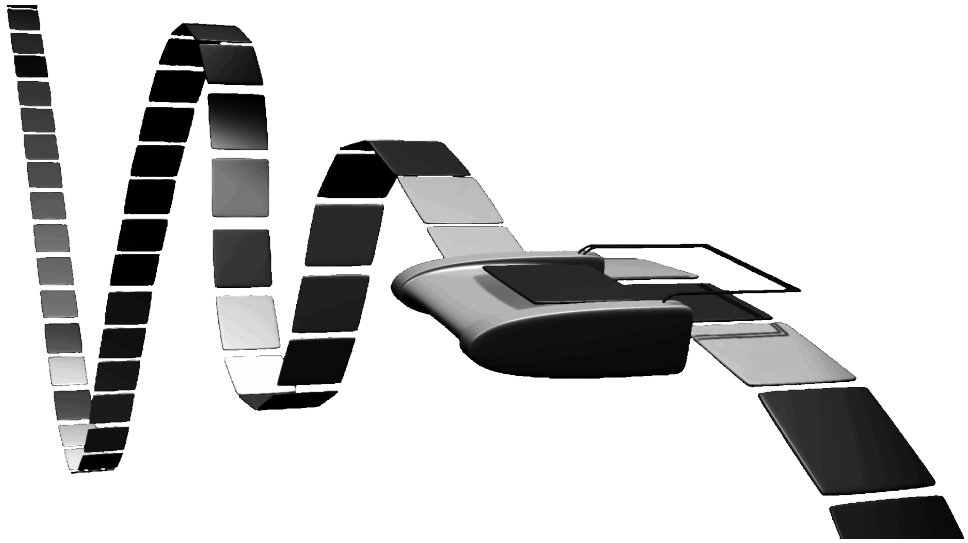
Your Turing machine or mine?

John Goldsmith

June 20, 2007

UTM = Universal Turing Machine



The range of functions computable by a Turing machine is far wider than anything we need to write a grammar.

A universal Turing machine is one that can "emulate" any Turing machine. There are many of them; some accept very short encodings to compute interesting and complex functions and others require spelling out in great detail.

What does this have to do with the learning of natural languages by human beings? A lot—and not very much!

The linguist wants to discover of language *can be learned*—while the psycholinguist wants to discover how language *is* learned.

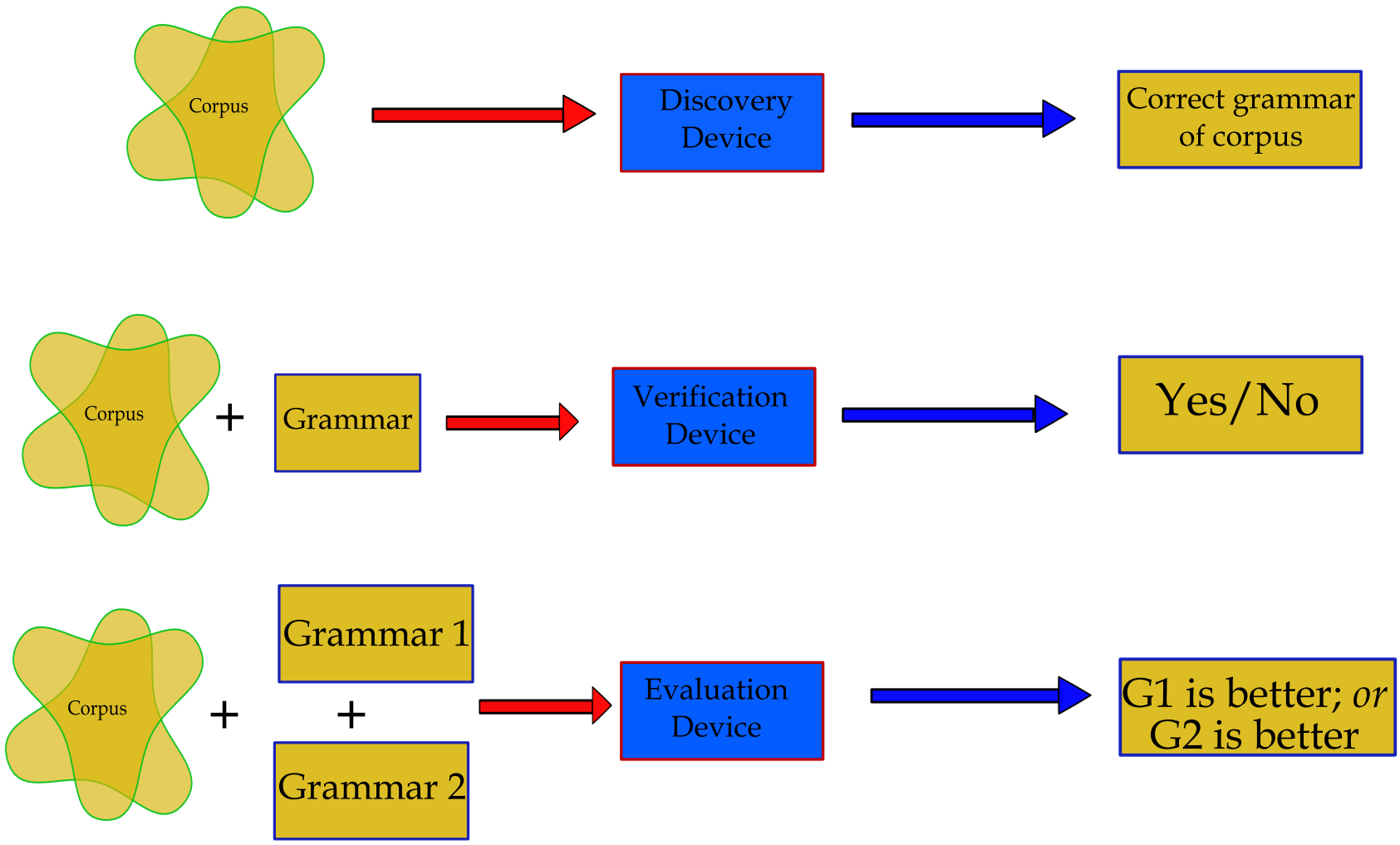
Obviously anything learned by the one is going to be very relevant to the other. Lurking behind this: it is very hard for the field of linguistics to figure out how to interpret and integrate findings that derive from psychological studies: those findings can *inspire* linguistic research, but psychological studies are already so embedded within particular views of what language is that it is extremely difficult to take them as some sort of boundary conditions for linguistic analysis.

My suggestion is that we develop and make explicit a strictly linguistic understanding of how language works, one which can be utilized by psycholinguists in turn.

- What linguistics is, mid 20th Century
- Ray Solomonoff and probabilistic grammars
- Algorithmic complexity: and there is no such thing as a free lunch
- Proposal: a way to avoid the arbitrary convention of UTM-choice

What is linguistic theory?

Chomsky's 3 models, from his *Logical Structure of Linguistic Theory* 1955



Chomsky's three views of linguistic theory

Chomsky believed that we could and should account for grammar selection on the basis of the formal simplicity of the grammar, and that the specifics of how that simplicity should be defined was a matter to be decided by studying actual languages in detail. In the last stage of classical generative grammar, Chomsky went so far as to propose that the specifics of how grammar complexity should be defined is part of our genetic endowment.

Chomsky proposed the following methodology, in two steps.

- Linguists should develop formal grammars for individual languages, and treat them as scientific theories, whose predictions could be tested against native speaker intuitions among other things. Eventually, in a fashion parallel to the way in which a theory of physics or chemistry is tested and improved, a consensus will develop as to the form and shape of the *right* grammar, for a certain number of human languages.

- Linguists will be formulating their grammars with an eye to what aspects of their fully specified grammars are universal and what aspects are language-particular. And here is where the special insight of generative grammar came in: Chomsky proposed that it should be possible to specify a higher-level language in which grammars are written which would have the special property that the *right* grammar was also the *shortest* grammar that was compatible with any reasonable sized sample of data from any natural language. If we could do that, Chomsky would say that we had achieved our final goal, *explanatory adequacy*.

There were two other assumptions made (though occasionally questioned, as we will see) by this early Chomskian approach.

- The *first* [Small role of data]: the role played by the data from the language was minor, and could be used in a simple manner to throw out from consideration any grammar which failed to generate the data.

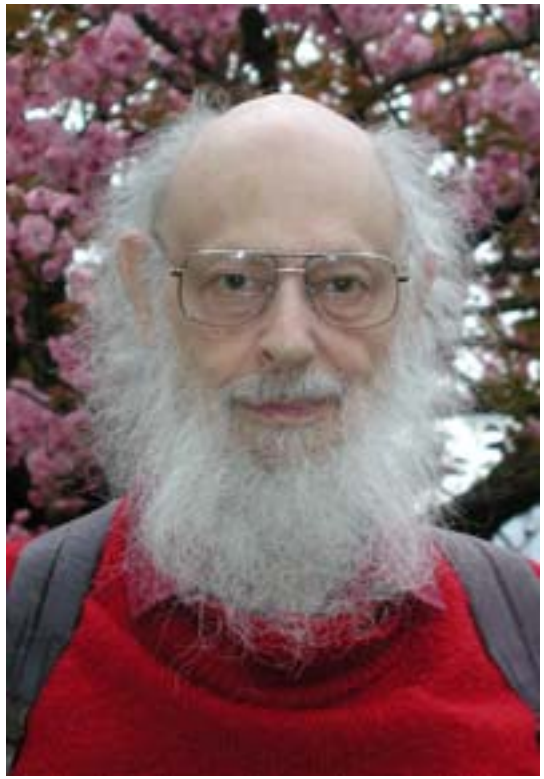
- The *second* [Ranking can be based on grammar length] was that once we eliminate grammars from consideration that do not generate the right data, we can rank them on the basis of some formal property of the grammars themselves; we develop a formalism in which program length provides that formal property: the shortest grammar will always be the one predicted to be the right one.

Two fatal flaws to this program:

- The Small role of data assumption is untenable, and there were people, like Ray Solomonoff, who were working to demonstrate this at exactly the same moment in the 1950s.

- The Universal Grammar is Free Grammar fallacy. Chomsky's actual method of theory development put a strong emphasis on developing the grammar-writing linguistic theory (Universal Grammar) right from the start, and the linguist needed to pay a certain "cost" for adding complexity to the grammars of individual languages, but there was no "cost" for adding complexity to the Univeral Grammar.

Solomonoff and the logic of confirmation



Chomsky in *Language and Mind, 1968* pp. 76-77:

> 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 sense. Notice that it is a great oversimplification to suppose thata 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.

Probabilistic models

$$\sum_{r \in \mathcal{R}} pr(r) = 1.0$$

Blending together Solomonoff's work with that other people's (notably Kolmogorov, Chaitin, and Rissanen):

We can naturally assign a probability distribution over grammars, based on their code length in some appropriate universal algorithmic language, such as that of a Universal Turing machine.

If such grammars are expressed in a binary encoding, and if they are "self-terminating", then we assign each grammar g the probability $2^{-|g|}$

If we accept the following assumptions:

- Our grammars are probablistic (they assign a distribution to the representations they generate);
- The goal, or one major goal, of linguistics is to produce gramamrs;
- There is a natural prior distribution over algorithms, possibly modulo some concern over choice of universal Turing machine or its equivalent;

then we can conclude:

- there is a natural formulation of the question, what is the best grammar, given the data D? The answer is:

$$\arg\max_{g} \; pr(g)pr(D|g)$$

where

$$pr(g) := 2^{-|g|}$$

Abstract and idealized view of linguistics: The linguist has a sample of data from a set of languages, $\mathcal{L}$ and a computer and a computer language for grammars of each language.

English and Japanese have different structures, and probability must be assigned according to different models. Some parts of the model are specifically set aside for treating sentences from English, some for treating sentences from Japanese, and other parts will be relevant for both.

Grammar-writers' premise: to develop a Linguistic Theory which is a way of writing grammars of any human language. This Linguistic Theory is in effect a higher-level computer language which, when given a complete grammar, can perform tasks that require knowledge of language, and only knowledge of language (like parsing, perhaps).

Rissanen and There ain't no such thing as a free UG

We noted above: two fatal flaws to the classical generative picture:

- Universal Grammar is Free Grammar fallacy: while the complexity of a particular grammar counts against it as a scientific hypothesis, the complexity of Universal Grammar has no cost associated with it from a scientific point of view. Its complexity may be the result of millions of years of evolutionary pressure—or not; the linguist neither knows nor cares.

- Failure to deal with the relationship between grammar and data (the "small role of data" flaw).

An *almost* perfect scientific linguistic world in which there is a competition between a certain number of groups of researchers, each particular group defined by sharing a general formal linguistic theory.

1. You adopt an approved Universal Turing machine (UTM^1). The set of such machines that have already been approved is $\mathcal{U}$.

2. You adopt a set of corpora which constitutes the data for various languages; everyone in the group must adopt all approved corpora.

3. The activities involved in this competition are the following. You will have access to the data $\mathcal{C}$, and you will select a Universal Turing Machine of your choice, you will come up with a Universal Grammar UG, and a set of grammars $\{\Gamma_l\}$ for each language. Again, a bit more slowly:

1. You will write a Universal Grammar which runs on the UTM that you have adopted, UTM^1 in your case, let's say. You will compute the length of your UG_i as it runs as a compiler on your UTM.

2. You will write a probabilistic grammar for each language $l \in \mathcal{L}$, and these grammars are written in the formal language required by your Universal Grammar. We will indicate a grammar by an upper case gamma Γ; a particular grammar is Γ_k. Every Universal Grammar must have the ability to calculate the length of any grammar proposed by any group—not just grammars that their own group proposes, and this length must be finite (this is a "broadness" condition).

3. You will use the probabilistic grammars to extract redundancies from the corpora, and then

4. You will calculate two quantities: (a) the length of the Universal Grammar UG_1, and (b) the length of the linguistic analyses (the "empirical term"): Symbolically, we can express the length of the linguistic analyses with an "empirical term" for a given triple, consisting of: UTM^{α}; a set of grammars $\{\Gamma_l\}$; and the length of the unexplained information in each corpus:

$$Emp(UTM^{\alpha}, UG^{i}, \{\Gamma_l\}, \mathcal{C}) = \sum_{l \in \mathcal{L}} |\Gamma_l|_{UG^i} - log\ pr_{\Gamma_l}(\mathcal{C}_l) \qquad (1)$$

This group is trying to minimize the quantity:

$$|UG^i|_{UTM^\alpha} + Emp(UTM^\alpha, UG^i, \{\Gamma_l\}, \mathcal{C}) \tag{2}$$

which is essentially the minimal description length of the data, given UTM_1.

Another way to put this is that we are doing standard MDL analysis, but restricting our consideration to the class of models where we know explicitly how to divide the models for each language into a universal part and a language-particular part.

We *might* then imagine you win the competition if you can demonstrate that the final sum that you calculate in (3) is the smallest of all the groups. But this won't work correctly, because it is perfectly possible that two competitors will each find that their own systems are better than their competitors' systems, because of the UTM that they use to find the minimum. That is, suppose we have two groups, Group 1 and Group 2, which utilize UTM^{α} and UTM^{β}. Then

It is perfectly possible (indeed, it is natural) to find that

$$\begin{aligned} &|UG^i|_{UTM^\alpha} + Emp(UTM^\alpha, UG^i, \{\Gamma_l\}^1, \mathcal{C}) \ < \\ &|UG^j|_{UTM^\alpha} + Emp(UTM^\alpha, UG^j, \{\Gamma_l\}^2, \mathcal{C}) \end{aligned}$$

and yet, for a value of β different from α:

$$\begin{aligned} &|UG^i|_{UTM^\beta} + Emp(UTM^\beta, UG^i, \{\Gamma_l\}^1, \mathcal{C}) \ > \\ &|UG^j|_{UTM^\beta} + Emp(UTM^\beta, UG^j, \{\Gamma_l\}^2, \mathcal{C}) \end{aligned}$$

This is because each group has a vested interest in developing a UTM which makes their Universal Grammar extremely small. This is just a twist, just a variant, on the problem described in the Universal Grammar is Free fallacy that I discussed above.

Proposed solution:

Which Turing machine? The one whose cost of emulation is the least.

The general problem

We have a set of N approved UTMs. A UTM is, by definition, a machine which can be programmed to emulate any Turing machine.

An emulator that makes $machine_i$ into a simulator of $machine_j$: $\{i \Rightarrow j\}$. Its length is $[i \Rightarrow j]$.

When a group wants to introduce a new UTM UTM_{N+1} to the group, they must produce emulators $\{i \Rightarrow N+1\}$ (for all $i \leq N$) and $\{N+1 \Rightarrow i\}$ (for all $i \leq N$).

What do we need to do to pick the best UTM? What we want to avoid is the case of a tricky linguist who develops a UTM (let's call it UTM^*) which makes his or her UG unreasonably short, but that property should be revealed by the way in which all the emulators for UTM^* are all long. That is, the rational choice is to pick the UTM_i which satisfies the condition:

$$\arg\min_{\alpha} \sum_{i} [i \Rightarrow \alpha] \tag{3}$$

A UTM^{α} which is unfair is one which cannot be easily reproduced by other UTMs

Real Universal Grammar

Linguists believe that children acquire language through the application of innate, universal principles of grammar. This also explains why most extraterrestrials speak colloquial English that is readily comprehensible to American audiences.