Chronicling Crises: Event Detection in Early Modern Chronicles from the Low Countries

Alie Lassche[1], Jan Kostkan[2] and Kristoffer Nielbo[2]

[1]Leiden University, Institute of History, Doelensteeg 16, 2311 VL Leiden, The Netherlands

[2]Center for Humanities Computing Aarhus, Jens Chr. Skous Vej 4, Building 1483, DK-8000 Aarhus C, Denmark

Abstract

Between the Middle Ages and the nineteenth century, many middle-class Europeans kept a handwritten chronicle, in which they reported on events they considered relevant. Discussed topics varied from records of price fluctuations to local politics, and from weather reports to remarkable gossip. What we do not know yet, is to what extent times of conflict and crises influenced the way in which people dealt with information. We have applied methods from information theory – dynamics in word usage and measures of relative entropy such as *novelty* and *resonance* – to a corpus of early modern chronicles from the Low Countries (1500–1820) to provide more insight in the way early modern people were coping with information during impactful events. We detect three peaks in the novelty signal, which coincide with times of political uncertainty in the Northern and Southern Netherlands. Topic distributions provided by Top2Vec show that during these times, chroniclers tend to write more and more extensively about an increased variation of topics.

Keywords

event detection, top2vec, doc2vec, information theory, novelty, resonance, relative entropy, early modern chronicles

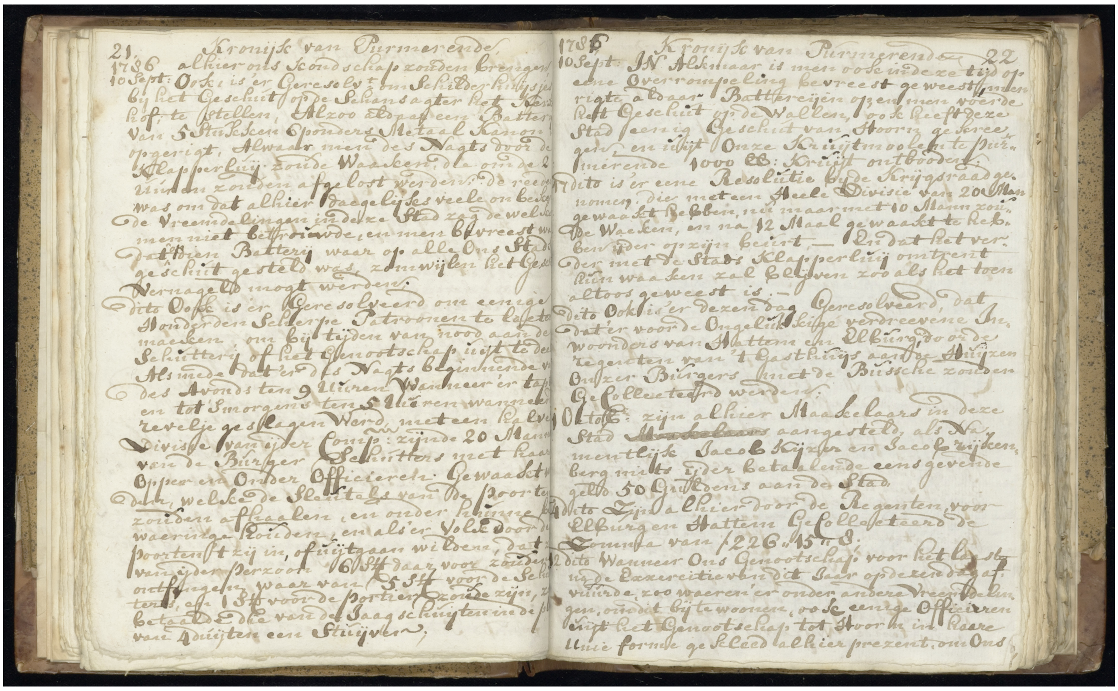
1. Introduction

Between the Middle Ages and the nineteenth century, many middle-class men (and a handful of women) in Europe kept a handwritten chronicle, in which they reported on current events in their communities, and on what they considered interesting or relevant. These texts were ordered chronologically, providing both the date and a report of a certain event (see Figure 1). Chronicles were rarely printed in the lifetime of their authors, but, despite their scribal form, they could still circulate in the localities, be read and continued by other authors, influencing future generations [13].

The described topics in chronicles were various. Chroniclers alternated between descriptions of political developments and records of price fluctuations of grain, butter, and milk, weather reports, mentions of their relatives' birthdays, gossip, religious developments, and reports on unusual, strange, or marvelous events both nearby and further away. Although chronicles did change over time, the type of information chroniclers selected to be included remained fairly stable, which enables us to study the genre across centuries. One subcategory

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium

a.w.lassche@hum.leidenuniv.nl (A. Lassche); jan.kostkan@cas.au.dk (J. Kostkan); kln@cas.au.dk (K. Nielbo)
0000-0002-7607-0174 (A. Lassche); 0000-0002-9707-7121 (J. Kostkan); 0000-0002-5116-5070 (K. Nielbo)



21 Kronijk van Purmerende,
1786 alhier ons kondschap zouden brengen,
10 Sept: Ook is er Geresolv: om Schilderhuijsjes
by het Geschut op de Schans agter het Kerk
hof te stellen, Alzoo aldaar een Battery
van 5 Stukken 6ponders Metaal Kanon
opgerigt, Alwaar men des Nagts door de
Klapperluij zoude Waaken, die om de 2
uuren zouden afgelost werden; de reeden
was om dat alhier daagelijks veele onbeken
de Vreemdelingen in deze Stad zag de wel
men niet betrouwde, en men bevreest wa
dat die Battery waar op alle Ons Stads
geschut gesteld was, zomwijlen het Gesch
vernageld mogt werden.
dito Ook is er Geresolveerd om eenige
Honderden Scherpe Patroonen te laeten
maeken om bij tijden van nood aan de
Schutterij of het Genootschap uijt te deel
Als mede dat er des Nagts beginnende va
des Avonds ten 9 Uuren Wanneer er tap
en tot 's morgens ten 5 Uuren wanneer
revelje geslagen Werd, met een halve
Divisie van ijder Comp: zijnde 20 Mann
van de Burger Schutters met haar
Opper en Onder Officieren Gewaakt w
den, welke de Sleutels van de poorte
zouden afhaalen, en onder hunne be
waeringe houden, en als er Volk door de
poorten It zij in, of uijtgaan wilden, dat z
van ijder perzoon 6 St daar voor zouden
ontfangen, waar van 5 St voor de Schu
ters, en 1 St voor de portiers zoude zijn, z
betaalde die van de Jaagschuijten in de pl
van 4 duijten een Stuijver,

1786 Kronijk van Purmerende 22
10 Sept: IN Alkmaar is men ook in deze tijd op
eene Overrompeling bevreest geweest, men
rigte aldaar Batterijen op, en men voerde
het Geschut op de Wallen, ook heeft deze
Stad eenig Geschut van Hoorn gekree
gen, en uijt Onze Kruijtmoolen te Pur
merende 1000 ℔: Kruijt ontbooden.
dito is er eene Resolutie by de Krijgsraad ge
nomen, die met een Heele Divisie van 20 Man
gewaakt hebben, nu maar met 10 Mann zou
de Waeken, en na 12 Maal gewaakt te heb
ben ijder op zijn beurt, — En dat het ver
der met de Stads Klapperluij omtrent
hun waaken zal blijven zoo als het toen
altoos geweest is. —
dito Ook is er dezen dag Geresolveerd, dat
dat er voor de Ongelukkige verdreevene In
woonders van Hattem en Elburg, door de
regenten van 't Gasthuijs aan de Huijzen
Onzer Burgers met de Bussche zouden
Gecollecteerd werden.
1 Oktob: zijn alhier Maakelaars in deze
Stad ~~Maakelaars~~ aangesteld, als Na
mentlijk Jacob Kijzer en Jacob Rijken
berg mits ijder betaalende eens gevende
geld 50 Guldens aan de Stad.
4 dito Zijn alhier door de Regenten, voor
Elburg en Hattem Gecollecteerd de
Somma van f 226 „ 15 „ 8.
12 dito Wanneer Ons Genootschap voor het laatst
na de Exercitie van dit Jaar op dezen dag af
vuurde, zoo waeren er onder andere Vreemdelin
gen, om dit bij te woonen, ook eenige Officieren
uijt het Genootschap tot Hoorn in haare
Uniforme gekleed alhier prezent, om Ons

Figure 1: Fragment of an eighteenth-century chronicle about Purmerend. (Albert Pietersz. Louwen, *Kronijk der stad Purmerende*, 1791. Noord Hollands Archief Haarlem.)

of chronicle texts emerged out of political crises, wars, and civil conflicts. During such times, many authors started to record public events that were upsetting their lives and to teach their readers the lessons to be learned from them [13, 14].

What we do not know, is how times of conflict and crises influenced the chronicler's way of reporting and dealing with information. Did chroniclers collectively write about the same topic during such a crisis, or were they more recipients of broader information? This paper will demonstrate how chroniclers were coping with information during impactful events. We aim to provide more insight into the way early modern people dealt with crises. We apply methods from information theory to a corpus of early modern chronicles from the Low Countries to find an answer to the question formulated above.

Related work will be discussed in Section 2. The used corpus will be introduced in Section 3, while Section 4 contains a description of the methods used in this study. In Section 5, the obtained results will be discussed, and in Section 6, we draw some conclusions from the results and do some suggestions about the direction future research should take.

2. Related work

2.1. The theory of the event

Chronicles can be considered as a collection of chronological events. However, few historians have been studying 'the event' as a theoretical category [16, p. 198] – even though the anti-evenementalism of social historian Fernand Braudel, who considered the history of events as the mere froth on the waves of history, was largely replaced by a return to writing about events in the 1970s. Anthropologist Marshall Sahlins has stated that '[e]vents can be distinguished from uneventful happenings only to the extent that they violate the expectations generated by cultural structures. The recognition of the event as the event, therefore, presupposes structure' [16, p. 199][15]. The same distinction is made by [8]: 'In an abstract sense, every occurrence can be described as an event', they state. However, in most contexts, the term *event* is not used in that way. There is time that is filled with events, but since these are embedded in routines, such times are not experienced as events in the narrower sense. Instead, 'only those incidences that strike us as noticeable ruptures with expected processes and routines are recognized as real events' [8, p. 78].

From a linguistic perspective, the question of whether an incident is important enough to be an event is not relevant: every state, change, or happening is considered an event. In linguistics, the term *event* is related to the concept of *eventuality*, which was introduced by the linguist Emmon Bach in 1986 and comprised states, processes, and events [2]. Many linguists indeed understand the term eventuality in the broadest sense, comprising events, processes, states, happenings, changes, episodes, etc., as is for example the case in a study on event detection and classification for historical texts [17].

The difference between the historical and linguistic perspectives on events is that linguists focus on the linguistic elements that are used, while historians focus on the result of an event. When linguists use subcategories to distinguish between events, it is based on these linguistic elements. Neither the historical perspective nor the linguistic perspective on events is completely suitable for the study of events in early modern chronicles. In these texts, both abstract and emphatic events are included. First and foremost, chroniclers report on happenings in their surroundings that they consider being noticeable ruptures of their daily routine. This can be the threat of war, the visit of a foreign king, a national conflict, or unusual weather phenomena. This could be considered an event in the emphatic sense. At the same time, however, chroniclers also include reports on daily or weekly routines such as checking the wind direction and temperature, summarizing the Sunday sermon, reporting on the prices on the market, or referring to weekly board meetings they attend. These are the abstract kind of events. In the context of this study, we consider an event to be every description that is linked to a specific date. For more details on methods see Section 4 and Appendix A.

2.2. Information theory

The methods used in this study are based on dynamics in word usage and measures of relative entropy. We know from previous work that word usage in newspapers is sensitive to the dynamics of socio-cultural events [5, 4]. Methods from complexity science, such as fractal analysis, have been used to identify distinct domains of newspaper content based on temporal

Table 1
Corpus statistics (last updated on 03-05-22).

corpus	# chronicles	# authors	# tokens
annotated	191	143	11,028,367
corrected	96	80	5,123,256

patterns in word use [18], and to distinguish cultural and social catastrophic events that display class-specific fractal signatures in, among other things, word usage in newspapers [5].

Previous studies have shown that entropy measures can be used to detect fundamental conceptual differences between distinct periods [7], opposite political movements [3], and the development of ideational factors in writing with a serial structure [11]. More specifically, several studies have applied windowed relative entropy to thematic text representations to generate signals that capture information *novelty*, which is the content difference from the past, and information *resonance*, which is the degree to which future information conforms to novelty. The methods have been successfully applied to parliamentary debates from the French Revolution [3], to Dutch newspapers from the second half of the 20st century [18], and to Danish newspapers from the COVID-19 pandemic [12].

3. Corpus

The data set applied in this study is a subset of a corpus that was collected and digitized in the context of the research project ‘Chronicling Novelty. New knowledge in the Netherlands, 1500-1850’.[1] The full corpus consists of about 320 early modern chronicles that are written in the Dutch language between 1500 and 1850. They are chronologically organized, cover events that happened in the lifetime of the author, and focus on local events more than national, individual, or familial. About 130 of these chronicles had been published before as a contribution to a journal, on the initiative of an archive, or in the private domain, and were digitized by the Digital Library for Dutch Literature (DBNL). The other chronicles are kept in libraries and archives throughout the Netherlands and Belgium.

Every manuscript page was scanned and transcribed with both the Handwritten Text Recognition tool Transkribus [9], and the help of volunteers on the online crowdsourcing platform *Vele Handen*. Afterwards, both content and layout was annotated by volunteers, using labels including `page number`, `date`, `location`, and `person name`.

Our data set consists of 191 chronicles that were fully transcribed and annotated. However, the `date` tag, which plays a pivotal role in this study, repeatedly was subject to many bugs and crashes. The tag therefore needed a manual inspection. We use all 191 annotated chronicles for training models (`corpus annotated`), but only the 96 chronicles in which the `date` label was manually checked, were used for analysis (`corpus corrected`). See Table 1 for more statistics on the used corpus.

[1] On http://www.chroniclingnovelty.com/kronieken/, an overview of the corpus can be found.

Table 2
Primitives statistics.

primitives	# documents	used in # step in pipeline
`annotated`	116,023	2
`corrected`	63,883	–
`corrected daily`	36,147	3, 4
`prototypes`	22,516	5

4. Pipeline

We developed a five-step research pipeline, additional details can be found in appendix A:[2]

1. **Chunk chronicles into primitives**. Since chronicles are chronologically structured and mentions of a date are labeled as such, we make a cut before every `date` label, to chunk the texts into smaller fragments that can be connected to a date label. This was done for both `corpus annotated` and `corpus corrected`. We call the resulting text chunks *primitives*. We furthermore made a subset of `primitives corrected daily`, only containing primitives with a fully specified date tag. Table 2 contains statistics of the distinctive data sets. The `primitives annotated` are used in step 2, while the `primitives corrected daily` are used in the other steps.
2. **Primitive representations**. We use a `Top2Vec` model to create both document representations of the `primitives annotated`, and topics [1]. The model provides two relevant outputs, which are the `doc2vec` embeddings of the documents [10], and the cosine similarities of the documents toward the estimated topic centroids. We reduced the trained model to 100 topics.
3. **Down-sampling (choosing prototypical primitives)**. In order to compute novelty and resonance signals, we want only one textual representation per day. In doing so, we avoid calculating novelty over primitives that are in fact descriptions of the same day. Instead, we enable keeping the temporal dimension of novelty to at least one month. We therefore cluster the `primitives corrected daily` per day and pick a prototypical primitive, based on cosine similarity: if a day has multiple primitives, the primitive with the shortest distance to the others is picked as a prototype, assuming that this primitive is the most representative one. This method also allows us to calculate the *uncertainty* of a prototypical primitive, which we express as the standard deviation of the mean distance of a prototypical event to the other primitives on that day.
4. **Diachronic topic analysis**. We group `primitives corrected daily` per year, and take the mean cosine similarities to every topic, to analyze topic fluctuation over time.
5. **Novelty detection**. We calculate *novelty* and *resonance* of our time series of `prototypes`. We expect peaks in the novelty signal to be indications of an event.

[2] Please refer to the git repository for the full code: https://github.com/centre-for-humanities-computing/dutch-chronicles.

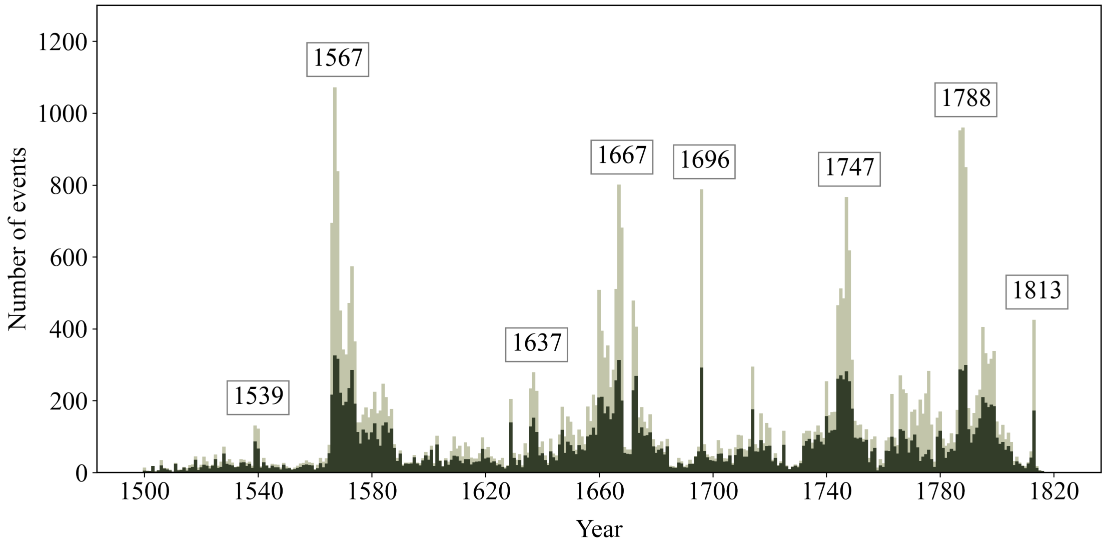


Figure 2: Frequency of `primitives corrected daily` and `prototypes` (1500-1820).

5. Results

5.1. General primitive and topic statistics

Figure 2 shows the frequency of primitives per year in light green. The time span 1500–1820 is chosen to exclude references to events that happened far before a chronicler's contemporary times, and avoid artifact patterns over time. We have labeled eight peaks in the frequency plot, of which the ones in 1567, 1667, 1747, and 1788 are the most outstanding. Keeping in mind that chronicle production peaked during times of uncertainty, these peaks are not surprising, because they mark key points in the history of the Netherlands: 1567 is just before the beginning of the Dutch Revolt against the Spanish king. In 1667, the War of Devolution is fought between France and Spain, concerning the rule of the Southern Netherlands (known then as the Spanish or Habsburg Netherlands). In 1747, the French army has invaded the southern parts of the Dutch Republic. Several cities are occupied, or threatened by occupation. It results in the appointment of Willem V as Stadtholder of all provinces of the Northern Netherlands, and thus the end of the Second Stadtholderless Period. The peak around 1788 marks the restoration of the Orange Stadtholder against the rise of the democratic Patriot movement during the period 1780–1787. These fluctuations in event frequency confirm that more chroniclers tend to write (or, chroniclers tend to write more often) during times of uncertainty.[3] Based on this graph alone, we cannot say anything about the content they are writing. To remedy this, we use a `Top2Vec` model to represent the chronicle content.

[3] Given this, it is striking that no peak appears in 1672, the Dutch *Rampjaar* (Disaster Year) in which the people were described as *redeloos* (irrational), the government as *radeloos* (distraught), and the country as *reddeloos* (beyond salvation). However, this absence is mainly due to a lack of data in the full corpus from that period.

Table 3
Topics in the categories *natural* (0), *social* (1), *linguistic* (2), *economic* (39), *political* (42), and *cultural* (48).

topic	words
0	wolkachtig motbuitjes verhelderd betrekt wolkens buiachtig verdund verdonkert betrokken
1	overleede hat do overleeden edl weeduwe mevrou weed juffrou haar jaaren niwe oostindise dri
2	par la dont dans dernier nouveau un alors le sur avec du francais on il etre une consequence les
39	weinig duur daardoor thans hooi mogelijk steeds aanzien vooral prijzen aardappels oorzaak
42	spaengiarden ducdalbe vuel ducdalf deestyt brusel vuele prinsche dagelycx scuyten gescut
48	plechtigheyd geluyd beyaerd triumph luyster feest autoriteyten bywezen magten musicq

Several categories can be discerned from qualitative inspection of the topics in the reduced Top2Vec model. There are natural, cultural, social, economic, and political topics – and some topics fall into multiple categories. A sixth category contains topics of words that are not clustered on semantic similarity, but on linguistic characteristics. In Table 3, we included an example of every category. Topic 0 (the most dominant topic within the corpus) clearly belongs to the category of natural topics, containing words related to the weather. Topic 1 falls into the category of social topics, with words such as 'overleeden' (*passed away*), 'weeduwe' (*widow*) and 'juffrou' (*miss*). Topic 2 is not really a topic, because its words are clustered on linguistic characteristics: they are all French. In topic 39, words related to prices and products, such as 'duur' (*expensive*), 'hooi' (*hay*), 'aardappels' (*potatoes*), and 'prijzen' *(prices*), are clustered, belonging to the economic category. The words 'spaengiarden' (*Spaniards*), 'ducdalbe' (*Duke of Alba*), 'prinsche' (*prince*), and 'gescut' (*artillery*) indicate that topic 42 can be considered a political one. Topic 48 contains words related to festivities ('plechtigheyd' (*ceremony*), 'beyaerd' (*carillon*), 'feest' (*party*), 'musicq' (*music*) and thus belongs to the cultural category.

Topical dynamics were represented by plotting the mean cosine similarity of all primitives in one year towards a certain topic. Some of the topics show a clear trend over time which can easily be linked to a social, cultural, or political situation at that time. Other topics demonstrate a repetitive pattern, remaining stable on the long term.

5.2. Chronicling novelty

The prototype primitives, of which the frequency is visualized in dark green in Figure 2, were used to compute the novelty signal. Three peaks can be observed in the novelty signal in Figure 3, which remain solid when we adjust the window size w. Peaks are visible around 1568, 1662, and 1789. These fluctuations in the signal indicate that a document vector is surprising compared to its preceding w vectors. The valley the novelty signal shows after a peak means that the following documents are less surprising because their vectors are more similar to the w previous ones. The slow ascent of the peaks in 1568 and 1789 point to a long period of steadily increasing surprise. The steep descent that follows, indicates a sudden decrease in content novelty of the documents.

The three peaks mark the earlier mentioned key points in the history of the Low Countries: the start of the Dutch Revolt (1568), the prelude to the War of Devolution (1662), and the end of the patriot movement with the return of the Stadtholder (1789). Furthermore, they approximately coincide with a peaking primitive frequency. This indicates that when more

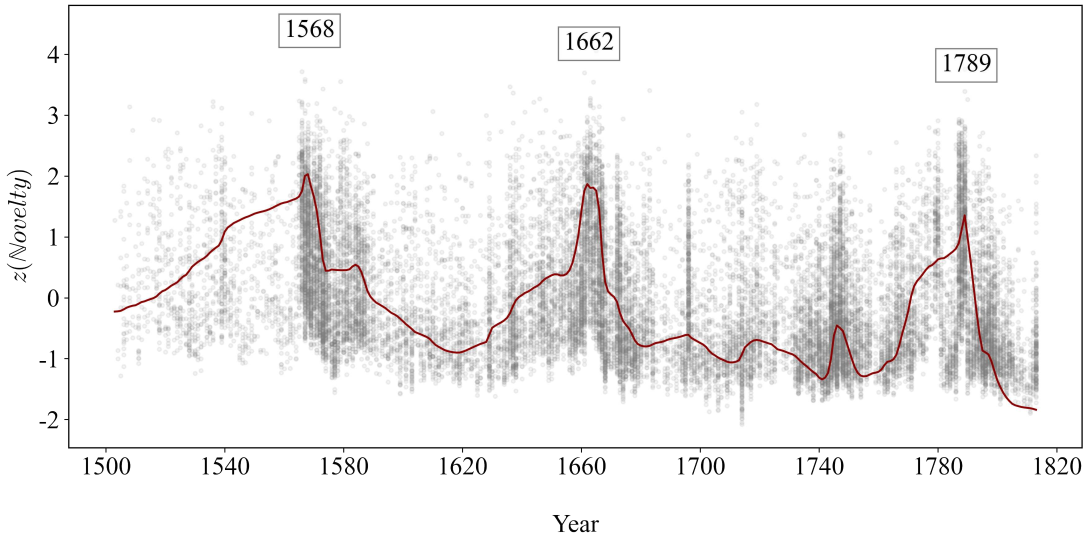


Figure 3: Novelty signal of `prototypes` (1500-1820). Standard scores of individual documents (grey). Trend line (red) estimated using a nonlinear adaptive filter (described in the Appendix).

Table 4
Dominant topics during novelty peaks (1568, 1662, and 1789).

year	topic	words
1568	10	steeds hooi daarvan thans aardappels weinig
	7	genoegsaem eenider naermiddag uijtvoer verscheijde
	9	commune besluit adjointe maire mairie zone prefect
1662	7	genoegsaem eenider naermiddag uijtvoer verscheijde
	4	vuyt spaengaerden comen recomen voirss alsdoen
	5	eenijghe guarnijsoen aprijl opden seere brussele
1789	56	paus innocentius roomen romijn gregorius clemens stoel
	49	solemnele kercke reliquien kerck misse gesongen
	0	wolkachtig motbuitjes verhelderd betrekt wolkens

chroniclers report (or chroniclers report more) in their chronicle, the content of their reports changes. A peak in the novelty signal can mean several things, for example, *(1)* a few topics become more dominant than others, *(2)* earlier dominant topics become less dominant, or *(3)* more diverse topics appear. The yearly mean cosine similarity per topic that was obtained with `Top2Vec` is used to get insight in the topics that are dominating at the time of the novelty peaks. Although one might expect that, during crises, chronicles would become more focused on a small number of topics directly related to the crisis, it turns out that the topic distribution is flatter during novelty peaks than in other years: even the topics with the highest mean cosine similarities are still close to or below 0. It indicates a large variety in described topics during such years. The top words of the three most dominant topics during novelty peaks are included in Table 4.

It must be said that document length is an important driver for the variation between high and low novelty. Longer documents are more novel than shorter documents, because long documents contain more varied information and therefore have average similarity to many topics, while short documents, containing less information, have high similarity with one topic, but uniformly low similarity with the rest.

A positive association between novelty and resonance would indicate an innovation bias, meaning that novelty introduced in the past leaves traces in the future. In our results, the resonance signal remains flat over time, which suggests that future information does not conform to the introduced novelty. It means that a newly introduced event in the corpus of chronicles does not have an effect on the content that is described afterward. In other words: big events do not impact the writing style and habits of chroniclers in the long term.

Different interpretations are possible regarding the obtained results. Increasing diversity in described topics during crises suggests that during times of high uncertainty, people are more open to new information. Alternatively, it can be an indication of an expanding mediascape, or a more thorough consumption of new media. Furthermore, it may be indicative of how early modern people understood crises in a different way than we expect. A crisis is not only about soldiers, sieges, and deaths (topics 4 and 5 in 1662), but it is also relevant what kind of food is available (topic 10 in 1568), how the weather might influence this (topic 0 in 1789), what the city government decides (topic 9 in 1568), how the situation in a foreign city evolves, and to not forget religious duties (topic 49 in 1789). This variety in topics could also show that not every chronicler was equally affected by these historical events. Still, they felt the need to start or intensify their recordings on events happening in their lives. An overload of information asks for a sufficient approach. After being exposed to an information explosion, people tend to select what is relevant to them, discarding the other topics. This is what the resonance signal shows: there is a short period of information overload, but soon, the variety of described topics decreases again.

6. Conclusion

We have presented a method to detect events in early modern Dutch chronicles, which has provided insight into the way chroniclers cope with information during impactful events, and how these happenings influence their way of writing. Our main conclusion is that early modern chronicles tend to write more and more extensively during times of political uncertainty. However, the topics they describe during such times also get more varied. This is shown by a peak in the novelty signal, and a flatter topic distribution. Furthermore, such an increase in event density and a change in the novelty signal does not influence future reported events. Soon, things get back to normal, and the (writing) life of the chronicler continues as it did before.

The representativeness of the corpus used for analysis is not unproblematic and is something that could be improved in future research. We have used about one-third of the full corpus of chronicles, mainly due to the fact that only this part was digitized and annotated at the time. Besides, the focus on only 'daily events' introduced a bias in the results. The date with the highest frequency of primitives (67), most of them were from one author, describing how the Stadtholder Willem V visited his hometown Purmerend. Other frequently reported dates pointed us as well to events happening on one day, rather than to events spread over a longer period of time. An exploration of the frequently mentioned dates would gain value when 'monthly dates' were also included.

In this study, we have used the corpus as a whole in computing novelty and resonance signals, showing that the genre remains stable over time, despite of several impactful events. Future work will focus on the novelty and resonance signals of individual authors, in order to see whether the elevated topic diversity during crises can also be observed here. This should furthermore provide more insight in their personal writing style, and whether certain events they experience have a lasting influence on their way of chronicling. Other future research should investigate whether the absence of a positive relationship between novelty and resonance is distinctive for the genre of chronicles, or more general something that can be observed in early modern texts. Early modern pamphlets or newspapers – specifically written for an audience – would be an interesting showcase to explore their difference from chronicles – since the latter are in the end more often a case of a private matter.

Acknowledgments

This research was executed at the Center for Humanities Computing Aarhus. The research stay of Alie Lassche was funded by the Dr. Catharine van Tussenbroek Fonds, and the Stichting de Fundatie van de Vrijvrouwe van Renswoude te 's-Gravenhage. Kristoffer Nielbo was supported by the Nordic e-Infrastructure Collaboration (NeIC) and the Danish e-Infrastructure Cooperation (DeIC) with grant DeiC-AU1-L-000001.

References

[1] D. Angelov. "Top2Vec: Distributed Representations of Topics". In: arXiv:2008.09470 [cs, stat] (2020). arXiv: 2008.09470 [cs, stat].

[2] E. Bach. "The Algebra of Events". In: *Linguistics and Philosophy* 9.1 (1986), pp. 5–16.

[3] A. T. J. Barron, J. Huang, R. L. Spang, and S. DeDeo. "Individuals, Institutions, and Innovation in the Debates of the French Revolution". In: *Proceedings of the National Academy of Sciences* 115.18 (2018), pp. 4607–4612. DOI: 10.1073/pnas.1717729115.

[4] J. Daems, T. D'haeninck, S. Hengchen, T. Zere, and C. Verbruggen. "'Workers of the World'? A Digital Approach to Classify the International Scope of Belgian Socialist Newspapers, 1885–1940". In: *Journal of European Periodical Studies* 4.1 (2019), pp. 99–114. DOI: 10.21825/jeps.v4i1.10187.

[5] J. Gao, J. Hu, X. Mao, and M. Perc. "Culturomics Meets Random Fractal Theory: Insights into Long-Range Correlations of Social and Natural Phenomena over the Past Two Centuries". In: *Journal of The Royal Society Interface* 9.73 (2012), pp. 1956–1964. DOI: 10.1098/rsif.2011.0846.

[6] K. L. Gray. *Comparison of trend detection methods*. University of Montana, 2007.

[7] J. Guldi. “The Measures of Modernity: The New Quantitative Metrics of Historical Change Over Time and Their Critical Interpretation”. In: *International Journal for History, Culture and Modernity* 7.1 (2019), pp. 899–939. DOI: 10.18352/hcm.589.

[8] T. Jung and A. Karla. “1. Times of the Event: An Introduction”. In: *History and Theory* 60.1 (2021), pp. 75–85. DOI: 10.1111/hith.12193.

[9] P. Kahle, S. Colutto, H. Hackl, and H. Mühlberger. “Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents”. In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 04. 9, pp. 19–24. DOI: 10.1109/icdar.2017.307.

[10] Q. Le and T. Mikolov. “Distributed representations of sentences and documents”. In: *International conference on machine learning*. Pmlr. 2014, pp. 1188–1196.

[11] K. L. Nielbo, K. F. Baunvig, B. Liu, and J. Gao. “A Curious Case of Entropic Decay: Persistent Complexity in Textual Cultural Heritage”. In: *Digital Scholarship in the Humanities* 34.3 (2019), pp. 542–557. DOI: 10.1093/llc/fqy054.

[12] K. L. Nielbo, F. Haestrup, K. C. Enevoldsen, P. B. Vahlstrup, R. B. Baglini, and A. Roepstorff. *When No News Is Bad News – Detection of Negative Events from News Media Content.* 2021. arXiv: 2102.06505 [cs].

[13] J. Pollmann. “Archiving the Present and Chronicling for the Future in Early Modern Europe”. In: *Past & Present* 230.suppl 11 (2016), pp. 231–252. DOI: 10.1093/pastj/gtw029.

[14] J. Pollmann. *Catholic Identity and the Revolt of the Netherlands, 1520-1635.* The Past & Present Book Series. Oxford; New York: Oxford University Press, 2011.

[15] M. Sahlins. “The Return of the Event, Again: With Reflections on the Beginnings of the Great Fijian War of 1843 to 1855 between the Kingdoms of Bau and Rewa”. In: *Clio in Oceania: Toward a Historical Anthropology.* Ed. by A. Biersack. Washington: Smithsonian Institution Press, 1991, pp. 37–100.

[16] W. H. Sewell Jr. *Logics of History: Social Theory and Social Transformation.* Chicago Studies in Practices of Meaning. Chicago: University of Chicago Press, 2005.

[17] R. Sprugnoli and S. Tonelli. “Novel Event Detection and Classification for Historical Texts”. In: *Computational Linguistics* 45.2 (2019), pp. 229–265. DOI: 10.1162/coli_a_00347.

[18] M. Wevers, J. Kostkan, and K. Nielbo. “Event Flow - How Events Shaped the Flow of the News, 1950-1995”. In: *Computational Humanities Research Conference.* Amsterdam, 2021, pp. 62–76.

A. Methods

Document segmentation

The corpus does not come reliably segmented into sentences or documents. We use provided date tags for our document segmentation. In general, we consider a document to be the text

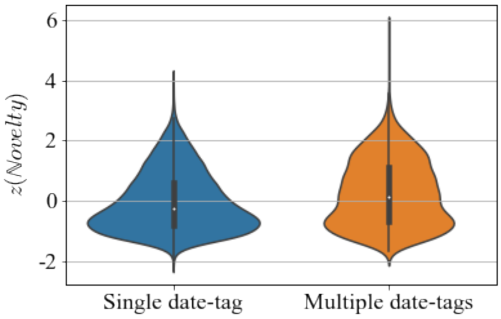


Figure 4: Comparison of novelty score distributions between documents with a single and multiple date-tags.

beginning with a date tag `<date A>`, which in turn serves to date the document. A document can span multiple lines and pages, and ends with a date tag `<date B>`, which indicates the beginning of the next document. However, we do not segment within lines – a document always contains an entire line of text. In doing so, we attempt to address the cases, when an event's dating does not appear at the beginning of the entry, but at some later point (e.g. in the middle of a line). As a result of this rule, lines with multiple date tags are considered one event. In such case, all date tags are recorded for later sanity checks, but the date tag appearing first is chosen as the dating of such event. Allowing an event to have multiple date-tags is useful when parsing chronicle entries dated with a range of dates. For example, an entry such as `` `between <date A> and <date B>, there was a heavy rainfall which influenced the harvest in a bad way' `` is not split into two, as long as both date tags appear on the same line.

The chronicles that were digitised by the DBNL pose a challenge here, since their lines do not match the actual lines in the original manuscript. Instead, their lines are artificially created when converting them to page-XML in order to upload them in Transkribus. Since chroniclers were very inconsistent in their use of punctuation, these lines sometimes have the length of one paragraph. This is not a problem as long as such a chronicle entry corresponds to a single dated event. However, in some cases one document can consist of multiple dated entries referring to different events. Luckily, the vast majority of the documents we used to fit the novelty signal have a single date-tag (20,372 out of 22,516 documents) and thus correspond to a single dated entry. Furthermore, whether a document has a single date-tag or multiple does not seem to predict novelty scores very well (see Figure 4). In summary, the sanity checks we conducted did not uncover a consequential bias to our analysis, resulting from our document segmentation strategy.

Preprocessing

Annotators have marked words that span multiple lines with a special character. First, we concatenate these into a single word and remove all the other special characters used by annotators. Next, unique IDs are assigned to documents, which can be used interchangeably in both the annotated and corrected corpus.

From here on, the processing steps differ for both corpora to reflect the task they are used for. The annotated corpus is used for training of the Top2Vec model and should as such offer both sufficient diversity and number of texts. For this reason, we only exclude documents shorter than 50 characters (mostly OCR artifacts, or non-documents, such as page numbers). In total, 116,023 documents are passed on to Top2Vec training with the average length of 577 characters (SD=1260).

On the other hand, the corrected corpus is used to fit the novelty signal and should in turn be reliably transcribed and dated. First we attempt to capture events that can be dated up to a day (i.e. having a fully specified date tag in a YYYY-MM-DD format, as opposed to e.g. YYYY-MM). This undoubtedly has an effect on the results, since events that are not connected to a specific date by the chronicler, but instead to e.g. a month, are excluded. This is further discussed in the final section of the paper. Second, we exclude documents shorter than 50 and longer than 5000 characters. With this additional upper limit on document length we attempt to exclude events, which contain verbatim copies of official documents and non-events, such as chronicle appendices. These very long documents are outliers in both corpora (less than 1% of the documents in the corrected corpus are longer than 5000 characters) and a majority of them contain a single date-tag, meaning they are not a concatenation of multiple short events. Finally, only documents dated in the period of interest (years 1500 through 1820, inclusive) are kept. In total, 36,147 documents are passed on to the next step of novelty detection with the average length of 525 characters (SD=635).

Event representations

We use a Top2Vec model to create both document representations and topics [1]. This model is based on the assumption that many semantically similar documents are indicative of an underlying topic. Consecutively, Top2Vec creates jointly embedded document vectors and word vectors using doc2vec, it creates lower dimensional embeddings of document vectors using UMAP, and it finds dense areas of documents using HDBSCAN. For each dense area, the centroid of the document vectors is then calculated, which is assumed to be the topic vector. Finally, it searches for the *n*-closest word vectors to the resulting topic vector, in order to create a topic.

An important difference between Top2Vec and traditional bag-of-word topic modeling methods such as LDA, is that the semantic embedding used in the first method has the advantage of learning the semantic association between words and documents. We consider Top2Vec therefore a more suitable method, since the corpus at hand contains large spelling variation, for which the semantic embedding approach can serve as a solution. Furthermore, LDA models topics as distribution of words, which are then used to recreate the original document word distributions with minimal error. This often necessitates uninformative words which are not topical to have high probabilities in the topics since they make up a large proportion of all text. A stopword list can be used to solve this problem, but expanding the list in order to get rid of these non-topical words can be a never ending iterating process. Top2Vec does not need a stop list, because high frequency words that occur in all documents will not be particularly close to any topic vector and thus not dominating in any topic.

We train a model on all primitives annotated that are longer than 50 characters. The

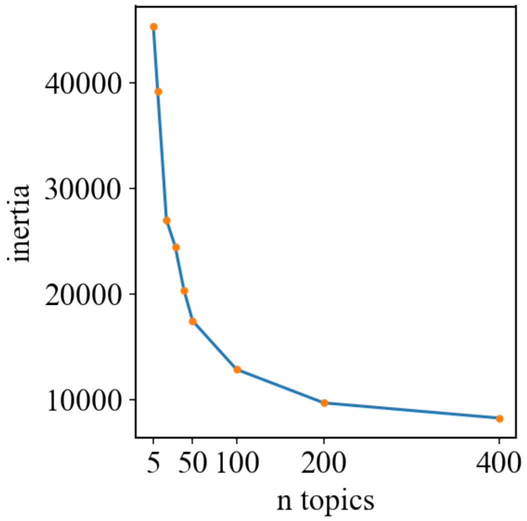


Figure 5: Inertia at different numbers of target topic centroids.

trained model contains 426 topics, but we reduce it to 100 topics (after using the elbow method to find the optimum number of topics, see Figure 5), using the hierarchical topic reduction function in Top2Vec, which finds the representative topics of the corpus by iteratively merging each smallest topic to the most similar topic until the number of 100 topics is reached. The model provides two relevant outputs, which are the vector representations of the documents, and the cosine similarities of the documents towards the topic centroids. Concerning the latter, it is important to note that Top2Vec, being a geometrical model, differs here from a probabilistic model such as LDA. The 'weight' is therefore not a probability between 0 and 1, but the cosine similarity between a document and a topic, which is a value between -1 and $+1$.

Choosing prototypical events

As was mentioned earlier, there is a sharp difference in the number of primitives across years in our corpora. In order to alleviate this problem, we pick a single 'prototypical' document for each day if there are multiple documents tied to that day. To acquire prototypes, we first group doc2vec document embeddings (acquired in the previous step) into daily subsets. For each subset, we then calculate pair-wise distances between embeddings. The embedding with the lowest average distance to all the other in the subset is then picked as the prototype. The distance metric used is cosine distance:

$$D_C = 1 - \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

Hereby we aim to capture the document that is most similar to other documents in a daily subset. For example, if multiple documents refer to the same event on the same day, only one will be picked to represent it. Furthermore, this step allows us to regularize the interval of

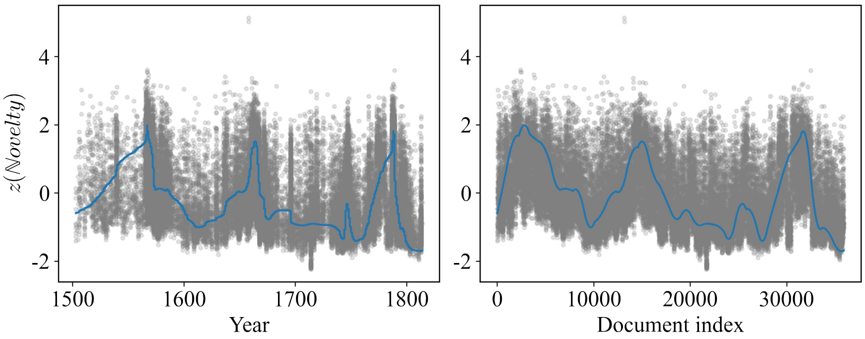


Figure 6: Novelty signal *without* choosing prototypes (calculated using the whole `corrected daily` subset of the corpus). The left graph shows novelty with date on the x-axis, while the right graph shows documents by index (in chronological order, but increments do not correspond to time intervals).

measurement (time elapsed between datapoints). Regular intervals are important for choosing the window (w) parameter in novelty detection, as well as interpreting the resulting novelty values; After choosing prototypical events, a primitive with high novelty can be considered novel in the context of w or more days, and not just an eventful afternoon with w records (e.g. Stadtholder Willem V visiting Purmerend).

Furthermore, a sanity check in which we did not choose prototypical primitives revealed that the novelty peaks remain practically unchanged. It is therefore very unlikely that the peaks are driven by the picking of irregular documents as prototypes.

Novelty detection

The following measures are calculated on 300-D `doc2vec` embeddings of the chosen prototypical events, ordered by date. First, embeddings are turned into a probability distribution using the softmax function:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2}$$

We then proceed calculate novelty, transience and resonance. With *novelty*, we refer to an event $s^{(j)}$'s reliable difference from past events $s^{(j-1)}$, $s^{(j-2)}$, ..., $s^{(j-w)}$ in window w:

$$\mathbb{N}_w(j) = \frac{1}{w} \sum_{d=1}^{w} JSD(s^{(j)} \mid s^{(j-d)}) \tag{3}$$

and *resonance* as the degree to which future events $s^{(j+1)}$, $s^{(j+2)}$, ..., $s^{(j+w)}$ conform to an event $s^{(j)}$'s novelty:

$$\mathbb{R}_w(j) = \mathbb{N}_w(j) - \mathbb{T}_w(j) \tag{4}$$

where $\mathbb{T}$ is the *transience* of $s^{(j)}$:

$$\mathbb{T}_w(j) = \frac{1}{w} \sum_{d=1}^{w} JSD(s^{(j)} \mid s^{(j+d)}) \tag{5}$$

This model for novelty and resonance was originally proposed in [3], but here we use the symmetrized and smooth version with the Jensen-Shannon divergence (JSD):

$$JSD(s^{(j)} \mid s^{(k)}) = \frac{1}{2} D(s^{(j)} \mid M) + \frac{1}{2} D(s^{(k)} \mid M) \tag{6}$$

where $M = \frac{1}{2}(s^{(j)} + s^{(k)})$ and D signifies Kullback–Leibler divergence [3, 18]:

$$D(s^{(j)} \mid s^{(k)}) = \sum_{i=1}^{K} s_i^{(j)} \log_2 \left(\frac{s_i^{(j)}}{s_i^{(k)}} \right) \tag{7}$$

In this case, Jensen-Shannon divergence is preferred over Kullback–Leibler divergence for a number of reasons. First, we maintain that JSD allows us to relax assumptions about the temporal order of observations, as it is a symmetric metric (meaning $JSD(P|Q) = JSD(Q|P)$ for probability distributions P and Q) [18, 12]. Information in our dataset is not always presented in a strictly chronological way: both events happening over a range of dates, and 'flashbacks' (recollections of past events presented out of order at a future date) are examples of cases where attributing a temporal order would be problematic. Second, the calculated JS divergences are a smoother version of KL divergences, with the maximum possible difference between probability distributions P and Q being 1 (if a base-2 logarithm is used). This propriety makes some downstream tasks such as peak detection easier, because extreme values will not be orders of magnitude greater than the mean (and therefore an extra normalization step is not required).

Nonlinear adaptive filtering

Nonlinear adaptive filtering is applied to the information signals because of the their inherent noisiness [6]. First, the signal is partitioned into segments (or windows) of length $w = 2n + 1$ points, where neighboring segments overlap by $n + 1$. The time scale is $n + 1$ points, which ensures symmetry. Then, for each segment, a polynomial of order D is fitted. Note that $D = 0$ means a piece-wise constant, and $D = 1$ a linear fit. The fitted polynomial for ith and $(i + 1)th$ is denoted as $y^{(i)}(l_1), y^{(i+1)}(l_2)$, where $l_1, l_2 = 1, 2, \ldots, 2n + 1$. Note that the length of the last segment may be shorter than w. We use the following weights for the overlap of two segments.

$$y^{(c)}(l_1) = w_1 y^{(i)}(l + n) + w_2 y^{(i)}(l), \quad l = 1, 2, \ldots, n + 1 \tag{8}$$

where $w_1 = (1 - \frac{l-1}{n})$, $w_2 = 1 - w_1$ can be written as $(1 - \frac{d_j}{n})$, $j = 1, 2$, where d_j denotes the distance between the point of overlapping segments and the center of $y^{(i)}, y^{(i+1)}$. The weights decrease linearly with the distance between point and center of the segment. This ensures that the filter is continuous everywhere, which ensures that non-boundary points are smooth.