Questions that linguistics should answer

- What kinds of things do people say?
- What do these things say/ask/request about the world?

Example: *In addition to this, she insisted that women were regarded as a different existence from men unfairly.*

- Text corpora give us *data* with which to answer these questions
- What words, rules, statistical facts do we find?
- Can we build programs that learn effectively from this data, and can then do NLP tasks?
- They are an externalization of linguistic knowledge

Corpora

- A *corpus* is a *body* of naturally occurring text, normally one organized or selected in some way
  - Greek: one corpus, two corpora
- A *balanced corpus* tries to be representative across a language or other domain
- Balance is something of a chimaera: What is balanced? Who spends what percent of their time reading the sports pages?

The Brown corpus

- Famous early corpus. Made by W. Nelson Francis and Henry Kučera at Brown University in the 1960s. A balanced corpus of written American English in 1960 (except poetry!).
- 1 million words, which seemed huge at the time.

  Sorting the words to produce a word list took 17 hours of (dedicated) processing time, because the computer (an IBM 7070) had the equivalent of only about 40 kilobytes of memory, and so the sort algorithm had to store the data being sorted on tape drives.
- Its significance has increased over time, but also awareness of its limitations.
- Tagged for part of speech in the 1970s
  - The/AT General/JJ-TL Assembly/NN-TL ,/, which/WDT adjourns/VBZ today/NR ,/, has/HVZ performed/VBN

Recent corpora

- British National Corpus. 100 million words, tagged for part of speech. Balanced.
- Newswire (*NYT* or *WSJ* are most commonly used): Something like 600 million words is fairly easily available.
- Legal reports; UN or EU proceedings (parallel multilingual corpora – same text in multiple languages)
- The Web (in the billions of words, but need to filter for distinctness).
- Penn Treebank: 2 million words (1 million *WSJ*, 1 million speech) of parsed sentences (as phrase structure trees).

Common words in *Tom Sawyer* (71,370 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

Frequencies of frequencies in *Tom Sawyer*

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

Zipf's law in *Tom Sawyer*

Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
Oh	116	90	10440
two	104	100	10400

Word	Freq. (f)	Rank (r)	$f \cdot r$
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
Could	2	4000	8000
Applausive	1	8000	8000

Zipf's law

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

There is a constant k such that

$$f \cdot r = k \tag{2}$$

(Now frequently invoked for the web too!

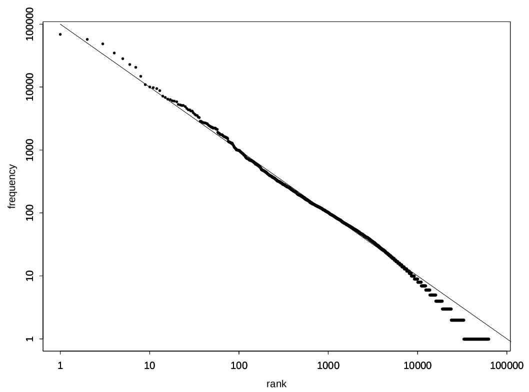
See http://linkage.rockefeller.edu/wli/zipf/)

Mandelbrot's law

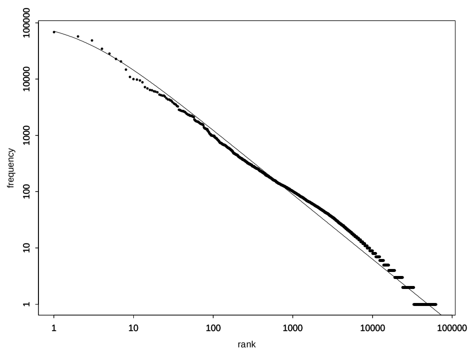
$$f = P(r + \rho)^{-B} \tag{3}$$

$$\log f = \log P - B \log(r + \rho) \tag{4}$$

Zipf's law for the Brown corpus



Mandelbrot's formula for the Brown corpus



$P = 10^{5.4}$, $B = 1.15$, $\rho = 100$

Commonest bigrams in the *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

Filtered common bigrams in the *NYT*

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

KWIC display

```
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 for *showed* in *Tom Sawyer*

$$\mathrm{NP}_{agent}\ \text{showed off}\ (\mathrm{PP}[\textit{with/in}]_{manner})$$

$$\mathrm{NP}_{agent}\ \text{showed}\ (\mathrm{NP}_{recipient}) \left(\begin{Bmatrix} \mathrm{NP}_{content} \\ \mathrm{CP}[\textit{that}]_{content} \\ \mathrm{VP}[\mathrm{inf}]_{content} \\ \textit{how}\ \mathrm{VP}[\mathrm{inf}]_{content} \\ \mathrm{CP}[\textit{where}]_{content} \end{Bmatrix}\right)$$

$$\mathrm{NP}_{agent}\ \text{showed}\ \mathrm{NP}[\textit{interest}]\ \mathrm{PP}[\textit{in}]_{content}$$

$$\mathrm{NP}_{agent}\ \text{showed}\ \mathrm{NP}[\textit{aversion}]\ \mathrm{PP}[\textit{to}]_{content}$$

Basic issues in text

- Upper and lower case
  - When is it useful to treat *black*, *Black* and *BLACK* the same or differently?
- Tokenization: what is a word?
  - Whitespace separated? *“I’m bored!”*
  - Oddities like *Micro$oft* or *C|net*, or even `:-)`
  - May want to have whitespace-separated single words: *New York* or *make up*. Or *by and large*?
  - Word segmentation is much harder in other languages (Chinese, Thai, etc. – no word breaks marked)

Basic issues in text

- Periods/Sentence boundaries
  - Normally ends of sentences. But also used for abbreviations: *Calif.*
  - Can have both functions at once (haplology): *I like you, etc.*
  - People have researched sentence segmentation.
- Single apostrophes
  - How does one treat *I’ll* or *isn’t* and/or *dog’s*
  - With final quotes like *the kids’ toys* you need longer distance context to tokenize well

Basic issues in text

- Hyphenation:
  - Traditionally, for text line breaks
  - *e-mail* or *co-operate* or *A-1-plus* [vs. *cooperate*]
  - *the aluminum-export ban* or *a text-based medium*
  - *the idea of a child-as-required-yuppie-possession must be motivating them*
  - In Dow Jones newswire get all of *data base*, *data-base* and *database* (with choice partly depending on section of paper).
- Homographs: *saw*

Much of the structure is implicit, traditionally

- Two carriage returns indicate a paragrah break
- Now, often SGML or XML gives at least some of the macro structure (sentences, paragraphs). Commonly not micro-structure
- <p><s>And then he left.</s>
  <s>He did not say another word.</s></p>
- <utt speak="Fred" date="10-Feb-1998">That is an ugly couch.</utt>
- May not be semantic markup:
  - <B><font="+3">Relevant prior approaches</B>

Distinctiveness of NLP as an ML problem

- Language allows the complex, compositional encoding of thoughts, ideas, feelings, . . . , intelligence.
- Most structure is hidden
- Long pipelines
- Large and strange, sparse, discrete distributions
- Large scale
- Feature-driven; performance-driven

Distinctiveness of NLP as an ML problem

- Much hidden structure; long processing pipelines
  - Long pipelines of probabilistic decompositions, through which errors can – and do – propagate
  - The problem has a relational/CSP nature. It's not just doing a series of (assumed iid) simple classification tasks. There are a lot of decisions to coordinate.
  - We are often dealing with hierarchical structures (branching processes), and often want to allow more complex forms of information sharing (dependencies).

Large and strange, sparse, discrete distributions

- Both features and assigned classes regularly involve multinomial distributions over huge numbers of values (often in the tens of thousands).
- The distributions are very uneven, and have fat tails
- Enormous problems with data sparseness: much work on smoothing distributions/backoff (shrinkage), etc.
- We normally have inadequate (labeled) data to estimate probabilities
- Unknown/unseen things are usually a central problem
- Generally dealing with discrete distributions though

Sparsity

- How often does an every day word like *kick* occur in a million words of text?
  - *kick*: about 10 [depends vastly on genre, of course]
  - *wrist*: about 5
- Normally we want to know about something bigger than a single word, like how often you *kick a ball*, or how often the conative alternation *he kicked* ***at*** *the balloon* occurs.
- How often can we expect that to occur in 1 million words?
- Almost never.
- "There's no data like more data" [if of the right domain]

Distinctiveness of NLP as an ML problem

- Large scale
  - Scale is extremely large: tens or hundreds of thousands of features/parameters.
  - Need to balance beauty and practicality. Dynamic programming techniques and speed issues are key. Need to exploit sparseness.
- Feature-driven; performance-driven
  - The emphasis is on developing and finding ways to effectively exploit useful features from a rich but sparse representation, rather than on trade-offs between machine learning methods (there's still linguistics there!)