

# Computational Linguistics

## 1. Introduction

Xiaojing Bai

Tsinghua University

<https://bxjthu.github.io/CompLing>

That the powerful play goes on,  
and you may contribute a verse.



That the powerful play goes on,  
and you may contribute a verse.

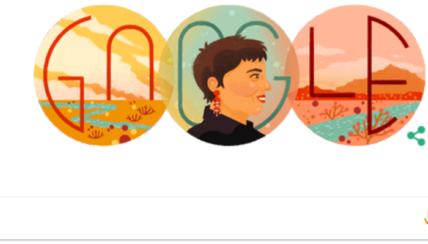
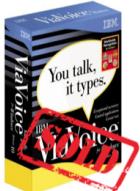
**What will your verse be?**

[A collection of readings on AI](#)

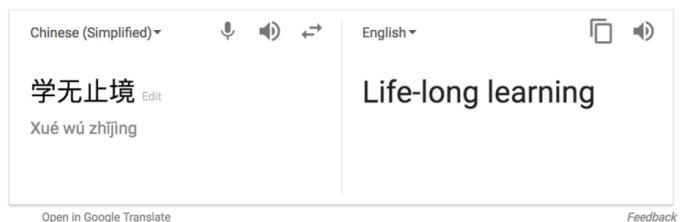
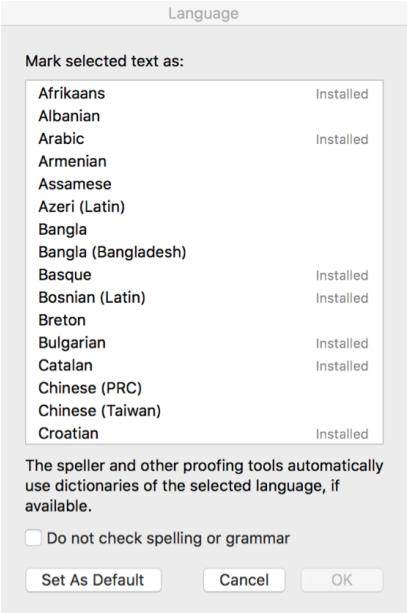
# At the end of this session you will

- understand what computational linguistic is in the general sense;
- understand what this course is for and what it focuses on;
- have some preliminary but important ideas about computational linguistic;
- get started with Python, NLTK, and PyCharm;
- learn basic data types and sequence operations in Python.

Giving computers the ability  
to process human language



Hans Rosling: I'm going to ask you three multiple choice questions. Use this device. Use this device to answer. The first question is, how did the number of deaths per year from natural disaster, how did that change during the last century? Did it more than double, did it remain about the same in the world as a whole, or did it decrease to less than half? Please answer A, B or C. I see lots of answers.



Giving computers the ability

to process human language

A vibrant interdisciplinary field  
with many names corresponding to its many facets

Overlapping fields in different departments

# What is *Computational Linguistics*?

Menu

---

- About the ACL ▶
- News ▶
- Journals ▶
- Conferences ▶
- Events ▶
- ACL Fellows ▶
- SIGs
- Anthology ▶
- Wiki
- Software Registry
- Education ▶
- Policies ▶
- Archives ▶



Association for  
Computational Linguistics

The premier international scientific and professional society for people working on computational problems involving human language

Founded in 1962 and originally named the Association for Machine Translation and Computational Linguistics

For more information: <https://www.aclweb.org/>

# What is *Computational Linguistics*?

Computational linguistics is the scientific study of **language** from a computational perspective.

Computational linguists are interested in providing **computational models** of various kinds of linguistic phenomena. These models may be "**knowledge-based**" ("hand-crafted") or "**data-driven**" ("statistical" or "empirical").

Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide **a computational explanation for a particular linguistic or psycholinguistic phenomenon**; and in other cases the motivation may be more purely technological in that one wants to provide **a working component of a speech or natural language system**.

# What is *Computational Linguistics*?

Indeed, the work of computational linguists is incorporated into many **working systems** today, including speech recognition systems, text-to-speech synthesizers, automated voice response systems, web search engines, text editors, language instruction materials, to name just a few.

*“Linguistics has a hundred-year history as a scientific discipline, and computational linguistics has a fifty-year history as a part of computer science. But it is only in the last decade or so that language understanding has emerged as an industry reaching millions of people, with information retrieval and machine translation available on the Internet, and speech recognition becoming popular on desktop computers.”*

Peter Norvig and Stuart Russell. *Foreword*. J+M 2nd

# Fire the *linguists*?

*“Every time I fire a linguist,  
the performance of the speech recognizer goes up.”*

Frederick Jelinek (1932 – 2010)

## Rationalism vs. Empiricism (rule-based vs. statistics-based)

### Suggested readings:

- Church, K. (2011). A pendulum swung too far. *Linguistic Issues in Language Technology*, 6(5), 1-27.
- Wintner, S. (2009). What science underlies natural language engineering?. *Computational Linguistics*, 35(4), 641-644.
- 宗成庆 (2008). 《统计自然语言处理》. 北京: 清华大学出版社.
- 冯志伟 (2008). 序言. 《统计自然语言处理》 (宗成庆著). 北京: 清华大学出版社.

## About this course

- Overview
- Aims
- Prerequisite
- Teaching
- Readings
- Grading
- Tentative schedule

# Preliminary but important ...

- Some brief history
- Knowledge in natural language processing
- Complexity of language
- Models and algorithms

Required reading: [J+M 1](#)

# Some brief history - a sketch

- 1940s – 1950s: foundational insights
- 1957 – 1970: the two camps
- 1970 – 1983: four paradigms
- 1983 – 1993: empiricism and finite-state models redux
- 1994 – 1999: the field comes together
- 2000 – 2008: the rise of machine learning

# Some brief history - old events

- 1949 Warren Weaver's memorandum
- 1950 Turing Test
- 1952 First conference on MT at MIT
- 1954 First MT demo by Georgetown University and IBM
- 1954 *Mechanical Translation* >>> 1980 *Computational Linguistics*
- 1957 Chomsky's *Syntactic Structures*
- 1962 The Association for Machine Translation and *Computational Linguistics*
- 1966 The (in)famous ALPAC report

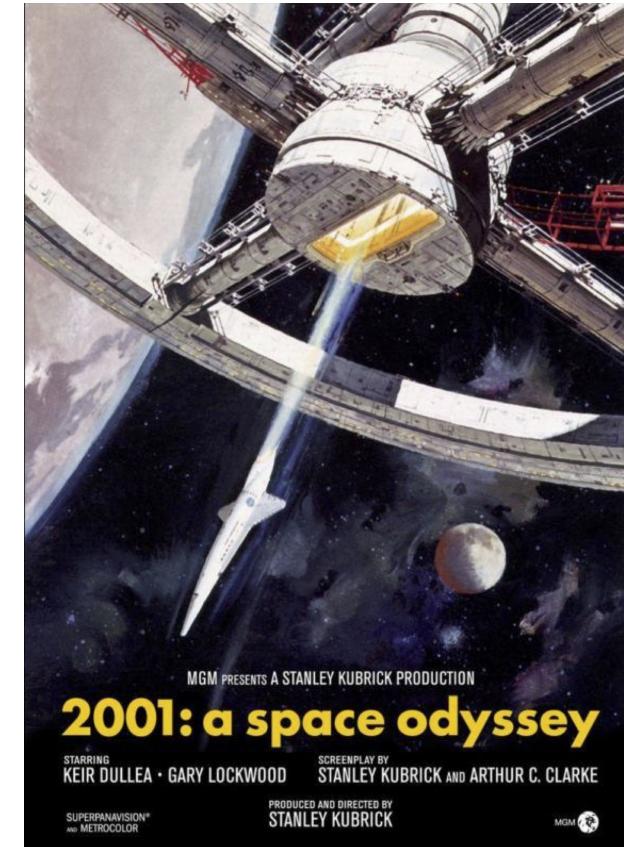
# Knowledge in natural language processing

Mankind finds a mysterious, obviously artificial, monolith buried on the moon and, with the intelligent computer HAL, sets off on a quest...

*Dave Bowman: Open the pod bay doors, HAL.*

*HAL: I'm sorry Dave, I'm afraid I can't do that.*

Stanley Kubrick and Arthur C. Clarke:  
*2001: A Space Odyssey*



# Knowledge in natural language processing

What would HAL need to know about language?

# Knowledge in natural language processing

## Homework 0: Task 1

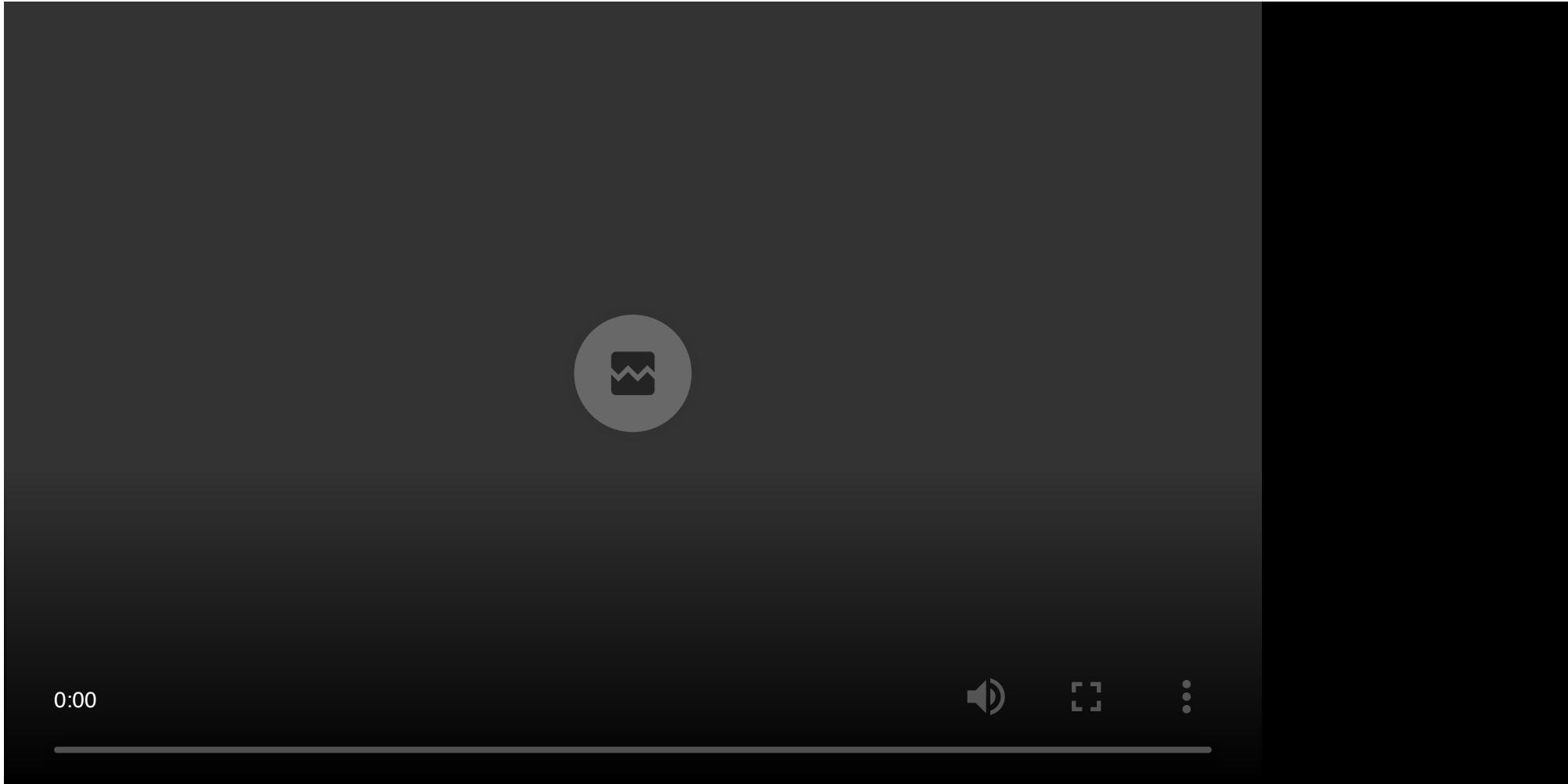
Question 1: With an electronic application in your daily life, which has the ability to process human language, analyze what it would need to know about language to perform the particular NLP task(s)?

# Knowledge in natural language processing

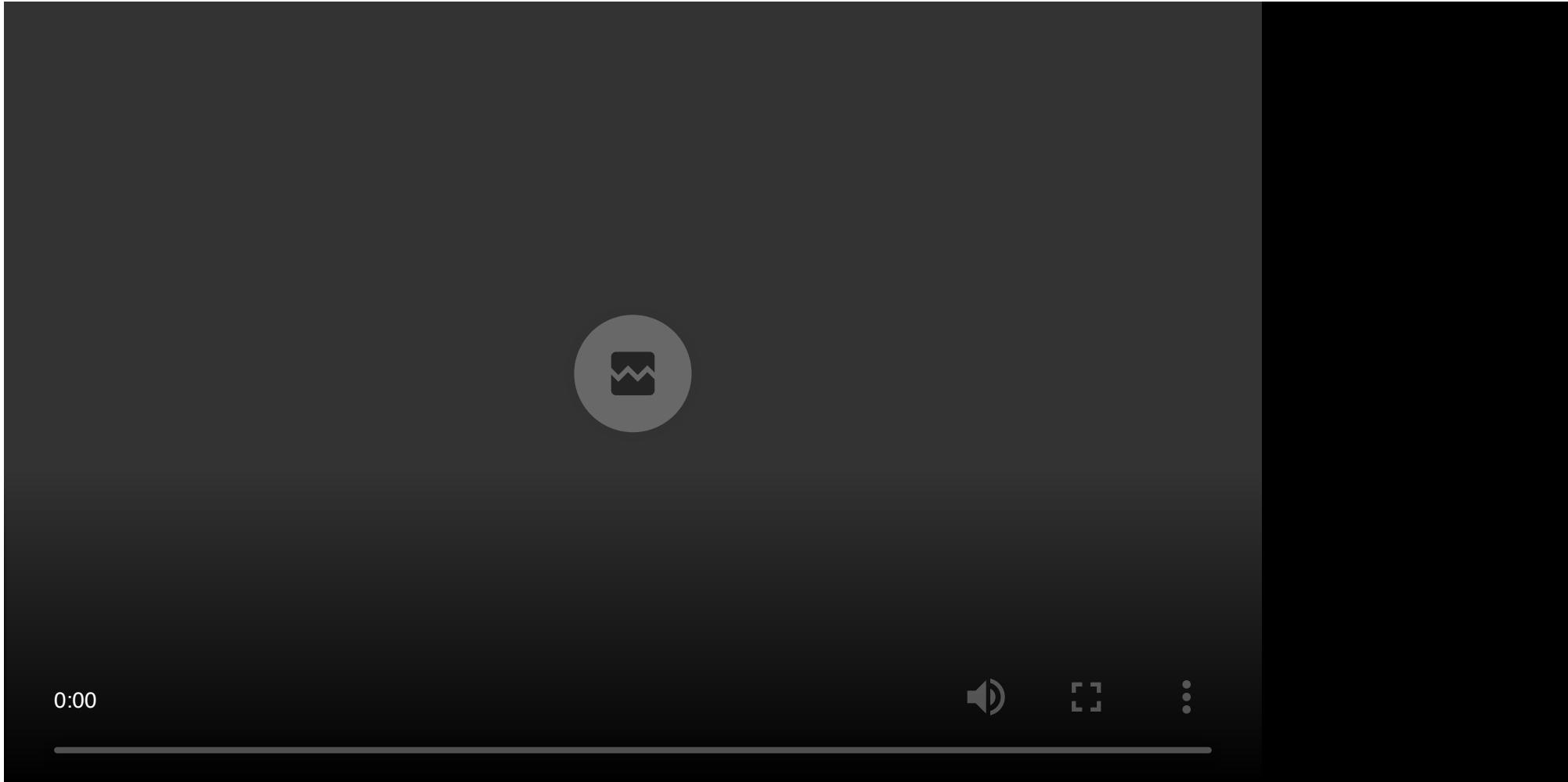
What would HAL need to know about language?

- **Phonetics and phonology:** knowledge about linguistic sounds
- **Morphology:** knowledge of the meaningful components of words
- **Syntax:** knowledge of the structural relationships between words
- **Semantics:** knowledge of meaning
- **Pragmatics:** knowledge of the relationship of meaning to the goals and intentions of the speaker
- **Discourse:** knowledge about linguistic units larger than a single utterance

# Knowledge in natural language processing



# Knowledge in natural language processing



# Complexity of language

Resolving ambiguity at different levels

e.g. *I made her duck.*

# Complexity of language

## Resolving ambiguity at different levels

e.g. *I made her duck.*

- I cooked waterfowl for her.
- I cooked waterfowl belonging to her.
- I created the (plaster?) duck she owns.
- I caused her to quickly lower her head or body.
- I waved my magic wand and turned her into undifferentiated waterfowl.

# Complexity of language

## Difficulties in Chinese information processing 中文信息处理

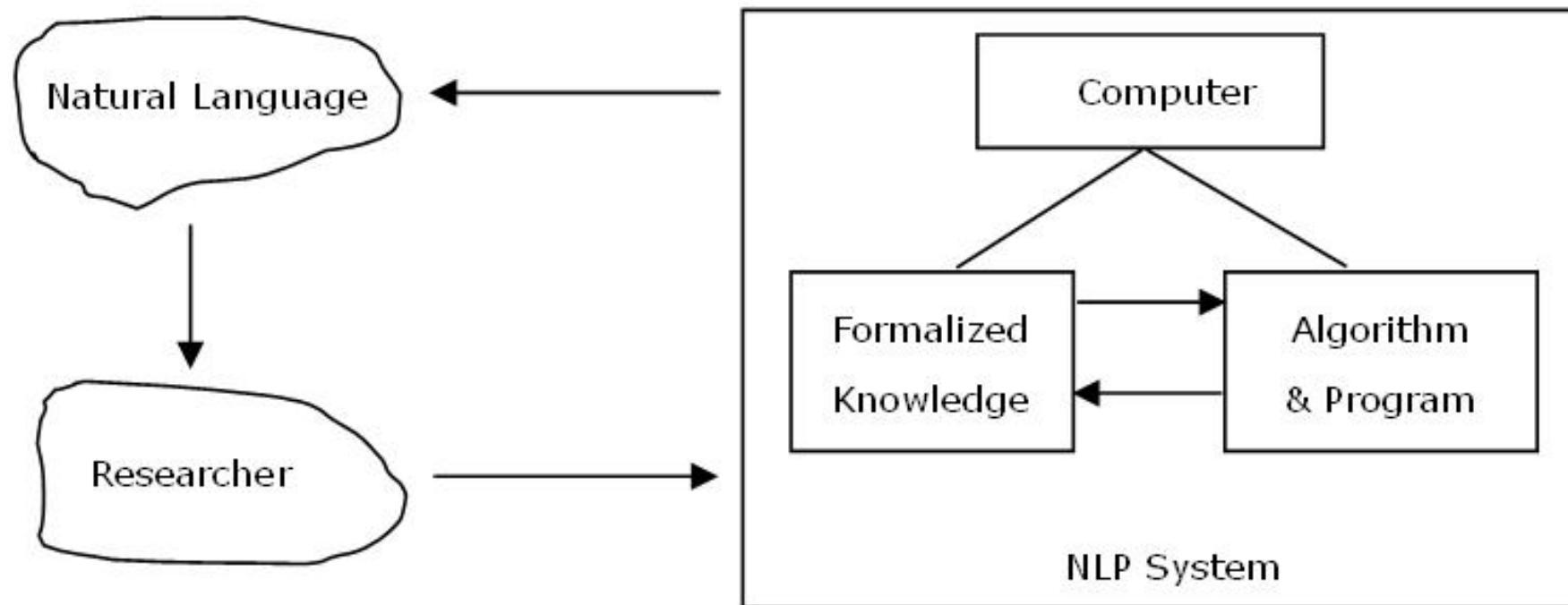
Suggested readings:

俞士汶等: [汉语自动分析的特殊困难与对策](#)

詹卫东: [现代汉语歧义现象举例](#)

# Models and algorithms

How do NLP systems work?



# Models

- A model of an **object** is a physical representation that shows what it looks like or how it works. The model is often smaller than the object it represents.
- A model of a **system or process** is a theoretical description that can help you understand how the system or process works, or how it might work.
- Language model: a **formal** description of linguistic knowledge
- Important models: state machines, rule systems, logic, probabilistic models, vector-space models

# Models - rule systems

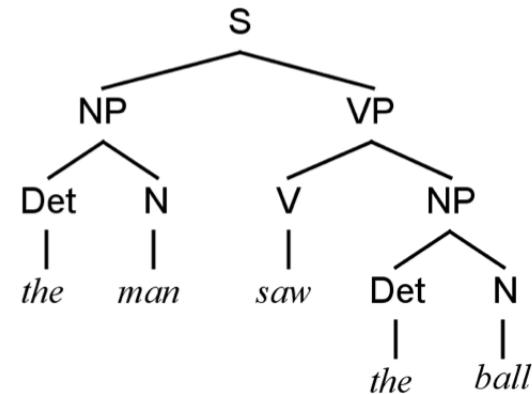
$$G = \{V_n, V_t, S, P\}$$

$$V_n = \{S, NP, VP, Det, N, V\}$$

$$V_t = \{\textit{the}, \textit{man}, \textit{boy}, \textit{ball}, \textit{saw}, \textit{hit}, \textit{took}\}$$

$$P = \left\{ \begin{array}{l} S \rightarrow NP\ VP \\ NP \rightarrow Det\ N \\ VP \rightarrow V\ NP \\ Det \rightarrow \textit{the} \\ N \rightarrow \textit{man} \mid \textit{boy} \mid \textit{ball} \\ V \rightarrow \textit{saw} \mid \textit{hit} \mid \textit{took} \end{array} \right\}$$

- ✓ the man saw the ball
- ✓ the boy took the ball
- ✗ man the ball
- ✗ the saw hit the man



# Models - rule systems and probabilistic models

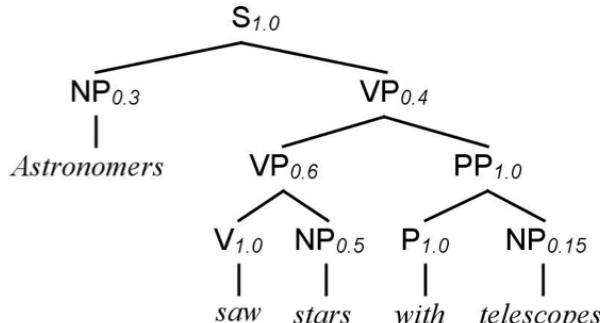
$$G = \{V_n, V_t, S, P\}$$

$$V_n = \{S, NP, VP, Det, N, V\}$$

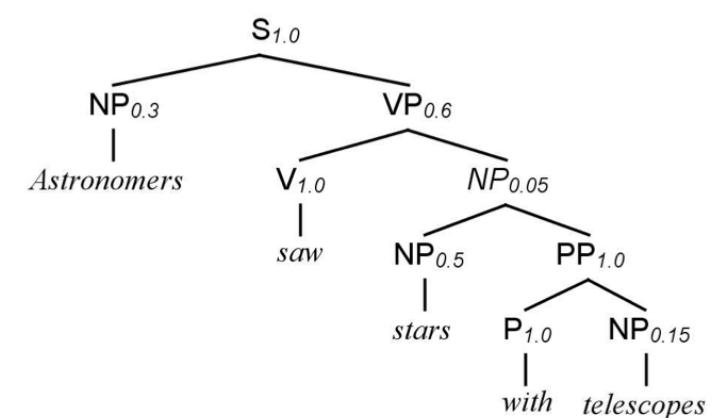
$$V_t = \{\text{the, man, boy, ball, saw, hit, took}\}$$

$$P = \left\{ \begin{array}{l} S \rightarrow NP\ VP \\ NP \rightarrow Det\ N \\ VP \rightarrow V\ NP \\ Det \rightarrow \text{the} \\ N \rightarrow man \mid boy \mid ball \\ V \rightarrow saw \mid hit \mid took \end{array} \right\}$$

$$\begin{array}{ll} S \rightarrow NP\ VP & 1.0 \\ VP \rightarrow VP\ PP & 0.4 \\ VP \rightarrow V\ NP & 0.6 \\ PP \rightarrow P\ NP & 1.0 \\ NP \rightarrow NP\ PP & 0.05 \end{array}$$



$$\begin{array}{ll} P \rightarrow with & 1.0 \\ V \rightarrow saw & 1.0 \\ NP \rightarrow astronomers & 0.3 \\ NP \rightarrow stars & 0.5 \\ NP \rightarrow telescopes & 0.15 \end{array}$$



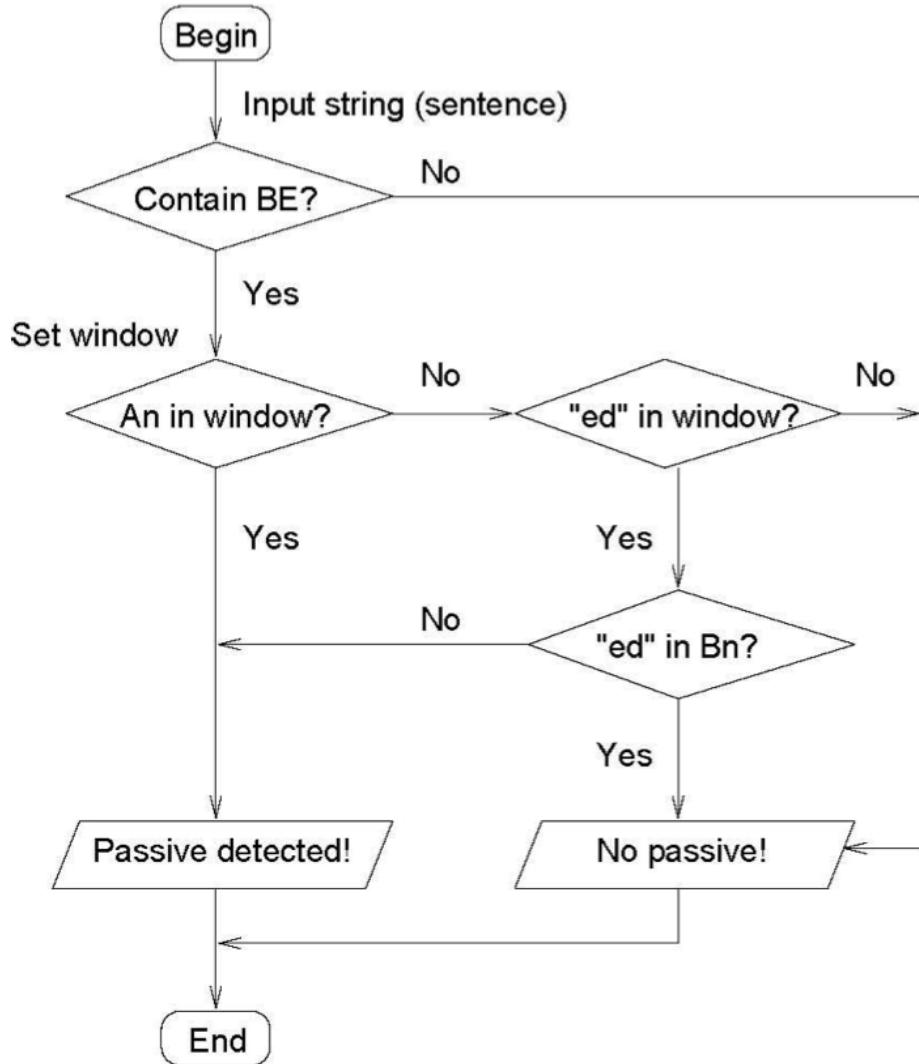
# Algorithms

## Homework 0: Task 1

Question 2:

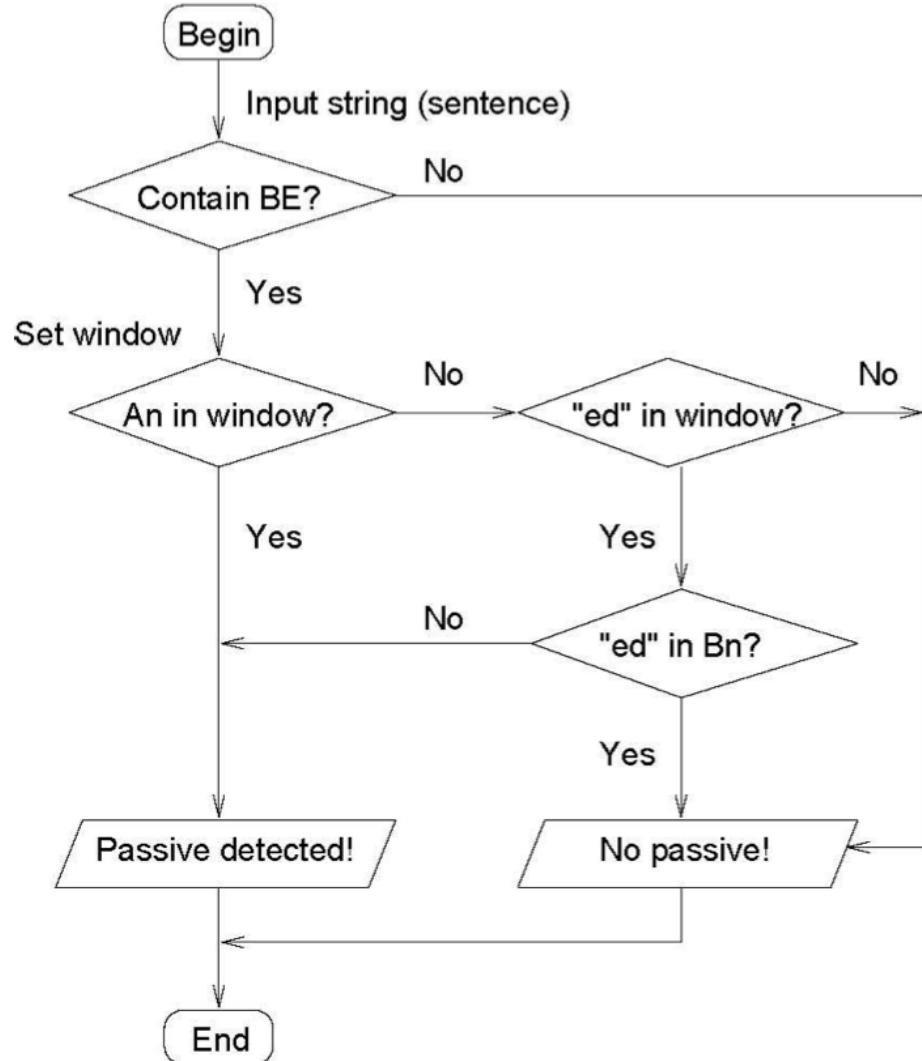
- 1) Read the algorithm in [the picture](#) carefully, which is supposed to help when one wants to know whether he should quit his job. Is the algorithm properly designed?
- 2) Design an algorithm likewise, which can identify English sentences that use the passive voice (*e.g. The book has been returned*).

# Algorithms



# Algorithms

- An algorithm is a series of mathematical **steps**, especially in a computer program, which will give you the answer to a particular kind of problem or question.
- Important algorithms: state space search algorithms, machine learning algorithms



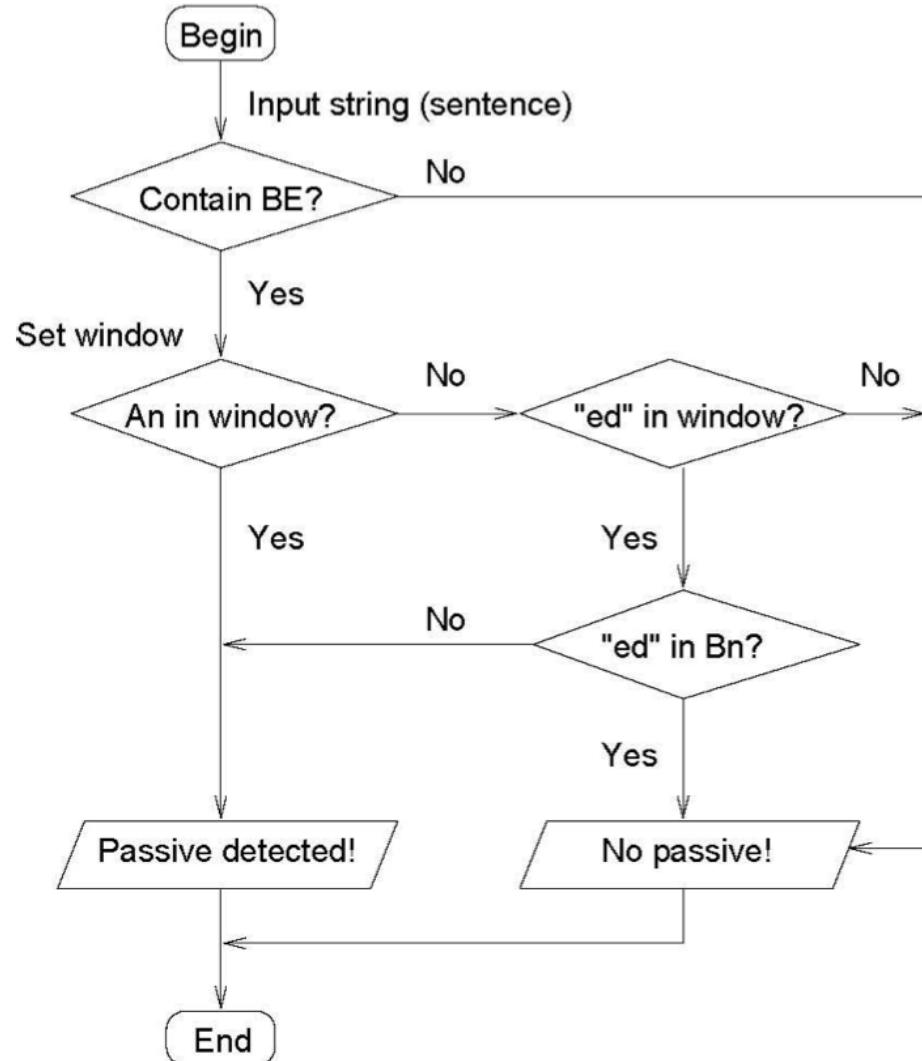
# Algorithms

## Post-class practice (optional)

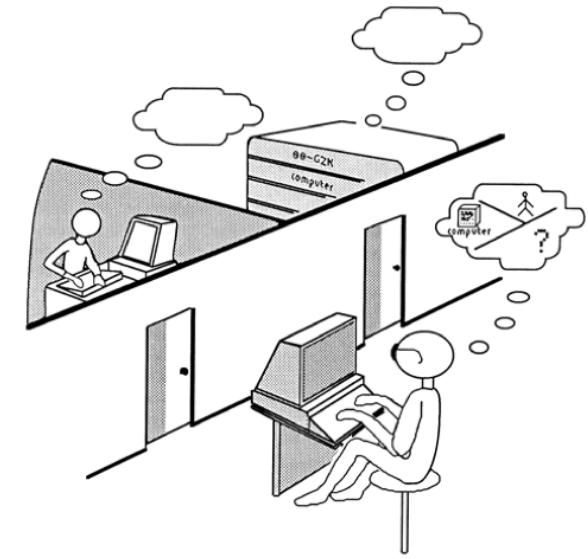
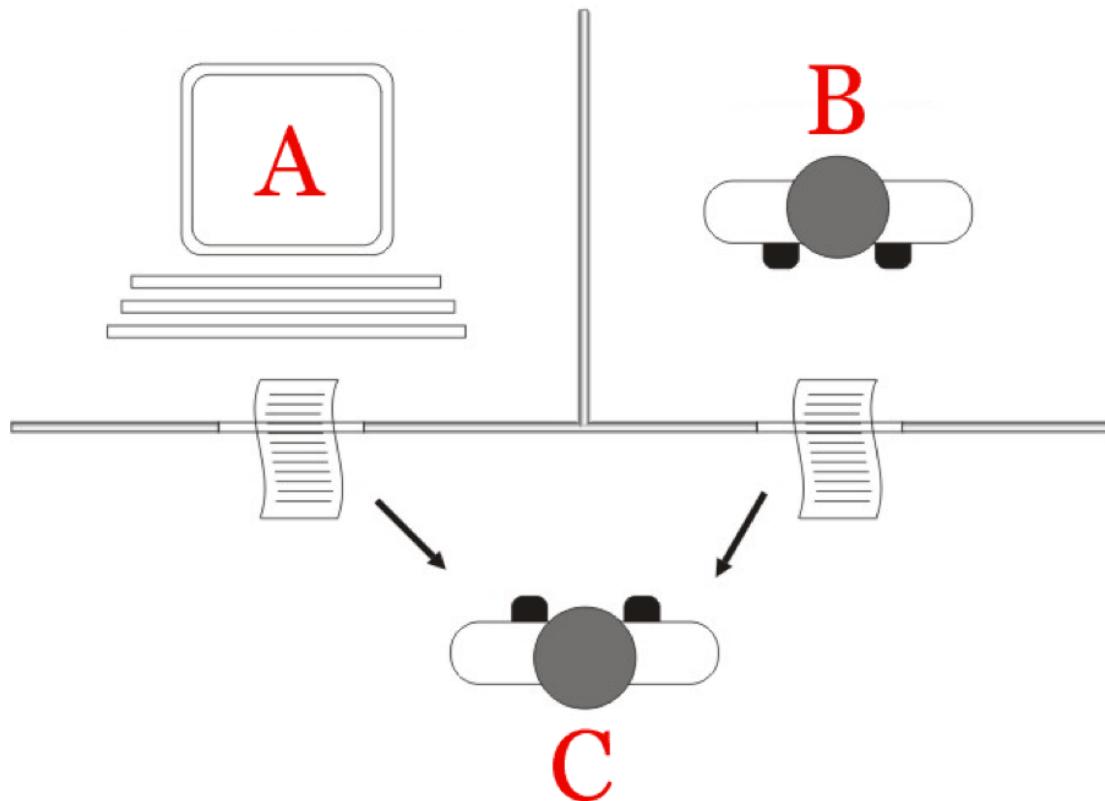
Write code to implement this algorithm, which can identify English sentences that use the passive voice.

*An*: a set of the past participles of irregular verbs

*Bn*: a set of words that end with "ed" but are not the past participles of regular verbs



# The Turing Test



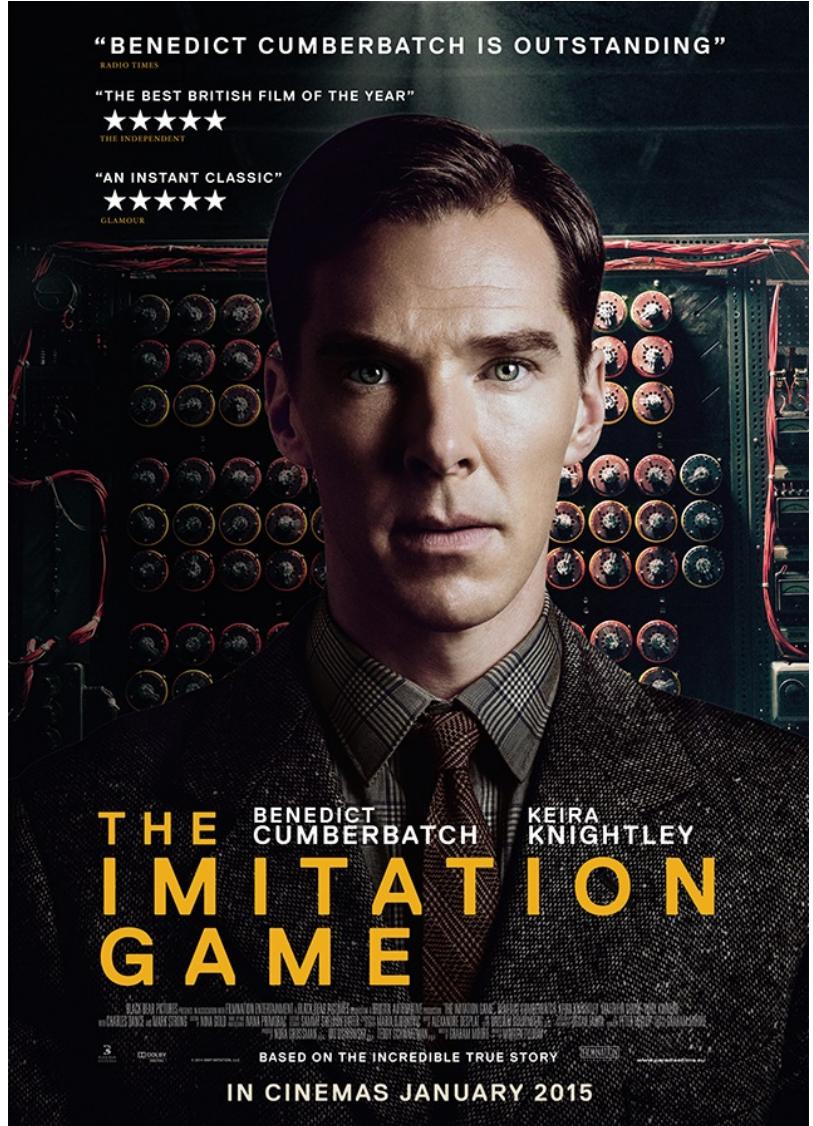
By Alan Turing in 1950

# The Turing Test

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then **teach it to understand and speak English**. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

**We can only see a short distance ahead, but we can see plenty there that needs to be done.**

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.



Directed by Morten Tyldum  
Written by Graham Moore  
Based on *Alan Turing: The Enigma* by Andrew Hodges

# At the end of this session you will

- understand what computational linguistic is in the general sense;
- understand what this course is for and what it focuses on;
- have some preliminary but important ideas about computational linguistic;
- get started with Python, NLTK, and PyCharm;
- learn basic data types and sequence operations in Python.

# Homework

- Review: [J+M 1](#) (Quiz 1 on Sept. 26, 2018)
- Read: [J+M 2](#)
- Read and practice: <http://www.nltk.org/book/ch01.html>

## Next session

Text Normalization, Finite State Transducers,  
and Morphological Parsing