Optimization is the answer. Now, what is the question?

John A Goldsmith

September 26, 2008

The Ohio State University

Work done together with

- Jason Riggle (Chicago)
- Aris Xanthos (Lausanne).

Introduction

Overview

- A brief look at 4 research projects
- A new look at data, and its role
- Some mathematics (all of it probability)
- A simple principle, which has a broad range of consequences, both abstract and concrete.

A simple idea:

- Let us find the most likely hypothesis h, given the data D.
- Let us find the h that maximizes pr(h|D).

Could this be an understanding of the science we do?

Probability 2

What could $pr(h|D)$ mean?
"The probability of a hypothesis, given the data that we have."

- Bayes' rule says: $pr(h|D) = \frac{pr(D|h)pr(h)}{pr(D)}$
- The difficult part to understand is $pr(h)$. And maybe understanding what $pr(D|h)$ is is not so easy, either.
- $pr(h)$: this is the prior probability of a grammar h. How can we conceive of such a thing?

Before answering that question, let's review what a *probability* is.
(It has nothing to do with fuzzy data.)

What is a probability?

- An itemization of the (infinite) set of possibilities $\{r_i\}$;
- A function pr that maps each of these to a non-negative number $pr : \{r_i\} \rightarrow [0, 1]$
- such that these numbers sum to 1.0: $\sum_i pr(r_i) = 1.0$
- The hard part is understanding this both at the level of the forms generated by the grammar (their probabilities sum to 1.0), and the probability assigned to grammars (the probability of all grammars also sums to 1.0).

Yes: A fully probabilistic model assigns a probability to the grammar chosen, and a probability of the data, given the grammar.

A probability distribution over grammars

- Kolmogorov complexity: a universal system of measurement.
- This work is an exploration of that hypothesis.
- It is the extreme opposite of the view that language is learnable only because human languages are selected from a small subset of possible algorithms.

Linguistics and learning

- The linguist wants to understand how language can be learned.
- The psychologist wants to understand how language is learned.

They sound very similar, and clearly they are related. But they are not the same.

Two families of views on how to understand learning:

- A procedural, deterministic algorithmic search procedure in a space representing grammars with a determinate ending point.
- A non-deterministic search procedure seeking a maximum (minimum) on some abstract ("energy") landscape. To learn a grammar is to search the space of grammars for one that has the best properties, given the data. How do we flesh out that notion?

Classical generative grammar took the position that all linguistics could do (and hence, should do) is define the mapping from grammar to height on the "complexity" landscape: consistent with the second view.

- Tell us which, of two grammars, is more highly valued: $f : \mathcal{G} \times \mathcal{G} \rightarrow \{1, 2\}$
- But Chomsky abandoned the program (1979).
- No effort was made to deal with the question of *fit* of model and data.
- It was replaced by the view that there is no learning to speak of in language-learning.
- The work presented here emphasizes both grammar complexity and probability assigned to observed data.

- The general strategy is to design probabilistic grammars, and to focus on grammars where we can plausibly approximate the Kolmogorov complexity of the grammar.
- This strategy can be converted to a hill-climbing strategy for learning.
- We can see whether the style of learning that emerges is one that reflects the structure that we as linguists recognize.

Roadmap

- Word discovery
- Morpheme discovery
- Sonority
- Vowel harmony

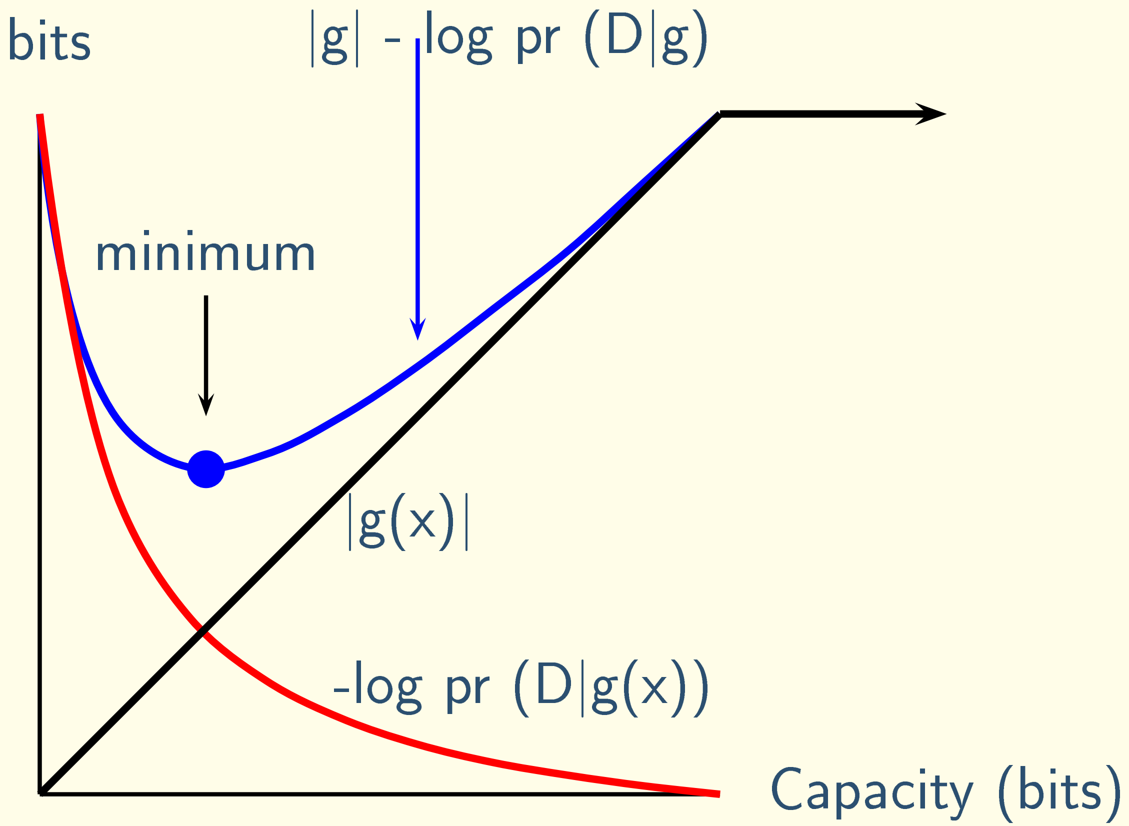
Word discovery

A long history...

- The problem: take a corpus C without breaks, and insert them in the right places with no prior knowledge of the language.
- Important step taken in the mid-1990s by Michael Brent and Carl de Marcken: using MDL.
- Minimum Description Length analysis.
- Looks for a happy medium between two extremes:
  - one extreme says all the letters are separate words;
  - one says the corpus is one long word.

Minimum description length (MDL)

- Description Length of Data D, using grammar g.
- In this case, the grammar is the lexicon, or list of words.
- Length of g + What is not accounted for in the data D by g.
- DL(D,g) = |g| - $log\, pr_g(D)$
- Find the grammar g^* that minimizes DL(D,g).


bits
|g| - log pr (D|g)
minimum
|g(x)|
-log pr (D|g(x))
Capacity (bits)

The most likely hypothesis h^*, given the data D.

$$pr(h|D) = \frac{pr(D|h)pr(h)}{pr(D)} \tag{1}$$

$$h^* = \arg\max_h \; pr(D|h)pr(h) \tag{2}$$

$$= \arg\max_h \left[log\, pr(D|h) + log\, pr(h)\right] \tag{3}$$

$$= \arg\min_h \left[\underbrace{-log\, pr(D|h)}_{compressedlength} \underbrace{-log\, pr(h)}_{lengthofh}\right] \tag{4}$$

$$= \arg\min_h \left[\begin{array}{c}\text{Length of data, given h +}\\ \text{length of h}\end{array}\right] \tag{5}$$

In English, choose the hypothesis h^* such that h minimizes the sum of the compressed length of the data plus the length of the hypothesis.

How do we determine the length of the hypothesis?

- Universal Kolmogorov prior: the length of an algorithm $\mathcal{A}$ is the length, in bits, of the shortest program that implements $\mathcal{A}$ on a chosen universal Turing machine.
- pr(g) $= 2^{-|g|} = \frac{1}{2^{|g|}}$
- In the case of a lexicon, the length of $\mathcal{L}$ is the number of bits it takes to specify a list of N letters in M words. Crude approximation: $log_2 27 \times (N + M)$
- Better approximation:

$$-\sum_{w \in \mathcal{L}} log_2 \sum_{i=1}^{|w|+1} log_2 \, pr(w[i]|w[i-1])$$

word discovery as optimization problem

Begin a hill climbing operation:

- Lexicon $\mathcal{L} \Leftarrow$ alphabet.
- Loop
  - Consider the two words w_i, w_j that occur together the most frequencly under the current analysis $\mathcal{L}$
  - Add $w_i w_j$ to the Lexicon to form $\mathcal{L}^*$.
  - If DL(D,$\mathcal{L}^*$) < DL(D,$\mathcal{L}$), set $\mathcal{L} \Leftarrow \mathcal{L}^*$.
- ...until stopping condition is satisfied.

NoSpaces

3749 sentences, 400,000 characters.
TheFultonCountyGrandJurysaidFridayaninvestigationofAtl
anta'srecentprimaryelectionproducednoevidencethatan
yirregularitiestookplace.f
Thejuryfurthersaidinterm-endpresentmentsthattheCityE
xecutiveCommittee,whichhadover-allchargeoftheelecti
on,deservesthepraiseandthanksoftheCityofAtlantaforthem
annerinwhichtheelectionwasconducted ...

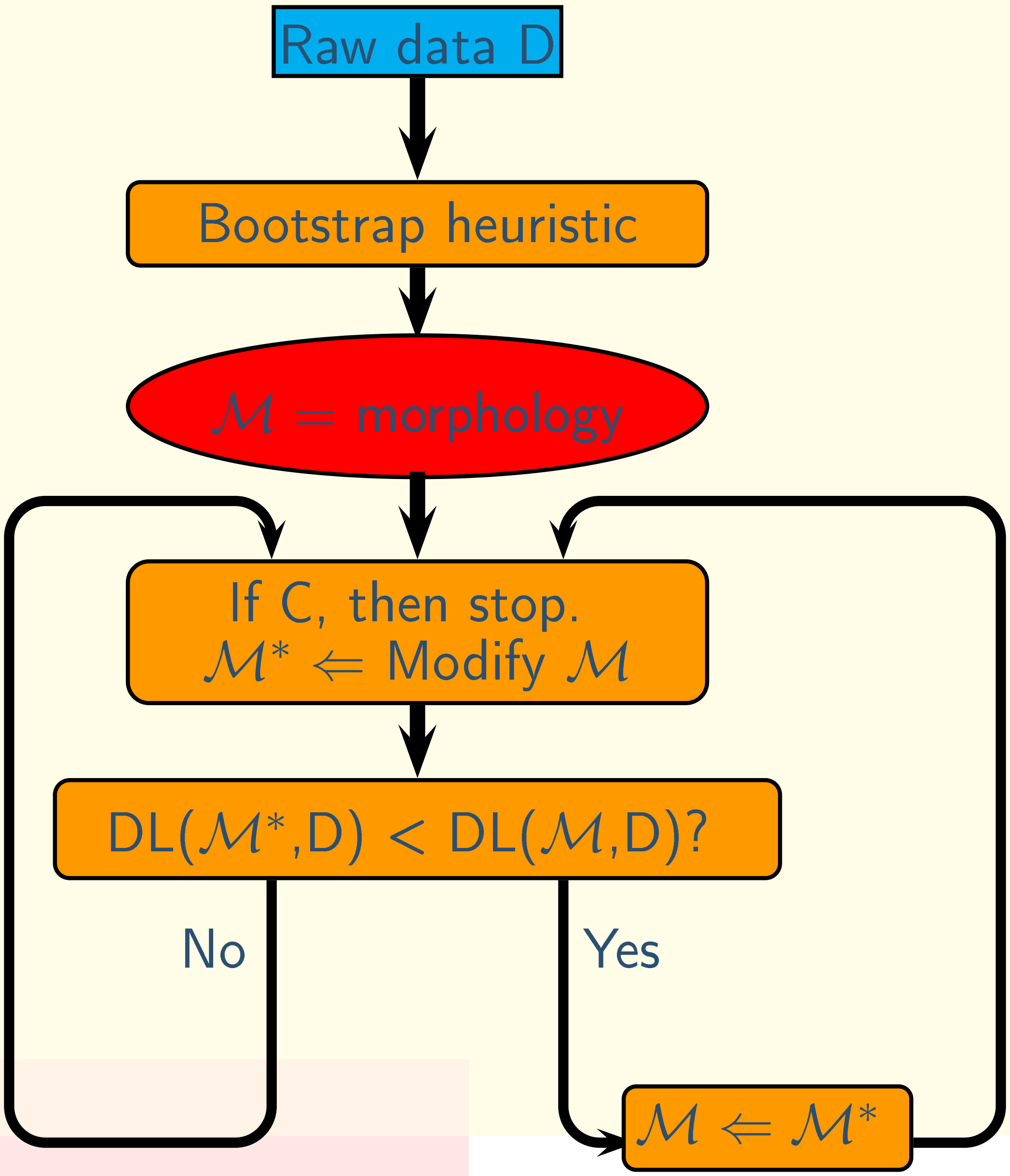
The Fulton County Grand Ju ry s aid Friday an investi gation of At l anta 's recent prim ary e lection produc ed no e videnc e that any ir regul ar it i es took place . Thejury further s aid in term - end present ment s thatthe City Ex ecutive Commit t e e ,which had over - all charg e ofthe e lection , d e serv e s the pra is e and than k softhe City of At l anta forthe man ner in whichthe e lection was conduc ted.

Problems for word learning

- The language model was too simple.
- We need a language model in which generalizations can be due to:
  - Existence of morphemes in the lexicon.
  - Existence of words with complex morphology.
  - Syntactic structure.
- But the example gives proof-of-concept plausibility to the enterprise.

Morphology discovery

- Linguistica Project linguistica.uchicago.edu
- Unsupervised learning of natural language morphologyf
- Employing MDL

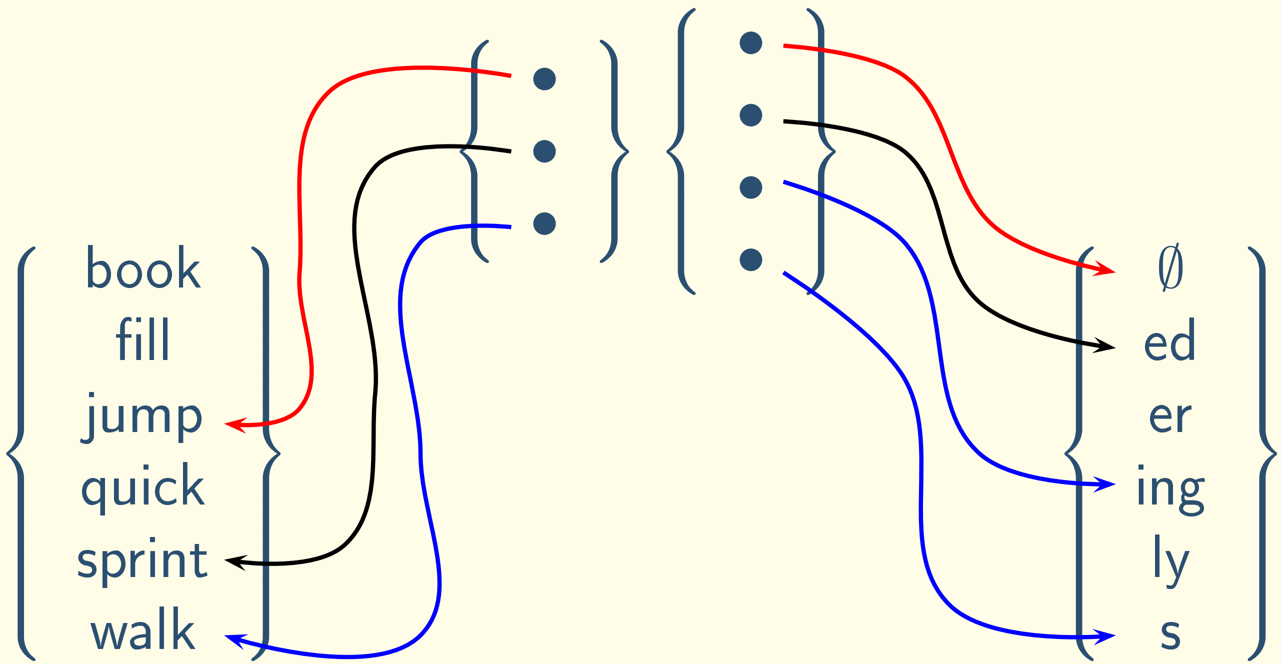

Raw data D
Bootstrap heuristic
$\mathcal{M}$ = morphology
If C, then stop.
$\mathcal{M}^* \Leftarrow$ Modify $\mathcal{M}$
DL($\mathcal{M}^*$,D) < DL($\mathcal{M}$,D)?
No
Yes
$\mathcal{M} \Leftarrow \mathcal{M}^*$

What is the complexity of a morphology? Sum of three things:

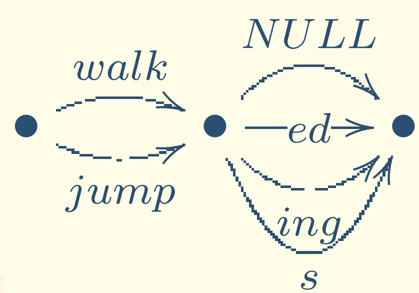
- Complexity of a list of stems;
- Complexity of a list of affixes;
- Complexity of the grammar of permitted combinations: $signatures$.

$$\left\{ \begin{array}{c} \text{jump} \\ \text{walk} \\ \text{sprint} \end{array} \right\} \left\{ \begin{array}{c} \emptyset \\ \text{ed} \\ \text{ing} \\ \text{s} \end{array} \right\}$$

Signature complexity



Bit-cost of link = -log pr of target



(6)

List of stems:

$$\sum_{t \in Stems} \sum_{i=1}^{|t|+1} -log\, pr(t_i | t_{i-1})$$

List of affixes:

$$\sum_{f \in Affixes} \sum_{i=1}^{|f|+1} -log\, pr(f_i | f_{i-1})$$

Signatures:

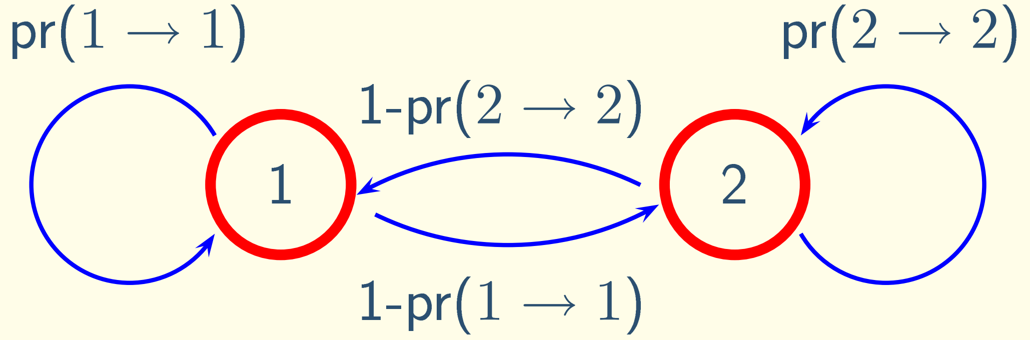
$$\sum_{\sigma \in Signatures} \left(\sum_{stem\, t \in \sigma} -log\, pr(t) + \sum_{suffix\, f \in \sigma} -log\, pr(f) \right)$$

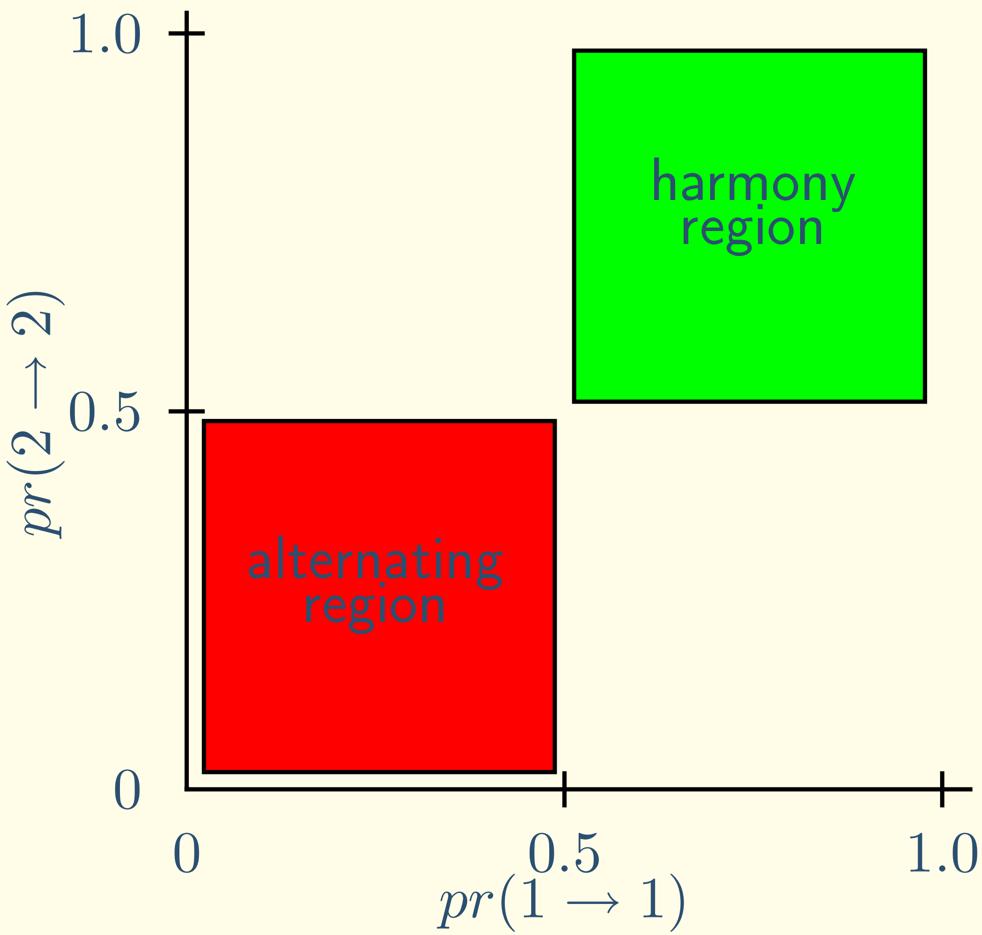
Phonology and probability

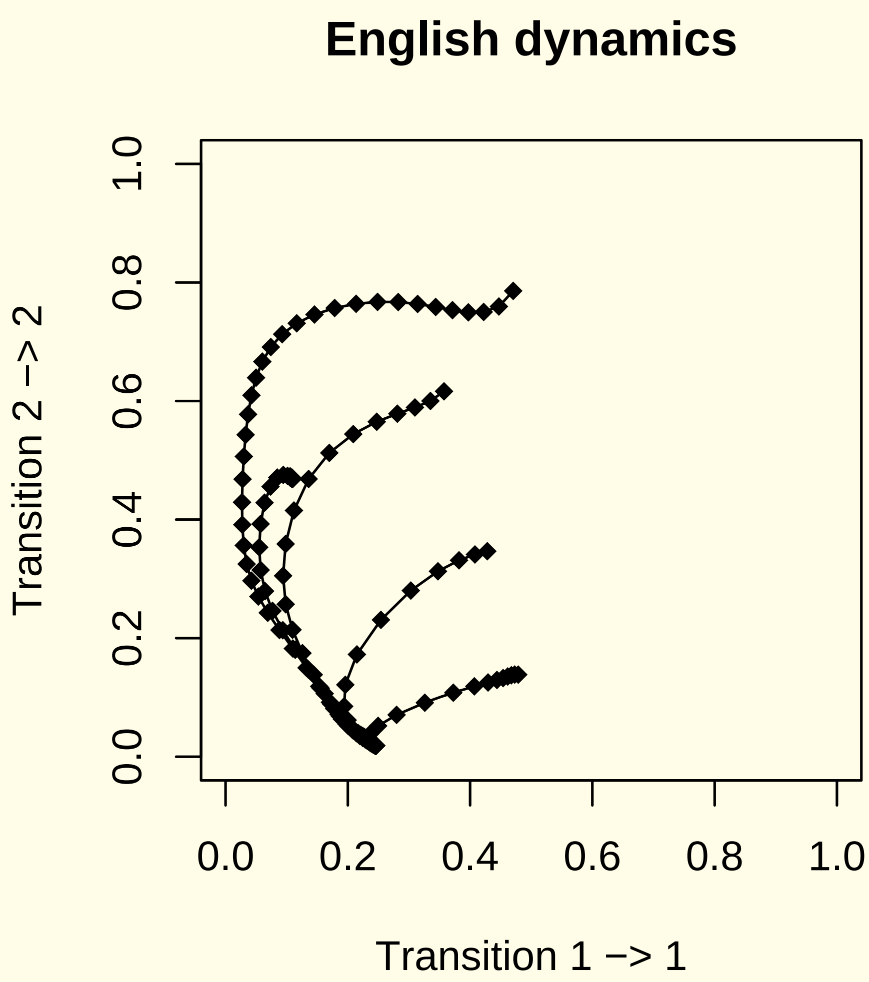
- A probabilistic approach to phonological representations is comfortable with complex representations.
- However, it is able to extract much more information from very simple (linear) structures.
- A few illustrations of the way in which "inverse log frequencies" (or information content) reflects linguistic intuitions of complexity and markedness.
- Mutual information $MI(a,b) \equiv log \frac{pr(a\&b)}{pr(a)pr(b)}$: positive if a and b attract, and negative if they repel.

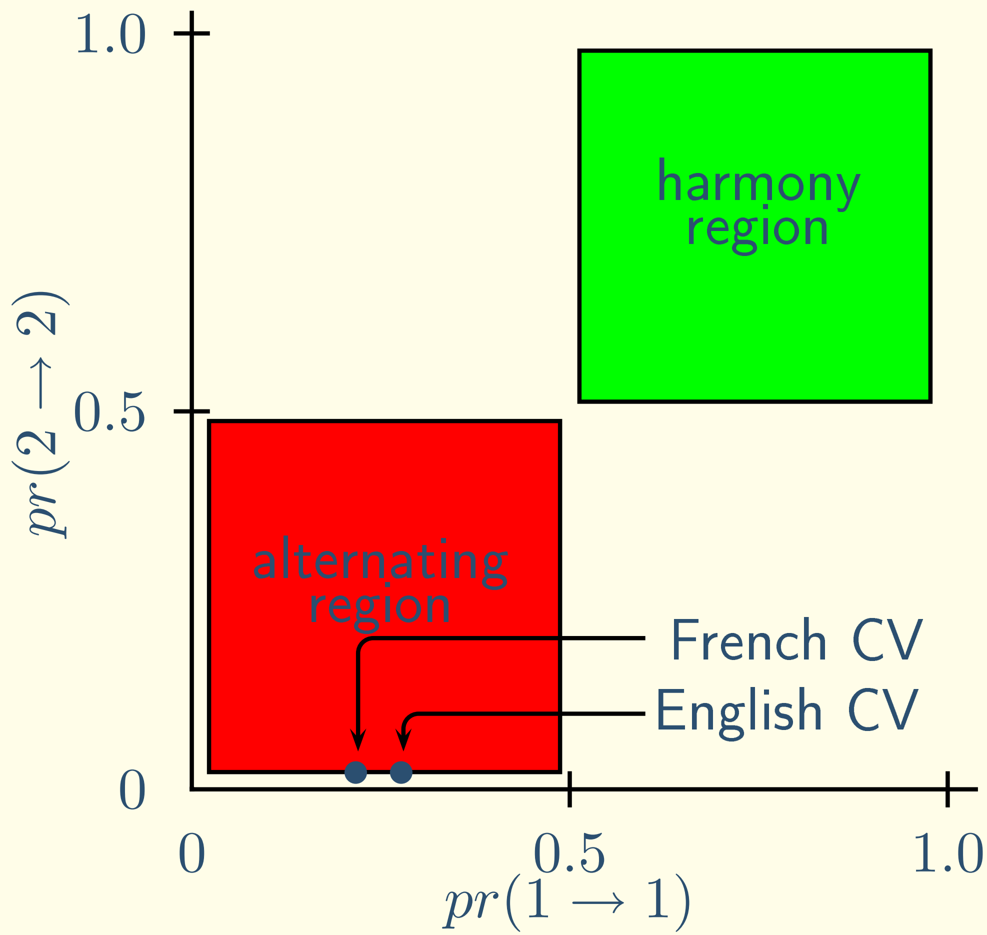
Sonority learning

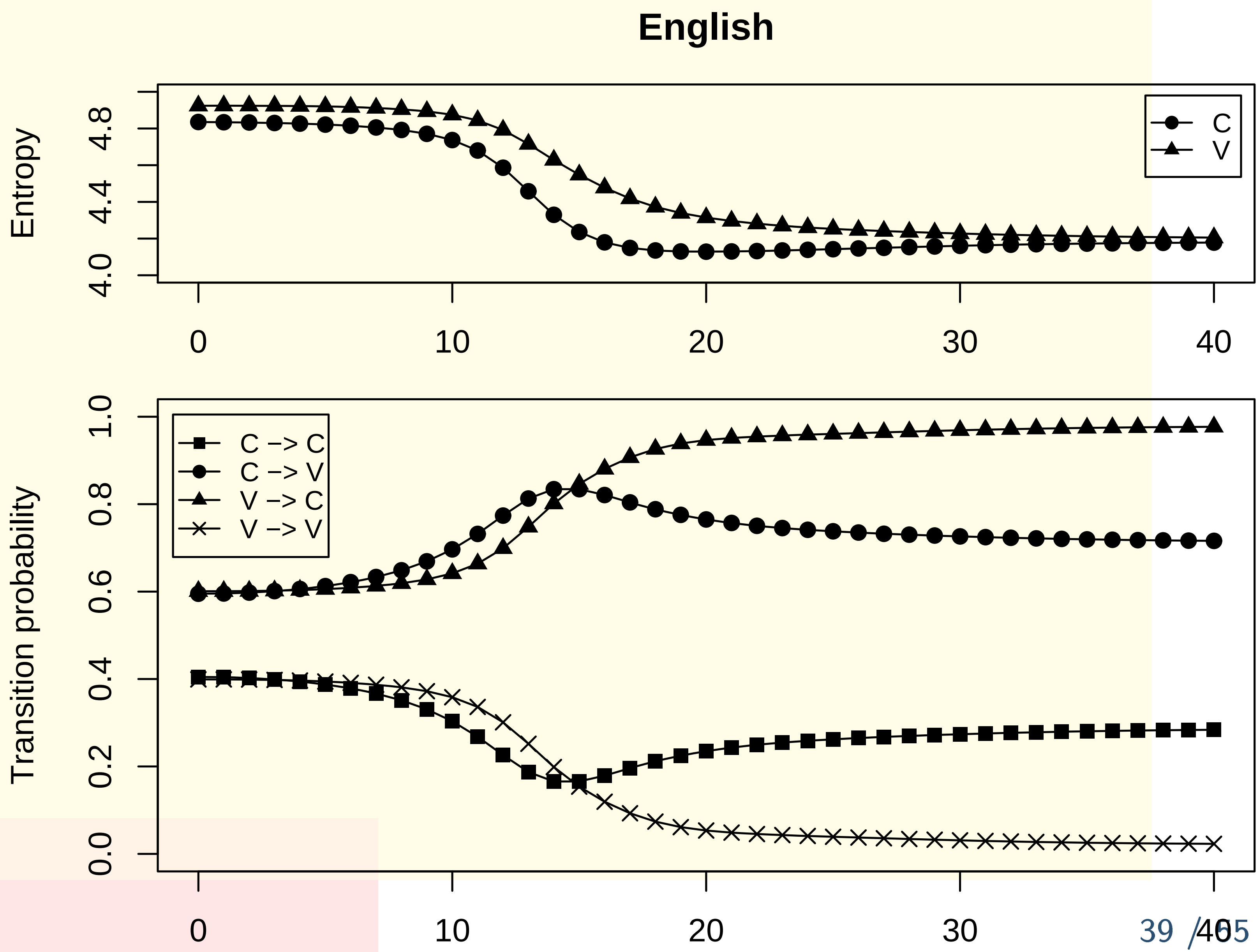
Work done jointly with Aris Xanthos (Lausanne)


pr($1 \to 1$)
pr($2 \to 2$)
1-pr($2 \to 2$)
1
2
1-pr($1 \to 1$)

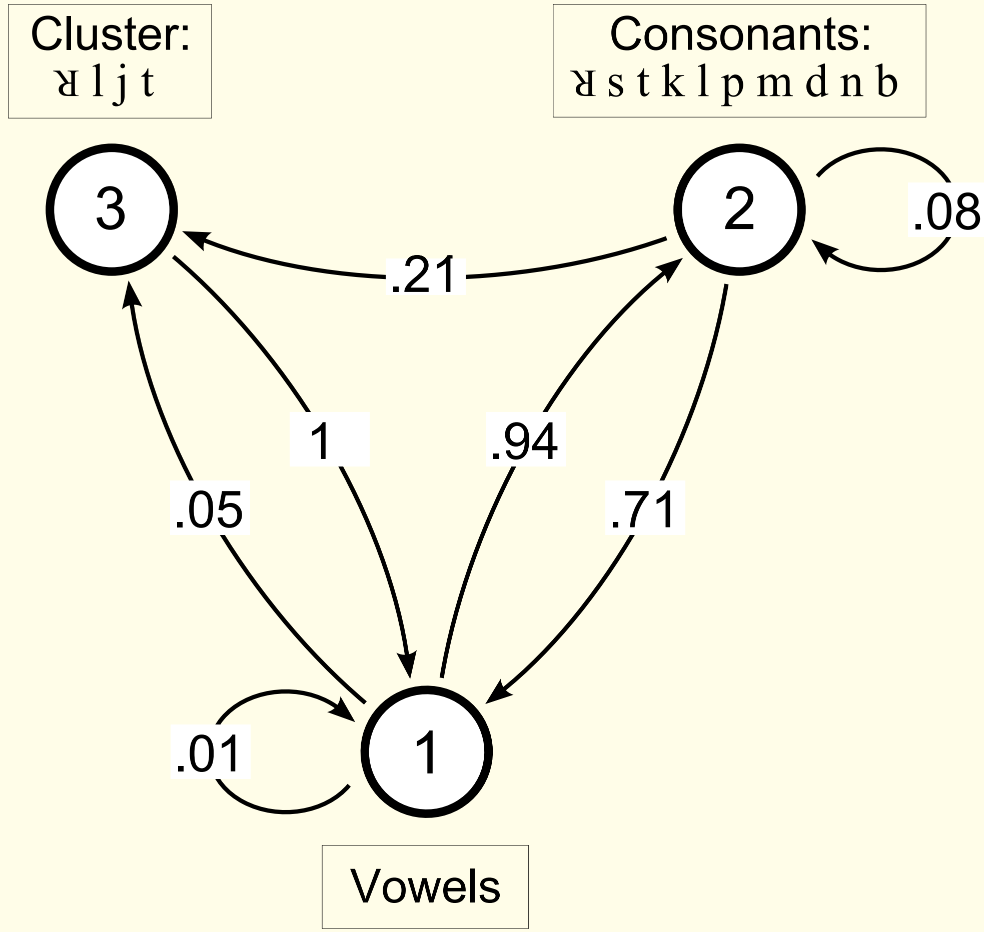

1.0
0.5
0
$pr(2 \rightarrow 2)$
harmony region
alternating region
0
0.5
1.0
$pr(1 \rightarrow 1)$

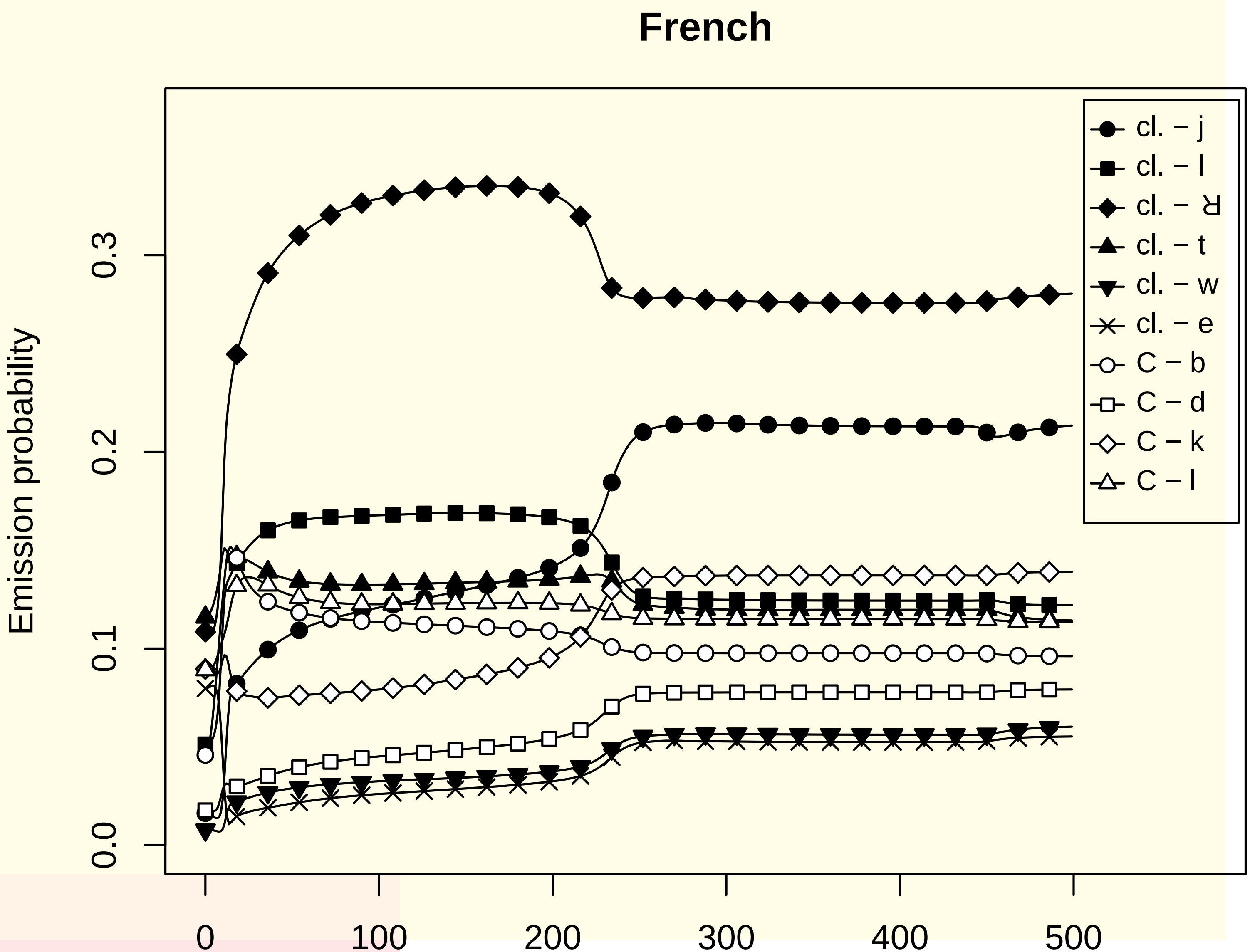

English dynamics
Transition 2 −> 2
0.0
0.2
0.4
0.6
0.8
1.0
Transition 1 −> 1


1.0
0.5
0
$pr(2 \rightarrow 2)$
harmony region
alternating region
French CV
English CV
0
0.5
1.0
$pr(1 \rightarrow 1)$

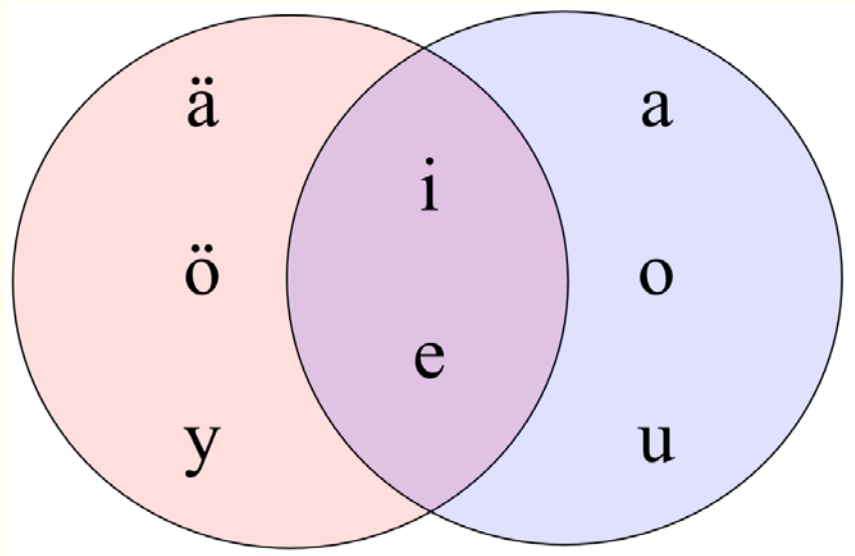

English
Entropy
4.0
4.4
4.8
C
V
0
10
20
30
40
Transition probability
0.0
0.2
0.4
0.6
0.8
1.0
C -> C
C -> V
V -> C
V -> V
0
10
20
30
40

French


Cluster:
ʁ l j t
Consonants:
ʁ s t k l p m d n b
3
2
1
.08
.21
1
.94
.05
.71
.01
Vowels



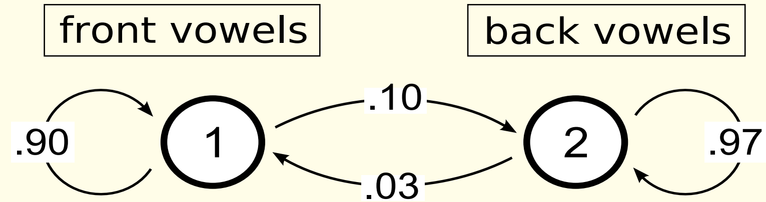
French
Emission probability
0.0
0.1
0.2
0.3
0
100
200
300
400
500
cl. − j
cl. − l
cl. − ʁ
cl. − t
cl. − w
cl. − e
C − b
C − d
C − k
C − l

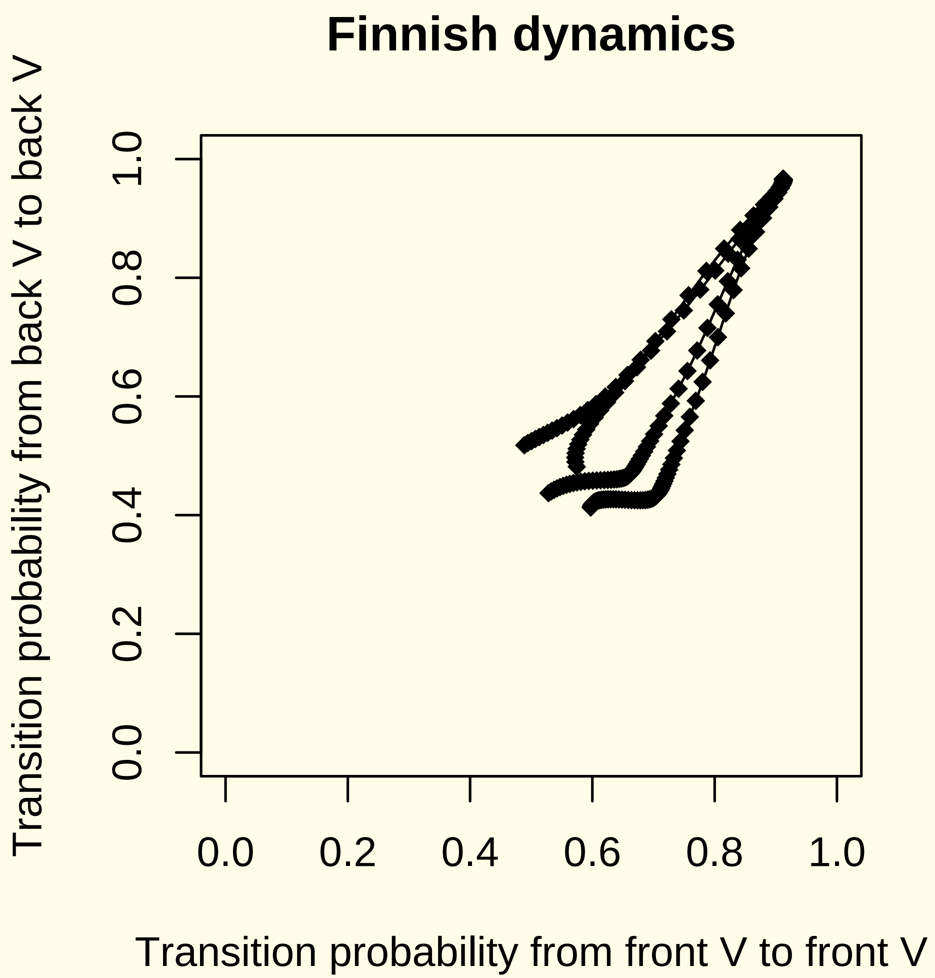


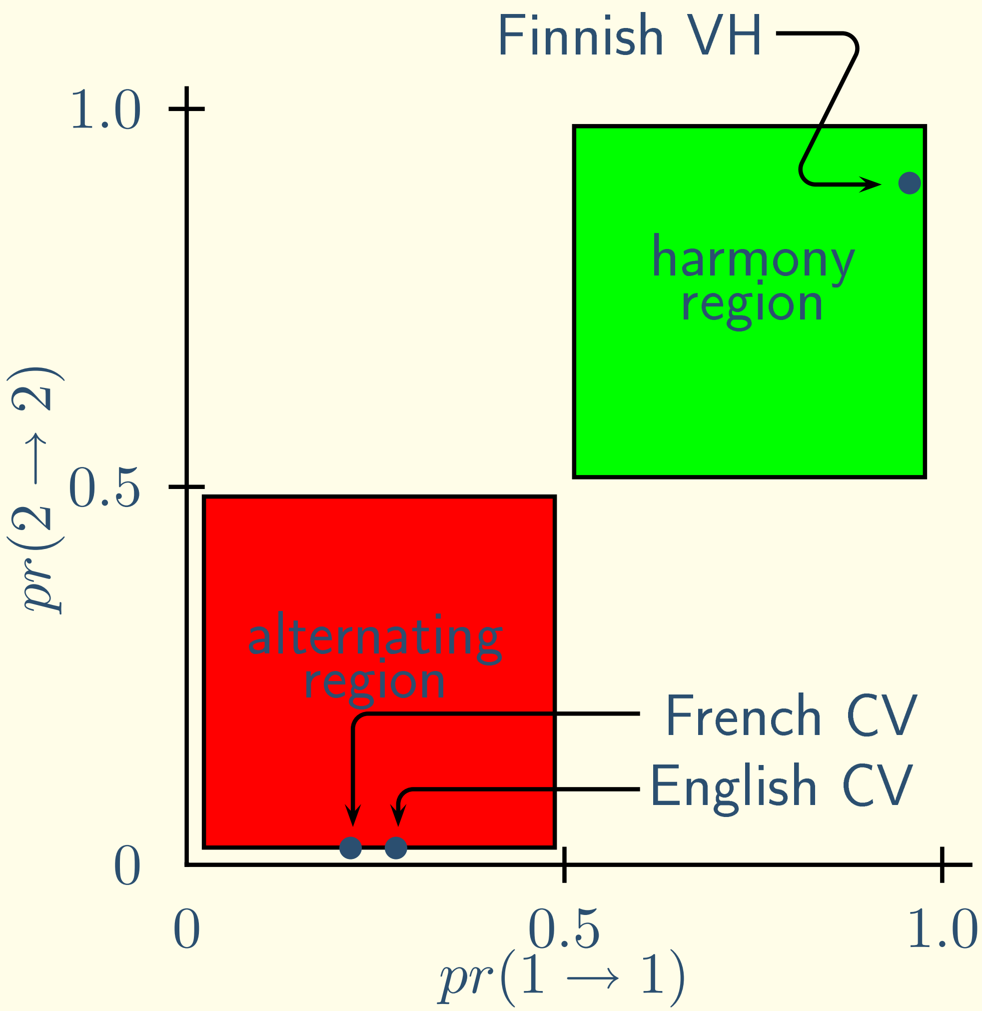
- We can easily separate vowels and consonants in Finnish.
- What if we find the best 2-state system to generate the vowel sequences?

Vowel	Log ratio
ö	999
ä	961
y	309
e	0.655
i	0.148

Vowel	Log ratio
o	-7.66
a	-927
u	-990




Finnish dynamics
Transition probability from back V to back V
Transition probability from front V to front V
0.0
0.2
0.4
0.6
0.8
1.0


Finnish VH
1.0
0.5
0
harmony region
alternating region
French CV
English CV
0
0.5
1.0
$pr(1 \to 1)$
$pr(2 \to 2)$

Vowel harmony learning

Work done jointly with Jason Riggle (Chicago).

- Q: How can information theory express compactly the regularity that linguists call *vowel harmony*?
- A: In a vowel harmony system, the number of choices for the next vowel is cut in half after the first vowel...
- if you allow the choice of vowel to be dependent on the previous vowel.

	ä	ö	y	e	i	a	o	u
ä	2.23	1.54	1.43	-0.35	0.14	-3.00	-2.64	-3.18
ö	1.29	2.54	1.53	-0.16	0.32	-1.23	-0.89	-1.29
y	1.61	3.42	2.35	0.19	0.21	-2.96	-2.71	-3.73
e	0.30	-1.85	-0.30	0.23	0.23	-0.67	-0.54	-0.20
i	0.02	0.51	-0.35	0.62	-0.15	0.11	0.06	-0.55
a	-3.54	-5.27	-3.33	-1.03	0.07	0.45	-0.36	0.02
o	-3.34	-4.33	-2.30	0.10	0.82	0.18	0.29	0.06
u	-2.93	-3.65	-2.22	0.15	-0.11	0.15	0.89	1.12

The numbers support the intuition.

Autosegmental model

$$\# \xrightarrow{2.46} V \xrightarrow{3.14} t \xrightarrow{0.46} V \xrightarrow{4.23} k \xrightarrow{0.38} V \xrightarrow{2.22} V \xrightarrow{6.34} d \xrightarrow{0.07} V \xrightarrow{2.80} n$$

$$\xrightarrow{6.19} \text{ö} \xrightarrow{3.68} \text{ö} \xrightarrow{3.68} \text{ö} \xrightarrow{1.84} i \xrightarrow{2.00} e$$

Inverse log conditional probability

päätuotetta (päätuote = main product)	
Word probability in unigram model: Plog sum in unigram model:	2.00×10^{-15} 48.8
Word probability in bigram model: Plog sum in bigram model:	6.37×10^{-12} 37.2
Word prob in autosegmental model: Plog sum in autosegmental model:	2.41×10^{-14} 45.2

- The real effect of vowel harmony appears much more strongly over a consonant than adjacent.
- The over-all mutual information between choice of C and V is greater than the effect of vowel harmony

$$
\begin{tikzcd}
\# \arrow[r, "2.46"] & \text{ö} \arrow[r, "3.14"] \arrow[rr, bend right, dotted, leftrightarrow, "(2.54)"'] & t \arrow[r, "0.46"] & \text{ö} \arrow[r, "4.23"] \arrow[rr, bend right, dotted, leftrightarrow, "(2.54)"'] & k \arrow[r, "0.38"] & \text{ö} \arrow[r, "2.22"] & i \arrow[r, "6.34"] \arrow[rr, bend right, dotted, leftrightarrow, "(0.62)"'] & d \arrow[r, "0.07"] & e \arrow[r, "2.80"] & n
\end{tikzcd}
$$

Our final model operates on a single tier, including adjacent mutual information and also vowel-to-vowel mutual information.

We develop a model in which inter-segmental forces vary from language to language with regard to their fall-off over distance.

The future of bayesian linguistics

A formal, empiricist view of linguistics:

- We compute the length of the grammar expressed in a yet unknown language: "UG" of classical generative grammar.
- We use Kolmogorov complexity for the grammar's length: a truly universal measure.
- Linguistics becomes a science of external linguistic facts, as gathered and archived by linguists (rather than an avowed subdiscipline of cognitive psychology).