UNIVERSITY OF
SAN FRANCISCO

Master of Science
in Analytics

Introduction

Natural Language Processing

What is NLP?

- The ability of a computer program to process human language
  - ***Computer program*** — early versions were FSAs/FSTs
  - ***Process*** — natural language understanding vs. natural language generation
  - ***Human language*** — artificial languages... (written) text... speech
- Fields of Study
  - Computer Science (1950s?) > Artificial Intelligence
  - Linguistics > Computational Linguistics

History (& Why You Should Care)

- Foundations (1940s - 1956)
  - Regular expressions & finite automata (Turing > Kleene)
  - Grammars (context-free grammars), formal languages (Chomsky, Backus, Naur)
  - Neurons, probabilistic models, markov models (McCulloh & Pitts)
  - Noisy channels (Shannon)
  - Speech Recognition
- NLP Splits (1957 - 1970)
  - Research in the *symbolic* camp (eg. Chomsky) concentrated on algorithms, formal AI
    - Success: Sentence parsing, programming languages
    - Critique: rules have exceptions, eg. English SVO violation in "*I thee wed*"
  - Research in the *stochastic* camp (eg.) ignited by including statistics, EE
    - Success: Federalist Papers (Mosteller & Wallace, 1964), speech recognition
    - Critique: p(*colourless green ideas sleep furiously*) > 0
    - Corpora and dictionaries
- Unification
  - Jelinek: *Every time I fire a linguist, the performance of the speech recognizer goes up*
  - Combining symbolic and stochastic approaches led to discourse models, etc.

@emilymbender

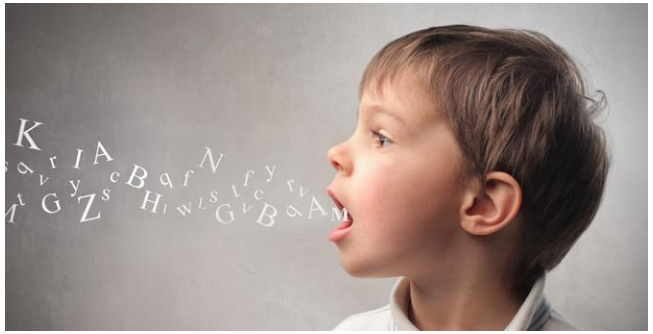
- Twitter reply to question, paraphrased: what should NLP people know?

  - Language has structure beyond linear order of words.
  - That structure is useful to leverage if we're interested in extracting meaning.
  - […]
  - The structure of any given language is fairly consistent across genres [...].
  - But languages vary in the structures that they use [...].
- Full thread: https://twitter.com/emilymbender/status/848607406925815808

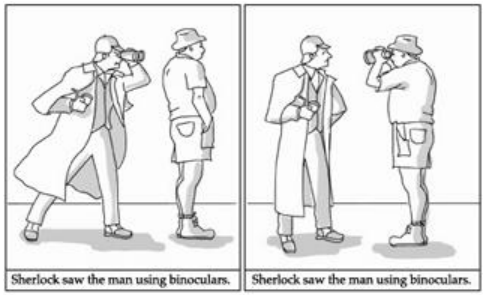
Structure of Language (NLP Areas)

Pragmatics

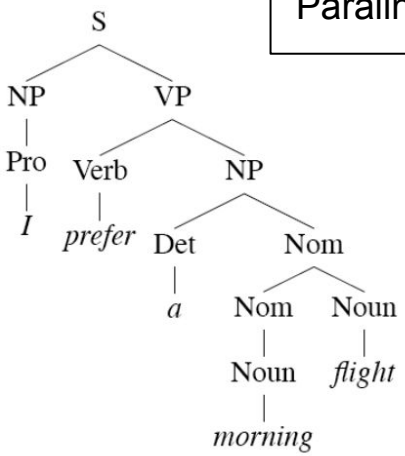
Paralinguistics



Phonetics, Phonology



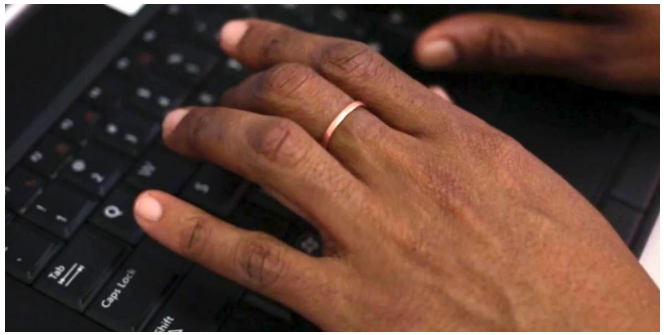
Semantics



Syntax

Words & Morphology

Keystroke Dynamics, Production



What is NLP (for Data Science)?

- Sentiment Analysis
  - “Voice of the customer”
  - How does a person (a group of people) feel about a thing
  - Used for marketing, customer service, etc.
- Topic Modelling
  - Given a (large) document, determine what topics are discussed
  - Topics may include identity / location of author*, success of a marketing campaign, etc.
- Information Retrieval / Extraction
  - Retrieve documents related to a query ala Google search engine
  - Retrieve data / documents related to a query, such as:

INPUT: *Today, San Francisco based Foo Inc. announced their acquisition of Bar Corp.*

TEMPLATE: acquisition_of(company$_1$, company$_2$)

RESULT: acquisition_of(Bar Corp, Foo Inc)

This Course

- What we'll study:
  - Human language (examples, etc.)
  - Linguistic and psychological theories
  - Algorithms
  - Applications (application / engineering)
- Topics
  - Words (Regular Expressions; Morphology; N-grams and LMs; Sequences & Viterbi)
  - Syntax (POS tagging, CFGs) and Applications (NLTK)
  - Semantics and Applications (Word clouds, word2vec)
  - Systems (Named Entity Recognition; Text Summarization; IE; Topic Models; Sentiment Analysis) and Applications (Mallet & gensim)
- Schedule — online as a Google Spreadsheet
- Grading
  - 50% = assignments
  - 50% = quizzes
  - 1 (group?) project in lieu of up to (1 assignment + 1 quiz)