Linguistics & Corpora

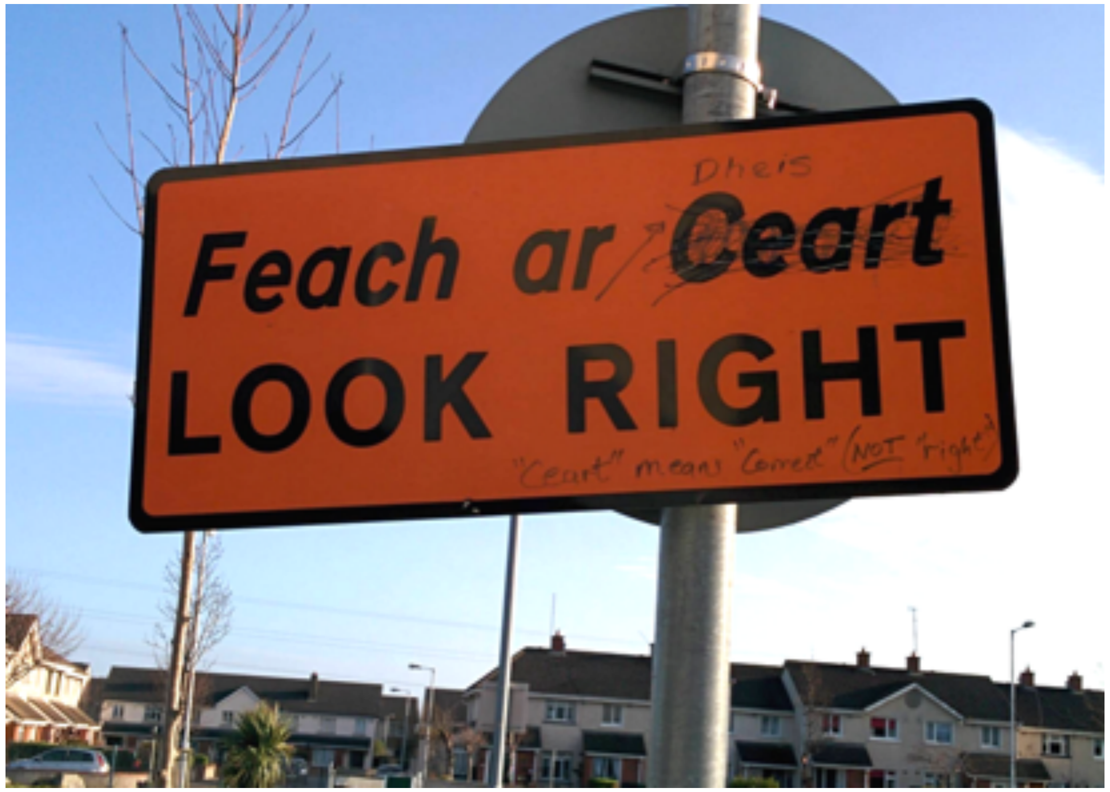
Wednesday, January 21, 2015

Plan for Today:

- Linguistics wrap-up
- Working with corpora

Help me remember:

- Fire drill at 10:40



http://languagelog.ldc.upenn.edu/nll/?p=17382

Last time

Sounds: Phonetics & Phonology

Letters: Orthography

Words: Morphology

Sentences: Syntax

Today

Finish our tour of linguistics

- Some complications in syntax
- Semantics

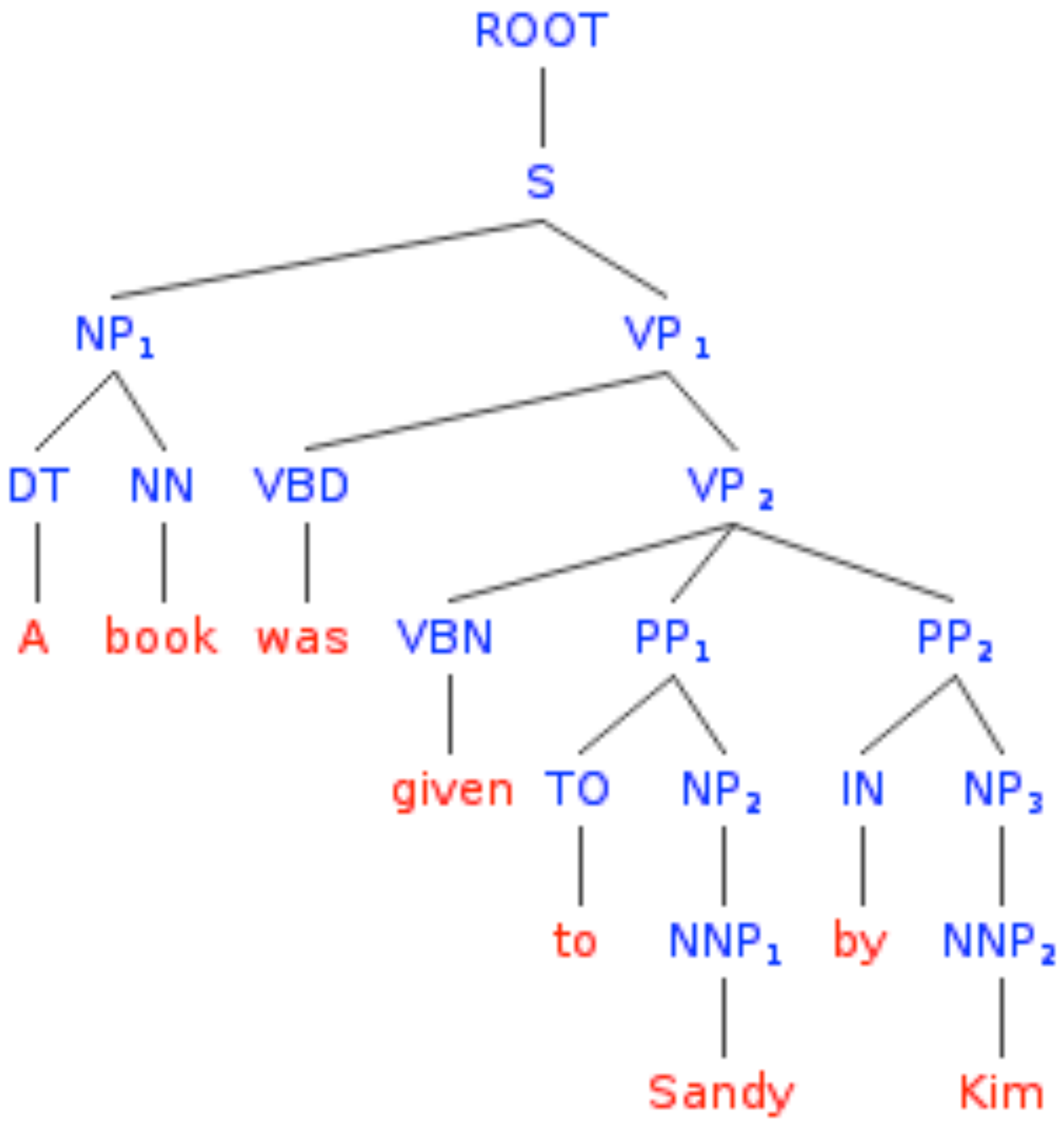
Working with Corpora

- General approach
- Python & NLTK

Syntax

Syntactic Structure

[ROOT [S [NP [DT A] [NN book]] [VP [VBD was] [VP [VBN given] [PP [TO to] [NP [NNP Sandy]]] [PP [IN by] [NP [NNP Kim]]]]]]]



Phrase Structure Rules

Context Free Grammars

- What assumptions are we making by having our rules be context free?
- When might that be a problem?

Dependencies

Language isn’t strictly context-free

- Differences between verbs
  I saw the cat.
  I put the cat on the table.
  * I put the cat.
  ? I saw.
- Subject-verb agreement
  I see the cat.
  He sees the cat.
  * I sees the cat.

Argument drop, aka null instantiation (Fillmore 1986)

- Definite null instantiation:

She promised.
I tried.
When did she leave?
They agreed.
She found out.
I forgot.

- Indefinite null instantiation:

I spent the afternoon baking.
We already ate.
What happened to my sandwich? *Fido ate.

Argument drop, aka null instantiation (Fillmore 1986)

- Lexically licensed: Possibility of an argument going missing depends on the lexical identity of the head (*eat* v. *devour*)

Fido ate.	*Fido devoured.
She promised.	*She pledged/vowed/guaranteed.
They accepted.	*They authorized.
She found out.	*She discovered.
He lost the race/his wallet.	He lost.

#98

- Systematic: Subjects (e.g., in Spanish) or any argument (e.g., Japanese) can be dropped, if supported by the discourse context

#99

Semantics

Introducing meaning…

Syntactic structure doesn’t directly give us meaning.

Syntactic grammaticality vs. semantic weirdness

- The dog barked.
- # The book barked.

Syntactic position vs. semantic role

- Kim saw Sandy.
- Sandy was seen by Kim.

Lexical semantics

Meaning of individual words

- How they combine
- How they relate to one another
  - Synonyms
  - Metonymy
    - “The White House issued a statement today…”
    - How is your Algs homework coming along?
  - Synechdoche: Part-to-whole
    - All hands on deck
    - He has many mouths to feed.

Language in 10 Presentations

Goal: Expose us all to different languages

- What's interesting about them?
- What should we know to process them automatically?

Expectations: 10 minutes, 3-6 slides

- Language facts (demographics, location)
- Linguistic characteristics (orthography, morphology, syntax)
- Computational efforts (resources, tools, papers)

Grading

- 20%: Adherence to time limit (within 2 minutes or so)
- 40%: Coverage of topics
- 40%: Polish of presentation

Working with corpora

How do we start to *use* all of this linguistics stuff?

When does linguistics help us?

How far can we get without it?

How do we get examples of language for our models?

What do we do with that data once we have it?

Class Activity: Alien Languages

arrat

bat

bichat

cat

dat

forat

gat

hilat

iat

iat lat pippat eneat hilat oloat at-yurp

NLTK

Python library for dealing with language corpora

Hides some of the messiness of dealing with text data, so that we can focus on the interesting algorithmic parts.

Python check-in

What is the difference between the following two lines? Which one will give a larger value?

Will this be the case for other texts?

```
>>> sorted(set(w.lower() for w in text1))
>>> sorted(w.lower() for w in set(text1))
```