Red Centre Software
Data Comes Alive

Quantitative Text Analysis of Wuthering Heights

Version: May 2008

Quantitative Text Analysis of Wuthering Heights

This document outlines a quantitative approach to the text of Wuthering Heights, using cross tabulation and time series techniques.

PREPARING THE TEXT

The text was downloaded from Project Gutenberg:

http://www.gutenberg.org/etext/768

For text preparation

- Remove all peripheral text, such as the Gutenberg licence
- Remove all non-text lines, such as *********
- Save as WutheringHeightsRaw.txt
- Run this script:

```
Sub Main

   Set rub = CreateObject("Ruby.App1")

   path = "D:\RubyData\WuthHghts\Source\"
   rub.System "RemoveBlankLines", path&"WutheringHeightsRaw.txt",_
                                  path&"WuthHghts.txt"
   rub.System "ImportTextAsMulti", path&”WuthHghts.txt",_
                                   path&"WuthHghts"

End Sub
```

This script creates the variable *WuthHghts*.

The *ImportTextAsMulti* method codes each word, then works line by line, replacing each word with its code. For example, input of

```
“The fat cat sat on the mat”
```

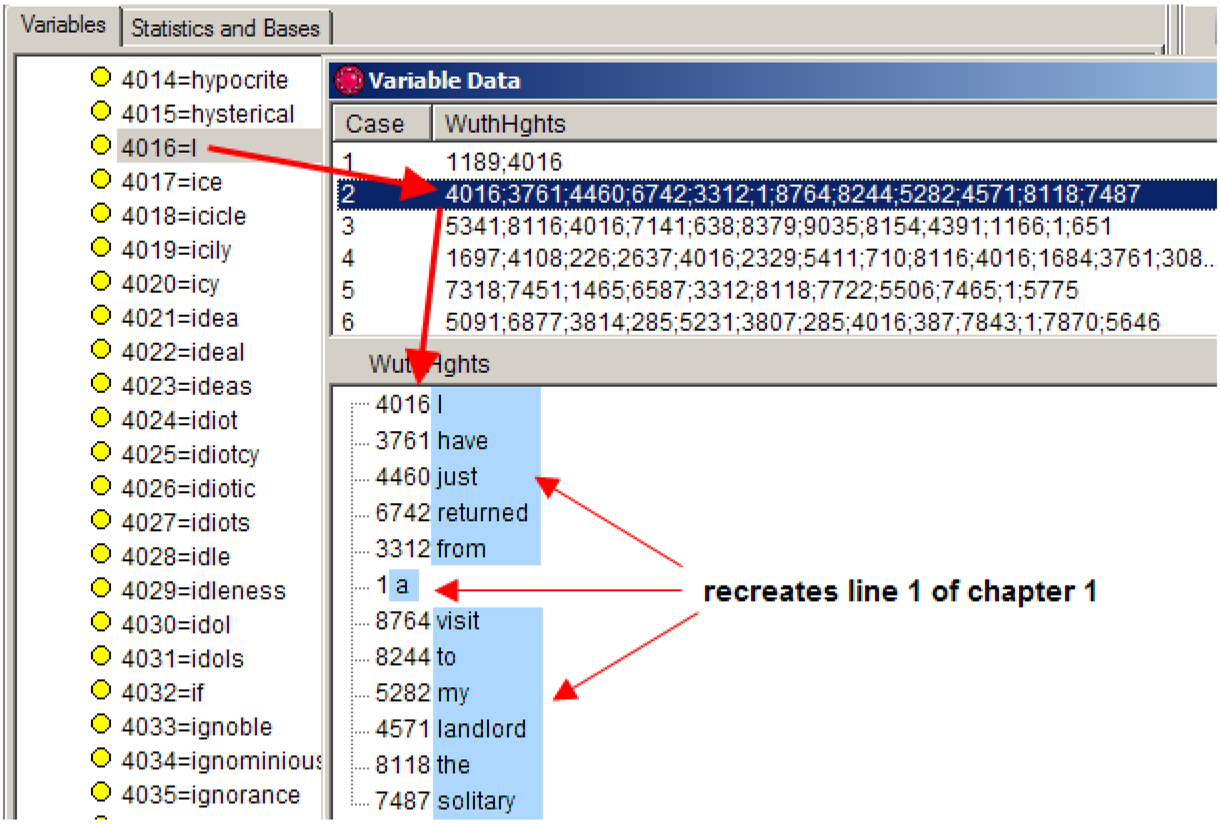
On a codeframe of

```
1=cat
2=fat
3=mat
4=on
5=sat
6=the
```

is coded as 6;2;1;5;4;6;3.

The first line of chapter 1 is represented as

```
4016;3761;4460;6742;3312;1;8764;8244;5282;4571;8118;7487
```



Variables | Statistics and Bases

○ 4014=hypocrite
○ 4015=hysterical
○ 4016=I
○ 4017=ice
○ 4018=icicle
○ 4019=icily
○ 4020=icy
○ 4021=idea
○ 4022=ideal
○ 4023=ideas
○ 4024=idiot
○ 4025=idiotcy
○ 4026=idiotic
○ 4027=idiots
○ 4028=idle
○ 4029=idleness
○ 4030=idol
○ 4031=idols
○ 4032=if
○ 4033=ignoble
○ 4034=ignominious
○ 4035=ignorance

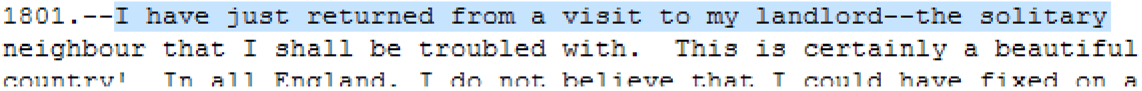
Variable Data

Case	WuthHghts
1	1189;4016
2	4016;3761;4460;6742;3312;1;8764;8244;5282;4571;8118;7487
3	5341;8116;4016;7141;638;8379;9035;8154;4391;1166;1;651
4	1697;4108;226;2637;4016;2329;5411;710;8116;4016;1684;3761;308..
5	7318;7451;1465;6587;3312;8118;7722;5506;7465;1;5775
6	5091;6877;3814;285;5231;3807;285;4016;387;7843;1;7870;5646

WuthHghts

4016	I
3761	have
4460	just
6742	returned
3312	from
1	a
8764	visit
8244	to
5282	my
4571	landlord
8118	the
7487	solitary

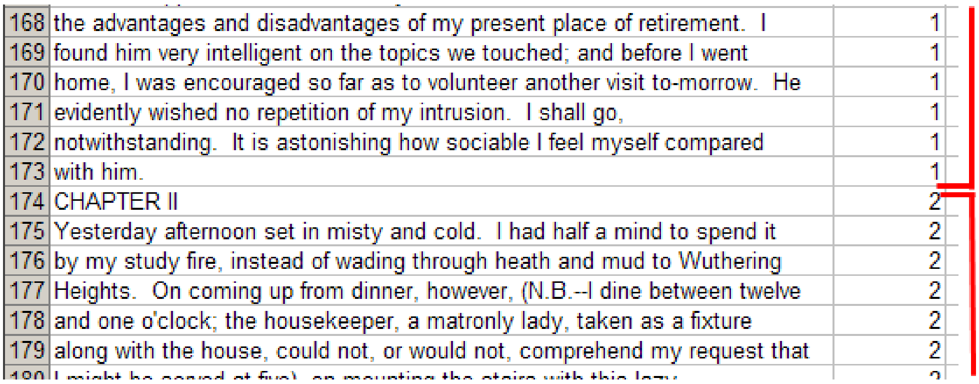
recreates line 1 of chapter 1



```
1801.--I have just returned from a visit to my landlord--the solitary
neighbour that I shall be troubled with.  This is certainly a beautiful
country!  In all England, I do not believe that I could have fixed on a
```

The complete text has thus been rendered as a set of multi-response items, stored in the variable *WuthHghts*, where each case (a *respondent* in survey terms) is a line of the text. Note that this process eliminates all punctuation.

Coding the lines by chapter was done manually in Excel. Searching for the text "Chapter" quickly isolated the boundary points. Excel's drag-fill feature was used to enter the chapter number against each line.



168	the advantages and disadvantages of my present place of retirement. I	1
169	found him very intelligent on the topics we touched; and before I went	1
170	home, I was encouraged so far as to volunteer another visit to-morrow. He	1
171	evidently wished no repetition of my intrusion. I shall go,	1
172	notwithstanding. It is astonishing how sociable I feel myself compared	1
173	with him.	1
174	CHAPTER II	2
175	Yesterday afternoon set in misty and cold. I had half a mind to spend it	2
176	by my study fire, instead of wading through heath and mud to Wuthering	2
177	Heights. On coming up from dinner, however, (N.B.--I dine between twelve	2
178	and one o'clock; the housekeeper, a matronly lady, taken as a fixture	2
179	along with the house, could not, or would not, comprehend my request that	2
180	I might be served at five) on mounting the stairs with this lazy	2

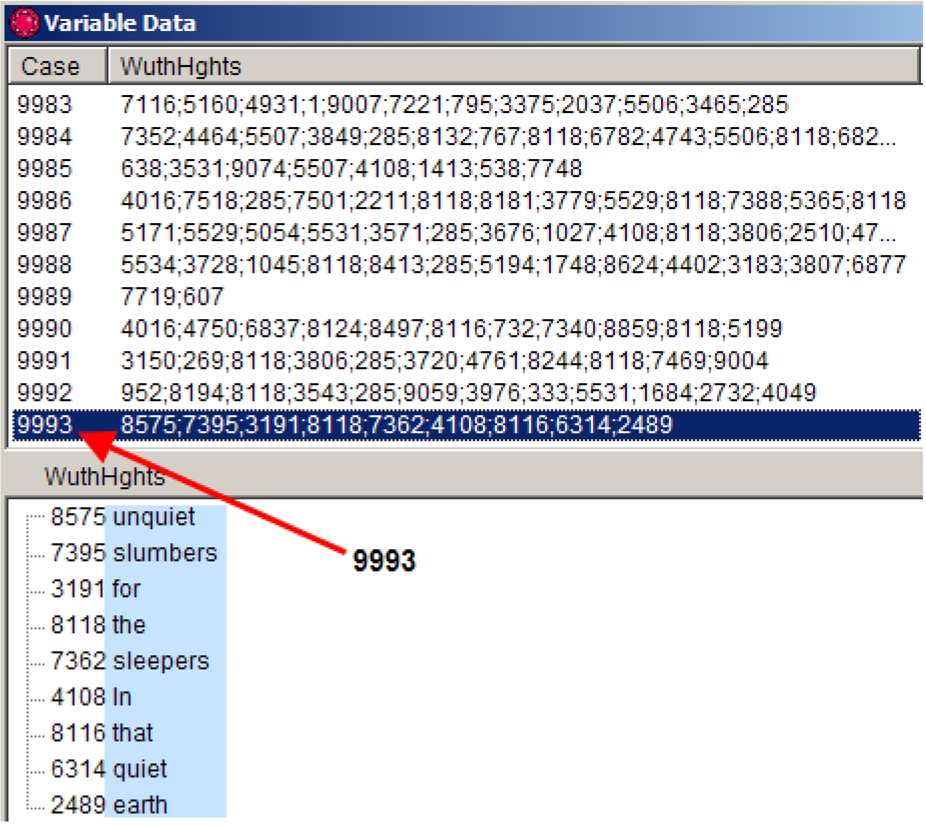
end of Chapter 1

start of chapter 2

The column was then copied to a text editor, and saved as Chapter.cd. The matching Chapter.met was created manually, and then coded using the Edit Variable form.

Basic Statistics

Number of lines, including the 34 chapter headings: 9993



Case	WuthHghts
9983	7116;5160;4931;1;9007;7221;795;3375;2037;5506;3465;285
9984	7352;4464;5507;3849;285;8132;767;8118;6782;4743;5506;8118;682...
9985	638;3531;9074;5507;4108;1413;538;7748
9986	4016;7518;285;7501;2211;8118;8181;3779;5529;8118;7388;5365;8118
9987	5171;5529;5054;5531;3571;285;3676;1027;4108;8118;3806;2510;47...
9988	5534;3728;1045;8118;8413;285;5194;1748;8624;4402;3183;3807;6877
9989	7719;607
9990	4016;4750;6837;8124;8497;8116;732;7340;8859;8118;5199
9991	3150;269;8118;3806;285;3720;4761;8244;8118;7469;9004
9992	952;8194;8118;3543;285;9059;3976;333;5531;1684;2732;4049
9993	8575;7395;3191;8118;7362;4108;8116;6314;2489

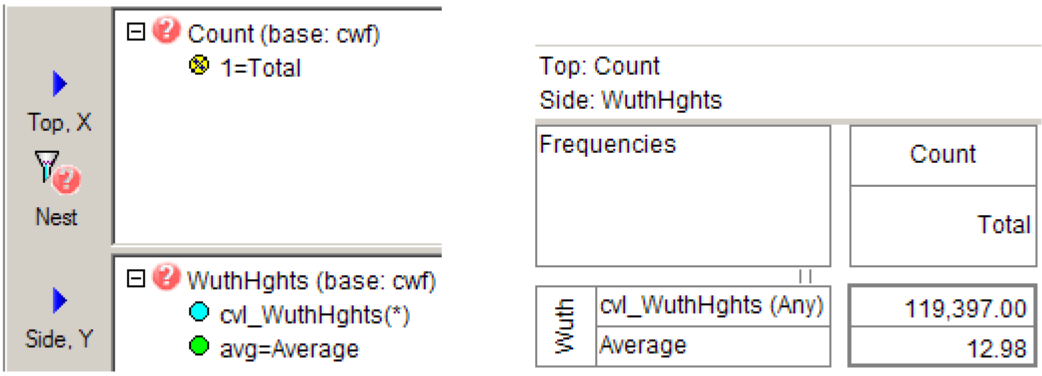
WuthHghts

8575	unquiet
7395	slumbers
3191	for
8118	the
7362	sleepers
4108	In
8116	that
6314	quiet
2489	earth

9993

Total number of words, using **c**ount of **val**ues cvl_: 119,397

Average number of words per line, using pseudo-code avg: 12.98



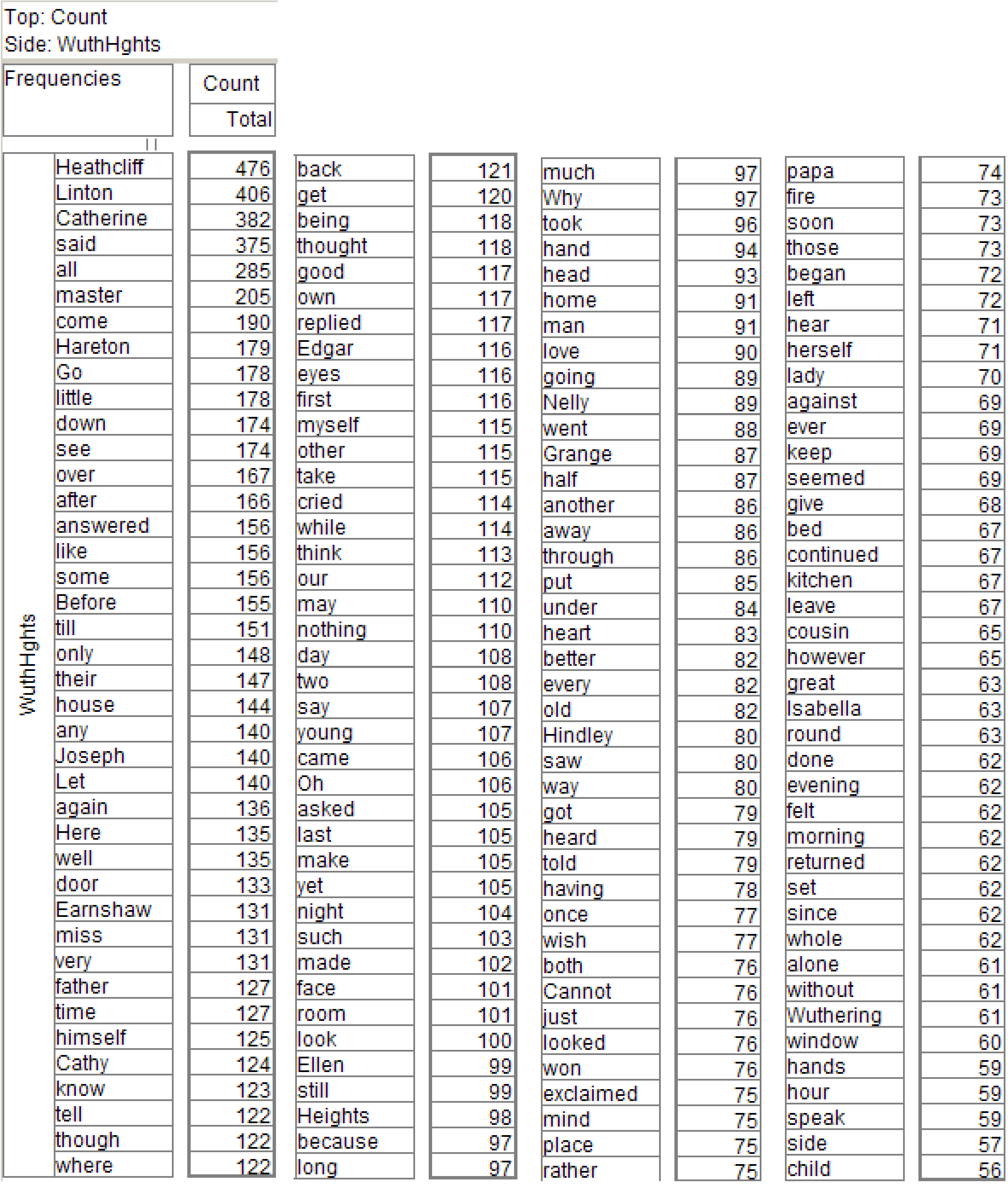
Count (base: cwf)
1=Total

WuthHghts (base: cwf)
cvl_WuthHghts(*)
avg=Average

Top: Count
Side: WuthHghts

Frequencies		Count Total
Wuth	cvl_WuthHghts (Any)	119,397.00
	Average	12.98

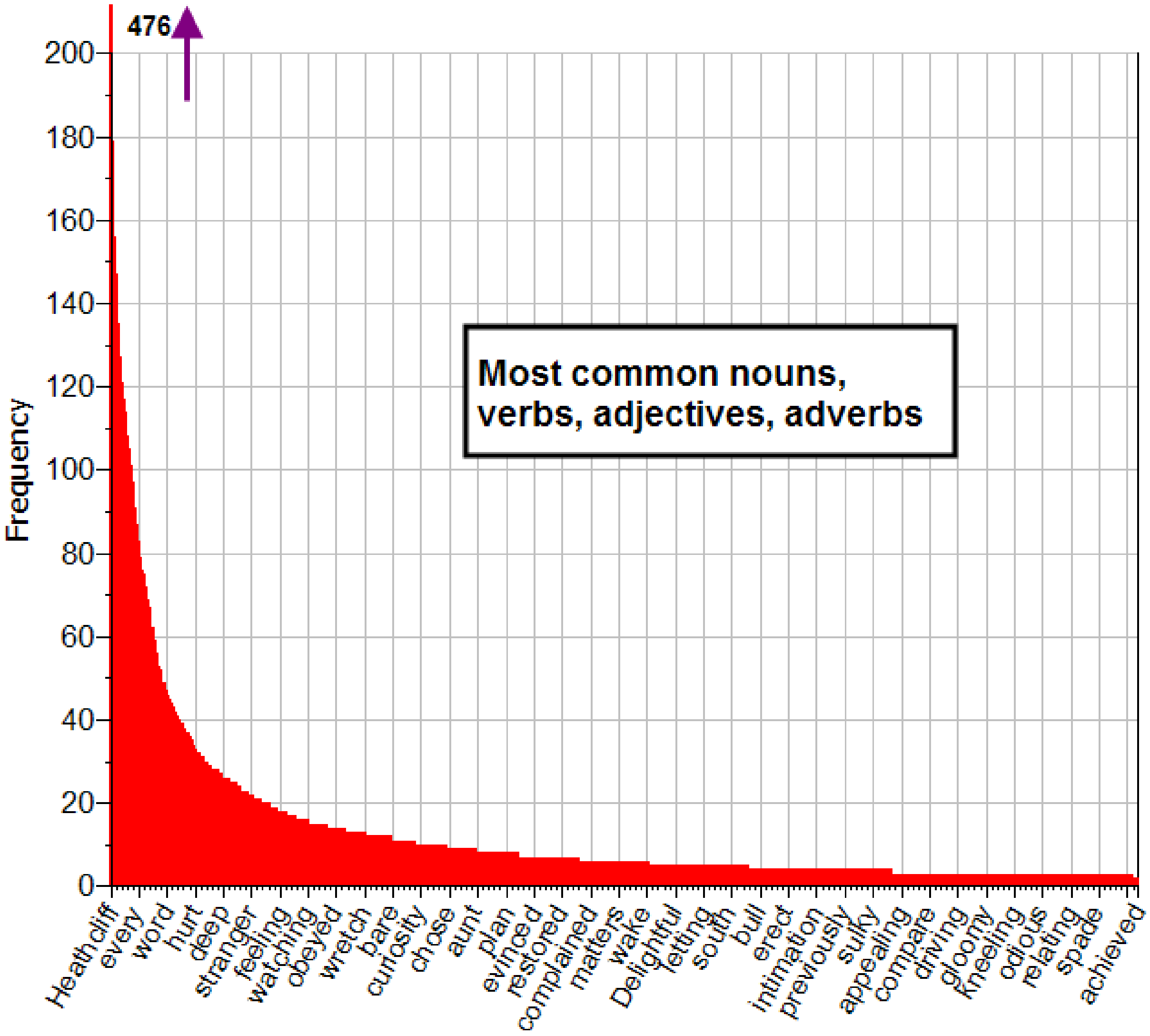
Most frequent words (first 160):



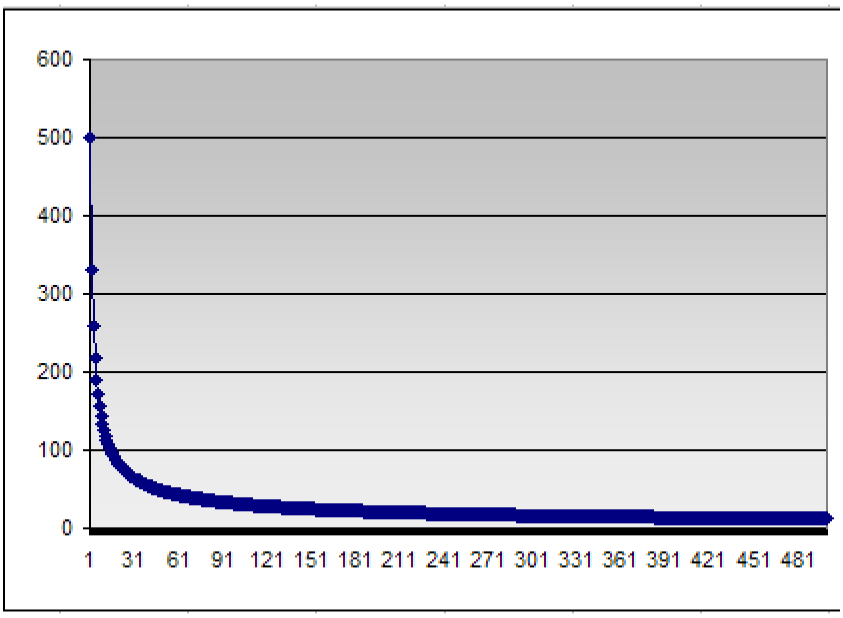
Top: Count
Side: WuthHghts

	Frequencies	Count
		Total
WuthHghts	Heathcliff	476
	Linton	406
	Catherine	382
	said	375
	all	285
	master	205
	come	190
	Hareton	179
	Go	178
	little	178
	down	174
	see	174
	over	167
	after	166
	answered	156
	like	156
	some	156
	Before	155
	till	151
	only	148
	their	147
	house	144
	any	140
	Joseph	140
	Let	140
	again	136
	Here	135
	well	135
	door	133
	Earnshaw	131
	miss	131
	very	131
	father	127
	time	127
	himself	125
	Cathy	124
	know	123
	tell	122
	though	122
	where	122
	back	121
	get	120
	being	118
	thought	118
	good	117
	own	117
	replied	117
	Edgar	116
	eyes	116
	first	116
	myself	115
	other	115
	take	115
	cried	114
	while	114
	think	113
	our	112
	may	110
	nothing	110
	day	108
	two	108
	say	107
	young	107
	came	106
	Oh	106
	asked	105
	last	105
	make	105
	yet	105
	night	104
	such	103
	made	102
	face	101
	room	101
	look	100
	Ellen	99
	still	99
	Heights	98
	because	97
	long	97
	much	97
	Why	97
	took	96
	hand	94
	head	93
	home	91
	man	91
	love	90
	going	89
	Nelly	89
	went	88
	Grange	87
	half	87
	another	86
	away	86
	through	86
	put	85
	under	84
	heart	83
	better	82
	every	82
	old	82
	Hindley	80
	saw	80
	way	80
	got	79
	heard	79
	told	79
	having	78
	once	77
	wish	77
	both	76
	Cannot	76
	just	76
	looked	76
	won	76
	exclaimed	75
	mind	75
	place	75
	rather	75
	papa	74
	fire	73
	soon	73
	those	73
	began	72
	left	72
	hear	71
	herself	71
	lady	70
	against	69
	ever	69
	keep	69
	seemed	69
	give	68
	bed	67
	continued	67
	kitchen	67
	leave	67
	cousin	65
	however	65
	great	63
	Isabella	63
	round	63
	done	62
	evening	62
	felt	62
	morning	62
	returned	62
	set	62
	since	62
	whole	62
	alone	61
	without	61
	Wuthering	61
	window	60
	hands	59
	hour	59
	speak	59
	side	57
	child	56

As a sorted distribution plot:



I am still attempting to discern the model for this. Y=500/x^0.6 is not too bad:

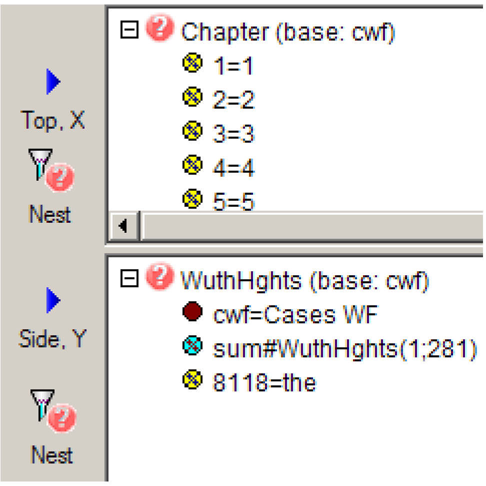


Maybe a log function would be better.

SOME INTERESTING CHARTS

Distribution of Articles

A chart which counts the number of occurances of *a* or *an* and *the* within each chapter against the total number of lines in each chapter can be specified as



Top, X

Nest

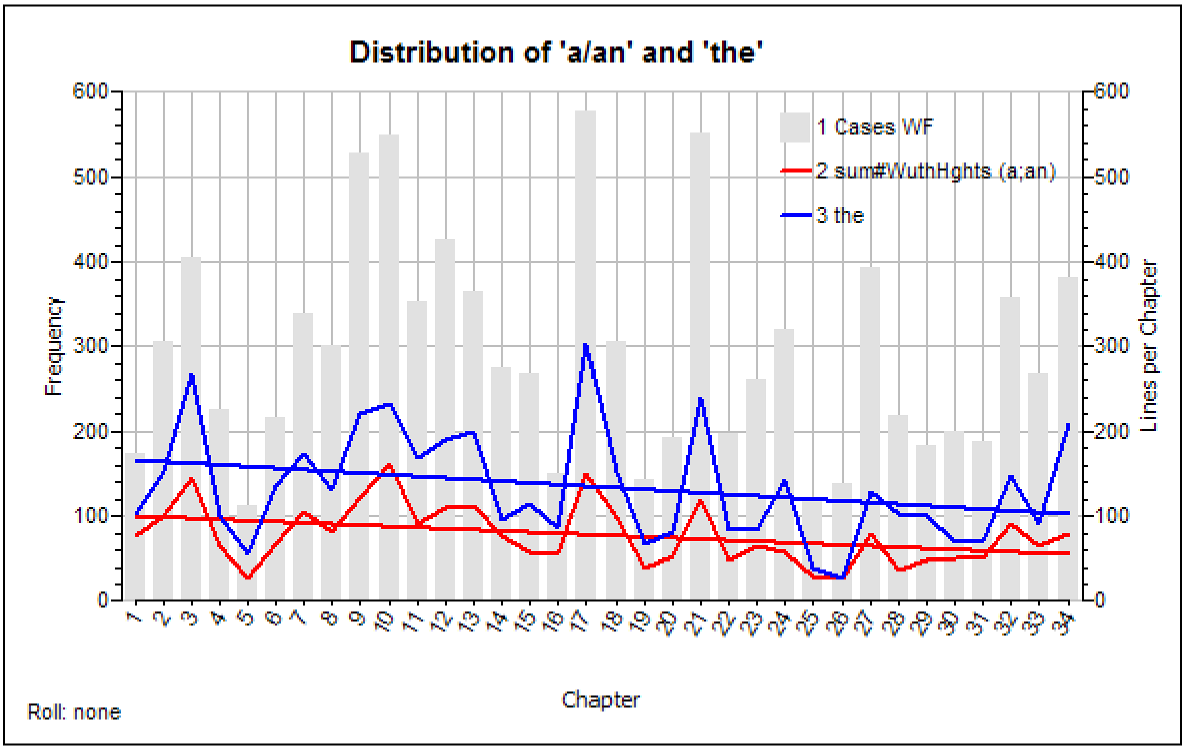
Chapter (base: cwf)
- 1=1
- 2=2
- 3=3
- 4=4
- 5=5

Side, Y

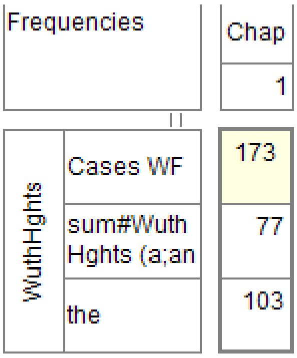
Nest

WuthHghts (base: cwf)
- cwf=Cases WF
- sum#WuthHghts(1;281)
- 8118=the

Putting the base (cwf = Cases Weighted Filtered) on Y2 as bars allows the proportions to be visually estimated, by inspecting how far up a bar the line series cross (halfway=50%).



For example, looking at chapter 1, the actual counts are



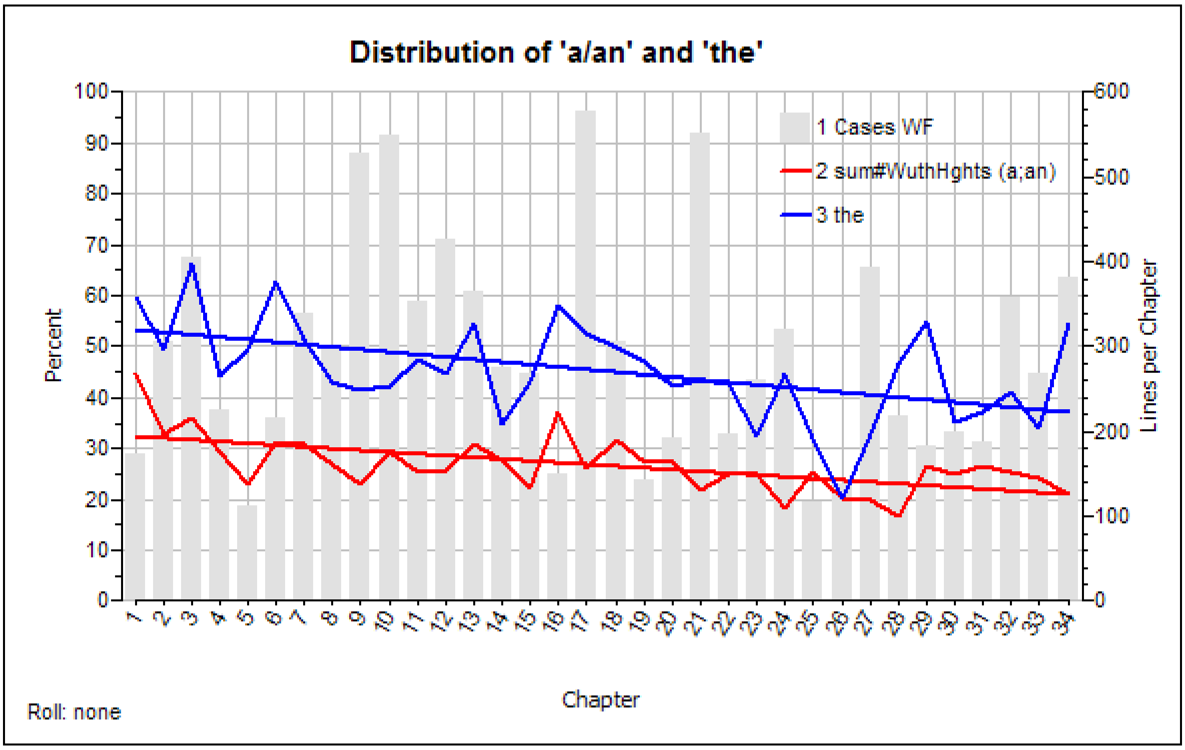
Frequencies		Chap
		1
WuthHghts	Cases WF	173
	sum#Wuth Hghts (a;an	77
	the	103

So, *a/an* occurs 77 times over 173 lines, giving 100*77/173 = 44.5%, and *the* occurs 103 times over 173 lines, giving 100*103/173 = 59.5%.

This chart tells us that

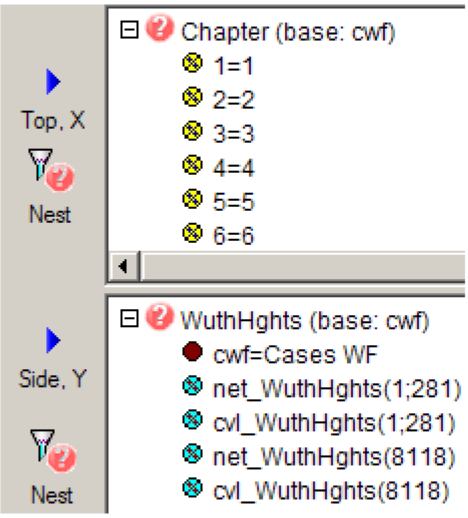
- *the* is more common than *a/an* everywhere except for chapter 26
- frequency per chapter declines with a slight convergence (the trend lines get a bit closer)

Looking at the percentage per chapter instead, the chart is



The drop in the trend lines is surely rather strange. The percentage for *the* drops from 53% in chapter 1 to 38% by chapter 34. The percentage for *a/an* drops from 32% to 21%. This is clearly not just statistical noise.

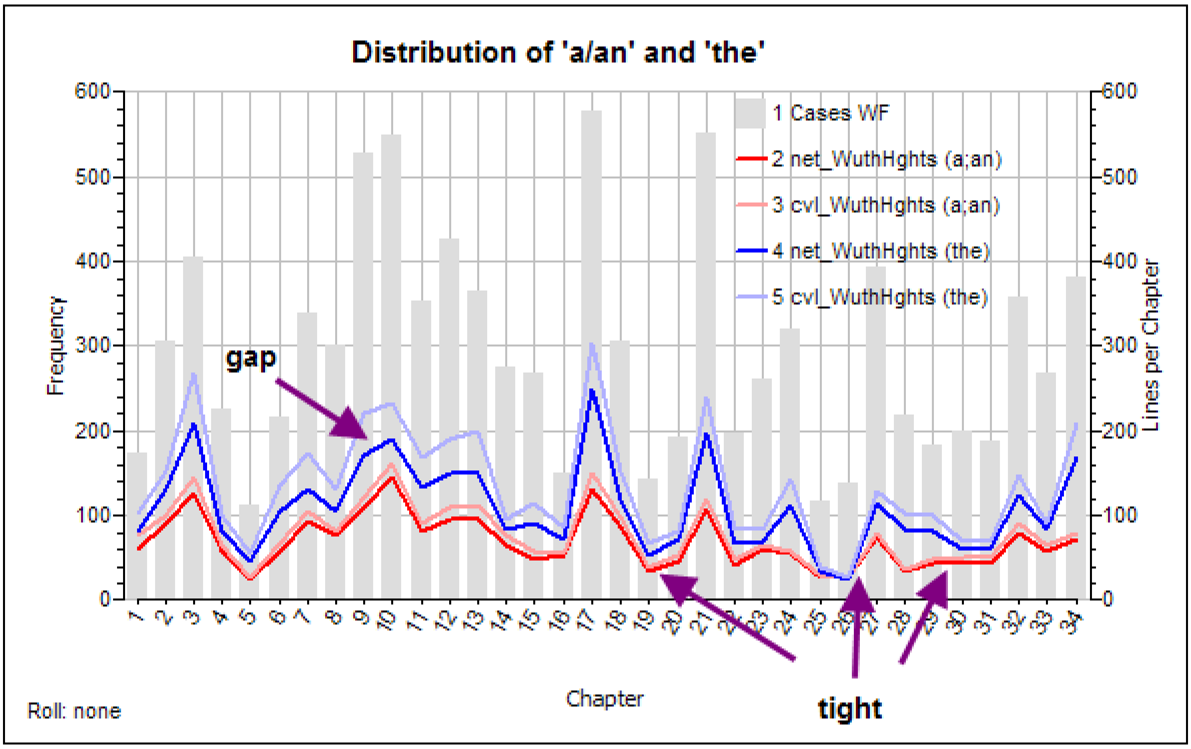
The two charts above give the number of times the articles occur as a proportion of the number of lines. Therefore, a line with ten instances of *the* will increase the numerator by 10. A measure of prose flacidity could be construed as the difference between the number of times an article occurs in a line, and the number of lines with at least one article. By this measure, a line with ten occurances of *the* is very flacid, and a line with just one or none, is very tight. The chart which shows the density of occurences can be specified as



⊟ Chapter (base: cwf)
1=1
2=2
3=3
4=4
5=5
6=6

⊟ WuthHghts (base: cwf)
cwf=Cases WF
net_WuthHghts(1;281)
cvl_WuthHghts(1;281)
net_WuthHghts(8118)
cvl_WuthHghts(8118)

The chart as frequencies is



The net_ series count one per line, regardless of the number of instances within that line. The cvl_ series count each instance. So, an input line like

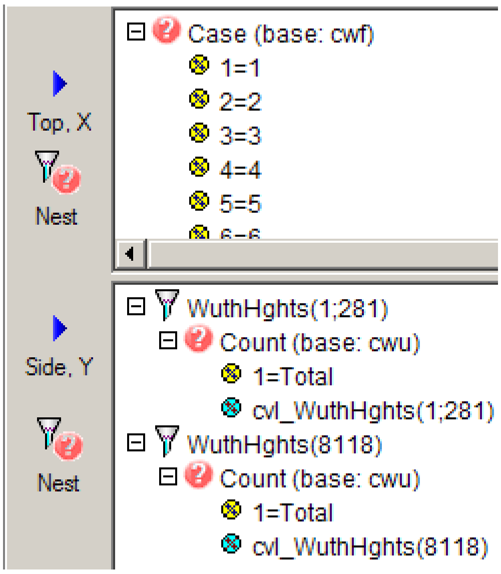
> The tree on the hillside was pleasant to the eye

would count as 1 for *net_WuthHghts(the)*, but as 3 for *cvl_WuthHghts(the)*.

This shows that multiple instances of *a/an* within a line are more common in chapters 1 to 17 (the gap between the dark blue and light blue series decreases from chapter 18), and that from chapters 18 to 30, *the* hardly ever occurs more than once within a line.

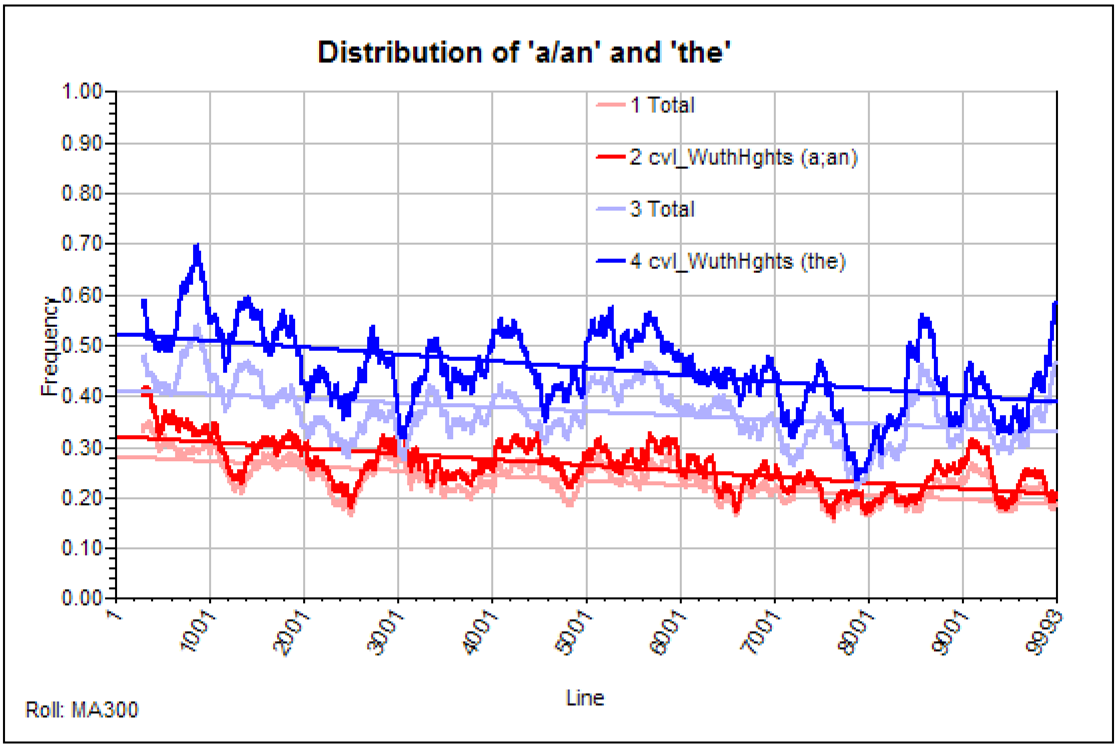
Thus, on this one criterion, one could say that the writing style gets tighter (less flacid) as the novel progresses.

Looking at the same series, but by line number instead of by chapter, the specification is



```
⊟ Case (base: cwf)
      1=1
      2=2
      3=3
      4=4
      5=5
      6=6
⊟ WuthHghts(1;281)
   ⊟ Count (base: cwu)
         1=Total
         cvl_WuthHghts(1;281)
⊟ WuthHghts(8118)
   ⊟ Count (base: cwu)
         1=Total
         cvl_WuthHghts(8118)
```

And the chart is



The Y1 axis now shows values from zero to 1, where 1 would mean 'in every single line'. The moving average is 300 lines. Light red and light blue plot the proportional number of lines which contain at least one instance of 'a' or 'an', and 'the'. The dark red and dark blue lines are the proportional number of times the words occur. For example

Over any window of 300 lines

If 150 lines contain ‘the’, then plot 150/300 = 0.5 (light)

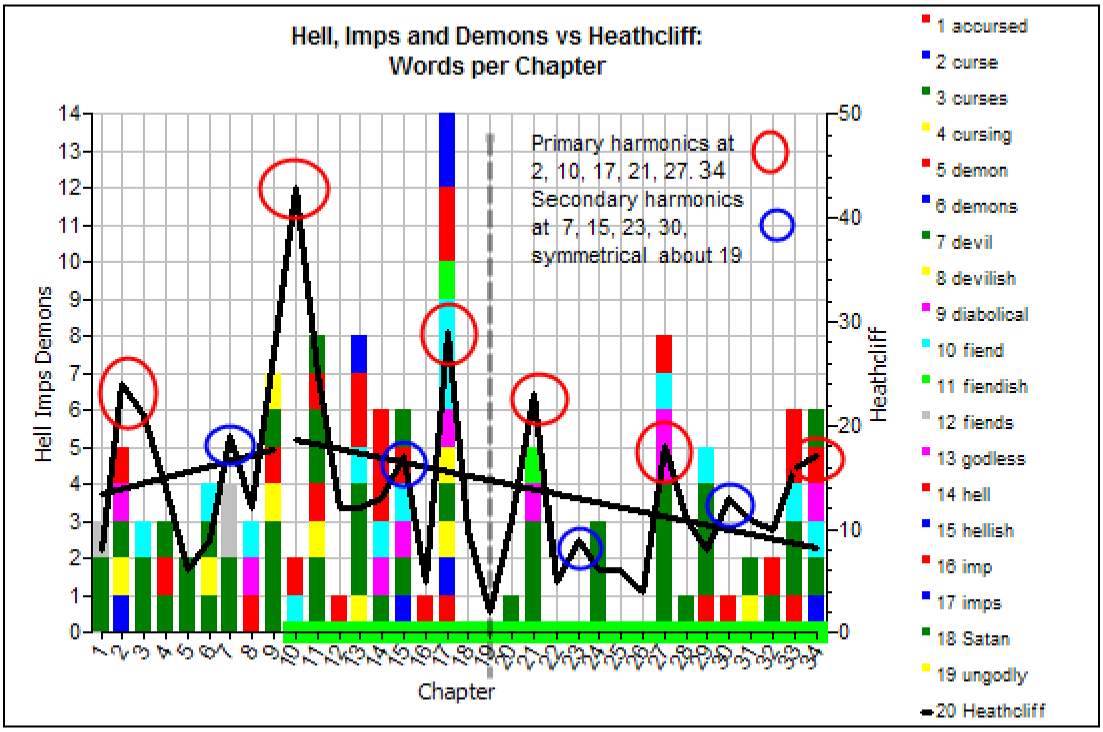
If 150 lines contain 200 instances of ‘the’, then plot 200/300 = 0.66666 (dark)

So why does the trend line sink by 34% for *a/an* (from 0.32 to 0.21) , and for *the,* by 27% (0.52 to 0.39)?

Does this suggest much tighter writing as the novel progresses?

The trend line pairs each have a slight convergence, indicating that the drop in instances is matched by a drop in the number of text lines with more than one instance (the gap between net and count of values is closing).

Hell, Imps and Demons



This chart shows that satanic imagery occurs mostly in chapters where Heathcliff is mentioned a lot. Note however chapter 10, where Heathcliff is mentioned 43 times, but satanic imagery occurs only twice, one *fiend*, and one *hell*.

Top: Chapter
Side: WuthHghts

Frequencies	Chapter																																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
accursed								1									1												1				1	
curse		1													1		1																	
curses			1		2	1			2		2													1										
cursing		1				1					1		1				1														1			
demon											1																							
demons																																		1
devil	2	1	1	1		1	2		1		2		3	1	1		1			1	3			2			4	1	3		1	1	2	1
devilish									1								1																	
diabolical		1						1						1	1		1				1						2							
fiend			1			1		1		1			1	1	1		3										1		1				1	1
fiendish																	1				1													
fiends	1						2																											
godless																																		1
hell		1							1	1	1	1	2	3	1	1	2										1			1		1	2	1
hellish													1				1																	
imp				1																														
imps																	1																	
Satan				1					1		1				1																			1
ungodly									1																									

This is the chapter where Heathcliff returns after many years. Neither instance is prejudicial.

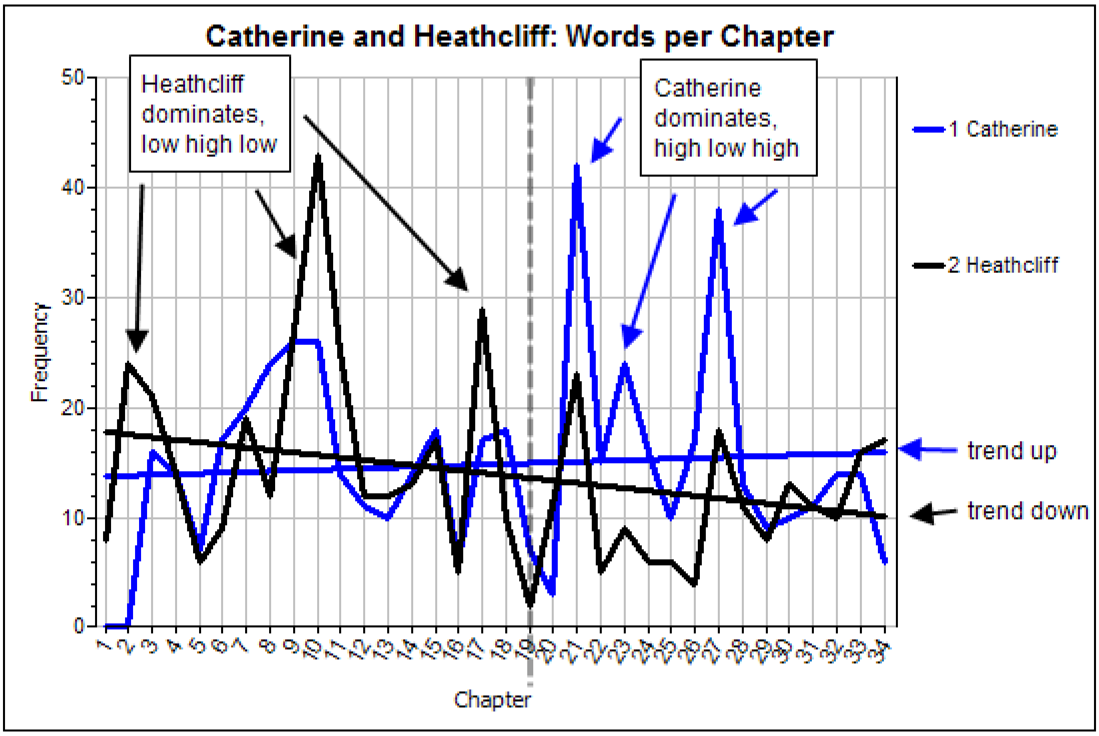
'Go and carry my message,' he interrupted, impatiently. 'I'm in hell till you do!'

'Mr. Heathcliff is not a fiend: he has an honourable soul, and a true one, or how could he remember her? '

For the chart, the sequence of peaks in the Heathcliff series is notable. Major peaks alternate with minor ones, symmetrically around the mid point at chapter 19. This is clearly a technique to engage reader attention – Heathcliff dominates, then recedes – first escalating to the crescendo of chapter ten, then abating in waves through to the end.

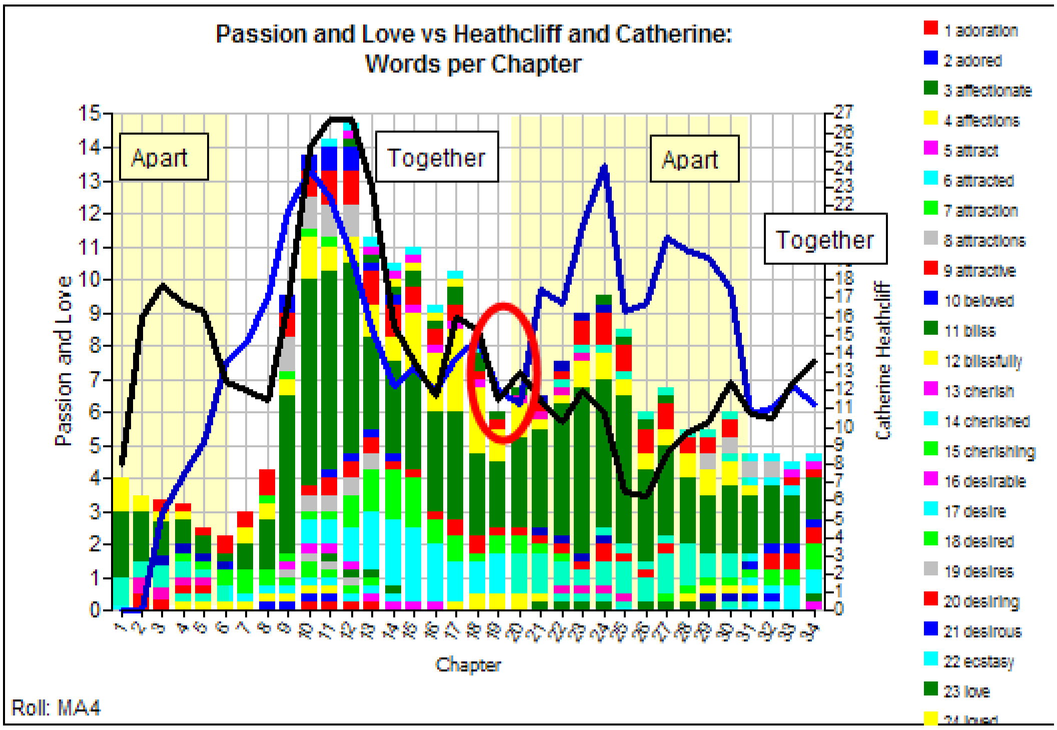
Catherine and Heathcliff

This chart shows the number of instances of *Catherine/Catherines/Cathy* and *Heathcliff* (appears to exist in no other form) in each chapter.



Of some note is the inverted pattern of peaks around the symmetric axis at chapter 19. The less intense low-high-low is extended over more chapters than the more intense high-low-high (more intense, because two highs for Catherine as opposed to only one high for Heathcliff), so impact-wise they could be said to balance out. By chapter 28, they both sink slowly from view. Overall, the trend line for Catherine rises where Heathcliff declines. From chapters 21 to 27, Catherine doubles Heathcliff.

Love and Passion



This chart is smoothed at MA4 to show where the structure in the interplay between Catherine (blue) and Heathcliff (black) balances with or diverges from the aggregate of love and passion.

The interplay between Catherine and Heathcliff is in a clear sequence of separation/convergence/separation/convergence, labelled as Apart/Together in the above chart. The two end sections (first Apart, chapters 1 to 6, last Together, chapters 31 to 34) are both short, and the two middle sections (chapters 7 to 19, and then 20 to 30) are both long, again a symmetry. The closest convergence for Catherine and Heathcliff is at the middle axis point, chapter 19 again. From chapters 7 to 19, all three dimensions – Heathcliff, Catherine and the love/passion aggregate – peak and ebb simultaneously. From chapters 20 to 30 there is disunity of purpose as the three dimensions diverge.

[end of document]