

# Natural Language Processing

## Lecture 7: Parts of Speech

- My cat who lives dangerously no longer has nine lives.

- My cat who **lives** dangerously no longer has nine **lives**.

- My cat who **lives** dangerously no longer has nine **lives**.
- **lives**: noun /laɪvz/
- **lives**: verb /lɪvz/

- Mr. Black used to have a black beard but it is less black now than it used to be. He might black out if he realizes this fact.

- Mr. Black used to have a black beard but it is less black now than it used to be. He might black out if he realizes this fact.

# Part-of-Speech Tagging Task

- Input: a sequence of word tokens  $w$
- Output: a sequence of part-of-speech tags  $t$ , one per word

The linguistic facts are considerably more complicated than this, but there are good reasons for keeping it simple.

# Example

|         |       |          |   |           |   |
|---------|-------|----------|---|-----------|---|
| Charlie | Brown | received | a | valentine | . |
|         |       |          |   |           |   |
|         |       |          |   |           |   |

# Example

|             |             |          |            |           |             |
|-------------|-------------|----------|------------|-----------|-------------|
| Charlie     | Brown       | received | a          | valentine | .           |
| proper noun | proper noun | verb     | determiner | noun      | punctuation |
|             |             |          |            |           |             |

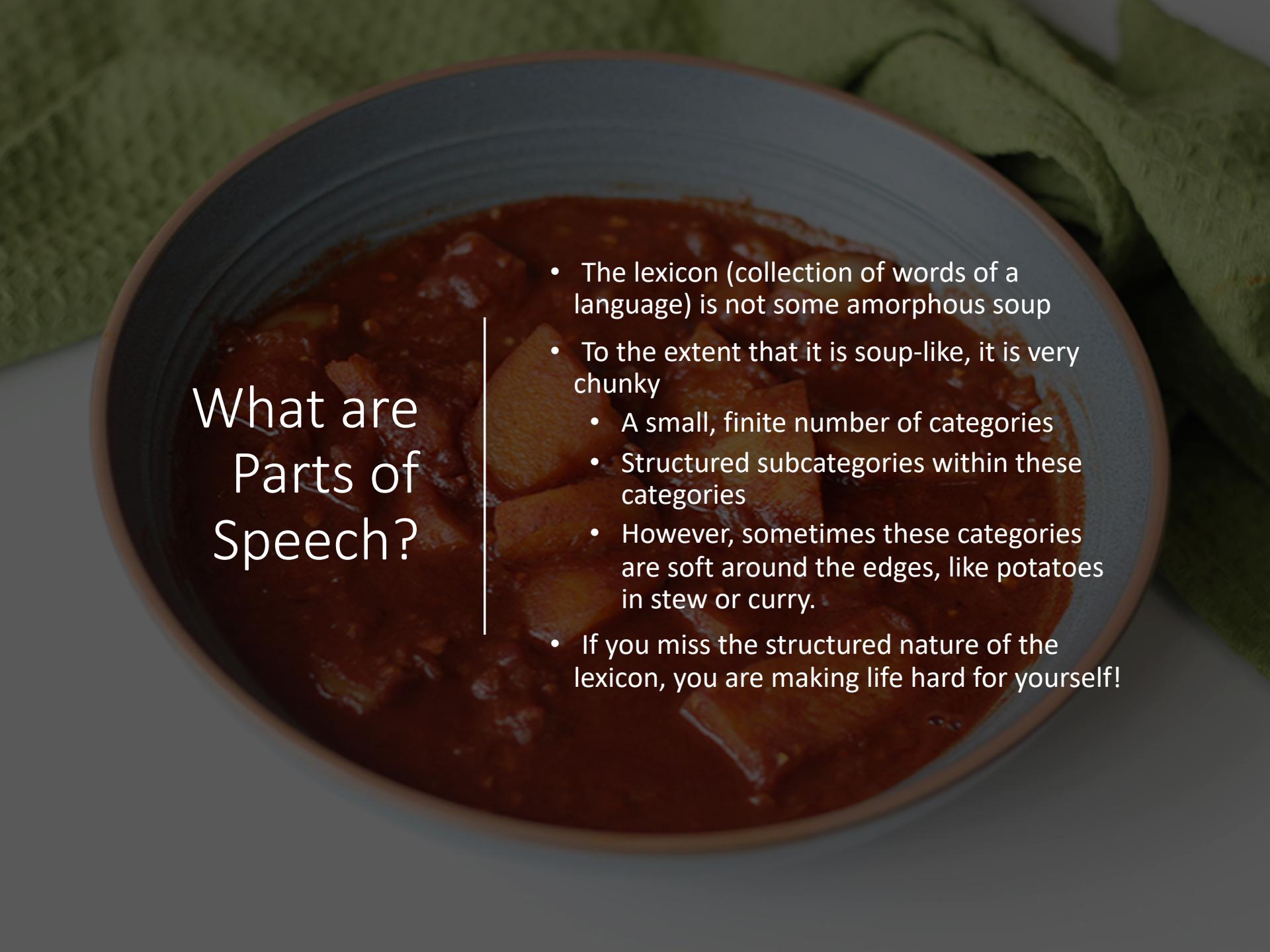
# Example

|   |  |                           |                         |                    |                                |
|---|--|---------------------------|-------------------------|--------------------|--------------------------------|
| Charlie                                     | Brown                                  | received                  | a                       | valentine          | .                              |
| proper noun                                 | proper noun                            | verb                      | determiner              | noun               | punctuation                    |
| name,<br>first name,<br>person<br>name, ... | name,<br>last name,<br>person<br>name, | past tense,<br>transitive | indefinite,<br>singular | singular,<br>count | end-of-<br>sentence,<br>period |



# Kaplan's Question

“So, you work on POS tagging!  
What’s a part of speech?”

A photograph of a bowl of soup with meat and vegetables, serving as the background for the slide. A vertical white line separates the title from the list.

# What are Parts of Speech?

- The lexicon (collection of words of a language) is not some amorphous soup
- To the extent that it is soup-like, it is very chunky
  - A small, finite number of categories
  - Structured subcategories within these categories
- However, sometimes these categories are soft around the edges, like potatoes in stew or curry.
- If you miss the structured nature of the lexicon, you are making life hard for yourself!



Q: What Do English  
Teachers Do?

A: Tell well-intentioned lies.

# What are Parts of Speech?

- A limited number of tags for word “class”
- **Distributional criteria**
  - Has the same contexts
  - Has the same syntactic functions (subject, object, modifier of nouns)
  - Occurs in the same positions in syntactic structure
- **Morphological criteria**
  - Allows the same suffixes, prefixes
- **Not about meaning**
  - We are suggesting that your English teacher lied to you
  - Get used to it

# Open-Class vs. Closed- Class

- Parts of speech can be divided into open-class and closed class
- **Open-class parts of speech**
  - Nouns
  - Verbs
  - Adjectives
  - and any class to which you can easily add new members
- **Closed-class parts of speech**
  - Pronouns
  - Determiners
  - Conjunctions
  - and any class to which you cannot easily add new members

# Some Open-Class Parts of Speech

# English Nouns

- Can be subjects and objects of verbs
  - *This book is about geography.*
  - *I read a good book.*
- Can be objects of prepositions
  - *I'm mad about books.*
- Can be plural or singular (*books, book*)
- Can have determiners (*the book*)
- Can be modified by adjectives (*blue book*)
- Can have possessors (*my book, John's book*)

# English Verbs

- Takes nouns phrases as arguments
  - At least a subject
    - *Dr. Mortensen parsed aggressively.*
  - Sometimes one or two objects
    - *Dr. Mortensen parsed the data.*
    - *Prof. Black passed [the function] [an argument].*
- Can take tense morphology (past/non-past)
- Can be modified by adverbs

# English Adjectives

- Modify nouns
  - his **pitiful code** (attributive)
  - His **code** is **pitiful**. (predicative)
- Can take comparative/superlative (-er/-est) suffixes when allowed by prosody
  - *big, bigger, biggest*
  - But *pitiful, more pitiful, most pitiful*
- Not all languages have adjectives—some languages (like Korean, Hmong, and Vietnamese) use verbs to modify nouns in this way

# English Adverbs

- Modify verbs, adjectives and other adverbs
  - He **erroneously** **concluded** that PHP is a real programming language **simply** because it is Turing complete.
  - He **concluded** **erroneously** that PHP is a real programming language.
  - The design of PHP is **exceptionally poor**.
  - My code **runs** **very slowly**.

# Some Closed-Class Parts of Speech

# English Prepositions

- Occur before noun phrases
- Relate noun phrase to some higher-level constituent
  - I scattered the data **from hell to breakfast**.
  - He lingered **in the depths of despair**.
- It is actually not difficult to characterize pronouns **formally**, but they are very difficult to characterize **semantically** (a good argument not to introduce semantic considerations into PoS categories)
- Also, they are often identical in spelling and pronunciation to **particles**

Talking  
Barbie  
says...

**MATH IS HARD**



**LET'S GO SHOPPING!**

memegene

NLP Barbie  
Says...

**PREPOSITIONS ARE HARD.**



**LET'S GO SHOPPING!**

memegenerator.net

# English Determiners

- Determiners are words that come at the beginning of noun phrases in English
- The most recognizable determiners are probably **articles** like *the*, *a*, and *an*
  - **The interpreter** choked on **an unknown identifier**.
- Other determiners include some demonstratives like *this* and *that*.
  - **That version of Python** really chaps my hide.

# English Pronouns

- Pronouns replace noun phrases, acting as a sort of shorthand for them
  - You code like a boy.
  - Your type system is not well-founded.
  - I'm not going to give the password to you.
  - Who knows Haskell, really?

# English Conjunctions

- Conjunctions join phrases, clauses, or sentences.
- Typically, the conjuncts joined by a conjunction are of the same time
- Coordinating conjunctions
  - *and, or, but...*
- Subordinating conjunctions
  - *if, because, though, while...*



# English Auxiliary Verbs

- “Helping verbs” that occur before main verbs
- Some occur as main verbs as well
  - *Be*
    - I **am** the type system. (**main verb**)
    - I **am working** on my project, you insensitive clod. (**aux. verb**)
  - *Have*
    - I **have** no qualms about criticizing your choice of languages. (**main verb**)
    - I **have written** a brilliant function that **will accomplish** just that! (**aux. verb**)
- Others (e.g. modals) occur only as auxiliary verbs
  - *would, will, could, can, might, must...*

# English Particles

- *Particle* is sometimes used as a grab-bag category for closed-class items that do not fit in another category
- Most often, in English, these resemble prepositions or adverbs and are used in combination with a verb
  - He **tore off** his shirt.
  - He **tore** his shirt **off**.

# Numerals

- Numerals have properties of both nouns and adjectives
  - They can be the subject and object of verbs:
    - Two will enter but only one will leave.
    - I bought twenty.
  - They can function both attributively and predicatively:
    - Two variables were undeclared.
    - We are three.
  - When then are used attributively, they come before any adjectives:
    - The two undeclared variables were the cause of much consternation.
    - \*The undeclared two variables were the cause of much consternation.

# Why have Parts of Speech

- There are too many words
  - You'd need a lot of data to train rules (but consider contemporary pretrained models)
  - Rules would be very specific
- PoS tags allow generalization of models
- Give useful reduction in model sizes
- There are many different tag sets
  - You want the right one for your task

# Broad POS categories

open classes

nouns

verbs

adjectives

adverbs

closed classes

prepositions

determiners

pronouns

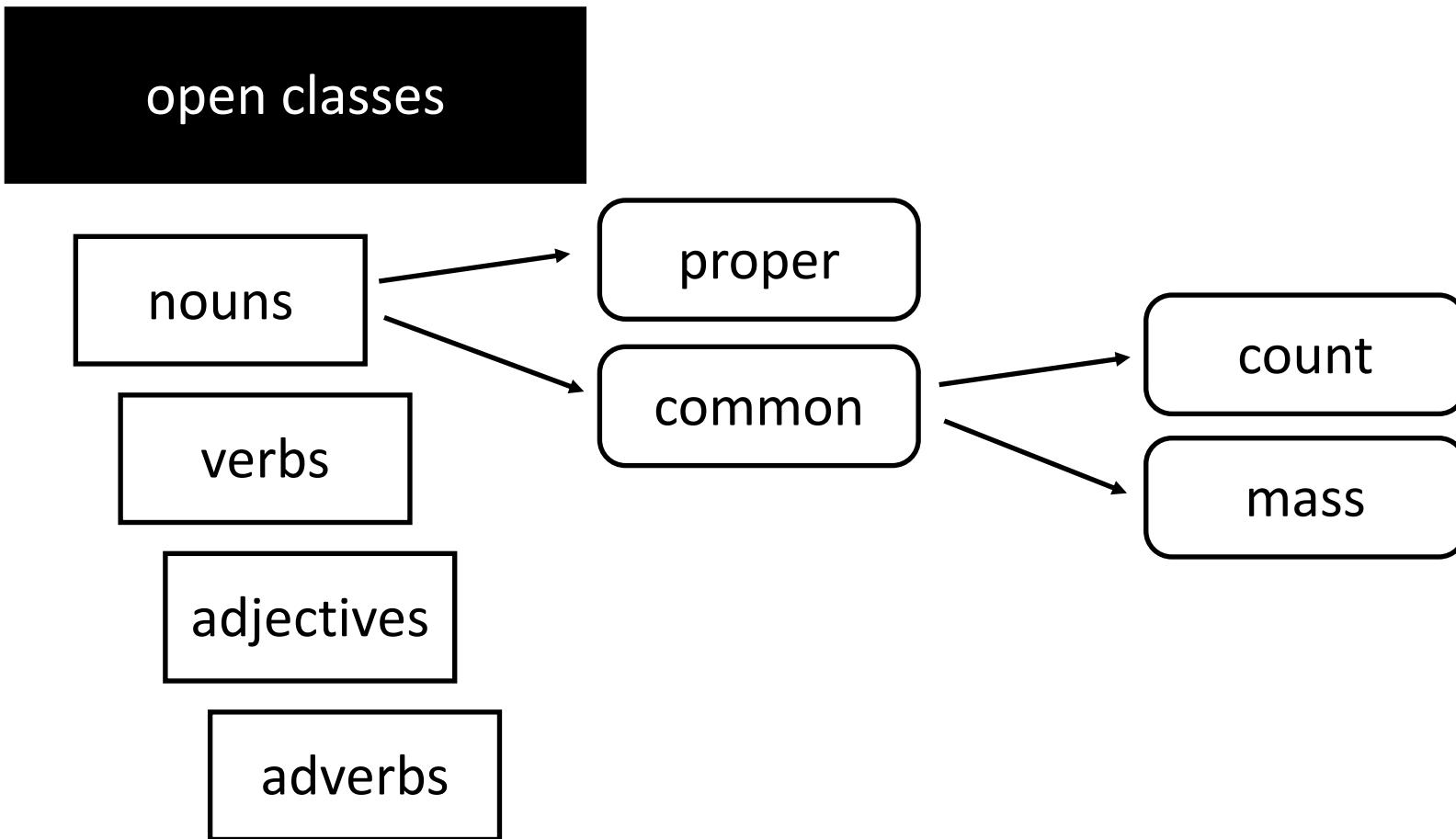
particles

numerals

conjunctions

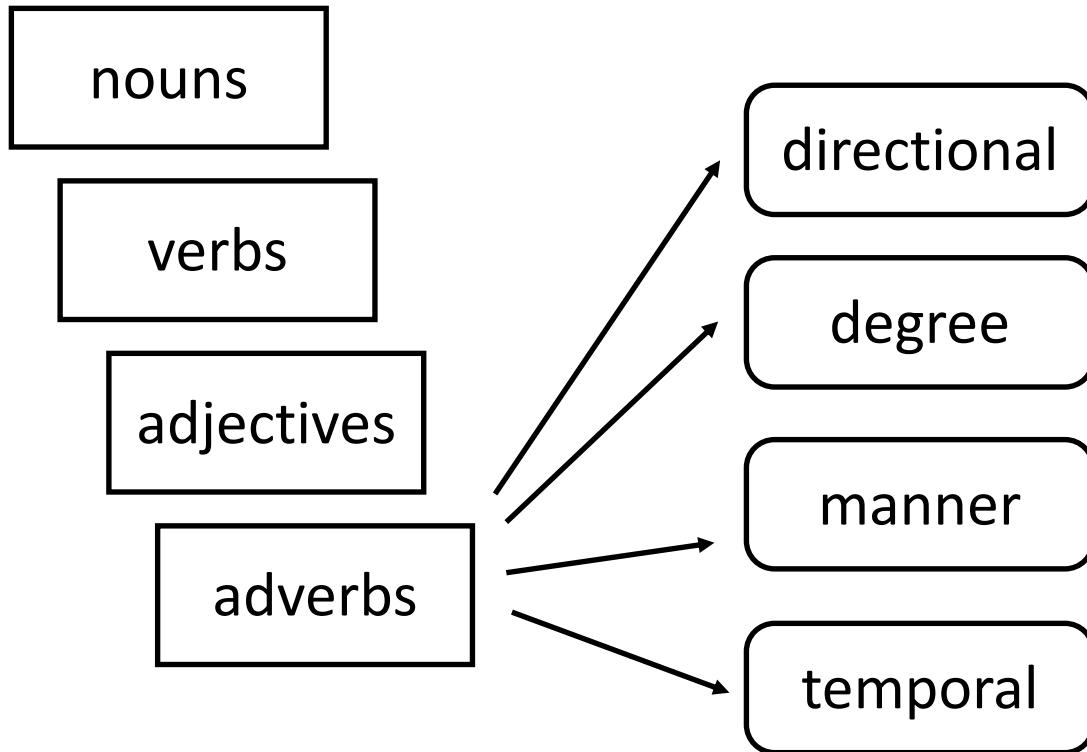
auxiliary verbs

# More Fine-Grained Classes



# More Fine-Grained Classes

open classes



# Hard Cases

- I will call up my friend
- I will call my friend up
- I will call my friend up in the treehouse
- Gerunds
  - I like walking.
  - I like apples.
  - His walking kept him fit.
  - His apples kept him fit.
  - His walking slowly kept him fit.
  - His apples slowly kept him fit.

But what do you want these for?

# Maybe?

- Interjections
- Negatives
- Politeness markers
- Greetings
- Existential *there*
- Numbers, Symbols, Money, ...
- Emoticon
- URL
- Hashtag

---

**ADJ:** adjective

---

**ADP:** adposition (preposition or postposition)

---

**ADV:** adverb

---

**AUX:** auxiliary

---

**CCONJ:** coordinating conjunction

---

**DET:** determiner

---

**INTJ:** interjection

---

**NOUN:** noun

---

**NUM:** numeral

---

**PART:** particle

---

**PRON:** pronoun

---

**PROPN:** proper noun

---

**PUNCT:** punctuation

---

**SCONJ:** subordinating conjunction

---

**SYM:** symbol

---

**VERB:** verb

---

**X:** other

# Some PTB Data (POS Tags)

IN In DT an NNP Oct. CD 19 NN review IN of `` `` DT The NN Misanthrope " " IN at  
NNP Chicago POS 's NNP Goodman NNP Theatre -LRB- -LRB- `` `` VBN Revitalized NNS  
Classics

VBP Take DT the NN Stage IN in NNP Windy NNP City , , " " NN Leisure CC & NNS  
Arts -RRB- -RRB- , , DT the NN role IN of NNP Celimene , , VBN played IN by NNP Kim NNP  
Cattrall , , VBD was RB mistakenly VBN attributed TO to NNP Christina NNP Haag . .

NNP Ms. NNP Haag VBZ plays NNP Elianti . .

NNP Rolls-Royce NNP Motor NNPS Cars NNP Inc. VBD said PRP it VBZ expects  
PRP\$ its NNP U.S. NNS sales TO to VB remain JJ steady IN at IN about CD 1,200 NNS cars IN  
in CD 1990 . .

DT The NN luxury NN auto NN maker JJ last NN year VBD sold CD 1,214 NNS cars  
IN in DT the NNP U.S.

# Why Tagging is Hard

- If every word by spelling (orthography) was a candidate for just one tag, PoS tagging would be trivial
  - How would you do it?
  - What problems do you foresee?
- As we've already seen, this won't always work
  - *lives* can be a noun or a verb
  - *black* can be a adjective, verb, proper noun, common noun, etc.
- But how bad is this problem, really?

# How bad is the ambiguity?

PoS tags per orthographic word in PTB (Penn Treebank)

|              |            |             |
|--------------|------------|-------------|
| 7 down       | 5 out      |             |
| 6 that       | 5 many     |             |
| 6 set        | 5 less     |             |
| 6 put        | 5 left     |             |
| 6 open       | 5 Japanese |             |
| 6 hurt       | 5 in       |             |
| 6 cut        | 5 hit      |             |
| 6 bet        | 5 half     | 317 down RB |
| 6 back       | 5 further  | 200 down RP |
| 5 vs.        | 5 forecast | 138 down IN |
| 5 the        | 5 fit      | 10 down JJ  |
| 5 spread     | 5 first    | 1 down VBP  |
| 5 split      | 5 East     | 1 down RBR  |
| 5 say        | 5 counter  | 1 down NN   |
| 5 's         | 5 cost     |             |
| 5 run        | 5 close    |             |
| 5 repurchase | 5 bid      |             |
| 5 read       | 5 beat     |             |
| 5 present    | 5 a        |             |

# “Down”

CD One CD hundred CC and CD ninety CD two JJ former NNS greats , , JJ near NNS greats , , RB hardly NNS knowns CC and NNS unknowns VBP begin DT a JJ 72-game , , JJ three-month NN season IN in NN spring-training NNS stadiums RB up CC and **RB down** NNP Florida

PRP He MD will VB keep DT the NN ball **RP down** , , VB move PRP it RB around IN As DT the NN judge VBD marched **IN down** DT the JJ center NN aisle IN in PRP\$ his VBG flowing JJ black NN robe , , PRP he VBD was VBN heralded IN by DT a NN trumpet NN fanfare

JJ Other NNP Senators VBP want TO to VB lower DT the **JJ down** NNS payments VBN required IN on JJ FHA-insured NNS loans

NNP Texas NNP Instruments , , WDT which VBD had VBN reported NNP Friday IN that JJ third-quarter NNS earnings VBD fell RBR more IN than CD 30 NN % IN from DT the JJ year-ago NN level , , VBD went **RBR down** CD 2 CD 1\8 TO to CD 33 IN on CD 1.1 CD million NNS shares

IN Because NNS hurricanes MD can VB change NN course RB rapidly , , DT the NN company VBZ sends NNS employees NN home CC and NNS shuts **VBP down** NNS operations IN in NNS stages : -- DT the RBR closer DT a NN storm VBZ gets , , DT the RBR more JJ complete DT the NN shutdown

NNP Jaguar POS 's JJ American NN depositary NNS receipts VBD were IN up CD 3\8 NN yesterday IN in DT a **NN down** NN market , , VBG closing IN at CD 10 CD 3\8

# “Japanese”

RB Meanwhile , , JJ Japanese NNS bankers VBD said PRP they VBD were RB still JJ hesitant IN about VBG accepting NNP Citicorp POS 's JJS latest NN proposal

CC And DT the NNPS Japanese VBP are JJ likely TO to VB keep RB close IN on NNP Conner POS 's NNS heels

DT The NN issue VBZ is RB further VBN complicated IN because IN although DT the NNS organizations VBP represent JJ Korean NNS residents , , DT those NNS residents VBD were RB largely VBN born CC and VBN raised IN in NNP Japan CC and JJ many VBP speak RB only NNP Japanese

CC And DT the NNP Japanese VBP make RB far JJR more NNS suggestions : -- CD 2,472 IN per CD 100 JJ eligible NNS employees CC vs. RB only CD 13 IN per CD 100 NNS employees IN in DT the (...)

DT The NNS Japanese VBP are IN in DT the JJ early NN stage RB right RB now , , VBD said NNP Thomas NNP Kenney , , DT a JJ onetime NN media NN adviser IN for NNP First NNP Boston NNP Corp. WP who VBD was RB recently VBN appointed NN president IN of NNP Reader POS 's NNP Digest NNP Association POS 's JJ new NNP Magazine NNP Publishing NNP Group

IN In CD 1991 , , DT the NNS Soviets MD will VB take DT a JJ Japanese NN journalist IN into NN space , , DT the JJ first NN Japanese TO to VB go IN into NN orbit

# How do we do this

- Pick the most frequent tag
  - Gives about 90% accuracy
- Look at the context
  - Preceeding (and succeeding) words
  - Preceeding (and succeeding) tags
  - The ...
  - To ...
  - John's blue ...
- We'll understand how we might look at local context better after we talk about **HMMs**

# The Out-of-Vocabulary Problem

- How do you handle cases where your dictionary does not include all of the words?
  - Proper names?
  - Borrowed words?
  - Neologisms?
- These are not generally a problem for you, as a language user
- How would you give a POS-tagger the same superpower?
- Stay tuned!

# Summary

- Here are a few important points:
  - Parts of speech are defined in terms of distribution and structure, not meaning (semantics)
  - Parts of speech may be open class or closed class
  - Parts of speech may differ across languages
  - PoS tags allow models to be more general
  - PoS tagging is non-trivial
    - Ambiguity—more than one PoS per orthographic word
    - Hard cases—more than one tag appropriate for a single sense of a word
    - Out of vocabulary (OOV) problem—your tagger will come across (open class) words it has never seen before