Measuring Distance between Language Varieties

Adam Kilgarriff, Jan Pomikalek, Pavel Rychly, Vit Suchomel

Supported by EU Project PRESEMT

How to compare language varieties

- Qualitative
- Quantitative

- ***Quantitative means corpus***
  - Corpus represents variety
  - Compare corpora

My big question

- How to compare corpora
  - How else can corpus methods/corpus linguistics be scientific
  - Roles
    - How do varieties contrast
    - How do corpora contrast
      - When we don't know if they are different
    - Find bugs in corpus construction

Corpus comparison

- Qualitative
- Quantitative

Qualitative

- Take keyword lists
  - (a-z){3,}
  - Lemma if lemmatisation identical, else word
  - C1 vs C2, top 100/200
  - C2 vs C1, top 100/200
  - study

Qualitative: example, OCC and OEC

- OEC: general reference corpus
- OCC: writing for children

Look at fiction only

Top 200 keywords (each way)

what are they?

Prep	*above across along down inside like off past round towards*	*as by during in throughout toward until upon within*
Pron	*everybody nobody them they us we*	
Verbs	**Aktionsart** *reach stop* **action/motion** *bend catch climb fly leap lift scramble swim swing twist* **general** *come eat fetch hurry* **modal** *can might must shall will* **perception** *check disappear hide listen peer point see* **reporting** *reckon say shout*	**American English** *gonna* **culture/writing** *edit publish write* **death** *bury die* **general** *attend accept acquire act base become consider continue contribute enter establish figure found include introduce involve obtain produce provide receive remain return serve sigh smile state support survive* **jobs** *appoint promote resign retire succeed* **public affairs** *develop elect review* **relationships** *kiss love marry date*
Other	*all there well*	*although hey oh okay since uh uhm which whom whose yeah*

Do it

- Sketch Engine does the grunt work
- It's ever so interesting

Quantitative

- Methods, evaluation
  - Kilgarriff 2001, *Comparing Corpora,* Int J Corp Ling
  - Then:
    - not many corpora to compare
  - Now:
    - Many
    - Ad hoc, from web
      - First question: is it any good, how does it compare
- Let's make it easy: offer it in Sketch Engine

Original method

- C1 and C2:
  - Same size, by design
  - Put together, find 500 highest freq words
- For each of these words
  - Freqs: f1 in C1, f2 in C2, mean=(f1+f2)/2
  - $(f1-f2)^2/mean$ (chi-square statistic)
- Sum
- Divide by 500: CBDF

Evaluated

- Known-similarity corpora
  - Shows it worked
  - Used to set parameter (500)
  - CBDF better than alternative measures tested

Adjustments for SkE

- Problem: non-identical tokenisation
  - Some awkward words: *can't*
  - undermine stats as one corpus has zero
- Solution
  - commonest 5000 words in each corpus
  - ***intersection only***
  - commonest 500 in intersection

Adjustments for SkE

- Corpus size highly variable
  - Chi-square not so dependable
  - Also not consistent with our keyword lists
    - Link to keyword lists – link quant to qual
- Keyword lists
  - nf = normalised (per million) frequencies
  - Keyword lists: nf1+k/nf2+k
  - Default value for k=100
  - We use: if nf1>nf2, nf1+k/nf2+k, else nf2+k/nf1+k
- Evaluated on Known-Sim Corpora
  - as good as/better than chi-square

	(BASE)	(BAWE)	Corpus	Brown	Family	e-flux	enTenTen	ample)	agging	BeebOx	OEC	l/Port	ukWaC	rewing	ickens	no2_en	PICAE
: Spoken English Corpus (BASE)		3.28	2.77	3.12	2.83	5.19	2.71	2.73	2.85	4.37	2.80	3.21	2.88	4.66	4.27	4.99	2.66
Written English Corpus (BAWE)	3.28		2.15	2.21	2.10	3.37	1.98	2.15	2.27	4.28	2.05	2.39	1.92	4.01	3.88	3.20	1.69
British National Corpus	2.77	2.15		1.59	1.32	3.61	1.51	1.69	1.58	2.70	1.45	1.64	1.63	3.89	2.56	3.79	1.72
Brown	3.12	2.21	1.59		1.35	3.56	1.58	1.79	1.72	2.74	1.58	1.90	1.87	3.92	2.58	3.77	1.89
Brown Family	2.83	2.10	1.32	1.35		3.52	1.47	1.70	1.56	2.62	1.42	1.67	1.70	3.89	2.39	3.75	1.76
e-flux	5.19	3.37	3.61	3.56	3.52		3.28	3.46	3.53	5.70	3.31	3.47	3.02	5.17	5.38	4.32	3.13
enTenTen	2.71	1.98	1.51	1.58	1.47	3.28		1.41	1.35	2.98	1.32	1.75	1.42	3.76	2.76	3.51	1.56
enTenTen2 (5G sample)	2.73	2.15	1.69	1.79	1.70	3.46	1.41		1.59	3.09	1.51	1.88	1.51	3.67	2.99	3.67	1.73
s + word family tagging	2.85	2.27	1.58	1.72	1.56	3.53	1.35	1.59		2.86	1.46	1.79	1.62	3.87	2.71	3.78	1.76
BeebOx	4.37	4.28	2.70	2.74	2.62	5.70	2.98	3.09	2.86		2.78	3.01	3.23	5.10	2.90	5.72	3.58
OEC	2.80	2.05	1.45	1.58	1.42	3.31	1.32	1.51	1.46	2.78		1.55	1.52	3.82	2.71	3.61	1.67
SiBol/Port	3.21	2.39	1.64	1.90	1.67	3.47	1.75	1.88	1.79	3.01	1.55		1.74	4.02	2.92	3.82	2.05
ukWaC	2.88	1.92	1.63	1.87	1.70	3.02	1.42	1.51	1.62	3.23	1.52	1.74		3.74	3.07	3.33	1.60
brewing	4.66	4.01	3.89	3.92	3.89	5.17	3.76	3.67	3.87	5.10	3.82	4.02	3.74		4.97	4.73	3.88
Dickens	4.27	3.88	2.56	2.58	2.39	5.38	2.76	2.99	2.71	2.90	2.71	2.92	3.07	4.97		5.53	3.26
volcano2_en	4.99	3.20	3.79	3.77	3.75	4.32	3.51	3.67	3.78	5.72	3.61	3.82	3.33	4.73	5.53		3.28

ge options

	enTenTen2 (5G sample)		enTenTen		
word	Freq	Freq/mill	Freq	Freq/mill	Score
n't	7145913.0	1320.3	0	0.0	14.2
loan	1442112.0	266.4	90516.0	27.7	2.9
online	2633924.0	486.6	377451.0	115.5	2.7
your	23521226.0	4345.7	5094910.0	1558.6	2.7
insurance	1508621.0	278.7	180503.0	55.2	2.4
credit	1868610.0	345.2	290987.0	89.0	2.4
loans	956579.0	176.7	69694.0	21.3	2.3
internet	1048243.0	193.7	130778.0	40.0	2.1
mortgage	758073.0	140.1	55578.0	17.0	2.1
marketing	1039826.0	192.1	147690.0	45.2	2.0
website	1504769.0	278.0	308979.0	94.5	1.9
business	3360052.0	620.8	907308.0	277.6	1.9
you	44831646.0	8283.0	14133031.0	4323.6	1.9
buy	1411770.0	260.8	298308.0	91.3	1.9
products	1563278.0	288.8	367225.0	112.3	1.8
skin	912979.0	168.7	153980.0	47.1	1.8
company	2525457.0	466.6	687907.0	210.4	1.8
weight	1032168.0	190.7	199215.0	60.9	1.8
wedding	581293.0	107.4	48899.0	15.0	1.8
home	3378302.0	624.2	991851.0	303.4	1.8
cash	838068.0	154.8	141777.0	43.4	1.8
help	3795317.0	701.2	1159162.0	354.6	1.8
debt	771292.0	142.5	125032.0	38.3	1.8
web	1444940.0	267.0	370450.0	113.3	1.7

What's missing

- Heterogeneity
- "how similar is BNC to WSJ"?
  - ***We need to know heterogeneity before we can interpret***
- The leading diagonal
- 2001 paper: randomising halves
  - Inelegant and inefficient
  - Depended on standard size of document

New definition, method (Pavel)

- Heterogeneity (def)
  - *Distance between most different partitions*
- Cluster to find `most different partitions'
- Bottom-up clustering
  - until largest cluster has over one third of data
  - Rest: the other partition
- Problem
  - nxn distance matrix where n > 1 million
  - Solution: do it in steps

Summary

- Corpus comparison
  - Qualitative: use keywords
  - Quantitative
    - On beta
    - Heterogeneity (to complete the task) to follow (soon)

Simple maths for keywords

	N	freq	Freq per m
Focus Corp	2m	80	40
Ref corp	15m	300	20
ratio			2

- Intuitive
- Nearly right but:
  - How well matched are corpora
    - Not here
  - Burstiness
    - Not here
  - Can't divide by zero
  - Commoner *vs.* rarer words

You can't divide by zero

	fc	rc	ratio
buggle	10	0	?
stort	100	0	?
nammikin	1000	0	?

- Standard solution: add one

	fc	rc	ratio
buggle	11	1	11
stort	101	1	101
nammikin	1001	1	1001

- **Problem solved**

High ratios more common for rarer words

	fc	rc	ratio	interesting?
spug	**10**	**1**	**10**	**no**
grod	**1000**	**100**	**10**	**yes**

- some researchers: grammar, grammar words
- some researchers: lexis content words

No right answer

Slider?

Solution: don't just add 1, add n

- n=1

word	fc	rc	fc+n	rc+n	Ratio	Rank
obscurish	10	0	11	1	11.00	1
middling	200	100	201	101	1.99	2
common	12000	10000	12001	10001	1.20	3

- n=100

word	fc	rc	fc+n	rc+n	Ratio	Rank
obscurish	10	0	110	100	1.10	3
middling	200	100	300	200	1.50	1
common	12000	10000	12100	10100	1.20	2

Solution

- n=1000

word	fc	rc	fc+n	rc+n	Ratio	Rank
obscurish	10	0	1010	1000	1.01	3
middling	200	100	1200	1100	1.09	2
common	12000	10000	13000	11000	1.18	1

- **Summary**

word	fc	rc	n=1	n=100	n=1000
obscurish	10	0	1st	2nd	3rd
middling	200	100	2nd	1st	2nd
common	12000	10000	3rd	3rd	1st