SCOTTISH GRADUATE SCHOOL FOR ARTS & HUMANITIES

Scottish Graduate School of Social Science
Sgoil Cheumnaichean Saidheans Sòisealta na h-Alba

Corpus Analysis

1400 Monday 16 April

marc.alexander@glasgow.ac.uk

Quantify Your Enjoyment

- Qualitative/quantitative distinction
  - Empirical, evidence, rigor, explicit, objective
  - Insight, interpretation, subjective, critical
- Quantitative very useful, but can be something most humanists avoid
- But quantitative can sometimes also be the only option!

Descriptive Statistics
(counting stuff)

- Raw numbers of occurrences
- Occurrences as percentage of the corpus
  - Relative context in the full corpus
- Normalised frequency
  - per thousand words/per million words
    - (divide your results by how many thousands or millions of words there are in the corpus)

Special Statistics
(counting stuff and doing maths)

Type/token ratio

Imagine a corpus of 400,000 words
=400,000 *tokens*
But lots of these words will be repeated over and over again!
Actually, there may be only 140,000 *different* words used
=140,000 *types*
Number of types divided by the number of tokens gives you the type/token ratio
In this case, 35%.

Special Statistics
(counting stuff and then doing maths)

Type/token ratio (TTR)

Why?

'lexical diversity'/'vocabulary richness'

As well as the obvious problems, very dependent on the size of the corpus.

Standardised TTR (STTR) does the same but averages the figure over 'chunks' of data (WordSmith, STELLA)

Special Statistics
(counting stuff and then doing maths)

Type/token ratio (TTR)

Fraser's workshop description (364 words) has 206 unique words, and so a TTR of 56.6%

Funnybones: 56 tokens, 18 types, TTR 32%

Brown Corpus: 1,023,243 tokens, 41,144 types, TTR 4%

Thomas Jefferson's writings: 2,392,159 tokens, 42,841 types, TTR 1.7%

Special Statistics
(counting stuff and then doing maths)

Type/token ratio (TTR)

So perhaps not as useful as many people made it out to be!

But a handy first statistic when comparing two similarly-sized corpora!

See also: Hoover, David. 2003. Another Perspective on Vocabulary Richness. *Computers and the Humanities*. 37(2), 2003: 151-78.

Significance Statistics
(counting stuff and then doing hard maths)

What is ‘significant’?

Statistical unlikelihood if we assume things are normal

Often approximated by ‘keyness’

‘Key words are those whose frequency is unusually high in comparison with some norm’ (Mike Scott)

Keyness

Keyness tests “...compare the difference between the actual frequencies observed in the corpus (the observed frequencies) and the frequencies we would expect if no factor other than chance had been operating to affect the frequencies (the expected frequencies).”

“The closer the expected frequencies are to the observed frequencies, the more likely it is that the observed frequencies are a result of chance.”

“On the other hand, the greater the difference between the observed frequencies and the expected frequencies, the more likely it is that the observed frequencies are being influenced by something other than chance.”

(McEnery and Wilson 2001: 84-85)

Wait, wait, wait…

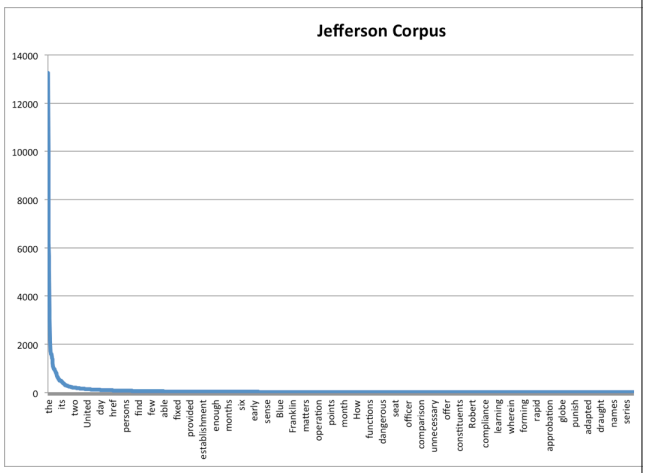
What does language look like, statistically?

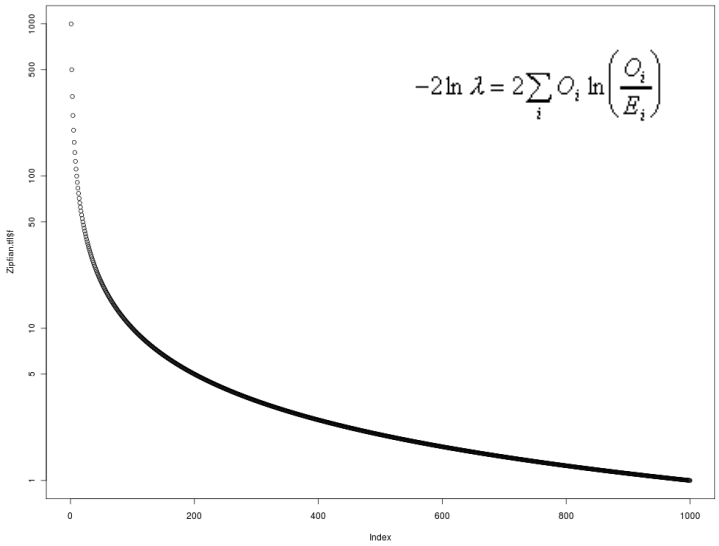
What’s “normal” in frequencies of words in natural language?

BNC Frequencies

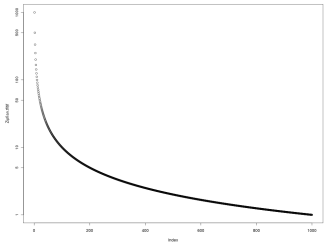
Word	PoS	Freq	Word	PoS	Freq	Word	PoS	Freq
the	Det	61847	as	Prep	1774	're	Verb	835
of	Prep	29391	them	Pron	1733	yeah	Int	834
and	Conj	26817	some	DetP	1712	three	Num	797
a	Det	21626	when	Conj	1712	good	Adj	795
in	Prep	18214	could	VMod	1683	back	Adv	793
to	Inf	16284	him	Pron	1649	make	Verb	791
it	Pron	10875	into	Prep	1634	such	DetP	763
is	Verb	9982	its	Det	1632	on	Adv	756
to	Prep	9343	then	Adv	1595	there	Adv	746
was	Verb	9236	two	Num	1561	through	Prep	743
I	Pron	8875	out	Adv	1542	year	NoC	737
for	Prep	8412	time	NoC	1542	over	Prep	735
that	Conj	7308	my	Det	1525	'll	VMod	726
you	Pron	6954	about	Prep	1524	must	VMod	723
he	Pron	6810	did	Verb	1434	still	Adv	718
be*	Verb	6644	your	Det	1383	even	Adv	716
with	Prep	6575	now	Adv	1382	take	Verb	715
on	Prep	6475	me	Pron	1364	too	Adv	701
by	Prep	5096	no	Det	1343	more	DetP	699
at	Prep	4790	other	Adj	1336	here	Adv	699
have*	Verb	4735	only	Adv	1298	own	DetP	695
are	Verb	4707	just	Adv	1277	come	Verb	695
not	Neg	4626	more	Adv	1275	last	Ord	691
this	DetP	4623	these	DetP	1254	does	Verb	687
's	Gen	4599	also	Adv	1248	oh	Int	684
but	Conj	4577	people	NoC	1241	say	Verb	679
had	Verb	4452	know	Verb	1233	no	Int	662
they	Pron	4332	any	DetP	1220	going*	Verb	658
his	Det	4285	first	Ord	1193	'm	Verb	658
from	Prep	4134	see	Verb	1186	work	NoC	653
she	Pron	3801	very	Adv	1165	where	Adv	628
that	DetP	3792	new	Adj	1145	erm	Uncl	627
which	DetP	3719	may	VMod	1135	us	Pron	623
or	Conj	3707	well	Adv	1119	government	NoC	622
we	Pron	3578	should	VMod	1112	same	DetP	615
's	Verb	3490	her*	Pron	1085	man	NoC	614
an	Det	3430	like	Prep	1064	might	VMod	614
–n't	Neg	3328	than	Conj	1033	day	NoC	610
were	Verb	3227	how	Adv	1016	yes	Int	606
as	Conj	3006	get	Verb	995	however	Adv	605
do	Verb	2802	way	NoC	958	put	Verb	596
been	Verb	2686	one	Pron	953	world	NoC	590
their	Det	2608	our	Det	950	over	Adv	584
has	Verb	2593	made	Verb	943	another	DetP	581
would	VMod	2551	got	Verb	932	in	Adv	573
there	Ex	2532	after	Prep	927	want	Verb	572
what	DetP	2493	think	Verb	916	as	Adv	567
will	VMod	2470	between	Prep	903	life	NoC	566
all	DetP	2436	many	DetP	902	most	Adv	565
if	Conj	2369	years	NoC	902	against	Prep	562
can	VMod	2354	er	Uncl	896	again	Adv	561
her*	Det	2183	've	Verb	891	never	Adv	559
said	Verb	2087	those	DetP	888	under	Prep	553
who	Pron	2055	go	Verb	881	old	Adj	544
one	Num	1962	being	Verb	862	much	DetP	531
so	Adv	1893	because*	Conj	852	something	Pron	526
up	Adv	1795	down	Adv	845	Mr	NoC	524

13263	**the**
10139	**of**
6175	**to**
5651	**and**
3400	**in**
3055	**a**
2247	**that**
2143	**be**
1741	**by**
1705	**it**
1585	**is**
1582	**which**
1563	**for**
1472	**as**
1428	**their**
1385	**on**
1216	**with**
1068	**our**
1034	**from**
1009	**have**

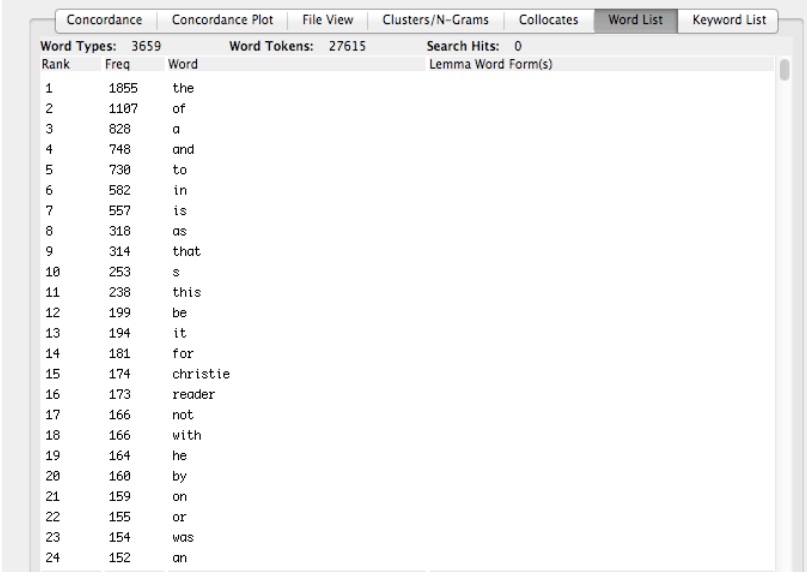


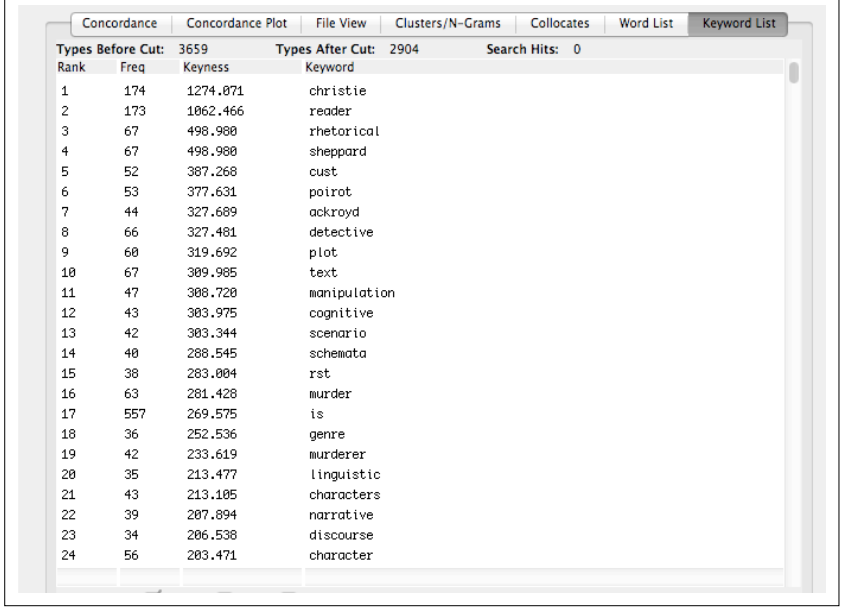


Rank	Freq	Word
1	69968	the
2	36449	of
3	28905	and
4	26232	to
5	23495	a
6	21402	in
7	10786	that
8	10105	is
9	9816	was
10	9800	he
11	9495	for
12	9094	it
13	7280	with
14	7249	as
15	6995	his
16	6760	on
17	6381	be
18	6252	s
19	5943	i
20	5376	at



Rank	Freq	Word
25323	2	westerner
25463	2	zurich
25669	1	aerodynamic
28022	1	citrus
29257	1	digestible
30536	1	filaments
37602	1	scooping
41132	1	zoe
41133	1	zombie
41134	1	zombies
41135	1	zoned
41136	1	zooey
41137	1	zoology
41138	1	zoomed
41139	1	zooming
41140	1	zooms
41141	1	zoooop
41142	1	zorrillas
41143	1	zu
41144	1	zwei





Rank	Freq	Keyness	Keyword
1	174	1274.071	christie
2	173	1062.466	reader
3	67	498.980	rhetorical
4	67	498.980	sheppard
5	52	387.268	cust
6	53	377.631	poirot
7	44	327.689	ackroyd
8	66	327.481	detective
9	60	319.692	plot
10	67	309.985	text
11	47	308.720	manipulation
12	43	303.975	cognitive
13	42	303.344	scenario
14	40	288.545	schemata
15	38	283.004	rst
16	63	281.428	murder
17	557	269.575	is
18	36	252.536	genre
19	42	233.619	murderer
20	35	213.477	linguistic
21	43	213.105	characters
22	39	207.894	narrative
23	34	206.538	discourse
24	56	203.471	character

Sampling

What does your corpus sample?

Does it represent ‘language’?

Biber says ‘everyday’ language should be around 90% conversation, 3% notes/letters [nowadays, emails?], and 7% things like press, academic prose, fiction, lectures, news, magazines, etc etc

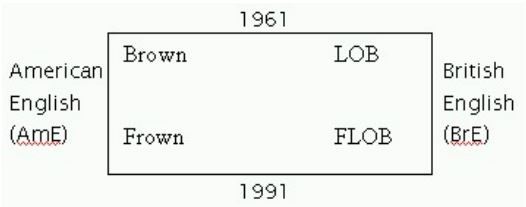
But some texts have a disproportionate influence on language and culture…

Sampling

Does it sample the possible language that there is out there?

Does it have a set structure that means it can be compared? (Prioritizing comparison over representativeness.)

Sampling



Also BLOB (before LOB): 1931 British English

http://www.lancs.ac.uk/fss/courses/ling/corpus/blue/l02_1.htm

Sampling

	Broad text category		Text category letter and description ("genre")	Number of texts			
				Brown	Frown	LOB	FLOB
Informative	Press	A	Press: Reportage	44	same as Brown		
		B	Press: Editorial	27	" " "		
		C	Press: Reviews	17	" " "		
	General Prose	D	Religion	17	" " "		
		E	Skills, Trades and Hobbies	36		38	
		F	Popular Lore	48		44	
		G	Belles Lettres, Biographies, Essays	75		77	
	Learned writing	H	Miscellaneous: Government documents, industrial reports etc.	30	same as Brown		
		J	Science	80	" " "		
Imaginative	Fiction	K	General Fiction	29	" " "		
		L	Mystery and Detective Fiction	24	" " "		
		M	Science fiction	6	" " "		
		N	Adventure and Western	29	" " "		
		P	Romance and Love story	29	" " "		
		R	Humour	9	" " "		

http://www.lancs.ac.uk/fss/courses/ling/corpus/blue/l02_1.htm

ICE corpora:
Canada*
East Africa*
Great Britain
Hong Kong*
India*
Ireland
Jamaica*
New Zealand
The Philippines*
Singapore*
Sri Lanka (written)
USA (written)*

* = freely available from the ICE website (ICE-GB is available in STELLA)

http://ice-corpora.net/ice/design.htm

The design of ICE corpora is as follows:

SPOKEN (300)	**Dialogues** (180)	**Private** (100)	Face-to-face conversations (90) Phonecalls (10)
		Public (80)	Classroom Lessons (20) Broadcast Discussions (20) Broadcast Interviews (10) Parliamentary Debates (10) Legal cross-examinations (10) Business Transactions (10)
	Monologues (120)	**Unscripted** (70)	Spontaneous commentaries (20) Unscripted Speeches (30) Demonstrations (10) Legal Presentations (10)
		Scripted (50)	Broadcast News (20) Broadcast Talks (20) Non-broadcast Talks (10)
WRITTEN (200)	**Non-printed** (50)	**Student Writing** (20)	Student Essays (10) Exam Scripts (10)
		Letters (30)	Social Letters (15) Business Letters (15)
	Printed (150)	**Academic writing** (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Popular writing (40)	Humanities (10) Social Sciences (10) Natural Sciences (10) Technology (10)
		Reportage (20)	Press news reports (20)
		Instructional writing (20)	Administrative Writing (10) Skills/hobbies (10)
		Persuasive writing (10)	Press editorials (10)
		Creative writing (20)	Novels & short stories (20)

Numbers in brackets indicate the number of 2,000-word texts in each category.

Sampling

The question you want to answer dictates what data you need to answer it

'Representativeness' differs in sampling:

Representative of times, of speakers, of people, of places, of another corpus, of...