Zooming in, zooming out : 30 years of Corpus Stylistics Bricolage

Anne Bandry-Scubbi

Zooming In, Zooming Out, 30 years of Corpus Stylistics Bricolage

Anne BANDRY-SCUBBI
University of Strasbourg, UR 2325 SEARCH
bandry@unistra.fr

Abstract

[This essay addresses the relationship between distant and close reading, advocating middle-ground reading, with a strong focus on text *per se*. In an evolutionary approach it reviews the pragmatic use over three decades of mechanically-enhanced reading of eighteenth-century fiction, with texts first seen as systems then as parts of embedded and overlapping corpora of varying sizes. Norm and typicality are explored with canonical and non canonical fiction.] 460

1 Introduction

In the words of the late John Burrows I "declare myself, first and last, a student of English literature," having taken up "computational stylistics" (2010: 13) because the methods and tools it provides make it possible to view texts from a somewhat different perspective, combining distant, middle-ground and close reading. I am no computer wizard and have almost never created the tools I use, but rather taken what was at hand, or within my grasp, technically, financially and intellectually. I therefore consider Lévi-Strauss's notion of bricolage an apt way to qualify thirty-odd years of research on British fiction from the eighteenth and early nineteenth century with the help of a computer, which has entailed the risk of not being taken seriously by specialist scholars. Being French, I was trained in the close-reading tradition (*explication de texte*) by heirs of structuralism: we saw the text as a "network of signifiers" (Barthes 1970), and our aim was to lay bare the logic of a text, or rather of a fragment of text, seen as a system, a coherent and dynamic whole. What I like to call "computer-aided textual analysis" extends this approach to complete literary works thanks to the non-linear reading enabled by considering a text as a "multi-dimensional space" (Rastier 2001: 93). Once literary texts became more widely available in digital form, this change of scale and of perspective provided "an enhanced contextualisation which changed the conception of textuality and intertextuality" with the possibility of examining a text within a set of corpora (Rastier 2001: 93). Gigantism is a temptation, taking into account the "99.5 per cent" which have not made it into the literary canon along with the .5 per cent that did (Moretti 2013: 66), reading them "distantly" but not as coherent narratives per se. The Stanford Literary Lab and others now provide the (very) big picture, usefully broadening the view of "the rise of the novel" into "one rise" or the gender-shift into several (Moretti 2005: 5, 27). "My" scale can be called middle-ground reading, at which the individual text, say, a novel, is analysed by zooming out onto a corpus, i.e. an organised set of texts encompassing it for a meaningful reason (time, genre, or criteria related to an hypothesis) and zooming in onto

some of its features which quantitative and qualitative comparison and contrast show to be relevant to that hypothesis. For technical, financial but mainly intellectual reasons, a coherent literary text forms my usual object of study. This essay focuses on the trials, errors and successes in an evolutionary approach similar to Martinet's (2020) and is therefore highly autobiographical with the embarrassment of a largely self-centred bibliography.

2 Texts as systems

In 1985 I was appointed as a teacher of English in the Science College of a small French university (Université de Haute-Alsace) and I began a PhD in British literature, on the texts written as reactions to Laurence Sterne's *Tristram Shandy*. The concomitance of these two events has shaped the whole of my research. It coincided with the implementation of the French national project *Informatique pour tous* (Computing for All), which made computers available for students and staff, but also in libraries, and led to private acquisitions of personal computers at affordable prices. I bought my first machine in 1987. Part of my teaching was English for computer science students. Trying to convince them that "if / then / else" could not be used in *real* English taught me how to communicate with programmers and I traded proofreading my colleagues' publications in English against devising a program which produced frequency indexes of the genuine and spurious *Shandy* volumes on university computers which I could then use on my own.

Software such as the Oxford Concordance Programme was out of my reach, as were digital texts, but networking before the internet, I strove to meet researchers in France and abroad who combined the study of literature, or at least humanities, and the use of computers. I visited the Oxford University Computing Services (OUCS) and Besançon linguists, and most usefully came in contact with Paris colleagues who had set up groups in their respective universities. I became a regular visitor to Liliane Gallet and Marie-Madeleine Martinet's CATI (*Cultures anglophones et technologies de l'information* / Anglophone Cultures and Information Technologies) in Paris IV Sorbonne and a regular contributor to Françoise Deconinck-Brossard's RAO group (*Recherche assistée par ordinateur* / Computer-aided research) in Paris X-Nanterre. These annual meetings along with frequent correspondence provided an ideal forum to discuss ideas, methodologies, protocols, analysis of data. Simultaneously, I was integrating the network of Sterne scholars, which gave me access to a printed concordance and frequency index of Sterne's sermons (thanks to Kenneth Monkman at Shandy Hall) and, even more preciously, a digital copy of Sterne's text. Somewhat like the original readers of *Shandy*, published in instalments from 1759 to 1767, I received the 6250 bpi magnetic tapes volume by volume over the year 1989 from an American scholar (Diana Patterson) who was typing them for a digital facsimile of the first edition. As is well known, Sterne exposes the workings of story-telling and of printing conventions, and by provoking readers invited reactions to his playful text, which were the focus of my PhD. By 1990 I had eleven volumes of *Shandy* in ASCII text: the nine genuine ones and the two spurious ones for which I had paid a typist but corrected her work very thoroughly as eighteenth-century English was beyond her skills. The OCR tests I had obtained from Strasbourg and Nice had

convinced me that typing could be a better method for pre-1830 texts.[1] I had also acquired *Textsearch* which ran on my home computer and added to indexing the possibility of concordancing.

I could then work in an approach not yet called Corpus Stylistics, naming it Computer-Aided Text-Analysis (ATAO in French: *Analyse textuelle assistée par ordinateur*) in discussions with Deconinck-Brossard, after computer-aided design, engineering, publishing, etc. If I was very much interested in technology, my aim was not to devise software but to use it in order to enrich my literary analysis of *Tristram Shandy* by looking at which traits of his writing were lampooned, parodied or pastiched. As Milic had written of Sterne, "a clever imitator could perhaps duplicate the imitable features of the vocabulary and thus render worthless the lexical criteria of style description and identification" (Milic 1967b). Statistics derived from frequency counts of vocabulary helped explain why the spurious third volume of *Shandy* read so badly while the spurious ninth was good enough to fool the first German translator. The most frequent words[2] gave insights not easily perceived by linear reading (my guides were Burrows on Austen (1987), Milic on Swift (1967a), Kenny 1986, Farringdon 1989), contrary to the shandean dashes and fragmented text which became the craze and were conspicuous on the page, whether genuine, spurious or simply written under the influence of Sterne. The clumsy spurious third volume had far less varied vocabulary despite its many narrative false starts, it resorted to the verb *be* far too often and did not manage to catch the correct ratio of *and* and *the*; on the contrary, the author of the 1766 spurious ninth volume did much better – but admittedly had more to work from. This study published in 1992 was expanded to three continuations of *A Sentimental Journey* (Sterne died shortly after having published Volumes I and II but had raked in subscriptions which announced III and IV, creating a potential market). I examined the most frequent adjectives and nouns to conclude, notably, that the least successful sequel contained the fewest terms designating males (Bandry 2000). Not many of the swarm of imitators managed to latch on to Sterne's ironic and deft sentimentalism. Their attempts at recreating his manner of writing provided me with the opportunity to explore writing taken as a quantitative norm (Sterne's) and the deviations of imitations. I also gave in to the attribution temptation with *The Clockmaker's Outcry against the Author of The Life and Opinions of Tristram Shandy*, in collaboration with my Sterne mentor Geoffrey Day. Circumstantial evidence proved far stronger than stylometric comparison.

I was lucky to be able to publish my findings in the journal which had hosted Deconinck-Brossard's "Confessions d'une dix-huitiémiste branchée" (Confessions of a Wired Eighteenth-Century Scholar), now *XVII-XVIII* (of which I became general editor from 2012 to 2018). We collaborated a few years later on Moll Flanders.[3] "On peut compter sur Moll" (You

[1] I visited Charles Muller in Strasbourg in 1990 and at his invitation met Etienne Brunet at a PhD viva a few months later. My questions were then largely about OCR.

[2] Being no linguist, I have always used a very basic definition of word: a sequence of letters between two blanks or punctuation marks. For the same reason I have never sought to lemmatize my texts automatically.

[3] In the meantime, I had been promoted to senior lecturer in the English department of Université de Haute-Alsace and the two of us we were teaching this novel in our respective universities. This computer-aided analysis tied in with work done by CATI on *Georgian cities*, in which Moll Flanders is one of the texts taken into account (for the evolution of this project see Martinet 2020).

Can Count on Moll, 1997) relies on frequency lists and keyword concordances, either quantitative or qualitative. By following some of the most frequent substantives throughout the text with concordances, such as *house*, we teased out some of the ways in which Defoe weaves his story and we challenged critical views of his text while confirming others.[4] *Gentlewoman* provides a striking example of how rewarding it is to follow a keyword along the entire text, as Defoe gives it from the start a tainted definition which the heroine adopts as a rule of conduct to the very end, through the vagaries of its 64 occurrences. No computer is needed to relish the irony of the final uses which seriously undermine the respectability of the denomination by the incongruousness of Moll's newly found adult son calling her a gentlewoman while she gives him a watch without "tell[ing] him [she] had stole it from a Gentlewoman's side." Yet following the progression with a concordance from each occurrence to the next puts in evidence the snowball effect by which each new use adds on meaning. Concordances helped us read the text as a system.

Using the text as a corpus required dividing it into parts. As Defoe's fiction has no chapters or other marked divisions, we divided it into 22 sections according to textual signals by which the narrator indicates the shift from one episode to the next. We did not want the software to decide on partition arbitrarily. We could then compare the parts in terms of vocabulary use. For the first time we plotted episodes and vocabulary on a map, established by correspondence analysis. From the low value of the first extraction and the very crowded graph of the first two factors, we concluded that Moll Flanders is a very homogeneous text but thereby exposed the readers of *XVII-XVIII* to greater complexity than they were accustomed to. We were trying out tools which became much easier to use (for me at least) with the program *Hyperbase*.

From 2001, *Hyperbase* became my main instrument, which I have been using on many corpora since, whereas Deconinck-Brossard applied different tools to her homiletic corpora. *Hyperbase* is the brainchild of Etienne Brunet, a French scholar and developer, who had started with published frequency indexes of an author or a period: *Le vocabulaire de Jean Giraudoux: structure et évolution* (1978), *Le vocabulaire de Proust* (1983), *Le vocabulaire français de 1789 à nos jours* (1981).[5] I invited him to my Master's seminar in 2007 and my students presented computer-aided analyses they had prepared on a corpus of modernist short stories as a basis for discussion. In a few clicks, *Hyperbase* provides user-friendly graphs.

The first of my textual analyses which relied on *Hyperbase* focused on *Gulliver's Travels*. In "Gulliver et la machine à compter: une étude de spécificités" (Gulliver and the Counting Machine: a study of quantitative keyness, Bandry 2001) I explored quantitative keyness by contrasting the four voyages. The French term "spécificités" is less of a challenge for literary scholars, particularly with the crystal-clear definition by Lebart & Salem: "forms 'abnormally' frequent in one part of the corpus" (1994: 260; the calculation relies on the

[4] I had done this with life and opinions in *Shandy* and some of the spinoffs (Bandry 1993). Very often, the last use of an important term comes with an ironic twist.

[5] *Hyperbase* uses *Le Trésor de la langue française* by default for external comparisons but this can be switched to the *BNC* for English; dictionaries for Italian and Portuguese are also available. The software can be used with any Latin alphabet.

hypergeometric model). Results obtained from *Gulliver's Travels* provided a modest proposal for a new way to look at a text which has given rise to a huge amount of critical work. Far less ambitious than Milic's *A Quantitative Approach to the Style of Jonathan Swift* (1967a), my contribution aimed at offering evidence of features of the text established from non-linear reading. The three analyses I drew from quantitative keyness are firstly that the expression of size contradicts expectations: terms of bigness appear more frequently in Lilliput and of smallness in Brobdingnag. Secondly, that the narrowness of the Houyhnhnms' world and vocabulary is all the more effective as it comes after the richness of the Voyage to Laputa, Balnibarbi, Luggnagg, Glubbdubdrib and Japan. This can be summed up by one fact: "the thing that is not" brings no new word, and certainly no *hapax* (the third Voyage has 15% more types and 22% more *hapax legomena* than expected if the other three are taken as a norm). The third main finding is that the receding use of first-person plural pronouns combined with the increasing third-person plurals over the four books conveys Gulliver's gradual alienation from the human race: by the fourth voyage, humans are *they* rather than *we*. This reading did not reveal anything new about Swift's famous book but showed how the writer achieves the effects created by the text. Going back and forth between book and data made it possible to apply techniques of close reading to a text of slightly over 100 000 words.

3 Corpus-based Approach

My next experiment was less convincing but an important step. Thanks to the digital turn taken in most universities and to a paper on my Sterne findings presented before the cross-border informal research group of English studies (EUCOR), I was given access to the treasure-trove of Chadwick-Healey's *Eighteenth-Century Fiction Database* which Basel University had acquired but nobody used – and most French university libraries could not afford. After a frenzy of downloading onto floppy disks, I came home with reliable electronic editions of nearly 100 texts I could use on my personal computer.[6] However, overwhelmed by quantity, I did not at first adopt a corpus approach. At several conferences I examined a key term or set of terms in a series of fictions from the eighteenth century to explore a theme but it took me a while to realise why I had lost the satisfaction of providing a view of what I like to call the texture of the text and could occasionally illustrate by textual imaging, a phrase coined after medical imaging (mainly bar charts at that point, produced by a spreadsheet or *Hyperbase*).

I therefore decided to explore in depth how one could use a corpus approach, with *Excel* and *Hyperbase* as my tools. I first concentrated on Defoe's fiction with eight texts,[7] starting from what I was comfortable with and pushing it somewhat further: lexical richness, distribution of most frequent words (pronouns, nouns, verbs), a factorial analysis map representing lexical connection which I could explain better as *Hyperbase* provided easier to read graphs, and a corpus constituted of eight parts was more manageable than the twenty-two we had devised

[6] Reliable electronic text had been a challenge so far, as Gutenberg.org editions were not always so, *Shandy* being a case in point. I had carefully proofread the digital *Moll Flanders* and *Gulliver's Travels* before using them.

[7] The semi-fictional status of the *Journal* was of particular interest. The debate on Defoe de-attributions had begun some 15 years earlier.

for *Moll Flanders*. A readily comprehensible paper had shown me the way: it used *Hyperbase* to compare Seneca's tragedies (Mellet 1998). In the same manner, the eight Defoe novels are positioned according to the vocabulary they share.[8] Colleagues in mathematics helped me by pointing out that if the first results made sense in terms of what I knew about the texts from literary history and traditional stylistic analysis, I could trust the data the software was producing and elaborate further analyses from it. The texts were positioned in ways that made sense both in their groupings (the two *Robinsons*, the two female stories) and in their differentiations (the texts really considered as fiction *vs* those with more ambivalent status, with *Journal of the Plague Year* in the middle). I then concentrated on quantitative keyness as I had done with *Gulliver's Travels* but pushed this to groups of texts in order to examine whether the congruence between the two *Robinsons* and that between *Moll* and *Roxana* were thematic or came from specific stylistic traits. Other eighteenth-century texts provided external elements of comparison, most notably *Gulliver*, *Joseph Andrews*, *Memoirs of a Woman of Pleasure*, *Betsy Haywood*.[9] In the background a more specific research question began to take shape: what differentiates stories of female characters from those of male ones, and how this combines with whether they were written by men or by women. Not unexpectedly, what drew the most interest from the very few readers of this unpublished essay was the comparative analysis of how the combined uses of the verbs *see* and *know* structure *Moll Flanders* and *Roxana*, which considers the texts as a system rather than as a corpus.[10] I had spotted these two verbs from the list of words whose frequency increases as the corpus unfolds, another measure provided by *Hyperbase*, and therefore made the methodological point that looking at text as data shows what to focus on: underlying features of style, not necessarily perceived in linear reading or traditional literary analyses.

The second essay on Eliza Haywood's fiction aimed to explore the differences and similarities between her 1720s racy texts and the ones written in the wake of – and reaction to – Richardson's *Pamela*, as Haywood, both hailed and derided as the "Great Arbitress of Passion," had been able to adapt to the changing taste and produce fiction that sold in the 1740s after having been a best-selling author twenty years earlier.[11] The first stage of the study was to identify Haywood's vocabulary in *Love in Excess* by contrast with that of the other 1719 bestseller *Robinson Crusoe* and the just as successful 1726 *Gulliver's Travels*. Each text was divided into two narratively logical sections so as to constitute a corpus of six parts roughly equal in length (from 61,000 to 41,000 tokens). Lexical richness established from a set of comparisons (type/token ratio, *hapax legomena*, successive 1000-word sections) ranked *Gulliver* first, *Love in Excess* second and *Robinson* third. It confirmed the interest of studying Haywood, whom recent research was pushing as a major forgotten author to be canonised in the feminist rewriting of literary history, in reaction to Ian Watt's founding

[8] A word contributes to drawing two texts together if it belongs to both, and to pulling them apart if it only occurs in one of them (Brunet 60).
[9] The pornographic content of Cleland's *Memoirs* (1748-49) does not prevent it from having a very strong formal similarity with contemporary fiction.
[10] The essay and the one on Haywood were part of my *Habilitation*. The analysis mentioned is now available in Bandry 2018.
[11] The qualification comes from "To Mrs Eliza Haywood on her Writing" (1732) while she had been one of Pope's victims in sexist terms (*Dunciad Variorum*, 1729).

critical text *The Rise of the Novel* and his infamous dismissal: "The majority of eighteenth-century novels were actually written by women, but this had long remained a purely quantitative assertion of dominance" before Burney and Austen (1957: 298).[12] Some of my conclusions confirmed well-known facts: Haywood's characteristics were her breathless syntax (the paucity of *and* and *or* is made up for by a superabundance of exclamation marks, parentheses and dashes, of which the recent paper edition provided a reliable proof) and the repetitive use of specific (and thematic) words: *love*, *passion*, *friendship*, *cry*, *opportunity*, *desire*.[13] It seems a truism to find these features by contrast with *Gulliver's Travels* and *Robinson Crusoe*. However, the comparison with the quantitative keywords of Defoe's stories of women shows that he did not adopt his rival novelist's vocabulary for his ventures into the feminine, concentrating instead on their survival: *house*, *money*, *husband* and female status (*girl*, *woman*, *mother*). Haywood is more interested in the relationships between characters, and particularly but not only, between women and men. In the rest of her fiction, both from the 1720s and the 1740s-50s, she explores different combinations. A quantitative keyword approach in a corpus comprising seven texts of varying size by Haywood shows how she varies her recipes with the same ingredients, toning down the raciness in the 1740s but keeping a strong interest in expressing female sensuality and agency. The difference in size between the parts of the corpus was taken into account as a possible factor for some of the stylistic analyses.[14] Despite this, the shortness of quantitative keywords lists for each part shows the strong homogeneity of the HAYWOOD corpus, in contrast with the long lists from DEFOE. Haywood varies the distribution of what is clearly *her* vocabulary. In this study (2004) and two later articles I then combined the corpus approach with a focus on one particular text (Bandry-Scubbi 2010, 2012). Like the view of *Gulliver's Travels* provided by the contrastive study of quantitative keywords, the back and forth movements between a text and a corpus to which it belongs, zooming in and zooming out, contextualisation and intertextualisation, gave me the opportunity of drawing out some stylistic features of texts which were becoming part of the literary canon.

4 Embedded and overlapping corpora

This dual level of analysis was set up on a larger scale for the study of another eighteenth-century novel, *Roderick Random* (Bandry-Scubbi 2009). Smollett's 1748 first published work (c. 200,000 words) was written in the context of "the rise of the novel," in rivalry with Fielding who had positioned his 1742 *Joseph Andrews* as a "new species of writing." The

[12] The seminal series of essays *The Passionate Fictions of Eliza Haywood* was published in 2000 by Kirsten T. Saxton and Rebecca P. Boccicchio (UP Kentucky, 2000) after Paula Backscheider, a renowned Defoe scholar, had called for the reinvestigation of her work: "we must […] problematize, complicate, and revise many of the commonly accepted opinions about Haywood's work" ("The Shadow of an Author: Eliza Haywood", *ECF* 11.1 1998: 90). David Oakleaf's 2000 edition of *Love in Excess, or the Fatal Enquiry* (Broadview) provided the necessary reliable paper copy against which to compare the electronic Chadwick-Healey version.

[13] A strongly marked preference for 'words' in Haywood and 'word' in Defoe and Swift tipped the scale in favour of not lemmatizing, a point advocated by linguists but which I have often found counterproductive for literary analysis. Hyperbase makes it very easy to draw up lists and so to gather the forms of a lemma easily when useful.

[14] Two long novels were divided into their initial volumes so that the parts of the corpus varied from 12,000 to 85,000 tokens.

third open contender was Richardson who had taken the literary market by storm in 1740 with his own "new species of writing," *Pamela*. Both Fielding and Haywood had reacted to Richardson's epistolary novel, the first with the pithy *Shamela* and the second with the somewhat rambling *Anti-Pamela*, both part of a vast movement now dubbed The Pamela Vogue.[15] Both then went on to write several examples of what came to be called novels. I therefore set up a series of corpora to draw out the stylistic features of *Roderick Random*. 1740S consists of fiction of comparable length from that decade with a balance between male and female authors as well as between stories of male or female protagonists. *Random* being written as a first-person narrative, 1STPERSON is composed of fictional autobiographies from the eighteenth century, with some novels also present in 1740S. The risk of bias caused by the prevalence of Defoe is balanced by the advantage of having several corpora of similar size including the novel under study, as *Random* also belongs to the SMOLLETT corpus which comprises all of Smollett's fiction. I therefore had three corpora of over a million words each, countering the objections of linguist colleagues that I did not have enough data to work from. Moreover, *Random* taken on its own makes up two different corpora: one in which it is divided into its 69 chapters and another into 9 parts, the main stages of the narrative. These overlapping corpora made it possible to reinvestigate the claims made in the 1970s by Smollett scholars who had examined his style without the help of computers but with the logic of samples, which also prevailed in the work of the first digital stylistic studies such as Milic's for reasons of available computer capacity. Indeed, the features of what was described as "writing in the superlative" (Boucé 1971) mainly occur in the specific samples they selected (Grant 1977). Yet these scholars dealt with what one of them called "language as projectile" (Grant 1982), the very fast pace of this story of an impulsive and lusty young male character depicted in a series of violent adventures so as to make him illustrate the name his author gave him, Random.

Focusing on this text set in several corpora can be likened to a kaleidoscope: changing the corpus in which *Random* is observed shows different features by comparison and contrast. Some traits specific to this text recur whatever the corpus, others show Smollett's conformity to the writing conventions of his contemporaries. Quantitative keyness drawn out from 1740S, 1STPERSON and SMOLLETT enabled me to zoom in on the stylistic means which make the reader feel rushed through the text. Syntax constituted a first set of features with sentence length, quantity of parentheses – examined with due caution –, the heavy use of WH-relative clauses, characteristic of Smollett's early fiction as the SMOLLETT corpus showed, along with the knowledge that he shortened his sentences when revising his second novel *Peregrine Pickle* (Boucé 1971). I then looked at indications of how "time is collapsed" (Stevick) with the accumulation of vocabulary indicating temporality, and at the way Smollett refers to the body (verbs and organs, body parts – only *Woman of Pleasure* ranks higher among the 19 novels taken into consideration). Another quantitative keyword led to understanding Smollett's way of relating a fictional autobiography: *my* is the most specific word of *Random* in 1STPERSON. It can be deduced from concordances that this fictional world is organised

[15] All these texts are available in image form in Chadwick Healey's *ECCO*. At DH2016 they outrageously proposed to provide all of the material in text form to subscribing libraries who would agree to pay an additional fee.

around the speaker who mentions "my" body, "my" situation and "my" acquaintances more than himself as "I" (Roderick is not the only user of the first person of course). To draw these features out, I experimented with lists, on a larger scale than in Bandry-Scubbi & Deconinck-Brossard 2005: WH-words, time indications, vocabulary of the body. A version of this work more focused on its use of *Hyperbase* got me into Brunet's session at JADT 2010. However, my strong stylistic bent put me in an awkward position among scholars like Jean-Marie Viprey who had moved from their inspiring analyses of an author's style (1997) to demonstrating the software they were fine-tuning. Meanwhile my work was deemed too technical to be published in volumes derived from conferences primarily concerned with literature or cultural studies at which I presented papers on the use of body words by female authors, with my corpora as a backdrop (what hands do in *Pamela*, *Evelina*, *Isabella* or *Pride and Prejudice*, for instance). I nearly gave Corpus Stylistics up.

However, its use by a doctoral student of mine to study the expression of space in children's fiction over a century confirmed its potential.[16] In the meantime, Deconinck-Brossard and I had compared the sermons and fiction of Sterne and Swift to test whether genre or author was the strongest criterion of authorship, with the interesting twist of Sterne having woven one of his published sermons into *Shandy* (2005). This study relied on our previous work and on Biber's categories. We looked at very frequent words and *hapax legomena*, quantitative keyness, verbs of persuasion and of assertion, the latter identified by Biber as indicative of narration. We divided *Shandy* into its chapters to have units of a size comparable to sermons, examined Swift and Sterne's sermons within a larger homiletic corpus and reached the expected conclusion that genre prevailed over authorship. Our interest resided in testing methods under the aegis of the French Society of English Stylistics. We had been discussing the notions of stylistic signature (Milic) or linguistic fingerprint ever since our first encounter in the early 1990s, talking with computer scientists and forensic linguists, taking into account the computer-assisted analysis of Romain Gary writing under the pseudonym of Emile Ajar which shows that an author can change the way he writes (Tirvengadum 1998).

In 2012 I discovered Chawton House Library's Novels Online collection. This "ongoing project [aims at] making freely accessible full-text transcripts of some of the rarest works in the Chawton House library collection," which consists in "works by women, mostly in English, and mostly within the period 1600-1830."[17] I had found my second treasure-trove, thanks to which I could combine all my experience, working with a large quantity of reliable electronic texts, both canonical and non-canonical. I set up a project entitled "Strategies of Writing: Women's Fiction in the Long Eighteenth-Century, a Corpus-Based Stylistic Analysis" and read up on critical work by scholars linked to Chawton as well as recent advances in what was now called Corpus Stylistics (Mahlberg, Fischer-Starcke, Jockers, Stanford Literary Lab). My findings were published by *ABO: Interactive Journal for Women in the Arts, 1640-1830*, whose aims fitted my objective: computer-aided literary analysis (Bandry-Scubbi 2015). The challenge I had set myself was to identify features which

[16] Caroline Orbann, "L'Espace imaginaire dans le roman de jeunesse britannique : de Water-Babies de Charles Kingsley (1863) à Charlie and the Great Glass Elevator (1973)," co-directed with Professor Monique Chassagnol, defended 2016.
[17] https://chawtonhouse.org/the-library/library-collections/womens-writing-in-english/novels-online/

constitute typicality within a set of 42 novels by women published between 1752 and 1834, 34 of which come from Chawton Novels Online and 8 are now part of the canon, by Jane Austen, Frances Burney, Maria Edgeworth and Eliza Haywood. They come under the heading of "feminine" novels, "domestic comedy, centring on a heroine, in which the critical action is an inward progress towards judgment" (Butler: 145), neither gothic nor historical, the other two main categories of fiction which developed in the period. A reference corpus was set up with 34 novels answering the same space and time criteria, half of them by male authors. The choice was determined in part by which texts were available online. The CHAWTON34 corpus is embedded within WOMEN42, which partly overlaps with CONTROL34, both comprising over 5 million words. 7 of the Chawton texts were issued by the (in)famous Minerva Press, which had a reputation for saturating the market with formulaic novels: they turned out to be indistinguishable from the rest on lexical connection maps. Correspondence analysis on types distinguishes female texts from male ones, indicating that specific vocabulary is used by each gender. Quite logically, the same process applied to tokens separates texts according to whether their protagonists are male or female. Quantitative keyness of texts by female authors in CONTROL34 shows that what characterises these texts is an interest in both genders (whereas fiction by men in the reference corpus takes females into account to a much lesser extent), a strong use of small-group interaction and of dialogue, with a particular liking for *cried*, which points to the prevalence of intensity, and a concern for feelings and emotions, along with family, marriage, and sight. Typicality can reside in the rate at which words are used, as one "eccentric" text becomes unexceptional when tokens are taken into account rather than types. The list of disproportionately frequent terms in *Three Weeks on the Downs* compared to the Chawton corpus shows that its originality comes only from its setting (a ship), not the way in which it uses vocabulary. The rest of the title gives away the reason for its presence in the corpus: *or Conjugal Fidelity Rewarded, Exemplified in the Narrative of Helen and Edmund*. WOMEN42's most central text in terms of tokens was chosen as a case study, for it presents the seeming paradox of being "a highly original tale" (Brown, Clements, Grundy 2006) told in unoriginal terms. *Rachel*, a didactic tale, exemplifies the norm of the corpus by playing with concepts which had become somewhat outmoded by the time it was published, such as sensibility. This quantitative analysis came to the same conclusion as one of the contemporary reviewers (Bandry-Scubbi 2015: 21-22). The use of tokens clusters most canonical texts at a safe distance from most Chawton novels. From this, it can be inferred that Austen, Burney, Edgeworth and Haywood (unsurprisingly) drew from the same stock of words as did their less famous female contemporaries, but they used them at different rates. The difference between these canonical texts and the rest of the Reference corpus provides a basis from which to identify common vocabulary traits. This approach makes the less famous fiction a benchmark, rather than providing a normative judgment of texts that left a small footprint in literary history.

Although the Chawton staff did not know of any other use made of these electronic texts, I was surprised to receive an article while I was giving a paper on "Women's Novels 1750s-1830s and the Company They Keep: A Computational Stylistic Approach" at DH2016 in Krakow. Using far more sophisticated techniques, Jan Rybicki had compared the Chawton texts to those by 'famous men' and 'famous women' authors of eighteenth- and early

nineteenth- century fiction to identify the gender of authors from multivariate analysis of the 100-1000 most frequent words. We reached the same main conclusion: women who became famous (Burney, Edgeworth, Austen) were the ones who wrote more like their male counterparts. This confirmed me in my choice of sticking to a number of novels I can actually read and focus on, navigating between close and middle-ground, rather than distant, reading, because my interest resides in how stories are told, how unfolding a text into a network of words (the French *explication de texte*) enables one to relish its texture.

I have since relied on this study as a basis to examine different themes, sometimes specific (the narrative use of miniatures in Bandry-Scubbi & Friant-Kessler 2018), sometimes wide (the modes and narrative use of leisure in a forthcoming article).[18]

5 Conclusion

"What Corpus Stylistics can do beyond the obvious provision of quantitative data is help with the analysis of an individual text by providing various options for the comparison of one text with groups of other texts to identify tendencies, intertextual relationships, or reflections of social and cultural contexts" (Mahlberg 2007 : 221). When I came upon Mahlberg's work, I related strongly to the new name given to the kind of research I had been doing all along; it states an equal status with corpus linguistics and shows the evolution of such work since the 1950s and 1960s pioneers who called it lexicometry or stylostatistics. The recent definition of style by Hermann *et al.* takes corpus stylistics into account, but not exclusively: "Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" (2015: 44). This shows, I think, the maturity of the discipline. All scales are needed: close reading, distant reading and, I claim, middle-ground reading. In the interplay between a specific text and a set of carefully constructed corpora, words "knock on the door" (Rastier 2001: 96) thanks to results such as quantitative keyness or a systematic look at concordances. Stylistic hypotheses can thereby be formulated, checked, proven wrong, lead to others and participate in the rewriting of literary history (Moretti 2013: 64) or more modestly, the critical view of a text. Hopefully, such tinkering often proves more fruitful than that of the Lagado professor: "he had emptyed the whole Vocabulary into his frame, and made the strictest Computation of the general Proportion there is in Books between the Number of Particles, Nouns, and Verbs, and other Parts of Speech" (Swift 2008: 172). The result derided by Gulliver should coalesce into major works "improving speculative Knowledge" but has not gone beyond the stage of "several Volumes in large Folio […] of broken sentences." With or without an Oulipian touch, mechanically-enhanced writing is now on its way. Mechanically-enhanced *reading* should not lose sight of the literary text as an entity with its own logic aiming to give the reader – whether or not a scholar – pleasure.

[18] I also supervise two doctoral theses on similar corpora, one with a quantitative approach and one without: Juliette Misset, "Reading the Didactic Mode: British Novels, 1778-1814"; Lucy-Anne Katgely, "Entre obscurité et renommée: trajectoires et chemins de traverse des romans britanniques féminins de 1789 à 1830."

References

Bandry A. Gulliver et la machine à compter: une étude de spécificités. XVII-XVIII 2001; 53(1): 145-157.

Bandry A. Les livres de Sterne: Suites et fins. XVII-XVIII 2000; 50(1): 115-136.

Bandry A, Deconinck-Brossard F. On peut compter sur Moll [Flanders]. XVII-XVIII 1997; 45: 171-190.

Bandry-Scubbi A. Chawton Novels Online, Women's Writing 1751-1834 and Computer-Aided Textual Analysis. ABO: Interactive Journal for Women in the Arts, 1640-1830 2018; 5(2), Article 1. DOI: http://dx.doi.org/10.5038/2157-7129.5.2.1 http://scholarcommons.usf.edu/abo/vol5/iss2/1

Bandry-Scubbi A. La difficile acceptation du foyer dans The History of Jemmy and Jenny Jessamy d'Eliza Haywood. In: Eric Lysøe editor. Signes du feu. Paris: L'Harmattan, 2010: 87-97.

Bandry-Scubbi A. Les mots de Haywood. In: DEconinck-Brossard F, editor. Recherche Assistée par Ordinateur (RAO). Paris: Université Paris-Nanterre: 2004.

Bandry-Scubbi A. Renaissance de/chez Eliza Haywood. In: Claire Bazin C, Leduc G, editors. Littérature anglo-saxone au féminin: (re)naissance(s) et horizons, XVIIIe-XXe siècles, Paris: L'Harmattan, 2012:17-39.

Bandry-Scubbi A. Roderick Random amidst Eighteenth-Century Fiction: a computer-aided textual analysis. XVII-XVIII 2009; 66: 205-225.

Bandry-Scubbi A. Roxana, réseaux de mots, prisons des corps. Journée d'étude spéciale agrégation d'anglais, 24 novembre 2018, Université de Lorraine. https://videos.univ-lorraine.fr/index.php?act=view&id=7240.

Bandry-Scubbi A, Deconinck-Brossard F. De la lexicométrie à la stylostatistique? Sterne et Swift: textes croisés. Bulletin de Stylistique anglaise 2005; 26: 67-85.

Bandry-Scubbi A, Friant-Kessler B. Peindre en corpus: Miniatures et roman anglais féminin (1751-1834). In: Deconinck-Brossard F, Gallet-Blanchard L, editors. Palette pour Marie-Madeleine Martinet, 2017.

Barthes R. S/Z. Paris: Seuil, 1970.

Biber D. Variation across Speech and Writing. Cambridge: Cambridge University Press, 1988.

Boucé P-G. Les romans de Smollett. Paris: Didier Erudition, 1971

Brown S, Clements P, Grundy I, editors. Orlando: Women's Writing in the British Isles from the Beginnings to the Present. Cambridge: Cambridge University Press Online, 2006. <http://orlando.cambridge.org/>.

Brunet E. Manuel de référence pour Hyperbase 9.0. Nice: Université de Nice Sophia-Antipolis, 2011.

Burrows J.F. Computation into Criticism: A Study of Jane Austen's Novels and an Experimental Method. Oxford: Clarendon, 1987.

Burrows J.F. Never Say Always Again: Reflections on the Number Game. In: McCarty Willard, editor. Text and Genre in Reconstruction. Cambridge: Open Book, 2010. 13-36.

Butler M. Jane Austen and the War of Ideas. Oxford: Clarendon, 1975.

Farringdon M.G, Farringdon J. A Stylometric Analysis. In: Battestin M, editor. New Essays by Henry Fielding: his contributions to The Craftsman (1734-1739) and Other Early Journalism. Charlottesville: University Pres of Virginia, 1989: 549-591.

Fischer-Starcke B. Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries. London: Continuum, 2010.

Grant D. Roderick Random: Language as Projectile. In: Bold A, editor. Smollett: Author of the First Distinction. London: Vision, 1982: 129-147.

Grant D. Tobias Smollett: A Study in Style. Manchester: Manchester University Press, 1977.

Herrmann B, van Dalen-Oskam K, Schöch C. Revisiting Style, a Key Concept in Literary Studies. JLT 2015; 9(1): 25-52.

Hyperbase: Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels. Version 9. 2011. Université de Nice.

Jockers M. Macroanalysis: Digital Methods & Literary History. Urbana, Chicago: University of Chicago Press, 2013.

Kenny A. A Stylometric Study of The New Testament. Oxford: Clarendon, 1986.

Lebart L, Salem A. Statistique textuelle. Paris:Dunod, 1994.

Lévi Strauss C. La pensée sauvage. 1962. Paris: Plon.

Mahlberg M. Corpus Stylistics: Bridging the Gap between Linguistic and Literary Studies. In: Hoey M, Mahlberg M, Stubbs M, Teubert W, editors. Text, Discourse and Corpora: Theory and Analysis. London: Continuum, 2007.

Martinet M-M. Digital Representation of City Cultural History: Feedback on the Twenty-year Long Interdisciplinary Experiment. ILCEA 2020; 39. DOI: https://doi.org/10.4000/ilcea.8645

Mellet S. Les tragédies de Sénèque vues à travers Hyperbase. In: Mellet S, Vuillaume M, editors. Mots chiffrés, mots déchiffrés: mélanges offerts à Etienne Brunet. Paris: Champion, 1998: 255-271.

Milic L.T. A Quantitative Approach to the Style of Jonathan Swift. The Hague: Mouton: 1967a.

Milic L.T. Information Theory and the Style of Tristram Shandy. In: Cash A, Stedmond J, editors. The Winged Skull: Papers from the Laurence Sterne Bicentenary Conference. London: Methuen, 1967b.

Moretti F. Distant Reading. London: Verso, 2013.

Moretti F. Graphs, Maps, Trees. London: Verso, 2005.

Rastier F. Arts et Sciences du texte. Paris: PUF, 2001.

Rybicki J. Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies. Digital Scholarship in the Humanities 2016; 31(4): 746–761. https://doi.org/10.1093/llc/fqv023

Swift J. Gulliver's Travels. 1726. Oxford: Oxford University Press, 2008.

Textsearch: A Full-Text Retrieval System for Humanities Research. Vers. 3.1. LinguaTECH, 1987.

Tirvengadum V. Linguistic Fingerprints and Literary Fraud. CHWP 1998; A.9. https://projects.chass.utoronto.ca/chwp/tirven/index.html

Viprey J-M. Dynamique du vocabulaire des Fleurs du Mal. Paris: Champion, 1997.