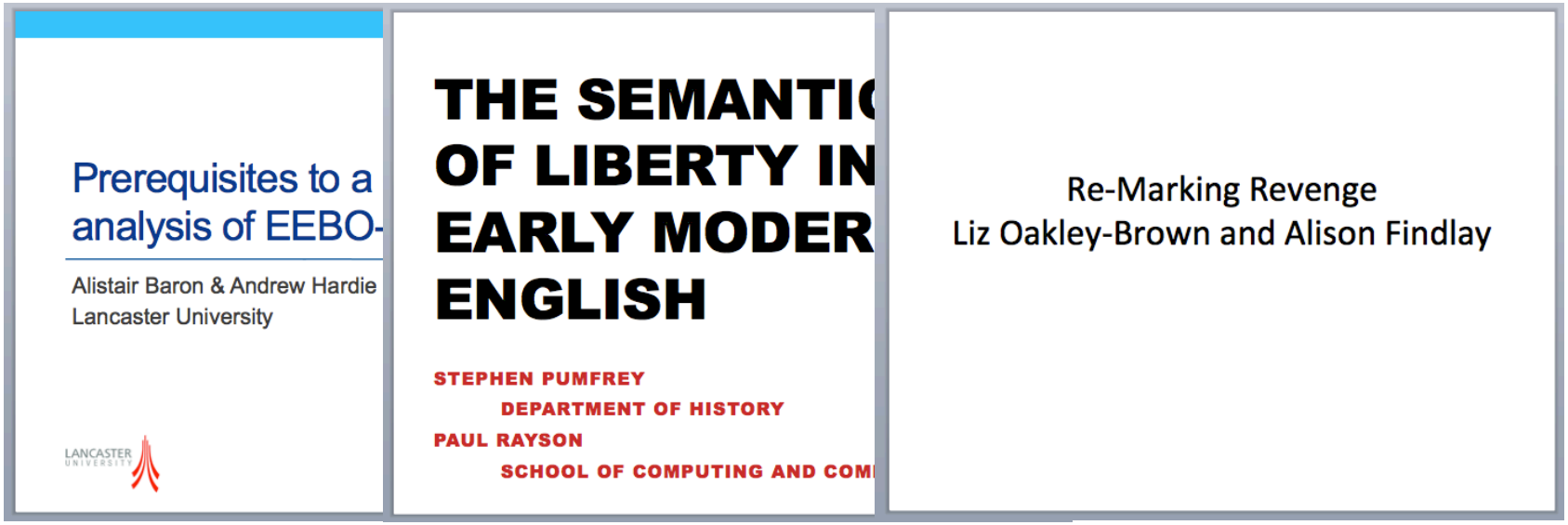
Transforming EEBO-TCP into a corpus

Paul Rayson (@perayson) & Alistair Baron (@al586)
School of Computing and Communications
Andrew Hardie (@HardieResearch)
Department of Linguistics and English Language
Lancaster University

CREME @ Lancaster
creme.lancs.ac.uk



"Experiments in 17th century English: manual versus automatic conceptual history." Stephen Pumfrey; Paul Rayson; John Mariani, Literary and Linguistic Computing 2012; doi: 10.1093/llc/fqs017

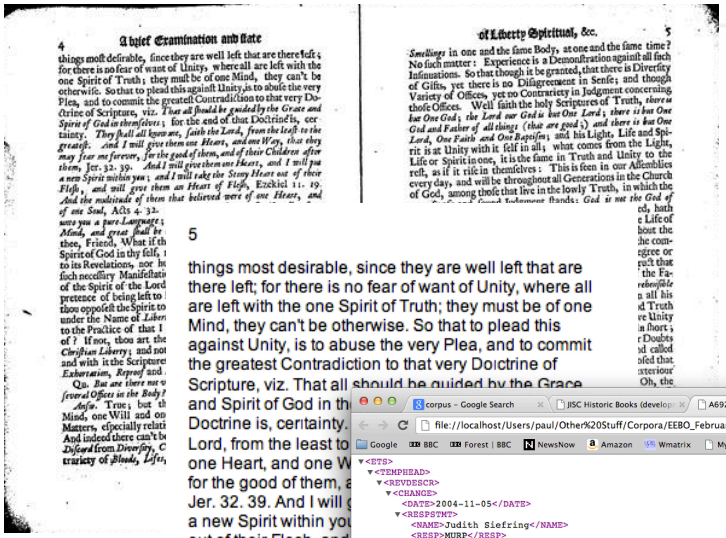
Big Data + Old History = Distant Reading

LANCASTER UNIVERSITY

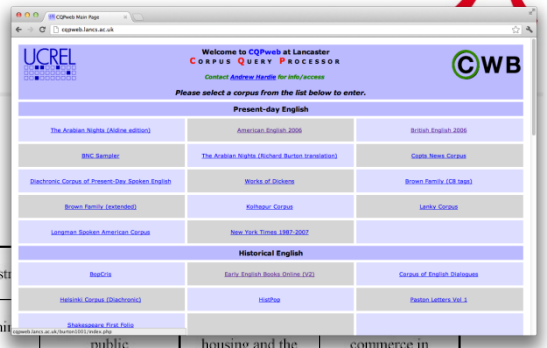
"Historians are now faced with more material than they could ever hope to read in a lifetime, or even 100 lifetimes."

With thanks to Adam Crymble, KCL, UK
http://youtu.be/tp4y-_VoXdA

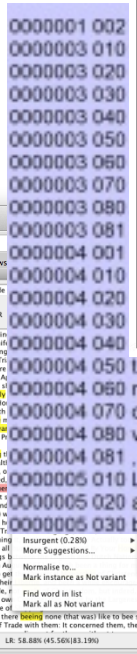
Corpus framework



things most desirable, since they are well left that are there left; for there is no fear of want of Unity, where all are left with the one Spirit of Truth; they must be of one Mind, they can't be otherwise. So that to plead this against Unity, is to abuse the very Plea, and to commit the greatest Contradiction to that very Doctrine of Scripture, viz. That all should be guided by the Grace and Spirit of God in th… Doctrine is, ceritainty. … Lord, from the least to … one Heart, and one W… for the good of them, a… Jer. 32. 39. And I will … a new Spirit within you … out of their Flesh, and … Ezekiel 11. 19. And th… were of one Heart, an… Unity too? I will restor… shall be of one Heart … be their Peace. There… What if thou wilt not be… God in thy self, nor wa… Revelations, nor humi… such necessary Manif… Counsel of the Spirit … under the pretence of … which means thou op… pleadest for Dis-unity, … thee, May not I exhort … moved to press thee t…



A general and abstract terms			
F food and farming		public	housing and the home	commerce in industry
K entertainment, sports and games	L life and living things	M movement, location, travel and transport	N numbers and measurement
O substances, materials, objects and equipment	P education	Q language and communication	S social actions, states and processes
T Time	W world and environment	X psychological actions, states and processes	Y science and technology
Z names and grammar			



Part of speech tagging

- CLAWS part-of-speech tagger
  - Hybrid rule-based and statistical methods
  - Trained and tested on 100 million-word British National Corpus
  - 97-98% accuracy across a variety of text types
  - http://ucrel.lancs.ac.uk/claws/

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)
AT	article (e.g. the, no)
AT1	singular article (e.g. a, an, every)
BCL	before-clause marker (e.g. in order (that),in order (to))
CC	coordinating conjunction (e.g. and, or)
CCB	adversative coordinating conjunction (but)
CS	subordinating conjunction (e.g. if, because, unless, so, for)
CSA	as (as conjunction)
CSN	than (as conjunction)
CST	that (as conjunction)
CSW	whether (as conjunction)
DA	after-determiner or post-determiner capable of pronominal
DA1	singular after-determiner (e.g. little, much)
DA2	plural after-determiner (e.g. few, several, many)
DAR	comparative after-determiner (e.g. more, less, fewer)
DAT	superlative after-determiner (e.g. most, least, fewest)
DB	before determiner or pre-determiner capable of pronominal
DB2	plural before-determiner (both)
DD	determiner (capable of pronominal function) (e.g any, some
DD1	singular determiner (e.g. this, that, another)
DD2	plural determiner (these,those)
DDQ	wh-determiner (which, what)
DDQGE	wh-determiner, genitive (whose)
DDQV	wh-ever determiner, (whichever, whatever)
EX	existential there
FO	formula
FU	unclassified word
FW	foreign word
GE	germanic genitive marker - (' or's)
IF	for (as preposition)
II	general preposition
IO	of (as preposition)
IW	with, without (as prepositions)
JJ	general adjective
JJR	general comparative adjective (e.g. older, better, stronger)
JJT	general superlative adjective (e.g. oldest, best, strongest)
JK	catenative adjective (able in be able to, willing in be willing
MC	cardinal number,neutral for number (two, three..)
MC1	singular cardinal number (one)
MC2	plural cardinal number (e.g. sixes, sevens)
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
MD	ordinal number (e.g. first, second, next, last)
MF	fraction,neutral for number (e.g. quarters, two-thirds)
ND1	singular noun of direction (e.g. north, southeast)

Semantic Tagging

- USAS (UCREL Semantic Analysis System)
  - Rule-based and knowledge-based system
  - Tagging coarse-grained sense in context
  - Trained and tested on wide variety of corpus types and domains
  - 91% accurate on 'general' language
  - Tagset based on Tom McArthur's Longman Lexicon
  - http://ucrel.lancs.ac.uk/usas/

I MONEY & COMMERCE
I1 Money generally
I1.1 Money: Affluence
I1.2 Money: Debts
I1.3 Money: Price
I2 Business
I2.1 Business: Generally
I2.2 Business: Selling
I3 Work and employment
I3.1 Work and employment: Generally
I3.2 Work and employment: Professionalism
I4 Industry
K ENTERTAINMENT, SPORTS & GAMES
K1 Entertainment generally
K2 Music and related activities
K3 Recorded sound etc.
K4 Drama, the theatre & show business
K5 Sports and games generally
K5.1 Sports
K5.2 Games
K6 Children's games and toys
L LIFE & LIVING THINGS
L1 Life and living things
L2 Living creatures generally
L3 Plants
M MOVEMENT, LOCATION, TRAVEL & TRANSPORT
M1 Moving, coming and going
M2 Putting, taking, pulling, pushing, transporting &c.
M3 Movement/transportation: land
M4 Movement/transportation: water
M5 Movement/transportation: air
M6 Location and direction
M7 Places
M8 Remaining/stationary
N NUMBERS & MEASUREMENT
N1 Numbers
N2 Mathematics
N3 Measurement
N3.1 Measurement: General
N3.2 Measurement: Size
N3.3 Measurement: Distance
N3.4 Measurement: Volume
N3.5 Measurement: Weight
N3.6 Measurement: Area
N3.7 Measurement: Length & height
N3.8 Measurement: Speed
N4 Linear order
N5 Quantities
N5.1 Entirety; maximum
N5.2 Exceeding; waste
N6 Frequency etc.
O SUBSTANCES, MATERIALS, OBJECTS & EQUIPMENT
O1 Substances and materials generally
O1.1 Substances and materials generally: Solid
O1.2 Substances and materials generally: Liquid
O1.3 Substances and materials generally: Gas
O2 Objects generally
O3 Electricity and electrical equipment
O4 Physical attributes
O4.1 General appearance and physical properties
O4.2 Judgement of appearance (pretty etc.)
O4.3 Colour and colour patterns
O4.4 Shape
O4.5 Texture
O4.6 Temperature
P EDUCATION
P1 Education in general
Q LINGUISTIC ACTIONS, STATES & PROCESSES
Q1 Communication
Q1.1 Communication in general
Q1.2 Paper documents and writing

A General and abstract terms	**B** The body and the individual	**C** Arts and crafts	**E** Emotion
F Food and farming	**G** Government and public	**H** Architecture, housing and the home	**I** Money and commerce in industry
K Entertainment, sports and games	**L** Life and living things	**M** Movement, location, travel and transport	**N** Numbers and measurement
O Substances, materials, objects and equipment	**P** Education	**Q** Language and communication	**S** Social actions, states and processes
T Time	**W** World and environment	**X** Psychological actions, states and processes	**Y** Science and technology
Z Names and grammar			

Problems for automatic methods …

- Spelling variation (also includes “change over time”, e.g. from “lybertye” to “liberty”)
- Grammatical change over time
- Meaning change over time (possibly marked by variants)

- Impact on corpus linguistics and computational methods
  - Simple searching for words and frequency lists.
  - Key words (Baron et al., 2009) and clusters (Palander-Collin & Hakala, 2011)
  - POS tagging (Rayson et al., 2007)
  - Semantic tagging (Archer et al., 2003).

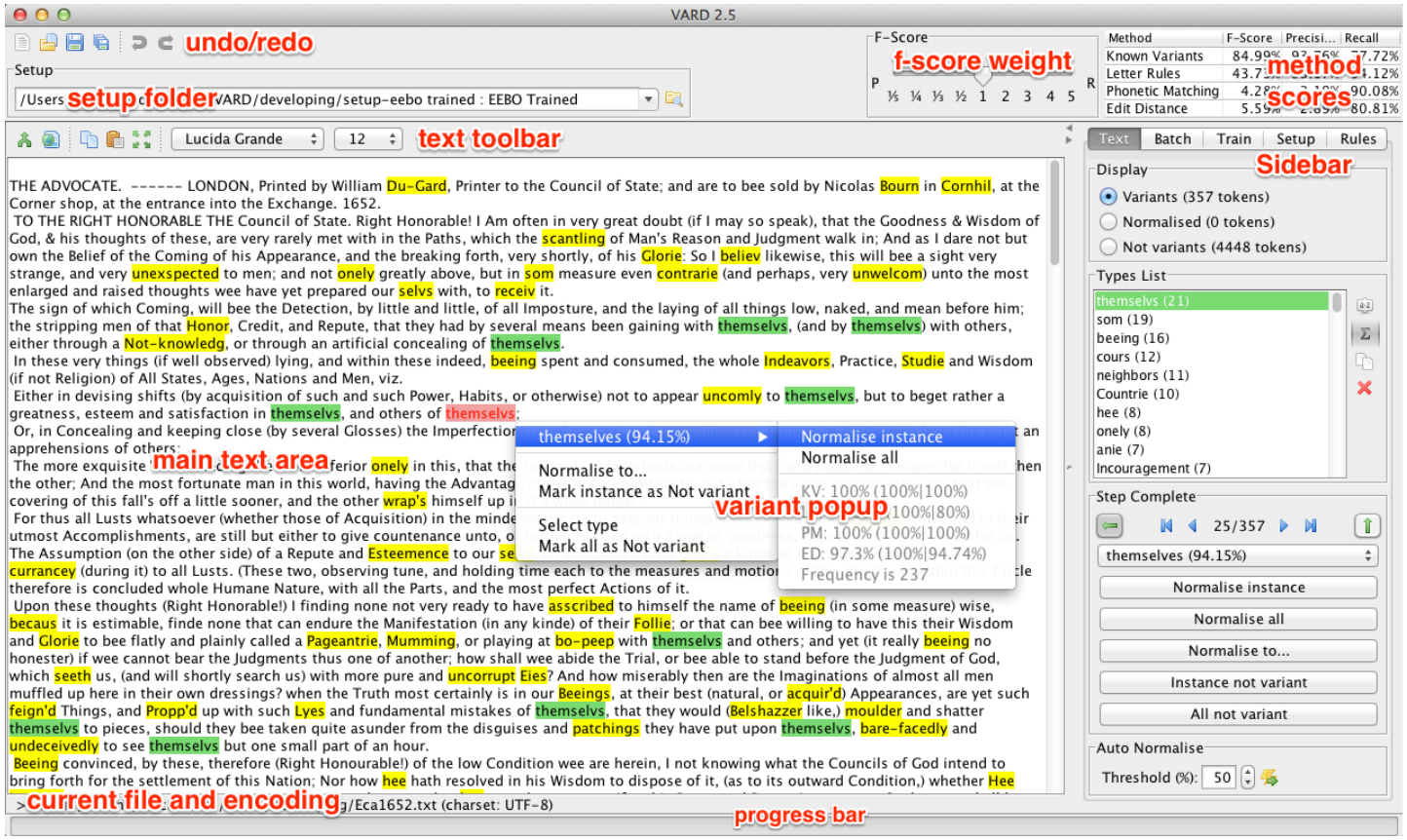
Solutions

- Historical variant spelling detection
- Need to use historically valid taxonomies or thesauri, or revise our existing modern tagsets and taggers
  - Historical Thesaurus of English (Christian Kay et al)
  - Shakespearean Thesaurii: Spevack (1993), Trussler (1986)

VARD (VARiant Detector)

http://ucrel.lancs.ac.uk/vard/

- VARD training – Innsbruck, EMEMT, CEEC, CED, newsbooks corpora



Pre-processing

LANCASTER UNIVERSITY

- Removal of many XML tags that split words; problems deciding whether to replace them with a space character
- Vowels with macrons were converted into equivalent HTML hex entities
- Sentence breaks
  - Because of things.</s> <s>That's right, because of things
- Metadata extraction

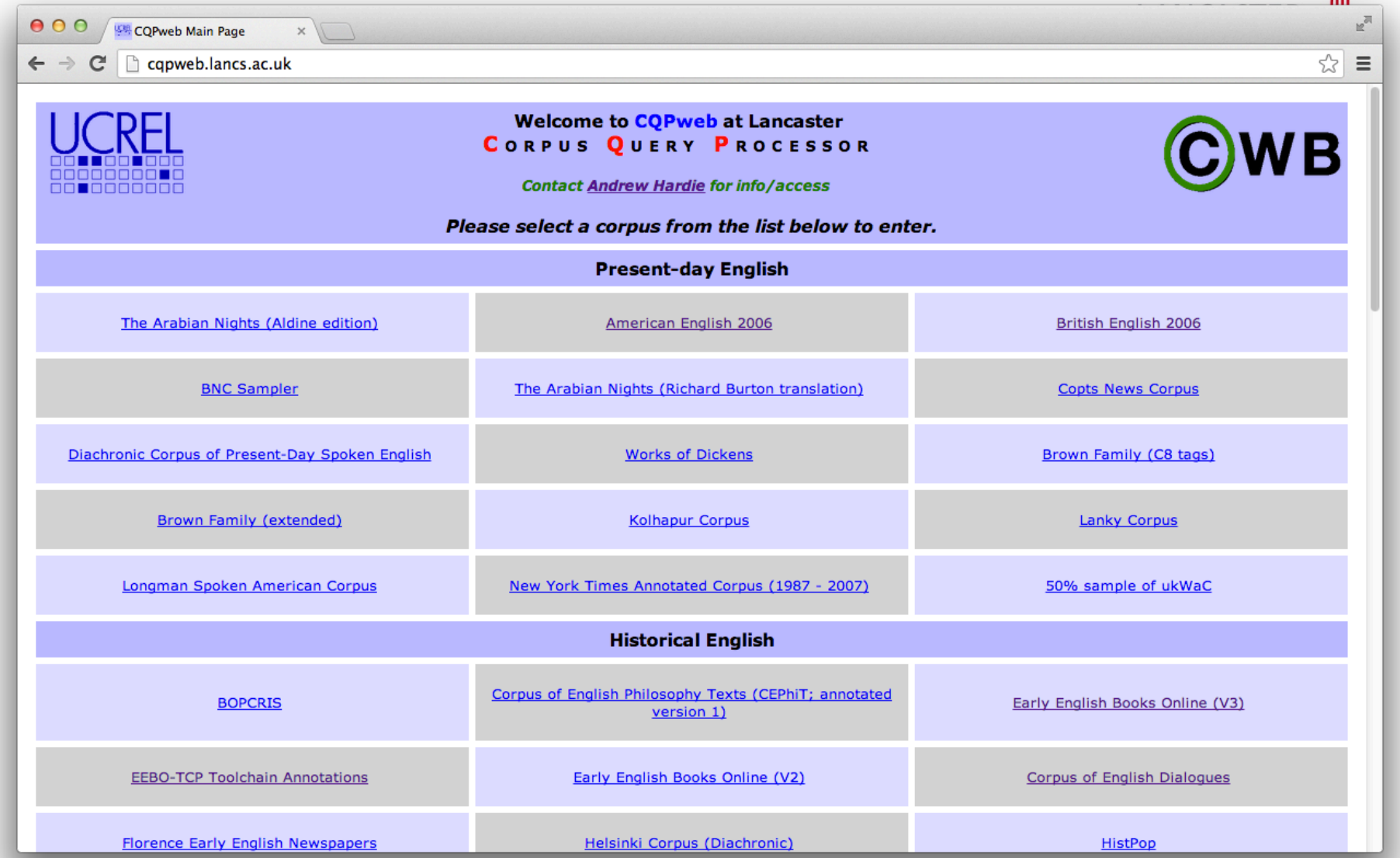
Processing Power

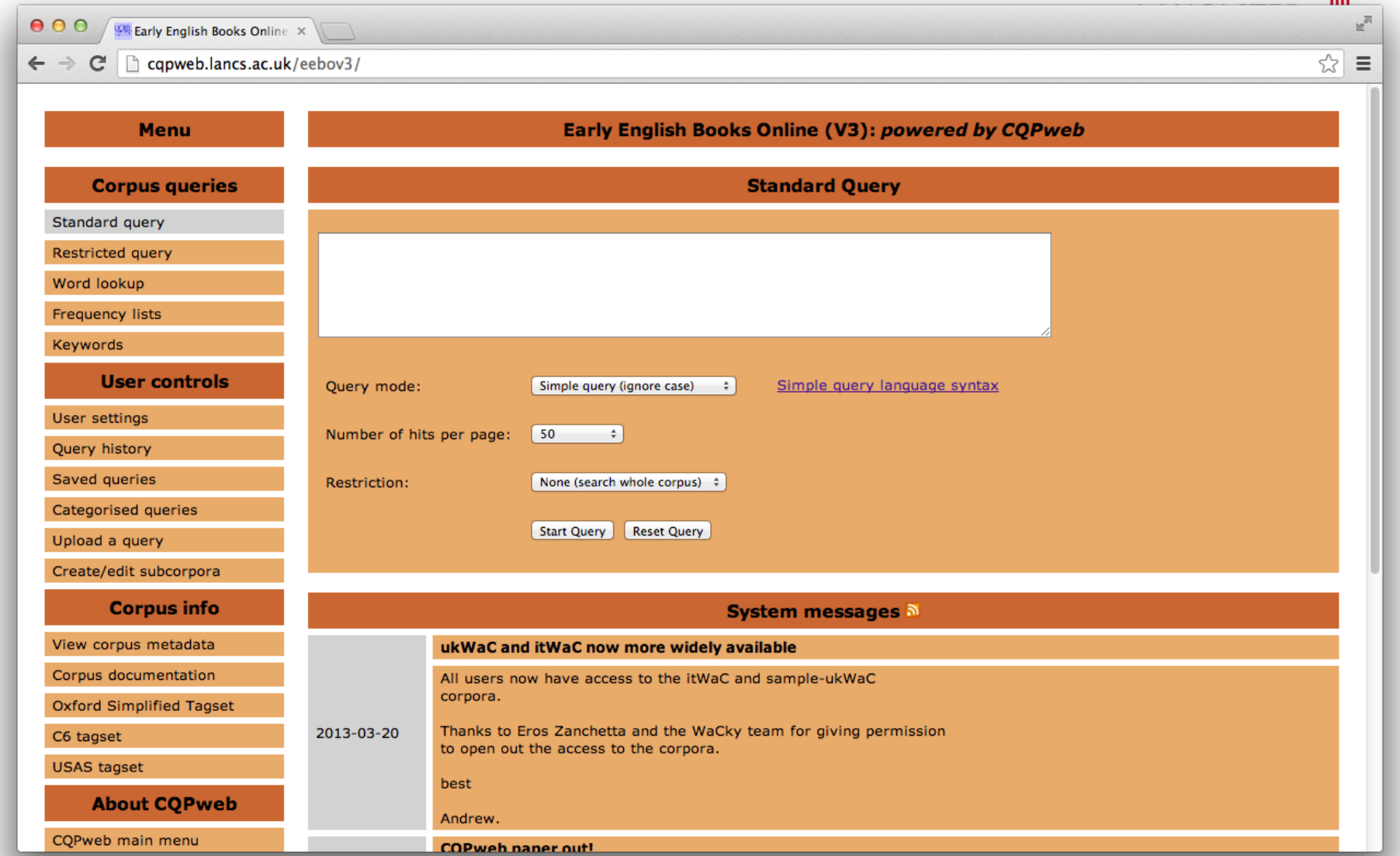
- 55 hours (tagging time) plus 18 hours (indexing time and frequency list preparation) …
- 8 to 11 processes running in parallel at any one time (tagging)
- 1 processor indexing (CQPweb doesn't allow parallel indexing)
- Two virtual machines running Debian OS
  - 8 core 8Gb RAM (VMware, network storage)
  - 8 core 5Gb RAM (Virtual box, local disks)

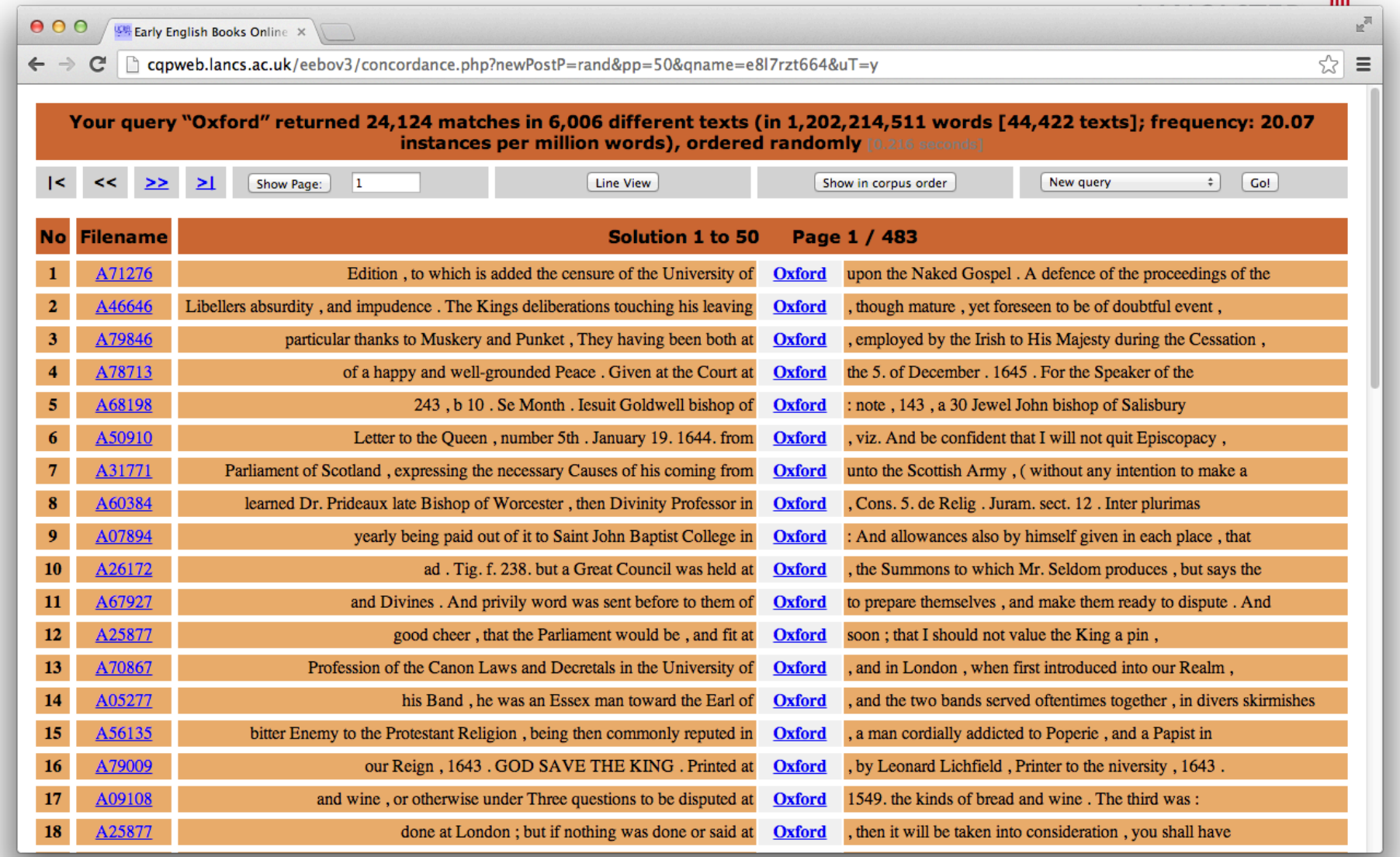
LIVE DEMO?!!!!!

LANCASTER UNIVERSITY

- CQPweb screenshots/demonstration showing
  - basic word search (Oxford)
  - basic modernised variant search ("would" vs variants of would)
  - "experiment" (distribution plot)
  - "liberty" (with semantic tags in context)
  - bar charts showing variation over time


CQPweb Main Page
cqpweb.lancs.ac.uk
UCREL
Welcome to CQPweb at Lancaster
CORPUS QUERY PROCESSOR
Contact Andrew Hardie for info/access
CWB
Please select a corpus from the list below to enter.
Present-day English
The Arabian Nights (Aldine edition)
American English 2006
British English 2006
BNC Sampler
The Arabian Nights (Richard Burton translation)
Copts News Corpus
Diachronic Corpus of Present-Day Spoken English
Works of Dickens
Brown Family (C8 tags)
Brown Family (extended)
Kolhapur Corpus
Lanky Corpus
Longman Spoken American Corpus
New York Times Annotated Corpus (1987 - 2007)
50% sample of ukWaC
Historical English
BOPCRIS
Corpus of English Philosophy Texts (CEPhiT; annotated version 1)
Early English Books Online (V3)
EEBO-TCP Toolchain Annotations
Early English Books Online (V2)
Corpus of English Dialogues
Florence Early English Newspapers
Helsinki Corpus (Diachronic)
HistPop

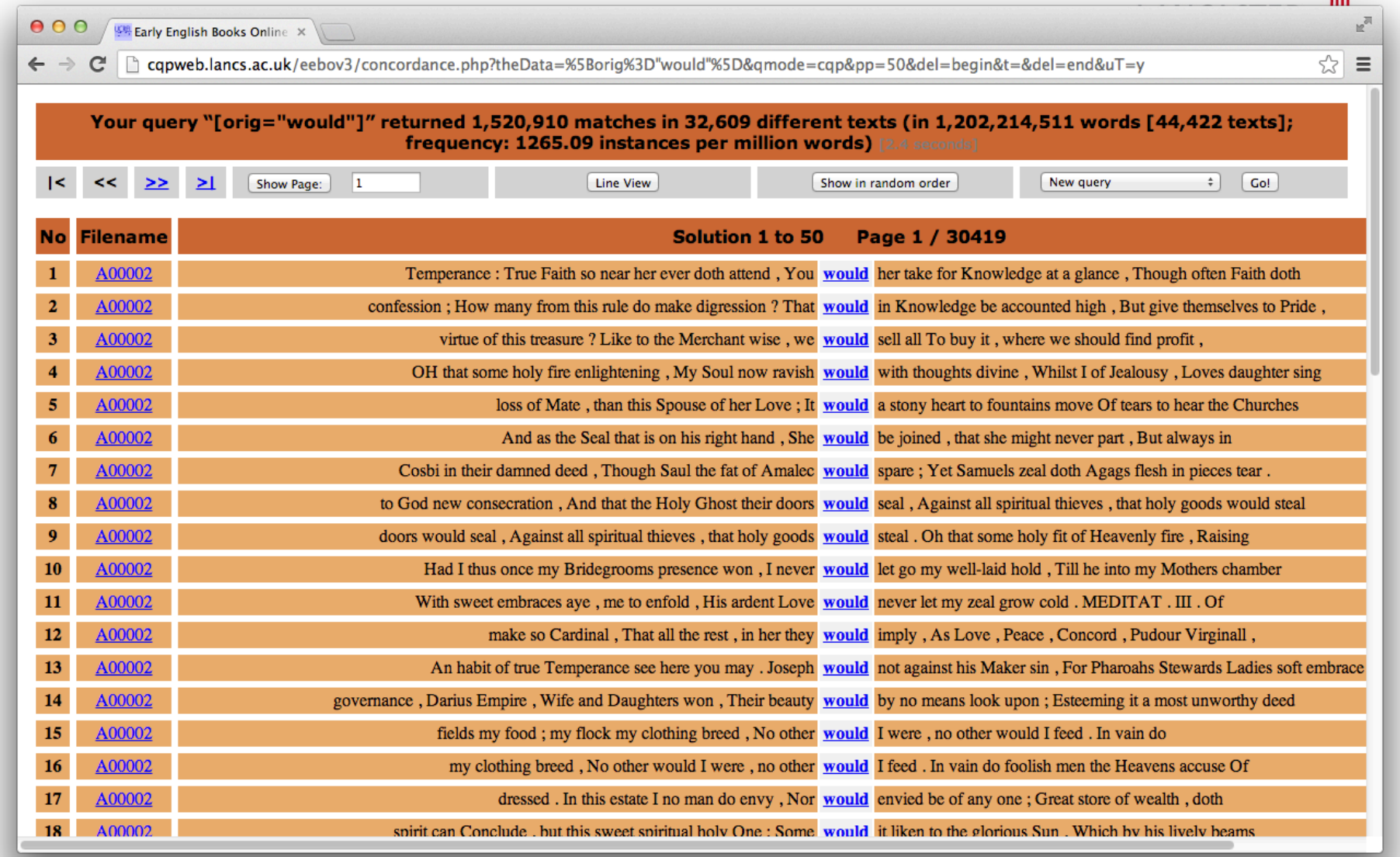

Early English Books Online
cqpweb.lancs.ac.uk/eebov3/
Menu
Early English Books Online (V3): powered by CQPweb
Corpus queries
Standard query
Restricted query
Word lookup
Frequency lists
Keywords
User controls
User settings
Query history
Saved queries
Categorised queries
Upload a query
Create/edit subcorpora
Corpus info
View corpus metadata
Corpus documentation
Oxford Simplified Tagset
C6 tagset
USAS tagset
About CQPweb
CQPweb main menu
Standard Query
Query mode:
Simple query (ignore case)
Simple query language syntax
Number of hits per page:
50
Restriction:
None (search whole corpus)
Start Query
Reset Query
System messages
2013-03-20
ukWaC and itWaC now more widely available
All users now have access to the itWaC and sample-ukWaC corpora.
Thanks to Eros Zanchetta and the WaCky team for giving permission to open out the access to the corpora.
best
Andrew.
CQPweb paper out!



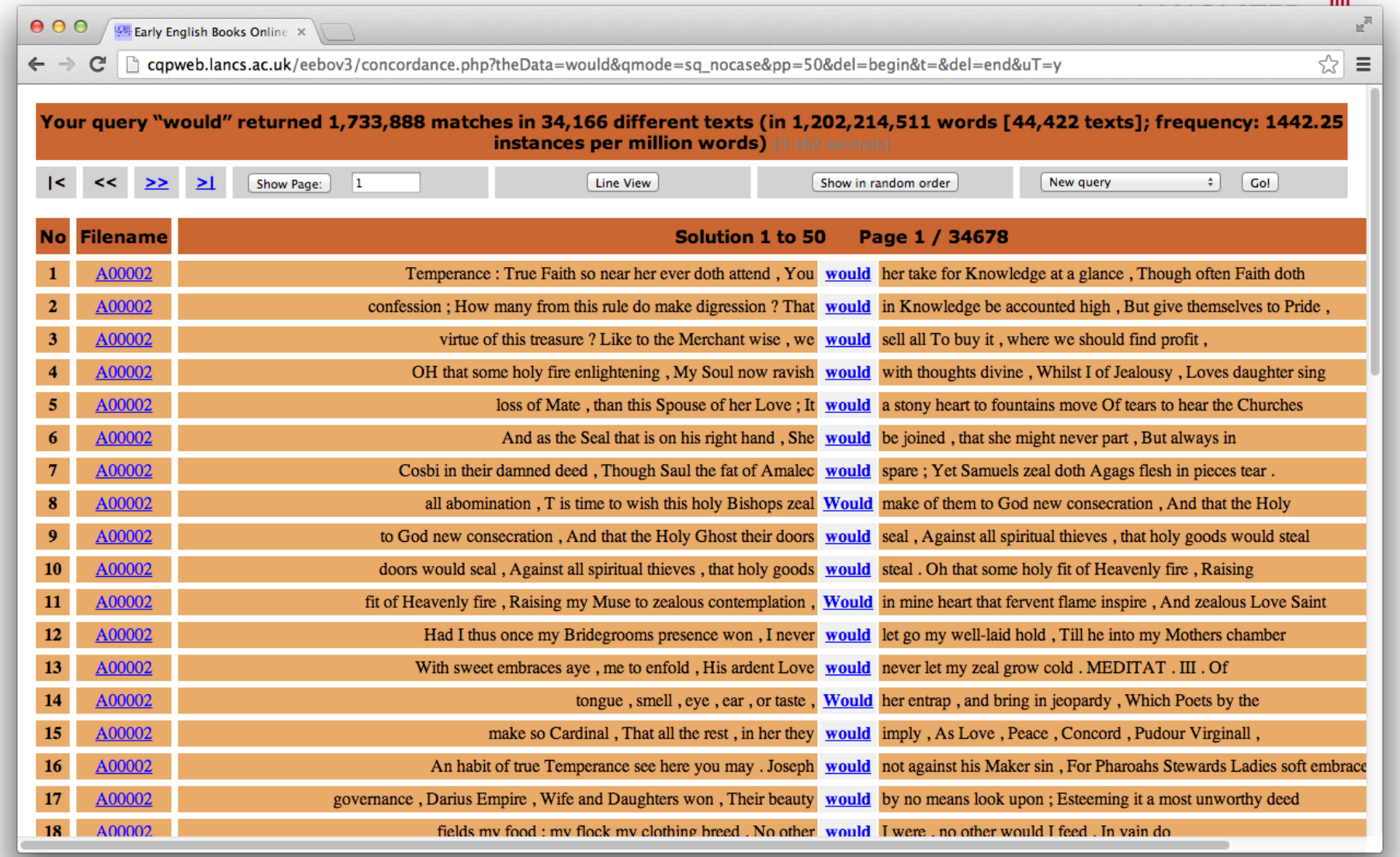
Your query "Oxford" returned 24,124 matches in 6,006 different texts (in 1,202,214,511 words [44,422 texts]; frequency: 20.07 instances per million words), ordered randomly [0.216 seconds]

|< << >> >| Show Page: 1 Line View Show in corpus order New query Go!

No	Filename	Solution 1 to 50 Page 1 / 483		
1	A71276	Edition , to which is added the censure of the University of	Oxford	upon the Naked Gospel . A defence of the proceedings of the
2	A46646	Libellers absurdity , and impudence . The Kings deliberations touching his leaving	Oxford	, though mature , yet foreseen to be of doubtful event ,
3	A79846	particular thanks to Muskery and Punket , They having been both at	Oxford	, employed by the Irish to His Majesty during the Cessation ,
4	A78713	of a happy and well-grounded Peace . Given at the Court at	Oxford	the 5. of December . 1645 . For the Speaker of the
5	A68198	243 , b 10 . Se Month . Iesuit Goldwell bishop of	Oxford	: note , 143 , a 30 Jewel John bishop of Salisbury
6	A50910	Letter to the Queen , number 5th . January 19. 1644. from	Oxford	, viz. And be confident that I will not quit Episcopacy ,
7	A31771	Parliament of Scotland , expressing the necessary Causes of his coming from	Oxford	unto the Scottish Army , (without any intention to make a
8	A60384	learned Dr. Prideaux late Bishop of Worcester , then Divinity Professor in	Oxford	, Cons. 5. de Relig . Juram. sect. 12 . Inter plurimas
9	A07894	yearly being paid out of it to Saint John Baptist College in	Oxford	: And allowances also by himself given in each place , that
10	A26172	ad . Tig. f. 238. but a Great Council was held at	Oxford	, the Summons to which Mr. Seldom produces , but says the
11	A67927	and Divines . And privily word was sent before to them of	Oxford	to prepare themselves , and make them ready to dispute . And
12	A25877	good cheer , that the Parliament would be , and fit at	Oxford	soon ; that I should not value the King a pin ,
13	A70867	Profession of the Canon Laws and Decretals in the University of	Oxford	, and in London , when first introduced into our Realm ,
14	A05277	his Band , he was an Essex man toward the Earl of	Oxford	, and the two bands served oftentimes together , in divers skirmishes
15	A56135	bitter Enemy to the Protestant Religion , being then commonly reputed in	Oxford	, a man cordially addicted to Poperie , and a Papist in
16	A79009	our Reign , 1643 . GOD SAVE THE KING . Printed at	Oxford	, by Leonard Lichfield , Printer to the niversity , 1643 .
17	A09108	and wine , or otherwise under Three questions to be disputed at	Oxford	1549. the kinds of bread and wine . The third was :
18	A25877	done at London ; but if nothing was done or said at	Oxford	, then it will be taken into consideration , you shall have

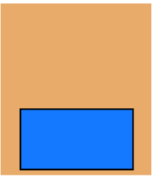


No	Filename	Solution 1 to 50 Page 1 / 30419		
1	A00002	Temperance : True Faith so near her ever doth attend , You	would	her take for Knowledge at a glance , Though often Faith doth
2	A00002	confession ; How many from this rule do make digression ? That	would	in Knowledge be accounted high , But give themselves to Pride ,
3	A00002	virtue of this treasure ? Like to the Merchant wise , we	would	sell all To buy it , where we should find profit ,
4	A00002	OH that some holy fire enlightening , My Soul now ravish	would	with thoughts divine , Whilst I of Jealousy , Loves daughter sing
5	A00002	loss of Mate , than this Spouse of her Love ; It	would	a stony heart to fountains move Of tears to hear the Churches
6	A00002	And as the Seal that is on his right hand , She	would	be joined , that she might never part , But always in
7	A00002	Cosbi in their damned deed , Though Saul the fat of Amalec	would	spare ; Yet Samuels zeal doth Agags flesh in pieces tear .
8	A00002	to God new consecration , And that the Holy Ghost their doors	would	seal , Against all spiritual thieves , that holy goods would steal
9	A00002	doors would seal , Against all spiritual thieves , that holy goods	would	steal . Oh that some holy fit of Heavenly fire , Raising
10	A00002	Had I thus once my Bridegrooms presence won , I never	would	let go my well-laid hold , Till he into my Mothers chamber
11	A00002	With sweet embraces aye , me to enfold , His ardent Love	would	never let my zeal grow cold . MEDITAT . III . Of
12	A00002	make so Cardinal , That all the rest , in her they	would	imply , As Love , Peace , Concord , Pudour Virginall ,
13	A00002	An habit of true Temperance see here you may . Joseph	would	not against his Maker sin , For Pharoahs Stewards Ladies soft embrace
14	A00002	governance , Darius Empire , Wife and Daughters won , Their beauty	would	by no means look upon ; Esteeming it a most unworthy deed
15	A00002	fields my food ; my flock my clothing breed , No other	would	I were , no other would I feed . In vain do
16	A00002	my clothing breed , No other would I were , no other	would	I feed . In vain do foolish men the Heavens accuse Of
17	A00002	dressed . In this estate I no man do envy , Nor	would	envied be of any one ; Great store of wealth , doth
18	A00002	spirit can Conclude , but this sweet spiritual holy One : Some	would	it liken to the glorious Sun . Which by his lively beams

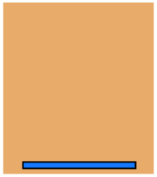


No	Filename	Solution 1 to 50 Page 1 / 34678		
1	A00002	Temperance : True Faith so near her ever doth attend , You	would	her take for Knowledge at a glance , Though often Faith doth
2	A00002	confession ; How many from this rule do make digression ? That	would	in Knowledge be accounted high , But give themselves to Pride ,
3	A00002	virtue of this treasure ? Like to the Merchant wise , we	would	sell all To buy it , where we should find profit ,
4	A00002	OH that some holy fire enlightening , My Soul now ravish	would	with thoughts divine , Whilst I of Jealousy , Loves daughter sing
5	A00002	loss of Mate , than this Spouse of her Love ; It	would	a stony heart to fountains move Of tears to hear the Churches
6	A00002	And as the Seal that is on his right hand , She	would	be joined , that she might never part , But always in
7	A00002	Cosbi in their damned deed , Though Saul the fat of Amalec	would	spare ; Yet Samuels zeal doth Agags flesh in pieces tear .
8	A00002	all abomination , T is time to wish this holy Bishops zeal	Would	make of them to God new consecration , And that the Holy
9	A00002	to God new consecration , And that the Holy Ghost their doors	would	seal , Against all spiritual thieves , that holy goods would steal
10	A00002	doors would seal , Against all spiritual thieves , that holy goods	would	steal . Oh that some holy fit of Heavenly fire , Raising
11	A00002	fit of Heavenly fire , Raising my Muse to zealous contemplation ,	Would	in mine heart that fervent flame inspire , And zealous Love Saint
12	A00002	Had I thus once my Bridegrooms presence won , I never	would	let go my well-laid hold , Till he into my Mothers chamber
13	A00002	With sweet embraces aye , me to enfold , His ardent Love	would	never let my zeal grow cold . MEDITAT . III . Of
14	A00002	tongue , smell , eye , ear , or taste ,	Would	her entrap , and bring in jeopardy , Which Poets by the
15	A00002	make so Cardinal , That all the rest , in her they	would	imply , As Love , Peace , Concord , Pudour Virginall ,
16	A00002	An habit of true Temperance see here you may . Joseph	would	not against his Maker sin , For Pharoahs Stewards Ladies soft embrace
17	A00002	governance , Darius Empire , Wife and Daughters won , Their beauty	would	by no means look upon ; Esteeming it a most unworthy deed
18	A00002	fields my food ; my flock my clothing breed . No other	would	I were . no other would I feed . In vain do

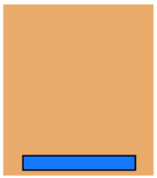
Distribution breakdown for query "[orig="experiment"]": this query returned 4,848 matches in 2,176 different texts

1590_1599	1600_1609	1610_1619	1620_1629	1630_1639	1640_1649	1650_1659	1660_1669	1670_1679	1680_1689	1690_1699	1700_1709
											
95	132	165	159	411	367	1022	697	717	574	274	28
39.94	57.27	61.92	55.75	63.5	87.48	168.91	112	118.17	142.07	128.49	17.23
2.38	2.3	2.66	2.85	6.47	4.2	6.05	6.22	6.07	4.04	2.13	1.63

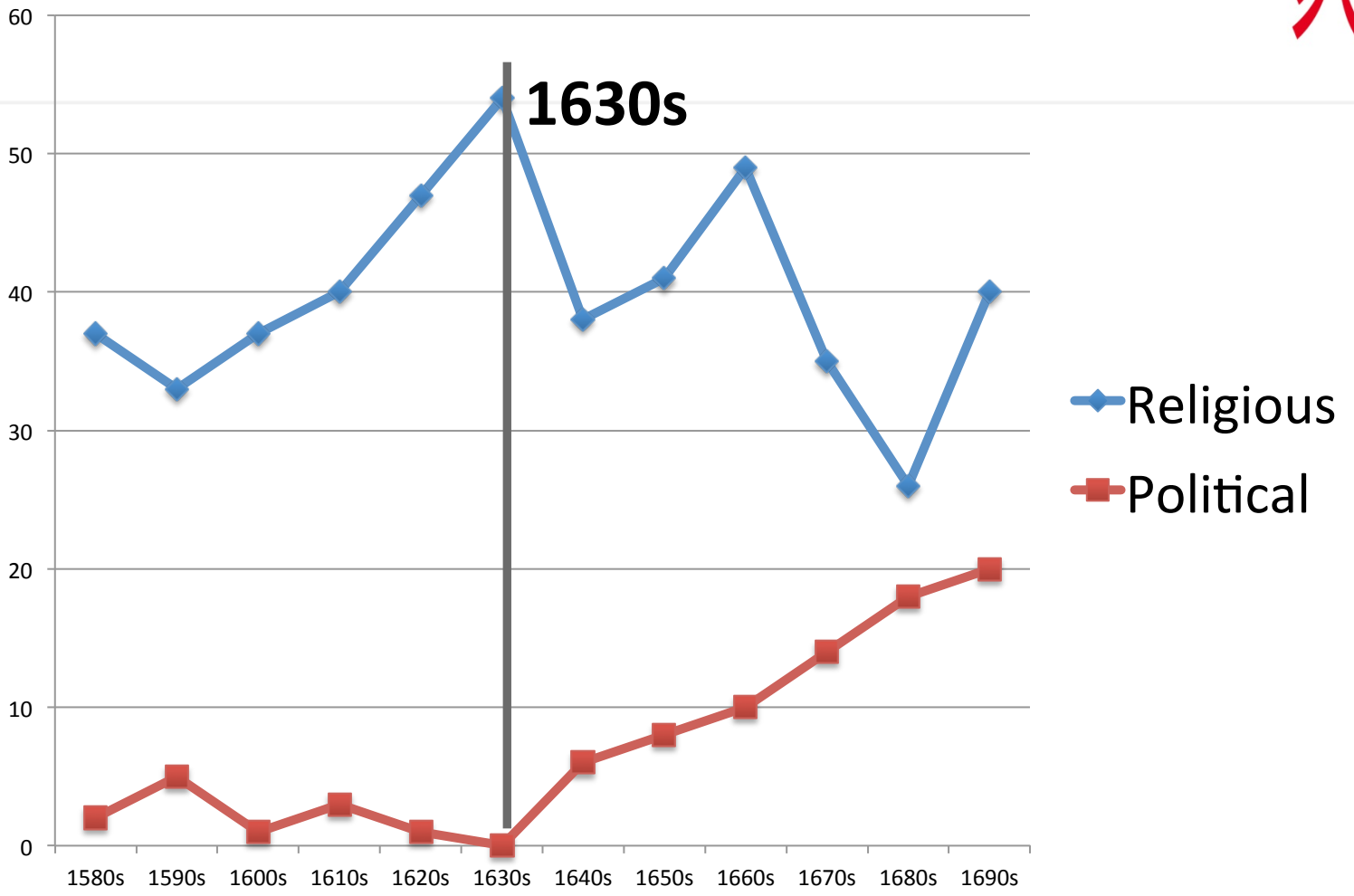
Distribution breakdown for query "experiment": this query returned 14,299 matches in 3,108 different texts

1590_1599	1600_1609	1610_1619	1620_1629	1630_1639	1640_1649	1650_1659	1660_1669	1670_1679	1680_1689	1690_1699	1700_1709
											
99	136	171	435	423	394	1517	3063	2188	2469	1727	1423
39.94	57.27	61.92	55.75	63.5	87.48	168.91	112	118.17	142.07	128.49	17.23
2.48	2.37	2.76	7.8	6.66	4.5	8.98	27.35	18.52	17.38	13.44	82.59

Distribution breakdown for query "{experiment}": this query returned 20,156 matches in 4,405 different texts

1590_1599	1600_1609	1610_1619	1620_1629	1630_1639	1640_1649	1650_1659	1660_1669	1670_1679	1680_1689	1690_1699	1700_1709
											
267	382	425	760	857	829	2613	3889	2912	3039	2016	1498
39.94	57.27	61.92	55.75	63.5	87.48	168.91	112	118.17	142.07	128.49	17.23
6.69	6.67	6.86	13.63	13.5	9.48	15.47	34.72	24.64	21.39	15.69	86.95

The most important result concerning discourses of liberty??
Decline of the religious; rise of the political (?)

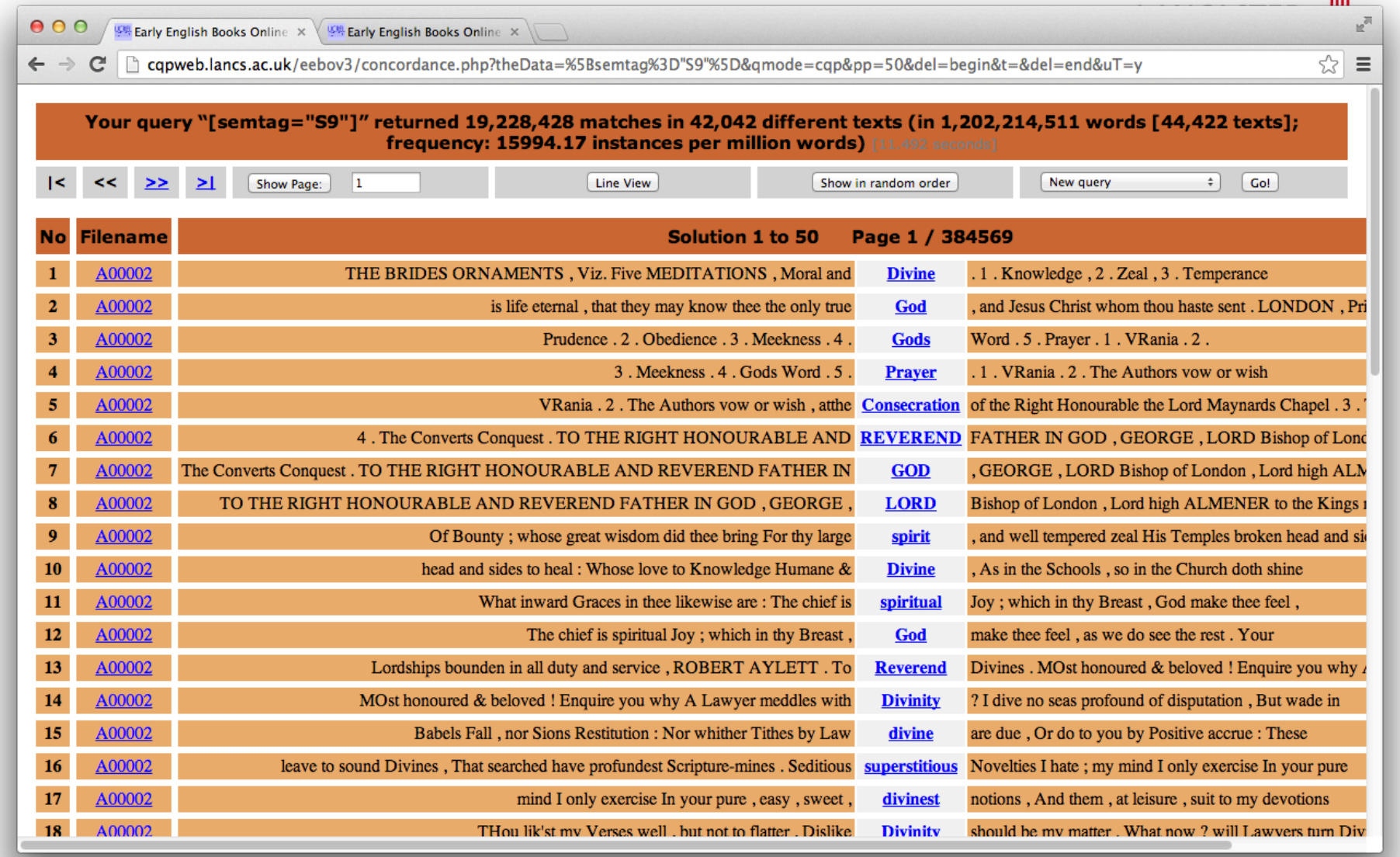


"Religious" and "Political" occurences of "Liberty" in the early modern period 1580s-1690s.

l[i,y]bert[e,ee,ie,ye,y]

LANCASTER UNIVERSITY

Category [↓]	Words in category	Hits in category	Dispersion (no. files with 1+ hits)	Frequency [↓] per million words in category
1580_1589	42240219	5053	469 out of 856	119.63
1590_1599	39940085	4987	570 out of 1046	124.86
1600_1609	57272438	8399	757 out of 1349	146.65
1610_1619	61922837	8091	769 out of 1465	130.66
1620_1629	55748866	6949	809 out of 1674	124.65
1630_1639	63496401	8357	865 out of 1572	131.61
1640_1649	87480996	18623	3096 out of 7424	212.88
1650_1659	168912439	30545	2724 out of 4690	180.83
1660_1669	111998646	18809	2108 out of 4227	167.94
1670_1679	118167747	19735	1962 out of 3920	167.01
1680_1689	142071417	26708	3094 out of 6916	187.99
1690_1699	128494904	22197	2648 out of 5588	172.75


Early English Books Online
Early English Books Online
cqpweb.lancs.ac.uk/eebov3/concordance.php?theData=%5Bsemtag%3D"S9"%5D&qmode=cqp&pp=50&del=begin&t=&del=end&uT=y
Your query "[semtag="S9"]" returned 19,228,428 matches in 42,042 different texts (in 1,202,214,511 words [44,422 texts]; frequency: 15994.17 instances per million words) [11.492 seconds]
|< << >> >|
Show Page: 1
Line View
Show in random order
New query
Go!
No Filename Solution 1 to 50 Page 1 / 384569
1 A00002 THE BRIDES ORNAMENTS , Viz. Five MEDITATIONS , Moral and Divine . 1 . Knowledge , 2 . Zeal , 3 . Temperance
2 A00002 is life eternal , that they may know thee the only true God , and Jesus Christ whom thou haste sent . LONDON , Pri
3 A00002 Prudence . 2 . Obedience . 3 . Meekness . 4 . Gods Word . 5 . Prayer . 1 . VRania . 2 .
4 A00002 3 . Meekness . 4 . Gods Word . 5 . Prayer . 1 . VRania . 2 . The Authors vow or wish
5 A00002 VRania . 2 . The Authors vow or wish , atthe Consecration of the Right Honourable the Lord Maynards Chapel . 3 .
6 A00002 4 . The Converts Conquest . TO THE RIGHT HONOURABLE AND REVEREND FATHER IN GOD , GEORGE , LORD Bishop of Lond
7 A00002 The Converts Conquest . TO THE RIGHT HONOURABLE AND REVEREND FATHER IN GOD , GEORGE , LORD Bishop of London , Lord high ALM
8 A00002 TO THE RIGHT HONOURABLE AND REVEREND FATHER IN GOD , GEORGE , LORD Bishop of London , Lord high ALMENER to the Kings
9 A00002 Of Bounty ; whose great wisdom did thee bring For thy large spirit , and well tempered zeal His Temples broken head and si
10 A00002 head and sides to heal : Whose love to Knowledge Humane & Divine , As in the Schools , so in the Church doth shine
11 A00002 What inward Graces in thee likewise are : The chief is spiritual Joy ; which in thy Breast , God make thee feel ,
12 A00002 The chief is spiritual Joy ; which in thy Breast , God make thee feel , as we do see the rest . Your
13 A00002 Lordships bounden in all duty and service , ROBERT AYLETT . To Reverend Divines . MOst honoured & beloved ! Enquire you why
14 A00002 MOst honoured & beloved ! Enquire you why A Lawyer meddles with Divinity ? I dive no seas profound of disputation , But wade in
15 A00002 Babels Fall , nor Sions Restitution : Nor whither Tithes by Law divine are due , Or do to you by Positive accrue : These
16 A00002 leave to sound Divines , That searched have profundest Scripture-mines . Seditious superstitious Novelties I hate ; my mind I only exercise In your pure
17 A00002 mind I only exercise In your pure , easy , sweet , divinest notions , And them , at leisure , suit to my devotions
18 A00002 THou lik'st my Verses well , but not to flatter . Dislike Divinity should be my matter . What now ? will Lawyers turn Div

No	Filename	Solution 1 to 50 Page 1 / 417		
1	A00005	's enfected with heresye as was Valens & other / wherefore the	church had no liberty	when valens was on live . A Synodus of a h ö
2	A00005	Charles the king of France / to whom it was granted the	liberty of eleccyon of the popes	and to ordeyne the appostles set . Leo the fourt
3	A00007	infected with heresy / as was Valens & other / wherefore y	church had no liberty	when Valens was on live . A synodus of an . C
4	A00007	Charles the king of France / to whom it was granted the	liberty of eleccyon of the pope	/ and to ordeyne y apostles set . Leo the fourth
5	A00011	of the Cathdrall , or Mother Churches , saving always the general	liberty lest the Bishop	by Law , during the time of the Administration
6	A00011	Only here is one help peradventure for it , if the	Bishop be pleased to give liberty	to Parishes to remove the Table into the midst r
7	A00011	help peradventure for it , if the Bishop be pleased to give	liberty to Parishes	to remove the Table into the midst r the Comm
8	A00011	liberty to Parishes to remove the Table into the midst r the	Communion ; this liberty	always Saved . So as here is all the hope of Par
9	A00089	of the Cathedral , or Mother Churches , saving always the general	liberty left to the Bishop	by Law , during the time of Administration of t
10	A00184	the Kings declaration in that behalf ? Or do any of your	Parish abuse that liberty	allowed them , by using unlawful sports , as Bo
11	A00233	matters of Divinity their ordinary Table-talke ? Or that under pretence of	holiness and edification , take the liberty	at their Trencher-meetings , or where several co
12	A00271	is incident , but hereunto can I tell you , that your	liberty of the late Gospel	has provided a more sufficient remedy , to avoi
13	A00275	well to those that be touchyd about the popishe apparel in this	liberty of Gods	truth , which is taught plainly without offencis
14	A00275	to God 's glory . Concurrythe with his word , edefyeth his	church , mayntaynyth Christyan liberty	: which conditions & circunstances if they wan
15	A00275	we should Minor . do that which is repngnaunt to Christian	liberty , maintain an opinion of holiness	, where none is , show consent to Idolatry , der
16	A00277	well to those that be touchyd about the popishe apparel in this	liberty of Gods	truth , which is taught plainly without offencis
17	A00277	to God 's glory . Concurrythe with his word , edefyeth his	church , mayntaynyth Christyan liberty	: which conditions & circumstances if they war
18	A00277	Minor . we should do that which is repugnant to Christian	liberty , maintain an opinion of holiness	, where none is , show consent to Idolatry , der

No	Filename	Solution 1 to 50 Page 1 / 16		
1	A00505	of the Senate , in case they did any thing diminish their	liberty , & should kepe the Senators	short , that they should not tiranniz
2	A00616	with what heart and courage they weere kendled to defend their	liberty . They strove with the Senators	, which of them loved the commor
3	A00619	divulged it to purpose . Read all , and use thy minds	liberty ; how thy suffrage	falls , I weigh not : For it was not v
4	A00627	this time they began to suspect to suspect their own estate &	liberty . Insomuch that the Senators	of Carthage seing the Carthaginian
5	A00637	infirmities of the people be a good reason to take away their	liberty in the Election	of their Ministers , then the contrar
6	A00658	of such whose parents diceassed , and they left to their own	election of liberty	and life with full Coffers , waste th
7	A00662	so honest a prison , to hope to settle himself eftsoones in	liberty : Oh seigneur	Pasquier says Phylopolo , are you
8	A01080	nature is declared : and that to ruele the people by government	politic is no yoke , but liberty	, and great security not only to the
9	A01160	Huguenots , which war takes from them . And what ? the	liberty of conscience . Many politic	and wise men , have acknowledge
10	A01516	all , and exempted from Argument ; The MIDDLR TERM to be	elected at the liberty	of every Mans Invention : The Rec
11	A01556	been deliuerit from ye hands of unfaithful tyrantis , set again in	liberty , Joseph . lib	13 cap . 18. de antiq. iud . and rest
12	A01567	this our present ordinance , that whosoever shall hereafter come , the	election ought to remain at liberty	, unto the consent of ye rest of our
13	A01685	it is no wisdom for a man that lends to tie his	liberty at another 's election	, nor conscience in him that borrov
14	A01760	libei in usu vel abstinentia rerum adiaphorarum . Calvine speaking of our	liberty in things indifferent , Instit. lib.	3 as nunc usurpare nunt omittere ir
15	A01802	the covent set down this order , that they should be at	liberty (Notwithstanding these elections	) to choose again , so they chose a
16	A01802	and so the Popes) the Chapter of Sarum had the	liberty of a fre election	, and chose one John Chaundler ,
17	A01802	made divers laws and statutes to restore churches and couents to the	liberty of their ancient elections	. When the king therefore had labo
18	A01802	remembering his promise often made , hereafter he would not impeach the	liberty of elections	by interposing his armed requests

Important considerations & future work

LANCASTER UNIVERSITY

- Looking at the whole library in one go?
- Representativeness
  - Should we just include all the texts?
  - First lesson for Steve Pumfrey was to normalise frequencies e.g. per decade
  - Should we select from the distribution of texts in EEBO-TCP to address issues of comparability across time, genre and text type?
  - What proportions of EEBO material per decade have been transcribed and how does this change across TCP releases?
  - Do historians consider this a good representative sample of books available at the time, for a general readership?

LOB & Brown family of corpora

LANCASTER UNIVERSITY

Text categories		Number of samples in each category		
		Brown Corpus	LOB Corpus	
A	Press: reportage	44	44	Press
B	Press: editorial	27	27	
C	Press: reviews	17	17	
D	Religion	17	17	General Prose
E	Skills, trades and hobbies	36	38	
F	Popular lore	48	44	
G	Belles lettres, biography, essays	75	77	
H	Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ)	30	30	
J	Learned and scientific writings	80	80	Learned
K	General fiction	29	29	Fiction
L	Mystery and detective fiction	24	24	
M	Science fiction	6	6	
N	Adventure and western fiction	29	29	
P	Romance and love story	29	29	
R	Humour	9	9	
Total		500	500	

British National Corpus

LANCASTER UNIVERSITY

	texts	words %
Spoken demographic	153	4.30
Spoken context-governed	755	6.27
Written books and periodicals	2685	80.55
Written-to-be-spoken	35	1.29
Written miscellaneous	421	7.56

	texts	words %
Imaginative	476	18.75
Informative: natural & pure science	146	4.34
Informative: applied science	370	8.15
Informative: social science	526	15.94
Informative: world affairs	483	19.60
Informative: commerce & finance	295	8.34
Informative: arts	261	7.47
Informative: belief & thought	146	3.45
Informative: leisure	438	13.91

Examples of balanced historical corpora

LANCASTER UNIVERSITY

- ARCHER (A Representative Corpus of Historical English Registers)
  - multi-genre corpus of British and American English 1600-1999
  - 2 million words, 50 year periods, target = 10 texts, c. 2,000w each, per genre and variety in each period
  - advertising, drama, fiction, sermons, journal, legal, medicine, news, early prose, science, letters, diary
- Helsinki corpus of English Texts
  - 1,572,800 words, c. 730–1710
  - 11 time periods (max 100 years), 400K Old English, 600K Middle English, 550K EmodE
  - socio-historical variation analysis: geographical dialect, type and register; gender, age, social rank

Corpus of Late Modern English Texts CLMET (1710-1920)

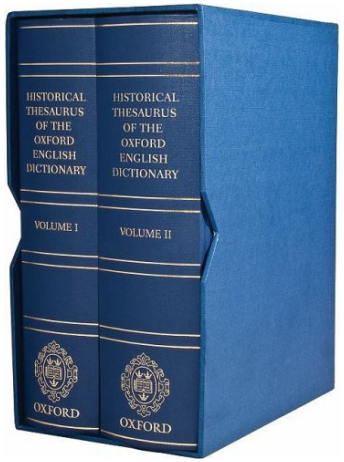
Sub-period	Number of authors	Number of texts	Number of words
1710-1780	51	88	10,480,431
1780-1850	70	99	11,285,587
1850-1920	91	146	12,620,207
TOTAL	212	333	34,386,225

Genre	1710-1780	1780-1850	1850-1920
Narrative fiction	4,642,670	4,830,718	6,311,301
Narrative non-fiction	1,863,855	1,940,245	958,410
Drama	407,885	347,493	607,401
Letters	1,016,745	714,343	479,724
Treatise	1,114,521	1,692,992	1,782,124
Other	1,434,755	1,759,796	2,481,247

- https://perswww.kuleuven.be/~u0044428/
- Diller, H., De Smet, H., Tyrkkö, J. (2011). A European database of descriptors of English electronic texts. The European English Messenger 19, 21-35.

Further work on the software

- Train VARD, CLAWS and USAS to improve accuracy on EEBO-TCP data
- Employ historically valid taxonomies or thesauri
  - Historical Thesaurus of English (Glasgow)
  - PhD student at UCLAN (supervised by Dawn Archer)



Thanks for your attention

LANCASTER UNIVERSITY

- http://ucrel.lancs.ac.uk/
- http://cqpweb.lancs.ac.uk/
- http://creme.lancs.ac.uk/

- p.rayson@lancaster.ac.uk
- @perayson

- Questions, comments?