Why NLP?

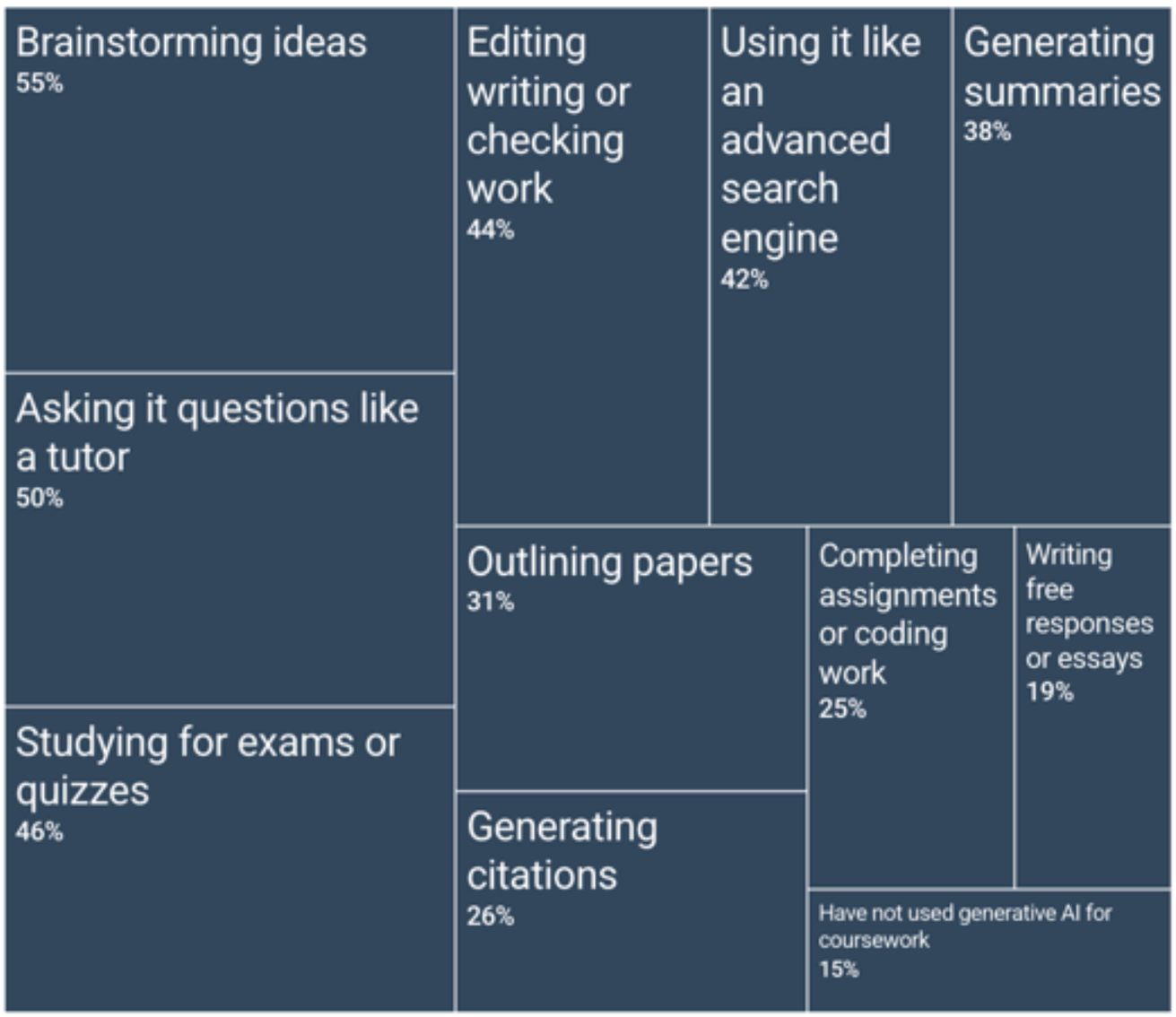
CS6120: Natural Language Processing
Northeastern University

David Smith

Using Generative AI for Coursework

Students on just how they've used generative AI for coursework in the last year, all and by institution type

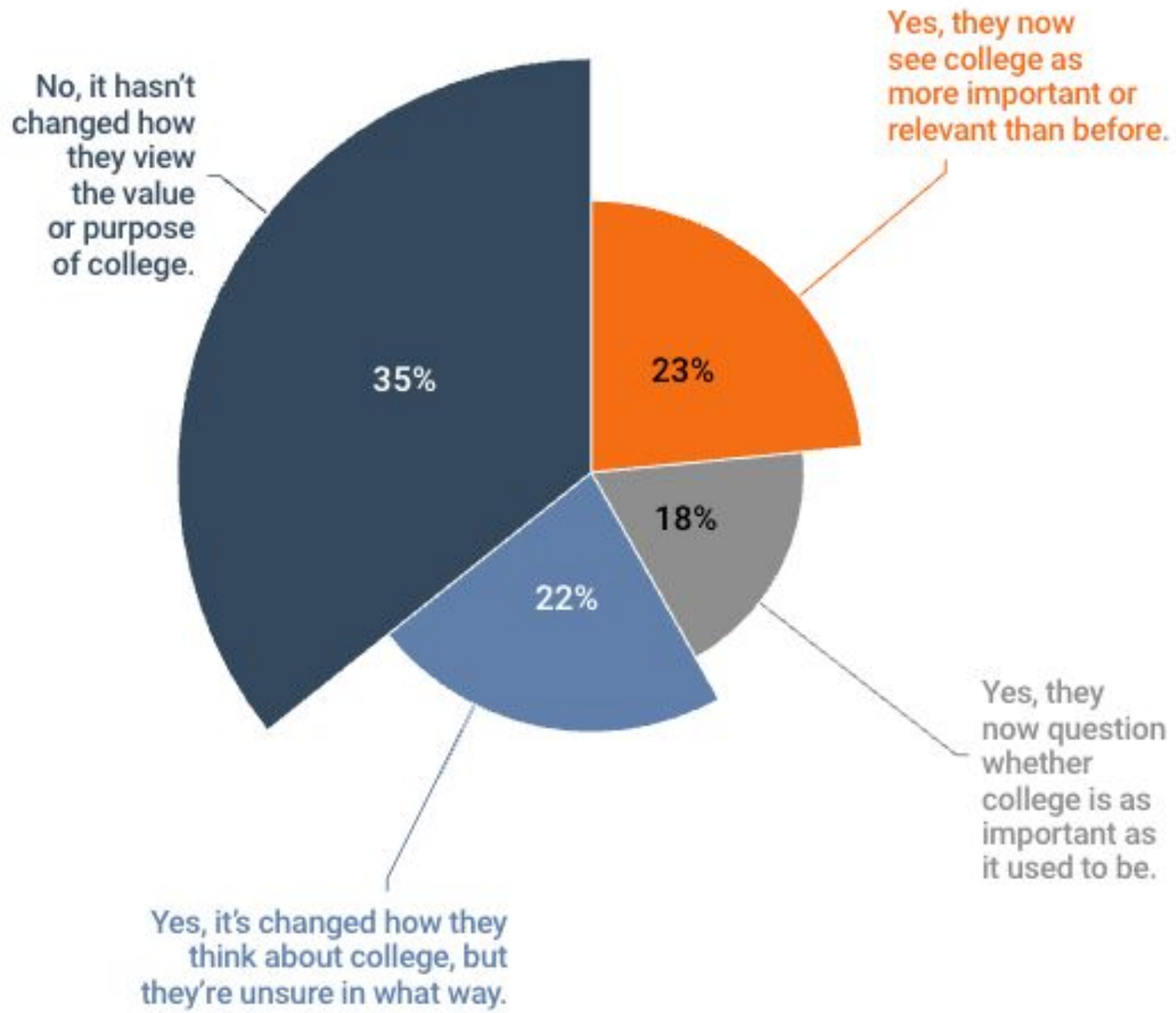
All (n=1,047) | Four-year (n=786) | Two-year (n=261)



Source: *Student Voice flash survey on AI, July 2025 • n=1,047* • Inside Higher Ed *x Generation Lab.* ***Question:*** *In which of the following ways have you used generative AI (e.g., ChatGPT) for your coursework in the past year? Select all that apply.* ***Note:*** *Up to 2% in each group chose other.*

AI and College Value

Students (all) say whether/how the rise of generative AI has changed how they think about the value or purpose of college



Source: *Student Voice flash survey on AI, July 2025 • n=1,047* • Inside Higher Ed *x Generation Lab.* ***Question:*** *Has the rise of generative AI (such as ChatGPT) influenced how you think about the value or purpose of a college education?* ***Note:*** *2% chose other.*

Please suggest a list of topics students should cover in an NLP class.

Here's a structured list of core and extended topics for a Natural Language Processing (NLP) class, moving from fundamentals to more advanced themes. You could use this as a syllabus backbone or checklist for student learning:

---

1. Foundations of NLP

- What is NLP? Scope, applications, and history
- Linguistic basics: morphology, syntax, semantics, pragmatics
- Text preprocessing: tokenization, normalization, stemming, lemmatization
- Representing text: bag-of-words, n-grams, TF-IDF

---

2. Classical NLP Methods

- Language modeling with n-grams and smoothing
- Part-of-speech tagging (Hidden Markov Models, Maximum Entropy models)
- Parsing: constituency and dependency parsing
- Named Entity Recognition (NER), chunking, and information extraction
- Sentiment analysis with traditional machine learning

---

3. Statistical and Machine Learning Approaches

ChatGPT is too traditionalist. We will cover more recent topics.

D Please suggest a list of topics students should cover in an NLP class.

Here's a comprehensive list of topics that should be covered in an NLP (Natural Language Processing) class:

Foundational Concepts

- Introduction to NLP and its applications
- Linguistic fundamentals (morphology, syntax, semantics, pragmatics)
- Text preprocessing (tokenization, normalization, stemming, lemmatization)
- Regular expressions and pattern matching
- N-grams and language modeling basics

Traditional NLP Techniques

- Part-of-speech (POS) tagging
- Named Entity Recognition (NER)
- Parsing (constituency and dependency parsing)
- Word sense disambiguation
- Information extraction
- Text classification (spam detection, sentiment analysis)
- Clustering and topic modeling (LDA, LSA)

Machine Learning for NLP

- Feature engineering for text

Claude is also too traditionalist.

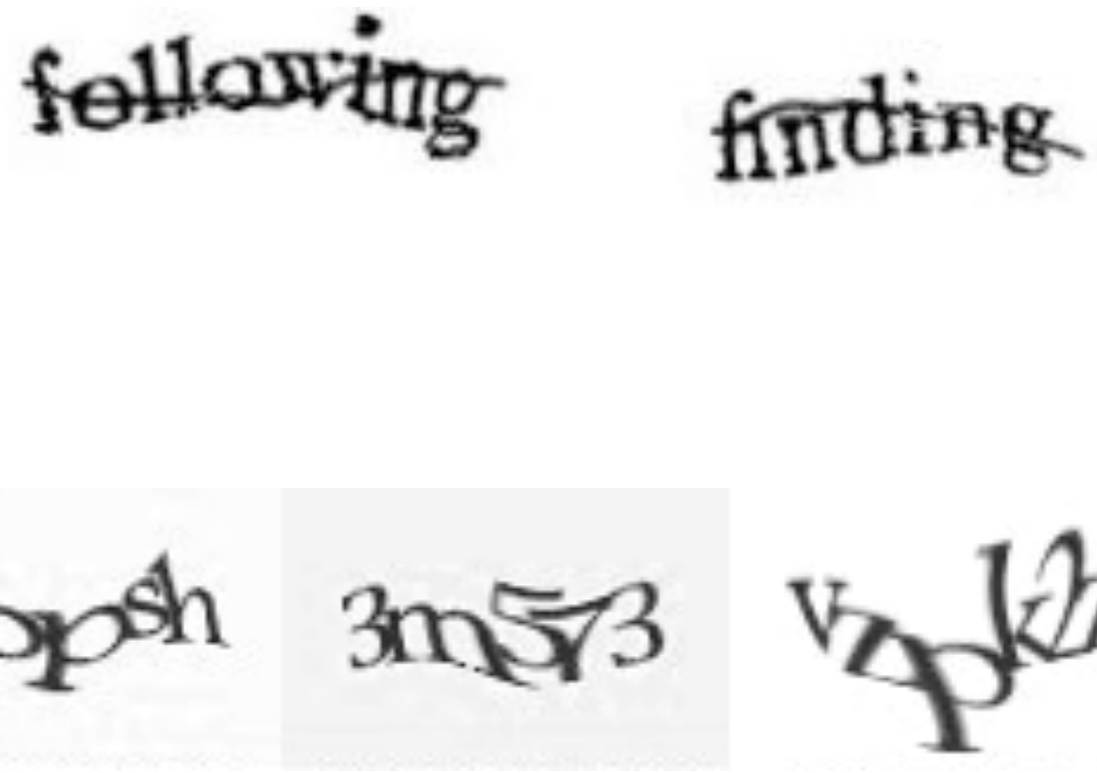
OpenAI, Altman sued over ChatGPT's role in California teen's suicide

By **Jody Godoy**

August 26, 2025 5:46 PM EDT · Updated August 26, 2025

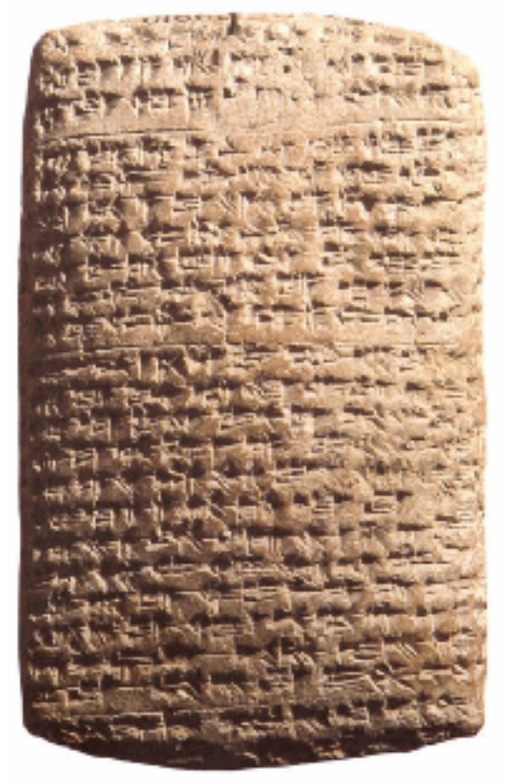
OpenAI CEO Sam Altman attends an event to pitch AI for businesses in Tokyo, Japan February 3, 2025. REUTERS/Kim Kyung-Hoon/File Photo Purchase Licensing Rights

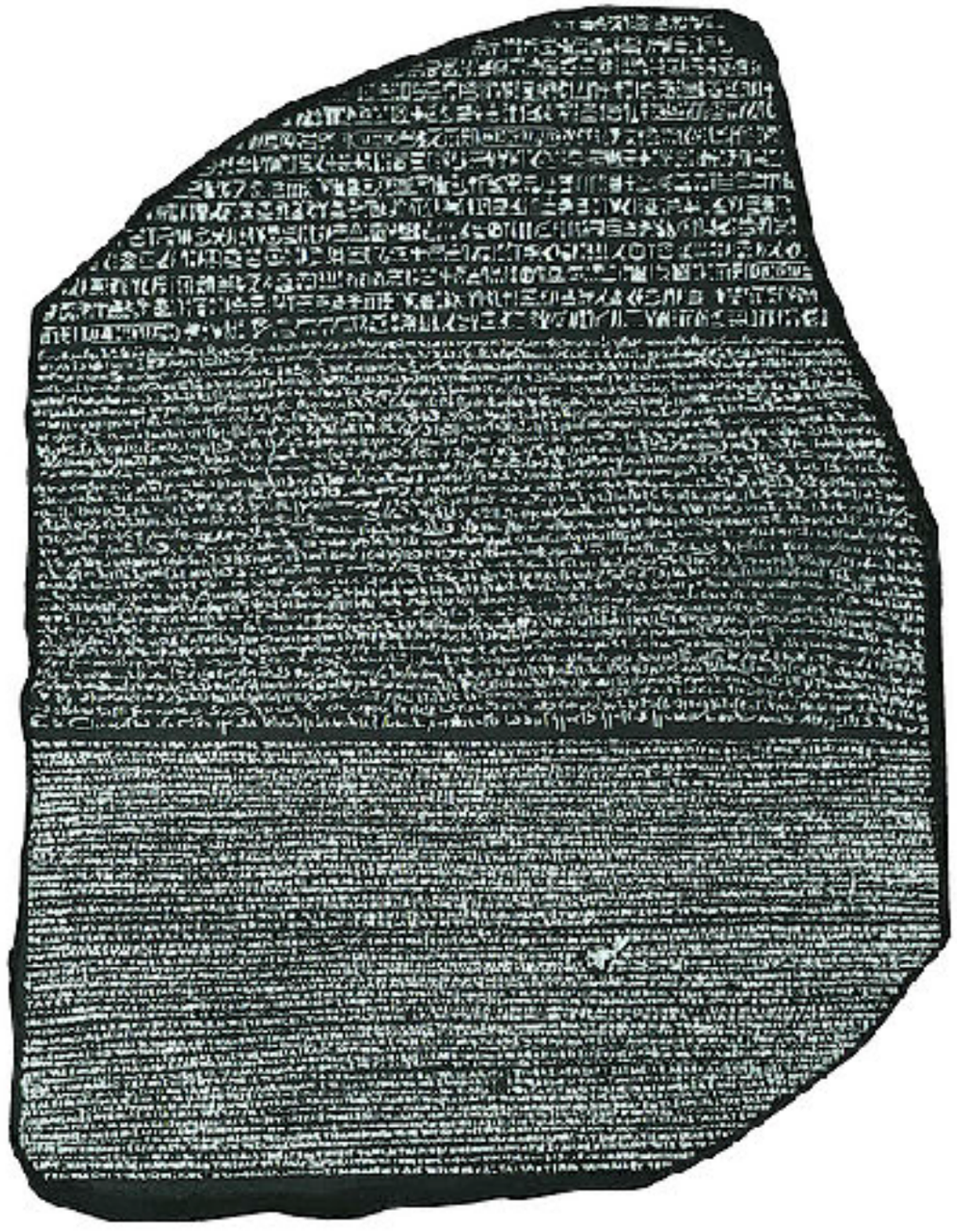
Codes



CAPTCHA

Completely Automated Public Turing test to tell Computers and Humans Apart





ON COMPUTABLE NUMBERS, WITH AN APPLICATION TO THE ENTSCHEIDUNGSPROBLEM

By A. M. Turing.

[Received 28 May, 1936.—Read 12 November, 1936.]

The "computable" numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. Although the subject of this paper is ostensibly the computable *numbers*, it is almost equally easy to define and investigate computable functions of an integral variable or a real or computable variable, computable predicates, and so forth. The fundamental problems involved are, however, the same in each case, and I have chosen the computable numbers for explicit treatment as involving the least cumbrous technique. I hope shortly to give an account of the relations of the computable numbers, functions, and so forth to one another. This will include a development of the theory of functions of a real variable expressed in terms of computable numbers. According to my definition, a number is computable if its decimal can be written down by a machine.

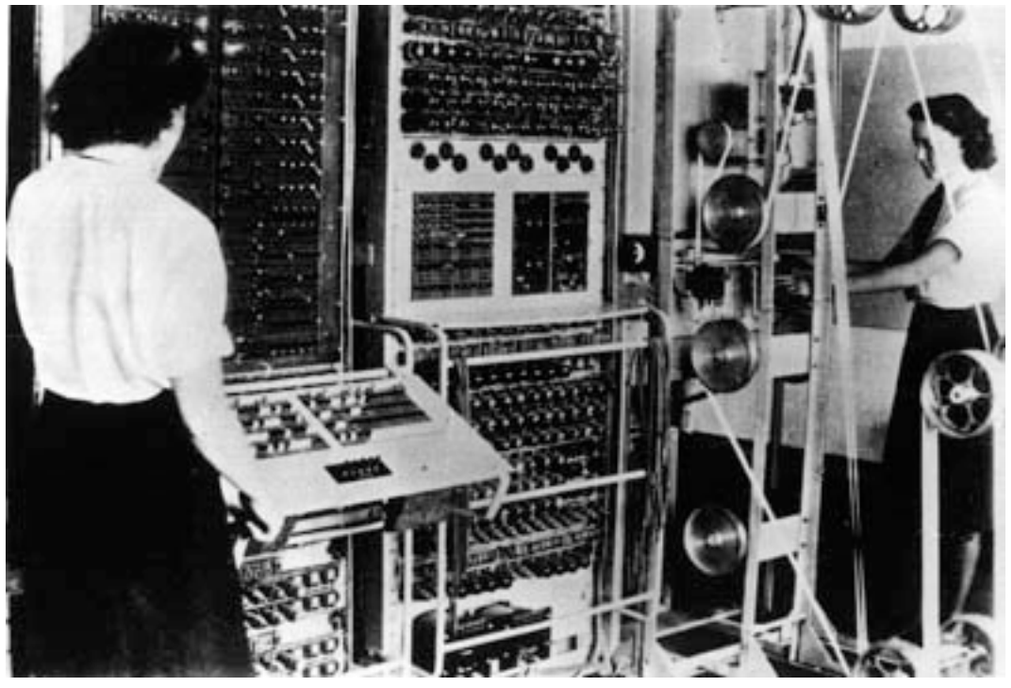
In §§ 9, 10 I give some arguments with the intention of showing that the

with the *m*-configuration written below the scanned symbol. The successive complete configurations are separated by colons.

```
  : ə ə 0  0 : ə ə 0  0 : ə ə 0  0 : ə ə 0  0    : ə ə 0  0  1 :
ɓ     ɔ       q         q         q         p
ə ə 0  0  1 : ə ə 0  0  1 : ə ə 0  0  1 : ə ə 0  0  1 :
    ƥ       ƥ         f         f
ə ə 0  0  1 : ə ə 0  0  1   : ə ə 0  0  1  0 :
        f         f         ɔ
ə ə 0  0  1 x 0 : ....
      ɔ
```

This table could also be written in the form

```
ɓ : ə ə ɔ 0  0 : ə ə q 0  0 : ...,                    (C)
```



Warren Weaver
to Norbert Wiener
4 March 1947

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: "**This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.**"

PAUL KIPARSKY

TENSE AND MOOD IN INDO-EUROPEAN SYNTAX*

1. THE HISTORICAL PRESENT

The 'historical' or 'dramatic' present tense used in narrating past events, which is common in many Indo-European languages, has always been interpreted in essentially semantic terms. A typical traditional formulation is

> it is quite mistaken to transfer it to the earlier stages of Indo-European. In Greek, Old Irish, and Old Norse, for example, the historical present has quite different syntactic and semantic properties, to which the traditional idea, or any of its variants[2], must utterly fail to do justice.

* This work was supported in part by the Joint Services Electronics Program under Contract DA36-039-AMC-03200(E); in part by the National Science Foundation (Grant GP-2495), the National Institutes of Health (Grant MH-04737-05), the National Aeronautics and Space Administration (Grant NsG-496), and the U.S. Air Force (ESD Contract AF 19 (628)-2487). – I thank Michael Connolly, Eric Hamp, Einar Haugen, George Lakoff, Calvert Watkins, and Roy Wright for offering valuable criticism and/or referring me to some of the examples cited here.

The Turing Test

Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?

Witness: It wouldn't scan.

Interrogator: How about "a winter's day," That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

Predictions

Predicting Language

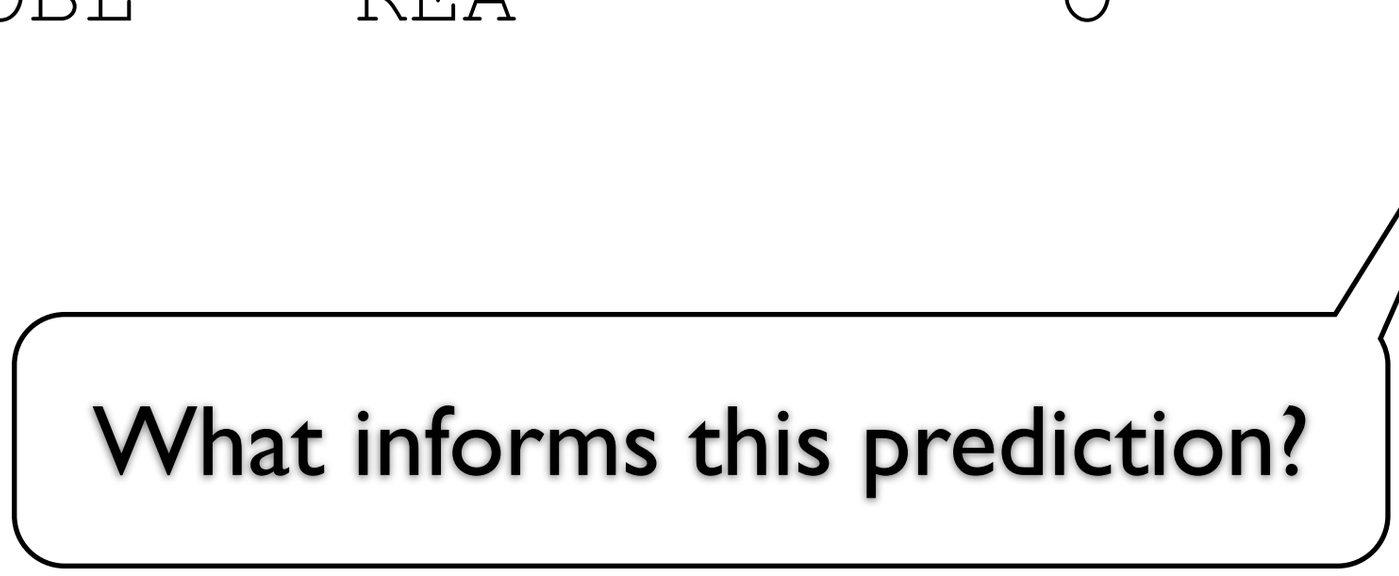
Dionyſius,* both of them Colophonians, with all the nerves and ſtrength one finds in them, appear to be too much labored, and ſmell too much of the lamp; whereas the paintings of Nicomachus† and the verſes of Homer, beſide their other excellencies and graces, ſeem to have

tense, and the glassy slags more fusible, and perhaps also more effectually decomposing the iron ore. The same quantity of fuel, applied at once to the furnace, would only prolong the duration of its heat, not augment its intensity.

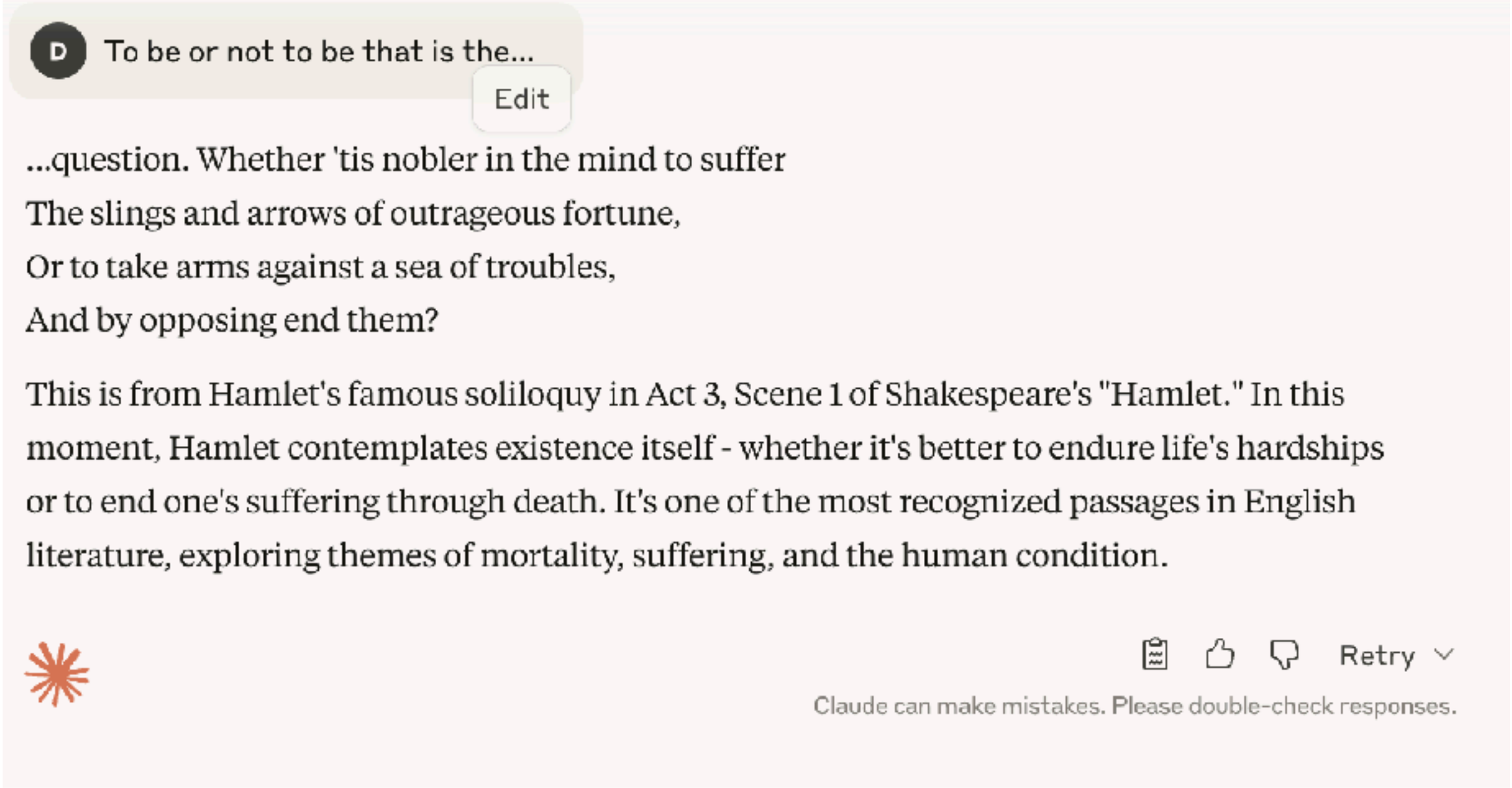
Predicting Language

```
A SMALL OBLONG READING LAMP ON THE DESK

--SM----OBL----REA----------O------D---
```



Predicting Language



Codes and Entropy

Codes Again

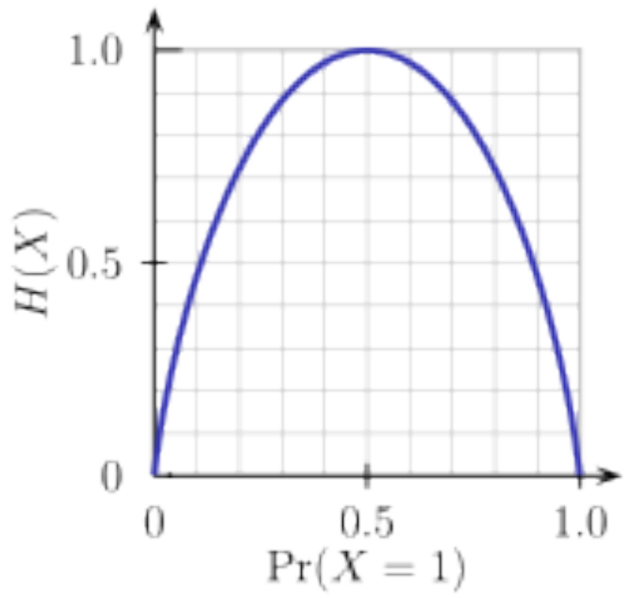
- How much *information* is conveyed in language?
- How uncertain is a classifier?
- How short of a message do we need to send to communicate given information?
- Basic idea of compression: common data elements use short codes while uncommon data elements use longer codes

Compression and Entropy

- **Entropy** measures "randomness"
  - Inverse of compressability

$$H(X) = -\sum_{i=1}^{n} p(X = x_i) \lg p(X = x_i)$$

- Lg (base 2): measured in *bits*
- Upper bound: lg *n*
- Example curve for binomial



Compression and Entropy

- Entropy bounds compression rate
  - Theorem: $H(X) \leq E[\ |\text{encoded}(X)|\ ]$
  - Recall: $H(X) \leq \lg n$
    - n is the size of the domain of X
- Standard binary encoding of integers optimizes for the worst case
- With knowledge of $p(X)$, we can do better:
- $H(X) \leq E[\ |\text{encoded}(X)|\ ] < H(X) + 1$
- Bound achieved by *Huffman codes*

Predicting Language

Claude Shannon

```
A SMALL OBLONG READING LAMP ON THE DESK

--SM----OBL----REA----------O------D---
```

Cloned Shannon makes the same guesses.

Predicting Language

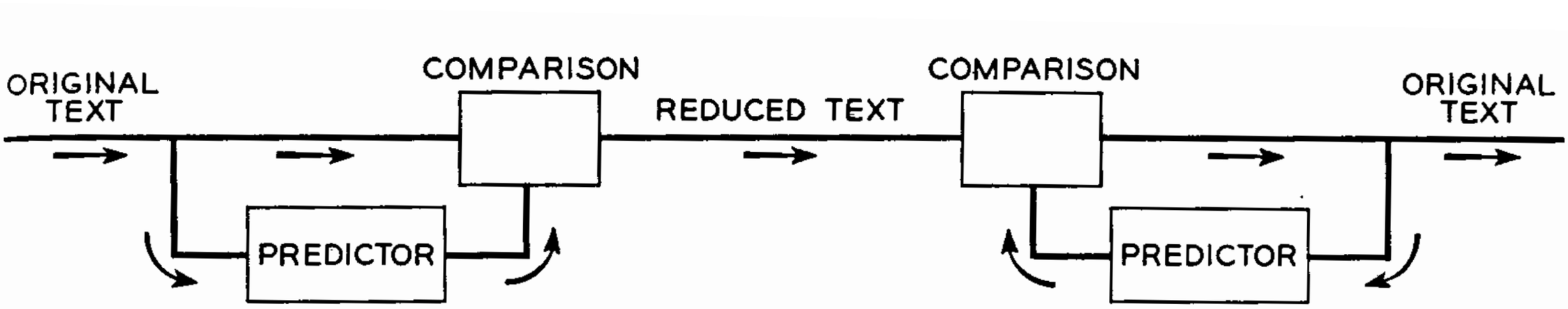


Fig. 2—Communication system using reduced text.

Claude Shannon. *Prediction and Entropy of Printed English*. 1950

Predicting Language

```
T H E R E   I S   N O   R E V E R S E   O N   A   M O T O R C Y C L E
- - - R - - I - - N - - R - V - - - E - O N - A   M - - - - C - - - -
1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1
```

http://www.ccs.neu.edu/home/dasmith/courses/cs6120/shannon/

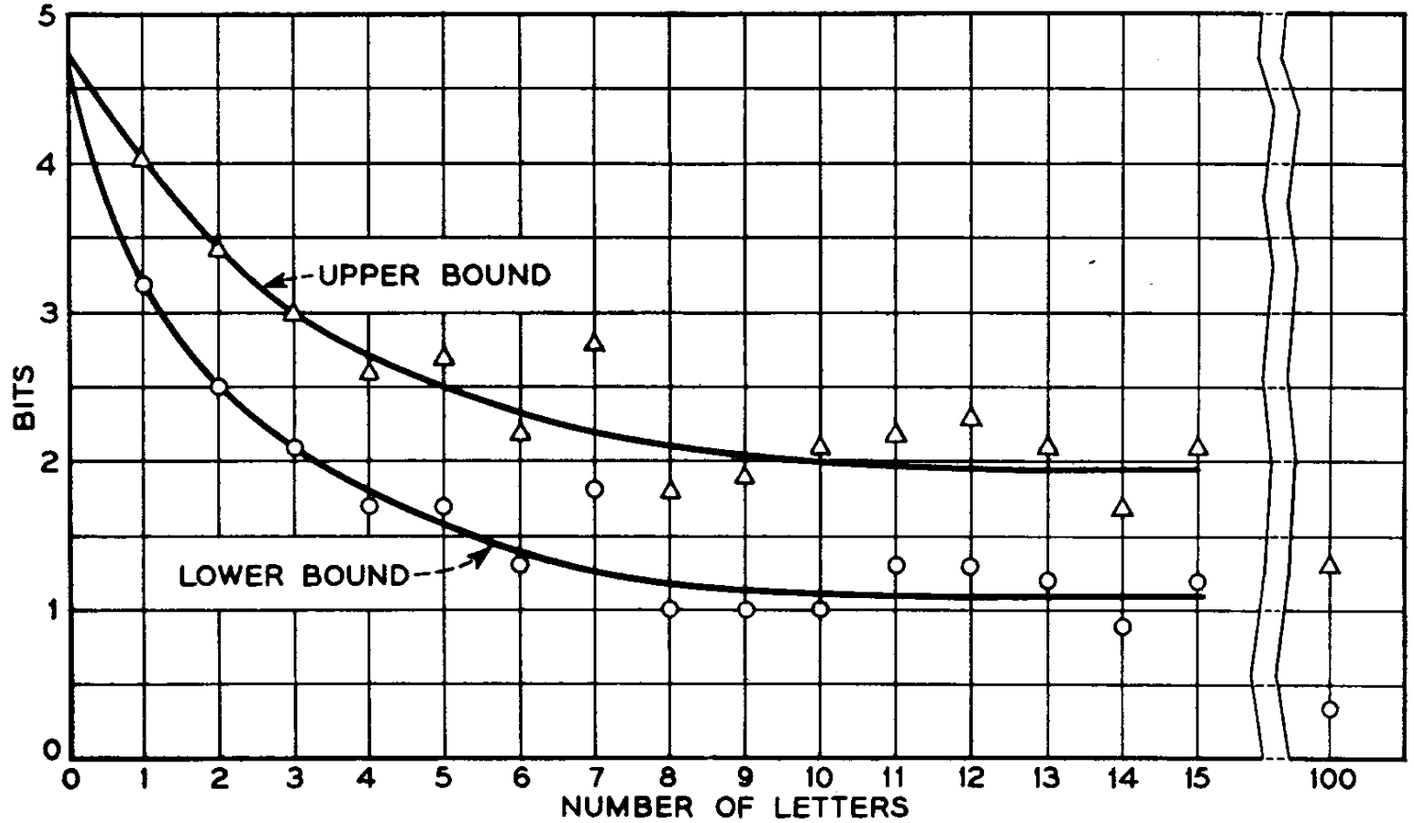


Fig. 4—Upper and lower experimental bounds for the entropy of 27-letter English.

What if we can’t ask a human?

Replace Human Predictions

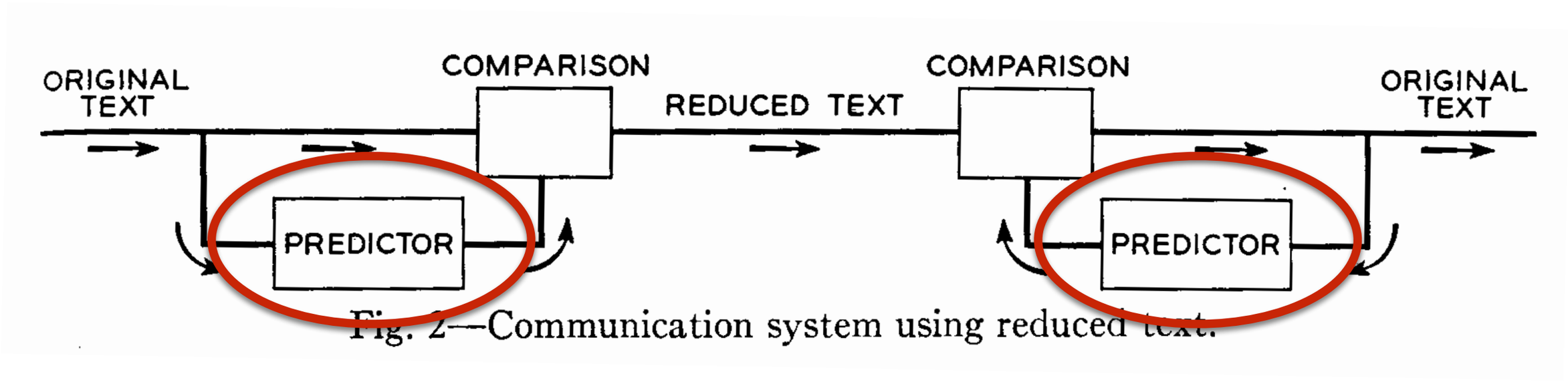


Fig. 2—Communication system using reduced text.

Claude Shannon. *Prediction and Entropy of Printed English*. 1950

A **language model**
is a function that
assigns a probability to
a string of text.

Language Models

- A simple definition
  - A **language model** is a function that assigns a probability to a string of text.
- A course full of questions
  - How do we compute those probabilities?
  - How is this function parameterized?
  - What if we only have some of the text?
  - A string of what kind of symbols?
  - What kinds of problems can we solve with LMs?

Syllabus

https://siwu.io/nlp-class/

What You'll Learn

- Markov (finite-state, n-gram) LMs
- Linear classifiers
- Word embeddings
- Neural classifiers
- Morphology, syntax, and semantics
- Sequences, Attention, and Transformer LMs

What You'll Learn

- A Taxonomy of Large Language Models
- Pretraining
- Generation
- Post-training: RLHF, DPO, and friends
- Prompting, In-context Learning
- Benchmarks and Experimental Design

What You'll Learn

- Retrieval, Retrieval-Augmented Generation
- Summarization
- Multilinguality and Translation
- Language in Social Context

Guest Lectures

- Alexander Spangher (10/7): narrative LMs
- Terra Blevins (11/4): multilingual learning
- Niloofar Mireshghallah (11/14): security & privacy
- Lucy Li (11/21): comp. social science

What You'll Do

- Five programming assignments (30%)
- Six quizzes (30%)
- Course project (40%)
- First two individually, last in groups of 1–4

Course Project

- Groups of 1–4
- Initial pitch
- Research plan
- Sample data for evaluation
- Grade contract
- Presentation
- Final report
- Feedback on each step

Course Staff

- **David Smith**: instructor for this section
- **Si Wu**: instructor for additional section
- **Divya Sri Bandaru**:TA
- **Tejus Dinesh**:TA
- *Announcing office hours soon*

Data

What if we can't ask a human?

Look at what they've already said in the past!

Digital Breadcrumbs

Slide from David Lazer

- Email
- Text messaging
- Social media
- Phone records
- Web links
- Web searches
- Smart cards
- License plates

Digital Detritus

- Social media posts
- Emails of government employees
- Supreme Court decisions
- Anti-vax message boards
- Newspaper ads for runaway slaves
- Nineteenth-century novels
- Letters by seventeenth-century scientists

There’s no data like more data.

Robert Mercer

The Roots of Big Data

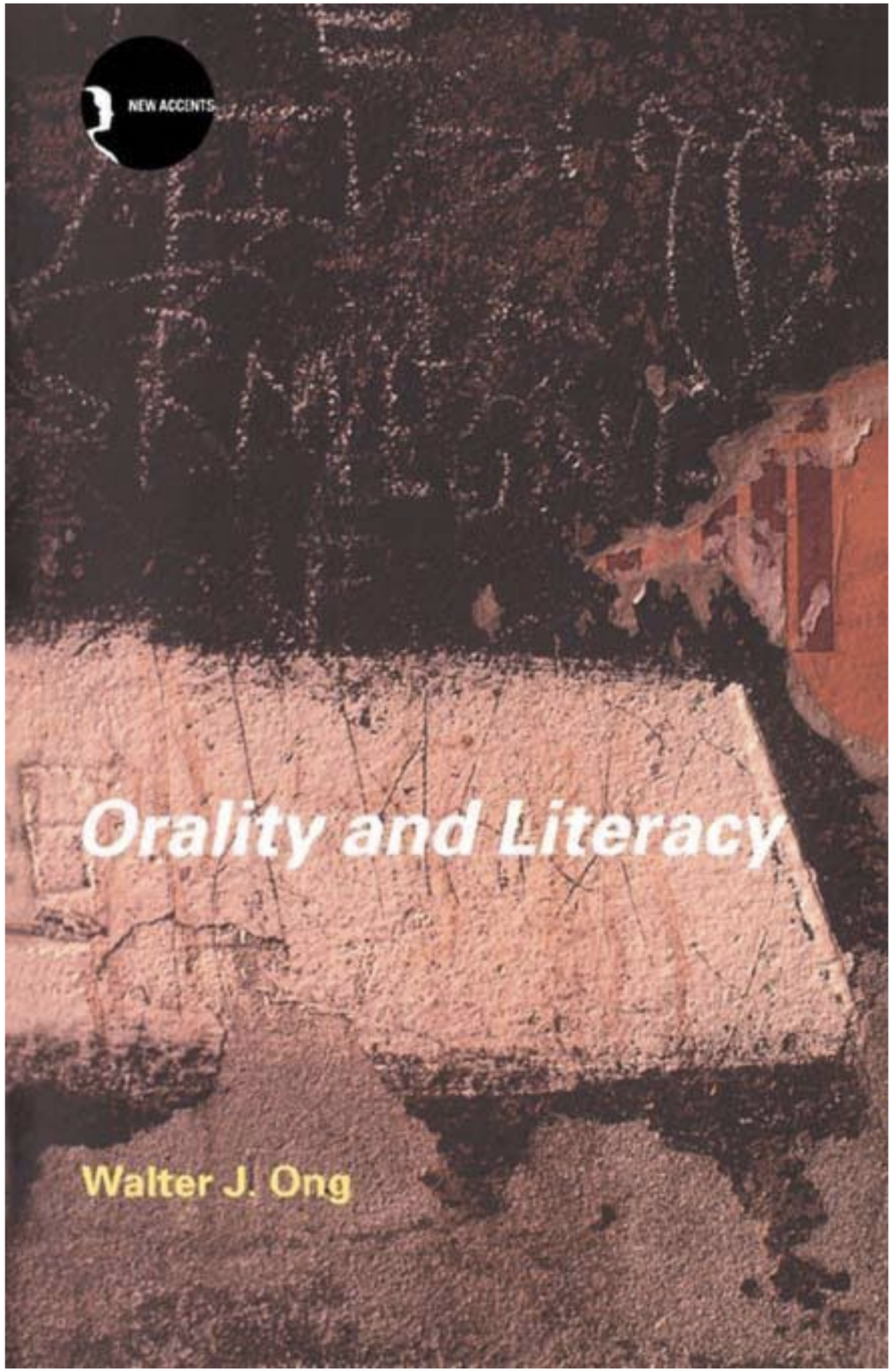
- Big Government, Business, Science
- Social change: Living online
- Digitizing the past


2007
2006
2005
2004
2003
2002
Call 800-578-1540 to Subscribe
www.real-data.com
CALL (603) 669-3822 or 800 578-1540
For more information

Metropolitan Museum of Art

Socrates: I heard, then, that at Naucratis, in Egypt, was one of the ancient gods of that country … and the name of the god himself was Theuth. He it was who invented numbers and arithmetic and geometry and astronomy, also draughts and dice, and, most important of all, letters. Now the king of all Egypt at that time was the god Thamus … To him came Theuth to show his inventions, saying that they ought to be imparted to the other Egyptians. But Thamus asked what use there was in each, and as Theuth enumerated their uses, expressed praise or blame, according as he approved or disapproved. The story goes that Thamus said many things to Theuth in praise or blame of the various arts, which it would take too long to repeat; but when they came to the letters, "This invention, O king," said Theuth, "will make the Egyptians wiser and will improve their memories; for it is an elixir of memory and wisdom that I have discovered." But Thamus replied, "Most ingenious Theuth, one man has the ability to beget arts, but the ability to judge of their usefulness or harmfulness to their users belongs to another; and now you, who are the father of letters, have been led by your affection to ascribe to them a power the opposite of that which they really possess. For this invention will produce forgetfulness in the minds of those who learn to use it, because they will not practice their memory. Their trust in writing, produced by external characters which are no part of themselves, will discourage the use of their own memory within them. You have invented an elixir not of memory, but of reminding; and you offer your pupils the appearance of wisdom, not true wisdom, for they will read many things without instruction and will therefore seem to know many things, when they are for the most part ignorant and hard to get along with, since they are not wise, but only appear wise.

Phaedrus: Socrates, you easily make up stories of Egypt or any country you please.


NEW ACCENTS
Orality and Literacy
Walter J. Ong

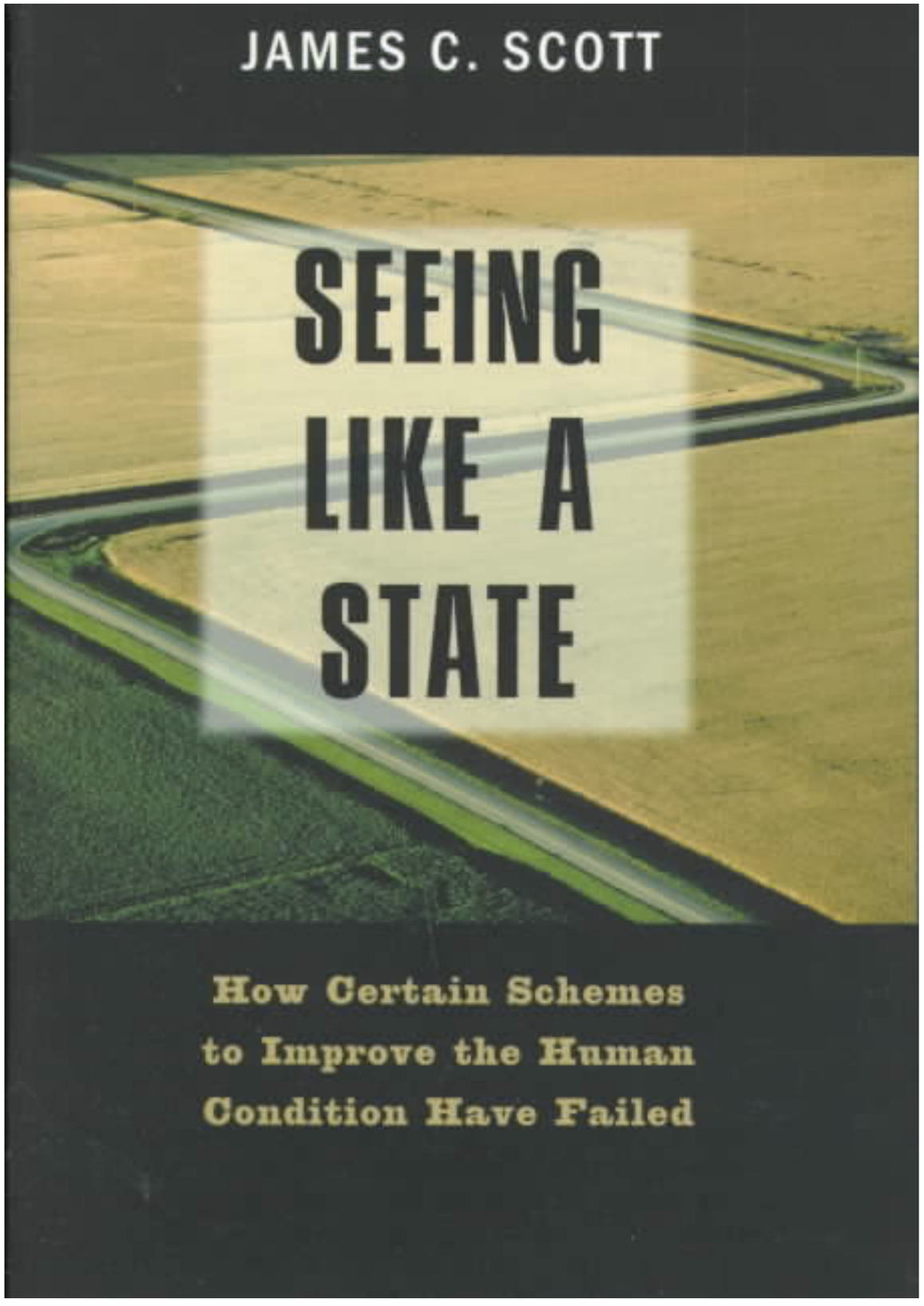
The printing press as an agent of change

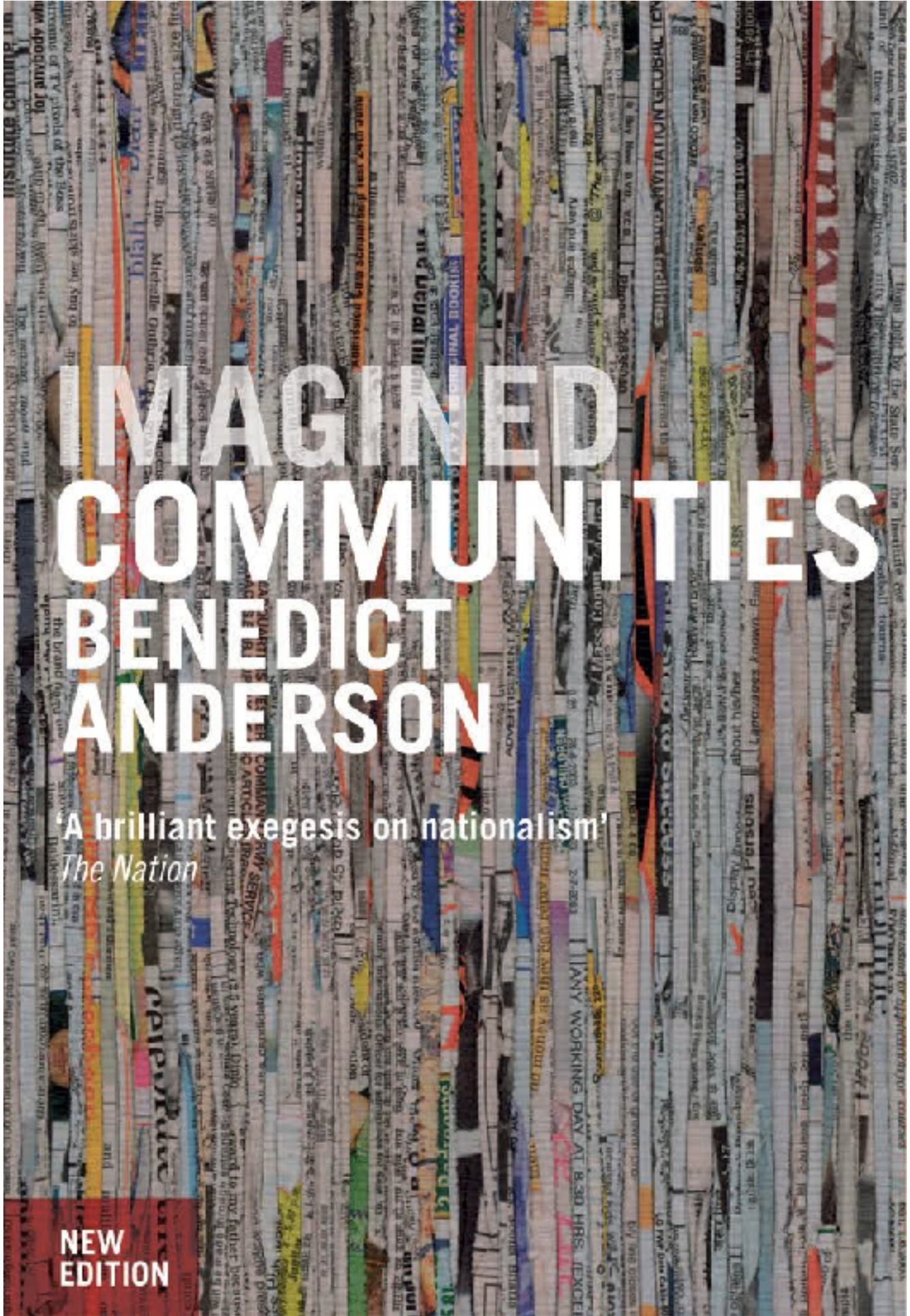
COMPLETE IN ONE VOLUME

Elizabeth L. Eisenstein

NATIONS AND NATIONALISM

ERNEST GELLNER


JAMES C. SCOTT
SEEING LIKE A STATE
How Certain Schemes to Improve the Human Condition Have Failed


IMAGINED
COMMUNITIES
BENEDICT
ANDERSON
'A brilliant exegesis on nationalism'
The Nation
NEW
EDITION

nothing but the
the history of the
at if ever we did
rties would be so
r the defeat of an
he defeated party
deserved calamity,
war in which we
ever likely to
continent more
when it began,
stance of a purely
e to have even a
for the purpose of
ppression, and not
retation of some
st as independent
n. If it be pos-
nd America, it is
e of doing so. It
s what we want.
ength is what we
etween two great
revent, or greatly
us, hopeless, and
ivil war in charac-
sies of national

, it would seem
good opportunity
more a cordial
least if we may
the other side of

events, unless the accounts from many quarters as to General Schenck's instructions are utterly belied, the new American Ambassador will bring us quite *reasonable,* though not perhaps wholly admissible demands,—demands which we certainly ought to consider most gravely, and of which we should do well to yield frankly and freely all that we should ourselves feel called upon, in the same circumstances, to press. If we do so, General Schenck's mission may make England safer and stronger than she has ever been since the close of the Civil War in 1865, and will give her a reputation for moderation and candour as well.

ENGLISH PUBLIC OPINION ON THE WAR.

Some of the philosophers should turn their attention from the subject of spectroscopic investigations and the invention of electrometers, galvanometers, hygrometers, and so forth, to the far more difficult problem of inventing a mode of measuring the intensity and diffusion of political wishes and convictions. No task at present is more difficult for a Statesman than this. There are, indeed, all sorts of shades of difference between the character of really prevalent and preponderant public opinions, of which no man, however acute, ever forms more than a purely conjectural impression, and of which, nevertheless, any respectably-accurate measure would be a matter of the highest political importance. For instance, there is at times a public opinion on one side of a question which is very widely diffused, but of very slight intensity,—which, in fact, amounts to nothing more than a wish in a particular direction without a will, and still more without any intention of submitting to a considerable sacrifice rather than not carry out the will into action. Again, there is such a thing as

Large AI models are cultural and social technologies

Implications draw on the history of transformative information systems from the past

HENRY FARRELL, ALISON GOPNIK, COSMA SHALIZI, AND JAMES EVANS Authors Info & Affiliations

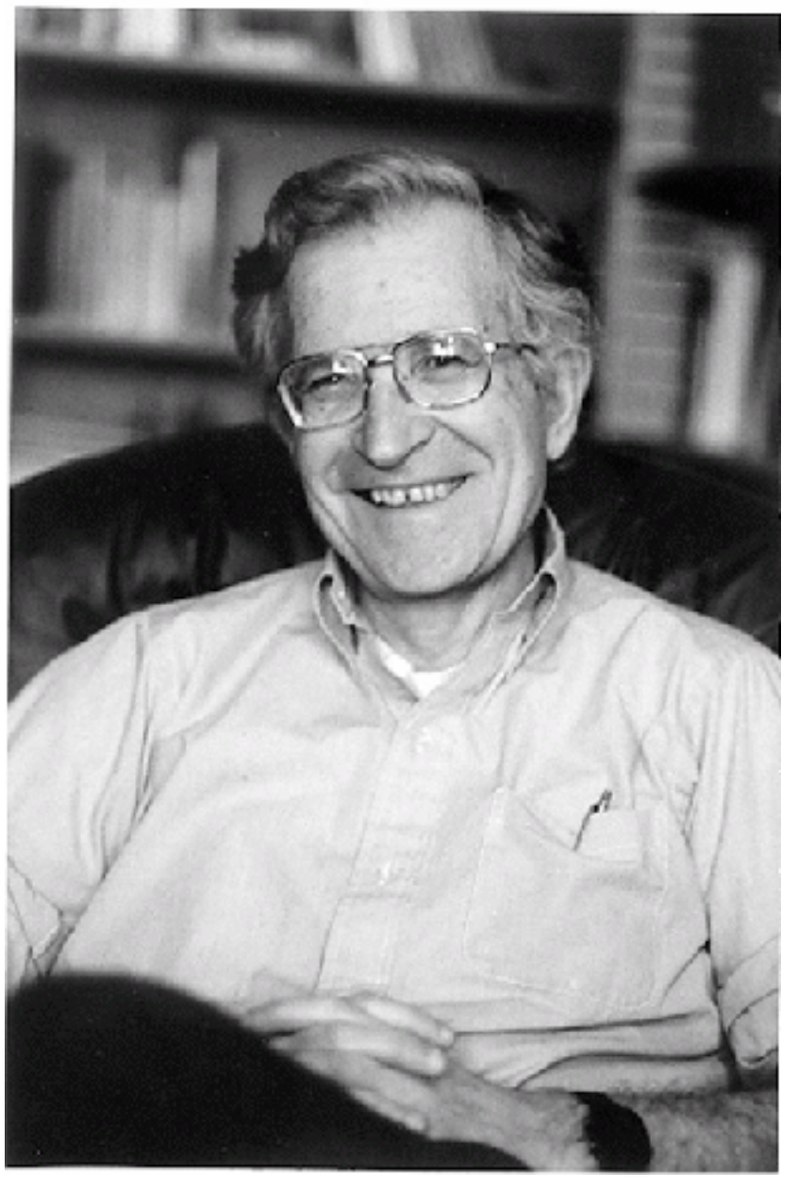
SCIENCE • 13 Mar 2025 • Vol 387, Issue 6739 • pp. 1153-1156 • DOI: 10.1126/science.adt9819

Abstract

Debates about artificial intelligence (AI) tend to revolve around whether large models are intelligent, autonomous agents. Some AI researchers and commentators speculate that we are on the cusp of creating agents with artificial general intelligence (AGI), a prospect anticipated with both elation and anxiety. There have also been extensive conversations about cultural and social consequences of large models, orbiting around two foci: immediate effects of these systems as they are currently used, and hypothetical futures when these systems turn into AGI agents—perhaps even superintelligent AGI agents. But this discourse about large models as intelligent agents is fundamentally misconceived. Combining ideas from social and behavioral sciences with computer science can help us to understand AI systems more accurately. Large models should not be viewed primarily as intelligent agents but as a new kind of cultural and social technology, allowing humans to take advantage of information other humans have accumulated.

Linguistic Description

Noam Chomsky



- Anarcho-syndicalist polemicist
- Inventor of several theories of "generative grammar" opposed to language models
- **Pioneer of formal language theory**

Linguistic Modules

- Phonetics and phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse
- *With lots of crossings between levels!*

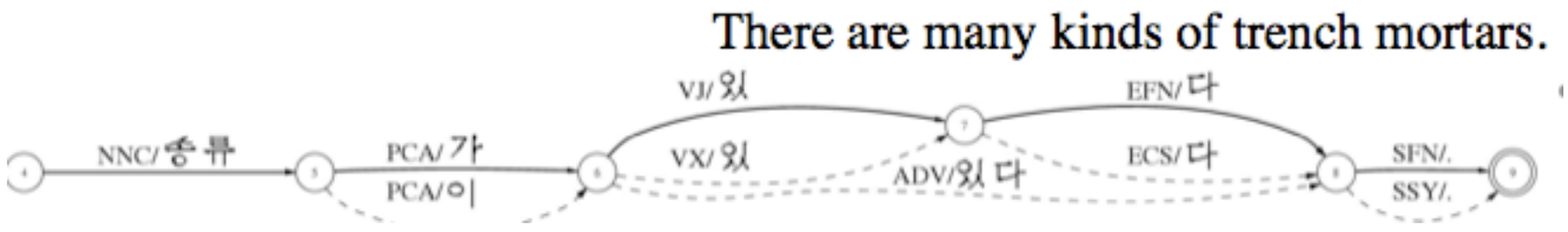
Phonetics and Phonology

- Phonetics: language sounds & their physiology
- Phonology: systems of discrete sounds in languages
  - E.g.: devoicing of *it is* to *it's*
  - E.g.: syllable structure: *sign, signify*

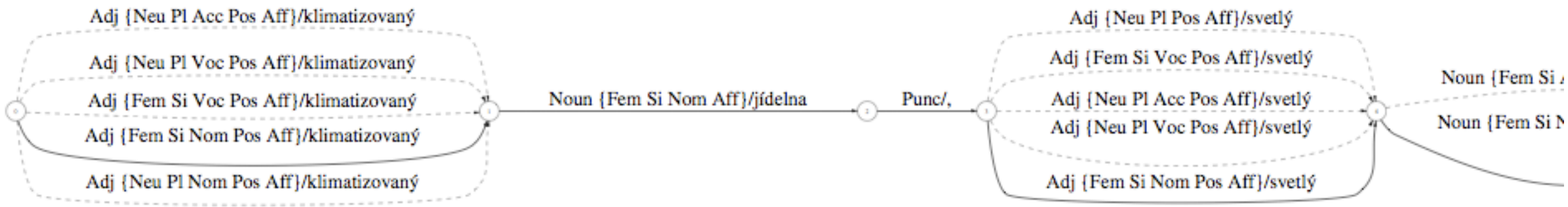
Morphology

- Inflectional (in some languages):
  - *love* →*loved*
- Derivational:
  - *tea-cup, un-helpful, with-stand, craisin*
- Turkish: *uygarlastiramadiklarimizdanmissinizcasina*
  - *uygar las tir ama dik lar imiz dan mis siniz casina*
  - *(behaving) as if you are among those whom we could not civilize*

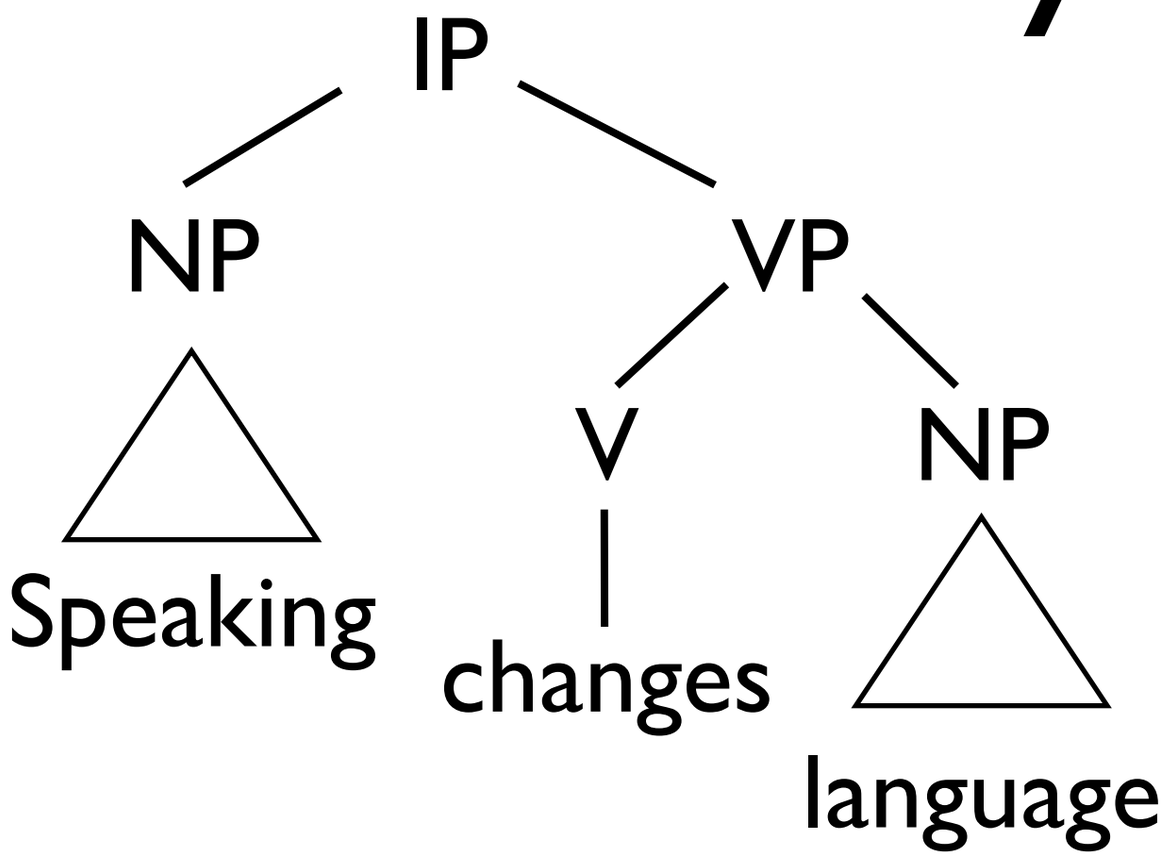
Morphological Tagging



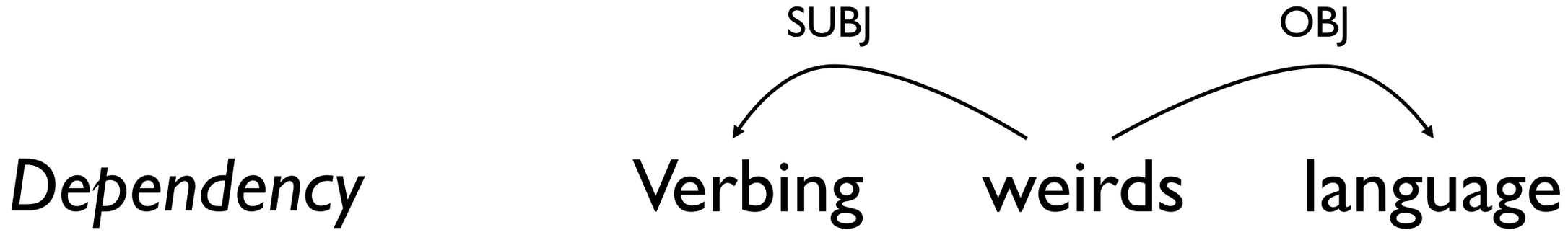
c. Klimatizovaná jídelna, světlá místnost pro snídaně. Air-conditioned dining room,



Syntax

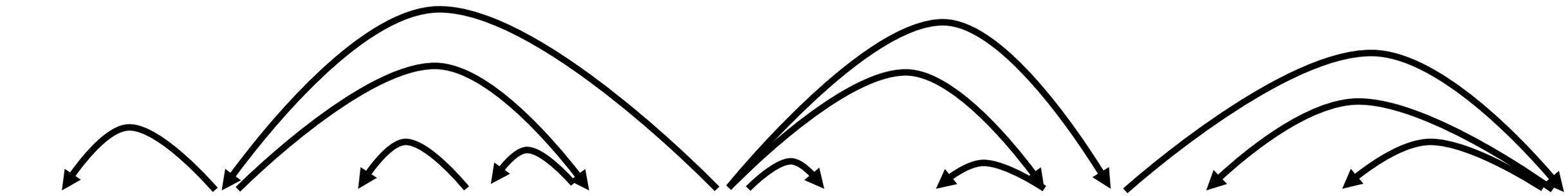


Constituency



Dependency

Semantics



Pierre Vinken , 61 years old , will join the board as a nonexecutive director

PropBank **join** predicate

ARG0	ARG1	ARG-PRD
Vinken	board	director

Pragmatics

- Context affects meaning
- Conversational implicature
  - *Is your mother at home? Yes.*
- Speech acts: "how to do things with words"
  - *I grant you permission to speak.*

Discourse

- Study of units larger than a single utterance
- Turn taking
- Coreference
- Organized exposition

To-Do List

- Check course website:
  - https://siwu.io/nlp-class/
- Click on links to:
  - Join Gradescope for homework
  - Join Ed for discussion

Thanks