
Pamphlet 1

→ January 15, 2011

# Literary Lab

## Quantitative Formalism: an Experiment

Sarah Allison

Ryan Heuser

Matthew Jockers

Franco Moretti

Michael Witmore

**Pamphlets of the Stanford** Literary **Lab**

IISSN 2164-1757 (online version)
IISSN 2164-3431 (print version)

Sarah Allison

Ryan Heuser

Matthew Jockers

Franco Moretti

Michael Witmore

# Quantitative Formalism: an Experiment

This paper is the report of a study conducted by five people – four at Stanford, and one at the University of Wisconsin – which tried to establish whether computer-generated algorithms could "recognize" literary genres. You take *David Copperfield*, run it through a program without any human input – "unsupervised", as the expression goes – and ... can the program figure out whether it's a gothic novel or a *Bildungsroman*? The answer is, fundamentally, Yes: but a Yes with so many complications that it is necessary to look at the entire process of our study. These are new methods we are using, and with new methods the process is almost as important as the results.

## 1. Prologue: Docuscope Reads Shakespeare

During the Fall of 2008, Franco Moretti was visiting Madison, where Michael Witmore introduced him to work he and Jonathan Hope had been doing on Shakespeare's dramatic genres, using a text tagging device known as Docuscope, a hand-curated corpus of several million English words (and strings of words) that had been sorted into grammatical, semantic and rhetorical categories.[1]

1 See Jonathan Hope and Michael Witmore, "The Very Large Textual Object: A Prosthetic Reading of Shakespeare," *Early Modern Literary Studies* 9.3 (January, 2004): 6.1-36; Witmore and Hope, "Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays" in *Early Modern Tragicomedy*, eds. Subha Mukherji and Raphael Lyne (London: Boydell and Brewer, 2007), 133-53; Hope and Witmore, "The Hundredth Psalm to the Tune of 'Green Sleeves': Digital Approaches Shakespeare's Language of Genre," *Shakespeare Quarterly* 61.3, "Special Issue: New Media Approaches to Shakespeare," ed. Katherine Rowe (Fall 2010): 357-90; and Witmore's blog, www.winedarksea.org.

Docuscope is essentially a smart dictionary: it consists of a list of over 200 million possible strings of English, each assigned to one of 101 functional linguistic categories called "Language Action Types" (LATs).[2] When Docuscope "reads" a text, it does so by looking for words, and strings of words, that it can "recognize" – that is to say, that it can match to one of its 101 LATs. When this happens, the associated LAT is credited with one appearance. For example, since Docuscope assigns "I" and "me" to the LAT "FirstPerson", their occurrence in a text is recorded as an appearance of the LAT "FirstPerson."[3]

Based on these counts, Hope and Witmore used unsupervised factor analysis – a factor, here, being a pattern that includes some categories, in variable proportions, and excludes others – to create portraits of received genre distinctions such as those made by the editors of the First Folio (Heminges and Condell), and of the genre of "late romances" that was first identified in the nineteenth century. Multivariate analyses and clustering techniques made groupings of the plays that corresponded not only to conventional genre groupings, but also picked out texts that critics had identified as outliers.[4] Thus, in clustering Shakespeare's Folio plays, the program managed to take *Henry VIII* out of the History plays cluster and place it near other "late plays," a re-adjustment from the initial Folio designations that later critics have advocated as well. One can see this grouping pattern in figure 1 below, taken from an early complete linkage clustering of the plays.

After seeing these results, Moretti asked Witmore whether he would consider clustering novelistic genres. Witmore agreed, and a meeting was planned for February 2009 at Stanford.

---

2 For Docuscope, see David Kaufer, Suguru Ishizaki, Brian Butler, Jeff Collins, *The Power of Words: Unveiling the Speaker and Writer's Hidden Craft* (Lawrence Erlbaum Associates: New Jersey and London, 2004). A fascinating discussion of how the program came to be designed and an early précis of its categories can be found at: http://www.betterwriting.net/projects/fed01/dsc_fed01.html, accessed 3 March 2010.

3 Because of the way they are used in the program, LATs must be given names without spaces. Obviously the characterization of the words that are contained in each of these categories is a matter of interpretation, as is the choice of those words themselves, which took place over the course of almost a decade of hand-coding. In general, Witmore and Hope use the categories or LATs to identify statistical patterns, then move from the categories to concrete textual instances in order to see how particular words are functioning in context.

4 They discovered, for instance, that Shakespeare's "late romances" were distinguished, linguistically, from those that went before them by word patterns that allowed speakers to narrate past action while highlighting their own emotional stance with respect to those actions (a process they called "focalized retrospection"). Specific linguistic features of these plays were responsible for this effect, for example (1) certain types of subordinated conjunction (a comma, followed by the word "which") and (2) past tense verb forms introduced by a past tense auxiliary form of the verb "to be." Comedies and histories were also shown to be significantly distinct from one another, with comedy possessing a high degree of first and second person pronouns (classed under the LATs FirstPerson and DirectAddress), a high degree of language expressing uncertainty (the LAT Uncertainty); an absence of nouns and verbs used to refer to motion, the properties of sensed objects, and sensed changes in objects (LATs labeled Motions, SenseProperty, SenseObject); an absence of first person plural pronouns (the LAT Inclusive); and an absence of words indicating social entities or expectations that must be shared or mutually acknowledged (the LAT CommonAuthority).

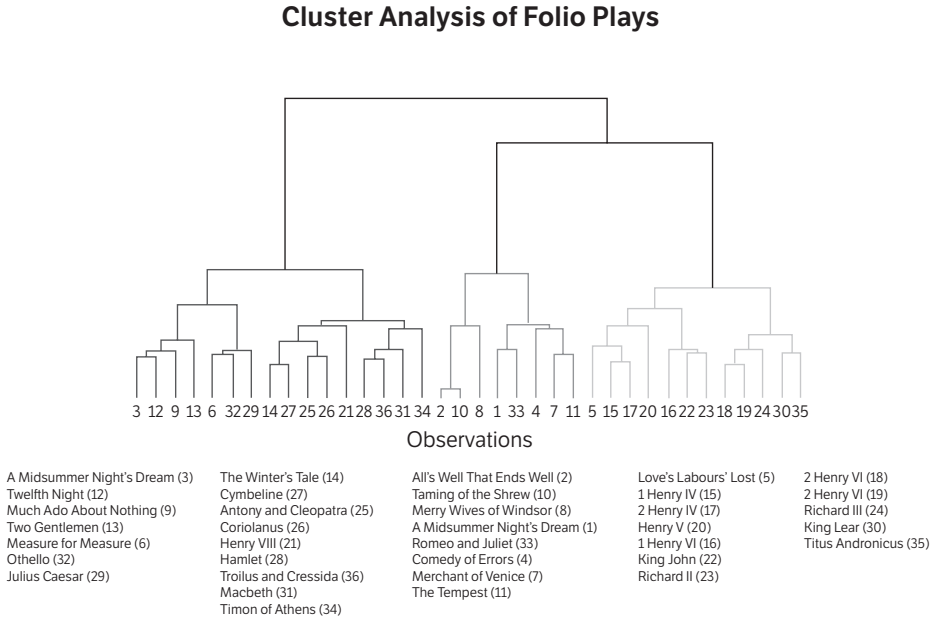



**Figure 1:** Dendrogram illustrating clustering of Shakespeare plays rated on Docuscope's Language Action Types (LATs) produced in 2003. Clustering method: complete linkage, Euclidean distances. Notice the presence of comedies in the first and third columns, late plays and tragedies in the second, and histories in the fourth and fifth. "Incorrect classifications" such as *Othello* and *Love's Labours' Lost* are discussed on Witmore's blog, www.winedarksea.org.

## 2. February 2009: Docuscope Recognizes Novelistic Genres

The starting point of the study was a corpus of 250 19th century British novels from the Chadwick-Healey collection.[5] Working with existing genre bibliographies, Moretti put together a sample of 36 texts loosely comparable to the Shakespeare corpus of the first Docuscope experiment, which comprised 12 genre sets, divided into two groups of 6. The first group (sets 1 through 6) included 4 gothic novels, 4 historical novels, 4 national tales, 4 industrial novels, 4 silver-fork novels, and 4 *Bildungsromane*. Of the 6 sets in the second group, 3 were also present in the first (sets 8, 9, and 12: 2 texts each from industrial novels, gothic novels, and *Bildungsromane*), whereas the other 3 were not (sets 7, 10, and 11: 2 texts each from anti-Jacobin, evangelical, and Newgate novels). Docuscope's task was to find and match the 3 sets from the second group that were also present in the first.[6]

5 We limited ourselves to this database because most other texts available on the web in 2006-8 appeared too unreliable for our purposes. Today, our assessment would be different, and a new initial pool would probably modify important aspects of our research.

6 This is the complete list of the texts: set 1 (gothic novels): *A Sicilian Romance, The Old Manor House, The Monk, and Melmoth the Wanderer*; set 2 (historical novels): *Waverley, Ivanhoe, The Entail,* and *Valperga*; set 3 (national tales): *Castle Rackrent, The Wild Irish Girl, The Absentee,* and *Marriage*; set 4 (industrial novels): *Shirley, Alton Locke, Hard Times,* and *North and South*; set 5 (silver-fork novels): *Glenarvon, Vivian Grey, Pelham,* and *Mrs Ar-*

To be sure he wouldn't unconsciously "tilt" his work on Docuscope's results in a pre-determined direction, Witmore asked to be told nothing about the texts he was receiving; title-pages were removed from the files ("they often provide giveaway clues that are less interesting than the microlinguistic moves that get made in the text"), and he literally walked into the meeting without knowing how Docuscope had performed. He was "hoping that Docuscope would fail at this test," he emailed us a few days before the meeting, "since I have a stake in arguing that it is material constraints on performance (in plays) that allows Docuscope to make intelligible genre discriminations when it comes to Shakespeare. If Docuscope turns out to be good at picking genres of novels as well, I am going to have to expand my notion of 'material constraint' in its relationship to language practices." (Later, though, he seemed pleased at how well Docuscope had done.)

Witmore used a variety of measures to match the genres from the two groups. For example, he assessed the degree to which multivariate statistical analysis could produce "factors" that would pry apart pairs from one another – a factor being a pattern of having certain LATs and lacking certain others.[7] He also compared each pairing against a collection of texts called the Frown Corpus (early 1990s American English) to see when they both exhibited identical elevated and depressed scores on LATs in comparison with the average score from Frown.[8] By combining these techniques, Witmore came up with the following matches: 2:9 (with 1:9 a close second), 4:8, and 6:12. When the curtain was lifted, it turned out that Docuscope's only mistake consisted in mis-matching group 9 (gothic novels) with group 2 (historical) rather than 1 (gothic): a mix-up most literary historians would consider venial, or maybe even inevitable, given the porous borders between these two genres. (And then, as Witmore wrote in his presentation, the correct 1-9 pairing was indeed "a close second.")

---

*mytage, or Female Domination*; set 6 (*Bildungsromane*): *Jane Eyre, The History of Pendennis, David Copperfield,* and *Daniel Deronda*; set 7 (anti-Jacobin novels): *Mordaunt*, and *Adeline Mowbray*; set 8 (industrial novels): *The Life and Adventures of Michael Armstrong, the Factory Boy,* and *Mary Barton*; set 9 (gothic novels): *The Mysteries of Udolpho*, and *Zofloya, or, The Moor*; set 10 (evangelical novels): *Coelebs in Search of a Wife*, and *Self-Control;* set 11 (Newgate novels): *Eugene Aram* and *Jack Sheppard*; set 12 (*Bildungsromane*): *Great Expectations* and *Middlemarch*.

Retrospectively, this list is odd – and flawed – in two opposite ways. First, the 36 texts were chosen so as to maximize variation within each given genre. Although quite wrong as a way to select a sample from a population, this choice was meant to increase the severity of the test: Docuscope had to prove it could "recognize" a genre even when given a quite disparate bundle of specimens. If this increased the difficulty of the enterprise, a second decision did the exact opposite: instead of giving Witmore 36 texts to be assigned to various generic classes, Moretti gave him discrete groups that were already subdivided into genres. This, clearly, made matters much easier, as the internal variation within any given genre could be averaged out by looking at the group as a whole.
These odd, antithetical decisions show how unprepared we were as a group – or should we say: as a discipline? – for this type of research. The idea of a random sample, for instance, never really crossed our minds...

7 One can think of a factor as a recipe for describing recurring patterns of variation in a larger collection of items. If each novel is a stack of cards, Docuscope examines all of the decks and counts what is in them. Then factor analysis goes through all of the contents of each stack and says, "whenever I see lots of red sixes, I see very few fours and fives of any kind." These recipes of "presences and absences" can then be tested against imposed groups of those stacks (genres) to see if the factors reliably distinguish items from each.

8 Use of a reference corpus seemed like a good idea, and since Frown had been used to test Docuscope in its development, those comparisons were built into the tool and so available for ready use. It turned out that the Frown comparisons were the most accurate in predicting literary critical genre judgments.

As the meeting was nearing its end, John Bender asked the hard question that was hanging in the air: Striking as these results were, did we think they had produced new knowledge? The answer, of course, was No: Docuscope had corroborated what literary scholars already knew – or at least were convinced of – i.e. that certain texts belonged to the same class. No new knowledge there. But that human judgment and unsupervised statistical analysis would agree on genre classification – this *was* a novelty that had emerged from the test. Just as Docuscope had corroborated existing scholarship, the latter had proven Docuscope's reliability. We wanted to know whether it could replicate its Shakespeare results in unfamiliar territory, and it could; that first experiment had not been a fluke. A computer could classify literary texts. And when Witmore – in passing, and almost as an afterthought – showed an old, unpublished chart from his Shakespeare study, the possibility seemed even richer in implications.

## 3. March 2009: Most Frequent Words Recognize Novelistic Genres

Docuscope had passed the test. Was it the only program that could do so? Matthew Jockers, who had been working on authorship studies for a while, wanted to see whether the methods he had been developing could be applied to genre recognition as well. In many ways, genre classification is akin to authorship attribution. But there is one important difference. With authorship problems, one attempts to extract a feature set that excludes context-sensitive features from the analysis, the consensus being that a set made up primarily of frequent, or closed-class, word features yields the most accurate results. For genre classification, however, one would intuitively assume that context words – say: "castle" in gothic novels – would be critical. Yet, Jockers's preliminary results suggested that an equally distinct genre "signal" may be detected from a small set of high-frequency features.

Using just 44 word and punctuation features – which we eventually ended up calling Most Frequent Words, or MFW – Jockers was able to classify the novels in the corpus as well as Witmore had done with Docuscope (and its far more complex feature set).[9] Using the "dist" and "hclust" functions in the open-source "R"[10] statistics application, Jockers clustered the texts in the dendrogram of figure 3.1:

9 To derive his feature set, Jockers lowercased the texts, counted and converted to relative frequencies the various feature types, and then winnowed the feature set by choosing only those features that have a mean relative frequency of .03% or greater. This resulted in a matrix consisting of the following 44 features (the prefix "p_" indicates a punctuation token type instead of a word token): "a", "all", "and", "as", "at", "be", "but", "by", "for", "from", "had", "have", "he", "her", "him", "his", "i", "in", "is", "it", "me", "my", "not", "of", "on", "p_apos", "p_comma", "p_exlam", "p_hyphen", "p_period", "p_ques", "p_quote", "p_semi", "said", "she", "so", "that", "the", "this", "to", "was", "which", "with", "you".

10 http://www.r-project.org/

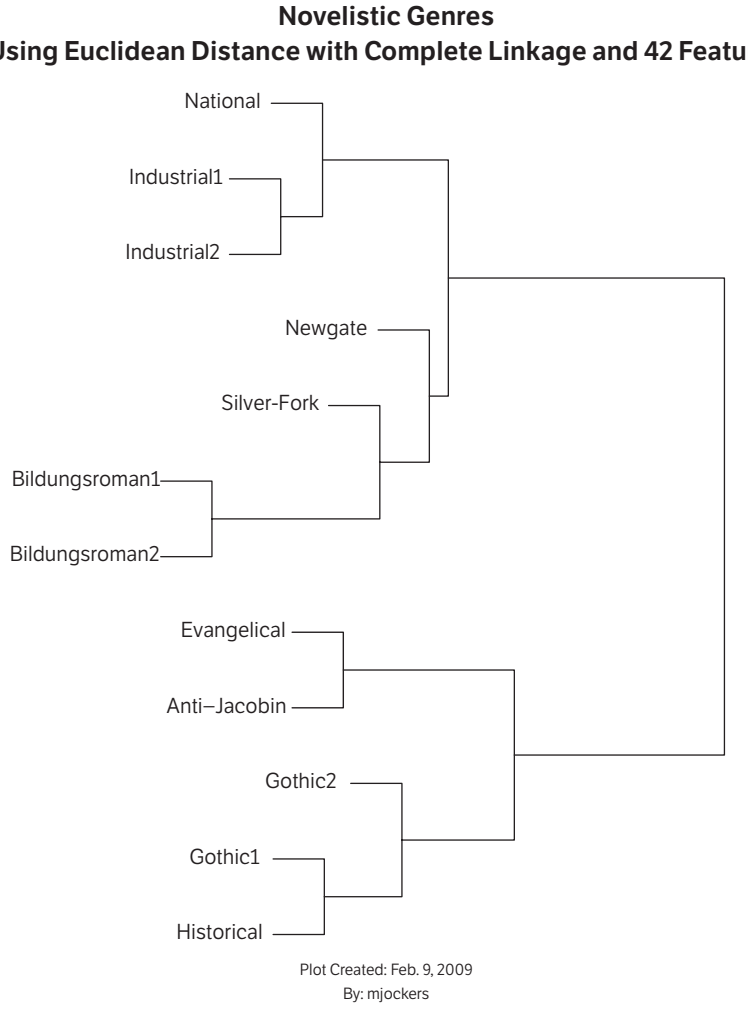



**Figure 3.1:** Cluster Dendrogram of novel genres using Most Frequent Words (MFW).

After Jockers shared his results with Witmore, Witmore suggested testing this methodology on the Shakespeare corpus. Once again, MFW accurately clustered the majority of Shakespeare's plays into the "tragedies," "comedies," "histories", and "late plays" of figure 3.2.

"Quantitative Formalism," reads the title of this article. Formalism, because all of us, in one way or another, were interested in the formal conventions of genre; and quantitative, because we were looking for more precise – ideally, measurable – ways to establish generic differences. So, we really wanted Docuscope and MFW to do well. But so well, no one had thought possible: not only were genre signals quite strong – they were *equally strong at wholly different textual levels*: just as recognizable by Docuscope's mix of grammar and semantics, as by the handful of function words of MFW. The convergence was so clear, it was almost spooky: it suggested that the logic of genre reached a depth that no one had imagined, and no one really knew how to explain. The frequency of articles and conjunctions which allowed the identification of Newgate novels or *Bildungsromane* in text after text – could this really be essential to the functioning of a genre? Why?

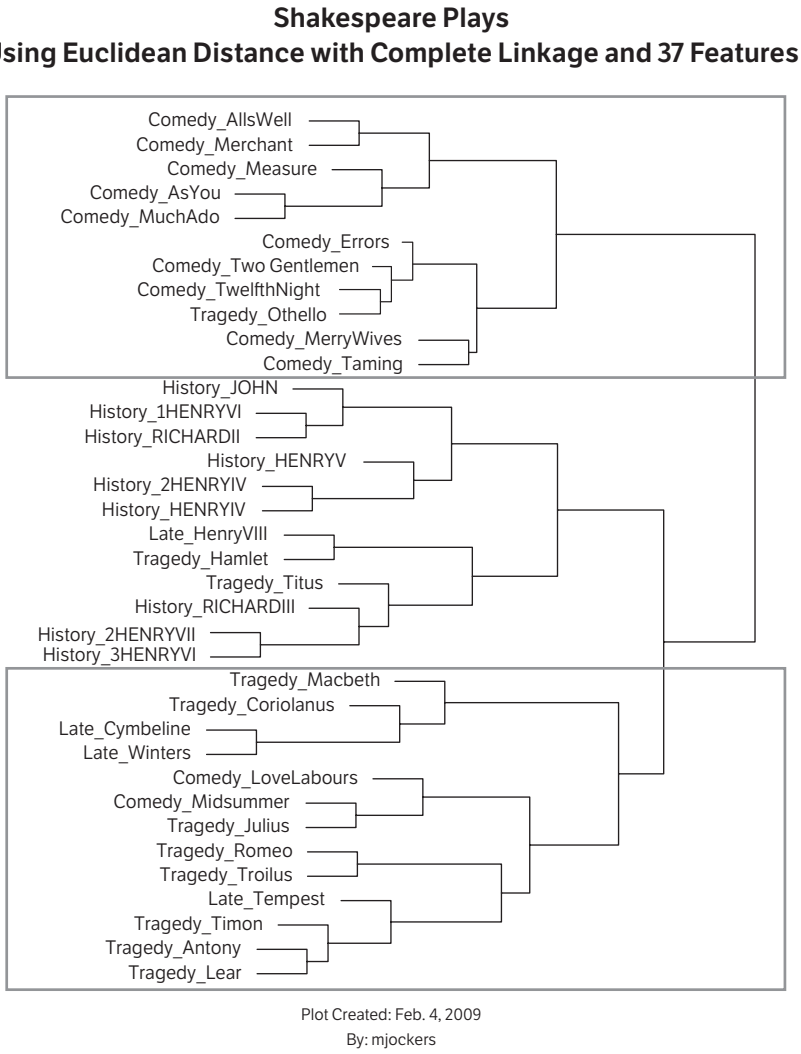



**Figure 3.2:** Dendrogram of Shakespeare First Folio plays using Most Frequent Words with major clusters highlighted. Here Jockers used the 37 features from the Shakespeare plays that had a mean relative frequency of greater than or equal to .03%. Note the similarity between this tree and Docuscope's diagram in fig. 1.1, with the close pairings of *Winter's Tale* and *Cymbeline*; *2 Henry VI* and *3 Henry VI*, and the proximity of *Coriolanus* to the *Cymbeline-Winter's Tale* pair.

As soon as school was over, we met again.

## 4. June 2009: Forking Paths

Our next meeting, at Stanford, began with Witmore showing a page that Docuscope had isolated as the most "gothic" of the entire corpus – that is to say, the one which presented an extremely high number of typically gothic features (figure 4.1):

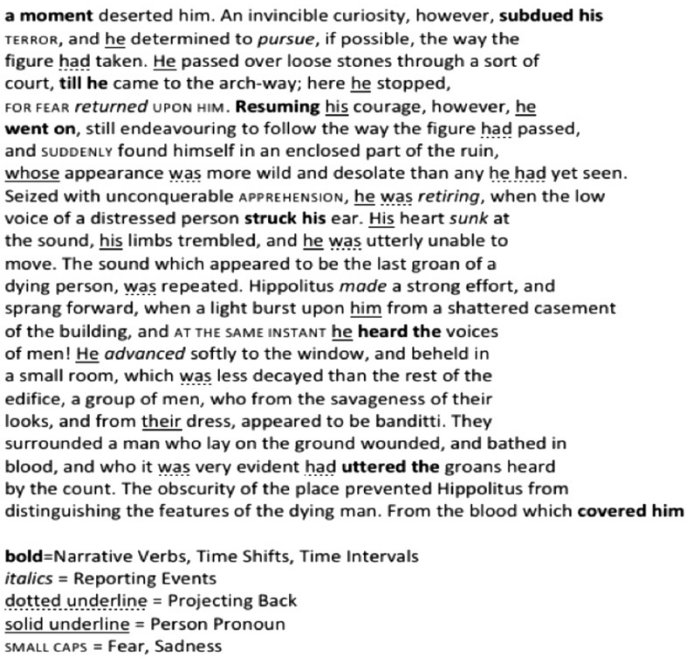



**Figure 4.1:** Docuscope screenshot of tokens differentiating the gothic from several other genres, drawn from Ann Radcliffe, *A Sicilian Romance* (1790). These differentiating bundles of LATs were identified through factor analysis and ANOVA, with factors winnowed through the Tukey test.

It was an interesting moment; not just because the idea of a genre's "typical" page was unusual and intriguing, but because, as Sarah Allison immediately pointed out, the gothic of Docuscope appeared to be quite different from that of "Humanscope" (as she called it): it was not the same gothic *we* saw. For us, that page was gothic because of the subdued terror and the archway, the ruin and apprehension and the limbs that trembled – not because of the "he" "him" "his" "had" "was" "struck the" and "heard the" which caught Docuscope's attention. Between the two approaches, there seemed to be nothing in common. Or perhaps, more precisely: nothing in common, *in terms of their units of analysis;* but everything in common *in terms of results*: whether via banditti and blood, or "uttered the" and "covered him", Humanscope and Docuscope agreed that this page belonged to the gothic, and to no other genre. And at this point, the idea that had first confusedly crossed our minds a few months earlier crystallized once and for all: genres, like buildings, possess distinctive features at every possible scale of analysis: mortar, bricks, and architecture, as Ryan Heuser, put it: the mortar, the grains of sand, of Most Frequent Words, the bricks of Docuscope's lexico-grammatical categories, and the architecture of themes and episodes that readers recognize. The three layers were not even overlapping; their signals were largely distinct from each other. Different as the three layers were among themselves, though, they were also different from the corresponding layers of other genres: the gothic "mortar" totally

unlike the "mortar" of the national tale, or the anti-Jacobin novel; the gothic "bricks" unlike the "bricks" used by other genres, and the same for the more visible architectural shapes.

We will return to the conceptual questions posed by these observations towards the end of this article. On that day in June, though, something else seemed even more inspiring: the chart we briefly mentioned at the end of section 2, which displayed all of Shakespeare plays along two orthogonal axes (figure 4.2: Shakespeare's Plays)

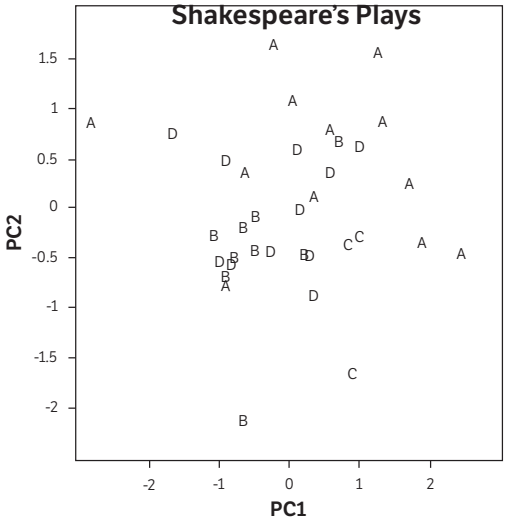



**Figure 4.2:** Scatterplot matrix in which Shakespeare's plays are rated on their first two principal components after having been counted by Docuscope and analyzed in terms of aggregates of LATs. PCA performed on the covariance matrix, unscaled data. Item key: A = comedy, B = History, C = Late Plays, D = Tragedies. Note how the two components place comedies in the upper right quadrant, histories in the lower left, and several late plays in the lower right (whereas tragedies, for some reason, are dispersed all over the field).

Witmore and Hope had abandoned the idea of publishing this diagram in a scholarly book of traditional literary criticism: they felt it would be more effective to make their point entirely with words. But the group saw in the chart the promise of an intuitive, synthetic view of the literary field, with each genre placed in relation to all the others. Moretti, in particular, was struck by the similarity between the chart and the principal components charts that Cavalli-Sforza (et.al.), in *The History and Geography of Human Genes*, had used to trace relationships among human populations.[11] Could narrative genres be similarly reduced to two basic variables? And would the ensuing distribution correlate with, say, Bourdieu's

11 See L. Luca Cavalli-Sforza, Paolo Menozzi, and Alberto Piazza, *The History and Geography of Human Genes*, Princeton UP 1994, especially pp. 39ff. Principal component analysis is a procedure, similar to factor analysis, which reduces the variance existing within a group of objects -- in our case, the linguistic-stylistic difference among literary texts -- to two orthogonal axes, called Principal Component 1 and 2 (PC1 and PC2). Principal Component 1 is the combination of features that expresses the maximum amount of variance available to a single component; Principal Component 2 displays a further increase of variance orthogonally with respect to PC1. Taken together, PC1 and PC2 are a very economical way of representing as much variance as it is possible on two dimensions; however, they never express the total amount of variance within a system, but, rather, a trade-off between high intuitive visibility and a (limited) loss of precision.

sociological (but highly subjective) map of the French literary field? Could we actually map morphology over social distinction?

Witmore's chart seemed perfect for all this. Even the fact that it wasn't perfect – with those tragedies fudging the more orderly patterns of the other genres – seemed a sign of reliability, as history is itself never perfect. So, we decided to repeat the attempt with novelistic genres. If the results were good, two further developments would become imaginable. First, the system of genres might turn from a hodge-podge of unrelated categories[12] to a single matrix of interconnected formal variables. And, second, it might become possible to chart the Great Unread – the vast, unexplored archive that lies underneath the narrow canon of literary history. One could give Docuscope and MFW thousands of texts of unknown generic affiliation, and see where they would fall in the gravitational field of better-known genres. One could envisage generation-by-generation maps of the literary universe, with galaxies, supernovae, black holes ...

With these questions running through our heads, we re-deployed the February and March data along the lines of figure 4.2. The first visualization, produced by MFW – figure 4.3 – turned out to be perfectly ambiguous: promising and perplexing in equal measure. There was certainly less clarity than in the Shakespeare case; but, we were charting twice as many genres, and over a much longer period. And then, some patterns *were* visible: with a few exceptions, gothic and historical novels lay on the negative side of principal component 1 (the left side of the horizontal axis), while the *Bildungsroman* and industrial novels were clearly on its positive side. For us, this was both good and bad news. Good, because a pattern is what one always looks for, in exploratory work. But bad, because the pattern was *chronological*, more than formal: one generation, then a second, more confused one, and then a third. Was principal component 1 capturing *genre* signals then – or historical ones? The latter seemed more likely, especially given how poorly those genres that flourished in the same years (gothic/historical; silver-fork/Newgate; industrial/*Bildungsroman*) were separated. History seemed definitely stronger than form.

But there were also some data that contradicted the historical alignment: in the crowded central section, which contained genres from two different generations, the vertical axis of PC2 – which separated anti-Jacobin and evangelical novels from Newgate stories – might be capturing genre signals after all.[13] Would it be possible to isolate such signals, and magnify them?

12 Right now, the very names of novelistic genres are a telling – even maddening – sign of categorical confusion highlighting now the novel's medium (the epistolary novel), now its content (historical, industrial), style (naturalist), protagonist (picaresque, pastoral), all the way to more or less fanciful metaphors (gothic, silver-fork).

13 Then again, with only two texts each for these genres, this could easily be the result of chance. Or not.

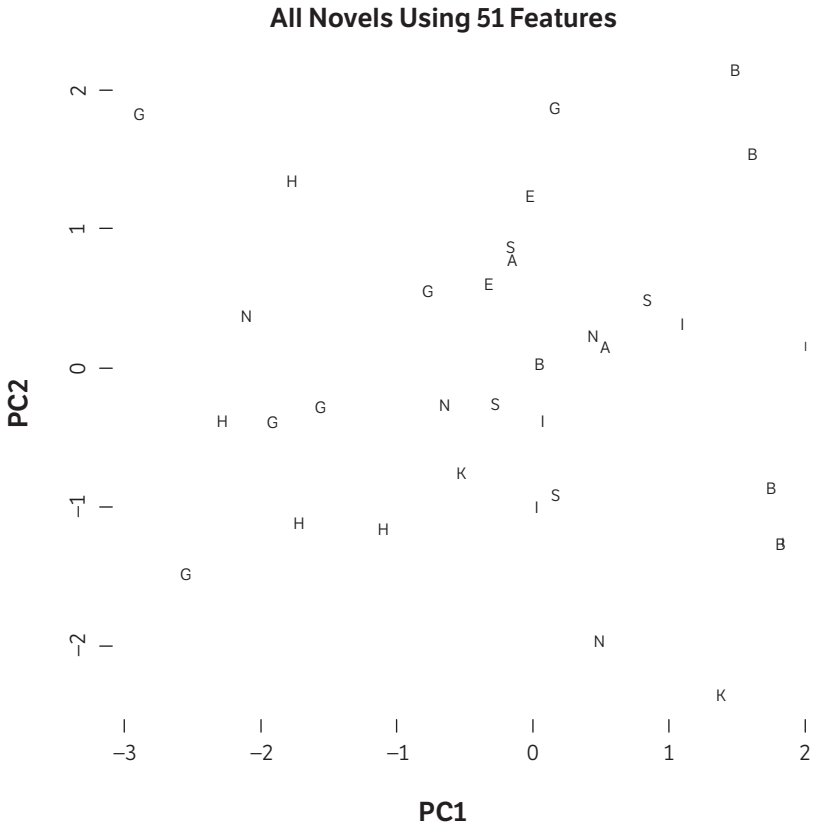



**Figure 4.3:** A graphical representation of the first two principal components in a PCA analysis of the Most Frequent Words (MFW). Each letter represents a single text (A=anti-Jacobin novels, B=*Bildungsromane*, E=evangelical novels, G=gothic novels, I=industrial novels, K=Newgate novels, N=national tales, S=silver-fork novels).

## 5. June-September 2009: Dead End

From June to September, Witmore and Jockers kept looking for ways to improve the early results of PC analysis. First, they segmented the texts to see whether smaller units would improve differentiation. All texts were divided into ten equal parts – but the results did not change much. Then, noticing that the segments' distribution was often very uneven – as in figure 5.1, where about one third of them fudge an otherwise good separation between gothic and historical novels – we decided to label all the segments: "Historical.8.1" would indicate the first segment of *Windsor Castle* (which happened to be the eighth historical novel in our corpus); "Gothic.1.10" the tenth segment of *Vathek* (which was the first gothic text), and so on. The overlap among different genres might turn out to be limited to specific portions of the texts (beginnings, or endings); if that were so, and genres became more distinctive – more "themselves", as it were – at specific moments in the plot, then one could focus on those moments and magnify their separation. It was a plausible, perhaps even an ingenious hypothesis. But – no. Some novels were most distinctive early on; others, late in the plot; or in the middle; or nowhere in particular.

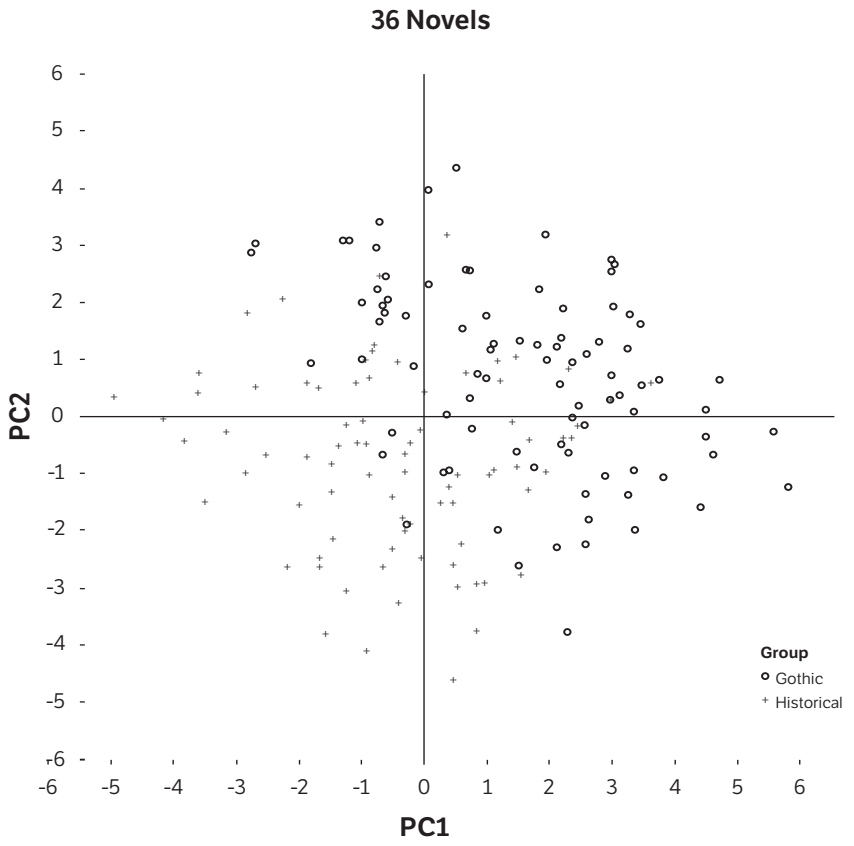



**Figure 5.1:** 8000-word segments of the first two groups of 36 novels, rated by Docuscope on first two principal components. In all PCA analyses below, data are scaled (i.e., PCA is performed on the correlation matrix of percentage scores).

Next, we turned to the composition of our corpus: as explained in footnote 6, the initial collection of 36 texts tended to exaggerate variation within each genre, making life unnecessarily hard for Docuscope and MFW. We returned to the Chadwyck-Healey database and added to the initial corpus all those texts that existing bibliographies had assigned to specific genres; included two new genres (Jacobin and sensation novels); and repeated all the calculations on the new corpus of 106 texts.[14]

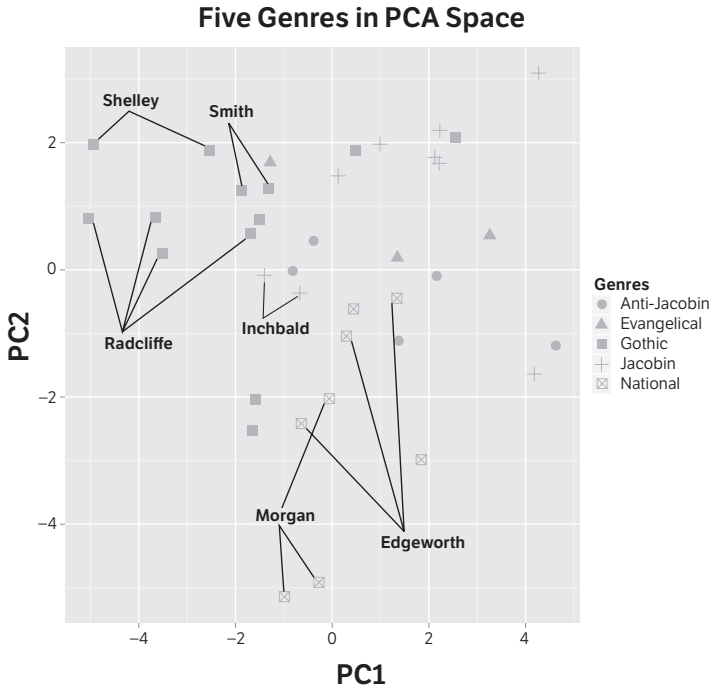
Nothing.

14 This second corpus also included a few texts, mostly from "minor" genres, scanned for us by the Stanford libraries. Since however the Chadwyck-Healey database remained the major source, canonical texts still predominated: of 28 historical novels, for instance, 14 were by Scott.

Maybe trying to chart eight decades at once was too much. We divided the corpus into three generations;[15] though of course less crowded, the new charts were just as indecisive. By the end of summer, it was clear that the results were no longer changing.

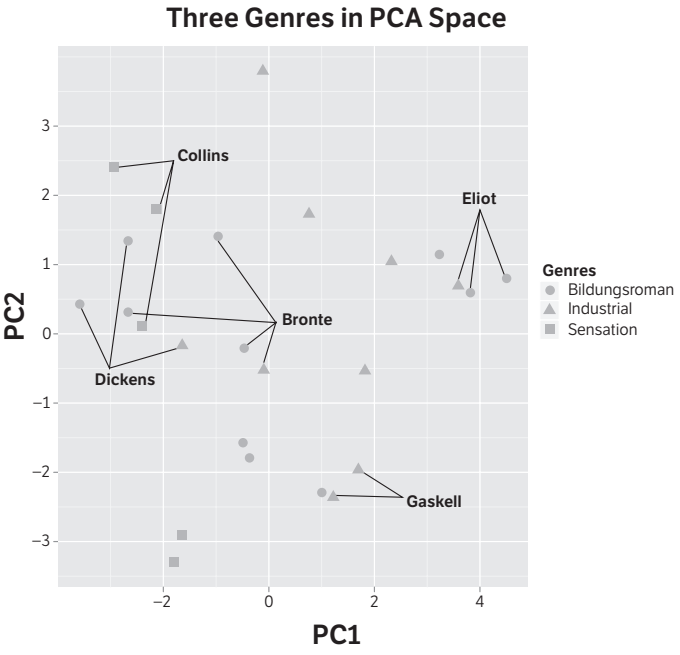
## 6. November 2009: Authors vs. Genres

In November, in the course of a teleconference which included the five authors and a few Stanford grad students, we looked again at the 3 generational maps, which now included all individual texts (figures 6.1-3), and all of a sudden realized how strong the "author" signal was. Remember, we didn't want authors; we wanted genres. But it was impossible not to notice that Docuscope and MFW clustered the former much better than the latter. With Dickens, Brontë, and Eliot, for instance – who had all written both industrial novels and *Bildungsroman* – the "pull" of the author in figure 6.3 was clearly much stronger than that of the genre; and the same was true for Bulwer-Lytton's *Last Days of Pompeii*, *Eugene Aram*, and *Pelham*, closely clustered together in figure 6.2, despite the fact that they belonged to the rather different genres of historical, Newgate, and silver-fork fiction.




**Figures 6.1-3:** Generational analysis of original 36 novels as rated by Docuscope on first two principal components. Notice the proximity among the texts by Inchbald, Smith, Radcliffe, Shelley, Morgan, and Edgeworth in 6.1; by Ainsworth, Porter, Lytton, Galt, and of course Scott, in 6.2; by Gaskell, Dickens, Brontë, Collins and Eliot in 6.3.

15 The first generation (ca. 1790-1820) included gothic, Jacobin, anti-Jacobin, national tales, and evangelical novels; the second (ca. 1815-1850) historical, silver-fork, and Newgate novels; the third (ca. 1845-1875) industrial, *Bildungsroman*, and sensation novels.


**Figures 6.2-3:** (See caption for 6.1)

Why should authors be so much more recognizable than genres? Probably, because Docuscope and MFW are very good at capturing something all writers do, whether they know it or not: using imperceptible linguistic patterns that provide an unmistakable stylistic "signature". Genres also have such stylistic signatures, of course; but genres have a

*narrative* signature too – their plot – which is at least as important. The episodes that so powerfully identify the *Bildungsroman* for instance – discussions with old mentors and young friends, false starts, disappointments, the discovery of one's vocation … – all this has no equivalent in a sensation novel; just as a sensation novel's mysteries and murders would make no sense in an industrial novel, and so on. So, what happens when the same writer moves from one genre to another – when, say, Dickens moves from the industrial novel *Hard Times* to the urban multiplot of *Little Dorrit*, the historical *Tale of Two Cities*, or the *Bildungsroman* of *Great Expectations* – what happens is that *his plots change, but his style doesn't.* Or not as much. The stories of Coketown, London, or Paris are much more different than the words Dickens uses to narrate them. His language remains basically the same.

Why did Docuscope and MFW recognize authors so well, then – and genres less well? Because they had been designed to recognize language, but not plot.[16] They were probably doing the best that could be done in separating genres on the sole basis of their language and style; but language and style are just not enough to delimit a genre from another. And after all, why should they be? In addressing their readers, genres use both style and plot (in the nineteenth century, probably, more plot than style): our programs were missing half of the structure, and it made sense that they should be only half successful. Half successful does not mean un-successful. But it does suggest that an analytical tool capable to quantify plot is still missing.[17] And as long as that is the case, the generic distribution effected by Docuscope and MFW was too random to support a good literary taxonomy, let alone an exploration of the archive. The Great Unread would, for the time being, remain unread.

## 7. December 2009: 220 Charts

In December, Allison, Heuser, and Moretti turned to a new set of visualizations: two series of charts that included all possible pairings among the 11 genres of the enlarged corpus (gothic/Jacobin, gothic/anti-Jacobin, gothic/national tale, and so on, all the way to the other end of the chronological spectrum). These charts came in two forms; the first showed the distribution of two genres based on MFW (figure 7.1) and Docuscope (figure 7.2). These were our basic tools, allowing us to intuitively grasp whether two specific genres separated well – as gothic and sensation novels in figures 7.1-2 – or not. (MFW and Docuscope, incidentally, turned out to be equally able – or unable, as the case may be – to separate genres from each other.)

16 They can certainly see how actions are described: with simple or complex sentences, stressing subjective mood or objective results, surprise or retrospection. But they can hardly see what actions consist of: a story's chronological (and semantic) chain largely eludes them.

17 This finding cheered Witmore, since it suggests that in novelistic representation, plot provides an avenue of generic differentiation that *has to be* less visible to Docuscope because it does not have to be tied to the physical limits of the medium, whereas Renaissance drama – constantly grappling with the difficulty of telling stories with real bodies in a few hours – might have this extra-stylistic avenue foreclosed, leading to more legible (because materially constrained) generic styles at the level of the sentence.

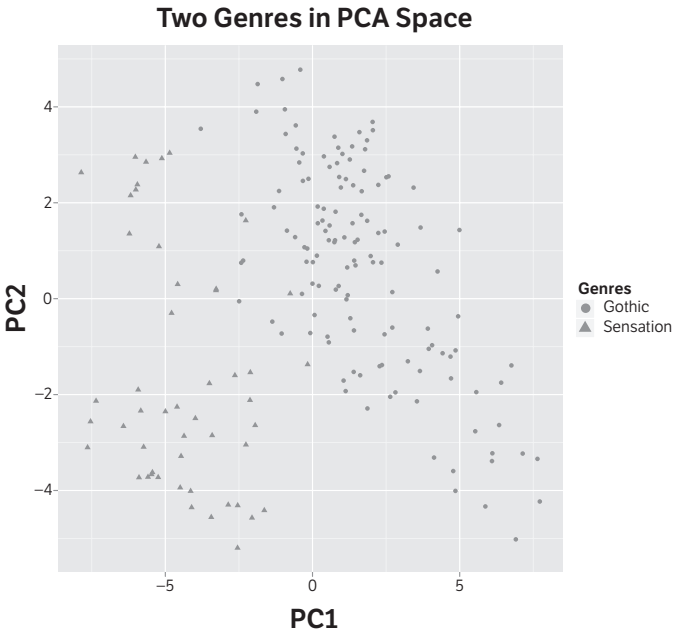



**Figure 7.1:** Most Frequent Word scatter plot. Here, and in all other PCA charts, each point (circle or triangle) on the plot stands for one segment (one tenth) of a text.

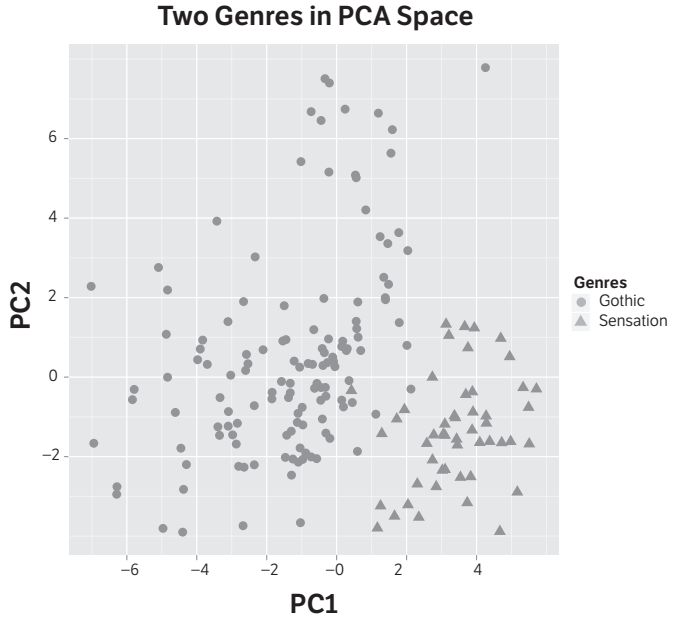



**Figure 7.2:** Docuscope scatter plot.

The second type of chart re-deployed the circles and triangles of figures 7.1-2 adding two further features. First, it tagged each segment, making explicit which (part of which) text it came from: the circles in the lower right corner of figure 7.1, for instance, turned out in figure 7.3 to belong to *Vathek*, thus bringing to light the "centrality" – or "eccentricity", as the case may be – of each text within its genre (an issue which

may have profound consequences for our knowledge of genre, and which we plan to investigate in the future). And then, figures 7.3-4 also indicated which traits of the two principal components contributed to the specific shape of a genre's distribution: which words, or Docuscope Dimensions exerted a stronger pull in separating gothic from sensation novels. So, for instance, the lower right quadrant of figure 7.3 highlights the definite article as an important differential feature of the gothic in MFW analysis (compare with figure 7.1); in figure 7.4, a similar role is played, in the lower left quadrant, by "Narrative VP", "Pronouns", and "Reporting Events" (compare with figure 7.2).[18]

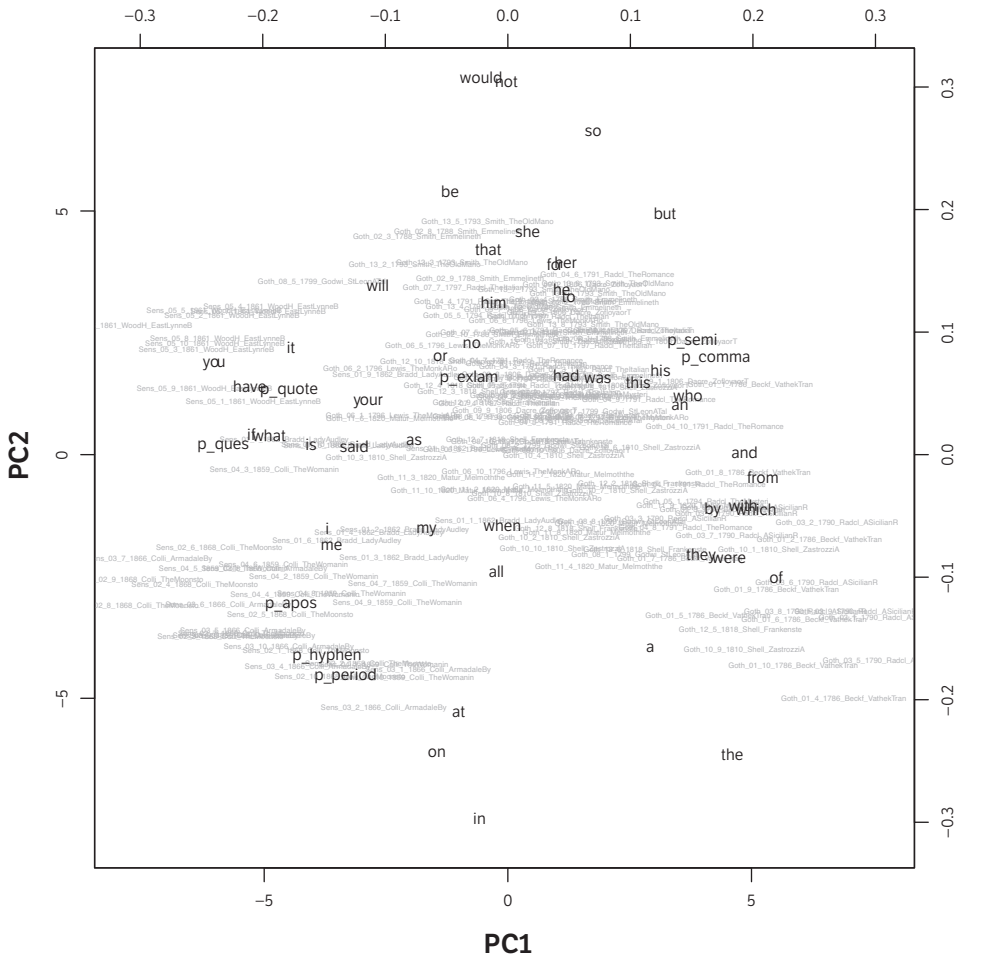



**Figure 7.3:** Most Frequent Word scatterplot (light grey titles) and component loadings (black).

18 As each of the 55 genre pairings appeared in this double form, we examined 110 charts produced by Docuscope, and 110 produced by MFW. The mapping technique used in figs. 7.3-4, in which differential traits become visible within the distribution of the data themselves, is described in Mick Alt, *Exploring Hyperspace: A Non-Mathematical Explanation of Multivariate Analysis*, McGraw-Hill, London-NY 1990, chapter 4.

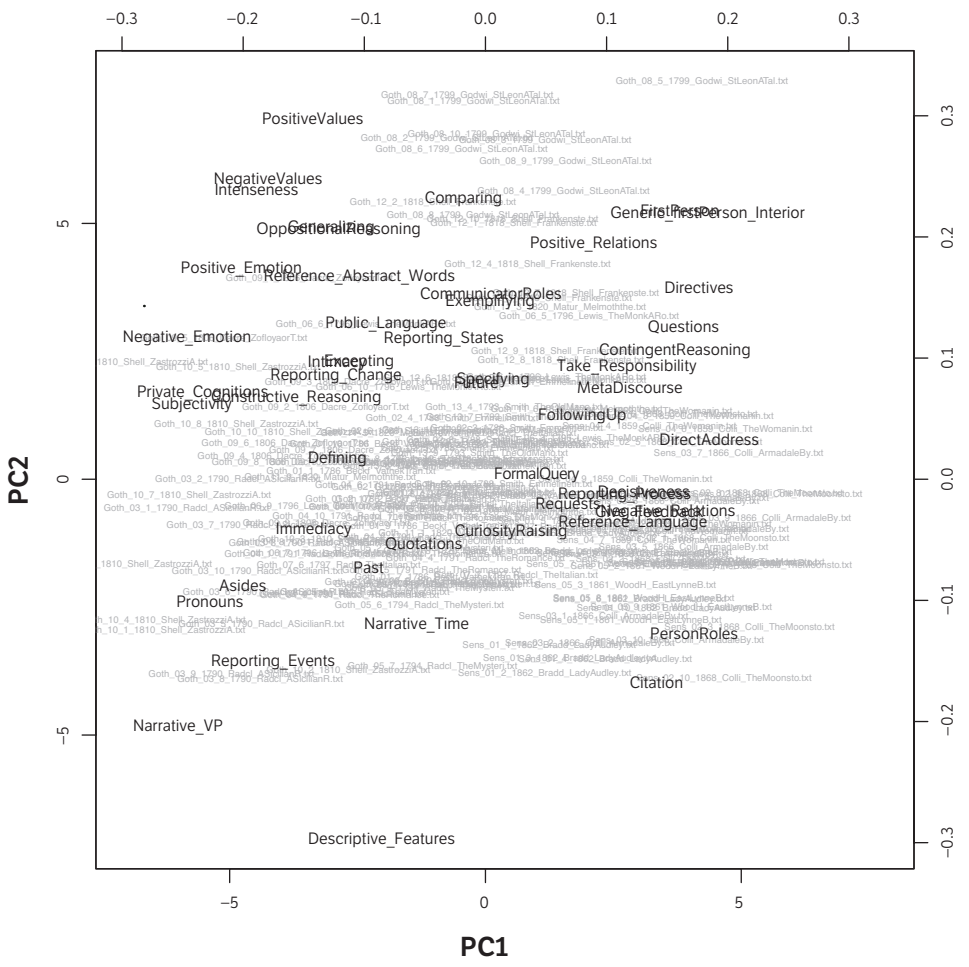



**Figure 7.4:** Docuscope scatterplot (light grey titles) and component loadings (black).

As we studied our charts, it became clear that they rested on two premises that were quite different from those of current genre theory: they never looked at a genre *per se*, in isolation, but always and only in relation to another genre; and they were not interested in those features that could add up to a synthetic ideal-type, but only in those that could *differentiate* one genre from an another. This relational-differential emphasis made for a very "realistic" approach, reminiscent of Bourdieu's "position-taking": just like authors or schools, genres engage in a struggle for recognition: one could almost feel, not just the difference, but the *conflict* of forms in those traits that pulled them in one direction or the other. And yet, this image of genre was clearly also incomplete, because differential features may tell us all we need to know in order to demarcate one form from another, and yet very little about that form's inner structure. If all men in an audience wore pink, and all women blue, the colours would differentiate them *perfectly*, and tell us *nothing* about them. We'll return to this point at the end of the article.

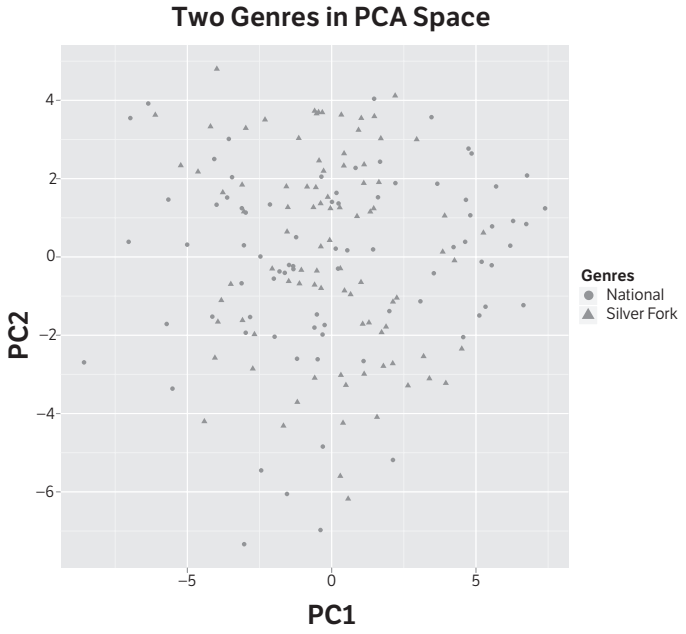



**Figure 7.5:** Most Frequent Word scatterplot of two genres rated on first two principal components.

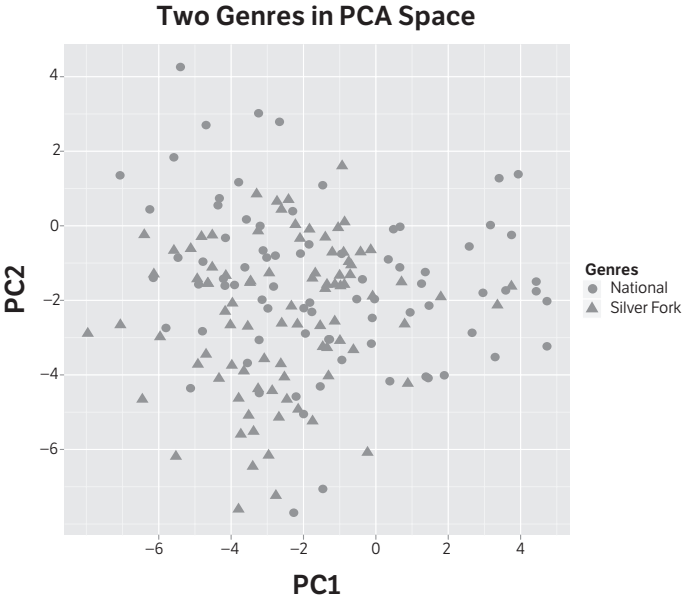



**Figure 7.6:** Docuscope scatterplot of two genres rated on first two principal components.

Now, one thing that the charts made clear was the *variability* of genre signals: quite strong in figures 7.1-2, for instance, but rather weak in about one fourth of the cases – like figures 7.5-6, where neither MFW nor Docuscope managed to extricate national tales from silver-fork novels. Why some genres should be so hard to separate – especially in a case like this, where the difference, intuitively, ought to be quite vivid – was an intriguing question; but we decided to leave it for another study, and focus instead on a group of charts where

the separation was rather good, and dependent on a recurring set of traits: the pairings of gothic novels with the three "ideological" genres – Jacobin, anti-Jacobin, and evangelical novels – that were their short-lived contemporaries.[19] Since the charts were all similar, we reproduce here only the gothic/Jacobin pairings: figures. 7.7-8, based on MFW, and figures 7.9-10, based on Docuscope and its Dimensions.

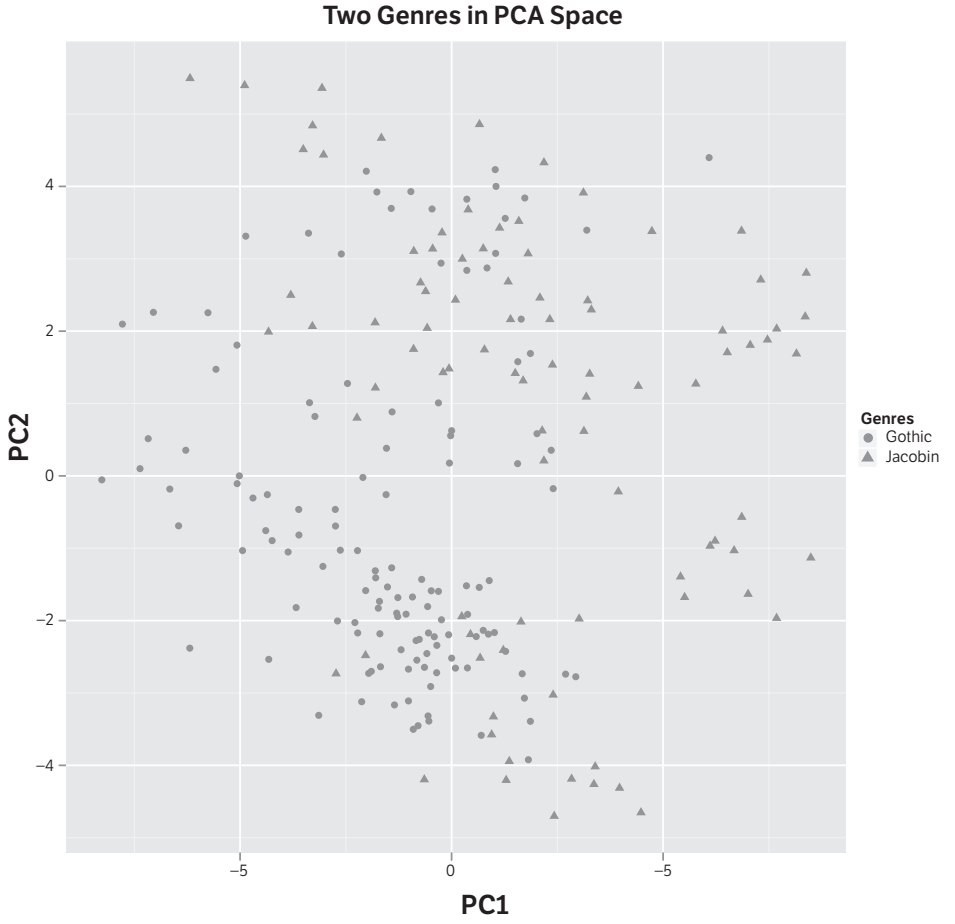



**Figure 7.7:** Most Frequent Word scatterplot of two genres rated on first two principal components.

To better understand the relationship between the two genres—and to begin to put the figures into language—we looked closely at the features that were particularly effective at separating gothic from Jacobin along the first principal component (PC1: the x-axis in figures 7.7-10). A principal component ranks the likelihood of certain features occurring, so texts are sorted according to the features they lack, as well as by the features they have.

19 One of our problems was that we had automated our comparisons, using only the first two (and therefore, most powerful) principal components to pull apart the genres. Of course, PCA generates multiple components and there are ways of establishing (for example, the Tukey test) whether any given component sorts two groups. But we wanted some raw measure of "sortability" among pairs, which is what led us to simply profile all of the pairs on their first two components and leave other – potentially quite powerful – components aside.

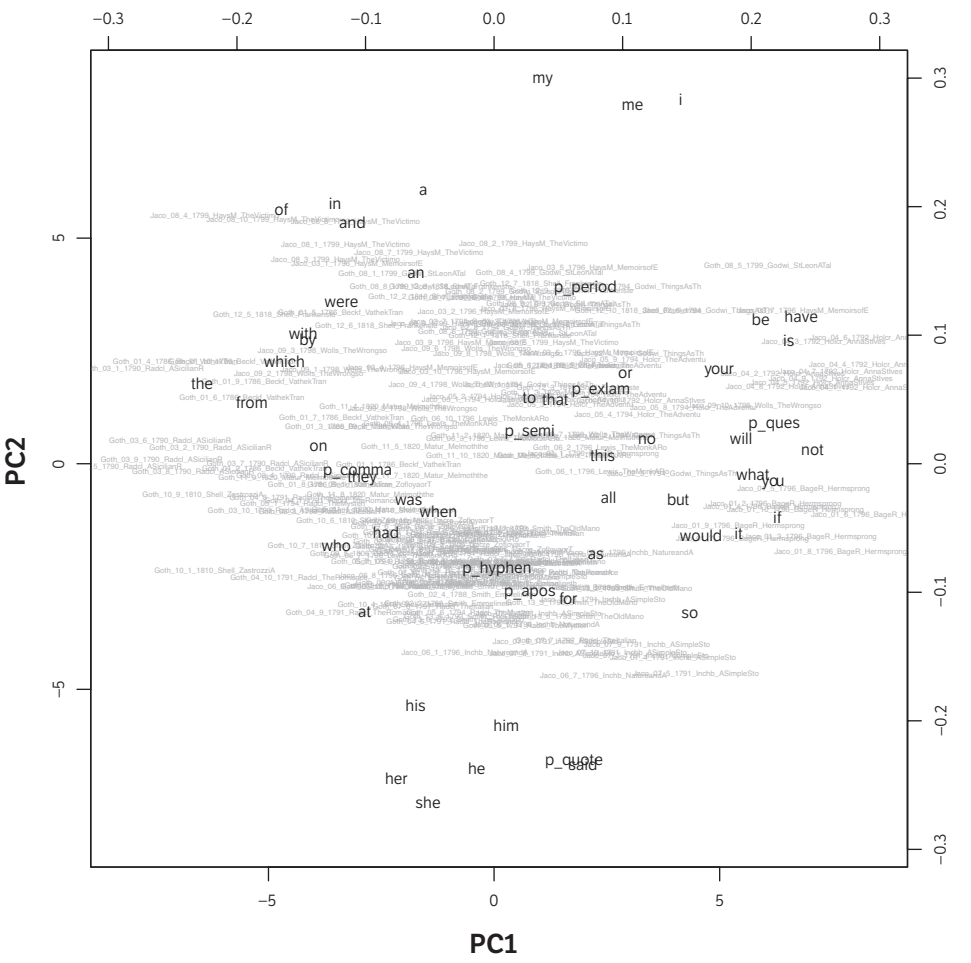



**Figure 7.8:** Most Frequent Word scatterplot with titles (light grey) and component loadings (black).

Roughly speaking, we found that the gothic novel averages less talk and more action than the Jacobin. Words and phrases that characterize gothic texts show a marked "narrative" inclination: past tense and pronouns; spatial prepositions; and words marked by Docuscope as "Narrative Time" (for example, "whilst," "when he," "as he"). See in MFW, on the left side of figure 7.8: was, had, who, she, he, her, his, they, the ubiquitous the, and the large group of locative prepositions from, on, in, at; in Docuscope, on the left side of figure 7.10, see Narrative VP (for example, "heard the," "reached the," "commanded the") and Pronouns. Markers associated with oral discourse, on the other hand, tend not to occur in gothic novels: a foregrounding of the addressee (you, your), questions, polemical markers (but, no), and verbs inflected in the present, future and conditional. In MFW, note on the righthand side of figure 7.8 the cluster of you, p_ques [?], but, if, not, is, will, and would; in Docuscope, on the right of figure 7.10, see Questions, Oppositional Reasoning ("not," "but," "however"), and, just below it, Directives ("should," "must," "you will soon"). Though Jacobin texts generally tack in this direction, they are more scattered than the gothic ones. What separates the genres here seems to be, not so much the absence of narrative in Jacobin texts, but the presence of talk, something like an argumentative style.


**Figure 7.9:** Docuscope scatterplot of two genres rated on first two principal components.

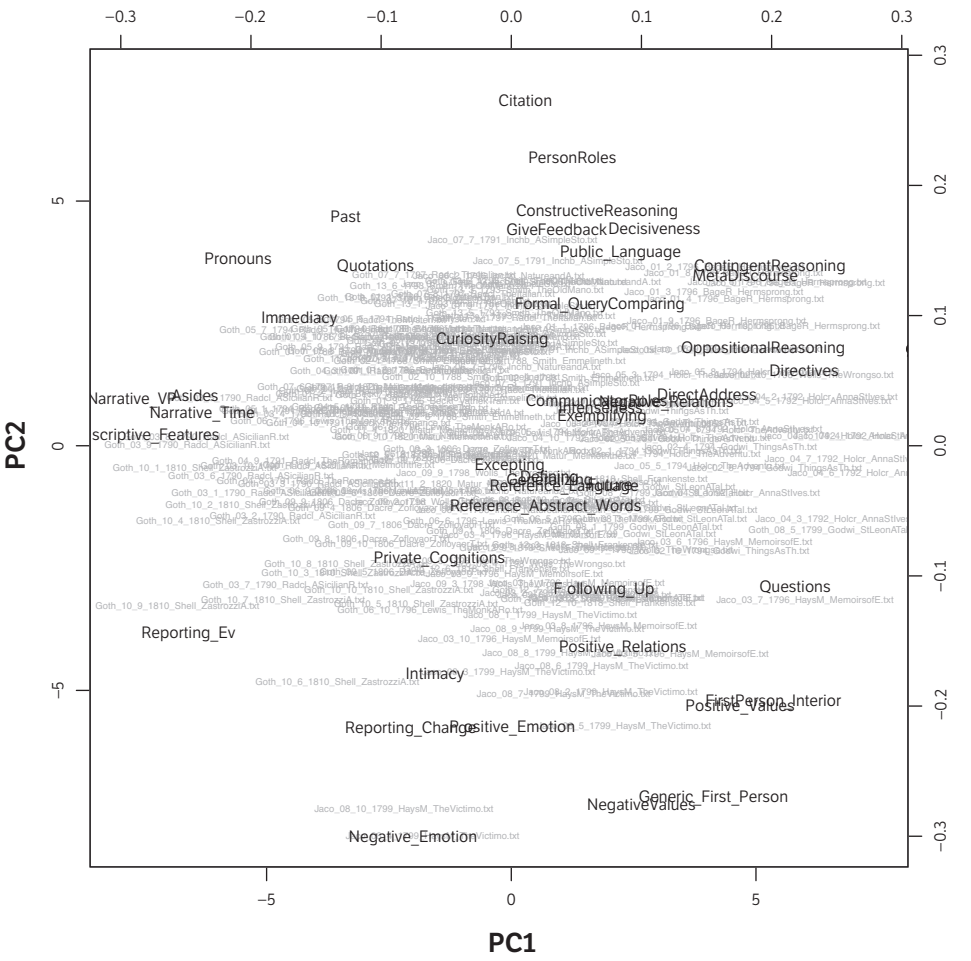



**Figure 7.10:** Docuscope scatterplot with titles (light grey) and component loadings (black).

Such, then, were the raw data that our analytic techniques had placed in front of us. Could they become good interpretive questions? We tried. Noticing, for instance, the high frequency of the conditional in the ideological genres – where, indeed, possibility is important – Jockers and Moretti compiled a list of the (more or less) 13,000 sentences that included "would"; looked at the associated pronouns, adjectives, and adverbs; at the types of verbs involved; at the negative forms, the past tense … A few results stood out: "would+never" occurred twice as often in the gloomy evangelical novels than elsewhere, for instance; and the impersonal pronoun "it" was 50% more frequent in Jacobin and anti-Jacobin novels – full of abstract discussions of principle – than anywhere else. Both findings made perfect sense. But were they also surprising? They certainly corroborated and enriched existing knowledge of the genres in question. Did they also *change* it?

## 8. March 2010: Experiments, Explorations, Hypotheses

In March, we met for one last retrospective glance at a year of work. Why had we turned to Docuscope and MFW in the first place? Because we were looking for an explicit, quantifiable way to assign texts to this or that genre. It was, in part at least, a matter of attribution. Attribution ... "To trace every piece to its real creator", writes Carlo Ginzburg,

> we should not depend (...) on the most conspicuous characteristics of a painting, which are the easiest to imitate: eyes raised towards the heavens in the figures of Perugino, Leonardo's smiles, and so on. We should examine, instead, the most trivial details that would have been influenced least by the mannerisms of the artist's school: earlobes, fingernails, shapes of fingers and of toes.

Earlobes, fingernails ... It is in these "involuntary signs," Ginzburg continues,

> in the "material trifles" – a calligrapher might call them "flourishes" – comparable to "favorite words and phrases" which "most people introduce into their speaking and writing unintentionally, often without realizing it", that Morelli recognized the surest clue to an artist's identity.[20]

Involuntary signs: this is certanly what MFW and LATs are. But are they *just* that? Because, clearly, there is a problem with earlobes and fingernails: good as they might be at identifying the author of a painting, they are worthless at explaining its meaning. In fact, they are good at the one *because* they are bad at the other: it's only because "trifles" have no structural function, that authors let go and "write unintentionally, without realizing it" – thereby betraying themselves. If those words were important, they would be more careful.

There is something paradoxical in these traits that classify so well, and explain so little. Especially so in our case: because, after all, MFW and LATs were in at least one respect the very opposite of earlobes and fingernails: instead of being rare and peripheral details, they were so frequent as to be almost ubiquitous. And how could such pervasive traits tell us nothing about the structure of genre? It was possible, of course, that it was all our fault; that, although we had managed to isolate the data, and were probably the first to "see" them, we just didn't know how to make sense of them. Possible; and we are ready to place our data at the disposal of others, who may obtain better results.

But there is also a simpler explanation: namely, that these features which are so effective at differentiating genres, and so entwined with their overall texture – these features cannot offer new insights into structure, *because they aren't independent traits, but mere consequences of higher-order choices.* Do you want to write a story where each and every room may be full of surprises? Then locative prepositions, articles and verbs in the past tense are bound to follow. They are the *effects* of the chosen narrative structure. And, yes, once Docuscope and MFW foreground them, making us fully aware of their presence, our knowledge is analytically enriched: we "see" the space of the gothic, or the link between action verbs and objects (highlighted by the frequency of articles), with much greate*r* clarity. But, for the time being, the gain seems to be comparative more than qualitative: greate*r* clarity, rather than clarity of a different type.

20 Carlo Ginzburg, "Clues", in *Clues, Myths, and the Historical Method*, Hopkins UP 1989, pp. 96-7, 118.

We started with an experiment: testing the classifying power of Docuscope in a new and controlled setting. The experiment then turned into an exploration: Docuscope and MFW, charting the field of novelistic genres, and their inner composition. "Exploratory Data Analysis", as John Tukey has called it: detective work, focusing on clues that lead to new questions, and a broader understanding of the data. Statistical findings, said Heuser, made us realize that genres are icebergs: with a visible portion floating above the water, and a much larger part hidden below, and extending to unknown depths. Realizing that these depths exist; that they can be systematically explored; and that they may lead to a multi-dimensional reconceptualization of genre: such, we think, are solid findings of our research. Now, more explorations are on the horizon: the switch from unsupervised to supervised techniques, for instance; or the explicit inclusion of semantic data, which we have so far mostly avoided so as to focus more strictly on the formal properties of genres. And then, at the end of it all, the great challenge of experimental work: the construction of hypotheses and models capable of explaining the data. This study is a step in that direction.

## About Us

The Stanford Literary Lab, directed by Matthew Jockers and Franco Moretti, discusses, designs, and pursues literary research of a digital and quantitative nature. The Lab is open to all students and faculty at Stanford - and, on a more ad hoc basis, to students and faculty from other institutions.

We envisage a variety of projects, ranging from dissertation chapters to courses, individual or group publications, conference papers and panels, and even short books. Ideally, research will take the form of a genuine "experiment," and extend over a period of one or two years. On our website (litlab.stanford.edu) you will find a list of our present activities, most of which gather together several projects, and are open to further collaboration. We plan to initiate two more experiments in 2010-11, and add another two in 2011-12.

At the Lab, all research is collaborative (even though some outcomes may end up having a single author). We hold regular group meetings to evaluate the progress of a specific experiment, the status of existing hypotheses, and future research developments. (If interested in these meetings, please contact Jockers or Moretti: as a rule, visitors are welcome.) Occasionally, we will have public presentations of our research, which will be announced on our website under "Events".

**January 2011**


litlab.stanford.edu