Towards a new empiricism in linguistics

John A. Goldsmith
The University of Chicago
A touch of history
Classical Rationalism vs Empiricism

Knowledge is individual

Rationalists

Leibniz 1646-1716

Locke 1632-1704

Spinoza 1632-1677

Empiricists

Hume 1711-1776

Berkeley 1685-1753

Descartes 1596-1650
20th Century

Logical positivism, logical empiricism

A new rationalism

Rudolf Carnap
Hans Reichenbach
Noam Chomsky
Finding a synthesis

I will present a new empiricism today---but there is a touch of irony in the name: The new empiricism must include all that was important in the old rationalism as well as the old empiricism.
Empiricism / Rationalism

1. Prototype of knowledge is sensory: vision.
2. Innate knowledge is not rich in information.
3. Frequency is relevant: occurrences of events can be counted and measured profitably.
4. Knowledge is always labeled by a degree of (un)certainty.

1. Prototype of knowledge is mathematical—timeless.
2. Innate knowledge is like any other kind of knowledge.
3. What is important does not occur at a particular moment.
4. Knowledge is certain, by definition.
1. Empiricism / Rationalism

1. Prototype of knowledge is sensory: vision.
   “I just saw a shooting star!“
   “Most subject NPs in English are "pronouns.

1. Prototype of knowledge is mathematical—timeless.
   “There are an infinite number of prime numbers."
   “Sentences in English take the form Subject-Verb-Object”
2. Empiricism / Rationalism

Innate knowledge is not rich in information. What we come to the world with is a set of general strategies for finding coherence of various kinds in experience.

Innate knowledge is like any other kind of knowledge. Human knowledge can be best modeled as a logical or mathematical proof. Some of the assumptions in the proof do not come from experience.
3. Empiricism / Rationalism

Frequency is relevant: occurrences of events can be counted and measured profitably.

What is important does not occur at a particular moment.
4. Empiricism / Rationalism

Knowledge is always labeled by a degree of (un)certainty.

Knowledge is certain, by definition.
Fundamental issues

1. **induction**: How do we construct a theory that projects *from* observed data *to* not-yet-observed predictions?

2. **disciplinary autonomy**: How does linguistics relate to psychology and other disciplines?

3. **richness of innate schemata**: How do we find the proper balance of the Learned and the Unlearned?

4. **data**: What is the nature of the data upon which linguistics rests?

5. **science**: What does it mean to take linguistics to be a science?
Some red herrings:

- **Behaviorism**: empiricists feel no desire to be behaviorists.
- **The search for explanation**: empiricists are just as interested in finding explanation and understanding.
- **Data fetishes**: empiricists feel free to be data fetishes, but no reason to urge others to be. They also feel free to be search for the simplest mathematical formula.
1. **induction**: How do we construct a theory that projects *from* observed data *to* not-yet-observed predictions?

2. **disciplinary autonomy**: How does linguistics relate to psychology and other disciplines?

3. **richness of innate schemata**: How do we find the proper balance of the Learned and the Unlearned?

4. **data**: What is the nature of the data upon which linguistics rests?

5. **science**: What does it mean to take linguistics to be a science?
1. Probability as answer to the problem of induction

The problem of induction:

**Q:** How can we pass from a belief about particulars to a belief in a generalization?

**A:** With a probabilistic account:

1. An enumeration of all possible outcomes \( \{e_i\} \), and

2. A weight assigned to each: \( \text{pr}(e_i) \)…. 
Probabilistic account

What is a probabilistic account?
1. An enumeration of all possible outcomes \( \{e_i\} \);
2. A weight assigned to each: \( \text{pr}(e_i) \);
3. All probabilities are greater than 0: \( \text{pr}(e_i) \geq 0 \); and
4. They sum to 1: \( \sum \text{pr}(e_i) = 1.0 \)
That may not be what you thought a probabilistic account was—
But it is.
Probabilistic accounts are not inherently fuzzy or informal.
They are inherently both formal and quantitative.
Probability is the quantitative theory of evidence.

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

James Clark Maxwell: 1850
A probabilistic grammar

...assigns a weight to each representation generated by the grammar.

Is it clear that the sum of an infinite number of terms can equal 1.0?

\[ 1 = 0.5 + 0.25 + 0.125 + 0.0625 + 0.03125 + \ldots \]

\[ 1 = 0.9 + 0.09 + 0.009 + 0.0009 + 0.00009 + \ldots \]
But probabilists prefer “inverse log probabilities” (plog)

<table>
<thead>
<tr>
<th>Probability</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>0.125</td>
<td>3</td>
</tr>
<tr>
<td>0.000 977</td>
<td>10</td>
</tr>
<tr>
<td>0.000 0305</td>
<td>15</td>
</tr>
<tr>
<td>0.000 000 953</td>
<td>20</td>
</tr>
<tr>
<td>0.000 000 000 931</td>
<td>30</td>
</tr>
</tbody>
</table>

Think of this as something like a measure of complexity.
The probabilist’s answer to the riddle of induction

First part of answer:
A probabilistic model $m$ assigns probability to possible sets of observations, but $m$ is just one of many possibilities.
We choose the particular model $m$ which allocates more of its probability to the actual, observed universe than any other model does.
The probabilist’s answer to the riddle of induction

We use probability to judge the model, not the data.
The probabilist’s answer to the riddle of induction

*Second* part of answer:
We also want the theory to be simple.
What’s *simple*?
What’s simple?

There are many parochial, local notions of simple, and only one general, universal notion of what is simple.

- The general, universal notion of what is simple only works for algorithms.
- While finding algorithms always requires creativity and insight…
- Evaluating them is deterministic and straightforward, and involves…
- Algorithmic complexity.
Algorithmic complexity

- The length of the shortest computer program for a universal computer that performs the task you are interested in.
- Kolmogorov, Solomonoff, Chaitin, and others.
How do we construct a number algorithmically?

- 0.10100100010000100001000001…

M=0 and \( n=0 \).

Loop indefinitely:

- Add \( n \) 0's to the right end of M;
- Add a 1;
- Add 1 to \( n \);
- continue with loop.

The simplicity of the description of the best method defines the simplicity of the number itself.
We do much the same thing when comparing grammars
A grammar assigns a probability to each string of symbols.

\[ g : \Sigma^* \rightarrow [0,1] \]

\[ \sum_{s \in \Sigma^*} g(s) = 1 \]
A prior over grammars

$$\pi_G : G \rightarrow [0,1]$$

$$\sum_{g \in G} \pi_G(g) = 1$$

A theory assigns a probability to each grammar.

We can have a truly Universal Grammar if we use algorithmic complexity.
What is a probabilistic grammar, really?

A probabilistic grammar’s primary goal in life is to evaluate grammars, not to evaluate data.

Take home message

Probabilities arise from a model (i.e., a theory); they are not simply read off of observations.
Bayesian reasoning and seeking the Minimum Description Length

The description length of a set of data D, given a grammar g, is
Length of grammar g + pLog probability of the D assigned by g

Both are measured in bits
Minimize the Description Length of a corpus

Find the grammar $g$ that minimizes:

$$\text{length}(g) + p \log \text{prob}(D \mid g)$$

This is equivalent to finding the grammar $g$ whose probability is the greatest, given the corpus. (We will see below that we are guaranteed that this is a positive number.)
The heart of the new empiricism

We need **skill** and **knowledge** to know how to obtain important data.

We need **skill** and **knowledge** to figure out how to develop probabilistic models for the data.

We need to minimize an expression which puts equal emphasis on *theory* and *data*:

\[
DL = \text{Grammar length} + p\log \text{prob (data)}
\]
Minimum description length

Extension of the work on algorithmic complexity. Developed notably by Jorma Rissanen.
Fundamental issues

1. **induction**: How do we construct a theory that projects *from* observed data *to* not-yet-observed predictions?

2. **disciplinary autonomy**: How does linguistics relate to psychology and other disciplines?

3. **richness of innate schemata**: How do we find the proper balance of the Learned and the Unlearned?

4. **data**: What is the nature of the data upon which linguistics rests?

5. **science**: What does it mean to take linguistics to be a science?
The rise of linguistics as a discipline

- 1870s: William Dwight Whitney
- 1924: Founding of the LSA and of the first Linguistics Departments.
  - The rise of a belief in the independence and legitimacy of linguistics’ **methods** as the best scientific methods in all of the social sciences.
The science of language, dealing with the most basic and simplest of human social institutions, is a human (or mental or, as they used to say) moral science. It is most closely related to ethnology, but precedes ethnology and all other human sciences in the order of growing complexity, for linguistics stands at their foot, immediately after psychology, the connecting link between the natural sciences and the human. The methods of linguistics resemble those of the natural sciences, and so do its results, both in their certainty and in their seeming by no means obvious, but rather, in many instances, paradoxical to the common sense of the time.
We are casting off our dependence on psychology, realizing that linguistics, like every science, must study its subject-matter in and for itself, working on fundamental assumptions of its own; that only on this condition will our results be of value to related sciences (especially, in our case, to psychology) and in the light of these related sciences in the outcome more deeply understandable.

In other words, we must study people’s habits of language—the way people talk—without bothering about the mental processes that we may conceive to underlie or accompany these habits. We must dodge this issue by a fundamental assumption, leaving it to a separate investigation, in which our results will figure as data alongside the results of the other social sciences.
Fundamental issues

1. **induction**: How do we construct a theory that projects *from* observed data *to* not-yet-observed predictions?

2. **disciplinary autonomy**: How does linguistics relate to psychology and other disciplines?

3. **richness of innate schemata**: How do we find the proper balance of the Learned and the Unlearned?

4. **data**: What is the nature of the data upon which linguistics rests?

5. **science**: What does it mean to take linguistics to be a science?
If we could look inside someone’s head to see how much of our knowledge of language was learned and how much was not... What would we see?
Non-learned

Learned
Which is it?

Non-learned

Learned

Non-learned

Learned
If most linguistic knowledge is not learned, then we need to develop methods to uncover that hidden knowledge.

If most of it is learned, then we need to understand the ways by which it can be learned.
Linguists and computer scientists have taken up that challenge, and developed methods for inducing linguistic knowledge from data.

I will talk about some of my work on this below.
1. **induction**: How do we construct a theory that projects *from* observed data *to* not-yet-observed predictions?

2. **disciplinary autonomy**: How does linguistics relate to psychology and other disciplines?

3. **richness of innate schemata**: How do we find the proper balance of the Learned and the Unlearned?

4. **data**: What is the nature of the data upon which linguistics rests?

5. **science**: What does it mean to take linguistics to be a science?
The nature of linguistic data

Linguists today are faced by a rich range of options:
On-line corpora, especially from the internet;
Powerful computers, which can handle complex hypotheses and probabilistic models with little sweat; and sets of data many orders of magnitude larger than had been possible in the past.
Fundamental issues

1. **induction**: How do we construct a theory that projects *from* observed data *to* not-yet-observed predictions?

2. **disciplinary autonomy**: How does linguistics relate to psychology and other disciplines?

3. **richness of innate schemata**: How do we find the proper balance of the Learned and the Unlearned?

4. **data**: What is the nature of the data upon which linguistics rests?

5. **science**: What does it mean to take linguistics to be a science?
Linguistics as a science

There are many ways to do linguistics. This is only one of them. The goal of linguistics is to find the shortest description of all of the linguistic data that has been collected. The description length is always positive; therefore there is a minimum.*
You have to build the simplest grammar you can;
I can tell you how to measure that simplicity, with just a little roughness around the edges;
And you are tested on how well your grammar accounts for all of the data that has been collected, and your grammar’s simplicity. With no subjectivity.
What kind of linguistics is that?

Is it scientific? Yes. Doing it right requires the same skills at grammar design that linguistics always has required.

Is it about the human brain? Maybe, but not in an obvious fashion. IMHO, it is unquestionably about the mind, but that opinion is irrelevant.
Is linguistics a branch of psychology?

As the earliest linguists argued: the answer is No.

But linguistics has much to offer psycholinguists: help in framing hypotheses.

Linguistics has no claim to determine the outcome of their results.

But theoretical linguistics is answering a different scientific question.
Chomsky’s argument

Either linguistics is a science, or it is not. If it is a science, then it is a science of something that exists in the physical world. If it is, then the only plausible candidate for that something is the human brain. The study of the functions of the brain is psychology. QED.
What's wrong with that?

“The only plausible candidate for that something is the human brain.”

Nothing else? Not linguistic data?

That's why Chomsky asserts that the study of E-language is incoherent. This is a scientific account of linguistics as the study of E-language.
In practice:

Linguistica

<table>
<thead>
<tr>
<th>Signatures</th>
<th>Exemplar</th>
<th>Corpus Count</th>
<th>Stem Count</th>
<th>Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL.s</td>
<td>abuse</td>
<td>1793</td>
<td>445</td>
<td>3967</td>
</tr>
<tr>
<td>ed</td>
<td>accelerate</td>
<td>1657</td>
<td>457</td>
<td>1114</td>
</tr>
<tr>
<td>ing</td>
<td>embezzl</td>
<td>1046</td>
<td>258</td>
<td>1047</td>
</tr>
<tr>
<td>NULL.ly</td>
<td>absolute</td>
<td>369</td>
<td>101</td>
<td>961</td>
</tr>
<tr>
<td>: ly</td>
<td>alarming</td>
<td>1119</td>
<td>148</td>
<td>294</td>
</tr>
<tr>
<td>er</td>
<td>14-pow</td>
<td>4726</td>
<td>424</td>
<td>958</td>
</tr>
<tr>
<td>NULL.ed.ing.s</td>
<td>account</td>
<td>484</td>
<td>35</td>
<td>798</td>
</tr>
<tr>
<td>: NULL.ed.ing</td>
<td>approach</td>
<td>263</td>
<td>40</td>
<td>649</td>
</tr>
<tr>
<td>: NULL.ed.s</td>
<td>affect</td>
<td>282</td>
<td>43</td>
<td>620</td>
</tr>
</tbody>
</table>

Stems:
- account
- appeal
- ask
- assault
- attack
- award
- belong
- board
- claim
- demand
- explain
- export
- extend
- fear
- happen
- interest
- kick
- lock
- market
- offer
- panel
- point
- record
- remain
- represent
- request
- result
- return
- staff
- succeed
- talk
- train
- want
- word
The Linguistica group at the University of Chicago draws its membership from the Department of Linguistics and the Department of Computer Science. Our core interest is unsupervised learning of natural language structure, but this interest has taken us to work in a number of other areas, including automatically obtaining corpora through the Internet, and the discovery of structure in bioinformatic databases.

This site contains a good number of details about the Linguistica
Linguistica Project

Open source C++ software which accepts a large text in any language and produces, as its output, a morphology. A morphology is a list of affixes, stems, and a finite state automaton that generates words with them, plus the morphophonemics.
The key is to build an automatic linguist who uses Minimum Description Length as its constant measuring stick for determining what is the best analysis of the data.

Linguistica looks for the shortest description length of the corpus, and we test its conclusions to see whether they match linguists’ understanding.
<table>
<thead>
<tr>
<th>Signatures</th>
<th>Exemplar</th>
<th>Descr. Length</th>
<th>Corpus Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL.s</td>
<td>Action</td>
<td>11653.3</td>
<td>8021</td>
</tr>
<tr>
<td>:s</td>
<td>bleacher</td>
<td>16.5742</td>
<td>2</td>
</tr>
<tr>
<td>'s.NULL</td>
<td>Adair</td>
<td>4878.06</td>
<td>6508</td>
</tr>
<tr>
<td>:'s</td>
<td>Airline</td>
<td>2563.54</td>
<td>494</td>
</tr>
<tr>
<td>NULL.ly</td>
<td>According</td>
<td>3836.48</td>
<td>2933</td>
</tr>
<tr>
<td>:ly</td>
<td>Actual</td>
<td>3647.92</td>
<td>1870</td>
</tr>
<tr>
<td>NULL.ed.ing.s</td>
<td>account</td>
<td>1047.81</td>
<td>2817</td>
</tr>
<tr>
<td>:ing</td>
<td>Represent</td>
<td>8157.98</td>
<td>2245</td>
</tr>
<tr>
<td>:ed</td>
<td>centraliz</td>
<td>14926.3</td>
<td>5762</td>
</tr>
<tr>
<td>:NULL.ed</td>
<td>Accept</td>
<td>1552.34</td>
<td>397</td>
</tr>
<tr>
<td>:ed.ing</td>
<td>Display</td>
<td>1150.94</td>
<td>262</td>
</tr>
<tr>
<td>:NULL.ed.s</td>
<td>Ask</td>
<td>777.912</td>
<td>874</td>
</tr>
<tr>
<td>:NULL.ing</td>
<td>Aid</td>
<td>1243.46</td>
<td>596</td>
</tr>
<tr>
<td>:NULL.ed.ing</td>
<td>View</td>
<td>616.543</td>
<td>1374</td>
</tr>
<tr>
<td>:NULL.ing.s</td>
<td>Fall</td>
<td>635.745</td>
<td>966</td>
</tr>
<tr>
<td>:ing.s</td>
<td>Record</td>
<td>213.994</td>
<td>33</td>
</tr>
<tr>
<td>:ed.ing.s</td>
<td>proclaim</td>
<td>35.869</td>
<td>9</td>
</tr>
<tr>
<td>er</td>
<td>Abn..</td>
<td>10214.5</td>
<td>8041</td>
</tr>
<tr>
<td>Words</td>
<td>Stem</td>
<td>Mini-Lexicon 3</td>
<td>Mini-Lexicon 2</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>decline</td>
<td>declin</td>
<td></td>
<td>e</td>
</tr>
<tr>
<td>declined</td>
<td>declin</td>
<td></td>
<td>ed</td>
</tr>
<tr>
<td>declines</td>
<td>declin</td>
<td></td>
<td>es</td>
</tr>
<tr>
<td>decolletage</td>
<td>decolletage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decor</td>
<td>decor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decorate</td>
<td>decor</td>
<td></td>
<td>at e</td>
</tr>
<tr>
<td>decorating</td>
<td>decor</td>
<td></td>
<td>at ing</td>
</tr>
<tr>
<td>decoration</td>
<td>decor</td>
<td></td>
<td>at ion</td>
</tr>
<tr>
<td>decorations</td>
<td>decor</td>
<td></td>
<td>at ion s</td>
</tr>
<tr>
<td>decorative</td>
<td>decor</td>
<td></td>
<td>at ive</td>
</tr>
<tr>
<td>decorator</td>
<td>decor</td>
<td></td>
<td>at or</td>
</tr>
<tr>
<td>decorators</td>
<td>decor</td>
<td></td>
<td>at or s</td>
</tr>
<tr>
<td>decrease</td>
<td>decrease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decree</td>
<td>decree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decreeing</td>
<td>decree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>decried</td>
<td>decri</td>
<td></td>
<td>ed</td>
</tr>
<tr>
<td>decries</td>
<td>decri</td>
<td></td>
<td>es</td>
</tr>
<tr>
<td>dedicated</td>
<td>dedicat</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Corpus

Pick a large corpus from a language -- 5,000 to 1,000,000 words.

Exactly how is MDL used to learn a grammar?
Corpus

Bootstrap heuristic

Feed it into the “bootstrapping” heuristic...
Corpus

Bootstrap heuristic

Morphology

Out of which comes a preliminary morphology, which need not be superb.
Corpus

→

Bootstrap heuristic

→

Morphology

→

incremental heuristics

Feed it to the incremental heuristics...
Out comes a modified morphology.
Corpus

Bootstrap heuristic

Morphology

Is the modification an improvement? Ask MDL!

modified morphology

incremental heuristics
If it is an improvement, replace the morphology...
Corpus

Bootstrap heuristic

modified morphology

Incremental heuristics

Send it back to the incremental heuristics again...
Continue until there are no improvements to try.
Proposition:

The correct morphology of a language is the FSA that provides the shortest description length of the data.

Find the morphology with the greatest probability, given the data.
Phonology

Sonority: consonant/vowel split
Vowel harmony
Syllable structure

What 2-state first-order device is most probable, given the data?
One that divides the segments into consonants and vowels
Finnish vowel harmony
Final thoughts on probability

The essence of the present theory is that no probability, direct, prior, or posterior, is simply a frequency.

Sir Harold Jeffreys: 1939
• Two philosophers who disagree about a point should, instead of arguing fruitlessly and endlessly, be able to take out their pencils, sit down amicably at their desks, and say "Let us calculate."

Gottfried von Leibniz
(1646 – 1716)
It is seen in this essay that the theory of probabilities is at bottom only common sense reduced to calculus; it makes us appreciate with exactitude that which exact minds feel by a sort of instinct without being able ofttimes to give a reason for it.

*Philosophical Essay on Probabilities* (1814)
Linguistics is still in the process of working out what it is.

There is no one single answer to that question anyway.

The relationship of data and theory remains a thorny question, to which MDL and Bayesianism gives a very appealing answer.