Probabilistic approaches to language and language learning

John Goldsmith

The University of Chicago

This work is based on the work of too many people to name them all directly.

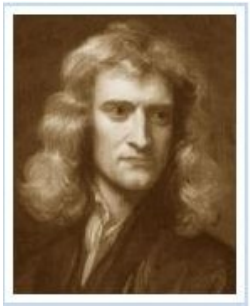
Nonetheless, I must specifically acknowledge Jorma Rissanen (MDL), Michael Brent and Carl de Marcken (applying MDL to word discovery), and Yu Hu, Colin Sprague, Jason Riggle, and Aris Xanthos, at the University of Chicago.

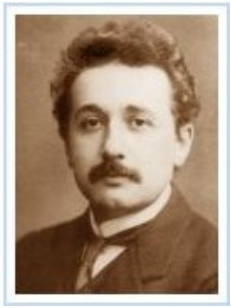
How can it be innovative —much less subversive — to propose to use statistical and probabilistic methods in a scientific analysis in the year 2006 Anno Domini?

1. Rationalism and empiricism—and modern science.
2. The mystery of the *synthetic apriori* is still lurking.
3. Universal grammar is a fine scientific hypothesis, but not a good *synthetic a priori.*
4. Grammar construction as *maximum a posteriori probability*.

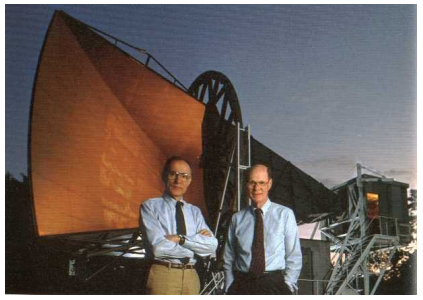
1. The development of modern science

The surprising effectiveness of mathematics in understanding the universe.





The reasonable effectiveness of understanding the universe by observing it carefully.



Rationalism	Empiricism
The effectiveness of mathematical models of the universe, and the mind's ability to develop abstract models, and make predictions from them. Trust the mind.	The effectiveness of observing the universe even when what we see is not what we expected. Especially then. Trust the senses.

Francis Bacon

Those who have handled sciences have been either men of experiment or men of dogmas.

The men of experiment are like the **ant**, they only collect and use; the reasoners resemble **spiders**, who make cobwebs out of their own substance.

But **the bee** takes a middle course: it gathers its material from the flowers of the garden and of the field, but transforms and digests it by a power of its own.

Not unlike this is the true business of **philosophy**; for it neither relies solely or chiefly on the powers of the **mind**, nor does it take the matter which it gathers from natural history and mechanical experiments and lay it up in the memory whole, as it finds it, but lays it up in the understanding altered and digested.

The collision of rationalism and empiricism

Kant's synthetic apriori:

The proposal that there exist contentful truths knowable independent of experience.

They are accessible because the very possibility of mind presupposes them.

Space, time, causality, induction.

2. Synthetic apriori

The problem is still lurking.

Efforts to dissolve it have been many.

One method, in both linguistics and psychology, is to *naturalize* it: to view it as a scientific problem.

"The problem lies in the object of study:

the human brain."

Synthetic apriori

The mind's construction of the world is its best understanding of what the senses provides it with.

$$World = \underset{world_i \in possible\,worlds}{\arg\max} \; pr(world_i \mid observations)$$

The real world is the one which is most probable, given our observations.

Bayesian,
maximum a posteriori reasoning

D = Data
H = Hypothesis

Bayes' Rule

$$pr(H \mid D) = \frac{pr(D \mid H)\, pr(H)}{pr(D)}$$

D = Data
H = Hypothesis

Bayes' Rule

$$pr(H \mid D) = \frac{pr(D \mid H)\, pr(H)}{pr(D)}$$

$$pr(H \mid D)\, pr(D) = pr(D \text{ and } H) = pr(D \mid H)\, pr(H)$$

Definition

Define pr(A|B) = pr(A&B)/pr(B)

D = Data
H = Hypothesis

Bayes' Rule

$$pr(H \mid D) = \frac{pr(D \mid H)\, pr(H)}{pr(D)}$$

$$pr(H \mid D)\, pr(D) = pr(D \text{ and } H) = pr(D \mid H)\, pr(H)$$

Definition

Definition

D = Data
H = Hypothesis

Bayes' Rule

$$pr(H \mid D) = \frac{pr(D \mid H)\, pr(H)}{pr(D)}$$

$$pr(H \mid D)\, pr(D) = pr(D \text{ and } H) = pr(D \mid H)\, pr(H)$$

Definition

Definition

$$pr(H \mid D)\, pr(D) = pr(D \mid H)\, pr(H)$$

D = Data
H = Hypothesis

Bayes' Rule

$$pr(H \mid D) = \frac{pr(D \mid H)\, pr(H)}{pr(D)}$$

$$pr(H \mid D)\, pr(D) = pr(D \mid H)\, pr(H)$$

$$\downarrow$$

$$pr(H \mid D) = \frac{pr(D \mid H)\, pr(H)}{pr(D)} \quad \checkmark$$

If reality is the most probable hypothesis, given the evidence...

we must find the hypothesis for which the following is a maximum:

D = Data
H = Hypothesis

$$pr(D \mid H)\, pr(H)$$

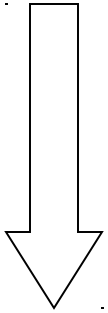
How do we calculate the probability
of our hypothesis about what reality is?

rationalism

How do we calculate the probability of our observations, given our understanding of reality?

empiricism

How do we calculate the probability
of our hypothesis about what reality is?

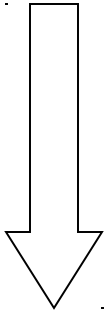


Assign a (“prior”) probability to all hypotheses, based on their coherence.

Measure the coherence.

Call it *an evaluation metric*.

How do we calculate the probability of our observations, given our understanding of reality?



Insist that your grammars be probabilistic: they assign a probability to their generated output.

Generative grammar

Construct an evaluation metric:

Choose the grammar which best satisfies the evaluation metric, as long as it somehow matches up with the data.

Generative grammar satisfies the rationalist need.

Generative grammar

Construct an evaluation metric:

Choose the grammar which best satisfies the evaluation metric, as long as it somehow matches up with the data.

Generative grammar satisfies the rationalist need.

It fails to say anything at all about the empiricist need.

Assigning probability to algorithms

after Solomonoff, Chaitin, Kolmogorov

The probability of an algorithm ...related to... the length of its most compact expression

$$\log \mathrm{pr}(A) = - \mathrm{length}(A)$$

$$\mathrm{pr}(A) = 2^{-\mathrm{length}(A)}$$

Assigning probability to algorithms

after Solomonoff, Chaitin, Kolmogorov

The probability of an algorithm

...related to...

the length of its most compact expression

$\log pr(A) = - length(A)$

$pr(A) = 2^{-length(A)}$

Assigning probability to algorithms

after Solomonoff, Chaitin, Kolmogorov

The probability of an algorithm

...related to...

the length of its most compact expression

$$\log \mathrm{pr}(A) = - \mathrm{length}(A)$$

$$\mathrm{pr}(A) = 2^{-\mathrm{length}(A)}$$

Assigning probability to algorithms

after Solomonoff, Chaitin, Kolmogorov

The probability of an algorithm

...related to...

the length of its most compact expression

$\log pr(A) = - length(A)$

$pr(A) = 2^{-length(A)}$

The promise of this approach is that it offers an *apriori* measure of complexity expressed in the language of probability

Let’s get to work and write some grammars.

We will make sure they all assign probabilities to our observations.

We will make sure we can calculate their length.

Then we know how to rationally pick the best one...

The real challenge for the linguist
is to see if this methodology will lead to the automatic discovery of structure that we already know is there.

To maximize
pr(Grammar)*pr(Data|Grammar)

we maximize

log pr(Grammar)+log pr (Data|Grammar)

or *minimize*

-log pr(Grammar)–log pr(Data|Grammar)

or minimize

Length(Grammar) – log pr(Data|Grammar)

An observation:

thedogsawthecatandthecatsawthedog

An observation:

thedogsawthecatandthecatsawthedog

What is its probability?

An observation:

thedogsawthecatandthecatsawthedog

What is its probability?

Its probability depends on
the model we propose.
The mind is active.
The mind chooses.

An observation:

thedogsawthecatandthecatsawthedog

What is its probability?

If we only know that the language has *phonemes*, we can calculate the probability based on *phonemes*.

Phonological structure

(1)The probability of a phoneme can be calculated independent of context; or

(2) We can calculate a phoneme's probability conditioned by the phoneme that precedes it.

Phonological structure

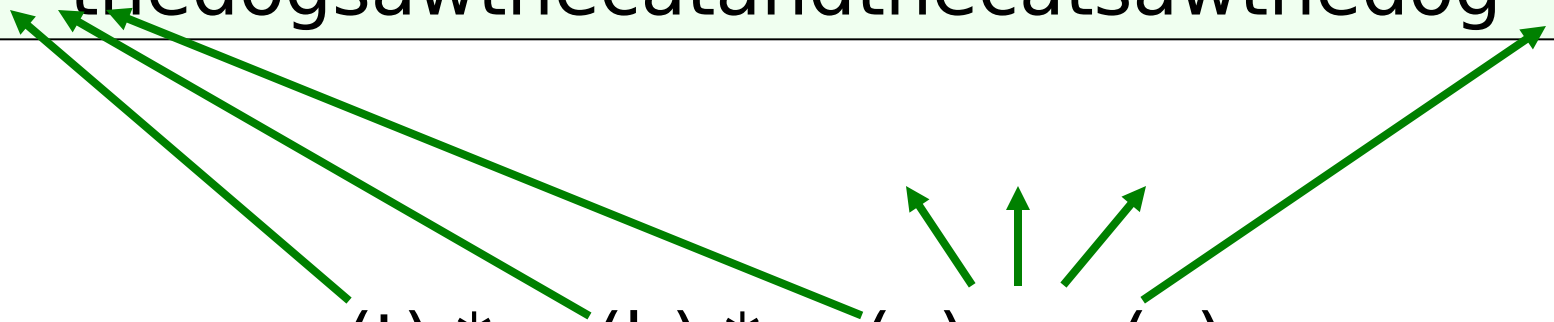
(1)The probability of a phoneme can be calculated independent of context; or

(2) We can calculate a phoneme's probability conditioned by the phoneme that precedes it.

To make life simple for now, we choose (1).

Probability of our observation:

thedogsawthecatandthecatsawthedog



pr(t) * pr(h) * pr(e)…pr(g)

Multiply the probability of all ***33*** letters.

$$=2.04*10^{-33}$$

D = Data
H = Hypothesis

$$pr(D \mid H)\, pr(H)$$

We have pr(D|H): probability of the data *given the phoneme hypothesis.*

What is the probability of the phoneme hypothesis: pr(H)?

D = Data
H = Hypothesis

$$pr(D|H)pr(H)$$

We have pr(D|H): probability of the data *given the phoneme hypothesis.*

What is the probability of the phoneme hypothesis: pr(H)?

We interpret that as the question:

What is the probability of a system with

11 distinct phonemes?

D = Data
H = Hypothesis

$$pr(D \mid H)\, pr(H)$$

We have pr(D|H): probability of the data *given the phoneme hypothesis.*

What is the probability of the phoneme hypothesis: pr(H)?

We interpret that as the question:

What is the probability of a system with 11 distinct phonemes?

Π(11) = Prob[Phoneme Inventory (Lg)=11]

D = Data
H = Hypothesis

$$pr(D|H)pr(H)$$

We have pr(D|H): probability of the data *given the phoneme hypothesis.*

What is the probability of the phoneme hypothesis: pr(H)?

And is there a better hypothesis available, anyway?

Yes, there is.

The *word* hypothesis:

There is a vocabulary in this language:

- the
- dog
- saw
- cat
- and

The *word* hypothesis:

There is a vocabulary in this language:

The *words* have frequencies:

the	4/11
dog	2/11
saw	2/11
cat	2/11
and	1/11

and the observation's probability is the produc
of *11* probabilities...

the dog saw the cat and the cat saw the dog

$$probability = \frac{4}{11}\frac{2}{11}\frac{2}{11}\frac{4}{11}\frac{2}{11}\frac{1}{11}\frac{4}{11}\frac{2}{11}\frac{2}{11}\frac{4}{11}\frac{2}{11}$$

$$= 5.74 * 10^{-8}$$

which is much, *much* bigger than **2.04*10-**

D = Data
H = Hypothesis

We *need* to calculate:

$$pr(D \mid H)\,pr(H)$$

on the word model

We just calculated:

$$pr(D \mid H)$$

so *now* we need to calculat

$$pr(H)$$

the probability of the lexicon

The probability of this lexicon:

the
dog
saw
cat
and

generated by this alphabet:

a	0.15
c	0.05
d	0.1
e	0.05
g	0.05
h	0.05
n	0.05
o	0.05
s	0.05
t	0.1
w	0.05
#	0.25

Probability = $1.29 * 10^{-20}$

D = Data
H = Hypothesis

We *need* to calculate:

$$pr(D \mid H)\, pr(H)$$

We just calculated:

$pr(D \mid H)$

the probability of the data, given the lexicon

$5.74 * 10^{-8}$

$pr(H)$

the probability of the lexicon

$1.29 * 10^{-20}$

D = Data
H = Hypothesis

We *need* to calculate:

$$pr(D \mid H)\, pr(H)$$

We just calculated:

$pr(D \mid H)$

the probability of the data, given the lexicon

$5.74 * 10^{-8}$

$pr(H)$

the probability of the lexicon

$1.29 * 10^{-20}$

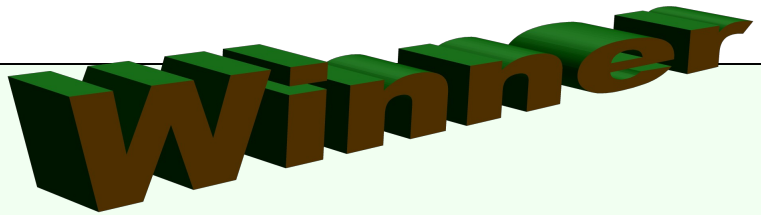
Probability of data under *lexicon* hypothesis

Product = $7.39 * 10^{-28}$

Probability of data under *letter* hypothesis

$2.04 * 10^{-33}$

D = Data
H = Hypothesis



Probability of data under *lexicon* hypothesis

$Product = 7.39 * 10^{-28}$

Probability of data under *letter* hypothesis

$2.04 * 10^{-33}$

How do we scale up to *grammar?*

0. Word discovery: Brent, de Marcken
1. Morpheme discovery
2. Phonology discovery
3. Word-category discovery
4. Grammar discovery

How do we scale up to *grammar?*

0. Word discovery
1. Morpheme discovery

   http://linguistica.uchicago.edu
2. Phonology discovery
3. Word-category discovery
4. Grammar discovery

Very high level overview of calculating the complexity of a morphology

A morphology is a *finite state device*, and

transitions between states are labeled by

morphemes.

Its *length* is much smaller than that of a corresponding word list (=lexicon).

Capturing redundancies shortens grammars

$$\left\{\begin{matrix} jump \\ walk \end{matrix}\right\}\left\{\begin{matrix} NULL \\ ed \\ ing \end{matrix}\right\}$$

length = 14

jump jumped
jumping

walk walked
walking

length = 34

Calculating the size of the morphology

$$\textit{Suffix list} \quad \sum_{f \in Suffixes} \left(\lambda^* |f| + \log \frac{[W_A]}{[f]} \right)$$

$$\textit{Stem list}: \quad \sum_{t \in Stems} \left(\lambda^* |t| + \log \left(\frac{[W]}{[t]} \right) \right)$$

Number of letters

structure

+ Signatures, which we'll get to on the next slide.

Information contained in the Signature component

$$\sum_{\sigma \in Signatures} \log \frac{[W]}{[\sigma]}$$

list of pointers to signatures

$$+ \sum_{\sigma \in \text{Signatures}} \log < \text{stems}(\sigma) > + \log < suffixes(\sigma) >$$

$$+ \sum_{\sigma \in Sigs} \left(\sum_{t \in Stems(\sigma)} \log \frac{[W]}{[t]} + \sum_{f \in Suffixes(\sigma)} \log \frac{[\sigma]}{[f in \sigma]} \right)$$

<X> indicates the number of distinct elements in X

How do we scale up to *grammar?*

[0. Word discovery]

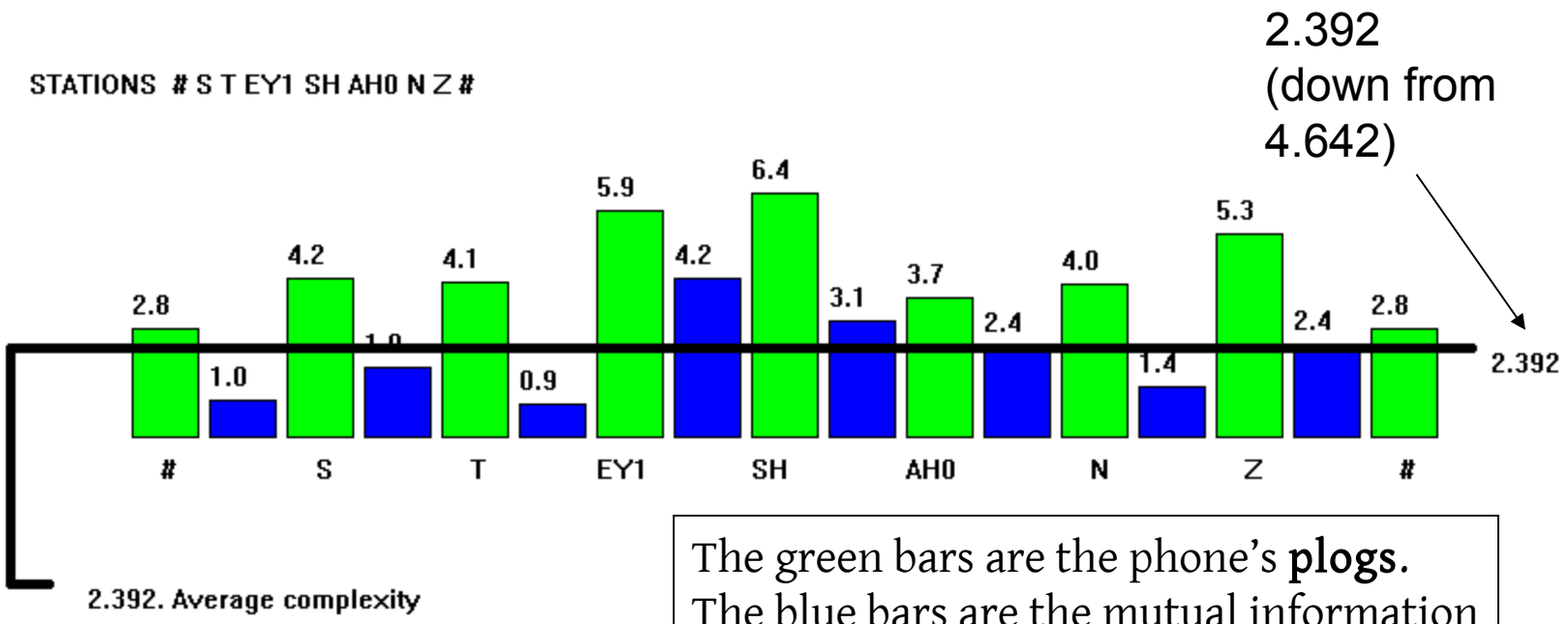
1. Morpheme discovery
2. Phonology discovery
3. Word-category discovery
4. Grammar discovery

Capturing phonological regularities increases the probability of the data.

Let's look at mutual information graphically

Every pair of adjacent phonemes
is attracted to every one of its neighbors.

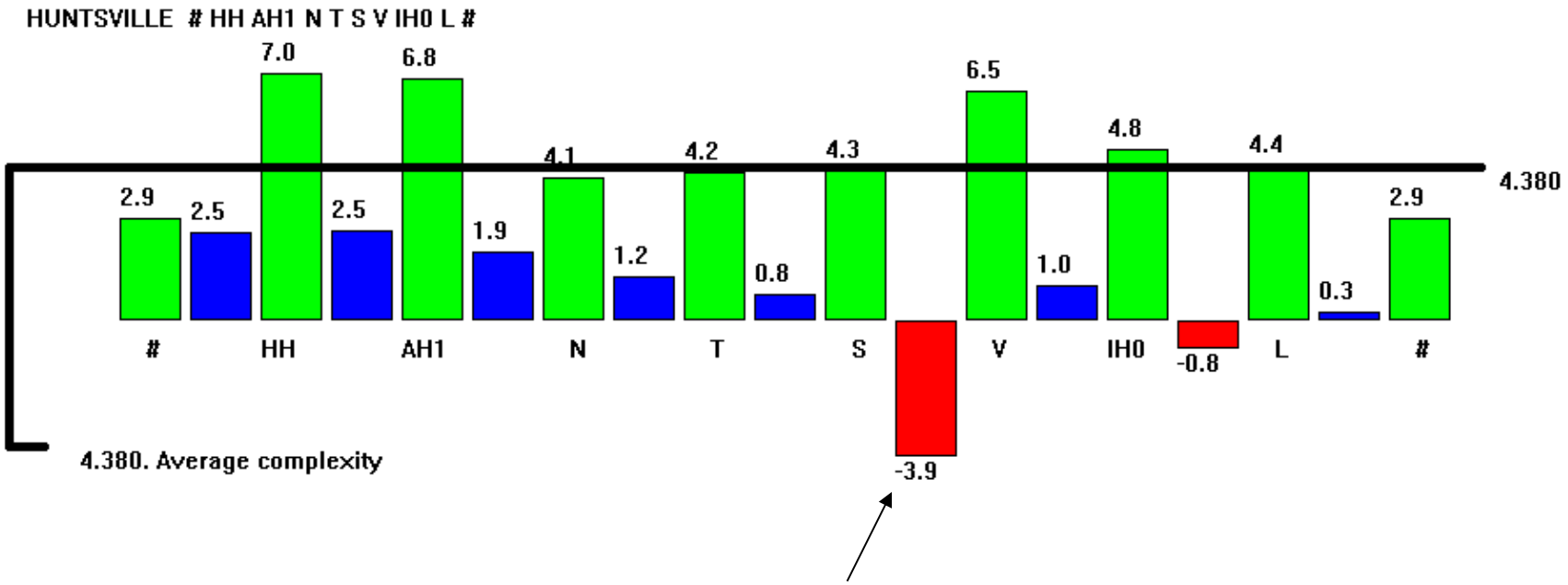
"stations"



The green bars are the phone's **plogs**.
The blue bars are the mutual information
(the stickiness) between the phones.

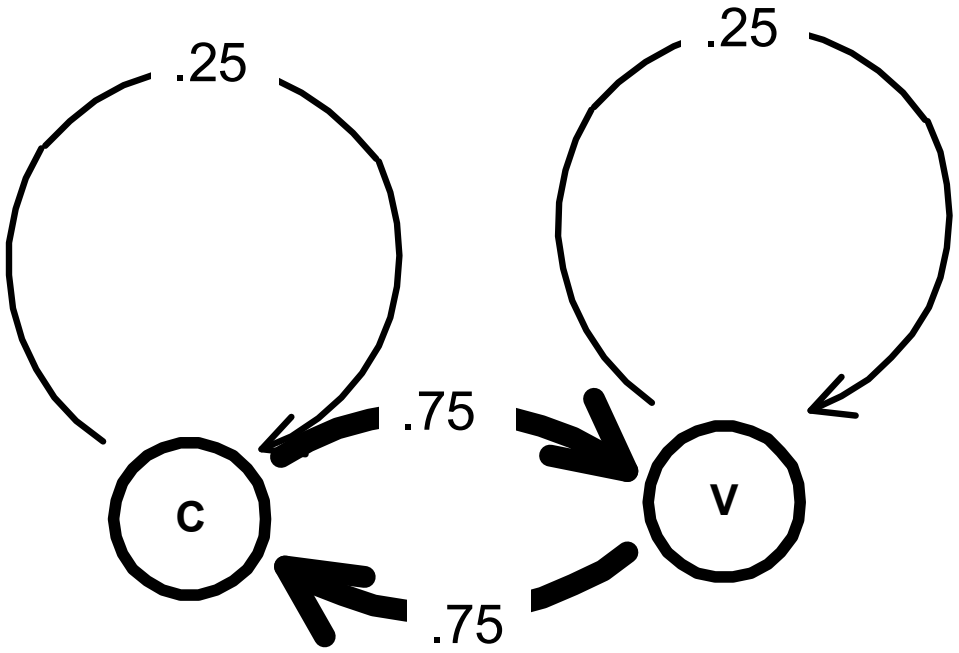
Example with negative mutual information:

"HUNTSVILLE"



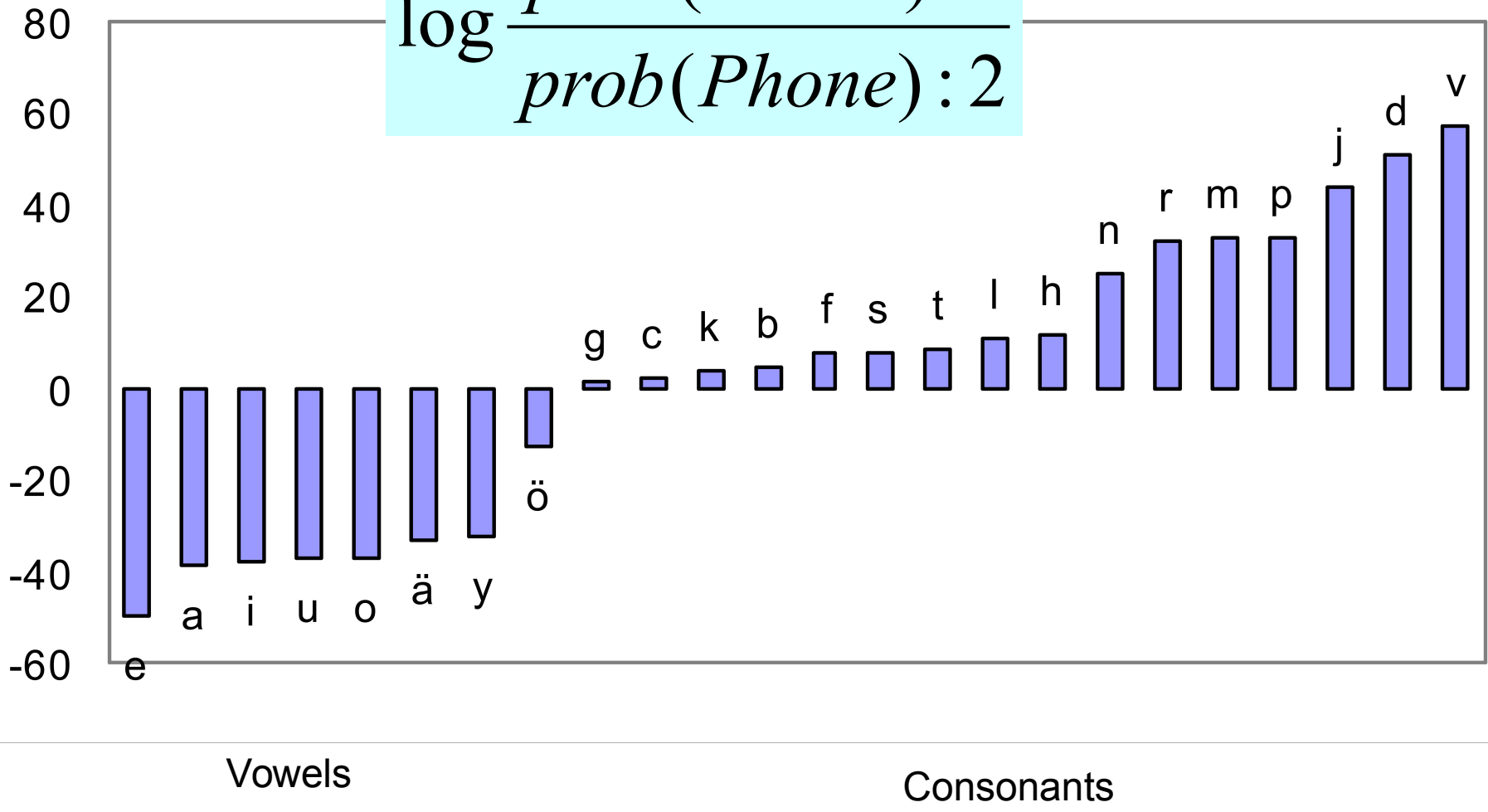
The mutual information can be negative – if the frequency of the phone-pair is *less than* would occur by chance.

Transition probabilities (Finnish): Learned by an HMM

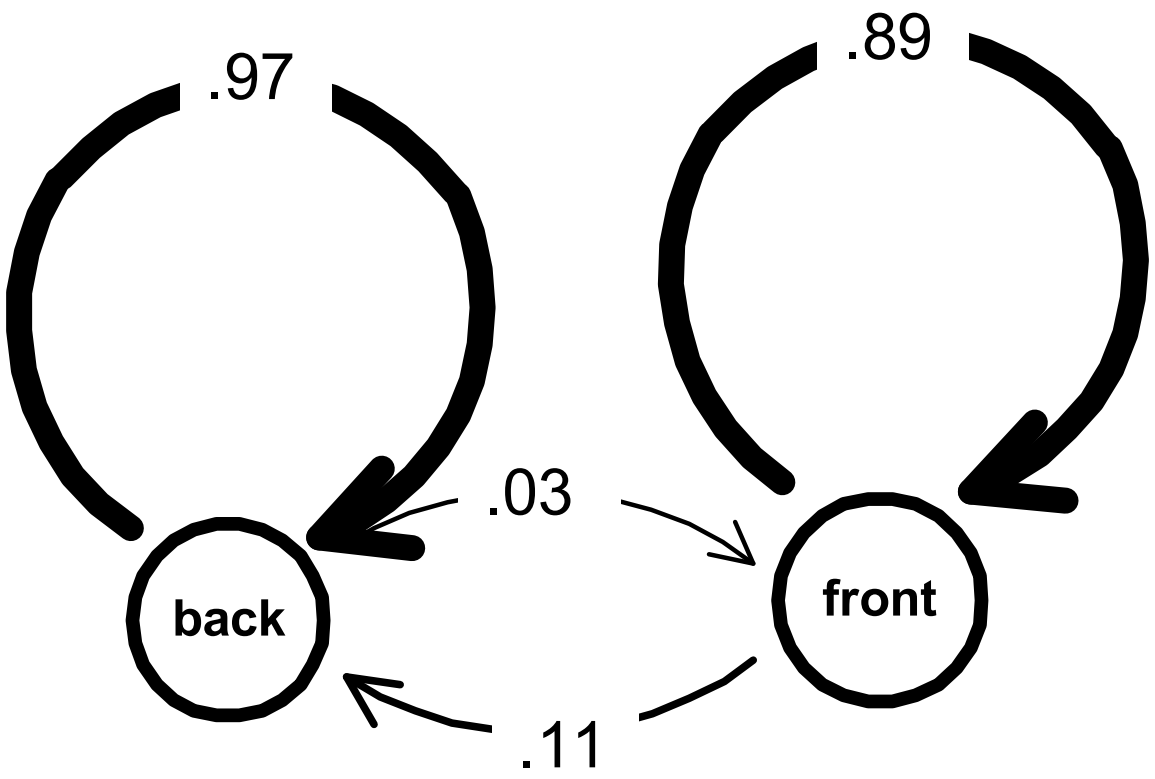


Log probability C/V Finnish segments

$$\log \frac{prob(Phone):1}{prob(Phone):2}$$



Vowel harmony



Find the best two-state Markov model to generate Finnish vowels

The HMM divides up the vowels like this:

Back vowels

State 1	Probability
a	0.353305
i	0.215194
u	0.158578
e	0.139881
o	0.133042
y	7.71E-15
ö	1.60E-18
ä	1.51E-18

Front vowels

State 2	Probability
i	0.266105
ä	0.255554
e	0.254647
y	0.157373
ö	0.050579
o	0.014794
a	0.000647
u	0.000302

Phonological models

They need not be “local”; they can be structural, and “distant”, in the sense of autosegmental and metrical phonology.

How do we scale up to *grammar?*

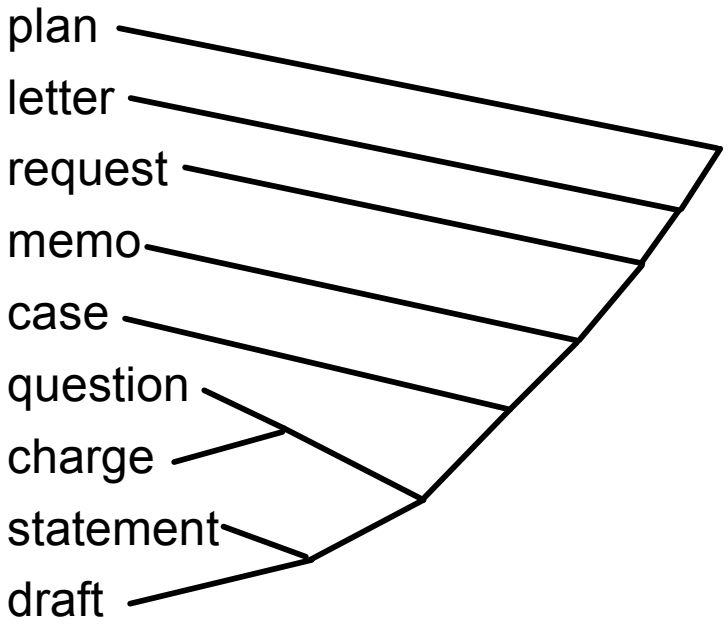
[0. Word discovery]

1. Morpheme discovery
2. Phonology discovery
3. Word-category discovery
4. Grammar discovery

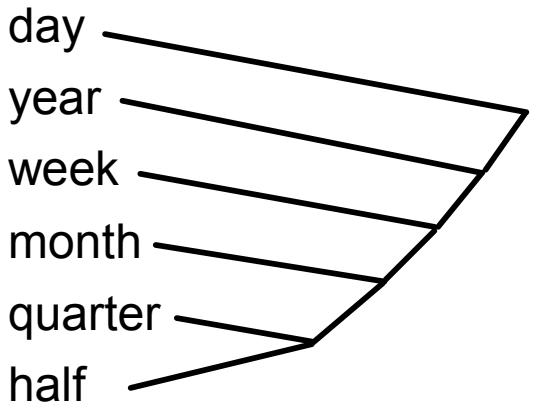
Category induction

Much of it in the context of hidden Markov models and statistical machine translation.

The first classic study by Brown *et al.*, the IBM statistical translation group:



Examples of categories induced by distribution (Brown et al.



How do we scale up to *grammar?*

[0. Word discovery]

1. Morpheme discovery
2. Phonology discovery
3. Word-category discovery
4. Grammar discovery

Much work here in the last 20 years

Much of it under the rubric of *language modeling;*

Some as *grammar induction.*

This is hard (but so is the rest).

Part of the problem is in inducing phrase-structure; part of it is dealing with the syntax of grammatical agreement patterns.

How can it be innovative —much less subversive — to propose to use statistical and probabilistic methods in a scientific analysis in the year 2006 Anno Domini?

Answer:

It is innovative and subversive:

not because we use probability— but because this allows in a new synthetic apriori, MAP (maximum aposteriori probability).

We can ***reject*** the false dilemma: either linguistics is psychology, or linguistics is a (silly) game.

Linguistics is a science of language data with one right, and many wrong, answers.

Conclusion

The linguistic question: can we use the principle:

Maximize the probability of the data

as our sole scientific maxim?

Can we thus dispense with the need for a substantive Universal Grammar? (Yes.)

What are the consequences for psychologists if this is so?

The End

Shift from generative grammar

Chomsky, Language and Mind (Future):

p. 76: No one who has given any serious thought to the problem of formalizing inductive procedures or “heuristic methods” is likely to set much store by the hope that such a system as a generative grammar can be constructed by methods of any generality.

p. 76-7: A third task is that of determining just what it means for a hypothesis about the generative grammar of a language to be “consistent” with the data of sense. Notice that it is a great oversimplification to suppose that a child must discover a generative grammar that accounts for all the linguistic data that has been presented to him and that “projects” such data to an infinite range of potential sound-meaning relations.…The third subtask, then, is to study what we might think of as the problem of “confirmation”—in this context, the problem of what relation must hold between a potential grammar and a set of data for this grammar to be confirmed as the actual theory of the language in question.