Towards a new empiricism in linguistics

John Goldsmith

May 26, 2007
Outline

1. Rationalism and empiricism
2. Learning and linguistics
3. Three models of linguistic theory
4. MDL: Minimum Description Length
This talk is relatively abstract.

It gives the background for the talk on morphology, which is very concrete.

A look ahead: the central question is the passage from the finite to the infinite—how should this be done? For generative grammar, the answer is with finite algorithms (“infinite languages described with finite means.”). For empiricism, the answer is extended to distributions, i.e., an infinite set of numbers that sum to 1.0.
Outline

1. Rationalism and empiricism
2. Learning and linguistics
3. Three models of linguistic theory
4. MDL: Minimum Description Length
Empiricism & Rationalism

The 17th and 18th century movement in natural philosophy against traditional authority of the Catholic church and Aristotelian scholastics.
Rationalism:
- Descartes
- Spinoza
- Leibniz

Empiricism
- Locke
- Hume
- Re-emergence of the term **empiricism** in 1920s: logical empiricism.
- Re-emergence of the term **rationalism** in 1960s: in linguistics.

The modern sense of empiricism came with an additional sense of skepticism concerning abstract theoretical objects. In the 17th century, this skepticism was shared by rationalists and empiricists. Today: skepticism with respect to Universal Grammar. The question: can we do everything that a linguist wants to do assuming nothing but logic and information theory—something that is truly universal.
Principal issues

Empiricist

- Prototype of knowledge is *sensory*, vision;
- Innate knowledge is not rich in information.
- Frequency is relevant: occurrences of events can be counted profitably.
- Knowledge is always labeled by a degree of (in)certainty

Rationalist

- Prototype of knowledge is *mathematics*, timeless;
- Innate knowledge is like any other kind of knowledge.
- What is important does not happen at a particular moment
- Knowledge is, by definition, certain.
Outline

1. Rationalism and empiricism
2. Learning and linguistics
3. Three models of linguistic theory
4. MDL: Minimum Description Length
Chomsky’s original message

- Understanding learning is hard.
- The problem of learning = problem of language acquisition = the central problem of linguistics.
- This parallels the problem of induction in the philosophy of knowledge, a problem that may not yet be solved (influence of Nelson Goodman).
- Consequence: emphasis on the study of the *unlearned*. 
The problem of induction is: how can we pass from a belief in particulars to a belief in a generalization?

How do we pass from an observation of a small number of cases to a belief in a generalization (with infinite set of predictions)?

**Data set 1**

Mapitunorama
The problem of induction is: how can we pass from a belief in particulars to a belief in a generalization?

How do we pass from an observation of a small number of cases to a belief in a generalization (with infinite set of predictions)?

**Data set 1**
Mapitunorama

**Data set 2**
Mapitunorama; ramatunomapi.
### The problem of induction

The problem of induction is: how can we pass from a belief in particulars to a belief in a generalization?

How do we pass from an observation of a small number of cases to a belief in a generalization (with infinite set of predictions)?

<table>
<thead>
<tr>
<th>Data set 1</th>
<th>Mapitunorama</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set 2</td>
<td>Mapitunorama; ramatunomapi.</td>
</tr>
<tr>
<td>Data set 3</td>
<td>Mapitunorama; ramatunomapi. Sutotunorama.</td>
</tr>
</tbody>
</table>
Data set 3
Mapitunorana; ramatunomapi. Sutotunorana.

How do we evaluate our success?
- Test predictions of unseen data. (Prediction: Ramatunosuto, right?)
- Evaluate how well our analysis treats the observed data.

These two ways are very different.
What is linguistics?

Linguistics is the study of the problem of induction, focused on linguistic data. Chomsky’s conclusion was that the most important side of this problem lies in what was known without having been learned: the Innate.
In its early, general formulation, what is unlearned (=not learned) is the evaluation metric, which is essential to deciding what the best generalization for treating any set of data. In virtually all of its actual formulations (from Haj Ross 1967 through *Conditions on Transformation to Government and Binding*), the Unlearned took on one of two styles:

- Styles of information flow (=“components”), or
- Algebraic statements (e.g., Specified Subject Condition)

Nothing that is language-particular could be Unlearned; thus nothing that is language-particular is of fundamental interest to (this kind of) linguistics.
But that strategy completely ignores the other side of the coin:

- What aspects of Language L are *learned* by the speaker?
- Learning is hard; how are those aspects learned?

This is the heart of an empiricist view of linguistic theory.
What is the proper balance of the Learned and the Unlearned?
What is the proper balance of the Learned and the Unlearned?
This issue is not about functionalism.
There are two complementary tasks for linguistics:

- Study what is universal and innate, and common to all languages.
- Study what is not universal in each language, and figure out how a learner could learn this.

Different linguists may have different opinions as to which task is more interesting, more difficult, more rewarding, and so on.

There is no theoretical reason why the first should be more scientifically interesting than the second.
Important sequence in *Syntactic Structures* Chomsky 1957

Three models of linguistic theory

- First model of linguistics: *Proposing grammars*: Project data to grammar
- Second model of linguistics: *Evaluating grammars*: Use data to prove or disprove grammar
- Third model of linguistics: *Ranking grammars*: Use data to rank grammars
Zellig Harris and some computational linguists today:

\[ \text{Data} \rightarrow \text{Device}_1 \rightarrow \text{Grammar} \]
Model 2

Data → Device$_2$ → “G is/is not correct for D”

Grammar

The device simply tells us if the grammar is the correct one, given the data D.
Model 3

Data → $Grammar_1$ → $Device_3$ → “$Grammar_2$ is better.”

The device does not output the grammar: it indicates which grammar is the better, in view of the data.
Origins of these notions

Contrast between $Device_1$ and $Device_2$ is the difference between

- Finding a proof
- Checking a proof
Finding versus checking a proof

By the early 1950s

Logicians had noticed that there was a parallel between

- determining whether a proposition was a logical implication of a set of axioms,
- determining whether a formula was a well-formed expression in a given formal language.

Is the following a theorem: \( A \rightarrow \mathcal{L} \)?

That’s always a hard question: does a valid proof exist?

Is the following proof valid?: \( A \rightarrow L_1 \rightarrow L_2 \rightarrow \mathcal{L} \)

That is usually an easy question to answer.
The Model-2 view of linguistic theory would indeed be easier to implement than Model-1, if we knew that there were local properties of a “correct” grammar that could be identified. If there is a deterministic method to go from data to grammar, then that method should be implementable into this model. (Is that what Harris was trying to find? That’s open to debate.)

Data → Device₂ → “G is/is not correct for D” → Grammar
The first mention of a linguistic device that evaluates and ranks a set of analyses, rather than proposing or evaluating is in Rulon Wells (1947) *Immediate Constituents*. 
In choosing the [analysis] that will be used in dividing any utterance into constituents, it is necessary to consider the whole system which these [classes], taken collectively, form with each other. This is why an analysis is not pronounced good or bad of itself, but only better or worse than some other....We call an IC-analysis **wrong** when there is another possible analysis of the same sequence that is better, and **right** when there is none. p. 88.

In descriptive linguistics, discovery consists in finding the best scheme in terms of which to describe the facts; it is not strictly part of the exposition to show that that scheme is the best. p. 101.
What’s rationalistic? What’s empiricist?

- Model 3, augmented with a reasonable (but not perfect) search procedure, provides us with Model 1.
- Why shouldn’t linguists attempt to develop such search procedures?
- Chomsky notes that it is the highest goal. That is what we try to do with automatic learning of morphology. Why not, after all?
Outline

1. Rationalism and empiricism
2. Learning and linguistics
3. Three models of linguistic theory
4. MDL: Minimum Description Length
3 concepts of probability

- Stochastic element is inherent in the object. Gambling. Our understanding of the system leads us to a mathematical model whose parameters are inferred from the observations.

- Noise in the data. Social science experiments. We assume a relatively simple model, and infer the parameters from a desire to maximize the fit to the observations, despite the presence of noise. The term “statistics” is often used in this context.

- Formal complexity of encoding. Not inherently statistical.
All learning and all explanation is based on finding simple generalizations that maximally predict the observations. This is implicit in most scientific hypotheses, but we must make it explicit.
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 a language. 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.
The logic of confirmation

Confirmation function $C(g, D)$

We can prove that if

- the degree of confirmation of a grammar $g$ by a set of data $D$ is inversely related to the value of the function $C$; and
- The function $C$ is “extensive” at the sentence level ($C(S_1) + C(S_2) = C(S_1 + S_2)$)

then this forms a probability distribution $\frac{e^{-C(g, S)}}{Z}$
Probabilistic grammars

- First proposed by Ray Solomonoff in the mid 1950s
- Precisely as a solution to the problem of induction.
- Developed into the concept of algorithmic complexity
How can we evaluate how well we have accounted for the facts?

**Answer 1**: Test the model on predictions not observed at time of hypothesizing.

**Answer 2**: Use a probabilistic model. A probabilistic model can generate an infinite set of possible observations (possible forms). But it is obliged to distribute a finite amount of “reality” (=probability mass) over that infinite set. The observed data form a finite subset of the possible observations. We will choose the theory that assigns the highest probability to the observed data (all other things being equal).
All other things being equal?

All other things are never equal.

We do not want a theory that merely *repeats* the observed data. We want a theory whose formal simplicity is the guarantee that we do not over-fit the data—that is, that we do not simply repeat the data.
Suppose the data is abababab. Possible analyses:

- abababab
- \((ab)^4\)
- \((ab)^+\) (that is, any number > 0 of copies of \(ab\))
- \(ab(\?)^*ab\) (that is, \(ab\)...anything...\(ab\)).
Suppose the data is $abababab$. Probabilistic analyses:

- $abababab$ prob: 1.0. No future predictions.
- $(ab)^4$ prob 1.0. No future predictions.
- $(ab)^+ \text{ prob } \frac{1}{16}$, if $\text{pr}(ab) = 0.5$.
- $ab(?)^*ab \text{ prob } \left(\frac{1}{8}\right)^4 \frac{1}{2} = \frac{1}{8192}$

The better analyses assign higher probability to the observed data, because they *notice* generalizations.
A probabilistic grammar is a generative grammar which generates an (infinite) set of representations, plus a non-negative number $pr(x)$ associated with each representation $x$ such that $\sum_x pr(x) = 1$. 
We will do the same thing (impose conditions of probability) over the infinite set of grammars, theories, etc.
Algorithmic complexity

- **Solomonoff**: Choose the analysis which maximizes the probability of the data. This equals (the probability of the data, given the theory) times (the probability of the theory). Calculate the probability of the theory as \(\frac{1}{2^{|G|}}\), where \(G\) = length of the grammar (in bits).

- **Chaitin**: You cannot use \(\frac{1}{2^{|G|}}\) unless the grammar has the “prefix property” (a certain kind of punctuation structure).

- **Kolmogorov**: A generalization that you notice is **only significant** structure if it can be better summarized as an algorithm than as arbitrary markings.

Within narrow limits, we can calculate the length, in bits, of any grammar. I will do this tomorrow for a morphology. It is not hard to do (it’s not trivial, either!).
Rissanen’s Minimum Description Length

We want to find the most likely grammar, $g$ given the data $D$: find the grammar $g$ that maximizes $\text{pr}(g|D)$.

$$\text{pr}(g|D) = \frac{\text{pr}(D|g) \text{ pr}(g)}{\text{pr}(D)}$$

Taking logarithms:

$$\log \text{pr}(g|D) = \log \text{pr}(D|g) + \log \text{pr}(g)$$

Multiply by -1, and minimize instead of maximize:

$$-\log \text{pr}(g|D) = -\log \text{pr}(D|g) - \log \text{pr}(g)$$

= Optimal compressed length of data $D$ + length of grammar $g$. 

The bottom line

This is the bottom line, in my opinion, and it is reasonable to call it *empiricist*. Our goal is to explicitly compute the algorithmic complexity of any theory that we wish to take seriously, and we compare alternative theories by measuring, for each theory of grammar, the following: for each chunk of language data, the sum of the grammar complexity plus the data complexity according to that grammar, choosing the grammar for which that sum is the smallest.

\[
\sum_{\text{all languages } l} \min(|\text{grammar } g(l)| + \log \frac{1}{\text{prob}(\text{data}_l|g(l))})
\]
Classical problem

My grammars

Grammar of English
Grammar of Swahili

Your grammars

Grammar of English
Grammar of Swahili

English
Swahili
Classical generative solution

My UG & grammars

Your UG & grammars
Problem for classical solution

My UG compiler

Grammar of English
Grammar of Swahili

Your UG compiler!

Grammar of English
Grammar of Swahili
Every linguist can propose a UTM to be considered. Each linguist can try to make his/her UTM as friendly to his/her theory as s/he wishes. Principle of equity: We choose the Universal Turing Machine $u$ for which the total cost of mapping all other UTMs into $u$ is the smallest.

$$\arg\min_u \sum_{v \in \mathcal{U}} [v \gg u]$$
Autonomy of linguistics as a discipline, as a method

Conclusion:

- The goal of linguistics is to learn how to find the best linguistic model of language data, where how good a model is is based on grammar length and on probability it assigns to the observed data.
- This is always an empirical study; there is always a best analysis, and we can prove that there is no automatic way to find that best analysis. We must be creative linguists if we want to make progress.
- There is no need to say that linguistics is really psychology, or that it is really biology.
- Linguistics is not really psychology, and it is not really biology.
- It is linguistics.