Encyclopaedia of Shakespeare's Language

Lancaster University

Text Hackathon: Extracting Knowledge from Big Digital Texts

(Centre for Textual Studies, De Montfort University, 10-12th November 2017)

From simple word counts to collocates and keywords

Jonathan Culpeper,

Lancaster University, UK

@ShakespeareLang

http://wp.lancs.ac.uk/shakespearelang

Arts & Humanities Research Council

CASS
Corpus Approaches to Social Science

The Queen's Anniversary Prizes
For Higher and Further Education
2015

UCREL

Encyclopaedia of Shakespeare's Language

Lancaster University

Text Hackathon: Extracting Knowledge from Big Digital Texts

(Centre for Textual Studies, De Montfort University, 10-12th November 2017)

Unlocking the meanings of words and the styles they create using corpus-based techniques

Jonathan Culpeper,

Lancaster University, UK

@ShakespeareLang

http://wp.lancs.ac.uk/shakespearelang

Arts & Humanities Research Council

CASS Corpus Approaches to Social Science

The Queen's Anniversary Prizes For Higher and Further Education 2015

UCREL

Overview

1. Counting words

2. Meanings and styles through:
   - ➢ Frequencies of words
   - ➢ Frequencies of word clusters (n-grams)
   - ➢ Concordances and collocates (statistically associated co-words)
   - ➢ Keywords (statistically distinctive words)

3. A note on programs I used, etc. (see handout)

Why bother to count linguistic items?

It's all about patterns:

- Patterns of language usage shape meanings, styles, cultures, etc.

Counting can:

- Reveal patterns you didn't know
- Confirm patterns you did had a hunch about

Counting also has the merit that:

- It does not rely on intuition
- It's relatively precise

Why use computers for counting?

Obvious advantages:

- They can count up more stuff than you could in several lifetimes
- They are systematic

Not so obvious disadvantages:

- Getting them to count even ‘simple’ words is not straightforward
- Different programs (with the same settings) will often give you different counts of the same thing
- Mistakes can lurk within the counts

And humans are never redundant:

- You decide the what – what data and what to count
- And you interpret what the results mean

What to count with a computer?

WORDS, WORDS, WORDS

Why words?

- Words carry a fairly large part of the meanings we wish to convey
- Words, especially some, carry at least part of the grammar of the language
- Words are a major part of styles (not just authorial)
- Words are many (difficult for a human to count in extensive data)
- Words pattern (cf. word choice)

Words

So, with words,

we are on to a winner!?

The word: Not so simple

Different words in Shakespeare: What can we 'learn' from the internet?

- In his collected writings, Shakespeare used **31,534** different words. (A misinterpretation of Efron and Thisted 1976; https://statistics.stanford.edu/sites/default/files/BIO%2009.pdf)
- Literary elites love to rep Shakespeare's vocabulary: across his entire corpus, he uses **28,829** words (https://pudding.cool/2017/02/vocabulary/)
- Unique words: There are **27,352** distinct spellings in Shakespeare (http://wordhoard.northwestern.edu/userman/scripting-example.html)
- Around **20,000** (David Crystal, and others)

Of course there is also the major issue of what counts as "Shakespeare"!!!

Do we count word-forms or lexemes?

Word-forms and lexemes (lemmas -- dictionary headword)

- Dictionary headword/lemma:

do

- Modern (morphological) word-forms:

do, does, doing, did, done

- Early modern (morphological) word-forms:

do, does, do(e)st, doth, doing, did, didst, done

Do we count word-forms or lexemes?

Word-forms and lexemes

Dictionary headword/lemma:
do **= 1**

Modern (morphological) word-forms:
do, does, doing, did, done **= 5**

Early modern (morphological) word-forms:
do, does, do(e)st, doth, doing, did, didst, done **= 8**

The word: Not so simple

Other problems with counting words

a) Can we simply adopt an orthographic definition of a word?

b) Would we want to include all such words?

c) Are the different ways of spelling words an issue?

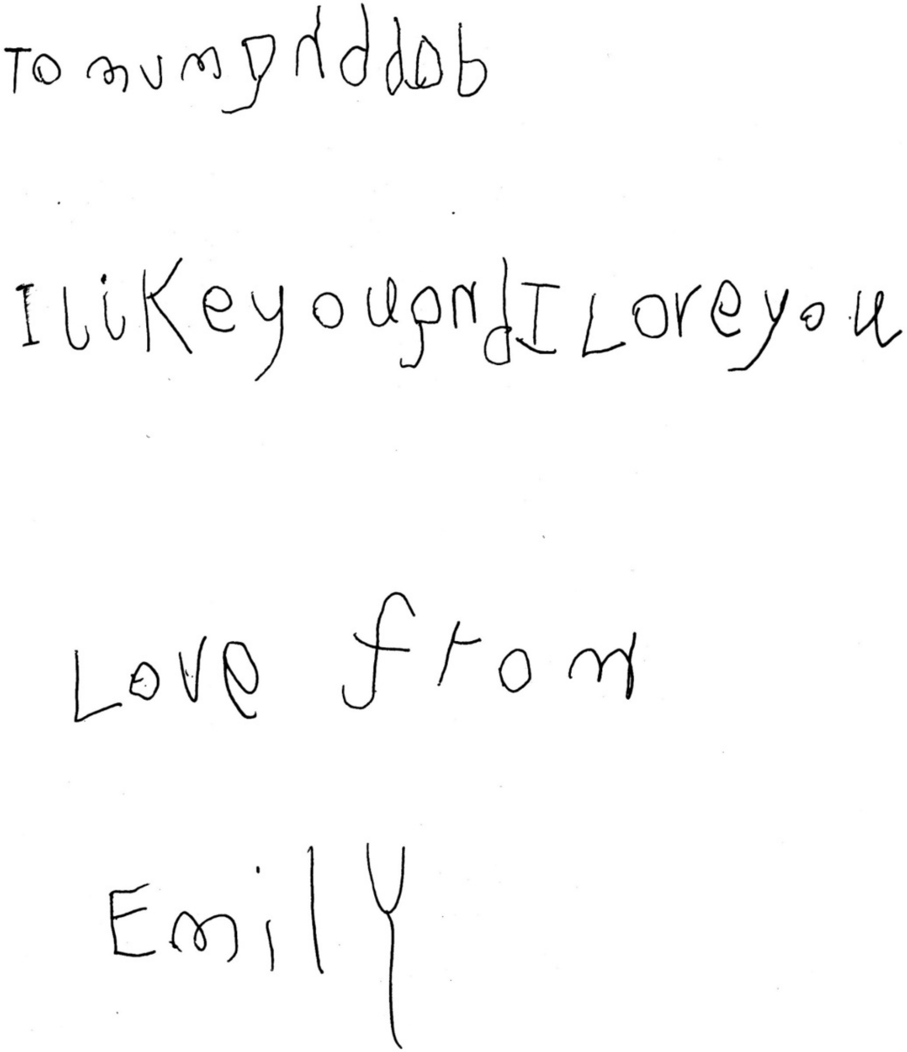
d) Are the words accurately transcribed in the first place?

The word: Apply the orthographic definition?

The usual way of defining a word in corpus linguistics:

orthographic word = 'a string of uninterrupted non-punctuation characters with white space or punctuation at each end' (Leech et al. 2001: 13-14)

The word: Apply the orthographic definition?



To mumpnddab

IliKeyoupndILoreyou

Love from

Emily

The word: Apply the orthographic definition?

Interference from other ways of defining words:

- Words in speech transposed to writing

Tybalt: Gentlemen, **good den**, a word with one of you.

Romeo and Juliet, III.1

The word: Apply the orthographic definition?

- Words as independent units of meaning

➢ *The plane landed* = 3 words?

➢ *The plane took off* = 3 words? (cf. phrasal verbs)

➢ *He kicked the bucket* = 2 words? (cf. idioms)

Compounds:

- *my self*, *well come*, etc.
- *hourglass / hour-glass / hour glass*

Contractions:

Present-day *gonna* < *going to* (BNC "gon-na");

Also: *can't*, *I'm*, *we'll*, etc.

The word: Do we include all words?

What about:

- Proper nouns
- Onomatopoeic words and noises: *Do de do de* (*King Lear*, 3.6)
- Errors: *aud* for *and*
- Malapropisms: [Quickly] *She's as fartuous a civil modest wife* (*Merry Wives* 2.2)
- 'Foreign words': *Monsieur*

The word: Are different ways of spelling words an issue?

You decide to study the use of the word *would* in a corpus. You type it into your search program … and look at the result.

But in historical texts you miss:

wold, *wolde*, *woolde*, *wuld*, *wulde*, *wud*, *wald*, *vvould*, *vvold*, etc., etc.

One orthographic word today; many in EModE ….

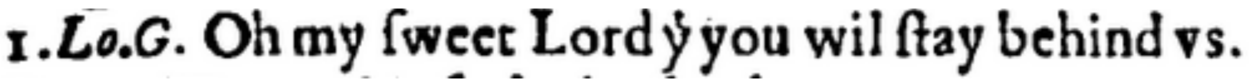
a **huge** problem!

Spelling is still an issue today.

The word: Are the words accurately transcribed?

Accuracy is problem for transcriptions of spoken data and historical texts.

- Manual transcriptions are error prone and costly.
- Double-keying is super-costly.
- For spoken data, voice-recognition programs are very limited.
- For historical data, OCR only works up to a point (see work by Amelia Joulain-Jay). For example, one particular problem is the long ‘s’, which resembles an ‘f’.



1.*Lo.*G. Oh my ſweet Lord yͭ you wil ſtay behind vs.

<u norm="1 Lord" label="1. Lo. G"> Oh my sweet Lord **CyC** you , wil stay behind vs.</u>

(Partial) Solutions?

Tokenization processing – to segment a text into orthographic words, deal with compounds and contractions, etc.

Spelling regularisation processing – to group spelling variants under word-forms (cf. VARD)

Lemmatization processing – to group word-forms under lemmas (‘headwords’)

No perfect solution.

Meanings and styles: Frequencies of words

- Are the words of Christina Aguilera's song *Beautiful* typical of pop song lyrics?

I am beautiful no matter what they say
Words can't bring me down
I am beautiful in every single way
Yes words can't bring me down, Oh no
So don't you bring me down today

- Need to characterize the style of pop song lyrics.
- Word **frequencies** – create a "word list" of pop song lyrics and compare with other genres.

Meanings and styles: Frequencies of words

Pop song lyrics	*An academic paper*	*Spoken English*	*Written English*
I	The	The	The
You	Of	I	Of
Me	And	You	And
And	In	And	A
The	To	It	In
My	A	A	To (inf.)
To	Is	‘s	Is
Is	That	to	To (prep.)
All	Language	of	Was
I’m	It	That	It

Meanings and styles: Frequencies of words

Content words vs. grammatical/function words

I **am beautiful** no **matter** what they **say**
Words can't **bring** me **down**
I **am beautiful** in every **single way**
Yes **words** can't **bring** me **down**, Oh no
So don't you **bring** me **down today**

Meanings and styles: Frequencies of words

Pop song lyrics	love, make, life, boyfriend, baby, know, need, down, come, time, said, goes, say, alone, end, look, ride, sad, bring, feel, feeling, rain, right, things
Academic writing	language, speech, writing, spoken, written, historical, communicative, types, example, English, text, features, texts, functions, medium, registers, linguistics, register, time, see, functional, interaction, Saussure, words, area

Meanings and styles: Frequencies of words

Simple frequencies of words in (relatively) big data -- **distribution**

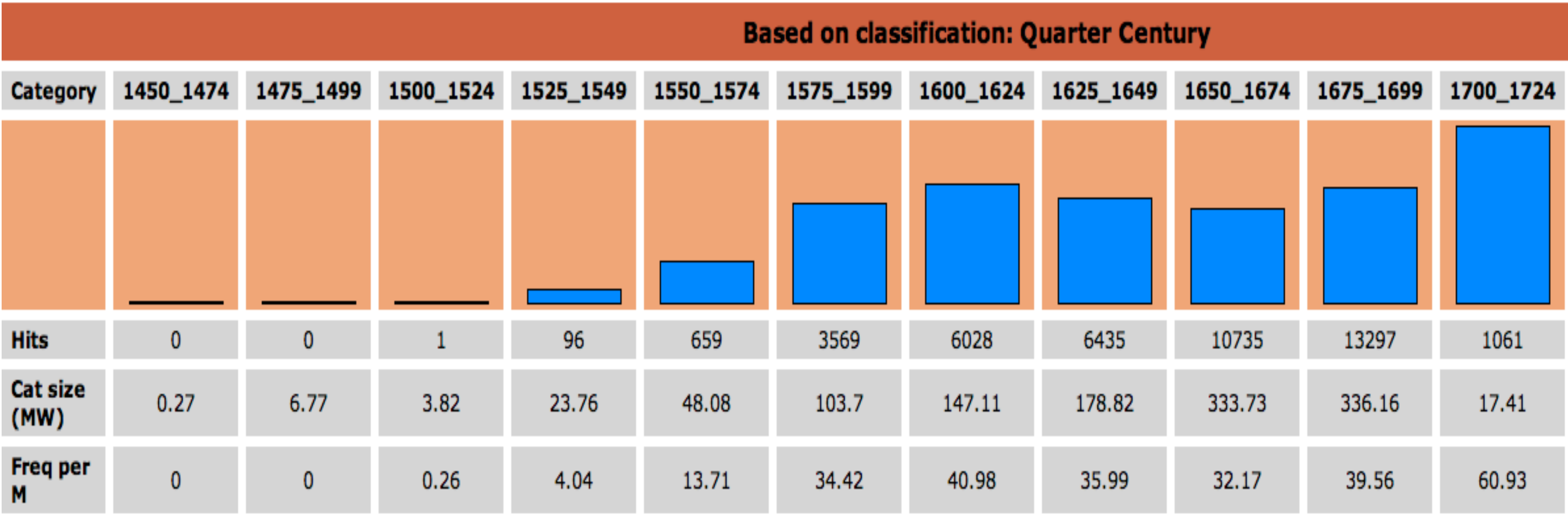
Two examples:

- Did the three Italian conduct or etiquette manuals published in English between 1561 and 1581 have much of an impact?

  Early English Books Online (EEBO-TCP) interrogated through CQPweb

Meanings and styles: Frequencies of words

- The frequencies of the word *manners*, 1450-1724



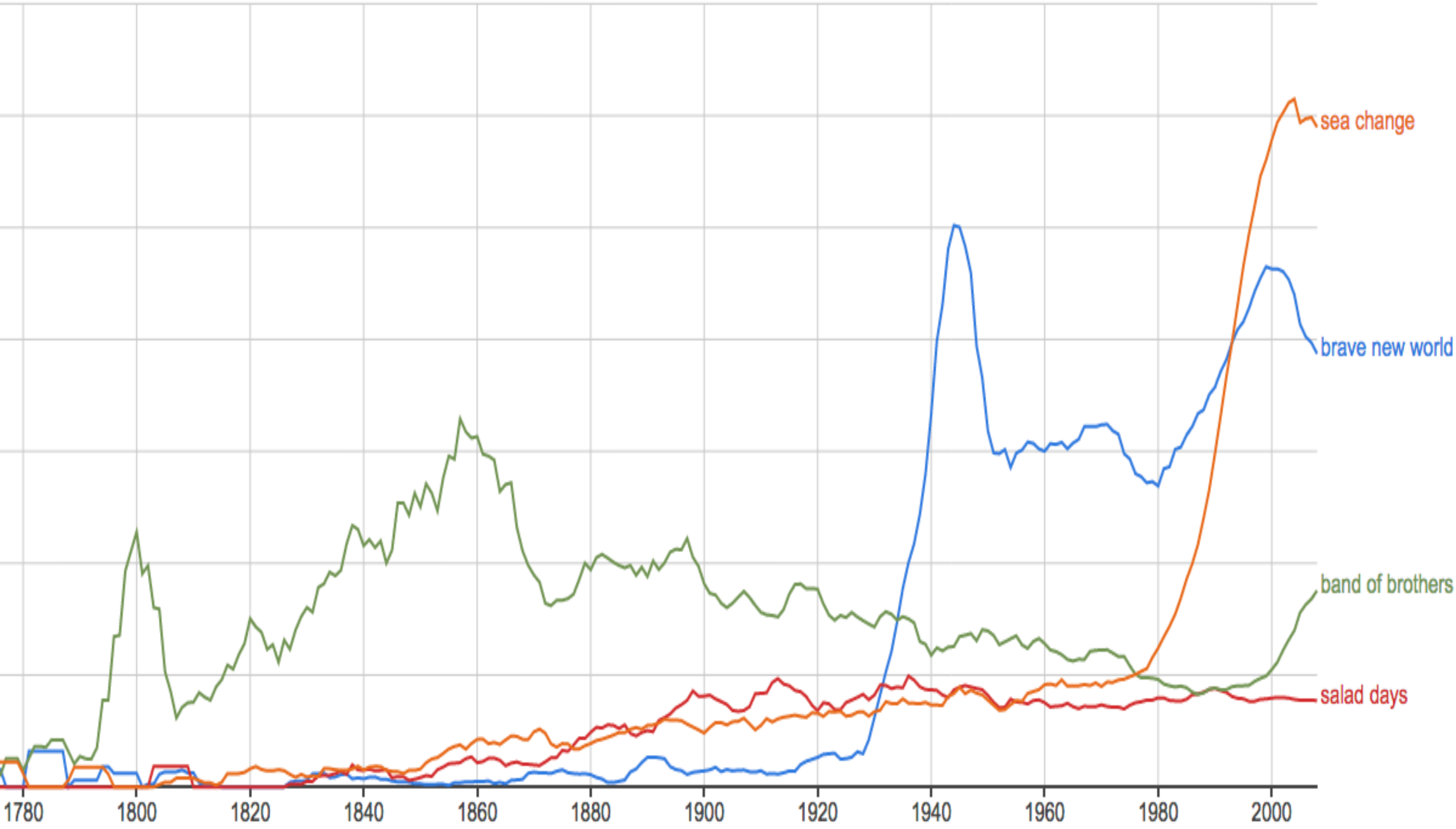
	Based on classification: Quarter Century										
Category	1450_1474	1475_1499	1500_1524	1525_1549	1550_1574	1575_1599	1600_1624	1625_1649	1650_1674	1675_1699	1700_1724
Hits	0	0	1	96	659	3569	6028	6435	10735	13297	1061
Cat size (MW)	0.27	6.77	3.82	23.76	48.08	103.7	147.11	178.82	333.73	336.16	17.41
Freq per M	0	0	0.26	4.04	13.71	34.42	40.98	35.99	32.17	39.56	60.93

Meanings and styles: Frequencies of words

- What happened to phrases associated with Shakespeare in subsequent phases of the development of English?

  Google books interrogated through Google's N-gram Viewer

Four phrases associated with Shakespeare and their use in printed material over the last 200 years (Google's N-Gram Viewer)



Meanings and styles: Frequencies of word clusters (n-grams)

Maybe the key to styles is certain **clusters** of words?

- Authorship attribution. E.g. The contribution made by other authors to “Shakespeare’s works”, and vice versa. Cf. Gary Taylor & Gabriel Egan (2016). *The New Oxford Shakespeare.* Christopher Marlowe credited as co-author of *Henry VI* plays, Thomas Middleton as co-author of *All’s Well That Ends Well*; *Arden of Faversham* added to Shakespeare’s 'çanon’.
- But also a means of characterizing all kinds of styles. E.g. work by Michaela Mahlberg.
- How do we identify the clusters, what are they anyway?

Meanings and styles: Frequencies of word clusters (n-grams)

I will finish this presentation shortly

I will

will finish

finish this

this presentation

presentation shortly **= 5 unique n-grams (5 types; 1 token each)**

I will finish

will finish this

finish this presentation

this presentation shortly **= 4 unique n-grams (4 types; 1 token each)**

Meanings and styles: Frequencies of word clusters (n-grams)

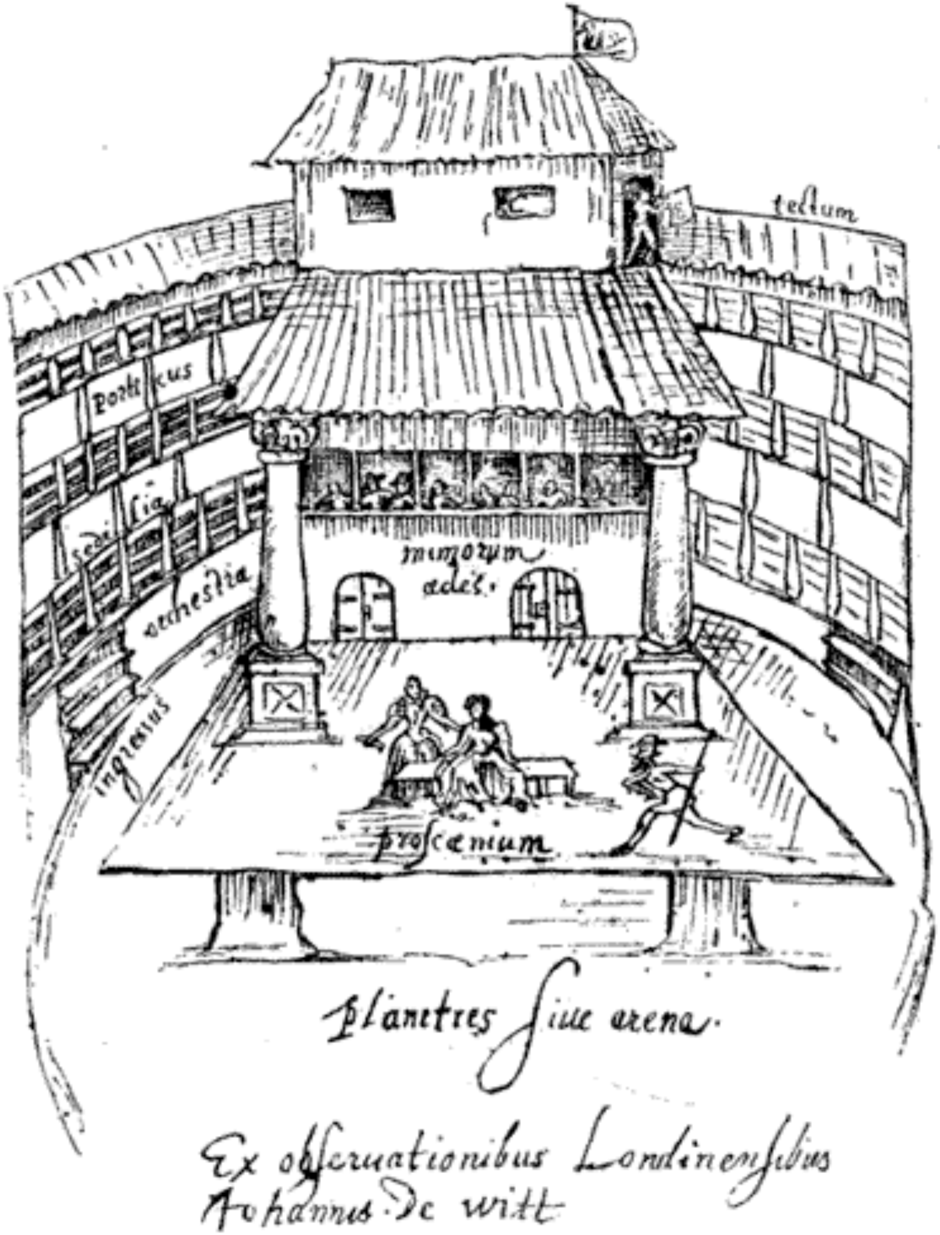
Shakespeare	EModE Plays	Present-day Plays
I pray you	it is a	I don’t know
I will not	what do you	what do you
I know not	and I will	I don’t want
I am a	it is not	do you think
I am not	I have a	do you want
my good lord	I will not	I don’t think
there is no	in the world	to do with
I would not	I tell you	do you know
it is a	I know not	going to be
and I will	I warrant you	don’t want to

Three-word N-grams in order of frequency (coloured items appear in another column)

Data in 2nd and 3rd columns draw from Culpeper and Kytö (2010)

Meanings and styles: Frequencies of word clusters (n-grams)

Meanings and styles: Frequencies of word clusters (n-grams)



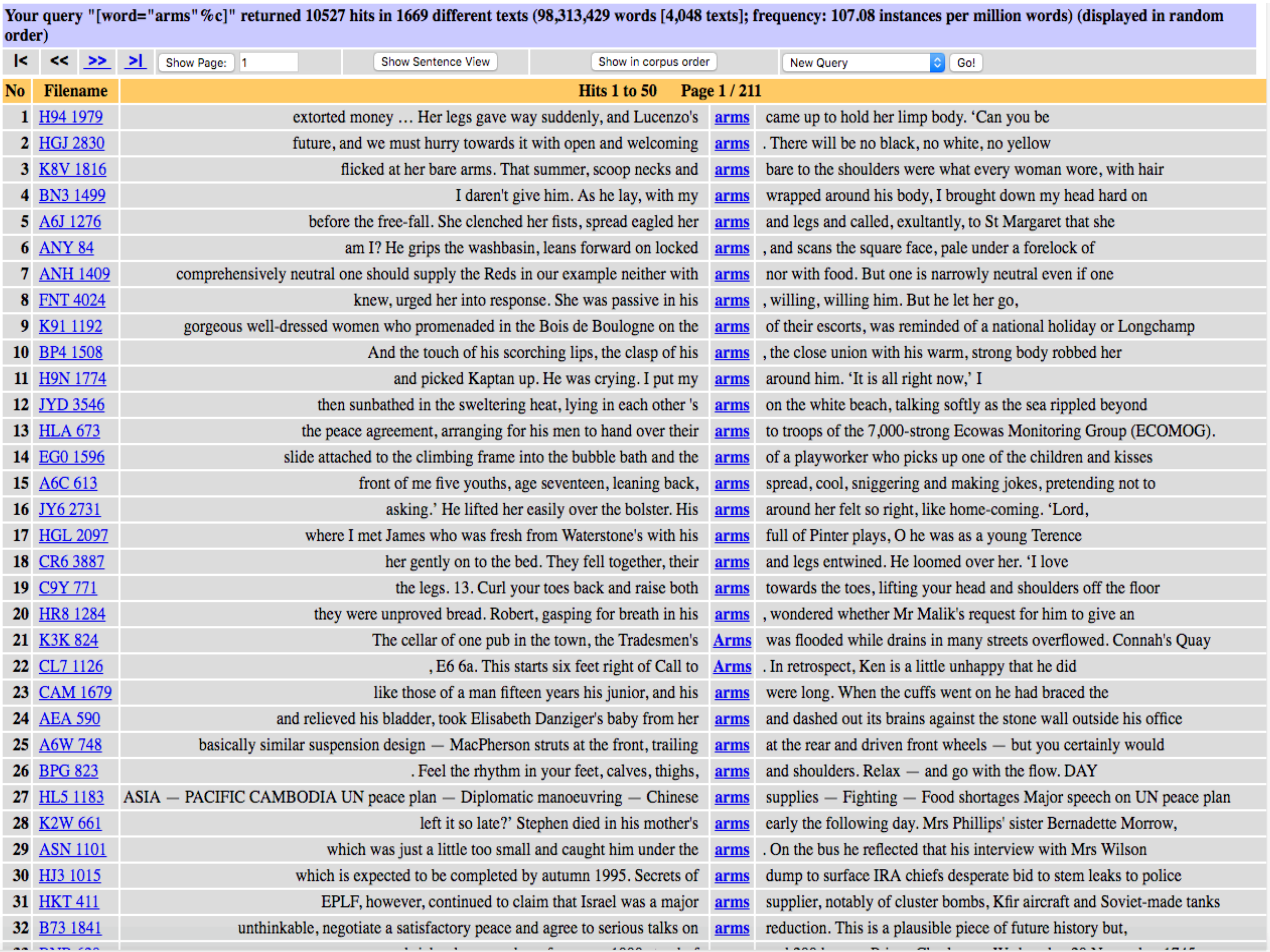
Purpose-built outdoor theatres:

The Theatre (1576),
The Curtain (1577),
The Rose (1587),
The Swan (1595),
The Globe (1599), and
The Fortune (1600).

Meanings and styles: Concordances Collocates

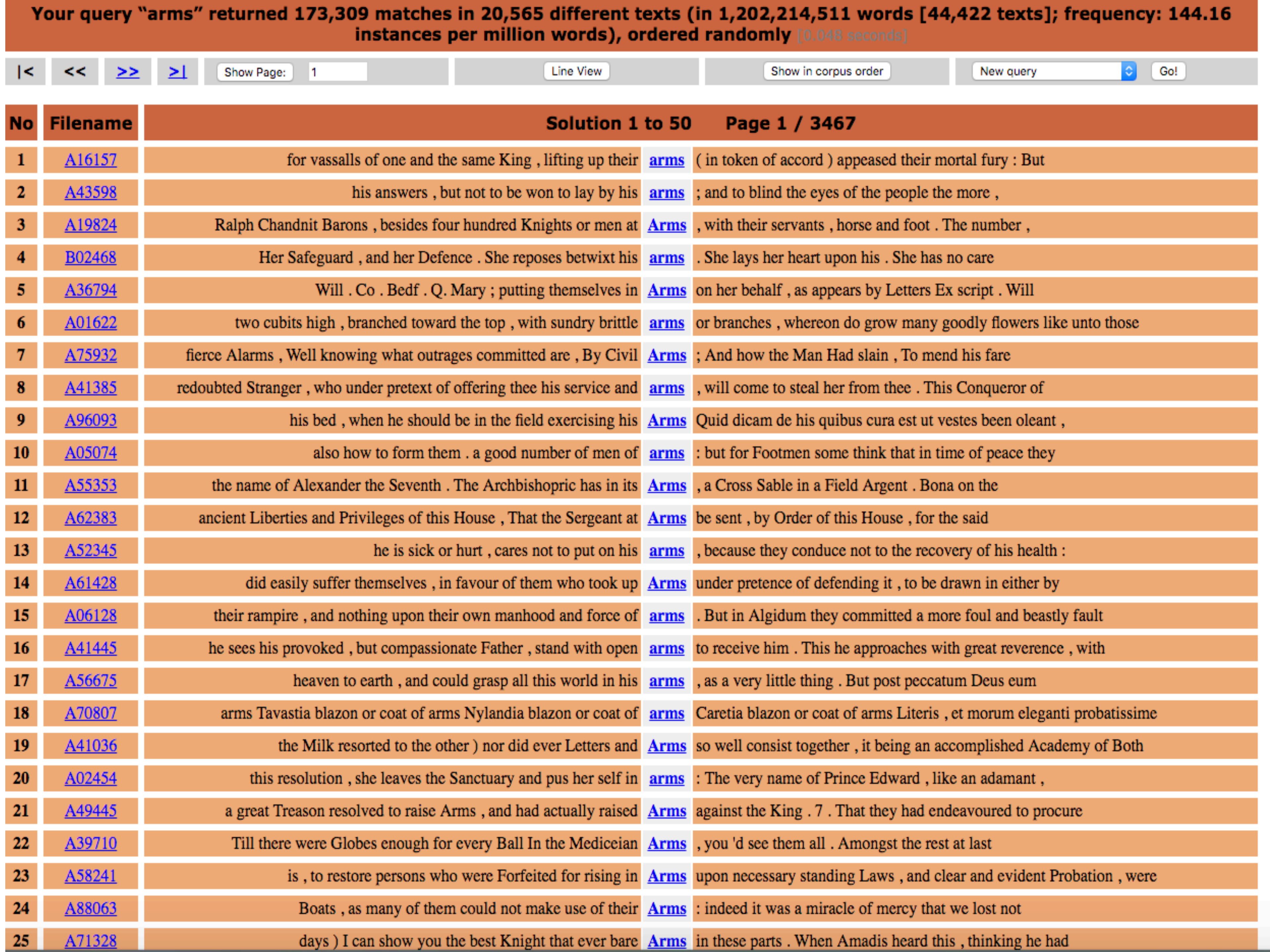
Has the word *arms* changed in meaning?

- A **concordance** in Early English Books Online (EEBO-TCP)
- A **concordance** in the British National Corpus



Your query "[word="arms"%c]" returned 10527 hits in 1669 different texts (98,313,429 words [4,048 texts]; frequency: 107.08 instances per million words) (displayed in random order)

No	Filename	Hits 1 to 50	Page 1 / 211	
1	H94 1979	extorted money … Her legs gave way suddenly, and Lucenzo's	arms	came up to hold her limp body. 'Can you be
2	HGJ 2830	future, and we must hurry towards it with open and welcoming	arms	. There will be no black, no white, no yellow
3	K8V 1816	flicked at her bare arms. That summer, scoop necks and	arms	bare to the shoulders were what every woman wore, with hair
4	BN3 1499	I daren't give him. As he lay, with my	arms	wrapped around his body, I brought down my head hard on
5	A6J 1276	before the free-fall. She clenched her fists, spread eagled her	arms	and legs and called, exultantly, to St Margaret that she
6	ANY 84	am I? He grips the washbasin, leans forward on locked	arms	, and scans the square face, pale under a forelock of
7	ANH 1409	comprehensively neutral one should supply the Reds in our example neither with	arms	nor with food. But one is narrowly neutral even if one
8	FNT 4024	knew, urged her into response. She was passive in his	arms	, willing, willing him. But he let her go,
9	K91 1192	gorgeous well-dressed women who promenaded in the Bois de Boulogne on the	arms	of their escorts, was reminded of a national holiday or Longchamp
10	BP4 1508	And the touch of his scorching lips, the clasp of his	arms	, the close union with his warm, strong body robbed her
11	H9N 1774	and picked Kaptan up. He was crying. I put my	arms	around him. 'It is all right now,' I
12	JYD 3546	then sunbathed in the sweltering heat, lying in each other 's	arms	on the white beach, talking softly as the sea rippled beyond
13	HLA 673	the peace agreement, arranging for his men to hand over their	arms	to troops of the 7,000-strong Ecowas Monitoring Group (ECOMOG).
14	EG0 1596	slide attached to the climbing frame into the bubble bath and the	arms	of a playworker who picks up one of the children and kisses
15	A6C 613	front of me five youths, age seventeen, leaning back,	arms	spread, cool, sniggering and making jokes, pretending not to
16	JY6 2731	asking.' He lifted her easily over the bolster. His	arms	around her felt so right, like home-coming. 'Lord,
17	HGL 2097	where I met James who was fresh from Waterstone's with his	arms	full of Pinter plays, O he was as a young Terence
18	CR6 3887	her gently on to the bed. They fell together, their	arms	and legs entwined. He loomed over her. 'I love
19	C9Y 771	the legs. 13. Curl your toes back and raise both	arms	towards the toes, lifting your head and shoulders off the floor
20	HR8 1284	they were unproved bread. Robert, gasping for breath in his	arms	, wondered whether Mr Malik's request for him to give an
21	K3K 824	The cellar of one pub in the town, the Tradesmen's	Arms	was flooded while drains in many streets overflowed. Connah's Quay
22	CL7 1126	, E6 6a. This starts six feet right of Call to	Arms	. In retrospect, Ken is a little unhappy that he did
23	CAM 1679	like those of a man fifteen years his junior, and his	arms	were long. When the cuffs went on he had braced the
24	AEA 590	and relieved his bladder, took Elisabeth Danziger's baby from her	arms	and dashed out its brains against the stone wall outside his office
25	A6W 748	basically similar suspension design — MacPherson struts at the front, trailing	arms	at the rear and driven front wheels — but you certainly would
26	BPG 823	. Feel the rhythm in your feet, calves, thighs,	arms	and shoulders. Relax — and go with the flow. DAY
27	HL5 1183	ASIA — PACIFIC CAMBODIA UN peace plan — Diplomatic manoeuvring — Chinese	arms	supplies — Fighting — Food shortages Major speech on UN peace plan
28	K2W 661	left it so late?' Stephen died in his mother's	arms	early the following day. Mrs Phillips' sister Bernadette Morrow,
29	ASN 1101	which was just a little too small and caught him under the	arms	. On the bus he reflected that his interview with Mrs Wilson
30	HJ3 1015	which is expected to be completed by autumn 1995. Secrets of	arms	dump to surface IRA chiefs desperate bid to stem leaks to police
31	HKT 411	EPLF, however, continued to claim that Israel was a major	arms	supplier, notably of cluster bombs, Kfir aircraft and Soviet-made tanks
32	B73 1841	unthinkable, negotiate a satisfactory peace and agree to serious talks on	arms	reduction. This is a plausible piece of future history but,



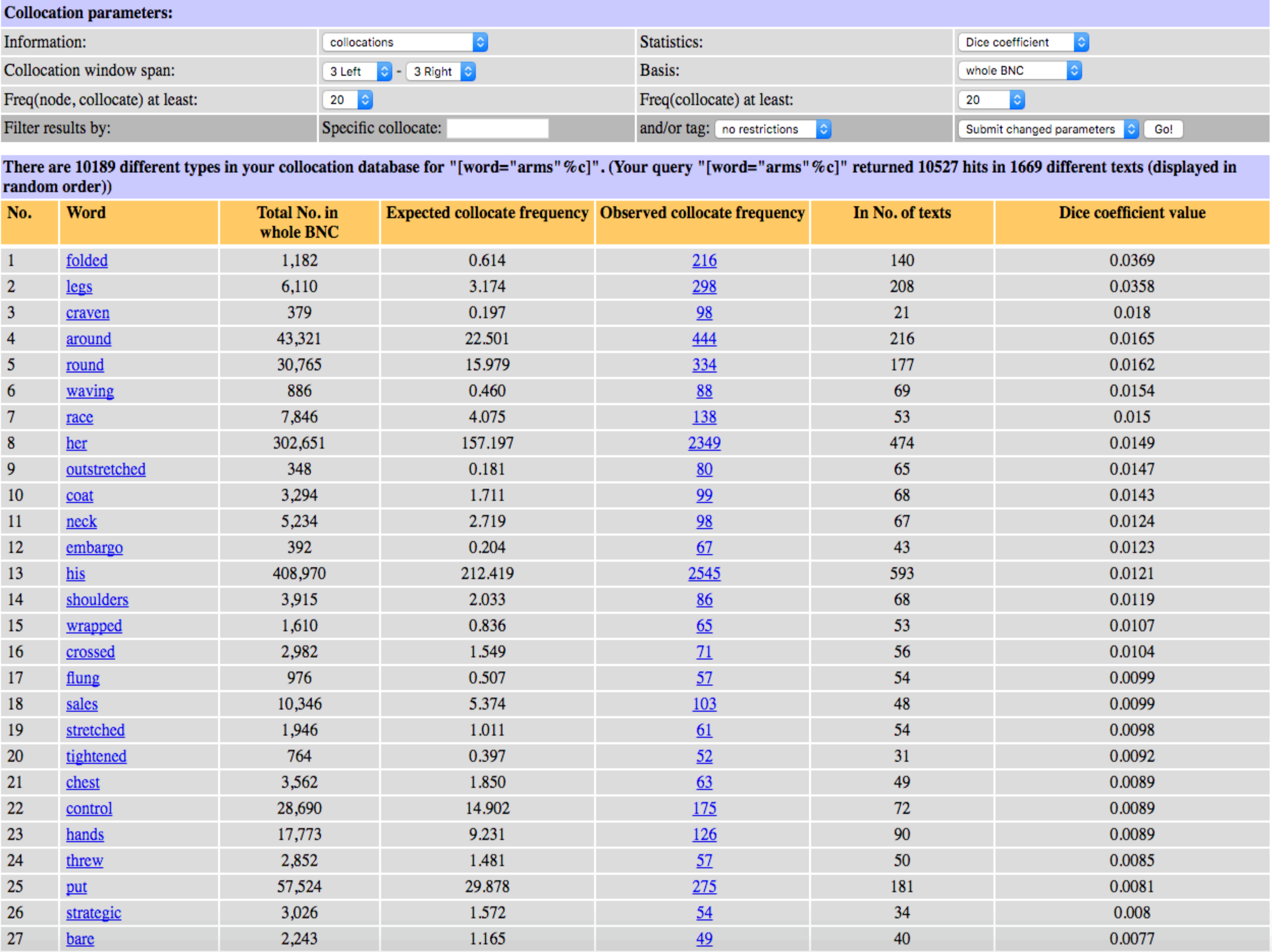
Your query "arms" returned 173,309 matches in 20,565 different texts (in 1,202,214,511 words [44,422 texts]; frequency: 144.16 instances per million words), ordered randomly [0.048 seconds]

No	Filename	Solution 1 to 50	Page 1 / 3467	
1	A16157	for vassalls of one and the same King , lifting up their	arms	(in token of accord) appeased their mortal fury : But
2	A43598	his answers , but not to be won to lay by his	arms	; and to blind the eyes of the people the more ,
3	A19824	Ralph Chandnit Barons , besides four hundred Knights or men at	Arms	, with their servants , horse and foot . The number ,
4	B02468	Her Safeguard , and her Defence . She reposes betwixt his	arms	. She lays her heart upon his . She has no care
5	A36794	Will . Co . Bedf . Q. Mary ; putting themselves in	Arms	on her behalf , as appears by Letters Ex script . Will
6	A01622	two cubits high , branched toward the top , with sundry brittle	arms	or branches , whereon do grow many goodly flowers like unto those
7	A75932	fierce Alarms , Well knowing what outrages committed are , By Civil	Arms	; And how the Man Had slain , To mend his fare
8	A41385	redoubted Stranger , who under pretext of offering thee his service and	arms	, will come to steal her from thee . This Conqueror of
9	A96093	his bed , when he should be in the field exercising his	Arms	Quid dicam de his quibus cura est ut vestes been oleant ,
10	A05074	also how to form them . a good number of men of	arms	: but for Footmen some think that in time of peace they
11	A55353	the name of Alexander the Seventh . The Archbishopric has in its	Arms	, a Cross Sable in a Field Argent . Bona on the
12	A62383	ancient Liberties and Privileges of this House , That the Sergeant at	Arms	be sent , by Order of this House , for the said
13	A52345	he is sick or hurt , cares not to put on his	arms	, because they conduce not to the recovery of his health :
14	A61428	did easily suffer themselves , in favour of them who took up	Arms	under pretence of defending it , to be drawn in either by
15	A06128	their rampire , and nothing upon their own manhood and force of	arms	. But in Algidum they committed a more foul and beastly fault
16	A41445	he sees his provoked , but compassionate Father , stand with open	arms	to receive him . This he approaches with great reverence , with
17	A56675	heaven to earth , and could grasp all this world in his	arms	, as a very little thing . But post peccatum Deus eum
18	A70807	arms Tavastia blazon or coat of arms Nylandia blazon or coat of	arms	Caretia blazon or coat of arms Literis , et morum eleganti probatissime
19	A41036	the Milk resorted to the other) nor did ever Letters and	Arms	so well consist together , it being an accomplished Academy of Both
20	A02454	this resolution , she leaves the Sanctuary and pus her self in	arms	: The very name of Prince Edward , like an adamant ,
21	A49445	a great Treason resolved to raise Arms , and had actually raised	Arms	against the King . 7 . That they had endeavoured to procure
22	A39710	Till there were Globes enough for every Ball In the Mediceian	Arms	, you 'd see them all . Amongst the rest at last
23	A58241	is , to restore persons who were Forfeited for rising in	Arms	upon necessary standing Laws , and clear and evident Probation , were
24	A88063	Boats , as many of them could not make use of their	Arms	: indeed it was a miracle of mercy that we lost not
25	A71328	days) I can show you the best Knight that ever bare	Arms	in these parts . When Amadis heard this , thinking he had

Meanings and styles: Concordances Collocates

Problem: Sometimes a concordance is too long and complex to see the patterns.

- So we can examine **collocates**.
- A collocation is a lexical co-occurrence pattern, a habitual co-occurrence between a "node" (e.g. *arms*) and the words or "collocates" that tend to co-occur with it within a particular span (e.g. 3 words to the left and 3 words to the right).
- Knotty problems attend not only the size of the span, but statistics used to identify that habitual co-occurrence pattern.

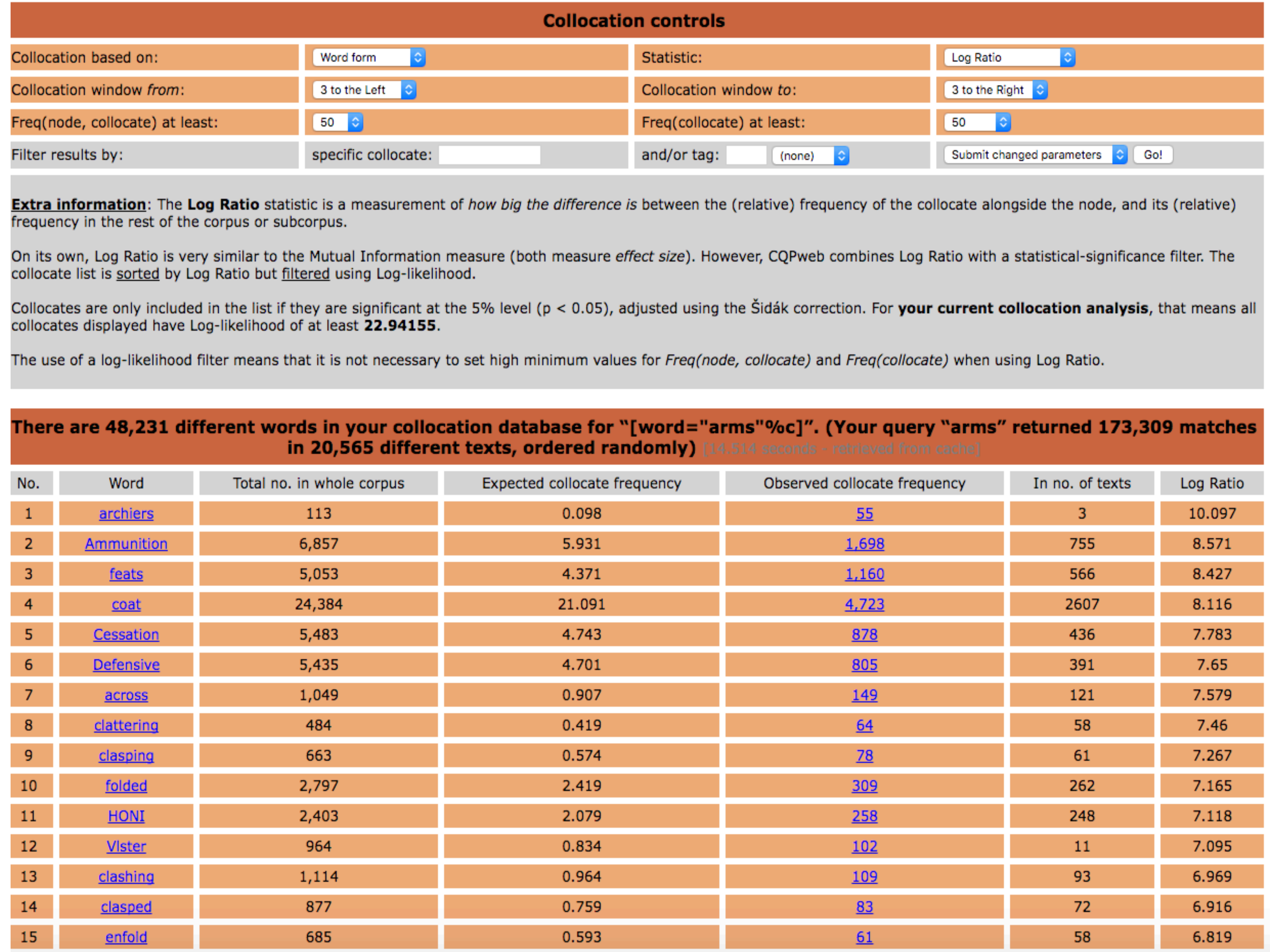


Collocation parameters:

Information:	collocations	Statistics:	Dice coefficient
Collocation window span:	3 Left - 3 Right	Basis:	whole BNC
Freq(node, collocate) at least:	20	Freq(collocate) at least:	20
Filter results by:	Specific collocate:	and/or tag: no restrictions	Submit changed parameters Go!

There are 10189 different types in your collocation database for "[word="arms"%c]". (Your query "[word="arms"%c]" returned 10527 hits in 1669 different texts (displayed in random order))

No.	Word	Total No. in whole BNC	Expected collocate frequency	Observed collocate frequency	In No. of texts	Dice coefficient value
1	folded	1,182	0.614	216	140	0.0369
2	legs	6,110	3.174	298	208	0.0358
3	craven	379	0.197	98	21	0.018
4	around	43,321	22.501	444	216	0.0165
5	round	30,765	15.979	334	177	0.0162
6	waving	886	0.460	88	69	0.0154
7	race	7,846	4.075	138	53	0.015
8	her	302,651	157.197	2349	474	0.0149
9	outstretched	348	0.181	80	65	0.0147
10	coat	3,294	1.711	99	68	0.0143
11	neck	5,234	2.719	98	67	0.0124
12	embargo	392	0.204	67	43	0.0123
13	his	408,970	212.419	2545	593	0.0121
14	shoulders	3,915	2.033	86	68	0.0119
15	wrapped	1,610	0.836	65	53	0.0107
16	crossed	2,982	1.549	71	56	0.0104
17	flung	976	0.507	57	54	0.0099
18	sales	10,346	5.374	103	48	0.0099
19	stretched	1,946	1.011	61	54	0.0098
20	tightened	764	0.397	52	31	0.0092
21	chest	3,562	1.850	63	49	0.0089
22	control	28,690	14.902	175	72	0.0089
23	hands	17,773	9.231	126	90	0.0089
24	threw	2,852	1.481	57	50	0.0085
25	put	57,524	29.878	275	181	0.0081
26	strategic	3,026	1.572	54	34	0.008
27	bare	2,243	1.165	49	40	0.0077



Collocation controls

Collocation based on:	Word form	Statistic:	Log Ratio
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	50	Freq(collocate) at least:	50
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

Extra information: The **Log Ratio** statistic is a measurement of *how big the difference is* between the (relative) frequency of the collocate alongside the node, and its (relative) frequency in the rest of the corpus or subcorpus.

On its own, Log Ratio is very similar to the Mutual Information measure (both measure *effect size*). However, CQPweb combines Log Ratio with a statistical-significance filter. The collocate list is sorted by Log Ratio but filtered using Log-likelihood.

Collocates are only included in the list if they are significant at the 5% level (p < 0.05), adjusted using the Šidák correction. For **your current collocation analysis**, that means all collocates displayed have Log-likelihood of at least **22.94155**.

The use of a log-likelihood filter means that it is not necessary to set high minimum values for *Freq(node, collocate)* and *Freq(collocate)* when using Log Ratio.

There are 48,231 different words in your collocation database for "[word="arms"%c]". (Your query "arms" returned 173,309 matches in 20,565 different texts, ordered randomly) [14.514 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log Ratio
1	archiers	113	0.098	55	3	10.097
2	Ammunition	6,857	5.931	1,698	755	8.571
3	feats	5,053	4.371	1,160	566	8.427
4	coat	24,384	21.091	4,723	2607	8.116
5	Cessation	5,483	4.743	878	436	7.783
6	Defensive	5,435	4.701	805	391	7.65
7	across	1,049	0.907	149	121	7.579
8	clattering	484	0.419	64	58	7.46
9	clasping	663	0.574	78	61	7.267
10	folded	2,797	2.419	309	262	7.165
11	HONI	2,403	2.079	258	248	7.118
12	Vlster	964	0.834	102	11	7.095
13	clashing	1,114	0.964	109	93	6.969
14	clasped	877	0.759	83	72	6.916
15	enfold	685	0.593	61	58	6.819

Meanings and styles: Concordances Collocates

The case of *good*

Crystal & Crystal (2004:201-202):

(1) [intensifying use] real, genuine (‘love no man in good earnest’).
(2) kind, benevolent, generous.
(3) kind, friendly, sympathetic.
(4) amenable, tractable, manageable.
(5) honest, virtuous, honourable.
(6) seasonable, appropriate, proper.
(7) just, right, commendable.
(8) intended, right, proper.
(9) high-ranking, highborn, distinguished.
(10) rich, wealthy, substantial.

Meanings and styles: Concordances Collocates

1. A polite address: '(my) good Lord/friend/Sir/Master/Lady/Madam/ etc.'. Typically used when meeting or parting, thanking or making suggestions. *But (good my Lord) do it so cunningly* TGV, III. 1.

2. Honest, truthful, principled; of high moral standards. (This sense also shapes the discourse markers '(in) good faith/sooth/troth', which mean truly or honestly). *a man of good repute, carriage, bearing, & estimation* LLL, I. 1.

3. Positive rather than negative. Typically, contrasted with 'bad'. *Is thy news good or bad?* ROM, II. 5.

4. In one's favour, especially favourable wishes or blessings. *The Gods be good to us* COR, V. 4.

5. A welcoming, cheerful manner. *Therefore for Gods sake entertain good comfort, And cheer his Grace with quick and merry eyes* R3, I. 3.

Meanings and styles: Concordances Collocates

The case of *Irish*

- Strongest collocate: *Irish rug*

"Show me a fair scarlet, a vvelch frise, a good Irish rug" (Eliot, 1595)

- Thematic groups (top 50 collocates)

Negative connotations (items below are relatively frequent & well dispersed)

Uncivilised: *savage, wild*

Hostile: *wars*, *enemies*, *against*

Ungovernable: *rebels*

Associated groups: *Scottish, Scots,* (*English)*

Insignificant??: *mere*

Political power: *nation, lords*

Language: *tongue, language, speak*

Meanings and styles: Keywords

- ‘**Keyness**’ is a matter of an item’s frequency in a body of data being statistically unusual relative to that item in a comparative body of data.
- Keywords are not keywords in the sense of Raymond Williams (1976), where they are cultural, social and political hotspots.
- Keywords are statistically based style markers.

Meanings and styles: Keywords

Lily James and Richard Madden.

(Photo: Johan Perrson)

- What language characterizes Romeo and what language Juliet?

Meanings and styles: Keywords

Rank-ordered keywords for Romeo and Juliet (raw frequencies in brackets)

Romeo	Juliet
beauty (10), love (46), blessed (5), eyes (14), more (26), mine (14), dear (13), rich (7), me (73), yonder (5), farewell (11), sick (6), lips (9), stars (5), fair (15), hand (11), thine (7), banished (9), goose (5), that (84)	if (31), be (59), or (25), I (138), sweet (16), my (92), news (9), thou (71), night (27), would (20), yet (18), that (82), nurse (20), name (11), words (5), Tybalt's (6), send (7), husband (7), swear (5), where (16), again (10)

Meanings and styles: Keywords

Juliet**:**

- **If** he **be** married, / Our grave is like to **be our** wedding-bed (I.v.)
- **If** they do see thee, they will murder thee (II.ii.)
- But **if** thou meanest not well (II.ii.)
- Is thy news good, **or** bad? answer to that; Say either, and I'll stay the circumstance: Let me be satisfied, is 't good **or** bad? (II.ii)
- Tis almost morning; I would have thee gone; And **yet** no further than a wanton's bird […] (II.ii.)

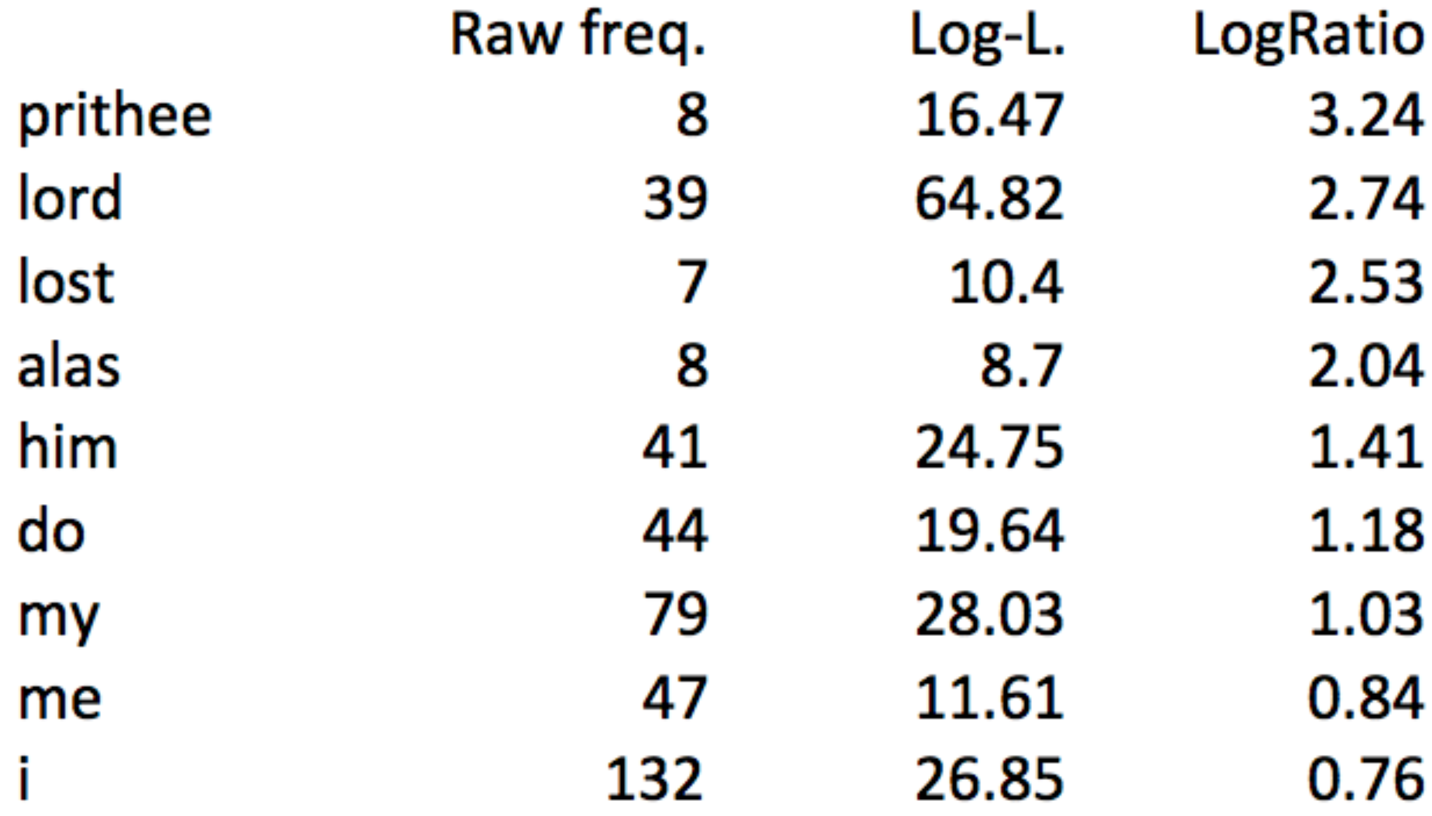
Meanings and styles: Keywords

How keywords move beyond simple frequency lists. The case of Shakespeare‘s Desdemona.

TOTAL	**2753**
I	132
my	79
and	61
you	60
to	57
not	48
me	47
do	44
the	41
him	41
lord	39
that	38

Meanings and styles: Keywords

Desdemona’s keywords



	Raw freq.	Log-L.	LogRatio
prithee	8	16.47	3.24
lord	39	64.82	2.74
lost	7	10.4	2.53
alas	8	8.7	2.04
him	41	24.75	1.41
do	44	19.64	1.18
my	79	28.03	1.03
me	47	11.61	0.84
i	132	26.85	0.76

For Othello: *I* is ranked 109, *me* 70 and *my* 74

A note on programs I used, etc.

See handout!

Concluding remarks

- Although I have focused on words, these techniques work for other items – phrases/expressions, grammatical tags, semantic tags, etc.
- The techniques will work for small datasets and large, although some techniques don’t produce anything sensible for really small datasets and computing power can be an issue for really large datasets.
- Techniques and tools are constantly being developed.
  - At Lancaster: e.g. LancsBox
  - Laurence Anthony