

About 'heart', 'inward' and MONK analytics

By Martin Mueller

Below are some musings about the words 'inward' and 'heart'. I report on a set of observations that I made about the distributions of these words in the NCF corpus. I make a basic assumption that quite a few people will be interested about longitudinal data in MONK. Call it fancier ways of looking up word histories in the OED. What MONK can add to the OED is better data about usage. In the world of words, Usage is King, and Usage is a function of when, where, and how often. Aristotle talks about this in the Poetics. Some ancient satirist (was it Lucian?) talk about Attic philologists quarrelling about 'keitai' or 'ou keitai' , whether a word 'occurs' or 'does not occur.'

The French, unlike the English, have a formal frequency-based dictionary of their language. Just as the dose makes the poison, so frequency makes the meaning. The nuanced recording of frequency information about various kinds of linguistic data is a major contribution that computers can make to literary studies.

My ramblings throught the words 'inward' and 'heart' have as their goal to identify some simple strategies through which users can do roughly similar things with a lot less hassle. I used R, Minitab, Excel, and Access. I don't think many literary scholars will want to do this. But think of this as an example of iterative data exploration with a variety of tools. The Monk task is to lower the hassle factor.

I began with the simple question whether male and female novelists differ interestingly in their use of words. The source was NCF in WordHoard. You make a work set of novels written by women and another written by men. You then use Dunning's log likelihood ratio.

Dunning's is a better version of the chi-square test, or so I am told. You define some text or texts as a "reference corpus." You then define an "analysis corpus," and you ask whether the distiributions of tokens in the analysis corpus differs interestingly from their distribution in the reference corpus. The tokens can be seen at different levels of abstraction as types of spellings, lemmata, or POS tags. Dunning's is interpreted like a chi-square test with two degrees of freedom. You get a separate result for every tpe. The comparison produces a list of token types that are used disproportionately often or rarely in the Analysis corpus compared with the Reference corpus.

The results of this test are quite striking. If you look at the nouns that are more common in women writers, you see virtually the entirely vocabulary of thoughts and emotions: heart, sorrow, joy, etc. There are a lot of sergeants, arch-deacons, and bishops in the characteristically male vocabulary.

Almost by accident, I hit on the word 'inward' as a word more common in women's writings. It occurs approximately ~1,500 times in a corpus of 40 million words, and it occurs in about two thirds of the 250 novels. WordHoard retrieves all instances of 'inward' in less than a minute, and I assume that MONK will do as well or better. I think it's at the outer boundary of a search scenario where it makes sense for the system to retrieve all instances of a word with KWIC output. Later I added the word 'heart'. This word occurs in every novel. There are some 30,000 instances of it in NCF. You almost certainly don't want to have all those instances in a first search. Apart from the fact that it takes five minutes to retrieve the data, what are you going to do with them once you have them? So this is a case where your first encounter with the data is almost certainly more productive if you get a good look at the forest before you look at any of the trees.

Following Phil Burns' advice, I began with a box plot. The box plot is a graphic technique that goes back to the late seventies. It is very primitive but quite powerful precisely because it involves no fancy statistics and makes no assumptions about the underlying distribution of the data. If you know eighth-grade math you can learn in less than fifteen minutes how to read box plots. And once you have seen a few of them, you begin to see characteristic differences.

Let us compare the box plots for the ubiquitous 'heart' and the less common 'inward'. I now make another assumption. A "search" command in MONK triggers the creation of a 'dataframe.' Dataframe is a technical term in R-speak. In its specific meaning, a dataframe is a table that the statistical program R can work on. It consists of a table with rows and columns. The columns consist either of counts or continuous variables (n occurrences or the frequency per 10K of 'inward' in *Daniel Deronda*) or they are 'factors' (author, sex, decade of publication, genre). This is standard statistical stuff.

Alternately, a dataframe can be seen as a general procedure for delivering data in a systematic fashion for subsequent manipulation. The "long data format" that statistical programs like as the input for their operations may not be a particularly efficient way of managing data on the server or delivering them to the client machine. What matter is that the client receives a general "dataframe" that can be transformed into particular dataframes that support this or that subsequent manipulating or 'analytic', using the term in its technical UIMA sense.

Here are five rows of the 'heart' dataframe in the 'long data format' that supports all operations described in this paper. ¹

Author	Title	Date	Origin	Sex	Genre	Count	TotCount
Hays	VictimofPrejudice	1799	British	f	fiction	161	48150
Lamb	Glenarvon	1816	British	f	fiction	486	151111
Hays	EmmaCourtney	1796	British	f	fiction	201	65159
Charles	Schoenberg-Cotta	1864	British	f	fiction	470	172334

¹ Neither 'origin' nor 'genre' are invoked as factors in the subsequent analyses because in the data set all authors are British (except for Scott) and all works are fiction.

Wollstonecraft	Mary	1788	British	f	fiction	64	23506
Shelley	M. Falkner	1837	British	f	fiction	392	152558

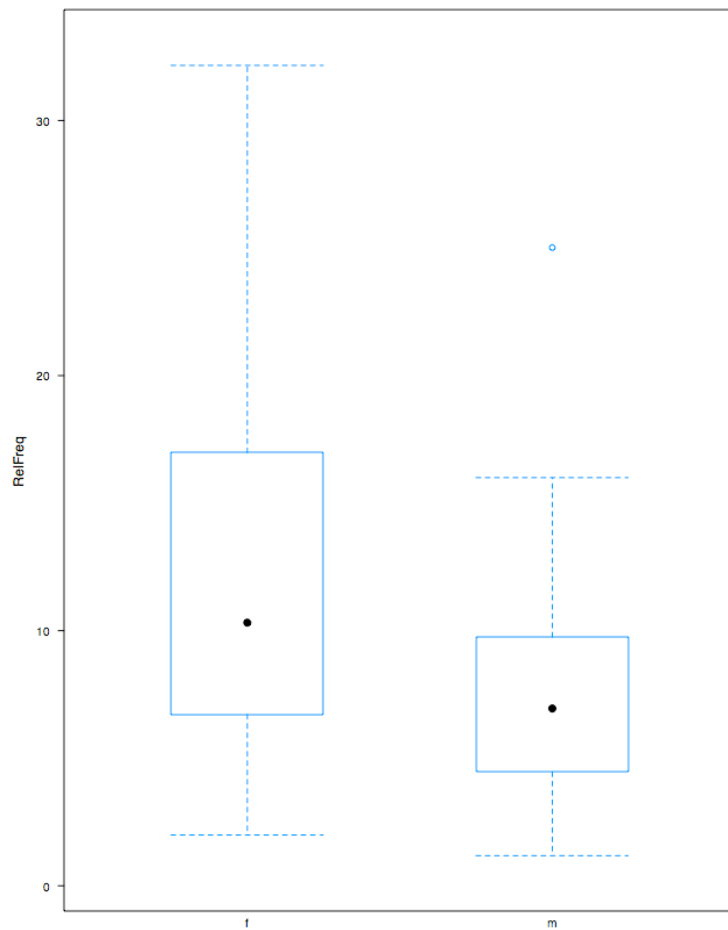
For any search of any kind, the critical thing is that these data are delivered in a manner suitable for subsequent manipulation.

The simplest form of manipulation may be actually be exporting the data to a third-party program, whether an Excel spreadsheet, a statistics program, or a visualization program like Many Eyes.

Dunning's test had shown me that male and female novelists differ in their use of words like 'heart' and 'inward'. How do they differ? R is not a program that casual users are likely to learn, but it has a cool command that lets you create box plots side-by-side. You cannot do that in Excel or Many Eyes.

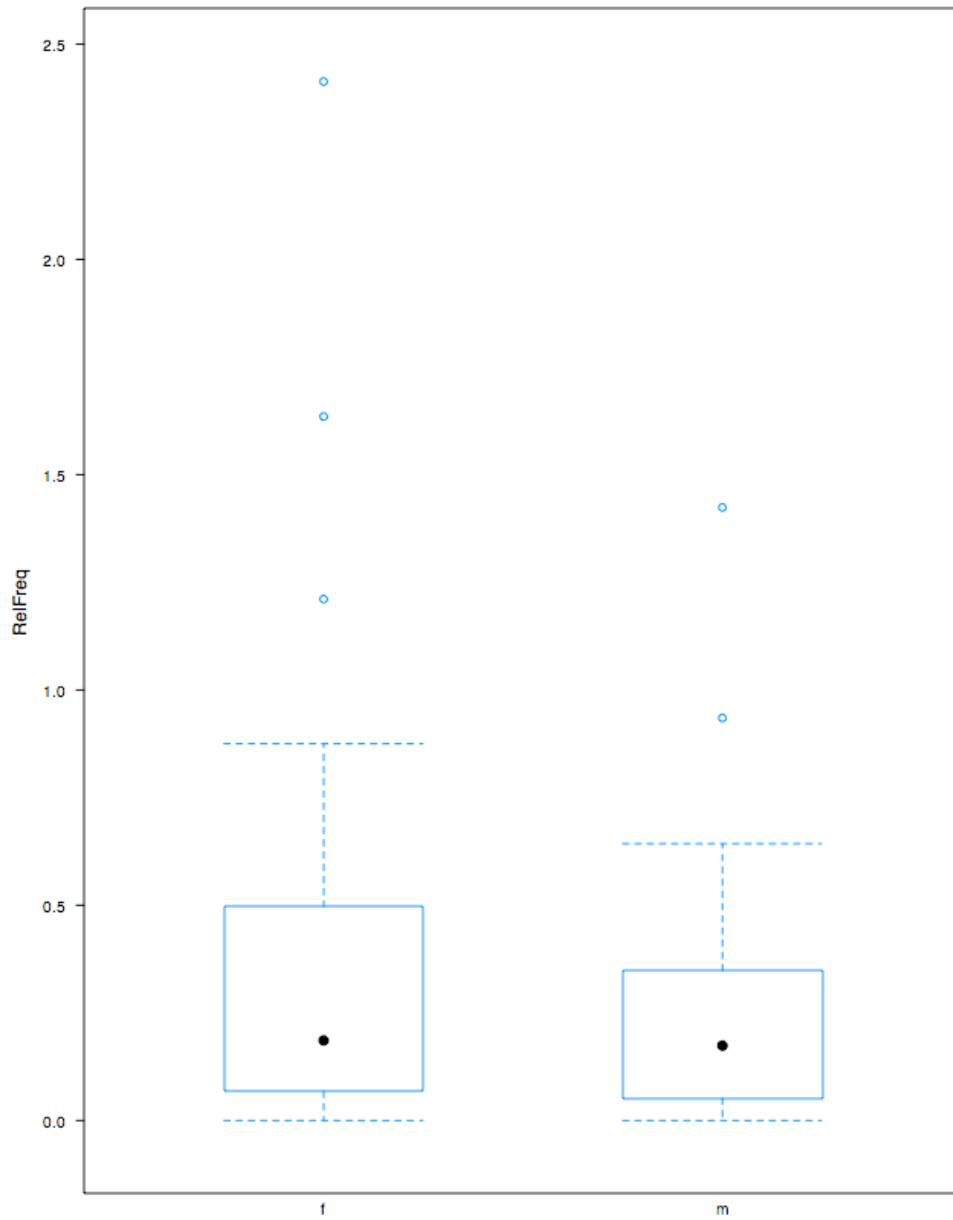
Let us look at the side-by-side box plots of 'heart' and 'inward'. In both cases we look at relative frequencies per 10K words per author, a figure that is easily computed from the 'count' and 'totcount' columns. The box shows the interquartile range, with a dot inside the box marking the average (For some reason the side by side display doesn't show you the median, which typically is drawn as a horizontal line through the interquartile box. The "whiskers" show the range of values that differ from the top or bottom of the interquartile range by a factor of 1.5. Values above or below those whiskers are shown as 'outliers'.

To interpret the box plots, it helps to know that of the 249 novels roughly two-third were written by men. (165:84). In the case of 'heart' you can see that the mean for women is distinctly higher (~12:7). You also see that the variance for 'heart' in women's writing is larger: the lowest value is somewhat higher, but the top whisker is considerably higher. There is one outlier on the male side, but its value is well below the top whisker value on the female side.



Let us now look at the box plot for 'inward', which tells a very different story. We see that the mean values barely differ, but the variance differs considerably. There are three outliers for women and two for men, but you need to remember that there are twice as many novels by male writers. The outliers are also much more pronounced. The top value is ~ five times the value of the top of the interquartile range whereas for "heart" the top value appears to be a little more than twice as high as the top interquartile value. Because "Inward" with its 1,500 occurrences is a lot less common than "heart" with its 33,000 occurrences, you want to treat those differences in proportion with caution. Still, it looks as if the outliers of 'inward' are of a more drastic kind.

If you turn from the box plot to the details, you see that the outlier for 'inward' is George Eliot. In fact, the 405 occurrences of 'inward' in her novels account for more than a quarter of all occurrences of the word. If you look at the male outliers, you see two texts with distinct theological concerns, James Hogg's *Confessions of a Justified Sinner* and Cardinal Newman's *Loss and Gain*. So you wonder whether the distribution of 'inward' may have less to do with gender and more with spirituality. But eight of the ten writers with the highest relative frequency of 'inward' are women. On the other hand, there are a lot of novels where the difference between men and women is negligible.



Box plots, then, are quite informative visualizations, and they can be made to tell stories with no technical statistical knowledge. You do need to know how they are constructed, and you do need to know something about your data.

Three useful numbers

I have always found it useful to know three numbers about the distribution of a word in a text or across a collection of texts:

1. The raw count
2. The relative frequency
3. The z-score or standardized value $((\text{value} - \text{average})/\text{standard deviation})$

The first of these tells you how much there is in the first place. This is especially important when the counts are low: if you count occurrences on the fingers of one hand, relative frequencies may not tell you much or be actively misleading. The second expresses the same data in proportional terms and lets you compare counts in one work with another. The third places a particular count more explicitly in a comparative environment and tells you how many standard deviations a text sits above or below an average conceived of as 0.

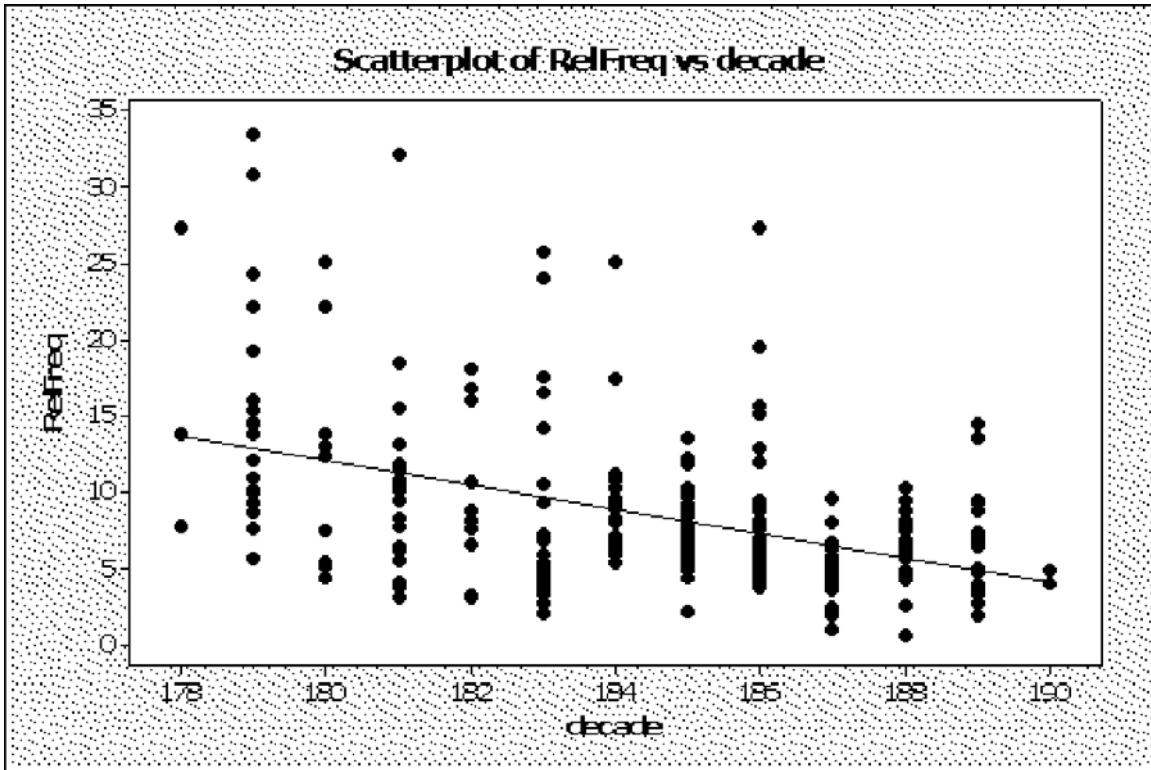
The standardized value is questionable because it belongs to the world of normal distributions and the Bell curve, when it is plainly the case that linguistic phenomena are typically not normally distributed. On the other hand, standardized values between +1 and -1 suggest broadly that you are not in a world of interesting difference, while values above 2 or below -2 begin to be interesting. Here is a table that shows the five top values for inward

Author	Sex	Count	TotalCount	RelFreq	Z-score
Eliot	f	405	1,678,227	2.41	5.79
Ward	f	51	311,921	1.64	3.68
Hogg	m	12	84,254	1.42	3.11
BronteC	f	83	685,562	1.21	2.5
Newman	m	10	106,945	0.94	1.78

That is a lot of figures to absorb, but it is hard to see how you could do with less. For instance, the raw counts tell you that in the case of the male authors Hogg and Newman you are dealing with individual works of moderate length, whereas with the female authors, and in particular with George Eliot and Charlotte Bronte, you are clearly in a world of stylistic habits across a number of works. Similarly, differences in relative frequencies might look modest, but if you think of z-scores above 2 as likely outliers, they are a good indicator of what values to look at more closely. The z-score of 'inward' for George Eliot may be said to shout at you.

'Heart' over time

Does the use of 'heart' change over the course of the nineteenth century? There are two ways of answering this question. First, you can make a scatter plot, showing the relative frequency by decade. This shows that the word is used more and with greater variance in the late eighteenth century.



A one-way Anova test, using relative frequency as a response and decade of publication as a factor is in some ways even more informative. Minitab has a very primitive, but quite powerful visualization of results. Remember what Anova does: it compares two or more series of measurements with regard to some factor. Each series of measurements is a sample. Anova analyses the variance of the sample and estimates the mean of the (indefinitely large) population from which the sample is drawn. Thus the Anova tests treats the collection of NCF novels as if they were random samples drawn from a larger population

Let us begin with an Anova test that compares male and female writers.

Difference between 'heart' in novels by men and women

Level	N	Mean	StDev
f	85	11.673	7.170
m	164	6.848	3.481

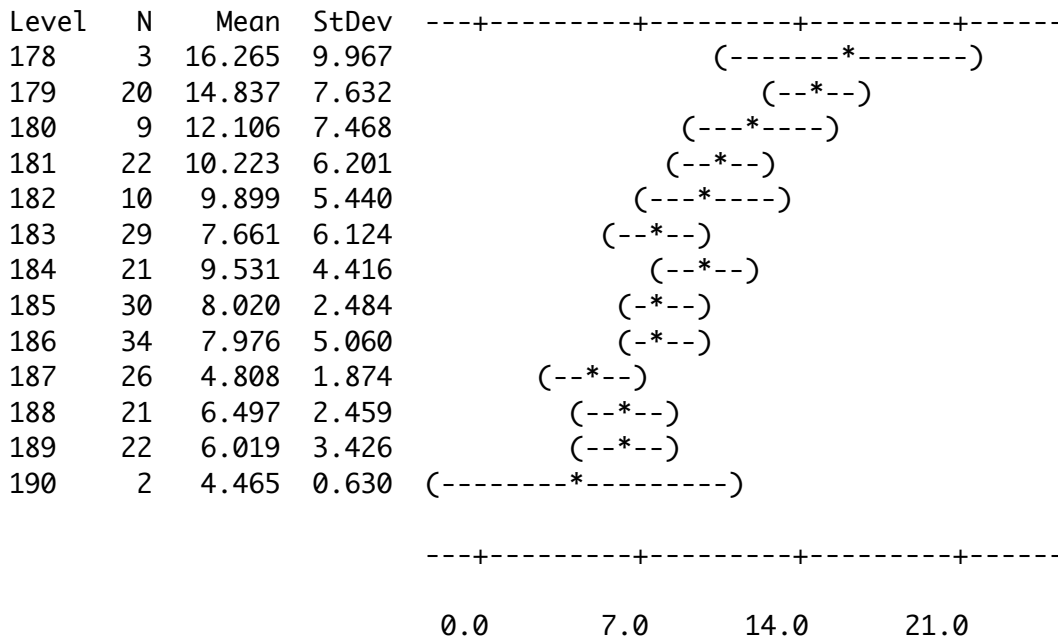
+-----+-----+-----+-----+
 (---*---) (---*---)
 +-----+-----+-----+-----+

6.0 8.0 10.0 12.0

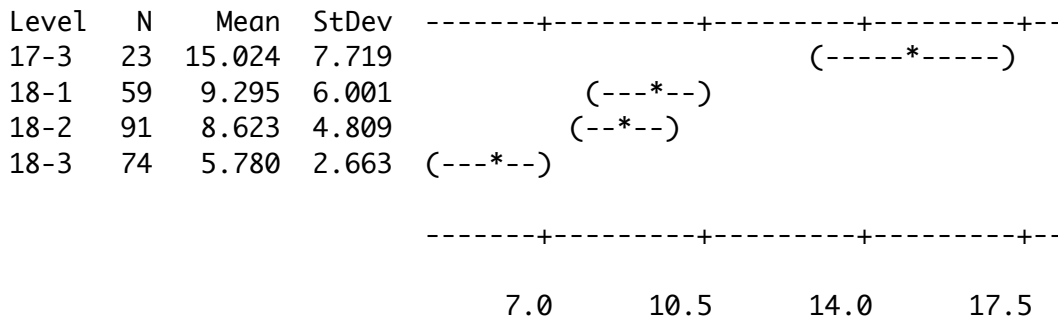
This is not a pretty summary, but it is very clear. There is a 95% chance that the female average for 'heart' lies between ~10.5 and 12 and that the male average sits between ~6.4 and 7.8. It is crystal clear that these estimates are very clearly separated.

Now we turn to longitudinal data and look at the use of 'heart' by decade. The visualization tells a rather wobbly story. You notice that the samples from the 1780's and 1900's are too small to be worth anything, and some of the decade samples are on the small side for reliable estimates. So we group by third of century rather than by decade.

Difference in use of 'heart' by decade



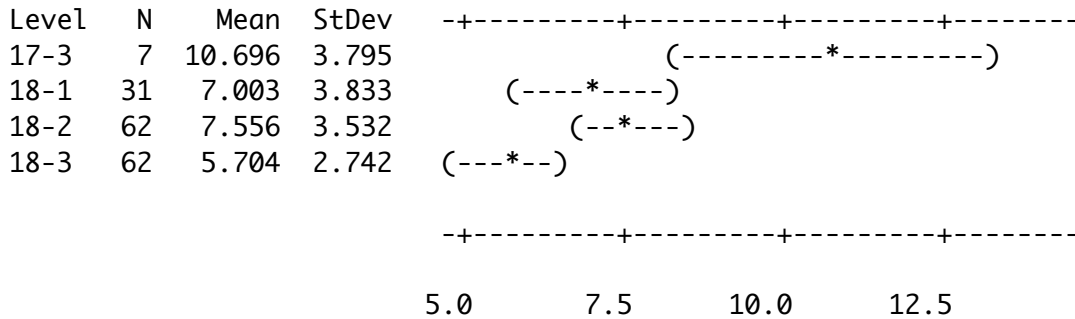
Difference in use of 'heart' by third of century



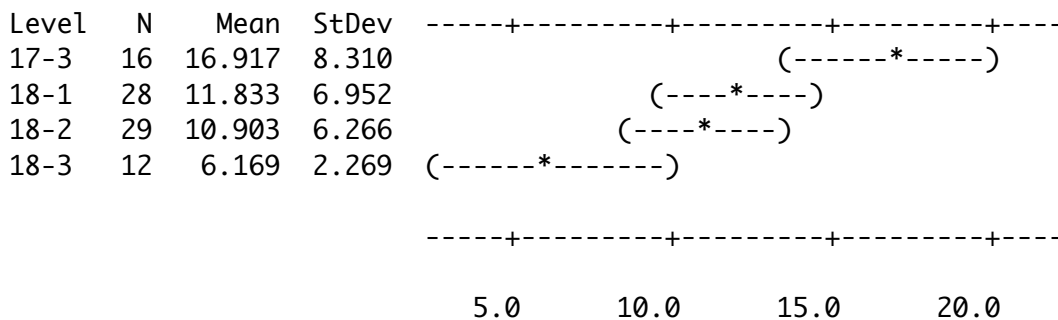
This tells quite a compelling story. 'Heart' is a lot more common in late eighteenth than in late nineteenth-century fiction. There is not much difference between the early middle and middle nineteenth century, but to the extent that there is a trend, the values from the middle of the century move towards the late part of the century.

But perhaps this has nothing to do with time and more to do with sex of author? So we run Anova tests separately on male and female authors. Despite the low sample size for late eighteenth century novels by men and late nineteenth-century novels by women, it is pretty clear that the change from the late eighteenth to the late nineteenth century has nothing to do with the sex of the author.

'Heart' in novels by male writers by third of century



'Heart' in novels by female writers by third of century



We learn from this exercise not only that 'heart' varies by time and sex of author, but that the differences are quite large and of the same order of magnitude. If we look at the list of women writers who use 'heart' a lot, we find Mary Wollstonecraft and her daughter, Mary Shelley, in the top group, which consists largely of lesser known writers. Jane Austen does not use the word a lot, especially in comparison to the women writers of her generation. In fact, her use of it (6.1 per 10K) sits a little below the average of male writers of that period (6.9 per 10K).

Can we get there more simply?

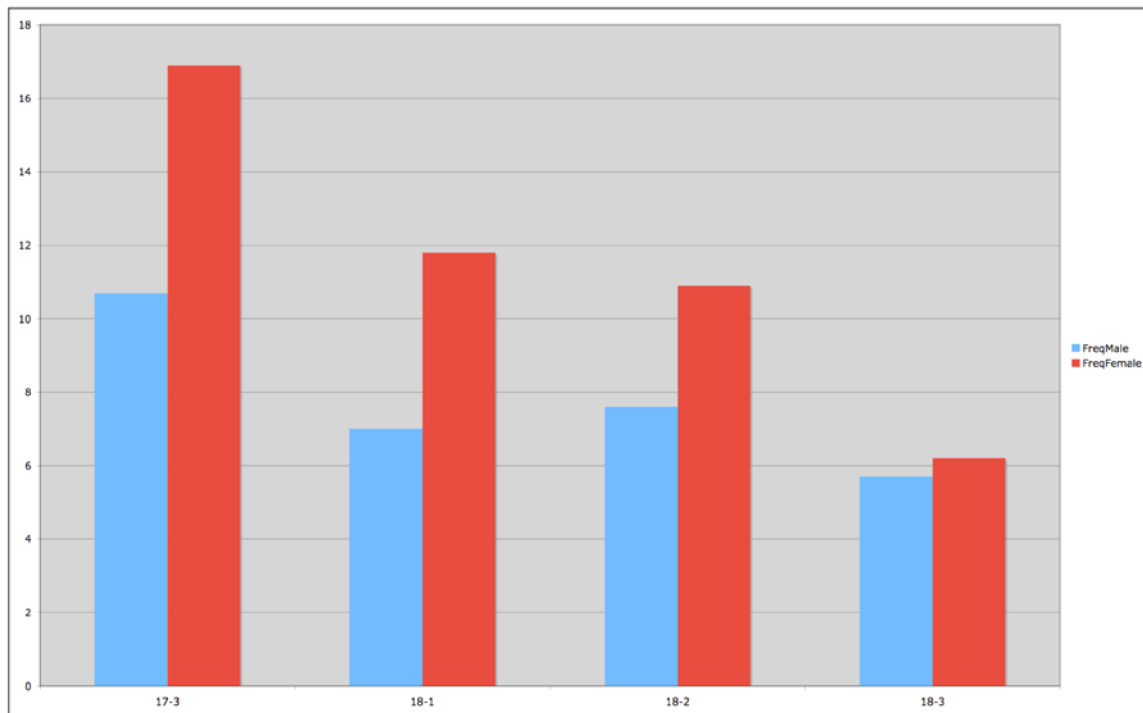
The upshot of this rather cumbersome set of operations involving four different software programs is that we have placed 'heart' in a co-ordinate space of gender and time. It is not a bad framework for starting an inquiry in which we look at individual passages in

individual texts. One may find it illuminating to do this inquiry for other words, such as 'sorrow', 'joy'. A dozen interestingly similar or contrasting patterns would build up a framework for analysis of keywords in certain kinds of novels. For instance, the male writers with high values for 'heart' and 'inward' are explicitly religious writers. Does that hold up for other words?

Can we get there more simply? My data exploration involved the following analytics or visualizations based on a standard dataframe created by a search:

1. Computing relative frequencies
2. Computing standardized values
3. Sorting by raw counts, relative frequencies, standardized values, author, date (at various levels of aggregation), sex of author
4. A scattergram with a regression line based on relative frequencies by decade
5. Side-by-side boxplots, separating data by the factor of sex of author
6. Analysis of variance operations
 - a. relative frequency by sex
 - b. relative frequency by decade
 - c. relative frequency by third of century
 - d. relative frequency for male writers by third of century
 - e. relative frequency for female writers by third of century

Would less be more? For instance, we can chart the 'heart' values for male and female writers by third of century as follows:



This Excel style sheet is more dramatic than the boring manual typewriter style visualizations in Minitab's Anova. It does a nice job of highlighting that women use 'heart' more. But the chart may be misleading for the last third of the nineteenth century. Are male and female writers becoming more like each other? Or is the low sample size to blame?

Who develops the analytics for these operations?

If some of these techniques or better versions of them are integrated into MONK, who does the development work for the 'analytics' that the interface group will need? There is some wobbliness about the use of 'analytic' in MONK discussions. In UIMA parlance, an 'analytic' is any procedure that is run on 'analysis data' or "the logical union of an artifact and its metadata." I take that to mean that an analytic is any procedure in which somebody does something with data in order to find out something. Standardizing a vector of values is an analytic, and so is creating a box plot. Analytics in UIMA parlance are simple or complex. An 'aggregate analytic' is an analytic that consists of several constituent analytics that are combined in a formal work flow.

In MONK speak, 'analytic' has tended to be used in the sense of "an aggregate analytic that is implemented through a D2K itinerary and performs some text mining operation on a set of documents." By that definition, nothing in this particular discussion counts as an 'analytic'. On the other hand, if we look at people actually doing text analysis in a broad sense, these are the 'analytical' operations they will perform much of the time.

This is a matter that will require clarification.