Text Analytics

Lecture 1: Introduction

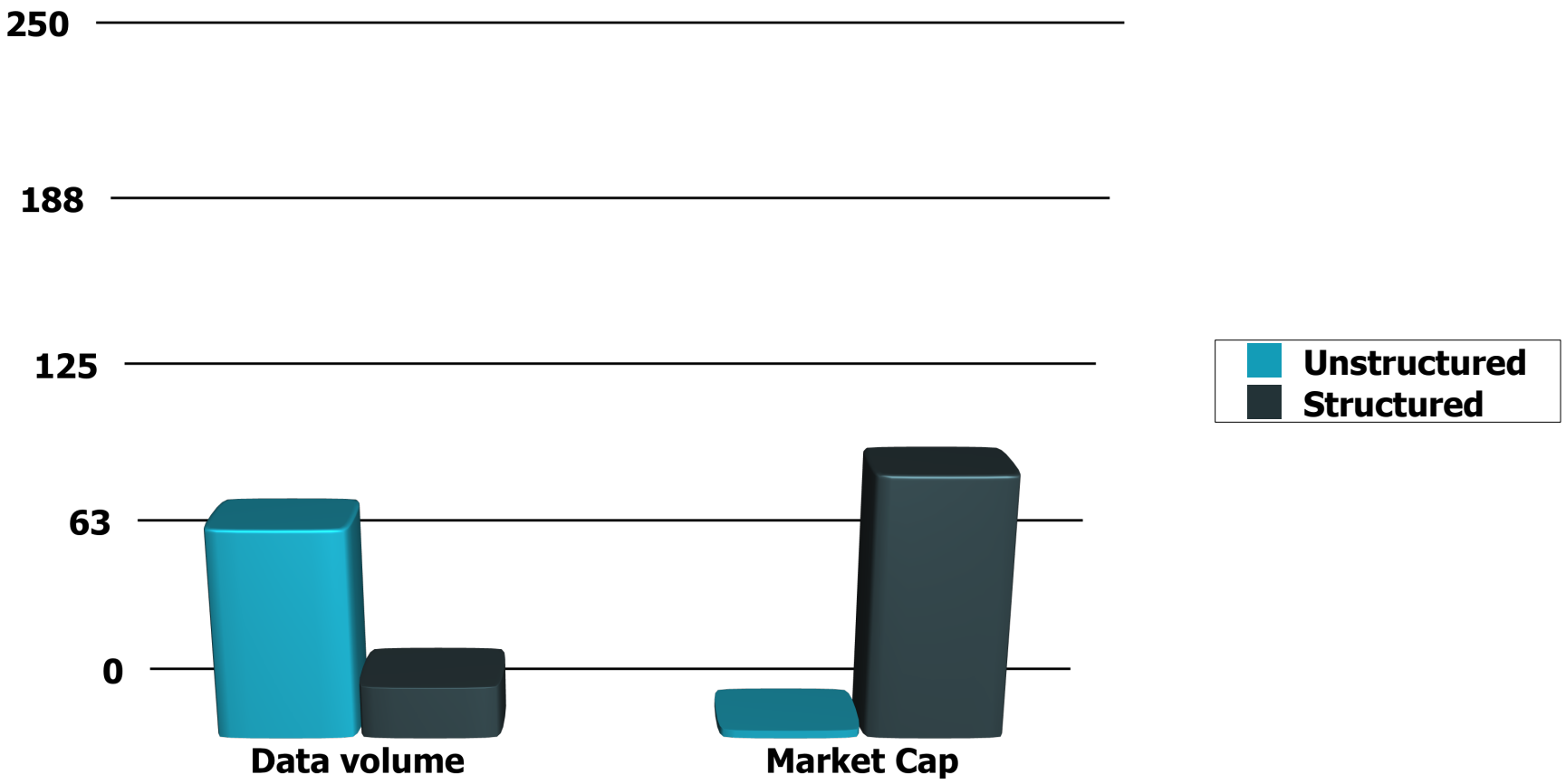
Klinton Bicknell

(borrowing from: Dan Klein, Roger Levy, Dan Jurafsky, and Jim Martin)

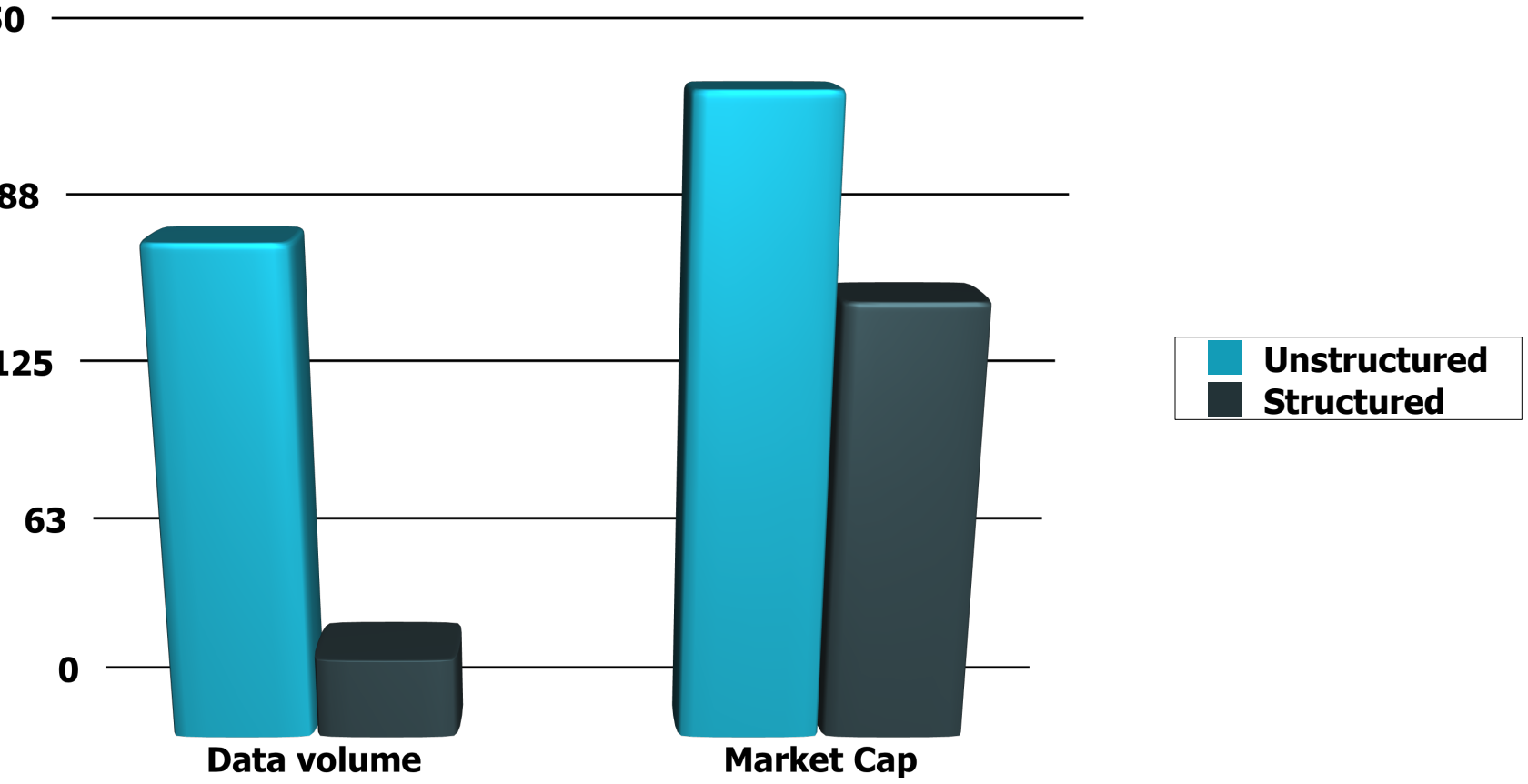
Text analytics

Text analytics is extracting useful information from unstructured language data

Unstructured (text) vs. structured (database) data in the mid-nineties



Unstructured (text) vs. structured (database) data today

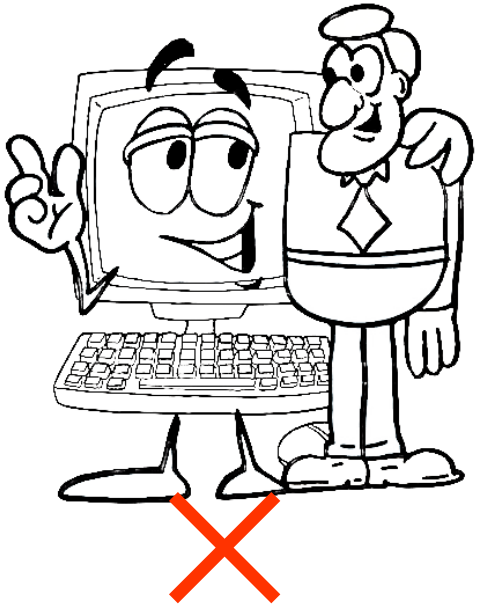


Text analytics

Why is this hard?

The mystery

- What's now impossible for computers (and any other species) to do is effortless for humans



The Dream

- It’d be great if machines could
  - "read" all the text and "listen" to all the speech
  - summarize them
  - answer questions about them
  - convert them to a database
- But they can’t (yet!):
  - Language is complex, ambiguous, flexible, and subtle
  - natural language processing (NLP) / computational linguistics (CL) are the fields that try to solve these problems

Instead…

- we can use what *does* already (mostly) work from NLP
  - what names (e.g., people, companies) are mentioned: **named entity recognition**
  - which words (e.g., nouns, pronouns) refer to the same thing: **coreference resolution**
  - which meanings (roughly) are words being used in: **part-of-speech tagging**
  - what is the structure (rough meaning) of sentences: **parsing**
  - what is the intended word: **spelling correction**
  - what is that in another language: **machine translation**
  - what was said: **automatic speech recognition**
  - …

Instead…

- … and creatively combine those results with machine learning
  - unsupervised
    - document clustering
    - frequency analysis
    - co-occurrence analysis
  - supervised
    - document categorization
    - sentiment analysis

This class

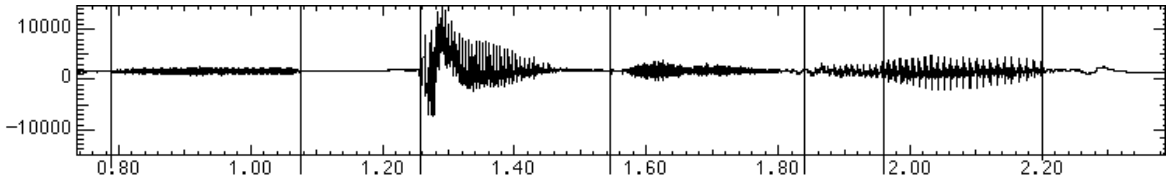
- practical knowledge of how to use state-of-the-art NLP and text analytics tools
  - esp. Stanford CoreNLP and Apache Lucene
- theoretical understanding of how these algorithms work
  - the state-of-the-art is far from perfect
  - prevents improper application
  - knowledge of limitations
  - knowledge of how to tweak parameters for desired results

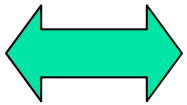
Text analytics

State-of-the-art

Speech Systems

- Automatic Speech Recognition (ASR)
  - Audio in, text out
  - SOTA: 0.3% for digit strings, 5% dictation, 50%+ TV





“Speech Lab”

Machine Translation

Atlanta, preso il killer del palazzo di Giustizia

ATLANTA - La grande paura che per 26 ore ha attanagliato Atlanta è finita: Brian Nichols, l'uomo che aveva ucciso tre persone a palazzo di Giustizia e che ha poi ucciso un agente di dogana, s'è consegnato alla polizia, dopo avere cercato rifugio nell'alloggio di una donna in un complesso d'appartamenti alla periferia della città. Per tutto il giorno, il centro della città, sede della Coca Cola e dei Giochi 1996, cuore di una popolosa area metropolitana, era rimasto paralizzato.

Atlanta, taken the killer of the palace of Justice

ATLANTA - The great fear that for 26 hours has gripped Atlanta is ended: Brian Nichols, the man who had killed three persons to palace of Justice and that a customs agent has then killed, s' is delivered to the police, after to have tried shelter in the lodging of one woman in a complex of apartments to the periphery of the city. For all the day, the center of the city, center of the Coke Strains and of Giochi 1996, heart of one popolosa metropolitan area, was remained paralyzed.

- Translation systems encode:
  - Something about fluent language
  - Something about how two languages correspond
- SOTA: for easy language pairs, better than nothing, but more an understanding aid than a replacement for human translators

Information Extraction

- Information Extraction (IE)
  - Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

  - SOTA: perhaps 70% accuracy for multi-sentence temples, 90%+ for single easy fields

Ambiguity

- computational linguists are obsessed with ambiguity
- ambiguity is a fundamental problem in CL
- resolving ambiguity is a crucial goal

Problem: Ambiguities

- Headlines:
  - Iraqi Head Seeks Arms
  - Ban on Nude Dancing on Governor's Desk
  - Juvenile Court to Try Shooting Defendant
  - Teacher Strikes Idle Kids
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
  - Hospitals Are Sued by 7 Foot Doctors

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

- I caused her to quickly lower her head or body
  - **Lexical category**: "duck" can be a N or V
- I cooked waterfowl belonging to her.
  - **Lexical category:** "her" can be a possessive ("of her") or dative ("for her") pronoun
- I made the (plaster) duck statue she owns
  - **Lexical Semantics:** "make" can mean "create" or "cook"

Ambiguity is Pervasive

- **Grammar**: Make can be:
  - **Transitive: (verb has a noun direct object)**
    - I cooked [waterfowl belonging to her]
  - **Ditransitive: (verb has 2 noun objects)**
    - I made [her] (into) [undifferentiated waterfowl]
  - **Action-transitive (verb has a direct object and another verb)**
  - I caused [her] [to move her body]

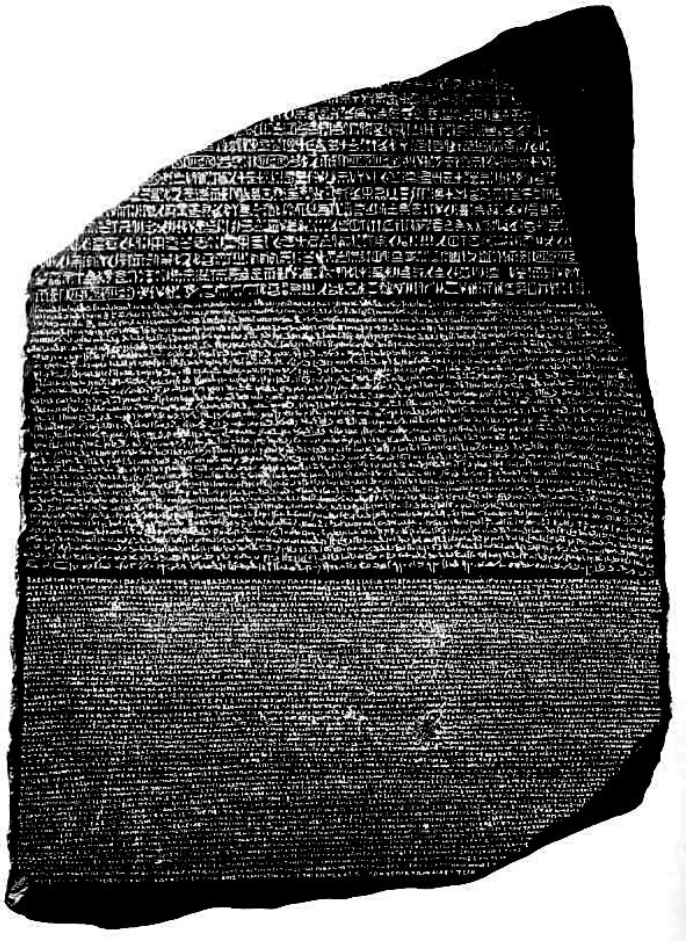
Ambiguity is Pervasive

- **Phonetics!**
  - I mate or duck
  - I'm eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I'm aid her duck
  - I mate her duck
  - I'm ate her duck
  - I'm ate or duck
  - I mate or duck

Ambiguity: no single right answer

- many interpretations could be correct
- but most interpretations are very unlikely
- goal: we want to assign probabilities to interpretations
- solution: what linguistic analysis did similar inputs receive before?
    - use corpora to *train* models
    - how do we formalize *similar*

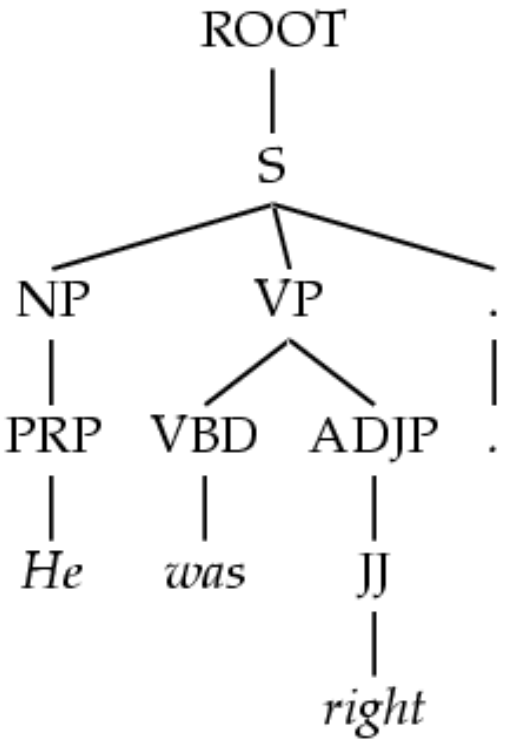
Corpora



- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what

Corpus-Based Methods

- A corpus like a treebank gives us three important tools:
  - It gives us broad coverage

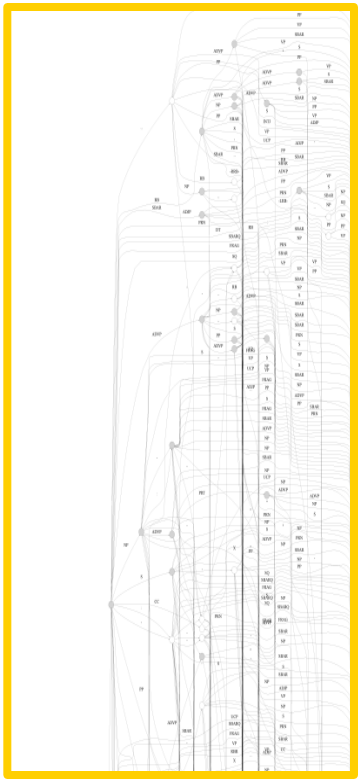


ROOT → S

S → NP VP .

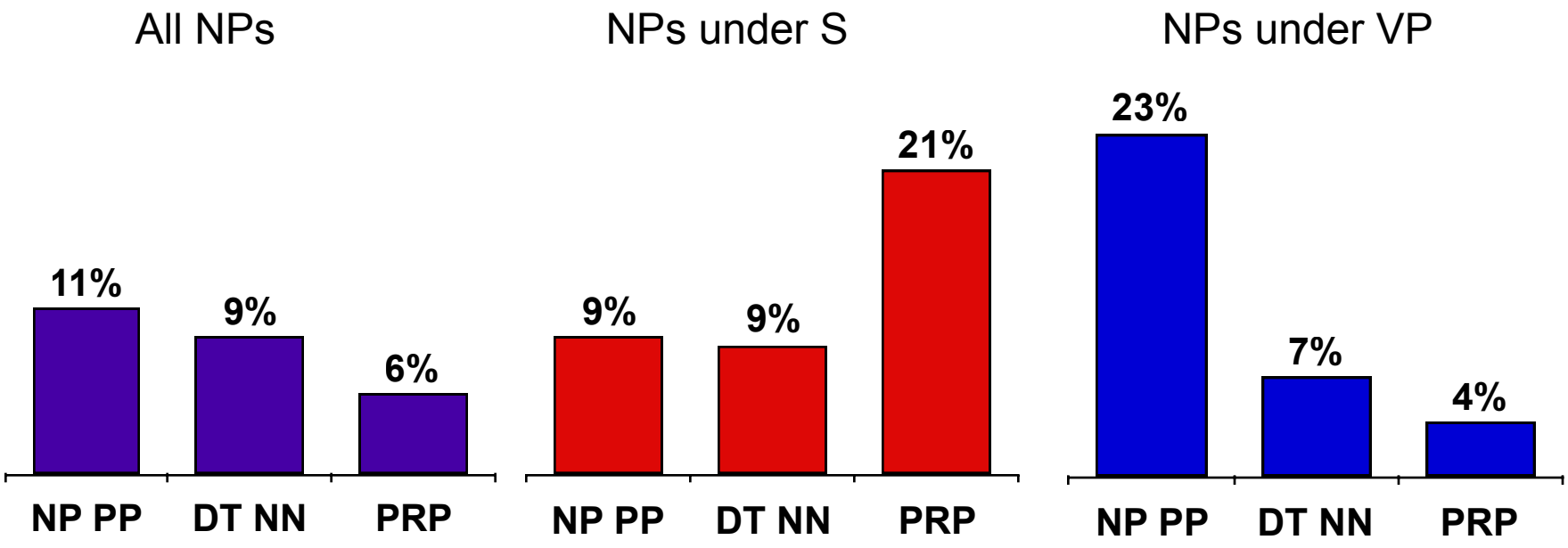
NP → PRP

VP → VBD ADJ



Corpus-Based Methods

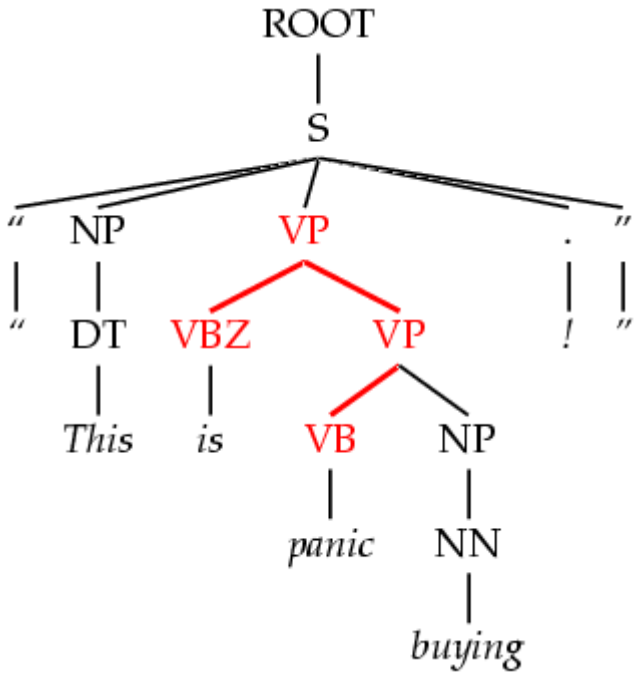
- It gives us statistical information
- "Subject-object asymmetry":

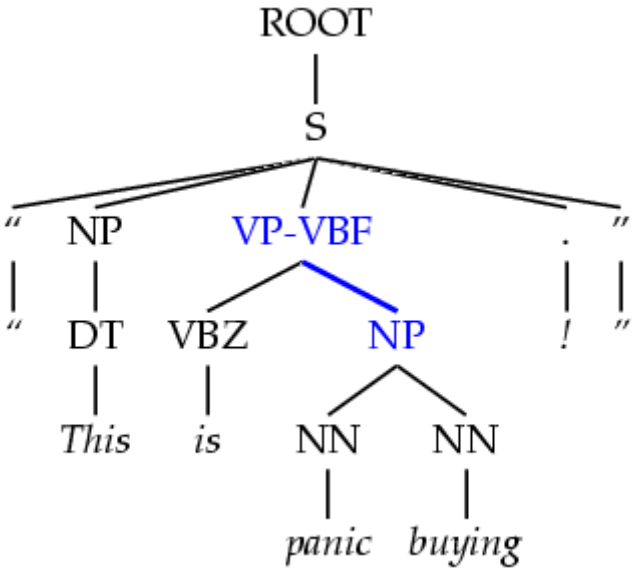


- *This is a very different kind of subject/object asymmetry than the traditional domain of interest for linguists*
- *However, there are connections to recent work with quantitative methods* (e.g., Bresnan, Dingare, Manning 2003)

Corpus-Based Methods

- It lets us check our answers!





Models and Algorithms

- By models we mean the formalisms that are used to capture the various kinds of linguistic knowledge we need.
- Algorithms are then used to manipulate the knowledge representations needed to tackle the task at hand.

Models

- State machines
- Rule-based approaches
- Logical formalisms
- Probabilistic models

Algorithms

- Many of the algorithms that we'll study will turn out to be transducers; algorithms that take one kind of structure as input and output another.
- Unfortunately, ambiguity makes this process difficult. This leads us to employ algorithms that are designed to handle ambiguity of various kinds

Paradigms

- In particular..
  - State-space search
    - To manage the problem of making choices during processing when we lack the information needed to make the right choice
  - Dynamic programming
    - To avoid having to redo work during the course of a state-space search
      - CKY, Earley, Minimum Edit Distance, Viterbi, Baum-Welch
  - Classifiers
    - Machine learning based classifiers that are trained to make decisions based on features extracted from the local context

State Space Search

- States represent pairings of partially processed inputs with partially constructed representations.
- Goals are inputs paired with completed representations that satisfy some criteria.
- As with most interesting problems the spaces are normally too large to exhaustively explore.
  - We need heuristics to guide the search
  - Criteria to trim the space

Dynamic Programming

- Don’t do the same work over and over.
- Avoid this by building and making use of solutions to sub-problems that must be invariant across all parts of the space.