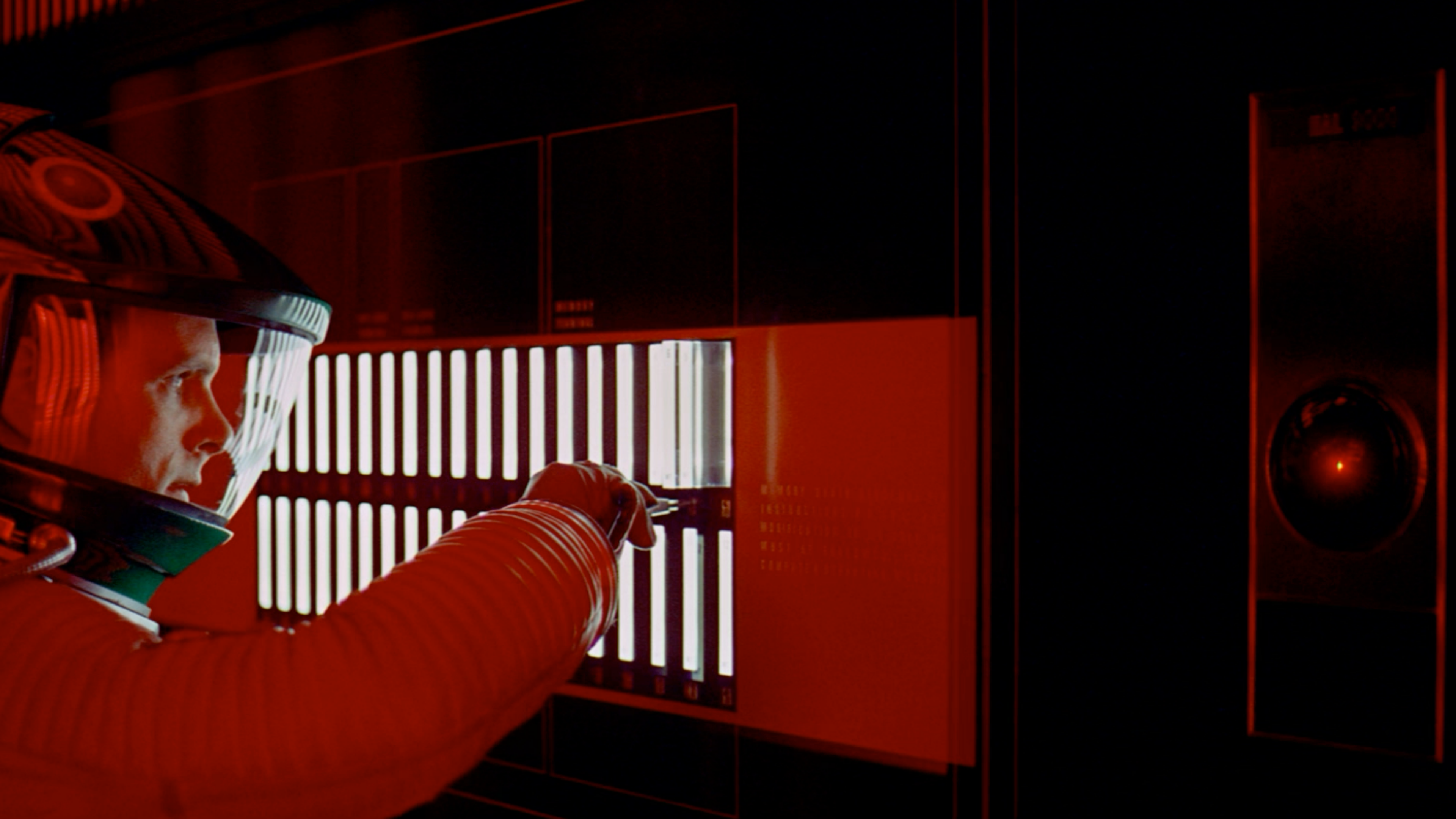
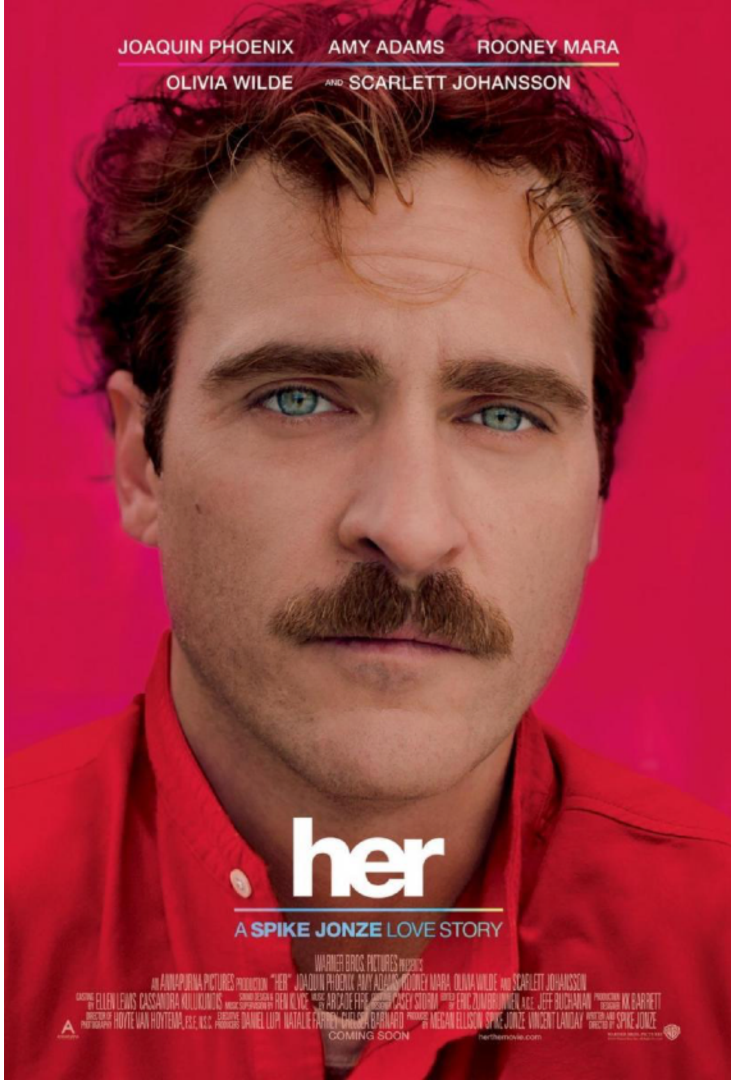
# Applied Natural Language Processing

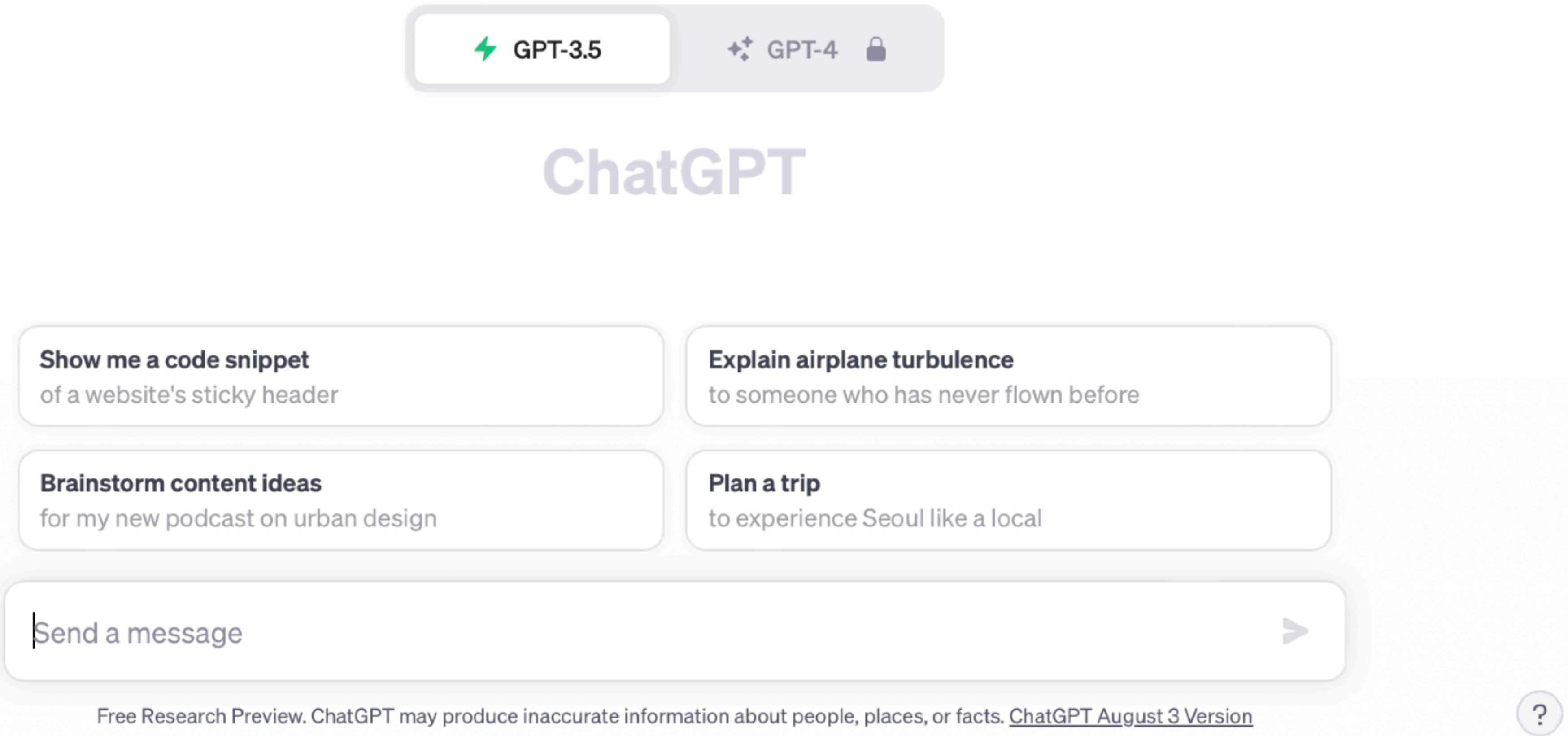
Info 256

Lecture 1: Introduction (August 23, 2023)

David Bamman, UC Berkeley

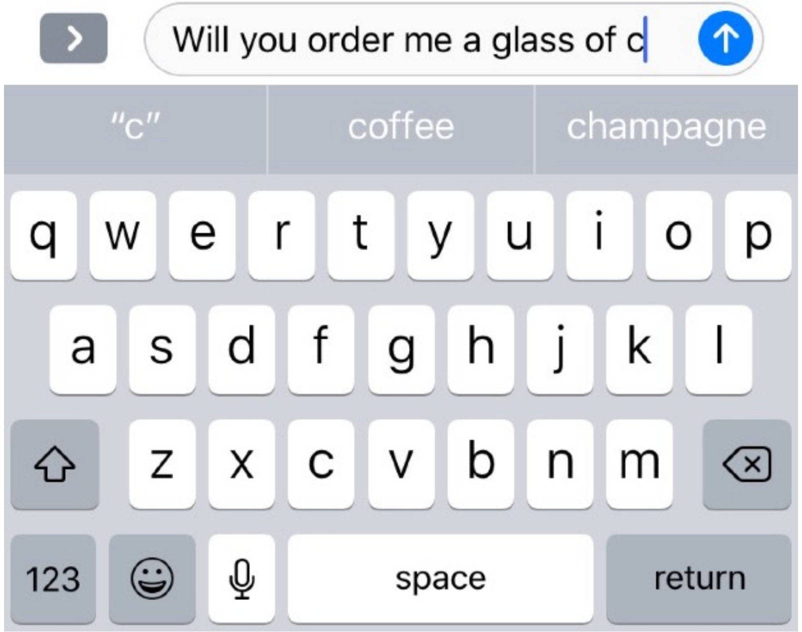

MEMORY
HAL 9000


JOAQUIN PHOENIX
AMY ADAMS
ROONEY MARA
OLIVIA WILDE
AND SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY
WARNER BROS. PICTURES PRESENTS
AN ANNAPURNA PICTURES PRODUCTION "HER" JOAQUIN PHOENIX AMY ADAMS ROONEY MARA OLIVIA WILDE AND SCARLETT JOHANSSON
CASTING BY ELLEN LEWIS CASSANDRA KULUKUNDIS SOUND DESIGN & MUSIC SUPERVISION BY REN KLYCE MUSIC BY ARCADE FIRE COSTUME DESIGNER CASEY STORM EDITED BY ERIC ZUMBRUNNEN, A.C.E. JEFF BUCHANAN PRODUCTION DESIGNER KK BARRETT
DIRECTOR OF PHOTOGRAPHY HOYTE VAN HOYTEMA, F.S.F., N.S.C. EXECUTIVE PRODUCERS DANIEL LUPI NATALIE FARREY CHELSEA BARNARD PRODUCED BY MEGAN ELLISON SPIKE JONZE VINCENT LANDAY WRITTEN AND DIRECTED BY SPIKE JONZE
COMING SOON
herthemovie.com


GPT-3.5
GPT-4
ChatGPT
Show me a code snippet
of a website's sticky header
Explain airplane turbulence
to someone who has never flown before
Brainstorm content ideas
for my new podcast on urban design
Plan a trip
to experience Seoul like a local
Send a message
?

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version

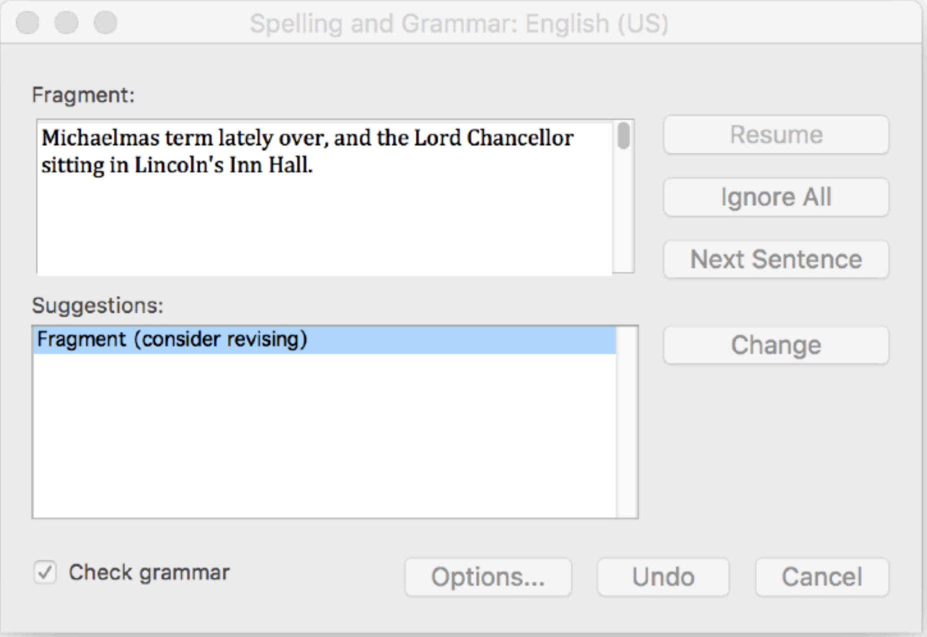
# Predictive text messaging



Will you order me a glass of c

# Grammar checking

London. Michaelmas term lately over, and the Lord Chancellor sitting in Lincoln's Inn Hall. Implacable November weather. As much mud in the streets as if the waters had but newly retired from the face of the earth, and it would not be wonderful to meet a Megalosaurus, forty feet long or so, waddling like an elephantine lizard up Holborn Hill.



Spelling and Grammar: English (US)

Fragment:

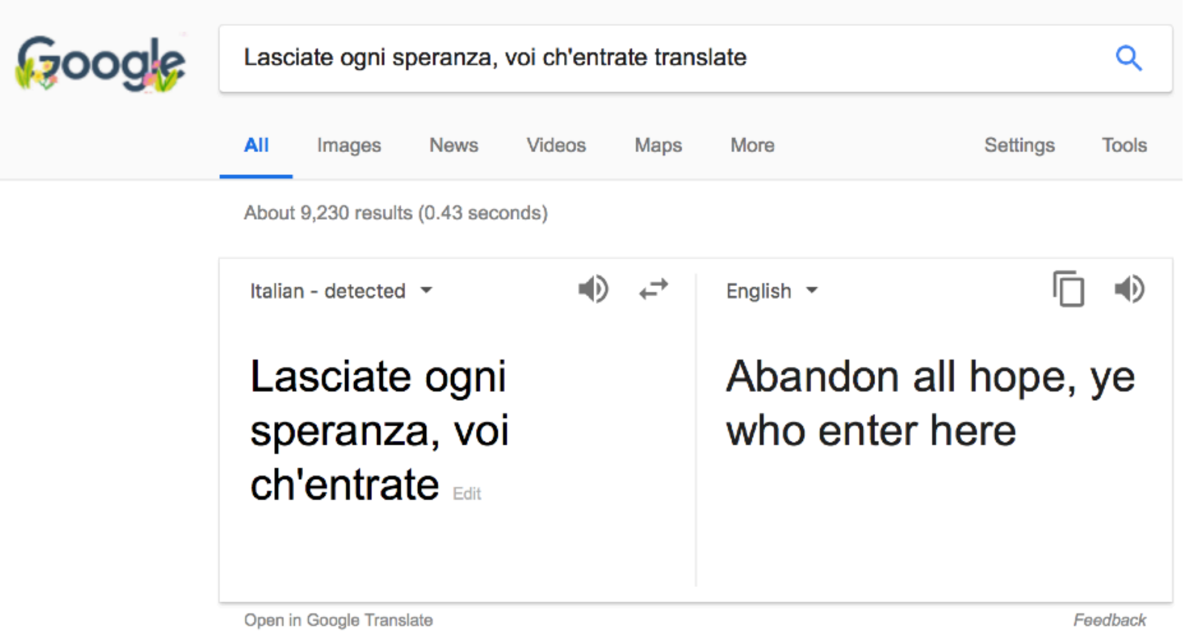
Michaelmas term lately over, and the Lord Chancellor sitting in Lincoln's Inn Hall.

Suggestions:

Fragment (consider revising)

☑ Check grammar

# Machine translation



Lasciate ogni speranza, voi ch'entrate translate

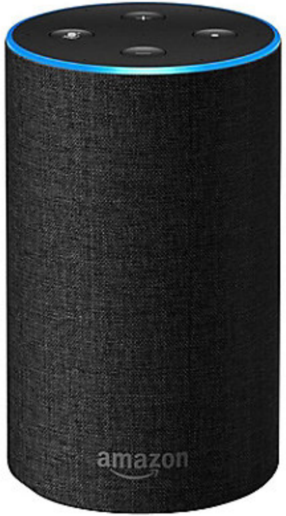
About 9,230 results (0.43 seconds)

Lasciate ogni speranza, voi ch'entrate

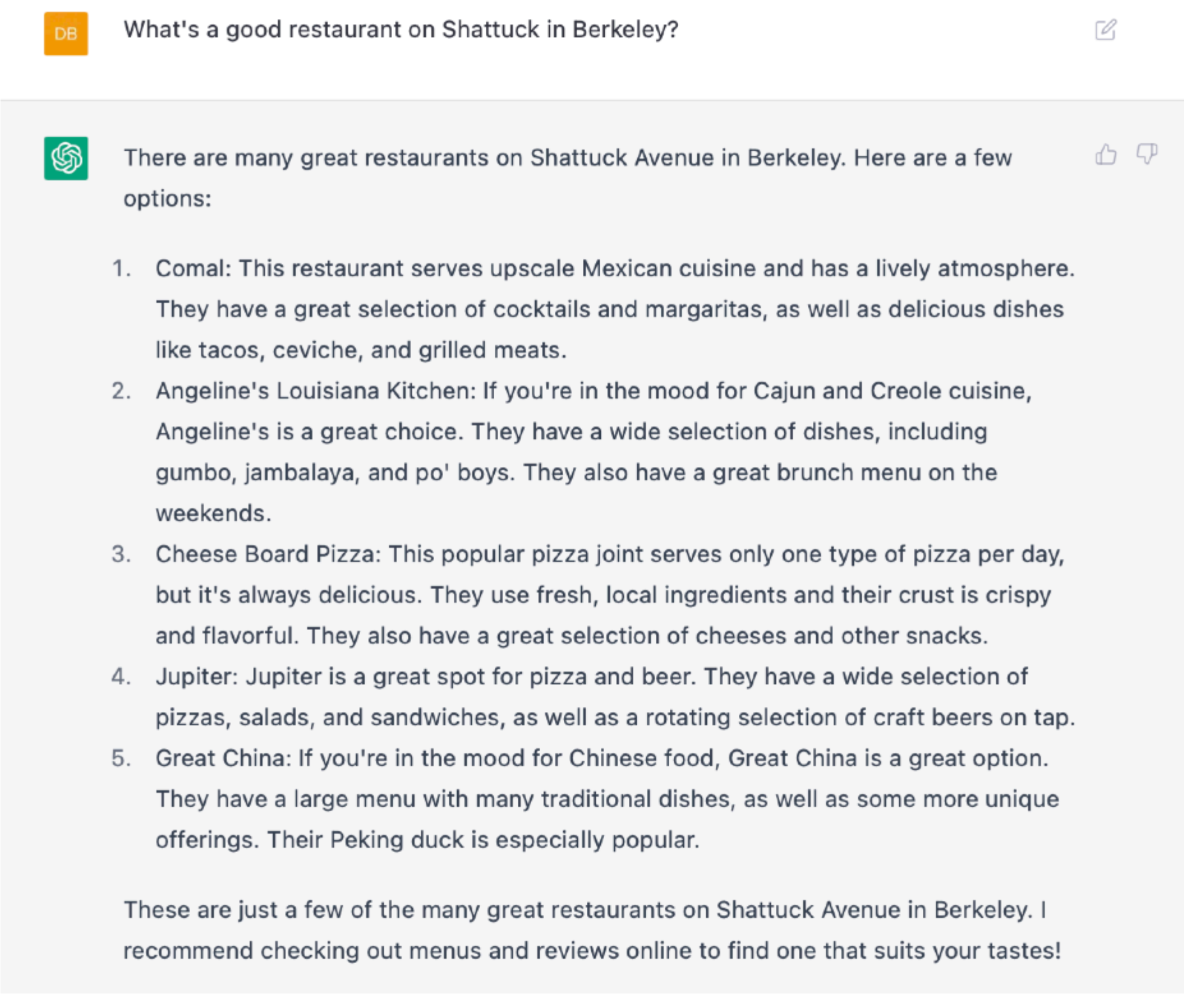
Abandon all hope, ye who enter here

# Speech Recognition

“Alexa, how many cups are in a quart?”



# Question Answering



What's a good restaurant on Shattuck in Berkeley?

There are many great restaurants on Shattuck Avenue in Berkeley. Here are a few options:

1. Comal: This restaurant serves upscale Mexican cuisine and has a lively atmosphere. They have a great selection of cocktails and margaritas, as well as delicious dishes like tacos, ceviche, and grilled meats.
2. Angeline's Louisiana Kitchen: If you're in the mood for Cajun and Creole cuisine, Angeline's is a great choice. They have a wide selection of dishes, including gumbo, jambalaya, and po' boys. They also have a great brunch menu on the weekends.
3. Cheese Board Pizza: This popular pizza joint serves only one type of pizza per day, but it's always delicious. They use fresh, local ingredients and their crust is crispy and flavorful. They also have a great selection of cheeses and other snacks.
4. Jupiter: Jupiter is a great spot for pizza and beer. They have a wide selection of pizzas, salads, and sandwiches, as well as a rotating selection of craft beers on tap.
5. Great China: If you're in the mood for Chinese food, Great China is a great option. They have a large menu with many traditional dishes, as well as some more unique offerings. Their Peking duck is especially popular.

These are just a few of the many great restaurants on Shattuck Avenue in Berkeley. I recommend checking out menus and reviews online to find one that suits your tastes!

# NLP

If you’re interested in the core methods and algorithms, take Info 159/259 (NLP) instead.

- language modeling
- sequence labeling
- phrase-structure parsing
- dependency parsing
- dynamic programming
- MT

# Applied NLP

How do we use the methodologies in NLP toward some end?

# Software/Libraries

NumPy

HUGGING FACE

SciPy

scikit learn

spaCy

PyTorch

Keras

**NLTK**

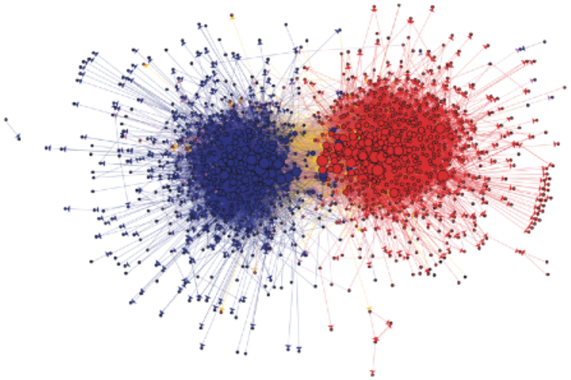
TensorFlow

# NLP is interdisciplinary

- Artificial intelligence
- Machine learning (ca. 2000—today); statistical models, neural networks
- Linguistics (representation of language)
- Social sciences/humanities (models of language at use in culture/society)

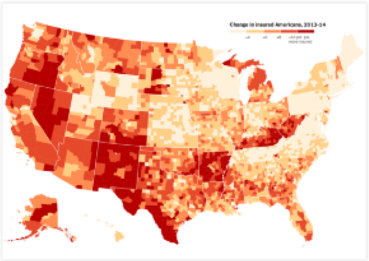
# NLP is interdisciplinary

Computational Social Science



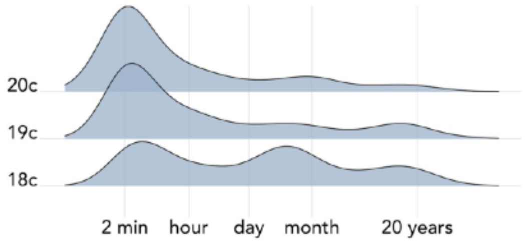
Adamic and Glance 2005

Computational Journalism



Change in insured Americans under the ACA,
NY Times (Oct 29, 2014)

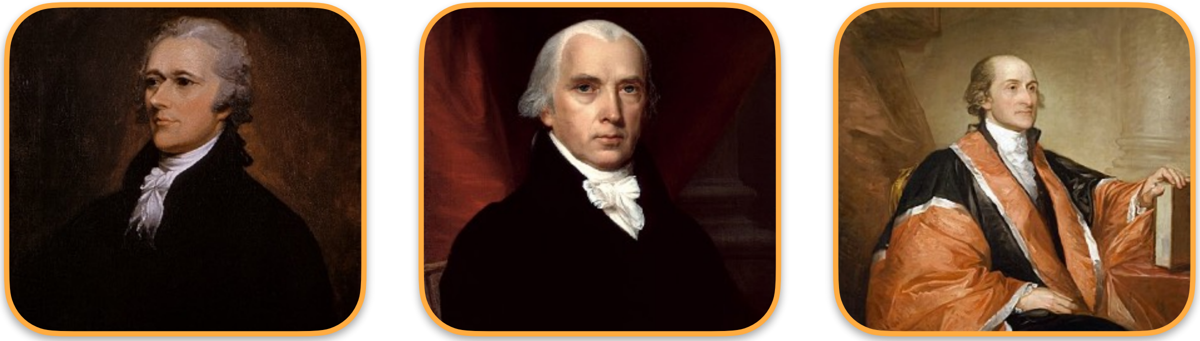
Computational Humanities

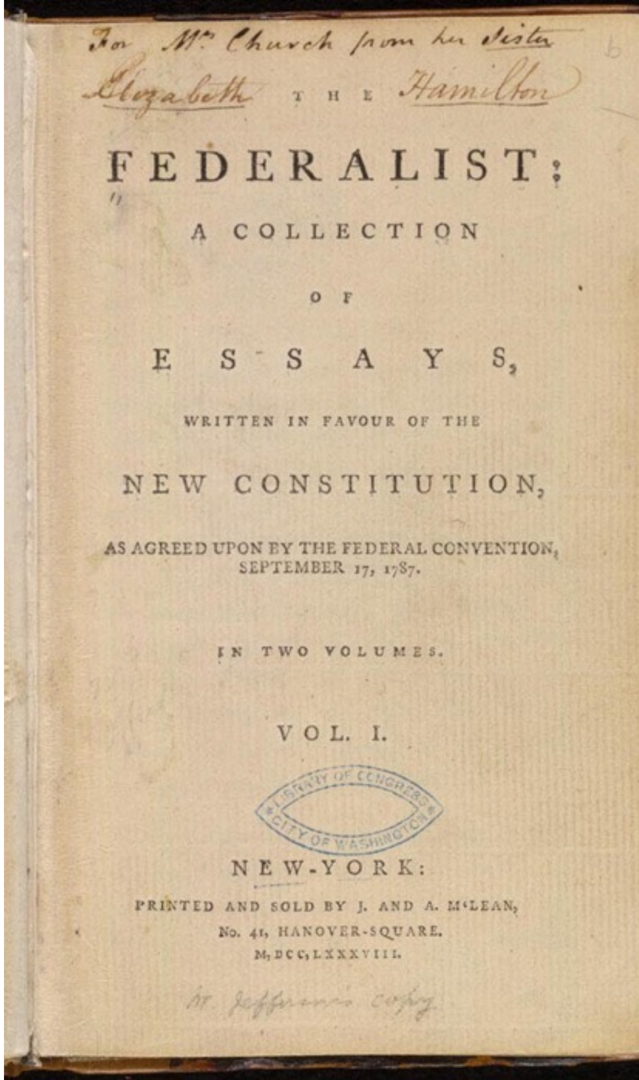


Underwood 2018

# Authorship attribution

- Mosteller and Wallace (1963) use Bayesian models over word counts to infer authorship of unknown or contested works in the *Federalist Papers* (between Alexander Hamilton, James Madison, John Jay)





For Mrs Church from her Sister
Elizabeth Hamilton

THE
FEDERALIST:
A COLLECTION
OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

IN TWO VOLUMES.

VOL. I.

LIBRARY OF CONGRESS
CITY OF WASHINGTON

NEW-YORK:
PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.
M,DCC,LXXXVIII.

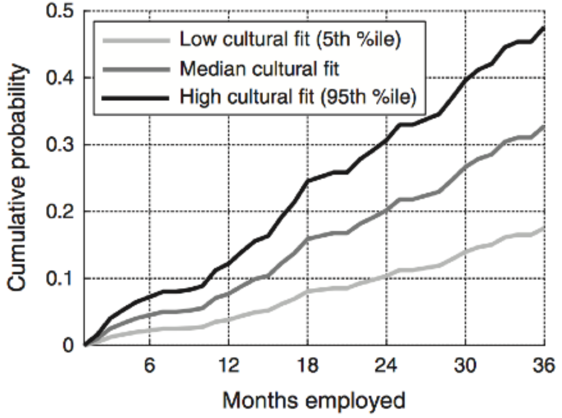
Mr Jefferson's copy

# Enculturation

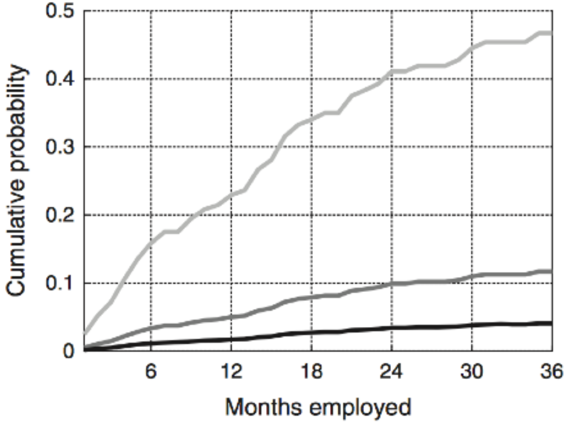
Input: employee emails

Output: promotion to manager, time to separation

(A) Promotion



(B) Involuntary exit

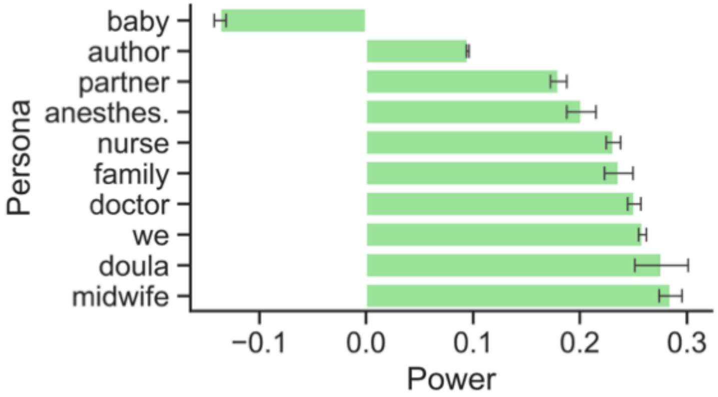


Srivastava et al. (2017), "Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations" (*Management Science*)

# Power in Birth Stories

- Topic modeling, connotation frames and personas to measure the structure of birth stories and representation of power within them.

“The doctor broke my water.”



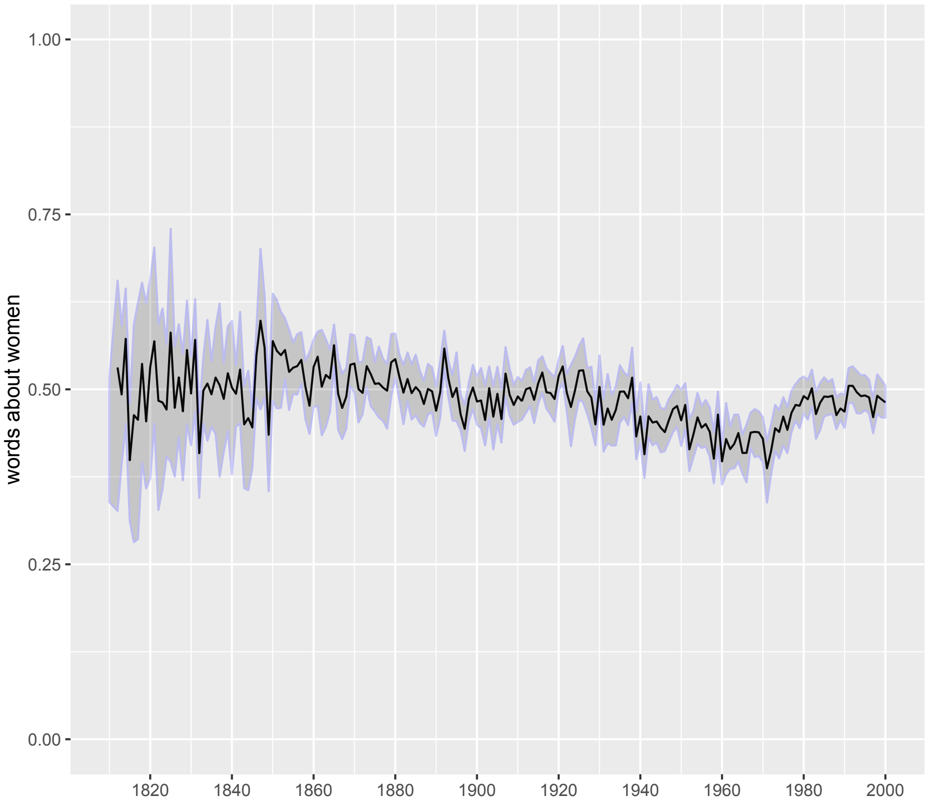
Antoniak et al. (2019), “Narrative Paths and Negotiation of Power in Birth Stories,” CSCW

# Respect

Input: transcripts of 981 OPD traffic stops (everyday interactions)

Output: measure of “respect” directed from officer to driver

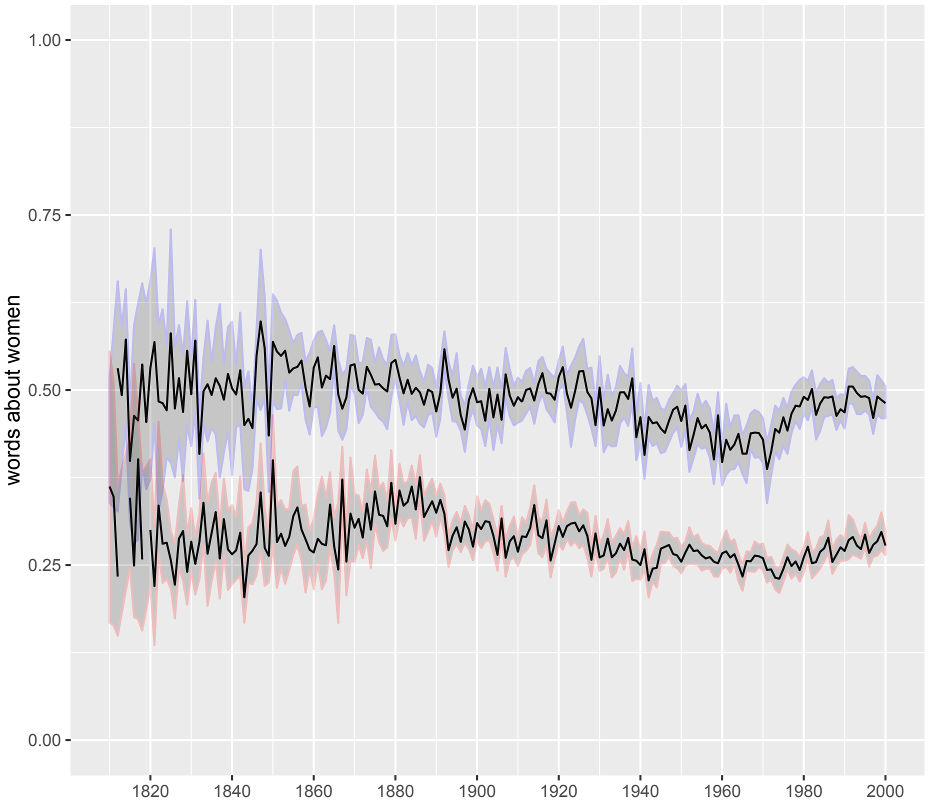
Voigt et al. 2017, “Language from police body camera footage shows racial disparities in officer respect”



written by women

Fraction of words about female characters

Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," (*Cultural Analytics*)



written by women

written by men

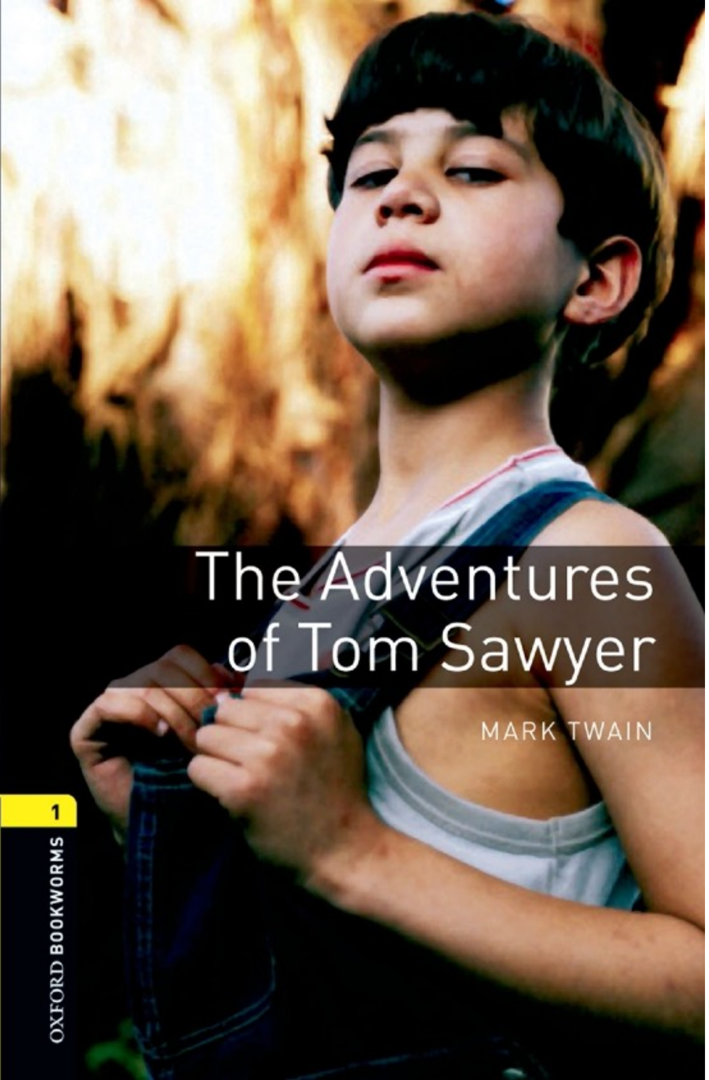
Fraction of words about female characters

Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," (*Cultural Analytics*)

# Measurement

This is fundamentally a problem of measurement: how do we design an algorithmic instrument that can transform a text into a quantity?

“TOM!” No answer. “TOM!” No answer. “What's gone with that boy, I wonder? You TOM!” No answer. The old lady pulled her spectacles down and looked over them about the room; then she put them up and looked out under them. She seldom or never looked *through* them for so small a thing as a boy; they were her state pair, the pride of her heart, and were built for “style,” not service--she could have seen through a pair of stove-lids just as well. She looked perplexed for a moment, and then said, not fiercely, but still loud enough for the furniture to hear: “Well, I lay if I get hold of you I'll--” She did not [redacted] ime she was bending down and punching under the bed with the 0.53 e needed breath to punctuate the punches with. She resurrecte cat. “I never did see the beat of that boy!” She went to the open door and stood in it and looked out among the tomato vines and “jimpson” weeds that constituted the garden. No Tom. So she lifted up her voice at an angle calculated for distance and shouted: “Y-o-u-u TOM!” There was a slight noise behind her and she turned just in time to seize a small boy by the slack of his roundabout and arrest his flight. “There! I might 'a' thought of that closet. What you been doing in there?” “Nothing.” “Nothing! Look at your hands. And look at your mouth. What *is* that truck?” “I don't know, aunt.”



"TOM!"

No answer.

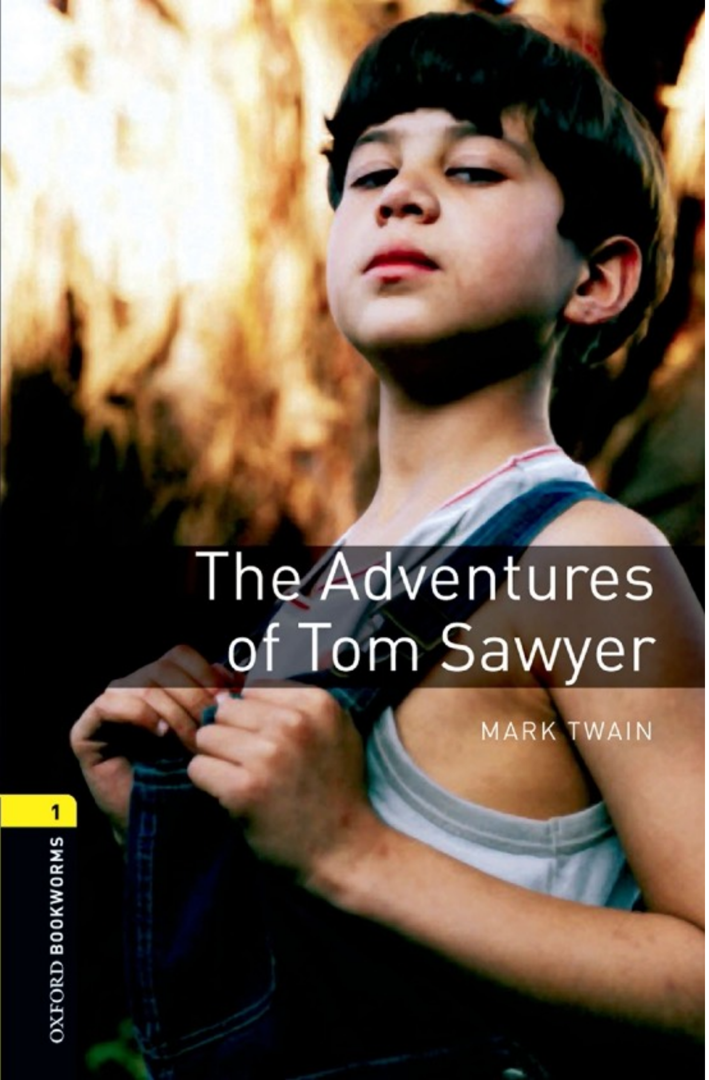
"TOM!"

No answer.

"What's gone with that boy,  I wonder? You TOM!"

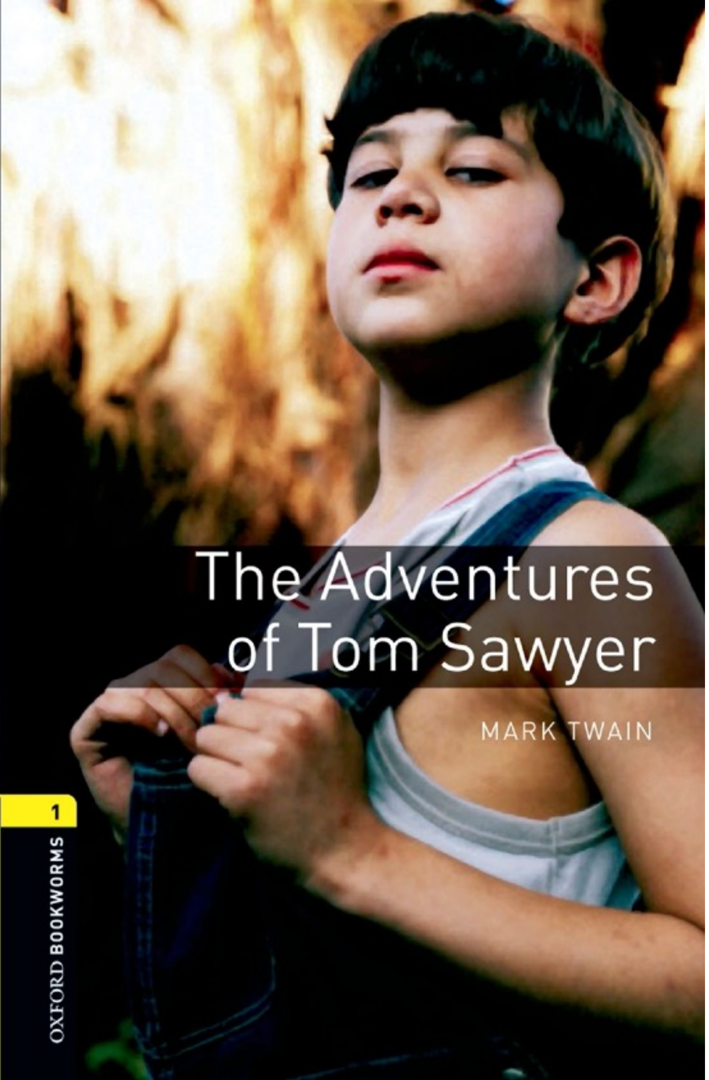
No answer.

The old lady pulled her spectacles down and looked over them about the room.



# Bag of Words

tom no answer tom no answer what's gone with that boy , I wonder ? you tom ! no answer the old lady pulled her spectacles down and looked over them about the room .



"TOM!"

No answer.

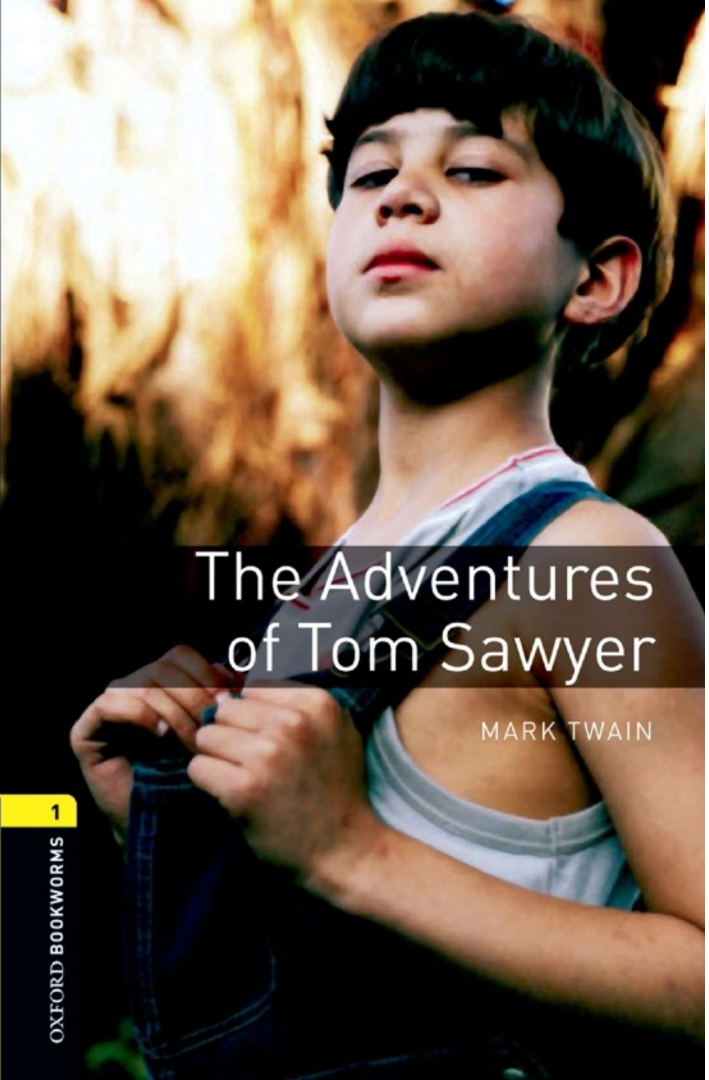
"TOM!"

No answer.

"What's gone with that boy,  I wonder?
You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.



nouns

"TOM!"

No answer.

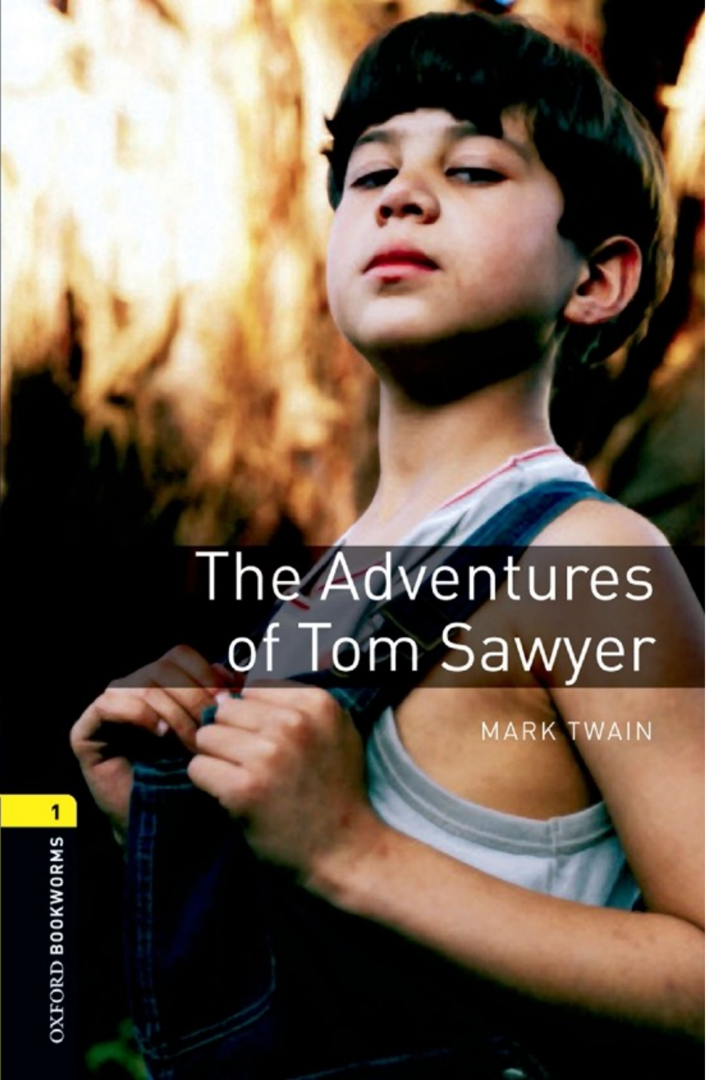
"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.



people

"TOM!"

No answer.

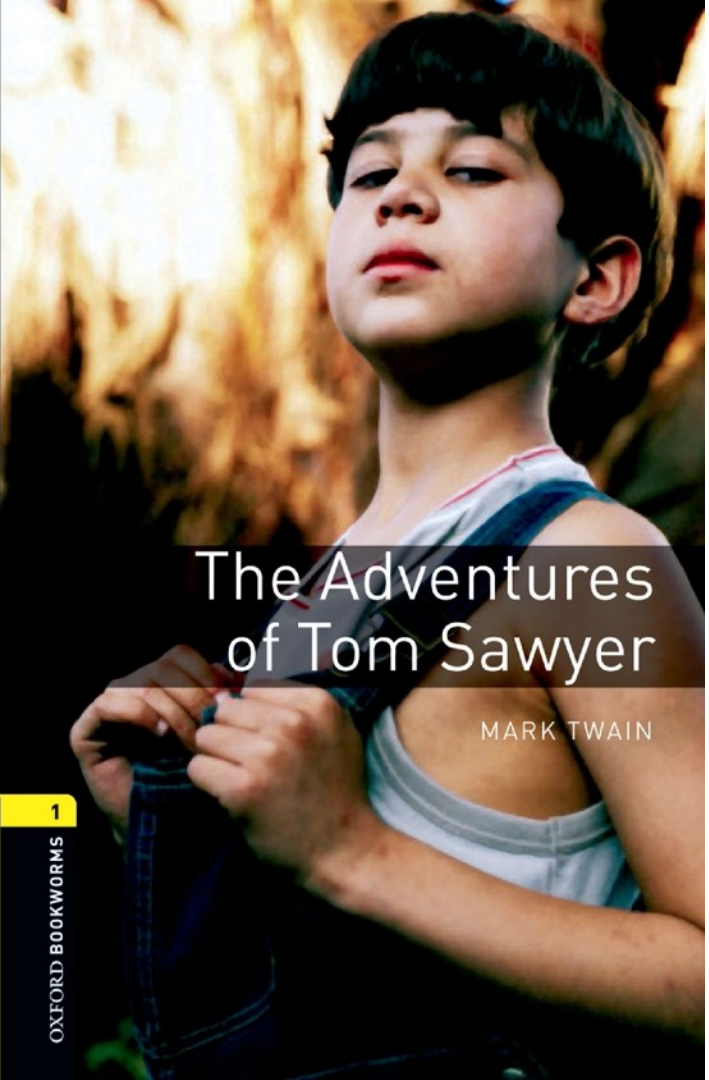
"TOM!"

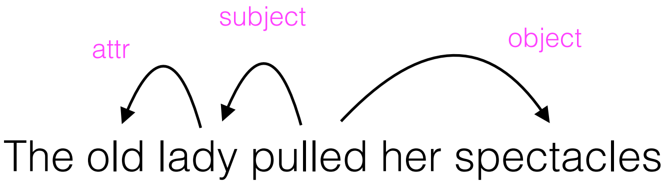
No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

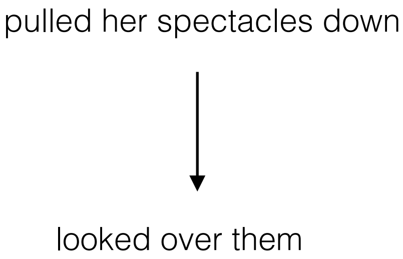
The old lady pulled her spectacles down and looked over them about the room.





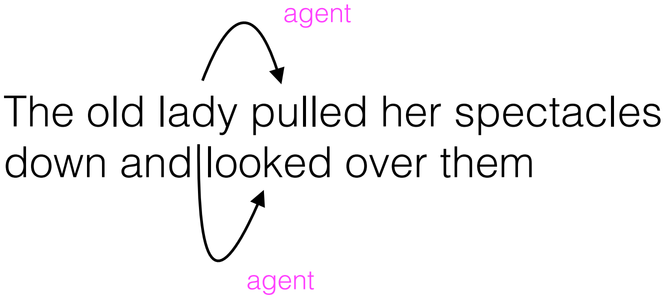
The old lady pulled her spectacles

# Temporal sequence



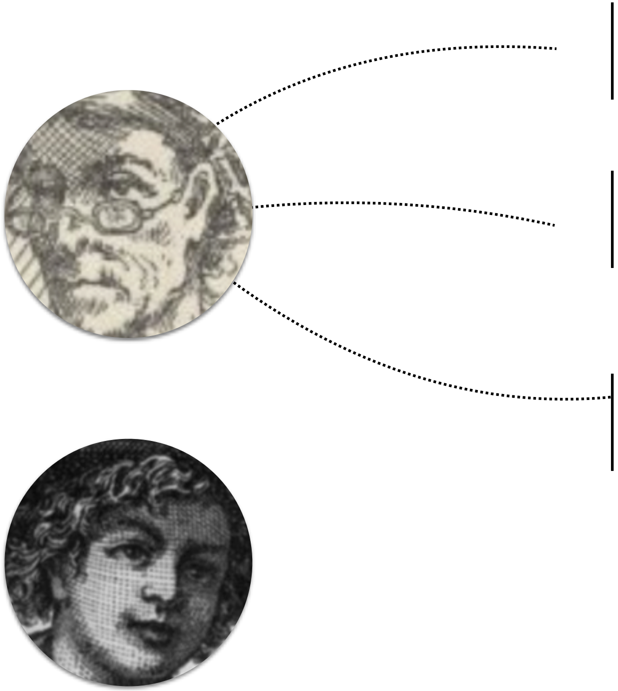
pulled her spectacles down

looked over them



The old lady pulled her spectacles down and looked over them

# Speaker identification



"TOM!"

No answer.

"TOM!"

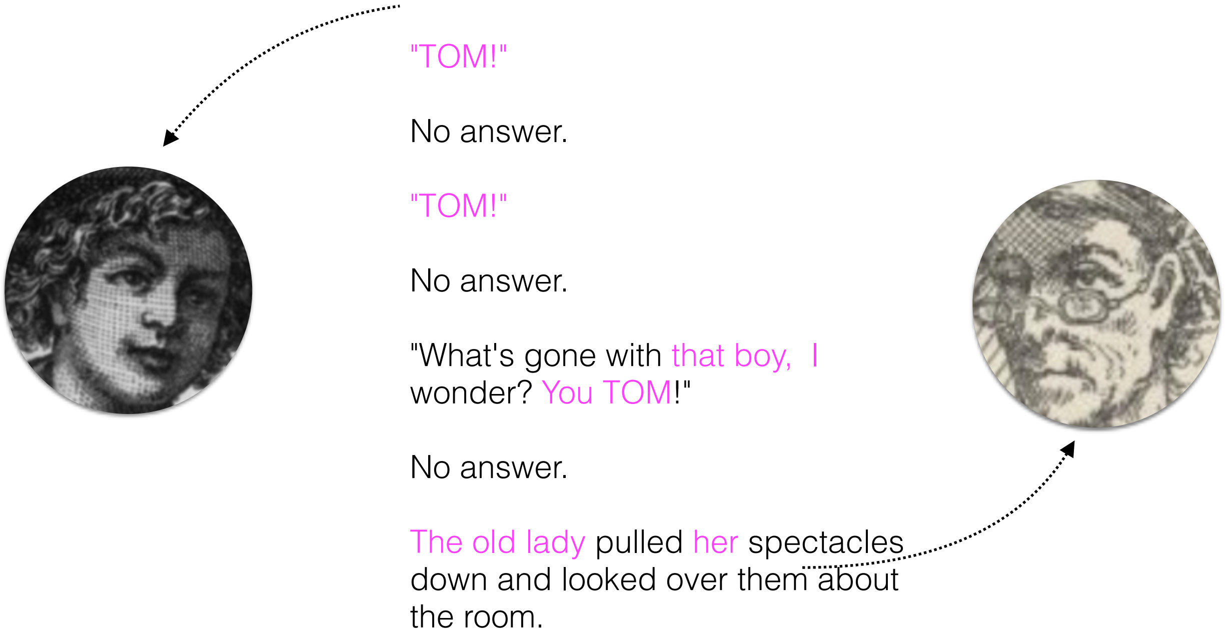
No answer.

"What's gone with that boy,  I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

# Coreference



"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.

# What makes language hard?

- Language is a complex social process
- Tremendous ambiguity at every level of representation
- Modeling it is AI-complete (requires first solving general AI)

# What makes language hard?

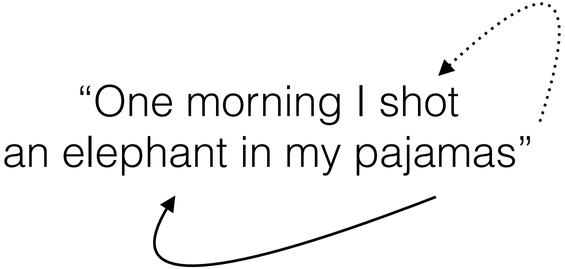
- Speech acts (“can you pass the salt?)
  [Austin 1962, Searle 1969]
- Conversational implicature (“The opera singer was amazing; she sang all of the notes”).
  [Grice 1975]
- Shared knowledge (“Warren is running for election”)
- Variation/Indexicality (“This homework is wicked hard”)
  [Labov 1966, Eckert 2008]

# Ambiguity

“One morning I shot
an elephant in my pajamas”

*Animal Crackers*

# Ambiguity



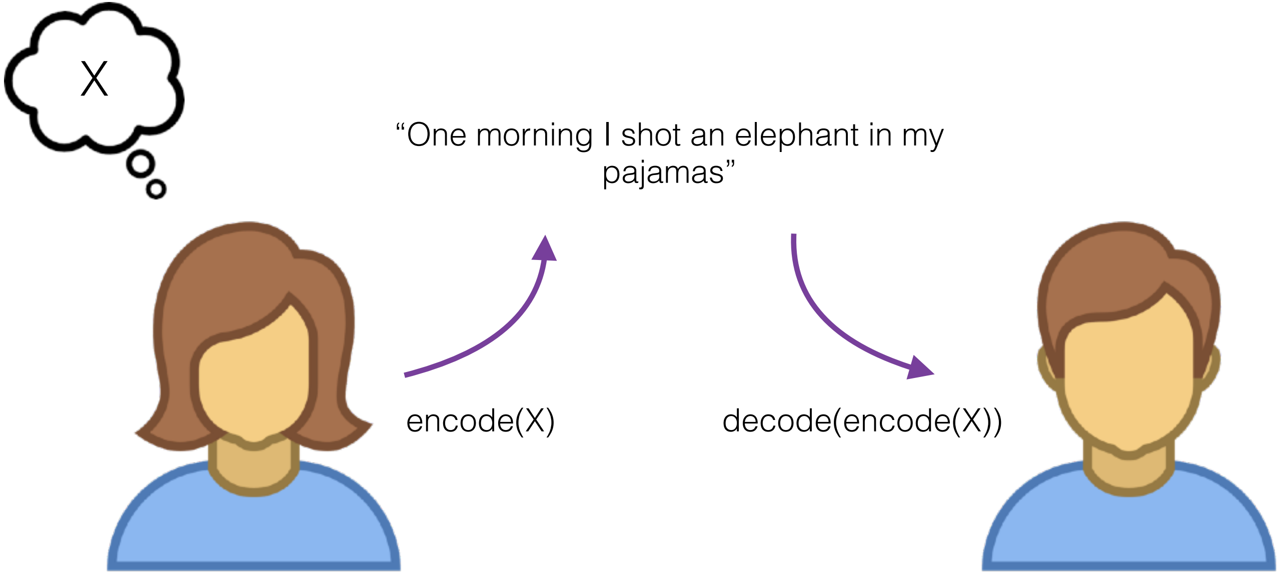
“One morning I shot
an elephant in my pajamas”

*Animal Crackers*

# Ambiguity

“One morning I shot
an elephant in my pajamas”

# Information theoretic view



X

“One morning I shot an elephant in my pajamas”

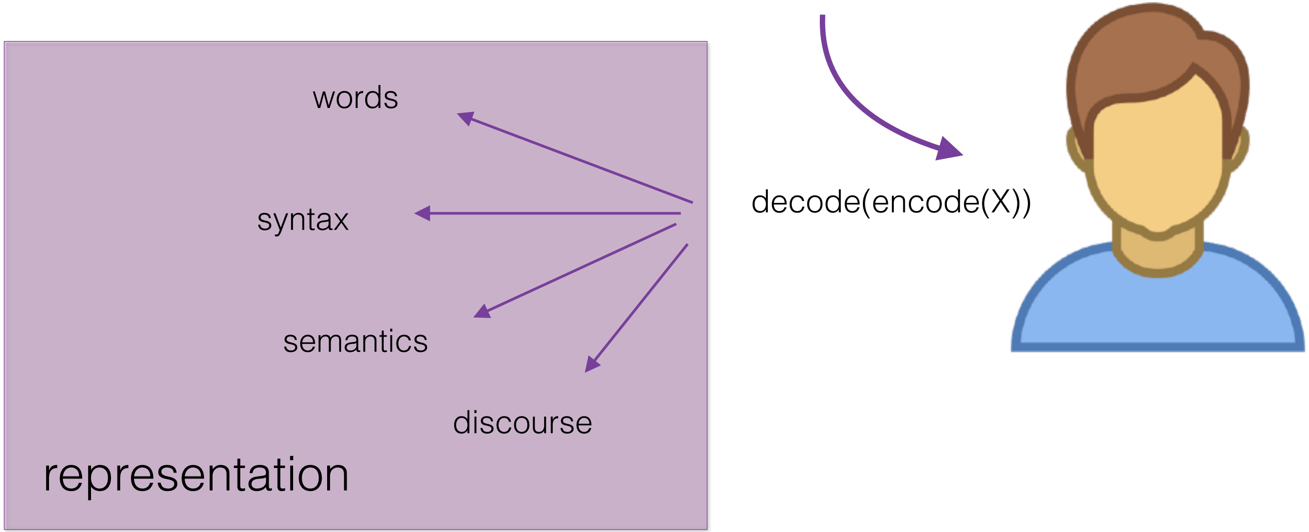
encode(X)

decode(encode(X))

Shannon 1948

# Decoding

“One morning I shot an elephant in my pajamas”



decode(encode(X))

words

syntax

semantics

discourse

representation

# “Raw” data

- We often want to make claims about the world using textual data.
- Data is not self-evident, neutral or objective
- Data is collected, stored, processed, mined, interpreted; each stage requires our participation.
- What is the process by which the data you have got to you?

Gitelman and Jackson (2013); D’Ignazio and Klein (2020)

# Administrivia

- David Bamman
  dbamman@berkeley.edu
- Office hours:
    - Wednesdays 10am-noon, 314 South Hall
- Kent Chang, TA
  kentchang@berkeley.edu

# Info 256

- Each class period will be divided between:
  - a short lecture; and
  - in-class lab work using Jupyter notebooks
- Students must prepare for each class and submit homeworks before class.

# Grading

- Homeworks (40%)
- Participation (10%)
- Group project (50%)

# Late submissions

- All homeworks are due on the date/time specified, before each class. We'll go over the homework in class, so no late homeworks.

- You can drop 3 homeworks.

# Homeworks

- Homeworks will be frequent; you are free to discuss them at a high level with your classmates, but all coding must be done individually.
- If you use or build on others' code (e.g., from StackOverflow), you must cite its source.
- UC Berkeley code of conduct: http://sa.berkeley.edu/code-of-conduct

# Participation

- Participation includes:
    - Coming to class and working in groups. Attendance is required!
    - Peer assessment of homework and project deliverables.

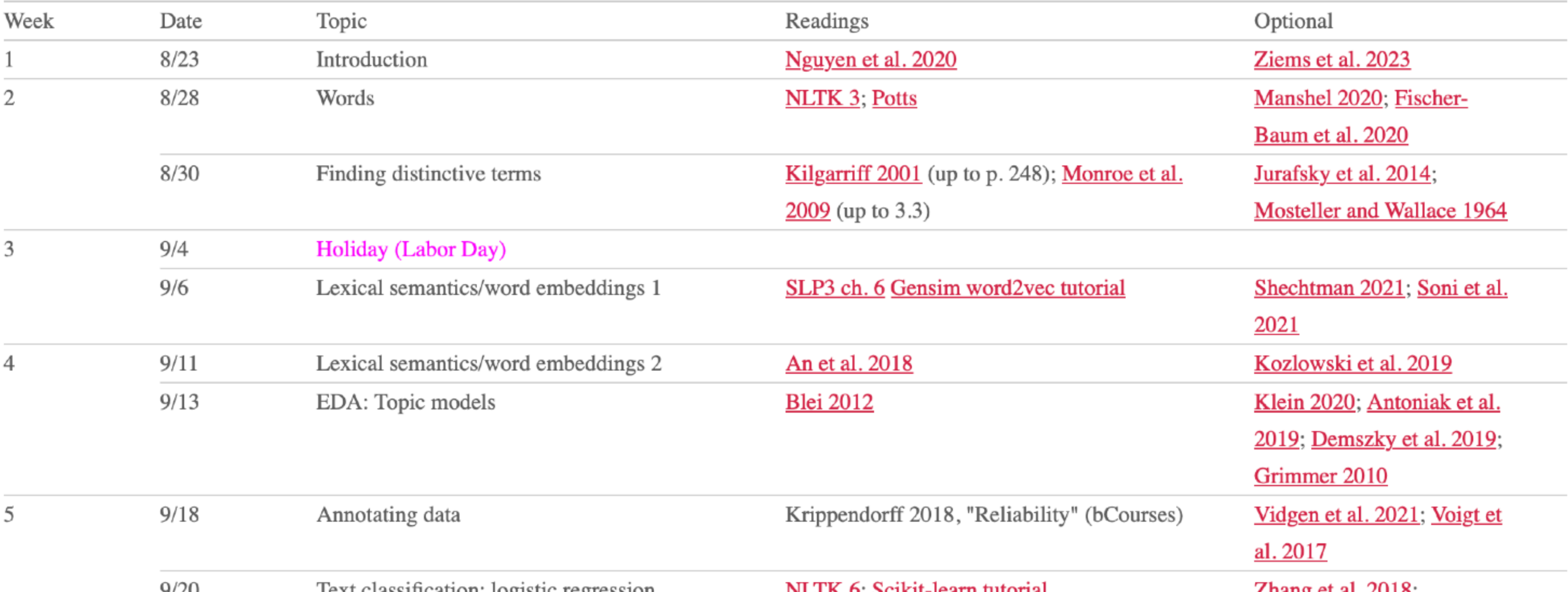
# Course project

- Semester-long project (involving 1-3 students) , involving natural language processing in support of an empirical research question.
    - Project proposal/literature review
    - Midterm report
    - 6-page final report, workshop quality
    - Project presentation

# ACL 2023 workshops

- The 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)
- The 7th Workshop on Online Abuse and Harms
- The 17th Workshop on Linguistic Annotation (LAW)
- The 22nd Workshop on Biomedical Natural Language Processing (BioNLP)
- The 5th Workshop on NLP for Conversational AI
- The 3rd Workshop on Trustworthy NLP
- The 5th Clinical Natural Language Processing Workshop (Clinical NLP)
- The 3rd Workshop on NLP for Indigenous Languages of the Americas
- The 5th Workshop on Narrative Understanding

# Syllabus



| Week | Date | Topic | Readings | Optional |
|---|---|---|---|---|
| 1 | 8/23 | Introduction | Nguyen et al. 2020 | Ziems et al. 2023 |
| 2 | 8/28 | Words | NLTK 3; Potts | Manshel 2020; Fischer-Baum et al. 2020 |
|  | 8/30 | Finding distinctive terms | Kilgarriff 2001 (up to p. 248); Monroe et al. 2009 (up to 3.3) | Jurafsky et al. 2014; Mosteller and Wallace 1964 |
| 3 | 9/4 | Holiday (Labor Day) |  |  |
|  | 9/6 | Lexical semantics/word embeddings 1 | SLP3 ch. 6 Gensim word2vec tutorial | Shechtman 2021; Soni et al. 2021 |
| 4 | 9/11 | Lexical semantics/word embeddings 2 | An et al. 2018 | Kozlowski et al. 2019 |
|  | 9/13 | EDA: Topic models | Blei 2012 | Klein 2020; Antoniak et al. 2019; Demszky et al. 2019; Grimmer 2010 |
| 5 | 9/18 | Annotating data | Krippendorff 2018, "Reliability" (bCourses) | Vidgen et al. 2021; Voigt et al. 2017 |
|  | 9/20 | Text classification: logistic regression | NLTK 6; Scikit-learn tutorial | Zhang et al. 2018; |

# Github

- Course homework will be on Github: https://github.com/dbamman/anlp23
- Sign up for an account right now if you don't have one!

# In class

- https://github.com/dbamman/anlp23
- anlp23/0.setup
- Install anaconda environment + libraries we'll use frequently.