

Machine Translation

Kevin Knight

USC/Information Sciences Institute
USC/Computer Science Department



Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

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

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

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** **clock** 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 eneak .
2a. ok-drubel ok-voon anak plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anak plok nok . / 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok . / 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghrok clock . X / 10b. wat nnat gat mat bat hilat . process of elimination
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** **clock** 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 eneak .
2a. ok-drubel ok-voon anak plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anak plok nok . / 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghrok clock . X / / 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 eneak .
2a. ok-drubel ok-voon anak plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anak plok nok . / 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok . / / 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghrok klok . / / / 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . / 5b. totat jjat quat cat .	11a. lalok nok crrok hihok yorok zanzanak . / └ / / / zero 11b. wat nnat arrat mat zanzanat . 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 .

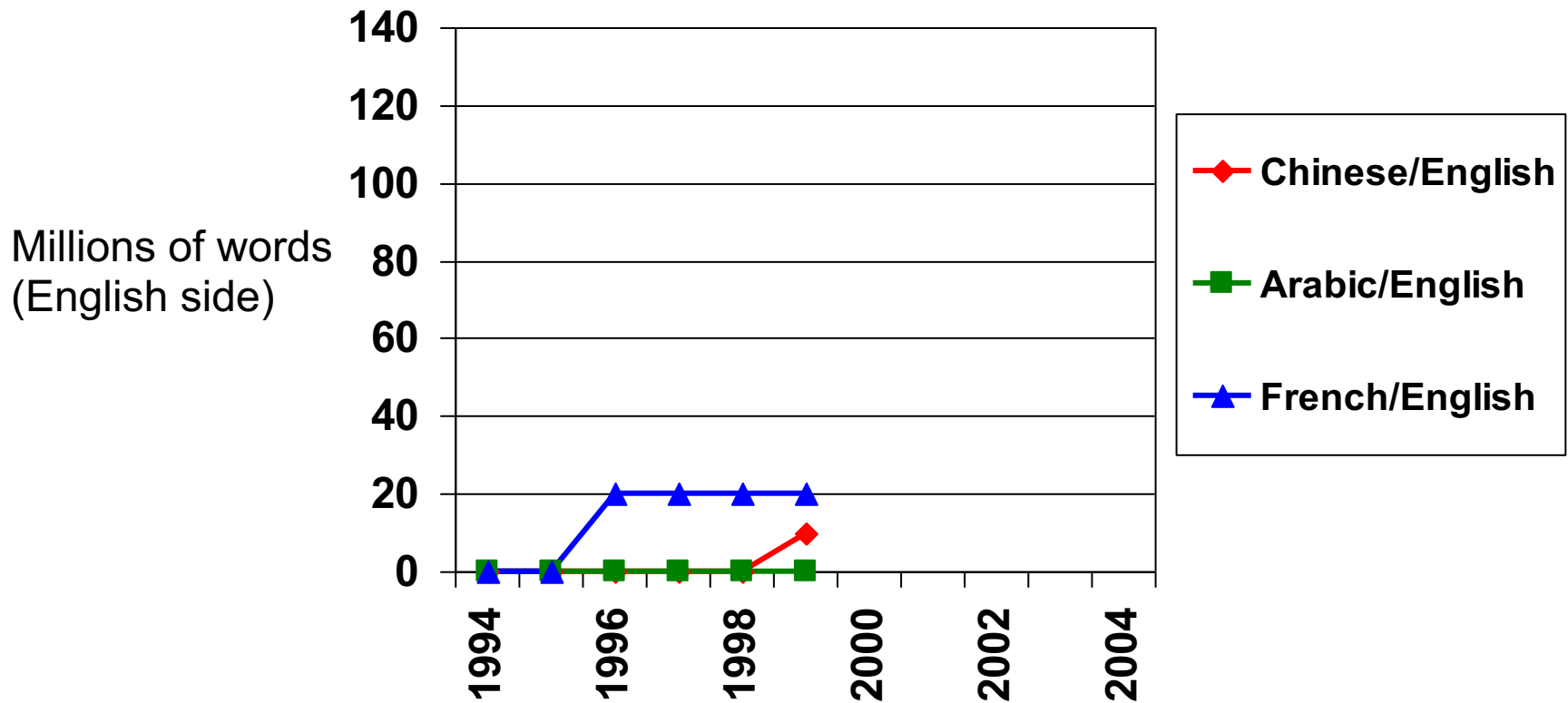
6a. the associates are also angry .

6b. los asociados tambien estan enfadados .

12a. the small groups are not modern .

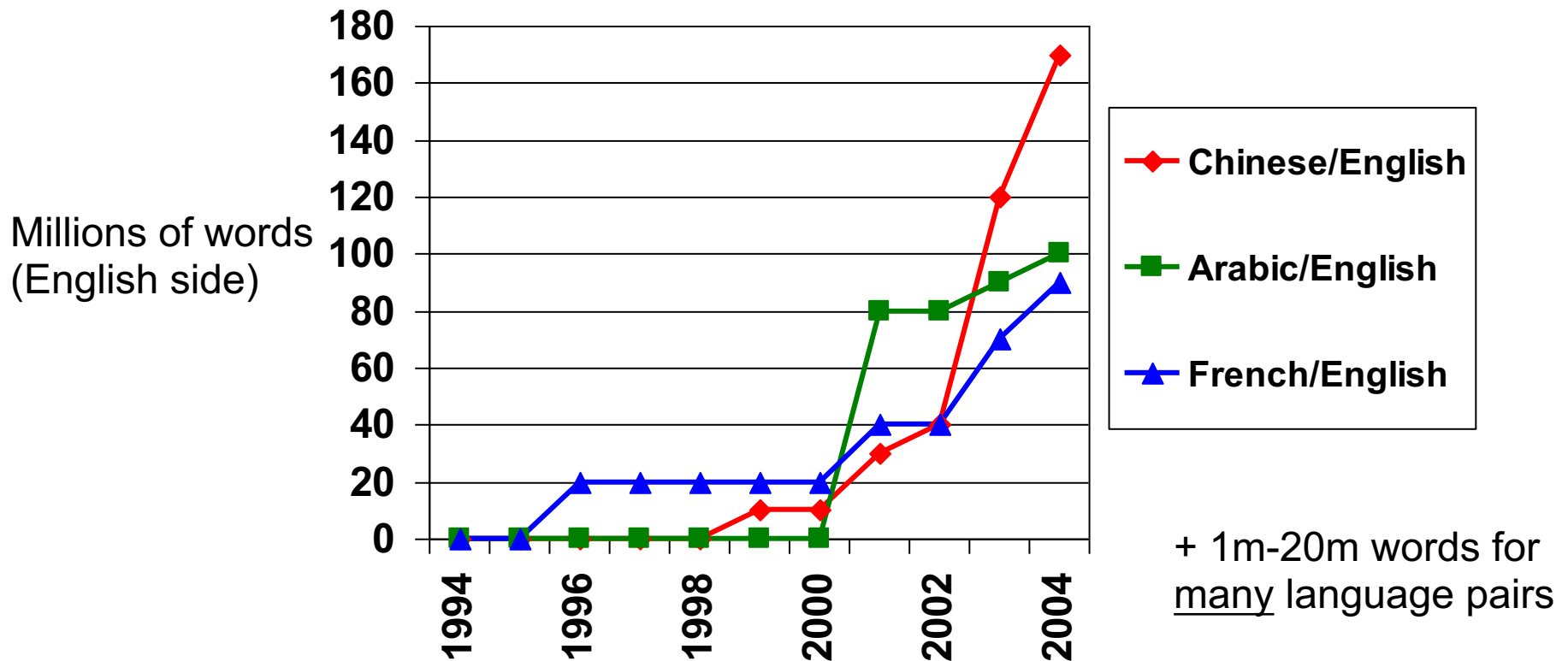
12b. los grupos pequenos no son modernos .

Ready-to-Use Online Bilingual Data



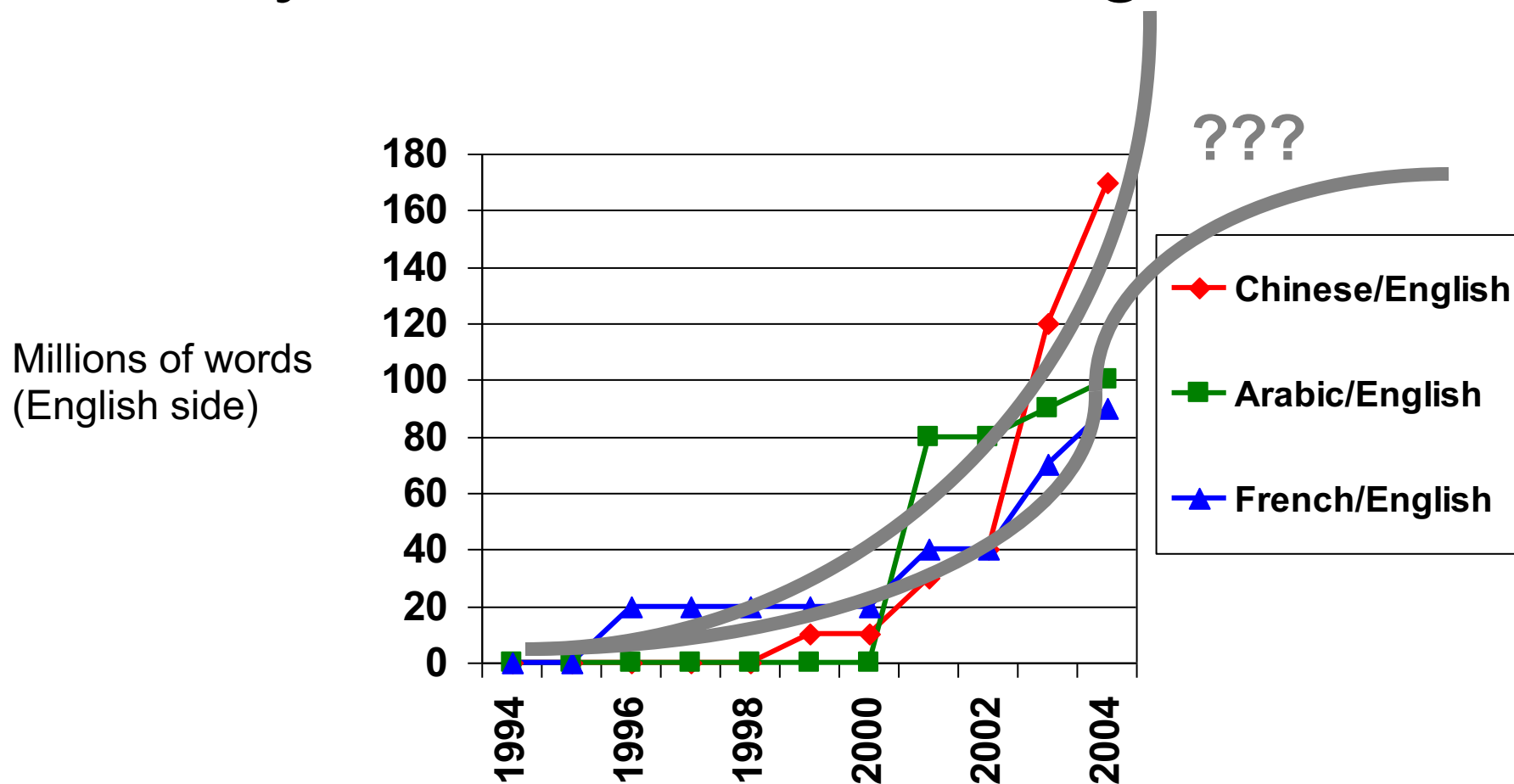
(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

Ready-to-Use Online Bilingual Data



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

Ready-to-Use Online Bilingual Data



→ One Billion?

From No Data to Sentence Pairs

- Easy way: Linguistic Data Consortium (LDC)
- Really hard way: pay \$\$\$
 - Suppose one billion words of parallel data were sufficient
 - At 20 cents/word, that's \$200 million
- Pretty hard way: Find it, and then earn it!
 - De-formatting
 - Remove strange characters
 - Character code conversion
 - Document alignment
 - **Sentence alignment**
 - **Tokenization (also called Segmentation)**

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.
- The diagram illustrates the alignment of five English sentences with three Spanish sentences. Arrows show the following connections: English sentence 1 aligns with Spanish sentence 1; English sentence 2 aligns with Spanish sentence 1; English sentence 3 aligns with Spanish sentence 2; English sentence 4 aligns with Spanish sentence 3; and English sentence 5 aligns with Spanish sentence 3.

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。