Information theory and phonology

John Goldsmith
The University of Chicago

All the particular properties that give a language its unique phonological character can be expressed in numbers.
-Nicolai Trubetzkoy

Outline

1. What is phonology? What is an information theoretic approach to phonology?
2. A brief history of Probability and Information theory
3. Trubetzkoy's conception of probabilistic phonology.
4. Basic notions: probability, positive log probability (plog); mutual information; entropy.
5. Establishing the "force field" of a language: Soft phonotactics.
6. Categorization through hidden Markov models: discovering Finnish vowel harmony automatically.

1. What is phonology?

- The study of
  - the inventory and possible combinations of discrete sounds in words and utterances in natural languages: in a word, ***phonotactics***;
  - the "modifications"—alternations–between sounds occasioned by the choice of morphs and phones in a word or utterance: in a word, ***automatic alternations***.
  - We will focus on the first, today.

Probabilistic analysis

- A probabilistic analysis aims at taking a set of data as its input, and
- Finding the optimal set of ***values*** for a ***fixed set of parameters***, where the investigator sets the fixed set of parameters ahead of time; the *method* allows one to find the *best values*, given the data. The analysis (set of values) then makes predictions beyond the input data.

Probabilistic phonology

- What it is:
  - Specification of a set of parameterized variables, with a built-in objective function (that which we wish to optimize): typically it is the *probability of the data*.
- What it is *not*:
  - An effort to ignore phonological structure.

Redundancies = generalizations

A set of data does not come with a probability written on it; that probability is derived from a model.

A model that "extracts" regularities from the data will—by definition—assign a higher probability to the data.

The goal is to find the model that assigns the highest probability to the data, all other things being equal.

The goal of automatic discovery of grammar

- My personal commitment is to the development of algorithmic approaches to automatic learning (by computer) of phonology, morphology, and syntax that do algorithmically what linguists do intuitively.
- I believe that the best way to accomplish this is to develop probabilistic models that maximize the probably of the data.
- http://linguistica.uchicago.edu
  Linguistica project

2. Brief history of probability

Blaise Pascal (1632-1662)

Pierre de Fermat (1601-1665)

Beginnings of work on probability: for *gambling*.

Pierre de Laplace

(1749-1827)

First application to major scientific problems:
theory of errors,
actuarial mathematics, and
other areas.

19th century: the era of probability in physics

The central focus of 19th century physics was on *heat*, *energy*, and the *states of matter* (gas, liquid, solid, e.g.).

Kinetic theory of gases vs. caloric theory of heat.

Principle of conservation of energy.

19th century physics

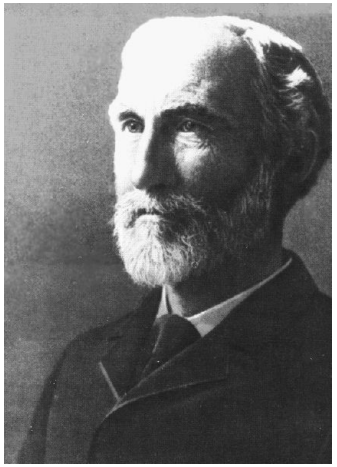
Rudolf Clausius:

development of notion of *entropy:* there exists no thermodynamic transformation whose sole effect is to extract a quantity of heat from a colder reservoir to a hotter one.

19th century

Ludwig Boltzmann (1844-1906): 1877: Develops a probabilistic expression for entropy.

Willard Gibbs: American (1839-1903)



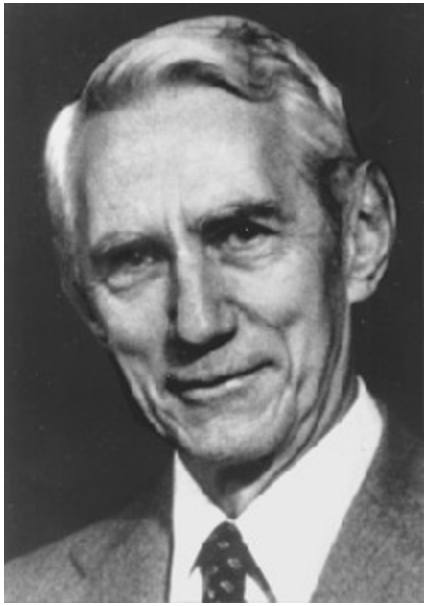
Quantum mechanics

- All observables are based on the probabilistic collapse of the wave function (Schrödinger): physics becomes probabilistic down to its core.

Entropy and information

Leo Szilard

Claude Shannon



Norbert Wiener

Shannon is most famous for

- His definition of *entropy*, which is the **average (positive) log probability** of the symbols in a system.

$$-\sum prob(w_i)\log prob(w_i)$$

- This set the stage for a quantitative treatment of symbolic systems.

Any expression of the form $\sum_i prob(w_i)F(x)$ is a weighted average of the F-values:

3. History of probabilistic phonology

- Early contribution by Count Nicolas Trubetzkoy
- Chapter 9 of *Grundzüge der Phonologie* (*Principles of Phonology*) 1939:
  - Chapter 7 "On statistical phonology"
  - Cites earlier work by Trnka, Twaddell, and George Zipf ("Zipf's Law").

Trubetzkoy's basic idea

"Statistics in phonology have a double significance. They must show the *frequency* with which an element (a phoneme, group of phonemes, [etc.]) appears in the language, and also the importance of the functional productivity of such an element or an opposition. (VII.1.)...

VI.4

"The absolute value of phoneme frequencies is only of secondary importance. Only the relationship between the [*observed*] frequency and the *expected* frequency [given a model] possesses a real value. This is why the *determination of frequencies* in a text must be preceded by a careful calculation of probabilities, taking neutralization into account.

Chechen

- “Consider a language where a consonantal distinction is neutralized word-initially and word-finally. Thus the marked value can only appear in syllable-initial position except word-initially. If the average number of syllables per word is α, we expect the frequency of the unmarked to the marked to be $\frac{\alpha+1}{\alpha-1}$.”

[This is the case for geminate consonants in Chechen] where the average number of syllables per word is 1.9 syllables; thus the ratio of the frequency of geminates to non-geminates should be 9/29 (about 1/3). In fact, we find:

Chechen

tt:t	12:90	11%
qq:q	6:45	12%
ćć:ć	25:59	30%
ll:l	16:32	33%
All	*59:226*	*20%* *predicted: 31%*

Trubetzkoy follows with a similar comparison of glottalized to plain consonants, where the glottalized consonant appears only word-initially.

"We must not let ourselves be put off by the difficulties of such a calculation, because it is only by comparing observed frequencies to predicted frequencies that the former take on value."

Trubetzkoy's goal:

$$\frac{observed\ \ frequency\,(phoneme)}{predicted\ \ frequency(phoneme)}$$

Thus, for Trubetzkoy, a *model* generates a set of expectations, and when reality diverges from those expectation, it means that the language has its own expectations that differ from those of the linguist at present: and therefore more work remains to be done.

Essence of probabilistic models:

- Whenever there is a choice-point in a grammar, we must assign degrees of *expectedness* of each of the different choices.
- And we do this in a way such that these quantitites add up to 1.0.
- These are probabilities.

Frequencies and probabilities

- **Frequencies** are numbers that we observe (or count);
- **Probabilities** are parameters in a theory.
- We can set our probabilities on the basis of the (observed) frequencies; but we do not *need* to do so.
- We often do so for one good reason:

Maximum likelihood

A basic principle of empirical success is this:

> Find the probabilistic model that assigns the highest probability to a (pre-established) set of data (observations).

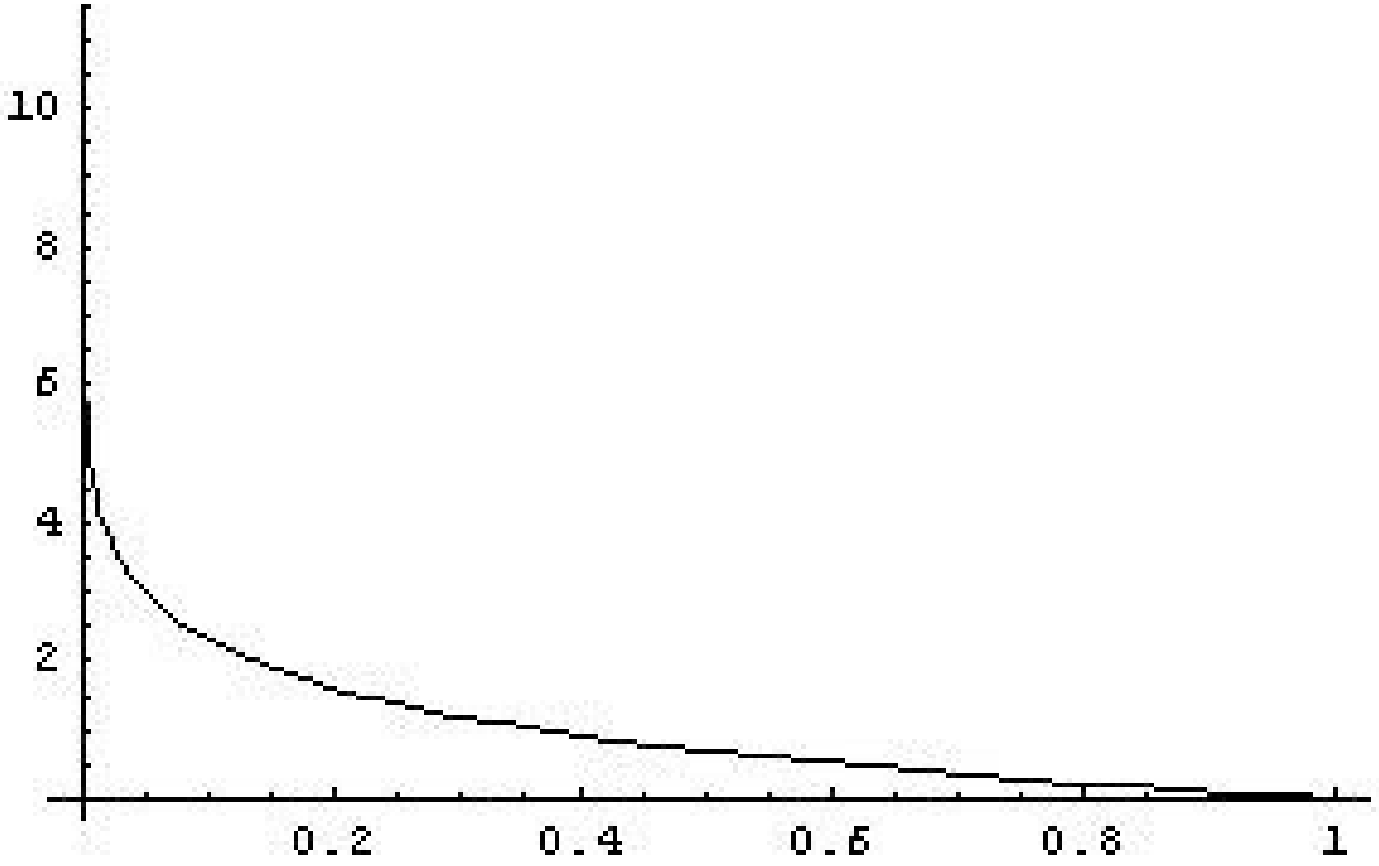
Maximize the probability of the data.

In simple models, this happens by setting the probability parameters to the observed frequencies.

Probability models as "scoring models"

- An alternative way to think of probabilistic models is as models that assign *scores* to representations: the higher the score, the worse the representation.
- The score is the logarithm of the inverse of the probability of the representation. We'll see why this makes intuitive sense....

Plog(x) = -log(x)=log(1/x)



The natural unit of plogs is the **bit.**

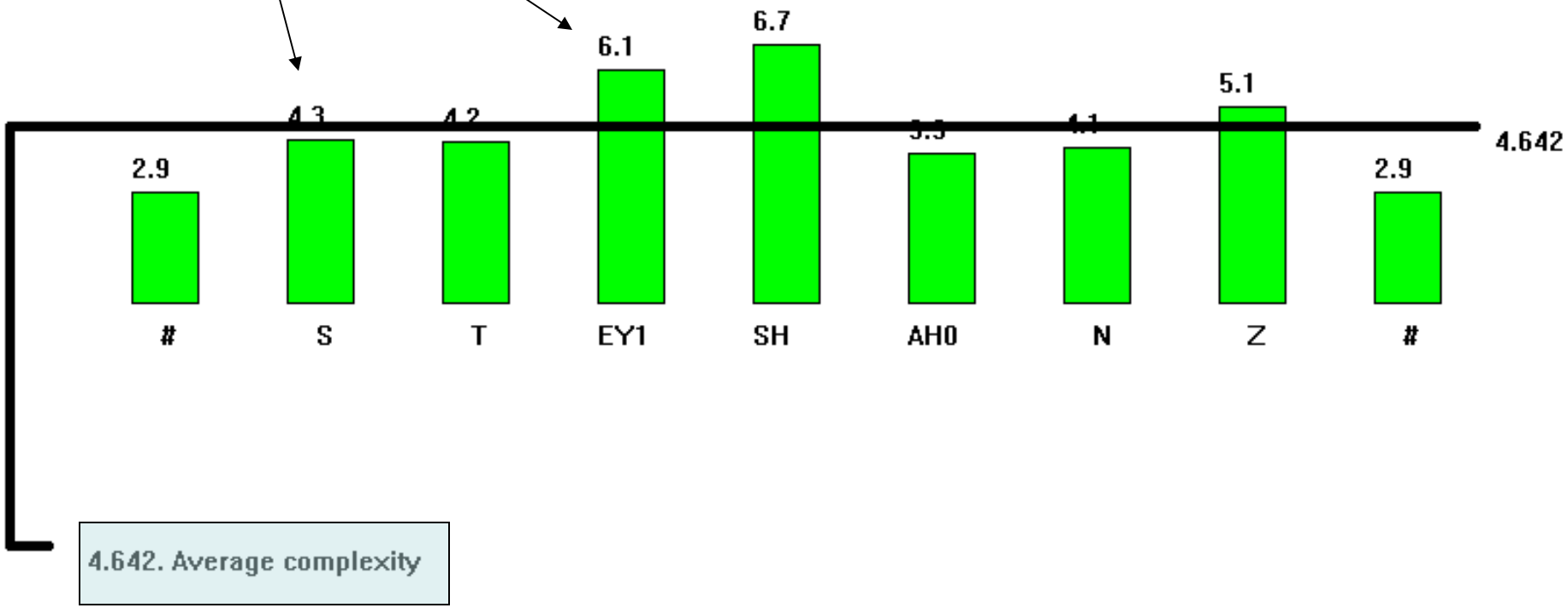
Don’t forget:
Maximize probability =
minimize plog

Unigram model
Type counts.

STATION'S # S T EY1 SH AH0 N Z #

Height of the bar indicates the "positive log frequency" of the phoneme:

$$\log \frac{1}{frequency(phoneme)} = -\log frequency(phoneme)$$



Phonemes of English

- Top of list

Phoneme	Count	P-Log Freq
#	63,023	2.9061
ə	32,374	3.871
n	28,494	4.055
t	25,975	4.189
s	24,885	4.2508
l	22,382	4.4037
r	22,250	4.4123
k	19.435	4.6074
d	17,062	4.7953

- Bottom of list

Phoneme	**Count**	**P-Log Freq**
h	**3778**	**6.9704**
uw	**3679**	**7.0087**
ǰ	**3308**	**7.1620**
ɚ	**2536**	**7.5406**
y	**2521**	**7.5540**
č	**2274**	**7.7028**
aw	**1534**	**8.2705**
θ	**1423**	**8.3791**
oy	**575**	**9.6864**

Simple segmental representations

- "Unigram" model for French (English, etc.)
- Captures only information about segment frequencies.
- The probability of a word is the product of the probabilities of its segments: or...
- The log probability is the sum of the log probabilities of the segments.
- Better still: the **complexity** of a word is its **average log probability**:

$$\frac{1}{length(W)} \sum_{i=1}^{length(W)} -\log_2 prob(w_i)$$

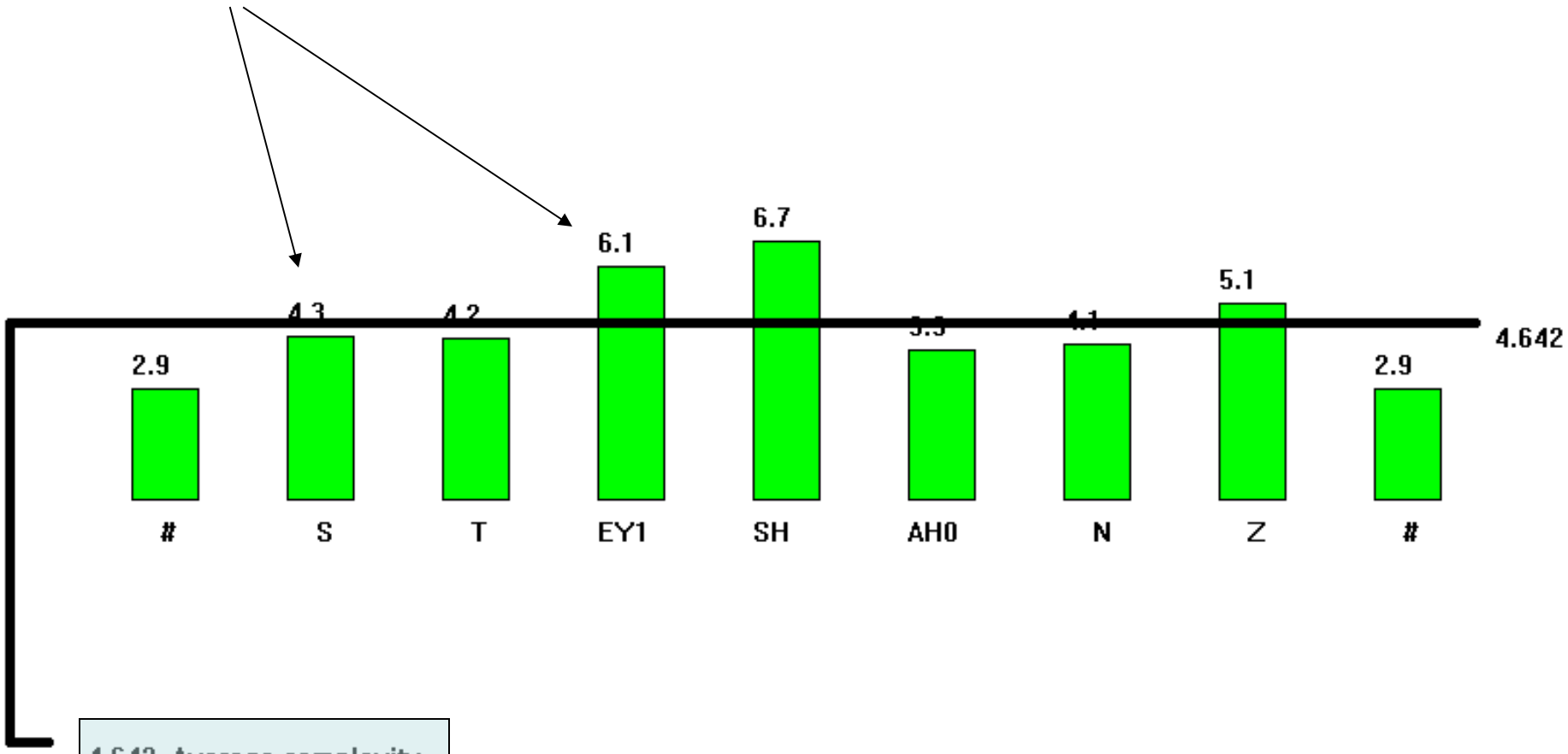
Let's look at that graphically...

- Because *log probabilities* are much easier to visualize.
- And because the log probability of a whole word is (in this case) just the sum of the log probabilities of the individual phones.
- The *plog* is a quantitative measure of *markedness* (Trubetzkoy would have agreed to that!).

Unigram model
Type counts.

STATION'S # S T EY1 SH AH0 N Z #

Height of the bar indicates the "positive log frequency" of the phoneme.



But we care greatly about the *sequences*

- For each pair, we compute:
  - the ratio of
    - the number occurrences *found* to
    - The number of occurrences expected (if there were no structure, i.e., if all choices were independent).

$$\frac{freq(ab)}{freq(a)\,freq(b)}$$

Trubetzkoy's ratio

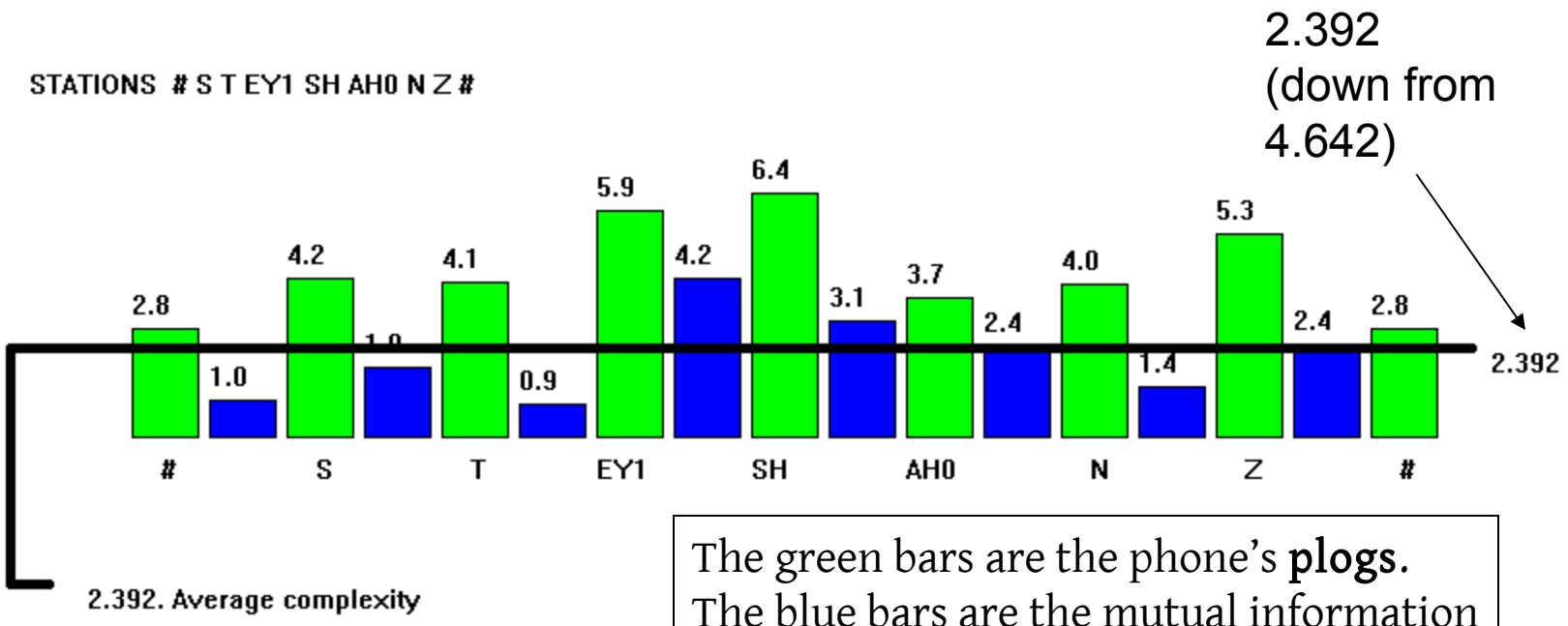
Or better still: $$\log \frac{freq(ab)}{freq(a)\,freq(b)}$$

Mutual information (a,b)

Let's look at mutual information graphically

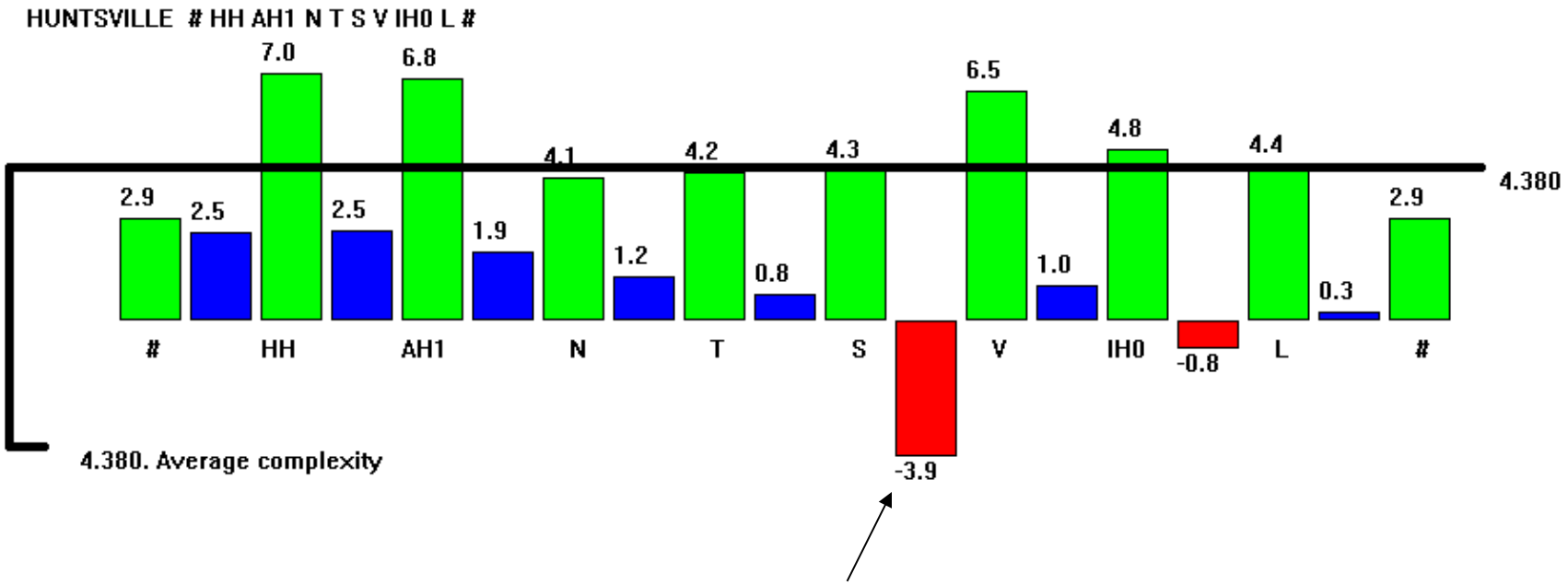
Every pair of adjacent phonemes
is attracted to every one of its neighbors.

"stations"

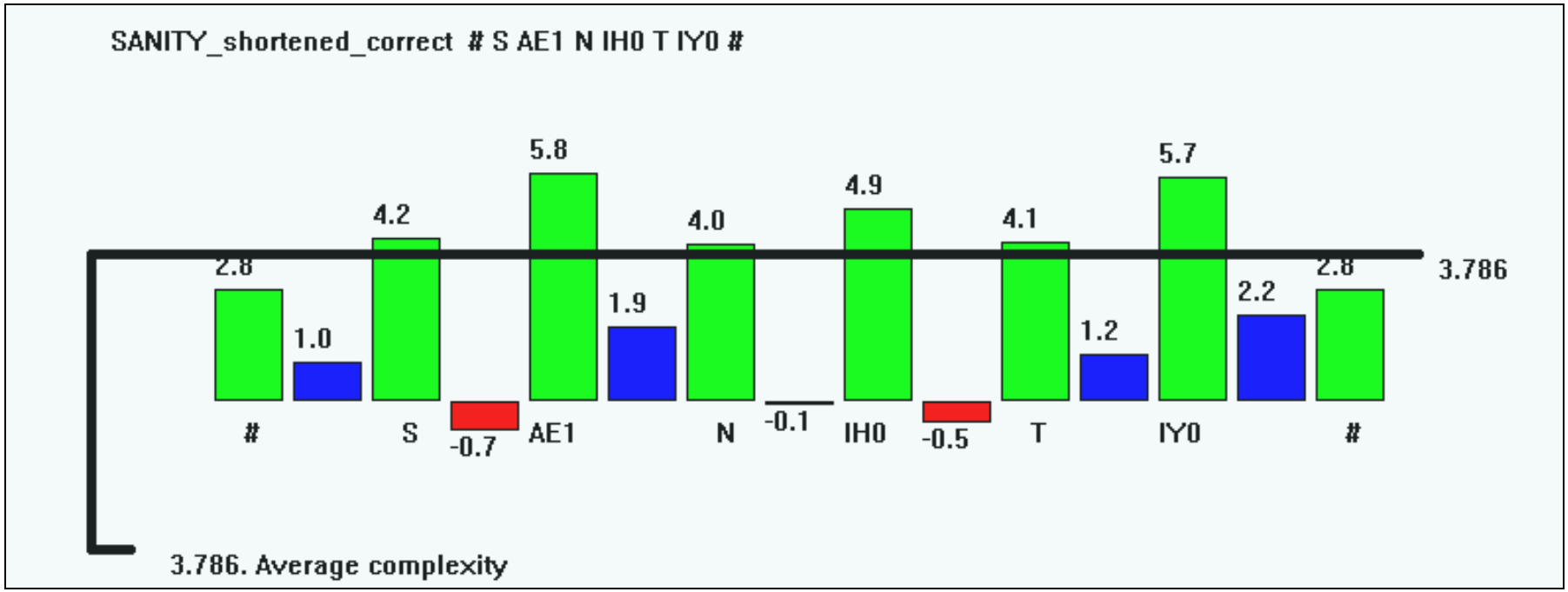


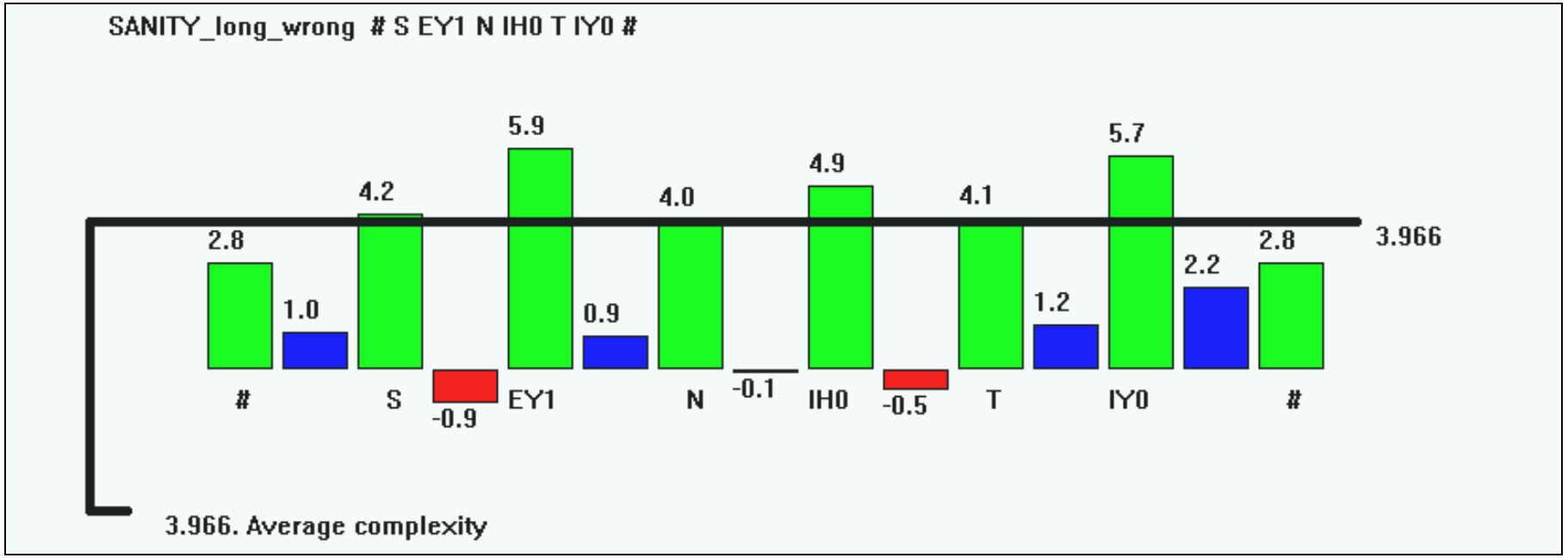
The green bars are the phone's **plogs**.
The blue bars are the mutual information
(the stickiness) between the phones.

Example with negative mutual information:



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


SANITY_shortened_correct # S AE1 N IH0 T IY0 #
2.8
1.0
4.2
-0.7
5.8
1.9
4.0
-0.1
4.9
-0.5
4.1
1.2
5.7
2.2
2.8
3.786
#
S
AE1
N
IH0
T
IY0
#
3.786. Average complexity


SANITY_long_wrong # S EY1 N IH0 T IY0 #
2.8
1.0
4.2
-0.9
5.9
0.9
4.0
-0.1
4.9
-0.5
4.1
1.2
5.7
2.2
2.8
3.966
#
S
EY1
N
IH0
T
IY0
#
3.966. Average complexity

Complexity = average log probability

- Rank words from a language by complexity:
  - Words at the top are the "best";
  - Words at the bottom are...what?

borrowings,
onomatopeia, rare phonemes,
short compounds,
foreign names,
and errors.

Top of the list:

- can
- stations
- stationing
- handing
- parenting
- warren’s
- station
- warring

Bottom of the list:

- A.I.
- yeah
- eh
- Zsa
- uh
- ooh
- Oahu
- Zhao
- oy
- arroyo

- We have, as a first approximation, a system with $P + P^2$ parameters: P plogs and P^2 mutual informations.
- The pressure for nativization is the pressure to rise in this hierarchy of words.
- We can thus define the direction of the phonological pressure...

Nativization of a word: a French example

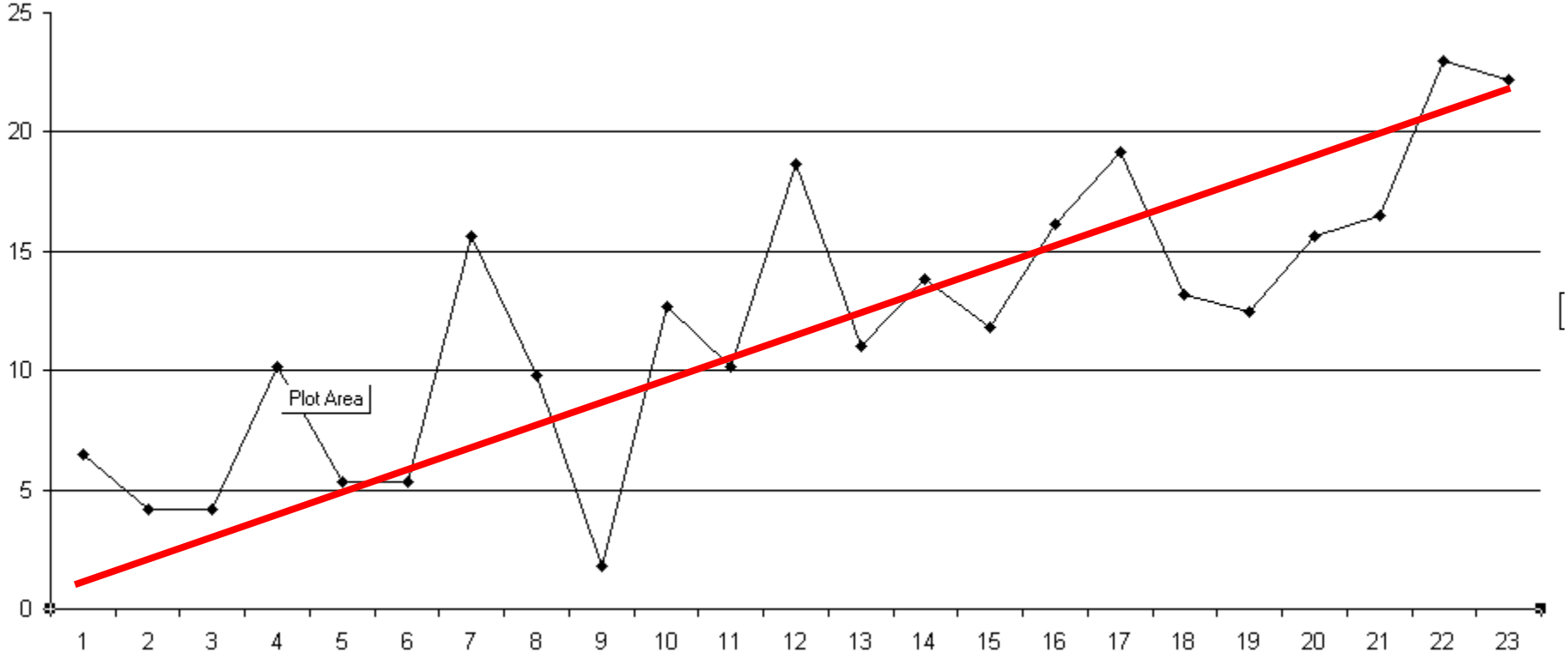
- *Gasoil* [gazojl] or [gazɔl]
- Compare average log probability (bigram model)
  - [gazojl] 5.285
  - [gazɔl] 3.979
- This is a huge difference.
- Nativization *decreases the average log probability of a word.*

Phonotactics

- Phonotactics include knowledge of 2nd order conditional probabilities.
- Examples from English...

This list was randomized, then given to students to rank:

1 stations
2 hounding
3 wasting
4 dispensing
5 gardens
6 fumbling
7 telesciences
8 disapproves
9 tinker
10 observant
11 outfitted
12 diphtheria
13 voyager
14 Schafer
15 engage
16 Louisa
17 sauté
18 zigzagged
19 Gilmour
20 Aha
21 Ely
22 Zhikov
23 kukje


25
20
15
10
5
0
Plot Area
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

Large agreement with average log probability (plog).

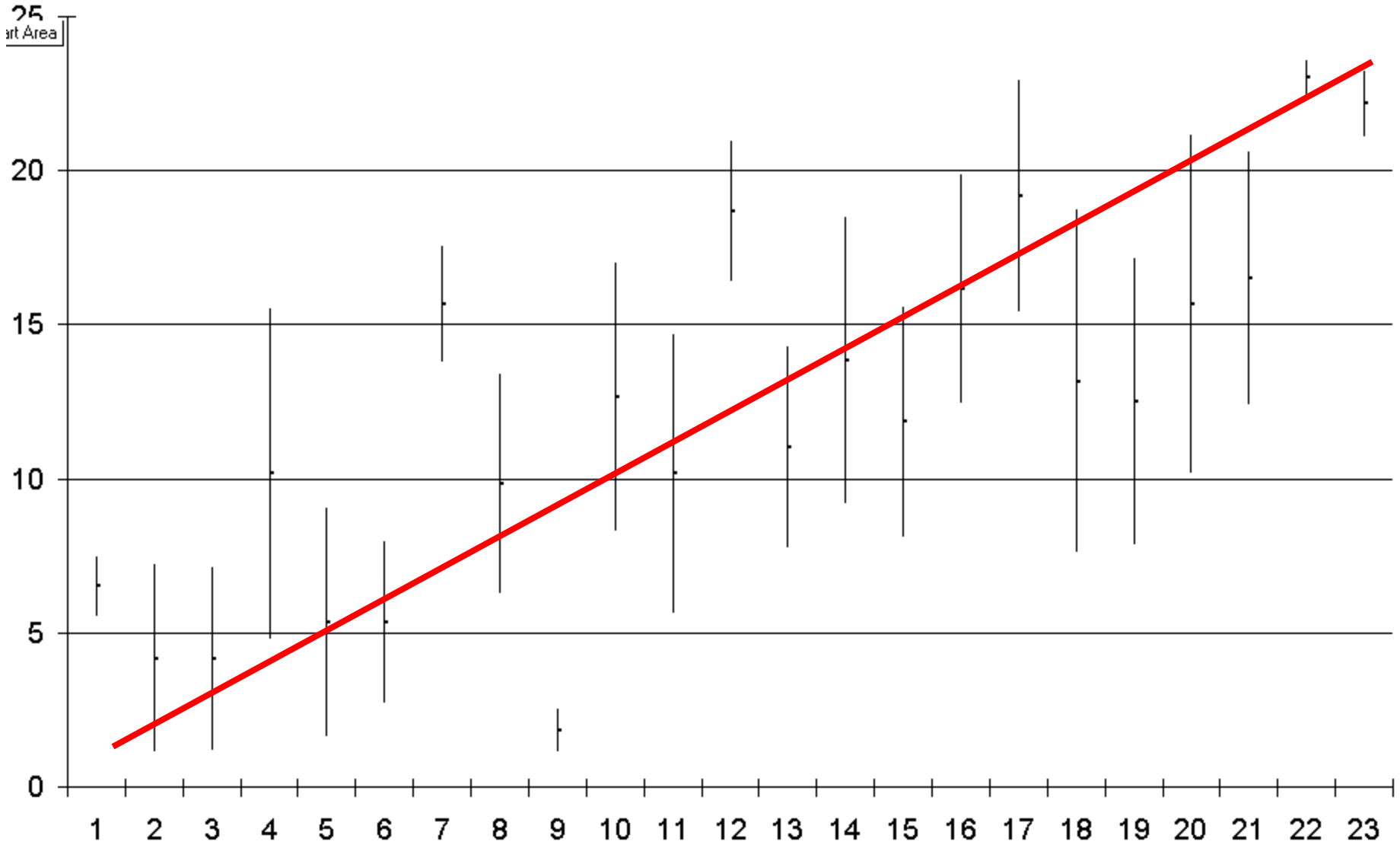
But speakers didn't *always* agree. The biggest disagreements were:

People liked this better than computer: **tinker**

Computer liked this better than people: **dispensing, telesciences, diphtheria, sauté**

Here is the average ranking assigned by six speakers:

and here is the same score, with an indication of one standard deviation above and below:



Categories

- So far we have made no assumptions about categories.
- Except that there are “phonemes” of some sort in a language, and that they can be counted.
- We have made no assumption about phonemes being sorted into categories.

Ask a 2-state HMM to find the device which assigns the highest probability to a sequence of phonemes

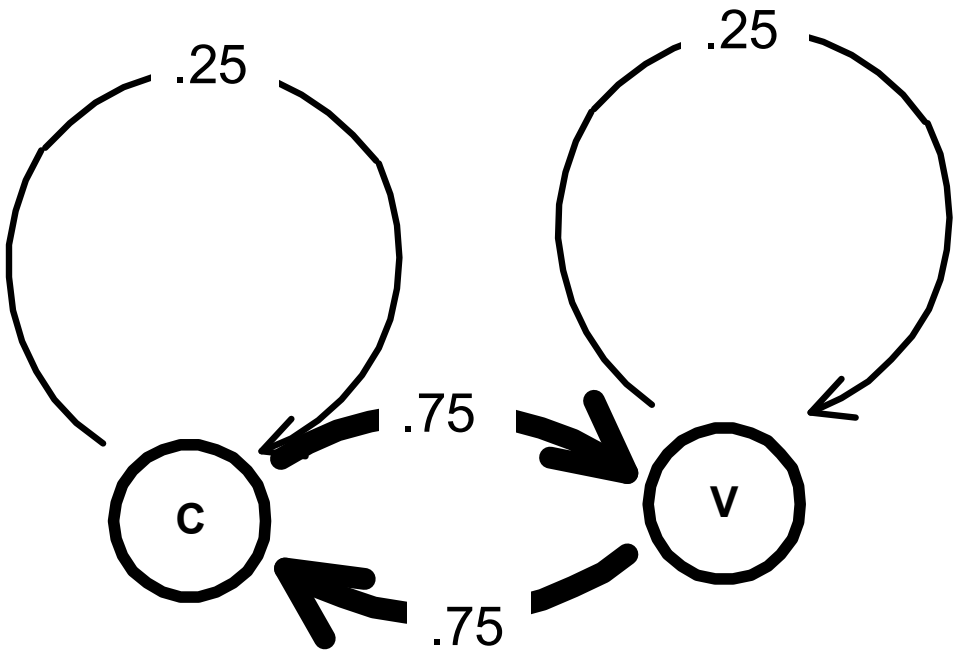
Let's apply the method to the phonemes in Finnish words: 44,450 words.

We begin with a finite-state automaton with 2 states: both states generate *all* the phonemes with roughly equal probability.

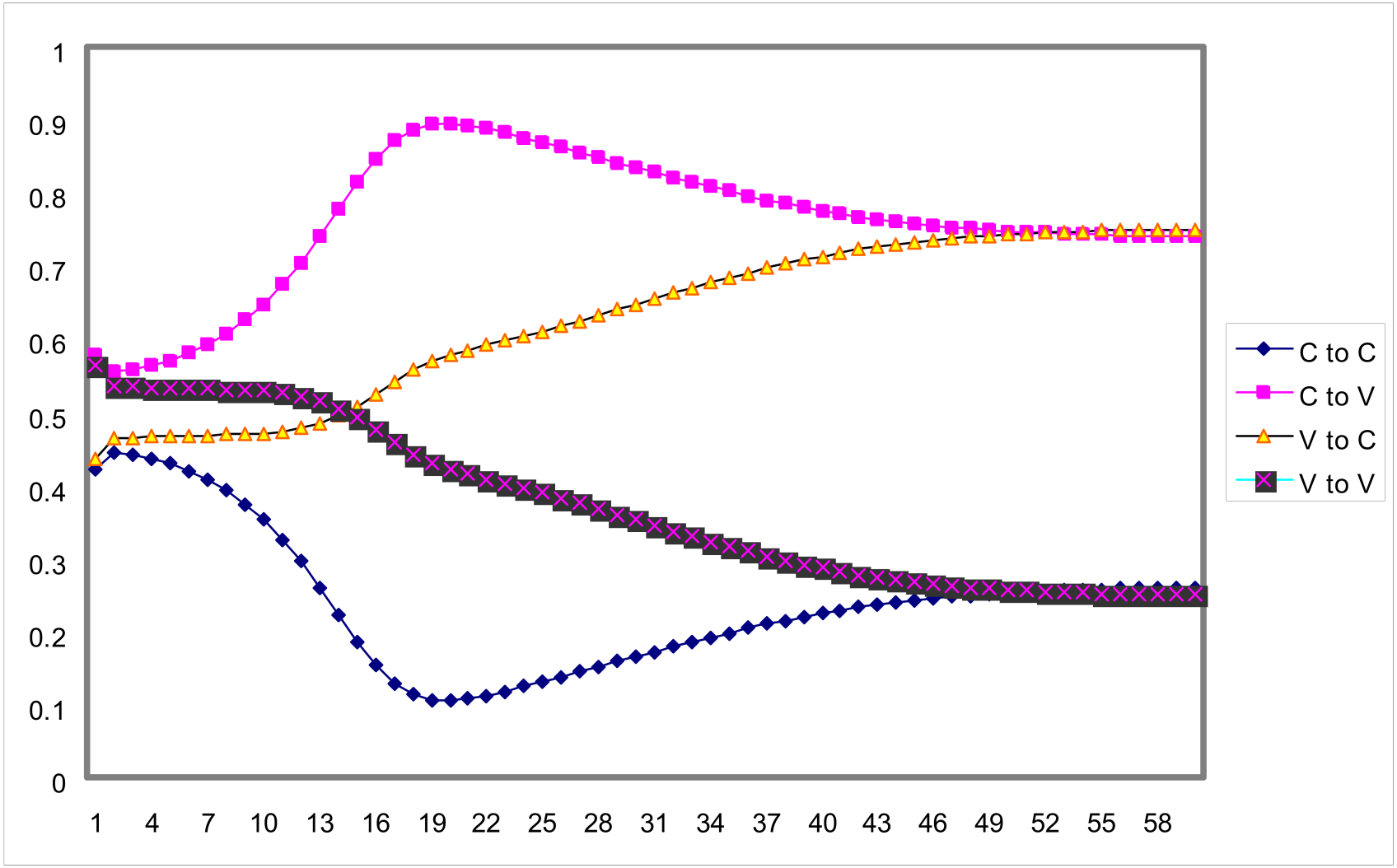
Both states begin with random transition probabilities to each other.

The system *learns* the parameters that maximize the probability of the data.

Transition probabilities (Finnish)

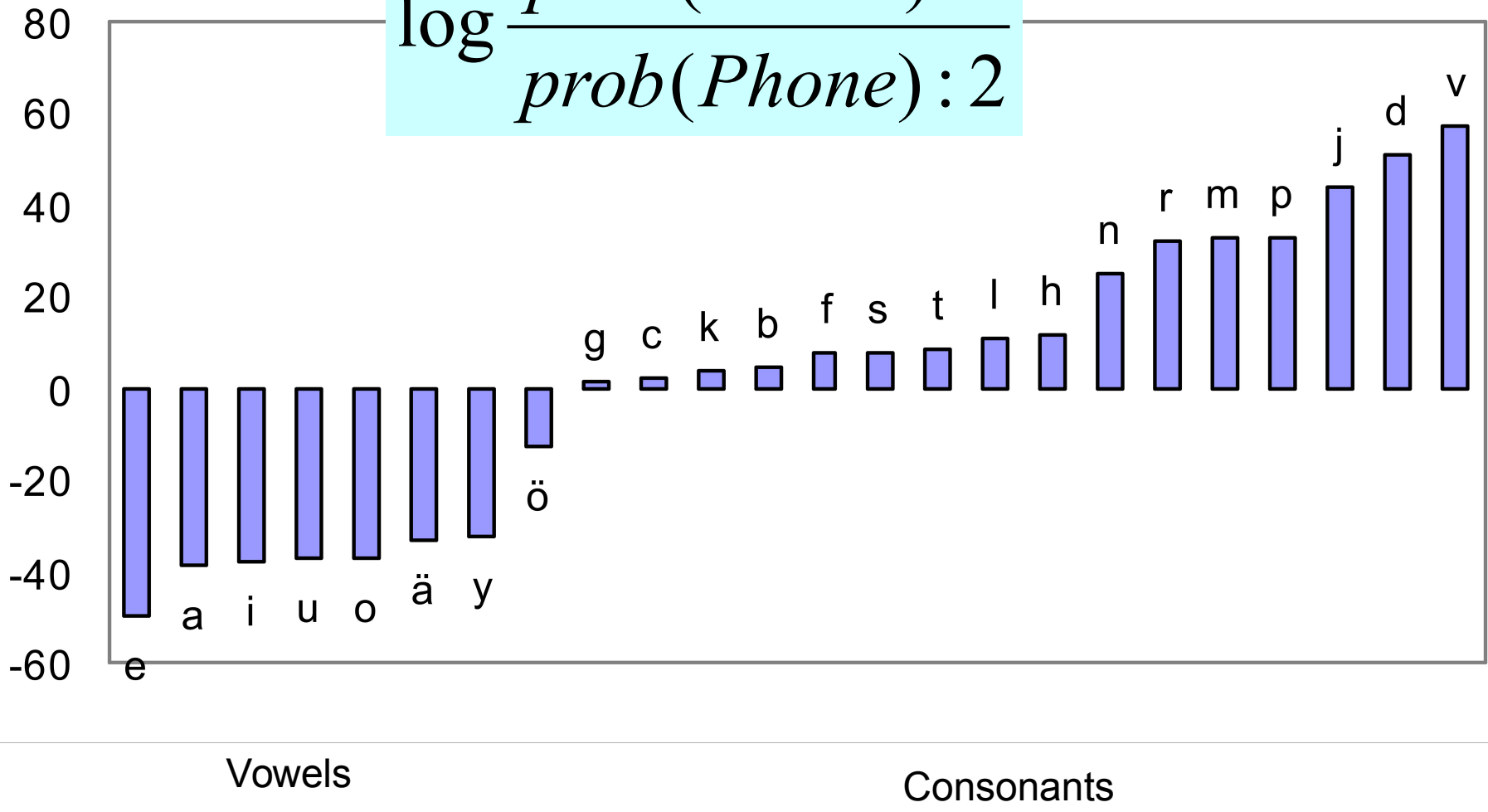


Finding Cs and Vs in Finnish

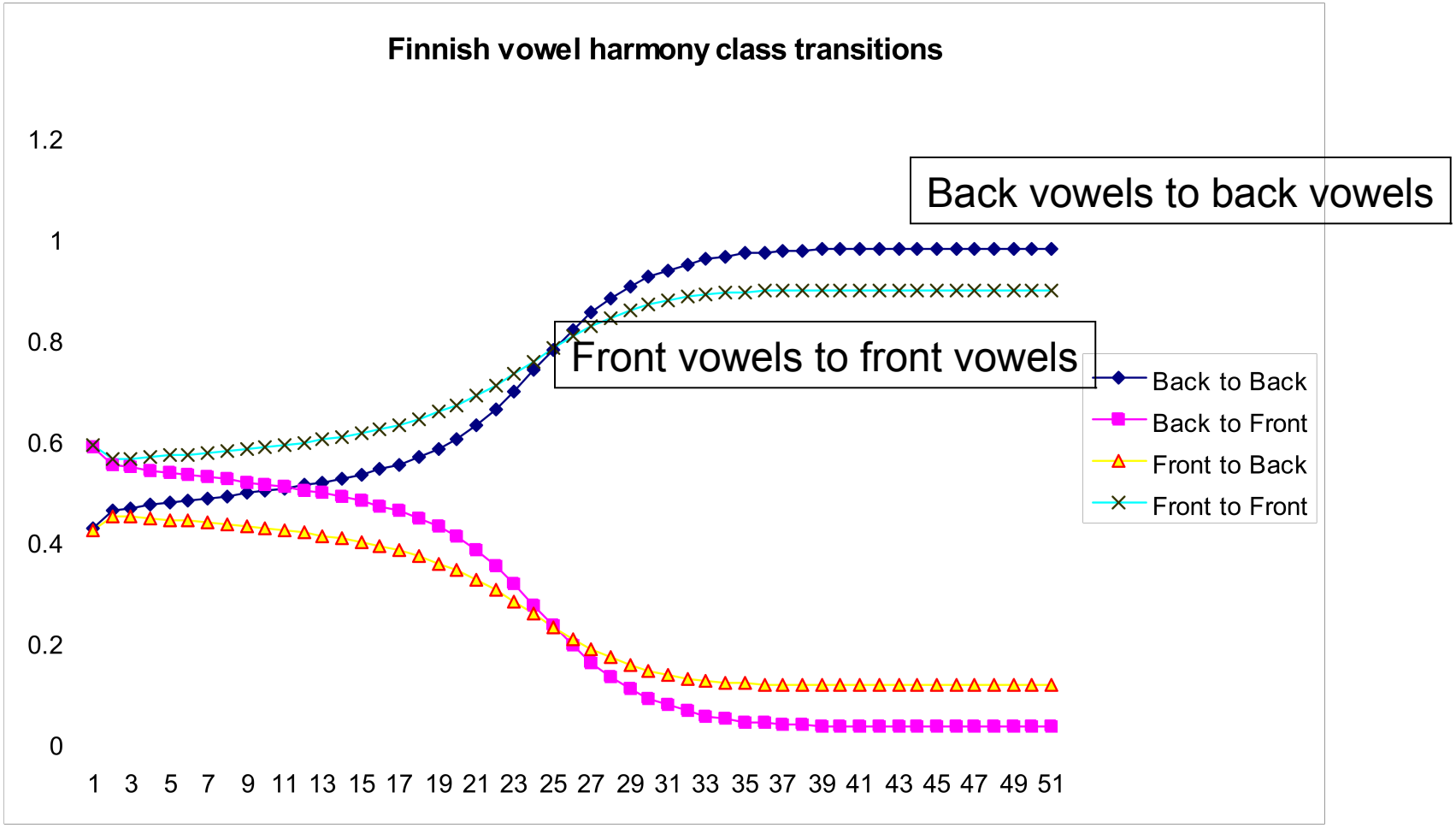


Log probability C/V Finnish segments

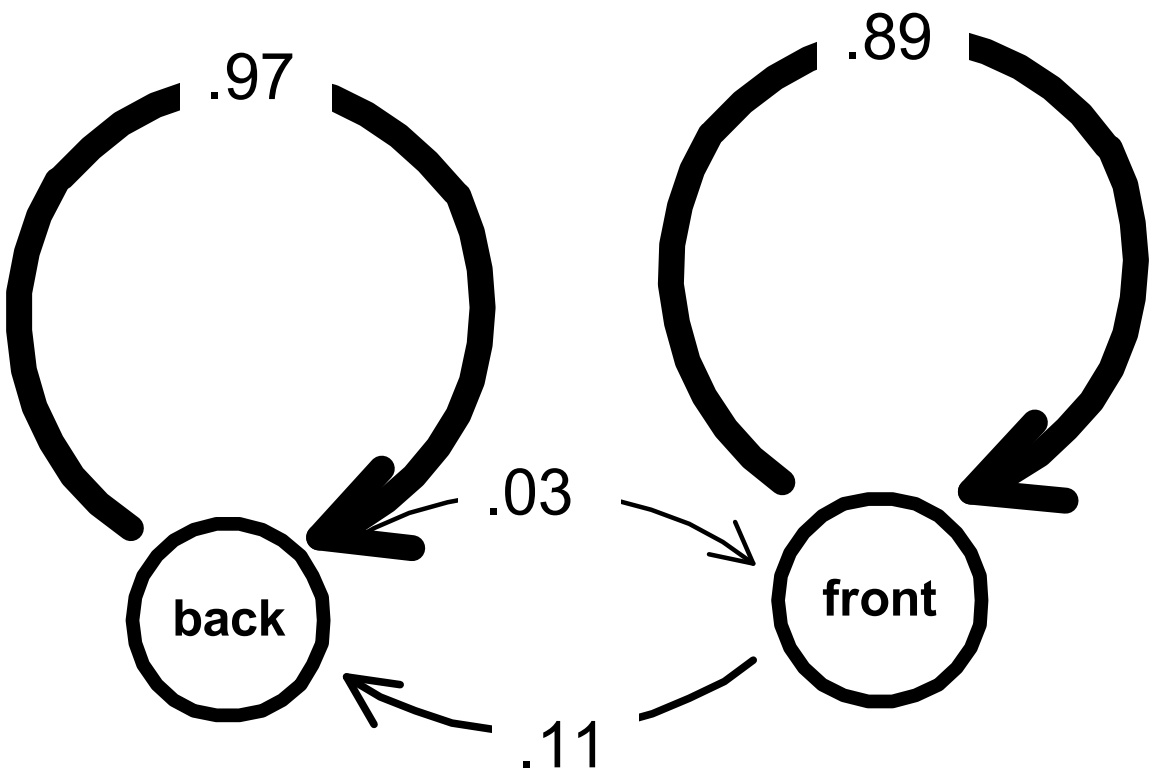
$$\log \frac{prob(Phone):1}{prob(Phone):2}$$



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



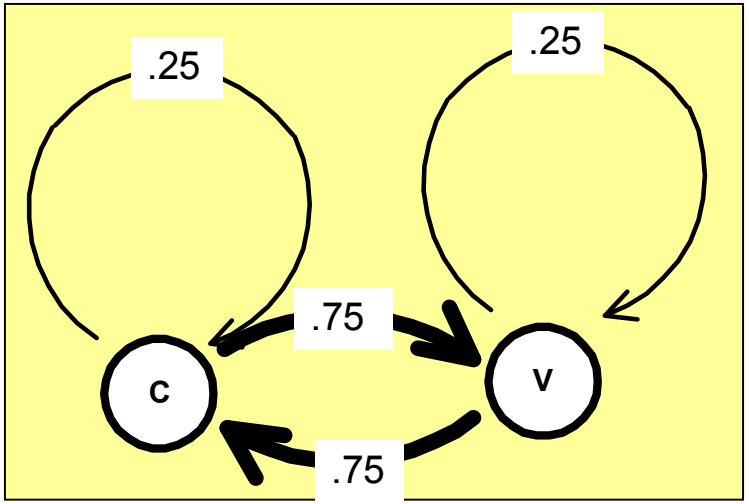
Vowel harmony



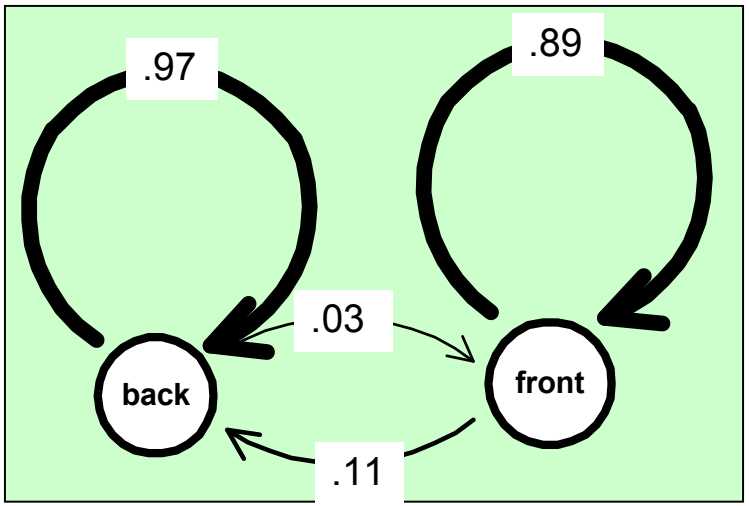
Vowel harmony classes in Finnish

	Vowel	State 1	State 2	Log ratio probability
Back vowels	a	0.353305	0.000647	9.093269
	u	0.158578	0.000302	9.038642
	o	0.133042	0.014794	3.168788
neutral vowels	i	0.215194	0.266105	-0.30636
	e	0.139881	0.254647	-0.8643
Front vowels	y	7.71E-15	0.157373	-44.2153
	ö	1.60E-18	0.050579	-54.8158
	ä	1.51E-18	0.255554	-57.2334

Contrast what was learned:



Splitting all segments into consonants and vowels.



Splitting all vowels into front and back vowels.

...from **exactly the same learning algorithm**, pursuing exactly the same goal: **maximize the probability of the data**.

Take-home message

The scientific goal of discovery of the best algorithmic model that generates the observed data is an outstanding one for linguists to pursue, and it requires no commitment to any particular theory of universal grammar rooted in biology.

It is deeply connected to theories of learning which are currently being developed in the field of machine learning.