DHQ: Digital Humanities Quarterly
2021
Volume 15 Number 4

Classifying and Contextualizing Edits in Variants with Coleto: Three Versions of Andy Weir's *The Martian*

Erik Ketzan <ketzane_at_tcd_dot_ie>, Trinity College Dublin
Christof Schöch <schoech_at_uni-trier_dot_de>, University of Trier

Abstract

This paper introduces Coleto, an automatic collation tool for the comparison of variant texts in English, German, or French, which separates edits from variant texts so that textual changes can be classified and contextualized. Coleto's proposed methodology for the classification of edits in variants includes: major/minor expansion, major/minor condensation, changes to numbers and whitespace, and common orthographic features. From this classification schema, Coleto generates: an aligned table of edits in the variants, visualizations of the frequency of classified edits, and a visualization of edit density across the progression of the texts. As a sample use case, we present mixed-method analyses of Andy Weir's science fiction bestseller, *The Martian*, aided by Coleto's functions and generated outputs. Code available at: https://github.com/dh-trier/coleto

1. Introduction

1 This paper began with the desire to combine close and distant reading to compare three versions (or *variants*) of Andy Weir's science fiction novel, *The Martian* (2011, 2014, 2016). We wished to formally describe the changes to the texts made by two professional edits, then interpret how these thousands of textual changes, some big but mostly small, alter readings of the texts. While a number of automatic collation and variant comparison tools exist, none of them easily provide information on the textual changes that we were most interested in when analyzing *The Martian*, including: (1) a visualization of edits across the course of the entire texts; (2) all of the edits aligned and isolated from their original textual contexts; and (3) automatic classification of these thousands of edits. As no available tool performs all of these functions, we created one, and now present Coleto, a Python suite which inputs two variant text files (in English, German, or French) and creates a number of outputs: a visualization of edits across the progression of the texts, which indicates sections of the texts that contain greater or fewer edits; a .tsv table containing all of the textual changes presented side-by-side and categorized by type of edit; and visualizations of frequencies of edit types.[1]

2 Interested users, even those with minimal coding experience, may use Coleto to quickly gain a variety of information about their variant texts. Comparing variant texts has a long scholarly history, yet to perform genetic/textual criticism requires a great amount of manual work: scholars place the variants side by side and read them through closely, pen or highlighter in hand. Coleto can reduce the massive effort of manual comparison and provide faster and easier access to all of the edits in variant texts, classified and presented in easy-to-use formats. With Coleto, philologists and scholars of genetic/textual criticism may gain speed, new information and new precision, at worst, and new insights, at best.

3 Coleto's primary methodological improvement to collation is our suggested classification of edits in variants. Collation tools have previously provided information on insertion, deletion, and (sometimes) transposition, but the study of some variant texts may be improved by additional rule-based classification: expansion major/minor, condensation major/minor, as well as common orthographic features including punctuation (e.g. commas), hyphenation (e.g. *large-ish* → *largish*), whitespace (e.g. *in to* → *into*), capitalization (e.g. *earth* → *Earth*), and italics (e.g. *again* → *again*). We also defined another type of edit, "numbers", for edits common in *The Martian*, e.g. *4* → *four*, or abbreviations involving numbers,

e.g. *80 km → 80 kilometers*.

2. Use case: *The Martian*

4 *The Martian* by Andy Weir is a best-selling science fiction novel about an astronaut named Mark Watney who is stranded alone on Mars. Using science and ingenuity, Watney secures food, water, oxygen, and shelter, establishes communication with NASA back on Earth, and finally reaches a spaceship that enables his rescue. The text of *The Martian* exists in three complete variants. Weir originally self-published *The Martian* on his personal website in 2011 (hereafter, *Martian1* or *M1*), then began selling it on Amazon.com in 2012, for 99 cents, where it sold 35,000 copies in three months [Alter 2017]. A major New York publisher, Crown, no doubt impressed by this self-publishing success, bought the rights from Weir, edited the text, and re-released it in 2014 (hereafter, *Martian2* or *M2*). Describing this edit from *Martian1* to *Martian2* in interviews, Weir has said, “The editing process was pretty smooth. It was not a lot of changes at all” [Savage 2015], as well as, conversely, “...there were a lot of edits and changes... No significant plot changes, nothing like that, but a lot of the wording. It’s much more polished” [Debic 2014]. A third variant was published by Crown in 2016, *The Martian: Classroom Edition*, which removes the text’s extensive profanity to market the novel to educational audiences (hereafter, *Martian3* or *M3*). Explaining this decision, Weir stated, “I got a lot of emails from science teachers who said, ‘Man I’d love to use your book as a teaching aid, but there’s so much profanity in it that we can’t really do that’” [Alter 2017].

5 Our experiments first seek to formally describe the textual changes in the professional edits to *The Martian*, especially as Weir’s inconsistent public statements on the first edit make it unclear how extensive this editing process was. Figure 1 summarizes the textual variation between *Martian1*, *Martian2*, and Martian3, as performed by the statistics feature of Wdiff [Von Gagern 2014].[2]

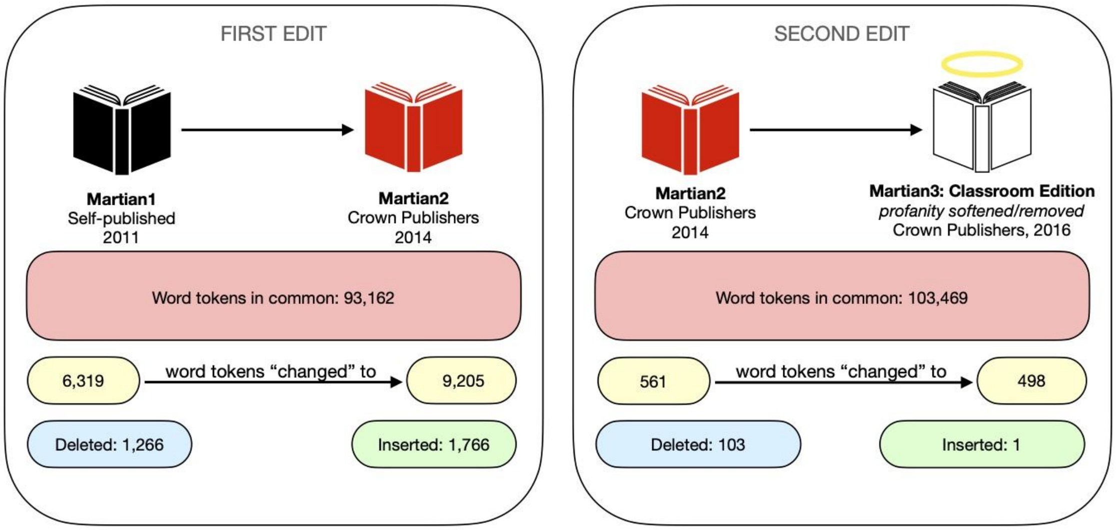

Figure 1. Edits from *Martian1* to *Martian2* and *Martian2* to *Martian3* as grouped by the statistics feature of Wdiff.[3]

6 Quantitatively, Wdiff’s collation statistics lends weight to Weir’s statement that there were “not a lot of changes at all” in the first edit to *The Martian*, as 92.4% of his self-published version is shared with the Crown Publishers edition. But 7.6% is not an insubstantial textual change, as changing even a few words in a literary text can potentially alter its interpretation. The basic collation provided by Wdiff thus leaves many questions about both edits to *The Martian* unanswered. Are there passages or sections of *The Martian* that received especially high amounts of edits? How many edits comprise the 6,319 words transformed to 9,205 words in the first edit, or the words changed in the second edit? Are the edits mostly minor corrections to e.g. punctuation, or more substantial changes such as added paragraphs? Are the edits long or short? How many of these edits can be automatically classified, and classified how? Did *Martian3*, the “Classroom Edition” (intended to remove profanity for student readers) alter any text *besides* the profanity? And moving

on to interpretation, how do the hundreds of minor textual changes, such as capitalization and the rendering of numbers and scientific abbreviations, alter readings of the texts? To begin to answer these, we apply Coleto to the texts.

3. Related Work

A number of tools exist to assist in comparison of textual variants. Wdiff compares text files on a word-per-word basis and provides statistics of insertions, deletions, and changes [Von Gagern 2014]. Archival and philological websites allow parallel visual inspection of variants with annotations such as linguistic information or highlighting [Van Dalen-Oskam 2015]. The Versioning Machine is a long-running open-source tool that displays variant texts and manuscript images side-by-side, with a number of changes highlighted [Schreibman et al. 2003]. Juxta and its web-based interface Juxta Commons provided user-friendly collation and visualization of changes in plain text files [Wheeles 2013]. Juxta Commons as a web service was shut down in September 2020, although its source code remains online.[4] CollateX is advanced software for textual collation providing users with alignment tables as well as a visualization of variant changes as a word-level graph [Dekker and Middell 2011]. TRAViz provides sophisticated collation and aims to improve on CollateX's visualization methodology [Jänicke et al. 2017] [Jänicke 2015]. Elisa Nury has introduced the PyCoviz tool to visualize collation based on CollateX [Nury 2019]. Other tools not specifically focused on variant comparison could also be useful for variant comparison tasks, for instance TRACER, a suite for detection of text re-use [Büchler et al. 2014]. 7

Different users and their research tasks will be drawn to one or more of these collation tools over the other. While existing collation suites generally approach collation visualization through side-by-side comparison of text chunks or sentence-level comparison, Coleto takes a different approach: to identify and separate all of the edits in variant texts, classified by rule-based methods and presented in output documents and visualizations. This assists our close and distant reading of *The Martian*, as we wished to interpret the cumulative effect of hundreds of minor textual changes. As digital humanities collation tools can and do build upon one another, Coleto's open-source code and methodology for classifying edits can be incorporated into any other collation suite. 8

In addition to work in digital genetic criticism (e.g. [Van Hulle 2008] [Ferrer 2011]), previous studies which apply digital methods to the interpretation of variants of relatively recent fiction include Yufang Ho's comparison of the 1966 and revised 1977 versions of John Fowles's novel *The Magus* [Ho 2011]. Martin Paul Eve examined differences in the U.S. and U.K. editions of David Mitchell's *Cloud Atlas*, including collation and comparison of a single character's narrative, arguing that "textual variance [is] an element of *Cloud Atlas* that can and should be *read*" [Eve 2016, 27], see also [Eve 2019]), an assertion that we follow. In addition, Thomas Crombez and Edith Cassiers have analyzed Luc Perceval's adaptation of Dostoyevsky's *The Brothers Karamazov* into a theater play, based on extensive materials from the director's Dropbox account and using Neil Fraser's "diff_match_patch" library to great effect [Crombez and Cassiers 2017]. 9

4. Advancing the classification of edits in variant texts

Existing schemas to classify edits in variant texts have so far been limited. The established methodology for classifying edits in scholarly editing is based on a distinction between *accidentals* and *substantives*, as defined by the widely-used Greg-Bowers tradition, which is also included, for instance, in the MLA Committee on Scholarly Editions' *Guidelines for Editors of Scholarly Editions* [Modern Language Association 2011].[5] Importantly for digital texts, there is no widely-applicable or widely-followed typology of edits in digital scholarly editing and collation, with different materials calling for different typologies [TEI-L 2016]. We therefore suggest the following categories of edits in variant texts, as integrated in Coleto (Table 1). 10

Script-Identifiable Edits	Semantically Open Edits
Numbers, e.g. *4 → four*, or abbreviations involving numbers, e.g. *80 km → 80 kilometers* **Hyphenation**, e.g. *large-ish → largish* **Punctuation**, e.g. commas **Whitespace**, e.g. *in to → into* **Capitalization**, e.g. *earth → Earth* **Italics**, e.g. *again → again*	**Insertion** **Deletion** **Expansion (Minor)** **Expansion (Major)** **Condensation (Minor)** **Condensation (Major)**

Table 1. Coleto’s classification schema for types of edits in variants of text.

11 Script-identifiable edits are those that can be confidently classified into types based on formal linguistic features alone; these include changes related to capitalization, whitespace, hyphenation, spelling of numbers, and changes to scientific abbreviations. For the rest of the edits that cannot be easily defined in a rule-based way, we classify them based on the nature and extent of the edit, under the proposed term, *Semantically Open Edits*: the well-established categories in existing collation tools of deletion and insertion, but supplemented by expansion and condensation categories. For the latter two, a distinction between small and large edits is beneficial, but the choice is (perhaps inevitably) an arbitrary one. For the following experiments, we defined *major* edits as having a Levenshtein distance greater than 5,[6] but users of Coleto may select their own preferred Levenshtein value in the configuration file.

12 The Gothenburg model of textual collation is described as Tokenization → Normalization → Alignment → Analysis/Feedback → Visualization (Birnbaum and Spadini 2020).[7] Coleto’s primary contribution to general collation methodology is to the Analysis step of the Gothenburg model (Figure 2).

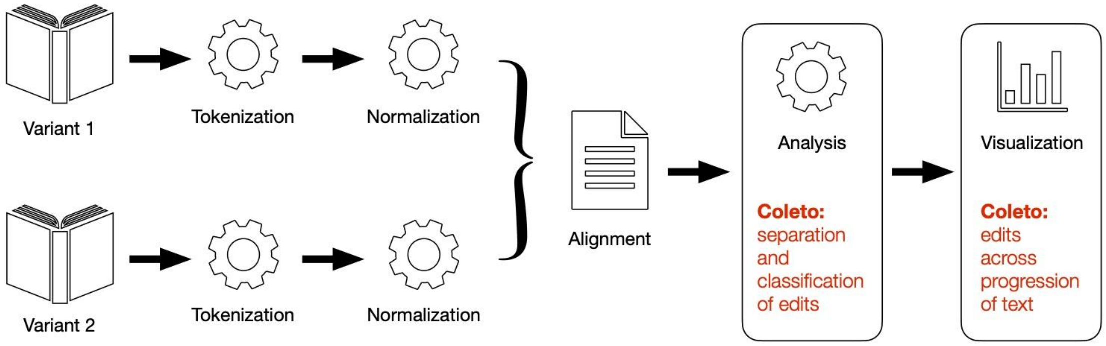

Figure 2. Coleto’s features within the Gothenburg model of collation.

5. Applying Coleto to *The Martian*

13 As its target audience is humanities researchers, Coleto is designed for ease of use, so users need only download the code from GitHub, place two variant texts in .txt format in the Data folder, choose configurations in the config.yaml file, and run the run_coleto Python script. Coleto then performs the following tasks: (1) Preprocessing of texts: splits texts into one sentence per line to prepare for alignment; (2) Align texts using Wdiff; (3) Perform analysis: rule-based detection and classification of edits; (4) Generate overview table; and (5) Generate statistics and visualizations.

14 Coleto also generates a number of outputs, the first being a data table in the .tsv format of all the edits in the variants, aligned in one convenient place (Table 2).

itemid	version1	version2	category	main-type	lev-dist	lev-dist-class	lendiff-chars
9298-1	495	494	script-identifiable	numbers	1	minor	0
9299-1	5	five	script-identifiable	numbers	4	minor	3
9299-2	Tau Event	dust storm	other	tbc	9	major	1
9307-1	Adicalia		other	deletion	8	major	-8
9308-1	3000m	3000 meters	script-identifiable	numbers	6	major	6
9308-2	500m	500 meters	script-identifiable	numbers	6	major	6

Table 2. Sample of table generated by Coleto comparing edits from *Martian1* to *Martian2*.

15 Coleto also creates three visualizations. A bar chart visualizes frequencies of *Semantically Open Edits,* or edits as classified by the statistics feature of Wdiff (insertion and deletion) alongside our suggested Levenshtein-based edit schema for Wdiff’s *changes,* which we subdivide into Expansion Major/Minor and Condensation Major/Minor (Figure 3).

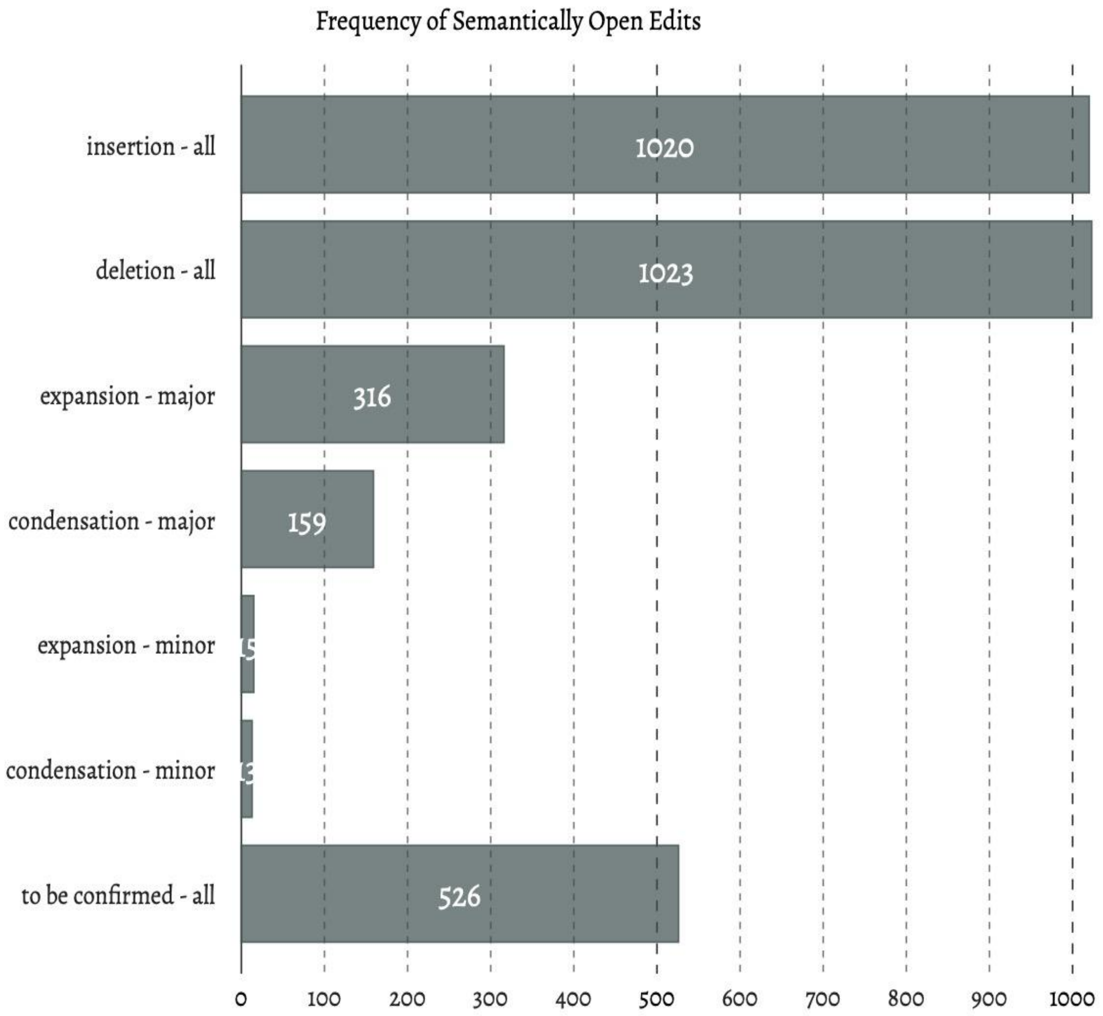

Figure 3. Coleto's visualization of edits between *Martian1* and *Martian2*, with Insertion and Deletion as classified by the statistics feature of Wdiff, alongside our Levenshtein-based edit schema for Expansion and Condensation.

16 This first visualization provides researchers with a bird's-eye view of changes between two variant texts. And if a researcher wishes to examine, for instance, all of the Expansion Major edits, these are tagged in the .tsv table generated by Coleto. The next visualization is of our suggested, more narrowly defined, script-identifiable edits. We have so far implemented: numbers, punctuation, capitalization, hyphenation, whitespace, and italics (Figure 4).

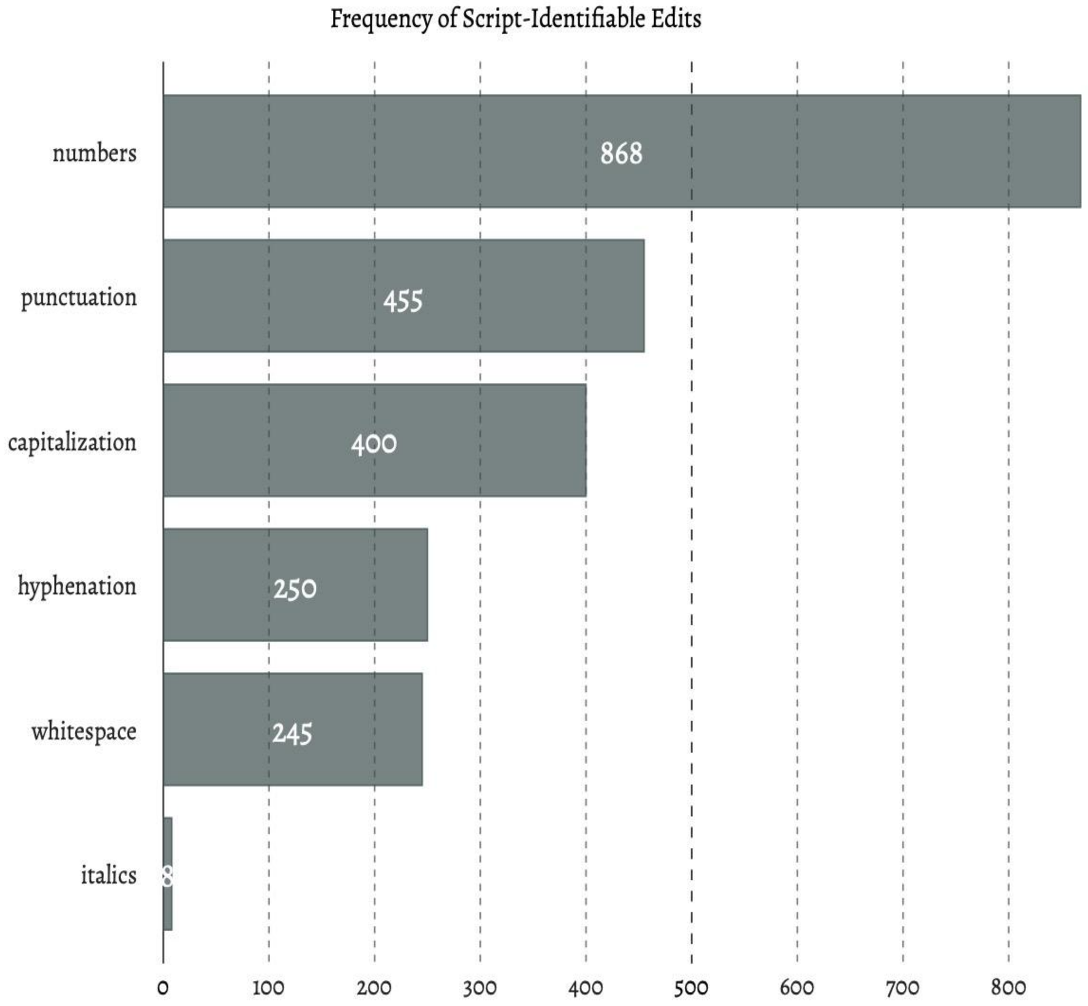

Figure 4. Coleto’s visualization of script-identifiable edits between *Martian1* and *Martian2*.

Figure 4 displays that 2,226 script-identifiable minor textual edits were detected in the editing process from *M1* to *M2*, which provide evidence for a commentator to reconcile author Andy Weir’s conflicting statements on the first edit: “It was not a lot of changes at all,” and “...there were a lot of edits and changes... No significant plot changes, nothing like that, but a lot of the wording.” While changes to hyphenation, capitalization, etc. may be of interest to some research tasks, they are undoubtedly of little interest to others, and the assignment of edits into finer classifications assists researchers in isolating the specific edits they wish to examine. 17

The third visualization generated by Coleto renders all of the edits across the progression of the texts (Figure 5). Script-identifiable and “other edits” are visualized as different series, with the value of each series calculated by the sum of the absolute Levenshtein distances for each sentence of the text, with Savitzky-Golay smoothing applied.[8] 18

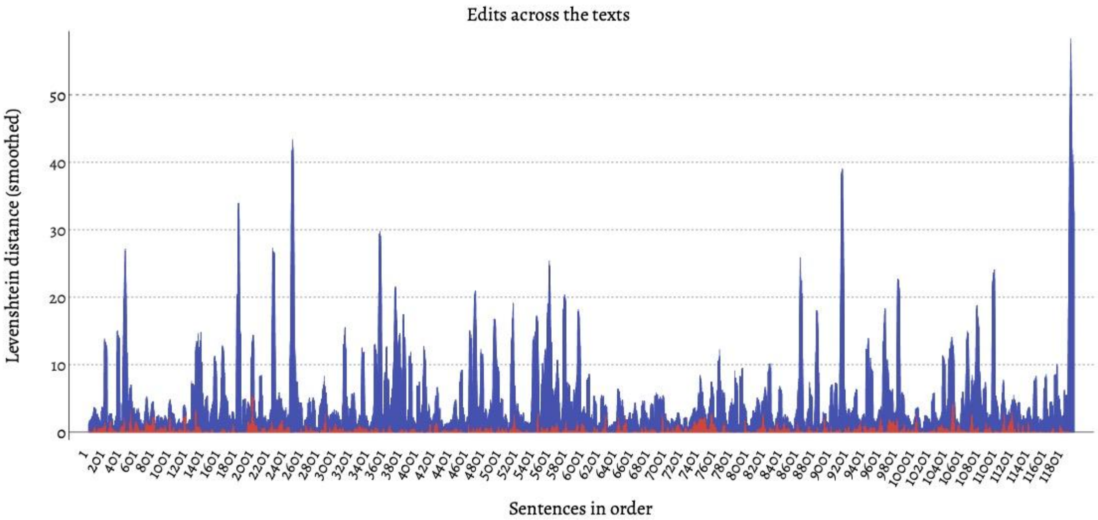

Figure 5. Visualization by Coleto of edits between *Martian1* and *Martian2* across the progression of the texts (Script-identifiable Edits in red, Semantically Open Edits in blue).

This visualization assists the researcher in obtaining a sense of how edits are distributed across a text, with notable spikes highlighting passages or sections which feature especially high amounts of edits, which may warrant closer inspection. 19

6. Investigating *The Martian* with Coleto

The data and visualizations generated by Coleto may assist in a variety of formal, hermeneutic, and mixed-method studies. In this section, we return to our original goal of investigating textual issues in variants of *The Martian*, which can also illustrate Coleto's benefits for researchers. 20

6.1 Quantifying and interpreting reductions in profanity

Profanity is a key stylistic feature of *The Martian*, as exemplified by the novel's opening lines in *M1* and *M2*: "I'm pretty much fucked. That's my considered opinion. Fucked." Profanity may be the most important stylistic feature in the textual history of *The Martian*, as the *raison d'être* for an entire variant, *M3*, the "Classroom Edition," was to remove the novel's copious profanity for educational audiences. Profanity is also a touchstone for issues of censorship and commercialization in discourse around the film adaptation of *The Martian*. While the film only explicitly includes two verbal instances of *fuck*, the offending word is also written on screen and alluded to a number of other times, as Jacob Brogan writes: "[*fuck* is] often suggested in the way that sex might be in a different sort of all-ages movie." [Brogan 2015]. Brogan reads *The Martian* film's limited use of *fuck* as a metafictional device, the film "pointedly nodding to its own limitations [... and] making reference to one of the most bizarre guidelines of the MPAA's PG-13 rating, the principle that a film can include only one, and in some very rare cases two, (non-sexual) uses of the word." Brogan also suggests Watney's frequent use of *fuck* is emblematic of the character's psychology: "Indeed, Watney's access to the one thing he's not supposed to say — and his willingness to keep saying it — indexes his indomitability." As another impression, *The Martian* foregrounds the use of conversational American English to render almost all of its narrative, a tone underscored by its frequent profanity. An improved formal understanding of profanity in *The Martian* texts could thus support a variety of research questions. 21

While locating profanity in texts is a matter of simple query, Coleto's generated tables bring additional speed and convenience to comparing the profanity in *M2* and *M3* (Table 3), while the generated visualizations provide additional context (Figure 6). 22

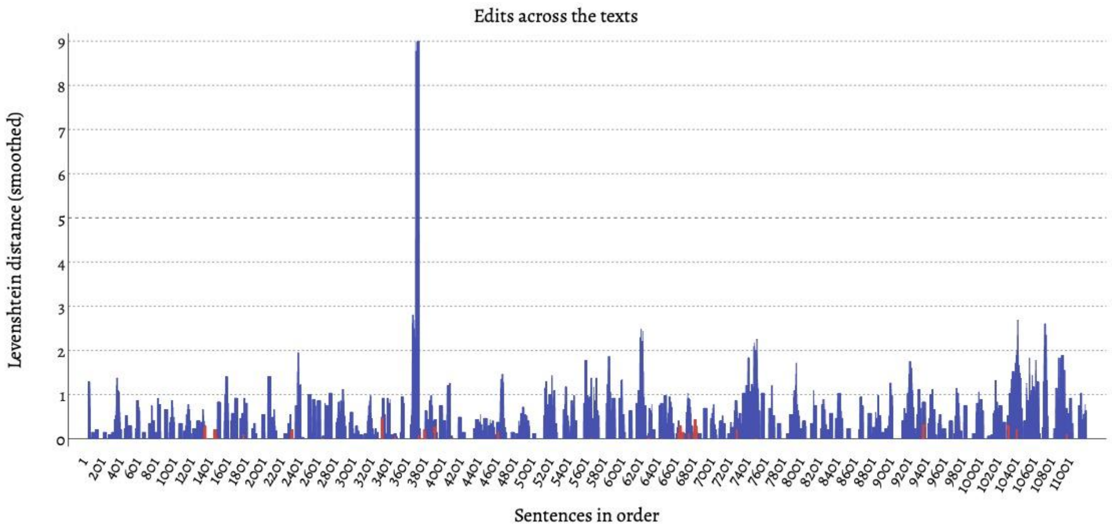

Figure 6. Visualization by Coleto of edits between *Martian2* and *Martian3* across the progression of the texts (Script-identifiable Edits in red, Semantically Open Edits in blue).

Note that the y-axis is much smaller here than in Figure 5, as the edit from *Martian2* to *Martian3* (removing and replacing profanity) was much smaller than the edit from *Martian1* to *Martian2*. The large spike around sentence 3600 in Figure 6 demonstrates how Coleto highlights areas of large edits which scholars may wish to investigate further. In this case, however, the explanation is mundane: small textual variation surrounding a particularly long string of code that NASA sends to Watney on Mars. 23

itemid	version1	version2	category	main-type	lev-dist	lev-dist-class	lendiff-chars
1-1	fucked.	screwed.	other	tbc	5	minor	1
3-1	Fucked.		other	deletion	7	major	-7
4-1		Screwed.	other	insertion	8	major	8
5-1	two months	month	other	condensation	5	minor	-5
48-1	lets	let	other	tbc	1	minor	-1
56-1	shit	crap	other	tbc	4	minor	0
95-1	fucking		other	deletion	7	major	-7
185-1	fucked.	screwed.	other	tbc	5	minor	1
241-1	piss.	pee.	other	tbc	3	minor	-1
280-1	damned	stupid	other	tbc	5	minor	0
309-1	botanist, damn it.	botanist!	other	condensation	10	major	-9

Table 3. Table 3. Selection of table generated by Coleto comparing edits from *Martian2* to *Martian3*.

Manually reviewing the separated edits in Coleto's generated .tsv table reveals that over 290 of the edits to *Martian3* remove or soften profanity, depending on how "profanity" is defined,[9] in ways presumably acceptable to school boards: *fuck* → *crap*, *ass* → *butt*, *piss* → *pee*, *shit* → *crap*, etc. As some examples beyond corporeal profanity, "Jesus Christ" is changed to a secular "Wow," "Thank God" is removed, and a line involving the epithet *gay* is also altered: 24

> [08:31] JPL: Good, keep us posted on any mechanical or electronic problems. By the way, the name of the probe we're sending is Iris. Named after the Greek goddess who traveled the heavens with the speed of wind. She's also the goddess of rainbows. 25

> [08:47] WATNEY: Gay probe coming to save me. Got it.
>
> ([Martian1] and [Martian2])

26 In *Martian3*, “Gay probe” is altered to “Pride parade probe.” These changes to *Martian3* were previously noted by Susan Ohanian (2016), who criticized the spirit of the edit and noted inconsistencies, such as another instance of *Jesus Christ* that remains in the text [Ohanian 2016].

27 None of the bowdlerization in *Martian3* is surprising, as the express purpose of the text was to remove or reduce profanity and other potentially offensive content. But we wished to also know whether *Martian3* altered any text *besides* potentially offensive content — perhaps Weir or his editors used the occasion of a new edition to make some final corrections? — and the table generated by Coleto can quickly answer this question, as well. By manual inspection and sifting of the dataframe, it emerges that *Martian3* also makes factual corrections to scientific details, as well as occasional replacement of continuous verbs and other minor stylistic features (Table 4).

M2	M3	Comment	Comment2	Levenshtein	Automatic Classification
eighteen	twenty-two	science?		9	
most	part	science?		3	
sixty seconds	five minutes	science?		10	
thirty-two-minute	twenty-two-minute	science?		3	
two months	month	science?		5	condensation
grumbled, crossing	grumbled and crossed		passive	7	
said, rubbing	rubbed		passive	9	condensation

Table 4. Selection of edits from *Martian2* to *Martian3*, including manually added comments.

28 By separating and classifying edits, Coleto can also serve as an exploratory research tool by assisting in the spotting of non-obvious trends. For instance, manually inspecting the data table of changes from the earlier *Martian1* to *Martian2*, we stumbled upon the discovery that profanity was cut in this first edit, as well. This was a surprise, given the large frequency and foregroundedness of profanity in *Martian2*. It emerges that in the first edit of *The Martian*, from self-published novel to major publishing house release, *fuck* and *shit* (and their word forms) were substantially reduced, by about 33% and 15%, respectively. Numerous other words and phrases were also softened with “lesser” profanity, more varied profanity, or non-profanity (e.g. “the shit hits the fan” → “all hell breaks loose”) in the first edit.

29 *The Martian* thus has the dubious distinction of having been bowdlerized three times: in *M2*, which reduced profanity significantly, the thorough scrubbing of profanity in *M3*, the *Classroom Edition,* and finally in the film adaptation, which abided by a (reported) film board’s code of two *fuck*s allowed, but winked at the audience through additional implied profanity. This textual history should inform readings of profanity in any variant or adaptation of *The Martian*, and perhaps any discussion of the protagonist Mark Watney’s nonconformist characterization.

6.2 Interpreting changes to numbers and abbreviations

30 When comparing variants manually, or even with existing collation tools, it can be difficult to gain a sense of the frequency and distribution of edits, especially very minor stylistic edits. For instance, in the first edit of *The Martian*, it is readily apparent that numerous minor stylistic/orthographic changes have been made to the way scientific abbreviations and numbers are written, such as *L* → *liter* and *8* → *eight,* but existing variant analysis tools provide no simple way to quantify these and gain a sense of this type of edit’s distribution.

31 From the script-identifiable edits of *The Martian*, Coleto classifies 868 edits as the rendering of numbers and scientific values in the first edit, from *M1* to *M2*. While individual instances of such numerical changes may seem interpretively

insignificant, hundreds of them scattered across the text have a *cumulative* effect upon readings of the text. In protagonist Mark Watney's narration in *Martian1*, the use of cardinal numbers and short scientific abbreviations support the fiction that Watney is a scientist working in dangerous conditions, too focused on surviving the life-threatening conditions on Mars to bother writing "kilometers" or "forty-one" out in full. Plausibility is a central concern of science fiction [Stockwell 2000], and when such abbreviations and numerals are expanded in *Martian2*, this may increase readability, but weakens the stylistic realism of Watney's voice.

32 In *Martian1*, however, Weir did not only write *L* for *liter* or *8* for *eight* in the voice of astronaut Mark Watney, but also in the third-person narration of the NASA scientists back on Earth, which has a jarring effect on readability (Table 5):

	Martian1	**Martian2**
Watney's first-person narration in personal journal style	That extra 18kwh of storage will be tough. I'll have to take 2 of the Hab's 9kwh fuel cells and load them on to the rover or trailer. [Martian1, 212]	That extra 18 kilowatt-hours of storage will be tough. I'll have to take two of the Hab's 9-kilowatt-hour fuel cells and load them onto the rover or trailer. [Martian2, 237]
Character dialogue in the context of third-person narration (NASA scientists)	"What's the biggest gap in coverage we have on Watney right now?" "Um," Mindy said. "Once every 41 hours, we'll have a 17 minute gap. The orbits work out that way." [Martian1, 76]	"What's the biggest gap in coverage we have on Watney right now?" "Um," Mindy said. "Once every forty-one hours, we'll have a seventeen-minute gap. The orbits work out that way. " [Martian2, 85]

Table 5. Selection of passages containing edits classified as "numbers" by Coleto.

6.3 Formal identification of fixed scientific details

33 In addition to plausibility, scientific rigor is another central concern of science fiction [Stockwell 2000, 78], especially the subgenre of "hard science fiction," and Weir pays careful attention to the accuracy of scientific and mathematical explanations throughout *The Martian*. Weir went as far as to code a computer program that plotted the trajectory of the fictional *Hermes* spacecraft to ensure scientific accuracy in *The Martian*. Weir's dedication to scientific accuracy was so great that a NASA scientist later reviewed Weir's trajectory calculations, and concluded that Weir had gotten them right [Burke 2015].

34 In the process of writing and publishing *Martian1* online, readers occasionally pointed out corrections to mathematics and science presented in the novel, which Weir then incorporated [Dickerson 2015]. While locating these scientific corrections in the variants of *The Martian* would certainly be possible manually, it would be a time-consuming task of scanning the text for single-character changes to numerical digits. The side-by-side data table generated by Coleto, however, greatly speeds up this manual search (Table 6).

Martian1	**Martian2**
The answer is a cool **1000**	The answer is about **1100**
The atmosphere is **98%** CO2	The atmosphere is **95** percent CO2
It's a closed system.	It's a closed system. Okay, technically I'm lying. The plants aren't entirely water-neutral. They strip the hydrogen from some of it (releasing the oxygen) and use it to make the complex hydrocarbons that are the plant itself.

Table 6. Examples of scientific corrections in the edit from *Martian1* to *Martian2*, as identified by Coleto's *numbers* edit classification.

6.4 Interpreting the altered epilogue

35 The most substantial spike in Coleto's visualization of edits between *M1* and *M2* in Figure 5, above, occurs at the very end of the novel: the removal of the 263-word epilogue at the end of *Martian1* and a 255-word addition at the end of *Martian2*, which substantially alters the novel's closure. In both *M1* and *M2*, Watney, finally safely aboard the rescue spacecraft, is happy to be alive and states, "This is the happiest day of my life." In *Martian1*, however, a brief epilogue follows this. Some time after his rescue in space, Watney, now safely back on Earth, is employed by NASA to train the crew of an upcoming space mission. Watney sits upon a bench, killing time, when a young boy recognizes him as the famous astronaut. Watney tolerates the boy's attention, until the boy poses a question:

> 36 "So Mr. Watney," the boy said, "If you could go to Mars again, like, if there was another mission and they wanted you to go, would you go?"
>
> Watney scowled at him. "You out of your fucking mind?"
>
> "Ok time to go," the mom said, quickly herding the boy away. They receded in to [sic] the crowded sidewalk.
>
> Watney snorted in their direction. Then he closed his eyes and felt the sun on his face. It was a nice, boring afternoon.
>
> [Martian1, 331–32]

37 This is a typical epilogue in fiction, as defined by the Russian Formalist Boris Eikhenbaum: it sets the perspective by a shift in time and provides some sort of after-history of the major characters [Torgovnick 1981, 11]. This epilogue of *M1* may be read to evince the lack of internal growth by Watney, who remains emotionally disengaged and contemptuous of playing the role of hero. Watney's harrowing adventures on Mars have not altered his nonconformist characterization, a nonconformity encapsulated throughout the text by — and on the final page of the novel, reiterated by — the most foregrounded stylistic feature in the novel: profanity. The epilogue of *M1* also shuts the door on potential sequels, which could be read as authorial and textual nonconformity with commercial expectations for science fiction novels.

38 *Martian2*, however, cuts this pessimistic epilogue and inserts a substantial amount of optimistic text just before the novel's end. In *M2*'s new ending, Watney expresses gracious appreciation for all the parties involved in his rescue — his crewmates, NASA, the Chinese space program, the billions of people who hoped for his survival — and affirms a widespread faith in human nature:

> 39 The cost for my survival must have been hundreds of millions of dollars. All to save one dorky botanist. Why bother?
>
> Well, okay. I know the answer to that. Part of it might be what I represent: progress, science, and the interplanetary future we've dreamed of for centuries. But really, they did it because every human being has a basic instinct to help each other out. It might not seem that way sometimes, but it's true.
>
> [Martian2, 368–69]

40 This greatly alters the tone of the novel's ending, revises Watney's growth and characterization, and now leaves open the possibility of sequels. The correlation between reduced profanity in *M2* and Watney's character is notable: as the text itself becomes more conformist, so does its protagonist.

7. Conclusion and future work

41 A traditional scholar who sat down to compare the first two variants of *The Martian* would possibly, but possibly not, have noticed the reduced profanity, although the changed epilogue is impossible to miss. If such a scholar were sufficiently rigorous and willing to invest the necessary time, they would undoubtedly locate the corrected science and

note the changes to abbreviations, numerals, and other minor stylistic aspects, as well. Coleto, like any collation tool, does not seek to replace such diligent genetic criticism, but to supplement it with speedier information and new precision, as well as provide a method that scales to longer texts which may challenge the time and patience of scholars performing manual comparison of variant texts.

42 Coleto could naturally be adapted per research topic, and may hopefully be useful for critical interest in, *inter alia*, variants of fiction, the role of the editor, and text re-use. Some tantalizing use cases for comparing variants of fiction come from the rapidly emerging global phenomenon of digital self-published best-sellers,[10] in which, like *The Martian*, texts often exist in a self-published variant as well as a later professionally-edited variant.

43 A significant weakness of Coleto is its inability to detect transpositions, as it relies on Wdiff for collation. For instance, Wdiff does not detect if a sentence is removed, but then replaced a few pages later in the novel — Coleto would classify this as an unrelated deletion and insertion. While Coleto could be integrated with transposition-detecting software in the future, for now, a pragmatic research suggestion would be to use Coleto in tandem with CollateX [Dekker and Middell 2011], as described e.g. in [Schöch 2016]. Comparing *Martian1* and *Martian2* in CollateX results in 98 transpositions detected. Of these, 10 involve punctuation and should be considered artefacts of the method. Another 33 represent transpositions of a single word, showing stylistic preferences on the word-order level. Yet another 21 concern multi-word expressions which change the overall construction of a sentence or paragraph more substantially. Finally, 34 concern segments of more than 5 words, typically in the range of a short phrase to several sentences. In all cases, the scope of movement is rather small, with transpositions remaining very localized phenomena. This inspection of transpositions in this test case shows that, quantitatively and above all qualitatively, transpositions were not a major part of the first edit to *The Martian*, but naturally this step might play a greater role in variants of other texts.

44 For further work on Coleto, we plan to pursue the quantitative investigation into the semantics of edits by using word embedding models, thereby quantifying the amount of actual semantic distance between two tokens in an edit, rather than the surface-level Levenshtein distance. Some notable challenges we expect are defining edits concerning spelling mistakes (where misspelled words may not be in a model's vocabulary) and edits involving different numbers of words (where it is not obvious which words establish semantic distance).

Acknowledgements

45 Thanks to Martin Paul Eve, the audience of our talk at Digital Humanities 2017 in Montreal, and the anonymous reviewers of this journal for excellent feedback on various versions of this project.

Notes

[1] Coleto is available on GitHub: https://github.com/dh-trier/coleto, https://doi.org/10.5281/zenodo.4569328.

[2] Wdiff, https://www.gnu.org/software/wdiff/

[3] Wdiff defines *changed* as "A changed word is one that has been replaced or is part of a replacement." Deletion and Insertion thus refer to words which are added or removed with words taking their place in the text. https://www.gnu.org/software/wdiff/manual/wdiff.html

[4] https://github.com/performant-software/juxta-service

[5] Greg-Bowers was originally designed for the disciplinary goal of interpreting authorial intention, and although authorial intention as the dominant paradigm of textual criticism and scholarly editing was disrupted in the 1980s, notably with Jerome J. McGann's *A Critique of Modern Textual Criticism* (1983), and scholars have continued to evolve the goals of textual criticism and scholarly editing, the fundamental methodological categories of the Greg-Bowers tradition have persisted. Scholars are not unanimous in supporting Greg-Bowers, for instance G. Thomas Tanselle, who found the terms *accidentals* and *substantives* to be "misleading and often untenable in their implication of a firm distinction in all cases" [Greetham 1992, 335–36].

[6] Levenshtein distance is a metric for quantifying the difference between two strings: this counts each character-level addition, deletion or transformation required to turn one string into another string. Each individual edit may concern several directly adjacent words, which is why many edits involve Levenshtein distances higher than 5. In the edit from *M1* to *M2*, 2634 (or 49.7%) of the edits are minor and 2664 (or 50.3%)

are major, according to our definition. The median Levenshtein difference we observed was 4. For a discussion of this and other methods of approximate string matching, see [Navarro 2001].

[7] The Interedition Development Group, “The Gothenburg Model”, https://collatex.net/doc/#gothenburg-model

[8] Savitzky-Golay filtering is a smoothing algorithm relying on a least-squares based, best-fit method, originally developed to filter out noise-induced variation in data while producing minimal distortions of the actual signal [Savitsky and Golay 1964].

[9] The definition of “profanity” varies by authority. In Tony McEnery’s computational linguistic investigation of profanity, *Swearing in English*, McEnery queries a list of “bad language words” or BLWs “partly guided by claims within the literature, partly by my own intuition.” [McEnery 2006, 30].

[10] Examples of this phenomenon include *Fifty Shades of Grey* (originally published as fan-fiction and later at author E.L. James’ website) and *Wu Kong* (a Chinese Internet novel turned big-budget action film). Some 40 self-published authors on Amazon had sold over a million copies of their e-books by 2016 [Alter 2016], while Hollywood is reportedly “snapping up” self-published authors in 2017 [Kean 2017].

Works Cited

Alter 2016 Alter, A. (2016). “Meredith Wild, a Self-Publisher Making an Imprint.” *The New York Times*, January 30, 2016, https://www.nytimes.com/2016/01/31/business/media/meredith-wild-a-self-publisher-making-an-imprint.html.

Alter 2017 Alter, A. (2017). “Andy Weir’s Best Seller ‘The Martian’ Gets a Classroom-Friendly Makeover.” *The New York Times*, February 24, 2017, https://www.nytimes.com/2017/02/24/business/andy-weirs-best-seller-the-martian-gets-a-classroom-friendly-makeover.html.

Andrews and Macé 2012 Andrews, T.L. and Macé, C. (2012). “Trees of Texts – Models and methods for an updated theory of medieval text stemmatology.” *Digital Humanities 2012 Conference Abstracts*. Hamburg: Hamburg University Press, 2012, at 120.

Andrews and van Zundert 2013 Andrews, T.L. and van Zundert, J. (2013). “An Interactive Interface for Text Variant Graph Models.” *Digital Humanities 2013 Conference Abstracts*, Lincoln: University of Nebraska, at 90.

Archer and Jockers 2016 Archer, J. and Jockers, M. (2016). *The Bestseller Code*. London: Allen Lane.

Birnbaum and Spandini 2020 Birnbaum and Spadini. (2020). “Reassessing the locus of normalization in machine-assisted collation”. *Digital Scholarship in the Humanities* 14:3. http://www.digitalhumanities.org/dhq/vol/14/3/000489/000489.html#interedition

Booker 2015 Booker, K. M. (2015). *Historical Dictionary of Science Fiction in Literature*. London: Rowman & Littlefield.

Brogan 2015 Brogan, J. (2015). “Ridley Scott’s *The Martian* Has Far Less Profanity Than the Book — but Its F-Bombs Are Perfect.” *Slate*, October 6, 2015. https://slate.com/culture/2015/10/profanity-in-ridley-scotts-the-martian-the-film-drops-so-many-fewer-f-bombs-than-andy-weir-s-book.html

Burke 2015 Burke, L. (2015). “An Examination of ‘The Martian’ Trajectory.” *NASA*, 1-13.

Büchler et al. 2014 Büchler, M., Burns, P. R., Müller, M., Franzini, E., Franzini, G. (2014) “Towards a Historical Text Re-use Detection.” In: Biemann, C. and Mehler, A. (eds.) *Text Mining, Theory and Applications of Natural Language Processing*. Springer International Publishing Switzerland.

Classroom Edition Weir, A. (2016). *The Martian: Classroom Edition.* New York: Crown Publishing Group.

Clute 1995 Clute, J. and Nicholls, P. (1995). *The Encyclopedia of Science Fiction*. New York: St. Martin’s Press.

Crombez and Cassiers 2017 Crombez, Th., Cassiers, E. (2017). “Postdramatic methods of adaptation in the age of digital collaborative writing”. *Digital Scholarship in the Humanities* 32.1, 17-35.

Debic 2014 Debic, B. (2014). “The Martian | Andy Weir | Talks at Google.” [Online Video]. 26 February 2014. Available from: https://www.youtube.com/watch?v=gMfuLtjgzA8. [Accessed: 29 December 2017].

Dekker and Middell 2011 Dekker, R. and Middell, G. (2011). “Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements.” *Supporting Digital Humanities 2011*. University of Copenhagen, Denmark. 17-18 November 2011.

Dickerson 2015 Dickerson, K. (2015). “Some of the trickiest science in 'The Martian' came from the book's biggest fans.”

Business Insider, October 8, 2015. https://www.businessinsider.com/andy-weir-the-martian-science-crowdsourcing-2015-10?r=US&IR=T

Eve 2016 Eve, M. P. (2016). "'You have to keep track of your changes': The Version Variants and Publishing History of David Mitchell's *Cloud Atlas*." *Open Library of Humanities* 2:2, https://olh.openlibhums.org/article/10.16995/olh.82/

Eve 2019 Eve, M. P. (2019). *Close Reading With Computers*. Stanford: Stanford University Press.

Fameli 2015 Fameli, J. (2015). "Adam Savage Interviews 'The Martian' Author Andy Weir - The Talking Room." [Online Video]. 11 June 2015. https://www.youtube.com/watch?v=5SemyzKgaUU

Ferrer 2011 Ferrer, D. (2011). *Logiques du brouillon: Modèles pour une critique génétique*. Paris: Seuil.

Greetham 1992 Greetham, D.C. (1992). *Textual Scholarship: An Introduction*. New York: Garland.

Ho 2011 Ho, Y. (2011). *Corpus Stylistics in Principles and Practice: A Stylistic Exploration of John Fowles' The Magus*. New York: Continuum.

James 1994 James, E. (1994). *Science Fiction in the 20th Century*. Oxford: Oxford University Press.

Jockers 2016 Jockers, M. (2016). "Introduction to the Syuzhet Package." https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

Jänicke 2015 Jänicke S., Geßner A., Franzini G., Terras M., Mahony S., Scheuermann G. (2015). "TRAViz: A Visualization for Variant Graphs." In Digital Scholarship in the Humanities , 30(suppl 1): i83–99.

Jänicke et al. 2017 Jänicke, S., and Wrisley, D.J. (2017). "Visualizing Mouvance: Toward a visual analysis of variant medieval text traditions." Digital Scholarship in the Humanities, Volume 32, Issue suppl_2, Pages ii106–ii123, https://doi.org/10.1093/llc/fqx033

Kean 2017 Kean, D. (2017). "'Show me the money!': the self-published authors being snapped up by Hollywood," *The Guardian*, May 15, 2017, https://www.theguardian.com/books/2017/may/15/self-published-authors-hollywood-andy-weir-the-martian-el-james

Martian1 Weir, A. (2011). *The Martian*. Self-published.

Martian2 Weir, A. (2014). *The Martian*. New York: Crown Publishing Group.

McEnery 2006 McEnery, T. (2006). *Swearing in English: Bad language, purity, and power from 1586 to the presen*t. London: Routledge.

McGann 1983 McGann, J. 1983. *A Critique of Modern Textual Criticism*. Chicago: University of Chicago Press.

Modern Language Association 2011 Modern Language Association (2011). "Reports from the MLA Committee on Scholarly Editions, Guidelines for Editors of Scholarly Editions." https://www.mla.org/Resources/Research/Surveys-Reports-and-Other-Documents/Publishing-and-Scholarship/Reports-from-the-MLA-Committee-on-Scholarly-Editions/Guidelines-for-Editors-of-Scholarly-Editions

Mohammad and Turney 2013 Mohammad, S. M. and Turney, P. D. (2013). "Crowdsourcing a Word-Emotion Association Lexicon." *Computational Intelligence*, 29(3): 436–65.

Navarro 2001 Navarro, G. (2001). "A guided tour to approximate string matching." *ACM Computing Surveys*. 33(1): 31–88. doi:10.1145/375360.375365.

Nury 2019 Nury, E. (2019). "Visualizing Collation Results." *Varia* 14. https://journals.openedition.org/variants/950

Ohanian 2016 Ohanian, S. (2016). "Classroom Edition: Dumb-Ass Fiddling with a Smart-Ass Book". *Schools Matter*. http://www.schoolsmatter.info/2016/07/classroom-edition-dumb-ass-fiddling.html

Savage 2015 Savage, A. (2015). "Adam Savage Interviews 'The Martian' Author Andy Weir - The Talking Room." [Online Video]. 11 June 2015. Available from: https://www.youtube.com/watch?v=5SemyzKgaUU. [Accessed: 29 December 2017].

Savitsky and Golay 1964 Savitzky, A. and Golay, M.J.E. (1964). "Smoothing and Differentiation of Data by Simplified Least Squares Procedures". *Analytical Chemistry* 38(8): 1727-1639. DOI: 10.1021/ac60214a047

Schreibman et al. 2003 Schreibman et al. (2003). The Versioning Machine. http://v-machine.org (accessed January 5, 2021).

Schöch 2016 Schöch, C. (2016). “Detecting Transpositions when Comparing Text Versions using CollateX.” *The Dragonfly’s Gaze*. http://dragonfly.hypotheses.org/954

Simpson 2017 Simpson, P. (2017). “Interview: Andy Weir.” *Sci-Fi Bulletin*. https://scifibulletin.com/books/science-fiction/interview-andy-weir/ (accessed December 1, 2017)

Stockwell 2000 Stockwell, P. (2000). *The Poetics of Science Fiction*. Oxon: Routledge.

Swafford 2015 Swafford, A. (2015). “Problems with the Syuzhet Package”. Anglophile in Academia: Annie Swafford's Blog. https://annieswafford.wordpress.com/2015/03/02/syuzhet/

TEI-L 2016 TEI-L (2016). Types of Edits. TEI-List. http://tei-l.970651.n3.nabble.com/Types-of-edits-tp4028495.html

Torgovnick 1981 Torgovnick, M. (1981). *Closure in the Novel*. Princeton: Princeton University Press.

Twenge et al. 2017 Twenge, J.M., Van Landingham, H., and Campbell, W.K. (2017). “The Seven Words You Can Never Say on Television: Increases in the Use of Swear Words in American Books, 1950-2008.” *SAGE Open*, July-September 2017, pp. 1-8.

Van Dalen-Oskam 2015 van Dalen-Oskam. (2015). “In Praise of the Variant Analysis Tool: A Computational Approach to Medieval Literature.” In *Texts, Transmissions, Receptions: Modern Approaches to Narratives*. Edited by André Lardinois, Sophie Levie, Hans Hoeken and Christoph Lüthy. Leiden: Brill, pp. 35-54.

Van Hulle 2008 van Hulle, D. (2008). *Manuscript Genetics, Joyce’s Know-How, Beckett’s Nohow*. Gainesville: University Press of Florida.

Von Gagern 2014 von Gagern. (2014). GNU Wdiff. https://www.gnu.org/software/wdiff/

Wheeles 2013 Wheeles D., Jensen K. (2013). “Juxta Commons.” In *Proceedings of the Digital Humanities 2013*. University of Nebraska-Lincoln, 17 July 2013. http://dh2013.unl.edu/abstracts/ab-142.html