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; 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 attendence + 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

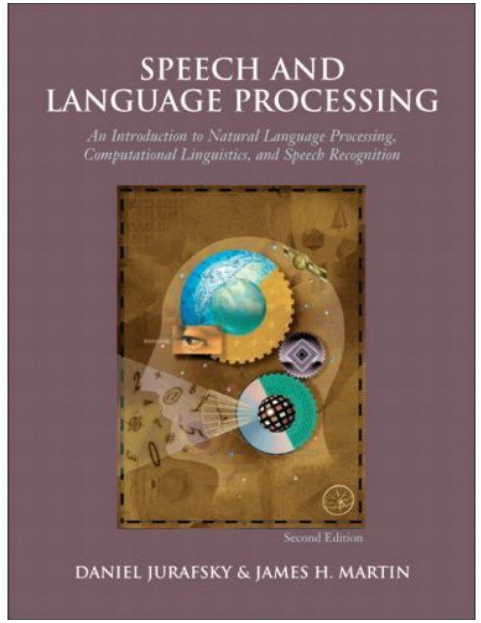
- *Slides*
- **Text books:**

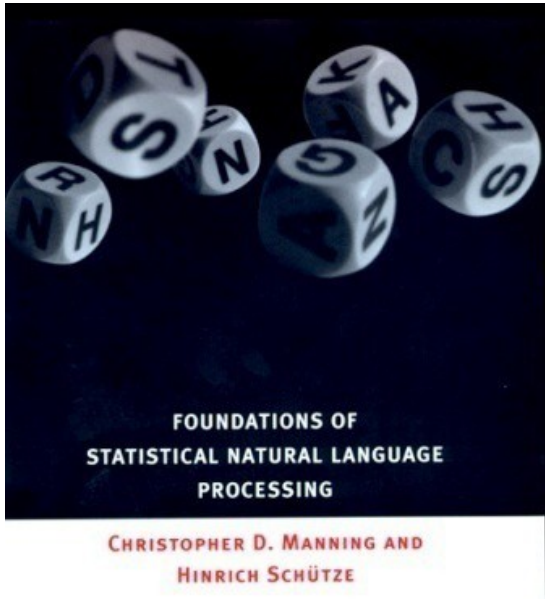
  1) Speech and Language Processing, Daniel Jurasky & James H. Martin, second edition, printed by Prentice Hall, 2009 (https://web.stanford.edu/~jurafsky/slp3/)

  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

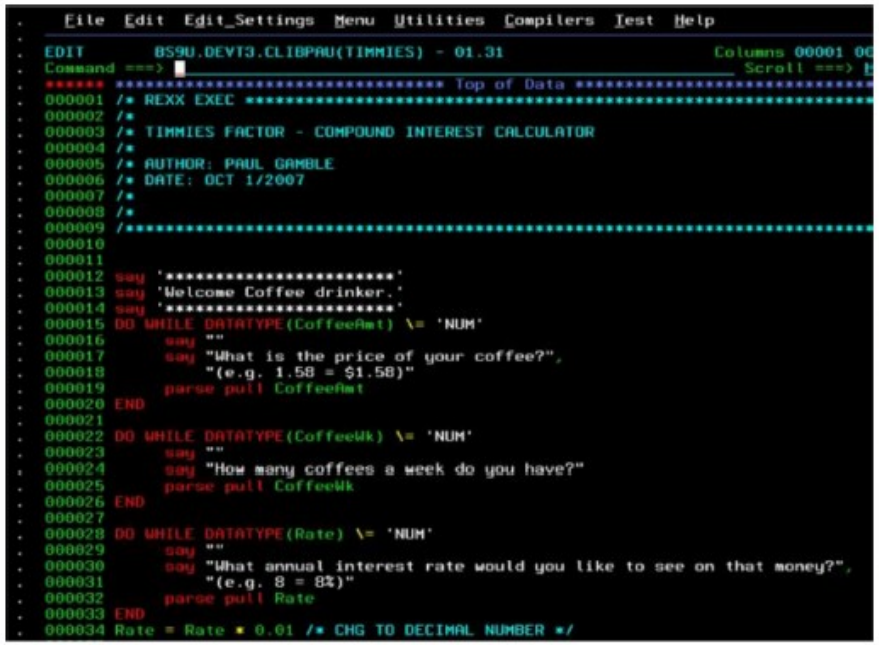




NLP in Industry

Communication With Machines

~50-70s

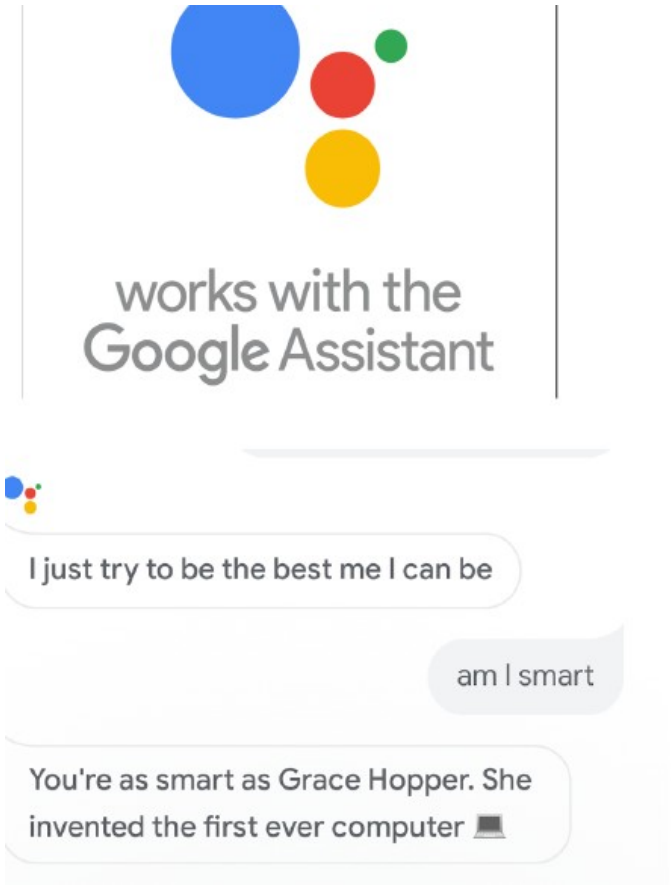


~80s

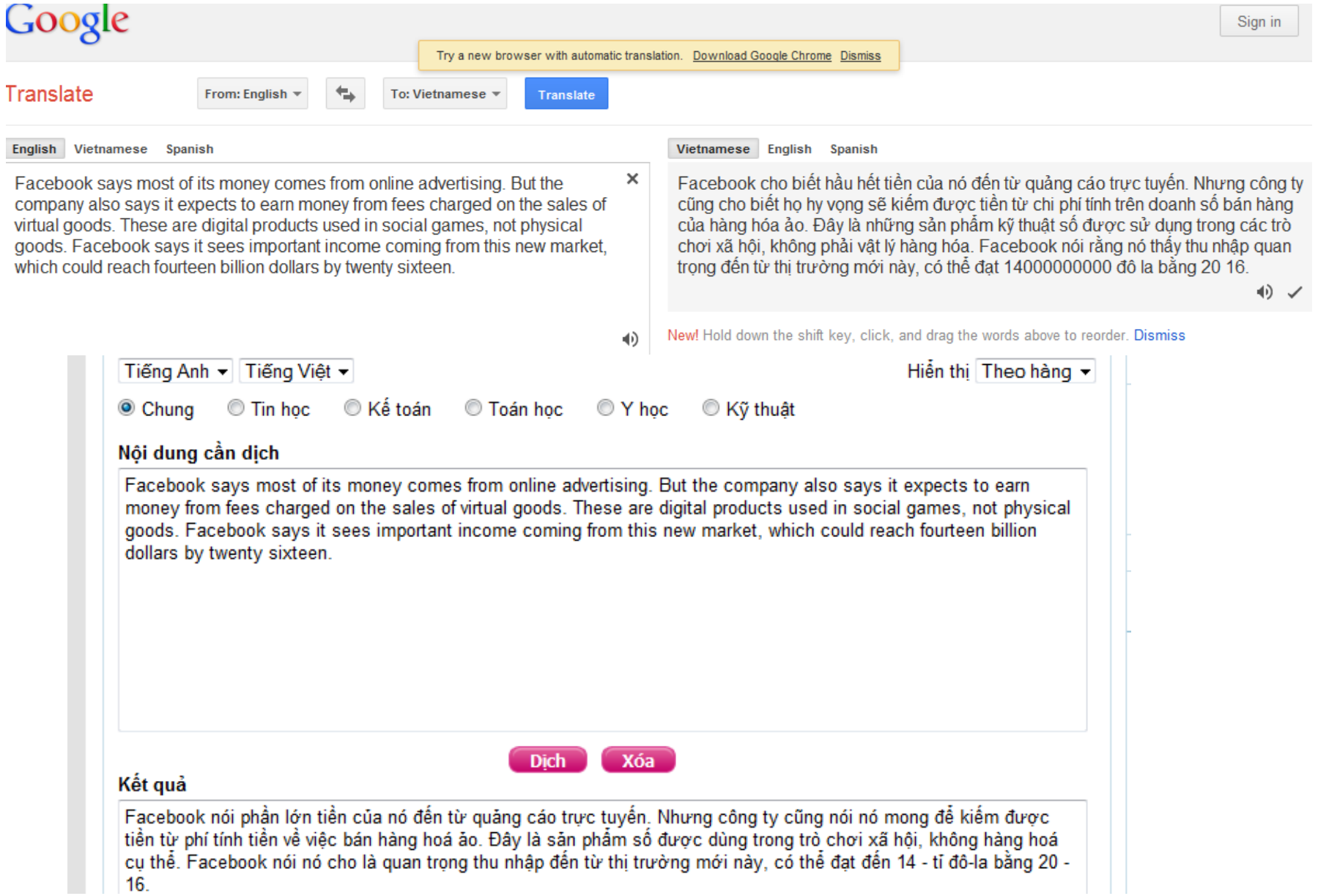
today

Virtual Assistant

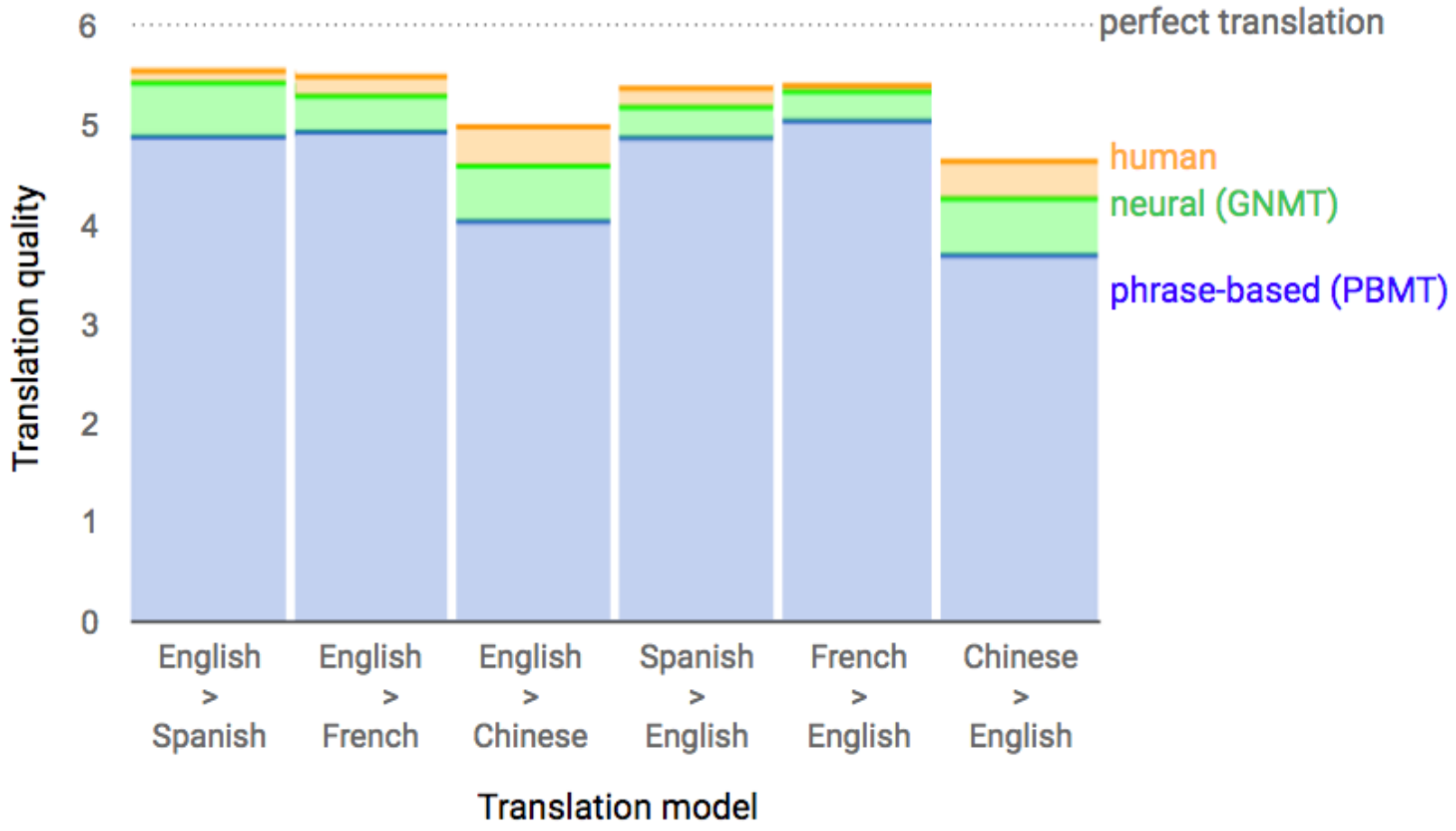
- **Conversational agents contain:**
  - Speech recognition
  - Language analysis
  - Dialogue processing
  - Information retrieval
  - Text to speech
- **Google now, Alexa, Siri, Cortana, VAV…**



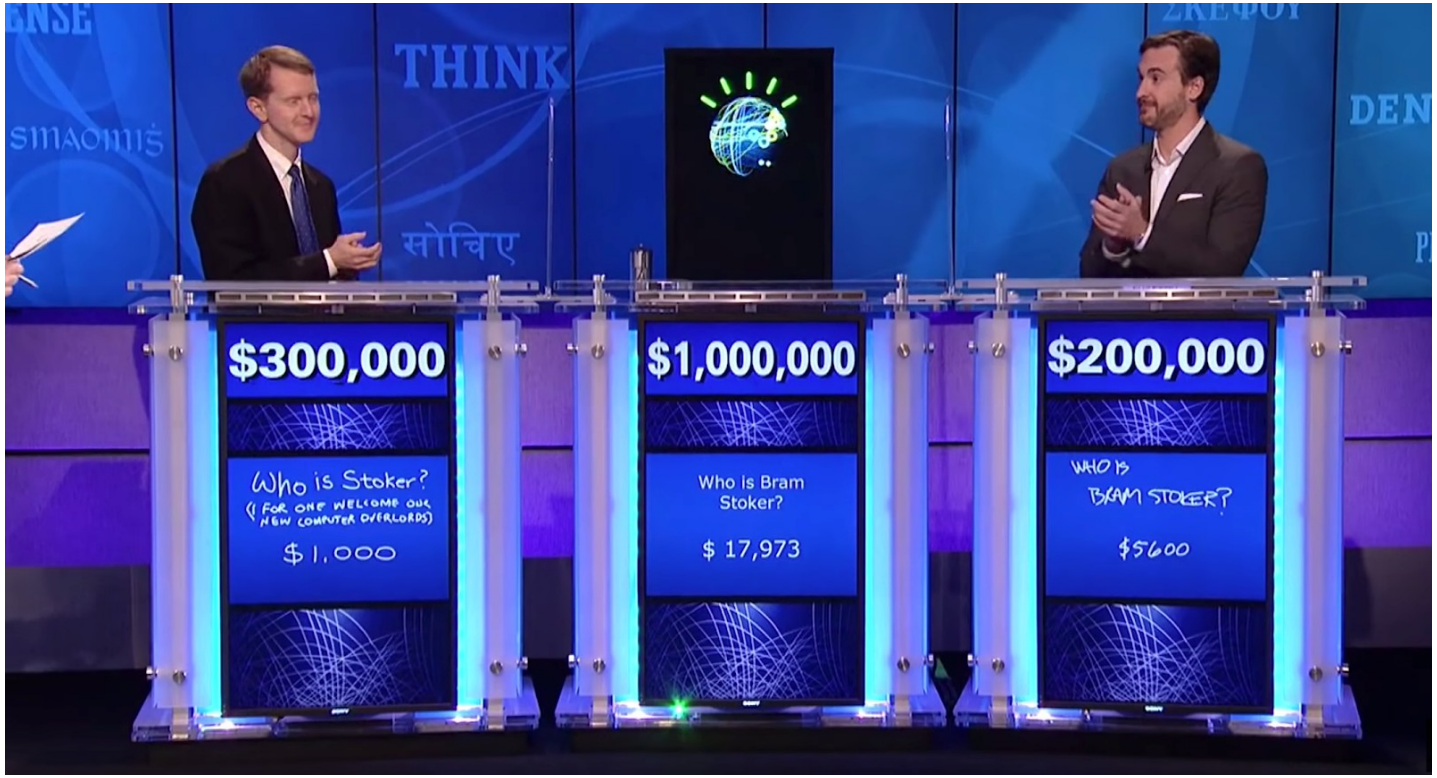
Google Translate & Vietgle Translate



Machine Translation vs. Human



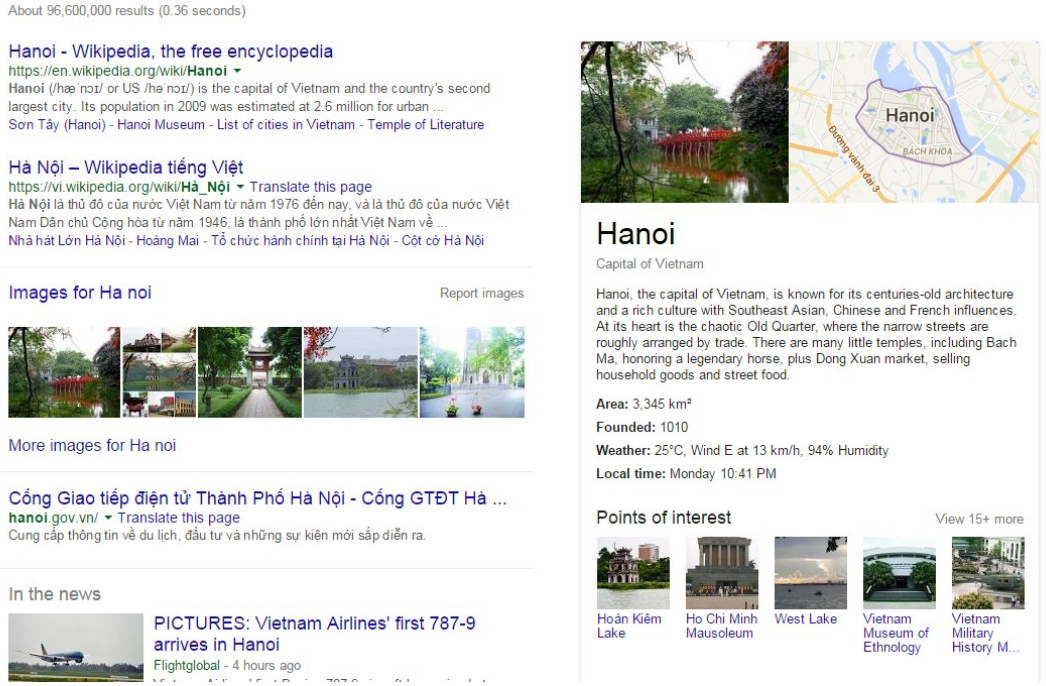
Watson system –IBM 2011 (Question-Answering)



- 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



Key Applicatons in 2019

- Computatonal linguistcs (i.e., modeling the human capacity for language computatonally)
- Informaton extracton, especially “open” IE
- Queston answering, **chatbot** (e.g., Watson, Google now)
- **Machine translaton**
- Summarizaton
- Opinion and sentment analysis
- Social media analysis
- Fake News Recogniton

NLP Careers: So hot!

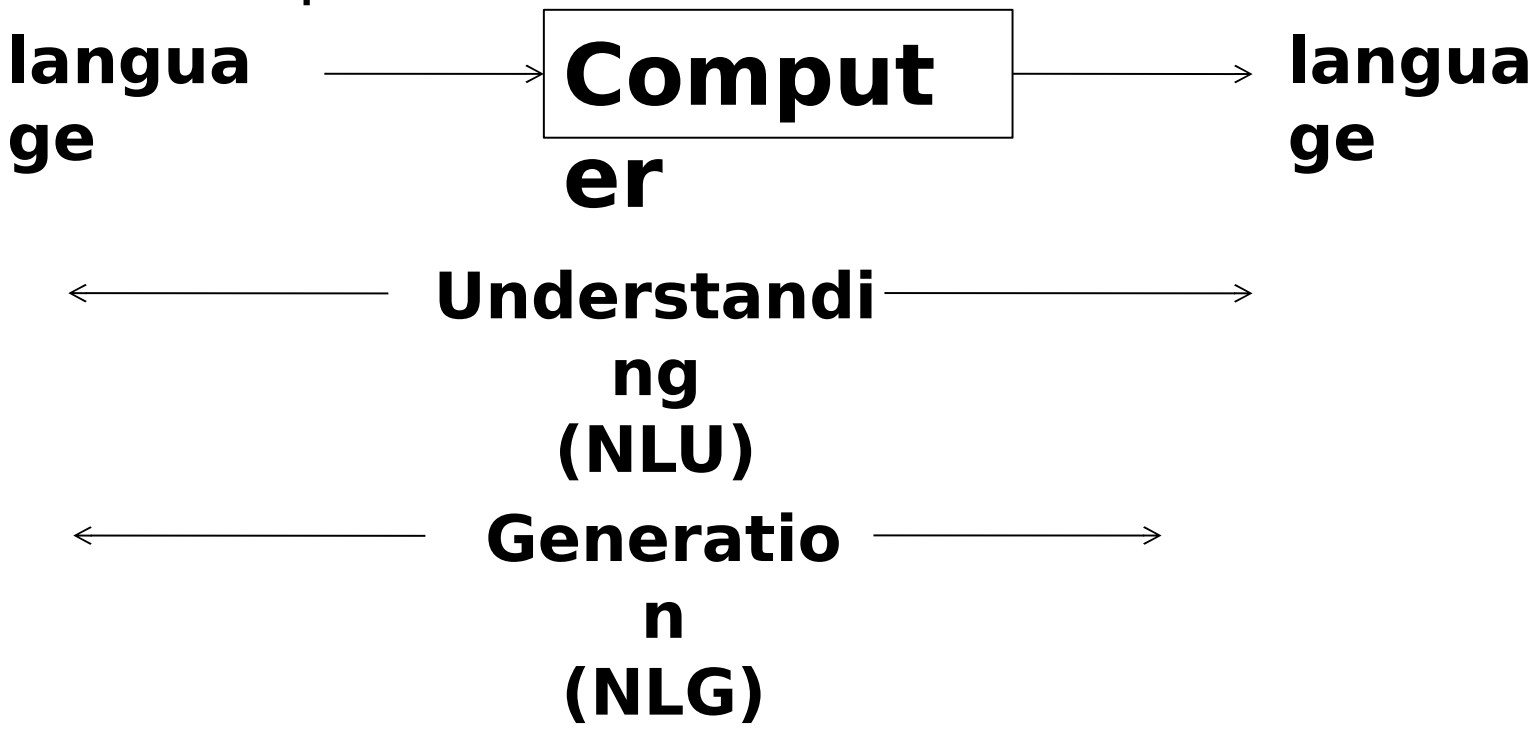
- Industry
- Government
- Academia

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

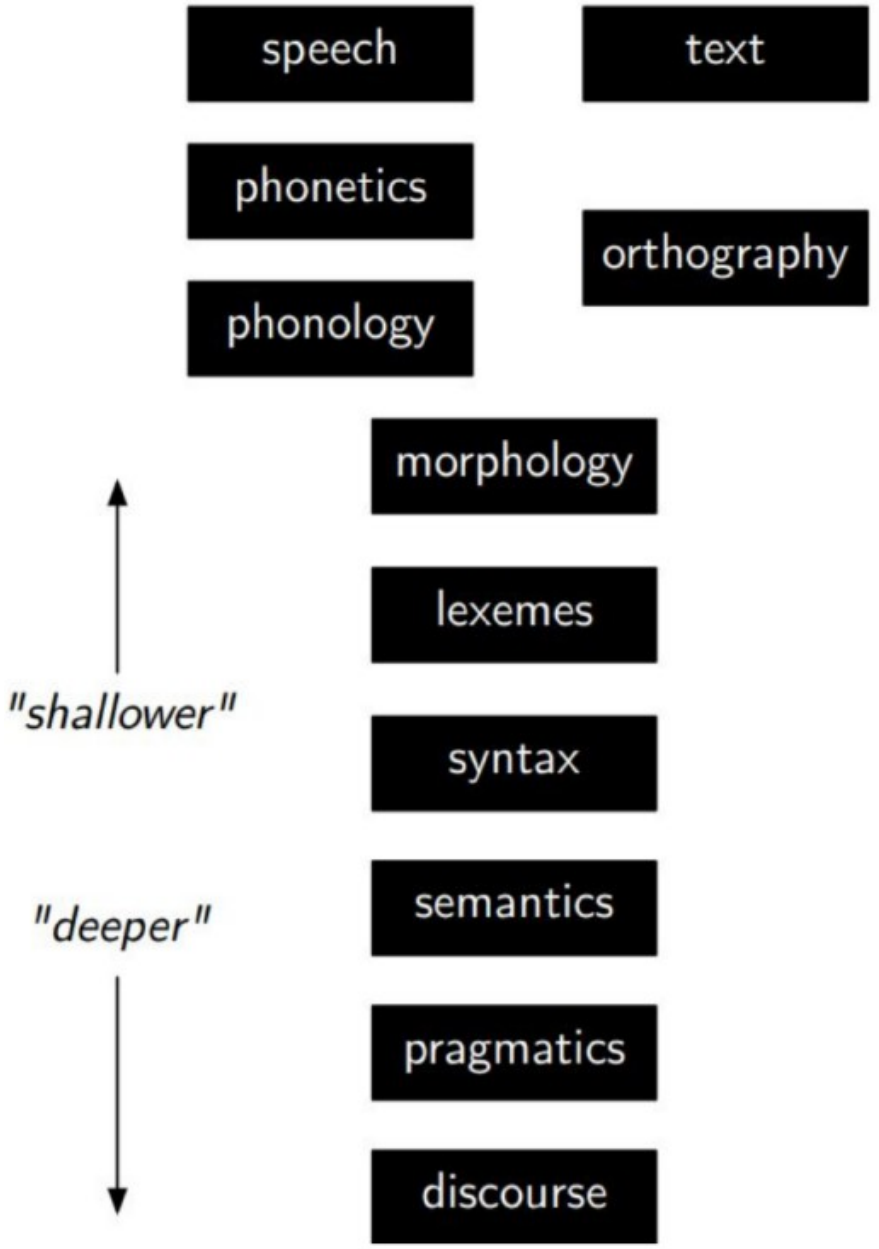


Natural language processing and computational linguistics

- **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

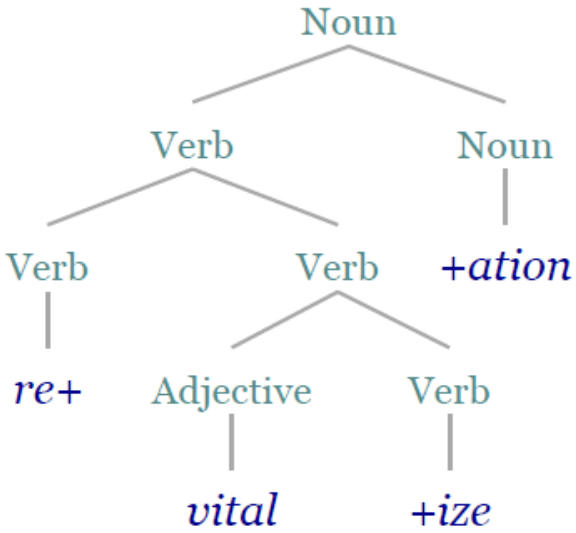
- *Phonetics (ngữ âm) studies the sounds of a language*

- *Phonology (âm vị học) studies the distributional properties of these sounds*

Morphology

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

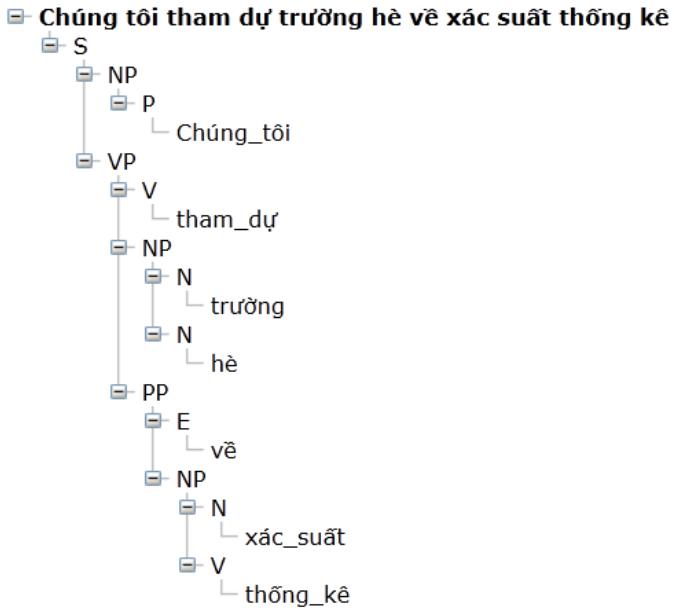
Example: re+vital+ize+ati



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

Syntax

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



- 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

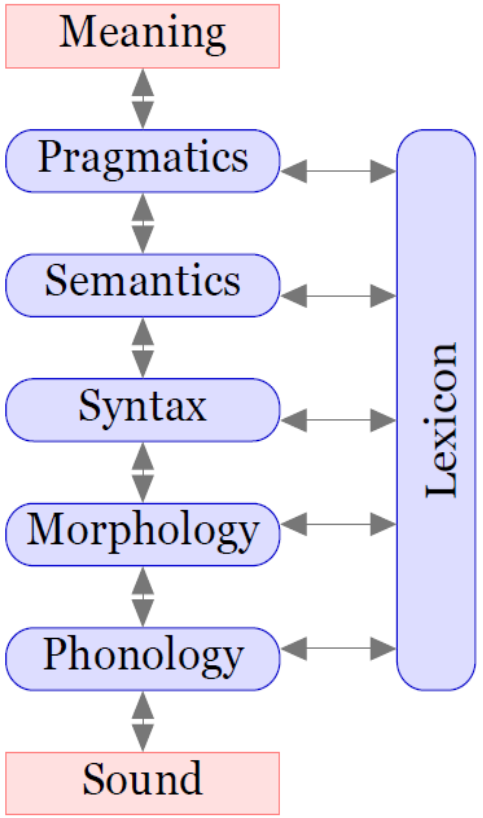
  Ex: I have a dinner in/for an hour

- 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

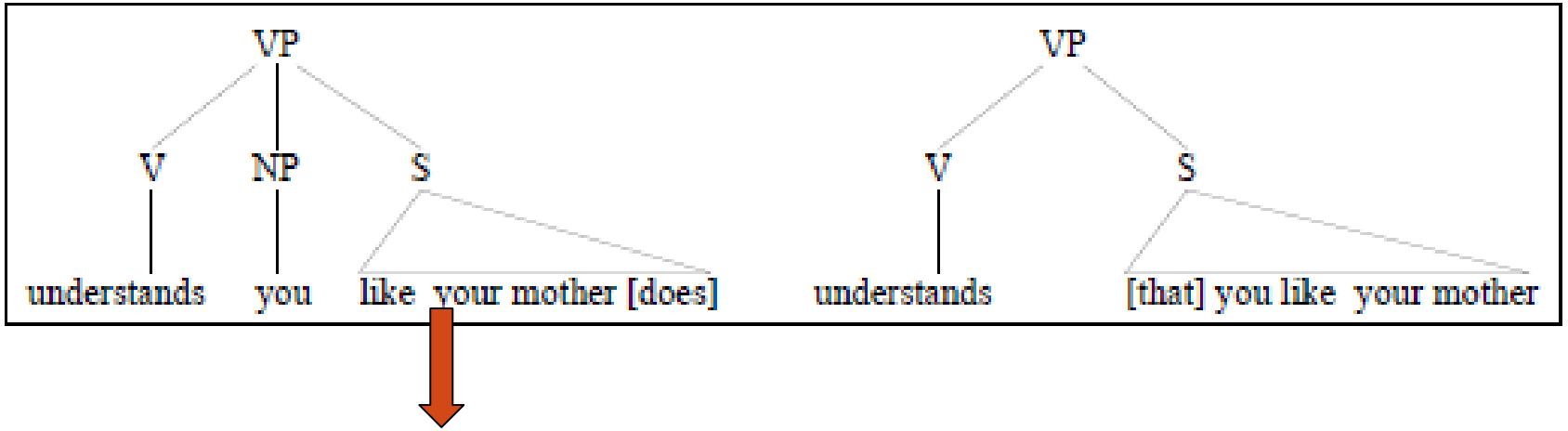
- “**At last, a computer that understands you like your mother**”
- It understands you as well as your mother understands you
- It understands (that) you like your mother
- It understands you as well as it understands your mother

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”

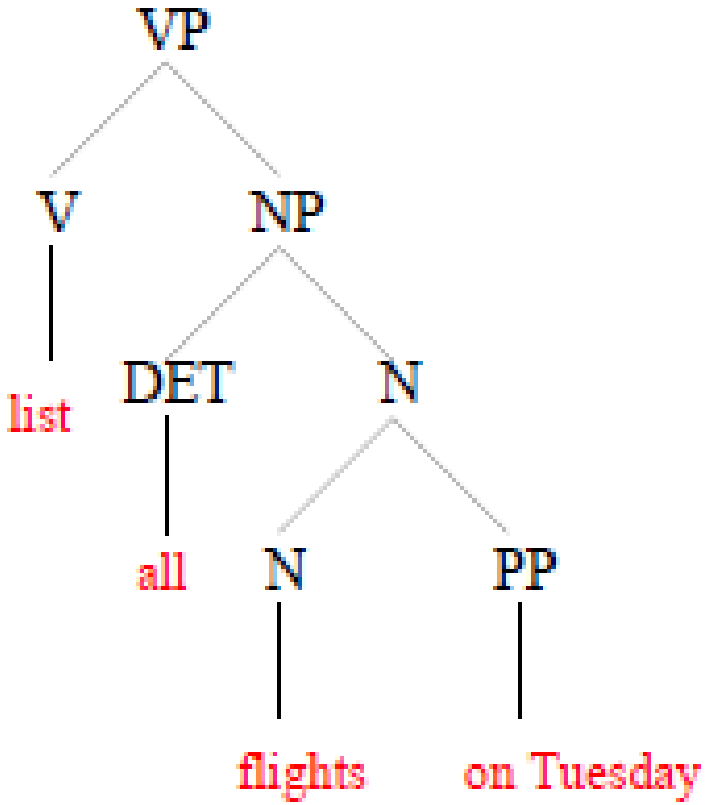
Ambiguity at Many Levels

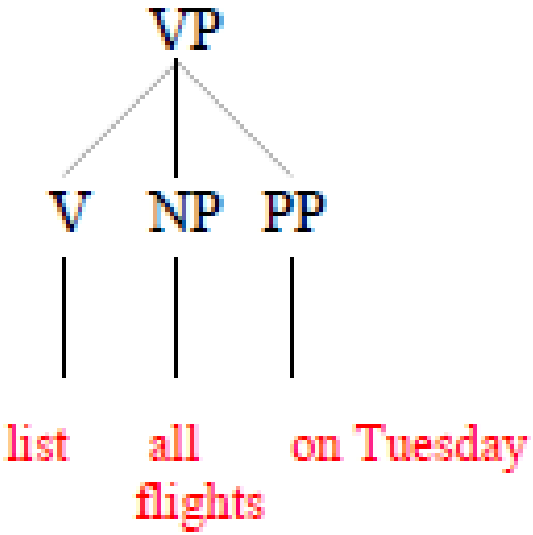
- At the **syntactic** level:



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

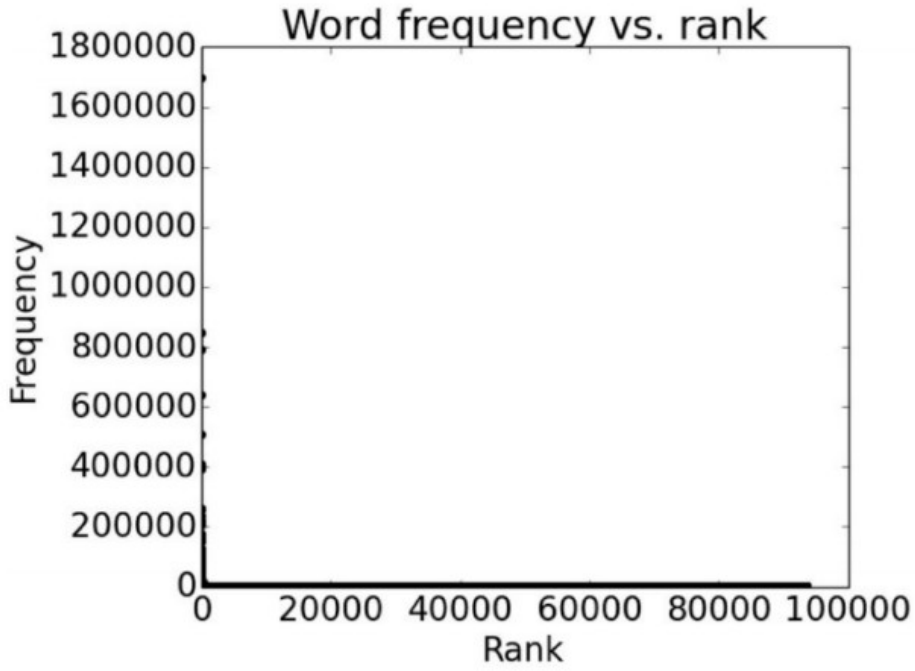
Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

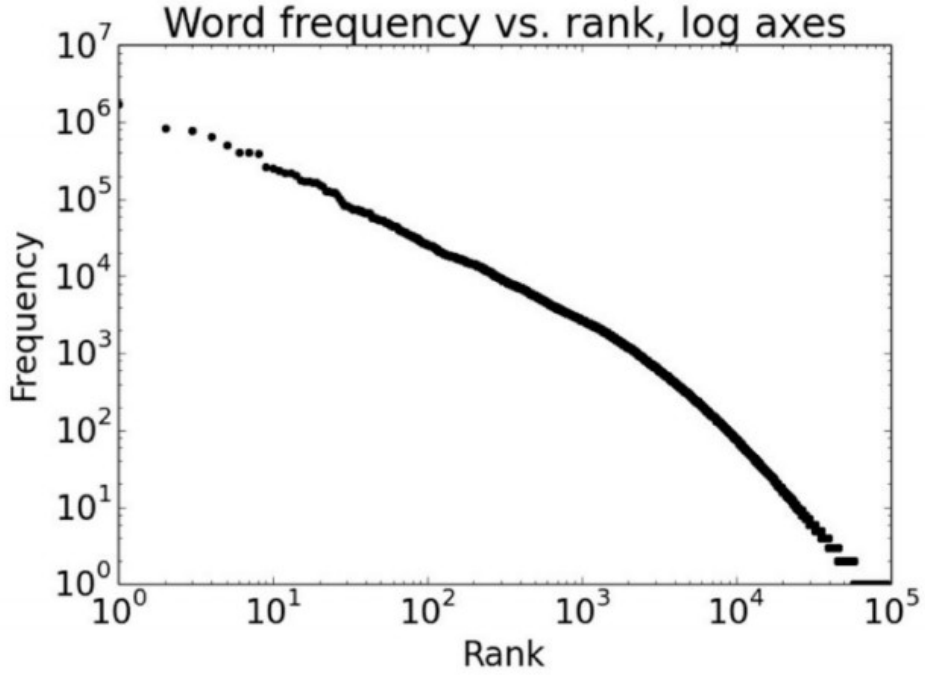
nouns

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 nth ranked word?

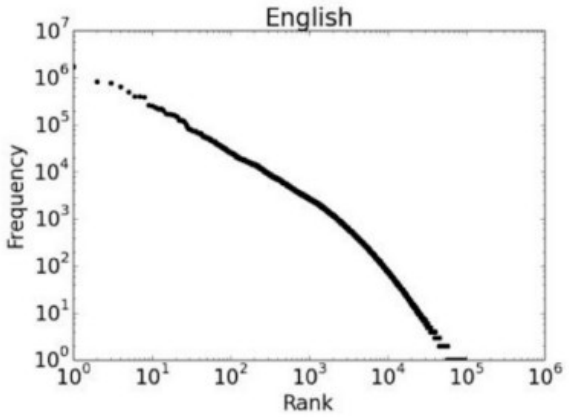


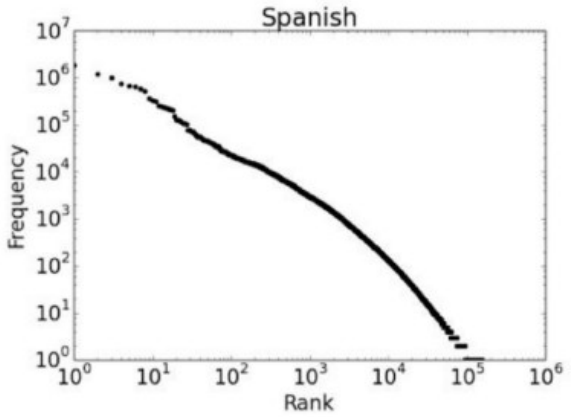


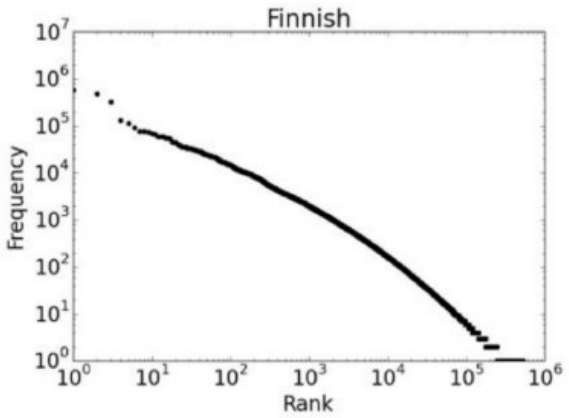
Sparsity

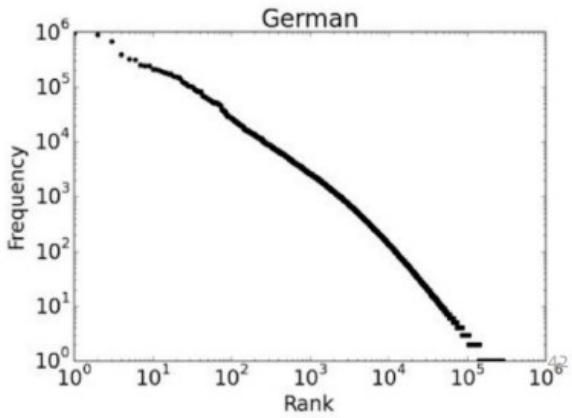
- Regardless of how large our corpus is, there will be a lot of infrequent words

- 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

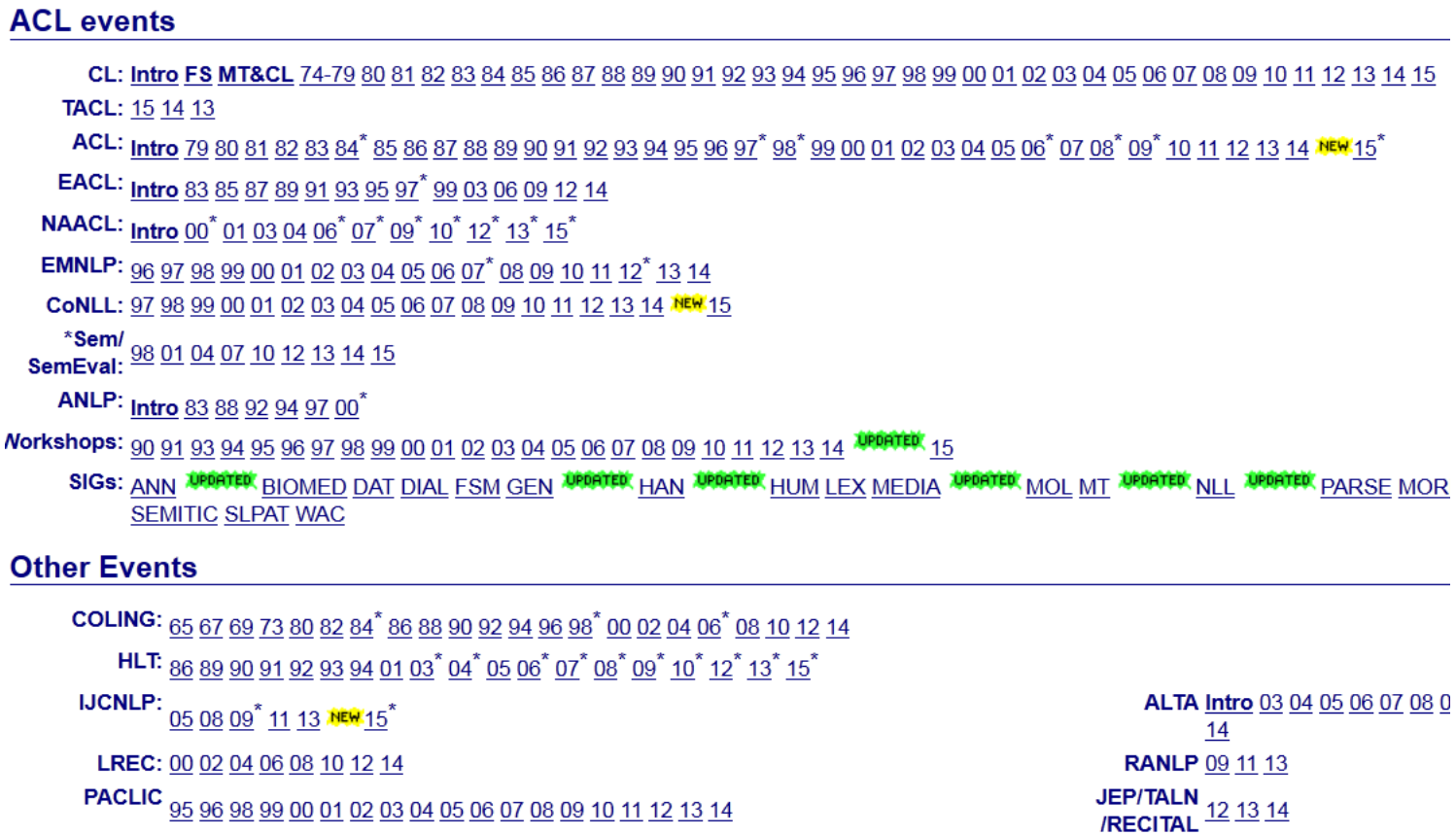
- Words and Sequences
  - Text classifications
  - Probabilistic language models
  - Vector semantics and word embeddings
  - Sequence labeling: POS tagging, NER
  - HMM
- Parsers
- Semantics
- Applications
  - Machine translation, Question Answering, Dialog Systems

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/



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