1. Introduction: When the humanities meets digital …

As a scholar engaged in digital humanities, I am often asked what difference there is, if any, between traditional humanities scholarship and digital humanities. My answer is rather straightforward: the essential nature of literary scholarship, or the interpretative nexus between evidence/findings and insight, remains the same, I would argue, whether studies are carried out in a digital or non-digital environment. On the other hand, the use of digital tools has made striking difference in the way texts are read: computers have made it possible to read texts non-linearly. Language is by its nature linear in the sense that a text is a linear sequence of words; a word is a linear sequence of letters; no one can utter two or more words simultaneously; the flow of information is uni-directional, etc. Human processing of language is thus linear, so is our reading of texts. Since literature is the art of language, it cannot help being constricted by the linearity of language. Digital humanities however allows us to step outside of these linear constraints.

What sort of advantages do we gain if we read a text/texts non-linearly? Once digitized, a text, normally an object of close reading and literary/aesthetic appreciation, can transform itself into a "bag of words" that could now be decomposed and processed non-linearly. At that point, the text becomes an entity that is free from the linearity of language. Just as a pair of chromosomes of our cells can be sequenced as an array of the genetic alphabets (ACGTs)[1], so a text can likewise be reduced to a vector of words, alphabets, and frequency counts, among others. That, in turn, can lend itself to statistical visualization and "distant reading". Just like genomes are sequenced and submitted to bioinformatic analysis to produce a dendrogram representing the divergence of modern taxonomic groups from their common ancestor, it is possible to transform a corpus of classic fiction into a set of high dimensional word-frequency vectors. Complex interrelationships between texts can then be visualised in the form of a cluster dendrogram or multi-dimensional scatter diagram.

Non-linear reading of texts enables:

- Counting, Sorting, Concordancing; Deconstructing, Decontextualising, and Bootstrap sampling
- Transforming a text into a vector of words, numbers, or frequencies, etc.
- Data/Text mining, Visualisation and "Distant reading" (Moretti, 2013)

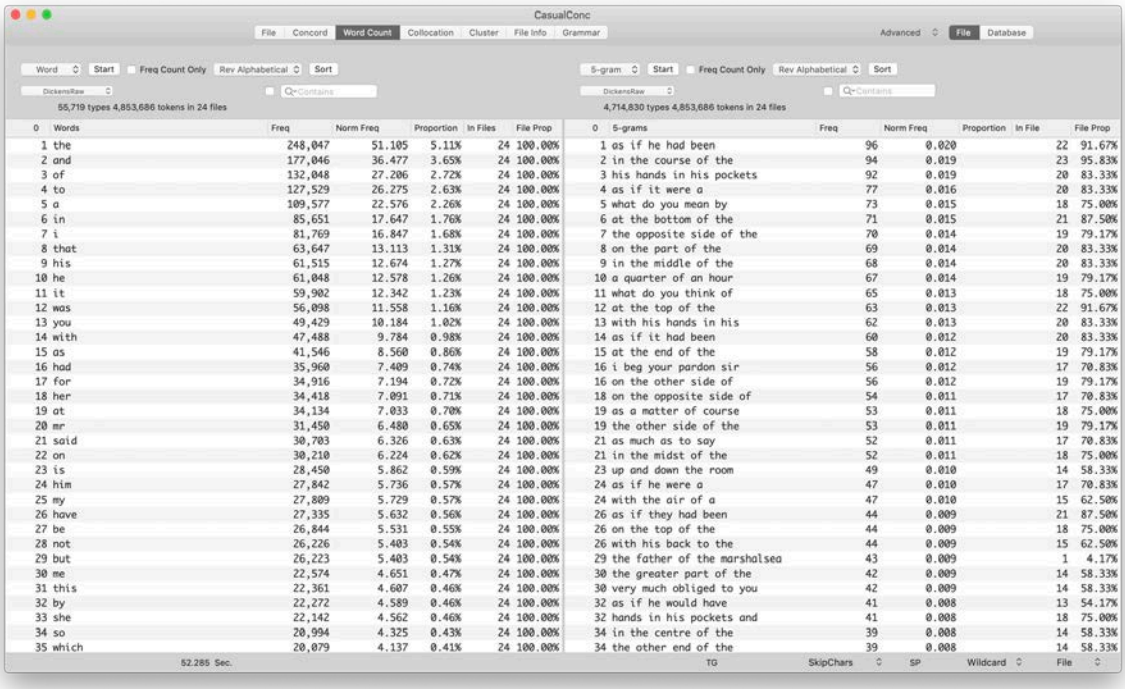
2. Non-linear reading as macro/distant reading

Figure 1 shows most frequent words and 5-grams (five-word sequence). Frequency counting is one of the most basic operations in digital humanities: texts are converted into vectors, fed to statistical analysis or machine learning. While a trained analyst will be able to precisely infer what genre or linguistics register the text in question belongs to judging from words ranked in the top 10, 20, or 30 of the frequency list, it is extremely difficult to guess what five-word sequences are the most frequent in a given corpus. Frequent *n*-grams (*n* being the number of words in sequence) are good markers of phraseological, referential, and stylistic features of texts in the corpus. If we focus on a 5-gram *his hands in his pockets*, a KWIC (KeyWordx-In-Context) concordance makes it possible to read texts "vertically". This vertical reading can lead to an

[1] Chromosomes of our cells are broken down to four bases: adenine, guanine, cytosine and thymine.

realization that the phrase is actually part of larger phraseological framework with his hands in his pockets, and (additional description of body(-parts)

of local textual functions played by the



55,719 types 4,853,686 tokens in 24 files

0	Words	Freq	Norm Freq	Proportion	In Files	File Prop
1	the	248,047	51.105	5.11%	24	100.00%
2	and	177,046	36.477	3.65%	24	100.00%
3	of	132,048	27.206	2.72%	24	100.00%
4	to	127,529	26.275	2.63%	24	100.00%
5	a	109,577	22.576	2.26%	24	100.00%
6	in	85,651	17.647	1.76%	24	100.00%
7	i	81,769	16.847	1.68%	24	100.00%
8	that	63,647	13.113	1.31%	24	100.00%
9	his	61,515	12.674	1.27%	24	100.00%
10	he	61,048	12.578	1.26%	24	100.00%
11	it	59,902	12.342	1.23%	24	100.00%
12	was	56,098	11.558	1.16%	24	100.00%
13	you	49,429	10.184	1.02%	24	100.00%
14	with	47,488	9.784	0.98%	24	100.00%
15	as	41,546	8.560	0.86%	24	100.00%
16	had	35,960	7.409	0.74%	24	100.00%
17	for	34,916	7.194	0.72%	24	100.00%
18	her	34,418	7.091	0.71%	24	100.00%
19	at	34,134	7.033	0.70%	24	100.00%
20	mr	31,450	6.480	0.65%	24	100.00%
21	said	30,703	6.326	0.63%	24	100.00%
22	on	30,210	6.224	0.62%	24	100.00%
23	is	28,450	5.862	0.59%	24	100.00%
24	him	27,842	5.736	0.57%	24	100.00%
25	my	27,809	5.729	0.57%	24	100.00%
26	have	27,335	5.632	0.56%	24	100.00%
27	be	26,844	5.531	0.55%	24	100.00%
28	not	26,226	5.403	0.54%	24	100.00%
29	but	26,223	5.403	0.54%	24	100.00%
30	me	22,574	4.651	0.47%	24	100.00%
31	this	22,361	4.607	0.46%	24	100.00%
32	by	22,272	4.589	0.46%	24	100.00%
33	she	22,142	4.562	0.46%	24	100.00%
34	so	20,994	4.325	0.43%	24	100.00%
35	which	20,079	4.137	0.41%	24	100.00%

4,714,830 types 4,853,686 tokens in 24 files

0	5-grams	Freq	Norm Freq	Proportion	In File	File Prop
1	as if he had been	96	0.020		22	91.67%
2	in the course of the	94	0.019		23	95.83%
3	his hands in his pockets	92	0.019		20	83.33%
4	as if it were a	77	0.016		20	83.33%
5	what do you mean by	73	0.015		18	75.00%
6	at the bottom of the	71	0.015		21	87.50%
7	the opposite side of the	70	0.014		19	79.17%
8	on the part of the	69	0.014		20	83.33%
9	in the middle of the	68	0.014		20	83.33%
10	a quarter of an hour	67	0.014		19	79.17%
11	what do you think of	65	0.013		18	75.00%
12	at the top of the	63	0.013		22	91.67%
13	with his hands in his	62	0.013		20	83.33%
14	as if it had been	60	0.012		20	83.33%
15	at the end of the	58	0.012		19	79.17%
16	i beg your pardon sir	56	0.012		17	70.83%
16	on the other side of	56	0.012		19	79.17%
18	on the opposite side of	54	0.011		17	70.83%
19	as a matter of course	53	0.011		18	75.00%
19	the other side of the	53	0.011		19	79.17%
21	as much as to say	52	0.011		17	70.83%
21	in the midst of the	52	0.011		18	75.00%
23	up and down the room	49	0.010		14	58.33%
24	as if he were a	47	0.010		17	70.83%
24	with the air of a	47	0.010		15	62.50%
26	as if they had been	44	0.009		21	87.50%
26	on the top of the	44	0.009		18	75.00%
26	with his back to the	44	0.009		15	62.50%
29	the father of the marshalsea	43	0.009		1	4.17%
30	the greater part of the	42	0.009		14	58.33%
30	very much obliged to you	42	0.009		14	58.33%
32	as if he would have	41	0.008		13	54.17%
32	hands in his pockets and	41	0.008		18	75.00%
34	in the centre of the	39	0.008		14	58.33%
34	the other end of the	39	0.008		14	58.33%

Figure 1 Frequency counting of single words and 5-grams