What Changed When Andy Weir's *The Martian* Got Edited?

Erik Ketzan
eketza01@mail.bbk.ac.uk
Birkbeck, University of London, United Kingdom

Christof Schöch
christof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

Introduction

The Martian is a best-selling science fiction novel by Andy Weir that became a hit film in 2015. The novel exists in two versions, or variants: Weir self-published *The Martian* on his personal website in 2011 (hereafter, "*Martian1*") and began selling it on Amazon.com in 2012. Crown Publishing subsequently bought the rights, edited the book, and re-released it (hereafter, "*Martian2*").

The research presented here investigates what exactly changed when *The Martian* got edited. At first glance, the two versions appear essentially the same, with no major changes to plot, character, or structure. A closer look using a combination of quantitative and qualitative methods, however, reveals a number of noteworthy changes, as well as notable changes that result from thousands of seemingly minor copyedits.

Aims

The aim of our research is to identify what changed between the two variants of *The Martian* using a combination of close reading and digital methods, analyze why those changes are important, and propose a methodology for comparing self-published and later-edited novels, an increasingly common phenomenon. We hypothesize that the editing process of a leading publishing house results in a novel that is more "mainstream", i.e. socialised, domesticated, and appealing to a general audience. In order to test this hypothesis, we explore a range of aspects, including style, content, and character. Our research also aims to bring a critical perspective to the strengths and weaknesses of a variety of qualitative and technical methods in identifying the edits and assessing their importance.

Related Work

In addition to work in digital genetic criticism (e.g. van Hulle 2008), a small number of studies use digital methods to explore variants of contemporary fiction. Yufang Ho (2011) compared the 1966 and revised 1977 versions of John Fowles's novel *The Magus*, while Martin Paul Eve (2016) looked at differences in the US and UK versions of David Mitchell's *Cloud Atlas*. As both Ho and Eve use different methods from one another and from us, it appears that no standard method has emerged so far for this type of research.

Data

The data used for this research is primarily two plain text files of the variants of *The Martian*. *Martian1* was obtained in PDF format from Andy Weir's website. *Martian2* was obtained by scanning a print copy, performing OCR with manual corrections. We consider this our best option given the legal issues regarding text protected by copyright.

Methods and Results

Basic collation

We used the Wdiff frontend to the "diff" algorithm (Hunt & McIlroy 1975) to produce a collated version of *Martian1* and *Martian2* and assess the number and extent of the edits. We then used bespoke Python scripts to classify the edits identified by Wdiff.

We found a total of 5146 edits were made to the novel. While 92% of the 101,000 words in *Martian1* remain unchanged in *Martian2*, the remaining 8% of the words undergo some type of edit, whether they are deleted or modified (Figure 1). The sheer number of edits calls for automatic means to classify them and detect any patterns.

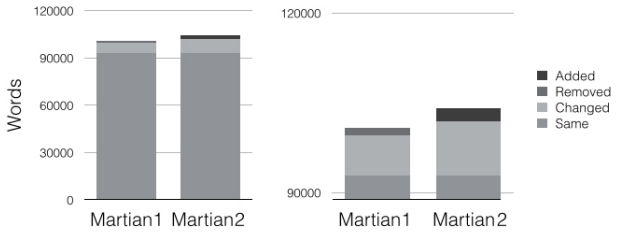

Figure 1: Visualization of edits to The Martian as grouped by Wdiff.

Automatic Classification of Edits

Edits were automatically classified into two broad categories: script-detectable copyedits, and all other edits. Script-detectable copyedits includes changes in capitalization, whitespace, hyphenation, spelling of numbers, abbreviations, or combinations thereof (Figure 2). All other edits were classified as insertion, deletion, expansion or condensation and as "minor" or "major", depending on the Levenshtein distance (Figure 3). Of the 5146 edits, 2863 (or 55%) were script-detectable copyedits, while 2283 (or 45%) comprised the rest. The code used as well as the collation data obtained are available on GitHub.

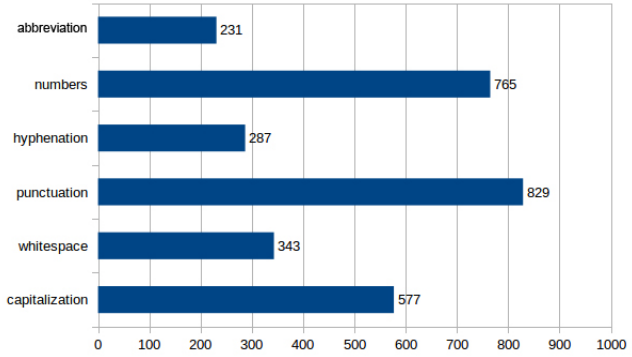

Figure 2: Script-identifiable copyedits to *The Martian.*

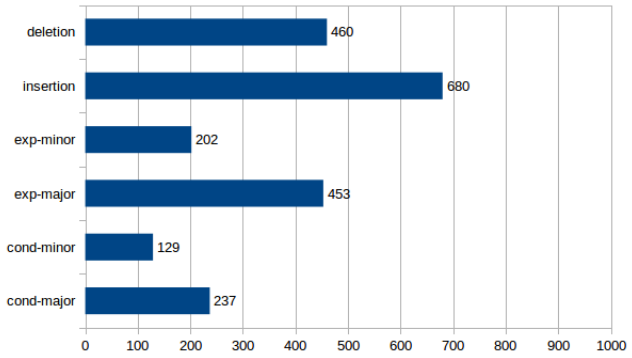

Figure 3: All other edits to *The Martian.*

Cumulative Effect of the Script-Identifiable Copyedits

Taken together, the 2863 script-identifiable copyedits have substantial effects upon the text. Weir's many misspellings and misuse of hyphens and capitalization are corrected. Numbers in *Martian1* are overwhelmingly written numerically, and 765 of these become words in *Martian2*, e.g. "8" becomes "eight". We found 231 instances of edits involving abbreviations, e.g. "L" becomes "liters".

The copyedits work together in different ways when they appear in protagonist Mark Watney's narration or in sections written in the third person (Figure 4). When Watney narrates, the hundreds of misspellings, numerals, and scientific abbreviations in *Martian1* support the fiction that he is a scientist working in extreme conditions. *Martian2* increases readability but eliminates the stylistic realism of Watney's text. When Weir uses, for instance, numerals in the dialogue of other characters, the effect can be jarring. *Martian2* corrects this for the better.

Martian1	*Martian2*
My idea is to make 600L of water (limited by the hydrogen I can get from the Hydrazine). That means I'll need 300L of liquid O2. I can create the O2 easily enough. It takes 20 hours for the MAV fuel plant to fill its 10L tank with CO2.	My idea is to make 600 liters of water (limited by the hydrogen I can get from the hydrazine). That means I'll need 300 liters of liquid O2. I can create the O2 easily enough. It takes twenty hours for the MAV fuel plant to fill its 10-liter tank with CO2.
"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."	"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."

Figure 4: Edits to numerals and scientific abbreviations in Watney's narration (top) and third-person character dialogue (bottom).

Detecting transpositions with CollateX

Wdiff does not detect transpositions, or text that has been moved to a different location in the novel. Using CollateX (Dekker & Middell 2011) as described in Schöch (2016) revealed a total of 126 transpositions. Twenty-eight (or 22%) involve punctuation and should be considered artefacts of the method; 43 (or 34%) represent transpositions of a single word, showing stylistic preferences on the word-order level; 55 (or 44%) concern multi-word expressions which change the overall construction of a sentence or paragraph more substantially.

Figure 5 shows a relatively minor transposition appearing in combination with a contraction of a sentence.

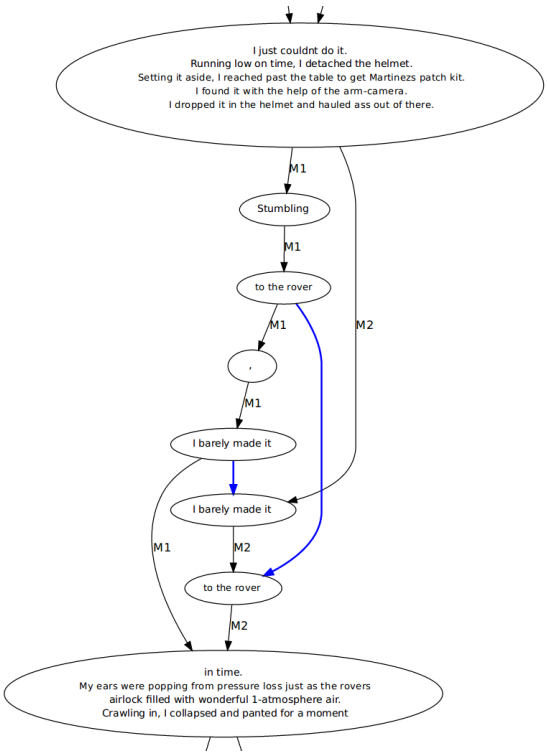

Figure 5: An example of a transposition identified by CollateX.

We conclude that, quantitatively and qualitatively, transpositions were not a major part of the edit to *The Martian*. However, future work could apply the same method to other, comparable variants of novels to gain better reference points.

Close Reading of Other Edits

When we grouped the other edits, placed them into a spreadsheet, and manually inspected them, a number of thematic and stylistic shifts between *Martian1* and *Martian2* became apparent.

Profanity is a key stylistic feature of *The Martian* that is substantially cut and softened by the edit. Words like "fuck" and "shit" are substantially reduced (by about 33% and 15%, respectively), while numerous other words and phrases are softened with "lesser" profanity or simple non-profanity (e.g. "the shit hits the fan" becomes "all hell breaks loose"). Figure 6 shows a selection of these edits. Similarly, crude and sophomoric humor is cut in key instances.

The plot of *The Martian* revolves around solving one problem after another to rescue an astronaut, Mark Watney, stranded on Mars, while relatively little text is devoted to Watney's emotions or inner world. In *Martian2*, however, Watney expresses significantly more emotion: he misses his family and friends more and expresses despair, loneliness, and introspection more often.

06762-1	before I fucked with it.	first.	other	condensation-major	tone-down
10002-1	Unlike my worn out shit, the	The	other	condensation-major	tone-down
10443-1	fucking deathtraps.	death traps.	other	condensation-major	tone-down
00107-1	the blood	it	other	condensation-major	tone-down
00541-1	gives a shit?	cares?	other	condensation-major	tone-down
00904-1	Fucker	It	other	condensation-major	tone-down
01655-1	"This is so fucked up,"	"That'll be fun,"	other	condensation-major	tone-down
02133-1	fucking	damned	other	condensation-major	tone-down
03414-1	fucking shit	shit,	other	condensation-major	tone-down
05303-1	fuck it up,	fail,	other	condensation-major	tone-down
06727-1	a trailer for all the shit I have to bring.	my cargo trailer.	other	condensation-major	tone-down
07084-1	this... oh fuck it.	this ...	other	condensation-major	tone-down
10358-1	my pee in to	water into	other	condensation-major	tone-down
10372-2	an piss-based	a steamy	other	condensation-major	tone-down
11390-1	cause I'M NOT FUCKING THERE!	because I'm not there!	other	condensation-major	tone-down
00698-2	shit in to	mass into	other	condensation-minor	tone-down
06716-1	shit	ass	other	condensation-minor	tone-down
00424-1	After that, things got disgusting.		other	deletion-major	tone-down
00425-1	I spent three hours spreading shit on Martian sand.		other	deletion-major	tone-down
00435-1	That smell's going to stick around for a while, too.		other	deletion-major	tone-down
00436-1	It's not like I can open a window.		other	deletion-major	tone-down
00437-1	Still, you get used to it.		other	deletion-major	tone-down
04989-1	It absolutely pissed the air out.		other	deletion-major	tone-down
04502-2	fuckwits!"	assholes !"	other	expansion-major	tone-down
05016-2	shit.	useless.	other	expansion-major	tone-down
07284-1	fucker	bastard	other	expansion-major	tone-down
07327-2	shit.	anything.	other	expansion-major	tone-down
09518-1	fuckton	shitload	other	expansion-major	tone-down
00641-1	fucking	screwing	other	expansion-minor	tone-down
01241-1	fuck	screw	other	expansion-minor	tone-down

Figure 6: examples of toned-down profanity in the editing of The Martian

Additionally, *Martian1* contains an epilogue that is completely cut in the edit. It portrays Watney, back on Earth, being openly and profanely rude to a young fan. In *Martian2*, meanwhile, text is added to have Watney express gracious appreciation for all the parties involved in his rescue and a widespread faith in human nature. The edit therefore alters the tone of the ending substantially.

We believe that all of these changes, analyzed together with close reading, serve to align Watney's character with our overall hypothesized goal of the edit: to make Watney more "relatable," "nice," and "human," and thus to appeal to a wider audience.

Edits Over the Course of the Novel

Patterns in the edits related to textual progression are revealed by measuring the absolute Levenshtein distance of the script-identifiable copyedits and other edits line by line (Levenshtein distance is a metric for measuring the difference between two sequences, see Navarro 2001).

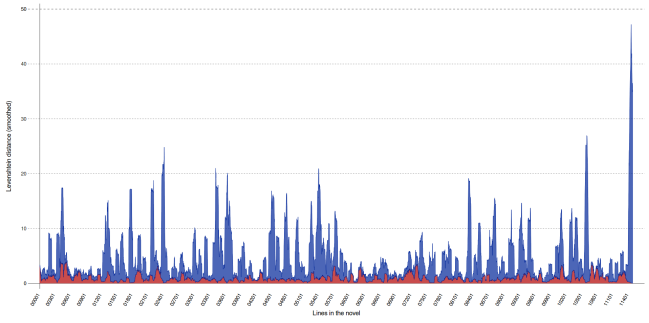

Figure 7: Sum of absolute Levenshtein distance per line over textual progression (script-identifiable copyedits in red, other edits in blue).

Figure 7 shows the sum of the absolute Levenshtein distances for each line of the novel (with Savitzky-Golay smoothing applied). The graph shows the substantial modifications to the ending of the novel, but also a large number of locations with smaller but nonetheless above-average modifications.

Conclusion and Further Research

We have identified and analyzed a number of key features that emerged from the editing of *The Martian*, notably on the level of style and character, which combine to make the novel more appealing to a wider audience.

Ongoing research into *The Martian* concerns the relative frequency and function of parts of speech, quantifying the amount of syntactic change, and the legal issues affecting the obtaining and processing of the texts. We hope to present these additional findings in the near future.

As for our typology of edits, an established methodology for classifying edits in the companion fields of textual analysis and scholarly editing is the distinction between the "accidentals" and "substantives" used by the Greg-Bowers tradition and included in the MLA Committee on Scholarly Editions' *Guidelines for Editors of Scholarly Editions* (Modern Language Association, 2011). Scholars are not unanimous, however, in supporting this. G. Thomas Tanselle, for instance, found these terms "misleading and often untenable in their implication of a firm distinction in all cases" (Greetham 1992, pp.335-336). Further, there appears to be no widely-applicable typology of edits in digital scholarly editing and collation, with different materials calling for different typologies (see TEI-L 2016).

Our typology of edits departs from previously proposed ones by focusing entirely on types which can be identified automatically, based on surface features. While limited in scope and excluding any semantic criteria, our typology may serve as a first approach to the edits of any text and allow quantitative comparison of some key phenomena. We believe that our method could be applied to other variants of fiction — by itself or incorporated alongside another taxonomy, including accidentals/substantives — particularly to novels which begin as self-published works but are later edited and re-released, an increasingly important phenomenon in contemporary fiction.

Bibliography

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.

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.* https://olh.openlibhums.org/article/10.16995/olh.82/

Greetham, D. (1992). *Textual scholarship: An introduction.* New York/London: Garland Publishing.

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

Hunt, J. W. & Mcilroy, M. D. (1975). An algorithm for differential file comparison. *Computer Science.*

Modern Language Association (2011). Reports from the MLA Committee on Scholarly Editions, Guidelines for Editors of Scholarly Editions, available at: 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

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

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

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

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

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

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