Mark Davies
Brigham Young University

The advantages of using relational databases for large historical corpora

Workshop on Historical Text Mining / U Lancaster / July 2006

---

1. VIEW: Variation in English Words and Phrases (http://view.byu.edu)

1.1 Original format:
<w AT0>the <w NN1>reluctance <w PRF>of <w AJ0>central <w NN1>government
<w TO0>to <w VVI>take <w AVP>on <w DT0>any <w AJ0>new <w NN1>expenditure

1.2 Convert to "vertical format" and add delimiters
AT0^the
NN1^reluctance
PRF^of
AJ0^central
NN1^government
TO0^to

Table 1. Import into MS SQL Server ; run SQL query to add adjacent words

ID	word1	pos1	word2	pos2	word3	pos3	word4	pos4	word5	pos5	word6	pos6	word7	pos7
1463998	the	AT0	reluctance	NN1	of	PRF	central	AJ0	government	NN1	to	TO0	take	VVI
1463999	reluctance	NN1	of	PRF	central	AJ0	government	NN1	to	TO0	take	VVI	on	AVP
1464000	of	PRF	central	AJ0	government	NN1	to	TO0	take	VVI	on	AVP	any	DT0
1464001	central	AJ0	government	NN1	to	TO0	take	VVI	on	AVP	any	DT0	new	AJ0
1464002	government	NN1	to	TO0	take	VVI	on	AVP	any	DT0	new	AJ0	expenditure	NN1
1464003	to	TO0	take	VVI	on	AVP	any	DT0	new	AJ0	expenditure	NN1	makes	VVZ
1464004	take	VVI	on	AVP	any	DT0	new	AJ0	expenditure	NN1	makes	VVZ	it	PNP
1464005	on	AVP	any	DT0	new	AJ0	expenditure	NN1	makes	VVZ	it	PNP	necessary	AJ0
1464006	any	DT0	new	AJ0	expenditure	NN1	makes	VVZ	it	PNP	necessary	AJ0	to	TO0
1464007	new	AJ0	expenditure	NN1	makes	VVZ	it	PNP	necessary	AJ0	to	TO0	seek	VVI
1464008	expenditure	NN1	makes	VVZ	it	PNP	necessary	AJ0	to	TO0	seek	VVI	funds	NN2

Table 1. Basic query: break * [n*]

DISTRIB	WORD/PHRASE	TOKENS REG1	PER MIL IN REG1 [100,000,000 WORDS]
1	BREAK THE LAW	119	1.19
2	BREAK THE NEWS	58	0.58
3	BREAK THE DEADLOCK	41	0.41
4	BREAK THE RULES	31	0.31
5	BREAK THE ICE	29	0.29
6	BREAK THE SILENCE	28	0.28
7	BREAK WITH TRADITION	28	0.28
8	BREAK THE HABIT	24	0.24

Table 2. By register (e.g. lexical verbs in [law texts] (W_ac_polit_law_edu)

DISTRIB	WORD/PHRASE	TOKENS REG1	PER MIL IN REG1 [4,640,346 WORDS]
1	MAKE	3133	675.17
2	TAKE	2276	490.48
3	GIVE	1792	386.18
4	PROVIDE	1347	290.28
5	APPLY	1278	275.41
6	SEE	1212	261.19
7	USE	1137	245.02
8	CONSIDER	1041	224.34

Table 3. Specific to register (e.g. compare to other ACAD) (x ≥ 2 in both)

DISTRIB	WORD/PHRASE	TOKENS REG1	TOKENS REG2	PER MIL IN REG1 [4,640,346 WORDS]	PER MIL IN REG2 [10,789,236 WORDS]	REG 1-2 RATIO
1	SUE	331	10	71.33	0.93	76.96
2	CERTIFY	27	2	5.82	0.19	31.39
3	ADJOURN	26	2	5.60	0.19	30.23
4	NOTIFY	38	3	8.19	0.28	29.45
5	WAIVE	38	3	8.19	0.28	29.45
6	DISCLOSE	190	16	40.95	1.48	27.61
7	OVERRULE	23	2	4.96	0.19	26.74
8	PROHIBIT	53	5	11.42	0.46	24.65
9	PLEAD	62	6	13.36	0.56	24.03
10	RESCIND	29	3	6.25	0.28	22.48

Table 4. See relative frequency in all 70 registers (*plead*)

#	REGISTER NAME	# PER MILLION	# TOKENS	# WORDS
1	**S_courtroom**	**102.0**	**13**	**127474**
2	W_fict_drama	43.7	2	45757
3	W_letters_prof	15.1	1	66031
4	W_ac_polit_law_edu	14.2	66	4640346
5	W_fict_poetry	13.5	3	222451
6	W_hansard	11.2	13	1156171
7	W_news_script	10.8	14	1292156
8	S_parliament	10.4	1	96239
9	W_email	9.4	2	213045
10	W_newsp_other_arts	8.4	2	239258
11	S_consult	7.2	1	138011
12	W_newsp_brdsht_nat_misc	5.8	6	1032943
13	W_newsp_brdsht_nat_arts	5.7	2	351811
14	W_newsp_tabloid	5.5	4	728413
15	W_religion	5.3	6	1121632

Table 5. Lexical bundles (ACAD vs FICT)

DISTRIB	WORD/PHRASE	TOKENS REG1	TOKENS REG2	PER MIL IN REG1 [15,429,582 WORDS]	PER MIL IN REG2 [16,194,885 WORDS]	REG 1-2 RATIO
1	THE USE OF	3179	170	206.03	10.50	19.63
2	AS A RESULT	2063	143	133.70	8.83	15.14
3	IN WHICH THE	2589	181	167.79	11.18	15.01
4	THE FORM OF	1487	116	96.37	7.16	13.45
5	THE NUMBER OF	2701	213	175.05	13.15	13.31
6	THE UNITED STATES	1768	140	114.59	8.64	13.25
7	THE EFFECT OF	1662	133	107.72	8.21	13.12
8	THAT THERE IS	1635	137	105.97	8.46	12.53
9	CAN NOT BE	2858	242	185.23	14.94	12.40

Table 6. Collocates: [*kitchen*] within 10 words of a noun (by frequency)

	WORD	# TIMES NEARBY	TOTAL IN CORPUS	% NEARBY
1	DOOR	518	23207	2.2%
2	ROOM	460	28443	1.6%
3	TABLE	425	18804	2.3%
4	HOUSE	261	47087	0.6%
5	BATHROOM	205	2323	8.8%

Table 7. [*kitchen*] within 10 words of a noun (by percentage / modified Z-score)

	WORD	# TIMES NEARBY	TOTAL IN CORPUS	% NEARBY
1	DRAINER	12	32	37.5%
2	SINK	92	285	32.3%
3	SCULLERY	32	176	18.2%
4	WORKTOP	13	75	17.3%
5	UTENSILS	22	127	17.3%
6	LIVING-ROOM	11	69	15.9%
7	PANTRY	18	155	11.6%

Table 8. CHARTS: frequency of [nn*] [nn*] – six "macro" registers

REGISTER	SPOKEN 4887	FICTION 6076	NEWS 11362	ACADEMIC 16485	NON-ACAD 18960	MISC 34146
	472.9	375.2	1,068.1	1,068.4	1,139.8	1,202.6

Table 9. CHARTS: Frequency of [nn*] [nn*] – sub-registers of SPOKEN

	DIS_	DOC_	NWS	CLS_	CST_	CNV_	CRT_	DEM_	INT_	ORH_	COM_	HUM_	NAT_	POL_	SOC_	MTG_	PRL_	DEB_	SRM_	SP+_	SP-_	SPO_	TUT_	UNC
	515	650	689	291	312	305	604	283	509	450	1,390	118	1,235	413	375	744	873	959	109	934	660	270	615	486

Table 10. Grouping by synonyms: {sheer/utter/absolute} [nn*]

	%	+	-	SHEER
1	1.00	60	--	weight
2	1.00	31	--	force
3	1.00	26	--	luck
4	1.00	23	--	quantity
5	1.00	13	--	cliff
6	1.00	12	--	cliffs
7	1.00	11	--	coincidence
8	1.00	10	--	enjoyment

	%	+	-	UTTER
1	1.00	19	--	confusion
2	1.00	5	--	condemnation
3	1.00	5	--	devastation
4	1.00	5	--	disregard
5	1.00	5	--	helplessness
6	1.00	4	--	loneliness
7	1.00	3	--	dejection
8	1.00	3	--	ruthlessness

	%	+	-	ABSOLUTE
1	1.00	98	--	majority
2	1.00	84	--	terms
3	1.00	54	--	zero
4	1.00	53	--	minimum
5	1.00	51	--	value
6	1.00	39	--	egalitarianism
7	1.00	33	--	right
8	1.00	27	--	price

Table 11. Comparing collocates: *man/woman* + ADJ

	MAN +	# MAN	# WOMAN	% MAN
2	UNITED	48	0	100.0%
3	MACHO	36	0	100.0%
4	BURLY	30	0	100.0%
5	SELF-MADE	29	0	100.0%
6	ARMED	26	0	100.0%
10	CIVILISED	19	0	100.0%
11	MUSCULAR	19	0	100.0%
12	SCIENTIFIC	17	0	100.0%
13	STEADY	16	0	100.0%
16	TALENTED	14	0	100.0%
21	CIVILIZED	12	0	100.0%

	WOMAN +	# WOMAN	# MAN	% WOMAN
1	CLEANING	28	0	100.0%
2	DUMPY	23	0	100.0%
4	LIBERATED	18	0	100.0%
5	MOTHERLY	17	0	100.0%
11	BUXOM	11	0	100.0%
12	RAPED	10	0	100.0%
13	VAGINAL	8	0	100.0%
20	ABUSED	6	0	100.0%
23	BOSSY	6	0	100.0%
24	CLINGING	6	0	100.0%
25	COMPLAINING	6	0	100.0%

Table 12. Comparing collocates: *chair* in FICTION and ACADEMIC

	WORD	# REG 1	# REG 2	% REG 1
1	HANDS	84	0	100.0%
2	LEGS	62	0	100.0%
3	FIRE	59	0	100.0%
4	FEET	58	0	100.0%
5	KITCHEN	55	0	100.0%
6	FACE	50	0	100.0%
7	FLOOR	47	0	100.0%
8	WOMAN	31	0	100.0%
9	END	29	0	100.0%
10	LEATHER	29	0	100.0%
11	COFFEE	28	0	100.0%
12	JACKET	28	0	100.0%

	WORD	# REG 2	# REG 1	% REG 2
1	COMMITTEE	18	0	100.0%
2	FIELD	17	0	100.0%
3	MEMBERSHIP	7	0	100.0%
4	PHILOSOPHY	7	0	100.0%
5	CIRCLE	4	0	100.0%
6	CO-ORDINATOR	4	0	100.0%
7	REFERENCE	4	0	100.0%
8	ROLE	4	0	100.0%
9	MATHEMATICS	3	0	100.0%
10	FRENCH	2	0	100.0%
11	ADMISSIONS	2	0	100.0%
12	INJURIES	2	0	100.0%

Table 13. WordNet queries

SYNONYM:	MORE GENERAL:	MORE SPECIFIC:	PART OF:	CONTAINS PART:
[=x]	[>x]	[<x]	[@x]	[&x]
[=small]	[>shriek]	[<woman]	[@house]	[&leg]
[=scream].[v*]	[>wallop].[v*]	[<hit].[v*]	[@body]	[&shelf]

Table 14: [=beat].[v*] the [nn*]

	%	+	-	BEAT
1	1.00	16	--	system
2	1.00	14	--	eggs
3	1.00	14	--	world
4	1.00	11	--	recession
5	1.00	10	--	egg
6	1.00	7	--	hell
7	1.00	6	--	clock
8	1.00	6	--	cream
9	1.00	6	--	shit
10	1.00	6	--	traffic

	%	+	-	CRUSH
1	1.00	4	--	rebellion
2	1.00	2	--	biscuits
3	1.00	2	--	coup
4	1.00	2	--	imagination
5	1.00	1	--	amaretti
6	1.00	1	--	bottle
7	1.00	1	--	cane
8	1.00	1	--	career
9	1.00	1	--	cockatrice
10	1.00	1	--	corn

	%	+	-	VANQUISH
1	1.00	1	--	alien
2	1.00	1	--	romans
3	0.50	1	1	past

Table 15: More specific words for walk as verb: [<walk].[v*]

18	STROLL	485	4.85
19	LIMP	466	4.66
20	TREAD	445	4.45
33	STALK	224	2.24
35	LURCH	201	2.01
36	STUMBLE	193	1.93
39	TIPTOE	144	1.44
41	STAGGER	128	1.28
47	STRUT	95	0.95
49	SWAGGER	77	0.77
54	PROWL	60	0.60
55	TRUDGE	60	0.60
62	SAUNTER	34	0.34
67	TRAIPSE	18	0.18
68	PRANCE	18	0.18