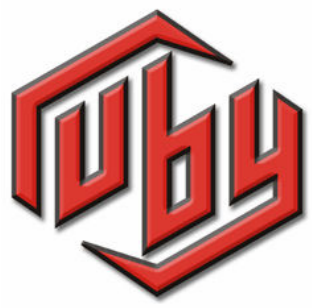


Text Analysis of Romeo and Juliet

Version: 31 October 2007

Text Analysis of Romeo and Juliet

TEXT ANALYSIS OF ROMEO AND JULIET

This document examines some approaches to the analysis of a literary text using cross tabulation and time series smoothing.

Preparation of the text

The case unit is a word, question mark or exclamation mark. The text was prepared for importing by

- removing all stage instructions, all Act and Scene delimiters, and all punctuation except ! and ?
- expanding all speaker indications to the full name (eg Merc. to Mercutio)
- prefixing 'dp', for dramatis personae, before all speaker indicators
- reorganising the text as a single column of words

The result to this point is to tranform the original text from, for example,

> Samp. Gregory, on my word, we'll not carry coals.
> Greg. No, for then we should be colliers.

to

dpSampson
Gregory
on
my
word
we'll
not
carry
coals
dpGregory
No
for
then
we
should
be
colliers

For rudimentary analysis, the character, act and scene is recorded against each case, (where a case is a single word, ! or ?). Rearranging the single column of text to get the speaker against each word was done using the following script.

```
''' turn lines like
   '''    dpJuliet
   '''    Romeo
   '''    Romeo
   '''    Wherefore
   '''    art
   '''    thou
   '''    Romeo
   ''' into lines like
   '''    Romeo        dpJuliet
   '''    Romeo        dpJuliet
   '''    Wherefore    dpJuliet
   '''    art          dpJuliet
   '''    thou         dpJuliet
   '''    Romeo        dpJuliet

   line = ""
   dp   = ""
   count = 1
   Do While Not sourcefile.AtEndOfStream
      line = sourcefile.ReadLine
      if Left(line, 2) = "dp" then
         dp = line                ''' remember dramatis personae
      else
         outfile.WriteLine line & vbTab & dp
      end if
   Loop

   outfile.Close
   sourcefile.Close
```

The output was then loaded into Excel, the Act and Scene data manually entered, and each column named according to the final desired variable. Here, for example, is the beginning of Act 1, Scene 2.

Text	DramPers	Act	Scene
But	dpCapulet	1	2
Montague	dpCapulet	1	2
is	dpCapulet	1	2
bound	dpCapulet	1	2
as	dpCapulet	1	2
well	dpCapulet	1	2
as	dpCapulet	1	2
I	dpCapulet	1	2

This was then saved as a tab-delimited text file and imported into Ruby, creating four source variables: Text, DramPers, Act and Scene.

Basic Statistics

1. Number of words, excluding ! and ?: 23,961

2. Number of ! and ?:

!	486
?	371

3. Main Characters (defined as >= 2%) as percent of text:

dpRomeo	19%
dpJuliet	18%
dpFriarLaurence	11%
dpNurse	10%
dpCapulet	9%
dpMercutio	9%
dpBenvolio	5%
dpPrince	2%
dpParis	2%
dpLadyCapulet	2%

Tybalt, though a major element of the plot, has only 1% of the text.

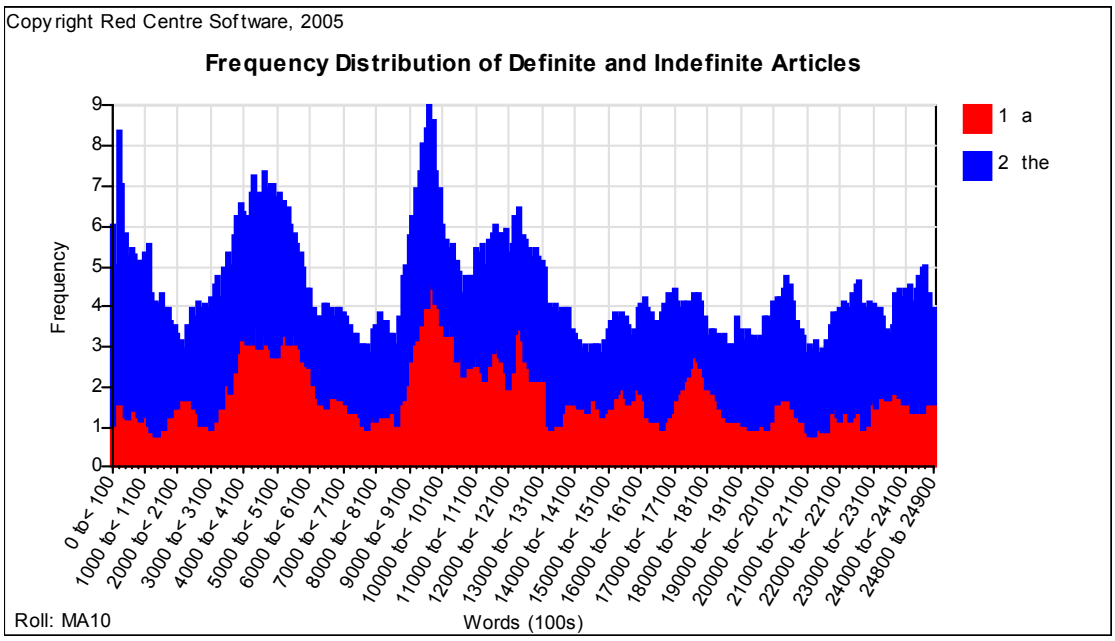
4. Words per Scene, counts and percents:

	Frequency	Percent
Total	24,818	100%
Act 1 Scene 1	1982	8%
Act 1 Scene 2	808	3%
Act 1 Scene 3	911	4%
Act 1 Scene 4	917	4%
Act 1 Scene 5	1333	5%
Act 2 Scene 1	360	1%
Act 2 Scene 2	1613	6%
Act 2 Scene 3	768	3%
Act 2 Scene 4	1702	7%
Act 2 Scene 5	706	3%
Act 2 Scene 6	291	1%
Act 3 Scene 1	1641	7%
Act 3 Scene 2	1190	5%
Act 3 Scene 3	1421	6%
Act 3 Scene 4	305	1%
Act 3 Scene 5	2078	8%
Act 4 Scene 1	1037	4%
Act 4 Scene 2	371	1%
Act 4 Scene 3	491	2%
Act 4 Scene 4	235	1%
Act 4 Scene 5	1174	5%
Act 5 Scene 1	707	3%
Act 5 Scene 2	228	1%
Act 5 Scene 3	2549	10%

Analysis

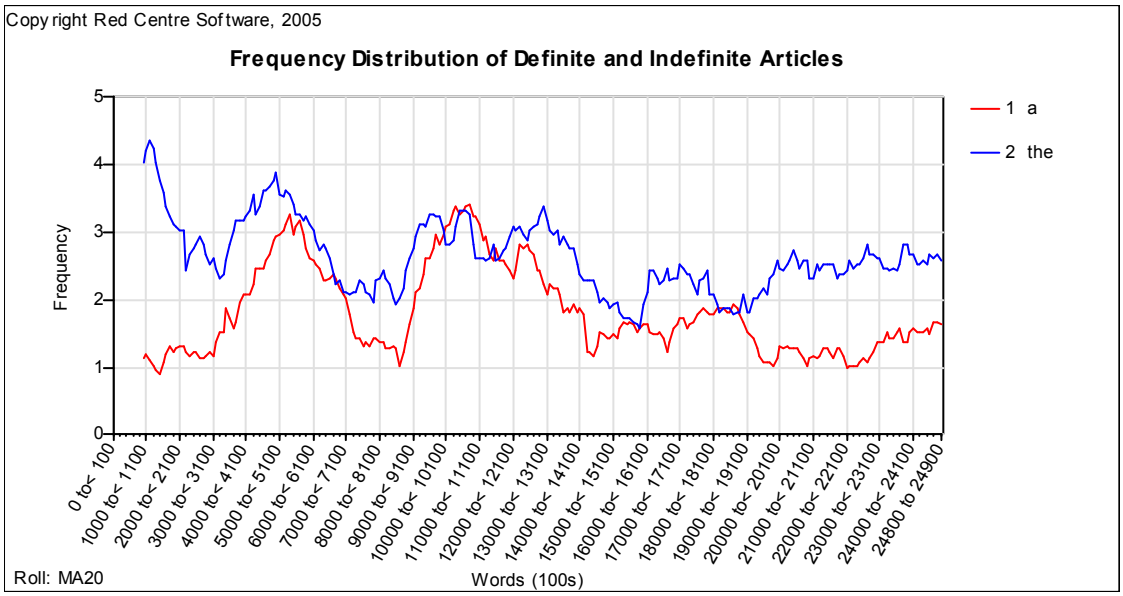
The text can be considered as a time series by either resolving as groups of words, where the minimum resolution is a single word and the maximum resolution is the entire text, or as a sequence of scenes. Both approaches have been used. For things such as the distribution of articles or pronouns, grouping words probably gives more meaningful output, whereas for thematic elements, such as love or conflict, scene by scene is often more appropriate.

1. Distribution of Articles

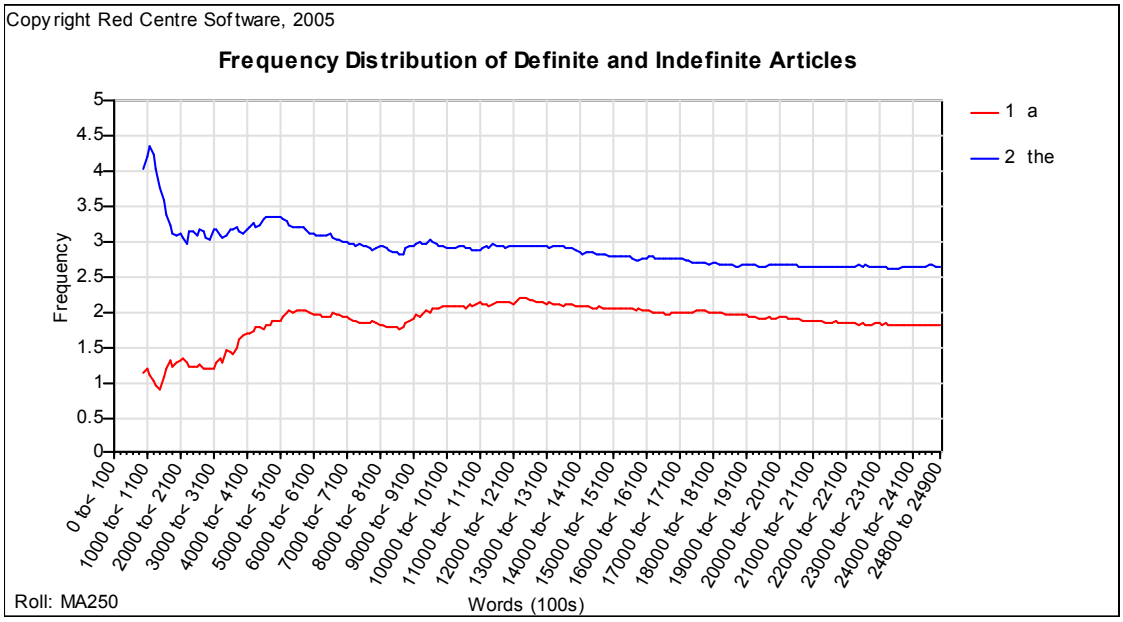


This chart shows the number of times a definite or indefinite article occurs in each group of 100 words, averaged over the current and preceding nine groups of 100 words. For example, if the frequencies per 100 words for ten groups are 4, 6, 4, 6, 3, 7, 3, 7, 5, 5 then the plotted value is (4+6+4+6+3+7+3+7+5+5)/10 = 5. Because each x axis point represents 100 words, the plot is effectively a percentaging, so about 4% of the text comprises articles.

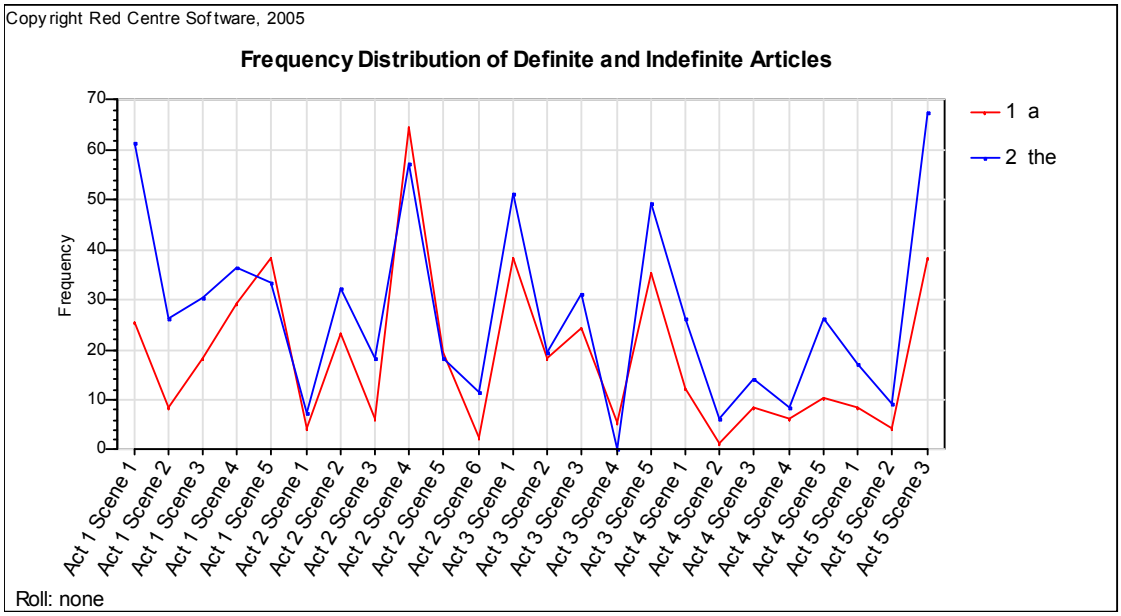
The distribution is not as smooth as one would expect. Over the first 13,000 words, the articles clump in quite a different pattern, with peaks at around 4,000 and 10,000, compared to the remainder of the text. The same data as line series, averaged over 20 groups of 100 words, shows that the frequencies rise and fall together, except for the sharp drop in 'a' counts at 19,500.



Increasing the moving average to 250 so that the last plotted points are the average across the entire text (because 250*100 = 25,000), shows the indefinite article at 1.8%, and the definite article at 2.6%, giving 4.4% combined.

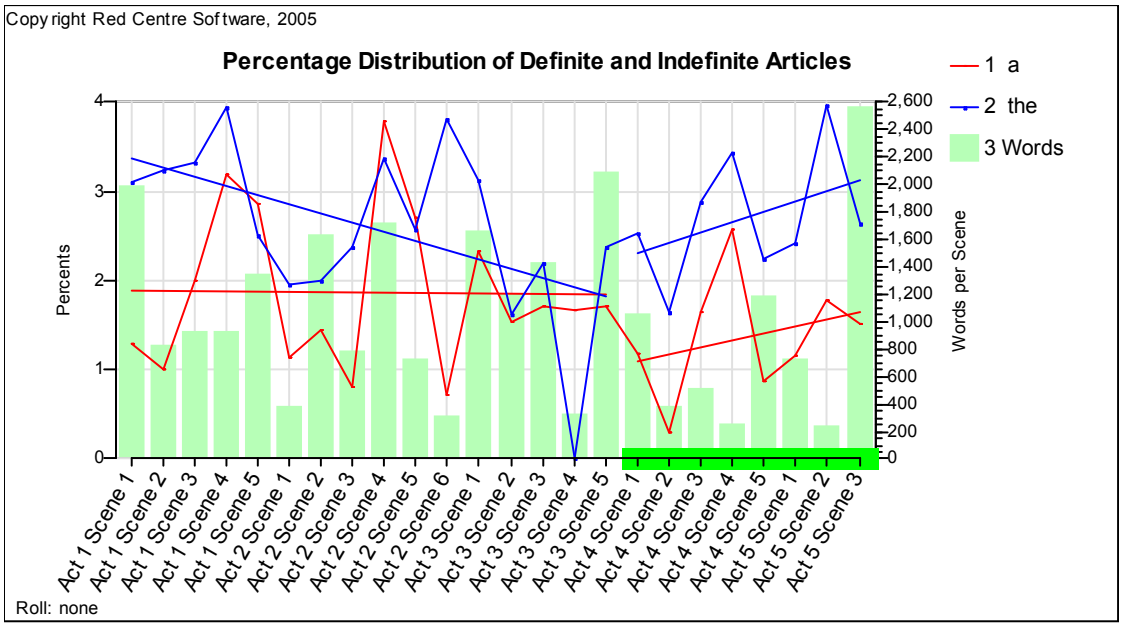


Plotting article frequency counts per scene (as opposed to per 100 words) reveals a tighter correlation.



Act 3 Scene 4 has no occurences of 'the'.

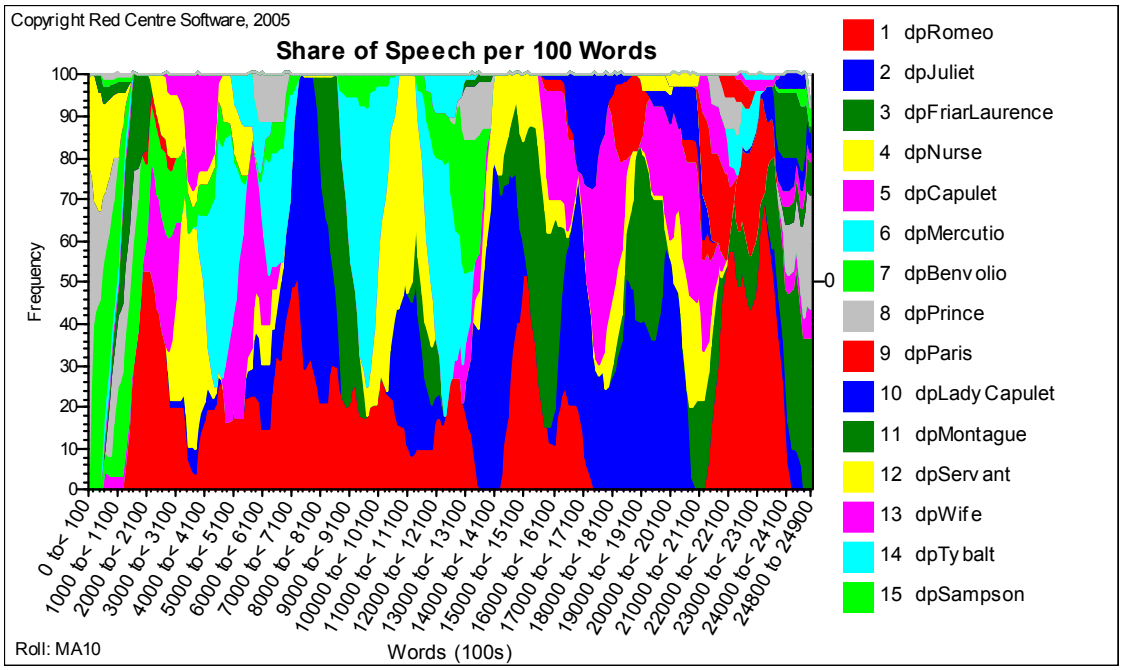
As percents, with the number of words per scene on Y2, shows that proportional divergence only happens for very short scenes, in particular Act 2 Scene 6 and Act 3 Scene 4.



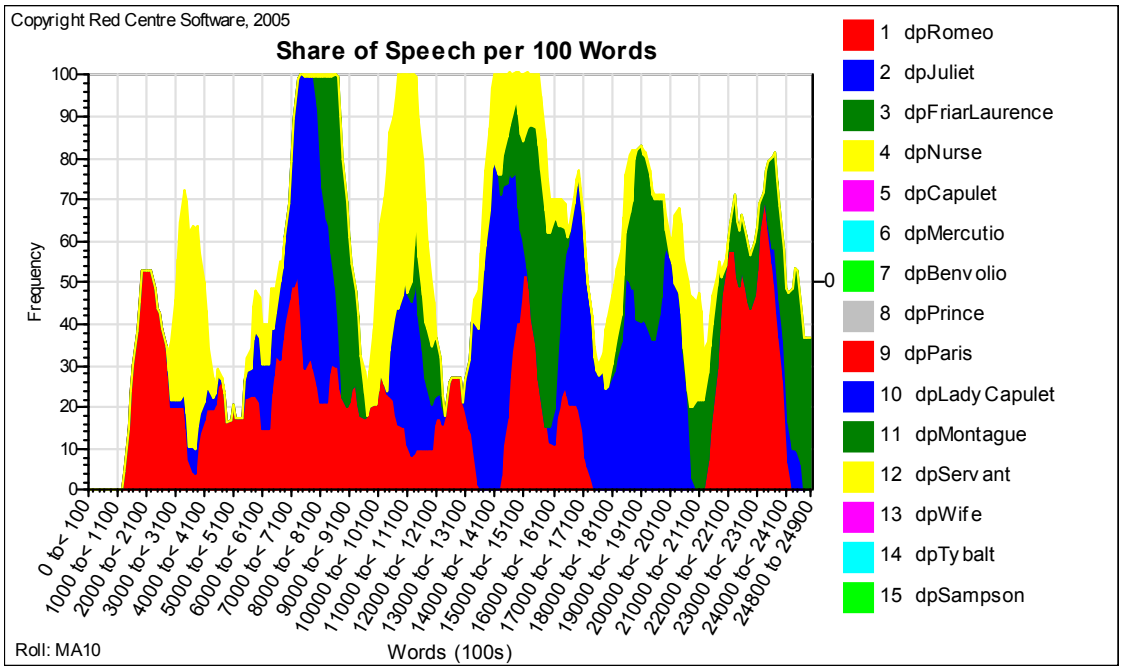
Strangely, the proportions are a lot more regular from Act 4, Scene 1, as the regression lines show. In fact, all earlier (from Act 1 Scene 1 to Act 3 Scene 5) looks rather chaotic by comparison. Does this suggest more careful re-working/re-drafting for the final two acts? It would be interesting to see if this were the case for other plays.

2. Share of Speech

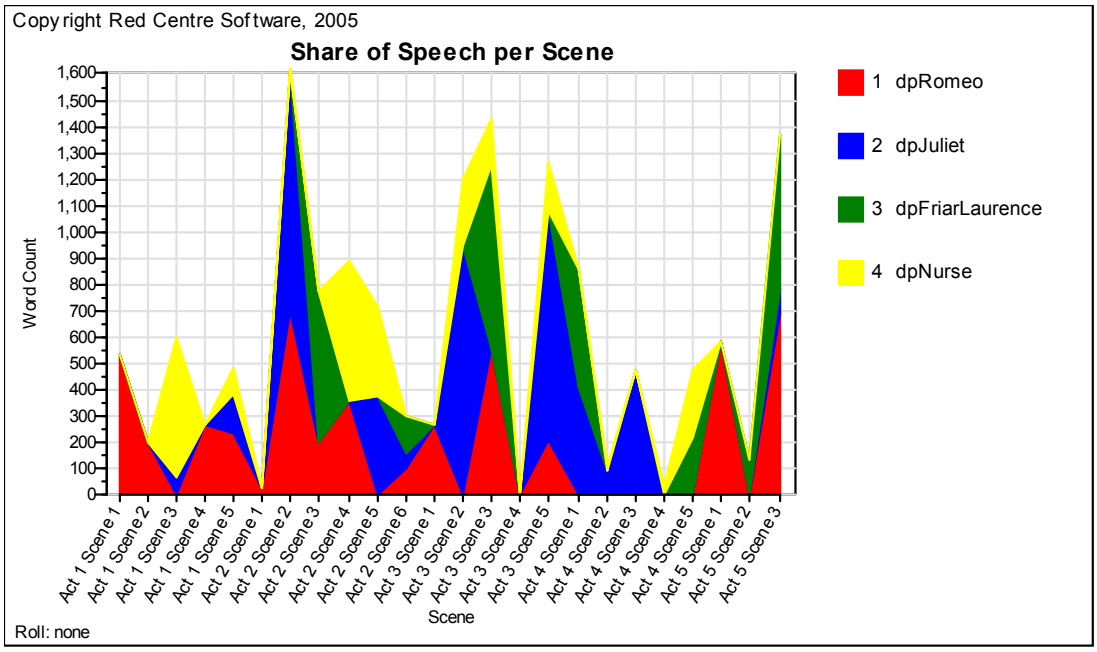
This chart shows that Romeo, Juliet, Friar Laurence and the Nurse do most of the talking, in that order.

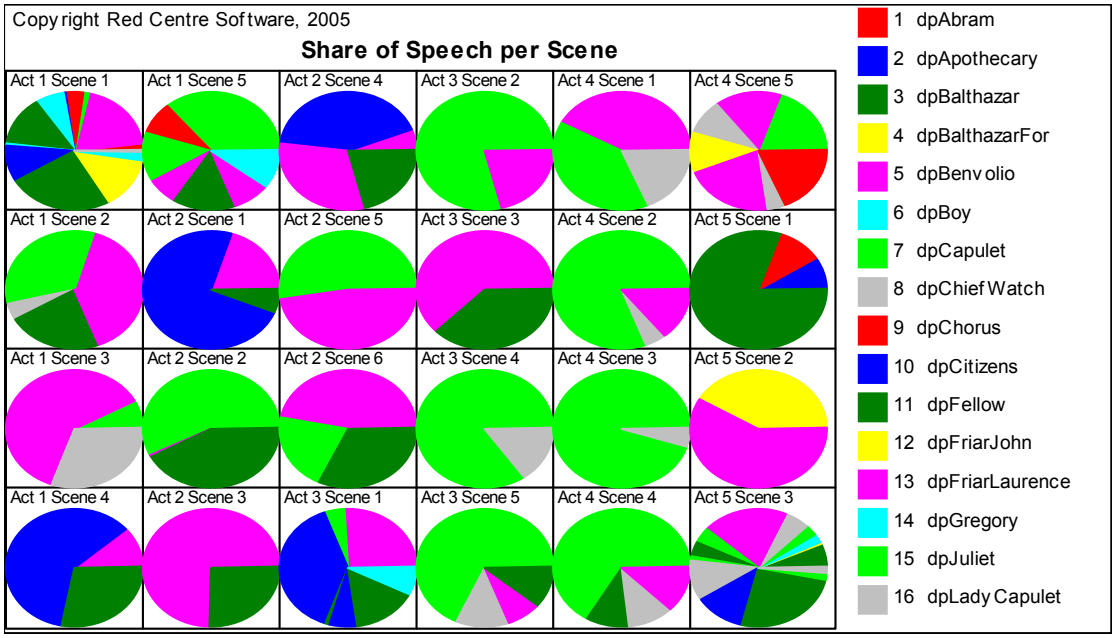


Showing just these four:

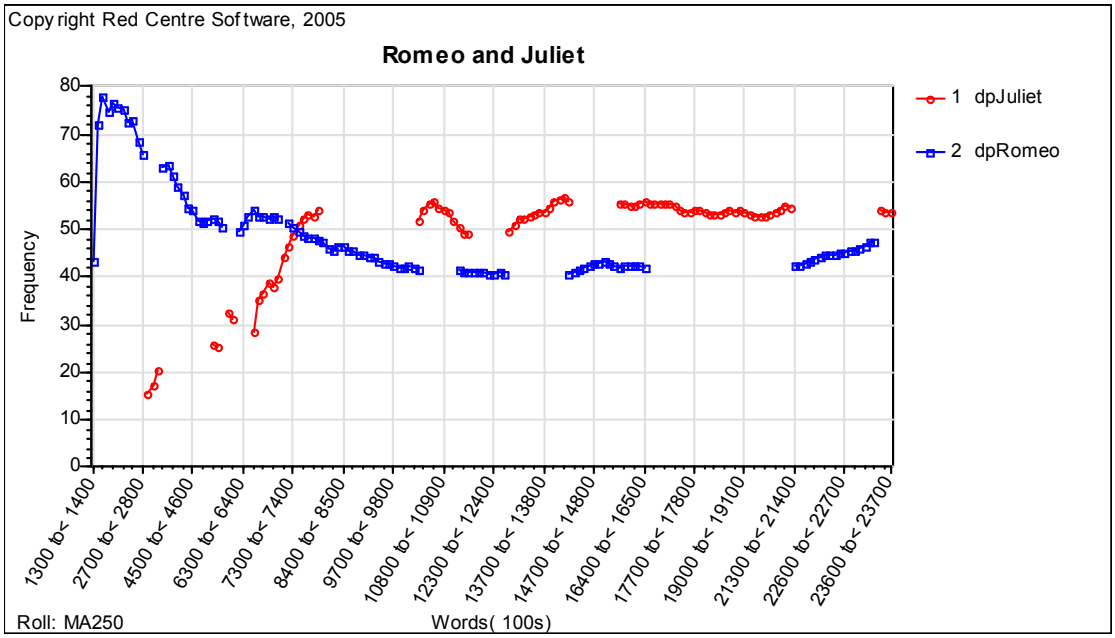


Oppositions of male/female, youth/age, passion/restraint, individual/institution (friar, nurse), death/life.



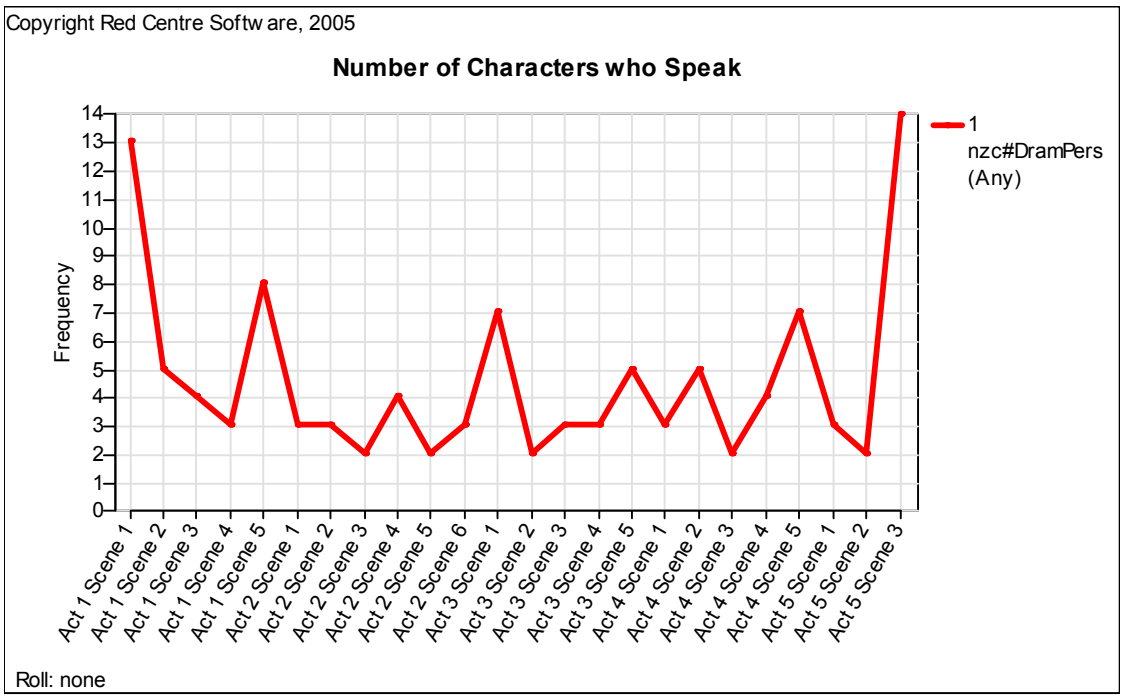


Romeo talks the most intially, but Juliet soon catches up.



3. Characters per Scene

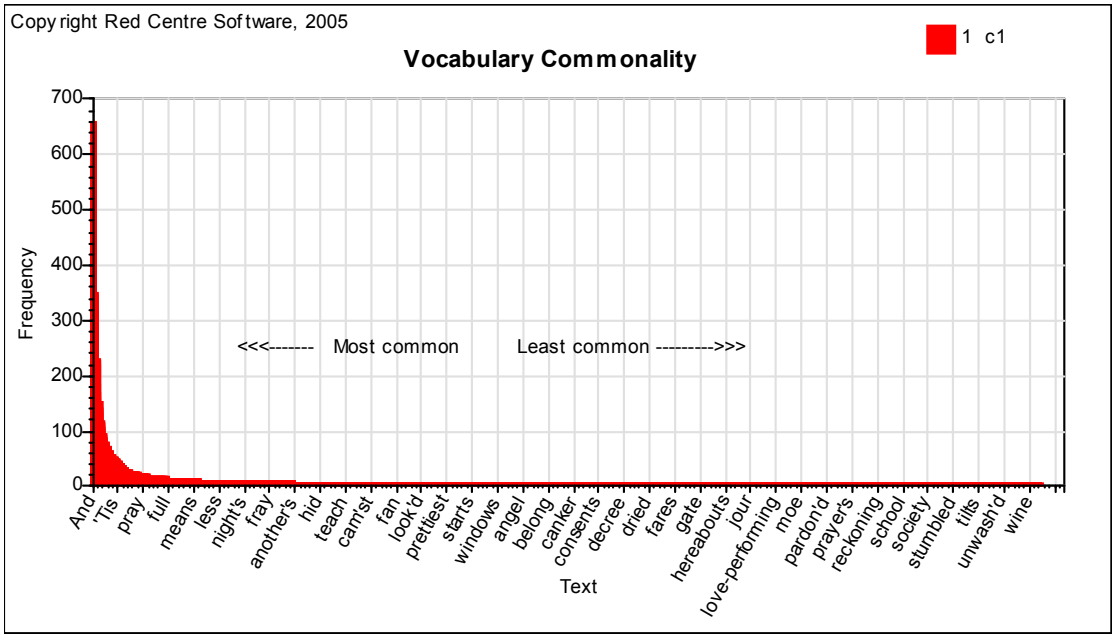
The number of characters per scene can be explicitly obtain by plotting the count of non-zero cells.



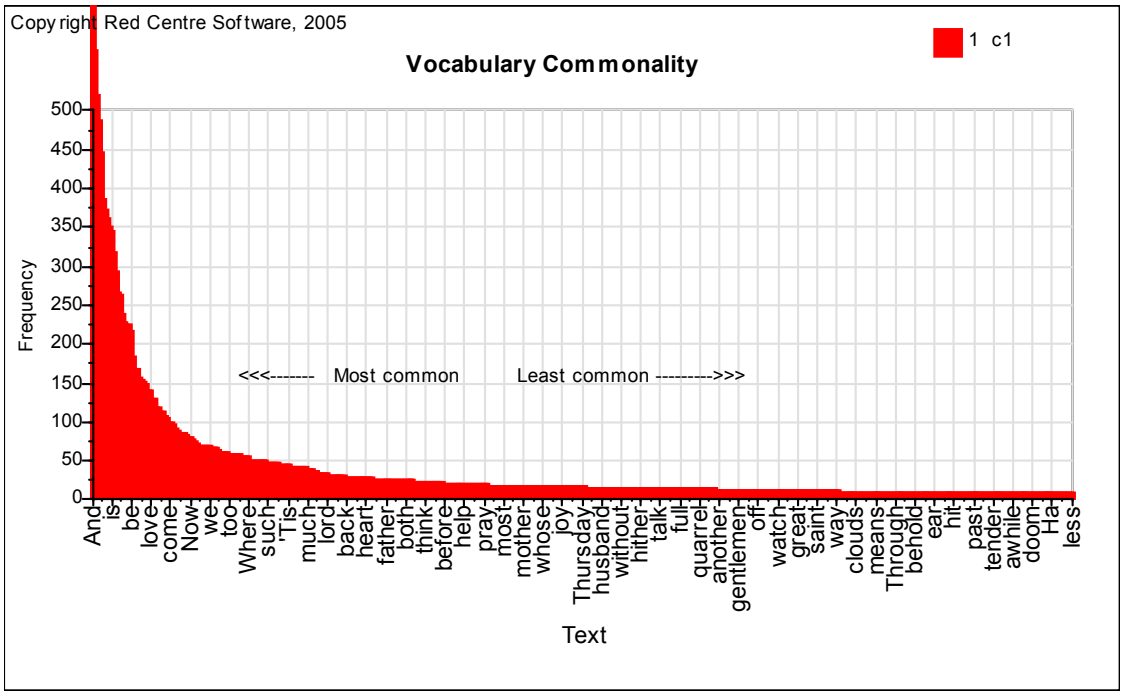
The peaks here are surprisingly symmetrical. A wave pattern intended to facilitate following the plot?

4. Vocabulary

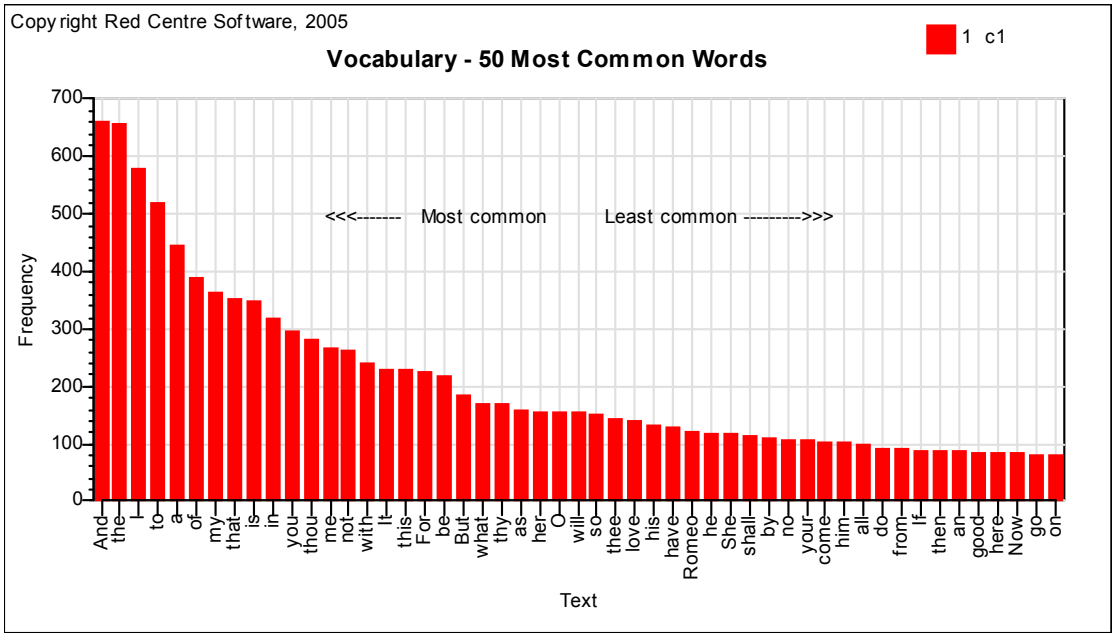
The overall distribution appears regular, with the number of uncommon words mirroring the frequencies of common words. The model is hypothesized as y=(1/x)/n, where large n means more words used less often, that is, a measure of variety.



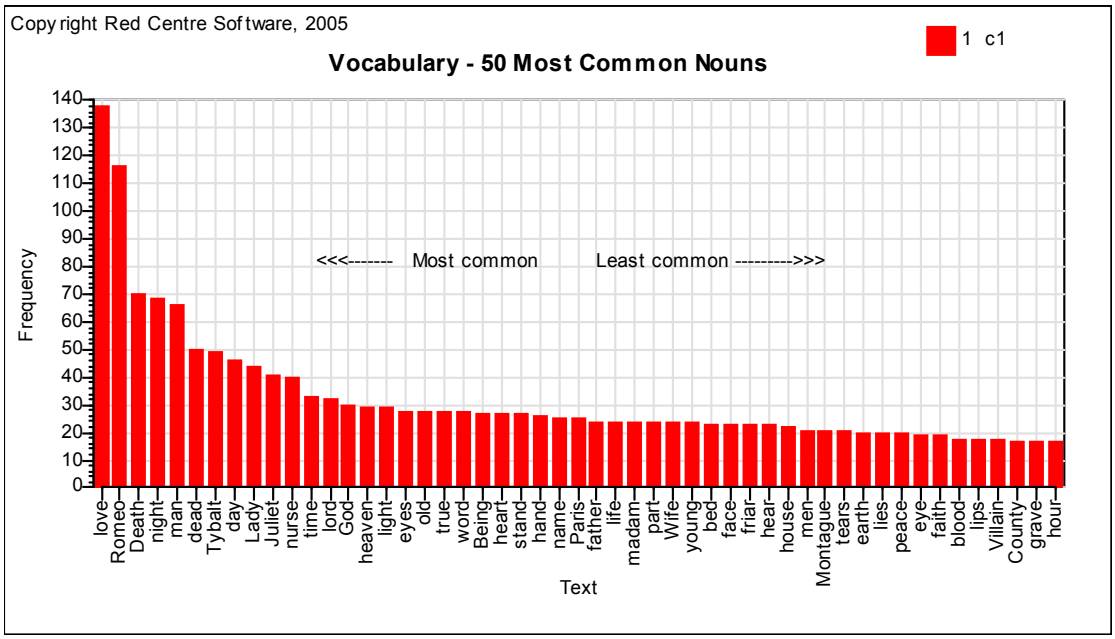
Zooming in on the bottom left corner, allowing 500 points on X and Y1, gives



Showing all first 50 X axis items:

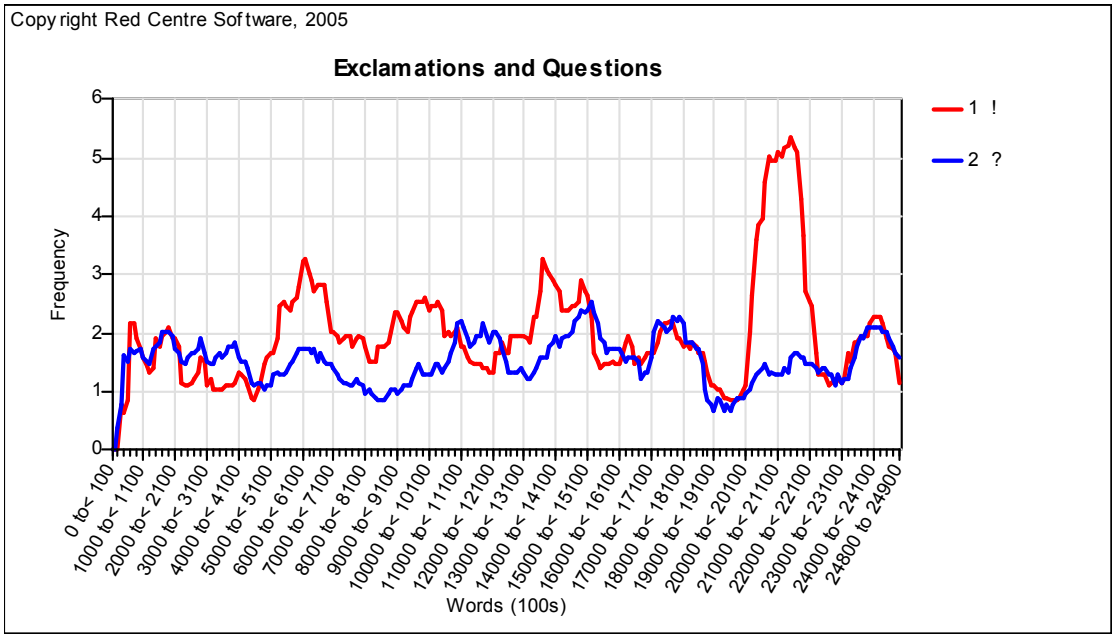


And for nouns only:

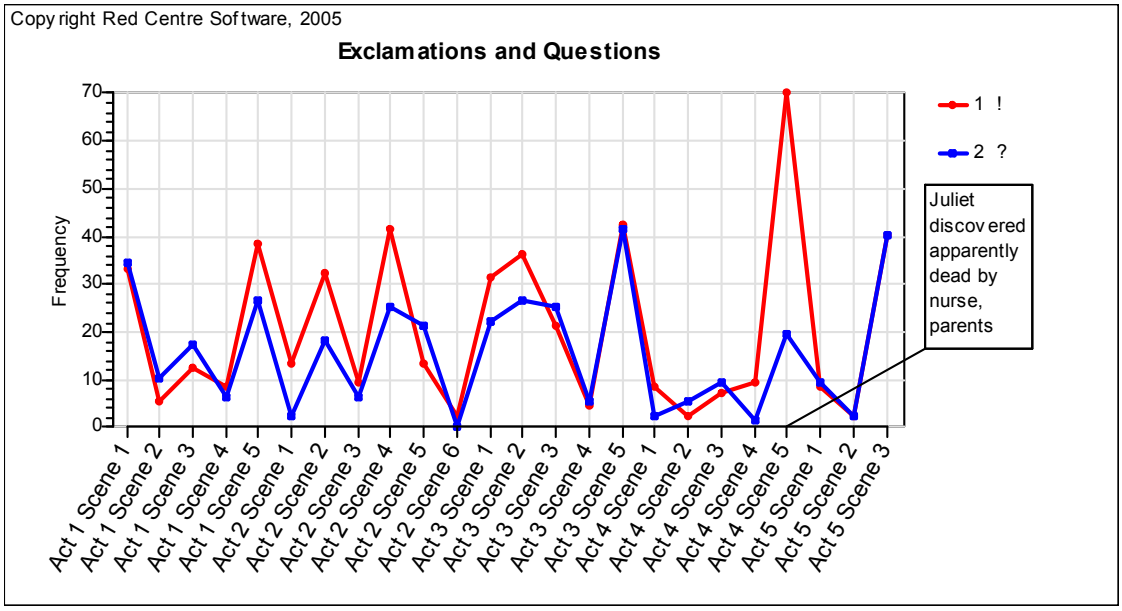


5. Exclamations and Questions

Generally more exclaming than questioning, with a very big blowout in exclamations from 21,000 ff, and a rather strange coincidence from 22,100 ff.



By scene, the distributions are



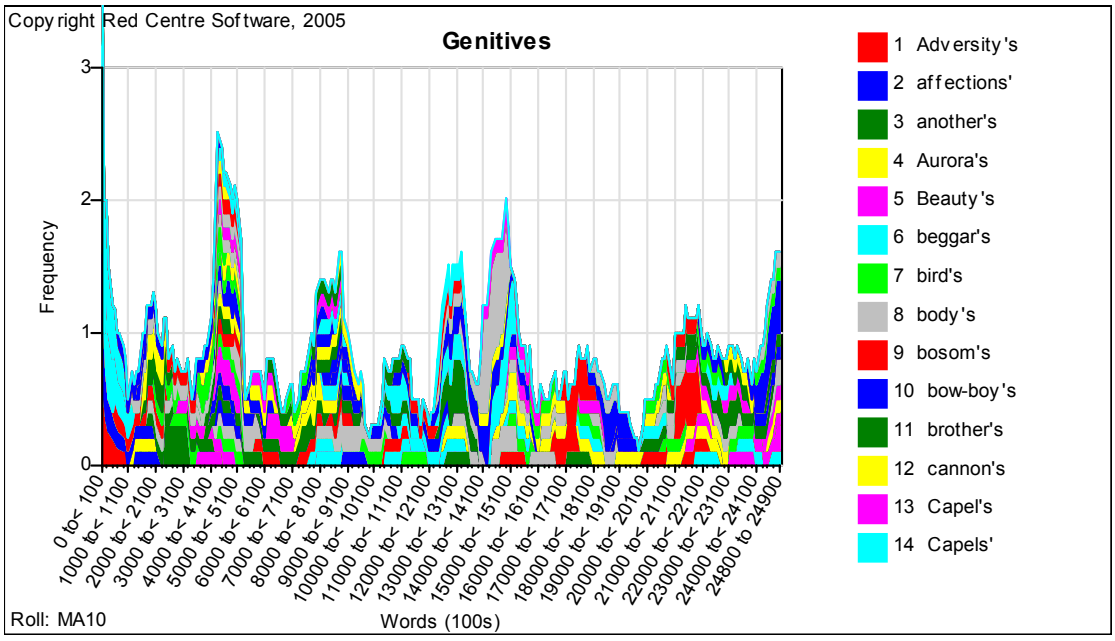
The large increase in exclamations at the discovery of Juliet's apparent corpse is understandable, but why should the number of questions equal the number of exclamations for each scene of Act 5? Hard to accept that this is accidental.

Top: AllScenes
Side: Text

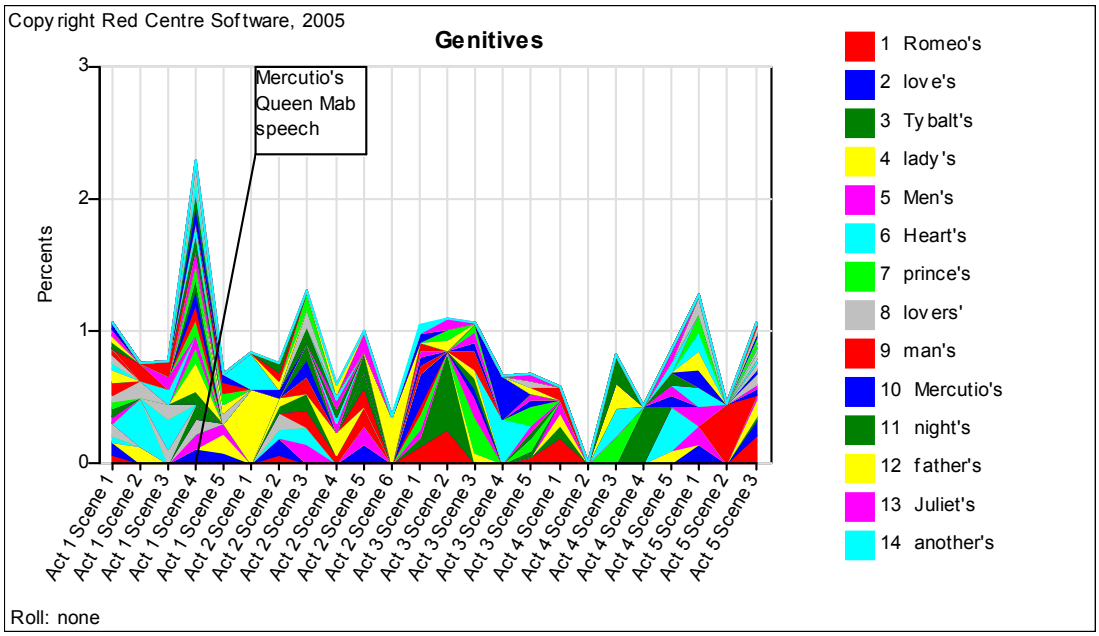
Frequencies		AllScenes			
		Act 4 Scene 5	Act 5 Scene 1	Act 5 Scene 2	Act 5 Scene 3
Text	!	70	8	2	40
	?	19	9	2	40

6. Possessive Apostrophes

Not sure if there is anything to be made of this.

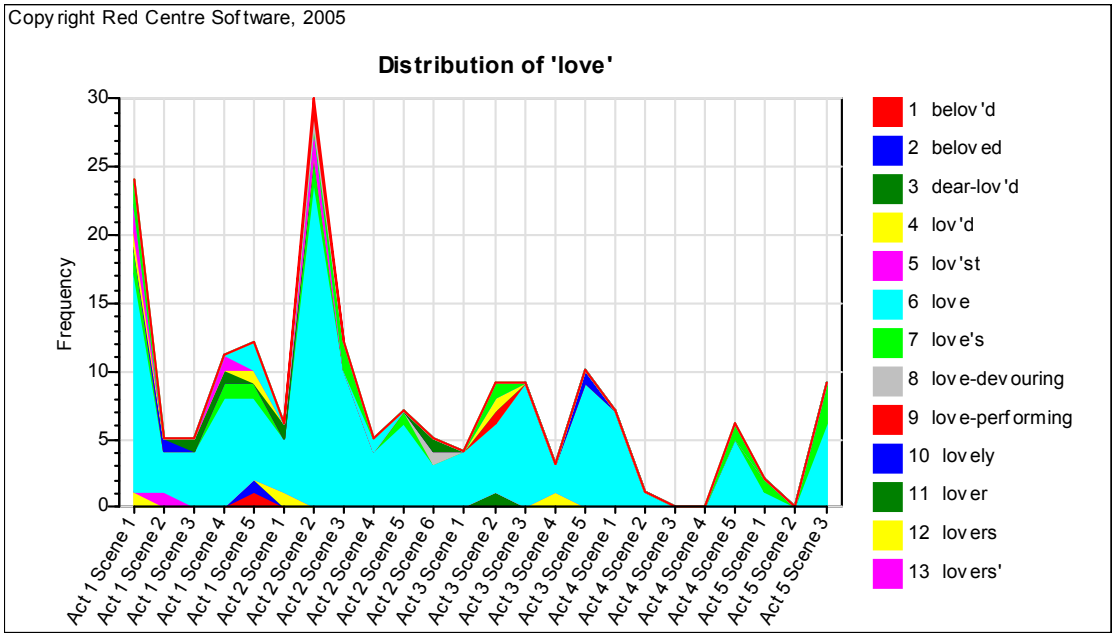


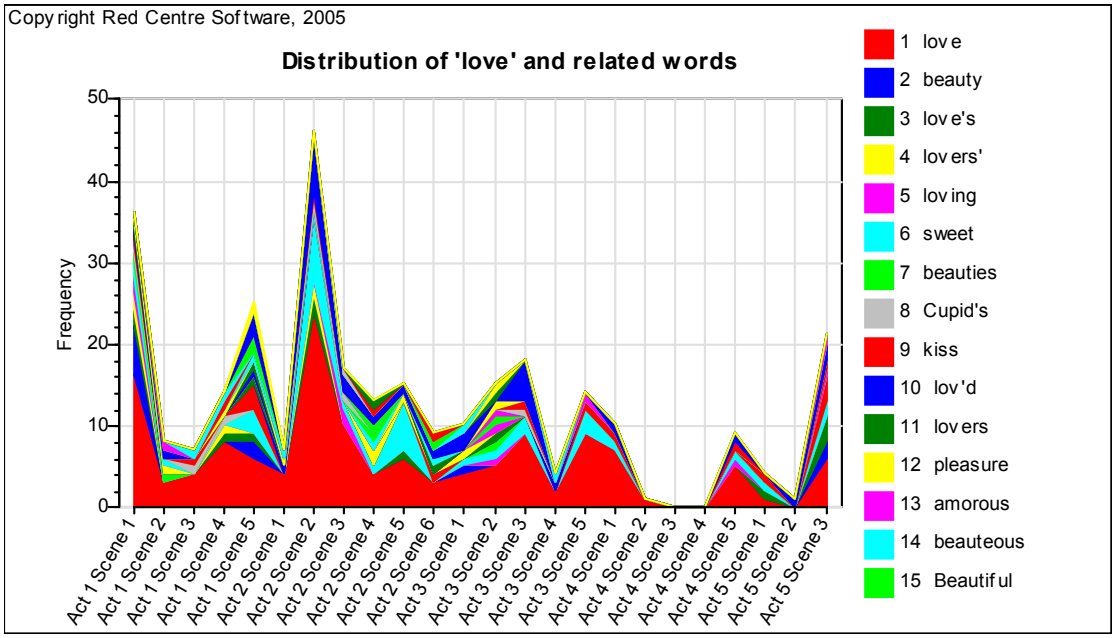
Mercutio's Queen Mab speech has the highest percentage, suggesting careful and precise drafting.



7. Thematic Distributions

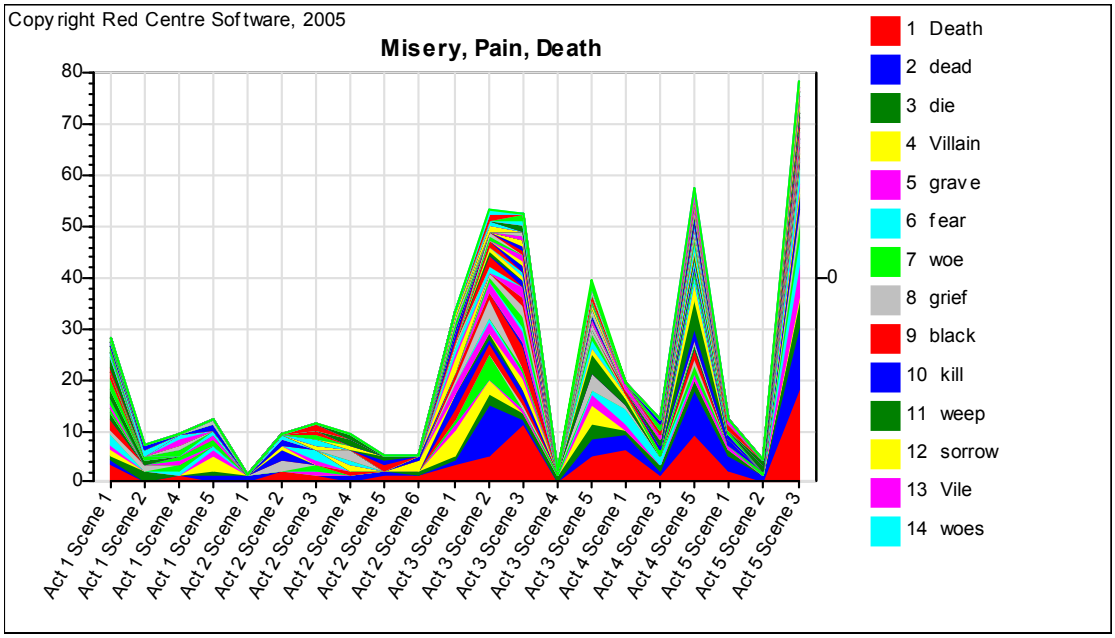
Love





Misery, Pain, Death

As 'love' words drop off towards the end, misery, pain and death words pick up from Act 3 and following, in escalating waves from Act 3 Scene 5.

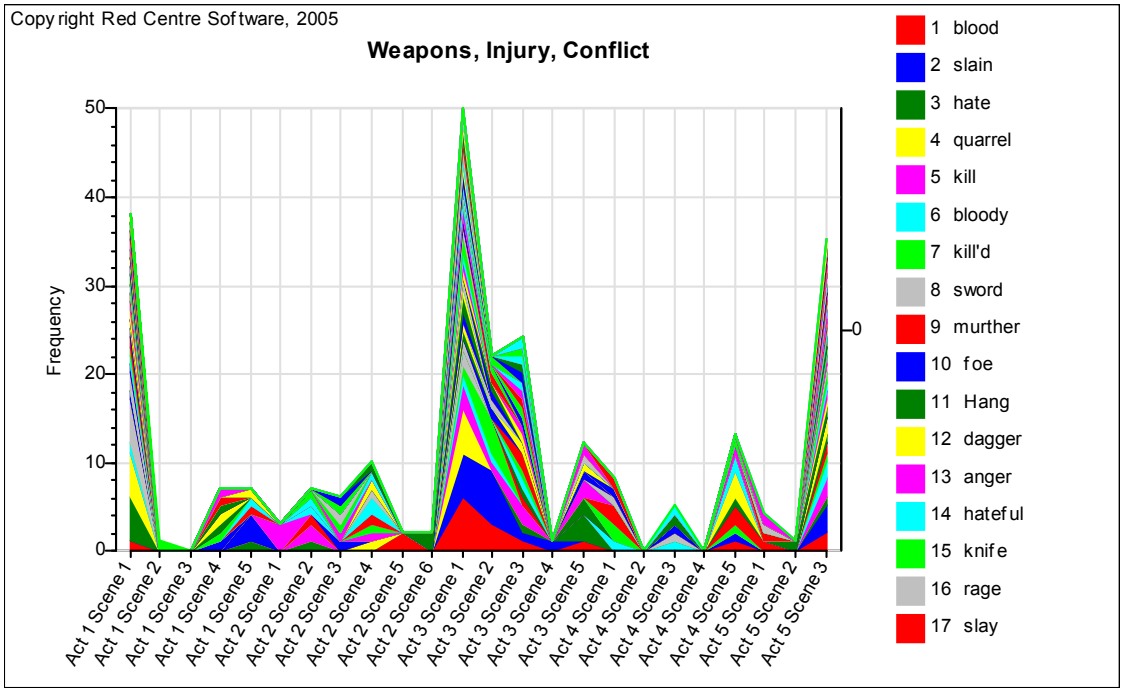


The full list of plotted words, together with number of occurences, is

death	69	dies	3	plagues	1	doleful	1
dead	49	discords	3	sore	1	dreadful	1
die	23	griefs	3	starv'd	1	fearest	1
Villain	17	death's	3	starve	1	fearfully	1
grave	16	dismal	2	starveth	1	fears	1
fear	15	lamentation	2	a-bleeding	1	funeral	1
woe	14	loathed	2	abhorred	1	grieve	1
grief	13	miserable	2	abhors	1	groan'd	1
black	10	misery	2	abus'd	1	groaning	1
kill	10	pains	2	abuse	1	hate's	1
weep	9	sorrows	2	abuses	1	hated	1
sorrow	8	carrion	2	Accurs'd	1	death-darting	1
Vile	8	groan	2	ache	1	death-mark'd	1
woes	8	groans	2	aches	1	deathbed	1
fearful	7	doomsday	2	aching	1	hatred	1
kill'd	7	despair	2	afflicted	1	hideous	1
doom	6	Despised	2	Affliction	1	Killing	1
murther	6	murd'red	2	anguish	1	Despis'd	1
woful	6	murders	2	baleful	1	murder	1
lamentable	5	murtherer	2	calamity	1	Murder'd	1
hell	5	Unhappy	2	canker	1	murdered	1
hateful	4	woeful	2	chaos	1	murderer	1
torture	4	horrible	1	died	1	torment	1
weeping	4	lament	1	dire	1	tormented	1
weeps	4	melancholy	1	direful	1	Torments	1
loathsome	3	misfortune	1	dirges	1	villanous	1
pain	3	misfortune's	1	displeas'd	1	wail	1
plague	3	moans	1	displeasure	1	wailing	1
deadly	3	pestilence	1	distraught	1	weep'st	1
devil	3	pestilent	1	distressed	1		

Weapons, Injury, Conflict

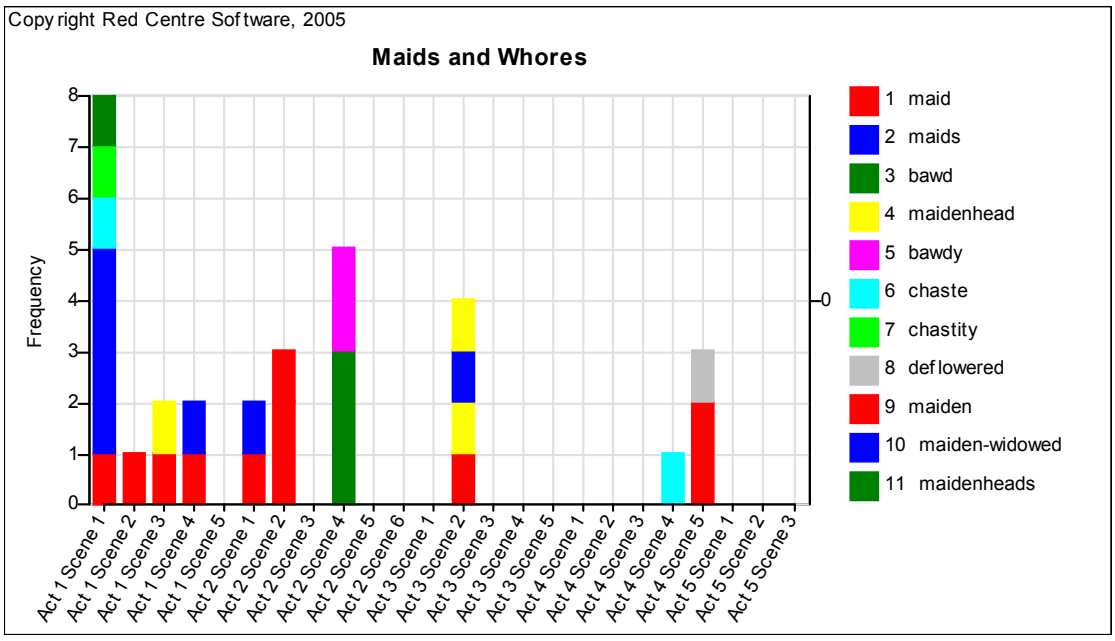
The distribution of Weapons, Injury and Conflict words is remarkably symmetrical.



The full list of words with frequencies is

blood	17	detestable	2	axe	1	hanging	1
slain	17	duellist	2	battlements	1	hate's	1
hate	12	enmity	2	blades	1	hated	1
quarrel	11	fights	2	blame	1	injur'd	1
kill	10	hangs	2	bleeding	1	injuries	1
bloody	7	mangled	2	bleeds	1	Killing	1
kill'd	7	murd'red	2	brawling	1	mangle	1
sword	7	murders	2	cannon's	1	mistempered	1
murther	6	murtherer	2	chariot	1	murder	1
foe	5	outrage	2	choking	1	Murder'd	1
Hang	5	quarrelling	2	club	1	murdered	1
dagger	5	slander	2	Clubs	1	murderer	1
anger	4	slaught'red	2	dismemb'red	1	punish'd	1
hateful	4	temper	2	dispute	1	punished	1
knife	4	violent	2	distemp'rature	1	quarrell'd	1
rage	4	weapon	2	distempered	1	quarrels	1
slay	4	weapons	2	fierce	1	raging	1
blows	3	rapier	2	fight	1	railest	1
charge	3	whip	2	fighting	1	rancour	1
fury	3	a-bleeding	1	foe's	1	revenge	1
slaughter'd	3	adversary	1	foes	1	sland'red	1
swords	3	Adversity's	1	furious	1	slays	1
arm'd	2	anger'd	1	gore-blood	1	stabb'd	1
Arms	2	angry	1	gory	1	violence	1
blade	2	argues	1	grievance	1	violently	1
blow	2	argument	1	grievances	1	rapier's	1
brawl	2	armour	1	grudge	1	Whipp'd	1
brawls	2	arrow	1	hang'd	1		

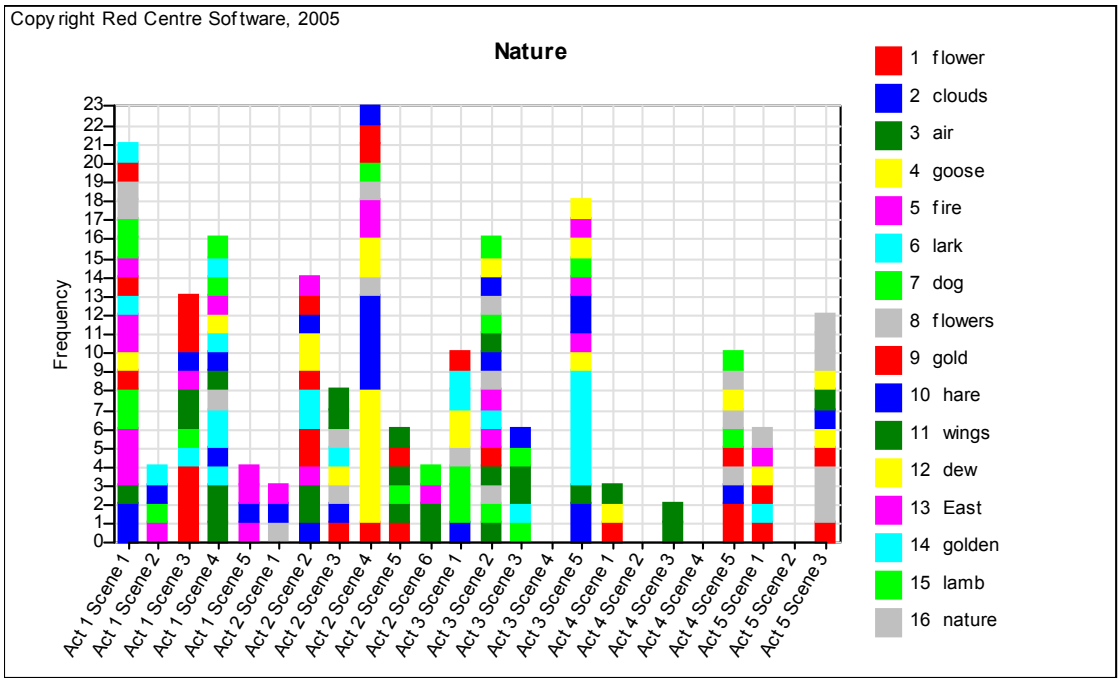
Maidens and Whores



Words and frequencies are

maid	10	deflowered	1
maids	6	maiden	1
bawd	3	maiden-widowed	1
maidenhead	2	maidenheads	1
bawdy	1	maidenhoods	1
chaste	1	whore	1
chastity	1	whoreson	1

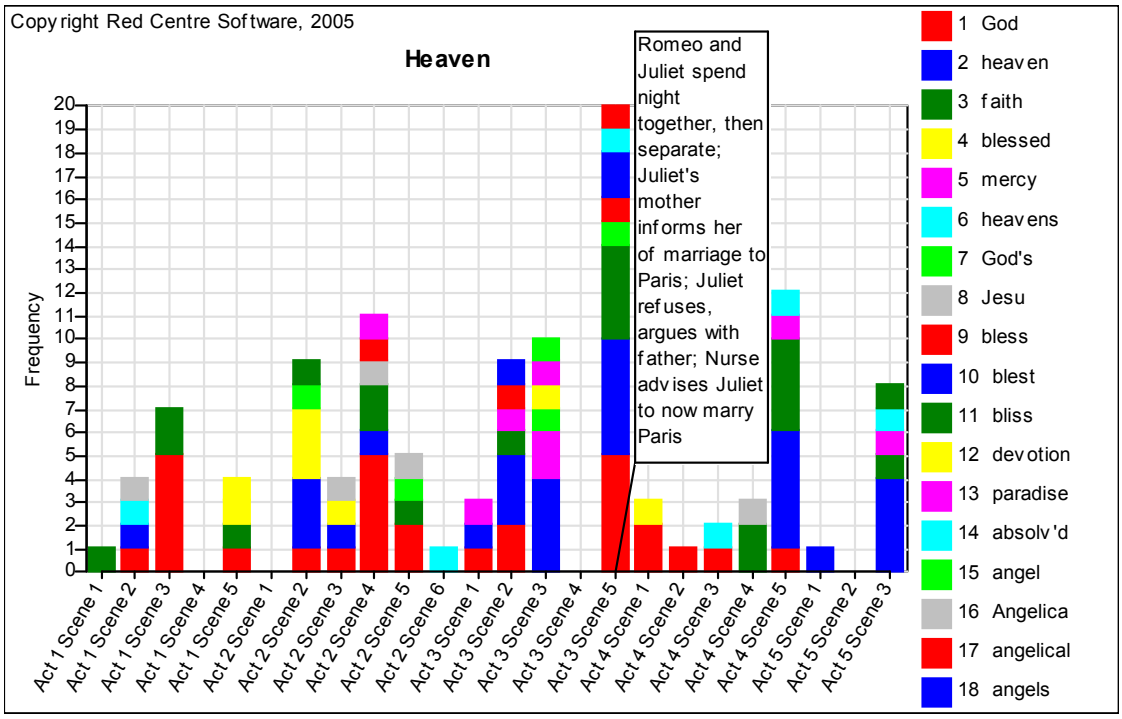
Nature



Word list and frequencies is

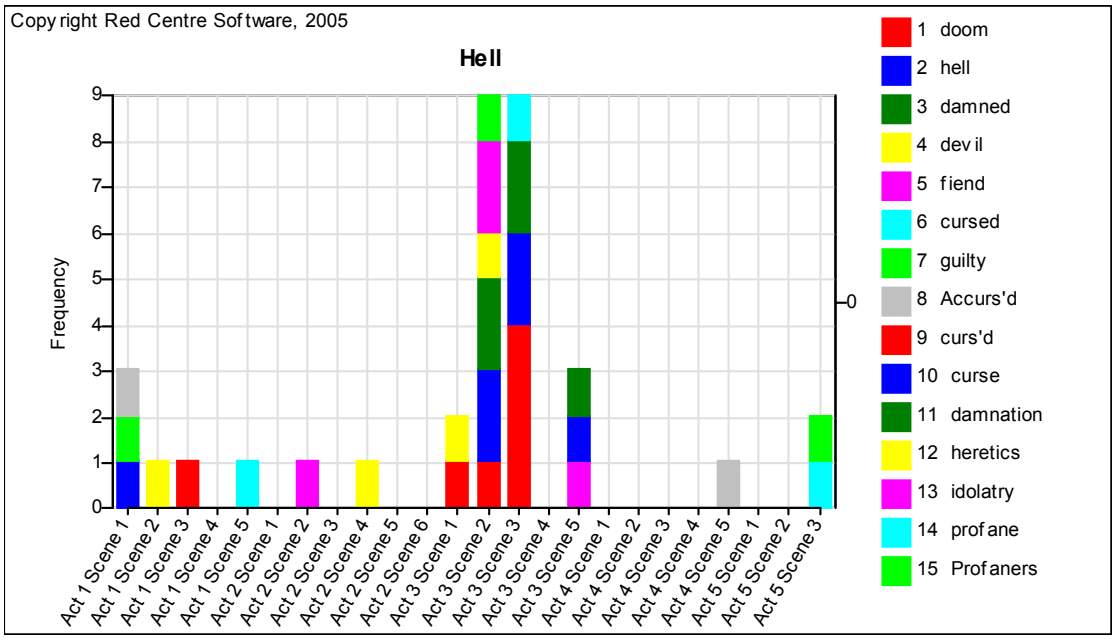
flower	9	dove	2	bird's	1	mountain	1
clouds	8	dovehouse	2	birds	1	pomegranate	1
air	7	egg	2	cloudy	1	rat	1
goose	7	fish	2	dew-dropping	1	raven	1
fire	6	horses	2	dog's	1	raven's	1
lark	6	leaves	2	Dove-feather'd	1	rose	1
dog	5	nature's	2	doves	1	serpent	1
flowers	5	roses	2	dragon	1	serpents	1
gold	5	stone	2	Eastern	1	snow	1
hare	5	stones	2	evening	1	snowy	1
wings	5	stony	2	fir'd	1	spider's	1
dew	4	storm	2	fire-ey'd	1	thorn	1
East	4	wind	2	fires	1	tigers	1
golden	4	winds	2	fishes	1	well-flower'd	1
lamb	4	worm	2	fishified	1	wild-goose	1
nature	4	wormwood	2	flow'r	1	wind-swift	1
airy	3	afire	1	flow'ring	1	wind:	1
bark	3	agate	1	gnat	1	winged	1
beast	3	alligator	1	grasshoppers	1	Winter	1
rosemary	3	ape	1	grove	1	wolvish-ravening	1
toad	3	aqua	1	grub	1	Worms	1
bird	2	aqua-vitae	1	grubs	1	worms'	1
cat	2	Aurora's	1	ladybird	1		
Cats	2	beasts	1	leaf	1		
cave	2	beetle	1	mandrakes	1		

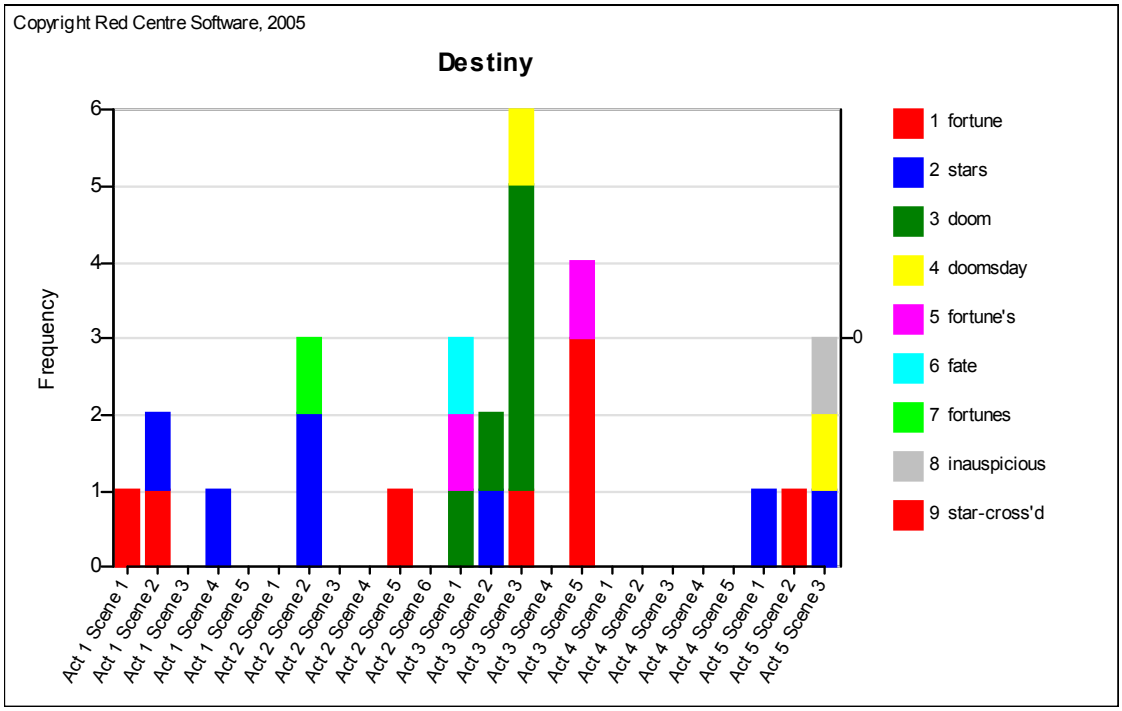
Heaven, Hell, Destiny



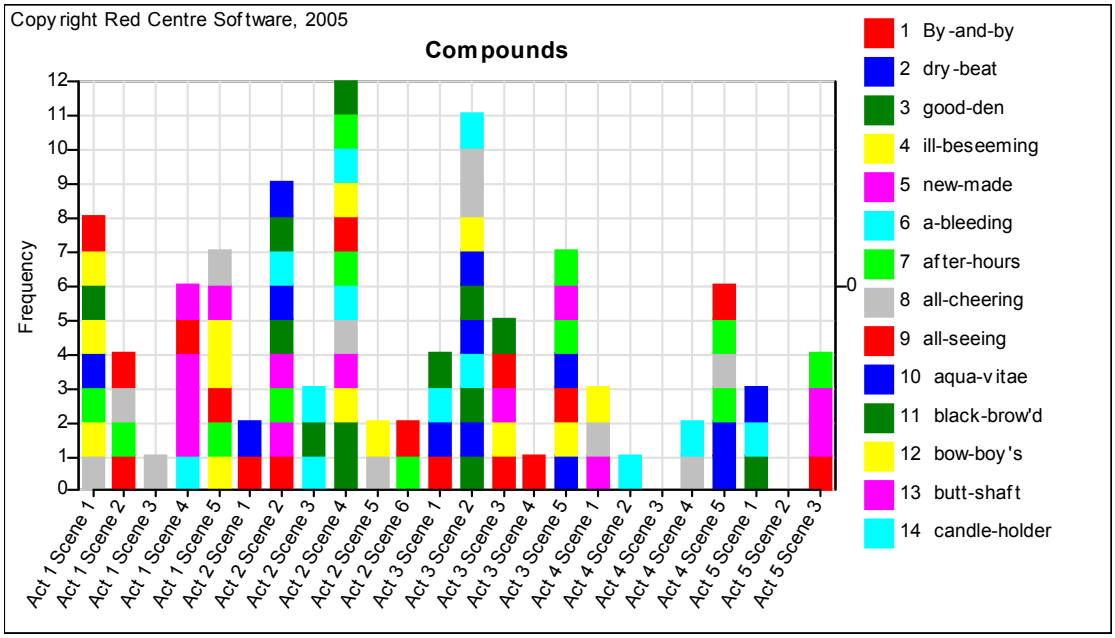
Frequencies are

God	29	absolv'd	1
heaven	28	angel	1
faith	18	Angelica	1
blessed	5	angelical	1
mercy	5	angels	1
heavens	4	baptiz'd	1
God's	3	blessing	1
Jesu	3	blessings	1
bless	2	devout	1
blest	2	divine	1
bliss	2	God-den	1
devotion	2	God-i-god-en	1
paradise	2	heavenly	1




Copyright Red Centre Software, 2005
Destiny
Frequency
1 fortune
2 stars
3 doom
4 doomsday
5 fortune's
6 fate
7 fortunes
8 inauspicious
9 star-cross'd

8. Compounds


Copyright Red Centre Software, 2005
Compounds
Frequency
1 By-and-by
2 dry-beat
3 good-den
4 ill-beseeming
5 new-made
6 a-bleeding
7 after-hours
8 all-cheering
9 all-seeing
10 aqua-vitae
11 black-brow'd
12 bow-boy's
13 butt-shaft
14 candle-holder

The full list of compounds is

By-and-by	field-bed	love-performing	soon-speeding
dry-beat	fiery-footed	loving-jealous	star-cross'd
good-den	fire-ey'd	maiden-widowed	Still-waking
ill-beseeming	five-and-twenty	Mist-like	tallow-face
new-made	flattering-sweet	mouse-hunt	tassel-gentle
a-bleeding	flirt-gills	neighbour-stained	tempest-tossed
after-hours	fruit-tree	nimble-pinion'd	three-hours
all-cheering	go-den	pardona-mi's	tithe-pig's
all-seeing	God-den	poor-John	truckle-bed
aqua-vitae	God-i-god-en	precious-juiced	true-love
black-brow'd	gore-blood	saint-seducing	two-and-forty
bow-boy's	green-sickness	savage-wild	unlook'd-for
butt-shaft	grey-coated	self-will'd	up-fill
candle-holder	grey-ey'd	serving-creature	well-flower'd
cock-a-hoop	hard-hearted	serving-creature's	well-govern'd
cot-quean	high-lone	sharp-ground	well-seeming
court-cubbert	hunt's-up	silver-sweet	white-upturned
dear-lov'd	ill-divining	sin-absolver	wild-goose
death-darting	ill-shaped	single-sold	wind-swift
death-mark'd	join-stools	sir-reverence	without-book
dew-dropping	lazy-pacing	skains-mates	wolvish-ravening
Dove-feather'd	life-weary	slug-abed	world-wearied
Earth-treading	long-experienc'd	sober-suited	
fashion-mongers	love-devouring	son-in-law	

[end of document]