COR: Corpus Linquistics

Lecture 5
Collocation, Frequency, Corpus Statistics

Francis Bond
Department of Asian Studies
Palacký University
https://fcbond.github.io/
bond@ieee.org

https://github.com/bond-lab/Corpus-Linguistics

Overview

- Frequency
- Corpus Statistics
- Collocations

Word Frequency Distributions

Lexical statistics & word frequency distributions

- Basic notions of lexical statistics
- Typical frequency distribution patterns
- Zipf's law
- Some applications

Lexical statistics

- Statistical study of the frequency distribution of types (words or other linguistic units) in texts
  - remember the distinction between **types** and **tokens**?
- Different from other categorical data because of the extreme richness of types
  - people often speak of **Zipf's law** in this context

Basic terminology

- N: sample / corpus size, number of tokens in the sample
- V: vocabulary size, number of distinct types in the sample
- V_m: spectrum element m, number of types in the sample with frequency m (i.e. exactly m occurrences)
- V_1: number of hapax legomena, types that occur only once in the sample (for hapaxes, #types = #tokens)
- Consider {c a a b c c a c d}
- $N = 9, V = 4, V_1 = 2$

- Rank/frequency profile:

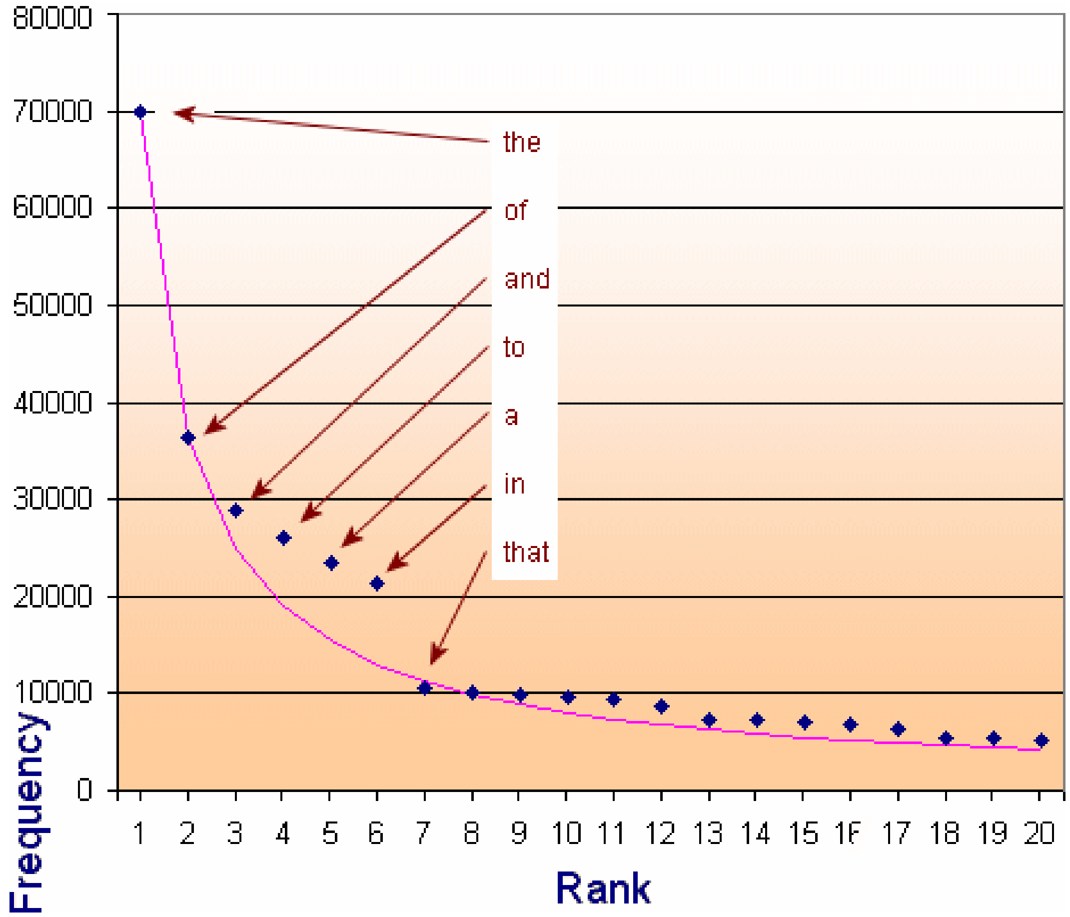
item	frequency	rank	
c	4	1	
a	3	2	
b	1	3	
d	1	3	(or 4)

  Expresses type frequency as function of rank of a type

Top and bottom ranks in the Brown corpus

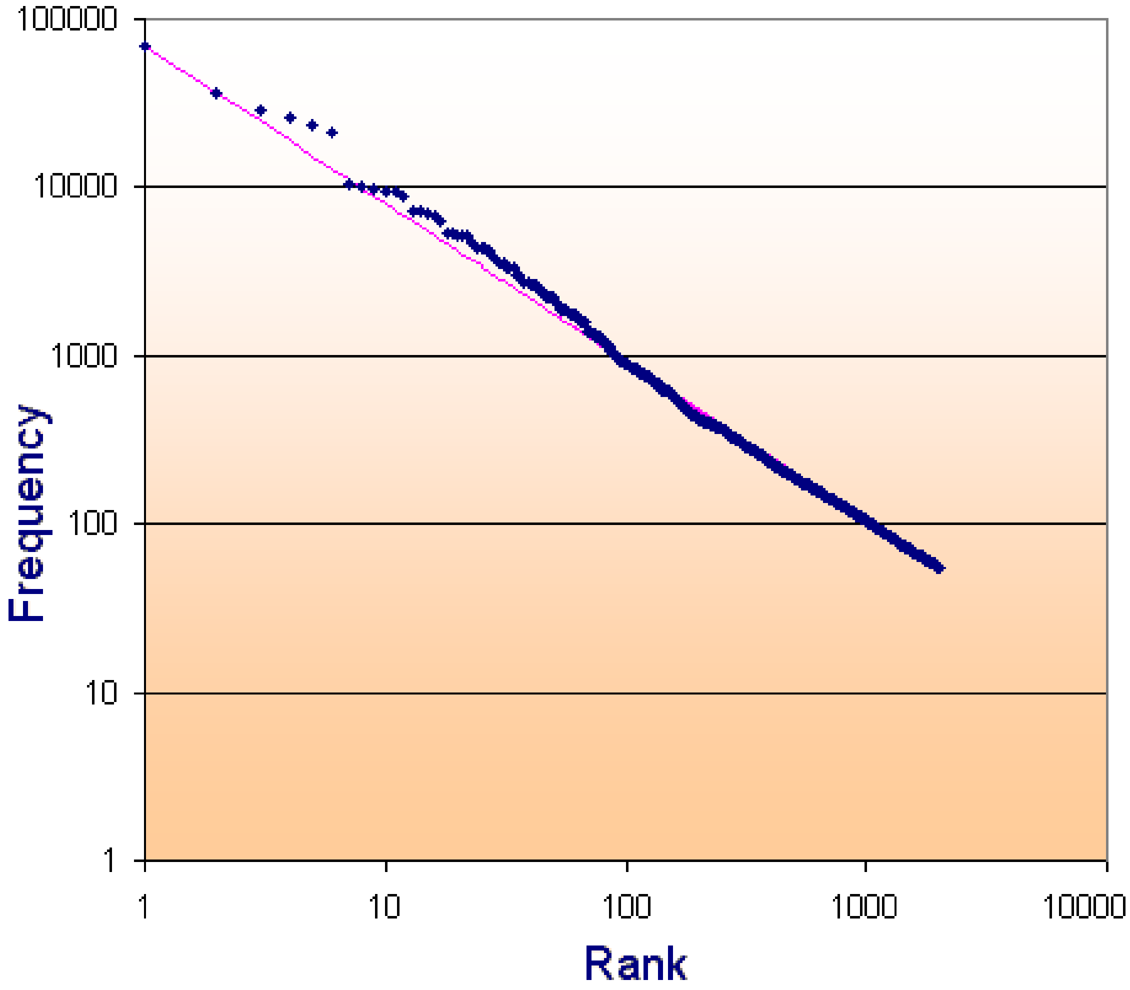
top frequencies			bottom frequencies		
r	f	word	rank range	f	randomly selected examples
1	62642	*the*	7967– 8522	10	*recordings, undergone, privileges*
2	35971	*of*	8523– 9236	9	*Leonard, indulge, creativity*
3	27831	*and*	9237–10042	8	*unnatural, Lolotte, authenticity*
4	25608	*to*	10043–11185	7	*diffraction, Augusta, postpone*
5	21883	*a*	11186–12510	6	*uniformly, throttle, agglutinin*
6	19474	*in*	12511–14369	5	*Bud, Councilman, immoral*
7	10292	*that*	14370–16938	4	*verification, gleamed, groin*
8	10026	*is*	16939–21076	3	*Princes, nonspecifically, Arger*
9	9887	*was*	21077–28701	2	*blitz, pertinence, arson*
10	8811	*for*	28702–53076	1	*Salaries, Evensen, parentheses*

Rank/frequency profile of Brown corpus



Is there a general law?

- Language after language, corpus after corpus, linguistic type after linguistic type, . . . we observe the same "few giants, many dwarves" pattern
- The nature of this relation becomes clearer if we plot $\log(f)$ as a function of $\log(r)$


100000
10000
1000
100
10
1
Frequency
1
10
100
1000
10000
Rank

Zipf's law

➢ Straight line in double-logarithmic space corresponds to power law for original variables

➢ This leads to Zipf's (1949, 1965) famous law:

$$f(w) = \frac{C}{r(w)^a} \quad \text{or} \quad f(w) \propto \frac{1}{r(w)} \tag{1}$$

$f(w)$: Frequency of Word w
$r(w)$: Rank of the Frequency of Word w (most frequent $=$ 1, ...)

➢ With a $=$ 1 and C $=$60,000, Zipf's law predicts that:
  - ∗ most frequent word occurs 60,000 times
  - ∗ second most frequent word occurs 30,000 times
  - ∗ third most frequent word occurs 20,000 times

* and there is a long tail of 80,000 words with frequencies
* between 1.5 and 0.5 occurrences(!)

Applications of word frequency distributions

- Most important application: extrapolation of vocabulary size and frequency spectrum to larger sample sizes
  - productivity (in morphology, syntax, ...)
  - lexical richness
    (in stylometry, language acquisition, clinical linguistics, ...)
  - practical NLP (est. proportion of OOV words, typos, ...)

- Direct applications of Zipf's law in NLP
  - Population model for Good-Turing smoothing
    If you have not seen a word before its probability should probably not be 0 but closer to $\frac{1}{N}$
  - Realistic prior for Bayesian language modelling

Other Zipfian (power-law) Distributions

- Calls to computer operating systems (length of call)
- Colors in images
  the basis of most approaches to image compression
- City populations
  a small number of large cities, a larger number of smaller cities
- Wealth distribution
  a small number of people have large amounts of money, large numbers of people have small amounts of money
- Company size distribution
- Size of trees in a forest (roughly)

Hypothesis Testing for Corpus Frequency Data

Some questions

- How many passives are there in English?
  What proportion of verbs are in passive voice?
  - a simple, innocuous question at first sight, and not particularly interesting from a linguistic perspective
- but it will keep us busy for many hours ...
- slightly more interesting version:
  - Are there more passives in written English than in spoken English?

More interesting questions

- How often is *kick the bucket* really used idiomatically? How often literally? How often would you expect to be exposed to it?
- What are the characteristics of **translationese**?
- Do Americans use more split infinitives than Britons? What about British teenagers?
- What are the typical collocates of *cat*?
- Can the next word in a sentence be predicted?
- Do native speakers prefer constructions that are grammatical according to some linguistic theory?

Back to our simple question

- How many passives are there in English?
  - American English style guide claims that
    - * "In an average English text, no more than 15% of the sentences are in passive voice. So use the passive sparingly, prefer sentences in active voice."
    - * http://www.ego4u.com/en/business-english/grammar/passive states that only 10% of English sentences are passives (as of June 2006)!
  - We have doubts and want to verify this claim

Problem #1

- Problem #1: What is English?
- Sensible definition: group of speakers
  - e.g. American English as language spoken by native speakers raised and living in the U.S.
  - may be restricted to certain communicative situation
- Also applies to definition of sublanguage
  - dialect (Bostonian, Cockney), social group (teenagers), genre (advertising), domain (statistics), ...

Intensional vs. extensional

- We have given an intensional definition for the language of interest
  - characterised by speakers and circumstances
- But does this allow quantitative statements?
  - we need something we can count
- Need extensional definition of language
  - i.e. language = body of utterances
    "All utterances made by speakers of the language under appropriate conditions, plus all utterances they could have made"

Problem #2

- Problem #2: What is "frequency"?

- Obviously, extensional definition of language must comprise an infinite body of utterances

  - So, how many passives are there in English?
  - ∞ ...infinitely many, of course!

- Only relative frequencies can be meaningful

Relative frequency

- How many passives are there ...
  - ...per million words?
  - ...per thousand sentences?
  - ...per hour of recorded speech?
  - ...per book?
- Are these measurements meaningful?

Relative frequency

- How many passives could there be at the most?
  - every VP can be in active or passive voice
  - frequency of passives is only interpretable by comparison with frequency of potential passives
- comparison with frequency of potential passives
  - What proportion of VPs are in passive voice?
  - easier: proportion of sentences that contain a passive
- Relative frequency = proportion π

Problem #3

- Problem #3: How can we possibly count passives in an infinite amount of text?

- Statistics deals with similar problems:
  - goal: determine properties of large population
    (human populace, objects produced in factory, ...
  - method: take (completely) random sample of objects, then extrapolate from sample to population
  - this works only because of random sampling!

- Many statistical methods are readily available

Statistics & language

- Apply statistical procedure to linguistic problem
  - take random sample from (extensional) language
  - What are the objects in our population?
    - words? sentences? texts? ...
  - Objects = whatever proportions are based on → unit of measurement
- We want to take a random sample of these units

Types vs. tokens

- Important distinction between types & tokens
  - we might find many copies of the "same" VP in our sample, e.g. *click this button (software manual)* or *includes dinner, bed and breakfast*
- sample consists of occurrences of VPs, called tokens
  - each token in the language is selected at most once
- distinct VPs are referred to as types
  - a sample might contain many instances of the same type
- Definition of types depends on the research question

Types vs. tokens

- Example: **Word Frequencies**
  - word type = dictionary entry (distinct word)
  - word token = instance of a word in library texts
- Example: **Passives**
  - relevant VP types = active or passive ($\rightarrow$ abstraction)
  - VP token = instance of VP in library texts

Types, tokens and proportions

- Proportions in terms of types & tokens
- Relative frequency of type v
  $=$ proportion of tokens t_i that belong to this type

$$p = \frac{f(v)}{n} \tag{2}$$

  - $f(v) =$ frequency of type
  - $n =$ sample size

Inference from a sample

- Principle of inferential statistics
  - if a sample is picked at random, proportions should be roughly the same in the sample and in the population
- Take a sample of, say, 100 VPs
  - observe 19 passives $\rightarrow p = 19\% = .19$
  - style guide $\rightarrow$ population proportion $\pi = 15\%$
  - $p > \pi \rightarrow$ reject claim of style guide?
- Take another sample, just to be sure
  - observe 13 passives $\rightarrow p = 13\% = .13$
  - $p < \pi \rightarrow$ claim of style guide confirmed?

Problem #4

- Problem #4: Sampling variation
  - random choice of sample ensures proportions are the same on average in sample and in population
  - but it also means that for every sample we will get a different value because of chance effects → **sampling variation**
- The main purpose of statistical methods is to estimate & correct for sampling variation
  - that's all there is to statistics, really ☺

Estimating sampling variation

- Assume that the style guide's claim is correct
  - the null hypothesis H_0, which we aim to refute

$$H_0 : \pi = .15$$

  - we also refer to $\pi_0 = .15$ as the null proportion

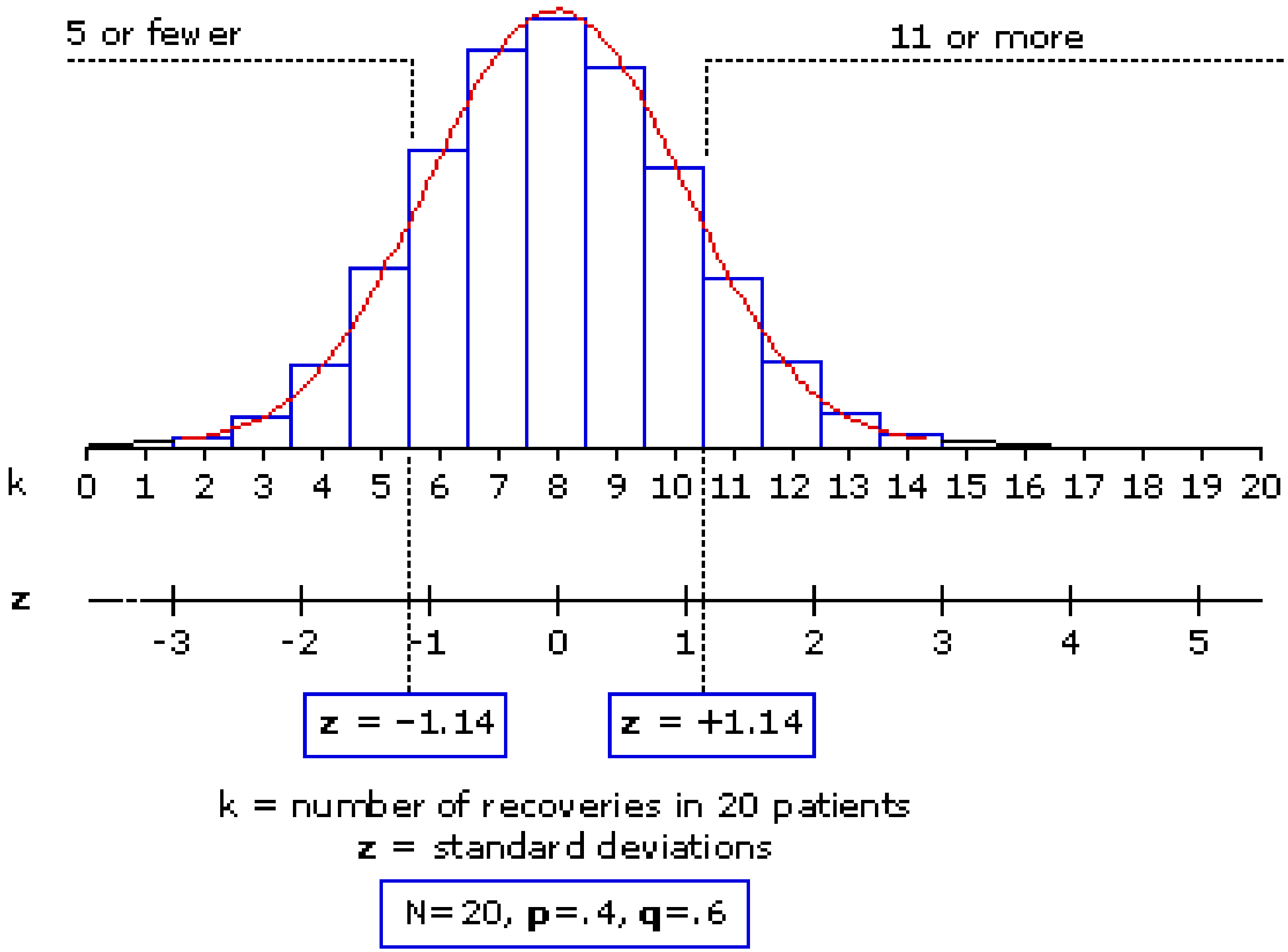
- Many corpus linguists set out to test H_0
  - each one draws a random sample of size $n = 100$
  - how many of the samples have the expected $k = 15$ passives, how many have $k = 19$, etc.?

Estimating sampling variation

- We don't need an infinite number of monkeys (or corpus linguists) to answer these questions
  - randomly picking VPs from our metaphorical library is like drawing balls from an infinite urn
  - red ball = passive VP / white ball = active VP
  - H_0: assume proportion of red balls in urn is 15
- This leads to a binomial distribution

$$\frac{(\pi_0)(1-\pi_0)}{N}$$

Binomial Sampling Distribution for $N = 20, \pi = .4$



Statistical hypothesis testing

- Statistical hypothesis tests
  - define a rejection criterion for refuting H_0
  - control the risk of false rejection (type I error) to a "socially acceptable level" (significance level)
  - p-value $=$ risk of false rejection for observation
  - p-value interpreted as amount of evidence against H_0

- Two-sided vs. one-sided tests
  - in general, two-sided tests should be preferred
  - one-sided test is plausible in our example

Error Types

System	Actual	
	target	not target
selected	tp	fp
not selected	fn	tn

Precision $= \frac{tp}{tp+fp}$; Recall $= \frac{tp}{tp+fn}$; $\mathrm{F}_1 = \frac{2PR}{P+R}$

tp True positives: system says Yes, target was Yes

fp False positives: system says Yes, target was No (Type I Error)

tn True negatives: system says No, target was No

fn False negatives: system says No, target was Yes (Type II Error)

Example: Similarity

- System says *eggplant* is similar to *brinjal*
  True positive

- System says *eggplant* is similar to *egg*
  depends on the application (both food), but generally not so good
  False positive

- System says *eggplant* is **not** similar to *aubergine*
  False negative

- System says *eggplant* is **not** similar to *laptop*
  True negative

Hypothesis tests in practice

- Easy: use online wizard
  - http://sigil.collocations.de/wizard.html
  - http://vassarstats.net/
- Or **Python**
  - One-tail test: `scipy.stats.binom.sf(k, n, p)`
    k = number of successes, p = number of trials,
    p = hypothesized probability of success
    returns p-value of the hypothesis test
  - Two-tail test: `scipy.stats.binom_test(k,n,p)`
- Or **R** http://www.r-project.org/

Confidence interval

- We now know how to test a null hypothesis H_0, rejecting it only if there is sufficient evidence
- But what if we do not have an obvious null hypothesis to start with?
  - this is typically the case in (computational) linguistics
- We can estimate the true population proportion from the sample data (relative frequency)
  - sampling variation $\rightarrow$ range of plausible values
  - such a confidence interval can be constructed by inverting hypothesis tests (e.g. binomial test)

Confidence intervals

- Confidence interval = range of plausible values for true population proportion
  We know the answer is almost certainly more than X and less than Y
- Size of confidence interval depends on sample size and the significance level of the test
- The larger your sample, the narrower the interval will be
  that is the more accurate your estimate is
  - $19/100 \rightarrow$ 95% confidence interval: [12.11% ... 28.33%]
  - $190/1000 \rightarrow$ 95% confidence interval: [16.64% ... 21.60%]
  - $1900/10000 \rightarrow$ 95% confidence interval: [18.24% ... 19.79%]
- http://sigil.collocations.de/wizard.html

Effect Size

- The difference between the two values is the effect size
- For something to be significant the two confidence intervals should not overlap
  - either a small confidence interval (more data)
  - or a big effect size (clear difference)
- lack of significance does not mean that there is no difference only that we are not sure
- significance does not meant that there is a difference only that it is unlikely to be by chance
- if we run an experiment with 20 different configurations and one is significant what does this tell us?

Frequency comparison

- Many linguistic research questions can be operationalised as a frequency comparison
  - Are split infinitives more frequent in AmE than BrE?
  - Are there more definite articles in texts written by Chinese learners of English than native speakers?
  - Does *meow* occur more often in the vicinity of *cat* than elsewhere in the text?
  - Do speakers prefer *I couldn't agree more* over alternative compositional realisations?

- Compare observed frequencies in two samples

Frequency comparison

k_1	k_2
$n_1 - k_1$	$n_2 - k_2$

19	25
81	175

- Contingency table for frequency comparison
  - e.g. samples of sizes $n_1 = 100$ and $n_2 = 200$, containing 19 and 25 passives
  - H_0: same proportion in both underlying populations
- Chi-squared X^2, likelihood ratio G^2, Fisher's test
  - based on same principles as binomial test

Frequency comparison

- Chi-squared, log-likelihood and Fisher are appropriate for different (numerical) situations

- Estimates of effect size (confidence intervals)
  - e.g. difference or ratio of true proportions
  - exact confidence intervals are difficult to obtain
  - log-likelihood seems to do best for many corpus measures

- Frequency comparison in practice
  - `http://sigil.collocations.de/wizard.html`

Do Particle verbs correlate with compound verbs?

- Compound verbs (VV): *hikari-kagayaku* "shine-sparkle"; *kaki-ageru* "write up (lit: write-rise)"
- Particle Verbs (PV): *give up*, *write up*
- Syntactically they are quite different
- Semantically they are both examples of verb+satellite
  - One verb varies a lot
  - The satellite is more constrained and is often spatial

For more details see Soh, Xue Ying Sandra (2015)

Do Particle verbs correlate with compound verbs? II

Look at all Japanese-English verb pairs from Wordnet

	PV	V	total
VV	1,777	5,885	7,662
V	10,877	51,137	62,014
total	12,654	57,022	69,676

- What is the confidence interval for the distribution of VV in Japanese?
  12.36% [12.10% ...12.62%]
- What is the confidence interval for the distribution of PV in English?
  14.92% [14.66% ...15.18%]
- How many PV=VV would you expect if they were independent?
  $0.1236 \times 0.1492 \times 69676 = 1,284$
- Is PV translated as VV more than chance?
  upper bound is $0.1262 \times 0.1518 \times 69676 = 1,334.8$

Do Particle verbs correlate with compound verbs? III

- You can test with a Chi Square Calculator for 2x2
- https://www.socscistatistics.com/tests/chisquare/default.aspx
  The chi-square statistic is 146.6123.
  The p-value is < 0.00001. Significant at p $< .05$.
- A chi-square test of independence was performed to examine the relation between verb satellites in Japanese (VV) and English (PV). The relation between these variables was significant, χ^2 (1, N = 69,676) = 146, p = $<$ 0.00001. VV are more likely to be translated as PV.

Collocations

Outline

- Collocations & Multiword Expressions (MWE)
  - What are collocations?
  - Types of cooccurrence
- Quantifying the attraction between words
  - Contingency tables

What is a collocation?

- Words tend to appear in typical, recurrent combinations:
  - *day* and *night*
  - *ring* and *bell*
  - *milk* and *cow*
  - *kick* and *bucket*
  - *brush* and *teeth*

- such pairs are called **collocations** (Firth, 1957)

- the meaning of a word is in part determined by its characteristic collocations

  “You shall know a word by the company it keeps!”

What is a collocation?

- Native speakers have strong and widely shared intuitions about such collocations
  - Collocational knowledge is essential for non-native speakers in order to sound natural
  - This is part of “idiomatic language”

An important distinction

- **Collocations** are an empirical linguistic phenomenon
  - can be observed in corpora and quantified
  - provide a window to lexical meaning and word usage
  - applications in language description (Firth, 1957) and computational lexicography (Sinclair, 1991)
- **Multiword expressions** = lexicalised word combinations
  - MWE need to be lexicalised (i.e., stored as units) because of certain idiosyncratic properties
  - non-compositionallity, non-substitutability, non-modifiability (Manning and Schütze, 1999)
  - not directly observable, defined by linguistic tests (e.g. substitution test) and native speaker intuitions
  - Sometimes called **collocations** but we will distinguish

But what are collocations?

- Empirically, collocations are words that show an attraction towards each other (or a **mutual expectancy**)
  - in other words, a tendency to occur near each other
  - collocations can also be understood as statistically salient patterns that can be exploited by language learners

- Linguistically, collocations are an epiphenomenon of many different linguistic causes that lie behind the observed surface attraction.

epiphenomenon "a secondary phenomenon that is a by-product of another phenomenon"

Collocates of *bucket* (n.)

noun	f	verb	f	adjective	f
water	183	throw	36	large	37
spade	31	fill	29	single-record	5
plastic	36	randomize	9	cold	13
slop	14	empty	14	galvanized	4
size	41	tip	10	ten-record	3
mop	16	kick	12	full	20
record	38	hold	31	empty	9
bucket	18	carry	26	steaming	4
ice	22	put	36	full-track	2
seat	20	chuck	7	multi-record	2
coal	16	weep	7	small	21
density	11	pour	9	leaky	3
brigade	10	douse	4	bottomless	3
algorithm	9	fetch	7	galvanised	3
shovel	7	store	7	iced	3
container	10	drop	9	clean	7
oats	7	pick	11	wooden	6

Collocates of bucket (n.)

- opaque **idioms** (*kick the bucket*, but often used literally)
- **proper names** (*Rhino Bucket*, a hard rock band)
- **noun compounds**, lexicalised or productively formed (*bucket shop, bucket seat, slop bucket, champagne bucket*)
- **lexical collocations** = semi-compositional combinations (*weep buckets, brush one's teeth, give a speech*)
- cultural **stereotypes** (*bucket and spade*)
- **semantic compatibility** (*full, empty, leaky <u>bucket</u>; throw, carry, fill, empty, kick, tip, take, fetch a <u>bucket</u>*)

- **semantic fields** (*shovel, mop*; hypernym *container*)
- **facts of life** (*wooden bucket; bucket of water, sand, ice, ...*)
- often sense-specific (*bucket size, randomize to a bucket*)

Operationalising collocations

- Firth introduced collocations as an essential component of his methodology, but without any clear definition

  > Moreover, these and other technical words are given their 'meaning' by the restricted language of the theory, and by applications of the theory in quoted works. (Firth 1957, 169)

- Empirical concept needs to be formalised and quantified
  - intuition: collocates are "attracted" to each other, i.e. they tend to occur near each other in text
  - definition of "nearness" $\rightarrow$ cooccurrence
  - quantify the strength of attraction between collocates based on their recurrence $\rightarrow$ cooccurrence frequency
  - We will consider word pairs (w_1, w_2) such as (*brush, teeth*)

Different types of cooccurrence

1. **Surface cooccurrence**
   - ➢ criterion: surface distance measured in word tokens
   - ➢ words in a **collocational span** (or **window**) around the node word, may be symmetric (L5, R5) or asymmetric (L2, R0)
   - ➢ traditional approach in lexicography and corpus linguistics

2. **Textual cooccurrence**
   - ➢ words cooccur if they are in the same text segment (sentence, paragraph, document, Web page, . . .)
   - ➢ often used in Web-based research ($\rightarrow$ Web as corpus)
   - ➢ often used in indexing

3. **Syntactic cooccurrence**

- words in a specific syntactic relation
  - adjective modifying noun
  - subject/object noun of verb
  - *N of N*
- suitable for extraction of MWEs (Krenn and Evert 2001)

* Of course you can combine these

Surface cooccurrence

- Surface cooccurrences of $w_1 =$ *hat* with $w_2 =$ *roll*
- symmetric window of four words (L4, R4)
- limited by sentence boundaries

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat . A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. [. . .] There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it . The wind puffed, and Mr. Pickwick puffed, and the hat rolled over and over as merrily as a lively porpoise in a strong tide ; and on it might have rolled, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

- coocurrence frequency $f = 2$
- marginal frequencies $f_1(\text{hat}) = f_2(\text{roll}) = 3$

Textual cooccurrence

- Surface cooccurrences of $w_1 =$ *hat* with $w_2 =$ *over*
- textual units = sentences
- multiple occurrences within a sentence ignored

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a hat.	hat	—
A man must not be precipitate, or he runs over it ;	—	over
he must not rush into the opposite extreme, or he loses it altogether.	—	—
There was a fine gentle wind, and Mr. Pickwick's hat rolled sportively before it.	hat	—
The wind puffed, and Mr. Pickwick puffed, and the hat rolled over and over as merrily as a lively porpoise in a strong tide ;	hat	over

- cooccurrence frequency $f = 1$

- marginal frequencies $f_1 = 3$, $f_2 = 2$

Syntactic cooccurrence

- Syntactic cooccurrences of adjectives and nouns
- every instance of the syntactic relation (A-N) is extracted as a pair token
- Cooccurrency frequency data for young gentleman:
  - *There were two gentlemen who came to see you.*
    (*two, gentleman*)
  - *He was no gentleman, although he was young.*
    (*no, gentleman*) (*young, he*)
  - *The old, stout gentleman laughed at me.*
    (*old, gentleman*) (*stout, gentleman*)
  - *I hit the young, well-dressed gentleman.*
    (*young, gentleman*) (*well-dressed gentleman*)
  - coocurrence frequency $f = 1$
  - marginal frequencies $f_1 = 2, f_2 = 6$

Quantifying attraction

- Quantitative measure for attraction between words based on their recurrence $\rightarrow$ **cooccurrence frequency**

- But cooccurrence frequency is not sufficient
  - bigram *is to* occurs $f = 260$ times in Brown corpus
  - but both components are so frequent ($f_1 \approx 10,000$ and $f_2 \approx 26,000$) that one would also find the bigram 260 times if words in the text were arranged in completely random order
  - take expected frequency into account as **baseline**

- Statistical model required to bring in notion of **chance cooccurrence** and to adjust for sampling variation

- bigrams can be understood either as syntactic cooccurrences (adjacency relation) or as surface cooccurrences (L1, R0 or L0, R1)

What is an n-gram?

- An n-gram is a subsequence of n items from a given sequence. The items in question are typically phonemes, syllables, letters, words or base pairs according to the application.
  - n-gram of size 1 is referred to as a **unigram**;
  - size 2 is a **bigram** (or, less commonly, a **digram**)
  - size 3 is a **trigram**
  - size 4 or more is simply called an n**-gram**
- **bigrams** (from the first sentence): BOS *An, An n-gram, n-gram is, is a, a subsequence, subsequence of, ...*
- 4**-grams** (from the first sentence): BOS *An n-gram is, An n-gram is a, n-gram is a subsequence, is a subsequence of, ...*

Attraction as statistical association

- Tendency of events to cooccur $=$ **statistical association**
  - statistical measures of association are available for contingency tables, resulting from a cross-classification of a set of "items" according to two (binary) factors cross-classifying factors represent the two events

- Application to word cooccurrence data
  - most natural for syntactic cooccurrences
  - "items" are pair tokens (x, y) $=$ instances of syntactic relation
  - factor 1: Is x an instance of word type w_1?
  - factor 2: Is y an instance of word type w_2?

Measuring association in contingency tables

	w_1	$\neg w_1$
w_2	both	one
$\neg w_2$	other	neither

- Measures of significance
  - apply statistical hypothesis test with null hypothesis H_0 : independence of rows and columns
  - H_0 implies there is no association between w_1 and w_2
  - association score $=$ test statistic or p-value
  - one-sided vs. two-sided tests

  $\rightarrow$ amount of evidence for association between w_1 and w_2

- Measures of effect-size
  - compare observed frequencies O_{ij} to expected frequencies E_{ij} under H_0
  - or estimate conditional prob. $\Pr(w_2 \mid w_1)$, $\Pr(w_1 \mid w_2)$, etc.
  - maximum-likelihood estimates or confidence intervals

  $\rightarrow$ strength of the attraction between w_1 and w_2

Interpreting hypothesis tests as association scores

- Establishing significance
  - p-value $=$ probability of observed (or more "extreme") contingency table if H_0 is true
  - theory: H_0 can be rejected if p-value is below accepted significance level (commonly .05, .01 or .001)
  - practice: nearly all word pairs are highly significant

- Test statistic $=$ significance association score
  - convention for association scores: high scores indicate strong attraction between words
  - Fisher's test: transform p-value, e.g.$-log_{10}$p
  - satisfied by test statistic X^2, but not by p-value
  - Also log-likelihood G^2

- In practice, you often just end up ranking candidates
  different measures give similar results
  there is no perfect statistical score

Acknowledgments

- Thanks to Stefan Th. Gries (University of California, Santa Barbara) for his great introduction *Useful statistics for corpus linguistics*: https://pdfs.semanticscholar.org/069f/d4b980d75af1125afa2ac3c8f5af2e66414d.pdf

- Many slides inspired by Marco Baroni and Stefan Evert's **SIGIL** *A gentle introduction to statistics for (computational) linguists* http://www.stefan-evert.de/SIGIL/

- Some examples taken from Ted Dunning's *Surprise and Coincidence - musings from the long tail* http://tdunning.blogspot.com/2008/03/surprise-and-coincidence.html