**BORNET — Anatomy**

<5 figures, 1 table>

**Anatomy of a Drop-Off Reading Curve**

Bornet, C. and Kaplan, F.

Not all readers finish the books they start reading. Electronic media allow us to measure more precisely how this ‘drop-off’ effect unfolds as readers are reading a book. A curve showing how many people have read each chapter of a book is likely to be progressively going down: the more we advance in the book, the less readers got to this point. This article is an initial study about the shape of these ‘drop-off’ reading curves.

The data we analyse were gathered from readers of a series of novels of different length written by one single author. During the last twenty years, Daniel de Roulet wrote a nine-book saga about the interleaving stories of two families over a time range of seventy years. In 2014, for the publication of the concluding tenth book, he decided to release an alternative retelling that constituted of a reordering of all the 297 chapters, forming a coherent saga titled *La simulation humaine*. This new saga was made available free of charge to the general public in the form of a website1 and a mobile iOS2 and Android3 app. It has been composed in a way that is readable from one end to the other. This creates the possibility of reading nine chosen subsequences, each one making for a self-standing story. Each possible reading path is characterised by its number of chapters and is also given a title, as shown in Table 1 below.

|  |  |
| --- | --- |
| Story title | Number of chapters |
| *Une fleur de cerisier* | 1 |
| *Sur les barricades* | 1 |
| *Les liquidateurs* | 1 |
| *Air Force One* | 4 |
| *Fukushima au début* | 5 |
| *Le pucelage d’un Helvète* | 10 |
| *L’ingénieur et la fillette* | 30 |
| *Fissions* | 33 |
| *La réalité, mais digitale* | 38 |
| *La simulation totale* | 174 to 297 |

Table 1. Stories of *La simulation humaine* and their respective sizes, in chapters.

**Methodology**

Formally, the collected data are very similar to state-of-the-art website analytics. Every time a chapter is loaded, it is possible to know the precise date, time, and type of device used. Additionally, each reader is uniquely identified using a browser session and a cookie on the website, or a device identifier on the mobile app.

Over a period of eight months, a total of 7,000 chapters were read, out of which approximately half were not relatable to regular readers (be it search-engine queries or users refusing to be tracked). For this study, we thus kept the subset of data where at least two chapters were related to the same person. We also left out the paths presenting clear test patterns (for instance, paths consisting of multiple repetitions of the same chapters) or no overall significant reading time. In all, we considered 310 unique reading paths.

Each path is represented by a chain of tuples (*ci,ti*), with *ci* being a chapter and *ti* the time the reader spent on it (that is, the interval in seconds between start of *ci* and *ci*+1). Given the average reading speed of an educated adult (Kershner et al., 1964) and the relative homogeneity of our corpus, we marked transitions as ‘skipped’ when the said time interval was less than 60 seconds, thus considering the chapter as not having been read. We then aggregated the total number of read chapters and plotted them according to their respective stories’ orderings, resulting in one *reading curve* for each story.

**Characterisation of Reading Curves**

All reading curves are expected to start with a strong drop-off effect, corresponding to the considerable number of readers who merely peek at the start of the story, but don’t finish it. Of course, this effect is especially noticeable in the context of digital and mobile content (‘Localytics Indexes’, *Q3* 2014).4

The classic ‘drop-off’ reading curve is characterised by an exponential drop-off as illustrated in Figure 1. In our samples, this phase was shown to typically last for the first three or four chapters of the story. Then the rate of the monotonic drop-off can vary depending on the attractiveness of the texts. Figure 1 shows two examples, one with a very high drop-off rate and another one where this effect is present but less important. Notice that not only is the drop lower in the second case but the effect also wears off earlier.

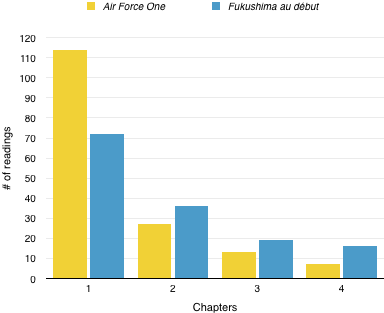


Figure 1. Two examples of exponential drop-offs with different rates/slopes.

More interestingly, some ‘drop-off reading curves’ are also characterised by a plateau that occurs typically once the first chapters are passed. This could be because readers who reach a certain point in the book are convinced to go on till the end. Figure 2 shows an example of this plateau characterised by an almost null drop-off rate once chapter 6 is passed.

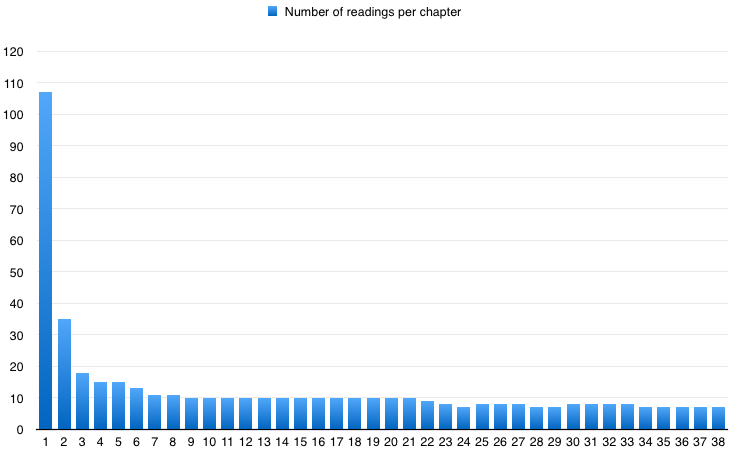


Figure 2. An example of a curve with a plateau in *La réalité, mais digitale*.

In some cases, the drop-off reading curves show unexpected gaps in the reading plateau. This case could typically be attributed to readers who skipped one or several chapters (hence resulting in low reading times that were filtered out of our data), but then resumed reading later on in the same story. This ‘skimming gap’ could thus be interpreted as a mild sign of boredom, not sufficiently strong to stop the reading but significant enough to speed it up. This phenomenon can be seen in Figures 2 and 3.

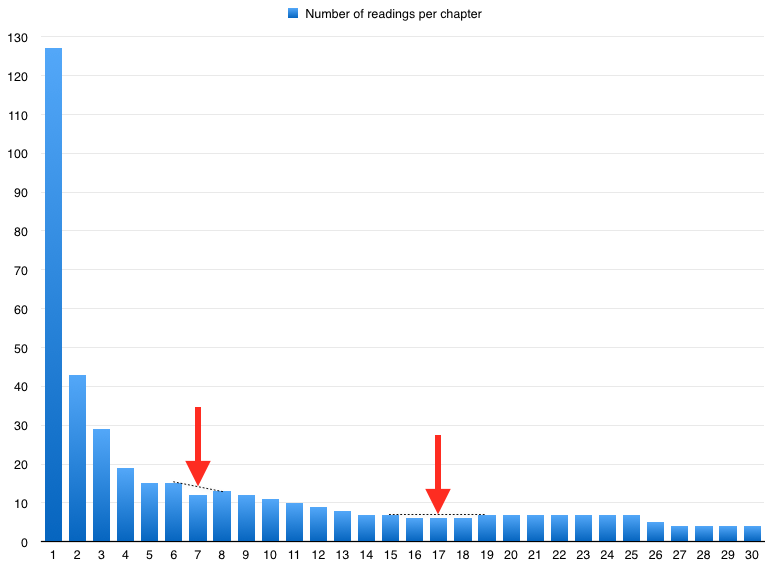


Figure 3. A skimming gap in *L’ingénieur et la fillette*.

Moreover, some other stories never achieve the plateau stage and continuously lose readers as the chapters go by. In the example shown in Figure 4, this decrease also features a couple of stop points, which could possibly be interesting starting points to investigate why the subsequent drops occur.

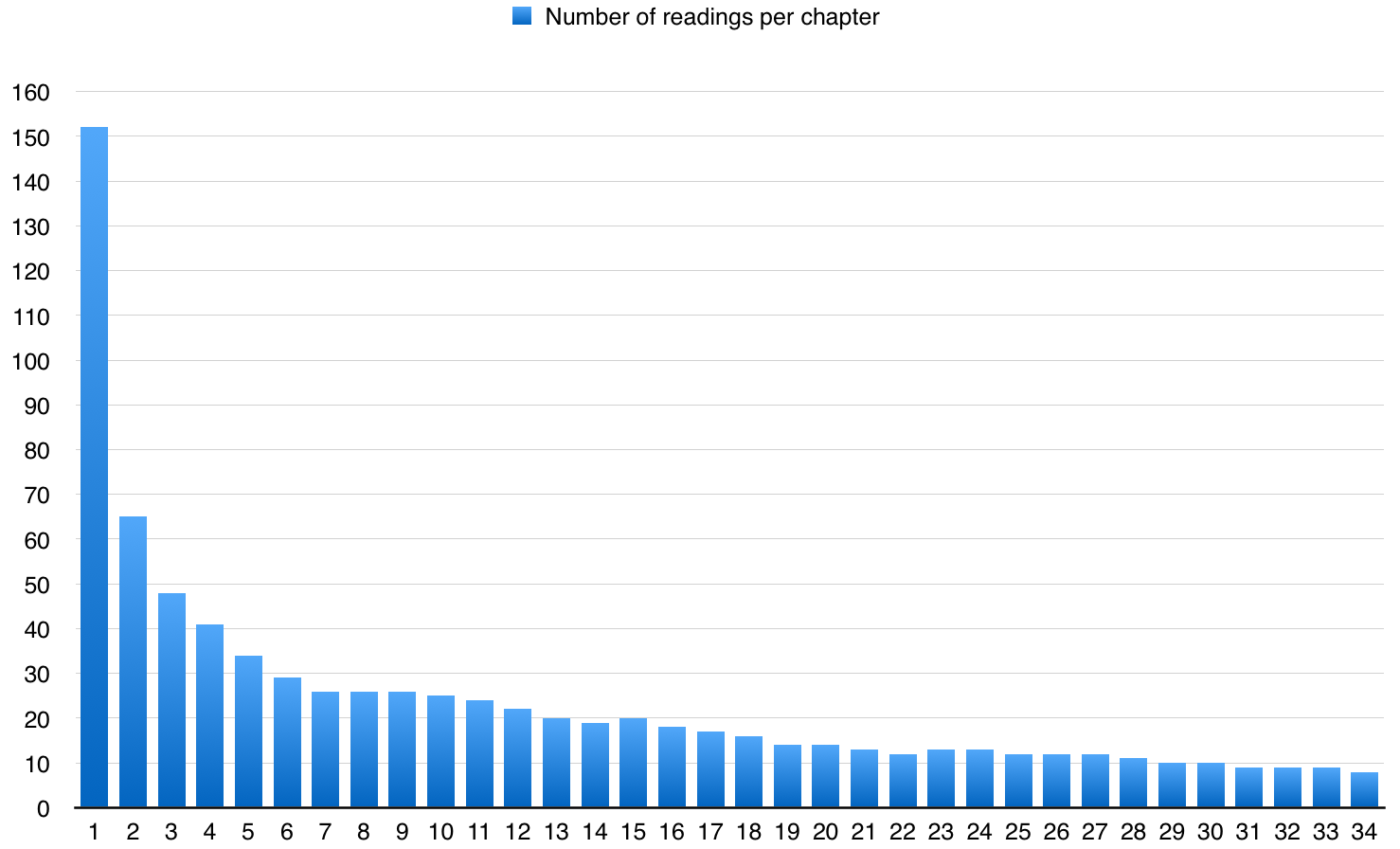


Figure 4. Continuously decreasing readers in *Fissions*.

**From Reading Curves to Chapters Classification**

These simple examples show the potential richness of drop-off curves among the various reading analytics curves. Two reading regimes can be identified from this initial study: the immersion mode, characterised by very small drop-off rate, and the critical mode, characterised by dropping and skimming behavior.

We segmented the chapters of our corpus using these two categories and trained a machine learning classifier to try to identify key features to predict whether a chapter is immersive or leading to a potential drop-off. We decided to use an implementation of a J48 pruned tree, given the generally good scores obtained by this method in computational linguistics (Pedersen, 2001; Youn and McLeod, 2007) and general purposes classification (Omid, 2011). A visual representation of the resulting tree is shown in Figure 5.

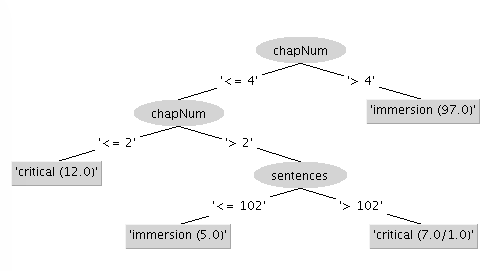


Figure 5. J48 tree predictor for critical and immersion reading modes.

This structure yields a 95% correct classification in a 10-fold cross validation on our dataset, which makes it overall a pretty accurate predictor. It basically confirms our empirical observations on the general shape of the reading curves, putting the start of the immersion phase latest after the fourth chapter. Interestingly, it also hints that using shorter chapters in the critical phase could hasten the immersion process. Furthermore, this result shows that the other features characterising the chapters—such as their position in the saga, length in words, or presence of main characters—had no significant influence on the drop-off rate prediction.

**Conclusion and Further Works**

We are aware that these preliminary results may be highly linked to the studied corpus and not generalizable. However, the concepts and tools that we proposed in this article could well be extended or adapted to look at more general reading patterns, with or without the active participation of the authors of the texts to be analysed.

As more people read *La simulation humaine*, we expect to collect more analytics and be able to refine our results, as well as confirm or falsify our hypotheses with deeper quantitative analyses. Additionally, we’ll aim to propose tools that may detect critical reading phases, suggest improvements to the author, and predict drop-offs for new stories.

**Acknowledgements**

This research was made possible by a very precious and close collaboration with Daniel de Roulet, Swiss architect, computer scientist, and author. This research is supported by the *Fondation Jan Michalski pour l’Ecriture et la Littérature*.

**Notes**

1. http://www.simulationhumaine.com.

2. http://appstore.com/la-simulation-humaine.

3. http://play.google.com/store/apps/details?id=com.simulationhumaine.app.

4. http://www.localytics.com/resources/app-indexes-q3-2014/.

**Bibliography**

**Alter, A.** (2012). Your E-Book Is Reading You. *Wall Street Journal*, 19 July, http://online.wsj.com/articles/SB10001424052702304870304577490950051438304.

**Kershner, A. M.** (1964). Speed of Reading in an Adult Population under Differential Conditions. *Journal of Applied Psychology,* **48**: 25–28, http://dx.doi.org/10.1037/h0047157.

**Omid, M.** (2011). Design of an Expert System for Sorting Pistachio Nuts through Decision Tree and Fuzzy Logic Classifier. *Expert Systems with Applications,* **38**: 4339–47, http://dx.doi.org/10.1016/j.eswa.2010.09.103.

**Pedersen, T.** (2001). A Decision Tree of Bigrams Is an Accurate Predictor of Word Sense. *Association for Computational Linguistics*, pp. 1–8, http://dx.doi.org/10.3115/1073336.1073347.

**Reading, Literacy & Education Statistics.** (n.d.). http://www.readfaster.com/education\_stats.asp#literacystatistics.

**Roulet, D.** (2013). *Ecrire la mondialité: Essays.* Baconnière, Genève.

**Youn, S. and McLeod, D.** (2007). A Comparative Study for Email Classification. In Elleithy, E. (ed.), *Advances and Innovations in Systems: Computing Sciences and Software Engineering*. Dordrecht: Springer Netherlands, pp. 387–91, http://link.springer.com/10.1007/978-1-4020-6264-3\_67.