Natural Language Processing

Berlin Chen 2005

Textbooks & References

- Textbooks
  - C. Manning and H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999
  - D. Jurafsky and J. H. Martin, Speech and Language Processing, Prentice-Hall, 2000

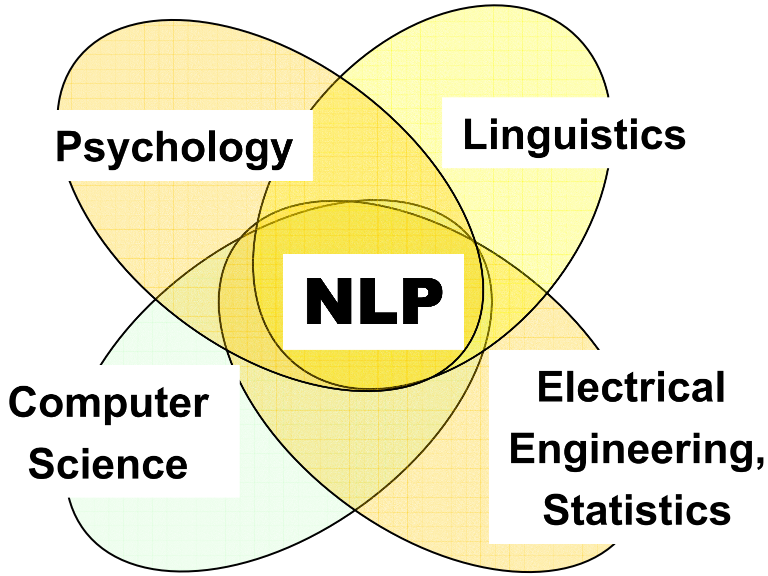
- References
  - J. Allen, Natural Language Understanding, Benjamin/Cummings Publishing Co, 1995
  - X. Huang, A. Acero, H. Hon, Spoken Language Processing, Prentice Hall, 2001

Motivation for NLP

- **Academic**: Explore the nature of linguistic communication
  - Obtain a better understanding of how language work
- **Practical**: Enable effective human-machine communication
  - Conversational agents are becoming an important form of human-computer communication
  - Revolutionize the way computers are used
    - More flexible and intelligent

Motivation for NLP

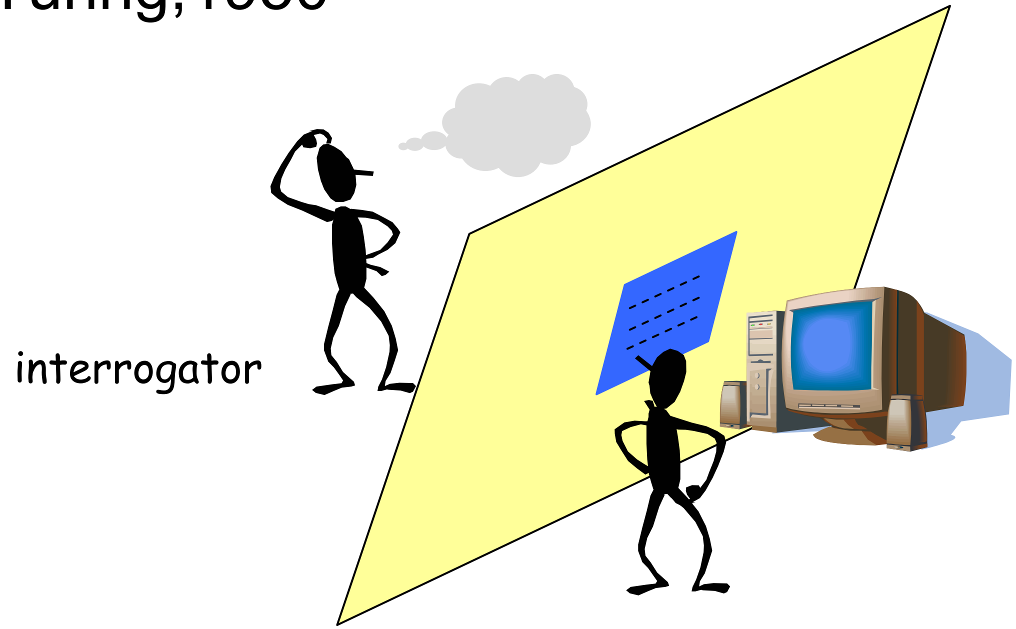
- Different Academic Disciplines: Problems and Methods
  - Electrical Engineering, Statistics
  - Computer Science
  - Linguistics
  - Psychology



- Many of the techniques presented were first develpoed for speech and then spread over into NLP
  - E.g. Language models in speech recognition

Turing Test

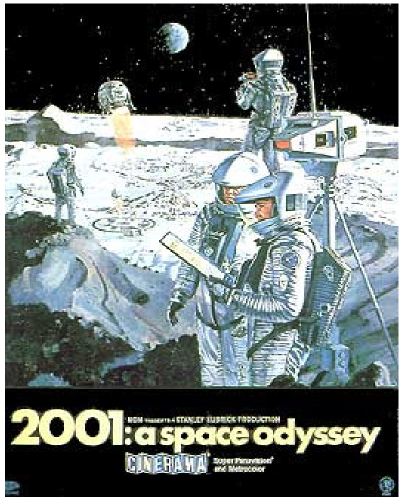
- Alan Turing,1950



  - Predicted at the end of 20 century a machine with 10 gigabytes of memory would have 30% chance of fooling a human interrogator after 5 minutes of questions
    - Does it come true?

Hollywood Cinema

- Computers/robots can listen, speak, and answer our questions
  - E.g.: HAL 9000 computer in “*2001: A Space Odyssey*” (2001太空漫遊)



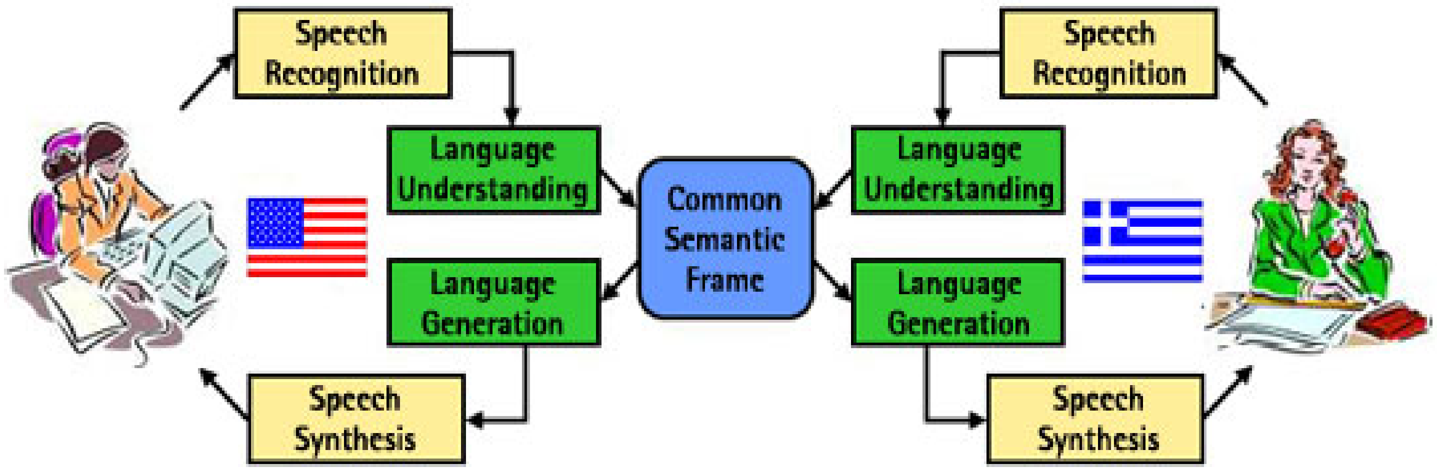
State of the Art

- Canadian computer program accepted daily weather data and generated weather reports (1976)

- Read student essays and grade them
- Automated reading tutor
- Spoken Dialogues
  - AT&T, How May I Help You?

State of the Art (cont.)

- MIT Spoken dialogue systems for information of restaurant, air travel, etc. (1991~)



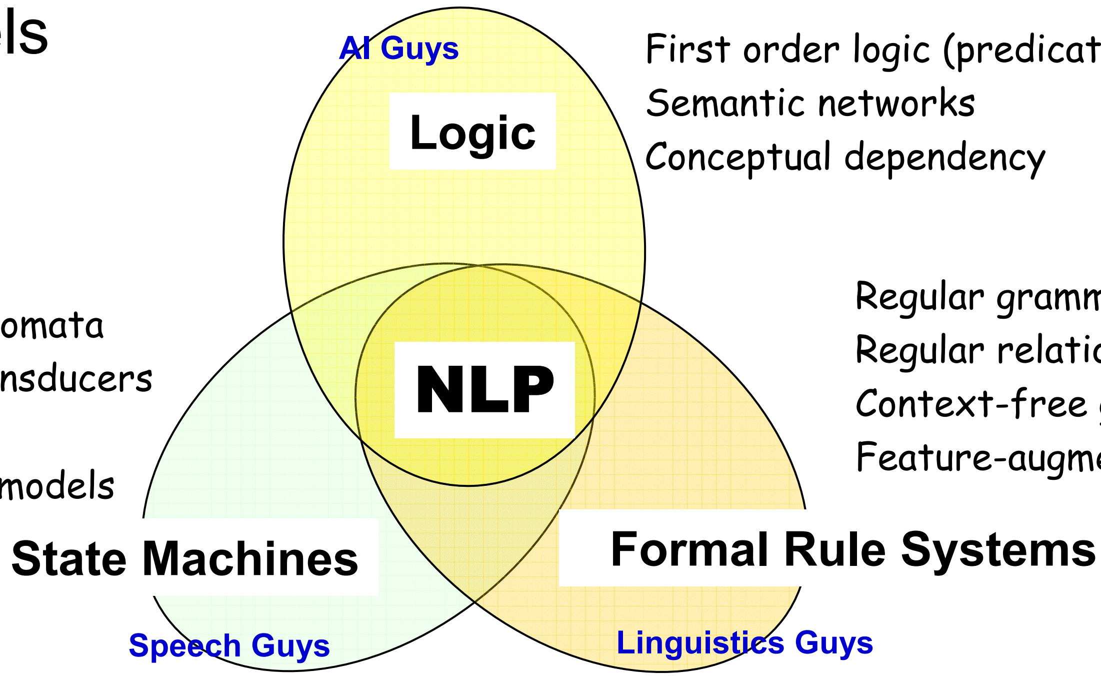
  - Speech recognition/synthesis
  - Natural language understanding/generation
  - Machine translation

State of the Art (cont.)

- CMU Universal Speech Interface

Models and Algorithms for NLP

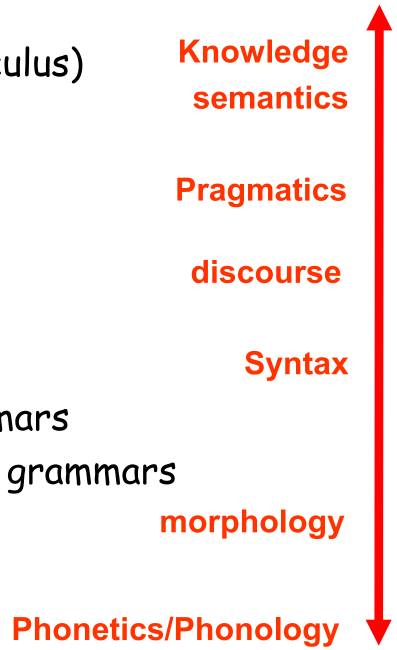
- Models



First order logic (predicate calculus)
Semantic networks
Conceptual dependency

Finite-state automata
Finite-state transducers
Markov models
Hidden Markov models

Regular grammars
Regular relations
Context-free grammars
Feature-augmented grammars



- Algorithms
  - Search:
    - Dynamic programming, depth-first search, best-first search, A* search
  - Learning/Training Methods

Major Topics for NLP

- Probability Theory/Statistics
  - Supervised/Unsupervised Machine Learning Techniques
- Words
  - Morphology
  - Regular expressions
  - Automata, Finite-State Transducers
- Syntax
  - Part-of-Speech Tagging
  - (Probabilistic) Context-Free Grammar
  - Parsing

Major Topics for NLP (cont.)

- Semantics/Meaning
  - Representation of Meaning
  - Semantic Analysis
  - Word Sense Disambiguation

- Pragmatics
  - Natural Language Generation
  - Discourse, Dialogue and Conversational Agents
  - Machine Translation

Dissidences

- Rationalists (e.g. Chomsky)
  - Humans are innate language faculties
  - (Almost fully) encoded rules plus reasoning mechanisms
  - Dominating between 1960's~mid 1980's

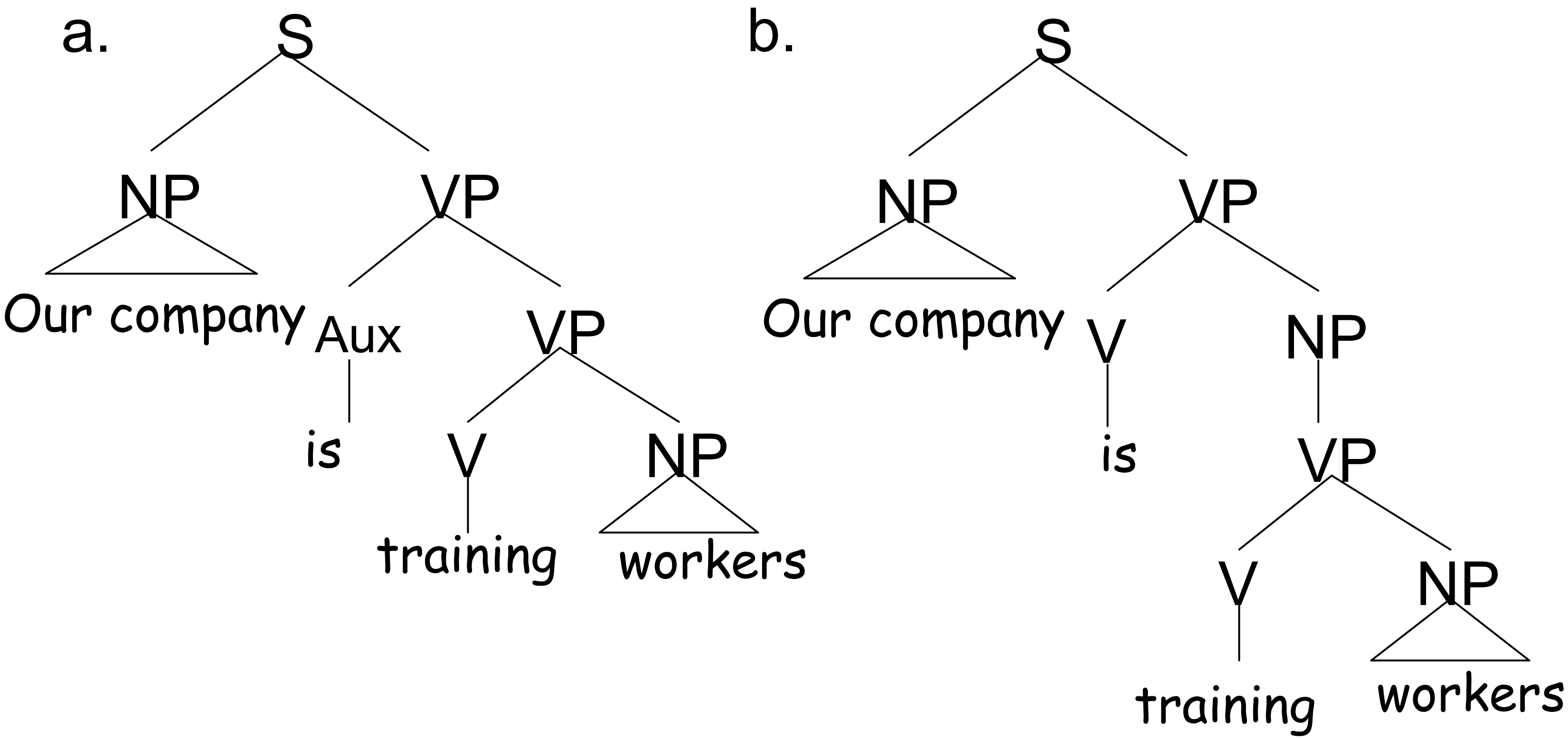
- Empiricists (e.g. Shannon)
  - The mind does not begin with detailed sets of principles and procedures for language components and cognitive domains
  - Rather, only general operations for association, pattern recognition, generalization etc., are endowed with
    - General language models plus machine learning approaches
  - Dominating between 1920's~mid 1960's and resurging 1990's~

Dissidences: Statistical and Non-Statistical NLP

- The dividing line between the two has become much more fuzzy recently
  - An increasing number of non-statistical researches use corpus evidence and incorporate quantitative methods
    - Corpus: "a body of texts" (大量的文稿)
  - Statistical NLP needs to start with all the scientific knowledge available about a phenomenon when building a probabilistic model, rather than closing one's eye and taking a clean-slate approach
    - Probabilistic and data-driven

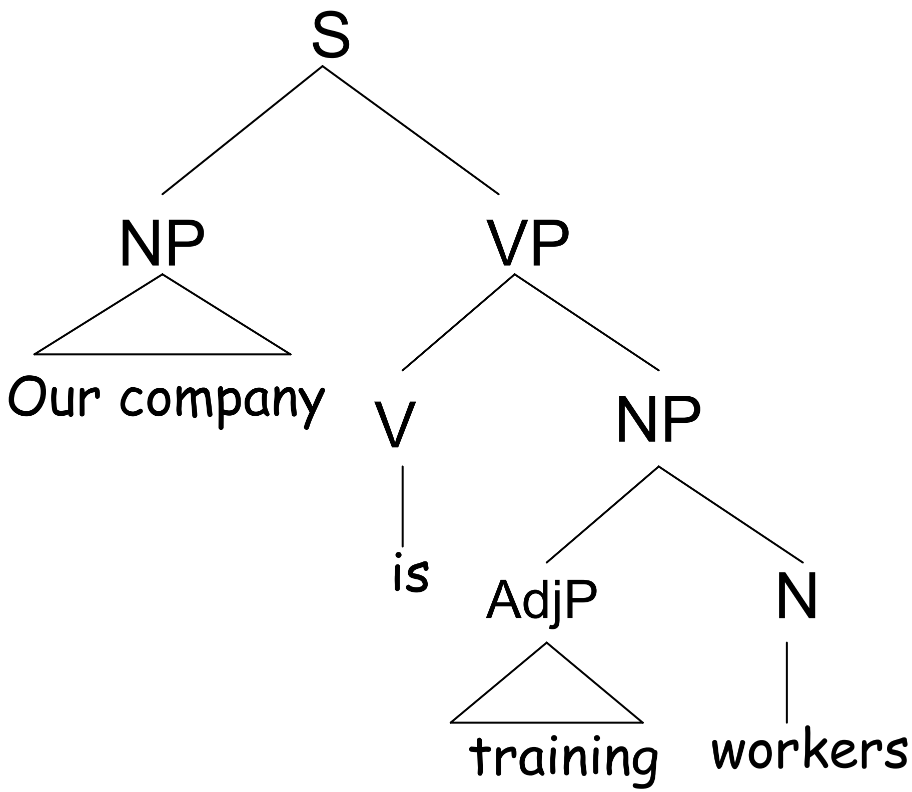
Ambiguity of Language

- A simple sentence, such as “Our company is training workers,” has 3 syntactic analyses (parses)



(Cf. Our problem is training workers.)

Ambiguity of Language (cont.)



(Cf. Those are training wheels.)

– The last two parses (b. and c.) are semantic anomalous!

Text Characteristics

- Word Counts

Tom Sawyers

- 71,370 word tokens
- 8,018 word types (distinct words)

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Table 1.1 Common words in *Tom Sawyer.*

  – Most common words are function words

Text Characteristics (cont.)

- Word Counts (count.)

Word Frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
> 100	102

3993/8018=49.8%

90%

Table 1.2 Frequency of frequencies of word types in *Tom Sawyer.*

  - The most common 100 words account for 50.9% of the word tokens

Text Characteristics (cont.)

Zipf, 1929

- Zipf's Law $f \propto \frac{1}{r} \Rightarrow f \cdot r = k$

Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

Table *1.3* Empirical evaluation of Zipf's law on Tom *Sawyer*.

Text Characteristics (cont.)

- Zipf's Law (cont.) $f \propto \frac{1}{r} \Rightarrow f \cdot r = k$

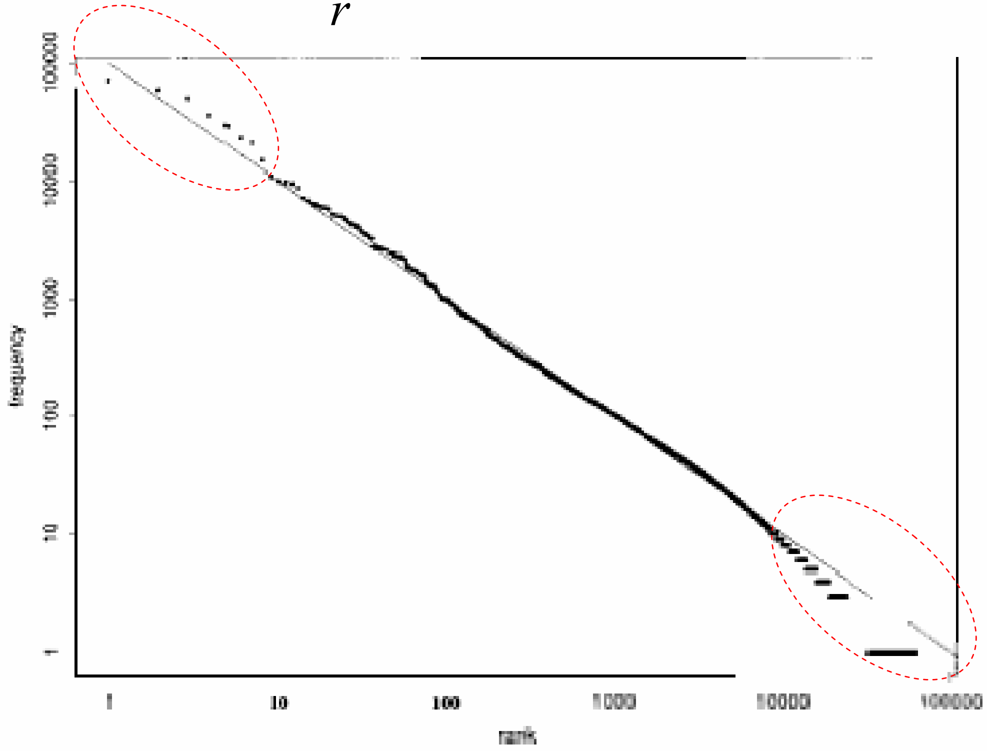


Figure 1.1 Zipf's law. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Zipf for k = 100,000, that is f x r = 100,000.

Text Characteristics (cont.)

- Zipf's Law (cont.) $f = p(r + \rho)^{-B}$

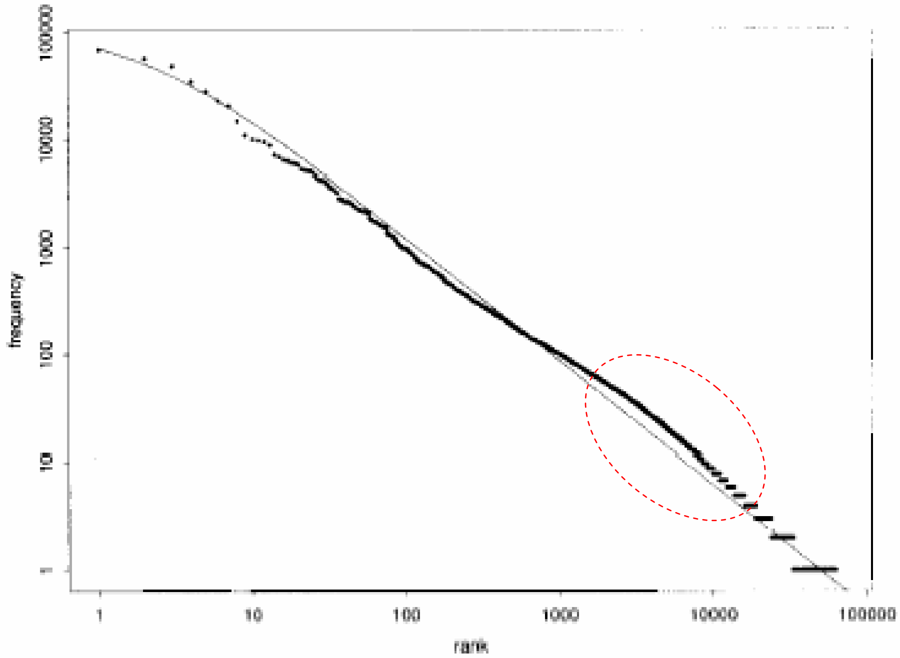


Figure 1.2 Mandelbrot's formula. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Mandelbrot's formula for $P = 10^{5.4}$, $B = 1.15$, $\rho = 100$.

Text Characteristics (cont.)

- Collocations
  - A collocation is an expression consisting of two or more words that correspond to some conventional way of saying thing
  - Where somehow the whole is perceived as having an existence beyond the sum of its parts
    - Expressions accompanied by certain connotations
  - E.g., (compound words, phrasal verbs, idioms, etc.)
    - Strong tea
    - Powerful heroin
    - Make up
    - Kick the bucket
    - Hear it through the grapevines
  - Important in areas like machine translation and information retrieval

Text Characteristics (cont.)

- Collocations (cont.)
  - E.g., finding the most common two-word sequences in a text

Frequency	Word 1	Word 2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table *1.4* Commonest bigram collocations in the New *York Times.*

Text Characteristics (cont.)

- Collocations (cont.)
  - Filtered by using “adjective noun” or “noun noun” constraints

Frequency	Word 1	Word 2	Part-of-speech pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

Table 1.5 Frequent bigrams after filtering. The most frequent bigrams in the New *York Times* after applying a part-of-speech filter.

Applications of NLP

- Speech Recognition
- Information Retrieval and Extraction
- Summarization
- Question Answering
- Conversational Agents
- Machine (Speech/Language) Translation
- Spelling Check
- Segmentation and Alignment
- Bioinformatics
- ….

Lexical Resources

- Corpora (Speech/Language Resources)
  - Refer speech waveforms, machine-readable text, dictionaries, thesauri as well as tools for processing them
  - International
    - Agents: e.g., LDC - Linguistic Data Consortium
    - Brown Corpus (1960's~70's, American English, balanced corpus)
      - 1 million words from 500 written text of different genres
    - Penn Treebank (Wall Street Journal, Parsed Sentences)
    - Canadian Hansards
    - CMU Lexicon
    - etc.
  - Domestic
    - Agents: e.g., The Association for Computational Linguistics and Chinese Language Processing
      - 中文詞庫
      - 中文語料庫、平衡語料庫等
      - Chinese Treebank

Research Resources

- Institutes/People
  - Foreign
    - MIT
    - CU
    - CMU
    - JHU
    - UMass
    - Cambridge
    - Microsoft
    - IBM
    - MITRE
    - HP
    - ……………

Research Resources (cont.)

- Conferences and Journals
  - **ACL**: Association for Computational Linguistics
  - **COLING**: International Conference on Computational Linguistics
  - **Computational Linguistics**
  - **Natural Language Engineering**

  - **ICSLP**: International Conference on Spoken Language Processing
  - **EUROSPEECH:** European Conference on Speech Communication and Technology
  - **ICASSP:** IEEE International Conference on Acoustics, Speech, Signal processing
  - **Speech Communication**
  - **Computer Speech and Language**
  - **IEEE Transactions on Speech and Audio Processing**

Topic List and Schedule

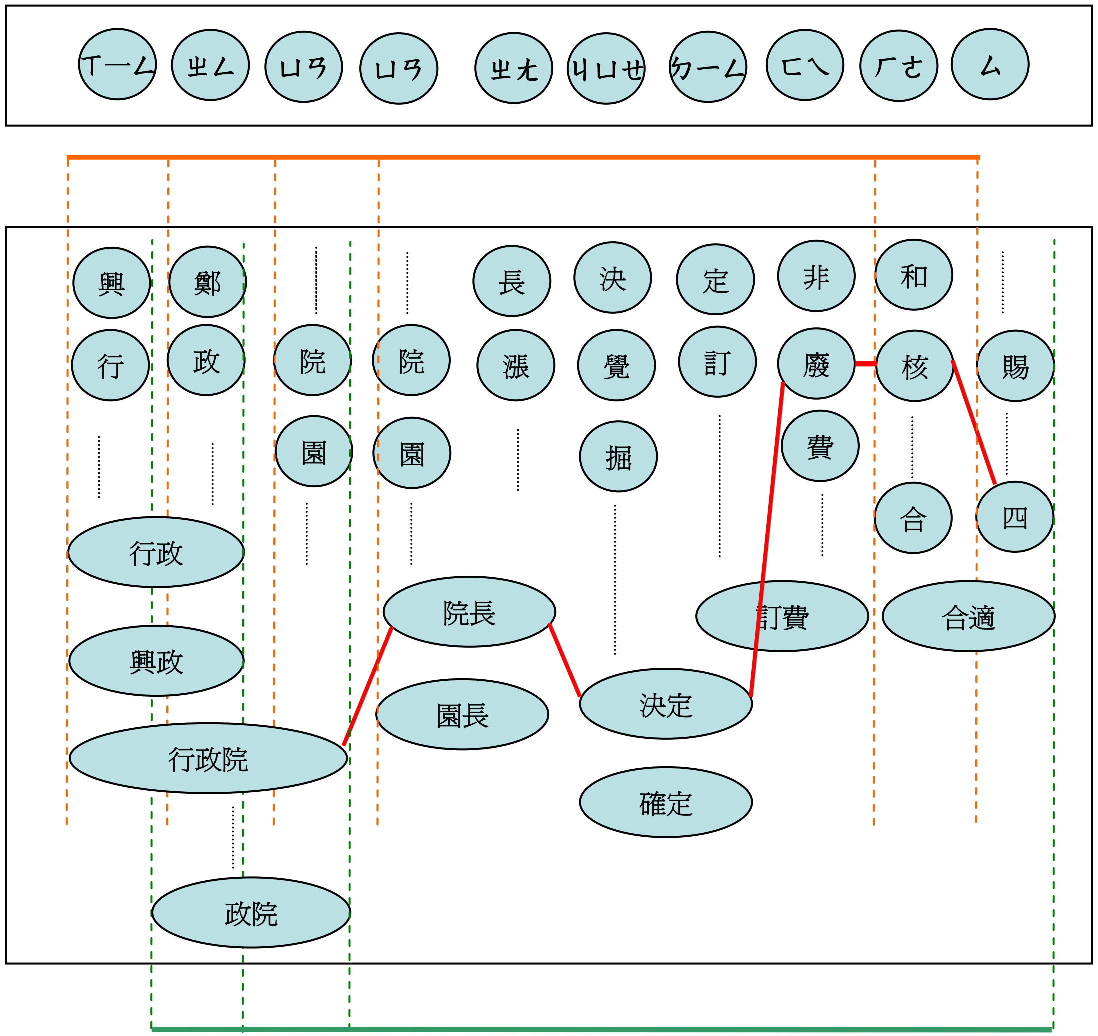
2/22	Course Overview & Introduction	
3/1	Mathematical Foundations*	
3/8	Linguistic Essentials	
3/15	Corpus-based Work	
3/22	Collections*	
3/29	N-gram Language Modeling*	
4/5	**Break**	
4/12	Part-of-Speech Tagging	
4/19	**Midterm**	
4/26	Parsing with Context-Free Grammars	
5/3	Word Sense Disambiguation*	
5/10	Probabilistic Context-Free Grammars	
5/17	Text Categorization*	
5/24	Lexical Acquisition	
5/31	Statistical Alignment and Machine Translation	
6/7	Information Extraction & Other Topics	
6/14	Paper Survey	
6/21	Paper Survey	
6/28	**FINAL**	

Homework to be Issued

- Chinese Input
- Part-of-Speech Tagging
- Syntactic Parsing
- Word Sense Disambiguation

Homework 1: Chinese Input

行 政 院 院 長 決 定 廢 核 四



Homework 1: Chinese Input (cont.)

- Requirements
  - Corpus
  - Lexicon
  - Bigram/Trigram Language Modeling
  - (Lattice) Search Algorithms
  - GUI (Graphical User Interface)
  - Performance Assessment/Error Analysis