# Natural Language for Communication

*Michael Rose*

*March 30, 2019*

*In which we see how humans communicate with one another in natural language, and how computer agents might join in on the conversation.*

This chapter looks at language models for communication. We start with grammatical models of the phrase structure of sentences, add semantics to the model, and then apply it to machine translation and speech recognition.

## Phrase Structure Grammars

A big issue for n-gram models of language is **data sparsity**. For example, with $10^5$ words, we get $10^{15}$ trigram probabilities to estimate. We can adress the problem of sparsity through generalization.

### Generative Capacity

Grammatical formalisms can be classified by their **generative capacity**; the set of languages they can represent. Chomsky decscribes four classes of grammatical formalisms that differ only in the form of the rewrite rules. Here is the hierarchy, most powerful class first:

- **Recursively enumerable** grammars use unrestricted rules. Both sides can use any number of terminal and nonterminal symbols (for example $ABC \rightarrow DE$). These grammars are equivalent to Turing machines in their expressive power.
- **Context Sensitive Grammars** - These are restricted only in that the right side must contain at least as many symbols as the left hand side. For example, $ABC \rightarrow AYB$, or $a^n b^n c^n$.
- **Context Free Grammars** - The left hand consists of a single nonterminal symbol. Thus each rule licenses rewriting the nonterminal as the right hand side in any context.
- **Regular Grammars** - The most restricted class. Every rule has a single nonterminal on the left hand side and a terminal symbol optionally followed by a nonterminal on the right hand side. Regular grammars are equivalent in power to finite state machines.

The grammars higher in the hierarchy have more expressive power, but the algorithms for dealing with them are less efficient.

A popular model of phrase structure is the **probabilistic context free grammar**, or PCFG. A **grammar** is a collection of rules that defines a language as a set of allowable strings of words.

Here is a PFCG rule:

$$\begin{aligned} VP &\rightarrow Verb \quad [0.70] \\ &\mid \quad VP \quad NP \quad [0.30] \end{aligned}$$

Here VP (verb phrase) and NP (noun phrase) are **nonterminal symbols**.