

Machine Translation

Computational Linguistics;

Ling 409;

Spring 2013

Some strange terminology...

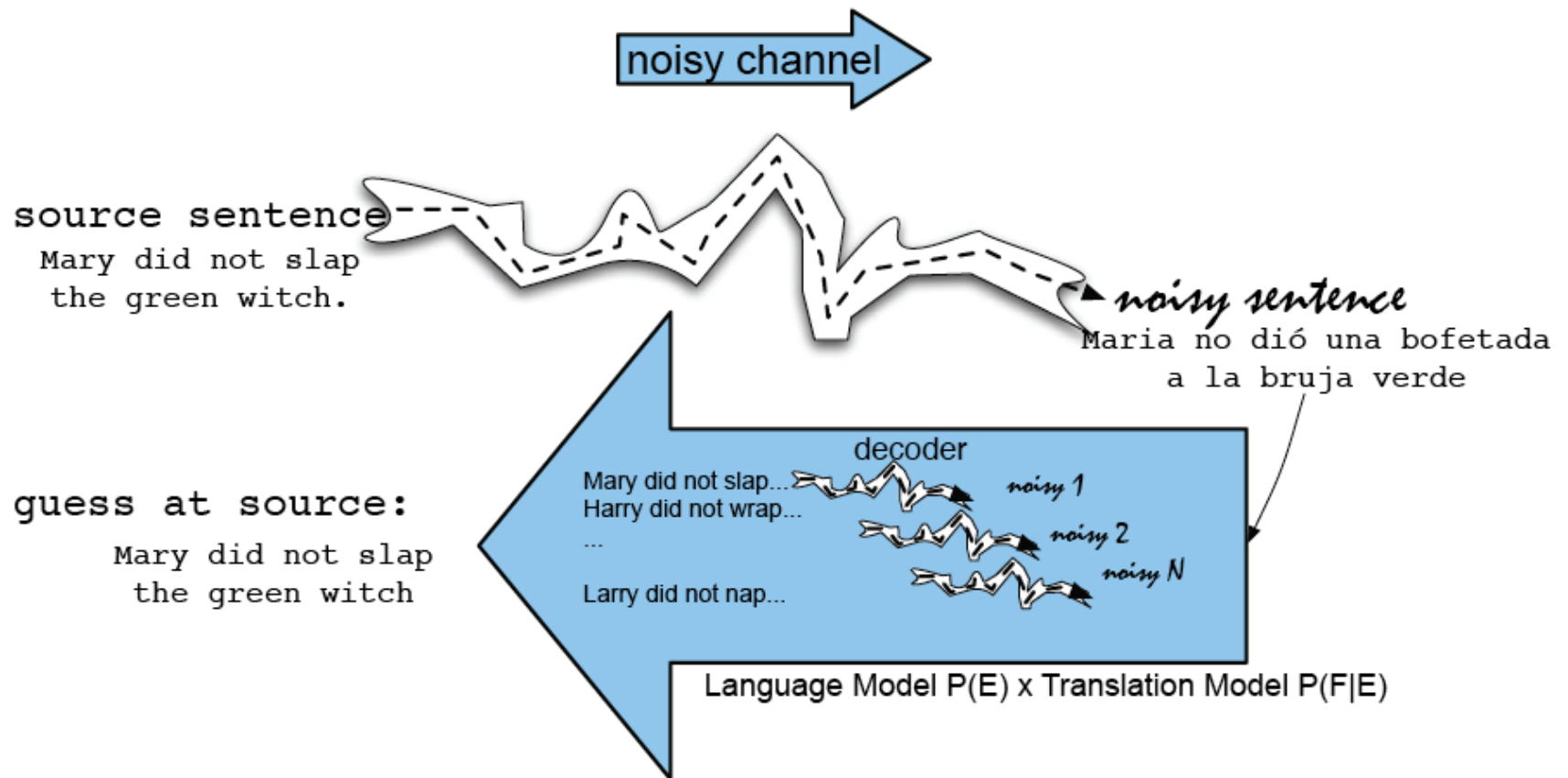
- ▶ If translating, for example, French to English, we're going to call the English the **source** and the French the **target**.
- ▶ But... ¿why?

Warren Weaver (1947)

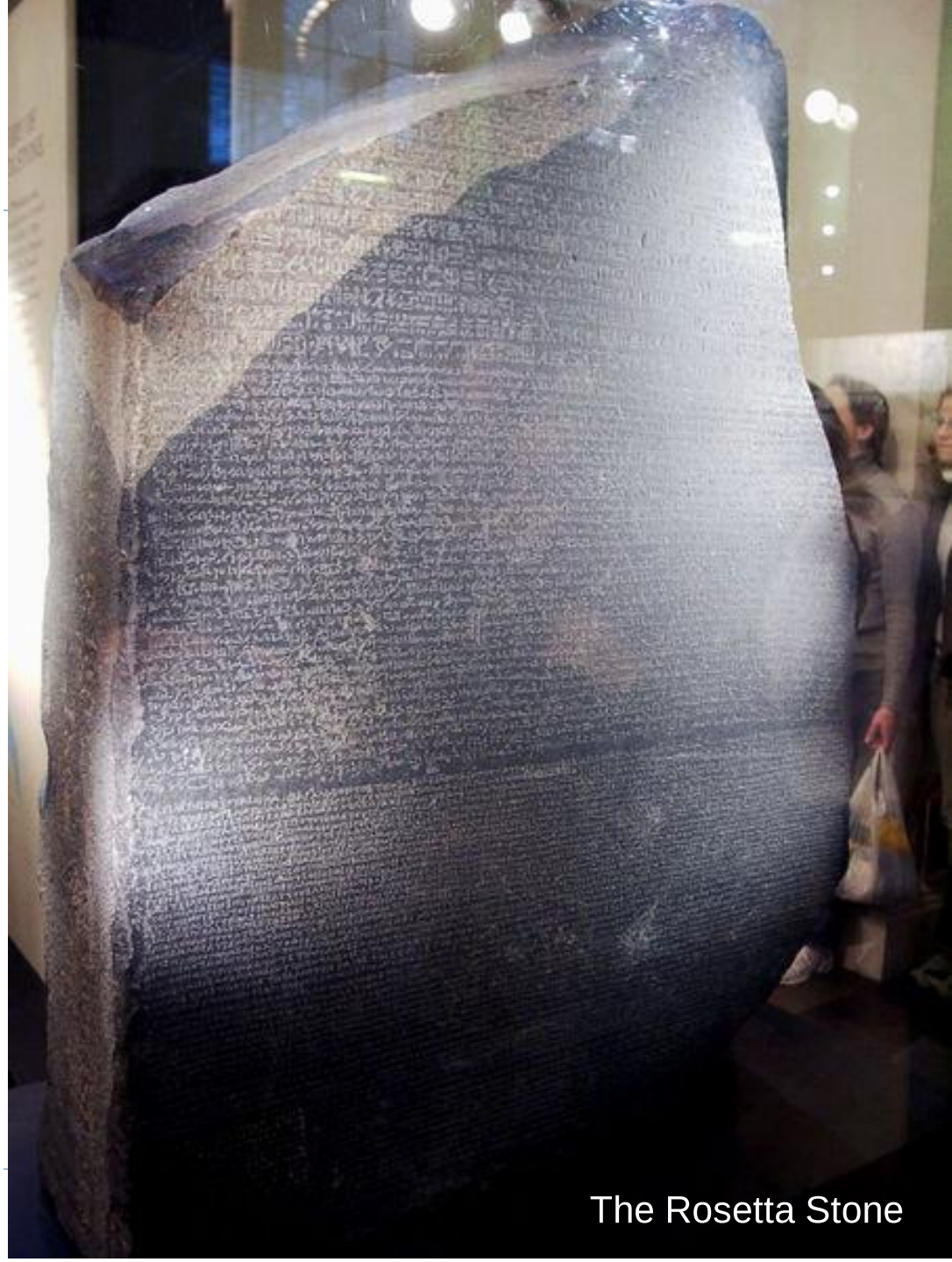
“ When I look at an article in Russian, I say to myself: **This is really written in English, but it has been coded in some strange symbols.** I will now proceed to decode.” [emphasis mine]



Noisy Channel



What is this?



Training Data!

Egyptian hieroglyphs

Demotic

Greek



Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

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 klok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok **farok** izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok **farok** ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

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 klok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok **farok** izok stok .

5b. totat **jjat** quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok **farok** ororok lalok sprok izok enemok .

7b. wat **jjat** bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok **crrrok** hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ??

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ??

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ??

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloak at-yurp .

10a. lalok mok nok **yorok** ghrok klok .

10b. wat nnat gat **mat** bat hilat .

11a. lalok nok crrrok hihok **yorok** zanzanak .

11b. wat nnat arrat **mat** zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat mat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat mat

Centauri/Arcturan (Knight 1997)

Your assignment, translate this to Arcturan:

crrok hihok yorok **clock** kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok clock .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat mat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok **clock** .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

???

Process of elimination

10a. lalok mok nok yorok ghirok **clok** .

10b. wat nnat gat mat bat hilat .

clok **must** align with
Something, right?

jjat ?? arrat mat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok **clock** .

process of
elimination

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat mat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok **clock** .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

process of
elimination

jjat ?? arrat mat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloak at-yurp .

10a. lalok mok nok yorok ghirok **clock** .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

process of
elimination

jjat ?? arrat mat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok **clock** .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

process of
elimination

jjat ?? arrat mat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok **clock** .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

process of
elimination

jjat ?? arrat mat bat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok **clock** .

10b. wat nnat gat mat **bat** hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat mat bat

Centauri/Arcturan (Knight 1997)

Your assignment, translate this to Arcturan:

ccrrrok hihok yorok klok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloak at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok ccrrrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ?? arrat mat bat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloak at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

???

jjat ... arrat mat bat

Centauri/Arcturan (Knight 1997)

Your assignment, translate this to Arcturan:

crrok hihok yorok klok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok

3b. totat dat arrat vat hilat

4a. ok-voon anak drok brok

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. lat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloak at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

zero fertility

???

jjat ... arrat mat bat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ... arrat mat bat oloat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat **oloot** at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ... arrat mat bat oloat

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

jjat ... arrat mat bat oloat at-yurp .

Centauri/Arcturan (Knight 1997)

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

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneak .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanak .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

Centauri/Arcturan: New assignment

- ▶ The next task is to arrange these Arcturan words in the correct, idiomatic Arcturan order:
 - ▶ { jjat, arrat, mat, bat, oloat, at-yurp }
- ▶ But we already know how to do that, right?
- ▶ What technology (or technologies) that we've covered could we use here?

The point of this exercise?

- ▶ What didn't we have?
 - ▶ No dictionary required for either language
 - ▶ No grammar
 - ▶ No part of speech tags
 - ▶ No meaning
 - ▶ No broader meaning or social context
- ▶ What did we have?
 - ▶ Just text pairs and a bunch of **really dubious** assumptions.
 - ▶ Such as?



Spanish/English text

Translate: Clients do not sell pharmaceuticals in Europe.

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 .



Statistical machine translation

- ▶ Current statistical MT systems are based on maximizing our old friend $P(E|F)$
- ▶ E is the target language, F is the source
- ▶ We'll start as usual by using Bayes

$$\hat{E} = \operatorname{argmax}_{E \in \text{English}} \underbrace{P(F|E)}_{\text{translation model}} \underbrace{P(E)}_{\text{language model}}$$



The three sub-problems of Stat MT

▶ Language model

- ▶ Given an English string e , assigns $P(e)$ by the usual methods
- ▶ good English string \rightarrow high $P(e)$
- ▶ random word sequence \rightarrow low $P(e)$

▶ Translation model

- ▶ Given a pair of strings $\langle f, e \rangle$, assigns $P(f | e)$ by formula
- ▶ $\langle f, e \rangle$ look like translations \rightarrow high $P(f | e)$
- ▶ $\langle f, e \rangle$ don't look like translations \rightarrow low $P(f | e)$

▶ Decoding algorithm

- ▶ Given a language model, a translation model, and a new sentence f ... find translation e maximizing $P(e) * P(f | e)$



The three sub-problems of Stat MT

- ▶ For the language model, just use normal language models.
- ▶ For the translation model there are lots of choices
 - ▶ Word-based (e.g. Centauri/Arcturan)
 - ▶ Phrase-based
 - ▶ Syntactic
- ▶ For the decoding model we'll focus on A^* like methods (heuristic search methods with a beam)



Word-Based Stat MT

- The basic premise in word-based MT is that the texts consists of words that
 - Need to be (1) translated
 - And (2) moved around

The green witch is at home this week

Diese Woche ist die grune Hexe zu Hause

Word-Based Stat MT

□ So in our Bayesian scheme...

$$\hat{E} = \operatorname{argmax}_{E \in \text{English}} \underbrace{P(F|E)}_{\text{translation model}} \underbrace{P(E)}_{\text{language model}}$$

□ Given the German sentence we need
argmax over the English possibilities

Ok but

- I indicated that we didn't have any resources other than the bilingual texts
 - In particular, no dictionaries, no bilingual dictionaries, and no probabilities
 - Fortunately, we can get all that from the bitexts

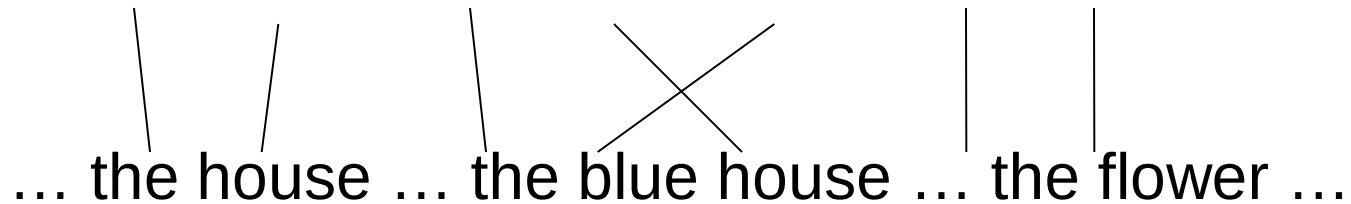
Word Alignments

□ Let's start with a simple alignment type.
From E to F with a 1 to 1 assumption

□ Each word in E aligns with 1 word in F

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

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

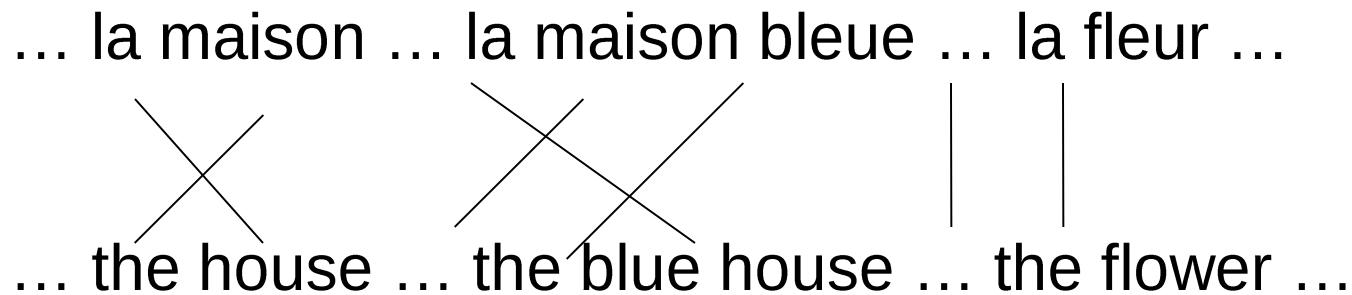


This is one possible alignment.

Word Alignments

- Let's start with a simple alignment type.
From E to F with a 1 to 1 assumption.

... la maison ... la maison bleue ... la fleur ...
... the house ... the blue house ... the flower ...



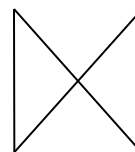
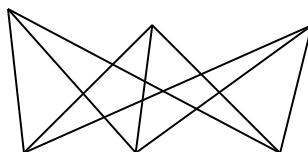
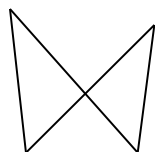
Here's another

Word Alignments

- 1 to 1 is not the only possible or useful type.
- Based on the language pairs, 1 to many, many to 1, 1 to none, etc. are all likely.
 - 1 to many is a word aligned to a phrase
 - Many to 1 is a phrase aligning to a word
 - 1 to none is a word that just isn't there in the other text

Alignment Probabilities

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

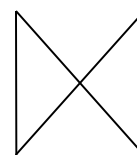
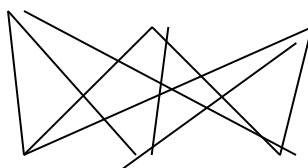
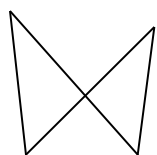


... the house ... the blue house ... the flower ...

- Assume that all word alignments equally likely.
- That is, that all $P(\text{french-word} \mid \text{english-word})$ are equal
- Recall that **we want $P(f|e)$**

Word Alignment

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



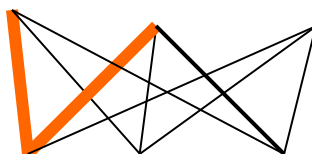
... the house ... the blue house ... the flower ...

Assume we're interested in $P(\text{la}|\text{the})$. "the" can co-occur with (aligns) with 4 distinct french words. If we make each of those equally likely then the $P(\text{la} | \text{the})$ is .25.

	La	Maison	Bleue	Fleur
the	.25	.25	.25	.25

Word Alignment

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

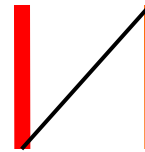
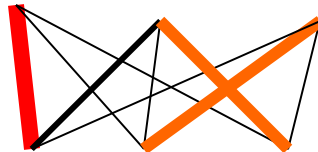
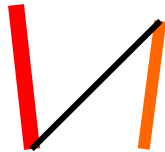


... the house ... the blue house ... the flower ...

But “la” and “the” are observed to co-occur more frequently than expected so $P(\text{la} \mid \text{the})$ should be higher. Meaning the others $P(x \mid \text{the})$ need to be lower (to still sum to 1).

Word Alignment

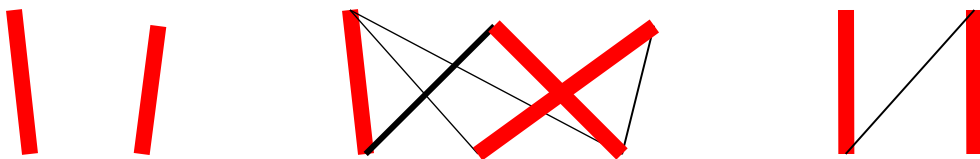
... la maison ... la maison bleue ... la fleur ...



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

Word Alignment


... la maison ... la maison bleue ... la fleur ...
... the house ... the blue house ... the flower ...



The diagram shows the alignment of words between two sentences. The French sentence is "... la maison ... la maison bleue ... la fleur ..." and the English sentence is "... the house ... the blue house ... the flower ...". Red vertical lines are placed under each word. Black lines connect aligned words: 'la' to 'the', 'maison' to 'house', 'bleue' to 'blue', and 'fleur' to 'flower'. A large red 'X' is drawn over the phrase 'la maison bleue' in the French sentence, indicating a misalignment or a specific alignment rule being demonstrated.

Word Alignment

... la maison ... la maison bleue ... la fleur ...
... the house ... the blue house ... the flower ...



What?

□ What was that?

EM

1. Start with equiprobable 1-1 word alignments
2. The $P()$ of an alignment is the product of the probability of the word alignments that make it up
3. Count the 1-1 word alignments and prorate by the $P()$ of the alignment from which they're gathered
4. Use those recomputed discounted scores to recompute the $P()$ of the alignments
5. Go to 3

MT Evaluation

Traditionally difficult because there is no single “right answer”.

20 human translators will translate the same sentence 20 different ways.

Evaluation Metric (BLEU)

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)

- Contra doesn't hold. Can find perfectly good improvements that hurt, or don't help, BLEU

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 in Action

枪手被警方击毙。

(Foreign Original)

the gunman was shot to death by the police .

(Reference Translation)

System Outputs

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	#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)

System Outputs

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	#7
al by the police .	#8
the ringer is killed by the police .	#9
police killed the gunman .	#10

NIST 2008 Results

Arabic to English (primary system) Results

Entire *Current* Evaluation Test Set

significance groups*	system	BLEU-4*	IBM BLEU	NIST	TER	METEOR
Constrained Training Track						
1	google_arabic_constrained_primary	0.4557	0.4526	10.8821	48.535	0.6857
2	IBM-UMD_arabic_constrained_primary	0.4525	0.4300	10.6183	48.436	0.6539
3	IBM_arabic_constrained_primary	0.4507	0.4276	10.5904	48.547	0.6530
3	bbn_arabic_constrained_primary	0.4340	0.4290	10.6590	49.599	0.6784

UnConstrained Training Track

17	google_arabic_unconstrained_primary	0.4772	0.4739	11.1864	46.853	0.6996
18	IBM_arabic_unconstrained_primary	0.4717	0.4527	11.0591	46.755	0.6902
19	apptek_arabic_unconstrained_primary	0.4483	0.4474	10.8420	48.263	0.7160
20	cmu-smt_arabic_unconstrained_primary	0.4312	0.4114	10.3617	50.082	0.6672

NIST 2008 Results

Chinese to English (primary system) Results

Entire *Current* Evaluation Test Set

significance groups*	system	BLEU-4*	IBM BLEU	NIST	TER	METEOR
Constrained Training Track						
1	MSR-NRC-SRI_chinese_constrained_primary	0.3089	0.2947	8.5059	58.460	0.5379
1	bbn_chinese_constrained_primary	0.3059	0.2959	8.2023	57.067	0.5468
1	isi-lw_chinese_constrained_primary	0.3041	0.2940	8.0950	57.734	0.5467
1	google_chinese_constrained_primary	0.2999	0.2887	8.5143	58.359	0.5567
2	MSR-MSRA_chinese_constrained_primary	0.2901	0.2766	8.1480	60.073	0.5171
3	SRI_chinese_constrained_primary	0.2697	0.2575	7.8942	61.622	0.5101
3	Edinburgh_chinese_constrained_primary	0.2608	0.2513	7.8117	60.654	0.5142

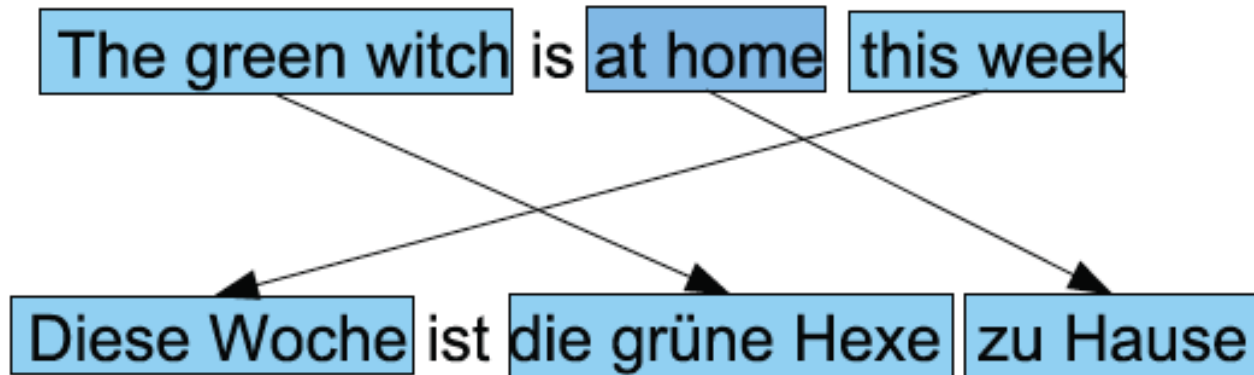
UnConstrained Training Track						
12	google_chinese_unconstrained_primary	0.3195	0.3069	8.8628	57.009	0.5707
13	cmu-smt_chinese_unconstrained_primary	0.2597	0.2474	8.0026	62.411	0.5363
14	NRC-SYSTRAN_chinese_unconstrained_primary	0.2523	0.2443	8.0473	63.002	0.5490
15	UKA_chinese_unconstrained_primary	0.2406	0.2323	7.4571	61.706	0.4916

Question

- What does the use of BLEU suggest as a strategy for the construction of these statistical systems?

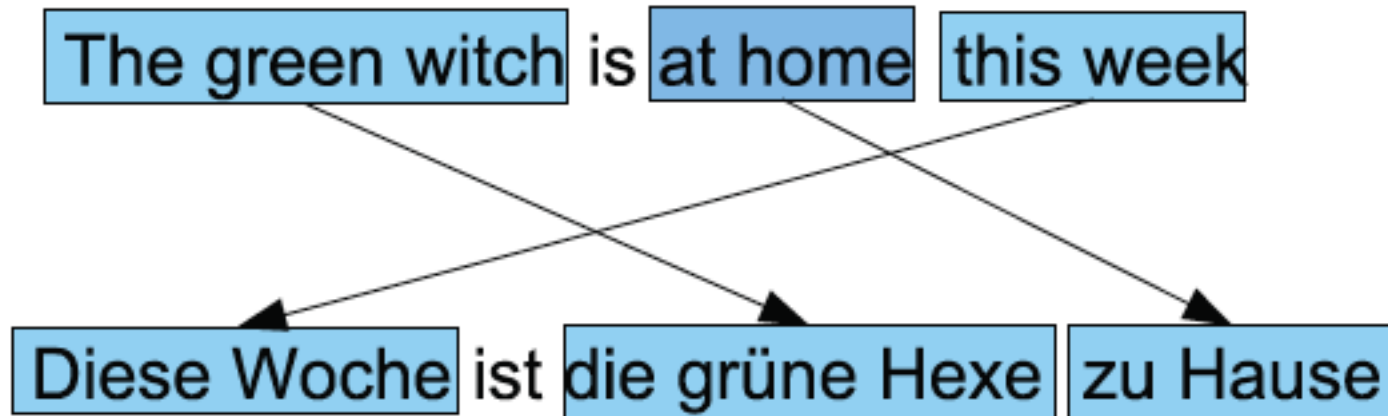
Phrase-Based Stat MT

- Turns out that word-based approaches don't work all that well.
- The basic premise in phrase-based MT is that the texts consists of phrases that
 - Need to be (1) translated
 - And (2) moved around



Phrase-Based MT

- The probability of such a translation is the product of the individual phrase translation probabilities and the the movement/dislocations probabilities.



Discovering Phrases

- Ok so now we have a good idea of which words translate to which other words.
- Now we need to use that to get phrases and phrase translation probabilities
- Lots of (ad hoc?) schemes for doing this...
- Symmetrizing alignments works by first aligning twice (E, F) and (F,E).

Discovering Phrases (1)

- Align both ways, then intersect to get high precision alignments.

Spanish to English

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

English to Spanish

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

Intersection

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

Discovering Phrases (2)

- From these high precision points, add word alignments from the union of the original alignments.

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

Discovering Phrases (3)

- These initial alignments phrase alignments can then be grown by fusing the word alignments such that...
 - Each proposed phrase alignment includes all the words in the component phrase alignments on each side (i.e. don't split adjacent alignment pairs).
 - Including words as necessary that were not in the original set.

Discovering Phrases (3)

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

Discovering Phrases (3)

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

Discovering Phrases (3)

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

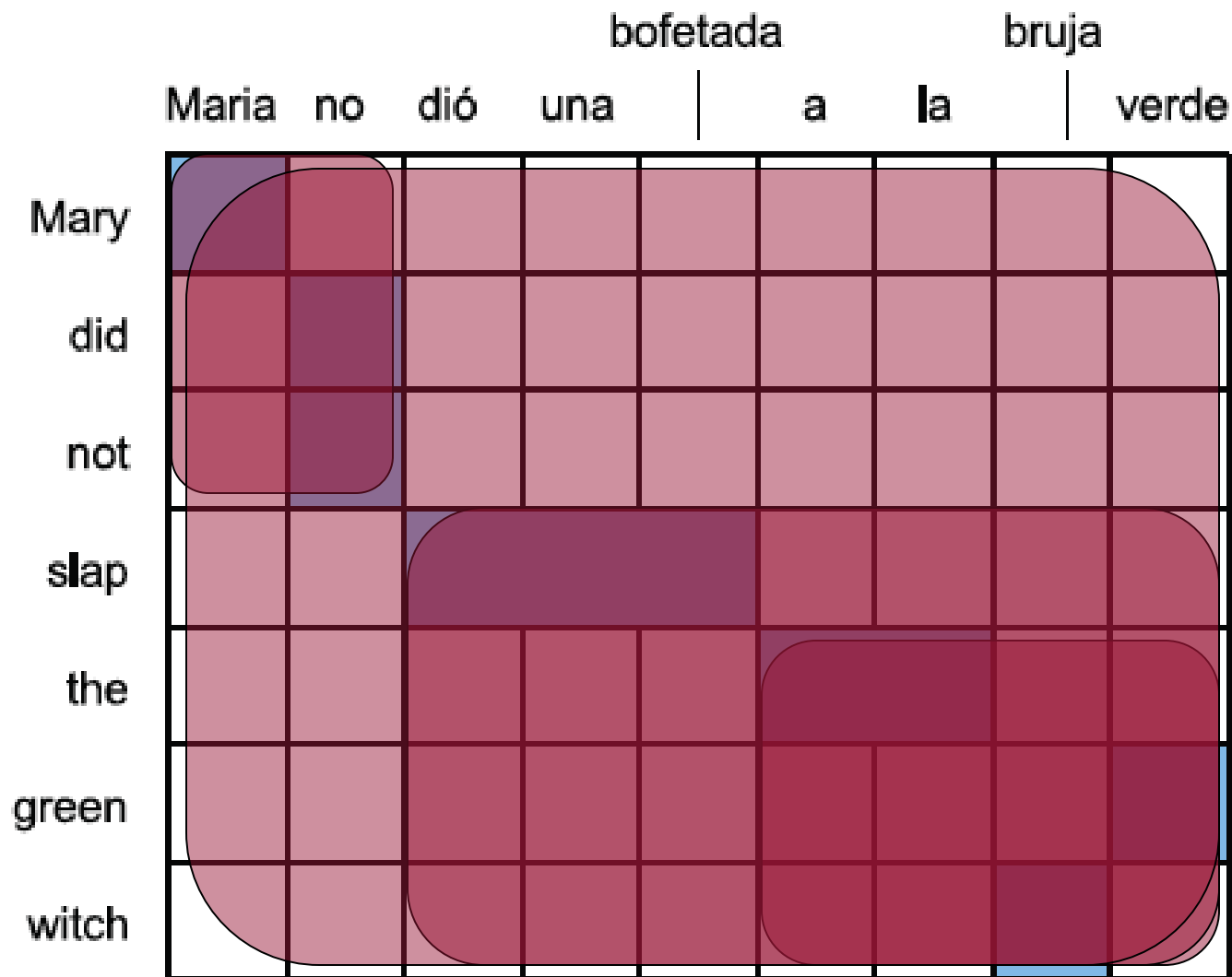
Discovering Phrases (3)

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

Discovering Phrases (3)

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

Discovering Phrases (3)



Phrase Translation

- Given such phrases we can get the required counts for our translation model from

$$\phi(\bar{f}, \bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f}, \bar{e})}$$

Next Time

- The gory details...