

# **Introduction to Natural Language Processing (NLP)**

**Tran Hong-Viet**  
**UET-VNU**

# Content

- Course Information
- Some achievements of NLP
- Overview of NLP
  - Linguistic levels of description
  - Why is NLP difficult?
- Conclusion

# Course information

- **Course:** Natural Language Processing (NLP)
- **Instructor:** Ass Prof. Nguyen Phuong Thai; Dr. Tran Hong Viet,  
Email: [thainp@vnu.edu.vn](mailto:thainp@vnu.edu.vn); [thviet@vnu.edu.vn](mailto:thviet@vnu.edu.vn)  
Tel: 0975486888

# Course information

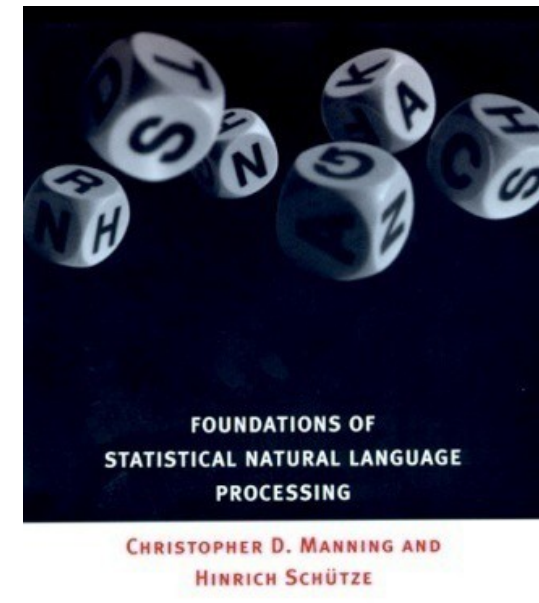
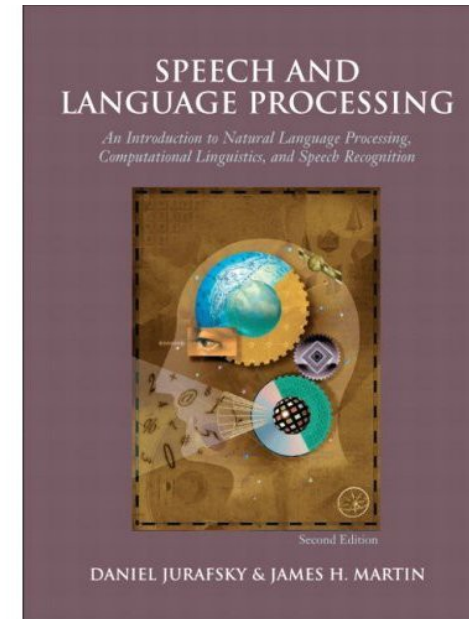
- **Course web page:** <https://courses.uet.vnu.edu.vn/>  
choose NLP course.
  - Up to date information
  - Lecture notes
  - Relevant dates, links, etc.
- **Prerequisites:** Programming principles, discrete mathematics for computing, software design and software engineering concepts, AI. Good knowledge of C++, Java, Python.
- **Python** required for programming assignments.
- **Grading:** 30% for (midterm + homeworks/assignments )  
+10% for attendance + 60% for final

# Policy & Practical issues

- Encourage discussion but assignments must be your individual work
- Codes copied from books or other libraries but be explicitly acknowledged
- Sharing or copying codes is strictly prohibited.

# Reference

- *Slides*
- **Text books:**
  - 1) *Speech and Language Processing*, Daniel Jurasky & James H. Martin, second edition, printed by Prentice Hall, 2009 ( <https://web.stanford.edu/~jurafsky/slp3/> )
  - 2) *Natural Language Processing* , Eisenstein, 2018  
(<https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf>)
  - 3) *Foundation of Statistical Natural Language Processing*, Christopher D. Manning & Hinrich Schutze, 2001



# NLP in Industry





# Communication With Machines



~50-70s

```
File Edit Edit_Settings Menu Utilities Compilers Test Help
EDIT BS9U.DEVT3.CLIPPAU(TIMMIES) - 01.31 Columns 00001 00
Command ==> Scroll ==>
***** Top of Data *****
000001 /* REXX EXEC
000002 /*
000003 /* TIMMIES FACTOR - COMPOUND INTEREST CALCULATOR
000004 /*
000005 /* AUTHOR: PAUL GAMBLE
000006 /* DATE: OCT 1/2007
000007 /*
000008 /*
000009 /*
000010
000011
000012 say '*****'
000013 say 'Welcome Coffee drinker.'
000014 say '*****'
000015 DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'
000016 say ""
000017 say "What is the price of your coffee?",
000018 "(e.g. 1.58 = $1.58)"
000019 parse pull CoffeeAmt
000020 END
000021
000022 DO WHILE DATATYPE(CoffeeWk) \= 'NUM'
000023 say ""
000024 say "How many coffees a week do you have?"
000025 parse pull CoffeeWk
000026 END
000027
000028 DO WHILE DATATYPE(Rate) \= 'NUM'
000029 say ""
000030 say "What annual interest rate would you like to see on that money?",
000031 "(e.g. 8 = 8%)"
000032 parse pull Rate
000033 END
000034 Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */
```

~80s

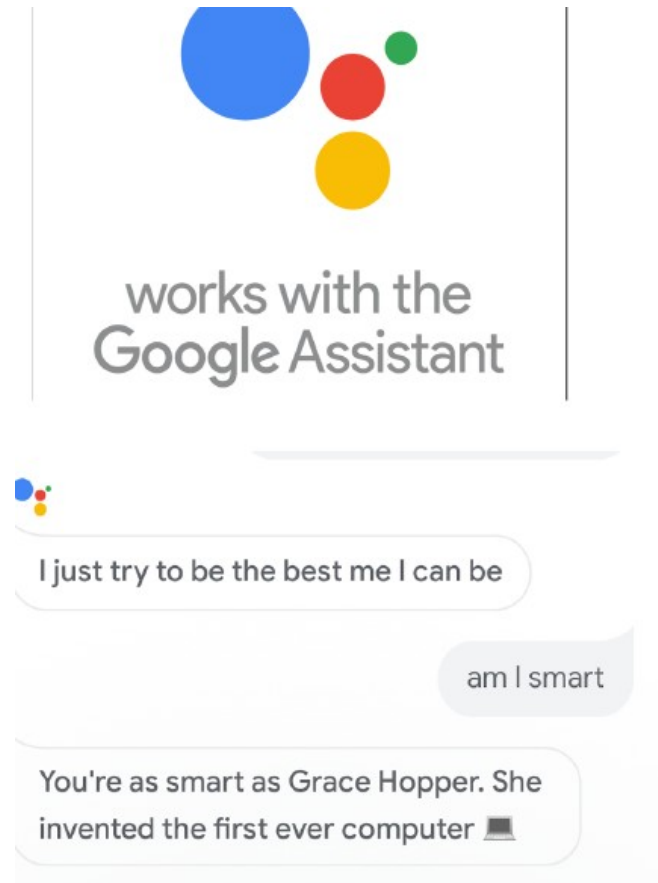


today




# Virtual Assistant

- **Conversational agents contain:**
  - Speech recognition
  - Language analysis
  - Dialogue processing
  - Information retrieval
  - Text to speech
- **Google now, Alexa, Siri, Cortana, VAV...**



# Google Translate & Vietgle Translate



Sign in

Try a new browser with automatic translation. [Download Google Chrome](#) [Dismiss](#)

Translate

From: English To: Vietnamese Translate

English Vietnamese Spanish

Facebook says most of its money comes from online advertising. But the company also says it expects to earn money from fees charged on the sales of virtual goods. These are digital products used in social games, not physical goods. Facebook says it sees important income coming from this new market, which could reach fourteen billion dollars by twenty sixteen.

Vietnamese English Spanish

Facebook cho biết hầu hết tiền của nó đến từ quảng cáo trực tuyến. Nhưng công ty cũng cho biết họ hy vọng sẽ kiếm được tiền từ chi phí tính trên doanh số bán hàng của hàng hóa ảo. Đây là những sản phẩm kỹ thuật số được sử dụng trong các trò chơi xã hội, không phải vật lý hàng hóa. Facebook nói rằng nó thấy thu nhập quan trọng đến từ thị trường mới này, có thể đạt 14000000000 đô la bằng 20 16.

New! Hold down the shift key, click, and drag the words above to reorder. [Dismiss](#)

Tiếng Anh Tiếng Việt

☒ Chung ☐ Tin học ☐ Kế toán ☐ Toán học ☐ Y học ☐ Kỹ thuật

Nội dung cần dịch

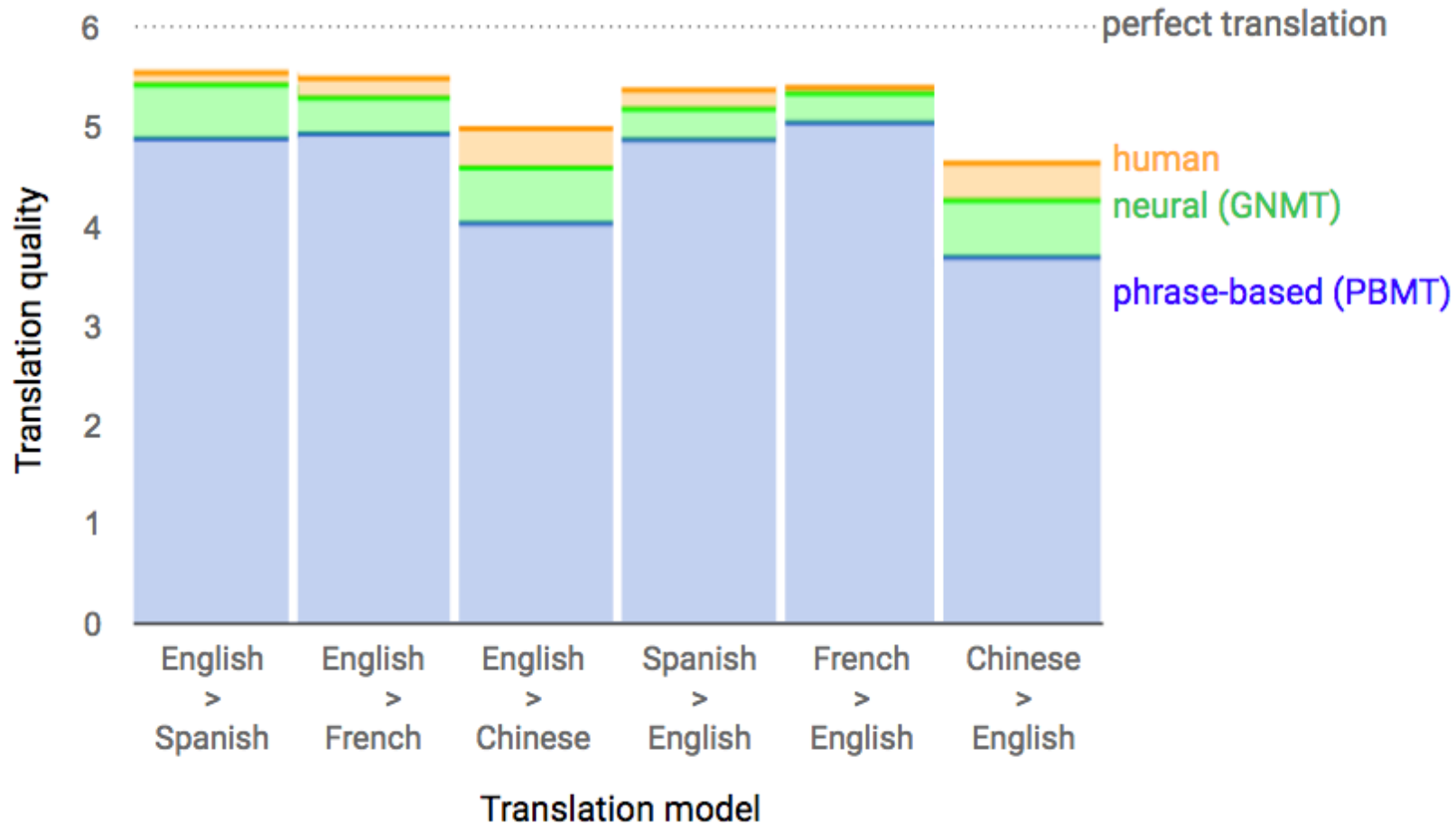
Facebook says most of its money comes from online advertising. But the company also says it expects to earn money from fees charged on the sales of virtual goods. These are digital products used in social games, not physical goods. Facebook says it sees important income coming from this new market, which could reach fourteen billion dollars by twenty sixteen.

Dịch Xóa

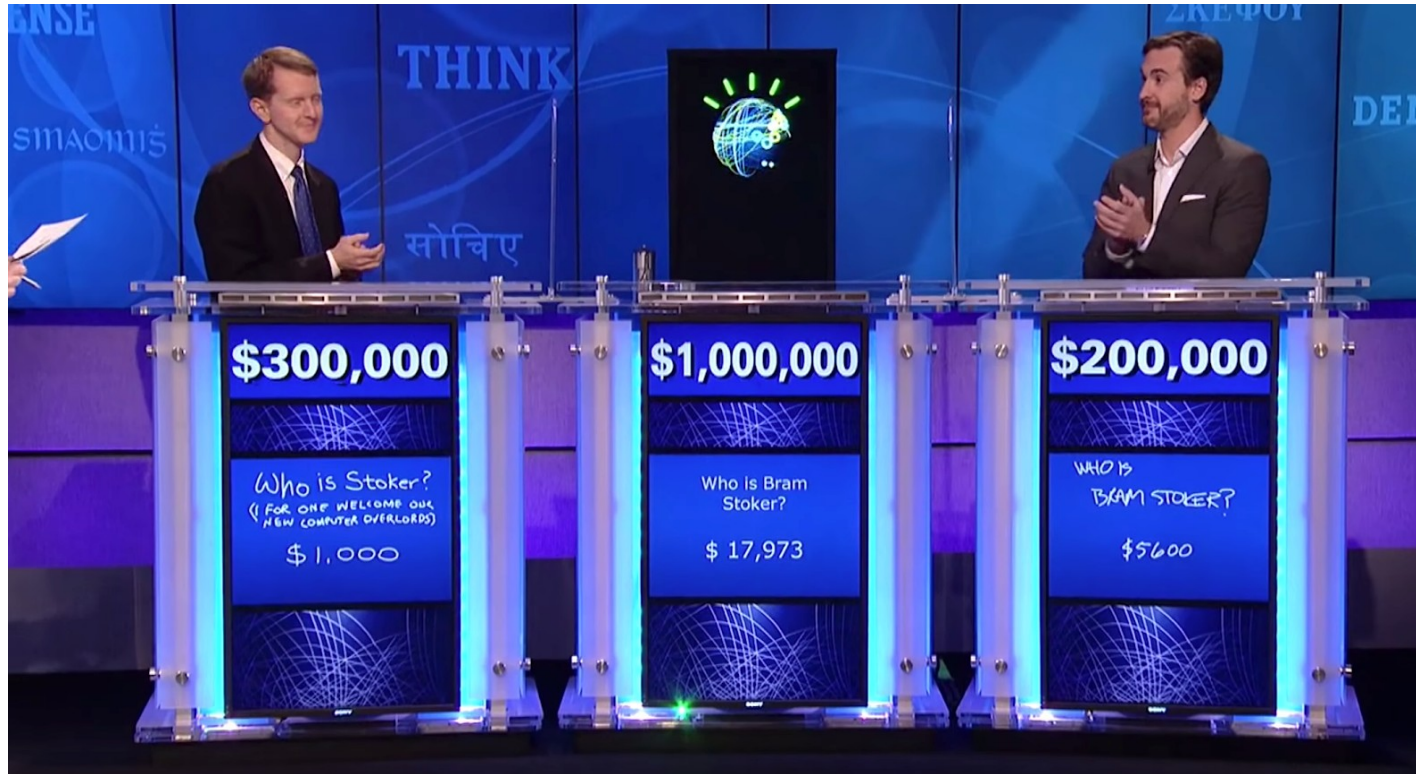
Kết quả

Facebook nói phần lớn tiền của nó đến từ quảng cáo trực tuyến. Nhưng công ty cũng nói nó mong để kiếm được tiền từ phí tính tiền về việc bán hàng hoá ảo. Đây là sản phẩm số được dùng trong trò chơi xã hội, không hàng hoá cụ thể. Facebook nói nó cho là quan trọng thu nhập đến từ thị trường mới này, có thể đạt đến 14 - tỉ đô-la bằng 20 - 16.

# Machine Translation vs. Human



# Watson system -IBM 2011 (Question-Answering )



- IBM built a computer that **won Jeopardy in 2011**
- Question answering technology built on **200 million text pages**, encyclopedias, dictionaries, thesauri, taxonomies, ontologies, and other databases

# Google's Knowledge Graph

- Goal: move beyond keyword search document retrieval to directly
  - easier for mobile device users
- Google's Knowledge Graph (Knowledge Graph ("*things not strings*")):
  - built on top of FreeBase
  - entries are synthesised from Wikipedia, news stories, etc.
  - Manually updating

About 96,600,000 results (0.36 seconds)

[Hanoi - Wikipedia, the free encyclopedia](#)

<https://en.wikipedia.org/wiki/Hanoi>

Hanoi (/həˈnoɪ/ or US /heɪˈnoɪ/) is the capital of Vietnam and the country's second largest city. Its population in 2009 was estimated at 2.6 million for urban ...  
[Sơn Tây \(Hanoi\)](#) - [Hanoi Museum](#) - [List of cities in Vietnam](#) - [Temple of Literature](#)

[Hà Nội - Wikipedia tiếng Việt](#)

[https://vi.wikipedia.org/wiki/Hà\\_Nội](https://vi.wikipedia.org/wiki/Hà_Nội) - [Translate this page](#)

Hà Nội là thủ đô của nước Việt Nam từ năm 1976 đến nay, và là thủ đô của nước Việt Nam Dân chủ Cộng hòa từ năm 1946, là thành phố lớn nhất Việt Nam về ...  
[Nhà hát Lớn Hà Nội](#) - [Hoàng Mai](#) - [Tổ chức hành chính tại Hà Nội](#) - [Cột cờ Hà Nội](#)

[Images for Hanoi](#)

[Report images](#)



[More images for Hanoi](#)

[Cổng Giao tiếp điện tử Thành Phố Hà Nội - Cổng GTĐT Hà ...](#)

[hanoi.gov.vn/](http://hanoi.gov.vn/) - [Translate this page](#)

Cung cấp thông tin về du lịch, đầu tư và những sự kiện mới sắp diễn ra.

[In the news](#)



[PICTURES: Vietnam Airlines' first 787-9 arrives in Hanoi](#)

[Flightglobal](#) - 4 hours ago



## Hanoi

Capital of Vietnam

Hanoi, the capital of Vietnam, is known for its centuries-old architecture and a rich culture with Southeast Asian, Chinese and French influences. At its heart is the chaotic Old Quarter, where the narrow streets are roughly arranged by trade. There are many little temples, including Bach Ma, honoring a legendary horse, plus Dong Xuan market, selling household goods and street food.

**Area:** 3,345 km²

**Founded:** 1010

**Weather:** 25°C, Wind E at 13 km/h, 94% Humidity

**Local time:** Monday 10:41 PM

[Points of interest](#)

[View 15+ more](#)



[Hoan Kiem Lake](#)



[Ho Chi Minh Mausoleum](#)



[West Lake](#)



[Vietnam Museum of Ethnology](#)



[Vietnam Military History M...](#)

# Key Applications in 2019

- Computational linguistics (i.e., modeling the human capacity for language computationally)
- Information extraction, especially “open” IE
- Question answering, **chatbot** (e.g., Watson, Google now)
- **Machine translation**
- Summarization
- Opinion and sentiment analysis
- Social media analysis
- Fake News Recognition

# NLP Careers: So hot!

- Industry
- Government
- Academia

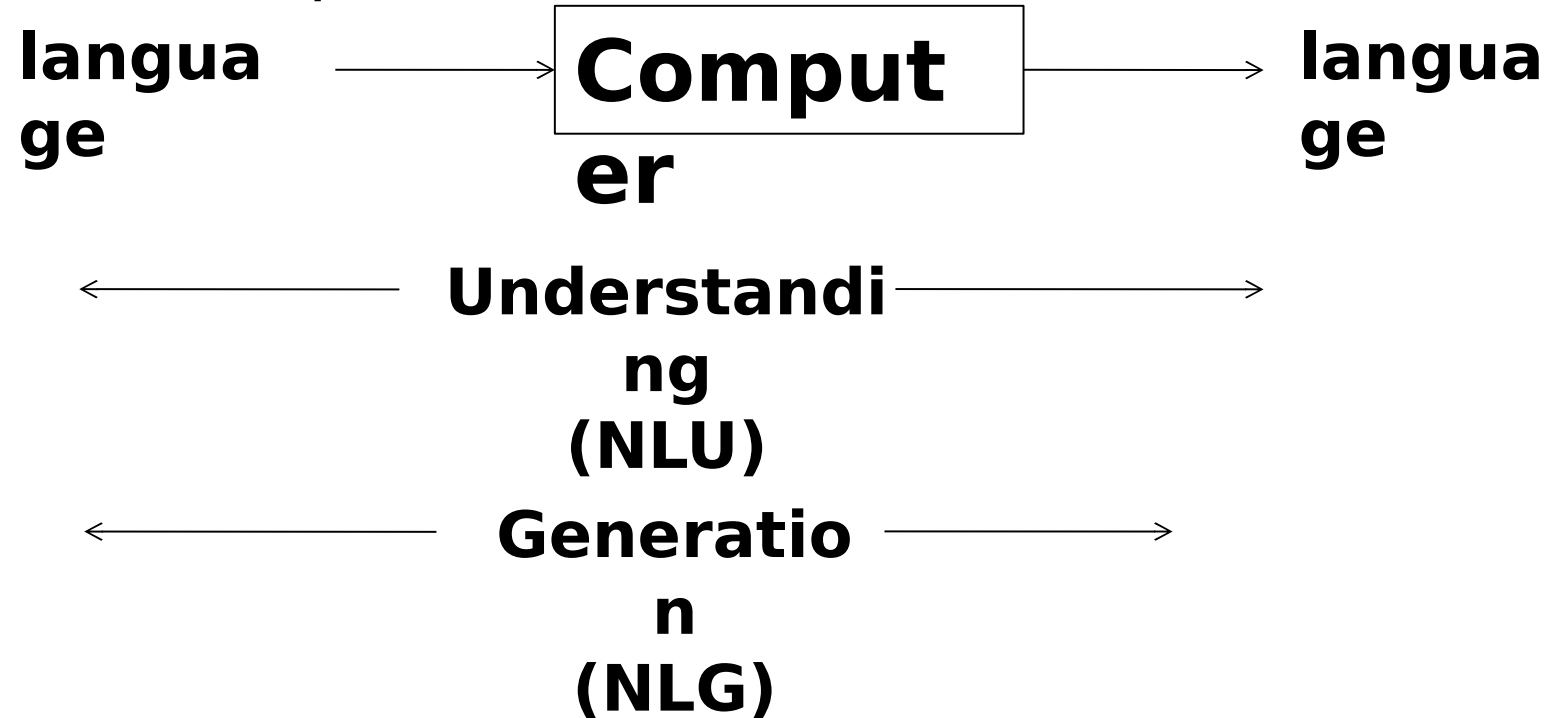


# What is NLP?

- **Natural language processing (NLP)** is a subfield of artificial intelligence and **computational linguistics**. It studies the problems of automated generation and understanding of **natural human languages**.
- **Natural-language-generation systems** convert information from computer databases into normal-sounding human language. **Natural-language-understanding systems** convert samples of human language into more formal representations that are easier for **computer** programs to manipulate.

# What is Natural Language Processing?

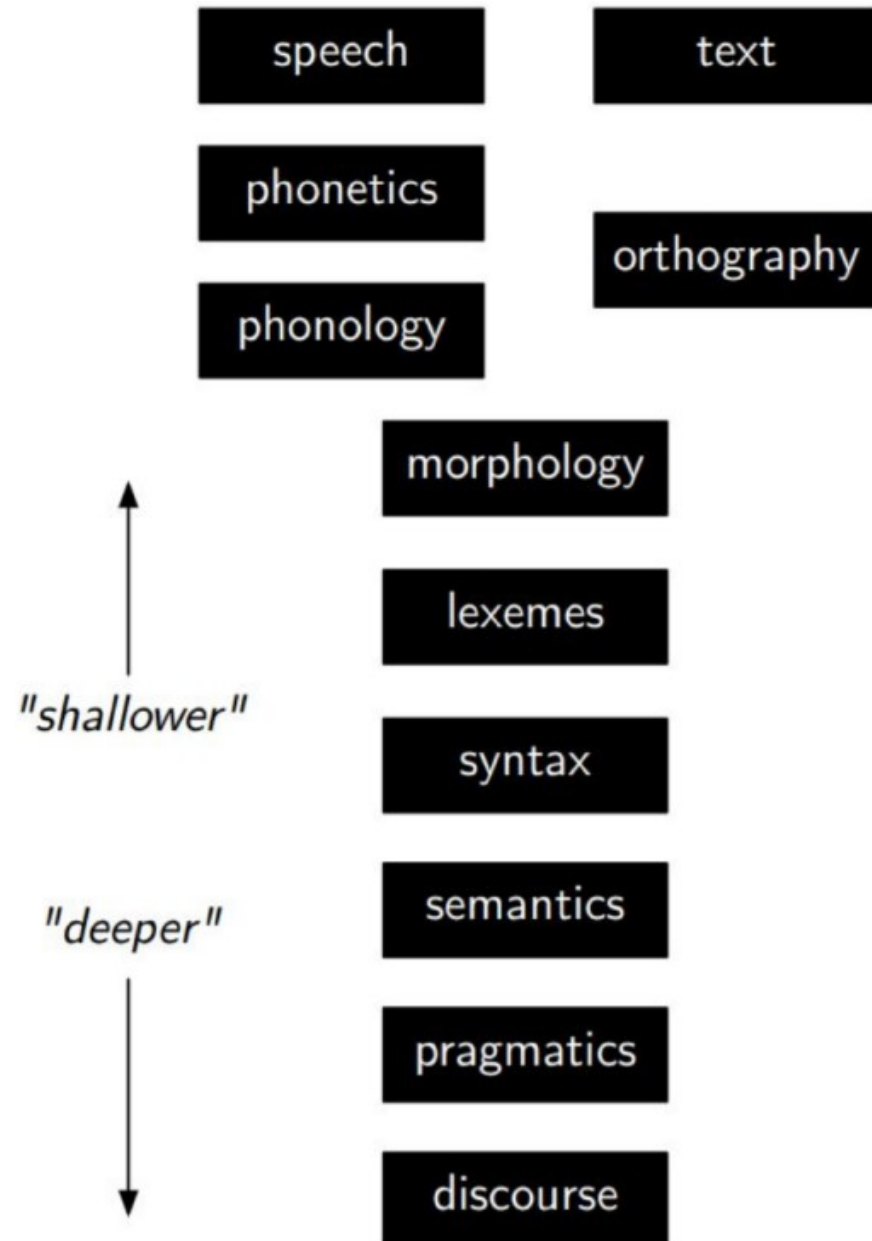
- Computers using natural language as input and/or output



# Natural language processing and computational linguistics

- **Natural language processing (NLP)** develops methods for solving practical problems involving language:
  - Automatic speech recognition
  - Machine Translation
  - Sentiment Analysis
  - Information extraction from documents
- **Computational linguistics (CL)** focused on using technology to support/implement linguistics:
  - how do we understand language?
  - how do we produce language?
  - how do we learn language?

# Level Of Linguistic Knowledge



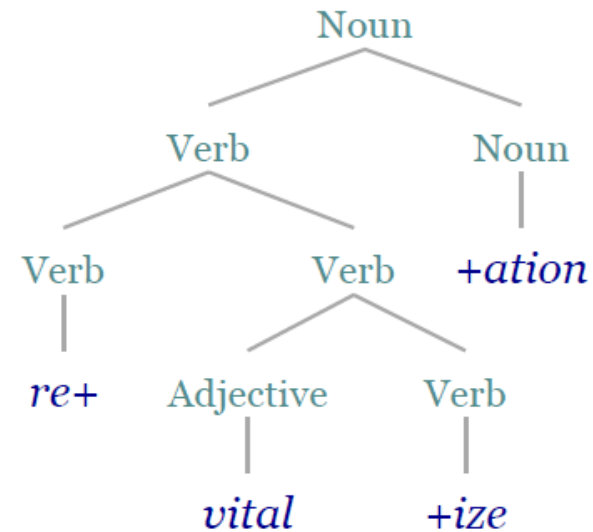
# Phonetics and phonology

- *Phonetics (ngữ âm) studies the sounds of a language*
- *Phonology (âm vị học) studies the distributional properties of these sounds*

# Morphology

- *Morphology studies the structure of words*
- Morphological derivation exhibits hierarchical structure

Example: re+vital+ize+ation

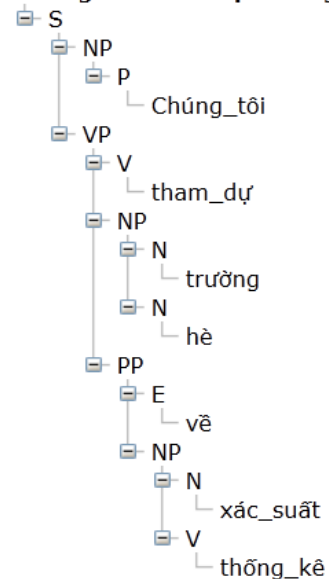


- *The suffix usually determines the syntactic category of the derived word*

# Syntax

- Syntax studies the ways words combine to form phrases and sentences

Chúng tôi tham dự trường hè về xác suất thống kê



- Syntactic parsing helps identify who did what to whom, a key step in understanding a sentence



# Semantics and pragmatics

- Semantics studies the meaning of words, phrases and sentences

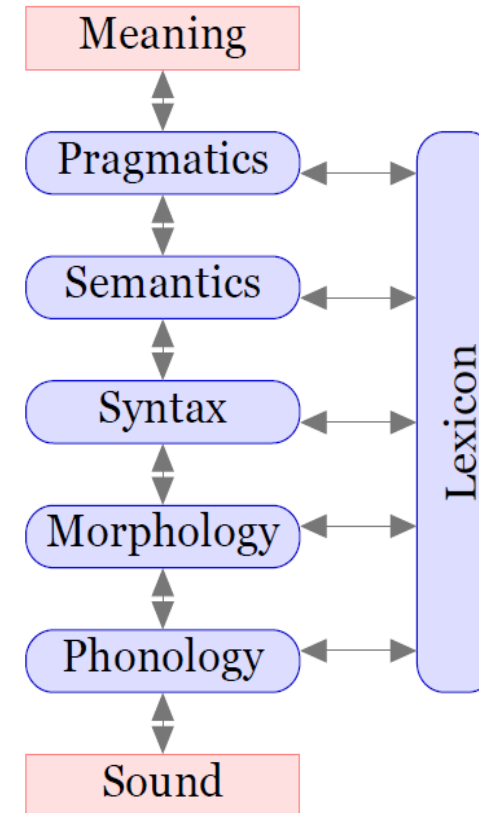
*Ex:* I have a dinner **in/for** an hour

- Pragmatics (Ngữ dụng) studies how we use language to do things in the world

*Ex:* Con vịt chạy đến Mary và liếm chân cô.

# The lexicon

- A language has a lexicon, which lists for each morpheme
  - how it is pronounced (phonology),
  - its distributional properties (morphology and syntax),
  - what it means (semantics), and
  - its discourse properties (pragmatics)
- The lexicon interacts with all levels of linguistic representation



# What's driving NLP and CL research?

- Tools for managing the "information explosion"
  - extracting information from and managing large text document collections
  - NLP is often free tools integrated with main products to sell more ads;  
Ex: speech recognition, machine translation, document clustering (news), etc.
- Mobile and portable computing
  - keyword search / document retrieval don't work well on very small devices
  - we want to be able to talk to our computers (speech recognition) and have them say something intelligent back (NL generation)

# Factors Changing NLP Landscape

- Increases in computing power
- The rise of the web, then the social web
- Advances in machine learning
- Advances in understanding of language in social context

# Natural Language Processing

- **Applications**

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

- **Core Technologies (NLP sub-problems)**

- Language modeling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Word sense disambiguation
- Semantic role labeling
- ...

**NLP lies at the intersection of computational linguistics and machine learning.**

# Why is NLP difficult?

- Ambiguity
- Sparsity
- Abstractly, most NLP applications can be viewed as prediction problems
  - Should be able to solve them with Machine Learning
- The label set is often the set of all possible sentences
  - infinite (or at least astronomically large)
- Training data for supervised learning is often not available
  - Unsupervised/semi-supervised techniques for training from available data
- Algorithmic challenges
  - vocabulary can be large (e.g., 50K words)
  - data sets are often large (GB or TB)

# Ambiguity ???

**“At last, a computer that understands you like your mother”**

**“Ông già đi nhanh quá”**



# Ambiguity

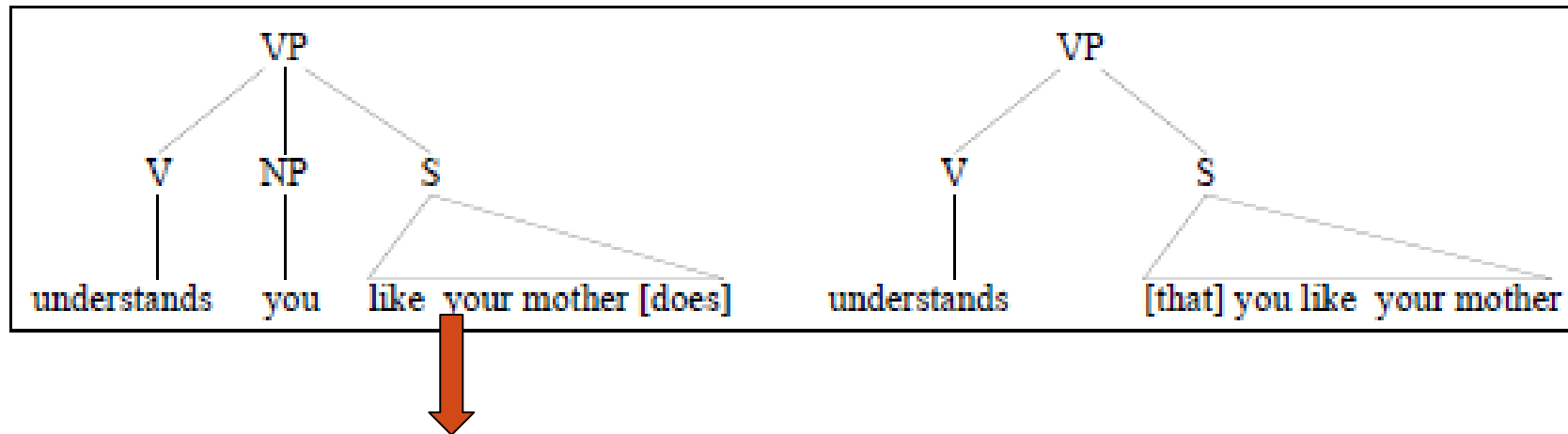
- “At last, a computer that understands you like your mother”
- It understands you as well as your mother understands you
- It understands (that) you like your mother
- It understands you as well as it understands your mother

# Ambiguity at Many Levels

- At the acoustic level (speech recognition):
- “... a computer that understands you like your mother”
- “... a computer that understands you lie cured mother”

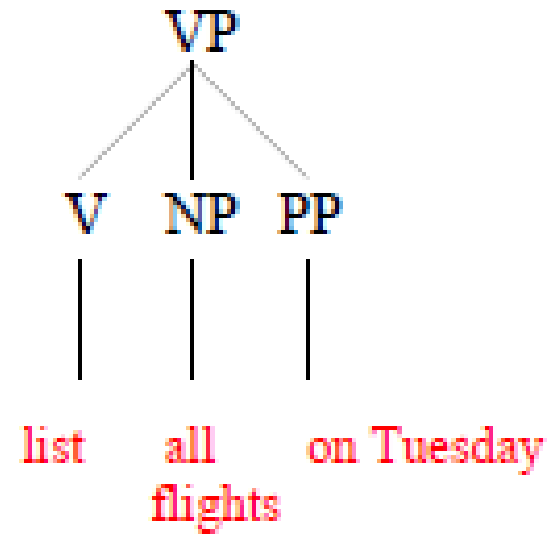
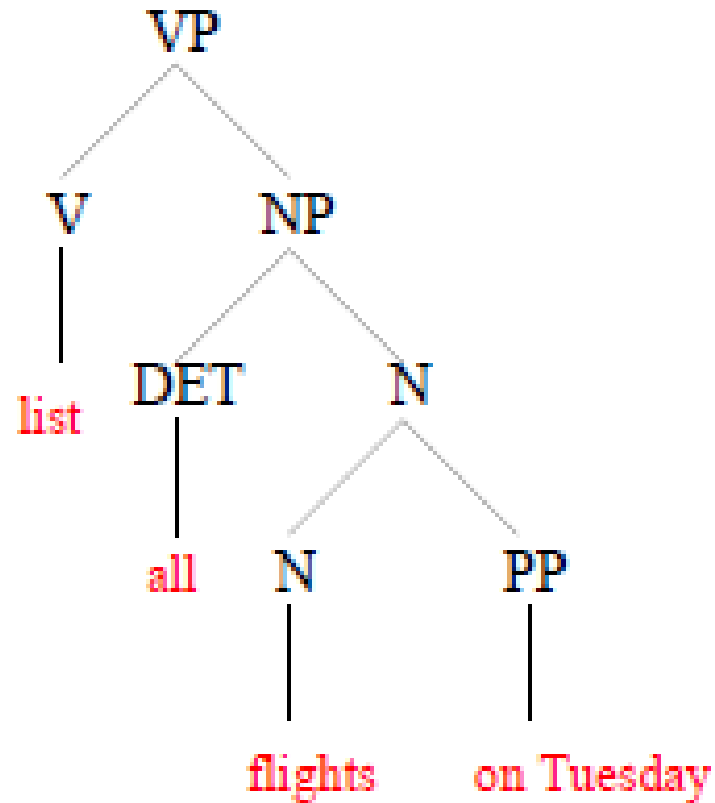
# Ambiguity at Many Levels

- At the **syntactic** level:



**Different structures lead to different interpretations**

# More Syntactic Ambiguity



# Ambiguity at Many Levels

- At the **semantic** (meaning) level:
  - Two definitions of “bank”
    - an organization where people and businesses can invest or borrow money, change it to foreign money, etc., or a building where these services are offered
    - sloping raised land, especially along the sides of a river
- This is an instance of **word sense ambiguity**

# More Word Sense Ambiguity

- At the **semantic** (meaning) level:
  - They put money in the bank
  - I saw her duck with a telescope

# Dealing with Ambiguity

- **How can we model ambiguity?**
  - Non-probabilistic methods (CKY parsers for syntax) return all possible analyses
  - Probabilistic models (HMMs for POS tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return **the best possible analyses**, i.e., the most probable one.
- But the “best” analysis is only good if our probabilities are accurate. Where do they come from?



# Corpora

- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
- Examples
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of French/English sentences
  - Yelp reviews
  - VLSP Corpus (Vietnamese)

# Statistical NLP

- Like most other parts of AI, NLP is dominated by statistical methods
- Typically more robust than rule-based methods
- Relevant statistics/probabilities are learned from data
- Normally requires lots of data about any particular phenomenon

# Sparsity

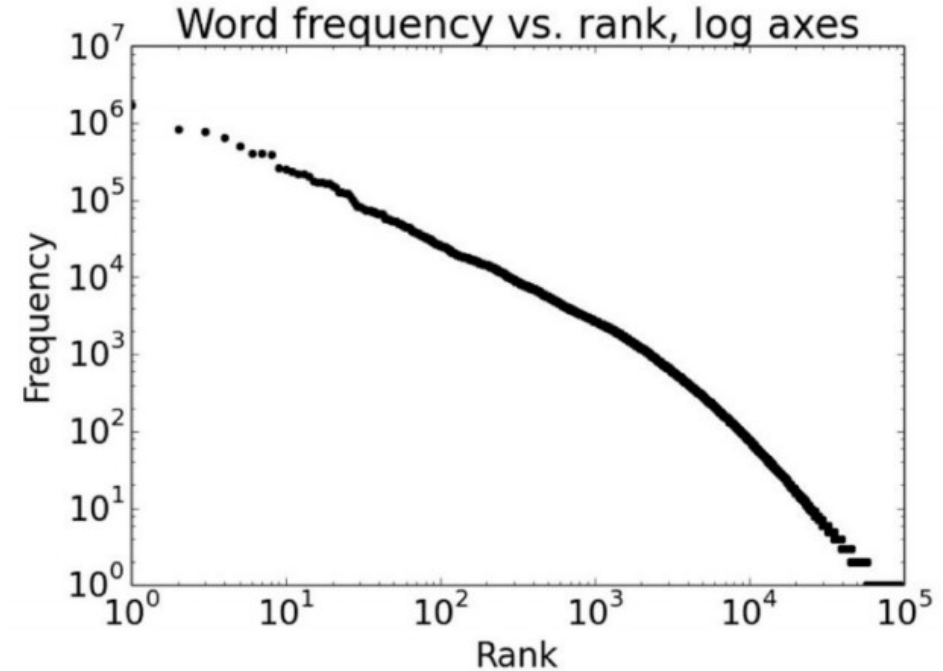
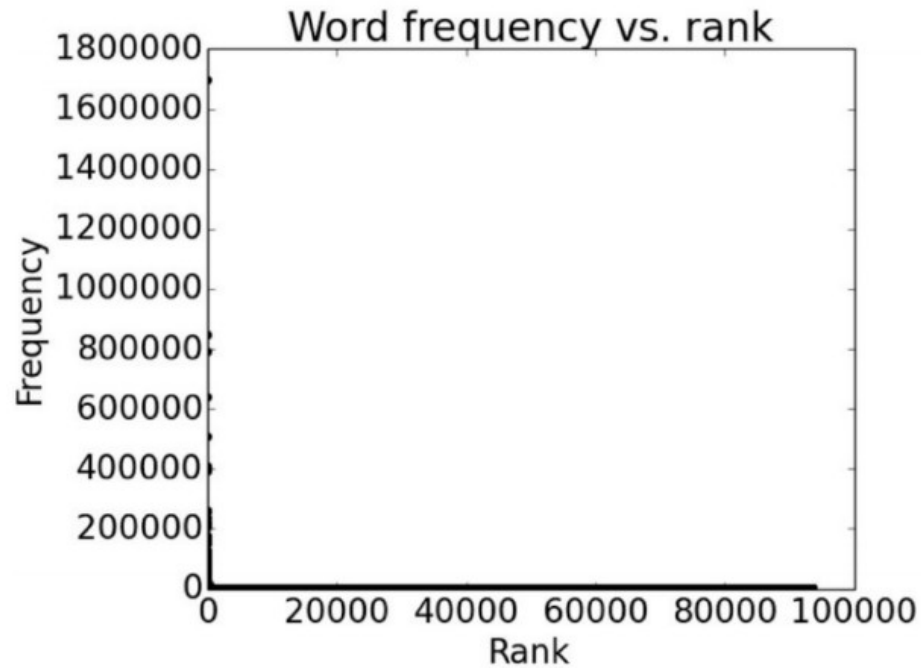
- Sparse data due to **Zipf's Law**
- Example: the frequency of different words in a large text corpus

any word	
Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

nouns	
Frequency	Token
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

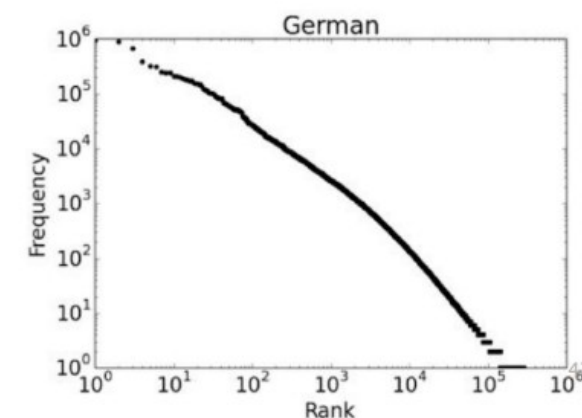
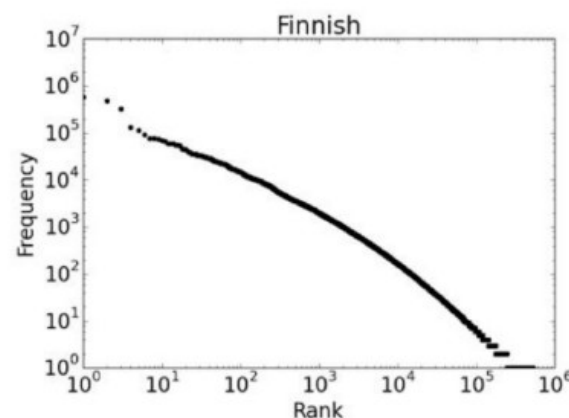
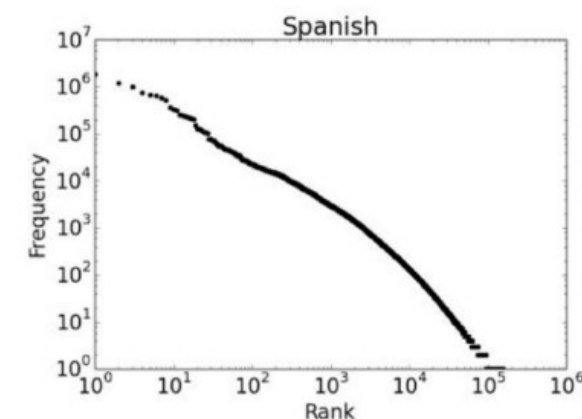
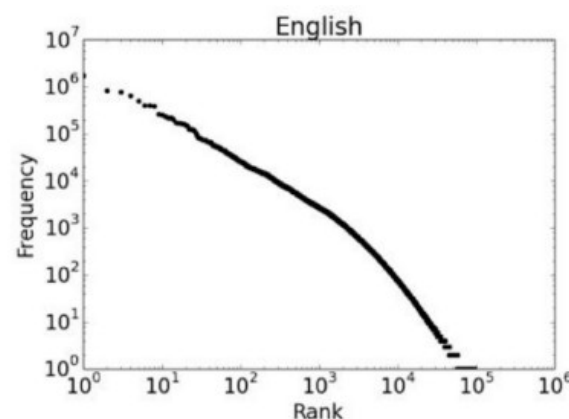
# Sparsity

- Order words by frequency. What is the frequency of  $n$ th ranked word?



# Sparsity

- Regardless of how large our corpus is, there will be a lot of infrequent words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen



# Fields with Connections to NLP

- Machine learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Data science
- Political science
- Psychology
- Economics
- Education

# Today's Applications

- Conversational agents
- Information extraction and question answering
- Machine translation
- Summarization
- Opinion and sentiment analysis
- Social media analysis
- Visual understanding
- Essay evaluation
- Mining legal, medical, or scholarly literature
- ...

# What is this course?

- **Linguistic Issues**

- What are the range of language phenomena?
- What are the knowledge sources that let us disambiguate?
- What representations are appropriate?
- How do you know what to model and what not to model?

- **Statistical Modeling Methods (almost Machine Learning)**

- Increasingly complex model structures
- Learning and parameter estimation
- Efficient inference: dynamic programming, search
- Deep neural networks for NLP: LSTM, CNN, Seq2seq, Transformer



# Outline of Topics

- Words and Sequences
  - Text classifications
  - Probabilistic language models
  - Vector semantics and word embeddings
  - Sequence labeling: POS tagging, NER
  - HMM
- Parsers
- Semantics
- Applications
  - Machine translation, Question Answering, Dialog Systems

# Goals of this Course

- Learn about the problems and possibilities of natural language analysis:
  - What are the major issues?
  - What are the major solutions?
- At the end you should:
  - Agree that language is difficult, interesting and important
  - Be able to assess language problems
    - Know which solutions to apply when, and how
    - Feel some ownership over the algorithms
  - Be able to use software to tackle some NLP language tasks
  - Know language resources
  - Be able to read papers in the field

# Journal and Conference in NLP

- <http://anthology.aclweb.org/>

## ACL events

---

CL: [Intro](#) [FS](#) [MT&CL](#) [74-79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#)  
TACL: [15](#) [14](#) [13](#)  
ACL: [Intro](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84](#)\* [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#)\* [98](#)\* [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#)\* [07](#) [08](#)\* [09](#)\* [10](#) [11](#) [12](#) [13](#) [14](#) [NEW](#) [15](#)\*  
EACL: [Intro](#) [83](#) [85](#) [87](#) [89](#) [91](#) [93](#) [95](#) [97](#)\* [99](#) [03](#) [06](#) [09](#) [12](#) [14](#)  
NAACL: [Intro](#) [00](#)\* [01](#) [03](#) [04](#) [06](#)\* [07](#)\* [09](#)\* [10](#)\* [12](#)\* [13](#)\* [15](#)\*  
EMNLP: [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#)\* [08](#) [09](#) [10](#) [11](#) [12](#)\* [13](#) [14](#)  
CoNLL: [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#) [12](#) [13](#) [14](#) [NEW](#) [15](#)  
\*Sem/  
SemEval: [98](#) [01](#) [04](#) [07](#) [10](#) [12](#) [13](#) [14](#) [15](#)  
ANLP: [Intro](#) [83](#) [88](#) [92](#) [94](#) [97](#) [00](#)\*  
Workshops: [90](#) [91](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#) [12](#) [13](#) [14](#) [UPDATED](#) [15](#)  
SIGs: [ANN](#) [UPDATED](#) [BIOMED](#) [DAT](#) [DIAL](#) [FSM](#) [GEN](#) [UPDATED](#) [HAN](#) [UPDATED](#) [HUM](#) [LEX](#) [MEDIA](#) [UPDATED](#) [MOL](#) [MT](#) [UPDATED](#) [NLL](#) [UPDATED](#) [PARSE](#) [MOR](#)  
[SEMITIC](#) [SLPAT](#) [WAC](#)

## Other Events

---

COLING: [65](#) [67](#) [69](#) [73](#) [80](#) [82](#) [84](#)\* [86](#) [88](#) [90](#) [92](#) [94](#) [96](#) [98](#)\* [00](#) [02](#) [04](#) [06](#)\* [08](#) [10](#) [12](#) [14](#)  
HLT: [86](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [01](#) [03](#)\* [04](#)\* [05](#) [06](#)\* [07](#)\* [08](#)\* [09](#)\* [10](#)\* [12](#)\* [13](#)\* [15](#)\*  
IJCNLP: [05](#) [08](#) [09](#)\* [11](#) [13](#) [NEW](#) [15](#)\*  
LREC: [00](#) [02](#) [04](#) [06](#) [08](#) [10](#) [12](#) [14](#)  
PACLIC [95](#) [96](#) [98](#) [99](#) [00](#) [01](#) [02](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [09](#) [10](#) [11](#) [12](#) [13](#) [14](#)  
ALTA [Intro](#) [03](#) [04](#) [05](#) [06](#) [07](#) [08](#) [0](#)  
[14](#)  
RANLP [09](#) [11](#) [13](#)  
JEP/TALN [12](#) [13](#) [14](#)  
/RECITAL

# Conclusion

- Computational linguistics and natural language processing:
  - were originally inspired by linguistics,
  - but now they are almost applications of machine learning and statistics
- We solve these problems using standard methods from machine learning:
  - Define a probabilistic model over the relevant variables
  - Factor the model into small components that we can learn
  - Ex: HMMs, SVM, CRFs and PCFGs
  - End2end: **Deep Learning**

# References

- Slides of NLP course from CMU, Toronto University
- Some Tutorials of NLP