Computation, Information, and Intelligence (COMS/ENGRI/INFO/COGST 172), Fall 2005
10/31/05: **Lecture 28 aid — Introduction to machine translation**

Agenda: Finish smoothing; begin discussion of machine translation.

Announcements: Ray Doyle's usual Monday 8pm office hours will *not* be changing.

Follow-up: Clarification/reminder: we are generally using the term "grammar" not to refer to the *prescriptivist* set of rules you learn in school, but to the (mostly) unconscious knowledge of language structure that native speakers of the language in question pretty much all exhibit, whether or not they have had any formal education.

I. The poverty of the stimulus, again The basic argument is that infants and young children do not hear enough samples of language to infer *from those samples alone* the complex rules of (natural) grammar that they master by a quite early age. Thus, the argument (roughly) goes, a substantial amount of the knowledge of language must be inborn.

After all, how else can we distinguish the "rightness" of two sentences, neither of which have ever be encountered in the history of the language?

II. The sparse data problem Computers don't have access to enough data, either. A standard dataset used in NLP has *95%* of the instances in the *test* data not occurring in the data one is allowed to learn from (Collins and Brooks, 1995).

Sometimes the situation is summed up as follows: "lack of evidence is not evidence of lack".

III. Jelinek-Mercer smoothing Set the probability of a rule $V_i \rightarrow w_i V_j$ (which, in our case, corresponds to the probability that if word w_i occurs then word w_j follows it) to

$$\lambda \frac{\#(w_i w_j)}{\#(w_i)} + (1-\lambda) \frac{\#(w_j)}{\sum_k \#(w_k)}$$

where λ is between 0 and 1 (usually non-inclusive).

IV. Bar-Hillel's (1960) example

1. "The pen is in the box."
2. "The box is in the pen."

V. Machine-translation paradigms Ordered by the depth of language analysis apparently required.

1. Direct replacement: word-for-word translation.
   But: "I am a fan [of this class]"

2. Syntactic transfer:
   Source-language utterance → source-language parse tree
   → target-language parse tree
   → target-language utterance
   But: "I like $\text{singing}_{\text{gerund}}$" vs. "Ich singe $\text{gern}_{\text{adverb}}$"

3. The interlingual approach:
   Source-language utterance → interlingua representation
   → target-language utterance
   But: What is "blue"?