[*CHWP* Titles] [CHC 2004]

Playing 'Shame': One Technique for Introducing Text Analysis to the Literary Studies Classroom

Playing with Text Analysis (A Joint Session of COCH/COSH and ACCUTE; Geoffrey Rockwell, organiser)

Ray Siemens

Canada Research Chair in Humanities Computing, University of Victoria

siemens@uvic.ca |||| http://web.uvic.ca/~siemens/ |||| About the Author

CHWP A.51, publ. April 2009.

[Abstract / *Résumé*]

KEYWORDS / *MOTS-CLÉS*: Computer-assisted text analysis, content abstraction, Google, teaching / *Analyse de texte assistée par ordinateur, abstraction de contenu, Google, enseignement*.

> Note: This text of this paper, below, was read at "Playing with Text Analysis," a joint session of COCH/COSH and ACCUTE organised at the 2004 HSSFC Congress by Geoffrey Rockwell.

Introduction

A former professor of mine, now gone to his just reward – a character who one might never imagine to find in a David Lodge novel, and yet he was noted in one as a poor soul banished in the late 1960s from glitzy, big-shoulder US academic culture to the pastoral Canadian prairies we all know and love – gave me some of the most useful pragmatic advice I'd ever received from an academic up to the point that I'd received it. He suggested that all of us concern ourselves as much with the expanding of our own knowledge as we do with concealing those areas in which we have little expertise or experience. This was heady stuff for me (I was quite a few years younger, then), but it was an apt observation. And when I think of the focus of this panel – 'playing with text analysis' – his words resonate.

They resonate today as I think about computer assisted text analysis because this technology appears, today, to be the technology integral both in expanding and consolidating knowledge – and in quickly masking deficiency in those darker areas. This may be arguable but, certainly, text analysis is the chief technology underlying most 'discovery' carried out on electronic materials: it underlies the catalogue we use in our libraries as much as it does the online directory services we use to locate phone numbers and addresses; and it has as much of a place in processes with names like "data mining" as it does in the searches we carry out on our electronic literary texts. Worth noting for our purposes also is that much of what's fundamental to this technology has its origins in the work of those interested in carrying out searches – simple and complex – on literary and historical texts; for humanists, what underlies such technology also underlies many of the assumptions of our research as much as it does what we bring to the classroom. Thus, the communication of this to our students is essential. It is on this

latter point that I will focus.

Geoffrey Rockwell has kindly invited me to discuss a technique I’ve used to introduce computer assisted text analysis to a literary studies class. It is based loosely on the literary parlour game made most famous as "Humiliation" (also called "Shame" by some of those that I know) in David Lodge's novel *Changing Places*, where a group of academics attempt to top each other by making public the most shameful gap in their professional knowledge. Readers of Lodge will recall that the professor who's never read *Hamlet* wins the game but loses his job as the result of his shameful admission. Rather than propagating the accumulation of embarrassing gaps in knowledge, my way of introducing text analysis techniques champions how one can use down-and-dirty text analysis techniques to help prevent those potentially shameful situations. I must say at the outset that this technique is neither good text analysis nor is it good academic research, but it has value in the classroom because it follows basic principles and it invariably yields results that have significant impact on the students. The impact of this game-playing lays a strong foundation for an appropriate academic address of the topic in the classroom.

Setting the Stage and Gathering Materials

I begin a lesson by jokingly admitting (and, I’m happy to say, pretending) that I’ve never read, something like, say, Hamlet – a play I should know well, because I teach Shakespeare from time to time. For the purposes of my example today, though, I’ll use Edmund Spenser’s *Faerie Queene* – in large part because it is a text that some schools of thought suggest that those of us with doctorates in English literature should all have read at one point or another, though precious few have read it clear through, and fewer yet are able to talk about it coherently beyond what takes place in Book I, Canto I, where a gentle knight goes pricking on a plain. It is a bit more believable, I think, to admit this sort of gap; and I don’t claim, myself, coherence of this text. (Another example I like to use is Milton’s *Paradise Lost*, for much the same reason; lots of people should have read it, though year by year fewer are exposed to it in a coherent fashion – and our memory fades over time, doesn’t it?) The focus of the example, I note, could just as easily be Eliot’s *Wasteland*, Morrison’s *Jazz*, Atwood’s most recent – or an author, or a literary theme, trope, or critical approach. It could even be a topic as non-literary as antibiotic-resistant viruses (“superbugs”), as a student suggested to me last fall; the results are equally-illuminating. I find that the book-one-should-know approach has the best impact.

What I might first do, after setting a scene in which I discuss the need to know more, in a great hurry, about my author and my text, is to turn to an internet search engine; most recently, I’ve used Google. Here, I enter the search terms “Faerie Queene” and “Spenser” (see figure 1). One of the nice things in using a work with a unique spelling, and an author who’s not so well recognised, is that it has the effect of limiting the results returned by the search engine to a manageable number. Anyone who has entered “Shakespeare” into a search engine will know that Shakespeares have done more than written plays and a few non-dramatic works: they’ve also built fishing reels and rods for generations, made fine handcrafted furniture, and offer a libido enhancer of the sort whose notice should be caught by our spam-filter. And, so, those searching Shakespeare have, actually, a better chance of coming across a commercial site for fishing or furniture (or otherwise) than they do the peer-reviewed materials published on Shakespeare in a proper academic journal, or an appropriate text of a play from the Internet Shakespeare Editions or Renaissance Electronic Texts and Representative Poetry – or, even, some Sparknotes summary of one of Shakespeare’s plays’ salient features.

Lucky for me – and luck has some part in it, I’m afraid – the first result of my search yields what is found in figure 2, among this a reliable text of the entire *Faerie Queene* from U Oregon’s Renascence Editions; these are public domain electronic texts that have been produced predominantly by academics and qualified enthusiasts, with some quality control, descriptions of transcription and encoding practices, often an out-of-copyright edited text as its source, and a common layout across a large group of texts. Here, I’ll save all 6 books of the text, plus the “Mutability Cantos.” And I’ll save them not as HTML format – which I might if I were to want to redisplay them – but, rather, I’ll save them as a plain text file . . . which is something that our analysis program will prefer.

I also see a number of other resources that will give me varying degrees of knowledge about the text; I’ll save them in the same manner. These include a number of good resources, chiefly articles that I’ve found linked to a page on a fairly good ‘clearance house’ website called *The Luminarium*, some pieces from journals like *Studies in English Literature*, and *Criticism*, and so on. Also included are some of the *Sparknotes* to the work (after

downloading 6 pages, I note that this service wants my credit card number, so I abandon this pursuit); another service claims to have a number of key studies available, but they want a subscription charge of $20 US for access, so I move on. I don't spend much time assembling the corpus, deliberately; perhaps 10 minutes maximum, but often as little as 4-5 minutes.

Preparing a Textbase, Discussing Basic Analysis

With a corpus of sorts assembled, I then move on to the analysis package. For a number of years, for this sort of work I would use the program that pioneered such discovery for many of us in Canada and well beyond: *Text Analysis Computing Tools* or TACT – produced at U Toronto's Centre for Computing in the Humanities by Ian Lancashire, Willard McCarty, Russ Wooldridge, John Bradley, and others. These days, I also use Stefan Sinclair's *Hyperpo* for some work of this kind for my own research, though in-class I tend to use a program called *Concordance*, written by R J C Watt of U Dundee and the U London Institute of English Studies. As the method I describe has been integrated into the Text Analysis Portal for Research (TAPoR), as the "Googlizer," I suspect that I'll use this in the future.

The text files I've downloaded amount to some 5 megabytes in size – roughly the equivalent of 4 Bible's worth of text, or about 200 John Grisham novels' worth – but it will take some time for the program to prepare those files for proper searching by "generating" a textbase that will be some 30 megabytes total in size. This is done by selecting a menu command, specifying a few options, and then letting the machine calculate, sort, and index for 5 minutes or so more (see figure 3).

During the time the computer generates the textual database, I discuss some of the basic concepts and terminology that one needs to know to understand even down-and-dirty text analysis like I'm carrying out. Specifically, I talk about how tokens (all instances of the same graphical forms of a word, or character string) differ from types (a single representative of that graphical form); i.e. how you may have the single type of word "Redcrosse" appear 164 times in the textbase and, so, there are 164 tokens for the type "Redcrosse." I talk about how, when we're interested in the content of a text, we tend to pay attention to open class or content words – words like nouns, irregular verbs, adjectives and adverbs – more than we pay attention to closed class or function words – words like articles, pronouns, prepositions, conjunctions, and regular verbs; content words, so goes one school's mantra, denote the content of a piece, while function words provide the mechanics necessary to make the expression of that content. Lastly, I'll suggest a simple maxim: one of the most straightforward ways of learning about what's important across a body of texts is [a] repetition of words that have some import to that body (word frequency) and [b] the repeated occurrence of significant words in close relation to one another (co-location or co-occurrence).

I'll then offer a strategy that will allow an opening for our exploration of the text corpus we've assembled via its content: I'll suggest that we first determine the ten most frequently occurring open class (content) words; then we'll follow that with an examination of each of these words' top 5 or so open class collocates; and, finally, we'll assemble a cluster map of the results, and see if we can glean anything about our subject.

Computer-Assisted Analysis

To determine the most frequent open class words, I structure the display of the software so that it displays words by their frequency of occurrence, and then I sort them manually into open and closed class. It takes me some time to get to the open class words; I note, first, that "the" appears 14, 851 times, but this word and others similar to it will be of little assistance to us here. Eventually, I hit the word "knight," with 797 occurrences, and it becomes my first of ten words that we will use to begin our analysis; the others, as seen in figure 4, will be Great (791), Faire (743), Long (620), Selfe (604), Forth (575), Love (567), Most (463), Life (457), and Full (450). These words' collocates are found below and in figure 5 and figure 6.

- Collocates of 'knight': against, false, Redcrosse, Sir, gentle, armed, noble, straunger, good, aliue, euer, great, Lady

- Further Collocates:
    - Great (791): passing, goodly, knight, powre, glory, ladie, huge, earth, distresse, store, desire, chaine, God, Queene, fortune
    - Faire (743): knight, goodly, handling, full, wondrous, ladies, face, Britomart, Florimell, Pastorell, Pastorella, Virgin, Vna, Amoret, Damzell
    - Long (620): selfe, great, labour, huge, prison, time, space, toyle, stood, vaine, sought, deep, day, epic, travelled, traversed, rage, rest
    - Selfe (604): day, couert, ground, death, nature, maker, Cupid, vertue, prepaire, deceiue, despite, daunger, fight, Braggadochio
    - Forth (575): knight, rest, Errour, balme, nature, passed, poured, blood, issewed, drew, right, rode, iourney, bowre, castle. dreadful
    - Loue (567): face, lore, losse, ladie, God, sweet, true, dear, owne, former, faire, deare, hate, delight, friendship, brest, betrayed, daunger
    - Most (463): curteous, faire, loue, mightie, beaten, binding, selfe, goodly, noble, sacred, glorious, strong, perfect, man, Virgin, account
    - Life (457): bloud, death, patrone, raise, tree, state, wretched, happie, loue, loathed, episode, depriued, light, long, world, brest
    - Full (450): happie, false, fresh, Duessa, deare, vomit, gaping, sore, loth, glad, riche, deepe, great, wrath, aduenture, Britomart, dismay

With this carried out, experience suggests that many of those involved in this sort of exercise will already be looking for connections that made sense in the context of the text being considering. Typically, they will already be looking for ways to attempt to use the results gleaned in this fashion to avoid the shame of admitting (as I've teased, about myself, at the outset of the exercise) that they don't know much about the work; they will be playing Lodge's game of Humiliation, I suppose, by to an end opposite that seen in *Changing Places*. A further investment of 5 or so minutes provides cluster diagrams – say, written quickly on the blackboard – of the results of this process for all the top 10 words (see figure 7).

Focused Discussion

Then, discussion. For that discussion, I might begin with a few leading assertions, perhaps first noting that there seem to be a considerable number of words we might associate with knights, quests, courtly love and chivalry, &c.; and then I'll encourage the students to come up with similar assertions based in the evidence , regardless of their familiarity with the text itself. Some of the group will have read the text or some part of it, and they'll be comparing the text-analysis results against their own recollection; others, with some small sense of the text, will be attempting to use this information to enhance their understanding; and others yet, who don't know the text at all, will be using what's at hand to attempt to come to terms with what might be the salient features of the larger text.

Inevitably, we'll end up with a discussion of the characters that one or two of the students might be able to identify, possible motivations and challenges. Just as inevitably, some of what is discussed will be in keeping with what tradition has encouraged us to think about Spenser's work – at least as much as discussion, in the absence of actually knowing the text in a way that would be more responsible, will tend to reflect that situation of not yet knowing the text. And, inevitable still, will be the observations that focus on the fact that there is so much that this type of analysis cannot hope to capture – especially when carried out in the quick-and-dirty way that I present it; this, I note, has often been pointed out by a student or two who are already familiar with discovery methods of this sort (and the number of those who are familiar with these sorts of methods, in my undergraduate classes, increases every year).

Conclusion

In conclusion, let me say that I think we would all agree that there is no substitute for thoughtful reading – and

so will most everyone in the group that would join me in this sort of quick computer-assisted textual analysis of a text. That said, with the limitations of this explicitly acknowledged at the outset, as I've done here, one can't help but point out some of the positive outcomes of this method. It fosters a number of very important things. It encourages students to use contemporary technologies as part of their discovery (and I note, again, that a quickly-growing number of our students are already using such technologies, whether we provide direction in the vein or not). It allows us to consider the basic tenets of computer-assisted text analysis, especially those that have a strong foundation and proper applications in humanities disciplines such as ours. It demonstrates how we can use this technology in the context of our own discipline to discuss the salient features of a literary text, at the same time as it works towards a consideration of the merits and limitations of this discovery method. And, last of my points, if presented appropriately, it should reinforce for us all that there is no substituted for the thoughtful, informed, close reading of literary texts that is the hallmark of our discipline; one may be able to use this sort of thing to bluff a bit, but the professor and student who's not read Hamlet will, like Lodge's character, eventually be found out – through bravado or otherwise.

About the exercised itself: as I do it, the whole process before discussion should take about 15 to 20 minutes; discussion can add up to an hour on top of this, depending how one might chose to lead it. And I note that this exercise is best followed up by a through consideration of the work itself. Not everyone will want to do the *Faerie Queene*, I know; happily, the technique is portable – across texts, across authors, across disciplines, and beyond.

Works Cited

- Lancashire, Ian, in collaboration with J. Bradley, W. McCarty, M. Stairs, and T. R. Wooldridge. (1996). *Using TACT with Electronic Texts: A Guide to Text Analysis Computing Tools*. New York: Modern Language Association.
- Sinclair, Stéfan. *HyperPo: Text Analysis and Exploration Tools*. <URL: http://huco.ualberta.ca/HyperPo/>.
- Text Analysis Portal for Research (TAPoR). Geoffrey Rockwell, project leader. <URL: http://www.tapor.ca/>.
- Watt, R.J.C. *Concordance*. <URL: http://www.concordancesoftware.co.uk/>.

Figure 1

Figure 2

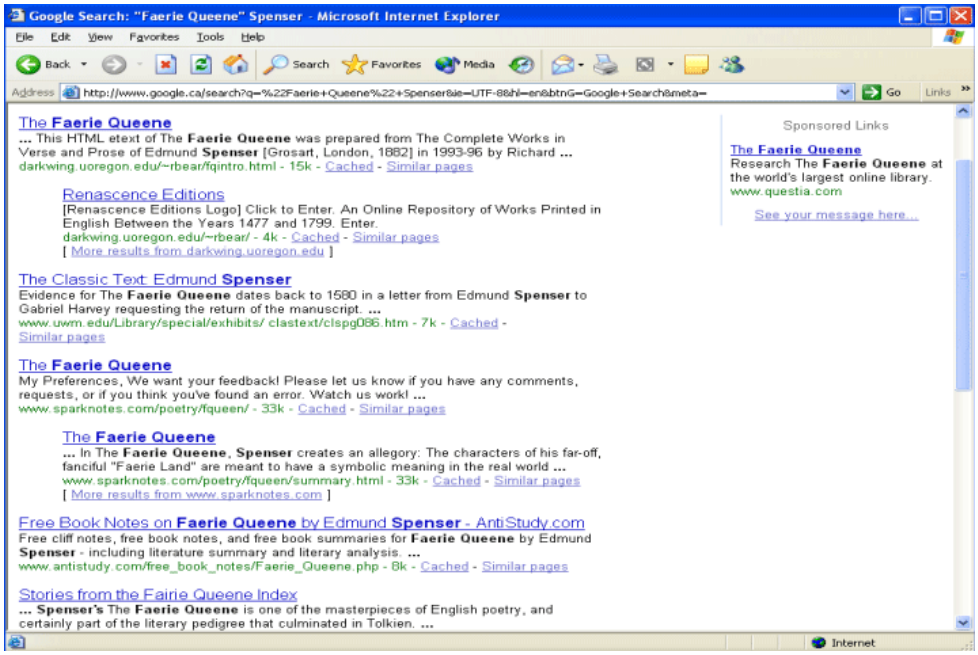

Figure 3

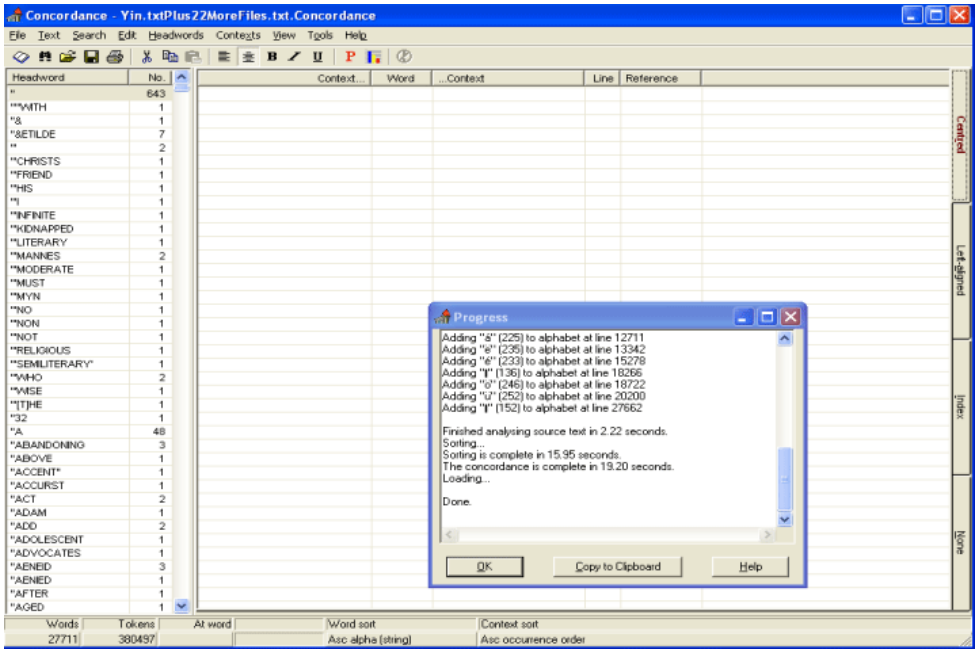

Figure 4

Analysis Process 2: 10 Most Frequent Open Class Words

Knight (797)
Great (791)
Faire (743)
Long (620)
Selfe (604)
Forth (575)
Loue (567)
Most (463)
Life (457)
Full (450)

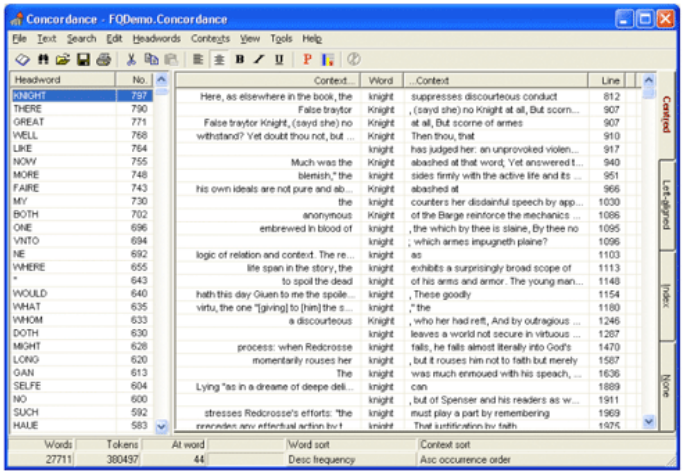

Figure 5

Analysis Process 3: Collocates of "knight"

Knight (797): against, false, Redcrosse, Sir, gentle, armed, noble, straunger, good, aliue, euer, great, Lady

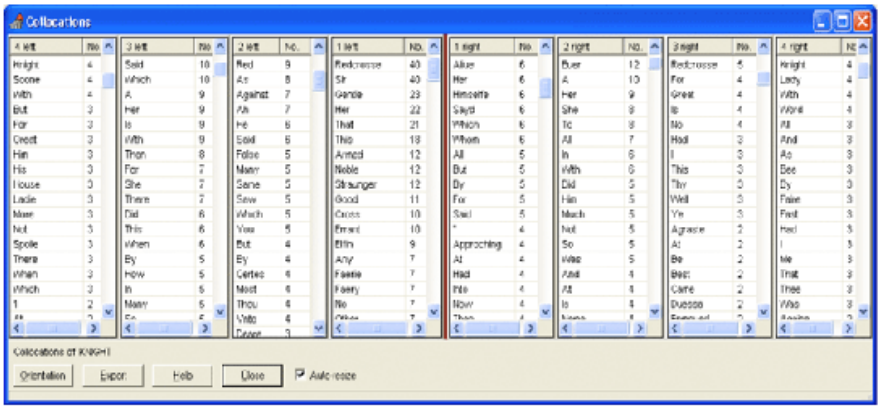

Figure 6

Analysis Process 4: Other Collocates

- **Great (791)**: passing, goodly, knight, powre, glory, ladie, huge, earth, distresse, store, desire, chaine, God, Queene, fortune
- **Faire (743)**: knight, goodly, handling, full, wondrous, ladies, face, Britomart, Florimell, Pastorell, Pastorella, Virgin, Vna, Amoret, Damzell
- **Long (620)**: selfe, great, labour, huge, prison, time, space, toyle, stood, vaine, sought, deep, day, epic, travelled, traversed, rage, rest
- **Selfe (604)**: day, couert, ground, death, nature, maker, Cupid, vertue, prepaire, deceiue, despite, daunger, fight, Braggadochio
- **Forth (575)**: knight, rest, Errour, balme, nature, passed, poured, blood, issewed, drew, right, rode, iourney, bowre, castle. dreadful
- **Loue (567)**: face, lore, losse, ladie, God, sweet, true, dear, owne, former, faire, deare, hate, delight, friendship, brest, betrayed, daunger
- **Most (463)**: curteous, faire, loue, mightie, beaten, binding, selfe, goodly, noble, sacred, glorious, strong, perfect, man, Virgin, account
- **Life (457)**: bloud, death, patrone, raise, tree, state, wretched, happie, loue, loathed, episode, depriued, light, long, world, brest
- **Full (450)**: happie, false, fresh, Duessa, deare, vomit, gaping, sore, loth, glad, riche, deepe, great, wrath, aduenture, Britomart, dismay

Figure 7

Analysis Process 5: Clustering (Example: "long")