Journal of Open
Humanities Data

Drama Critiques' Database

DATA PAPER

]u[ubiquity press

MYLÈNE MAIGNANT
DAMIEN PELLÉ
GAËTAN BRISON
THIERRY POIBEAU
*Author affiliations can be found in the back matter of this article

ABSTRACT

Drama Critiques' dataset was produced as part of Mylène Maignant's PhD project at the Ecole Normale Supérieure (Ulm) between 2018 and 2022. Her research aims to explore the reception of contemporary London theatre (2010 to 2020) by analysing a large corpus based on the reviews written by two distinct groups of theatre critics: the journalistic one on the one hand, and the digital one on the other hand (bloggers). By relying partly on the data made available by *Theatre Record* (https://www.theatrerecord.com/) and by automatically extracting the content of 28 blogs, we built a corpus constituting more than 40,000 theatre reviews. Only 36,000 are open access, as some of the bloggers have not given their authorisation yet. The purpose of this data collection consisted of exploring the similarities and differences between these two literary communities. We were interested in better understanding the cultural discourse both journalists and bloggers construct. Given the amount of data, we relied on digital technologies to investigate this field. Using various digital techniques, such as computational linguistics, sentiment analysis, and Geographic Information Systems, we conducted a number of different analyses to map this cultural phenomenon. If some publications have already tackled the literary theme of English digital theatre criticism, none of them have examined it from a computational perspective. Drama Critiques' dataset is then the first corpus which not only offers so many contemporary reviews based on journalists and bloggers' publications, but which also proposes a study of its content (https://doi.org/10.5281/zenodo.6799656).

Repository location: 10.5281/zenodo.6799656.

CORRESPONDING AUTHOR:
Mylène Maignant
Laboratoire LATTICE, CNRS & ENS-PSL & Université Sorbonne nouvelle, Montrouge, France
mylene.maignant@ens.psl.eu

KEYWORDS:
Theatre; Dataset; Journalism; Criticism; Visualisation; Blog

TO CITE THIS ARTICLE:
Maignant, M., Pellé, D., Brison, G., Poibeau, T. (2022). Drama Critiques' Database. *Journal of Open Humanities Data*, 8: 21, pp. 1–14. DOI: https://doi.org/10.5334/johd.81

1. CONTEXT AND MOTIVATION – DRAMA CRITIQUES: EXAMINING LONDON THEATRE CRITICISM THROUGH COMPUTER SCIENCE

1.1 DIGITAL HUMANITIES AND THEATRE STUDIES: STATE-OF-THE-ART

While projects in Digital Humanities have grown at a very fast pace during the last 10 years, theatre studies are still underrepresented in this field. In order to better understand the theoretical and practical contours of this discipline to which this paper belongs, we rely on the arguments proposed by Bardiot (2017). The three sections she identifies are particularly relevant to situate this project within the community of theatre studies and Digital Humanities.

The first category deals with projects that examine theatre history as a global phenomenon thanks to computational methods. Miller's analysis of Broadway theatres is probably one of the most emblematic examples of this approach. Basing his study on the Internet Broadway Database[1] and on the Playbill Vault dataset,[2] Miller reinterprets a century of the history of the great Broadway shows through a socio-economic perspective (Miller, 2016). On a different level, IbsenStage[3] is another case in point regarding this category. IbsenStage has been developed in co-operation with AusStage, The National Library of Norway, as well as the University of Oslo. IbsenStage is a quantitative research tool which has gathered more than 20,000 records on the reception of Henrik Ibsen's plays from 1850 to the present day. This massive collection of data gave birth to two publications at the crossroad between Computational Sciences and Sociological/Literary theories. While the first one focused on only one play (Holledge et al., 2016), the other examined a number of Ibsen's plays in a specific spatio-temporal framework (Hanssen, 2020).

The second category Bardiot identifies concerns projects that focus on the texts of theatre plays. The studies of Shakespeare's masterpieces, thanks to computational stylistics, are the most well-known examples. Stylometry has been employed to determine whether some of Shakespeare's plays were written by him alone or were the products of a collaboration of several authors (Craig & Kinney, 2009). Another recent example in this category is The Samuel Beckett Digital Manuscript Project.[4] By digitizing all of Beckett's drafts, reading notes, translations, and plays, Mark Nixon and Dirk Van Hulle contributed to building a wider understanding of the Beckettian heritage (Van Hulle et al., 2016). An important piece of work based on the Text Encoding Initiative (TEI)[5] has also been carried out to share these documents with a greater audience (Van Hulle et al., 2016).

Finally, the third category encompasses projects that focus on the interaction with visual data. The Simulated Environment for Theatre (SET),[6] for instance, is software which enables one to work simultaneously with the text of one play and with the representation of a stage (Roberts-Smith et al., 2013). Rekall[7] is another open-source environment which facilitates the process of documenting, annotating, and sharing visual archives (Bardiot et al., 2014).

Drama Critiques is situated at the intersection between the first and the second categories. Similar to projects such as Visualizing Broadway[8] or IbsenStage, statistics and literary theories are associated in order to model a part of theatre history to provide quantified answers to literary questions. This popular approach in Digital Humanities takes root in what Moretti (2013) calls "distant reading":

> The reality of the text undergoes a process of deliberate reduction and abstraction. 'Distant reading', I have once called this type of approach; where distance is however not an obstacle, but a *specific form of knowledge*: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models (Moretti, 2013, p. 1).

Far from replacing humans, computational techniques helped us organise, sort, and analyse the data we had collected. They enabled us to have an overall vision of the various connections between

[1] https://www.ibdb.com/ (last accessed: 25 July 2022).

[2] https://search.library.wisc.edu/database/UWI50183 (last accessed: 25 July 2022).

[3] https://ibsenstage.hf.uio.no/ (last accessed: 25 July 2022).

[4] https://www.beckettarchive.org/ (last accessed: 25 July 2022).

[5] https://bdmpmanual.uantwerpen.be/ (last accessed: 25 July 2022).

[6] http://www.arts.uwaterloo.ca/~j33rober/set.html (last accessed: 25 July 2022).

[7] http://www.rekall.fr/ (last accessed: 25 July 2022).

[8] https://visualizingbroadway.com/ (last accessed: 25 July 2022).

journalists and bloggers. For some of our observations, we also relied on "close reading" by using technical tools. To study the structure of the reviews, for instance, we annotated 1,000 reviews manually to train an algorithm to label the rest of the corpus automatically (Maignant et al., 2021). Hence why this project is situated in the first and the second categories Bardiot identifies. We studied this dataset both from macro and micro perspectives, thanks to computational techniques. Our corpus can be found on our Zenodo repository: https://doi.org/10.5281/zenodo.6799656.

1.2 REDEFINITION OF CLASSIC THEATRE CRITICISM IN THE DIGITAL ERA

Our project deals with London theatre criticism from 2010 to 2020. Whether it is in literature, theatre, or cinema, critics are often represented as parasites. Austere, smug, and frustrated, they are those shadowy figures who judge and moralise artworks at dusk. However, the etymology of the word reminds us that this profession does not consist of destroying the production of others, but of analysing it. "To criticize" means "to discern, to separate, to sort".[9] The purpose of the critic is to select some characteristics of a cultural event in order to dissect it, "to sieve" it, as its Greek etymology indicates, and to give an analysis of it.

The end of the 20th century marks a shift in the role and the status of the critic. The arrival of the Internet redefined the modalities of expression of theatre critics. In 1997, the first two theatre review websites were created: British Theatre Guide[10] and What's On Stage[11] (Radosavljevic, 2016). In the 2010s, a myriad of blog platforms emerged thanks to the popularity of digital tools such as WordPress[12] and Blogger.[13] They enabled anyone without any skills in programming to start a blog platform. Students, theatre professionals, but also mere amateurs began publishing their own theatre reviews online.

A diversity of independent voices from multiple backgrounds flourished in the landscape of theatre criticism. Published by independent authors, these reviews tend to focus on marginal theatres and plays that are less frequently represented by journalists. These digital platforms share the promulgation of amateur writers who are not paid for their activity but who keep writing out of passion for theatre.

The platform titled *A Younger Theatre* is quite explicit in this regard. It is about opening a digital space for younger generations to give a fresh and critical opinion on the English contemporary stage. The affirmation of their digital identity is clear:

> [A Younger Theatre is] a platform for those who are often unheard. We champion the emerging generation. We celebrate excellence and pride ourselves on professionalism, while also leaving room for risk, failure and learning.[14]

The name of other blog platforms, such as View from the Cheap Seat[15] or Partially Obstructed View,[16] highlight the fact that the amateur does not necessarily have a reasonably good seat while a professional will. Since bloggers are not paid to attend shows, their placement in the auditorium is affected. Their experience of the reception of one play is then necessarily altered. Someone placed in front of a pole or at the back of the hall will not see or hear a theatre show in the same way as someone sitting close to the stage.

1.3 POLARISATION BETWEEN THE LITERARY CANON (JOURNALISTS) AND ITS MARGINS (BLOGGERS)

The emergence of this blogosphere provoked sharp debates within the critics' community. For Michael Billington, who is a long-standing critic working for *The Guardian*, a blog is more an "informal letter" than a true review (Billington, 2007). Danielle Tarento, co-founder of the Menier

[9] https://www.larousse.fr/encyclopedie/divers/critique/187226 (last accessed: 25 July 2022).

[10] https://www.britishtheatreguide.info/ (last accessed: 25 July 2022).

[11] https://www.whatsonstage.com/ (last accessed: 25 July 2022).

[12] https://wordpress.com/ (last accessed: 25 July 2022).

[13] https://www.blogger.com/dashboard/reading (last accessed: 25 July 2022).

[14] Description of the page "About" from the Website https://www.ayoungertheatre.com/. Retrieved from https://www.ayoungertheatre.com/about/ (last accessed: 2 February 2022).

[15] https://viewfromthecheapseat.com/ (last accessed: 25 July 2022).

[16] http://partially-obstructed-view.blogspot.com/ (last accessed: 25 July 2022).

Chocolate Factory theatre,[17] goes as far as to claim that bloggers are not genuine writers as "they do not have the intellectual background or historical background or time to know what they are writing about" (Hemley, 2016). The provocative title of Ronan McDonald's essay, *The Death of the Critic* (2007), embodies the core of this controversy. On the one hand, some of the professional reviewers who write for the most popular UK newspapers deny the legitimacy of the bloggers. On the other hand, these new voices in the digital space demand their speaking right.

Above all, these tensions shed light on communities that seem to consider theatre from two different perspectives. The staging of two plays crystallises the polarity of these debates. In 2007, the technicality of the staging of Martin Crimp's play *Attempts on her Life* (1997), performed at the National Theatre and directed by Katie Mitchell, was hailed with admiration by the blogosphere. Conversely, the journalistic critics received it with great hostility. For Georgina Brown, critic for the *Mail on Sunday*, "[that was] the worst play [she had] ever seen" (Vaughan, 2020, p. 44). Mark Shenton, chief theatre critic of the *Sunday Express*, wrote that he was "seriously contemplating making an attempt on [his] own life" (Vaughan, 2020, p. 44). Five years later, Simon Stephens' play *Three Kingdoms* (2012), performed at the Lyric Hammersmith and directed by Sebastian Nübling, provoked a similar wave of dissent.

This polarisation highlights the dichotomy between the centre of theatre criticism and its peripheries. While authors' voices who belong to the canon, such as broadsheet papers like *The Guardian, The Independent*, or *The Times* are easily heard, bloggers still struggle to gain visibility. Their presence is nevertheless necessary. As the blogger Megan Vaughan articulates it:

> It does kinda feel like those of us working on the Internet have a responsibility to exercise all the freedoms it gives us to play with words and structure and form, because criticism should be a LANDSCAPE. Digital criticism, for me, is the freedom to be different, but implicit in that is an obligation to be different, for the sake of a healthy culture of discourse, now and in the future (Radosavljevic, 2016, p. 23).

The purpose of this project thus lies in better understanding how these two communities interact. In other words, what kind of cultural discourse do they construct? Is the opposition between the centre and its peripheries reflected in their writing, in their aesthetic preferences, or in the theatres they tend to go? And which computational techniques shall we use to model these differences and similarities?

2. METHOD – CREATING AND STRUCTURING THE DATABASE: GATHERING INFORMATION FROM TWO DIFFERENT TYPES OF SOURCES

2.1 JOURNALISTIC CRITICISM: THEATRE RECORD, FROM THE PDF FORMAT TO THE TEXTUAL FILE

Our first sub-corpus was created thanks to the online database *Theatre Record*[18] and can be found on our Zenodo repository at this address: https://zenodo.org/record/6799656#.YsSEFYTP25c. *Theatre Record* was originally a biweekly paper magazine created by Ian Herbert in 1981 which reprints, in full, all the national drama reviews of the productions in London and its regions. Its archives were digitised in 2019 by Julian Oddy. Each newspaper published is now available online in PDF format for those who have a subscription to *Theatre Record*. All the newspapers' issues have the same characteristics. For each of the shows, a certain number of reviews is given, accompanied by details about the production, such as the cast, the credits, and the photographs. The theatre in which the play was performed, the opening and closing dates of the show, the director, and the theatre company are recorded as well.

Out of the 84 newspapers available on *Theatre Record*, we selected 23 of them in total. Since this corpus focuses on printed newspapers only, online news websites were excluded. We also removed newspapers whose reviews were not about London performances and all the newspapers which had too small a number of reviews (25 reviews). All the selected newspapers are well-known among the general public.[19]

[17] https://www.menierchocolatefactory.com/ (last accessed: 25 July 2022).

[18] https://www.theatrerecord.com/ (last accessed: 25 July 2022).

[19] For the complete list of the 28 newspaper and blog platforms, see Appendix A.

The first step of our work consisted of converting the images of the PDF files into a textual format. While this task seemed to present no obstacle at first, it turned out to be a time-consuming exercise since the reviews were arranged in columns. This particular layout complicated the process of conversion for Optical Character Recognition softwares. After testing a number of solutions (Onlineocr, GROBID (GeneRation of BIbliographic Data), Transkribus, Kraken, and the package Pdftools in R), we chose ABBYY FineReader.[20] FineReader is an Optical Character Recognition software that converts images of PDFs into a wide range of formats (.txt, .png, .csv, etc.) in more than 10 different languages. The conversion took about 10 minutes per file, which represents 38 hours in total since we had 228 PDF files. The software was able to detect the relationships between the columns and so retain the proper reading sequence of the text.

Once each of the PDF files were converted, we separated the metadata from the body of the text using the text editor Sublime Text.[21] We then started cleaning the dataset. We first used regular expressions with Python to clean the reviews automatically. This task included the removal of the number of pages and email addresses. We then spent a considerable amount of time cleaning the rest of the dataset manually. This part included the removal of unnecessary information within the text of the reviews (such as the opening and closing dates of the show, acknowledgments, URL links, etc.) and the uniformisation of the metadata. Finally, cleaned files were integrated into a Tidy Text Format, that is, a table in which each variable is a column and each observation is a row. More than one year in total was necessary to build this corpus, and about 1,050 hours of work. Figure 1 summarises all the necessary steps we went through to create this first sub-corpus.

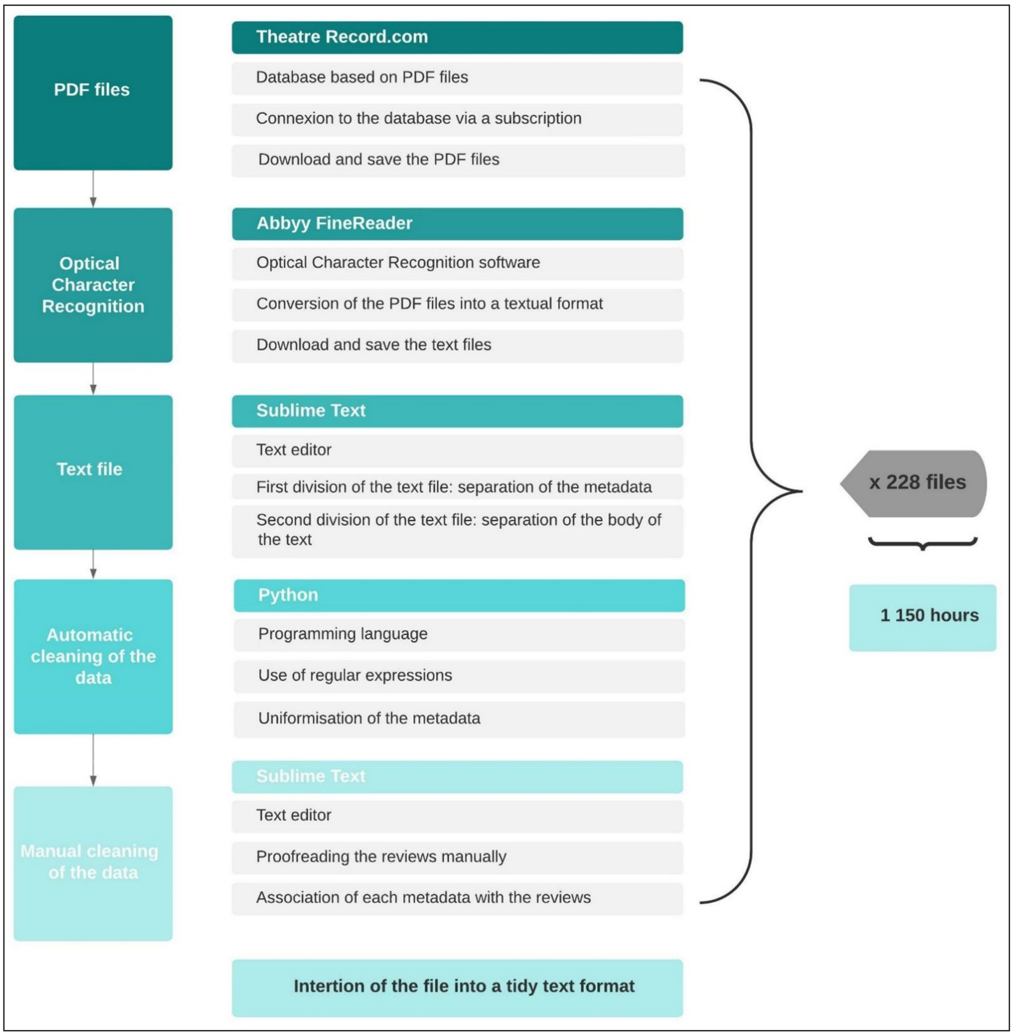
PDF files
Optical Character Recognition
Text file
Automatic cleaning of the data
Manual cleaning of the data
Theatre Record.com
Database based on PDF files
Connexion to the database via a subscription
Download and save the PDF files
Abbyy FineReader
Optical Character Recognition software
Conversion of the PDF files into a textual format
Download and save the text files
Sublime Text
Text editor
First division of the text file: separation of the metadata
Second division of the text file: separation of the body of the text
Python
Programming language
Use of regular expressions
Uniformisation of the metadata
Sublime Text
Text editor
Proofreading the reviews manually
Association of each metadata with the reviews
Intertion of the file into a tidy text format
x 228 files
1 150 hours

Figure 1 Diagram summarising the different steps for the creation of the first sub-corpus.

[20] https://pdf.abbyy.com/fr/ (last accessed: 25 July 2022).

[21] https://www.sublimetext.com/ (last accessed: 25 July 2022).

2.2 DIGITAL CRITICISM: THE BLOGOSPHERE, WEB SCRAPING THE CONTENT OF THE BLOGS

The second sub-corpus is constituted by the 28 most popular English blog platforms whose authors' publications deal with London plays only.[22] The selection of these blogs was based on two websites: the top 10 of the most visited blogs in the UK established by Vuelio in 2020,[23] and by the platform MyTheatreMates.[24] MyTheatreMates.com was co-founded by Mark Shenton and Terri Paddock, two professional reviewers. In order to give more visibility to the peripheries of English theatre criticism, they created this web platform to enable bloggers' voices to be heard. The following four conditions are required if one wants their review to be published on this blog platform:

1. "You have your own personal website"
2. "You post original theatre-related content on your personal website at least once a fortnight"
3. "You can provide three professional arts references (e.g. artists you have interviewed or, if you review, producers or publicists who already regularly provide you with complimentary press tickets to shows)"
4. "You are active on Twitter"[25]

Whether these blog platforms come from Vuelio or MyTheatreMates.com, they all have the same characteristics: they have no printed equivalent, their content is entirely free, and their authors are not paid for their activity.

The content of these websites was extracted with Python web scraping techniques (Mitchell, 2015). Web scraping consists of writing a script which creates an artificial user who will automatically copy and paste the HTML information of a web page. We wrote 28 different scripts in total, as each of the structures of the blog platforms were different. Once the data was collected, the same work of cleaning, uniformisation, and structuring the reviews was carried out on this second sub-corpus. While a part of it could be cleaned automatically with regular expressions using Python, the other part had to be read manually. Creating this corpus based on digital reviews was less time-consuming than the first one because the reviews were already in a textual format. However, it still represented more than 250 hours of work in total. Figure 2 illustrates all the steps taken to create the second sub-corpus.

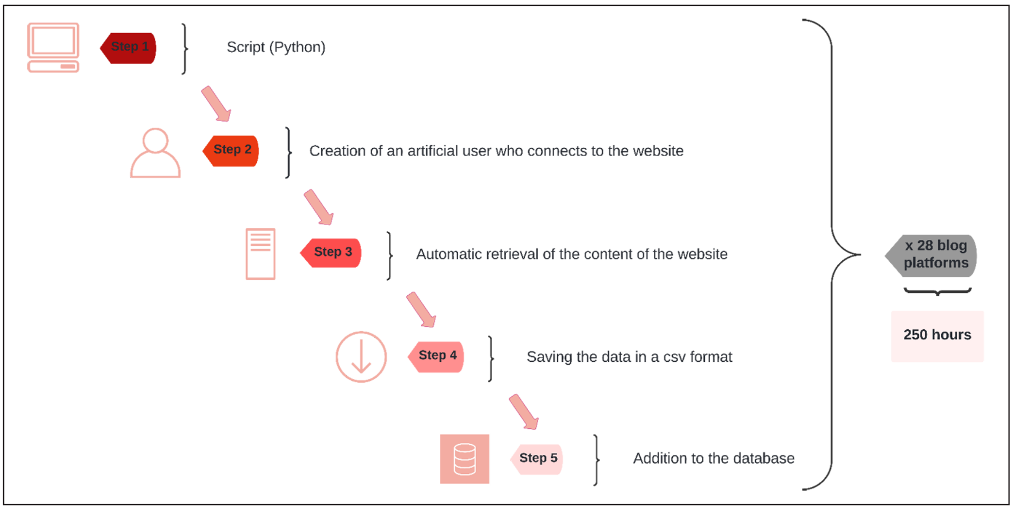
Step 1
Script (Python)
Step 2
Creation of an artificial user who connects to the website
Step 3
Automatic retrieval of the content of the website
Step 4
Saving the data in a csv format
Step 5
Addition to the database
x 28 blog platforms
250 hours

Figure 2 Diagram summarising the different steps for the creation of the second sub-corpus.

2.3 ENRICHING THE CORPUS WITH VARIOUS METADATA

We also enriched the corpus with metadata to broaden the possibilities of exploration and exploitation of the corpus. Here is the list of all the metadata we added to our file:

[22] To find the complete list of the 28 blog platforms, see Appendix A.

[23] https://www.vuelio.com/uk/social-media-index/top10-theatre-blogs/ (last accessed: 25 July 2022).

[24] https://mytheatremates.com/ (last accessed: 25 July 2022).

[25] Description of the page "Want to join MyTheatreMates" from the Website https://mytheatremates.com/. Retrieved from https://mytheatremates.com/about-us/authors/ (last accessed: 02 February 2022).

- The name of the newspaper/blog
- The creation date of the newspaper/blog
- The publication date of the review
- The name and the gender (masculine, feminine, or androgynous) of the reviewer, which was automatically determined with the package "genderizeR"
- The title of the play
- The name of the playwright
- The star rating of the review (where this information was available)
- The theatre in which the play was performed
- The URL of the website of the theatre
- The type of venue[26]
- The seating capacity of the theatre (where this information was available)
- The coefficient of each theme within each review[27]
- The coefficient of positivity, neutrality, and negativity within each review which was calculated with the algorithm XLNet
- The coefficient of subjectivity, anger, fear, happiness, sadness, and surprise within each review which was calculated with the algorithm text2emotion

3. RESULTS AND DISCUSSION – THE FIRST CORPUS TO GATHER NUMEROUS JOURNALISTIC AND DIGITAL THEATRE REVIEWS: A NEW DATASET IN OPEN ACCESS IN THE LANDSCAPE OF DIGITAL HUMANITIES

3.1 MAIN CHARACTERISTICS OF THE OVERALL CORPUS

Table 1 summarises the main characteristics of our corpus. A more detailed version is available on Zenodo: https://zenodo.org/record/6799656#.YsSEFYTP25c.

Table 1 Main characteristics of the Drama Critiques' corpus.

	CORPUS I JOURNALISTIC CRITICISM	CORPUS II DIGITAL CRITICISM	CORPUS I AND CORPUS II
Source	PDF	HTML	PDF/HTML
Time creation (hrs)	1,150	250	1,400
Number of newspapers/blogs	23	28	51
Number of reviewers	454	655	1,069
Number of reviews	21,717	21,326	43,043
Number of words	7,689,704	11,416,428	19,106,132

3.2 ANALYSIS OF THREE RESEARCH FIELDS AT THE CROSSROAD BETWEEN ALGORITHMS AND LITERATURE

The creation of this corpus enabled us to conduct three experiments to map the relationship between the canon of theatre criticism (journalistic reviews) and its peripheries (digital reviews). These three experiences correspond to three literary questions associated with three technical fields. The first one focused on computational stylistics. We wanted to understand the extent to which the literary style of these authors differs from one corpus to another. We relied on Holmes' (1985) definition of literary style, who defines it as an ensemble of "several variables which may be used as stylistic fingerprints" (p. 1). We found out that bloggers tend to prone a

[26] For the complete list of 24 type of venue categories, see Appendix B.

[27] For the complete list of the 20 themes, see Appendix C. These coefficients were calculated by counting the number of synonyms related to the name of the theme.

more subjective way of writing, since the pronoun "I" appears twice as much in their reviews than in the ones written by journalists.

Our analysis of the structure of the reviews also confirmed this idea. We worked on text zoning, a method which consists of tagging sentences to reveal text structures. We manually annotated 1,000 random reviews according to the eight different categories that constitute a review: the introduction, the description of the plot, the performance of the actors, the audio and visual details of the stage, the observations of the audience, the remarks about the literary structure of one performance, the reviewer's subjective analysis, and the conclusion. After having trained an algorithm to label the rest of the corpus automatically, we realised that bloggers tended to focus on categories related to affect.[28] "Visual and audio details", "performance of actors", and remarks "related to the audience", were represented more in the corpus based on digital reviews. The common characteristic of these three categories lies in the fact that they all put to the front the human aspects of one play. These results appear to confirm the idea that bloggers seek to affirm their own voice. The recurrent use of the pronoun "I" also tends to confirm this trend. While journalists position themselves as objective observants, bloggers favour a more personal and subjective approach.

The second experiment was based on sentiment analysis. We wanted to examine the aesthetic preferences of these two communities to better understand which kind of plays they favoured. We focused on the notion of "success", which can be understood in two ways. The first definition deals with the number of times a play was seen. It was noteworthy that in this case, both journalists and bloggers mostly attended Shakespeare's adaptations (*A Midsummer Night's Dream, Macbeth, Hamlet*, among others). In other words, both groups attended the classics of English literature. The second definition of success included the percentage of positivity calculated by the algorithm. Except for one play (*Rosenbaum's Rescue* (2019), played at the Park Theatre and directed by A. Bodin Saphir), all the reviews that had obtained the highest score of positivity were praised for the sense of intimacy they conveyed. Counter to our assumptions, no real gap separated the canon from its peripheries regarding their aesthetic preferences. These results led us to conclude that bridges exist between these two communities, despite what they affirm.

Finally, we explored the dichotomy "centre versus margins" from a geographical point of view. By relying on Geographic Information Systems (GIS), we created a digital map which enabled us to visualise the reviewers' journeys through London. This experiment led us to conclude that bloggers covered a wider spectrum of places, with journalists having visited 210 different places in London from 2010 to 2020, and bloggers 446 different places. The types of places also varied from one corpus to another. By classifying them into 24 different categories (e.g., theatres, pubs, academic or cultural institutions, concert halls),[29] we found out that bloggers and journalists have different preferences. Religious places, for instance, are absent from the places visited by bloggers, whereas they are the most frequented places by journalists. On the contrary, places like museums, art associations, and bowling alleys are popular among the digital community. Different from the theatrical institutions, we can guess that these places offer a wider variety of shows. This map and this classification thus make it possible to deduce that bloggers seem to cover a wider field of theatres, and probably honour a more diversified culture.

3.3 AN ORIGINAL YET LIMITED CONTRIBUTION

Although we are proposing a completely new corpus, some of its aspects could be improved. One of the limits of Drama Critiques' dataset lies in the limited number of reviews that are made available in open access. The initial corpus is constituted of more than 43,000 theatre reviews, whereas only 36,000 are freely accessible. This is due to the fact that some bloggers have not given us their authorisation yet. Table 2 summarises the main characteristics of the current version available on our website and on Zenodo.[30]

[28] The scripts of this experiment are available on our GitHub repository: https://github.com/MyleneM/ML_Project.

[29] For the complete list of the 24 categories, see Appendix B.

[30] https://zenodo.org/record/6799656#.YsSEFYTP25c.

	CORPUS I IN OPEN ACCESS JOURNALISTIC CRITICISM	CORPUS II IN OPEN ACCESS DIGITAL CRITICISM	CORPUS I AND II IN OPEN ACCESS
Number of newspapers/blogs	23	22	45
Number of reviewers	454	420	874
Number of reviews	21,717	15,048	36,765
Number of words	7,689,704	7,771,451	15,461,155

Table 2 Main characteristics of the Drama Critiques' corpus in open access.

4. IMPLICATIONS/APPLICATIONS – PRESENTING AND SHARING THE DATASET

4.1. IMPLEMENTATION OF A SEARCH ENGINE

In order to make our database available to all, we implemented a search engine based on SQL language within our website.[31] We added a number of filters so that users can access our data more easily. Below is the list of the filters we implemented:

- Journalistic/digital corpus
- Name of the newspaper/blog
- Date of publication of the review
- Reviewer's gender
- Reviewer's name
- Title of the play
- Location of the play
- Themes of the play

Once one or multiple options are selected, a list of all the corresponding reviews appear on another page. For each of the reviews, except the themes of the play, all the options mentioned above are indicated. Whether users are looking for the reviews written by women only or if they want to discover plays dealing with the theme "Family", the interface enables them to navigate quickly among the data.

4.2 "BEHIND THE MACHINE": THE ONLINE APPLICATION TO BRIDGE THE GAP BETWEEN "DIGITAL" AND "HUMANITIES"

We also considered the pedagogical aspect when we created this dataset. An online application describing the different steps to analyse a text from a computational perspective has been implemented.[32] This project emerged after having presented Drama Critiques at a couple of conferences related to literature only. For those who do not have a technical background, "reading" a theatre review with a computer can be a complicated task to understand. The use of new technologies can sometimes raise questions such as "Is the computer going to replace human beings?" or "What do you mean by 'algorithm'?"

It is to answer these questions that we designed "Behind the Machine". Composed of multiple sections, the user starts by choosing a journalistic review within a table on the first page. The dashboard placed on the left of the page displays the different and necessary steps to make the raw text readable by a computer (e.g., tokenisation, lemmatisation). By clicking on each of these categories, the selected review appears with both the explanation and the illustration of the chosen process.

For the section "Named Entity Recognition" for instance, a small text reads:

> In Natural Language Processing, Named Entity Recognition (NER) is a process where a sentence or a chunk of text is parsed to find entities that can be put under categories like names, organizations, locations, quantities, monetary values, percentages, etc. Traditional NER algorithms include only names, places, and organizations.[33]

[31] https://dramacritiques.com/en/database/.

[32] https://dramacritiques.com/en/behind-the-machine/.

[33] https://deepai.org/machine-learning-glossary-and-terms/named-entity-recognition (last accessed: 25 July 2022).

An image with different labels (e.g., date, place, name) on the words appear so that the users can clearly understand what the process of Named Entity Recognition entails. By unravelling the mechanisms at work behind the machine, we hope that it will help democratise computational methods in the humanities.

4.3 OPENING RESEARCH POSSIBILITIES FOR COMPUTER SCIENTISTS AND ACADEMICS IN HUMANITIES

Finally, a part of Drama Critiques' dataset and all the programming scripts that enabled us to carry out the technical analyses are in open access on GitHub[34] and Zenodo.[35] Anyone can thus run the algorithms on the whole corpus again to better understand the results or adapt them to their own data. Sharing the database and the scripts enables us to ensure the transparency of our data and our results. It also enables us to contribute to the field of English literature by proposing the first reusable dataset to offer numerous theatre reviews on journalistic and digital criticism.

This corpus could be useful for data scientists for a number of reasons. Testing Natural Language Processing algorithms, training machine learning models, or building new digital tools based on language, requires a lot of clean and structured data. Furthermore, the geolocation of each of these theatres also paves the way for anyone who would be interested in GIS to use this dataset to further explore this domain. Whether it is to examine cultural institutions in England or to create any new geographical applications, data are needed.

Concerning the literary aspect of this project, many possibilities remain open to any academics eager to examine reception/theatre studies in greater depths. We have already shed light on three different technical fields (computational stylistics, sentiment analysis, and GIS) corresponding to three different literary questions ("How do these authors write?", "What are their aesthetic preferences?", "Where do they go in London?"). However, one could either further investigate these domains (apply other sentiment analysis models or improve the map, for instance), or explore new questions related to drama studies.

One could, for example, carry out an analysis of theatre props based on this corpus. One could also think of comparing each newspaper and blog platform with one another to see the differences and similarities within one literary community. Building a second corpus with reviews written by regional newspapers and blog platforms would be a relevant idea as well. It would enable one to compare London criticism with its regions, or with other English-speaking countries. The possibilities are numerous, and only a few of them were listed here.

APPENDICES

Appendix A List of all the newspapers and blog platforms represented in Drama Critiques' Dataset.

NUMBER	NEWSPAPER	BLOG PLATFORM
1	Daily Express	Aleks Sierz Blog
2	Daily Mail	A Younger Theatre
3	Daily Telegraph	Breaking the Fourth Wall Blog
4	Evening Standard	Cultural Capital Blog
5	Financial Times	Everything Theatre Blog
6	Guardian	Exeunt Blog
7	Herald Tribune	Kate in Brockley Blog
8	Independent	London Theatre Review Blog
9	Independent on Sunday	Lou Review Blog
10	Jewish Chronicle	Mind the Blog
11	Mail on Sunday	Monkey Matters Blog
12	Metro London	Musical Theatre Review Blog

(Contd.)

[34] https://github.com/MyleneM.

[35] https://zenodo.org/record/6799656#.YsSEFYTP25c.

NUMBER	NEWSPAPER	BLOG PLATFORM
13	Observer	Ought To Be Clowns Blog
14	Spectator	Partially Obstructed View Blog
15	Stage	Pocket Size Theatre Blog
16	Sunday Express	Rev Stan Blog
17	Sunday Telegraph	Rewrite This Story Blog
18	Sunday Times	Scatter of Opinion Blog
19	Telegraph	Stage Review Blog
20	Time Out	Susan Elkin Blog
21	Times	The Blog of Theatre Things
22	Tribune	The Plays the Thing UK
23	Variety	Theatre Cat Blog
24		View From the Cheap Seat
25		Webcow Girl Blog
26		West End Whingers Blog
27		West End Wilma Blog
28		2nd From Bottom Blog

Appendix B List of the 24 spatial categories represented in Drama Critiques' Dataset.

NUMBER	TYPE OF VENUE
1	Academic Place
2	Arts Association
3	Arts Centre
4	Building
5	Charity Association
6	Cinema
7	Concert Place
8	Cultural Centre
9	Gallery
10	Hotel
11	Library
12	Miscellaneous
13	Museum
14	Natural Space
15	Online
16	Pub
17	Pub Theatre
18	Public Space
19	Reception venue
20	Religious Place
21	Restaurant Area
22	Studio
23	Theatre
24	Town Hall

NUMBER	THEME
1	Body
2	Childhood
3	Cultural Difference/Race
4	Death
5	Disability
6	Education
7	Environment
8	Family
9	Friendship
10	Identity
11	LGBT/Queer/Sexuality/Gender
12	Love
13	Memory
14	Politics
15	Relationship
16	Religion
17	Science
18	Supernatural
19	Violence
20	Women/Feminism

Appendix C List of the 20 themes represented in Drama Critiques' Dataset.

NUMBER	NEWSPAPER	BLOG PLATFORM
1	Daily Express	Aleks Sierz Blog
2	Daily Mail	A Younger Theatre
3	Daily Telegraph	Breaking The Fourth Wall Blog
4	Evening Standard	Cultural Capital Blog
5	Financial Times	Everything Theatre Blog
6	Guardian	London Theatre Review Blog
7	Herald Tribune	Lou Review Blog
8	Independent	Mind the Blog
9	Independent on Sunday	Monkey Matters Blog
10	Jewish Chronicle	Musical Theatre Review Blog
11	Mail on Sunday	Ought To Be Clowns Blog
12	Metro London	Rev Stan Blog
13	Observer	Scatter of Opinion Blog
14	Spectator	Stage Review Blog
15	Stage	Susan Elkin Blog
16	Sunday Express	The Blog of Theatre Things
17	Sunday Telegraph	The Plays the Thing UK
18	Sunday Times	Theatre Cat Blog
19	Telegraph	View From the Cheap Seat
20	Time Out	Webcow Girl Blog
21	Times	2nd From Bottom Blog
22	Tribune	
23	Variety	

Appendix D List of all the newspapers and blog platforms represented in Drama Critiques' open access Dataset.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Drama Critiques' Corpus.** Drama Critiques' corpus available in csv format. DOI: https://doi.org/10.5334/johd.81.s1

ACKNOWLEDGEMENTS

We would like to thank Alison Cook, editorialist at *Theatre Record,* for her great support and her warm enthusiasm. We are also very grateful to each of the 21 bloggers who agreed on having their reviews published. Without their consent, Drama Critiques' open access dataset would not be as rich as it currently is.

We would also like to acknowledge Frédérique Mélanie and Clément Plancq's participation, both engineers at the Lattice Laboratory for their help.

FUNDING INFORMATION

Mylène Maignant is supported by the EUR (Ecole Universitaire de Recherche) Translitteræ (programme "Investissements d'avenir" ANR-10-IDEX-0001-02 PSL and ANR-17-EURE-0025). Thierry Poibeau is supported in part by PRAIRIE (programme 3IA, by the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001) — (PRAIRIE 3IA Institute). Lastly, this research also benefited from the support of the CNRS International Research Network Cyclades (Corpora and Computational Linguistics for Digital Humanities).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

- Mylène Maignant: Project Manager, Data Collector and Data Analyst
- Damien Pellé: Data Collector
- Gaëtan Brison: Data Analyst
- Thierry Poibeau: PhD Advisor

AUTHOR AFFILIATIONS

Mylène Maignant orcid.org/0000-0001-8046-7783
Laboratoire LATTICE, CNRS & ENS-PSL & Université Sorbonne nouvelle, Montrouge, France

Damien Pellé
French Department, National University of Ireland, Galway, Ireland

Gaëtan Brison
HI! PARIS Department, Institut Polytechnique de Paris, Palaiseau, France

Thierry Poibeau
Laboratoire LATTICE, CNRS & ENS-PSL & Université Sorbonne nouvelle, Montrouge, France

REFERENCES

Bardiot, C. (2017). Arts de la scène et culture analytics. *Revue d'historiographie du théâtre: Études théâtrales et humanités numériques*, 4, 11–20.

Bardiot, C., Coduys, T., Jacquemin, G., & **Marais, G.** (2014). Rekall: un environnement open-source pour documenter, analyser les processus de création et simplifier la reprise des œuvres scéniques. *Journées d'Informatique Musicale*, 119–129.

Billington, M. (2007). Who needs reviews? *The Guardian* (online). Retrieved from https://www.theguardian.com/stage/theatreblog/2007/sep/17/whoneedsreviews (last accessed: 02 February 2022).

Craig, H., & Kinney, A. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Crimp, M. (1997). *Attemps on her Life*. London: Faber and Faber.

Description of the page "Want to join MyTheatreMates" from the Website https://mytheatremates.com/. Retrieved from https://mytheatremates.com/about-us/authors/ (last accessed: 02 February 2022).

Hanssen, J-M. (2020). Ibsen on the German Stage 1876–1918. *Ibsen Studies*, *20*(1), 94–100. DOI: https://doi.org/10.1080/15021866.2020.1757298

Hemley, M. (2016). Danielle Tarento: Unpaid bloggers often lack 'intellectual background' to write theatre reviews. *The Stage* (online). Retrieved from https://www.thestage.co.uk/news/danielle-tarento-unpaid-bloggers-often-lack-intellectual-background-to-write-theatre-review (last accessed: 02 February 2022).

Holledge, J., Bollen, J., Helland, F., & **Tompkins, J.** (2016). *A Global Doll's House: Ibsen and Distant Visions*. London: Palgrave Macmillan. DOI: https://doi.org/10.1057/978-1-137-43899-7

Holmes, D. I. (1985). The analysis of literary style – A review. *Journal of the Royal Statistical Society*, *148*(4), 328–341. DOI: https://doi.org/10.2307/2981893

Maignant, M., Brison, G., & **Poibeau, T.** (2021). Text Zoning of Theater Reviews: How Different are Journalistic from Blogger Reviews? [Conference session]. *Workshop on Natural Language Processing for Digital Humanities*, Dec 2021, Sichar, India. hal-03498270.

McDonald, R. (2007). *The Death of the Critic*. London, New York: Continuum.

Miller, D. (2016). Average Broadway. *Theatre Journal*, *68*, 529–553. DOI: https://doi.org/10.1353/tj.2016.0105

Mitchell, R. (2015). *Web Scraping with Python: Collecting Data from the Modern Web*. Sebastopol: O'Reilly Media.

Moretti, F. (2013). *Distant Reading*. London: Verso. 1.

Radosavljevic, D. (2016). *Theatre Criticism*. London: Bloomsbury. DOI: https://doi.org/10.5040/9781472577122

Roberts-Smith, J., Ruecker, S., DeSouza-Coelho, S., Dobson, T., Gabriele, S., Rodriguez, O., Sinclair, S., Akong, A., Bouchard, M., Jakacki, D., Lam, D., & **Northam, L.** (2013). Visualizing theatrical text: From watching the script to the simulated environment for theatre (SET). *Digital Humanities Quaterly*, *7*(3). Retrieved from http://www.digitalhumanities.org/dhq/vol/7/3/000166/000166.html (last accessed: 02 February 2022).

Stephen, S. (2012). *Three Kingdoms*. London: Methuen Drama.

Van Hulle, D., Nixon, M., & **Neyt, V.** (2016). *The Beckett Digital Library*: *A Digital Genetic Edition* (Series 'The Beckett Digital Manuscript Project'). Brussels: University Press Antwerp. Retrieved from http://www.beckettarchive.org (last accessed: 02 February 2022).

Vaughan, M. (2020). *Theatre Blogging*. London: Bloomsbury. DOI: https://doi.org/10.5040/9781350068858

TO CITE THIS ARTICLE:
Maignant, M., Pellé, D., Brison, G., Poibeau, T. (2022). Drama Critiques' Database. *Journal of Open Humanities Data*, 8: 21, pp. 1–14. DOI: https://doi.org/10.5334/johd.81

Published: 25 August 2022