**Hogwarts High**

**Using computer-assisted collation to compare the UK and US versions of Harry Potter**

**Fluid works**

In the previous chapter, we discussed the notion of the ‘text’. In this chapter, we need to briefly reflect on the concept of the ‘work’ before we embark on the more technical aspects of comparing textual versions with the aid of digital collation tools. The ‘work’ is not some kind of Platonic ideal; nor is it the *sum* of all its different versions. In ‘A Practical Theory of Versions’, Jack Stillinger did argue somewhat tautologically that ‘A *work* is constituted by all known versions of the *work*’ (Stillinger 2015: 171), but this does not mean that all known versions need to be added up to get the work as their sum total. Instead, the work is *implied by* all these versions (Shillingsburg 1996: 176).

In *The Fluid Text*, John Bryant defined the work as ‘a flow of energy’ (2002: 61). Although this is the most inclusive of definitions, it does emerge from a profound critique of previous editorial theories, most of which reduced the work to ‘a product or set of products’ (61); instead, he sees the work as a process. Bryant sees textual fluidity as ‘the inherent condition of any written document’: ‘writing is fundamentally an arbitrary hence unstable hence variable approximation of thought. Moreover, we revise words to make them more closely approximate our thoughts, which in turn evolve as we write. And this condition and phenomenon of textual fluidity is not a theoretical supposition; it is fact’ (Bryant 2002: 1).

Bryant, however, was not the first to introduce this notion of the work as process. French genetic critics had been employing the metaphor of fluidity and insisting on the inclusion of the temporal dimension since the 1960s.[[1]](#footnote-1) Still, Bryant preferred to distinguish his theory from genetic criticism – in spite of the manifest affinities – because he saw genetic criticism as focusing too narrowly on the writing process, whereas his ‘fluid text theory’ encompasses more than the genesis *stricto sensu*. Even though genetic criticism also acknowledges that the genesis continues after publication, which is referred to as a work’s ‘epigenesis’ (Van Hulle 2007), Bryant insisted on the need for an even broader definition of the work, one that ‘embraces all versions of a text, including sources and adaptations’ and ‘all forms of revision, both authorial and cultural’ (2013: 47).

This is a bold statement. In the case of *Harry Potter*, this would imply that all the sources of inspiration for J. K. Rowling and all the adaptations of the novels, including the movie and stage adaptations, as well as all the subsequent fanfiction, would be part of the ‘work’. The fluid text encompasses what a culture invests in a work and takes into account the ‘adaptations of texts beyond the writer’s authority or control’ (Bryant 2002: 62). In other words, ‘a poetics of the fluid text is a poetics of revision, whether that change is induced by an individual writer, a social demand, or as is often the case, a combination of the two’ (62). Adaptation theorists such as Linda Hutcheon have built on Bryant’s idea of ‘reception-generated changes’ to suggest a ‘reception continuum’ (Hutcheon 2013: 171). The film versions of the *Harry Potter* books certainly have their place on this reception continuum. But there is a difference between seeing adaptations in terms of a continuum and including them in the notion of the ‘work’. J.K. Rowling will not be the only person objecting to the idea that pornographic *Harry Potter* fanfiction would be considered part of her work. Nonetheless, even pornographic *Harry Potter* fanfiction has its place on the reception continuum. The notion of ‘process’ is clearly a more accepted notion in 21st-century criticism and textual scholarship than it was in the previous century, but the question to what extent this process coincides with the ‘work’ is a matter of debate.

In 2009, Sally Bushell also emphasized the notion of process, but instead of applying it to the ‘work’ she rather emphasized the notion of ‘*text* as process’. Ever since structuralist and post-structuralist thinkers such as Roland Barthes criticized the notion of ‘work’ in favour of ‘text’, the ‘work’ was regarded as a ‘manifestly relegated concept’ (Sutherland 1996: 16). According to Barthes the work was ‘caught up in a process of filiation’: ‘The author is regarded as the father and the owner of his work.’ (Barthes 1979: 78). For Barthes, the ‘text’ was the alternative ‘without the father’s signature’ (78). From a 21st-century perspective, the gendered nature of this statement strikes us as remarkably dated, especially in the case of J.K. Rowling – who famously used her initials, rather than full first name, to avoid being identified as a female writer and potentially scare off readers of different genders. Post-structuralist theory had quite an impact on literary criticism and textual scholarship, so it took a few decades before the concept of the ‘work’ was rehabilitated.

One of the editorial theorists who helped revive the concept of the ‘work’ is Paul Eggert. In the Anglo-American tradition of scholarly editing, the work had often been seen as an ‘idea’ in an almost Platonic sense. For editors who took the final authorial intention as their guiding principle, what the (late) author had intended was the ideal work, which was seldom perfectly realised due to circumstances and agents of textual change (such as censors, amanuenses, typists, compositors, typesetters, publishers’ editors introducing alterations to the text). But writers often rely on the publisher’s editor to correct mistakes, so this textual agency has to be accounted for in a definition of the work. Moreover, the work is a relation between the material document and meaningful text. Text only becomes meaningful thanks to readers, and so readers also need to be accounted for in a definition of the work. Eggert therefore defines the work in terms of a ‘negative dialectic’ – an interaction between the material thesis (document) and the meaningful antithesis (text), but which does not transcend into some sort of Hegelian synthesis of an ‘ideal’ work. The dialectic is explicitly ‘negative’ in Theodor Adorno’s sense: it is not an idealism, but a continuing co-dependence; not an object, but an abstract entity with an ongoing existence. The work thus emerges as a regulative concept that encompasses the ceaseless interaction of the text-document dialectic, involving agency of textual change and taking place over time. In other words, the notion of the ‘work’ models the development of meaning (text) from a material basis (document) over time. From Eggert’s perspective, the regulative aspect means that the work is neither a substance nor ‘a flow of energy’, but a concept that helps us name and differentiate an entity that we refer to as, for instance, *Harry Potter and the Chamber of Secrets*.

**The Americanisation of Harry Potter**

Of course, the previous paragraph raises the question: what about *Harry Potter and the Philosopher’s Stone*? In the US, the first volume of the Harry Potter series did not appear under its original, British title. It was renamed *Harry Potter and the Sorcerer’s Stone*. And this is certainly not the only variant. Does this mean that these are two different works? Or are they variant versions of the same work? Most definitions would agree that they are the same work. But then the question arises how these versions related to each other. Various scholars have manually compared the British and American editions of the first volume to discuss the extent of the revisions, the (un)desirability, and what they suggest about the implied readership and different concepts of childhood of the British and American publishers (Bloomsbury and Scholastic, respectively).

The American publisher, Scholastic, decided to ‘translate’ the Harry Potter books from British English to American English. Daniel Radosh quotes Scholastic’s editor of the Harry Potter books, Arthur A. Levine, as saying:

‘I wasn’t trying to, quote, ‘Americanize’ them. What I was trying to do was translate, which is something different. I wanted to make sure that an American kid reading the book would have the same literary experience that a British kid would have.’ (Levine in Radosh 1999: 56)

In reaction to this statement, Philip Nel rightly asks the rhetorical question: ‘Were it possible to create “the same literary experience” for children from different countries, why would it be desirable?’ (Nel 2002: 261). As Nel suggests, it is telling that this question does not seem to occur to the editor, which buttresses Sukanta Chaudhuri’s accusation of ‘cultural reappropriation’ (Chaudhuri 2000: 5). A year before the first Harry Potter novel came out, Jane Whitehead had already studied the phenomenon of the Americanisation of British children’s literature (Whitehead 1996). One of her conclusions was that the practice is so common that it remains largely unnoticed. But that changed with the Harry Potter novels. The books are read and reread by so many people, many of whom share their reading experiences online, that the differences between the editions, but also continuity errors and typographical mistakes, are often spotted by readers rather than by the copy-editors at the publishing houses.

Philip Nel notes that, to be fair, it should be said that Levine did confer with the author on his ‘translation’: to some extent, the American version should therefore be considered “authorial”. Consequently, if one does *not* regard the American edition as a translation but as another version (as one would generally do with the American editions of British fiction), one could even argue that the Scholastic editions represent J.K. Rowling’s ‘final authorial intention’ (at least so far – for the author may of course still decide to further revise her texts). For instance, the American editor spotted the following continuity error: in volume 1, the author of *A History of Magic* is called Bathilda Bagshot, whereas according to volume 3 (in the Bloomsbury edition) it is written by Adalbert Waffling (*Azkaban*, Bloomsbury, 7). Scholastic corrects the continuity error, changing Adalbert Waffling back to Bathilda Bagshot (*Azkaban*, Scholastic, 1). As Nel points out, some of the changes to the American edition imply an authorial decision and ‘ought to be incorporated into future Bloomsbury editions as well’ (Nel 2002: 264). Needless to say, this would complicate the theoretical debate even further.

The example that Nel uses to illustrate this point is an addition that is not part of the ‘translation’ or ‘Americanisation’ process, but a substantial change that introduces the magical skill of Apparating much earlier (in book 2, *The Chamber of Secrets*) than in the British editions (where it is introduced in chapter 9 of book 3, *The Prisoner of Azkaban*), when Ron suggests flying his parents’ car:

‘Even under-age wizards are allowed to use magic if it’s a real emergency, section nineteen or something of the Restriction of Thingy…’

‘*But your Mum and Dad…’ said Harry, pushing against the barrier again in the vain hope that it would give way. ‘How will they get home?’*

*‘They don’t need the car!’ said Ron impatiently. ‘They know how to Apparate! You know, just vanish and reappear at home! They only bother with Floo powder and the car because we’re all underage and we’re not allowed to Apparate yet….’*

Harry’s feeling of panic turned suddenly to excitement. (*Chamber of Secrets*, Scholastic 69; the passage in italics is not in the British edition)

In the American edition, Harry phrases a question that readers may have had at this point of the story. The bigger point Nel makes is that, in spite of the benefits of the American editions, revised in consultation with the author, the disadvantages of the ‘translation’ outweigh the advantages: ‘replacing British vernacular with what Americans think of as British vernacular diminishes the novels’ realism’ (267), resulting in what he calls ‘British simulacra’ enacting ‘a kind of stealthy vandalism on the source texts’ (268).

**Collation**

*‘Collation’ in Bibliography*

What in textual criticism is called ‘collation’ (the careful comparison of textual versions) is often considered a boring job for uninspiring nitpickers, but Nel’s analysis shows it can be an effective tool to analyse literature(s) and culture(s) at large, concluding in this particular case that ‘Americans should develop an awareness of cultures other than their own’ (272) and endorsing Chaudhuri’s assessment that even spelling changes can be a reflection of (particularly American) ‘global arrogance’, linked by Nel to multinational capitalism. From that perspective, it is telling that an apparently insignificant variant between the British edition’s ‘jelly’ and the American editions ‘Jell-O’ turns a reference to food into a brand name (272). This type of ‘translation’ is therefore a ‘damaging practice’, Nel concludes (283).

Since Philip Nel convincingly argues that the translation was unnecessary, one would expect that the English and American editions were treated as versions of the same book, especially in the case of the later novels, since from the fourth Harry Potter novel onwards, Scholastic reduced its textual interventions. Philip W. Errington’s *J. K. Rowling: A Bibliography 1997-2013* (2015) also treats the transatlantic editions of the second and third novels as versions of the same work. But *The Philosopher’s Stone* and *The Sorcerer’s Stone* do feature as two separate works in his bibliography. They are respectively referred to as A1 (first published in 1997) and A4 (first published in 1998). Errington explains that ‘The decision to separate the English *Harry Potter and the Philosopher’s Stone* from the American *Harry Potter and the Sorcerer’s Stone* is correct within the strictest bibliographical rules since the two titles are different (and so, in occasional places, is the text)’ (xxv).

These ‘strictest bibliographical rules’ are nominalist in the philosophical sense that universals (in this case the notion of the ‘work’) are regarded as mere names without any corresponding reality; only particular objects exist. These objects are described by Errington in a thorough, bibliographical manner. They are ‘merely the carriers of information’, as Errington notes, ‘the physical form on which readers first experienced the imaginary world of J.K. Rowling’. Still, even from this ‘merely’ material point of view the Harry Potter phenomenon is impressive: ‘Never before did a series start with 500 copies in hardback and conclude with a matching edition of over eight million copies’ (Errington 2015: xxi).

Errington provides a bibliographical ‘collation’ (a description of the material aspects of a book edition) for every work by J.K. Rowling. For the first English edition of *Harry Potter and the Philosopher’s Stone*, this collation reads ‘112 unsigned leaves bound in indeterminate gatherings; 197 by 128mm’, followed by a list of the numbered and unnumbered pages: ‘1-7 8-18 19 20-27 28 …’ (Errington 2015: 1). For people who want to know if their copy is a first edition and potentially valuable, Errington gives a few tips and tricks. ‘For our purposes, a first edition is a first impression of a new edition … In the simplest terms, unless your copy includes a strike-line of numbers (usually ’10 9 8 7 6 5 4 3 2 1’) or notes ‘First Edition’, it is a reprint’ (Errington 2015: xxi).

But one of the best tricks is to compare versions. For instance, in the first British hardback edition (500 copies) and the first British paperback edition (5,150 copies) – both of which came out simultaneously on 26 June 1997 – there is a small mistake in the seventh line of the publisher’s imprint page, mentioning the copyright of the cover illustration by Thomas Taylor: there is no space between ‘Taylor’ and ‘1997’. And on page 53, which lists the equipment that pupils at Hogwarts need when they first come to school, the list of ‘Other Equipment’ mentions ‘1 wand’ twice. As Errington notes, J.K. Rowling once wrote in someone’s copy of the paperback edition (sold at Christie’s South Kensington on 30 November 2010), on page 53, next to the final ‘1 wand’: ‘this was taken out in the next print of the book (a mistake). Keep this copy, it might be valuable one day!’ (Errington 2015: 8). To facilitate this sort of comparison, the bibliographer adds a list of 15 variants between the first edition and later reprints and subsequent new editions, including:

Page 53, line 9 deletion of second ‘1 wand’ from list

This brings bibliography very close to scholarly editing, and indeed, according to Philip Gaskell in *A New Introduction to Bibliography*, ‘the heart of the matter’ of bibliography is ‘the use of an understanding of books as material objects in the production and distribution of accurate texts’ (Gaskell 1985: 311). A critical edition usually contains a so-called ‘critical apparatus’, listing textual variants between different versions. The difference between the bibliographer’s list and a critical apparatus is that the former is not exhaustive. Its main purpose is to help discern various editions or reprints. What is slightly confusing, though, is that bibliographers and scholarly editors use the term ‘collation’ to denote different things.

*‘Collation’ in Scholarly Editing*

As we have seen, the bibliographer’s collation is a description of the book’s size, the number of leaves, the way they are bound and numbered (or not). The critical editor’s collation is a comparison of textual versions. The result of this comparison has often been the dullest part of a scholarly edition. The main problem of the critical apparatus is that textual variants that may be valuable for interpretation are buried among hundreds of minor textual changes. That is why self-critical editors often use the metaphor of a ‘graveyard of variants’ to describe the apparatus – a place where variants, once they have been detected and identified, may rest in peace. The opposite should be the case: once the variants are identified, the apparatus should invite researchers to exploit the variants’ hermeneutic potential, for even a differently placed comma can have interpretive consequences.

Previous collations of the US and UK versions of Harry Potter – in the editorial sense – have been published. Apart from Gaskell’s more scattered textual observations or Nel’s more detailed inventory for a number of books, the publishing houses involved have released a survey of the editorial interventions in the jungle of editions that have so far been issued.[[2]](#footnote-2) However, many of these lists are incomplete or lack the necessary scholarly and bibliographic references to make the collations “reproducible” or “replicable” from a quantitative or scientific point of view. Most of these side-by-side comparisons have been carried out manually and people have lamented the painstakingly slow nature of the collation process. It is also easy to miss something.

Developments in digital or automatic collation are making at least two things possible at this moment: first of all, to make sure comparison does no longer need to be done manually (which is a painstakingly slow process), and secondly, to envisage ways to turn the dullest part of a critical edition (the apparatus) into one of the most exciting tools for hermeneutics. A few examples of digital collation tools are Juxta and CollateX. The latter, which we will use below, presents itself as software to

1. read multiple (≥ 2) versions of a text, splitting each version into parts (tokens) to be compared,
2. identify similarities of and differences between the versions (including moved/transposed segments) by aligning tokens, and
3. output the alignment results in a variety of formats for further processing, for instance
4. to support the production of a critical apparatus or the stemmatical analysis of a text’s genesis. (https://collatex.net, 2017)

With this automatic collation tool, we can now test Philip Nel’s conclusions on a larger, empirical scale, and explore ways to measure the tendency towards simplification that critics have addressed on the basis of smaller-scale comparisons. In concrete terms, the question we would like to address is how we can use automatic collation to compare the British and American editions and interpret the differences.

**A computational approach to collation**

In technical terms, collation is a form of **sequence alignment**: one compares two (or more) lists of symbols and tries to find an optimal alignment of the corresponding subsequences between them. Naturally, the sequences which do not match across both sequences also become easily visible in this manner. In biology, algorithms for sequence alignment are commonly used for comparing DNA sequences and establishing family trees of different species that reflect how they are genetically related. In a well-known *Nature* letter from 1998, a team of scientists and scholars nevertheless showed that the same computational procedures can be used to compare literary texts.[[3]](#footnote-3) They used automatic sequence alignment to analyse the 58 extant fifteenth-century manuscripts of “The Wife of Bath's Prologue” from Chaucer’s *Canterbury Tales*. The phylogenetic tree which their work produced corresponded closely to earlier **stemma** which had been produced of this text, using more conventional philological analysis.

Especially in earlier times, producing a *stemma codicum* was often done using the so-called **Lachmannian method**, named after its inventor, the nineteenth-century German critic Karl Lachmann. According to this method, it is possible to determine which manuscripts were copied from each other, through inspecting the errors which they have in common with respect to the original text (i.e. the **Principle of the Common Errors**). While still influential, this view is not entirely uncontested nowadays, partly because it can be difficult for modern readers to determine what exactly constitutes an “error”. In Rowling’s case, for instance, the Lachmannian logic is hard to apply, as some of the corrections introduced by Scholastic, were also introduced in subsequent UK re-editions. In more conventional editorial theory, this phenomenon is known as **contamination**, i.e. the genetic process where (initially independent) lines of textual offspring start to influence or contaminate each other. This should again raise the awareness of the reader that the UK version should not necessarily be considered more “original” or “authentic” with respect to the US version, since one tradition cannot easily claim precedence over the other.

In the past years, automated collation has become an increasingly valuable instrument in editorial scholarship and much methodological progress has been booked. Below, we will apply automated collation to compare the UK and US versions of the Harry Potter series. While a full collation of all versions listed by Gaskell would be a welcome scholarly endeavour, it is also an intractable one: many of these editions are no longer on the market and the extant copies of these versions have become extremely rare (and thus expensive). Additionally, most versions are not readily available in a digital format and those versions that are available, are notoriously hard to trace back to a specific printed edition. Note that Gaskell, perhaps partly for the same reason, does not even cover the digital editions. The Scholastic editions, in particular, are very hard to obtain in a digital format these days, and we suspect that their circulation is being deliberately limited by the publisher for commercial strategies that we will return to below. For all these reasons, we chose to work with the following two complete editions of the series:

1. J.K. Rowling, *Harry Potter. The Complete Collection (1-7)*. First edition. Pottermore Publishing (2015). ISBN13: 9781781106464.
2. J.K. Rowling, *Harry Potter: The Complete Series*. 7 vols. Scholastic, 2013. ISBN 978-0-545-59627-5.

For both items, we chose to work with full *omnibus* editions, to maximize the internal coherence of the individual books in each series. (1) represents the latest version of the UK so-called “Bloomsbury series” as officially available from pottermore.com in a digital format. (2) represents the youngest physically printed and full edition of the Scholastic series that was commercially available. This series was digitized using optical character recognition and manually post-corrected.[[4]](#footnote-4) To the best of our knowledge, both series represent the latest and most accurate versions of both the UK and US book tradition at the time of writing. Both series have been divided into a list of plain text files per chapter – note that the number of chapters is the same in both series and has not changed over time.

Before we delve into a more detailed side-by-side comparison, let us first have a look at some general statistics that we can extract from the series. Below goes a simple line chart that shows the total number of characters in each of the UK books.

A close up of a logo

Description automatically generated

This plot visually confirms our expectation: ever since the notably shorter *Philosopher's Stone*, the books have grown considerably larger. Although the *Half-Blood* Prince somewhat countered this trend, the finale of the series grows longer again – and both concluding volumes are clearly longer than the opening three items in the series. The fifth item, the *Phoenix* reaches the maximum length in characters – and, as will we see later, this is not the only respect in which the *Phoenix* seems to play an exceptional role in the development of the series. Interestingly, Rowling recently tweeted that in her Cormoran Strike series too, the fourth book would be the longest as well:



As nitty-gritty scholars, we should note that our character counts would suggest that the *Phoenix*, and not the *Goblet*, is in fact the most voluminous instalment in the series.[[5]](#footnote-5)

Counting the characters in an edition, however, does not yield a very sophisticated perspective on the Harry Potter books. Using plain character counts to compare the UK and US versions can moreover be misleading, because both series use very different typographical conventions: the UK edition makes use of single quotes, for instance, to mark direct speech, where the US version uses double quotes. More problematic is the issue of suspension points: in the UK edition, these are marked by a single glyph (“…”) whereas the US edition uses three characters, between which it inserts whitespaces (“. . .”), thus distorting the number of characters. Likewise, myriad other spelling differences have been introduced when preparing the books for their American delivery.

A more meaningful level to compare both texts would be the level of sentences and words. As expanded upon elsewhere in this book, the process of segmenting a stream of characters into sentences and tokens, is called **(sentence) tokenization**. Being able to discern sentences and words in the novels would enable us, for instance, to estimate the average length of sentences. Here, we come to touch upon the concept of **readability**, which has been a thorny issue in international Rowling criticism, especially in the context of the difference between the UK and US versions. Reviewers, critics and readers alike have noted that many aspects of the US version seem to have been somewhat “simplified” in comparison to the UK version. Some critics have gone as far as claiming that this would constitute a pure act of “dumbing down” the books for a less intellectual American readership. Others might counter this naive view by arguing that such interventions might actually boost the stylistic qualities of the books. One crude, yet insightful way to quantify such differences would be to inspect the average length of sentences in the books. Would the length of sentences (i.e. the number of word tokens they contain), on average, indeed be shorter in the UK than in the US version?

For this analysis, we started from a prose-only version of all chapters of the UK and US editions in a plain text format. The original text has been segmented into sentence and tokens using the standard sentence and word tokenizer for English, offered by the Natural Language Toolkit (version 3.4).[[6]](#footnote-6) Using string replacements via regular expressions, we have tried to normalize different orthographic conventions in the texts where necessary, such as with the suspension points, making sure that the token counts between both series would be maximally comparable. Applying the same sentence segmentation approach to both series was also necessary to guarantee, as much as technically possible, that direct speech would be treated in the exact same way (as this is a confusing phenomenon for most sentence tokenizers). Finally, we have calculated the average number of tokens in a sentence in both versions and the results are presented below.



A close up of a logo

Description automatically generated

These results were obtained by a fully automated procedure and because such results are never perfect, we should remain careful when interpreting the outcome. Nevertheless, the results suggest a striking difference: we can indeed observe that the US sentences in each book are shorter on average than their British counterparts. The difference is smallest in the (less heavily edited?) *Philosopher’s Stone* but grows more pronounced in the later and longer books, where the US editors (and Rowling herself, apparently) seem to have felt an even stronger need to bring down on the average sentence length. Already this naive sentence level comparison suggests a high level of editorial interference by Scholastic.

**From micro to macro**

So far, we have been relying on fairly crude and superficial metrics to compare the UK and US versions, i.e. average sentence length. This is indeed typical of much work in digital literary studies, where scholars adopt a **distant or macroanalytical perspective** – “dumb down, to scale up”, so to speak. In this section, however, we turn to a methodology that allows us to zoom in again on the text, in order to get a grasp of the *actual* microlevel differences that exist between both versions. Let us start with a simple example and collate the first chapter, “Owl Post” of the *Prisoner of* *Azkaban* in the UK and US versions respectively.

|  |
| --- |
| ***Do it yourself***  Compare the first paragraphs of “Owl Post”, the first chapter of the *Azkaban*, in both traditions. Use a marker to manually spot differences in the text that is offered by both versions. Next, you can compare your own collation with the output of CollateX. Did you miss any? Did the algorithm miss any? Can you propose a classification for the variant readings which you spotted? |

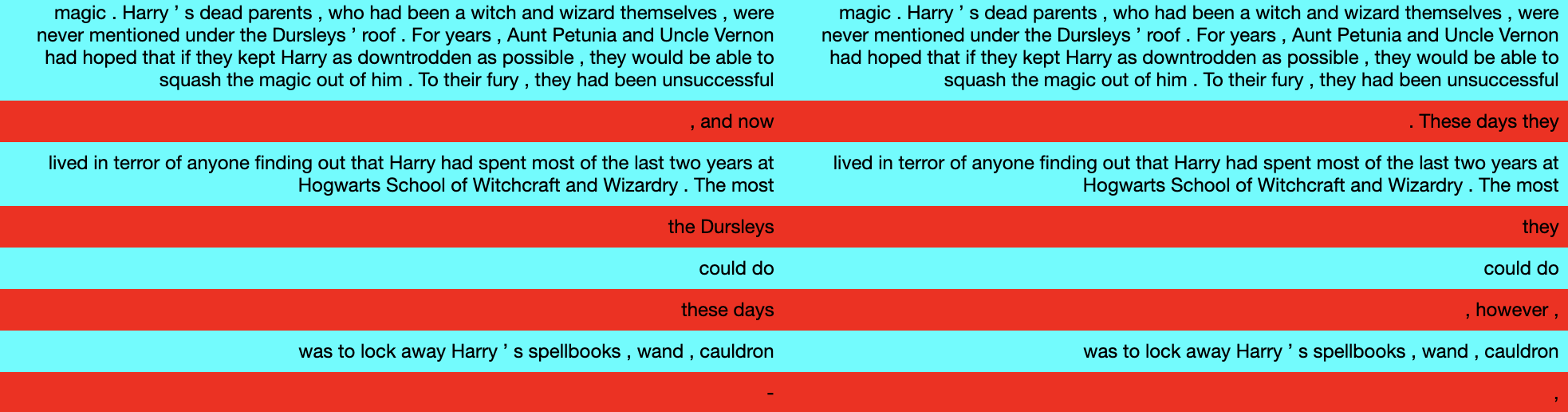
CollateX is a piece of software that does not come with a graphical user interface, but which is easy enough to use from the command line. We point it to two plain text versions of the opening chapter of the series and ask to write the result to a JSON file, that will contain the output in a structured format. Throughout this chapter, we will use CollateX with its default settings to make our analyses more easily replicable. The result of this process is visualized as a table below (for the first paragraphs only):



As we can glean from the table, CollateX outputs a highly detailed side-by-side comparison. The blue cells show parts that are common to the two textual “witnesses”, as parallel versions of the same text are often called by scholarly editors. The red cells contain passages that are different and offer so-called “variant” readings. The American version, for instance, replaces the British word “torch”, with the more colloquial “flashlight”. As you can see, even the most minute difference in orthography gets picked up by the algorithm: the use of punctuation for instance, shows many subtle differences in both versions, with a comma for instance being swapped for a hyphen in the example above.

When perusing the differences detected in this first run, it quickly becomes clear that not all textual differences are equally meaningful. The bulk of deviations are fairly straightforward differences in orthography and punctuation that are not extremely worthwhile from a literary point of view: “neighbours” becomes “neighbors”, for instance, but this does not constitute a very meaningful change. Note that these changes are not necessarily without implications however: in the chapter on stylometry, we will see that “towards” is a typical Rowling-word, which the author frequently uses where contemporary colleagues almost invariably give preference to the more modern spelling of the word (“toward”). If an analysis uses the American version of the books, one should therefore be aware that an important stylistic marker has been consistently removed from the text.

A second category of changes involves clear instances of Americanization: “flashlight” can to some extent be considered a synonym of “torch”, but for an American audience, the latter word might have somewhat different associations (i.e. a long burning wooden stick as opposed to an electric device). The collation abounds in such instances, where synonyms are introduced that can be fairly easily explained when taking the cultural differences of the intended audiences into account. (Note that we do not claim that these differences are uninteresting: we only claim that they can be easily explained.) Interestingly, however, not all differences are so easily accounted for. Many interventions in the text are much harder to interpret from a purely semantic or cultural point of view. Consider the following passage, for instance:



The variants readings which emerge from the collation of these sentences are much harder to explain from a purely semantic perspective. The American edition chooses to start a new sentence after “unsuccessful”, although the text does not provide a clear motivation for that, apart from mere sentence length of course. But the changes go beyond mere sentence splitting: note how “the Dursleys” are replaced by the less specific demonstrative pronoun “they”, although there seems to be no specific reason for this. The same can be said for the variant “these days” versus “however”, the former being moved to the beginning of the newly formed sentence in the US version. Such textual manipulations are intriguing because they no longer involve straightforward semantic synonyms: they reflect *purely stylistic* interventions in the text that are much harder to account for interpretatively.

**Analyzing the variants**

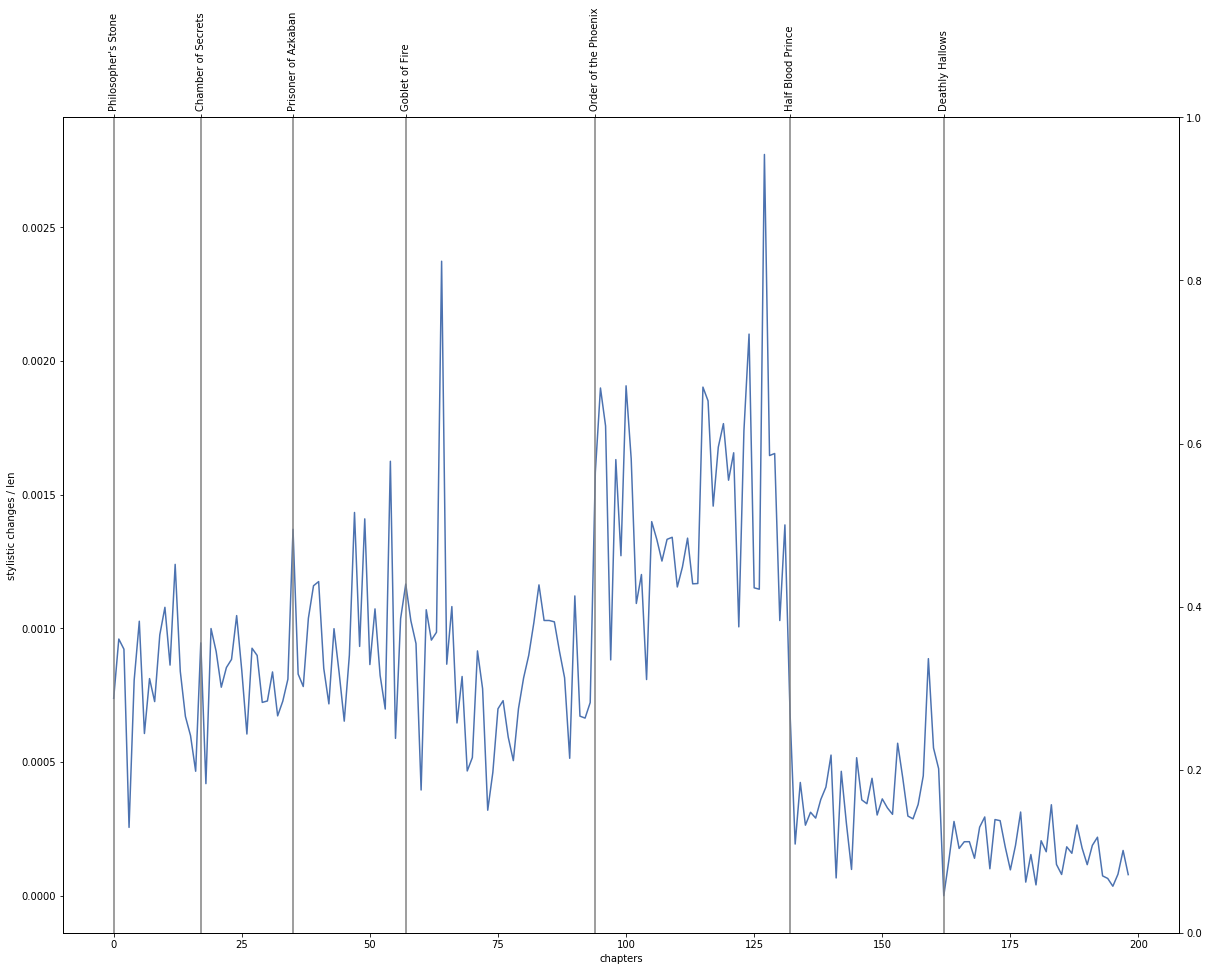
Whereas previous work, such as Nel’s, has mainly focused on semantic differences, the stylistic distortions are equally surprising, and shed more light on the editorial process. The main problem for a computational approach, however, is that we would need a way filter out the non-stylistic interventions and separate these from cases that involve a more transparent synonymy. **Spell checking** is the first filter we apply. The different orthographic traditions in British are American English are well understood and involve predictable character-level replacements, such as “grey” vs “gray”, “neighbours” vs “neighbors” or “realize” vs “realise”. We use a list of common double spellings in English to detect variant readings that can be considered simple spelling variants.[[7]](#footnote-7) Here, we adopt a fairly greedy definition of spelling variants, in the sense that we see the pair “toward” and “towards” are as an orthographic “double”, although one could argue that these are different word types. Additionally, variant readings that are caused by difference in mere punctuation are also assigned to this category.

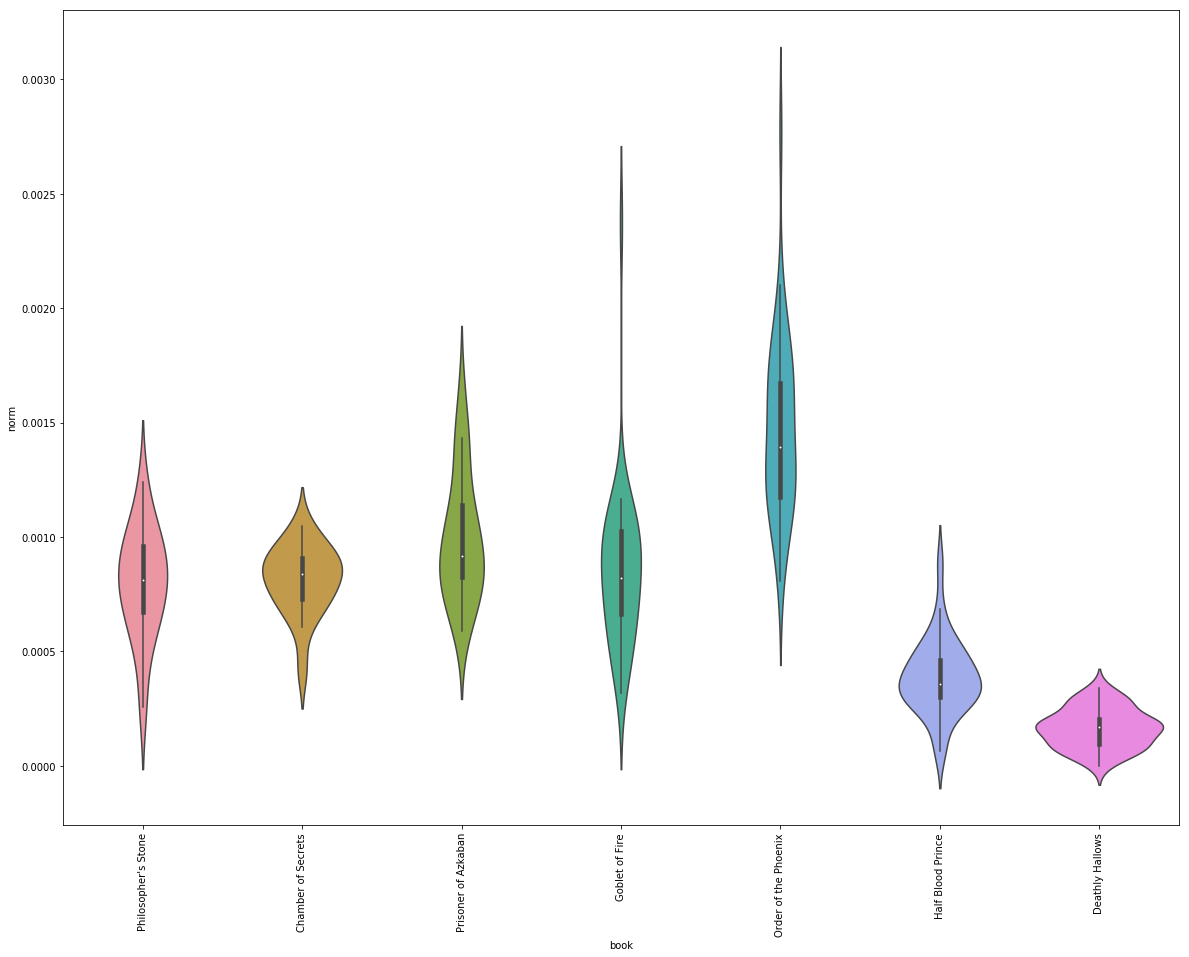
We use a second “filter” to identify **synonyms** or words with a meaning that is so close that they could be considered semantically interchangeable, such as “torch” and “flashlight”. Here, one could wonder whether such terms are truly close enough to be considered “pure synonyms”, since as Philip Nel convincingly argues, such words do come with a different set of associations on both sides of the Ocean. One interesting resource to help up resolve such issues in a more objective and replicable manner is **WordNet**.[[8]](#footnote-8) Wordnet is a lexical resource that provides a large database or ontology of words. In this gigantic tree structure, nouns, verbs, adjectives and adverbs are grouped into sets of synonyms that are called **synsets**. Each of these sets covers an abstract concept and will hold, amongst other, the various synonyms that can be used to express this concept. The English WordNet is a wonderful resource, but one should keep in mind that its coverage is not exhaustive: “torch” and “flashlight” have been recorded in the same synset, as are “cutting” and “clipping”, for example, but “pitch” and “field” have not been registered as synonyms. In such cases, it is hard to determine whether the creators of WordNet have a good reason for this, or whether this goes back to a simple lack of coverage.

We have applied this fairly simple filtering process to the result of the automatic collation in order to obtain a categorization of the variant readings identified. One final category that we distinguished in this respect was the so-called “EOS” or “**end-of-sentence**” category: we marked each sentence boundary in the data with a special symbol (‘<eos>’). Variant readings were assigned to this category, if the UK variant did not contain this symbol, but the US version did: obviously, those variants are instances of new sentence boundaries introduced in the US version. Below goes a coloured table that illustrates the four-way classification of the variant readings, i.e. “neutral”, “spelling”, “synonym”, “eos”, and finally, “style”.

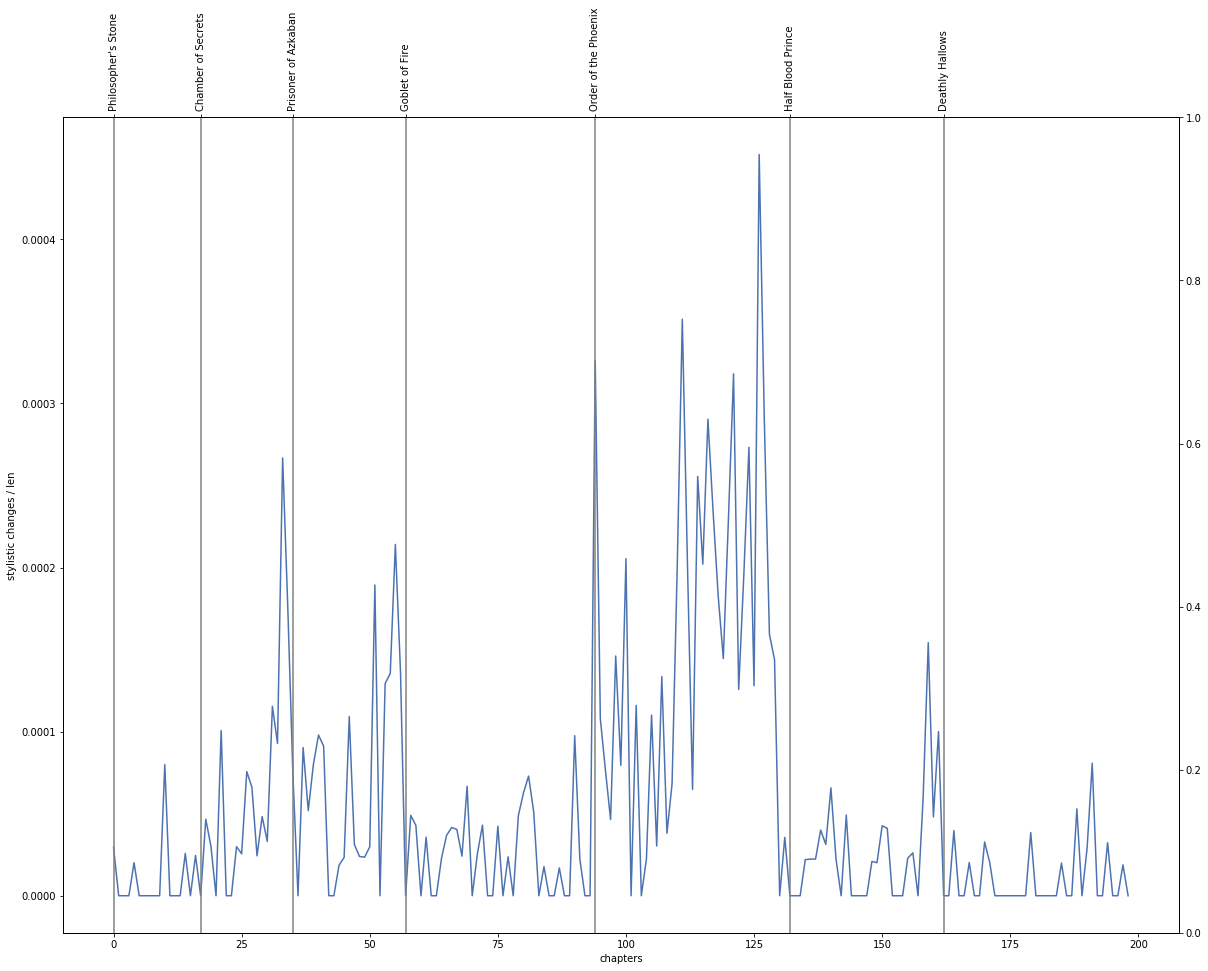


We are ready to start plotting the presence of these four kinds of variant readings throughout the series. We can take the absolute frequency of each phenomenon in a particular chapter and normalize that count and divide it by the number of characters in the UK version of the chapter – so that we can more meaningfully compare chapters of different lengths. Below we show a plot where each point on the horizontal axis represents a single chapter (199 chapters in total): its score on the vertical axis reflects the normalized number of stylistic interventions. The book divisions are shown as vertical lines, to increase the interpretability of the graph. As one can see, the degree of stylistic interventions remains relatively constant for the first four books but shows a remarkable spike for the *Order of the Phoenix*. The final two books then show a remarkable drop and in fact seem to have invited a much lower degree of interference than the first books in the series. Because we see much individual variation between the chapters inside a single book, it is also worthwhile to aggregate these scores in so-called violin plots that better capture the variation inside a single book. The violin plot below, however, does not really alter the view.





Further below, we show a similar plot for the EOS-category. A similar trend emerges, at least again for the Order of the Phoenix, because we see that the longest book also invited the highest number of sentence boundary manipulations. All these measurements indicate how Rowling seems to have pushed the length of her narrative to the extreme in the fifth book, with the American editorial team trying to normalize this tendency as much as possible. In this graph, however, the second and third instalments show more pronounced interventions in this area than the closing pair of novels.



How should we interpret this evolution? Objectively, we have quantitative evidence to argue that the fifth and longest book required the most editorial effort: after the *Phoenix*, it seems like both traditions diverge much less strongly than before. In all likelihood, the “streamlining” of the final instalments in the series were a conscious act towards more uniformity between both traditions. As the series was reaching the height of its *global* popularity, it would make sense that Rowling’s publishing found it decreasingly attractive to maintain two entirely separate geographic traditions for the book series. Both practically and economically, it is much more difficult to maintain two distinct product lines. Another explanation is the rising age of the intended readership as the series progresses: since the final volumes were marketed as books for older readers than the first, adaptations to ease the readers’ efforts may also have been deemed less necessary.

**Bibliography**

Barthes, Roland. 1979. ‘From Work to Text’. In Josue V. Harari, ed., *Textual Strategies: Perspectives in Poststructuralist Cricism*. Ithaca, NY: Cornell UP. 73-81.

Boot, Peter, et al. 2017. *Advances in Digital Scholarly Editing.* Sidestone Press.

Bordalejo, Barbara. 2010. ‘The Encoding System’. *The Commedia of Dante Alighieri: A Digital Edition*. Ed. Prue Shaw. Saskatoon: Scholarly Digital Editions.

Bryant, John. 2002. *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. Ann Arbor: University of Michigan Press.

——. 2013. ‘Textual Identity and Adaptive Revision: Editing Adaptation as a Fluid Text’. In *Adaptation Studies: New Challenges, New Directions*. Ed. Jörgen Bruhn, Anne Gjelsvik and Eirik Frisvold Hanssen. London: Bloomsbury Academic. 47– 68.

Chaudhuri, Sukanta. ‘Harry Potter and the Transfiguration of Language’. *New Straits Times*, Malaysia (9 August 2000): 5.

Errington, Philip W. 2015. *J. K. Rowling: A Bibliography 1997-2013*. London: Bloomsbury.

Gaskell, Philip. 1985. *A New Introduction to Bibliography*. Oxford: Clarendon Press.

Hutcheon, Linda. 2013. *A Theory of Adaptation*. 2nd edn. Abingdon, UK: Routledge.

Nel, Philip. 2002. ‘You Say “Jelly,” I Say “Jell-O”?: Harry Potter and the Transfiguration of Language.’ *The Ivory Tower and Harry Potter: Perspectives on a Literary Phenomenon*. Ed. Lana A. Whited. Columbia and London: University of Missouri Press. 261-84.

Radosh, Daniel. 1999. ‘Why American Kids Don’t Consider Harry Potter an Insufferable Prig’, The New Yorker (20 Sept. 1999): 56.

Robinson, Peter. 2012. ‘Towards a Theory of Digital Editions’. *Variants* 10 (2012): 105-131.

Schmidt, Desmond and Robert Colomb. 2009. ‘A data structure for representing multi-version texts online’. *International Journal of Human`Computer Studies* 67/6 (2009): 497-514.

Schmidt, Desmond. 2010. ‘The role of markup in the digital humanities’. *Historical Social Research* 37 (2010): 125-46.

Shillingsburg, Peter L. 1996. Scholarly Editing in the Computer Age: Theory and Practice (3rd Edition). Ann Arbor: University of Michigan Press.

Stillinger, Jack. 1994. “A Practical Theory of Versions”. *Coleridge and Textual Instability: The Multiple Versions of the Major Poems*. Oxford: Oxford UP. 118-141. Rpt. in *The Broadview Reader in Book History*. Ed. Michelle Levy and Tom Mole. Peterborough: Broadview. 165-80.

Sutherland, Kathryn. 1996. ‘Looking and Knowing: Textual Encounters of a Postponed Kind’. In *Beyond the Book: Theory, Culture and the Politics of Cyberspace*. Ed. Warren Chernaik, Marilyn Deegan and Andrew Gibson. Oxford, UK: Office for Humanities Communication. 11–22.

Whitehead, Jane. 1996. ‘“This Is NOT What I Wrote!”: The Americanization of British Children’s Books’. Part I. *Horn Book Magazine* 72.6: 687-693.

1. Pierre-Marc de Biasi employed the metaphor of fluidity in “Qu’est-ce qu’un brouillon” (“What Is a Literary Draft?”; 1996) and Nicolas Cavaillès subsequently used the metaphor of a river and its affluents to describe the relationship between “endogenesis” (the writing of drafts) and “exogenesis” (the author’s interaction with external source texts): “La rivière de l’exogenèse se jette dans le fleuve de l’endogenèse, tout élément exogénétique étant progressivement incorporé au texte en cours jusqu’à s’y fondre et y suivre un destin commun avec les éléments endogénétiques qui l’environnent” “The river of exogenesis flows into the stream of endogenesis as each exogenetic element is progressively incorporated in the work in progress until it is fully integrated and shares a common destiny with the endogenetic elements surrounding it.” (Cavaillès 2007). [↑](#footnote-ref-1)
2. See, for instance, <http://www.philnel.com/2012/07/22/hpusa/#chart> or https://www.hp-lexicon.org/differences-changes-text/. [↑](#footnote-ref-2)
3. https://www.nature.com/articles/29667 [↑](#footnote-ref-3)
4. These files were scanned, digitized (with *ABBYY FineReader 12 Professional*) and manually corrected by a single proofreader (dra. Lindsey Geybels) in the summer of 2018. [↑](#footnote-ref-4)
5. A number of fans in this thread on Twitter were quick to point this out: <https://twitter.com/jk_rowling/status/958376528643002370?ref_src=twsrc%5Etfw>. [↑](#footnote-ref-5)
6. <https://www.nltk.org/>. Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O’Reilly Media Inc. [↑](#footnote-ref-6)
7. Taken from http://www.tysto.com/uk-us-spelling-list.html (on 22 December 2017). We added a number of new spelling variants. [↑](#footnote-ref-7)
8. George A. Miller (1995). WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11: 39-41. <https://wordnet.princeton.edu/> [↑](#footnote-ref-8)