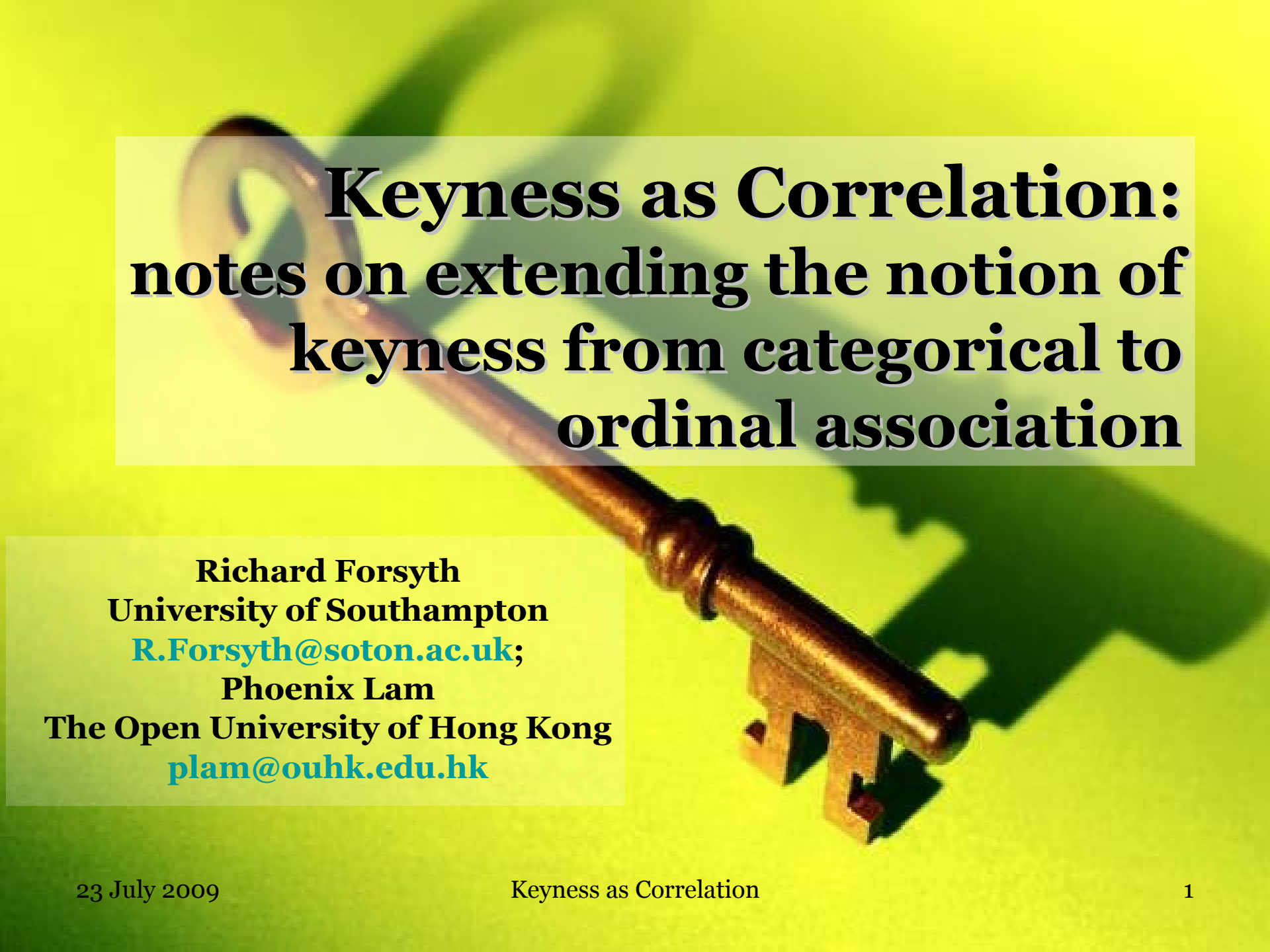


Keyness as Correlation: notes on extending the notion of keyness from categorical to ordinal association

Richard Forsyth
University of Southampton
R.Forsyth@soton.ac.uk;
Phoenix Lam
The Open University of Hong Kong
plam@ouhk.edu.hk

Outline

- Background
  - Current approaches to “keyness”
  - Motive for extension
- Research question
  - Are other indices more suitable than adapting G^2?
- Test corpora used
- Keyness indices tested
- How to evaluate a keyness index?
- Results of comparing keyness indices
- Discussion
  - Next steps
  - Feedback from this audience?

Current approaches to "keyness"

- Comparisons of Corpus A with Corpus B
- Or particular document versus "reference" corpus
- Examples:
  - American National Corpus versus British National Corpus
  - Merchant of Venice versus Jew of Malta
  - Miss Marple novel(s) versus Hercule Poirot novel(s)
  - Spam versus legitimate emails
  - Etc.
- More details, see:
  - Scott (1997), Rayson & Garside (2000), Kilgarriff (2001)

Example keyness list

KeyWords

File Edit View Compute Settings Windows Help

	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P	Lemmas	Set
1	GWENDA	134	0.97	0		178.98	0.0000000000		
2	MARPLE	57	0.41	0		75.98	0.0000000000		
3	GILES	50	0.36	0		66.63	0.0000000000		
4	SHE	242	1.76	96	0.74	57.96	0.0000000000		
5	MISS	70	0.51	7	0.05	56.49	0.0000000000		
6	HOUSE	76	0.55	15	0.12	41.46	0.0000000000		
7	HELEN	21	0.15	0		27.96	0.0000001206		
8	HENGRAVE	21	0.15	0		27.96	0.0000001206		
9	DILLMOUTH	20	0.15	0		26.63	0.0000002432		
10	ENGLAND	24	0.17	1		25.01	0.0000005685		
11	MAN	10	0.07	43	0.33	-24.08	0.0000009222		
12	MY	49	0.36	112	0.86	-29.16	0.0000000637		
13	ME	34	0.25	96	0.74	-34.55	0.0000000012		
14	I	231	1.68	483	3.71	-108.45	0.0000000000		

KWs plot links clusters filenames notes source text

14 Type-in

Current approaches (cont.)

- Essentially categorical association
  - Which terms associate with which corpus (A/B)?
  - Effectively each document has a label (A/B)
- Association measures derived from (relative) frequencies
- Most popular indices:
  - Chi-squared
  - Log-likelihood (G^2)
    - $G^2 = 2 * \Sigma_j(f_j * \ln(f_j/e_j))$
  - (both asymptotically equivalent, but Dunning (1993) showed G^2 to be more robust with highly unbalanced frequencies)

What if y-variable is numeric (not nominal)?

- When documents have scores, not category labels, e.g.:

Area	Text type	Y-variable
Education	Student writing	Assessed grade
Finance	News story	Share price rise/fall
Medicine	Patient transcript	Severity index
Politics	Campaign speech	Policy position rating

Research question

- G^2 can be adapted to relate term frequencies to ordinal or interval y-values
  - (just chop y-scale at median and treat as binary)
- But this ignores information, so…
- **Would other keyness indices have advantages with this sort of data?**

Test area chosen

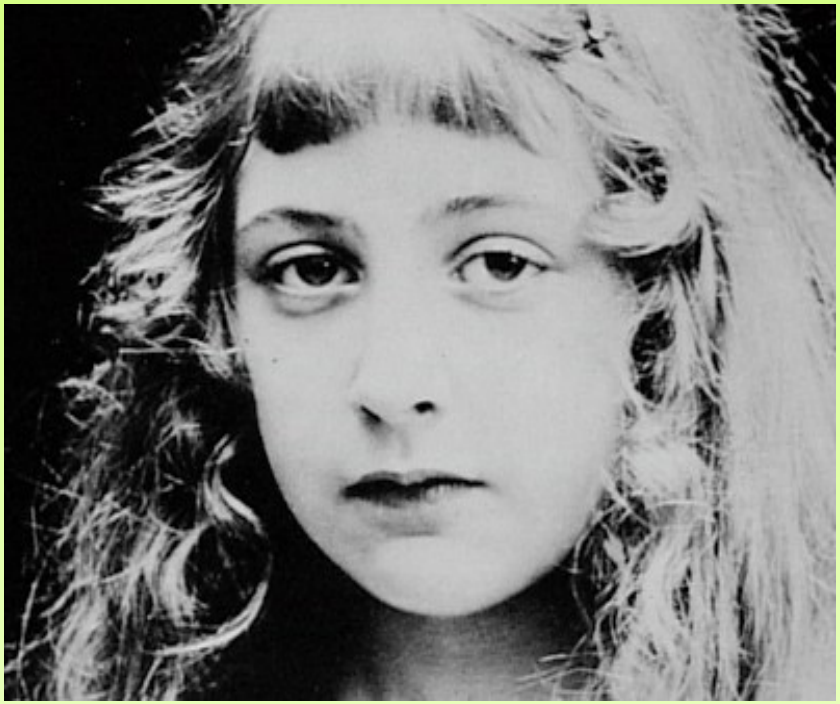
- Stylochronometry
  - Associating date (of composition) with a text
  - $\hat{y}$ = f(text)
  - y is year or age of author
  - In principle can be extended to other y variables
    - emotive score of emails
    - patient severity index

Basic idea: as the artist ages …

does his art “age” too?

Lily pond, 1926

From the lines on her face …



... to her lines on the page

- We acquiesced and followed him out of the room. John strode on ahead and I took the opportunity of whispering to Poirot:
- "There will be an inquest then?"
- Poirot nodded absently. He seemed absorbed in thought; so much so that my curiosity was aroused.
- "What is it? You are not attending to what I say."
- "It is true, my friend. I am much worried."
- "Why?"
- "Because Mademoiselle Cynthia does not take sugar in her coffee."
- "What? You cannot be serious?"
- "But I am most serious. Ah, there is something there that I do not understand. My instinct was right."
- "What instinct?"
- "The instinct that led me to insist on examining those coffee-cups. Chut! no more now!"
- We followed John into his study, and he closed the door behind us.

- 'No, indeed,' said Henry. 'No, indeed. I am wondering really - yes, our time's very short you know - whether we hadn't better - well, give up this tour at this point here. Not continue with it. It seems to me that there's bound to be a bit of difficulty resuming things until we know definitely. If this was - well - I mean, if this should be so serious that it could prove fatal, there might - well - I mean there might have to be an inquest or something of that kind.'
- 'Oh Henry, don't say dreadful things like that!'
- 'I'm sure,' said Miss Cooke, 'that you are being a little too pessimistic, Mr Butler. I am sure that things couldn't be as serious as that.'
- In his foreign voice Mr Caspar said: 'But yes, they are serious. I hear yesterday. When Mrs Sandbourne talk on telephone to doctor. It is very, very serious. They say she has concussion bad - very bad. A special doctor he is coming to look at her and see if he can operate or if impossible. Yes - it is all very bad.'
- 'Oh dear,' said Miss Lumley.

Test corpora used

Name	Content	Docs	Words	Meansize	Vocsize
AC	Chapters by Agatha Christie	52	141093	2713.33	10160
IM	Chapters* by Iris Murdoch	52	213492	4105.62	13190
WY	Poems by WB Yeats	89	19919	223.81	3834
Augs	US presidential inaugural speeches	39	102012	2615.69	8080
Xmas	Christmas broadcasts by QE 2	57	37797	663.11	4104

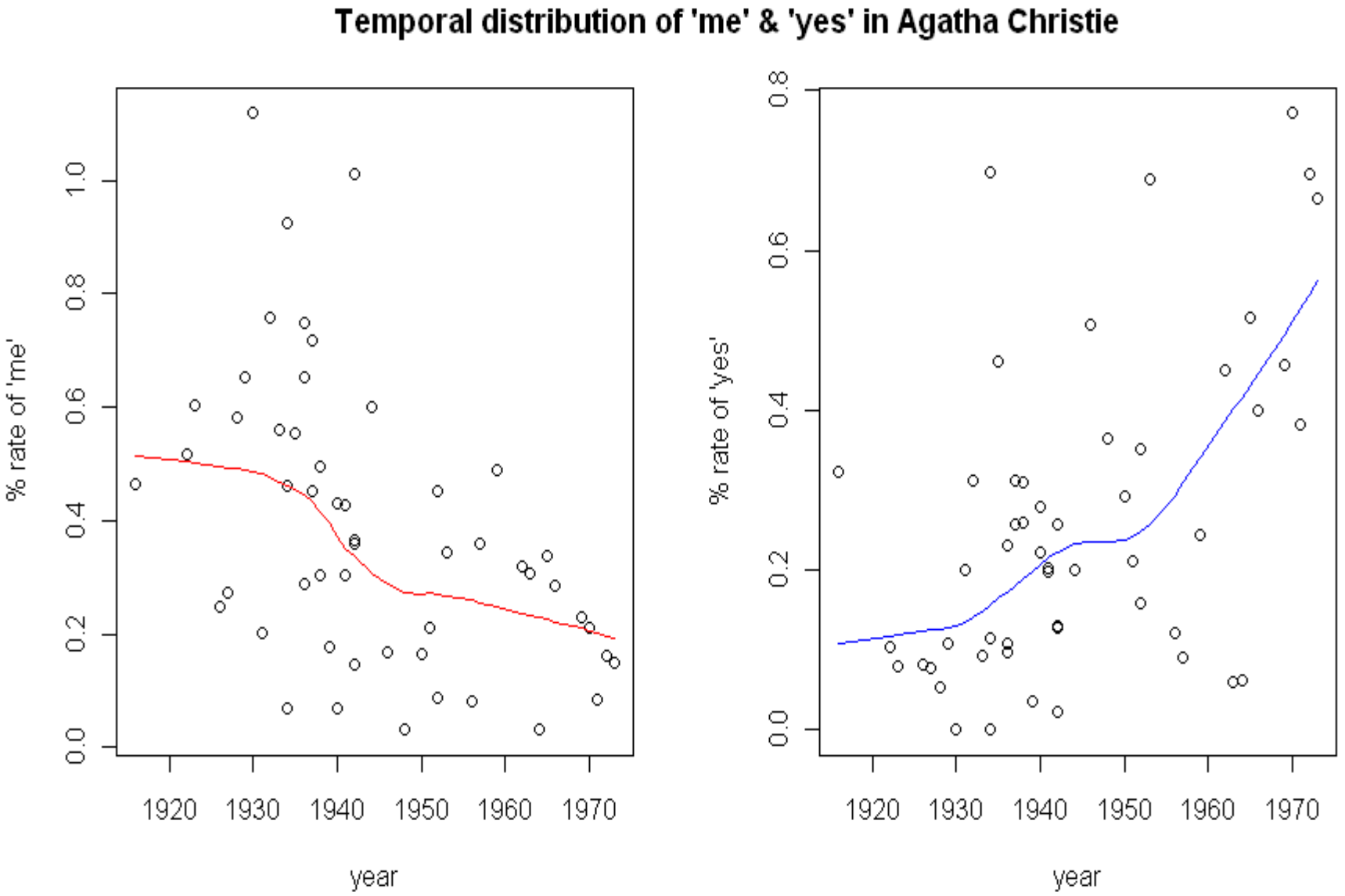
Term selection procedure

- Consider all terms with document frequency >= 2
- Rank them according to a keyness index
  - Some sort of x-y association:
    - x = term frequency
    - y = dependent variable value (e.g. year)
- Keep most extreme at either end
  - how many to keep?
    - [sqrt(vocsize) in present experiments]

Example term-list (keymode 0)

Term	Keyness	Rank
poirot	-104.03601	-1
:	-86.17161	-2
!	-69.80582	-3
charles	-67.33293	-4
lee	-65.50910	-5
monsieur	-57.30462	-6
mademoiselle	-56.00497	-7
julius	-46.58004	-8
m	-46.51233	-9
me	-41.94469	-10

Term	Keyness	Rank
blake	40.23065	10
canon	40.23065	9
dermot	42.91270	8
inspector	44.81696	7
victoria	45.67093	6
oliver	49.99694	5
craddock	83.14335	4
henry	89.84846	3
miss	107.03291	2
marple	196.14689	1


Temporal distribution of 'me' & 'yes' in Agatha Christie
% rate of 'me'
0.0
0.2
0.4
0.6
0.8
1.0
1920
1930
1940
1950
1960
1970
year
% rate of 'yes'
0.0
0.2
0.4
0.6
0.8
1920
1930
1940
1950
1960
1970
year

Indices tested

- Nine "association" indices tried
  - 0. Log-likelihood (G^2) with median as cutpoint
  - 1. Pearson's r, correlation coefficient
  - 2. "Riditized" correlation: corr(ridit(x),ridit(y))
    - Bross (1958)
  - 3. Goodman & Kruskal's Gamma
  - 4/5. Minor mods to K&R Gamma
    - (slow and very similar to #3 so not reported here)
  - 6. Frequency-adjusted z-score (FAZS)
  - 7. G^2 with square-roots of counts
  - 8. botched attempt at faster G&K Gamma!

Example term-list (keymode 6)

Term	Keyness	Rank	Term	Keyness	Rank
the	-0.32726	-1	*don't*	0.06510	10
i	-0.17372	-2	*or*	0.07592	9
:	-0.11599	-3	*think*	0.07701	8
!	-0.11582	-4	*know*	0.08755	7
poirot	-0.11480	-5	*very*	0.08763	6
me	-0.11458	-6	*marple*	0.08976	5
is	-0.10483	-7	*yes*	0.09782	4
my	-0.10448	-8	*miss*	0.12443	3
his	-0.06948	-9	*said*	0.13760	2
.	-0.06102	-10	*,*	0.20446	1

Example term-list (keymode 0)

Term	Keyness	Rank
poirot	-104.03601	-1
:	-86.17161	-2
!	-69.80582	-3
charles	-67.33293	-4
lee	-65.50910	-5
monsieur	-57.30462	-6
mademoiselle	-56.00497	-7
julius	-46.58004	-8
m	-46.51233	-9
me	-41.94469	-10

Term	Keyness	Rank
blake	40.23065	10
canon	40.23065	9
dermot	42.91270	8
inspector	44.81696	7
victoria	45.67093	6
oliver	49.99694	5
craddock	83.14335	4
henry	89.84846	3
miss	107.03291	2
marple	196.14689	1

Example term-list (keymode 6)

Term	Keyness	Rank
the	-0.32726	-1
i	-0.17372	-2
:	-0.11599	-3
!	-0.11582	-4
poirot	-0.11480	-5
me	-0.11458	-6
is	-0.10483	-7
my	-0.10448	-8
his	-0.06948	-9
.	-0.06102	-10

Term	Keyness	Rank
don't	0.06510	10
or	0.07592	9
think	0.07701	8
know	0.08755	7
very	0.08763	6
marple	0.08976	5
yes	0.09782	4
miss	0.12443	3
said	0.13760	2
,	0.20446	1

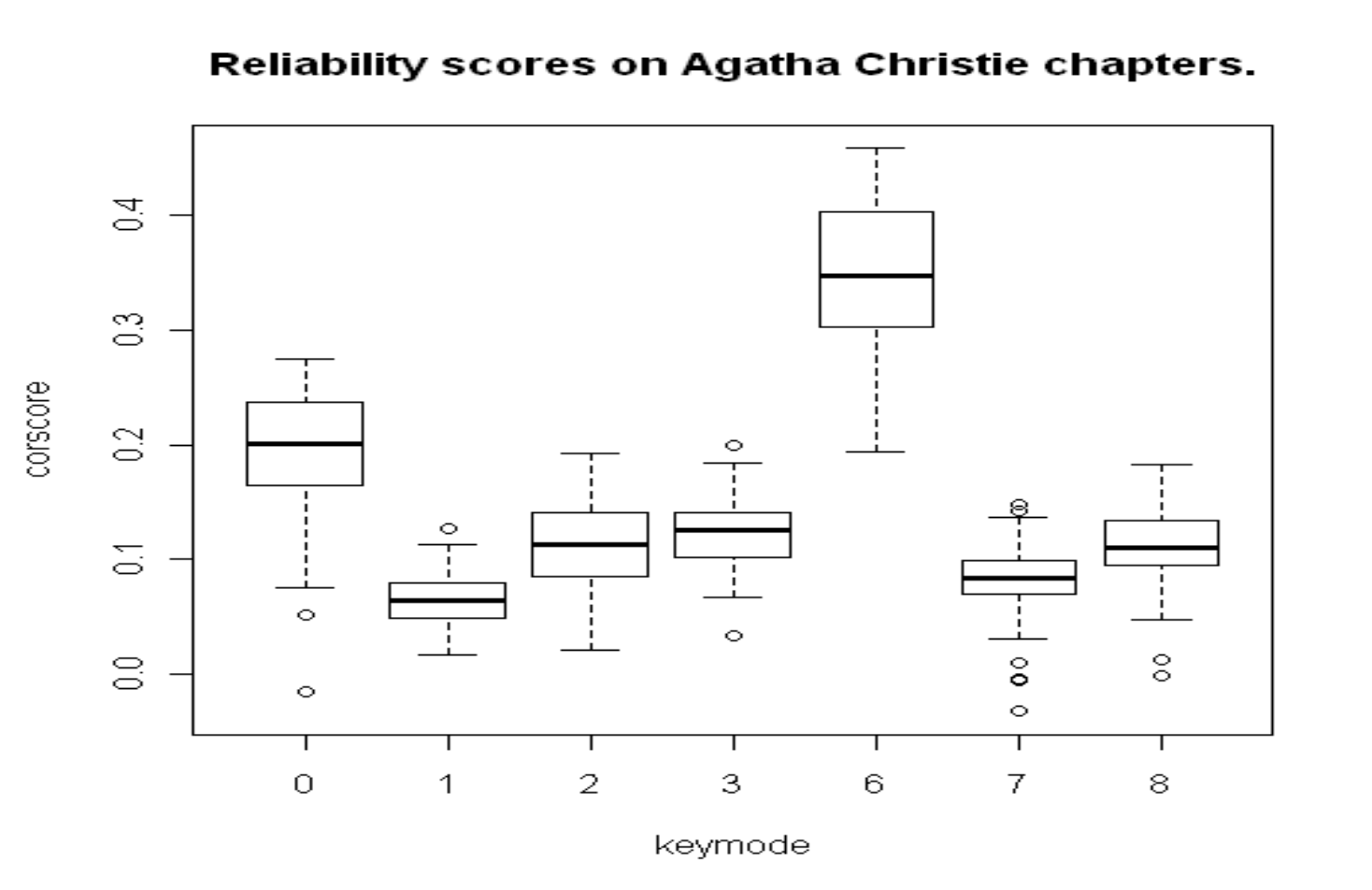
Desiderata for a keyness index

- Reliability
  - Consistent / stable across data
- Validity
  - Predictive accuracy on unseen data
    - (if used as features in a predictive model)
- Simplicity
  - Cheap / easy to compute
- Serendipity
  - Promotes human insight?
- N.B. Only reliability testing reported here

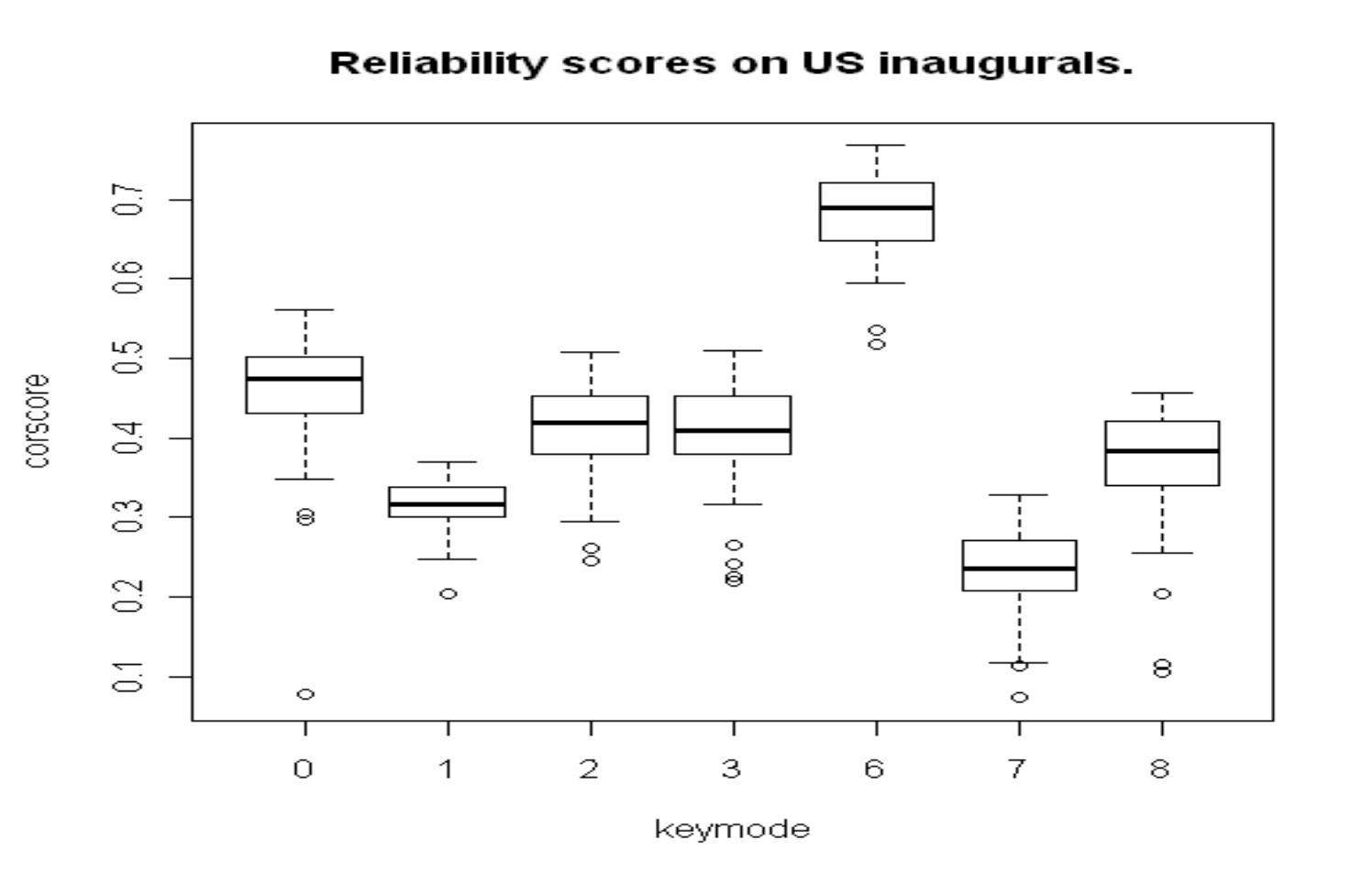
How to measure reliability?

- Split-half testing:
  - 5 corpora, 7 indices, 60 repetitions:
    - Random 50% subsamples
    - Term-list built and ranked on both halves (A/B)
    - Union of terms given ranking on both lists
      - (terms found in only 1 list given midrank)
    - Correlation of ranks
  - => “corscore”

Boxplot for Agatha Christie chapters



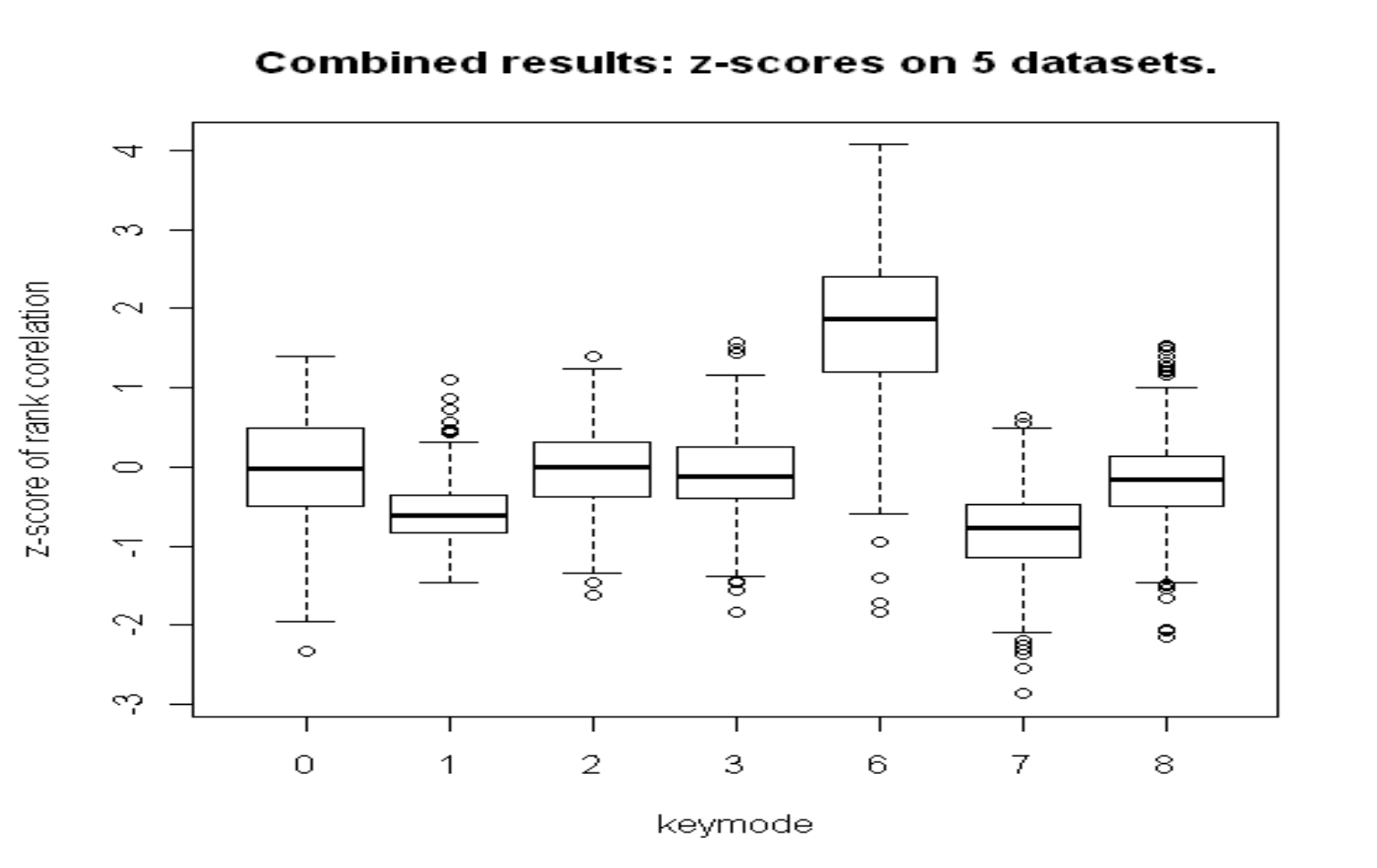
Boxplot for US inaugural speeches



Corscores aggregated using z-scores

- All five plots very similiar
- So corscores converted to z-scores within each dataset
  - "common currency"
- Then 5 datasets merged

Results for all 5 corpora combined



Conclusions

- Keymode 6 clearly “best”
  - (i.e. most stable)
- Mode 2 and mode 0 next
  - (worth considering)
- Modes 1 and 7 worst
  - (to be avoided)
- Mode 2 always better than mode 1
- Mode 0 always better than mode 7
- Mode 3 slows dramatically with large data sets

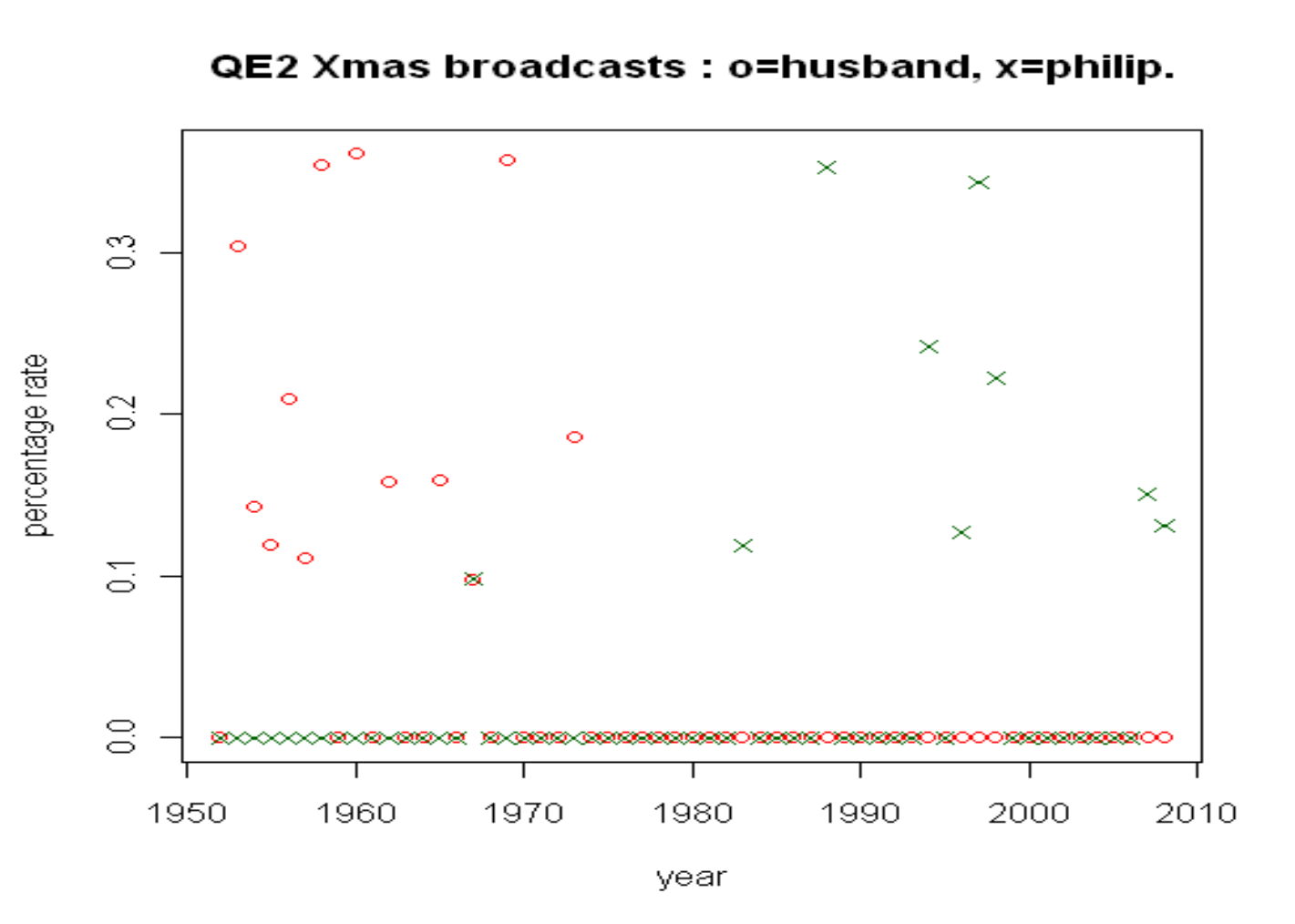
Formula for FAZS (keymode 6)

- Frequency-adjusted z-score:
- $FAZS = 100 * Z / N_d$
- where:
- $Z = \Sigma(w_t * w_y)$
- w_t = term-rate
  - relative frequency of term in document
- w_y = z-score of document y-value
  - standard score on y dimension
- N_d = number of documents

Discussion

- Still to do
  - Other problem domains
  - Predictive testing
  - More work on size of term-list
  - Tags as terms
  - Term combinations
    - See, for example, Cheng, Greaves and Warren 2006
  - Meaningful groupings of terms
    - To assist interpretation

"My Philip and I"



Concordance of "husband"

Concord

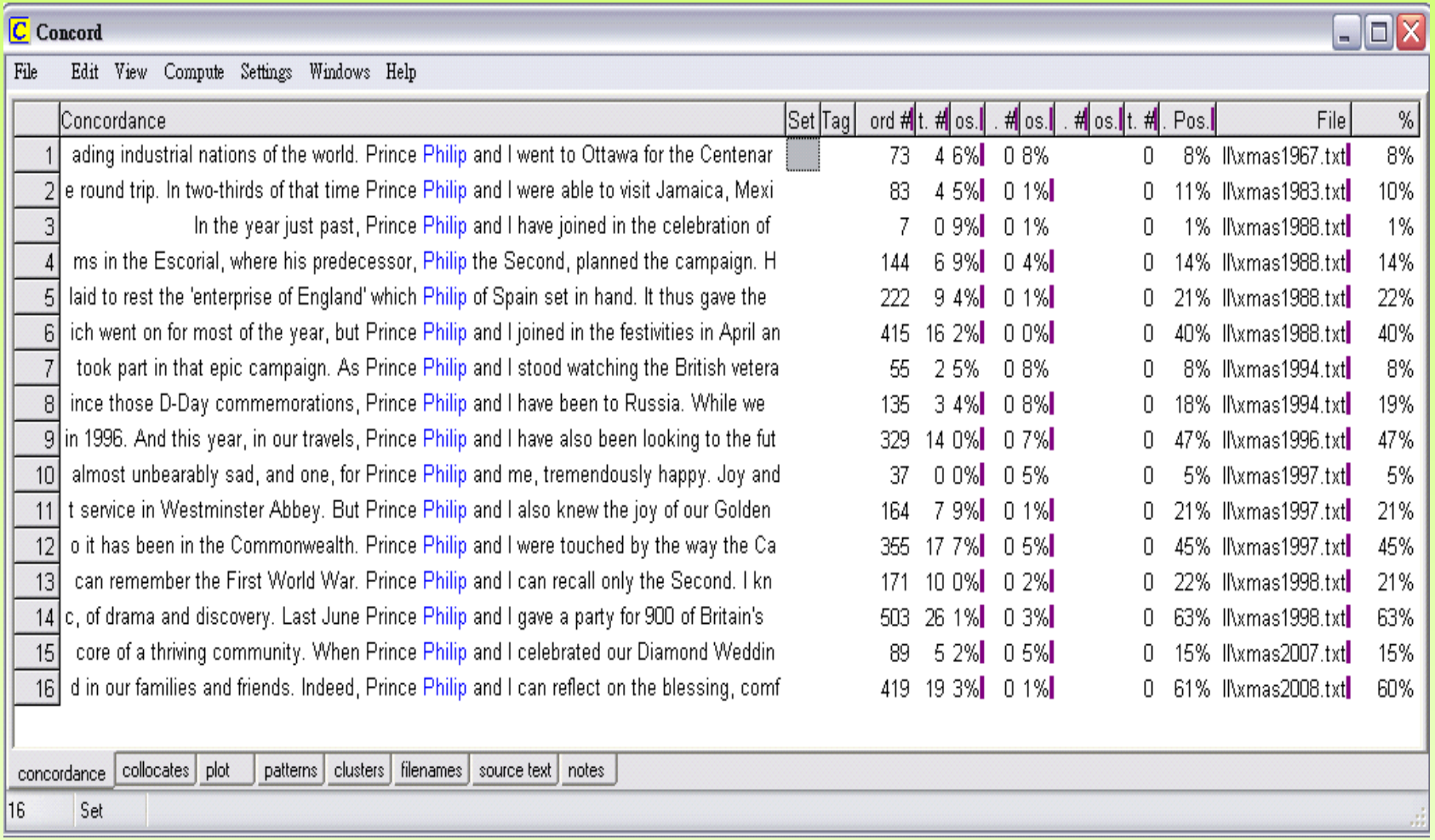
File Edit View Compute Settings Windows Help

	Concordance	tg	ord #	. #	. Pos.	#	Pos.	#s.	#	Pos.	File	%
1	It is now two years since my husband and I spent Christmas with our children.		8	0	53%	0	1%	0	1%		ll\xmas1954.txt	1%
2	wealth and of sending you my good wishes. My husband and our children, together with the other		29	1	10%	0	6%	0	6%		ll\xmas1960.txt	5%
3	that come to me from all parts of the world. My husband and children join me in thanking all of y		30	2	12%	0	4%	0	4%		ll\xmas1958.txt	4%
4	ow, our daughter joins us for Christmas with her husband and we are celebrating the festival this		35	2	37%	0	7%	0	7%		ll\xmas1973.txt	7%
5	think of each other, on Christmas Day. For my husband and myself and for our children, the yea		53	2	17%	0	7%	0	7%		ll\xmas1955.txt	6%
6	ange. It may be the first Christmas for many as husband and wife, or the first Christmas with gra		54	3	58%	0	9%	0	9%		ll\xmas1965.txt	10%
7	embers and we wish them all good fortune. My husband and I are greatly looking forward to re-vi		87	6	16%	0	15%	0	15%		ll\xmas1962.txt	16%
8	ime comes they, too, will be great travellers. My husband and I left London a month ago, but we h		124	5	11%	0	14%	0	14%		ll\xmas1953.txt	13%
9	s a great satisfaction and comfort to me and my husband to know that they have won a place in y		140	4	54%	0	53%	0	53%		ll\xmas1969.txt	51%
10	children and myself greater joy than that of my husband. To him I say: "From all the members of		146	4	100%	0	17%	0	17%		ll\xmas1956.txt	18%
11	icult to understand. It is because of this that my husband and I are so greatly looking forward to o		161	6	25%	0	32%	0	32%		ll\xmas1960.txt	31%
12	of you than ever before. In three weeks' time my husband goes to India and Pakistan and then on		231	11	39%	0	30%	0	30%		ll\xmas1958.txt	28%
13	went to the islands, they were a colony and my husband was serving with the Mediterranean Fle		307	15	71%	0	32%	0	32%		ll\xmas1967.txt	32%
14	brations of the state of Queensland. In June, my husband and I will be going to Canada once agai		316	15	36%	0	41%	0	41%		ll\xmas1958.txt	40%
15	reat Tudor forbear, who was blessed with neither husband nor children, who ruled as a despot and		346	12	56%	0	38%	0	38%		ll\xmas1953.txt	36%
16	and their children there is room at the Inn. If my husband cannot be at home on Christmas Day, I		383	12	13%	0	44%	0	44%		ll\xmas1956.txt	43%
17	m of my Canadian people. Also during 1957 my husband and I paid visits to Portugal, France, De		496	26	30%	0	60%	0	60%		ll\xmas1957.txt	61%
18	ut the Commonwealth, who will not join with my husband and me in sending to those who mourn		870	33	59%	0	97%	0	97%		ll\xmas1953.txt	97%

concordance | collocates | plot | patterns | clusters | filenames | source text | notes

18 Set

Concordance of "Philip"


Concord

File Edit View Compute Settings Windows Help

	Concordance	Set	Tag	ord #	t. #	os.	. #	os.	. #	os.	t. #	. Pos.	File	%
1	ading industrial nations of the world. Prince Philip and I went to Ottawa for the Centenar			73	4	6%	0	8%			0	8%	ll\xmas1967.txt	8%
2	e round trip. In two-thirds of that time Prince Philip and I were able to visit Jamaica, Mexi			83	4	5%	0	1%			0	11%	ll\xmas1983.txt	10%
3	In the year just past, Prince Philip and I have joined in the celebration of			7	0	9%	0	1%			0	1%	ll\xmas1988.txt	1%
4	ms in the Escorial, where his predecessor, Philip the Second, planned the campaign. H			144	6	9%	0	4%			0	14%	ll\xmas1988.txt	14%
5	laid to rest the 'enterprise of England' which Philip of Spain set in hand. It thus gave the			222	9	4%	0	1%			0	21%	ll\xmas1988.txt	22%
6	ich went on for most of the year, but Prince Philip and I joined in the festivities in April an			415	16	2%	0	0%			0	40%	ll\xmas1988.txt	40%
7	took part in that epic campaign. As Prince Philip and I stood watching the British vetera			55	2	5%	0	8%			0	8%	ll\xmas1994.txt	8%
8	ince those D-Day commemorations, Prince Philip and I have been to Russia. While we			135	3	4%	0	8%			0	18%	ll\xmas1994.txt	19%
9	in 1996. And this year, in our travels, Prince Philip and I have also been looking to the fut			329	14	0%	0	7%			0	47%	ll\xmas1996.txt	47%
10	almost unbearably sad, and one, for Prince Philip and me, tremendously happy. Joy and			37	0	0%	0	5%			0	5%	ll\xmas1997.txt	5%
11	t service in Westminster Abbey. But Prince Philip and I also knew the joy of our Golden			164	7	9%	0	1%			0	21%	ll\xmas1997.txt	21%
12	o it has been in the Commonwealth. Prince Philip and I were touched by the way the Ca			355	17	7%	0	5%			0	45%	ll\xmas1997.txt	45%
13	can remember the First World War. Prince Philip and I can recall only the Second. I kn			171	10	0%	0	2%			0	22%	ll\xmas1998.txt	21%
14	c, of drama and discovery. Last June Prince Philip and I gave a party for 900 of Britain's			503	26	1%	0	3%			0	63%	ll\xmas1998.txt	63%
15	core of a thriving community. When Prince Philip and I celebrated our Diamond Weddin			89	5	2%	0	5%			0	15%	ll\xmas2007.txt	15%
16	d in our families and friends. Indeed, Prince Philip and I can reflect on the blessing, comf			419	19	3%	0	1%			0	61%	ll\xmas2008.txt	60%

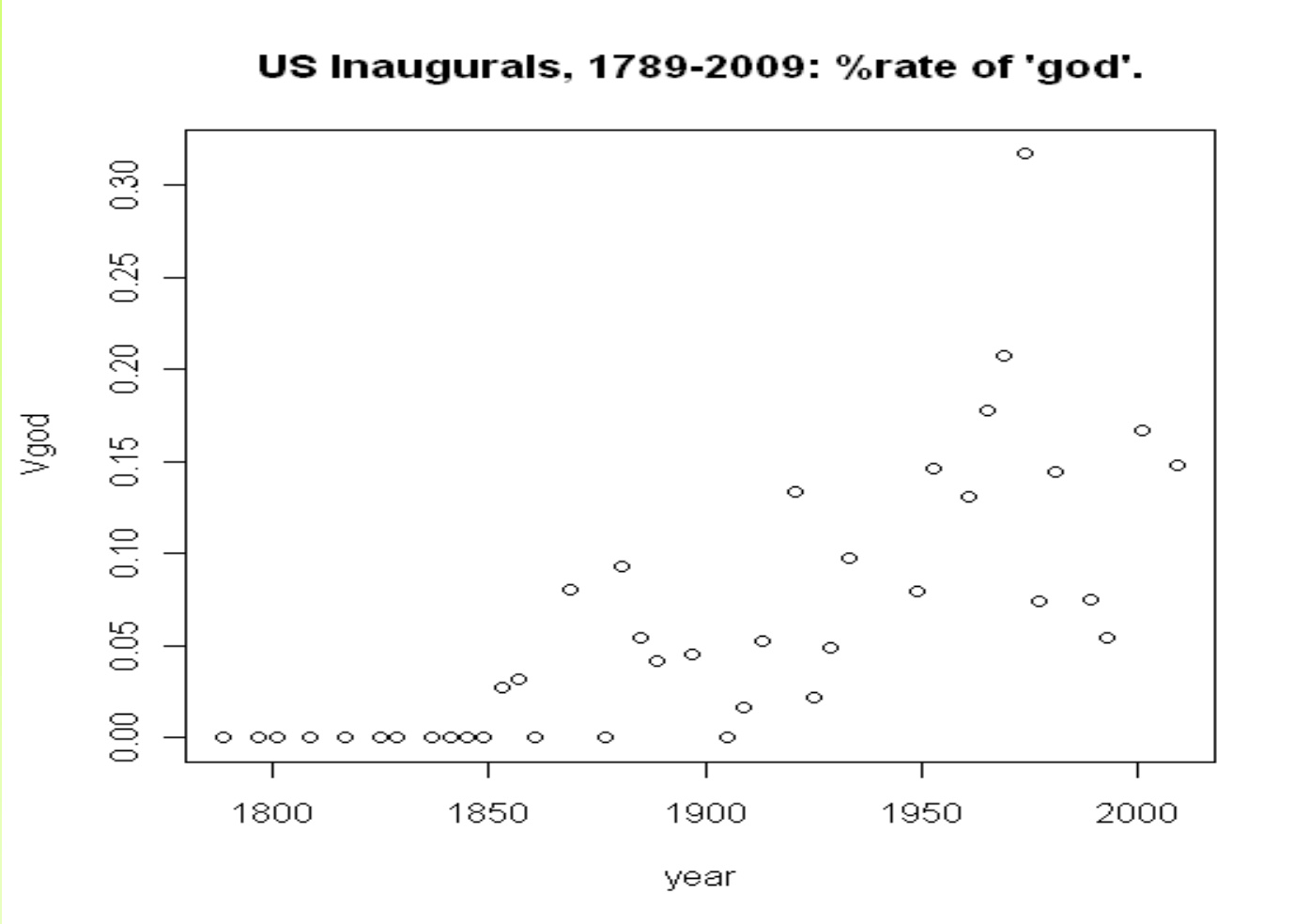
concordance | collocates | plot | patterns | clusters | filenames | source text | notes

16 Set

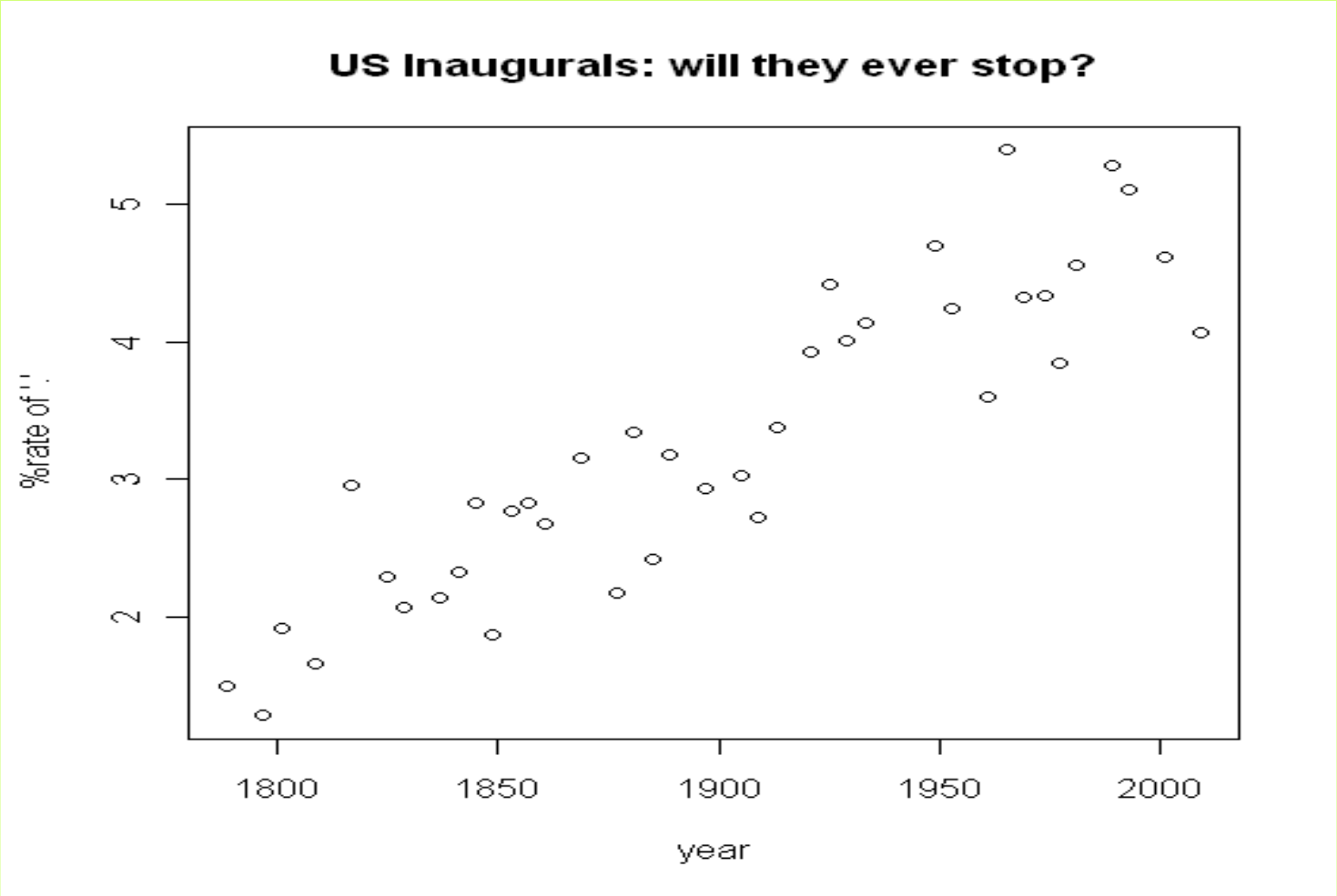
References

- **Bross, I. D. J.** (1958). How to use ridit analysis. *Biometrics,* 14:18-38.
- **Cheng, W., Greaves, C. & Warren, M.** (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics,* 11(4): 411-433.
- **Dunning, T.** (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics,* 19(1): 61-74.
- **Kilgarriff, A.** (2001). Comparing corpora. *International Journal of Corpus Linguistics,* 6(1): 1-37.
- **Rayson, P. & Garside, R.** (2000). Comparing corpora using frequency profiling.
  - http://www.comp.lancs.ac.uk/computing/users/publications/rg_acl2000.pdf
- **Scott, M.** (1997). PC analysis of key words – and key key words. *System,* 25(2): 233-245.
- http://ucrel.lancs.ac.uk/llwizard.html

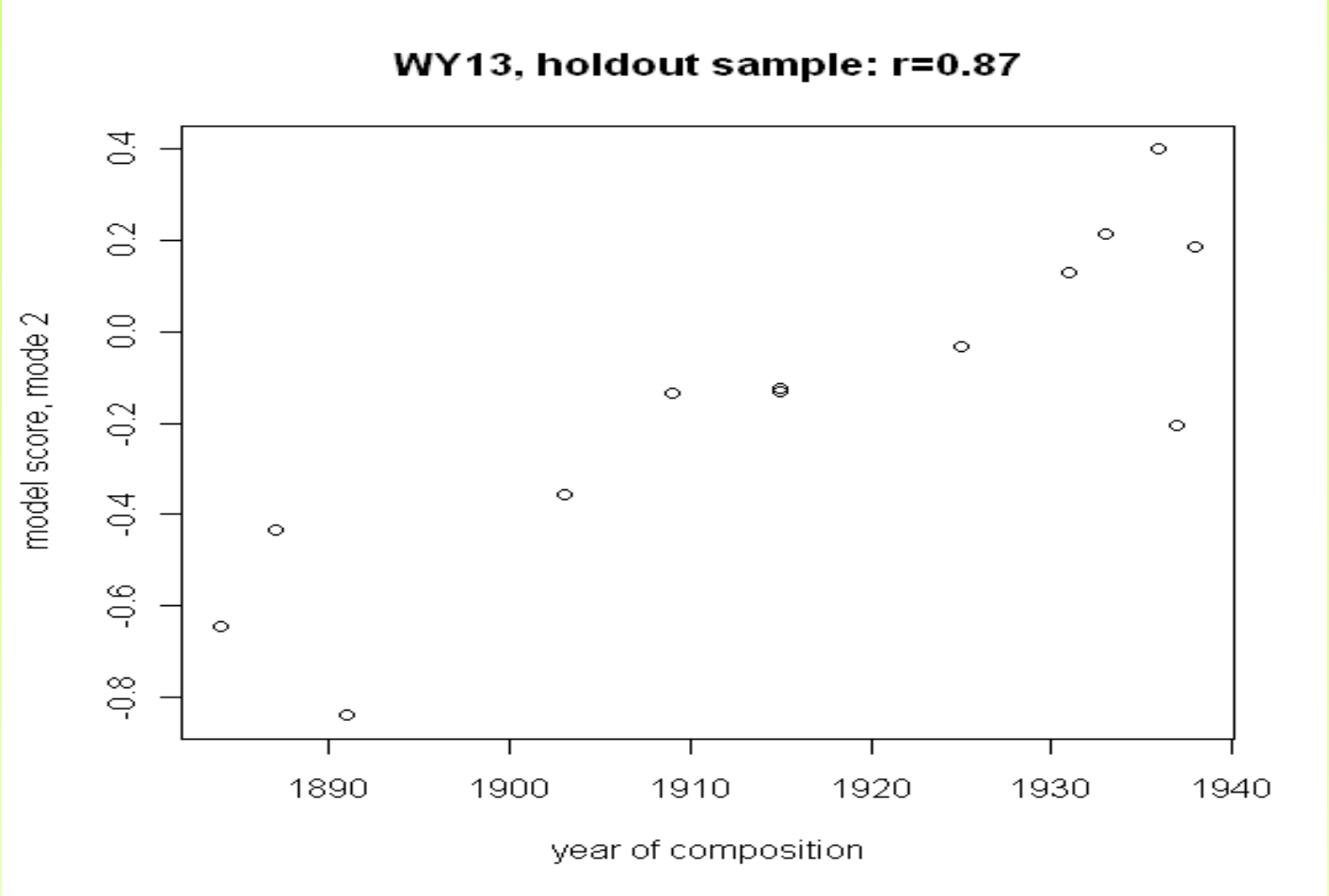
In "god" we trust

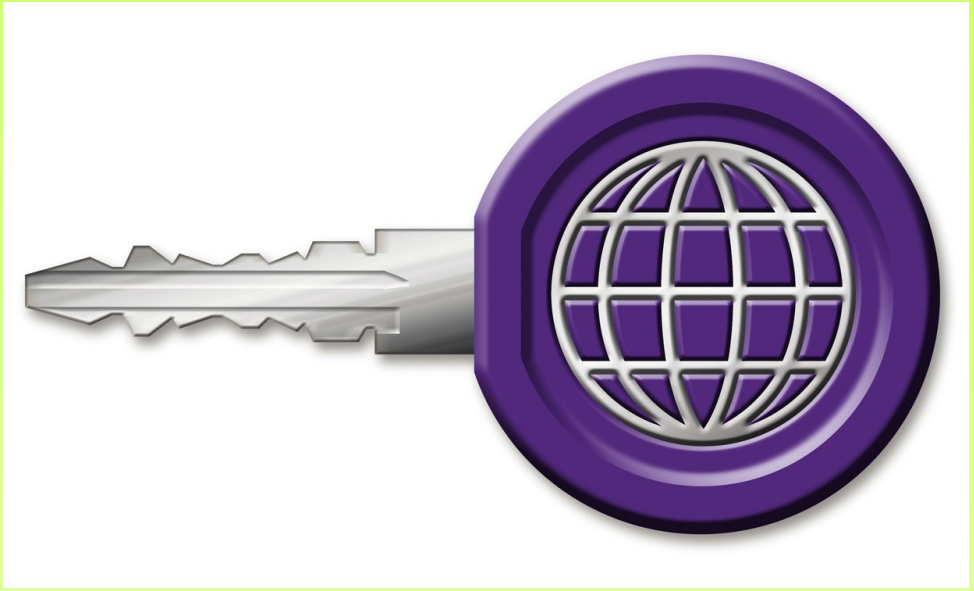


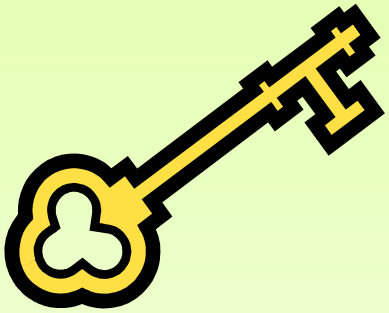
The periods tell you the period ;-)

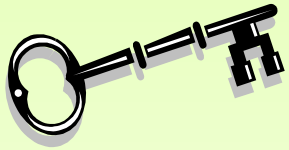


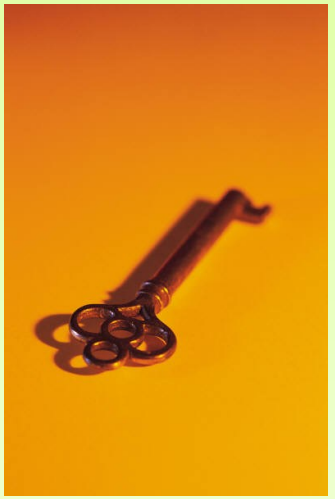
A poet young & old











What if y-variable is numeric (not nominal)?

- When documents have scores, not category labels, e.g.:

Area	**Text type**	**Y-variable**
Education	Student writing	Assessed grade
Finance	News story	Share price rise/fall
Medicine	Patient transcript	Severity index
Politics	Campaign speech	Policy position rating

Test corpora used

Name	Content	Docs	Words	Mean size	Vocsize
AC	Chapters by Agatha Christie	52	141093	2713.33	10160
IM	Chapters* by Iris Murdoch	52	213492	4105.62	13190
WY	Poems by WB Yeats	89	19919	223.81	3834
Augs	US presidential inaugural speeches	39	102012	2615.69	8080
Xmas	Christmas broadcasts by QE2	57	37797	663.11	4104

Example term-list (keymode 6)

the	-0.32726	-1	don't	0.06510	10
i	-0.17372	-2	or	0.07592	9
:	-0.11599	-3	think	0.07701	8
!	-0.11582	-4	know	0.08755	7
poirot	-0.11480	-5	very	0.08763	6
me	-0.11458	-6	marple	0.08976	5
is	-0.10483	-7	yes	0.09782	4
my	-0.10448	-8	miss	0.12443	3
his	-0.06948	-9	said	0.13760	2
.	-0.06102	-10	,	0.20446	1

Example term-list (keymode 0)

poirot	-104.03601	-1	blake	40.23065	10
:	-86.17161	-2	canon	40.23065	9
!	-69.80582	-3	dermot	42.91270	8
charles	-67.33293	-4	inspector	44.81696	7
lee	-65.50910	-5	victoria	45.67093	6
monsieur	-57.30462	-6	oliver	49.99694	5
mademoiselle	-56.00497	-7	craddock	83.14335	4
julius	-46.58004	-8	henry	89.84846	3
m	-46.51233	-9	miss	107.03291	2
me	-41.94469	-10	marple	196.14689	1

Example term-list (keymode 6)

the	-0.32726	-1	about	0.04707	20
i	-0.17372	-2	-	0.04745	19
:	-0.11599	-3	mean	0.04795	18
!	-0.11582	-4	really	0.04803	17
poirot	-0.11480	-5	who	0.04860	16
me	-0.11458	-6	a	0.05119	15
is	-0.10483	-7	they	0.05285	14
my	-0.10448	-8	pikeaway	0.05513	13
his	-0.06948	-9	oliver	0.05841	12
.	-0.06102	-10	people	0.06132	11
not	-0.05752	-11	don't	0.06510	10
m	-0.05743	-12	or	0.07592	9
mr	-0.05604	-13	think	0.07701	8
then	-0.05076	-14	know	0.08755	7
was	-0.04793	-15	very	0.08763	6
...	-0.04738	-16	marple	0.08976	5
will	-0.04413	-17	yes	0.09782	4
at	-0.04025	-18	miss	0.12443	3
as	-0.03483	-19	said	0.13760	2
face	-0.03432	-20	,	0.20446	1