Quantifying Text Reuse in Seventeenth-Century Rotterdam Chronicles

Max van Winden, Faculty of Humanities, Utrecht University

Introduction

In July 1563, a fire raged in Rotterdam, destroying over 60 ships and 250 houses. This event is described in five 17th-century Rotterdam chronicles, which attribute the devastation to a combination of dry winds and thatched roofs. Two chroniclers mention a barrel-making workshop as the ignition site, and one reports that the most regrettable loss was the lives of those perished in the fire. The similarities between these descriptions are evident, and in 1895, a genealogical study established that these texts were derived from transcripts of an unknown original by Jan Gerritsz. Waerschut (?-1623) (Unger and Bezemer 1895). Despite their strong similarity, the chronicles vary in length and starting date, as indicated by table 1. Moreover, they differ significantly in the contents of the addendums following Waerschut's final entry in 1623.[1]

Considering the observed similarities and differences, this study aims to systemically compare chronicle contents to gain insight into how information from chronicle sources was used. Such analysis can enhance our understanding of how the Rotterdam transcripts were created. Earlier, the topic of text reuse was studied on a small scale, indicating that chroniclers worked as editors, making considered decisions while compiling their texts (Caers 2020). This study seeks to enable the systematic investigation of copying behaviour by developing a method to quantify text reuse. This may contribute to studies on the historic use of chronicles for local history, and the dissemination of knowledge within a growing 17th-century media landscape (Lassche and Morante 2021; Pollmann 2016).

ID*	Author	Start	End	Size (tokens)
A	Anonymous	1426	1648	108,418
B	Anonymous	1426	1658	106,806
C	Jacob Lois	1020	1672	551,748
D	Anonymous	89	1687	322,607
E	Anonymous	1426	1690	51,837

* A = Anonymous. 1648. Kroniek van Rotterdam
B = Anonymous, 1658. Kroniek van Rotterdam
C = Jacob Lois. 1671. Cronycke Ofte Een Corte Waare Oude Beschrivinge Der Stadt Rotterdam, Beschreven Door Jacob Lois, Schepen Derselve Stad Beginnende van Den Jare 1270 Tot Den Jare 1664 En Voorts Vervolgt Anno 1671, Vergaderd Uit Veel Oude Memorien, Contracten, Hantvesten, Prevelegien Ende Geschreve Notitien, Ende by Hem Veel Genooteert Etc. Ende Int Kordt Byeengevoucht
D = Anonymous, 1687. Kronijk, Inhoudende Den Opgang En Voortgang van de Scheepryke Wijdvermaarde Koopstad Rotterdam, Beschreven Door Jan Gerritsz. van Waarschut, Bakker Overleden Int Jaar 1623. Met Een Vervolgh Tot Het Jaar 1663
E = Anonymous, 1690. Beschryvinge Der Stad Rotterdam Mitsgaders Geschiedenissen Zoo Binnen de Stad Als Elders Voorgevallen, van Den Jare 1426 Tot Den Jare 1690

Table 1: Corpus of Rotterdam Chronicles

Method

Digital transcripts of the Rotterdam texts were retrieved from a corpus of early modern Dutch chronicles, collected as part of the research project *Chronicling Novelty. New Knowledge in the*

1. Three of five chroniclers refer to Waerschut's death in 1623, indicating his authorship of the contents before this date.

Netherlands, 1500-1850.[2] The text files are enriched with labels for locations, personal names, and dates (Kuijpers 2022). Existing text reuse detection methods have successfully clustered similar texts by comparing all texts within a given corpus (Vesanto et al. 2017).[3] As for chronicles, stylometric experiments have been successful in distinguishing source and author text within this genre (Smith et al. 2022).

This study utilizes annotated 'date'-labels to make localized text similarity comparisons using the Jaccard index.[4] This text distance measure expresses similarity as the ratio of the number of shared items to the total number of unique items in two sets (Leskovec, Rajaraman, and Ullman 2019). By limiting comparisons to year-level data, the proposed method avoids unnecessary calculations, as descriptions concerning different years are expected to be less similar. As such, the chronicling-genre allows for a more efficient approach to text similarity measurement.

To convert chronicles into text sets, I make use of an XML-parser developed in a recent study by Lassche, Kostkan and Nielbo (2022). This parser splits chronicles into individual events based on the position of the 'date'-label in the text.[5] After parsing, for each chronicle, events concerning the same year are aggregated and split into series of overlapping sets of *n*-size items, called shingles. For example, "anno 1426" is represented as the set 'ann', 'nno', 'no1', 'o14', '142', '426'. Subsequently, year-descriptions are compared with the Jaccard index. This includes collecting data on which year descriptions appear in the chronicles. Applying this method yields a dataset containing similarity scores between year-pairs. The resulting values range from 0, indicating no similarity, to 1 indicating full similarity. Irregularities, such as added or deleted text, affect this output, which can be used for follow-up analyses.

Results

To determine "low" similarity, experiments with a chronicle from Ghent were performed on the corpus.[6] A shingle size of 3 and a similarity threshold of 0.3 were found to be optimal parameters for identifying (dis)similar pairs in the corpus. A smaller shingle size caused Rotterdam events to be incorrectly identified as similar to those of Ghent. Conversely, a larger shingle size failed to reflect similarities within the corpus. The similarity threshold was established by comparing the contents of the descriptions in the corpus.

Table 2 shows the distribution of similarity scores, indicating a high similarity between chronicles A and B. Additionally, appendix 1 shows that only six out of 67 years were not shared between them, indicating that both chroniclers relied on a similar source. Their contents do not overlap after the year 1623, as shown by appendix 2. This divergence applies to all

2. https://chroniclingnovelty.com/
3. See also, for example, the Yale Digital Humanities Lab Intertext project, which provides a comprehensive interface for viewing the probability of intertextuality between two text fragments within a given corpus, https://dhlab.yale.edu/projects/intertext/
4. Please refer to the GitHub repository for the full code and corpus: https://github.com/mvanwinden/chronicling-similarity
5. The XML-parser for annotated chronicles is developed at the *Center for Humanities Computing at Aarhus University*, available at https://github.com/centre-for-humanities-computing/dutch-chronicles
6. Experiments were performed with a corpus of 24 year descriptions, distributed over five Rotterdam chronicles and a chronicle of Ghent from the same time-period, which is written by Justus Billet.

chronicles. This proves the unlikelihood of chroniclers copying each other's work. Regarding chronicle E, there are already deviations after 1593. Possibly, the author of this transcript intervened in the text. Nonetheless, prior to 1623, chronicle E shows similarities the other chronicles, suggesting that chronicle E is based on a less complete transcript.

	Chronicle pairs									
Jaccard Similarity	**A, B**	**A, C**	**A, D**	**A, E**	**B, C**	**B, D**	**B, E**	**C, D**	**C, E**	**D, E**
< 0.1	0	4	1	8	8	3	11	17	15	11
0.1-0.2	1	16	10	2	25	16	3	33	17	6
0.2-0.3	3	17	18	1	14	17	2	37	12	6
0.3-0.4	4	22	19	8	21	22	8	12	7	8
0.4-0.5	21	6	11	11	5	9	16	3	1	3
0.5-0.6	31	0	2	8	0	1	4	0	0	2
0.6-0.7	2	0	0	1	0	0	0	0	0	4
> 0.7	0	0	0	0	0	0	0	0	0	2
Total	62	65	61	39	73	68	44	102	52	42

Table 2: Distribution of Jaccard distance measurements between chronicle pairs

Between 1426 and 1623, there is overlap between all chronicles, both in the years present and the Jaccard index. Upon examining pairs scoring below threshold, it was evident that dissimilarity arose from added information. In other words, every chronicle provides specific details about a given year, while some chronicles (mostly C and D) offer additional information, as is also shown in the introductory note on the event of 1563. This suggests that the information provided by a Waerschut transcript was followed carefully, and chroniclers copied without much consideration.

Conclusion and discussion

This paper presented a computational method to quantify text reuse in five seventeenth-century Rotterdam chronicles. The analysis of Jaccard distance measurement reveals that chroniclers followed the (incomplete) Waerschut transcripts carefully. There is no pattern of specific information being adapted. These findings provide insight into the reliability chroniclers attributed to their sources. Visualizations of (dis)similarity proved useful to get an impression of the textual relationship between the chronicles. These confirm the 1895 view that the chronicles were copied from different transcripts.

Follow-up studies could identify themes in the information added or withheld, and potentially shed light on the author's intervention in the absence or adaptation of year descriptions. Closer analysis of the studied material could also help improve robustness of this method. In this regard, the influence of standardised spelling and the avoidance of loan words through language purism could be considered. In closing, measuring text reuse on a year level could prove useful for studies on source use in chronicles, as other dated sources such as newspaper articles can be included. Comparing by year ensures that this remains a manageable task.

Bibliographical references

Secondary works

Caers, Bram. 2020. *Vertekend Verleden: Geschiedenis Herschrijven in Vroegmodern Mechelen (1500-1650)*. Uitgeverij Verloren.

Kuijpers, Erika. 2022. 'De Informatiebronnen van Albert Louwern (1722-1798), Kroniekschrijver Te Purmerend'. In *Makelaars in Kennis: Informatie Verzamelen, Verwerken & Verspreiden in de Vroegmoderne Nederlanden*, 131–57. Universitaire Pers Leuven.

Lassche, Alie, Jan Kostkan, and Kristoffer Nielbo. 2022. 'Chronicling Crises: Event Detection in Early Modern Chronicles from the Low Countries'. *Proceedings http://Ceur-Ws.Org ISSN* 1613 (2022): (0073): 215–30.

Lassche, Alie, and Roser Morante. 2021. 'The Early Modern Dutch Mediascape. Detecting Media Mentions in Chronicles Using Word Embeddings and CRF'. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 1–10. Punta Cana, Dominican Republic (online): Association for Computational Linguistics.

Leskovec, Jure, Anand Rajaraman, and Jeffrey D Ullman. 2019. *Mining of Massive Datasets*. Cambridge university press.

Pollmann, Judith. 2016. 'Archiving the Present and Chronicling for the Future in Early Modern Europe'. *Past & Present* 230 (suppl 11): 231–52. https://doi.org/10.1093/pastj/gtw029.

Smith, Eleanor, Lianne Wilhelmus, Erika Kuijpers, Alie Lassche, and Roser Morante. 2022. 'Identifying Copied Fragments in an 18th Century Dutch Chronicle'. *Proceedings of the Thirteenth Language Resources and Evaluation Conference.*, 5865–78.

Unger, Johan H. W., and Willem Bezemer. 1895. *Bronnen Voor de Geschiedenis van Rotterdam: II De Oudste Kronieken Beschrijvingen van Rotterdam En Schieland*. Rotterdam: P. van Waesberge & Zoon.

Vesanto, Aleksi, Asko Nivala, Heli Rantala, Tapio Salakoski, Hannu Salmi, and Filip Ginter. 2017. 'Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910'. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*.

Primary works

Anonymous. 1648. "Kroniek van Rotterdam." Rotterdam. Manuscript collection 33.01 inv.nr. 1552. Stadsarchief Rotterdam.

Anonymous. 1658. "Kroniek van Rotterdam." Rotterdam. Manuscript collection 33.01 inv.nr. 1553. Stadsarchief Rotterdam.

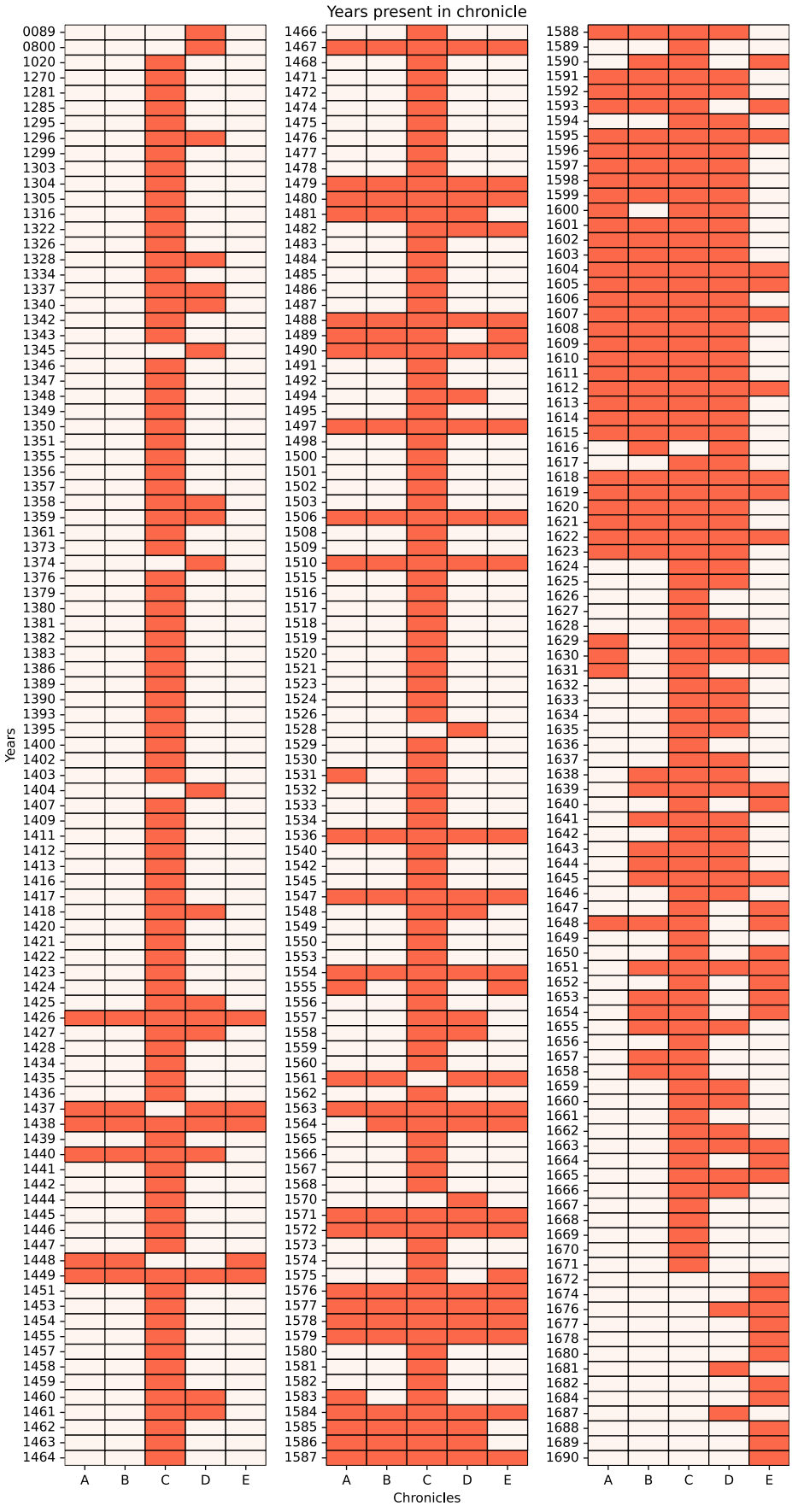
Anoymous. 1690. "Beschryvinge Der Stad Rotterdam Mitsgaders Geschiedenissen Zoo Binnen de Stad Als Elders Voorgevallen, van Den Jare 1426 Tot Den Jare 1690". Afschrift Uit Het Einde Der 17e Eeuw." Rotterdam. Manuscript collection 33.01 inv.nr. 1577. Stadsarchief Rotterdam.

Anonymous. 1663. "Kronijk, Inhoudende Den Opgang En Voortgang van de Scheepryke Wijdvermaarde Koopstad Rotterdam, Beschreven Door Jan Gerritsz. van Waarschut,

Bakker Overleden Int Jaar 1623. Met Een Vervolgh Tot Het Jaar 1663". Rotterdam. Manuscript collection 33.01 inv.nr. 1555. Stadsarchief Rotterdam.

Lois, Jacob. 1671. "Beschrijving: "Cronycke Ofte Een Corte Waare Oude Beschrivinge Der Stadt Rotterdam, Beschreven Door Jacob Lois, Schepen Derselve Stad Beginnende van Den Jare 1270 Tot Den Jare 1664 En Voorts Vervolgt Anno 1671, Vergaderd Uit Veel Oude Memorien, Contracten, Hantvesten, Prevelegien Ende Geschreve Notitien, Ende by Hem Veel Genooteert Etc. Ende Int Kordt Byeengevoucht"." Rotterdam. Manuscript collection 33.01 inv.nr. 1562. Stadsarchief Rotterdam.

Appendix 1 Distribution of year descriptions in the corpus.



Appendix 2 Jaccard index heat map for year descriptions shared between chronicles.