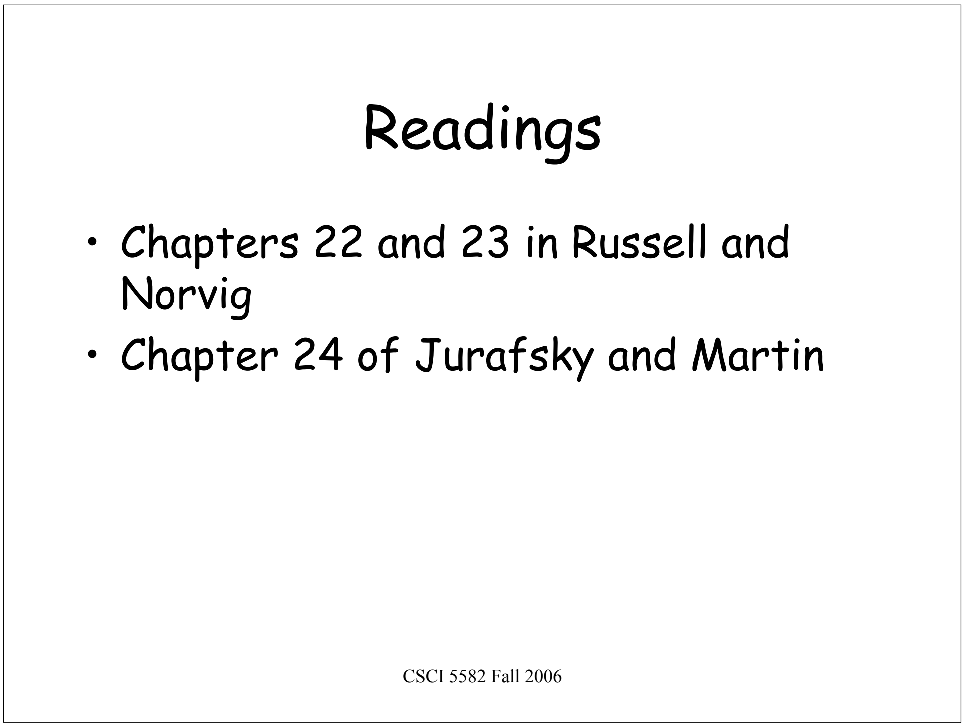
CSCI 5582
Artificial Intelligence

Lecture 24

Jim Martin

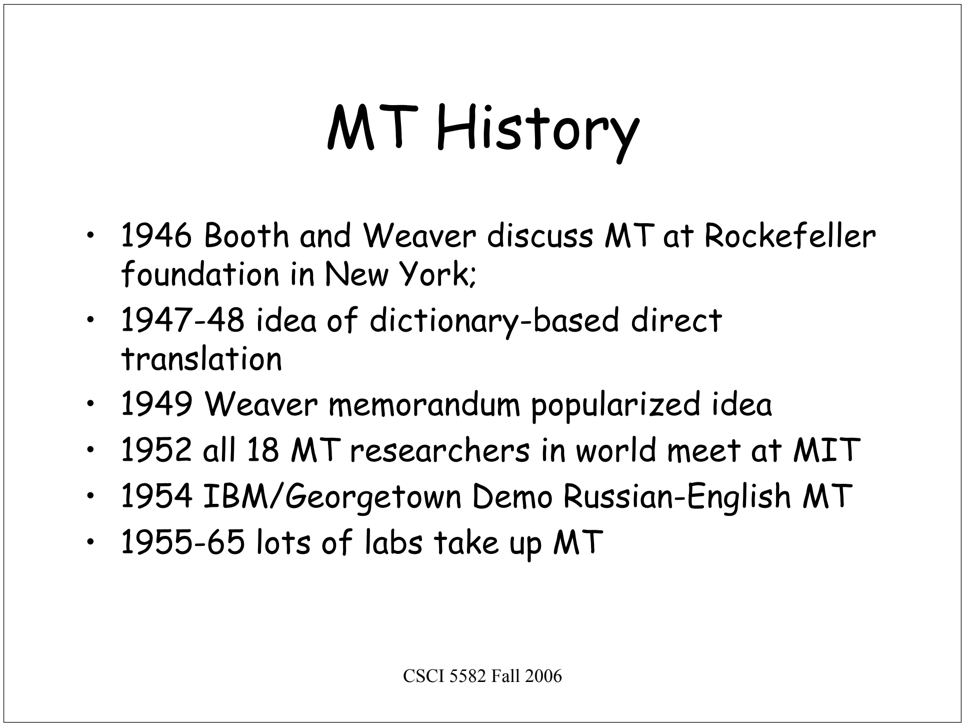
Today 12/5

- Machine Translation
  - Background
  - Why MT is hard
  - Basic Statistical MT
    - Models
    - Training
    - Decoding



Readings

- Chapters 22 and 23 in Russell and Norvig
- Chapter 24 of Jurafsky and Martin



MT History

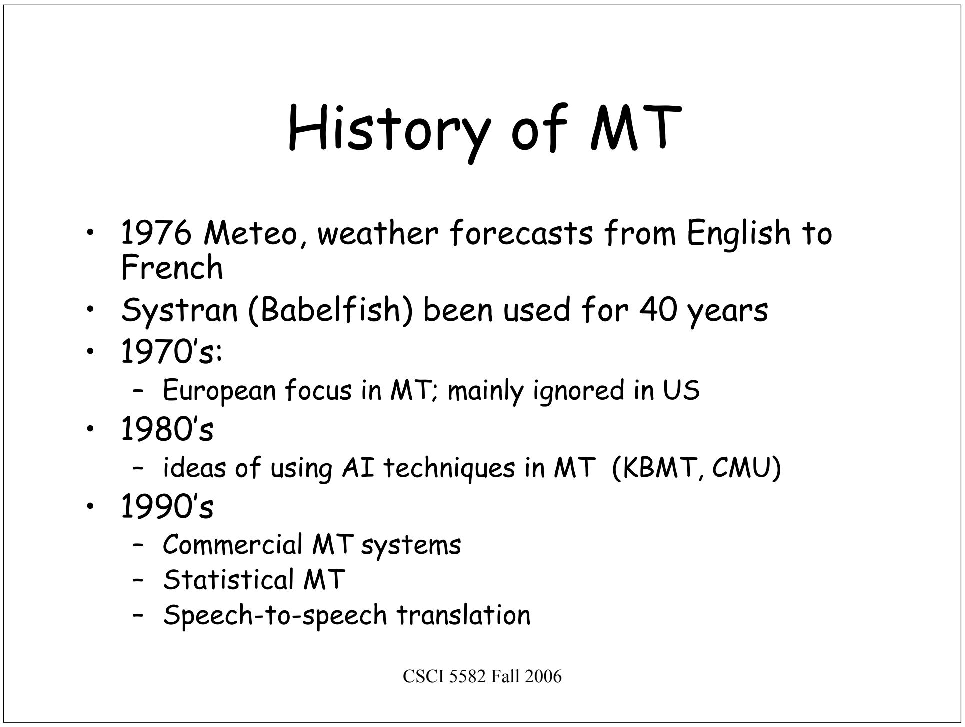
- 1946 Booth and Weaver discuss MT at Rockefeller foundation in New York;
- 1947-48 idea of dictionary-based direct translation
- 1949 Weaver memorandum popularized idea
- 1952 all 18 MT researchers in world meet at MIT
- 1954 IBM/Georgetown Demo Russian-English MT
- 1955-65 lots of labs take up MT

History of MT: Pessimism

- 1959/1960: Bar-Hillel "Report on the state of MT in US and GB"
  - Argued FAHQT too hard (semantic ambiguity, etc)
  - Should work on semi-automatic instead of automatic
  - His argument
    Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.
  - Only human knowledge let's us know that 'playpens' are bigger than boxes, but 'writing pens' are smaller
  - His claim: we would have to encode all of human knowledge

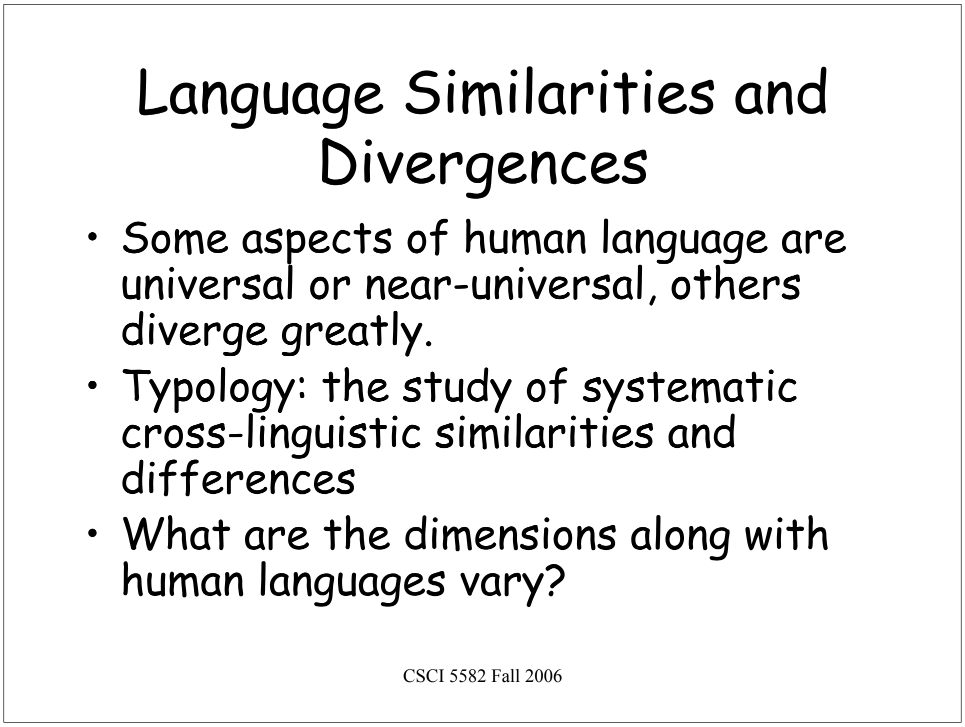
History of MT: Pessimism

- The ALPAC report
  - Headed by John R. Pierce of Bell Labs
  - Conclusions:
    - Supply of human translators exceeds demand
    - All the Soviet literature is already being translated
    - MT has been a failure: all current MT work had to be post-edited
    - Sponsored evaluations which showed that intelligibility and informativeness was worse than human translations
  - Results:
    - MT research suffered
      - Funding loss
      - Number of research labs declined
      - Association for Machine Translation and Computational Linguistics dropped MT from its name



History of MT

- 1976 Meteo, weather forecasts from English to French
- Systran (Babelfish) been used for 40 years
- 1970's:
  - European focus in MT; mainly ignored in US
- 1980's
  - ideas of using AI techniques in MT (KBMT, CMU)
- 1990's
  - Commercial MT systems
  - Statistical MT
  - Speech-to-speech translation



Language Similarities and Divergences

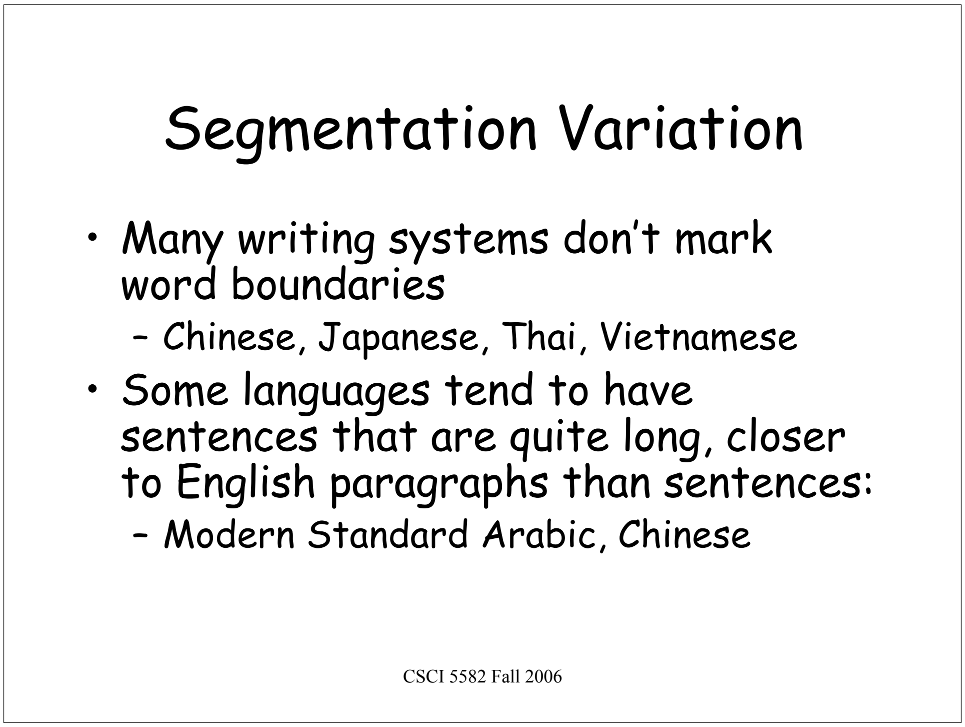
- Some aspects of human language are universal or near-universal, others diverge greatly.
- Typology: the study of systematic cross-linguistic similarities and differences
- What are the dimensions along with human languages vary?

Morphological Variation

- Isolating languages
  - Cantonese, Vietnamese: each word generally has one morpheme
- Vs. Polysynthetic languages
  - Siberian Yupik (`Eskimo'): single word may have very many morphemes
- Agglutinative languages
  - Turkish: morphemes have clean boundaries
- Vs. Fusion languages
  - Russian: single affix may have many morphemes

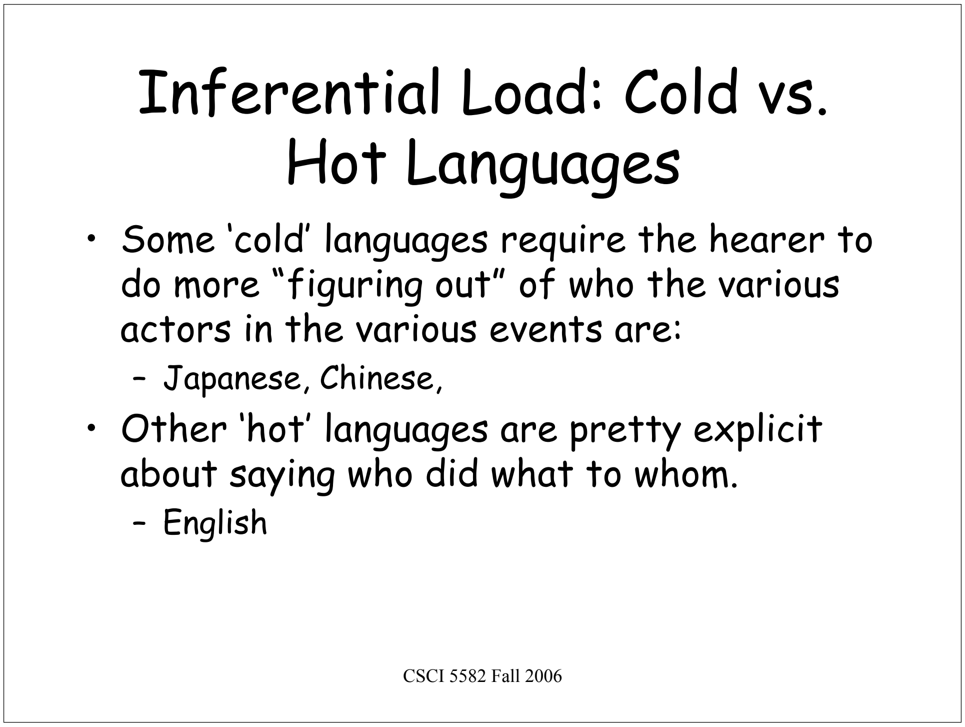
Syntactic Variation

- SVO (Subject-Verb-Object) languages
  - English, German, French, Mandarin
- SOV Languages
  - Japanese, Hindi
- VSO languages
  - Irish, Classical Arabic
- Regularities
  - SVO languages generally have prepositions
  - VSO languages generally have postpositions



Segmentation Variation

- Many writing systems don’t mark word boundaries
  - Chinese, Japanese, Thai, Vietnamese
- Some languages tend to have sentences that are quite long, closer to English paragraphs than sentences:
  - Modern Standard Arabic, Chinese

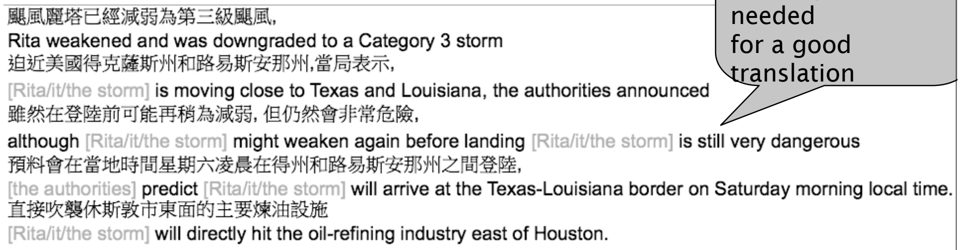


Inferential Load: Cold vs. Hot Languages

- Some ‘cold’ languages require the hearer to do more “figuring out” of who the various actors in the various events are:
  - Japanese, Chinese,
- Other ‘hot’ languages are pretty explicit about saying who did what to whom.
  - English

Inferential Load (2)

Noun phrases in blue do not appear in Chinese text … But they are needed for a good translation



颶風麗塔已經減弱為第三級颶風,
Rita weakened and was downgraded to a Category 3 storm
迫近美國得克薩斯州和路易斯安那州,當局表示,
[Rita/it/the storm] is moving close to Texas and Louisiana, the authorities announced
雖然在登陸前可能再稍為減弱, 但仍然會非常危險,
although [Rita/it/the storm] might weaken again before landing [Rita/it/the storm] is still very dangerous
預料會在當地時間星期六凌晨在得州和路易斯安那州之間登陸,
[the authorities] predict [Rita/it/the storm] will arrive at the Texas-Louisiana border on Saturday morning local time.
直接吹襲休斯敦市東面的主要煉油設施
[Rita/it/the storm] will directly hit the oil-refining industry east of Houston.

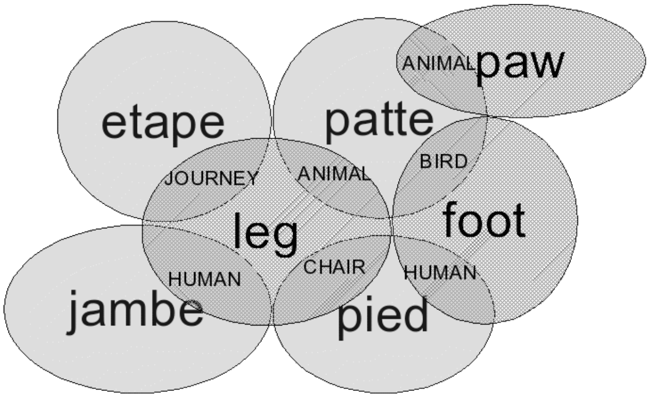
Lexical Divergences

- Word to phrases:
  - English “computer science” = French “informatique”
- POS divergences
  - Eng. ‘she likes/VERB to sing’
  - Ger. Sie singt gerne/ADV
  - Eng ‘I’m hungry/ADJ
  - Sp. ‘tengo hambre/NOUN

Lexical Divergences: Specificity

- Grammatical constraints
  - English has gender on pronouns, Mandarin not.
    - So translating “3rd person” from Chinese to English, need to figure out gender of the person!
    - Similarly from English “they” to French “ils/elles”
- Semantic constraints
  - English ` brother’
  - Mandarin ‘gege’ (older) versus ‘didi’ (younger)
  - English ‘wall’
  - German ‘Wand’ (inside) ‘Mauer’ (outside)
  - German ‘Berg’
  - English ‘hill’ or ‘mountain’

Lexical Divergence: many-to-many

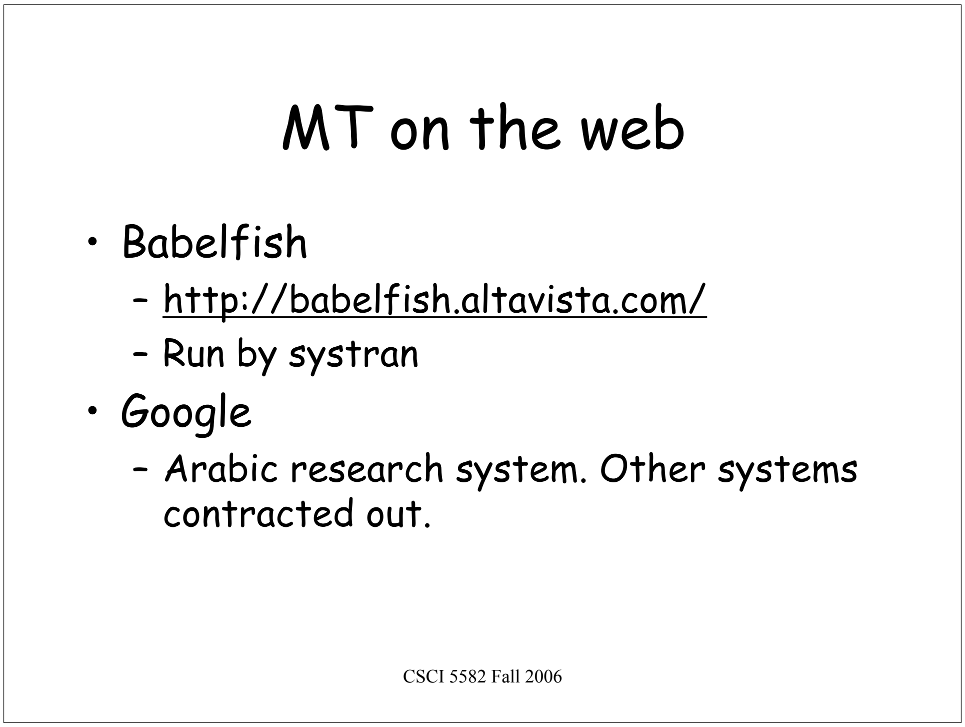


Lexical Divergence: Lexical Gaps

- Japanese: no word for privacy
- English: no word for Cantonese ‘haauseun’ or Japanese ‘oyakoko’ (something like `filial piety’)
- English ‘cow’ versus ‘beef’, Cantonese ‘ngau’

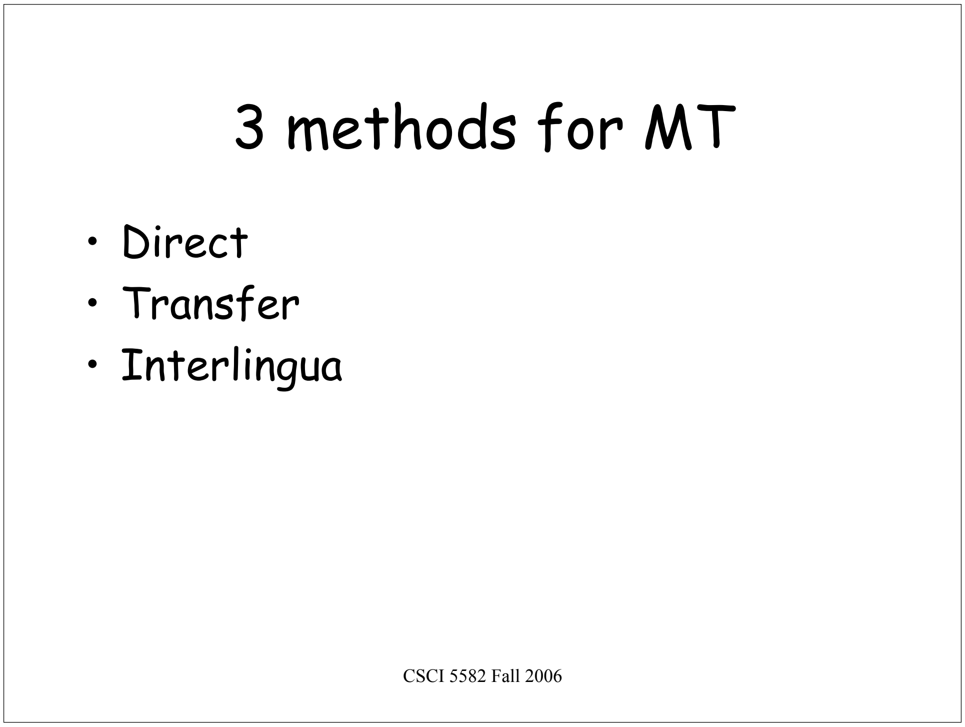
Event-to-argument divergences

- English
  - The bottle floated out.
- Spanish
  - La botella salió flotando.
  - The bottle exited floating
- Verb-framed lg: mark direction of motion on verb
  - Spanish, French, Arabic, Hebrew, Japanese, Tamil, Polynesian, Mayan, Bantu familiies
- Satellite-framed lg: mark direction of motion on satellite
  - Crawl out, float off, jump down, walk over to, run after
  - Rest of Indo-European, Hungarian, Finnish, Chinese



MT on the web

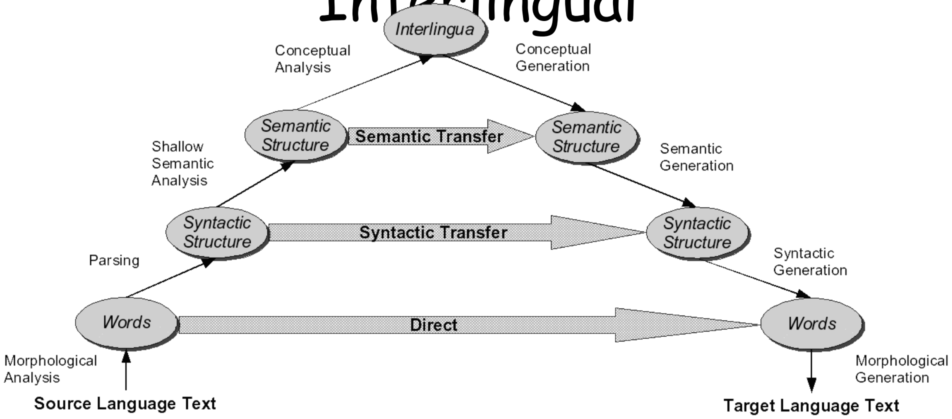
- Babelfish
  - http://babelfish.altavista.com/
  - Run by systran
- Google
  - Arabic research system. Other systems contracted out.



3 methods for MT

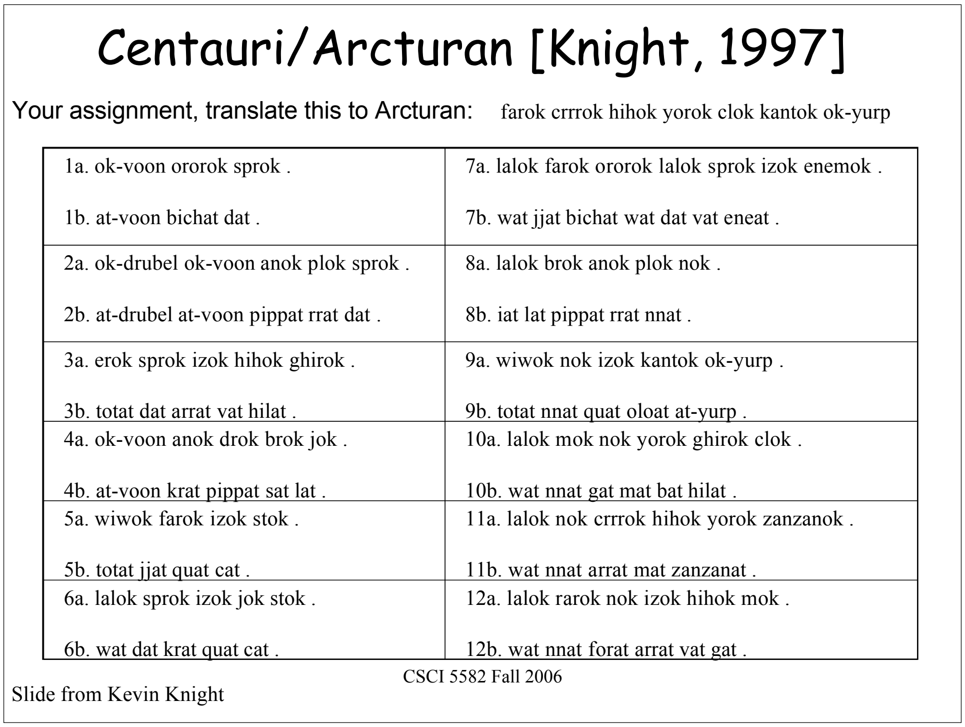
- Direct
- Transfer
- Interlingua

Three MT Approaches: Direct, Transfer, Interlingual



Centauri/Arcturan [Knight, 1997]

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

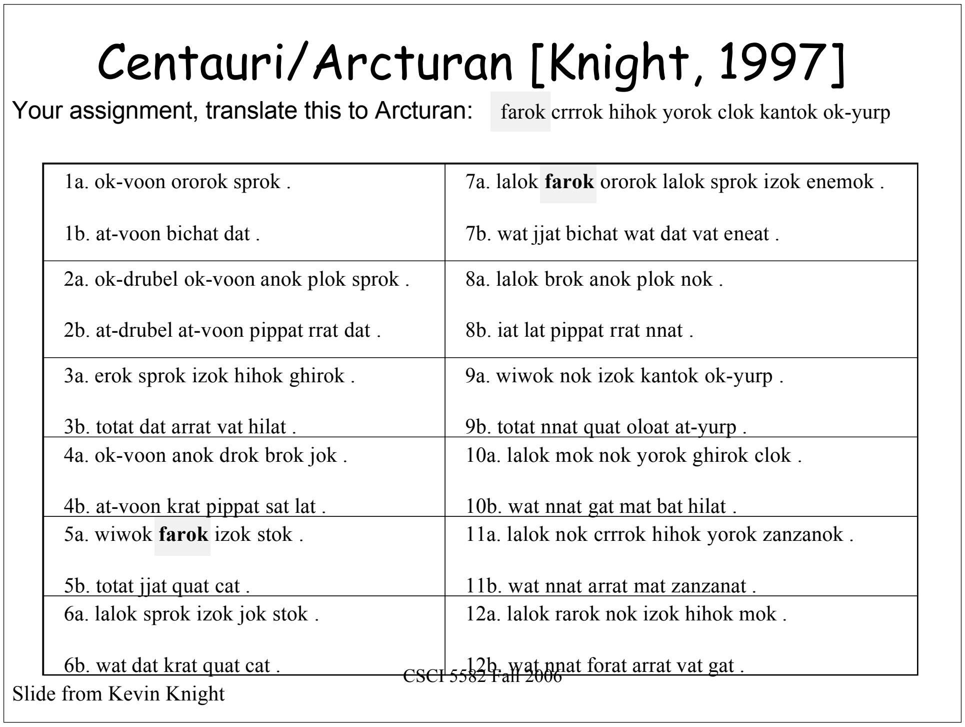


Centauri/Arcturan [Knight, 1997]

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 .

Slide from Kevin Knight

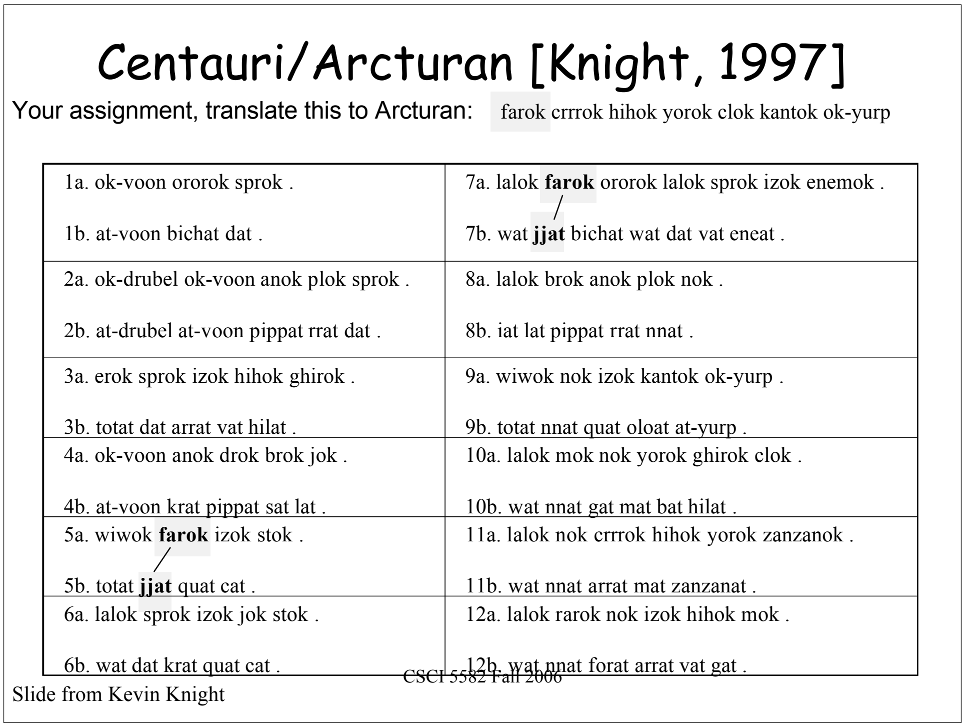


Centauri/Arcturan [Knight, 1997]

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 .

Slide from Kevin Knight



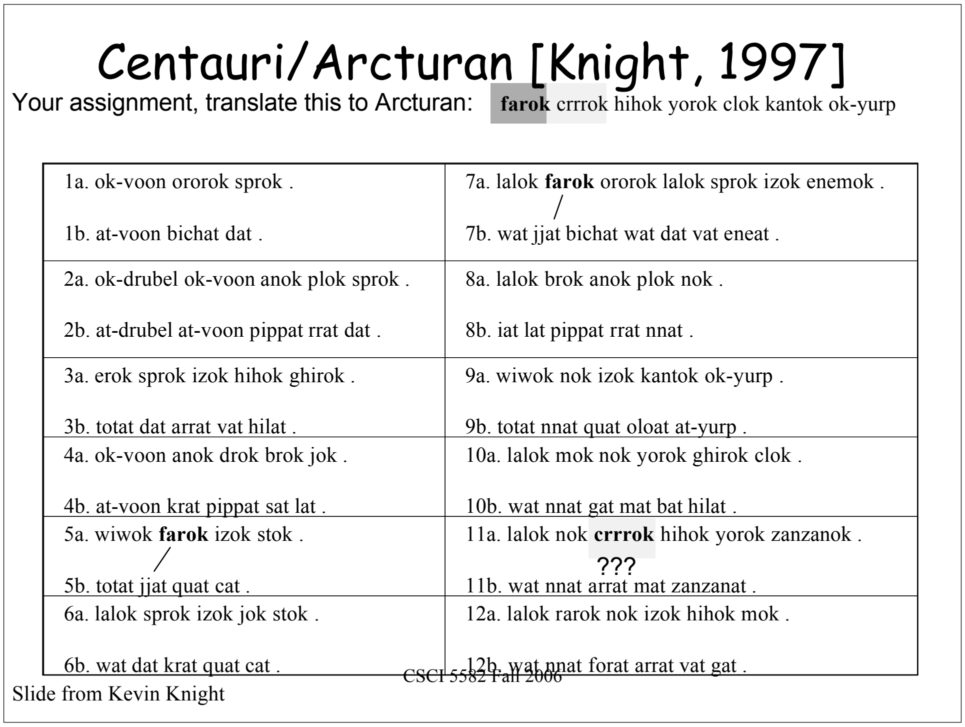
Centauri/Arcturan [Knight, 1997]

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 .

CSCI 5582 Fall 2006

Slide from Kevin Knight



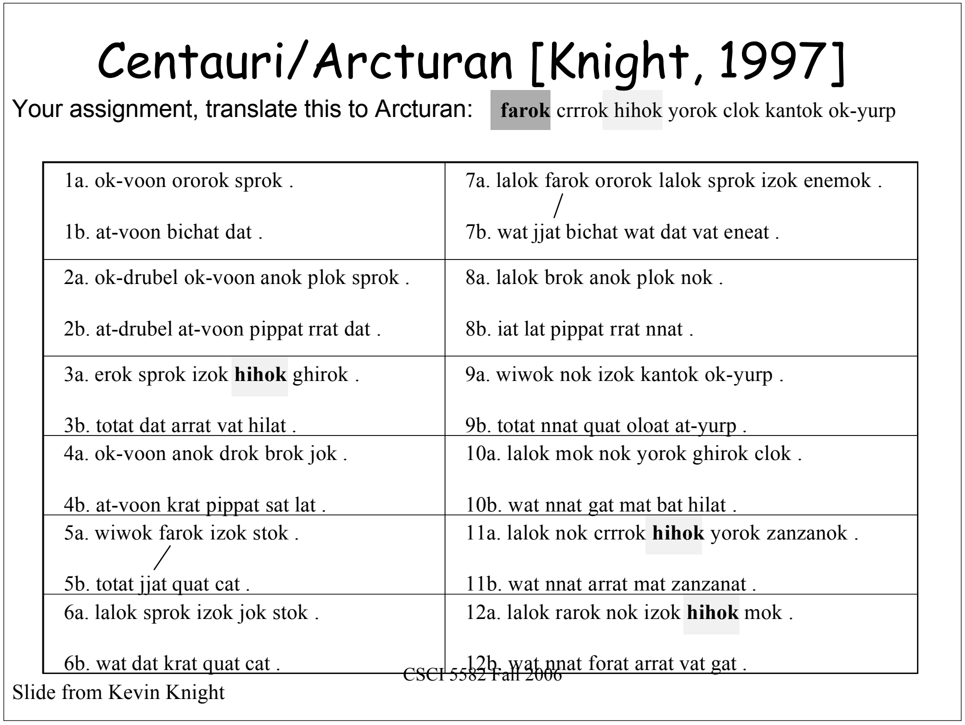
Centauri/Arcturan [Knight, 1997]

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 .

CSCI 5582 Fall 2006

Slide from Kevin Knight

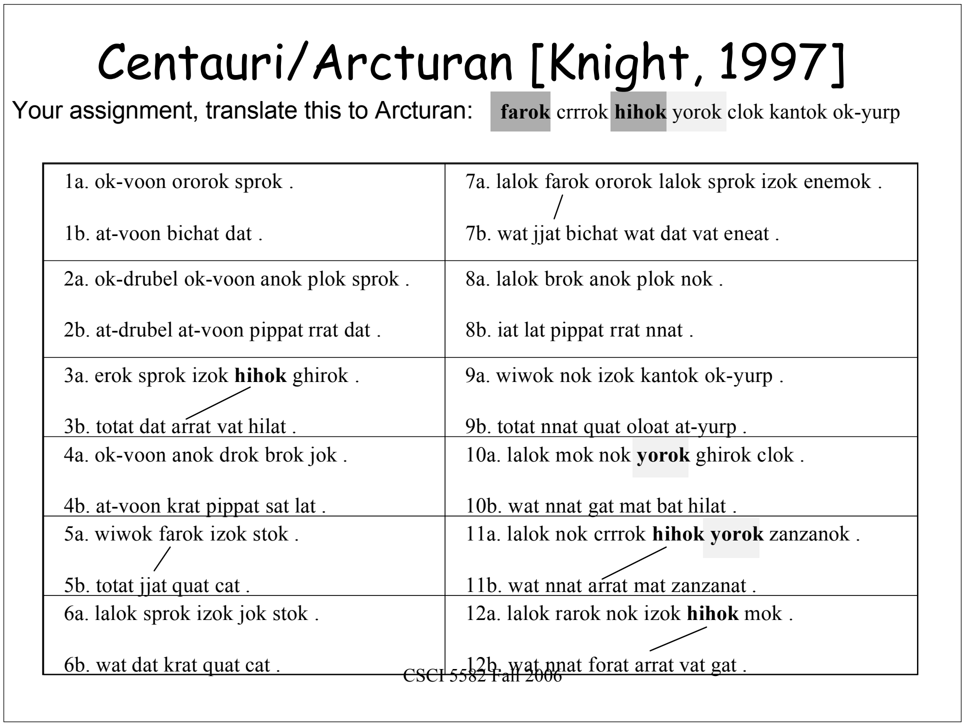


Centauri/Arcturan [Knight, 1997]

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 .

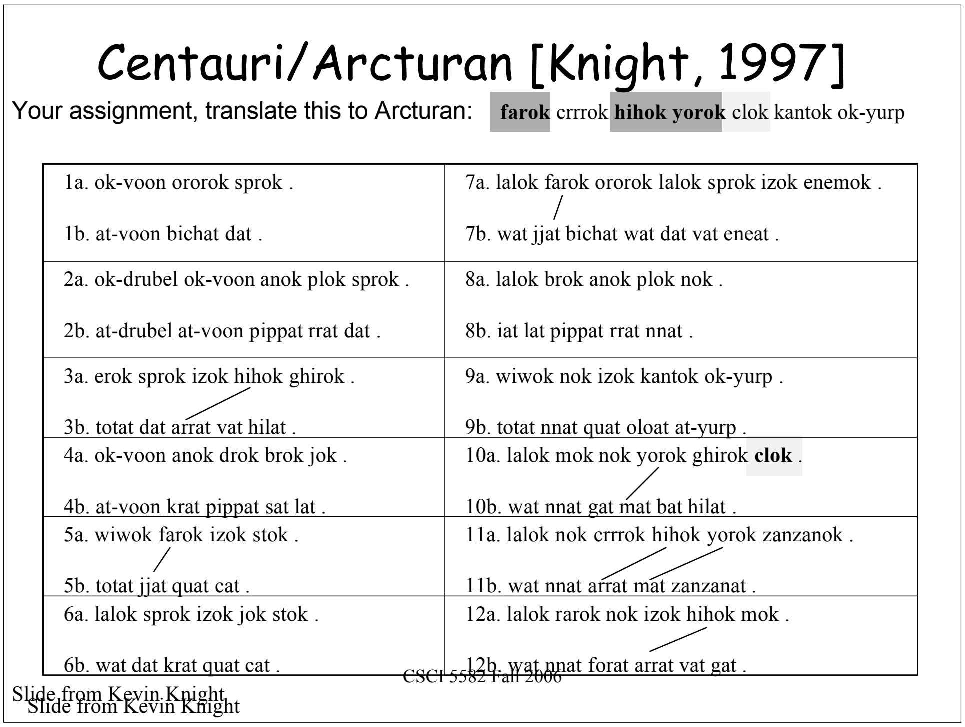
Slide from Kevin Knight



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 .

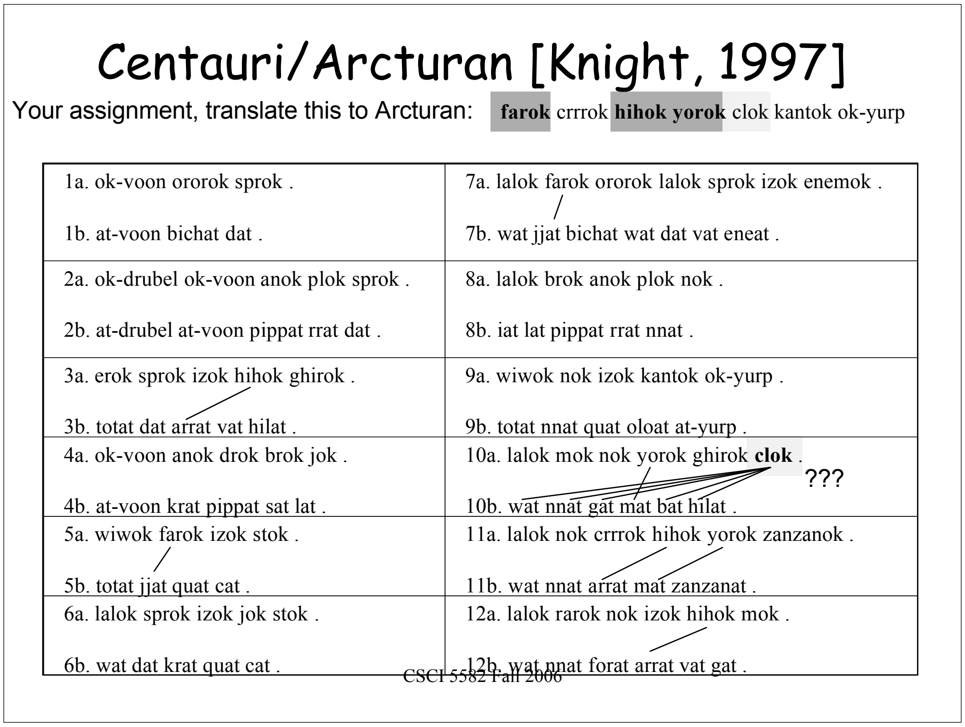


Centauri/Arcturan [Knight, 1997]

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 .

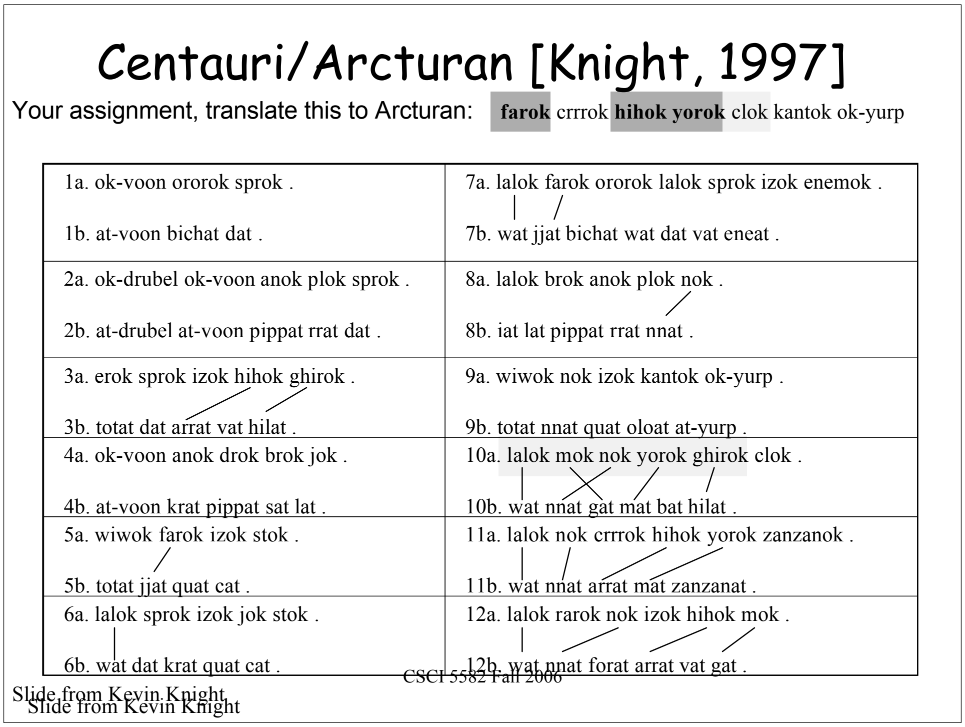
Slide from Kevin Knight
Slide from Kevin Knight



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 .

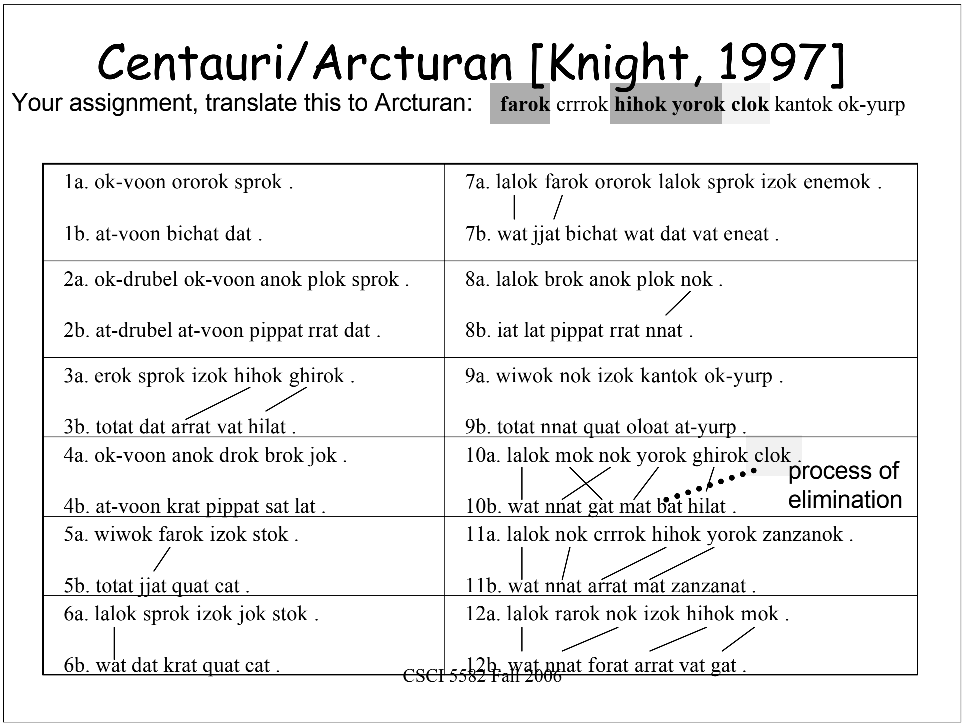


Centauri/Arcturan [Knight, 1997]

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 .

Slide from Kevin Knight
Slide from Kevin Knight

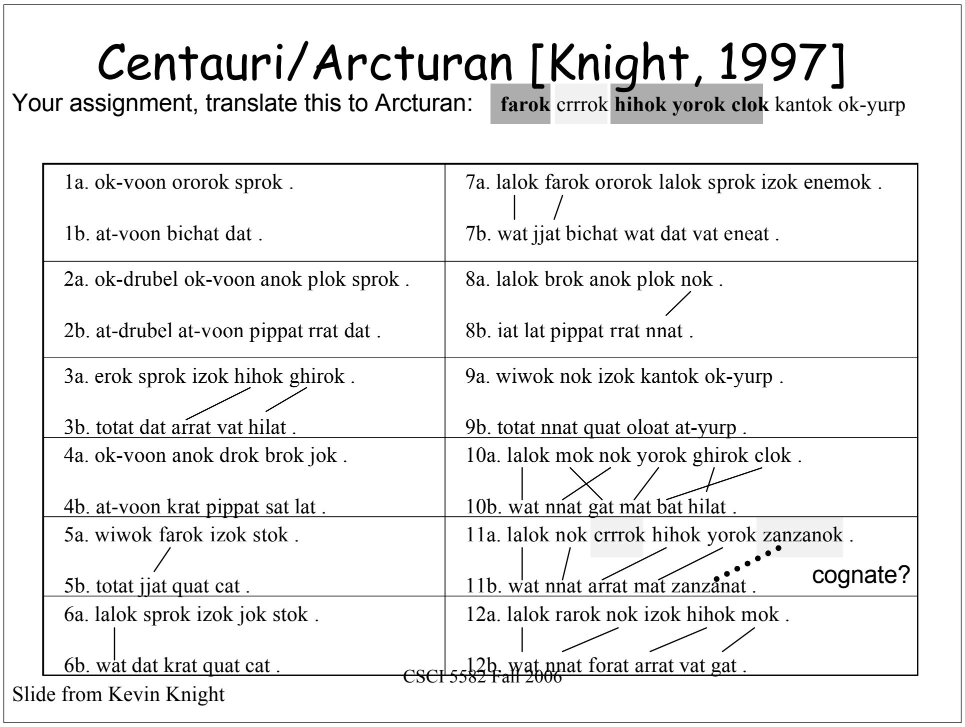


Centauri/Arcturan [Knight, 1997]

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]

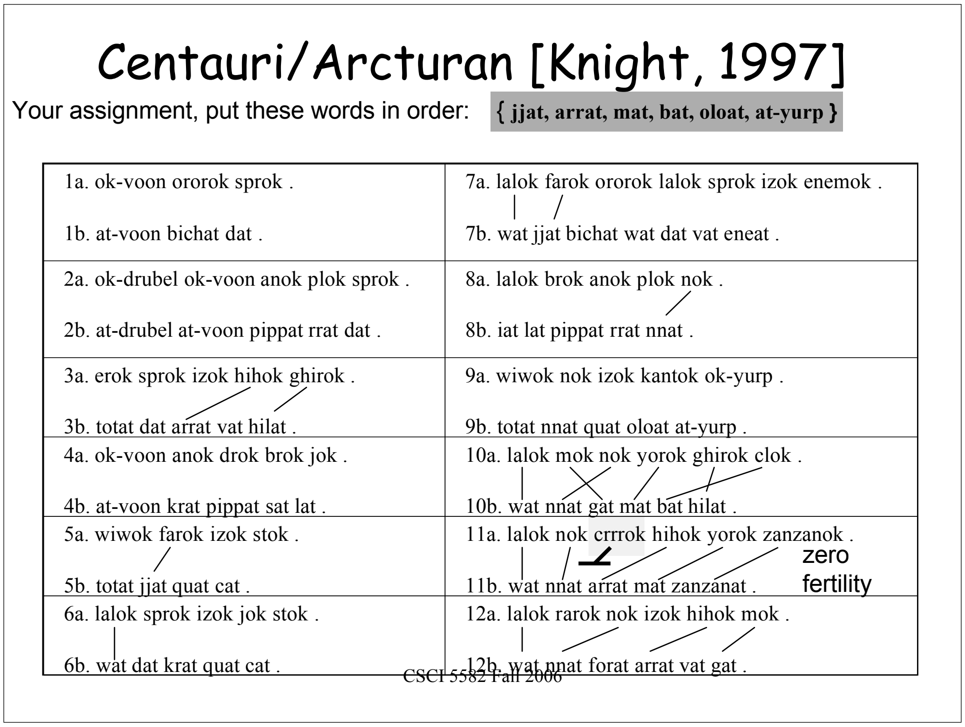
Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

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 .
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 .

cognate?

CSCI 5582 Fall 2006

Slide from Kevin Knight



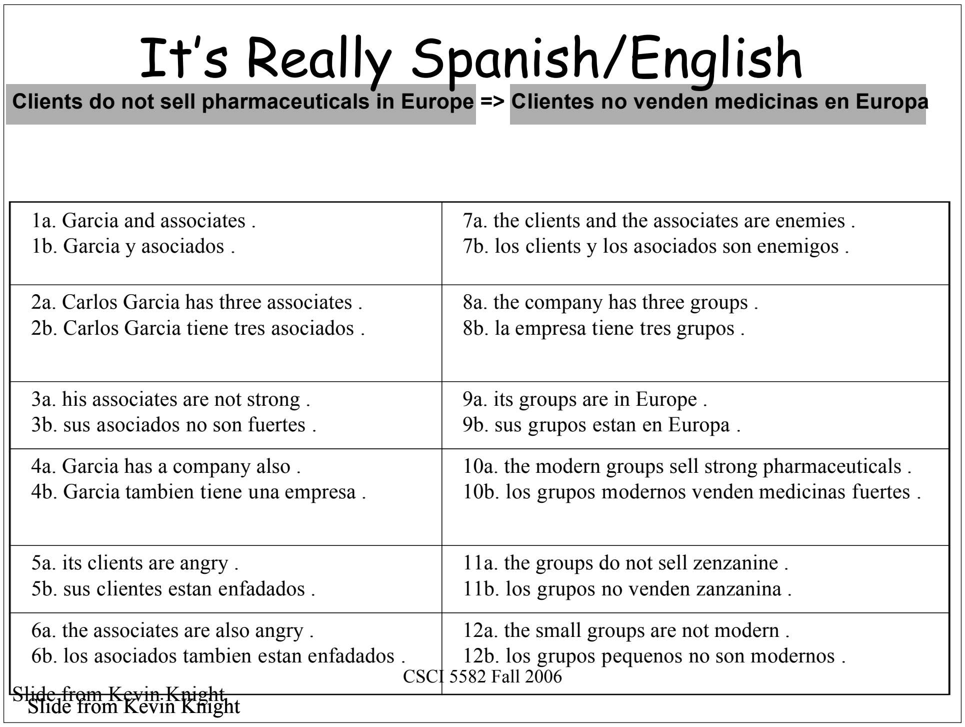
Centauri/Arcturan [Knight, 1997]

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 .
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 .

zero fertility

CSCI 5582 Fall 2006



It’s Really Spanish/English

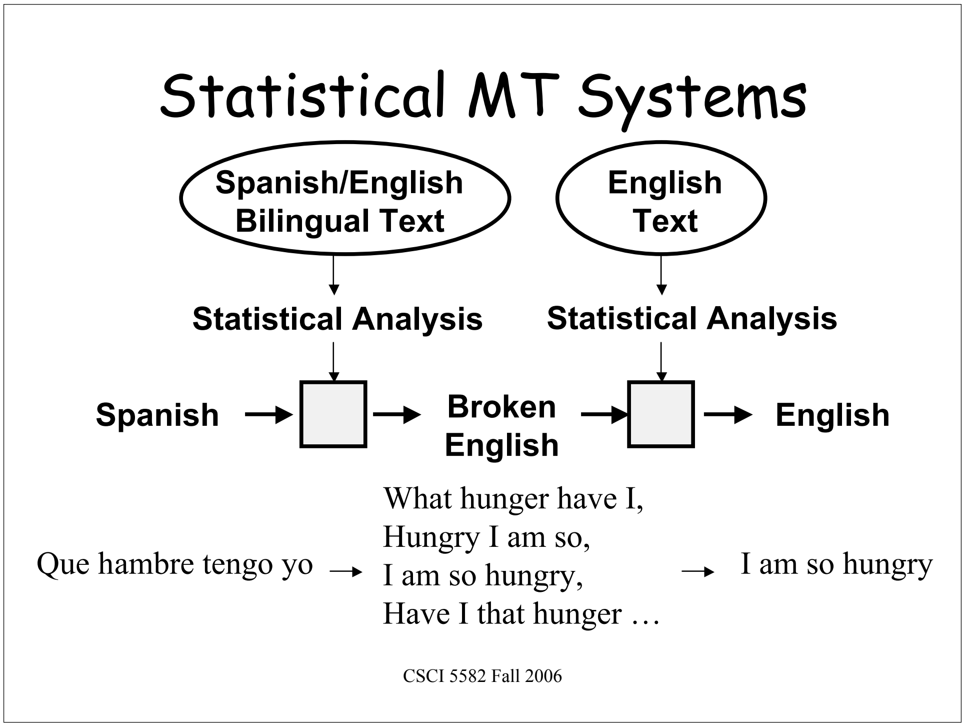
Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

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 .

Slide from Kevin Knight

Slide from Kevin Knight

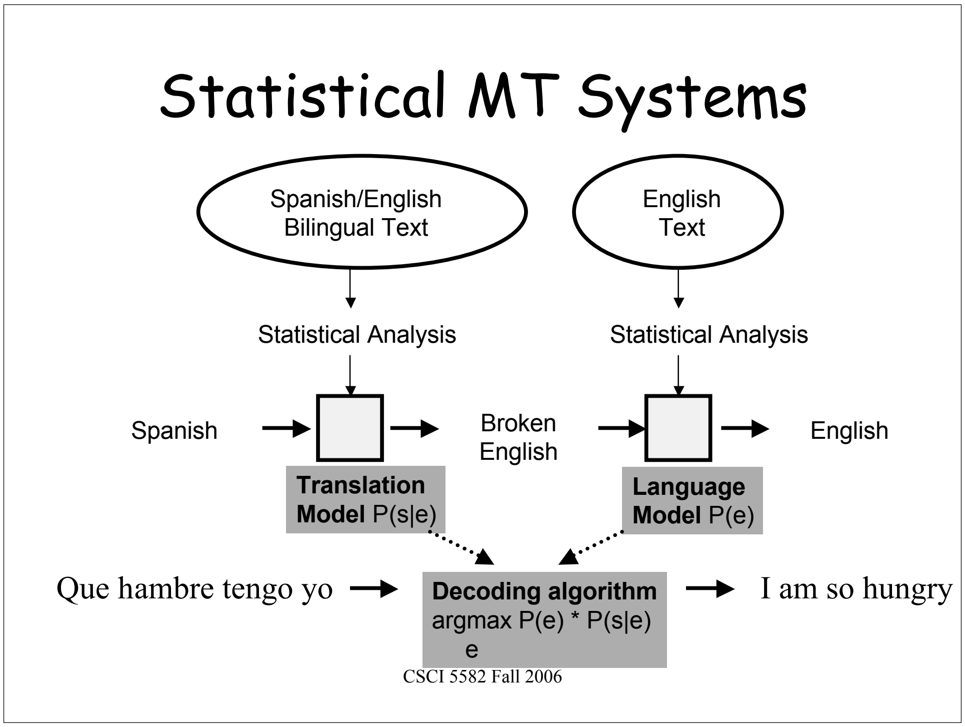
CSCI 5582 Fall 2006



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

CSCI 5582 Fall 2006



Statistical MT Systems

Spanish/English Bilingual Text

English Text

Statistical Analysis

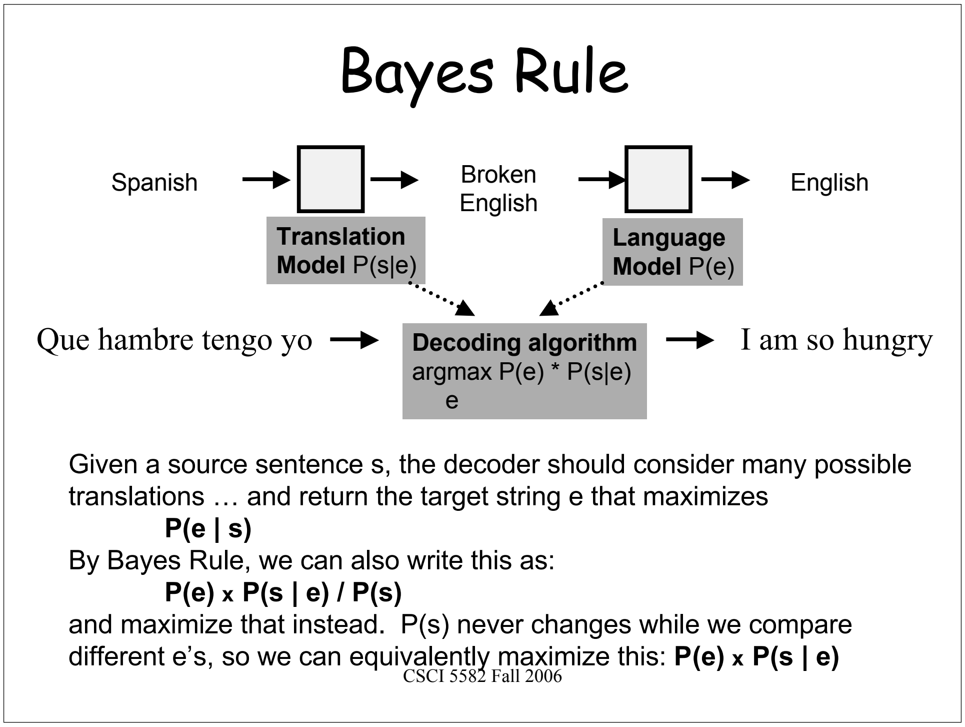
Statistical Analysis

Spanish → → Broken English → → English

Translation Model P(s|e)

Language Model P(e)

Que hambre tengo yo → **Decoding algorithm** argmax P(e) * P(s|e) e → I am so hungry



Bayes Rule

Spanish → → Broken English → → English

Translation Model P(s|e)

Language Model P(e)

Que hambre tengo yo → **Decoding algorithm** argmax P(e) * P(s|e) e → I am so hungry

Given a source sentence s, the decoder should consider many possible translations … and return the target string e that maximizes

P(e | s)

By Bayes Rule, we can also write this as:

P(e) x P(s | e) / P(s)

and maximize that instead. P(s) never changes while we compare different e’s, so we can equivalently maximize this: **P(e) x P(s | e)**

Four Problems for Statistical MT

- Language model
  - Given an English string e, assigns P(e) by the usual methods we've been using sequence modeling.
- Translation model
  - Given a pair of strings <f,e>, assigns P(f | e) again by making the usual markov assumptions
- Training
  - Getting the numbers needed for the models
- Decoding algorithm
  - Given a language model, a translation model, and a new sentence f … find translation e maximizing

    P(e) * P(f | e)

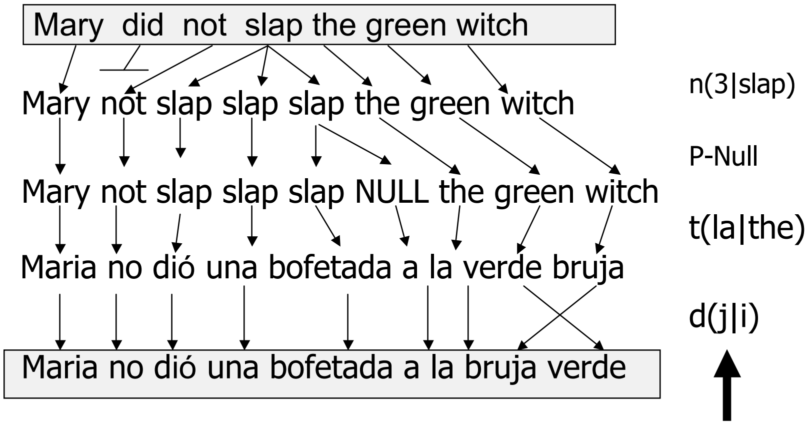
3 Models

- IBM Model 1
  - Dumb word to word
- IBM Model 3
  - Handles deletions, insertions and 1-to-N translations
- Phrase-Based Models (Google/ISI)
  - Basically Model 1 with phrases instead of words

IBM Model 3

Brown et al., 1993

Generative approach:



Mary did not slap the green witch

Mary not slap slap slap the green witch

Mary not slap slap slap NULL the green witch

Maria no dió una bofetada a la verde bruja

Maria no dió una bofetada a la bruja verde

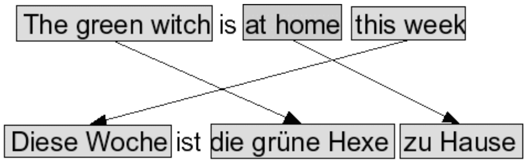
n(3|slap)

P-Null

t(la|the)

d(j|i)

Phrase-based translation



The green witch is at home this week

Diese Woche ist die grüne Hexe zu Hause

- Generative story here has three steps
  1) Discover and align phrases during training
  2) Align and translate phrases during decoding
  3) Finally move the phrases around

Alignment Probabilities

- Recall what of all of the models are doing

  Argmax P(e|f) = P(f|e)P(e)

  In the simplest models P(f|e) is just direct word-to-word translation probs. So let's start with how to get those, since they're used directly or indirectly in all the models.

Training alignment probabilities

- Step 1: Get a parallel corpus
  - Hansards
    - Canadian parliamentary proceedings, in French and English
    - Hong Kong Hansards: English and Chinese
- Step 2: Align sentences
- Step 3: Use EM to train word alignments. Word alignments give us the counts we need for the word to word P(f|e) probs

Step 2: Sentence Alignment

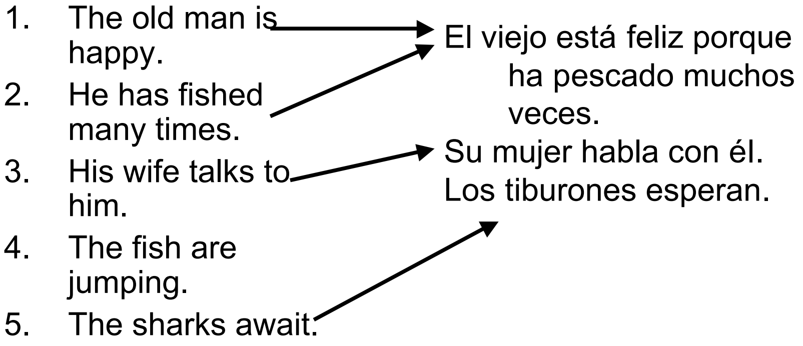
The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

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

Intuition:

- use length in words or chars
- together with dynamic programming
- or use a simpler MT model

Sentence Alignment



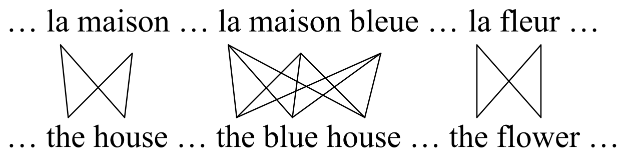
1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

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

Step 3: Word Alignments

- Of course, sentence alignments aren't what we need. We need word alignments to get the stats we need.
- It turns out we can bootstrap word alignments from raw sentence aligned data (no dictionaries)
- Using EM
  - Recall the basic idea of EM. A model predicts the way the world should look. We have raw data about how the world looks. Start somewhere and adjust the numbers so that the model is doing a better job of predicting how the world looks.

EM Training: Word Alignment Probs



… la maison … la maison bleue … la fleur …

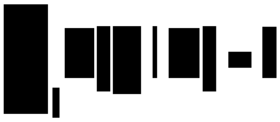
… the house … the blue house … the flower …

All word alignments equally likely

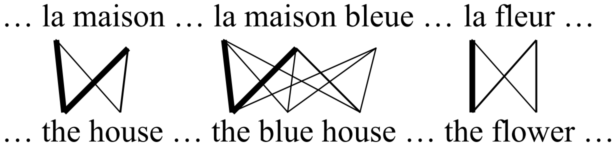
All P(french-word | english-word) equally likely.

EM Training Constraint

- Recall what we’re doing here… Each English word has to translate to some french word.
- But its still true that



EM for training alignment probs

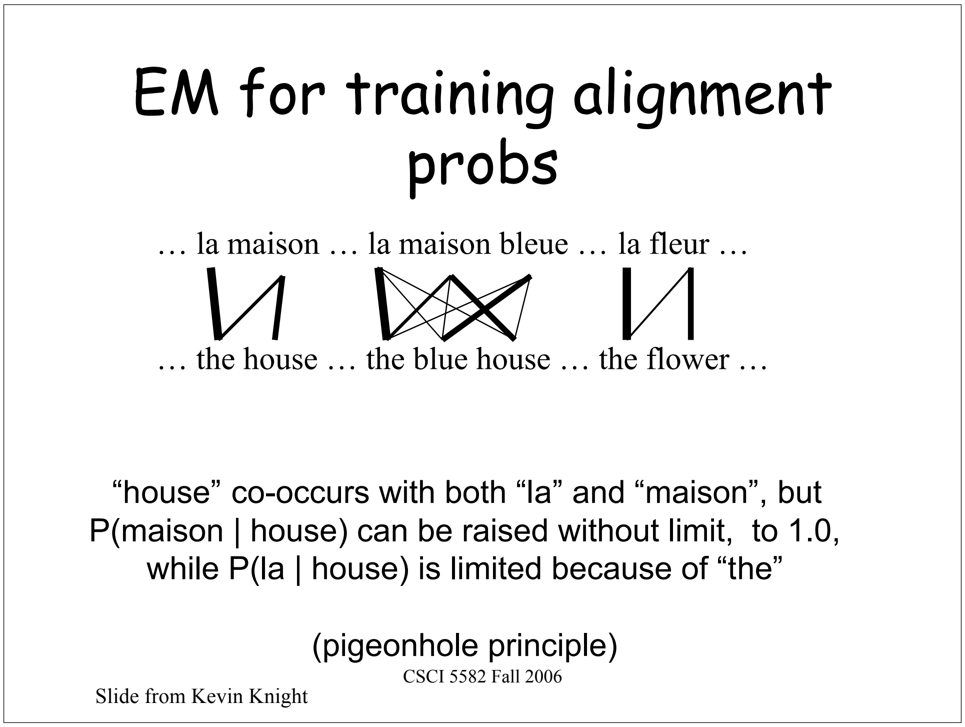


… la maison … la maison bleue … la fleur …

… the house … the blue house … the flower …

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

Slide from Kevin Knight



EM for training alignment probs

… la maison … la maison bleue … la fleur …

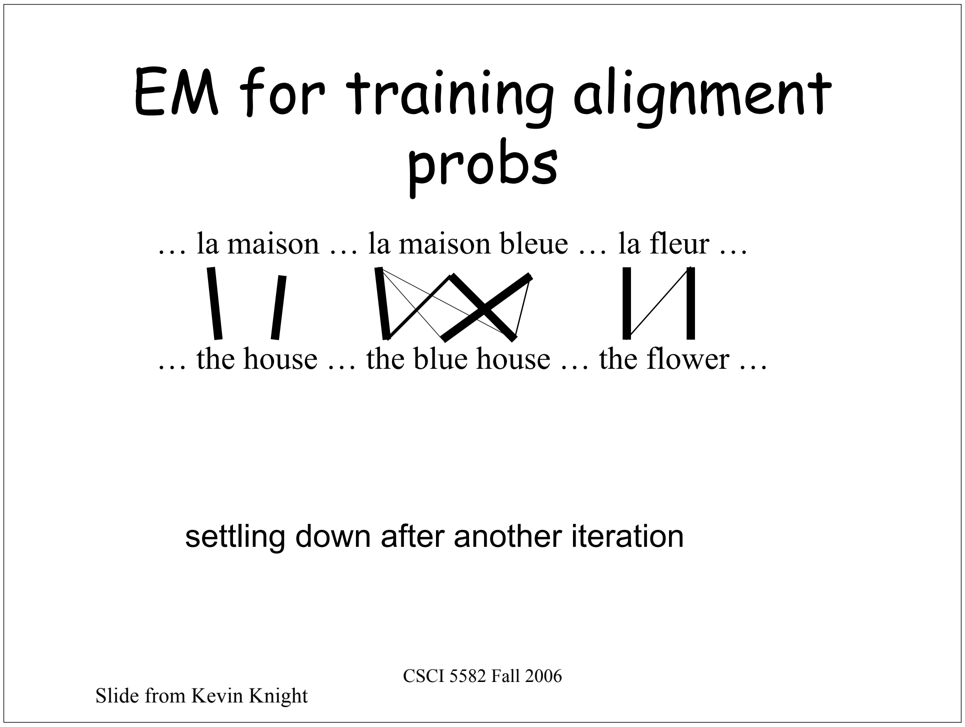
… the house … the blue house … the flower …

"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)

CSCI 5582 Fall 2006

Slide from Kevin Knight



EM for training alignment probs

… la maison … la maison bleue … la fleur …

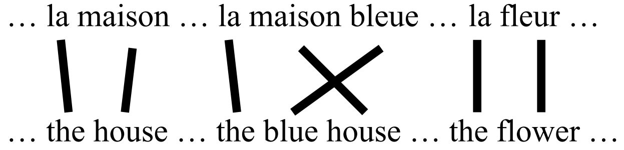
… the house … the blue house … the flower …

settling down after another iteration

CSCI 5582 Fall 2006

Slide from Kevin Knight

EM for training alignment probs



… la maison … la maison bleue … la fleur …

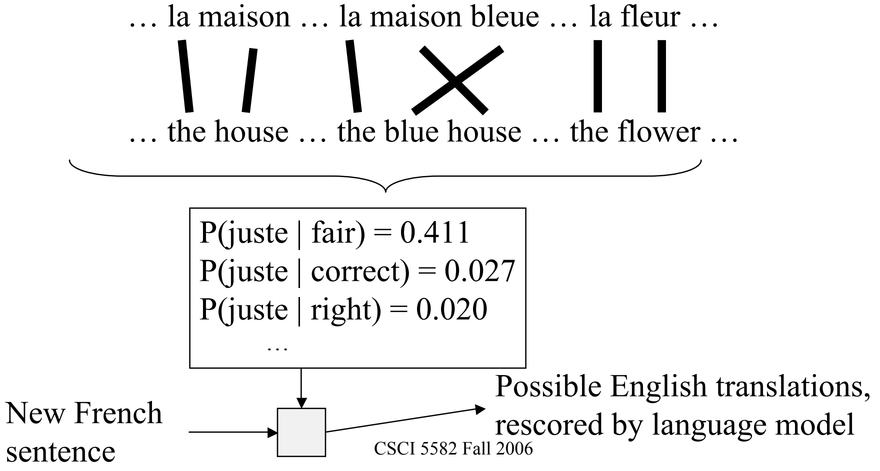
… the house … the blue house … the flower …

Inherent hidden structure revealed by EM training!
For details, see:

- Section 24.6.1 in the chapter
- "A Statistical MT Tutorial Workbook" (Knight, 1999).
- "The Mathematics of Statistical Machine Translation" (Brown et al, 1993)
- Free Alignment Software: GIZA++

Slide from Kevin Knight

Direct Translation



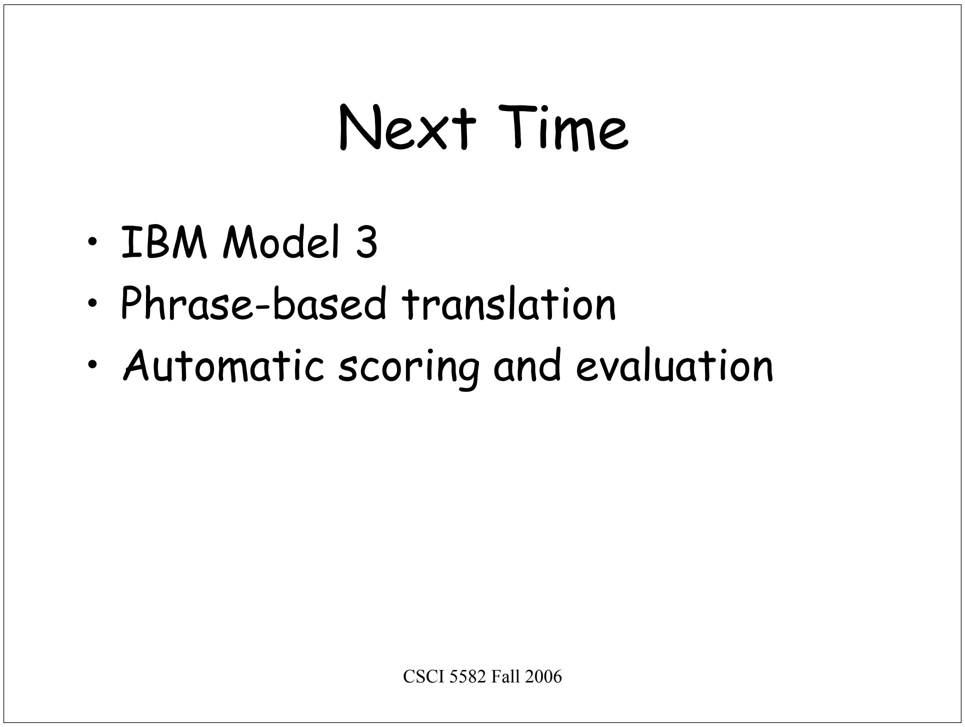
… la maison … la maison bleue … la fleur …

… 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, rescored by language model



Next Time

- IBM Model 3
- Phrase-based translation
- Automatic scoring and evaluation