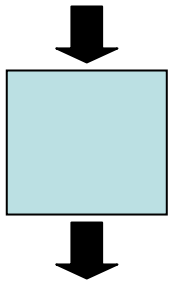
Statistical Machine Translation

Kevin Knight

CS562
August 22, 2006

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



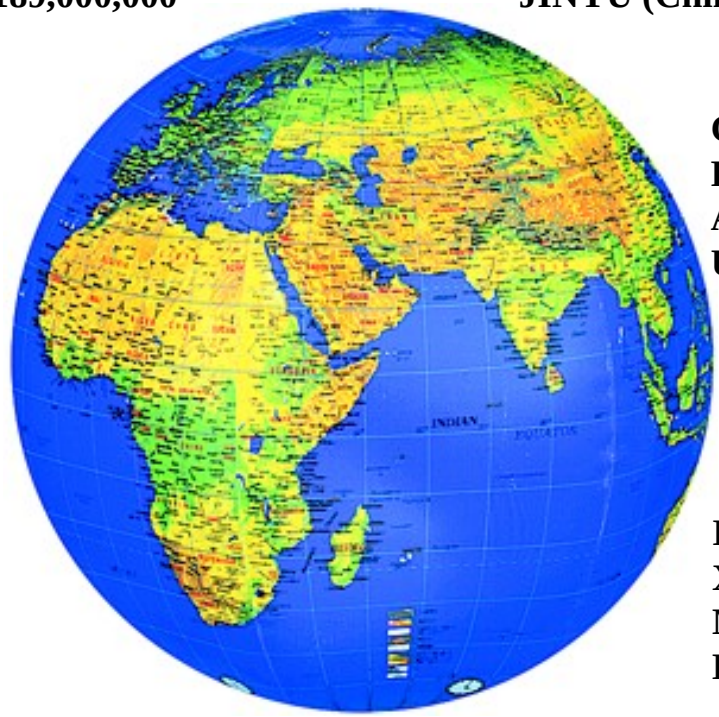
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Why People Get Into This Field

- Passion about understanding how human language works
  - What makes one sequence of words grammatical, and another not?
- Interest in foreign languages
  - What’s the difference between English and Chinese?
- Desire to change the world
  - How will the world be different when the language barrier disappears?

Thousands of Languages Are Spoken

MANDARIN	885,000,000
SPANISH	332,000,000
ENGLISH	322,000,000
BENGALI	189,000,000
HINDI	182,000,000
PORTUGUESE	170,000,000
RUSSIAN	170,000,000
JAPANESE	125,000,000
GERMAN	98,000,000
WU (China)	77,175,000
JAVANESE	75,500,800
KOREAN	75,000,000
FRENCH	72,000,000
VIETNAMESE	67,662,000
TELUGU	66,350,000
YUE (China)	66,000,000
MARATHI	64,783,000
TAMIL	63,075,000



TURKISH	59,000,000
URDU	58,000,000
MIN NAN (China)	49,000,000
JINYU (China)	45,000,000
GUJARATI	44,000,000
POLISH	44,000,000
ARABIC	42,500,000
UKRAINIAN	41,000,000
ITALIAN	37,000,000
XIANG (China)	36,015,000
MALAYALAM	34,022,000
HAKKA (China)	34,000,000
KANNADA	33,663,000
ORIYA	31,000,000
PANJABI	30,000,000
SUNDA	27,000,000

Source: Ethnologue

Why It's Hard

- Each word has tons of meanings
  - I'll **get** a cup of coffee → ?
  - I didn't **get** that joke → ?
  - I **get** up at 8am → ?
  - I **get** nervous → ?
  - Yeah, I **get** around … → ?
- Each word has zillions of contexts
- Word order is different

Why It's Hard

- Output must be a grammatical, sensible, never-before-uttered sentence!
- Computers **consume** lots of human language
  - Google, Yahoo, Altavista …
  - Speech recognizers …
- More challenging to also **produce** human language
  - What makes one sequence of words grammatical, and another not?

Rapid Progress in the Field of Statistical Machine Translation

1999 → **2002** → **2003** → **2004**

insistent Wednesday may recurred her trips to Libya tomorrow for flying

Cairo 6-4 (AFP) - an official announced today in the Egyptian lines company for flying Tuesday is a company " insistent for flying " may resumed a consideration of a day Wednesday tomorrow her trips to Libya of Security Council decision trace international the imposed ban comment .

Egyptair Has Tomorrow to Resume Its Flights to Libya

Cairo 4-6 (AFP) - said an official at the Egyptian Aviation Company today that the company egyptair may resume as of tomorrow, Wednesday its flights to Libya after the International Security Council resolution to the suspension of the embargo imposed on Libya.

Statistical MT system results in annual US government NIST evaluations.

2005

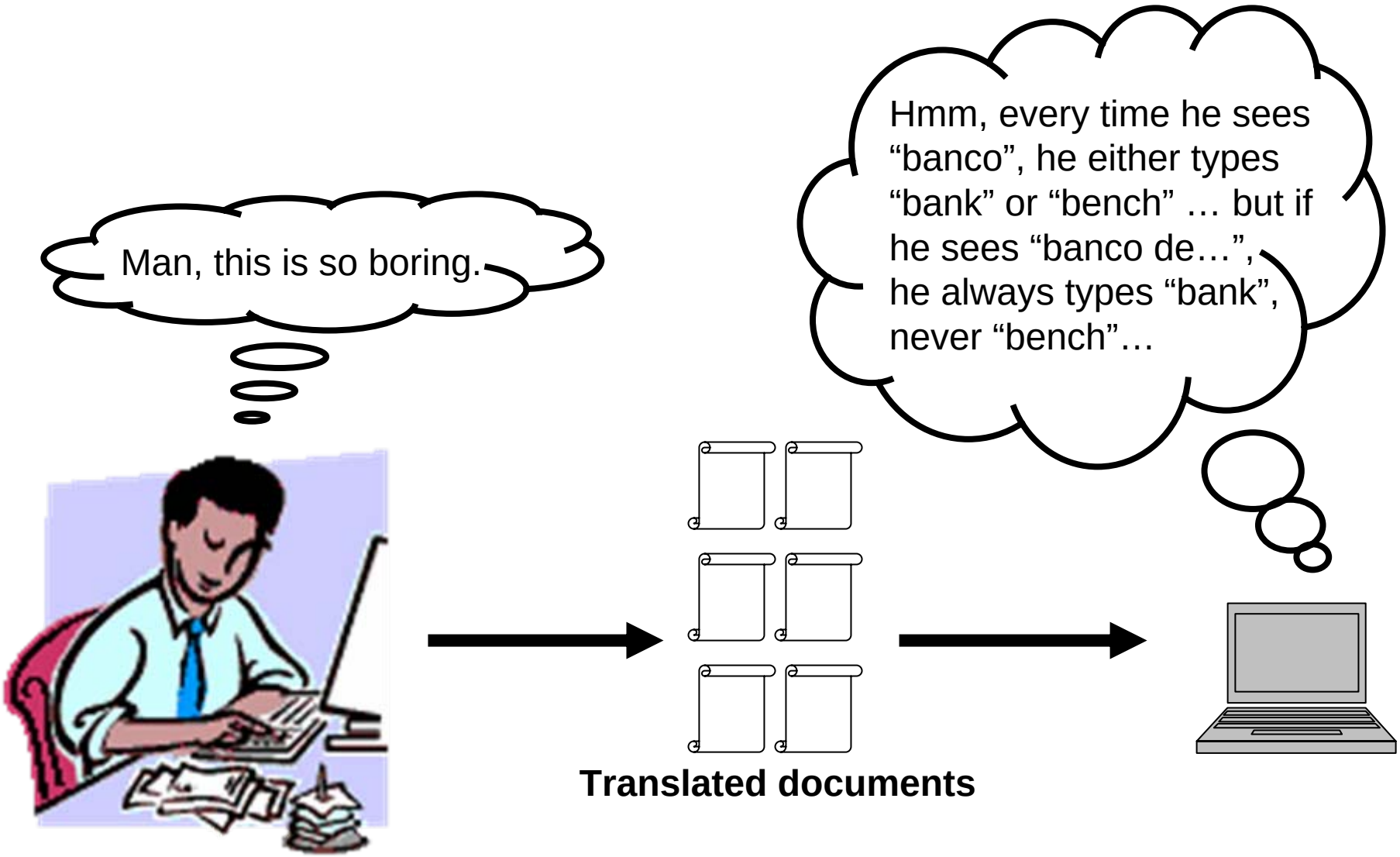
news broadcast

foreign language speech recognition

English translation

searchable archive

Data-Driven Machine Translation



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat **jjat** bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat **jjat** quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . / 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . / 5b. totat jjat quat cat .	11a. lalok nok **crrrok** hihok yorok zanzanok . ??? 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok **yorok** ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** **yorok** zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . ??? 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

process of elimination

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . cognate?
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . zero fertility
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

It’s Really Spanish/English

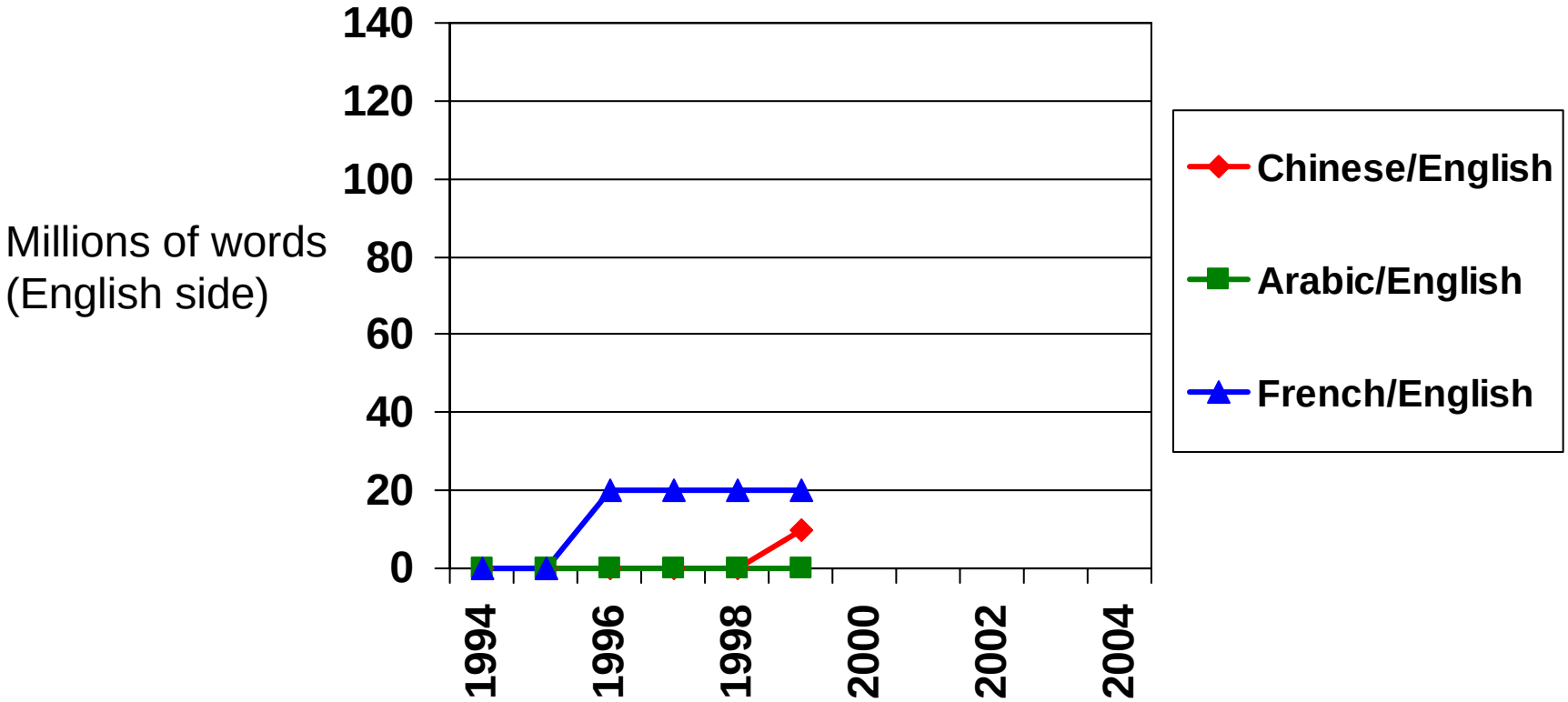
Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Data for Statistical Machine Translation

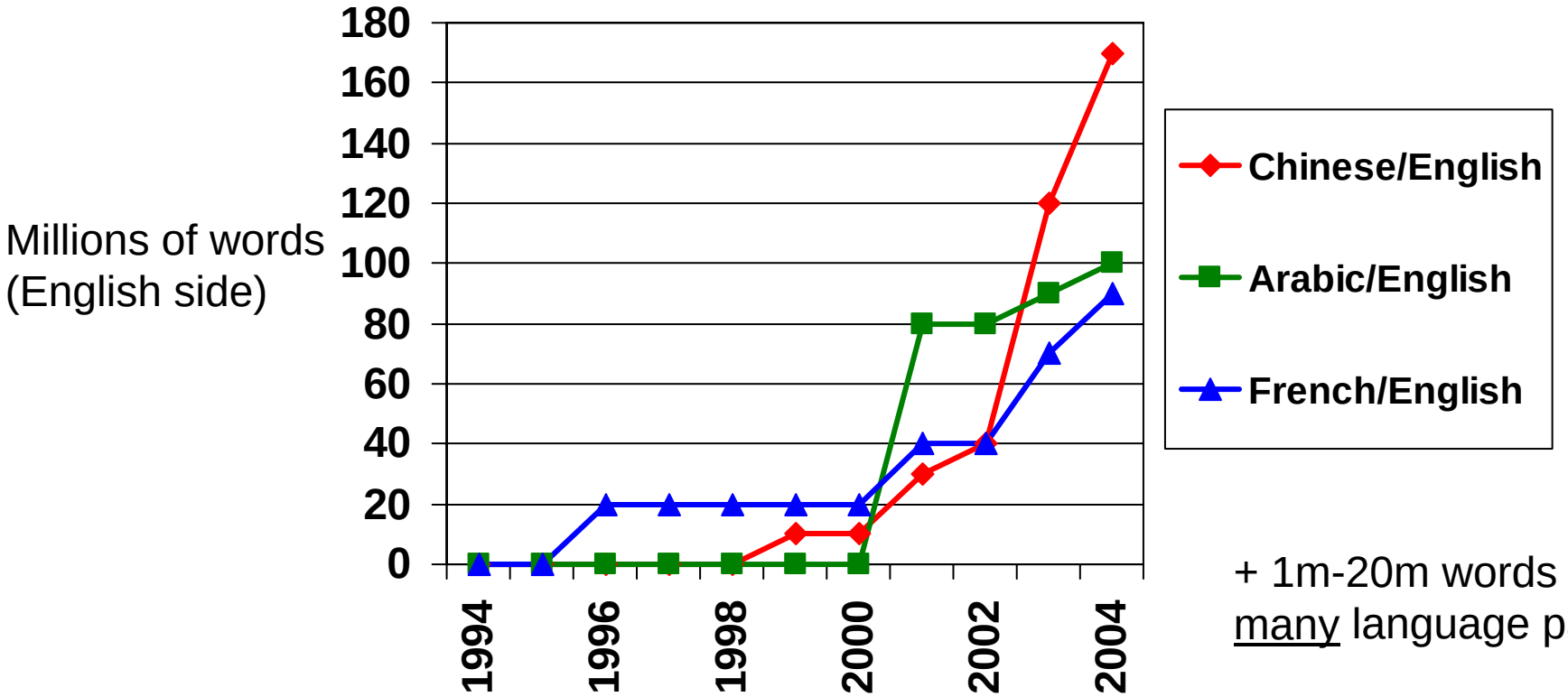
and data preparation

Ready-to-Use Online Bilingual Data



(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

Ready-to-Use Online Bilingual Data



+ 1m-20m words for <u>many</u> language pairs

(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

Sentence Alignment

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.

Sentence Alignment

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.

Sentence Alignment

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.

Sentence Alignment

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

Note that unaligned sentences are thrown out, and sentences are merged in n-to-m alignments (n, m > 0).

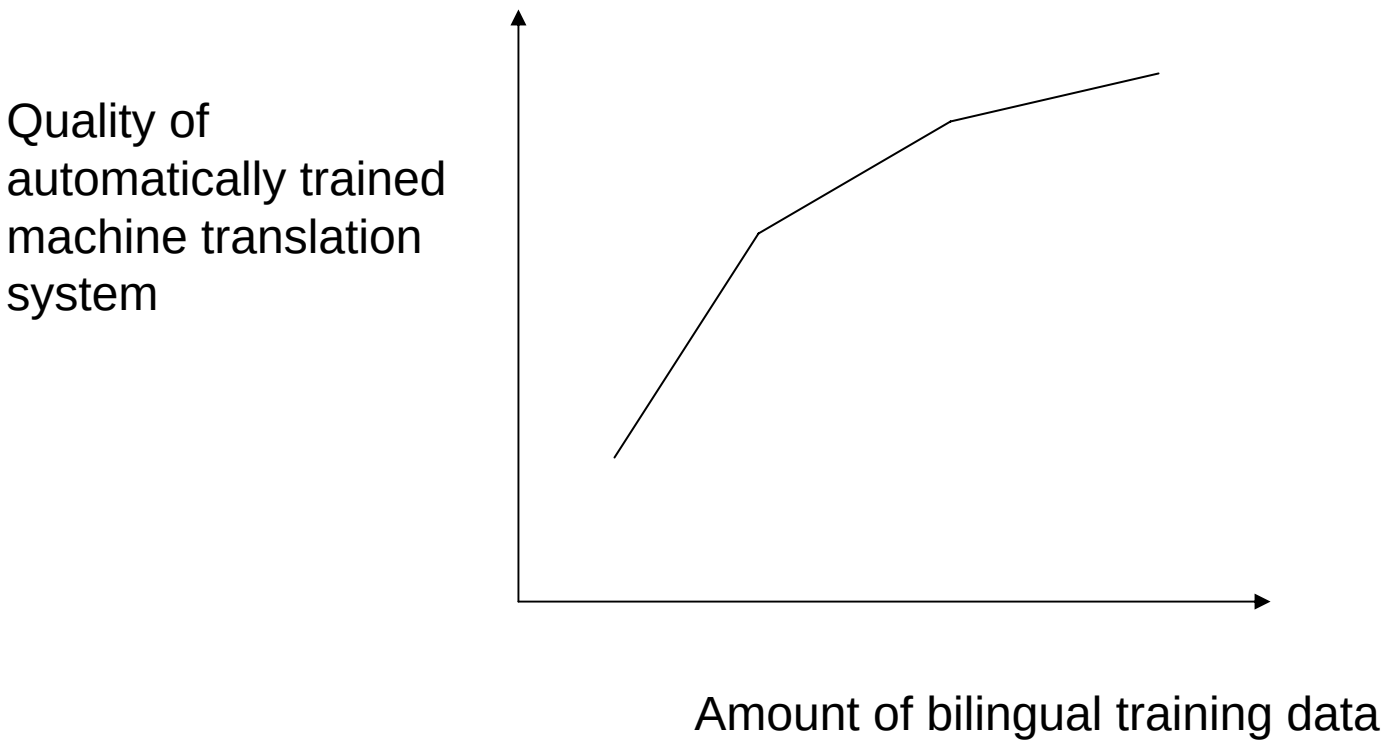
Tokenization (or Segmentation)

- English
  - Input (some byte stream):

    `"There," said Bob.`
  - Output (7 “tokens” or “words”):

    `" There , " said Bob .`
- Chinese
  - Input (byte stream): 美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件。
  - Output: 美国 关岛国 际机 场 及其 办公室均接获 一名 自称 沙地 阿拉 伯富 商拉登 等发 出 的 电子邮件。

It Is Possible to Draw Learning Curves: How Much Data Do We Need?



MT Evaluation

MT Evaluation

- No single right answer to compare against!
- Manual Evaluation:
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - Error categorization
- Testing in an application that uses MT as one component
  - Question answering from foreign language documents
- Automatic:
  - WER (word error rate)
  - **BLEU (Bilingual Evaluation Understudy)**

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the the")
- Brevity penalty
  - Can't just type out single word "the" (precision 1.0!)

*** Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- BLEU4 formula
  (counts n-grams up to length 4)

exp (1.0 * log p1 +
0.5 * log p2 +
0.25 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1,
0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Multiple Reference Translations

Reference translation 1:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

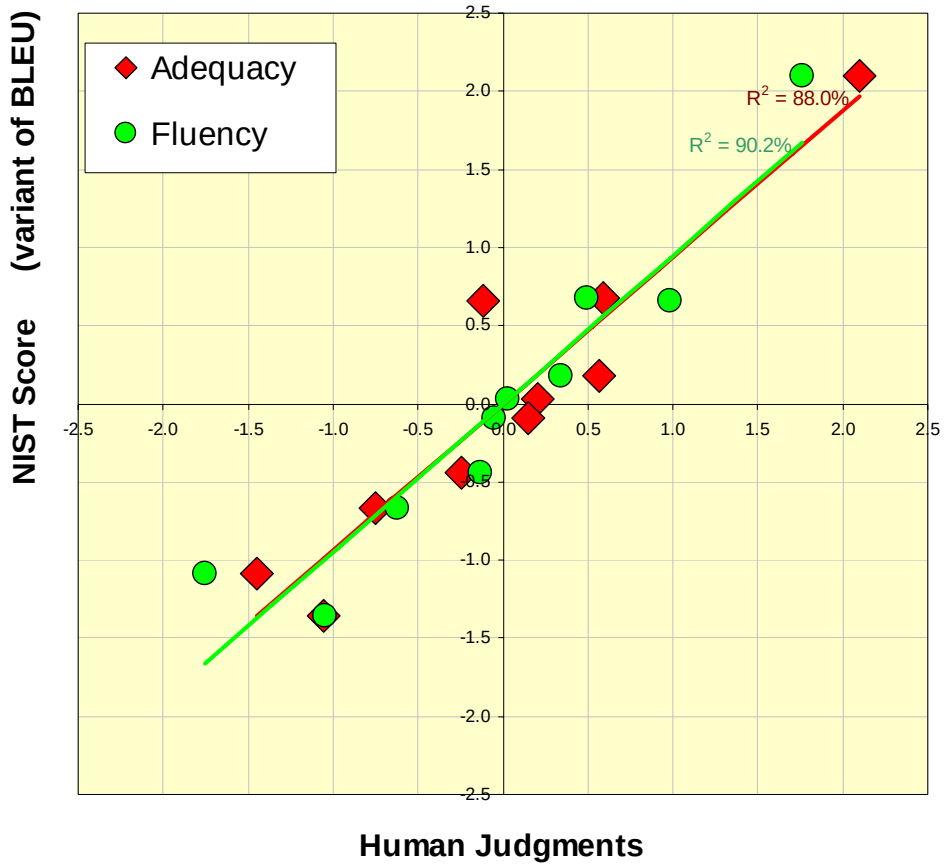
Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

BLEU Tends to Predict Human Judgments



slide from G. Doddington (NIST)

BLEU in Action

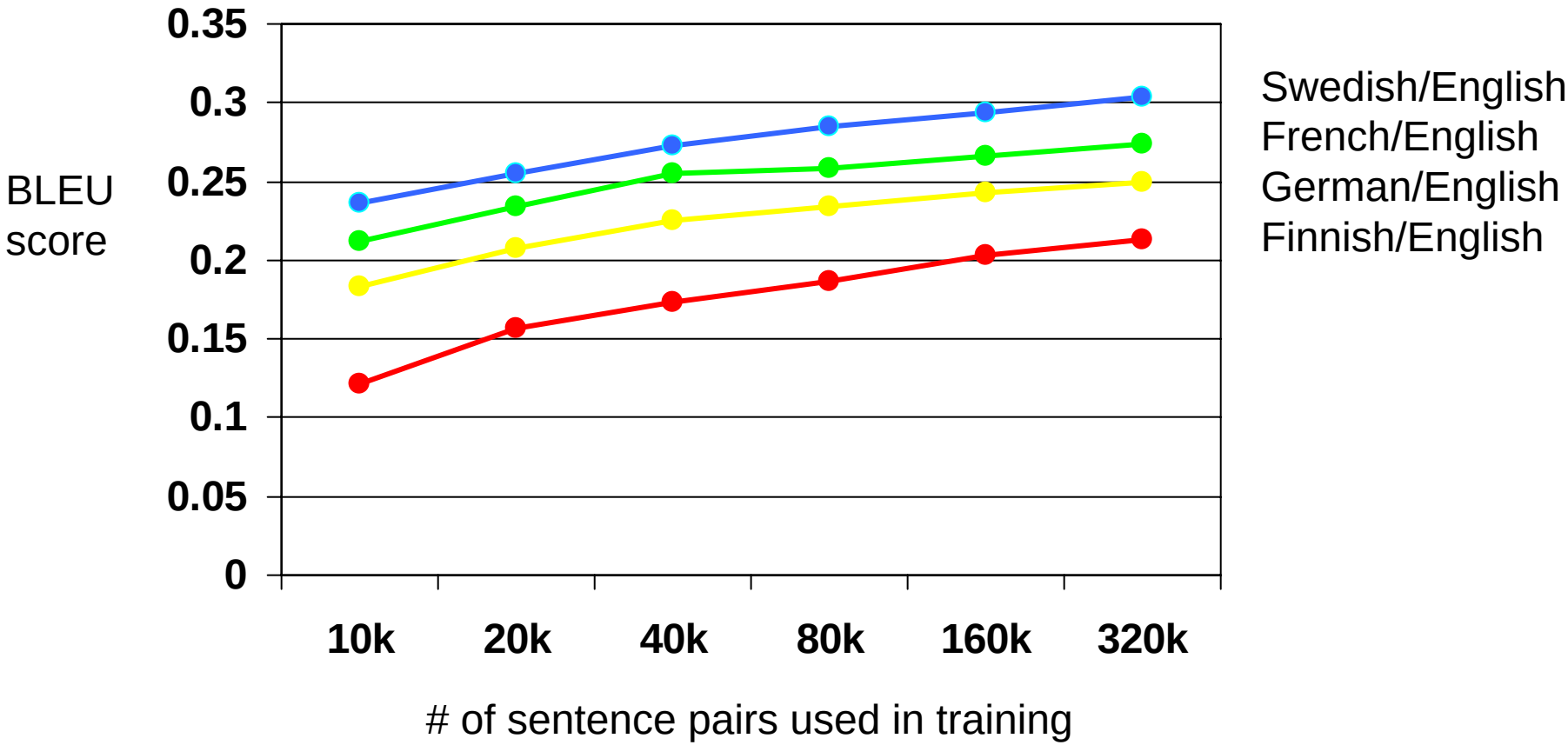
枪手被警方击毙。	(Foreign Original)
the gunman was shot to death by the police .	(Reference Translation)
the gunman was police kill .	#1
wounded police jaya of	#2
the gunman was shot dead by the police .	#3
the gunman arrested by police kill .	#4
the gunmen were killed .	#5
the gunman was shot to death by the police .	#6
gunmen were killed by police ?SUB>0 ?SUB>0	#7
al by the police .	#8
the ringer is killed by the police .	#9
police killed the gunman .	#10

BLEU in Action

枪手被警方击毙。	(Foreign Original)
the gunman was shot to death by the police .	(Reference Translation)
the gunman was police kill .	#1
wounded police jaya of	#2
the gunman was shot dead by the police .	#3
the gunman arrested by police kill .	#4
the gunmen were killed .	#5
the gunman was shot to death by the police .	#6
gunmen were killed by police ?SUB>0 ?SUB>0	#7
al by the police .	#8
the ringer is killed by the police .	#9
police killed the gunman .	#10

green	= 4-gram match	(good!)
red	= word not matched	(bad!)

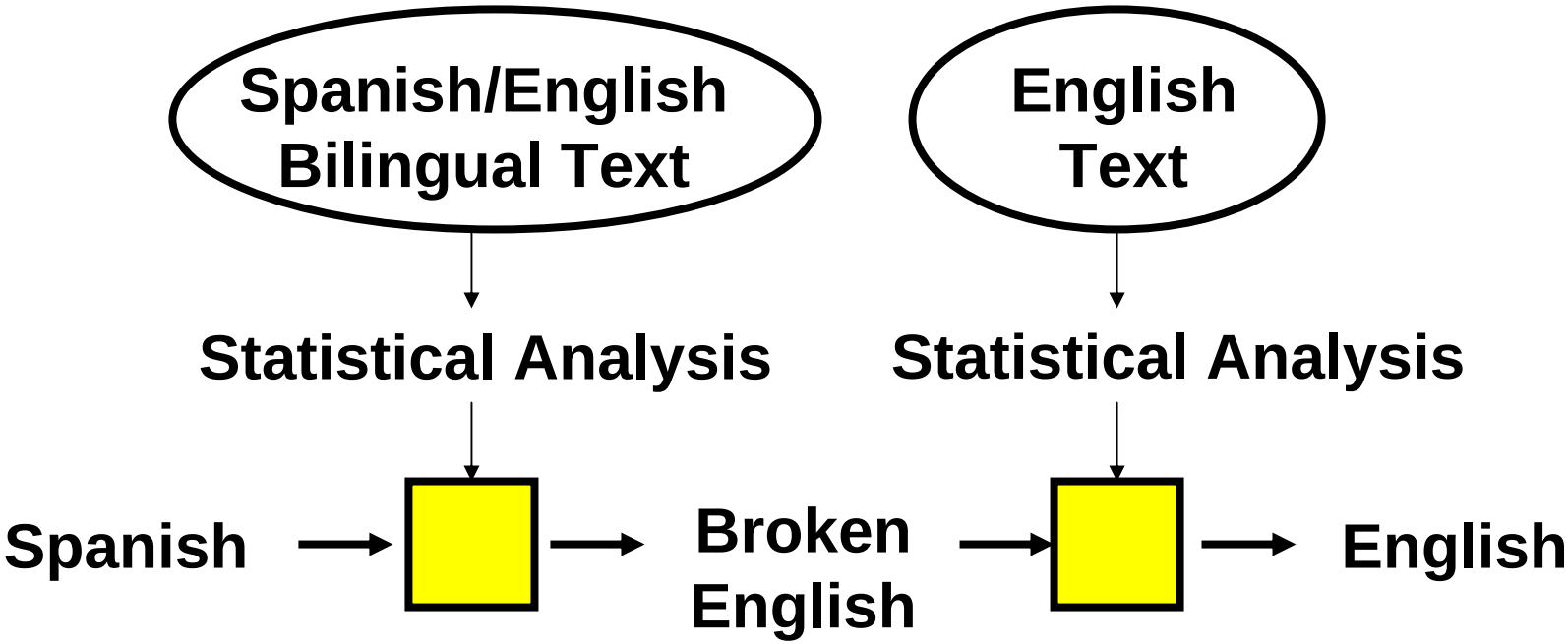
Sample Learning Curves



Experiments by
Philipp Koehn

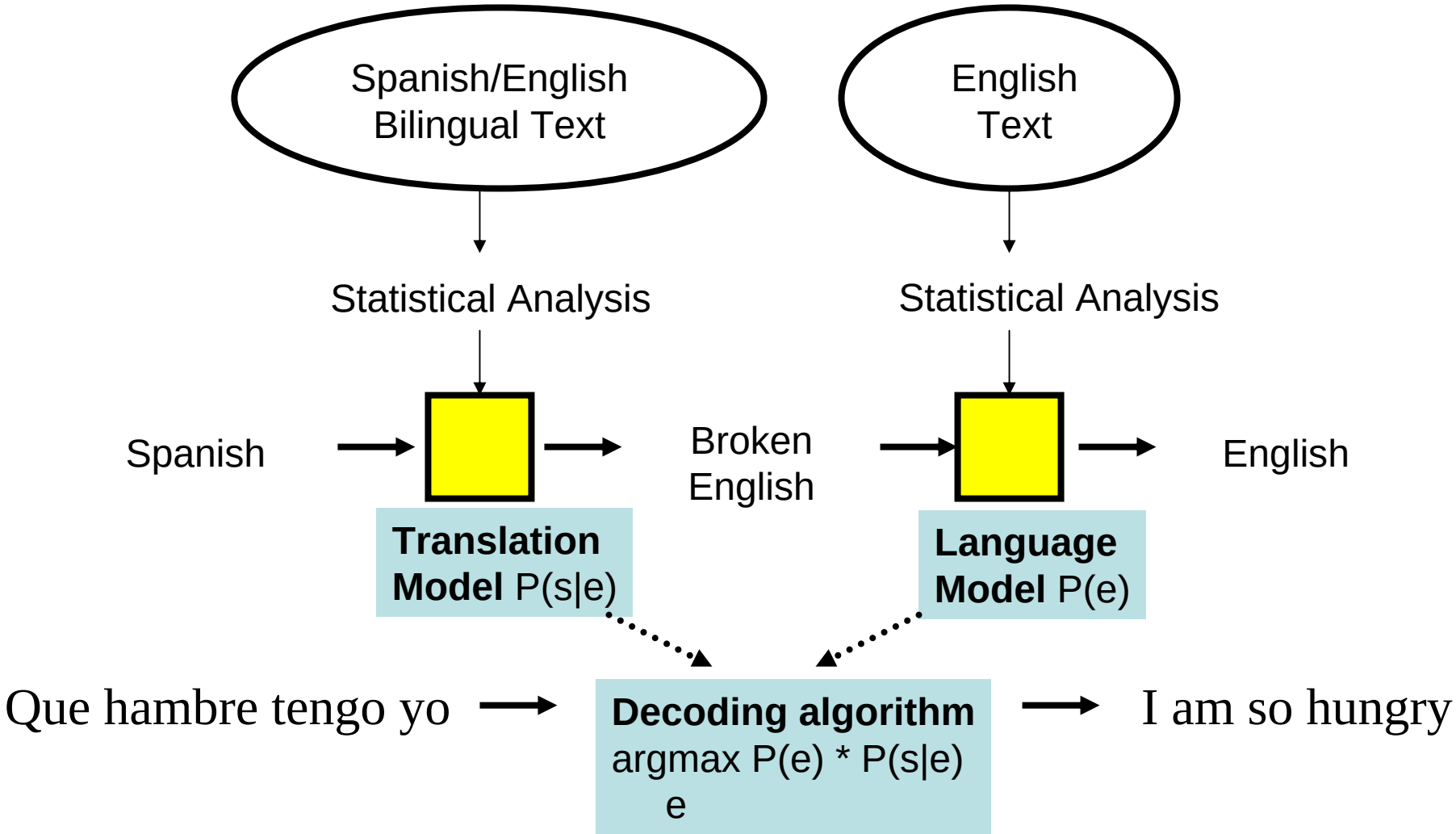
Word-Based Statistical MT

Statistical MT Systems



Que hambre tengo yo → What hunger have I,
Hungry I am so,
I am so hungry,
Have I that hunger … → I am so hungry

Statistical MT Systems



Three Problems for Statistical MT

- Language model
  - Given an English string e, assigns P(e) by formula
  - good English string -> high P(e)
  - random word sequence -> low P(e)

- Translation model
  - Given a pair of strings <f,e>, assigns P(f | e) by formula
  - <f,e> look like translations -> high P(f | e)
  - <f,e> don't look like translations -> low P(f | e)

- Decoding algorithm
  - Given a language model, a translation model, and a new sentence f … find translation e maximizing P(e) * P(f | e)

Language Modeling

Goal of a language model for MT:

He is on the soccer field
He is in the soccer field

Is table the on cup the
The cup is on the table

American shrine
American company

Need to make these decisions, because translation model may not have a lot of context information!

The Classic Language Model

Word Bigrams

Process model of English:

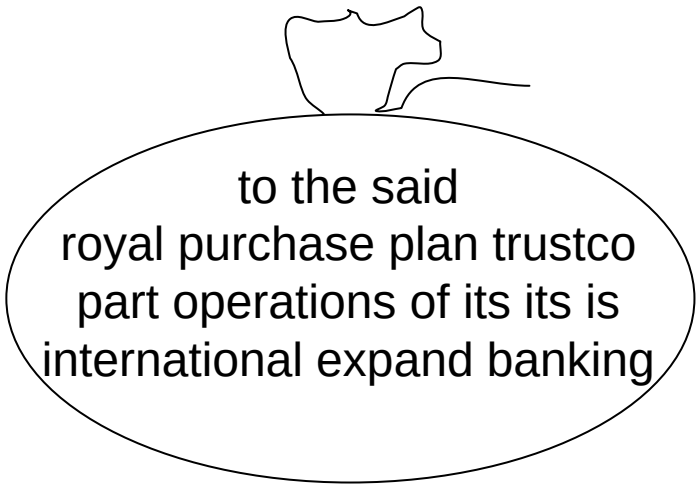
Generate each word based only on the previous word.

P(I saw water on the table) =

P(I | START) ·
P(saw | I) ·
P(water | saw) ·
P(on | water) ·
P(the | on) ·
P(table | the) ·
P(END | table)

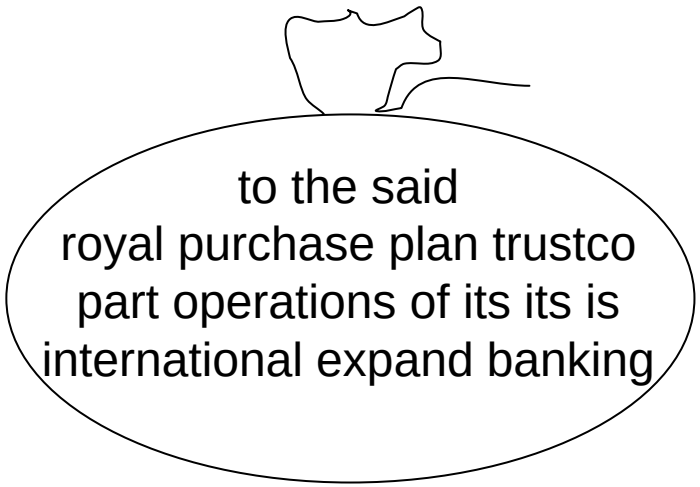
Probabilities can be tabulated from an online English corpus … just like Weaver's Turkish case.

Trigram Language Model



[Soricut & Marcu, 05]

Trigram Language Model

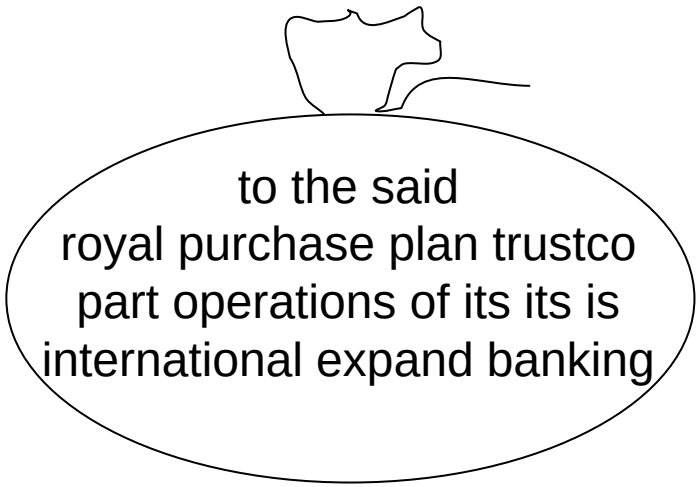


random order

the banking trustco is said to
expand its purchase part of its
royal international plan operations

[Soricut & Marcu, 05]

Trigram Language Model



trigram
best
order

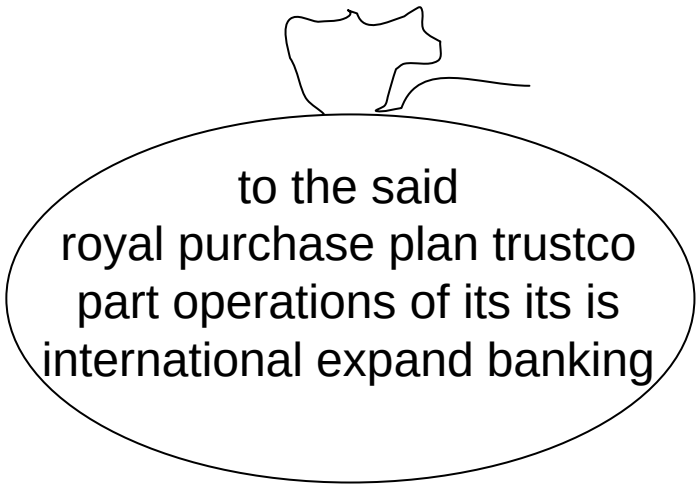
the banking trustco is said to
expand its purchase part of its
royal international plan operations

royal trustco said the purchase is
part of its plan to expand
its international banking operations

N-grams have a lot of
semantics in them!

[Soricut & Marcu, 05]

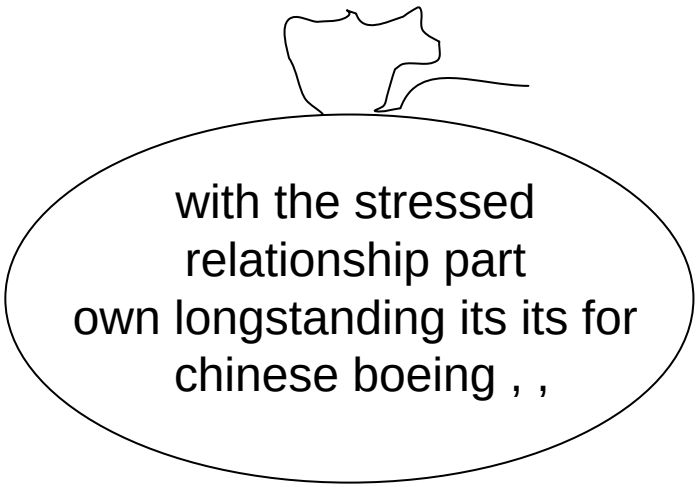
Trigram Language Model



trigram best order →

the banking trustco is said to
expand its purchase part of its
royal international plan operations

royal trustco said the purchase is
part of its plan to expand
its international banking operations



trigram best order →

for its part, stressed the longstanding
relationship with its own, chinese boeing

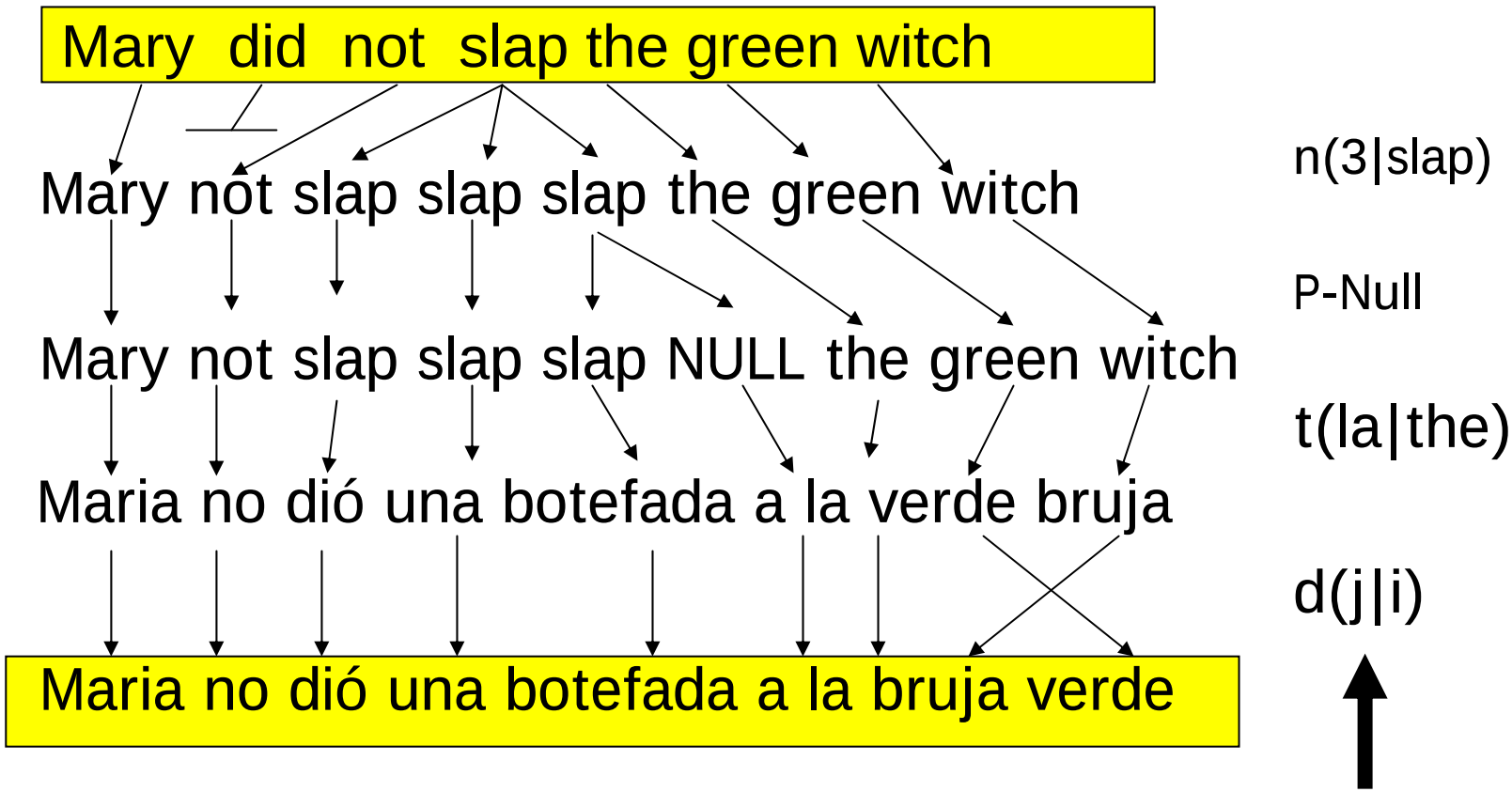
boeing, for its part, stressed its own
longstanding relationship with the chinese

[Soricut & Marcu, 05]

The Classic Translation Model

Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

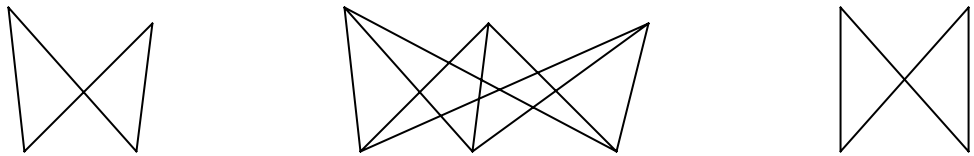
Generative story:



Probabilities can be learned from raw bilingual text.

Statistical Machine Translation

… la maison … la maison bleue … la fleur …



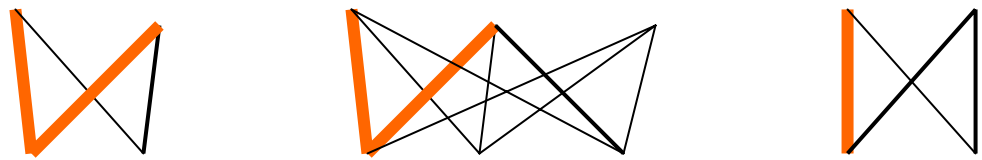
… the house … the blue house … the flower …

All word alignments equally likely

All P(french-word | english-word) equally likely

Statistical Machine Translation

… la maison … la maison bleue … la fleur …

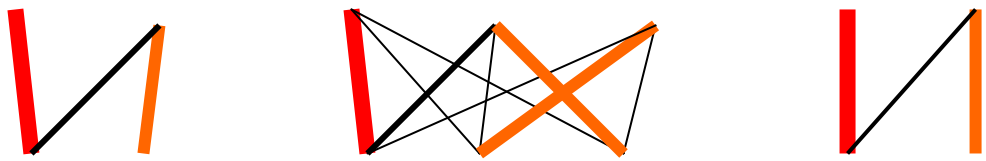


… the house … the blue house … the flower …

“la” and “the” observed to co-occur frequently,
so P(la | the) is increased.

Statistical Machine Translation

… la maison … la maison bleue … la fleur …



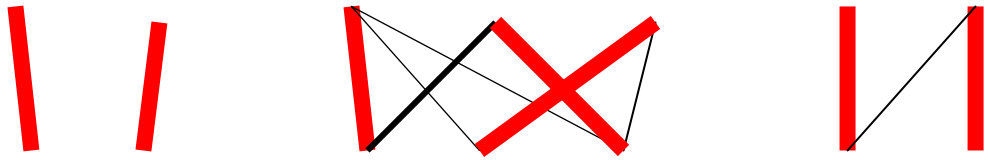
… the house … the blue house … the flower …

“house” co-occurs with both “la” and “maison”, but P(maison | house) can be raised without limit, to 1.0, while P(la | house) is limited because of “the”

(pigeonhole principle)

Statistical Machine Translation

… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

settling down after another iteration

Statistical Machine Translation

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

Inherent hidden structure revealed by EM training!

For details, see:

- “A Statistical MT Tutorial Workbook” (Knight, 1999).
- “The Mathematics of Statistical Machine Translation” (Brown et al, 1993)
- Software: GIZA++

Statistical Machine Translation

… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

P(juste | fair) = 0.411
P(juste | correct) = 0.027
P(juste | right) = 0.020
…

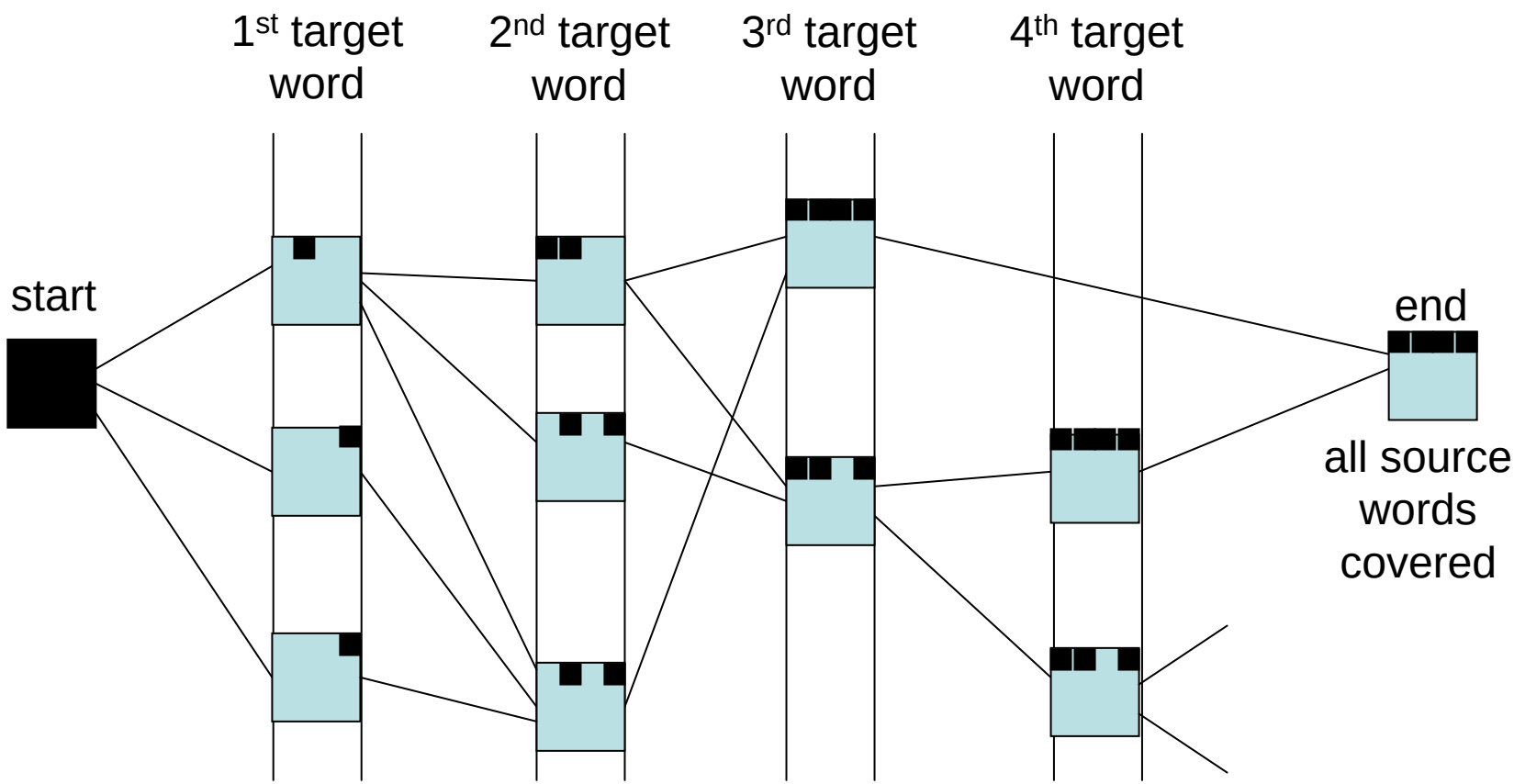
new French sentence

Possible English translations, to be rescored by language model

Decoding for “Classic” Models

- Of all conceivable English word strings, find the one maximizing $P(e) \times P(f \mid e)$
- Decoding is an NP-complete challenge
  - (Knight, 1999)
- Several search strategies are available
- Each potential English output is called a *hypothesis*.

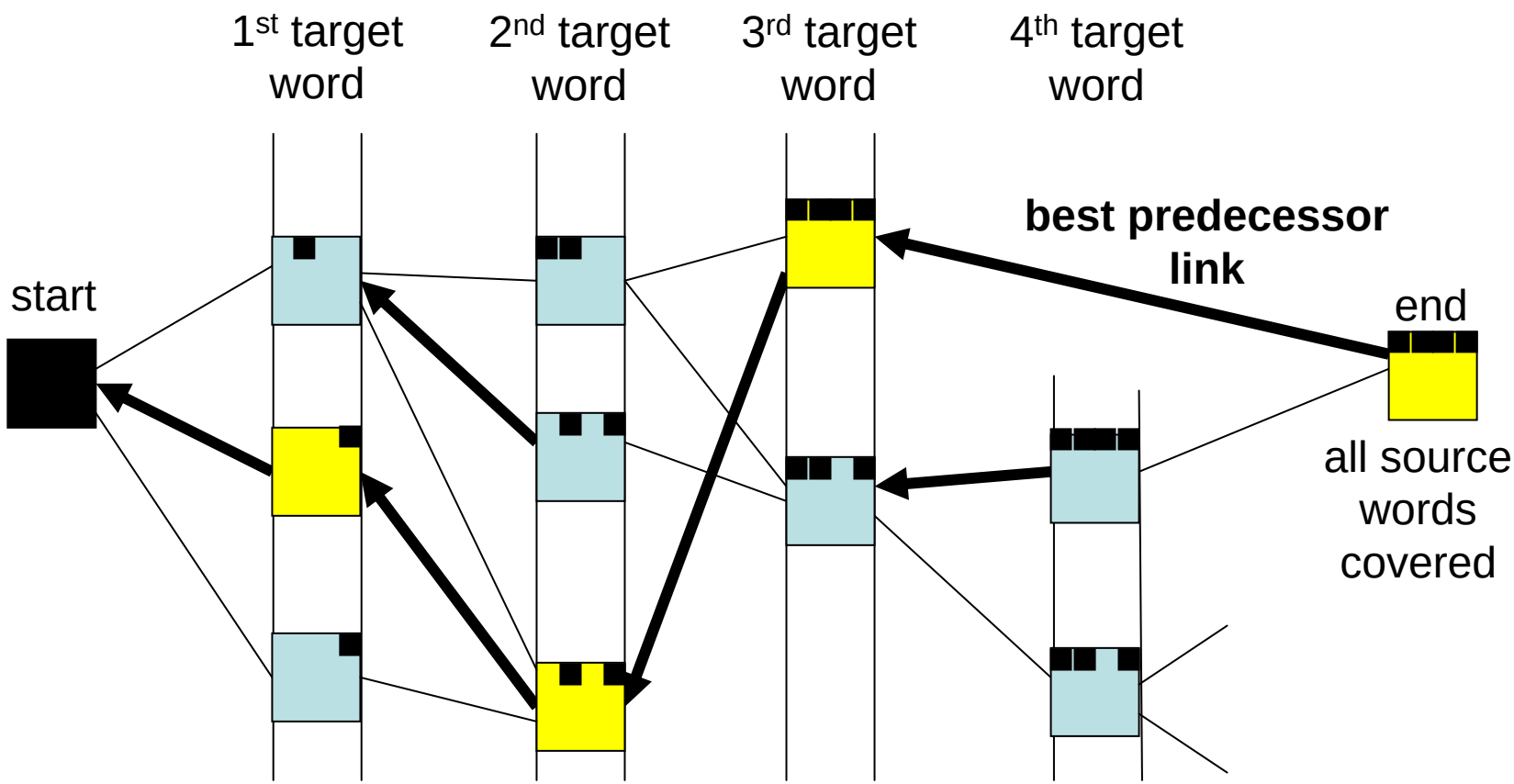
Dynamic Programming Beam Search



Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969;
Brown et al, 1996 US Patent;
(Och, Ueffing, and Ney, 2001]

Dynamic Programming Beam Search



Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969;
Brown et al, 1996 US Patent;
(Och, Ueffing, and Ney, 2001]

The Classic Results

- *la politique de la haine .* *(Foreign Original)*
- politics of hate . (Reference Translation)
- the policy of the hatred . (IBM4+N-grams+Stack)

- *nous avons signé le protocole .* *(Foreign Original)*
- we did sign the memorandum of agreement . (Reference Translation)
- we have signed the protocol . (IBM4+N-grams+Stack)

- *où était le plan solide ?* *(Foreign Original)*
- but where was the solid plan ? (Reference Translation)
- where was the economic base ? (IBM4+N-grams+Stack)

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

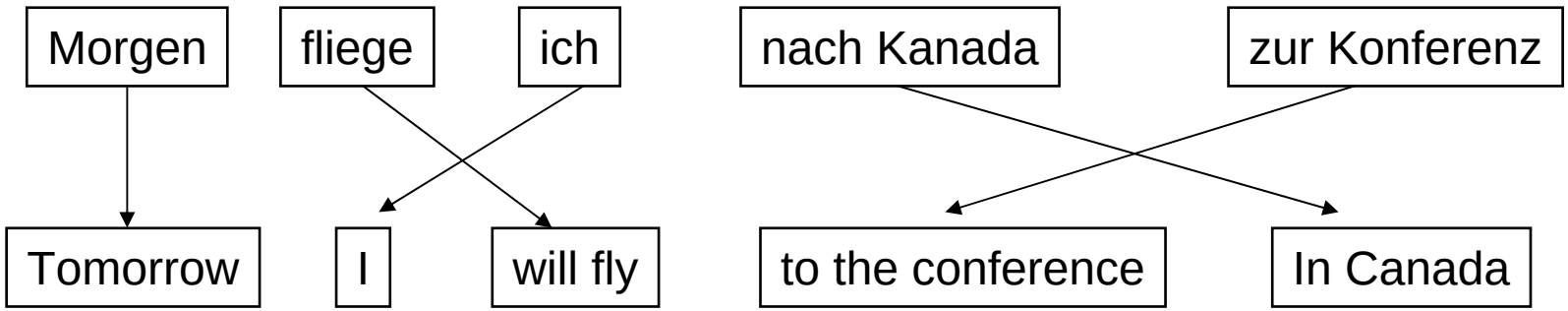
the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and

Flaws of Word-Based MT

- Multiple English words for one French word
  - IBM models can do one-to-many (fertility) but not many-to-one
- Phrasal Translation
  - “real estate”, “note that”, “interest in”
- Syntactic Transformations
  - Verb at the beginning in Arabic
  - Translation model penalizes any proposed re-ordering
  - Language model not strong enough to force the verb to move to the right place

Phrase-Based Statistical MT

Phrase-Based Statistical MT



- Foreign input segmented in to phrases
  - “phrase” is any sequence of words
- Each phrase is probabilistically translated into English
  - P(to the conference | zur Konferenz)
  - P(into the meeting | zur Konferenz)
- Phrases are probabilistically re-ordered

See [Koehn et al, 2003] for an intro.

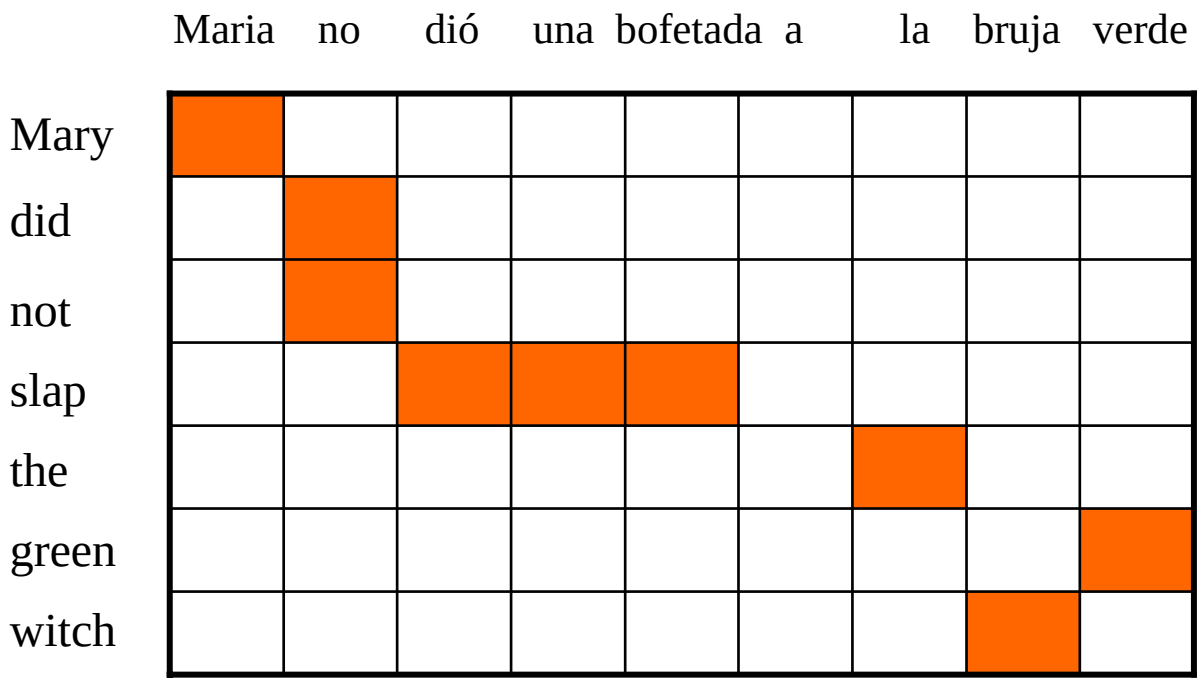
This is state-of-the-art!

Advantages of Phrase-Based

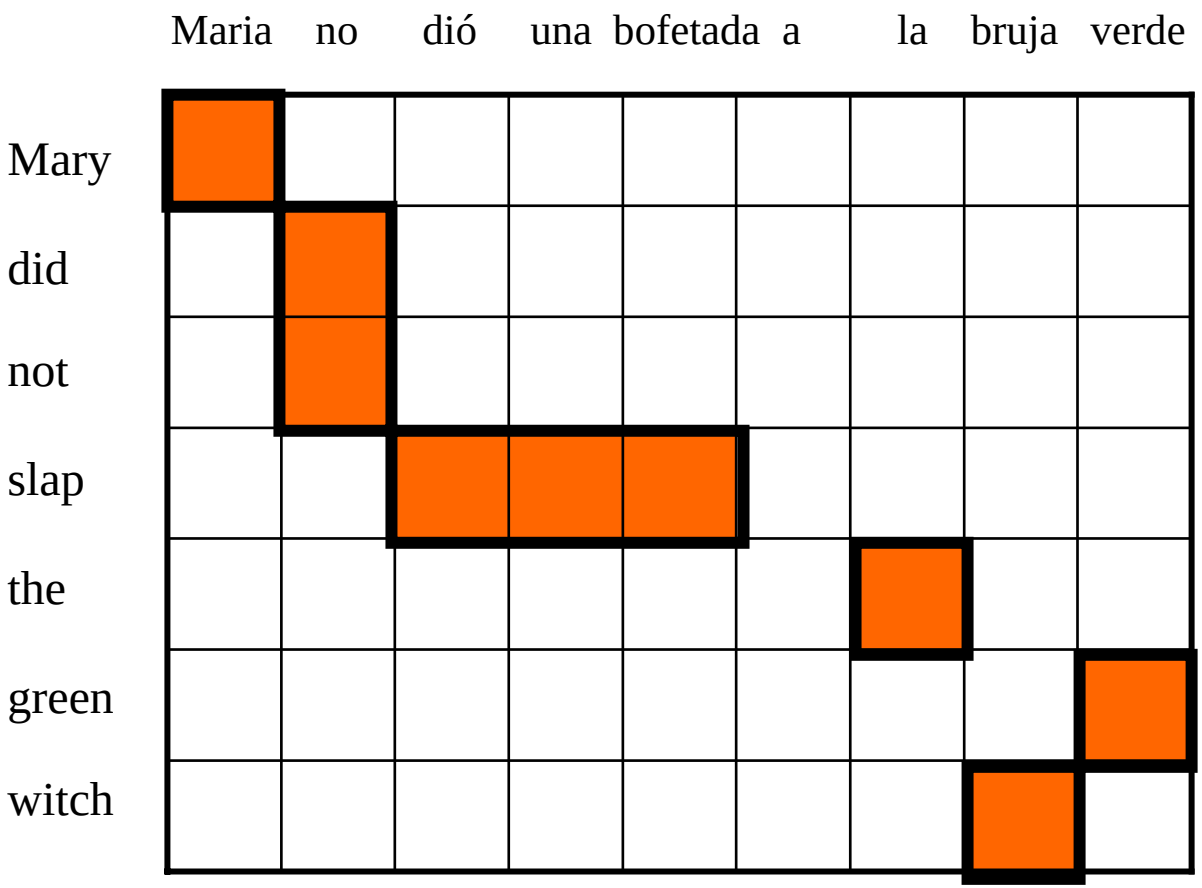
- Many-to-many mappings can handle non-compositional phrases
- Local context is very useful for disambiguating
  - “Interest rate” → …
  - “Interest in” → …
- The more data, the longer the learned phrases
  - Sometimes whole sentences

How to Learn the Phrase Translation Table?

- Start with word alignment, build phrases from that.

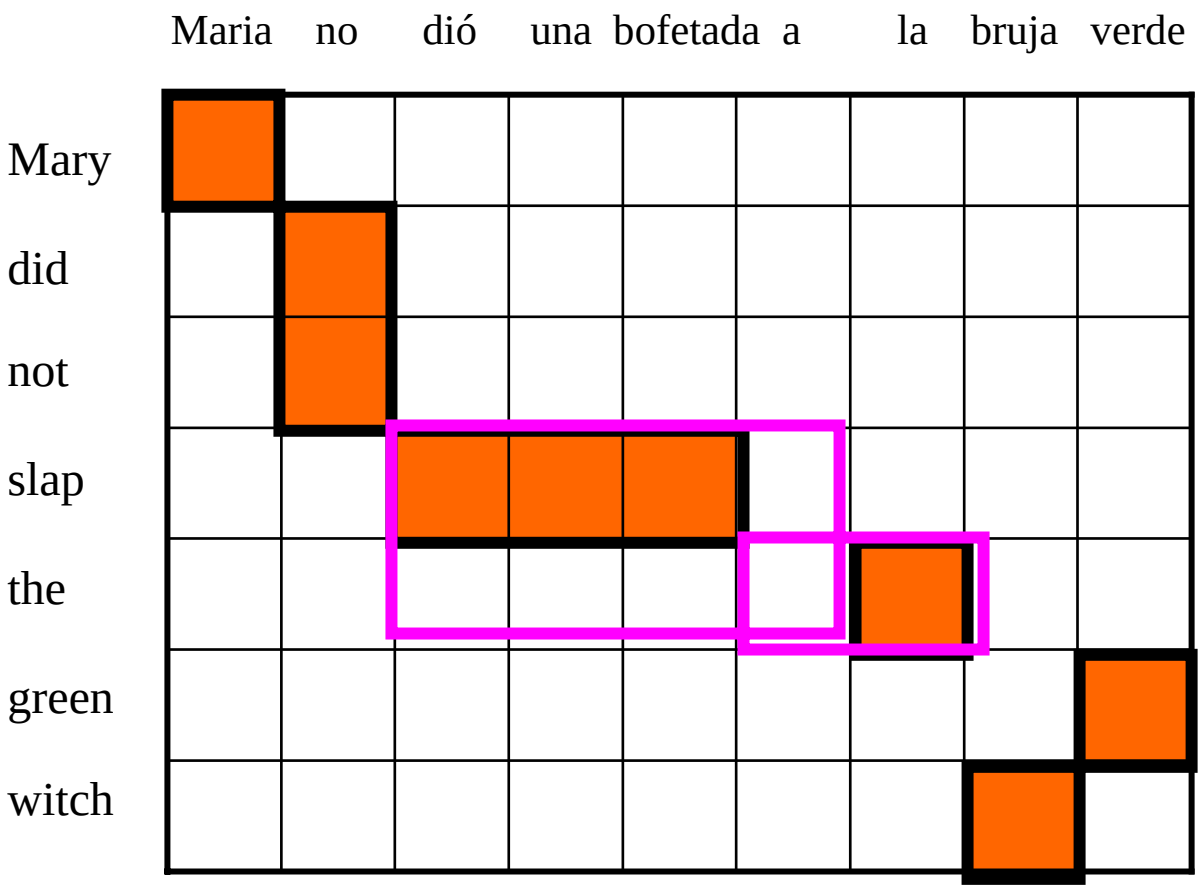


Word Alignment Induced Phrases



(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

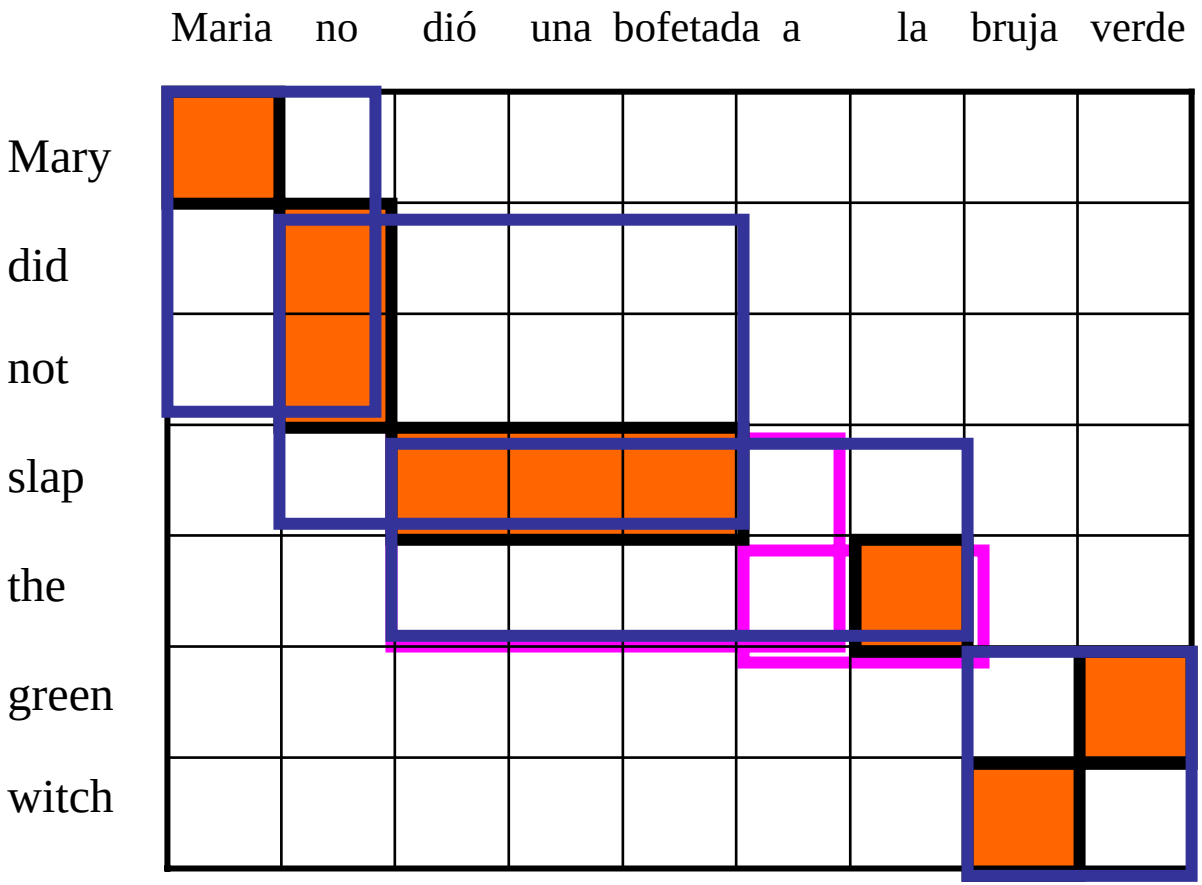
Word Alignment Induced Phrases



(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(a la, the) (dió una bofetada a, slap the)

Word Alignment Induced Phrases



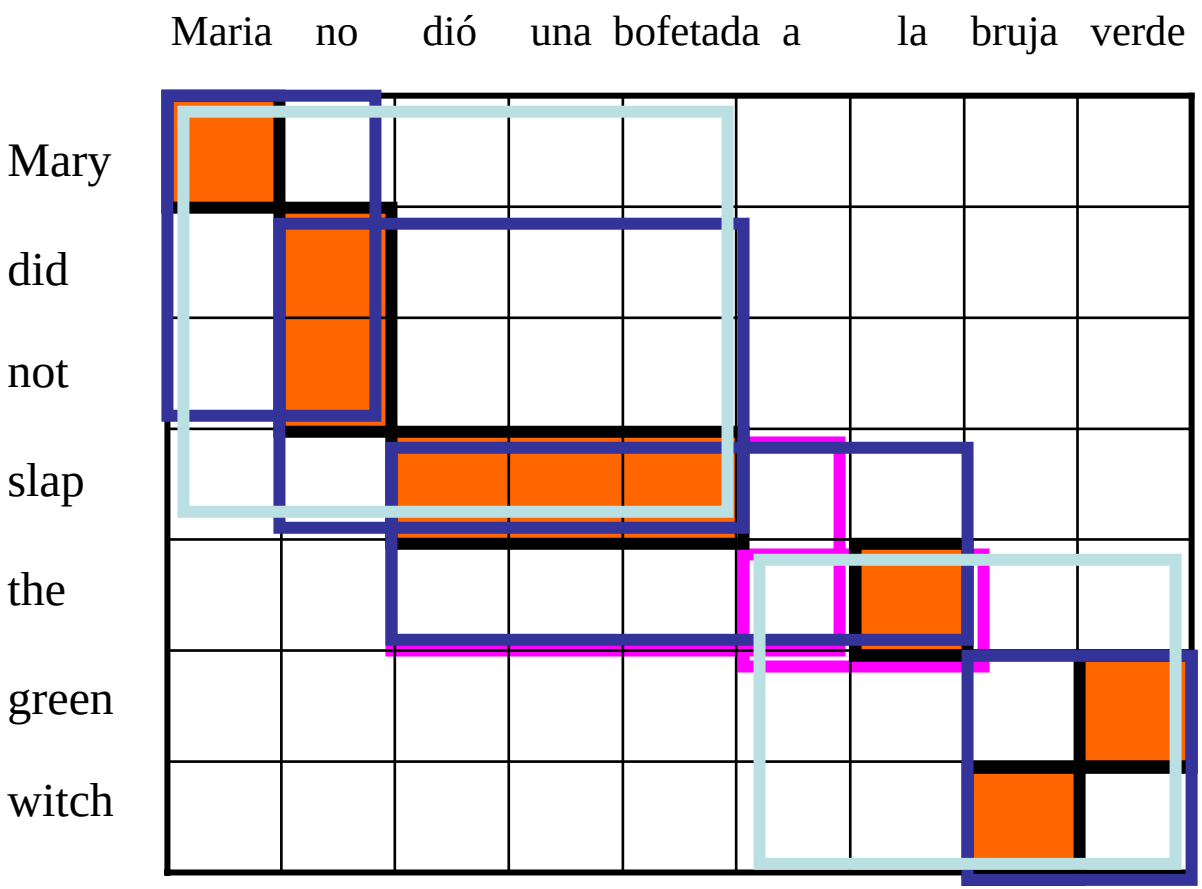
(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(a la, the) (dió una bofetada a, slap the)

(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)

(bruja verde, green witch)

Word Alignment Induced Phrases



(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

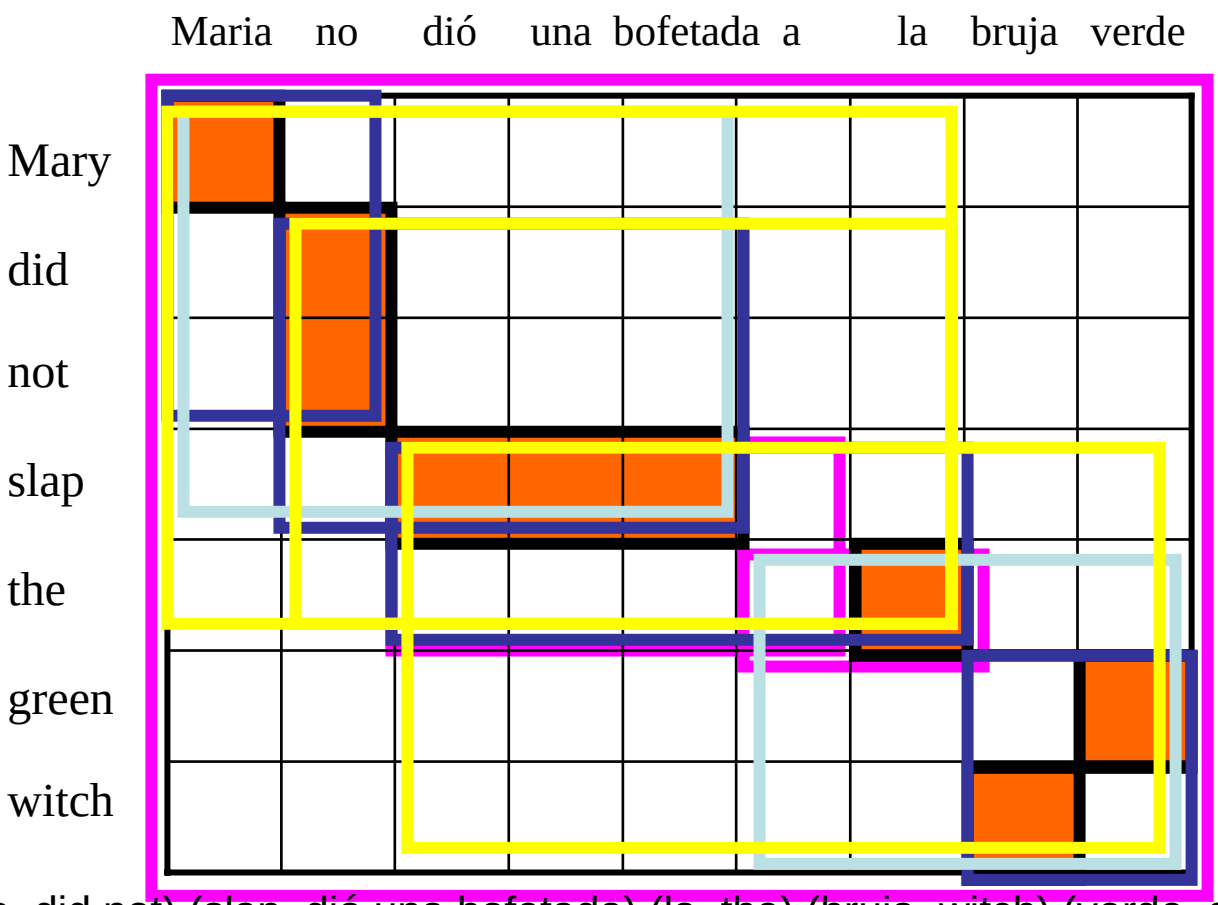
(a la, the) (dió una bofetada a, slap the)

(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)

(bruja verde, green witch) (Maria no dió una bofetada, Mary did not slap)

(a la bruja verde, the green witch) …

Word Alignment Induced Phrases



(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(a la, the) (dió una bofetada a, slap the)

(Maria no, Mary did not) (no dió una bofetada, did not slap), (dió una bofetada a la, slap the)

(bruja verde, green witch) (Maria no dió una bofetada, Mary did not slap)

(a la bruja verde, the green witch) …

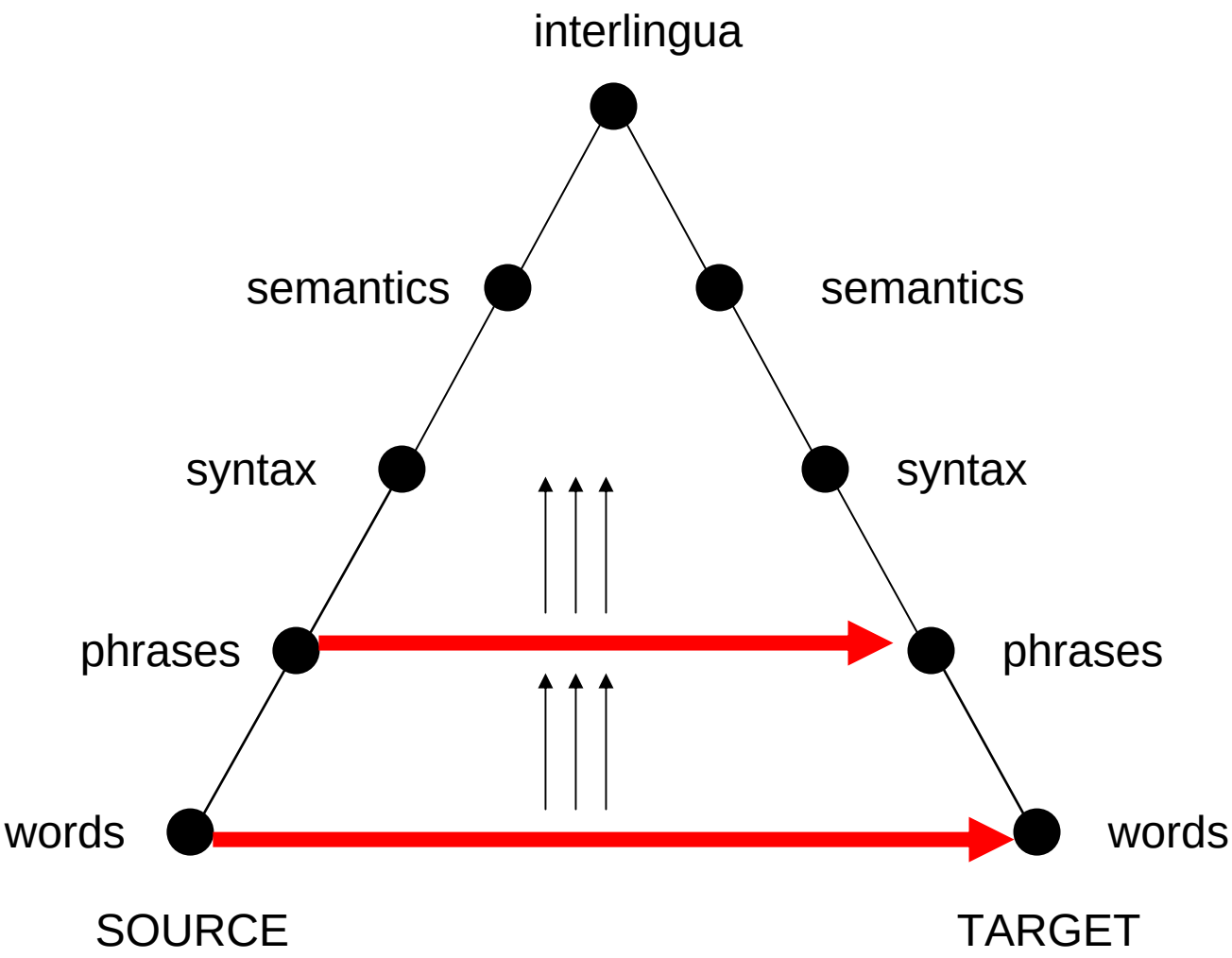
(Maria no dió una bofetada a la bruja verde, Mary did not slap the green witch)

Phrase-Based Models

- So, now we have a vast list of phrase pairs with frequencies
  - P(F | E) = count(F, E) / count(E)
- Billions of phrase pairs ready in our database!
- Translation accuracy is much better than with word-based methods

Syntax and Semantics in Statistical MT

MT Pyramid

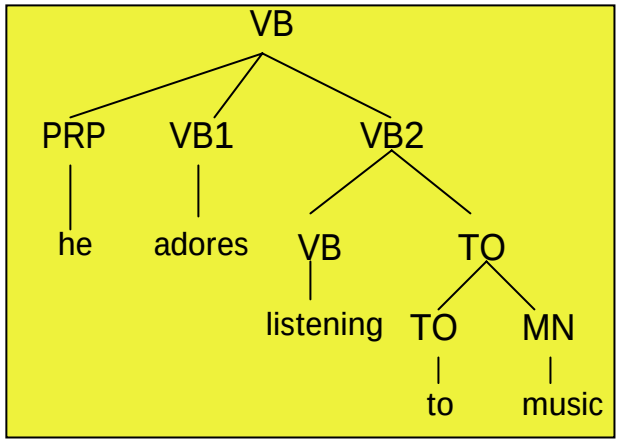

interlingua
semantics
semantics
syntax
syntax
phrases
phrases
words
words
SOURCE
TARGET

Why Syntax?

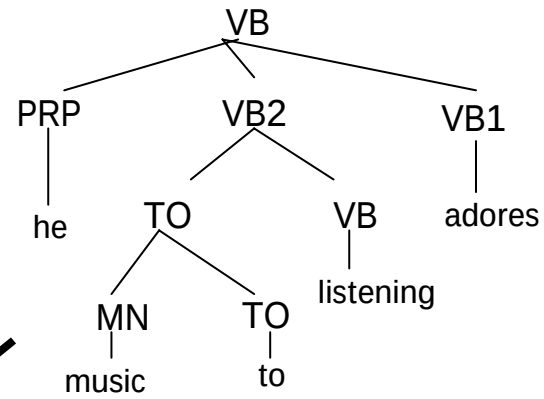
- Need much more grammatical output
- Need accurate control over re-ordering
- Need accurate insertion of function words
- Word translations need to depend on grammatically-related words

Yamada/Knight 01: Modeling and Training

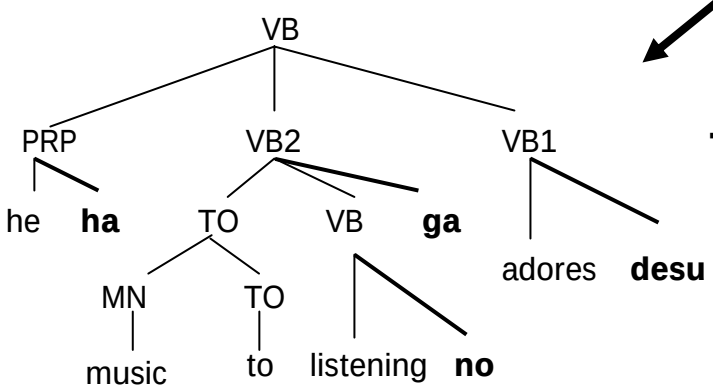
Parse Tree(E)



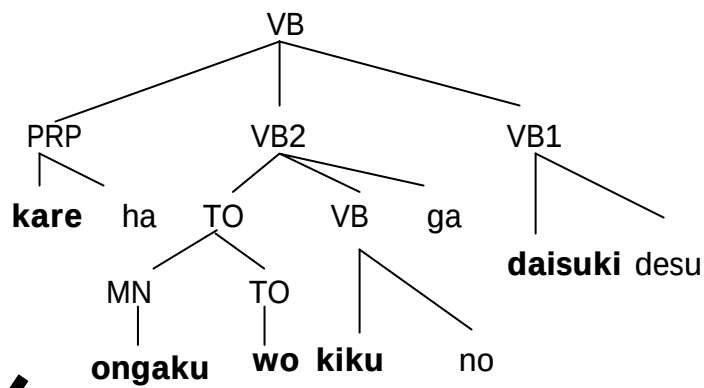
Reorder



Insert



Translate



Take Leaves

Sentence(J)

Kare ha ongaku wo kiku no ga daisuki desu

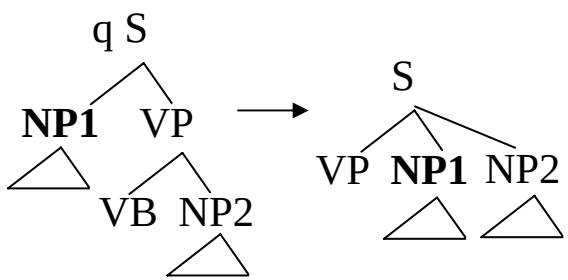
Japanese/English Reorder Table

Original Order	Reordering	P(reorder|original)
PRP VB1 VB2	PRP VB1 VB2	0.074
	PRP VB2 VB1	**0.723**
	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
	TO VB	**0.893**
TO NN	TO NN	0.251
	NN TO	**0.749**

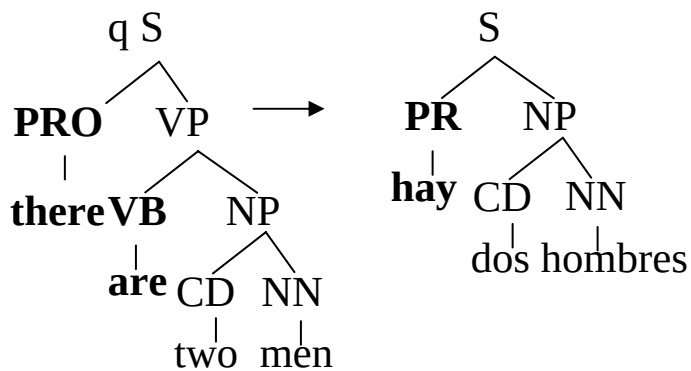
For French/English, useful parameters like P(N ADJ | ADJ N).

Syntax MT Models are Tree Automata [Graehl & Knight 04]

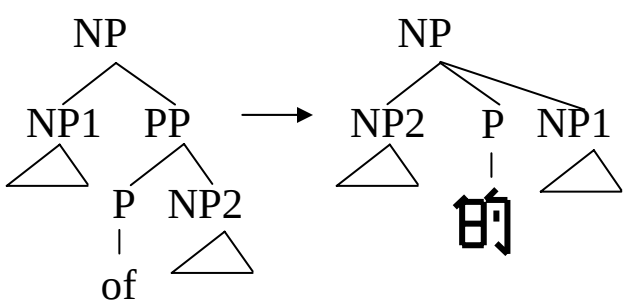
Non-local Re-Ordering (*English/Arabic*)



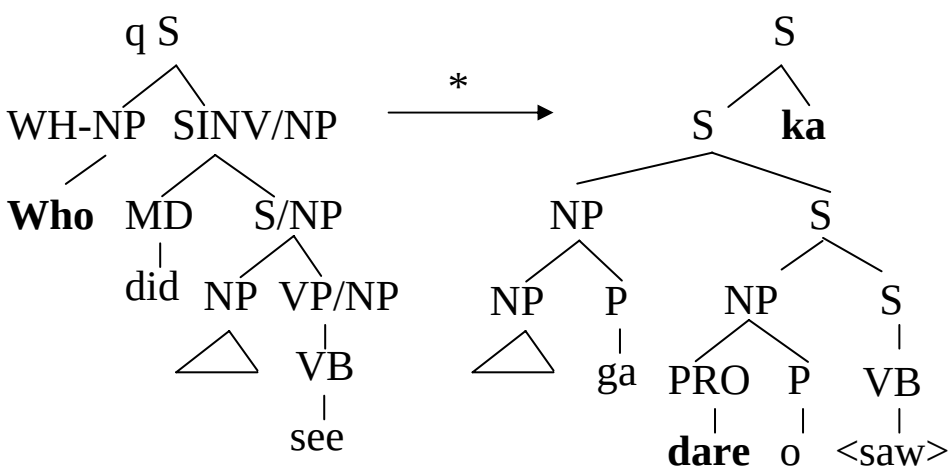
Non-constituent Phrasal Translation (*English/Spanish*)



Lexicalized Re-Ordering (*English/Chinese*)



Long-distance Re-Ordering (*English/Japanese*)



Present and Future

- Phrase-based models have been state-of-the-art for six years
  - Word alignments
  - Phrase pair extraction & probabilities
  - N-gram language models
  - Beam search decoding
  - Feature functions & learning weights

- In 2006, statistical models using syntax outperformed phrase-based models for the first time! Only just beginning…

the end