Quantitative study of corpora

Frequency lists and collocations

Serge Sharoff

Centre for Translation Studies
University of Leeds

s.sharoff@leeds.ac.uk

Data science for translators

- ChatGPT and other AI models: Large Language Models

→ Large corpora: 90,000 years of reading for ChatGPT

- Statistics for understanding how LLMs and MT engines work
- Future careers for engaging with model developers

History of corpus development

孝 – Critical studies in China: about 400 BC

- Concordances: Eusebius of Caesaria (320) on canon tables
- Early corpora: stenography (Käding), language learning (GSL)
- Computer corpora (1960s, 1MW): Brown Corpus (American)
- Megacorpus era (1990s, 100+MW): Bank of English and the British National Corpus (BNC)
- Internet corpora (2000s, 1+GW): ukWac, en10^{10}

	BC	BNC(a)	BNC(ipm)	ukWac	en10^{10}
soccer	1	1,321	13	8	8
integrity	10	1467	15	16	23
undermine integrity	0	11	0.11	0.07	0.10
year	1,589	163,930	1,639	1,631	1,425

How to count words

- How many words are there in English?
- A cat is on a mat. 32,200 cats are on a place-mat.
- Tokens: sequences separated by punctuation (12-14)

?? N-acetyl-p-aminophenol, trimethyl-xanthine-dione

?? translation-oriented, Hong Kong, d'Arcy, John's, I'm

?? Other languages: Arabic, Chinese, German
Fachhochschulratspräsident
点击进入联合国安全理事会网站，了解更多信息。

- Types: a set of equivalent tokens

?? word forms or lemmas: cat vs. cats, *be*/*are*
or lemmas+POS: kind.NN vs. kind.JJ

- Ranking types in frequency lists

?? I do uh main- mainly business management

Vocabulary size

Corpus	Tokens = N	Types = $\lvert V \rvert$
Shakespeare	884 K	31 K
Brown corpus	1 M	38 K
Switchboard	2.4 M	20 K
BNC	100M	665K
ukWac	2GW	11 M

- Around 100,000 in ranking (247 examples in ukWac):
  Foch Havard deliriously genotypic under-15 wana zucchini
- Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch:
  Rank in BNC: 543226, 1 example
  Rank in ukWac: 294309, 41 examples

→ We need many texts for good lexical coverage
Texts: 100MW=290,000 pages (350 W/p)
Speech: 1,000,000 hr (100 W/m)

- 2810 instances of f**k in BNC (spoken)
  Does this mean that people swear frequently?

The BNC frequency list

Rank	Lemma	POS	AbsFrq	Frq(ipm)	Frq%	Coverage	Cov%
1	the	det	6187267	61872.67	6.187%	61872.67	6.19%
2	be	v	4239632	42396.32	4.240%	104268.99	10.43%
3	of	prep	3093444	30934.44	3.093%	135203.43	13.52%
4	and	conj	2687863	26878.63	2.688%	162082.06	16.21%
5	a	det	2186369	21863.69	2.186%	183945.75	18.39%
6	in	prep	1924315	19243.15	1.924%	203188.90	20.32%
7	to	inf-to	1620850	16208.50	1.621%	219397.40	21.94%
8	have	v	1375636	13756.36	1.376%	233153.76	23.32%
9	it	pron	1090186	10901.86	1.090%	244055.62	24.41%
10	to	prep	1039323	10393.23	1.039%	254448.85	25.44%
2000	connect	v	4510	45.10	0.005%	775928.30	77.59%
2001	fundamental	a	4508	45.08	0.005%	775973.38	77.60%
2002	plane	n	4505	45.05	0.005%	776018.43	77.60%
2003	height	n	4505	45.05	0.005%	776063.48	77.61%
2004	opening	n	4504	45.04	0.005%	776108.52	77.61%
2005	lesson	n	4503	45.03	0.005%	776153.55	77.62%
2006	similarly	adv	4502	45.02	0.005%	776198.57	77.62%
2007	shock	n	4502	45.02	0.005%	776243.59	77.62%
2008	rail	n	4502	45.02	0.005%	776288.61	77.63%

Problems with frequency lists

- the object of counting (colour, gonna, with respect to)
- the “whelks” problem (Adam Kilgarriff)

  A marine gastropod mollusc of the genus Buccinum, having a turbinate shell, esp. B. undatum, common on the European and North American coasts, much used for food. (OED)
- Frequency spikes (whelk problem);
  comply, therapy, exhaust, gastric, swimming, darling, celebration
  mushroom, outrage, presently, absorptance, retrieve, dirt, skipper

  Journal of Gastroentorology and Hepatology: 713 kW in the BNC
  peptide, endoscopy: the top 3000 BNC words
- Frequency drops (banana-toothbrush problem)
  anchor, instrumental, sodium, banana, tilt, hunter, armour
  leer, enthrall, sheaf, toothbrush, dungeon, stocky, lawsuit
- reliability: what happens with another corpus?

Frequency lists from three corpora

the	22905	the	7802100	the	6187267
of	12710	be	4523108	be	4239632
be	10686	to	3409757	of	3093444
a	9952	of	3338835	and	2687863
and	8323	a	3337996	a	2186369
in	7010	and	3174355	in	1924315
to	6502	in	2622013	to	1620850
that	4392	have	1623255	have	1375636
price	3080	that	1594191	it	1090186
for	2912	for	1296688	to	1039323
it	2674	say	1126948	for	887877
we	2534	it	1097742	i	884599
have	2514	he	1013629	that	760399
cost	2251	on	972005	you	695498
by	2034	with	924460	he	681255
this	2003	not	912954	on	680739
demand	1944	as	784007	with	675027
on	1882	at	739731	do	559596

BNC vs. New York Times

Rank	Frq	Lemma	Rank	Frq	Lemma
3983	2560	environmentalist	3983	1822	accent
3984	2559	casual	3984	1822	elder
3985	2559	scratch	3985	1822	twentieth
3986	2557	troy	3986	1822	vietnam
3987	2556	petition	3987	1821	unnecessary
3988	2555	pipe	3988	1821	underneath
3989	2554	roast	3989	1819	invent
3990	2554	genre	3990	1819	timing
3991	2554	merchant	3991	1819	recipe
3992	2551	canyon	3992	1818	hungry
3993	2551	flip	3993	1818	morgan
3994	2550	automatic	3994	1817	autonomy
3995	2549	efficient	3995	1816	cave
3996	2549	grind	3996	1815	delegation
3997	2549	bug	3997	1815	tactic
3998	2548	ongoing	3998	1814	diagram
3999	2547	fatal	3999	1814	influential

Comparing frequencies

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Corpus size	c	d	c+d

- $loglike = 2(a\ln(\frac{a}{E1}) + b\ln(\frac{b}{E2})); E1 = c\frac{a+b}{c+d}; E2 = d\frac{a+b}{c+d}$

Word	Corpus1	Corpus2	LL-score
price	147048	26741	910+
you	39749	603306	6005-
put	38897	61016	51-

		BNC (spoken)	BNC (written)	ukWac
HW	f**k	2810	1603	16309
	Corpus size	10M	90M	2000M

A keyword list

More in BNC	LL-score	More in NYT	LL-score
you	6005.14	say	8559.54
I	5271.42	percent	4513.35
she	3334.57	bush	2364.29
be	2411.89	gore	1982.47
do	1610.71	president	1518.25
they	1502.79	atlanta	1468.84
your	1282.15	game	1258.34
can	1191.74	clinton	1240.37
what	1090.53	york	1214.84
my	1023.56	news	1199.25

A keyword list

More frequent in I-DE			More frequent in IDS		
Word form	Gloss	LLscore	Word form	Gloss	LLscore
ich	I	1,789.63	Mark	Mark	796.69
dass	that (new)	1226.98	Uhr	hour	476.57
mir	me$_{dat}$	533.53	Prozent	percent	302.65
wir	we$_{nom}$	515.32	daß	that (old)	307.32
Sie	you$_{pol}$	469.46	sei	be-subjunc	291.95
du	you$_{fam}$	376.29	dpa	dpa	262.05
mich	me$_{acc}$	458.73	bis	to-temporal	258.87
oder	or	432.71	Millionen	millions	235.37
Ich	I	416.26	gestern	yesterday	225.47
du	you$_{fam}$	297.20	SPD	SPD	181.97
kann	can	295.89	sagt	said	177.19
uns	us$_{acc}$	284.49	Franken	franc	127.02
the	-	282.68	taz	taz	120.24

Putting frequencies into dictionaries

- Frequency counts for words (GSL, CCED, LDOCE)
- Different indications:

	CCED	LDOCE
go	♦♦♦♦♦	S1, W1
significant	♦♦♦♦	S3, W1
calm	♦♦♦	S3, W3
polish	♦♦	-,-
bungalow	♦	S3, -
sanction	♦♦♦♦	-, -

Comparing frequencies and translations

- Intellitext
  `http://corpus.leeds.ac.uk/itweb/`
- Kelly database
  `http://kelly.sketchengine.co.uk/`
- Loglikelihood calculator:
  `http://ucrel.lancs.ac.uk/llwizard.html`
- The SketchEngine: `https://app.sketchengine.eu/`

Basic points

- Size matters: 100 MW for moderately frequent words
- Counting words: tokens and types
- Relative frequency and coverage
- Reliability of frequency lists
- Comparing frequencies using Intellitext

For the next session

Find frequency lists for your languages

To read:

Ch. 3 of McEnery, Wilson, Corpus Linguistics
Unit A6 of McEnery, Xiao, Tono, Corpus-based language studies: an advanced resource book
Ch. 2 and 3 of Biber, Reppen, The Cambridge Handbook of English Corpus Linguistic