

Quantitative linguistics

- Much of theoretical and descriptive linguistics is *qualitative*
- Computational corpus linguistics opens up possibilities for *quantitative* work
- New methodology: data collection, statistics, hypothesis testing

Frequency

- *Absolute frequency* is the number of times a type occurs in a corpus.
- Zipf's Law (1949) predicts that frequency times rank is a constant:

$$f \times r = c$$

- *Relative frequency* is absolute frequency divided by corpus size

Hypothesis testing

- Paired sign test
- Significance
- Chi squared

Chi squared

- Federalist papers, published under the pseudonym 'Publius' in 1787–8
- Written by James Madison, John Jay, and Alexander Hamilton
- Jay wrote 5 essays, Hamilton wrote 43, Madison wrote 14, another 12 were disputed
- Mosteller and Wallace selected a set of words which appeared with different relative frequencies in Hamilton's and Madison's essays:
upon, also, an, by, of, on, there, this, to, although, both, enough, while, whilst, always, though, commonly, consequently, considerably, according, apt, direction, innovation, language, vigor, kind, matter, particularly, probability, work
- Used log odds to compare frequencies

Chi squared

- Observed counts:

	Madison	Hamilton	total
<i>whilst</i>	12	1	13
not <i>whilst</i>	39207	117155	156362
total	39219	117156	156375

- Expected values (row total \times column total \div grand total):

	Madison	Hamilton	total
<i>whilst</i>	3.26	9.74	13
not <i>whilst</i>	39215.74	117146.26	156362
total	39219	117156	156375

- Calculate chi squared:

$$\begin{aligned} \chi^2 &= \frac{(12 - 3.26)^2}{3.26} + \frac{(1 - 9.74)^2}{9.74} + \frac{(39207 - 39215.74)^2}{39215.74} + \frac{(117155 - 117146.26)^2}{117146.26} \\ &= 23.43 + 7.84 + 0.002 + 0.00065 \\ &= 31.27 \end{aligned}$$

Chi squared

- A problem:

	Madison	Hamilton	total
<i>the</i>	3948	10867	14815
not <i>the</i>	35271	106289	141560
total	39219	117156	156375

- Calculate chi squared:

$$\begin{aligned} \chi^2 &= \frac{(3948 - 3715.6)^2}{3715.6} + \frac{(10867 - 11099.4)^2}{11099.4} + \\ &\quad \frac{(35271 - 35503.4)^2}{35503.4} + \frac{(106289 - 106056.6)^2}{106056.6} \\ &= 14.54 + 4.87 + 1.52 + 0.51 \\ &= 21.44 \end{aligned}$$

- Differences in frequent words are almost always significant, but are they meaningful?

Chi squared

- The chi squared test depends on the assumption that words are independent and identically distributed (i.i.d.)
- Words in real texts are not i.i.d., so the frequencies are always non-random. And, the more common a word is, the more evidence the chi squared test has for non-randomness.
- Since words tend to occur in bursts, we can get misleading results even for less frequent words.
- Despite its shortcomings, we can still use χ^2 as a measure of similarity between corpora.

Chi square

	male %	female %	χ^2		male %	female %	χ^2
fucking	0.08	0.01	1233.1	she	0.42	0.87	3109.7
er	0.56	0.36	945.4	her	0.14	0.28	965.4
the	2.60	2.20	698.0	said	0.29	0.47	872.0
yeah	1.29	1.10	310.3	n't	1.44	1.70	443.9
aye	0.07	0.03	291.8	I	3.24	3.58	357.9
right	0.36	0.27	276.0	and	1.73	1.94	245.3
hundred	0.09	0.05	251.1	to	1.37	1.54	198.6
fuck	0.02	0.00	239.0	cos	0.20	0.26	194.6
is	0.79	0.67	233.3	oh	0.78	0.90	170.2
of	0.81	0.69	203.6	Christmas	0.02	0.04	163.9

Chi square

- Listing words by X^2 doesn't prove anything, but identifies possible questions for further research.
- The chi square metric can also be used to compare corpora on the basis of the n most frequent words (construct an $n \times 2$ contingency table).
- Dunning's log likelihood statistic is similar to X^2 , but sometimes better for small counts:

$$G^2 = 2 \sum_{i,j} O_{ij} \ln \frac{O_{ij}}{E_{ij}}$$

- Other tests are less sensitive to differences in absolute frequency (Mann Whitney or Wilcoxon rank test).

Collocations

- Collocations are groups of words that occur together frequently
 - fixed expressions (*by and large*)
 - idioms (*spill the beans*)
 - cliched language (*as soon as possible*)
 - lexical bundles (*the extent to which*)
 - quotes (*Beam me up, Scotty*)
 - 'snowclones' (*And I, for one, welcome our new insect overlords*)
 - compounds (*zip code*)
 - names (*Las Vegas*)
 - verb/particle constructions (*pick up*)
 - light verbs (*have breakfast*)
 - quantifiers (*a glass of water*)
 - selectional restrictions (*vivacious girl*)

Collocations

- Collocational knowledge is important for language learners, especially for sentence generation:
 - *strong tea* vs. *powerful tea*
 - *strong car* vs. *powerful car*
- On-line language production heavily influenced by collocational patterns
- Collocations are (probably) the right unit to store in translation databases
- Collocation inventories can be constructed by quantitative analysis of corpora

Information Theory

- Alan Turing, Claude Shannon
- A mathematical theory of communication and how messages convey information
- Originally developed for decoding messages during WW2, later applied to improving telegraph and telephone communication
- Based on probabilities: how would an optimal betting strategy change after you find out a roulette wheel is rigged?
- Fundamental concept = information entropy

Entropy

- Entropy is the average surprise on finding out the outcome of some random variable X

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

- If $P(x)$ is the probability that the outcome is x , then $-\log P(x)$ is how surprised we would be if the outcome were x
- Since $P(x)$ ranges from 0 (for impossible events) to 1 (for certain events), the surprise ranges from infinity (for impossible events) to 0 (for certain events)
- Entropy is the weighted average of the surprise for all possible outcomes

Entropy

- Entropy is never less than zero (no surprise, total information) but has no upper limit
- More possible outcomes and more even probabilities raises surprise and entropy, fewer outcomes or skewed probabilities lowers surprise and entropy
- Source Coding Theorem: an optimal binary encoding for S uses on average $H(S) \pm 1$ bits
- one such optimal coding scheme is Huffman coding, created in 1954 as homework in an MIT information theory class

Entropy

- Shannon game: guess the next letter in an English text

Model	Entropy (bits)
zeroth order	4.76
first order	4.03
second order	2.8
humans	1.3

- More information helps a lot – ‘optimal’ code depends on the amount of context
- Note that typical text encodings use 8 bits per character, so there’s lots of room for compression
- Also compare with the number of bits required for audio (speech) recordings