An OCR Pipeline and Semantic Text Analysis for Comics

Rita Hartel(✉) and Alexander Dunst

Paderborn University, Warburger Straße 100, 33098 Paderborn, Germany
rst@upb.de, dunst@mail.upb.de

Abstract. Optical character recognition has remained a challenge for comics, given the high variability of placement of text on the page, the wide variety of frequently handwritten fonts, and the limited availability and small size of datasets. This paper reports on currently on-going work on an OCR pipeline that includes text spotting with the help of a U-Net based fully convolutional neural network and OCR training with the open-source software Calamari, which was performed on the "Graphic Narrative Corpus" of book-length graphic novels written in English. Based on the results of the OCR training, we then present an analysis of the textual properties of 129 graphic novels correlated with page length, historical development, and genre affiliation.

Keywords: Comics · Graphic novels · OCR · Text spotting · Semantic analysis

1 Introduction: Context & Previous Work

Computational comics analysis has made significant strides in recent years (for a recent overview see [1]), both in document analysis and content generation. While automated text recognition has become routine for many types of documents, however, it represents an on-going challenge for comics and its subtypes, from comic strips and Japanese manga to French bande dessinée and graphic novels. Figure 1 shows some examples of texts within captions or balloons of a graphic novel that are difficult to detect automatically. There are several reasons for this status quo: Both visually and textually, comics are characterized by the high variability typical of artistic production. Text fonts, the types of texts used, and their placement on the page may differ from one strip or book to another, or even within a single work. Traditionally, text was handwritten by authors or, in the case of more industrially-produced comics, by specialized letterers. Today, they are as likely to be digitized by using samples of such handwriting, or produced from existing fonts that aim to give the impression of being handwritten, particularly in translation or in serial comics.

The variability of text fonts extends to text types. Speech and thought representation are usually placed inside balloons that indicate their source via a tail. In addition, onomatopoeia, or sequences of letters that imitate sounds, and diegetic text—from street signs to characters reading newspapers—may be found anywhere on the comics page. Graphic novels further include a substantial amount of narrative text, commonly placed

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12666, pp. 213–222, 2021.
https://doi.org/10.1007/978-3-030-68780-9_19

in captions. As indicated, there are graphic conventions for these different types, but authors may choose to disregard them to make stylistic statements or pursue narrative goals. Because some of these conventions and their application differ significantly between subtypes, the transfer of heuristics presents significant difficulties. Similarly, pretrained OCR engines based on neural network architectures lead to relatively high error rates [2]. As a direct consequence, semantic and syntactic text analysis has remained unusually limited for these documents, drawing on small sets of manually annotated texts or so-called paratexts, such as letters to the editor [3, 4]. Based on a preliminary version of the OCR pipeline presented below, Hartel & Dunst presented an exploratory bag-of-words based analysis of a small number of graphic novels written in English [5].

Fig. 1. Examples for texts in graphic novels that are hard to detect automatically

To the best of our knowledge, no information on even the basic textual properties of comics text is currently available. Among other potential uses, such data would allow for a comparison with other textual or multimodal media, from films and television to novels and dramatic plays, or within the aforementioned subtypes of comics. This information can also be used to compare visual and textual features, the two semiotic modes present in most comics. In a step towards the further document analysis of this medium, the paper details the construction of an OCR pipeline via a U-Net based page segmentation recently developed by Dubray & Laubrock [6] and the open-source OCR software Calamari [7], trained on the Graphic Narrative Corpus," which currently consists of 255 book-length graphic novels [8]. Some of the visual properties of this copyright-protected corpus had been analyzed in an earlier paper [9], but automatic text recognition and analysis had heretofore eluded us. We then report on the methodology and results of the OCR training for a sample of 131 graphic novels from the GNC. Finally, we include an overview of basic textual properties and an analysis of the textual complexity of different genres contained in the GNC.

2 Methodology and Dataset

Two basic avenues exist for optical character recognition. Bottom-up approaches aim at separating text from graphic elements without prior page segmentation. This has

the advantage that text types such onomatopoeia or diegetic text that do not follow established graphic conventions in their placement may more easily be included in recognition tasks. However, this approach has led to high error rates in the past [2]. Here we follow the opposite path, that is to say, we begin with page segmentation but exclude onomatopoeia and diegetic text. This decision is motivated by the expectation that prior segmentation will lead to better recognition results but also accords with the specific properties of the documents under discussion. Graphic novels are book-length comics, often running to a length of several hundred pages. As their name implies, they share some properties of other print literature, including a large amount of narrative text. Onomatopoeia plays a much smaller role than in manga and related traditions. While diegetic text helps create atmosphere and may contain clues for detailed interpretation, it is usually of little importance for an understanding of the story.

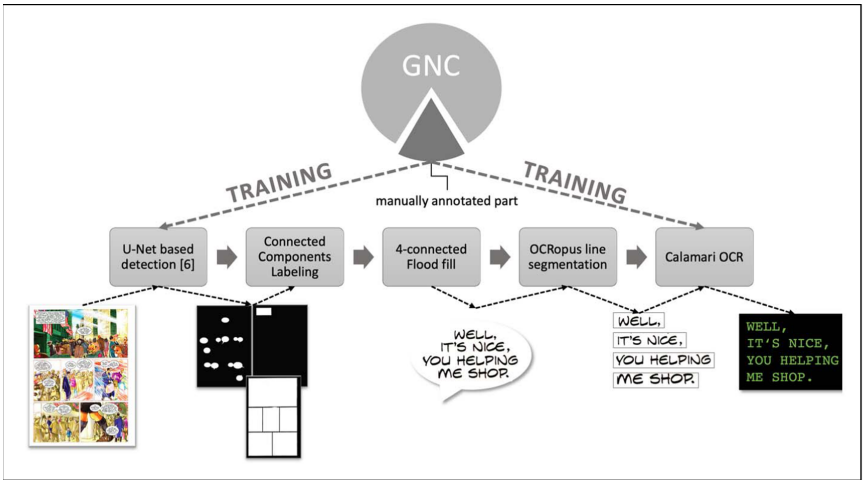

Fig. 2. Pipeline for automated GNML annotation

For page segmentation, Dubray & Laubrock employed a U-Net-based fully neural network model that distinguishes between speech balloons, captions, and panels. This architecture was trained on 3,430 manually annotated pages taken from 200 graphic novels in the GNC [8]. The F1 scores were 95% for speech balloons, 91.8% for captions, and 98.6% for panels, significantly above the results reported by Nguyen et al. [10].

In a first phase, we then train Calamari OCR [7] on a training set consisting of 12,838 lines of text extracted from the manually annotated pages. We chose Calamari over the LSTM-based Tesseract, as in our preliminary tests Calamari achieved better results. We evaluated the training with a test set consisting of 4,726 lines of text (i.e., 80% training data and 20% test data).

Figure 2 shows how the pipeline automatically generates GNML annotations for scans of graphic images. Input to the pipeline is a scan in form of a PNG image. The **U-NET based detection** uses Dubray & Laubrock's model that classifies which pixels in the scanned graphic novels are part of a panel, a caption, a balloon, or belong to other

components of the comics image. In more precise terms, the output of the trained system are three bitmaps: one for panels, one for captions, and one for speech balloons. Each pixel has a (grayscale) value between 0 and 255, where the higher the value, the higher the probability that the pixel belongs to the specified class (panel, caption, balloon).

These bitmaps are analyzed in a post-processing step. We first perform **Connected Component Labeling**, i.e., we search for connected components within the bitmaps. Looking specifically at the background of the black-and-white version of the original scan, we then perform a **4-connected flood-fill** approach on each connected component within the bitmap. This process leads to more precise shapes than if we based our analysis only on the machine learning approach. In order to produce input for the next step in the pipeline—OCR with the help of the trained Calamari model—we removed noise that is likely to belong to the border of the speech balloon or the caption. For this step, we once again employed the 4-connected flood-fill approach to remove all areas that consist of border pixels (in most cases these were black in the black-and-white version of the original scan) that touch the border of the detected polygon.

After extracting the polygons that included the text objects (caption or balloon) as binary images, we applied the **line segmentation tool of OCRopus/ocropy** [12] to split these images into separate text lines.

Finally, the trained **Calamari OCR** proceeded with detecting the text within each text line image. After OCR processing is completed, all of the retrieved data (panel polygons, caption and balloon polygons including detected text) were transformed into an GNML file, an XML dialect developed for the analysis of the graphic novels contained in our corpus. These files were then analyzed together with the data from manually annotated graphic novels.

In a second phase, we extended the OCR pipeline described above to graphic novels that had not yet been annotated. Automated analyses were corrected manually, which was still far less effort than annotating text completely by hand. This second phase resulted in an extended training set of 58,608 lines and a test set of 14,597 lines of texts. Training yielded an average character error rate of 3,47% on the test set. In further semantic analyses for a digital humanities project, we focus on methods that are performed not on the texts themselves but on the bag-of-words for individual graphic novels, meaning that word order does not matter. As a consequence, we measure the overall quality of the OCR system in terms of what we call the Bag Error Rate (BER) instead of the more character or word error rates [5]. The BER is computed by comparing the bag-of-words of the original text and of the recognized text, calculating the number of different entries, and dividing it by the total number of words contained in the original text. Figure 3 shows a histogram of the BER for phases 1 and 2. As can be seen from the leftward movement on the X-axis of the histogram, the BER decreased significantly as we moved from phase 1 to phase 2 in the OCR pipeline, with the vast majority of texts now showing an error rate between 0–20% and occurrences of very high error rates much lower than was the case after phase 1. In our pipeline, text lines are often read in the wrong sequence due to the shape of individual balloons. Therefore, one frequent error concerns incorrect word order. Consequently, the CER is a bit worse than the BER. Yet, the overall trend in phase 2 is similar, as evidenced by the significant leftward movement after phase 2 that results in an average CER of 28%.

3 OCR Pipeline: Results and Discussion

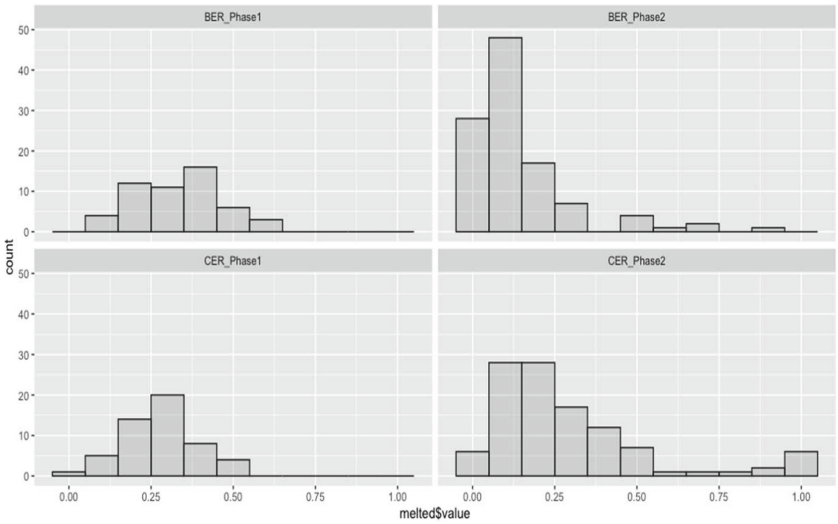

Fig. 3. Comparison of BER (Bag Error Rate, top) and CER (Character Error Rate, bottom) after phase 1 (left) and after phase 2 (right)

Of the 255 full-length graphic novels currently available as retro-digitized book scans in the GNC, we have manual annotations of the entire text in captions and balloons for 15 graphic novels, plus further annotations of roughly 10% of the pages of 106 graphic novels. After the first phase of the pipeline, we semi-automatically annotated 10% of the pages of 25 additional graphic novels. That is to say, we ran our trained system to produce automatic text annotations and checked them for accuracy manually. Automatic annotations that had few errors were then corrected by hand to add to the training set.

We used the partial annotations of these 131 graphic novels to evaluate the quality of our automatic annotations and computed that 70 of these met our conditions for further semantic text analysis. In addition to publication in North America, either in translation or in the original English, these requirements include that at least 80% of the objects (panels, caption, balloons) were recognized correctly and that the BER amounted to less than 20%. While earlier tests [5] showed that bag-of-words analyses can produce meaningful results starting at a BER of up to 50%, the higher threshold was chosen to meet recognition rates comparable to those used in research on major text formats, from historical newspapers to other print literature. Furthermore, we computed automated annotations for the remaining graphic novels for which no ground truth existed. 10% of the pages were manually checked by trained research assistants to see whether most of the objects (once again around 80%) were detected correctly and whether the OCR confidence for these texts reached a minimum of 80%. In total, 114 automated annotations fulfilled our quality requirements. To these were added 15 manually annotated graphic novels. Taken together, then, 129 texts were used for the analysis of textual properties described below.

4 Some Textual Properties of 129 Graphic Novels

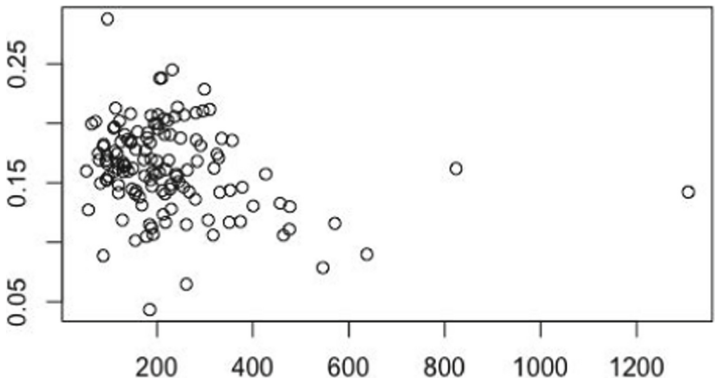

Fig. 4. Relation between page length and textual density, with a statistically significant weak negative correlation coefficient of –0.244

The dataset provided as part of this paper contains basic statistical information on the textual properties of 129 graphic novels. This information is given for each page of these book-length graphic novels (29,926 pages in total), with the exception of pages that only contain bibliographical or additional texts such as forewords and do not contribute to the overall narrative. Pages are numbered starting with the cover page of each graphic novel. Bibliographic information, including the ISBN numbers that enable identification of the editions used for digitization, are available in a separate database. The linguistic measurements include: syllable and sentence length, the number of unique words, normalized type-token ratio (cttr), and the number of words per page. While the first are standard measurements in computational linguistics, the last makes use of the single comics page as a stable, functional and semantic, unit. The data was extracted from the full text using the text analysis package KoRpus [11] and can be downloaded from our homepage.[1]

For our text analyses, we further calculated a complexity score for each graphic novel, consisting of the averages of syllable and sentence length, cttr, and words per page. The first two are widely used to calculate the readability of written text, for instance in the Gunning Fog index and Flesh reading ease test. Sentence length and variants of normalized type-token ratio have also become standard measures in the digital humanities, specifically in digital literary studies [13, 14]. Finally, the number of words per page was added because cognitive studies of comics have shown that text consumes the vast majority of a reader's attention, with the time spent looking at panels increasing with the amount of text [15]. The complexity scores were then correlated with information about page length, year of publication, and genre affiliation, which can be found in the aforementioned GNC database. Figure 4 shows a weak inverse relationship between page length and textual density. In other words, the more page numbers a title has, the

[1] https://groups.uni-paderborn.de/graphic-literature/gncorpus/TextFeaturesGNC.csv.zip.

lower its textual density will be. This accords with recent studies that have found a similarly negative correlation between the complexity of novels and the productivity of their authors [16].

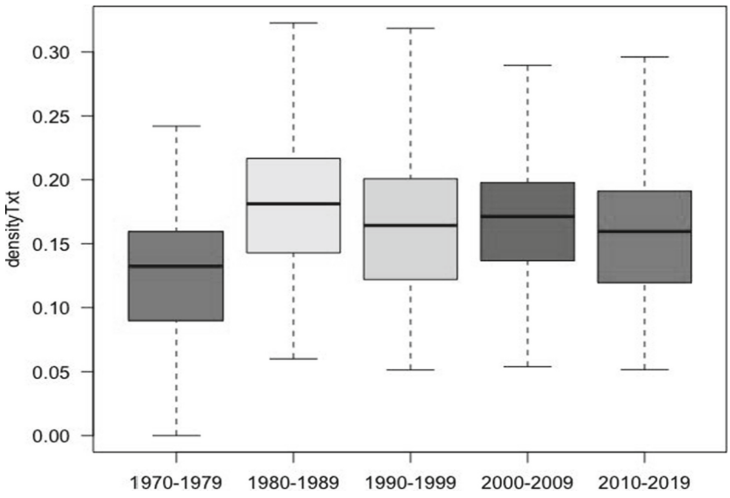

Fig. 5. Textual Density scores of 129 graphic novels by decade

Graphic novels have steadily increased their page length over the last two decades, a development that has come at a time when they are dominantly sold through general market bookstores rather than specialized comic shops. As the combination of Fig. 4 and Fig. 5 indicates, this has led to lower textual complexity, both in relation to page length and during the last decade. A possible reason could lie in the attempt to appeal to readers that are not used to the complex text-image combinations found in comics and might struggle with the cognitive demands placed on them as a consequence. However, more detailed study will be needed to support this or other potential explanations.

Genre affiliation has been shown to constitute an influential determinant for the visual style of comics [9]. The analysis of textual complexity shows this to be the case for its linguistic components as well. Figure 6 compares four large groups represented in the GNC: graphic fantasy, an umbrella term for the genres of superhero, science fiction, fantasy, and horror narratives; graphic memoirs; graphic novels in the narrow sense of literary fiction in the comics medium; and other non-fiction, specifically graphic journalism, travel narratives, and historical non-fiction. It is this last group which records the highest textual complexity. Graphic non-fiction is followed by graphic memoirs. Somewhat surprisingly, the fictional genre of the graphic novel features the second-lowest textual complexity, a result that goes against common assumptions about the particular complexity of more prestigious texts. However, Jannidis, Konle, and Leinen had already found that literary novels are not necessarily more complex linguistically than dime novels [13]. Our results support this view and extend it to book-length comics. Less

surprising, perhaps, is the complexity score of the popular genres classed as graphic fantasy: These are narratives of visual spectacle and show a significantly lower investment in complex texts.

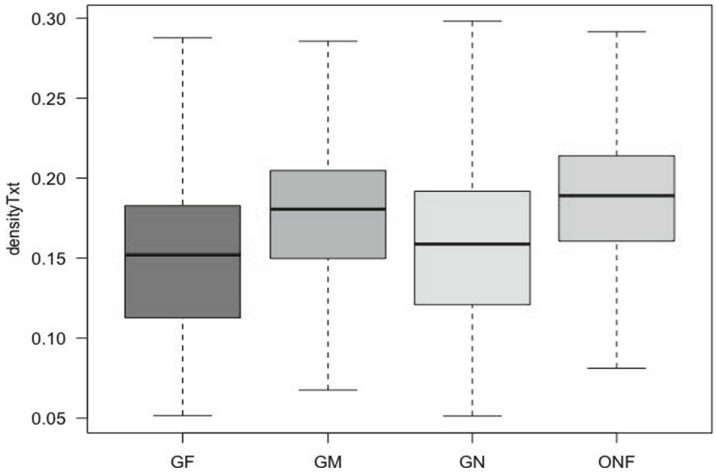

Fig. 6. Textual Density scores by macro genres in the GNC

5 Conclusion and Future Work

In this paper, we presented an OCR pipeline that builds on prior work on text spotting for comics and the open-source OCR software Calamari, which was then trained on the GNC. The results show that optical character recognition for comics and graphic novels has now reached a point where it can lead to meaningful analysis of comics text beyond manually-annotated datasets. Given the specific challenges of comics as a document type, this pipeline involves significant post-processing steps and currently excludes onomatopoeia and diegetic background text. The significant error rates and the exclusion of almost half of the corpus even with the comparatively low threshold chosen for this paper also indicate that much work remains to be done. This future research includes training the neural network architecture employed for text spotting in this paper to recognize onomatopoeia and diegetic text as additional classes. Further training with larger training sets, time-intensive manual correction of OCR results, and the refinement of available OCR software are likely to form part of future improvements in text recognition for comics as well. Beyond these immediate steps, the production of synthetic training data may offer a way to overcome copyright restrictions and the limited representativeness of existing but small data sets.

As this discussion indicates, a significant factor in these on-going challenges continues to be the limited availability of publicly accessible datasets. Unfortunately, the GNC is no exception in this regard, given that all included works are protected by copyright. In addition to prohibiting the dissemination of the retro-digitized scans, these restrictions

also mean that we currently cannot share the OCR model developed for this paper. However, variants of semantic text analysis may proceed without full texts of copyrighted works becoming available if detailed statistical information is provided on a page-by-page level instead. Such text segments are usually chosen on the basis of word length for print literature such as novels or plays. The fact that individual pages represent a stable, basic unit for comics makes identification of text segments, and comparison with the visual components of specific pages, far easier. Such data is not subject to copyright restrictions because the original works cannot be reconstructed on this basis and therefore represents one avenue of sharing research results. The data provided as part of this paper makes a first step in this direction. Given the needs of different communities within computer science and related disciplines, for instance in graphic document analysis, machine learning, and the digital humanities, this can only be a partial solution. Others will have to be found for research on copyrighted comics to advance further in the coming years.

References

1. Laubrock, J., Dunst, A.: Computational approaches to comics analysis. Top. Cogn. Sci. **12**(1), 1–37 (2020). https://doi.org/10.1111/tops.12476
2. Rigaud, C., Burie, J., Ogier, J.: Segmentation-free speech text recognition for comic books. In: 14th IAPR International Conference on Document Analysis and Recognition, vol. 3, pp. 29–34. IEEE, Los Alamitos, CA (2017). https://doi.org/10.1109/ICDAR.2017.288
3. Unser-Schutz, G.: Influential or influenced? the relationship between genre, gender and language in manga. Gend. Lang. **9**, 223–254 (2015). https://doi.org/10.1558/genl.v9i2.17331
4. Walsh, J.A., Martin, S., St. Germain, J.: The spider's web: an analysis of fan mail from amazing spider-man, 1963–1995. In: Dunst, A., Laubrock, J., Wildfeuer, J. (eds.) Empirical Comics Research: Digital, Cognitive, and Multimodal Methods, pp. 62–84. Routledge, New York (2018)
5. Hartel, R., Dunst, A.: How good is good enough? establishing quality thresholds for the automatic text analysis of retro-digitized comics. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.-H., Vrochidis, S. (eds.) MMM 2019. LNCS, vol. 11296, pp. 662–671. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05716-9_59
6. Dubray, D., Laubrock J.: Multi-class semantic segmentation of comics: a U-Net based approach. In: Graphics Recognition (GREC) Workshop, International Conference on Document Analysis and Recognition (ICDAR), Sydney, pp. 5–6 (2019)
7. Wick, C., Reul C., Puppe F.: Calamari—A High Performance Tensorflow-based Deep Learning Package for Optical Character Recognition. https://arxiv.org/ftp/arxiv/papers/1807/1807.02004.pdf
8. Dunst, A., Hartel, R., Laubrock, J.: The graphic narrative corpus (GNC): design, annotation, and analysis for the digital humanities. In: 2nd International Workshop on coMics Analysis, Processing, and Understanding, 14th IAPR International Conference on Document Analysis and Recognition, Kyoto, Japan (2017). https://doi.org/10.1109/ICDAR.2017.286
9. Dunst, A., Hartel, R.: The quantitative analysis of comics: towards a visual stylometry of graphic narrative. In: Dunst, A., Laubrock, J., Wildfeuer, J. (eds.) Empirical Comics Research: Digital, Multimodal, and Cognitive Methods, Chap. 12, pp. 239–263. Routledge, New York (2018)

10. Nguyen, N.-V., Rigaud, C., Burie, J.-C.: Multi-task model for comic book image analysis. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.-H., Vrochidis, S. (eds.) MMM 2019. LNCS, vol. 11296, pp. 637–649. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05716-9_57
11. Michalke, M.: koRpus. An R package for text analysis. https://reaktanz.de/?c=hacking&s=koRpus (2020)
12. Breuel, T.M.: The OCRopus open source OCR system. DRR (2008). https://doi.org/10.1117/12.783598
13. Jannidis, F., Konle, L, Leinen, P.: Makroanalytische Untersuchung von Heftromanen. In: Sahle, P. (ed.) DHd 2019 Book of Abstracts, pp. 167–173 (2019). https://zenodo.org/record/2596095
14. Jones, E., Nulty, P.: Quantitative measures of lexical complexity in modern prose fiction. Digit. Scholarsh. Hum. **34**, 914–937 (2019). https://doi.org/10.1093/llc/fqz020
15. Kirtley, C., Murray, C., Vaughan, P.B., Tatler, B.W.: Reading words and images: factors influencing eye movements in comics reading. In: Empirical Comics Research [9], pp. 264–283, Routledge, New York (2018)
16. Liddle, D.: Could fiction have an information history? statistical probability and the rise of the novel. J. Cult. Anal. (2019). https://doi.org/10.22148/16.033