Natural Language Processing

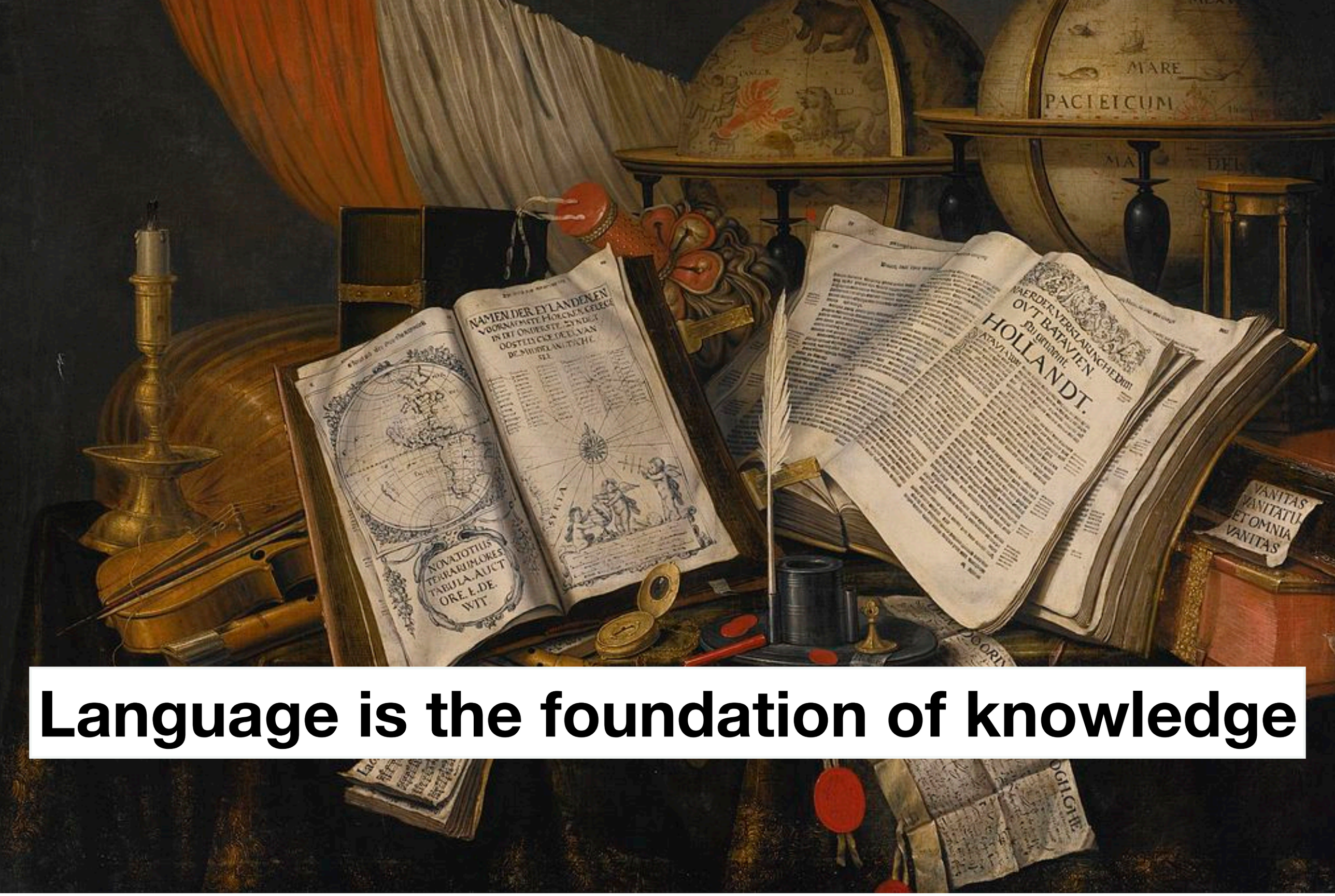
Kevin Duh
Johns Hopkins University

Fall 2019
https://kevinduh.github.io/nlp-course/

Language is what makes us human

PUBLICDOMAIN

“Cro-Magnon artists painting” by CHARLES R. KNIGHT
http://commons.wikimedia.org/wiki/File:Font-de-Gaume.jpg



Language is the foundation of knowledge

PUBLICDOMAIN

“Vanitas still life” by Edwaert Collier

https://commons.wikimedia.org/wiki/File:EDWAERT_COLLIER_VANITAS_STILL_LIFE.jpg

Natural Language Processing (NLP) studies how computers can interpret and manipulate text

This course will teach you the **fundamental** models & techniques for computer processing of human language

Today's Outline

1. **Survey of NLP Problems**

2. **Adminstration:**

    - Course expectations
    - Procedures (wait list, etc.)

3. **Two Major Themes**

    - Language has structure
    - Language processing involves ambiguity resolution

Self-Introductions

Instructor:
Kevin Duh

JOHNS HOPKINS UNIVERSITY

国立大学法人
奈良先端科学技術大学院大学
NARA INSTITUTE of SCIENCE and TECHNOLOGY

NTT

UNIVERSITY of WASHINGTON

RICE

Self-Introductions

Head TA:
Arya McCarthy

TA:
Suzanna Sia

CA: TBD, I'll introduce them next time

Today's Outline

1. **Survey of NLP Problems**
2. **Adminstration:**
    - Course expectations
    - Procedures (wait list, etc.)
3. **Two Major Themes**
    - Language has structure
    - Language processing involves ambiguity resolution

NLP research tackles a variety of problems

- **Applications**
  - Dialog Systems
  - Question Answering
  - Sentiment Analysis
  - Information Extraction
  - Machine Translation
- **Analysis of Linguistic Structure**
  - Word-level
  - Sentence-level
  - Document-level

Dialogue Systems

Bot (ELIZA): Is something troubling you?

User: Men are all alike.

Bot: What is the connection, do you suppose?

User: They are always bugging us about something or other.

Bot: Can you think of a specific example?

From: Güzeldere and Franchi. “dialogues with colorful personalities of early ai.” https://web.stanford.edu/group/SHR/4-2/text/dialogues.html

Question Answering

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail. Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud.

Question:
What causes precipitation to fall?

Answer: gravity

Question:
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

Answer:

From: Rajpurkar et. al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP2016. https://aclweb.org/anthology/D16-1264

Sentiment Analysis

User Review:

As a backpacking student Ryanair was really my only option (being the cheapest!). I think their bad reputation is built mostly on people who simply don't read the terms of their flight. It's a cheap A-B service that saves you a tonne of cash if you stick to their rules. I actually found the staff to be very friendly and helpful. No complaints there! What I will complain about is the food on offer. I understand plane food is bad in general but this is another level of bad. The food is disgusting overpriced unfortunately. Also, the seats are uncomfortable for people like me who are over 6 feet tall. It was difficult to relax at times. My flight arrived on time which was great. This meant I made my connecting train. Happy days! Overall I was very pleased with Ryanair. The price I paid was really cheap in comparison to the others on offer and as someone looking to save money, this was really all I was after – a cheap flight.

What is the user's opinion on…

Food: 

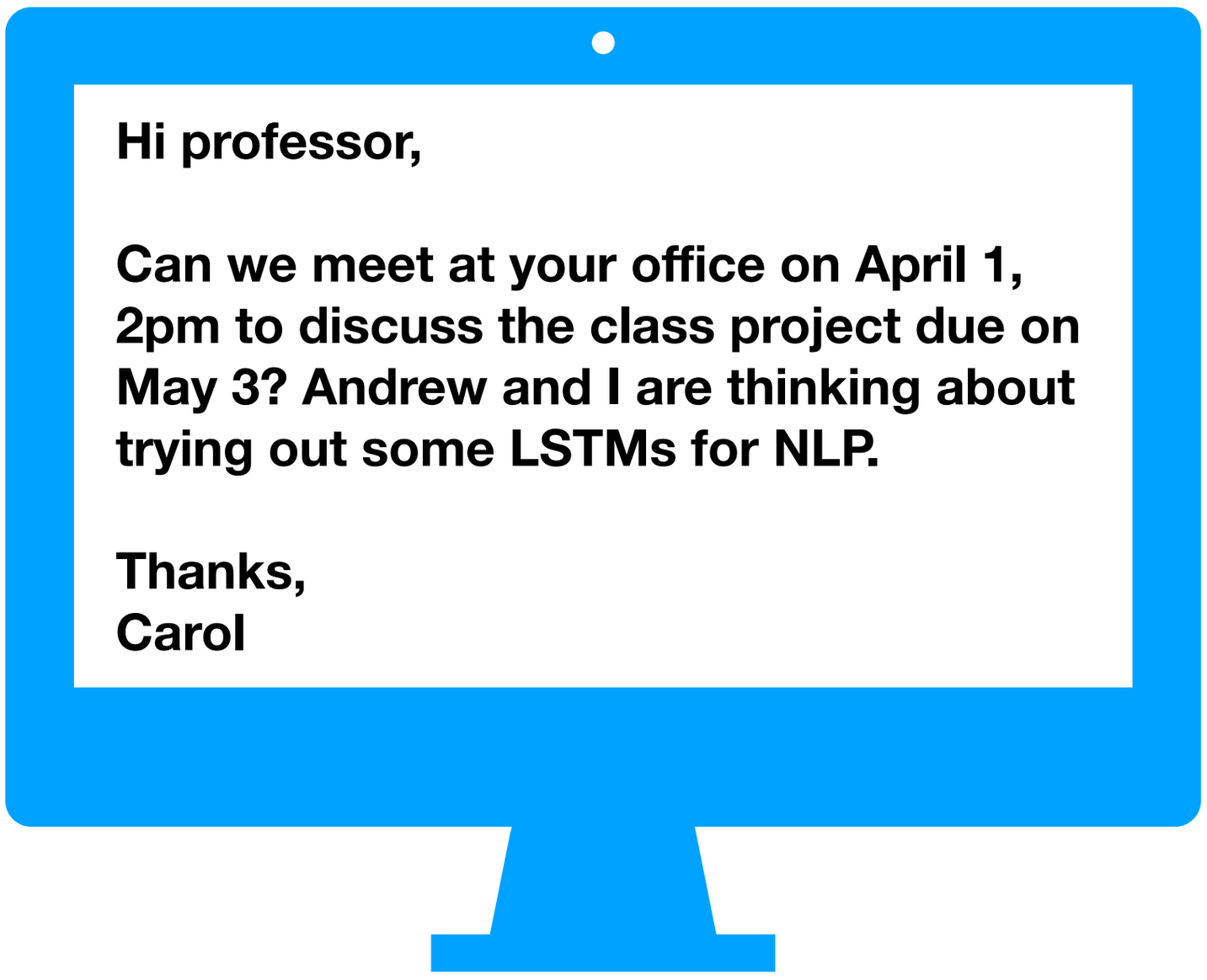
Staff: 

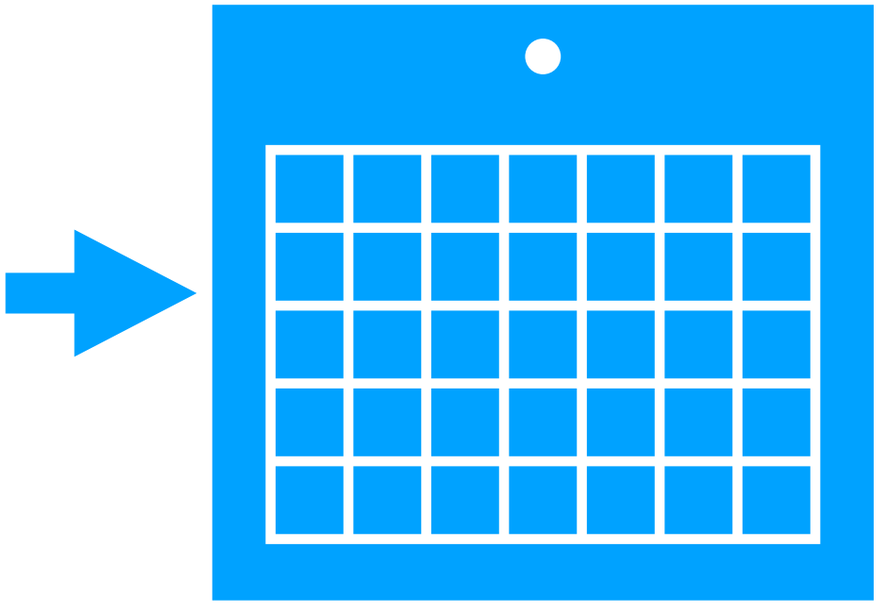
Punctuality:

Comfort:

From: https://aylien.com/text-api/aspect-based-sentiment-analysis/

Information Extraction

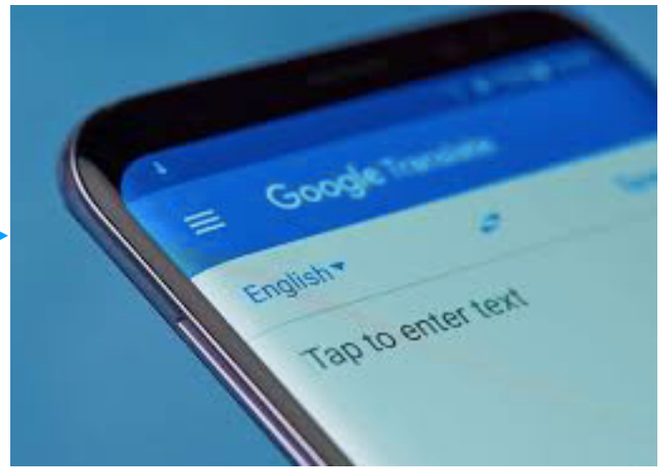




Subject: Discuss the class project
Time: April 1, 2pm
People: Carol and Andrew

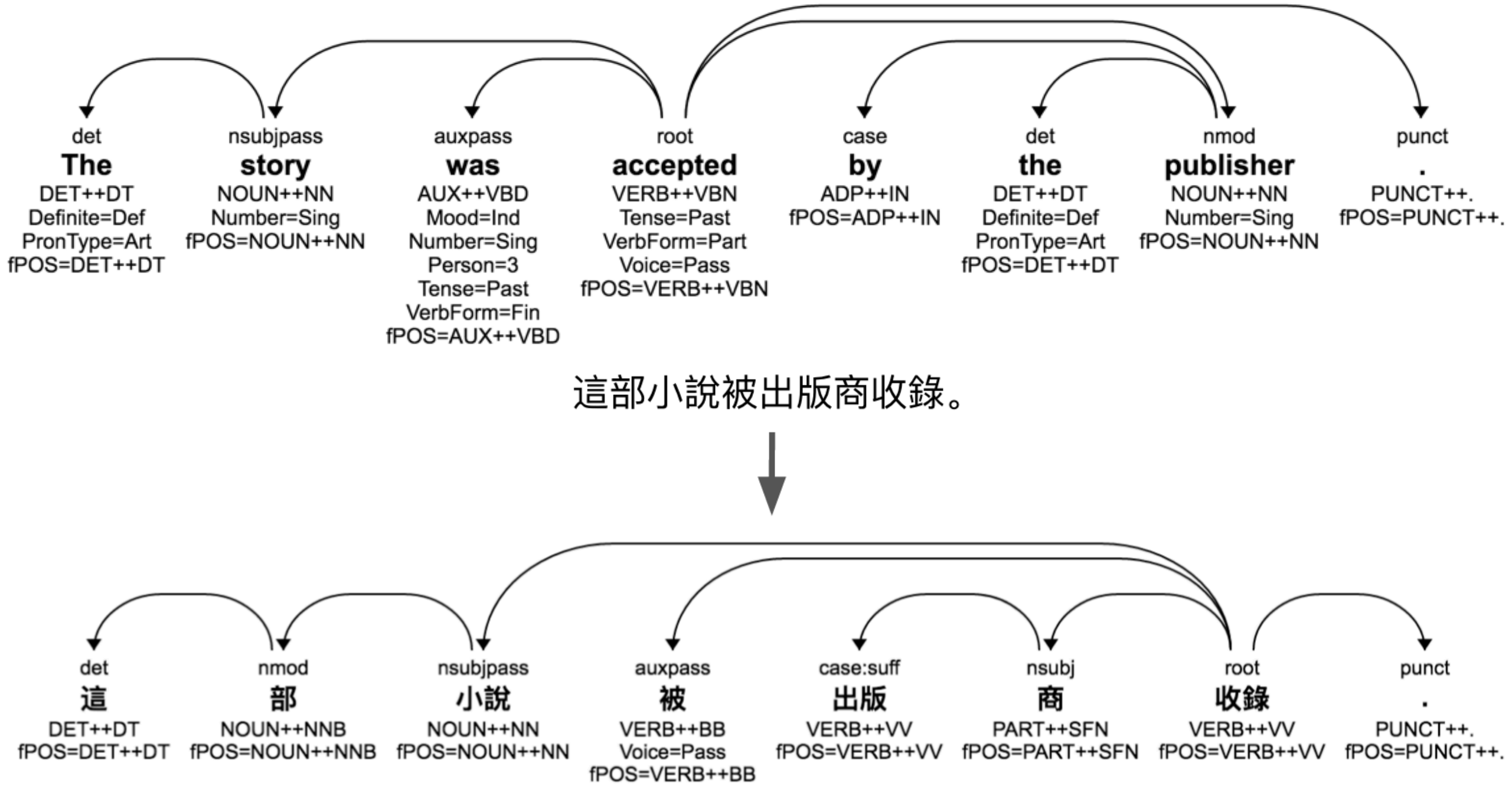
Machine Translation

Cogito, ergo sum



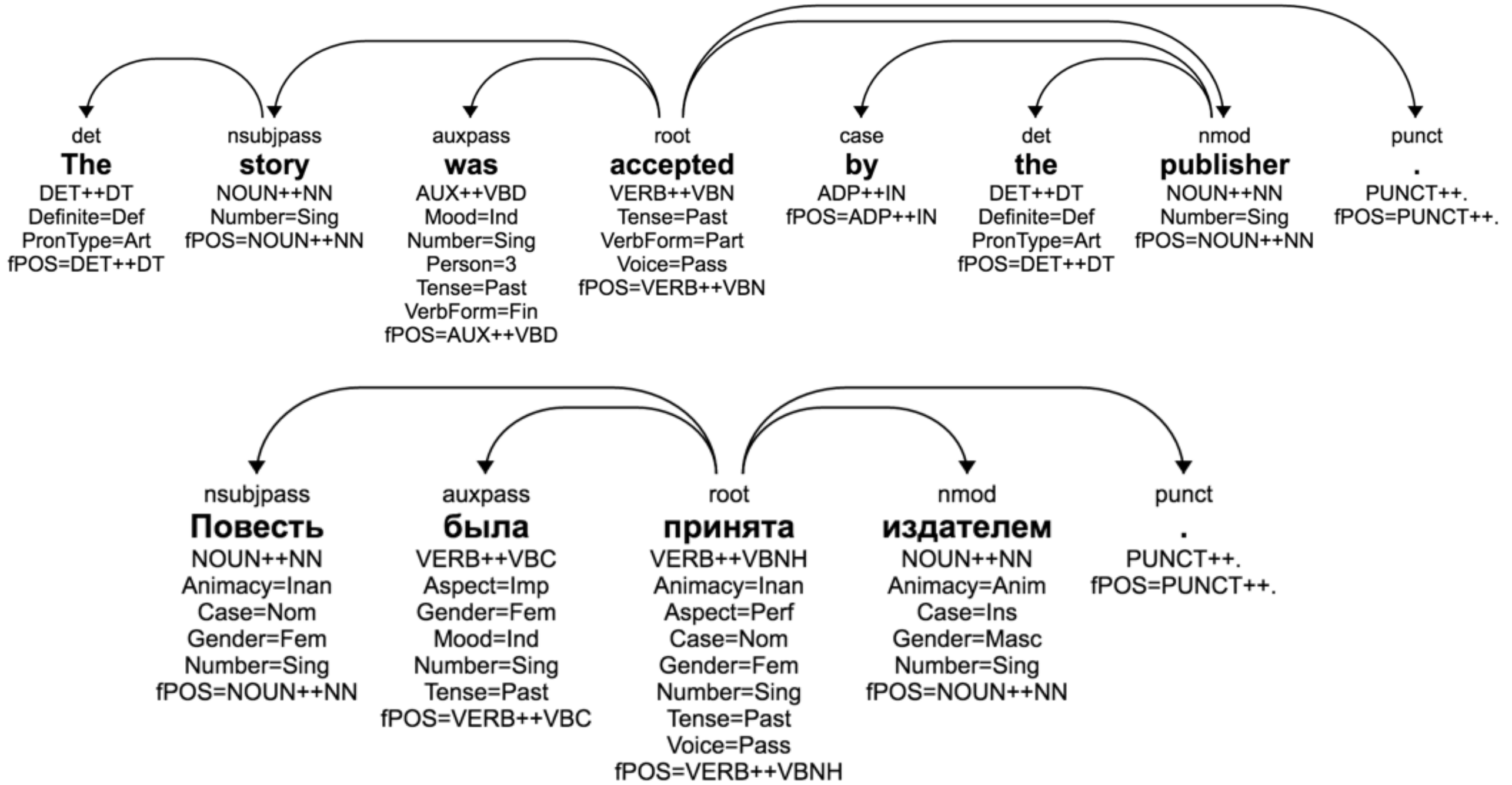
I think, therefore I am

Linguistic Analysis: Morphology & Syntax



From: https://ai.googleblog.com/2016/08/meet-parseys-cousins-syntax-for-40.html

Linguistic Analysis: Morphology & Syntax



From: https://ai.googleblog.com/2016/08/meet-parseys-cousins-syntax-for-40.html

Linguistic Analysis: Semantics

1. The story was accepted by the publisher

2. The publisher accepted the story

- What's going on: someone accepted something
- Who's someone: the publisher
- What's something: the story

Today's Outline

1. **Survey of NLP Problems**

2. **Adminstration:**

   - Course expectations

   - Procedures (wait list, etc.)

3. **Two Major Themes**

   - Language has structure

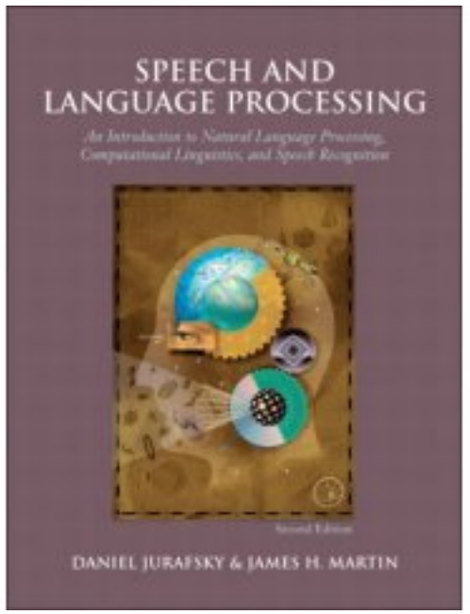
   - Language processing involves ambiguity resolution

Course organization

- Website: https://kevinduh.github.io/nlp-course/

- We'll use Piazza for discussions & questions, and Gradescope for homework hand-in.

- Textbook: Jurafsky & Martin, Speech and Language Processing, 2nd ed. (P98.J87 2009 in Science Ref section on C-Level)



- Class: Shaffer 301, MWF 11:00-11:50

- Office Hours: M 10-11, Th 18:45-20:00

Grading

- **6 Homework assignments: 60%** (10% each)
  - These are non-trivial, so start early and come to office hours/Piazza. If you can do these proficiently, then you have mastered the material
  - Late policy: You're allowed up to 10 late days throughout the term. Rather than ask me for an extension, just use a late day. After 10 days, we'll have to give you zeros.
    - You can work in teams, but do your own work and follow integrity code
- **Participation: 5%** (in-class and on Piazza)
- **Midterm exam: 15%** (October 16, in-class)
- **Final exam: 20%** (during Finals week, date TBD)
- ***My goal is you all master the material and pass with flying colors!***

Recipe for Success

- Sleep well and come to lecture ready to engage. Otherwise, take a nap at home and catch up later.

- Read textbook and slides before and/or after class. The human brain usually requires multiples passes to solidify knowledge.

- Ask questions. If you don’t understand something, you should ask. If you understand it, then you’ll naturally have follow-up questions.

- Befriend your neighbor. Work in teams and teach each other.

- Start your homework early. Learn good programming practices while you’re at it.

- Have fun!

Lecture Modules

1. Intro: Modeling grammar*
2. Language Models*
3. Text Classification*
4. Linguistics 101
5. Tree Parsing*
6. Neural Networks
7. Sequence Tagging*
8. Topic Models
9. Finite State Transducers*
10. Semantics
11. Structured Prediction

Plus five research talks!

**** module has corresponding homework assignment***

Related Courses

- This course is based on Prof. Jason Eisner's NLP course offered previously: http://www.cs.jhu.edu/~jason/465/

- Prerequisite: Data structures (601.226 or 600.226)

- Knowledge of linguistics, statistics, machine learning, and automata helps but is not assumed.

- This course can be taken together with: Intro to Human Language Technology (601.467/667), an omnibus course organized by Prof. Philipp Koehn that focuses on applications

If you’re on the wait list…

- Send me email (kevinduh @ cs. jhu.edu) by Sept. 5 with the following:
  - Title: Request to register for Natural Language Processing (601.465/665)
  - Body:
    - Name
    - Major (CS, EE, Cog Sci, etc.)
    - Year (Undergrad junior, Master’s 1 year, PhD 7th year, etc.)
    - One sentence: why you want to take the course
- Note the class is pretty full but I will try my best to accomodate

Questions so far?

Today's Outline

1. **Survey of NLP Problems**

2. **Adminstration:**

   - Course expectations
   - Procedures (wait list, etc.)

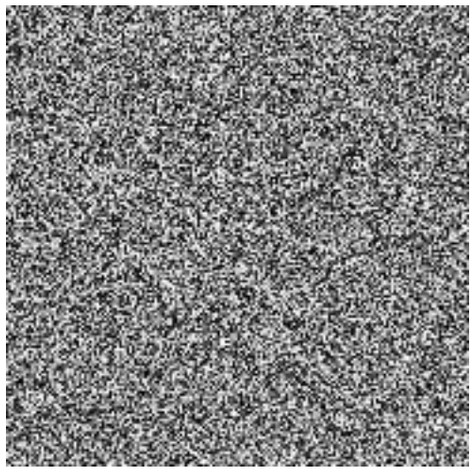
3. **Two Major Themes**

   - Language has structure
   - Language processing involves ambiguity resolution

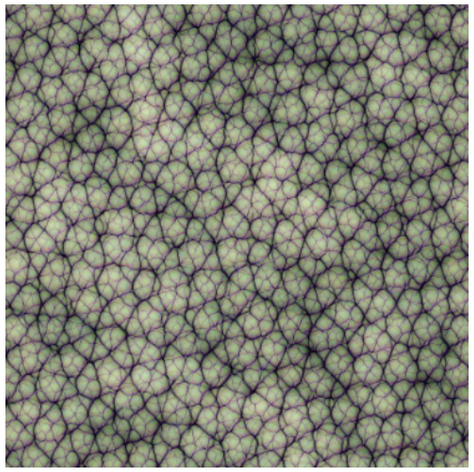
Two major themes run throughout the course

- Language has structure
  - There are patterns in what we say; this can be exploited this for more efficient learning and inference
- Language processing involves ambiguity resolution
  - There is ambiguity in what we say; this has to be resolved, e.g. by probabilistic models

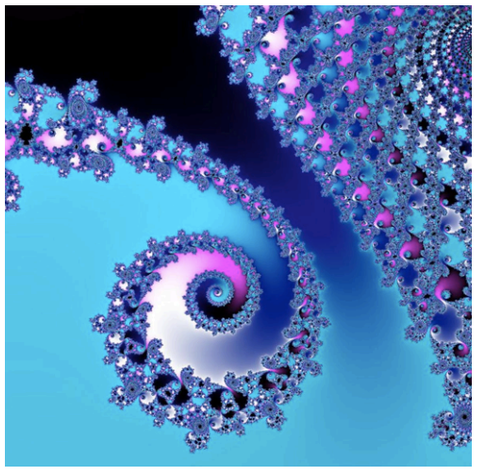
What is structure?



No apparent structure



Some structure



Some structure

Structure = there is some pattern, not just randomness

What is structure?

sdfpfkgpowkpork
opvsdkofaewpewmd
fdfadffpkbwkr

No apparent structure

abc abc abc
efg efg efg
xyz xyz xyz

Some structure

How do you describe this image?

There are infinite set of sentences:

a cute dog
a very cute dog
super cute puppy
adorable puppy looking at me
....

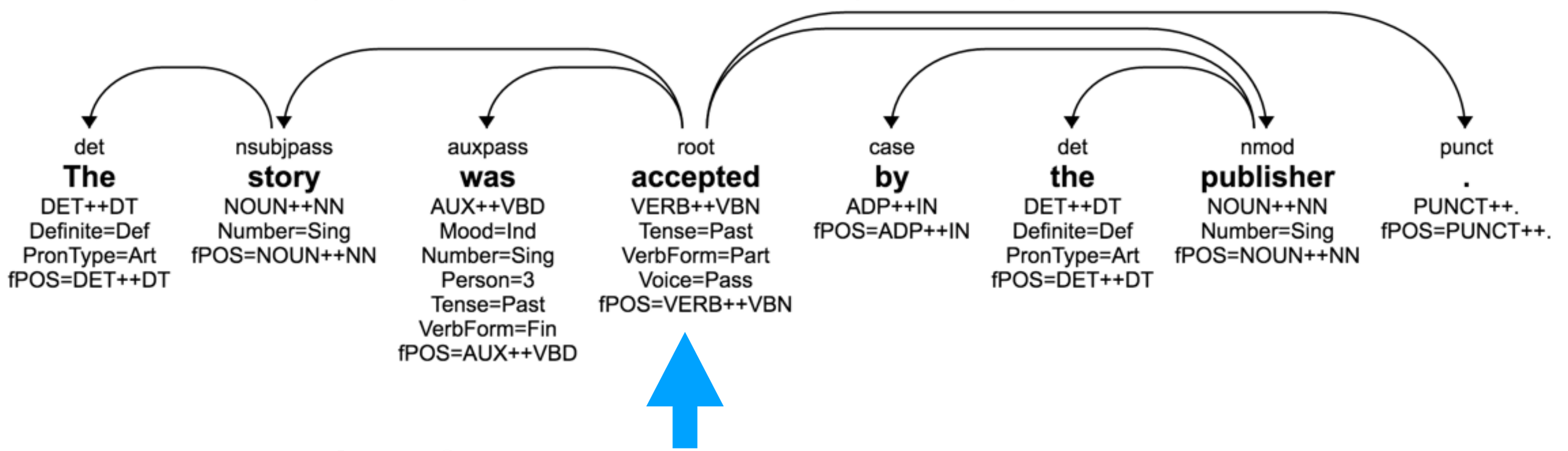
But not all are likely:
dog cute a
dog cut a very
puppy cute super
me at puppy looking adorable at
....

NLP Problems may have structure in both input and output

- Standard setup in machine learning:
  - Input **x** is a vector in R^D
  - Output **y** is a label from {class1, class2, class3, … classK}
- Characteristics of NLP problems:
  - **x** is a word or sentence: discrete input, with structure
  - **y** may be label from {class1, class2, class3, … classK}
  - **y** may also be an instance of a large structured space

Structure in Syntactic Analysis

- **<u>Output Tree structure</u>**
  - For each word, predict its “head” (arrow)
  - For N words, there are N+1 possible labels per word
  - But not all (N+1)xN prediction is a valid dependency tree



Input Sentence:
The story was accepted by the publisher .

From: https://ai.googleblog.com/2016/08/meet-parseys-cousins-syntax-for-40.html

Structure in Question Answering

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail. Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud.

Question:
What causes precipitation to fall?

Answer: gravity

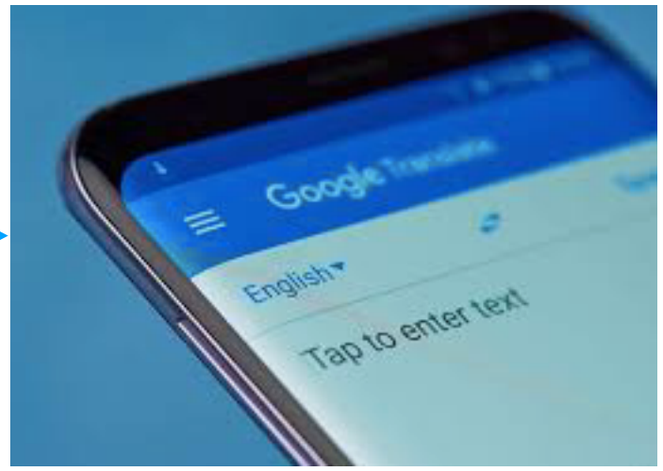
Question:
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

Answer:

From: Rajpurkar et. al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP2016. https://aclweb.org/anthology/D16-1264

Structure in Machine Translation

Cogito, ergo sum



I think, therefore I am

Exploiting structure

- Many of the algorithms we will learn will exploit structure, e.g. sequence or tree structure for efficiency

- We'll learn about Viterbi algorithm for decoding sequences, Eisner algorithm for parsing trees, etc.

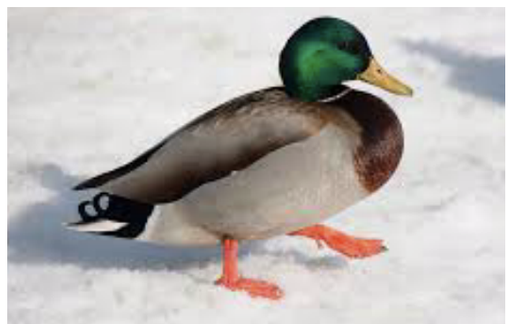
Language is full of ambiguity

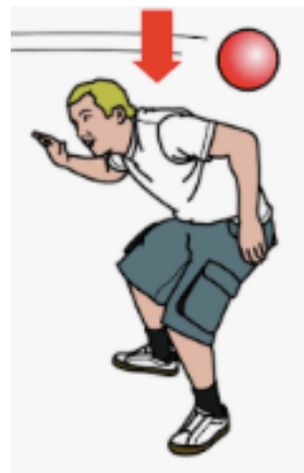
- “I made her duck”

From: Jurafsky and Martin, Speech and Language Processing, 2nd ed.

To resolve ambiguity, we'll be exploring different probabilistic models

- Counting statistics
  - e.g. how many times "duck" means





- Linear (or log-linear) models
  - e.g. extract features: "I", "made", "her", "duck" and combine with weights
- Neural network models
  - can be viewed as a logical extension of linear models

A note on machine learning

- Machine learning powers many NLP models. But generic machine learning methods by itself is not sufficient for good performance.

- In this class, I hope you'll come to appreciate the intricacies of language and see how we can see how we exploit its structure for better machine learning.

Today's Outline

1. **Survey of NLP Problems**

2. **Adminstration:**

    - Course expectations
    - Procedures (wait list, etc.)

3. **Two Major Themes**

    - Language has structure
    - Language processing involves ambiguity resolution

Next

- Grammaticality:
  - What are allowed structures in an English sentence
  - How do we model them? (Context Free Grammar, CFG)
- First homework