Language Stuff

Hal Daumé III

Computer Science
University of Maryland

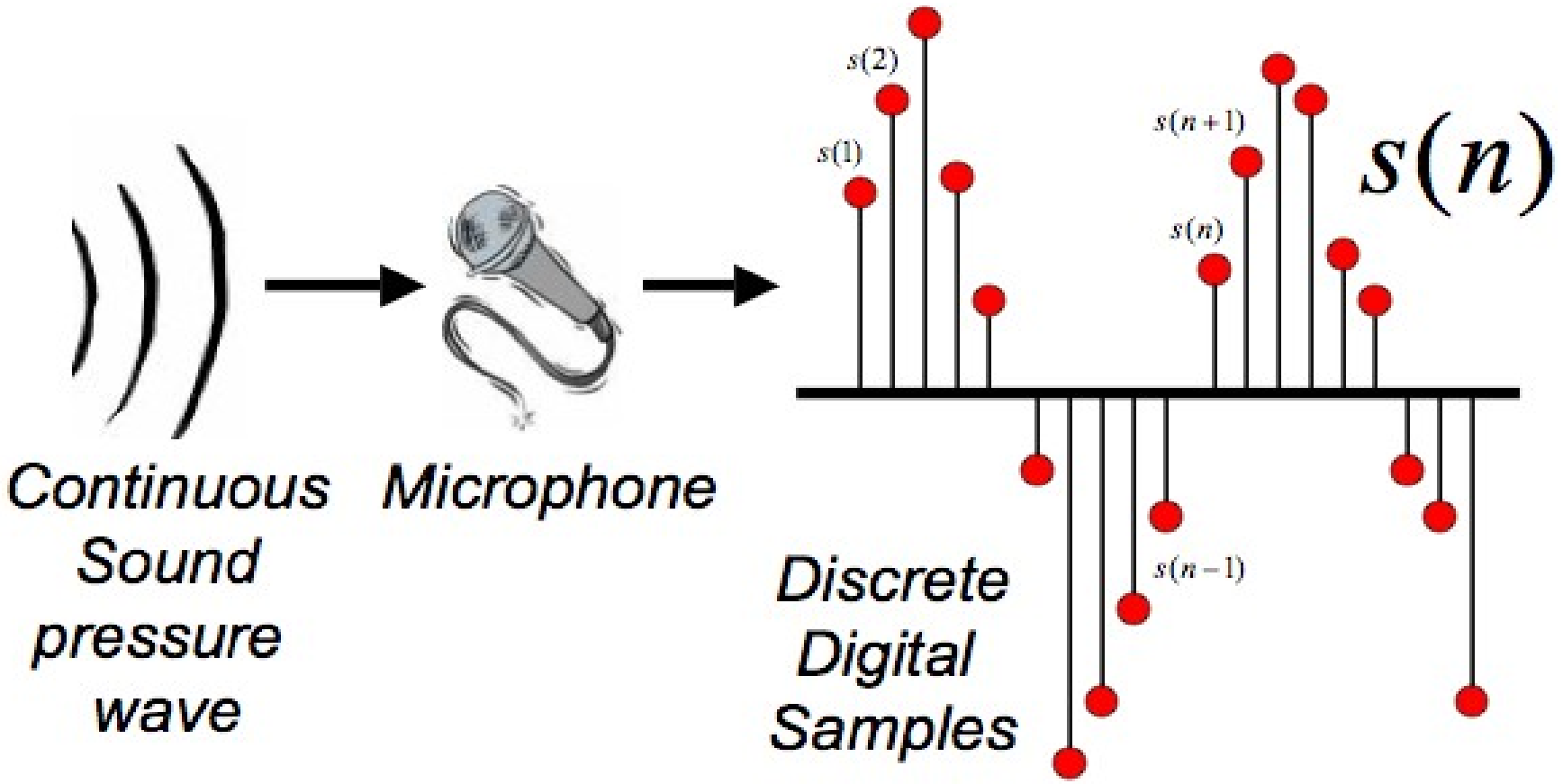
me@hal3.name

CS 421: Introduction to Artificial Intelligence

26 Apr 2012

Many slides courtesy of
Dan Klein, Stuart Russell,
or Andrew Moore

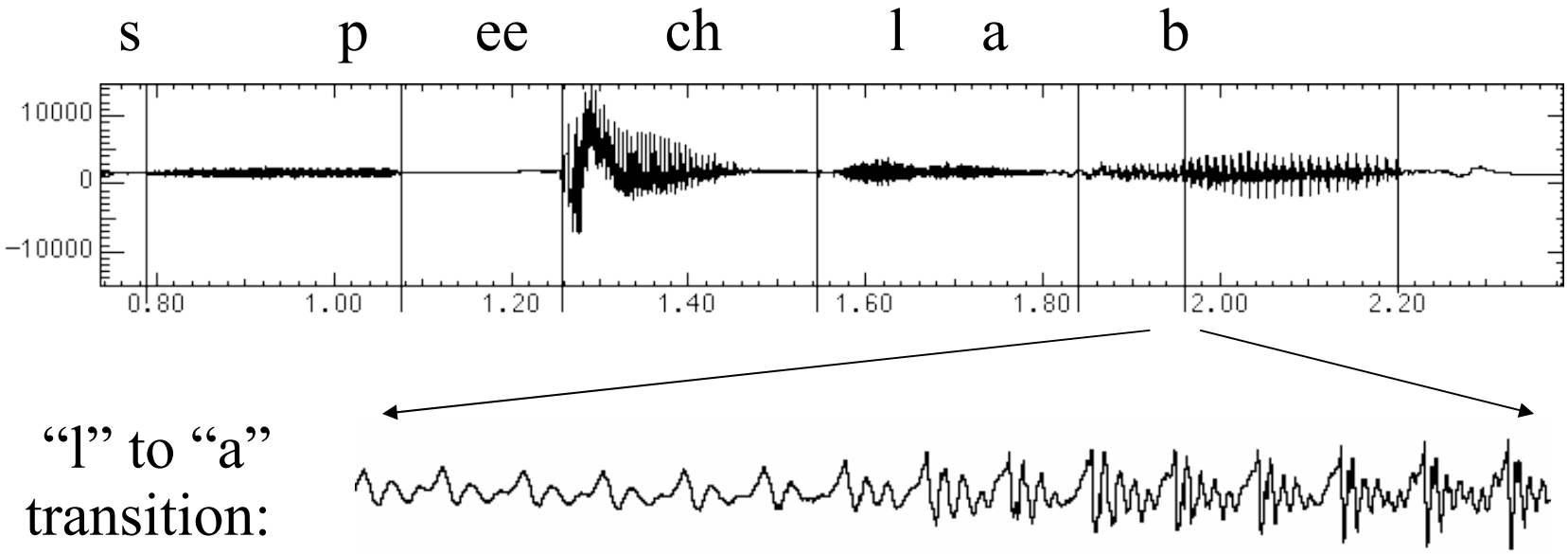
Digitizing Speech



Thanks to Bryan Pellom for this slide!

Speech in an Hour

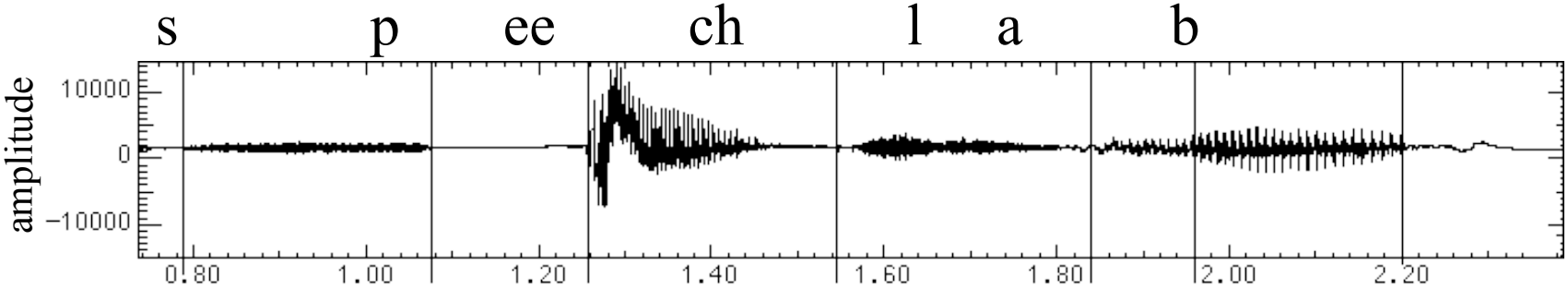
- Speech input is an acoustic wave form



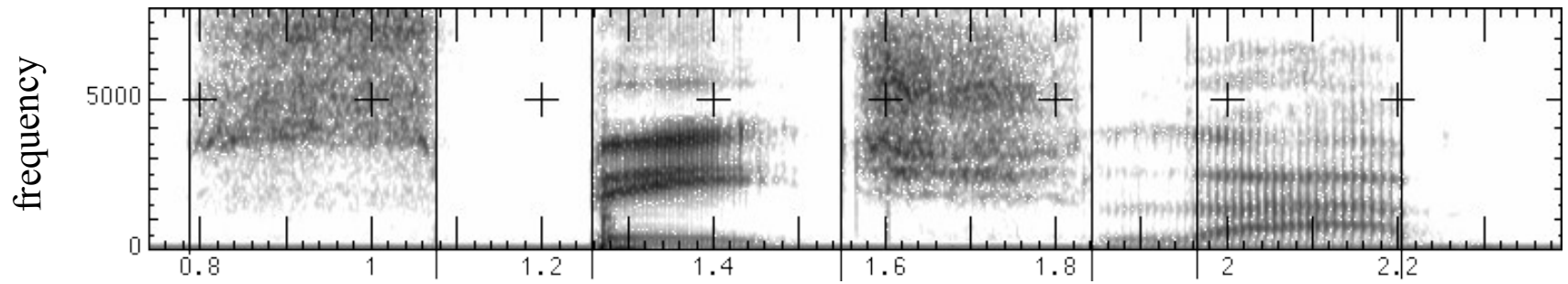
Graphs from Simon Arnfield’s web tutorial on speech, Sheffield:
http://www.psyc.leeds.ac.uk/research/cogn/speech/tutorial/

Spectral Analysis

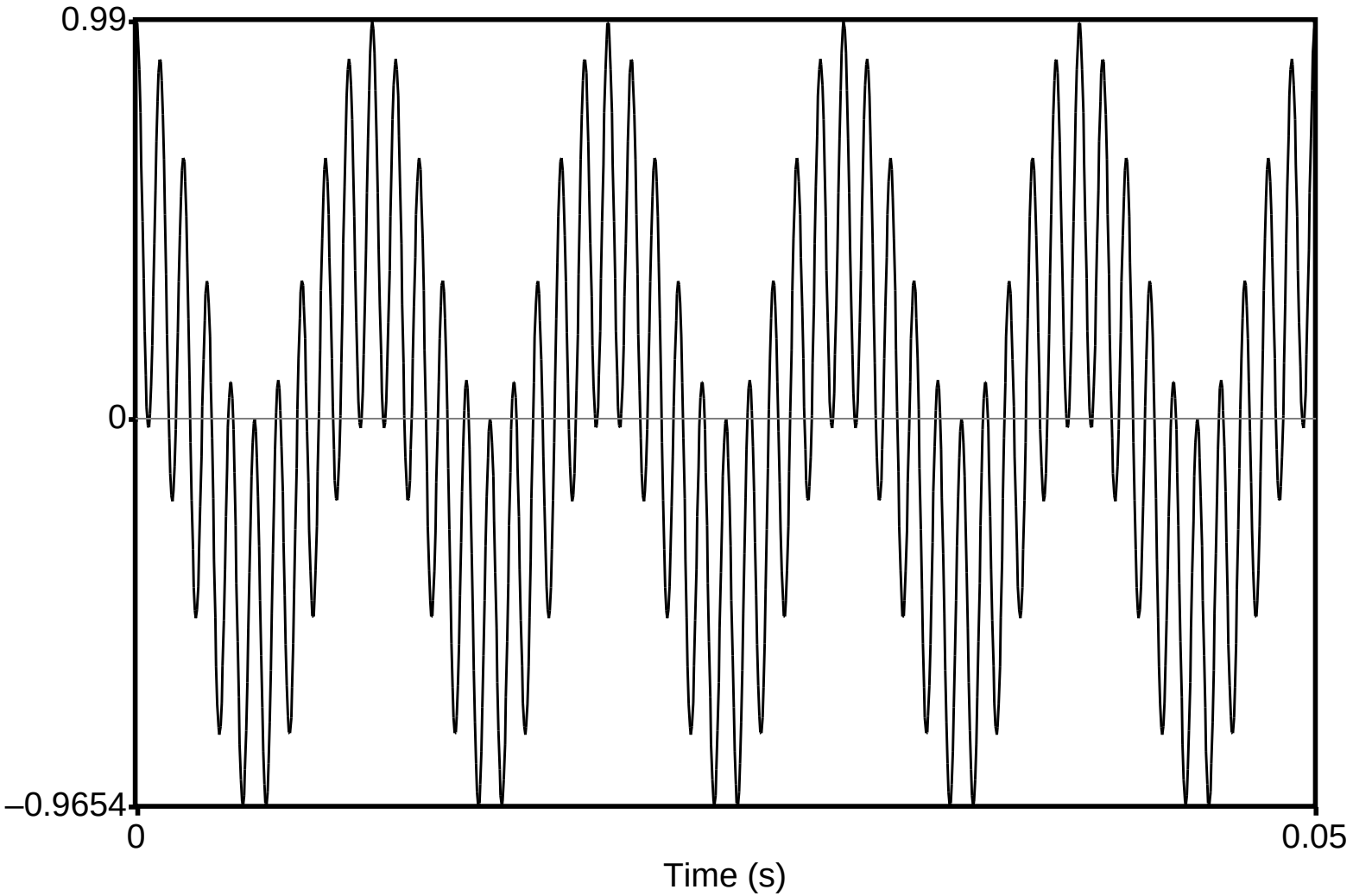
- Frequency gives pitch; amplitude gives volume
  - sampling at ~8 kHz phone, ~16 kHz mic (kHz=1000 cycles/sec)



- Fourier transform of wave displayed as a spectrogram
  - darkness indicates energy at each frequency

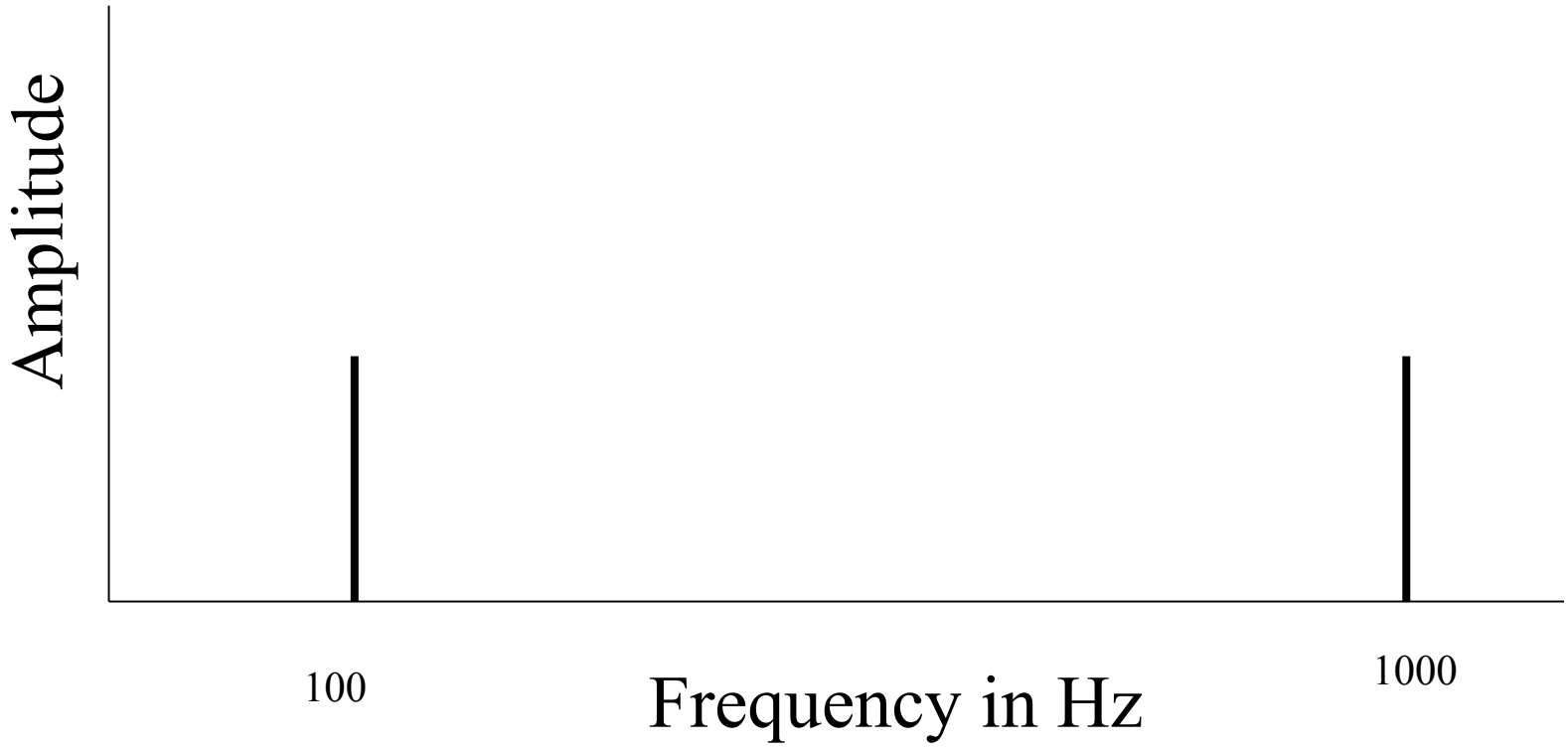


Adding 100 Hz + 1000 Hz Waves

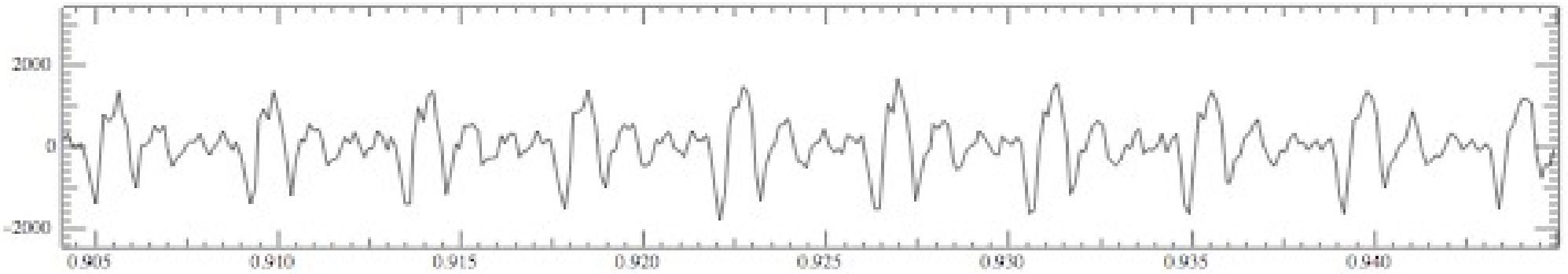


Spectrum

Frequency components (100 and 1000 Hz) on x-axis



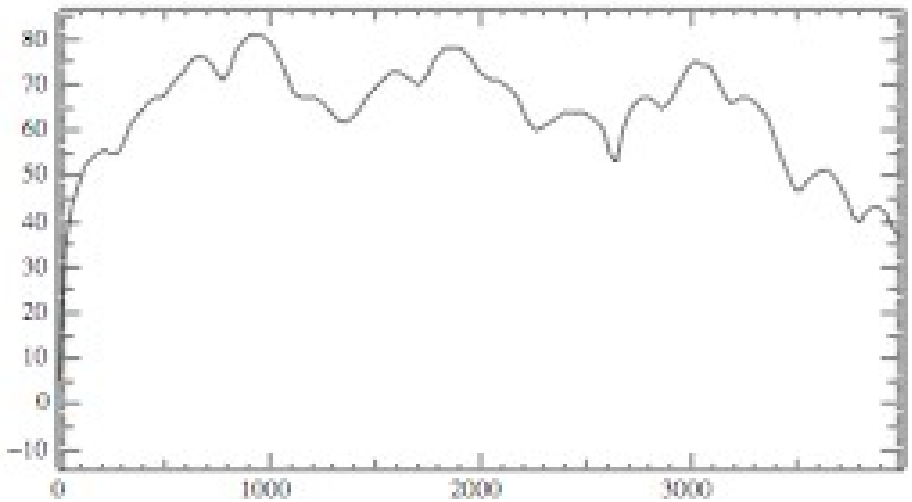
Part of [ae] from “lab”



- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

Back to Spectra

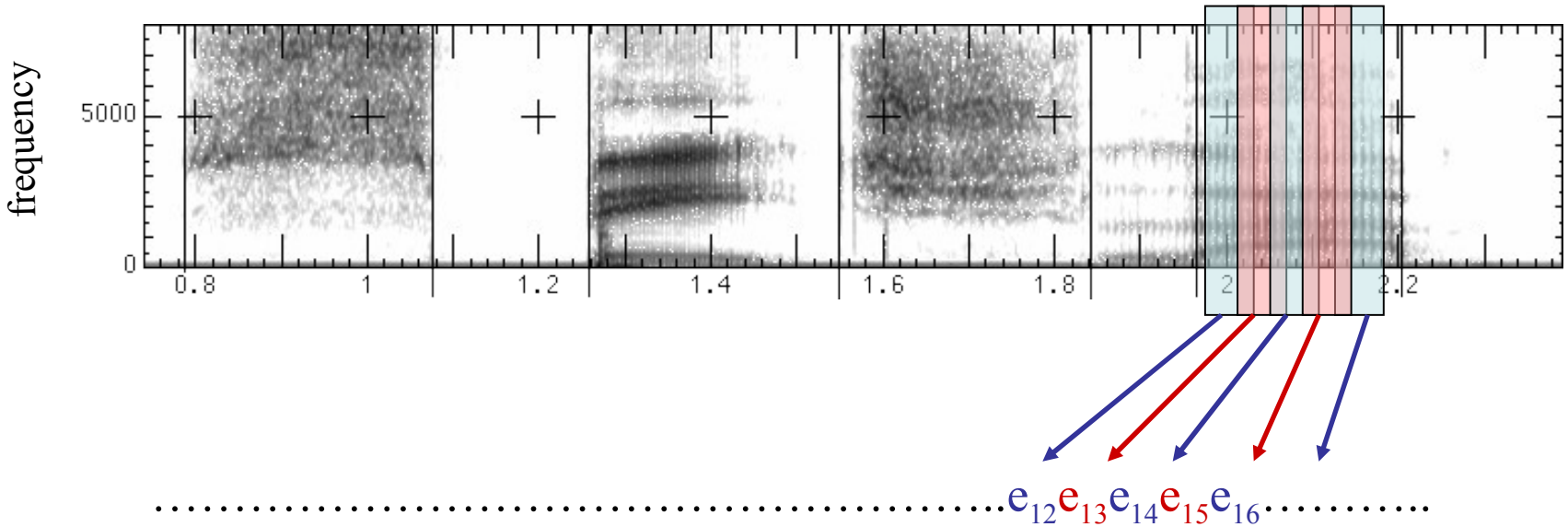
- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.



- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

Acoustic Feature Sequence

- Time slices are translated into acoustic feature vectors (~39 real numbers per slice)



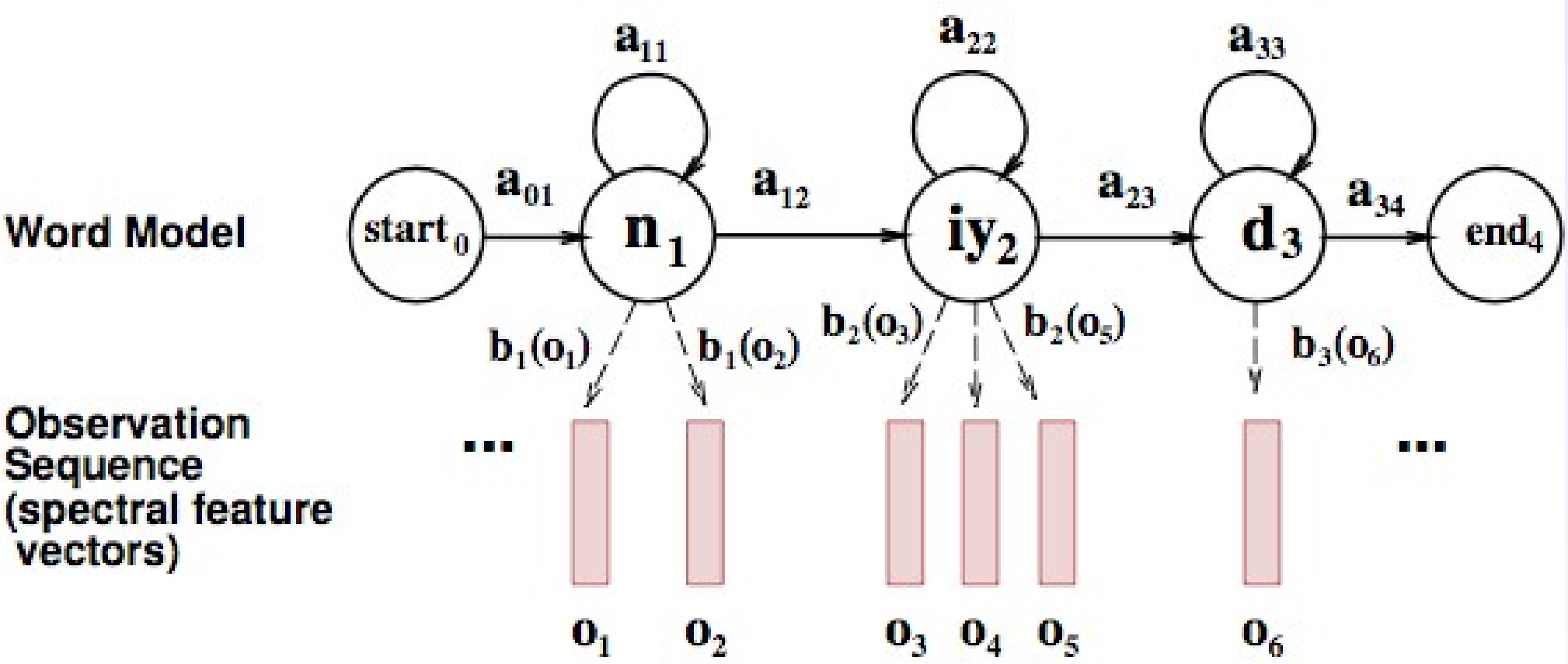
- These are the observations, now we need the hidden states X

State Space

- P(E|X) encodes which acoustic vectors are appropriate for each phoneme (each kind of sound)

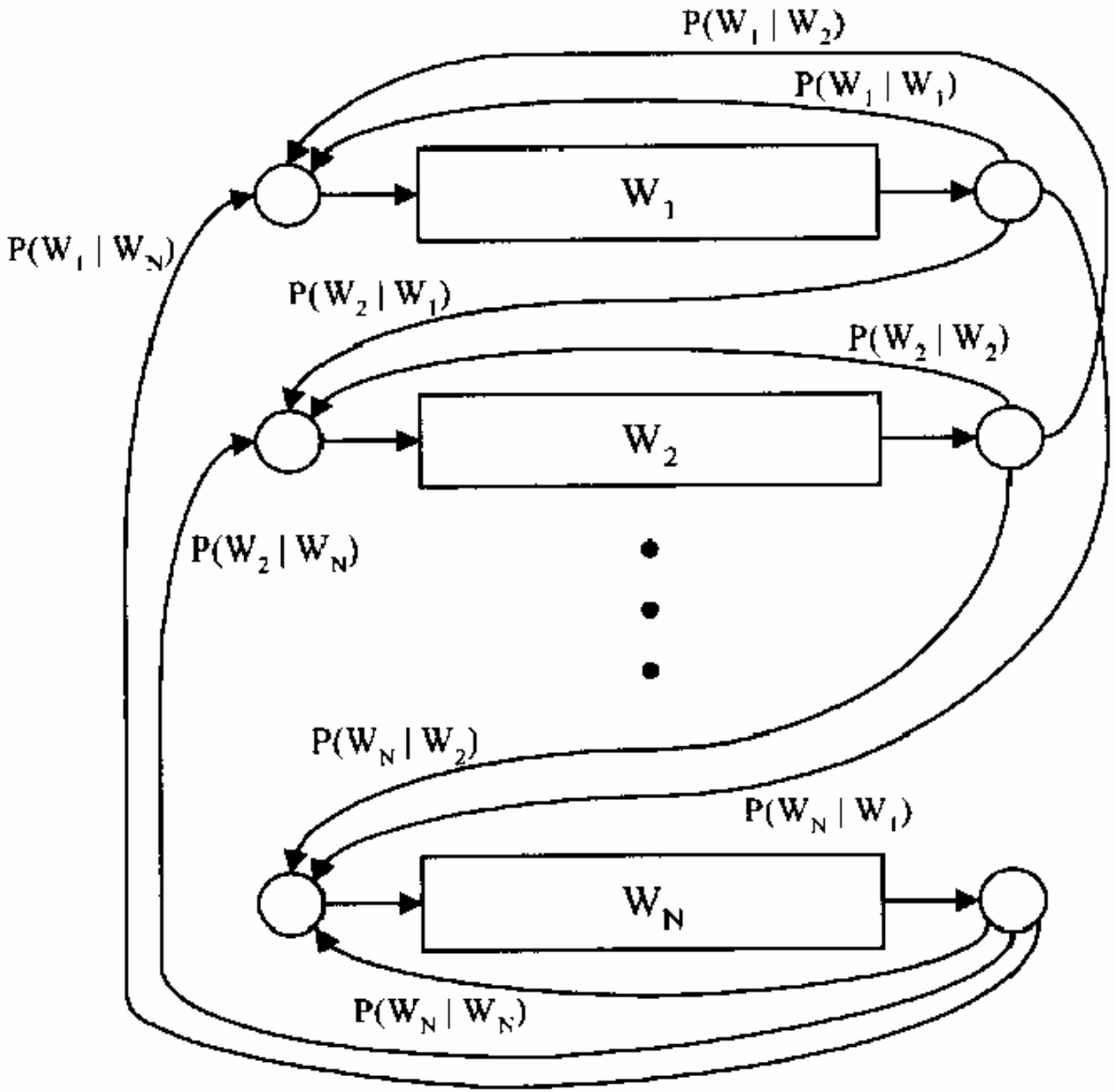
- P(X|X') encodes how sounds can be strung together
- We will have one state for each sound in each word
- From some state x, can only:
  - Stay in the same state (e.g. speaking slowly)
  - Move to the next position in the word
  - At the end of the word, move to the start of the next word
- We build a little state graph for each word and chain them together to form our state space X

HMMs for Speech



Markov Process with Bigrams

Figure from Huang et al page 618



Decoding

- While there are some practical issues, finding the words given the acoustics is an HMM inference problem
- We want to know which state sequence $x_{1:T}$ is most likely given the evidence $e_{1:T}$:

$$x^*_{1:T} = \arg\max_{x_{1:T}} P(x_{1:T}|e_{1:T})$$

$$= \arg\max_{x_{1:T}} P(x_{1:T}, e_{1:T})$$

- From the sequence x, we can simply read off the words

Training (aka “preview of ML”)

- Two key components of a speech HMM:
  - Acoustic model: p(E | X)
  - Language model: p(X | X')

  - Where do these come from?

- Can we estimate these models from data:
  - p(E | X) might be estimated from transcribed speech
  - p(X | X') might be estimated from large amounts of raw text

n-gram Language Models

- Assign a probability to a sequences of words

$$p(w_1, w_2, \ldots, w_I) = \prod_{i=1}^{I} p(w_i \mid w_1, \ldots, w_{i-1})$$

$$\approx \prod_{i=1}^{I} p(w_i \mid w_{i-k}, \ldots, w_{i-1})$$

- If I gave you a copy of the web, how would you estimate these probabilities?

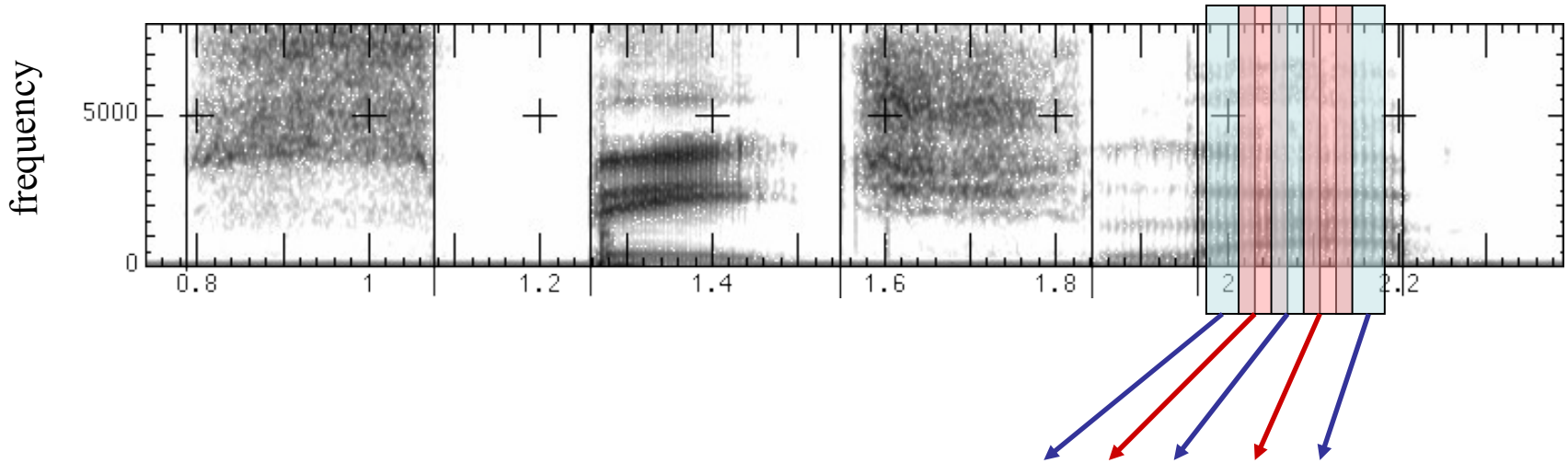
Need to "smooth" estimates intelligently to avoid zero probability *n*-grams.

Language modeling is the art of good smoothing.

See [Goodman 1998], [Teh 2007]

Acoustic models

- What if I gave you data like:



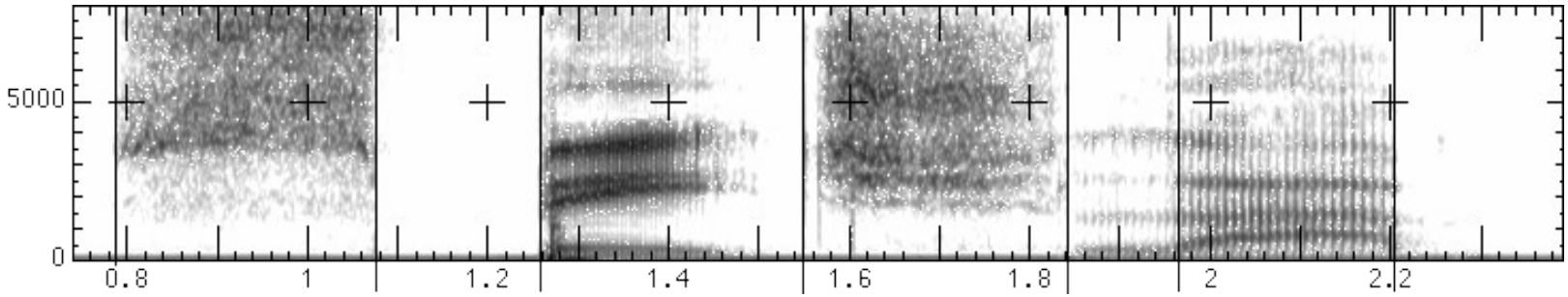
……..……..sp ee ch l ae b.....

- How would you estimate p(E|X)?
- What's wrong with this approach?

Acoustic models II

- What does our data really look like:

Acc:



W: yesterday I went to visit the speech lab

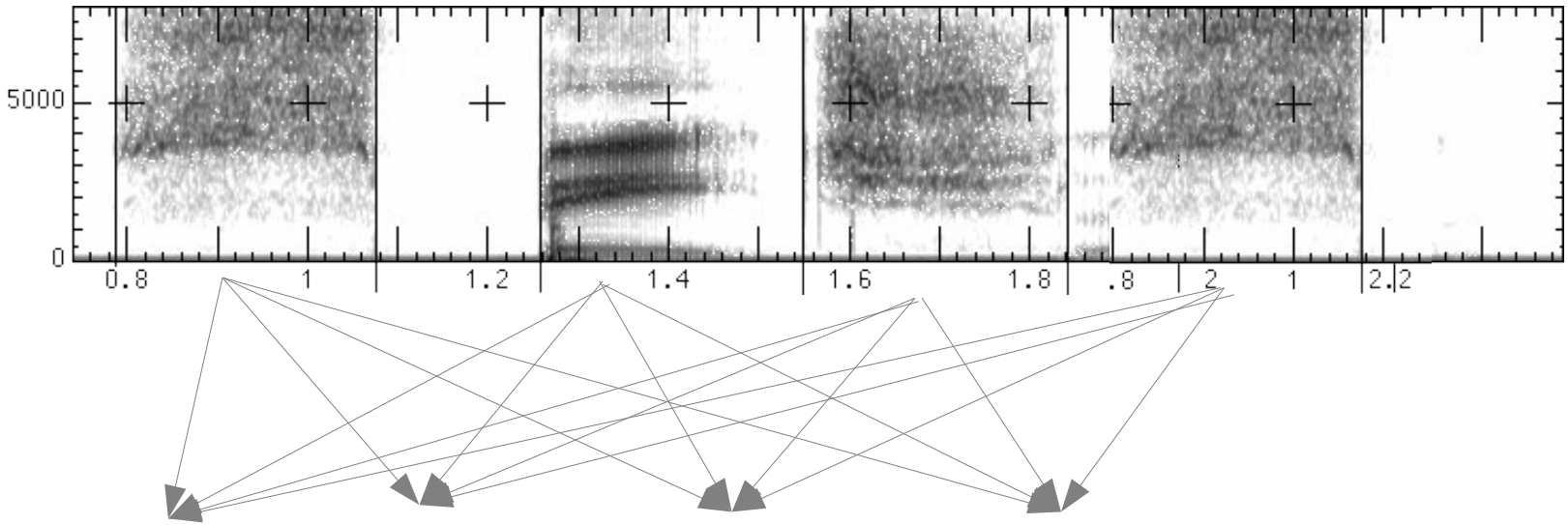
- We'd like to know *alignments* between transcript and waveform

- Suppose someone gave us a good speech recognizer.... could we figure out alignments from that?

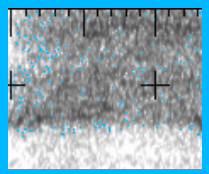
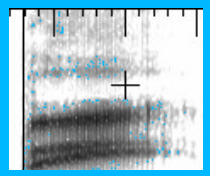
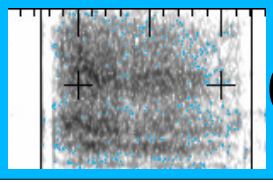
Expectation Maximization

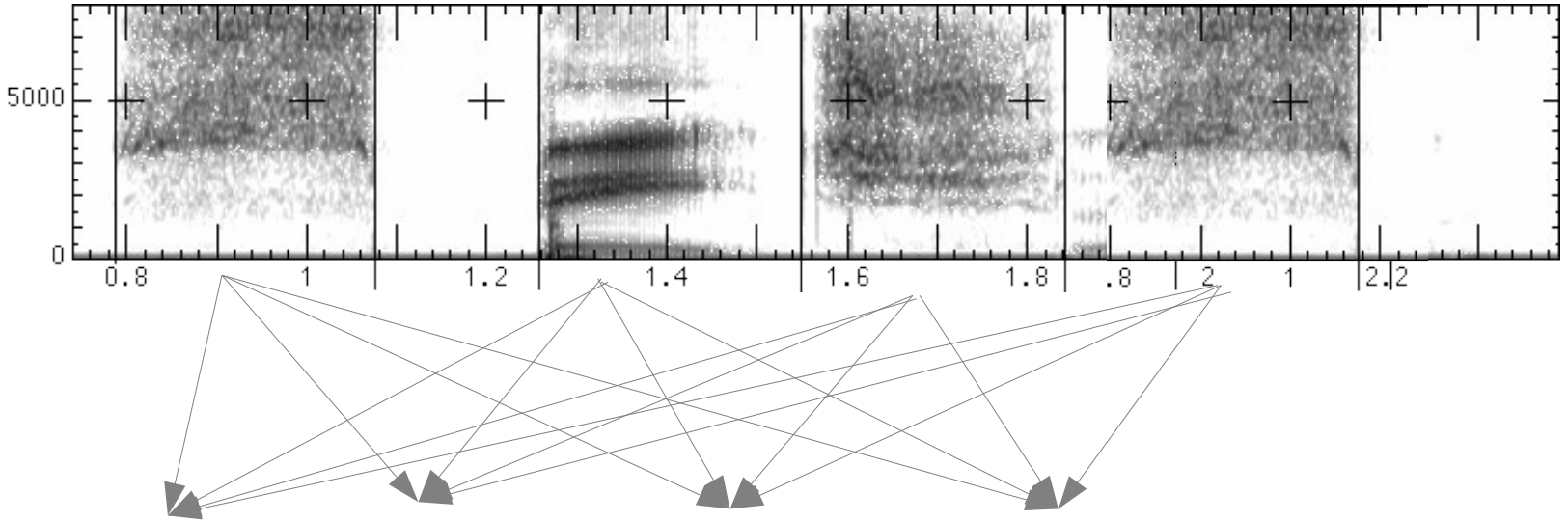
- A general framework to do parameter estimation in the presence of hidden variables

- Repeat ad infinitum:
  - E-step: make probabilistic guesses at latent variables
  - M-step: fit parameters according to these guesses



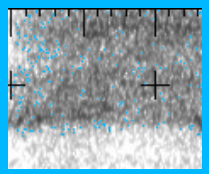
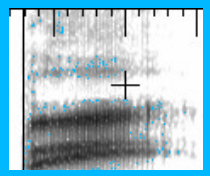
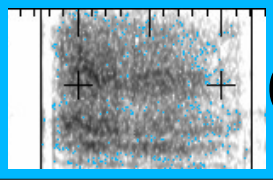
Expectation Maximization

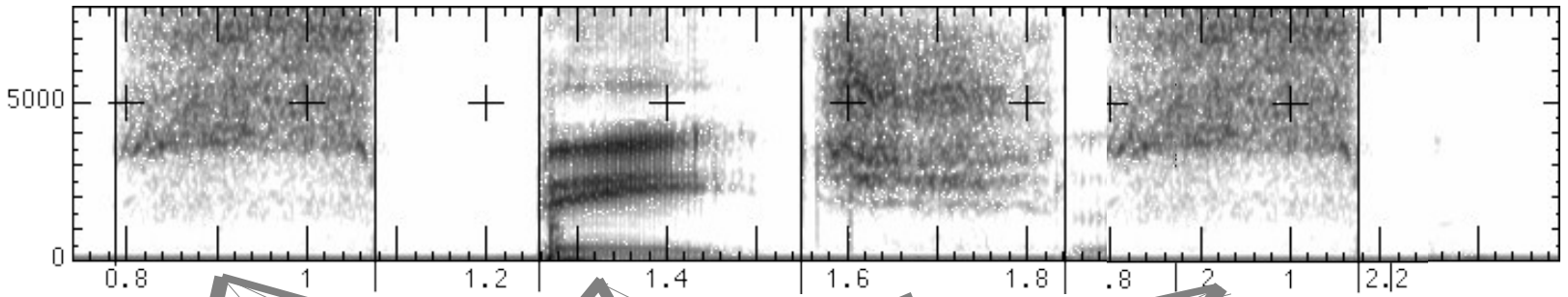
e	p(e | "I")	p(e | "LIKE")	p(e | "A")
	0.33 → 2	0.33 → 1	0.33 → 1
	0.33 → 1	0.33 → 1	0.33 → 1
	0.33 → 1	0.33 → 1	0.33 → 1



W: I LIKE A I

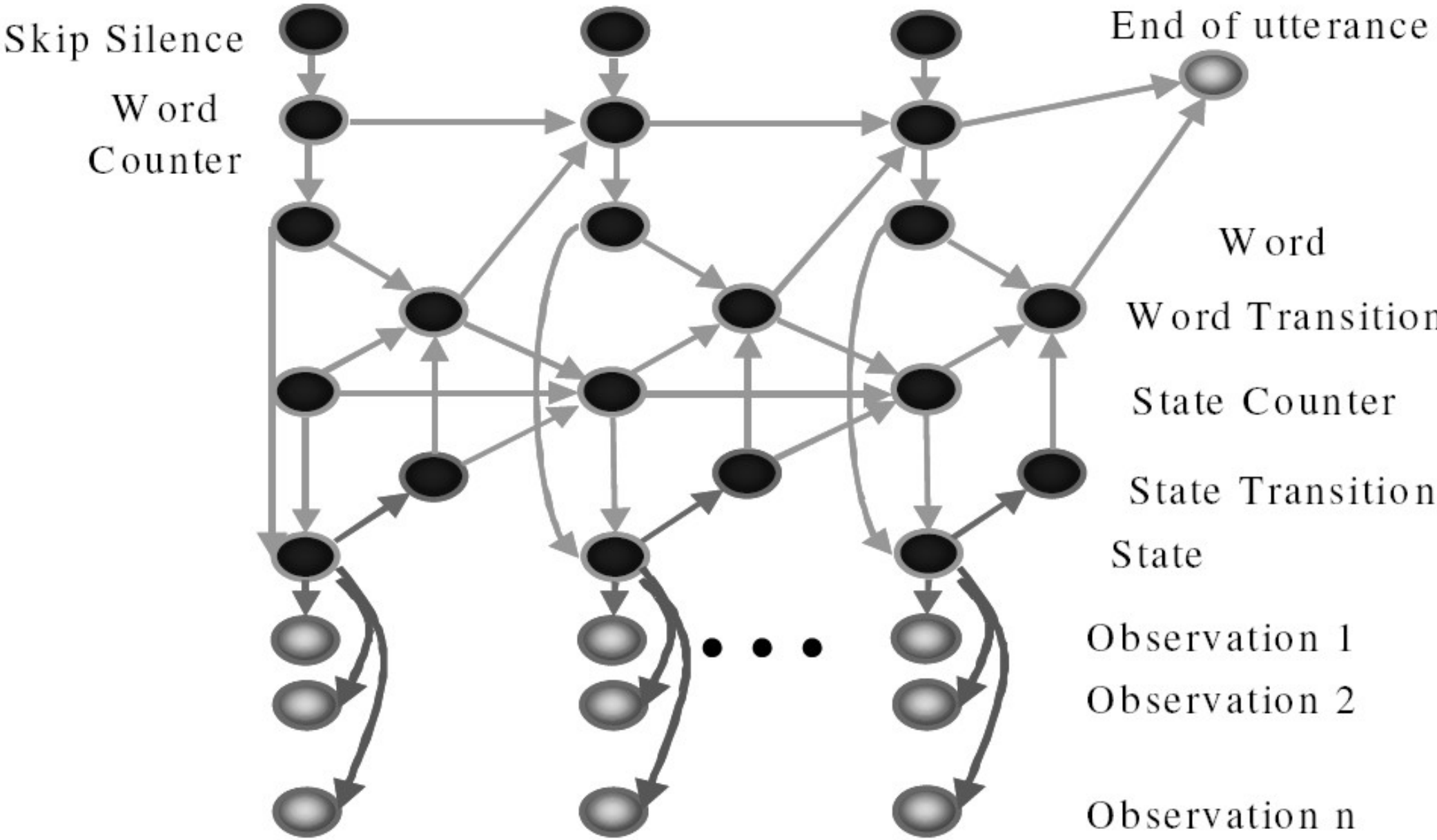
Expectation Maximization

e	p(e | "I")	p(e | "LIKE")	p(e | "A")
	0.5 → 4	0.33 → 1	0.33 → 1
	0.25 → 1	0.33 → 2	0.33 → 2
	0.25 → 1	0.33 → 2	0.33 → 2



W: I LIKE A I

State of the Art DBNs for Speech



Summary

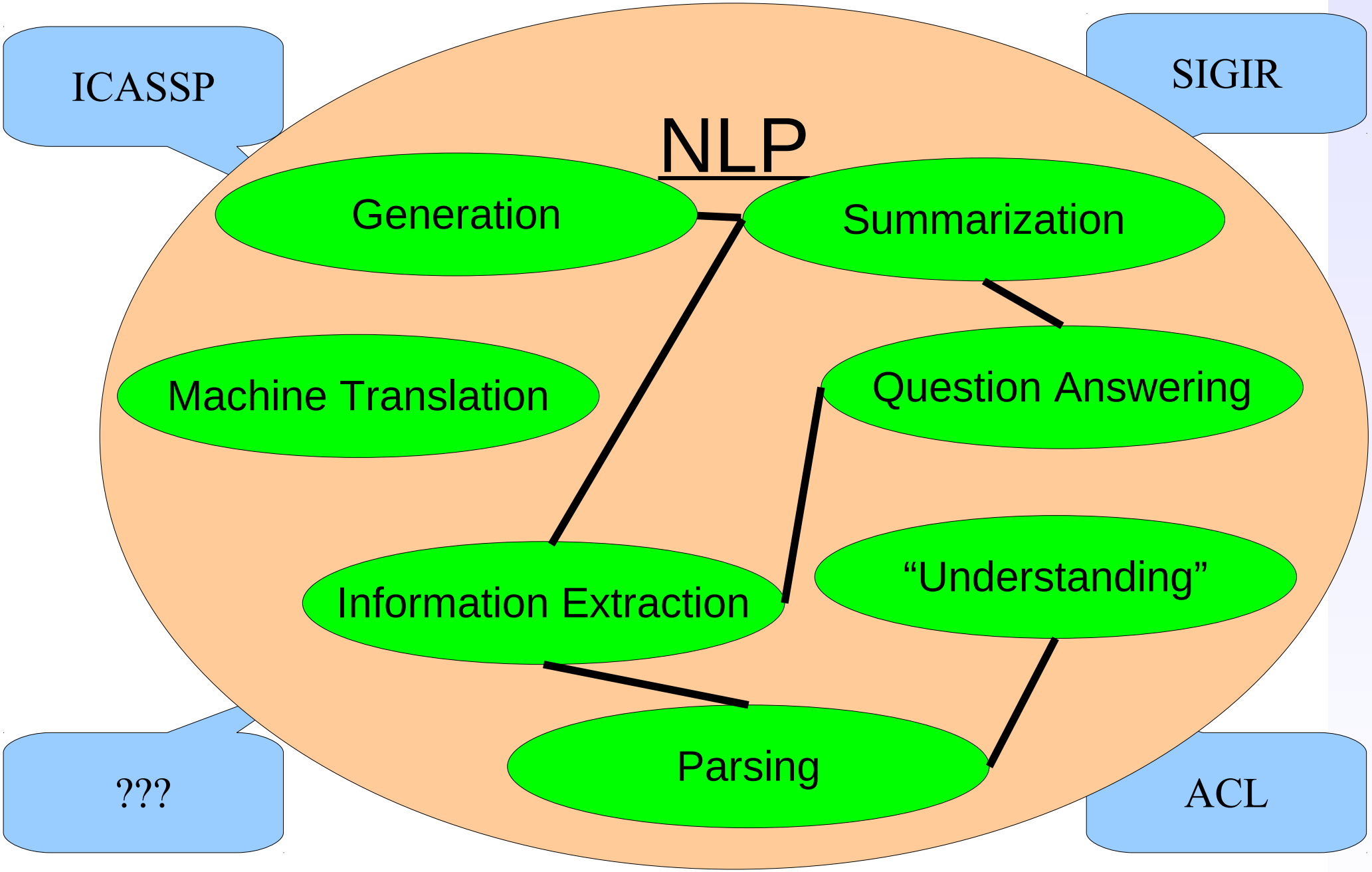
- HMMs allow us to “separate” two models:
  - acoustic model (how does what I want to say sound?)
  - language model (what do I want to say)
- Speech recognition is “just” decoding in an HMM/DBN
  - Plus a heck of a lot of engineering
- Expectation maximization lets us estimate parameters in models with hidden variables
- Most research today focuses on language modeling

Translate Centauri -> Arcturan

Your assignment, translate this Centauri sentence 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 .

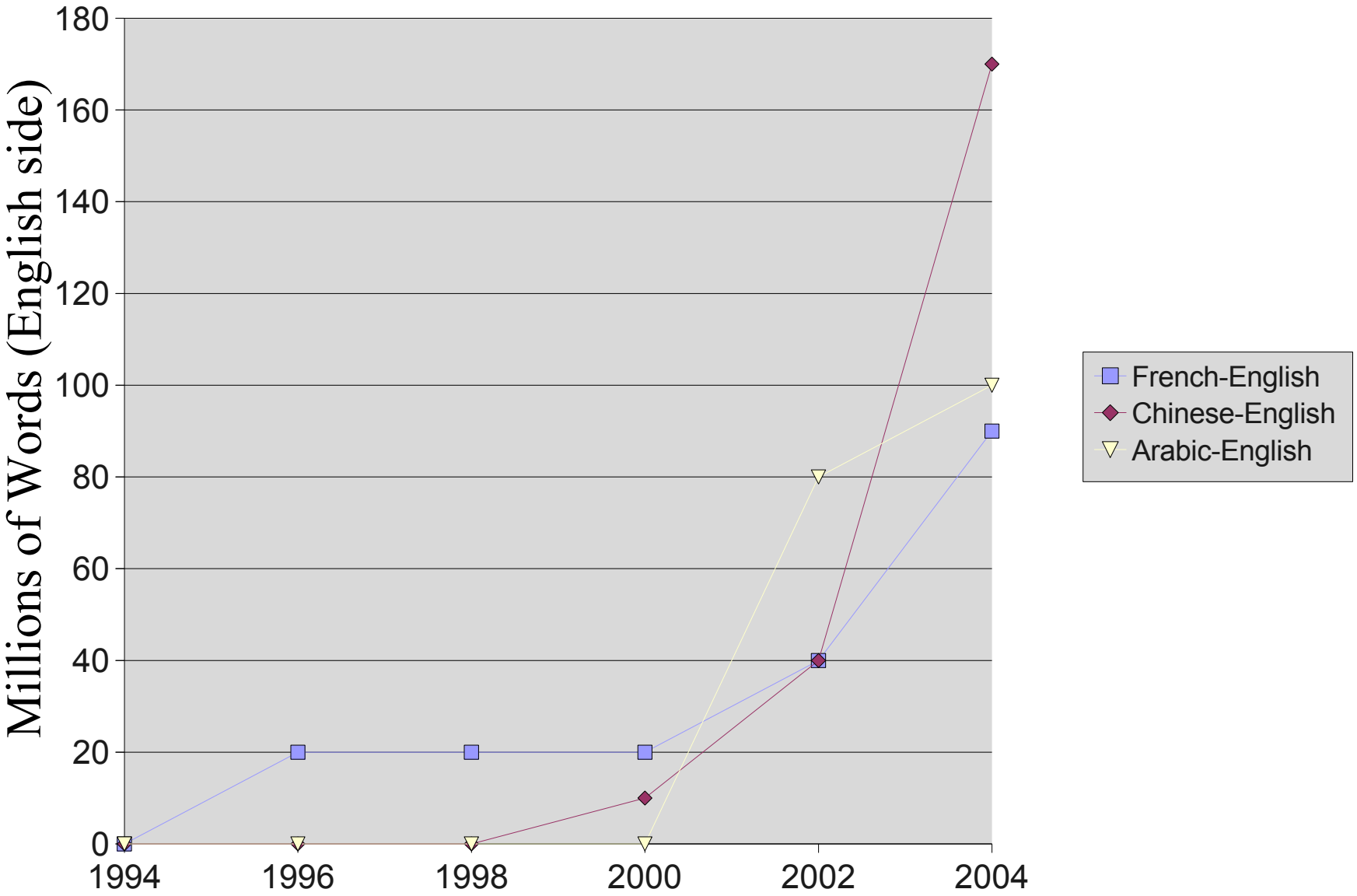
Topology of the Field



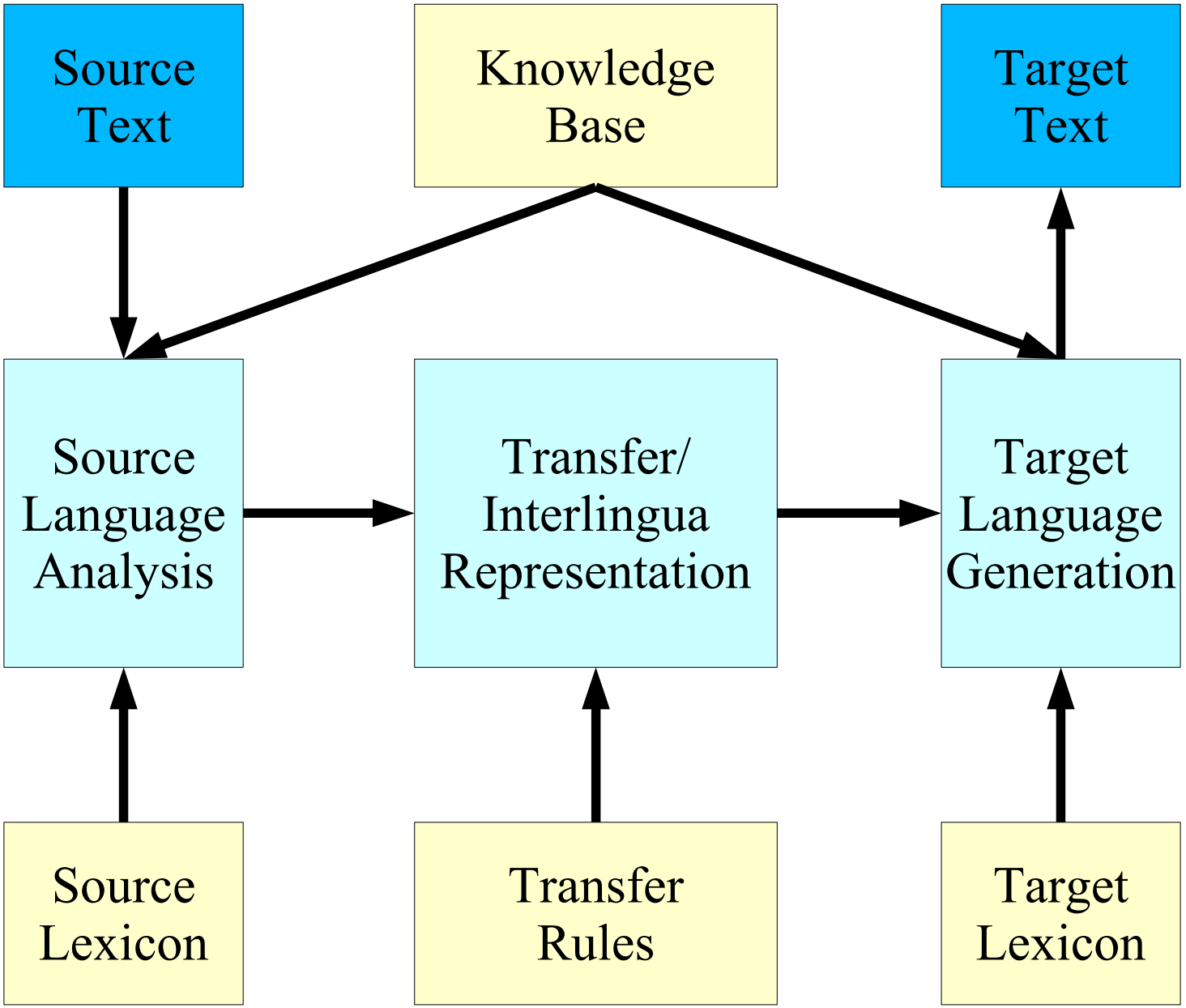
A Bit of History

1940s	Computations begins, AI hot, Turing test
	Machine translation = Code-breaking?
1950s	Cold war continues
1960s	Chomsky and statistics, ALPAC report
1970s	Dry spell
1980s	Statistics makes significant advances in speech
1990s	Web arrives
	Statistical revolution in machine translation, parsing, IE, etc
	Serious “corpus” work, increasing focus on evaluation
2000s	Focus on optimizing loss functions, reranking
	How much can we automate?
	Huge process in machine translation
	Gigantic corpora become available, scaling
	New challenges

Ready-to-use Data

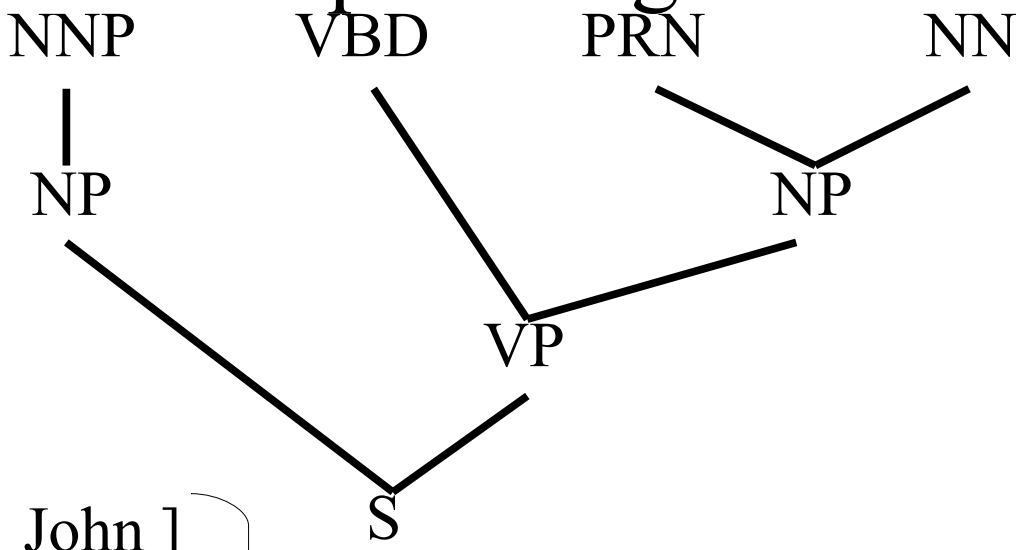


Classical MT (1970s and 1980s)



Layers of complexity

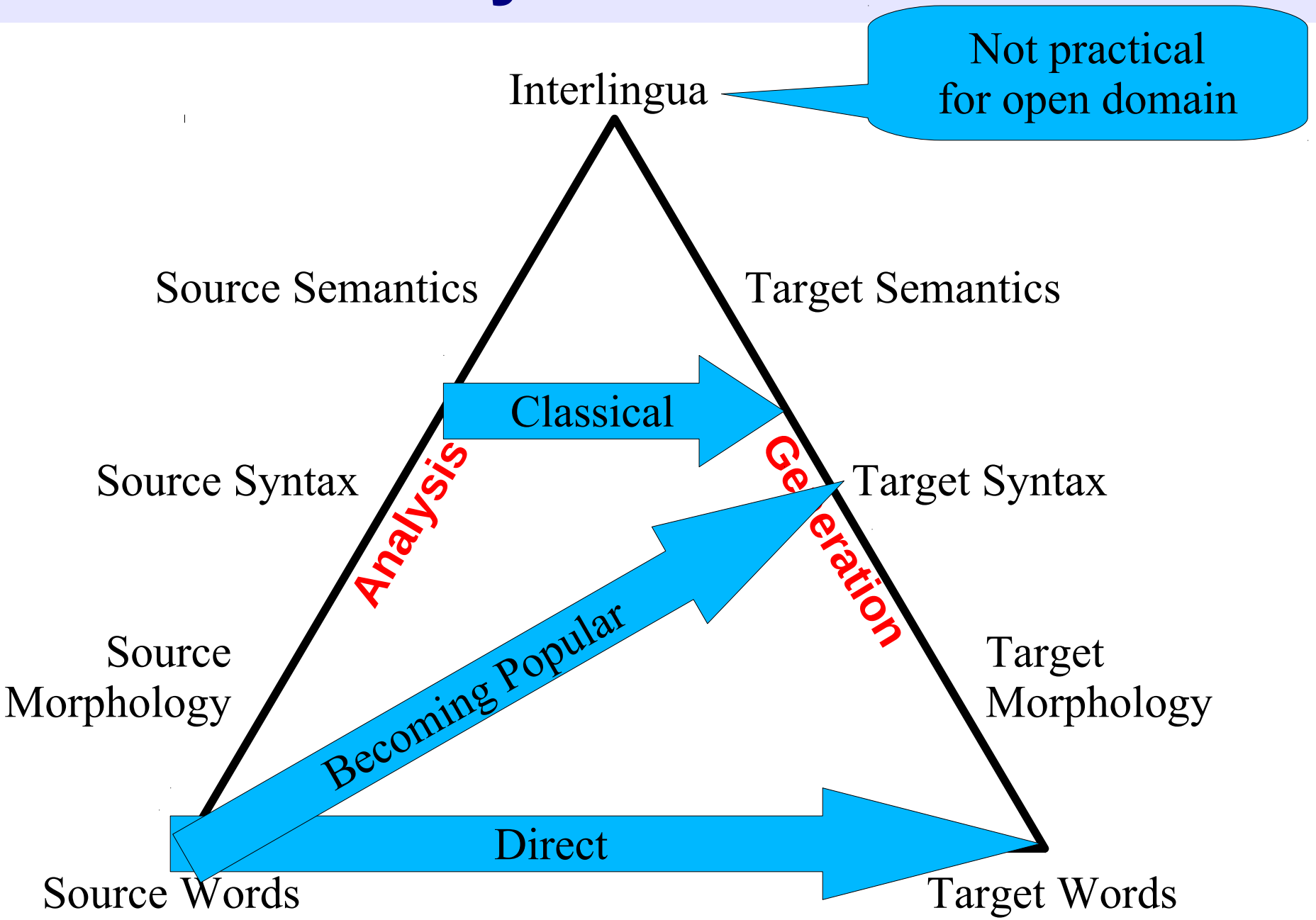
- Text: John saw his brother
- Morphology: John see he brother +past +gen
- Syntax:



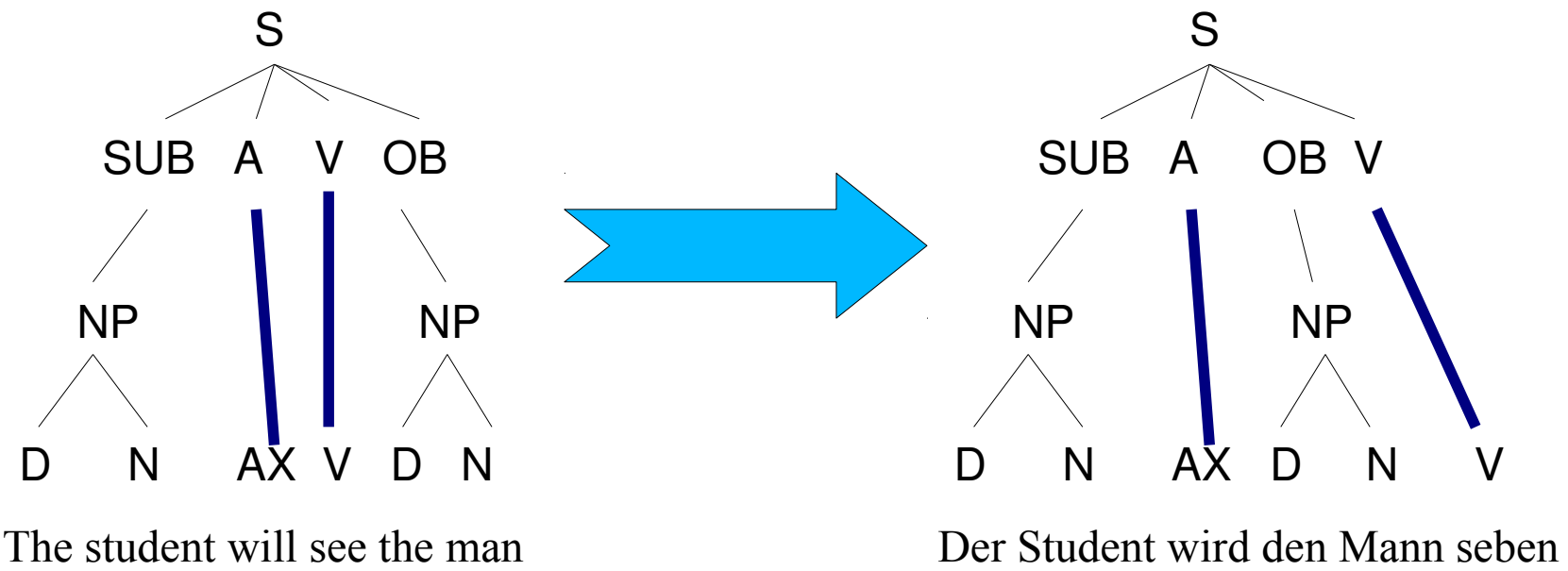
- Semantics:

Agent [Person John]
Event see (+past)
Patient Person brother
Poss *

How Much Analysis?

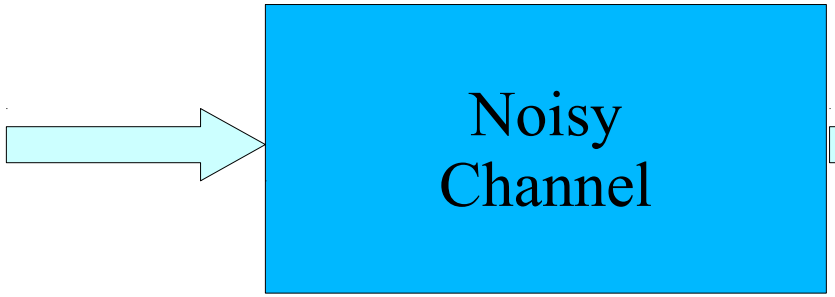


Syntactic Transfer



- Now, just get a bunch of linguists to sit down and write rules and grammars

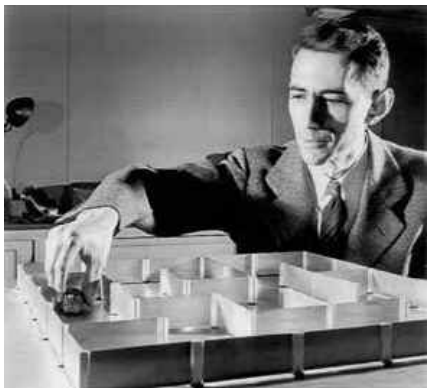
While We're Busy Writing Grammars



$p(s)$ $p(o \mid s)$ $p(s \mid o) \propto p(s) * p(o \mid s)$

"Imagined" Words | Speech Process | Acoustic Signal

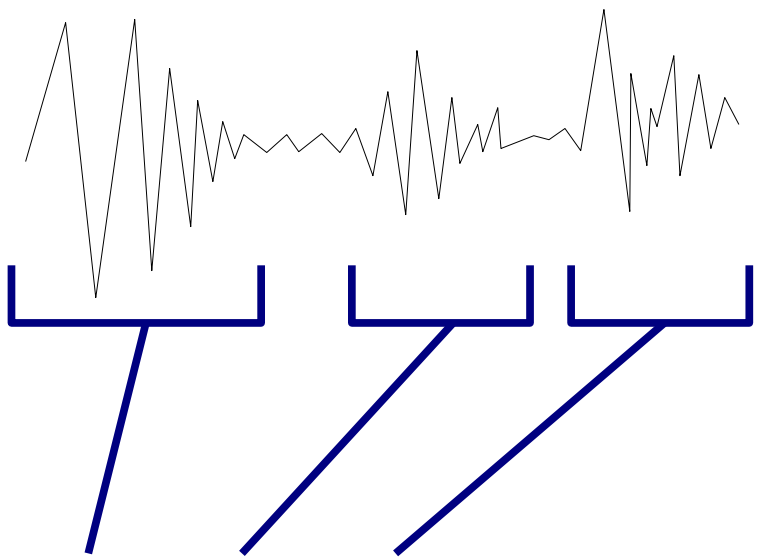
Need: *p(word sequence)*
and *p(signal | word sequence)*



Claude Shannon

Acoustic Modeling: p(a | w)

Signal:



Transcription: the man ate

Key notion: acoustic-word *a*

Chicken-and-egg problem that we can solve using Expectation Maximization (EM)

Speech Rec = Machine Translation?

- Peter F. Brown
- Stephen A. Della Pietra
- Vincent J. Della Pietra
- Robert Mercer
- *The Mathematics of Statistical Machine Translation: Parameter Estimation*
- Computational Linguistics 19 (2), June 1993

- Probably the most important paper in NLP in the last 20 years

"Brown 93"

Centauri/Arcturan [Knight 97]

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 97]

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 97]

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 97]

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 97]

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 97]

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 97]

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 97]

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 97]

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 97]

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 . 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 97]

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

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

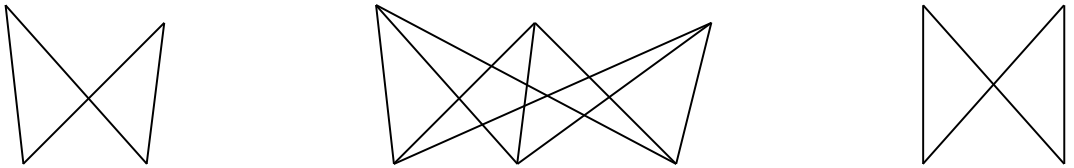
Centauri/Arcturan [Knight 97]

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

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

Unsupervised EM Training

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

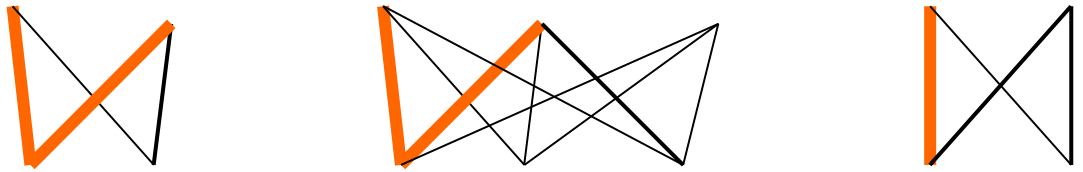


… the house ….. the blue house ..…. the flower …

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

Unsupervised EM Training

… la maison …… la maison bleue …… la fleur …

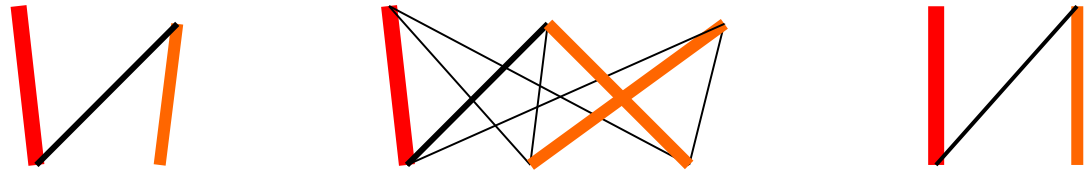


… the house …… the blue house …… the flower …

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

Unsupervised EM Training

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



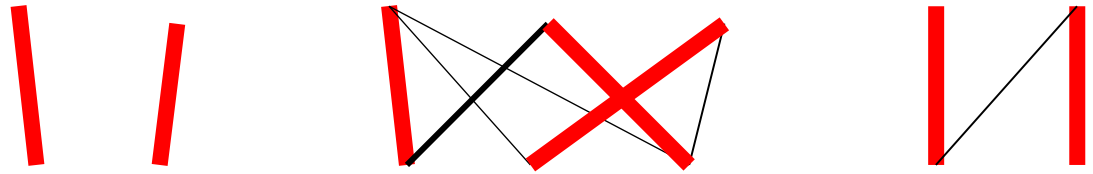
… the house …… the blue house …… the flower …

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

(pigeonhole principle)

Unsupervised EM Training

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



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

settling down after another iteration

Unsupervised EM Training

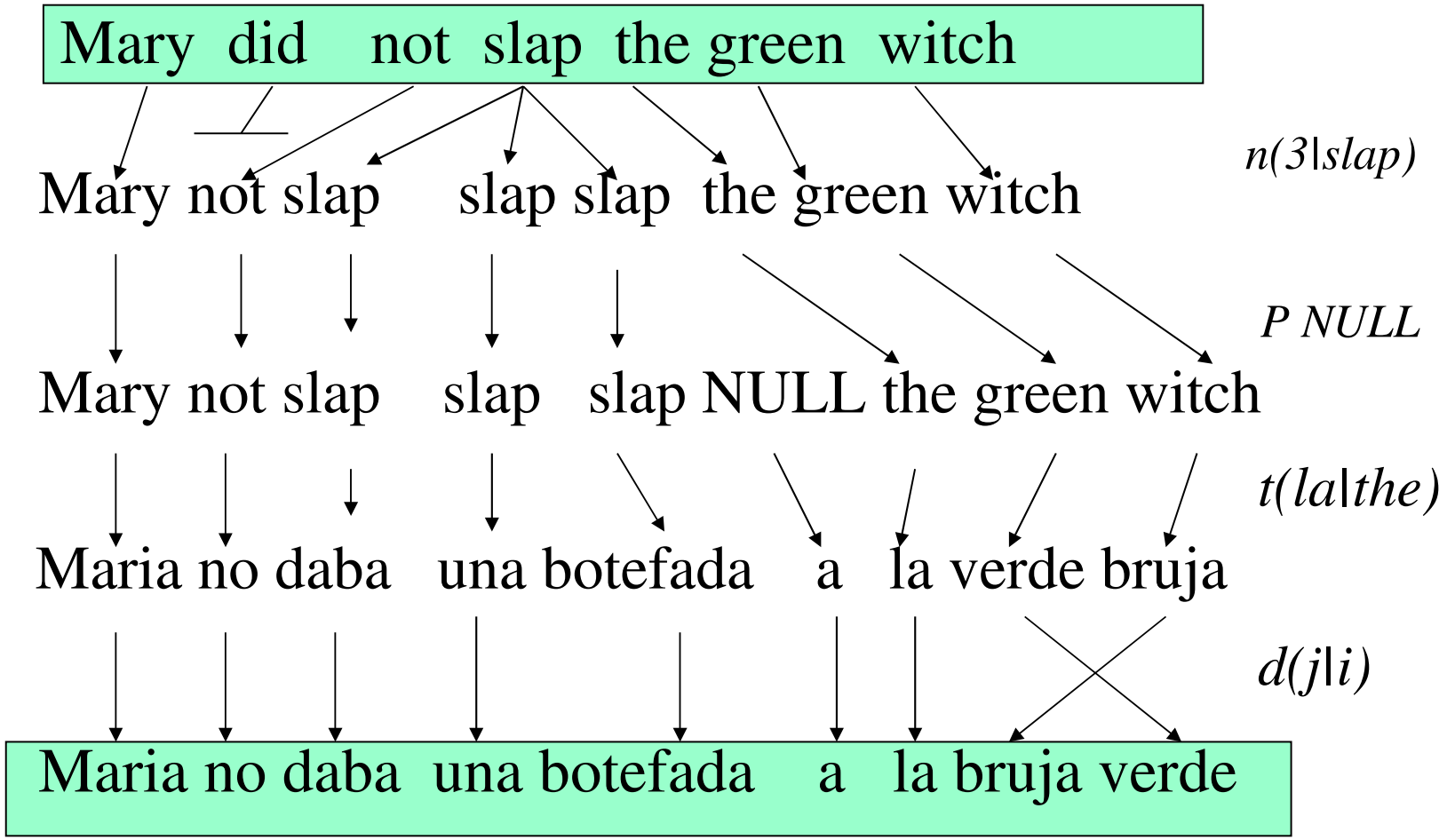
… la maison ….. la maison bleue …... la fleur …

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

Inherent hidden structure revealed by EM training!

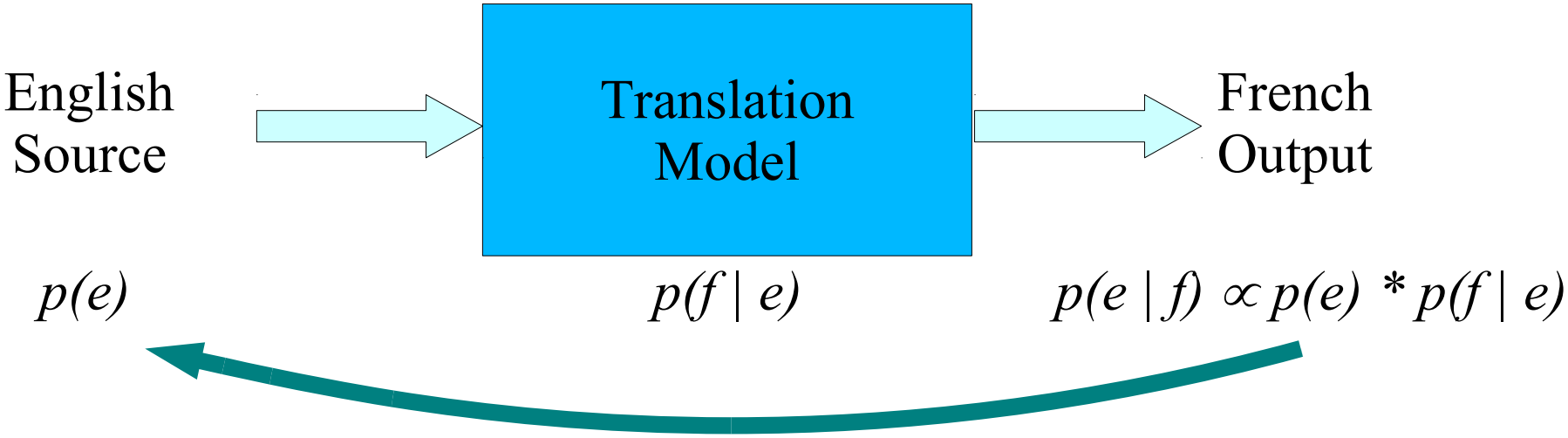
- "A Statistical MT Tutorial Workbook" (Knight, 1999). Promises free beer.
- "The Mathematics of Statistical Machine Translation" (Brown et al, 1993)
- Software: GIZA++

The IBM Model [Brown et al., 1993]



Use the EM algorithm for training the parameters

Decoding for Machine Translation



$$\text{Decoding: } \hat{e} = \underset{e}{argmax}\ p(e)\ p(f \mid e)$$

Problem in NP-hard; use search:

Greedy Search

Integer Programming

Beam Search

A* Search

Progress in Statistical MT

slide from C. Wayne, DARPA

2002

insistent Wednesday may recurred her trips to Libya tomorrow for flying

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

And said the official " the institution sent a speech to Ministry of Foreign Affairs of lifting on Libya air , a situation her receiving replying are so a trip will pull to Libya a morning Wednesday " .

2003

Egyptair Has Tomorrow to Resume Its Flights to Libya

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

" The official said that the company had sent a letter to the Ministry of Foreign Affairs, information on the lifting of the air embargo on Libya, where it had received a response, the first take off a trip to Libya on Wednesday morning ".

Automatic Evaluation of Translation

Reference 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.

Tri-gram match

Bi-gram matches

"Bleu" metric

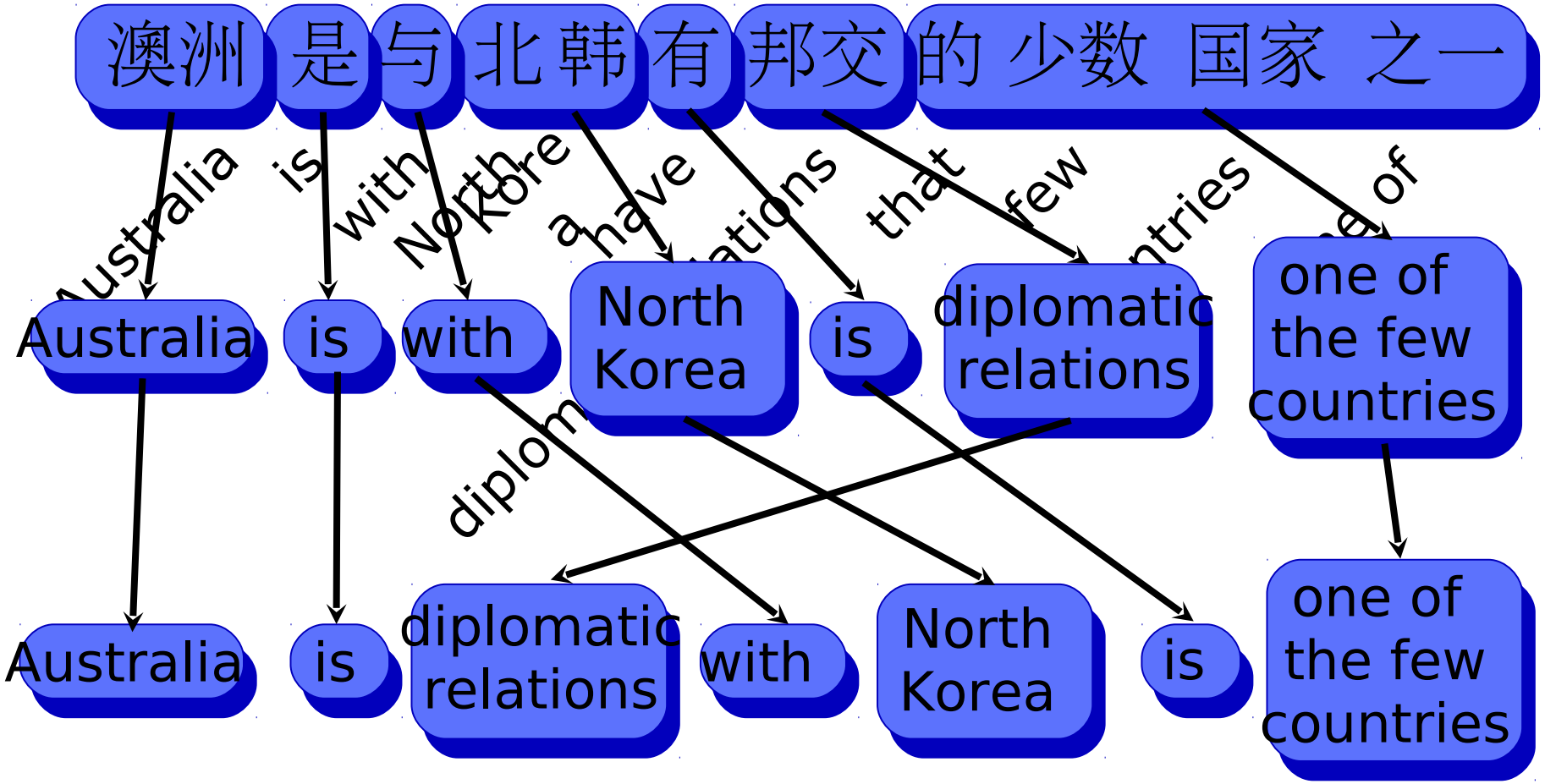
Minimum Error Rate Training for MT

- Desire MT system with high BLEU/??? scores [Och, ACL03]

- Algorithm:
  - Build MT system based on generative parameters
  - Decode development corpus to get n-best lists (~10k best)
  - Optimize parameters to get high BLEU scores on n-best lists
  - Repeat until converged

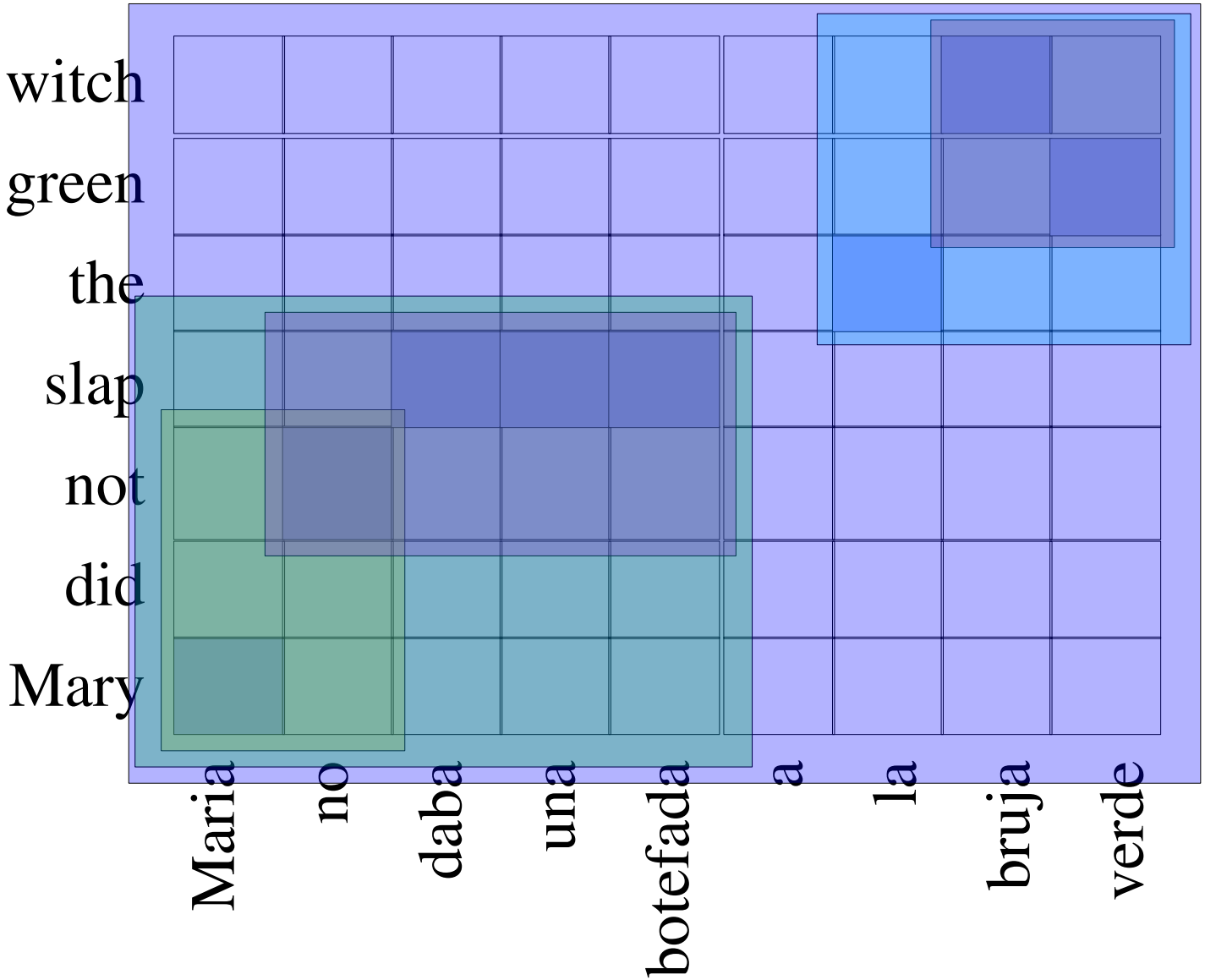
Phrase-Based Translation

[Koehn, Och and Marcu, NAACL03]



Training Phrase-Based MT Systems

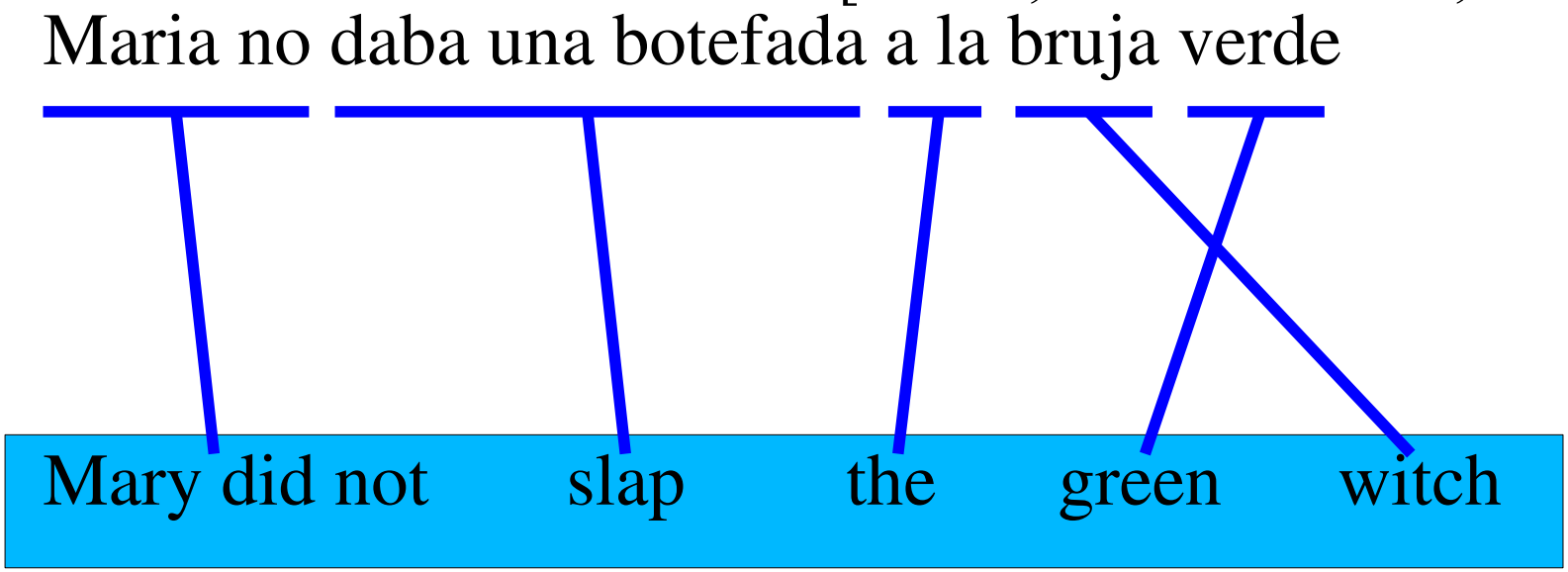
[Koehn, Och and Marcu, NAACL03]



57

Decoding Phrase-Based MT

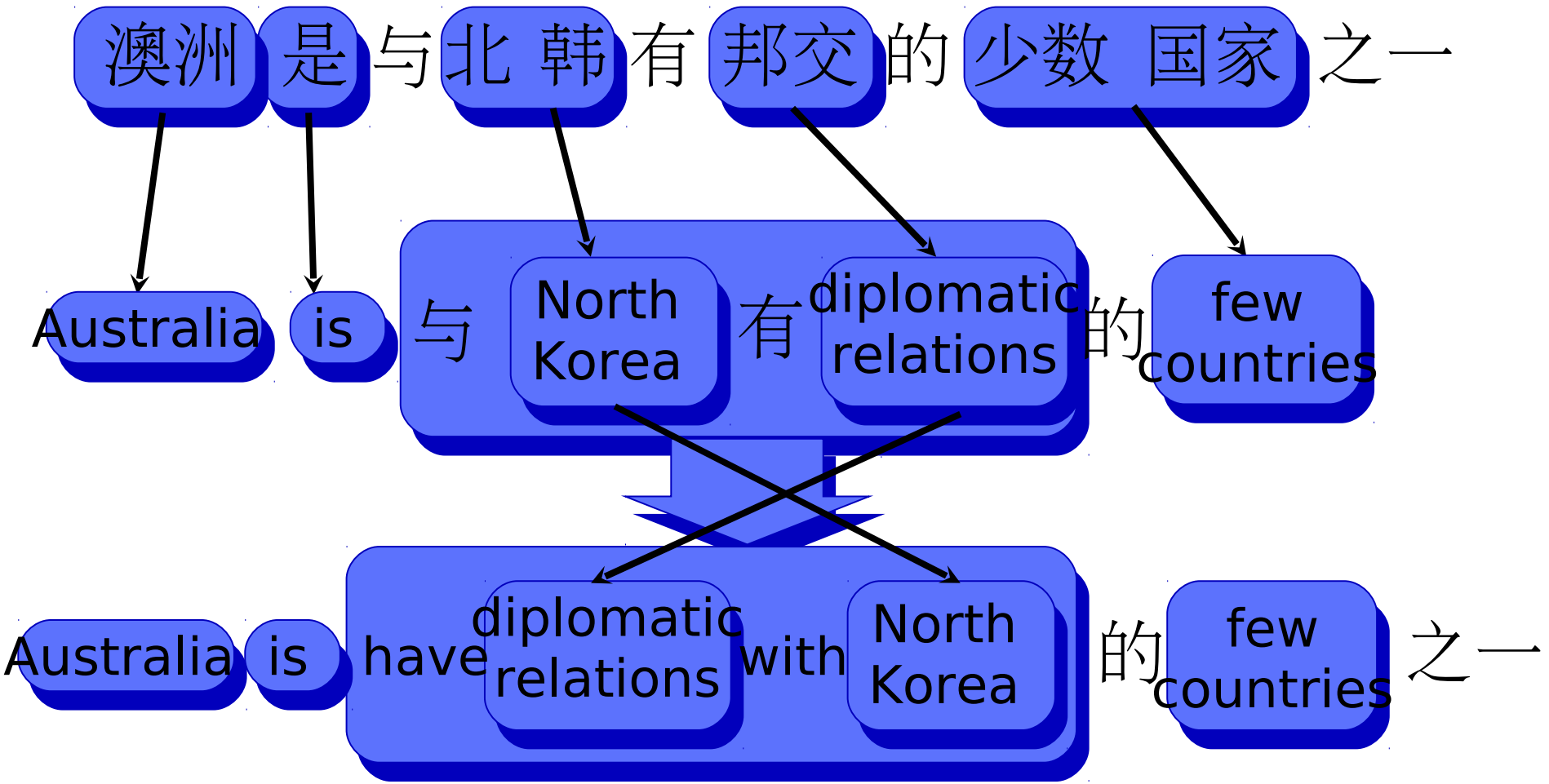
[Koehn, Och and Marcu, NAACL03]



- Each step induces a cost attributed to:
  - Language model probability: *p(slap | did not)*
  - T-table probability: *p(the | a la)* and *p(a la | the)*
  - Distortion probability: *p(skip 1)* [for a la --> verde]
  - Length penalty
  - ...

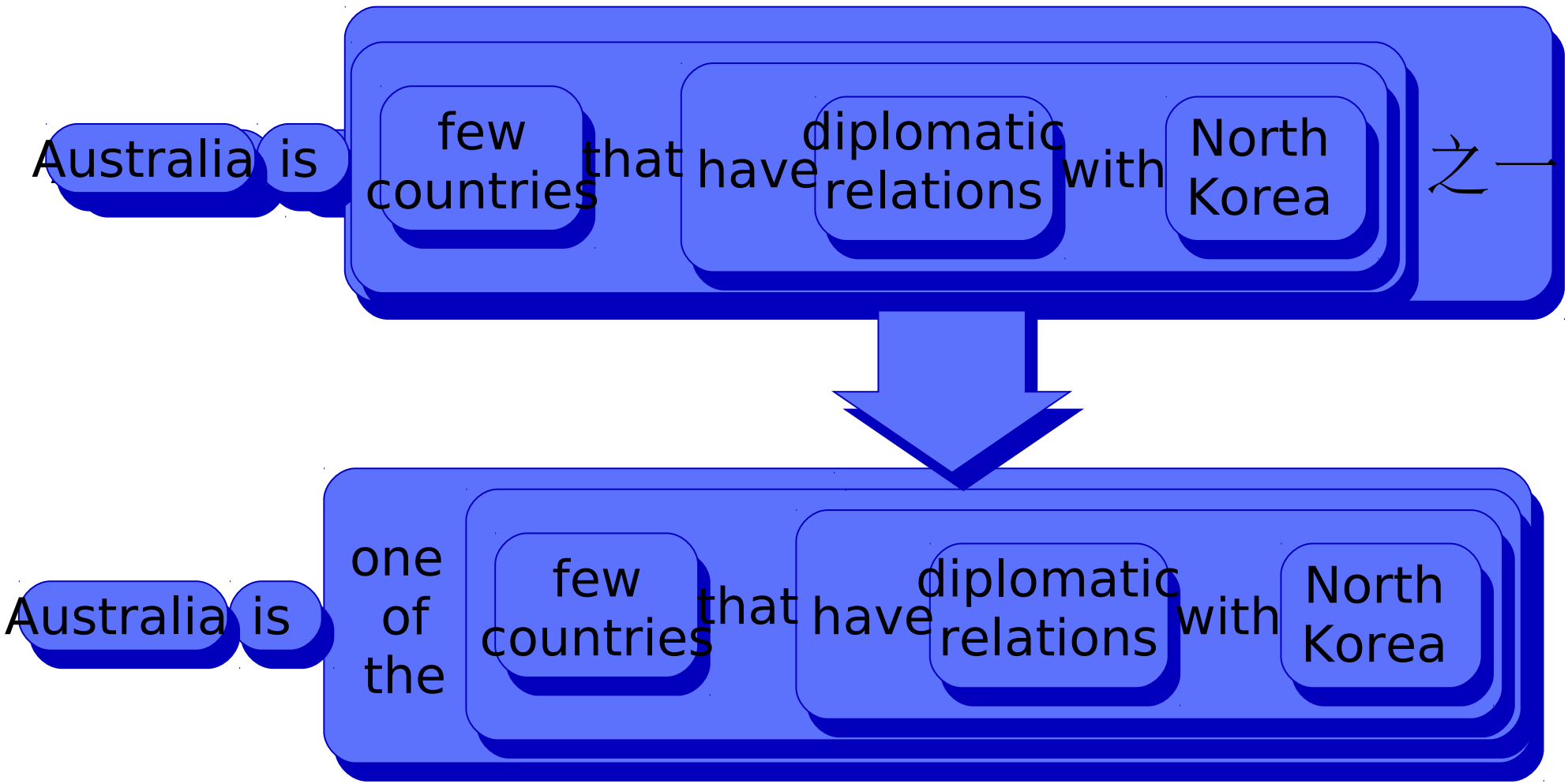
Hierarchical Phrase-Based MT

[Chiang, ACL05]



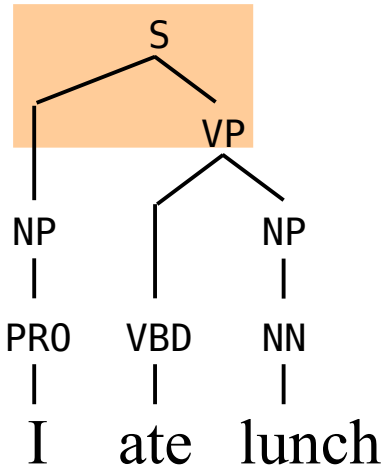
Hierarchical Phrase-Based MT

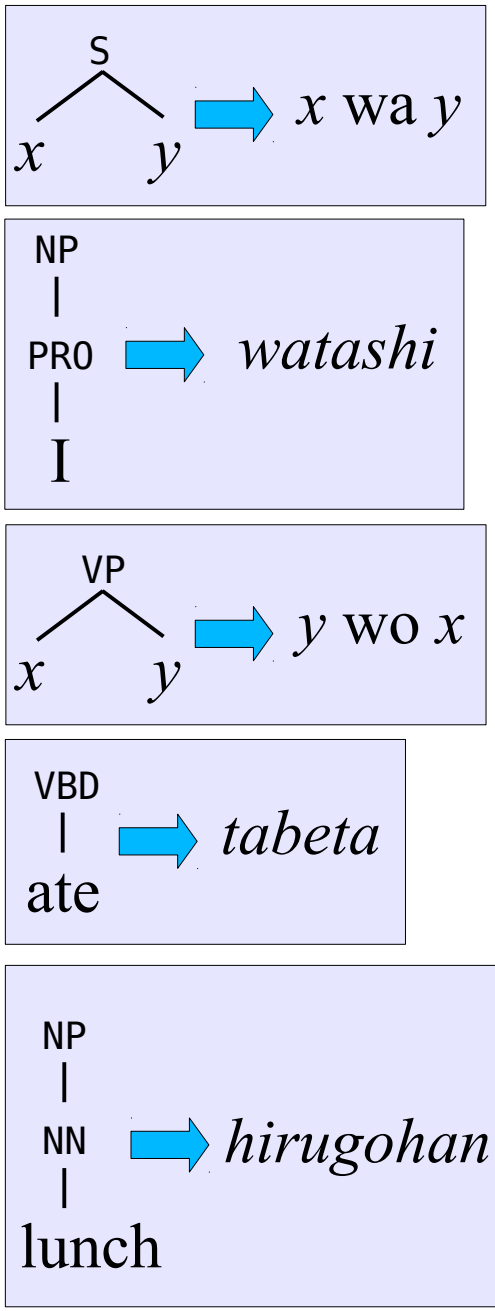
[Chiang, ACL05]

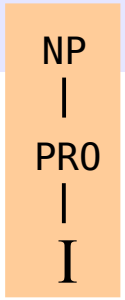


60

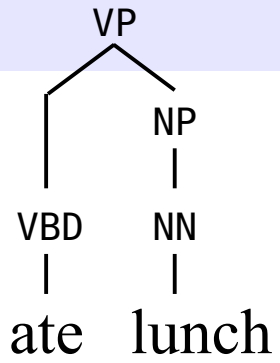
Syntax for MT



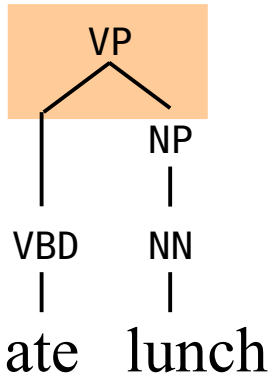




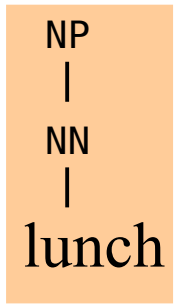
wa



watashi wa



watashi wa



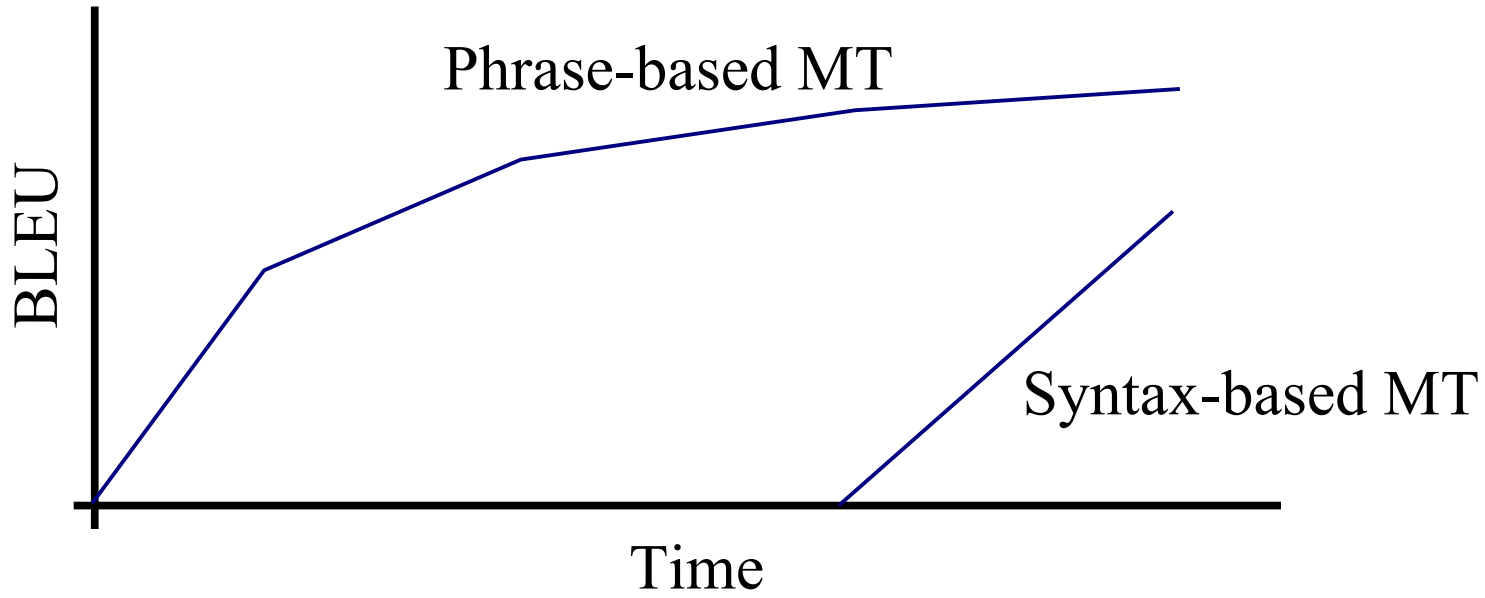
wo

watashi wa hirugohan wo tabeta

Kevin Knight,
Daniel Marcu,
Ignacio Thayer,
Jonathan Graehl,
Jon May,
Steve DeNeefe

Syntax for MT

- Decoding:
  - Tree-to-tree/string automata
  - CKY parsing algorithm
- Rule learning:
  - Parsed English corpus
  - Aligned data (GIZA++)
  - Extract rules and assign probabilities



Summary

- Old school tranislation = interlingua
  - Works well for limited domains
  - Costs a lot of money
- New school translation = statistical
  - Started out naïve
  - Becoming more linguistically motivated every year
- Translation is currently the “hot topic” in NLP
  - It looks like linguistics really is going to help, after all
  - (so long as you use it wisely in conjunction with statistics)