Vive la différence: tracing the (authorial) gender signal by multivariate analysis of word frequencies.

Jan RYBICKI
Institute of English Studies, Jagiellonian University, Kraków, Poland

Motto:
(a) Oh dear, you've put the peanut butter in the refrigerator again.
(b) Shit, you've put the peanut butter in the refrigerator again.
(Lakoff 1973, 50)

Introduction

Quite recently, I was hired, if that is the correct word, to check if a corpus of 18th-century English novels by women might not contain, among its several anonymous works, one or more books by a man; and then to see in what way the women's novels in the same corpus, now virtually unknown, differed from the more famous female writers of the same era. The commission came from the COST Action IS0901 "Women Writers in History: Toward a New Understanding of European Literary Culture" (2009-2013) headed by Prof. Suzan van Dijk, to whom this paper is gratefully dedicated. A part of this text is derived from my papers given at two of the COST Action conferences: "Transcultural, Transnational, Trans-disciplinary Perspectives on Women's Literary History," Poznań, 26-28. Nov. 2012, and "European Female Authorship: Networks and Obstacles," The Hague, 19-21 June 2013.

Any discussion of the nature of the differences – and, indeed, of their very existence – between men and women is risky business, and that for a variety of reasons; in fact, as I write these words, my mind rebels against the phrase "men and women" thrust upon me by the original language of this paper, English: were it not a set phrase (which it is), it should have been the other way round, as it almost always is in my native Polish: "*kobiety i mężczyźni*." After all, this happens to be the Polish title of Françoise Giroud's and Bernard-Henri Lévy's celebrated book, *Les Hommes et les femmes* (1993); its translator, Kalina Szymanowska, must have felt like I do now. Still, there is more than one reason to try to eschew, in this paper, all ideological pre-assumptions, choosing instead the path of manly empiricism: the experiment first, the results, if any, second, the discussion, third.

And while this, too, will be construed as an ideological decision, it makes sense to show where multivariate word-frequency analysis stands on the presence and the whereabouts of the gender signal – which, together with those of genre, chronology, theme – often shows up uninvited when all a stylometrist wants to see is authorial attribution. In fact, separating and identifying all these signals is a major challenge for computational stylistics in the nearest future, and there is still much work to do despite Matt Jockers's full chapter focused on this issue in his *Macroanalysis* (2013).

Since there is little in terms of a theoretical basis that would explain why and how most frequent word frequencies usually work so much better in authorial attribution than any other features, one has to do with what one has. What stylometrists do have outside stylometry when the gender signal is concerned is James Pennebaker's *The Secret Life of Pronouns* (2011). No wonder, then, that it has been received by the stylometric community with some enthusiasm, as evidenced by a recent review in this journal:

> The book deserves a review in LLC because it pays attention to linguistic and literary interests, and especially because it adds an interpretive dimension to stylometry (the study of style using exact techniques), which has been underdeveloped to-date …. As readers of this journal know, stylometry has come to focus increasingly on authorship attribution as an objective validation of its work, and has come to accept … that function word distributions are the most interesting indicators of authorship. I will criticize Pennebaker a bit later in the text for largely ignoring the stylometric literature, but I will focus on what he does contribute, and that is a great deal (Nerbonne 2014, 140).

Indeed, Pennebaker comes to the issue central to this paper in his Chapter 3, "The Words of Sex, Age, and Power," when he says: "Women use first-person singular, cognitive, and social words more; men use articles more; and there are no meaningful differences between men and women for first-person plural or positive emotion words" (Pennebaker 2011, 40). But then he returns to this list and extends it somewhat: "Men use more big words, nouns (which is another way of saying that they use more articles, comment mine), prepositions, numbers, swear words. Women use more personal pronouns, verbs (including auxiliary verbs), negative emotion (especially anxiety), negations (no, not, never), certainty words (always, absolutely), hedge phrases ("I think," "I believe")" (43). More importantly, he takes the discussion from real life to literature (more precisely, to drama and film scripts) and proposes "a nine-point masculinity-femininity language scale" (49). Measured on that,

> Shakespeare and Tarantino are males and write like males. Their male and female characters use function words the ways males do. The two writers may share the same stealth word usage, but they clearly differ in the content and breadth of what they write. Shakespeare is of interest because he

> brilliantly conveys real-life themes and concerns women have. But his use of function words, much like Tarantino's, suggests that he fails at getting inside the minds of women (56).

In this, Pennebaker makes a vital point. It is obvious that, in literature, unlike in life, we must expect shifts from the author's gender signal to that of his or her narrator, and more obviously from character to character; authors might or might not be able to conceal their gendered language from frequent-word analysis. One could even risk making value judgements based on that; clearly, Pennebaker feels justified to do so, and the authors of *Pericles* and *Pulp Fiction* both seem to have failed the test. This of course makes one wonder whether anyone might pass at all.

Within stylometry, the gender signal has been traced, with success, in anything from political speeches (Dahllöf 2012, Yu 2014), spoken (Singh 2001, Iyeiri et al. 2011) and formal written language (Argamon et al. 2003, Mikros 2013) to blogs (Schler et al. 2006, Mikros 2013a) and celebrity tweets (Mikros 2013b). In fiction, the most notable work has probably been by Koppel, Argamon and Shimoni (2002), who achieved a ca. 80% success rate in identifying authorial gender. Interestingly, their results for fiction did not differ very much from those for non-fiction, and this suggests that Pennebaker may be right: most authors seem unable to fake gender signals; males cannot "write women," and vice versa. I am personally and painfully aware of this myself: of the thirty novels I have translated from English to Polish in my previous lifetime, only three were by women, and all three (especially Nadine Gordimer's *None to Accompany Me*) were accompanied by a constant fear, in the hapless translator, of writing "male" rather than "female." Of course, stylometric research, too, makes one very conscious of that – the more so as it is, more often than not, applied to complete novels, while *cherchez la femme* should be more reasonably focused on character idiolects. It is a truth universally acknowledged that this is where the adventure of modern stylometry beyond authorship attribution really began with *Computation into Criticism* (Burrows 1987; see also Rybicki 2006). But this creates even more problems: in a standard novel, very few characters speak more than 10,000 or even 5,000 words (Rybicki 2008), respectively the really-safe and almost-safe limits for the most-frequent-words approach (Eder 2013), and it is often very difficult to find a heroine who would come close to this value in, say, historical romances – if the book was written by a man, and the same is true of Shakespearean women (Rybicki 2007). In fact, judging authors on their success in transgressing the linguistic gender divide seems best reserved to multiple-narrator novels: it takes *Wuthering Heights* or *Bleak House* or *Ulysses* to give the stylometrist enough material for idiolectic comparison between genders.

But that is already a slightly different story. If we were to return to our muttons and to the task at hand, any study dealing with similar issues must mention the paper by Mark Olsen, who traced early female writing and its deliberate creation of "a distinct female literary voice" in his "*Écriture*

féminine: Searching for an Indefinable Practice?" (Olsen 2005, 147). Olsen performed a comprehensive analysis of the problem from frequencies of works in a selection of genres for the two genders to identify "male" and "female" words; a chronological perspective was also adopted, so that these phenomena were observed over five centuries.

Then there is Matt Jockers's cautionary tale: when he searched for the relative impact of various signals in stylometric analysis, he found that "classification tests and … linear regression tests showed gender to be a bit player," since it only accounted for some 8% of the overall results (Jockers 2013, 99). And while he states that "when it comes to the nineteenth-century novel, it is not terribly difficult to separate the men from the women" (this is not at all counterintuitive, since genders, also literary, were still then kept in fairly strict *apartheid*), he provides a list of those writers of the 19th century who were notoriously difficult to gender-classify – and this list includes some of the protagonists of this study: William Beckford, Maria Edgeworth, Matthew Lewis and William Godwin (94-95).[1]

On the surface, the task given to me by COST Action IS0901 was much simpler: it was *cherchez l'homme* in what was supposed to be a women-only corpus of 18th- and early 19th-century texts, and then to see if there were any stylometric differences between women who made it at one point into the English literary canon (before it was gone with the wind), and those who did not. "At one point" is not just a cliché here: the canonized women include Austen and Burney, but their literary fortunes followed very different paths. The author of *Pride and Prejudice* is now a veritable saint of English literature and its film adaptations, and a major object of academic study; the author of *Cecilia*, by contrast, has lost the advantage she once had over her rival. To illustrate: a popular handbook of

[1] In fact, Lewis and Godwin were both such inveterate troublemakers in the initial phases of this research that they had to be rejected from the reference corpus altogether. Lewis tended to join all other Gothic writers (Walpole's *Castle of Otranto* had to go for the same reason), male and female, who constantly formed a separate group in most initial analyses. In terms of numbers, the Gothic novel was – at least in my set of texts – strongly dominated by women, and its genre/theme signal successfully obfuscated that of gender. Beckford remained in the corpus to demonstrate this phenomenon as evident in his Gothic *Vathek*, and because of the interesting behaviour of his *Azemia*, a parody of the genre represented by many of the Chawton House novels, which stuck to its parodees throughout the study. This is nothing new, for the stylometric behaviour of parodies has already been described, as most of the good things in the field, by Burrows (2005). Godwin's behaviour was even stranger in a corpus that also contained his daughter's works (after all, *Frankenstein* is supposed to incorporate some of Mary Shelley's childhood experiences), and this will be dealt with in a separate paper, but because of even the slightest suspicion of parental meddling he, too, had to go. There are other potentially interesting children/parent issues associated with this corpus. It might be worthwhile to trace the influence, on Mary Shelley, of her mother, Mary Wollstonecraft, as well as of her father; and Maria Edgeworth's writing tends to become very Protean around the time when she collaborated on her father's autobiography…

English literature devoted, in its 1874 edition, a short paragraph to Burney and not a word to Austen; in 1891, the same short paragraph on the former was more than matched by almost three times as much on the latter (Backus 1874, 1891). Still, both writers mattered and matter more than any of the Chawton House novelists; the famous males' corpus used in this study was very well represented and described at much more length in the same venerable publication. Also, canons can change in space as well as in time: from the Polish perspective, for instance, the English canon of the 18th century must include Jane Porter, author of just two novels (apart from her play and her shorter fiction); but one of these novels is *Thaddeus of Warsaw*; it is set in Poland and is probably the first historical romance that deals with the dramatic events of Polish history – and was published before their first native literary representations in the genre.

Material and method

The corpus that I was given to study had been collected by the Chawton House Library, the renowned centre of research on early women's writing from 1600-1830, very aptly located for the place's Austenian history. It consisted of forty-six novels written between 1723 and 1830; of these, thirty-four had named authors, five of whom wrote two novels each. Twelve were anonymous. These texts were compared with two reference corpora: one contained the more famous female rivals: Austen, Radcliffe, Burney, Edgeworth, Shelley (a total of 22 novels); the other, the famous male writers of the era: Swift, Johnson, Richardson, Fielding, Sterne, Smollett, Goldsmith, Beckford, Peacock (21 novels). To answer the main two questions, various combinations of these were used.

Various combinations were also used of the functions available in *stylo* (Eder et al. 2013), the stylometric package for R, the open-source statistical programming environment (R Core Team 2014), combined, for network analysis, with *Gephi* (Bastian et al. 2009). The workflow consisted in producing bootstrap consensus trees and/or bootstrap consensus networks of cluster-analysed classic Delta distances (Burrows 2002) between most-frequent-word frequencies in texts compared (Eder 2014; Rybicki 2014, 2014a; Eder forthcoming); frequencies of medium-frequency words characteristic for the various sets of texts obtained through Burrows's Zeta procedure (2006) as modified by Craig (2009) and incorporated in *stylo*'s "oppose" function were also used as input for both types of cluster analysis consensus visualizations (c.f. Hoover[2]). The network analysis itself applied the Force Atlas 2 algorithm that is particularly useful for visualization of differences, i.e. distances, between individual texts; the input – cross-validated strength of cluster-analysis linkages

[2] This is truly embarrassing. A quick survey among fellow stylometrists around the globe has confirmed my suspicions that it was David Hoover who first voiced the idea of using Zeta words in a Delta-like procedure, so there is a consensus on that; but we could not agree (not even David) when or where exactly that happened.

– is represented in two ways: by the size of edges, or links, between the texts, and by the distance between the nodes that represent them. To quote its makers,

> ForceAtlas2 is a force directed layout: it simulates a physical system in order to spatialize a network. Nodes repulse each other like charged particles, while edges attract their nodes, like springs. These forces create a movement that converges to a balanced state. This final configuration is expected to help the interpretation of the data (Jacomy et al. 2014).

It should be noted here that I would like to avoid the ongoing running battle between proponents of various statistical methods that might or might not be optimal, or state-of-the-art, or simply fashionable in stylometric research. The sad truth is that there exists no universal consensus; that comparative studies conyinue to ascribe improvements of a percent point one way or another; and that while certain computer-intensive methods (such as SVM) might have a slight advantage over certain less intensive ones (pseudo-bootstrapped Delta-based cluster-analysis, as above), the significance of the choice between the methods in a *literary* (as opposed to a *methodological*) study like the present one is probably nil. In my honest opinion it is more important to use stable methodology, even if this means losing a percent of attribution success rate here and there, rather than to multiply unknown variables by testing ever-new algorithms while trying to solve a *literary* (or linguistic) question.

Results

The first thing to check was whether the gender signal is at all visible *the easy way*, i.e. in simple most-frequent-word analysis. Some hope in that respect can be garnered from Pennebaker, but it should be borne in mind that the list of MFWs for any collection of texts is rarely identical to one made up *a priori* of function words alone. The cluster-analysis consensus tree for the "famous men" and the "famous women" sets is presented in Fig. 1, and it is not very helpful: it does some good authorship attribution, and it only goes wrong when the authors go Gothic (see Note 1). Adding the Chawton House novels (including the anonymous ones) to the corpus does nothing, too. This is exactly what a consensus tree is supposed to do: to eliminate all signals except the strongest one, and the strongest signal is usually that of the author.

Figure 1. Consensus tree for texts by "famous men" (red; prefixed M) and "famous women" (green; prefixed W), obtained for 100-1000 most frequent words.

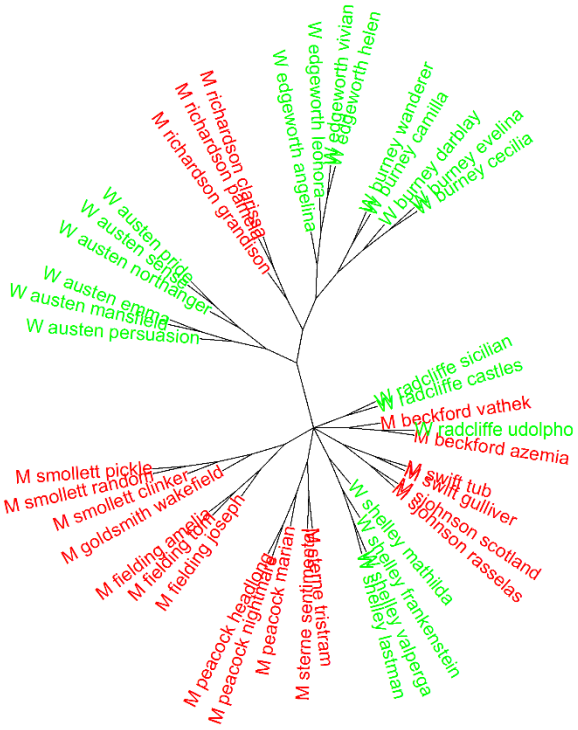
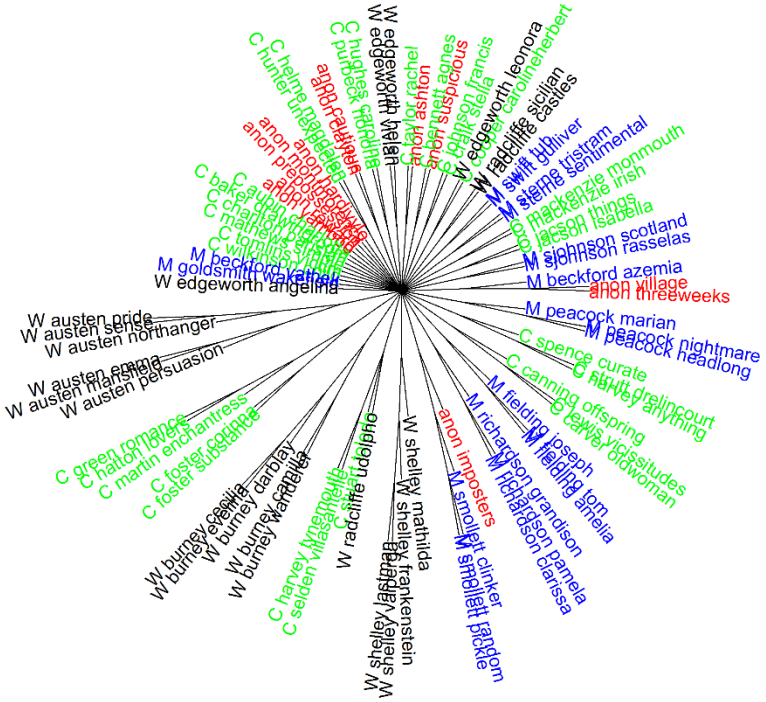
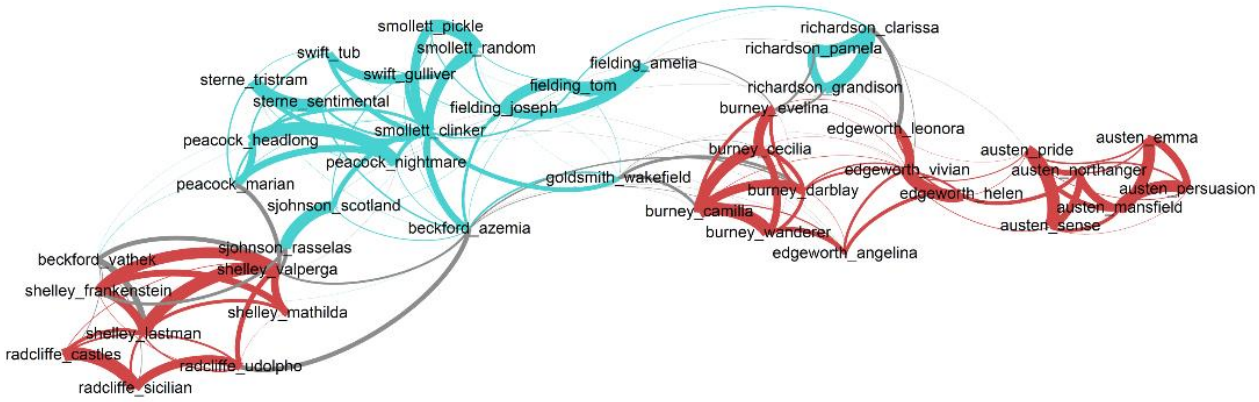


Figure 2. Consensus tree for texts by "famous men" (blue; prefixed M), "famous women" (black; prefixed W), Chawton House novels by known (prefixed C, green) and unknown (red; prefixed anon) authors, obtained for 100-1000 most frequent words.



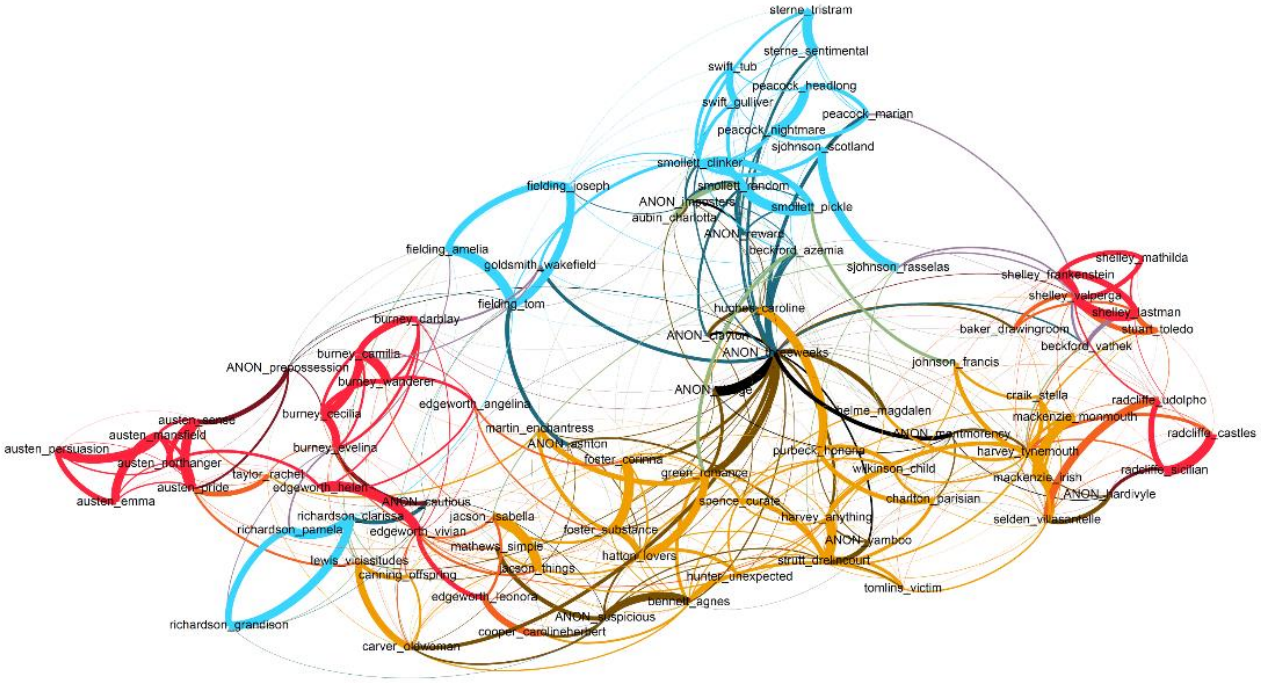
Things become somewhat more interesting when the consensus cluster analysis output is used to generate a network visualization in Gephi. Since the latter brings out more than just one signal, more than just the strongest clusterings, the power of authorial connections is now mitigated by other signals, and, in Fig. 3, two more factors can be discerned: that of genre/theme, which makes Shelley accompany Radcliffe and *Vathek* (right), and that of gender, which brings the men together in one node in the middle, and the remaining women in another (left). The one misgendered author, Richardson, can be partially excused: his *Pamela* and *Clarissa* might be in fact instances of successful gender language stylization prompted by the epistolary form; the dubious presence, in the closest vicinity, of *Grandison*, might only suggest that it is dominated at the onset by Harriet Byron rather than by the later fortunes of the eponymous hero – or, perhaps more reasonably, by the fact that while *Pamela* and *Clarissa* have been attracted to the female circle by gender-associated similarities of vocabulary, the authorial signal in the third Richardson novel drew it, in turn, into the area.

Figure 3. Network analysis graph for texts by "famous men" (blue) and "famous women" (red), obtained for 100-1000 most frequent words.



The addition of the Chawton House novels to the reference sets of "famous men" and "famous women" seems to provide quite promising results in terms of gender identification of the 12 anonymous texts (Fig. 4). The position of the males does not change; Richardson is still an outlier; but now some of the anonymous texts take interesting positions, and this is especially true of *The Imposters detected: or, the Life of a Portuguese, in which The Artifices and Intrigues of Romish Priests are humourously displayed* (1760), and a very Richardsonian work, *The Reward of Virtue; or, the History of Miss Polly Graham* (1769). If, then, the evidence of the most frequent words is to be trusted, these would be the first two suspects for male authorship.

Figure 4. Network analysis graph for texts by "famous men" (blue), "famous women" (red), Chawton House novels by known (orange) and unknown (black; prefixed ANON) authors, obtained for 100-1000 most frequent words.



This evidence would be even more trustworthy if there existed a more theoretical model of the male/female difference in lexical choice. The one that immediately comes to mind is the list (or, more precisely, the categories) already quoted from Pennabaker above. The categories, then, were converted into a list of some 300 words matching Pennebaker's specifications, but no visible separation between men and women were observed. Still, this was perhaps too much to hope for: the corpus in this study was that of 18th- and early-19th-century texts; Pennebaker's – a much more general one. It has already been mentioned that Jockers's analyses feature some of the authors in my reference corpus, so another attempt was made with the words from his list of "Features best distinguishing male and female authors" (94). There was some overlap between this list and that of Pennebaker's, but also with that of the most frequent words as produced by my study, but no improvement in terms of gender distinction.

The search for the optimal wordlist obviously leads to another Burrowsian method, Zeta (2006), which divides texts into equal-sized samples and then looks for words that appear consistently within one text or group of texts and consistently do *not* appear in another. This produces a set of medium-frequency-range words (roughly speaking, similar to keywords obtained with the classical log-likelihood approach, minus function words); among the chief attractions of this approach is the fact that such words are much more "meaningful" than the high frequency function words and tend to make sense from a traditionally-literary point of view. Thus, in order to generate a male/female distinction, the "famous men" and "famous women" texts were used as the two groups to find their

consistently-preferred words using the "oppose" function in *stylo*. The resulting 600 words or so were then used on the combined Chawton House and "canonical" writers corpus, with the latter serving as benchmark to identify the optimal number of Zeta words used in the analysis. Then, only those cluster analysis graphs were accepted as significant that correctly divided the reference set of "famous" men and "famous women" in an act of honest and well-founded cherry-picking.

Fig. 5 presents such a graph for 284 medium frequency "male" and "female" words; results for longer wordlist were very similar up to 490. First of all, the texts that served to produce the wordlist have been divided by gender with near-perfection. There is just one exception: Beckford's *Azemia* remains on the female part of the cluster analysis tree. As has already been noted, it is a parody of "female" writing of the time. Although Beckford makes his opinions and his objectives clear in the novel's subtitle, *Imitations of the Manner, both in Prose and Verse, of Many of the Authors of the Present Day*, he also compounds the illusion by inventing a feminine author persona, "Jacquette Agenta Mariana Jenks," and by beginning his parodistic harassment with the dedication:

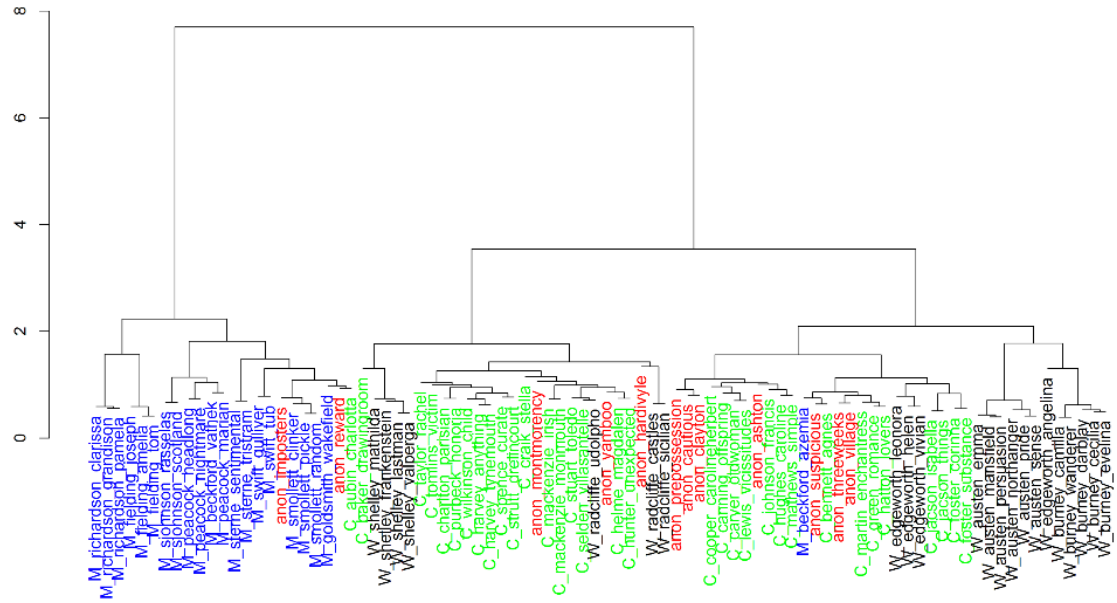
> Despairing as I do to give to these pages the acumen, brilliance, consistency, delicacy, elevation, fancy, genius, humour, judgment, illumination, knowledge, luxuriance, merriment, naiveté, omniscience, pathos, quickness, raillery, suavity, tenderness, urbanity, vivacity, wit, 'xcellence, youthfulness, and zest, which corruscate from the etincellant pen of your Ladyship, I yet venture to flatter myself that this debut in literature, which I have thus the honour to place under your protective kindness, may, rather owing to your smiling approbation (dear to literary spirits), than to any individual merit, serve to arouse, not unacceptably, the elegant leisure of the amiable fair, in that superior region of the British atmosphere where your Ladyship sparkles a benignant and irradiating planet.

While the gender of none of the "famous women" is misattributed, a single Chawton House corpus text with known authorship is also listed as "male": Penelope Aubin's *The Life of Charlotta Du Pont, an English Lady; Taken from her own MEMOIRS* (1723). The author undoubtedly existed; had a husband and three children; and produced a number of novels and translations; and thus she cannot be simply dismissed as another spurious Jenks. I should add, with not too much conviction, that her genderial misattribution might be blamed on her highly adventurous content (including some Madagascar Pirates who seem to have operated in the wrong Ocean), parts of which were derived from the author's brother's experience in the colonies:

> Giving an Account how she was trepan'd by her Stepmother to Virginia, how the Ship was taken by some Madagascar Pirates, and retaken by a Spanish Man of War. Of her Marriage in the Spanish West-Indies, and Adventures whilst she resided there, with her return to England. And the History of several Gentlemen and Ladys whom she met withal in her Travels; some of whom had been Slaves in Barbary,

and others cast on Shore by Shipwreck on the barbarous Coasts up the great River Oroonoko: with their Escape thence, and safe Return to France and Spain.[3]

Figure 5. Cluster analysis tree for texts by "famous men" (blue; prefixed M), "famous women" (black; prefixed W), Chawton House novels by known (green; prefixed C) and unknown (red; prefixed anon) authors, obtained for 284 "famous men"/"famous women" Zeta keywords.



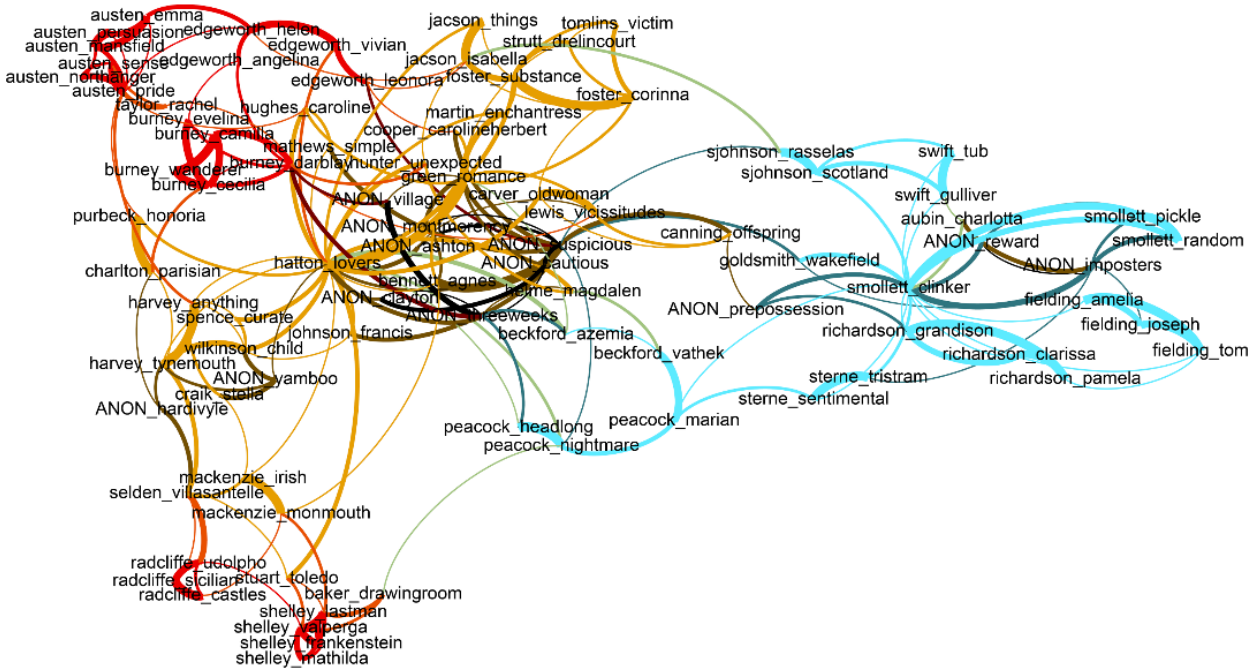
More *à propos* of this study, two anonymous Chawton House works also – and consistently at that – refuse to cluster with the women writers, and they have already done so when analysed on function words: *The Imposters detected* and *The Reward of Virtue*. In a network graph performed on the same data (Fig. 6), this effect persists over a background of good separation between the men (left) and the women (right). Another Chawton House anonymous novel approaches the male nodes: *Prepossession*, and it is anyone's guess whether this is due to the skill of the unknown female author who penned this book, subtitled *Memoirs of Count Touloussin. Written by Himself*, or whether, much less probably, there indeed was a person who called himself Count Touloussin. It is also worth noting that the other anonymous texts cluster elegantly with the rest of the Chawton collection.

Even more interesting is the composition of the wordlist that has produced the above results. Many of the words it contains make sense in terms of Pennebaker's and Jockers's findings. Its female part

[3] There is another authorial skeleton in this closet, although of little help in this particular case: almost a century later, in 1770, someone (apparently a bookseller) changed a few names in Aubin's work (she had already been dead for several decades) and published it anonymously as *The Inhuman Stepmother; or the History of Miss Harriot Montague* (Kulik 2000).

is especially striking, as it contains – apart from a much greater proportion of verbs – tokens that seem to be in direct relation to the subject-matter and the general mood of these novels, and they seem to confirm the most stereotypical views: *feelings, felt, idea, feel, exclaimed, anxious, feeling, party, oh, alone, painful, carriage, anxiety, attention, society, beautiful, surprise, object, occupied, voice, quitted, suddenly, remain, appeared, manners, kindness, reached, regret, existence, smile, hastily, spoke, attachment, listened, deeply, yes, agitated, followed, excited, seated, interesting, happiness, emotion, wished, agitation, amiable, wishes, tone, admiration, instantly, completely, affection, immediately, listen, situation, moment, recollection, silence, intelligence, waiting, occurred, silent, attentions, chance, meant, evidently, events, announced, smallest, eager, sought, aware, perfectly, ah, moments, wholly, feared, ma, ceased, sunk, elegant, form, change, forgotten, vain, sorrow, sensations, interrupted, struck, hurried, agony, minutes, alarm, experienced, extreme, lovely, attached, interested, deep, residence, surprised, passing, hastened, propriety, explanation, evening, fortitude, eagerly, apparent, emotions, quick, influence, solicitude, cried, countenance, render, sensibility, animated, fixed, drew, sorrows, expression, exertion, misery, glance, plans, unable, lost, fearful, circumstance, seek, joined, disappointment, distant, calm, accompanied, scene, melancholy*.

Figure 6. Network analysis graph for texts by "famous men" (blue), "famous women" (red), Chawton House novels by known (orange) and unknown (black; prefixed ANON) authors, obtained for "famous men"/"famous women" Zeta keywords.



The men in this experiment use words perhaps less directly associated with the content of their stories, but they like virtues such as *help, honest, honour, favour, forgive, deserve, merit, order, reputation, quality*; they are courteous to other men: *gentlemen, squire, fellow*; they probably have many things to say about the other *sex*: *pretty, clothes, reputation*; they favour archaisation: *hath, thou, thee, thy, dost, hast, tis, doth, wilt, methinks, nay* (note the presence of *yes* among the "female" words above); and they are partial to contractions: they like to address their *reader*; and, although they do invoke *devil* and *god* on occasion, nevertheless they have their characters swear without being too explicit about it: *cursed, swore*. They speak of *body* and body parts: *mouth, nose*; and they definitely care about money and numbers: *piece, expense, expensive, six, twenty, three*.

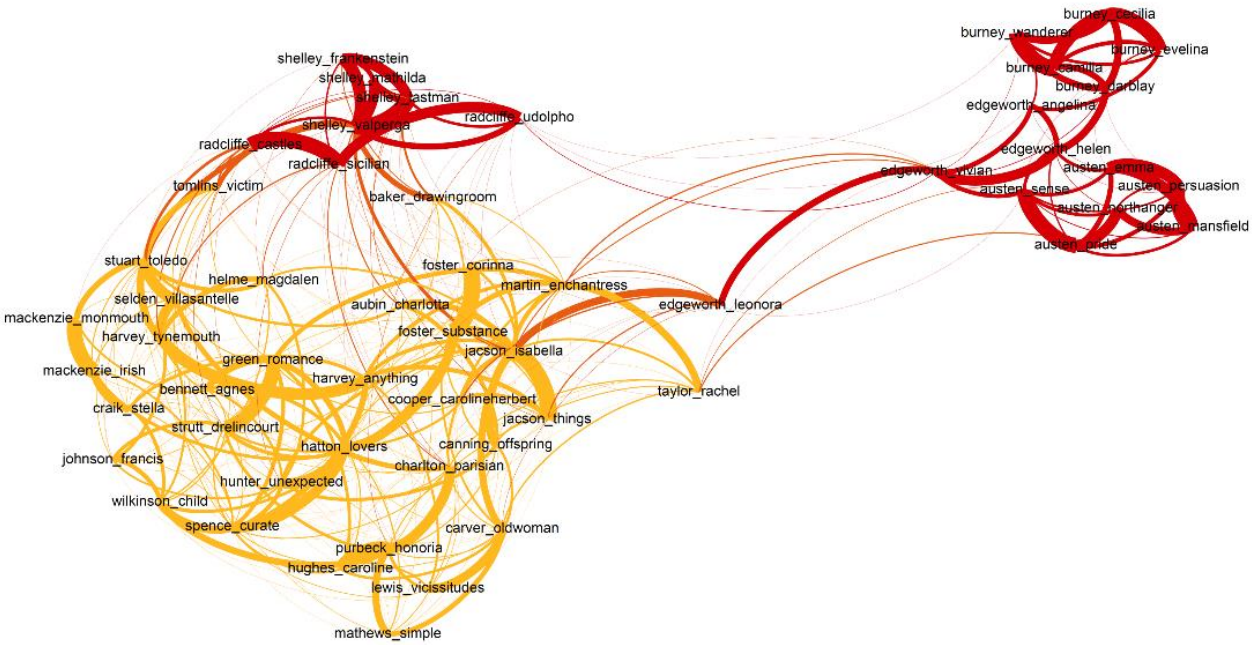
But there is more. This figure also shows that there is some division of the left-hand side of the network in Fig. 6. The centre of that portion of the graph contains the Chawton House novels, anonymous and non-anonymous; their famous rivals recede into the peripheries, with a separate cluster for the Gothic genre (bottom left). There is a way to examine this phenomenon more precisely: by assessing the difference between the "canonized" women in relation to the men and the "uncanonised" Chawton House authors. And while a Delta-distance network analysis of these two groups based on their respective Zeta words might seem something of a tautology (not entirely, perhaps, since Delta and Zeta are two independent methods dealing with very different lists of words, Fig. 7), the crucial difference here is best understood by comparing the lists of Zeta words characteristic for the two groups. The Chawton House authors exhibit a decided preference for words that would be most readily and stereotypically associated with sentimentalist feminine fiction; these words include (in the order of significance by Zeta score): *bosom, lovely, worthy, respecting, husband, beheld, requested, protection, child, female, sentiments, beloved, cheek, parent, virtue, madam, unfortunate, wishes, thy, charms, tender, fortune, fatal, affection, conduct, daughter, virtuous, retired, acquainted, providence, tears, inform, rectitude, supposed, arms, lover, conceal, wife, death, friendship, whilst, principles, thou, amiable, parents, innocent, replied, cause, elegant, form, behold, health, introduced, request, calculated, fate, woman, breast, expressed, greatly, soul, girl, deprived, discovered, possessed, father's, heaven, sex, infant, lay, peace, intention, religion, heroine, christian, description, attachment, receive, attended, laid, terms, england, residence, particular, inclined, passion, thee, sacred, accompanied, pursue, polite, tear, gratify, attend, entertained, innocence, chamber, guilty, humble, angel, servant, departure, grateful, endeavour, possession, prevented, handsome, faithful, act, contents, attached, agitated, generally, mercy, hearts, features, son, although, period, threw, task, pale, ardent, convinced, intended, unhappy, father, sigh, oh, vile, tenderness, pious, fashionable, sensible, charming, honest, domestic, lips, errors, consequences, accompany, proper, events, enjoy, gained, safety, story, affectionate, valuable, birth, husband's, chose, flattered, adopted, appointed, interested, evinced, remain,*

stranger, attendant, indulge, guilt, hours, paid, injured, possess, different, beautiful, extended, until, seat, necessary, obligations, attentions, instance, fell, consequence, benevolent, hermitage, heir, companion, vice, actions, almighty, proof, accustomed, embrace, resumed, duty, truly, anguish, weak, art, readers, dispatched, example, assistance, acted, permitted, reward, bed, reasons, marriage, contained, presented, reach, probable, countess, remarked, ear, wretch, supported, lordship, arrival, lamented, visible, requesting, widow, captain, reflections, previous, language, friendly, experience, entertain, treated, bless, mentally, lady's, acknowledge, particularly, loss, gratification, procure, concealed, sum, prove, deprive, arm, indebted, presence, indulgence, pair, adieu, senses, derived, pleasing, beauty, guests, gain, artful, victim, interesting, sufficiently, wretched, rank, fascinating, inclination, case, witness, pensive, convince, composed, fair, relative, dress, dictates, frame, reverend, furnished, age, event, support, confess, prudence, virtues, holy, union, human, confined, amongst, major, under, veil, society, amusements, experienced, mrs, offered, married, alas, resided, arguments, beings, add, sorrows, offspring, conversation, information, discovery, fortunate, free. This long list can be easily and meaningfully divided into several semantic categories of great import in sentimentalist fiction: sentimentally-tinged parts of the anatomy and actions performed therewith; family and social terms; abstract terms; exclamations and terms of praise; uncertainty and negative emotions. Particularly striking is the way simple statistics like Zeta highlight the five key terms of sentimentalist fiction: *bosom, lovely, worthy, respecting, husband*.

In comparison, the preferred words of the canonical women writers seem much more mundane (again, the order is by decreasing Zeta score): *hardly, anything, else, sort, nobody, cried, looking, talk, presently, everything, coming, beginning, it's, talked, among, forced, talking, forward, walked, everybody, won't, went, courage, suddenly, o, afterwards, ago, away, frightened, don't, glad, difficulty, sorry, that's, word, quiet, merely, meanwhile, laughing, looked, listen, concerning, can't, whether, get, tis, meant, came, quite, anybody, eager, speak, tried, things, together, begged, quick, whatever, wholly, speech, waiting, got, makes, because, ashamed, stairs, general, touched, half, begun, mere, afraid, spoke, gone, imagine, quietly, moved, beg, go, play, therefore, hard, forth, worse, utterly, seem, there's, worth, answered, species, come, angry, comes, new, work, eagerness, surprised, standing, run, yes, immediate, ask, begin, honour, something, risk, especially, speaking, back, began, amazed, hurried, struck, air, feeling, hurry, people, instantly, hear, while, help, understand, revived, listened, rest, changed, seemed, he's, ran, stay, aloud, serious, sit, spent, wonder, i'm, shame, provoked, news, sure, to-day, low, spoken, followed, enjoyment, around, clear, greater, assure, eagerly, midst, less, hate, sight, shocked, counsel, watching, seems, kindness, perceived, laughed, voice, party, except, you'll, open, i've, openness, simply, ay, right, further, notion, try, seized, she's, going, none, stayed, change, whither, minute, deal, walking, odd, however,*

running, startled, wait, somebody, pause, wrong, arranged, desire, myself, joined, window, mingled, desired, one's, fresh, calling, done, grew, shut, terrible, doing, wanted, abruptly, somewhat, conference, encouragement, consciousness, involuntarily, praise, does, attack, full. In fact, they either seem like parts of a simple most-frequent-word list or share some categories (contractions, positive emotions and features) with the "male" wordlists of previous experiments. And this speaks volumes of the possible mechanism of canonisation: a female writer enters the generally-accepted canon if she writes more like a man.

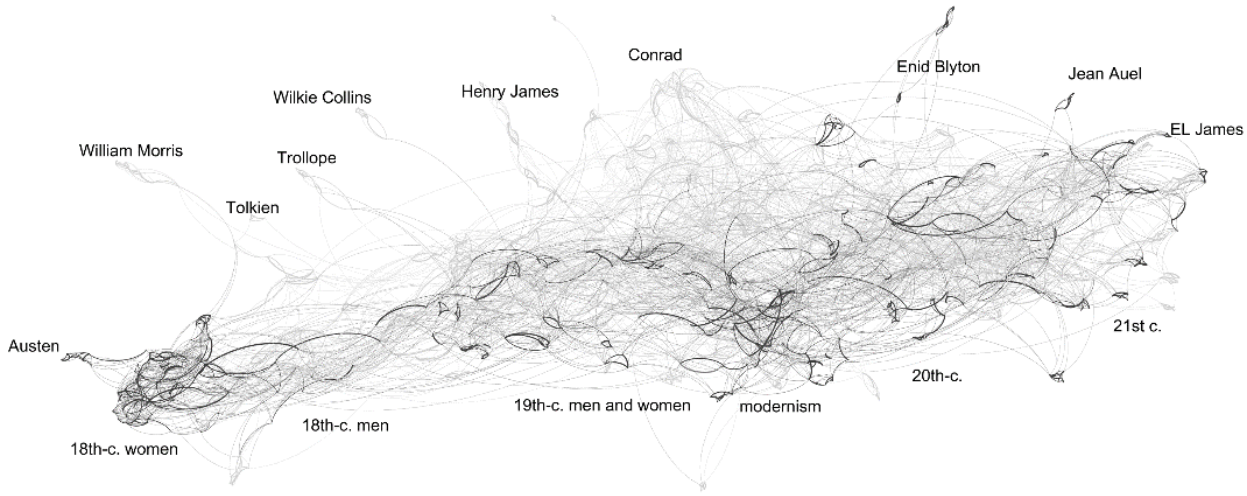
Figure 7. Network analysis graph for texts by "famous women" (red) and Chawton House novels by known authors (orange), obtained for their respective Zeta keywords.



Since gender divisions seem to become clearer and clearer in the material studied so far – material derived from the crucial time of the emergence of the novel throughout the 18th century – it is tempting to take this discussion one step further to see whether these (or other) "male" and "female" words continue to function further into the 19th century; whether they survive the 20th, and whether they are at all visible in the 21st. Despite Pennebaker, there is some suggestion that they would not, since the rigidly demarcated gender roles of the 18th century have been increasingly transgressed, and since much of the transgression was owed to literature (although literature's enthusiastic involvement in the demarcation cannot be ignored as well). The problem is compounded by a phenomenon that has been observed in this study – by the fact that *a priori*-made gender-sensitive wordlists tend to fail when they have not been derived from the material at hand. It is somewhat problematic, then, whether using a wordlist compiled from 18th-century novels would

have any significance for the gender divide in later times in any quantitative study, and worse: whether it is still a valid feature for gender identification at all for historical and social reasons. To dispel or confirm these fears, I conducted a series of network analyses on a corpus of 1000 novels: a total of a little more than 111 million word-tokens in 635 works by men and 365 by women. The corpus included all of the above-studied texts to represent the 18th century; a good representation of the 19th-century novel, and an even greater proportion of books from the 20th and the 21st centuries. The first network graph (Fig. 8) was to check if, in such an extensive corpus, most-frequent words alone can provide any gender identification. The elongated quasi-one-dimensional diagram is dominated by the chronological signal on the macro scale (the texts are aligned from the earliest to the most recent, or, in this case, left to right), and the authorial signal (the texts group by author) on the micro scale. The former is quite strong indeed: the first group from the left comprises Jane Austen and then other 18th-century women writers (both canonical and not), with some overlap with the adjacent group of the male writers of the same century. But then the gender signal is no longer discernible in the later centuries, as the male light grey and the female dark grey travel along what can be construed as the horizontal time axis. There is a slightly denser concentration of female texts that corresponds to the modernism of Woolf, Hall, West and Mansfield, but there are plenty male modernists around them as well. Some interesting outliers have also been marked in the figure, and a spirited commentator could venture that what began, in female writing in English, with Jane Austen, now ends with E.L. James…

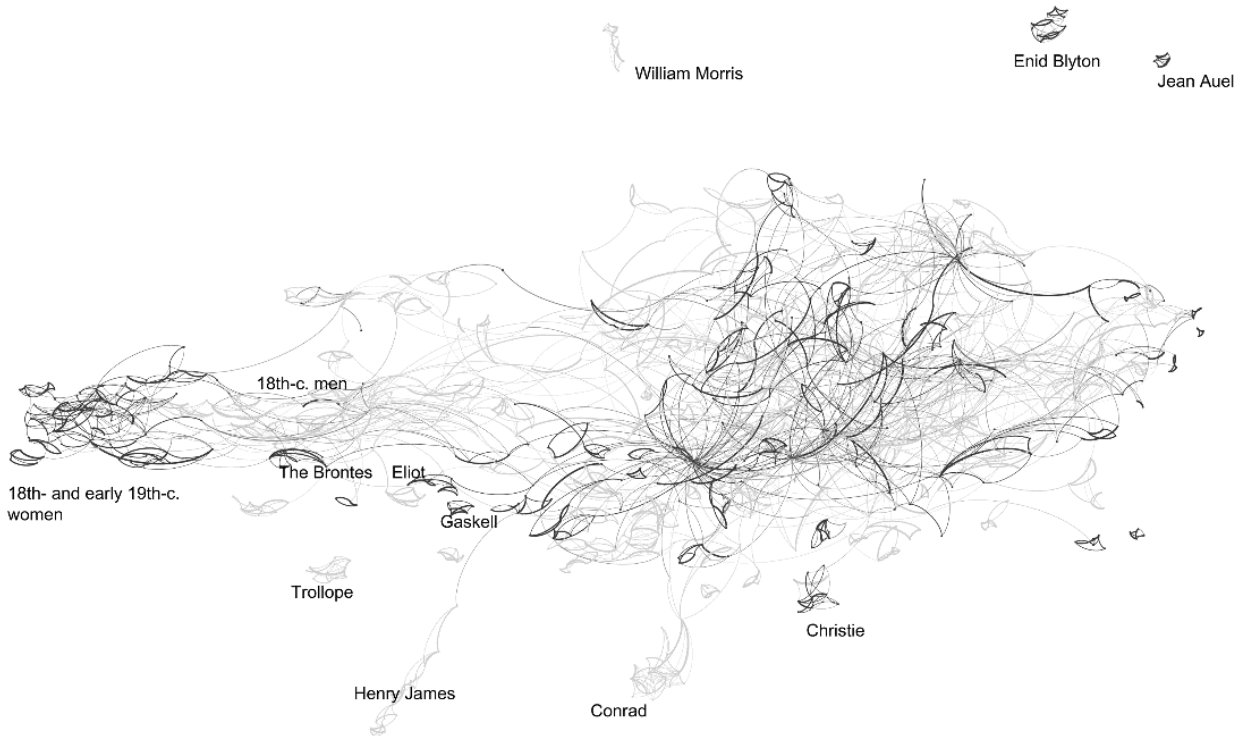
Figure 8. Network analysis graph for 1000 novels by men (light grey) and women (dark grey), obtained for 100-1000 most frequent words.



To gain some insight into the validity of 18th-century gender-signal keywords, they need to be checked against the same 1000-novel corpus (Fig. 9). Ominously, nothing much happens, and the plot seems only to gain in vertical size. Closer inspection reveals a slightly better removal of the

18th-century male writers from Austen, Burney and the Chawton House novels and, perhaps, a more solitary evolutionary line: the Brontës-George Eliot-Gaskell; but then the dark grey links between female writers become submerged in the lighter-grey sea of male writing, and the male and female evolutions move together, steadily, into the 21st century. This suggests that, if anything, the 18th-century gender marker words have become obsolete; that they are now less useful than mere most frequent words for the gender signal in the entire corpus. Also, many of the outliers marked in the previous plot return in the present one, and this is another point of interest: quite clearly, it is quite immaterial which words are taken for this sort of analysis, as long as they are fairly frequent and there is enough of them – in vindication of some of the findings of Rybicki and Eder (2011).

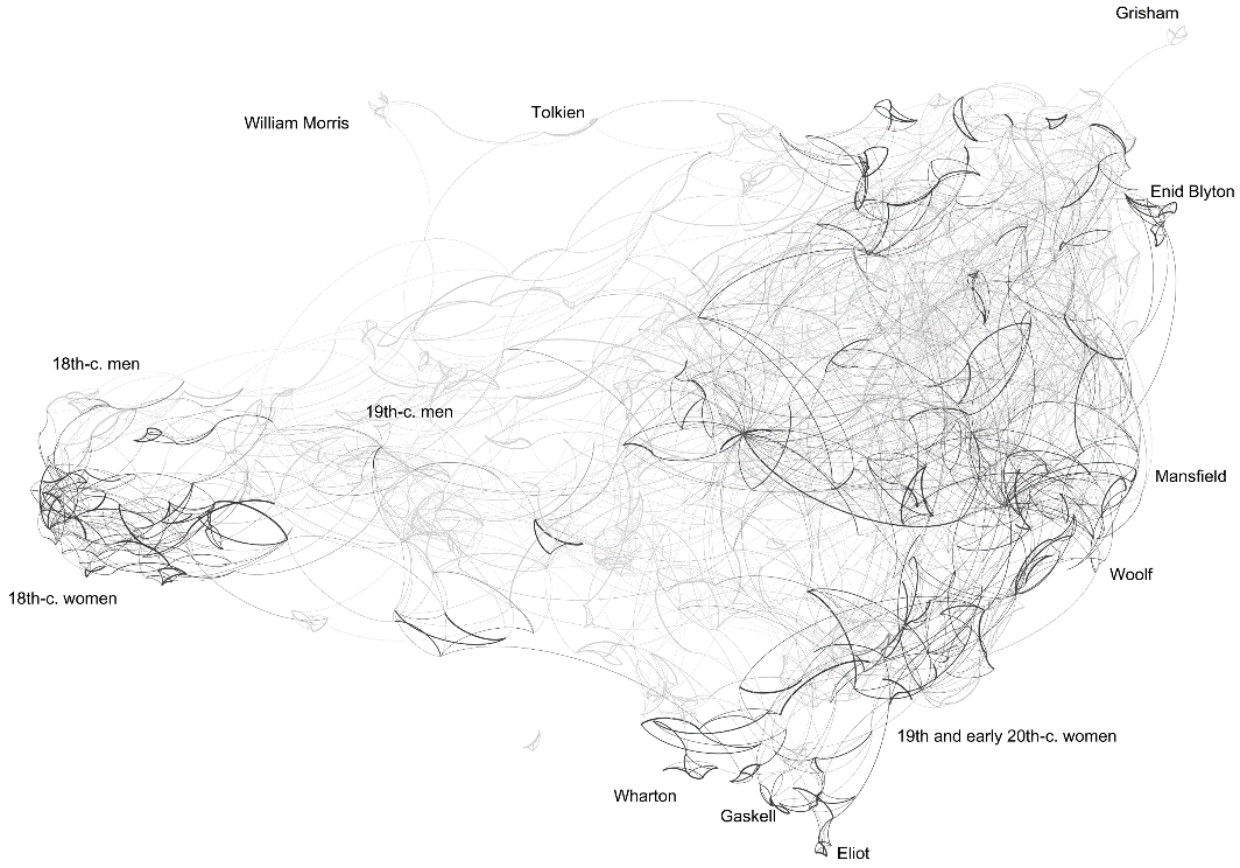
Figure 9. Network analysis graph for 1000 novels by men (light grey) and women (dark grey), obtained for 18th-c. "famous men"/"famous women" Zeta keywords.



To come up with a somewhat fresher gender-sensitive wordlist, a 100-text subcorpus was used that included novels written by men (67) and women (33) between 1839 and 1939. Zeta words for both genders were obtained. Those for women included a much more complex combination in comparison to the favourite words of their predecessors – sensory and cognitive terms: *listening, watched, watching, wondered, paused, absorbed, consciousness, conscious, expression;* both negative and positive states and emotions: *feelings, agreeable, tender, difficult, wide, quietly, suffering, shut, passionate, dared, liked, sunshine, loving, lifted, busy, roused, shone, bent, effort,*

pale, apt; reading: *reading, books;* "feminine" and/or household objects: *silk, flowers, roses, curls, chair;* colours: *crimson, brown.* By contrast, male vocabulary, if anything, became more stereotypical; it now came to reflect the political turmoil of the era and a strong domination of nouns: *enemy, honour, pardon, battle, captain, officer, sword, shoot, shot, fight, army, officers, killed, armed, smoke, wrath;* there are numbers and/or money: *thousand, dozen, paid, fifty*; suddenly, lower extremities seem to dominate the human body: *legs, heels*; finally, women are referred to as *female*; there is swearing, more outspoken this time: *swear, swore, oath, devil, deuce*; and drinking: *bottle, drunk*. The resulting network (Fig. 10) departs even more than the previous one from the linear order of the one based on most frequent words and finally disturbs the hegemony of the chronological order somewhat by placing 19th- and early 20th-century women writers in the periphery of the graph, bottom centre; significantly, this set of nodes represents other texts than just those used to create the wordlist in the first place.

Figure 10. Network analysis graph for 1000 novels by men (light grey) and women (dark grey), obtained for 1839-1939 male/female Zeta keywords.



There is yet another element of the difference between the male and female linguistic universes in literature, but it only becomes evident in more inflected languages than English. A similar experiment on a similar 100-text corpus of Polish novels, while yielding similar semantic categories for the

genders, produced another class of words: past-tense verb forms with masculine and feminine suffixes. Indeed, the only verbs that appeared in lists of some 300 Zeta words obtained for each gender from the mid-19th- to early 20th-century Polish texts were those with the respective genderial suffixes. The presence of such a phenomenon is not a surprise in itself, for the dominance of the male perspective in books by male writers and of the female perspective in those by women is not only a reasonable expectation but also a well-established literary fact. Heroines in 19th-century historical romances speak very little when compared to heroes; Jane Austen never has two men in a scene without a woman present. What *is* surprising is the scale of the phenomenon: the other gender has very little to do, or say.

Conclusions

It might seem that this study has met its primary objectives. Quite consistently, two different methods of gender identification by word frequencies have pointed out two potential suspects in the search for a possible male among the anonymous authors in the Chawton House corpus of 18th- and early 19th-century English women's writing. In particular, *The Imposters detected: or, the Life of a Portuguese* seems a good catch, as it departs from the sentimentalism that dominates the collection – into part picaresque, part anti-Catholic satire. If, indeed, *The Imposters* are the work of a male imposter, he might have made a small but potentially revealing blunder, not so much in failing to simulate the feminine way of using functions words or of selecting keywords as in this very male sentence from his preface to the story: "None but *womanish* and weak minds will take offence at the stories related in this little work." While other texts in the 1000-strong corpus use the offending word, only *The Imposters* are so rude in direct authorial address to the reader. The combination of the evidence of multivariate analysis, or distant reading, and of the evidence of a single word in the preface, or close reading – make the anonymous author a very likely suspect. There are more suspect things about this book, by the way. Does not the editor pretend that the story is a translation, from the French, of a manuscript found in Padua? Not that it is rarity in those times: half of the English Gothic novels pretend, or at least once pretended, to be translations. The French themselves take no responsibility: their *Annales typographiques ou notice du progrès des connoissances humaines* of 1760, vol. 2, published in Paris, list the original English title, provide the French translation, and very candidly comment on the quality of the *ouvrage*: "*De toutes les brochures auxquelles les dernieres affaires du Portugal ont donné lieu, il n'y en a pas eu de plus mauvaise que celle dont on vient de lire le titre*."[4]

[4] "Of all the pamphlets that deal with the recent affairs of Portugal, there are none that would be as bad as the one whose title we have just provided" (translation mine).

The other recurring suspect, *The Reward of Virtue; or, the History of Miss Polly Graham*, is somewhat more mysterious, for it provides no telling preface and no other clues. At least its philosophy is somewhat more constructive, for the final chapter of the work describes a useful institution, an early version of a Dickensian workhouse: Bounty Hall is a place where "a company of ladies, having considered the great inconveniencies which many virtuous women of family labour under from being reduced, through unavoidable misfortunes, to poverty, took the generous resolution of providing an asylum for those unhappy persons." Sadly, this noble piece of writing is dismissed by an anonymous critic of *The Monthly Review; or, Literary Journal* (1769) as "a jumble of improbable and ill-connected tales."

The second direct question of this paper – the difference between the Chawton House women writers and their more fortunate rivals such as Austen or Shelley – can be answered in a way that makes sense in the context of the canon wars that have been erupting at least for the last half-century. It is interesting how the view that seems to emerge from this stylometric analysis – that women become part of the canon if/when they write a little more like men - might join the fray between Harold Bloom's defence of the Western canon and what he calls, rudely, "Schools of Resentment." Yet perhaps traditional literary scholarship has never voiced it as clearly as quantitative research has that the notions of "writing like a man" and "writing like a woman" are in such a constant flux that they risk becoming, in fact, quite problematic. This is what seems to be suggested by the inability, in this study, to produce a stable "canon" of male and female keywords that would survive a change of corpus or shifts in literary evolution, or both, either basing on statistical analysis or on *a priori*-defined lists and categories.

At the same time, the individual keywords make sense in terms of traditional literary history and can be used to vindicate the traditionally suspect bag-of-words model: it seems it is something that can be trusted as long as the statistics is robust and the methods are sound. Another thing that makes sense is the very change, over time, in the keywords – as evidenced by the shift in the gender-sensitive Zeta words from the vocabulary of sentimentality that made the Chawton House corpus such a consistent collection of texts, to the much less one-sided collection of words that define women's writing a century later. The evolution from once-sided to complex can be turned on its head, too, as the changes in male keywords that happened between the publication of *Oliver Twist*, the earliest book in the 100-novel corpus (1839), and the present boundary of the public-domain (1939), were those from a variety of semantic fields to a vocabulary undoubtedly and specifically dominated by the wars of the time: the Boer, the Great, and the impending Second.[5]

[5] This must be compared with the wonderful immunity to the outside world apparent in the novels of Jane Austen (as opposed to her correspondence), whose keywords seem quite innocent of the Napoleonic wars.

One thing that is quite clear from this exercise in gendered language is that there are greater powers than this that influence the evolution of literary lexicon. Chief among them is time, the hero of another of John Burrows's classical papers, "Tiptoeing into the Infinite" (1996): the chronological signal even appears unwanted when the focus of an analysis is on gender (and, in fact, on anything else). Jockers estimates the influence of his "decade" category at 14% (98); I wonder if time, perhaps treated more generally, is not much, much more important. There is, it must be said, an interesting dualism about the chronological signal in literary language, since it concerns single-lifespan and single-author collections of texts as well as large and long-span multi-author corpora, and both phenomena cannot be both blamed on the same mechanism of linguistic change. But that is an entirely different story – to translate back into the original the Bowdlerized Polish translation of Kipling's final line of the first Jungle Book.

And yet there are pieces of information in her work for which French Intelligence might have paid good money: while the removal of Wickham's regiment from Longbourne is important chiefly for the evolution of several plot lines within *Pride and Prejudice*, the fact that his unit is moved to Brighton – on the English Channel – suggests a possibility of an invasion. Austen has long been entangled in colonialism (c.f. Said 1993, 80-96), why not the Napoleonic wars as well?

References

Argamon, S., Koppel, M., Fine, J., Shimoni A.R. (2003). Gender, genre, and writing style in formal written texts. *Text* 23 (3): 321-346.

Backus, T.J. (1874, 1891). *Shaw's New History of English Literature.* New York and Chicago: Sheldon & Co.

Bastian M., Heymann S., Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Burrows, J. (1987). *Computation into criticism: a study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon.

Burrows, J. (1996). Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative. In Susan Hockey and Nancy Ide (Eds.), *Research in Humanities Computing* 4. Oxford: Clarendon, 1-33.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17: 267-287.

Burrows, J. (2005). Who wrote Shamela? Verifying the Authorship of a Parodic Text. *Literary and Linguistic Computing*, 20 (4): 437-450.

Burrows, J. (2006). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing* 22: 27-47.

Craig, H. and Kinney, A. F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Dahllöf, M. Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability. *Literary and Linguistic Computing* 27 (2): 139-153.

Eder, M. (2013). Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*, first published online November 14, doi:10.1093/llc/fqt066.

Eder, M. (2014). Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii. *Teksty Drugie* 2014/2: 90-105.

Eder, M., Kestemont, M., and Rybicki, J. (2013). Stylometry with R: a suite of tools. *Digital Humanities 2013. Conference Abstracts*, University of Nebraska-Lincoln, 487–89.

Eder, M., Kestemont, M. and Rybicki, J. (2014). *Computational Stylistics*. <https://sites.google.com/site/computationalstylistics/>

Giroud, F. and Lévy B.-H. (1994). *Kobiety i mężczyźni*, Warszawa: Puls.

Jacomy, M., Venturini, T., Heymann, S. and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* 9(6): e98679. doi:10.1371/journal.pone.0098679

Iyeiri, Y., Yaguchi, M. and Baba, Y. (2011). Principal component analysis of turn-initial words in spoken interactions. *Literary and Linguistic Computing* 26 (2): 139-152

Koppel, M., Argamon, A. and Shimoni A.R. (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing* 17 (4): 401-412.

Kulik, Maggie. What the Bookseller Did: A Case of Eighteenth-Century Plagiarism. *Female Spectator* 4, No. 4 (2000): 9-10.

Lakoff, R. (1973). Language and Woman's Place, *Language in Society*, Vol. 2 (1): 45-80.

Mikros, G.K. (2013). Systematic stylometric differences in men and women authors: a corpus-based study. In R. Köhler and G. Altmann (Eds.), *Issues in Quantitative Linguistics 3. Dedicated to Karl-Heinz Best on the occasion of his 70th birthday*. Lüdenscheid: RAM-Verlag, 206-223.

Mikros, G.K. (2013a). Authorship Attribution and Gender Identification in Greek Blogs. In I. Obradović, E. Kelih and R. Köhler (Eds.), *Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO)* in Belgrade, Serbia, April 16-19, 2012. Belgrade: Academic Mind.

Mikros, G.K. and Perifanos, K. (2013b). Authorship attribution in Greek tweets using multilevel author's n-gram profiles. In E. Hovy, V. Markman, C. H. Martell and D. Uthus (Eds.), *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext."* 25-27 March 2013, Stanford. Palo Alto: AAAI Press, 17-23.

Nerbonne, J. (2014). The Secret Life of Pronouns. What Our Words Say About Us. James Pennebaker (review). *Literary and Linguistic Computing* 29 (1): 149-141.

Olsen, M. (2005). *Écriture féminine*: Searching for an Indefinable Practice? *Literary and Linguistic Computing* 20 (Suppl): 147-164.

Pennebaker, J. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. New York, Bloomsbury Press.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wien, http://www.R-project.org/.

Rybicki, J. (2006). Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. *Literary and Linguistic Computing* 21: 91-103.

Rybicki, J. (2007). Twelve Hamlets: a stylometric analysis of major character's idiolects in three English versions and nine translations. *Digital Humanities 2007: Conference Abstracts*. University of Illinois, Urbana-Champaign, 191-92.

Rybicki, J. (2008). Does Size Matter? A Reexamination of a Time-proven Method. *Digital Humanities 2008*, 184.

Rybicki, J. (2014). Pierwszy rzut oka na stylometryczną mapę literatury polskiej. *Teksty Drugie* 2014(2): 106-128.

Rybicki, J. (2014a). Visualizing Literature: Artistic Statistics. In *Art in Literature, Literature in Art*. Ed. Magdalena Bleinert, Izabella Curyło-Klag, Bożena Kucała. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, 135-146.

Rybicki, J. and Eder, M. (2011). Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing* 26 (3): 315-321.

Said, E. (1993). *Culture and Imperialism*. London: Chatto and Windus.

Schler, J., Koppel. M., Argamon, S. and Pennebaker, J. (2006). Effects of Age and Gender on Blogging. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* 6, 199-205.

Singh, S. (2001). A Pilot Study on Gender Differences in Conversational Speech on Lexical Richness Measures. *Literary and Linguistic Computing* 16 (3): 251-264

Yu, B. (2014). Language and gender in Congressional speech. *Literary and Linguistic Computing* 29 (1): 118-132