Introduction to Natural Language Processing

What is Natural Language Processing

- Natural Language Processing (NLP)
  - It is the study of the computational treatment of natural (human) language.
  - In other words, teaching computers how to understand (and generate) human language.

Motivation for NLP

- Understand language analysis & generation
- Communication
- Language is a window to the mind
- Data is in linguistic form
- Data can be in Structured (table form), Semi structured (XML form), Unstructured (sentence form).

Modern Applications

- Search engines
  - Google, Yahoo, Bing, Baidu
- Question answering
  - IBM's Watson
- Natural language assistants
  - Apple's Siri, MS Cortana
- Translation systems
  - Google Translate
- News digest
  - Yahoo!
- Automated earthquake reports
  - LA Times
- Automated stock market reports
  - Narrative Science

Notes

- Computers are confused by (human) language
  - Specific techniques are needed
- NLP draws on research in many fields
  - Linguistics, Theoretical Computer Science, Mathematics, Statistics, Artificial Intelligence, Psychology, Databases, etc.

Two Contrasting Views of Language

- Language as a phenomenon
- Language as a data

Language Processing

- *Level 1* – Speech sound (*Phonetics & Phonology*)
- *Level 2* – Words & their forms (*Morphology, Lexicon*)
- *Level 3* – Structure of sentences (*Syntax, Parsing*)
- *Level 4* – Meaning of sentences (*Semantics*)
- *Level 5* – Meaning in context & for a purpose (*Pragmatics*)
- *Level 6* – Connected sentence processing in a larger body of text (*Discourse*)

Examples of Levels

- L1 : sound
- L2 : Dog - Dog(*s*), Dog(*ged*)
  Lady – Lad(*ies*)

  Should we store all forms of words in the lexicon?
- L3 : Ram goes to market ***(right)***
  goes Ram to the market ***(wrong)***
- L4 : translation from unstructured to structured representation

  go : (event)
  agent : Ram
  source : ?
  destination : market

Example (Contd.)

- L5 : User situation & context

  "*Is that water?*" – the action to be performed is different in a chemistry lab and on a dining table.
- L6 : Backward & forward references –
- Coreference resolution

  "*The man went near the dog. It bit him.*"

  Often co reference & ambiguity go together as in –

  "*The dog went near the cat. It bit it.*"

Linguistic Model of Natural Languages

- Linguists used to believe that there are rules the structure linguistic expressions which can differentiate wellformed speeches from ill-formed utterances
- “All grammars leak” (Edward Sapir, 1921)
  - It is just not possible to provide an exact and complete characterization of wellformed utterances that cleanly divides them from all other sequences of words, which are regarded as ill-formed utterances!
  - people always stretch and bend the ‘rules’ to meet their communicative needs!!
- It was found to be partially useful and needed to be loosenned to allow creativity in language

Language: Rationalist vs Empiricist Approaches

- **The rationalist approach postulate (Noam Chomsky):** The key parts of language are innate - hardwired in the brain at birth as part of the human genetic inheritance
  - Children can learn something as complex as a natural language from the limited input (of variable quality and interpretability) that they hear during their early years!

- **Empiricism approach postulate:** we can learn the complicated and extensive structure of language by specifying an appropriate general language model, and then inducing the values of parameters by applying statistical, pattern recognition, and machine learning methods to a large amount of language use
  - A baby's brain begins with general operations for association, pattern recognition, and generalization, and that these can be applied to the rich sensory input available to the child to learn the detailed structure of natural language

Probabilistic Approach of Language

- What are the common patterns that occur in language use?
  - Typically, a question of statistical nature!
- Statistical models of language are built and successfully used for many natural language processing (NLP) tasks
- Empiricist approach to NLP:
  - An appropriate general language model, plus
  - Induce the values of parameters by applying statistical, pattern recognition, and machine learning methods to a large amount of language use
- Use textual context as a surrogate for situating language in a real world context
- Corpus (pl. corpora) --- Body of texts
- “You shall know a word by the company it keeps” (J. R. Firth, 1957)

Statistical Concerns

- L1 : speech (make sense of sound)

  Approach –
  - Learning based
  - Probabilistic

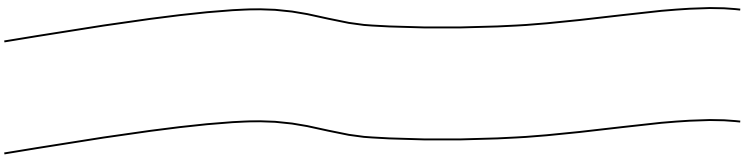
Statistical Concerns

- Statistical NLP draws from the work of Shannon, where the aim is to assign probabilities to linguistic events, so that we can say which sentences are ‘usual’ and which are ‘unusual’

- Interested in good descriptions of the associations and preferences that occur in the totality of language use.

Noisy Channel Metaphor

Speech Signal



Text

Noisy

- I want food.

- It is cold today.

Data-Driven Approach

The issues in this approach are -

- Corpora collection (coherent piece of text)
- Corpora cleaning – spelling, grammar, strange characters' removal
- Annotation
  - Named entity recognition
  - POS detection
  - Parsing
  - Meaning

The biggest challenge for NLP is *Ambiguity.*

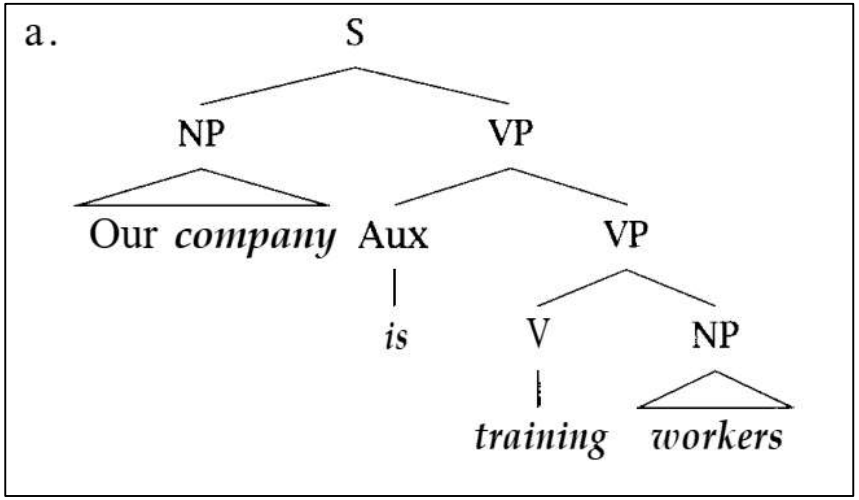
Ambiguity in Natural Language

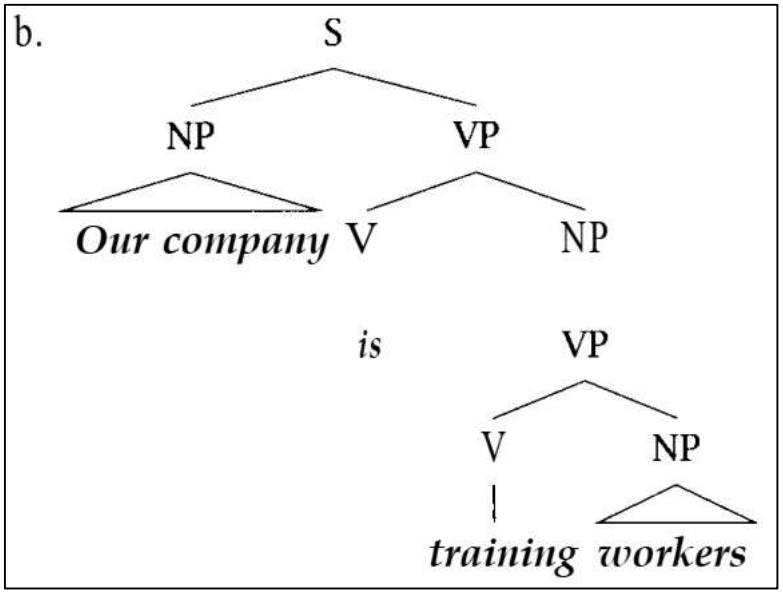
Ambiguity can be of 2 types –

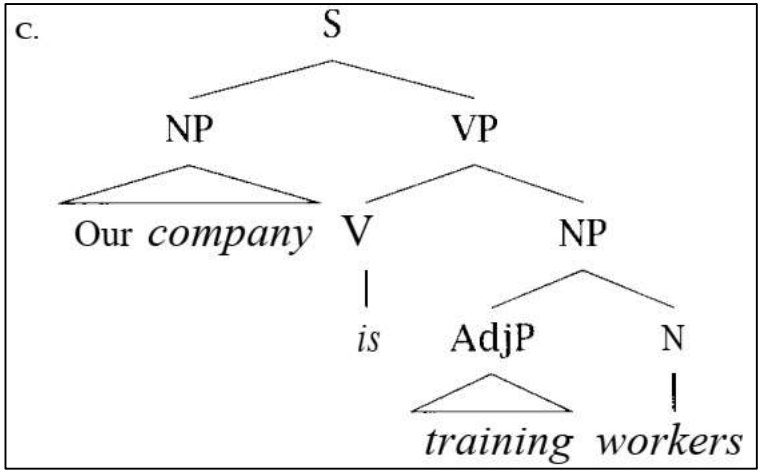
- Lexical – multiple meanings of words
  - It is dealt with in "*lexical semantics*"
    - Ex - "*The bank organized a loan mela on the bank of the river*"

- Structural –
  - It is dealt with in parsing.
    - Ex – "*I saw the boy with a telescope*"
    - Ex - "Our company is training workers"

Ambiguity in Natural Language







A practical NLP system must be good at making disambiguation decisions of word sense, word category, syntactic structure, and semantic scope

Ambiguity in Natural Language

- The goal of maximizing coverage while minimizing resultant ambiguity is fundamentally inconsistent with symbolic NLP systems
- Extending the coverage of the grammar to obscure constructions simply increases the number of undesired parses for common sentences and vice versa
- Hand-coded syntactic constraints and preference rules are time consuming to build, do not scale up well, and are brittle in the face of the extensive use of metaphor in language
- A Statistical NLP approach seeks to solve ambiguity problems by automatically learning lexical and structural preferences from corpora

Roadmap to Language Processing

- Collection the following: Some lexical resources - machine-readable text, dictionaries, thesauri
  - **Brown corpus:** A tagged corpus of about a million words that was put together at Brown University in the 1960s and 1970s; a balanced corpus; unfortunately, not a free one
  - **Lancaster-Oslo-Bergen (LOB) corpus:** A British English replication of the Brown corpus
  - **Susanne corpus:** A 130,000 word subset of the Brown corpus; a free corpus
  - **Penn Treebunk:** A corpus of syntactically annotated sentences; text taken from the Wall Street Journal; again a paid one
  - **Canadian Hunsurds:** A bilingual corpus; text from the proceedings of the Canadian parliament; also a paid corpus
  - **WordNet:** an electronic dictionary of English; words are organized into a hierarchy; a free dictionary

Roadmap to Language Processing

- Assumption: The input text represented as a list of words
- What are the most common words in the text?

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

- Word Tokens: 71370
- Token Types: 8018
- Ratio of tokens to types = 8.9

Roadmap to Language Processing

- In a corpus, word types are often distributed very unevenly.

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

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

- 12 Words occur 700 or more times
- Most common 100 words account for slightly over half (50.9%) of the word tokens
- Rest (49.8%) of the word types occur only once in the corpus

Zipf's Law

- Principle of Least Effort: people will act so as to minimize their probable average rate of work
- Zipf's law: If we count up how often each word (type) of a language occurs in a large corpus, and then list the words in order of their frequency of occurrence, we can explore the relationship between the frequency of a word f and rank its position in the list, known as its *rank r*. Zipf's law says that:

$$f \propto \frac{1}{r}$$

- Empirically, 50th most common word should occur with three times the frequency of the 150th most common word

Zipf's Law Illustrated on Tom Sawyer

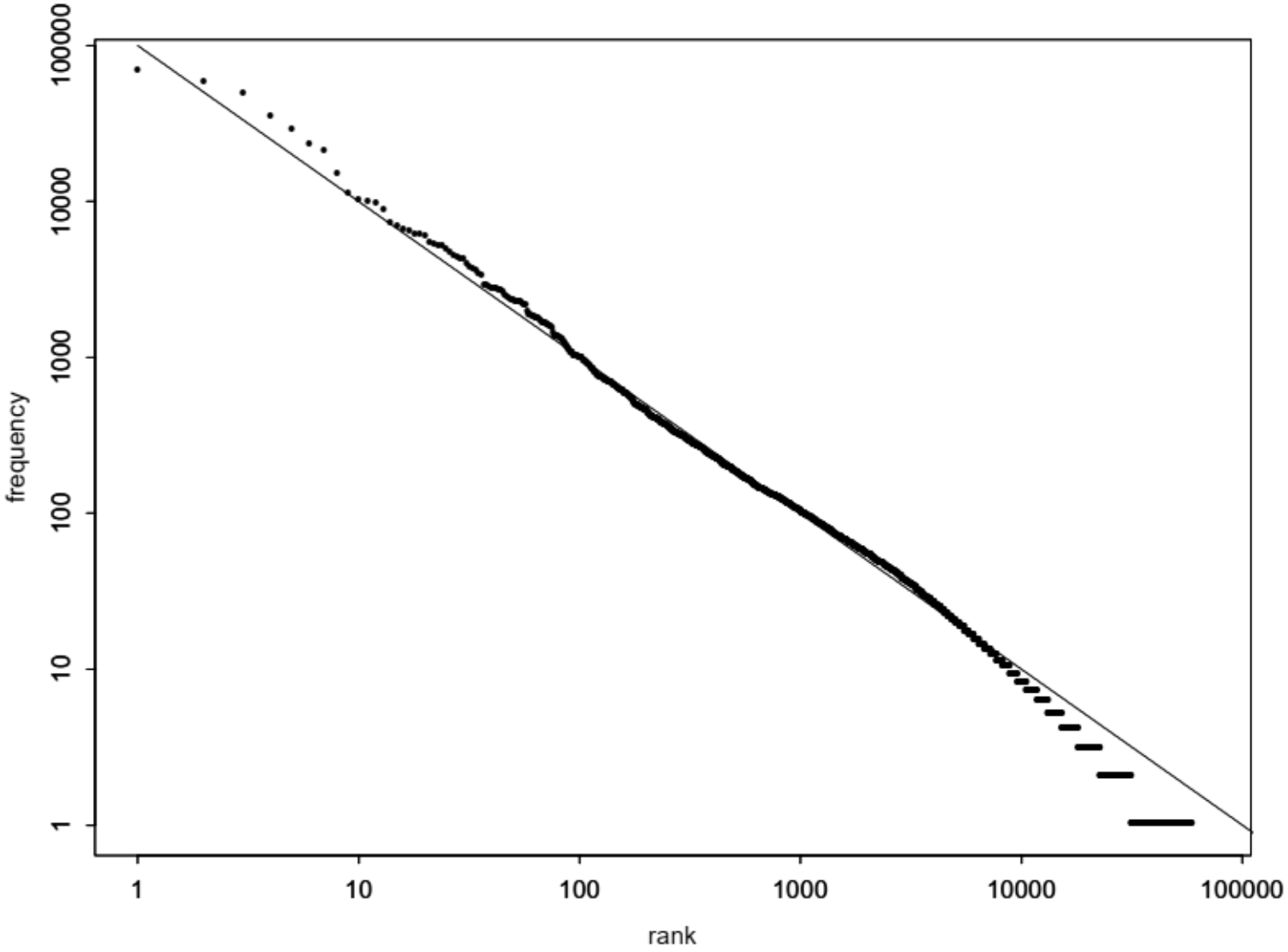
Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	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*.

Zipf's Law Illustrated

- Both the speaker and the hearer are trying to minimize their effort
- Speaker's effort is conserved by having a small vocabulary of common words
- Hearer's effort is lessened by having a large vocabulary of individually rarer words (so that messages are less ambiguous)
- Maximally economical compromise between these competing needs is argued to be the kind of reciprocal relationship between frequency and rank

Zipf's Curve for Brown Corpus



Mandelbrot's Refinement

- **The Revised Law:**

$$f = P(r + \rho)^{-B}$$

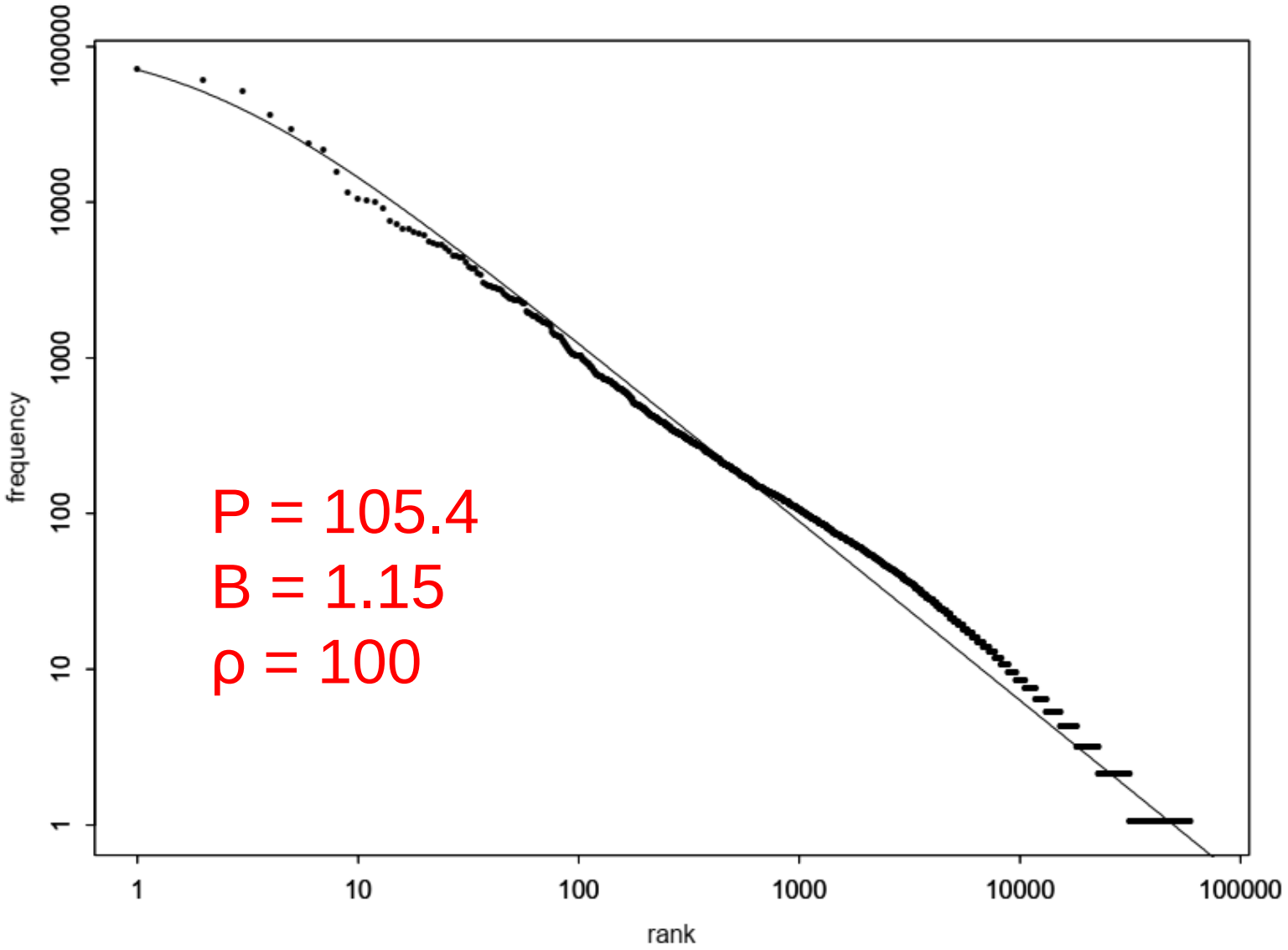
or

$$log\ f = log\ P - B\ log\ (r + \rho)$$

Here, P, B and ρ are corpus parameters

- **Note:** If B = 1 and ρ = 0, the the law reduces to Zipf's law

Mandelbrot's Curve for Brown Corpus



Zipf: A Second Law

- **Zipf's Observation:** The number of meanings (m) of a word is correlated with its frequency (f)
  - Conservation of speaker effort would prefer there to be only one word with all meanings
  - Conservation of hearer effort would prefer each meaning to be expressed by a different word
- **Zipf's Proposal:** $m \propto \sqrt{f}$ or $m \propto \frac{1}{\sqrt{r}}$
- **Empirical Support:**
  - In Zipf's own study, words of frequency rank about 10,000 average about 2.1 meanings, words of rank about 5000 average about 3 meanings, and words of rank about 2000 average about 4.6 meanings

Zipf: A Third Law

- **Zipf's Observation: Content words tend to clump**
  - Content words are words that have meaning
  - They can be compared to grammatical words, which are structural
  - Nouns, main verbs, adjectives and adverbs are usually content words; whereas, Auxiliary verbs, pronouns, articles, and prepositions are usually grammatical words
- **Study:** Zipf measured the number of lines or pages between each occurrence of the word in a text, and then calculate the frequency (F) of different interval sizes (I)
- **Observation:** For words of frequency at most 24 in a 260,000 word corpus, Number of intervals of a certain size is inversely related to the interval size

$$F \propto I^{-p}$$

- In Zipf's studies, *p* ranged between 1.0 and 1.3

Significance of Power Law

- It is quite Surprising that Power laws can be used to describe many natural phenomena
- Zipf's laws are examples of a Power laws
- A Random Experiment: A random generator that produces random characters from the 26 letters and thee blank character
- Probability of a word of length n being generated = $\left(\frac{26}{27}\right)^n \times \frac{1}{27}$ --- the probability of generating non-blank characters n times followed by the blank after that

Significance of Power Law

- The above expression may be simplified to a Mandelbrot-like power law form
- Key insights from the expression: Two opposing trends got combined
  - There are 26 times more words of length n + 1 than of length n
  - There is a constant ratio by which words of length n are more frequent than words of length n + 1

Collocation of Words

- The whole is perceived to have an existence beyond the sum of the parts!
  - Compounds; Ex. --- Disk Drive
  - Phrasal Verbs; Ex. --- Make up
  - Other stock phrases; Ex. --- Bread and Butter
  - May be several words long; Ex. --- International Best Practices
  - May be discontinuous; Ex. --- make [something] up
- Important in areas of Statistical NLP such as machine translation and information retrieval
  - A word may be translated differently according to the collocation it occurs in
  - One may want to index only ‘interesting’ phrases, that is, those that are collocations

Collocation of Words: Commonest Bigrams in NYT

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

- **Assumption:** Collocations tend to be frequent in usages
- Most frequent bigrams not necessarily correspond to collocations
- Part-of-speech pattern based collocation filter is needed to remove irrelevant bigrams
- Freqently-used part-of-speech pattern based collocation filter
  - Noun-Noun (NN)
  - Adjective-Noun (AN)

Commonest Bigrams in NYT: Filtered

Frequency	Word 1	Word 2	POS 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

Concordance

- Concordance refers to an alphabetic list of all instances of a certain word or phrase in a corpus in its immediate context (before and after)
- It is most commonly displayed as a 'keyword in context' (KWIC), a computer-generated set of concordance lines consisting of the target word and its linguistic context, which may be used to explore various uses or senses of the feature of interest
- Sometimes one would like to search for series of phrases that contain particular keyword in a passage or corpus --- Concordance programs in NLP are used to locate and print series of phrases that contain the target keyword

Syntactic Frames and Key Word In Context (KWIC) Display

- A Syntactic frame represents synthetically a set of possible syntactic structures associated with the verb
- In such a representation, all occurrences of the word of interest are lined up beneath one another, with surrounding context shown on both sides
- A Key Word In Context (KWIC) Concordance program program can find from a corpus a display as the one as shown below

1	could find a target. The librarian	“showed	off” - running hither and thither w
2	elights in. The young lady teachers	“showed	off” - bending sweetly over pupils
3	ingly. The young gentlemen teachers	“showed	off” with small scoldings and other
4	seeming vexation). The little girls	“showed	off” in various ways, and the littl
5	n various ways, and the little boys	“showed	off” with such diligence that the a
6	t genuwyne?” Tom lifted his lip and	showed	the vacancy. “Well, all right,” sai
7	is little finger for a pen. Then he	showed	Huckleberry how to make an H and an
8	ow’s face was haggard, and his eyes	showed	the fear that was upon him. When he
9	not overlook the fact that Tom even	showed	a marked aversion to these inquests
10	own. Two or three glimmering lights	showed	where it lay, peacefully sleeping,
11	ird flash turned night into day and	showed	every little grass-blade, separate
12	that grew about their feet. And it	showed	three white, startled faces, too. A
13	he first thing his aunt said to him	showed	him that he had brought his sorrows
14	p from her lethargy of distress and	showed	good interest in the proceedings. S
15	ent a new burst of grief from Becky	showed	Tom that the thing in his mind had
16	shudder quiver all through him. He	showed	Huck the fragment of candle-wick pe

Syntactic Frames from Tom Sawyer

Syntactic frames for *showed* in *Tom Sawyer*

NP_{agent} showed **off** $(PP[with/in]_{manner})$

NP_{agent} showed $(NP_{recipient})$ $\left(\left\{ \begin{array}{l} NP_{content} \\ CP[that]_{content} \\ VP[inf]_{content} \\ how\ VP[inf]_{content} \\ CP[where]_{content} \end{array} \right\} \right)$

NP_{agent} showed $NP[interest]$ $PP[in]_{content}$

NP_{agent} showed $NP[aversion]$ $PP[to]_{content}$