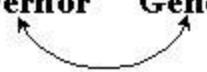
NLP: The Main Issues

- Why is NLP difficult?
  - many "words", many "phenomena" --> many "rules"
    - **OED: 400k words; Finnish lexicon (of forms): ~2 . 10^7**
    - **sentences, clauses, phrases, constituents, coordination, negation, imperatives/questions, inflections, parts of speech, pronunciation, topic/focus, and much more!**
  - irregularity (exceptions, exceptions to the exceptions, ...)
    - **potato -> potato es (tomato, hero,...); photo -> photo s,** and even: both **mango -> mango s** or **-> mango es**
    - Adjective / Noun order: **new book, electrical engineering, general regulations, flower garden, garden flower, ...:** but **Governor General**



Difficulties in NLP (cont.)

- ambiguity
  - **books: NOUN** or **VERB?**
    - you **need** many books *vs.* she **books** her flights online
  - **No left turn weekdays 4-6 pm / except transit vehicles** (*Charles Street at Cold Spring*)
    - when may transit vehicles turn: Always? Never?
  - **Thank you for not smoking, drinking, eating or playing radios without earphones.** *(MTA bus)*
    - Thank you for not eating without earphones??
    - or even: Thank you for ~~not~~ drinking without earphones!?
  - **My neighbor's hat was taken by wind. He tried to catch it.**
    - ...catch the wind *or* ...catch the hat ?

(Categorical) Rules or Statistics?

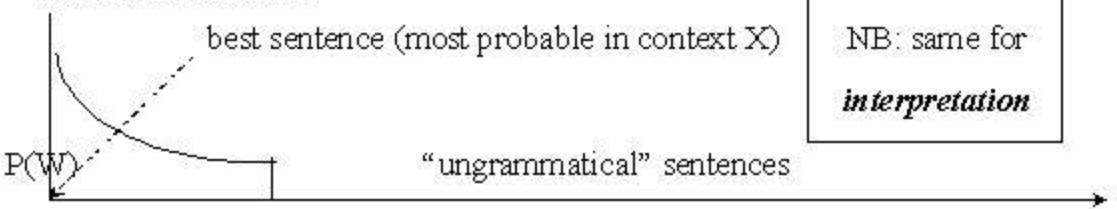
- Preferences:
  - clear cases: context clues: she books --> books is a verb
    - **rule: if an ambiguous word (verb/nonverb) is preceded by a matching personal pronoun -> word is a verb**
  - less clear cases: pronoun reference
    - **she/he/it refers to the most recent noun or pronoun (?) (but maybe we can specify exceptions)**
  - selectional:
    - **catching hat >> catching wind (but why not?)**
  - semantic:
    - **never thank for drinking in a bus! (but what about the earphones?)**

Solutions

- Don't guess if you know:
  - **morphology (inflections)**
  - **lexicons (lists of words)**
  - **unambiguous names**
  - **perhaps some (really) fixed phrases**
  - **syntactic rules?**
- Use statistics (based on real-world data) for preferences (only?)
  - **No doubt: but this is the big question!**

Statistical NLP

- Imagine:
  - Each sentence W = { w_1, w_2, ..., w_n } gets a probability P(W|X) in a context X (think of it in the intuitive sense for now)
  - For every possible context X, sort all the imaginable sentences W according to P(W|X):
  - Ideal situation:



NB: same for ***interpretation***

Real World Situation

- Unable to specify set of grammatical sentences today using fixed "categorical" rules (maybe never, cf. arguments in MS)
- Use statistical "model" based on **REAL WORLD DATA** and care about the best sentence only (disregarding the "grammaticality" issue)