



TRIPLE LETTER SCORE

TRIPLE LETTER SCORE

TRIPLE LETTER SCORE

DOUBLE LETTER SCORE

DOUBLE LETTER SCORE

DOUBLE LETTER SCORE

DOUBLE LETTER SCORE

L<sub>1</sub>

DOUBLE LETTER SCORE

N<sub>1</sub> A<sub>1</sub> T<sub>1</sub> U<sub>1</sub> R<sub>1</sub> A<sub>1</sub> L<sub>1</sub>

TRIPLE WORD SCORE

DOUBLE LETTER SCORE

DOUBLE LETTER SCORE

DOUBLE LETTER SCORE

DOUBLE LETTER SCORE

N<sub>1</sub>

TRIPLE LETTER SCORE

TRIPLE LETTER SCORE

TRIPLE LETTER SCORE

TRIPLE LETTER SCORE

G<sub>2</sub>

DOUBLE WORD SCORE

DOUBLE WORD SCORE

U<sub>1</sub>

DOUBLE LETTER SCORE

DOUBLE WORD SCORE

DOUBLE LETTER SCORE

DOUBLE LETTER SCORE

A<sub>1</sub>

DOUBLE WORD SCORE

C<sub>3</sub> O<sub>1</sub> M<sub>3</sub> P<sub>3</sub> U<sub>1</sub> T<sub>1</sub> I<sub>1</sub> N<sub>1</sub> G<sub>2</sub>

DOUBLE WORD SCORE

DOUBLE WORD SCORE

TRIPLE LETTER SCORE

TRIPLE LETTER SCORE

DOUBLE WORD SCORE

E<sub>1</sub>

TRIPLE WORD SCORE

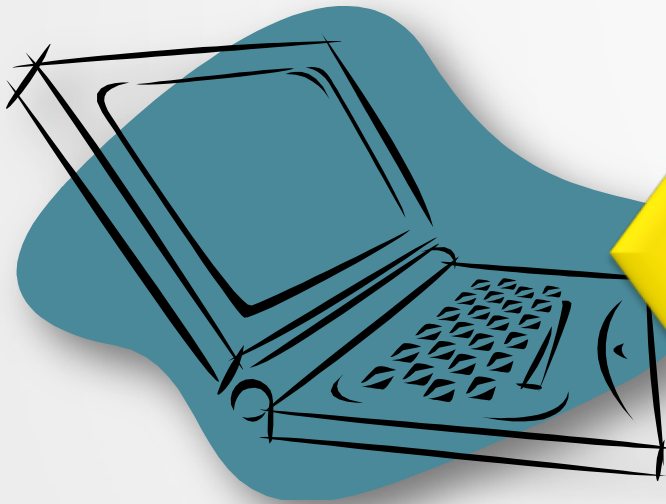
DOUBLE LETTER SCORE

TRIPLE WORD SCORE

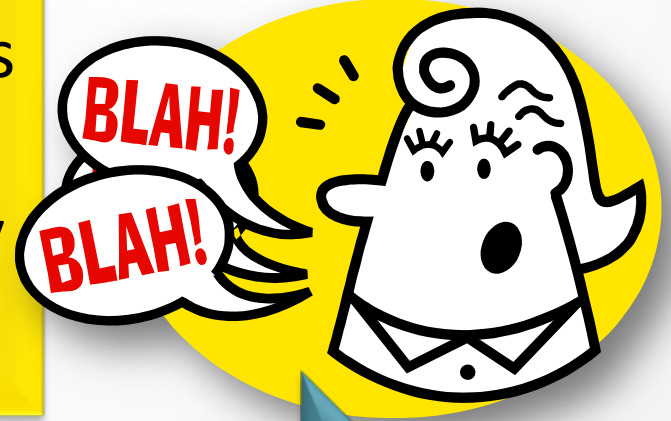
DOUBLE LETTER SCORE

TRIPLE WORD SCORE

# What is natural language computing?



Getting computers to understand everything we say and write.



In this class (and in the field generally), we are interested in learning the *statistics of language*.

Increasingly, computers give insight into how humans process language, or generate language themselves.

# Today

- Basic definitions in **natural language processing (NLP)**.
- Applications
  - Translating between languages
  - Speech recognition
  - Answering questions
  - Engaging in dialogue
- Course logistics.



Examples

# What can natural language do?

The ultimate in **human-computer interaction**.

“translate *Also Sprach Zarathustra*”

“take a memo...”

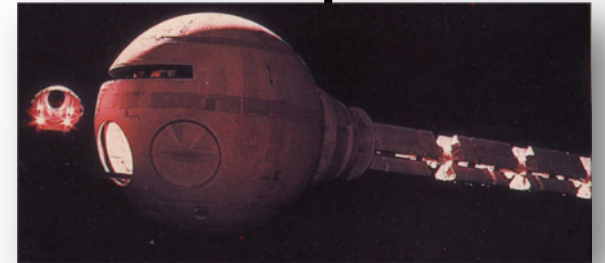
“open the pod bay doors”

“how far until Jupiter?”

“Can you summarize *2001: A Space Odyssey*?”




`open (podBay.doors) ;`




We're making progress, but why are these things *still* hard to do?

# A little deeper

- Language has *hidden structures*, e.g.,
  - How are **sounds** and **text** related?
    - e.g., why is this:  not a 'ghoti' (*enough*, *women*, *nation*)?
  - How are words **combined** to make sentences?
    - e.g., what makes '*colourless green ideas sleep furiously*' **correct** in a way **unlike** '*furiously sleep ideas green colourless*'?
  - How are words and phrases used to produce **meaning**?
    - e.g., if someone asks '*do you know what time it is?*', why is it **inappropriate** to answer '*yes*'?
- We need to organize the way we think about language...

# Categories of linguistic knowledge

- Phonology: the study of patterns of speech sounds.  
e.g., “read” → /r iy d/
- Morphology: how words can be changed by inflection or derivation.  
e.g., “read”, “reads”, “reader”, “reading”, ...
- Syntax: the ordering and structure between words and phrases (i.e., grammar).  
e.g., *NounPhrase* → *article adjective noun*
- Semantics: the study of how meaning is created by words and phrases.  
e.g., “book” → 
- Pragmatics: the study of meaning in contexts.  
e.g., explanation span, refutation span

# Ambiguity – Phonological

- Phonology: the study of patterns of speech sounds.

Problem for  
speech synthesis

“read” → /r iy d/

as in ‘I like to **read**’

“read” → /r eh d/

as in ‘She **read** a book’

“object” → /aa<sup>1</sup> b jh eh<sup>0</sup> k t /

as in ‘That is an **object**’

“object” → /ah<sup>0</sup> b jh eh<sup>1</sup> k t /

as in ‘I **object!**’

Problem for  
speech recognition

“too” ← /t uw/

as in ‘**too** much’

“two” ← /t uw/

as in ‘**two** beers’

- Ambiguities can often be **resolved** in context, but not always.
  - e.g., /h aw t uw r eh<sup>1</sup> k ah ?? n ay<sup>2</sup> z s (b|p) iy ch/
    - ‘how to recognize speech’
    - ‘how to wreck a nice beach’

# Resolution with syntax

- If you hear the sequence of speech sounds

*/b ah f ae l ow b ah f ae l ow b ah f ae l ow b ah f ae l ow ...  
b ah f ae l ow b ah f ae l ow b ah f ae l ow b ah f ae l ow/*

which word sequence is being spoken?

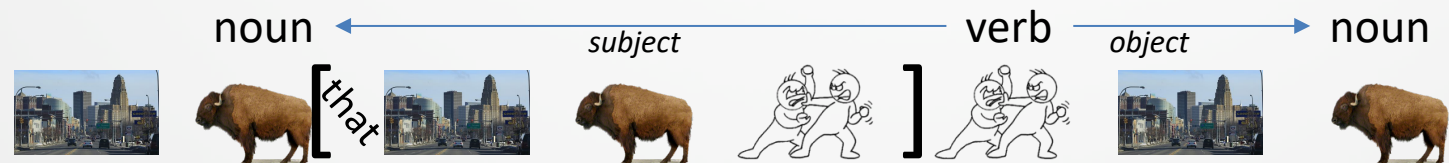
→ “Buff a low buff a lobe a fellow Buff a low buff a lobe a fellow...”

→ “Buffalo buff aloe buff aloe buff aloe buff aloe buff aloe ...”

→ “Buff aloe buff all owe Buffalo buffalo buff a lobe ...”

→ “Buff aloe buff all owe Buffalo buff aloe buff a lobe ...”

→ **“Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo”**



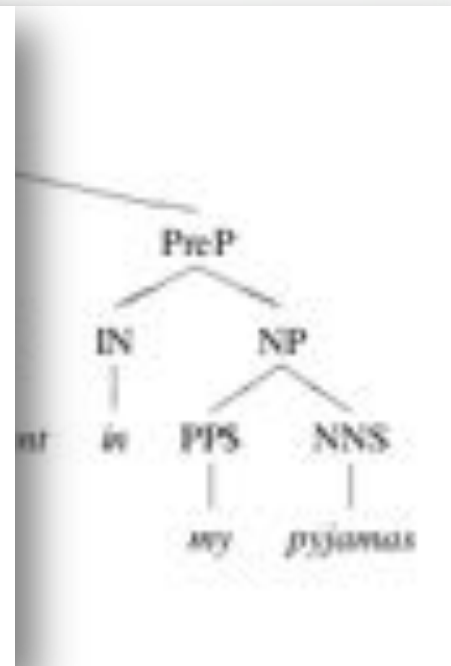
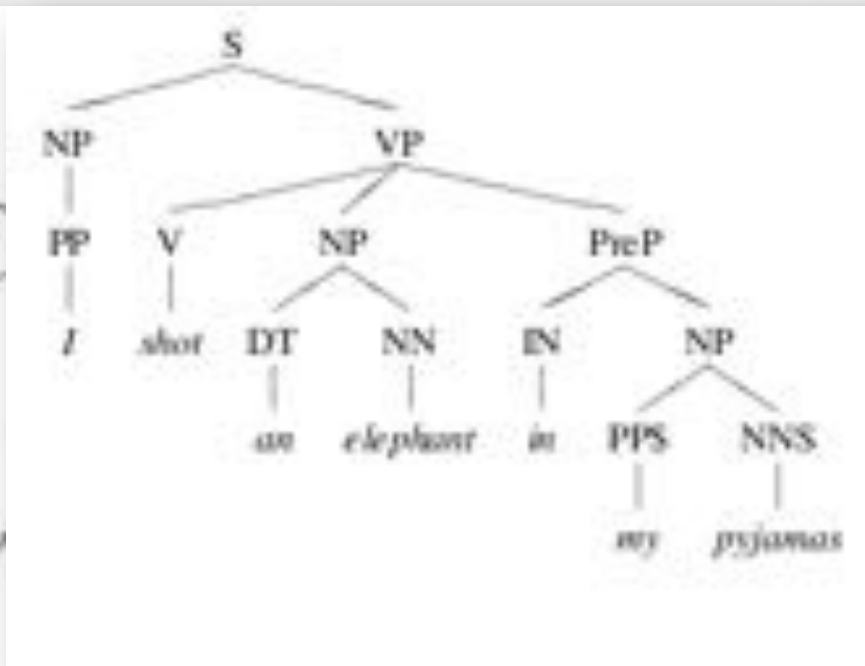
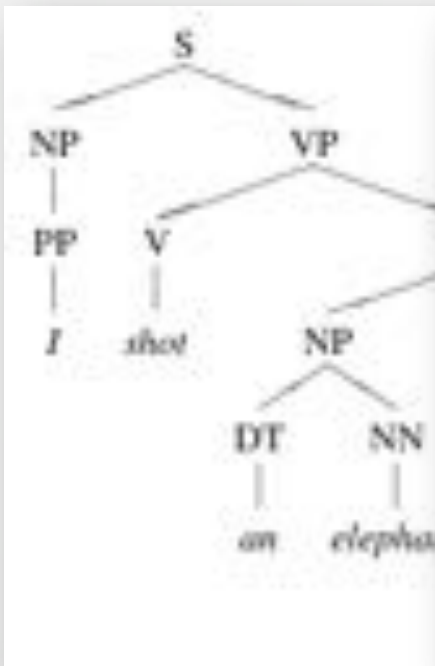
- It's obvious (to us) that the last option is most likely because we have knowledge of **syntax**, i.e., grammar.



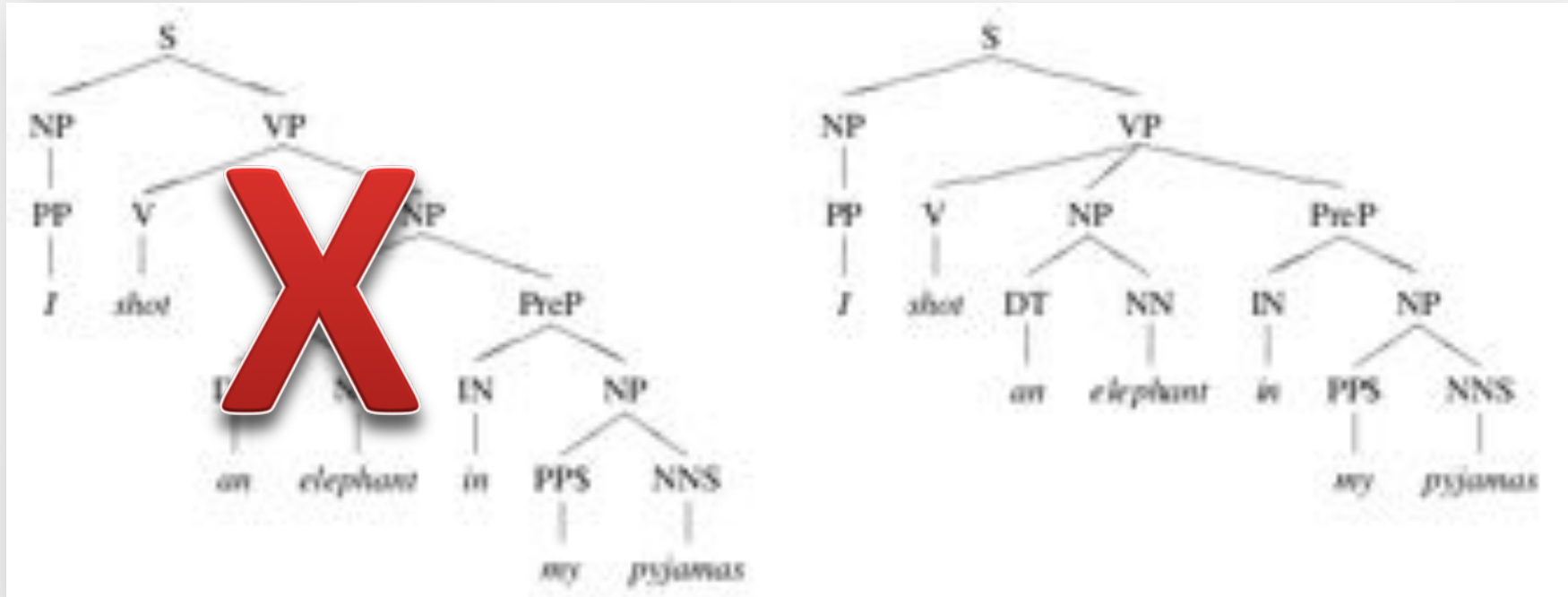
# Ambiguity – Syntactic

- Syntax: the ordering and structure between words. Words can be grouped into ‘parse tree’ structures given grammatical ‘rules’.

e.g., “I shot an elephant in my pyjamas”



# Resolution with semantics



- It's obvious (to us) that the elephants don't wear pyjamas, and we can discount one option because of our knowledge of **semantics**, i.e., meaning.

# Ambiguity – Semantic

- Semantics: the study of how meaning is created by the use of words and phrases.
- 

- “Every man loves a woman”

→  $\forall x \text{ man}(x) \exists y: (\text{woman}(y) \wedge \text{loves}(x, y))$

→  $\exists y: \text{woman}(y) \wedge \forall x (\text{man}(x) \rightarrow \text{loves}(x, y))$

- “I made her duck”

→ I cooked waterfowl meat for her to eat.

→ I cooked waterfowl that belonged to her.

→ I carved the wooden duck that she owns.

→ I caused her to quickly lower her head.

- “Give me the pot”

→ It’s time to bake.

→ It’s time to get baked.

# Resolution with pragmatics

- It's obvious (to us) which meaning is intended given **knowledge** of the **context** of the conversation or the **world** in which it takes place.

- “Every man loves a woman”

→  $\forall x \text{ man}(x) \exists y: (\text{woman}(y) \wedge \text{loves}(x, y))$

~~→  $\exists y: \text{woman}(y) \wedge \forall x (\text{man}(x) \rightarrow \text{loves}(x, y))$~~

If you know that no one woman is so popular

- “I made her duck”

→ I cooked waterfowl meat for her to eat.

~~→ I cooked waterfowl that belonged to her.~~

~~→ I carved the wooden duck that she owns.~~

~~→ I caused her to quickly lower her head.~~

If the question was “*what type of food did you make for her?*”

- “Give me the pot”

~~→ It's time to bake.~~

→ It's time to get baked.

If the conversation is taking place in Canada

# Ambiguity – miscellaneous

- Newspaper headlines (spurious or otherwise)

**Kicking Baby Considered to  
be Healthy**

...

**Squad Helps Dog Bite Victim**

...

**Canadian Pushes Bottle Up Germans**

...

**Milk Drinkers are Turning to Powder**

...

**Grandmother of Eight Makes  
Hole in One**

...

**Kids Make Nutritious Snacks**

...

**Juvenile Court Tries Shooting  
Defendant**

...

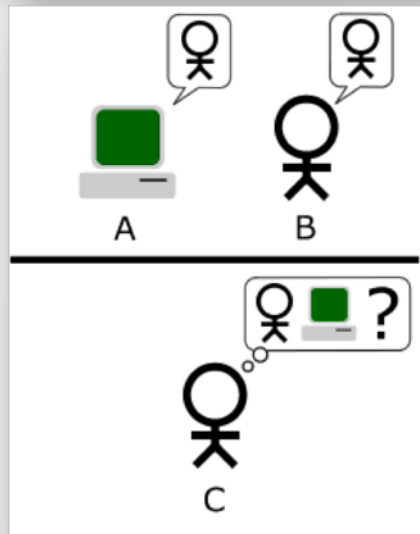
**Local High School Dropouts  
Cut in Half**

...

# NLP as Artificial Intelligence

- NLP involves **resolving ambiguity** at all levels.
  - **Reasoning with world knowledge.**
    - In the early days knowledge was **explicitly encoded** in artificial **symbolic** systems (e.g., context-free grammars) by **experts**.
  - Now, algorithms learn using **probabilities** (or pseudo-probabilities) to distinguish subtly different competing hypotheses.
    - E.g., is *Google* a **noun** or a **verb**?
    - Examples where *Google*  $\in$  *Nouns* ("**Google** makes Android"), does **not** mean that Google is never a verb ("**Go** Google yourself").
    - $P(\textit{Google} \in \textit{Nouns}) > P(\textit{Google} \in \textit{Verbs}) > 0$

# The Turing Test



- **First** and most **fundamental** test of machine intelligence.
- A machine (A) imitates a human using nothing but a text-based instant messenger.
  - If a human interrogator (C) cannot reliably differentiate a real human (B) from the machine, that machine is said to be ‘intelligent’.
- Turing, Alan M. (1950) Computing machinery and intelligence. *Mind*, **59**, pp. 433-460.

# Aside – Chatbots

- ELIZA (Weizenbaum, 1966): simple pattern matching to imitate a psychiatrist.
- Surprisingly effective despite **no linguistic knowledge.**
- e.g.,

**User:** Men are all alike.

**ELIZA:** In what way?

**User:** They're always bugging us about something or other.

**ELIZA:** Can you think of a specific example?

**User:** My boyfriend made me come here.

**ELIZA:** Your boyfriend made you come here.

(Jurafsky and Martin, 2009)





# Course outline (approximate)

- Introduction, linguistic data, language models (3 lectures)
- Features and classification (1 lecture) \*
- Neural language models (2 lectures) \*
- Hidden Markov models (3 lectures) \*
- Machine translation (3 lectures) \*\*
- Articulatory and acoustic phonetics (2 lectures) \*
- Automatic speech recognition (2 lectures) \*\*
- Speech synthesis (1 lecture) \*\*
- Entropy and information theory (2 lectures) \*
- Information retrieval (1 or 2 lectures) \*\*
- Dialogue and chatbots (1 lecture) \*\*
- Review (1 lecture)

\* techniques

\*\* applications



# Preview: Machine translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

- One of the most prized applications in NLP.
- Requires both **interpretation** and **generation**.
- Over \$100B spent annually on human translation.

# Preview: Machine translation

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

<b>Human</b>	According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959B US dollars of foreign capital, including 40.007B US dollars of direct investment from foreign businessmen.
<b>IBM4</b>	The Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007B US dollars today provide data include that year to November China actually using foreign 46.959B US dollars and
<b>Yamada/Knight</b>	Today's available data of the Ministry of Foreign Trade and Economic Cooperation shows that China's actual utilization of November this year will include 40.007B US dollars for the foreign direct investment among 46.959B US dollars in foreign capital.

# Preview: Machine translation

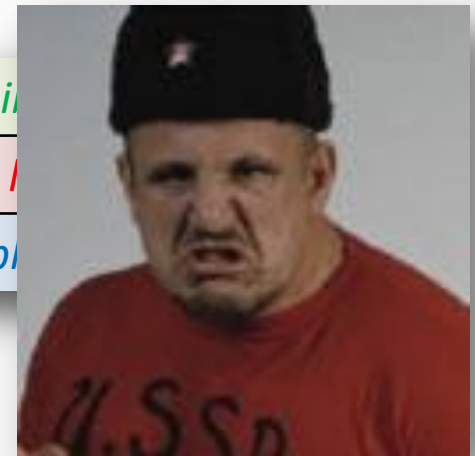
- In the 1950s and 1960s direct **word-for-word** replacement was popular.
  - Due to semantic and **syntactic ambiguities** and **differences** in source languages, results were mixed.



US English

“The spirit is willing, but the  
flesh is weak”

“The vodka is good, but the  
meat is rotten”



Russian

# Preview: Machine translation

- One problem is disparity of meanings in languages.



Stephen  
Harper

Former Prime Minister of Canada

**nation** *n.* a large body of people, associated with a particular **territory**, that is sufficiently conscious of its **unity** to seek or to possess a **government** of its own

**nation** *n.* an aggregation of persons of the same **ethnic family**, often speaking the same **language** or cognate **languages**



Pauline  
Marois

Former Première Ministre du Québec

# Preview: Machine translation

- Solution: automatically learn statistics on parallel texts

... citizen of Canada has the right to vote in an election of members of the House of Commons or of a legislative assembly and to be qualified for membership ...



... citoyen canadien a le droit de vote et est éligible aux élections législatives fédérales ou provinciales ...

e.g., the *Canadian Hansards*:  
bilingual Parliamentary proceedings

# Statistical machine translation

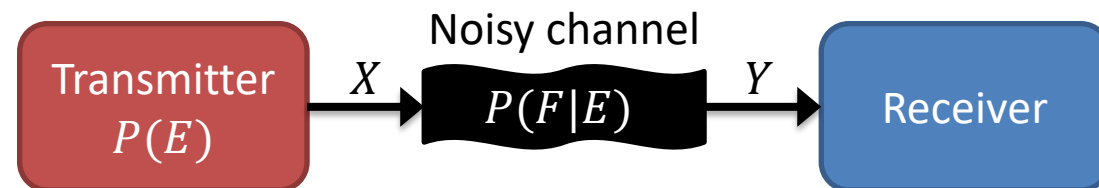
- Much of modern statistical machine translation is based on the following perspective...



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**.'

Warren Weaver

March, 1947



Claude Shannon

July, 1948

# Aside – Machine translation

- <http://www.translationparty.com> uses Google Translate to go back and forth between English and Japanese until we get two consecutive identical English phrases.

Start with an English phrase:

I want to learn about natural language computing	let's go!
私は自然言語コンピューティングを勉強したい	into Japanese
I want to learn natural language computing	back into English
私は自然言語コンピューティングを勉強したい	back into Japanese
I want to learn natural language computing	back into English

**Equilibrium found!**  
You've heard about [Question Party](#) right?



# Preview: Machine translation

Start with an English phrase:

that's one small step for a man, one giant leap for mankind

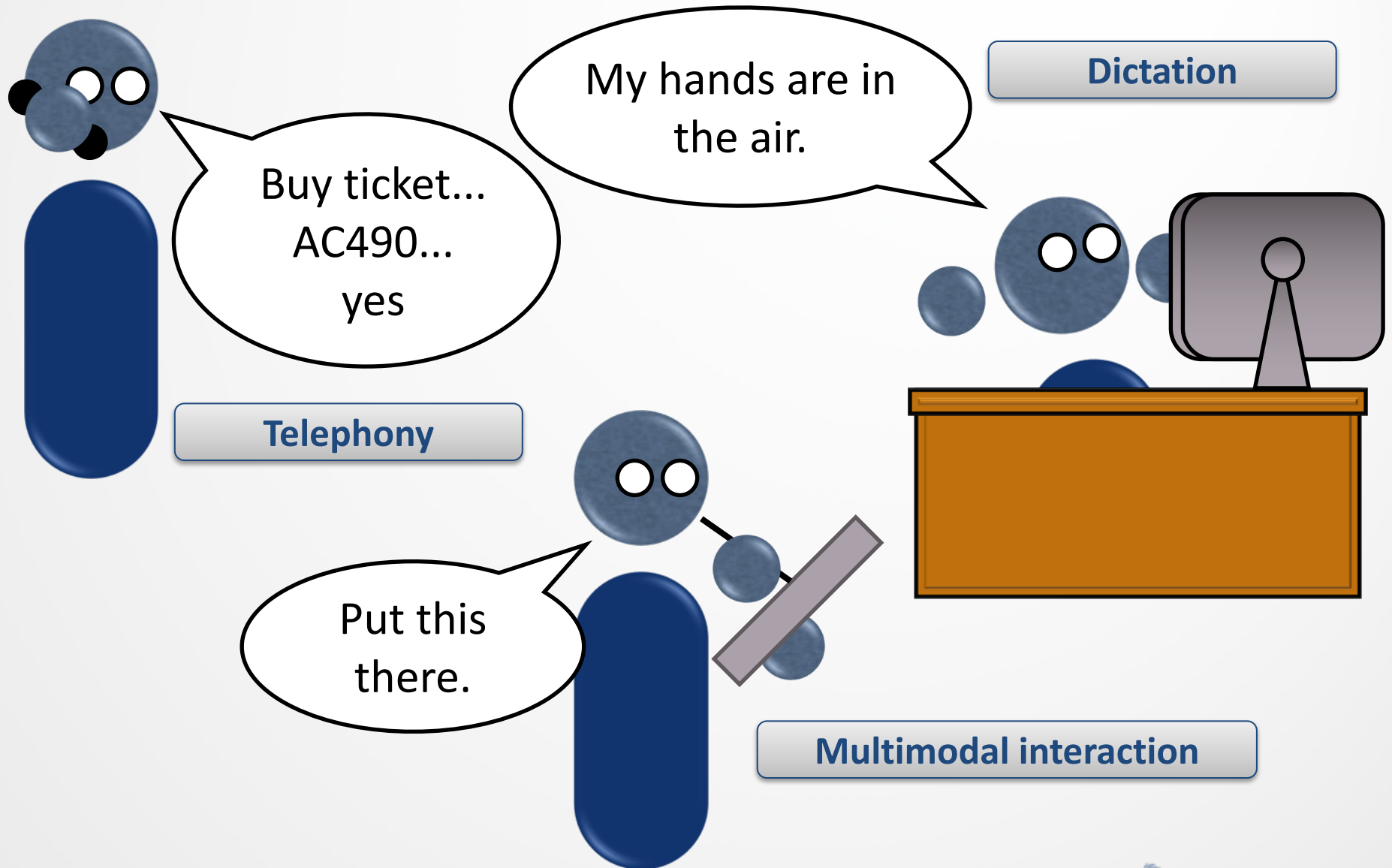
find equilibrium

that's one small step for a man, one giant leap for mankind let's go!

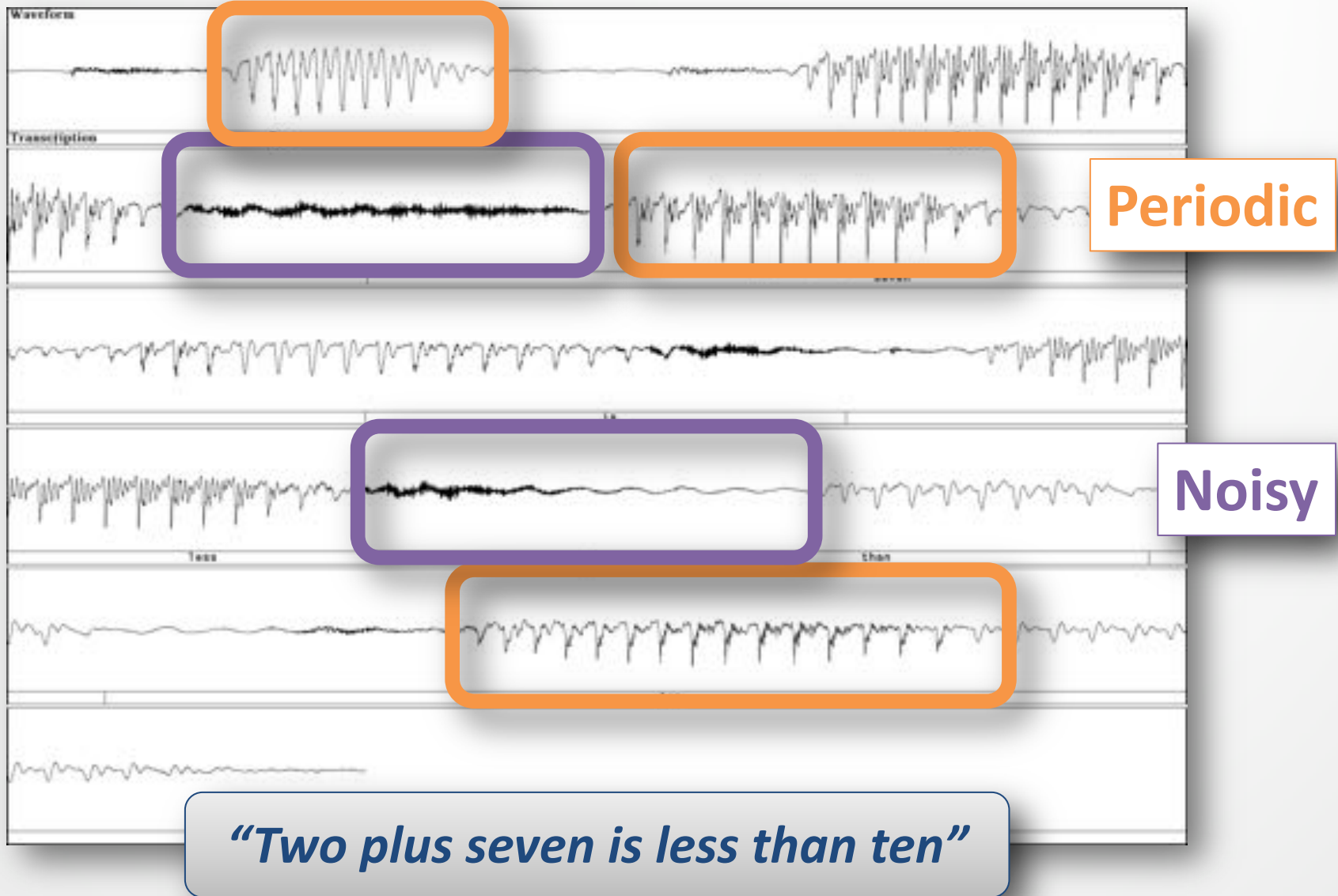
それは人間にとっては小さな一歩だが、一歩	人類にとっては小さな一歩ステップは、男性にとって理想的なホテルです	<a href="#">back into Japanese</a>
It is but one small step for man, o mankind	Step One small step for mankind, this hotel is ideal for men	<a href="#">back into English</a>
それは人間にとっては小さな一歩一歩、しいが、さ	人類にとっては小さな一歩ステップ、このホテルは、男性にとって理想的なホテルです	<a href="#">back into Japanese</a>
It is step by small step for man, b humanity, the	Step One small step for mankind, this hotel is ideal for men	<a href="#">back into English</a>

**Equilibrium found!**  
Okay, I get it, you like Translation Party.

# Preview: Speech recognition

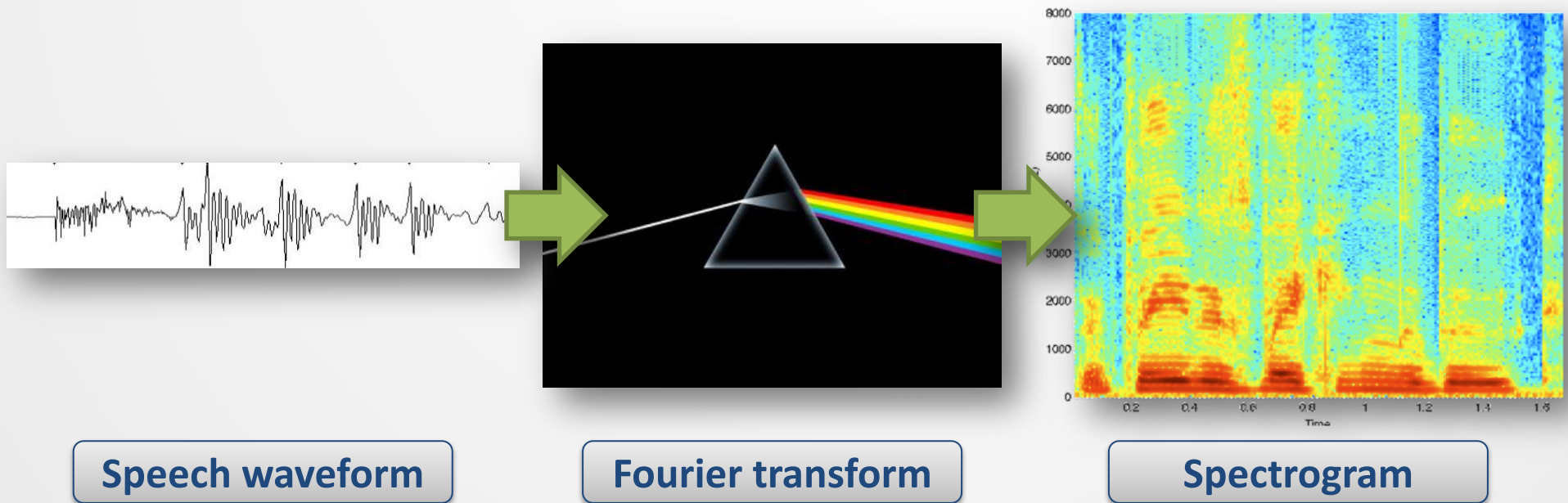


# Speech waveforms

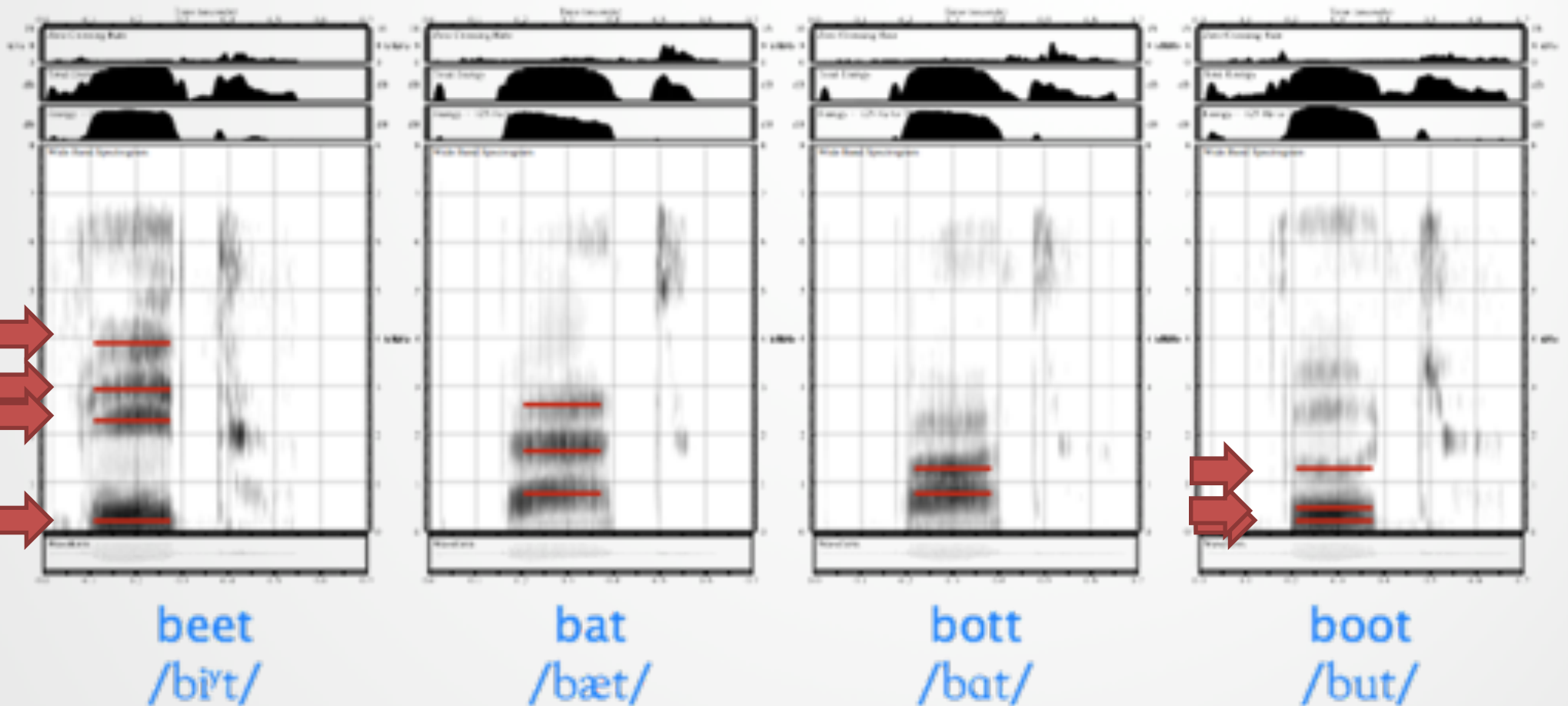


# Spectrograms

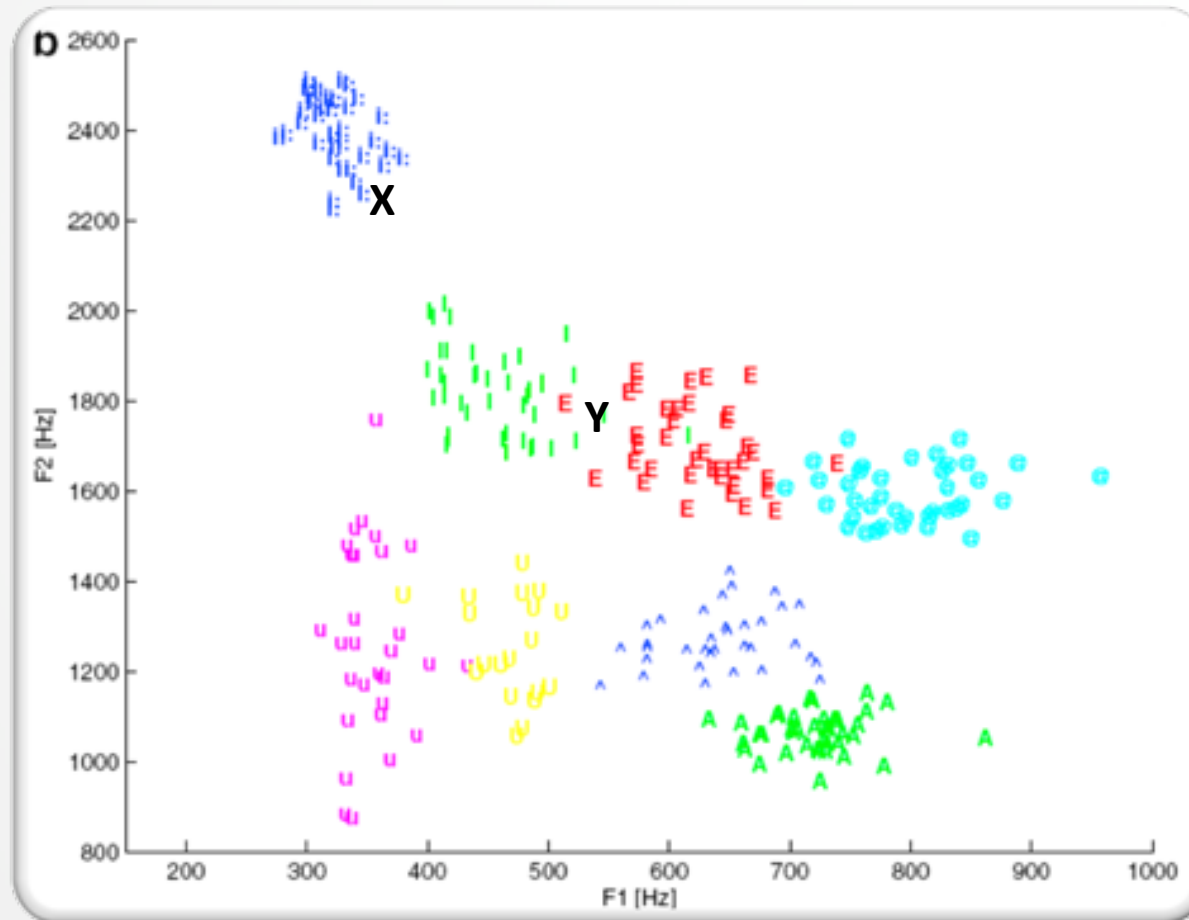
- Speech sounds can be thought of as overlapping **sine waves**.
  - Speech is **split apart** into a 3D graph called a ‘**spectrogram**’.
  - Spectrograms allow machines to extract **statistical features** that differentiate between different kinds of sounds.



# Speech recognition



# Preview: Speech recognition



What is Y?

- In order to classify an unknown observation (e.g., **X**), we need a **statistical** model of the distribution of sounds

# Preview: Questions and answers



Which woman has won more than 1 Nobel prize?



This one



(Marie Curie)

- **Question Answering (QA)** and **Information Retrieval (IR)** involve many of the same principles.

# Preview: Information retrieval



what woman won more than one nobel prize

All News Videos Images Shopping More Settings Tools

About 4,000,000 results (0.49 seconds)

**Marie Curie** won the Nobel prize in 1903 for Physics and 1911 in Chemistry; Linus Pauling in 1954 (for Chemistry) and 1962 (for Peace); John Bardeen in 1956 (for Physics) and 1972; Frederick Sanger in Chemistry in 1958 and 1980. Who has won more than one Nobel prize? Apr 1, 2007

[Who has won more than one Nobel prize? - Times of India](#)  
timesofindia.indiatimes.com/home/...won-more-than-one-Nobel-prize/.../1839923.cms

About this result Feedback

People also ask

- Who has won Nobel Prize twice?
- What women won the Nobel Prize?
- How many women have won the Nobel Prize?
- How many women have been awarded the Nobel Peace Prize?

Feedback



which woman has won more than 1 nobel prize?

Using closest WolframAlpha interpretation: nobel prize

WolframAlpha computational knowledge engine.

what woman won more than one nobel prize?

Web Apps Examples Random

Using closest WolframAlpha interpretation: **won more than one**

More interpretations: nobel prize woman

Assuming Korean won for "won" | Use North Korean won instead

2010	Richard F. Heck	chemistry	United States	United States
2010	Christopher A. Pissarides	economics	United Kingdom	Cyprus
2010	Dale T. Mortensen	economics	United States	United States
2010	Peter A. Diamond	economics	United States	United States
2010	Mario Vargas Llosa	literature	Peru	Peru



# Aside – Question answering

WolframAlpha™ computational knowledge engine

How much potassium is in 450,000 cubic kilometers of bananas?

Input interpretation:

banana	amount	450 000 km <sup>3</sup> (cubic kilometers)	potassium
--------	--------	--	-----------

Result:

$1.5 \times 10^{12}$  t (metric tons)

AT&T 3:33 PM 100%

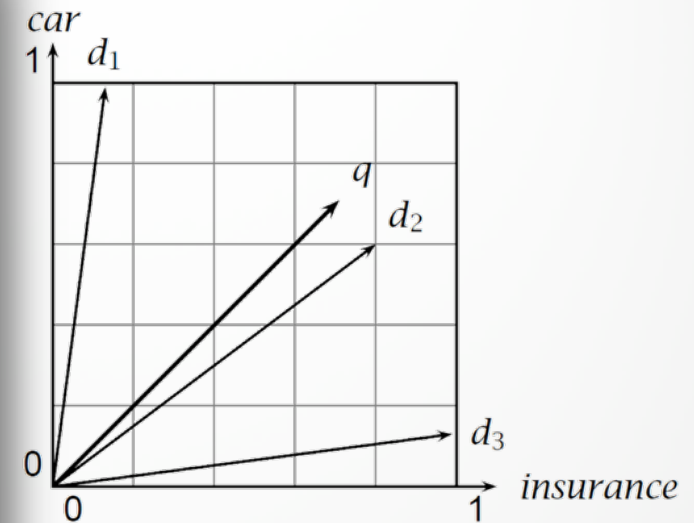
“Should I bring an umbrella next Monday?”  
tap to edit

There's no rain in the forecast next Monday:

New York  
Weekly Forecast

Monday		50 36
Tuesday		48 36
Wednesday		50 32
Thursday		43 32
Friday		39 30
Saturday		37 34
Sunday		37 30
Monday		36 27
Tuesday		37 30
Wednesday		45 36

# Answer questioning?



$$\cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^n q_i d_i}{\sqrt{\sum_{i=1}^n q_i^2} \sqrt{\sum_{i=1}^n d_i^2}}$$

- Retrieving information can be a **clever combination** of many very **simple concepts** and algorithms.

# Overview: NLP

- Is natural language processing (the discipline) hard?
  - **Yes**, because **natural language**
    - is highly ambiguous at all levels,
    - is complex and subtle,
    - is fuzzy and probabilistic,
    - involves real-world reasoning.
  - **No**, because **computer science**
    - gives us many powerful statistical techniques,
    - allows us to break the challenges down into more manageable features.
- Is Natural Language Computing (the course) hard?
  - More on this soon...

# NLP in industry



# Natural language computing

- **Instructor**: Frank Rudzicz and Sean Robertson ([csc401-2020-01@cs](mailto:csc401-2020-01@cs))
- **TAs**: Julia Watson, KP Vishnubhotla, Mohamed Abdalla, Shuja Khalid, Amanjit Kainth, Zhining Zhu
- **Meetings**: MF (lecture, MB128), W (tutorial, SF1105) at 10h-11h
- **Languages**: English, Python.
- **Website**: <http://www.cs.toronto.edu/~frank/csc401/>
- **You**: Understand basic **probability**, can **program**, or can pick these up as we go.
- **Syllabus**: Key **theory** and **methods** in statistical natural language computing.  
Focus will be on *Markov and neural models, machine translation, and speech recognition.*

# Office hours

- **Time:**
  - Mondays, 11h00-12h00
- **Location:**
  - The Vector Institute (MaRS West, Suite 710)
  - The streets



# Evaluation policies

- **General**: Three assignments : **15%, 20%, 25%** (ranked by your mark)  
Final exam : **40%**
- **Lateness**: **10%** deduction applied to electronic submissions that are 1 minute late.  
Additional **10%** applied every 24 hours up to 72 hours total, at which point grade is **zero**.
- **Final**: If you **fail** the final exam, then you **fail** the course.
- **Ethics**: Plagiarism and unauthorized collaboration can result in a grade of **zero** on the homework, **failure** of the course, or **suspension** from the University. See the course website.

# Theme – NLP in a post-truth society

- The **truth** is the most important thing in the Universe.
  - At the very least, the truth allows us to rationally **optimize** legal, political, and personal decisions.
- The truth can sometimes be obscured deliberately via **deception**, or inadvertently through **bias**, **fallacy**, or intellectual **laziness**.
  - Nowhere is this perhaps more obvious than on **social media** or in **pseudo-journalism**.
- Natural language processing gives us **tools** to combat this scourge.



# Assignments

- Assignment 1: Corpus statistics, sentiment analysis
  - task: analyze bias on Reddit
  - learn: statistical techniques, features, and classification.
- Assignment 2: Statistical machine translation
  - task \*: translate between political extremes
  - learn: neural seq2seq and language models.
- Assignment 3: Automatic speech recognition
  - task: detect lies in speech
  - learn: signal processing, phonetics, and hidden Markov models.

# Assignment 1 – Bias in social media

- Involves:
  - Working with social media data  
(i.e., gathering statistics on some data from Reddit),
  - Part-of-speech tagging (more on this later),
  - Classification.
- **Announcements:** Piazza forum, email.
- You should get an early start.



# Projects – graduate students only

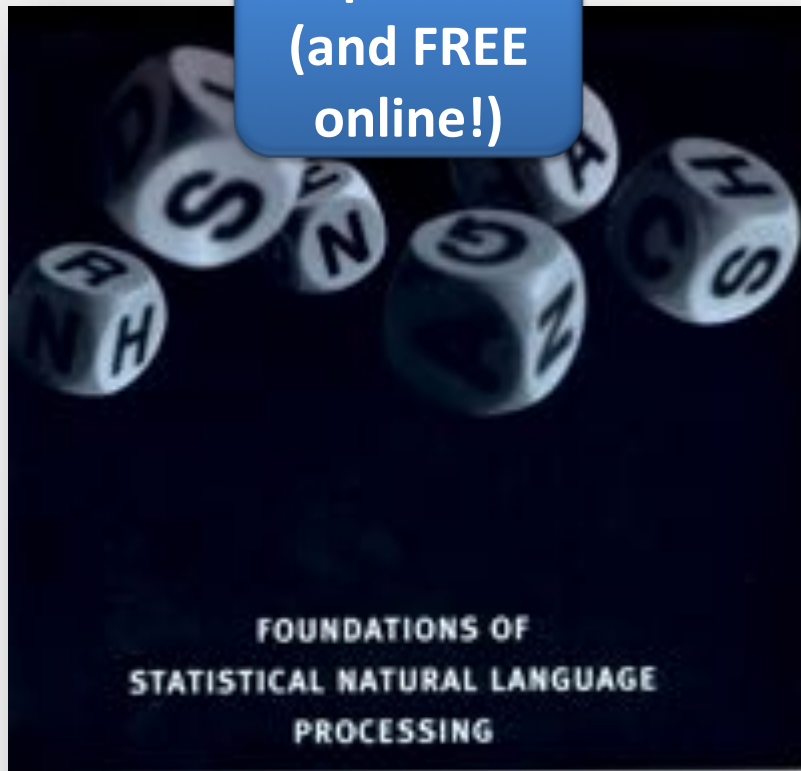
- Graduate students can **optionally** undertake a full-term **project** worth **60%** of their grade **instead** of the assignments.
  - Good for those, e.g., who prefer to work in teams.
- Teams must consist of 1 or 2 humans (no more, no fewer).
- Projects must contain a significant **programming** and **scientific** component.
- Projects must be **relevant** to the course.

# Projects – graduate students only

- Some possible ideas for projects include:
  - A deception filter for news media online.
  - A novel method of using data in language  $A$  to train a classification system in language  $B$  for  $A \neq B$ .
- If you decide to take this option, you have to notify us by email about your team by **18 January!**
- You will need to periodically submit **checkpoints** that build on their antecedents.
  - See course webpage for detailed requirements!

# Reading

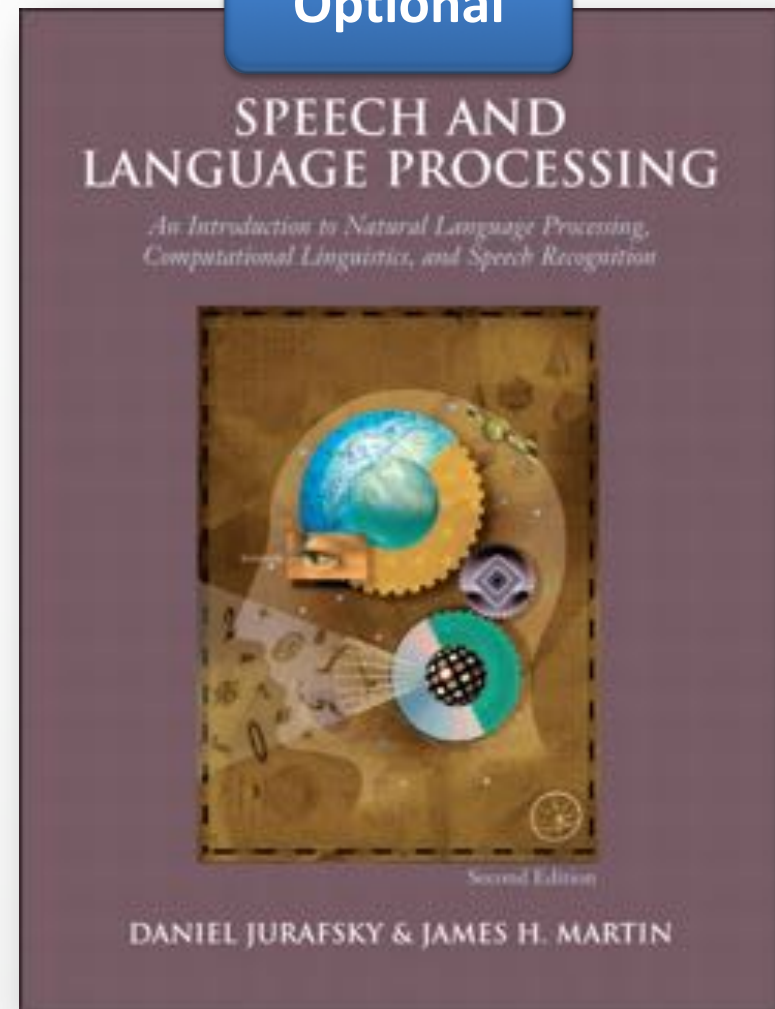
Optional  
(and FREE  
online!)



CHRISTOPHER D. MANNING AND  
HINRICH SCHÜTZE

<https://search.library.utoronto.ca/details?10552907>

Optional



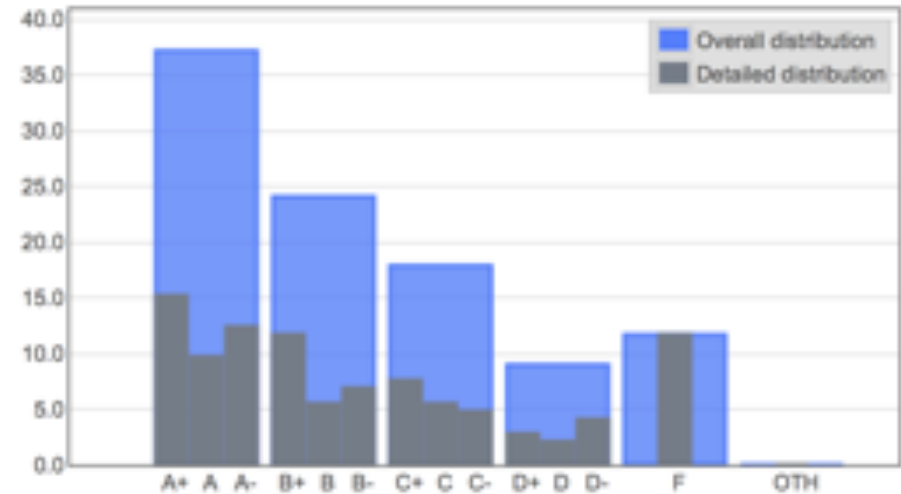
DANIEL JURAFSKY & JAMES H. MARTIN

# Stats from 2017-2019

2017

A	37.2%	B	24.1%	C	17.9%	D	9%
A+	15.2%	B+	11.7%	C+	7.6%	D+	2.8%
A	9.7%	B	5.5%	C	5.5%	D	2.1%
A-	12.4%	B-	6.9%	C-	4.8%	D-	4.1%
F	11.7%	OTH	0%	Average		Median	
F	11.7%	OTH	0%	70.01 %		76 %	

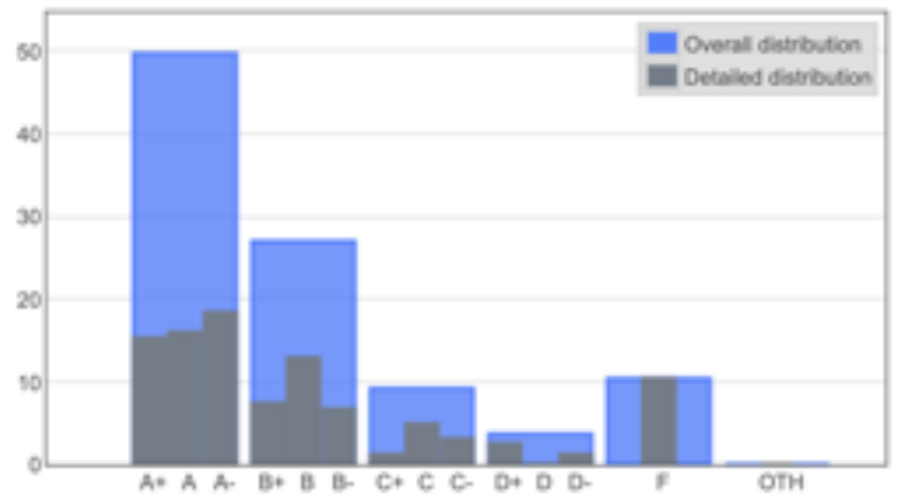
Class average excluding exam no shows: 75.20%  
 Fails excluding exam no shows: 3.79%



2019

A	49.7%	B	27%	C	9.2%	D	3.7%
A+	15.3%	B+	7.4%	C+	1.2%	D+	2.5%
A	16%	B	12.9%	C	4.9%	D	0%
A-	18.4%	B-	6.7%	C-	3.1%	D-	1.2%
F	10.4%	OTH	0%	Average		Median	
F	10.4%	OTH	0%	73.54 %		79 %	

Class average excluding exam no shows: 77.52%  
 Fails excluding exam no shows: 4.58%



# Assignment 1 and reading

- **Assignment 1** available by Friday (on course webpage!)
  - Due 11 February
  - TAs: Julia Watson;  
KP Vishnubhotla.
- **Reading:**
  - Manning & Schütze: Sections 1.3—1.4.2,  
Sections 6.0—6.2.1.