**Title:** Periodizing Samuel Beckett’s Works: A Stylochronometric Approach

**Abstract:** We report the first analysis of Samuel Beckett’s prose writings using stylometry or the quantitative study of writing style. A novelty of this paper is that we focus on grammatical function words in Beckett’s works, a linguistic category which has not been studied before in Beckett scholarship. To these function words, we apply methods from computational stylistics to model the stylistic evolution in Beckett’s oeuvre. Our analyses reveal a number of interesting novelties that shed new light on existing periodizations in the secondary literature of Beckett’s oeuvre in an ‘early’, ‘middle’, and ‘late’ period. We analyse Beckett’s prose writings in both English and French, demonstrating notable symmetries and asymmetries between both languages.

**Author names and affiliations:**

\* Dirk van Hulle, [dirk.vanhulle@uantwerp.be](mailto:dirk.vanhulle@uantwerp.be), University of Antwerp, Prinstraat 13, room S.D.137, Prinsstraat 2000, Antwerp, BELGIUM.

Mike Kestemont, [mike.kestemont@uantwerp.be](mailto:mike.kestemont@uantwerp.be), University of Antwerp, Prinstraat 13, room S.D.118, Prinsstraat 2000, Antwerp, BELGIUM.

**\* = Corresponding author**

**3-5 Highlights:**

* We offer the first stylometric analysis of Samuel Beckett’s prose writings.
* We use quantitative methods to study the evolution and periodization of Beckett’s oeuvre.
* We focus on grammatical function words, a linguistic category which has not been studied before in Beckett scholarship.
* Our analyses reveal a number of interesting novelties that shed new light on existing periodizations.
* We examine Beckett’s prose in two languages, demonstrating notable symmetries and asymmetries.

**Keywords**

Samuel Beckett, Stylometry, Function words, Stylochronometry, periodization

**Acknowledgements:** Dirk Van Hulle’s research leading to these results has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 313609. Mike Kestemont was funded for this research as a post-doctoral researcher for the Research Foundation of Flanders. The authors would like to thank Folgert Karsdorp for kindly sharing software implementations of various methods used in this papers (e.g. https://github.com/fbkarsdorp). Other software libraries which we explicitly would like to acknowledge include: *librosa* (McFee *et al.*, 2015), *sklearn* (Pedregosa *et al.*, 2011), *matplotlib* (Hunter, 2007), *seaborn* (doi:http://dx.doi.org/10.5281/zenodo.12710) and *nltk* (Bird *et al.*, 2009).

**Author biography**

**Dirk van Hulle**, PhD, is Professor of English Literature and director of the Centre for Manuscript Genetics at the University of Antwerp. He is the editor of the new *Cambridge Companion to Samuel Beckett* (2015) and a trustee of the International James Joyce Foundation. With Mark Nixon, he is co-editor of the *Beckett Digital Manuscript Project* and editor-in-chief of the *Journal of Beckett Studies*. He is the author of *Textual Awareness* (2004), *Manuscript Genetics: Joyce’s Know-How, Beckett’s Nohow* (2008), *Modern Manuscripts: The Extended Mind and Creative Undoing* (2014), co-author with Mark Nixon of *Samuel Beckett’s Library* (2013) and with Shane Weller of *The Making of Samuel Beckett’s L’Innommable / The Unnamable* (2014).

**Mike Kestemont**, PhD, is a postdoctoral researcher for the Research Foundation of Flanders at the University of Antwerp. He specializes in computational text analysis for the Humanities, in particular stylometry or computational stylistics. He has published on the topic of authorship attribution in various fields, such as Classics or medieval European literature. Mike actively engages in the debate surrounding the Digital Humanities and attempts to merge methods from Artificial Intelligence with traditional scholarship in the Humanities. His website (www.mike-kestemont.org) contains pointers to his recent scholarly activities, including an open access scientific documentary about stylometry and Hildegard of Bingen (vimeo.com/70881172).

**1. Introduction**

Probably best known as the author of *En attendant Godot* / *Waiting for Godot*, Samuel Beckett was not only a bilingual playwright, but also a poet, translator, essayist and novelist. Notably his prose fiction will be the focus of this article, in which we propose a quantitative method to delineate a periodization of Beckett’s œuvre. In art studies in general, there is a tradition of distinguishing an ‘early’ and ‘late’ period in an artist’s work, sometimes with a distinct ‘middle’ period in between. The late Beethoven sonatas are a good example, or the early Rembrandt’s ‘smooth’ style versus the rough paint surfaces of the late Rembrandt. In literature, the ‘early’ novels of Jane Austen (such as *Elinor and Marianne*, *First Impressions* and *Susan*) are distinguished from their published counterparts (*Sense and Sensibility*, *Pride and Prejudice*, *Northanger Abbey*), and even earlier works are referred to as ‘youthful writings’ by ‘Young Jane’ (Byrne, 2015). Shakespeare’s plays are commonly grouped into the ‘early’, ‘middle’ and ‘late plays’. But it is often difficult to determine exactly when an author’s work moves from, say, the ‘early’ to the ‘middle’ stage.

In Beckett studies, we find the same pattern of periodization, but also the difficulty of clearly determining where one period ends and the next one begins. Peter Boxall problematizes the idea of periodizing Beckett’s œuvre, because the neatness of such a narrativization entails the danger of doing injustice to the singularity of the individual works. But he admits that it is hard not to parcel his work into a beginning, a middle and an end. And his version of this narrative runs ‘from the Joycean extravagance of his early, mannered work, through the comic agony of frenzied becoming in his middle period, to the bleached impossibility of his later prose’ (Boxall, 2015: 34). He even sketches an outline of these three phases: the early period up to and including the novel *Watt*, written during the Second World War; the rich middle period up to and including *The Unnamable*; and the later, ‘stunted’ and ‘halting’ prose after the close of *The Unnamable* (33).

In ‘Early Beckett’, his recent chapter in *The* *New Cambridge Companion to Samuel Beckett*, John Pilling suggests a break that is marked by Beckett’s decision to write in French:

Eventually, though it took a long time to emerge, it became expedient for Beckett to insist upon a break between his pre-war writing and his post-war writing, as expressed not only in conversation with Charles Juliet but with many others. By abandoning his native English, the line became easier to draw (Pilling, 2015: 28).

This periodization, however, is more complex than it may seem. During the war, Beckett wrote a novel in English: *Watt* is therefore neither ‘pre-war’ nor ‘post-war’. Moreover, the (end of the) Second World War did not coincide exactly with Beckett’s decision to start writing in French. He had already written almost a dozen poems in French in 1937 (*Dieppe*) and 1938. This essay will concentrate on Beckett’s prose fiction, and – with this focus in mind – ‘early Beckett’ can be said to consist of the collection of stories *More Pricks Than Kicks* (which Beckett started writing in 1931, published in 1934), *Dream of Fair to Middling Women* (written in 1932, published posthumously in 1992), the novels *Murphy* (started in 1935, published in 1938) and *Watt* (started during the War in 1941, published in 1953).

The ‘middle Beckett’ would start with the *Nouvelles* and the novel *Mercier et Camier*. But again, this should be nuanced. Beckett started writing the first of the *Nouvelles* in English. He ‘drew the line’ between English and French by literally drawing a line in the middle of his story *La Fin* (originally called *Suite*). The manuscript held at Boston College (BC MS 1991-001, ref 53, 31r) shows how Beckett first started in English and, after about 28 pages, suddenly drew a line (in March 1946) and started writing in French. After this story, he wrote the novel *Mercier et Camier* (started in 1946, published in 1970 in French, in 1974 in Beckett’s English translation) and the other *Nouvelles*: *L’Expulsé* (1946, published in 1946; *The Expelled*, published in 1962); *Premier Amour* (1946; published in 1970; *First Love*, published in 1973); *Le Calmant* (1946, published in 1955; *The Calmative*, published in 1967).

The remarkably creative five-year period after the War was referred to (by Beckett himself) as ‘the siege in the room’ (Knowlson, 1996; Bair, 1978: 346). In this period, he also wrote the three novels *Molloy* (started in 1947, published in 1951; English version published in 1955), *Malone meurt* (started in 1948; published in 1951; *Malone Dies*, published in 1956) and *L’Innommable* (started in 1949, published in 1953; *The Unnamable*, published in 1958). These three novels have been published as a so-called ‘trilogy’.[[1]](#footnote-1)

Still, that unity did not prevent critics from categorizing the novels in separate periods. In Brian McHale’s periodization of Beckett’s works the border runs between *Molloy* and *Malone meurt* / *Malone Dies*. In *Constructing Postmodernism*, McHale indicated a difference between modernism and postmodernism that was marked by ‘the distinction between the cognitivist and the postcognitivist Beckett’ (McHale, 1992: 34). McHale thus implied a distinction between an early Beckett and a later Beckett, that is ‘the Beckett who is still preoccupied with modernist issues of reliability and unreliability of narrators, radical subjectivity, and multiplicity of perspectives, as in *Watt* and *Molloy*’ (*ibidem*: 34) and ‘the Beckett who focuses instead on the status of fictional worlds, the power (and impotence) of language to make and unmake worlds, and the relationship between fictional being and elusive “real” being, as in *Malone Dies*, *The Unnamable*, and many of the later short texts’ (*ibidem*: 34).

H. Porter Abbott situates the border after the *The Unnamable*: ‘*Watt*, *Mercier et Camier*, the *Nouvelles*, and the ‘trilogy’, all conform to the quest structure, despite the manifold incompetence of the questers.’ But the next work, *Textes pour rien* (started in 1951, published in 1955; *Texts for Nothing*, published in 1967), ‘marks a pause in the story of the œuvre’ (Abbott, 1996: 89). According to this periodization, it is the ‘willful shredding of narrative linearity within the *Texts*’ (90) that marks the end of a period, which coincides with the end of the so-called ‘great creative period’ (Federman & Fletcher, 1970: 63). The *Texts for Nothing* were followed by *Foirades* (the earliest of which was written in 1954; published in 1973; *Fizzles* published in 1976) and *From an Abandoned Work* (started in 1955, published in 1956; *D’un ouvrage abandonné* published in 1967).

*Comment c’est* (started in 1958, published in 1961; *How It Is*, published in 1964) is considered another ‘pivotal work’. The title has been interpreted as a ‘pun on beginnings (*comment c’est* is *commencer*)’ (Abbott, 1996: 102), and would in that sense mark the beginning of the late period. Again, the notion of a (failed) beginning was emphasized in the very short *Faux départs* (1964, published in 1965). The subsequent prose pieces were usually so short that – simply in quantitative terms – most of them can barely be used for stylometric analysis: *All Strange Away* (started in 1964, published in 1976); *Imagination morte imaginez* (written and published in 1965; *Imagination Dead Imagine*, published in 1965); *Assez* (written in 1965, published in 1966; *Enough*, published in 1967); *Bing* (written and published in 1966; *Ping*, published in 1967); *Sans* (written and published in 1969; *Lessness*, published in 1970); *Le Dépeupleur* (written and published in 1970; *The Lost Ones*, published in 1972); *Abandonné* (started in 1970, published in 1972); *Still* (started in 1972, published in 1975); *Sounds* (started in 1972, published in 1978); *Still 3* (started in 1973, published in 1978); *As the Story Was Told* (written and published in 1973); *La Falaise* (started in 1974, published in 1975); *Un Soir* (started in 1979, published in 1980; *One Evening*, published in 1980); *The Way* (written and published in 1981); Ceiling (written in 1981; published in 1985; *Plafond*, published in 1985). The more slightly longer prose texts are sometimes referred as the ‘second trilogy’ (although Beckett did not like this term) and were published both separately and together under the title *Nohow On*: *Company* (started in 1976, published in 1980; *Compagnie*, published in 1980); *Mal vu mal dit* (started in 1980, published in 1981; *Ill Seen Ill Said*, published in 1981) and *Worstward Ho* (started in 1981, published in 1983), followed by the short text *Stirrings Still* (started in 1983, published in 1989; *Soubresauts*, published in 1989).

The last text, *Comment dire* / *what is the word*, is again very short and usually treated as a poem, although it might just as well be regarded as a prose text – a piece about a sentence that does not manage to get written. It ends with the words ‘comment dire’, or ‘what is the word’ in Beckett’s translation. Beckett thus made his œuvre end in the middle of sentence – a poetical statement indicating his view on his work in terms of continuous incompletion. The Joycean notion of a ‘work in progress’ is not entirely suitable to denote this continuing incompletion as it implies the notion of improvement (‘progress’), whereas Beckett preferred to think in terms of ‘mere gress’ because of its ‘purity from destination and hence from schedule’ (186) (Beckett, 2009: 186).

In spite – and because – of Beckett’s aversion to ‘schedule’ and ‘destination’, scholars have tried to at least determine stages in his aesthetic ‘gress’. John Bolin (2013), for instance, focuses on the voices of the professor and the poet (the voice of ironic commentary versus a poetic voice that is open to the contingent) arguing that the latter is given greater prominence for the first time in the novel *Watt*. From this perspective, Beckett’s last novel in English before he switched to writing fiction in French thus becomes a pivotal work that closes the early period, separating it from his mature work, more or less in accordance with Pilling’s periodization. But other critics already discern aspects of the later work in the writings of the 1930s. For instance, Fintan O’Toole finds ‘the seed of Beckett’s mature work’ in passages from the unpublished story ‘Echo’s Bones’ (written in the Fall of 1933, published posthumously in 2014 in an annotated edition by Mark Nixon). Chiara Montini suggests yet another periodization: a four-phase division, based on Beckett’s use of different languages: (1) *le monolinguisme polyglotte* (1929-1937); (2) *le biblinguisme à dominance Anglophone* (1937-1945); (3) *le bilinguisme à dominance francophone* (1946-1953); (4) *le bilinguisme mixte* (Montini, 2008: 65).

The various opinions may lead us to a relativistic conclusion that periodization is merely a way of giving in to an urge to turn time into a plot, which Frank Kermode famously described in *The Sense of an Ending* as the ‘tick-tock’ model, an organization which humanizes time by giving it a form (Kermode, 1967: 45). But perhaps this tendency to categorize or periodize an œuvre is more than a subjective feeling. It is possible that it is based on stylistic developments that the human interpreter ‘senses’ but cannot always illustrate with empirical data. The hermeneutic approach thus resorts to explanations such as the ones described above, resulting in almost as many different periodizations as the number of critics that devise them. We therefore wondered what a non-human ‘interpreter’ would come up with as a periodization by means of stylometry.

**2. Stylometry and Stylochronometry**

Stylometry or computational stylistics is a multidisciplinary research domain that uses computational methods to study writing style (Holmes, 1998). To some extent, it can be considered a form of Artificial Reading, in which computers are used to artificially simulate and potentially enhance or complement human reading. Because of its heavy reliance on quantification, stylometry is commonly considered a part of the Digital Humanities or Humanities Computing (Schreibman & Siemens, 2008) and it is indeed in this domain that stylometry saw some of its seminal applications, such as the well-known study of authorship in the *Federalist Papers* by Mosteller and Wallace (1964) or John Burrows’s pioneering study of the speech of the different characters in Jane Austen’s novels (1987).

Stylometric research is typically concerned with the relationship between a text’s writing style and its metadata. In the popular field of authorship attribution (Love, 2002; Juola, 2008; Stamatatos, 2009; Koppel, 2009), for instance, researchers attempt to establish a quantitative link between an author’s identity and his/her unique writing style. In many cases, these methodologies have been able to attribute anonymous texts to a known author. Especially in the field of in the fields of ancient and medieval literature (which abound in anonymous texts), stylometry has generated a number of thought-provoking results, for instance, as to the attribution of disputed texts, such as the Shakespeare-Marlowe controversy (Burrows, 2012). Further research has focused on other forms of ‘author profiling’, such as the automatic detection of the age or gender of a text’s author (Daelemans, 2013). In literary studies, stylometry has moreover proved useful to arrive at a better understanding of the stylistic (dis)similarities between texts from different genres or periods (Jannidis & Lauer, 2014).

This study of periodization in Samuel Beckett’s works sets out from a slightly less studied subfield of stylometry, usually referred to as ‘stylochronometry’ (Stamou, 2008), in which a text’s writing style is studied as a function of its date of composition. For instance, research into medieval charters has shown that quantitative measurements allow us to date undated charters on the basis of stylistic features (Feuerverger *et al.*, 2008). Other recent research in stylochronometry has used diachronic models to study the development of the writing style in the œuvres of individual authors. A text book example is David Hoover’s recent analysis of Henry James’s œuvre (Hoover, 2014), in which Hoover was able to detect striking diachronic trends. In the field of Anglo-Saxon literature, other studies have reported similar findings for the oeuvres of W. B. Yeats (Forsyth, 1999) or Jack London (Juola, 2007).

Most stylometric research today focuses on stylistic features that can be automatically extracted from (digitized) texts. Whereas seminal studies focused on the analysis of rather superficial characteristics, such as average word or sentence length, a variety of more advanced stylistic features has been explored in the recent literature, including stylistic characteristics related to semantics, morpho-syntax, metaphors, intertextuality etc. (Stamatatos, 2009). In this article, we focus on a feature type that has proven surprisingly successful to model writing style across various languages, authors and genres: function words (Binongo, 2003; Rybicki & Eder, 2011).

Apart from the practical advantage that function words can normally be easily identified by computer algorithms there are also a number of theoretical considerations that help explain the attractiveness of function words for research in computational text analysis (Kestemont, 2014). Word frequencies in human languages follow a so-called “Zipfian” distribution: a few high-frequency words tend to occur very commonly in (not only literary) texts, whereas a much larger and diverse group of words occur much less frequently. The former group of high-frequency words is generally referred to as ‘function words’ by linguists: they cover a small set of (typically short) words, such as articles, prepositions, conjunctions or pronouns, which serve a predominantly grammatical function and do not carry any straightforward meaning when used in isolation, as opposed to, for instance, the much more straightforward semantics of most nouns or verbs.

Function words appear frequently throughout texts, which makes them an attractive and rich basis for quantitative research (Binongo, 2003). The fact that most texts written in a particular language will have to use the same set of function words – one can hardly write an English text without using *the* or *and* – makes them a reliable base unit for textual comparison. Empirical studies have shown that precisely function words capture a variety of stylistic choices, giving important clues as to an author’s gender, identity or sentiment (Stamatatos, 2009). Finally, research in authorship attribution suggests that authors use function words to a large extent *unconsciously* – this seems especially the case in longer prose texts (Pennebaker, 2011). This makes function words a refreshing object of research, because they lend themselves well to modeling stylistic aspects that are less deliberately constructed by authors. In Beckett studies, stylistic research has so far mainly focused on rhetoric (Clément, 1989) and self-translation (Sardin, 2002; Montini, 2007; Mooney, 2011). None of this research, however, focuses specifically on function words. The following section delineates the methodology by means of which we applied the study of function words to a stylochronometric analysis of Beckett’s works.

**3. Materials and Preprocessing**

In this study, we analyzed Beckett’s prose fiction, in both French and English. Naturally, the limitation to these (lengthier) prose writings does not do justice to Beckett’s achievements as a playwright, poet and essay writer, although future research might easily expand the scope of our investigations to include other genres. We obtained Beckett’s œuvre in digital form, mostly through purchasing the prose works in epub format and extracting the plain text. From this corpus, we removed all non-authorial paratexts, such as prefaces or page numbers. To further process these texts, we identified individual words in these materials using a standard tokenizer (Bird *et al.*, 2009).[[2]](#footnote-2) To maximize the comparability across texts, we have only considered lower-case, alphabetical character strings, ignoring for instance punctuation marks and digits.

Table 1 gives an overview of the materials we collected, including the publication dates of the editions we used,[[3]](#footnote-3) the text’s size (token counts after tokenization and preprocessing), as well as other metadata for each text. In terms of chronology, we distinguish between a text’s official publication date in French and/or English, as well as the fairly exact moment in time Beckett is known to have started working on a text in either language (Van Hulle & Verhulst, 2015). Finally, Table 1 lists a convenient abbreviation for each text (in each language), which we will use as a shorthand to identify texts in the graphs shown below. While international copyright law does not allow us to redistribute the original texts, all software code, necessary to replicate our findings, is freely available from the following software repository: (**link**). Note that this repository also holds high-resolution versions of the plots in this paper, which will be much more readable on screen.

[Insert Table 1 here: **Table of the analyzed corpus.** For each text (in each language), we list the original title, an abbreviated title, the language in which Beckett first composed a text, the text’s length (word count after tokenization), as well as the publication date of the first edition. The ‘start of composition’ column provides an indication as to when Beckett started writing a particular text (in either language). Some texts were originally published as part of a larger collection, which we identify in the ‘container text’ column. ‘NA’s indicate missing information, e.g. the French title of a text that was only written in English.]

As is common across a variety of studies in stylometry (Burrows, 2002; Hoover, 2003), we defined a relevant list of function words by extracting an initial list of the 300 most frequent words (MFW) from both the English and the French corpus. This number of MFW seems an acceptable choice in the light of previous studies (Stamatatos, 2009) and allows us to mine the most common function words from Beckett’s oeuvre. From this list, we have manually removed non-grammatical words, which might correlate too strongly with the topic of particular texts, instead of their writing style (Hoover, 2003). Removed items include the names of characters, numerals, common nouns or verbs etc. We refrained from removing common auxiliary verbs, such as inflected forms of *have*, because they are interestingly tied to a text’s narrative perspective, as will become clear below. Likewise, we have not removed any personal pronouns from these word lists (as is common in attribution studies) because we are especially interested in the stylistic shifts this category of words might capture. After this procedure, we were left with 162 function words for the English corpus and 169 for the French, the frequencies of which form the basis of our analysis (listed in Table 2).

[Table 2 (see *mfw\_en\_culled.txt* and *mfw\_fr\_culled.txt)*: **List of the most frequent words in the French and English corpus**, i.e. the function word vocabulary which forms the basis of all subsequent analyses in this paper. Words preceded by a hash tag were manually removed from the data to reduce the effect of, for instance, topic-related features which are less related to writing style.]

**4. Preliminary analyses**

In this section we describe a number of exploratory analyses of the material, using two established methods from stylometry to visualize the main variation in the data: hierarchical agglomerative clustering (HAC) and principal components analysis (PCA). Both techniques are unsupervised procedures, in the sense that they do not require any additional information other than the texts themselves to produce a result. We should therefore emphasize that, in this section, we do not yet explicitly integrate any chronological information in the analyses. This will help us establish to what extent Beckett’s œuvre might have a ‘natural’ chronological structure when it comes to the author’s writing style. Previous research in stylochronometry suggests that some œuvres indeed display an important diachronic structure, which can be detected without explicit supervision by scholars (Juola, 2007; Hoover, 2014).

The prose texts in the Beckett corpus differ significantly in length, ranging from novel-size works to the *Nouvelles* and shorter texts (consult Table 1). To compensate for this skewedness, we sliced the texts into consecutive, non-overlapping samples of a fixed size (expressed in token length), which is a common ‘windowing procedure’ in segmentation studies (e.g. Brooke *et al.*, 2012). We dropped trailing words at the end of texts, if they did not constitute an entire sample anymore. While longer texts, inevitably, still have a quantitative advantage (they contribute more samples), this procedure allowed us to study Beckett’s prose writings as a collection of equally sized text slices.

We produced a purely numerical representation of these slices of text, which can serve as the input for quantitative algorithms. Each text slice is represented as a list of real values (or ‘vector’), which will hold the relative frequencies of the items in the function word vocabulary we consider. We obtained a word’s relative frequency by dividing its absolute frequency in a slice by the slice’s total word length. In technical terms, we ‘vectorized’ these textual samples under a so-called bag-of-words assumption (Manning *et al.*, 2008): because we only collect the frequency of individual words – as if the text were a randomly jumbled ‘bag of words’ – this representation is agnostic of a term’s original position in a document, or even the original word order. This is a common and efficient vectorization strategy in e.g. Information Retrieval, which has only been shown to work well in literary analyses in stylometry (Stamatatos, 2009). After vectorization, the corpus can now be viewed as a two-dimensional matrix in which the number of rows will be equal to the number of slices and in which the number of columns will amount to the number of most frequent words we consider. Each cell holds the relative frequency of a single function word in a particular text slice. The advantage of such a numeric text representation is that texts can easily be studied in a quantitative fashion, or even plotted as points in a geometric space.[[4]](#footnote-4)

To get a first glance at the internal structure of Beckett’s prose writings, we now turn to a Hierarchical Agglomerative Clustering (HAC). As its name suggests, this technique from multivariate statistics attempts to establish a cluster hierarchy of the slices in our corpus, which will hopefully produce a meaningful first division of Beckett’s works. The result of a HAC is typically graphically represented using a ‘dendrogram’ or tree representation (see Figs. 1-2). As opposed to divisive clustering, agglomerative clustering is a ‘bottom-up’ procedure, which will start at the outer twigs of the tree, i.e. the original text slices (Hoover, 2003). The cluster analysis sets out from a square distance table, which indicates the Euclidean distances between all the slices in our corpus. Based on these distances, the HAC algorithm will start merging the most similar text slices into new nodes at the next level in the tree.[[5]](#footnote-5) Iteratively, these newly created twigs will in turn be merged into higher-level nodes, until all nodes have eventually been joined in a single root node or ‘stem’.

The result of this operation can now be visualized as a dendrogram, such as the one in Figure 1: in this figure, the distance between the pair of items joined under a particular node is reflected in the branch length of branches. This form of tree visualization is especially common in phylogenetic studies, where such trees are used, for instance, to detect ‘clades’ of closely related species. Increasingly, such representations are also common in stemmatological manuscript studies (e.g. Van Reenen *et al.*, 2004). Each slice in this tree is annotated by a label, which consists of the initial characters of the work’s title followed by an index (e.g. ‘mercier\_6’ represents the sixth slice we extracted from *Mercier and Camier*).

[Insert Figures 1 and 2 here. **Dendrograms of the English and French prose works.** These dendrograms are based on the Euclidean distance between the MFW vectors obtained for equal-length slices from Beckett’s œuvre. We used agglomerative clustering and Ward linkage. The smaller version of the plot uses coloring to highlight the three main branches detected.]

In Figures 1 and 2 (for the English and French corpora respectively) we have plotted the dendrograms for the function word vocabulary described above. For this analysis, we tested samples of several sizes (1000 words, 2500 words, 5000 words, etc.). Even though a larger sample size might imply the exclusion of a few very short texts in the visualization, the trends we observed were similar in all cases. For the visualization (see Figures 1 and 2) we use a 3761-word sample size in order not to clutter the image, while still allowing samples of *From an Abandoned Work* (3761 words) and *Worstward Ho* to be included in the tree. The top branches in this tree inform us as to the main ‘natural’ divisions this procedure has detected among the slices.

For the English data, 3 natural ‘branches’ clearly emerge at the base of the tree. Under a first branch, we find a cluster of texts that is dominated by samples from Beckett’s earlier works, such as *Murphy, Dream of Fair to Middling Women* and *More Pricks Than Kicks*. Next, we observe a second cluster, dominated by samples from *The Unnamable*, but also includes samples from *Texts for Nothing* and as well as the single sample from *Worstward Ho* and two from *Malone Dies*. The rest of samples are collected in the third branch, and primarily belong to Beckett’s middle period, including the majority of samples from works such as *Malone Dies, Molloy, Watt* or *Mercier and Camier*. Interestingly, this third cluster also has a clade that links the *Nouvelles* to a branch of *Molloy* samples.

For the parallel analysis of the French corpus, we can make similar observations. A first branch of French prose is dominated by samples from *Molloy* and *La Fin*, clustered together with the other *Nouvelles*, such as *Premier amour*. The second cluster is dominated by samples from *L’Innommable* and *Textes pour rien*, paired at a higher level with *Comment c’est,* which forms a relatively tight clade of its own. The third cluster, finally, primarily contains early Beckett, such as *Murphy* and *Watt*, but also *Malone meurt*, the shorter *Foirades* and *Mal vu mal dit* – like their counterparts in the English tree. In both languages, we see how a number of texts resist a hard classification into one of three clusters: most notably, *Watt* and *Malone meurt* are often scattered across different branches.

Overall, it is clear that slices extracted from the same texts have a tendency to be merged together at an early stage in the tree, often before being joined with samples from other works – which attests to the stylistic cohesion inside works. Figures 1 and 2 already suggest that the use of function words in Beckett’s writings produces a clustering, which might roughly correspond to some of the periodization schemes suggested in the traditional literature. Interestingly, however, we see that this initial, rough division into ‘three Becketts’ is far from obvious: we see how samples from *Watt* for instance, can occur at fairly different positions in the tree. Additionally, it is worrying that the relatively larger sample size we selected, leaves out several of Beckett’s shorter writings from the later period, simply because they are too short to fit the sample size. This highlights the need – which we will address in the next section – to introduce analyses that can explicitly take into account the chronological structure of both corpora, also for smaller sample sizes.

To arrive at a better understanding of how specific MFWs underpin the results of this, we now turn to a Principal Components Analysis of the material. PCA, too, is an unsupervised procedure, which we can use to study clusters of samples in Beckett’s writings (Binongo & Smith, 1999). In stylometry, it is common practice to produce a scatterplot of the original samples as points in a two-dimensional scatterplot, which can be said to capture the main stylistic variation in a data set (see Figures 3-4). As an aid to interpretation, we have performed a cluster analysis on top of the samples’ position in the first components in order to identify the four primary clusters of samples in the scatterplots; we have colored the samples accordingly.[[6]](#footnote-6)

While PCA provides us with a slightly less detailed visualization of the material, the technique has the considerable advantage that the samples in the plot can be overlaid with the so-called ‘loadings’ of the original word features (in a smaller, grey-colored font). These loadings offer an indication of how the original vocabulary contributed to the structure that emerges from this plot. Samples plotted in a similar region of the scatterplot can be said to be more stylistically similar. The overlaid loadings of the individual word features (in grey) reveal how individual words contribute to the position of samples in the scatterplot. For instance, word loadings plotted to the far left are typical of samples that can be found in the same area.

[Insert Figures 3-4 here: **PCA components analysis of the (culled) most frequent words** selected above, for the English and French corpus respectively. We plot the position of 3,761-word slices; label names indicate the work from which a sample was extracted. Word loadings have been overlaid in grey. A similar temporal structure is evident for both the English and French data, especially in the horizontal dimension.]

From the resulting scatterplots (see Figures 3-4), we can draw a number of interesting observations that partially mirror the results of the HAC. For the English data, we see how the horizontal spread is dominated by a threefold clustering, with some of Beckett’s earlier works clustering in the far left. Interestingly, the loadings reveal that these works are characterized by a relatively elevated frequency of words related to a third-person narrative perspective (*he*, *she*, *his*, *has*, etc.). To the far right, we find a relatively tight sample cloud corresponding to some of Beckett’s post-war works, such as *The Unnamable* and *Texts for Nothing*. According to the loadings, these texts can be characterized by the use of first-person pronouns (*I*, *me*) in combination with impersonal pronouns such as *it* and *there*, which suggests that these texts focus on the relation between an authorial ‘I’ and his non-personal surroundings. We see how texts from the in-between period (such as *Molloy* and *Watt*) also hold the middle in the horizontal distribution of samples. Interestingly, the visualization shows that a single sample jumps out with respect to the vertical dimension of the plot: *Worstward Ho* (and to a lesser extent, *How it is*) is characterized by what seems, at first sight, a rich mix of fairly abstract quantifiers and determiners, often with an indeterminate semantics ((*n)ever*, *no/yes,* …).

As to the French scatterplot in Figure 4, is it surprising how similar the structure is that arises from the French prose, albeit in a horizontally inversed version. To the far right, we find some of the earlier works (*Watt*, *Murphy*), which are also characterized by a dominant third-person perspective (cf. pronouns *ses*, *sa*, *son*, …). These works are placed in opposition to the sample cloud containing the *L’Innommable* samples to the far left, in which the use of the first-person perspective is striking (*j’*, *je*, *me*, *suis*, *moi*, etc.), again in combination with the impersonal *c’est*. For the French data, too, we find that the vertical variation focuses on displaying aspects of Beckett’s later writings, most notably *Comment c’est*, and to a lesser extent *Foirades* and *Mal vu, mal dit*. Here too, the PCA reveals a shift in Beckett’s vocabulary to a set of indeterminate adverbs and articles, including abstract oppositions such as *toujours* vs *jamais* or *oui* vs *non*. Thus, interestingly, both scatterplots horizontally create an opposition between earlier and later writings in Beckett’s oeuvre, focusing on an opposition between a first-person and third-person perspective. With respect to the vertical dimension, both analyses show a tendency to focus on a vocabulary shift, which seems to occur in Beckett’s later, shorter writings, towards a more abstract and indeterminate vocabulary, dominated by adverbs and particles with a very bleak and abstract semantics.

**5. Chronological analyses**

In the previous section, we have shown how exploratory statistical techniques can be used to analyze the stylistic structure of Beckett’s prose fiction. However, these analyses were agnostic of the diachronic structure of the data, meaning that samples from Beckett’s early works could in principle just as easily cluster with his later writings. This prevented us so far from identifying clear turning points in Beckett’s career. Therefore, we now move to quantitative methods that are able to take into account the actual chronology of Beckett’s oeuvre.

Variability-Based Nearest Neighbour Clustering (VNC) is a clustering technique which has recently been introduced by Gries & Hilpert (2012) in the context of the diachronic analysis of corpus linguistic data. For this research, Gries & Hilpert considered corpora with an important diachronic component, such as child language acquisition data. Their aim was to identify distinct temporal stages, by pinpointing the main turning points. Gries & Hilpert proposed a simple adaptation of existing clustering techniques. In traditional cluster analyses, each node can be freely combined with any other node in the tree, thus potentially scrambling the original chronological order of the data. In their VNC approach, Gries and Hilpert add the constraint that only consecutive nodes, which are immediately adjacent in time, can form new clusters. This restriction allows analyses in which the chronological structure of the data is reflected in the top branches of trees, representing the main diachronic stages in the data.[[7]](#footnote-7)

We have run a VNC on our data, using the same experimental settings as for Figures 3-4. The resulting trees offer a much clearer insight into the chronological structure of Beckett’s œuvre. Figure 5, for the English prose, displays a clear initial cluster of Beckett’s earliest two novels, *More Pricks than Kicks* and *Dream of Fair to Middling Women*. This initial cluster is of course absent from the French prose’s VNC dendrogram in Figure 6, since these texts lack a translated counterpart.[[8]](#footnote-8) Otherwise, the structures of Figures 5 and 6 run remarkably parallel. *Murphy* and *Watt* constitute the second chronological cluster of works. Only at a higher level are these texts paired with a cluster consisting of the *Nouvelles*, *Mercier and Camier*, *Molloy* and finally *Malone Dies*. The last major branch for both languages holds the tight clade representing *The Unnamable*, *Texts for Nothing*, *How It Is*, and the shorter series of late works. In the English tree, it is worth noting that *Worstward Ho* again occupies a fairly pronounced position, emphasizing again the special status of this work.

[Insert Figures 5-6 here: **A Variability-Based Neighbour Analysis of Beckett’s prose writings (French and English)**. These dendrograms offer an analysis that improves upon the analyses in Figures 1-2 by adding the constraint that only adjacent nodes can be merged to form a new node. The integration of this chronological knowledge yields a much more insightful result.]

The results of the VNC analysis bring us closer to answering our original research question. But although the VNC analysis offers a visualization, it does not yet identify the exact turning points we are after. To achieve this, we took the VNC one step further, and instead of visualizing a tree we made the VNC merely output the *n* data samples it considers the most significant turning points. Apart from the ordering of the original data points, this procedure only required us to predetermine the number of diachronic stages or ‘segments’ we would like to obtain; e.g. for the setting *n*=3, we would obtain three diachronic stages, which will of course hopefully correspond to an early, middle and late Beckett. We derived this segmentation approach from parallel studies in audio research, where efficient libraries exist to divide audio fragments such as songs into a number of consecutive segments (McFee *et al.*, 2015). In Table 3 we list the results obtained for various values of *n*.

[Insert Table 3 here: **Segmentation of Beckett’s French and English writings through time**. This analysis considers all MFW listed above, and offers the results for various values of *n* (2 <= *n* <= 5), i.e. the parameter which controls how many diachronic segments we wish to obtain).]

|  |  |  |
| --- | --- | --- |
| ***n* (number of segments)** | **turning points English prose** | **turning points French prose** |
| **2** | calmative\_1 | amour\_1 |
| **3** | calmative\_1, unnamable\_1 | amour\_1, abandonne\_1 |
| **4** | calmative\_1, unnamable\_1, ned\_1 | amour\_1, innommable\_1, abandonne\_1 |
| **5** | calmative\_1, unnamable\_1, abandoned\_1, ho\_1 | amour\_1, innommable\_1, abandonne\_1, comment\_1 |

When asked to identify a single turning point in Beckett’s oeuvre, the analysis for both languages singles out a sample from the *Nouvelles*, when considering the entire MFW vocabulary. Interestingly, the English analysis opts for the first sample of *The Calmative*, which is only the second *Nouvelle*, indicating a slight difference in stylistic status for *Premier amour* / *First Love* in both languages. Our analyses suggest that the English version of *Premier amour* / *First Love* behaves more like Beckett’s early prose in English than in French, which is remarkable, since *Premier amour* was written in 1946, whereas the English translation was made more than two decades later, in the early 1970s. At first sight, this result is largely in line with previous scholarship which had the start of the ‘middle Beckett’ coincide with the *Nouvelles*, even though these periodizations are based on different grounds. If we zoom in, however, there stylometric analyses do show remarkable stylistic oddities that nuance the rough-and-ready, 3-phase periodization.

The analysis is also interesting with reference to the French *Nouvelles*. On the one hand, this stylistic pattern is surprising, since these four short texts were written at roughly the same time and *Premier Amour* (*First Love*) was not written first (the chronological order of composition was *La Fin*, *L’Expulsé*, *Premier Amour*, *Le Calmant*). On the other hand, it is remarkable that Beckett seems to have sensed a difference between the cluster of three and *Premier Amour*, for he chose not to publish *Premier Amour* together with the other three. He only published it in 1970, almost 25 years later, shortly after he received the Nobel Prize, when publishers urged him to look in his drawers for publishable texts.

When asked to choose two additional turning points, both analyses identify the start of *The Unnamable* and *From an Abandoned Work*, however not in the same order. *L’Innommable* / *The Unnamable* may not be the least surprising turning point, as most of the existing periodizations mention it as a pivotal work. But usually it is the end of this work that is pinpointed as the watershed. This periodization is partly inspired by the novel’s famous closing line ‘I can’t go on, I’ll go on’, marking both an end (‘I can’t go on’ in the present tense) and a beginning (‘I’ll go on’ in the future tense). These semantic connotations do not influence the stylometric periodization, which interestingly marks the *beginning* of *L’Innommable* / *The Unnamable* as a turning point, thereby breaking up the ‘unit’ of the so-called ‘trilogy’. Moreover, in combination with the preceding and following turning points, an interesting pattern emerges: Beckett’s career as a writer of prose fiction spans a period of more than fifty years; in this half century, the three major stylistic turning points occur within a comparatively limited interval of only a decade (1946-1956).

Finally, when identifying a fourth turning point, the analysis singles out *Comment c’est* from the French prose samples and *Worstward Ho* from the English series. Especially for the English corpus, this is a remarkable result, given the limited length of Beckett’s final prose work, stressing the exceptional nature of *Worstward Ho*.

To examine how robust this segmentation is, we have complemented this analysis with a bootstrap analysis. Previous research in stylometry has noted that some clustering techniques are highly dependent on the exact number of function words considered in an analysis. Eder (2013) showed that even small differences in the MFWs under consideration, could sometimes introduce fairly large differences in the cluster results. Eder proposed an iterative bootstrap procedure borrowed from phylogenetics. In each iteration, the clustering would be restricted to a random subset of all MFW available, thus each time producing a slightly different result. At the end of these iterations, a consensus would be extracted from all results. For our segmentation analysis, such a procedure has the advantage that we can test whether the proposed segmentation is not too strongly dependent on a specific subset of individual MFW, which might coincidentally differ strongly across works.

We therefore applied a similar bootstrapping to our segmentation procedure from the previous section: we repeated the segmentation procedure a fixed number of times (1,000 iterations), using a randomly select subset (50%) of our MFW frequencies. In Figures 7-8, we have reproduced bar plots which represent the outcome of this bootstrapping experiment for both the English and French corpus for *n*=3 segments (Figure 7, i.e. so that we can target an ‘early’, ‘middle’ and ‘late’ Beckett) and *n*=4 (Figure 8). The colored bars indicate how often a particular slice (absolute counts) was chosen as a particular segmentation boundary. This analysis serves to test the robustness of the segmentation reported in Table 3.

For the French data, the previous results appear extremely stable across different sets of MFW, since again, the first of samples of *Premier Amour*, *D’un ouvrage abandonné* and *L’Innommable* are almost consistently singled out as the main turning points in Beckett’s French prose. For the English data, the same is true for *The Unnamable* and *From an Abandoned Work*, but interestingly, the first boundary seems much more unstable in the English prose: when varying the set of MFW, the *Nouvelles* clearly compete with samples from *Watt*, which appear to constitute a slightly more stable boundary across the bootstrapping.

[Insert Figures 7a-b and 8a-b here: **Bar plot visualizations of bootstrapped segmentation analysis of the English and French corpus**. We have repeated the segmentation experiment reported in Table 3 1,000 times; each time the segmentation was carried out on a randomly impaired subset of the original MFW feature space (for 4 diachronic segments). The colored bars indicate how often a particular slice (absolute counts) was chosen as a particular segmentation boundary. The number of colors depends on the number of segmentations requested. Only the first sample’s name is plotted for each work on the vertical axis, to unclutter the bar plots.]

**6. Discussion**

In the course of our experiments, a number of interesting trends have emerged, which might offer a valuable framework to re-inspect some of the periodization issues discussed in Beckett scholarship. The VNC analysis generally supported the general periodization of Beckett’s oeuvre into an early, middle and late ‘cluster’. In English these periods would cover, firstly, Beckett’s early works, *More Pricks than Kicks* and *Dream of Fair to Middling Women*; secondly the mid-career works, ranging from *Murphy* to *Malone dies*; and thirdly, a series of later works starting with *The Unnamable*. From the French prose, a similar periodization arises (although it clearly reflects the absence of any translated counterparts for *More Pricks than Kicks* and *Dream of Fair to Middling Women*). Here, we could distinguish a first period roughly coinciding with the French versions of *Murphy* and *Watt*, a middle period starting with the *Nouvelles* and, finally, the late works starting with *L’Innommable*. Interestingly, the *Nouvelles* also seemed to coincide with a minor turning point in the second period in the English prose.

The initial periodization roughly coincides with some of the periodization schemes proposed in the earlier literature, albeit on completely different grounds. An important difference, however, is that our analyses invariably point to the beginning of *L’Innommable* / *The Unnamable* as a major stylistic turning point, thus breaking up the unity of the so-called post-war ‘trilogy’. Indeed, in terms of function word frequencies, *L’Innommable* / *The Unnamable* fits in better with Beckett’s later works than the first two novels of the trilogy, which themselves appear to be more similar to the *Nouvelles*.

Apart from the VNC analyses, we have also attempted to obtain a “hard” diachronic segmentation of Beckett’s oeuvre. For the French corpus, this segmentation yielded stable results, with the start of *Premier amour* and *L’Innommable* being singled out as the most significant fault lines. Additionally, *D’un ouvrage abandonné* / *From an abandoned work* – notwithstanding its short length – also often emerged as a major watershed in both languages, leading the way for Beckett’s later, experimental works. This might be a valuable pointer for further research, since this particular text does not seem to have played a significant role in the periodization debate so far. In fact, this work is arguably a blind spot in Beckett studies. John Pilling calls it one of Beckett’s oddest ‘odds’ and refers to the critical literature on this work as ‘what little there is’ (Pilling, 2007: 173). The only existing periodization that takes this work into account is Chiara Montini’s, whose fourth phase (marked by mixed bilingualism) begins with *From an Abandoned Work*.

Another remarkable result is that *Worstward Ho*’s unparalleled style is also reflected in many of our results. As a work that was written in the last decade of Beckett’s career, it constitutes no less than a fourth turning point, which indicates that, even when he was in his late seventies, Beckett still kept reinventing himself stylistically.

Interestingly, the segmentation approach for the English corpus, also singled out *The Unnamable* and *From an Abandoned Work* as fault lines, but had a harder time identifying a clear break between the early and middle period: both *Watt*, to a lesser extent *Murphy*, and especially some of the *Nouvelles* seem to compete in this respect. As such, our particular methodology finds it surprisingly more difficult in English than in French to model the transition from a young to a middle Beckett. This result possibly reflects the fact that the original, English version of Watt was written relatively early (during the war), whereas its translation was made much later (in cooperation with Agnès and Ludovic Janvier) and was not published in French until 1968. In any case, this particular result offer grounds for a re-examination of the difference in evolution between Beckett’s French and English prose production, and in particular the role of *Watt* as a transitional novel.

**7. Conclusion and perspectives**

In this paper, we reported the results of a quantitative analysis of the stylistic evolution of Samuel Beckett’s prose writings. The main novelty of this work lies in the fact that it is restricted to the frequencies of the most frequent function words, an inconspicuous category of words, which has not yet been closely analyzed in Beckett scholarship. Indeed, as pioneering researcher John Burrows once said: ‘It is a truth not generally acknowledged that, in most discussions of works of English fiction, we proceed as if a third, two-fifths, a half of our material were not really there’ (1987: 1). The advantage of taking this neglected category into account is that it allows us to nuance and adjust existing periodizations. In this sense, the qualitative analysis of the existing periodizations and the quantitative method proposed in this article do not overlap, but complement each other. Especially the clear watershed of the *beginning*, rather than the ending, of *L’Innommable* / *The Unnamable* is notable. The discovery of the pivotal importance of *From an Abandoned* *Work* calls for a reappraisal of this underrated work. But also the more fuzzy result regarding the first turning point is important as it indicates that the decisive moment that separates the early work from the mature work is not unequivocally linked to the transition from writing in English (*Watt*) to writing prose fiction in French (the *Nouvelles*). The stylochronometric analyses thus help us reassess certain critical commonplaces and prompt us to keep questioning not just our periodizations, but also our urge to periodize *tout court*. Kermode’s ‘tick-tock’ model, mentioned above, not only applies to plots in fiction but also to critics’ attempts at periodizing an author’s œuvre. By trying to humanize time and make sense of an œuvre, we also turn it into a plot. Especially in Beckett’s case, the dangerous neatness of such a narrative with a beginning, a middle and an end tends to obscure the continuous incompletion of his work. That does not mean we should stop periodizing; if anything, it is an invitation to keep trying and failing. More periodizations imply more reassessments. The more nuanced the plots become, the better they may cope with the plotless continuum of his work’s ‘mere gress’.

**Bibliography**

Abbott, H. P., 1996. Beckett Writing Beckett: The Author in the Autograph. Cornell.

Bair, D., 1978. Samuel Beckett: A Biography. Jonathan Cape.

Beckett, S., 2009. The Letters of Samuel Beckett, vol. I, 1929-1940, Martha Dow Fehsenfeld and Lois More Overbeck (eds.). Cambridge University Press.

Binongo, J.N., 2003. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance 16 (2), 9-17.

Binongo, J.N. & Smith, W., 1999. The application of principal components analysis to stylometry. Lit. Linguist. Computing 14 (4), 445-466.

Bird, S., Klein, E. & Loper, E., 2009. Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit. O’Reilly.

Bolin, J., 2013. Beckett and the Modern Novel. Cambridge University Press.

Boxall, P., 2015. Still Stirrings: Beckett’s Prose from Texts for Nothing to Stirrings Still. In: Van Hulle, D. (ed.), The New Cambridge Companion to Samuel Beckett. Cambridge University Press, 33-47.

Brooke, J., Hammond, A. & Hirst, G., 2012. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. In: Elson, D., Kazantseva, A., Mihalcea, R. & Szpakowicz, S. (eds.), Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature. Association for Computational Linguistics, 26-35.

Burrows, J.F., 1987. Computation into Criticism: A Study of Jane Austen’s Novels. Oxford University Press.

Burrows, J.F., 2002. ‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship. Lit. Linguist. Computing 17 (3), 267-287.

Burrows, J.F., 2012. A Second Opinion on “Shakespeare and Authorship Studies in the Twenty-First Century”. Shakespeare Quarterly 63 (3), 355-392.

Byrne, P., 2015. ‘Young Jane’. Times Literary Supplement, 20 February 2015, 9-10.

Daelemans, W., 2013. Explanation in Computational Stylometry. In: Gelbukh, A. (ed.), 2013. Computational Linguistics and Intelligent Text Processing, 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II. Springer 2013 Lecture Notes in Computer Science, 451-462, 2013.

Eder, M., 2013. Computational stylistics and Biblical translation: how reliable can a dendrogram be? In: Piotrowski, T. & Grabowski, Ł. (eds.), The translator and the computer. WSF Press, 155-170.

Federman, R. & Fletcher, J., 1970. Samuel Beckett: His Works and His Critics. University of California Press.

Feuerverger, A., Hall, P., Tilahun, G. & Gervers, M., 2008). Using statistical smoothing to date medieval manuscripts. In: Balakrishnan, N., Pena, E. & Silvapulle, M.J. (eds.), Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen. Inst. Math. Stat. Collect. 1, 321-331.

Forsyth, R.S., 1999. Stylochronometry with substrings, or: a poet yound and old. Lit. Linguist. Computing 14 (4), 467-478.

Gries, S.T. & Hilpert, M., 2012. Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In: Nevalainen, T. & Traugott, E.C. (eds.), The Oxford Handbook of the History of English. Oxford University Press, 134-144.

Holmes, D.I., 1998. The Evolution of Stylometry in Humanities Scholarship. Lit. Linguist. Computing 13 (3), 111-117.

Hoover, D.L., 2003. Multivariate Analysis and the Study of Style Variation. Lit. Linguist. Computing 18 (4), 341-360.

Hoover, D.L., 2014. A Conversation Among Himselves. Change and the Styles of Henry James. In: Hoover, D.L., Culpeper, J. & O’Halloran, K. (eds.), Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama. Routledge, 90-119.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Computing in Science and Engineering 9 (3), 90-95.

Jannidis, F. & Lauer, G., 2014. Burrows’s Delta and Its Use in German Literary History. In: Erlin, M. & Tatlock, L., Distant Readings. Topologies of German Culture in the Long Nineteenth Century, Camden House, 29-54.

Juola, P., 2007. Becoming Jack London. Journal of Quantitative Linguistics 14 (2), 145-147.

Juola, P., 2008. Authorship attribution. Foundations and Trends in Information Retrieval 1 (3), 233-334.

Kermode, F., 1967. The Sense of an Ending: Studies in the Theory of Fiction. Oxford University Press.

Kestemont, M., 2014. Function Words in Authorship Attribution. From Black Magic to Theory? In: Feldman, A., Kazantseva, A. & Szpakowicz, S. (eds.), Proceedings of the Third Computational Linguistics for Literature Workshop, co-located with EACL 2014. Association for Computational Linguistics, 59-66.

Knowlson, J., 1996. Damned to Fame: The Life of Samuel Beckett. Bloomsbury.

Koppel, M., Schler, J., & Argamon, S., 2008. Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology, 60 (1), 9-26.

Love, H., 2002. Attributing Authorship: An Introduction. Cambridge University Press.

Manning, C.D., Raghavan, P. & Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press.

McFee, B. *et al.*, 2015. librosa: Audio and Music Signal Analysis in Python. In: Proceedings of the 14th Python in Science Conference (Scipy 2015).

McHale, B., 1992. Constructing Postmodernism. Routledge.

Mooney, S., 2011. A Tongue Not Mine: Beckett and Translation. Oxford University Press.

Montini, C., 2007. La Bataille du soliloque: Genèse de la poétique bilingue de Samuel Beckett. Rodopi.

Montini, C., 2008. Sinking in the Mud: From an Abandoned Work et le difficile retour à l’anglais. Samuel Beckett Today / Aujourd’hui 20, 63-9.

Mosteller, F. & Wallace, D.L., 1964. Inference and Disputed Authorship: The Federalist. Addison-Wesley.

O’Toole, F., 2015. Samuel Beckett: The Private Voice. The New York Review of Books, 19 March 2015, 34-6.

Pedregosa, F., *et al.*, 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830.

Pennebaker, J., 2011. The Secret Life of Pronouns: What our Words Say about us. Bloomsbury.

Pilling, J., 2007. From an Abandoned Work: All the Variants of the One. Samuel Beckett Today / Aujourd’hui 18, 173-83.

Pilling, J., 2015. Early Beckett: The One Looking Through His Fingers. In: Van Hulle, D. (ed.), The New Cambridge Companion to Samuel Beckett. Cambridge: Cambridge University Press, 3-18.

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

Sardin, P., 2002. Samuel Beckett auto-traducteur ou l’art de ‘l’empêchement’: Lecture bilingue et génétique des textes courts auto-traduits (1946-1980). Artois Presses Université.

Schreibman, S. & Siemens, R. (eds.), 2004. A Companion to Digital Literary Studies. Blackwell.

Stamatatos, E., 2009. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60 (3), 538-556.

Stamou, C., 2008. Stylochronometry: stylistic development, sequence of composition, and relative dating. Lit. Linguist. Computing 23 (2), 181-199.

Tucker, D., 2014. Introduction: ‘No One Wander Unpunished’: Revisiting Molloy, Malone meurt / Malone Dies and L’Innommable / The Unnamable. Samuel Beckett Today / Aujourd’hui 27, 11-24.

Van Hulle, D. & Weller, S., 2014. The Making of Samuel Beckett’s L’Innommable / The Unnamable. Bloomsbury Academic.

Van Reenen, P., Den Hollander, A. & Van Mulken, M. (eds.), 2004. Studies in Stemmatology II. John Benjamins.

1. The term ‘trilogy’ was used for the 1959 Olympia Press edition of the three novels *Molloy, Malone Dies* and *The Unnamable* (Van Hulle & Weller, 2014: 81). Beckett however expressed his dislike of this term at several occasions, for instance in letters to John Calder, Aidan Higgins and Barbara Bray (Tucker, 2014: 23). [↑](#footnote-ref-1)
2. For the French texts, this tokenization strategy introduces a number of minor issues (e.g. *aujourd’hui* will be considered as two tokens (*aujourd* and *hui*). Because we limit analyses to function words, however, this has not proven to impose a major problem. [↑](#footnote-ref-2)
3. For the French texts, we used the publications by *Les Éditions de Minuit*; for the English the most publications by Faber & Faber (2009-2010). [↑](#footnote-ref-3)
4. Additionally, we scale each column to lie between 0 and 1, which is a recommended practice (Eder, 2013). [↑](#footnote-ref-4)
5. We use Ward’s linkage to construct our cluster hierarchy, a common metric (Eder, 2013) that continually monitors the overall reduction in variance in a tree when selecting which nodes to merge next. [↑](#footnote-ref-5)
6. Like in Figures 1-2, we again used HAC with Ward linkage and Euclidean distances. [↑](#footnote-ref-6)
7. For these analyses, we rely on an implementation by Folgert Karsdorp, available at: https://github.com/fbkarsdorp/HACluster. [↑](#footnote-ref-7)
8. They were not translated by the author himself. [↑](#footnote-ref-8)