Welcome to CS395C!

The field of Natural Language Processing is one of the oldest in computer science; as soon as they invented computers they wanted to use them to translate between Russian and English.

An apocryphal story from/about that era well illustrates why NLP is hard. Supposedly someone fed a sentence into an English-Russian translation system, then fed the result of that into a Russian-English system. The input was “The spirit is willing, but the flesh is weak.” The output was “The vodka is good, but the meat is rotten.” So it goes.

NLP is hard because to get it well and truly right, you’d need full world knowledge. We say (informally) that it’s “AI-complete”—the NLP problem won’t be solved until we have true AI, and vice versa. However, we can make significant inroads on a lot of subproblems.

There are two main styles of NLP, reflecting philosophies of linguistics. Rationalists believe that language is so complex it must be largely hard-wired into our brains, so that learning language as infants is just a matter of configuring parameters. Empiricists believe that infants are just really good at pattern recognition and extrapolation from minimal data—there is an inborn faculty for language, but no actual grammar or anything.

The corresponding schools of NLP, then, believe that more or less linguistic information needs to go into the system from the start. Rationalists will find an expert linguist to write out a grammar of, say, English, and then type that in to the computer. Effort is spent on finding better experts, being able to handle larger amounts of expert input, and in facilitating the usage of that information. Empiricists, on the other hand, will set up a very basic system and then feed boatloads of actual text in, gathering statistics as they go. They work with corpora (sg. ‘corpus’—from the Latin for ‘body’, it refers to a body of data used to train empirical systems).

More on why NLP is tricky:
Time flies like an arrow.
Fruit flies like a banana.

Without knowing the definitions of the words and some properties about the world, it’s difficult to say why the sentence structure of the second sentence
here is so different from the first.

Another example:
They won the game in zingzonk.
If you don’t know a word (‘zingzonk’), you don’t even know what sort of question this sentence would answer—it could be a place (like ‘Chicago’), or a time (like ‘April’), or even an adverbial of manner (like ‘overtime’).

Language is a moving target. Even with the best text corpus from 1989, for instance, we’re not going to see words like ‘internet’, or ‘DVD’, or ‘blingbling’.

**Zipf’s Law**

An example of a neat observation that has come out of NLP. Fairly robustly, within a given text the frequency (total number of occurrences) times the rank (e.g. 1 for the most common) of a given word ends up being a constant.

According to this observation, a word with twice the rank of another should be half as common: the 20th most common word should only occur half as often as the 10th most common. This turns out to be true more often than not.