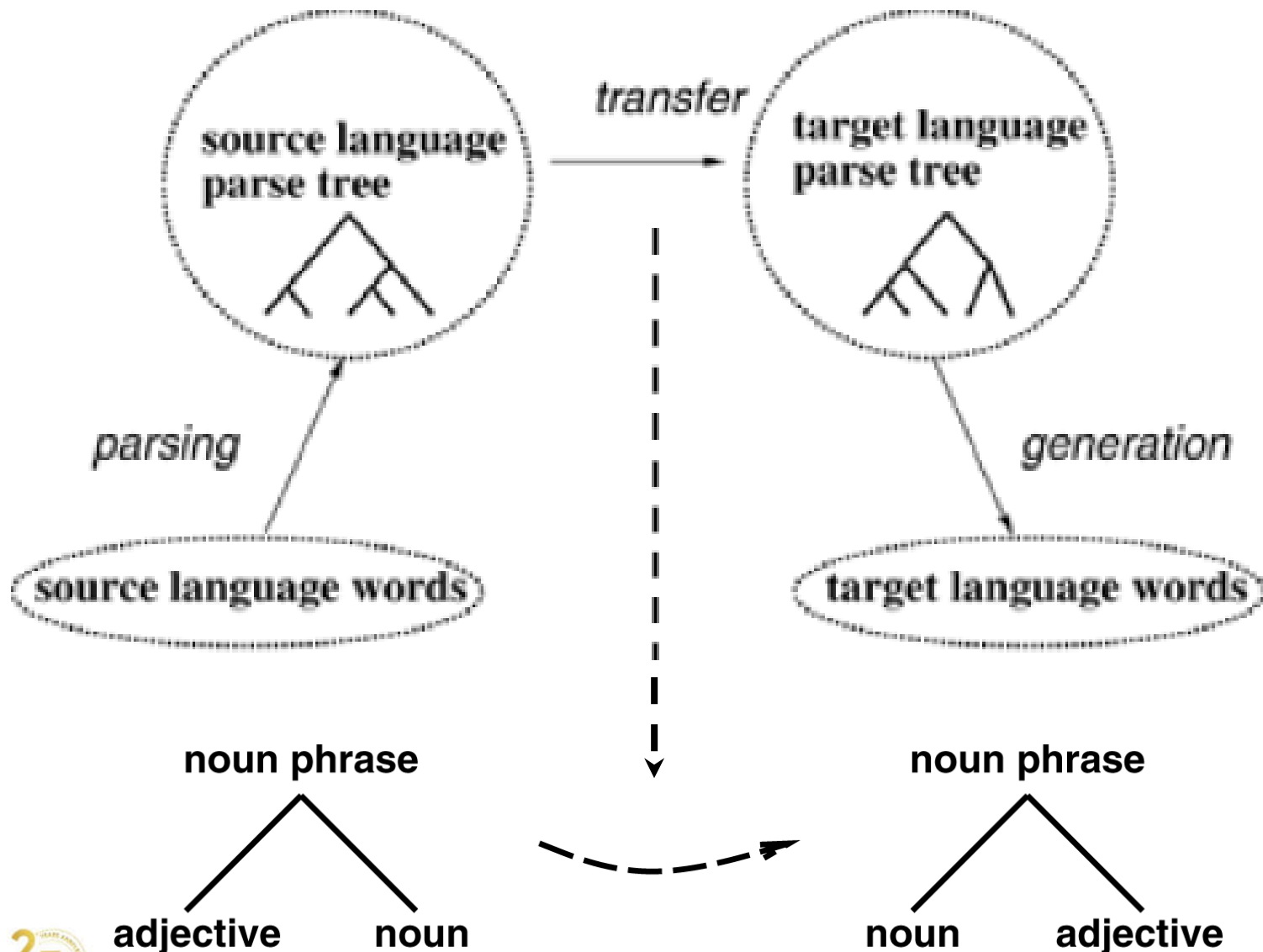
$$a = a(s)$$

$$g = f(a(s)); f - \text{hàm chuyển đổi}$$

$$t = g(f(a(s)))$$



# Sơ đồ chuyển đổi



# Luật chuyển đổi

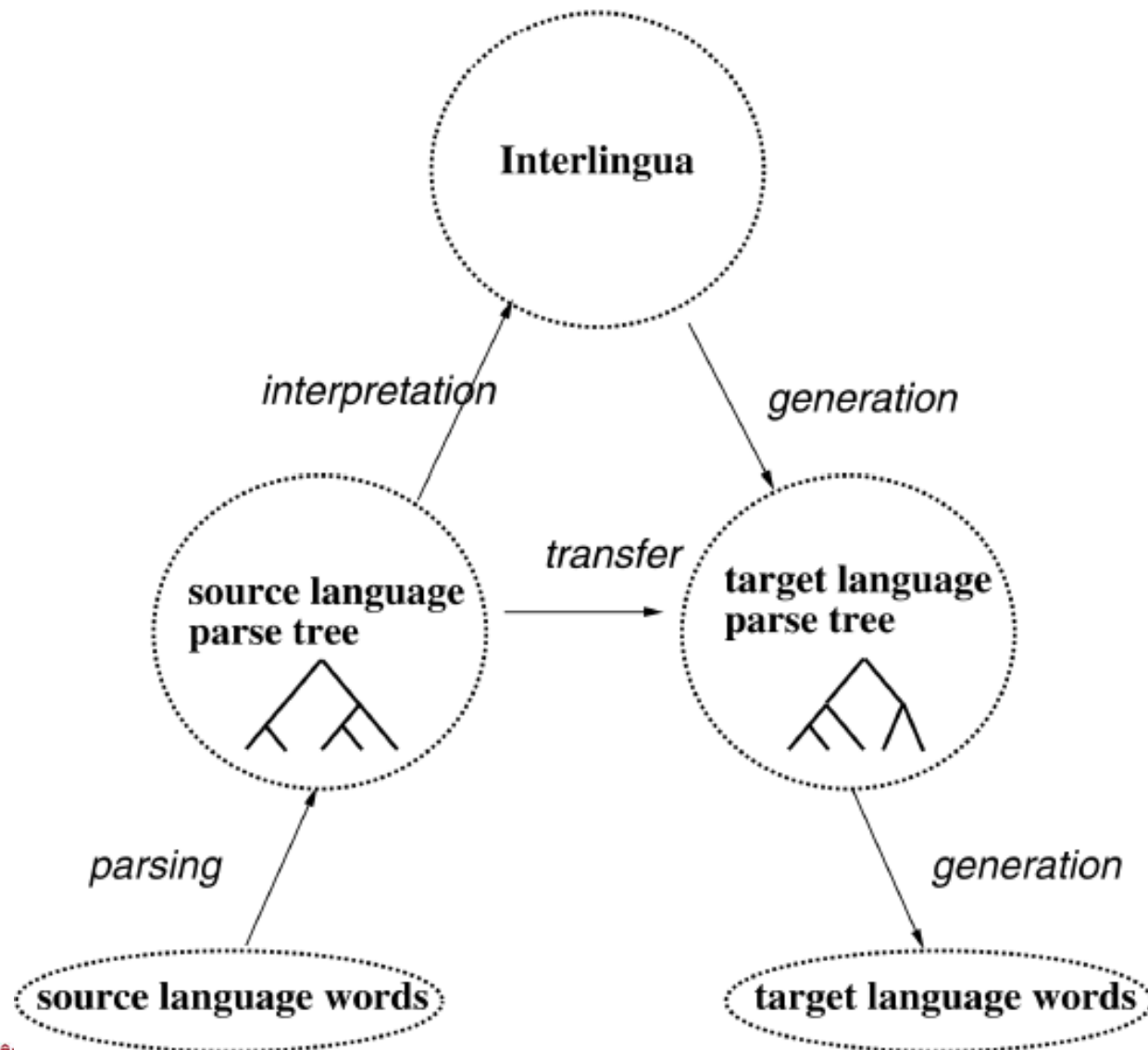
English to French:

1. NP  $\rightarrow$  Adjective<sub>1</sub> Noun<sub>2</sub>  
 $\Rightarrow$   
NP  $\rightarrow$  Noun<sub>2</sub> Adjective<sub>1</sub>

Japanese to English:

2. Existential-There-Sentence  $\rightarrow$  There<sub>1</sub> Verb<sub>2</sub> NP<sub>3</sub> Postnominal<sub>4</sub>  
 $\Rightarrow$   
Sentence  $\rightarrow$  (NP  $\rightarrow$  NP<sub>3</sub> Relative-Clause<sub>4</sub>) Verb<sub>2</sub>
3. NP  $\rightarrow$  NP<sub>1</sub> Relative Clause<sub>2</sub>  
 $\Rightarrow$   
NP  $\rightarrow$  Relative-Clause<sub>2</sub> NP<sub>1</sub>

# Sơ đồ chuyển đổi



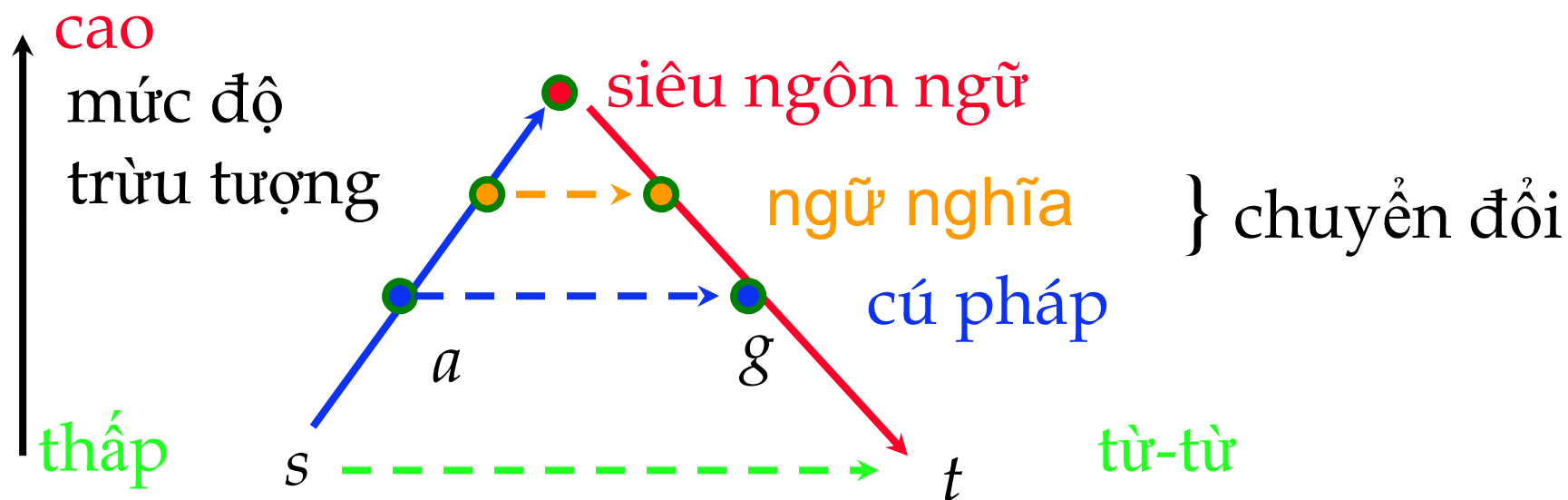
# Cách tiếp cận siêu ngôn ngữ: sử dụng nghĩa

- Chuyển đổi: các luật chuyển đổi từ ngôn ngữ này sang ngôn ngữ khác
- Đối tượng/sự kiện (ontology)

|              |  |     |  |        |    |              |       |
|--------------|--|-----|--|--------|----|--------------|-------|
| event        | gardening  |     |  |        |    |              |       |
| agent        | <table><tr><td>man</td><td></td></tr><tr><td>number</td><td>sg</td></tr><tr><td>definiteness</td><td>indef</td></tr></table> | man |  | number | sg | definiteness | indef |
| man          |  |     |  |        |    |              |       |
| number       | sg   |     |  |        |    |              |       |
| definiteness | indef  |     |  |        |    |              |       |
| aspect       | progressive  |     |  |        |    |              |       |
| tense        | past   |     |  |        |    |              |       |

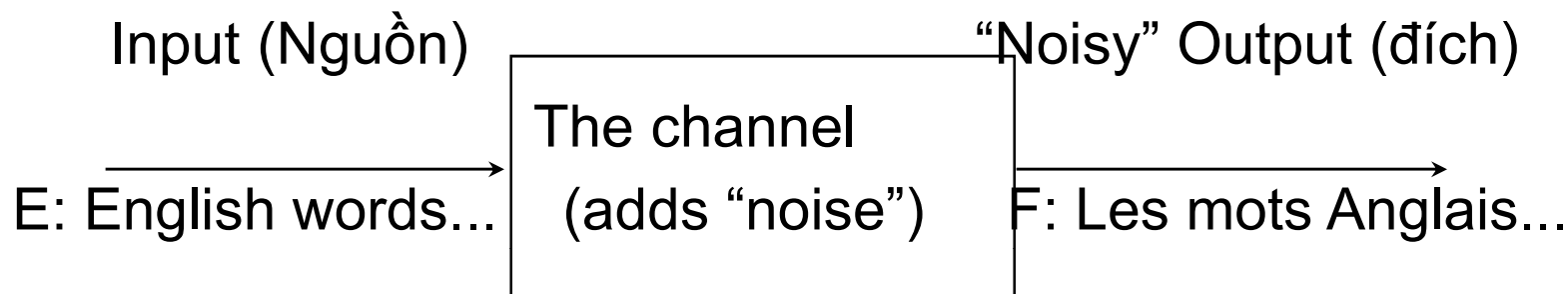
# Dịch máy thống kê

# Các kiểu dịch máy



# ý tưởng

- Coi việc dịch như bài toán kênh có nhiễu

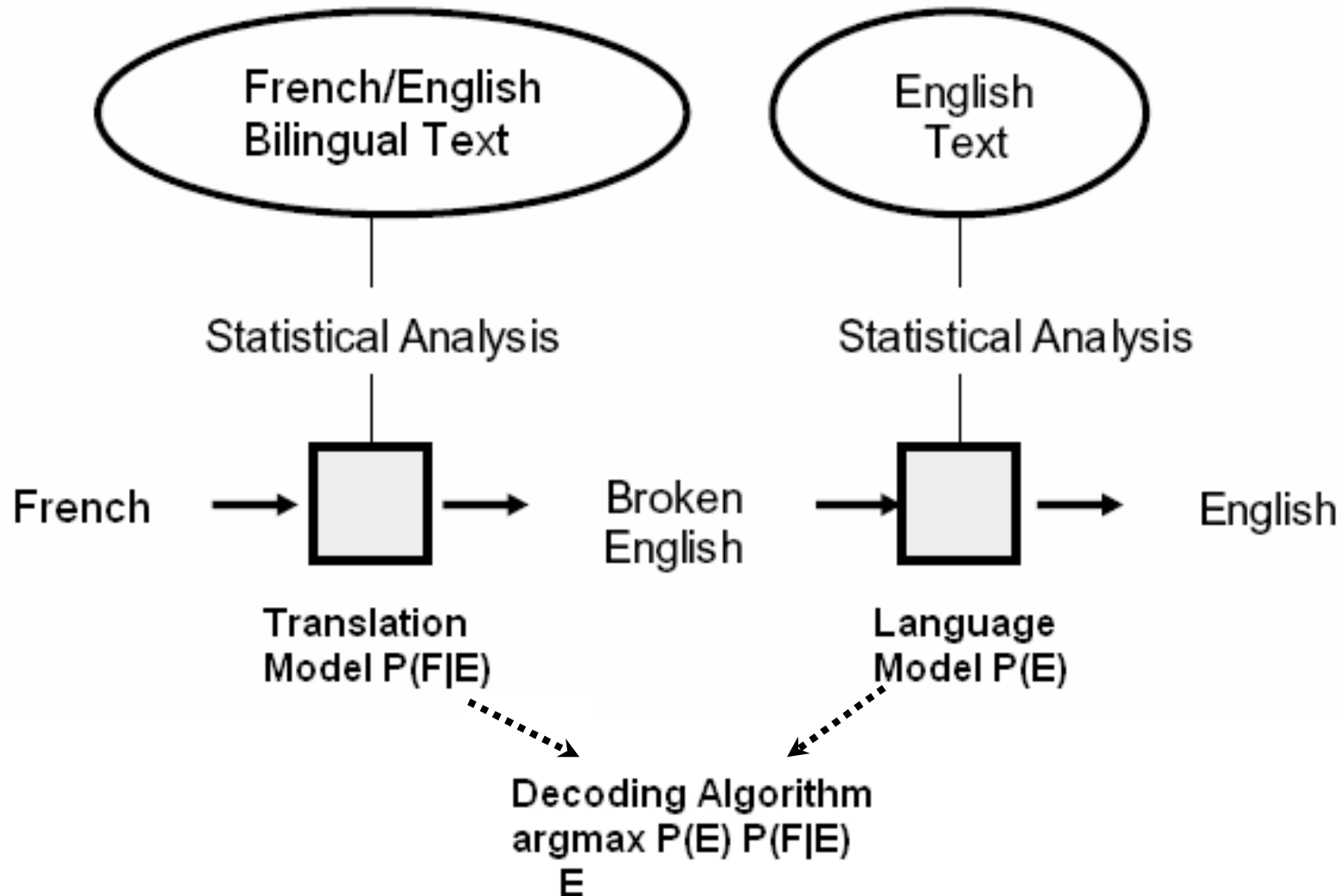


- Mô hình dịch:  $P(E|F) = P(F|E) P(E) / P(F)$
- Khôi phục lại  $\underline{E}$  khi biết  $\underline{F}$ :

Sau khi đơn giản hóa ( $P(F)$  không đổi):

$$\operatorname{argmax}_E P(E|F) = \operatorname{argmax}_E P(F|E) P(E)$$

# Dịch máy thống kê





# Các yếu tố

- **Mô hình ngôn ngữ - Language Model (LM)**: xác suất thấy 1 câu tiếng Anh (E) (xác suất tiên nghiệm):  
 $P(E)$
- **Mô hình dịch - Translation Model (TM)**: câu đích trong tiếng Pháp (F) khi có câu tiếng Anh:  
 $P(F|E)$
- Thủ tục tìm kiếm:
  - Cho F, tìm E tốt nhất sử dụng mô hình ngôn ngữ LM và mô hình dịch TM.
- Vấn đề: thiếu dữ liệu!
  - Ta không thể tạo từ điển câu  $E \leftrightarrow F$
  - Thậm chí bình thường ta không thấy 1 câu lặp lại 2 lần

# Ý tưởng giống hàng

- Mô hình dịch TM không quan tâm đến chuỗi đúng các từ tiếng Anh
- Sử dụng cách tiếp cận gán nhãn:
  - 1 từ tiếng Anh (“tag”) ~ 1 từ tiếng Pháp (“word”)  
→ không thực tế: thậm chí số từ trong 2 câu không bằng nhau  
→ sử dụng “giống hàng”.

# Ý tưởng giống hàng

- Các tập ngữ liệu sử dụng giả thiết:
  - Dữ liệu song song (dịch  $E \leftrightarrow F$ )
- Giống hàng câu
  - Phát hiện câu
  - Giống hàng câu
- Giống hàng từ
  - Tách từ
  - Giống hàng từ (với 1 số ràng buộc)

# Giống hàng câu

The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.

# Giống hàng câu

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

# Giống hàng câu

- |                              |  |  |
|------------------------------|--|--|
| 1. The old man is happy.     |  | 1. El viejo está feliz porque ha pescado muchos veces. |
| 2. He has fished many times. |  | 2. Su mujer habla con él.                              |
| 3. His wife talks to him.    |  | 3. Los tiburones esperan.                              |
| 4. The fish are jumping.     |  |  |
| 5. The sharks await.         |  |  |

## Khó khăn:

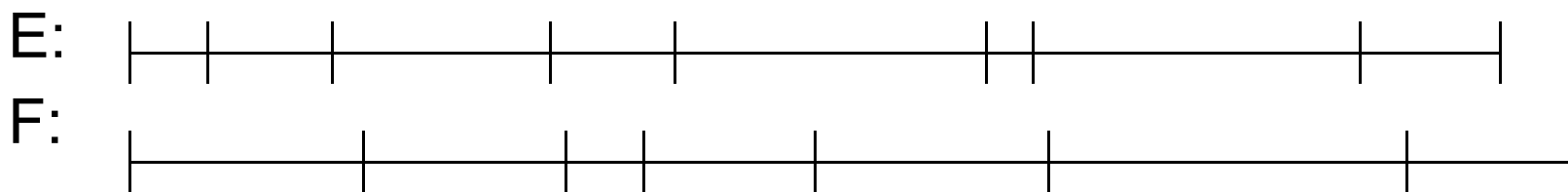
- ☛ **Sự liên quan chéo:** trật tự câu thay đổi khi dịch

# Phát hiện biên của câu

- Sử dụng luật, danh sách liệt kê:
  - Dấu kết thúc câu:
    - Dấu ngắt đoạn (nếu được đánh dấu)
    - 1 số ký tự: ?, !, ;
    - Ván đề: dấu chấm '.'
      - Kết thúc câu (... left yesterday. He was heading to...)
      - Dấu chấm thập phân : 3.6 (three-point-six)
      - Dấu chấm hàng nghìn: 3.200
      - Viết tắt: cf., e.g., Calif., Mt., Mr.
      - Vân vân: ...
      - 1 số ngôn ngữ: 2nd ~ 2.
      - Ký hiệu đầu: A. B. Smith
- Phương pháp thống kê: vd Maximum Entropy

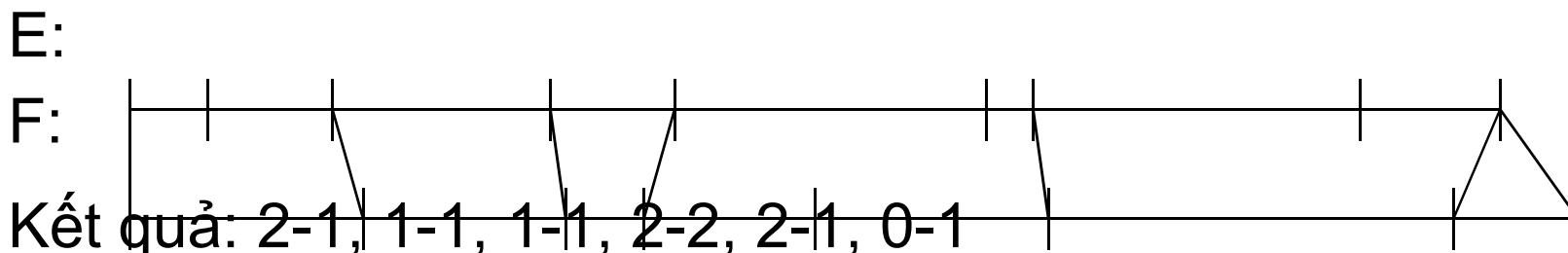
# Giống hàng câu

- Vấn đề với phát hiện biên của câu:



- Đầu ra mong đợi**: Các phân mảnh với cùng số lượng mảnh liên tiếp nhau.

- Giống hàng:



- Kết quả: 2-1, 1-1, 1-1, 2-2, 2-1, 0-1



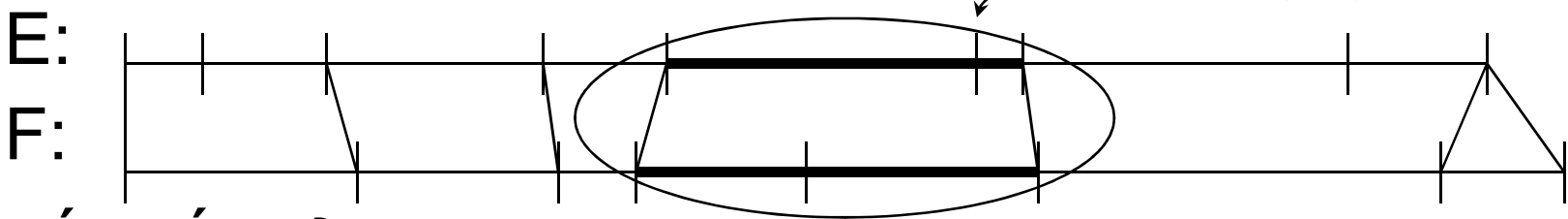
# Các phương pháp giống hàng

- Nhiều phương pháp (xác suất hoặc không)
  - Dựa trên độ dài ký tự
  - Dựa trên độ dài từ
  - “cùng gốc” (sử dụng nghĩa từ)
    - Sử dụng từ điển (F: prendre ~ E: make, take)
    - Sử dụng khoảng cách từ (độ tương tự): tên, số, từ vay mượn, từ gốc Latin
- Kết quả tốt nhất:
  - Thống kê, dựa trên từ hoặc dựa trên ký tự

# Giống hàng dựa trên độ dài

- Định nghĩa bài toán như việc tính xác suất:  
 $\operatorname{argmax}_A P(A|E,F) = \operatorname{argmax}_A P(A,E,F)$  ( $E,F$  cố định)

- Định nghĩa 1 “bead”:



- Lấy xấp xỉ:

$$P(A,E,F) \cong \prod_{i=1..n} P(B_i),$$

Trong đó  $B_i$  là 1 bead;  $P(B_i)$  không phụ thuộc vào phần còn lại của  $E,F$ .

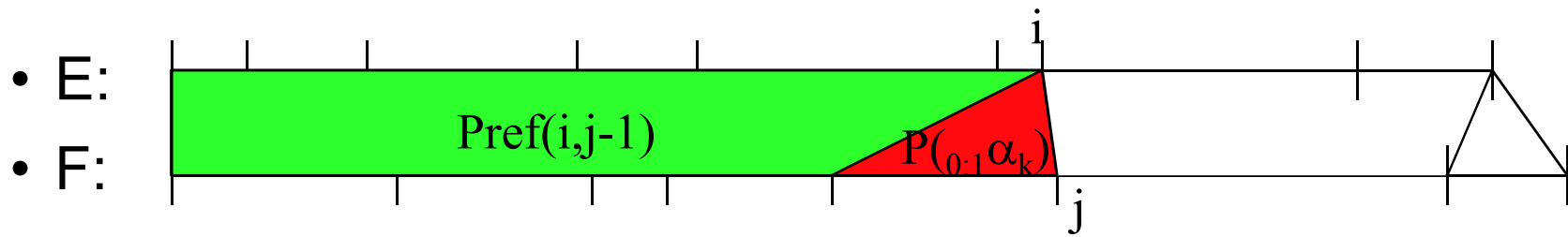
# Nhiệm vụ giống hàng

## Định nghĩa:

- Cho  $P(A, E, F) \cong \prod_{i=1..n} P(B_i)$ ,  
tìm cách chia  $(E, F)$  thành  $n$  bead  $B_{i=1..n}$ , sao cho  
tối đa xác suất  $P(A, E, F)$  trên tập luyện.
- $B_i =_{p:q} \alpha_i$ , với  $p:q \in \{0:1, 1:0, 1:1, 1:2, 2:1, 2:2\}$   
mô tả phép giống hàng
- $\text{Pref}(i, j)$  – xác suất của cách giống hàng tốt nhất từ  
điểm đầu cho đến  $(i, j)$

# Định nghĩa đệ quy

- Khởi tạo:  $\text{Pref}(0,0) = 1$ .
- $\text{Pref}(i,j) = \max ($   
 $\text{Pref}(i,j-1) P_{(0:1}\alpha_k), \text{Pref}(i-1,j) P_{(1:0}\alpha_k), \text{Pref}(i-1,j-1) P_{(1:1}\alpha_k),$   
 $\text{Pref}(i-1,j-2) P_{(1:2}\alpha_k), \text{Pref}(i-2,j-1) P_{(2:1}\alpha_k), \text{Pref}(i-2,j-2) P_{(2:2}\alpha_k) )$



# Xác suất của 1 Bead

- Định nghĩa  $P_{(p:q)}\alpha_k$ :
  - $k$  đề cập đến “bead” kế tiếp, với các đoạn của câu  $p$  và  $q$ , độ dài  $l_{k,e}$  và  $l_{k,f}$ .
- Sử dụng phân bố chuẩn cho các độ dài khác nhau:
$$P_{(p:q)}\alpha_k = P(\delta(l_{k,e}, l_{k,f}, \mu, \sigma^2), p:q) \cong P(\delta(l_{k,e}, l_{k,f}, \mu, \sigma^2))P(p:q)$$
$$\delta(l_{k,e}, l_{k,f}, \mu, \sigma^2) = (l_{k,f} - \mu l_{k,e}) / \sqrt{l_{k,e} \sigma^2}$$
- Đánh giá  $P(p:q)$  từ tập dữ liệu nhỏ, hoặc đoán và đánh giá lại sau khi giống hàng
- Từ có thể được dùng như dấu hiệu tốt hơn để định nghĩa  $P_{(p:q)}\alpha_k$ .

# Giống hàng từ - Mức dễ

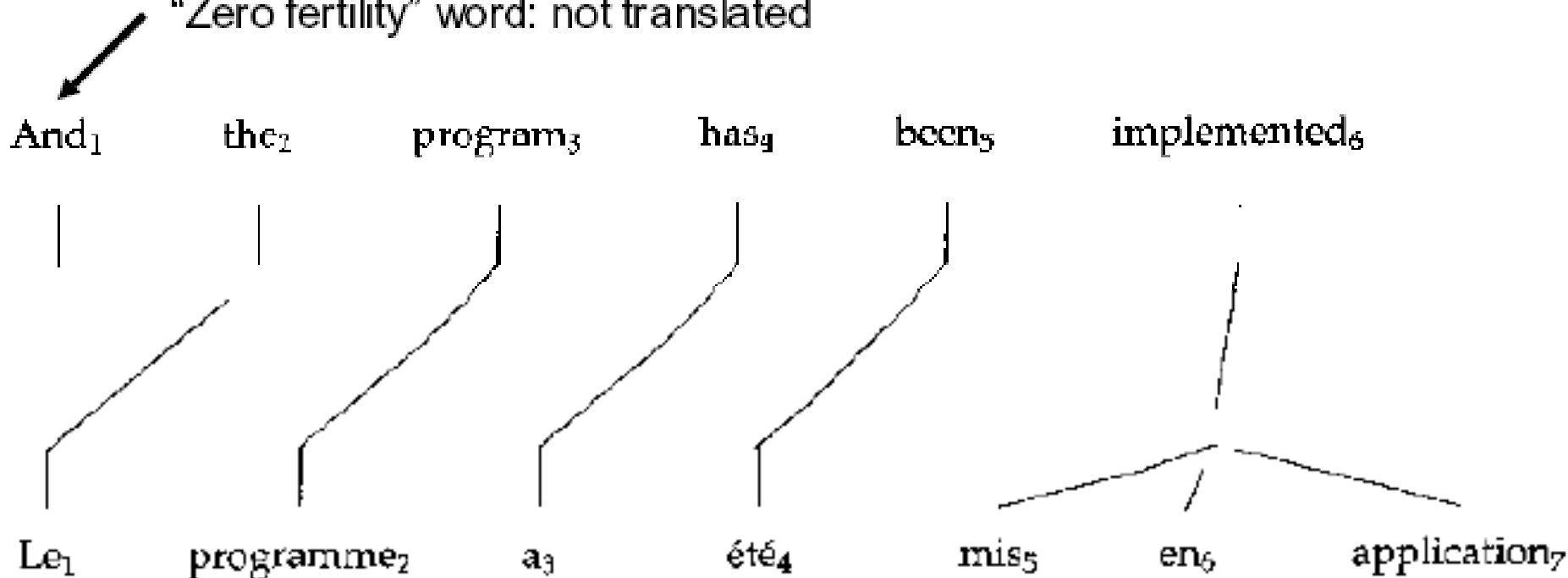
Japan shaken by two new quakes|

Le Japon secoué par deux nouveaux séismes

Extra word appears in French: “spurious” word

# Giống hàng từ - Khó hơn

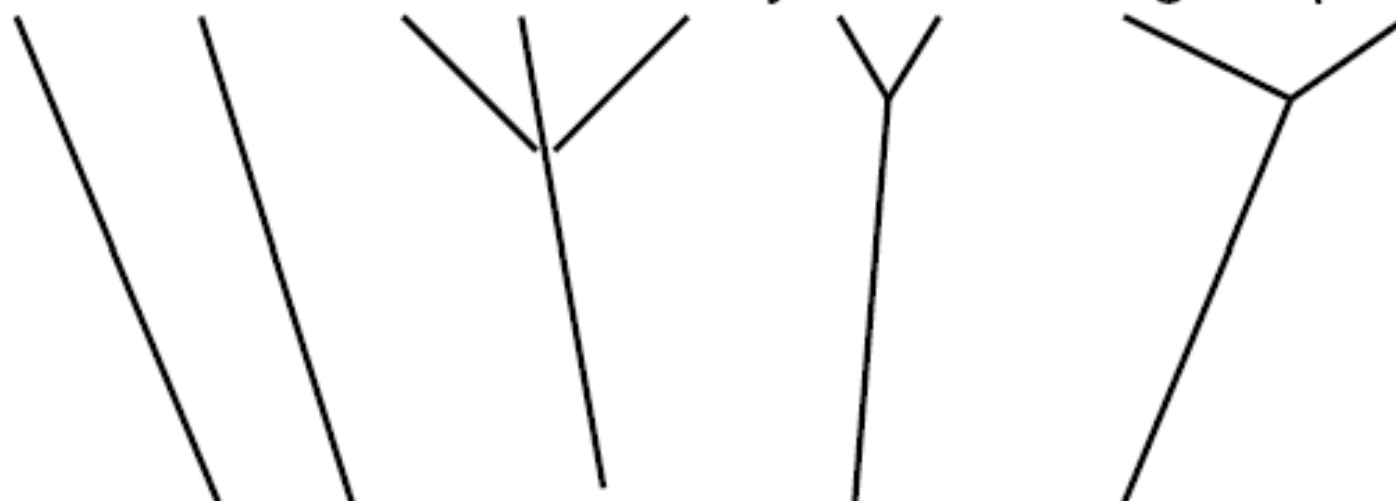
“Zero fertility” word: not translated



One word translated as several words

# Giống hàng từ - Khó hơn

The balance was the territory of the aboriginal people



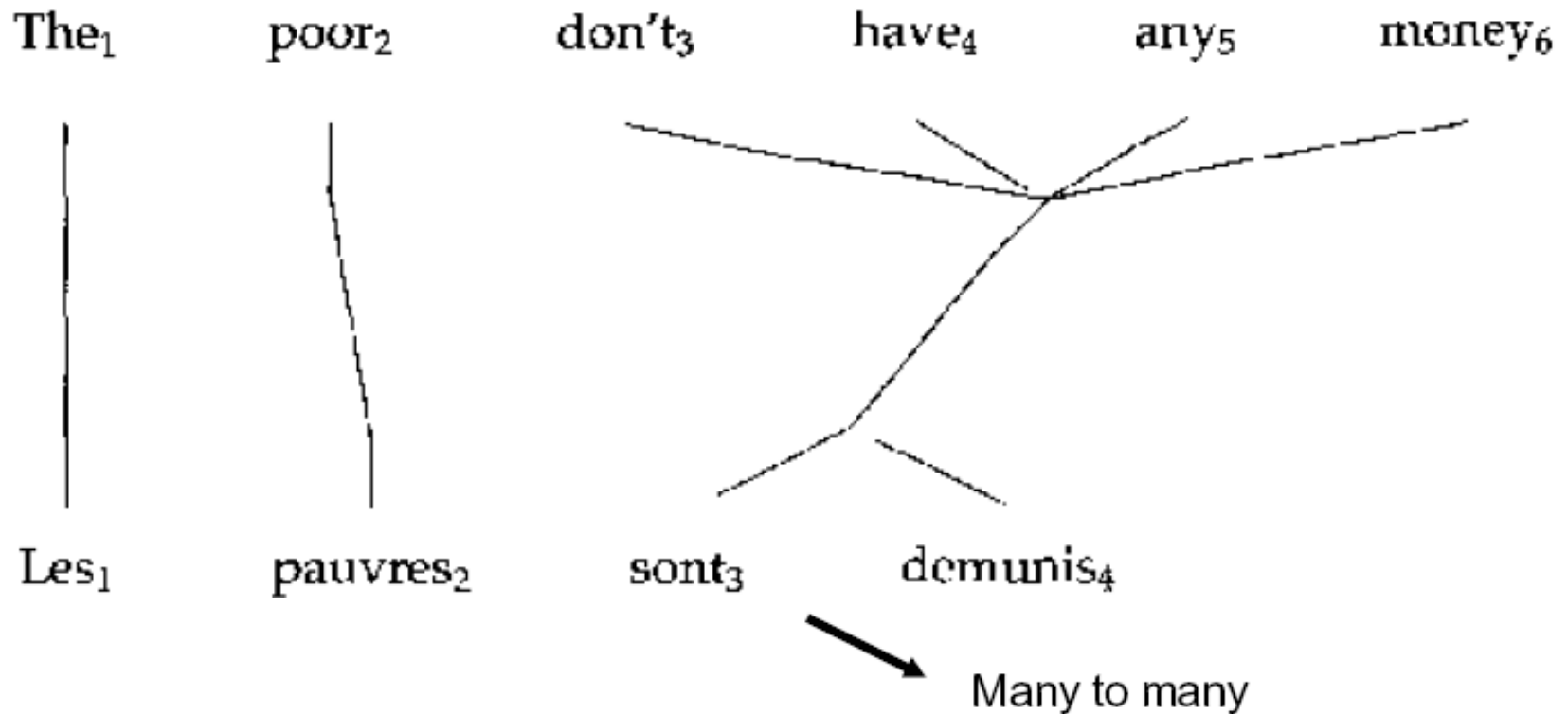
Le reste appartenait aux autochtones



Several words translated as one



# Giống hàng từ - Khó



- A line group linking a minimal subset of words is called a 'ceptr' in the IBM work

# Giống hàng từ - Mã hóa

0    1    2        3        4    5        6

- $e_0$  And the program has been implemented

- $f_0$  Le programme a été mis en application

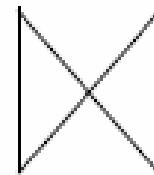
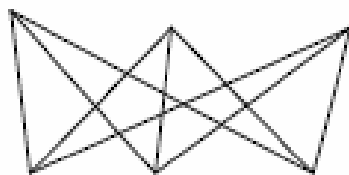
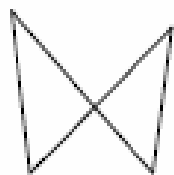
0    1        2        3    4    5    6        7

- Gán thông tin tuyến tính:

- $f_0(1)$  Le(2) programme(3) a(4) été(5) mis(6) en(6) application(6)
- $e_0$  And(0) the(1) program(2) has(3) been(4) implemented(5,6,7)

# Học việc giống hàng từ sử dụng EM

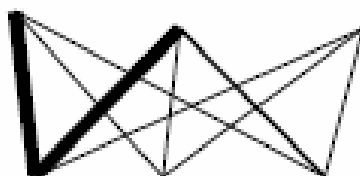
... la maison ... la maison bleue ... la fleur ...



... the house ... the blue house ... the flower ...



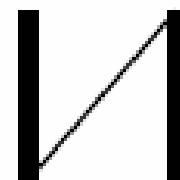
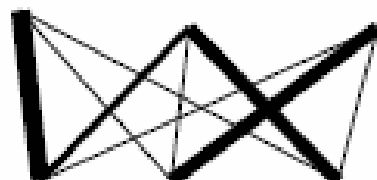
... la maison ... la maison bleue ... la fleur ...



... the house ... the blue house ... the flower ...

# Học việc giống hàng từ sử dụng EM

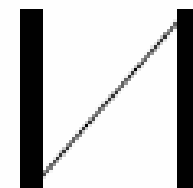
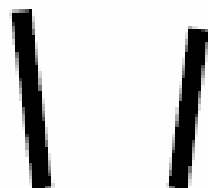
... la maison ... la maison bleue ... la fleur ...



... the house ... the blue house ... the flower ...



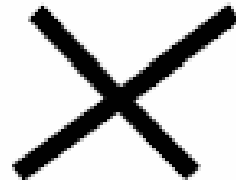
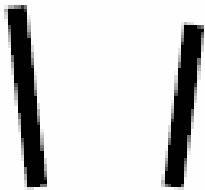
... la maison ... la maison bleue ... la fleur ...



... the house ... the blue house ... the flower ...

# Học việc giống hàng từ sử dụng EM

... la maison ... la maison bleue ... la fleur ...



... the house ... the blue house ... the flower ...

# Các thành phần của mô hình dịch

- Giả thiết
  - Việc dịch các dữ liệu độc lập với nhau
  - 1 từ tiếng Anh – n từ tiếng Pháp
  - 1 từ tiếng Pháp - (0-1) từ tiếng Anh

$$P(f | e) = \frac{1}{Z} \sum_{a_1}^l \cdots \sum_{a_m=0}^l \prod_{j=1}^m P(f_j | e_{a_j})$$

- $f_j$  - từ  $j$  trong  $f$ ;
- $a_j$  - vị trí trong  $e$  được giống hàng với  $f_j$
- $e_{a_j}$  - từ trong  $e$  được giống hàng với  $f_j$
- $Z$  là hằng số chuẩn hóa
- $a_j = 0$ : từ  $j$  trong câu tiếng Pháp được giống hàng với một từ rỗng (không dịch sang)
- $m$  - độ dài của  $f$

# Ví dụ

- $P(\text{Jean aime Marie} \mid \text{John loves Mary})$
- Giống hàng(Jean, John), (aime, loves), (Marie, Mary), ta có 3 xác suất  
 $P(\text{Jean} \mid \text{John}) \times P(\text{aime} \mid \text{loves}) \times P(\text{Marie} \mid \text{Mary})$

# Giải mã

$$\begin{aligned}\bar{e} &= \arg \max_e P(e | f) \\ &= \arg \max_e \frac{P(e)P(f | e)}{P(f)} \\ &= \arg \max_e P(e)P(f | e)\end{aligned}$$

Vấn đề: không gian tìm kiếm vô hạn

Mẹo:

- tìm kiếm dùng ngăn xếp: xây dựng dần, lưu trong stack các phần đã dịch
- sử dụng một số độ đo về độ phù hợp, vd., *chamber/house*, (nhưng có thể đi sai đường nếu 1 từ thường xuất hiện với từ khác, như *commune/house*, vì có *Chambre de Communes* (hạ nghị viện)





# Học mô hình dịch




- ☀ Ta muốn đánh giá xác suất dịch  $p(f|e)$  từ tập dữ liệu song ngữ
- ☀ ... nhưng không có thông tin về giống hàng
- ☀ Bài toán con gà quả trứng
  - ☐ nếu ta có giống hàng --> có thể đánh giá các tham số của mô hình
  - ☐ nếu ta có các tham số của mô hình --> có thể đánh giá giống hàng

# Thuật toán EM

## Dữ liệu không đầy đủ

-  Nếu có dữ liệu đầy đủ --> có thể đánh giá mô hình
-  Nếu có mô hình --> có thể lấp lỗ hổng của dữ liệu

## EM:

-  khởi tạo các tham số của mô hình
-  gán xác suất cho phần dữ liệu thiếu
-  đánh giá các tham số của mô hình từ phần dữ liệu đủ

## Lặp lại quá trình

# Thuật toán EM

## Expectation-Step: áp dụng mô hình vào dữ liệu

- thiếu thông tin về một phần của dữ liệu (giống hàng)
- sử dụng mô hình, gán xác suất với 1 giá trị nào đó

## Maximization-Step: đánh giá mô hình từ dữ liệu

- dùng các giá trị được gán như giá trị đúng
- đếm sự xuất hiện của các tham số trong mô hình (với trọng số là xác suất)
- đánh giá mô hình từ phép đếm

## Lặp đến khi hội tụ

# Thuật toán EM

## Expectation-step

- Khởi tạo giá trị  $P(w_f | w_e)$  ngẫu nhiên
- Tính số lần tìm thấy  $w_f$  trong tiếng Pháp khi có  $w_e$  trong tiếng Anh

$$z_{w_f, w_e} = \sum_{(e, f) s.t. w_e = e, w_f = f} P(w_f | w_e)$$

## Maximization-step

- Đánh giá lại xác suất dịch P từ giá trị z trên:

$$P(w_f | w_e) = \frac{z_{w_f, w_e}}{\sum_v z_{v, w_e}}$$

tổng được tính trên tất cả các từ tiếng Pháp v

# Thuật toán giống hàng từ

Khởi tạo với tập ngữ liệu giống hàng câu.

Cho  $(E, F)$  là 1 cặp câu (là 1 bead).

1. Khởi tạo ngẫu nhiên  $p(f|e)$ ,  $f \in F$ ,  $e \in E$ .
2. Đếm trên tập ngữ liệu:

$$c(f, e) = \sum_{(E, F); e \in E, f \in F} p(f|e)$$

với  $\forall$  cặp giống hàng  $(E, F)$ , kiểm tra xem  $e$  có trong  $E$  và  $f$  có trong  $F$  không. Nếu đúng, bổ sung  $p(f|e)$ .

3. Đánh giá lại:

$$p(f|e) = c(f, e) / c(e) \quad [c(e) = \sum_f c(f, e)]$$

4. Lặp đến khi  $p(f|e)$  thay đổi ít.

# Cách giống hàng tốt nhất

Với mỗi cặp (E,F), tìm

$$A = \operatorname{argmax}_A P(A|F,E) = \operatorname{argmax}_A P(F,A|E)/P(F)$$

$$= \operatorname{argmax}_A P(F,A|E)$$

$$= \operatorname{argmax}_A (\varepsilon / (l+1)^m \prod_{j=1..m} p(f_j|e_{a_j}))$$

$$= \operatorname{argmax}_A \prod_{j=1..m} p(f_j|e_{a_j})$$

- Sử dụng thuật toán lập trình động theo kiểu Viterbi.
- Tính lại  $p(f|e)$

# Đánh giá

Đánh giá dựa trên tập ngữ liệu Hansard:

- 48% câu tiếng Pháp được dịch đúng
- 2 loại lỗi:
  - Dịch sai nghĩa:
    - Permettez que je donne un exemple à chambre
    - Let me give an example in the House (incorrect decoding)
    - (Let me give the House an example)
  - Dịch sai ngữ pháp:
    - Vous avez besoin de toute l'aide disponible
    - You need all of the benefits available (ungrammatical decoding)
    - (You need all the help you can get)

# Lý do

- **Hiện tượng méo**: từ tiếng Anh ở đầu câu được giống hàng với từ tiếng Pháp ở cuối câu – hiện tượng này giảm xác suất giống hàng
- **Hiện tượng sinh (fertility)**: sự tương ứng giữa từ tiếng Anh và tiếng Pháp (1-to-1, 1-to-2, 1-to-0, ...),
  - Vd, fertility(**farmers**) trong tập ngữ liệu = 2, vì từ này khi dịch sang tiếng Anh thường gồm 2 từ : **les agriculteurs**
  - **To go** → **aller**



# Lý do

- **Các giả thiết độc lập:** các câu ngắn được ưu tiên hơn vì có ít xác suất hơn (khi nhân)  
 $\Rightarrow$  nhân kết quả với 1 hằng số tỉ lệ thuận với độ dài câu
- **Phụ thuộc dữ liệu luyện:** 1 thay đổi nhỏ trong dữ liệu luyện gây ra thay đổi lớn trong các giá trị ước lượng tham số  
Vd,  $P(le|the)$  thay đổi từ 0.610 xuống 0.497
- **Tính hiệu quả.** Bỏ các câu  $> 30$  từ, vì làm không gian tìm kiếm tăng theo cấp số mũ
- **Thiếu tri thức ngôn ngữ**

# Thiếu tri thức ngôn ngữ

- Không lưu thông tin về các ngữ: ví dụ không giống hàng được “to go” và “aller”
- Không có ràng buộc cục bộ:

Eg, *is she a mathematician*

- Âm vị. Các từ tạo bởi các âm vị khác nhau được coi là các ký hiệu riêng biệt
- Dữ liệu thưa. Các đánh giá cho các từ ít gặp không chính xác

# Open sources

- GIZA++: công cụ dịch máy thống kê để huấn luyện mô hình IBM 1-5 cho giống hàng từ
- MOSES: công cụ dịch máy thống kê
- Moses có 2 kiểu dịch: **phrase-based** và **tree-based**

# Ví dụ

- Cuộc\_sống đẹp
- Beautiful life
  
- cuộc\_sống của tôi
- my life

# Dịch máy sử dụng cú pháp

# Tại sao dùng cú pháp

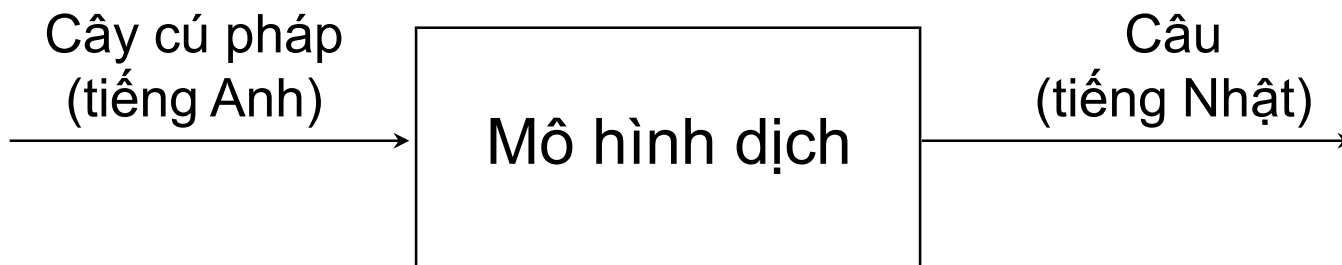
- Cần thông tin ngữ pháp
- Cần các ràng buộc khi sắp lại câu
- Khi chèn các từ chức năng vào câu, cần đặt ở vị trí chính xác
- Khi dịch từ cần sử dụng từ có cùng từ loại với nó

# Yamada and Knight (2001): Lý do cần cú pháp

- He adores listening to music.

彼は音楽を聞くのが大好きです  
Kare ha ongaku wo kiku no ga daisuki desu

# Mô hình dựa trên cú pháp

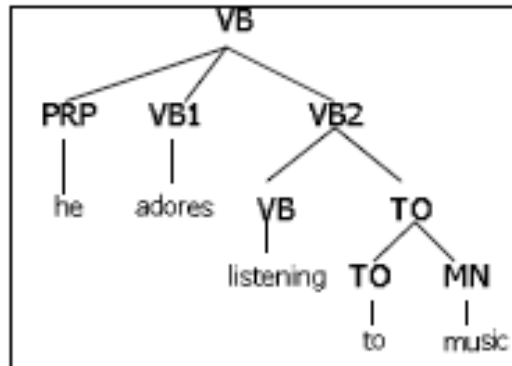


- Tiền xử lý câu tiếng Anh bằng bộ PTCP
- Thực hiện các phép tính xác suất trên cây cú pháp
  - Sắp lại trật tự các nút
  - Chèn nút mới vào
  - Dịch các từ ở lá



# Cây cú pháp (Anh) → câu (Nhật)

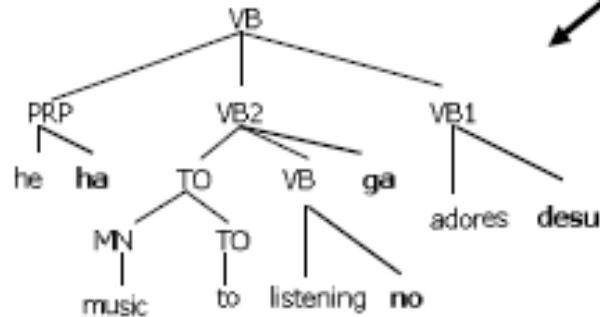
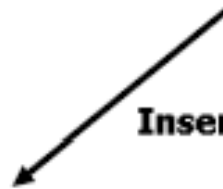
Parse Tree(E)



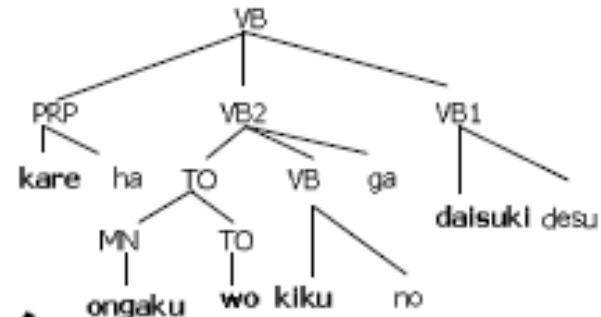
Reorder



Insert



Translate



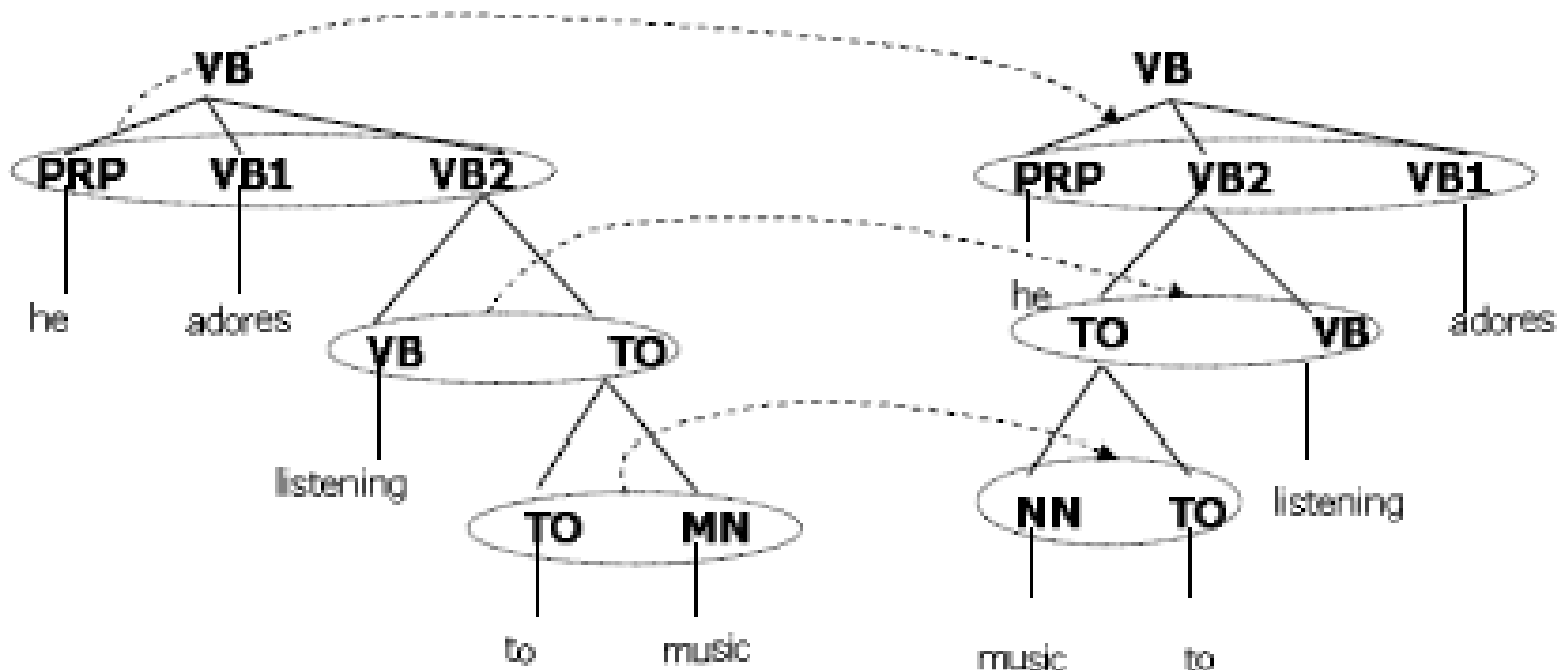
Take Leaves



Sentence(J)

*Kare ha ongaku wo kiku no ga daisuki desu*

# 1. Sắp lại trật tự



$$P(\text{PRP VB1 VB2} \mid \text{PRP VB2 VB1}) = 0.723$$

$$P(\text{VB TO} \mid \text{TO VB}) = 0.749$$

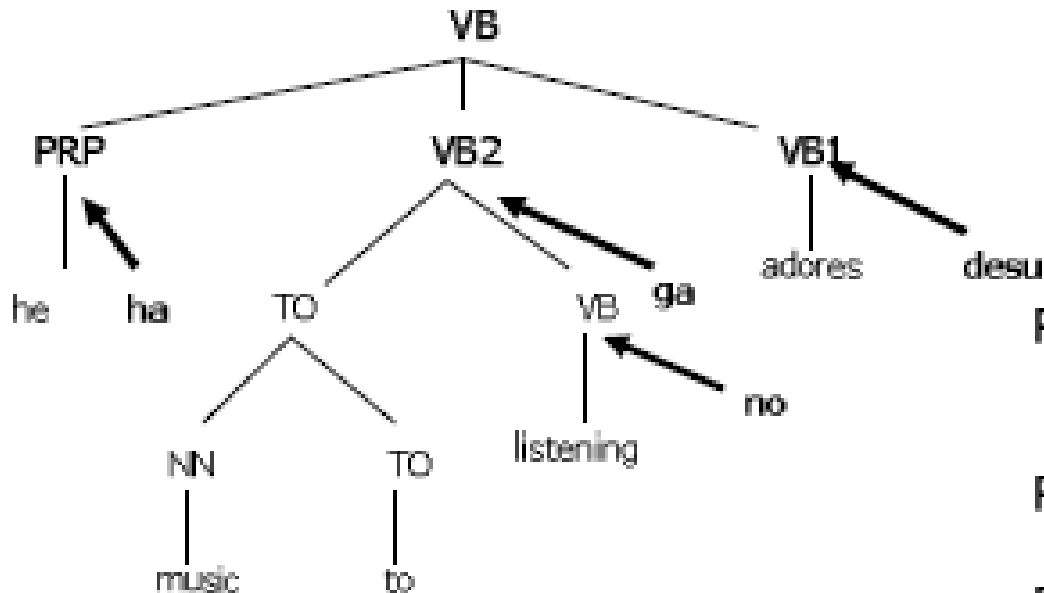
$$P(\text{TO NN} \mid \text{NN TO}) = 0.893$$

Đặc trưng điều kiện = dãy các nhãn con

# Bảng tham số: sắp lại

| Trật tự gốc | Sắp lại            | P(Sắp lại   Trật tự gốc) |
|-------------|--------------------|--------------------------|
| PRP VB1 VB2 | PRP VB1 VB2        | 0.074                    |
|             | <b>PRP VB2 VB1</b> | <b>0.723</b>             |
|             | VB1 PRP VB2        | 0.061                    |
|             | VB1 VB2 PRP        | 0.037                    |
|             | VB2 PRP VB1        | 0.083                    |
|             | VB2 VB1 PRP        | 0.021                    |
| VB TO       | VB TO              | 0.107                    |
|             | <b>TO VB</b>       | <b>0.893</b>             |
| TO NN       | TO NN              | 0.251                    |
|             | <b>NN TO</b>       | <b>0.749</b>             |
|             |                    |                          |

## 2. Chèn



$$P(\text{none}|\text{TOP-VB}) = 0.735$$

⋮

$$P(\text{right}|\text{VB-PRP}) * P(\text{ha}) = 0.652 * 0.219$$

$$P(\text{right}|\text{VB-VB}) * P(\text{ga}) = 0.252 * 0.062$$

⋮

$$P(\text{none}|\text{TO-TO}) = 0.900$$

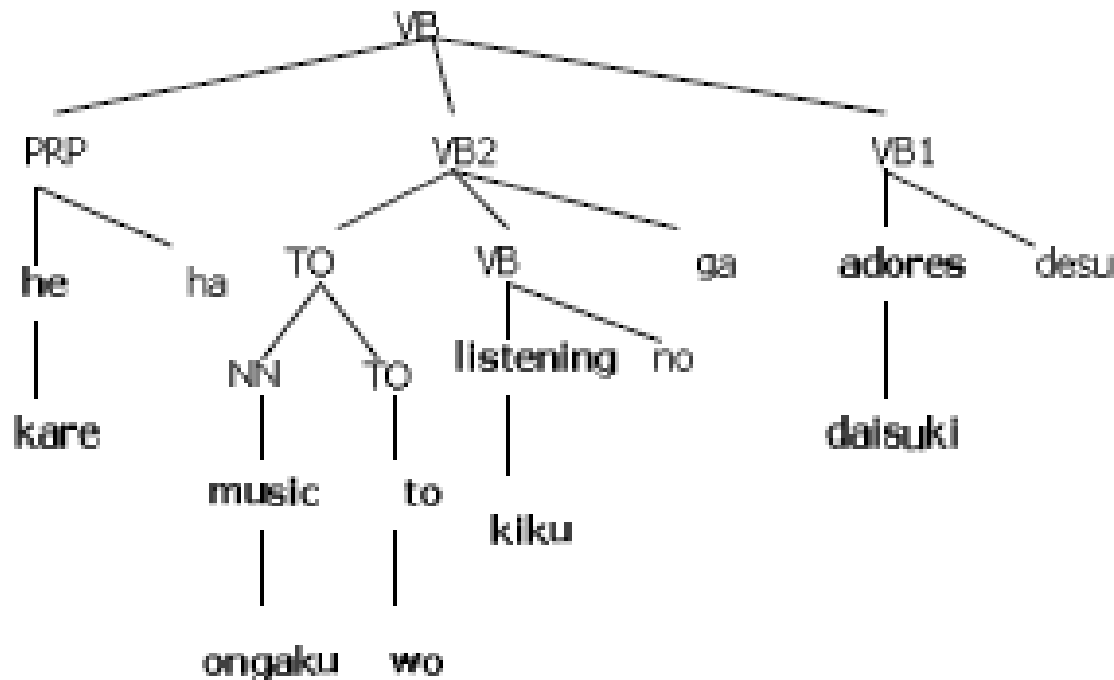
Đặc trưng điều kiện = nhãn cha & nhãn nút (vị trí) & none (là từ)

# Bảng tham số: chèn

| Parent label<br>node level | TOP<br>VB | VB<br>VB | VB<br>TO | TO<br>TO | TO<br>NN | TO<br>NN |
|----------------------------|-----------|----------|----------|----------|----------|----------|
| P (none)                   | 0.735     | 0.687    | 0.344    | 0.700    | 0.900    | 0.800    |
| P (left)                   | 0.004     | 0.061    | 0.004    | 0.030    | 0.003    | 0.096    |
| P (right)                  | 0.260     | 0.252    | 0.652    | 0.261    | 0.097    | 0.104    |

| W    | P (insert-w) |
|------|--------------|
| ha   | 0.219        |
| ta   | 0.131        |
| wo   | 0.099        |
| no   | 0.094        |
| ni   | 0.090        |
| te   | 0.079        |
| ga   | 0.062        |
|      |              |
| desu | 0.0007       |
|      |              |

# 3. Dịch



$P(\text{he} \text{---} \text{kare}) = 0.952$   
 $P(\text{music} \text{---} \text{ongaku}) = 0.900$   
 $P(\text{to} \text{---} \text{wo}) = 0.038$   
 $P(\text{listening} \text{---} \text{kiku}) = 0.333$   
 $P(\text{adore} \text{---} \text{daisuki}) = 1.000$

Đặc trưng điều kiện = từ (tiếng Anh)

# Bảng tham số: Dịch

| E | adores        | he  | listening                           | music                      | to   |
|---|---------------|---|-------------------------------------|----------------------------|--|
| J | daisuki 1.000 | kare 0.952<br>NULL 0.016<br>nani 0.005<br>da 0.003<br>shi 0.003<br> | kiku 0.333<br>kii 0.333<br>mi 0.333 | ongaku 0.900<br>naru 0.100 | ni 0.216<br>NULL 0.204<br>to 0.133<br>no 0.046<br>wo 0.038<br> |

Ghi chú: Dịch thành NULL → xóa

# Thử nghiệm

- Dữ liệu luyện: 2000 cặp câu J-E
- J: tách từ sử dụng Chasen
- E: PTCP sử dụng bộ PTCP Collins
  - Luyện trên 40000 câu từ Treebank, độ cx ~90%
- E: từ cây cú pháp, xác định trật tự từ và chuyển đổi (SVO  $\leftrightarrow$  SOV)
- Luyện sử dụng EM: 20 vòng lặp



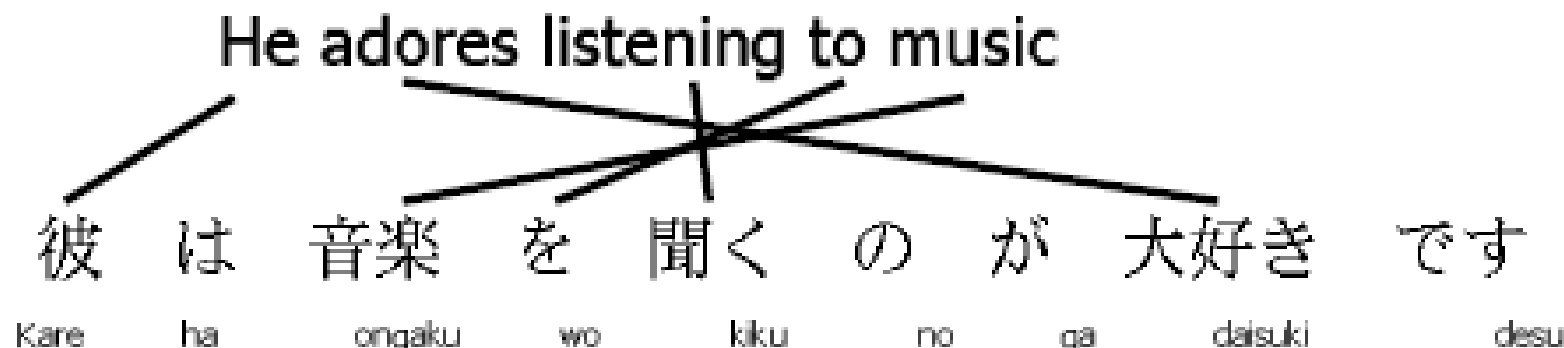
# Kết quả

|             | Điểm trung bình | #câu |
|-------------|-----------------|------|
| Y/K model   | 0.582           | 10   |
| IBM model 5 | 0.431           | 0    |

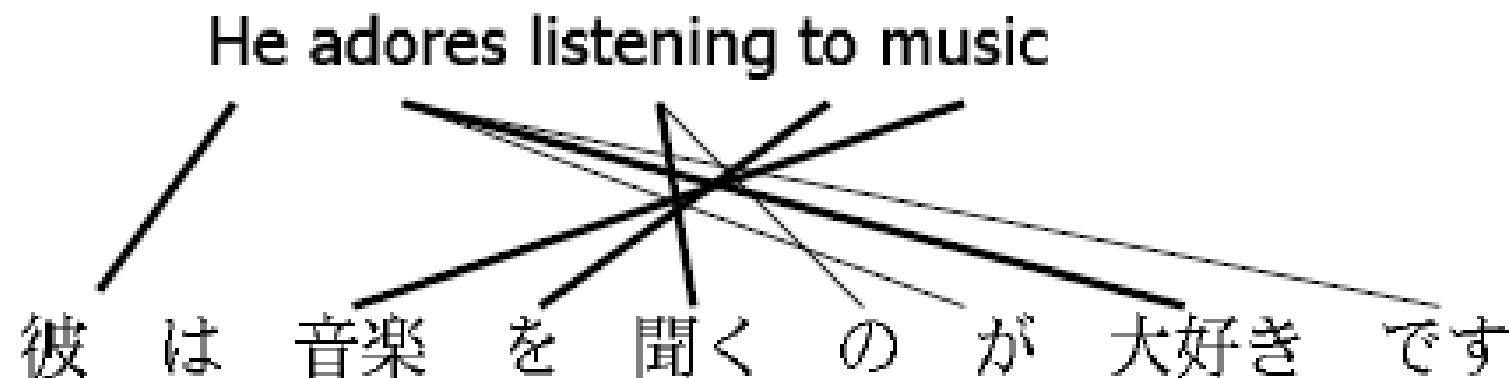
- Điểm trung bình được tính trên 3 người với 50 câu
- ok(1.0), không chắc (0.5), sai (0.0)
- chỉ tính độ chính xác

# Kết quả: giống hàng 1

Syntax-based Model



IBM Model 3

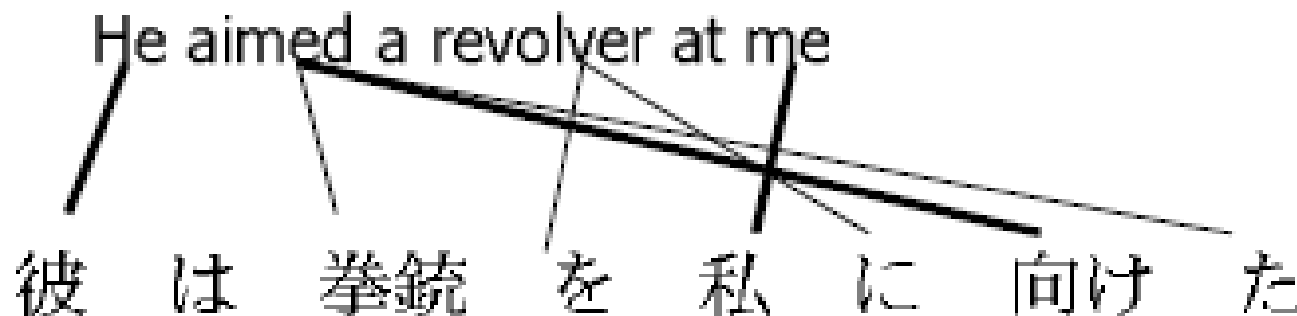


# Kết quả: giống hàng 2

Syntax-based model



IBM Model 3



# Một số mã nguồn mở

- Xem <http://fosmt.org/>
  - Moses
  - Giza++

# Một số hệ thống dịch máy trên Internet

- [http://www.google.com/language\\_tools?hl=en](http://www.google.com/language_tools?hl=en)
- <http://www.systransoft.com/index.html>
- <http://babelfish.altavista.digital.com/>



# Google d ch

Th r trình duyệt mới c  t nh năng dịch tự động.  
[T i xuống Google Chrome](#)

T r: Tiếng Anh Sang: Tiếng Việt Dịch

Dịch mọi trang web

|                        |                      |                      |                 |                   |                |
|------------------------|----------------------|----------------------|-----------------|-------------------|----------------|
| Ph t hiện ngôn ngữ     | Tiếng Basque         | Tiếng Galicia        | Tiếng M  Lai    | Tiếng S c         | Tiếng X c-bi   |
| Tiếng  -r p            | Tiếng B lar t        | Tiếng George         | Tiếng Macedonia | Tiếng Slovak      | Tiếng X r Wale |
| Tiếng Agiecbaigi ng    | Tiếng B  Đào Nha     | Tiếng Hà Lan         | Tiếng Malta     | Tiếng Slovenia    | Tiếng          |
| Tiếng Ai-len           | Tiếng Bungary        | Tiếng H n            | Tiếng Na Uy     | Tiếng Tây Ban Nha | Tiếng Yiddish  |
| Tiếng Aix len          | Tiếng Catalan        | Tiếng Hin- i ( n   ) | Tiếng Nam Phi   | Tiếng Th i        |                |
| Tiếng An-ba-ni         | Tiếng Creole   Haiti | Tiếng Hungari        | Tiếng Nga       | Tiếng Th  Nh  Kỳ  |                |
| <b>Tiếng Anh</b>       | Tiếng Croatia        | Tiếng Hy Lạp         | Tiếng Nhật      | Tiếng Thụy  iển   |                |
| Tiếng Armenia          | Tiếng Đan Mạch       | Tiếng Indonesia      | Tiếng Pháp      | Tiếng Trung Qu c  |                |
| Tiếng Ba Lan           | Tiếng Do Th i        | Tiếng Latinh         | Tiếng Phần Lan  | Tiếng Ukraina     |                |
| Tiếng Ba Tư            | Tiếng   c            | Tiếng Latvia         | Tiếng Philippin | Tiếng Urdu        |                |
| Tiếng Bantu (  ng Phi) | Tiếng Ext nia        | Tiếng Litواني        | Tiếng Rumani    | Tiếng Việt        |                |



Inside the USA » Blog Archive » April Fools - Microsoft Internet ...

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

Address <http://insidetheusa.net/2008/04/02/april-fools/> Go Links

Pennsylvanie Etats Unis  
Visitez-vous Pennsylvanie? Comparez prix & critiques d'hôtels

Krankenversicherung USA  
Unkomplizierter, hoher Kostenschutz vom US-Spezialisten! Div. Lösungen.

Annonces Google

## April Fools

par Jerome ITU ~ 02/04/2008, 09:22 . Classé dans : Humour, Politique US :

La journée de mardi a été riche en poissons de tout genre.

Dans la catégorie écolo, on nous a présenté le tout nouveau [Air Force One](#), un modèle hybride, "15 à 20% plus économique".

Dans la catégorie politique, Hillary fend l'armure et propose, au vu des récentes [performances](#) du sénateur, un défi [au bowling](#) à Obama pour décider du nominé démocrate. Ainsi "les américains sauront que si le téléphone sonne à 3 heures du matin, ils auront un président prêt à jouer au bowling dès le premier jour".

Dans la catégorie sport, c'est Chabal qui a fait les frais de l'humour du jour. Les sites spécialisés ont relayé son [départ dans la NFL](#) américaine, aux New England Patriots, pour un contrat de 15 millions de dollars pour 3 ans. On attend toujours la confirmation de l'homme qui soulève les foules en Nouvelle-Zélande.

Et, enfin, dans la catégorie blog, Superfrenchie a révélé un pan de sa [généalogie](#). Il serait apparenté à un certain... Bill O'Reilly. "My cousin Billy". Là, c'est gros quand même !

Merci pour cette imagination débordante, en tout cas.

De la part d'un internaute bloqué devant son écran toute la journée, la faute à de maudits troubles digestifs...

Internet

Translated version of <http://insidetheusa.net/2008/04/02/april...>

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

Address <http://translate.google.com/translate?u=http%3A%2F%2Finsidetheusa.net%2F2008%2F04%2F02%2Fapril-fools%2F> Go Links

Google™ This page was [automatically translated](#) from French. [View original web page](#) or mouse over text to view original language.

Ads by Google

## April Fools

by Jerome ITU ~ 02/04/2008, 09:22. Filed under: Funny, U.S. policy.

The day Tuesday was rich in fish of all kinds.

In the green category, we introduced the brand new [Air Force One](#), a hybrid "15 to 20% more economical."

In the political category, Hillary fend armor and offers, given the recent [performances](#) of the senator, a challenge [bowling](#) to Obama to decide the Democratic nominee. Thus "the Americans know that if the phone rings at 3 o'clock in the morning, they will have a president ready to play bowling from the first day."

In the sport category is Chabal who has borne the brunt of humour of the day. The specialized sites have relayed his [departure in the NFL](#) American, the New England Patriots, for a contract of 15 million dollars for 3 years. It is still awaiting confirmation from the man who raised the crowds in New Zealand.

And, finally, in the category blog, Superfrenchie revealed a pan of its [genealogy](#). It would be akin to a certain... Bill O'Reilly. "My cousin Billy." There is still big news.

Thank you for your imagination, anyway.

On the part of a visitor blocked in front of his screen all day, the fault of cursed digestive disorders...

Some anecdotes crispy, you who you are delivered to your workplace on Tuesday...

Internet





## Small Business

- ▶ [SYSTRAN 7 Business Translator](#)
- ▶ [SYSTRAN 7 Premium Translator](#)
- ▶ [SYSTRAN Enterprise Server 7](#)
- ▶ [SYSTRANLinks](#)

## Enterprise Solutions

- ▶ [SYSTRAN 7 Premium Translator](#)
- ▶ [SYSTRAN 7 Business Translator](#)
- ▶ [SYSTRAN Enterprise Server 7](#)
- ▶ [SYSTRANLinks](#)
- ▶ [Professional Services](#)

[What is automated translation?](#)[Who we are](#)

Translate with **SYSTRAN**

**Text** **Web Page**

English French

URL:

**Translate**

**SYSTRAN**

DERNIÈRES  
NOUVELLES**« Centaines tuées » en Côte d'Ivoire**

Au moins 800 personnes ont été tuées dans la ville occidentale de la Côte d'Ivoire de Duekoue cette semaine, le Comité International de la Croix Rouge dit.

**» Plus des nouvelles de BBC****▼ Nouvelles**

Éditez



Le chef de l'ONU  
condamne des  
massacres afghans

il y a 24 minutes

**▼ Sport**

Éditez



Ensemble de  
Muralitharan pour défier  
la blessure

il y a environ 6 heures

**▼ Projecteur**

AFFAIRES DE SPORT



Le golf est important  
de défis sont-ils cette  
dollar ?

- Pièces en t d'affa
- Le Ryder Cup do  
de £82.4m »
- Plus des affaires