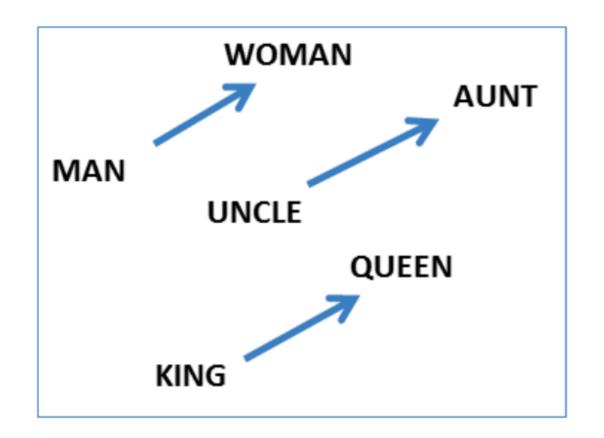
## Historical data, modern problems

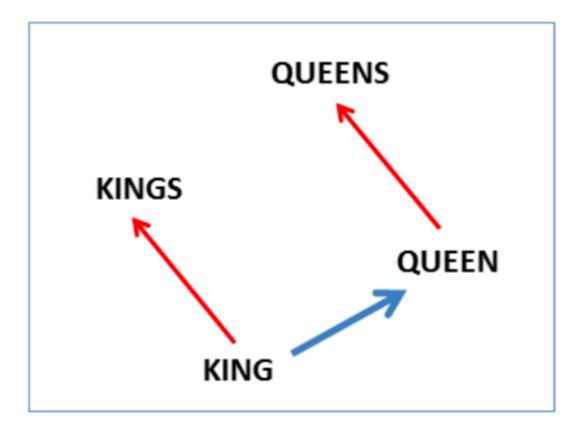
Dr Heather Froehlich Penn State University @heatherfro

## Language is always changing

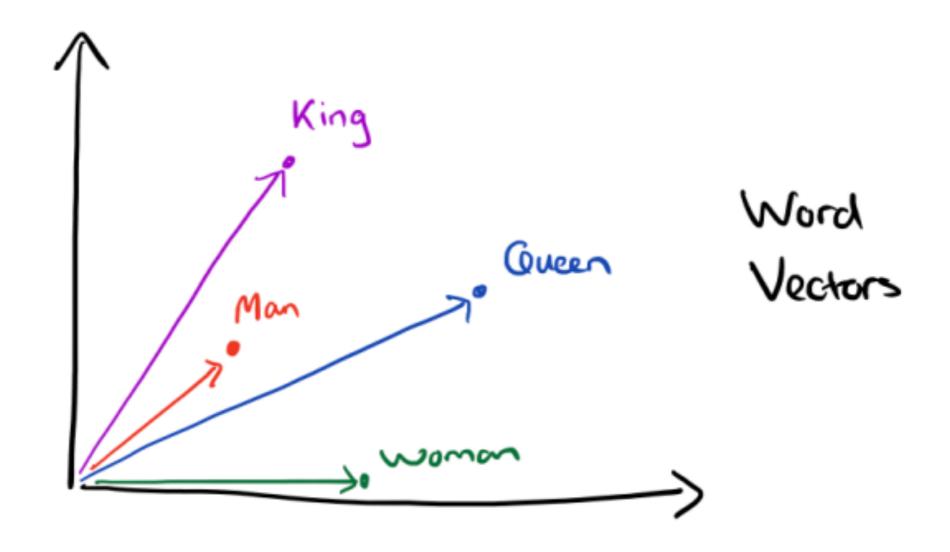
# machine learning methods are not sensitive to historical change!

(they only know the information they've been presented)

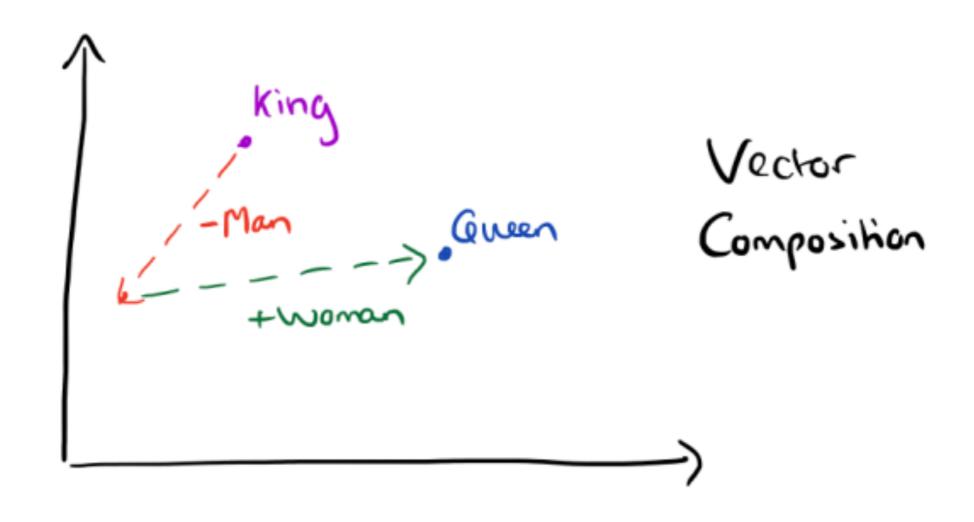




https://blog.acolyer.org/2016/04/21/the-amazingpower-of-word-vectors/ (Mikolov et al, NAACL HLT 2013)



https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/



https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

## creates word pairings based on input

'a is to b as c is to?'

## but, words fall in & out of fashion



#### corpus.byu.edu

home corpora users related resources my account upgrade help

Created by Mark Davies, BYU. Overview, search types, looking at variation, corpus-based resources, updates.

The most widely used online corpora -- more than 130,000 distinct researchers, teachers, and students each month.

English	# words	language/dialect	time period	compare
News on the Web (NOW) (See the 2017 Word of the Year)	5.7 billion+	20 countries / Web	2010-yesterday	
Global Web-Based English (GloWbE)	1.9 billion	20 countries / Web	2012-13	
Wikipedia Corpus	1.9 billion	English	-2014	Info
Hansard Corpus	1.6 billion	British (parliament)	1803-2005	Info
Early English Books Online	755 million	British	1470s-1690s	
Corpus of Contemporary American English (COCA)	560 million	American	1990-2017	****
Corpus of Historical American English (COHA)	400 million	American	1810-2009	**
Corpus of US Supreme Court Opinions	130 million	American (law)	1790s-present	
TIME Magazine Corpus	100 million	American	1923-2006	
Corpus of American Soap Operas	100 million	American	2001-2012	*
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993	**
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s	
CORE Corpus	50 million	Web registers	-2014	
Other languages				
Corpus del Español (see also)	2.1 billion	Spanish	1200s-1900s	*
Corpus do Português (see also)	1.1 billion	Portuguese	1300s-1900s	

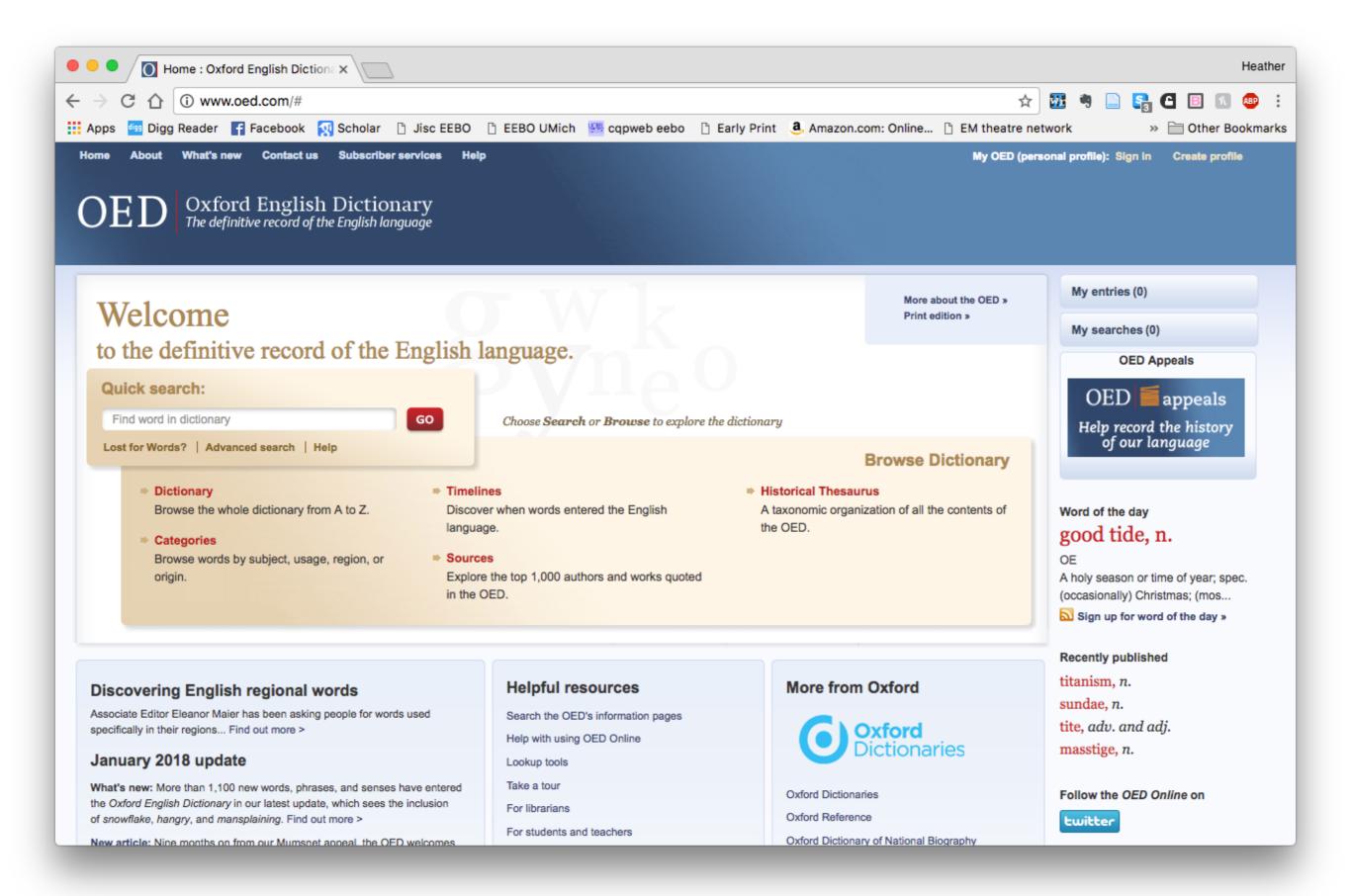
The corpus is composed of more than 400 million words of text in more than 100,000 individual texts. The major sources for each genre are as follows:

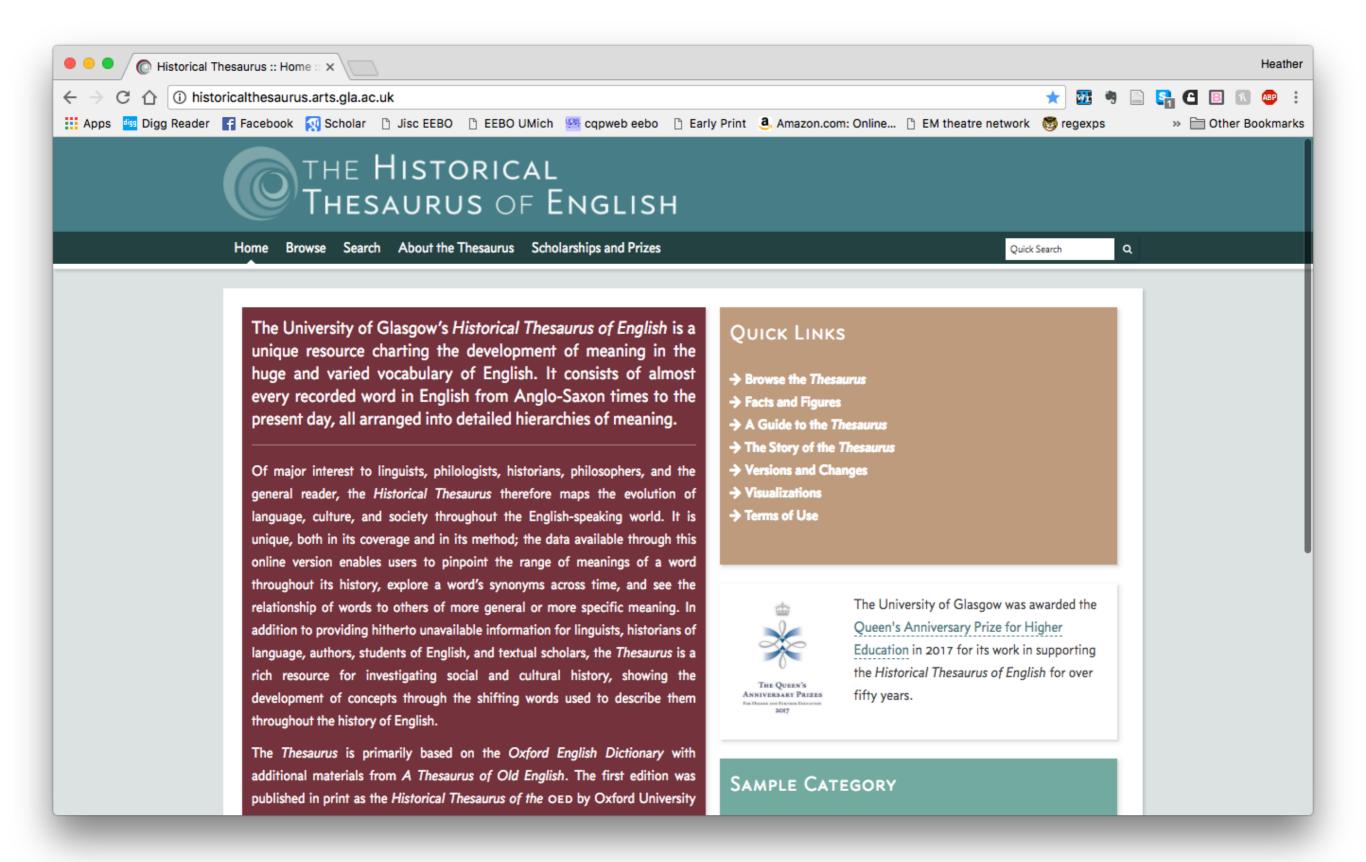
Fiction	Project Gutenberg (1810-1930), Making of America (1810-1900), scanned books (1930-1990), movie and play scripts, COCA (1990-2010)
Magazine	Making of America (1810-1900), scanned and PDF (1900-1990), COCA (1990-2010)  - In each decade, the magazines are balanced across at least ten magazines (with equivalent sub-genres for the 1900s)
Newspaper	PDF > TXT from historical archives of newspapers (1850-1980), COCA etc (1990-2010)
Non-fiction	Project Gutenberg (1810-1900), www.archive.org (1810-1900), scanned books (1900-1990), COCA (1990-2010) - In each decade, the non-fiction is balanced across the Library of Congress classification system

The corpus is balanced by genre across the decades. For example, fiction accounts for 48-55% of the total in each decade (1810s-2000s), and the corpus is balanced across decades for sub-genres and domains as well (e.g. by Library of Congress classification for non-fiction; and by sub-genre for fiction -- prose, poetry, drama, etc). This balance across genres and sub-genres allows researchers to examine changes and be reasonably certain that the data reflects actual changes in the "real world", rather than just being artifacts of a changing genre balance.

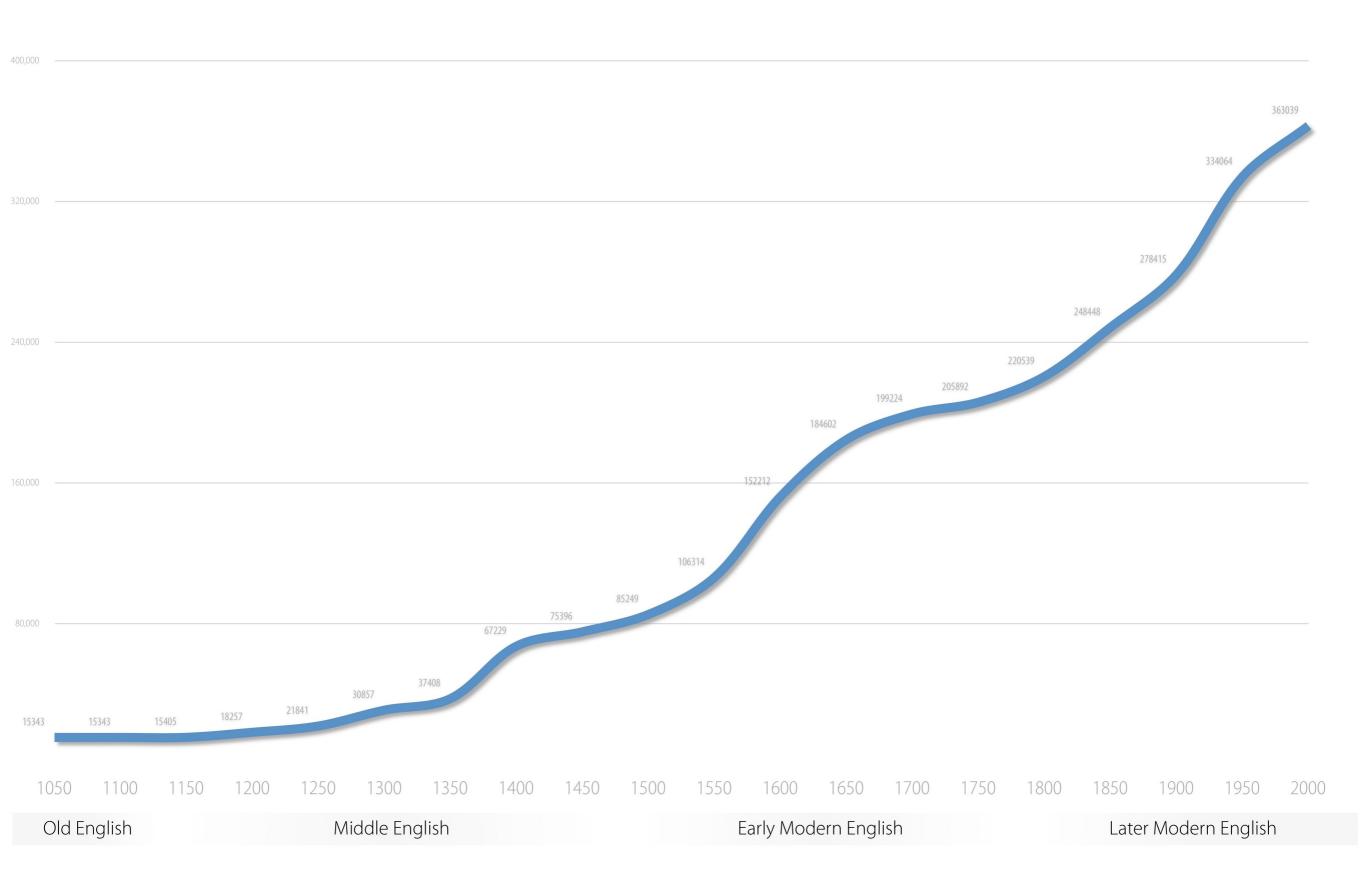
Download all 115,000 texts, for use on your own computer.

DECADE	FICTION	POPULAR MAGAZINES	NEWSPAPERS	NON-FICTION BOOKS	TOTAL	% FICTION
1810s	641,164	88,316	0	451,542	1,181,022	0.54
1820s	3,751,204	1,714,789	0	1,461,012	6,927,005	0.54
1830s	7,590,350	3,145,575	0	3,038,062	13,773,987	0.55
1840s	8,850,886	3,554,534	0	3,641,434	16,046,854	0.55
1850s	9,094,346	4,220,558	0	3,178,922	16,493,826	0.55
1860s	9,450,562	4,437,941	262,198	2,974,401	17,125,102	0.55
1870s	10,291,968	4,452,192	1,030,560	2,835,440	18,610,160	0.55
1880s	11,215,065	4,481,568	1,355,456	3,820,766	20,872,855	0.54
1890s	11,212,219	4,679,486	1,383,948	3,907,730	21,183,383	0.53
1900s	12,029,439	5,062,650	1,433,576	4,015,567	22,541,232	0.53
1910s	11,935,701	5,694,710	1,489,942	3,534,899	22,655,252	0.53
1920s	12,539,681	5,841,678	3,552,699	3,698,353	25,632,411	0.49
1930s	11,876,996	5,910,095	3,545,527	3,080,629	24,413,247	0.49
1940s	11,946,743	5,644,216	3,497,509	3,056,010	24,144,478	0.49
1950s	11,986,437	5,796,823	3,522,545	3,092,375	24,398,180	0.49
1960s	11,578,880	5,803,276	3,404,244	3,141,582	23,927,982	0.48
1970s	11,626,911	5,755,537	3,383,924	3,002,933	23,769,305	0.49
1980s	12,152,603	5,804,320	4,113,254	3,108,775	25,178,952	0.48
1990s	13,272,162	7,440,305	4,060,570	3,104,303	27,877,340	0.48
2000s	14,590,078	7,678,830	4,088,704	3,121,839	29,479,451	0.49
TOTAL	207,633,395	97,207,399	40,124,656	61,266,574	406,232,024	0.51





http://historicalthesaurus.arts.gla.ac.uk

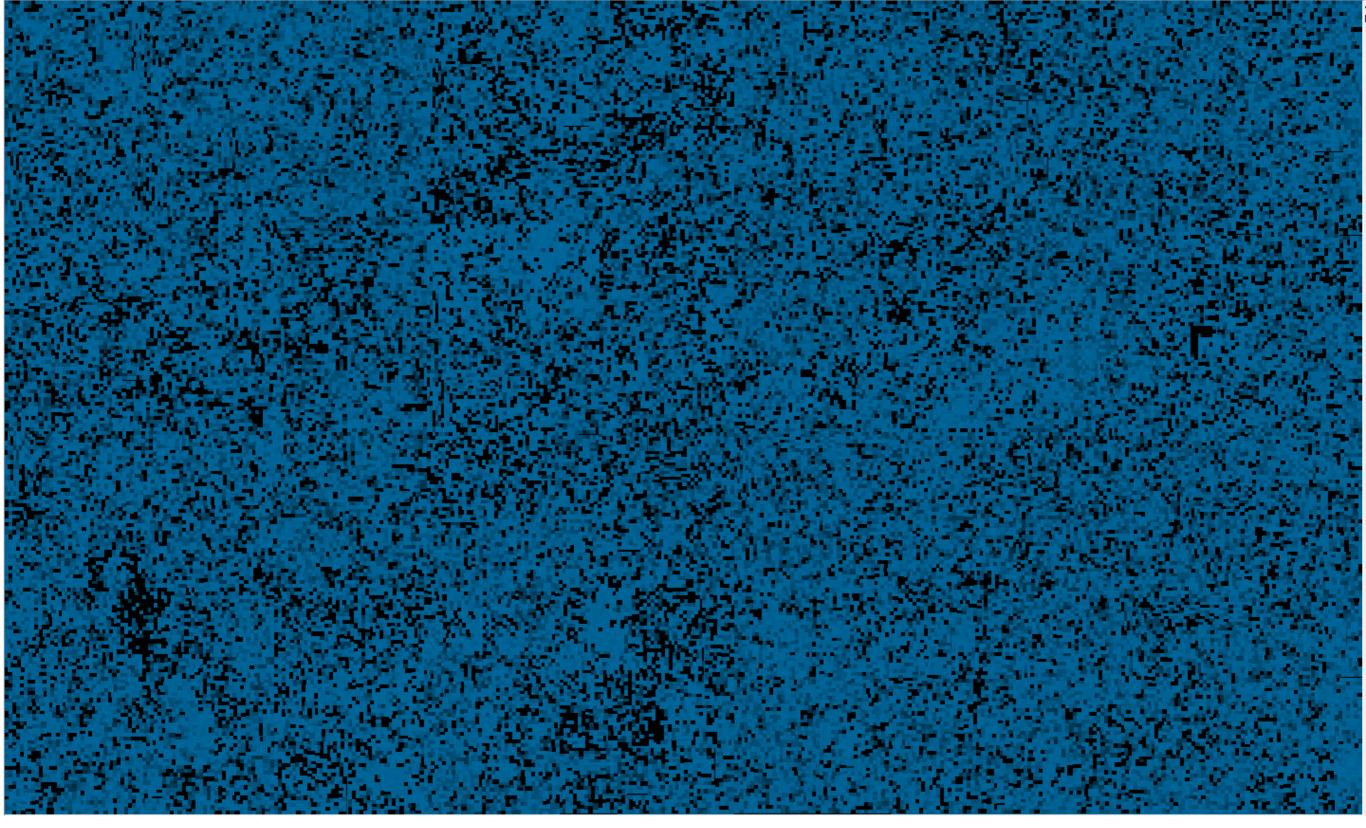


### http://historicalthesaurus.arts.gla.ac.uk

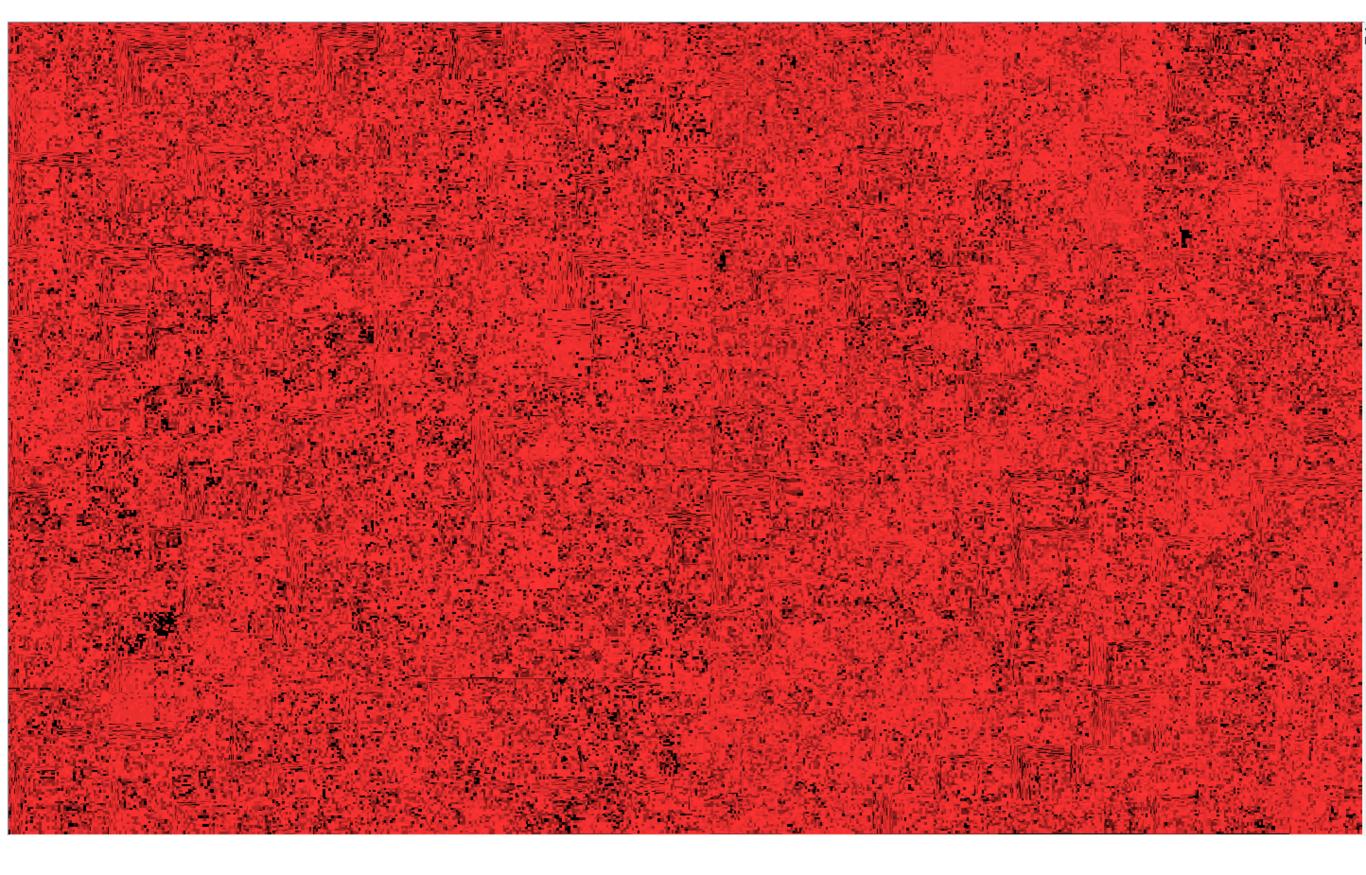


Modern English (469,470 words displayed here in all), from <a href="http://historicalthesaurus.gla.ac.uk">http://historicalthesaurus.gla.ac.uk</a>

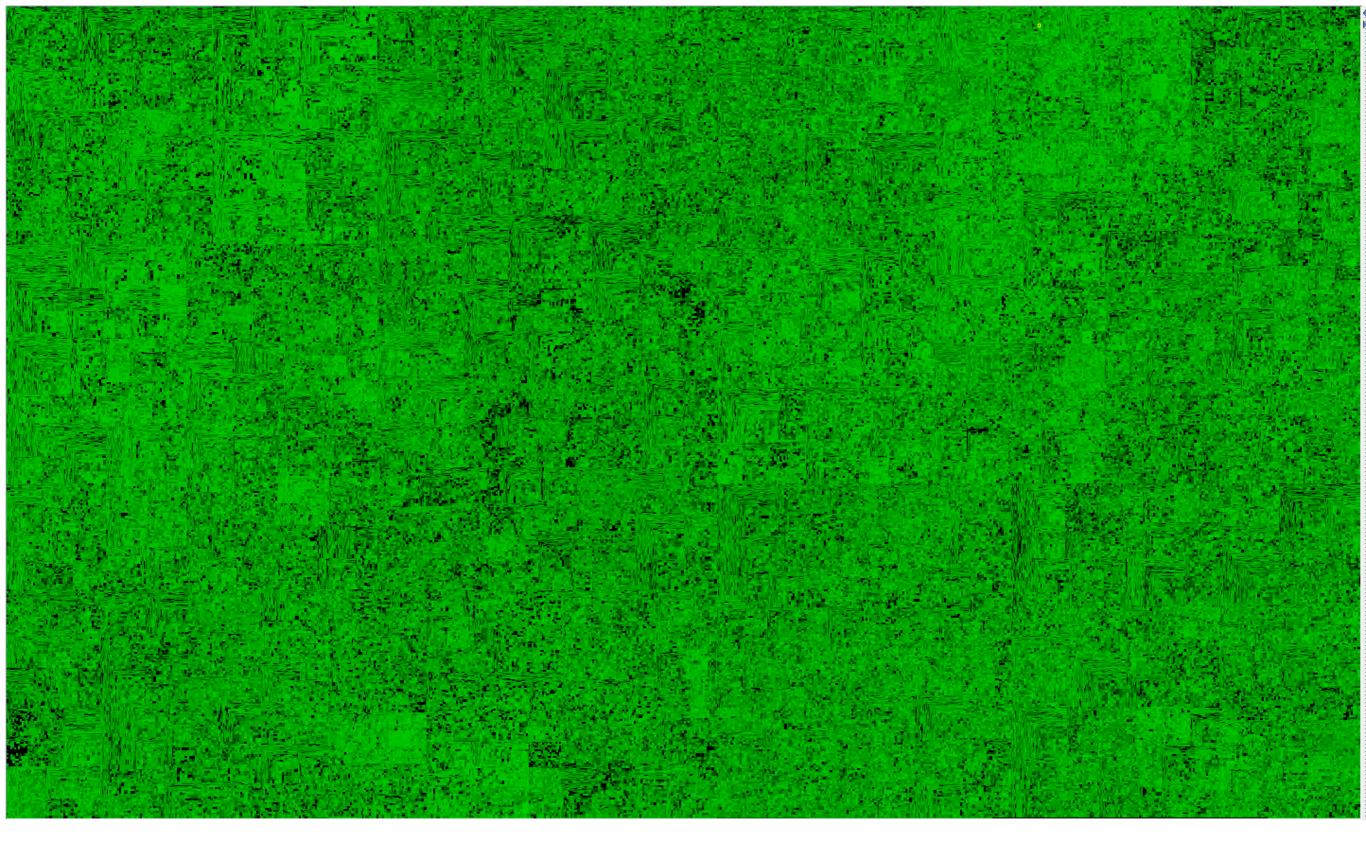
.Food and Drink	Health and Disea	ase The E	Body Biology	Leisure		Work
Animals	Plants	Textiles beo	ple Clothing  Samuel Death (Other)	Authority	Communication	Armed Hostility
Action	our Physics	Chemistry	Wholeness Quantity and	Travel	Faith	Society Dwelling  Morality Education
Space Movem	Properties of	onstitution of Matter	Relative Properties Physical	Mental	Emotion	Possession Faculty of Will
Time Existence Creation Causatio			Sensibility The Supernatural	Capacity	Language	Aesthetics Philosophy Refusal and Denial



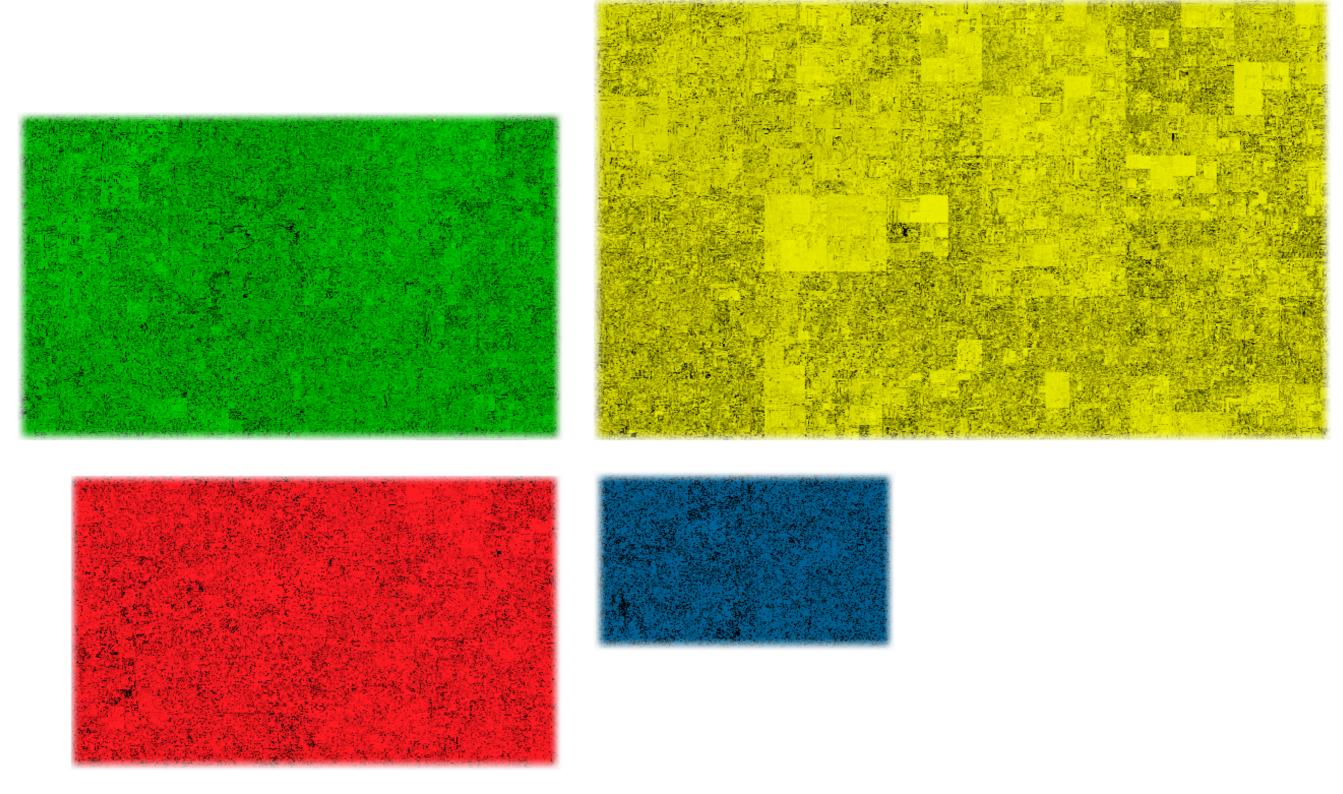
Recorded English in the Age of Geoffrey Chaucer (c.1340-1400, approximately 73,000 words) from <a href="http://historicalthesaurus.gla.ac.uk">http://historicalthesaurus.gla.ac.uk</a>



Recorded English in the Age of William Shakespeare (1564-1616, approximately 208,000 words)



Recorded English in the Age of Samuel Johnson (1709-1784, approximately 248,000 words)



Comparative Sizes of the Modern English, Johnson, Shakespeare, and Chaucer Slices

#### THE HISTORICAL THESAURUS OF ENGLISH

Home

Browse Search About the Thesaurus Scholarships and Prizes

Quick Search

Q

#### Home > Browse >

- The world
- The mind
  - 02.01 Mental capacity
  - 1 02.02 Attention and judgment
  - ♣ 02.03 Goodness and badness
  - ₽ 02.04 Emotion

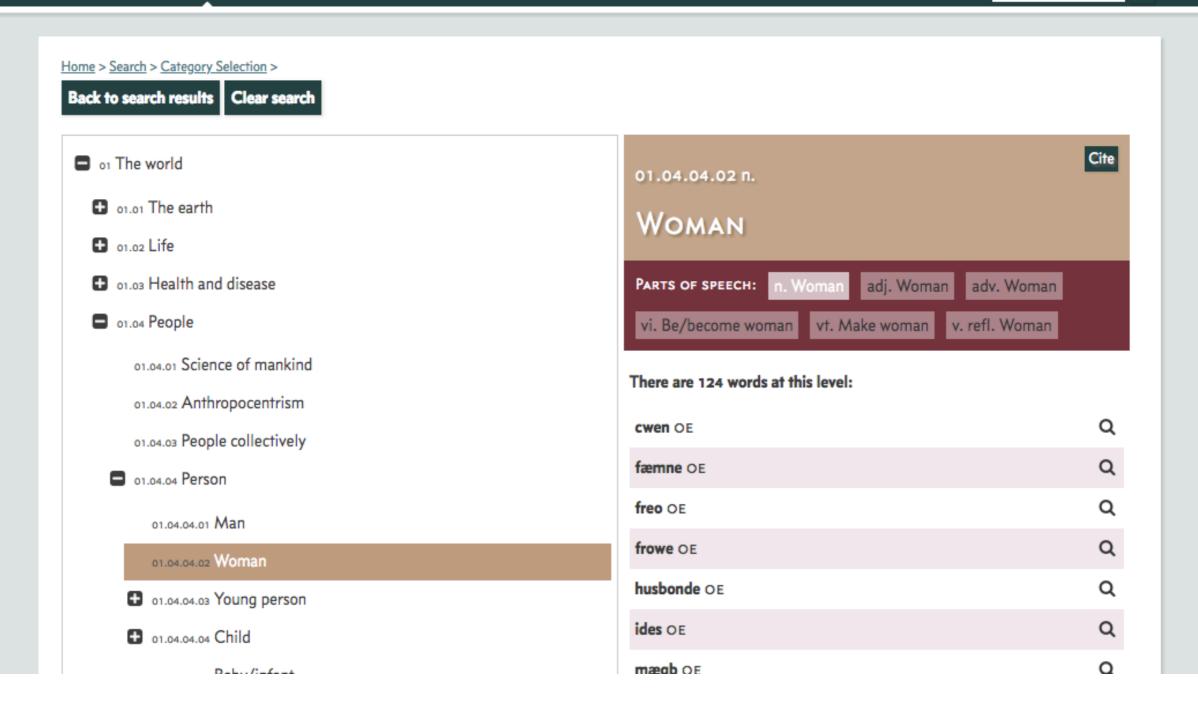
  - ⊕ 02.06 Possession
  - € 02.07 Language
- ⊕ o₃ Society

Press on a plus icon (1) to expand a category. Press on a category heading to view the parts of speech, words and subcategories in the category.

http://historicalthesaurus.arts.gla.ac.uk/

## THE HISTORICAL THESAURUS OF ENGLISH

Home Browse Search About the Thesaurus Scholarships and Prizes Quick Search Q



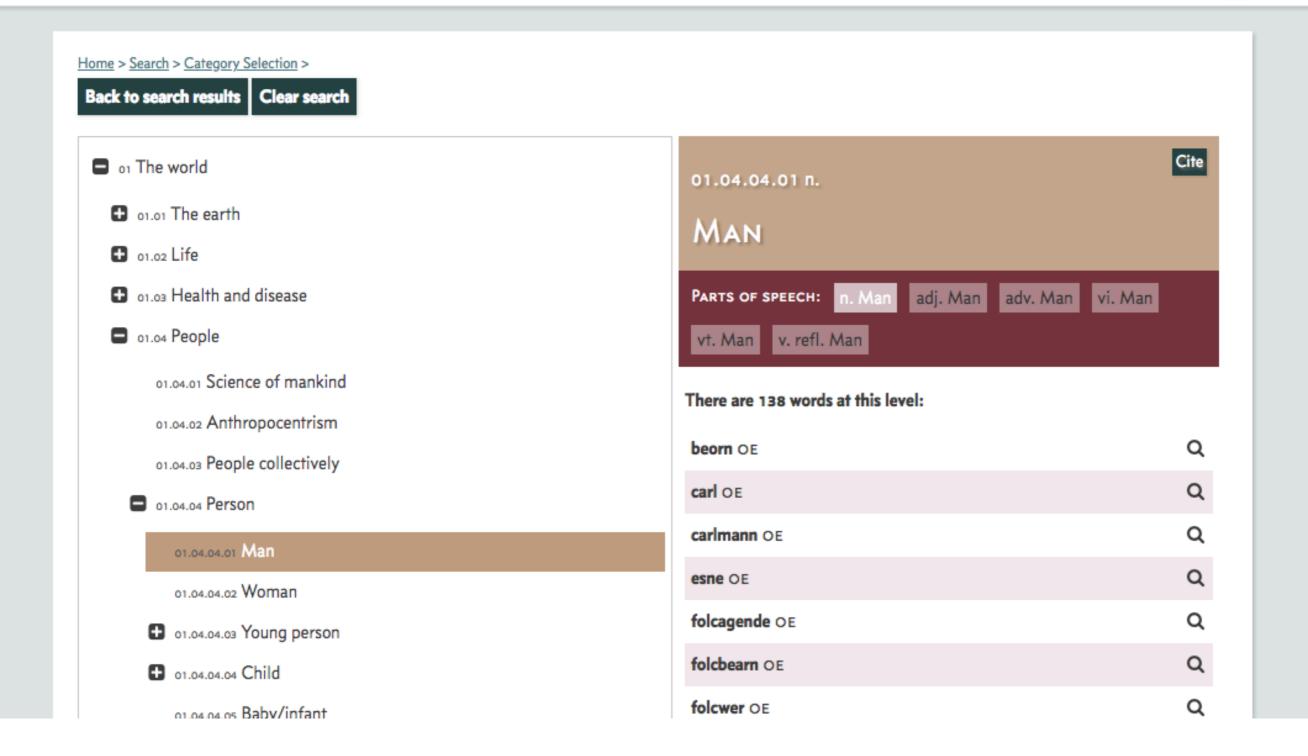
#### THE HISTORICAL THESAURUS OF ENGLISH

Home Browse

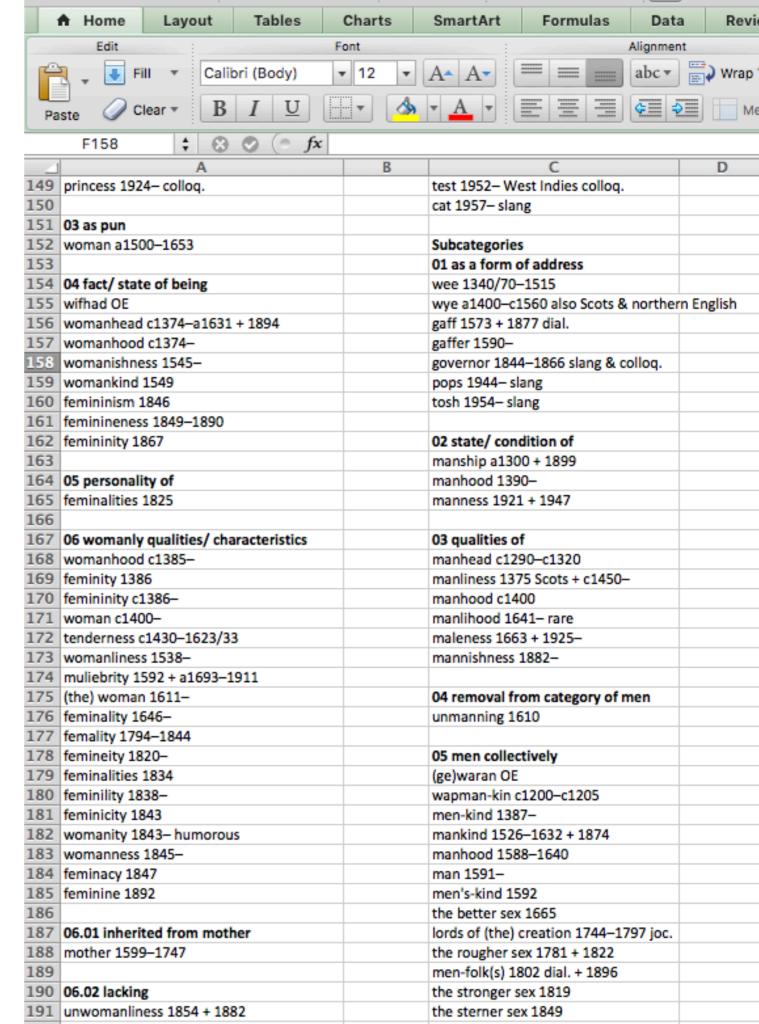
Search About the Thesaurus Scholarships and Prizes

Quick Search

Q



Harvest terms for further analysis



## now, this search apparatus is not perfect

(especially if you are used to running your own scripts on your own data)

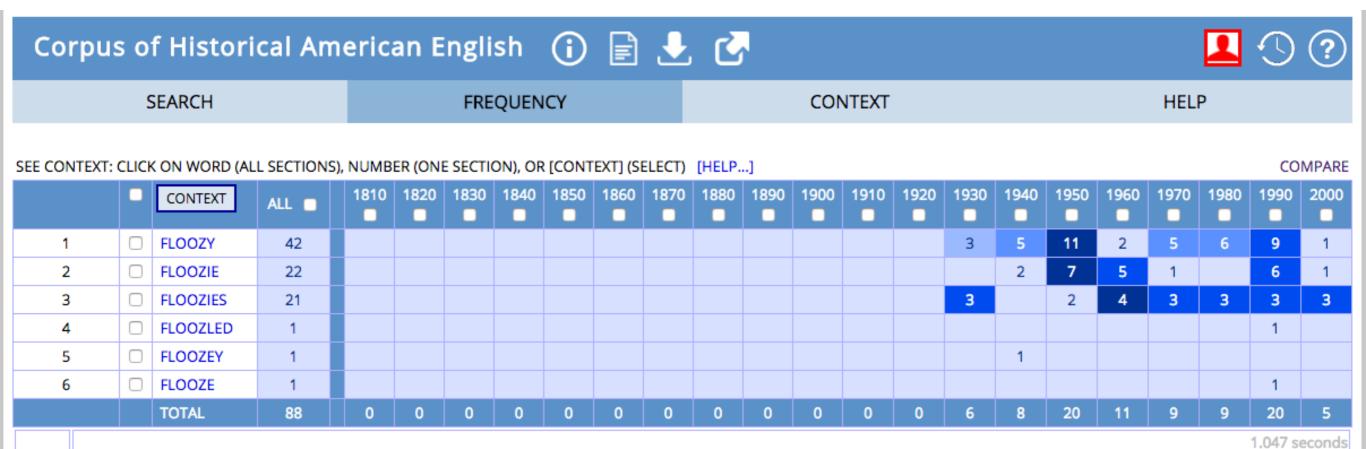
	1																						
SEE CONTEXT: CL	SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [HELP]																						
		CONTEXT	ALL	1810	1820	_	1840	1850	1860	1870	1880	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
1		WOMAN	171682	420	2383	3783	4881	5647	7170	9407	8563	11097	11848	10563	11951	10484	8965	8957	7918	9922	10906	13123	13694
2		WOMEN	127678	165	1057	1719	<u>1996</u>	2208	3653	4712	4751	7381	6949	7885	11225	8696	7755	7077	6084	7967	10411	12838	13149
3		WOMANHOOD	2156		17	41	94	154	209	223	205	198	133	312	146	70	58	49	40	57	40	48	62
4		WOMANLY	1566		14	63	58	123	211	219	170	176	104	75	59	36	23	30	42	43	39	28	53
5		WOMANKIND	318	2	7	23	7	17	32	40	33	22	26	27	19	14	9	8	8	6	8	7	3
6		WOMANISH	236	3	22	17	17	8	8	24	11	13	14	8	12	6	20	11	8	9	12	9	4
7		WOMENFOLK	168							1	3	4	4	-11	11	28	29	22	17	14	3	9	12
8		WOMANLINESS	131				1	7	8	17	15	22	13	4	16	2	4	6	3	4	1	7	1
9		WOMENS	100	1		1	1				1	5	1	6	6	7	16	6	6		21	8	14
10		WOMEN-FOLKS	87		1	4	5	4	12	6	5	11	5	9	14	10			1				
11		WOMENFOLKS	75			1			3	1	1	2		2	9	15	18	7	5	2	7	2	
12		WOMANLIKE	73			2	1	2	6	2	6	7	8	11	8	3	7	3	4	2		1	
13		WOMAN-LIKE	69			1	3	5	8	10	9	5	9	3	7	2	2	2	2	1			
14		WOMANIZER	57														2		6	6	16	10	17
15		WOMEN-FOLK	52			1		1	2	1	7	2	4	15	14	3		2					
16		WOMANIZING	51													2		1		6	17	-11	14
17		WOMANS	51		1	1	3		2	1	7	4		1	4	3	3	2	2	4	5	2	6
18		WOMANI	41													3	6	4	2	9	10	6	1
19		WOMAN-HATER	35	3	1			1	2	2	2	6	5	3	2	7	1						
20		WOMAN-SUFFRAGE	32							1	3	20	2	6									
21		WOMENKIND	27					1	1	1	6	1	5	4	4	2	2						
22		WOMAN-CHILD	24					1	1	2			12				1		2	1	1	2	1
23		WOMANLESS	24										2	1	6	2	3	2	1	4	3		
24		WOMANYOU	24													2	1	2	2	6	7	4	
25		WOMAN-	21						2		1		4		2			1	3	1	3	1	3
https://corpus.byu.ed	du/co	ha/x3.asp?xx=2&w11=women&r=4	4			3	2	2	1	2	1	1	2		1								

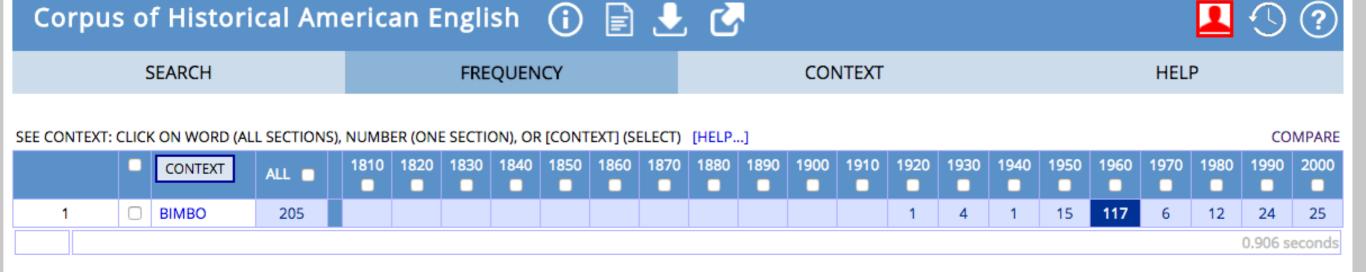
http://corpus.byu.edu/coha

### 01.04.04.02 n.

Woman

Tom 1906- slang Austral.	Q
frail 1908- slang, chiefly US	Q
bit of stuff 1909- chiefly slang	Q
floozie/floozy 1911- colloq.	Q
muff 1914- slang, orig. US	Q
babe 1915- slang, chiefly US	Q
bird 1915– slang, colloq., & derog.	Q
number 1919-	Q
rossie 1922- derog. Irish	Q
tit 1922- dial.	Q
<b>bit</b> 1923–1953 slang	Q
bimbo 1929-	Q
person 1935 contemp. + 1939 contemp.	Q
quim 1935 N. Amer. slang + 1974 N. Amer. slang	Q
scupper 1935- vulgar slang	Q
split 1935- N. Amer.	Q
rye mort 1936 slang	Q
totsy 1938- slang	Q
skinny 1941– Austral.	Q
Richard 1950- slang	Q
potato/potater 1957- Austral. slang	Q





#### (there are two uses of the word 'bimbo' visible here)

C	Corpus of Historical American English 🕠 🖹 🕹 🕜											
			SEARCH				FREQUENCY	CONTEXT	HELP			
ьв	1969	HC	Play:OspertsCareer	Α	R	L	nad come to intercede with the governmen	t to get you released from the Desert. BIMBO Mr. Rud	oit nas been overthrown: a new and liberal regime has			
69	1969	FIC	Play:OsbertsCareer	Α	В	С	Bimbo have just landed outside; the parach	ute lies beside them on the ground. Bimbo gathers it	up and Osbert detaches the ropes from his belt. Bimbo			
70	1936	FIC	TAintRight	Α	В	С	strong boxes bursting with jewels and gold.	From time to time, out of Bimbo's incalculable horde	, he passed out new fortunes to Andy and the family.			
71	1992	FIC	Play:BrotherAndres	Α	В	С	to the very pay cheque he's selling himself i	for. Counsellor Troi is a <mark>bimbo</mark> in spandex! (Pause.) p.	47 DAVE Commander Riker is not a wiener			
72	1969	FIC	Play:OsbertsCareer	Α	В	С	to keep to our beats. THE GUARD Then why	are you in here? BIMBO I'm here as an inside observ	er. The Guard lets down a rope-ladder.			
73	1989	FIC	RedPhoenix	Α	В	С	barked, " plastic's only good for two things	model airplanes and some bimbo's shopping trip. "	He remembered that his aide had smiled gamely and p			
74	1969	FIC	Play:OsbertsCareer	Α	В	С	No, buddy: not a jail an industrial training	camp. p. 28 BIMBO Industrial training camp be damr	ned! I was brought here against my will! THE			
75	1969	FIC	Play:OsbertsCareer	Α	В	С	cause of my imprisonment is plain! DR. DIS	H What do you mean? BIMBO I was trying to sell axes.	. I thought they were still chopping down trees			
76	1956	FIC	Boonisland	Α	В	С	solemnly, " It's a great big garfish. " She was	seated on Bimbo's knee, and was staring into his fac	e like a devotee before an idol			
77	1969	FIC	Play:OsbertsCareer	Α	В	С	were still chopping down trees. DR. DISH No	othing could be more ill-advised! BIMBO Well, I have s	sold French post-cards to the English German painting			
78	1994	NEWS	SanFran	Α	В	С	Fillmore in May. After that show he took a r	ide across town to sample Bimbo's charm, and booke	ed a return gig. # " I saw Bimbo's			
79	2006	NEWS	USAToday	Α	В	С	referred to as " a Bond girl. " # " It sounds li	ke a <mark>bimbo</mark> or something, " says Green, 25. # It's noth	ing like the			
80	1969	FIC	Play:OsbertsCareer	Α	В	С	(In the course of this speech, the Bear has	s been sniffing at <mark>Bimbo</mark> 's ankles, and the speaker now	w turns to the Captain.) This bear			
81	1969	FIC	Play:OsbertsCareer	Α	В	С	to rebuild! OSBERT Some people believe the	at the Government causes the explosions itself. BIMB	O Well, the cause of my imprisonment is plain! DR. DISH			
82	1969	FIC	Play:OsbertsCareer	Α	В	С	the Birds begin to squawk so loudly that he	can not make himself heard. BIMBO (to the Captain o	of the Guard). Can't those confounded birds be			
83	1969	FIC	Play:OsbertsCareer	Α	В	С	splendid wall! How high is it? THE GUARD E	ghty bisters, buddy. Bimbo pretends to admire the w	all, while the Guard is opening the door, but			
84	1999	FIC	FantasySciFi	Α	В	С	his feelings. Then I remembered the way he	e'd brushed me off for some <mark>bimbo</mark> in a motel room. I	Maybe it was the wine, the night, his			
85	1955	FIC	Harpers	Α	В	С	as, I'm sorry to say, you'll find out, my poor	Bimbo, in a very few years from now. " Tenderly he b	rushed his check			
86	1988	NEWS	Chicago	Α	В	С	be culled from Georgetown' s Charles Smith	n, Virginia Tech' s Vernell " <mark>Bimbo</mark> " Coles and UC- San	ta Barbara' s Brian Shaw. Pittsburgh' s Charles			
87	1969	FIC	Play:OsbertsCareer	Α	В	С	I prefer a comprehensive and well-administ	ered one, such as Mr. Rudolf gives us. BIMBO (vehem	ently). If humanity had wished such a fate, we should b			
88	1969	FIC	Play:OsbertsCareer	Α	В	С	: he has always had an original turn of mind	l. My father, Mr. <mark>Bimbo</mark> , when he came to succeed Mr	Rudolf, with a characteristically generous gesture,			
89	1969	FIC	Play:OsbertsCareer	Α	В	С	in a holster on his hip. There is a short viole	nt struggle of which Bimbo gets the worse. The Guard	d shoves him off and leaves. THE GUARD (			
90	1969	FIC	Play:OsbertsCareer	Α	В	С	we are given a holiday. The different grades	are used for different purposes. BIMBO And about h	ow long does it take to sift out the seventy sizes? DR			

#### Corpus of Historical American English

164 1969 FIC

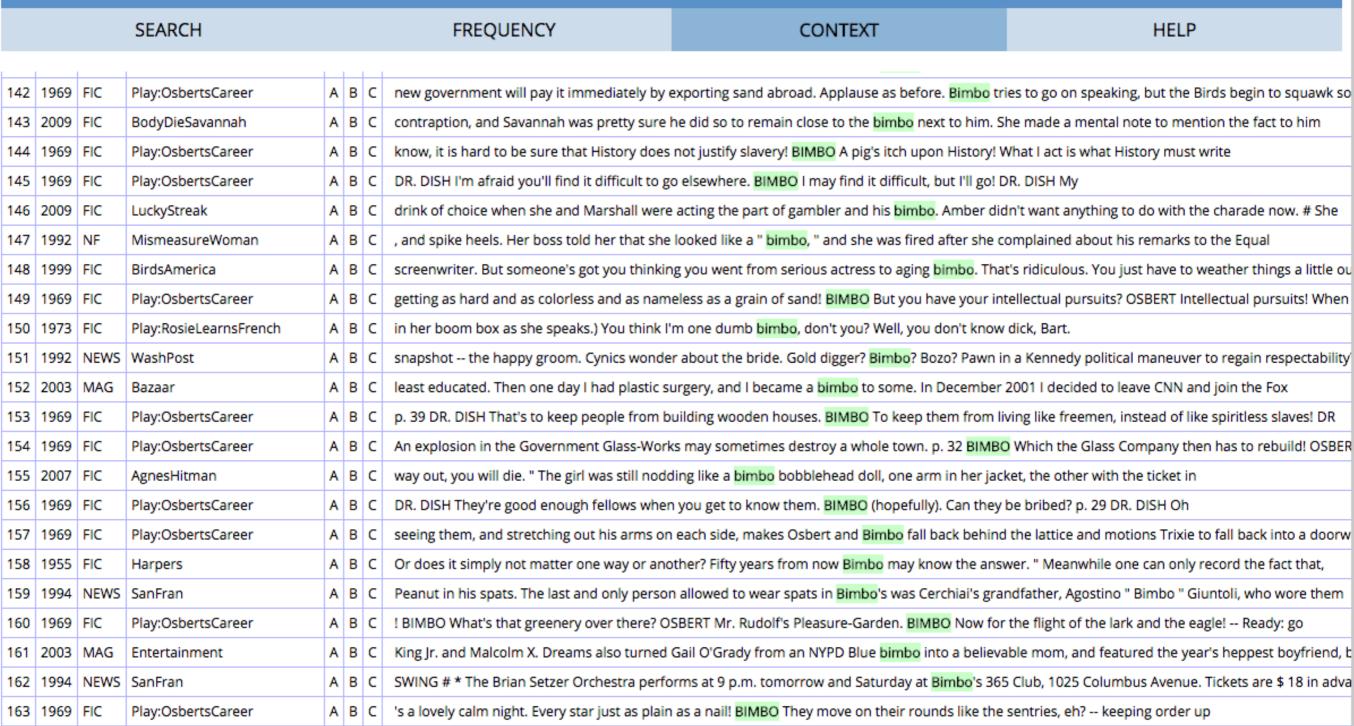
Play:OsbertsCareer











old lips were young, to kiss that pretty pout away! OSBERT (to Bimbo). Well, did you enjoy the oasis? BIMBO (sitting down in







Corpus of Historical American English 🕠 🖹 🕹 🕜											
			SEARCH			FREQUENCY	CONTEXT	CONTEXT +			
74	1557	i i c	г шулосрроосы	^	U	was a go out for my fair man pine me	IIOOZICS	of the town . So that when I was married to which			
43	1978	FIC	FoolsDie	Α	В	C bars and pick up men . That she was a middle-ago	ed floozy	but that he liked her . He had made her stop drinking			
44	1954	FIC	SweetThursday	Α	В	C . " She would n't want to congregate with no mo	re floozies	How about me?" said Doc ."			
45	1960	FIC	Mov:OceansEleven	Α	В	C every time you feel like making a grab at some ne	floozie	OCEANWhat do you plan to do about it?			
46	1960	MAG	Time	Α	В	C heroes " consist mainly of Lopez , Pardo and nightcle	lb	scarcely fazes ERRA , the Equatorian Refugee Relief			
47	1994	FIC	ContempFic	Α	В	C , " she said to herself . " You dumb	ld floozy	bitch , now look what you done . You fell down the			
48	1945	FIC	LionIsInStreets	Α	В	" Old jades ! " exclaimed Hank . " Sounds plun	floozy	They were on the gallery now and Jules opened his			
49	1991	MAG	Time	Α	В	C Midler , on the other hand , is our great show-t	oiz	and Allen personifies the anxious urban intellect . It is hard			
50	1994	FIC	HarpersMag	Α	В	c eighteen; then after the war he left her for son	ne floozy	and she had a nervous breakdown and we had to look after			
51	1993	FIC	ShiningOnes	Α	В	to be avoiding me . He acted like   was son	ne floozy	the one time I asked him downstairs for a friendly little drink			
52	1985	MAG	Newsweek	Α	В	of a redhead careering through a kooky party . A tarted-	floozy	flirting in a gondola . Dressed in lingerie or lace or in			
53	1953	FIC	Good-for-Nothing	Α	В	C GOOD-FOR-NOTHING 49older . And she 's probably some territ	<mark>le</mark> floozy	is n't she ? That 's always the case with the			
54	1953	FIC	Good-for-Nothing	Α	В	C a sweet little number who 'll make you forget about th	at floozy	in no time at all . My God , what else is			
55	1978	FIC	Whistle	Α	В	to walk about . " She left him . Is th	at	servicing every man aboard ? she thought . If so , whyshould			
56	1937	FIC	Play:BeppoBeth	Α	В	C way I'd done back in Galesburg. p. 9 BETH Th	at floozy	who finally grabbed you off would certainly have been right in			
57	1958	FIC	Play:EpitaphBluebird	Α	В	Out at the boy and exits. p. 93 MAMA LOU Th	at	AUSTRALIA Yelling offstageHey ! Hey , you , baby ! Wait			
58	1958	FIC	Play:EpitaphBluebird	Α	В	C Now maybe we 'll have a decent neighborhood again [] Th	at floozie	's turned it into a regular harlot 's paradice . I 've			
59	1951	MAG	Time	Α	В	C thriller about a middle-aged railroad watchman and the	ne floozy	he marries ( TIME , Aug. 27 ) . # The Whistle			
61	1937	FIC	Play:BeppoBeth (1)	Α	В	C incest-taboo about you and took to going out with	ne floozies	in Paris just the way I'd done back in Galesburg. p.			
62	1991	FIC	FeministStud	Α	В	C difficulty and fragmentation but also of possibility . #	FLOOZIE	IN THE JACUZZI/ The Problematics of Culture and Identity for an			

long distance to come down, do something,

like boys -- others , like their mothers or

2218

buy

THE

the

the

**FLOOZIE** 

floozie

floozies

IN THE JACUZZI # TITLE # The title is not one ,

see in movies or on TV . As I am ,

. Joe would n't stir off his backside in New York

A B C

A B C

HermanHadTwoDaughters A B C

63 1994 FIC

64 1968 FIC

65 1989 FIC

Iris

OtherSide

```
01.04.04.04.02 n.
Person :: Girl slut (1664-1884 joc.)
01.05.19.05.03.01 05 n.
Mammals :: Dog :: female slut (1845- perh. orig. US)
01.09.10.06.01 adj.
Clean/dirty:: Dirty (of persons) slut (1500/20-1638)
01.09.10.06.01 vt.
Make like dirty person slut (a1618 + 1645)
01.09.10.06.01 02 n.
Cleanness/dirtiness :: Dirty person :: woman/girl slut (1402-)
01.10.08.07.01.03 09.06 n.
Light :: Candle :: made from animal fat :: made from a rag dipped in fat slut (1609-1886)
01.10.08.07.01.03 20 n.
Light :: Candle :: guttering slut (a1864)
01.15.11 01.02 n.
Difficulty :: that which is difficult :: a difficult thing/person slut (c1460)
01.16.03.02.01.02 02 adj.
Orderly :: Untidy :: of persons slut (1500/20-1638)
01.16.03.02.01.02 04.01 n.
Order :: Untidiness :: an untidy person :: woman slut (1402-)
02.02.02.02.03 02.01 n.
Inattention :: Light-mindedness :: person :: female slut (1664-)
02.04.19.07 05.03 n.
Pride :: Impudence :: impudent person :: girl/woman slut (c1450-)
```

```
01.10.08.07.01.03 20 n.
Light :: Candle :: guttering slut (a1864)
01.15.11 01.02 n.
Difficulty :: that which is difficult :: a difficult thing/person slut (c1460)
01.16.03.02.01.02 02 adj.
Orderly:: Untidy:: of persons slut (1500/20-1638)
01.16.03.02.01.02 04.01 n.
Order :: Untidiness :: an untidy person :: woman slut (1402-)
02.02.02.02.03 02.01 n.
Inattention :: Light-mindedness :: person :: female slut (1664-)
02.04.19.07 05.03 n.
Pride :: Impudence :: impudent person :: girl/woman slut (c1450-)
03.04.09.02.01.01 01 vi.
Be under authority :: Types of servant :: act as menial slut (1829)
03.04.09.02.01.01 | 10 n.
Subjection :: Types of servant :: menial servant/drudge slut (1855)
03.04.09.02.01.02.01 | 16.11 n.
Subjection :: Domestic servant :: maid/housemaid :: kitchen-maid slut (c1450)
03.06.05.07.02 03 vi.
Be unchaste/sexually indulgent :: of woman: make herself available slut (about) (1829-)
03.06.05.07.02 03.03.01 n.
Moral evil :: Unchastity :: unchaste behaviour of woman :: sexual indulgence :: unchaste/loose woman slut (c1450-)
03.11.04 14 vi.
Carry on an occupation/work :: work in other specific ways/conditions slut (about) (1829)
```

1. we need historically-sensitive analyses of large scale data

we need historically-sensitive analyses of large scale data
 assuming the computer can do it is dangerous

- 1. we need historically-sensitive analyses of large scale data
  - 2. assuming the computer can do it is dangerous
    - 3. there are resources to help us do (1) and (2)

## thank you!

@heatherfro // hfroehli.ch