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

John A Goldsmith

April 11, 2008
Discovering and Representing Phonological Patterns
University of Konstanz
Work done together with

- Jason Riggle (Chicago)
- Aris Xanthos (Lausanne).
Introduction
- A brief look at 4 research projects
- A new look at data, and its role
- An attempt to start over and start afresh.
- Some mathematics (all of it probability)
- A lot of graphic visualization for that reason.
A simple idea:

- Let us find the most likely hypothesis $h$, given the data $D$.
- Let us find the $h$ that maximizes $\text{pr}(h|D)$.
What could \( pr(h|D) \) mean?

- 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?
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 $pr : \{r_i\} \to [0, 1]$ 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).
- 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.
Two families of views:

- 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.

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: $f : \mathcal{G} \times \mathcal{G} \to \{1, 2\}$

But Chomsky abandoned the program (1979).

No effort was made to deal with the question of fit of model and data.

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.
- Word discovery
- Morpheme discovery
- Sonority
- Vowel harmony
Word discovery
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 \( DL(D,g) \).
Description Length

\[ \text{bits} \quad |g| - \log \text{pr (D|g)} \]

\[ \text{minimum} \]

\[ |g(x)| \]

\[ -\log \text{pr (D|g(x))} \]

Capacity (bits)
The most likely hypothesis $h^*$, given the data $D$.

\[
pr(h|D) = \frac{pr(D|h)pr(h)}{pr(D)}
\]  

(1)

\[
h^* = \arg \max_h pr(D|h)pr(h)
\]  

(2)

\[
h^* = \arg \max_h [\log pr(D|h) + \log pr(h)]
\]  

(3)

\[
= \arg \min_h \left[ \begin{array}{c}
-log pr(D|h) \\
-compressed length
\end{array} \right] +
\begin{array}{c}
-log pr(h) \\
-length of h
\end{array}
\]  

(4)

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

(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:

\[
- \sum_{w \in \mathcal{L}} \log_2 \sum_{i=1}^{\lfloor |w|/2 \rfloor} \log_2 \Pr(w[i]|w[i-1])
\]
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 frequently 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.
3749 sentences, 400,000 characters.
The Fulton County Grand Jury said Friday an investigation of Atlanta’s recent primary election produced no evidence that an irregularity took place.
The jury further said interim-end presentation that the City Executive Committee, which had overall charge of the election, deserves the praise and thank of the City of Atlanta for the manner in which the election was conducted . . .
The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced no evidence that any irregularities took place. The jury further said in the end presentments that the City Executive Committee, which had oversaw all charges of the election, did serve as the primary and that the City of Atlanta forethe man ner in which the election was conducted.
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 morphology
- Employing MDL
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.

\[
\begin{align*}
\{ \text{jump} & \} & \{ \emptyset & \} \\
\{ \text{walk} & \} & \{ \text{ed} & \} \\
\{ \text{sprint} & \} & \{ \text{ing} & \} \\
\end{align*}
\]
Signature complexity

Bit-cost of link = -log pr of target

(6)
List of stems:

\[ \sum_{t \in \text{Stems}} \sum_{i=1}^{\mid t \mid + 1} -\log pr(t_i | t_{i-1}) \]

List of affixes:

\[ \sum_{f \in \text{Affixes}} \sum_{i=1}^{\mid f \mid + 1} -\log pr(f_i | f_{i-1}) \]

Signatures:

\[ \sum_{\sigma \in \text{Signatures}} \left( \sum_{\text{stem } t \in \sigma} -\log pr(t) + \sum_{\text{suffix } f \in \sigma} -\log pr(f) \right) \]
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 → 1) 1-pr(2 → 2) pr(2 → 2)

1 2

1-pr(1 → 1)
acknowledgements

Introduction

Word discovery

Morphology discovery

Phonology and probability

Sonority learning

Sonority

Phase space

Random starting points

Phase space

Learning...

French

French dynamics

Finnish vowels

Finnish vowels

Random starting points

Phase space

Vowel harmony learning

The future of bayesian linguistics

Phase space
English dynamics

Random starting points
Acknowledgements
Introduction
Word discovery
Morphology discovery
Phonology and probability
Sonority learning
Sonority
Phase space
Random starting points
Phase space
Learning...
French
French dynamics
Finnish vowels
Finnish vowels
Random starting points
Phase space
Vowel harmony learning
The future of bayesian linguistics
French dynamics

Acknowledgements
Introduction
Word discovery
Morphology discovery
Phonology and probability
Sonority learning

Sonority
Phase space
Random starting points
Phase space
Learning...

French

French dynamics

Finnish vowels
Finnish vowels
Random starting points
Phase space

Vowel harmony learning

The future of bayesian linguistics
- We can easily separate vowels and consonants in Finnish.

- What if we find the best 2-state system to generate the vowel sequences?
<table>
<thead>
<tr>
<th>Vowel</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>ö</td>
<td>999</td>
</tr>
<tr>
<td>ä</td>
<td>961</td>
</tr>
<tr>
<td>y</td>
<td>309</td>
</tr>
<tr>
<td>e</td>
<td>0.655</td>
</tr>
<tr>
<td>i</td>
<td>0.148</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Log ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>o</td>
<td>-7.66</td>
</tr>
<tr>
<td>a</td>
<td>-927</td>
</tr>
<tr>
<td>u</td>
<td>-990</td>
</tr>
</tbody>
</table>

**Front vowels**

1. \( \text{C1} \) to \( \text{C2} \)
   - 0.90

2. \( \text{C2} \) to \( \text{C1} \)
   - 0.97

**Back vowels**

1. \( \text{C3} \) to \( \text{C4} \)
   - 0.10

2. \( \text{C4} \) to \( \text{C3} \)
   - 0.03
Transition probability from front V to front V

Transition probability from back V to back V

Finnish dynamics
Phase space

Finnish VH

harmony region

alternating region

French CV

English CV

\( pr(2 \rightarrow 2) \)

\( pr(1 \rightarrow 1) \)
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.
The numbers support the intuition.
In the context of autosegmental models, the inverse log conditional probability of certain transitions can be calculated. For instance, the probability of transitioning from a consonant (C) to a vowel (V) might be given by a formula like:

\[ P(C \rightarrow V) = \frac{1}{1 + e^{-\lambda(C \rightarrow V)}} \]

Where \( \lambda(C \rightarrow V) \) is the log odds of the transition. This function allows for a smooth transition between segments based on the log odds, facilitating the modeling of sonority and other phonological features in a probabilistic framework.
<table>
<thead>
<tr>
<th>päätuotetta \textit{(päätuote = main product)}</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Word probability in unigram model:</td>
<td>$2.00 \times 10^{-15}$</td>
</tr>
<tr>
<td>Plog sum in unigram model:</td>
<td>48.8</td>
</tr>
<tr>
<td>Word probability in bigram model:</td>
<td>$6.37 \times 10^{-12}$</td>
</tr>
<tr>
<td>Plog sum in bigram model:</td>
<td>37.2</td>
</tr>
<tr>
<td>Word prob in autosegmental model:</td>
<td>$2.41 \times 10^{-14}$</td>
</tr>
<tr>
<td>Plog sum in autosegmental model:</td>
<td>45.2</td>
</tr>
</tbody>
</table>
- 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
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).